H24/3041

MONASH UNIVERSITY
THESIS ACCEPTED IN SATISFACTION OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
ON.... 3 August 2001 ......
.... [redacted]
Sec. Ph.D. and Scholarships Committee

Under the copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing for the purposes of research, criticism or review. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

Addendum

Page 22, 1st sentence should read: The nucleus is a charged particle which has a magnetic moment, μ, and a spin. Applying a magnetic field removes the degeneracy of the spin state.

Page 26, last paragraph, 2nd sentence should read: An amide group, a derivative of an amino acid consists of an amino group, a carboxyl group and a side chain (R) bound to the carbon atom, Figure 1.4.

Pages 27 and 28, Text should refer to symmetric (a) and asymmetric (b) phosphate stretching vibrations. Caption for Figure 1.7 should read: Schematic representation of the symmetric (a) and asymmetric (b) phosphate stretching vibrations of phosphate.

Page 29, Table 1.2 and page 142, Figure 142. In reference to the absorption band arising ~1080 cm^{-1}, this band is an overlap from contributions of phosphodiester linkages in nucleic acids as well as C-O vibrations from glycogen.

Page 32, 2nd paragraph, 3rd sentence should read: This experiment remained inconclusive because diffraction effects for an infrared microscope fitted with a mercury-cadmium-telluride (MCT) normally prevents the use of physical apertures no less than 10 µm with the use of a 15*x* objective.

Page 63, equation 2.12 should read: $A = -\log_{10}\dfrac{I}{I_0} = \log_{10}(1/T)$

Page 114, 1st sentence should read: A minimum of six transmission spectra were recorded for each sample with the effective aperture reduced to 50 x 50 µm.

Page 116. PCA was utilised to determine which samples were likely to be a mixture of endocervical and ectocervical cells. This was achieved by performing the analysis of spectra known to contain only ectocervical or endocervical cells. These samples were inspected visually and either kept or discarded.

Page 140, Section 5.1.1.4, 1st sentence: data points should read data point spacing.

Page 167, last sentence should read: The high reproducibility of these spectra post clean-up, as compared with spectra pre clean-up, is further illustrated in Figure 5.22.

Page 159, Figure 5.14

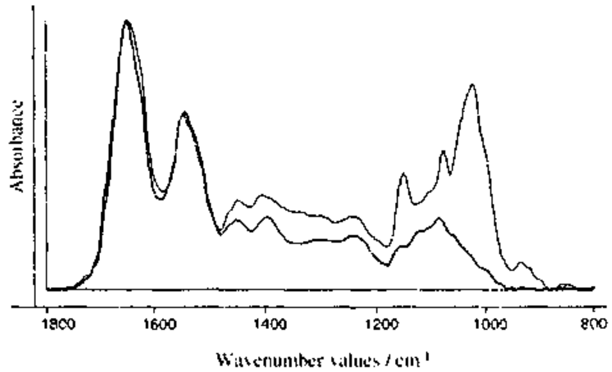

Figure 5.14 Infrared spectra of HeLa cells (blue) and normal squamous epithelial cells (black).

INFRARED MICROSPECTROSCOPY AND MULTIVARIATE STATISTICS IN THE DIAGNOSIS OF CERVICAL CANCER

Melissa Joanne Romeo, B.Sc. (Hons)
School of Chemistry, Monash University

A dissertation submitted in accordance with the requirements for the
Degree of Doctor of Philosophy

December 2000

Contents

1 An introduction to cervical cancer and diagnostic techniques ... 1

1.1 Carcinoma of the cervix ... 1

1.1.1 George Papanicolaou and the cervical smear ... 1

1.1.2 The Papanicolaou (Pap) or cervical smear ... 2

1.1.2.1 The squamocolumnar junction and transformation zone ... 4

1.1.2.2 Staining ... 5

1.1.2.3 The Bethesda system for reporting cytological diagnoses of cervical smears ... 5

1.1.3 Histology of the female genital tract ... 6

1.1.4 Carcinogenesis ... 7

1.1.5 Cytology of neoplasia and cervical lesions ... 8

1.1.6 Human Papilloma Virus, smoking and cervical cancer ... 10

1.1.7 The accuracy of the Pap smear ... 12

1.1.8 Gold standards ... 13

1.2 Adjuncts to conventional cervical smear screening ... 14

1.2.1 PAPNET® ... 14

1.2.2 ThinPrep® ... 15

1.2.3 Cervicography ... 17

1.3 Non-conventional approaches to cervical cancer diagnosis ... 18

1.3.1 Polarprobe ... 18

1.3.2 Spectroscopy ... 19

1.3.2.1 Raman spectroscopy ... 19

1.3.2.2 Magnetic resonance spectroscopy ... 21

1.3.2.3 Fluorescence spectroscopy ... 25

1.3.3 Infrared spectroscopy ... 26

1.3.3.1 Characteristic vibrations of interest ... 26

1.3.3.2 Infrared spectroscopy in the diagnosis of cancer ... 29

1.3.3.3 Characteristic infrared spectral changes between normal and malignant cervical cells ... 29

1.3.3.4 The role of multivariate statistics ... 33

1.4 The importance of identifying potential confounding variables ... 35

1.4.1 The menstrual cycle ... 36

1.4.1.1 Cytology of the ovulatory cycle ... 37

1.4.2 Endocervical cells ... 38

1.4.3 Benign cellular changes (BCCs) ... 39

1.4.3.1 Inflammation, lymphocytes and other blood components ... 39

1.4.3.2 Metaplasia ... 40

1.4.4 Mucins ... 41

1.4.5 Connective tissue ... 41

1.4.6 Semen ... 42

1.4.7 Fixatives ... 42

1.5 A new diagnostic technique for cervical cancer ... 42

1.6 References ... 45

2 An introduction to infrared spectroscopy and multivariate statistics ... 56

2.1 Theory of infrared spectroscopy ... 56

2.1.1 The dipole moment ... 57

2.1.2 Rotation ... 58

2.1.3 Vibration ... 59

2.1.4 Vibrational modes ... 61

2.1.5 ABSORPTION OF LIGHT. 62
2.2 THE INFRARED SPECTROMETER.. 63
2.2.1 MICHELSON INTERFEROMETER 64
2.2.2 FOURIER TRANSFORMATION 64
2.2.3 SOURCES AND DETECTORS 67
2.2.4 NOISE 67
2.2.4.1 SOURCES OF INSTRUMENTAL NOISE 68
2.2.4.2 SIGNAL-TO-NOISE ENHANCEMENT 68
2.3 FTIR MICROSCOPE. 69
2.4 INSTRUMENTAL PARAMETERS 70
2.5 THEORY OF MULTIVARIATE STATISTICS 72
2.5.1 PATTERN RECOGNITION 72
2.5.2 CLASSIFICATION 72
2.5.3 K-NEAREST NEIGHBOURS (K-NN) 73
2.5.4 DISCRIMINANT ANALYSIS 75
2.5.5 PRINCIPAL COMPONENT ANALYSIS (PCA) 77
2.5.6 SOFT INDEPENDENT MODELLING OF CLASS ANALOGY (SIMCA) 80
2.5.7 ARTIFICIAL NEURAL NETWORKS (ANNS) 83
2.5.7.1 NEURODES 83
2.5.7.2 BACK PROPAGATION 87
2.5.7.3 NETWORK OPTIMISATION 89
2.5.7.4 BAYESIAN REGULARISED NEURAL NETWORKS 90
2.6 REFERENCES 92

3 MATLAB ROUTINES FOR PRE-PROCESSING IR SPECTRA 95

3.1 MAIN MENU 96
3.1.1 LOAD EXISTING DATABASE 96
3.1.2 START NEW DATABASE 96
3.1.3 ADD TO EXISTING DATABASE 97
3.2 PRE-PROCESSING MENU 97
3.2.1 DEFINE SPECTRAL REGION 97
3.2.2 ACCOUNT FOR NONLINEARITIES OF THE MCT 98
3.2.3 SIGNAL-TO-NOISE RATIO (SNR) 99
3.2.4 DERIVATIVES (SAVITZKY-GOLAY) 101
3.2.5 NORMALISATION 102
3.2.5.1 VECTOR NORMALISATION 103
3.2.5.2 MAXIMUM NORMALISATION 103
3.2.5.3 MEAN NORMALISATION 103
3.2.5.4 RANGE NORMALISATION 103
3.2.6 BASELINE CORRECTION 104
3.2.7 PLOT DATA 105
3.2.8 ASSIGN DIAGNOSIS 105
3.2.8.1 NORMAL/ABNORMAL 107
3.2.8.2 ABS NORMAL/ABNORMAL 108
3.2.8.3 BIOPSY/CYTOLOGY 108
3.2.8.4 NORMAL/DYSPLASIA 108
3.2.8.5 ABS NORMAL/NORMAL/ABNORMAL 108
3.2.9 VIEW DATABASE DETAILS 108
3.3 MULTIVARIATE ANALYSIS 108
3.4 REFERENCES 110

4 AN INVESTIGATION INTO THE INFLUENCE OF HORMONES AND NUCLEIC ACIDS IN THE DIAGNOSIS OF CERVICAL CANCER 111

4.1 HORMONAL STUDY 111
4.1.1 METHODOLOGY 112

4.1.1.1 PARTICIPANTS 112
4.1.1.2 SAMPLE COLLECTION 113
4.1.1.3 SAMPLE PREPARATION 113
4.1.1.4 MULTICAVITY INFRARED CELL 113
4.1.1.5 INFRARED MICROSPECTROSCOPY 113
4.1.1.6 DATA TREATMENT 114
4.1.1.7 BLOOD PREPARATION 114
4.1.2 RESULTS AND DISCUSSION 114
4.1.2.1 PROGESTERONE ASSAY 114
4.1.2.2 CONTAMINATION 115
4.1.2.3 SPECTRAL SUBTRACTION 116
4.1.2.4 ECTOCERVICAL CELLS 116
4.1.2.5 NAVICULAR CELLS 119
4.1.2.6 ENDOCERVICAL CELLS 120
4.1.2.7 NUCLEIC ACIDS 122
4.1.3 CONCLUSION 124
4.2 AN INVESTIGATION INTO NUCLEIC ACID CONTRIBUTIONS IN THE IR SPECTRA OF NORMAL CERVICAL EPTITHELIAL AND HELA CELLS 124
4.2.1 METHODOLOGY 128
4.2.1.1 PREPARATION OF HOMOGENATE 128
4.2.1.2 ISOLATION OF NUCLEI 128
4.2.2 RESULTS AND DISCUSSION 129
4.2.3 CONCLUSION 135
4.3. REFERENCES 136

5 A DIAGNOSTIC TECHNIQUE FOR CERVICAL CANCER 139

5.1 PRELIMINARY STATISTICAL ANALYSIS 139
5.1.1 METHODOLOGY 139
5.1.1.1 SAMPLE COLLECTION 139
5.1.1.2 SAMPLE PREPARATION 140
5.1.1.3 INFRARED SPECTROSCOPY 140
5.1.1.4 DATA TREATMENT 140
5.1.1.5 SOFTWARE 141
5.1.2 RESULTS AND DISCUSSION 141
5.1.2.1 CONTAMINATION 142
5.1.2.2 PRINCIPAL COMPONENT ANALYSIS 142
5.1.2.3 K-NEAREST NEIGHBOURS 144
5.1.2.4 SOFT INDEPENDENT MODELLING OF CLASS ANALOGY 145
5.1.2.5 ARTIFICIAL NEURAL NETWORKS 148
5.1.2.6 LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS 149
5.1.3 CONCLUSION 152
5.2 FINAL STATISTICAL ANALYSIS 152
5.2.1 METHODOLOGY 152
5.2.1.1 DATABASE 152
5.2.2 RESULTS AND DISCUSSION 153
5.2.2.1 PRINCIPAL COMPONENT ANALYSIS 153
5.2.2.2 BAYESIAN REGULARISED ARTIFICIAL NEURAL NETWORKS 157
5.2.3 CONCLUSION 161
5.3 CLEAN-UP OF CERVICAL SMEARS 162
5.3.1 METHODOLOGY 163
5.3.1.1 SAMPLE COLLECTION AND PREPARATION 163
5.3.1.1.1 THINPREP® 163
5.3.1.1.2 RED AND WHITE CELL LYSIS BUFFER 163
5.3.1.2 CYTOLOGICAL ANALYSIS 164
5.3.2 RESULTS AND DISCUSSION 164

5.3.2.1 THINPREP® 164
5.3.2.2 BLOOD CELL LYSIS BUFFERS 165
5.3.3 CONCLUSION 173
5.4 CONFOUNDING VARIABLES AND MULTIVARIATE STATISTICS 173
5.4.1 METHODOLOGY 175
5.4.1.1 SAMPLE COLLECTION 175
5.4.1.2 DATA TREATMENT 175
5.4.2 RESULTS AND DISCUSSION 176
5.4.2.1 ENDOCERVICAL CELLS 176
5.4.2.2 INFLAMMATION 182
5.4.2.3 CANDIDA ALBICANS 185
5.4.2.4 BACTERIAL VAGINOSIS 188
5.4.2.5 OTHER BENIGN CELLULAR CHANGES 190
5.4.3 CONCLUSION 191
5.5 REFERENCES 192

6 CONCLUSIONS AND FUTURE DIRECTIONS 195

APPENDIX A 198

GLOSSARY 198

APPENDIX B 201

LIGHT MICROSCOPY OF CERVICAL SMEARS 201

APPENDIX C 205

GAUSSIAN DISTRIBUTION 205

APPENDIX D 206

MATLAB ROUTINES 206
CERVJOIN.M 206
JCAMPDBJOIN.M 209
JCAMPCERVJOIN.M 212
DEFREGJOIN.M 214
NONLINJOIN.M 216
SNRJOIN.M 217
SVGLJOIN.M 220
VNORMJOIN.M 222
MNORMJOIN.M 223
MEANORMJOIN.M 224
RANGENORMJOIN.M 225
BLJOIN.M 226
DIAGNOSISRWH.M 229
DIAGRWHNA.M 231
NORMAB1.M 233
DIAGNOSISFPV.M 234
LDACERV.M 236
QDACERV.M 242

APPENDIX E 249

JCAMP.DX FORMAT 249

APPENDIX F 250

RECRUITMENT POSTER 250
EXPLANATORY STATEMENT 251
CONSENT FORM 252

APPENDIX G **253**

PROCEDURE FOR REMOVING HELA CELLS FROM A MONOLAYER SUSPENSION 253
PREPARATION OF EHRLICH'S HAEMOTOXYLIN 253
PREPARATION OF SOLUTIONS NEEDED FOR SUBCELLULAR FRACTIONATION 253

APPENDIX H **254**

THE NITROGEN BOMB 254

APPENDIX I **255**

MAKE COMPATIBLE 255

APPENDIX J **256**

CONSENT FORM AND EXPLANATORY STATEMENT FOR RWH 256
CONSENT FORM, EXPLANATORY STATEMENT AND QUESTIONNAIRE FOR FPV 258

APPENDIX K **261**

INTERPRETATION OF SIMCA RESULTS [12] 261
  SAMPLE TO MODEL DISTANCE 261
  MODEL DISTANCE 261
  DISCRIMINATION POWER 261
  MODELLING POWER 262
  OUTLIER DETECTION 262

REFERENCES (APPENDICES) **263**

LIST OF FIGURES

1.1 A cervical smear is obtained by insertion of a speculum, which widens the vagina to reveal the cervix 3
1.2 Schematic representation of the squamocolumnar junction and transformation zone. 4
1.3 Schematic representation of the scattering effects possible with the induction of polarised states 20
1.4. Schematic representation of the amide functional group 27
1.5. Schematic representation of the vibrations contributing to the amide I infrared mode. 27
1.6. Schematic representation of the vibrations contributing to the amide II infrared mode. 27
1.7. Schematic representation of the asymmetric and symmetric stretching vibrations of phosphate. 28
1.8 The reproductive cycle of the human female 36
2.1 The electromagnetic spectrum 56
2.2 Plane polarised electromagnetic radiation 56
2.3 The rigid rotor 58
2.4 A model for the vibration of a diatomic molecule 59
2.5 Harmonic oscillator potential well 60
2.6 The anharmonic oscillator potential and harmonic oscillator 61
2.7 Possible stretching (a) and bending (b) vibrational modes of infrared absorption 62
2.8 Schematic diagram of a Michelson interferometer 64
2.9 Apodisation functions and resulting instrumental lineshapes 66
2.10 Signal-to-noise ratio 67
2.11 General layout of an FTIR microscope 70
2.12 Schematic of the Bruker IFS 55 Spectrometer. 71
2.13 Schematic representation of the circle with radius r about an unclassified object containing three nearest neighbours 74
2.14 Plot of observations on two variables X_1, X_2 and the same observations plotted with respect to their principal components PC_1, PC_2. 78
2.15 Schematic representation of the construction of a class box around two classes 81
2.16 Illustration of effect of different probability levels for deciding the boundary between a class and variable space 81
2.17 Upper and lower limits of principal component scores enables the model to be closed in variable space 82
2.18 Schematic representation of the generation of an output from a neurode. 84
2.19 Schematic representation of a fully connected feed-forward neural network 86
2.20 Schematic presentation of weight correction by back-propagation of errors 87
2.21 Schematic illustration of the training and validation errors as a function of the iteration step 89
3.1 Main menu for the Matlab routine *Cervjoin.m* 96
3.2 Load menu for the Matlab routine *Cervjoin.m*. 96
3.3 Pre-processing menu for the Matlab routine *Cervjoin.m*. 97
3.4 Define region menu for the Matlab routine *Defregjoin.m*. 98
3.5 Spectral regions obtained through define spectral region option 98
3.6 Illustration of the removal of nonlinearity effects. 99
3.7 Matlab plots representing the steps involved in the calculation of the RMS noise component for calculation of SNRs 100

3.8 Illustration of derivative techniques 102
3.9 Illustration of normalisation techniques 104
3.10 Illustration of the baseline correction process. 105
3.11 Diagnosis menu for the Matlab routine *Cervjoin.m* 107
3.12 Multivariate statistics menu for the Matlab routine *Cervjoin.m*. 109
4.1 Multicavity IR cell 113
4.2 A Spectrum of ectocervical cells contaminated with Ayre spatula debris, uncontaminated ectocervical cells and a spectrum of Ayre spatula debris 115
4.3 Infrared spectra obtained from a woman not taking oral contraception 117
4.4 Infrared spectra of a woman, not taking oral contraception, collected over four cycles. 118
4.5 Infrared spectra of a woman taking monophasic oral contraception 118
4.6 Scores plots showing discrimination between normal and dysplastic cervical cells including menstrual data 119
4.7 Microscopic image of a cervical smear showing the presence of navicular cells 120
4.8 Infrared spectra of ectocervical cells exhibiting a high concentration of glycogen . 120
4.9 Infrared spectra of endocervical cells from a woman not taking oral contraception 121
4.10 Infrared spectra of ectocervical cells obtained from a woman not taking oral contraception highlighting the phosphodiester region. Part I 122
4.11 Infrared spectra of ectocervical cells obtained from a woman not taking oral contraception highlighting the phosphodiester region. Part II 123
4.12 Infrared spectra obtained from a woman not taking oral contraception, highlighting the amide II band. 124
4.13 Chemical structure of iodixanol 127
4.14 Schematic representation of the purification of nuclei from mammalian cells 129
4.15 Infrared spectra of whole cells and cellular material believed to be nuclei collected from the 30/35% OptiPrep™ interface 132
4.16 Microscopic image of isolated nuclei obtained from the 30/35% OptiPrep™ interface. 132
5.1 Averaged infrared spectra of cervical cells exhibiting various grades of abnormalities compared with normal cervical cells. 142
5.2 Loadings plots of the first three principal components, illustrating the variance of 501 wavenumber values 143
5.3 A Scores plot illustrating the discrimination obtained between IR spectra of normal and abnormal cervical cells 144
5.4 Coomans plots resulting from classification of unknown samples based on models formed from normal and abnormal data 147
5.5 Coomans plot showing the effects of pre-processing data using different algorithms... 148
5.6 Results of linear and quadratic discriminant analysis 151
5.7 Results of linear and quadratic discriminant analysis using non-subjectively chosen data 151
5.8 PC1 versus PC2 scores plot of normal and abnormal high-grade dysplasia, CIS and SCC samples. 154
5.9 PC1 versus PC2 scores plot of normal and abnormal cells diagnosed by cytology. 154
5.10 PC1 versus PC2 scores plot of averaged normal and abnormal spectra diagnosed by cytology. 155
5.11 PC1 versus PC2 scores plot of normal and abnormal samples illustrating the effects on clustering of normalisation techniques. 156

5.12 Loadings plots of the first 3 principal components resulting from the PCA of maximum, vector, mean and range normalisation techniques. 156
5.13 Plots comparing the predicted outputs of the Bayesian regularised artificial neural networks with the expected outputs 158
5.14 Infrared spectra of HeLa cells and normal squamous epithelial cells 159
5.15 PC1 versus PC2 scores plot of the infrared spectra of cervical samples diagnosed by cytology as normal, normal but exhibiting signs of inflammation, other blood components, bacterial infection and benign cellular changes, including metaplasia, CIN I and HPV effects and abnormal (CIN II, III, CIS and SCC, including HPV effects). 160
5.16 Infrared spectrum of the ThinPrep® PreservCyt™ solution after desiccation. 164
5.17 Infrared spectrum of red cell lysis buffer after desiccation. 165
5.18 Infrared spectrum of cervical cells contaminated with red cell lysis buffer. 166
5.19 Infrared spectrum of white cell lysis buffer after dessication 166
5.20 Infrared spectra showing the results of the clean-up of cervical samples. Part I 167
5.21 Light microscopy slides of cervical smears used in the clean-up process 168
5.22 Infrared spectra showing the results of the clean-up of cervical samples. Part II 168
5.23 Scores plot of distribution of clean-up samples before and after the clean-up process.. 169
5.24 Scores plot of normal and abnormal samples,including after clean-up samples 171
5.25 Averaged IR spectra of normal and abnormal ectocervical and normal endocervical cells. 175
5.26 Infrared spectrum of normal endocervical cells and abnormal ectocervical cells 177
5.27 PC1 versus PC2 scores plot showing a separation between normal ectocervical cells, abnormal ectocervical cells and normal endocervical cells 178
5.28 Loadings plots of the first three principal components 178
5.29 PC1 versus PC2 scores plot of data in the regions 1096-1062 cm^{-1} and 1060-994 cm^{-1} of normal ectocervical, abnormal ectocervical and normal endocervical 179
5.30 PC1 versus PC2 scores plot of data in the region 1096-1062 cm^{-1} showing a separation between normal ectocervical and endocervical cells and between abnormal ectocervical and normal endocervical cells 180
5.31 PC1 versus PC2 scores plot of data in the region 1060-994 cm^{-1} showing a separation between normal ectocervical and endocervical cells and between abnormal ectocervical and normal endocervical cells 181
5.32 Averaged infrared spectra of smears containing normal ectocervical, abnormal ectocervical cells and inflammatory effects 185
5.33 PC1 versus PC2 scores plot showing a separation between normal diagnosed smears and abnormal diagnosed smears with inflammation in the region 1800-800 cm^{-1}... 183
5.34 PC1 versus PC2 scores plot showing a separation between normal diagnosed smears and abnormal diagnosed smears with inflammation in the region 1096-1062 cm^{-1}. 184
5.35 PC1 versus PC2 scores plot showing a separation between normal diagnosed smears and abnormal diagnosed smears with inflammation in the region 1060-994 cm^{-1}... 184
5.36 Averaged infrared spectra of normal, abnormal and *Candida* diagnosed smears. ... 186
5.37 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with *Candida* diagnosed smears in the 1800-800 cm^{-1} region. 186
5.38 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with *Candida* diagnosed smears in the 1096-1062 cm^{-1} region. 187

5.39 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with *Candida* diagnosed smears in the 1060-994 cm^{-1} region. 187
5.40 Averaged infrared spectra of normal and abnormal diagnosed cervical smears with smears diagnosed with bacterial vaginosis. 188
5.41 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with bacterial vaginosis diagnosed smears in the 1800-800 cm^{-1} region. 189
5.42 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with bacterial vaginosis diagnosed smears in the 1096-1062 cm^{-1} region. 190
5.43 PC1 versus PC2 scores plot showing a separation of normal diagnosed smears and abnormal diagnosed smears with bacterial vaginosis diagnosed smears in the 1060-994 cm^{-1} region. 190
5.44 PC1 versus PC2 scores plot of normal cervical smears and smears diagnosed with metaplasia, atypia and keratinisation 191

APPENDICES

B.1 Parabasal cells 201
B.2 Intermediate cells 201
B.3 Superficial cells 201
B.4 Endocervical cells 202
B.5 CIN I (mild dysplasia) 202
B.6 CIN II (moderate dysplasia) 202
B.7 CIN II (moderate dysplasia) with bacterial infection 202
B.8 CIN III (severe dysplasia) 203
B.9 CIS (carcinoma *in situ*) 203
B.10 Invasive SCC (squamous cell carcinoma) 203
B.11 Invasive undifferentiated carcinoma 203
B.12 Adenocarcinoma 204
B.13 Koilocytes caused by HPV 204
B.14 Squamous metaplasia 204
B.15 Atrophy 204
C.1 Standardised Gaussian probability curve 205
H.1 Nitrogen bomb 254
K.1 Coomans plot interpretation 261

LIST OF TABLES

1.1 Example of validity measures of a screening test 13
1.2 Infrared band assignments of biological molecules 28
2.1 Infrared spectral regions 56
3.1 Numerical codes and corresponding diagnosis assigned to spectra 106
4.1 Summary of participation in the hormonal study. 114
4.2 Physical properties of iodixanol in water (20°C) 127
5.1 Results of *K*-NN with *K* values ranging from *K* = 1 to *K* = 9. 145
5.2 Results of SIMCA analysis 146
5.3 Results of ANN analysis 149
5.4 Results of discriminant analysis. 150
5.5 Results of applying a BRANN to the infrared spectra of cervical samples. 157
5.6 Summary of the cytological diagnoses use to form four groups of infrared spectra for PCA 160
5.7 Diagnostic results for samples used in clean-up experiment 169
5.8 Biopsy and Cytology diagnosis from clean-up samples, indicating the presence of an endocervical component and inflammatory exudate. 172
5.9 Histology and cytology results for the cervical smears of 24 samples used in the clean-up process 172
5.10 Sensitivity and specificity of SIMCA, LDA and *K*-NN in the classification of cervical smears principally containing ectocervical cells 174
5.11 Number of IR spectra of each cell or diagnostic type used for PCA. 175
5.12 Presence of leukocytes in cervical smears in relation to hormonal stimulation. 185

ABSTRACT

This dissertation investigates the use of infrared (IR) microspectroscopy and multivariate statistics in the diagnosis of cervical cancer. During the course of the candidature a significant number of IR spectra of epithelial cells from cervical smears were collected. These spectra had either a cytological and/or a histological diagnosis. Multivariate statistical techniques were investigated as a means of objectively diagnosing IR spectra and also to determine the effects of confounding variables.

A pre-processing routine was written in Matlab code to allow bulk spectral pre-processing to be performed within one program, thus saving time and the inconvenience of importing and exporting data between existing programs. The pre-processing routine objectively removes spectra exhibiting nonlinearity effects or noise outside an acceptable limit, as well as performing baseline correction, normalisation and derivative calculations.

The spectroscopic effects of hormonal influences on cervical epithelium were investigated. IR spectra of ectocervical and endocervical cells were obtained from women taking monophasic oral contraception and women not taking oral contraception. IR spectra reflected the cyclical changes occurring in the squamous epithelium. These are seen by an increase in the glycogen peak at 1025 cm^{-1} towards ovulation, and a subsequent decrease following ovulation. The contribution of nucleic acids to these also appeared cyclically related. As expected the IR spectra of ectocervical cells obtained from women taking monophasic contraceptives did not exhibit cyclical variation. IR spectra of endocervical cells from both groups did not exhibit cyclical variation. Principal component analysis indicates that cyclical variation is not a confounding variable in the spectroscopic diagnosis of cervical cancer.

IR spectra of isolated nuclei were recorded to determine the contribution of nucleic acids to the spectra of cervical epithelial cells. IR spectra of isolated nuclei exhibited no features characteristic of nucleic acids.

Principal component analysis (PCA), soft independent modelling of class analogy (SIMCA), *K*-nearest neighbours (*K*-NN), linear and quadratic discriminant analysis and artificial neural networks (ANNs) were investigated for their potential to classify and predict IR spectra of cervical cells diagnosed by cytology and histology as normal or abnormal (high-grade dysplasia or malignancy). Bayesian regularised (BR) ANNs

performed the best out of these techniques in the preliminary analysis of a subsection of the collected data. When more data was included and cytological diagnosis was used as the expected output, the BRANN performed poorly, and was unable to train or predict the IR spectra of normal and abnormal diagnosed cervical cells.

The presence of non-epithelial cells in cervical smears, as well as benign changes of cells is a problem in spectroscopic analysis. Chemical removal of blood components from abnormal diagnosed cervical smears was investigated. Lymphocytes were successfully removed from samples, demonstrated through light and IR microscopy, however IR spectra obtained from some abnormal samples after the removal of lymphocytes exhibited spectral features similar to those of normal diagnosed cervical cells. PCA was performed on the spectra of samples before and after the clean-up process (compared with normal and abnormal diagnosed spectra from the databank). The majority of post clean-up samples were grouped with normal diagnosed samples.

The influence of endocervical cells and benign cellular changes associated with inflammation, bacterial and yeast infections were investigated. Spectral differences between these groups from normal and abnormal samples were exhibited in the phosphodiester and carbohydrate regions. PCA was able to separate normal ectocervical cells from normal endocervical cells, and samples diagnosed with inflammation, *Candida* or bacterial vaginosis. Abnormal ectocervical cells were separated from normal endocervical cells, and samples diagnosed with inflammation or bacterial vaginosis.

The work embodied in this dissertation was performed during 1997 – 2000 in the School of Chemistry, Monash University. It contains no material which has been accepted for the award of any other degree or diploma in any university and, to the best of my knowledge, no material previously published or written by another person, except where due reference is made in the text of the dissertation.

[redacted]

Melissa Romeo

The following publications are related to the work presented in this dissertation:

M. J. Romeo, F. Burden, M. Quinn, B. Wood and D. McNaughton, Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer., *Cellular and Molecular Biology*, **44**(1): 179-187. 1998.

B. R. Wood, M. Quinn, B. Tait, M. Ashdown, T. Hislop, M. J. Romeo and D. McNaughton, A FTIR microspectroscopic study of cell types and potential confounding variables in screening for cervical malignancies, *Biospectroscopy*, 4(2): 75-91. 1998.

M. J. Romeo, B. Wood and D. McNaughton, Observing the cyclical changes in cervical epithelium using infrared microspectroscopy, *Vibrational Spectroscopy*, in press.

All work embodied in this thesis was performed personally by the author except where acknowledgment of assistance is indicated under the Section entitled "Acknowledgments".

ACKNOWLEDGMENTS

There are many people I am indebted to for their help in making my PhD possible. First of all I would like to thank my supervisor Dr. Don McNaughton. Thank you for sharing your spectroscopic knowledge, guidance, advice and for the daily chats towards the end. Cheers. My co-supervisor Prof. Michael Quinn for advice from a gynaecological perspective. Prof. Frank Burden, one of my chemometrics gurus, for your guidance with Matlab programming, sense of humour, support, encouragement and motivation.

Dr Elizabeth McNaughton and Dr Chris Blandy for donating their valuable time to perform the cervical smears used in the hormonal study. Dr Ben Canney for performing progesterone assays. I especially thank the women who participated in this study, without your generosity the investigation would not have been possible.

The people from Monash University who assisted with technical advice and support: Ian Boundy and Robyn Kelly-Young for preparation in slides, cytological analysis and your interest and enthusiasm in my project. Dr Brian Cooke for the use of your microscope and digital camera. Steve Morton for photography and scanning. Finlay Shanks for maintenance on the microscope, always being available to assist with problems and for allowing me to temporarily adopt the zip drive. Rod Hall for proof-reading and the interesting discussions in the third year laboratory. Don McGilvery for computer expertise and last minute assistance with printing.

To all the staff at the Royal Women's Hospital Dysplasia Clinic and Medical Records, Victorian Cytology Service and Family Planning Victoria whom I met along the way, thank you for your friendliness, advice and answers to my questions. In particular I would like to acknowledge Dr Geoff Tan, Vicki Pumo, and Debra Riech. Special mention goes to Julenne Hallo from the Royal Women's Hospital. Thank you for preparing thousands of smears for me. If it weren't for your help I would still be sitting at the microscope. Thank you also for your encouragement and support.

The guys in the McNaughton spectroscopy group: Dr Bayden Wood for your, advice, encouragement and pep talks, and along with Phil Heraud and Tony Eden for the wicked Upwey parties. Thank you also for providing the normal diagnosed spectra for the comparison of baseline effects in SIMCA. Dom Appadoo for being such an inspiration and for help and advice with programming, particularly with translating discriminant analysis equations into Matlab code. Mustafa Kansiz and Bernd Giese for, amongst many other things, the memories of Tokyo, Enschede and Amsterdam, ensuring life in the office is never boring and for your patience and consideration when I took over your computer and office in the last week. Fei-sy Ooi for organising group events and the fabulous contributions to morning tea, especially when I was vegan.

Thank you to my wonderful friends, I could write a chapter about each of you individually but will refrain. I hope you know how much I love you and how much you mean to me. Thankyou for your love, faith in me, your support and encouragement when things went askew and for all the experiences (good and bad) we have shared: Michelle Tolliday, Olivia Dwyer, Sophie Ping, Alexandra Colvin, Corey Evans, Kylie Butcher, Craig Kingham, Miranda Phealan, Ross Gray, Chris, Derek Huxley, Fiona Wallis, Helen Webb, Fiona Scholes, Terrance, Michel Cabrie, Andrew and Gabrielle Watkins, Michelle Goodman, Gavin Williams, Rachel, Jonathon, Ian Fenton, Stine Moe and Luke Schmidt. Renée Otmar who is well aware of the impact she has had in my life. Particularly for helping me to achieve something I thought impossible for nine years. Lynton Tideman, Sean Turnbull and Zoltan Deak for bailing me out of computer situations. Michelle Kumarich for proof-reading the entire manuscript, your comments and recommendations were very appreciated. Nathan Boyle, who is a fantastic flatmate, for the balcony discussions - they mean a lot to me.

Last but by no means least to my family. Mum and Dad for your unconditional love, support and faith in me. Steven, Selina, Kimberley, Darren and Jenny for taking holidays at very appropriate times, thanks for the use of your respective houses, and of course for your love and support.

ABBREVIATIONS

ADP	adenosine diphosphate
ANN	artificial neural network
ASCUS	atypical squamous cells of undetermined significance
ATP	adenosine triphosphate
BCC	benign cellular change
BP	back propagation
BRANN	Bayesian regularised artificial neural network
CIN	cervical intraepithelial neoplasia
CIS	carcinoma *in situ*
DFT	discrete Fourier transform
DNA	deoxyribonucleic acid
FN	false negative
FP	false positive
FPV	Family Planning Victoria
FT	Fourier transform
FTIR	Fourier transform infrared
HGEA	high-grade epithelial abnormality
HPV	Human Papilloma Virus
HSIL	high-grade squamous intraepithelial neoplasia
IR	infrared
KBr	potassium bromide
K-NN	*K*-nearest neighbours
KRS-5	potassium thallium iodide
LDA	linear discriminant analysis
LEEP	loop electrosurgical excision procedure

LGEA	low-grade epithelial abnormality
LSIL	low-grade squamous intraepithelial neoplasia
MCT	mercury cadmium telluride
MRI	Magnetic Resonance Imaging
MRS	magnetic resonance spectroscopy
N/C ratio	nuclear–to–cytoplasmic ratio
NIR	near infrared
NMR	nuclear magnetic resonance
Pap smear	Papanicolaou smear
PBS	phosphate buffered saline
PC	principal component
PCA	principal component analysis
PCr	phosphocreatine
PCR	polymerase chain reaction **and** principal component regression
PDE	phosphodiester
PE	Perkin Elmer
Pi	inorganic phosphate
PLS	partial least squares
PO_2^-	phosphate
PME	phosphomonoester
PMN	polymorphonuclear leukocyte
QDA	quadratic discriminant analysis
RBC	red blood cell
RBF	radial basis function
RCLB	red cell lysis buffer

RF	radiofrequency
RMS	root-mean-square
RNA	ribonucleic acid
RSD	residual standard deviation
RWH	Royal Women's Hospital
SCC	squamous cell carcinoma
SCJ	squamocolumnar junction
SET	standard error of training
SEP	standard error of prediction
SIL	squamous intraepithelial neoplasia
SIMCA	soft independent modelling of class analogy
SNR	signal–to–noise ratio
TN	true negative
TP	true positive
UV	ultraviolet
UVRR	ultraviolet resonance Raman
VCS	Victorian Cytology Service
WBC	white blood cell
WCLB	white cell lysis buffer

SPECTROSCOPIC NOTATION

A	absorbance
B	magnetic field of electromagnetic radiation
c	speed of light
D_e	the depth of the potential well for an anharmonic oscillator
D_o	dissociation energy
δ	deformation mode **and** chemical shift
δ	optical path difference
$\delta\overline{v}$	resolution
E_o	zero point energy
E_v	vibrational energy
∈	molar absorption/extinction coefficient
ε	electric field of electromagnetic radiation
f	restoring force
h	Plank's constant
I	moment of inertia **and** nuclear spin quantum number
I	intensity of light passing through sample
I_o	intensity of light hitting sample
$I(x)$	amplitude of the interferogram
J	coupling constant
k	force constant
l	length of cell
L	distance travelled by fixed mirror
λ	wavelength
m_1 and m_2	mass of atoms 1 and 2 respectively

μ	dipole moment, reduced mass **and** magnetic moment
n	number of scans
N	noise
ν	stretching mode
$\tilde{\nu}$	wavenumber value
q	charge on atoms
r	distance between two atoms
r_e	equilibrium separation
S	signal
$S(\nu)$	smoothing function
T	transmittance
T_1	spin-lattice relaxation time
T_2^*	spin-spin relaxation time
υ	frequency
V	potential energy

STATISTICAL NOTATION

A	number of principal components
a	eigenvalue
ϑ	bias
$\boldsymbol{Cov}$	covariance matrix (unless otherwise indicated)
δ	correction constant
D	training set
$d_{1,2}$	distance between classes 1 and 2
$d_{A,B}$	distance between classes A and B
$d_A(x) = d_B(x)$	quadratic discriminant line
$\boldsymbol{E}(g)$	residual matrix
e_i	residual distance
$f_A(x) = f_B(x)$	linear discriminant line
F_{crit}	critical limit used to determine class boundaries at a given probability level
$f(x)$	sigmoidal function
H_i	Bayesian regularised artificial neural network model
η	learning rate
λ	eigenvalue
M	number of variables (columns)
$\boldsymbol{Out}$	output vector
μ	population mean **and** momentum constant
N	number of samples or objects (rows)
Net	input function of a neurode
$\mathbf{p_a}$	loadings vector

$P(D\|H_i)$	evidence
$P(D\|\mathbf{w},H_i)$	likelihood
$P_{(G(A))}$	unconditional or *a priori* probability of occurrence
$P(\mathbf{w}\|D,H_i)$	posterior probability of the weight parameters
$P(\mathbf{w}\|H_i)$	prior probability distribution
$P_{(x\|G(A))}$	conditional probability of the pattern vector x arising from group A
r_k	radius of circle or sphere containing *K*-nearest neighbours
Σ	sum
σ^2	population variance
s^2	sample variance
s_i	residual standard deviation of sample *i* **and** signals of a neural network
s_{max}	limit of class boundary given a probability level of $p = 0.05$ or $p = 0.01$ to determine F_{crit}
s_t	spread of principal component scores
^{T}	transpose (also denoted by `)
τ	validation error
$\hat{\tau}$	minimum validation error
$\mathbf{t_a}$	scores vector
t_{lower}	defines the lower limit of a SIMCA class envelope
t_{max}	maximum principal component score
t_{min}	minimum principal component score
t_{upper}	defines the upper limit of a SIMCA class envelope
$V_{K,x}$	volume of space containing *K* nearest neighbours
ΔW	increment by which the weights are corrected

W	weight vector
w_j	weights
$W^{(new)}$	corrected weight vector
$W^{(old)}$	uncorrected weight vector
X	multivariate signal vector
$\bar{x}$	sample mean
$x(g)$	vector defining the group centroid
Y	target vector

To my beautiful niece Kimberley
Never lose your inquisitiveness, the whole world awaits
your exploration

CHAPTER 1

AN INTRODUCTION TO CERVICAL CANCER AND DIAGNOSTIC TECHNIQUES

1 AN INTRODUCTION TO CERVICAL CANCER AND DIAGNOSTIC TECHNIQUES

1.1 CARCINOMA OF THE CERVIX

Carcinoma of the cervix is the second most common occurring cancer in women worldwide [1], and was the ninth most frequently occurring cancer in Australian females in 1996. In that year, 923 Australian women were diagnosed with cancer of the cervix and 301 died from cancer of the cervix [2]. There can be no doubt that since the introduction of mass screening programs, the incidence and mortality of cervical cancer has been reduced dramatically. The fact that the prognostic outlook on cervical cancer has improved can be traced back to George N. Papanicolaou's contribution to the field of exfoliative cytology, the Pap smear.

1.1.1 GEORGE PAPANICOLAOU AND THE CERVICAL SMEAR

George N. Papanicolaou has been referred to as the "founding father of modern day exfoliative cytology" [3]. In 1917, whilst working under the direction of Charles Stockard at Cornell University, George Papanicolaou established a correlation of the cytology of vaginal smears with the ovarian and uterine cycles in guinea pigs. In the 1920s he began investigating women with menstrual functional problems where he discovered exfoliated cancer cells. These findings, entitled "New Cancer Diagnosis" were presented to the Third Race Betterment Conference in 1928. The presentation was "weakly received and almost rebuffed by many, especially the pathologists" [4]. Not to be deterred, he concentrated all his efforts on the detection of cancer by vaginal smear and in 1943, published with Herbert Traut "Diagnosis of Uterine Cancer by the Vaginal Smear" [5].

The discovery of exfoliated cancer cells led to the development of a screening procedure for the diagnosis of cancer of the cervix [5]. The technique involved the collection of vaginal fluid by aspiration with a rubber bulb attached to a glass pipette which sampled the posterior fornix of the vagina. The fluid was spread onto the surface of a glass slide and fixed in equal parts 95% alcohol and ether. Slides were stained with haematoxylin to ensure nuclear staining and analysed with light microscopy. The most characteristic features of the abnormal cells were the atypical form and the nuclei, which were often very large, irregularly shaped and *hyperchromatic*[1].

By 1943 the concepts of early cancer and carcinoma *in situ* were widely understood [6], and the potential of the Pap smear for cancer prevention was finally appreciated. Papanicolaou had studied vaginal pool secretions which were easy to obtain but tedious to screen and not applicable in a clinical setting [7]. Ayre [8] used a wooden spatula (now referred to as the Ayre spatula) to directly scrape cells from the cervix [9]. By the late 1940s, cytology laboratories were opening and by the 1950s, Pap smear screening was widespread, even before clinical trials could be performed [10].

It is not well known that the Romanian pathologist, Aureli Babes, had introduced the concept of cytologic sampling of the uterine cervix for the diagnosis of cervical cancer in 1928, two years before the discovery of exfoliated cancer cells by Papanicolaou [11].

1.1.2 THE PAPANICOLAOU (PAP) OR CERVICAL SMEAR

A cervical smear is a test to detect abnormalities of the cells in the cervix. These abnormal cells are the first warning signs of cervical cancer, which if left undiscovered and untreated, may be fatal. If detected early, pre-cancerous changes can be treated.

To obtain a cervical smear a speculum is inserted into the vagina and the handles are squeezed together, widening the jaws to open the vagina and reveal the cervix, Figure 1.1. When the cervix is visible it is examined for any signs of infection or warts. An Ayre spatula is inserted through the speculum and, using the *external os* as a fulcrum, the ectocervix is gently scraped to remove a thin layer of squamous cells [12]. The endocervical or columnar component of the cervical smear is obtained with a Cytobrush™ (MedScand, Inc, Hollywood, Fla., USA), which is inserted, via the speculum, into the cervical canal and rotated to ensure all areas are sampled [12]. The scraped cells are smeared onto a microscope slide, fixed and sent to a cytology laboratory for screening.

If atypical cells are detected, a *colposcopy* is performed and a surgical biopsy specimen is taken from the most abnormal area or areas observed. This specimen is examined histologically to assess the grade of dysplasia or *cervical intraepithelial neoplasia* (*CIN*) and to confirm or exclude the presence of invasive cancer. Colposcopy is the most sensitive method for the detection of abnormalities, however the specificity of the

---

[1] Refer to the glossary for definitions of words appearing in italics (excluding bacterial strains).

technique is quite low. Colposcopy is an expensive, time-consuming technique and is therefore not practical for the routine screening of cervical cancer [13]. If the abnormality is considered histologically to be mild, it will be treated with laser ablation, *cryotherapy* or a *LEEP* procedure. If the abnormality is severe, a cone biopsy or even a hysterectomy may be performed.

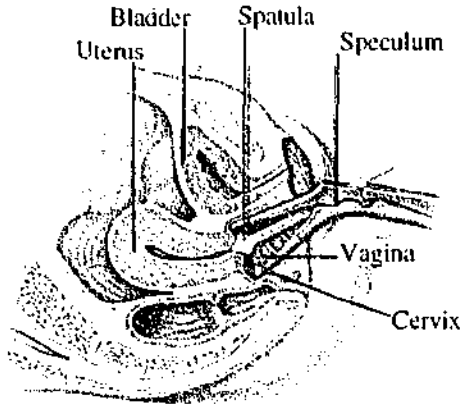

Figure 1.1 A cervical smear is obtained by insertion of a speculum, which widens the vagina to reveal the cervix. A spatula is rotated against the surface of the cervix to scrape off a thin layer of ectocervical squamous cells. A Cytobrush™ (not shown) is then used to collect endocervical squamous cells.

Two types of cancer can develop in the uterine cervix: squamous cell carcinoma, which develops from the squamous epithelium; and adenocarcinoma, which arises from the glandular lining of the endocervical canal. Approximately 85-90% of cervical cancers are squamous cell carcinomas. Squamous cell carcinoma (SCC) is preceded by well recognised epithelial changes and precancerous lesions, which develop through several grades: cervical intraepithelial neoplasia (CIN) I to III; or squamous intraepithelial lesions (SIL), low to high grade.

Cervical intraepithelial neoplasia is graded in terms of the amount of *differentiation* of the neoplasia, with CIN I showing the most differentiation. Progression of precancerous changes and development of invasive carcinoma usually takes about 10-12 years [14]. It has been reported that the regression of CIN I is 60%, persistence 30%, progression to CIN III 10% and to invasion 1% [15]. Similarly, the percentage of CIN II that will regress is 40%, persist 40%, progress to CIN III 20% and to invasion 5%. The likelihood of CIN III regressing is 33% and progressing to invasion greater than 12%. From this data it is clear that the probability of atypical epithelium becoming invasive increases with the severity of the atypia.

As well as identifying cancer and precancerous changes, the cervical smear test is able to detect infections including thrush (*Candida albicans*), *Gardnerella*, *Trichomonas*, *Actinomyces*, wart virus and genital herpes.

1.1.2.1 THE SQUAMOCOLUMNAR JUNCTION AND TRANSFORMATION ZONE

The squamocolumnar junction is the point in the cervix where the squamous epithelium meets the columnar epithelium. This junction undergoes several changes during the life cycle of a female [16]. In young women the majority or the entire transformation zone is ectocervical. As a woman ages, the transformation zone recedes to the endocervical canal [13]. Morphologically there are two types of squamocolumnar junction (SCJ), Figure 1.2. The original SCJ is the border where the original squamous epithelium meets the outermost limit of the developing transformation zone. The present SCJ is the innermost border where the maturing squamous metaplasia meets the mucus secreting columnar epithelium. The transformation zone is the area of actively maturing epithelium between the SCJs and is composed of both squamous and columnar epithelium [17].

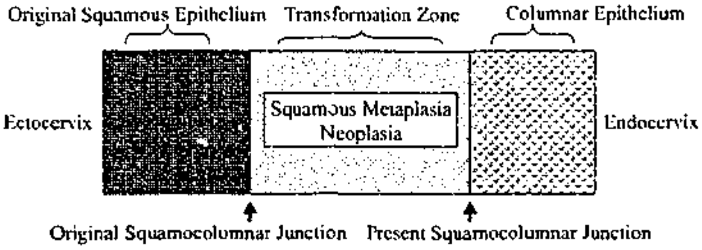
Original Squamous Epithelium
Transformation Zone
Columnar Epithelium
Ectocervix
Squamous Metaplasia
Neoplasia
Endocervix
Original Squamocolumnar Junction
Present Squamocolumnar Junction

Figure 1.2 Schematic representation of the squamocolumnar junction and transformation zone. Redrawn from [17].

For the purposes of cytological analysis a fully satisfactory specimen must contain both squamous (ectocervical) and columnar (endocervical) or squamous metaplastic cells [18, 19]. These cellular elements form the microscopic basis for the assumption that the transformation zone has been sampled. The transformation zone is the site at which cervical neoplasia is likely to arise [9]. Data published in the current literature is inconclusive in relation to the endocervical component as a measure of specimen adequacy. Cross-sectional studies have repeatedly demonstrated that smears with endocervical cells have a significantly higher frequency and higher grade of squamous epithelial abnormalities than do smears lacking such cells [20-22]. However longitudinal studies have demonstrated no increase in the frequency of such lesions on follow up among women whose earlier smear lacked an endocervical component [18, 23]. Since endocervical and ectocervical cells are usually sampled with different instruments the presence of both ectocervical and endocervical cells does not guarantee adequate sampling of the transformation zone [23]. It has been suggested that the presence of endocervical mucus is a more important component for assessing the adequacy of a cervical smear as abnormal cells from the endocervical canal can become trapped in the mucus [24].

1.1.2.2 STAINING

The Papanicolaou stain contains three cytoplasmic dyes: orange G, eosin Y and light green, with haematoxylin for nuclear detail. The Pap stain is able to stain cells blue or blue-green (basophilia, cyanophilia), pink (acidophilia, eosinophilia), orange (orangeophilia) or indeterminate (gray-blue). Cytoplasmic cyanophilia is associated with metabolically active cells, staining RNA, including ribosomes (free ribosomes in the cytoplasm are associated with protein synthesis for internal consumption and are often seen in rapidly proliferating or neoplastic cells). Cytoplasmic acidophilia is associated with an abundance of non-ribosomal organelles (mitochondria, lysosomes, neuroendocrine granules, filaments and smooth endoplasmic reticulum).

1.1.2.3 THE BETHESDA SYSTEM FOR REPORTING CYTOLOGICAL DIAGNOSES OF CERVICAL SMEARS

The Bethesda system was first introduced in the USA in 1988 and was devised from the obvious need to have uniform nomenclature and terminology in the reporting of cytologic findings of Pap smears [25]. The system introduces statements about the adequacy of the specimen ("satisfactory for evaluation", "satisfactory for evaluation but limited by…." and "unsatisfactory for evaluation") as well as detailed criteria for the inclusion of a smear under different criteria based on cellular changes. Specimen adequacy is limited by factors such as blood, inflammation, thick areas of cells, poor fixation and air-drying artifacts [23]. Benign cellular changes include inflammation, bacterial or yeast infection, metaplasia and atrophy. What was previously termed mild dysplasia or CIN I, changes associated with Human Papilloma Virus (HPV), and ASCUS (atypical squamous cells of undetermined significance) are reported as low-grade squamous intraepithelial lesions (LSIL). Higher grades of dysplasia, previously known as (CIN II-III), and CIS are reported as high-grade squamous intraepithelial lesions (HSIL) [23].

In Australia a similar system for reporting the findings of cervical smears is used [19]. The general categories are as follows: unsatisfactory, negative, low-grade epithelial abnormality, inconclusive and high-grade epithelial abnormality. CIN nomenclature is retained in favour of LSIL and HSIL. Benign reactive changes are reported as negative rather than having a separate category as with the Bethesda system.

1.1.3 HISTOLOGY OF THE FEMALE GENITAL TRACT

The uterus is comprised of two main parts: the cervix (neck) and the body (corpus). The cervix is a tubular structure measuring about 4 cm in length and 3 cm in diameter. About half the total length of the cervix lies in the vagina and is called the portico vaginalis (ectocervix), the remainder (endocervix) is continuous with the body of the uterus [26].

The ectocervix lies external to the primary squamocolumnar junction (external os) and is covered by stratified squamous epithelium [27]. Squamous epithelium protects the cervix and vagina from physical, chemical and microbiologic damage [9].

Squamous cells can be divided into four types: superficial[2], intermediate, parabasal and basal cells. As the name implies, superficial cells originate from the superficial layer of squamous epithelium and are the most common epithelial cells at the preovulatory phase of reproductive women [27]. Superficial cells are polyhedral in shape and are reported to be between 35-50 μm in diameter [26, 27] with a nucleus that is small (5-7 μm), shrunken, round or oval in shape. Most superficial cells contain *pyknotic* nuclei, which represent the final stage in the maturation process [27]. Superficial cells are not capable of further growth [26]. Cytoplasm of squamous epithelium stains either eosinophilic (indication of maturation [26]) or cyanophilic [28].

Intermediate cells originate from the middle layer of squamous epithelium and are the most common epithelial cells at the post ovulatory progesterone phase of reproductive women [27]. Intermediate cells (40-50 μm) are polygonal with a round or oval nucleus (9-11 μm) larger than superficial cells [27]. Intermediate cells have cyanophilic cytoplasm and a *nuclear-to-cytoplasmic* (*N/C*) *ratio* less than 1:6 [28]. A frequent variant of intermediate cells are navicular cells which fill with glycogen deposits and are associated with pregnancy and the late menstrual phase [9].

Parabasal cells originate from the deep layer of the squamous epithelium and are smaller (15-30 μm) than superficial or intermediate cells. The nucleus (8-12 μm) is round

---

[2] Refer to Appendix B for light microscopic images of normal epithelial cells, and cells exhibiting various stages of abnormality, be it benign, precancerous or malignant.

or oval [27]. These cells exhibit cyanophilic staining and have a thick, dense cytoplasm and a N/C ratio varying from 1:3 to 1:6 [28]. The occurrence of these cells is associated with childhood, post-partum and menopause, although they may be seen in immature squamous metaplasia [27].

Basal cells are small, round, about 12 μm in diameter, cyanophilic and derived from the basal layer [26]. These cells have an N/C ratio between 1:2 and 1:3 and are not found in cytologic smears under normal conditions [28]. The basal cells are the only cell in the epithelium capable of regeneration, i.e. able to undergo mitosis [26].

Glycogen is present in normal squamous epithelium and absent in atypical squamous epithelium. Basal cells are almost totally devoid of glycogen and the concentration increases towards the surface with superficial cells containing the highest concentration. The quality of glycogen in the epithelium appears to be hormonally controlled and serves to maintain the acid pH of the vagina. Bacterial vaginal flora (*Lactobacillus vaginalis*) metabolise the glycogen released by the *cytolysis* of cells to form lactic acid [16].

The endocervix is 2-3 mm thick and lined with a single layer of tall columnar epithelium. The nuclei are round or oval and generally situated in the lower third of the cell [16]. Columnar cells may be ciliated or mucus secreting. Unlike basal epithelial cells, which are only able to differentiate into squamous cells, endocervical reserve cells have the ability to differentiate in either glandular or squamous cells [9].

1.1.4 CARCINOGENESIS

The development of cancer is a slow process over many years, originating from a mutation of a gene or a series of changes in cellular DNA in one cell and leading to invasive carcinoma, capable of metastasising (spreading) throughout the body [29].

Mutations can occur which result in the formation of an epithelial cell capable of mitosis. These cells may develop the ability to replicate under conditions when the replication of normal cells is inhibited. The uncontrolled *proliferation* associated with cancer results when the enzymatic mechanisms governing cell replication are disturbed and no longer respond to normal regulating agents [26]. The mutated cell and its descendants reproduce at an abnormal rate, whilst still appearing normal, a condition referred to as *hyperplasia*. Over time one in a million of these cells will undergo another genetic mutation, further increasing cell proliferation. In addition to excessive proliferation descendants of this cell appear abnormal in shape and orientation, known as dysplasia. With time another mutation altering cell behaviour occurs. Affected cells become increasingly abnormal in appearance and growth. The tumour may remain contained (*in situ* cancer) indefinitely, not having broken through boundaries between tissues, and some cells may undergo further mutations. A tumour is considered malignant when invasion of underlying tissue and shedding of cells into the blood or lymphatic system occurs (invasive carcinoma). These cells are likely to form new tumours (metastases) in the body [30]. The development of cancer depends on the body's defense mechanisms such as apoptosis and phagocytosis and it is likely that mutated cells will die before cancer can be established [26].

In the course of reshuffling DNA, caused by abnormal mitosis, cell differentiation and function are usually seriously affected [26]. Uncontrolled cell division results in a loss of coordination between the nucleus and the cytoplasm, which leads to the characteristic increase in N/C ratio [9].

1.1.5 CYTOLOGY OF NEOPLASIA AND CERVICAL LESIONS

Morphologic nuclear abnormalities are intimately associated with cancer. The following morphologic changes may be observed in the nuclei of cancer cells [29]:

1. Nuclear enlargement. The nuclei occupy a larger volume than benign cells of similar size.
2. Anisonucleosis or variation in nuclear size and shape. Cancer cells of similar size may contain nuclei of various sizes.
3. Relationship of the cytoplasmic to nuclear volume is often significantly altered in favour of the nucleus, therefore the N/C ratio is increased in cancer cells.
4. Hyperchromasia or the visually perceptible darker staining of the nuclei of cancer cells with nuclear stain (haematoxylin).
5. Abnormal appearance and distribution of *chromatin*. Cancer cells often display large, coarse and irregular, dark staining chromatin granules, significantly larger than normally occurring chromocentres.
6. Nuclear abnormalities. Nuclei are often enlarged, often presenting multiple *nucleoli* or irregular shapes.

7. Mitoses in cancer cells. Numerous mitoses are frequently observed in cancer, many of which are abnormal. Defects arising from abnormal mitosis prevent the proper separation and uneven distribution of *chromosomes* during metaphase and anaphase [26]. The result of abnormal cell division is either cells with abnormal numbers of chromosomes, or gigantic abnormal tumour cells with multiple nuclei.

The nuclei of cancer cells usually contain more DNA compared to normal cells. Nuclear DNA is contained in chromosomes, hence it follows that an increase in nucleic acids (and degree of hyperchromasia) would be proportional to the number and size of chromosomes. It is possible in cancer that not only the absolute concentration of DNA but also the physical state of DNA or DNA associated proteins may be modified in a manner as yet undetermined [29]. The volume, and therefore, to a limited extent the size of the normal nuclei reflects closely the content of DNA and DNA associated protein [26].

The variability in the size of the cytoplasm is a reflection of changes occurring in the RNA and ultimately in the DNA. Variability in shape may result from cell crowding caused by rapid growth [29].

The nucleus generally reflects the health of the cell (normal, inflamed, and degenerative) and nuclear size is an indication of the functional activity of the cell, with larger nuclei associated with more active cells [9].

The following is a summary of the diagnostic features of malignancy used by cytologists. No single feature is diagnostic of malignancy and virtually any feature listed can be found in benign cells [9].

1. Cells – usually numerous, disordered, crowded groups (chaotic architecture), single intact atypical cells, cannibalism, pleomorphism, anisocytosis, abnormal shapes, increased N/C ratio.
2. Nucleus – disorderly, loss of polarity, crowded, enlargement, pleomorphic size and shape, multinucleation, naked, molding, irregular nuclear membrane, thick nuclear membrane, hyperchromatic, irregular abnormal chromatin, prominent multiple irregular or macro nuclei, mitotic figures.
3. Cytoplasm – pleomorphism, loss of cell boundaries, abnormal staining, abnormal cellular products (keratin, mucin).
4. Background – *necrosis*, blood, *tumour diathesis.*

The diagnosis of dysplasia or CIN relies on comparison of nuclear size and shape with surrounding "normal" epithelial cells. Cells arising from mild dysplasia (CIN I) are usually of superficial or intermediate size. The nucleus of CIN I is the largest of the neoplastic and cancerous cells due to the high level of differentiation [9]. CIN I is diagnosed by a comparison of the nucleus with those of surrounding intermediate cells. CIN I is diagnosed if nuclei are larger and exhibit darker staining patterns [28]. Moderate dysplasia (CIN II) show cellular patterns intermediate between mild and severe dysplasia. Most CIN II resembles parabasal size cells with big, dark non-uniform nuclei. Severe dysplasia (CIN III) resembles mainly parabasal or basal cells showing big, dark nuclei with dense uniform chromatin. Cells of carcinoma *in situ* (CIS) are poorly differentiated. The essential difference between dysplasia and CIS is the presence or absence respectively, of any visible signs of squamous differentiation in the abnormal cells [9]. For invasive cancer to be diagnosed, the presence of necrosis or tumour diathesis must be observed. Koilocytes are mature epithelial cells with *cytoplasmic halos* and atypical big, dark nuclei. Differentiating between dysplasia and HPV is often very difficult unless koilocytes are present, indicating HPV infection.

1.1.6 HUMAN PAPILLOMA VIRUS, SMOKING AND CERVICAL CANCER

Epidemiological studies demonstrate that the major risk factor for the development of preinvasive or invasive carcinoma of the cervix is HPV infection [31-36]. Infection with the HPV virus far outweighs other known risk factors such as high parity, promiscuity, young age at first intercourse, low socioeconomic status, and a history of smoking [31, 35, 37]. Studies show that patients with low-risk cytology and high-risk HPV infection with types 16, 18, 31 and 33 are more likely to have cervical intraepithelial neoplasia [11, 38]. Acute infection with HPV types 16 and 18 confer an 11 to 16.9 fold risk of development of high grade CIN [37, 39]. HPV types 6 and 11 have not been associated with the development of neoplasia or cancer and as such are considered low risk types [35].

The oncogenic potential of the virus has been attributed to the *E6* and *E7* genes. The products of these genes stimulate cell proliferation by activating the cell-cycle-specific proteins and interfere with the functions of cellular growth-regulatory tumour suppressor genes *p53* and *pRB* [36]. Whilst it has been established that infection of specific types of HPV is essential for the development of cervical cancer, progression to malignancy requires the involvement of other risk factors and/or cellular events. Important cofactors may be immunity to HPV, age at exposure to HPV, type of exposure, presence of other sexually transmitted diseases and hormonal status [40].

The HPVs are small viruses (about 55 nm in diameter) with a circular double stranded DNA genome. HPV viruses initiate infection and replication in squamous epithelial cells. Viral transcription and replication occurs as cells become more differentiated. The Papillomavirus *E*2 protein affects cellular processes and acts as a transcriptional regulator. Malignant transformation is usually accompanied by disruption of the *E*2 gene, causing deregulation of *E*6 and *E*7 expression. The increased expression of these two genes produces HPV-transformed cells, which are less prone to apoptosis [33].

Conventional diagnostic techniques relying on cytological and histological examination are unable to detect the Papilloma virus. The presence of HPV is associated with the formation of koilocytes in the squamous epithelium. Molecular methods such as Southern blotting, dot/slot blotting, *in situ* hybridisation and polymerase chain reaction (PCR) are being employed for HPV detection. Of these methods, PCR show the most sensitivity and specificity and is able to detect a single molecule of HPV-DNA out of a million cells [36]. Although the literature suggests that HPV is a principal *aetiologic* agent for the development of cervical cancer [11, 38], not all HPV infections lead to cervical cancer and some cancers arise without HPV infection. The incorporation of HPV-DNA testing in conjunction with conventional Pap test screening could identify women likely to have progression to invasive cancer if they harbour infection of HPV, especially high risk types.

Whilst it is difficult to assess the causal effect of smoking on invasive cervical cancer due to the potential confounding factor of sexual behaviour, epidemiological studies have consistently found a relationship between the two [41]. Nicotine and cotinine (a nicotine metabolite) concentrations found in cervical mucus correlate with cigarette consumption [42], and it is thought that smoking may act as a co-carcinogen with HPV being the primary cause. Smoking may depress the immune response allowing HPV infection to persist. The persistence of this infection is believed to allow the development of cervical cancer. Immunosuppression is associated with an increased risk in cervical neoplasia and there is evidence to suggest that smoking may induce immunological changes in the cervix [41, 43].

1.1.7 THE ACCURACY OF THE PAP SMEAR

Since the Pap smear was first introduced in 1943, there have been many reports regarding the accuracy, sensitivity and specificity of this technique. Although the Pap smear has been widely accepted, there continues to be considerable debate as to cost effectiveness and screening intervals required [44].

Given the undesirable effects of misdiagnosis for both patients and laboratories, it is important that the problem of false negative results are addressed in the development of any clinical diagnostic tool.

A false negative result, in the area of cervical cancer, is the failure to detect abnormal cells in a sample from a subject with cervical malignancy or neoplasia. False negative smears give false reassurance and may delay the discovery of a carcinoma until a later stage [45]. Estimates of false negative rates of the Pap smear have been reported to range from 6 to 69% [12, 44, 46-48].

The opposite of a false negative result and just as undesirable, is a false positive result which involves the detection of abnormal cells in a sample from a subject without cervical malignancy or neoplasia. False positive smears lead to unnecessary diagnostic procedures and anxiety in screened women [49]. False positive rates of the Pap smear have been estimated to be as high as 32% [50].

There are many factors which affect the detection rates in cervical smears, including the sampling technique used, preparation of the patient, fixation and staining of the smears, accuracy of screening by the cytotechnologist and interpretation by the cytopathologist [44].

Gay *et al.* [44] reviewed the false negative rate in their laboratory, rescreening the cytologic slides of cases where negative Pap smears were obtained within one year prior to the identification of a malignant tissue diagnosis. The review was conducted without the knowledge of the pathologic diagnosis and false negative cases were placed in the following categories:

1. Sampling error: no malignant or dysplastic cells found on review.
2. Screening error: malignant or dysplastic cells present but not marked by cytotechnologist.

3. Interpretation error: malignant or dysplastic cells present and marked by screener, but their significance misinterpreted by the cytopathologist.

The results obtained showed a 20% false negative rate over the four year period. Of this rate, sampling errors accounted for 62%, screening errors for 16% and interpretation errors for 22%. False negatives may also be caused by vaginal douching, the use of intravaginal drugs and coitus prior to examination or collection of cervical cells [51].

It has been estimated that between 43 and 96% of cervical cells collected by a cervical smear are lost when conventional Pap smears are made [52]. Furthermore the smears may show a different pattern of cellular size and distribution indicating that the cell population in the smear is not representative of the original collection; as a nonrandom portion of cells on the sampling device are transferred to the microscope slide. If these cells are not representative of the original cell population, there may be a significant chance of failing to identify abnormalities in the Pap smear.

1.1.8 GOLD STANDARDS

In order to determine the accuracy of any technique, a second test for verification is generally required, called the gold standard. Histologic diagnosis through biopsy is often taken as the gold standard in cytology [9]. Validity measures (Table 1.1) such as true positive, true negative, false positive and false negative need to be carried out to determine the overall ability of a test result to indicate the disease state of a patient.

Table 1.1 Example of validity measures of a screening test[3], where TP, TN, FP and FP represent true positive, true negative, false positive and false negative respectively.

Test	Patient with disease	Patient without disease	Total
Positive	TP	FP	(TP+FP)
Negativ	FN	TN	(FN+TN
Total	(TP+FN)	(FP+TN)	

A true positive test result is a true indication of the presence of disease, for example diagnosing abnormal when an abnormality is present. Conversely, a true negative test result is an indication of the absence of disease, for example diagnosing normal when no abnormalities are present.

---

[3] Adapted from [9, 53].

The term sensitivity refers to the ability of a technique to detect the presence of disease, i.e. the detection rate. The sensitivity is measured as the number of true positives divided by the total number of positives, i.e. (TP)/(TP+FN) [9, 53] and relates to the ability to detect true positives at the expense of including false positives.

Specificity, on the other hand, refers to the ability of a technique to rule out the presence of disease and can also refer to the ability of a technique to diagnose the particular type of disease present, i.e. CIN II rather than CIN, degree unknown. Specificity is measured as the number of true negatives divided by the total number of negatives, i.e. (TN/TN+FP) [53]. Sensitivity and specificity are tradeoffs, and increasing the sensitivity will probably result in a decrease in specificity.

1.2 ADJUNCTS TO CONVENTIONAL CERVICAL SMEAR SCREENING

The Pap smear, despite wide reports of low sensitivity (high false negative rates), is highly specific with regard to detection of high-grade squamous intraepithelial lesions (HSILS) and cancer [13].

A screening test must be highly sensitive, even if specificity is compromised. The drive to improve sensitivity, thus reducing the false negative rate, whilst maintaining or even improving the specificity is the primary objective behind the development of new technology for the detection of cancer and precancerous lesions [13].

1.2.1 PAPNET®

Manual screening of cervical smears is fatiguing, time-consuming and difficult [54]. Screening involves the microscopic search for the relatively few abnormal cells on a slide. As more than 90% of cytological smears are negative, psychological habituation can easily occur [54]. An automated cytological analysis could eliminate the habituation of manual screening thereby reducing the error rate.

The introduction of the PAPNET® cytological screening system (Neuromedical Systems, Inc., Suffren, NY) is the first step towards the automation of cervical screening. The system serves as a device for locating potentially abnormal cells, which are then interpreted by a cytopathologist [55]. The complete system includes two units: a scanner and a review station. The scanner uses two neural networks[4] to analyse each smear, identifying 128 abnormal appearing cells on the slide [56], and digitally storing colour images of each cell scene. The exact location (*x* and *y* coordinates) of the abnormal cells on the actual slide are also displayed [13]. One neural network is trained to be sensitive to the detection of single abnormal cells, whilst the other is sensitive to the detection of groups of abnormal cells [56]. The images produced are viewed by a cytologist who categorises the slides under 'negative' or 'review,' in which case further microscopic examination is required.

PAPNET® has been shown by several investigators to be a useful tool in the screening and rescreening of cervical smears [54-59]. The system is sensitive to the identification of small numbers of neoplastic epithelial cells and detached malignant cells and is able to detect abnormalities missed during manual screening. PAPNET® utilises the robustness and error tolerance of neural networks to detect abnormal appearing cells in smears that have poor fixing or staining, inflammation, blood and even tumour diathesis.

Despite the high sensitivity of the PAPNET® system to detect abnormalities, the ultimate diagnosis still remains with the cytopathologist and hence human judgement.

1.2.2 THINPREP®

It is reported that to prepare a slide for cytological diagnosis there are four critical parameters that need to be optimised: cellular morphology, clarity, density and uniformity [60]. The conventional methods for the preparation of a slide for diagnosis have little control over these sample components. Often cellular distribution on the slide is uneven, cells can be damaged in the drying process and the presence of blood and mucins can make accurate diagnosis difficult. Furthermore, only a small percentage of cells are transferred onto the microscope slide in the preparation of conventional Pap smears [52, 61, 62]. This subsection of cells is not randomly selected from the sampling device and may not be representative of the cellular composition of the cervix [52].

The ThinPrep® processor (Cytyc Corporation, Marlborough, MA, USA) allows the automated preparation of slides from cells collected in a fluid suspension and is able to control cell density, maintain uniform cellular distribution and enhance the presentation of

---

[4] Refer to Section 2.5.7 for a detailed explanation of the theory of artificial neural networks.

cellular morphology [63]. Rinsing the cells in a fluid suspension ensures that virtually all the cells from the collection device are transferred and preserved [60]. Homogenisation ensures that cells are thoroughly mixed and slides are representative of the population of collected cells.

With the ThinPrep® method, cells to be diagnosed are obtained by gently scraping the transformation zone of the cervix as per conventional smears. The difference between the two methods arises in the preparation of the cells for cytological analysis. Rather than smearing the cellular material directly onto a glass microscope slide, the sampling instruments containing cellular material are rinsed into a vial of preservative (PreservCyt™) solution that preserves the morphology of the cells. PreservCyt™ (Cytyc Corporation, Marlborough, MA, USA) is a buffered alcohol solution able to lyse red blood cells and kill microbiologic elements. The vial is placed inside the ThinPrep® processor where a filter cylinder inserted into the vial is spun at high speed resulting in the breaking up of large clumps of mucus and cellular clusters and homogenisation of the suspension. The cells are collected on a polycarbonate filter membrane that minimises red blood cells, mucins and non-diagnostic debris. A vacuum is applied to the cylinder and the rate at which the pressure difference across the filter membrane changes is monitored by a microprocessor to estimate the percentage coverage of the filter and in turn the number of cells on the filter. An evenly dispersed layer of cells is deposited onto a glass microscope slide in a 20 mm circle and cells are preserved by immersion in a fixative bath containing alcohol. The ThinPrep® process results in a slide containing a thin, uniform layer of cells retaining diagnostic clusters with preserved morphology. These slides are stained for cytological analysis in a routine manner.

Several studies involving the ThinPrep® processor have been undertaken [57, 60, 63-69]. In each trial both conventional and ThinPrep® slides were prepared. This involved the clinician first making a conventional smear on a microscope slide and then rinsing the instruments containing the remaining cellular material into a vial containing PreservCyt™ solution to make a ThinPrep® slide as outlined above. Whilst these studies reported improved specimen quality with the ThinPrep® technique and a high correlation in diagnostic ability compared to conventional screening, no significant difference in the detection of disease with the two techniques was reported. ThinPrep® was found to be superior to the conventional technique in the detection of low-grade epithelial abnormalities [60, 63, 66, 67].

In a study comparing the ThinPrep® technique with histology (gold standard), ThinPrep® was shown to be significantly more sensitive than the conventional method for the detection of low-grade SIL and more severe disease [70]. The ThinPrep® method showed no significant difference in specificity from the conventional method.

Although ThinPrep® offers an improvement in the quality of cervical smears for cancer screening, there are several disadvantages in using monolayer preparations [13]. The interpretation of monolayers is more difficult and requires the retraining of cytopathologists. The cost of ThinPrep® is substantially more than the cost of preparing conventional slides, in terms of the cost of the preservative fluid used and the extra time taken to prepare the slides. This increase in cost may however be offset by the shorter time required to read the slides given the smaller surface area covered by the cells in the monolayer.

1.2.3 CERVICOGRAPHY

Cervicography (National Testing Laboratories, Fenton, Mo, USA) was first proposed by Stafl [71] in 1981 as an adjunct of cervical screening and intended to complement cytologic sampling [72]. Cervicography is the process of interpreting an ectocervical photographic image of the cervix based on colposcopic principles. The cerviscope is a specially designed hand-held 35 mm camera with a telephoto macrolens fitted with an illumination and flash system, which enables a panoramic photograph (cervicogram) to be taken of the cervix [72]. To obtain a cervicogram, the cervix is visualised with a speculum, cleaned with dry gauze and moistened with 4-5% acetic acid [71].

Cervicography, like colposcopy, is most useful in younger women whose entire squamocolumnar junction (transformation zone) can be visualised; and may be of limited value for women whose squamocolumnar junction is located within the endocervical canal (for example in elderly and post-menopausal women) [72].

Ferris *et al.* [72] compared the diagnostic ability of cervicography with histology, when performed alone and in conjunction with conventional screening. Cervicography was found to detect twice as many cases of premalignant cervical disease when compared with the Pap smear alone. When used in conjunction with the Pap smear, cervicography identified nearly two and a half times the number of women with dysplasia as compared to the Pap smear alone. This investigation was potentially limited by the fact that most of the subjects were young and the squamocolumnar junction was easily visualised. Sensitivity and specificity measurements could not be determined, as colposcopy was not performed on women with negative cytologic results.

Earlier studies also reported an increase in the detection of cervical dysplasia compared to the Pap smear [73, 74]. However a significant number of positive cervicograms were found to be false positives. These studies found cervicography to be more sensitive and less specific than the Pap smear. False positive cervicograms typically result from over-interpretation of the significance of acetowhite epithelium [72]. Cervicography false negative failures may represent disease located in the endocervical canal and not visible to the evaluator [72].

Cervicography has been shown to be a very sensitive technique in younger women whose transformation zone is predominantly ectocervical. The sensitivity of cervicography decreases however, when the transformation zone recedes into the endocervical canal [13].

1.3 NON-CONVENTIONAL APPROACHES TO CERVICAL CANCER DIAGNOSIS

1.3.1 POLARPROBE

Polarprobe is an electronic prototype for the detection of cancers and pre-cancers of the cervix [75]. A pen-sized probe tip is placed on the cervix prior to colposcopy and methodically moved across the cervical tissue, stimulating it with electrical and optical pulses for a 2 minute period. The tissue response to the pulses is detected and relayed to an electronics module, which assembles the detector signals and data. Computer software interprets the tissue-response signals and these are compared to a catalogue of 14 tissue types determined by mathematical models.

Coppleson *et al.* [75] collected tissue-responses from 106 volunteers to develop tissue recognition algorithms, which were then applied to an additional 77 volunteers. The algorithm was capable of recognising 14 tissue types, categorised as: Human Papillomavirus (HPV), minor atypia, cervical intraepithelial neoplasia (CIN) grades 1, 2 and 3, micro-invasive/invasive carcinoma, transformation zone edge, columnar cell types 1 and 2, immature metaplasia, mature metaplasia, cervical squamous epithelium, vaginal squamous epithelium and regenerative squamous epithelium. When compared to colposcopy/histology, Polarprobe managed a concordance of 85% on low-grade abnormalities (HPV, minor atypia, CIN 1), 90% for high-grade abnormalities (CIN II or

III) and 99% for invasive cancer. These results were obtained by amalgamating the tissue types listed above into 6 categories: Atypia-CIN I, CIN II-III, invasive cancer, squamous epithelium, physiologic metaplasia and columnar epithelium.

These findings, whilst giving a good indication of the tissue mapping accuracy of the probe, do not directly indicate diagnostic accuracy. This technique is further limited by the size of the probe, which excludes evaluation of tissue within the cervical os [75].

1.3.2 SPECTROSCOPY

The premise that biochemical changes occurring in cells undergoing transformation from normal to cancerous will precede morphological changes [76] is the basis for the investigation of spectroscopy as a possible tool for the diagnosis of cervical cancer. Raman and Infrared (IR) are the two main types of spectroscopic methods based on vibrations of atoms in a molecule. Vibrations that lead to changes in the dipole moment of a molecule can be detected using IR spectroscopy, whereas Raman is sensitive to vibrations that modulate band polarisability. The number of vibrations for a non-linear molecule containing n atoms is $3n$-6 ($3n$-5 for linear molecules). For a biomolecule there are many vibrations resulting in a complex spectrum. Many of the vibrations can be grouped to specific bonds with typical functional groups of interest including C=O, -COOH, O-H and S-H. [77]. Some vibrational modes do not represent a single type of bond oscillation but are instead coupled to neighbouring bonds. A classic example of this are the regions in the IR spectrum termed the amide modes (Section 1.3.3.1) which are characteristic of the IR absorption of proteins.

1.3.2.1 RAMAN SPECTROSCOPY

All molecular species possess polarisability, which is a measure of the ease at which electrons can be induced to move within a molecule under the influence of an applied field [78]. Application of an electric field can induce a dipole resulting from a distortion of the electron cloud. The induced dipole in a molecule vibrating with frequency ν_{vib} and irradiated with frequency ν_0 will vary as a function of ν_0, $(\nu_0+\nu_{vib})$ and $(\nu_0-\nu_{vib})$. The polar state is more energetic than the relaxed state and so polarisation is an endothermic process with spontaneous relaxation accompanied by a release of energy. This process results in the emission of radiation, which is described as scattering, and is the basis for the investigation of molecules via Raman spectroscopy [78].

Polarisation is not an energy state of the molecule and there is no change in the electronic configuration. The polarised state is described as virtual, Figure 1.3, and relaxation from this state can occur in three ways. The most probable is elastic or Rayleigh scattering, where relaxation occurs with no change in vibrational quantum number. Relaxation which occurs with a change in vibrational quantum number produces inelastic or Raman scattering and is termed Stokes if $\Delta v = +1$ and anti-Stokes if $\Delta v = -1$. The intensity of the Raman scatter is related to the population of the initial state of the molecule. For Stokes scatter the intensity is normally related to the population of the ground state, $v = 0$, and for anti-Stokes to that of the first excited vibrational level, $v = 1$.

The Raman scattering efficiency depends on the fourth power of the frequency of the light being scattered. Since the virtual state is not a fixed level, Raman scattered light can be produced from an excitation source with any wavelength, and so Raman spectra are presented as the intensity versus frequency shift from this excitation source [78].

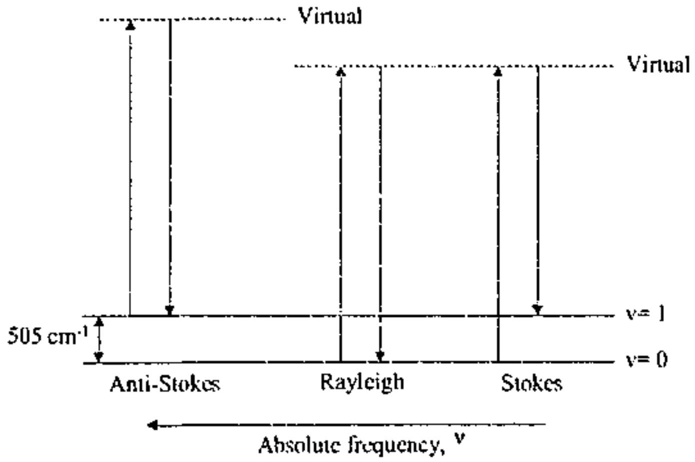

Figure 1.3 Schematic representation of the scattering effects possible with the induction of polarised states. Redrawn from [78].

A conventional Raman system consists of a laser source illuminating the sample and collection optics to gather the scattered light and pass it into the detection system. A detailed description of Raman spectroscopy theory and instrumentation can be found in Hendra *et al.* [78].

Raman spectroscopy has been investigated as a potential technique for the diagnosis of cancers including breast [79-83], skin [81, 83-86] and gynaecological cancer [83, 84].

The potential of near infrared (NIR)-Raman spectroscopy was first investigated by Liu *et al.* [84]. A NIR laser source was used to reduce the fluorescence arising from chromophores inherent in biological specimens. Three main spectral differences were observed between malignant, normal and benign tissue samples. In cancerous tissue, the intensity of the amide I stretching vibration band at 1657 cm^{-1} was less than the intensity of the C-H bending vibrational band at 1445 cm^{-1}. The amide III band at 1262 cm^{-1} was broadened in cancerous lesions; and an additional peak at 934 cm^{-1} was observed only in normal and benign cervical samples.

The reduction in the intensity of the amide I band (1656 cm^{-1}) in precancerous tissue compared to normal, using NIR Raman spectroscopy, was used to form a diagnostic algorithm [87]. The algorithm differentiated precancerous tissues from other tissue categories with a false negative of 9% and a false positive of 12%. The amide I band was also used to discriminate between HSIL and LSIL with a false negative and false positive of 14% and 4% respectively. Cytological changes associated with inflammation and metaplasia were also separated from precancerous lesions.

Ultraviolet resonance Raman (UVRR) spectroscopy at 257 nm excitation was used to study suspensions of normal and malignant cultured cell lines [83]. Cell spectra closely resembled that of DNA, with peaks at 1580, 1480 and1330 cm^{-1}arising due to nucleotide bases. Strong contributions from tryptophan and tyrosine appeared in the 1670-1520 cm^{-1} region. The ratios of the Raman spectral peaks 1480/1614 cm^{-1} and 1480/1540 cm^{-1}, which are sensitive to the concentration of nucleic acids relative to cell proteins, were found to be higher in malignant than normal cells. Despite these promising results, the use of UVRR *in vivo* is precluded by the possible mutagenic effects of UV radiation [88].

1.3.2.2 MAGNETIC RESONANCE SPECTROSCOPY (MRS)

Nuclear magnetic resonance (NMR) spectroscopy gives information about the number of magnetically distinct atoms of the type being studied, with hydrogen and carbon being the most common. When hydrogen nuclei (protons) are studied, for example, the number of distinct types of hydrogen nuclei as well as information regarding the immediate environment of each type can be determined. Many atomic nuclei have a property called spin whereby the nuclei behave as if they were spinning. The most common nuclei that possess spin and which can therefore be studied through this technique include $^{1}_{1}H$, $^{2}_{1}H$, $^{13}_{6}C$, $^{14}_{7}N$, $^{17}_{8}O$, $^{19}_{9}F$ and $^{31}_{15}P$. An atomic nucleus that possesses an odd mass, odd atomic number or both, has a quantised spin angular momentum and a magnetic moment. The number of spin states is determined by the nuclear spin quantum number, I, and there are $2I + 1$ allowed spin states with integral differences ranging from $+I$ to $-I$ [89].

When a magnetic field is applied, the nucleus (which is a charged particle) has a magnetic moment, μ, generated by its charge and spin. The resonance effect occurs when nuclei are induced to absorb energy and change their spin orientation with respect to an applied field. As energy absorption is a quantised process, the energy absorbed equals the energy difference between the spin states. This energy difference is a function of the strength of the applied magnetic field. Nuclear magnetic resonance is so versatile because not all nuclei in a molecule have resonance at the same frequency. This variability arises because the nuclei in a molecule are surrounded by electrons and exist in slightly different electronic environments [89].

The basic requirements for a NMR spectrometer are a radiofrequency (RF) source and a magnetic field. The sample is placed in a probe that is positioned between the poles of the magnet. A coil on the probe transmits the RF radiation and either the magnetic field or the RF is slowly varied. When the resonance condition of the nuclei under investigation is satisfied the sample absorbs energy from the RF radiation and the resulting signal is recorded [90].

Several kinds of "spectra" can be obtained using MRS. A one-dimensional (1D) spectrum can be analysed using five main parameters. The chemical shift (δ) describes the location of each moiety in the frequency scale and is expressed in dimensionless units of ppm (parts per million). The coupling constant J corresponds to the interaction of neighbouring moieties, is expressed in Hertz and the area of each resonance corresponds to the number of resonant nuclei. The spin-spin relaxation time $T2^*$ is inversely proportional to the linewidth at half height of the resonance. Lastly, the spin-lattice relaxation time T1 often acts as a partial saturation factor [91, 92].

2D NMR experiments either observe two different parameters of the same nucleus observation such as the COSY method which provides both δ and J on a 2D map, or two different nuclei, such as ^{13}C and ^{1}H, which are studied simultaneously. 3D methods are also available but these are time consuming, difficult to analyse, and not really appropriate for a clinical setting [91, 92]. For a more detailed explanation of the theory and instrumentation of MR spectroscopy, the reader is referred to texts by Abraham *et al.* [90] and Pavia *et al.* [89].

Magnetic resonance spectroscopy has been investigated as a technique for the diagnosis of various cancers [91-99] including cancer of the cervix [100-103] and ovaries [104, 105].

Proton magnetic resonance (^{1}H MR) spectroscopy allows the detection and identification, by means of the chemical shift (frequency), of molecules in cells and tissues that have sufficient molecular motion to be visible on the MR time scale. MR spectra have been shown to provide information on the chemical and biological characteristics of cells, with the most prominent feature emerging being the appearance of triglyceride in the spectra of malignant cervical cells and tissue [94, 100, 106]. MR spectroscopy is a non-destructive technique allowing histologic classification of specimens after spectra have been recorded [101].

^{1}H MR spectra of malignant cervical tissue are characterised by an intense resonance at 1.3 ppm, which arises primarily from methylene protons of acyl chains in mobile neutral lipids with contributions from methyl protons of lactate and threonine [101-103]. High-grade dysplastic (preinvasive) specimens lack the intense methyl resonance at 1.3 ppm and there is an increase in the broad featureless resonance between 3.4 and 4.2 ppm relative to the spectra of invasive specimens. This resonance arises mainly from protons on carbohydrate, protein and phospholipid metabolites. Plotting the CH_2/CH_3 ratio against the CH/CH_2 ratio resulted in a separation between invasive and preinvasive with a sensitivity and specificity of 94 and 98% respectively [102].

Wallace *et al.* [104] used MRS coupled with linear discriminant analysis (LDA) to distinguish ovarian cancer from normal tissue with a sensitivity of 100%, a specificity of 95% and an accuracy of 98%. 1D ^{1}H spectra from normal, benign, borderline and cancerous biopsy specimens of ovarian tissue were dominated by resonances from lipid methylene (1.3 ppm) and methyl (0.9 ppm) groups, cholines (3.2 ppm), creatines (3.0 ppm) and lysine and polyamines (1.7 ppm). The spectra from the benign neoplasms resembled the spectra of normal samples, both having less intense lipid signals than the borderline and malignant tumours. Smith *et al.* [96] used similar methods and chemometric techniques to classify normal and malignant tissues with a sensitivity of 100% and a specificity of 95%.

^{31}P has been used to study the cellular physiology and metabolism of malignant and normal specimens [92, 97, 98, 107]. *In vivo* ^{31}P spectra typically have three peaks due to nucleotide triphosphates (predominantly ATP), a peak due to phosphocreatine (PCr) and an inorganic phosphate (Pi) peak [98]. Human cancers are characterised by elevated PME (phosphomonoester compounds), elevated PDE (phosphodiester compounds) and an alkaline pH [92].

^{13}C MRS has been applied *in vivo* for studying cancer by monitoring, non-invasively, glycolysis and other metabolic pathways and the flux of metabolites in tumours relative to normal tissue. The presence (or absence) of several important metabolites including glucose, lactate, citrate and alanine, can be detected, along with their concentration, mobility and relaxation characteristics [92].

In vivo studies using MRS are limited by the highest magnetic field strength permitted in clinical practice, 1.5-4 Tesla. *In vivo* pre-clinical studies of tumours by ^{1}H MRS allows the observation of several metabolites that are not detected by MRS of other nuclei. The disadvantage of their use *in vivo* however are the high concentrations of tissue water and lipids which produce intense background signals and must therefore be suppressed in order to observe the metabolites of interest [92]. Metabolites within living tissue or tissue samples are NMR detectable only if they are highly mobile in the cytosol or interstitial spaces. Membrane associated molecules are usually not seen due to line broadening from dipolar coupling. *In vivo* tumour spectra contain only a few resonances and the resolution is much lower giving rise to peaks that overlap and are poorly resolved. In principle the area under each peak is proportional to the concentration of the substance(s) that give rise to it. However peak overlap, distortion and irregular baseline present a problem in analysing and quantifying spectra [97].

The main limitations for the introduction of MRS in a clinical setting is the relatively high cost of equipment, lack of procedural standardisation and low sensitivity of the technique (in terms of amount of sample required for analysis) [91]. Low sensitivity arises because molecules of interest are present in low concentrations and because nuclei other than ^{1}H give rise to weaker signals. For *in vivo* techniques to become clinical practice, it will be necessary to improve detection sensitivity, to automate spectral analysis and to obtain such spectra in a clinically reasonable time at an affordable cost [96].

Whilst the studies mentioned above have indicated the potential of MRS in the diagnosis of invasive cervical carcinoma, MRS appears unable to differentiate between normal and pre-invasive neoplasms [101]. MRS could be a useful adjunct to histology in the clinical management of cervical cancer, where the diagnosis of a neoplasm as invasive depends on the presence of diagnostic debris such as necrotic material, but seems an unlikely technique for routine clinical screening of cervical cancer. ^{1}H MR spectroscopy may have the capacity to allow objective diagnosis of invasive cancer based on cellular chemistry rather than morphologic criteria.

Magnetic Resonance Imaging (MRI) has been used successfully to measure tumour volumes of cervical neoplasia with a high degree of precision [108]. An endovaginal receiver coil placed around the cervix detected tumour volumes as low as 0.2 cm^3 and may be useful for precancer management of patients.

1.3.2.3 FLUORESCENCE SPECTROSCOPY

Laser induced fluorescence spectroscopy is a non-invasive real-time technique that has the potential for quantitatively analysing the biochemical and morphological changes that occur in neoplasia [109, 110].

The technique works by placing a fibre-optic probe on the cervix and illuminating the tissue with low-power, monochromatic light. The system measures tissue fluorescence at excitation wavelengths of 337, 380, and 460 nm. The fluorescent light emitted by the tissue is collected and a fluorescence spectrum recorded, the shape of which is based on the number and type of fluorophores in the tissue. Different levels of fluorescence recorded from the tissue of patients with normal, pre-neoplastic and neoplastic cervices form the basis of discrimination for this technique [110]. Using the excitation wavelengths mentioned above, Turner *et al.* [109] used connectionist methods such as multilayered perceptrons, and radial basis function (RBF) networks to create algorithms more reliable, direct and accurate in precancer detection than those achieved by human experts or multivariate statistical algorithms.

Mitchell *et al.* [110] reported for the diagnosis of SIL, sensitivities of 87% for squamous epithelium, 96% for columnar epithelium, and 78% for the transformation zone. Using the same excitation wavelengths in conjunction with the application of acetic acid to the cervix, Ramanujam *et al.* [111] developed a diagnostic algorithm with 82% sensitivity and 68% specificity in the discrimination of SILs from non-SILs, and 79% sensitivity and 78% specificity in the discrimination of LSIL from HSIL. Acetic acid was found to enhance the discrimination achieved between normal and precancerous cervical tissue [112].

1.3.3 INFRARED SPECTROSCOPY

Infrared (IR) spectroscopy[5] is the study of the interaction of infrared light with molecules and measures the absorption of this light as chemical bonds vibrate. The wavelength of light absorbed depends on the molecules involved in the bond, the type of vibration and the environment [113]. IR spectroscopy is a well-established technique for the study of molecular structure. It is useful in biology and medicine for identification and quantification of compounds with unknown chemical structure and to study the nature of inter- and intra-molecular interactions and conformations of systems. The vibrational spectrum of a compound is analogous to the human fingerprint and is highly characteristic of its physical properties [113].

Biological molecules can be grouped into four major classes: proteins, nucleic acids, lipids and carbohydrates (which are the main constituents of the four elements of interest in diagnostic cytopathology: the nucleus, cytoplasm, membrane and extracellular matrix). Interactions occurring among and between these molecules are usually weak, and comprise of hydrogen bonding and electrostatic and van der Waals interactions [76].

With simple organic molecules it is possible, through theoretical studies such as *ab initio* calculations, to assign absorptions to IR spectra. This is not the case with complex molecules such as biological specimens. Assignment of infrared bands of human tissues, for example, relies on the extrapolation of the infrared spectrum of lipids, carbohydrates, proteins and polynucleotides recorded in isolation and in other systems [76].

1.3.3.1 *CHARACTERISTIC VIBRATIONS OF INTEREST*

Following is a brief description of the main infrared absorptions arising from biological tissues and cells. For a more detailed list of infrared absorptions of biological molecules refer to Table 1.2.

The amide region

Amino acids are the basic structural units of all proteins, peptides and polypeptides. An amino acid consists of an amino group; a carboxyl group and a side chain (R) bound to the carbon atom, Figure 1.4. The spectral region from 1700-1500 cm^{-1} contains the amide I and II vibrations of the amide bonds of the protein components.

---

[5] Refer to Chapter 2 for a detailed explanation on the theory and instrumentation of infrared spectroscopy.

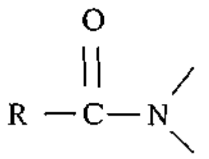

Figure 1.4. Schematic representation of the amide functional group.

The amide I region, near 1650 cm^{-1} arises predominantly from the C=O stretching vibration of the amide group [76, 77] with contributions from peptide N-H in plane bending [114], Figure 1.5.

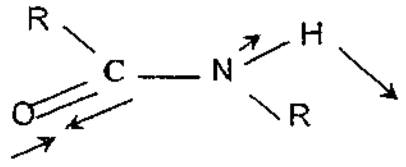

Figure 1.5. Schematic representation of the vibrations contributing to the amide I infrared mode.

The amide II region (1560-1500 cm^{-1}) is primarily an N-H bending vibration coupled to a C-N stretching vibration [76, 77], Figure 1.6. These two vibrational modes of tissue proteins are also able to provide information concerning conformational structure [115].

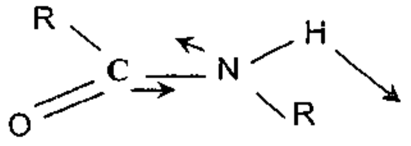

Figure 1.6. Schematic representation of the vibrations contributing to the amide II infrared mode.

The amide III (1350-1250 cm^{-1}) absorption arises primarily from N-H in plane bending and C-N stretching vibrations with significant contributions from CH_2 wagging vibrations [76, 77].

The phosphodiester region

The spectral region between 1000-1250 cm^{-1} contains the vibrational modes of phosphate groups. In nucleic acids, phosphodiester linkages of the polynucleotide chain lead to two strong IR bands: asymmetrical ($\nu_{as}PO_2^-$, 1244 cm^{-1}, Figure 1.7a) and symmetrical ($\nu_sPO_2^-$, 1080 cm^{-1}, Figure 1.7b) phosphate stretching vibrations [76, 116]. A weak band at 1082 cm^{-1} arising from glycogen contributes to the intensity of the symmetric phosphate stretch [116]. PO2$^-$ groups from phospholipids contribute minimally to these bands [116].

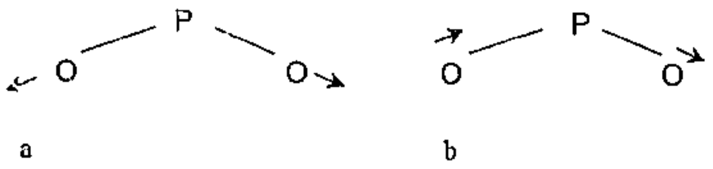

Figure 1.7. Schematic representation of the asymmetric (a) and symmetric (b) stretching vibrations of phosphate.

The carbohydrate region

The major absorption of carbohydrate occurs in the infrared region 1200-1000 cm^{-1} and has been attributed to C-O stretching vibrations. The infrared region between 2800-1800 cm^{-1} is generally devoid of absorptions in biological materials.

Table 1.2. Infrared band assignments of biological molecules [76, 77, 114-123]

Frequency (cm^{-1})	**Functional Group**	**Vibrational Mode**	**Arising from**
3015	=CH	ν	Lipids
~ 2960 2957 2956	CH_3	ν_{as}	Methyl end groups of membrane lipids as well as methyl side chains in cellular proteins
2922	CH_2	ν_{as}	Lipids
2874	CH_3	ν_s	
2852	CH_2	ν_s	Methylene chains in membrane lipids
1741 1730	C=O	ν	Acyl chain lipids
1650	Amide I	νC=O δN-H, in plane	Amide functional groups of amino acids and proteins
1642	OH	δ	Water
1500-1560	Amide II	δN-H with νC-N	Amide functional groups of amino acids and proteins
1485	$(CH_3)_3N$	δ	
1463-1473	CH_2	δ, scissoring	Lipids
1457	CH_3	δ	Methyl groups of proteins, also seen in collagen
1452	CH_3	ν_{as}	Methyl groups of proteins
1405	$(CH_3)_3N^+$	δ	
1404	CH_3	δ	Collagen
1401	CH_3	δ	Methyl groups of proteins
1399 1378	CH_3	δ	Methyl groups of proteins
1318 1339			Sharp bands from connective tissue
1200-1400	CH_2	δ, wagging	
1250-1350	Amide III	δN-H, νC-N and δCH_2, wagging	Amide functional groups of amino acids and proteins

1283 1204			Sharp bands from connective tissue
1244 1240 1228	PO_2^-	ν_{as}	Phosphodiester linkages in nucleic acids
1205			Collagen
1170	CO-O-C	ν_s	Glycogen
1155	C-O	ν	C-OH of proteins and C-O of carbohydrates
1122	C-O	ν	Mucins
~1080	PO_2^-	ν_s	Phosphodiester linkages in nucleic acids
1070	CO-O-C	ν_s	
1047	C-O-P	ν	
1047	C-OH C-O	ν δ	Carbohydrates, in particular glycogen
1043	CH_2OH	ν	Mucins
1031			Collagen
1025	$-CH_2OH$	ν	Glycogen
1023	C-O	ν	Glycogen
971	PO_4^-	ν_s	Dianionic phosphate monoesters of phosphorylated proteins and cellular nucleic acids

1.3.3.2 INFRARED SPECTROSCOPY IN THE DIAGNOSIS OF CANCER

Infrared spectroscopy has been extensively applied to study changes at the molecular level of various human cancers. Several groups have investigated the use of IR spectroscopy in the diagnosis of colon [117, 120, 124], cervical [116, 118, 121, 123, 125-131], lung [132, 133], and liver [119] cancers, as well as leukemia [134-136].

1.3.3.3 CHARACTERISTIC INFRARED SPECTRAL CHANGES BETWEEN NORMAL AND MALIGNANT CERVICAL CELLS

The initial work in the application of infrared spectroscopy in the detection of cervical cancer was undertaken by Wong *et al.* [118]. Several changes in the infrared spectra were found to be common to cancers including colon, stomach, skin, esophagus, liver, cervix and vagina.

After collecting infrared spectra of exfoliated cervical cells from women with normal or dysplastic cytology, Wong *et al.* [116] found several observable spectral differences, and noted the following differences in malignant from normal cells:

1. Significant changes in intensity of bands at 1303, 1244, 1155, 1082, 1047 and 1025 cm^{-1}.
2. Significant shifts of peaks normally appearing at 1244, 1155 and 1082 cm^{-1}.
3. Additional peak at 970 cm^{-1}.

They also noted four prominent features characteristic of a dysplastic spectrum:

1. The intensity of the glycogen band is intermediate between those of normal and malignant samples.
2. The $\nu_sPO_2^-$ peak of 1082 cm^{-1} is not shifted.
3. The centre of gravity of the band at 1155 cm^{-1} is not shifted to the same extent as in cervical cancer.
4. The additional band at 970 cm^{-1} is less intense than in cervical cancer.

The ratio of the peak intensities of bands at 1025 cm^{-1} (glycogen) and 1082 cm^{-1} (phosphodiester groups of nucleic acids) were found to differ greatly between normal and malignant cells [116]. The decrease in intensity of the νC-OH band at 1155 cm^{-1} and the νC-O band at 1023 cm^{-1}, which disappears with malignancy, occurs due to a reduction in the glycogen level in abnormal cells [116]. It is well known that cells undergoing neoplastic and malignant transformation exhibit a reduction in glycogen [6, 133]. The relative intensity of the phosphodiester stretching ($\nu_{as}PO_2^-$) band at 1240 cm^{-1}, with respect to infrared bands originating from the vibrations of proteins and lipids, became stronger in abnormal cells due to an increase in the N/C ratio [116]. This correlates with the increase in N/C ratio that occurs with the uncontrolled proliferation associated with cancer. Growth control mechanisms have been inhibited and cells devote all their energy reserves to division, which leads to a decrease in the ability of these cells to differentiate and form a normal cytoplasm [9].

Pressure dependence studies and deconvolution techniques were used in an attempt to correlate the spectral findings with possible structural changes occurring in malignant cervical cells [115, 116]. It was revealed that in malignant tissue there were extensive changes in the degree of hydrogen bonding of phosphodiester groups of nucleic acids and C-OH groups of proteins, as well as changes in the degree of disorder of methylene chains of lipids.

These findings were further substantiated and found to be applicable to malignant cervical tissue [115]. In 1996 Yazdi *et al.* [123] identified the following features common to all types of cancer:

1. Increase in hydrogen bonding of the phosphodiester groups of nucleic acids.
2. Decrease in the hydrogen bonding of the C-OH groups of proteins.
3. Enhanced molecular packing of nucleic acids.
4. Increase in hypomethylation.
5. Reduced glycogen levels in glycogen-rich tissues.
6. Increase in the disorder of the methylene chains of membrane lipids.

Fung *et al.* [53] compared FTIR spectroscopy in the screening of cervical cells with conventional Pap smears using colposcopy directed biopsy as the gold standard. Specificity and sensitivity were reported for FTIR (98.8% and 98.6%) and for the Pap smear (90.5% and 86.6%). Infrared spectra were classified as abnormal if they contained any of the spectral features based on the earlier work of Wong *et al.* [116, 123].

Although these findings seem to indicate that IR spectroscopy is a powerful tool in the (visual) discrimination of normal and malignant cervical epithelial cells and tissue, it is becoming increasingly apparent that there may be other factors contributing to the spectral changes assumed to be arising from neoplastic processes and malignancy. Section 1.4 provides a more detailed discussion of some of these factors.

Yazdi *et al.* [123] demonstrated that changes in the intensity, frequency and band shape of many bands in the IR spectra of abnormal cervical specimens were altered from those of normal cervical specimens [123]. The majority (95%) of HSIL cases investigated showed dramatic changes in IR spectra. However, only 54% of LSIL cases and 33% of ASCUS cases exhibited significant spectral changes. The remaining cases exhibited only slight spectral changes, which were considered to be a reflection of the heterogeneity of lesions classified as LSIL and ASCUS.

In 1998 Diem's group [128] conducted a series of experiments in an attempt to demonstrate that IR spectroscopy could be used as a marker of maturation and differentiation in cervical squamous epithelium. The spectral differences between basal, parabasal, intermediate and superficial layers arose mainly in the 1200-900 cm^{-1} region. The differences were seen as an increase in glycogen concentration towards the surface, i.e. as cells matured from the basal layer they accumulated more glycogen. Cervical squamous epithelial cells accumulate glycogen as a process of maturation, the concentration being hormone dependent and peaking around ovulation [137]. Differences were also noted in the amide I/amide II ratio, believed to be a result of nucleic acid contributions.

It was hypothesised [128] that the nucleus in a superficial cell, which has reached maturity and is therefore tightly compacted, is infrared "opaque" and that any spectral features of nucleic acids observed from these cells are a result of RNA rather than DNA contributions. A 5 μm infrared beam focused on a single cell nucleus produced no observable light at the detector, whereas a similar size beam focused on the cytoplasm produced a weak IR spectrum. This experiment remained inconclusive because diffraction effects for an infrared microscope fitted with a mercury-cadmium-telluride (MCT) normally prevents the use of an aperture less than 30 μm [138]. A recent study in the same laboratory has concluded that DNA in the nucleus has an optical density too high to allow transmission of IR radiation [136].

Despite the differences that are observed in normal squamous epithelial cells as a product of maturation, Cohenford and Rigas [127] found that the spectra of cytologically normal intermediate and superficial cervical squamous cells from women with dysplasia or cancer differed from those of the cells of normal women[6].

Exfoliated cervical cell samples are dominated by the presence of superficial and intermediate cells, with parabasal cells occurring less frequently and basal cells rarely observed [28]. Superficial and intermediate cells were found to give rise to two spectral patterns [127]. The first pattern (A) was characterised by bands at approximately 1652, 1544, 1242, 1153, 1105, and 1080 cm^{-1} and an intense glycogen band at 1027 cm^{-1}. The second spectral pattern (B) was characterised by a significant reduction in the intensity of the 1027 cm^{-1} band. Other bands were observed at approximately 1653, 1544, 1171, 1114, and 1079 cm^{-1}. The band seen at 1241 cm^{-1} exhibited an increase in intensity compared to pattern A. The position of the peaks between the two spectral patterns did not differ significantly except for the peaks at 1153 and 1105 cm^{-1} which were shifted to 1171 and 1114 cm^{-1} in the pattern B spectra. A spectral pattern intermediate between patterns A and B were observed in about 5% of the cells. Glycogen was believed to be the primary cause

---

[6] Refer to Section 1.3.3.4 for a further discussion of these results.

in the difference between the two spectral patterns, with peaks associated with glycogen (1155, 1080 and 1026 cm^{-1}) all observed in the infrared spectra exhibiting spectral pattern A. This was further proven when cells exhibiting spectral pattern A stained positive for glycogen with *Lugol's reagent*. No staining was observed in any of the cells yielding spectral pattern B [127]. The spectral patterns displayed by parabasal, endocervical, koilocytic, dysplastic and malignant cells were all similar to pattern B spectra.

The infrared spectra of dysplastic and malignant cells were found to have spectral features similar to those observed in basal cells [131]. A loss in spectral detail in the low frequency peaks associated with DNA was found as the cells progressed towards cancer. In addition to the loss of spectral detail in the 1200-1000 cm^{-1} region, dysplastic tissue spectra exhibited a small increase in the intensity between 1500 and 1150 cm^{-1} and an increase in the amide II peak [131]. Given that basal cells are very rarely observed in cervical smears, a spectral pattern representing these characteristics could be used as an indication of the presence of dysplasia of malignancy.

The majority of studies involving the application of infrared spectroscopy to the investigation of cervical cancer have been based on results of samples containing a large amount of cells. The spectral differences observed with infrared spectroscopy can only be attributed to large changes in the total sample, rather than changes occurring as a result of the few abnormal cells generally present in cervical smears [139]. Lowry [139] employed mapping techniques in an attempt to understand the nature of these effects. Results show that abnormal changes occur across the whole sample, rather than from just the presence of a few abnormal cells. This observation provides further evidence in support of the argument that significant biochemical changes are occurring in the cells before morphological changes of the disease state can be visually detected.

1.3.3.4 THE ROLE OF MULTIVARIATE STATISTICS

The discrimination between the infrared spectra of normal and malignant cervical cells reported by Wong *et al.* [116, 118] is not ideal because results were based on visual inspection of the spectra and the use of peak ratio comparison. Visual inspection of infrared spectra introduces subjective bias, and the technique of peak ratio measurements is insensitive to interference from extraneous factors and subtle differences between spectra [126]. Differences in the thickness of the sample can also contribute to peak ratio bias and if these are to be used as criteria for discrimination, sample thickness must be compensated for [139].

Biological specimens are inherently variable in nature and as such discrimination between IR spectra of cervical specimens requires the use of robust and sensitive methods. These methods must be able to model for nonlinearities arising from various sources including sample processing errors, baseline shifts, batch-to-batch variations, and the presence of non-diagnostic debris [126]. Methods also need to be sensitive to the presence of 'outlier' spectra which may result from samples with less than optimal numbers of cells, or specimens containing blood, mucus or other non diagnostic debris.

Infrared spectroscopy has been coupled to various multivariate statistical techniques to create effective models and classification tools in the investigation of cervical cancer [125-127, 140].

Wood *et al.* [125] used principal component analysis (PCA) to identify 7 key wavenumber values contributing to the majority of the variance between the infrared spectra of normal and abnormal cervical cells. Infrared spectra were assigned into two groups with type 1 spectra exhibiting a spectral profile characteristic of normal epithelial cells and type 2 spectra exhibiting features of dysplasia. Type 1 spectra were characterised by an intense glycogen band at 1022 cm^{-1} and a pronounced $\nu_sPO_2^-$ band. Type 2 spectra showed pronounced $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands and a reduction in the band arising from glycogen. These spectral types are in agreement with those described by Wong *et al.* [115] as normal (type 1) and dysplastic or malignant (type 2). Further evidence to support type 2 spectra as displaying spectral profiles consistent with dysplasia or malignancy was seen when IR spectra obtained from HeLa cells (a malignant cell line) were shown to cluster with type 2 spectra in a 2D principal component scores plot. Comparison of the two spectral profiles with cytology (Pap smear) and histology (biopsy) revealed that 86% of the spectra exhibiting type 1 spectral profiles were diagnosed normal by Pap smear and 87% exhibiting type 2 spectral profiles were diagnosed histologically as showing dysplasia or HPV effects [125].

Cohenford *et al.* [126] utilised principal component regression (PCR) to achieve a separation between normal and malignant cervical cells based on the presence (normal) or absence (malignant) of a peak attributed to glycogen at ~1025 cm^{-1}. They also reported findings suggesting that cells of atrophic cervical samples shared important structural features with neoplastic cells. It has been well documented that cells in cervical cancer and cervical atrophy have a reduction in glycogen content [6]. The discrimination of normal and malignant cells based solely on the glycogen region is of questionable value given the inherent amount of variation seen in the glycogen levels of normal squamous epithelium.

Cohenford and Rigas [127] used partial least squares (PLS) as a means of achieving a separation between malignant and cervical squamous epithelial cells. Chemometric analysis was concentrated in the spectral regions 1200-1000 cm^{-1} and 3000-2800 cm^{-1}. Calibration curves were fitted using the spectral data of normal, dysplastic and malignant samples exhibiting pattern A spectra. Calculated average predicted values were found to differ significantly between normal and dysplastic and also between normal and malignant. Similar calibration curves were fitted for pattern B spectra. No statistically significant difference was noted between the normal and dysplastic groups, although differences between normal and malignant and dysplastic and malignant groups were statistically significant. These findings suggest that when a morphologically defined neoplasia develops in the cervix, normal-appearing cells surrounding the abnormal cells have extensive structural, chemical or metabolic changes, which become apparent using IR spectroscopy. This indicates that changes associated with the neoplastic process may occur earlier than presently recognised morphologically.

1.4 THE IMPORTANCE OF IDENTIFYING POTENTIAL CONFOUNDING VARIABLES

The cytology of a cervical smear is very complex and comprises a variety of cell types including endocervical and ectocervical epithelial cells, erythrocytes, leukocytes, and platelets (thrombocytes). Smears may also contain bacteria, yeast, mucins, semen and other contaminants [140].

It is becoming increasingly apparent that changes seen in IR spectroscopy are not always indicative of a disease process. There are many variables that could potentially contribute to spectral changes observed between healthy and abnormal samples, and changes seen are not always indicative of the progress of disease [130]. In order for the application of IR spectroscopy in the diagnosis of cervical cancer to be highly sensitive and specific, it is of vital importance to identify possible confounding variables in cervical smears and to assess their influence on the IR spectra of exfoliated cervical cells.

1.4.1 THE MENSTRUAL CYCLE

The menstrual cycle consists of three phases (flow, proliferative and secretory), paralleled to the follicular, ovulatory and luteal phase of the ovarian cycle, Figure 1.8, which provides hormonal stimulation (in the form of estrogen and progesterone) to the endometrium and epithelium of the vagina and cervix [141]. Menstrual bleeding occurs in the flow phase, with the first day of bleeding used as a reference point for the cycle. Follicles in the ovary begin to grow during the follicular phase, culminating in the maturation of one follicle, leading to ovulation. In the ovary, the follicular tissue that remains after ovulation is transformed into the corpus luteum (luteal phase). The corpus luteum is an endocrine tissue that secretes estrogen and progesterone [141].

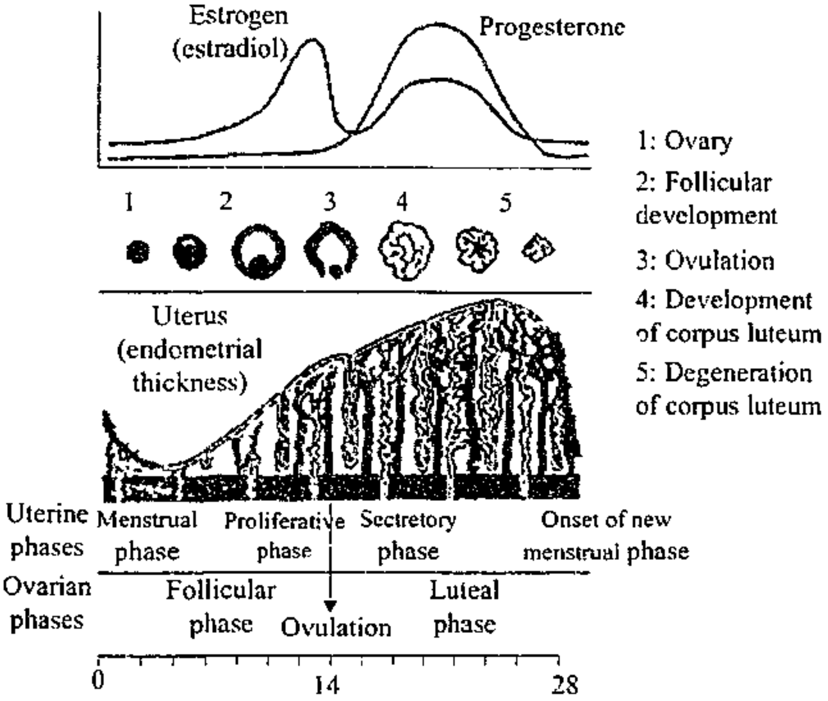

Figure 1.8 The reproductive cycle of the human female [141].

During the follicular phase the ripening follicle produces only estrogen, causing development of the epithelium. With continuing estrogen stimulation, a separate layer of intermediate cells differentiates from the parabasal layer. These cells contain vacuoles rich in glycogen [137]. Towards the end of the follicular stage, full estrogenisation is established and the epithelium has developed into a thick structure, with an outer covering of superficial cells. A characteristic feature of the completion of this maturation process is the appearance of pyknotic nuclei in the superficial cells. Only estrogen is capable of producing this degree of proliferation. Following ovulation, the corpus luteum secretes both progesterone and estrogen. Progesterone causes the highly proliferated epithelium to regress back to intermediate proliferation [142]. The relative numbers of superficial and intermediate squamous cells vary, depending on the phase of the menstrual cycle [143].

Endocervical cells also participate in cyclic hormonal changes, although no morphological changes are observed by light microscopy [144]. During estrogenic activity the endocervical epithelium proliferates and an increase in secretory activity is observed [16]. Cervical mucus, which is thick for the majority of the menstrual cycle becomes liquid for 3 or 4 days prior to, during and after ovulation [26]. The number of secretory and ciliated cells in the endocervix is variable and the presence or absence of either may be related to the phase of the menstrual cycle [16]. Doornewaard *et al.* [22] reported that the day of the menstrual cycle on which cervical smears were obtained was influential on the presence or absence of endocervical cells, with the absence of endocervical cells more likely in the second half of the menstrual cycle.

1.4.1.1 CYTOLOGY OF THE OVULATORY CYCLE

1. Follicular phase: occurs from menstruation to ovulation. Cervical smears taken during this phase consist of large, flat superficial cells and intermediate squamous cells. Under normal cyclic conditions the number of polymorphonuclear leukocytes (PMNs) decreases and the percentage of superficial cells increases throughout the phase [28]. The increase in the number of superficial cells is directly related to the maturation effect of estrogen on squamous epithelium.

2. Ovulatory phase: refers to the midcycle days when ovulation occurs (theoretically day 14 in an ideal 28 day cycle). Cervical smears are "clean" with little, if any, mucus or PMNs. The percentage of superficial cells is at a peak due to estrogenic stimulation. A postovulatory reaction is expressed by folding of cells and appearance of small clusters, sometimes accompanied by mucus and PMNs [28].

3. Luteal phase: occurs post ovulation and continues to the onset of menstruation. Progesterone activity causes abundant exfoliation and squamous maturation reaches the intermediate cell level seen through an increase in intermediate cells. This pattern continues until 2-4 days before menstruation when an increase in superficial cells may occur. During the premenstrual period abundant PMN and granular mucus are present [28].

4. Menstrual phase: may show cellular smear pattern similar to late luteal or early follicular. Numerous blood cells and groups of endometrial cells are present [28].

Oral contraceptives act primarily through pituitary inhibition to prevent ovulation and suppress endogenous estrogen and progesterone [142]. The two most prescribed oral contraceptives are monophasic and triphasic. The monophasic pill has a fixed combination of estrogen and progesterone throughout the cycle. Cellular patterns are characterised by the absence of cyclical changes and presence of the features usually associated with the second half of the cycle [142]. The triphasic pill varies the concentrations of estrogen and progesterone in an effort to mimic the cyclical pattern of endogenous hormone secretion [145].

1.4.2 ENDOCERVICAL CELLS

The presence of endocervical columnar cells in cervical smears is usually indicative of adequate sampling of the transformation zone however endocervical cells have been found by several groups to exhibit IR spectral patterns similar to those seen in dysplasia and malignancy [121, 127, 129, 140]. Columnar cells differ markedly from squamous cells in structure, molecular composition of the cytoplasm and physiological functions.

Endocervical cell spectra are characterised by an increase in the relative intensity ratios of $\nu_{as}PO_2^-$ (1238 cm^{-1}) and $\nu_sPO_2^-$ (1082 cm^{-1}) with respect to δCH_3 (1401 cm^{-1}), compared to ectocervical cells. Significant changes in the band shape and relative intensity of the C-O stretching vibration at 1155 cm^{-1} have also been noted as well as the appearance of a band at 971 cm^{-1} which is absent in ectocervical epithelial cell spectra. The intensity of the methylene band increases whilst the methyl band decreases in endocervical compared to ectocervical, suggesting that the number of methyl groups with respect to ethylene groups is lower in columnar cells than in squamous cells [121]. Differences between columnar and squamous epithelium are most pronounced in the 1100-1000 cm^{-1} region, with columnar cells showing a characteristic broad peak with maxima at 1076 and 1040 cm^{-1} and a weak shoulder at approximately 1120 cm^{-1}. These peaks have been attributed to the carbohydrate moiety of glycoproteins in cervical mucus, which accumulates in columnar epithelium [129]. Glycogen is not accumulated in columnar cells [16].

Columnar cells can be differentiated from malignant squamous cells by shoulder bands occurring at 1122 and 1052-1043 cm^{-1} which are stronger in normal endocervical than malignant squamous cells. The presence of mucins in endocervical cells contribute to the intensities of these shoulder bands

1.4.3 BENIGN CELLULAR CHANGES (BCCS)

There are many processes occurring in the epithelium of the cervix that are considered to result in benign changes. Two examples of benign cellular changes, which could potentially confound the IR detection of cancer and precancer, are inflammation and metaplasia.

Yazdi *et al.* [123] demonstrated that changes in IR spectra arose in cervical cells from preinvasive and other conditions such as BCC and ASCUS. All samples with BCCs exhibited abnormal IR spectra. Fifty-nine percent of these samples, which had cytologic changes associated with inflammation, exhibited IR spectral features similar to those seen in malignancy.

1.4.3.1 INFLAMMATION, LYMPHOCYTES AND OTHER BLOOD COMPONENTS

Inflammation occurs as a tissue reaction to injury and the injured tissue is associated with the presence of leukocytes. There are three main types of inflammation [29]:

1. Acute: characterised by necrosis and breakdown of tissues with the predominance of polymorphonuclear leukocytes [26]. Liquification of necrotic tissue and dead leukocytes results in exudate (pus).
2. Subacute: less tissue breakdown with a predominance of leukocytes and lymphocytes.
3. Chronic: slight tissue breakdown with a predominance of leukocytes and lymphocytes with occasional plasma cells.

Macrophages (cell capable of engulfing large particles) or histiocytes participate in the bodies defenses by phagocytosis. They may be mono- or multi-nucleated with round or kidney shaped nuclei. The cytoplasm is usually filled with small vacuoles [26].

Wood *et al.* [140] found that in general, the IR spectra of leukocytes (a type of lymphocyte or white blood cell (WBC)) exhibit features suggestive of malignant transformation in the phosphodiester region. Spectra are characterised by a reduction in glycogen at 1050 cm^{-1} and 1024 cm^{-1} and pronounced $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands at 1240 and 1080 cm^{-1} respectively. These spectral differences were observed in the IR spectra of cytologically inflamed samples, although several features in the spectra enabled their differentiation from malignancy [123].

Thrombocyte (platelet) numbers become significant as a consequence of a cervical lesion or tissue damage. Thrombocyte aggregation arises as part of the initial clotting mechanism and IR spectra are characterised by intense $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ peaks at 1240 and 1080 cm^{-1} respectively, possibly due to a combination of phosphorylated proteins (ADP) found in thrombocyte granules. Thrombocytes exhibit characteristic bands at 980 cm^{-1} and 935 cm^{-1} that can be used as a marker of contamination [140].

Chiriboga *et al.* [130] found the most common contaminant in the IR spectra of cervical tissue and cells arose from the presence of PMNs. PMN contamination lead to spectra that were different from those of pure epithelial cells. The spectral features of PMNs could possibly mask those of diseases such as cervical dysplasia.

Erythrocytes (red blood cells (RBCs)) are frequently observed in smears and their presence may result from cervical sampling during menstruation (flow phase), tissue damage caused from sampling or symptomatic bleeding of a cervical lesion [140]. The IR spectrum indicates a lack of glycogen, but erythrocyte spectra are readily discerned from abnormal spectra by examination of distinctive phosphate peaks. Erythrocytes exhibit only diminutive phosphate peaks due to a deficiency in nucleic acids.

1.4.3.2 METAPLASIA

Metaplasia occurs as a response to stimuli such as pH or endocrine changes, trauma or inflammation. Metaplastic cells are derived from columnar basal cells, which differentiate into squamous cells in order to protect the delicate glandular epithelium. Metaplasia frequently occurs at puberty and at pregnancy when the size and shape of the cervix increases [9]. Immature metaplastic cells resemble parabasal cells.

The IR spectra of 19 out of 36 samples, diagnosed by cytology as 'within normal limits', demonstrated abnormal spectra [123]. All of these patients had previous histories of abnormalities of the cervix and the authors believe that these spectra could be indicative of the occurrence of molecular changes before morphological abnormalities are observable in cytology. However careful review of these smears showed the presence of squamous metaplastic cells, parabasal cells or mild reactive cell changes in some of the smears that might have caused changes in the IR spectra.

1.4.4 MUCINS

The consistency of cervical mucus can vary considerably depending on hormonal influences such as ovulation, at which time the mucus becomes thinner enabling the reception, transport and nutrition of spermatozoa. The major component of mucus is a highly characteristic epithelial glycoprotein with smaller amounts of protein [146]. Cervical mucus contains a high proportion of sugars (70-90%) and low nitrogen content. Large amounts of the amino acids proline, serine and threonine are present whereas aromatic and sulfur-containing amino acids are present in low concentration or absent altogether. The constituent sugars are N-acetylglucosamine, N-acetlygalactosamine, N-acetyl neuraminic acid, fucose, galactose and ester sulfate. The most probable structure of mucus is a polypeptide backbone to which a large number of branched sugar chains are attached via O-seryl and O-threonyl glycosidic linkages. Neuraminic acid and fucose are present at either the ends of the sugar chains or as non-reducing terminal side chains. The protein component of mucus is believed to act as a cross-linking reagent between the long threads of glycoprotein. It has been suggested that the cross-linkages occur via the neuraminic acid residues through a combination of ionic and hydrogen bonds. Low molecular weight components include sodium and chloride ions and organic constituents include glucose [146].

The presence of mucus represents a problem in IR spectroscopy since its high viscosity prevents the easy separation from epithelial cells, and both the viscosity and the spectral pattern of mucus depends markedly on the menstrual cycle [130]. IR spectra arising from mucins are characterised by a broad peak with maxima at 1076 and 1040 cm^{-1} and a weak broad shoulder at approximately 1120 cm^{-1}. These peaks have been attributed to the carbohydrate moiety of glycoproteins, the major constituent of mucus [130].

1.4.5 CONNECTIVE TISSUE

The problem with sampling cervical tissue is the presence of connective tissue below the epithelium, which absorbs in the IR region 1100-950 cm^{-1}, similar to the region where changes associated with malignancy have been noted [115]. The purpose of connective tissue is to provide a matrix, which connects and binds cells and organs. The major constituent of connective tissue is the extra cellular matrix composed of collagen, protein fibres, an amorphous ground substance and tissue fluid [115, 122]. Connective tissue becomes less of a problem when studying exfoliated cervical cells, as only cells are scraped in the collection process. Connective tissue may become problematic if endocervical cells are sampled [121], as the epithelium of the endocervix is only a single layer thick. If a smear is taken with force, it is foreseeable that connective tissue may be present. Connective tissue can be differentiated from normal and malignant tissue and cells in the region 1500-1200 cm^{-1} (Table 1.2).

1.4.6 SEMEN

Infrared spectra of semen are characterised by an intense amide II band (1550 cm^{-1}); a prominent band at 1400 cm^{-1} associated with the COO^- groups of fatty acids and amino acids; an intense broad band at 1084 cm^{-1}; and a distinctive doublet appearing at 981 and 968 cm^{-1}. Semen does not seem to present a problem as a confounding variable because the distinctive spectral doublet could be used as a marker for its presence [140].

1.4.7 FIXATIVES

For analysis of cells it is desirable to preserve structure and cellular constituents with the least possible distortion. The fixative used must be able to quickly penetrate the cell membrane and stop all biochemical and mechanical activity. The direct effect of most fixatives is on cell proteins and protein-lipid compounds, which become denatured and coagulated [26]. Alcohol is a coagulative fixative and may cause up to 70% shrinkage in cells [9]. Isotonic saline can cause the precipitation of glycogen from the cytoplasm [140].

1.5 A NEW DIAGNOSTIC TECHNIQUE FOR CERVICAL CANCER

The investigations of objective spectroscopic techniques for diagnosing cervical cancer performed in this dissertation are a continuation of the Honours project undertaken by the author. The main outcome of that project was a trained artificial neural network that predicted the IR spectra of 20 unknown normal and abnormal samples with 100% sensitivity and 90% specificity [147]. The project also presented preliminary results of a study of the IR spectra of cervical smears obtained during the menstrual cycle.

Matlab was utilised to write routines that would enable the objective pre-processing of spectral data. Chapter Three gives a detailed account of the Matlab routines written for pre-processing and analysis. Unless otherwise stated, all routines presented in this chapter were written by the author of this dissertation. Having had no real experience in computer programming, Matlab code was self-taught under the guidance of Prof. Frank Burden, a chemometrician in the School of Chemistry at Monash University (Clayton, Australia).

Chapter Four of this dissertation presents an investigation into the spectroscopic effects of hormonal stimulation and nucleic acids in the diagnosis of cervical cancer. IR spectra of cervical smears comprise molecular contributions of the major components of cells. The intensity of peaks arising from cellular components present on cervical smears differ according to stage of maturation, normality of the cell, environment of the cervix and stage of the menstrual cycle, to name a few. It is necessary to identify the contributions of individual components of cervical cells in order to understand the spectral changes seen in the manifestation of abnormality. Weekly cervical smears were obtained from participants to study the hormonal influences on cervical epithelium. Subcellular fractionation of HeLa cells and epithelial cells of cervical smears was performed to isolate nuclei and obtain IR spectra. This was to assess the contribution of nucleic acids in IR spectra of cervical smears.

Chapter Five is divided into four sections:

1. A continuation of the investigation into multivariate statistical techniques started as a collaboration with Bayden Wood, who presented the initial findings as part of his PhD dissertation [148].
2. A continuation of the investigation of the ability of artificial neural networks to classify and predict IR spectra of normal and abnormal cervical smears with high sensitivity and specificity, presented in Romeo *et al.* [147].
3. An investigation of the chemical removal of potential confounding variables from cellular deposits by lysing blood components, including leukocytes associated with inflammation and platelets, which cause cell aggregation and inhomogeneity of sample deposits.
4. An investigation into the spectroscopic and statistical influences of the presence of confounding variables in cervical smears. This is a continuation of our earlier study, which identified leukocytes including B- and T- lymphocytes, macrophages, polymorphs and monocytes, fibroblasts and connective tissue as potential confounding variables in the diagnosis of cervical cancer by IR spectroscopy

[140]. The present study investigates the potential of endocervical cells, benign cellular changes including inflammation, bacterial and yeast infections to confound spectroscopic diagnosis of cervical cancer.

1.6 REFERENCES

1. *Cancer facts and figures*. 1995, American Cancer Society. (Cited in: Tumer, K., *et al.*, Ensembles of radial basis function networks for spectroscopic detection of cervical precancer. *IEEE Transactions on Biomedical Engineering*, 1998. **45**(8): p. 953-961).
2. *Cancer Stats*, Australian Institute of Health and Welfare. 1999. http://www.aihw.gov.au/
3. Milder, J., Introductory remarks from the conference on early cervical neoplasia, 1968. *Obstet. and Gynecol. Surv.*, 1969. **24**: p. 679-680.
4. Marchetti, A., Biographic and personal recollections of George N. Papanicolaou. *Obstet. and Gynecol. Surv.*, 1969. **24**: p. 680-684.
5. Papanicolaou, G. and H. Traut, The diagnostic value of vaginal smears in carcinoma of the uterus. *Am. J. Obstet. Gynecol.*, 1941. **42**: p. 193-206.
6. Schiller, W., Early diagnosis of carcinoma if the cervix. *Surgery, Gynecology and Obstetrics*, 1933. **56**: p. 210-222.
7. Wied, G., Importance of the site from which vaginal cytologic smears are taken. *Am. J. Clin. Pathol.*, 1955. **25**: p. 742-750.
8. Ayre, J.E., *Cancer Cytology of the Uterus: Introducing a Concept of Cervical Cell Pathology*, New York: Grune & Stratton, Inc. 1951.
9. DeMay, R., *The Art and Science of Cytopathology*. Vol. 1. Chicago: American Society of Clinical Pathology Press 1995.
10. Koss, L., Cervical (Pap) smear: new directions. *Cancer*, 1993. **71**(4 Suppl): p. 1406-1412.
11. Koss, L., The Papanicolaou test for cervical cancer detection: A triumph and a tragedy. *JAMA*, 1989. **261**: p. 737-743.
12. Frost, J., Diagnostic accuracy of "cervical smears". *Obstet. Gynecol. Surv.*, 1969. **24**: p. 893-908.
13. Spitzer, M., Cervical screening adjuncts: recent advances. *Am. J. Obstet. Gynecol.*, 1998. **179**(2): p. 544-556.
14. *CancerNet*. The National Cancer Institute. 1996. http://cancer.med.upenn/pdq/100103.html

15. Oster, A., Natural history of cervical intraepithelial neoplasia: A critical review. *Int. J. Gynecol. Pathol.*, 1993. **12**(2): p. 186-192.

16. Hafez, E.S.E., Structural and ultrastructural parameters of the uterine cervix. *Obstetrical and Gynecological Survey*, 1982. **37**(8): p. 507-516.

17. *Glossary of terms used in colposcopy*. Oncolink. 1994. http://www.oncolink.upenn.edu/

18. Kivlahan, C., Papanicolaou smears without endocervical cells: Are they inadequate? *Acta Cytologica*, 1986. **30**(3): p. 258-260.

19. Sterrett, G., *The Australian (modified Bethesda) system for reporting gynaecological (cervical) cytology.*. National Cervical Screening Program: Melbourne. 1996. Courtesy of the Victorian Cytology Service.

20. Elias, A., *et al.*, The significance of endocervical cells in the diagnosis of cervical epithelial changes. *Acta Cytologica*, 1983. **27**(3): p. 225-229.

21. Vooijs, G., *et al.*, The influence of sample takers on the cellular composition of cervical smears. *Acta Cytol.*, 1986. **30**(3): p. 251-257.

22. Doornewaard, H. and Y. van der Graaf, Contribution of the Cytobrush to determining cellular composition of cervical smears. *J. Clin. Pathol.*, 1990. **43**: p. 393-396.

23. Kurman, R. and D. Solomon, *The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses*. New York: Springer-Verlag 1994.

24. Koss, L., Cytologic evaluation of the uterine cervix: Factors influencing its accuracy. *Pathologist*, 1982. **36**: p. 401-407.

25. Isacson, C. and R. Kurman, The Bethesda System: A new classification for managing Pap smears. *Contemporary Ob/Gyn*, 1995. **40**(6): p. 67-74.

26. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 3 ed. Philadelphia: J B Lippincott Company 1979.

27. Takahashi, M., *Color Atlas of Cancer Cytology*. 2 ed. Tokyo: IGAKU-SHOIN Medical Publishers, Inc. 1981.

28. Riotton, G. and W. Christopherson, *Cytology of the Female Genital Tract*. Vol. 8. Geneva: World Health Organisation 1973.

29. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 2 ed. Philadelphia: J B Lippincott Company 1968.

30. Weinberg, R., How cancer arises. *Scientific American*, 1996. **September**: p. 32-47.

31. Schiffman, M., *et al.*, Epidemiologic evidence that Human Papillomavirus causes most cervical intraepithelial neoplasia. *J. Natl. Cancer Inst.*, 1993. **85**: p. 958-64.

32. Kjaer, S.K., *et al.*, Human Papillomavirus - the most significant risk determinant of cervical intraepithelial neoplasia. *Int. J. Cancer*, 1996. **65**: p. 601-606.

33. Vernon, S., E. Unger, and W. Reeves, Human Papillomavirus and cervical cancer. *Curr. Probl. Obstet. Gynecol. Fertil.*, 1998. **21**(4): p. 104-124.

34. Franco, E., T. Rohan, and L. Villa, Epidemiologic evidence and Human Papillomavirus infection as a necessary cause of cervical cancer. *Journal of the National Cancer Institute*, 1999. **91**(6): p. 506-511.

35. Murthy, N.S. and A. Mathew, Risk factors for pre-cancerous lesions of the cervix. *European Journal of Cancer Prevention*, 2000. **9**: p. 5-14.

36. Das, B.C., *et al.*, Cancer of the uterine cervix and Human Papillomavirus infection. *Current Science*, 2000. **78**(1): p. 52-63.

37. Brisson, J., C. Morin, and M. Fortier, Risk factors for cervical intraepithelial neoplasia: differences between low- and high-grade lesions. *American Journal of Epidemiology*, 1994. **140**(8): p. 700-710.

38. Cuzick, J., *et al.*, Human Papillomavirus type 16 DNA in cervical smears as predictor of high-grade cervical cancer. *The Lancet*, 1992. **339**: p. 959-960.

39. Koutsky, L., K. Holmes, and C. Critchlow, A cohort study of the risk of cervical intraepithelial neoplasia grade 2 or 3 in relation to Papillomavirus infection. *New England Journal of Medicine*, 1992. **327**(18): p. 1272-1278.

40. Mitchell, M., *et al.*, Cervical Human Papillomavirus infection and intraepithelial neoplasia: A review. *Journal of the National Cancer Institute Monographs*, 1996. **21**: p. 17-25.

41. Szarewski, A., *et al.*, Effect of smoking cessation on cervical lesion size. *The Lancet*, 1996. **347**: p. 941-943.

42. Prokopczyk, B., *et al.*, Identification of Tobacco-specific carcinogen in the cervical mucus of smokers and nonsmokers. *Journal of the National Cancer Institute*, 1997. **89**(12): p. 868-873.

43. Waggoner, S. and X. Wang, Effect of nicotine on proliferation of normal, malignant, and Human Papillomavirus-transformed human cervical cells. *Gynecologic Oncology*, 1994. **55**: p. 91-95.

44. Gay, D., L. Donaldson, and J. Goellner, False-negative results in cervical cytologic studies. *Acta Cytologica*, 1985. **29**(6): p. 1043-1046.

45. Smolka, H. and H. Soost, *An Outline and Atlas of Gynaecological Cytodiagnosis.* 2 ed. London: Edward Arnold. 1965.

46. Morell, N., *et al.*, False-negative cytology rates in patients in whom invasive cervical cancer subsequently developed. *Obstet. Gynecol.*, 1982. **60**: p. 41-45.

47. van der Graaf, Y., *et al.*, Screening errors in cervical cytologic screening. *Acta Cytologica*, 1987. **31**(4): p. 434-438.

48. van der Graaf, Y. and G. Vooijs, False negative rate in cervical cytology. *J. Clin. Pathol.*, 1987. **40**: p. 438-442.

49. Vooijs, P., Y. van der Graaf, and A. Elias, Cellular composition of cervical smears in relation to the day of the menstrual cycle and the method of contraception. *Acta Cytologica*, 1987. **31**(4): p. 417-426.

50. Fahey, M.T., L. Irwig, and P. Macaskill, Meta-analysis of pap test accuracy. *Am. J. Epidemiol.*, 1995. **141**: p. 680-9.

51. Rubio, C., False negatives in cervical cytology: Can they be avoided? *Acta Cytologica*, 1981. **25**(2): p. 199-201.

52. Tezuka, F., *et al.*, Numerical accounts of epithelial cells collected, smeared and lost in the conventional Papanicolaou smear preparation. *Acta Cytologica*, 1995. **39**: p. 837-838.

53. Fung, M.F.K., *et al.*, Comparison of Fourier-transform infrared spectroscopic screening of exfoliated cervical cells with standard Papanicolaou screening. *Gynecologic Oncology*, 1997. **66**: p. 10-15.

54. Mango, L., Computer-assisted cervical cancer screening using neural networks. *Cancer Letters*, 1994. **77**: p. 155-162.

55. Boon, M. and L. Kok, Neural network processing can provide means to catch errors that slip through human screening of Pap smears. *Diagnostic Cytopathology*, 1993. **9**(4): p. 411-416.

56. Boon, M. and L. Kok, Classification of cells in cervical smears, in *Applications of Neural Networks*, A. Murray, Editor. Kluwer Academic Publishers: Netherlands. 1995

57. Sherman, M., *et al.*, PAPNET analysis of reportedly negative smears preceding the diagnosis of a high-grade squamous intraepithelial lesion or carcinoma. *Mod. Pathol.*, 1994. **7**(5): p. 578-581.

58. Schechter, C., Cost-effectiveness of rescreening conventionally prepared cervical smears by PAPNET testing. *Acta Cytologica*, 1996. **40**(6): p. 1272-1282.

59. Doornewaard, H., *et al.*, Negative cervical smears before CIN 3/carcinoma. *Acta Cytologica*, 1997. **41**(1): p. 74-78.

60. Zahniser, D. and P. Sullivan, CYTYC Corporation. *Acta Cytologica*, 1996. **40**(1): p. 37-44.

61. Hutchinson, M., *et al.*, Study of cell loss in the conventional Papanicolaou smear. *Acta Cytologica*, 1992. **36**(4): p. 577.

62. Goodman, A. and M. Hutchinson, Cell surplus on sampling devices after routine cytologic smears. *The Journal of Reproductive Medicine*, 1996. **41**(4): p. 239-240.

63. Hutchinson, M., C. Cassin, and H. Ball, The efficacy of an automated preparation device for cervical cytology. *Am. J. Clin. Pathol.*, 1991. **96**(3): p. 300-305.

64. Tezuka, F., *et al.*, Diagnostic efficacy and validity of the ThinPrep method in cervical cytology. *Acta Cytologica*, 1996. **40**(3): p. 513-518.

65. Ferenczy, A., *et al.*, Conventional cervical cytologic smears vs. ThinPrep smears. *Acta Cytologica*, 1996. **40**(6): p. 1136-1142.

66. Lee, K., *et al.*, Comparison of conventional Papanicolaou smears and a fluid-based, thin-layer system for cervical cancer screening. *Obstetrics and Gynecology*, 1997. **80**(2): p. 278-284.

67. Roberts, J., *et al.*, Evaluation of the ThinPrep pap test as an adjunct to the conventional pap smear. *M.J.A.*, 1997. **167**(3): p. 466-469.

68. Sherman, M., *et al.*, Cervical specimens collected in liquid buffer are suitable for both cytologic screening and ancillary Human Papillomavirus testing. *Cancer*, 1997. **81**(2): p. 89-97.

69. Linder, J., Recent advances in thin-layer cytology. *Diagnostic Cytopathology*, 1998. **18**(1): p. 24-32.

70. Sheets, E.E., *et al.*, Colposcopically directed biopsies provide a basis for comparing the accuracy of ThinPrep and Papanicolaou smears. *Journal of Gynecologic Techniques*, 1995. **1**(1): p. 27-33.

71. Stafl, A., Cervicography: A new method for cervical cancer detection. *Am. J. Obstet. Gynecol.*, 1981. **139**(7): p. 815-825.

72. Ferris, D., *et al.*, Cervicography: Adjunctive cervical cancer screening by primary care clinicians. *The Journal of Family Practice*, 1993. **37**(2): p. 158-164.

73. Tawa, K., *et al.*, A comparison of the Papanicolaou smear and the cervigram: sensitivity, specificity and cost analysis. *Obstet. Gynecol*, 1988. **71**: p. 229-35.

74. Gundersen, J.H., C.W. Schauberger, and N.R. Rowe, The Papanicolaou smear and the cervigram, a preliminary report. *J. Reprod. Med.*, 1988. **33**: p. 46-48.

75. Coppleson, M., *et al.*, An electronic approach to the detection of pre-cancer and cancer of the uterine cervix: a preliminary evaluation of Polarprobe. *Int. J. Gynecol. Cancer*, 1994. **4**: p. 79-83.

76. Mantsch, H. and M. Jackson, Molecular spectroscopy in biodiagnostics (from Hippocrates to Herschel and beyond). *Journal of Molecular Structure*, 1995. **347**: p. 187-206.

77. Haris and Chapman, Does Fourier transform infrared spectroscopy provide useful information on protein structures? *TIBS*, 1992. **17**.

78. Hendra, P., C. Jones, and G. Warres, *FT Raman Spectroscopy: Instrumentation and Chemical Applications*. New York: Ellis Harwood. 1991.

79. Redd, D.B., *et al.*, Raman spectroscopic characterization of human breast tissues: implications for breast cancer diagnosis. *Applied Spectroscopy*, 1993. **47**(6): p. 787-791.

80. Frank, C.J., *et al.*, Characterization of human breast biopsy specimens with near-IR Raman spectroscopy. *Anal. Chem.*, 1994. **66**: p. 319-326.

81. Kline, N.J. and P.J. Treado, Raman chemical imaging of breast tissue. *Journal of Raman spectroscopy*, 1997. **28**: p. 119-124.

82. Manoharan, R., *et al.*, Raman spectroscopy and fluorescence photon migration for breast cancer diagnosis and imaging. *Photochemistry and Photobiology*, 1998. **67**(1): p. 15-22.

83. Yazdi, Y., *et al.*, Resonance Raman spectroscopy at 257 nm excitation of normal and malignant cultured breast and cervical cells. *Applied Spectroscopy*, 1999. **53**(1): p. 82-85.

84. Liu, C.H., *et al.*, Raman, fluorescence and time resolved light scattering as optical diagnostic techniques to separate diseased and normal biomedical media. *J. Photochem. Photobiol.*, 1992. **16**: p. 187-209.

85. Gniadecka, M., H.C. Wulf, and N.N. Mortensen, Diagnosis of basal cell carcinoma by Raman spectroscopy. *Journal of Raman Spectroscopy*, 1997. **28**: p. 125-129.

86. Gniadecka, M., *et al.*, Distinctive molecular abnormalities in benign and malignant skin lesions: studies by Raman spectroscopy. *Photochemistry and Photobiology*, 1997. **66**(4): p. 418-423.

87. Mahadevan-Jansen, A., *et al.*, Near-infrared Raman spectroscopy for the detection of cervical precancers. *Photochemistry and Photobiology*, 1998 **68**(1):123-132.

88. Mahadevan-Jansen, A. and R. Richards-Kortum, Raman spectroscopy for the detection of cancers and precancers. *Journal of Biomedical Optics*, 1996. **1**(1): p. 31-70.

89. Pavia, D., G. Lampman, and G. Kriz, *Introduction to Spectroscopy*. 2 ed. Orlando: Saunders College publishing. 1996.

90. Abraham, R.J., J. Fisher, and P. Loftus, *Introduction to NMR Spectroscopy*. 3 ed. New York: John Wiley & Sons. 1998.

91. Decertaines, J., High resolution nuclear magnetic resonance spectroscopy in clinical biology: application in oncology. *Anticancer Research*, 1996. **16**: p. 1325-1332.

92. Robinson, S.P., *et al.*, Nuclear magnetic resonance spectroscopy of cancer. *The British Journal of Radiology*, 1997. **70**: p. S60-S69.

93. Halliday, K.R., L.O. Sillerud, and C.M. Fenoglio-Preiser, In vitro ^{13}C MRS of tumor tissues, cells and plasma ^{13}C MRS of excised tissues. *Adv. Pathol.*, 1989. **2**: p. 223-229.

94. Holmes K., and C. Mountford, Identification of triglyceride in malignant cells. *Journal of Magnetic Resonance*, 1991. **93**: p. 407-409.

95. Kriat, M., *et al.*, Analysis of plasma lipids by NMR spectroscopy: application to modifications induced by malignant tumors. *Journal of Lipid Research*, 1993. **34**: p. 1009-1019.

96. Smith, I. and D. Blandford, Diagnosis of cancer in humans by ^{1}H NMR of tissue biopsies. *Biochem. Cell Biol.*, 1998. **76**: p. 472-476.

97. Stubbs, M. and J.R. Griffiths, Monitoring cancer by magnetic resonance. *British Journal of Cancer*, 1999. **80**(Suppl 1): p. 86-94.

98. Doyle, V.L., S.J. Barton, and J.R. Griffiths, ^{31}P and ^{1}H MRS of human cancer. *Current Science*, 1999. **76**(6): p. 772-776.

99. Lenk, R., NMR studies of the physiological states in cancer cells and tissues. *Archives des Sciences*, 1996. **49**: p. 51-57.

100. Mountford, C., *et al.*, Uterine cervical punch biopsy specimens can be analysed by ^{1}H MRS. *Magnetic Resonance in Medicine*, 1990. **13**: p. 324-331.

101. Delikatny, J., *et al.*, Proton MR and human cervical neoplasia: ex vivo spectroscopy allows distinction of invasive carcinoma of the cervix from carcinoma in situ and other preinvasive lesions. *Radiology*, 1993. **188**: p. 791-796.

102. Mountford, C., *et al.*, Human cancers detected by proton MRS and chemical shift imaging ex vivo. *Anticancer Research*, 1996. **16**: p. 1521-1532.

103. Mountford, C., *et al.*, Cancer pathology in the year 2000. *Biophysical Chemistry*, 1997. **68**: p. 127-135.

104. Wallace, J.C., *et al.*, Classification of ^{1}H MR spectra of biopsies from untreated and recurrent ovarian cancer using linear discriminant analysis. *MRM*, 1997. **38**: p. 569-576.

105. Massuger, L., *et al.*, ^{1}H-Magnetic resonance spectroscopy. *Cancer*, 1998. **82**(9): p. 1726-1730.

106. Kunnecke, B., *et al.*, Proton magnetic resonance and human cervical neoplasia. II. Ex vivo chemical-shift microimaging. *Journal of Magnetic Resonance*, 1994. **104**: p. 135-142.

107. Cohen, J.S., *et al.*, A history of biological applications of NMR spectroscopy. *Progress in Nuclear Magnetic Spectroscopy*, 1995. **28**: p. 53-85.

108. deSouza, N.M., *et al.*, Value of magnetic resonance imaging with an endovaginal receiver coil in the pre-operative assessment of Stage I and IIa cervical neoplasia. *British Journal of Obstetrics and Gynecology*, 1998. **105**: p. 500-507.

109. Tumer, K., *et al.*, Ensembles of radial basis function networks for spectroscopic detection of cervical precancer. *IEEE Transactions on Biomedical Engineering*, 1998. **45**(8): p. 953-961.

110. Mitchell, M.F., *et al.*, Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstetrics and Gynecology*, 1999. **94**(5, Part 2): p. 889-896.

111. Ramanujam, N., *et al.*, Cervical precancer detection using a multivariate statistical algorithm based on laser induced fluorescence spectra at multiple excitation wavelengths. *Photochem. Photobiol.*, 1996. **64**(4): p. 720-735.

112. Agrawal, A., *et al.*, Fluorescence spectroscopy of the cervix: influence of acetic acid, cervical mucus and vaginal medications. *Lasers in Surgery and Medicine*, 1999. **25**: p. 237-249.

113. Jackson, M. and H. Mantsch, IR spectroscopy: An insight into diseased tissue. *Analysis*, 1995. **October/November**: p. S10-S15.

114. Mantsch, H. and McElhaney, Applications of infrared spectroscopy to biology and medicine. *Journal of Molecular Structure*, 1990. **217**: p. 347-362.

115. Wong, P., R. Wong, and M.F.K. Fung, Pressure-tuning FT-IR study of human cervical tissues. *Applied Spectroscopy*, 1993. **47**(7): p. 1058-1063.

116. Wong, P., *et al.*, Infrared spectroscopy of exfoliated human cervical cells: Evidence of extensive structural changes during carcinogenesis. *Proc. Natl. Acad. Sci.*, 1991. **88**: p. 10988-10992.

117. Wong, P. and B. Rigas, Infrared spectra of microtome sections of human colon tissues. *Applied Spectroscopy*, 1990. **44**(10): p. 1715-1718.

118. Wong, P., M. Cadrin, and S. French, Distinctive infrared spectral features in liver tumor tissue of mice: Evidence of structural modifications at the molecular level. *Experimental and Molecular Pathology*, 1991. **55**: p. 269-284.

119. Wong, P., E. Papavassiliou, and B. Rigas, Phosphodiester stretching bands in the infrared spectra of human tissues and cultured cells. *Applied Spectroscopy*, 1991. **45**(9): p. 1563-1567.

120. Wong, P., S. Lacelle, and H. Yazdi, Normal and malignant human colonic tissues investigated by pressure-tuning FT-IR spectroscopy. *Applied Spectroscopy*, 1993. **47**(11): p. 1330-1836.

121. Wong, P., *et al.*, Characterization of exfoliated cells and tissues from human endocervix and ectocervix by FTIR and ATR/FTIR spectroscopy. *Biospectroscopy*, 1995. **1**: p. 357-364.

122. Jackson, M. and H. Manstch, Biomembrane structure from Fourier transform infrared spectroscopy. *Spectrochimica Acta*, 1993. **15**(1): p. 53-69.

123. Yazdi, H., M. Bertrand, and P. Wong, Detecting structural changes at the molecular level with Fourier transform infrared spectroscopy. *Acta Cytologica*, 1996. **40**(4): p. 664-668.

124. Rigas, B., *et al.*, Human colorectal cancers display abnormal Fourier-transform infrared spectra. *Proc. Natl. Acad. Sci.*, 1990. **87**: p. 8140-8144.

125. Wood, B., *et al.*, An investigation into FTIR spectroscopy as a biodiagnostic tool for cervical cancer. *Biospectroscopy*, 1996. **2**: p. 1-11.

126. Cohenford, M., *et al.*, Infrared spectroscopy of normal and abnormal cervical smears: Evaluation by principal component analysis. *Gynecologic Oncology*, 1997. **66**: p. 59-65.

127. Cohenford, M. and B. Rigas, Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: Implications for tumor biology. *Proc. Natl. Acad. Sci.*, 1998. **95**: p. 15327-15332.

128. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix. *Biospectroscopy*, 1998. **4**(1): p. 47-53.

129. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. III. Spectral differences between squamous and columnar tissue and cells from the human cervix. *Biospectroscopy*, 1997. **3**(4): p. 253-257.

130. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. II. A comparative study of spectra of biopsies of cervical squamous epithelium and of exfoliated cervical cells. *Biospectroscopy*, 1998. **4**(1): p. 55-59.

131. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. IV. Detection of dysplastic and neoplastic changes of human cervical tissue via infrared microspectroscopy. *Cellular and Molecular Biology*, 1998. **44**(1): p. 219-229.

132. Benedetti, E., *et al.*, A new approach to the study of human solid tumor cells by means of FT-IR microspectroscopy. *Applied Spectroscopy*, 1990. **44**(8): p. 1276-1280.

133. Yano, K., *et al.*, Evaluation of glycogen level in human lung carcinoma tissues by an infrared spectroscopic method. *Cancer Letters*, 1996. **110**: p. 29-34.

134. Benedetti, E., *et al.*, Infrared characterization of nuclei isolated from normal and leukemic (B-CLL) lymphocytes: Part III. *Applied Spectroscopy*, 1986. **40**(1): p. 39-43.

135. Benedetti, E., *et al.*, Determination of the relative amount of nucleic acids and proteins in leukemic and normal lymphocytes by means of Fourier transform infrared microspectroscopy. *Applied Spectroscopy*, 1997. **51**(6): p. 792-797.

136. Boydston-White, S., *et al.*, Infrared spectroscopy of human tissue. V. Infrared spectroscopic studies of Myeloid Leukemia (ML-1) cells at different phases of the cell cycle. *Biospectroscopy*, 1999. **5**: p. 219-227.

137. Rogers, J., Physiology of menstruation, in *Endocrine and Metabolic Aspects of Gynecology*. W B Saunders: London. p. 1-15. 1963.

138. Cournoyer, R., J. Shearer, and D. Anderson, Fourier transform infrared analysis below the one-nanogram level. *Analytical Chemistry*, 1977. **49**(14): p. 2275-2277.

139. Lowry, S.R., The analysis of exfoliated cervical cells by infrared microspectroscopy. *Cellular and Molecular Biology*, 1998. 44(1): p. 169-177.

140. Wood, B., *et al.*, FTIR microspectroscopic study of cell types and potential confounding variables in screening for cervical malignancies. *Biospectroscopy*, 1998. 4(2): p. 75-91.

141. Campbell, N.A., Animal Reproduction, in *Biology*. Benjamin/Cummings Publishing Co. Inc.: Sydney. p. 930-954. 1990.

142. Wachtel, E.G., Exfoliative cytology in gynaecological practice. London: Butterworth & Co. 1969.

143. Patten, S.F., The normal uterine cervix, in *Diagnostic Pathology of the Uterine Cervix*. S Karger: Sydney. p. 30-41. 1978.

144. Smolka, H. and H.J. Soost, *An Outline and Atlas of Gynecological Cytodiagnosis*. 2 ed. London: Edward Arnold. 1965.

145. Reynolds, J.E.F., *Martindale the Extra Pharmacopoeia*. 30 ed. London: The Pharmaceutical Press. p1168-1178. 1993.

146. Gibbons, R.A. and P. Mattner, Some aspects of the chemistry of cervical mucus. *International Journal of Fertility*, 1966. **11**(4): p. 366-372

147. Romeo, M., *et al.*, Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. *Cellular and Molecular Biology*, 1998. 44(1): p. 179-187.

148. Wood, B.R., *Biomedical Applications of Fourier Transform-Infrared Microspectroscopy*. PhD Dissertation, in Chemistry. Monash University, Australia.1998.

CHAPTER 2

AN INTRODUCTION TO INFRARED SPECTROSCOPY AND MULTIVARIATE STATISTICS

2 AN INTRODUCTION TO INFRARED SPECTROSCOPY AND MULTIVARIATE STATISTICS

2.1 THEORY OF INFRARED SPECTROSCOPY

Spectroscopy is the study of the interaction of electromagnetic radiation with matter and deals with transitions of molecules from one state or energy level to another. Figure 2.1 shows the electromagnetic spectrum for different types of radiation.

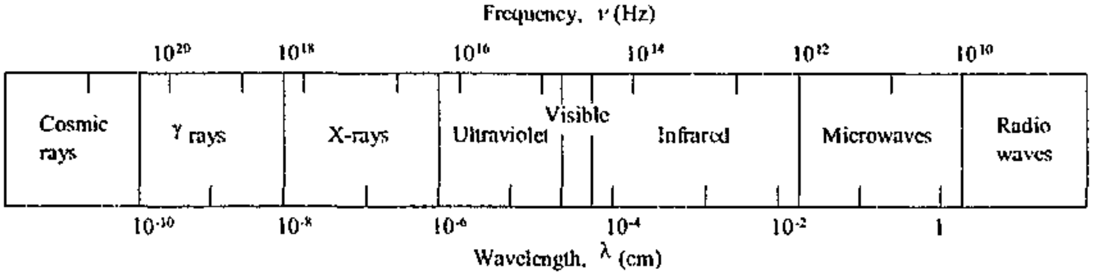

Figure 2.1 The electromagnetic spectrum. Redrawn from[1].

All molecules undergo vibrational motion and the energy of most of these molecular vibrations corresponds to that of the IR region of the electromagnetic spectrum, Table 2.1.

Table 2.1 Infrared spectral regions [2].

Region	**Wavelength (λ) Range, μm**	**Wavenumber ($\bar{\nu}$) Range, cm^{-1}**	**Frequency (ν) Range, Hz**
Near	0.78-2.5	12800-4000	3.8 x 10^{14}-1.2 x 10^{14}
Middle	2.5-50	4000-200	1.2 x 10^{14}-6.0 x 10^{12}
Far	50-1000	200-10	6.0 x 10^{12}-3.0 x 10^{11}

Electromagnetic radiation can be thought of as having properties of particles and waves, and consists of perpendicular oscillating electric (ε) and magnetic (B) fields, Figure 2.2.

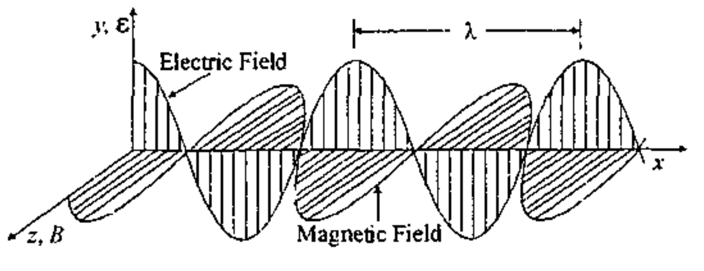

Figure 2.2 Plane polarised electromagnetic radiation of wavelength λ propagating along the *x*-axis. Redrawn from [3].

2.1.1 THE DIPOLE MOMENT

In order to absorb infrared radiation, a molecule must undergo a net change in dipole moment as a consequence of its vibrational motion. The electric dipole moment, μ, is given by $\mu = qr$ where q is the charge on the atoms separated by a distance r. The dipole moment is a vector quantity and is measured in SI units of coulomb meter (Cm). An electric dipole moment is present in a molecule when there is a difference between the centres of positive and negative charge. An electric dipole may be permanent, due to differences in electronegativities among the atoms in the molecule, or induced (temporary), caused by distortions of the molecule due to either interactions with other molecules or to intramolecular motions [4]. The dipole moment of a vibrating diatomic molecule is the sum of two components, a permanent dipole moment, μ_o, which is due to the partial electronic charge on each of the atoms when they are at equilibrium, and a component that changes as the molecule undergoes vibration, $\mu(q)$: $\mu = \mu_o + \mu(q)$ where q is the displacement from equilibrium. Since the dipole moment μ is the product of the partial charge on the atoms and the internuclear separation, it should go to zero at small internuclear separations or when the atoms are separated. If this is represented by a curve, then the slope of the curve, $d\mu/dq$, is essentially constant over the amplitude of the oscillation. This is represented by $(d\mu/dq)_o$ which is the dipole moment change per unit displacement from equilibrium. The dipole moment term is written as [4]:

$$\mu = \mu_o + \left(\frac{d\mu}{dq}\right)_o q \qquad \textbf{Equation 2.1}$$

If the frequency of the vibration of the electromagnetic radiation matches the natural frequency of the molecule a net transfer of energy occurs, resulting in a change in the amplitude of the molecular vibration and absorption of the radiation.

Homonuclear diatomic and polyatomic molecules with a centre of inversion cannot have permanent dipole moments since nuclei attract the electrons equally. Heteronuclear diatomic and unsymmetrical molecules have permanent dipole moments since one atom will be more electronegative than the other/s and will have a net negative charge [3].

The molecular dipole moment of a heteronuclear molecule oscillates about equilibrium as two atoms with net negative and positive charges move back and forth. This oscillating moment is able to absorb energy from an electric field if the oscillations of the field occur at the same frequency. Different molecules absorb infrared radiation at different frequencies and therefore IR spectroscopy is used to identify 'compounds' and possibly their structure by the frequencies of the absorptions of the molecules present. The absorptions of each type of bond or functional group are found in certain regions of the infrared range and as such absorption in a particular region is indicative of a bond or functional group.

Excitation can result in a molecule undergoing an increase in vibrational amplitude, or rotational frequency. Since the rotational energies of molecules are smaller than vibrational energies, vibration and rotation usually occur simultaneously [2]. Models explaining the mechanics of rotation and vibration in diatomic molecules have been formed in order to understand the spectra produced from the excitation of molecules with infrared radiation.

2.1.2 ROTATION

The simplest model for explaining the rotation of a diatomic molecule is the rigid rotor, Figure 2.3.

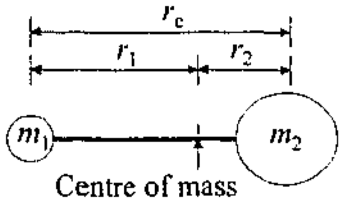

Figure 2.3 The rigid rotor. Redrawn from [1].

This model supposes that the two nuclei are fixed at r_e, equilibrium separation. If the nuclei have masses m_1 and m_2, the molecule will rotate about the centre of mass defined such that $m_1r_1 = m_2r_2$.

It is possible to completely resolve the fine structure due to rotational transitions in the infrared region using gaseous samples. The infrared bands of liquid or solid samples however, are often broadened due to rotational coupling with vibration [5].

Rotational energies of a diatomic molecule can be calculated using the reduced mass, μ, (not the same as the dipole moment, Equation 2.10) and the moment of inertia, I, of the molecule. The theory of rotational vibration is presented in Harris *et al* [3].

2.1.3 VIBRATION

A simple model for the vibration of a diatomic molecule, Figure 2.4, can be formed if the bond between the two nuclei is thought to behave as a spring that obeys Hooke's Law, Equation 2.2.

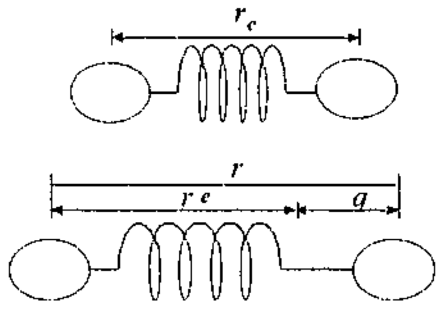

Figure 2.4 A model for the vibration of a diatomic molecule. Called a harmonic oscillator because it obeys Hooke's Law. Redrawn from [3].

$$\text{restoring force} = f = -kq$$ **Equation 2.2**

where k is the force constant (SI units is newtons per metre). The potential energy is given by:

$$V = (1/2)kq^2$$ **Equation 2.3**

The harmonic oscillator model predicts molecules to have discrete vibrational energy levels characterised by the quantum number v:

$$E_v = (v + 1/2)\frac{h}{2\pi}\sqrt{k/\mu} \equiv (v + 1/2)h\nu$$ **Equation 2.4**

where:

$$\nu = \frac{1}{2\pi}\sqrt{k/\mu}$$ **Equation 2.5**

In the lowest vibrational state, the ground state ($\upsilon = 0$), the molecule has zero point energy, $E_0 = (1/2)h\nu$, whereas rotational energy levels have a ground state energy of zero. The harmonic oscillator model, Figure 2.5, predicts a diatomic molecule to have equally spaced vibrational energy levels, starting $(1/2)h\nu$ from the bottom of the potential well with the spacing between levels equal to $h\nu$. The mid IR vibrational spectra of diatomic molecules usually result from excitation from the $\upsilon = 0$ to the $\upsilon = 1$ energy levels. Therefore the difference in energy, ΔE, between these two levels is $h\nu$. This value can be used to calculate the force constant (k) of a chemical bond, which is an indication of bond strength [3].

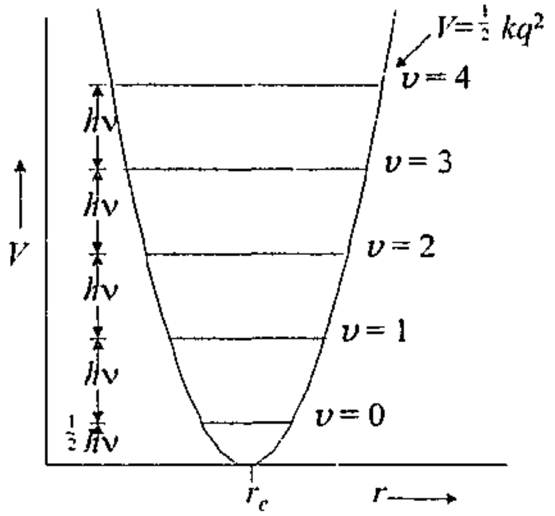

Figure 2.5 Harmonic oscillator potential well showing energy levels. Redrawn from [3].

The parabolic potential well generated by the harmonic oscillator is a poor representation of the force between diatomic molecules. A real molecule should have an asymmetric potential well, Figure 2.6. As q (Figure 2.4) decreases the nuclei come together and repel each other. At small values of q the repulsion is very strong, but as q increases the restoring force equilibrates and the molecule dissociates [3].

In vibrational spectroscopy wavenumber (cm^{-1}) units are usually used in preference to wavelength, energy or frequency. Energy (joules) can be converted to wavenumber values/(cm^{-1}) using Equation 2.6:

$$\frac{E(\text{joules})}{h(\text{joules/sec}) \cdot c(\text{m/s}) \cdot 100(\text{cm/m})} = \overline{E}(\text{cm}^{-1}) \qquad \textbf{Equation 2.6}$$

where the bar over the symbol $\overline{E}$ emphasises that the units are in wavenumbers. The wavenumber scale is preferred because of its linearity and direct proportionality with energy (Equation 2.7) and frequency (Equation 2.8):

$$E = hc\overline{\nu} \qquad \textbf{Equation 2.7}$$

$$\overline{\nu} = \frac{1}{\lambda} = \frac{\nu}{c} \qquad \textbf{Equation 2.8}$$

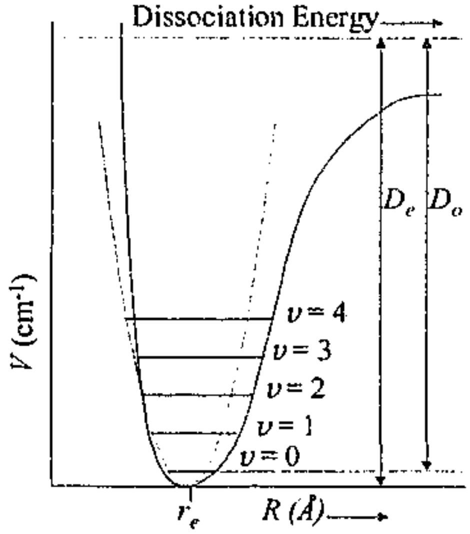

Figure 2.6 The anharmonic oscillator potential (solid line) and harmonic oscillator (dashed line). Redrawn from [3].

If a diatomic molecule is assumed to behave as an harmonic oscillator, the natural frequency of the vibration is given by:

$$\bar{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \qquad \textbf{Equation 2.9}$$

which is derived from Hooke's Law (Equation 2.2). The reduced mass, μ, is given by:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \qquad \textbf{Equation 2.10}$$

The value of k varies from one bond to another. The force constant of triple bonds are approximately three times those of single bonds, whilst the force constant of double bonds are approximately twice that of single bonds [5].

Consequently bond strength and mass of the atoms in a molecule affect the frequency at which IR absorption occurs. Since stronger bonds have a larger force constant they will vibrate at a higher frequency than weaker bonds. Bonds between atoms of higher masses will have a larger reduced mass and will vibrate at lower frequencies than bonds between lighter atoms [5].

2.1.4 VIBRATIONAL MODES

The simplest modes of vibrational motion, which gives rise to absorption in an infrared-active molecule are stretching and bending, Figure 2.7. Stretching modes take the form of symmetric and asymmetric stretches. Asymmetric stretches generally occur at higher frequencies than symmetric stretching vibrations. Bending vibrations are commonly referred to as scissoring, rocking, wagging or twisting, and occur at a lower frequency than stretching vibrations because the value of k is lower [5].

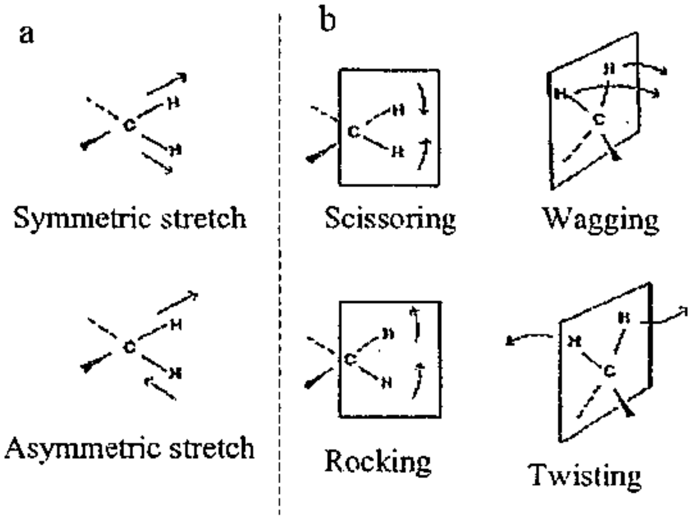

Figure 2.7 Possible stretching (a) and bending (b) vibrational modes of infrared absorption [5]. Scissoring and rocking are termed in-plane bending vibrations, whilst wagging and twisting are termed out-of-plane bending vibrations.

The absorptions mentioned above are termed fundamental absorptions because they are a result of excitation from the ground state to the lowest excited state. Whilst these absorptions give rise to strong infrared bands, weaker overtone, combination and difference bands may also be observed. Fermi resonance effects may also be seen and occur as a result of coupling between a fundamental and an overtone or combination band. Fermi resonance is often observed in carbonyl compounds [5].

2.1.5 ABSORPTION OF LIGHT.

In spectroscopy a sample is illuminated with light and the amount of light absorbed by the sample is measured as a function of the energy of the light. The amount of light absorbed by a sample can be expressed as transmittance (Equation 2.11) or absorbance (Equation 2.12). If the intensity of the light striking a sample is I_0, and the intensity of the light that passes out of the other side of the sample is I, the transmittance, T is the ratio:

$$T = I/I_o$$ **Equation 2.11**

The absorbance, A, is defined as:

$$A = \log_{10}\frac{I}{I_o} = \log_{10}(1/T) \qquad \textbf{Equation 2.12}$$

Absorbance is a useful measure of light absorption because at low absorbance values it is directly proportional to the molar concentration of the sample, c, and the length of the light path through the sample cell, ℓ, as expressed by the Beer-Lambert Law:

$$A = \in c\ell \qquad \textbf{Equation 2.13}$$

The proportionality constant, $\in$, is known as the molar absorption [6] or extinction coefficient (units M^{-1} cm^{-1}) and is a measurement of how strongly a particular sample absorbs light at a given wavelength [3].

2.2 *THE INFRARED SPECTROMETER*

An infrared spectrometer, or spectrophotometer, is used to measure the spectrum of a compound. The two types of infrared instruments commonly used are dispersive and Fourier transform (FT). Dispersive instruments make use of diffraction gratings which act as monochromators to separate polychromatic radiation into monochromatic components [7]. As a consequence dispersive spectrometers are time intensive as the detector is only able to receive information about individual spectral elements at any given time. There are two main advantages of a FT spectrometer over traditional dispersive instruments. Dispersive instruments are restricted due to the ability of the slits to receive information about a narrow band at a given time. The interferometer of the FT system receives information about the entire spectral region in each scan. This is known as the Fellget or multiplex advantage. The Jacquinot or throughput advantage is the ability of interferometers to collect large amounts of energy. The grating spectrometer requires long and narrow slits whereas the interferometer has a much larger area for the same resolving power and less attenuation of infrared radiation. Advantages resulting from this include large resolving power, high wavenumber accuracy, fast scanning time, and large scan range [8]. The advantages of FTIR so outweigh the grating spectrometers that dispersive instruments in the mid IR are no longer used to any great extent.

2.2.1 MICHELSON INTERFEROMETER

The essential instrument for modern infrared spectroscopy is a Michelson interferometer, Figure 2.8.

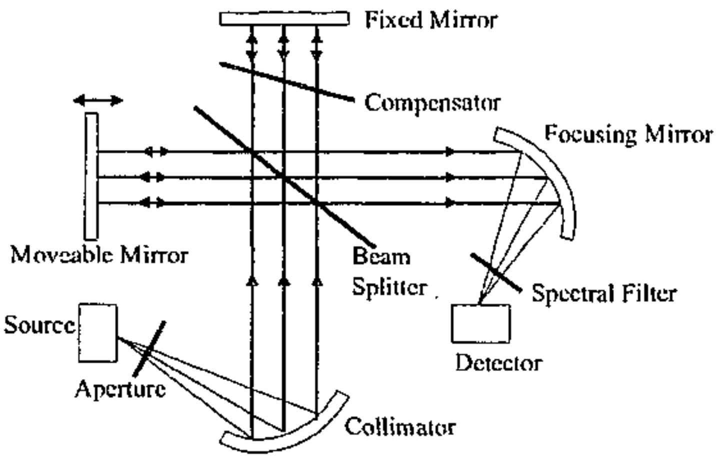

Figure 2.8 Schematic diagram of a Michelson interferometer.

Infrared light is emitted by a source and directed to a beam splitter. Ideally, the beam splitter allows half of the light to pass through, whilst the other half is reflected. The reflected beam strikes a moving mirror and returns to the beam splitter. The transmitted beam travels to a fixed mirror through distance L, is reflected and returns to the beam splitter after a total path length of $2L$. A stepper motor, the precision of which is governed by the modulated output of a HeNe laser, moves the reflecting mirror around L by successive distances x resulting in a beam with a total path length of $2(L+x)$. The beams recombine at the beam splitter, with the partial waves interfering constructively or destructively as a consequence of the path length difference ($2x$) between the wave trains. Maximum detector signal arises if the partial waves interfere constructively, i.e. when the optical retardation is an exact multiple of the wavelength λ: $2x = n\lambda$ ($n = 0,1,2,\ldots$). Minimum detector signal arises from destructive interference when $2x$ is an odd multiple of $\lambda/2$.

2.2.2 FOURIER TRANSFORMATION

The beam leaving the interferometer is passed through the sample and focussed onto the detector. The quantity measured by the detector is the intensity $I(x)$ of the combined infrared beams as a function of the mirror displacement (x) and is known as the interferogram. The dependence of the intensity $I(x)$ on the mirror displacement (x) is given by the cosine function:

$$I(x) = S(\nu)\cos(2\pi\nu x)$$ **Equation 2.14**

The interferogram is converted from an optical path difference or time domain into a frequency domain ($S(\nu)$) by means of a mathematical operation called Fourier transformation [8]:

$$S(\nu) = \int_{-\infty}^{+\infty} I(x)\cos 2\pi\nu x \cdot dx$$ **Equation 2.15**

If the interferogram is sampled, as is the case where sampling points are determined by the interference pattern of a monochromatic HeNe laser, it consists of N discrete equidistant points and discrete Fourier transformation (DFT, Equation 2.16) must be performed [8]. As a consequence of performing discrete rather than continuous Fourier transformation of the interferogram $I(x)$, the continuous variables scan length x and frequency ν become the discrete variables $n.\ \Delta x$ and $k.\ \Delta\nu$ [9].

$$S(k \cdot \Delta\nu) = \sum_{n=0}^{N-1} I(n\Delta x)\exp(i2\pi n \cdot \frac{k}{N})$$ **Equation 2.16**

The mathematical process of Fourier transformation assumes infinite boundaries, whereas discrete Fourier transformation (DFT) performs integration over a finite range. DFT can lead to the spectral artifacts known as the picket fence effect, aliasing and leakage [8].

The picket fence effect is seen when the interferogram contains frequencies that do not coincide with the frequency sample points $k*\Delta\nu$. This effect can be avoided by zero filling, which causes spectral interpolation by adding zeroes to the end of the interferogram [8].

The process of DFT produces a spectrum and its mirror image or alias. The first N/2 points of the DFT represent the spectrum whilst the remainder represents the mirror image. The point at which the mirror image begins is called the folding or Nyquist number. Aliasing presents a problem if an overlap occurs between the spectrum and its alias or if the spectrum is non-zero above the Nyquist number. Alias overlap can be avoided by increasing the number of sampling points in the interferogram, i.e. reducing Δx [8].

The truncation of the interferogram at finite optical path difference causes convolution of the true interferogram with a boxcar function, resulting in leakage or the appearance of positive and negative sidelobes in the instrumental lineshape. Side lobes can be removed by apodisation, where the interferogram is multiplied by a suitable function before the FT is carried out [10]. The problem of leakage can be avoided by truncating the interferogram less abruptly than with the rectangular boxcar function. Apodisation functions and the instrumental lineshape produced are presented in Figure 2.9. The use of apodisation functions other than the boxcar results in a loss of resolution.

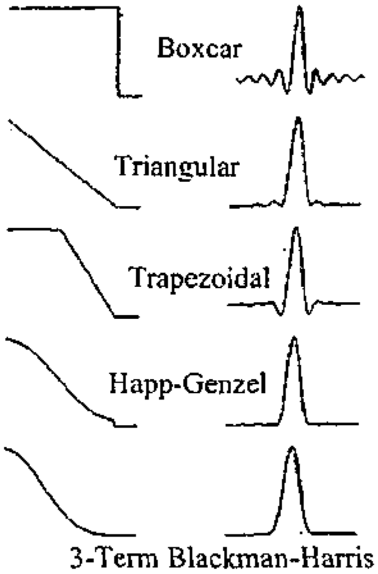

Figure 2.9 Apodisation functions and resulting instrumental lineshapes [8].

The measured interferogram is generally not symmetrical about the centreburst, $x = 0$. Asymmetry is caused as a result of sampling positions not coinciding with zero path difference, the measurement of a 'one sided' interferogram, or wavenumber dependent phase delays [8]. Fourier transformation of the asymmetrical interferogram produces a complex spectrum rather than a real spectrum. Phase correction is employed to extract the spectrum $S(\nu)$ from the complex output of the FT. The procedure known as 'multiplicative phase correction' or the 'Mertz method' is able to extract the spectrum without amplification of noise [8].

Fringes are the appearance of sinusoidal modulations on the baseline of infrared spectra. Fringing results from multiple reflections of the IR beam between two parallel surfaces in the spectrometer's optical path [8]. The effects of fringing can be avoided by using a lower resolution for data acquisition or by removal or adaptation of the offending element.

2.2.3 SOURCES AND DETECTORS

The infrared source is usually an inert solid which, when heated electrically to temperatures between 1500 and 2400 K, is able to emit continuous radiation approximating that of a black body. Sources most often used are globars (SiC), Nichrome wires, or ceramic Nernst glowers (mixtures of zirconium, yttrium and erbium oxides).

Infrared detectors can be classified as thermal or semiconductive. Thermal detectors measure the temperature change caused by the infrared radiation and include thermocouples, bolometers and pyroelectric detectors. Semiconductor detectors include photoconductive and photovoltaic devices that have faster response times than thermal detectors. In photoconductive detectors such as HgCdTe (mercury, cadmium, and telluride (MCT)), the infrared photon promotes an electron across the band gap between the valence and conductivity band which is measured as a change in current across the detector. In photovoltaic detectors such as InSb an electric current is produced that induces a charge that is directly proportional to the light intensity [7]. Semiconductor detectors are usually cooled with liquid nitrogen to reduce noise arising from thermal sources.

2.2.4 NOISE

Noise is an effect that is visible as fluctuations of the baseline in a signal, Figure 2.10. Random noise usually cancels out of a spectrum after many iterations, and the signal-to-noise ratio (SNR) improves as a function of the square root of the number of scans, *n* [5]:

$$\frac{S}{N} = f\sqrt{n}$$

Equation 2.17

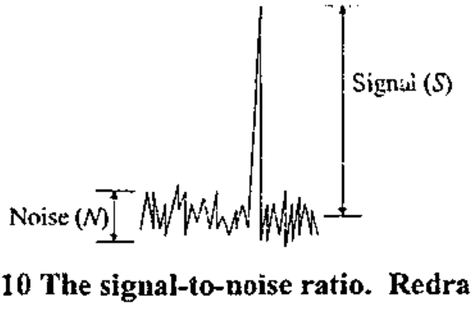

Figure 2.10 The signal-to-noise ratio. Redrawn from [5].

The detection limit for an instrument is the point at which a signal cannot be discriminated from the noise level and is frequently considered to be at a signal to noise ratio of 2 [6] or 2-3 [2]. The SNR is a measure of the signal in a peak measured relative to the noise (background signal) and is useful for describing the quality of the instrument or the instrumental method.

For a dc signal, noise takes the form of a time variation of the signal about the mean. Noise can therefore be defined as the standard deviation of the signal where signal is given by the mean [2]. The SNR is statistically defined as the ratio of signal strength to root-mean-square (RMS) noise [11]:

$$\frac{S}{N} = \frac{\text{mean}}{\text{standard deviation}} = \frac{1}{\text{relative standard deviation}}$$ **Equation 2.18**

2.2.4.1 SOURCES OF INSTRUMENTAL NOISE

Noise is associated with each component of an instrument and is a complex composite arising from several sources. Instrumental noise can be divided into four general categories. Johnson or thermal noise, shot noise, environmental or interference noise and flicker or 1/*f* noise. Johnson or thermal noise arises due to thermal agitation of electrons or other charge carriers in resistive elements of an instrument including resistors, capacitors and radiation detectors. Johnson noise is dependent on the frequency bandwidth but independent of frequency and is often referred to as white noise. Narrowing the bandwidth with filters and lowering the temperature of the detector, usually with liquid nitrogen, can reduce Johnson noise. Shot noise occurs wherever a current requires the movement of electrons or other charged particles across a junction. Flicker (1/*f*) noise has a magnitude that is inversely proportional to the frequency *f* of the signal and its presence is not well understood. Environmental or interference noise occurs as conductors in the instrument extract electromagnetic radiation from the surroundings and convert it to a signal. Temperature fluctuations over time also contribute to noise [2].

2.2.4.2 SIGNAL-TO-NOISE ENHANCEMENT

Several techniques can be employed to enhance the SNR. Hardware methods improve the SNR through incorporating different components into the instrument design; whilst software methods extract signals from noisy environments. Shielding or grounding the circuits can reduce noise arising from environmental electromagnetic radiation.

Difference amplifiers can be used to attenuate noise generated in transducer circuits. Analog filtering uses low-pass filters to remove any high frequency component arising from Johnson or shot noise. High-pass filters reduce the effect of drift and other low frequency flicker noise. The process of modulation reduces the effects of flicker noise at low frequencies by converting the dc signal from the transducer to a higher frequency. The chopper amplifier employs an electrical or mechanical chopper to convert the input signal to a square-wave form. This process removes only noise occurring after chopping the signal so it is preferable to chop the signal close to the source.

Infrared spectroscopy, i.e. dispersive instruments, use mechanical choppers to reduce noise, whereas in FT instruments the signal is modulated by the interferometer. Noise can be a problem in infrared spectroscopy because the source intensity and the detector sensitivity are low. This results in an electrical signal that is generally small and requires amplification. Thermal radiation and environmental noise may also present problems. Boxcar averaging is a digital procedure for smoothing irregularities, assumed to be arising from noise, in the waveform. Digital filtering is a method of smoothing that assumes a linear or polynomial relationship exists between the points being sampled in the boxcar procedure. Digital filtering can also be used to convert the original signal that varies as a function of time (time-domain signal) to a frequency-domain signal in which frequency is the independent variable. This is accomplished by a Fourier-transform procedure [2]. As seen in Equation 2.17 noise can also be reduced by simple signal averaging.

2.3 FTIR MICROSCOPE.

The FTIR microscope combines an optical microscope and an FTIR spectrophotometer. The FTIR microscope, Figure 2.11, utilises cassegrain optics for visible light and the infrared beam; therefore the infrared spectrum obtained corresponds to the visually selected area. The optical microscope enables a 120 times magnification of the visible light image of a sample.

A microscope enables the collection of spectra from several areas of one sample. This allows spectral reproducibility to be verified on samples of known heterogeneity, and on that basis allows sample differentiation in samples of unknown heterogeneity. The microscope aperture is important for isolating groups of cells and potential contaminants for further investigation. Infrared spectra can be collected from samples as small as 20μm^2.

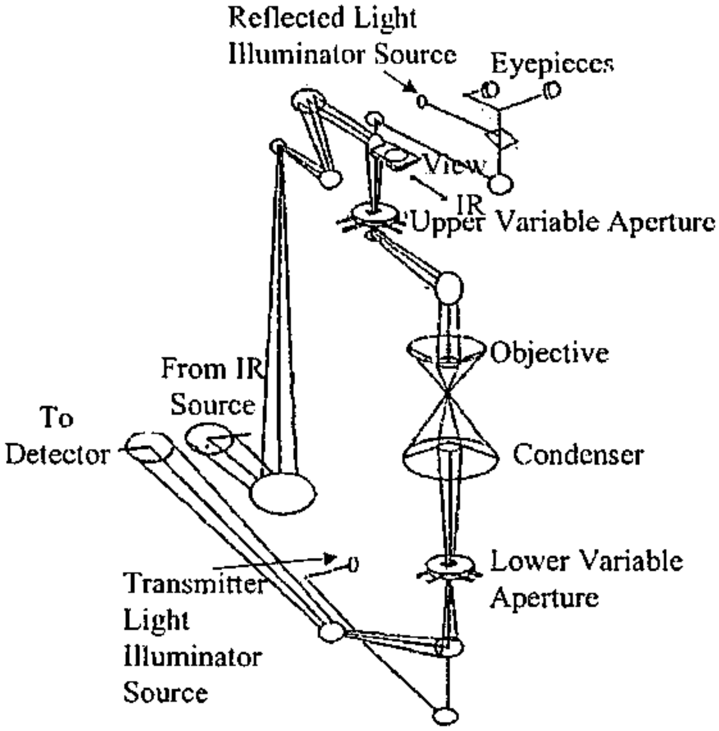

Figure 2.11 The general layout of an FTIR microscope [12].

2.4 *INSTRUMENTAL PARAMETERS*

A Bruker IFS-55 Spectrophotometer (Figure 2.12) coupled to an A590 infrared microscope was used to collect the majority of the spectral data. A Perkin Elmer (PE) 1600 spectrophotometer coupled to a PE infrared microscope was used to collect the infrared spectra for the menstrual study and a small proportion of the spectra used for the database.

Fifty scans were accumulated for each spectrum at a resolution of 8 cm^{-1}. The Bruker IR Microscope utilises a globar (MIR) source, a Ge multilayer coating on a KBr beam splitter, a MCT detector and an aperture setting of 5.0 mm. The MCT detector has an operating temperature of 77 K, is enclosed in a liquid nitrogen filled dewar and gives spectra with far less noise than those obtained with detectors that operate at room temperature. A low pass filter operating at 16 Hz and 12638 cm^{-1} was employed to remove any high frequency components arising from Johnson or shot noise. The acquisition mode was single sided, fast return. The frequency range of data collection was 3650-700 cm^{-1} with a phase resolution of 32. Mertz phase correction was employed as was a 3 term Blackman-Harris apodisation function and a zero filling factor of 2. To reduce the effects of environmental water vapour and carbon dioxide, the microscope was enclosed in a Perspex box purged with nitrogen.

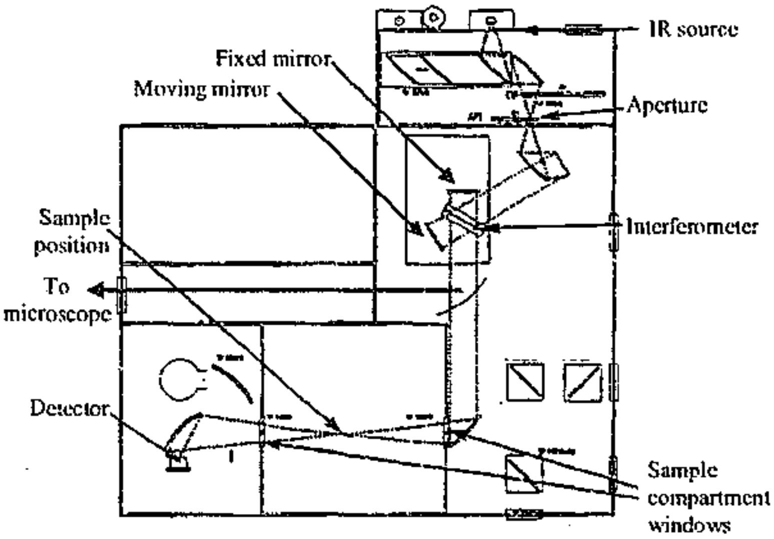

Figure 2.12 Schematic of the Bruker IFS 55 Spectrometer. Courtesy of Bruker (Australia).

Detectors such as MCT photoconductive detectors generate a nonlinear response with respect to concentration and intensity. A nonlinear response results in non-zero intensities appearing in spectral regions where zero intensity is expected [13]. Refer to Section 3.2.2 for an explanation of removal of nonlinearity effects.

In order to reduce the measurement time, resolution should not be higher than necessary as doubling the resolution results in a 4-fold increase in measurement time to maintain the same signal-to-noise ratio. Resolution is related to the optical path difference of the wave train and therefore the mirror displacement by the following relationship:

$$\text{Resolution} = \frac{1}{\text{Optical path difference}} = \frac{1}{2} \times \text{Mirror displacement} \qquad \textbf{Equation 2.19}$$

To achieve a particular resolution ($\delta\bar{\nu}$) at a certain wavenumber ($\bar{\nu}$), the aperture diameter (A) must be a small value as defined by:

$$\text{A(mm)} < 2 * \text{F(mm)} * \left(\frac{\delta\bar{\nu}}{\bar{\nu}}\right)^{\frac{1}{2}} \qquad \textbf{Equation 2.20}$$

Where F = instrument focal length. The focal length of the IFS-55 is 69 mm [13] and the aperture size used for collection of spectra was 0.6 mm.

2.5 THEORY OF MULTIVARIATE STATISTICS

The term multivariate data analysis incorporates many different methods and techniques and can be divided into two different types: clustering methods and ordinal methods. Clustering methods are algorithmic approaches that aim to divide or collect samples into groups, depending on similarities. Ordinal methods, such as principal component analysis (PCA), use mathematical properties to decompose data matrices. These methods work by obtaining new coordinates describing the variance, or covariance, in the data set. This enables complex data consisting of many variables to be reduced to a lower dimensionality [14].

2.5.1 PATTERN RECOGNITION

Pattern recognition refers to the ability to assign an object to one of several possible categories according to the values of measured parameters or variables. In chemometrics, pattern recognition can be further divided into two groups: unsupervised and supervised. Unsupervised pattern recognition includes cluster analysis and hierarchical techniques and the interpretation of the number of clusters and populations can often be subjective, as this information is not known prior to analysis. In supervised pattern recognition techniques such as classification or discriminant analysis the number of groups is already known and there are representative samples of each group. Classification uses information from known samples to identify and categorise future samples. The means of deriving the classification rules from previously classified samples is referred to as discrimination [15].

If the data observed for one object is measured by M variables, it can be represented as a vector and as a point in M-dimensional space by giving each variable one coordinate axis. A class of objects is represented by a swarm of points in M-space and several classes as distinct or overlapping swarms. Pattern recognition can be seen as the methodology to describe these swarms quantitatively and enables the calculation of which class a new object is to be assigned to [16].

2.5.2 CLASSIFICATION

The concept of distance is very important in classification procedures and follows from the assumption that proximity in multivariate space is indicative of similarity between samples. Therefore samples that are near in variable space are considered to have the same characteristics, whereas a large separation is suggestive of different characteristics.

The most commonly used method for determining similarity between samples is to measure the Euclidean distance of samples in variable space [17]:

$$d_{1,2} = \left[\sum_{j=1}^{M} (x_{1j} - x_{2j})^2 \right]^{\frac{1}{2}} \qquad \textbf{Equation 2.21}$$

where M is the number of variables.

Classification on the lowest level operates under the assumption that all objects in the training and test sets belong to one of the initially defined classes. If the reference or training sets of the classes can be separated from each other by a surface, new objects are classified according to which side of this surface they fall. Two examples of this type of classification are K-nearest neighbours (Section 2.5.3) and linear discriminant analysis (Section 2.5.4). It is often unrealistic to assume that all objects belong to one of the defined classes and these methods of classification do not allow for the presence and detection of outliers or the possibility that an object might belong to an unknown class [16].

The next level of classification operates by containing each class in a closed envelope (mathematical structure) in M-space. These class envelopes are constructed so that an object falling within an envelope is considered to be a member of that class and objects falling outside all envelopes are considered to be outliers to all classes. An example of this level of classification is called SIMCA (soft independent modelling of class analogy, Section 2.5.6) [16].

2.5.3 K-NEAREST NEIGHBOURS (K-NN)

K-nearest neighbours is referred to as a non-parametric technique and searches primarily for similarity within classes, making no assumptions about the distribution of the data. Unknown samples are classified by measuring the Euclidean or Mahalanobis distance from the unknown to members in the training set. An object is assigned to the class to which the majority of nearest neighbours belong.

For two objects characterised by multivariate pattern vectors x_1 and x_2 defined by:

$$x_1 = (x_{11}, x_{12}, \ldots, x_{1M})$$

$$x_2 = (x_{21}, x_{22}, \ldots, x_{2M}) \qquad \textbf{Equation 2.22}$$

where M is the number of variables, the Euclidean distance between objects 1 and 2 is given by Equation 2.2.1. This equation can be expressed in vector notation as:

$$d_{AB} = [(a-b)^T \cdot (a-b)]^{1/2}$$ **Equation 2.23**

where **a** and **b** represent x_1 and x_2 respectively.

The Mahalanobis distance is a weighted distance measure and is defined as:

$$d_{AB} = [(a-b)^T \cdot Cov^{-1}(a-b)]^{1/2}$$ **Equation 2.24**

where ***Cov*** if the variance-covariance matrix for the original data.

Application of these distance equations to K-NN defines a circle or sphere about the unclassified sample in point space containing K nearest neighbours, with radius r_K, which is the distance to the K^{th} nearest neighbour. This is shown schematically in Figure 2.13.

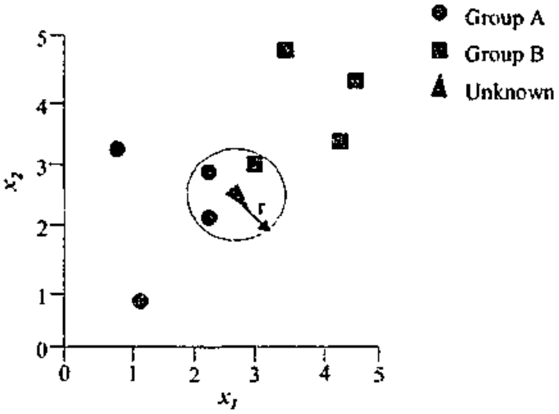

Figure 2.13 Schematic representation of the circle with radius *r* about an unclassified object containing three nearest neighbours. The unknown sample is assigned to group A. Redrawn from [15].

The volume of the sphere is used as an estimate of $P_{(x|Gi)}$ or the conditional probability of the pattern vector x arising from group i:

$$P_{(x|G_i)} = \frac{\sum k_i}{n_i} \cdot \frac{1}{V_{K,x}}$$ **Equation 2.25**

where n_i is the number of samples known to belong to each group i, k_i is the number of nearest neighbours in group i and $V_{K,x}$ is the volume of space which contains the K nearest neighbours.

Using Equation 2.25 in Bayes' rule (Section 2.5.4) an object is assigned to group i if:

$$\frac{P_{(G_i)} \cdot k_i}{n_i} > \frac{P_{(G_j)} \cdot k_j}{n_j}, \text{ for all } j \neq i$$ **Equation 2.26**

If the number of objects in each training set, n_i, is proportional to the unconditional probability of occurrence of the groups, $P_{(Gi)}$, the equation is further simplified to, assign to group i if:

$$ki > k_j$$ **Equation 2.27**

The choice of distance measurement used, be it Euclidean or Mahalanobis depends on correlation between samples and variables and the ratio of variables to sample objects. When variables are correlated, the Euclidean distance leads to distorted conclusions. If the number of variables is lower than the number of objects, the Mahalanobis distance may be applied because this distance takes account of correlation existing within objects in the class. When the number of variables is comparable or higher than the number of objects, and when the variables are correlated, the Mahalanobis distance cannot be calculated because the covariance matrix is singular [18].

The choice of K is arbitrary but for data in which the classes overlap, $K = 3$ or 5 have been shown to provide good classification [15]. *K-NN* is a useful statistical technique in that it drastically reduces the amount of storage and computation requirements. However the classification of a new object requires recalculation of all distances, and if another class is added, *K-NN* criteria must be recomputed. *K-NN*s is sensitive to unequal numbers of objects in the training sets and the classification of an unknown object can differ depending on K. *K-NN*s is superior to some other pattern recognition techniques in that each sample is unambiguously assigned to a single group.

2.5.4 DISCRIMINANT ANALYSIS

The central idea behind discriminant analysis, like most classification techniques, is to assign an observation, x, of unknown origin to a distinct group on the basis of the value of the observation.

Discriminant analysis is based on Bayes' theorem, which states "*a sample or object should be assigned to that group having the highest conditional probability*" [15]. The application of this rule to parametric classification provides discriminating ability. An unknown sample is assigned to, for example, Group A or G(A) on the condition that:

$$P_{(G(A)|x)} > P_{(G(B)|x)}$$ **Equation 2.28**

Determining these conditional probability values involves the analysis of all samples in the parent population, which is quite obviously unrealistic. Bayes' theorem provides an indirect means of estimating the conditional probability, $P_{(G(A)|x)}$. According to Bayes' theorem:

$$P_{(G(A)|x)} = \frac{P_{(x|G(A))} \cdot P_{(G(A))}}{P_{(x|G(A))} \cdot P_{(G(A))} + P_{(x|G(B))} \cdot P_{(G(B))}}$$ **Equation 2.29**

$P_{(G(A))}$ and $P_{(G(B))}$ are the *a priori* probabilities and represent the probabilities of a sample belonging to A and B in the absence of having data. $P_{(x|G(A))}$ is a conditional probability that expresses the chance of a vector pattern x arising from Group A. This probability can be estimated by sampling the population of Group A. Whilst $P_{(x|G(A))}$ and $P_{(x|G(B))}$ can be estimated through the analysis of large numbers of samples, if the variables contributing to the vector pattern are assumed to follow a normal distribution, the conditional probability values can be calculated from:

$$P_{(x|G(A))} = \frac{1}{2\pi|\mathbf{Cov}|^{1/2}} \exp[-1/2(x-\mu_A)^{\mathrm{T}} \cdot \mathbf{Cov}_A^{-1} \cdot (x-\mu_A)]$$ **Equation 2.30**

In quadratic discriminant analysis, a sample is assigned to Group A if:

$$\ln P_{(G(A))} - 0.5\ln(|Cov_A|) - 0.5(x - \mu_A)^{\mathrm{T}}\, Cov_A^{-1}\, (x - \mu_A) >$$

$$\ln P_{(G(B))} - 0.5\ln(|Cov_B|) - 0.5(x - \mu_B)^{\mathrm{T}}\, Cov_B^{-1}\, (x - \mu_B)$$ **Equation 2.31**

The discriminant function, $d_A(x)$ is defined by:

$$d_A(x) = 0.5\ln(|Cov_A|) + 0.5(x - \mu_A)^{\mathrm{T}}\, Cov_A^{-1}\, (x - \mu_A)$$ **Equation 2.32**

and an object is assigned to Group A if:

$$d_A(x) < d_B(x)$$ **Equation 2.33**

If the prior probabilities can be assumed to be equal, i.e. $P_{(G(A))} = P_{(G(B))}$ the discriminant line between Groups A and B is given by:

$$d_A(x) = d_B(x)$$ **Equation 2.34**

A further simplification to Bayes' classifier can be made if the covariance matrices for both groups are assumed to be equal. Equal covariance implies that the correlations between variables are independent of the group to which the objects belong [15]. In these cases the groups are linearly separable and linear discriminant analysis is performed. With the assumption of equal covariance matrices, an object is assigned to Group A if:

$$\ln P_{(G(A))} - 0.5(x - \mu_A)^T(\boldsymbol{Cov}^{-1})(x - \mu_A) >$$

$$\ln P_{(G(B))} - 0.5(x - \mu_B)^T(\boldsymbol{Cov}^{-1})(x - \mu_B) \qquad \textbf{Equation 2.35}$$

where $\boldsymbol{Cov} = \boldsymbol{Cov}_A = \boldsymbol{Cov}_B$

This further simplifies to:

$$(\mu_A{}^T\, \boldsymbol{Cov}^{-1}\boldsymbol{x}) - 0.5(\mu_A{}^T\, \boldsymbol{Cov}^{-1}\mu_A) > (\mu_B{}^T\, \boldsymbol{Cov}^{-1}\boldsymbol{x}) - 0.5(\mu_B{}^T\, \boldsymbol{Cov}^{-1}\mu_B) \qquad \textbf{Equation 2.36}$$

or

$$f_A(x) > f_B(x) \qquad \textbf{Equation 2.37}$$

As with the quadratic discriminant analysis (Equation 2.34), the discriminant line between Groups A and B is given by:

$$f_A(x) = f_B(x) \qquad \textbf{Equation 2.38}$$

Like K-nearest neighbours, discriminant analysis performs poorly when the ratio of objects to variables is small [19]. To overcome this problem principal component analysis can be utilised as a means of variable reduction and the resulting principal components can be used instead of the original variables.

2.5.5 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis is one of the most widely used multivariate statistical techniques for the extraction and interpretation of information from multivariate data [20]. The aim of PCA is to reduce a large number of variables down to a small number of summary variables, or principal components (PCs), that explain most of the variance in the data. All PCs are orthogonal and each successive component expresses decreasing amounts of variation with most of the variation explained by the first few components. This enables the multi-dimensional data to be represented in two or three dimensions, which are easily visualised. The technique works by transforming the original variables onto a new set of axes in the direction of the greatest variation in the data, Figure 2.14.

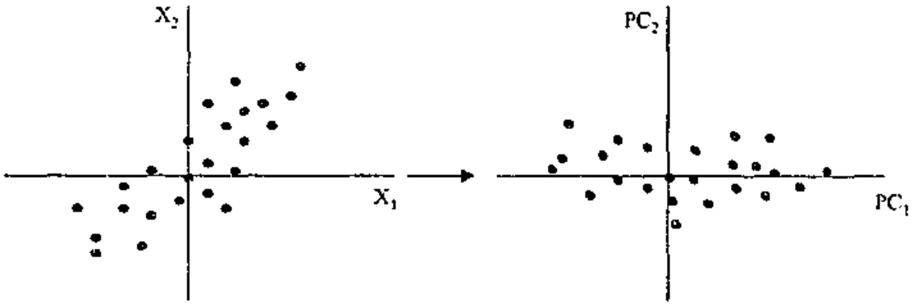

Figure 2.14 Plot of observations on two variables X_1, X_2 (left) and the same observations plotted with respect to their principal components PC_1, PC_2 (right). Adapted from [20].

The first component is oriented along the axis of greatest variance of the variables in the data matrix about their means. The second PC is independent of (orthogonal to) the first PC and is the vector along the axis of next largest variance in the data. Succeeding PCs can be calculated which will be orthogonal to the preceding ones and which explain some of the remaining variance. The PCs are linear combinations of the original variables, which are fitted in the least squares sense through the points in measurement space. These new variables usually result in a reduction of variables from the original set and often can be correlated with physical or chemical factors [21].

If x is a vector of p variables then algebraically the first PC is a linear combination of $x_1, x_2, \ldots, x_p$:

$$\mathbf{PC_1} = a_{11}x_1 + a_{12}x_2 + \ldots\ldots + a_{1p}x_p = \sum_{i=1}^{p} a_{1i}x_i \qquad \textbf{Equation 2.39}$$

The variables, x_i can be either deviation from mean scores or standardised scores, and the variance of PC_1 is maximised given the normalisation constraint that the sum of the squared weights is equal to one ($\sum_{i=1}^{p} a_{1i}^2 = 1$).

The second PC involves finding a second weight vector ($a_{21}, a_{22}, \ldots\ldots a_{2p}$) such that the variance of:

$$\mathbf{PC_2} = a_{21}x_1 + a_{22}x_2 + \ldots\ldots + a_{2p}x_p = \sum_{i=1}^{p} a_{2i}x_i \qquad \textbf{Equation 2.40}$$

is maximised subject to the normalisation constraint, $\sum_{i=1}^{p} a_{2i}^{2} = 1$ and the constraint that it is uncorrelated with the first PC, $\sum_{i=1}^{p} a_{1i}a_{2i} = 0$. The independence condition is specified by the constraint that the second PC has the next largest sum of squared correlations with the original variables. The sum of squared correlations with the original variables, or the variance of the PCs get smaller with the extraction of successive PCs. The sum of the variance of the PCs is equal to the sum of the variance of the original variables:

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_i^2 \qquad \textbf{Equation 2.41}$$

where λ_i is the variance of the i^{th} PC [22].

The main statistics resulting from PCA are the variable weight vectors or latent vectors (eigenvectors), $a = a_1, a_2, \ldots a_p$, from each PC and its associated variance or latent root (eigenvalue), λ. The pattern of variable weights for a particular PC are used to interpret that component and the magnitude of the variance of the PCs provide an indication of how well they account for the variability in the data [22].

Principal components are easier to interpret when the elements of the latent vector are transformed to correlations of the variables with the particular PCs. These correlations are called loadings. Loadings are calculated by multiplying each of the elements of a particular latent vector, a_i by the square root of the associated latent root, $\sqrt{\lambda_i}$. Thus the correlations of the variables with the i^{th} PC is $\sqrt{\lambda_i} a_i$. Variables that correlate highly with a particular PC give meaning to that component. The relative magnitudes of the elements in the eigenvector or loadings for a particular PC indicate the relative contribution of the corresponding variable to the variance of that PC. The first PC usually has large correlations with all the variables and is essentially a weighted average of the standardised variable scores. The PC scores for any pair of PCs can be plotted. The reasons for doing this include checking for outlying observations, searching for clusters and, in general, understanding the structure of the data [22].

2.5.6 SOFT INDEPENDENT MODELLING OF CLASS ANALOGY (SIMCA)

A major advantage of soft modelling techniques, such as SIMCA, is that objects are not forced into discrete classes. This is useful for the detection of outliers, which are samples that belong to none of the predefined groups [15].

SIMCA allows the development of a separate mathematical description for each class independently. A new object is classified according to its position in the pattern space with relation to the class boxes. PCA is used to form models for individual classes.

Each sample or object is represented by a $1 \times M$ row vector containing the M measurements on the sample. The row vectors (x_i) for N samples together form the $N \times M$ data matrix X. By the use of the constraints outlined in Section 2.5.5 the matrix X can be decomposed into principal components. A PC model can be written by collecting the results from PCA, eigenvalues, a scores vector t_a, and a loading vector p_a (orthogonal to t_a) containing N and M elements respectively, into matrices [17]:

$$X = \mathbf{1}\bar{x}'(g) + \mathbf{TP}' + \mathbf{E}(g) \qquad \textbf{Equation 2.42}$$

where **1** is of dimension $N \times 1$, g represents the class and the vector $x(g)$, which defines the centroid for each variable, is of dimension $M \times 1$. The matrix $E(g)$ contains the residuals, which is the difference between the data and the model for class g [17]. The group centroids define the models so that the group centroid x_g plus a residual distance e_i describe each sample. The optimal number of components, A, to use in each model is determined by double cross-validation[7].

An A-dimensional hyperplane is fitted to each class, which can be visualised (Figure 2.15) as the construction of a class box for each class. A new object is classified according to its position in the pattern space with relation to the class boxes.

The classification rule for assigning a new object to classes is based on the distances, s_i, from the object to the class [24]. The object-class distances are calculated as squared residuals

$$s_i^2 = \frac{e_i'(g)e_i(g)}{M - A} \qquad \textbf{Equation 2.43}$$

---

[7] A validation set is used to check how well a model will perform on future samples taken from the same population as the calibration or training samples. Cross validation is a validation method where some samples are kept out of the calibration and used for prediction. This process is repeated until all the samples have been kept out once [23].

where s_i is termed the residual standard deviation (RSD) of the sample i.

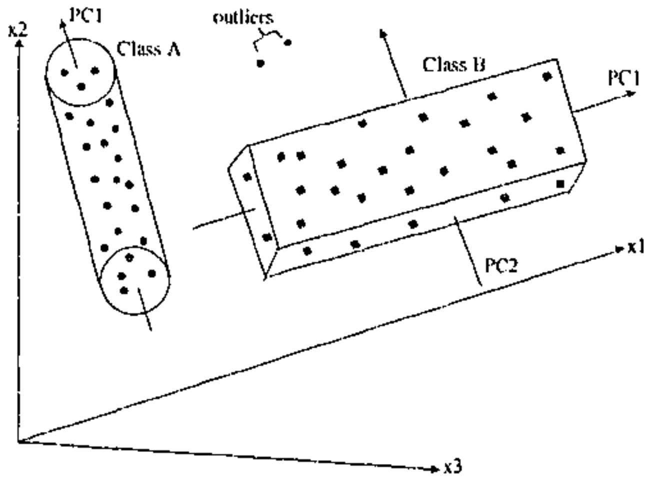

Figure 2.15 Schematic representation of the construction of a class box around two classes. Redrawn from [24].

If the RSD for the samples are collected into a $N\times1$ vector, s, the mean residual standard deviation for the class can be determined:

$$s_g^{\ 2} = \frac{s's}{N-A-1}$$ **Equation 2.44**

Comparison of sample i RSD with the mean RSD of the class using F-tests gives RSD limits for inclusion in the class, which effectively determines a class boundary around the principal components between the class and variable space.

$$s_{\max}^2 = s_g^{\ 2} F_{crit}$$ **Equation 2.45**

The difference in class boundaries between using a probability level p = 0.05 or 0.01 to determine F_{crit} is illustrated in a one-component model in Figure 2.16.

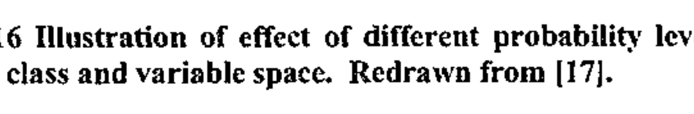

Figure 2.16 Illustration of effect of different probability levels for deciding the boundary between a class and variable space. Redrawn from [17].

To close the class model along each principal component the extreme scores $t_{min,a}$ and $t_{max,a}$ and their spread $s_{t,a}$ are used to define lower and upper limits for the scores [25]. This is illustrated for a one component system in Figure 2.17.

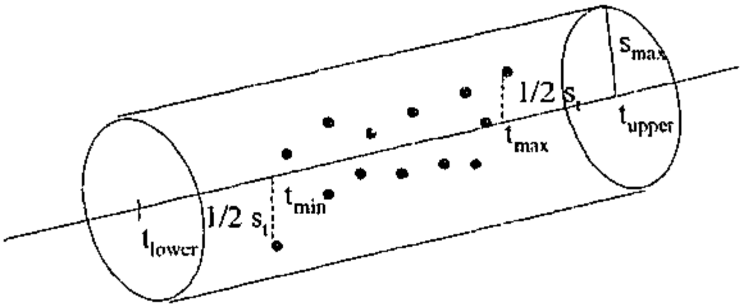

Figure 2.17 Using the sample scores along each principal component upper and lower limits are defined which enables the model to be closed in variable space. Redrawn from [17].

The upper and lower limits are defined as:

$$t_{lower,a} = t_{min,a} - \tfrac{1}{2}\, s_{t,a}$$ **Equation 2.46**

$$t_{upper,a} = t_{max,a} + \tfrac{1}{2}\, s_{t,a}$$ **Equation 2.47**

Class envelopes, as well as allowing for the detection of outliers, provide information of the relevance of each variable and measures of inter-classes distances. The relevance of a variable can be measured by modelling power and discriminating power. The modelling power of a variable is related to its contribution to the description of the classes and is related to the within-class variation of a variable compared with the total variation of that variable over the whole training set. In terms of class envelope this is seen as the average thickness of the envelopes along the variable coordinate axis compared with the total range of the variable. The discriminating power of a variable relates to the contribution of that variable in discrimination between the classes. This is measured as the average distance between class envelopes along the coordinate axis of the variable compared with the average envelope thickness along the same axis. Inter-class distance is measured as the distance between two envelopes relative to their average thickness [16].

Two further advantages of the SIMCA approach compared to purely distance-based cluster approaches are firstly; modelling into principal components separates structure from noise. Amongst others this provides a basis for the rejection of irrelevant variables and outlying samples [17]. Secondly, the use of PCA to build separate models for classes enables the introduction of new classes without the need for recomputation of existing classes.

2.5.7 ARTIFICIAL NEURAL NETWORKS (ANNS)

Artificial neural networks are mathematical models and algorithms, which have been designed to mimic the information processing and knowledge acquisition methods of the human brain. Neural networks have significant advantages over standard computer methods, especially for pattern recognition applications [26]. Neural networks do not use rules, rather they 'learn' from 'training sets' in a similar way to humans. ANNs have the ability to learn directly from data, and can process data that only broadly resembles that on which they were trained [27]. After learning the patterns of inputs and outputs they are able to classify patterns and make predictions based upon new patterns of inputs [28]. Although ANNs are programs and therefore software, the number of inputs to a network is limited by available hardware and processing time required. Problems handled by artificial neural networks can generally be divided into four groups [26]:

1. Association (auto or hetero): in auto association, the system is able to reconstruct correct patterns if the pattern learned is incomplete or corrupted. In hetero association the system makes a one-to-one association between members of two sets of patterns.
2. Classification: the goal of classification is to assign all given objects to appropriate classes of objects based on one or more properties that categorise a given class.
3. Transformation: involves the transformation or mapping of a multivariate space into another space of the same or lower dimensionality.
4. Modelling: is the search for an analytical function or model that will give a specified n-variable output for any m-variable input. Whereas standard modelling techniques require the mathematical function to be known in advance, the nonlinearity and large number of variable parameters (weights) enables the neural network to adapt to any relation between input and output data without prior knowledge of the mathematical function.

2.5.7.1 NEURODES

The basic unit in an ANN is the neuron or neurode, which has many input paths each modified by a weight. The generation of an output from a neurode for a given input involves two steps. The first step is the evaluation of the net input *Net*, the second step involves a nonlinear transformation of *Net*.

The net input or decision function of a neurode is a function of the weights w_i and all the signals s_i that arrive at a neurode:

$$Net = w_1 s_1 + w_2 s_2 + ... + w_m s_m$$ **Equation 2.48**

Equation 4.28 employs a linear transformation on a multivariate signal X using the weight vector W to obtain a one-variable signal, where X is a multidimensional vector whose components are the individual signals. This procedure has been referred to as a linear learning machine because the input vector is linearly proportional to the corrected result, *Net* [26].

The most commonly used transfer function for the nonlinear transformation of *Net* into an output, is the sigmoidal function $f(x)$:

$$f(x) = \frac{1}{1 + e^{-x}}$$ **Equation 2.49**

The generation of an output is shown schematically in Figure 2.18.

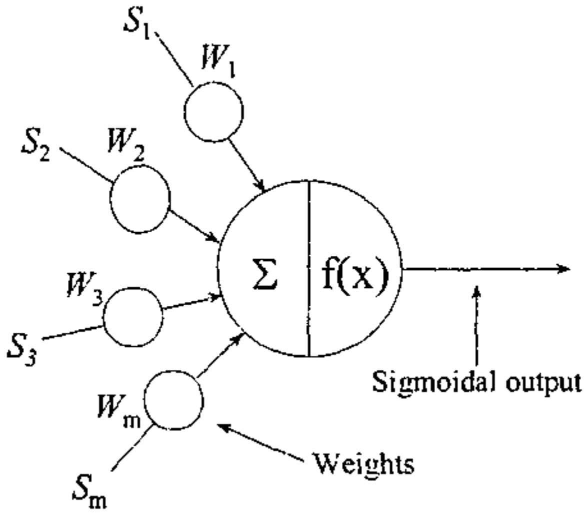

Figure 2.18 Schematic representation of the generation of an output from a neurode.

The nonlinearity of the transfer function enables the network to be flexible in adjusting to different learning situations. The derivative of the sigmoidal transfer function is important to the way in which the neural network is able to learn and is used to determine the gradient for finding the surface minimum:

$$\frac{df(x)}{dx} = f(x)[1 - f(x)]$$ **Equation 2.50**

The initial weights associated with individual neurodes in the network are usually randomly chosen, and then improved iteratively. An increment for W is calculated as the difference between a new, corrected weight vector $W^{(new)}$ and the old, uncorrected vector $W^{(old)}$:

$$\Delta W = W^{(new)} - W^{(old)}$$ **Equation 2.51**

This correction is achieved by employing the delta rule. In order to improve the decision vector, W, correction should be proportional to a certain parameter δ, (which is proportional to the error) and to the input X for which the wrong answer was obtained. After correction, the new weight vector should classify the vector X if not correctly, then with a smaller error than before:

$$\Delta W = \eta \delta X$$ **Equation 2.52**

where δ is the correction constant and η is a constant of proportionality, or learning rate.

In order to ensure that large changes in the decision vector do not lead to previously correctly classified objects becoming falsely classified, η is usually kept less than 1. If the learning rate is too high, the network tends to oscillate and not learn the correct mapping from the inputs to targets. If the weight adjustments are too small, learning will take a long time [26].

If the decision function (Net) is taken to be the dot product between the representation of the object X and the weight vector W:

$$Net = \mathrm{W} \cdot \mathrm{X} + \vartheta = \sum_{i=1}^{m} w_i x_i + \vartheta$$ **Equation 2.53**

The offset parameter, ϑ, is called a bias and it increases the adaptability of the decision function to the problem it is designed to solve. The bias acts as an extra weight and always receives an input of 1. The addition of a bias moves the problem from the two-dimensional space, where the solution would be a line, to a three-dimensional space, where the solution is a plane [26].

An artificial neural network consists of many neurodes organised into groups called layers or slabs [28]. Each neurode in the layer has the same number of m weights and receives the same m-dimensional input signal simultaneously. The neurodes of two layers can be fully, partially or randomly connected. Full connection means that each neurode in one layer is connected to all the neurodes in the next layer.

The topographical data of the network including number of inputs and outputs, number of layers, number of neurodes in each layer, number of weights associated with each neurode and the interconnections of the layer(s) form what is known as the architecture of the network (Figure 2.19). All neurodes in one layer receive the same number of inputs, including the input connected to the bias. The number of signals arriving from the previous layer determines the number of weights in each neurode.

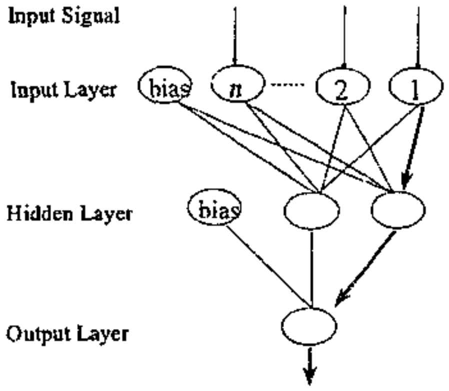

Figure 2.19 Schematic representation of a fully connected feed-forward[8] neural network, with neurodes represented by circles and connections between the neurodes represented by lines.

The input neurodes do not modify the signal, as there are no weights or transfer functions associated with them. These neurodes act to distribute the input and as such are referred to as a non-active layer [26].

The layer(s) below the input layer are referred to as the hidden layer(s) because they are not connected to the outside world. The layer of neurodes that yields the final signal is called the output layer.

The number of hidden layers and neurodes, and thus the number of weights in the neural network, is governed by the number of data points in the training set. Ideally, there should be at least twice as many samples as there are weights [29]. For one hidden layer, the number of weights (accounting for the bias node in the input and hidden layer) is equal to:

$$[(i+1)\times n]+[(n+1)\times j] \qquad \textbf{Equation 2.54}$$

---

[8] Feed-forward is the term used to describe neural networks in which the data passes through the network in one direction (from the input layer through any hidden layer/s to the output layer).

where:

i is the number of nodes in the input layer (1,2,3.........).

j is the number of nodes in the output layer (1,2,3........).

n is the number of nodes in the hidden layer (1,2,3.......).

2.5.7.2 *BACK PROPAGATION*

The most widely used learning or training method for an artificial neural network is called back-propagation of errors. Back propagation (Figure 2.20) employs a modification of the delta rule, applying the equations for the correction of weights throughout the layers of the network, starting with the weights in the output layer and continuing back towards the input layer [27]. Training requires showing the network many data input sets thousands of times before the ANN adjusts its internal weights enough to give accurate output responses, using least-squares, to input data [28].

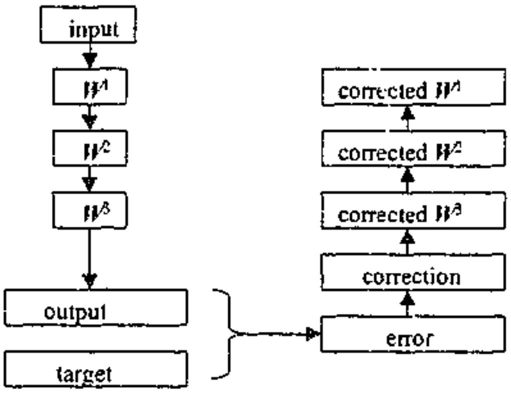

Figure 2.20 Schematic presentation of weight correction by back-propagation of errors. Redrawn from [26].

Back propagation is a supervised learning method and requires a set of pairs of inputs X_s and targets Y_s. One advantage of training a neural network by back propagation is that there is no need to know the exact form of the mathematical function on which the model is built. In a fully connected multi-layer network, where each input has an influence on all the weights, it is virtually impossible to test the influence of the weights on the final output because it would require determining the effects of each individual input on each weight. This becomes difficult with a complex architecture of many weights and sensitivity analysis must be performed to enable interpretation of internal parameters. The architecture of a back propagation neural network, i.e. the number of layers, the number of neurodes in each layer and the connection of the neurodes, is the main feature influencing the flexibility of the model [26].

Correction of weights can be made after each new input, termed immediate correction, or after all the inputs have been tested, deferred correction. With immediate correction, correction is made as soon as an error is detected, and the accumulated error of the entire training set is used for correction. Most applications of neural networks use immediate correction.

During learning, the object *X* (input vector) is presented to the neural network and the output vector *Out* is compared with the target vector *Y*, which is the correct output for *X*. Once the error produced by the network is known, weights are corrected throughout the neural network using a modification of the delta rule seen in Equation 2.52.

$$\Delta w^{l}_{ji} = \eta \delta^{l}_{j} Out^{l-1}_{i} + \mu \Delta w_{ji}^{l(\text{previous})}$$ **Equation 2.55**

where *l* is the index of the current layer, *j* identifies the current neuron, and *i* is the index of the input source, i.e. the index of the neuron in the upper layer.

The correction of weights in the *l*-th layer is composed of two terms, which pull in opposite directions: the first term tends towards a fast "steepest descent" convergence, whilst the second is a longer-range function that prevents the solution from being trapped in a local minima. The constant η is called the learning rate and μ is called the momentum constant or factor, which can be used to speed network training. The momentum factor adds a proportion of the previous weight changes to the current weight changes and appropriate selection can prevent the network from oscillating and speed learning [26].

In a back-propagation neural network, where the output is obtained directly from the neurons in the output layer, it is advisable to scale each component of the target to lie between 0 and 1. Due to the nonlinear character of the transfer function, it is better to scale the entire output to lie between 0.1 and 0.9. Scaling offers three advantages [26]:

1. Easier comparison of the output and target data.
2. Proper calculation of the RMS (root-mean-square) error.
3. Later recalculation of the correct answer from the output neuron.

2.5.7.3 NETWORK OPTIMISATION

Some inherent problems of feed-forward, back-propagation neural networks include overtraining, overfitting and network architecture optimisation.

The network architecture (the number of neurodes in the hidden layer, the number of inputs, the number of epochs (training cycles) and the choice of values for the learning rate and momentum factor) all contribute to the ability of the neural network to learn. Increasing the number of epochs will enable the network to be trained longer, but can sometimes result in a network that has lost the ability to generalise. This is known as overtraining and occurs when the network has "memorised" the input patterns and is unable to accurately predict outcomes from new objects. Overtraining can be prevented by reducing the number of epochs or by employing a validation set. As described earlier, training a network involves an iterative reduction of the error function defined with respect to a set of training data. The error generally decreases as a function of the number of iterations. When the error is measured with respect to independent data, as is the case with a validation set, there is often a decrease at first followed by an increase as the network starts to overtrain (Figure 2.21). The network architecture at the point of smallest error with respect to the validation data is expected to have the best generalisation ability [30].

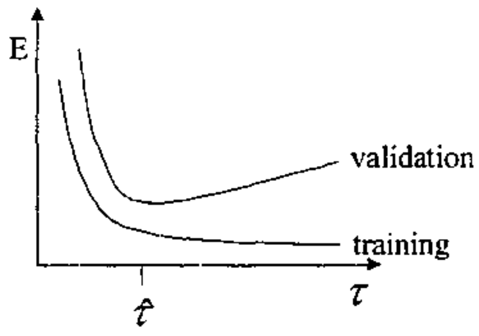

Figure 2.21 Schematic illustration of the training and validation errors as a function of the iteration step τ . To achieve a network with the best predictive performance training should be stopped at the point corresponding to the minimum validation error $\hat{\tau}$. Redrawn from [30].

Validation procedures such as the leave *N* out method produce a family of network models and it is not always clear which model gives the best predictive ability. It is also not obvious which architecture results in the best model further necessitating architecture optimisation [31].

Overfitting occurs when the number of weights is larger than the number of objects and is a consequence of parameter redundancy, i.e. the network has more parameters than are needed to find a solution to the problem. Overfitting can be avoided by employing PCA to reduce redundant information. The number of principal components that gives the lowest standard error of prediction should be chosen. Furthermore the use of a test set allows selection of models with best predictivity [31]. The test set should comprise data previously unseen by the network, i.e. data not included in either the training or validation sets.

Several methods can be employed to optimise network architecture and control the effective complexity of the network model. The complexity can be varied by changing the number of adaptive parameters in the network, known as structural stabilisation [30]. Currently there are no objective methods for architecture optimisation using this technique and the best network is usually found using trial-and-error or rule-of-thumb [32].

Another approach to controlling the complexity of a network model involves the addition of a penalty term to the error function, known as regularisation. The simplest form of regulariser is called weight decay and the addition of this term to the error function encourages small weights, which are less likely to result in overfitting. For a detailed explanation of regularisation and weight decay refer to Bishop [30].

2.5.7.4 BAYESIAN REGULARISED NEURAL NETWORKS

Bayesian regularised artificial neural networks (BRANNs) are a special type of neural network based on a Gaussian approximation[9] to the posterior weight distribution and offer several advantages compared with conventional back-propagation neural networks [31].

BRANNs were introduced in 1992 by MacKay [33, 34] and are essentially a mathematical formulation of Occam's Razor. Occam's Razor employs the principle of economy in explanations and states that if several theories account for a phenomenon, the simplest one, which describes the data sufficiently well, should be used [35].

The main benefits of BRANNs are [35]:

1. Weight decay parameters are adjusted automatically during training to near optimal levels resulting in the best generalisation.

---

[9] Refer to Appendix C for an explanation of the Gaussian distribution.

2. The Bayesian approach estimates the evidence (Equation 2.56) for each model, which is a measure of how probable the model is with respect to the data (assuming equal prior probability). The evidence is used as a quality measure to select the best model and enables comparison of models with different architectures as well providing an objective stopping criterion.

3. BRANNs do not require a separate validation set, therefore all available data can be used for training, which results in better models.

Whereas standard back-propagation neural network training methods use a single set of parameters, the Bayesian approach to neural network modelling considers all possible values of network parameters weighted by the probability of each set of weights. Bayesian inference is used to find the posterior probability of the weight parameters, w, and related properties using the prior probability distribution formed by the training set D using the BRANN model, H_i [31, 33].

$$P(\mathbf{w} \mid D, H_i) = \frac{P(D \mid \mathbf{w}, H_i) P(\mathbf{w} \mid H_i)}{P(D \mid H_i)} \qquad \textbf{Equation 2.56}$$

or in words

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

The Gaussian approximation is the main weakness of this approach and problems with the approximation are encountered when the number of weights exceeds one third of the number of objects in the training set [34].

For a detailed description of the theory and derivation of Gaussian approximations and Bayesian neural networks the reader is referred to the texts of MacKay [33, 34] and Thodberg [35].

2.6 REFERENCES

1. McMurry, J., Structure Determination, in *Fundamentals of Organic Chemistry*, Y. Howell, Editor. Wadsworth Inc.: Belmont, California. p. 372-379. 1990.
2. Skoog, D., *Principles of Instrumental Analysis*. 3 ed. Orlando: Saunders College Publishing. 1985.
3. Harris, D. and M. Bertolucci, *Symmetry and Spectroscopy An Introduction to Vibrational and Electronic Spectroscopy*. New York: Dover Publications, Inc. 1989.
4. Graybeal, J.D., *Molecular Spectroscopy*. New York: McGraw-Hill Book Company. 1998.
5. Pavia, D., G. Lampman, and G. Kriz, *Introduction to Spectroscopy*. 2 ed. Orlando: Saunders College publishing. 1996.
6. Denney, R.C., *A Dictionary of Spectroscopy*. 2 ed. New York: Wiley-Science. 1982.
7. Diem, M., *Introduction to Modern Vibrational Spectroscopy*. New York: John Wiley and Sons, Inc. 1993.
8. Herres, W. and J. Gronholz, *Understanding FT-IR data processing*. 1989, Bruker Analytische Messtechnik GmbH: Karlsruhe. p. 1-22.
9. Schrader, B., *IR and Raman Spectroscopy. Methods and Applications*. New York: VCH Publishers Inc. 1995.
10. Stuart, B., B. George, and P. McIntyre, *Modern Infrared Spectroscopy*, ed. D.J. Ando. New York: John Wiley & Sons. 1996.
11. Brereton, R., Fourier transforms: Use, theory and applications to spectroscopic and related data. *Chemometrics and Intelligent Laboratory Systems*, 1986. 1: p. 17-31.
12. Messerschmidt, R.G., Minimising optical nonlinearities in infrared microspectroscopy, in *Infrared Microspectroscopy Theory and Applications*, R.G. Messerschmidt and M.A. Hancock, Editors., Marcel Dekker, Inc.: New York. p. 1-19. 1998.

13. *Bruker Operating Manual*. Bruker Analytische Messtechnik GMBH: Germany. 1995.

14. Vogt, N., Soft modelling and chemosystemics. *Chemometrics and Intelligent Laboratory Systems*, 1987. **1**: p. 213-231.

15. Adams, M., *Chemometrics in analytical chemistry*. Cambridge: The Royal Society of Chemistry. 1995.

16. Albano, C., *et al.*, Four levels of pattern recognition. *Analytica Chimica Acta*, 1978. **103**: p. 429-443.

17. Kvalheim, O.M. and T.V. Karstang, SIMCA - Classification by means of disjoint cross validated principal components models, in *Data Handling in Science and Technology*, R. Brereton, Editor, Elsevier: Netherlands. 1992.

18. Wu, W. and D.L. Massart, Regularised nearest neighbour classification method for pattern recognition of near infrared spectra. *Analytica Chimica Acta*, 1997. **349**: p. 253-261.

19. Wu, W., *et al.*, Comparison of regularised discriminant analysis, linear discriminant analysis and quadratic analysis, applied to NIR data. *Analytica Chimica Acta*, 1996. **329**: p. 257-265.

20. Joliffe, I.T., *Principal Component Analysis*. New York: Springer-Verlag. 1986.

21. Scott, D., Determination of chemical classes from mass spectra of toxic organic compounds by SIMCA pattern recognition and information theory. *Anal. Chem.*, 1986. **58**: p. 881-890.

22. Dunteman, G.H., *Principal Components Analysis*. Quantitative Applications in the Social Sciences, ed. M.S. Lewis-Beck. Vol. 69. Newbury Park, CA: Sage Publications. 1989.

23. *The Unscrambler User Manual*. CAMO ASA: Sweden. 1998.

24. Frank, I. and R. Todeschini, *The Data Analysis Handbook: Data Handling in Science and Technology*. Vol. 14. Netherlands: Elsevier. 1994.

25. Wold, S., Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 1976. **8**: p. 127-139.

26. Zupan, J. and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*. 2 ed. New York: Wiley-VCH. 1999.

27. Hammerstrom, D., Neural networks at work. *IEEE Spectrum*, 1993. **June**: p. 26-32.

28. Maddalena, D., Applications of neural networks in chemistry. *Chemistry in Australia*, 1993. **May**: p. 218-221.

29. Burden, F., *The Analysis of Chemical Data*. Frank Burden: Monash University. 1996.

30. Bishop, C.M., *Neural Networks for Pattern Recognition*. New York: Oxford University Press. 1995.

31. Burden, F.R. and D.A. Winkler, Robust QSAR models using Bayesian regularised neural networks. *J. Med. Chem*, 1999. **42**(16): p. 3183-3187.

32. Burden, F.R., *et al.*, The use of automatic relevance determination in QSAR studies using Bayesian neural networks. In press, 2000.

33. MacKay, D.J.C., Bayesian interpolation. *Neural Computation*, 1992. **4**: p. 415-447.

34. MacKay, D.J.C., A practical Bayesian framework for backpropagation networks. *Neural Computation*, 1992. **4**: p. 448-472.

35. Thodberg, H.H., A review of Bayesian neural networks with an application to nea infrared spectroscopy. *IEEE Transactions on Neural Networks*, 1996. **7**(1): p. 56-72.

CHAPTER 3

MATLAB ROUTINES FOR PRE-PROCESSING IR SPECTRA

3 MATLAB ROUTINES FOR PRE-PROCESSING IR SPECTRA

Matlab (The Mathworks, Inc, MA, USA) is a software package that enables the user to utilise programs from an existing toolbox, or to write code for programs. Unless otherwise stated, the author of this dissertation wrote all programs used for pre-processing raw spectra prior to analysis. This chapter gives an explanation of the pre-processing routines written during this candidature.

The three main aims of pre-processing data are as follows [1]:

1. To reduce the amount of data and to eliminate irrelevant data.
2. To preserve or enhance sufficient information within the data in order to achieve a desired goal.
3. To extract information in, or transform the data to, a form suitable for further analysis.

Pre-processing the data ensures that all the spectra can be compared with one another because they have all been manipulated in the same way. Pre-processing also enables the objective removal of spectra that are inappropriate for inclusion in the analysis because they may be, for example, saturated or too noisy. Matlab was also used to perform several multivariate statistical techniques.

The pre-processing and multivariate statistical programs are controlled through an interactive front-end routine called *Cervjoin.m*[10]. This routine offers the user various menus, each containing several options. The main menu (Figure 3.1) is designed for importing data into the Matlab environment and the pre-processing menu (Figure 3.3) offers several standard functions. Although programs for most of these functions already existed, they were usually in different software and so data manipulation required importing and exporting data between various programs before the analysis could begin. This routine offers the user all the functions in one program. The analysis menu (Figure 3.12) calls multivariate statistical functions, and once again has the advantage of offering standard statistical functions in the one program.

---

[10] Words appearing in italics and ending with .m are Matlab program files. Refer to Appendix D for the code to each of these routines.

3.1 MAIN MENU

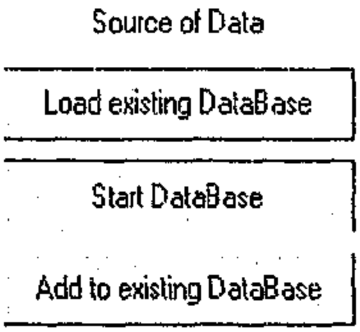

Figure 3.1 Main menu for the Matlab routine *Cervjoin.m*

3.1.1 LOAD EXISTING DATABASE

Selection of this option brings up a menu (Figure 3.2) offering choices of which Database to open. The first two options loads the Database in the form of a $n \times 1475$ matrix, where n represents the number of samples, i.e. the number of spectra, each with 1475 absorbance measurements in the region 3648 – 700 cm^{-1}. The third option is interactive and loads the chosen Database in the form of a $n \times m$ matrix, where m is equal to the number of wavenumber values plus a column of filenames and a column containing the diagnosis.

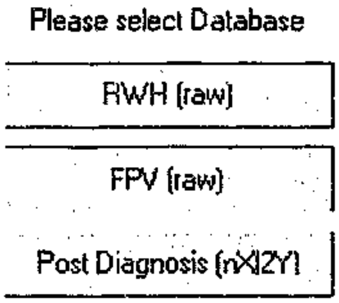

Figure 3.2 Load menu for the Matlab routine *Cervjoin.m*.

3.1.2 START NEW DATABASE

When this option is chosen the program *Jcampdbjoin.m* is called. This program searches the specified directory for '.dx' files, which are files in a JCAMP format[11]. The program imports each file individually, reading the data line by line to create a matrix in the form outlined above. Filenames, which are the patient numbers to allow diagnosis, are stored in

---

[11] Refer to Appendix E for an example of a spectrum in JCAMP format.

another matrix. The order of the filenames corresponds to the order of the spectra in the data matrix.

3.1.3 ADD TO EXISTING DATABASE

This option calls the program *Jcampcervjoin.m*, which imports JCAMP files into a pre-existing Matlab array, **DB**. Filenames for the newly imported files are imported into the pre-existing **Filename** matrix.

3.2 PRE-PROCESSING MENU

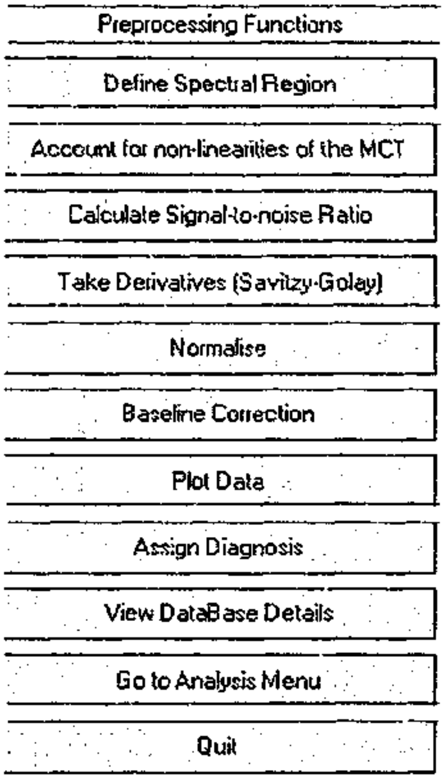

Figure 3.3 Pre-processing menu for the Matlab routine *Cervjoin.m*.

3.2.1 DEFINE SPECTRAL REGION

It is important with any pre-processing technique that the range of data can be manipulated or reduced. Selection of this option calls the program *Defregjoin.m* and allows the user to choose various regions, both interactively and pre-set. The user can choose from three options (Figure 3.4).

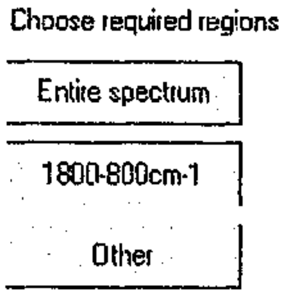

Figure 3.4 Define region menu for the Matlab routine *Defregjoin.m*.

The third option is interactive and enables the user to choose up to three spectral regions. This option is useful in principal component analysis where certain wavenumber regions are more influential than others in describing the variance. Figure 3.5 illustrates the resultant spectra for each of these options.

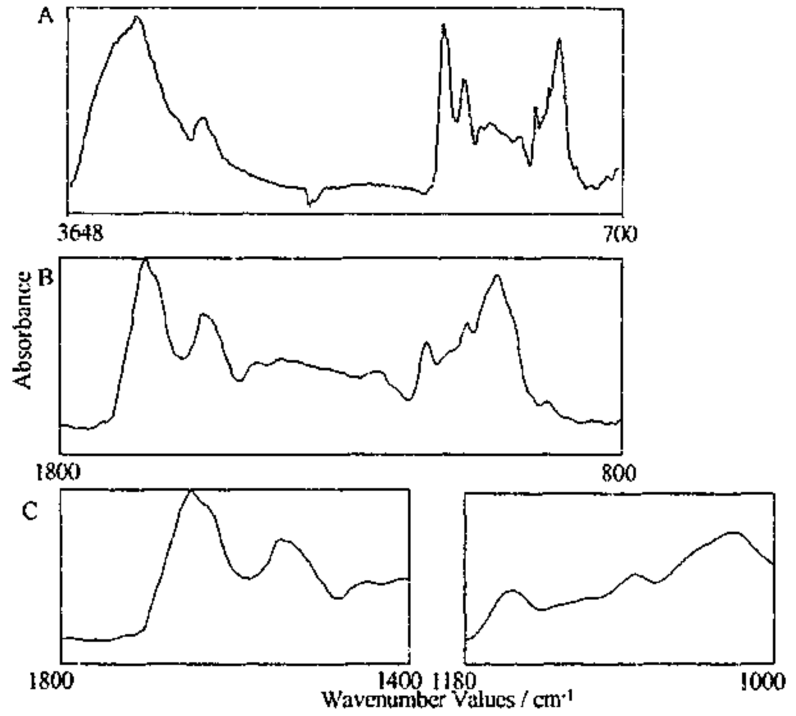

Figure 3.5 Spectral regions obtained through define spectral region option. A. Entire spectrum, B. 1800-800 cm^{-1} and C. Other.

3.2.2 ACCOUNT FOR NONLINEARITIES OF THE MCT

As discussed in Section 2.4, photoconductive detectors such as MCT detectors generate a non-linear response with respect to concentration and intensity. The Beer-Lambert Law, Equation 2.13, which describes the absorption of a substance with respect to concentration, is linear at small absorbance but becomes non-linear with absorbance greater than unity. To remove the effects of these two types of nonlinearities, spectra with absorbance greater than unity were removed from the database using the Matlab routine *nonlinjoin.m*. The routine finds the maximum absorbance in each spectrum and discards spectra with maximum absorbance greater than or equal to one, illustrated in Figure 3.6.

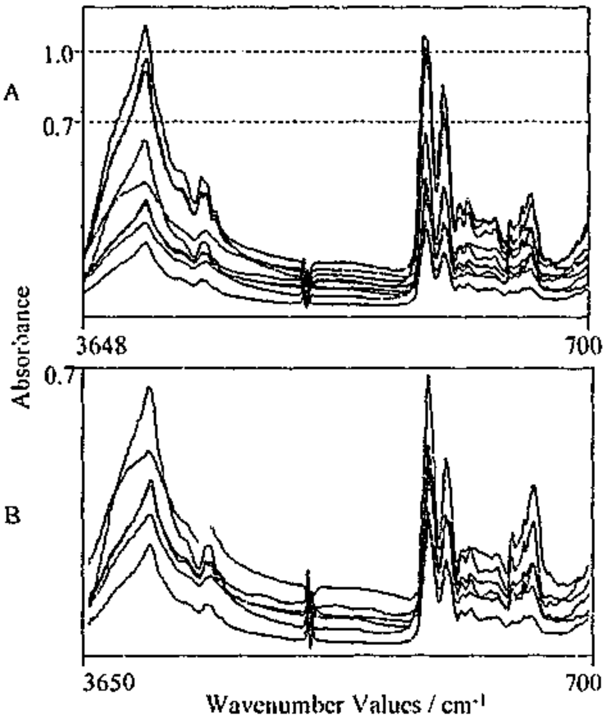

Figure 3.6 Spectra with a maximum absorbance greater than 1.0 will be discarded when this option is chosen.

3.2.3 SIGNAL-TO-NOISE RATIO (SNR)

Noise is a random effect that is visible as fluctuations of the baseline in a signal and was introduced in Section 2.2.4. The detection limit for an instrument is frequently considered to be a SNR of 2 - 3. To remove the effects of noisy spectra in the database, a program was written, *SNRjoin.m*, which calculates the SNR of a spectrum using the amide II band and removes all spectra from the database with a SNR less than 10. The SNR is calculated using the RMS (root mean square) method:

$$S/N = \frac{\text{average signal magnitude}}{\text{rms noise}}$$ **Equation 3.1**

The rms (root mean square) noise is the square root of the average deviation of the signal, x_i, from the mean noise value:

$$\text{rms noise} = \sqrt{\frac{\Sigma(\bar{x} - x_i)^2}{n-1}}$$ **Equation 3.2**

The first step in the calculation of SNR is to define a region of the spectrum free of peaks. This region was chosen to be 2100 – 1900 cm^{-1}. To overcome the problem of a non-zero baseline, a third order polynomial was fitted to this region of the spectrum and the defined region subtracted from the polynomial. This served two functions: first to enable the baseline to be flattened and offset to zero, and secondly to show the noise component of the spectrum to enable calculation of the RMS. Figure 3.7 gives a graphical representation of this process.

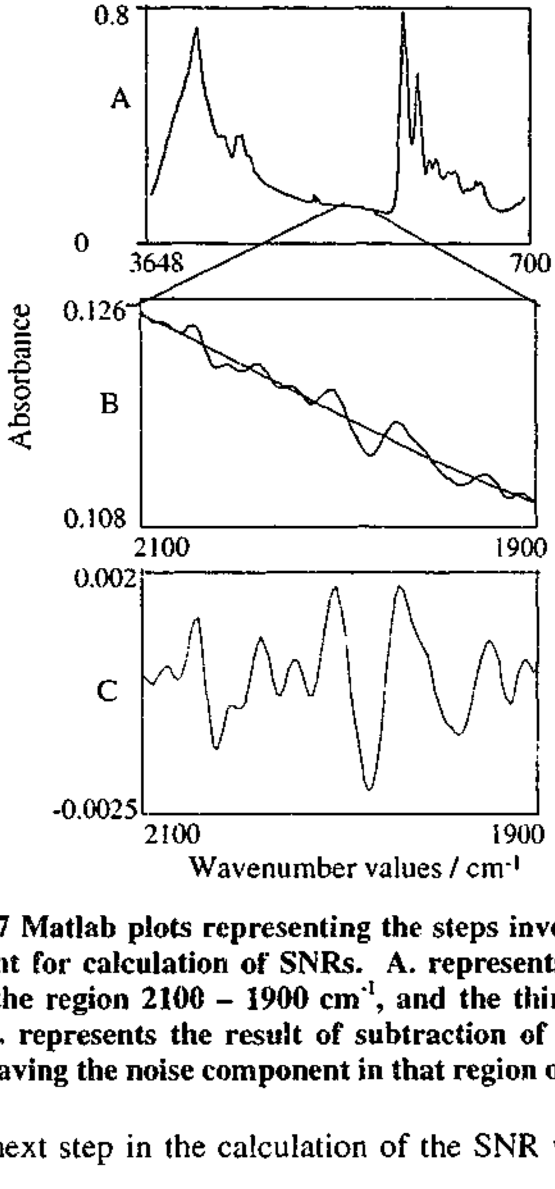

Figure 3.7 Matlab plots representing the steps involved in the calculation of the RMS noise component for calculation of SNRs. A. represents the entire spectrum. B. represents the noise in the region 2100 – 1900 cm^{-1}, and the third order polynomial (black) fitted to the noise. C. represents the result of subtraction of the fitted polynomial from the specified region, leaving the noise component in that region of the spectrum.

The next step in the calculation of the SNR was to determine the signal component. The amide II peak was chosen for this purpose, because the amide I peak is often used for normalisation. Once again, to overcome the problem of a non-zero baseline, baseline correction of the spectra in the region 2100-700 cm^{-1} was performed. A detailed description of baseline correction is given in Section 3.2.6. Once the spectrum had been baseline corrected, the signal component was taken to be the height of the amide II peak and the SNR was calculated by dividing this value by the RMS of the noise component.

3.2.4 DERIVATIVES (SAVITZKY-GOLAY)

A useful way of presenting and interpreting infrared spectra is by using differentiation. Infrared spectra are zero'th order but with the use of mathematical functions first, second and higher-order derivatives can be generated. Taking the derivative of a spectrum offers an apparent increase in resolution of the differential data compared with the original spectrum, as well as removing baseline effects. Derivative spectra emphasise changes in slope that are difficult to detect in the zero'th order spectrum. However noise, which is often comprised of high frequency components, may be amplified by differentiation [1].

There are many mathematical procedures that may be employed to differentiate spectral data. Assuming data is recorded at evenly spaced intervals along the λ axis, the simplest method to produce the first-derivative spectrum is by difference:

$$\frac{dy}{d\lambda} = \frac{y_{i+1} - y_i}{\Delta\lambda}$$ **Equation 3.3**

or

$$\frac{dy}{d\lambda} = \frac{y_{i+1} - y_{i-1}}{2\Delta\lambda}$$ **Equation 3.4**

and for the second derivative:

$$\frac{d^2 y}{d\lambda^2} = \frac{y_{i+1} - 2y_0 - y_{i-1}}{\Delta\lambda^2}$$ **Equation 3.5**

where y represents the spectral intensity or absorbance. The use of polynomial derivatives, such as the Savitzky-Golay algorithms [2], which utilises an array of weighted coefficients as a smoothing function to convolute the spectral data reduces the problems of amplified noise. Using a quadratic polynomial and a five-point moving window, the first derivative is given by

$$\frac{dy}{d\lambda} = \frac{1}{10\Delta\lambda}(-2y_{i-2} - y_{i-1} + y_{i+1} + 2y_{i+2})$$ **Equation 3.6**

and for the second derivative

$$\frac{d^2 y}{d\lambda^2} = \frac{1}{7\Delta\lambda^2}(2y_{i-2} - y_{i-1} - 2y - y_{i+1} + 2y_{i+2})$$ **Equation 3.7**

the use of additional terms sampling extra points from the data provides a better approximation compared with the results obtained using Equations 3.3, 3.4 and 3.5 [1].

A Savitzky-Golay smoothing and differentiation routine from the Matlab toolbox was adapted for use in the *Cervjoin.m* routine. Examples of the resultant first- and second-order derivative spectra using the *svgljoin.m* routine are shown in Figure 3.8B and Figure 3.8C respectively. Figure 3.8A is the original spectrum.

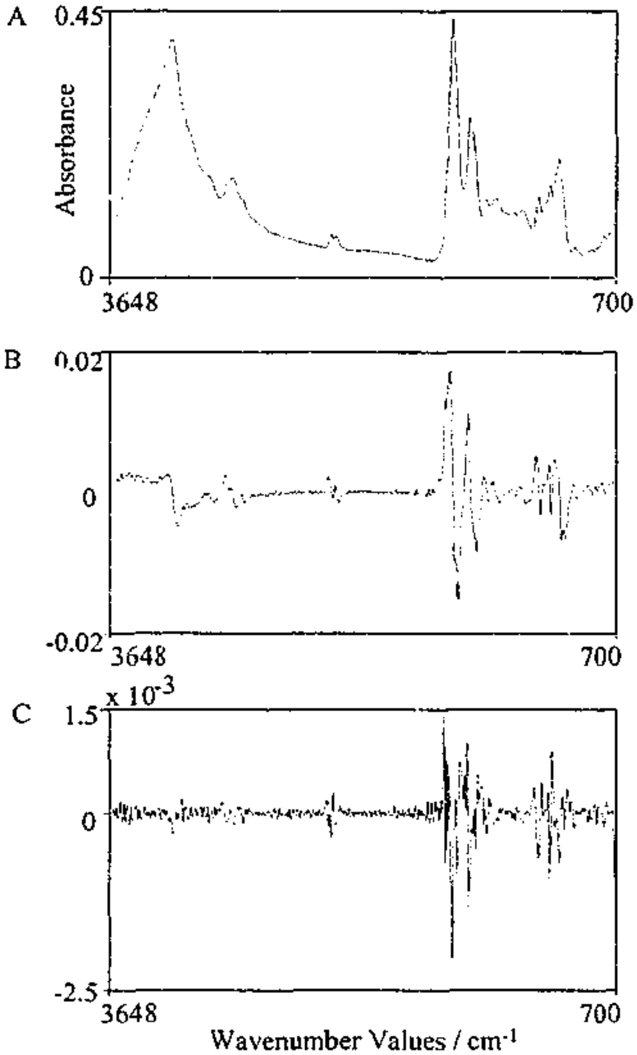

Figure 3.8 A. Absorbance spectrum. B. First order derivative, quadratic polynomial with 9 smoothing points. C. Second order derivative, cubic polynomial with 9 smoothing points.

3.2.5 NORMALISATION

In order to compare different spectra it is often useful to perform a normalisation function on the spectra so that they are "scaled" in order to achieve specific properties.

There are several different normalisation techniques that can be used to manipulate absorbance spectra. These include vector, maximum, mean, and range normalisation.

3.2.5.1 VECTOR NORMALISATION

In vector normalisation the average y-value (absorbance) of the spectrum is calculated first. This average value is then subtracted from the spectrum so that the middle of the spectrum is pulled down to y = 0. The sum of the squares of all y-values is then calculated and the spectrum is divided by the square root of this sum. The vector norm of the result spectrum is 1 [3].

$$X(i,k) = \frac{(X(i,\bullet) - \overline{X}(i,\bullet)}{\sqrt{\sum (X(i,\bullet))^2}} \qquad \textbf{Equation 3.8}$$

3.2.5.2 MAXIMUM NORMALISATION

This technique divides each sample by its maximum absolute value and is relevant only if all values of the curve have the same sign. If all values of the curve are positive, the maximum value becomes +1 [4].

$$X(i,k) = \frac{X(i,k)}{\max(|X(i,\bullet)|)} \qquad \textbf{Equation 3.9}$$

3.2.5.3 MEAN NORMALISATION

This is the most classical case of normalisation and the area under the curve becomes the same for each sample that has been normalised using this technique. This technique works by dividing each sample in a data matrix by its average and is the same as replacing the original variables by a profile centred about 1. Only the relative values of the variables are used to describe the sample. This transformation is not relevant if all values of the curve (spectrum) do not have the same sign [4].

$$X(i,k) = \frac{X(i,k)}{|\overline{X}(i,\bullet)|} \qquad \textbf{Equation 3.10}$$

3.2.5.4 RANGE NORMALISATION

In this technique each sample is divided by its range, i.e. maximum value minus minimum value. With range normalised samples the curve span becomes 1 [4].

$$X(i,k) = \frac{X(i,k)}{\max(i,\bullet) - \min(i,\bullet)} \qquad \textbf{Equation 3.11}$$

Figure 3.9A-D illustrates the effects on a spectrum of performing vector, maximum, mean and range normalisation respectively. The same spectrum (blue) was used for each normalisation technique.

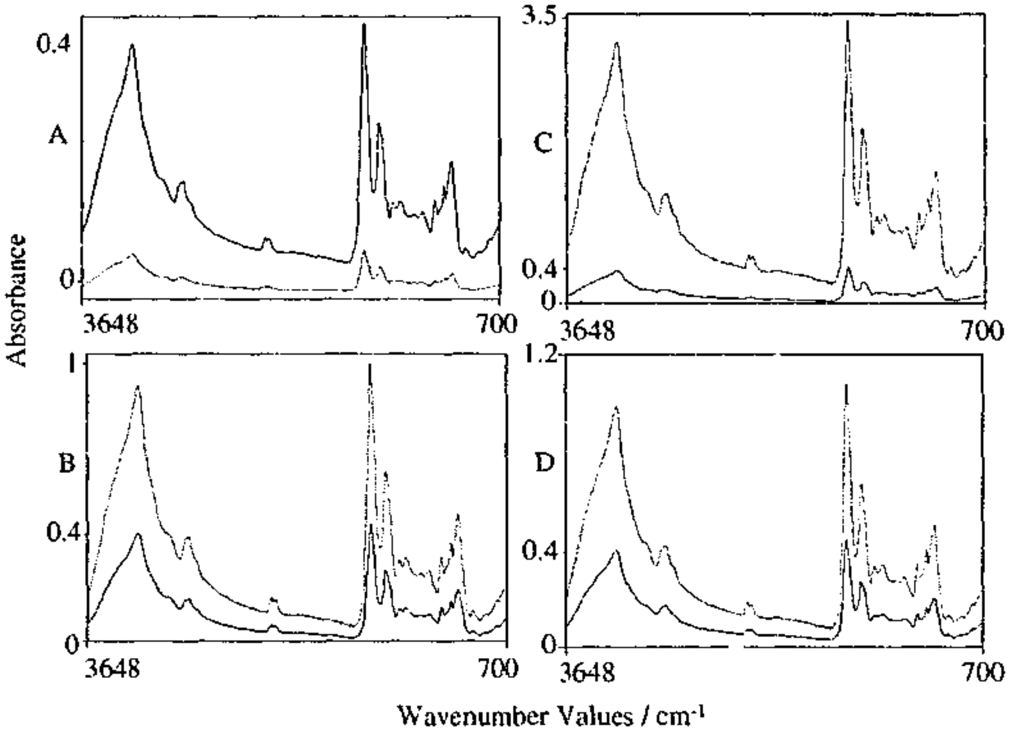

Figure 3.9 Illustration of normalisation techniques. The blue spectrum in each plot is un-normalised and the green spectrum in each plot is the result of performing vector (A), maximum (B), mean (C) and range (D) normalisation.

3.2.6 BASELINE CORRECTION

Baseline correction is an important pre-processing technique. Spectra produced from inhomogeneous samples, and from samples with non-uniform thickness will often have a non-zero baseline. Scattering effects from the solvents used may also cause a non-zero baseline. Many of the spectra recorded from samples collected in saline solution, for example show this effect as the infrared radiation is scattered by the salt crystals that are formed in the desiccation process. The baseline correction option in the pre-processing routine performs a linear regression of the baseline of the raw spectrum and uses the linear line produced to offset the entire spectrum to zero. Three regions free of absorbance peaks are chosen and the minimum absorbance in each region is determined. Two linear regression lines are fitted between these minima and the raw spectrum is subtracted from these lines to offset the entire spectrum to zero. The three regions chosen for detection of minima are 3648-3646 cm^{-1}, 2100-1800 cm^{-1} and 1000-700 cm^{-1}. Figure 3.10 illustrates this process.

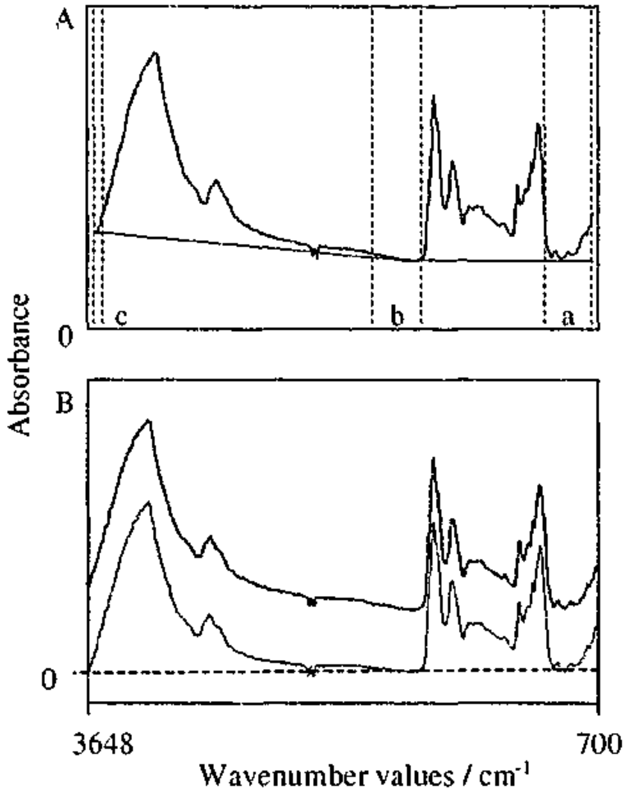

Figure 3.10A Spectrum to be baseline corrected (blue). Three regions are chosen: a (3648-3646 cm^{-1}), b (2100-1800 cm^{-1}) and c (1000-700 cm^{-1}) and the minimum absorbance for each region is determined. Linear regression is performed and a linear line (green) is fitted between the minimum in each region. The original spectrum (blue) is subtracted from this line to give the baseline corrected spectrum illustrated in B (red).

3.2.7 PLOT DATA

Selection of this option plots all the spectra in the Database against the wavenumber values. All plots in this chapter were generated using this option.

3.2.8 ASSIGN DIAGNOSIS

In order to perform statistical analysis on the spectra imported into the database, it was necessary to assign each spectrum a diagnosis. Most of the spectra obtained from samples collected from the Royal Women's Hospital (RWH) had both histological (biopsy) and cytological (Pap smear) results. Spectra obtained from samples collected from Family Planning Victoria (FPV) had only cytological results. Each sample was assigned a numerical code according to Table 3.1. This table was generated according to numerical codes used for reporting histological and cytological results at the RWH. Often more than one code was applicable to the samples and so the diagnosis for some samples consisted of a string of the applicable codes joined together.

Table 3.1 Numerical codes and corresponding diagnosis assigned to spectra.

Numerical Code	Diagnosis
01	Normal, negative
02	Normal metaplasia (mature)
03	Atypical metaplasia
04	Mild dysplasia (CIN I)
05	Moderate dysplasia (CIN II)
06	Severe dysplasia (CIN III)
07	Carcinoma *in situ* (CIS)
08	Minimal stromal invasion
09	Frank stromal invasion
10	Atypia
11	Dysplasia, unspecified
12	Atrophy
13	Inflammation
14	Normal metaplasia (immature)
15	Condyloma
16	Herpes
17	Vaginal adenosis
18	Hyperplasia (endometrial)
19	Hyperplasia (endocervical)
20	Polyp
21	Wart virus changes (HPV)
22	Hyperkeratosis
23	Not done, not applicable
24	Epithelium absent or denuded
25	High grade epithelial changes (HGEA)
26	Low grade epithelial changes (LGEA)
27	Bacterial vaginosis
28	Candida
29	Cervicitis
30	Inconclusive
31	Unsatisfactory
32	Bacteria, unspecified
33	Vault (vaginal smear)
34	Actinomyces
35	Trichomonas
36	Endocervical only
37	Endometriosis
38	Endocervical / glandular component absent
39	Keratinisation
55	Cytolysis
65	Koilocytes

Table 3.1 indicates the diversity of possible diagnoses resulting from Pap smears and biopsies. Whilst the majority of these diagnoses are not abnormal, in terms of dysplasia and malignancy, as discussed in Chapter 1, inflammation, metaplasia and other benign cellular changes (BCCs) may cause possible confounding results in infrared spectroscopy. Therefore it was decided that the analysis would be performed using several different options for diagnosis, as outlined in Figure 3.11.

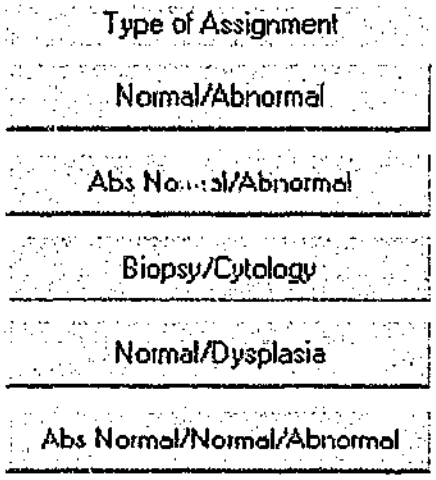

Figure 3.11 Diagnosis menu for the Matlab routine *Cervjoin.m*.

When the assign diagnosis option in the *Cervjoin.m* routine is selected, the user can choose between assigning a diagnosis for RWH or FPV data. *DiagnosisRWH.m* or *DiagnosisFPV.m* is called and this program loads a text file containing the filenames of the spectra and their corresponding diagnosis. The filenames in the database are compared with the filenames in the text file and each spectrum is assigned a diagnosis according to the chosen option from the menu illustrated in Figure 3.11.

3.2.8.1 NORMAL/ABNORMAL

Selection of this option assigns spectra a diagnostic value of 1 (abnormal) for samples with high-grade dysplasia (CIN II and III), dysplasia unspecified, CIS, invasive carcinoma or HGEA, excluding samples exhibiting CIN I, or LGEA and 0 (normal) for all other diagnoses.

3.2.8.2 ABS NORMAL/ABNORMAL

Selection of this option assigns spectra a diagnostic value of 1 (abnormal), as outlined in Section 3.2.8.1 and 0 (normal) for samples with a diagnosis of normal or negative only, excluding samples exhibiting BCC or metaplasia.

3.2.8.3 BIOPSY/CYTOLOGY

Selection of this option assigns spectra the exact numerical code given in Table 3.1. FPV spectra will only have a cytological diagnosis.

3.2.8.4 NORMAL/DYSPLASIA

Selection of this option assigns a diagnostic code of 1, 2, or 3 for samples exhibiting CIN I, CIN II, CIN III respectively, 4 for samples exhibiting CIS, 5 for samples exhibiting minimal or frank stromal invasion and 0 (normal) as outlined in Section 3.2.8.2.

3.2.8.5 ABS NORMAL/NORMAL/ABNORMAL

Selection of this option assigns a diagnostic code of 0 (abs normal) for samples with a diagnosis of normal or negative only, 1 (normal) for samples which have been diagnosed normal or negative but exhibit BCC or metaplasia, 2 (abnormal) for samples exhibiting LGEA, CIN 1 and/or HPV effects and 3 (abnormal) for samples exhibiting high-grade dysplasia, CIS, HGEA, dysplasia unspecified or invasion.

3.2.9 VIEW DATABASE DETAILS

Selection of this option returns a list of all the variables in the routine and their size. This option is useful for determining how many spectra are in the Database, particularly after performing a task that causes spectra to be discarded, for example the non-linearity and SNR routines.

3.3 MULTIVARIATE ANALYSIS

An introduction to the multivariate statistical techniques called by this menu (Figure 3.12) is given in Section 2.5. Matlab code for all techniques, except discriminant analysis, were obtained from the Matlab toolbox with only minor changes necessary for incorporation into *Cervjoin.m*. The discriminant analysis routines were modified considerably as the results from the existing routine gave ambiguous results. The codes for LDA (*LDAcerv.m*) and QDA (*QDAcerv.m*) can be found in Appendix D.

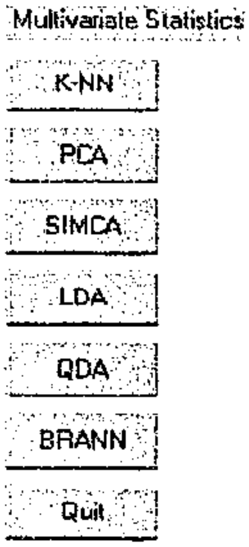

Figure 3.12 Multivariate statistics menu for the Matlab routine *Cervjoin.m*.

3.4 REFERENCES

1. Adams, M., *Chemometrics in analytical chemistry*. Cambridge: The Royal Society of Chemistry. 1995.
2. Savitzky, A. and M. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 1964. **36**(8): p. 1627-1639.
3. *OPUS / IR Version 2.0 Reference Manual*. Bruker Analytische Messtechnik GMBH: Germany. 1995.
4. *The Unscrambler User Manual*. CAMO ASA: Sweden. 1998.

CHAPTER 4

AN INVESTIGATION INTO THE INFLUENCE OF HORMONES AND NUCLEIC ACIDS IN THE DIAGNOSIS OF CERVICAL CANCER

4 AN INVESTIGATION INTO THE INFLUENCE OF HORMONES AND NUCLEIC ACIDS IN THE DIAGNOSIS OF CERVICAL CANCER

This chapter is divided into two sections. The first section investigates the influence of hormonal stimulation on IR spectra of cervical smears in an effort to determine if smears taken during certain phases of the menstrual cycle may confound diagnosis. The second section investigates subcellular fractionation as a means of isolating the nuclei from epithelial cells. This was to determine the exact contribution of the nucleus to IR spectra of cervical smears.

4.1 HORMONAL STUDY

Despite the earlier work of Wong *et al.* [1, 2], Yazdi *et al.* [3], and Fung *et al.* [4], which indicated that infrared spectroscopy is a powerful tool in the discrimination of normal and malignant cervical cells, it is becoming increasingly apparent that there may be other factors contributing to the spectral changes assumed to be arising from neoplastic processes and malignancy. Possible contributing factors and/or confounding variables have been identified. They include benign cellular changes (BCC), specifically metaplasia and inflammation [3]; erythrocytes [5] and lymphocytes [5, 6]; endocervical cells [5] and mucins [5-7].

Recently Diem's group [8] conducted a series of experiments demonstrating that infrared spectroscopy could be used to monitor maturation and differentiation in cervical squamous epithelium. The observed spectral differences between the basal, parabasal, intermediate and superficial layers of the squamous epithelium arose mainly in the 1200-900 cm^{-1} region. The spectral differences observed showed an increase in glycogen concentration towards the surface, i.e. as cells matured from the basal layer they accumulated more glycogen. Differences were also noted in the amide I/amide II ratio, believed to be a result of nucleic acid contributions. Despite these differences, Cohenford and Rigas [9] found that spectra of cytologically normal intermediate and superficial squamous cells from women with dysplasia or cancer were different from spectra of intermediate and superficial squamous cells in cytologically normal smears.

Multivariate statistics have been utilised by several groups to obtain a separation between the infrared spectra of normal, dysplastic and malignant samples. Wood *et al.* [10] used principal component analysis (PCA) to achieve a separation between the infrared spectra of normal and dysplastic cells. Cohenford *et al.* [11] employed principal component regression (PCR) to achieve a separation between normal and malignant cervical cells. Romeo *et al.* [12] used PCA coupled with ANNs to classify unknown dysplastic and normal samples.

Morphologically there are many changes occurring in cervical epithelium as a direct result of hormonal stimulation from the menstrual and ovarian cycles. A detailed description of these changes is given in Section 1.4.1. Cervical squamous epithelial cells accumulate glycogen as a process of maturation, the concentration of which is hormone dependent, peaking around ovulation [13]. Given the findings by Chiriboga *et al.* [8], outlined above, it would seem likely that the infrared spectra of cervical cells sampled throughout the menstrual cycle would exhibit spectral differences.

4.1.1 METHODOLOGY

4.1.1.1 PARTICIPANTS

Information for potential volunteers for this study took the form of a poster (Appendix F) placed around the Monash University Clayton Campus. Participants in this study were required to be pre-menopausal non-smokers with a history of normal Pap smears, the most recent within the last twelve months. Because nicotine has been found to affect cellular proliferation of the cervix [14], smokers were not included in this study to reduce the number of contributing variables. Interested women were informed about the project, given an explanatory statement and required to provide written consent to their participation[12].

Inclusion in the study was limited to women taking either monophasic or no form of oral contraception. The women taking monophasic contraception were asked to have a cervical smear once a week. Women not taking oral contraception were required to have four cervical smears each cycle corresponding to: postmenstrual, preovulatory, postovulatory, and premenstrual phases of the menstrual cycle. These women were also required to have blood taken once each cycle (postovulatory). Progesterone assays were undertaken to ensure that ovulation had occurred and that the women had functional menstrual cycles. Cervical smears were obtained at the Health Service on campus.

---

[12] Refer to Appendix F for explanatory statement and informed consent form.

4.1.1.2 Sample collection

Cervical cells were taken from the transformation zone of the cervix with both an Ayre spatula to ensure collection of ectocervical cells, and a Cytobrush™ (MEDSCAND, Hollywood, Fl, USA) to ensure collection of endocervical cells. Sampling instruments were agitated in separate 50 cm^3 centrifuge tubes containing 10 cm^3 absolute ethanol to collect the cells and then stored at –70°C until required.

4.1.1.3 Sample preparation

Samples were centrifuged at 2500 r.p.m. for 10 minutes and the ethanol supernatant removed with an automated pipette leaving a cellular pellet. Ultra-clean water was added and the tubes were then vortexed to re-suspend and clean the cellular material. This washing procedure was carried out three times and the cellular material was then pipetted into a KRS-5 multicavity cell and desiccated under vacuum.

4.1.1.4 Multicavity infrared cell

The multicavity infrared cell was purpose built for the analysis of biological samples [10]. The cell consists of a KRS-5 (thallium bromide iodide) crystal (50×30×5 mm) mounted between two aluminium plates (Figure 4.1). The top plate has fourteen regularly spaced 5-mm diameter cavities. The cavities are countersunk to allow the positioning of O-rings. The top plate has a dual function, acting as both a seal and a surface protector.

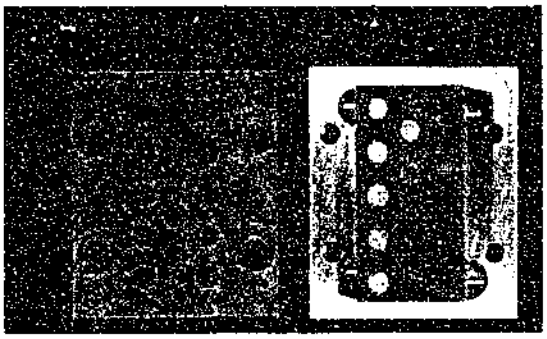

Figure 4.1 Multicavity IR cell comprising two aluminium plate, one with 14 countersunk holes, and a KRS-5 infrared substrate.

4.1.1.5 Infrared microspectroscopy

Following removal of the aluminium plates used to form sample wells, the KRS-5 multicavity IR cell containing fourteen cervical samples was placed on the sampling platform of a Perkin-Elmer IR microscope coupled to a Perkin-Elmer 1600 spectrophotometer. A minimum of six transmission spectra were recorded for each sample with the knife-edge aperture reduced to 50 x 50 μm. For each spectrum, 16 scans were co-added at a resolution of 8 cm^{-1}, with a total recording time for each spectrum of 20 seconds.

4.1.1.6 DATA TREATMENT

Infrared spectra were transferred via a Macintosh computer in JCAMP format onto an OPUS (Bruker Messtechnik, Karlsruhe, Germany) operating platform where the spectra from each sample were re-scaled, baseline corrected, normalised to the amide I peak (1650 cm^{-1}), averaged and converted into an ASCII format recognised by *Unscrambler II* (CAMO ASA, Oslo, Norway). Unscrambler enabled visual inspection of the spectra and multivariate statistical analysis.

4.1.1.7 BLOOD PREPARATION

Blood samples (10 cm^3) were centrifuged at 2500 r.p.m. for 15 minutes. Centrifugation separated the blood into three components: red blood cells, white blood cells, and serum. The serum was pipetted into a 10 cm^3 centrifuge tube and frozen until the assay was performed.

4.1.2 RESULTS AND DISCUSSION

Eleven non-smoking women participated in this study for periods of between four and twelve weeks. Table 4.1 summarises the length of participation.

Table 4.1 Summary of participation in the hormonal study.

Number of Women	Length of participation (weeks)	Oral Contraception
2	12	Monophasic
2	12	None
1	8	Monophasic
1	8	None
3	4	Monophasic
2	4	None

4.1.2.1 PROGESTERONE ASSAY

All of the serum samples collected from participants showed progesterone, indicating that ovulation had occurred and the women had functional ovulatory cycles.

4.1.2.2 CONTAMINATION

A thin, white substance covered a high proportion of the ectocervical cell sample deposits. The infrared spectra of these deposits, shown in Figure 4.2A, appeared unusual and were characterised by a doublet at 1053 and 1036 cm^{-1}, and peaks at 1730, 1323, 1235, 1160 and 1108 cm^{-1}.

The origin of this contaminant was considered to have arisen from loose fibres on the Ayre spatula, removed by agitation of the instrument in ethanol when collecting cellular material. An Ayre spatula was vigorously agitated in a solution of absolute ethanol, and this solution was centrifuged and the resulting pellet pipetted into a KRS-5 infrared cell and desiccated under vacuum. The resulting spectrum, also shown in Figure 4.2, showed similarities with the contaminated ectocervical cell spectrum, with peaks at 1730, 1323, 1232, 1157, 1108 and 1033 cm^{-1}.

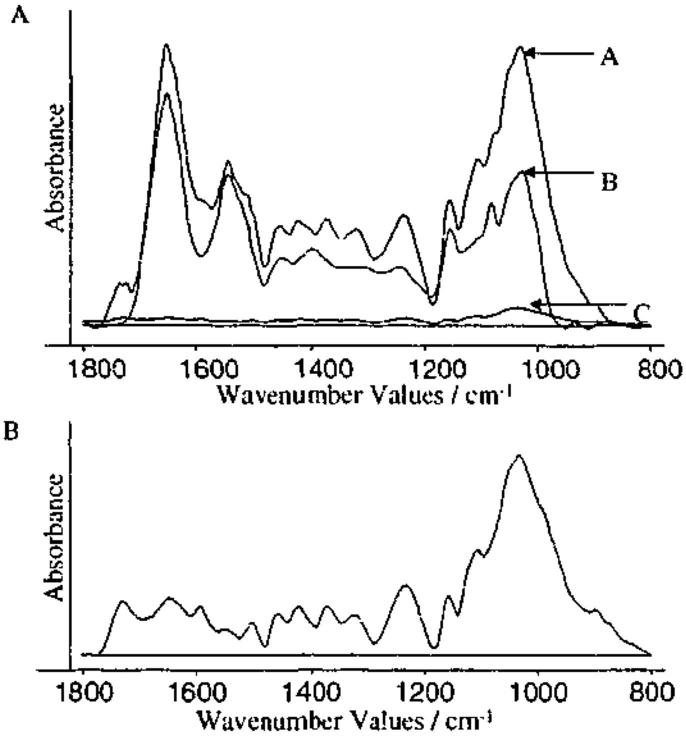

Figure 4.2A Spectrum of ectocervical cells contaminated with Ayre spatula debris (A, black), uncontaminated ectocervical cells (B, blue) and spectrum of Ayre spatula debris (C, red). Spectra have not been normalised. B. Ayre spatula debris, normalised to the most intense band.

Sampling instruments were initially kept in centrifuge tubes to maximise cell collection, however once the contaminant was identified as Ayre spatula debris, sampling instruments were briefly agitated in the ethanol solution immediately after collection and then discarded.

A further source of contamination identified were mixed populations of ectocervical and endocervical cells. Cervical smears are obtained from the transformation zone of the cervix, which is the area of the cervix where the squamous epithelium (ectocervical) and columnar epithelium (endocervical) meet and where neoplasia is likely to occur [15]. It is sometimes difficult to sample squamous and columnar cells separately because the transformation zone may not always be visible to the person taking the smear. The location of the transformation zone changes throughout the lifetime of a female and depends on age, reproductive status and pregnancy [16]. Consequently samples which showed IR spectra characteristic of both ectocervical and endocervical epithelium were discarded to minimise the chance of spectra resulting from a mixture of these two components confounding the effects occurring from cyclical changes.

4.1.2.3 SPECTRAL SUBTRACTION

To remove the effects of Ayre spatula contamination in the ectocervical cell spectra, spectral subtraction was performed on contaminated spectra. Spectral subtraction was performed iteratively using Grams 3.2 software (Warsash, Sydney, Australia). Because the relative intensities of the spatula spectrum compared to the ectocervical cell spectra were so low, direct subtraction of the spatula could not be performed due to scaling factors. Instead, spectral subtraction was performed by subtraction of a non-contaminated spectrum from a contaminated spectrum. Whilst this method is less than ideal, the same spectrum was used for all the subtractions.

4.1.2.4 ECTOCERVICAL CELLS

Ectocervical cells are characterised by:

1. Amide I and amide II bands at 1651 and 1544 cm^{-1} respectively.
2. Peaks at 1543 and 1394 cm^{-1} arising from deformation modes of methyl groups in proteins.
3. A very weak amide III band at 1318 cm^{-1}.
4. $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands at 1242 and 1081 cm^{-1} respectively.
5. A band at 1154 cm^{-1} arising from νC-O of proteins and carbohydrates.

6. A peak at 1027 cm^{-1} arising from the –CH_2OH stretching vibration of glycogen.

The infrared spectra of ectocervical cells exhibited variation throughout the cycle and an example is shown in Figure 4.3. The main differences were observed in the carbohydrate region (1200-1000 cm^{-1}). Increases in the intensities of the peaks attributable to glycogen would be expected towards mid-cycle, as a result of glycogen accumulation in intermediate cells [13]. This increase is clearly seen with the infrared spectra resulting from days 8 and 12 showing marked increases in the 1025 cm^{-1} band, attributed to glycogen, compared to days 19 and 26. Glycogen concentrations are expected to peak around ovulation, which in this 30 day cycle would occur at day 16.

Of the seventeen cycles of women not taking oral contraception examined in this study, all but two cycles exhibited this spectral pattern, i.e. the glycogen band at 1025 cm^{-1} increased towards mid-cycle and decreased following ovulation.

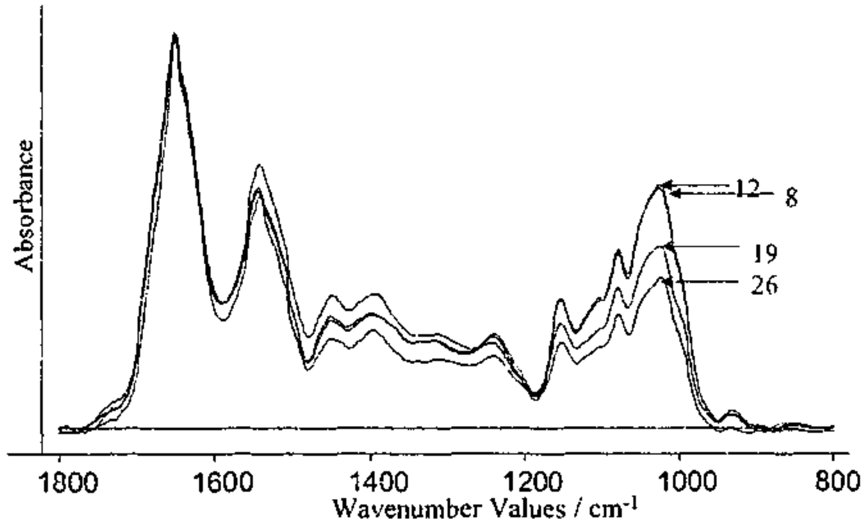

Figure 4.3 Infrared spectra obtained from a woman, not taking oral contraception, over one cycle. The length of this cycle was 30 days.

The consistency of these changes is also observed when similar days from different cycles are examined. Figure 4.4 shows the mid-cycle days from a woman not taking oral contraception. The slight differences in these spectra result from different cycle lengths, and it is known that cycle length both between and within women can differ substantially [17].

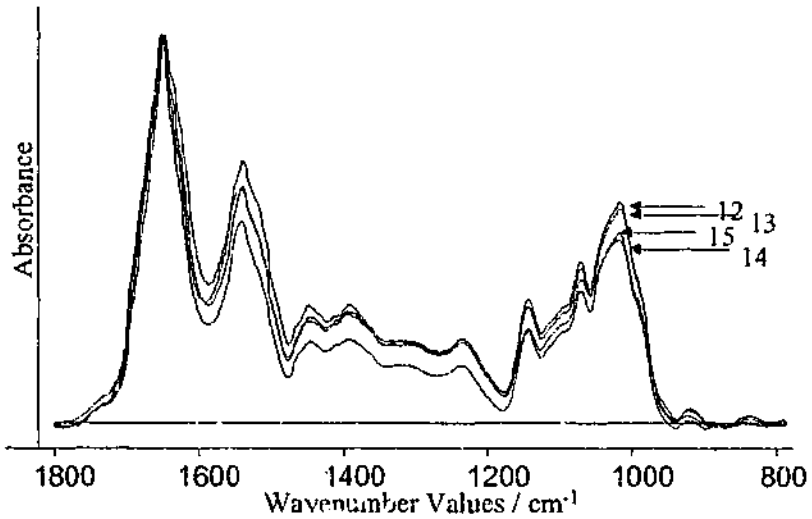

Figure 4.4 Infrared spectra of a woman, not taking oral contraception, collected over four cycles.

Figure 4.5 shows the IR spectra of ectocervical cells collected from a woman taking monophasic oral contraception. As expected the spectra do not exhibit large variation throughout the cycle, with small differences observed in the carbohydrate region. All of the cycles recorded from the monophasic participants showed similar spectral patterns, although at the end of the cycle, glycogen levels were found to decrease. This is believed to be due to the withdrawal of monophasic contraception to allow for menstruation.

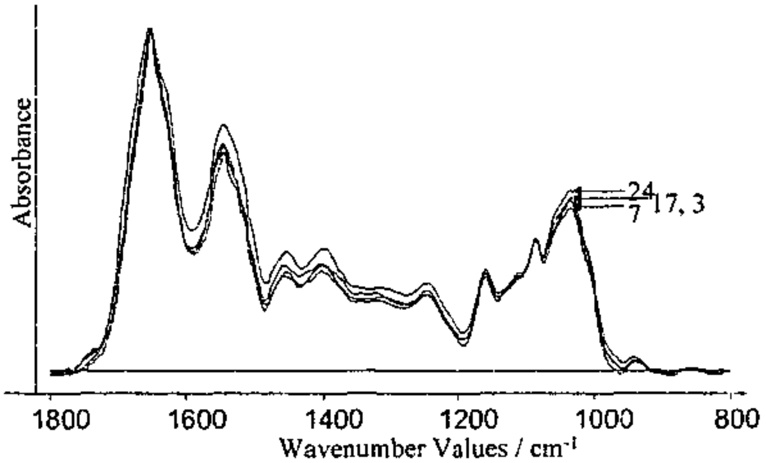

Figure 4.5 Infrared spectra of a woman taking monophasic oral contraception collected over one cycle.

Whilst it is obvious that there is a large degree of variability in cervical cells throughout the menstrual cycle, these cells are essentially normal and would be expected to group with normal cervical cells in multivariate statistics.

PCA was performed on two hundred and forty infrared spectra with known biopsy results. Figure 4.6A shows a clear separation between normal and dysplastic samples on a PC1 versus PC2 scores plot. Sixty-six infrared spectra from all participants were added to the data set and PCA was performed a second time. The resulting PC1 versus PC2 scores plot is shown in Figure 4.6B, with M and N representing monophasic and no contraception respectively and the number indicating the day of the cycle the cells were obtained. A separation between the abnormal and the normal/menstrual samples is observed indicating spectral changes arising throughout the menstrual cycle do not confound IR diagnosis of high-grade dysplastic and carcinoma *in situ* (CIS) samples.

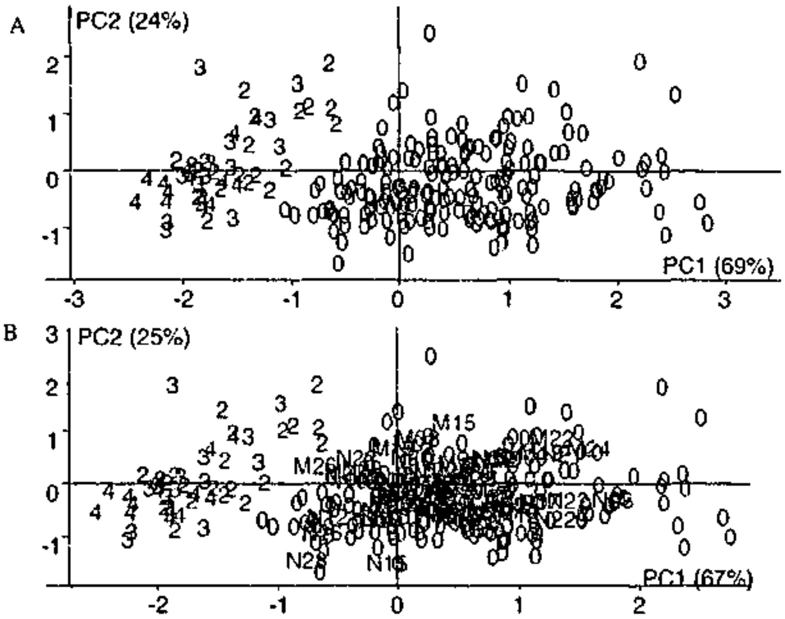

Figure 4.6 Scores plots showing discrimination between normal (green, 0) and dysplastic (blue) cervical cells (A) and with menstrual data included (B). The spectra were diagnosed by biopsy as normal (0), CIN II (2), CIN III (3) or CIS (4). The menstrual samples (red) are represented by an M or N, for monophasic or no contraception respectively, with the number indicating the day of the cycle the smear was obtained.

4.1.2.5 NAVICULAR CELLS

Navicular cells are a common variant of intermediate cells and are sometimes seen in cervical smears of pregnant women and smears taken during the late menstrual phase [15]. These cells are characterised by large deposits of glycogen and are shown in Figure 4.7. Several spectra, recorded between days 8 and 11 and shown in Figure 4.8, exhibited glycogen peaks with intensities equal to or greater than the intensity of the amide I band. These spectra may be the result of the presence of navicular cells.

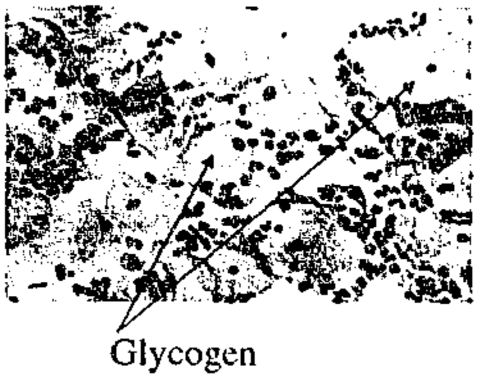

Figure 4.7 Microscopic image of a cervical smear showing the presence of navicular cells. These cells are of intermediate maturation and are filled with vacuoles of glycogen.

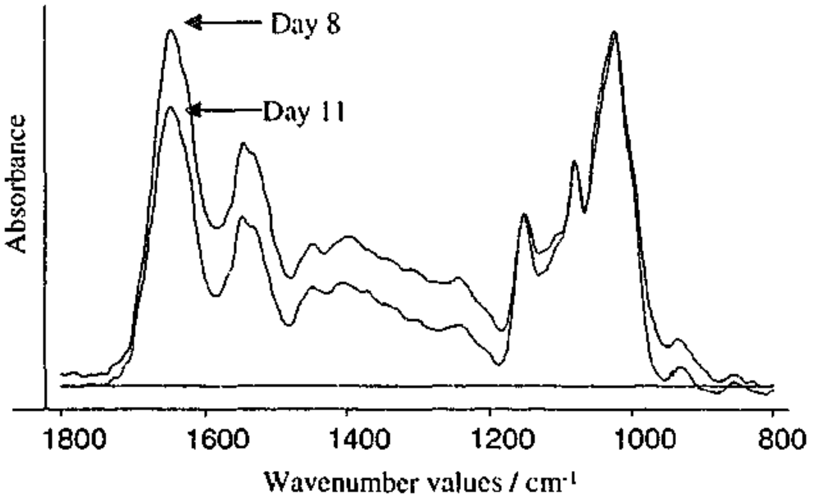

Figure 4.8 Infrared spectra of ectocervical cells sampled at day 11 of a 30 day cycle and day 8 of a 27 day cycle.

4.1.2.6 ENDOCERVICAL CELLS

The infrared spectra of endocervical cells differ somewhat from ectocervical cell spectra. Differences occur in the carbohydrate region (1200-1000 cm^{-1}) because endocervical cells do not contain glycogen [16, 18] and so lack the 1025 cm^{-1} peak arising from glycogen. The majority of endocervical cells present in cervical smears are mucus secreting [16] and spectra are differentiated from ectocervical cells by a band at 1047 cm^{-1} and a shoulder at 1120 cm^{-1} arising from the C-O and -CH_2OH stretching vibrations respectively of mucins.

Figure 4.9 illustrates the minimal variation in the IR spectra of endocervical cells throughout the cycle. The amount of protein remains relatively constant throughout the cycle, indicated by the intensity of the amide II band at 1544 cm^{-1}. The band shape and peak position does not vary throughout the cycle and the main differences noted occur in the carbohydrate region, with the peak at 1047 cm^{-1} and the shoulder at 1120 cm^{-1}, both attributed to stretching vibrations of carbohydrate moieties in mucus, showing changes in intensity.

These changes however do not seem to be related to the cycle. Changes associated with endocervical cells would be thought to coincide with mucus production in the cervix. For the majority of the menstrual cycle mucus is thick and viscous however in the days leading up to, during and after ovulation, the mucus becomes less viscous to facilitate the passage of sperm into the uterus [19]. If changes occurring in cervical mucus were manifested in the spectra of endocervical cells, changes would be expected around the time of ovulation. Ovulation usually occurs at about day 14 of a 28 day cycle. The length of the cycle in Figure 4.9 was 30 days. If changes in the mucins affected the infrared spectra of endocervical cells the spectrum from day 14 should be different from those of the other days. As can be seen in Figure 4.9 this is not the case, with the spectra arising from days 14 and 28 showing very similar spectra. The changes observed may arise from differences in the concentration of mucus in the epithelial cells, but they may also be an averaging artifact from the two populations of endocervical cells in the cervix, secretory and ciliated.

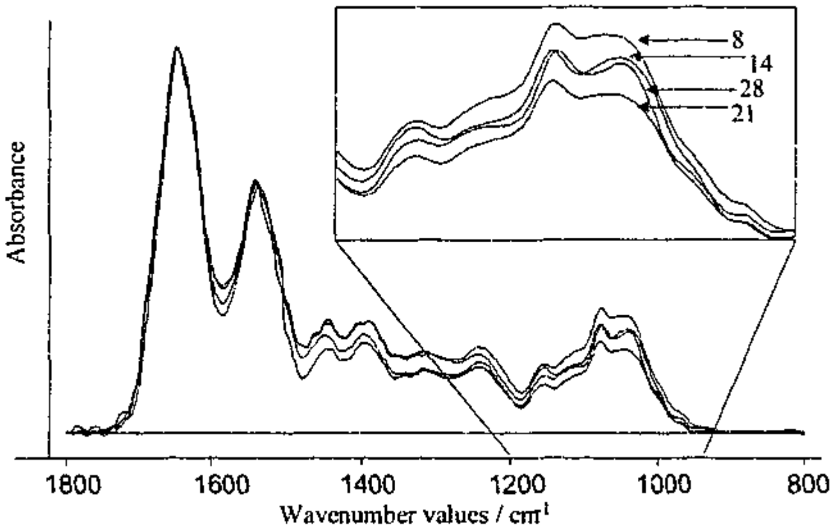

Figure 4.9 Infrared spectra of endocervical cells from a woman not taking oral contraception. These spectra were baseline corrected and normalised to the Amide I band.

Infrared spectra of endocervical cells from women taking monophasic contraception were of similar appearance to those obtained from women not taking oral contraception. This further supports the evidence that cyclical changes are not occurring in the spectra as a result of mucus because women taking monophasic contraception do not ovulate and so the consistency of the cervical mucus should remain constant throughout the cycle.

4.1.2.7 NUCLEIC ACIDS

As the nucleus of a superficial squamous cell becomes compact, so too does the nuclear material, which is believed to be too compact to absorb infrared radiation [8, 20]. Superficial cells appear in cervical smears as a result of estrogen stimulation and occur around ovulation as well as in the days leading up to menstruation [21]. The spectral effects of this phenomenon are seen by reductions in the bands associated with nucleic acids, namely the $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands arising from phosphodiester linkages at 1240 and 1080 cm^{-1} respectively. Figure 4.10 shows the phosphodiester region of ectocervical cell spectra from a woman not taking oral contraception. The spectrum resulting from a cervical smear obtained from day 15, where one expects a high proportion of superficial cells, exhibits a reduction in intensity of the $\nu_{as}PO_2^-$ compared to the spectra arising from days 8 and 22 where the proportion of intermediate cells is greater. Although this trend is reversed upon inspection of the $\nu_sPO_2^-$ band, this peak also has contributions from carbohydrate moieties and so the exact contribution from nucleic acids is unknown. Further evidence to support this theory is shown in Figure 4.11, where the spectra arising from day 36 also shows a reduction in the $\nu_{as}PO_2^-$ region.

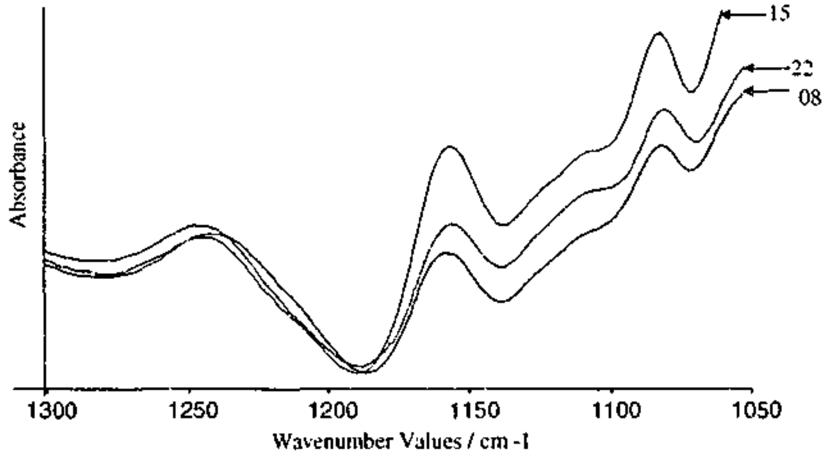

Figure 4.10 Infrared spectra of ectocervical cells obtained from a woman not taking oral contraception highlighting the phosphodiester region. The length of cycle was 29 days.

Even though there is a reduction seen in the $\nu_{as}PO_2^-$ bands with spectra associated with ovulation and premenstruation, the band does not disappear completely because phosphodiester linkages of RNA [20], which occur in the cytoplasm of cells, absorb radiation in this region.

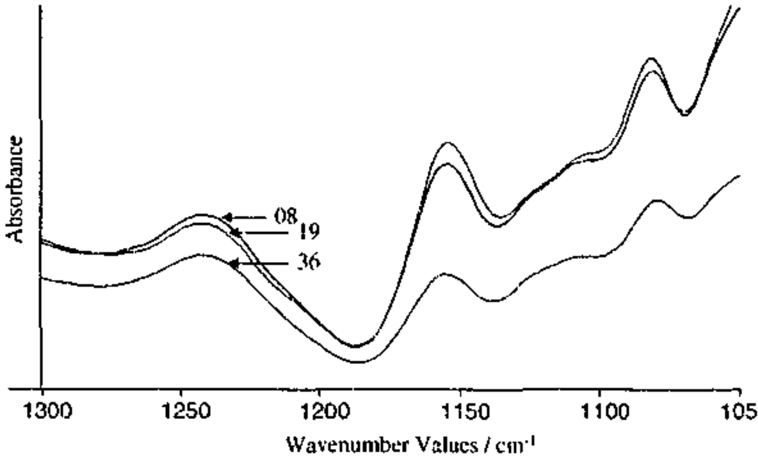

Figure 4.11 Infrared spectra of ectocervical cells obtained from a woman not taking oral contraception highlighting the phosphodiester region. The length of cycle was 36 days.

The consistency of the amide II peak intensities throughout the cycle, which indirectly show contributions from nucleic acids [8], may also be used as an indicator of nucleic acid changes. Chiriboga *et al.* [8] found that the intensity of the amide I/amide II ratio changed depending on the contributions of DNA. Morphologically, the N/C ratio changes throughout the cycle as seen with the appearance of pyknotic nuclei around ovulation. The actual concentration of the nuclear material in the cells does not vary rather it becomes more compact. If the protein and nucleic acid concentration of the cells were undergoing cyclic changes these would be seen as a difference in intensity of the amide II band.

Figure 4.12 shows the amide II bands of ectocervical cells, sampled from a woman not taking oral contraception. The intensity of these bands can be related to the contributions of DNA in cells. This is exhibited through changes in amide I/amide II ratio, as DNA exhibits a broad carbonyl stretch at the same frequency as the amide I peak arising from protein. It can be considered that DNA contributions would result in an increase in the amide I/amide II ratio, which in a spectrum normalised to the amide I band would result in an increase in the intensity of the amide II band with respect to the amide II band of spectra not showing DNA contributions. Therefore, the spectrum resulting from day 15 can be seen to have less contributions from nucleic acids than the spectra from days 8 and 22, indicating a higher presence of superficial cells with pyknotic nuclei.

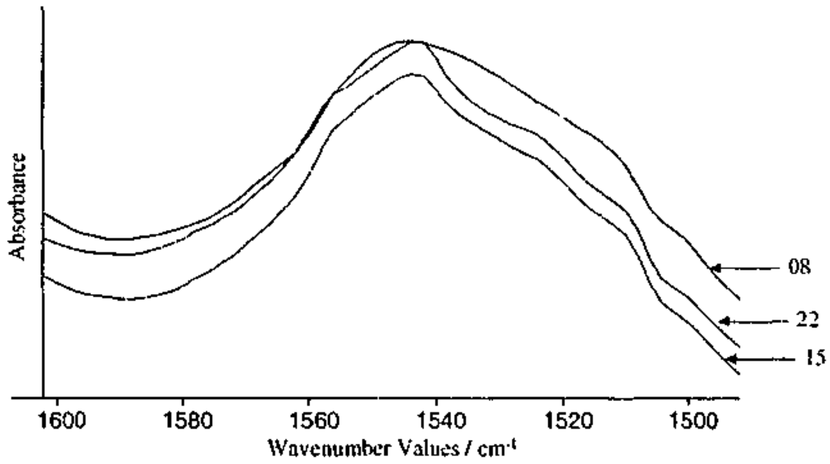

Figure 4.12 Infrared spectra obtained from a woman not taking oral contraception, highlighting the amide II band.

4.1.3 CONCLUSION

In conclusion it appears that there are definite changes occurring in ectocervical cells throughout the menstrual cycle. The main differences are seen in the carbohydrate region and can be attributed to the accumulation of glycogen in intermediate cells as a result of estrogenic stimulation. Ectocervical cell spectra obtained from women taking monophasic oral contraception did not exhibit the same degree of variation. Despite these cyclical differences, PCA was able to demonstrate that high-grade dysplasia could be separated from normal samples collected at different phases of the menstrual cycle. The infrared spectra of endocervical cells from both groups did not show cyclical variation. Contributions from nucleic acids, in particular DNA, also caused cyclical changes. Changes were manifested in a reduction of the $\nu_{as}PO_2^-$ and amide II bands from the infrared spectra of ectocervical cells obtained around ovulation and prior to menstruation.

4.2 AN INVESTIGATION INTO NUCLEIC ACID CONTRIBUTIONS IN THE IR SPECTRA OF NORMAL CERVICAL EPTITHELIAL AND HELA CELLS.

Given the complex nature of biological molecules and systems, it is often difficult to make band assignments based on the functional groups involved. Therefore it is necessary to obtain IR spectra of the constituent molecules and extrapolate these findings to IR spectra of biological molecules. The four main constituents of biological molecules are proteins, carbohydrates, lipids and nucleic acids. IR spectra of these macromolecules can be obtained by recording spectra of these constituents in isolation. To do this it is often necessary to isolate these components from whole cells by a technique known as subcellular fractionation.

Subcellular fractionation is essentially separating out the organelles of cells. The membrane, nucleus and other organelles can be isolated and purified allowing IR spectra to be obtained to determine which peaks in spectra are attributable to the various components of cells.

Cytologically the main differences in cell morphology of normal, dysplastic, malignant and benign cellular changes all involve changes in the nucleus, for example size (measured by N/C ratio) and shape (refer to Section 1.1.5) [15, 21-23]. Subcellular fractionation was utilised to extract the nuclei of whole cells in an effort to determine the contributions of nuclei, in particular nucleic acids to IR spectral differences seen between normal and other disease states, be it benign, precancerous or malignant.

The first step in subcellular fractionation is the formation of a cell homogenate by the rupture of the cell wall. This may be achieved by a number of methods including sonication, nitrogen cavitation or mechanical shearing. Generally, the homogenisation procedure should be able to produce at least 90% cell breakage [24].

A gaseous shear such as nitrogen cavitation involves the exposure of a cell suspension to nitrogen gas at about 800 psi (5516 kPa) at 4°C for about 15 minutes within a stainless steel pressure vessel. The suspension is then forced through a needle valve by the gas pressure, and cell rupture occurs by a combination of the sudden expansion of gas dissolved within the cytosol and the formation of bubbles of nitrogen gas in the medium.

Once the cell has been ruptured, centrifugation techniques employing density gradients or differential pelleting are used to recover the components. The concept of differential centrifugation, or separation of particles on the basis of size, utilises the principle that large particles settle faster than small ones. A cell homogenate can be centrifuged at progressively higher g-forces to produce pellets of partially purified particles. The nucleus, which is the largest and fastest sedimenting organelle, and the mitochondria can be obtained with reasonable purity by this method, although the pellet must be washed several times to remove trapped smaller particles. This may cause damage to and loss of the organelles [25].

Rate zonal centrifugation also uses the property of particle size to achieve separation of the organelles. The sample is placed on top of a continuous density gradient and the particles move down through the gradient as discrete zones depending on particle size. This technique is rarely employed in subcellular fractionation because of the restrictions imposed on sample size, which should be no more than 10% of the total gradient volume [25].

Organelles may also be separated on the basis of their buoyant density, through the use of pre-formed or self-generated density gradients. The sample is loaded at the bottom or middle of a discontinuous gradient or throughout a continuous gradient to minimise particle aggregation and artifactual banding and clumping of material [25]. The particles then either sediment or float upward to their isopycnic[13] point during centrifugation. It is important to maintain constant osmolarity throughout the gradient to preserve organelle integrity and structure. Exposure of cells and cell organelles to changing tonicity will alter their volume and hence density and may also affect viability.

It is therefore better to perform gradient centrifugation with a medium that is iso-osmotic with mammalian fluids. Sucrose has traditionally been very popular because it is widely available at a relatively low cost. The use of sucrose gradients however, has several disadvantages in terms of osmolarity and viscosity. The high osmolarity causes membrane-bound, osmotically sensitive particles to lose water and shrink, thus altering their buoyant densities. This can lead to co-banding of organelles, which under iso-osmotic conditions would have distinctive buoyant densities [26]. The high viscosity of sucrose solutions leads to low sedimentation rates and long periods at high centrifugation speeds.

Iodixanol (Nycomed Pharma AS, Oslo, Norway) or OptiPrep™ is an iso-osmotic density gradient medium that is resistant to bacterial degradation, non-toxic and not metabolised by mammalian cells. Iodixanol, Figure 4.13, is a non-ionic, dimeric hexaiodinated compound with a molecular weight of 1550, a density of 2.08 g/ml and a melting point between 221-244°C [27].

---

[13] The isopycnic point is the point at which the bouyant density is equal to the density of the medium.

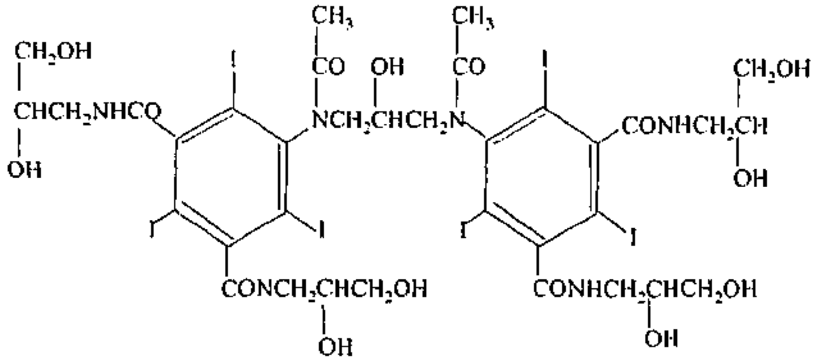

Figure 4.13 Structure of iodixanol, or 5,5'-[(2-hydroxy-1-3-propanediyl)-bis(acetylamino)] bis [N,N'-bis (2,3-dihydroxypropyl)-2,4,6-triiodo-1,3-benzenedicarboxamide]. Redrawn from [27].

The high density comes from the presence of two substituted triiodobenzene rings that are linked to a number of hydrophilic groups. Iodixanol is commercially available as a sterile 60% (w/v) aqueous solution with a density of 1.320 g/ml, an osmolarity of 260 mOsm, a refractive index of 1.4287 and does not contain buffers or other additives. The high solubility of iodixanol in water is attributed to the hydrophilic amide side chains and the hydroxylated carbon atoms between the two aromatic rings [27]. The physico-chemical properties of different concentrations of iodixanol are summarised in Table 4.2.

Iodinated density gradient media are able to form solutions dense enough to band subcellular organelles isopycnically without subjecting them to the damaging effects of high osmotic stress. The low osmolarity of these solutions maintains physiological osmolality and allows organelles to retain normal shape and volume, which results in better resolution of organelles [27].

Table 4.2 Physical properties of iodixanol in water (20°C) [27].

Iodixanol % (w/v)	**Molar concentration**	**Refractive Index (η)**	**Density, g/ml**	**Osmolarity, mOsm**
0	0	1.3330	0.998	0
10	0.0645	1.3490	1.052	38
20	0.1290	1.3649	1.105	80
30	0.1935	1.3809	1.159	115
40	0.2581	1.3968	1.213	150
50	0.3226	1.4128	1.266	200
60	0.3871	1.4287	1.320	260

Nitrogen cavitation and centrifugation using OptiPrep™ was performed on cervical epithelial cells and HeLa cells to isolate and obtain infrared spectra of nuclei.

4.2.1 METHODOLOGY

The cervical epithelial cells used in this experiment were obtained from the dysplasia clinic of the Royal Women's Hospital (Melbourne, Australia). Cultured HeLa cells were obtained from the Monash Medical Centre (Clayton, Australia). HeLa cells were removed from the culture flasks according to the protocol outlined in Appendix G.

4.2.1.1 PREPARATION OF HOMOGENATE

The preparation of a homogenate requires temperatures of less than 4°C. All operations were carried out at temperatures between 0-4°C, requiring the nitrogen pressure vessel[14] to be pre-cooled in ice. The cells were centrifuged at 1000 g for 10 minutes. and washed twice with phosphate buffered saline (PBS). After decantation of the supernatant the pellet was resuspended in PBS (at least 10 volumes PBS to 1 volume of cell pellet). Cells were resuspended by gentle inversion or the use of a Pasteur pipette, centrifuged at 1000 g for 10 minutes, and resuspended in a sucrose medium (solution D). Refer to Appendix G for details on preparation of solutions A to D. The suspension was cooled to 0°C and transferred to the stainless steel cavity of the nitrogen pressure vessel. Oxygen free nitrogen was introduced to a pressure of ~20 MPa and allowed to equilibrate for 20 minutes. With the delivery tube leading to a beaker, the delivery valve was opened slowly, ensuring that the tip of the delivery tube was kept above the level of the collected homogenate in the beaker. Once all the suspension had been collected the delivery valve was closed and any remaining gas vented. The homogenate was gently stirred to allow the foaming to subside.

The progress of the homogenisation was monitored by light microscopy to give an indication of the percentage cell rupture and to ensure that nuclear integrity had been maintained. A small amount of the homogenate was pipetted onto a microscope slide and stained with Ehrlich's Haemotoxylin[15]. The nuclei should appear dark red and unswollen.

4.2.1.2 ISOLATION OF NUCLEI[16]

Once the homogenate has been prepared, organelles such as the nucleus are isolated by means of centrifugation. Two gradient solutions of 30% (w/v) and 35% (w/v) iodixanol

---

[14] For more information regarding the nitrogen pressure vessel, or bomb, refer to Appendix H.

[15] Refer to Appendix G for the preparation of Ehrlich's Haemotoxylin.

[16] Methodology from [25]

were prepared by diluting solution C with solution D. (6 vol. C and 4 vol. D; 7 vol. C and 3 vol. D respectively). The entire homogenate or a crude nuclear pellet, obtained by centrifugation of the homogenate at 1000 g for 10 minutes and resuspended in solution D, was used. Equal volumes of the sample and solution C were mixed and 10-15 ml was transferred to a suitable centrifuge tube (40-50 ml) for a swinging bucket rotor of a high-speed centrifuge. The sample was underlayered with 10 ml of the 30% iodixanol and 5-10 ml of the 35% iodixanol and centrifuged at 10,000 g for 20 minutes. The nuclei banded at the 30/35% interface as illustrated in Figure 4.14.

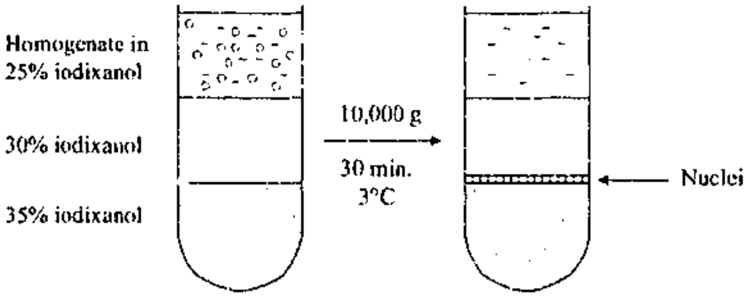

Figure 4.14 Schematic representation of the purification of nuclei from mammalian cells. Redrawn from [28].

Cellular material banding at the interface was carefully recovered with a Pasteur pipette and placed in an Eppendorf tube. The material was centrifuged at 1,000 g for 10 minutes and washed twice with 70% ethanol.

The relative centrifugal force (RCF) was determined using the following equation:

$$\mathrm{RCF} = 11.18 \times r \times \left(\frac{\mathrm{r.p.m.}}{1000}\right)^2$$ **Equation 4.1**

where r is the radius (cm) of the centrifuge, measured as the distance from the middle of the centrifuge to the bottom of the centrifuge tube (for a swinging bucket rotor) and the distance from the middle of the centrifuge to the middle of the top of the centrifuge (for a fixed rotor).

4.2.2 RESULTS AND DISCUSSION

A problem was initially encountered with performing separation of subcellular organelles because established protocols were predominantly for the homogenisation of liver and/or cultures of liver cells, rather than epithelial or HeLa cells. The first step in refining the protocol was to find out how the densities of organelles differ according to the type of cell and with respect to the gradient density medium used for the separation. One of the major advantages of OptiPrep™ is that it is able to maintain an iso-osmotic environment over all concentration and density ranges [27], meaning that the densities of organelles in a solution of OptiPrep™ will not be altered.

The receptacle chosen for delivery of homogenate from the nitrogen pressure vessel (a beaker) was inappropriate because the homogenate was expelled with very high force and the solution was sprayed out of the receiving beaker. A 50 ml centrifuge tube and lid were adapted to fit over the hose to minimise leakage[17]. Homogenisation is never 100% successful as there will be a certain proportion of whole cells that will not rupture. It is generally accepted that 90% rupture is adequate to perform organelle isolation [24].

A small sample of each homogenate produced was pipetted onto a glass microscope slide, stained and examined under a light microscope. The first few attempts at homogenisation were unsuccessful as only a small percentage of the cells were ruptured, and the nuclei from the ruptured cells still had cytoplasm attached. Higher pressures were then used and this resulted in a greater proportion of cell rupture and organelle liberation. HeLa cells posed a major problem for homogenisation, with cells either not rupturing, or rupturing and destroying the nuclei. It was thought that this problem might have been due to the cells being too old (they had been cultured over a few weeks and it was believed that the growth medium used was contaminated). Due to the difficulty with culturing enough HeLa cells for the process, it was decided that the protocol would be refined using epithelial cells from cervical smears. After experimenting with different equilibration times and pressures, the homogenisation rate increased to between 80-85%.

The first few attempts at gradient density centrifugation were also unsuccessful. Figure 4.14 shows a schematic representation of the centrifugation tubes before and after centrifugation. In an isotonic environment the nuclear pellet should have a density of 1.12 g/ml[18] and was expected to band at the 30/35% OptiPrep™ interface, but instead all the cellular material collected as a pellet at the bottom of the tube. Initially it was thought that the densities of the prepared solutions of OptiPrep™ were incorrect, so the refractive

---

[17] Refer to Appendix H for a schematic representation of the nitrogen pressure vessel.

[18] The homogenate was in a low concentration sucrose solution so it was expected that the density would increase slightly due to osmosis.

indices of the 30 and 35% OptiPrep™ solutions were calculated, and hence density calculated according to Equation 4.2 [27]:

$$\rho = 3.298\eta_{20^\circ C} - 3.396 \qquad \textbf{Equation 4.2}$$

The refractive indices of the 30% and 35% OptiPrep™ solutions were measured as 1.380 and 1.3889 respectively, which corresponds to densities of 1.155 g/ml (1.159 g/ml expected) and 1.184 g/ml (1.186 g/ml expected) respectively. Please refer to Table 4.2 for the physical properties of aqueous solutions of OptiPrep™. As the recorded density measurements were similar to that of the expected OptiPrep™ densities, it was thought that analysis of the pelleted cells by light microscopy might be useful in explaining the failure of the nuclei to band at the 30/35% interface.

Analysis by light microscopy revealed a mixture of whole cells, liberated nuclei and nuclei with cytoplasm still attached. The next homogenate was prepared and instead of only having two different densities of OptiPrep™, it was decided to use a 40% solution as well, to encourage the whole cells to settle at the 35/40% interface. Solutions were centrifuged at 10,000 g for 25-30 minutes, and two bands could be seen at the 30/35% and the 35/40% interfaces with the 35% solution containing a particulate suspension. These fractions were collected in Eppendorf tubes, centrifuged at 1,000 g for 10 minutes, and a small amount of cellular material fixed and stained for microscopic analysis. At the 30/35% interface, the majority of the cells had been ruptured as could be seen by the high proportion of liberated nuclei, but some cytoplasm was also present. The 35/40% interface had a high proportion of whole cells, a few had been ruptured but nuclei were still attached to the cytoplasm. The 35% suspension contained a mixture of whole cells, liberated nuclei and cytoplasm. The collected cellular material from the two interfaces as well from the 35% suspension were washed twice in 70% ethanol and prepared for infrared analysis. Figure 4.15 shows the spectra of nuclei and whole cells. The spectrum of the nuclei is the averaged result of three separate fractionation experiments and represents the cellular material at the 30/35% interface. Figure 4.16 shows the microscopic image of isolated nuclei from the 30/35% OptiPrep™ interface.

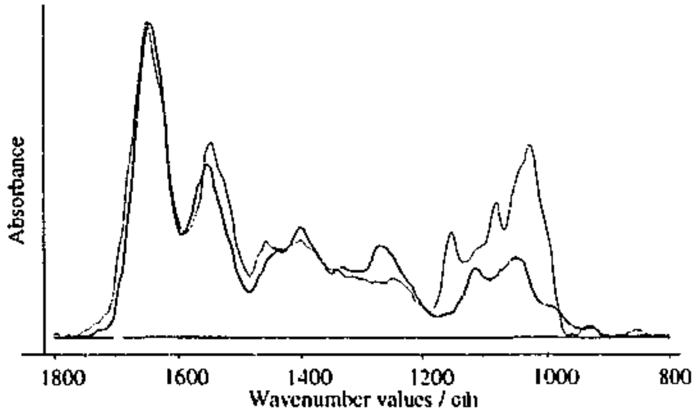

Figure 4.15 Infrared spectra of whole cells (red) and cellular material believed to be nuclei (blue) collected from the 30/35% OptiPrep™ interface.

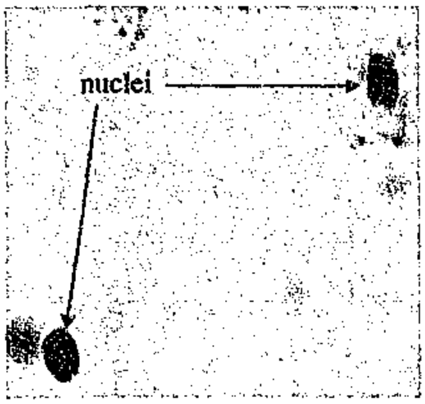

Figure 4.16 Microscopic image of isolated nuclei obtained from the 30/35% OptiPrep™ interface.

The infrared spectrum obtained from the 30/35% interface exhibits a shift in the amide I and II bands from 1650 to1646 cm^{-1} and 1544 to1550 cm^{-1} respectively compared to the whole cell. The spectrum obtained from isolated nuclei exhibits bands at 1432, 1400, 1334, 1270, 1114 and 1048 cm^{-1}. The nuclei spectrum lacks the 1240 and 1086 cm^{-1} bands arising from $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ of phosphodiester linkages in DNA respectively, reported by Wong *et al.* [1, 2]. The reported infrared spectra were obtained from homogenisation of a cultured colon adenocarcinoma cell line rather than cervical epithelial cells.

The bands at 1270 and 1114 cm^{-1} could be arising from $\nu_{as}PO_2^-$ and νC-O in RNA respectively, although it is unlikely that the expected $\nu_{as}PO_2^-$ band would have shifted from 1244 cm^{-1} [29]. Indications these bands are not due to RNA is further exhibited with the absence of a $\nu_sPO_2^-$ band at 1084 cm^{-1}. The presence of a carbohydrate peak at 1048 cm^{-1} may be due to residual glycogen from the small amount of cytoplasm not separated from the nuclei. Since food materials are stored in the cytoplasm of cells [30], carbohydrate bands were not expected in the IR spectrum of isolated nuclei.

Although nitrogen cavitation is highly reproducible, isolated nuclei can be very fragile [24]. Using cultured cells rather than tissue also presents problems in subcellular fractionation, especially if the density of the Optiprep™ solution is not correct. Cultured cells require higher shearing forces to cause cell disruption, or lysing. This increased pressure increases the chance of nuclear rupture. DNA released from even a few cells will result in severe aggregation of material [25].

It is possible that the absence of characteristic nucleic acid bands in the IR spectra of isolated nuclei were caused by DNA released by nuclear rupture and discarded with the supernatant. Light microscopy of the isolated nuclei however did not reveal aggregation, indicating that nuclear rupture had not occurred. The IR spectra presented in Figure 4.15 were obtained from ectocervical smears. Therefore the pressures required to cause cell rupture should be closer to those of cells collected from tissue samples. Furthermore the density of the OptiPrep™ solutions used were checked using refractive index measurements and found to be consistent with recommended densities [27]. If the nuclei have remained intact during this experiment, then there has to be another reason as to why contributions from nucleic acids, in particular DNA are not seen in the IR spectra of isolated nuclei.

Nucleic acids and proteins, known as polyelectrolytes and polyampholytes respectively, are classed as macroions and may carry a substantial charge depending on pH [31]. In solution, nucleic acid molecules repel each other and proteins are soluble at pH values above or below the *isoelectric point*. When positively and negatively charged macromolecules are mixed together electrostatic attraction results in molecular association. In the *chromosomes* of the nucleus of higher organisms the negatively charged DNA is strongly associated with the positively charged proteins, called *histones*, forming a complex called chromatin [31].

Eukaryotic cells contain an enormous amount of DNA. The *diploid* content of a human cell is about 8×10^9 base pairs, corresponding to a total length of nearly 3 metres. This DNA is packed into a nucleus about 10 μm in diameter [31], and the amount doubles as the cell undergoes mitosis. The optical density of the DNA-histone chromatin complex is thought to be too high to allow the transmission of IR radiation, hence phosphodiester bands arising from DNA are not always seen in IR spectra [20]. The study undertaken in the Diem laboratory [20] investigated the IR spectra of nuclei at different stages of the cell cycle in cultured myeloid leukemic (ML-1) cells and found an extremely low contribution of nucleic acid features in the IR spectra of ML-1 cells in the G1 and G2 phases, especially given that cells in the G2 phase are *tetraploid.* Furthermore, some of the spectral differences observed for the phases of the cell cycle resembled differences between normal and abnormal exfoliated cells.

The cell cycle will not affect the spectral features of normal cervical smears, as basal cells and reserve cells are the only epithelial cells to undergo mitosis and these are rarely seen in smears [21]. However the size of the nucleus and hence the degree of chromatin packing changes throughout the maturation of squamous epithelial cells. The diameter of the pyknotic nucleus associated with a mature superficial cell is 5-7 μm compared to a 9-11 μm diameter for the nucleus of an intermediate cell. Spectral differences in the phosphodiester region have been observed between the maturation stages of squamous epithelial cells [8, 32]. Isolated nuclei from superficial squamous cells of cervical smears would be expected to have less spectral contributions from DNA in the phosphodiester regions due to chromatin compactness than isolated nuclei from intermediate cells. This suggests that the epithelial cells in the cervical smears used for this experiment were predominantly of superficial maturation. The lack of RNA contributions may also be explained by the pyknotic nuclei of superficial cells. The highly compacted DNA in the small nucleus may increase the optical density of the entire nucleus, reducing spectral contributions from other macromolecules, including RNA.

Spectral differences observed between normal and neoplastic tissue have been linked to the optical density of the nucleus in these cells [33]. Increased N/C ratio and cell division associated with neoplasia may contribute to increased DNA features in IR spectra of neoplastic cells, brought about by decreased chromatin packing in larger nuclei. Benign cellular changes in squamous epithelial cells can also cause an increase in N/C ratio, the implications of this are investigated in Section 5.4.

A study investigating spectral differences between nucleated and anucleated cells undertaken in our laboratory also observed interesting changes in the phosphodiester region [34]. IR spectra of human red blood cells, which lack a nucleus, and chicken red blood cells, which contain a nucleus, were recorded. Difference spectroscopy revealed no significant differences between the two cell types in the $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands at 1240 and 1080 cm^{-1} respectively. Nuclei isolated from chicken red blood cells showed an increase in PO_2^- absorption at 1244 and 1080 cm^{-1} when collected in water and PBS.

Whilst the increased absorption was attributed to nuclei expansion, changes in electrostatic interaction between DNA and histones could also contribute to increased absorption. Dissociation of the chromatin complex in nuclei suspended in water may cause dispersion of DNA molecules throughout the nucleus. This would explain the increased intensity of PO_2^- bands. The reduction in PO_2^- band intensity in the IR spectra of nuclei suspended in PBS, relative to those in water may be arising from less dissociation of the chromatin complex. Although PBS is an isotonic solution, exposed nuclei can be fragile and swelling can occur. The use of PBS containing cations, as used in the study presented in this chapter, help to maintain nuclear integrity and reduce the possibility of swelling or rupture [24].

4.2.3 CONCLUSION

Subcellular fractionation via nitrogen cavitation was successful as confirmed by light microscopy of collected fractions. The recorded IR spectra of isolated nuclei did not correspond to those reported in the literature [1, 2], exhibited by a lack of nucleic acid contributions. It is possible that the extreme density of chromatin in the nucleus precludes the absorption of radiation by the nucleic acids present [20]. These findings would then support those of the Diem laboratory [8, 20], where the presence of nucleic acids in the nucleus of cervical squamous epithelial cells and myeloid leukemic cells respectively were not always detected using infrared spectroscopy. The study undertaken in our laboratory by Khurana [34] found no observable spectral differences between red blood cells with and without a nucleus. Phosphodiester contributions increased when nuclei were isolated, and the intensity of these bands depended on the solution used to rupture the red blood cells and collect the nuclei. Nuclear integrity was maintained in this study of isolated nuclei of cervical squamous epithelial cells. This was demonstrated through light microscopy of isolated nuclei. The refractive indices of the OptiPrep™ solutions used and the inclusion of PBS^+ maintained an isotonic environment for isolated nuclei. It is possible that the electrostatic interaction between DNA and histones plays a part in the contributions of nucleic acids to IR spectra.

4.3 REFERENCES

1. Wong, P., E. Papavassiliou, and B. Rigas, Phosphodiester stretching bands in the infrared spectra of human tissues and cultured cells. *Applied Spectroscopy*, 1991. **45**(9): p. 1563-1567.
2. Wong, P., *et al.*, Infrared spectroscopy of exfoliated human cervical cells: Evidence of extensive structural changes during carcinogenesis. *Proc. Natl. Acad. Sci.*, 1991. **88**: p. 10988-10992.
3. Yazdi, H., M. Bertrand, and P. Wong, Detecting structural changes at the molecular level with Fourier transform infrared spectroscopy. *Acta Cytologica*, 1996. **40**(4): p. 664-668.
4. Fung, M.F.K., *et al.*, Comparison of Fourier-transform infrared spectroscopic screening of exfoliated cervical cells with standard Papanicolaou screening. *Gynecologic Oncology*, 1997. **66**: p. 10-15.
5. Wood, B., *et al.*, FTIR microspectroscopic study of cell types and potential confounding variables in screening for cervical malignancies. *Biospectroscopy*, 1998. **4**(2): p. 75-91.
6. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. II. A comparative study of spectra of biopsies of cervical squamous epithelium and of exfoliated cervical cells. *Biospectroscopy*, 1998. **4**(1): p. 55-59.
7. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. III. Spectral differences between squamous and columnar tissue and cells from the human cervix. *Biospectroscopy*, 1997. **3**(4): p. 253-257.
8. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix. *Biospectroscopy*, 1998. **4**(1): p. 47-53.
9. Cohenford, M. and B. Rigas, Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: Implications for tumor biology. *Proc. Natl. Acad. Sci.*, 1998. **95**: p. 15327-15332.
10. Wood, B., *et al.*, An investigation into FTIR spectroscopy as a biodiagnostic tool for cervical cancer. *Biospectroscopy*, 1996. **2**: p. 1-11.

11. Cohenford, M., *et al.*, Infrared spectroscopy of normal and abnormal cervical smears: Evaluation by principal component analysis. *Gynecologic Oncology*, 1997. **66**: p. 59-65.

12. Romeo, M., *et al.*, Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. *Cellular and Molecular Biology*, 1998. **44**(1): p. 179-187.

13. Rogers, J., Physiology of menstruation, in *Endocrine and Metabolic Aspects of Gynecology*. W B Saunders: London. 1963.

14. Szarewski, A., *et al.*, Effect of smoking cessation on cervical lesion size. *The Lancet*, 1996. **347**: p. 941-943.

15. DeMay, R., *The Art and Science of Cytopathology*. Vol. 1. Chicago: American Society of Clinical Pathology Press. 1995.

16. Hafez, E.S.E., *Structural and ultrastructural parameters of the uterine cervix*. Obstetrical and Gynecological Survey, 1982. **37**(8): p. 507-516.

17. Campbell, N.A., Animal Reproduction, in *Biology*. Benjamin/Cummings Publishing Co. Inc.: Sydney. p. 930-954. 1990.

18. Gibbons, R.A. and P. Mattner, Some aspects of the chemistry of cervical mucus. *International Journal of Fertility*, 1966. **11**(4): p. 366-372.

19. Davajan, V., R. Nakamura, and K. Kharma, Spermatozoan transport in cervical mucus. *Obstet. and Gynecol. Surv.*, 1970. **25**: p. 1-43.

20. Boydston-White, S., *et al.*, Infrared spectroscopy of human tissue. V. Infrared spectroscopic studies of Myeloid Leukemia (ML-1) cells at different phases of the cell cycle. *Biospectroscopy*, 1999. **5**: p. 219-227.

21. Riotton, G. and W. Christopherson, *Cytology of the Female Genital Tract*. Vol. 8. Geneva: World Health Organisation. 1973.

22. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 2 ed. Philadelphia: J B Lippincott Company. 1968.

23. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 3 ed. Philadelphia: J B Lippincott Company. 1979.

24. Graham, J.M., *Homogenisation of tissues and cells*, in *Subcellular Fractionation: A Practical Approach*. Oxford University Press: New York. p. 1-29. 1997.

25. *Isolation of cell organelles*. 1998. http://www.nycomed-diagnostics.com/gradmed/cellorg/

26. Wattiaux, R., *et al.*, Isolation of rat liver lysosomes by isopycnic centrifugation in a metrizamide gradient. *Journal of Cell Biology*, 1978. **78**: p. 349-368.

27. Ford, T., J. Graham, and D. Rickwood, Iodixanol: a nonionic iso-osmotic centrifugation medium for the formation of self-generated gradients. *Analytical Biochemistry*, 1994. **220**: p. 360-366.

28. Ford, T., J. Graham, and D. Rickwood, The preparation of subcellular organelles from mouse liver in self-generated gradients of iodixanol. *Analytical Biochemistry*, 1994. **220**: p. 367-373.

29. Benedetti, E., *et al.*, Determination of relative amount of nucleic acids and proteins in leukemic and normal lymphocytes by mean of Fourier transform infrared microspectroscopy. *Applied Spectroscopy*, 1997. **51**: p. 792.

30. Roberts, M.B.V., *Structure and Function of Cells*, in *Biology. A Functional Approach*. Thomas Nelson and Sons Ltd: Edinburgh. p. 13-31. 1982.

31. Mathews, C.K. and K.E.van Holde, *Biochemistry*. CA: The Benjamins/Cummings Publishing Company. 1990.

32. Chiriboga, L., H. Yee, and M. Diem, Infrared spectroscopy of human cells and tissue. Part VI: A comparative study of histopathology and infrared microspectroscopy of normal, cirrhotic and cancerous liver tissue. *Applied Spectroscopy*, 2000. **54**(1): p. 1-8.

33. Chiriboga, L., H. Yee, and M. Diem, Infrared spectroscopy of human cells and tissue. Part VII: FT-IR microspectroscopy of DNase- and RNase-treated normal, cirrhotic and neoplastic liver tissue. *Applied Spectroscopy*, 2000. **54**(4): p. 480-485.

34. Khurana, J.S., The analysis of blood components using Fourier transform infrared spectroscopy. *Honours Thesis*, in Chemistry. Monash University: Australia. 1999.

CHAPTER 5

A DIAGNOSTIC TECHNIQUE FOR CERVICAL CANCER

5 A DIAGNOSTIC TECHNIQUE FOR CERVICAL CANCER

Multivariate statistics is an important tool for spectroscopists as it enables the objective analysis of data and the extraction of information in the data not possible from visual inspection. The term multivariate statistics encompasses a broad range of techniques, but the methods used in this analysis are limited to those described in Section 2.5. The purpose of the preliminary analysis was to ensure that discrimination between normal and abnormal (high-grade dysplastic and malignant) samples could be achieved before proceeding with collection of vast amounts of data.

The results of the final statistical analysis employing Bayesian regularised artificial neural networks to train and predict the data collected during the course of this research is presented and discussed in Section 5.2.

5.1 PRELIMINARY STATISTICAL ANALYSIS

5.1.1 METHODOLOGY

5.1.1.1 SAMPLE COLLECTION

Cervical cells used in the database were obtained from the Royal Women's Hospital Dysplasia Clinic (Melbourne, Australia) and from Family Planning Victoria (Melbourne, Australia). Samples were initially collected from all patients attending the Dysplasia Clinic. However, given the inaccuracies of the Pap smear it was decided that samples would only be collected from patients undergoing biopsy. Patients who attended the Dysplasia Clinic through referral[19] or existing patients with an abnormal Pap smear had biopsies. Therefore, the spectra collected that have associated biopsy results are sampled from an abnormal population and are not representative of the normal population.

Samples from patients attending the FPV Clinic were collected to ensure sampling of a normal population. Unfortunately only patients with Pap smears that are classified as abnormal are required to have a biopsy, a procedure not undertaken at this clinic. Therefore none of the FPV samples have a histological diagnosis although statistically 90% of the samples would be expected to be normal [1].

---

[19] Patients attending other clinics who have abnormal Pap smears are often referred to the dysplasia clinic for a Pap smear and biopsy.

Cervical cells were collected from the transformation zone of the cervix with an Ayre spatula and a Cytobrush™ to obtain ectocervical and endocervical cells respectively. Sampling instruments were agitated in the same 50 cm^3 centrifuge tube containing 10 cm^3 ethanol to collect the cellular material. The samples used to record IR spectra therefore comprised mixed populations of ectocervical and endocervical cells.

5.1.1.2 SAMPLE PREPARATION

Centrifuge tubes containing cervical cells were centrifuged at 2500 r.p.m. for 10 minutes. The ethanol supernatant was removed with an automated pipette leaving a cellular pellet. The cellular material was pipetted into the KRS-5 multicavity cell (Figure 4.1, Section 4.1.1.4) and desiccated under vacuum.

5.1.1.3 INFRARED SPECTROSCOPY

Following removal of the aluminium plates used to form sample wells, the KRS-5 infrared substrate, containing fourteen cervical samples was placed on the sampling platform of a Bruker IFS-55 infrared microscope system. Fifty scans were co-added at a resolution of 8 cm^{-1}, sampling 766 data points between 3648.788 and 698.129 cm^{-1}. A minimum of six transmission spectra were recorded for each sample, unless samples exhibited gross spatula contamination or there was not enough cellular material to record spectra with an acceptable signal to noise ratio.

5.1.1.4 DATA TREATMENT

The data points of the IR spectra recorded on the Bruker spectrometer were not integer values. To overcome this, the 'make compatible'[20] option in OPUS was utilised. Using a spectrum recorded on the PE spectrometer (1476 data points between 3650 and 700 cm^{-1}) as a comparison, interpolation was used to convert the spectra to 1475 data points between 3648 and 700 cm^{-1}.

Spectra were baseline corrected, averaged, normalised to the amide I band (using maximum normalisation (Section 3.2.5.2) and converted into JCAMP.dx format to enable importing into Unscrambler and Matlab for analysis.

---

[20] Refer to Appendix I for an explanation of the make compatible function.

5.1.1.5 SOFTWARE

The statistical analyses carried out in this chapter were undertaken using Unscrambler (CAMO, Oslo), Propagator (ARD, Columbia, USA) and Matlab (The Mathworks, Inc, MA, USA) software. Unscrambler was used for spectral averaging and visualisation of data, and was also employed for PCA and SIMCA. Propagator, a commercial artificial neural network package was used to perform initial neural network calculations. *K*-nearest neighbour, discriminant analysis, ANN and BRANN calculations were performed in Matlab, using existing and purpose written programs.

5.1.2 RESULTS AND DISCUSSION

Infrared spectra used in the following analyses were chosen because both the cytological and histological results were in agreement. Normal samples were included if they were diagnosed as negative without inflammation, bacterial infections, metaplasia or BCCs. Abnormal samples were included if they were diagnosed with high-grade abnormalities (CIN II, CIN III, CIS, SCC or invasion) with and without HPV effects.

Following this selection process, spectra were further subjected to visual processing and grouped according to spectral characteristics described by Wong *et al.* [2, 3], illustrated in Figure 5.1. Figure 5.1 compares averaged IR spectra of cervical cells diagnosed as normal with those exhibiting various grades of abnormality ranging from CIN I to CIS. It is now widely accepted that these spectra are not representative of the infrared spectra seen in routine spectroscopy of cervical smears [4, 5]. However the preliminary analyses were performed throughout the candidature, and the spectra chosen for inclusion in the analysis reflect the findings reported in the literature at the time [3, 6].

Replicate spectra of each sample were averaged according to the criteria discussed above. Spectra exhibiting features of contamination, Section 5.1.2.1, were also removed. In total, 211 normal samples and 181 abnormal samples were used in the preliminary analysis.

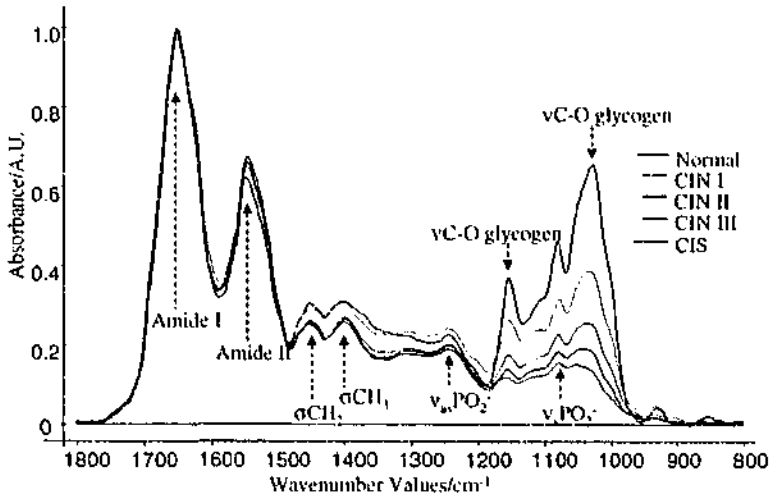

Figure 5.1 Averaged infrared spectra of cervical cells exhibiting various grades of abnormalities compared with normal cervical cells.

5.1.2.1 CONTAMINATION

The problem of spectral contamination by spatula debris discussed in Section 4.1.2.2. also presented a problem in the initial stages of data collection. Ethanol appeared to extract a residue from the wooden spatulas, which were initially kept stored in centrifuge tubes containing 70% ethanol to maximise cell collection. The use of KBr (potassium bromide) infrared substrates was trialed due to the expense and toxicity of KRS-5 substrates. KBr is very hygroscopic necessitating the use of absolute ethanol as a cell collection medium. The higher concentration of ethanol enhanced the extraction of spatula residue and an increase in spatula contamination was noted. Plastic spatulas were introduced to the RWH and FPV to minimise the loss of data due to contamination. Compliance was poor due to the fact that those using them complained the plastic spatulas were too flexible and caused patient discomfort. Wooden spatulas were re-introduced and cells were collected in 70% ethanol with KRS-5 as an infrared substrate. To minimise the effects of contamination, the spatulas were only briefly agitated in the ethanol solution and then discarded. Whilst there was still a small percentage of contaminated samples, these could easily be identified due to the characteristic bands discussed in Section 4.1.2.2.

5.1.2.2 PRINCIPAL COMPONENT ANALYSIS

PCA was utilised to identify the key wavenumber values accounting for the majority of the variance in the infrared spectra and thus reduce the number of variables. Each spectrum had 501 data points (1800 - 800 cm^{-1}) and with the aid of PCA, spectra were reduced to 7 data points. Loadings plots, Figure 5.2, were analysed to determine which wavenumber values were responsible for the discrimination. Wavenumber values with a loading higher than 0.1 were chosen. The resulting 7 wavenumber values (1620 – 1614 cm^{-1} and 1028 – 1024 cm^{-1}) are indicated by boxes marked (a) and (b) respectively. The box marked (c) corresponds to the amide I region. This region was excluded because the variance seen is most likely a result of normalisation, as the amide I band was the peak chosen to normalise to.

PCA formed the basis for each analytical technique either as a means of data reduction or to form models of the classification groups. Figure 5.3A illustrates the separation between the infrared spectra of normal and high-grade dysplastic and CIS diagnosed cervical cells using all 501 wavenumber values, whilst Figure 5.3B illustrates the separation when the number of variables or wavenumber values was reduced to 7. Variable reduction corresponded to wavenumber values in the range 1620 – 1614 cm^{-1} and 1028 – 1024 cm^{-1} inclusive. Reducing the number of variables resulted in tighter clusters of the two groups, and a slight increase in discrimination.

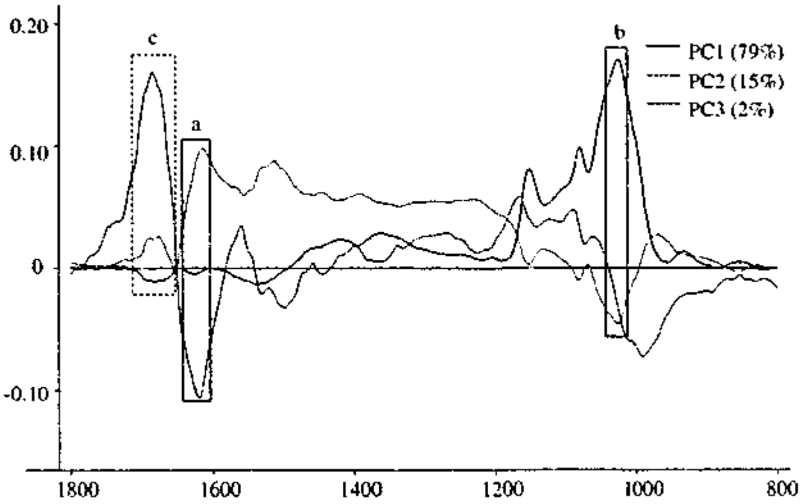

Figure 5.2 Loadings plots of the first three principal components, illustrating the variance of 501 wavenumber values. Seven variables with a loading greater than 0.1, shown by boxes marked (a) and (b) were chosen for a second PCA. The variables contained in box (c) were not included because this region was used for normalisation.

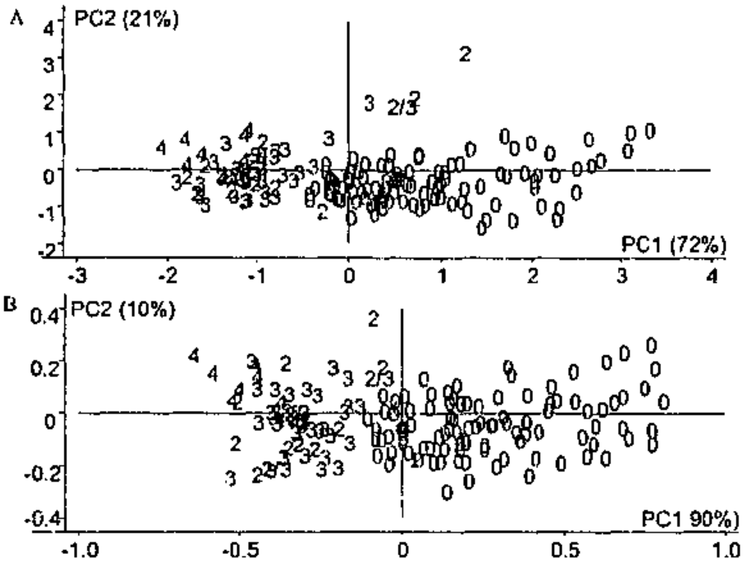

Figure 5.3A Scores plot illustrating the discrimination obtained between IR spectra of normal (0) and abnormal (2: CIN II, 2/3: CIN II-III, 3: CIN III and 4: CIS) cervical cells using 501 variables. B. Number of variables reduced to 7, note the tighter clustering and a slight increase in separation between the normal and abnormal samples.

5.1.2.3 K-NEAREST NEIGHBOURS

PCA coupled to *K*-nearest neighbours formed the next multivariate technique. Spectra were grouped into classes of normal and abnormal according to histological diagnosis, and PCA was performed. The scores of the principal components represent each sample, and values of *K* ranging from 1 – 9 were investigated in terms of sensitivity and specificity. The theory of *K*-nearest neighbours was discussed in Section 2.5.3, but to recap, the Euclidean or Mahalanobis distance between each unknown sample and members in the training set is calculated, and the unknown object is classified according to the majority of *K* nearest-neighbours to which it belongs. This can be visualised as the formation of a sphere around the unknown object, with the distances of the *K* nearest-neighbours acting as the boundary of the sphere. The object is assigned to the group with the most objects in the sphere.

K-NN was performed on normalised data with 501 wavenumber values, second derivative normalised data with 501 wavenumber values, and normalised data with 7 wavenumber values. The results of the analysis are summarised in Table 5.1. The spectrum of each sample included in the *K*-NN analysis was selected according to the criteria outlined in Section 5.1.2. The best results, in terms of the highest sensitivity and specificity were obtained using a *K* value of 3 (with normalisation and 501 wavenumber values) and 1 (with normalisation and 7 wavenumber values). Analysis using the second derivatives of infrared spectra resulted in lower sensitivity and specificity.

Table 5.1 Results of *K*-NN analysis with *K* values ranging from *K* = 1 to *K* = 9.

K	Pre-processing[a]	Variables	PCs (variance)	Sensitivity[b] (%)	Specificity[c] (%)
1	MN	501	4 (98%)	95	98
3	MN	501	4 (98%)	95	99
5	MN	501	4 (98%)	94	99
7	MN	501	4 (98%)	95	99
9	MN	501	4 (98%)	95	99
1	MN, 2Dv	501	4 (81%)	94	91
3	MN, 2Dv	501	4 (81%)	93	92
5	MN, 2Dv	501	4 (81%)	92	92
7	MN, 2Dv	501	4 (81%)	92	90
9	MN, 2Dv	501	4 (81%)	91	88
1	MN	7	4 (100%)	98	98
3	MN	7	4 (100%)	98	98
5	MN	7	4 (100%)	98	98
7	MN	7	4 (100%)	99	97
9	MN	7	4 (100%)	98	96

[a] MN represents spectra normalised to the amide I peak using maximum normalisation and 2Dv represent second derivative spectra obtained by employing the Savitzky-Golay algorithm.

[b] $$\text{sensitivity} = \frac{\text{number of samples predicted as abnormal}}{\text{number of abnormal samples}}$$

[c] $$\text{specificity} = \frac{\text{number of samples predicted as normal}}{\text{number of normal samples}}$$

Whilst the results of the *K*-NN analysis are promising in terms of sensitivity and specificity, there are some disadvantages to using hard modelling techniques such as *K*-NN. The most obvious disadvantages are that classification of a new object requires re-calculation of all distances, and addition of a new class involves recomputing *K*-NN criteria. *K*-NNs are sensitive to unequal numbers of objects in a training set, and classification results can differ depending on *K*.

5.1.2.4 SOFT INDEPENDENT MODELLING OF CLASS ANALOGY

Individual PC models were formed for the normal and abnormal groups of IR spectra. One hundred normal samples and 60 abnormal samples (13 CIN II, 2 CIN II-III, 37 CIN III and 8 CIS/SCC) made up the data for modelling. Once models had been formed, classification was performed on 36 normal samples and 29 abnormal samples (12 CIN II, 4 CIN II-III and 13 CIN III). Initial classification was undertaken using 501 variables, and 7 variables. However, when the data was randomised the results for both 501 and 7 variables changed. This was thought to be due to the different populations of samples making up training and classification groups. If the data used in the models are representative of the parent population, then the model will be able to account for the variability inherent in these samples. If the models are not representative of the parent population, then the model may not have strong modelling and discrimination powers and will be unable to classify unknown samples with a high degree of accuracy. This is illustrated in the results of SIMCA analysis on randomised and non-randomised data.

The results of the SIMCA analysis are summarised in Table 5.2 and illustrate the differences in sensitivity and specificity depending on the number of wavenumber values and randomisation of data.

Table 5.2 Results of SIMCA analysis

Class Parameters	**Sensitivity (%)**	**Specificity (%)**
501 wavenumber values	90	60
7 wavenumber values	85	92
501 wavenumber values, randomised data	56	96
7 wavenumber values, randomised data	85	92

Randomising the data improved the sensitivity and specificity of the data reduced to 7 wavenumber values and the specificity of the data using 501 wavenumber values. The sensitivity of the data using 501 wavenumber values was reduced dramatically with randomisation.

SIMCA models were created including spectra of samples diagnosed with CIN I and CIN I-II. A reduction in sensitivity was noted when CIN I and CIN I-II samples were included. Second derivatives of the spectra were calculated and SIMCA repeated. No improvement in the specificity or sensitivity was noted.

Figure 5.4 represents the results of the SIMCA analysis expressed in Coomans[21] plots, which give an indication of sample-to-model distances and class membership [7]. Despite the fact that the PCA scores plot in Figure 5.3 showed a separation between normal and abnormal samples, and the distance between the two models was 2.4 for 501 variables and 5 for 7 variables, SIMCA appears unable to achieve a good separation between the two models, with the majority of samples from the abnormal model, and a high proportion of samples from the normal model belonging to both models (indicated by the presence of these samples in the bottom left quadrant of the Coomans plot).

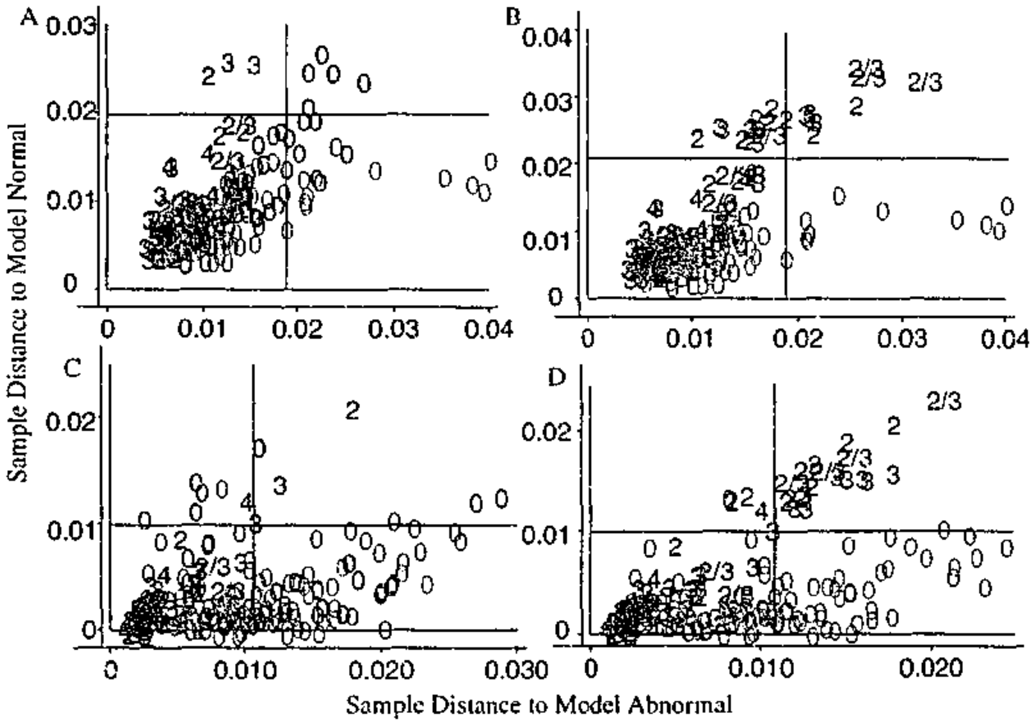

Figure 5.4 Coomans plots resulting from classification of unknown (green) samples based on models formed from normal (pink) and abnormal (blue) data. A. and B. represent classification results of normal and abnormal data respectively using 501 wavenumber values. C. and D. represent the same samples with reduced variables.

The modelling and discriminating power of the variables were examined as a means of eliminating redundant variables to improve the models. Variables with a discriminating power greater than 2 and variables with a modelling power greater than 0.5 were included and new PC models of the data were formed. No increase in specificity or sensitivity was noted. Outliers were removed from the PC models and new models calculated. Classification based on these new models performed poorly, with a resulting decrease in sensitivity and specificity. This result further highlights the need to have a broad spectrum

---

[21] Refer to Appendix K for an explanation of the interpretation of these plots and the relevance of the results obtained.

of data with which to create models for SIMCA analysis. The SIMCA models could possibly be improved by including more samples in each class. The robustness of the models and their separation would be improved, enhancing the prediction capabilities.

If the SIMCA models could be improved, this soft modelling technique possesses benefits as an analytical tool. First of all, it is possible to add new classes without changing the overall model, this arises from the formation of individual PCA models for each class. Secondly, objects are not forced into discrete classes allowing the detection of outliers.

The importance of ensuring all data is pre-processed in an identical manner is illustrated in Figure 5.5. Normal and abnormal data were collected separately and pre-processed using different baseline correction algorithms. PCA models were formed and the separation and classification results from differences in pre-processing methods rather than from spectroscopic differences between the two groups.

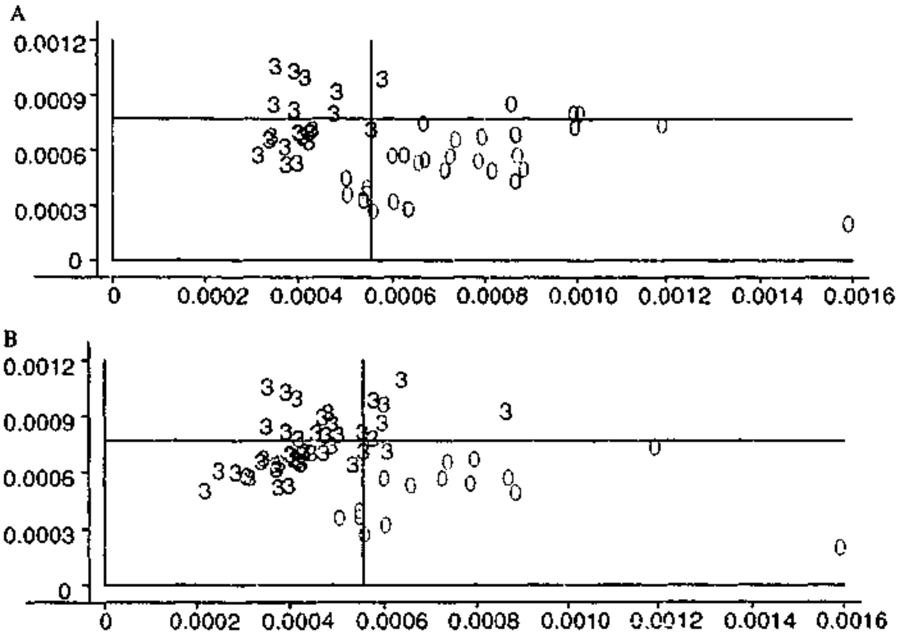

Figure 5.5 Coomans plot resulting from classification of unknown (green) samples based on models formed from normal (pink) and abnormal (blue) data. A. and B. represent classification results of normal and abnormal data respectively using 7 wavenumber values. The normal and abnormal diagnosed samples were pre-processed separately.

5.1.2.5 *ARTIFICIAL NEURAL NETWORKS*

Feed-forward, fully connected, back propagation neural networks were trained on the infrared spectra of cervical samples, again using all 501 wavenumber values and then the 7 wavenumber values identified by PCA. Different architectures were trialed, by changing network parameters such as the number of nodes and layers, the number of training cycles (epochs), learning rate and momentum factor.

The type of abnormal samples included for training and testing were also varied. Abnormal samples included high-grade dysplastic and malignant samples. When samples with low-grade dysplasia were included in the training set, the sensitivity was found to decrease as shown in Table 5.3. This correlates well with our previous study using neural networks in which samples exhibiting CIN I were predicted by the network as intermediate between normal and abnormal [8].

Table 5.3 Results of Neural Network Analysis

Neural Network Classes	Sensitivity (%)	Specificity (%)
Abnormal, including CIN I and CIN I-II	77	100
Abnormal, excluding CIN I	81	100
Abnormal, excluding CIN I and CIN I-II	84	100
BRANN (pre-selected data)	100	100

Whilst neural networks look like a promising technique for the diagnosis of cervical cancer, the data used for training and testing were subjectively chosen based on the spectral profiles seen in Figure 5.1 and discussed in Section 5.1.2. Although subjectivity is undesirable for a technique which aims to replace human judgement, it is important when forming data sets for training in pattern recognition that the data is representative of, in this case, the disease state [9]. Otherwise the network will perform poorly in prediction.

The network with the best performance (BRANN), as given in Table 5.3 had a standard error of prediction (SEP) of 0.0065 and a correlation of 0.99, which corresponded to 100% sensitivity and 100% specificity using pre-selected data.

5.1.2.6 LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS

Discriminant analysis, both linear and quadratic, was performed on the cervical data. Once again spectra were chosen because there was agreement in histology and cytology diagnosis, and because the infrared spectra were visually representative of the disease state given by the diagnosis. The results of discriminant analysis of pre-selected data are illustrated in Figure 5.6. A second discriminant analysis was performed including the data that had previously been selected out of the training set, refer to Section 5.1.2, due to the presence of confounding variables and the results are illustrated in Figure 5.7. Table 5.4 compares the sensitivity and specificity from linear and quadratic discriminant analysis with the average sensitivity and specificity reported for Pap smears in the literature as well as from our databank. Although the sensitivity and specificity of both discriminant analysis techniques has been reduced, the results are still promising in the sense that there is an improvement in sensitivity compared with the Pap smear.

Table 5.4 Results of discriminant analysis compared with values of the Pap smear reported in the literature and calculated from our study.

Technique	**Sensitivity (%)**	**Specificity (%)**
LDA (pre-selected data)	100	97
LDA	89	68
QDA (pre-selected data)	98	99
QDA	91	63
Pap smear [10-14]	80	68
Pap smear (from our study)[a]	59	79

[a] Two hundred samples with histology and cytology results were randomly chosen from the databank. Sensitivity and specificity were calculated for the Pap smear using histology as the gold standard.

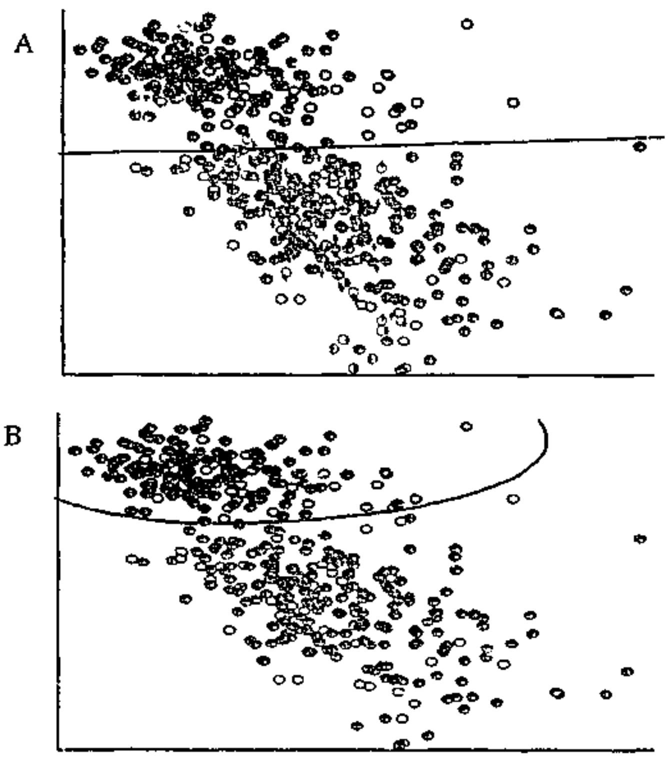

Figure 5.6 Results of linear discriminant analysis (A) and quadratic discriminant analysis (B), where +, +, o, o, o, and o represent normal original data, abnormal original data, normal classified data, abnormal classified data, normal predicted data and abnormal predicted data respectively. A red cross in a blue circle represents a sample misclassified as abnormal, and a blue cross in a red circle represents a sample misclassified as normal.

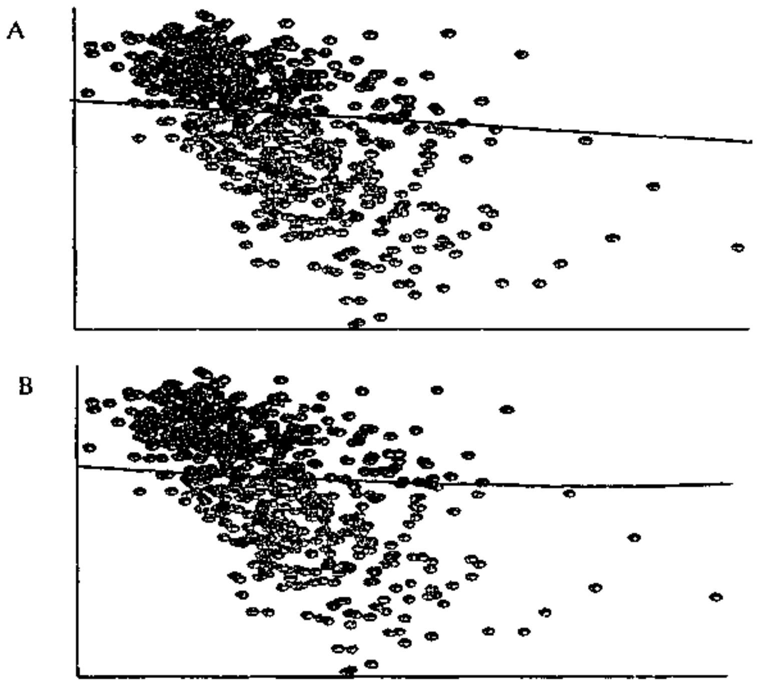

Figure 5.7 Results of linear (A) and quadratic (B) discriminant analysis using non-subjectively chosen data, where +, +, o, and o, represent normal original data, abnormal original data, normal classified data and abnormal classified data respectively. A red cross in a blue circle represents a sample misclassified as abnormal, and a blue cross in a red circle represents a sample misclassified as normal.

5.1.3 CONCLUSION

Multivariate statistical techniques were investigated to determine which method is best able to classify and predict normal and abnormal diagnosed cervical cells. *K*-NNs, SIMCA, ANNs (including BRANNs) and linear and quadratic discriminant analysis gave sensitivities and specificities ranging between 80-100%. These results were based on data pre-selected due to cytological and histological agreement. Spectra exhibiting signs of mild dysplasia or HPV were excluded as were spectra showing obvious signs of contamination or the presence of confounding variables.

Whilst the pre-selection processes employed in this investigation do not remove the subjectivity of the Pap smear, or rather introduce a new form of subjectivity by means of visual inspection of the spectra, the processes used reflect the feeling and attitudes in the literature at the time. It is likely that the values for sensitivity and specificity reported above indicate the best that could be achieved if cervical smears could be "cleaned up" in some way, i.e. if there was a way of removing the effects of non-diagnostic debris such as blood components and mucins. This is investigated further in Section 5.3.

5.2 FINAL STATISTICAL ANALYSIS

5.2.1 METHODOLOGY

The methodology used for sample collection, preparation and recording of infrared spectra was described in Section 5.1.1. Infrared spectra were made compatible in OPUS, refer to Section 5.1.1.4, and converted into JCAMP format for importing into Matlab where pre-processing and analysis by BRANNs could be undertaken. Once IR spectra had been imported into Matlab, spectra with a maximum absorbance greater than unity were discarded, as were spectra with a signal-to-noise ratio less than 10. Spectra were baseline corrected and reduced to the wavenumber region 1800 – 800 cm^{-1}. Other pre-processing techniques used are discussed in Section 5.2.2 below.

5.2.1.1 DATABASE

By the end of this candidature, an extensive database of infrared spectra with either cytological and/or histological diagnosis had been obtained. There were approximately 4700 individual samples collected, each with replicate spectra. The total number of samples collected from the RWH was approximately 2800, nearly one thousand of these had both cytology and histology results. About 900 samples were collected from FPV, all of which had a cytology result. Eighteen percent of spectra were discarded due to nonlinearity effects, one percent of spectra had a SNR of less than 10, and a further ten percent of spectra were discarded due to spatula contamination. A further twenty-four percent of spectra were excluded from analysis due to a diagnosis other than negative or high-grade dysplasia and malignancy, i.e. negative with BCC (3%), and CIN I/HPV (21%) effects. Fifty-four percent of the samples used for the final analysis were diagnosed negative and twenty-two percent were diagnosed with high-grade dysplasia.

5.2.2 RESULTS AND DISCUSSION

5.2.2.1 PRINCIPAL COMPONENT ANALYSIS

Given that biopsy is considered the gold standard in diagnosis of cervical cancer, it was decided that histology results would form the basis of diagnosis for IR spectra of cervical cells. Due to the inhomogeneity of many of the samples, indicated by different spectral profiles obtained from the same sample, it was initially decided that replicate spectra for each sample would not be averaged, and the analysis would proceed using all spectra, with multiple spectra representing one sample. A principal component scores plot of the data, Figure 5.8, exhibits a seemingly random distribution of normal and abnormal cells, with a few samples forming clusters along the PC1 axis. The normal component of the data comprised samples with a histological diagnosis of normal, excluding samples exhibiting inflammation, bacterial infection, HPV effects and benign cellular changes. The abnormal component comprised samples diagnosed as high-grade dysplasia, CIS and SCC, with and without HPV effects. Samples exhibiting mild dysplasia were excluded from the data. This was to determine if high-grade and malignant abnormalities could be discriminated from normal samples, given that cytologically the greatest differences exist between these disease states and normal.

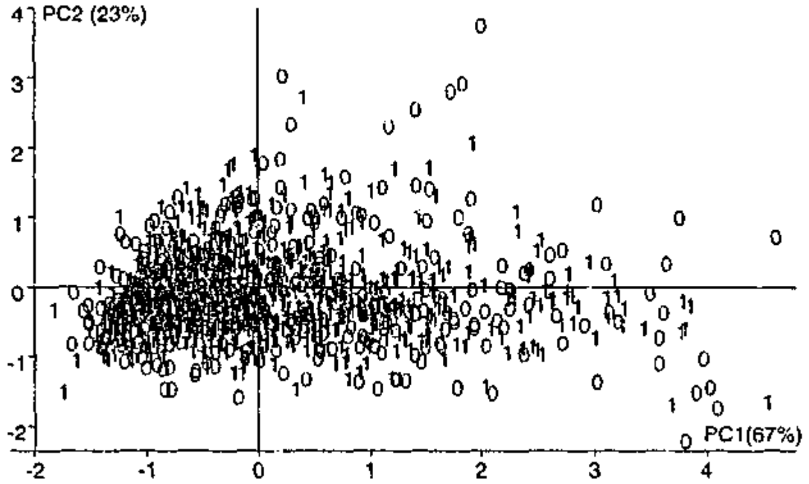

Figure 5.8 PC1 versus PC2 scores plot of normal (0) and abnormal (1) high-grade dysplasia, CIS and SCC samples.

It is undesirable to introduce the errors of the Pap smear into the statistical analysis. The Pap smear has a high degree of inaccuracy associated with it, Section 1.1.7, and 62% of this inaccuracy has been attributed to sampling errors [11], i.e. the sample obtained is not representative of the cytology of the cervix. The samples collected for IR analysis were from the same population of cells used for cytological diagnosis, and whilst they may not be reflective of the cytology of the cervix, or of the biopsy sample, the analytical technique should be based on the cytology of the cells forming the IR spectra.

PCA was performed on the same data, using cytology as the diagnosis. A PC scores plot, Figure 5.9, exhibits more distinct groups of normal and abnormal cells, although there is still overlap and spread of both diagnostic types throughout the plot.

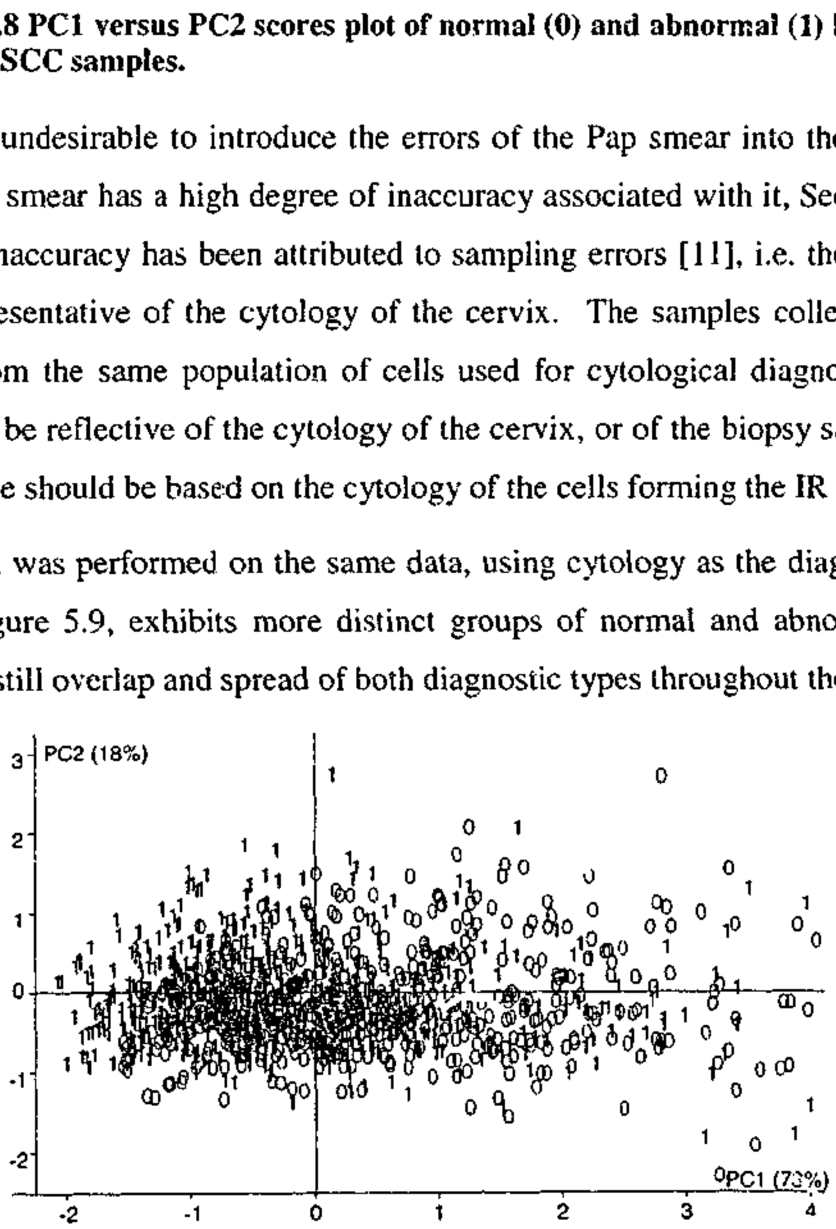

Figure 5.9 PC1 versus PC2 scores plot of normal and abnormal cells with cytology as the diagnosis.

The distribution of abnormal within the normal samples could be occurring as a result of the presence of clusters of normal cells in the abnormal diagnosed deposit. PCA was performed on all the spectra of the samples, so it is likely that samples diagnosed abnormal contain spectra of normal cells given that only a small proportion of cells in a smear are likely to be abnormal [15].

Replicate spectra of each sample were averaged. To remove subjectivity all spectra of each sample were averaged, regardless of spectral profile. Therefore the average spectrum reflects the overall chemistry of the cells in the sample and contains influences from all cell populations within the sample, be it normal epithelial cells at various stages of maturation, abnormal cells, blood components, endocervical cells, mucins or bacteria. The only spectra removed exhibited spatula contamination based on the presence of peaks in the spectral regions as discussed in Section 4.1.2.2.

PCA was performed on the averaged spectra, and the results illustrated in Figure 5.10. The number of spectra have been significantly reduced from the plots seen in Figure 5.8 and Figure 5.9 because each sample is represented by one spectrum rather than by replicate spectra.

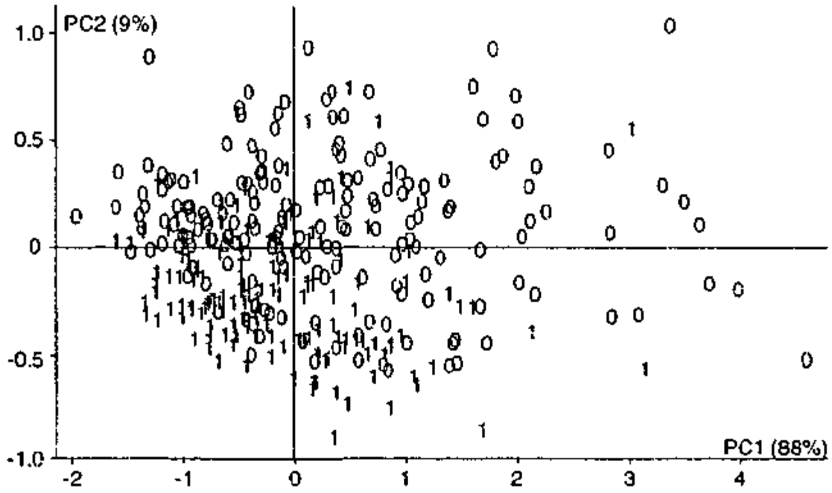

Figure 5.10 PC1 versus PC2 scores plot of averaged normal (0) and abnormal (1) spectra, diagnosed by cytology.

The next stage in the analysis involved determining the best pre-processing methods to enhance the discrimination between the classes of normal and abnormal. Figure 5.11 A – D illustrates the effects of maximum, vector, mean and range normalisation respectively on PCA of the data.

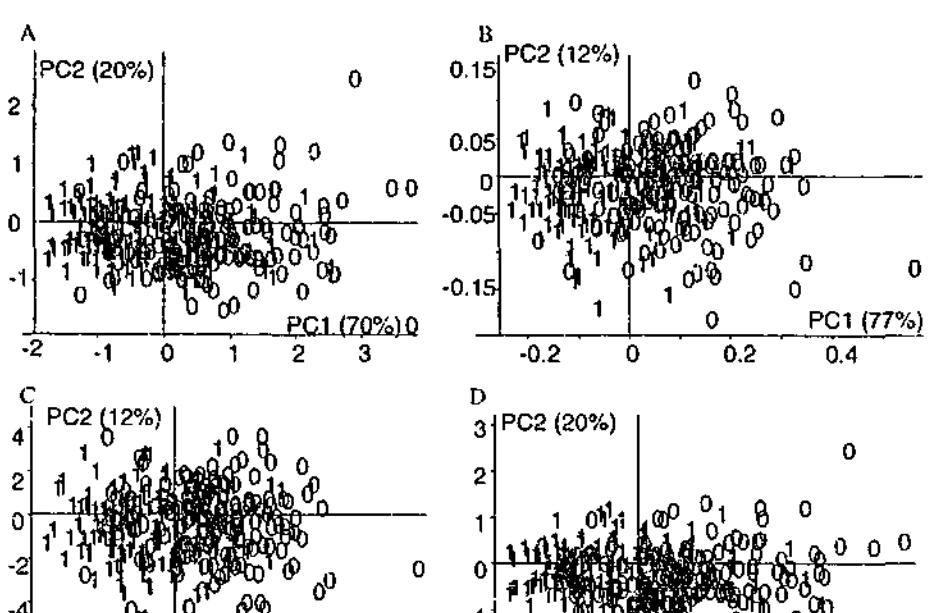

Figure 5.11 PC1 versus PC2 scores plot of normal and abnormal samples illustrating the effects on clustering of normalisation techniques. A – D represent the effects of maximum, vector, mean and range normalisation respectively.

Clusters of normal and abnormal samples are formed with each technique, but there is still a scattered distribution of normal and abnormal samples. Inspection of the loadings plots of these techniques shown in Figure 5.12 show that differences in the glycogen region are the strongest influence contributing to the first PC, which accounts for over 70% of the variance in the data.

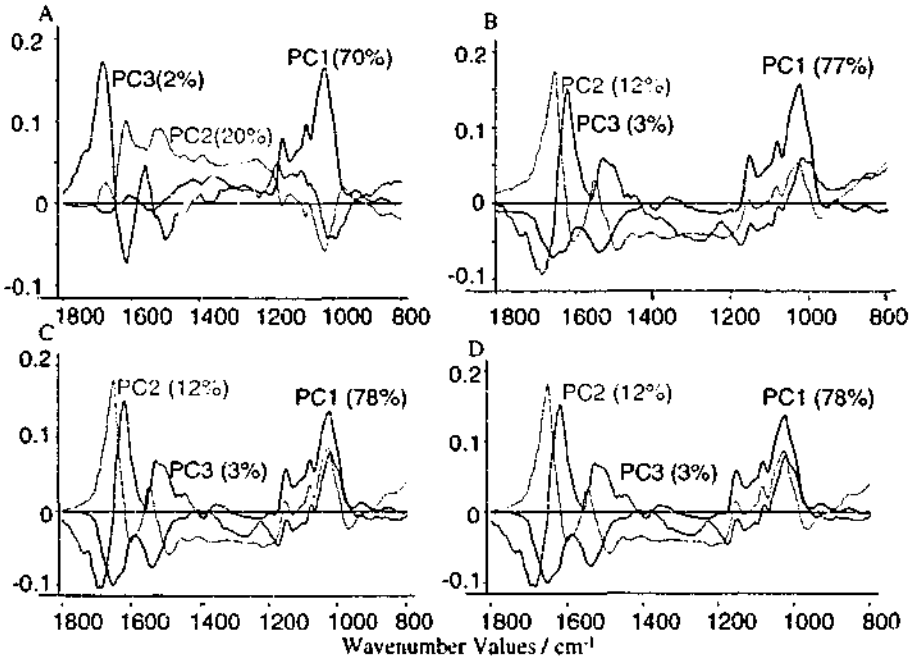

Figure 5.12 Loadings plots of the first 3 principal components resulting from the PCA of maximum (A), vector (B), mean (C) and range (D) normalisation techniques.

5.2.2.2 BAYESIAN REGULARISED ARTIFICIAL NEURAL NETWORKS

The data in each PCA analysis discussed in the previous section was used to form training and test sets for input into the BRANN. Each spectrum represented one sample and consisted of 501 wavenumber values. Network architecture was varied by changing the number of PCs, the inputs to the network, and the number of nodes in the hidden layer.

Basing the diagnosis on cytology results enabled the inclusion of more samples, previously excluded because a biopsy was not performed. FPV and Dysplasia Clinic data without biopsies were added to classes based on cytological diagnosis and the same criteria used for selection of normal and abnormal samples from histological results, discussed at the beginning of Section 5.2.2.1. It was important to include samples from FPV as they represented sampling from a normal population. The new data was included in network training and prediction, as were the results of calculating second order Savitzky-Golay derivative analysis employing a cubic polynomial and 9 smoothing points to remove baseline effects in the spectra.

The results of network training and testing, expressed in terms of correlation (R^2) between the classes of normal and abnormal, where a correlation of 1 indicates ideal discrimination of the classes, are summarised in Table 5.5. The standard error of training (SET) and the standard error of prediction (SEP) give an indication of the error in the network.

Table 5.5 Results of applying a BRANN to the infrared spectra of cervical samples.

Input Data^a	**Spectra**	**PCs**	**Hidden nodes**	**Epochs**	**Training**		**Prediction**	
					SET	**R^2**	**SEP**	**R^2**
H	1130	51	6	2	0.28	0.63	0.46	0.17
C	1116	51	6	3	0.29	0.65	0.48	0.20
C, Av	330	21	4	5	0.40	0.33	0.44	0.16
C, Av, MxN	330	21	4	4	0.40	0.31	0.43	0.19
C, Av, VN	330	21	4	1	0.39	0.32	0.39	0.33
C, Av, MeN	330	21	4	1	0.40	0.31	0.40	0.31
C, Av, RN	330	21	4	4	0.40	0.30	0.41	0.27
Call, Av, D	828	41	6	5	0.21	0.76	0.51	0.23
Call, Av, VN, D	828	46	6	3	0.21	0.77	0.55	0.10

[a] H: histology, C: cytology, MxN: maximum normalisation, VN: vector normalisation, MeN: mean normalisation, RN: range normalisation, Av: averaged data, Call: includes samples without biopsy results, D: second derivative.

The number of inputs and hidden nodes in the architecture hence the number of weights in the network were limited by the number of samples. Ideally the number of weights should be no more than about one third of the number of samples in the training set (Section 2.5.7.4). The number of epochs used for training was not increased because there was no improvement in the evidence (Section 2.5.7.4).

A plot of the samples comparing the predicted outputs of the BRANN with the expected outputs for the network with the largest correlation in training is shown in Figure 5.13A (training) and B (prediction). The largest correlation in training (0.77) arose from vector normalisation and calculation of Savitzky-Golay derivatives of baseline corrected averaged spectra with cytology diagnosis.

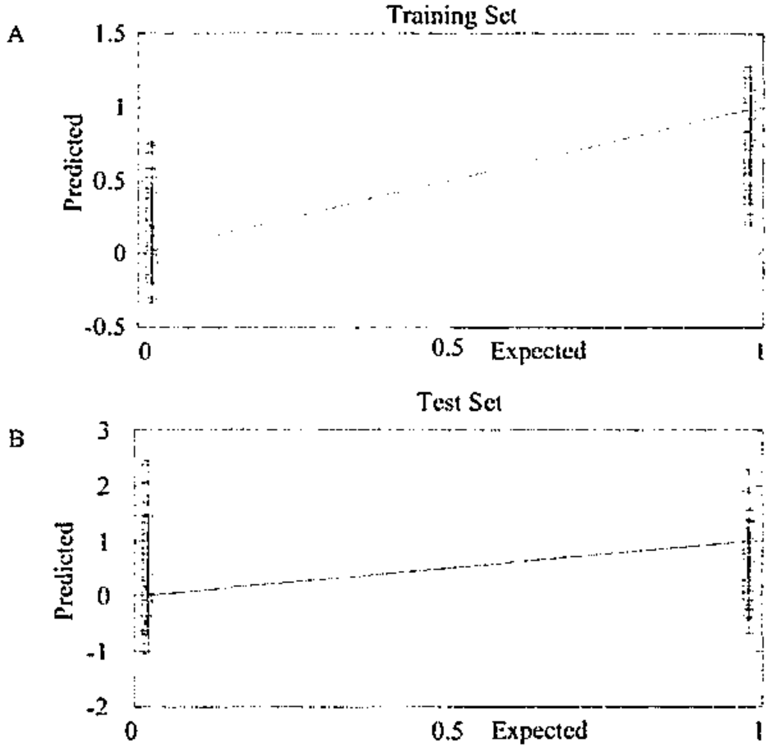

Figure 5.13 Plots comparing the predicted outputs of the BRANN with the expected outputs for the network with the largest correlation of training (A) and prediction (B).

The data is very scattered and the slight clustering of the two classes exhibited for the training set of Figure 5.13A is not distinct enough to allow discrimination between the groups. The correlation of this training set indicates that the network is able to extract patterns from the training set, but is unable to generalise these patterns to form predictions of unknown data, indicated by a prediction correlation of 0.1.

Even though the artificial neural network was unable to classify infrared spectra of normal and abnormal cervical cells, there have been numerous studies using infrared spectroscopy and chemometrics in the diagnosis of cervical cancer that show there is variability between the spectra of normal and abnormal cervical cells [6, 9, 16, 17]. The problems with these studies, as with the analysis performed in Sections 5.1 and 5.2 is that, though the technique used for analysis is objective, the data is pre-selected in some way. Infrared spectra that do not exhibit expected spectral features [6, 9] or spectra without conferring histology and cytology results [18] are removed from the training set. From a statistical point of view, this is a valid decision. Samples used for training and forming models should represent the differences between the two classes, and this is based to a certain extent on the spectral differences seen between normal epithelial cells and HeLa cells, illustrated in Figure 5.14. The main differences occur in the glycogen (1200 – 1000 cm^{-1}) and phosphate (1250 – 1000 cm^{-1}) regions, which are also where a number of bands due to confounding variables occur.

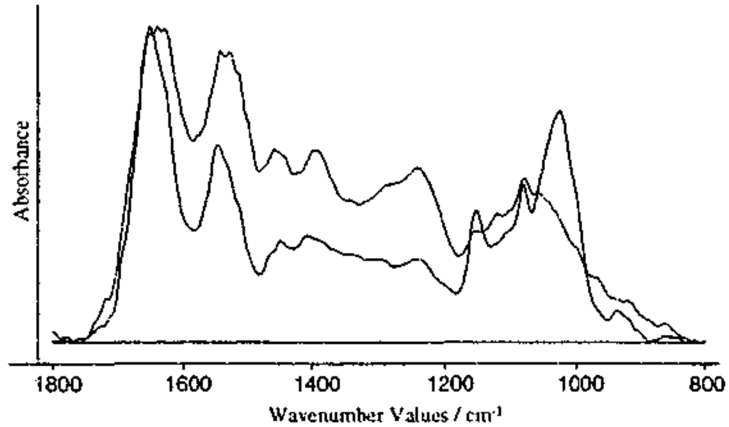

Figure 5.14 Infrared spectra of HeLa cells (black) and normal squamous epithelial cells (blue).

Although spectral differences between abnormal and isolated confounding variables exist [16, 17, 19-22] the differences are subtler than the differences between normal and cancer. As a result, multivariate statistical methods that are based on extracting information about the variance between spectra assigned a cytological or histological diagnosis of normal and abnormal use variables that account for the greatest variance. Perhaps statistical analysis based on differentiating between normal and everything else (abnormal and confounding variables) and then separating out abnormalities from spectra exhibiting the presence of confounding cells and mucin would be a better way to proceed. This is investigated further in Section 5.4.

Whilst the presence of inflammation and other confounding variables are themselves indicators of processes occurring in the cervix, they are present in smears which contain normal cells as well as those exhibiting specific disease states such as dysplasia and malignancy.

If spectroscopy is to be used as a diagnostic technique, there needs to be an analytical method capable of extracting the information from spectra without the exclusion of samples showing the presence of confounding variables. Otherwise the inconclusive rate of screening would be unacceptable.

Figure 5.15 illustrates a PC1 versus PC2 scores plot when all samples were included in the PCA. Spectra were divided into four groups based on cytological diagnosis, summarised in Table 5.6.

Table 5.6 Summary of the cytological diagnoses use to form four groups of infrared spectra for PCA.

Group	Cytology Diagnosis
0	Negative
1	Negative with inflammation, bacterial infection, metaplasia, erythrocytes and other BCCs
2	CIN I, HPV effects and low-grade epithelial abnormalities
3	CIN II-III, CIS, SCC and high-grade abnormalities, including HPV effects

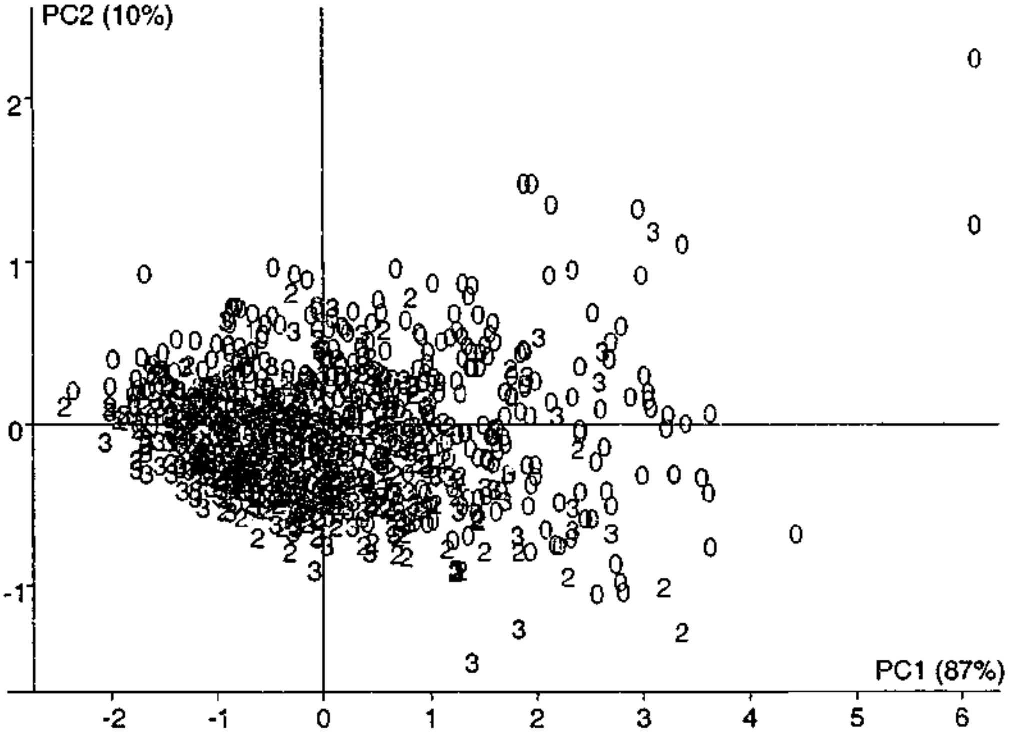

Figure 5.15 PC1 versus PC2 scores plot of IR spectra of cervical samples diagnosed by cytology as normal (0, green); normal but exhibiting signs of inflammation, other blood components, bacterial infection and benign cellular changes, including metaplasia (1, orange); CIN I and HPV effects (2, red); and abnormal, CIN II, III, CIS and SCC, including HPV effects (3, blue).

The scores plot indicates the diversity of IR spectra according to cytology diagnosis and PCA seems unlikely to be able to extract enough information from the spectra to differentiate between the groups.

Even if it was possible to extract this information from spectra, the inhomogeneity of the sample still presents a problem. It is not possible to record spectra of individual cells and time constraints and poor differentiation of unstained cells on the infrared substrate prevent searching for clusters of cells to sample. Therefore, infrared spectra will often represent the chemistry of both epithelial, be it normal or abnormal, cells and confounding cells and mucins.

An alternative to post-spectral extraction of confounding cells and mucins would be chemical removal before spectral collection. The advent of the ThinPrep® processor has helped reduce the influence of inflammatory effects by lysing white blood cells. Erythrocytes are also lysed and the epithelial cells are filtered from non-diagnostic debris. The results of chemical removal of the influences of inflammation and erythrocytes are presented in Section 5.3.

The low proportion of abnormal cells compared to normal cells in a deposit may also present a problem. It is possible to record many infrared spectra of one deposit, but if abnormal cells are missed, or if normal cells surround the abnormality, then spectra will be a composite of normal as well as normal and abnormal cells. This would account for spectra of diagnosed abnormal cells that exhibit large glycogen bands. Infrared imaging systems are available that record spectra of 4096 pixels on a sample. If the whole sample deposit could be analysed, this would minimise the risk of missing abnormal cells, and spectral regions of different areas on the deposit could be ratioed as an objective means of determining the best spectrum to use for analysis.

5.2.3 CONCLUSION

BRANNs were unable to predict unknown spectra of normal and abnormal diagnosed cervical samples when cytology was used as a diagnostic standard and spectra were not visually pre-selected.

Multivariate statistical techniques were able to differentiate between IR spectra of normal and abnormal cells on pre-selected data with histological and cytological agreement, excluding the effects of inflammation, bacterial infection, and benign cellular changes. Methods to reduce the effects of these confounding variables need to be investigated, either by removal of these spectral contaminants before spectroscopy or through other statistical techniques after spectroscopy.

5.3 CLEAN-UP OF CERVICAL SMEARS

Biological samples, in particular cervical smears, are intrinsically variable in nature. The population of epithelial cells, exhibiting various stages of maturation, is dependent on endogenous hormones and therefore cyclically dependent. Cervical smears, as well as containing populations of intermediate and superficial cells, may contain parabasal cells, in the case of atrophy associated with the post menopausal stage, as well as metaplastic squamous cells and endocervical cells. Smears may also contain platelets, erythrocytes (red blood cells, RBC), white blood cells (WBC) including polymorphonuclear leukocytes (PMNs), leukocytes and thrombocytes, which occur as an inflammatory response to tissue damage or trauma. Bacteria is naturally present in the cervix to maintain acidic pH, although there are several other strains such as *Candida albicans*, *Actinomyces* and *Trichomonas* which can cause infection. Mucins are also present in the smears, the viscosity of which is cyclically influenced [23].

Given the intrinsic variability of cervical smears, it would be beneficial to the infrared spectroscopic technique if some of the non-diagnostic debris associated with cervical smears could be removed. The presence of erythrocytes and other blood components may mask the spectral patterns of the epithelial cells, potentially confounding diagnosis.

The introduction of the ThinPrep® processor is a novel method of "cleaning up" cervical smears and producing homogenous, evenly deposited smears for cytological analysis. This is achieved by rinsing the collected cells in an alcohol buffered preservative solution able to lyse red blood cells and kill microbiological elements. Centrifugation at high speed breaks up large clumps of mucus and cellular clusters ensuring homogenisation of the suspension. The suspension is then filtered to minimise the presence of white blood cells, mucins and non-diagnostic debris. Refer to Section 1.2.2. for a detailed explanation.

In an attempt to reduce some of the confounding variables found in cervical smears, ThinPrep® solution (known commercially as PreservCyt™), white cell lysis buffer (WCLB) and red cell lysis buffer (RCLB) were investigated for their ability to reduce the presence of confounding variables, non-diagnostic debris and to improve the homogeneity of cellular deposits.

5.3.1 METHODOLOGY

Cervical smears used for the development of the "clean-up" protocol were obtained from the Royal Women's Hospital (Melbourne, Australia), Dysplasia Clinic and from theatre patients at the hospital undergoing laser ablation for cervical abnormalities such as dysplasia.

5.3.1.1 SAMPLE COLLECTION AND PREPARATION

Despite the fact that the presence of an endocervical component in cervical smears is deemed necessary for a satisfactory smear in cytology, the presence of these cells for IR analysis may cause confounding results to diagnosis. Due to this fact, ectocervical and endocervical cells were collected separately for the clean-up process. Samples used for these experiments were collected from women undergoing laser ablation therapy for cervical dysplasia of varying degrees.

5.3.1.1.1 THINPREP®

Ectocervical cells, obtained with an Ayre spatula, and endocervical cells, obtained with a Cytobrush™ were collected, after conventional Pap smears were made, in separate 50 cm^3 centrifuge tubes containing 10 ml of PreservCyt™ solution. The tubes were centrifuged at 3000 r.p.m. for 10 minutes, the supernatant removed and the resultant cellular pellet deposited into the multicavity infrared cell and desiccated under vacuum.

5.3.1.1.2 RED AND WHITE CELL LYSIS BUFFER

Ectocervical and endocervical cells were collected, after conventional Pap smears were made, in separate 50 cm^3 centrifuge tubes containing 10 ml of physiological saline solution (0.9% w/v). The tubes were centrifuged at 3000 r.p.m. for 10 minutes, the supernatant removed and a 10 μl portion of the cellular material deposited onto the IR substrate with the remaining cells resuspended in 10 ml of WCLB, vortexed and incubated at 37°C for 15 minutes. Following incubation the tubes were centrifuged for a further 10 minutes, and washed twice in saline to remove platelets. A 20 μm aliquot of the resultant vortexed suspension from each sample was spread over a glass microscope slide with a pipette and the slide sprayed with ethanol fixative. A further 10 μm aliquot was deposited on the KRS-5 infrared substrate for subsequent infrared analysis. If the cervical smears showed the presence of erythrocytes, RCLB was added to samples at the incubation stage.

5.3.1.2 CYTOLOGICAL ANALYSIS

Independent cytological analysis of the cells used in this experiment was undertaken in the Histology Department of Monash University.

5.3.2 RESULTS AND DISCUSSION

5.3.2.1 THINPREP®

Although the precise chemical makeup of the ThinPrep® PreservCyt™ solution is not available to the public, the solution comprises of an alcohol buffered solution containing EDTA (ethylenediaminetetraacetic acid or [ethylenedinitrilo] tetraacetic acid) coupled to a dihydrated disodium salt. Figure 5.16 shows the IR spectrum of the ThinPrep® solution after desiccation.

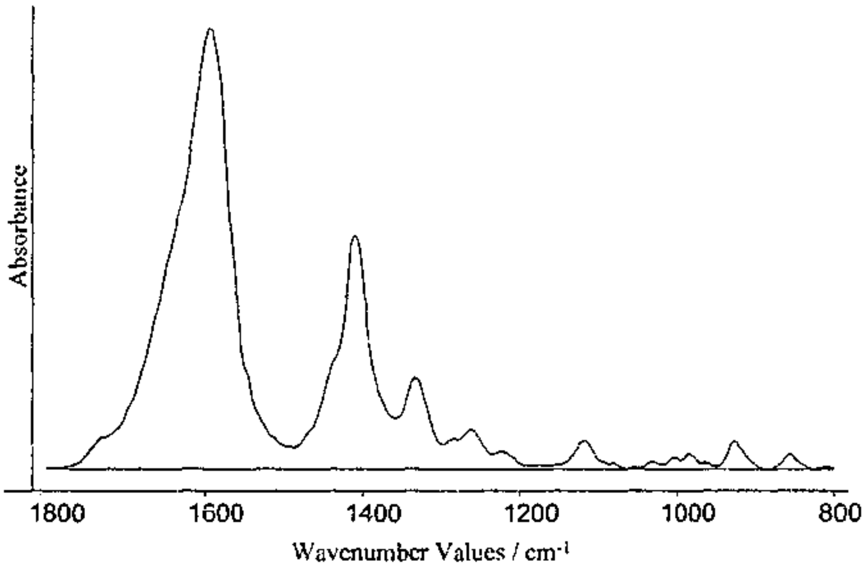

Figure 5.16 Infrared spectrum of the ThinPrep® PreservCyt™ solution after desiccation.

In a blind study, samples suspended in PreservCyt™ solution were obtained for IR analysis. Samples were not washed prior to deposition and IR spectra of these samples appeared to be contaminated with PreservCyt™ solution. Due to the cost of this solution, and the similar chemical make-up to WCLB, see below, no further investigations were performed using this solution.

5.3.2.2 BLOOD CELL LYSIS BUFFERS

White cell lysis buffer contains 10 mM Tris HCl, pH 8.2, 400 mM NaCl, 2 mM Na_2EDTA and red cell lysis buffer contains 77.03 g NH_4Cl and 0.84 g $NaHCO_3$ made up to one litre with distilled water.

Initial attempts at cell clean-up proved futile because the original protocol involved four stages of washing the cells in saline solution to remove the presence of buffers and platelets, which can cause aggregation and reduce the homogeneity of the deposit. Unfortunately so many washing stages also removed most of the cellular material leaving spectra that were very noisy and not suitable for further analysis. Consequently the number of washing steps was reduced to two, with no contamination seen from the presence of the white cell lysis buffer. Unfortunately two washings was not enough to remove the presence of red cell lysis buffer. The infrared spectrum of red cell lysis buffer, after desiccation, is shown in Figure 5.17 and a cervical cell spectrum contaminated with the buffer in Figure 5.18.

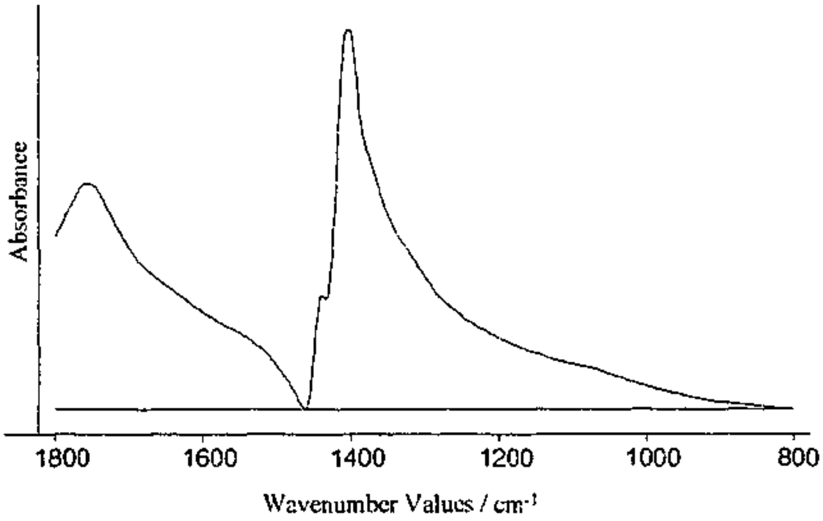

Figure 5.17 Infrared spectrum of red cell lysis buffer.

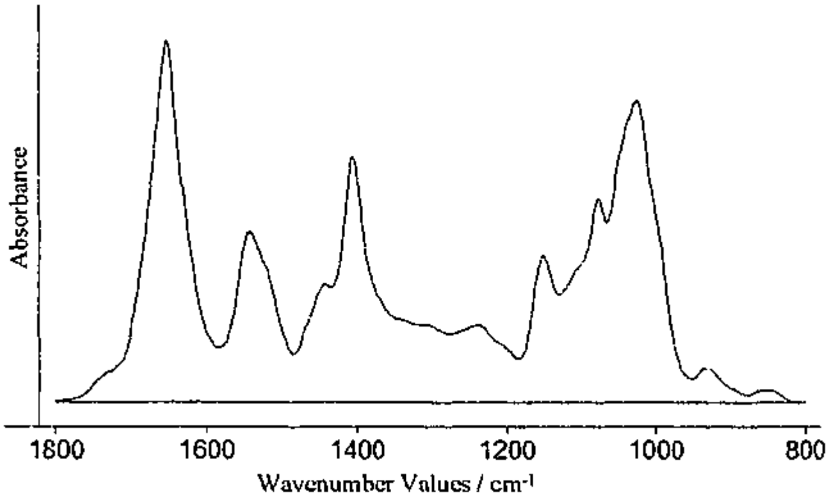

Figure 5.18 Infrared spectrum of cervical cells contaminated with red cell lysis buffer.

Figure 5.19 shows the infrared spectrum of white cell lysis buffer after desiccation. Due to its similar chemical makeup, it is not surprising that there are several similar peaks between this spectrum and the one seen in Figure 5.16 of the PreservCyt™ solution.

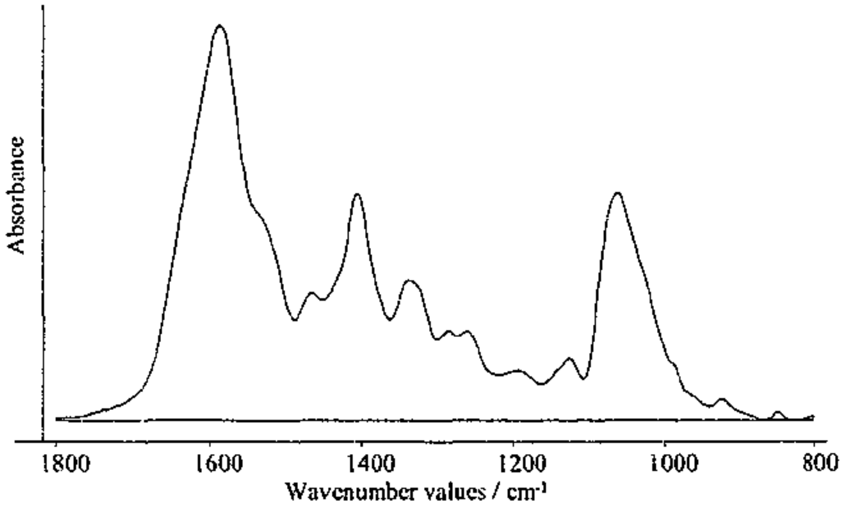

Figure 5.19 Infrared spectrum of white cell lysis buffer

Initial results from the revised protocol indicate that this process was successful in the removal of white blood cells from the cervical smears. This was demonstrated spectroscopically, Figure 5.20, and also by light microscopy of the before and after microscope slides (Figure 5.21), which were fixed and stained according to conventional Pap smear protocol.

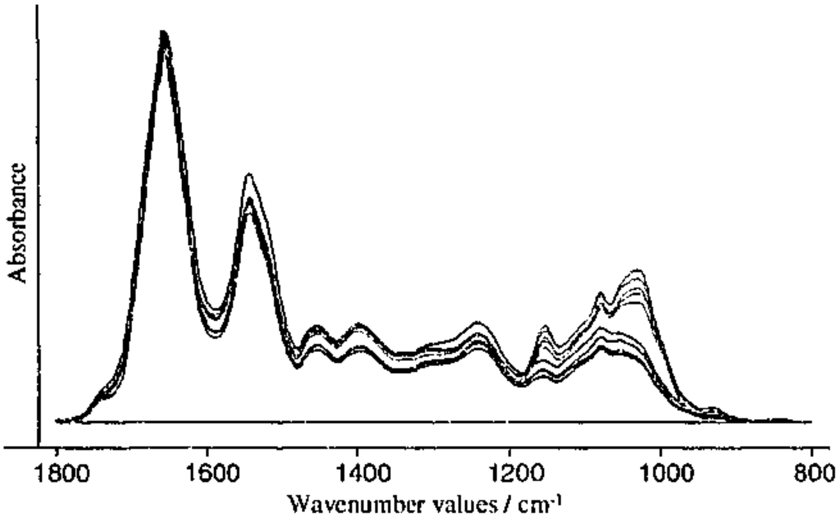

Figure 5.20 Infrared spectra showing the results of the clean-up protocol. Black spectra represent cervical cells before the clean-up process, and red spectra represent cervical cells after the clean-up process

The infrared spectra shown in Figure 5.20 show two different spectral patterns. The spectral pattern characteristic of the cervical cells before the clean-up process resembles spectra of lymphocytes, with characteristic $v_{as}PO_2^-$ and $v_sPO_2^-$ bands at 1240 and 1080 cm^{-1} respectively [24]. The spectra of cervical cells after the clean-up process has been performed show the characteristic bands associated with cervical epithelial cells. These findings are confirmed on inspection of the Pap smears, before clean-up (Figure 5.21A) and post clean-up (Figure 5.21B) under light microscopy, which indicate that the presence of white blood cells, in particular polymorphonuclear leukocytes have been removed.

The spectra also indicate that this technique has improved the homogeneity of the sample, which is seen by the high degree of reproducibility throughout the sample and the small degree of variation or spread of the spectra. The high reproducibility of these spectra post clean-up is further illustrated in Figure 5.22.

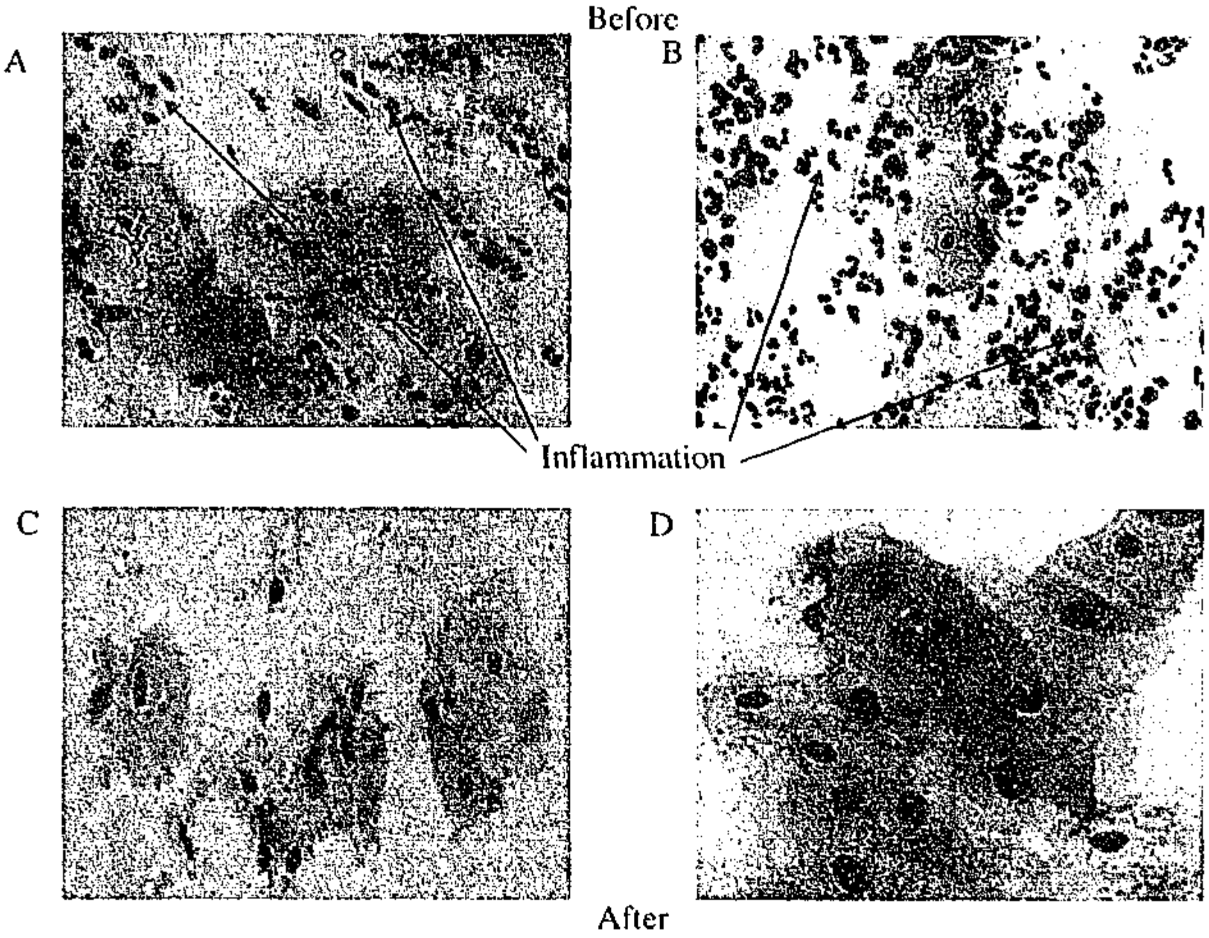

Figure 5.21 Light microscopy slides of cervical smears used in the clean-up process. A. and B. Cervical smears before the clean-up process showing a high proportion of white blood cells. C. and D Corresponding smears post-clean-up exhibiting no white blood cells.

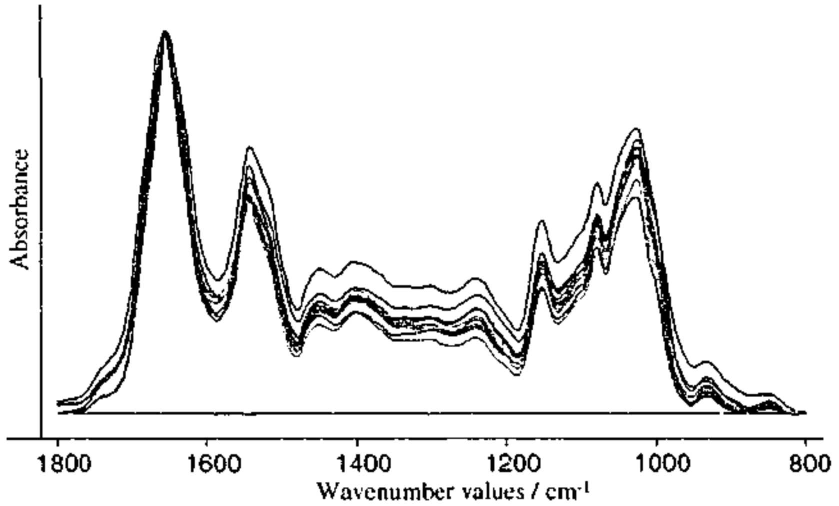

Figure 5.22 Infrared spectra showing the results of the clean-up protocol. Black spectra represents cervical cells before the clean-up process, and red spectra represent cervical cells after the clean-up process.

Principal component analysis (PCA) was performed on a small subset of the samples in the existing database with the spectra of the cleaned up cervical samples included. This was to determine if cleaning up the samples offered any benefits in terms of increased discrimination between normal and abnormal diagnosed samples. The resultant PC1 versus PC2 scores plot is shown in Figure 5.23. The cytological and histological diagnosis for the clean-up samples is given in Table 5.7.

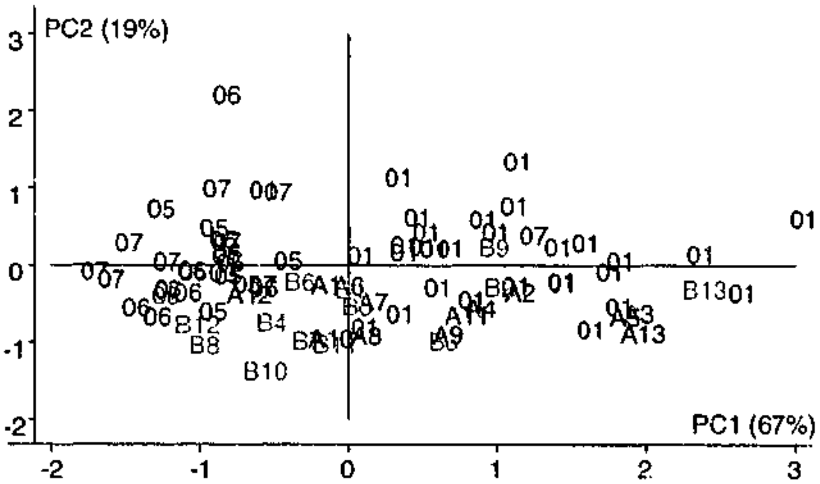

Figure 5.23 Scores plot showing distribution of clean-up samples before (B, orange) and after (A, purple) with normal (01, green) and abnormal (05, 06, 07 representing CIN II, III and CIS respectively, blue).

Table 5.7 Diagnostic results for samples used in clean-up experiment[22].

Sample Number	Cytology Result	Histology (Biopsy) Result
1	0421	0421
2	01	0421
3	0421	0421
4	0421	0421
5	01	0521
6	0421	0521
7	30	0421
8	01	0421
9	21	0421
10	01	0421
11	01	0521
12	040521	0521
13	01	0421

Whilst there appears to be slight differences in the distribution of the before and after clean-up samples, there is not necessarily an increase in the discrimination of the samples compared to normal and abnormal. Infrared spectra of before clean-up samples were recorded to ensure that the buffers used in this experiment were not affecting the integrity of the cells. When it was shown that this was not the case, through light microscopy and spectroscopy, infrared spectra were only obtained for samples post clean-up.

---

[22] Refer to Table 3.1 for an explanation of the diagnostic codes.

The majority of the clean-up spectra used for the PCA were the result of cellular material collected from the Ayre spatula and therefore would be expected to contain only ectocervical cells. Since the samples from the database were the result of mixed populations of ectocervical and endocervical cells, it is possible that the different populations of cells could have affected the discrimination. A second PCA was performed including spectra obtained from ectocervical cells only and no improvement in discrimination was achieved.

Even though the clean-up samples were supposedly ectocervical only, independent cytological analysis revealed that 89% of the samples used in the above PCA contained the presence of endocervical cells, despite cervical smears being obtained by spatula only. Cytological analysis also revealed that 92% of the samples appearing in Figure 5.23 exhibited signs of inflammation, ranging from mild to marked[23]. Since it has been noted that inflammation exhibits spectral features similar to those displayed in dysplasia and malignancy [17], it is likely that the differences seen between the before and after clean-up samples is due to the removal of inflammatory exudate.

The lack of discrimination between the clean-up samples could also be due to the majority of samples used in this analysis having a diagnosis of CIN I. Since the main aims of this research were to achieve discrimination between high-grade dysplastic and malignant samples from normal samples, only these samples from the database were used in the PCA. Assuming that the process of neoplasia is continuous from CIN I to CIN III and CIS, the clean-up samples would be expected to appear in between the two groups, as in seen the majority of samples in Figure 5.23. It is also possible that the differences seen between the abnormal (blue) and normal (green) samples are just a result of differences in inflammatory response often exhibited with dysplastic and malignant samples [25].

PCA was performed on all the post clean-up samples to compare the biopsy results with an independent cytological analysis. The resulting PC1 versus PC2 scores plot is given in Figure 5.24.

---

[23] Inflammatory infiltrate is recorded as mild when it covers ⅓ of the slide, moderate when it covers ½ of the slide and marked when it covers ¾ to all of the slide.

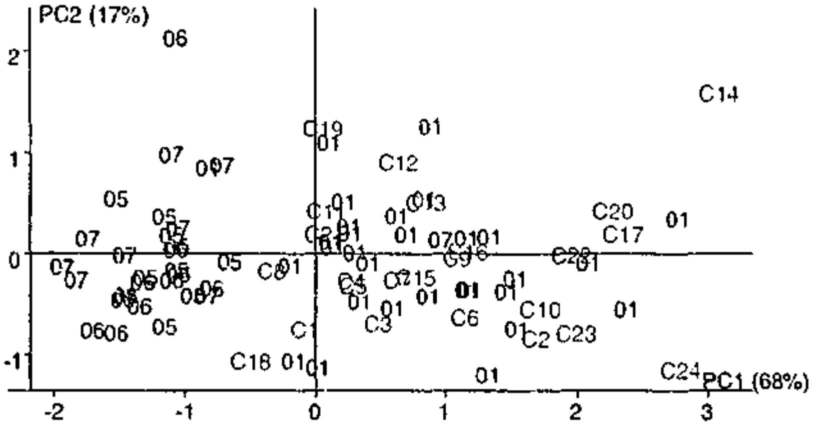

Figure 5.24 Scores plot of normal (green), abnormal (blue) and samples after clean-up (red).

As was seen in Figure 5.23, the scores plot shown in Figure 5.24 also shows post clean-up samples grouping with normal samples. The diagnostic results for these samples are given in Table 5.8. Eighty three percent of these samples exhibit signs of inflammation and 58% contain an endocervical component, despite 88% of the samples being obtained through spatula only. Table 5.8 also highlights the disparity seen between cytological and histological (biopsy) diagnosis, summarised in Table 5.9. It is possible, as stated previously that the discrimination seen between normal and abnormal cells arises from the inflammatory response associated with dysplasia and malignancy, rather than due to individual differences between the two disease states. Although there is a high proportion of samples diagnosed as CIN I/HPV (0421), a second PCA excluding these samples did not improve the discrimination.

Despite the fact that biopsy is considered the gold standard for diagnosis of cervical cancer, biopsies from the patients participating in this study were performed up to 6 weeks prior to surgery. Whilst it is highly unlikely that these abnormalities have regressed, a more plausible explanation for the lack of discrimination of abnormal clean-up samples and the lack of agreement with cytology results is due to the smears being taken at different times from the biopsy. Therefore, since the samples used in this experiment were of the same cell population as the smears sent for cytological analysis, it makes sense to use cytology results rather than biopsy for the definitive diagnosis. Even using this rationale, however, the two cytologically (high-grade) abnormal samples (C10 and C20) are grouped with the normal samples.

Table 5.8 Biopsy and Cytology diagnosis from clean-up samples, indicating the presence of an endocervical component and inflammatory exudate.

No.	Biopsy[a]	Cytology	Ectocervical only or mixed	Endocervical cells present	Inflammation
C1	0421	01	Mixed	Yes	Mild
C2	0421	01	Mixed	Yes	Mild
C3	0421	04	Mixed	No	Moderate
C4	0421	21	Ectocervical only	Yes	
C5	0421	01	Ectocervical only	No	Moderate
C6	0421	0421	Ectocervical only	Yes	Marked
C7	0421	0421	Ectocervical only	Yes	Marked
C8	0421	0421	Ectocervical only	Yes	
C9	0421	01	Ectocervical only	Yes	Marked
C10	0521	0521	Ectocervical only	Yes	Marked
C11	0521	0421	Ectocervical only	Yes	Marked
C12	0521	01	Ectocervical only	No	
C13	0521	01	Ectocervical only	Yes	Marked
C14	040521	04	Ectocervical only	No	Marked
C15	0421	0421	Ectocervical only	Yes	Mild
C16	0421	01	Ectocervical only	Yes	Moderate
C17	0421	21	Ectocervical only	No	Mild
C18	0421	04	Ectocervical only	Yes	Marked
C19	0621	01	Ectocervical only	Yes	Moderate
C20	050621	0521	Ectocervical only	No	Moderate
C21	0421	0421	Ectocervical only	No	Marked
C22	0421	04	Ectocervical only	No	Marked
C23	0421	01	Ectocervical only	No	
C24	0421	01	Ectocervical only	No	Marked

[a] A description of the diagnostic codes for cytology and histology are given in Table 3.1.

Table 5.9 Histology (biopsy) and cytology results for the cervical smears of 24 samples used in the clean-up process.

Diagnosis (code)	Biopsy (%)	Cytology (%)
Negative (01)	0	42
HPV (21)	0	8
CIN I (04)	0	17
CIN I, HPV (0421)	71	25
CIN I-II, HPV (040521)	4	0
CIN II, HPV (0521)	17	8
CIN II-III, HPV (050621)	4	0
CIN III, HPV (0621)	4	0

5.3.3 CONCLUSION

Red and white cell lysis buffers were investigated for their ability to remove blood components from cervical smears. Even though the use of white cell lysis buffer has demonstrated that it is an effective method for increasing spectral reproducibility and sample homogeneity and reducing the presence of inflammatory exudate, in particular PMNs from cervical smears, it appears that the clean-up process has reduced the ability of these samples to be discriminated from samples with normal cytology. Rather, the reduction of PMNs appears to be causing abnormal samples to be grouped with normal samples. The extent of this phenomenon is difficult to ascertain given that 42% of the samples investigated were diagnosed cytologically as normal, and 50% of the samples were diagnosed cytologically as having low-grade (CIN I and/or HPV) abnormalities.

5.4 CONFOUNDING VARIABLES AND MULTIVARIATE STATISTICS

Our group has previously investigated potential confounding variables in the spectroscopic diagnosis of cervical cancer [17]. The following were investigated: endocervical cells, erythrocytes, leukocytes, platelets, fibroblasts, connective tissue, mucin, semen and bacterial and yeast infections. The study identified endocervical cells, leukocytes, fibroblasts, connective tissue, mucins and semen as possible confounding variables. The infrared spectra of semen exhibited a characteristic doublet at 981 and 968 cm^{-1} allowing removal by visual analysis. It was concluded that erythrocytes, platelets and bacterial infections would not influence the diagnostic technique and were therefore considered as nonconfounding variables. *Candida albicans*, a common yeast infection of the cervix, was considered a possible confounding variable depending on severity of infection in the cervix.

Other groups have noted spectral differences between normal diagnosed cervical smears and those containing metaplastic cells [20], parabasal cells [16, 20] and PMNs [22]. The IR spectra of these cells are similar to those of high-grade dysplasia and malignancy.

Many of the cellular changes brought about by the identified confounding variables are diagnosed cytologically under the general term benign cellular changes (BCCs). The Bethesda System reports infection by *Trichomonas, Candida* and *Actinomyces* and cellular changes associated with inflammation and atrophy under this category [26].

The study by Wood *et al.* [17] also employed PCA as a means of separating normal from abnormal diagnosed cervical cells, investigating the separation achieved using IR spectra obtained from cervical smears containing only ectocervical cells and from smears containing only endocervical cells. Tighter clusters and a better separation between normal and abnormal diagnosed samples was found using the spectra of smears containing only ectocervical cells [17]. The results of multivariate statistical analysis of cervical smears containing only ectocervical cells is presented in the PhD Dissertation by Wood [24]. One hundred and two infrared spectra of normal (69) and abnormal (CIN I (3), CIN I/II (3), CIN II (11), CIN II/III (5) and CIN III (11)) diagnosed samples were classified utilising SIMCA, LDA and *K*-NN. Table 5.10 summarises the sensitivity and specificity of the three techniques.

Table 5.10 Sensitivity and specificity of SIMCA, LDA and *K*-NN in the classification of cervical smears principally containing ectocervical cells.

Technique	Sensitivity (%)	Specificity (%)
SIMCA	82	88
LDA	100	91
K-NN (*K*=1)	92	100

Despite the fact that this analysis shows high sensitivity and specificity for the classification of normal and abnormal diagnosed ectocervical smears, it was decided[24] that samples collected for the analysis presented in Sections 5.1 and 5.2 would consist of both ectocervical and endocervical cells. Because the presence of both cell types is deemed necessary by cytologists for a satisfactory smear representative of the transformation zone (Section 1.1.2.1).

It is important to determine the exact influence of endocervical cells in the IR spectra of cervical smears. Whilst there are obvious visual spectral differences between abnormal diagnosed cervical smears and endocervical cells, it is necessary to ascertain if multivariate statistics is able to distinguish between these cell types. Multivariate statistics was also employed to investigate the spectroscopic effects of BCCs in cervical smears.

---

[24] Under the advice of Professor Michael Quinn, Obstetrician and Gynaecologist at the Royal Women's Hospital.

5.4.1 *METHODOLOGY*

5.4.1.1 *SAMPLE COLLECTION*

The preparation of cervical smears for infrared spectroscopy has been described in Sections 4.1.1 and 5.1.1. Endocervical cells in this study were obtained from the hormonal study data. Smears diagnosed as normal but exhibiting BCCs including inflammation, bacterial or yeast infection were obtained from the database of smears. To increase the number of samples diagnosed with inflammation, smears diagnosed negative with inflammation effects ranging from mild to marked were obtained from the clean-up study. These spectra were obtained from smears before the clean-up process was performed. Normal and abnormal diagnosed (histological and cytological agreement) smears were randomly obtained from the database. Table 5.11 summarises the number of samples for each cell type or diagnosis investigated in this study.

Table 5.11 Number of IR spectra of each cell or diagnostic type used for PCA.

Symbol	**Cell type or diagnosis**	**Number of samples**
N	Negative (normal)	30
05	CIN II	10
06	CIN III	10
07	CIS	10
E	Endocervical	67
I	Inflammation	7
Y	*Candida albicans*	13
B	Bacterial vaginosis	9
C	Cervicitis	2
A	Atypia	1
K	Keratinisation	2
M	Metaplasia (immature)	1

5.4.1.2 *DATA TREATMENT*

Infrared spectra were baseline corrected and spectra for each sample averaged. The averaged spectra were normalised to the amide II band. Spectra exhibiting spatula contamination were removed prior to averaging.

The reason why there were low numbers of samples representing each diagnostic type was three-fold:

1. Cervical smears exhibiting severe cases of any of the above diagnoses may have been classified by the cytologist as inconclusive or unsatisfactory and discarded from the spectral databank.
2. These diagnoses may have been made in conjunction with a diagnosis of abnormality, be it dysplasia or malignancy, and grouped accordingly.
3. Infrared spectra may have been discarded due to nonlinearity effects or a low signal-to-noise ratio.

The number of normal and abnormal (CIN II, CIN III and CIS) samples were restricted in order to approximately match the numbers of samples with BCCs investigated in this study.

5.4.2 RESULTS AND DISCUSSION

5.4.2.1 ENDOCERVICAL CELLS

The spectral differences between normal ectocervical, abnormal ectocervical and normal endocervical cells are shown in Figure 5.25. The spectra presented in this figure are the result of averaging all the samples in each type (refer to Table 5.11).

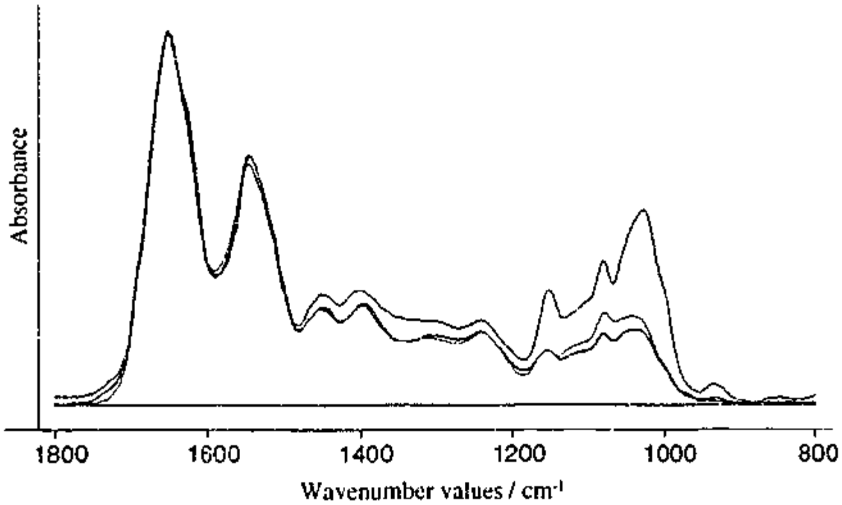

Figure 5.25 Averaged IR spectra of normal ectocervical (green), abnormal ectocervical (blue) and normal endocervical (red) cells.

The infrared spectra of endocervical cells and abnormal ectocervical cells are almost identical, except for differences in the $\nu_sPO_2^-$ and νC-O bands appearing at approximately 1080 and 1040 cm^{-1} respectively. The band at 1080 cm^{-1} is shifted in the endocervical spectrum compared to 1082 cm^{-1} in the ectocervical spectrum. This band arises from

$\nu_sPO_2^-$ of nucleic acids and $\nu_sCO\text{-}O\text{-}C$ of glycogen. The band shape arising from C-O vibrations of carbohydrate moieties is also different in the two spectra. This would be expected given the different types of carbohydrates present in the two cell types. Endocervical cells contain mucin, which exhibit a νCH_2OH band at 1043 cm^{-1}. Ectocervical cells contain glycogen, which exhibit νC-OH and δC-OH bands at 1047 cm^{-1} and a νCH_2OH band at 1025 cm^{-1}. Figure 5.26 highlights the spectral differences between normal endocervical cells and abnormal ectocervical cells in the lower phosphodiester and carbohydrate regions.

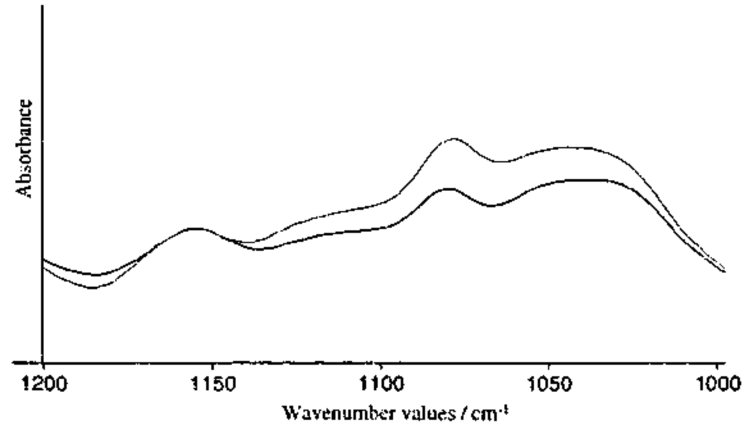

Figure 5.26 IR spectrum of normal endocervical cells (red) and abnormal ectocervical cells (blue).

In comparison with these two spectra, the IR spectrum of normal ectocervical cells exhibits an increase in intensity of the bands arising from:

1. $\nu_{as}CH_3$ of lipids and proteins (1450 cm^{-1}).
2. δCH_3 of proteins (1400 cm^{-1}).
3. $\nu_{as}PO_2^-$ (1240 cm^{-1}) and $\nu_sPO_2^-$ (1080 cm^{-1}) of phosphodiester linkages in nucleic acids.
4. νC-OH of proteins and νC-O of carbohydrates (1154 cm^{-1}), $\nu_sCO\text{-}O\text{-}C$ of glycogen (1080 cm^{-1}) and νC-O of glycogen (1028 cm^{-1}).

The intense, broad glycogen band at 1028 cm^{-1} overlap any contributions from νC-OH and δC-OH of glycogen at 1047 cm^{-1}. The δCH_3 band is shifted in the spectrum of the normal ectocervical cells (1402 cm^{-1}) compared with the abnormal ectocervical and normal endocervical cells (1396 cm^{-1}).

PCA was performed on the spectra of normal and abnormal ectocervical cells and normal endocervical cells to determine if a separation between the groups could be achieved. Figure 5.27 shows the PC1 versus PC2 scores plot of the three groups. Whilst there is a good separation between normal and abnormal ectocervical cells, and between normal ectocervical and endocervical cells, there is no separation between abnormal ectocervical cells and normal endocervical cells.

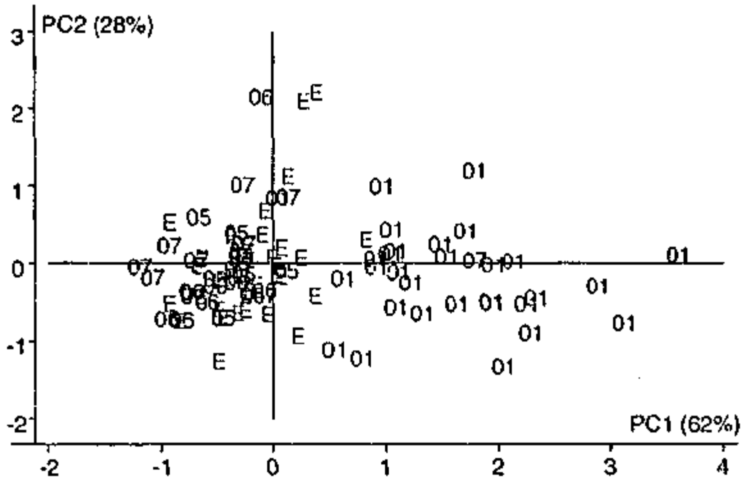

Figure 5.27 PC1 versus PC2 scores plot showing a separation between normal ectocervical cells (green, 01) and abnormal ectocervical cells (blue, 05-07) and normal endocervical cells (red, E).

The loadings plots of the first three principal components, Figure 5.28, were inspected to identify the wavenumber values contributing to the majority of the variance. The boxes marked (a) and (b) highlight the wavenumber values chosen for further PCA: 1096-1062 cm^{-1} and 1060-994 cm^{-1} respectively.

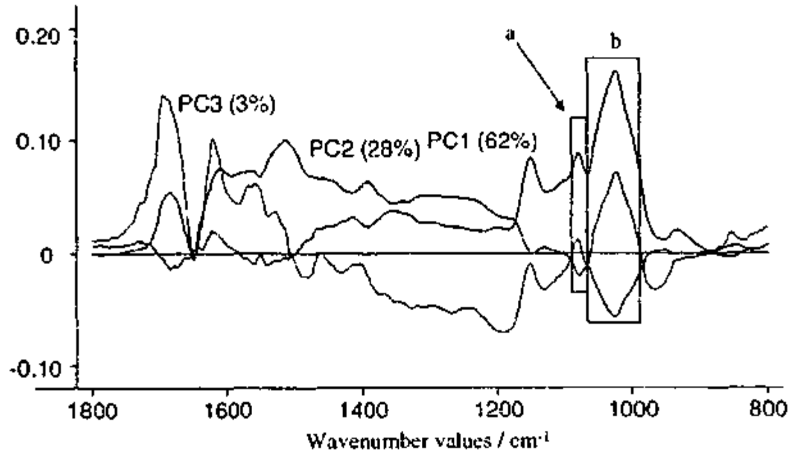

Figure 5.28 Loadings plots of the first three principal components. The boxes marked (a) and (b) highlight the wavenumber values chosen for further PCA, 1096-1062 cm^{-1} and 1060-994 cm^{-1} respectively.

Despite large principal component loadings for PC3 in the region 1700-1500 cm^{-1} IR spectral differences were not noted in this region. PCA was performed on spectra reduced to either (A) 1096-1062 cm^{-1} or (B) 1060-994 cm^{-1}, where the major variance occurs, and the resultant scores plots are shown in Figure 5.29. These two regions were chosen because spectral differences were noted between the IR spectra of abnormal ectocervical and normal endocervical cells at approximately 1080 cm^{-1} and between normal ectocervical and endocervical cells at 1040 cm^{-1}. The first two PCs account for 100% of the variance in both scores plots, and whilst there is a separation from normal in both plots, the groups form tighter clusters in the PCA of data reduced to variables in the region 1095-1062 cm^{-1}.

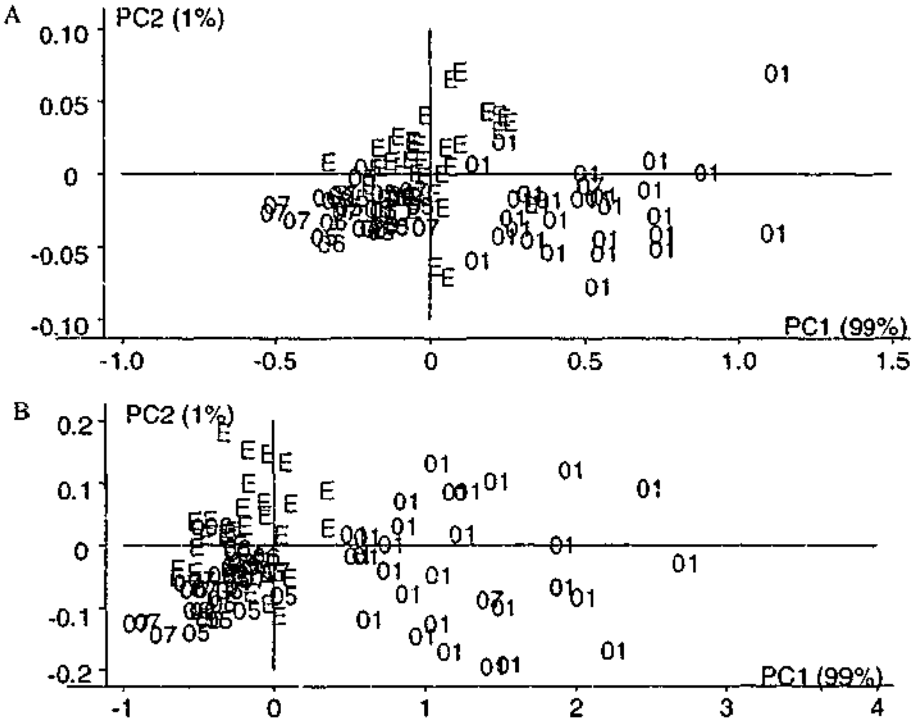

Figure 5.29 PC1 versus PC2 scores plot of data in the regions 1096-1062 cm^{-1} (A) and 1060-994 cm^{-1} (B) of normal ectocervical (green, 01), abnormal ectocervical (blue, 05-07) and normal endocervical (red, E).

PCA was also performed on normal ectocervical and endocervical cells and on abnormal ectocervical and normal endocervical cells separately to determine which variable region gives a better discrimination between the two groups. The PC1 versus PC2 scores plots for the 1096-1062 cm^{-1} region are shown in Figure 5.30. (A) shows the PC scores plots for normal endocervical and normal ectocervical cells and (B) shows the PC scores plot for normal endocervical and abnormal ectocervical cells. The slight overlap in the separation between normal endocervical cells and abnormal ectocervical cells, is less than the overlap seen when the entire wavenumber value region (1800-800 cm^{-1}) was used for analysis. Figure 5.31 shows the PC1 versus PC2 scores plots for the 1060-994 cm^{-1} region. Where (A) represents the scores plot of normal ectocervical and endocervical cells and (B) represents the scores plot of abnormal ectocervical and normal endocervical cells.

There is a separation between the spectra of normal ectocervical and endocervical cells and between abnormal ectocervical and normal endocervical cells. The separation between normal ectocervical and endocervical cells is more distinct in the 1060-994 cm^{-1} region although the cluster of normal spectra is tighter in the 1096-1062 cm^{-1} region.

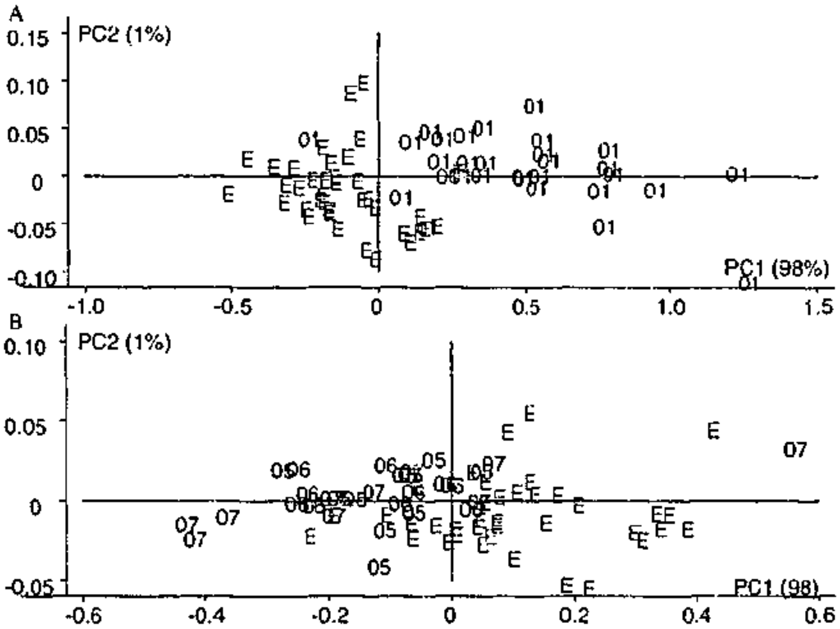

Figure 5.30 PC1 versus PC2 scores plot of data in the region 1096-1062 cm^{-1} showing a separation between normal ectocervical and endocervical cells (A) and between abnormal ectocervical and normal endocervical cells (B). Normal ectocervical cells (green) are represented by 01, abnormal ectocervical cells (blue) by 05, 06 and 07 and normal endocervical cells (red) by E.

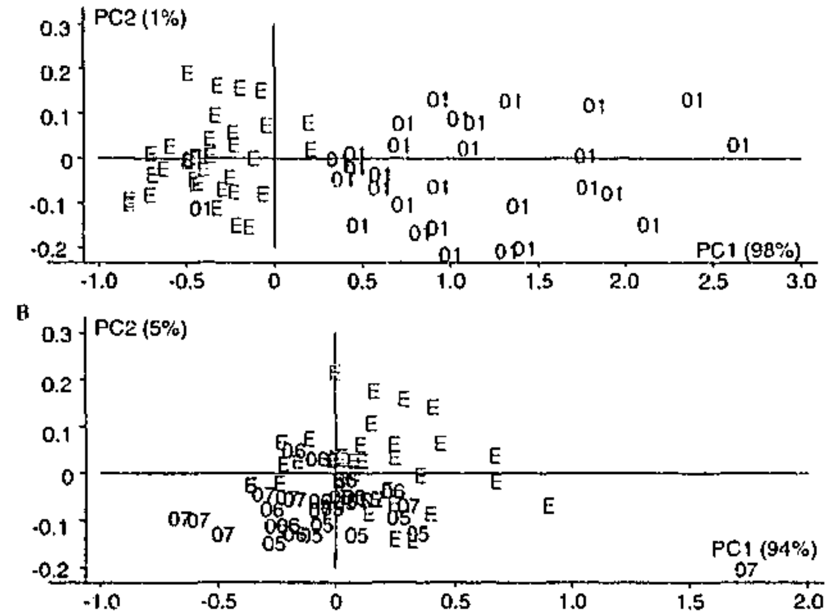

Figure 5.31 PC1 versus PC2 scores plot of data in the region 1060-994 cm^{-1} showing a separation between normal ectocervical and endocervical cells (A) and between abnormal ectocervical and normal endocervical cells (B). Normal ectocervical cells (green) are represented by 01, abnormal ectocervical cells (blue) by 05, 06 and 07 and normal endocervical cells (red) by E.

PCA in the regions 1096-1062 cm^{-1} and 1060-994 cm^{-1} could be employed as a means of identifying and isolating IR spectra of cervical smears which exhibit contributions from endocervical cells. Whilst IR spectra of endocervical cells could potentially confound spectroscopic diagnosis of cervical smears, their presence in smears is important for ensuring a representative sample. As was discussed in Section 5.3, a proportion of cervical smears obtained using only an Ayre spatula will contain the presence of endocervical cells. Since it is difficult and undesirable to sample only ectocervical cells, one has to rely on multivariate statistical tools to extract information about cell populations present in the data and the influence they impart on models formed. It would be advisable to utilise PCA in these regions prior to spectral averaging to identify and remove spectra with major contributions from endocervical cells. The fact that cervical lesions are predominantly of the squamous (ectocervical) cell variety [27] also supports the removal of endocervical cells once the smear has been deemed representative.

5.4.2.2 *INFLAMMATION*

The spectral profiles arising from cervical smears diagnosed as exhibiting an inflammatory response[25] are not necessarily comprised of contributions from white blood cells alone, but are rather a composite of different cell types and a reflection of cell populations present in the cervix at the time of sampling. Figure 5.32 shows the spectral differences exhibited between the averaged spectra of normal and abnormal ectocervical cells and cervical smears with an inflammatory response.

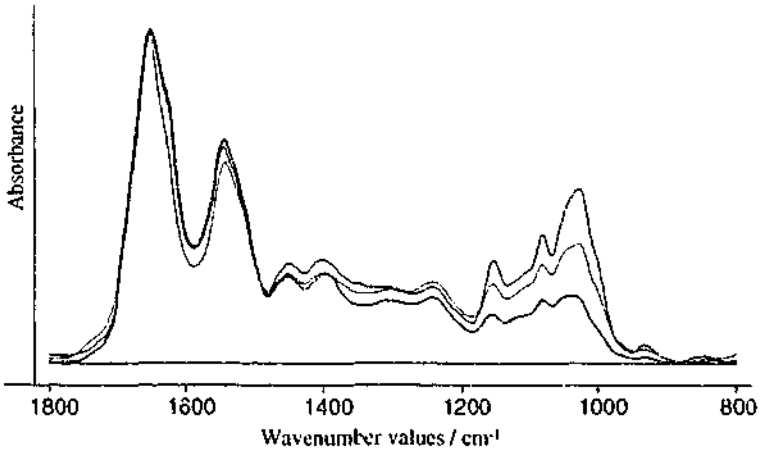

Figure 5.32 Averaged IR spectra of smears containing normal ectocervical (green), abnormal ectocervical (blue) cells and inflammatory effects (red).

As was noted in the comparison of endocervical cells with normal and abnormal ectocervical cells, spectra of inflamed cervical smears exhibit spectral differences in the phosphate and carbohydrate regions. The IR spectrum of the inflamed cervical smear does not exhibit any of the characteristic features of leukocytes described in our previous study [17]. Leukocytes, a type of white blood cell, exhibit pronounced $\nu_{as}PO_2^-$ and $\nu_sPO_2^-$ bands at 1240 and 1078 cm^{-1} respectively and a reduction in glycogen band intensity at 1024 and 1050 cm^{-1} compared with normal ectocervical cells. Contributions from inflammation depend on severity and the population of cell types sampled when obtaining the infrared spectrum.

PCA was employed as a means of discriminating cervical smears with inflammation from normal and abnormal diagnosed cervical smears. PCA was performed on data in the region 1800-800 cm^{-1} as well as in the regions 1096-1062 cm^{-1} and 1060-994 cm^{-1}. The PC1 versus PC2 scores plot for the entire wavenumber region is presented in Figure 5.33.

---

[25] Inflammatory cells include lymphocytes, PMNs, macrophages and plasma cells.

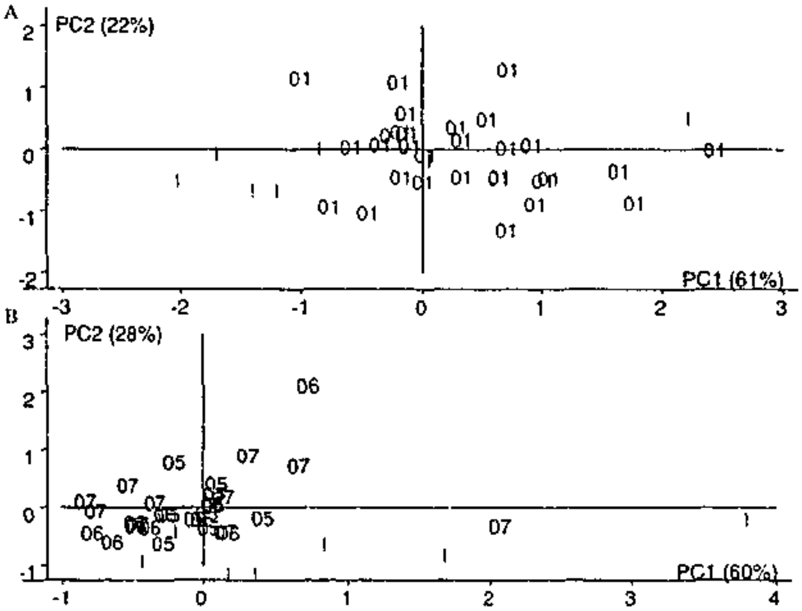

Figure 5.33 PC1 versus PC2 scores plot showing a separation between normal (green) diagnosed cervical smears (A) and abnormal (blue) diagnosed smears (B) with inflammation (red) in the region 1800-800 cm^{-1}.

Whilst there appears to be discrimination between the two groups in each plot, it is difficult to ascertain the extent of separation between the groups given the small number of samples diagnosed with inflammatory effects.

Figure 5.34 and Figure 5.35 show the principal component scores plots in the regions 1096-1062 cm^{-1} and 1060-994 cm^{-1} respectively. Tighter clusters are formed of the groups by reducing the number of variables, although variable reduction does not enhance the discrimination between normal and inflammation or abnormal and inflammation. Other regions were investigated and offered no benefit in terms of increased separation or overlap reduction. More samples of spectra exhibiting inflammatory effects are needed before the benefits of PCA in the identification and removal of inflammatory effects can be determined.

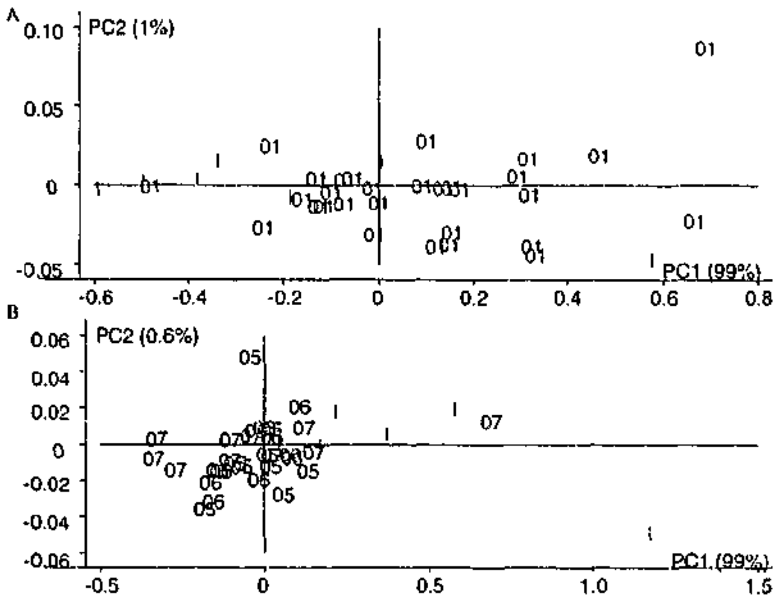

Figure 5.34 PC1 versus PC2 scores plot showing a separation between normal (green) diagnosed cervical smears (A) and abnormal (blue) diagnosed smears (B) with inflammation (red) in the region 1096-1062 cm^{-1}.

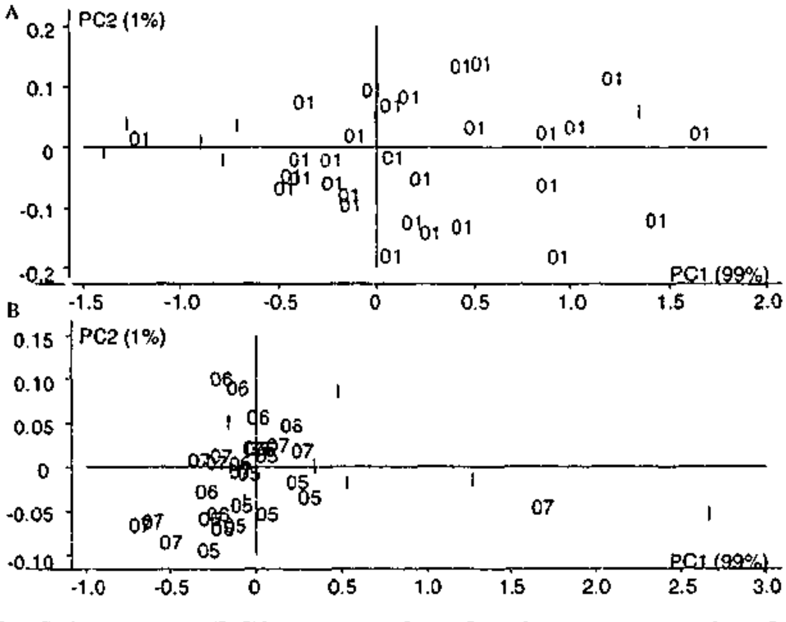

Figure 5.35 PC1 versus PC2 scores plot showing a separation between normal (green) diagnosed cervical smears (A) and abnormal (blue) diagnosed smears (B) with inflammation (red) in the region 1060-994 cm^{-1}.

The number of PMNs in a cervical smear cannot be correlated with inflammation, rather it is reflective of the menstrual cycle [28]. Cervical smears obtained when estrogen levels are high are "clean", whereas smears obtained after ovulation are accompanied by mucus, large numbers of leukocytes and plasma cells [29]. Table 5.12 summarises the predominant cell type and presence of leukocytes in relation to hormonal stimulation.

Table 5.12 Presence of leukocytes in cervical smears in relation to hormonal stimulation.

Phase	Predominant cell component	Other cell components
Menstrual phase	Erythrocytes	Degenerate endometrial cells and leukocytes
Early proliferative phase	Intermediate cells	Occasional endometrial cells and/or histiocytes
Late proliferative phase (ovulatory phase)	Superficial cells	Clean
Early secretory phase	Superficial cells	Some intermediate cells and leukocytes
Late secretory phase	Intermediate cells	Abundant mucus and leukocytes, cytoplasmic degeneration
Pregnancy	Intermediate cells of "navicular type"	
Menopause	Intermediate and parabasal cells	

Aside from the presence of PMNs associated with an inflammatory response to trauma or infection, inflammation causes changes in the cellular epithelium. Cervical smears associated with acute inflammation are characterised by the presence of neutrophilic leukocytes, degenerative or necrotic cells, and cellular debris, usually of intermediate or superficial maturation [27]. Cellular changes associated with chronic inflammation mimic malignancy and the diagnosis of benign relies on the N/C ratio [28]. Inflammation can also cause non hormonal maturation of the squamous epithelium [30].

Given the success in reducing the presence of white blood cells from cervical smears reported in Section 5.3 it may not be necessary to employ multivariate statistics to identify smears exhibiting inflammation.

5.4.2.3 CANDIDA ALBICANS

Candida albicans or thrush, as it is often referred, is a yeast infection common in the cervices of females. Infection occurs when there is a change in the balance of normal vaginal flora [15]. The IR spectrum of this microorganism has been presented elsewhere [17]. Spectra are characterised by an intense, broad band in the polysaccharide region (1150-900 cm^{-1}), and pronounced $\nu_{as}PO_2^-$ and δCH_3 bands. The averaged spectrum of all *Candida* diagnosed samples presented in Figure 5.36 does not exhibit any of these characteristic features. The IR spectrum of a sample is an averaged reflection of the cell types present in the cervix, and so the spectral features of *Candida* diagnosed samples will depend on the severity of the infection.

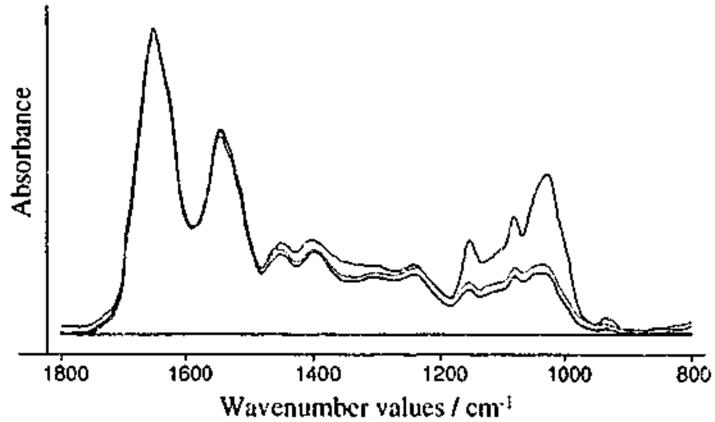

Figure 5.36 Averaged IR spectra of normal (green), abnormal (blue) and *Candida* (red) diagnosed cervical smears.

PCA was performed on the data in the regions described previously (1800-800 cm^{-1}, 1096-1062 cm^{-1} and 1060-994 cm^{-1}). The principal component scores plot for each analysis is presented in Figure 5.37, Figure 5.38 and Figure 5.39.

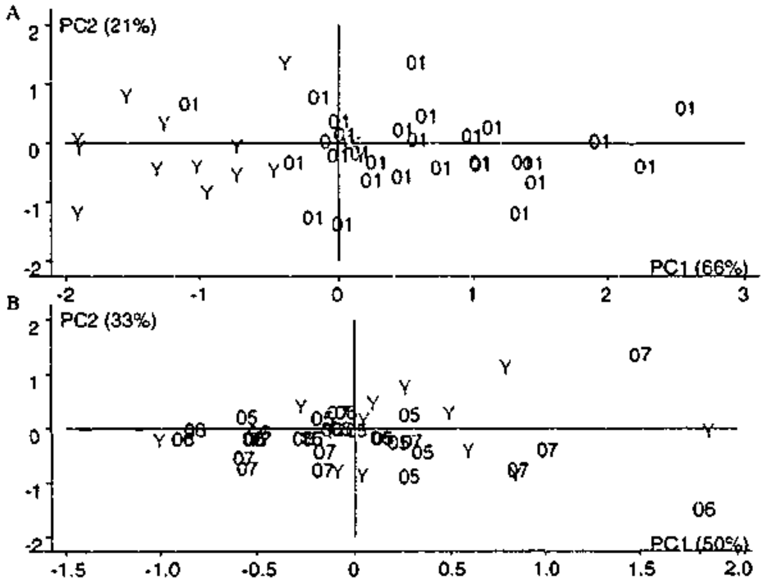

Figure 5.37 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with *Candida* (red) diagnosed smears in the 1800-800 cm^{-1} region.

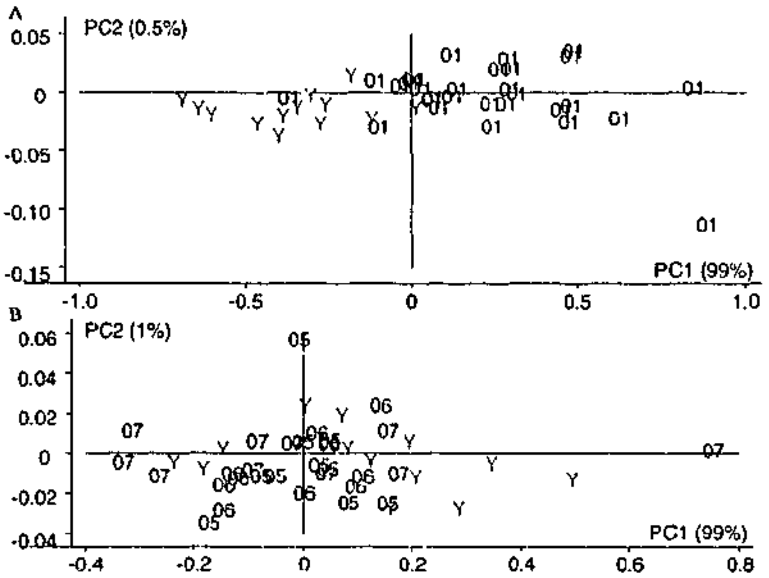

Figure 5.38 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with *Candida* (red) diagnosed smears in the 1096-1062 cm^{-1} region.

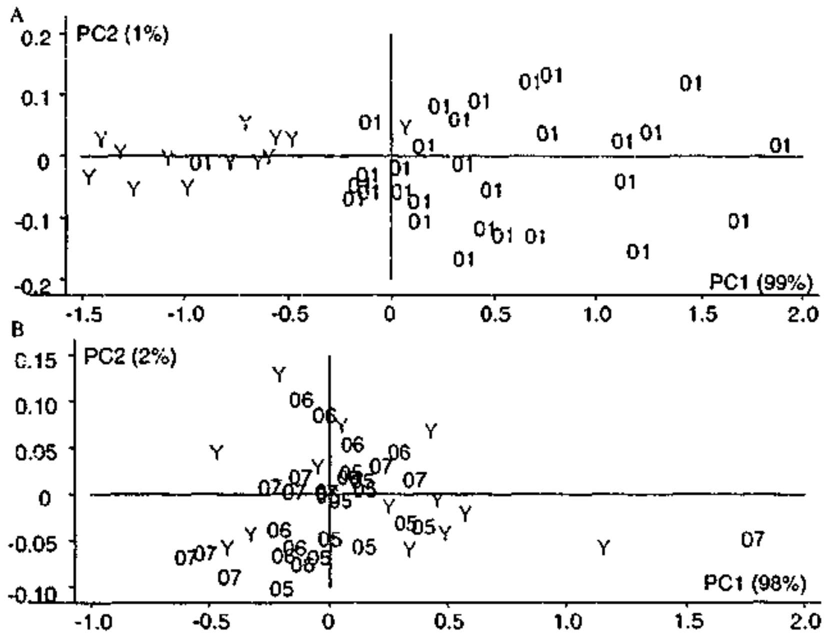

Figure 5.39 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with *Candida* (red) diagnosed smears in the 1060-994 cm^{-1} region.

Candida diagnosed samples, denoted Y, were separated from normal diagnosed samples in all three scores plots. Reducing the number of variables resulted in the formation of tighter clusters of individual groups but did not enhance the separation. It is important for the specificity and sensitivity of a technique that infection, inflammation and cellular changes not related to dysplasia or malignancy can be correctly identified. It is necessary to have distinction of benign changes from both normal and abnormal samples. The separation between *Candida* and normal diagnosed samples will result in high specificity, however the inability of PCA to separate *Candida* from abnormal diagnosed samples, in any of the spectral regions investigated, will result in low sensitivity.

Even if the presence of *Candida* is not high enough to warrant the spectral changes noted previously [17] the cellular changes brought about by the microorganism may explain the separation of IR spectra of these samples from normal diagnosed samples. Cellular changes, with and without inflammation, include nuclear enlargement and perinuclear clearing as seen with HPV [15].

5.4.2.4 BACTERIAL VAGINOSIS

Bacterial vaginosis is one of the few infections of the female genital tract not associated with an inflammatory response [15]. Bacterial vaginosis arises from an overgrowth of vaginal flora, including *Gardineralla vaginalis* and anaerobic *streps.* that thrive in conditions of pH 7-7.5[26]. These bacteria cause slight nuclear enlargement of cells.

The averaged IR spectra of normal and abnormal smears, and smears diagnosed with bacterial vaginosis are shown in Figure 5.40. Once again spectroscopic changes between the groups are exhibited in the phosphodiester and carbohydrate regions (1250-1000 cm^{-1}).

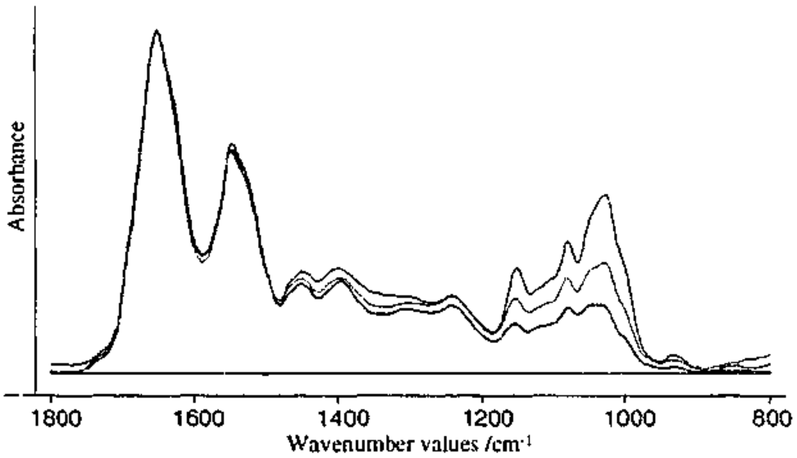

Figure 5.40 Averaged IR spectra of normal (green) and abnormal (blue) diagnosed cervical smears with smears diagnosed with bacterial vaginosis (red).

---

[26] The natural pH of the vagina and ectocervix is ~ 4.

PCA was performed on samples diagnosed with bacterial vaginosis in the regions chosen for the other BCCs investigated. Figure 5.41, Figure 5.42 and Figure 5.43 show the principal component scores plots of the entire spectral region (1800-800 cm^{-1}) and reduced spectral regions (1096-1062 cm^{-1} and 1060-994 cm^{-1}) respectively. Slight overlap is noted between bacterial vaginosis and the comparative normal and abnormal groups, but overall PCA is able to produce separation of the groups in all wavenumber value regions investigated.

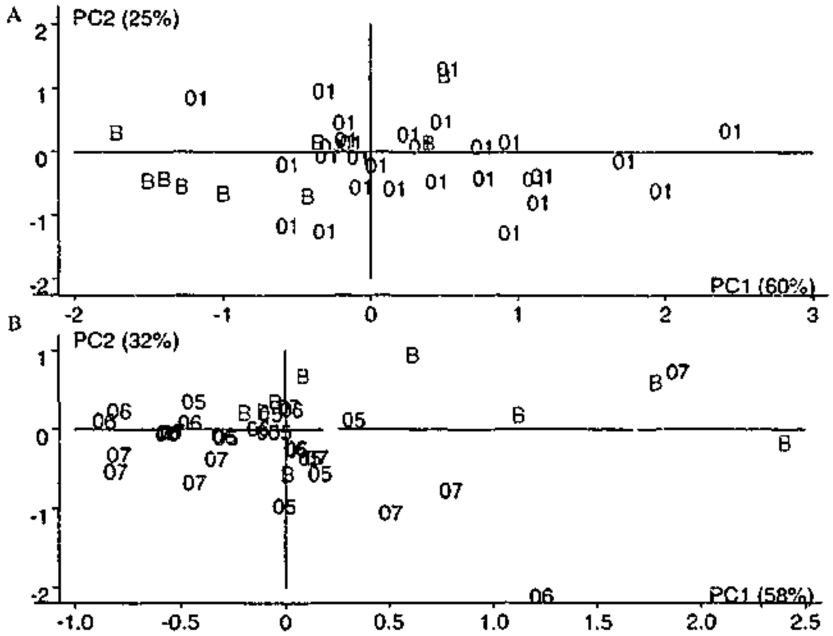

Figure 5.41 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with bacterial vaginosis (red) diagnosed smears in the 1800-800 cm^{-1} region.

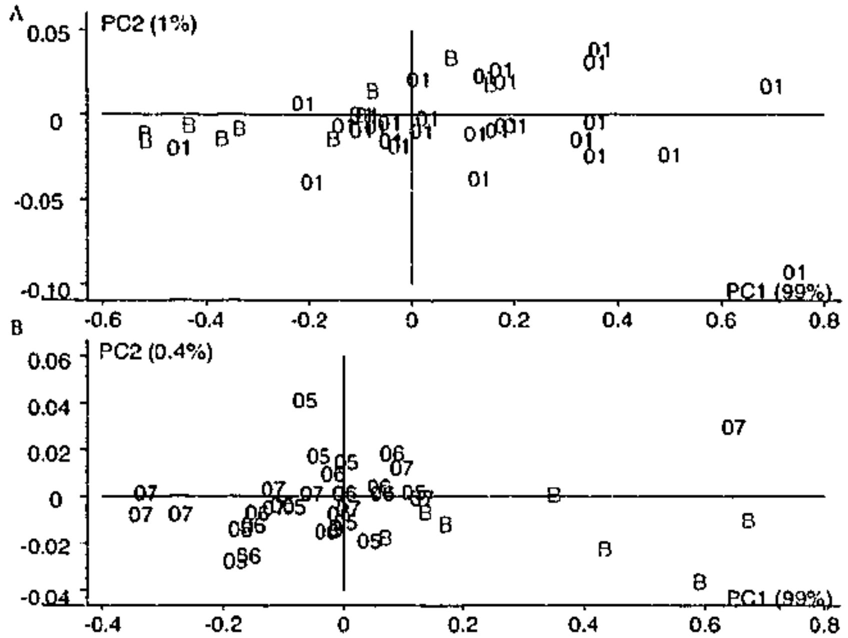

Figure 5.42 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with bacterial vaginosis (red) diagnosed smears in the 1096-1062 cm^{-1} region.

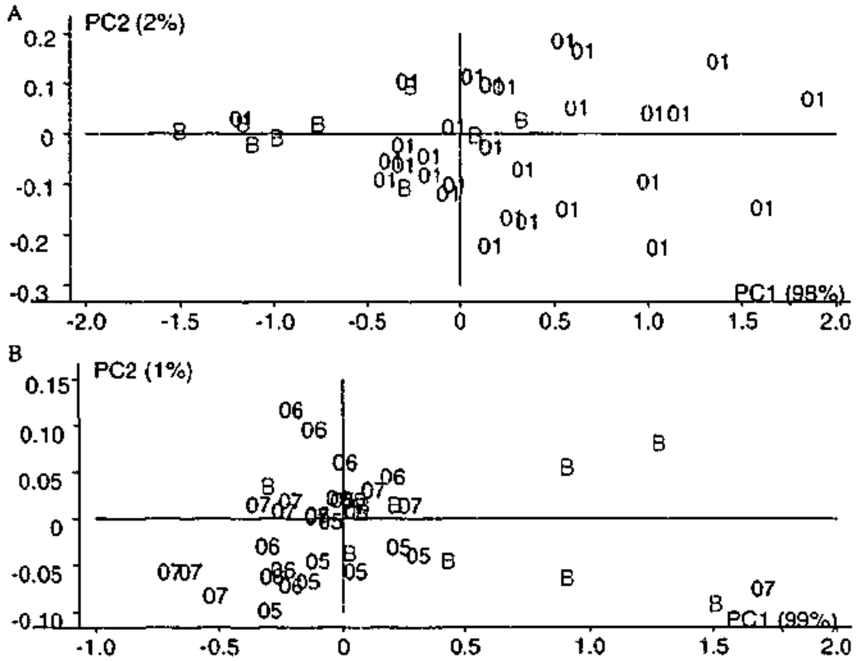

Figure 5.43 PC1 versus PC2 scores plot showing a separation of normal (green) diagnosed smears (A) and abnormal (blue) diagnosed smears (B) with bacterial vaginosis (red) diagnosed smears in the 1060-994 cm^{-1} region.

5.4.2.5 OTHER BENIGN CELLULAR CHANGES

Spectroscopic influences of atypia, metaplasia and keratinisation could not be fully investigated due to insufficient samples of each type. Figure 5.44 shows the principal component scores plot of normal cervical smears and other BCC diagnosed smears. More

IR spectra of these cellular changes are needed to determine the influence on spectroscopic diagnosis.

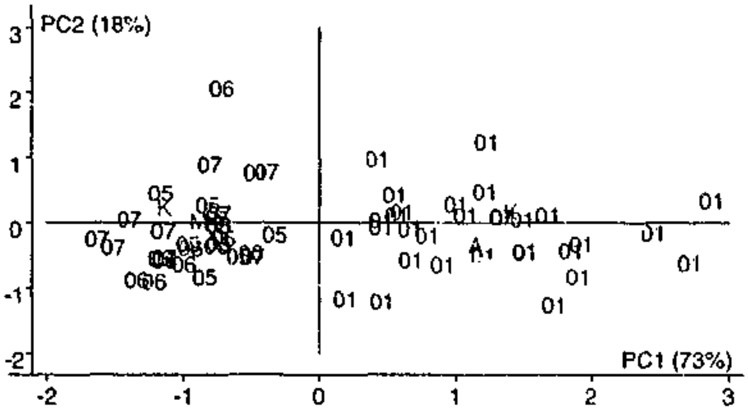

Figure 5.44 PC1 versus PC2 scores plot of normal (green) cervical smears and smears diagnosed with metaplasia (M), atypia (A) and keratinisation (K), red.

5.4.3 CONCLUSION

Infrared spectroscopy and principal component analysis of endocervical cells, and smears diagnosed with benign cellular changes were investigated to determine their influence as potential confounding variables in the diagnosis of cervical cancer.

Spectral differences in all cell and diagnostic types investigated were found in the phosphodiester and carbohydrate regions. Spectral differences in other bands were not distinct enough to allow differentiation between groups.

PCA was successfully used to obtain a separation of normal ectocervical smears from normal endocervical cells, and smears diagnosed with inflammation, *Candida*, and bacterial vaginosis in the regions 1800-800 cm^{-1}, 1096-1062 cm^{-1} and 1060-994 cm^{-1}). PCA obtained a separation with slight overlap of abnormal ectocervical smears from normal endocervical cells, inflammation, and bacterial vaginosis. *Candida* was not separated from abnormal ectocervical smears with any success.

Insufficient numbers of samples representing metaplasia, atypia and keratinisation prevented thorough statistical analysis of spectroscopic differences brought about by cellular changes exhibited in these groups.

5.5 *REFERENCES*

1. Mango, L., Computer-assisted cervical cancer screening using neural networks. *Cancer Letters*, 1994. **77**: p. 155-162.
2. Wong, P., *et al.*, Infrared spectroscopy of exfoliated human cervical cells: Evidence of extensive structural changes during carcinogenesis. *Proc. Natl. Acad. Sci.*, 1991. **88**: p. 10988-10992.
3. Wong, P., R. Wong, and M.F.K. Fung, Pressure-tuning FT-IR study of human cervical tissues. *Applied Spectroscopy*, 1993. **47**(7): p. 1058-1063.
4. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix. *Biospectroscopy*, 1998. **4**(1): p. 47-53.
5. Shaw, R.A., *et al.*, Analysis of biomedical spectra and images: from data to diagnosis. *Journal of Molecular Structure*, 2000. **500**: p. 129-138.
6. Wood, B., *et al.*, An investigation into FTIR spectroscopy as a biodiagnostic tool for cervical cancer. *Biospectroscopy*, 1996. **2**: p. 1-11.
7. *The Unscrambler User Manual*. CAMO ASA: Sweden. 1998.
8. Romeo, M., *et al.*, Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. *Cellular and Molecular Biology*, 1998. **44**(1): p. 179-187.
9. Cohenford, M., *et al.*, Infrared spectroscopy of normal and abnormal cervical smears: Evaluation by principal component analysis. *Gynecologic Oncology*, 1997. **66**: p. 59-65.
10. Spitzer, M., Cervical screening adjuncts: recent advances. *Am. J. Obstet. Gynecol.*, 1998. **179**(2): p. 544-556.
11. Gay, D., L. Donaldson, and J. Goellner, False-negative results in cervical cytologic studies. *Acta Cytologica*, 1985. **29**(6): p. 1043-1046.
12. Morell, N., *et al.*, False-negative cytology rates in patients in whom invasive cervical cancer subsequently developed. *Obstet. Gynecol.*, 1982. **60**: p. 41-45.

13. van der Graaf, Y., *et al.*, Screening errors in cervical cytologic screening. *Acta Cytologica*, 1987. **31**(4): p. 434-438.

14. van der Graaf, Y., and G. Vooijs, False negative rate in cervical cytology. *J. Clin. Pathol.*, 1987. **40**: p. 438-442.

15. Riech, D., Cytopathologist, Victorian Cytology Service (Melbourne, Australia). 2000, Personal communication.

16. Cohenford, M. and B. Rigas, Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: Implications for tumor biology. *Proc. Natl. Acad. Sci.*, 1998. **95**: p. 15327-15332.

17. Wood, B., *et al.*, FTIR microspectroscopic study of cell types and potential confounding variables in screening for cervical malignancies. *Biospectroscopy*, 1998. **4**(2): p. 75-91.

18. Shaw, R.A., *et al.*, Infrared spectroscopy of exfoliated cervical cell specimens: Proceed with caution. *Analytical and Quantitative Cytology and Histology*, 1999. **21**(4): p. 292-302.

19. Wong, P., *et al.*, Characterization of exfoliated cells and tissues from human endocervix and ectocervix by FTIR and ATR/FTIR spectroscopy. *Biospectroscopy*, 1995. **1**: p. 357-364.

20. Yazdi, H., M. Bertrand, and P. Wong, Detecting structural changes at the molecular level with Fourier transform infrared spectroscopy. *Acta Cytologica*, 1996. **40**(4): p. 664-668.

21. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. III. Spectral differences between squamous and columnar tissue and cells from the human cervix. *Biospectroscopy*, 1997. **3**(4): p. 253-257.

22. Chiriboga, L., *et al.*, Infrared spectroscopy of human tissue. II. A comparative study of spectra of biopsies of cervical squamous epithelium and of exfoliated cervical cells. *Biospectroscopy*, 1998. **4**(1): p. 55-59.

23. Davajan, V., R. Nakamura, and K. Kharma, Spermatozoan transport in cervical mucus. *Obstet. and Gynecol. Surv.*, 1970. **25**: p. 1-43.

24. Wood, B.R., *Biomedical Applications of Fourier Transform-Infrared Microspectroscopy*. PhD Dissertation, in Chemistry. Monash University. 1998.

25. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 2 ed. Philadelphia: J B Lippincott Company. 1968.

26. Kurman, R. and D. Solomon, *The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses*. New York: Springer-Verlag. 1994.

27. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 3 ed. Philadelphia: J B Lippincott Company. 1979.

28. Takahashi, M., *Color Atlas of Cancer Cytology*. 2 ed. Tokyo: IGAKU-SHOIN Medical Publishers, Inc. 1981.

29. Riotton, G. and W. Christopherson, *Cytology of the Female Genital Tract*. Vol. 8. Geneva: World Health Organisation. 1973.

30. DeMay, R., *The Art and Science of Cytopathology*. Vol. 1. Chicago: American Society of Clinical Pathology Press. 1995.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

6 CONCLUSIONS AND FUTURE DIRECTIONS

The results of this dissertation pose some very interesting questions as to the influence of confounding variables and non-diagnostic debris in the discrimination between normal, dysplastic and malignant cervical cells. It was found that changes in squamous epithelial cells brought about by hormonal stimulation of ovulating women could be monitored spectroscopically. PCA demonstrated that despite cyclical differences, the infrared spectra of these cells could be separated from those of dysplastic and malignant samples.

Multivariate statistical techniques were able to differentiate between the infrared spectra of normal and abnormal cells on pre-selected data with histological and cytological agreement, excluding the effects of inflammation, bacterial infection, and benign cellular changes. However, if this technique is to be introduced for routine screening methodologies capable of handling this degree of variation need to be developed.

The problem of confounding cells and non-diagnostic debris in cervical smears needs to be addressed, and techniques that help reduce the presence of these spectroscopic contaminants need to be further investigated. The results from the clean-up trial indicate that it is possible to remove inflammatory cells from cervical smears via chemical methods without affecting the integrity of epithelial cells. Although the removal of blood components from cervical samples was successful, the resulting samples were separated with normal by PCA. Whilst yet to be the standard procedure, it should be noted that use of the ThinPrep® processor in routine screening of cervical smears by cytologists is becoming more prevalent. A major limiting factor preventing its widespread use is the current prohibitive cost of the process. However with increased awareness of its benefits in the future it will become the standard procedure. Alternative methods to ThinPrep® to remove non-diagnostic debris should be investigated to determine if there are less expensive alternatives.

To further reduce the influence of PMNs, cervical smears could be obtained at a particular time of the ovarian cycle when the presence of PMNs is low and cervical cells are considered "clean". This would correspond to the late follicular phase, prior to the onset of ovulation when the influence of estrogen provides an environment for the maturation of squamous epithelial cells and reduced presence of PMNs.

PCA was successfully utilised to separate normal and abnormal ectocervical cells from normal endocervical cells and samples diagnosed with some benign cellular changes. The similarities noted between IR spectra of abnormal, inflamed and samples diagnosed with benign cellular changes raise the issue of what is being detected by spectroscopy. Are there spectroscopic differences between normal and abnormal cervical samples, or are differences due to non-specific changes occurring in the cell?

Spectral changes exhibited in the phosphodiester region as a result of maturation or abnormality should be further investigated. The presence and intensity of characteristic nucleic acid bands could be used as a marker for nuclear size, which reflects the maturity, health and vitality of the cell.

It would be useful to compare the infrared spectra of inflammatory responses brought about by infection such as abnormality or bacteria with the presence of PMNs in the cervix due to hormonal influences. This would require sampling both pre- and post-menopausal women, as post-menopausal women would not show signs of PMNs unless there was an inflammatory response.

This work was carried out using IR microspectroscopy and as such time constraints only permitted collection of spectra from random areas on the deposit, rather than recording spectra of the entire deposit. Infrared imaging systems are available that record spectra of 4096 pixels on a sample. If the whole sample deposit could be analysed, this would minimise the risk of missing abnormal cells, and spectral regions of different areas on the deposit could be ratioed as an objective means of determining the best spectrum to use for analysis. The employment of imaging spectroscopy would also help reduce the problem of introducing inaccuracies into the diagnostic method as thin sections of biopsied tissue could be used for imaging, giving a direct correlation with the diagnosis and sample site.

Finally, techniques such as flow cytometry that are capable of separating particles according to size and density should be investigated. The presence of confounding variables presents an obstacle in the development of spectroscopy in the diagnosis of cervical cancer, and populations of cells exhibiting various stages of maturation, as well as different types of epithelial cell may also cause similar problems. If more homogenous cervical cell deposits can be made, then this technique will clearly be more successful than the present techniques that lack homogeniety. It is hoped that this will facilitate increased success in the diagnosis of cervical cancer.

APPENDICES

APPENDIX A

GLOSSARY

aetiology:	study of causes, especially inquiry into the origin of disease.
chromatin:	chromatin is the complex formed from the electrostatic interaction between DNA and histone [1]. Chromatin granules are the precursors of chromosomes [2].
chromosomes:	are made up of genes which are the hereditary material determining the organism's characteristics [2].
CIN:	or cervical intraepithelial neoplasia. This term was introduced to enable a uniform nomenclature of precancerous legions, regardless of histologic type, requiring colposcopy and biopsy evaluation and treatment [3]. CIN are graded I to III according to severity, with CIN III being the most severe type.
colposcopy:	a colposcope is a microscope used for looking at the cervix and the inside of the vagina in detail.
cryotherapy:	uses a cryoprobe to rapidly remove heat from the tissue to produce controlled destruction of premalignant epithelial lesions [4].
cytolysis:	is the dissolution of the cytoplasm and is frequently associated with *Lactobacillus vaginalis*, which thrives on cells with high glycogen content [5].
cytoplasmic halos:	characteristic sign for the presence of HPV infection. A noticeable clearing around the nucleus is seen (refer to Appendix B).
differentiation:	process of morphological and functional specialisation of cells [5].
diploid:	a diploid cell contains two copies of each chromosome [1].
epidemiology:	the study of the incidence and distribution of epidemic disease in a community.
eukaryotic:	term for cells of organisms including multicellular plants, animals, fungi and some unicellular organisms.

external os:	squamocolumnar junction.
histones:	the major class of DNA-binding protein of chromatin [1].
hyperchromatic:	relates to staining pattern of cells. Hyperchromatic cells are so termed because they have a darker staining pattern.
hyperplasia:	is the abnormal proliferation of benign cells within a tissue and is associated with an absolute increase in amount of tissue [6].
isoelectric point:	the pH at which the net charge on an ampholyte (in this case amino acids and proteins) is zero [1].
LEEP:	or loop electrosurgical excision procedure. An electrical current is passed through a thin wire loop that acts as a knife to remove abnormal tissue [7].
Lugol's reagent:	aqueous solution of iodine and potassium iodide. Stains cells blue-violet in the presence of glycogen.
necrosis:	localised cell death.
neoplasm:	meaning 'new things formed' (or tumours) is a result of the growth of tissue that has escaped the controls governing normal proliferation and regeneration of cells. The term neoplasm is confined to clearly abnormal proliferation of tissue forming a visible or palpable swelling or tumour affecting part of an organ. Neoplasms may be benign or malignant [6].
nucleoli:	or nucleolus (singular) is the dense body in the nuclear membrane surrounding the nucleus [2] and is necessary for the production of *r*RNA.
N/C ratio:	or nuclear-to-cytoplasmic ratio. Is used to measure the maturation of a cell. The higher the N/C ratio, the less mature the cell. Cellular immaturity, measured by the N/C ratio, reaches its peak at carcinoma *in situ* (completely undifferentiated cells), even though mildly dysplastic cells have the largest nucleus [8].
proliferation:	rapid growth or reproduction.

pyknotic nuclei:	a nucleus is defined as pyknotic when it is less than 5 μm in diameter [9].
tetraploid:	a tetraploid cell contains four copies of each chromosome.
tumour diathesis:	comprised of protein, fibrin, blood, dead cells, and debris [8], and is cytologically indicative of invasive carcinoma.

APPENDIX B

LIGHT MICROSCOPY OF CERVICAL SMEARS[27]

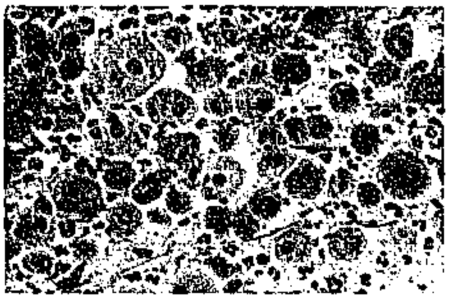

Figure B.1 Parabasal squamous epithelial cells

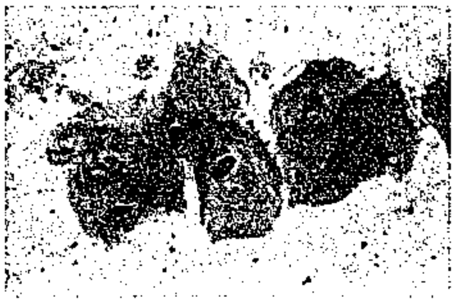

Figure B.2 Intermediate squamous epithelial cells

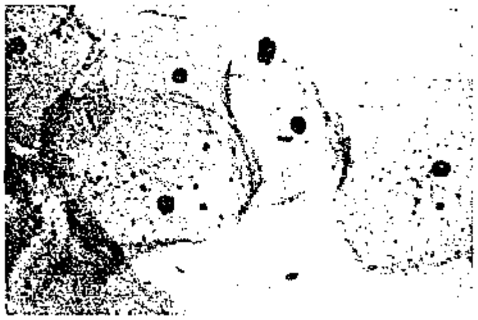

Figure B.3 Superficial squamous epithelial cells

---

[27] Courtesy of the Victorian Cytology Service (Melbourne Australia)

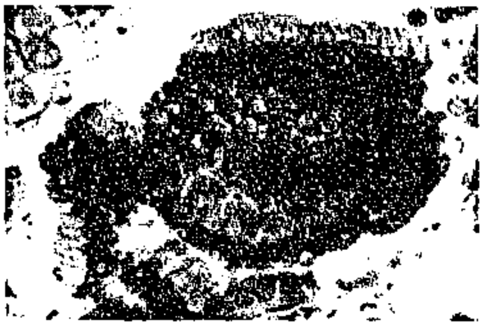

Figure B.4 Endocervical columnar epithelial cells

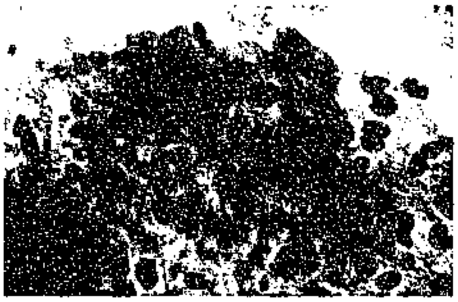

Figure B.5 CIN I (mild dysplasia)

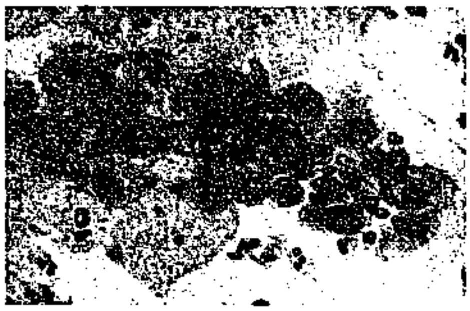

Figure B.6 CIN II (moderate dysplasia)

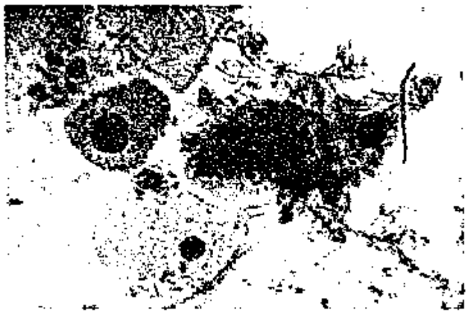

Figure B.7 CIN II (moderate dysplasia) with bacterial infection (seen by the rod-like structures in the cells.

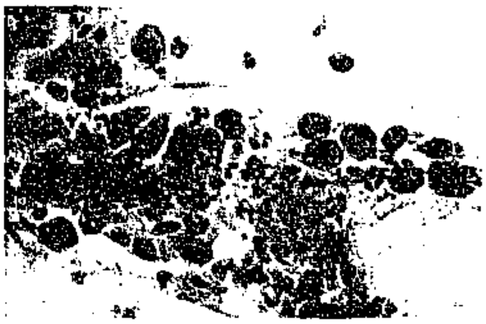

Figure B.8 CIN III (severe dysplasia)

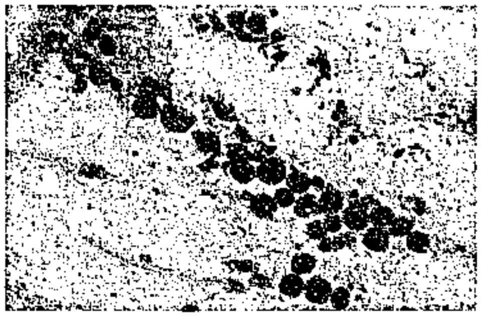

Figure B.9 CIS (carcinoma *in situ*)

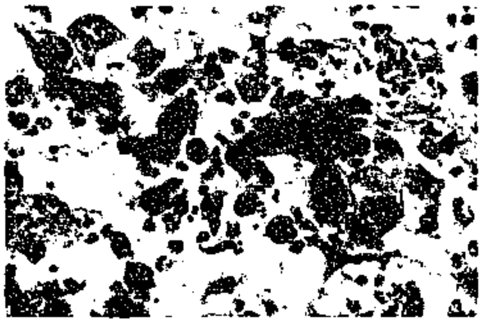

Figure B.10 Invasive SCC (squamous cell carcinoma)

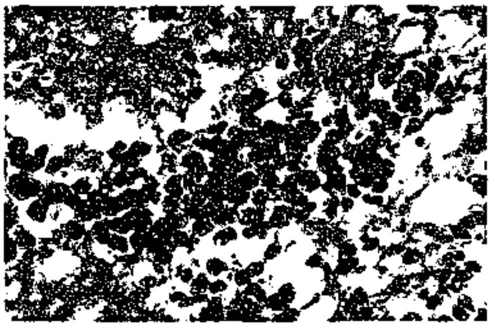

Figure B.11 Invasive undifferentiated carcinoma

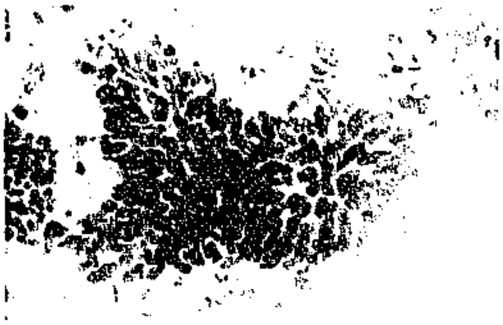

Figure B.12 Adenocarcinoma

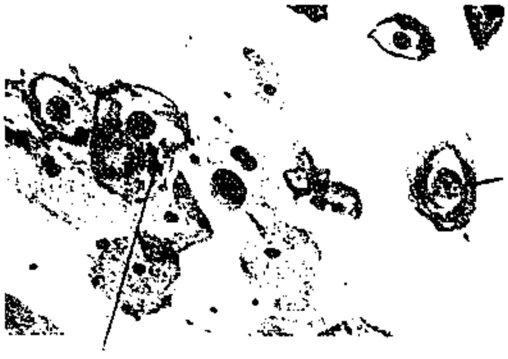

Figure B.13 Koilocytes caused by HPV

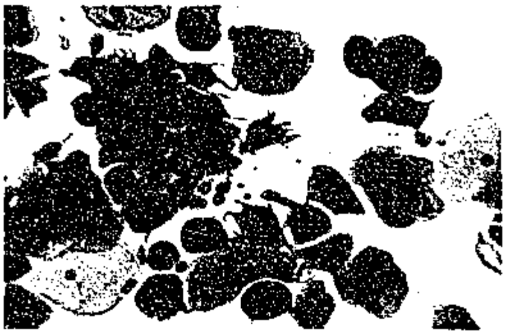

Figure B.14 Squamous metaplasia

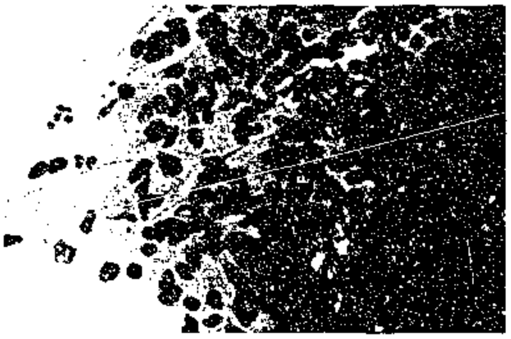

Figure B.15 Atrophy

APPENDIX C

GAUSSIAN DISTRIBUTION

The Gaussian or Normal distribution function is the most important distribution for continuous data due to its wide range of practical applications. The distribution approximates most measurements of physical characteristics, with their associated random errors and natural variations. The shape of this function, referred to as the Gaussian probability curve, is illustrated in Figure C. 1. For a single measurement variable, x, the mathematical model describing the Gaussian distribution function is:

$$f(x)=\frac{1}{\sigma\sqrt{(2\pi)}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \qquad \textbf{Equation C.1}$$

The curve is symmetric about the mean, μ; and the variance, σ^2, or standard deviation,σ, explain the spread about the mean. The curve is often standardised so that the area is equal to unity, and $f(x)$, the height of the curve gives the probability of observing a value within the specified ranges of x. The idealised distribution function is obtained from an infinite number of samples, called a parent population. As it is difficult to sample the entire population, a finite number of samples, n, are used to give an estimate of the mean, variance and standard deviation, denoted $\bar{x}$, s^2 and s respectively [10].

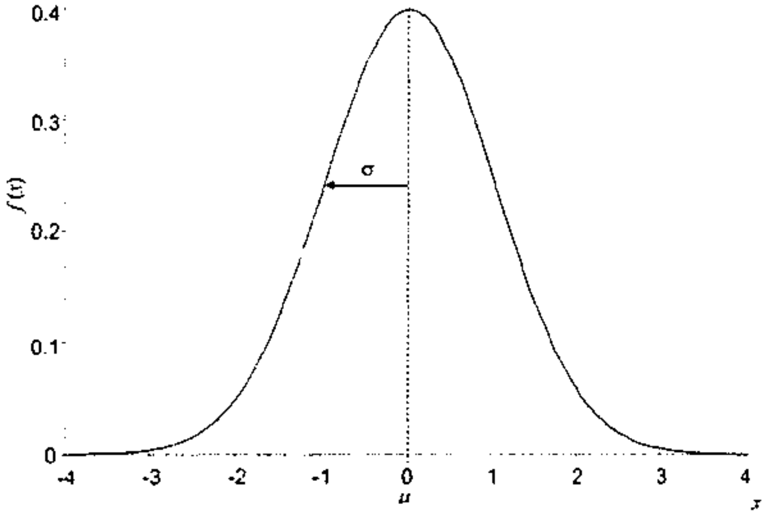

Figure C. 1 Standardised Gaussian probability curve.

APPENDIX D

MATLAB ROUTINES

CERVJOIN.M

```
% General Cervical DataBase Front End

% Author: Melissa Romeo, Monash University, 1999

clc;

clear all;

close all;

%------------- Get the Data -----------------------
Data=menu('Source of Data','Load existing DataBase','Start DataBase',...
'Add to existing DataBase');

if Data==1;
  WhichDB=menu('Please select Database','RWH','FPV','Other');

  if WhichDB==1;
    load DBRWH;

  elseif WhichDB==2;
    load DBFPV;

  elseif WhichDB==3;
    other=menu('Please select DataBase','RWH','FPV');
    if other==1
      load naallav1b;
    elseif other==2
      load fpv8blred;
    end
  end


elseif Data==2
  [DB,Filenames,WaveMatrix]=jcampdbjoin;

elseif Data==3
  [DB,Filenames,WaveMatrix]=jcampcervjoin;

end
```

```
%----------- Preprocessing ----------------------
c=0;
whilst c<1;
  Prepro=menu('Preprocessing Functions','Define Spectral Region',...
    'Account for non-linearities of the MCT','Calculate Signal-to-noise Ratio',...
    'Take Derivatives (Savitzky-Golay)','Normalise','Baseline Correction',...
    'Plot Data','Assign Diagnosis','View DataBase Details','Quit');

  if Prepro==1
    [DB,WaveMatrix]=defregjoin(DB,WaveMatrix);

        elseif Prepro==2
    [DB,Filenames]=nonlinjoin(DB,Filenames);

  elseif Prepro==3
    [SNR,DB,Filenames,SNRFN]=SNRjoin(DB,WaveMatrix,Filenames);

  elseif Prepro==4
        [DB]=svgljoin(DB);

        elseif Prepro==5
    norm=menu('What type of Normalisation?','Vector Normalisation','Max
Normalisation',...
      'Mean Normalisation','Range Normalisation');
    if norm==1
      [DB]=vnormjoin(DB);

    elseif norm==2
      [DB]=maxnormjoin(DB);

    elseif norm==3
      [DB]=meannormjoin(DB);

    elseif norm==4
      [DB]=rangenormjoin(DB);

    end

  elseif Prepro==6
    [DB,WaveMatrix]=bljoin(DB,WaveMatrix);

  elseif Prepro==7
    plot(WaveMatrix,DB);

  elseif Prepro==8
    Diag=menu('Which database do you want to perform diagnosis on?','RWH','FPV');
    if Diag==1;
      [DBXSpec,DCYSpec,needBX,needCY]=DiagRWHCY(DB,Filenames);
      %OrigDB=DB;
      %DB=DBXSpec(:,1:1475);
      %Filenarnes=DBXSpec(:,1476);
    elseif Diag==2;
      [DBFPV,DiagFPV,needFPV]=DiagnosisFPV(DB,Filenames);
    end

  elseif Prepro==9
    whos

  elseif Prepro==10
    c=c+1;
    break
  end
end

pack
%save Data1
```

JCAMPDBJOIN.M

```
function [DB,Filenames,WaveMatrix]=jcampdbjoin
% This routine is designed to import JCAMP.DX infrared spectral files and saves the
% results under 'DataBase' where DB is the matrix containing the absorbances for the
% spectral files. Filenames are stored in a matrix called Filenames. To add more samples
% to DataBase, run jcampcervjoin.m.

% Author: Melissa Romeo, Monash University, 1999

cd c:\Matlab\Mjr\Jcampfiles\Cerv;

D=dir;

[N,M]=size(D);

m=0;

c=1;

for IN=2:N;
        File=D(IN).name;

  [PATH,NAME,EXT,VER] = FILEPARTS(File);

  xtn=EXT;
  count=0;
    if strcmp(xtn,'.dx');
      eval('fid1=fopen(File,"r");');
      m=m+1;

      if count==0;
        count=count+1;
        % What is the first wavenumber value?
                        for i=1:10;
                                S=fgetl(fid1);
                            end

                            [FirstX]=sscanf(S,'%s');
                            FirstX=strrep(FirstX,'##FIRSTX=',");
                            str2num(FirstX);
                            [FirstX]=sscanf(FirstX,'%d');

                            %What is the last wavenumber value?
                            S=fgetl(fid1);
                            [LastX]=sscanf(S,'%s');
                            LastX=strrep(LastX,'##LASTX=',");
                            str2num(LastX);
                            [LastX]=sscanf(LastX,'%d');
```

```
                          %What is the delta value?
                          S=fgetl(fid1);
                          [DeltaX]=sscanf(S,'%s');
                          DeltaX=strrep(DeltaX,'##DELTAX=','');
                          str2num(DeltaX);
                          [DeltaX]=sscanf(DeltaX,'%d');

                          %Create a matrix of wavenumber values.
  [WaveMatrix]=FirstX:DeltaX:LastX;

  for i=12:15;
          S=fgetl(fid1);
                          end

 else
  for i=1:15;
          S=fgetl(fid1);
                          end
end

                  %What is the Y-factor?
                  [YFactor]=sscanf(S,'%s');
                  YFactor=strrep(YFactor,'##YFACTOR=','');
                  str2num(YFactor);
                  [YFactor]=sscanf(YFactor,'%f');

                  for i=16:18;
                          S=fgetl(fid1);
                  end

                  %Get the absorbances.
                  k=0;
          whilst 1
          S=fgetl(fid1);
              if strcmp(sscanf(S,'%s'),'##END=');
              break
          end

                          [Output]=sscanf(S '%i');
                          [Output]=[Output]';
                          l=length(Output);

                          for j=2:l;
                                  M1(m,k+j-1)=Output(j)*YFactor;

                          end

                          k=k+l-1;
end
          %What are the sample numbers?
```

```
        [NAME]=sscanf(NAME,'%d');
        Filenames(c)=NAME;
        c=c+1;

      %Close read files
      fclose(fid1);

        end
end
DB=M1
save DataBase DB Filenames WaveMatrix
```

JCAMPCERVJOIN.M

```
function [DB,WaveMatrix,Filenames]=jcampcervjoin
% This program is for importing jcamp files into a pre-existing Matlab array
% (DataBase, DB) containing infrared spectra of cervical cells.

% Author: Melissa Romeo, Monash University, 1999

cd c:\Matlab\Mjr\Jcampfiles\Cerv;
D=dir;
[N,M]=size(D);
m=0;
c=1;
load DataBase;
DB=DB';

for IN=2:N;
    File=D(IN).name;

 [PATH,NAME,EXT,VER] = FILEPARTS(File);

 xtn=EXT;
 count=0;
  if strcmp(xtn,'.DX')
   eval('fid1=fopen(File,"r");')
   m=m+1;
   count=count+1;

    for i=1:16
        S=fgetl(fid1);
            end

            %What is the Y-factor?
            [YFactor]=sscanf(S,'%s');
            YFactor=strrep(YFactor,'##YFACTOR=','');
            str2num(YFactor);
            [YFactor]=sscanf(YFactor,'%f');

            for i=17:19
                S=fgetl(fid1);
            end

            %Get the absorbances.
            k=0;
        whilst 1
        S=fgetl(fid1);
          if strcmp(sscanf(S,'%s'),'##END=');
          break;
        end;
```

```
                                [Output]=sscanf(S,'%i');
                                [Output]=[Output]';
                                l=length(Output);

                                for j=2:l;
                                        M2(m,k+j-1)=Output(j)*YFactor;

                                end

                                k=k+l-1;
          end
          %What are the sample numbers?
          [NAME]=sscanf(NAME,'%d');
          Filenames1(c)=NAME;
          c=c+1;

          %Close files after reading.
          fclose(fid1);

    end

  end
  Filenames=[Filenames Filenames1];
  M2T=M2';
  DB=[DB M2T];
  DB=DB';
  save DataBase DB WaveMatrix Filenames;
```

DEFREGJOIN.M

```
function [DB,WaveMatrix]=defregjoin(DB,WaveMatrix)
% This routine is interactive and allows the user to define regions of the spectrum to
% process further.

% Author: Melissa Romeo, Monash University, 1999

SpecReg=menu('Choose required regions','Entire spectrum','1800-800cm-1','Other');

        % In entire region is chosen, nothing happens.
        if SpecReg==1;
    RedDB=DB;
    RedWM=WaveMatrix;
        end

% If 1800-800cm-1 is chosen, DB is reduced down to a matrix of 501 wavenumbers.
        if SpecReg==2;
    RedDB=DB(:,925:1425);
    RedWM=WaveMatrix(925:1425);
  end
  % A user specified Spectral Region is chosen.  User can define 1 or multiple
  % spectral regions.
        if SpecReg==3
        SpecReg1=menu('User Specific Spectral Regions','1 Region','>1 Region');
    if SpecReg1==1
      upper=input('Please enter upper limit of region (WM)','s');
      lower=input('Please enter lower limit of region (WM)','s');

      upper=str2num(upper);
      lower=str2num(lower);
      RedDB=DB(:,upper:lower);
      RedWM=WaveMatrix(upper:lower);
    elseif SpecReg1==2
      Howmany=input('Please enter the number of regions?,(2 or 3)','s');
      HM=str2num(Howmany);
      if HM==2
        up1=input('Please enter upper limit of first region (WM)','s');
        lw1=input('Please enter lower limit of first region (WM)','s');
        up2=input('Please enter upper limit of second region (WM)','s');
        lw2=input('Please enter lower limit of second region (WM)','s');

        up1=str2num(up1);
        lw1=str2num(lw1);
        up2=str2num(up2);
        lw2=str2num(lw2);

        DB1=DB(:,up1:lw1);
        DB2=DB(:,up2:lw2);
      RedDB=[DB1 DB2];
      WM1=WaveMatrix(up1:lw1);
      WM2=WaveMatrix(up2:lw2);
      RedWM=[WM1 WM2];
    elseif HM==3
      up1=input('Please enter upper limit of first region (WM)','s');
      lw1=input('Please enter lower limit of first region (WM)','s');
      up2=input('Please enter upper limit of second region (WM)','s');
      lw2=input('Please enter lower limit of second region (WM)','s');
      up3=input('Please enter upper limit of third region (WM)','s');
      lw3=input('Please enter lower limit of third region (WM)','s');

      up1=str2num(up1);
      lw1=str2num(lw1);
      up2=str2num(up2);
      lw2=str2num(lw2);
      up3=str2num(up3);
      lw3=str2num(lw3);

      DB1=DB(:,up1:lw1);
      DB2=DB(:,up2:lw2);
      DB3=DB(:,up3:lw3);
      RedDB=[DB1 DB2 DB3];
      WM1=WaveMatrix(up1:lw1);
      WM2=WaveMatrix(up2:lw2);
      WM3=WaveMatrix(up3:lw3);
      RedWM=[WM1 WM2 WM3];
      end

  end

end

save defreg RedDB RedWM

DB=RedDB;
WaveMatrix=RedWM;
```

NONLINJOIN.M

```
function [DB,Filenames]=nonlinjoin(DB,Filenames)
% This routine is for the removal of data that falls outside the range of acceptable
% absorbance, ie max absorbance >=1.  It is important to remove samples which have
% max absorbance of >= 1 due to nonlinearities of the MCT and the Beer-Lambert Law.

% Author: Melissa Romeo, Monash University, 2000

Y=Filenames;

RedDB=DB';
[NrDB,NcDB]=size(RedDB);

MaxRedDB=max(RedDB);

p=1;
q=1;
for k=1:NcDB;
  if MaxRedDB(k)<=1;
    NewDB(:,p)=RedDB(:,k);
    NewY(p)=Y(k);
    p=p+1;

  else MaxRedDB(k)>1;
    BadDB(:,q)=RedDB(:,k);
    BadY(q)=Y(k);
    q=q+1;
  end
end

save Nonlin BadDB BadY MaxRedDB NewDB NewY RedDB
Filenames=NewY;
DB=NewDB;
```

SNRJOIN.M

```
function [SNR,DB,Filenames,SNRFN]=SNRjoin(DB,WaveMatrix,Filenames)
% Program to collate the SNR (signal-to-noise ratio) obtained from the subroutine
% SNRjoin1.

% Author: Melissa Romeo, Monash University, 2000

[DBm DBn]=size(DB);
t=1;
for i=1:DBm
  DBM=DB(i,:);
  [SNRa]=SNRjoin1(DBM,WaveMatrix);
  SNR(t)=SNRa;
  t=t+1;
end

p=1;
q=1;
Y=Filenames;
DB=DB';
[nrDB,ncDB]=size(DB);

% Cocantenate Filenames and SNR matrices so that there is a list of the SNR for
% each filename.

Y=Filenames';
SNR=SNR';
SNRFN=[Y SNR];
Y=Filenames';
SNR=SNR';

% Once the SNR for all spectra have been calculated, spectra with an SNR less than or
% equal to 10 are removed from the DataBase.

for k=1:ncDB
  if SNR(k)>=abs(10)
    NewDB(:,p)=DB(:,k);
    NewY(p)=Y(k);
    p=p+1;

  elseif SNR(k)<abs(10)
    BadDB(:,q)=DB(:,k);
    BadY(q)=Y(k);
    q=q+1;
  end
end

Filenames=NewY;
```

```
DB=NewDB';


function [SNRa]=SNRjoin1(DBM,WaveMatrix)
% Routine for calculating the signal to noise ratio of spectra

[m n]=size(DBM);

% reduce the spectra down to noise region (2100-1900 cm-1) (DBN)

DBN=DBM(:,775:875);

WMN=WaveMatrix(775:875);

% fit a third order polynomial to the noise region
for i=1:m
  [p(i,:),s(i,:)]=polyfit(WMN,DBN(i,:),3);

  % use the coefficients to plot the polynomial
  y(i,:)=polyval(p(i,:),WMN(i));

  %plot(WMN,y,WMN,DBN)

  % subtraction of the fitted polynomial from the original noise region will
  % give the residual noise
  noise(i,:)=DBN(i,:)-y(i,:);

  %plot(WMN,y,WMN,DBN,WMN,noise)

  % Calculate the amount of noise using the RMS
  RMSN(i)=std(noise(i,:));
end

% To calculate the signal component, the spectra need to be baseline corrected. The
% minimum absorbance in the range 2100-1800cm-1 and 1000-700cm-1 is found and a
% linear regression line is fitted. The regression line is subtracted from the original
% spectrum to give a baseline corrected spectrum. The signal is calculated from the
% Amide II band, and so is given by the maximum absorbance in the region 1560-
% 1520cm^-1.

% Find the minimum absorbance and corresponding wavenumber value in the specified
% region and calculate the gradient (M) and the intercept (C) and draw the line. The
% regression line is then subtracted from the specified region of the spectrum and the
% signal from the amide II band is calculated.

DBM=DBM';
WM=WaveMatrix';
[m n]=size(DBM);

for j=1:n
```

```
  DBWM=[DBM(:,j),WM];
  r1=DBWM(775:925,:);
  r2=DBWM(1325:1475,:);

  min1=min(r1);
  min2=min(r2);

  y1(j)=min1(1);
  y2(j)=min2(1);
  x1(j)=min1(2);
  x2(j)=min2(2);

  diff=x2(j)-x1(j);

  M(j)=y2(j)-y1(j);
  M(j)=M(j)/diff;
  C(j)=y2(j)-M(j)*x2(j);

  WMN=[x1(1):-2:x2(1)];
  [wr wc]=size(WMN);

  for k=1:wc
    RL(j,k)=(WMN(k)*M(j))+C(j);
  end

  DBM=DBM';
  WM=WM';

  for l=1:m
    RLall(j,l)=(WM(l)*M(j))+C(j);
    BL(j,l)=DBM(i,l)-RLall(j,l);
  end

  DBS(i,:)=DBM(i,775:1065);
  maxAm2(i)=max(DBS(i,:));
  SNRa(i)=maxAm2(i)./RMSN(i),
%   plot(WM,RLall,'b',WaveMatrix,DBM,'k',WM,BL,'r')

end
```

SVGLJOIN.M

```
function [DBDeriv,DB]=svgljoin(DB)
% Front routine for the Barry Weise Savitzky-Golay Smoothing Derivative Function.
% Allows interactive user input to specify the nature of the filter required, ie the number
% of smoothing points (width), the order of the polynomial required (order) and the order
% of the derivative (deriv).

width=input('Please enter the number of smoothing points required (odd number):','s');
order=input('Please enter the order of the polynomial required:','s');
deriv=input('Please enter the order of the derivative required:','s');

width=str2num(width);
order=str2num(order);
deriv=str2num(deriv);

DBDeriv=savgolm(DB,width,order,deriv);

DB=DBDeriv;

function y_hat = savgolm(y,width,order,deriv)
% SAVGOL Savitzky-Golay smoothing and differentiation.
% Inputs are the matrix of ROW vectors to be smoothed (y), and the optional variables
% specifying the number of points in filter (width), the order of the polynomial (order),
% and the derivative (deriv). The output is the matrix of smoothed and differentiated
ROW
% vectors (y_hat). If number of points, polynomial order and derivative are not specified,
% they are set to 15, 2 and 0, respectively.
%
%  Example: if y is a 5 by 100 matrix then savgolm(y,11,3,1) gives the 5 by 100 matrix of
% first-derivative row vectors resulting from a 11-point cubic Savitzky-Golay smooth of
% each row of y.
%
%  I/O format is: y_hat = savgolm(y,width,order,deriv);
%
% Sijmen de Jong Unilever Research Laboratorium Vlaardingen Feb 1993
% Modified by Barry M. Wise 5/94
% Modified by: Melissa Romeo, Monash University, 1999

[m,n] = size(y);
y_hat = y;
% set default values: 15-point quadratic smooth
if nargin<4
 deriv= 0;
 disp(' '), disp('Derivative set to zero')
end
if nargin<3
 order= 2;
 disp(' '), disp('Polynomial order set to 2')
```

```
end
if nargin<2
  width=min(15,floor(n/2));
  s = sprintf('Width set to %g',width);
  disp(' '), disp(s)
end
% In case of input error(s) set to reasonable values
w = max( 3, 1+2*round((width-1)/2) );
if w ~= width
  s = sprintf('Width changed to %g',w);
  disp(' '), disp('Width must be >= 3 and odd'), disp(s)
end
o = min([max(0,round(order)),5,w-1]);
if o ~= order
  s = sprintf('Order changed to %g',o); disp(' ')
  disp('Order must be <= width -1 and <= 5'), disp(s)
end
d = min(max(0,round(deriv)),o);
if d ~= deriv
  s = sprintf('Derivative changed to %g',d); disp(' ')
  disp('Deriviative must be <= order'), disp(s)
end
p = (w-1)/2;
% Calculate design matrix and pseudo inverse
x = ((-p:p)'*ones(1,1+o)).^(ones(size(1:w))'*(0:o));
for k = 1:m
  weights = (x'*x)\x';
% Smoothing and derivative for bulk of the data
  for i=p+2:n-p-1
    y_hat(k,i) = weights(d+1,:)*y(k,i-p:i+p)';
  end
  % Smoothing and derivative for tails
  weights = weights*[y(k,1:w)', y(k,n-w+(1:w))']; % full polynomial model
  for i=1:d
    weights = diag(1:o+1-i)*weights(2:o+2-i,:); % or its d'th derivative
  end
  y_hat(k,1:p+1)     = (x(1:p+1,1:1+o-d)*weights(:,1))'; % fitting the tails
  y_hat(k,n-p+(0:p)) = (x(p+1:w,1:1+o-d)*weights(:,2))';
end
```

VNORMJOIN.M

```
function [DB]=vnormjoin(DB)
% Vector normalisation routine.  Calculates the average y-value which is then subtracted
% from the spectrum so that the middle is pulled down to y=0.  The sum of the squares of
% all y-values is then calculated and the spectrum is divided by the square root of this
sum.

% Author: Melissa Romeo, Monash University, 1999

[m n]=size(DB);

for i=1:m
  m(i)=mean(DB(i,:));
  for j=1:n
    sqDB(i,j)=(DB(i,j)^2);
    sumsqDB(i)=sum(sqDB(i,:));
    sqrtDB(i)=sqrt(sumsqDB(i));
  end

  for k=1:n
    vnormDB(i,k)=(DB(i,k)-m(i))./sqrtDB(i);
  end
end
DB=vnormDB;
```

MNORMJOIN.M

```
function [DB]=Mnormjoin(DB)
% This program finds the peak of maximum absorbance in the range of the data, this peak
% is given an arbitary absorbance unit of 1 and all the other absorbances are adjusted
% accordingly.  ie, the maximum absorbance is divided by itself to become 1 and then all
% the other absorbances are divided by the maximum absorbance.

% Author: Melissa Romeo, Monash University, 1999

x=DB';
[rows,cols]=size(x);
for k=1:cols;
  Newx(:,k)=x(:,k)/max(x(:,k));
end
DB=Newx';

%axis tight; hold on;
%title ('Spectra of max-normalised (green) and un-normalised (blue) data');
%xlabel ('Wavenumber Values cm-1)');
%ylabel ('Absorbance (A.U.)');
%plot (Wave,Abs,'b');
%plot (Wave,Norm,'g');
%legend('unnorm','max-norm',2);hold off;
```

MEANORMJOIN.M

```
function [DB]=meannormjoin(DB)
% This program finds the mean absorbance in the range of each spectrum, then each
% absorbance is divided by the mean.

% Author: Melissa Romeo, Monash University, 2000

x=DB';
[rows,cols]=size(x);
for k=1:cols;
  Newx(:,k)=x(:,k)/mean(x(:,k));
end
DB=Newx';

%axis tight; hold on;
%title ('Spectra of max-normalised (green) and un-normalised (blue) data');
%xlabel ('Wavenumber Values cm-1)');
%ylabel ('Absorbance (A.U.)');
%plot (Wave,Abs,'b');
%plot (Wave,Norm,'g');
%legend('unnorm','max-norm',2);hold off;
```

RANGENORMJOIN.M

```
function [DB]=rangenormjoin(DB)
% This program finds the range of each spectrum (ie max absorbance - min absorbance)
% and each absorbance is divided by the range.

% Author: Melissa Romeo, Monash University, 2000

x=DB';
[rows,cols]=size(x);
for k=1:cols;
  Newx(:,k)=x(:,k)/((max(x(:,k)))-(min(x(:,k))));
end
DB=Newx';

%axis tight; hold on;
%title ('Spectra of max-normalised (green) and un-normalised (blue) data');
%xlabel ('Wavenumber Values cm-1)');
%ylabel ('Absorbance (A.U.)');
%plot (Wave,Abs,'b');
%plot (Wave,Norm,'g');
%legend('unnorm','max-norm',2);hold off;
```

BLJOIN.M

```
function [DB,WaveMatrix]=bljoin(DB,WaveMatrix)
% Frontend routine for baseline correction of spectra. The length of DB is determined, ie
% how many wavenumber values and calls the appropriate routine. Bljoinall is called if
% there are more than 501 wavenumbers (whole spectrum) and bljoinred if there are 501
% wavenumbers (spectrum 1800-800 cm-1).

% Author: Melissa Romeo, Monash University, 2000

[DBm DBn]=size(DB);
t=1;
[m n]=size(WaveMatrix);

for i=1:DBm
  DBM=DB(i,:);
  if n>501
    [DBM,WaveMatrix]=bljoinall(DBM,WaveMatrix);

  elseif n==501
    [DBM,WaveMatrix]=bljoinred(DBM,WaveMatrix);
  end

  DB(t,:)=DBM;
  t=t+1;
end

function [DBM,WaveMatrix]=bljoinall(DBM,WaveMatrix)
% Program for returning the indices of selected regions of wavenumbers. Calls the
% subroutine lamseljoin.m, which returns indices for each range.

freqs=WaveMatrix';
DBM=DBM';
inds1=lamseljoin(freqs,[3648 3646]);
inds2=lamseljoin(freqs,[2100 1800]);
inds3=lamseljoin(freqs,[1000 700]);

DB1=DBM(inds1);
WM1=freqs(inds1);

DB2=DBM(inds2);
WM2=freqs(inds2);

DB3=DBM(inds3);
WM3=freqs(inds3);

DBWM1=[DB1 WM1];
DBWM2=[DB2 WM2];
DBWM3=[DB3 WM3];
```

```
min1=min(DBWM1);
min2=min(DBWM2);
min3=min(DBWM3);

r1=lamseljoin(freqs,[min1(2) min2(2)]);
r2=lamseljoin(freqs,[min2(2) min3(2)]);

WMr1=freqs(r1);
WMr2=freqs(r2);

% Calculate the gradient (M1 and M2) and the intercept (C1 and C2) and fit the line to the
% data

M1=min2(1)-min1(1);
M1=M1/(min2(2)-min1(2));
C1=min2(1)-(M1*min2(2));

RL1=WMr1*M1+C1;

M2=min3(1)-min2(1);
M2=M2/(min3(2)-min2(2));
C2=min3(1)-(M2*min3(2));

RL2=WMr2*M2+C2;

%plot(WMr2,RL2,WMr1,RL1,WaveMatrix,DB)

% once the regression lines have been fitted, each region of the spectrum is subtracted
% from the regression line for that region resulting in a baseline corrected spectrum.

DBr1=DBM(r1);
DBr2=DBM(r2);

BL1=DBr1-RL1;
BL2=DBr2-RL2;

BL1=BL1';
BL2=BL2';
BLDB=[BL1 BL2];
DBM=BLDB;
%plot(WaveMatrix,DBM,WaveMatrix,BLDB)


function [DBM,WaveMatrix]=bljoinred(DBM,WaveMatrix)
% Program for returning the indices of selected regions of wavenumbers.  Calls the
% subroutine lamseljoin.m, which returns indices for each range.

freqs=WaveMatrix';
DBM=DBM';
```

```
inds1=lamseljoin(freqs,[1800 1700]);
inds2=lamseljoin(freqs,[1000 800]);

DB1=DBM(inds1);
WM1=freqs(inds1);

DB2=DBM(inds2);
WM2=freqs(inds2);

DBWM1=[DB1 WM1];
DBWM2=[DB2 WM2];

min1=min(DBWM1);
min2=min(DBWM2);

r1=lamseljoin(freqs,[min1(2) min2(2)]);

WMr1=freqs(r1);

% Calculate the gradient (M) and the intercept (C) and fit the line to the data

M=min2(1)-min1(1);
M=M/(min2(2)-min1(2));
C=min2(1)-(M*min2(2));

RL=WaveMatrix*M+C;

%plot(WaveMatrix,RL,WaveMatrix,DBM)

% once the regression lines have been fitted, each region of the spectrum is subtracted
% from the regression line for that region resulting in a baseline corrected spectrum.

RL=RL';
BL=DBM-RL;

BL=BL';

DBM=BL;

%plot(WaveMatrix,DBM,WaveMatrix,BL)
```

DIAGNOSISRWH.M

```
function [DBXSpec,DCYSpec,needBX,needCY]=DiagnosisRWH(DB,Filenames)
% This program is designed to assign a histological (biopsy) and cytological (Pap smear)
% diagnosis to the infrared spectra in DBRWH.  The program works by comparing the
% filenames of the spectral files with the patient number and returning a numerical
% diagnosis.

% Author: Melissa Romeo, Monash University, 2000

load RWHdiag2.txt;
RWH=RWHdiag2;
FN=Filenames';
[rRWH cRWH]=size(RWH);
[rF cF]=size(FN);

for i=1:rF
  S1=FN(i);

  for j=1:rRWH
    S2=RWH(j,1);

    if S1==S2
      DiagBX(i)=RWH(j,2);
      DiagCY(i)=RWH(j,3);
    end
  end
end

DiagBX=DiagBX';
DiagBX=[FN DiagBX];
DBXSpec=[DB DiagBX];

DiagCY=DiagCY';
DiagCY=[FN DiagCY];
DCYSpec=[DB DiagCY];

% Remove spectra that don't have a diagnosis (assigned 0).
function [DBXSpec,DCYSpec,needBX,needCY]=remzero(DBXSpec,DCYSpec);

assign=menu('Type of Assignment','Normal/Abnormal','AbsNormal/Abnormal',...
  'Normal/Dysplasia','AbsNormal/Normal/Abnormal');
if assign==1
  [DBXSpec,DCYSpec]=DiagRWHNA(DBBX,DBCY,FNBX,FNCY);

elseif assign==2
  [DBXSpec,DCYSpec]=DiagRWHNAX(DBBX,DBCY,FNBX,FNCY);

elseif assign==3
```

```
  [DBXSpec,DCYSpec]=DiagRWHND(DBBX,DBCY,FNBX,FNCY);

elseif assign==4
  [DBXSpec,DCYSpec]=DiagRWHNNAA(DBBX,DBCY,FNBX,FNCY);
end
```

DIAGRWHNA.M

```
function [DBXSpec,DCYSpec]=DiagRWHNA(DBBX,DBCY,FNBX,FNCY);

% This program is designed to assign a cytological (Pap smear) diagnosis to the infrared
% spectra in DBRWH.  The program works by comparing the filenames
% of the spectral files with the patient number and returning a numerical diagnosis.
% The diagnosis is then cocantenated to the DB file so that each spectrum has a diagnosis.

% Author: Melissa Romeo, Monash University, 2000

%clear all

%load om1normabpd
load CytolCYNNAA.txt;
RWH=CytolCYNNAA;
[rRWH cRWH]=size(RWH);
[rFNBX cFNBX]=size(FNBX);
[rFNCY cFNCY]=size(FNCY);

for i=1:rFNBX
  S1=FNBX(i);

  for j=1:rRWH
    S2=RWH(j,1);

    if S1==S2
      DiagBX(i)=RWH(j,2);
    end
  end
end

for k=1:rFNCY
  S3=FNCY(k);
  for l=1:rRWH
    S4=RWH(l,1);

    if S3==S4
      DiagCY(k)=RWH(l,3);
    end
  end
end


DiagBX=DiagBX';
DiagBX=[FNBX DiagBX];
DBXSpec=[DBBX DiagBX];

DiagCY=DiagCY';
```

```
DiagCY=[FNCY DiagCY];
DCYSpec=[DBCY DiagCY];

save om1RWHDiagNA

[DBXSpecN,DCYSpecN]=normab1(DBXSpec,DCYSpec);
%DBXSpec=DBXSpecN;
%DCYSpec=DCYSpecN;
```

NORMAB1.M

```
function [DBXSpecN,DCYSpecN]=normab1(DBXSpec,DCYSpec);

% Author: Melissa Romeo, Monash University, 2000

[rB cB]=size(DBXSpec);
[rC cC]=size(DCYSpec);

BX=DBXSpec(:,cB);
FNB=DBXSpec(:,cB-1);
CY=DCYSpec(:,cC);
FNC=DBXSpec(:,cC-1);


s=1;
t=1;
u=1;
v=1;

for i=1:rB
  if BX(i) <90;
    newBXSpec(s,:)=DBXSpec(i,:);
    s=s+1;
  elseif BX(i) >=90;
    badBXSpec(t,:)=DBXSpec(i,:);
    t=t+1;
  end
end

for j=1:rC
  if CY(j) <90;
    newCYSpec(u,:)=DCYSpec(j,:);
    u=u+1;
  elseif CY(j) >=90
    badCYSpec(v,:)=DCYSpec(j,:);
    v=v+1;
  end
end

DBXSpecN=newBXSpec;
DBBX=DBXSpecN(:,1:cB-2);
FNBX=DBXSpecN(:,cB-1);

DCYSpecN=newCYSpec;
DBCY=DCYSpecN(:,1:cC-2);
FNCY=DCYSpecN(:,cC-1);

save om1normabdiag
```

DIAGNOSISFPV.M

```
function [DBFPV,DiagFPV,needFPV]=DiagnosisFPV(DB,Filenames)
% This program is designed to assign a cytological (Pap smear) diagnosis to the infrared
% spectra in DBFPV.  The program works by comparing the filenames
% of the spectral files with the patient number and returning a numerical diagnosis.
% The diagnosis is then cocantenated to the DB file so that each spectrum has a diagnosis.

% Author: Melissa Romeo, Monash University, 2000

load fpvdiag.txt;
FPV=fpvdiag;
FN=Filenames';
[rFPV cFPV]=size(FPV);
[rF cF]=size(FN);

% need to convert the numbers of the array FN and Biopsy to strings to enable
comparison.
% (doesn't work because then only the first character is compared, need to extract
filename and
% then convert to string.

for i=1:rF
  S1=FN(i);

  for j=1:rFPV
    S2=FPV(j,1);

    if S1==S2
      DiagFPV(i)=FPV(j,2);
    end
  end
end

DiagFPV=DiagFPV';
%DiagFPV=[FN DiagFPV];
%DBFPV=[DB DiagFPV];

save FPVDiag

%assign=menu('Type of Assignment','Normal/Abnormal','True Normal/Abnormal',...
%'BX/CY','Normal/Dysplasia');
%if assign==1
%   [DBFPV,needFPV]=Fnormab(DBFPV);

%elseif assign==2
%   [DBFPV,needFPV]=Ftnormab(DBFPV);

%elseif assign==3
```

```
%   [DBFPV,needFCY]=Fcy(DBFPV);

%elseif assign==4
%   [DBFPV,needFCY]=Fnormdys(DBFPV);

%end
```

Please Note: The diagnosis programs for different types of diagnosis are all very similar. Therefore an example of the programs have been given in DiagRWHNA.m and Normab1.m above.

LDACERV.M

```
function
[mvect,pcov,tscor,pscor,tmislist,pmislist,dist_sum,tpct,ppct]=mjrlda(X1,Y1,X2,Y2,pscal)

% Linear Discriminant Analysis-Mahalanobis
% assigns ungrouped items to closest group centre, using the Mahalanobis
% distance measure (i.e., minimum distance classifier).
% Based on the explanation given in Adams [9].
% Author: Melissa Romeo, 1998

%-------------------------------------------------------
[X1,xmean,xstd]=pscale(X1,pscal);
if min(Y1)==0;Y1=Y1+1;end;
PCprint=1;
NPC=input('Number of PCs  [0=No PCA] = ');
if isempty(NPC),NPC=0;end;

        if NPC>0
                scl=0;
                [train,LD,ssq,res,q,tsq]=frbpca(X1,0,scl,NPC,PCprint);
        else
                train=X1;
        end;
%
----------------------------------------------------------
j=1;
k=1;
m=1;
n=2;

X1=train;
train=X1;
[Rt Ct]=size(train);
[RY1 CY1]=size(Y1);
trainT=train';

%-------- What is the Range of the data? ----------------------
PC1=train(:,1);
minx1=min(PC1);
maxx1=max(PC1);

PC2=train(:,2);
minx2=min(PC2);
maxx2=max(PC2);

%-------- Separate the data into classes -----------------------
p=1;
q=1;
```

```
for i = 1:RY1
  Y(i)=Y1(i);
  if Y(i)==1;
    Xa(p,1:Ct)=train(i,1:Ct);
    p=p+1;
    [RXa CXa]=size(Xa);
  elseif Y(i)==2;
    Xb(q,1:Ct)=train(i,1:Ct);
    q=q+1;
    [RXb CXb]=size(Xb);
  end
end

XaT=Xa';
XbT=Xb';

%----  Calculate the vector of variable means for each class ------
meanXa=MEAN(Xa);
meanXb=MEAN(Xb);
%----  Calculate the pooled covariance matrix (and the inverse) --------------------
Covt=cov(train);
invCovt=inv(Covt);
%----  Claculate fA(x) and fB(x) and assign the objects to classes ----------------
CA0 = 0.5*meanXa*invCovt*meanXa';
CA1 = meanXa*invCovt;
CB0 = 0.5*meanXb*invCovt*meanXb';
CB1 = meanXb*invCovt;

corA=0;
missedA=0;
cmissA=0;
corB=0;
missedB=0;
cmissB=0;

for i=1:Rt
  x=trainT(1:2,i);
  fA = CA1*x - CA0;
  fB = CB1*x - CB0;

  if fA>fB
    ClassA(j)=fA;
    MatrixA(1:2,j)=x;
    j=j+1;
    if Y1(i)==1
      corA=corA+1;
    else
      cmissB=cmissB+1;
      PrmissB(cmissB)=i;
      missedB=missedB+1;
    end
  end

  if fA<fB
    ClassB(k)=fB;
    MatrixB(1:2,k)=x;
    k=k+1;
    if Y1(i)==2
      corB=corB+1;
    else
      cmissA=cmissA+1;
      PrmissA(cmissA)=i;
      missedA=missedA+1;
    end
  end

  if fA==fB
    ClassM(m)=fA;
    MatrixM(1:2,m)=x;
    m=m+1;
  end

end
fprintf('\rTraining Classification Performance \r');
percorA=(corA/RXa)*100;
fprintf('Class1 Percentage Correct %7.4f\n',percorA);
percorB=(corB/RXb)*100;
fprintf('Class2 Percentage Correct %7.4f\n',percorB);

if cmissA>0;
  fprintf('\rThe sample number/s of the mis-assigned sample in Class A is %d\n',PrmissA);
end
if cmissB>0;
  fprintf('\rThe sample number of the mis-assigned sample in Class B is %d\n',PrmissB);
end


%----   Calculate the Linear Discriminant Function ------------------
CA11=CA1(1,1);
CA12=CA1(1,2);
CB11=CB1(1,1);
CB12=CB1(1,2);

Lsol(1)=((CA0-CB0)-minx2*(CA12-CB12))/(CA11-CB11);
Lsol(2)=((CA0-CB0)-maxx2*(CA12-CB12))/(CA11-CB11);

x2(1)=minx2;
x2(2)=maxx2;
```

```
%----  Original Data -------------------------
XA1=Xa(:,1);
XA2=Xa(:,2);
XB1=Xb(:,1);
XB2=Xb(:,2);

%----  Classified Data ----------------------
ClA1=MatrixA(1,:);
ClA2=MatrixA(2,:);
ClB1=MatrixB(1,:);
ClB2=MatrixB(2,:);

%----  Plot classification results ------------
hold on
axis([(minx1-0.1) (maxx1+0.1) (minx2-0.1) (maxx2+0.1)])
plot(XA1,XA2,'r+')
plot(ClA1,ClA2,'ro')
plot(XB1,XB2,'b+')
plot(ClB1,ClB2,'bo')
plot(Lsol,x2,'k')

%----  New compute prediction results -------------------------
%----  Predict which class an unknown sample belongs to -------

disp(' Input test data if desired ');
[X1,Y1,NA,NCol,Ncy]=getdata;

[X2]=usepscal(X1,xmean,xstd);
pred=X2*LD;
if min(Y1)==0;
  Y1=Y1+1;
end;
[RP,CP] = size(pred);
predT=pred';
%pdlist=Y1;
%pcor = 0;
%missed = 0;
%for i = 1:RP
        %for k = 1:ncls
                %pscor(i,k) = (mvect(k,:)-pred(i,:))*c*(mvect(k,:)-pred(i,:))';
        %end
        %temp(1:ncls) = pscor(i,1:ncls);
        %[junk,winner] = min(temp);
        %if (winner == pdlist(i))
                %pcor = pcor + 1;
        %else
                %missed = missed + 1;
                %pmislist(missed) = i;
        %end
%end
```

```
%temp = sum(tcor);
%fprintf('\n');
%tpct = (temp/nrow_t)*100;
%ppct = (pcor/RP)*100;
%fprintf('Training performance %7.4f \n',tpct);
%fprintf('Prediction performance %7.4f \n',ppct);

%----  "Unknown" Data Classes --------------------------------
p=1;
q=1;
[RY CY]=size(Y1);
for i = 1:RY
  Y(i)=Y1(i);
  if Y(i)==1;
    XPa(p,1:Ct)=pred(i,1:Ct);
    p=p+1;
  elseif Y(i)==2;
    XPb(q,1:Ct)=pred(i,1:Ct);
    q=q+1;
  end
end

j=1;
k=1;
m=1;

for i=1:RP
  ux=predT(1:2,i);
  fA = CA1*ux - CA0;
  fB = CB1*ux - CB0;

  if fA > fB
    PredA(j) = fA;
    PredMatrixA(1:2,j)=ux;
    j=j+1;
  end

  if fA < fB
    PredB(k)=fB;
    PredMatrixB(1:2,k)=ux;
    k=k+1;
  end

  if fA==fB
    PredM(m)=fA;
    PredMatrixM(1:2,m)=ux;
    m=m+1;
  end

end
```

```
%-------- What is thc Range of the data? ----------------------
P1=pred(:,1);
minPx1=min(P1);
maxPx1=max(P1);

P2=pred(:,2);
minPx2=min(P2);
maxPx2=max(P2);
%----  Original "Unknown" Data ------------------------
%XPA1=XPa(:,1);
%XPA2=XPa(:,2);
%XPB1=XPb(:,1);
%XPB2=XPb(:,2);

%----  Predicted Data ------------------------------
[RPa CPa]=size(PredMatrixA);
[RPb CPb]=size(PredMatrixB);
PA1=PredMatrixA(1,:);
PA2=PredMatrixA(2,:);
PB1=PredMatrixB(1,:);
PB2=PredMatrixB(2,:);

hold on
axis([(minx1-0.1) (maxx1+0.1) (minx2-0.1) (maxx2+0.1)])
plot(XA1,XA2,'r+')
plot(ClA1,ClA2,'ro')
plot(XB1,XB2,'b+')
plot(ClB1,ClB2,'bo')
%plot(XPA1,XPA2,'m+')
plot(PA1,PA2,'mo')
%plot(XPB1,XPB2,'c+')
plot(PB1,PB2,'co')
plot(Lsol,x2,'k')
```

QDACERV.M

```
function
[mvect,pcov,tscor,pscor,tmislist,pmislist,dist_sum,tpct,ppct]=mjrlda(X1,Y1,X2,Y2,pscal)
```

PCA routine modified by Frank Burden, Monash University. Originally written by Barry Wiese.

```
%------------------------------------------------------
[X1,xmean,xstd]=pscale(X1,pscal);
if min(Y1)==0;Y1=Y1+1;end;
PCprint=1;
NPC=input('Number of PCs  [0=No PCA] = ');
if isempty(NPC),NPC=0;end;

        if NPC>0
                scl=0;
                [train,LD,ssq,res,q,tsq]=frbpca(X1,0,scl,NPC,PCprint);
        else
                train=X1;
        end;
%-------------------------------------------------------
%
% Quadratic Discriminant Analysis-Mahalanobis
% assigns ungrouped items to closest group centre, using the Mahalanobis
% distance measure (i.e., minimum distance classifier).

% Based on the explanation given in Adams [9].
% Author: Melissa Romeo, 1993
%-------- Routine for QDA ----------
j=1;
k=1;
m=1;
n=2;

X1=train;
train=X1;

[Rt Ct]=size(train);
[RY1 CY1]=size(Y1);
trainT=train';

%-------- What is the Range of the data? ----------------------
PC1=train(:,1);
minx1=min(PC1);
maxx1=max(PC1);

PC2=train(:,2);
minx2=min(PC2);
```

```
maxx2=max(PC2);


%-------- Separate the data into classes -------------------------

p=1;
q=1;
for i = 1:RY1
  Y(i)=Y1(i);
  if Y(i)==1;
    Xa(p,1:Ct)=train(i,1:Ct);
    p=p+1;
    [RXa CXa]=size(Xa);
  elseif Y(i)==2;
    Xb(q,1:Ct)=train(i,1:Ct);
    q=q+1;
    [RXb CXb]=size(Xb);
  end
end

XaT=Xa';
XbT=Xb';

%-------        What is the vector of variable means for each group?
meanXa=MEAN(Xa);
meanXb=MEAN(Xb);

%-------        What is the variance-covariance matrix for each group?
CovXa=cov(Xa);
CovXb=cov(Xb);

%-------        and their inverse matrices
invCovXa=inv(CovXa);
invCovXb=inv(CovXb);

%-------        What is the determinant of each matrix?
detCovXa=det(CovXa);
detXa=detCovXa;
detCovXb=det(CovXb);
detXb=detCovXb;

%-------        Calculate the discriminant functions, dA(x) and dB(x), for each sample in
the training set.
% let Ma = (x-meanXa)
% let Mb = (x-meanXb)

corA=0;
missedA=0;
cmissA=0;
corB=0;
```

```
missedB=0;
cmissB=0;

for i=1:Rt
  x=trainT(1:2,i);
  Ma=(x-meanXa');
  Mb=(x-meanXb');

  dA = 0.5*log(detCovXa) + 0.5*Ma'*invCovXa*Ma;

  dB = 0.5*log(detCovXb) + 0.5*Mb'*invCovXb*Mb;

   if dA < dB
     ClassA(j) = dA;
     MatrixA(1:2,j)=x;
     j=j+1;
     if Y1(i)==1
       corA=corA+1;
    else
      cmissB=cmissB+1;
      PrmissB(cmissB)=i;
      missedB=missedB+1;
    end

  end

  if dA > dB
    ClassB(k)=dB;
    MatrixB(1:2,k)=x;
    k=k+1;
    if Y1(i)==2
      corB=corB+1;
    else
      cmissA=cmissA+1;
      PrmissA(cmissA)=i;
      missedA=missedA+1;
    end

  end

  if dA==dB
    ClassM(m)=dA;
    MatrixM(1:2,m)=x;
    m=m+1;
  end
end
fprintf('\rTraining Classification Performance \r');
percorA=(corA/RXa)*100;
fprintf('Class1 Percentage Correct %7.4f\n',percorA);
percorB=(corB/RXb)*100;
```

```
fprintf('Class2 Percentage Correct %7.4f\n',percorB);

if cmissA>0;
  fprintf('\rThe sample number(s) of the mis-assigned sample in Class A is
%d\n',PrmissA);
end
if cmissB>0;
  fprintf('\rThe sample number of the mis-assigned sample in Class B is %d\n',PrmissB);
end

%----  Solve the quadratic equation, where dA(x)=dB(x), solve for x where x=sol.
%where: dA(x) = x(1)^2a(1,1) + 2x(1)x(2)a(2,1) + x(2)^2a(2,2)+ ln(|CovXa|)
% and a(i,j) are the values from the invCov matrix.
%because there is only one equation, and two unknowns, assume that x(2)is known.
%therefore solve the quadratic equation for x(1) and then replace into the original
equation to solve for x(2).
%d = b^2 - 4ac in the quadratic

ai=invCovXa;
a11=ai(1,1);
a12=ai(1,2);
a21=ai(2,1);
a22=ai(2,2);

bi=invCovXb;
b11=bi(1,1);
b12=bi(1,2);
b21=bi(2,1);
b22=bi(2,2);

mua1=meanXa(1,1);
mua2=meanXa(1,2);
mub1=meanXb(1,1);
mub2=meanXb(1,2);

d=log(detXa)-log(detXb);

alp1=a11*mua1+a21*mua2;
alp2=a22*mua2+a21*mua1;

beta1=b11*mub1+b21*mub2;
beta2=b22*mub2+b21*mub1;

Ca=a11*(mua1)^2+a22*(mua2)^2+ 2*mua1*mua2*a21;
Cb=b11*(mub1)^2+b22*(mub2)^2+ 2*mub1*mub2*b21;

w=1;
div=(maxx2-minx2)/10;
for i=minx2:0.0001:maxx2
  x2=i;
  A=a11-b11;
  B=2*((a21-b21)*x2-alp1+beta1);
  C=(a22-b22)*(x2)^2-(alp2-beta2)*2*x2+Ca-Cb+d;
  e=(B^2)-4*(A*C);
  if e > 0
    Qsol(w,1)=(-B+sqrt(e))/(2*A);
    Qsol(w,2)=(-B-sqrt(e))/(2*A);
    xx2(w)=x2;
    w=w+1;
  end
end
[RQ CQ] = size(Qsol);
%minx2;
%maxx2;
%div;
Qsol;
xx2';

%----  Plotting the data --------------------

%----  Original Data -----------------------
XA1=Xa(:,1);
XA2=Xa(:,2);
XB1=Xb(:,1);
XB2=Xb(:,2);

%----  Classified Data ---------------------
ClA1=MatrixA(1,:);
ClA2=MatrixA(2,:);
ClB1=MatrixB(1,:);
ClB2=MatrixB(2,:);

hold on
axis([minx1-0.1 maxx1+0.1 minx2-0.1 maxx2+0.1])
plot(XA1,XA2,'r+')
plot(ClA1,ClA2,'ro')
plot(XB1,XB2,'b+')
plot(ClB1,ClB2,'bo')
plot(Qsol,xx2,'k')
%plot(Qsol(:,2),xx2,'m')
%plot(Qsol(:,1),xx2,'g')

%
%----  New compute prediction results -----------------------
%----  Predict which class an unknown sample belongs to -------

disp(' Input test data if desired ');
[X1,Y1,NA,NCol,Ncy]=getdata;

[X2]=usepscal(X1,xmean,xstd);
```

```
pred=X2*LD;
[RP,CP] = size(pred);
predT=pred';

if min(Y1)==0;
  Y1=Y1+1;
end;

for i=1:RP
  ux=predT(1:2,i);
  Ma=(ux-meanXa');
  Mb=(ux-meanXb');

  dA = 0.5*log(detCovXa) + 0.5*Ma'*invCovXa*Ma;

  dB = 0.5*log(detCovXb) + 0.5*Mb'*invCovXb*Mb;

   if dA < dB
     PredA(j) = dA;
     PredMatrixA(1:2,j)=ux;
     j=j+1;
   end

  if dA > dB
    PredB(k)=dB;
    PredMatrixB(1:2,k)=ux;
    k=k+1;
  end

  if dA==dB
    ClassM(m)=dA;
    MatrixM(1:2,m)=ux;
    m=m+1;
  end
end
%-------- What is the Range of the data? ----------------------
P1=pred(:,1);
minPx1=min(P1);
maxPx1=max(P1);

P2=pred(:,2);
minPx2=min(P2);
maxPx2=max(P2);
%----  Original "Unknown" Data ------------------------
%XPA1=XPa(:,1);
%XPA2=XPa(:,2);
%XPB1=XPb(:,1);
%XPB2=XPb(:,2);
```

```
%----  Predicted Data --------------------------------
[RPa CPa]=size(PredMatrixA);
[RPb CPb]=size(PredMatrixB);
PA1=PredMatrixA(1,:);
PA2=PredMatrixA(2,:);
PB1=PredMatrixB(1,:);
PB2=PredMatrixB(2,:);

hold on
axis([(minx1-0.1) (maxx1+0.1) (minx2-0.1) (maxx2+0.1)])
plot(XA1,XA2,'r+')
plot(ClA1,ClA2,'ro')
plot(XB1,XB2,'b+')
plot(ClB1,ClB2,'bo')
%plot(XPA1,XPA2,'m+')
plot(PA1,PA2,'mo')
%plot(XPB1,XPB2,'c+')
plot(PB1,PB2,'co')
plot(Qsol,xx2,'k')
```

APPENDIX E

JCAMP.DX FORMAT

```
##TITLE=sample
##JCAMP-DX=4.24
##DATA TYPE=INFRARED SPECTRUM
##DATE=10/11/1998
##TIME=9:32:46
##DATA PROCESSING=no operation
##XUNITS=1/CM
##YUNITS=ABSORBANCE
##RESOLUTION=8
##FIRSTX=3648
##LASTX=700
##DELTAX=-2
##MAXY=0.35364914
##MINY=0.0079337647
##XFACTOR=1
##YFACTOR=3.2936143e-010
##NPOINTS=1475
##FIRSTY=0.15247789
##XYDATA=(X++(Y..Y))
3648+462950053+459379072+458404064+460037344+463060576+467417728+472176
224
3634+477231296+481464992+484786848+485771168+485487296+486241888+487875
424
3620+491510976+495722464+498586464+501224512+504022592+507604064+513067
328
3606+518151296+521789536+524453280+525867872+527819648+530787968+534113
024
3592+537768704+541275392+544601536+548018368+551502528+554634752+557532
480
3578+560743360+564380672+569294400+575186944+581985536+588830464+593840
896
3564+597611136+599918400+601973760+604996032+608684160+613151808+618207
360
↓
↓
↓            Wavenumber and Absorbance Values
↓
↓
738+104561816+106976208+108915000+110471512+111640256+112741912+1137680
80
724+113702448+112613680+110632456+108169360+107549456+108679392+1126616
16
710+118773392+126022648+133859840+141024016+147377024+152507760
##END
```

APPENDIX F

RECRUITMENT POSTER

Department of Chemistry
Monash University

Non-smoking women urgently needed to participate in Chemistry Research Project

Please contact:
Melissa Romeo
Room 109C Chemistry
Ph. 9905 4557, Ext. 54557

EXPLANATORY STATEMENT

Fourier-Transform Infrared Spectroscopy in the Diagnosis of Cervical Cancer

My name is Melissa Romeo and I am studying for my Bachelor of Science (Honours) degree at Monash. A research project is an important component of the course and I am undertaking my research project under the supervision of Dr Don McNaughton, a lecturer in the Department of Chemistry.

The aim of this project is to investigate the cellular changes of cervical cells during the menstrual cycle using infrared spectroscopy. Infrared spectroscopy is generating substantial interest as a technique for the diagnosis of cervical cancer, and it is anticipated that the findings of this research project will contribute to the eventual implementation of infrared spectroscopy into cervical cancer screening.

I am seeking non-smoking women who are prepared to undergo cervical smears. The procedure will take approximately 10-15 minutes each week over a period of three months, and will be undertaken at the Health Service in the Union Building by a qualified medical practitioner. Women taking oral contraception (monophasic) are not excluded from this research. Women not on the pill will be required to give blood samples for a hormonal assay once every cycle. The women, if possible, will need to refrain from sexual intercourse at least 48 hours before each smear, and abstain from vaginal douches and baths 24 hours before each smear.

No findings that could identify any individual participant will be published. The anonymity of your participation is assured. Access to data is restricted to my supervisor and myself. Coded data are stored for five years, as prescribed by University regulations.

Participation in this research is entirely voluntary, and you will be paid for time and inconvenience. If you agree to participate, you may withdraw your consent at any time.

If you have any queries or would like to be informed of the aggregate research finding, please contact 9905 4557 or fax 9905 4597.

Should you have any complaint concerning the manner in which this research is conducted, please do not hesitate to contact The Standing Committee on Ethics in Research on Humans at the following address:

The Secretary
The Standing Committee on Ethics in Research on Humans
Monash University
Wellington Road
Clayton Victoria 3168
Telephone (03) 9905 2052 Fax (03) 9905 1420

Thank you.

Melissa Romeo
9905 4557

Informed Consent Form

Fourier-Transform Infrared Spectroscopy in the Diagnosis of Cervical Cancer

I agree to take part in the above Monash University research project. I have had the project explained to me, and I have understood and read the Explanatory Statement, which I retain for my records.

I understand that there may be a slight discomfort experienced during the cervical smear. Standard procedures for collecting cervical smears will be adhered to, and so there is no danger of additional risks.

I understand that results from my participation in the project are confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party.

I also understand that my participation is voluntary, that I can choose not to participate, and that I can withdraw my participation at any stage of the project.

Name:..(please print)

Signature:...Date:.....................

Independent witness to participant's voluntary and informed consent.

Name:..(please print)

Signature:...Date:...................

Address:..

....

....

APPENDIX G

PROCEDURE FOR REMOVING HELA CELLS FROM A MONOLAYER SUSPENSION

1. Tip off media into a centrifuge tube;
2. Pipette 3 ml Hank's BBS (balanced salt solution, room temp.) into flask and rock;
3. Tip off into a centrifuge tube;
4. Repeat steps 2 and 3;
5. Add 3 ml Trypsin-Versine solution (room temp.), rock flask;
6. Incubate 35°-37°C for 5 minutes;
7. Shake flask to loosen cells;
8. Collect solution in a centrifuge tube;
9. Centrifuge tubes at 1000 g for 10 minutes;
10. Collect pellets and resuspend in PBS^+.

PREPARATION OF EHRLICH'S HAEMOTOXYLIN

1. Dissolve 2 g haematoxylin in 100 ml of 95% ethanol;
2. Add 100 ml distilled water, 100 ml glycerin, 3 g ammonium or potassium alum and 10ml glacial acetic acid;
3. May be ripened by addition of 0.1 g sodium iodate;
4. Stain for 2-5 minutes.

PREPARATION OF SOLUTIONS NEEDED FOR SUBCELLULAR FRACTIONATION

Solution A:	60% w/v Iodixanol (obtained from Sigma Aldrich, Australia).
Solution B:	Diluent: 150 mM KCl, 30 mM $MgCl_2$, 120 mM Tricine NaOH, pH 7.8.
Solution C:	Working Solution (50% w/v Iodixanol): Mix 5 volumes solution A with 1 volume of solution B.
Solution D:	Homogenisation Medium: 0.25 M sucrose, 25 mM KCl, 5 mM $MgCl_2$, 20 mM Tricine NaOH, pH 7.8.

APPENDIX H

THE NITROGEN BOMB

Nitrogen cavitation is one way of producing cell rupture. A suspension of cells is placed in a sealed stainless steel vessel and nitrogen gas is introduced to pressurise the headspace above the cell suspension. The cells are allowed to equilibrate with the pressure and when the valve is released the difference in the pressure at the valve causes the cells to rupture. The level of disruption is controlled using different equilibration times and pressures.

The nitrogen bomb, Figure H.1, consists of a needle valve that is screwed into the bottom of the bomb and a Teflon ball, which acts as a seal. The chamber of the bomb is filled with the cell suspension (up to 15 ml) and the top section is screwed down. A seal is achieved through an o-ring contact. The beaker was replaced with a centrifuge tube with a hole drilled into the lid to minimise loss of homogenate.

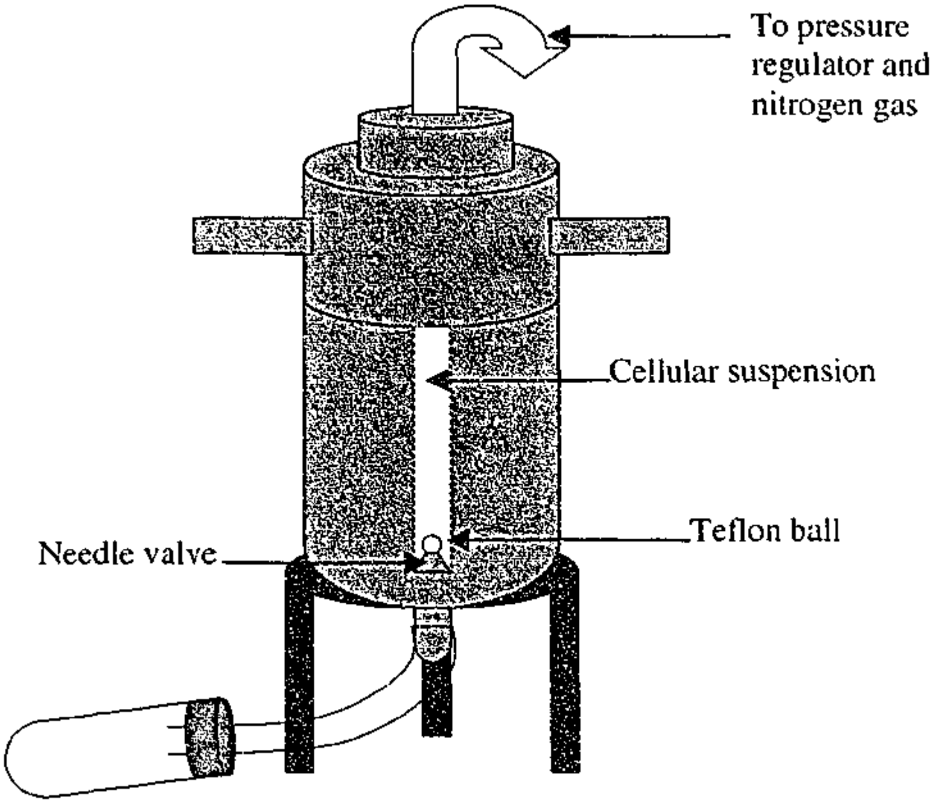

Figure H.1 Schematic representation of the nitrogen bomb apparatus used to produce a cell homogenate.

APPENDIX I

MAKE COMPATIBLE

This function changes the datapoint spacing of the selected files to match that of the 'principal file'. If the file limits of a selected spectrum lie outside the file limits of the principal file, the selected file is cut accordingly. The 'principal file' is the file to which all other files are made compatible [11].

The use of this function was necessary in order to compare spectra collected on the Perkin Elmer (PE) Spectrometer with spectra collected on the Bruker Spectrometer, as well as producing spectral files with integer wavenumber values. Spectral files recorded on the PE Spectrometer had 1476 data points ranging from 3650 – 700 cm^{-1}. Spectra recorded on the Bruker had 764 points ranging from 3648.789 - 705.3439 cm^{-1}. After Bruker spectra were made compatible using a method of quadratic interpolation they had 1475 points ranging from 3648 – 700 cm^{-1}.

APPENDIX J

CONSENT FORM AND EXPLANATORY STATEMENT FOR RWH

THE ROYAL WOMEN'S HOSPITAL

RESEARCH COMMITTEE

EXPLANATION AND CONSENT FORM AND QUESTIONNAIRE

1.0 Title of Project: FOURIER TRANSFORM INFRARED SPECTROSCOPY IN THE DIAGNOSIS OF CERVICAL CANCER

2.0 Chief Investigator: Dr. D McNAUGHTON/PROF. MA QUINN

3.0 Description of Project ("LAY TERMS")

The present methods of screening for cervical cancer produce a number of false negatives and false positives and around 40% of these arise from errors in laboratory diagnosis. These tests also involve complicated procedures, are time consuming and expensive. Fourier Transform Infrared (FTIR) Spectroscopy has recently been shown to be very promising as an inexpensive and rapid tool for the diagnosis of cervical cells. We wish to explore the possibility of using FTIR spectroscopy as a screening tool for cervical cancer by examining the cells left on instruments used to take your smear test.

4.0 Possible Risks, inconveniences and discomforts:

NONE

5.0 I, the undersigned ……………………………………………………

Hereby consent to my involvement in the research project no ………………

Titled: FOURIER TRANSFORM INFRARED SPECTROSCOPY IN THE DIAGNOSIS OF CERVICAL CANCER

1. Patient Number: ……………………………………
2. Date of Birth: ……………………………………
3. Method of Contraception (if oral contraception is it mono- or triphasic)?:

……………………………………………………………………………………………

4. Number of Pregnancies: ……………………………………
5. Smoker (YES/NO): ……………………………………
6. What day of your cycle are you currently? ……………………………………
(1st day of bleeding is taken as day 1)

7. Usual length of cycle (including menstruation): ....

8. Usual length of menstruation: ....

9. Do you use tampons? (YES/NO): ....

10. Have you engaged in sexual intercourse in the last 48 hours?:................................

All information is anonymous and confidential. Thank you for your cooperation.

5.1 I acknowledge that the nature, purpose and contemplated effects of the project so far as it affects me have been fully explained to my satisfaction by the research worker and my consent is given voluntarily.

5.2 The detail of the procedure proposed has also been explained to me, including the anticipated length of time it will take, the frequency with which the procedure will be performed and an indication of any discomfort that may be expected.

5.3 Although I understand that the purpose of this research project is to improve the quality of medical care, it has also been explained that my involvement may not be of any benefit to me.

5.4 I have been given the opportunity to have am member of my family or a friend present whilst the project was explained to me.

5.5 I am informed that no information regarding my medical history will be divulged and the results of any tests involving me will not be published so as to reveal my identity.

5.6 I understand that my involvement in this project will not affect my relationship with my medical advisers in their management of my health. I also understand that I am free to withdraw from the project at any stage.

I consent to be included in this research study

Signature: .... Date: / /

Witness: .... Date: / /

I, ... being the investigator named above, certify that I have explained the nature and abject of the investigations and have made clear that declining to participate would bear no adverse consequences.

Name and phone number for emergency contact: PROF. MA QUINN
(W) 9344 2130
(H) 9816 9387

CONSENT FORM, EXPLANATORY STATEMENT AND QUESTIONNAIRE FOR FPV

FAMILY PLANNING VICTORIA
EXPLANATION, CONSENT AND QUESTIONAIRE

1.0 Title of Project: FOURIER TRANSFORM INFRARED SPECTROSCOPY IN THE DIAGNOSIS OF CERVICAL CANCER.

2.0 Chief Investigators:

Dr. D McNaughton, Department of Chemistry, Monash University, Clayton. Ph: 9905 4525. Fax: 9905 4975.

Prof. Michael Quinn, Department of Obstetrics and Gynaecology, Melbourne University. Ph: 9344 2000.
Fax: 9347 1761.

3.0 Description of Project:

The present methods of screening for cervical cancer produce a number of false negatives and false positives and around 40% of these arise form errors in laboratory diagnosis. These tests also involve complicated procedures, are time consuming and expensive. Fourier Transform Infrared (FTIR) Spectroscopy has recently been shown to be very promising as an inexpensive and rapid tool for the diagnosis of cervical cells. A NH&MRC Grant has enabled us to explore the possibility of using FTIR spectroscopy as a screening tool for cervical cancer by examining the cells left on instruments used to take your smear test.

4.0 Participation in the study:

Your participation in this study is completely voluntary. You have the right to withdraw from the study at any time.

5.0 Confidentiality of Records:

All records in this study will be kept confidential. Your identity will not be known to anyone except those caring for you at Family Planning Victoria. The investigators performing this study will work with samples identified by a code kept at Family Planning Victoria.

6.0 Possible risks, inconveniences and discomforts;

NONE.

7.0 Benefits:

The study does not offer any direct benefits to you, but may eventually help other women by enabling the introduction of an objective screening method for the detection of cervical cancer.

8.0 Understandings:

I have chosen to take part in this research study and give my consent on the understanding that:

- The research will be carried out in a manner conforming to the principles set out by the National Health and Medical Research Council.
- I have received, read and understood information contained in the attached documents about the general purpose of the study, its methods, requirements, possible risks, inconveniences and discomforts.
- I understand that refusal to take part in the study will not affect the quality of my further medical care.
- I am volunteering to take part in this study and understand that I may withdraw at any time.
- The Family Planning Victoria Inc. Ethics Committee has approved this research.
- I have had the opportunity to ask questions about this study and the answers given have been to my satisfaction.
- If at any time during the study I need to obtain further information I am free to telephone Dr. Don McNaughton (9905 4525), or Melissa Romeo (9905 5721).

I, (name) ....

of (address)..

have received, read and understood detailed information about the above study and have decide to participate as confirmed by my signature below.

You will receive a copy of this form for your record.

Signature .... Date..............................

Witnessed by... Date..............................

PATIENT QUESTIONAIRE

Title: FOURIER TRANSFORM INFRARED SPECTROSCOPY IN THE DIAGNOSIS OF CERVICAL CANCER.

1. Patient Number: ....
2. Date of Birth: ....
3. Method of contraception (if oral contraception is it mono- or triphasic)?:
   ....
4. Number of Pregnancies: ....
5. Smoker (YES/NO): ....
6. Day of cycle (1st day of bleeding is taken as day 1): ....
7. Usual Length of cycle (including menstruation): ....
8. Usual Length of menstruation: ....
9. Do you use tampons? (YES/NO): ....
10. Have you engaged in sexual intercourse in the last 48 hours?: ....

All information is anonymous and confidential. Thank you for your cooperation.

APPENDIX K

INTERPRETATION OF SIMCA RESULTS [12]

SAMPLE TO MODEL DISTANCE

The Coomans plot, Figure K.1, shows the orthogonal distances from the new objects to two different classes or models. The membership limits are indicated for a user defined level of significance. Samples that fall within the membership limits are said to belong to that class.

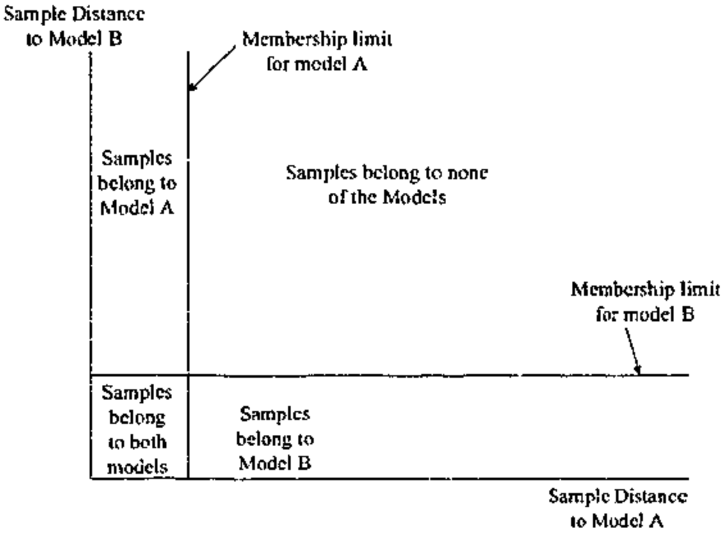
Sample Distance to Model B
Membership limit for model A
Samples belong to Model A
Samples belong to none of the Models
Membership limit for model B
Samples belong to both models
Samples belong to Model B
Sample Distance to Model A

Figure K.1 Coomans plot interpretation

MODEL DISTANCE

The model distance can be used to visualise the distance between one class and the other class or classes. By definition, the distance from a class or model to itself is 1. The distance to other classes should be greater than 3 for good separation between classes.

DISCRIMINATION POWER

The discrimination power of a variable shows how much each *X* variable (wavenumber value) contributes towards separating two classes. A discrimination power near 1 indicates the variable is of no use in separating the classes. Discrimination power should be greater than 3.

MODELLING POWER

The modelling power gives an indication of the relevance of a variable, ie how much the variance of the variable is used to describe the class model. Modelling power is between 0 and 1. A variable with a modelling power higher than 0.3 is said to model the class well.

OUTLIER DETECTION

PCA utilises three methods for identifying and removing outliers from data:

1. Scores plots show sample patterns according to one or two components. Samples lying far away from other samples are likely to be outliers.
2. Residuals measure how well samples or variables fit the model. A sample with a high residual is poorly described by the model and is an outlier.
3. Leverages measure the distance from the projected sample (model approximation) to the centre. Samples with high leverages have a stronger influence on the model than other samples, but may not necessarily be an outlier. An influential outlier has both a high residual and leverage and should be removed.

REFERENCES (APPENDICES)

1. Mathews, C.K. and K.E. van Holde, *Biochemistry*. CA: The Benjamins/Cummings Publishing Company. 1990.
2. Roberts, M.B.V., Structure and Function of Cells, in *Biology. A Functional Approach*. Thomas Nelson and Sons Ltd: Edinburgh. 1982.
3. Koss, L., The Papanicolaou test for cervical cancer detection: A triumph and a tragedy. *JAMA*, 1989. **261**: p. 737-743.
4. Mayeaux, E.J., S.D. Spigener, and J.A. German, Cryotherapy of the uterine cervix. *The Journal of Family Practice*, 1998. **47**(2): p. 99-102.
5. Riotton, G. and W. Christopherson, *Cytology of the Female Genital Tract*. Vol. 8. Geneva: World Health Organisation. 1973.
6. Koss, L., *Diagnostic Cytopathology and its Histopathologic Basis*. 2 ed. Philadelphia: J B Lippincott Company. 1968.
7. *Cervical Cancer*. National Cancer Institute. 1996. http://www.nci.nih.gov/
8. DeMay, R., *The Art and Science of Cytopathology*. Vol. 1. Chicago: American Society of Clinical Pathology Press. 1995.
9. Wied, G.L., Evaluation of endocrinologic condition by exfoliative cytology, in *Textbook of Gynecologic Endocrinology*, J. Gold, Editor. New York: Hoeber Medical Division. p. 133-184. 1968.
10. Adams, M., *Chemometrics in analytical chemistry*. Cambridge: The Royal Society of Chemistry. 1995.
11. *OPUS / IR Version 2.0 Reference Manual*. Bruker Analytische Messtechnik GMBH: Germany. 1995.
12. *The Unscrambler User Manual*. CAMO ASA: Sweden. 1998.