First published in *Physics in Medicine and Biology*, volume 51, issue 10 (2006).
Published by Institute of Physics Publishing Ltd.

Wavelet-based feature extraction applied to small-angle x-ray scattering patterns from breast tissue: a tool for differentiating between tissue types.

G Falzon †‡, S Pearson†, R Murison‡,C Hall§, K Siu‖ *, A Evans¶, K Rogers+, R Lewis‖ **

† Physics and Electronics, School of Biological, Biomedical and Molecular Sciences, University of New England, Armidale, NSW 2351, Australia.
‡ School of Mathematics, Statistics and Computer Science, University of New England, Armidale, NSW 2351, Australia.
§Daresbury Laboratory, Warrington, Cheshire, WA4 4AD, UK.
‖ School of Physics, Monash University, Victoria 3800, Australia
* Department of Medical Imaging and Radiation Sciences, Monash University, Victoria 3800, Australia
** Monash Centre for Synchrotron Science, Monash University, Victoria 3800, Australia
¶City Hospital Nottingham, NG5 1PB, UK.
+ Department of Materials and Medical Sciences, Cranfield University, Swindon, SN6 8LA, UK.

E-mail: gfalzon@pobox.une.edu.au

Abstract. This paper reports on the application of wavelet decomposition to SAXS patterns from human breast tissue produced by a synchrotron source. The pixel intensities of SAXS patterns of normal, benign and malignant tissue types were transformed into wavelet coefficients. Statistical analysis found significant differences between the wavelet coefficients describing the patterns produced by different tissue types. These differences were then correlated with position in the image and have been linked to the supra-molecular structural changes that occur in breast tissue in the presence of disease. Specifically, results indicate that there are significant differences between healthy and diseased tissue in the wavelet coefficients that describe the peaks produced by the axial d-spacing of collagen. These differences suggest that a useful classification tool could be based upon the spectral information within the axial peaks.

AMS classification scheme numbers: 42C40

Submitted to: *Phys. Med. Biol.*

1. Introduction

The association between the supra-molecular structural order of the protein collagen and breast cancer is receiving an increasing amount of scientific attention. Collagen, a major structural component of the extra-cellular matrix (ECM) in breast tissue, is known to be heavily involved in the progression of breast cancer (Raymond & Leong 1991; Kaupilla *et al* 1998; Pucci-Minafra 1998; Wang *et al* 2002). Small-Angle X-ray Scattering (SAXS), a well-established technique that provides structural information on length scales from 10-1000 nm, is highly sensitive to this supra-molecular structure, particularly when experiments are performed using the high flux provided by a synchrotron X-ray source. Differences between normal, benign and malignant breast tissue SAXS patterns have been related to the supra-molecular structural order of collagen (Lewis *et al* 2000; Fernandez *et al* 2002). It has been suggested that this method offers insight into the mechanisms of tissue invasion and has the potential to be used as a diagnostic test for malignancy. The research reported in this paper presents a novel method for analyzing the SAXS images produced, whilst also providing a means for tissue classification.

Typical SAXS patterns from healthy breast tissue contain sharp maxima in the diffraction pattern in the meridional (vertical) direction, produced by coherent scattering from the staggered arrangement of collagen molecules. Results indicate that these peaks are reduced in height above the background, and increased in width when malignant tissue is present (see for instance the SAXS patterns contained within Lewis *et al* (2000)). This suggests a significant reduction in the ordered collagen structure in the vicinity of breast tumours. There is also the possibility that the spread in the peak is due to the formation of a novel form of collagen, OF/LB collagen (Pucci-Minafra 1993). Analysis using mathematical wavelets (Mallat 1998) is ideally suited to these patterns, efficiently describing the sharp axial peaks whilst also allowing analysis of the pattern at several levels of spectral resolution. Indeed wavelet-based spectral analysis of SAXS diffraction patterns has already been used to identify breast tissue malignancy (Erickson 2003), the detection of malignancy being based upon the magnitude of the wavelet coefficients at a particular resolution level. It is highly desirable, however, that a link be formed between the wavelet classification and the systematic structural differences observed in the SAXS patterns. Understanding the fundamental differences in the wavelet decomposition for different tissue states (ie normal or malignant), informs the wavelet-based classification process. Furthermore, a classification based upon specific structural changes in breast tissue will aid the clinical acceptability of such a diagnostic system.

Our research hypothesis has been developed upon the following facts:

i Significant structural changes occur in collagen in the presence of breast tissue malignancy (Fernandez *et al* 2002; Wang *et al* 2002).

ii Significant differences are observed in the intensity of the diffraction maxima of SAXS patterns produced by normal, benign and malignant breast tissue (Lewis *et al* 2000; Fernandez *et al* 2002).

iii Research suggests that the differences observed in the SAXS patterns of normal, benign and malignant breast tissue can be linked to structural changes in collagen (Lewis *et al*

2000; Fernandez *et al* 2002).

iv Successful identification of breast tissue malignancy can be achieved using wavelets (Erickson 2003).

This work aims to identify spatially where significant differences in wavelet coefficients of the different diffraction patterns are occurring, linking this with pathological changes in the collagen present. It is proposed that significant differences occur in the intensity of the diffraction maxima between tissue types and that this is a key feature in the success of the wavelet classifier. Extraction of the wavelet coefficients associated with the diffraction peaks and the examination of the statistical descriptors allow an identification of any such differences. Furthermore, statistical projection techniques may be used to demonstrate that useful information on the differences between tissue types can be obtained using wavelets.

This paper reports on results from the above analysis. Section 2 discusses the use and appropriateness of wavelet decomposition for this application. Section 3 then details the statistical analyses used, giving a justification for choice of wavelet basis and an examination of the wavelet decomposition of the various tissue types in order to determine where differences occur. Statistical test results are presented that indicate significant differences occur between the wavelet coefficients of the different tissue types used. These differences occur amongst features that are produced by the supra-molecular structure of the protein collagen.

2. Method

2.1. Data set

The SAXS patterns used in this study were those reported in Lewis *et al* (2000).
The data consisted of:

- 20 SAXS patterns of apparently normal human breast tissue. This is labelled the 'Normal' data set.
- 22 SAXS patterns of breast tissue with invasive carcinoma present. This is labelled the 'Malignant' data set.
- 7 SAXS patterns of breast tissue containing fibroadenoma. This is labelled the 'Benign' data set.

The SAXS experiments were performed in September 1999 at the Daresbury Synchrotron Radiation Source, Station 2.1. The tissue samples were either core-cut biopsy specimens or tissue samples from mammoplasty patients, informed consent and ethics approval having been given. Full experimental details may be found in Lewis *et al* (2000). Wavelet decomposition was applied to 512 x 512 pixel digital images of the SAXS patterns. Statistical analysis of the wavelet coefficients was conducted in order to compare tissue groups. This section provides an overview of the sequential steps in the analysis.

2.2. Wavelet-based Feature extraction

The observable in SAXS patterns is the intensity due to coherently scattered radiation registered on a detector at a sufficiently large enough distance away to be considered in the Fraunhofer diffraction region. The intensity can be given as a function of scattering vector $\boldsymbol{h}$ as,

$$I(\boldsymbol{h}) = F(\boldsymbol{h})F^*(\boldsymbol{h}) = \int\int \rho(\boldsymbol{r_1})\rho(\boldsymbol{r_2}) \exp[i\boldsymbol{h}\cdot(\boldsymbol{r_1}-\boldsymbol{r_2})]dV_1 dV_2 \qquad (1)$$

(Koch 1991)
where $h = 4\pi\ \sin\theta/\lambda$, 2θ is the scattering angle, λ is the incident x-ray wavelength and $\rho(\boldsymbol{r})$ is the electron charge distribution of the scattering object. The observable on the detector can be considered the power spectrum of the electron density distribution of the scattering object under examination. The SAXS pattern is therefore a spectral representation of the object under study.

As has been mentioned, results presented in Lewis *et al* (2000) and Fernandez *et al* (2002) highlighted systematic differences in features within SAXS patterns from normal, benign and malignant breast tissue samples. These features have been specifically related to the supra-molecular structural arrangement of collagen. Several of these features are of interest when considering the use of wavelets as a means of analysing the patterns. The first set of features of interest arise from the well known periodic axial d-spacing, a consequence of the staggered arrangement of tropocollagen molecules within the fibrils (Bigi & Roveri 1991). For the case where the collagen fibres are partially aligned in the vertical direction in the capillary tube (as in our experiment), scattering from the axial d-spacing structure results in a series of maxima and minima in the meridional direction of the pattern. These features are referred to as the axial peaks. The second set of peaks, occurring in the equatorial region of the pattern, arises from the lateral packing of the fibrils in a quasi-hexagonal lattice (Eikenberry *et al* 1982a, 1982b). The diffracted intensity in the equatorial region can be approximated by a Bessel function. This set of features will be referred to as the Bessel peaks. The SAXS patterns acquired using tissue containing normal or malignant breast tissue show distinct differences in the sharpness and hence spectral content of the axial and Bessel features.

The wavelet decomposition allows the SAXS patterns to be decomposed into a variety of spatial frequency bandwidths whilst still allowing particular peaks to be identified. Therefore, changes in the spectral content of a particular feature between tissue types can be clearly identified. The discrete wavelet transform (DWT) effectively decorrelates the image data and inference about features in the wavelet domain is done using a decomposition of energy. This is akin to the statistical technique, analysis of variance. Energy is defined in the wavelet domain as,

$$E = \frac{1}{N}\Sigma_{j=1}^{N}||d_j||^2 \qquad (2)$$

where d_j is the j_{th} wavelet coefficient describing the image. The advantage of processing in the wavelet domain as compared to the image domain is that a greater proportion of energy is

compacted into a few coefficients in the wavelet domain. The results of this property is that it is relatively straightforward to identify those coefficients that describe features of interest in the SAXS pattern. The statistical analysis is therefore based upon the differences in the energy of the wavelet decomposition of the SAXS patterns.

2.3. Overview of Statistical Analysis

Ten wavelet bases from the Daubechies family, with various degrees of smoothness were used to decompose the SAXS patterns, the objective being to find the most 'suitable' basis to apply to these patterns. Selection of the most suitable basis was determined to aide the feature extraction process.
Criterion for the choice of wavelet basis included:

(i) How effectively the basis captured the information present in the pattern.
(ii) The absence of 'artifacts' produced by the wavelet decomposition in the coefficient matrices.
(iii) The basis that provided the greatest apparent difference between features in the tissue groups.

One measure of basis efficiency is the integrated squared-error (ISE), which is defined as

$$ISE = \int\int (f_{i,j} - \hat{f}_{i,j})^2 dx\, dy \approx \sum_{i,j=1}^{n_i,n_j} (f_{i,j} - \hat{f}_{i,j})^2, \tag{3}$$

where $f_{i,j}$ = intensity value at a given pixel in the image and $\hat{f}_{i,j}$ = intensity value at a given pixel using the wavelet reconstruction. The ISE is therefore the sum of squares of the error between the original image and the wavelet reconstruction. To determine how efficiently a particular wavelet basis represented the pattern, all wavelet coefficients below the pth quantile were set to zero, the remaining wavelet coefficients then being used to reconstruct the image and the ISE calculated.

Wavelet coefficient matrices were also examined for artifacts. Several artifacts were found when using smoother wavelets bases such as the Daubechies wavelet with nine vanishing moments. These oscillations can be confused with edges in a pattern, and artifacts of this form are undesirable because they bias the estimates of the energy of the features. Each wavelet decomposition of a SAXS pattern produces a total of 262144 coefficients, describing eight resolution levels that each have three matrices describing three different directions (horizontal, vertical and diagonal). Perspective plots of wavelet coefficient matrices were used to identify:

(i) The basis with the greatest apparent differences between tissue groups.
(ii) Those sets of coefficient matrices most likely to contain significant differences between tissue groups.

Based on these criterion, the Haar basis was selected for the decomposition. Apparent differences between tissue groups were identified in the horizontal and vertical coefficient matrices from resolution level four upwards.

The important features that would explain the 'apparent' differences between tissue types were selected in a two-stage process. The first stage was to use the Kolomogrov-Smirnov (K-S) test to compare the distributions of the energies of the tissue groups across resolution levels. This approach was able to screen the data to indicate which wavelet coefficient matrices deserve closer scrutiny. The K-S test examines the maximum vertical difference between two empirical distributions and is used to estimate the probability that the observed difference is due to random sampling (or alternatively due to tissue types). Significant contrasts were identified in many of the wavelet coefficient matrices, indicating that differences in the energy of specific features should be examined in these matrices.

To probe further, we used the wavelet energy of specific crystals (Bruce & Gao 1999). The wavelet decomposition can be described in terms of 'atoms' and 'crystals' by analogy to lattice structures. An 'atom' is an individual frequency-location (coefficient at a particular resolution level and position) and a 'crystal' is a collection of 'atoms'. The densities of the energies of crystals corresponding to axial and Bessel features were examined for each tissue type. Apparent differences between tissue types were found in both the axial and Bessel features from resolution levels four to eight.

The second step involved determining whether the observed differences in the energies amongst tissue types are real or simply a result of random sampling. The wavelet coefficients corresponding to specific features in the scattering pattern were analysed using a Generalised Linear Model (GLM) (McCullagh & Nelder 1990), the expected value of the response being related to the linear predictor by the log-link function. The energies of 49 features describing the axial and Bessel peaks (found in those matrices selected by exploratory analysis and the K-S test) were anlaysed. In all cases, model checking diagnostics indicated that the model was reliable. Results from this analysis are presented in the next section. The third and fifth axial axial peaks showed significant differences amongst all groups from resolution levels four to eight, suggesting that these features may be useful for classification purposes.

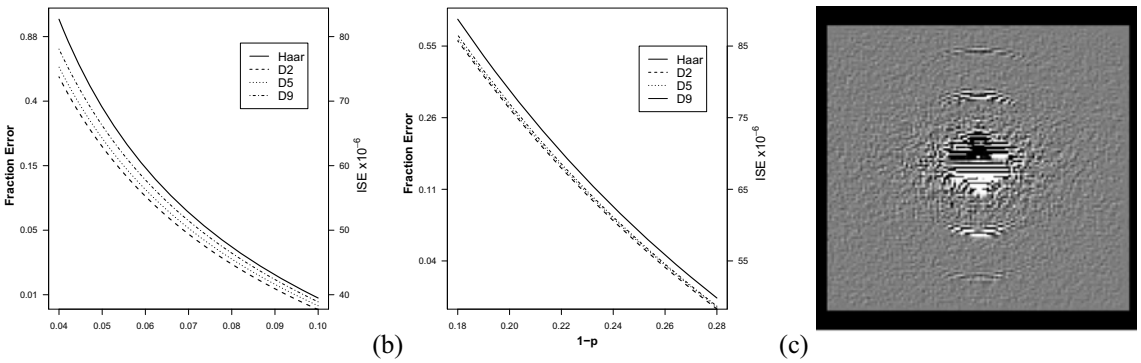
Projection Pursuit was used to find interesting low-dimensional projections of the high-dimensional multivariate feature set. The projection pursuit index was defined such that an 'interesting' projection corresponds to one in which there are large differences between observations between different groups. Further details of the algorithm may be found in Lee *et al* (2002). Our results clearly demonstrated that projections of the data can be found that provide discrimination between tissue types and also projections that provide interesting insights into the data. The next section presents results from this analysis.

3. Results and Discussion

3.1. Choice of Wavelet Basis

Results from the test for basis efficiency (calculated using the ISE) are shown in Figure 1(a)-(b), the curve of ISE vs 1-p decaying quadratically to zero as more coefficients are included in the image reconstruction. The basis with the curve that approaches zero (perfect reconstruction) the fastest is deemed the most efficient. The plot in Figure 1(a) displays one

Figure 1. Comparing wavelet bases with the ISE:
(a) ISE over selected quantiles:Normal Tissue.
(b) ISE over selected quantiles:Malignant Tissue.
(c)Artifacts produced using the Daubechies Nine basis.



such curve for Normal tissue over the range of quantiles likely to be associated with features of interest. Note that the x-axis is in units of $1-p$, so that the value of $1-p=0.05$ corresponds to the removal of all the wavelet coefficients below the 95th quantile in magnitude. A small ISE value at a particular quantile implies that the reconstructed image is a good approximation to the original. Therefore, removal of the wavelet coefficients below this quantile does not degrade the image quality significantly and the discarded coefficients are unlikely to contain important information. In contrast, a large ISE implies that the wavelet coefficients that were removed convey important information. Figures 1(a) and (b) indicate that an acceptably small value of ISE is reached for all bases using the upper 10% $(1-p=0.1)$ of all the wavelet coefficients for Normal tissue and the upper 28% for Malignant tissue. Inspection of both plots suggests that the Daubechies wavelet basis with two vanishing moments (D2) may have a significant advantage over the other bases.

Artifacts were observed when the patterns were decomposed using the Daubechies wavelet basis with nine vanishing moments (D9). Figure 1(c) displays the level 7 horizontal wavelet coefficients. The important axial features are evident but 'Gibbs-like' oscillations are also present (Mallat 1998). Examination of wavelet coefficient matrices for different bases suggests that these oscillations are related to the smoothness of the wavelet basis used in the analysis; the D9 basis having the greatest number of visual artifacts whilst the Haar basis the least.

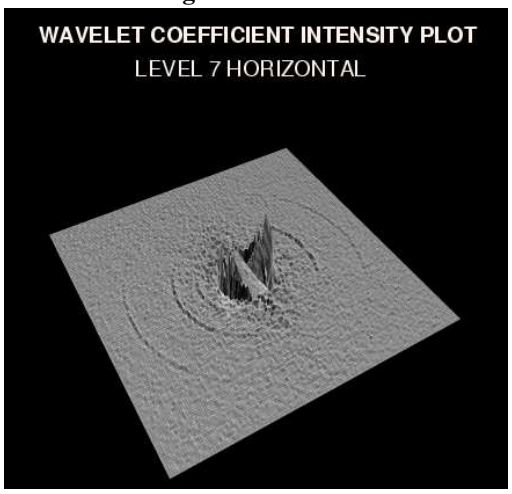
ISE curves were created and inspected for all SAXS patterns in this study. A similar result to Figures 1(a)-(c) was found to exist, and the Haar basis was selected since: it has the smallest support; it is the least smooth of all wavelet bases considered (and thereby believed to most accurately describe the sharp peaks in Normal tissue data); the use of the Haar basis removed artifacts found with the other bases. This last consideration was highly important since one can then be confident that the wavelet coefficients extracted are an accurate reflection of the true feature values. The cost of using the Haar basis is that there are many large coefficients across several resolution levels. This makes it difficult to isolate those spectral components that are different between tissue types. A highly competitive choice is

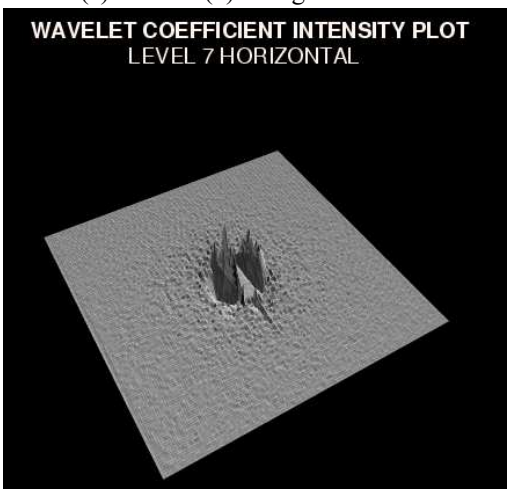
the D2 wavelet basis, and future work will examine the issue of basis selection further.

A number of other criteria for the selection of a wavelet basis exist such as the optimal removal of noise and suitability for image compression. These criteria were not addressed in this study but will also be the subject of future research.

3.2. Exploratory Data Analysis

Figure 2. Wavelet Coefficient Matrix Plots: (a) Normal (b) Malignant

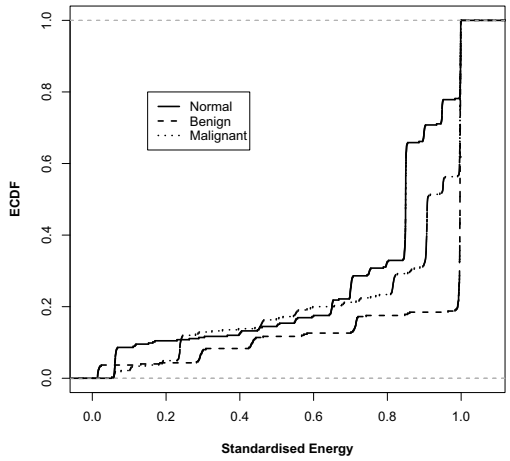
(a) 

(b) 

The perspective plot of wavelet coefficient matrices is a useful exploratory tool. For example, Figures 2(a)-(b) display the horizontal wavelet coefficients for a Normal and Malignant sample at the 7th resolution level. The magnitude of the wavelet coefficient at each pixel is represented by it's height in the plot. It is clear that differences in these plots are associated with the axial peaks. Closer inspection at other projections suggests that differences also exist in the Bessel peaks. This is encouraging as differences in these two peaks can be interpreted in terms of changes in the collagen fibres. Similar differences were found in resolution levels four to eight. This result indicates that the wavelet coefficients highlight edges in the patterns and that differences in sharpness of these edges, represented by differences in wavelet coefficient magnitude, may be used as a means of differentiating between tissue states. The next two sections describe statistical methods used to test whether these differences are significant and effective as a method of classification.

3.3. Kolomogrov-Smirnov Test

The empirical cumulative distribution function (ecdf) is an estimator of the underlying distribution function $P\{X \leq x\}$. It is the fraction of those values in the sample that are less than or equal to the value, x. In this case, the value of x is the energy of the wavelet coefficients. Figures 3(a)-(b), display the ecdf over a range of energy values for the level 5 vertical and horizontal wavelet coefficients averaged over the sample values in each group. The empirical cumulative distribution functions of two samples are compared by analysing the



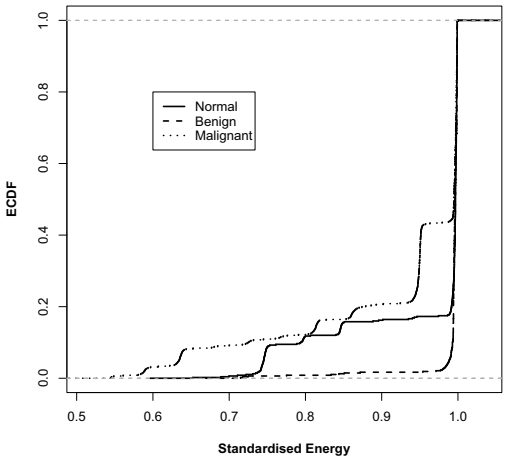


Table 1. K-S test statistics for significant contrasts
(where * indicates that the probability of a difference arising due to chance is less than 0.001).

	Horizontal					**Vertical**				
Level	4	5	6	7	8	4	5	6	7	8
N-M	0.07	0.11*	0.09*	0.08*	0.09*	0.16*	0.12*	0.11*	0.08*	0.04*
N-B	0.12	0.07*	0.10*	0.08*	0.04*	0.17*	0.08*	0.11*	0.08*	0.04*
B-M	0.10	0.12*	0.08*	0.06*	0.05*	0.16	0.06*	0.07*	0.04*	0.06*

D-statistic which is the maximum vertical distance between the values of the sample ecdfs. That is,

$$|D_n| = |\max(\mathbf{F}_{\Pi_1}(x) - \mathbf{F}_{\Pi_2}(x))|, \tag{4}$$

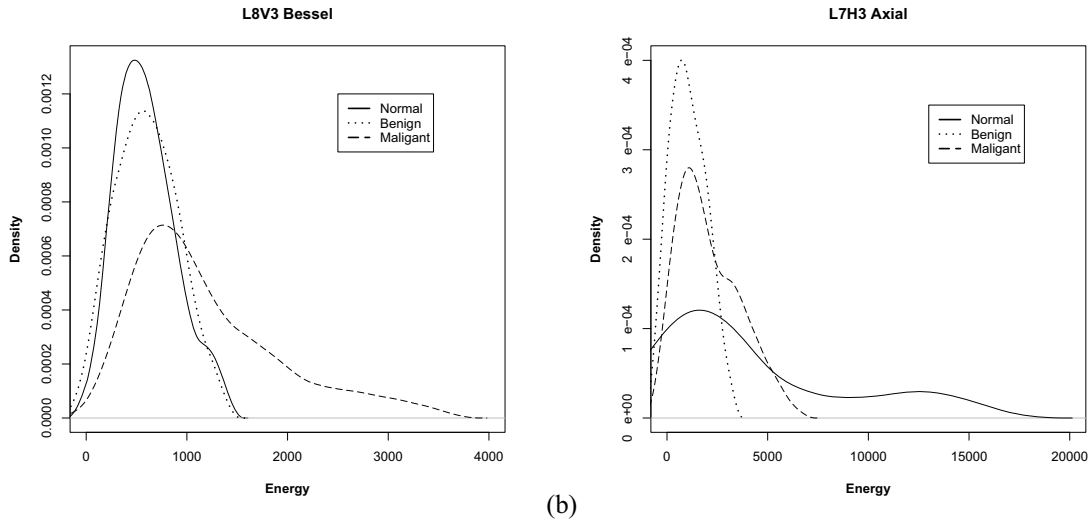
where **F** refers to the empirical distribution function, Π_1 and Π_2 refer to groups one and two respectively.

The pairwise test of differences between distributions of wavelet coefficients from the three tissue types are summarised in Table 1. This table shows the D-statistic for the contrasts, 'Normal' and 'Malignant' (N-M), 'Normal' and 'Benign' (N-B) and 'Benign' and 'Malignant' (B-M), for the horizontal and vertical wavelet coefficient matrices at the 4th, 5th, 6th, 7th and 8th resolution levels. The larger the D-statistic the larger the maximum difference in the ecdfs between the two groups. For example, the B-M level 5 horizontal contrast in Table 1 displays a statistically significant difference of D = 0.12. Therefore, the level 5 horizontal ecdfs between the benign and malignant groups differ in values of some point by up to twelve percent of the total. As seen in Figure 3 this implies that the ecdfs between benign and malignant tissue have a different rate of change as the energy increases. In other

words, one group has more wavelet coefficients of a certain absolute value of magnitude than the other. This then implies a spectral difference at this resolution level. Inspection of Table 1 indicates that there is evidence for very strong statistically significant differences between groups from resolution level four upwards. No strong statistical differences were found at any level in the diagonal direction. Comparisons were not performed below level four because the resolution is too low to adequately identify features of interest. From this analysis, we can focus our modelling using wavelet coefficients from frequency level four upwards.

3.4. Generalised Linear Models

Figure 4. Density Plot of Spectral Energy for the (a)third bessel and (b) third axial peak.



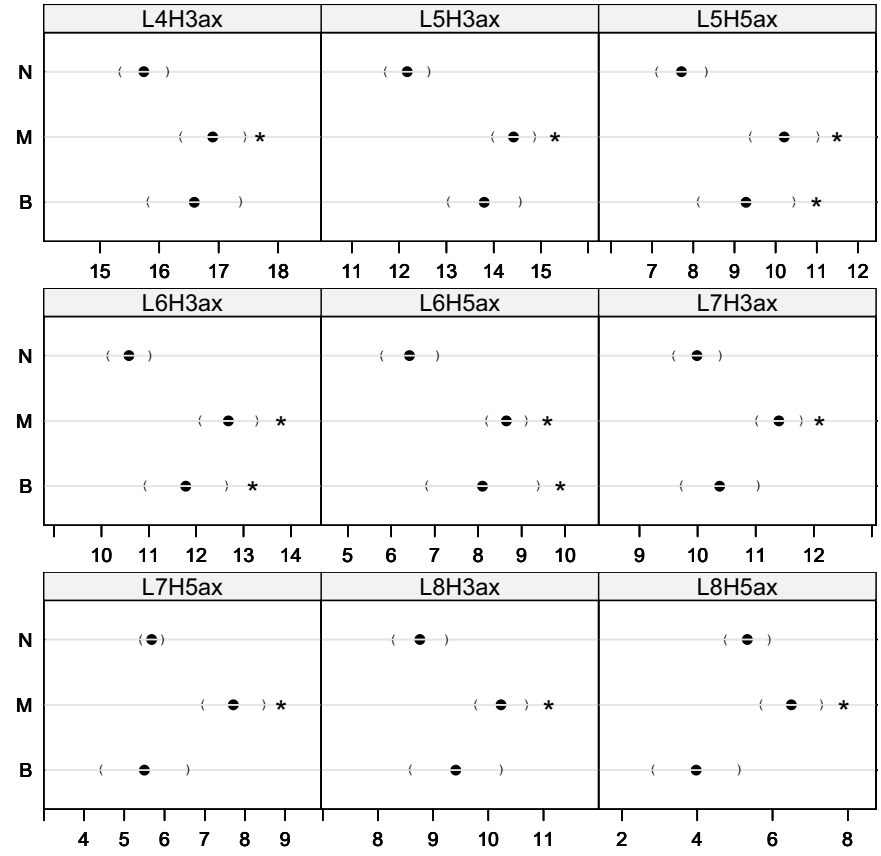
(a) (b)

Figures 4(a)-(b) show the distribution of spectral energies for Normal, Benign and Malignant tissue using wavelet coefficients from (a) the third Bessel peak and (b) the third axial peak region of the SAXS images. These figures illustrate that the energy reveals differences amongst tissue types at different sites (Bessel and axial peaks) and resolutions (level 8 in (a) and 7 in (b)). Interpretation of Figure 4(a) suggests that the malignant group contains high-frequency spectral components in this Bessel feature that are of larger magnitude than the other groups. A physical interpretation of this evidence is that there is a high frequency structure present in the 3rd-order Bessel peak, which may be more sharply-defined than for the other tissue types. It is noted that the density observed in the Malignant tissue samples actually overlaps the other groups. Possible explanations for this could be related to the amount of collagen in a particular sample or that some of the samples only contain a small percentage of malignant tissue and when imaged most of the signal is in fact from normal tissue.

In contrast, Figure 4(b) suggests that the Normal group has much greater high-frequency spectral components in the axial features than malignant tissue, indicating that at this

frequency it is probable that normal tissue has a much more sharply defined edge in the axial feature. Once again there is overlap of the density distributions, which will have implications in future work on classification. It is likely that any future classification system based upon spectral energy features will have to consider multiple features to be accurate. Apparent spectral differences between tissue types were found in both the axial and Bessel features from resolution levels four to eight.

Figure 5. 95% Confidence Intervals for the logarithm of energy statistics of features found to have significant differences,
* indicates a significant difference from the Normal contrast at the $p = 5\text{x}10^{-5}$ level.



To determine whether the observed group differences are real or simply a result of random sampling, the wavelet coefficients corresponding to specific features were analysed with a Generalised Linear Model (GLM). The GLM was used to find significant differences in the third and fifth axial peaks between all tissue groups at a range of resolution levels. Results are presented in Figure 5. The key reads as follows, L(N) indicates the Nth resolution level, (H,V,D) indicates the direction (horizontal, vertical or diagonal), (3,5) gives the orders of the peaks concerned whilst (ax and B) distinguishes the axial from the Bessel peaks. Since changes in the axial d-spacing of collagen are linked with malignant conditions, the finding of

significant differences in wavelet energy amongst tissue types suggests a wavelet classification model based on physical structures that are known to change in the presence of malignancy. These figures indicate that the lower resolutions, levels four to six, appear to be useful in distinguishing Normal tissue from Malignant tissue, whilst higher resolution levels (seven and eight) appear to be useful in distinguishing Benign from Malignant tissue.

No convincing evidence was found for significant differences between tissue types that are associated with the Bessel peaks above resolution level four. Exploratory work, as in Figure 4(a) suggests that the Bessel peaks may be a useful indicator when a sufficiently large proportion of malignant tissue is present in the sample. The differences or lack thereof may be related to the choice of wavelet basis. The use of the Haar basis may have made the differences in the axial peaks apparent but at the cost of the detection of any differences in the Bessel peaks. Analysis was also limited by the fact that it was not possible to isolate specific features related to the structure of collagen in wavelet coefficient matrices below resolution level four. The above issues will form the basis of future research. Results are now presented that demonstrate that the extracted features are useful for classification.

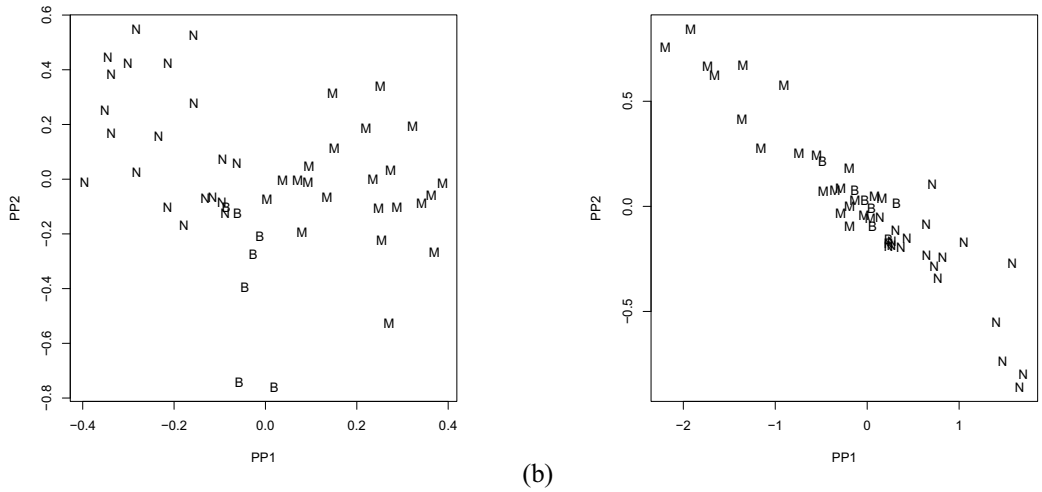
3.5. Exploratory Supervised Classification

In the GLM, the response was the wavelet energies and the explanatory variable was the tissue type. For classification we wish to predict the category of a future observation based upon the wavelet energy of selected features. Projection pursuit regression (PPR) was used to find interesting low-dimensional projections of the high-dimension multivariate feature data set. Such projections can yield important insights into the data set. Figure 6(a) displays a two-dimensional projection of the entire data set based upon the maximum of the LDA-index (Lee *et al* 2002). We see that this projection separates the tissue classes quite well, suggesting that an accurate classification system based upon these features may be able to be produced. Interpretation of the combination of features that yielded a particular projection is quite difficult, but the dominant effects will be stated. Projection one (PP1) in Figure 6(a) includes the dominant effects of the amorphous background scatter and the level four axial peaks (L4H3ax & L4H5ax) interacting with the high-resolution (L5,L7,L8) Bessel peaks. This combination appears to be an excellent discriminator of tissue type. The second projection (PP2) in Figure 6(a) corresponds to a mixture of both high and low resolution axial (L6H5ax,L2H135ax) and Bessel features (L8V1B,L8V2B,L8V3B). This projection direction appears to differentiate the Benign class from the Normal. These results suggest that classification of tissue type can be achieved by considering the interaction of the axial features, Bessel and background scatter features.

Figure 6(b) is another ‘interesting’ two-dimensional projection based upon a local maximum of the LDA index. It is seen that a linear relationship has been found between the two projections that approximately separates the Normal and Malignant tissue classes. The first projection direction (PP1) corresponds to a dominant influence of the Level 7 Vertical 3rd Bessel feature (L7V3B), other strong influences includes high-resolution axial features(particularly L6H3ax) and the Bessel feature, L6V23B. The second projection

direction has very strong axial feature influence (particularly L8H1ax) but again includes high resolution Bessel peak features. This projection suggests that a representation can be found that progressively 'charts' the deterioration of normal breast tissue towards abnormal histopathology as a function of the information present in breast tissue SAXS patterns.

Figure 6. Projections found using exploratory projection pursuit.
(a) Global Maximum of LDA index. (b) Local Maximum of LDA index.



(a) (b)

The GLM indicated that the axial peak features were important for classification. Projection Pursuit has shown that both the low-resolution wavelet coefficients (that cannot be attributed to a specific physical structure) and the interaction of Bessel and axial features provide important information about the changes that occur in breast tissue in the presence of malignant disease.

4. Conclusion

This paper has reported on the application of wavelet decomposition analysis to SAXS patterns as a means of uncovering statistically significant differences between patterns from 'Normal', 'Benign' and 'Malignant' tissue. Wavelet coefficients have been used to seek correlations between the SAXS patterns and the disease state of the tissue. These have been combined with statistical tests in order to indicate whether the differences between tissue state are statistically significant and indeed related to changes in collagen structure. Significant differences in the spectral energy of the axial features of 'Normal', 'Benign' and 'Malignant' tissue types have been found. Furthermore, projection pursuit regression suggests that the low-resolution wavelet coefficients and the interactions between axial and Bessel peak features may provide additional discriminatory power. These results suggest that a useful classification tool can be based upon the spectral content of these peaks. This work also supports the hypothesis that differences in SAXS patterns for different tissue states are produced by changes in the collagen structure. The next stage in this project is the

development of an accurate wavelet-based classification system using the SAXS patterns in which the classification is linked to the supra-molecular structural changes that are known to occur in breast tissue in the presence of disease.

Acknowledgments

The authors would like to acknowledge the support of the CCLRC Daresbury Synchrotron Radiation Source, and Dr E Kendall and Ms C Erickson of the University of Saskatchewan, Canada for their initial work and support in this area. This research was partially supported by the Access to Major Research Facilities Program, funded by the Commonwealth of Australia, and ongoing support has been provided by the University of New England, Australia. This work was prepared using the R statistical environment (R Development Core Team 2004) using the lattice (Deepayan 2004), Wavethresh (Nason *et al* 2004) and ClassPP (Lee 2002b) packages. Finally, the primary author would like to thank and give appreciation to the many people who have assisted in the preparation of this document.

References

Bigi A and Roveri N 1991 Fibre diffraction: collagen *Handbook of Synchrotron Radiation* edited by S.Ebashi, M.Koch & E.Rubenstein **4** (Amsterdam: North- Holland).

Bruce A and Gao HY 1996 *Applied Wavelet Analysis with S-PLUS*(New York:Springer-Verlag).

Deepayan S 2004 *lattice: Lattice Graphics*. R package version 0.10-14.

Eikenberry EF, Brodsky BB, Craig AS, Parry DAD 1982a Collagen fibril morphology in developing chick metatarsal tendon: 1. X-ray diffraction studies *Int. J. Biol. Macromol.* **4** 322-28.

Eikenberry E F, Brodsky B and Parry D A D 1982b Collagen fibril morphology in developing chick metatarsal tendon: 2. Electron microscope studies *Int J Biol Macromol* **4** 393-98

Erickson C 2003 Automated Detection of Breast Cancer Using SAXS Data and Wavelet Features, PhD Progress Report (Division of Biomedical Engineering, University of Saskatchewan, Canada).

Fernandez M, Keyrilinen J, Serimaa R, Torkkeli M, Karjalainen-Lindsberg M-L, Tenhunen M, Thomlinson W, Urban V, Suortti P 2002 Small-angle X-ray scattering studies of human breast tissue samples *Phys. Med. Biol.* **47** 577-92.

Kauppila S, Stenbäck F, Risteli J, Jukkola A, Risteli J 1998 Aberrant type I and type III collagen gene expression in human breast cancer in vivo *J Pathol* **186(3)** 262-8.

Koch MHJ 1991 *Scattering from Non-Crystalline Systems: Handbook on Synchrotron Radiation*, edited by S.Ebashi, M.Koch & E.Rubenstein, **4**, 241-67. Amsterdam: Elsevier.

Lee E-K, Cook D, Klinke S, Lumley T 2005 Projection Pursuit for Exploratory Supervised Classification *SFB 649 Discussion Paper*. Available at http://141.20.100.9/papers/pdf/SFB649DP2005-026.pdf

Lee E-K, Cook D, Klinke S, Lumley T 2002 *classPP: Projection Pursuit for supervised classification.* R package version 1.0.

Lewis RA, Rogers KD, Hall CJ, Towns-Andrews E, Slawson S, Evans A, Pinder SE, Ellis IO, Boggis CRM, Hufton AP, Dance DR 2000 Breast cancer diagnosis using scattered X-rays, *J. Synchrotron Rad.* **7** 348-52

Mallat S 1998 *A Wavelet Tour of Signal Processing* (London:Academic Press).

McCullagh P and Nelder JA 1989 *Generalised Linear Models. Second Edition.* (London:Chapman and Hall).

Nason G, Kovac A and Maechler M 2004 *Wavethresh: Software to perform wavelet statistics and transforms*.

Pucci-Minafra I, Andriolo M, Basiricò L, Alessandro R, Luparello C, Buccellato C, Garbelli R, Minafra S 1998 Absence of regular α_2(I) collagen chains in colon carcinoma biopsy fragments *Carcinogenesis* **19(4)** 575-84.

Pucci-Minafra I, Luparello C, Andriolo M, Basiric'o L, Aquino A and Minafra S 1993 A new form of tumor and fetal collagen that binds laminin *Biochemistry* **32(29)** 7421-7.

R Development Core Team 2004. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Raymond W and Leong A-Y 1991 Assessment of invasion of in breast lesions using antibodies to basement membrane components and myoepithelial cells *Pathology* **32** 291-97.

Venables WN and Ripley BD 1998, *Modern Applied Statistics with S-Plus, 2nd Edition* (New York:Springer-Verlag).

Vidakovic B 1999 *Statistical Modeling by Wavelets*(New York:Wiley).

Wang W, Wyckoff JB, Frohlich VC, Oleynikov Y, H*ü*ttelmaier S, Zavadil J, Cermak L, Bottinger EP, Singer RH, White JG, Segall JE, Condeelis JS 2002 Single cell behaviour in metastatic primary mammary tumours correlated with gene expression patterns revealed by molecular profiling *Cancer Research* **62** 6278-88.

First published in *Physics in Medicine and Biology*, volume 51, issue 10 (2006).
Published by Institute of Physics Publishing Ltd.