

Biomedical Physics & Engineering Express



NOTE

A constituent-based preprocessing approach for characterising cartilage using NIR absorbance measurements

OPEN ACCESS

RECEIVED
27 July 2015REVISED
13 November 2015ACCEPTED FOR PUBLICATION
30 November 2015PUBLISHED
18 January 2016Cameron P Brown¹ and Minsi Chen²¹ Botnar Research Centre, NDORMS, University of Oxford, Old Road, Oxford OX3 7LD, UK² Department of Computing and Mathematics, University of Derby, Kedleston Road, Derby DE22 1GB, UKE-mail: cameron.brown@ndorms.ox.ac.uk**Keywords:** near infrared spectroscopy, cartilage, osteoarthritis

Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

**Abstract**

Near-infrared spectroscopy is a widely adopted technique for characterising biological tissues. The high dimensionality of spectral data, however, presents a major challenge for analysis. Here, we present a second-derivative Beer's law-based technique aimed at projecting spectral data onto a lower dimension feature space characterised by the constituents of the target tissue type. This is intended as a preprocessing step to provide a physically-based, low dimensionality input to predictive models. Testing the proposed technique on an experimental set of 145 bovine cartilage samples before and after enzymatic degradation, produced a clear visual separation between the normal and degraded groups. Reduced proteoglycan and collagen concentrations, and increased water concentrations were predicted by simple linear fitting following degradation (all $p \ll 0.05$). Classification accuracy using the Mahalanobis distance was $>98\%$ between these groups.

1. Introduction

Osteoarthritis (OA) is a degenerative joint disease that carries substantial personal and societal burden in its later stages. The development of OA involves a changes across the whole joint, with the loss of structural integrity of articular cartilage a central factor in the early disease process. Damage or degeneration detectable by visible light in common diagnostic procedures such as arthroscopy, however, are generally indicative of more severe structural degradation. Thus, detection of early stage degradation in articular cartilage is of special clinical interest to allow intervention before the disease burden manifests.

Spectroscopy-based techniques have been widely applied to characterise biological tissues, with Raman and mid-infrared spectroscopy commonly used [1]. Near infrared (NIR) spectroscopy, which measures harmonics and combinations of the fundamental molecular vibrations in the infrared energy range, is gaining popularity [2–6]. NIR compares favourably with other techniques for *in vivo* application [7], and has strong potential as a clinical tool. As NIR spectroscopy works with standard xenon light sources and fibreoptics, measurements can be taken with minimal instrument modification and without extra radiation delivered to the patient.

A major restriction to the use of NIRS, however, is the difficulty in interpretation in tissues due to the wide, overlapping spectral bands of the main extracellular matrix components, and thus a lack of independent peaks from which to base analysis. The high dimensionality of spectral data poses a further challenge to avoid overfitting, and requires very high sample numbers for robust training. A common and plausible solution is to pre-select optimal subsets of the spectrum for training. Selection methods based on wavelet coefficient regression and a genetic algorithm [8], mutual information and B-spline compression [9, 10], and mutual information and a modified genetic algorithm [11] have been successfully implemented.

Here, we take a relatively simple and physically based approach, using a Beer's Law dimensionality reduction method to resolve spectral measurements into chemical concentrations. Intended as an input to more sophisticated predictive tools, this was coupled with linear least squares fitting, combining the second derivative spectra of matrix components to estimate their relative concentrations in the tissue. We tested our method's capability to probe matrix composition in articular cartilage, first using a sample set of normal and enzymatically degraded bovine cartilage-on-bone.

2. Methods

2.1. NIR spectroscopy experiment

Absorbance spectra from 145 sample points on 6 macroscopically normal bovine patellae samples were collected using a Nicolet FT-NIR system (Thermo-Nicolet, UK) with a standard Michelson interferometer light path. The 4 mm diameter fibre optic probe was coupled to the FT-NIR system via a Grasby SPECAC NIR fibre port accessory (SPECAC, Orpington, UK), offset from the surface by ≈ 1.5 mm, based on the maximum signal from a diffuse reflectance standard. The spectrum was acquired from 4000 to 12 500 cm^{-1} , with each measurement averaged over 124 scans (≈ 40 s). Patellae were immersed in 0.15 M saline between measurements. Samples were then enzymatically digested in 0.1 mg ml^{-1} of trypsin (from bovine pancreas T4665, Sigma–Aldrich, Australia) dissolved in 0.15 M phosphate buffered saline for 4 hours at 37 °C and retested. Removal of proteoglycans in the trypsin-treated samples was confirmed by safranin-O staining [12].

Spectra from isolated matrix components were collected using a Nicolet iS50 FT-NIR system with a standard SabIR fibre optic probe (Thermo Fisher Scientific, UK), offset from the surface by ≈ 4 mm, again based on maximum signal. Spectra of collagen (C9879, Sigma Aldrich, UK), chondroitin sulphate (C9819, Sigma Aldrich, UK) and distilled water were recorded and averaged over 64 scans (≈ 10 s). Cartilage and component spectra were taken in different laboratories, thus necessitating the use of different spectrometers. All spectra were reduced to the 4000–10 000 cm^{-1} (1–2.5 μm) region, with 779 data points.

2.2. Data pre-processing

Data smoothing and derivation

The discrete absorbance measurements were first fitted with a cubic B-spline as discussed in [13]. The number of knots and their initial placement were determined by the uniqueness of each data point, with two consecutive data points x_i and x_{i+1} considered unique if their difference is within a specified tolerance τ , i.e. $|x_{i+1} - x_i| \leq \tau$. An empirically obtained noise model and an explicitly defined degree of freedom were used to preserve desired data characteristics. Equation (1) is the piecewise cubic B-spline fitting, c_i is a control point derived from the data points and $B_{i,3}$ is the i th cubic basis function

$$A(\lambda) = \sum_{i=1}^N c_i B_{i,3}(\lambda). \quad (1)$$

The use of piecewise fitting allows us to control the preservation of certain localised features in the data. Furthermore, since our subsequent dimensionality reduction method is dependent on the second derivative of the spectral data, this can be more conveniently

achieved by taking the desired derivative of the basis function (see equation (2))

$$\frac{d^2 A}{d\lambda^2} = \sum_{i=1}^N c_i B_{i,1}(\lambda). \quad (2)$$

Dimensionality reduction

One of the most challenging aspects of performing classification on spectroscopic data is its high dimensionality. The NIR spectrum covers the range between 800 and 2500 nm, and a typical input per spectrum is of the order of 1000 data points. The parameterisation of each sample over the NIR spectrum inevitably results in a large feature space. This also increases the number of training samples required for training a classifier.

Our approach aimed to reduce dimensionality by basing our analysis on the main constituents of the tissue: water, collagen and proteoglycan. Given the absorptivity of each individual component, the relationship between a mixture and its constituents can be described using Beer's law. Equation (3) describes the specific formulation for modelling the absorbance of cartilage

$$A(\lambda_i) = c_w l \epsilon_w(\lambda_i) + c_c l \epsilon_c(\lambda_i) + c_p l \epsilon_p(\lambda_i) \quad (3)$$

$A(\lambda_i)$ is the absorptivity at wavelength λ_i ; ϵ_w , ϵ_c and ϵ_p are the absorptivity of water, collagen and proteoglycan respectively; c is the concentration of each corresponding component, i.e. c_w for water, c_c for collagen and c_p for proteoglycan; l is the path length. The second order derivative of A is:

$$\frac{d^2 A}{d\lambda^2} = c_w l \frac{d^2 \epsilon_w}{d\lambda^2} + c_c l \frac{d^2 \epsilon_c}{d\lambda^2} + c_p l \frac{d^2 \epsilon_p}{d\lambda^2}. \quad (4)$$

If the absorbance of a mixture and the absorptivity its constituents are known across the NIR spectrum, the concentration of each component can be calculated by solving the following linear system:

$$\begin{bmatrix} \frac{d^2 \epsilon_w}{\lambda_0^2} & \frac{d^2 \epsilon_c}{\lambda_0^2} & \frac{d^2 \epsilon_p}{\lambda_0^2} \\ \frac{d^2 \epsilon_w}{\lambda_1^2} & \frac{d^2 \epsilon_c}{\lambda_1^2} & \frac{d^2 \epsilon_p}{\lambda_1^2} \\ \vdots & \vdots & \vdots \\ \frac{d^2 \epsilon_w}{\lambda_{k-1}^2} & \frac{d^2 \epsilon_c}{\lambda_{k-1}^2} & \frac{d^2 \epsilon_p}{\lambda_{k-1}^2} \end{bmatrix} \begin{bmatrix} c_w l \\ c_c l \\ c_p l \end{bmatrix} = \begin{bmatrix} \frac{d^2 A}{\lambda_0^2} \\ \frac{d^2 A}{\lambda_1^2} \\ \vdots \\ \frac{d^2 A}{\lambda_{k-1}^2} \end{bmatrix}. \quad (5)$$

The least-square solution of the overdetermined linear system in equation (5) was obtained using QR factorisation.

Our use of Beer's law omitted the path length (i.e. $l = 1.0$) because the variance of our bovine cartilage thickness is small (1.66 ± 0.25 mm) and thickness information is unavailable in proposed arthroscopic applications of the technique. This does imply that the solution for the component concentration levels is subjected to scaling. However, we demonstrate in the

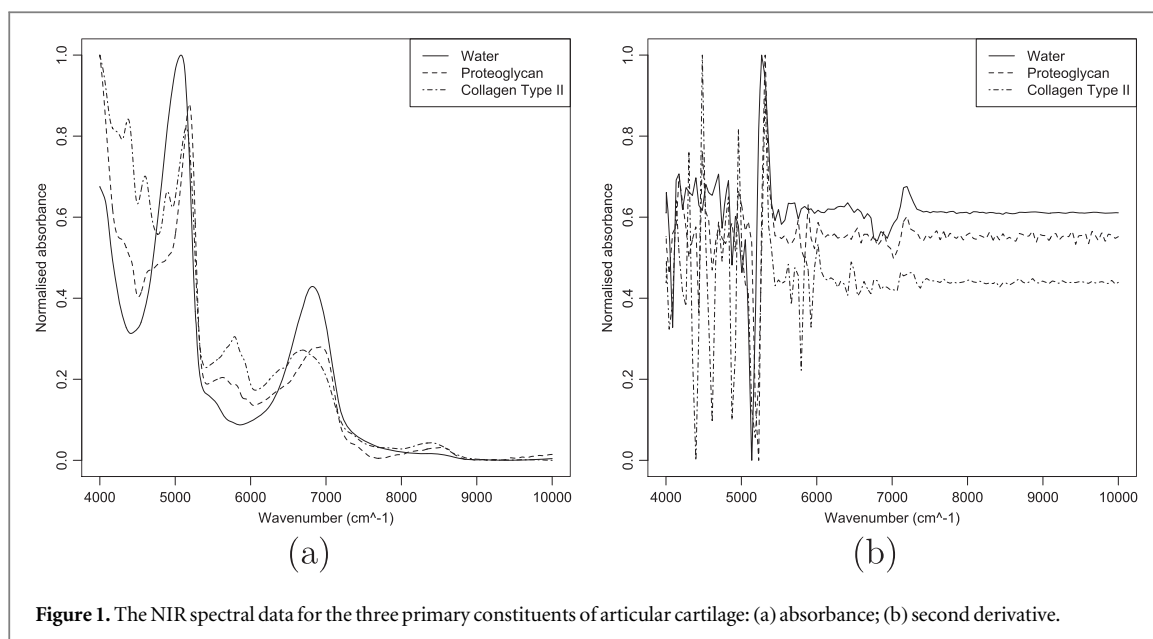


Figure 1. The NIR spectral data for the three primary constituents of articular cartilage: (a) absorbance; (b) second derivative.

result section that this scaling does not affect the accuracy of our classification method.

By estimating the concentration of individual components using the aforementioned approach, the spectral data can be mapped into a lower dimensional feature space consisting of the primary constituents. In the case of cartilage, the salient components water, collagen and proteoglycan allow each sample to be parameterised by a feature vector $\vec{c} = (c_w, c_c, c_p)$.

2.3. Classification

Based on results from unsupervised linear fitting described above, classification into normal and enzymatically degraded groups was performed based on the Mahalanobis distance. For a sample x , this was calculated as $D^2 = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the mean of each group and \mathbf{S} is the covariance matrix. Samples were classed as belonging to the training set to which the Mahalanobis distance was shortest.

The data pre-processing and classification methods described here were implemented in R 3.1.2 (<https://www.r-project.org>).

3. Results and discussion

Spectra of the matrix components (water, proteoglycan and collagen) are presented in figure 1. Spectra of cartilage-on-bone samples before and after enzymatic digestion are provided in figure 2. Raw spectra of normal and trypsin-treated samples are offset, however this offset is likely to include systematic experimental variations between measurements. To circumvent this effect, second-derivative spectra were used due to intrinsic normalisation as well as the improved separation of spectral peaks. Even with second-derivative spectra, the lack of peak independence can be observed in figure 2, despite the clear

differences in the isolated component spectra, thus prompting the mixture approach. The reader is referred to [4] for a description of peak assignments.

Unsupervised linear fitting of the three component spectra to sample spectra using equation (4) produced a clear visual separation of normal from enzymatically degraded samples. Using the Mahalanobis distance, normal and enzymatically degraded groups were classified with one false negative (0.34%) and four false positives (1.4%).

To examine the influence of omitting the thickness of individual sample, a path length randomly drawn from a normal distribution ($\mu = 1.66$, $\sigma = 0.25$) was assigned to each cartilage sample. This is equivalent to applying an arbitrary scaling to the solution of equation (5). This test was repeated $N = 100$ times where N was chosen to be 2/3 of the sample size in each cluster. Table 1 shows that a small variance can be observed from the centre of each cluster in the feature space when path lengths were factored in. However, this variance has a negligible impact on the stability of our classification method as shown in table 2. The mode of false positive and false negative count remained identical to the results where the path length was postulated to be one. It should be noted that degradation by trypsin has a negligible effect on sample thickness [14].

Visual trends (figure 3) for proteoglycan and water predictions conformed to expectations: enzymatic degradation reduced predicted proteoglycan content and increased predicted water content (both $p \rightarrow 0$), filling the space left by proteoglycan removal. Predicted proteoglycan and water content was negatively correlated ($r = -0.74$) for pooled sample groups. Predicted collagen content also decreased after enzymatic degradation ($p \rightarrow 0$). We interpret this as a secondary effect of trypsin on collagen by either restructuring in

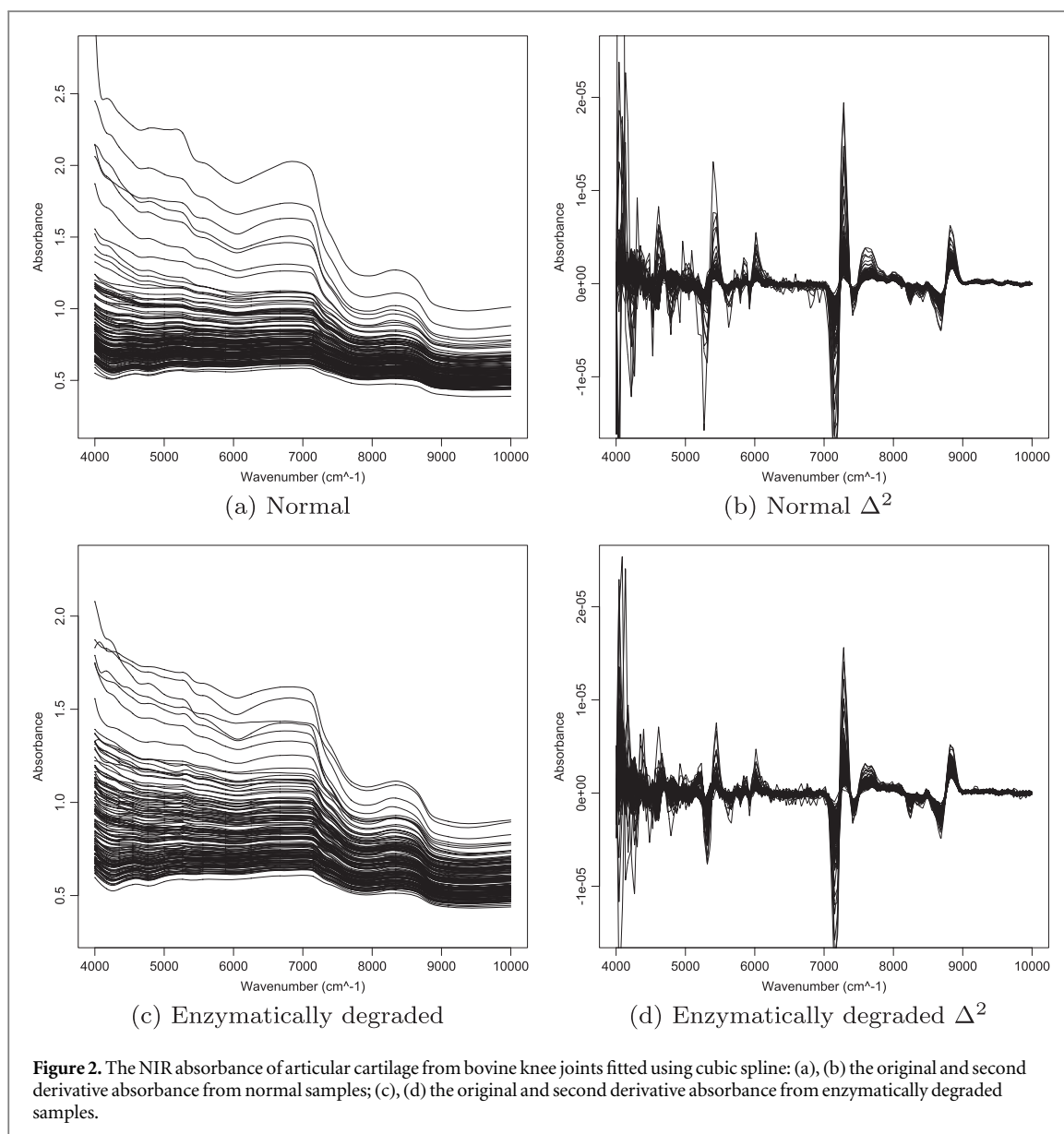


Table 1. The influence of scaling on the mean of each cluster $N = 100$.

Cluster	Mean c_w	Mean c_c	Mean c_p
Normal	0.1418 ± 0.0169	0.5949 ± 0.0446	0.5571 ± 0.0517
Degraded	0.4052 ± 0.0346	0.2486 ± 0.0316	0.2237 ± 0.0237

Table 2. The influence of scaling factor on the stability of classification performance; the worst case for false positive is 2.1% and for false negative is 1.7% $N = 100$.

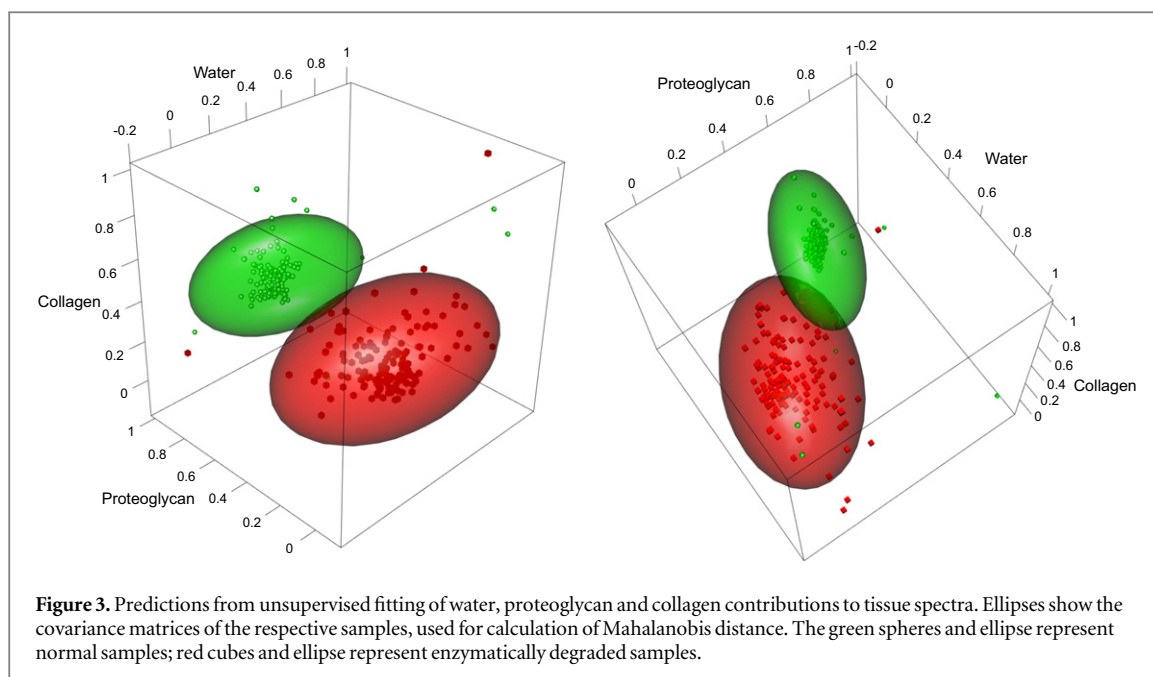
classification	Mode	Min	Max
False pos. # (%)	4 (1.4%)	3 (1.1%)	6 (2.1%)
False neg. # (%)	1 (0.34%)	0 (0%)	5 (1.7%)

the absence of proteoglycan affecting scattering or cleavage reducing collagen content.

It should be noted that study is limited by a lack of ground truth measurements of constituent concentrations in each sample. Safranin-O staining confirmed

loss of proteoglycan content following enzymatic degradation, however this was not quantified and pre-degradation staining was not possible. Solution stability in least squares fitting of bovine samples was confirmed by subsampling with spectral windows. Negligible fluctuation was observed in estimated concentrations, however, the optimisation and minimisation of spectral range for prediction was not explored here. Spectral optimisation and comparison of predictions with ground truth biochemical data will be undertaken in future work.

Applying Beer's law to second derivative spectra, we have developed a preprocessing technique with a



physical basis for dimensionality reduction to estimate concentrations of tissue matrix components from spectroscopy. Demonstrating its application to NIR spectroscopy of cartilage, we found that simple linear fitting of isolated second derivative spectra of water, proteoglycan and collagen could be used to separate normal from enzymatically degraded cartilage. We propose that this approach will provide a physically-based, low dimensionality input into more advanced classification strategies [15] such as an artificial neural network to predict absolute concentrations of matrix components. This approach may further be used in a range of soft tissues, such as cornea, skin, inter vertebral disk and tendon, in which the main matrix components are near infrared-active.

Acknowledgments

We gratefully acknowledge financial support from Arthritis Research UK (grant 20299 and Oxford EOTC), and the Oxford NIHR BRU in musculoskeletal disease.

References

- [1] Krafft C and Sergo V 2006 Biomedical applications of Raman and infrared spectroscopy to diagnose tissues *Spectroscopy* **20** 195–218
- [2] Spahn G, Plettenberg H, Nagel H, Kahl E, Klinger H M, Mückley T, Günther M, Hofmann G O and Mollenhauer J A 2008 Evaluation of cartilage defects with near-infrared spectroscopy (nir): an *ex vivo* study *Med. Eng. Phys.* **30** 285–92
- [3] Brown C P, Bowden J C, Rintoul L, Meder R, Oloyede A and Crawford R W 2009 Diffuse reflectance near infrared spectroscopy can distinguish normal from enzymatically digested cartilage *Phys. Med. Biol.* **54** 5579
- [4] Brown C P, Jayadev C, Glyn-Jones S, Carr A J, Murray D W, Price A J and Gill H S 2011 Characterization of early stage cartilage degradation using diffuse reflectance near infrared spectroscopy *Phys. Med. Biol.* **56** 2299
- [5] Padalkar M, Spencer R and Pleshko N 2013 Near infrared spectroscopic evaluation of water in hyaline cartilage *Ann. Biomed. Eng.* **41** 2426
- [6] Palukuru Uday P, McGoverin Cushla M and Pleshko N 2014 Assessment of hyaline cartilage matrix composition using near infrared spectroscopy *Matrix Biol.* **38** 3–11
- [7] Brown C P, Oloyede A, Crawford R W, Thomas G E R, Price A J and Gill H S 2012 Acoustic, mechanical and near-infrared profiling of osteoarthritic progression in bovine joints *Phys. Med. Biol.* **57** 547
- [8] Depczynski U, Jetter K, Molt K and Niemöller A 1999 Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm *Chemometr. Intell. Lab. Syst.* **47** 179–87
- [9] Rossi F, Lendasse A, François D, Wertz V and Verleysen M 2006 Mutual information for the selection of relevant variables in spectrometric nonlinear modelling *Chemometr. Intell. Lab. Syst. / J. Math. Background Chemometr. Intell. Lab. Syst.* **80** 215–26
- [10] Rossi F, François D, Wertz V, Meurens M and Verleysen M 2007 Fast selection of spectral variables with b-spline compression *Chemometr. Intell. Lab. Syst.* **86** 208
- [11] Fei Q, Li M, Wang B, Huan Y, Feng G and Ren Y 2009 Analysis of cefalexin with nir spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection *Chemometr. Intell. Lab. Syst.* **97** 127–31
- [12] Moody H R, Brown C P, Bowden J C, Crawford R W, McElwain D L S and Oloyede A O 2006 In vitro degradation of articular cartilage: does trypsin treatment produce consistent results? *J. Anat.* **209** 259–67
- [13] Gans P and Gill J B 1984 Smoothing and differentiation of spectroscopic curves using spline functions *Appl. Spectrosc.* **38** 370–6
- [14] Nieminen H J, Töyräs J, Rieppo J, Nieminen M T, Hirvonen J, Korhonen R and Jurvelin J S 2002 Real-time ultrasound analysis of articular cartilage degradation *in vitro* *Ultrasound Med. Biol.* **28** 519–25
- [15] Larrañaga P *et al* 2006 Machine learning in bioinformatics *Briefings Bioinformatics* **7** 86–112