2014

Selection of Preprocessing Methodology for Multivariate Regression of Cellular FTIR and Raman Spectra in Radiobiological Analyses

Aidan Meade
Technological University Dublin, aidan.meade@tudublin.ie

Colin Clarke
Technological University Dublin, Colin.Clarke@tudublin.ie

Hugh Byrne
Technological University Dublin, hugh.byrne@tudublin.ie

Fiona Lyng
Technological University Dublin, Fiona.lyng@tudublin.ie

Recommended Citation

Meade, A., Clarke, C. & Byrne, H. (2014). Selection of Preprocessing Methodology for Multivariate Regression of Cellular FTIR and Raman Spectra in Radiobiological Analyses. *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2-5 November 2014. doi:10.1109/BIBM.2014.6999164

Selection of preprocessing methodology for multivariate regression of cellular FTIR and Raman spectra in radiobiological analyses

Aidan D. Meade[1,2*], Colin Clarke[2], Hugh J. Byrne[3], Fiona M. Lyng[2]

1. School of Physics, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland.

2. Radiation and Environmental Science Centre, Focas Research Institute, Dublin Institute of Technology, Camden Row, Dublin 8, Ireland.

3. Focas Research Institute, Dublin Institute of Technology, Camden Row, Dublin 8, Ireland.

**Author to whom correspondence should be addressed:* aidan.meade@dit.ie

Abstract— **Vibrational spectra of biological species suffer from the influence of many extraneous interfering factors that require removal through preprocessing before analysis. The present study was conducted to optimise the preprocessing methodology and variable subset selection during regression of and confocal Raman microspectroscopy (CRM) and Fourier Transform Infrared microspectroscopy (FTIRM) spectra against ionizing radiation dose. Skin cells were γ-irradiated in-vitro and their Raman and FTIRM spectra were used to retrospectively predict the radiation dose using linear and nonlinear partial least squares (PLS) regression algorithms in addition to support vector regression (SVR). The optimal preprocessing methodology (which comprised combinations of spectral filtering, baseline subtraction, scaling and normalization options) was selected using a genetic algorithm (GA) with the root mean squared error of prediction (RMSEP) used as the fitness criterion for selection of the preprocessing chromosome (where this was calculated on an independent set of test spectra randomly selected from the dataset on each pass of the algorithm). The results indicated that GA selection of the optimal preprocessing methodology substantially improved the predictive capacity of the regression algorithms over baseline methodologies, although the optimal preprocessing chromosomes were similar for various regression algorithms, suggesting an optimal preprocessing methodology for radiobiological analyses with biospectroscopy. Feature selection of both FTIRM and CRM spectra using genetic algorithms and multivariate regression provided further decreases in RMSEP, but only with non-linear multivariate regression algorithms.**

Keywords—genetic algorithm, preprocessing, vibrational spectra, multivariate regression, radiobiology

I. INTRODUCTION

Evidence has accumulated in the recent literature of the wide variety of applications of vibrational spectroscopy in the elucidation and modelling of the effects of complex processes in the cell (viral transfection, changes to the extracellular matrix, effects of chemotherapeutic agents etc.) on its total biochemical composition[1-5]. Recent studies have also confirmed the applicability of infrared and Raman spectroscopies for the analysis of radiobiological effects at the cellular level[6, 7], particularly in relation to the retrospective prediction of radiation dose from Fourier Transform Infrared microspectroscopy (FTIRM) spectra of γ-irradiated cells[8].

In the retrospective prediction of radiation dose from vibrational spectra of the cell, as with many other applications, removal of spectral features that are not related to the biochemical composition of the cell is required. In FTIRM spectra, it is common to observe a broad oscillating baseline that has its origin in optical scattering effects (such as resonant and non-resonant Mie scattering) from subcellular organelles and other structures [9-13]. It has been shown that such effects can be modeled and extracted from the spectra using the extended multiplicative scatter correction (EMSC), including resonant effects [11, 14, 15]. Corrections for the absorptions of water vapour and CO_2 may be performed using machine-specific algorithms [16]. Spectral normalization, scaling and filtering may be used to account for point-to-point variations in biochemical composition in the sample and spectral noise respectively. Similar procedures are employed in respect of the spectral pre-treatment of confocal Raman microspectroscopic (CRM) data of the cell [1, 2, 17].

Preprocessing methods themselves have been demonstrated to affect the results of a classification [18, 19] or regression [20, 21] analysis of FTIRM or CRM spectral data and implies that an optimal preprocessing strategy must be employed. Selection of the optimal preprocessing strategy from a set of preprocessing options can proceed iteratively [18-20], or using evolutionary algorithms [21] which reduce the overall solution search time substantially. In this paper a genetic algorithm (GA) was used in an evolutionary search to establish an optimal preprocessing methodology and define the optimal set of preprocessing options for treatment of FTIRM and CRM with regression by various models against radiation dose. In addition, a multivariate analysis which employs the GA as a feature selection technique was used to further optimize the regression by the elimination of spurious variables. Three regression algorithms were chosen; a partial least squares regression (PLSR) algorithm and a non-linear version (NLPLSR) which respectively analysed spectral effects having a linear and quadratic relationship to radiation dose. The third

algorithm chosen was a support-vector regression algorithm (SVR) which analysed non-linear spectral effects occurring with dose, where those non-linearities could adopt any functional form. It was found that the SVR algorithm outperformed both PLSR and NLPLSR algorithms in prediction of radiation dose with feature selection, highlighting the non-linear nature of the spectral variation with dose and time after irradiation [8]. The change in the performance of the algorithms as a result of these treatments is highlighted.

II. Methods

2.1 Cell Culture and Sample Preparation

Human keratinocytes (HaCaT) were cultured in Dulbecco's MEM:F12 (1:1) whole medium (Sigma, Dorset, UK) supplemented with 10% fetal calf serum (Gibco, Irvine, UK), 1% penicillin-streptomycin solution 1,000 IU (Gibco, Irvine, UK), 2 mM L-glutamine (Gibco, Irvine, UK) and 1 μg/mL hydrocortisone (Sigma, Dorset, UK) in an incubator at 37˚C with 95% relative humidity and 5% CO_2. The cells were routinely subcultured at 80% confluency using a 1:1 solution of 0.25% trypsin and 1mM versene at 37˚C. Triplicate samples for FTIRM were prepared on MirrIR slides as detailed elsewhere[8] and were analysed at 6, 12, 24, 8 and 96 hours after irradiation with ten γ-radiation doses over the range from 0Gy to 5Gy. They were fixed in 4% formalin in phosphate buffered saline at each time point after irradiation and were stored in a desiccator until the time of analysis.

Triplicate samples for CRM were also prepared by depositing suspensions of 2.5×10^4 HaCaT cells onto fused quartz disks coated in a sterile solution of 2% w/v gelatin in dH_2O (the preparation of the coating and its polymerization on the quartz substrate is detailed elsewhere[3]) and cultured in DMEM-F12 with all supplements. The cells were allowed to effect initial attachment to the substrate for two hours and were then covered in fresh DMEM-F12 with all supplements. Approximately 24 hours after initial sample preparation the cells for FTIRM and CRM analysis were given γ-radiation doses over the range from 0Gy to 5Gy, and were fixed in 4% formalin at 96 hours after irradiation. Samples for CRM were stored in dH_2O at 4˚C until analysis.

2.2 FTIRM and CRM Measurements

FTIRM measurements were performed as detailed elsewhere[8]. Briefly, a Perkin-Elmer GX-II spectrometer was employed to record cell spectra over the 4000 to 720 cm^{-1} wavenumber range, using an aperture size of 100 μm ×100 μm, with a spectral resolution of 4 cm^{-1} and with 64 scans per spectrum. All spectra were recorded in transreflection mode with 300 spectra recorded at each dose and time point.

CRM data were acquired using a Horiba-Jobin Yvon HR-800 CRM spectrometer with a 785 nm laser as source. Spectra were acquired using a confocal hole diameter of 100μm and dispersion from a grating ruled with 300 lines/mm. The instrument was calibrated using the 520.7 cm^{-1} line of silicon. A spectrum of a neon lamp source was also taken as a reference for verification of the wavelength calibration of the spectrometer CCD detector. A water immersion objective with a ×100 magnification (Olympus LUMPlanFL 0.9 NA) was used for all spectral measurements, which were taken in dH_2O. Spectra of the quartz substrate were acquired in triplicate prior to, and at the end of, each measurement. CRM spectra of HaCaT cells at each dose point were acquired in a line scan across the cell with a step interval of 3μm such that spectra of the cell nucleus, cytoplasm and membrane were recorded. The spatial resolution of the system was determined to be approximately ±1.6 μm in separate measurements[22]. In the initial pre-processing of the CRM spectra, the signature of the quartz background was subtracted from all spectra and a rubberband algorithm, developed in house, was used to remove any residual baseline[2]. The line-scan spectra were then averaged for each cell to reduce further the measurement noise and provide spectra whose content comprised components from the membrane, cytoplasm and nucleus. No further preprocessing of either the FTIRM or CRM spectra was performed, although outliers were then removed in each dose category using Grubb's multivariate test for outliers[23].

2.3 Multivariate Regression and Genetic Algorithms

Multivariate regression against dose was performed using PLSR, NLPLSR and SVR regression algorithms. PLSR and NLPLSR algorithms were implemented in the Matlab v.7.2 environment (The Mathworks Inc., USA) with the PLS Toolbox v.5.0.3 (Eigenvector Research, Wenatchee, WA, USA). The SVR was implemented using the LIBSVM Toolbox [24]. Genetic algorithms were constructed using the Genetic Algorithm and Optimisation Toolbox [25], which allows the incorporation of binary and real valued genes within the GA chromosome.

2.4 Selection of Preprocessing Parameters

Selection of preprocessing parameters for multivariate regression was performed according to the method described by Jarvis and Goodacre [21]. A genetic algorithm constructed in Matlab was used to select from preprocessing options, whereby the GA chromosome contained genes that coded for each preprocessing option using a combination of binary digits and integers as shown in table 1. All of the preprocessing options were available for the preprocessing of FTIRM spectra, while all but the 'EMSC' option were made available to the GA for preprocessing of CRM spectra, as it was assumed that the rubberband correction algorithm removed much of the slowly varying background from the cellular Raman spectra.

TABLE I. STRUCTURE OF THE CHROMOSOME USED IN THE SELECTION OF OPTIMAL PREPROCESSING STEPS FOR PLSR, NLPLSR AND SVR WITH FTIRM AND CRM DATA; SG=SAVITSKY-GOLAY FILTERING, MA=MOVING AVERAGE FILTERING

Preprocessing Options	Gene Type	Possible Values
Derivation	Real Integer	1 – Off; 2 – 1st Order; 3 – 2nd Order
EMSC	Binary	0 – Off; 1 – On;
Filtering	Binary	0 – Off; 1 – On;
Filtering Type	Binary	0 – SG; 1 – MA;
SG window	Real Integer	Number from 7 to 21
SG order	Real Integer	Either 3 or 5
MA window	Real Integer	Number from 2 to 21
Normalisation	Real Integer	1 – Vector Area; 2 – Vector Length.; 3 – Min-Max
Scaling	Binary	0 – Off; Auto-scaling; 1 - Range-scaling;

In the GA algorithm, 60% of the total spectral data matrix was randomly selected for calibration of each of the regression models and the remaining 40% was retained for testing of the model with unseen data. This process was repeated for every execution of the regression algorithms. The RMSEP on the test set was used for evaluation of the performance of each regression model with a particular set of preprocessing parameters, and this constituted the fitness of the preprocessing chromosome, whereby the RMSEP was minimized during each evolution of the algorithm. At each initialization of the GA, the values of each of the genes were assigned randomly, and a total of twenty-five individual chromosomes with minimum RMSEP were selected for further evaluation. In total the GA was run for fifty separate initializations on each dataset, with thirty crossovers (p=0.6) and fifty mutations (p=0.05) per generation. The overall best chromosome of preprocessing parameters was determined from the median of the GA chromosomes giving the lowest RMSEP for each regression algorithm at the end of evolution. In defining the best value for the SVR regression parameters the gamma, γ (defining the regression kernel width) and penalty, C (defining an acceptable loss function for implementation of the regression) parameters were also assigned by the GA during evolution, via the incorporation of two extra genes to the chromosome in table 1.

2.6 Feature Selection Approaches

Feature selection was performed with PLSR, NLPLSR and SVR using a genetic algorithm (GA) constructed in Matlab. The method of Yoshida et al. [26, 27] was used for variable selection to prevent the overfitting that has been previously observed when GA's are used with a large search domain[26]. Briefly, a number of short GA runs were implemented for an evolution for 20 generations with 30 crossovers per generation (p>0.9) and 50 mutations per generation (p<0.05) to minimize the feature set selected by the algorithm. After this, a more extensive search (for 50 generations with the crossover and mutation rates as above) was performed on the feature set most often selected by the GA at the initial stage. A subset of one hundred spectra was randomly selected for calibration of the multivariate regression models with each preprocessing chromosome, and a separate one hundred spectra were also randomly selected for testing of the performance of the chromosome. Variables were encoded as binary digits. The fitness criterion for testing of the chromosome was the RMSEP of the regression with the unseen testing set of spectra. In total, the GA was run on fifty separate occasions.

III. RESULTS

3.1 Investigation of the effect of preprocessing parameters on regression performance

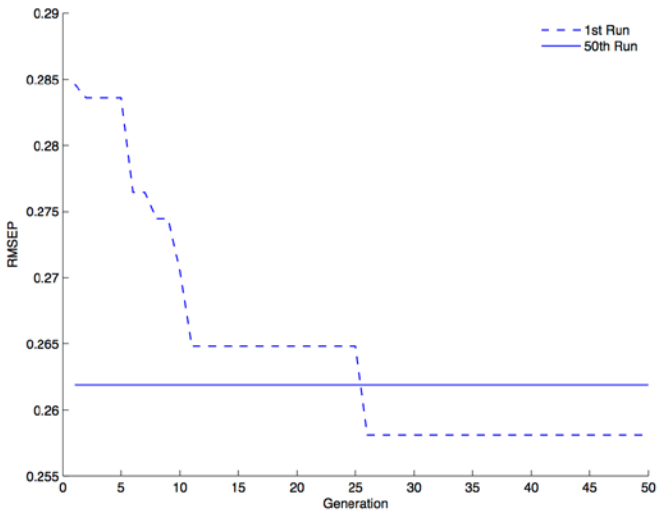

Figure 1. Typical evolution of a GA for selection of pre-processing parameters with PLSR RMSEP as fitness criterion utilizing FTIRM data at 6 hours post-irradiation. In the 50th execution of the GA, the optimal solution is found in the first generation of its evolution, and subsequently does not improve.

The content of the GA chromosomes that were selected for regression of FTIRM data versus radiation dose using PLSR, NLPLSR and SVR are shown in table 2. In this table a 'consensus' from the analysis (giving the most often chosen preprocessing parameters) is taken as the median of the GA chromosomes at the end of evolution at each time point. In addition, the *NA* entry in the tables indicates a preprocessing option that is not applicable by virtue of the parent option not having been selected (eg. the order of SG filtering is irrelevant if filtering has not been selected as an option by the GA). The associated change in regression performance with FTIRM spectra over baseline RMSEP values (taken from earlier work[8]) is shown in tables 3, 4 and 5 (where *SD* denotes the standard deviation on the mean). An example of the evolution in the RMSEP from two separate GA-PLSR executions is provided in Fig. 1 for illustration. This demonstrates that the algorithm approaches a consistent level of performance over the course of its evolution regardless of the randomly assigned values of the chromosome genes at the initiation of the algorithm, and as such implies that the preprocessing solution generated by the GA is optimal and consistent.

TABLE II. STRUCTURE OF THE CHROMOSOME USED IN THE SELECTION OF OPTIMAL PREPROCESSING STEPS FOR PLSR, NLPLSR AND SVR WITH FTIRM AND CRM DATA; SG=SAVITSKY-GOLAY FILTERING, MA=MOVING AVERAGE FILTERING

Time (hours)	6	12	24	48	96	Consensus
Preprocessing						
Derivation	Off	Off	Off	Off	Off	Off
EMSC	On	Off	On	On	On	On
Filtering	Off	Off	Off	Off	On	Off
Filtering Type	NA	NA	NA	NA	SG	NA
SG window	NA	NA	NA	NA	9	NA
SG order	NA	NA	NA	NA	5	NA
MA window	NA	NA	NA	NA	NA	NA
Normalisation	Vector	Min/Max	Vector	Min/Max	Vector	Vector
Scaling	Auto	None	Auto	Auto	Auto	Auto

TABLE III. IMPROVEMENT IN PLS RMSEP FOR FTIRM DATA THROUGH THE SELECTION OF OPTIMAL PREPROCESSING WITH THE GA. THE ORIGINAL PLS DATA IS TAKEN FROM EARLIER WORK[8]

Time	PLS RMSEP (Preprocessing) (Gy)	SD	Original PLS RMSEP (Gy)	SD	Percentage Change
6	0.26	0.01	0.31	0.02	-16
12	0.72	0.01	0.79	0.03	-8.8
24	0.33	0.01	0.33	0.02	0
48	0.52	0.01	0.46	0.02	+11
96	0.35	0.03	0.37	0.01	-13.5

TABLE IV. IMPROVEMENT IN NLPLS RMSEP FOR FTIRM DATA THROUGH THE SELECTION OF OPTIMAL PREPROCESSING WITH THE GA. THE ORIGINAL NLPLS DATA IS TAKEN FROM EARLIER WORK[8]

Time	NLPLS RMSEP (Preprocessing) (Gy)	SD	Original NLPLS RMSEP (Gy)	SD	Percentage Change
6	0.35	0.01	0.48	0.05	-27
12	0.64	0.19	0.76	0.06	-13
24	0.38	0.10	0.40	0.04	-5
48	0.43	0.15	0.46	0.03	-6.5
96	0.44	0.15	0.52	0.02	-15

TABLE V. IMPROVEMENT IN SVR RMSEP FOR FTIRM DATA THROUGH THE SELECTION OF OPTIMAL PREPROCESSING WITH THE GA. THE BASELINE SVR DATA WAS DETERMINED THROUGH REGRESSION OF SPECTRAL DATA AGAINST RADIATION DOSE, WITHOUT ANY FURTHER PREPROCESSING.

Time	SVR RMSEP (Preprocessing) (Gy)	SD	Baseline SVR RMSEP (Gy)	SD	Percentage Change
6	0.62	0.01	0.88	0.07	-29
12	0.94	0.01	1.08	0.06	-13
24	0.39	0.01	0.51	0.04	-23
48	0.54	0.01	0.70	0.05	-23
96	0.31	0.01	0.46	0.04	-33

The optimal preprocessing solutions for regression of the CRM data against radiation dose at 96 hours after irradiation are shown in table 6, together with the mean RMSEP after evolution of the GA-PLSR, GA-NLPLSR and GA-SVR algorithms in table 7 (where *SD* denotes the standard deviation on the mean). Each regression is performed separately with each individual algorithm, and a consensus estimate of the optimal preprocessing solution is again determined as a median of the GA solutions for each of the individual algorithms. Baseline performance for all algorithms is established through multiple evaluations (10 times each for PLSR and NLPLSR algorithms and 50 times for the SVR algorithm) with each regression algorithm on the raw spectral data. The spectral data matrix was randomly sorted on each pass of the algorithm

In respect of preprocessing of FTIRM data for multivariate regression, the consensus from table 2 is that the optimal solution is provided by using vector normalized and auto-scaled raw spectral data (i.e. not first or second derivative spectra) subjected to the extended multivariate scatter correction without filtering. Similarly, the consensus from table 6 is that the optimal preprocessing of CRM data for multivariate regression is provided by the use of raw spectral data that is not filtered. The consensus in relation to normalization of the data is in favour of the use of vector normalization.

TABLE VI. PREPROCESSING PARAMETERS FOR MULTIVARIATE REGRESSION OF CRM DATA AT 96 HOURS AFTER IRRADIATION SELECTED BY THE GA.

Time (hours)	PLS	NLPLS	SVR	Consensus
Preprocessing				
Derivation	Off	Off	1st Order	Off
EMSC	Off	Off	Off	Off
Filtering	Off	Off	On	Off
Filtering Type	NA	NA	SG	NA
SG window	NA	NA	19 points	NA
SG order	NA	NA	5	NA
MA window	NA	NA	NA	NA
Normalisation	Vector	Vector	None	Vector
Scaling	None	None	Auto	None

TABLE VII. PLS, NLPLS AND SVR RMSEP FOR CM DATA AFTER THE SELECTION OF OPTIMAL PREPROCESSING PARAMETERS WITH THE GA. THE BASELINE DATA WAS DETERMINED THROUGH REGRESSION OF THE RAW SPECTRAL DATA AGAINST RADIATION DOSE, AND ARE MEANS OF THE RMSEP FOR MULTIPLE EXECUTIONS OF EACH ALGORITHM AS DESCRIBED IN THE TEXT.

Algorithm	RMSEP (Gy)	SD	Baseline RMSEP (Gy)	SD	Percentage Change
PLS	0.32	0.01	0.43	0.02	-11
NLPLS	0.34	0.006	0.53	0.03	-36
SVR	0.26	0.03	0.40	0.03	-35

The improvement in regression performance accruing through the employment of selection of optimal preprocessing methodology is quite substantial in some instances, ranging from 5% to in excess of 30% of baseline RMSEP depending on the time point after irradiation and the regression algorithm in

question. This demonstrates that the identification of the optimal preprocessing methodology can improve the overall performance of the regression algorithm and should be considered as a component in the use of vibrational spectroscopic data for non-invasive radiological dosimetry. The consensus spectral processing methodologies for both FTIRM and CRM data have been used in treatment of both sets of data for the feature selection studies that follow.

3.2 Change in prediction of radiation dose with feature selection by GA

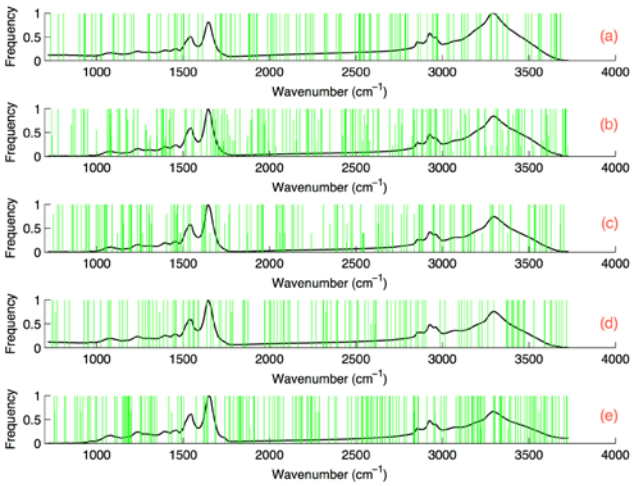

Figure 2. (a-e) Features of FTIRM spectra selected by GA-SVR from 6 hours (a) to 96 hours (e) after irradiation. The spectrum in black is the mean spectrum of control cells (0Gy) at each time point.

Selection of spectral features with the genetic algorithm involves minimization of the RMSEP with an independent set of test spectra, which should lead to an overall improvement in the prediction of radiation dose at each time point using each of the PLSR, NLPLSR and SVR algorithms. In the present analysis, the overall effect on prediction performance with GA feature selection and regression using either the GA-PLSR or GA-NLPLSR approaches was either marginal improvement or disimprovement of their performance. Contrastingly, the performance of the GA-SVR algorithm with GA feature selection increased after selection of the optimal preprocessing options. Features selected by the GA-SVR algorithms in regressing FTIRM data against radiation dose are shown in Fig. 2, while those selected by the GA-PLSR, GA-NLPLSR and GA-SVR algorithms in regressing CRM data versus radiation dose are shown in Fig. 3. In these figures the frequency with which a variable is selected by the algorithm is represented by the height of the bar.

From the data in table 8, it is clear that feature selection with the SVR algorithm increases the prediction performance at each time point for FTIRM data. This correlates well with the performance characteristics seen previously[8], where non-linear regression algorithms were seen to outperform both linear and linear-quadratic approaches in regressing spectral data versus radiation dose. It is clear therefore that selection of features that vary either linearly or according to a simple non-linear model with radiation dose only captures a small part of the spectral variation with dose, as it appears that most of the spectrum, or many of the spectral features at certain dose points, vary in a higher order non-linear manner with dose.

TABLE VIII. IMPROVEMENT IN SVR RMSEP FOR FTIRM DATA THROUGH VARIABLE SELECTION WITH THE GA, WHERE THE SELECTED VARIABLES AT EACH TIME POINT ARE THOSE DISPLAYED IN FIG. 2 (A-E). THE REFERENCE RMSEP VALUES ARE THOSE OBTAINED AFTER SELECTION OF PREPROCESSING METHODOLOGY (FROM TABLE 6)

Time	SVR RMSEP (Variable Selection) (Gy)	SD	SVR RMSEP (Preprocessing) (Gy)	SD	Percentage Change
6	0.30	0.02	0.62	0.01	-51
12	0.59	0.04	0.94	0.01	-37
24	0.31	0.02	0.39	0.01	-21
48	0.43	0.05	0.54	0.01	-20
96	0.35	0.03	0.31	0.01	+13

TABLE IX. IMPROVEMENT IN SVR RMSEP FOR CRM DATA THROUGH VARIABLE SELECTION WITH GA-SVR, WHERE THE SELECTED VARIABLES AT EACH TIME POINT ARE THOSE DISPLAYED IN FIG. 3(C). THE REFERENCE RMSEP VALUES ARE THOSE OBTAINED AFTER SELECTION OF PREPROCESSING METHODOLOGY (FROM TABLE 6)

Time	SVR RMSEP (Variable Selection) (Gy)	SD	SVR RMSEP (Preprocessing) (Gy)	SD	Percentage Change
96	0.096	0.004	0.26	0.03	-63

This is confirmed by the analysis of the CRM data at 96 hours after irradiation (table 9) in which a significant improvement in the performance of the prediction of dose with the SVR algorithm is observed after feature selection. A similar improvement in performance with the PLSR and NLPLSR algorithms was not seen.

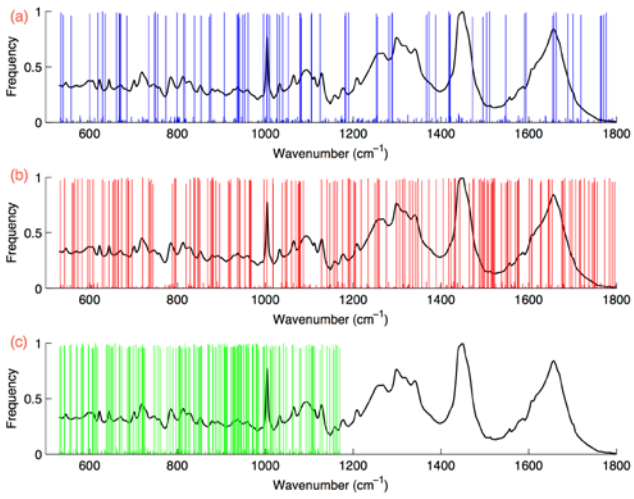

Figure 3. Features selected by (a) GA-PLSR, (b) GA-NLPLSR and (c) GA-SVR of CRM spectra against radiation dose at 96 hours after irradiation. The spectrum in black is the mean spectrum of control cells (0Gy) at 96 hours after irradiation.

Several important and interesting characteristics are apparent from the variable selection exercise that may have significance for the types of spectral effects observable after radiological damage of cells. In Fig. 2 variables within the FTIR spectra are selected by the GA-SVR algorithm that are both positioned at the peaks of the spectral bands and across their breadth also. This is suggestive of radiological damage having an effect on the breadth of spectral bands (broadening or narrowing) rather than positional shifts in the position of the peak of the band. In addition, a number of variables are selected which are associated with the remaining baseline in the spectra (between ~1780 cm^{-1} and ~2500 cm^{-1}) where no features of biochemical origin are present. A broad undulating feature in the baseline of FTIR spectra has previously been seen, having its origin in non-resonant and resonant Mie scattering effects [12, 13]. This scattering, when non-resonant, produces a broad curved baseline over the whole spectrum, whose curvature has a dependence on the diameter of the transparent scattering object within the cell (which can be any cellular organelle) [12]. It is well known that radiation damage can generate transparent subcellular membrane-bound bodies termed 'blebs' which encapsulate components of the cell and may scatter IR light in a similar manner to that observed with non-resonant Mie scattering. In the present work the EMSC algorithm was intentionally employed for scatter correction without correction for resonant Mie effects. It is possible, therefore, that the selection of spectral variables associated with the baseline in the in the 1780 cm^{-1} to 2500 cm^{-1} region is due to Mie scattering as a result of radiation-induced cellular blebbing.

It is also a point of interest in Fig. 3 that the features selected by GA-PLSR and GA-NLPLSR algorithms are distributed across the Raman spectrum while those selected by the GA-SVR algorithm are concentrated in the region containing strong modes of vibration associated with nucleic acids and their residues. This is not the case for the corresponding FTIR data in figure 2, where variables are selected corresponding to all molecular species including protein, lipid and nucleic acids. However, it has been demonstrated that variables selected in data with a high degree of covariation are highly dependent on the classification or regression algorithm and the wrapping algorithm [28]. In addition GA's do not consider any relationship between adjacent spectral variables but merely attempt to minimize a target classification or regression variable. In this context the GA selection of any particular set of variables in the FTIR data would not be expected to correlate molecularly with those selected in the Raman data. Overall the Raman variables selected in Fig. 3c suggest that molecular changes having a non-linear relationship to radiation dose at 96hrs after irradiation are predominantly associated with nucleic acids, and are perhaps due to structural modifications in DNA that are connected to mechanisms of ionizing radiation damage and repair.

These results highlight that selection of an optimal preprocessing methodology and selection of a feature subset can improve the performance of regression algorithms for radiobiological dosimetry using vibrational spectra.

IV. Conclusion

This study demonstrates that FTIRM and CRM, in addition to their potential in cytometry and tissue pathology, provide a platform form the non-invasive measurement of radiobiological damage as they are sensitive to the complex series of molecular responses produced in the cell. It has been demonstrated that powerful multivariate techniques can offer the means to analyse the changes in the biochemical fingerprint occurring with dose and time after irradiation as a platform for retrospective biological dosimetry. It has also been demonstrated that a suitable choice of preprocessing parameters and spectral variables can result in substantial increases in prediction performance of multivariate regression algorithms when used for biodosimetry with FTIRM and CRM spectra of irradiated cells. The study raises questions regarding the nature of the non-linearities in these changes that are suggested by the performance of the SVR algorithm in modelling the biochemical fingerprint, which will be the subject of future reports.

Acknowledgment

This research was supported by the National Biophotonics and Imaging Platform (NBIP) Ireland, funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycle 4, co-funded by the Irish Government and the European Union Structural Fund.

References

1 Bonnier, F., Knief, P., Lim, B., Meade, A.D., Dorney, J., Bhattacharya, K., Lyng, F.M., and Byrne, H.J.: 'Imaging live cells grown on a three dimensional collagen matrix using Raman microspectroscopy', Analyst, 2010, 135, (12), pp. 3169-3177

2 Bonnier, F., Meade, A.D., Merzha, S., Knief, P., Bhattacharya, K., Lyng, F.M., and Byrne, H.J.: 'Three dimensional collagen gels as a cell culture matrix for the study of live cells by Raman spectroscopy', Analyst, 2010

3 Meade, A.D., Lyng, F.M., Knief, P., and Byrne, H.J.: 'Growth substrate induced functional changes elucidated by FTIR and Raman spectroscopy in in-vitro cultured human keratinocytes', Anal Bioanal Chem, 2007, 387, (5), pp. 1717-1728

4 Nawaz, H., Bonnier, F., Knief, P., Howe, O., Lyng, F.M., Meade, A.D., and Byrne, H.J.: 'Evaluation of the potential of Raman microspectroscopy for prediction of chemotherapeutic response to cisplatin in lung adenocarcinoma', Analyst, 2010, 135, (12), pp. 3070-3076

5 Ostrowska, K.M., Malkin, A., Meade, A., O'Leary, J., Martin, C., Spillane, C., Byrne, H.J., and Lyng, F.M.: 'Investigation of the influence of high-risk human papillomavirus on the biochemical composition of cervical

cancer cells using vibrational spectroscopy', Analyst, 2010, 135, (12), pp. 3087-3093
6 Matthews, Q., Jirasek, A., Lum, J., Duan, X., and Brolo, A.G.: 'Variability in Raman spectra of single human tumor cells cultured in vitro: correlation with cell cycle and culture confluency', Appl Spectrosc, 2010, 64, (8), pp. 871-887
7 Meade, A.D., Clarke, C., Draux, F., Sockalingum, G.D., Manfait, M., Lyng, F.M., and Byrne, H.J.: 'Studies of chemical fixation effects in human cell lines using Raman microspectroscopy', Anal Bioanal Chem, 2010, 396, (5), pp. 1781-1791
8 Meade, A.D., Clarke, C., Byrne, H.J., and Lyng, F.M.: 'Fourier transform infrared microspectroscopy and multivariate methods for radiobiological dosimetry', Radiat Res, 2010, 173, (2), pp. 225-237
9 Bassan, P., Byrne, H.J., Bonnier, F., Lee, J., Dumas, P., and Gardner, P.: 'Resonant Mie scattering in infrared spectroscopy of biological materials - understanding the 'dispersion artefact'', Analyst, 2009, 134, (8), pp. 1586-1593
10 Bassan, P., Byrne, H.J., Lee, J., Bonnier, F., Clarke, C., Dumas, P., Gazi, E., Brown, M.D., Clarke, N.W., and Gardner, P.: 'Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells', Analyst, 2009, 134, (6), pp. 1171-1175
11 Bassan, P., Kohler, A., Martens, H., Lee, J., Byrne, H.J., Dumas, P., Gazi, E., Brown, M., Clarke, N., and Gardner, P.: 'Resonant Mie Scattering (RMieS) Correction of Infrared Spectra from Highly Scattering Biological Samples', Analyst, 2010, (DOI: 10.1039/b921056c)
12 Mohlenhoff, B., Romeo, M., Diem, M., and Wood, B.R.: 'Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy', Biophys J, 2005, 88, (5), pp. 3635-3640
13 Romeo, M., Mohlenhoff, B., and Diem, M.: 'Infrared micro-spectroscopy of human cells: Causes for the spectral variance of oral mucosa (buccal) cells', Vibrational Spectroscopy, 2006, 42, (1), pp. 9-14
14 Kohler, A., Sule-Suso, J., Sockalingum, G.D., Tobin, M., Bahrami, F., Yang, Y., Pijanka, J., Dumas, P., Cotte, M., van Pittius, D.G., Parkes, G., and Martens, H.: 'Estimating and correcting Mie scattering in synchrotron-based microscopic Fourier transform infrared spectra by extended multiplicative signal correction', Appl. Spectrosc., 2008, 62, (3), pp. 259-266
15 Thennadil, S.N., Martens, H., and Kohler, A.: 'Physics-based multiplicative scatter correction approaches for improving the performance of calibration models', Appl. Spectrosc., 2006, 60, (3), pp. 315-321
16 Bruun, S.W., Kohler, A., Adt, I., Sockalingum, G.D., Manfait, M., and Martens, H.: 'Correcting attenuated total reflection-Fourier transform infrared spectra for water vapor and carbon dioxide', Appl Spectrosc, 2006, 60, (9), pp. 1029-1039
17 Meade, A.D., Byrne, H.J., and Lyng, F.M.: 'Spectroscopic and chemometric approaches to radiobiological analyses', Mutat Res, 2010, 704, (1-3), pp. 108-114
18 Gaigneaux, A., Ruysschaert, J.M., and Goormaghtigh, E.: 'Cell discrimination by attenuated total reflection-Fourier transform infrared spectroscopy: the impact of preprocessing of spectra', Appl Spectrosc, 2006, 60, (9), pp. 1022-1028
19 Heraud, P., Wood, B.R., Beardall, J., and McNaughton, D.: 'Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells', Journal of Chemometrics, 2006, 20, (5), pp. 193-197
20 Afseth, N.K., Segtnan, V.H., and Wold, J.P.: 'Raman spectra of biological samples: A study of preprocessing methods', Appl Spectrosc, 2006, 60, (12), pp. 1358-1367
21 Jarvis, R.M., and Goodacre, R.: 'Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data', Bioinformatics (Oxford, England), 2005, 21, (7), pp. 860-868
22 Puppels, G.J., Colier, W., Olminkhof, J.H.F., Otto, C., Demul, F.F.M., and Greve, J.: 'DESCRIPTION AND PERFORMANCE OF A HIGHLY SENSITIVE CONFOCAL RAMAN MICROSPECTROMETER', J. Raman Spectrosc., 1991, 22, (4), pp. 217-225
23 Grubbs, F.: 'Procedures for detecting outlier observations in samples', Technometrics, 1969, 11, (1), pp. 1-21
24 Chih-Chung, C., and Chih-Jen, L.: 'LIBSVR; A library for support vector machines', Software available at http://www.csie.ntu.edu.tw/~cjlin/libSVR, 2001
25 Houck, C., Joines, J., and Kay, M.: 'Genetic Algorithm and Optimisation Toolbox', North Carolina State University, USA, v5
26 Varmuza, K., and Filzmoser, P.: 'Introduction to Multivariate Statistical Analysis in Chemometrics' (CRC Press, Taylor and Francis Group, 2009. 2009)
27 Yoshida, H., Leardi, R., Funatsu, K., and Varmuza, K.: 'Feature selection by genetic algorithms for mass spectral classifiers', Anal. Chim. Acta, 2001, 446, (1-2), pp. 485-494
28 Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L., and Goodacre, R.: 'A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data', Anal Chim Acta, 2014, 829, pp. 1-8