Analyst

Model based variable selection as a tool to highlight biological differences in Raman spectra of cells.

Journal:	*Analyst*
Manuscript ID:	AN-ART-04-2014-000731.R1
Article Type:	Paper
Date Submitted by the Author:	24-Jun-2014

Complete List of Authors:	Hedegaard, Martin; University of Southern Denmark, Department of Chemical Engineering, Biotechnology and Environmental Technology Cloyd, Kristy; Imperial College London, Department of Materials Horejs, Christine-Maria; Imperial College London, Department of Materials Stevens, Molly; Imperial College London, Department of Materials

Journal Name

RSCPublishing

ARTICLE

Model based variable selection as a tool to highlight biological differences in Raman spectra of cells.

Cite this: DOI: 10.1039/x0xx00000x

Martin A. B. Hedegaard,[a,b] Kristy L. Cloyd,[a] Christine-Maria Horejs,[a] and Molly M. Stevens[a*]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

In vitro Raman spectroscopy used for non-invasive, non-destructive characterization of single cells and tissues has proven to be a powerful tool for understanding the complex biochemical processes within these biological systems. Additionally it enables the comparison of a wide range of *in vitro* model systems by discriminating them based on their biomolecular differences. However, one persistent challenge in Raman spectroscopy has been the highly complex structure of cell and tissue spectra, which comprise signals from lipids, proteins, carbohydrates and nucleic acids, which may overlap significantly. This leads to difficulty in discerning which molecular components are responsible for the changes seen between experimental groups. To address this problem, we introduce a technique to highlight the significant biochemical changes between sample groups by applying a novel approach using Partial Least Squares – Discriminant Analysis (PLS-DA) Variable Importance Projection (VIP) scores normally used for variable selection as heat maps combined with group difference spectra to highlight significant differences in Raman band shapes and position. To illustrate this method we analyzed single HeLa cells in their live, fixed, fixed and ethanol dehydrated, to the fixed, dehydrated and then rehydrated states respectively. Fixation, ethanol dehydration and rehydration are known to induce molecular changes in the lipids and proteins within each cell.

Introduction

Raman micro-spectroscopy has been used successfully to characterize and compare a wide variety of biological samples ranging from single cells to whole tissues both *in vitro* and *in vivo*[1-4]. The technique has proven to be an ideal tool for the non-invasive measurement of lipids, proteins, carbohydrates, nucleic acids, and mineral in both living and fixed systems[5-7], allowing for characterization and comparison of biologically interesting samples without damage to the sample.

The versatile nature of Raman spectroscopy enables it to be used in a variety of configurations, ranging from fiber optical probes to microscope-based measurements, thus tailoring measurement collection to the constraints of the system of interest. Examples include *in vivo* studies using high throughput fiber probes and diffraction limited Raman imaging of cells and tissues rapidly capturing a wealth of biomolecular information rapidly. Additionally Raman spectroscopy works very well with hydrated samples furthering its compatibility with biological systems.

Previous studies have shown that multivariate analysis methods work well for the analysis of Raman spectra taken from biological systems. Examples include discriminating isogenic cancer cells[4], circulating tumor cells[5] and the *in vitro* calcification of different cell types[6], cell activity studies[7], and cell differentiation[8]. Methods used for supervised classification of biological Raman spectra include unsupervised methods such as cluster analysis and principal component analysis. Additionally a wide range of supervised methods have been applied including Linear Discriminant Analysis, Partial Least Squares – Discriminant Analysis (PLS-DA), Support Vector Machines and Neural Networks[4,9-13]. In order to fully utilize the power of these analytical techniques in biological systems, the experimental design must ensure the requirements for each technique are met and the interpretation of the results must be translatable into biological parameters.

Ultimately the analysis of the Raman spectra, whether simple or complex, strives to reveal the biochemical differences between

experimental groups of interest. Complications in both analysis and the interpretation may arise from multiple overlapping Raman bands and the difficulty in demonstrating which changes are statistically significant in distinguishing between the trial groups. Indeed overlapping Raman signals make it difficult to know whether differences between sample groups originate from changes in minerals, lipids, proteins, carbohydrates or nucleic acids, or if there are combined effects which act together to discriminate the groups of interest.

To improve this situation we suggest a new use of variable selection methods. Variable selection has seen extensive use in other fields for a variety of data types. A number of different approaches have been used ranging from model based variable importance such as Variable Importance Projection (VIP) scores and selectivity ratios over interval based methods and genetic algorithms[20]. All methods have the ability to select different regions that if used can improve the resulting model.

It has long been known that variable selection methods can be used to validate the model and help to find important regions of the spectra that contribute most to the differences.

To improve the interpretation of the resulting Raman spectra, we have selected PLS-DA Variable Importance Projection (VIP) scores to highlight the spectral changes that significantly contribute when discriminating between sample groups. Overlaying the difference spectrum between groups with a heat map defined by the VIP score clearly highlights the bands that contribute most to the model's ability to distinguish between experimental groups and thus the most statistically significant molecular changes between them. In addition the positive and negative contributions can be seen in the difference spectra.

To illustrate the ability of PLS-DA VIP score heat maps in conjunction with difference spectra to expose the critical molecular changes in biological systems, we apply this method to study single cell spectra of HeLa cells their live, fixed, fixed and ethanol dehydrated, to the fixed, dehydrated and then rehydrated states respectively.

Materials and Methods

HeLa Cell Culture

HeLa cells were cultured in DMEM Glutamax medium supplemented with 1% (v/v) antibiotic–antimycotic, 10% (v/v) fetal bovine serum (FBS) (Invitrogen, U.K.). All HeLa cells were culutured on MgF_2 glass slides (Global Optics) and seeded at 1 x104 cells/cm^2. Prior to cell seeding, MgF_2 glass slides were incubated in FBS for 6 hours. Raman spectral collection was then performed after 2 days. Cells from two independent batches were included in this study.

Fixation, Dehydration, and Rehydration

After 2 days in culture and with a maximum of 30 minutes live cell Raman imaging, HeLa cultures were fixed in 3.7% (v/v) formaldehyde (FA) for 40 minutes at 4°C and then rinsed 3x in phosphate buffered saline (PBS) and kept in PBS for subsequent Raman imaging. Cells were then dehydrated in a graded ethanol from 50%, 70%, 90% and 100% (v/v) series and Raman spectra were collected on the dried cells. Lastly cells were re-hydrated by submerging MgF_2 substrate with dried and fixed HeLa cells in PBS for 30 minutes before Raman imaging.

Raman Spectroscopy

Raman spectra were measured with a Renishaw InVia (Renishaw, Wotton-under-Edge, U.K.) spectrometer connected to a Leica (Wetzlar, Germany) microscope. The spectrometer uses a high power 785 nm diode spot laser (~30 mW at sample; Renishaw) for excitation. The laser was focused on individual cells by a 60× (NA = 1.0) long working distance (2 mm) water immersion objective (Nikon) or a 100× (NA 0.9) objective (Leica) resulting in a spot size of approx. 5 µm. Spectra of living HeLa cells were measured in PBS (Invitrogen) maintained at 37°C with a heated stage. Spectra of fixed and rehydrated cells were collected in PBS at room temperature and dehydrated cells were measured dry on the MgF_2 slide. For each cell a single spectrum was taken. The volume covered by laser spot (approx. 5x5x15µm) should sufficiently cover the whole cell thickness in all cases. All spectra were taken individually and care was taken to cover equivalent areas of each cell.

For all samples spectra of the fingerprint region (620–1720 cm^{-1}) were recorded at a resolution of ~1–2 cm^{-1}, with 3 accumulations of spectra each with a 5 second integration time. Backscattered radiation was collected by the same objective then passed through a 785 nm edge filter to block Rayleigh scattering and reflected laser light, before being directed through a 50 µm slit into the spectrometer equipped with a 1200 lines/mm grating, and finally detected by a deep-depletion charge-coupled device detector.

A total of 2080 spectra were collected from individual cells consisting of 384 from live cells, 714 from fixed cells, 338 from dehydrated cells and 644 from dehydrated cells.

Data Analysis

All data analysis was performed in MatLab R2013a (The Mathworks, Natick, MA, U.S.A.) with in house written scripts in combination with the PLS_toolbox 7.0 (Eigenvector Research, Wenatchee, WA, U.S.A.).

Spectra for the cell study were background corrected by subtracting the spectrum of PBS and substrate for the hydrated samples and the background of the substrate from the dehydrated samples. All spectra were normalized using extended multiplicative signal correction (EMSC)[14,15] using mean spectrum as reference and smoothed using a Savitzky-Golay filter (five points, second-order polynomial)[16]. The EMSC correction also assures that the remaining background signal in the spectra does not interfere with the analysis as they are corrected towards the same mean.

For classification we applied Partial Least Squares – Discriminant Analysis (PLS-DA). PLS-DA uses the properties of normal PLS regression to rotate a principal component analysis (PCA) model to best explain the differences between groups. Using the model to predict a matrix of zeros and ones depending on the experimental group allows for a rotation of a PCA model so it best discriminates based on group variances.

To highlight the variables, which contribute the most to the discrimination, we calculate the Variable Importance Projection (VIP) score. The VIP score works as a summary of the importance of the projection when finding the latent variables[17-20].

The VIP variable for the j^{th} variable can be expressed as

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}}$$

where w_{jf} is the weight value for variable j component f, SSY_f is the sum of squares of explained variance for the f^{th} component and J the number of variables. SSY_{total} is the total sum of squares explained of the dependent variable, and F is the total number of components. The inclusion of the weights of the PLS-DA model enable the VIP score to describe both how well the dependent variable is explained and also how important the variable is for the modeling of independent variables. This is due to the fact that the weights of a PLS-DA model reflect the covariance between the independent and dependent variables. Traditionally a VIP score lower than one denotes a non-important variable.

One PLS-DA model and corresponding VIP scores were calculated for each of the following four cases: live vs. fixed HeLa cells, fixed vs. dehydrated HeLa cells, and dehydrated vs. rehydrated HeLa cells. Each model was cross validated using leave out random subsets with 10 data splits and 20 iterations. The number of components was selected using RMSECV finding the first bend in the RMSECV curve. For each model the average and difference spectra were calculated to analyse the spectral differences. The difference spectra were overlaid with a heat map of the VIP score for each model. It should be noted that the selectivity ratio can be used as well and is defined as the ratio between the explained variance of each variable and the residual variance. Similar to VIP scores a high value indicates variables with a good predictive performance. It should be noted that the VIP scores do not indicate the sign of the variation, only the position. To obtain the sign of the variation it has to be combined with the difference spectra.

Results

Cultured HeLa cells were tested after 2 days of culture on the MgF_2 substrates and had the normal appearance of HeLa cells. Single HeLa cells were selected for Raman imaging, as the HeLa cells were not confluent at the time of testing. Figure 1 shows the mean Raman spectra of a) live cells, b) fixed cells, c) dehydrated cells and d) rehydrated cells. All spectra contain bands expected from single cells including RNA/DNA (785, 811, 1320, 1337, 1375 and 1575cm^{-1}) Amino Acids (Phe: 622, 1002, 1032, 1602 cm^{-1}, Trp: 760 1554 cm^{-1}, Tyr 645, 828cm^{-1}), lipids (877, 1095, 1124cm^{-1}) and the Amide III at 1257-1300 cm^{-1} and Amide I at 1655 cm^{-1}.

There are significant visual differences in the average spectrum in the amino acid Phe band at 1002 cm^{-1} for the dehydrated cells in Figure 1c and some changes in the shape of the Amide I band when compared to the three other experimental groups. A figure showing the average spectra and their standard derivation is included in supplementary material (Figure s1).

A PLS-DA model was employed to compare live vs. fixed, fixed vs. dehydrated and dehydrated vs. rehydrated HeLa cells and the resulting model details are shown in Table 1. The resulting classification plots for each can be found in supplementary material (Figure s2). For each model the VIP scores were calculated and plotted as a heat map on the background of the difference spectra shown in Figure 2 showing the average difference spectra between a) live-fixed, b) fixed-dehydrated and c) dehydrated-rehydrated overlaid with the VIP score heat map. The more intense the green band, the more significant that Raman spectral band was in the PLS-DA model's ability to distinguish between the groups.

Model	Components	Sensitivity	Selectivity	RMSECV
Live vs. Fixed	3	0.914	0.962	0.2597
Fixed vs. Dehydrated	2	1.00	1.00	0.0989
Dehydrated vs. Rehydrated	2	1.00	1.00	0.1271

Table 1: PLS-DA model results for the live vs. fixed, fixed vs. dehydrated and dehydrated vs. rehydrated sample groups.

Features highlighted in Figure 2a between live and fixed cells are mostly related to phospholipids (720, 1092cm^{-1}) and lipids (877, 935, 983 and 1370cm^{-1}). It is important to note that the most statistically significant changes highlighted by the VIP scores are not necessarily the largest variations in the difference spectrum.

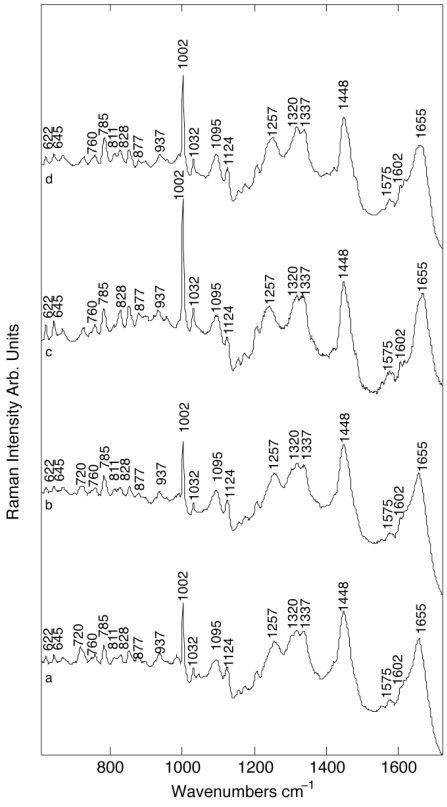

Figure 1: Average spectra of HeLa cells in a) live, b) fixed, c) ethanol dehydrated, and d) rehydrated form. Visual changes can be seen in the Amide I band at 1655cm^{-1} and the intensity of the Phe band at 1002 cm^{-1}.

Differences between fixed and ethanol dehydrated cells are also distinctly visible in Figure 2b with the most significant differences being related to unfolding of proteins, visible by the large change in band shape in the Amide I band at 1630-1696cm^{-1} and the exposure of phenylalanine due to break up in α-helixes and β-sheets resulting in higher intensities of the 622, 1002 and 1032cm^{-1} bands. In addition there are visible changes in the DNA related vibration at 780 cm^{-1} and a decrease in lipid content 877, 1227 and 1430cm^{-1}. Rehydrating the cells in PBS refolds the proteins to a certain degree reversing these changes as seen in Figure 2c, however this reversal does not return proteins to their native state and leaves visible changes in the live and fixed cell spectra.

Discussion

Unveiling the biological significance behind changes in information rich Raman spectra collected from biological samples can be challenging due to the complex nature of the molecular bonds identified by Raman spectroscopy. This translation however, is critical in applying this powerful non-invasive tool to biological investigations. Raman spectra collected from biological samples include signals from lipids, proteins, carbohydrates and minerals and details of their environmental influence on each other.

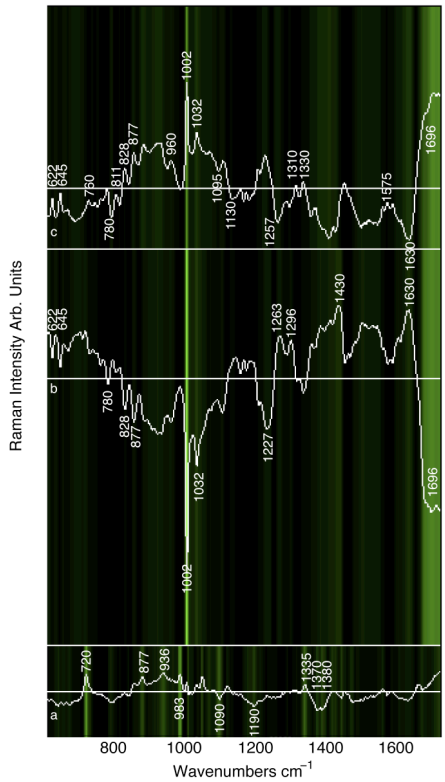

Figure 2: Difference spectra (white) overlaid by VIP scores (green heat map) for average spectra of a) live subtracted by fixed, b) fixed subtracted by dehydrated and c) dehydrated subtracted by rehydrated cell spectra. The brightness of the green background indicates the value of the VIP scores with the greater the colour intensity corresponding to a higher VIP score and thereby a greater significance of those bands to the PLS-DA models as shown in Table 1.

In this paper we have introduced a novel way of using PLS-DA VIP scores to visually highlight the significant differences between biological Raman datasets. The usefulness of this analysis is presented using a system comparing cultured HeLa single cells in their live, fixed, fixed and ethanol dehydrated and fixed, dehydrated and rehydrated forms respectively. It should also be noted that the VIP scores are not used for variable selection to reduce the dataset, but rather to understand what variables contribute to the current model.

The PLS-DA VIP scores highlighted Raman spectral features, which distinguished between HeLa cells depending on their experimental state. When comparing live HeLa cells to fixed HeLa cells, changes in spectral bands corresponding to differences in lipid composition due to the crosslinking of phospholipids were observed however no significant change in protein structure was resolved by Raman spectroscopy. It is well reported in previous studies on formaldehyde fixation that proteins are generally well preserved in the fixation process with lipids may be altered or lost[21,22]. In the cases of fixed HeLa cells vs. dehydrated HeLa cells and dehydrated HeLa cells vs. rehydrated HeLa cells, the main changes in the

molecular structure are due to the unfolding and refolding of proteins resulting in exposure of phenylalanine and breaking and reformation of β-sheets and α-helixes. In all cases the absolute intensity of the difference spectra is not necessarily indicative of the importance of those bands as a distinguishing factor between groups. The VIP scores highlight the statistically important differences quickly and clearly identify specific bands and peaks of significance.

It should be noted that the VIP score is always positive. That also means that the sign of the variation is not visible in the VIP score, but combined with the difference spectra it is clearly visible if the variation is positive or negative between two groups. The VIP scores also have a limitation in this approach in that they do not show the sign of the variation when more than two classes are included in the model. To address this one could use the regression vector in a similar way, showing negative and positive values of the regression vector in different colours, and thereby using the regression vector in a similar way to Beleites et. al.[23] We also want to highlight that this approach is not limited to PLS-DA, but in principle could be used for similar classification models.

Conclusions

In conclusion we have applied PLS-DA VIP scores to highlight significant chemical differences in biological spectra comparing single cells tested in their live, fixed, fixed and dehydrated versus fixed, dehydrated and rehydrated forms respectively. In the study, applying PLS-DA VIP scores could be used as an important tool for highlighting the biomolecular bonds, which distinguished one sample group from another. This visualisation thus helped in translating the changes observed to the biological differences between the sample groups. We believe this new analytical tool will find great applications across a wide breath of future Raman spectroscopic studies of biosystems.

Funding

Corresponding Author

*Molly M. Stevens, m.stevens@imperial.ac.uk

Notes and references

[a] Departments of Materials and Bioengineering and the Institute of Biomedical Engineering, Imperial College London, SW7 2AZ, UK. E-mail: m.stevens@imperial.ac.uk; Tel: +44 (0) 20 7584 6804
[b] Department of Chemical Engineering, Biotechnology and Environmental Technology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. E-mail: marhe@kbm.sdu.dk; Tel: +45 65507413

†Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

1. C. Krafft, B. Dietzek, J. Popp. *Analyst*, 2009,**134**, 1046-1057
2. C. Krafft, B. Belay, N. Bergner, B. F. M. Romeike,R. Reichart, R. Kalff, J. Popp. *Analyst*, 2012, DOI: 10.1039/c2an36083g
3. M. S. Bergholt, W. Zheng, Z. Huang, *J. Raman Spectrosc.* 2012,**43**, 255–262
4. M. Hedegaard, C. Krafft, H. J. Ditzel, L. E. Johansen, S. Hassing, J. Popp. *Anal. Chem*, 2010, **82** (7), 2797–2802
5. U. Neugebauer, T. Bocklitz, J. H. Clement, C. Krafft, J. Popp. *Analyst*, 2010, **135**, 3178-3182
6. E. Gentleman, R. J. Swain, N. D. Evans, S. Boonrungsiman, G. Jell, M. D. Ball, T. A. V. Shean, M. L. Oyen, A. Porter, M. M. Stevens. *Nat. Mater*. 2009, **8**, 763-770
7. R. J. Swain, S. J. Kemp, P. Goldstraw, T. D. Tetley, M. M. Stevens. *Biophys J*. 2008 **95**(12): 5978-5987
8. R. J. Swain, S. J. Kemp, P. Goldstraw, T. D. Tetley, M. M. Stevens. *Biophys J*. 2010 **98**(8): 1703–1711
9. T. Boklitz, M. Putsche, C. Stüber, J. Käs, A. Niendorf, P. Rösch, J. Popp. *J. Raman Spectrosc*. 2009, **40**(12), 1759-1765
10. M. Sattlecker, R. Baker, N. Stone, C. Bessant. *Chemometr. Intell. Lab*. 2011, **107**(2), 363-370
11. M. Sattlecker, N. Stone, J. Smith, C. Bessant. *J. Raman Spectrosc.* 2011, **42**(5), 897-903
12. C. Krafft, R. Salzer, S. Seitz, C. Ern, M. Schieker. *Analyst*, 2007, **132**, 647–653
13. P. Lasch, W. Haensch, D. Naumann, M. Diem. *BBA-Mol. Basis Dis*. 2004, **1688**(2) 176-186
14. H. Martens, J. Pram Nielsen, S. Balling Engelsen. *Anal. Chem*. 2003; **75** (3), 394–404
15. A. Kohler, C. Kirschner, A. Oust, H. Martens. *Appl. Spectrosc*. 2005, **59**, 707– 716
16. A. Savitzky, M. J. E Golay. *Anal. Chem*. 1964, **36**(8) 1627-1639
17. I. Chong, C. Jun. *Chemometr. Intell. Lab*., 2005, **78**, 103-112
18. O. M. Kvalheim. *J. Chemometrics*, 2009, **24**, 496-504
19. T. Rajalahti, R. Arneberg, F. S. Berven, K. Myhr, R. J. Ulvik, O. M. Kvalheim. *Chemometr. Intell. Lab*. 2009, **95**, 35-48
20 C. M. Andersen, R. Bro, *J. Chemometrics*, **24**, 728-737
21. A. D. Meade, C. Clarke, F. Draux, G. D. Sockalingum, M. Manfait, F.M. Lyng, H. J. Byrne. *Anal Bioanal Chem*. 2010, **396**, 1781-1791
22. M. M. Mariani, P. Lampen, J. Popp, B. R. Wood, V. Deckert. *Analyst*, 2009, **134**, 1154-1161
23. C. Beleites, K Geiger, M. Kirsch, S. B. Sobottka, G. Schackert, R. Salzer, *Anal Bioanal Chem*, 2011, **400**:2801-2816