Quantification of casein phosphorylation with conformational interpretation using Raman spectroscopy

Roger M. Jarvis,*[ab] Ewan W. Blanch,[ac] Alexander P. Golovanov,[ac] James Screen[d] and Royston Goodacre[ab]

Received 27th February 2007, Accepted 16th July 2007
First published as an Advance Article on the web 7th August 2007
DOI: 10.1039/b702944f

Raman spectroscopy is emerging as a powerful method for obtaining both quantitative and qualitative information from biological samples. One very interesting area of research, for which the technique has rarely been used, is the detection, quantification and structural analysis of post-translational modifications (PTMs) on proteins. Since Raman spectra can be used to address both of these questions simultaneously, we have developed near infrared Raman spectroscopy with appropriate chemometric approaches (partial least squares regression) to quantify low concentration (4 μM) mixtures of phosphorylated and dephosphorylated bovine α_s-casein. In addition, we have used these data in conjunction with Raman optical activity (ROA) spectra and NMR to assess the structural changes that occur upon phosphorylation.

Introduction

One of the main challenges in post-genomic biochemistry for proteomics lies in the study of post-translational modifications (PTMs):[1–3] these biochemical processes are important for cell regulation, and can often be driven by very subtle functional substitutions. Many different categories of PTM (300 or more) have been defined and are involved in numerous physiological activities within the cell. One very important PTM is phosphorylation, which is associated with signal transduction networks. The phosphorylation by a kinase or dephosphorylation by a phosphatase of a protein is of great importance in modifying and controlling its biological function, through conformational change. However, whilst the DNA gene sequence can be translated into the amino acid sequence of the protein, any PTMs must be discovered by analysing the protein directly, and this has classically been achieved by mass spectrometry (MS[4,5]). For an excellent review on recent analytical approaches in this field see ref. 6.

In addition to the detection and sequencing of PTM species it is desirable to propose a structure for the molecule. For proteins with highly ordered secondary or tertiary structure, conformational information can be obtained by one or a combination of X-ray crystallography,[7–9] nuclear magnetic resonance (NMR[10–12]) spectroscopy and circular dichroism (CD[13,14]). However, these methods do not appear to be incisive probes of proteins with little ordered secondary structure, due to the conformational flexibility of such molecules.

Various forms of MS are used very effectively to identify phosphorylation sites, in general by measuring the mass difference between phosphorylated and dephosphorylated forms, but obtaining quantitative information is often difficult.[1,15–17] However, a vibrational spectroscopic approach may provide opportunities for accurate quantitative analysis, because the area of a band due to a functional group is directly proportional to the concentration of that group. In addition, Raman spectroscopy may provide an advantage over infrared spectroscopy in this application due to its insensitivity to water.

In his 1999 review,[18] Carey discussed the potential for Raman spectroscopy to be utilized in structural biological studies. Indeed, over the past few years, many structural studies have been performed with conventional Raman spectroscopy[19–22] as well as more exotic methods, such as deep ultraviolet resonance Raman spectroscopy[23–28] and Raman optical activity (ROA[29,30]). The latter, also known as chiral Raman,[31,32] is particularly well suited to the study of proteins with little secondary structure, as has been demonstrated to great effect in refs. 33–38.

This article describes the development of Raman spectroscopy with conventional and chemometric analyses to quantify the phosphorylation of a protein in low concentration solutions. In addition, using Raman spectral analysis, conformational changes that occur upon phosphorylation are discussed. In this study, the α_s-casein protein isolated from cow's milk was used as a model post-translationally modified protein. Caseins are low molecular weight ($\sim$25 kDa) proteins whose biological function is thought to be to deliver calcium; for a recent review on casein structure and function see ref. 39. Naturally occurring bovine α_s-casein can exhibit a number of subtle sequence differences, but is generally phosphorylated on 8 or 9 serine amino acid sites[40] and has little secondary order structure. The conformation of this particular protein has previously been analysed in great detail,[41–44] although the dephosphorylated form has been ignored thus far. Therefore, it is not the intention of this article to repeat this very thorough structural work, but to demonstrate emerging analytical technologies that can be used to obtain rapidly,

[a]*Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, UK M1 7ND. E-mail: roger.jarvis@manchester.ac.uk; Fax: +44 (0)161 306 4480; Tel: +44 (0)161 306 4414*
[b]*School of Chemistry, The University of Manchester, PO Box 88,, Sackville Street, Manchester, UK M60 1QD*
[c]*Faculty of Life Sciences, The University of Manchester, The Mill, PO Box 88, Manchester, UK M60 1QD*
[d]*Renishaw plc, Old Town, Wotton-under-Edge, Gloucestershire, UK GL12 7DW*

detailed structural information on post-translationally modified molecules.

Experimental procedures

Materials

For quantification, phosphorylated and dephosphorylated α_s-casein (Sigma, UK) was dissolved in distilled water to prepare three independent 200 μM solutions of each protein. From these three sets, 21 solutions were prepared by mixing the phosphorylated and dephosphorylated forms in different ratios. Starting from a 100% phosphorylated solution, the percentage of the phosphorylated protein was decreased in each successive solution by 5%, down to a 0% phosphorylated solution. A 5 μL aliquot of each solution was then spotted on to a hydrophobic SpectRIM™ (Tienta Sciences Inc., Indianapolis, IN, USA) slide using a micropipette and allowed to dry at room temperature for approximately 2 h; each slide was capable of holding up to 25 samples. The total amount of deposited protein in each spot was approximately 1 nmol, or 25 μg. For the subsequent structural studies, 50 mg ml^{-1} solutions of phosphorylated and dephosphorylated α_s-casein were prepared from the same stock of lyophilized protein.

Near infrared (NIR) Raman spectroscopy

NIR Raman spectra were collected using a Renishaw RM2000 Raman microscope (Renishaw plc., Wotton-under-Edge, Gloucestershire, UK[45,46]), with a low power 785 nm diode laser delivering ~8 mW of power to the sample at a spectral resolution of ~6 cm^{-1}. Instrument control and data capture was performed using the GRAMS WiRE software package (Galactic Industries Corporation, Salem, NH, USA) running under Windows 95. Raman spectra were collected over a wavenumber shift range of 400–1800 cm^{-1}, and each sample was represented by a spectrum containing 1492 points relating to the number of Raman scattered photons (see Fig. 1A for examples). Four replicate Raman spectra were obtained from each sample, giving a total of 252 spectra; the total individual spectral integration time was 3 min.

ASCII data were exported from the GRAMS WiRE software into MATLAB® (The Mathworks, USA). In order to remove unavoidable baseline shifts, the spectra were optimised prior to calibration modelling by background correcting the signal with extended multiplicative signal correction (EMSC[47]) using a polynomial of order 5, following which a linear mean filter with a 5 bin window (30 cm^{-1}) was applied. For direct comparison of NIR Raman spectra of the phosphorylated and dephosphorylated protein, the twelve spectral replicates were simply averaged and no further corrections were applied.

Raman optical activity (ROA)

ROA spectra were collected using a ChiralRAMAN spectrometer (BioTools Inc, USA). The instrument operates with a backscattering geometry, a Nd:YVO4 laser at 532 nm and utilises the scattered circular polarization (SCP) strategy developed by Hug.[48,49] The incident laser power was 600 mW at the sample, 20 h of data were collected for phosphorylated α_s-casein and 17 hours of data were collected for dephosphorylated α_s-casein. The ROA spectra are presented as circular intensity differences $I^R - I^L$, where I^R and I^L are the Raman scattered intensities in right and left-circularly polarized incident light, respectively. These spectra were scaled relative to the band at 1440 cm^{-1} in the Raman spectra obtained in parallel with the ROA data.

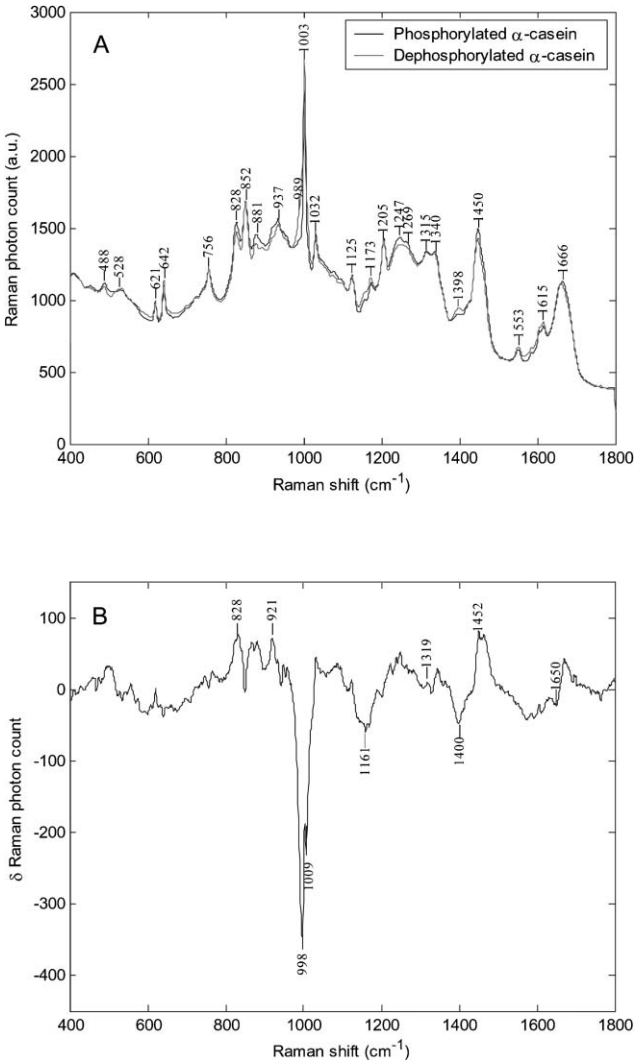

Fig. 1 (A) Mean spectra of twelve Raman spectra acquired independently from dephosphorylated and phosphorylated α_s-casein prepared on a hydrophobic slide. (B) The difference spectrum of the row normalised spectra of phosphorylated α_s-casein minus dephosphorylated α_s-casein.

Nuclear magnetic resonance (NMR)

The NMR spectra were obtained at 200 °C on a Bruker DRX 600 MHz spectrometer equipped with the Cryoprobe (Bruker Biospin, DE). Using WATERGATE (Bruker Biospin, DE) for water signal suppression, 1D spectra of phosphorylated and de-phosphorylated α_s-casein were collected using 32 scans in identical conditions, with an acquisition time for each

scan of 0.114 s. The 2024 complex data points collected in the time domain were multiplied by a QSINE (squared sine multiplication) window function (SSB (sinebell shift) = 2.8) and zero-filled to 16 384 points before Fourier transformation. 2D NOSEY (nuclear Overhauser enhancement) spectra (2048 by 150 complex data points) were collected with 16 scans and a mixing time of 0.15 s. Data in the time domain were multiplied by QSINE window functions with SSB = 3 and SSB = 2.5, in F2 (x-axis) and F1 (y-axis) dimensions respectively, and zero-filled to 1024 points in the F1 dimension before Fourier transformation.

Multivariate calibration

To generate a model predicting the percentage concentration of phosphorylated α_s-casein based on the Raman spectra of casein mixtures, the supervised multivariate method of partial least squares regression (PLSR[50,51]) was used, as described previously.[52] PLSR is a linear fixed regressor model of the form

$$Y = AX + B$$

which effectively performs a series of projections and rotations on the spectral data, to generate new axes represented by latent variables, on which a multivariate linear regression can be performed. The model was constructed using full cross-validation,[53,54] whereby PLSR was trained on a subset of data representing points on the calibration curve. The number of latent variables to be used were then optimized by projection of a second set of hold-out data, and finally the model was tested on a third set of independent data.

Feature subset selection

In order to determine the spectral bands most important for the quantification of the phosphorylated form of α_s-casein, a directed search using a genetic algorithm (GA[55,56]) coupled to PLSR was used, as described previously.[57]

A GA is an evolutionary computing technique that can be used to solve problems efficiently for which there are many possible 'good' solutions.[55,56,58,59] This class of algorithm is described as heuristic, which means that it searches for a range of possibilities, not necessarily finding the best solution but a range of sub-optimal (so called 'good') solutions. With GAs, the initial step is to generate a random population (array), consisting of a predefined number of individuals (rows) and variables (columns). Each individual represents a subset of the original variables within the larger superset of data under analysis. The next step in the GA is analogous to the process of Darwinian evolution, whereby through the processes of crossover, mutation and survival of the fittest, individuals are selected for the next generation until a particular stopping criterion has been reached. The GA uses an algorithm called a fitness function to assess the robustness of the solution proposed by each individual. This can take the form of a minimization function; therefore the fittest individuals are those with the lowest fitness value. The GA-PLSR algorithm used in this study takes the root mean squared error of prediction (RMSEP) for the cross-validation data as a measure of fitness. The RMSEP is given by

$$\mathrm{RMSEP} = \sqrt{\sum_{i=1}^{n} \left(y_{\mathrm{act}} - y_{\mathrm{pred}}\right)^2 \Big/ n}$$

where y_{act} are the actual dependent variables, y_{pred} are the predicted dependent variables and n is the number of objects.[60]

For this GA-PLSR experiment, the optimal starting conditions were initially determined by exploratory analysis, as described in ref. 54. Therefore, 250 independent GA runs were performed. For each run a population size of 500 individuals with 4 variables was used, with probabilities of mutation and crossover set to 0.4 and 0.8, respectively. Finally, the GA was programmed to terminate after the optimal solution was repeated for 40 subsequent runs.

Results and discussion

Quantification of α_s-casein

For this study hydrophobic slides were used for sample presentation. These produce a pre-concentration effect, since the hydrophobic coating on the slide causes aqueous samples to bead up and dry into a "coffee ring" pattern, concentrating the analyte at the edge of the ring. Spectra collected from this edge are therefore more intense than those from equivalent solutions on steel, glass or quartz.[61] From this study, as was seen previously,[62] it was estimated that the Raman scattering intensity for α_s-casein on a hydrophobic slide is about 5–10 times that of the protein on a stainless steel slide with a limit of detection below 4 μM.

The NIR Raman spectra obtained across the range of mixtures from deposits of phosphorylated and dephosphorylated α_s-casein are quite similar to the naked eye (Fig. 1A). Amongst the main features of the spectra are the band at 1003 cm^{-1}, which is due to phenylalanine, and the broad regions between 1220 cm^{-1} to 1300 cm^{-1}, 1420 cm^{-1} to 1480 cm^{-1} and 1640 cm^{-1} to 1700 cm^{-1}, assigned to amide III, C–H deformations and amide I vibrations, respectively,[63] although in Fig. 1B the difference spectrum shows clearly that there are many quantitative differences between the two spectral profiles. Inferences that can be made from these data regarding protein conformational differences are discussed in more detail below.

The Raman band assignments relating to the vibrational modes of phosphate in both its hydrated and dehydrated forms are given in Table 1. In Fig. 1 it is clear that there is no evidence in the Raman spectra of the strong ν_1 phosphate band

Table 1 List of typical band positions for phosphate related vibrations

Raman shift/cm^{-1}	Assignment
960	PO_4 (ν_1)
435	PO_4 (ν_2)
1075	PO_4 (ν_3)
590	PO_4 (ν_4)
1002	HPO_4 (ν_1)
875	H_2PO_4 (ν_2)

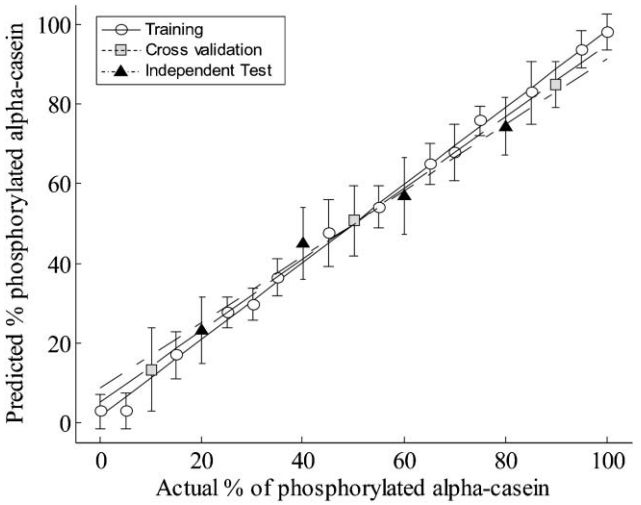

Fig. 2 Partial least squares predictions for the percentage of phosphorylated α_s-casein in an equimolar mixture (4 μM) of both the phosphorylated and non-phosphorylated forms. Mean prediction for each concentration are plotted with standard error bars.

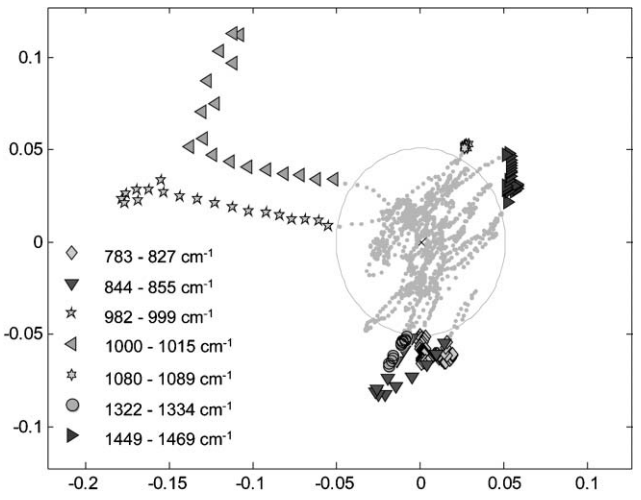

Fig. 3 A plot of the first two dimensions of the latent variable loadings. For clarity the centroid is plotted (marked by a cross), with an ellipse representing the 95% confidence interval: any points outside of this boundary are plotted explicitly and encoded with symbols according to the spectral region that they represent. The value of this depiction is that it can indicate regions of importance to the model.

at 960 cm^{-1}, which is resonance enhanced at 785 nm excitation, or indeed any of the weaker protonated bands. This phenomenon has been noted before,[62] and is most likely due to vibrational constraints engendered by hydrogen bonding interactions in this proton rich environment.[64] This is a problem that must be overcome if Raman spectroscopy is also to be developed for the detection of phosphopeptides, an analytical role that is currently performed using mass spectrometry.

The PLSR predictions for casein concentration are plotted in Fig. 2: note these are the mean of the predictions for each concentration with standard error bars. Clearly, from this result and from examination of the RMS error values in Table 2, it has been shown that Raman spectroscopy can provide an accurate means of determining the relative concentrations of a mixture of these two proteins. One example of where this approach could be applicable to a real biological problem is in quantitative modelling of the over- and under-production of specific proteins or peptides in mutant biological strains. However, this would likely depend upon the quantity of material recovered from the cell, and certainly involve some sample clean-up or development of a methodolgy to deal with interfering additives introduced through this process.

Decomposition of the spectral data during the application of PLSR yields a matrix of spectral loadings, which give an indication of the relative importance of the spectral variables for derivation of each latent variable.[50] The spectral loadings for the first two PLSR latent variables, which account for the majority of the variance in the data, are plotted against one another in Fig. 3. This indicates that the region around 1000 cm^{-1} is particularly important for this calibration model (see Table 4). By comparing the Raman spectra of the phosphorylated and dephosphorylated forms of α_s-casein (Fig. 1A), it is apparent that the strong band in this region is much broader for dephosphorylated casein. This could indicate that tryptophan side chains are promoted when phosphate is removed, clearly indicated by the prominent difference band at 1009 cm^{-1} in Fig. 1B. Another possibility is that an underlying quantitative relationship to the HPO_4 (ν_1) vibration exists under this band. Regardless of the biochemical meaning behind this observation, what it also suggests is that this band is a likely candidate to be used to formulate a calibration model based on the Raman scattering cross-section. However, integrating the area under this peak does not result in a correlation to concentration (data not shown) and has an R^2 value of 0.05. Additional spectral regions also appear outside of the 95% confidence limit, and may therefore be considered to represent important spectral information. These include the regions around 828, 989 and 1450 cm^{-1} that were also singled out by the conventional analysis discussed below and summarized in Table 4.

Feature subset selection

Interpretation of spectral loadings plots should in theory reveal the underlying variables that are most important to the model. Unfortunately, as was demonstrated above, for continuous spectral data of the type reported here loadings plots can be extremely difficult to interpret, because often the candidate short-list of potentially important 'peaks' can extend

Table 2 Percentage root mean square error values for the PLSR calibration model

Calibration error (%)	Cross-validation error (%)	Independent test error (%)	Number of PLS factors (latent variables) used
9.9	9.1	10.9	6

Table 3 Principal differences observed for the NIR Raman spectral profile of phosphorylated α_s-casein *versus* dephosphorylated α_s-casein, with tentative assignments to protein structure

Raman band/cm^{-1}	Phosphorylated α_s-casein	Dephosphorylated α_s-casein
850/830	Increase in Fermi doublet ratio from tyrosine residues	Decrease in Fermi doublet ratio from tyrosine residues
989	Decreased tyrosine signature	Increased tyrosine signature
1003	Decreased phenylalanine and increased tryptophan signatures, potential increase in HPO_4 (ν_1)	Increased phenylalanine and decreased tryptophan signatures, potential decrease in HPO_4 (ν_1)
1250 & 1265	Increase indicating more ordered structure	Decrease indicating less ordered structure
1400	Decrease in band intensity from histidine side chain conformation	Increase in band intensity from histidine side chain conformation
1655–1670	Increase in band maximum	Decrease in band maximum

Table 4 The spectral regions most frequently selected by GA-PLS, listed in order of the density of variables selected within those regions

Raman shift/cm^{-1}	Assignment	Phosphorylated α_s-casein	Dephosphorylated α_s-casein
ca. 1398	Histidine side chain conformation	Decrease	Increase
ca. 1577	W2 vibrational mode of tryptophan side chains	Decrease	Increase
ca. 1490	Relates to rate of deceleration of the band at *ca.* 1450 cm^{-1}	Increase	Decrease
ca. 1650	Shifts in the amide I maxima	Increase	Decrease

into the hundreds.[65,66] In addition, it is not sufficient, with a multivariate problem involving continuous data, to assume that single spectral bands or regions can be used to define the optimal solution, especially since relevant chemical information may be spread across the whole of the spectral space. Rather, it is necessary to explore combinations of variables that best describe the solution, and this cannot best be achieved through the interpretation of loadings.[67]

To address this issue much attention has been focused on using data mining methodologies to discover subsets of variables that can formulate good solutions to problems such as PLSR calibration.[54,57,68] The benefit of this method is that by reducing the solution from a complex multivariate problem down to only a handful of independent variables, one is able to make deductions about the biochemistry which defines the problem. One method in particular for feature selection that has attracted widespread attention is genetic algorithm (GA) optimization, which can very easily be coupled to PLSR and other multivariate models.[69,70]

Therefore, we applied GA-PLSR to the problem of discovering which bands in the Raman spectra of casein were most important for quantification. The benefit of this approach is that the spectral bands utilized with greatest frequency by the GA are very likely to relate to structural differences between the phosphorylated and dephosphorylated forms of the protein being studied.

Fig. 4 shows a scatter plot representing the Raman bands selected in the fittest individuals from each of the 250 independent GA runs. This plot indicates which regions of the spectrum were selected most frequently by the GA, and therefore contributed most to the calibration model. Clearly, only a small subset of spectral regions can be used to generate optimal PLSR results, and these are elucidated in Table 3. This strongly indicates that the major conformational changes upon dephosphorylation result in greater constraint of histidine side chains (*ca.* 1398 cm^{-1}), a change in the W2 vibrational mode of tryptophan side chains (*ca.* 1577 cm^{-1})[71] and finally a change in the α-helical structure of the molecule (*ca.* 1650 cm^{-1}). The region around 1490 cm^{-1} offers no obvious

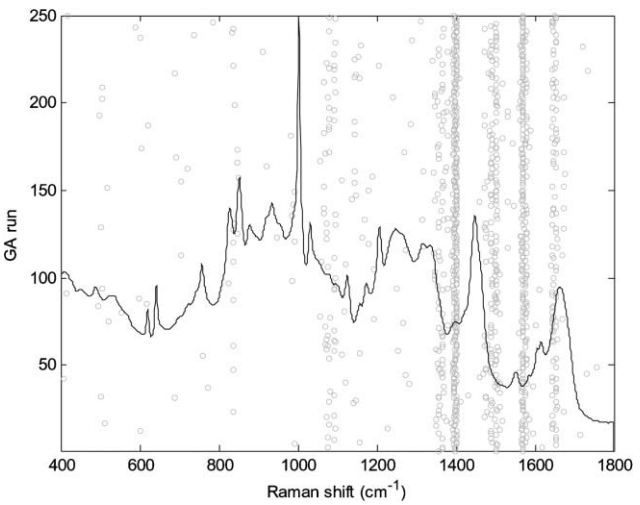

Fig. 4 A scatter plot of Raman shift *versus* the most optimal result from each independent GA run. The density of points about certain regions of the spectrum indicates bands that are important to the quantification of the different casein forms. The average spectrum of phosphorylated α-casein is overlaid for clarity.

structural assignment and needs to be investigated further in order to deduce any structural observations. In addition, it should be noted that this particular analysis highlighted two spectral regions of importance (*ca.* 1490 and 1577 cm^{-1}) that were missed by the more conventional spectral analysis. Further structural inferences that can be made from this information are expanded upon below.

Comparison of phosphorylated and dephosphorylated α_s-casein structure by Raman spectroscopy

Although the targetted analysis afforded by GA feature selection points to the most pronounced structural changes that occur during the dephosphorylation process, additional information can be mined by a more traditional manual interpretation of the spectra. Therefore, in Table 4

observations taken from the NIR Raman spectra in Fig. 1A and their difference spectra (Fig. 1B) are listed, with interpretations of the structural differences that these indicate.

The tyrosine Fermi doublet ratio at 850/830 can provide an important indicator of the average hydrogen bonding state of the tyrosines with respect to this protein.[72] These bands are perturbed by the local environment of the phenol ring in tyrosine and the intensity ratio of 850/830 is sensitive to hydrophobicity. The ratio ranges from 0.3, when the phenolic OH is a strong H-bond donor, to 2.5 when it is a strong H-bond acceptor. For these data, this band indicates that upon dephosphorylation the ratio 850/830 decreases (Fig. 1B) and that one or more of the tyrosines becomes a stronger H-bond donor/weaker H-bond acceptor, an observation which corresponds to the removal of phosphate groups. The tyrosine band at 989 cm^{-1} is promoted in dephosphorylated casein, which corresponds to the intensity increase in the 830 cm^{-1} band of the Fermi doublet noted above.

Also apparent is that upon dephosphorylation conformational changes are occurring that result in a large decrease ($\sim$6 cm^{-1}) in the position of the amide I maximum (1655–1670 cm^{-1}). It is not clear as to the identity of these conformational changes, due to a lack of detailed band assignments in the literature. In addition, the major bands at 1250 and 1265 cm^{-1} in the phosphorylated case indicate more ordered secondary structure compared with the dephosphorylated case. This supports the observations discussed above, and see Table 4.

This general observation is also confirmed by our ^{1}H NMR analysis (data not shown). Plotting the NMR spectra of phosphorylated and dephosphorylated α_s-casein together, so that intensities of the peaks present in both spectra were the same, showed a decrease in amide signal dispersion. In addition, a relative decrease of NOE cross-peak intensity in the amide–amide and amide–aromatic regions for the dephosphorylated α_s-casein sample was noted, in comparison with that of the phosphorylated form. Finally, changes in the histidine side chain band at 1400 cm^{-1} are also noted in Table 4, which at this stage merely confirms that structural changes are occurring upon dephosphorylation.

Raman optical activity (ROA) was introduced above as a particularly useful and accurate method for investigating protein structure. The method measures a small difference in Raman scattering from chiral molecules in right- and left-circularly polarized light.[29,30] Therefore, ROA has the same relation to conventional Raman spectroscopy as ultraviolet circular dichroism (UVCD) does to conventional ultraviolet absorption spectroscopy. As ROA spectra are sensitive to chirality they are an incisive probe of protein structure because only the few vibrational coordinates that most directly sample the skeletal chirality, such as those in the peptide backbone, make the largest contributions to ROA spectra. ROA bands are sensitive to secondary structure and can be particularly informative for studying unfolded proteins.[33–36] The intensity of the ROA amide I bands are also sensitive to tertiary structure, and these decrease as tertiary interactions decrease. Therefore, to complement the conventional Raman data, ROA spectra for phosphorylated and dephosphorylated α_s-casein were also obtained.

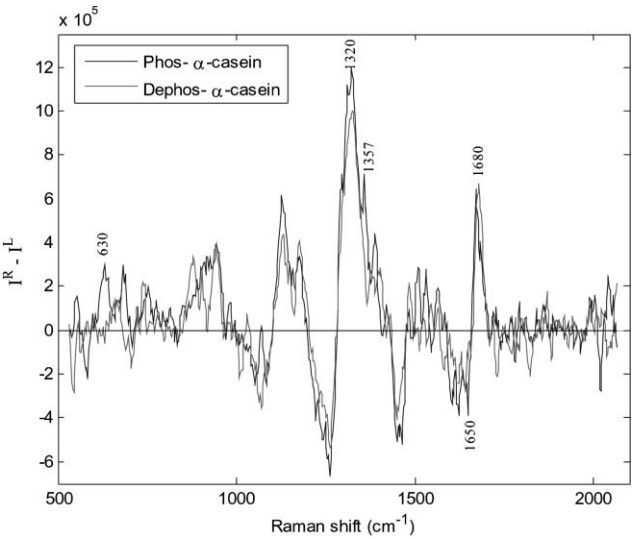

Fig. 5 Scaled and baseline corrected ROA spectra of phosphorylated and dephosphorylated α_s-casein acquired using 532.5 nm excitation. The dephosphorylated α-casein ROA spectrum was normalised to the same experimental conditions used for phosphorylated α-casein.

Fig. 5 shows the ROA spectra that were obtained from both phosphorylated and dephosphorylated α_s-casein. Chiral Raman spectra of phosphorylated α_s-casein have been reported previously in,[37] and these new ROA data very closely match, the published spectra, although the spectra reported here have a higher signal-to-noise ratio and are of better quality due to a lower fluorescence background. Several structural assignments are afforded by ROA analysis. Of these, a dominant band at 1320 cm^{-1} is present for both phosphorylated and dephosphorylated casein, indicating a high polyproline II (PPII) helix content. This is consistent with previous ROA studies of caseins.[37,38] As for the Raman spectra shown in Fig. 1, the ROA spectra for phosphorylated and dephosphorylated α_s-casein are similar; however several differences are apparent. One obvious difference, the weak band at 1357 cm^{-1}, supports the observation that a promotion of tryptophan residues occurs upon dephosphorylation. It is also apparent that the amide I couplet, $-$ve at 1650 and +ve at 1680 cm^{-1}, is weaker for dephosphorylated α_s-casein, indicating a decrease in tertiary structure with respect to phosphorylated α_s-casein. A band shift in the position of the positive component of this couplet also occurs, further indicating conformational change.

However, one additional observation can be made that was not available from the NIR Raman and NMR studies. The +ve band at *ca.* 630 cm^{-1} in the ROA spectrum of dephosphorylated casein, is near to a feature that has recently been associated with C–S stretching modes in cysteine.[73] This feature appears to be sensitive to the conformation of cysteine residues and is consistent with the Swiss-Prot (www.expasy.org/sprot/) sequence search for α_{s1}-casein, which identifies a single cysteine residue at position 8. Specifically, a positive ROA band at 658 cm^{-1} was found to correspond to the H–C_α bond of cysteine, being trans to the adjacent C_β–S bond.[73] At this stage we are not able to associate the band observed in Fig. 5 to a specific conformation; but we ascribe

this band to the stereochemistry of the cysteine residue. Interestingly, this band directly relates to an amino acid residue that is present in the molecule, and not a secondary structural motif. The relatively strong intensity of this particular band in this datum is perhaps rather extraordinary, but with ROA it is possible that signals from low abundance features are more intense due to an absence of signal averaging from enantiomeric structures. However, further work needs to be undertaken to determine why this band appears to be absent in the spectrum of the phosphorylated form. A possible reason is that subtly different forms of α_s-casein do exist, and it is perhaps the case that in the phosphorylated sample the proportion of this particular α_s-casein sequence is much lower.

Conclusions

This study clearly demonstrates the application of Raman spectroscopy to the exploratory analysis of PTMs, as demonstrated by investigating protein phosphorylation. Although a potential weakness of the method is the lack of ability to detect directly P=O vibrations in phosphorylated proteins, we have shown that NIR Raman spectroscopy can be used to quantify the relative concentrations of mixtures of phosphorylated and dephosphorylated α_s-casein accurately. In addition, we have demonstrated that advanced data mining strategies, such as GA-PLSR, have the ability to pinpoint the major biochemical changes that allow for quantification of these systems. Furthermore, in conjunction with ROA (and NMR) spectroscopy we have shown that Raman spectroscopic methods are highly descriptive of protein conformation and can be used to define the structural changes that occur upon dephosphorylation.

Acknowledgements

RMJ and RG are indebted to the UK's BBSRC for funding and EWB would like to thank the UK's EPSRC for funding. We would also like to thank Tienta Sciences, Indianapolis, USA, for supplying the SpectRIM™ slides.

References

1 M. Mann, S. Ong, M. Gronborg, H. Steen, O. N. Jensen and A. Pandey, *Trends Biotechnol.*, 2002, **20**, 261–268.
2 D. Tsur, S. Tanner, E. Zandi, V. Bafna and P. Pevzner, *Nat. Biotechnol.*, 2005, **23**, 1562–1567.
3 B. Williamson, J. Marchese and N. Morrice, *Mol. Cell. Proteomics*, 2006, **5**, 337–346.
4 M. Washburn, D. Wolters and J. Yates, *Nat. Biotechnol.*, 2001, **19**, 242–247.
5 H. Zhou, J. Ranish, J. Watts and R. Aebersold, *Nat. Biotechnol.*, 2002, **20**, 512.
6 M. Mann and O. N. Jensen, *Nat. Biotechnol.*, 2003, **21**, 255–260.
7 F. Schotte, M. Lim, T. Jackson, A. Smirnov, J. Soman, J. Olson, G. Phillips, M. Wulff and P. Anfinrud, *Science*, 2003, **300**, 1944–1947.
8 D. Picot, P. J. Loll and R. M. Garavito, *Nature*, 1994, **367**, 243–249.
9 J. Yano, M. Wester, G. Schoch, K. Griffin, C. Stout and E. Johnson, *J. Biol. Chem.*, 2004, **279**, 38091–38094.
10 R. Zahn, A. Liu, T. Lührs, R. Riek, C. v. Schroetter, F. L. García, M. Billeter, L. Calzolai, G. Wider and K. Wüthrich, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 145–150.
11 F. L. García, R. Zahn, R. Riek and K. Wüthrich, *Biophysics*, 2000, **97**, 8334–8339.
12 A. Golovanov, A. Lomize, A. Arseniev, Y. Utkin and V. Tsetlin, *Eur. J. Biochem.*, 1993, **213**, 1213–1223.
13 S. Provencher and J. Glockner, *Biochemistry*, 1981, **20**, 33–37.
14 W. C. Johnson, Jr., *Proteins: Struct., Funct., Genet.*, 1990, **7**, 205–214.
15 R. Banks, M. Dunn, D. Hochstrasser, J. Sanchez, W. Blackstock and D. Pappin, *Lancet*, 2000, **356**, 1749–1756.
16 O. Jensen, in *Proteomics: A Trends Guide*, 2000, pp. 36–42.
17 A. Marshall, C. Hendrickson and G. Jackson, *Mass Spectrom. Rev.*, 1998, **17**, 1–35.
18 P. Carey, *J. Biol. Chem.*, 1999, **274**, 26625–26628.
19 H. Vogel and F. Jahnig, *J. Mol. Biol.*, 1986, **190**, 191–199.
20 A. Desormeaux, J. Blochet, M. Pezolet and D. Marion, *Biochim. Biophys. Acta*, 1992, **1121**, 137–152.
21 S. Sane, S. Cramer and T. Przybycien, *Anal. Biochem.*, 1999, **269**, 255–272.
22 B. Bussian and C. Sander, *Biochemistry*, 1989, **28**, 4271–4277.
23 A. Mikhonin, S. Bykov, N. Myshakina and S. Asher, *J. Phys. Chem. B*, 2006, **110**, 1928–1943.
24 K. V. Pimenov, S. V. Bykov, A. V. Mikhonin and S. A. Asher, *J. Am. Chem. Soc.*, 2005, **127**, 2840–2841.
25 J. Clarkson, C. Sudworth, S. I. Masca, D. N. Batchelder and D. A. Smith, *J. Raman Spectrosc.*, 2000, **31**, 373–375.
26 J. Clarkson and D. A. Smith, *FEBS Lett.*, 2001, **503**, 30–34.
27 J. Clarkson, D. N. Batchelder and D. A. Smith, *Biopolymers*, 2001, **62**, 307–314.
28 M. Nagai, S. Kaminaka, Y. Ohba, Y. Nagai, Y. Mizutani and T. Kitagawa, *J. Biol. Chem.*, 1995, **270**, 1636–1642.
29 L. Nafie, *Annu. Rev. Phys. Chem.*, 1997, **48**, 357–386.
30 L. Barron and L. Hecht, in *Circular Dichroism, Principles and Applications*, eds. N. Berova, K. Nakanishi and R. Woody, John Wiley and Sons, New York, 2000, pp. 667–701.
31 L. Barron and A. Buckingham, *Molecular Physics*, 1971, **20**, 1111–1119.
32 L. Barron, M. Bogaard and A. Buckingham, *J. Am. Chem. Soc.*, 1973, **95**, 603–605.
33 E. Blanch, L. Morozova-Roche, D. Cochran, A. Doig, L. Hecht and L. Barron, *J. Mol. Biol.*, 2000, **301**, 553–563.
34 L. Barron, L. Hecht and E. Blanch, *Adv. Protein Chem.*, 2002, **62**, 51–90.
35 E. Blanch, D. Kasarda, L. Hecht, K. Nielsen and L. Barron, *Biochemistry*, 2003, **42**, 5665–5673.
36 E. Blanch, A. Gill, A. Rhie, J. Hope, L. Hecht, K. Nielsen and L. Barron, *J. Mol. Biol.*, 2004, **343**, 467–476.
37 E. Smyth, C. Syme, E. Blanch, L. Hecht, M. Vasak and L. Barron, *Biopolymers*, 2001, **58**, 138–151.
38 C. Syme, E. Blanch, C. Holt, R. Jakes, M. Goedert, L. Hecht and L. Barron, *Eur. J. Biochem.*, 2002, **269**, 148–156.
39 D. S. Horne, *Curr. Opin. Colloid Interface Sci.*, 2002, **7**, 456–461.
40 J. Mercier, F. Grosclaude and B. Ribadeau-Dumas, *Eur. J. Biochem.*, 1971, **23**, 41–51.
41 L. Creamer, T. Richardson and D. Parry, *Arch. Biochem. Biophys.*, 1981, **211**, 689–696.
42 D. Curley, T. Kumosinski, J. Unruh and H. Farrell, *J. Dairy Sci.*, 1998, **81**, 3154–3162.
43 D. Byler and H. Susi, *Ind. Microbiol.*, 1988, **3**, 73–88.
44 T. F. Kumosinski, E. M. Brown and H. M. Farrell, Jr., *J. Dairy Sci.*, 1991, **74**, 2889–2895.
45 K. P. J. Williams, G. D. Pitt, D. N. Batchelder and B. J. Kip, *Appl. Spectrosc.*, 1994, **48**, 232–235.
46 K. P. J. Williams, G. D. Pitt, B. J. E. Smith, A. Whitley, D. N. Batchelder and I. P. Hayward, *J. Raman Spectrosc.*, 1994, **25**, 131–138.
47 H. Martens, J. P. Nielsen and S. B. Engelsen, *Anal. Chem.*, 2003, **75**, 394–404.
48 W. Hug, in *Handbook of Vibrational Spectroscopy*, John Wiley and Sons, Chichester, UK, 2002, vol. 1, pp. 745–758.
49 W. Hug and G. Hangartner, *J. Raman Spectrosc.*, 1999, **30**, 841–852.
50 H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, 1989.
51 S. Wold, H. Martens and H. Wold, Proceedings of the Conference on Matrix Pencils, Heidelburg, 1982.

52 M. Winson, R. Goodacre, A. Woodward, É. Timmins, A. Jones, B. Alsberg, J. Rowland and D. Kell, *Anal. Chim. Acta*, 1997, **348**, 273–282.
53 R. Brereton, *Chemometrics: data analysis for the laboratory and chemical plant*, John Wiley & Sons Ltd, Chichester, 1st edn., 2003.
54 R. Jarvis and R. Goodacre, *Bioinformatics*, 2005, **21**, 860–868.
55 D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
56 J. H. Holland, *Adaptation in natural and artificial systems*, MIT Press, Boston, 1992.
57 D. Broadhurst, R. Goodacre, A. Jones, J. Rowland and D. Kell, *Anal. Chim. Acta*, 1997, **348**, 71–86.
58 T. Bäck, D. B. Fogel and Z. Michalewicz, *Handbook of Evolutionary Computation*, IOP Publishing/Oxford University Press, Oxford, 1997.
59 M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Boston, 1995.
60 D. M. Allen, *Technometrics*, 1971, **13**, 469–475.
61 P. Cipriani and D. Ben-Amotz, *J. Raman Spectrosc.*, 2005, **36**, 1052–1058.
62 D. Zhang, Y. Xie, M. Mrozek, C. Ortiz, J. Davisson and D. Ben-Amotz, *Anal. Chem.*, 2003, **75**, 5703–5709.
63 I. Degen, *Tables of Characteristic Group Frequencies for the Interpretation of Infrared and Raman Spectra*, Acolyte Publications, Harrow, UK, 1997.
64 G. Sauer, W. Zunic, J. Durig and R. Wuthier, *Calcif. Tissue Int.*, 1994, **54**, 414–420.
65 D. Kell, *Curr. Opin. Microbiol.*, 2004, **7**, 296–307.
66 D. B. Kell and S. G. Oliver, *Bioessays*, 2004, **26**, 99–105.
67 W. Krzanowski, *Principles of Multivariate Analysis: a User's Perspective*, Oxford University Press, 1988.
68 J. Handl and J. Knowles, *Exploiting the Trade-off—The Benefits of Multiple Objectives in Data Clustering*, 3410 edn., 2005.
69 H. S. Tapp, M. Defernez and E. K. Kemsley, *J. Agric. Food Chem.*, 2003, **51**, 6110–6115.
70 E. Kinoshita, Y. Ozawa and T. Aishima, in *Flavonoids in the Living System*, Plenum Press, New York, 1998, vol. 439, pp. 117–129.
71 M. Tsuboi, T. Ueda, K. Ushizawa, Y. Ezaki, S. Overman and G. Thomas, *J. Mol. Struct.*, 1996, **379**, 43–50.
72 T. Miura and G. Thomas, Jr., in *Subcellular Biochemistry. Vol. 24: Proteins: Structure, Function and Engineering*, eds. B. B. Biswas and S. Roy, Plenum Press, New York, 1995, pp. 55–99.
73 J. Kapitan, V. Baumruk and V. Gut, *Collect. Czech. Chem. Commun.*, 2005, **70**, 403–409.

