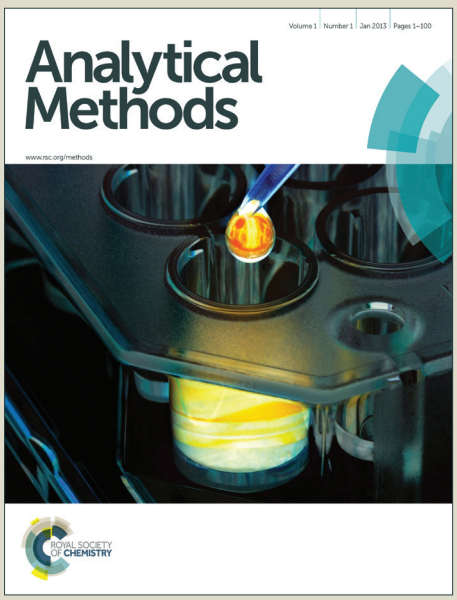
# Analytical Methods

Accepted Manuscript





ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Journal Name

RSCPublishing

ARTICLE

# Quantification of Protein Secondary Structure Content by Multivariate Analysis of Deep-Ultraviolet Resonance Raman and Circular Dichroism Spectroscopies

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

**Olayinka O. Oshokoya[a], Carol A. Roach[a] and Renee D. JiJi*[a]**

Determination of protein secondary structure (α-helical, β-sheet, and disordered motifs) has become an area of great importance in biochemistry and biophysics as protein secondary structure is directly related to protein function and protein related diseases. While NMR and x-ray crystallography can predict the placement of each atom in a protein to within an angstrom, optical methods (i.e. CD, Raman, and IR) are the preferred techniques for rapid evaluation of protein secondary structure content. Such techniques require calibration data to predict unknown protein secondary structure content where accuracy may be improved with the application of multivariate analysis. Here, a comparison of the protein secondary structure predictions obtained from multivariate analysis of ultraviolet resonance Raman (UVRR) and circular dichroism (CD) spectroscopic data using classical least squares (CLS), partial least squares (PLS), and multivariate curve resolution-alternating least squares (MCR-ALS) is made. Results of the multivariate analysis suggest that CD measurements provide more accurate prediction of protein α-helical content whereas UVRR more accurately predicts β-sheet content, an observation that is consistent with previous studies. Based on this analysis it is suggested that the best approach to rapid and accurate protein secondary structure determination is to combine both CD and UVRR spectroscopic data.

## 1. Introduction

In biochemistry and biophysics protein secondary structure is the arrangement of a subset of the amino acids in a repeating pattern, generally referred to as α-helices, β-sheets, and disordered motifs. Protein secondary structure may directly impact tertiary (entire protein) and quaternary (protein-protein) structure, and thus give important insight into protein function and diseases caused by protein misfolding [1, 2]. Protein secondary structural motifs are designated by the φ and ψ dihedral angles of the amide backbone, categorized as helical (α-helical (φ = -57°, ψ= -47°) and $3_{10}$-helical (φ = -49°, ψ= -26°)), β-sheet (antiparallel (φ = -139°, ψ= 135°) and parallel (φ = -120°, ψ= 115°)), or disordered (unfolded and structures having non-repetitive φ and ψ angles, e.g., turns, loops, etc...)[3-5]. Due to the importance of secondary structure motifs in protein function several techniques with varying levels of accuracy and complexity have been developed to quantify these structural features. Exact structure determination methods such as X-ray crystallography (XRC) and nuclear magnetic resonance (NMR) allow determination of the three-dimensional placement of each atom in a protein structure to within sub-angstrom resolution, however such methods may require extensive preparatory work and data analysis [6, 7]. When only the total or change in secondary structure content of a protein is desired, simple and rapid methods, such as conventional Raman, ultraviolet resonance Raman (UVRR), infrared (IR) absorption and circular dichroism (CD), are preferred because structural information is available without the delay of lengthy data analysis[8-14]. Additionally, studies have shown that quantification of secondary structure content is possible by combining multivariate methods with these simple and rapid spectroscopic techniques and a limited set of standard proteins [15-17], albeit with prediction errors as high as 10-15%[10, 15, 16, 18].

The origin of the protein secondary structural sensitivity of CD and Raman spectroscopies derives from the absorption of

photons by the amide backbone. CD is the current standard in secondary structure analysis of proteins and UVRR is an up-and-coming technique. In UVRR, the vibrational amide modes (I, II, III, and S) of the protein are enhanced, and shifts in position and intensity differences in these modes exist because of the limited molecular motions allowed by each secondary structural motif (Figure 1A)[19-21]. In particular, the amide S mode only appears in a UVRR spectrum if there is disordered or β-sheet structure within the protein[21, 22]. Use of the UVRR amide modes to predict secondary structure content can be complicated by the presence of aromatic amino acids (phenylalanine, tryptophan, and tyrosine) with vibrational modes that overlap the amide bands. UVRR has also been able to determine and monitor π – and $3_{10}$–helices using the amide III region of spectra at both 194 and 204 nm.[23] CD spectroscopy measures the difference in absorption of left and right handed circularly polarized light by a sample, which is related to the different structural motifs present in a protein.[9, 24, 25] The CD spectra for α-helix, β-sheet, and disordered protein structures are quite different (Figure 1B) and the dominant structural feature of the protein often dominates the acquired spectrum. For instance, the CD spectra from α- and π- helices are very similar making them very difficult to distinguish mathematically.[26] The spectral response (*s*) of a protein for both techniques is the sum of the relative responses ($\mathbf{s}_\alpha$, $\mathbf{s}_\beta$ and $\mathbf{s}_\tau$) and fractional amounts ($f_\alpha$, $f_\beta$ and $f_\tau$) of each secondary structure type;

$$\mathbf{S} = f_\alpha \mathbf{s}_\alpha + f_\beta \mathbf{s}_\beta + f_\tau \mathbf{s}_\tau \quad (1)$$

where α designates α-helical, β designates β-sheet, and τ designates disordered related variables.

When more than one secondary structure is present in a protein, as is often the case, the spectral features become convoluted and quantification of each individual motif may be better addressed with the use of multivariate methods. Multivariate calibration methods assume a linear relationship between spectral intensity (variable response) and the relative amounts of analytes in a mixture. In the case of proteins, the measured spectra (**X**) can be modelled as the product of the secondary structure content (**C**) and the underlying pure spectral profiles of each type of secondary structure (**S**) plus an error matrix (**E**) according to Equation 2:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (2)$$

A wide variety of multivariate analysis techniques have been developed for obtaining structural information from UVRR and CD spectra of proteins [9, 16, 17, 25, 27, 28]. However, the relative performance of various multivariate calibration methods on the prediction of secondary structure content has only been studied to a limited extent and mostly on IR-CD combined data sets[29, 30]. Herein the performance of a partial least squares (PLS), classical least square (CLS), and multivariate curve resolution-alternating least squares (MCR-ALS) method on both UVRR and CD spectra of a set of nine globular proteins are compared. The accuracy of each multivariate method is assessed by comparison to the secondary structure content determined by XRS and NMR as listed on the protein data bank (PDB). These multivariate calibration methods have been extensively reviewed in the literature[31-37].

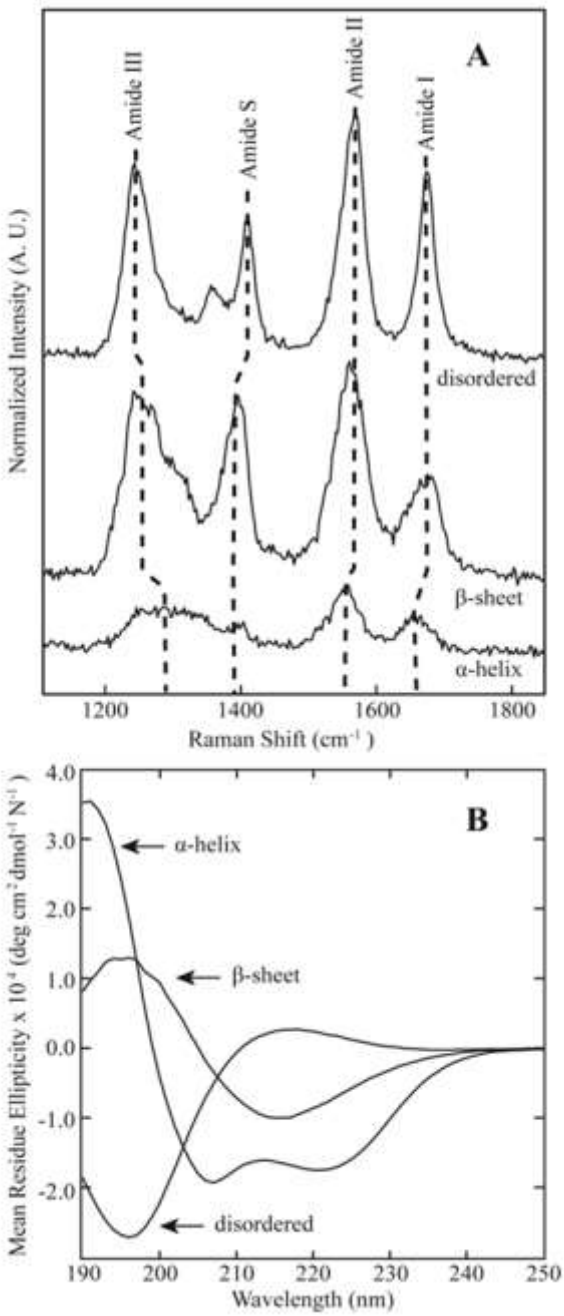



**Fig. 1** UVRR (A) and CD (B) spectra of poly-L-lysine in α-helix (25°C, pH 11.0), β-sheet (52°C, pH 11.3) and disordered (25°C, pH 4.0) conformations.

## 2. Experimental

### 2.1. Sample Preparation

Nine globular proteins with varying secondary structure content (Table 1), poly-L-lysine (70,000-150,000 g $mol^{-1}$) and amino acids L-phenylalanine (F) and L-tyrosine (Y) were purchased from Sigma Aldrich (St. Louis, MO) and used without further purification. The proteins and amino acids were prepared in phosphate buffer (pH 7.2). α-Helical and disordered poly-L-lysine were prepared by dissolving the peptide in pH 11.3 and pH 4 phosphate buffer, respectively. α–Helical poly-L-lysine was converted to β-sheet structure by heating the sample to 52°C. Concentrations were verified by UV-Visible absorption using a Hewlett Packard 8453 spectrometer (Palo Alto, CA), and were 0.5 mg $mL^{-1}$ for protein and peptide solutions, and 200 μM for the amino acids.

The proteins chosen for this study were globular proteins that could be obtained at low cost and readily soluble in water-based phosphate buffer. The experimental design took into consideration a range of proteins that showed a trend of increasing helical content and overall a well-proportioned combination of the major secondary structures. The experimental design also strives to prove that a limited amount of proteins can also be used to achieve secondary structure determination using multivariate analysis.

**Table 1** Secondary structure content (%) of proteins used as found on the Protein Data Bank.

| Protein | Abbreviation | Helix | β-sheet | Disordered |
|---|---|---|---|---|
| Bovine Serum Albumin | BSA[38] | 74.0 | 0.0 | 26.0 |
| Carbonic Anhydrase | CAH[39] | 17.8 | 29.0 | 53.2 |
| Chymotrypsinogen A | CTG[40] | 13.5 | 32.0 | 54.5 |
| Cytochrome C | CYC[41] | 41.0 | 1.0 | 58.0 |
| Glucose Oxidase | UOX[42] | 34.5 | 19.6 | 46.0 |
| Lysozyme | LSZ[43] | 41.9 | 6.2 | 51.9 |
| Myoglobin | MBN[44] | 73.9 | 0.0 | 26.1 |
| Ovalbumin | OVA[45] | 32.7 | 31.9 | 35.3 |
| Trypsinogen | TGN[46] | 10.1 | 31.4 | 58.5 |

### 2.2. Instrumentation

The UVRR instrument used to collect protein spectra has been previously described.[47] Briefly, a Nd:YLF pumped Ti:Sapphire laser is frequency quadrupled (Coherent Inc., Santa Clara, CA) to provide a 197 nm excitation source. Sample is circulated through two nitinol wires (Small Parts Inc., Miramar, FL) to create a thin film under a nitrogen purge, and is temperature controlled by a water-jacketed reservoir (Mid Rivers Glassblowing, Saint Charles, MO) using a bath recirculator (Isotemp 3016D, Fisher Scientific, Pittsburgh, PA). Scattering is collected in the 135° backscattering geometry and directed into a 1.2 m spectrometer (Horiba Jobin Yvon Inc., Edison, NJ) equipped with a Symphony CCD detector and spectra collected using Synerjy software (Horiba Jobin Yvon Inc., Edison, NJ). Each spectrum was the sum of 3 hours of signal collection and run in triplicate.

Circular dichroism spectra were obtained using a Model 62DS spectrometer (Aviv, Lakewood, NJ) from 190-250 nm. The instrument temperature control program was used for poly-L- lysine collection in order to maintain sample temperature and β-sheet composition. Protein and peptide samples were diluted to 0.2 mg $mL^{-1}$ for CD measurements, and signal was collected for 5 s at each wavelength and averaged over 5 scans to produce one spectrum for a total of 3 spectra for each sample.

### 2.3. Data Processing

Analysis of all data was carried out in MATLAB (version 7.11, Mathworks, Natick MA). Cosmic rays were removed using an in-house program, base-lined using the MATLAB curve-fitting toolbox, and each spectrum truncated to the 1266-1759 $cm^{-1}$ range.[28] Contributions to spectra from aromatic side chains were removed using the phenyalanine band at 1003 $cm^{-1}$ (F12) and tyrosine band at 853 $cm^{-1}$ (Y1) (Figure 2).

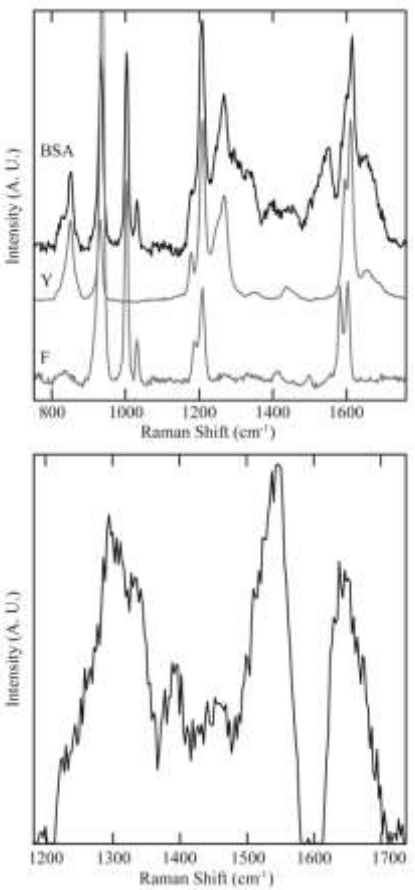



**Fig. 2** Top: BSA, phenylalanine and tyrosine UVRR spectra. Phenylalanine and tyrosine spectra are scaled to the bands at 1003 cm-1 (F12) and 853 cm-1 (Y1), respectively. Bottom: BSA spectrum with phenyalanine and tyrosine contributions subtracted.

Contributions from tryptophan were disregarded due to its negligible intensity in deep-UVRR spectra. Areas that appeared

to be negative in the spectrum after aromatic subtraction were set to zero. For CD data averaging of the 5 spectra collected was performed with no other pre-processing.

It was expected that the UVRR and CD protein spectra would be dominated by at least three principal components: the α-helical, β-sheet, and disordered conformations. Principal components analysis of the data via a singular value decomposition[33, 48] based scree plot suggested as few as three or as many as five components in the data matrix. Modelling of both data matrices was therefore conducted for three components (α-helical, β-sheet, and disordered), four components with $3_{10}$-helices, four components with β-turns, and 5 components with $3_{10}$-helices and β-turns. The models were evaluated for percentage relative error (%RE)

$$\%\ \mathrm{RE} = \frac{\left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right]^{1/2}}{\left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i)\right]} \times 100 \qquad (3)$$

where $n$ is the number of proteins, $y_i$ is the secondary structure content obtained from the PDB structures, and $\hat{y}_i$ is the value predicted. Comparison of each model's %RE values (Figure 3) shows that the UVRR error is lowest for the three component model, suggesting not all secondary structural types (helices, antiparallel and parallel sheets, different classes of turns and bends) are independent variable[49, 50]. On the other hand, for CD the five component model had the lowest average %RE.

Figure 3 shows a breakdown of the individual %RE of the different considered components in each model for both UVRR and CD. For UVRR, an increase in the number of components does not improve the predictive capability of the model for any of the structures; rather, it diminishes the predictive capability especially for disordered structure types. For CD, the high average %RE's are as a result of the method's poor predictive capability for β-sheet structure (Fig. 3). Figure 3 also shows that while an increase in the number of components reduces the %RE of the β-sheet structure, the %RE for β-sheet prediction is still very high (>110%) and an increase in the number of components does not improve the prediction of helical structure. The increase in the number of components in the CD model also diminishes the %RE for disordered structure prediction. Therefore, all UVRR and CD data was further processed with only three components.

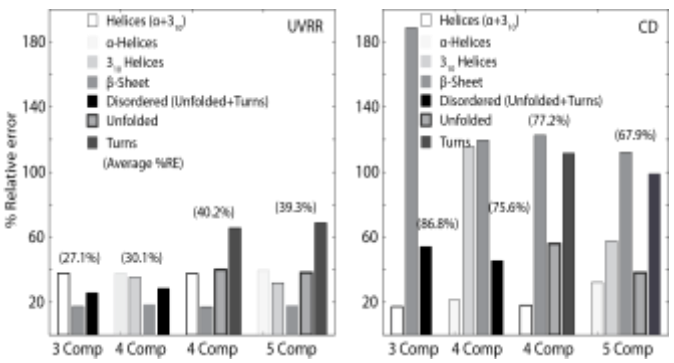



**Fig. 3** %RE of the different considered components in each model for both UVRR and CD.

For both UVRR and CD analysis, the triplicate spectra were compiled so that 27 individual spectra became the data matrix. From the data matrix, 22 spectra were randomly selected as the training set; the five remaining spectra were designated as the test set. For each multivariate method (CLS, PLS and MCR-ALS), the secondary structure content of the test set proteins were calculated using the model built from the training set. The process was repeated 30 times for each multivariate method in order to obtain a mean prediction error for the technique using root mean squared error of cross-validation (RMSECV) such that:

$$\mathrm{RMSECV} = \left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \qquad (4)$$

where $n$ is the number of proteins, $y_i$ is the secondary structure content obtained from the PDB structures, and $\hat{y}_i$ is the value predicted by the algorithm. Algorithms written in-house were employed for CLS [31, 36] and PLS analyses. The MCR-ALS algorithm was developed by Tauler *et.* al.[51] and is freely available. An in-house rotation matrix algorithm was used to optimize the output profiles from the MCR-ALS analysis of the UVRR and CD data. Briefly, the Raman protein data matrix (**X**) is related to the secondary structure content (**C**) and pure secondary structure spectra (**S**) as per equation 1, therefore pure secondary structure spectra may be obtained by:

$$\mathbf{S} = \mathbf{X}\mathbf{C}^{+} \qquad (5)$$

where + denotes the matrix pseudo-inverse. However, due to noise in the spectral measurement (the error matrix, **E**), the **S** obtained from MCR-ALS ($\mathbf{S}_{\mathrm{MCR}}$) is only an approximation of the pure secondary structure, and if used to determine the known concentrations of the model does not give the original concentration matrix **C** such that:

$$\mathbf{C}_{\mathrm{MCR}} = \mathbf{X}\mathbf{S}_{\mathrm{MCR}}^{+} \qquad (6)$$

where $\mathbf{C}_{\mathrm{MCR}}$ is only an approximation of the original concentration matrix. Both the approximate concentration, $\mathbf{C}_{\mathrm{MCR}}$, and pure secondary structure spectra, $\mathbf{S}_{\mathrm{MCR}}$, are related to the actual concentrations, **C**, and pure secondary structure spectra, **S**, by a rotation matrix, **W**:

$$\mathbf{C} = \mathbf{C}_{\mathrm{MCR}}\mathbf{W} \qquad (7)$$

$$\mathbf{S} = \mathbf{S}_{\mathrm{MCR}}\mathbf{W}^{-1} \qquad (8)$$

Such that:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} = \mathbf{C}_{\mathrm{MCR}}\mathbf{W}(\mathbf{S}_{\mathrm{MCR}}\mathbf{W}^{-1}) = \mathbf{C}_{\mathrm{MCR}}\mathbf{S}_{\mathrm{MCR}} \qquad (9)$$

where $\mathbf{W}\mathbf{W}^{-1}$ is an identity matrix. Therefore, the error in the estimate of the actual concentrations can be minimized by using equation 7 on all predicted concentrations.[52-55]

Occasionally, the predicted content for a particular secondary structure, helical for UVRR and β-sheet for CD, fell below zero. Given that the predicted amounts of secondary structure should be zero or greater, these values were set to zero. The sum of the predicted amounts of each secondary structure type was set to unity.

## 3. Results

For both UVRR and CD spectroscopic methods, the most accurate prediction (lowest RMSECV) is obtained for the secondary structure type with the strongest spectral intensity, β-sheet for UVRR and α–helix for CD (Table 2). In order to compare the ability of CLS and MCR-ALS to resolve the pure underlying secondary structure UV Raman profiles, the resolved profiles were compared to the homo polypeptide poly-L-lysine (Figures 5 and 7). The PLS algorithm does not produce resolved pure spectra so the spectrum of the protein with the largest predicted content for each structure type is presented along with the associated predicted protein spectrum. These proteins are designated by their three letter abbreviation.

Reference spectra obtained from manipulation of poly L-lysine into the three major protein secondary structure conformations was chosen to evaluate the results of spectral resolution by CLS, PLS and MCR-ALS. It is quite possible for the disordered form of poly L-lysine to possess some residual local chain order or any other conformation for that matter.[16, 56] The inability to obtain total conformity to one secondary structure from globular or membrane proteins at large led to the decision of picking poly L-lysine as the polypeptide of choice for result evaluation. As a result, poly L-lysine spectra were only used for evaluating spectral resolution and not included in the modelling of the data or for prediction of secondary structure.

### 3.1. Results for UVRR

Overall, each multivariate method performed similarly with an average prediction error of approximately 10% (Table 2). The RMSECV was lowest for predicted amounts of β-sheet content, typically less than 5%. The error in prediction of α-helical content ranged from 14-16% before normalization. After normalization, the error in prediction of helical content dropped and ranged from 9-12%. A similar reduction in RMSECV was observed for the predicted amounts of disordered structure after normalization. In general, normalization improved secondary structure estimation from UVRR spectra.

The predicted percentages of each secondary structure type show a linear correlation with the known secondary structure composition (Figure 4). For the MCR-ALS algorithm, significant under predictions were observed for disordered structural content of both lysozyme and cytochrome *c*. To compensate for these under estimations in disordered structure, the helical contents (Figure 4) of those same proteins were over estimated. The common factor between lysozyme and cytochrome c is an absence of β-sheet structure.

Figure 4 presents the pure secondary structure spectra obtained from the multivariate analysis, with the exception of PLS where the protein with the largest predicted percentage of any one secondary structure is present instead. The predicted pure UVRR α-helical spectrum from CLS is unrealistic with both positive and negative features. In contrast, the predicted α-helical spectrum from MCR-ALS is the most interesting in that the amide S (1390 $cm^{-1}$) is absent and the amide III (~1240 $cm^{-1}$) modes are significantly reduced. These two amide modes are markers of non-helical structure.[16, 22] Only the MCR-ALS algorithm effectively removes these contributions from the pure secondary structure Raman spectrum (PSSRS). The position of the amide I (1648 $cm^{-1}$), II (1544 $cm^{-1}$) and III (1299 $cm^{-1}$) bands in the PSSRS from MCR-ALS are slightly lower than observed with α- poly L-lysine (Table 4) but are still within the expected region for a helical protein. Bovine serum albumin is predicted to be 82% helical by PLS. The predicted spectrum of BSA obtained from the PLS method appears similar to α-helical poly-L-lysine spectrum.

For β-sheet PSSRS, the spectrum obtained from the CLS algorithm is most similar to the β-sheet poly-L-lysine spectrum. For all three algorithms, the predicted amide I (1668-1673 $cm^{-1}$) and amide II (1552-1560 $cm^{-1}$) positions fall within the expected regions (~1668 $cm^{-1}$ for amide I, ~1563 $cm^{-1}$ for amide II) for a β-sheet protein [8, 22] (Table 3). The amide S mode is predicted to be downshifted to 1389 $cm^{-1}$ (CLS, PLS) and 1392 $cm^{-1}$ (MCR) from that of poly L-lysine which occurs at 1408 $cm^{-1}$. The amide III band of the CLS β- sheet spectrum (1240 $cm^{-1}$) is closest in position and shape to that of the β-sheet poly L-lysine spectrum. Whereas the amide III band in the predicted β- sheet spectrum from MCR is broad ranging from 1229-1271 $cm^{-1}$.

**Table 2** RMSECV (%) values calculated

| Algorithm | helix | β-sheet | Disordered | Average |
|---|---|---|---|---|
| UVRR | | | | |
| CLS | 14.4 | 3.3 | 11.0 | 9.5 |
| Normalized CLS | 9.0 | 2.6 | 9.0 | 6.9 |
| PLS | 16.3 | 4.0 | 10.7 | 10.3 |
| Normalized PLS | 12.1 | 5.8 | 9.1 | 9.0 |
| MCR-ALS | 15.7 | 4.0 | 14.2 | 11.3 |
| Normalized MCR-ALS | 12.2 | 4.0 | 12.0 | 9.4 |
| CD | | | | |
| CLS | 6.4 | 31.8 | 22.5 | 20.2 |
| Normalized CLS | 16.3 | 17.9 | 9.3 | 14.5 |
| PLS | 4.4 | 14.1 | 18.7 | 12.4 |
| Normalized PLS | 15.5 | 9.6 | 10.1 | 11.6 |
| MCR-ALS | 5.8 | 14.8 | 14.2 | 11.6 |
| Normalized MCR-ALS | 10.8 | 11.4 | 6.7 | 9.6 |
| CD + UVRR | | | | |
| CLS | 6.4 | 3.3 | 5.6 | 5.1 |
| PLS | 4.4 | 4.0 | 4.2 | 4.2 |
| MCR-ALS | 5.8 | 4.0 | 5.4 | 5.1 |

All the multivariate methods produced a disordered spectrum very similar to that of disordered poly L-lysine. Specifically, all the spectra have two distinct features in the

amide III region that occur at approximately 1240 and 1280 $cm^{-1}$. These features occur at approximately 1260 and 1300 $cm^{-1}$ in poly-L-lysine. The difference in the predicted positions versus disordered poly L- lysine may be due to the difference in amino acid composition between a globular protein and a homo polypeptide. The amide S mode is predicted to be 3-5 $cm^{-1}$ lower for disordered structure with respect to β-sheet structure regardless of multivariate method, similar to poly-L-lysine (Table 3). The predicted amide I and II bands also occur in the expected experimental amide I (1548–1561$cm^{-1}$) and amide II (1661–1682 $cm^{-1}$) regions[8].

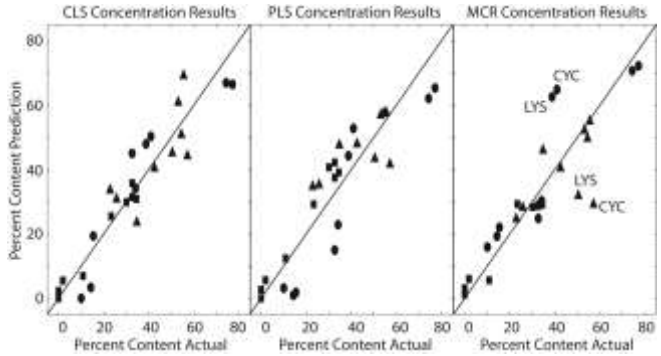



**Fig. 4** The actual versus predicted percentage composition for UVRR of α-helical (circles), β-sheet (squares), and disordered (triangles) structures as a percentage of content. The (1,1) line is shown to illustrate the deviations in the prediction.

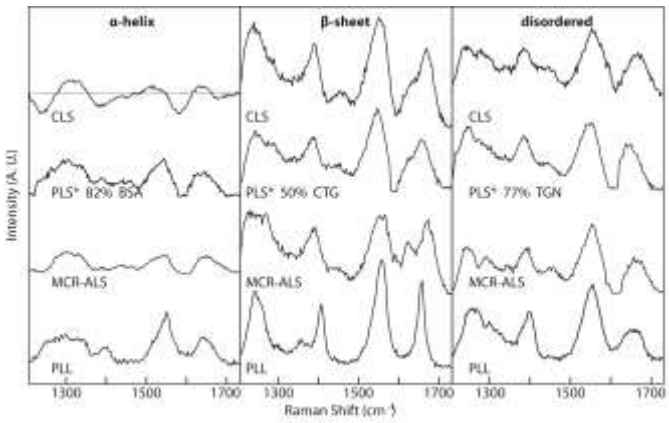



**Fig. 5** The PSSRS obtained from the various methods compared to the poly-L-lysine (PLL) pure conformer spectra. The dotted line is used to indicate the zero line for the spectrum that has a negative region. PLS does not produce PSSRS, so the largest predicted content for each structure type obtained during the iterative calculations is presented along with the protein spectrum (designated by the 3 letter abbreviation) associated with the prediction.

### 3.2. Results for CD

For CD, the most accurate predictions were obtained for the amount of α-helical content in each protein. The PLS algorithm predicted the helical content most accurately with an RMSECV of 4.4%. MCR-ALS performed nearly as well with a RMSECV of 5.8% (Table 2). Regardless of multivariate method (CLS, PLS, MCR-ALS), a linear correlation between the known secondary structure composition and the predicted amounts of each type of secondary structure was obtained (Figure 5). However, the predicted percentages of each secondary structure type from PLS and MCR-ALS cluster more tightly on the (1,1) line indicating a greater error in the predicted secondary structure compositions for CLS. Overall, the RMSECV for prediction of secondary structure compositions from CD spectra are significantly higher for β-sheet and disordered structure (Table 2). While normalization improves prediction of β-sheet and disordered contents, it appears to degrade the prediction of α-helical content from CD spectra.

**Table 3.** Frequencies ($cm^{-1}$) of amide bands in the resolved UVRR spectra for secondary structure obtained from CLS, PLS, MCR-ALS and the poly-L-lysine (PLL) pure conformer spectra.

| | CLS | PLS | MCR | PLL[16] |
|---|---|---|---|---|
| Helix | | | | |
| Amide III | - | 1257 | - | 1253 |
| Amide III | 1308 | 1299 | 1299 | 1291 |
| Amide S | - | *1386* | - | *1401* |
| Amide II | 1516 | 1549 | 1544 | 1552 |
| Amide I | 1647 | 1656 | 1648 | 1650 |
| β-Sheet | | | | |
| Amide III | 1240 | 1240 | 1229 | 1247 |
| Amide III | - | - | 1271 | - |
| Amide S | 1389 | 1389 | 1392 | 1408 |
| Amide II | 1552 | 1558 | 1560 | 1563 |
| Amide I | 1670 | 1668 | 1673 | 1668 |
| Disordered | | | | |
| Amide III | 1237 | 1240 | 1240 | 1260 |
| Amide III | 1280 | 1282 | 1288 | 1298 |
| Amide S | 1384 | 1384 | 1389 | 1398 |
| Amide II | 1552 | 1558 | 1552 | 1560 |
| Amide I | 1668 | 1659 | 1668 | 1667 |

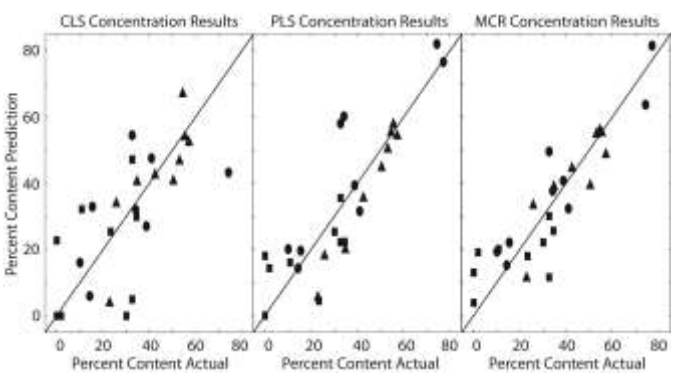



**Fig. 6** The predicted versus actual percent composition of secondary structure from CD analysis of helical (circles), β-sheet (squares), and disordered (triangles) structures as a percentage of protein content. The (1,1) line is shown to illustrate the deviations in the prediction.

The resolved pure CD spectra from CLS and MCR are shown in Figure 6. As mentioned above, the PLS algorithm



does not produce resolved pure spectra. Therefore, the predicted protein spectrum with the largest predicted content for each structure type is presented along with the protein spectrum. Each PLS spectrum is designated by the representative protein's three letter abbreviation. The predicted α-helical spectrum from each method appears similar to the pure CD spectrum of α-helical poly-L-lysine. The resolved pure β-sheet CD spectra from both the CLS and MCR-ALS are inconsistent with the CD spectrum of β-sheet structured poly-L-lysine. For CLS, the minimum is shifted to 205 nm from 212 nm for the CD spectrum of β-sheet structured poly L-lysine. The predicted pure β-sheet CD spectrum from MCR-ALS is unrealistic with a minimum of 200 nm versus the expected minimum of 217 nm for pure β-sheet structure. Therefore, this factor likely represents a mixture of β-sheet and disordered content.

The pure resolved disordered CD spectra from CLS and MCR have minima at 191 nm, 5 nm lower than the minima of disordered poly-L-lysine. The resolved pure disordered spectra also have positive features as expected for an unfolded protein but the positive features are unrealistically broad. Thus, neither algorithm sufficiently predicts the pure disordered CD spectrum. Chymotrypsinogen is predicted to have the greatest amount of disordered (49%) and β-sheet (34%) structure via PLS. Indeed, the predicted CD spectrum is characteristic of a protein with large amounts of disordered and β-sheet structure with a broad minimum at 202 nm that extends out to almost 230 nm.

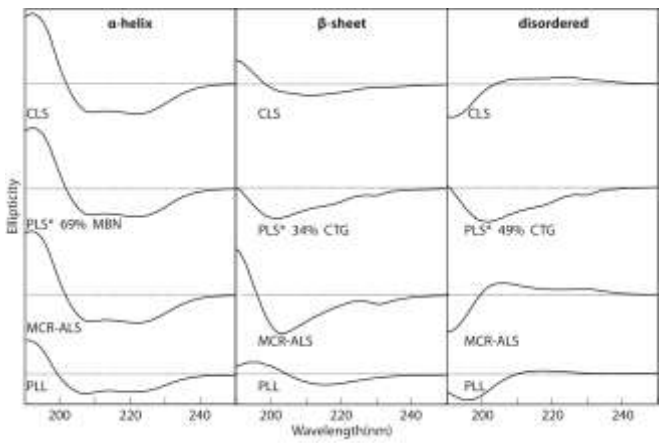



**Fig. 7** The CD pure spectra obtained from the various methods compared to the poly-L-lysine (PLL) pure conformer spectra. The dotted line is used to indicate the zero line for the spectrum that has a negative region. PLS does not produce PSSRS, so the largest predicted content for each structure type obtained during the iterative calculations is presented along with the protein spectrum associated with the prediction.

### 3.3. Results for UVRR + CD - Improving prediction of disordered structure

An accuracy of about 5% can be achieved when predicting helical content with CD and β-sheet content with UVRR. However, the error in the prediction of disordered (unfolded) structure remains around 10% with a minimum of 6.7% (normalized MCR-ALS/CD) and a maximum of 22.5% (CLS/CD). In order to improve the prediction of the fraction of disordered structure ($f_D$), the predicted percentages of α-helical ($f_\alpha$) and β-sheet ($f_\beta$) structure from CD and UVRR were combined, where $f_D = 100 - (f_\alpha + f_\beta)$. Prediction of disordered structure was improved and the RMSECV for disordered structure lowered to about 5% for each multivariate method, a significant improvement to other multivariate approaches where CD and IR spectroscopic data is combined with an average error of approximately 7%.[29]

A plot of the predicted amount of disordered structure versus the amount determined from the PDB structure for MCR illustrates how the values cluster more tightly to the (1,1) line when both types of spectroscopy are incorporated into the prediction (Figure 8).

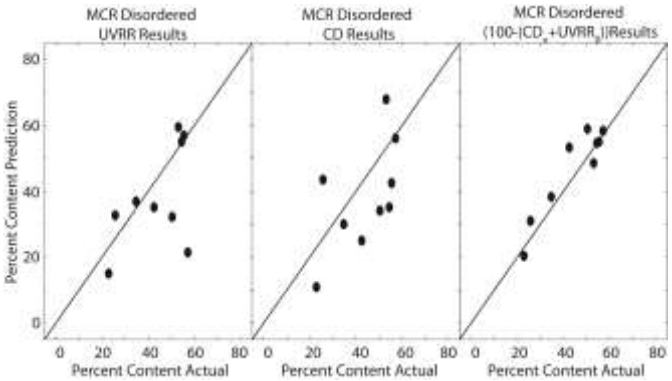



**Fig. 8** The predicted versus actual percent composition of disordered secondary structure from UVRR analysis, CD analysis, and (100 - ($CD_\alpha$ + $UVRR_\beta$)). The (1,1) line is shown to illustrate the deviations in the prediction.

## 4. Discussion and Conclusion

It is not surprising that the algorithms give the best predictions for β-sheet content using the UVRR data set, given that the β-sheet structured poly-L-lysine has the most intense UVRR spectrum and thus the greatest signal-to-noise ratio. It is interesting that the CLS algorithm predicts the β-sheet content more accurately than the other algorithms given that it is the simplest used here. However, the difference in prediction errors between all the algorithms is small.

It is also intriguing that though all the algorithms predict the β-sheet content more accurately, none of them give the smallest error for the highest β-sheet content protein in the data set (trypsinogen). In contrast, analysis of the CD data has PLS and MCR-ALS very close in prediction ability while CLS is poor comparatively. Additionally, despite the fact that all algorithms give the best predictions for the helical content as the α-helix has the largest signal in CD data, none of the algorithms predicts the highest α-helical content protein (myoglobin) with the most accuracy.

None of the multivariate methods were able to accurately predict the amount of disordered content from either UVRR or CD spectra. Combining the predicted amounts of helical and β-sheet contents enabled a more accurate estimation of the

disordered content. When the results of the two data sets are combined, the average RMSECV for PLS is slightly lower (4.2%) than CLS and MCR-ALS (5.1%).Thus, more accurate predictions of secondary structure content can be achieved when multiple techniques are employed, much as seen with CD+ IR spectroscopy[10], because of the difference in structural sensitivity of each technique. A slight improvement in RMSECV was observed when combining CD+UVRR (~5%) versus CD+IR (7.23%)[10], despite the smaller protein data set employed in the CD+UVRR analysis. The addition of the amide S and III regions that are visible in UVRR spectra but not in IR spectra, likely improved our RMSECV values.

Both CLS and MCR-ALS can be used for resolution of pure secondary structure profiles. CLS outperformed MCR-ALS when resolving pure secondary structure profiles from CD spectra of proteins. However, MCR-ALS outperformed CLS when resolving pure secondary structure profiles from UVRR spectra of proteins. This might be attributed to the application of non-negative constraints during the ALS optimization, which could not applied when analysing the CD spectra via MCR-ALS.

Multivariate techniques may be used to model a limited protein data set and predict unknown protein secondary structure content based on the model in both UVRR and CD spectroscopy, and is most accurate when both techniques are used in unison. An advantage of employing CD and UVRR is that the same sample can be used for both techniques as water does not contribute significantly to UVRR spectra. Normalization should be used with caution as it seriously degraded prediction of helical content from CD spectra.

## Acknowledgements

The authors thank Dr. Michael Henzl and Dr. Anmin Tan for help with the CD measurements. The authors also thank the NSF (Grant # CHE-1151533), University of Missouri Research Council, University of Missouri Research Board and University of Missouri Department of Chemistry for funding.

## Notes

*a University of Missouri, Department of Chemistry, Columbia, MO, USA. Fax: +1 (573) 882 2754; Tel: +1 (573) 882 8949; E-mail: jijir@missouri.edu*

## References

1. E. Herczenik and M. F. B. G. Gebbink, *The FASEB Journal*, 2008, **22**, 2115-2133.
2. A. Moglich, X. Yang, R. A. Ayers and K. Moffat, *Annual Review of Plant Biology*, 2010, **61**, 21-47.
3. D. Voet and J. G. Voet, *Biochemistry*, 3rd edn., John Wiley & Sons, Inc., Hoboken, NJ, 2004.
4. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *Journal of Computational Chemistry*, 2004, **25**, 1605-1612.
5. V. Mitaksov, S. M. Truscott, L. Lybarger, J. M. Connolly, T. H. Hansen and D. Fremont, *Chemistry and Biology*, 2007, **14**, 909-922.
6. A. Higashiura, T. Kurakane, M. Matsuda, M. Suzuki, K. Inaka, M. Sato, T. Kobayashi, T. Tanaka, H. Tanaka, K. Fujiwara and A. Nakagawa, *Acta Crystallographica Section D-Biological Crystallography*, 2010, **66**, 698-708.
7. F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein and H. Oschkinat, *Nature*, 2002, **420**, 98-102.
8. C. A. Roach, J. V. Simpson and R. D. JiJi, *Analyst*, 2011, **137**, 555-562.
9. N. J. Greenfield, *Nature Protocols*, 2006, **1**, 2876-2890.
10. S. Navea, R. Tauler, E. Goormaghtigh and A. de Juan, *Proteins*, 2006, **63**, 527-541.
11. R. Y. Yada, R. L. Jackman and S. Nakai, *International journal of peptide and protein research*, 1988, **31**, 98-108.
12. T. G. Spiro and C. A. Grygon, *Journal of Molecular Structure*, 1988, **173**, 79-90.
13. R. A. Copeland and T. G. Spiro, *Biochemistry*, 1987, **26**, 2134-2139.
14. J. T. Pelton and L. R. McLean, *Analytical Biochemistry*, 2000, **277**, 167-176.
15. C. Y. Huang, G. Balakrishnan and T. G. Spiro, *Journal of Raman Spectroscopy*, 2006, **37**, 277-282.
16. Z. Chi, X. G. Chen, J. S. W. Holtz and S. A. Asher, *Biochemistry*, 1998, **37**, 2854-2864.
17. V. A. Shashilov and I. K. Lednev, *Chemical Reviews*, 2010, **110**, 5692-5713.
18. S. Navea, R. Tauler and A. De Juan, *Analytical Chemistry*, 2006, **78**, 4768-4778.
19. S. A. Oladepo, K. Xiong, Z. Hong and S. A. Asher, *Journal of Physical Chemistry Letters*, 2011, **2**, 334-344.
20. S. A. Asher, Z. Chi and P. Li, *Journal of Raman Spectroscopy*, 1998, **29**, 927-931.
21. S. Song and S. A. Asher, *Journal of the American Chemical Society*, 1989, **111**, 4295-4305.
22. Y. Wang, R. Purrello, T. Jordan and T. G. Spiro, *Journal of the American Chemical Society*, 1991, **113**, 6359-6368.
23. Z. Ahmed and S. A. Asher, *Biochemistry*, 2006, **45**, 9068-9073.
24. N. Greenfield and G. D. Fasman, *Biochemistry*, 1969, **8**, 4108-4116.
25. N. J. Greenfield, *Analytical Biochemistry*, 1996, **235**, 1-10.
26. B. A. Wallace and R. W. Janes, *Current Opinion in Chemical Biology*, 2001, **5**, 567-571.
27. J. V. Simpson, G. Balakrishnan and R. D. Jiji, *Analyst*, 2009, **134**, 138-147.
28. J. V. Simpson, O. Oshokoya, N. Wagner, J. Liu and R. D. Jiji, *Analyst*, 2011, **136**, 1239-1247.
29. K. A. Oberg, J. M. Ruysschaert and E. Goormaghtigh, *European Journal of Biochemistry*, 2004, **271**, 2937-2948.
30. S. Navea, R. Tauler and A. D. Juan, *Analytical Biochemistry*, 2005, **336**, 231-242.
31. R. G. Brereton, *Chemometric: data analysis for the laboratory and chemical plant*, John Wiley & Sons, Inc., Hoboken, NJ, 2003.
32. P. K. R. Beebe, R. J., Seasholts, M. B., *Chemometrics: A Practical Guide*, Wiley, New York, NY, 1998.
33. E. R. Malinowski, *Factor Analysis in Chemistry*, John Wiley & Sons, New York, NY, 2002.



34. S. Wold, M. Sjöström and L. Eriksson, *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**, 109-130.
35. R. Tauler and A. de Juan, *Practical Guide to Chemometrics (Chapter 12)*, Taylor & Francis Group, LLC, 2006.
36. M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley- VCH, 2007.
37. A. de Juan and R. Tauler, *Critical Reviews in Analytical Chemistry*, 2006, **36**, 163-176.
38. K. A. Majorek, P. J. Porebski, A. Dayal, M. D. Zimmerman, K. Jablonska, A. J. Stewart, M. Chruszcz and W. Minor, *Molecular Immunology*, 2012, **52**, 174-182.
39. R. Saito, T. Sato, A. Ikai and N. Tanaka, *Acta Crystallographica Section D-Biological Crystallography*, 2004, **60**, 792-795.
40. P. E. Pjura, A. M. Lenhoff, S. A. Leonard and A. G. Gittis, *J. Mol. Biol.*, 2000, **300**, 235-239.
41. G. W. Bushnell, G. V. Louie and G. D. Brayer, *J. Mol. Biol.*, 1990, **214**, 585-595.
42. G. Wohlfahrt, S. Witt, J. Hendle, D. Schomburg, H. M. Kalisz and H. J. Hecht, *Acta Crystallographica Section D-Biological Crystallography*, 1999, **55**, 969-977.
43. R. Diamond, *J. Mol. Biol.*, 1974, **82**, 371-391.
44. H. C. Watson, *Progress in Stereochemistry*, 1969, **4**, 299.
45. P. E. Stein, A. G. Leslie, J. T. Finch and R. W. Carrell, *J. Mol. Biol.*, 1991, **221**, 941-959.
46. A. A. Kossiakoff, J. L. Chambers, L. M. Kay and R. M. Stroud, *Biochemistry*, 1977, **16**, 654-664.
47. M. Wang and R. D. Jiji, *Biophysical Chemistry*, 2011, **158**, 96-103.
48. S. D. Brown, *Appl. Spectrosc.*, 1995, **49**, 14A-31A.
49. P. Pancoska, M. Blazek and T. A. Keiderling, *Biochemistry*, 1992, **31**, 10250-10257.
50. A. Toumadje, S. W. Alcorn and W. Curtis Johnson Jr, *Analytical Biochemistry*, 1992, **200**, 321-331.
51. J. Jaumot, R. Gargallo, A. de Juan and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2005, **76**, 101-110.
52. M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi and M. Maeder, *Journal of Chemometrics*, 2006, **20**, 302-310.
53. C. A. Roach and S. L. Neal, *Applied Spectroscopy*, 2010, **64**, 1145-1153.
54. H. Abdollahi and R. Tauler, *Chemometrics and Intelligent Laboratory Systems*, 2011, **108**, 100-111.
55. C. A. Roach, *Analyst*, 2011, **136**, 2770-2774.
56. R. W. Woody, *Adv. Biophys. Chem.*, 1992, **2**, 37-79.

.