1
2
3
4 “Parallel Factor Analysis of Multi- Excitation Ultraviolet Resonance Raman Spectra for Protein
5 Secondary Structure Determination”,
6
7

8 Olayinka O. Oshokoya^a and Renee D. JiJi^a
9

10
11 ^aUniversity of Missouri, Department of Chemistry, Columbia, MO
12 601 S. College Avenue.
13 Columbia, MO 65211 USA
14
15

16
17 Author to whom correspondence should be made:

18 Renee D. JiJi

19 jjir@missouri.edu

20 573-882-8949

21 University of Missouri

22 Department of Chemistry

23 601 S. College Ave.

24 Columbia, MO 65211 USA
25
26
27

28
29 Keywords: Parallel factor analysis; Ultraviolet resonance Raman; Protein secondary structure;
30 Multiway; Trilinear.
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Graphical abstract

Abstract

Protein secondary structural analysis is important for understanding the relationship between protein structure and function, or more importantly how changes in structure relate to loss of function. The structurally sensitive protein vibrational modes (amide I, II, III and S) in deep-ultraviolet resonance Raman (DUVRR) spectra resulting from the backbone C-O and N-H vibrations make DUVRR a potentially powerful tool for studying secondary structure changes. Experimental studies reveal that the position and intensity of the four amide modes in DUVRR spectra of proteins are largely correlated with the varying fractions of α -helix, β -sheet and disordered structural content of proteins. Employing multivariate calibration methods and DUVRR spectra of globular proteins with varying structural compositions, the secondary structure of a protein with unknown structure can be predicted. A disadvantage of multivariate calibration methods is the requirement of known concentration or spectral profiles. Second-order curve resolution methods, such as parallel factor analysis (PARAFAC), do not have such a requirement due to the “second-order advantage.” An exceptional feature of DUVRR spectroscopy is that DUVRR spectra are linearly dependent on both excitation wavelength and secondary structure composition. Thus, higher order data can be created by combining protein DUVRR spectra of several proteins collected at multiple excitation wavelengths to give multi-excitation ultraviolet resonance Raman data (ME-UVRR). PARAFAC has been used to analyze ME-UVRR data of nine proteins to resolve the pure spectral, excitation and compositional profiles. A three factor model with non-negativity constraints produced three unique factors that were correlated with the relative abundance of helical, β -sheet and poly-proline II dihedral angles. This is the first empirical evidence that the typically resolved “disordered” spectrum represents the better defined poly-proline II type structure.

1 Introduction

The relationship between protein structure and function have made determination and monitoring of protein secondary structure an area of great importance in biochemical and biophysical research. This increased interest in resolving and quantifying protein secondary structure content also stems from the observation that secondary structure changes without changes in the primary structure are involved in some protein based diseases.[1-5]

Protein structure has four levels and a change at any level may result in changes in protein function. The primary structure of a protein is the amino acid sequence while the secondary structure refers to the structures within the protein that are defined by the phi (ϕ) and psi (ψ) dihedral angles of the amide backbone (Figure 1). The three dimensional arrangement of these secondary structure motifs is the tertiary structure and finally the arrangement of protein tertiary subunits to each other in larger complexes that function as a single unit is the quaternary structure.[6, 7] Because of the scope of this article, discussion would be limited to secondary structure.

Protein secondary structure is determined by the presence of specific hydrogen bonds as well as the set of dihedral angles (ϕ , ψ) which define spatial orientation of the peptide backbone. When these backbone dihedral angles have repeating values, the peptide forms regular secondary structure. Of the different secondary structures, the helical[8] and β -sheet[8, 9] structures are the most common with dihedral angles corresponding to [α - (-60°, 45°), 3_{10} - (-49°, -26°) and π - (-57°, -69°)] and [parallel (-139°, 135°) and anti-parallel (-119°, 113)] respectively. Another category in protein secondary structure is the unfolded conformation, in which each amino acid in the backbone randomly samples all sterically-allowed ϕ and ψ angles[10]. Recent evidence[11] suggests that these seemingly disordered proteins could have local regions of order that adopt

1
2
3
4 turn-like conformations or those of poly-proline II[12] (PPII, a left handed 3_1 -helix formed by
5
6 trans-L-poly-proline). Turns in a protein[13] (where the protein reverses its general direction) are
7
8 non-repetitive and occur over only a few residues but are abundant in protein structures and have
9
10 conformation flexibilities which are much greater than helices and β -sheets.
11
12

13
14 Traditional methods for protein secondary structure quantification such as x-ray
15
16 crystallography,[14, 15] nuclear magnetic resonance,[16, 17] and circular dichroism[18, 19] are
17
18 now complimented by vibrational methods like IR, conventional and resonance Raman.[20-
19
20 27] [ENREF 12 ENREF 12](#) Deep-ultraviolet resonance Raman (DUVRR) particularly, has
21
22 proven useful for quantification due to the structural sensitivity of the observed backbone amide
23
24 modes. Some advantages of DUVRR include increased signal intensity (10^2 - 10^6 times) versus
25
26 conventional Raman spectroscopy, minimal contribution from solvent water bands, elimination
27
28 of background fluorescence and selective enhancement of the peptide backbone modes derived
29
30 from the various vibrations of the backbone amide group (-CO-NH-) as a result of the π_2 to π_3^*
31
32 dipole- allowed transition.[26, 28-32] DUVRR sensitivity to protein secondary structure is
33
34 observed in the shifting and intensity changes of the four observable amide modes.[33-38] The
35
36 position and intensity of the four amide modes: amide I, II, III and S are dependent upon the
37
38 secondary structure of the protein with their relative contributions being proportional to the
39
40 relative amount of each secondary structure conformation. Similarly, each secondary structure
41
42 type exhibits distinct absorption profiles leading to a compositional dependence in resonance
43
44 enhancement versus excitation wavelength.[29] Therefore, the position and intensity of amide
45
46 modes change with varying excitation wavelengths and secondary structure composition.[39]
47
48
49
50
51
52
53
54

55 Initially, quantification studies of protein secondary structure employing DUVRR
56
57 focused on univariate calibration methods of single amide modes,[26] but have evolved to
58
59
60
61
62
63
64
65

1
2
3
4 include multivariate calibration and multivariate curve resolution methods and other advanced
5
6 statistical analyses of all observable amide modes.[22, 35, 36, 40-45] In these methods, first
7
8 order data are decomposed using bilinear models showing that individual protein spectra, $\mathbf{x}$, are a
9
10 linear combination of the different underlying pure secondary structure motifs, $\mathbf{s}$, and their
11
12 respective fractional amounts, c (Equation 1);
13
14

$$\mathbf{x} = c_{\alpha}\mathbf{s}_{\alpha} + c_{\beta}\mathbf{s}_{\beta} + \dots \quad (1)$$

15
16
17 where α designates α -helical and β designates β -sheet related variables. Other structures that
18
19 have been defined include α -L, turns, random coil or disordered and 3_{10} helices.
20
21

22
23 Spectra from different proteins with different secondary structure compositions can be combined
24
25 into a matrix and decomposed to give a matrix of known secondary structural content of each
26
27 protein and a matrix of underlying pure secondary. The pure underlying secondary structure
28
29 spectra may then be calculated via multivariate least square regression which in turn can be used
30
31 to determine the structural content of a new protein.
32
33
34

35
36 The “known” concentration profiles are typically determined from proteins whose atomic
37
38 level structure has been characterized by x-ray crystallography or NMR spectroscopy and
39
40 published in the protein database, PDB (www.rcsb.org)[46]. However, the conditions under
41
42 which the structures were determined do not necessarily represent the conditions under which the
43
44 spectroscopic measurements are being conducted. This is especially true of crystal structures,
45
46 which are static in nature. “Known” spectral profiles can be estimated from model
47
48 polypeptides, such as poly-L-lysine or poly-L-glutamic acid, that adopt “pure” secondary
49
50 structures with varying environments.[40] Given that these pre-determined or “known” inputs
51
52 are not taken from the proteins themselves in their relevant environments, they can themselves
53
54 introduce errors or biases into the multivariate analysis.
55
56
57
58
59
60
61
62
63
64
65

Trilinear methods are not prone to the rotational ambiguity of bilinear methods and reduce or eliminate the need for good initial estimates of either the **C** or **S** matrices. Multi-excitation UVRR (ME-UVRR) spectra of multiple proteins results in trilinear data that can be modeled with higher order algorithms such as parallel factor analysis (PARAFAC). This is particularly important for proteins whose disordered regions are not well characterized by NMR or x-ray crystallography but appear to have increasingly important functional significance.[47-51] Using a combination of ME-UVRR of nine globular proteins and PARAFAC, the pure spectral profiles of folded, helical and β -sheet structures are resolved along with the spectral profiles corresponding to the disordered portions of the proteins. Next, a univariate regression was performed for the secondary structures by relating the loadings to the relative abundance of helical (α - +₃₁₀), β -sheet and poly-proline II type (PPII) structure.

2 Materials and methods

2.1 Sample preparation

Nine proteins with varying secondary structure content (Table 1), amino acids L-phenylalanine (F) and L-tyrosine (Y) were obtained from Sigma Aldrich (St Louis, MO) and used without further purification. The proteins and amino acids were dissolved in 10 mM phosphate buffer (pH 7.2). Protein and aromatic amino acid concentrations were determined by UV-Visible absorption using a Hewlett Packard 8453 spectrometer (Palo Alto, CA), and were 0.5 mg·ml⁻¹ for protein solutions and 200 μ M for amino acid solutions. The selected proteins are readily soluble in aqueous solution, have a well-distributed combination of the major secondary structures and are relatively inexpensive, making them an ideal set of calibration proteins.

2.2 Deep-UV resonance Raman (DUVRR) spectroscopy

The DUVRR instrument used to collect protein spectra has been previously described. [52] Briefly, the fourth harmonic of a tunable Ti:Sapphire laser (Coherent Inc., Santa Clara, CA) was employed to generate excitation wavelengths from 197 to 205 nm in 1 nm increments. The sample was circulated through two nitinol wires (Small Parts Inc., Miramar, FL) to create a thin film under a nitrogen purge to remove ambient oxygen. The temperature of the sample was held at 22°C in a water-jacketed reservoir (Mid Rivers Glassblowing, Saint Charles, MO) using a bath recirculator (Isotemp 3016D, Fisher Scientific, Pittsburgh, PA). Raman scattering was 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, which was controlled by Synerjy software (Horiba Jobin Yvon Inc., Edison, NJ). Each spectrum was the sum of 1 hour of signal collection.

2.3 Data preprocessing and analysis

All computations were performed in the Matlab environment (Mathworks, Natick, MA). Raw spectra were preprocessed to remove cosmic rays using an in-house program. Sample spectra collected at each excitation wavelength were averaged and base-lined using the Matlab curve fitting toolbox. Spectra collected at different excitation wavelengths have different resolutions.[44] Each spectrum was interpolated to match the number of variables in the 197 nm spectra. Phosphate buffer spectra were subtracted from each protein spectra. Cyclohexane was used as a calibration standard and perchlorate was used as an internal intensity standard.

The aromatic amino acid contributions (phenylalanine and tyrosine) were quantitatively subtracted as described by Oshokoya et al.[45] Spectra of phenylalanine and tyrosine were subtracted from corresponding protein spectra collected at the same excitation wavelength.

Spectra were then truncated so that only the four amide regions (1200-1750 cm⁻¹) were used for calibration.

2.4 Calculation of secondary structure content

Protein coordinate files for the nine proteins were downloaded from the protein data bank, PDB (www.rscb.org)[46] and a dihedral angle calculator readily available online (<http://cib.cf.ocha.ac.jp/bitool/DIHED2/>)[53] was used to determine the relative abundance of the (ϕ , ψ) dihedral angles in each protein for secondary structure content distribution as displayed in Table 1. The overall distribution of (ϕ , ψ) angles are shown in Figure 2 along with the structural assignments.

2.5 Parallel factor analysis of ME-UVRR data

The parallel factor analysis (PARAFAC) model, in the chemical sciences, is used to decompose trilinear multi-way data arrays and facilitate the identification and quantification of independent underlying components without prior knowledge of sample composition.[54-58] ME-UVRR spectra of a single protein can be referred to as second-order or bilinear data. Trilinear DUVRR data is generated when the ME-UVRR spectra of several proteins with varying structural compositions are combined into a single data array (Figure 3). PARAFAC analysis of ME-UVRR data was first applied to the amide I region of ME-UVRR protein spectra (Simpson, et al. Analyst 2011)[44]. However, quantification of disordered regions was not possible.

PARAFAC is based on the trilinear model[54], as shown in the Equation (2)

$$r_{ijk} = \sum_{n=1}^N \hat{x}_{i,n} \hat{y}_{j,n} \hat{z}_{k,n} + e_{i,j,k} \quad (2)$$

In Equation 2, r_{ijk} represents the measured response of k^{th} protein at the i^{th} excitation wavelength and j^{th} wavenumber. In this study, the trilinear dataset was created by combining DUVRR

spectra of protein samples with varying secondary structure composition collected at nine different excitation wavelengths. Therefore, each k^{th} slice of the trilinear data cube, represents ME-UVRR spectra of a series of proteins and has the dimension $I \times J$, where I is a product of the spectral range and resolution in cm^{-1} and J is the number of excitation wavelengths. The number of components, N , chosen for the system is defined by the user and determination of the correct number of components was based on split half analysis [59, 60] of the data and also by carrying out core consistency diagnostics [61] of the ME-UVRR data. The PARAFAC algorithm used in this work has been described previously [62].

The data was arranged in three-way arrays with, 211 spectral points in cm^{-1} , in the rows or in the X-dimension, 9 excitation wavelengths in nm, in the columns or in the Y-dimension and 9 protein samples in the slices or in the Z-dimension resulting in a $211 \times 9 \times 9$ data cube (Figure 3). Although the orientation of the data array should not have an impact on the final solution for an ideal data set, it was found that the spectral information needed to be in either the X or Y dimension. This was because the PARAFAC algorithm initialized its operation using the information in the X and Y dimensions of the three-way ME-UVRR data. For the ME-UVRR data used, most of the variance in the data exist in the spectral and sample information while less variance occurs in the excitation dimension. We believe that initializing with the spectra and sample information in either the X or Y dimensions of the datacube leads to convergence to a more realistic steady state solution. Also, the use of non-negativity on the spectral dimension affects the convergence of the algorithm. It should be noted that the maximum number of iterations was never reached for convergence for any of the permutations used.

To evaluate the effect of data permutations on the resolved profiles, a series of models were created and analyzed. Four different permutations of the ME-UVRR data were created by

switching the dimensions in which the spectral, excitation wavelength and sample information occur. For instance, when spectral information, excitation wavelength and sample information are placed in the X-, Y-, and Z-dimensions respectively, we get data permutation XspYexZsa.

Table 2 includes a full list of model permutations.

When using PARAFAC, an initial definition of the number of factors is necessary. It is also possible to apply constraints[63] such as non-negativity, unimodality or closure. In this work, only non-negativity was employed as a constraint. PARAFAC models of the data array were developed using 2-, 3-, 4-, and 5-components and split-half analysis, core consistency and percent root mean squared error of calibration (%RMSEC, Equation 3) were used to select the number of factors:

$$\%RMSEC = \sqrt{\frac{\sum(\hat{c}_i - c_i)^2}{I}} \times 100 \quad (3)$$

where $\hat{c}_i$ is the predicted secondary structure concentration of the i th sample, c_i is the experimental secondary structure composition as determined from the (ϕ, ψ) dihedral angles of the i th sample and I is the number of samples.

PARAFAC analysis of the ME-UVRR data yields three loading matrices one of which corresponds to the underlying “pure” secondary structure Raman spectra (PSSR), the other, corresponding to the pure relative cross section of each spectral component of the different secondary structures present. The third loading matrix contains the relative secondary structure compositions of the proteins used; in the calibration step, these loadings are regressed against the experimental secondary structure compositions of the different proteins to get a linear calibration line. In the prediction step, this regression line is then used for prediction of protein secondary structure.

2.6 Determining the number of factors

2.6.1 Core consistency

Core consistency diagnostic can be used for the determination of the proper number of components for multi-way models and applies especially to the PARAFAC model. A detailed explanation of the principle and theory behind core consistency diagnostics has been described by Bro and Kiers [61].

The structural model for a three-way PARAFAC model can be stated as

$$\mathbf{R} = \mathbf{X}(\mathbf{Y}|\otimes|\mathbf{Z})^T \quad (4)$$

It may also be stated as a restricted Tucker3 model [64-67]

$$\mathbf{R} = \mathbf{X}\mathbf{T}^{(F \times FF)}(\mathbf{Z} \otimes \mathbf{Y})^T \quad (5)$$

where the core array $\mathbf{T}$ is a binary array with zeros in all places except for the superdiagonal which contains only ones and F is the number of components. After having fitted the PARAFAC model $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, verification that the trilinear structure is appropriate can be obtained by calculating the least-squares Tucker3 core given $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, according to

$$\mathbf{R} - \mathbf{X}\mathbf{G}(\mathbf{Y} \otimes \mathbf{Z})^T \quad (6)$$

A regression model can be constructed to solve the problem based on the loss of function

$$\text{vec}\mathbf{R} - (\mathbf{Z} \otimes \mathbf{Y} \otimes \mathbf{X})\text{vec}\mathbf{G} \quad (7)$$

If the PARAFAC model is valid then $\mathbf{G}$ should resemble $\mathbf{T}$, if the data can't approximately be described by a trilinear model or too many components are used then the core $\mathbf{G}$, will differ from $\mathbf{T}$. A simple way to assess if the model structure is reasonable is therefore to monitor the distribution of superdiagonal and off-diagonal elements of $\mathbf{G}$. if the superdiagonal elements are all close to one, and the off-superdiagonal elements are close to zero the model is not overfitting. If, on the other hand, this is not the case, then too many components have been used. It is

possible to calculate the superidentity of $\underline{\mathbf{G}}$ to obtain a single parameter for the model quality which is the core consistency (Equation 8). This is defined as the percentage of the alikeness expressed by one minus the sum of the squares of the elements of $\underline{\mathbf{G}} - \underline{\mathbf{T}}$ divided by the sum of the squares of the elements of $\underline{\mathbf{T}}$ and it reflects how well the Tucker3 core fits to the assumptions of the model.

$$\text{Core Consistency} = 100 \left(\frac{1 - \sum_{d=1}^F \sum_{e=1}^F \sum_{f=1}^F (g_{def} - t_{def})^2}{\sum_{d=1}^F \sum_{e=1}^F \sum_{f=1}^F (t_{def}^2)} \right) \quad (8)$$

When a sequence of models is run with an increasing number of components, the core consistency tends to start high, near 100%, then drop abruptly at the point when too many components are selected. The number of components is determined to equal the number in the largest model still having a high core consistency [61, 64].

The core consistency diagnostic was implemented using the *N*-way Toolbox for Matlab [68]. Figure 4 shows the core consistency plots of a 2- (A), 3- (B), 4- (C), and 5-component (D) PARAFAC model of the ME-UVRR protein data. In the plots, the circles are the superdiagonal elements of $\underline{\mathbf{G}}$ and hence should ideally be one. The squares are the off-superdiagonal elements which should ideally be zero. The line segment is made from the elements of $\underline{\mathbf{T}}$ and is hence the target that $\underline{\mathbf{G}}$ should resemble. In Figure 4, the core consistency plot of the two-component model of the ME-UVRR data is perfect (100%), the core elements on the superdiagonal are close to one and all off-diagonal elements are zero. The core consistency of the three-component model (76.6%) while not perfect is not an abrupt drop from the perfect core consistency of the two-component model and so is still considered appropriate. The abrupt drop in core consistency occurs when more than three components are used therefore making three components most suitable for the ME-UVRR data.

2.6.2 Split-half analysis

To further support the results from the core consistency diagnostics for the proper number of components to be used, split-half analysis[69, 70] of the data was performed. The split-half analysis is a type of jack-knife approach where different subsets of the data are analyzed independently. Due to the uniqueness of the PARAFAC model, the same result or loadings will be obtained in the non-split modes from models of any suitable subsets of the data, if the correct number of components is chosen. If too many or too few components are chosen, the model parameters will differ if the model is fitted to different data sets.

The 3-way ME-UVRR data was divided into four groups: A, B, C and D with group A consisting of the first quarter of the samples, group B, the second quarter of the samples, group C, the third quarter of the samples and group D, the last three of the samples. Halves were then created by making all possible combinations of the quarters i.e. AB, AC, AD, BC, BD and CD. A model was then fitted to each of the datasets using different number of components and the average sum of the squares of the residuals were compared. Figure 5 shows a comparison of the average sums of the squares of the residuals (SSR) of the PARAFAC models using 2-, 3-, 4- and 5- components. The error bars signifies the standard deviations of the averaged values. Very high averages and standard deviations of the SSR shows that 4- and 5-components are not suitable for modelling the ME-UVRR data (Figure 5). Using 3-components gives the lowest average SSR (2.38×10^7) and lowest standard deviation (9.59×10^7) therefore making it the suitable number of components.

3 Results and Discussion

Based on the results of the core consistency diagnostics and split-half analysis, it was concluded that the ME-UVRR data should be modeled with 3-components. In addition to β -sheet/strand and helical (α -helical and 3_{10} helical) structures, proteins adopt several other types of secondary structures including PPII, α -L and turn structures. The structural definitions in Table 1 are based on the distribution of (ϕ , ψ) dihedral angles around the ideal as shown in Figure 2. Turn structures are not included as turn structures have non-repetitive (ϕ , ψ) dihedral angles.

On average, prediction of β -sheet secondary structure showed the lowest %RMSEC for the 3-component PARAFAC models (Figure 6). Irrespective of data orientation, %RMSEC for β -sheet structure prediction was 3.4% for unconstrained 3-component models (Figure 6).

Application of non-negativity constraints to the models resulted in a modest increase in %RMSEC for β -sheet prediction

Percent RMSEC's for helical secondary structure were generally higher than those for β -sheet structure (Figure 6). Before application of non-negativity constraints, all models showed a lower %RMSEC (14.5%) and little increase was observed in %RMSEC when non-negativity constraints were applied with the exception of model XexYspZsa which had a %RMSEC of over 60%.

The third factor in the 3-component PARAFAC models showed very strong correlation to PPII type structure composition. For the unconstrained models, the %RMSEC for PPII type structure ranged from 17 to 22%, when non-negativity constraints were applied to the 3-component models, the %RMSEC decreased to 3.2-6.2% with model XspYsaZex having the lowest %RMSEC (3.2%). While the application of non-negativity constraints slightly increases

the %RMSEC for helical and β -sheet secondary structures in 3-component models, there is a significant decrease in the %RMSEC for PPII-type structure.

Examination of the resolved spectral profiles for each of the 3-component models provides further insight into the optimal model. The resolved spectral profiles for each model without non-negativity constraints were essentially identical (Figure 7A). The resolved helical spectrum from Figure 7A has a strong amide S band at 1390 cm^{-1} , an amide III feature at $1230\text{--}1240\text{ cm}^{-1}$, and an amide I band at 1667 cm^{-1} all of which are associated with non-helical structures. The higher %RMSEC of PPII-type structure when no non-negativity constraints are applied may be attributed to the incomplete resolution of helical and PPII-type structures. Application of non-negativity constraints to 3-component models with the samples in the Z-dimension results in the absence of an amide I feature for PPII-type structure, which is deemed unreasonable for protein DUVRR spectra (Figure 7C and 7D). Thus, the most reasonable model is produced when a 3-component model is employed with non-negativity constraints and excitation information is placed in the Z-dimension (Figure 7B).

The predicted helical spectrum from the PARAFAC analysis of the 3-factor model with excitation in the Z-dimension (Figure 8) shows an absence of the amide S ($1374\text{--}1406\text{ cm}^{-1}$) mode and two features in the amide III region (1297 cm^{-1} and 1336 cm^{-1}) as described in the literature.[33, 35, 71] [ENREF 27](#) The disappearance of the amide S mode is a key markers of helical secondary structure. Also, the positions of the amide I (1655 cm^{-1}), and II (1548 cm^{-1}) bands in the resolved secondary structure spectra from PARAFAC are within the expected range for helical proteins recorded in the literature [33, 35, 71]. For the β -sheet resolved pure spectral profile, the predicted amide I (1675 cm^{-1}), II (1556 cm^{-1}), S (1395 cm^{-1}) and III (1235 cm^{-1}) positions all fall within the expected range for β -sheet structures, which occur at $1670\text{--}1675\text{ cm}^{-1}$,

1
2
3
4 1550-1564 cm^{-1} , 1395-1406 cm^{-1} , and 1220-1241 cm^{-1} respectively.[37, 71] The amide I and II
5
6 of the disordered resolved spectral profiles appear at 1675 cm^{-1} and 1566 cm^{-1} falling within the
7
8 expected regions of 1660-1682 cm^{-1} and 1548-1561 cm^{-1} respectively.[35, 71] The amide III and
9
10 S modes are slightly upshifted (1250 cm^{-1}) and downshifted (1393 cm^{-1}) from the β -sheet
11
12 spectrum as expected for PPII-type structure.
13
14

15
16 The resolved secondary structure components from the 3-component models with
17
18 excitation in the Z- dimension are essentially identical and show good correlation with reported
19
20 helical, β -sheet and PPII-type contents for the sample proteins. A linear regression model was
21
22 built using the resolved intensity for the different secondary structure components and the
23
24 calculated secondary structure compositions given in Table 1 with the exception of ovalbumin.
25
26 Inclusion of ovalbumin significantly skewed all of the regression models for one or more
27
28 secondary structures. For the optimal 3-component model XspYsaZex with non-negativity
29
30 constraints, the %RMSEC decreased from 12% to 5% for β -sheet structure after removing
31
32 ovalbumin. The resolved intensities of the secondary structures obtained from the PARAFAC
33
34 analysis were then used to re-predict the secondary structure content using the linear regression
35
36 model (Figure 9).
37
38
39
40
41
42

43 The excitation profiles of the secondary structures obtained from the non-negativity
44
45 constrained 3-component PARAFAC model (Figure 10) showed strong resemblance to those
46
47 previously published in the literature[29, 44] especially for β -sheet which has a single maxima at
48
49 approximately 200 nm. The helical excitation profile has two maxima appearing at
50
51 approximately 198 nm and 204 nm. The PPII-type structure excitation profile has a maxima
52
53 below 197 nm and a shoulder at approximately 202 nm.
54
55
56
57
58
59
60
61
62
63
64
65

4 Conclusion

PARAFAC analysis was carried out on ME-UVRR data of a set of nine proteins with varying secondary structure compositions. It was established that the orientation of the three-way data, especially with excitation wavelength information in the Z-dimension, played a great role in the ability of the PARAFAC algorithm to resolve the underlying spectral profiles. A 3-component model resulted in the best resolution of helical, β -sheet and PPII-type structures. This is the first experimental evidence indicating that the third component typically resolved with multivariate methods represents PPII- type structure as calculated from (ϕ , ψ) dihedral angles. The prediction of PPII-type structure was significantly improved with a combination of three-way data, non-negativity constraints and PARAFAC analysis. The encouraging results of this exploratory analysis thus suggest that higher order data can be used for quantification of protein secondary structure without prior knowledge of the underlying spectral or compositional profiles.

Acknowledgements

The authors 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.

References

- [1] A. Moglich, X. Yang, R.A. Ayers, K. Moffat, Structure and function of plant photoreceptors, *Annual review of plant biology*, 61 (2010) 21-47.
- [2] E. Herczenik, M.F.B.G. Gebbink, Molecular and cellular aspects of protein misfolding and disease, *FASEB Journal*, 22 (2008) 2115-2133.
- [3] C. Weissmann, The state of the prion, *Nature Reviews Microbiology*, 2 (2004) 861-871.
- [4] S.B. Prusiner, Prions, *Proceedings of the National Academy of Sciences*, 95 (1998) 13363-13383.
- [5] C.C. Blake, M.J. Geisow, S.J. Oatley, B. Rerat, C. Rerat, Structure of prealbumin: secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å, *Journal of molecular biology*, 121 (1978) 339-356.
- [6] D. Voet, J.G. Voet, *Biochemistry*, 3rd ed., John Wiley & Sons, Inc., Hoboken, NJ, 2004.
- [7] J.T. Edsall, P.J. Flory, J.C. Kendrew, A.M. Liquori, G. Nemethy, G.N. Ramachandran, H.A. Scheraga, A proposal of standard conventions and nomenclature for the description of polypeptide conformation, *Journal of Biological Chemistry*, 241 (1966) 1004-1008.
- [8] L. Pauling, R.B. Corey, H.R. Branson, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain, *Proceedings of the National Academy of Sciences of the United States of America*, 37 (1951) 205-211.
- [9] F.R. Salemme, D.W. Weatherford, Conformational and geometrical properties of β -sheets in proteins. I. Parallel β -sheets, *Journal of molecular biology*, 146 (1981) 101-117.
- [10] D.A. Brant, P.J. Flory, The Configuration of Random Polypeptide Chains. II. Theory, *Journal of the American Chemical Society*, 87 (1965) 2791-2800.
- [11] R. Schweitzer-Stenner, Conformational propensities and residual structures in unfolded peptides and proteins, *Molecular BioSystems*, 8 (2012) 122-133.
- [12] P.M. Cowan, S. McGavin, Structure of Poly-L-Proline, *Nature*, 176 (1955) 501-503.
- [13] E.G. Hutchinson, J.M. Thornton, A revised set of potentials for beta-turn formation in proteins, *Protein Science : A Publication of the Protein Society*, 3 (1994) 2207-2216.
- [14] A. Higashiura, K. Ohta, M. Masaki, M. Sato, K. Inaka, H. Tanaka, A. Nakagawa, High-resolution X-ray crystal structure of bovine H-protein using the high-pressure cryocooling method, *Journal of Synchrotron Radiation*, 20 (2013) 989-993.
- [15] G. Blucz, M. Miller, R. Harrison, N. Thanki, G.L. Gilliland, C.M. Ogata, S.H. Kim, A. Wlodawer, Structure of monellin refined to 2.3 Å resolution in the orthorhombic crystal form, *Acta Crystallographica Section D: Biological Crystallography*, 53 (1997) 713-719.
- [16] F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein, H. Oschkinat, Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy, *Nature*, 420 (2002) 98-102.
- [17] D.S. Wishart, B.D. Sykes, F.M. Richards, An improved synthesis of α - ^{13}C glycine and heteronuclear NMR studies of its incorporation into thioredoxin, *Journal of Labelled Compounds and Radiopharmaceuticals*, 31 (1992) 1019-1028.
- [18] N.J. Greenfield, Using circular dichroism spectra to estimate protein secondary structure, *Nature Protocols*, 1 (2006) 2876-2890.
- [19] N.J. Greenfield, G.D. Fasman, Computed circular dichroism spectra for the evaluation of protein conformation, *Biochemistry*, 8 (1969) 4108-4116.
- [20] A. Dong, P. Huang, W.S. Caughey, Protein secondary structures in water from second-derivative amide I infrared spectra, *Biochemistry*, 29 (1990) 3303-3308.

- [21] W.K. Surewicz, H.H. Mantsch, D. Chapman, Determination of protein secondary structure by Fourier transform infrared spectroscopy: A critical assessment, *Biochemistry*, 32 (1993) 389-394.
- [22] C.A. Roach, J.V. Simpson, R.D. Jiji, Evolution of quantitative methods in protein secondary structure determination via deep-ultraviolet resonance Raman spectroscopy, *Analyst*, 137 (2012) 555-562.
- [23] S. Navea, R. Tauler, E. Goormaghtigh, A. de Juan, Chemometric tools for classification and elucidation of protein secondary structure from infrared and circular dichroism spectroscopic measurements, *Proteins*, 63 (2006) 527-541.
- [24] R.Y. Yada, R.L. Jackman, S. Nakai, Secondary structure prediction and determination of proteins--a review, *International journal of peptide and protein research*, 31 (1988) 98-108.
- [25] T.G. Spiro, C.A. Grygon, Applications of ultraviolet resonance raman spectroscopy to proteins, *Journal of Molecular Structure*, 173 (1988) 79-90.
- [26] R.A. Copeland, T.G. Spiro, Secondary structure determination in proteins from deep (192-223-nm) ultraviolet Raman spectroscopy, *Biochemistry*, 26 (1987) 2134-2139.
- [27] J.T. Pelton, L.R. McLean, Spectroscopic methods for analysis of protein secondary structure, *Analytical Biochemistry*, 277 (2000) 167-176.
- [28] A.V. Mikhonin, N.S. Myshakina, S.V. Bykov, S.A. Asher, UV resonance Raman determination of polyproline II, extended 2.5 1-helix, and β -sheet ψ angle energy landscape in poly-L-lysine and poly-L-glutamic acid, *Journal of the American Chemical Society*, 127 (2005) 7712-7720.
- [29] K. Rosenheck, P. Doty, The far ultraviolet absorption spectra of polypeptide and protein solutions and their dependence on conformation, *Proceedings of the National Academy of Sciences of the United States of America*, 47 (1961) 1775-1785.
- [30] S.A. Asher, Z. Chi, P. Li, Resonance Raman Examination of the Two Lowest Amide $\pi\pi^*$ Excited States, *Journal of Raman Spectroscopy*, 29 (1998) 927-931.
- [31] S. Song, S.A. Asher, UV resonance Raman studies of peptide conformation in poly(L-lysine), poly(L-glutamic acid), and model complexes: The basis for protein secondary structure determinations, *Journal of the American Chemical Society*, 111 (1989) 4295-4305.
- [32] S.A. Oladepo, K. Xiong, Z. Hong, S.A. Asher, Elucidating peptide and protein structure and dynamics: UV resonance raman spectroscopy, *Journal of Physical Chemistry Letters*, 2 (2011) 334-344.
- [33] Y. Wang, R. Purrello, T. Jordan, T.G. Spiro, UVRR spectroscopy of the peptide bond. 1. Amide s, a nonhelical structure marker, is a C α H bending mode, *Journal of the American Chemical Society*, 113 (1991) 6359-6368.
- [34] V.A. Shashilov, V. Sikirzhytski, L.A. Popova, I.K. Lednev, Quantitative methods for structural characterization of proteins based on deep UV resonance Raman spectroscopy, *Methods*, 52 (2010) 23-37.
- [35] C.Y. Huang, G. Balakrishnan, T.G. Spiro, Protein secondary structure from deep-UV resonance Raman spectroscopy, *Journal of Raman Spectroscopy*, 37 (2006) 277-282.
- [36] J.V. Simpson, G. Balakrishnan, R.D. Jiji, MCR-ALS analysis of two-way UV resonance Raman spectra to resolve discrete protein secondary structural motifs, *Analyst*, 134 (2009) 138-147.

- [37] J.C. Austin, T. Jordan, T.G. Spiro, Ultraviolet resonance Raman studies of proteins and related model compounds, *Adv. Spectrosc.* (Chichester, U. K.), 20 (1993) 55-127.
- [38] S.A. Asher, A.V. Mikhonin, S. Bykov, UV Raman demonstrates that α -helical polyalanine peptides melt to polyproline II conformations, *Journal of the American Chemical Society*, 126 (2004) 8433-8440.
- [39] R.A. Copeland, T.G. Spiro, Ultraviolet Raman hypochromism of the tropomyosin amide modes: a new method for estimating α -helical content in proteins, *J. Am. Chem. Soc.*, 108 (1986) 1281-1285.
- [40] Z. Chi, X.G. Chen, J.S.W. Holtz, S.A. Asher, Uv resonance raman-selective amide vibrational enhancement: Quantitative methodology for determining protein secondary structure, *Biochemistry*, 37 (1998) 2854-2864.
- [41] Z. Chi, S.A. Asher, UV resonance determination of protein acid denaturation: Selective unfolding of helical segments of horse myoglobin, *Biochemistry*, 37 (1998) 2865-2872.
- [42] I.K. Lednev, A.S. Karnoup, M.C. Sparrow, S.A. Asher, Nanosecond UV Resonance Raman Examination of Initial Steps in α -Helix Secondary Structure Evolution, *J. Am. Chem. Soc.*, 121 (1999) 4076-4077.
- [43] A. Ozdemir, I.K. Lednev, S.A. Asher, Comparison between UV Raman and circular dichroism detection of short α helices in bombolitin III, *Biochemistry*, 41 (2002) 1893-1896.
- [44] J.V. Simpson, O. Oshokoya, N. Wagner, J. Liu, R.D. Jiji, Pre-processing of ultraviolet resonance Raman spectra, *Analyst*, 136 (2011) 1239-1247.
- [45] O.O. Oshokoya, C.A. Roach, R.D. Jiji, Quantification of protein secondary structure content by multivariate analysis of deep-ultraviolet resonance Raman and circular dichroism spectroscopies, *Anal. Methods*, 6 (2014) 1691-1699.
- [46] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Res.*, 28 (2000) 235-242.
- [47] D. Kovacs, B. Szabo, R. Pancsa, P. Tompa, Intrinsically disordered proteins undergo and assist folding transitions in the proteome, *Arch. Biochem. Biophys.*, 531 (2013) 80-89.
- [48] P. Tompa, Intrinsically disordered proteins: A 10-year recap, *Trends Biochem. Sci.*, 37 (2012) 509-516.
- [49] V.N. Uversky, Introduction to intrinsically disordered proteins (IDPs), *Chemical Reviews*, 114 (2014) 6557-6560.
- [50] R. Van Der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of intrinsically disordered regions and proteins, *Chemical Reviews*, 114 (2014) 6589-6631.
- [51] M.P. Williamson, J.R. Potts, Intrinsically disordered proteins: Administration not executive, *Biochem. Soc. Trans.*, 40 (2012) 945-949.
- [52] M. Wang, R.D. Jiji, Resolution of localized small molecule-A β interactions by deep-ultraviolet resonance Raman spectroscopy, *Biophysical Chemistry*, 158 (2011) 96-103.
- [53] Andres Liljas, Lars Liljas, Jure Piskur, Göran Lindblom, Poul Nissen, M. Kjeldgaard, *TEXTBOOK OF STRUCTURAL BIOLOGY*, World Scientific 2009.
- [54] R. Bro, PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, 38 (1997) 149-171.

- [55] M.A.B. Levi, I.S. Scarminio, R.J. Poppi, M.G. Trevisan, Three-way chemometric method study and UV-Vis absorbance for the study of simultaneous degradation of anthocyanins in flowers of the *Hibiscus rosa-sinensis* species, *Talanta*, 62 (2004) 299-305.
- [56] R.D. Jiji, G.G. Andersson, K.S. Booksh, Application of PARAFAC for calibration with excitation - Emission matrix fluorescence spectra of three classes of environmental pollutants, *Journal of Chemometrics*, 14 (2000) 171-185.
- [57] R.A. Harshman, {Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis}, *UCLA Working Papers in Phonetics*, 16 (1970) 84.
- [58] R.A. Harshman, M.E. Lundy, PARAFAC: Parallel factor analysis, *Comput. Stat. Data Anal.*, 18 (1994) 39-72.
- [59] R.A. Harshman, W.S. DeSarbo, in *Research methods for multimode data analysis*, ed. H. G. Law, J. C. W. Snyder, J. Hattie, R. P. McDonald,, Praeger, New York, 1984, pp. 602-642.
- [60] R.A. Harshman, in *Research methods for multimode data analysis*, ed. H. G. Law, J. C. W. Snyder, J. Hattie, R. P. McDonald,, Praeger, New York, 1984, pp. 566-591.
- [61] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, 17 (2003) 274-286.
- [62] R.D. Jiji, K.S. Booksh, Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra, *Analytical Chemistry*, 72 (2000) 718-725.
- [63] M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley- VCH2007.
- [64] R. Bro, *Multi-way analysis in the food industry. Models, algorithms, and applications*, PhD Thesis,, University of Amsterdam, 1998
- [65] R.A. Harshman, M.E. Lundy, The PARAFAC model for three-way factor analysis and multidimensional scaling, *Research Methods for Multimode Data Analysis*, (1984) 122-215.
- [66] A.K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. Medium-rank second-order calibration with restricted Tucker models, *Analytical Chemistry*, 66 (1994) 3345-3351.
- [67] A.K. Smilde, Y. Wang, B.R. Kowalski, Theory of medium-rank second-order calibration with restricted-Tucker models, *J. Chemom.*, 8 (1994) 21-36.
- [68] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB, *Chemometrics and Intelligent Laboratory Systems*, 52 (2000) 1-4.
- [69] R.A. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multimodal factor analysis, *UCLA Working Papers in Phonetics*, 16 (1970) 1-84.
- [70] R.A. Harshman, W.S. De Sarbo, An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques, *Research Methods for Multimode Data Analysis*, (1984) 602-642.
- [71] R.D. Jiji, G. Balakrishnan, Y. Hu, T.G. Spiro, Intermediacy of Poly(L-proline) II and β -Strand Conformations in Poly(L-lysine) β -Sheet Formation Probed by Temperature-Jump/UV Resonance Raman Spectroscopy, *Biochemistry*, 45 (2006) 34-41.

Table 1: Secondary structure content (%) of proteins used calculated from (ϕ , ψ) dihedral angles as found on the Research Collaboratory for Structural Bioinformatics (RSCB) Protein Data Bank. Helices include both α -helical and 3_{10} -helical (ϕ , ψ) dihedral angles.

Protein	Abbreviation	Helices	Sheet/Strand	Total disordered			
				PPII	Unordered	α -L	Undefined
Bovine serum albumin	BSA	80.0	5.6	8.5	4.4	0.5	1.0
Carbonic anhydrase	CAH	27.2	37.5	21.2	8.4	3.3	2.3
Chymotrypsinogen A	CTG	26.0	33.3	23.5	7.4	5.3	4.5
Cytochrome c	CYC	52.0	15.7	18.6	4.9	6.9	2.0
Glucose oxidase	UOX	44.3	25.9	15.5	5.9	5.5	2.9
Lysozyme	LSZ	50.4	14.2	12.6	7.9	8.7	6.3
Myoglobin	MBN	85.4	5.3	2.6	2.0	2.6	2.0
Ovalbumin	OVA	43.4	32.8	14.6	3.8	3.4	2.1
Trypsinogen	TGN	23.6	32.7	23.6	6.4	5.9	7.7

Table 2: Orientation permutations for 3-component models

# of factors	Model	I	J	K
3	XspYexZsa	Spectra	Excitation λ	Sample
	XexYspZsa	Excitation λ	Spectra	Sample
	XspYsaZex	Spectra	Sample	Excitation λ
	XsaYspZex	Sample	Spectra	Excitation λ

Figure 1: Peptide backbone of protein showing phi (ϕ) and psi (ψ) dihedral angle.

Figure 2: Ramachandran plot showing distribution of (ϕ, ψ) dihedral angles around the ideal dihedral angles for each type of secondary structure. Undefined dihedral angles occur in regions outside the labeled areas.

Figure 3: Trilinear data array produced by multi-excitation UVRR spectra. Myoglobin (MBN), glucose oxidase (UOX) and trypsinogen (TGN) are shown.

Figure 4. Core consistency plots of a (A) two-, (B) three-, (C) four- and (D) five-component PARAFAC model of the ME-UVRR data.

Figure 5. Split-half analysis results showing the average sum of the squares of residuals for different component PARAFAC models.

Figure 6: Percent RMSEC for the 3-component models for each secondary structure type with (w/NN) and without (w/o NN) non-negativity constraints.

Figure 7: Resolved underlying spectra obtained from each 3-component PARAFAC model for each secondary structure type. (A) XspYsaZex, without non-negativity constraints, (B) XspYsaZex, with non-negativity constraints, (C) XspYexZsa, with non-negativity constraints and (D) XexYspZsa, with non-negativity constraints.

Figure 8: The pure secondary structure Raman spectra (PSSRS) obtained from the three component PARAFAC analysis of the trilinear ME-UVRR dataset.

Figure 9: The experimental versus predicted percentages of secondary structure composition for the 3-component model (XsYsaZex).

Figure 10: Excitation profiles for each secondary structure obtained from the 3-component XspYsaZex PARAFAC model. For visualization, the estimated excitation profiles based on a Gaussian fit of the resolved excitation profile are shown as dashed lines.

Table 1: Secondary structure content (%) of proteins used calculated from (ϕ , ψ) dihedral angles as found on the Research Collaboratory for Structural Bioinformatics (RSCB) Protein Data Bank. Helices include both α -helical and 3_{10} -helical (ϕ , ψ) dihedral angles.

Protein	Abbreviation	Helices	Sheet/Strand	Total disordered			
				PPII	Unordered	α -L	Undefined
Bovine serum albumin	BSA	80.0	5.6	8.5	4.4	0.5	1.0
Carbonic anhydrase	CAH	27.2	37.5	21.2	8.4	3.3	2.3
Chymotrypsinogen A	CTG	26.0	33.3	23.5	7.4	5.3	4.5
Cytochrome c	CYC	52.0	15.7	18.6	4.9	6.9	2.0
Glucose oxidase	UOX	44.3	25.9	15.5	5.9	5.5	2.9
Lysozyme	LSZ	50.4	14.2	12.6	7.9	8.7	6.3
Myoglobin	MBN	85.4	5.3	2.6	2.0	2.6	2.0
Ovalbumin	OVA	43.4	32.8	14.6	3.8	3.4	2.1
Trypsinogen	TGN	23.6	32.7	23.6	6.4	5.9	7.7

Table 2: Orientation permutations for 3- and 4- factor models

# of factors	Model	I	J	K
3	XspYexZsa	Spectra	Excitation λ	Sample
	XexYspZsa	Excitation λ	Spectra	Sample
	XspYsaZex	Spectra	Sample	Excitation λ
	XsaYspZex	Sample	Spectra	Excitation λ