DEVELOPMENT OF ADVANCED CHEMOMETRIC METHODS FOR ANALYSIS OF DEEP-ULTRAVIOLET RESONANCE RAMAN AND CIRCULAR DICHROISM SPECTROSCOPIC DATA FOR PROTEIN SECONDARY STRUCTURE DETERMINATION.

A dissertation

presented to

the Faculty of Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

OLAYINKA OSHOKOYA

Dr. Renee D. JiJi, Dissertation Supervisor

MAY, 2015

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

DEVELOPMENT OF ADVANCED CHEMOMETRIC METHODS FOR ANALYSIS OF DEEP-ULTRAVIOLET RESONANCE RAMAN AND CIRCULAR DICHROISM SPECTROSCOPIC DATA FOR PROTEIN SECONDARY STRUCTURE DETERMINATION

presented by Olayinka Oshokoya,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Renee JiJi

Professor Susan Lever

.

Professor Michael Greenlief

Professor Jianguo Sun

This work is dedicated to my magnificent wife, Oluwadamilola Ayo and son Enoch Oluwatobiloba. Everything I do, I do for you

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Almighty God for seeing me through to the completion of my Ph.D. degree, may your name be forever praised above all.

I would like to thank my graduate advisor, Dr. Renee Jiji, for her encouragement, patience and guidance throughout my graduate career. Without the efforts of Dr. Jiji, this work would not have been completed and I would not have found the drive for research which is so critical in a scientist. I also thank Dr. JiJi for advice and direction in UV Raman instrumentation, data interpretation and presentation, critiques of publication manuscripts, oral and poster presentations and beyond.

I would also like to thank the members of my graduate committee, Dr. Michael Greenlief, Dr. Susan Lever, and Dr. Jianguo Sun, as well as Dr. Jason Cooley, who have provided crucial guidance to me throughout my graduate career. I deeply appreciate their accommodation and willingness to help me succeed at the University of Missouri. I am very grateful to the Department of Chemistry at the University of Missouri for giving me the opportunity to work towards my doctoral degree. The financial support of the Chemistry Department has made both my graduate career and research possible.

I would like to thank all the members of the Jiji and Cooley groups for their help through the years. Their input on my publications, presentations, and research has served to improve my ability as a scientist.

Finally, I would like to thank my parents and siblings for their emotional support and prayers, my friends; Richard Osibanjo and Anthony Omosule for encouragement and advice when needed and my fellow graduate students at the chemistry department, University of Missouri. I also thank Jesus House Columbia and The Crossing Church. The people I have met here have

provided me with the support and encouragement which is so necessary for a successful graduate career. They have provided a sense of community that has been one of the greatest joys of my graduate career. Through their presence and interaction, my time in graduate school has become a wonderful time in my life that I will reflect on fondly for years to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 – Multivariate analysis of spectroscopic data for protein secondary structure determination.

1.1 Introduction

Protein structure determination has become a field of great significance in biophysics and biochemistry as it relates directly to protein function and changes in structure that consequently may lead to certain diseases[1, 2]. Deep-ultraviolet resonance Raman (DUVRR) spectra is sensitive to secondary structural motifs but, similar to circular dichroism (CD) and infrared spectroscopy, require the application of multivariate and advanced statistical analysis methods to resolve the pure secondary structure Raman spectra (PSSRS) for determination of secondary structure composition[3, 4].

A number of techniques have been developed for resolving protein secondary structure. X-ray crystallography (XRC) is one of such techniques which requires suitable crystalline samples of protein for structure determination, and producing protein crystals for some samples can be challenging. However, complete protein structures can be resolved to within fractions of an angstrom by this method[5-7]. Solid-state nuclear magnetic resonance (NMR) spectroscopy has emerged as a powerful tool for determining complete protein structures without the need to first produce a crystalline protein, although the complexity of the data can make interpretation a laborious process[8-10]. Circular dichroism which is an absorption based technique, has long been used for the evaluation and quantification of protein secondary structure. When more than one secondary structure type is present in a protein, deconvolution of CD data for quantification of protein secondary structure is subject to having a reliable reference set of pure spectra[11-14].

Vibrational spectroscopies have also been applied to quantify protein secondary structure, as specific secondary structure motifs have been found to correspond with certain vibrational bands. Studies involving Raman spectroscopy and infrared (IR) spectroscopy[15-17] have been applied to quantitation of protein secondary structure in terms of α-helix, β-sheet, and disordered structure. Temperature dependent Raman, vibrational circular dichroism (VCD), and Raman optical activity (ROA) spectroscopies have the additional advantage over other techniques of distinguishing between different types of α-helix and β-sheet peptide and protein structures[18-25].

The success of previous methods to quantify protein secondary structure content using deep ultraviolet resonance Raman spectra suggests that future multivariate analyses will be unique to the experimental data acquired, as such data continues to grow in complexity. The use of two or more spectroscopic techniques complimentarily and a combination of mathematical techniques, such as variations of component analysis[26], multivariate analysis approaches[3, 4, 27], use of data fusion techniques[28-31] and variable selection techniques[32, 33] suggest that future research will be able to look at more complex experimental data and extract useful protein secondary structure information.

The rest of this chapter focuses on giving a background to protein structure, ultraviolet resonance Raman and circular dichroism sensitivity to protein secondary structure, ultraviolet resonance Raman and circular dichroism instrumentation for better comprehension of chapters 3, 4 and 5.

1.2 Protein structure terminology

Protein structure has four levels; primary, secondary, tertiary, and quaternary. The primary structure refers to the simple amino acid sequence while the secondary structure refers to the structural motifs within the protein that are defined by the φ and ψ dihedral angles of the amide backbone (Figure 1.1). The three-dimensional arrangement of the secondary structures is the tertiary structure and quaternary structure is the arrangement of protein subunits to each other in larger complexes that function as a single unit[34]. The fact that secondary structure changes impact the tertiary and quaternary structure of a protein, as well as the observation that secondary structure changes unaccompanied by changes to the primary structure are involved in protein based diseases and altered protein function, has led to an increased interest in resolving and quantifying protein secondary structure content[35-37]. The discussion in this chapter would therefore be limited to secondary structure alone.

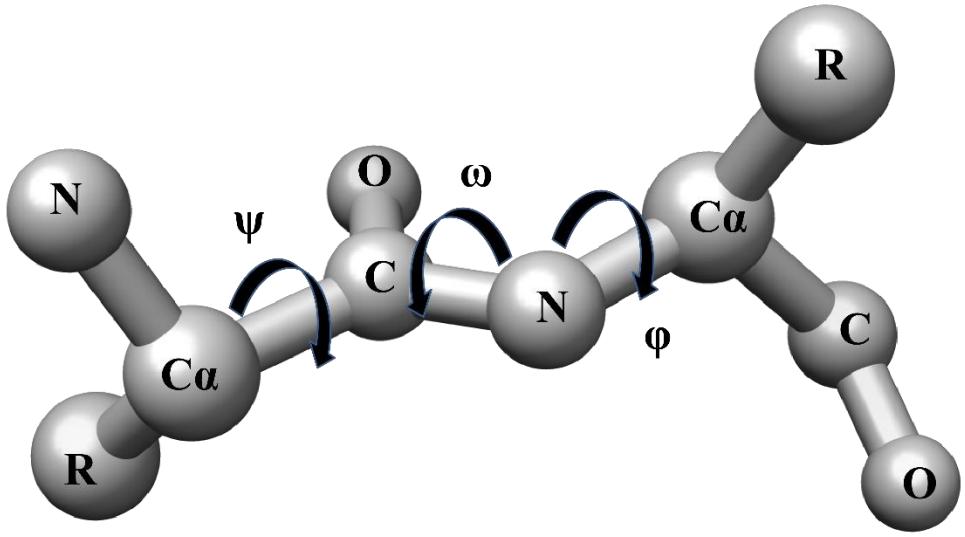

Fig. 1.1 Amide backbone of polypeptide showing phi (φ) and psi (ψ) angles.

1.2.1 Secondary structure

The secondary structure of a protein is determined by the set of dihedral angles (ψ, φ and ω), which define the spatial orientation of the peptide backbone, and the presence of specific hydrogen bonds. Given the planar character of the amide bond, the ω (along the C_α-C-N-C_α bonds) either adopts the more stable trans configuration of 180° or the cis configuration of 0°. The φ and ψ angles however, have more flexibility. When the backbone dihedral angles (φ, ψ) have repeating values, the peptide forms regular secondary structure. Standard terminology and description of the characteristics have been published[38-44] and serve as standards for defining specific secondary structure types, however, these standards are not always strictly followed.

1.2.1.1 Helices

Most helical structures feature a hydrogen bond three to five amino acid residues away from the carbonyl oxygen to the amide hydrogen which causes the backbone to trace a right-handed screw as one looks down its axis from the amino terminal end. The most common and stable helical form is the right-handed α-helix[45]. For every turn along the helical axis there are 3.6 amino acid residues that span 1.5 Å along the axis. The φ and ψ angles center around -60° and -45°, respectively. Other helical structures have been identified or theorized including the 3_{10}-helix (-49°, -26°) and the π-helix (-57°, -69°).

1.2.1.2 β-sheet

β-sheet structures[46] are generally more extended than α-helices. The backbone carbonyls and amides alternate their orientation along this extended conformation which gives rise to two hydrogen bonding schemes between individual strands: the parallel and anti-parallel[47]. As the protein continues from the *N* to *C* terminus, parallel strands propagate in the same direction while

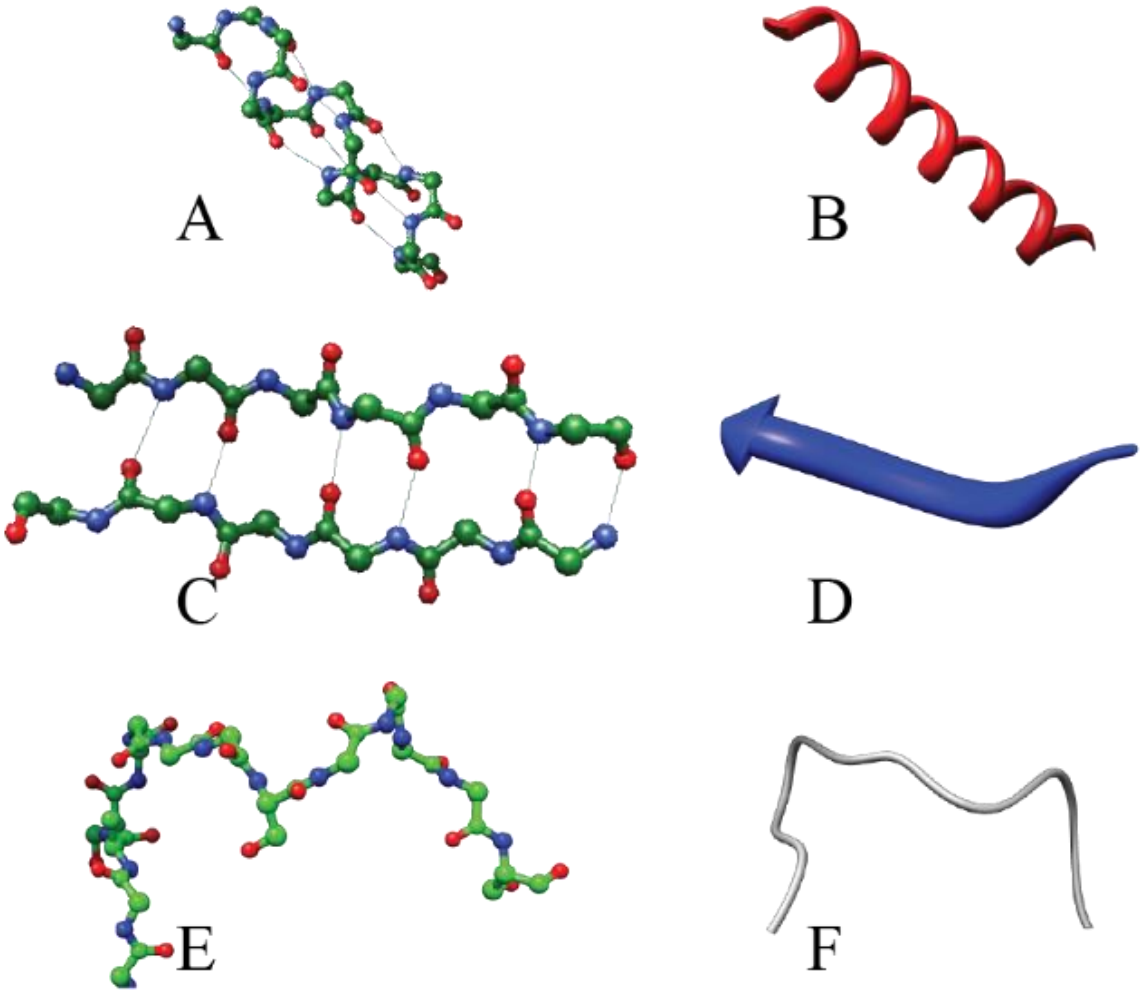

Fig. 1.2 Different secondary structures of protein. (A) α-helix (B) α-helix in ribbon representation (C) β-sheet (D) β-sheet in ribbon representation (E) unfolded structure (F) unfolded structure in ribbon representation.

anti-parallel strands alternate their direction as the β-sheet continues. The C_α does not lie in the plane defined by the extended backbone. Instead, it alternately lies above and below the plane which gives the designation for the pleated β-sheet. Hydrogen bonding in anti-parallel beta sheets is perpendicular to the strand. Parallel strands offset the carbonyl oxygen and amide nitrogen and its hydrogen bonds are more regularly spaced than those in anti-parallel strands. Parallel β-strands, although less stable, also tend to have a closer distribution around -119° and 113° for the φ and ψ angles, respectively[41]. The dihedral angles of anti-parallel strands are more widely scattered about -139° and 135°.

1.2.1.3 Unfolded structure

The final category in protein secondary structure is the disordered or unfolded conformation, in which each amino acid randomly samples all sterically-allowed φ and ψ angles[41, 48]. Historically this has also been called random coil. Recent evidence[49] suggests that these seemingly disordered proteins do have local regions of order that adopt turn-like conformations or those of poly-proline II[50] (PPII, a left-handed 3_{10}-helix formed by trans-L-poly-proline). Turns in a protein[51], (where the protein reverses its general direction) are also non-repetitive and occur over only a few residues but are abundant in protein structures. Their conformational flexibility is much greater than helices and β-sheets. The amino acid's position in the turn becomes more important in determining the allowed dihedral angles.

1.3 Optical methods for secondary structure determination

1.3.1 Circular dichroism

Circular dichroism (CD) is a spectroscopic technique that measures the difference in absorption of circularly (left and right) polarized light by an analyte. The phenomenon occurs when an analyte is intrinsically chiral, linked to a chiral center or placed in an asymmetric environment. Circular dichroism is widely applied to the study of protein structure because it is especially suitable for monitoring conformational changes[11, 13, 14]. The bonds of the peptide backbone absorb in the 190-250 nm range, giving us useful information about protein secondary structure. Each secondary structure motif has a signature spectral shape (Figure 1.3).

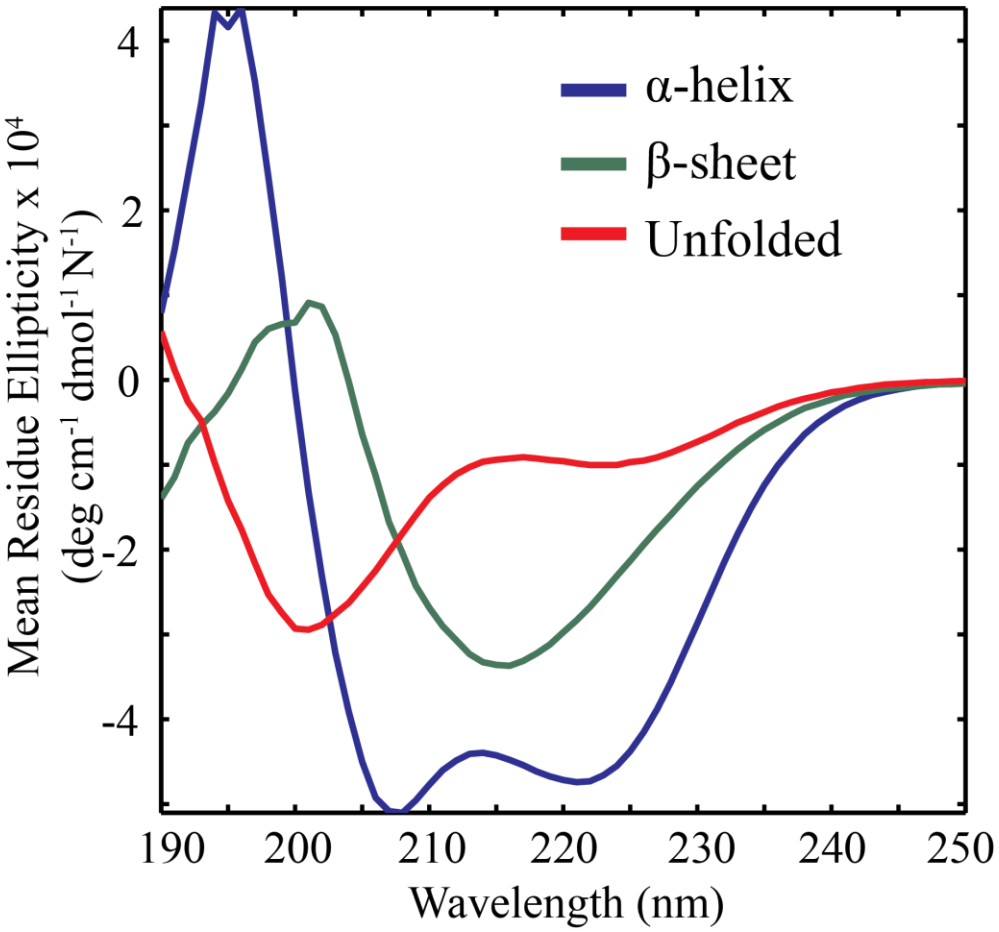

Fig. 1.3 Circular dichroism spectra showing the different protein secondary structure.

The α-helix is characterized by two negative features at 208 and 222 nm and one positive feature at 193 nm. The β-sheet has a distinct positive feature around 194 nm and a negative feature around 217 nm. The poly-proline type II (PPII) structure is characterized by a negative feature at 196 nm and a broad positive feature between 210 and 217 nm. Environmental information about any present side chains, especially aromatic amino acids, can be obtained in the 260 nm to 320 nm region if required.

1.4 Vibrational methods for secondary structure determination

Vibrational spectra for protein structure analysis may be obtained either by infrared absorption (IR) or Raman scattering spectroscopy and both techniques provide complimentary information.

1.4.1 Infrared spectroscopy

Infrared spectroscopy (IR) is one of the oldest and well established experimental techniques for secondary structure analysis of proteins and polypeptides[52-55]. It is non-destructive, requires very little sample preparation and can be used under a wide variety of conditions. This makes it a valuable tool for the investigation of protein structure[56-62] of the molecular mechanism of protein reactions[63-69] and of protein folding, unfolding and misfolding[70-73].

Protein IR spectra results from the absorption of IR radiation by vibrating chemical bonds (primarily stretching and bending motions) of the protein or polypeptide repeating amide backbone. This gives rise to nine characteristic IR absorption bands namely, amide A, B and I-VII. Of these, the amide I and II bands are the most prominent of the protein backbone vibrational modes[16, 53, 54]. The most sensitive spectral region to the protein secondary structural components is the amide I band (1700−1600 cm^{-1}) due almost entirely to the carbonyl stretch vibrations of the

peptide linkages. The frequencies of the amide I band components are found to be correlated closely to each secondary structural element of the proteins. The amide II band (1575-1480 cm^{-1}), in contrast, derives mainly from in-plane NH bending and from the CN stretching vibration, shows much less protein conformational sensitivity than the amide I[53]. The remaining amide vibrational bands are very complex and are therefore of less practical use in protein conformational studies[74].

Previously, the use of IR spectroscopy for protein secondary structure analysis was severely limited by factors including low sensitivity of the instrument, interfering absorption from aqueous solvent, and the lack of understanding of the correlations between specific backbone folding types and the individual component bands. IR spectrum was difficult unless D_2O was used as a solvent as water absorbs strongly in the most important spectral region at approximately 1640 cm^{-1}.[75] Even in D_2O solution, usually only qualitative information was obtained because the components of absorption bands associated with specific substructures, such as α-helix and β-sheet, could not be resolved. Later, as more protein structures were solved by X-ray crystallography, and the computational procedures for the resolution enhancement of broad IR bands of protein were developed, the IR spectra of polypeptides and proteins in both H_2O and D_2O solution could be assigned to secondary structure[15, 54, 55, 75-79] (Table 1.1).

The computerization of IR (Fourier transform) instrumentation has improved the signal-to-noise ratio and allowed extensive data manipulation. Also, new band narrowing methods have not only enriched the qualitative interpretation of the IR spectra and also provided a basis for the quantitative estimation of protein secondary structure.[15, 80-84]

Table 1.1 Assignment of amide I band positions to secondary structure[17, 54, 75]

Secondary structure	Band position in H_2O (cm^{-1})	Band position in D_2O (cm^{-1})
α-helix	1648-1657	1642-1660
β-sheet	1623-1641	1615-1638
	1674-1695	1672-1694
Turns	1662-1686	1653-1691
Disordered	1642-1657	1639-1654

1.5 UVRR sensitivity to secondary structure

The structural sensitivity of vibrational spectroscopic methods, such as infrared absorption spectroscopy, non-resonance Raman spectroscopy and ultraviolet resonance Raman spectroscopy primarily arises from the signal of the polypeptide backbone amide group. This signal results in several discrete vibrational modes (Figure 1.4): the amide I, II, III and S modes[85, 86]. A survey of these modes are discussed below and a combination of all four modes simultaneously make ultraviolet resonance Raman a very sensitive technique for secondary structure determination.

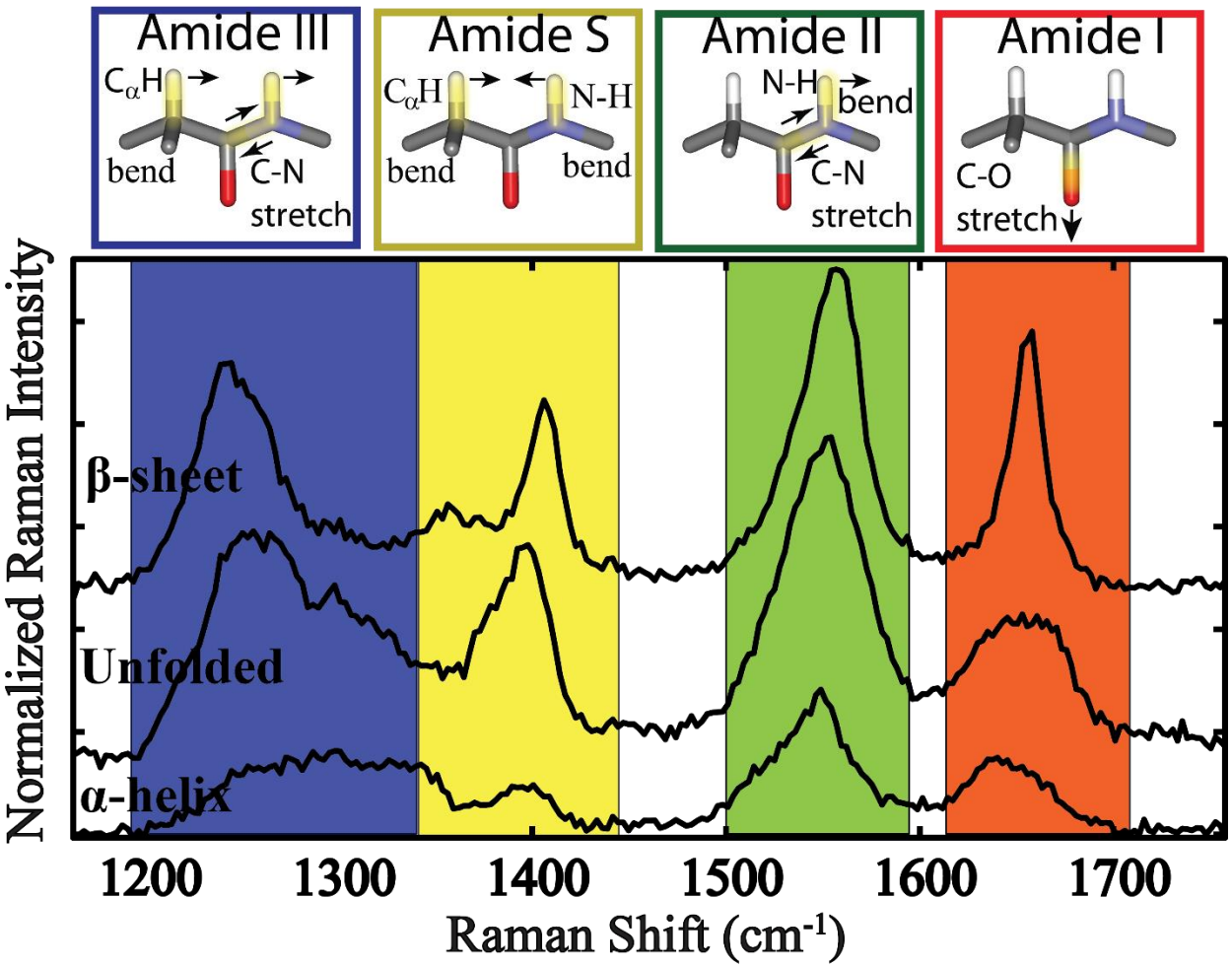

Fig. 1.4 UVRR amide modes of polypeptide backbone.

1.5.1 Amide I

The amide I is predominantly a carbonyl stretch with some NH bending contribution in the 1600 to 1700 cm^{-1} region[85]. The hydrogen bonding capability of the carbonyl makes the amide I a very structurally-sensitive mode because each secondary structure has a unique hydrogen bonding geometry. α-helices have lower frequency amide I modes (1645-1655 cm^{-1}) because of intramolecular hydrogen bonding and hydrogen bonding to water[87]. Amide I modes of β-sheet and disordered peptides overlap each other between 1660-1680 cm^{-1}. In soluble protein spectra, however, the amide I is usually the least intense of the amide modes (unless the protein has a very high helical content).

1.5.2 Amide II

The amide II is an out-of-phase combination of C-N stretching and N-H in-plane bending and appears between 1500 and 1600 cm^{-1}. α-helices have two distinct modes at lower wavenumbers (1520-1555 cm^{-1}) while β-sheet and disordered structures are at higher wavenumbers (1548-1564 cm^{-1})[87]. Its intensity is inversely proportional to helical content in a protein[87].

1.5.3 Amide S

The amide S, alternatively called the $C_\alpha H$ bending mode in literature, was the last of the amide modes to be identified and therefore not numerical in sequence. The "S" notation was proposed by Spiro for "secondary structure-sensitive"[88]. Finally, the $C_\alpha H$ bending coupled to NH bending was determined to be its source[86]. The amide S appears from 1374-1397 cm^{-1} for disordered structures and 1395-1406 cm^{-1} for beta sheet structures[87]. Like the amide II, its intensity is also inversely proportional to helical content[87]. Fully α-helical peptides show no amide S in their

Raman spectra because the α-helical conformation forces the C_α hydrogen in close proximity to the amide hydrogen which decouples the two vibrations[89, 90].

1.5.4 Amide III

The amide III shares the same C-N stretch and N-H in-plane bending modes as the amide II, only these are in phase and appear between 1200 and 1350 cm^{-1}. α-helices have an amide III mode at higher wavenumbers (1254-1345 cm^{-1}), β-sheets at the lower end (1220 to 1241 cm^{-1}), and disordered in between (1240-1279 cm^{-1}). Furthermore, three main sub-bands have been identified in the literature, called the amide III_1, III_2, and III_3 bands, in order of high to low energy. The latter is the "classical" amide III mode described in prior literature before the other two components were identified[91]. Most of the recent DUVRR structural analysis literature has focused on the amide III as it is the most sensitive to secondary structure. Amide III frequency is sinusoidally dependent upon the φ and ψ dihedral angles of the protein backbone[92, 93], which had been shown earlier[94] with non-resonance Raman spectroscopy. Therefore, ψ angle distributions can be quantified. However, the amide III is often overlapped with aromatic residue vibrations (assuming aromatic residues are present in the primary protein sequence), which are typically subtracted out in fitting analyses of the mode. The amide III is very rich in structural information when there is not significant spectral overlap in this region.

1.5.5 Aromatic vibrational modes

The amide modes are certainly adequate for structural analysis, but aromatic vibrational modes report on hydrogen bonding and environmental polarity and are largely responsible for reporting on tertiary and quaternary contacts in a protein. The aromatic amino acid, phenylalanine, tyrosine, and tryptophan contribute to the UVRR spectra of proteins[95]. Like the signal of the amide

backbone, the signals arising from the aromatic amino acids appear in several distinct modes and are also subject to resonance enhancement., with maximum enhancement occurring at excitation wavelengths beyond the optimum range of the amide backbone (>210 nm) for resonance enhancement. The modes arising from the aromatic amino acid side chains heavily overlap with the amide modes. The contribution from the aromatic modes will depend on the abundance of each aromatic amino acid and the environment of each individual residue[95].

UVRR spectra will have contributions from each secondary structure type, proportional to the relative amount of each structure type, in each amide mode, altering the position and intensity of each. In addition, the degree of resonance enhancement of each mode is directly related to the absorption profile of the chromophore at any selected excitation wavelength[96]. Each secondary structure has a unique absorption profile and, as a result[97], the excitation profile of each mode is conformation dependent[98-101]. As a result of the aforementioned relationships, varying the secondary structure content, pH, temperature, or excitation wavelength for a single protein will result in a bilinear data matrix, which is well suited to a multivariate data analysis approach. Alternatively, a series of proteins with varying secondary structure contents will also result in a bilinear data matrix.

Great interest exists for further application of UVRR, for instance in protein dynamics. However, interpretation of UVRR spectra can be extremely challenging due to spectral overlap. Often the barrier to structurally meaningful information is not experimental, but the methods used to analyze the spectra. As a result, a number of advanced chemometric methods have already been applied to UVRR spectra including: factor analysis, least squares, multivariate curve resolution and single point calibrations[26, 101-103].

1.6 UVRR instrumental overview

Ultraviolet resonance Raman (UVRR) is a technique with a wide array of applications including protein secondary structure determination and protein dynamics studies. While a powerful technique, UVRR instruments are not commercially available, therefore, each instrument must be designed and constructed by the intended users with the specific requirements of the field of interest in mind. UVRR studies of proteins have very specific requirements of the instrument in terms of excitation light source and due to the nature of resonance enhancement, photons of sufficient energy which will result in an electronic transition must be generated.

In the case of protein spectroscopy, the primary chromophore of interest is the amide backbone, however, the aromatic amino acids are also of great interest[85]. Resonance enhancement of amide modes rapidly increases below 220 nm as you approach the $\pi^*_3 \leftarrow \pi_2$ dipole allowed transition at 188 nm, leading to greater resonance enhancement as you decrease excitation wavelength[104]. This requires a laser source for the UVRR instrument capable of generating light in what is commonly called the "deep" ultraviolet region (180-280 nm)[105]. The laser source for deep UV light must be created from a multiple harmonics of a tunable Ti:Sapphire laser pumped by a frequency doubled Nd:YLF laser[106]. The Ti:Sapphire beam (IR wavelength near 800 nm) is frequency doubled at the second harmonic lithium triborate (LBO) crystal to create blue light. Both the residual collinear IR and blue light are frequency mixed at the third harmonic beta barium borate (BBO) crystal to produce violet light. This, in turn, is collinear with the residual IR and frequency mixed with another BBO crystal to produce the desired deep UV light from 195 to 206 nm. BBO crystals are moisture-sensitive, therefore the harmonics cavity is kept under a positive pressure of nitrogen gas.

The output of the harmonics cavity is directed to the sample chamber with a series of mirrors. In addition, just before the sample chamber, the laser passes through a biconvex lens to focus the laser to a pinpoint. At the sample chamber, the sample is irradiated with the focused excitation beam generating the Raman scatter.

The flow cell is pumped by a peristaltic pump to circulate the small (0.5 – 5 mL) sample volume which can be cooled by a water jacket surrounding the sample reservoir from 4° to 35° C, although measurements are typically taken at room temperature. The sample flows down two thin nitinol wires as a thin film of approximately 1 mm width. The focused laser beam is incident on this film. The atmosphere is nitrogen purged to avoid contributions to the Raman spectrum from molecular oxygen. The Raman scattered light is collimated and focused into a spectrograph which spatially separates the Rayleigh scattering and directs the Stokes scattering on a liquid nitrogen-cooled charge-coupled device (CCD). Figure 1.3 shows a simple schematic of the UVRR instrumentation used.

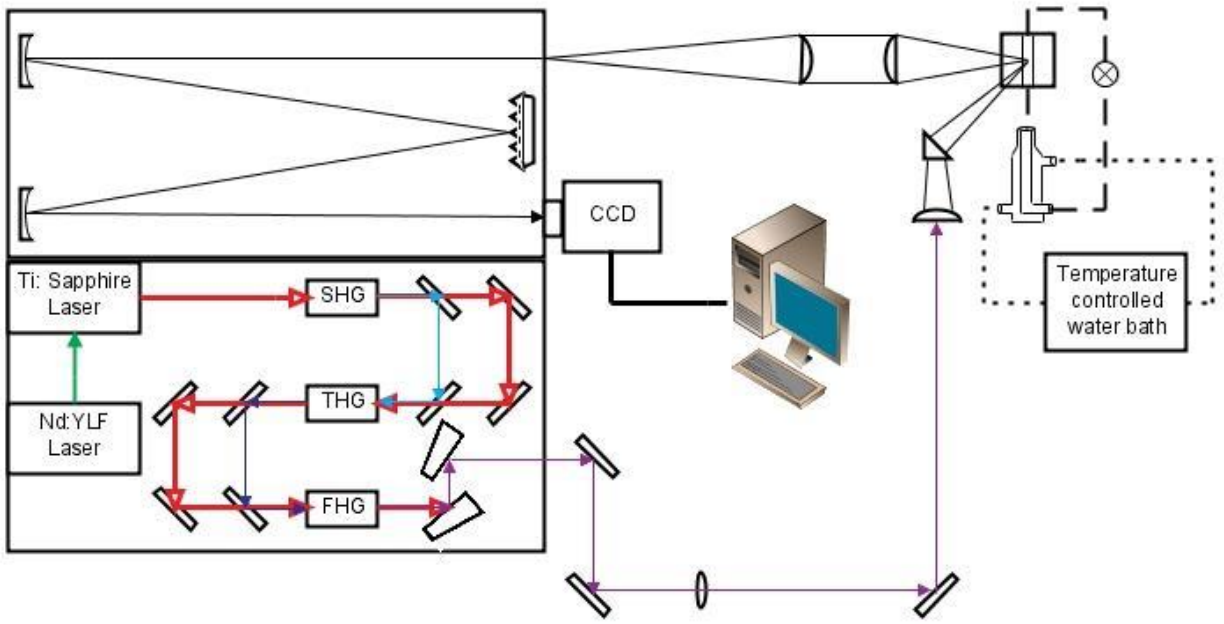

Fig. 1.5 Schematic of UVRR instrument.

1.7 References

1. E. Herczenik and M. F. B. G. Gebbink, *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. V. A. Shashilov and I. K. Lednev, *Chemical Reviews*, 2010, **110**, 5692-5713.

4. V. A. Shashilov, V. Sikirzhytski, L. A. Popova and I. K. Lednev, *Methods*, 2010, **52**, 23-37.

5. G. Bujacz, M. Miller, R. Harrison, N. Thanki, G. L. Gilliland, C. M. Ogata, S. H. Kim and A. Wlodawer, *Acta crystallographica. Section D, Biological crystallography*, 1997, **53**, 713-719.

6. H. Eklund, B. Nordstrom, E. Zeppezauer, G. Soderlund, I. Ohlsson, T. Boiwe, B. O. Soderberg, O. Tapia, C. I. Branden and A. Akeson, *Journal of molecular biology*, 1976, **102**, 27-59.

7. A. Higashiura, K. Ohta, M. Masaki, M. Sato, K. Inaka, H. Tanaka and A. Nakagawa, *Journal of Synchrotron Radiation*, 2013, **20**, 989-993.

8. F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein and H. Oschkinat, *Nature*, 2002, **420**, 98-102.

9. T. A. Cross and S. J. Opella, *Journal of the American Chemical Society*, 1983, **105**, 306-308.

10. D. S. Wishart, B. D. Sykes and F. M. Richards, *Journal of Labelled Compounds and Radiopharmaceuticals*, 1992, **31**, 1019-1028.

11. N. J. Greenfield, *Analytical Biochemistry*, 1996, **235**, 1-10.

12. N. J. Greenfield, 2004, vol. 383, pp. 282-317.

13. N. J. Greenfield, *Nature Protocols*, 2006, **1**, 2876-2890.

14. N. J. Greenfield and G. D. Fasman, *Biochemistry*, 1969, **8**, 4108-4116.

15. A. Dong, P. Huang and W. S. Caughey, *Biochemistry*, 1990, **29**, 3303-3308.

16. W. K. Surewicz, H. H. Mantsch and D. Chapman, *Biochemistry*, 1993, **32**, 389-394.

17. A. Barth and C. Zscherp, *Q Rev Biophys*, 2002, **35**, 369-430.

18. T. A. Keiderling, in *Circular Dichroism Principles and Applications*, eds. N. K., B. N. and R. W. Woody, VCH, New York, 1994, pp. 497-516.

19. T. A. Keiderling, in *Circular Dichroism: Principles and Applications*, eds. Berova N., Nakanishi K. and W. R. A., Wiley-VCH, New York, 2 edn., 2000, pp. 621-666.

20. T. A. Keiderling, *Current Opinion in Chemical Biology*, 2002, **6**, 682-688.

21. M. N. Kinalwa, E. W. Blanch and A. J. Doig, *Analytical Chemistry*, 2010, **82**, 6347-6349.

22. S. A. Oladepo, K. Xiong, Z. Hong and S. A. Asher, *Journal of Physical Chemistry Letters*, 2011, **2**, 334-344.

23. C. R. Jacob, S. Luber and M. Reiher, *Chemistry – A European Journal*, 2009, **15**, 13491-13508.

24. T. Weymuth and M. Reiher, *The Journal of Physical Chemistry B*, 2013, **117**, 11943-11953.

25. C. R. Jacob, S. Luber and M. Reiher, *The Journal of Physical Chemistry B*, 2009, **113**, 6558-6573.

26. M. Xu, V. A. Shashilov, V. V. Ermolenkov, L. Fredriksen, D. Zagorevski and I. K. Lednev, *Protein Science*, 2007, **16**, 815-832.

27. J. V. Simpson, O. Oshokoya, N. Wagner, J. Liu and R. D. Jiji, *Analyst*, 2011, **136**, 1239-1247.

28. David Lee Hall and S. A. H. McMullen, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA, 2004.

29. Martin Liggins II, David Hall and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*, CRC Press, Boca Rotan,FL, 2008.

30. H. B. Mitchell, *Multi-Sensor Data Fusion: An Introduction*, Springer, Berlin, 2007.

31. L. A. Klein, *Sensor and Data Fusion Concepts and Applications*, 2nd ed edn., SPIE Optical Engineering Press, Bellingham, 1999.

32. Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes and M. Hanpin, *Analytica Chimica Acta*, 2010, **667**, 14-32.

33. T. Mehmood, K. H. Liland, L. Snipen and S. Sæbø, *Chemometrics and Intelligent Laboratory Systems*, 2012, **118**, 62-69.

34. D. Voet and J. G. Voet, *Biochemistry*, 3rd edn., John Wiley & Sons, Inc., Hoboken, NJ, 2004.

35. C. C. Blake, M. J. Geisow, S. J. Oatley, B. Rerat and C. Rerat, *Journal of molecular biology*, 1978, **121**, 339-356.

36. S. B. Prusiner, *Proceedings of the National Academy of Sciences*, 1998, **95**, 13363-13383.

37. C. Weissmann, *Nature Reviews Microbiology*, 2004, **2**, 861-871.

38. *Journal of molecular biology*, 1966, **20**, 589.

39. J. T. Edsall, P. J. Flory, J. C. Kendrew, A. M. Liquori, G. Nemethy, G. N. Ramachandran and H. A. Scheraga, *Journal of Biological Chemistry*, 1966, **241**, 1004-1008.

40. J. T. Edsall, P. J. Flory, J. C. Kendrew, A. M. Liquori, G. Némethy, G. N. Ramachandran and H. A. Scheraga, *Journal of molecular biology*, 1966, **15**, 399-407.

41. G. N. Ramachandran and V. Sasisekharan, *Advances in Protein Chemistry*, 1968, **23**, 283-438.

42. J. S. Richardson, *Advances in Protein Chemistry*, 1981, **34**, 167-339.

43. G. D. Rose, L. M. Gierasch and J. A. Smith, *Advances in Protein Chemistry*, 1985, **37**, 1-109.

44. W. Kabsch and C. Sander, *Biopolymers - Peptide Science Section*, 1983, **22**, 2577-2637.

45. L. Pauling, R. B. Corey and H. R. Branson, *Proceedings of the National Academy of Sciences of the United States of America*, 1951, **37**, 205-211.

46. F. R. Salemme and D. W. Weatherford, *Journal of molecular biology*, 1981, **146**, 101-117.

47. L. Pauling and R. B. Corey, *Proceedings of the National Academy of Sciences of the United States of America*, 1951, **37**, 729-740.

48. D. A. Brant and P. J. Flory, *Journal of the American Chemical Society*, 1965, **87**, 2791-2800.

49. R. Schweitzer-Stenner, *Molecular BioSystems*, 2012, **8**, 122-133.

50. P. M. Cowan and S. McGavin, *Nature*, 1955, **176**, 501-503.

51. E. G. Hutchinson and J. M. Thornton, *Protein Science : A Publication of the Protein Society*, 1994, **3**, 2207-2216.

52. A. Elliott and E. J. Ambrose, *Nature*, 1950, **165**, 921-922.

53. S. Krimm and J. Bandekar, *Adv Protein Chem*, 1986, **38**, 181-364.

54. H. Susi and D. M. Byler, *Methods Enzymol*, 1986, **130**, 290-311.

55. W. K. Surewicz and H. H. Mantsch, *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular*, 1988, **952**, 115-130.

56. J. L. R. Arrondo, A. Muga, J. Castresana and F. M. Goñi, *Progress in Biophysics and Molecular Biology*, 1993, **59**, 23-56.

57. F. Siebert, *Methods in Enzymology*, 1995, **246**, 501-526.

58. M. Jackson and H. H. Mantsch, *Critical Reviews in Biochemistry and Molecular Biology*, 1995, **30**, 95-120.

59. E. Goormaghtigh, V. Cabiaux and J. M. Ruysschaert, *Sub-cellular biochemistry*, 1994, **23**, 329-362.

60. E. Goormaghtigh, V. Cabiaux and J. M. Ruysschaert, *Sub-cellular biochemistry*, 1994, **23**, 363-403.

61. E. Goormaghtigh, V. Cabiaux and J. M. Ruysschaert, *Sub-cellular biochemistry*, 1994, **23**, 405-450.

62. J. L. R. Arrondo and F. M. Goñi, *Progress in Biophysics and Molecular Biology*, 1999, **72**, 367-405.

63. J. Breton, *Biochimica et Biophysica Acta - Bioenergetics*, 2001, **1507**, 180-193.

64. A. Barth, in *Protein Structures: Methods in Protein Structure and Stability Analysis*, ed. E. A. P. E. V.N. Uversky, Nova Science Publishers, 2006.

65. H. Fabian and W. Mäntele, in *Handbook of Vibrational Spectroscopy*, ed. P. R. G. E. J.M. Chalmers, John Wiley & Sons, Chichester, 2002, pp. pp. 3399–3426.

66. W. Mäntele, in *Anoxygenic Photosynthetic Bacteria*, ed. M. T. M. E. Blankenship, C.E. Bauer (Eds.), Kluwer Academic Publishers, Dordrecht, 1995, pp. pp. 627–647.

67. K. Gerwert, *Biological Chemistry*, 1999, **380**, 931-935.

68. X. M. Liu, S. Sonar, C. P. Lee, M. Coleman, U. L. RajBhandary and K. J. Rothschild, *Biophysical Chemistry*, 1995, **56**, 63-70.

69. K. J. Rothschild, *Journal of Bioenergetics and Biomembranes*, 1992, **24**, 147-167.

70. H. Fabian, H. H. Mantsch and C. P. Schultz, *Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**, 13153-13158.

71. D. Reinstädler, H. Fabian and D. Naumann, *Proteins: Structure, Function and Genetics*, 1999, **34**, 303-316.

72. S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff and R. B. Dyer, *Biochemistry*, 1996, **35**, 691-697.

73. R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff and R. B. Dyer, *Biochemistry*, 1997, **36**, 15006-15012.

74. J. Bandekar, *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular*, 1992, **1120**, 123-143.

75. D. M. Byler and H. Susi, *Biopolymers*, 1986, **25**, 469-487.

76. H. Susi and D. Michael Byler, *Biochemical and Biophysical Research Communications*, 1983, **115**, 391-397.

77. N. N. Kalnin, I. A. Baikalov and S. Y. Venyaminov, *Biopolymers*, 1990, **30**, 1273-1280.

78. S. Y. Venyaminov and N. N. Kalnin, *Biopolymers*, 1990, **30**, 1243-1257.

79. S. Y. Venyaminov and N. N. Kalnin, *Biopolymers*, 1990, **30**, 1259-1271.

80. A. Dong, P. Huang and W. S. Caughey, *Biochemistry*, 1992, **31**, 182-189.

81. A. Dong, B. Caughey, W. S. Caughey, K. S. Bhat and J. E. Coe, *Biochemistry*, 1992, **31**, 9364-9370.

82. B. E. Bowler, *Biochemistry*, 1993, **32**, 183-190.

83. B. E. Bowler, *Biochemistry*, 1994, **33**, 2402-2408.

84. A. Dong, J. M. Malecki, L. Lee, J. F. Carpenter and J. C. Lee, *Biochemistry*, 2002, **41**, 6660-6667.

85. J. C. Austin, T. Jordan and T. G. Spiro, *Adv. Spectrosc. (Chichester, U. K.)*, 1993, **20**, 55-127.

86. Y. Wang, R. Purrello, T. Jordan and T. G. Spiro, *Journal of the American Chemical Society*, 1991, **113**, 6359-6368.

87. C. A. Roach, J. V. Simpson and R. D. Jiji, *Analyst*, 2012, **137**, 555-562.

88. Y. Wang, R. Purrello and T. G. Spiro, *Journal of the American Chemical Society*, 1989, **111**, 8274-8276.

89. T. Jordan and T. G. Spiro, *Journal of Raman Spectroscopy*, 1994, **25**, 537-543.

90. R. Schweitzer-Stenner, F. Eker, Q. Huang, K. Griebenow, P. A. Mroz and P. M. Kozlowski, *The Journal of Physical Chemistry B*, 2002, **106**, 4294-4304.

91. A. V. Mikhonin, Z. Ahmed, A. Ianoul and S. A. Asher, *The Journal of Physical Chemistry B*, 2004, **108**, 19020-19028.

92. A. Ianoul, M. N. Boyden and S. A. Asher, *Journal of the American Chemical Society*, 2001, **123**, 7433-7434.

93. S. A. Asher, A. Ianoul, G. Mix, M. N. Boyden, A. Karnoup, M. Diem and R. Schweitzer-Stenner, *Journal of the American Chemical Society*, 2001, **123**, 11775-11781.

94. R. C. Lord, *Applied Spectroscopy*, 1977, **31**, 187-194.

95. S. A. Asher, M. Ludwig and C. R. Johnson, *Journal of the American Chemical Society*, 1986, **108**, 3186-3197.

96. S. A. Asher, *Annual Review of Physical Chemistry*, 1988, **39**, 537-588.

97. K. Rosenheck and P. Doty, *Proceedings of the National Academy of Sciences of the United States of America*, 1961, **47**, 1775-1785.

98. R. A. Copeland and T. G. Spiro, *J. Am. Chem. Soc.*, 1986, **108**, 1281-1285.

99. C. Y. Huang, G. Balakrishnan and T. G. Spiro, *Journal of Raman Spectroscopy*, 2006, **37**, 277-282.

100. B. Sharma, S. V. Bykov and S. A. Asher, *The journal of physical chemistry. B*, 2008, **112**, 11762-11769.

101. R. A. Copeland and T. G. Spiro, *Biochemistry*, 1987, **26**, 2134-2139.

102. V. A. Shashilov, M. Xu, V. V. Ermolenkov and I. K. Lednev, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 2006, **102**, 46-61.

103. J. V. Simpson, G. Balakrishnan and R. D. Jiji, *Analyst*, 2009, **134**, 138-147.

104. R. D. JiJi, G. Balakrishnan, Y. Hu and T. G. Spiro, *Biochemistry*, 2006, **45**, 34-41.

105. D. J. Elliot, *Ultraviolet Laser Technology and Application*, Academic Press Inc, 1995.

106. G. Balakrishnan, Y. Hu, S. B. Nielsen and T. G. Spiro, *Applied Spectroscopy*, 2005, **59**, 776-781.

Chapter 2 - Multivariate Data Analysis

Multivariate analysis encompasses a set of techniques devoted to the analysis of data sets with more than one variable. It involves the application of statistical processes to data systems in order to expose the inherent structure and meaning between and within sets of variables contained in the data system. Linear algebra, also known as matrix algebra, is a study of linear sets of equations and their transformation properties and provides some of the basic mathematical concepts for solving linear systems utilized in multivariate data analysis. Therefore, an understanding of linear algebraic rules is essential to the interpretation of data that would be subjected to multivariate analysis. Basic information on select concepts in linear algebra and multivariate data analysis is presented here to provide context for the data analysis performed in subsequent chapters. A complete survey of both linear algebra and multivariate data analysis is beyond the scope of this work and is extensively reviewed in literature[1-11].

2.1 Linear Algebra

Linear algebra is an extension of scalar mathematics dealing with multi-dimensional quantities. Multi- dimensional data may be represented as vectors or matrices (section 2.1.1) and may be subjected to basic algebraic operations (section 2.1.2). Basic terminology and linear algebraic rules are presented in the following subsections.

2.1.1 Vectors and Matrices

A matrix is a set of numbers presented in two dimensions, rows and columns. When referring to matrix size rows are always listed before columns such that **Y**, an n × m, matrix contains n rows and m columns. A matrix that contains only one row or column it is referred to as a vector, as examples x and y illustrate:

$$\mathbf{x} = [j_1 \quad j_2 \quad \cdots \quad j_m] \tag{2-1}$$

$$\mathbf{y} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix} \tag{2-2}$$

where **x** is a vector that contains m columns with values j_1 through j_m, and **y** is a vector that contains n rows with values k_1 through k_n. Similarly, a matrix may be considered as a concatenated series of vectors, as with matrices **X** and **Y**:

$$\mathbf{X} = \begin{bmatrix} j_{11} & j_{12} & \cdots & j_{1m} \\ j_{21} & j_{22} & \cdots & j_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ j_{n1} & j_{n2} & \cdots & j_{nm} \end{bmatrix} \tag{2-3}$$

$$\mathbf{Y} = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & k_{n2} & \cdots & k_{nm} \end{bmatrix} \tag{2-4}$$

where matrix **X** contains n rows and m columns of values j_{11} through j_{nm}, and matrix **Y** contain n rows and m columns of values k_1 through k_{nm}. One may construct matrix **X** by concatenating n rows of x-sized vectors, or matrix **Y** by concatenating m columns of y-sized vectors.

2.1.2 Matrix Mathematics

Operations used in basic algebra all have corresponding matrix operations in linear algebra, such as addition and multiplication. A few mathematical operations unique to linear algebra exist as well, such as the matrix transpose. Some basic matrix mathematics utilized in the multivariate techniques of Chapters 3, 4 and 5 are presented in the following subsections.

2.1.2.1 Addition and Subtraction

Matrix addition is an element-by-element operation; this feature requires both matrices to have the same row and column dimensions. Addition of matrices **X** and **Y** (Equations 2-3 and 2-4 respectively) to form a new matrix **Z** is performed as:

$$\mathbf{X}+\mathbf{Y}=\begin{bmatrix} j_{11}+k_{11} & j_{12}+k_{12} & \cdots & j_{1m}+k_{1m} \\ j_{21}+k_{21} & j_{22}+k_{22} & \cdots & j_{2m}+k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ j_{n1}+k_{n1} & j_{n2}+k_{n2} & \cdots & j_{nm}+k_{nm} \end{bmatrix}=\mathbf{Z}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nm} \end{bmatrix} \quad (2\text{-}5)$$

where the values of *l* are the sums of their respective *j* and *k* values, therefore **Z** has dimensions equal to **X** and **Y**. Subtraction of values may be achieved in a similar fashion, where some or all values of *j* or *k* are negative.

2.1.2.2 Multiplication

Multiplication of matrices may be achieved in several ways, including scalar multiplication, point-by-point multiplication, and matrix multiplication. Scalar multiplication involves multiplying an entire matrix, **Y**, by a single value, *a*, to give a new matrix **Z**:

$$\mathbf{Y}\times a=\begin{bmatrix} a\times k_{11} & a\times k_{12} & \cdots & a\times k_{1m} \\ a\times k_{21} & a\times k_{22} & \cdots & a\times k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a\times k_{n1} & a\times k_{n2} & \cdots & a\times k_{nm} \end{bmatrix}=\mathbf{Z}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nm} \end{bmatrix} \quad (2\text{-}6)$$

where the elements *l* are the values of the multiplication of a and the corresponding *k*.

Each value of **Y** may be multiplied by a different value in point-by-point matrix multiplication, and throughout this text is indicated as a period between matrices. For matrices **X** and **Y** point-by-point multiplication is performed as:

$$\mathbf{Y}.\mathbf{X} = \begin{bmatrix} j_{11} \times k_{11} & j_{12} \times k_{12} & \cdots & j_{1m} \times k_{1m} \\ j_{21} \times k_{21} & j_{22} \times k_{22} & \cdots & j_{2m} \times k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ j_{n1} \times k_{n1} & j_{n2} \times k_{n2} & \cdots & j_{nm} \times k_{nm} \end{bmatrix} = \mathbf{Z} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nm} \end{bmatrix} \quad (2\text{-}7)$$

where the elements of **Z**, *l*, are the result of the respective multiplication of the elements of **X** and **Y**. In this case, **X** and **Y** must have the same dimensions, and are commutative, where **X**.**Y** = **Y**.**X**. Matrix multiplication involves row-by-column multiplication. Given matrices **X** and **Y** matrix multiplication yields:

$$\mathrm{XY} = \begin{bmatrix} (j_{11} \times k_{11} + j_{12} \times k_{21} + \cdots j_{1m} \times k_{n1}) & \cdots & (j_{11} \times k_{1m} + j_{12} \times k_{2m} + \cdots j_{1m} \times k_{nm}) \\ \vdots & \ddots & \vdots \\ (j_{n1} \times k_{11} + j_{n2} \times k_{21} + \cdots j_{nm} \times k_{n1}) & \cdots & (j_{n1} \times k_{1m} + j_{n2} \times k_{2m} + \cdots j_{nm} \times k_{nm}) \end{bmatrix}$$

$$= \boldsymbol{Z} = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nm} \end{bmatrix} \quad (2\text{-}8)$$

so that the first element of the first row of matrix **X** is multiplied by the first element of the first column of matrix **Y** and added to the product of the second element of the first row of matrix **X** multiplied by the second element of the first column of matrix **Y**, and so on to produce the first value of the first row in **Z**. Matrix multiplication will have no special notation in this text, so **XY** denotes the matrix multiplication of matrices **X** and **Y**, rather than scalar or point-by-point multiplication. Matrix multiplication is not commutative, so **XY** ≠ **YX**, and order of multiplication is important. However, matrix multiplication is associative, where **X**(**XY**) = (**XX**)**Y**. Because matrix multiplication is a row-by-column process, the number of rows of matrix **X** must equal the number of columns of matrix **Y** for multiplication to be possible, however the columns of **X** and the rows of **Y** need not be equal.

2.1.2.3 Matrix Transpose

A matrix transpose converts rows of a matrix to columns and columns of a matrix to rows. Given matrix **Y** (Equation 2-4), the transpose, $\mathbf{Y}^T$, is:

$$\mathbf{Y}^T = \begin{bmatrix} k_{11} & k_{21} & \cdots & k_{n1} \\ k_{12} & k_{22} & \cdots & k_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1m} & k_{2m} & \cdots & k_{nm} \end{bmatrix} \quad (2\text{-}9)$$

where **Y** has dimensions $n \times m$ and $\mathbf{Y}^T$ has dimensions $m \times n$. Any matrix may undergo a matrix transpose.

2.1.2.4 Matrix Pseudoinverse

The inverse of matrix **Y** is a matrix $\mathbf{Y}^{-1}$, where $\mathbf{Y}\mathbf{Y}^{-1} = \mathbf{I}$, and **I** is the identity matrix. Strictly, invertible matrices only exist for square matrices that have non-zero determinants. When matrix **Y** is not a square matrix, a pseudoinverse matrix may be implemented. One such pseudoinverse is described as $(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$, and is equivalent to $\mathbf{Y}^{-1}$ for non-square matrices.

2.2 Multivariate Techniques

Multivariate data analysis techniques have existed for over 70 years[12-20]. Application of multivariate techniques to chemical data has been referred to as chemometrics for over 40 years[15, 21-23]. Several variations on algorithms and techniques in multivariate data analysis exist, a few of which are discussed here in the context of data analyses performed in Chapters 3, 4 and 5. Several books and journal articles cover a much wider array of multivariate data analysis techniques in greater detail than are presented herein[1, 5, 8, 9, 11, 24].

Multivariate methods such as classical least squares (CLS), and partial least squares (PLS) assume a linear relationship between spectral intensity and the analyte concentrations in a mixture[25]; PLS can also be applied to non-linear systems [26-28]. Each method involves a calibration step, where the relationship between the spectra and the concentrations of the components is deduced from a set of reference samples followed by a prediction step in which the results of the calibration are used to determine the concentrations of the components from the spectra of the analyzed samples. For clarity, matrices are represented as bold type uppercase letters, vectors are represented as bold type lowercase letters and scalars are presented in italics according to the standard notation reviewed by Kiers[29]. A superscripted "T" and "+" denote the transpose and pseudoinverse function of a matrix respectively.

In the case of proteins, the deep ultraviolet resonance Raman (DUVRR) and circular dichroism (CD) spectra can be represented as the sum of the fraction of α-helical (c_α), β-sheet (c_β) and any other secondary structure content multiplied by the underlying pure secondary structure profiles for each type of secondary structure plus an error term (Equation 2-10)

$$\mathbf{x} = c_\alpha \mathbf{s}_\alpha + c_\beta \mathbf{s}_\beta + \ ... + e \qquad (2\text{-}10)$$

Where **x** represents the measured spectrum and $\mathbf{s}_\alpha$, and $\mathbf{s}_\beta$ represent the pure secondary structure profiles for α-helix, β-sheet and unordered respectively.

For a series of proteins, the matrix **X** can be represented as the product of two smaller matrices plus an error matrix **E**, (Equation 2-11, Figure 2.1). The matrix, **S**, includes the pure secondary structure profiles and the matrix C contains the fractional amounts of each secondary structure.

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \qquad (2\text{-}11)$$

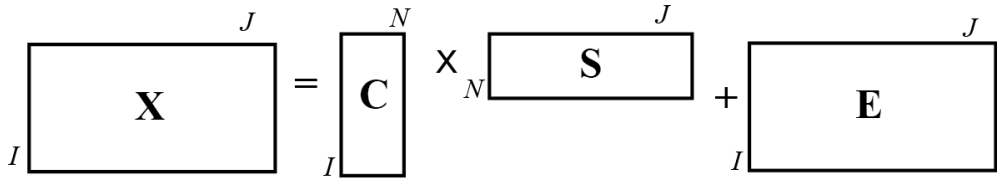

Fig. 2.1 Bilinear model for multivariate analysis of data for protein secondary structure determination.

Multivariate curve resolution (MCR) is a multivariate technique that can resolve multicomponent mixtures into a simple model consisting of a composition weighted sum of the signals of the pure compounds. It is a bilinear method to resolve an experimental data matrix into the product of two matrices that are usually associated with concentration and spectra[30-32].

2.2.1 Classical least squares

Classical least squares (CLS) has been previously used for secondary structure prediction of proteins from infrared and circular dichroism data[33]. It has also been shown that DUVRR spectra of proteins can be resolved into their underlying pure secondary structure Raman spectra (PSSRS)[34]. The PSSRS can be determined from the measured DUVRR spectra of a set of proteins with known secondary structure compositions by rearranging Equation 2-11 to give Equation 2-12

$$\hat{\mathbf{S}}^{\mathbf{T}} = (\mathbf{C}^{\mathbf{T}}\mathbf{C})^{-\mathbf{1}}\mathbf{C}\mathbf{X} \qquad (2\text{-}12)$$

where **X** denotes the matrix containing spectral data and **C** the concentration matrix. $\widehat{\boldsymbol{S}}$ represents the matrix of PSSRS. Once the PSSRS have been determined, the secondary structure content can then be determined using the PSSRS recovered from Equation 2-12 in Equation 2-13

$$\hat{\mathbf{C}}_{\mathbf{unk}} = \mathbf{X}_{\mathbf{unk}}^{\mathrm{T}}\hat{\mathbf{S}}^{\mathrm{T}}\left(\hat{\mathbf{S}}\hat{\mathbf{S}}^{\mathbf{T}}\right)^{-1} \qquad (2\text{-}13)$$

2.2.2 Principal component regression

Principal component regression[35] is a regression method that uses principal component analysis (PCA) in estimating regression coefficients. That is, instead of regressing the independent variables on the dependent variables directly, the principal components of the independent variables are used. The data matrix **X** is first decomposed into two smaller matrices referred to as

scores (**P**) and loadings (**T**), where $\mathbf{X}=\mathbf{P}\mathbf{T}^{T}$. The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular component in the data) and loadings (the weight by which each standardized original variable should be multiplied to get the component score). Using the first n principal components, an estimate of **X**, ($\widehat{\boldsymbol{X}}$) is then employed to determine the underlying PSSRS using Equation 2-14 a-c.

$$\hat{\mathbf{S}}_{\alpha} = \widehat{\boldsymbol{X}}^{+} c_{\alpha} \qquad (2\text{-}14a)$$

$$\hat{\mathbf{S}}_{\beta} = \widehat{\boldsymbol{X}}^{+} \mathrm{c}_{\beta} \qquad (2\text{-}14b)$$

$$\hat{\mathbf{S}}_{u} = \widehat{\boldsymbol{X}}^{+} \mathrm{c}_{u} \qquad (2\text{-}14c)$$

where $\widehat{\boldsymbol{X}}^{+}$is the pseudoinverse of the estimated data matrix. The secondary structural composition of an unknown protein can then be determined from the DUVRR spectrum and the resolved PSSRS according to Equation 2-15 a-c.

$$\hat{\mathrm{c}}_{\alpha,\mathrm{unk}} = r_{\mathrm{unk}}^{\mathrm{T}} \hat{\mathbf{S}}_{\alpha} \qquad (2\text{-}15a)$$

$$\hat{\mathrm{c}}_{\beta,\mathrm{unk}} = r_{\mathrm{unk}}^{\mathrm{T}} \hat{\mathbf{S}}_{\beta} \qquad (2\text{-}15b)$$

$$\hat{\mathrm{c}}_{\mathrm{u,unk}} = r_{\mathrm{unk}}^{\mathrm{T}} \hat{\mathbf{S}}_{\mathrm{u}} \qquad (2\text{-}15c)$$

where r_{unk} is the spectra of a protein with unknown secondary structure composition.

2.2.3 Partial least squares

In partial least squares[36], the response matrix (spectra), **X**, is decomposed in a fashion similar to principal component regression, generating a matrix of scores and loadings or factors. A similar analysis is also carried out on the concentration matrix, **C**, producing separate scores and loadings matrices. PLS strives to model all the constituents forming **X** and **C** so that residuals for

X and **C** are approximately zero. An inner relationship is also constructed that relates the scores of **X** and the scores of **C**. The general underlying model of multivariate PLS is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (2\text{-}16)$$

$$\mathbf{C} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \qquad (2\text{-}17)$$

$$\mathbf{U} = \mathbf{T}\mathbf{W} \qquad (2\text{-}18)$$

$$\widehat{\boldsymbol{S}} = \boldsymbol{P}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P})^{-1}\boldsymbol{W}\boldsymbol{Q}^{\mathrm{T}} \qquad (2\text{-}19)$$

$$\widehat{\boldsymbol{C}} = \boldsymbol{X}\widehat{\boldsymbol{S}} \qquad (2\text{-}20)$$

where **X** and **C** are the spectral and concentration matrices respectively, **T** and **U** are the score matrices for **X** and **Y** respectively, **P** and **Q**, loading matrices for **X** and **Y** respectively, **E** and **F**, error matrices for **X** and **Y** respectively and $\widehat{\boldsymbol{S}}$ plays the role of the regression coefficients like in CLS and PCR.

2.2.4 Multivariate Curve Resolution –Alternating least squares

Multivariate curve resolution is a branch of multivariate data analysis that focuses on separating overlapped spectra. It is a bilinear method which resolves multicomponent mixtures into a simple model consisting of a composition-weighted sum of the signals of the pure compounds[37, 38]. Experiments presented in Chapters 3 and 4 use multivariate curve resolution techniques to separate DUVRR as well as CD spectra.

MCR-ALS is described by the model shown below.

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad (2\text{-}21)$$

where **X** is the original data matrix, **C** and $\mathbf{S}^{\mathrm{T}}$ are the matrices containing the pure-component profiles related to the data variation in **X**, i.e., concentration and spectra respectively and **E** is the error matrix.

In MCR-ALS, the **C** and the $\mathbf{S}^T$ matrix have equal priority and both are optimized in iterative cycles. First the number of compounds/components in **X** are determined either by principal component analysis or already known from experience. Then initial estimates of the **C** matrix are calculated. Using the estimate of **C**, the $\mathbf{S}^T$ matrix is calculated under appropriately chosen constraints and then using the $\mathbf{S}^T$ estimates, the **C** matrix is again calculated under appropriately chosen constraints. From the product of **C** and $\mathbf{S}^T$ found in the above steps of an iterative cycle, an estimate of the original data matrix, **X**, is calculated. The steps are repeated until the matrix of residuals or error matrix becomes very small.

2.3 Principal component analysis

Some multivariate techniques require that the data be partitioned into a number of principal components (PCs) before further processing. The PCs are orthogonal to each other and represent a reduced dimensionality of a data matrix that describes most, but usually not all, of the variance displayed in the original data matrix. The first PC describes, as a vector in terms of the factor space, the direction of greatest variance in the data, and the second PC describes the direction of the next greatest variance in the data, and continues through to all PCs contained in the data. Any physical components whose contributions to the variance of the data matrix are correlated will be contained in one PC, and several PCs may be present within a data matrix. Non-principal components, those components that do not substantially contribute to the observed variance of the data set, typically describe noise in the data.

One way to find the number of PCs in a data matrix, **X**, is singular value decomposition (SVD), such that:

$$\mathbf{X} = \mathbf{U}\sum\mathbf{V}^T \qquad (2\text{-}22)$$

where **U** may be referred to as the row-singular vectors, **Σ** is a diagonal matrix that describes the size of (variance described by) each component in **U** and **V**, **V** may be referred to as the column-singular vectors, and **U** and **V** are orthogonal matrices (orthogonal meaning $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$, and similarly for **V**).

SVD is a preferred method because it is applicable to non-square matrices of data. For any non-square matrix **X**, the total number of components is equal to the smallest side of matrix **X**. The PCs of matrix X are selected by the analyst from the total number of components, and may be determined through several graphical means, such as a y-axis log plot of **Σ** values to determine component contribution to overall data matrix variance. Non-graphical methods may simply report the percentage of variance explained by each component, and the analyst must determine when a significant contribution to the data variance may possibly be a PC. In the y-axis log plot of **Σ** values, values that substantially differ from subsequent values contain more of the data variance and thus are chosen as PCs; often there is a "bend" or "elbow" in the plot that is used to initially guess at the number of PCs.

2.4 Multiway data analysis

The models considered so far (CLS, PLS and MCR-ALS) are useful to decompose two-way data, i.e., a data matrix, **X**, consisting of *I* rows and *J* columns. Typically, a set of *I* samples with their DUVRR spectra measured at *J* excitation wavelength provide this kind of data. If the DUVRR spectra of the *I* samples are measured at multiple excitation wavelength, *K*, the dimensions of the resulting data matrix, $\underline{\mathbf{X}}$, are now *I* X *J* X *K* representing a three-way array. Direct analysis of a three-way data array or trilinear data set as described is feasible by parallel factor analysis (PARAFAC).

2.4.1 Parallel factor analysis

The PARAFAC model was developed in 1970 and has since been increasingly used in chemometrics[39-44]. It has gained increasing attention in chemometrics due to its structural resemblance with many physical models of common instrumental data and its unique ability to decompose trilinear data into individual contributing components[45]. The PARAFAC model for an element x_{ijk} of a three-way array $\underline{\mathbf{X}}$ with dimensions I X J X K is as follows:

$$x_{ijk} = \sum_{n=1}^{N} a_{in}\, b_{jn} c_{kn} + e_{ijk} \qquad (2\text{-}23)$$

where a_{in}, b_{jn} and c_{kn} are the elements of the matrices of modes A, B and C, e_{ijk} represents the residual and N is the number of factors. The PARAFAC model can also be represented graphically as illustrated in Figure 2.2.

2.5 Summary

Linear algebra is a basic part of multivariate analysis, making a comprehension of linear algebra methods necessary in order to properly apply many multivariate analysis techniques. A brief survey of techniques and terminology in linear algebra and multivariate analysis has been presented. This Chapter serves as an introduction to those multivariate analysis methods utilized in Chapters 3, 4 and 5, though applications beyond ultraviolet resonance Raman spectroscopy and circular dichroism exist.

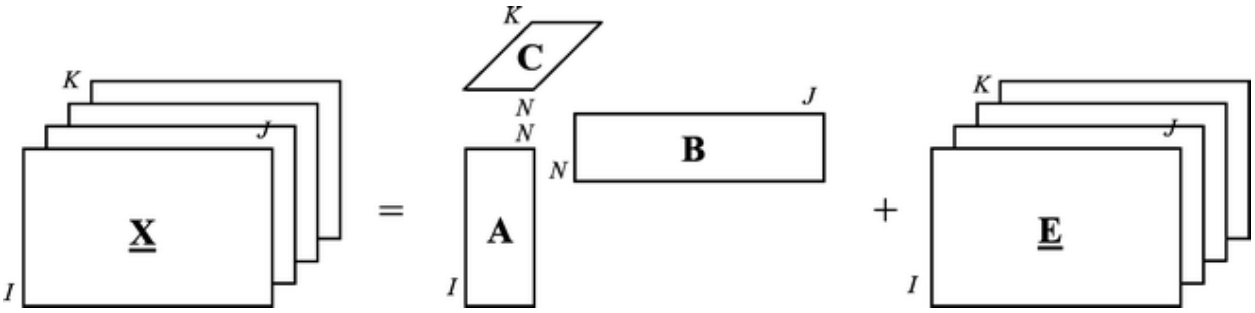

Fig. 2.2 The graphical representation of parallel factor analysis (PARAFAC) with N factors.

2.6 References

1. R. G. Brereton, *Chemometric: data analysis for the laboratory and chemical plant*, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

2. D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Sysems Theory*, Princeton University Press, Princeton, NJ, 2005.

3. S. Lipschutz and M. Lipson, in *fourth ed.*, McGraw-Hill Companies, Inc, New York, NY, 2009.

4. R. Bhatia, *Matrix Analysis*, Springer-Verlag New York, Inc., New York, NY, 1997.

5. K. R. Beebe, R. J. Pell and M. B. Seasholtz, *Chemometrics: A Practical Guide*, Wiley, New York, NY, 1998.

6. B. N. Datta, *Numerical Linear Algebra and Applications*, second edn., The Society for Industrial and Applied Mathematics, Philadelphia, PA, 2010.

7. S. Roman, *Advanced Linear Algebra*, Second edn., Springer Science + Business Media, Inc., New York, NY, 2005.

8. S. H. C. Du Toit, A. G. W. Steyn and R. H. Stumpf, *Graphical Exploratory Data Analysis*, Springer-Verlag, New York, NY, 1986.

9. E. R. Malinowski, *Factor Analysis in Chemistry*, John Wiley & Sons, New York, NY, 2002.

10. R. A. Usmani, *Applied Linear Algebra*, Marcel Dekker, Inc, New York, NY, 1987.

11. A. Bryman, *Quantitative Data Analysis with Minitab: A Guide to Social Scientists*, Routledge, London, UK, 1996.

12. S. L. Bieber and D. V. Smith, *Chemical Senses*, 1986, **11**, 19-47.

13. S. Geisser, *The Annals of Mathematical Statistics*, 1965, **36**, 150-159.

14. G. G. Koch, D. H. Freeman, Jr. and J. L. Freeman, *International Statistical Review / Revue Internationale de Statistique*, 1975, **43**, 59-78.

15. B. R. Kowalski, *Journal of Chemical Information and Computer Sciences*, 1975, **15**, 201-203.

16. P. Nomikos and J. F. MacGregor, *Technometrics*, 1995, **37**, 41-59.

17. K. C. S. Pillai, *The Annals of Mathematical Statistics*, 1955, **26**, 117-121.

18. C. R. Rao, *Biometrika*, 1948, **35**, 58-79.

19. P. R. Rider, *Econometrica* 1936, **4**, 264-268.

20. J. Stoehlmacher, D. J. Park, W. Zhang, D. Yang, S. Groshen, S. Zahedy and H. J. Lenz, *British journal of cancer*, 2004, **91**, 344-354.

21. I. E. Frank and J. H. Friedman, *Technometrics*, 1993, **35**, 109-135.

22. P. C. Maria, J. F. Gal, J. De Franceschi and E. Fargin, *Journal of the American Chemical Society*, 1987, **109**, 483-492.

23. J. Trygg, E. Holmes and T. Lundstedt, *Journal of proteome research*, 2007, **6**, 469-479.

24. C. Weihs, *Journal of Chemometrics*, 1993, **7**, 305-340.

25. V. A. Shashilov, V. Sikirzhytski, L. A. Popova and I. K. Lednev, *Methods (Amsterdam, Neth.)*, 2010, **52**, 23-37.

26. T. Naes, T. Isaksson and B. Kowalski, *Analytical Chemistry*, 1990, **62**, 664-673.

27. S. Wold, N. Kettaneh-Wold and B. Skagerberg, *Chemometrics and Intelligent Laboratory Systems*, 1989, **7**, 53-65.

28. N. B. Vogt, *Chemometrics and Intelligent Laboratory Systems*, 1989, **7**, 119-130.

29. H. A. L. Kiers, *Journal of Chemometrics*, 2000, **14**, 105-122.

30. R. Tauler and B. Kowalski, *J. Chemom.*, 1995, **9**, 31-58.

31. R. Tauler, *Chemom. Intell. Lab. Syst.*, 1995, **30**, 133-146.

32. J. A. de and R. Tauler, *Anal. Chim. Acta*, 2003, **500**, 195-210.

33. V. A. Shashilov and I. K. Lednev, *Chem. Rev. (Washington, DC, U. S.)*, 2010, **110**, 5692-5713.

34. Z. Chi, X. G. Chen, J. S. W. Holtz and S. A. Asher, *Biochemistry*, 1998, **37**, 2854-2864.

35. M. Otto, *Chemometrics: Statistics and computer application in analytical chemistry*, Wiley - VCH, 2007.

36. P. Geladi and B. R. Kowalski, *Analytica Chimica Acta*, 1986, **185**, 1-17.

37. R. Tauler, *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**, 133-146.

38. A. De Juan and R. Tauler, *Analytica Chimica Acta*, 2003, **500**, 195-210.

39. D. Baunsgaard, L. Munck and L. Nørgaard, *Applied Spectroscopy*, 2000, **54**, 1684-1689.

40. S. Bijlsma, D. J. Louwerse and A. K. Smilde, *Journal of Chemometrics*, 1999, **13**, 311-329.

41. R. Bro and H. Heimdal, *Chemometrics and Intelligent Laboratory Systems*, 1996, **34**, 85-102.

42. W. P. Gardner, R. E. Shaffer, J. E. Girard and J. H. Callahan, *Analytical Chemistry*, 2001, **73**, 596-605.

43. R. D. Jiji, G. G. Andersson and K. S. Booksh, *Journal of Chemometrics*, 2000, **14**, 171-185.

44. A. Marcos, M. Foulkes and S. J. Hill, *Journal of analytical atomic spectrometry*, 2001, **16**, 105-114.

45. N. D. Sidiropoulos and R. Bro, *Journal of Chemometrics*, 2000, **14**, 229-239.

Chapter 3 - Quantification of Protein Secondary Structure Content by Multivariate Analysis of Deep-Ultraviolet Resonance Raman and Circular Dichroism Spectroscopies[1]

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. 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] Adapted with permission from Oshokoya, O. O., Roach, C. A., JiJi, R. D.; Quantification of Protein Secondary Structure Content by Multivariate Analysis of Deep- Ultraviolet Resonance Raman and Circular Dichroism Spectroscopies. *Analytical Methods.* 2014, 6 (6), 1691-1699. Copyright 2014 Royal Society of Chemistry.

3.1 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 3.1A)[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 3.1B) 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_α, f_β and f_τ) of each secondary structure type;

$$\mathbf{S} = f_\alpha \mathbf{s}_\alpha + f_\beta \mathbf{s}_\beta + f_\tau \mathbf{s}_\tau \qquad (3\text{-}1)$$

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 3-2:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad (3\text{-}2)$$

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].

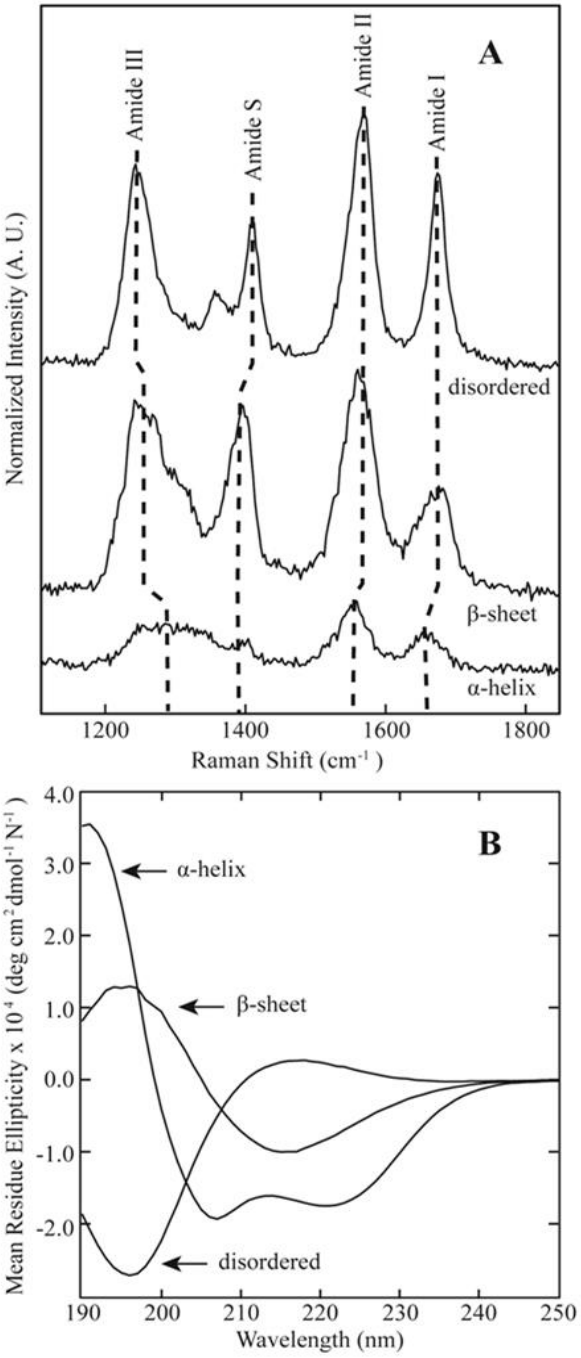

Fig. 3.1 UVRR (A) and CD (B) spectra of poly-L-lysine in α-helix (25°C, pH 11.0), β-sheet (52°C, pH 11.3) and disordered (25°C, pH 4.0) conformations.

3.2 Experimental

3.2.1 Sample Preparation

Nine globular proteins with varying secondary structure content (Table 3-1), 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 3-1 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

3.2.2 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.

3.2.3 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 3.2). 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.

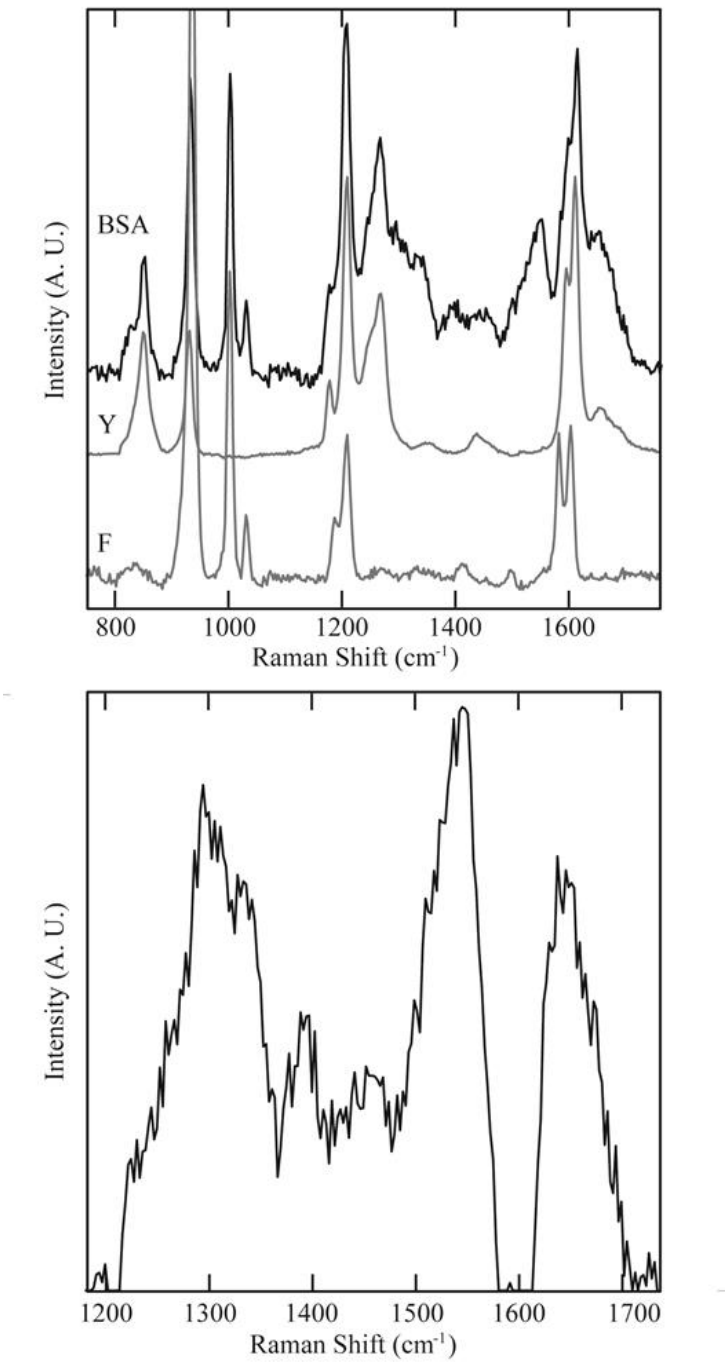

Fig. 3.2 Top: BSA, phenylalanine and tyrosine UVRR spectra. Phenylalanine and tyrosine spectra are scaled to the bands at 1003 cm^{-1} (F12) and 853 cm^{-1} (Y1), respectively. Bottom: BSA spectrum with phenyalanine and tyrosine contributions subtracted.

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)

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

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 3.3) 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 3.3 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 (Figure 3.3). Figure 3.3 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.

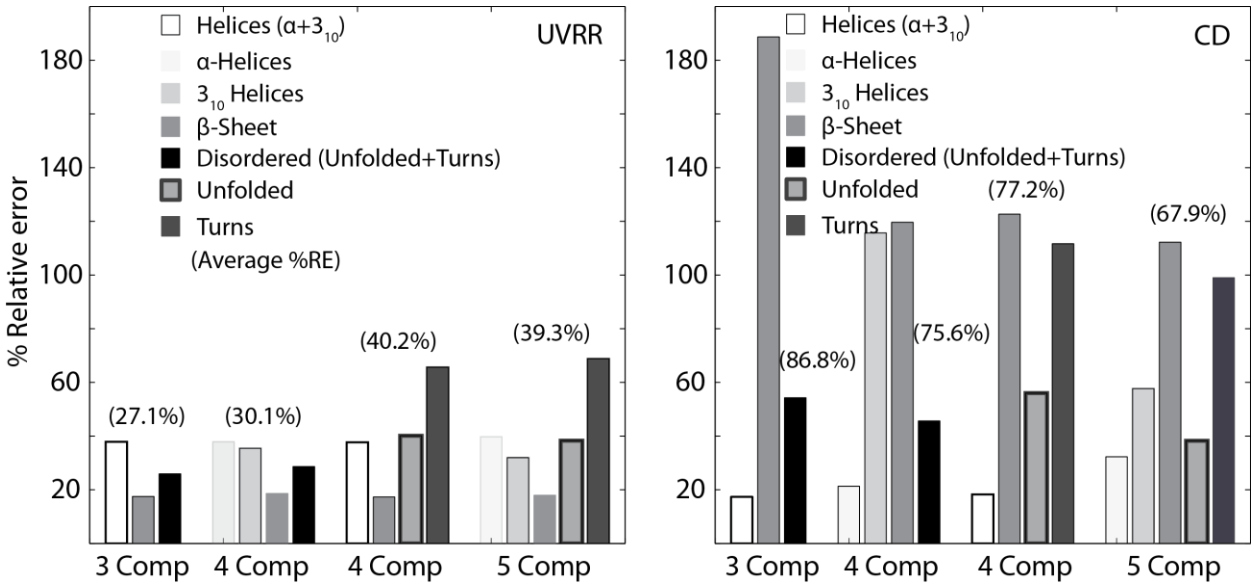

Fig. 3.3 %RE of the different considered components in each model for both UVRR and CD.

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:

$$\mathrm{RMSECV} = \left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \qquad (3\text{-}4)$$

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:

$$\mathbf{S} = \mathbf{X}\mathbf{C}^{+} \qquad (3\text{-}5)$$

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:

$$\mathbf{C}_{\mathrm{MCR}} = \mathbf{X}\mathbf{S}_{\mathrm{MCR}}^{+} \qquad (3\text{-}6)$$

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

$$\mathbf{C} = \mathbf{C}_{MCR}\mathbf{W} \qquad (3\text{-}7)$$

$$\mathbf{S} = \mathbf{S}_{MCR}\mathbf{W}^{-1} \qquad (3\text{-}8)$$

Such that:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{T} = \mathbf{C}_{MCR}\mathbf{W}(\mathbf{S}_{MCR}\mathbf{W}^{-1}) = \mathbf{C}_{MCR}\mathbf{S}_{MCR} \qquad (3\text{-}9)$$

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 3-7 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.

3.3 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 3-2). 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 3.5 and 3.7). 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.

3.3.1 Results for UVRR

Overall, each multivariate method performed similarly with an average prediction error of approximately 10% (Table 3-2). 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 3.4). 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 3.4) of those same proteins were over estimated. The common factor between lysozyme and cytochrome c is an absence of β-sheet structure.

Table 3-2 RMSECV (%) values calculated

UVRR				
Algorithm	helix	β-sheet	Disordered	Average
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

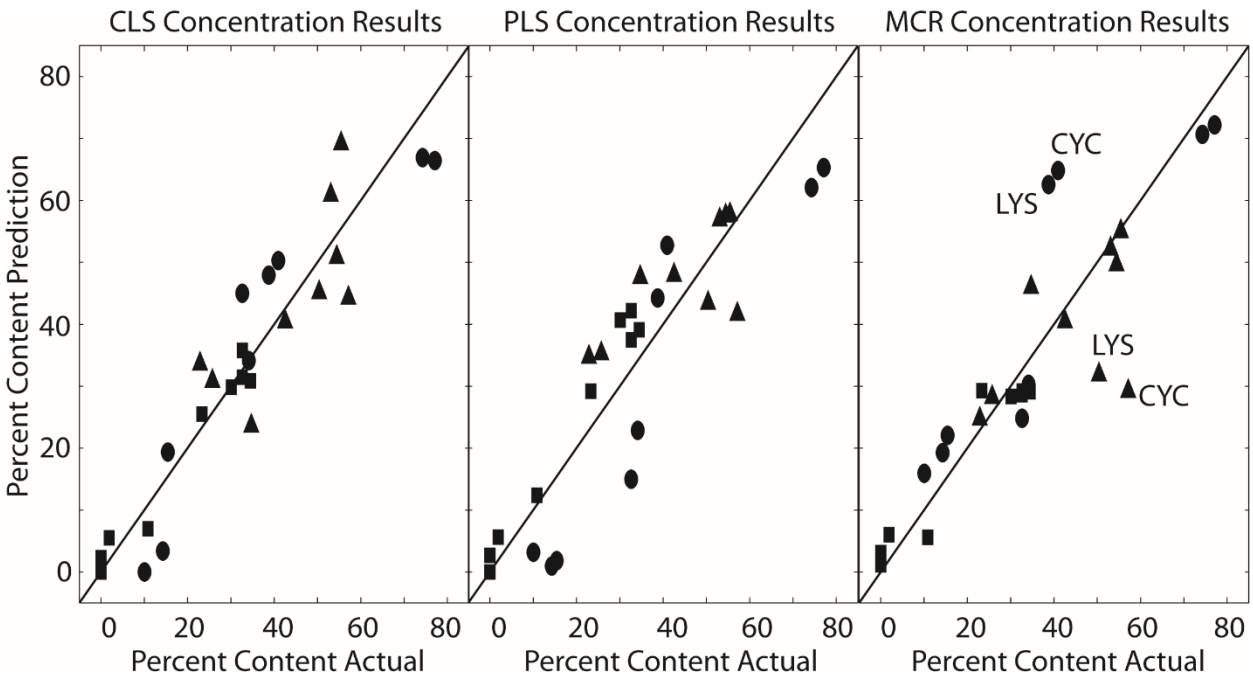

Fig. 3.4 The actual versus predicted percentage composition for UVRR of α-helical (circles), β-sheet (squares), and disordered (triangles) structures as a percentage of content. The (1,1) line is shown to illustrate the deviations in the prediction.

Figure 3.5 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 3-3) 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.

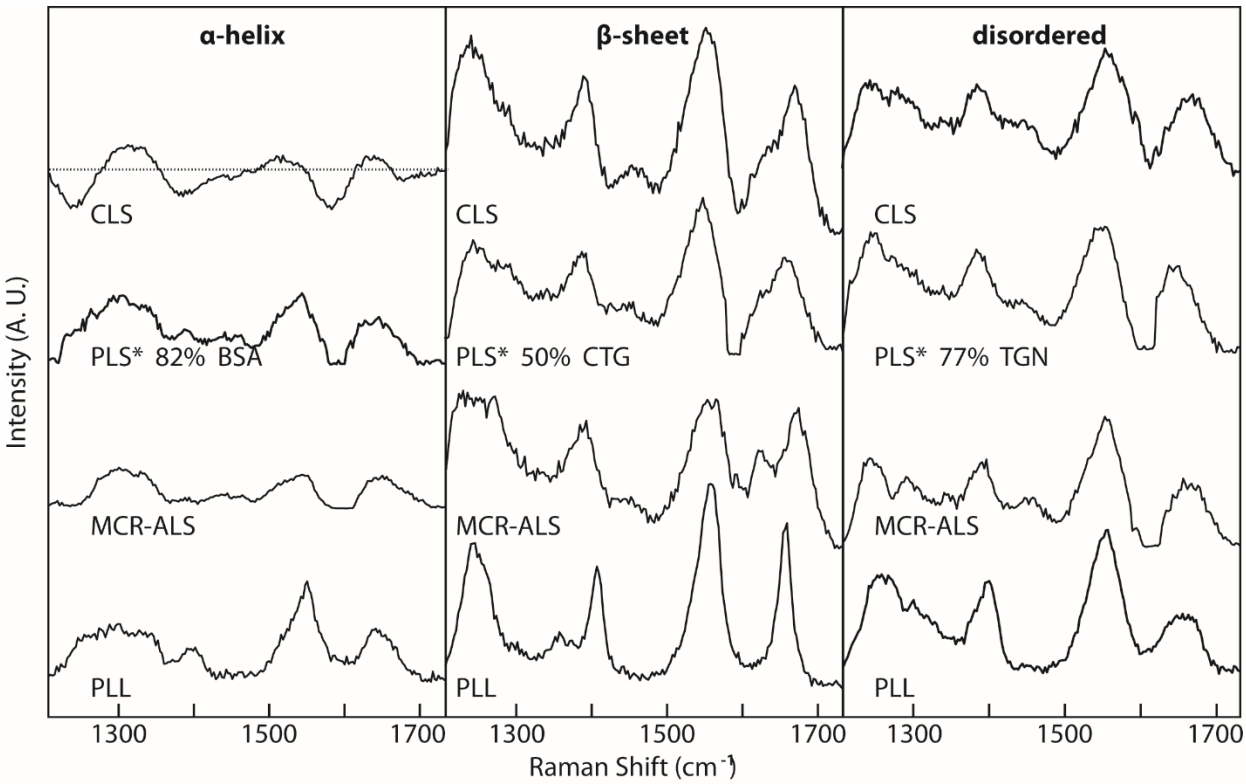

Fig. 3.5 The PSSRS obtained from the various methods compared to the poly-L-lysine (PLL) pure conformer spectra. The dotted line is used to indicate the zero line for the spectrum that has a negative region. PLS does not produce PSSRS, so the largest predicted content for each structure type obtained during the iterative calculations is presented along with the protein spectrum (designated by the 3 letter abbreviation) associated with the prediction.

Table 3-3 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

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 3-3). 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}.

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 3.3). The predicted amide I and II bands also occur in the expected experimental amide I (1548–1561cm^{-1}) and amide II (1661–1682 cm^{-1}) regions[8].

3.3.2 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

3-2). 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 3.6). 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 3-2). While normalization improves prediction of β-sheet and disordered contents, it appears to degrade the prediction of α-helical content from CD spectra.

The resolved pure CD spectra from CLS and MCR are shown in Figure 3.7. As mentioned previously, the PLS algorithm does not produce resolved pure spectra.

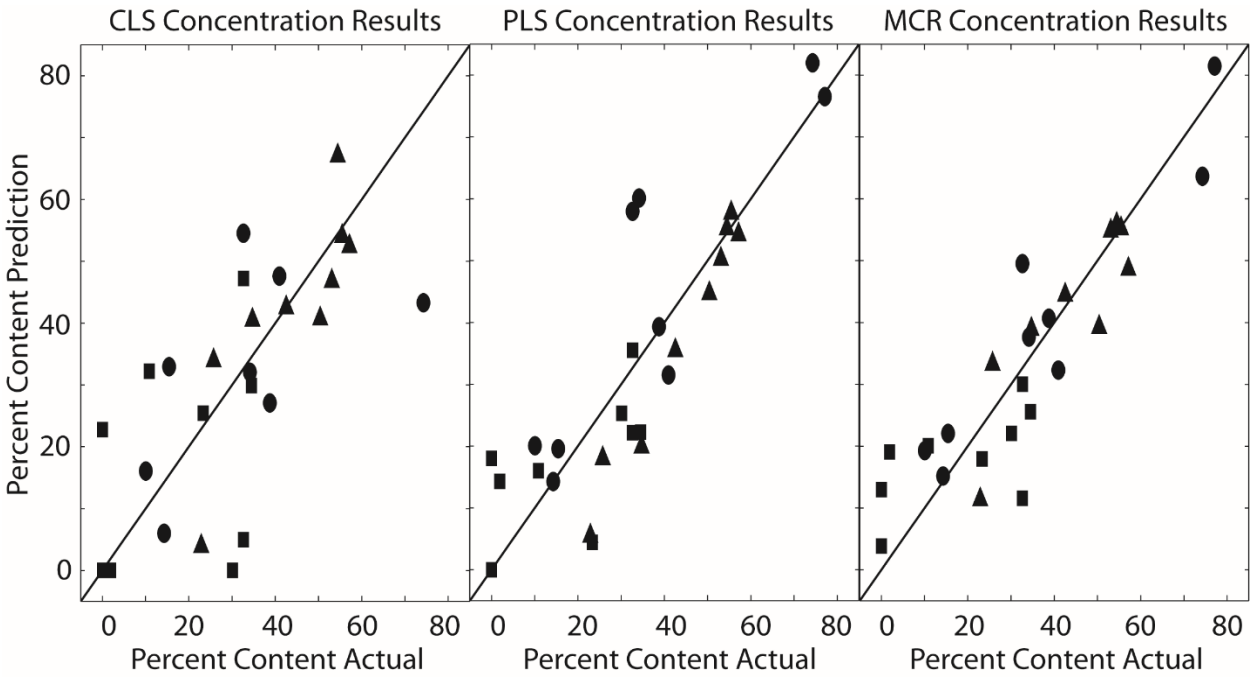

Fig. 3.6 The predicted versus actual percent composition of secondary structure from CD analysis of helical (circles), β-sheet (squares), and disordered (triangles) structures as a percentage of protein content. The (1,1) line is shown to illustrate the deviations in the prediction.

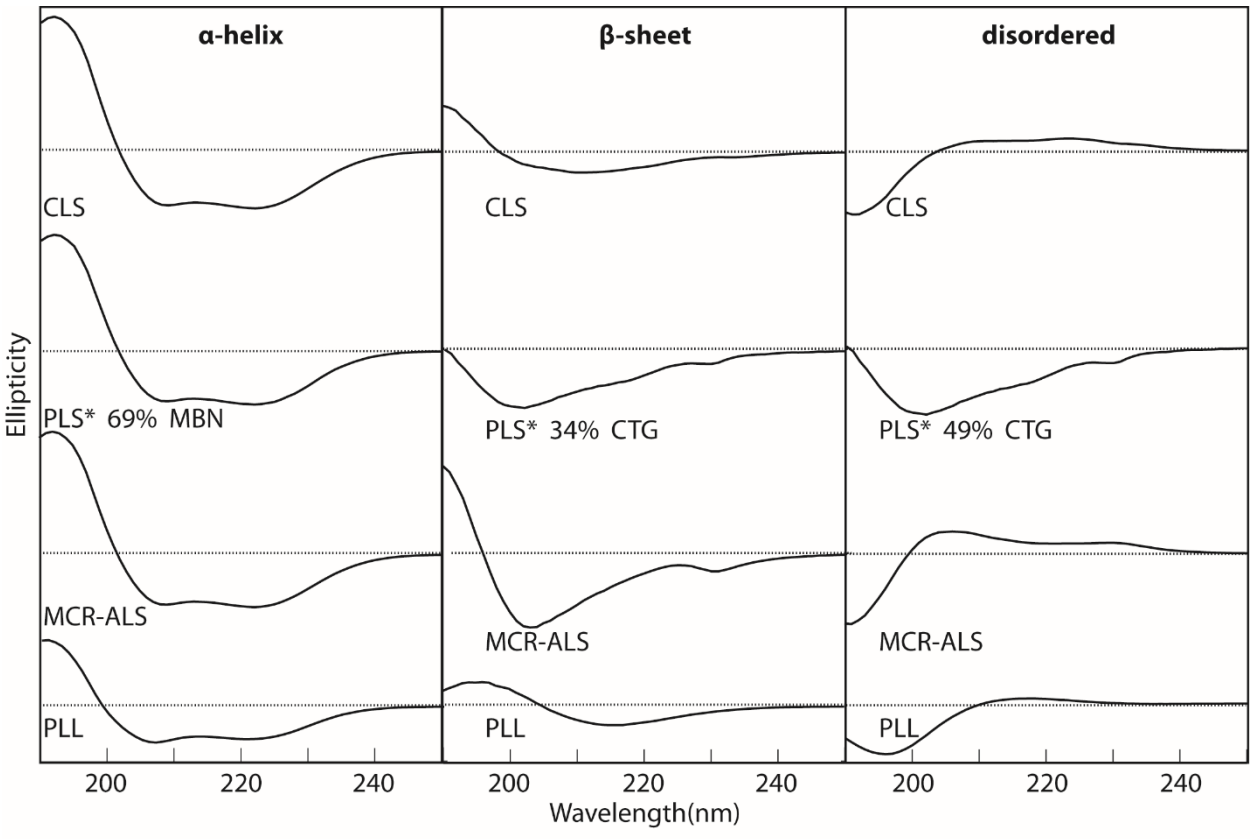

Fig. 3.7 The CD pure spectra obtained from the various methods compared to the poly-L-lysine (PLL) pure conformer spectra. The dotted line is used to indicate the zero line for the spectrum that has a negative region. PLS does not produce PSSRS, so the largest predicted content for each structure type obtained during the iterative calculations is presented along with the protein spectrum associated with the prediction.

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.

3.3.3 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_α) and β-sheet (f_β) 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 3.8).

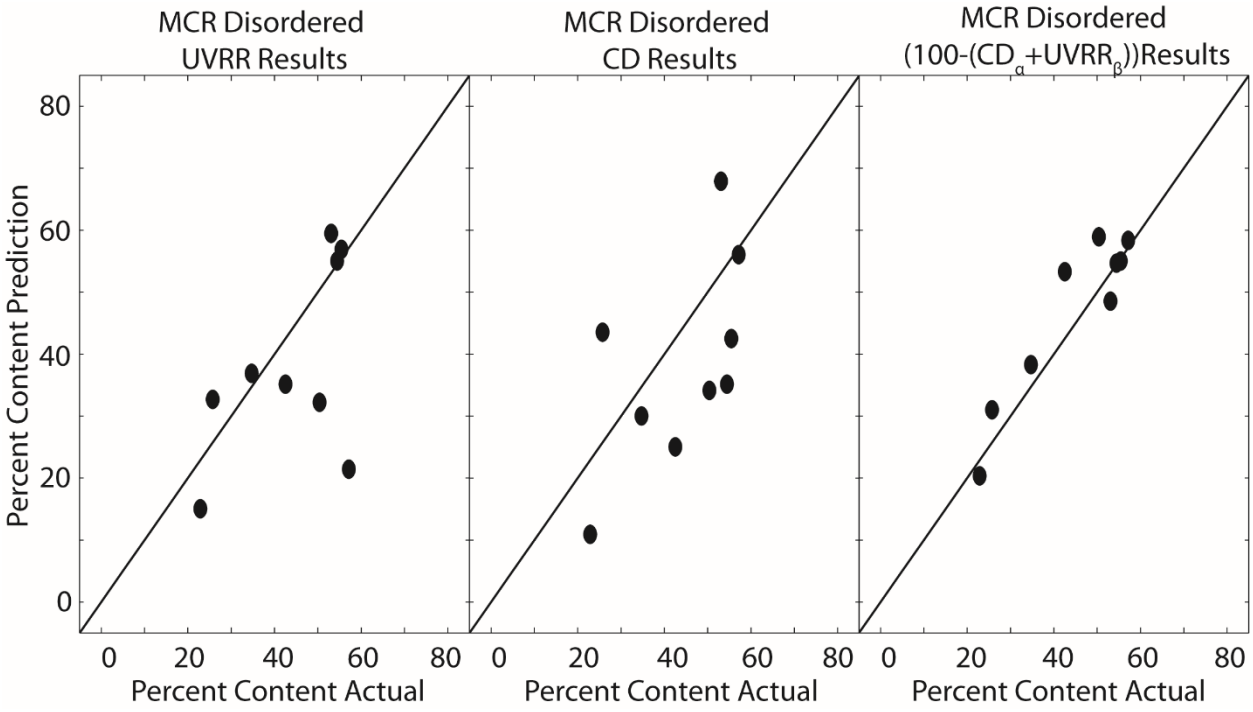

Fig. 3.8 The predicted versus actual percent composition of disordered secondary structure from UVRR analysis, CD analysis, and (100 - (CD_α + $UVRR_\beta$)). The (1,1) line is shown to illustrate the deviations in the prediction.

3.4 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.

3.5 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.

Chapter 4 - Fusing spectral data to improve protein secondary structure analysis: Data fusion

The determination of protein secondary structure has become an area of great significance as this knowledge is important for understanding relationships between protein structure and, more importantly, how the changes in structure affect function. Previous studies suggest that a complementary use of spectroscopic data from optical methods such as circular dichroism (CD), infrared (IR) and ultraviolet resonance Raman (UVRR) coupled with multivariate calibration techniques like multivariate curve resolution-alternating least squares (MCR-ALS) is the preferred route for real-time and accurate evaluation of protein secondary structure. This study presents a new strategy for the improvement of secondary structure determination of proteins by fusing CD and UVRR spectroscopic data. Also, a new method for determining the structural composition of each protein is employed, which is based on the relative abundance of the (φ,ψ) dihedral angles of the peptide backbone as they correspond to each type of secondary structure. Comparison of the predicted protein secondary structures from MCR-ALS analysis of CD, UVRR and fused data with definitions obtained from dihedral angles of the peptide backbone, yields lower overall root mean squared errors of calibration for helical, β-sheet, poly-proline II type and total unfolded secondary structures with fused data.

4.1 Introduction

Protein secondary structure quantification has become an area of intense biochemical and biophysical research due to the effects of secondary structure on tertiary and quaternary protein structure. There are four levels of protein structure and a change at any level can result in changes in protein function. The primary structure of a protein is the amino acid sequence while the secondary structure refers to the structural motifs within the protein that are defined by the phi (φ) and psi (ψ) dihedral angles of the amide backbone (Figure 4.1).

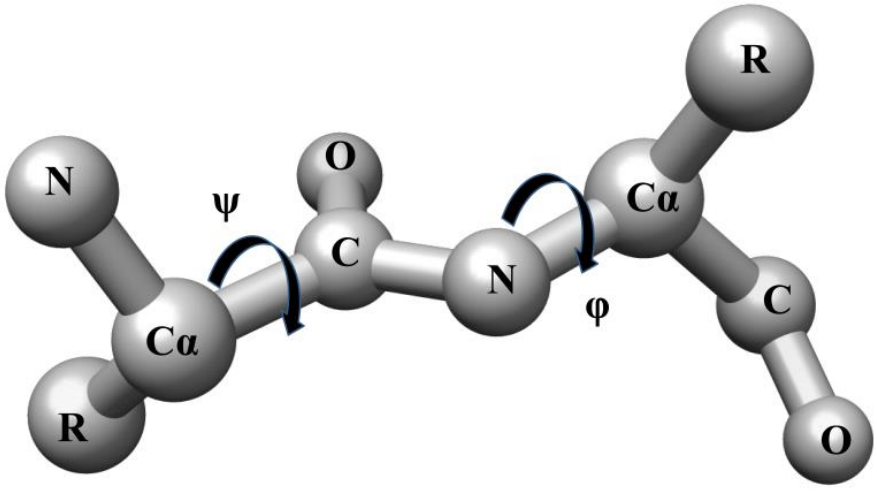

Fig. 4.1 Peptide backbone showing phi (φ) and psi (ψ) dihedral angles.

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.[1-6] Since secondary structure plays a significant role in protein function and select diseases, it is therefore of substantial interest to rapidly and accurately quantify protein secondary structure especially in an environment that mimics physiological conditions. Traditional methods of protein secondary structure quantification such as x-ray crystallography (XRC)[7], nuclear magnetic resonance (NMR)[8] and circular dichroism (CD)[9-11] are now complimented by a host of vibrational methods, in particular, ultraviolet resonance Raman (UVRR) spectroscopy which has proven useful due to its structural sensitivity to the amide backbone. CD is the current standard in secondary structure analysis of proteins and UVRR is an up-and-coming technique.[12]

Previous studies show that multivariate analysis of CD and UVRR data results in relatively accurate prediction of helical (α- (-57°, -47°), + 3_{10}- (-49°, -26°)) and β-sheet (anti parallel (-139°, 135°), + parallel (-119°, 113°)) secondary structures in proteins, respectively, but relatively poor prediction of the other secondary structures. [13] This is because α-helical secondary structure has the largest relative signal intensity in CD spectra of proteins whilst β-sheet secondary structure has the highest relative signal intensity in UVRR spectra of proteins. Combining the predicted amounts of helical and β-sheet contents from CD and UVRR enables a more accurate estimation of the disordered content, thus, more accurate predictions of secondary structure content.[13]

In this study, we describe a new addition to the toolbox for protein secondary structure determination by taking advantage of the partial selectivity's of both CD and UVRR spectroscopies. The best estimates by MCR-ALS analysis are achieved with fused data from both spectroscopic techniques. Data fusion refers to methods that combine multiple data types into a

single data array, with the expectation that the resulting fused data will be more informative than the individual input sources.[14-17] Generally, performing data fusion offers advantages which include improved detection, confidence and reliability.[18-24] Data fusion can be executed in one of three fashions; data level fusion, where the raw data generated by multiple sources are combined directly, or after appropriate normalization has been carried out so that the data are commensurate; feature-level fusion, where feature extraction methods are used to generate representations of the raw data which are then combined; and decision level fusion which involves combining decisions that have been arrived at independently by the available sources.[17] In this study, we utilize data level fusion, fusing the raw or preprocessed UVRR and CD data before any other analysis is carried out.

We have compared different preprocessing methods for the fused data to determine which method improves protein secondary structure prediction. We have also defined the structural classifications of secondary structure based on the relative distribution of (φ,ψ) dihedral angles of the amide backbone in each protein. We show that by redefining secondary structure based on dihedral angles and application of data fusion to CD and UVRR spectroscopic data, we can improve the determination of not only the helical or β-sheet contents of proteins but also other secondary structures most notably the poly-proline II (PPII) type structure.

PPII-type structure was first identified by Tiffany and Krimm [25-27] in poly-L-lysine and poly-L-glutamic acid and has since been shown to be the predominant structure in unfolded or disordered protein regions. PPII-type structure has (-79°, 150°) dihedral angles and is stabilized by water hydrogen bonding with the peptide backbone. Unfortunately, this structure is not defined in the protein data bank and thus difficult to quantify and distinguish from other unfolded or less prevalent structures. Less prevalent structures include left handed α-helices (57°, 47°) and turns,

which typically make up less than 5% of the protein's secondary structure. Turns are more complicated as the (φ, ψ) dihedral angles are not repetitive and differ depending on the type of turn. Thus, for quantitative purposes, it makes more sense to define each protein's structural composition based on the abundance of (φ, ψ) dihedral angles.

4.2 Materials and Methods

4.2.1 Sample preparation

Nine globular proteins with varying secondary structure content (Figure 4.2), amino acids L-phenylalanine and L-tyrosine 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 solution (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 μM for amino acid solutions for UVRR analysis and 0.2 mg $ml^{-1}$for CD measurements. Protein coordinate files for the nine proteins were downloaded from the protein data bank, PDB (www.rscb.org)[28] and a dihedral angle calculator readily available online (http://cib.cf.ocha.ac.jp/bitool/DIHED2/)[29] was used to determine the relative abundance of the (φ, ψ) dihedral angles in each protein for secondary structure content distribution as displayed in Figure 4.2. 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.

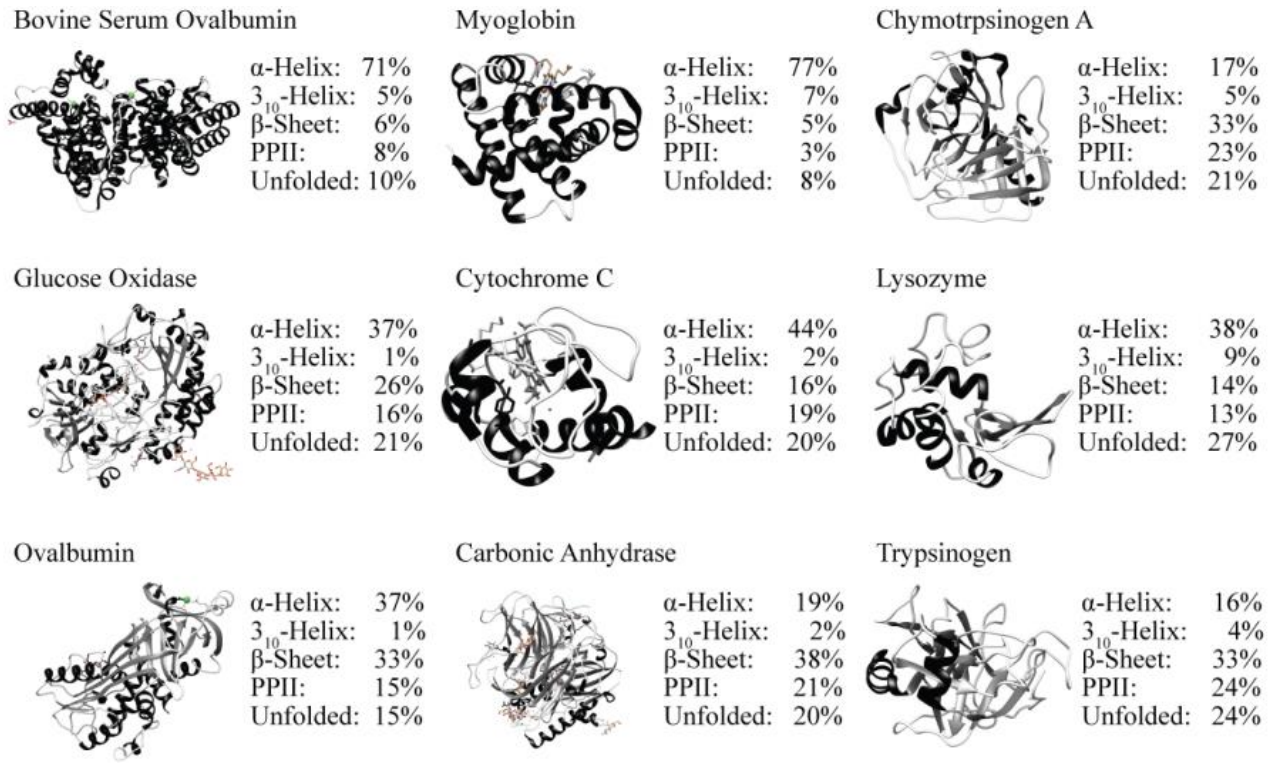

Fig. 4.2 Secondary structure content (%) of proteins used calculated from (φ, ψ) dihedral angles as found on the Research Collaboratory for Structural Bioinformatics (RSCB) Protein Data Bank.

4.2.2 UVRR Spectra acquisition

The UVRR instrument used to collect protein spectra has been previously described.[30] Briefly, the fourth harmonic of a tunable Ti:Sapphire laser (Coherent Inc., Santa Clara, CA) was employed to generate an excitation wavelength of 197 nm. The sample was circulated by a Minipuls2 peristaltic pump (Gilson Inc., Middleton, WI) 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 4°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 3 hours of signal collection. A small aliquot of a 1 M sodium perchlorate solution was added to each sample, for a final concentration of 200 mM, as an internal intensity standard. All spectra were collected in triplicate and calibrated using a standard cyclohexane spectrum.[31, 32]

4.2.3 CD spectra acquisition

All samples used for UVRR analysis were additionally measured for their corresponding CD spectra. An AVIV 62DS circular dichroism (Aviv Biomedical Inc., Lakewood Township, NJ) spectrometer and a quartz cell with a 1 mm optical path length (Hellman USA, Plainview, NY) were used to collect CD spectra. All spectra were collected between 190 and 250 nm with a resolution of 0.1 nm at room temperature. Every sample was measured five times with a scan speed of 1 nm/5 s and averaged. Each experiment was repeated in triplicate. Corresponding background spectra were collected in the same manner and subtracted from sample spectra**.**

4.2.4 Data processing

All data analyses were carried out in MATLAB (version 7.11, Mathworks, Natick MA). Cosmic rays in the UVRR spectra were removed using an in-house program,[33] and the spectra were base-lined using the MATLAB curve-fitting toolbox. Contributions to spectra from aromatic side chains were subtracted using the phenylalanine band at 1003 cm^{-1} (F12) and tyrosine band at 853 cm^{-1} (Y1) as previously described.[13] Contributions from tryptophan were disregarded due to its negligible intensity in deep-UVRR spectra (λ_{ex} <210 nm). Areas that appeared to be negative in the spectrum after subtraction of aromatic contribution were set to zero and each resulting spectrum truncated to the 1266–1759 cm^{-1} spectral range so that only the amide regions were used for modeling. For CD data, the mean residue ellipticiity (Θ_{MRE}) was calculated as previously described.[30]

A MCR-ALS algorithm was used based on that outlined by Bro and Sidiropoulous.[34] MCR-ALS was selected because on average it performed better in previous studies [13] compared to classical least squares and partial least squares for spectral resolution and secondary structure prediction.

For both UVRR and CD, the triplicate spectra were compiled to obtain 27 individual spectra (Figure 3). The UVRR and CD data were then fused according to the model in Figure 4 to give a single data matrix. To evaluate the potential predictive ability of the MCR-ALS models, the root mean squared error of calibration (RMSEC) was used (Equation 4-1).

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

In Equation 4-1, n is the number of samples, y_i is the abundance of each secondary structure element obtained from the (φ, ψ) dihedral angles as displayed in Figure 4.2 and $\hat{y}_i$ is the estimated

value obtained from least squares regression of the resolved composition profiles from the MCR-ALS algorithm.

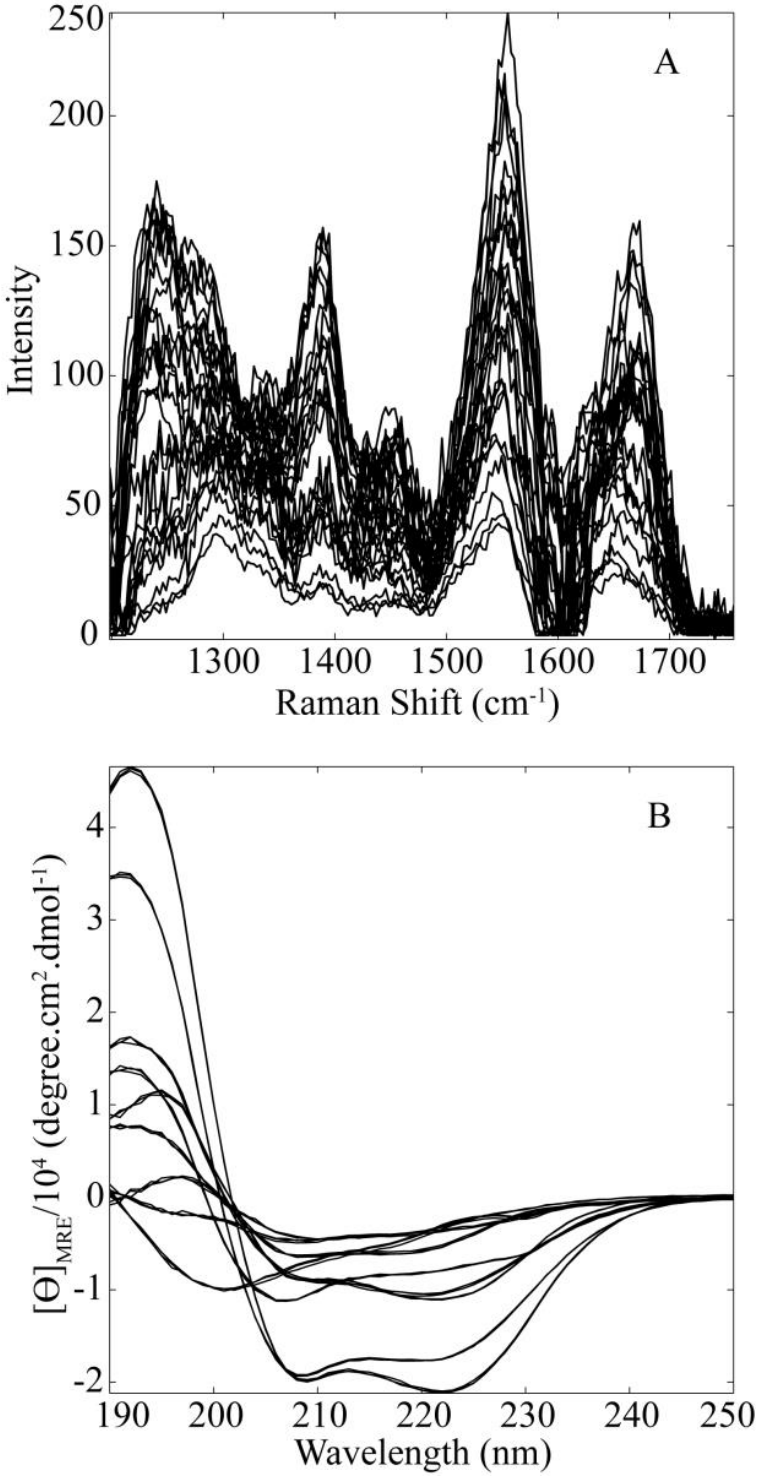

Fig. 4.3 UVRR (A) and CD (B) spectra of nine proteins used for multivariate analysis.

4.3 Results and Discussion

4.3.1 Protein secondary structure and UVRR and CD spectra

Ideally, proteins with similar secondary structure contents should have similar CD and UVRR spectra. However, while protein UVRR and CD spectra are highly reproducible, proteins with similar secondary structural content can have very different spectra. For instance, while carbonic anhydrase and chymotrypsinogen A have similar secondary structure distributions with high β-sheet and relatively low helical contents (Figure 4.4), there is a clear difference in their CD spectra but their UVRR spectra are overlapped. Bovine serum albumin and myoglobin also have similar secondary structure distributions but with high helical contents and no β-sheet structure; the CD spectra for both proteins are quite similar but in this case their UVRR spectra, while similar in shape are clearly differing in overall intensity.

It can be concluded that greater differences in the measured spectra of proteins with similar structural compositions will be observed when the dominant secondary structure type has a low relative signal intensity as compared to the other types of secondary structure. Therefore, poorer prediction of these structures is almost certain if only one technique is employed. In order to take advantage of the predictive capabilities of each technique (CD and UVRR), a data fusion approach was employed.

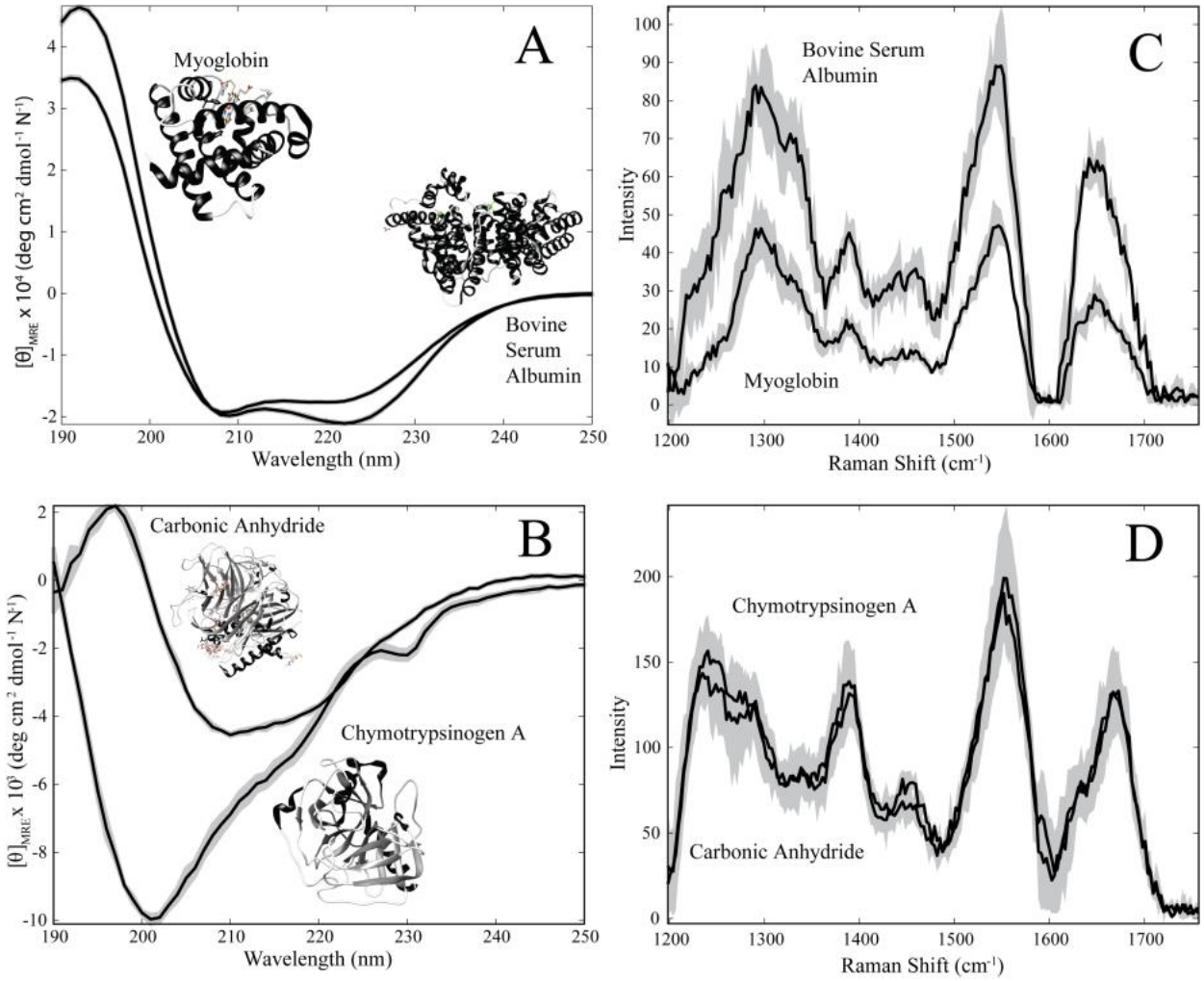

Fig. 4.4 CD (A and B) and UVRR (C and D) spectra of proteins with similar secondary structure compositions. The shaded gray area about the lines represent the standard deviation of three measurements at each variable.

4.3.2 Effect of preprocessing on estimation of composition profiles

The UVRR and CD data were fused according to the model in Figure 4.5 to yield a single data matrix. MCR-ALS was employed to resolve the underlying compositional and spectral profiles prior to preprocessing, and after normalization, auto scaling and variance scaling (Figure 4.6).

A 4-component model was employed because a 5-component model resulted in poorer predictions of the three most prominent structures; helical, β-sheet and PP-II and did not enable resolution of α- and 3_{10}-helical structure or parallel and antiparallel structures. The composition profiles from MCR-ALS analysis assigned to (helical (α- + 3_{10}-helices), β-sheet/strand, PP-II and unfolded (everything else)) were regressed using the secondary structure compositions obtained from the relative abundance of the (φ,ψ) dihedral angles of the peptide backbone for each protein (Figure 4.2). The resultant regression model was then used to re-predict secondary structure of the test protein samples.

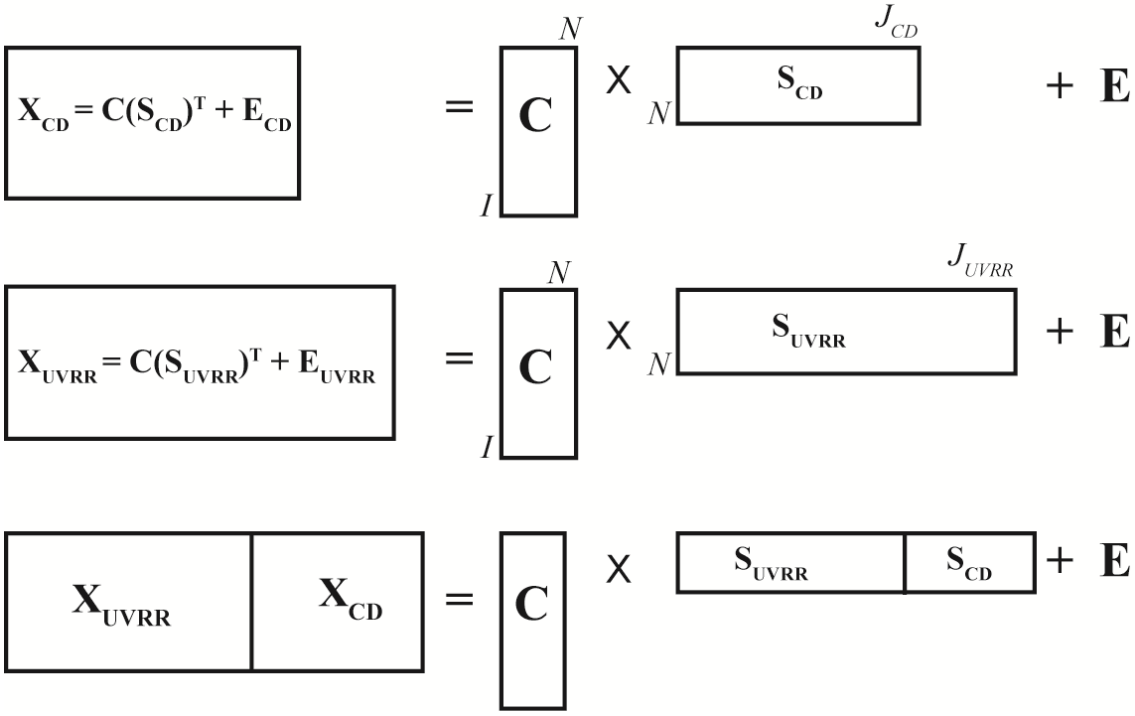

Fig. 4.5 Data fusion model for multivariate analysis for protein secondary structure determination.

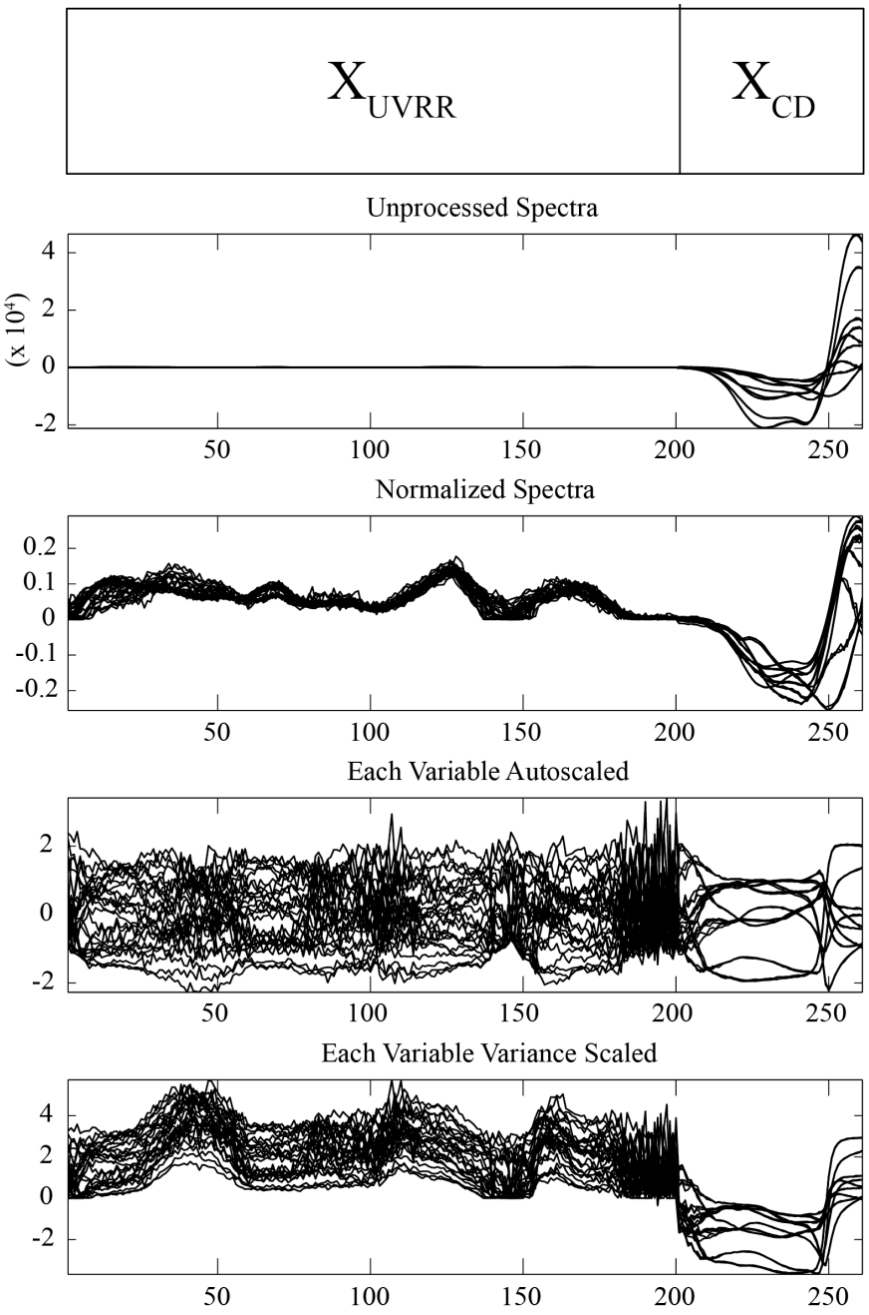

Figure 4.6 Fused CD and UVRR spectra after application of each preprocessing method.

Prediction accuracies of about 5% can be achieved for either helical content using CD or β-sheet content using UVRR (Figure 4.7). A comparison of the RMSEC values versus each preprocessing method is summarized in Table 1. When no preprocessing was employed, the results were similar to the use of CD data alone (Table 4-1 and Figure 4.7). RMSEC of β-sheet is high when no preprocessing is employed because the greater intensity of the CD spectra has a greater influence on the model (Figure 4.6), and the UVRR information is lost. Auto scaling of the fused data resulted in significantly higher RMSEC's for all secondary structure types. Normalization to unit variance did not improve prediction of β-sheet structure (RMSEC = 40%) and appeared to worsen prediction of helical structure, essentially doubling the RMSEC. Ultimately variance scaling improved prediction of β-sheet structure (RMSEC = 5.4%) without much penalty to the prediction of the other structures (Table 4-1 and Figure 4.7).

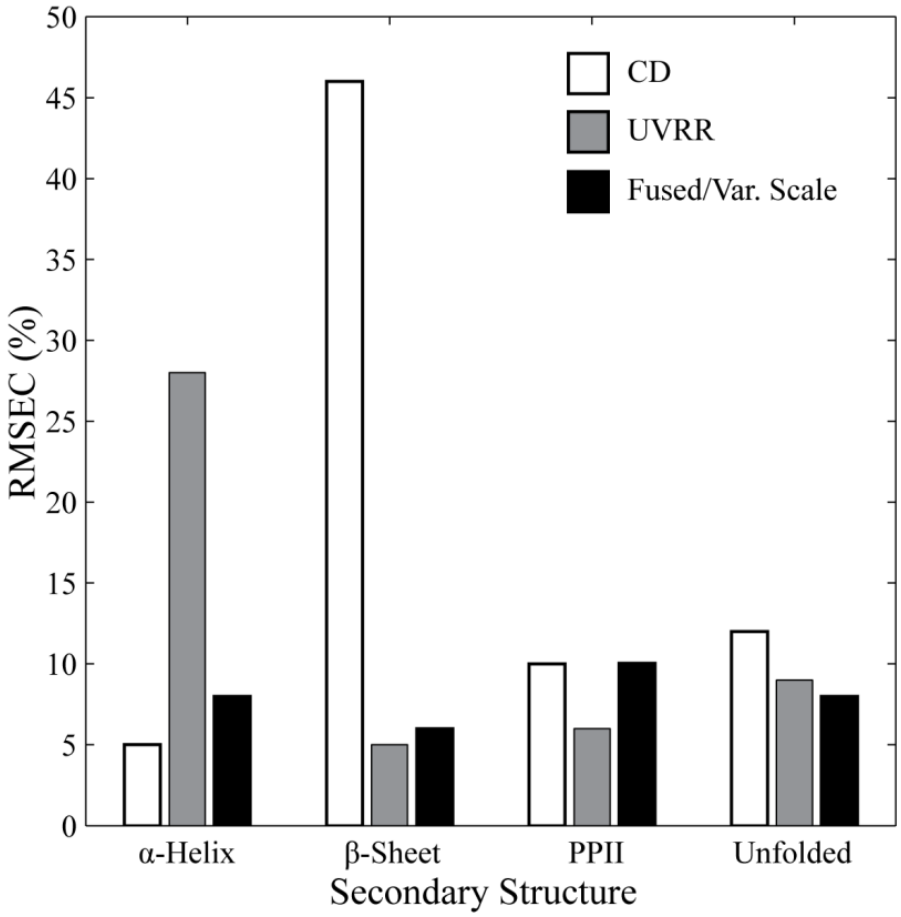

Figure 4.7 Root mean square error of calibration (RMSEC) for prediction of protein secondary structure using CD, UVRR and fused CD-UVRR spectroscopic data.

Table 4-1 Root mean square error of calibration (RMSEC) of MCR-ALS model employing different preprocessing methods

Pre-processing method	Helix	Sheet	PPII	Unfolded
Unprocessed	5.7%	54.2%	8.7%	11.0%
Normalized	12.9%	40.2%	3.7%	8.0%
Auto-scaled	296.4%	57.0%	82.3%	27.1%
Variance-scaled	6.6%	5.4%	10.7%	8.6%

4.4 Conclusions

In this work, a new approach to protein secondary structure determination by applying multivariate analysis to fused spectroscopic data was developed. The advantage to this approach where CD and UVRR data are fused over individual analysis of both spectroscopic methods is that we can exploit the selective predictive capabilities of each technique (helical structure for CD and β-sheet structure for UVRR) and further improve predictions of other secondary structures including the PPII-type structure. We have also demonstrated that the most appropriate preprocessing method prior to multivariate analysis is the variance scaling method.

While helical structure prediction is improved using multivariate analysis of the fused data, the limitation of separating α- and 3_{10}-helical structures still looms. This is because both structures have very similar spectra both in CD and UVRR hence making them statistically indistinguishable. Also, less prevalent structures like turns and α-L, which occur in very small quantities, are not yet quantifiable as their CD and UVRR spectra are not distinct enough. Expansion to include other structurally sensitive techniques such as Raman optical activity or vibrational circular dichroism may increase the number of quantifiable secondary structures.

4.5 References

1. C. C. Blake, M. J. Geisow, S. J. Oatley, B. Rerat and C. Rerat, *Journal of molecular biology*, **1978**, 121, 339-356.

2. E. Herczenik and M. F. B. G. Gebbink, *FASEB Journal*, **2008**, 22, 2115-2133.

3. A. Moglich, X. Yang, R. A. Ayers and K. Moffat, *Annual review of plant biology*, **2010**, 61, 21-47.

4. S. B. Prusiner, *Proceedings of the National Academy of Sciences*, **1998**, 95, 13363-13383.

5. C. Weissmann, *Nature Reviews Microbiology*, **2004**, 2, 861-871.

6. D. Voet and J. G. Voet, *Biochemistry*, 3rd edn., John Wiley & Sons, Inc., Hoboken, NJ, **2004**.

7. A. Higashiura, K. Ohta, M. Masaki, M. Sato, K. Inaka, H. Tanaka and A. Nakagawa, *Journal of Synchrotron Radiation*, **2013**, 20, 989-993.

8. F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein and H. Oschkinat, *Nature*, **2002**, 420, 98-102.

9. N. J. Greenfield, *Analytical Biochemistry*, **1996**, 235, 1-10.

10. N. J. Greenfield, *Nature Protocols*, **2006**, 1, 2876-2890.

11. N. J. Greenfield and G. D. Fasman, *Biochemistry*, **1969**, 8, 4108-4116.

12. C. A. Roach, J. V. Simpson and R. D. Jiji, *Analyst*, **2012**, 137, 555-562.

13. O. O. Oshokoya, C. A. Roach and R. D. Jiji, *Anal. Methods*, **2014**, 6, 1691-1699.

14. David Lee Hall and S. A. H. McMullen, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA, **2004**.

15. Martin Liggins II, David Hall and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*, CRC Press, Boca Rotan,FL, **2008**.

16. H. B. Mitchell, *Multi-Sensor Data Fusion: An Introduction*, Springer, Berlin, **2007**.

17. L. A. Klein, *Sensor and Data Fusion Concepts and Applications*, 2nd ed edn., SPIE Optical Engineering Press, Bellingham, **1999**.

18. A. Ardeshir Goshtasby and S. Nikolov, *Inf. Fusion*, **2007**, 8, 114-118.

19. I. Bloch, *IEEE Trans Syst Man Cybern Pt A Syst Humans*, **1996**, 26, 52-67.

20. I. Corona, G. Giacinto, C. Mazzariello, F. Roli and C. Sansone, *Inf. Fusion*, **2009**, 10, 274-284.

21. D. L. Hall and J. Llinas, *Proc. IEEE*, **1997**, 85, 6-23.

22. G. L. Rogova and V. Nimier, Proceedings of the Seventh International Conference on Information Fusion, FUSION 2004, Stockholm, **2004**.

23. D. Smith and S. Singh, *IEEE Trans Knowl Data Eng*, **2006**, 18, 1696-1710.

24. J. Yao, V. V. Raghavan and Z. Wu, *Inf. Fusion*, **2008**, 9, 446-449.

25. M. L. Tiffany and S. Krimm, *Biopolymers - Peptide Science Section*, **1968**, 6, 1379-1382.

26. M. L. Tiffany and S. Krimm, *Biopolymers - Peptide Science Section*, **1968**, 6, 1767-1770.

27. M. L. Tiffany and S. Krimm, *Biopolymers*, **1969**, 8, 347-359.

28. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, **2000**, 28, 235-242.

29. Andres Liljas, Lars Liljas, Jure Piskur, Göran Lindblom, Poul Nissen and M. Kjeldgaard, *TEXTBOOK OF STRUCTURAL BIOLOGY*, World Scientific, **2009**.

30. M. Wang and R. D. Jiji, *Biophysical Chemistry*, **2011**, 158, 96-103.

31. F. J. R. and N. K., *Introductory Raman Sepctroscopy*, Academic Press, Inc, San Diego, CA, **1994**.

32. J. C. Austin, T. Jordan and T. G. Spiro, *Adv. Spectrosc. (Chichester, U. K.)*, **1993**, 20, 55-127.

33. J. V. Simpson, O. Oshokoya, N. Wagner, J. Liu and R. D. Jiji, *Analyst*, **2011**, 136, 1239-1247.

34. R. Bro and N. D. Sidiropoulos, *Journal of Chemometrics*, **1998**, 12, 223-247.

Chapter 5 - Parallel Factor Analysis of Multi- Excitation Ultraviolet Resonance Raman Spectra for Protein Secondary Structure Determination.

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 α-helical (-57°, -47°), β-sheet (-119°, 113°) and poly-proline II type (-79°, 150°) dihedral angles. This is the first empirical evidence that the typically resolved "disordered" spectrum represents the better defined poly-proline II type structure.

5.1 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] Traditional methods for protein secondary structure quantification such as x-ray crystallography,[6, 7] nuclear magnetic resonance,[8, 9] and circular dichroism[10, 11] are now complimented by vibrational methods like IR, conventional and resonance Raman.[12-19]

Deep-ultraviolet resonance Raman (DUVRR) particularly, has proven useful for quantification due to the structural sensitivity of the observed backbone amide modes. Some advantages of DUVRR include increased signal intensity (10^2-10^6 times) versus conventional Raman spectroscopy, minimal contribution from solvent water bands, elimination of background fluorescence and selective enhancement of the peptide backbone modes derived from the various vibrations of the backbone amide group (-CO-NH-) as a result of the π_2 to π_3^* dipole- allowed transition.[18, 20-24] DUVRR sensitivity to protein secondary structure is observed in the shifting and intensity changes of the four observable amide modes.[25-30] The position and intensity of the four amide modes: amide I, II, III and S are dependent upon the secondary structure of the protein with their relative contributions being proportional to the relative amount of each secondary structure conformation. Similarly, each secondary structure type exhibits distinct absorption profiles leading to a compositional dependence in resonance enhancement versus excitation wavelength.[21] Therefore, the position and intensity of amide modes change with varying excitation wavelengths and secondary structure composition.[31]

Initially, quantification studies of protein secondary structure employing DUVRR focused on univariate calibration methods of single amide modes,[18] but have evolved to include multivariate calibration and multivariate curve resolution methods and other advanced statistical analyses of all observable amide modes.[14, 27, 28, 32-37] In these methods, first order data are decomposed using bilinear models showing that individual protein spectra, **x**, are a linear combination of the different underlying pure secondary structure motifs, **s**, and their respective fractional amounts, *c* (Equation 5-1);

$$\mathbf{x} = c_\alpha \mathbf{s}_\alpha + c_\beta \mathbf{s}_\beta + \cdots \qquad (5\text{-}1)$$

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

Spectra from different proteins with different secondary structure compositions can be combined into a matrix, **X**, and decomposed as described in Equation 5-2;

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} \qquad (5\text{-}2)$$

C being the matrix of known secondary structural content of each protein, **S** being the matrix of underlying pure secondary structure and superscript T denotes the matrix transpose. The pure underlying secondary structure spectra may then be calculated via multivariate least square regression;

$$\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-\mathbf{1}} \qquad (5\text{-}3)$$

where superscript T denotes the matrix transpose and $\mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathbf{C}^{\mathrm{T}})^{-1}$ is the pseudoinverse of **C** to give **S**. To determine the structural content of a new protein based on this analysis, Equation 5-2 can be rearranged to solve for $\mathbf{c}_{\mathrm{new}}$, the fractional amount of each secondary structure within the new

protein such that;

$$\mathbf{c}_{new} = \mathbf{x}_{\text{new}}(\mathbf{S}^{\text{T}}\mathbf{S})^{-1}\mathbf{S}^{\text{T}} \qquad (5\text{-}4)$$

where $\mathbf{x}_{\text{new}}$ is the spectrum of the new protein.

The "known" concentration profiles are typically determined from proteins whose atomic level structure has been characterized by x-ray crystallography or NMR spectroscopy and published in the protein database, PDB (www.rcsb.org)[38]. However, the conditions under which the structures were determined do not necessarily represent the conditions under which the spectroscopic measurements are being conducted. This is especially true of crystal structures, which are static in nature. "Known" spectral profiles can be estimated from model polypeptides, such as poly-L-lysine or poly-L-glutamic acid, that adopt "pure" secondary structures with varying environments.[32] Given that these pre-determined or "known" inputs are not taken from the proteins themselves in their relevant environments, they can themselves introduce errors or biases into the multivariate analysis.

Trilinear methods are not prone 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.[39-43] Using a combination of ME-UVRR of nine globular proteins and PARAFAC, the pure spectral profiles of folded α-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- $+3_{10}$), β-sheet and poly-proline II type (PPII) structure.

5.2 Materials and methods

5.2.1 Sample preparation

Nine proteins with varying secondary structure content (Table 5-1), 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 $mg \cdot ml^{-1}$ for protein solutions and 200 μM 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.

5.2.2 Deep-UV resonance Raman (DUVRR) spectroscopy

The DUVRR instrument used to collect protein spectra has been previously described.[28, 44] 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 4°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.

5.2.3 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.[36] Each spectrum was interpolated to match the number of variables in the 197nm 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.[37] 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.

Table 5-1 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

5.2.4 Calculation of secondary structure content

Protein coordinate files for the nine proteins were downloaded from the protein data bank, PDB (www.rscb.org)[38] and a dihedral angle calculator readily available online (http://cib.cf.ocha.ac.jp/bitool/DIHED2/)[45] 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 5-1. The overall distribution of (φ, ψ) angles are shown in Figure 5.1 along with the structural assignments.

5.2.5 Parallel factor analysis (PARAFAC) of ME-UVRR data

PARAFAC can be applied to multi-way or trilinear data arrays and has an advantage of resolving underlying components without prior knowledge of sample composition or pure spectral features from trilinear data.[46-48] 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 5.2). PARAFAC analysis of ME-UVRR data was first applied to the amide I region of ME-UVRR protein spectra (Simpson, et al. Analyst 2011)[36]. However, quantification of disordered regions was not possible.

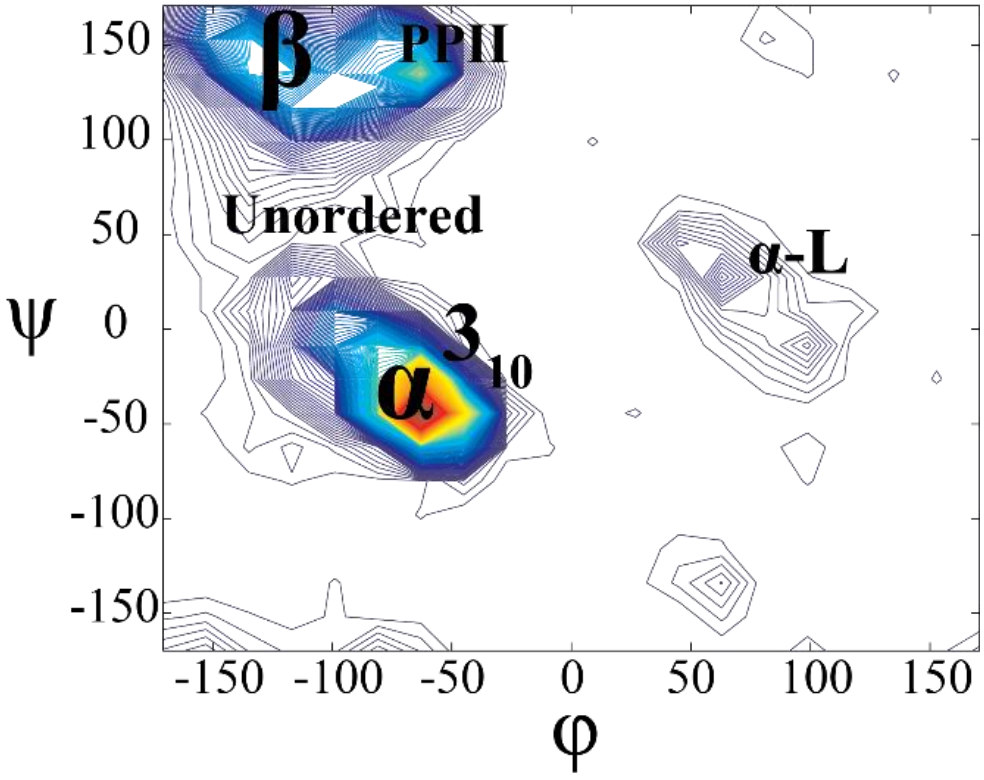

Fig 5.1 Ramachandran plot showing distribution of (φ, ψ) dihedral angles around the ideal dihedral angles for each type of secondary structure. Undefined dihedral angles occur in regions outside the labeled areas.

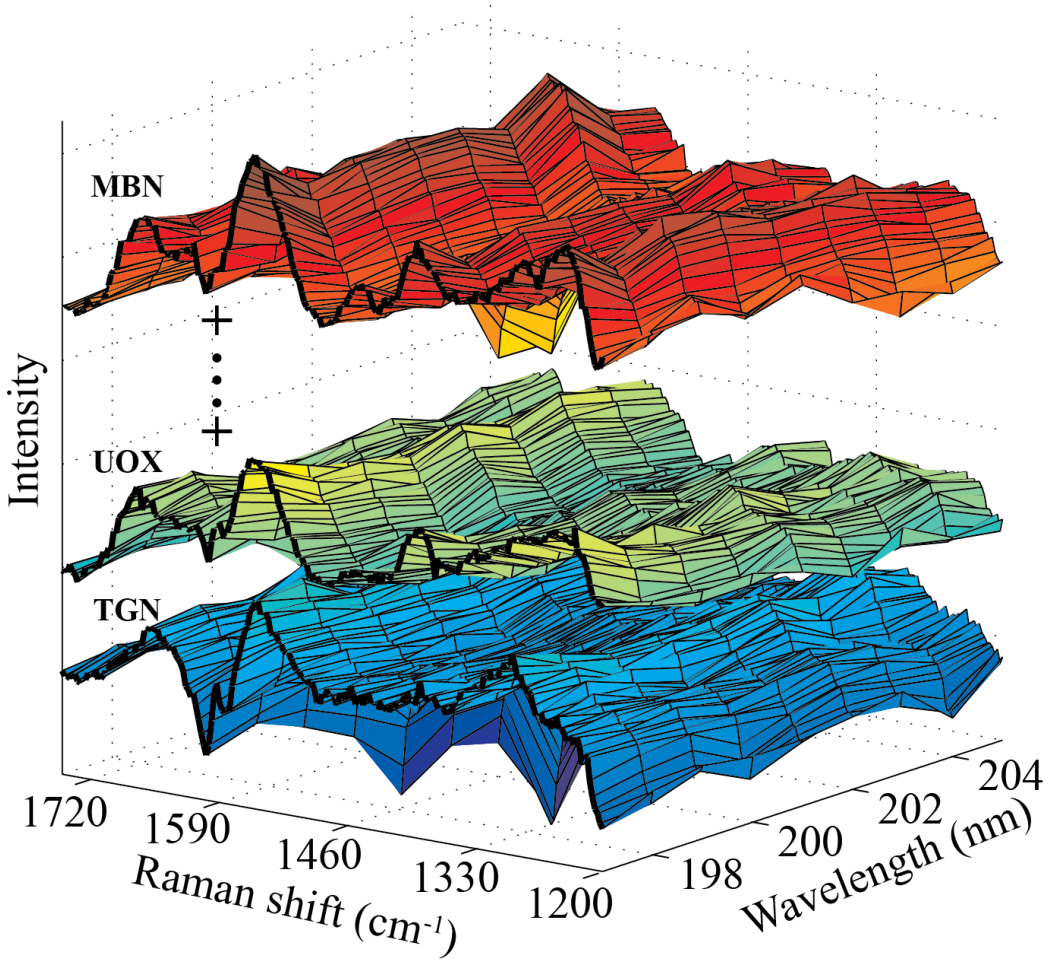

Fig 5.2 Trilinear data array produced by multi-excitation UVRR spectra. Myoglobin (MBN), glucose oxidase (UOX) and trypsinogen (TGN) are shown.

PARAFAC is based on the trilinear model[46], as shown in the Equation 5-6

$$r_{ijk} = \sum_{n=1}^{N} \hat{x}_{i,n}\ \hat{y}_{j,n}\ \hat{z}_{k,n} + \ e_{i,j,k} \qquad (5\text{-}6)$$

In Equation 5-6, 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 $\mathbb{R}$, represents ME-UVRR spectra of a series of proteins and has the dimension I x 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 the %RMSEC and reasonability of spectral profiles for a series of models. The weighted-PARAFAC algorithm[49] used in this work has been described previously.[50, 51]

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 x 9 x 9 data cube (Figure 5-2). 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. 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 5-2 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[52] 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 3, 4, and 5 factors and %RMSEC (Equation 5-7) was used to select the number of factors:

$$\%\text{RMSEC} = \sqrt{\frac{\sum(\hat{c}_i - c_i)^2}{I}} \times 100 \qquad (5\text{-}7)$$

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.

Table 5-2 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
4	XspYexZsa	Spectra	Excitation λ	Sample
	XexYspZsa	Excitation λ	Spectra	Sample
	XspYsaZex	Spectra	Sample	Excitation λ
	XsaYspZex	Sample	Spectra	Excitation λ

5.3 Results and Discussion

5.3.1 Determining the number of factors

The 3-, 4- and 5-factor PARAFAC models of the ME-UVRR data were evaluated for their predictive capability as well as the reasonability of the resolved spectral profiles. None of the resolved compositional profiles from the 5-factor model were correlated with either β-sheet/strand or helical structure. Further, the resolved spectral profiles were unreasonable in that there were duplicate factors. Therefore, it was concluded that the data should be modeled with fewer than five factors. 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 5-1 are based on the distribution of (φ, ψ) dihedral angles around the ideal as shown in Figure 5.1. 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 both 3- and 4- factor models (Figure 5.3). Irrespective of data orientation, %RMSEC for β-sheet structure prediction was 3.4% for unconstrained 3-factor models and 3.7% for 4-factor models (Figure 5.3). Application of non-negativity constraints to both 3- and 4-factor models resulted in a modest increase in %RMSEC. However, significant increases in %RMSEC were observed for the 4-factor models with non-negativity constraints and excitation data in the Z-dimension.

Percent RMSEC's for helical secondary structure were generally higher than those for β-sheet structure (Figure 5.3). Before application of non-negativity constraints, all 3-factor models showed a lower %RMSEC (14.5%) than 4-factor models (29%). For 3-factor models, little increase was observed in %RMSEC when non-negativity constraints were applied with the exception of model XexYspZsa which had a %RMSEC of over 70%. When non-negativity

constraints were applied to the 4-factor models, a decrease in %RMSEC was observed in all models except for model XspYexZsa which had a %RMSEC of 65%.

The third factor in the 3-factor models showed very strong correlation to PPII type structure composition. For the unconstrained 3-factor models, the %RMSEC for PPII type structure ranged from 17 to 22%, when non-negativity constraints were applied to the 3-factor 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-factor models, there is a significant decrease in the %RMSEC for PPII-type structure. For four factor models, the remaining two factors did not show strong correlation to any other secondary structure type including PPII-type structure.

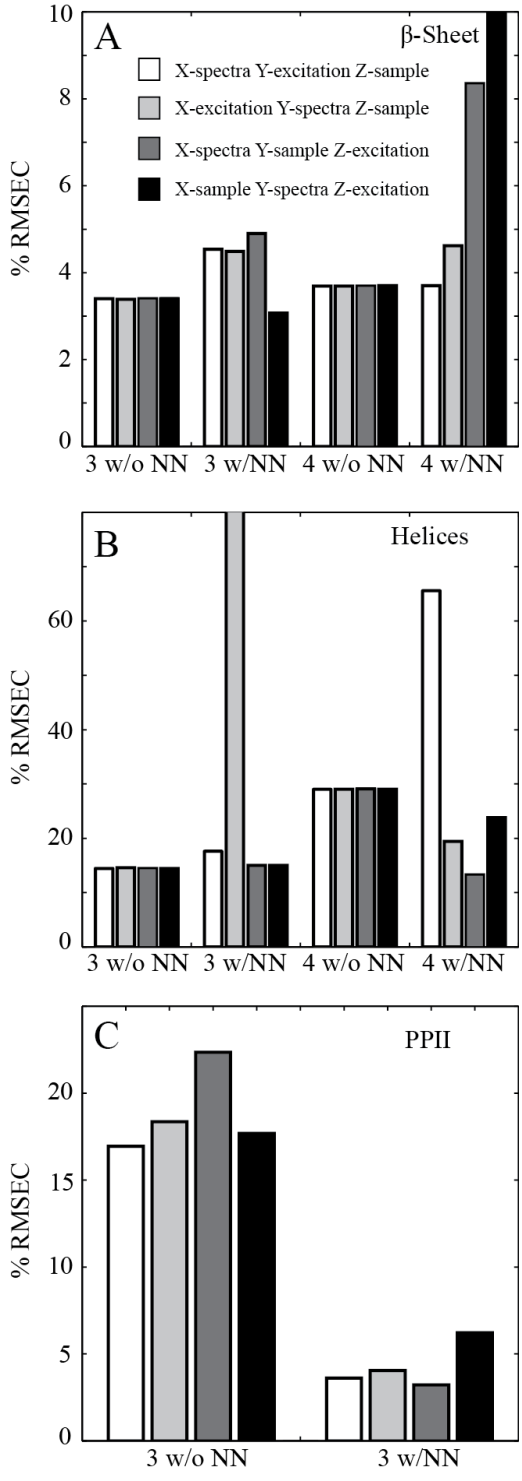

Fig 5.3 Percent RMSEC for the 3-factor and 4-factor models for each secondary structure type with (w/NN) and without (w/o NN) non-negativity constraints.

Examination of the resolved spectral profiles for each of the 3-factor models provides further insight into the optimal model. The resolved spectral profiles for each model without non-negativity constraints were essentially identical (Figure 5.4A). The resolved helical spectrum has a strong amide S band at 1390 cm^{-1}, an amide III feature at 1230-1240 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-factor 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 5.4C and 5.4D). Thus, the most reasonable model is produced when a 3-factor model is employed with non-negativity constraints and excitation information is placed in the Z-dimension (Figure 5.4B).

The predicted helical spectrum from the PARAFAC analysis of the 3-factor model with excitation in the Z-dimension (Figure 5.5) shows an absence of the amide S (1374-1406 cm^{-1}) mode and two features in the amide III region (1297-1336 cm^{-1}) as described in the literature.[25, 27, 53] 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[25, 27, 53]. 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 1670-1675 cm^{-1}, 1550-1564 cm^{-1}, 1395-1406 cm^{-1}, and 1220-1241 cm^{-1} respectively.[29, 53] The amide I and II of the disordered resolved spectral profiles appear at 1675 cm^{-1} and 1566 cm^{-1} falling within the expected regions of 1660-1682 cm^{-1} and 1548-1561

cm^{-1} respectively.[27, 53] The amide III and S modes are slightly upshifted (1250 cm^{-1}) and downshifted (1393 cm^{-1}) from the β-sheet spectrum as expected for PPII-type structure.

The resolved secondary structure components from the 3-factor models with excitation in the Z- dimension are essentially identical and show good correlation with reported helical, β-sheet and PPII-type contents for the sample proteins. A linear regression model was built using the resolved intensity for the different secondary structure components and the calculated secondary structure compositions given in Table 5-1 with the exception of ovalbumin. Inclusion of ovalbumin significantly skewed all of the regression models for one or more secondary structures. For the optimal 3-factor model XspYsaZex with non-negativity constraints, the %RMSEC decreased from 12% to 5% for β-sheet structure after removing ovalbumin. The resolved intensities of the secondary structures obtained from the PARAFAC analysis were then used to re-predict the secondary structure content using the linear regression model (Figure 5.6).

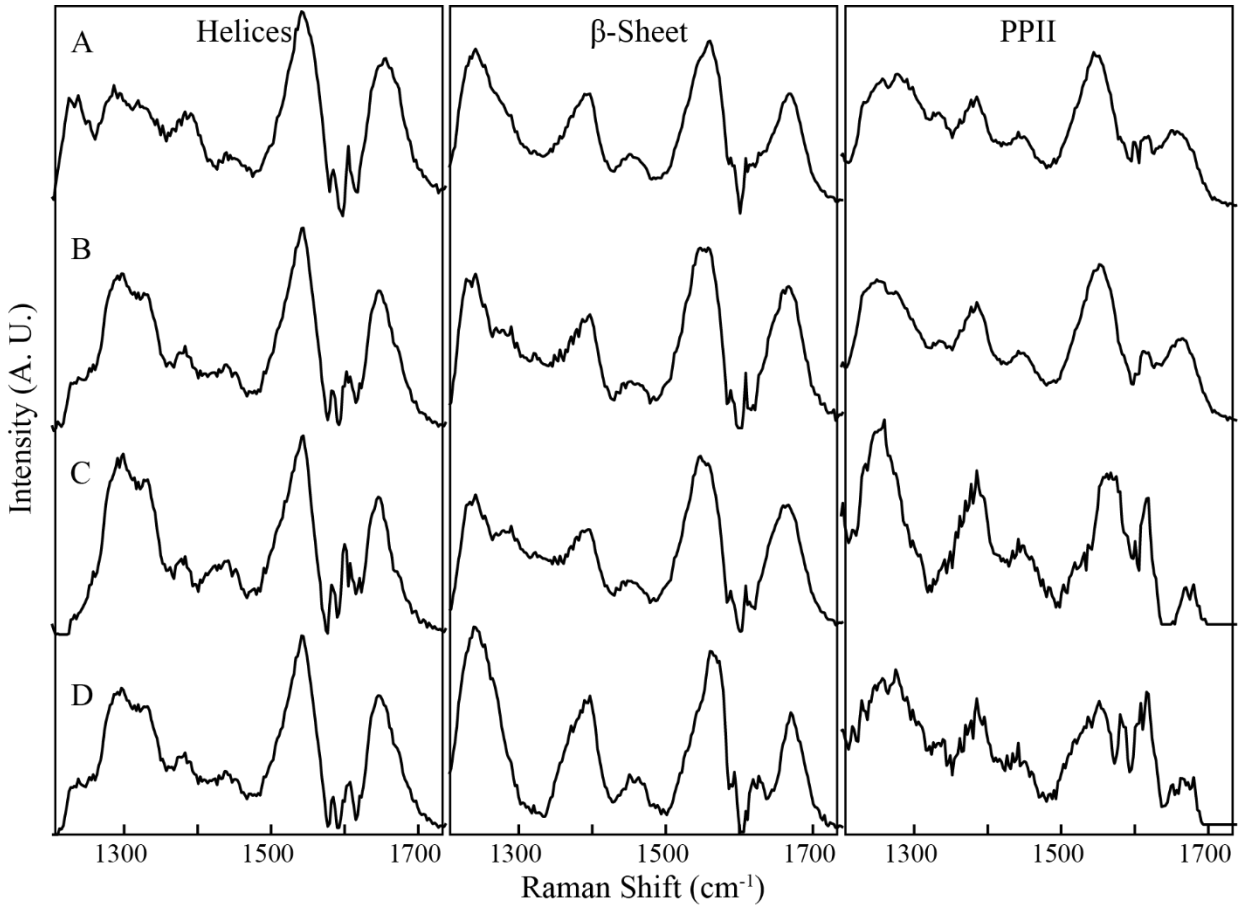

Fig 5.4 Resolved underlying spectra obtained from each 3-factor PARAFAC model for each secondary structure type. (A) XspYsaZex, without non-negativity constraints, (B) XspYsaZex, with non-negativity constraints, (C) XspYexZsa, with non-negativity constraints and (D) XexYspZsa, with non-negativity constraints.

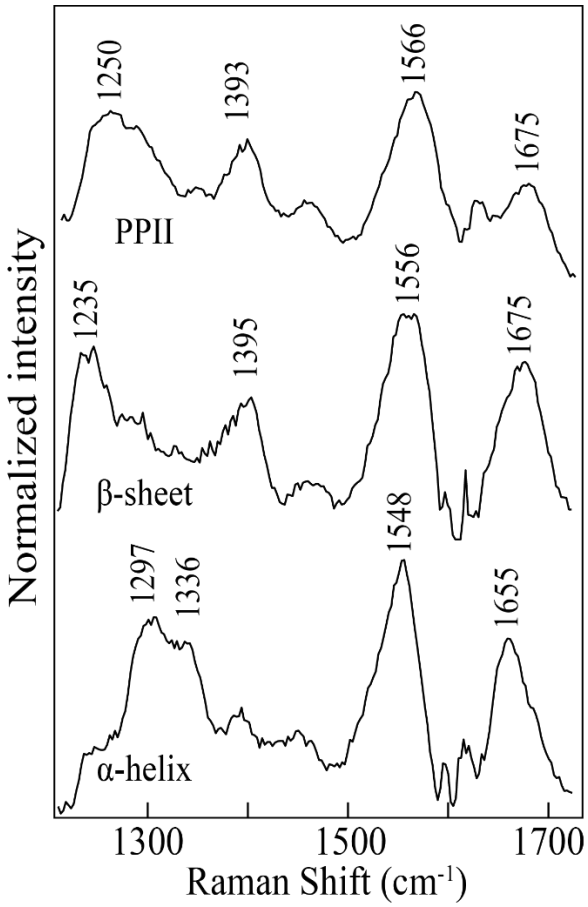

Fig 5.5 The pure secondary structure Raman spectra (PSSRS) obtained from the three component PARAFAC analysis of the trilinear ME-UVRR dataset.

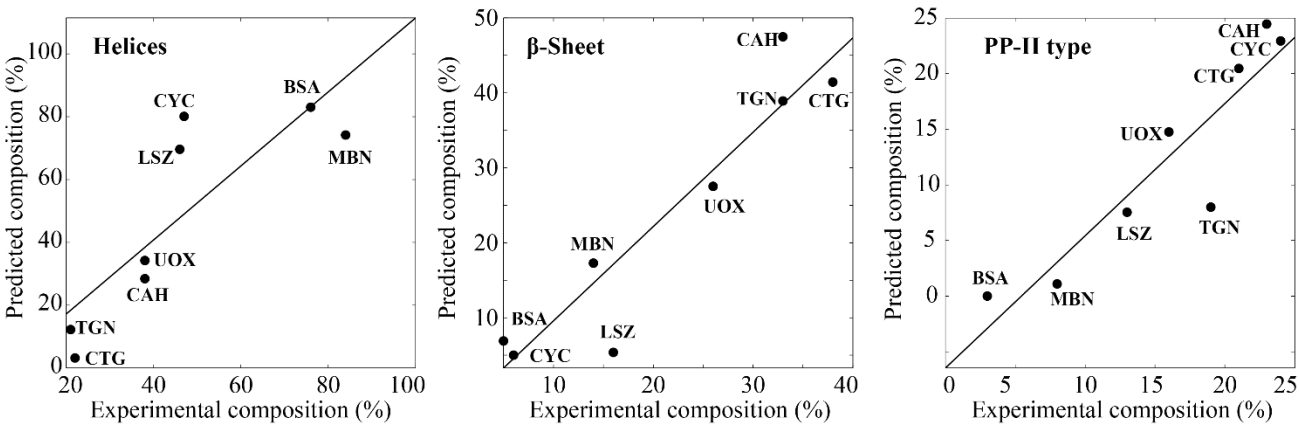

Fig 5.6 The experimental versus predicted percentages of secondary structure composition for the 3-factor model (XsYsaZex).

The excitation profiles of the secondary structures obtained from the non-negativity constrained 3-factor PARAFAC model (Figure 7) showed strong resemblance to those previously published in the literature[21, 36] especially for β-sheet which has a single maxima at approximately 200 nm. The helical excitation profile has two maxima appearing at approximately 198 nm and 204 nm. The PPII-type structure excitation profile has a maxima below 197 nm and a shoulder at approximately 202 nm.

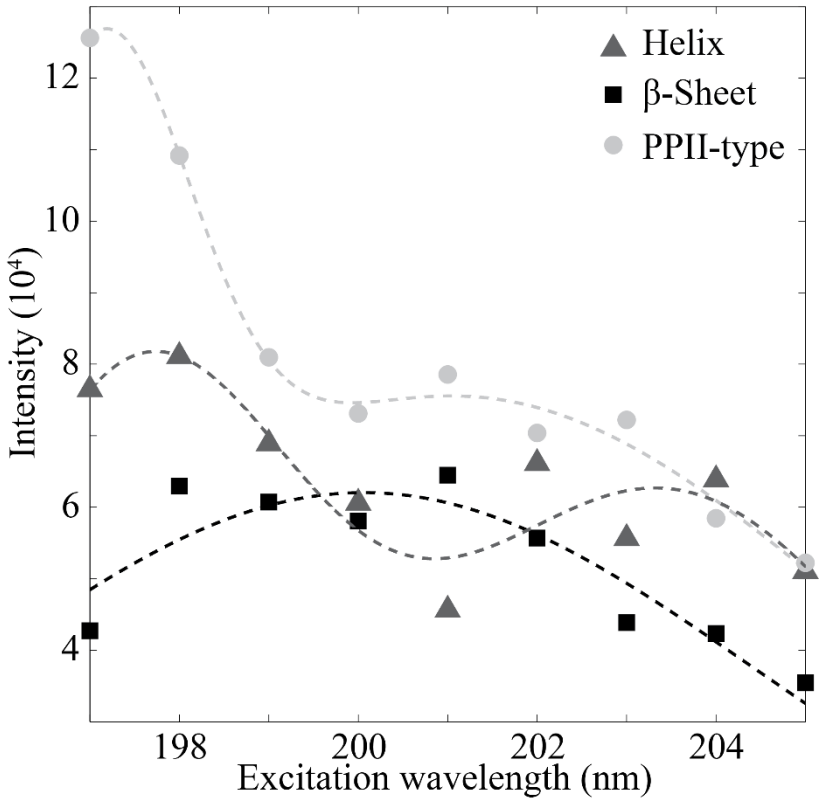

Fig 5.7 Excitation profiles for each secondary structure obtained from the 3-factor XspYsaZex PARAFAC model. For visualization, the estimated excitation profiles based on a Gaussian fit of the resolved excitation profile are shown as dashed lines.

5.4 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-factor 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.

5.5 References

1. A. Moglich, X. Yang, R. A. Ayers and K. Moffat, *Annual review of plant biology*, 2010, **61**, 21-47.
2. E. Herczenik and M. F. B. G. Gebbink, *FASEB Journal*, 2008, **22**, 2115-2133.
3. C. Weissmann, *Nature Reviews Microbiology*, 2004, **2**, 861-871.
4. S. B. Prusiner, *Proceedings of the National Academy of Sciences*, 1998, **95**, 13363-13383.
5. C. C. Blake, M. J. Geisow, S. J. Oatley, B. Rerat and C. Rerat, *Journal of molecular biology*, 1978, **121**, 339-356.
6. A. Higashiura, K. Ohta, M. Masaki, M. Sato, K. Inaka, H. Tanaka and A. Nakagawa, *Journal of Synchrotron Radiation*, 2013, **20**, 989-993.
7. G. Bluacz, M. Miller, R. Harrison, N. Thanki, G. L. Gilliland, C. M. Ogata, S. H. Kim and A. Wlodawer, *Acta Crystallographica Section D: Biological Crystallography*, 1997, **53**, 713-719.
8. F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein and H. Oschkinat, *Nature*, 2002, **420**, 98-102.
9. D. S. Wishart, B. D. Sykes and F. M. Richards, *Journal of Labelled Compounds and Radiopharmaceuticals*, 1992, **31**, 1019-1028.
10. N. J. Greenfield, *Nature Protocols*, 2006, **1**, 2876-2890.
11. N. J. Greenfield and G. D. Fasman, *Biochemistry*, 1969, **8**, 4108-4116.
12. A. Dong, P. Huang and W. S. Caughey, *Biochemistry*, 1990, **29**, 3303-3308.
13. W. K. Surewicz, H. H. Mantsch and D. Chapman, *Biochemistry*, 1993, **32**, 389-394.
14. C. A. Roach, J. V. Simpson and R. D. Jiji, *Analyst*, 2012, **137**, 555-562.
15. S. Navea, R. Tauler, E. Goormaghtigh and A. de Juan, *Proteins*, 2006, **63**, 527-541.

16. R. Y. Yada, R. L. Jackman and S. Nakai, *International journal of peptide and protein research*, 1988, **31**, 98-108.

17. T. G. Spiro and C. A. Grygon, *Journal of Molecular Structure*, 1988, **173**, 79-90.

18. R. A. Copeland and T. G. Spiro, *Biochemistry*, 1987, **26**, 2134-2139.

19. J. T. Pelton and L. R. McLean, *Analytical Biochemistry*, 2000, **277**, 167-176.

20. A. V. Mikhonin, N. S. Myshakina, S. V. Bykov and S. A. Asher, *Journal of the American Chemical Society*, 2005, **127**, 7712-7720.

21. K. Rosenheck and P. Doty, *Proceedings of the National Academy of Sciences of the United States of America*, 1961, **47**, 1775-1785.

22. S. A. Asher, Z. Chi and P. Li, *Journal of Raman Spectroscopy*, 1998, **29**, 927-931.

23. S. Song and S. A. Asher, *Journal of the American Chemical Society*, 1989, **111**, 4295-4305.

24. S. A. Oladepo, K. Xiong, Z. Hong and S. A. Asher, *Journal of Physical Chemistry Letters*, 2011, **2**, 334-344.

25. Y. Wang, R. Purrello, T. Jordan and T. G. Spiro, *Journal of the American Chemical Society*, 1991, **113**, 6359-6368.

26. V. A. Shashilov, V. Sikirzhytski, L. A. Popova and I. K. Lednev, *Methods*, 2010, **52**, 23-37.

27. C. Y. Huang, G. Balakrishnan and T. G. Spiro, *Journal of Raman Spectroscopy*, 2006, **37**, 277-282.

28. J. V. Simpson, G. Balakrishnan and R. D. Jiji, *Analyst*, 2009, **134**, 138-147.

29. J. C. Austin, T. Jordan and T. G. Spiro, *Adv. Spectrosc. (Chichester, U. K.)*, 1993, **20**, 55-127.

30. S. A. Asher, A. V. Mikhonin and S. Bykov, *Journal of the American Chemical Society*, 2004, **126**, 8433-8440.

31. R. A. Copeland and T. G. Spiro, *J. Am. Chem. Soc.*, 1986, **108**, 1281-1285.

32. Z. Chi, X. G. Chen, J. S. W. Holtz and S. A. Asher, *Biochemistry*, 1998, **37**, 2854-2864.

33. Z. Chi and S. A. Asher, *Biochemistry*, 1998, **37**, 2865-2872.

34. I. K. Lednev, A. S. Karnoup, M. C. Sparrow and S. A. Asher, *J. Am. Chem. Soc.*, 1999, **121**, 4076-4077.

35. A. Ozdemir, I. K. Lednev and S. A. Asher, *Biochemistry*, 2002, **41**, 1893-1896.

36. J. V. Simpson, O. Oshokoya, N. Wagner, J. Liu and R. D. Jiji, *Analyst*, 2011, **136**, 1239-1247.

37. O. O. Oshokoya, C. A. Roach and R. D. Jiji, *Anal. Methods*, 2014, **6**, 1691-1699.

38. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235-242.

39. D. Kovacs, B. Szabo, R. Pancsa and P. Tompa, *Arch. Biochem. Biophys.*, 2013, **531**, 80-89.

40. P. Tompa, *Trends Biochem. Sci.*, 2012, **37**, 509-516.

41. V. N. Uversky, *Chemical Reviews*, 2014, **114**, 6557-6560.

42. 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 and M. M. Babu, *Chemical Reviews*, 2014, **114**, 6589-6631.

43. M. P. Williamson and J. R. Potts, *Biochem. Soc. Trans.*, 2012, **40**, 945-949.

44. M. Wang and R. D. Jiji, *Biophysical Chemistry*, 2011, **158**, 96-103.

45. Andres Liljas, Lars Liljas, Jure Piskur, Göran Lindblom, Poul Nissen and M. Kjeldgaard, *TEXTBOOK OF STRUCTURAL BIOLOGY*, World Scientific, 2009.

46. R. Bro, *Chemometrics and Intelligent Laboratory Systems*, 1997, **38**, 149-171.

47. M. A. B. Levi, I. S. Scarminio, R. J. Poppi and M. G. Trevisan, *Talanta*, 2004, **62**, 299-305.

48. R. D. Jiji, G. G. Andersson and K. S. Booksh, *Journal of Chemometrics*, 2000, **14**, 171-185.

49. R. D. Jiji and K. S. Booksh, *Analytical Chemistry*, 2000, **72**, 718-725.

50. R. A. Harshman, *UCLA Working Papers in Phonetics*, 1970, **16**, 84.

51. J. D. Carroll and J. J. Chang, *Psychometrika*, 1970, **35**, 283-319.

52. M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley- VCH, 2007.

53. R. D. JiJi, G. Balakrishnan, Y. Hu and T. G. Spiro, *Biochemistry*, 2006, **45**, 34-41.

Chapter 6 – Conclusions

The quantification of protein secondary structure has become an area of intense biochemical and biophysical research due to the effects of secondary structure on tertiary and quaternary protein structure, as well as the role of secondary structure in protein function and select diseases. Traditional methods of quantifying protein secondary structure, such as X-ray crystallography (XRC), nuclear magnetic resonance (NMR) and circular dichroism (CD), are now complimented by a host of vibrational methods, in particular, deep-ultraviolet resonance Raman (DUVRR) as it has proven useful due to its structural sensitivity of the amide modes.

Typically, DUVRR and CD have been used independently to categorize and quantify protein secondary structure into three types, α-helix, β-sheet, and disordered structure. We have been able to demonstrate that their complimentary use leads to the improved prediction of disordered secondary structural content in proteins. Also, the experimental design utilized allowed us evaluate the performance of common multivariate methods (partial least squares, classical least squares and multivariate curve resolution- alternating least squares) for their predictive abilities using a limited protein data set. The results of these studies are of importance to researchers as they show that good estimations of secondary structural composition can be obtained with only a limited set of standard proteins.

A new approach to protein secondary structure was developed by the application of multivariate analysis to fused CD and DUVRR data. The rationale behind the application of data fusion to both spectra types before multivariate analysis was to exploit the selective predictive capabilities of each technique so as to further improve predictions of other secondary structures present including PPII-type structure. Limitations still exist with separation of different helical

structural types and prediction of less prevalent structures like turns which occur in very small quantities. Expansion of the data fusion methodology to include other structurally sensitive techniques like vibrational circular dichroism (VCD) or Raman optical activity (ROA) may increase the accuracy and number of quantifiable structures.

Finally, the combination of three-way DUVRR and parallel factor (PARAFAC) analysis with non-negativity constraints has been used here for prediction of protein secondary structure including PPII-type structure without initial assumptions of concentration/compositional or spectra profiles. The encouraging results obtained from 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. The importance of estimation without prior knowledge is that the NMR and X-ray structures that composition profiles are based on, don't necessarily represent the conditions that the optical spectra are being measured under. This allows us to move away from using NMR and X-ray information as initialization profiles.

These spectroscopic studies, and the application of chemometric methods to them, serve to highlight the value of advanced statistical methods for chemical analysis. Chemometric methods were shown to be powerful tools for the analysis of protein spectra, suggesting that future research will be able to look at more complex experimental data and extract useful protein secondary structure information thereby meriting continued research.

VITA

Olayinka Oshokoya was born on February 26th, 1984, in Lagos, Nigeria. He attended Federal Government College, Ijanikin (Lagos, Nigeria), and then studied at University of Lagos (Lagos, Nigeria), where he obtained his Bachelor's degree in Chemistry in 2004. He also obtained a Master's degree in Analytical Chemistry and Environmental Science in 2006 from Loughborough University (Leicestershire, United Kingdom). He has obtained a Ph.D. in Chemistry under the supervision of Dr. Renee D. JiJi, at the University of Missouri-Columbia, USA, with an anticipated graduation in May 2015.