DEVELOPMENT OF ADVANCED CHEMOMETRIC METHDODS FOR THE ANALYSIS
OF DEEP-UV RESONANCE RAMAN SPECTRA OF PROTEINS

_____

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

_____

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

_____

by

JOHN SIMPSON

Dr. Renee Jiji, Dissertation Supervisor

SEPTEMBER 2009

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

dissertation entitled

DEVELOPMENT OF ADVANCED CHEMOMETRIC METHODS FOR

THE ANALYSIS OF DEEP-UV RESONANCE RAMAN SPECTRA OF

PROTEINS.

Presented by John Simpson,

a candidate for the degree of Doctor of Philosophy,

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

_____

Dr. Renee Jiji

_____

Dr. C. Michael Greenlief

_____

Dr. Carol Deakyne

_____

Dr. Christie Spinka

I would like to dedicate this work to my father David, my mother Diane, my brothers James and Jay, and my girlfriend Charity. Thank you all for all your love and support.

ACKNOWLEDGEMENTS

I would like to thank my graduate advisor, Dr. Renee Jiji, for her encouragement 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 would also like to thank Dr. Jason Cooley who has given been kind enough to spend many hours reviewing my written documents and presentation, in addition to his important contributions to my research.

I would also like to thank the members of my graduate committee, Dr. C. Michael Greenlief, Dr. Carol Deakyne, and Dr. Christe Spinka, who have provided crucial guidance to me throughout my graduate career. I deeply appreciate their understanding 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. They have always been willing to collaborate with me in my research and have even helped me to collect some of the data that this dissertation is based upon. Their input on my publications, presentations, and research has served to improve my ability as a scientist.

Finally, I would like to thank my friends and fellow graduate students at the University of Missouri. The people I have met at the University of Missouri 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

DEVELOPMENT OF ADVANCED CHEMOMETRIC METHDODS FOR THE ANALYSIS
OF DEEP-UV RESONANCE RAMAN SPECTRA OF PROTEINS

John Simpson

Dr. Renee Jiji, Dissertation Supervisor

ABSTRACT

Ultra violet resonance Raman (UVRR) is a powerful spectroscopic technique for determining the secondary structural content of proteins in solution. However, the analysis of UVRR spectra can be problematic due to the difficulty of deterring the pure secondary structure Raman spectra, compounded by the presence of aromatic side chains which overlap in the amide bands. The use of multi-excitation datasets can help to alleviate the difficulty in determining the pure secondary structure Raman spectra, but due to increasing spectral resolution as excitation wavelength increases, multi-excitation datasets are notoriously difficult to align. In addition, the subtraction of the water band, which is heavily overlapped with the amide I vibration al mode can be difficult when relying on an internal intensity standard.

To address these difficulties we demonstrate the use of a series of chemometric methods. To determine the pure secondary structure Raman spectra, we demonstrate the use of multivariate curve resolution using the alternating least squares algorithm (MCR-ALS) with a multi-excitation data set. To alleviate mis-alignment in multi-excitation data, we demonstrate the use of correlation optimized warping (COW). Finally, we also propose a new water band subtraction method which will reliably determine the water band concentration and remove it, without an over subtraction.

Finally, we demonstrate the use of parallel factor analysis (PARAFAC) for the characterization of the behavior of the flavonoid quercetin in solution with the protein bovine serum albumin.

Chapter 1: Chemometric methods for the analysis of protein secondary structure by UV resonance Raman

1.1. Introduction

Protein structure determination is a challenging area of increasing importance in biochemistry and biophysics. Structural changes often occur in proteins as they perform their function in biological systems, such as the unfolding of an α-helix when signaled through the absorption of blue light by the LOV2 domain of phototropins.[1] Traditional methods for structural determination, such as X-ray crystallography (XRC) and nuclear magnetic resonance (NMR), can provide a wealth of structural information resolved at the atomic level. However, these methods, when applied to proteins, can be very time and labor intensive.[2] Crystallographic structures are often less meaningful than the more biologically relevant aqueous structures determined by NMR. However, NMR structure determinations have an upper limit in terms of protein size, which is far lower than that of XRC.[3] While these methods provide an enormous amount of data, a protein's secondary structure content is not always easily resolved with these techniques. Often, the secondary structural content is not static enough to be easily determined on the time scales of these techniques, limiting their application.

Classical spectroscopic methods such as circular dichroism (CD), infrared spectroscopy (IR), and Raman based methods are being revisited to fill this niche role in biochemical systems. However, the spectral signatures of each type of secondary structure are highly overlapped in these methods. This has led to the application of

chemometric methods to resolve the spectral features and relative proportions of α-helix, β-sheet and disordered secondary structures. This review will focus on the application of chemometric methods to ultraviolet resonance Raman (UVRR) spectral data for protein structure analysis in aqueous media.

UVRR has several advantages for protein spectroscopy. UVRR can be used to determine relative amounts of secondary structural content [4,5] and monitor protein dynamics on the ns to μs timescales.[6-10] In addition, UVRR spectra of proteins can be collected in an aqueous, non-deuterated environment, with little interference from water, which is a weak Raman scatterer. A unique advantage of UVRR spectroscopy is the selective excitation of the chromophore of interest. The ability to preferentially excite the chromophore, such as the amide backbone and/or the aromatic side chains, makes UVRR an attractive method to probe protein dynamics. Deep-UV excitation below 220 nm results in resonance enhancement of both the amide modes and aromatic side chains.[11] These advantages, in addition to its high degree of structural sensitivity, make UVRR an ideal technique for probing conformational changes in biochemical systems. [11-15] The advantages and subtle nuances of UVRR spectroscopy has already enabled its' use for the study of structural transition kinetics (temperature and pH induced)[16,17,9,7,18,19,10] as well as folding and unfolding studies of proteins (varying pH and ionic strength)[20-24].

1.2. UVRR structural sensitivity

The structural sensitivity of vibrational spectroscopic methods, such as infrared absorption spectroscopy, non-resonance Raman spectroscopy, and UVRR spectroscopy primarily arises from the signal of the polypeptide backbone amide group. This signal results in several discrete vibrational modes: the amide I, II, III and S modes.[11] Each of

these discrete vibrational modes have been assigned to specific vibrational motions. The amide I mode arises mainly from the C=O stretching motion. The amide II and III modes are composed of the out-of-phase and in-phase combinations of N-H in plane bending and C-N stretching, respectively. The amide S mode, S for structurally sensitive, is composed of the C_α-H bending coupled to the N-H bending motion of the amide III mode.[23]

Each secondary structure possible in a protein can be defined through a combination of φ and ψ dihedral angles along the amide backbone. The frequency of the vibrational modes of the amide groups will shift in energy and intensity with variations in the φ and ψ dihedral angles of the polypeptide backbone.[25] Non-resonant Raman studies has shown characteristic wavenumber ranges exist in the amide I and amide III which correspond to the major secondary structural groups, namely the α-helical, β-sheet, and random coil secondary structures. These wavenumber ranges demonstrate that in the amide I region shifts to lower energies with a high α-helical content (1645-1660 cm^{-1}), while high β-sheet content shifts the amide I region to higher energies (1665-1680 cm^{-1}). The wavenumber range of amide I maxima falls in the middle (1660-1665 cm^{-1}).[26] The opposite trend is observed in the amide III region where a high α-helical content shifts the region to higher energies (1260-1330 cm^{-1}), lower energies with the β-sheet (1230-1240 cm^{-1}). Again, the amide III mode falls in the middle of the two (1240-1250 cm^{-1}).[26] In the amide II mode, a shift in energy is not observed with variation in secondary structure. However, the amide II does display intensity variations with secondary structure content which has been utilized.

While the amide backbone is the chromophore which conveys secondary structure information, other chromophores are present in proteins which can also convey structural and conformational information. The aromatic amino acids; phenylalanine, tyrosine, and tryptophan will contribute to the UVRR spectra of proteins. Histidine, which frequently interacts with metal centers, shows poor UVRR enhancement when compared to phenylalanine, tyrosine, and tryptophan.[27] However, it is possible to enhance some histadine signals when the histadine residue is deuterated [28], the imidizole is protonated [29], or if the histadine is bound to a metal center.[30,31] 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) , with maximum enhancement occurring beyond the optimum range of the amide backbone 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 (hydrophobic or hydrophilic environment).[32]

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, because the degree of resonance enhancement of each mode is directly related to the absorption profile of the chromophore at the selected excitation wavelength.[33] Each secondary structure ha a unique absorption profile and as a result [34], the excitation profile of each mode is conformation dependent.[35-37,5] As a result of the aforementioned relationships, varying

either the secondary structure content or through changes in pH and 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 to 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 square, MCR-ALS and single point calibrations.[38-41,37]

1.3. Early work

Early UVRR studies utilized the heme chromophore groups in biologically significant proteins such as retinal[42], cytochrome C [43-45], myoglobin [46] and vitamin B_{12} [47]. Many of these studies focused on the signal arising from the porphyrin ring to which an active metal center is bound using the 457.9 nm, 496.5 nm, and 514.5 nm emission lines of an argon ion laser. Later work began to correlate the state of the chromophore with the conformation of the protein by monitoring the aromatic amino acid modes. As the proteins entered difference conformational states, spectral shifts could be observed in the aromatic amino acid modes.[48,24]

The utilization of excitation wavelengths below 210 nm allowed monitoring of gross structural changes which led to qualitative conclusions regarding structure directly from the amide modes. Changes in secondary structure were inferred based on shifts in the wavenumbers of the maxima of any given amide mode.[24] The amide I mode, due to

its well documented behavior in non-resonance enhanced spectra, was of particular interest. [49] The use of excitation wavelengths below 210 nm resulted in enhancement of the amide II mode. This in turn led to increased interest in its structural sensitivity [50]; the intensity of the amide II mode is extremely weak in non-resonance enhanced Raman.[51,52]

Interest in the amide II mode led to some of the first secondary structural determination methods for UVRR spectra. A comparison of the molar scattering ratio of the amide II mode to the α-helical content of a series of proteins containing primarily α-helical and random coil structure revealed a negative linear relationship. (Figure 1-1)[5] Similar negative correlations were also found between the molar scattering ratios of the amide III and amide II′ bands (deuterium substituted) with the α-helical content. A negative correlation was also observed between α-helical content and the molar scattering ratio of the amide S band. [53,23]

Efforts to expand quantification of secondary structure from the amide II mode to proteins which included β-sheet domains led to the use of multiple excitation wavelengths. The absolute Raman cross section (σ_N) of the amide II mode was treated as a linear combination of variables where each term was defined as the product of the fraction of a major secondary structure, α-helical, β-sheet, and unordered (formerly termed random coil), and the cross section of that secondary structure type at each excitation wavelength.[37]

$$\sigma_N = f_\alpha \sigma_\alpha + f_\beta \sigma_\beta + f_\tau \sigma_\tau \quad (1)$$

The cross section values of each secondary structure type were determined for both 192 nm and 200 nm excitation using a training set composed of proteins that had been well characterized through XRC.[37] The choice of excitation wavelengths is significant as 192
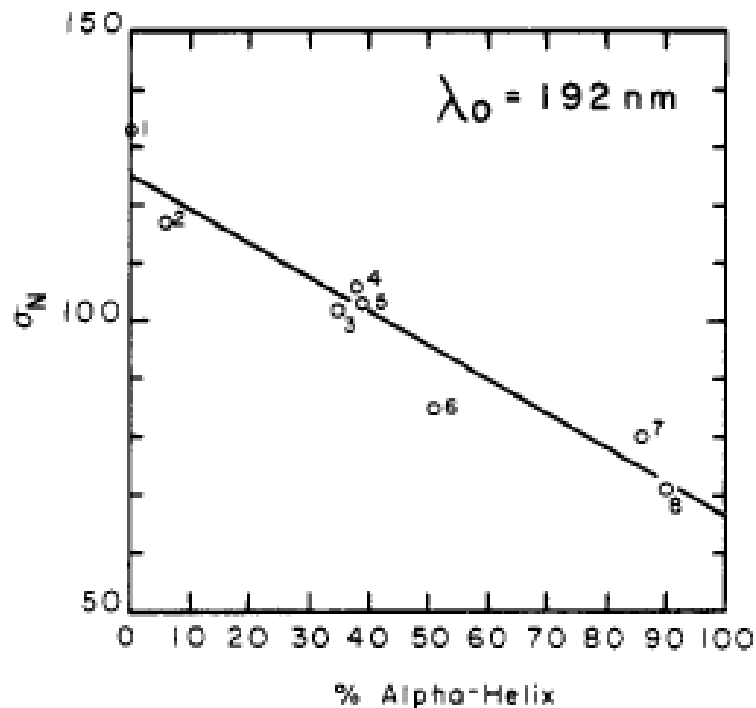
Fig 1-1 Plot of amide II Raman cross sections with 192-nm excitation vs. percent α-helix for the following: (1) tropomyosin, pH 12.5; (2) superoxide dismutase; (3) P-purothionin; (4) a,-purothionin; (5) cytochrome c; (6) insulin, pH 10; (7) myoglobin; (8) tropomyosin, pH 7.0. Reproduced with permission from Copeland, R. A. and T. G. Spiro, *Biochemistry* **26,** 2134 (1987).[37] Copyright 1987 American Chemical Society.

nm is the isobestic point of the absorption profiles of random coil and β-sheet structures. Therefore, these two structures will have identical cross section values at that wavelength, allowing the calculation of the α-helical content as had been done in previous studies.[5] With knowledge of the α-helical content, the remaining composition may be calculated from the amide II intensity in the 200 nm excitation spectrum. While this method performed well in predicting secondary structural content for a given set of proteins, the majority of the data contained within the UVRR spectra was ignored.

1.4. Prediction of protein secondary structure composition

The methods discussed above suffer from a major limitation in that the majority of the information contained in a UVRR spectrum is excluded, focusing on the maxima of a single amide mode. While the amide II mode showed great promise for secondary

structural determination for α-helix, β-sheet, and disordered structures, not all secondary

structures contribute strongly in the amide II region, limiting its application to important

sub types of secondary structure such as β-turn and different forms of helices.[37] Through

the application of the bilinear model to UVRR spectra, a more accurate and robust

method has been developed by using all the amide modes in a UVRR spectrum. The

bilinear model has become a common method in literature by which the secondary

structure of proteins may be determined from UVRR spectra.[54,4,20,10,55,56]

$$[D] = [R][C] \qquad (2)$$

The first application of the bilinear model to UVRR data was performed by Asher

and coworkers[4] on a data set consisting of the 206.5 nm excitation UVRR spectrum of a

series of well characterized proteins with different secondary structure compositions,

resulting in a bilinear data set. Equation 2 shows the relationship between the measured

UVRR spectra, D, and the pure component matrices, R and C, which represent the pure

secondary structure Raman spectra and secondary structure compositions, respectively.

Using the model shown in Equation 2, Asher et al.[4] calculated the pure Raman spectra of

each major secondary structure type using the known secondary structure compositions

of the proteins composing the data set by solving for R using Equation 3.

$$[R] = [D][C^T]([C][C^T])^{-1} \qquad (3)$$

Employing the calculated PSSRs, the secondary structure composition of new proteins

could then be determined using Equation 4, where D_1 and C_1 represent the experimental

UVRR spectrum and estimated secondary structure composition of the new protein,

respectively.

$$[C_1] = ([R]^T[R])^{-1}[R]^T[D_1] \quad (4)$$
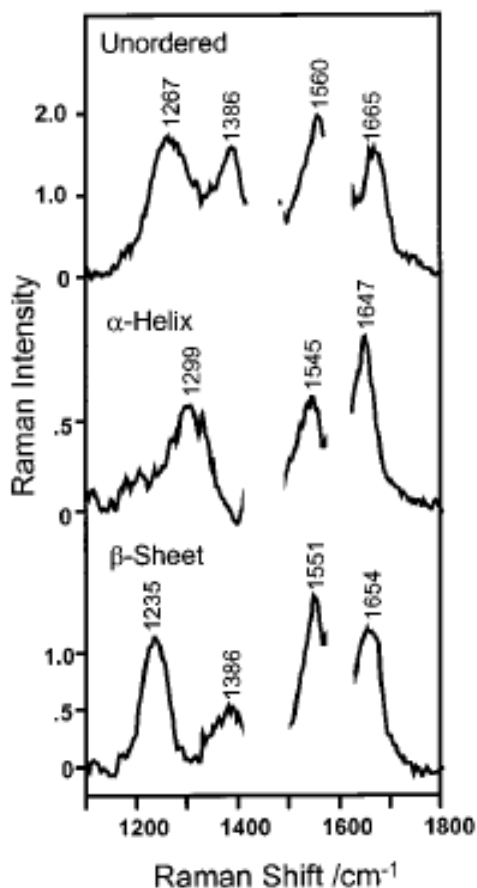
Fig. 1-2 Calculated pure secondary structure Raman spectra (PSSRS) obtained from aqueous solution 206.5-nm protein Raman spectra. The regions around 1610 and 1450 cm-1 are excluded due to interference from aromatic amino acid ring vibrations and side chain vibrations. Reproduced with permission from Chi, Z., Chen, X.G., Holtz, J, and S. A. Asher, *Biochemistry* **37**, (2854). Copyright 1998 American Chemical Society.

To discern the validity of the calculated PSSRs, they were compared to the resonance Raman spectra of poly-L-glutamic acid (PGA) and poly-L-lysine (PLL), proteins which can be induced to fold into primarily α-helical, β-sheet, and unordered confirmations through variations in pH and incubation at elevated temperatures, producing a control study[57,53,58]. Comparison to the resonance Raman spectra of PGA and PLL in each confirmation showed remarkable similarities. To further characterize the error in the calculated model, the calculated PSSRs, shown in Figure 1-2, were then used to determine the secondary structure compositions of the proteins used in the original

data set, with the error in the reconstruction of the data being less than 10% for the majority of proteins in the data set.[4]

Improvements to the use of the bilinear model have been made through the use of a lower excitation wavelengths lower than 206.5 nm.[36] Resonance enhancement of Raman scatter increases proportionally to the molar absorptivity of each secondary structure.[11] Spiro *et al.* demonstrated that with the use of 197 nm excitation spectra, the pure Raman spectrum and content of the β-turn secondary structure could be determined in a large protein dataset due to the greater excitation of the amide modes.[54] In addition, the use of a lower excitation wavelength resulted in a decrease in the error of the model.[54] This relationship between excitation wavelength and relative enhancement of various secondary structures will be key to expanding the types of secondary structure that can be quantified using UVRR.

1.5. Analysis of 2D multi-excitation UVRR spectra

Few studies have taken advantage of the relationship between resonance enhancement and excitation wavelength for structure determination. Each secondary structure type has a unique absorption profile[34], as shown in Figure 1-3, and will be differentially excited at each excitation wavelength, resulting in unique UVRR excitation profiles.[35,37] Exploiting this principle, the pure Raman spectra of the α-helical and non helical secondary structure was successfully determined from multi-excitation spectra of myoglobin, as shown in Figure 1-4.[38] Myoglobin is a predominantly (74%) α-helical protein.
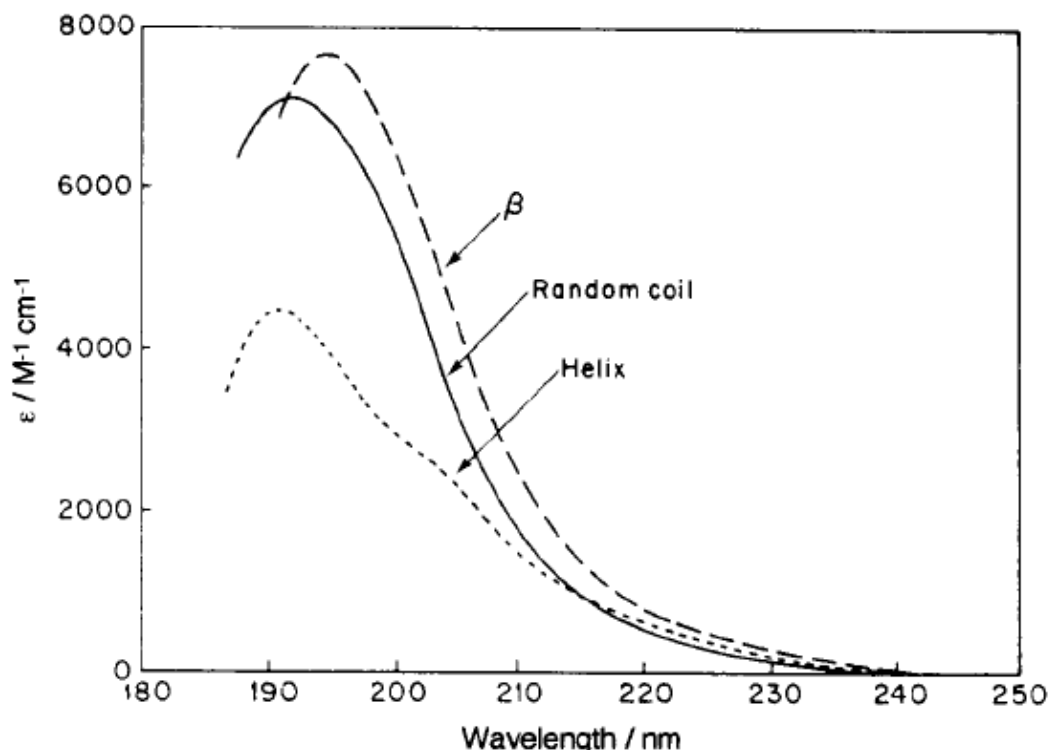
Fig. 1-3 Absorption spectra of poly-L-lysine in α-helical, β-sheet, and random coil conformations. Reprinted with permission from Copeland and Spiro, *Biochemistry* **26**, 2134. Copyright (1987) American Chemical Society.

In addition, this method did not require a subtraction of the signal from the aromatic amino acids. A key part of the decomposition method is the treatment of the aromatic amino acid signals, which overlap the amide I, II, and III regions. Previous to this multi-excitation study, the aromatic amino acids were subtracted from the dataset based on the intensity of the aromatic peaks.[36,4] With the use of a more advance deconvolution method, featuring a unique shape constraint, it was unnecessary to subtract the aromatic signal intensity to determine the pure α-helical secondary structure Raman spectra.

1.6. Modeling two-state processes

The application of the bilinear model marked the beginning of the use of more advanced statistical methods for secondary structure determination in UVRR spectra. In

11
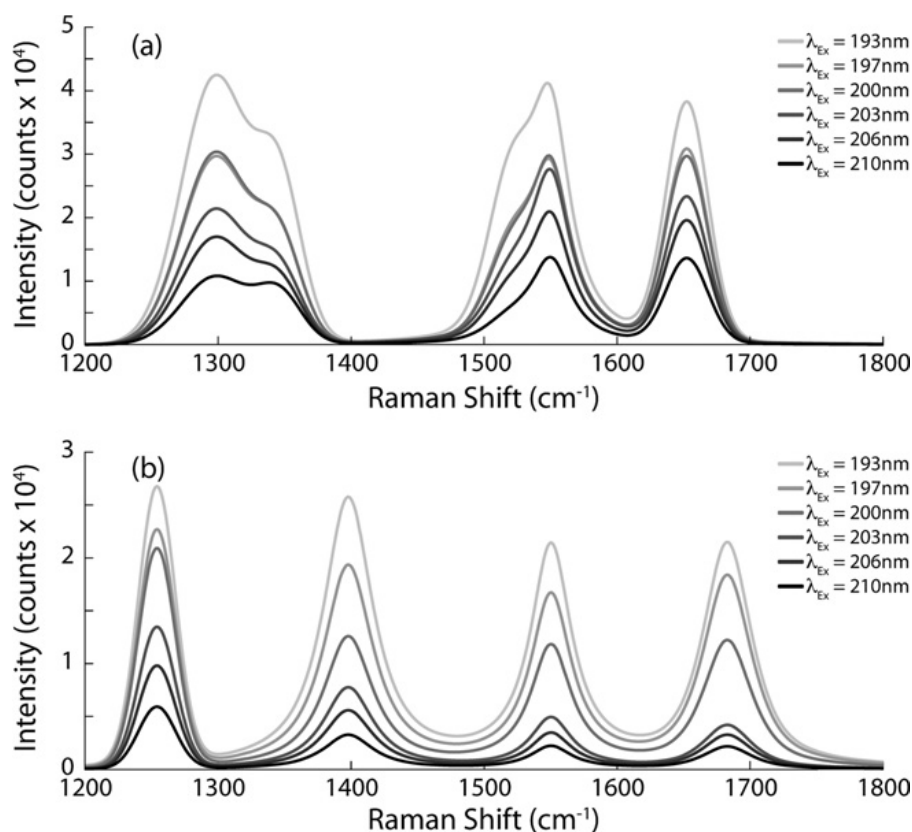
Fig. 1-4: Estimated pure UVRR spectral profiles of the (a) α-helical and (b) non-helical portions of myoglobin at each excitation wavelength. [Simpson, J. V., G. Balakrishnan, and R. D. Jiji, *The Analyst* **134** (138), 2009] – Reproduced with permission of The Royal Society of Chemistry.

addition, it has greatly simplified the monitoring of dynamic processes through UVRR.

One process, which is poorly understood, is the formation of protein fibrils in solution.

Through electrospray ionization mass spectrometry, it has been determined that this is

essentially a two state process.[39] However, relating this transition to secondary structure

was not possible through mass spectrometry. UVRR, coupled with chemometric

methods, have successfully related this process to secondary structure content.[59,56,40]

Hen egg white lysozyme is a good model protein for fibril formation. With

prolonged incubation, lysozyme under goes a transition from native lysozyme to a

primarily β-sheet structure which aggregate into fibrils, shown in Figure 1-5.[60] Utilizing

independent component analysis, a blind sorting technique, and two pure variable

methods, simple-to use interactive self-modeling mixture analysis (SIMPLISMA) and

stepwise maximum angle calculation (SMAC), Lednev and coworkers successfully

decomposed a data set containing the UVRR spectra of lysozyme collected at varying

time points during the fibrilization process into three component matrices.[40] These

component matrices were essentially the UVRR spectra of native lysozyme, unfolded

lysozyme, and fibrillar lysozyme, indicating that all spectra contained within the dataset

could be modeled as one of these three species. The concentrations of the pure Raman

spectra in the dataset indicated that native lysozyme unfolds during the transition to β-

sheet.

Further determination of the exact order of events in this process, and the relative

time scale of the events, was accomplished later through the use of 2D correlation UVRR
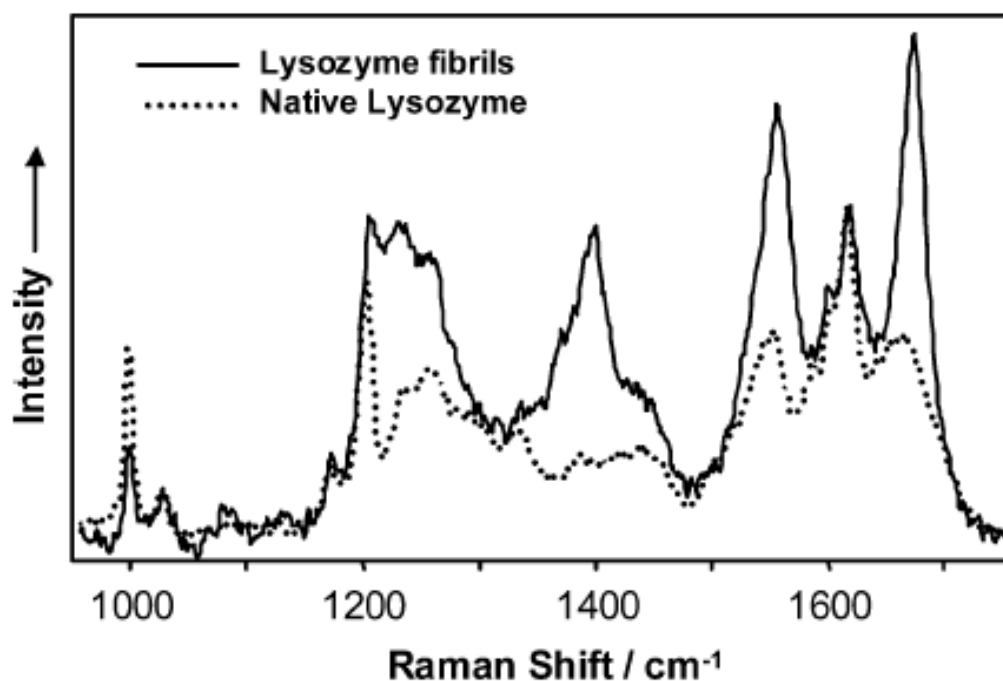
Fig. 1-5 UVRR spectra (197-nm) of native HEWL (pH 2, dotted line) and HEWL fibrils (solid line) measured at room temperature. The spectra were normalized for comparison using the phenylalanine-tyrosine band at ~1,600 cm^{-1}. Reproduced with permission from Springer Science and Business Media: *Analytical and Bioanalytical Chemistry, Deep-UV Raman spectrometer tunable between 193 and 205 nm for structural characterization of proteins.* **381**(2), 2005, p. 431-437. Lednev, Igor K., Ermolenkov, V., He, Wei, and Ming Xu.. Figure 5.

spectroscopy. Using 2D correlation UVRR spectroscopy, it was determined that the disordered structure is a transient species which is quickly formed relative to α-helix melting before converting to β-sheet conformation.[61]

1.7. Protein Dynamics

One of the most important applications of UVRR is the study of protein dynamics. UVRR spectroscopy has the potential to allow the characterization of secondary structure changes on a wide range of time scales, illuminating the relationship between structure and function as proteins interact with their environment. In addition, UVRR spectroscopy may allow the determination of the kinetic rates of such transformations, enabling a more complete understanding of protein dynamics. UVRR is highly applicable to folding dynamics of globular proteins, which in general follow the two-state model without significant intermediates.[62-66]

Studies that have probed rapid transformations of secondary structure domains have classically utilized laser induced temperature jumps in small volumes.[6,10,25,7] This requires minor alterations to the experimental setup.[67,68] A variety of systems have been probed using this method including: the transition between the two affinity states of hemoglobin[69], the unfolding of apomyoglobin[7], and characterizing the thermal denaturation of various homo-peptides such as poly-L-lysine, poly-L-proline, and poly-alanine.[9,10,6] The results of these studies have resulted in a proposed mechanism for the formation of β-sheets from meta-stable α-helices[9], the activation barrier of the folding and unfolding of α-helices[6], and revealed evidence for the order of events in protein unfolding.[25]

The application of UVRR spectroscopy to protein dynamics has been limited, but extremely revealing. Potential future applications of UVRR to protein dynamics are vast, but will require more analytic methodologies to resolve the complex underlying features. The development of novel chemometric methods will increase the number of systems that can be investigated using UVRR.

1.8. Refrences

1. Christie, J. M. *Annual Review of Plant Biology* **2007,** *58*, 21-45.

2. Sergiy O. Garbuzynskiy, B. S. M., Michail Yu. Lobanov, Alexei V. Finkelstein, Oxana V. Galzitskaya. *Proteins: Structure, Function, and Bioinformatics* **2005,** *60*, 139-147.

3. Wuthrich, K. *J. Biol. Chem.* **1990,** *265*, 22059-22062.

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

5. Copeland, R. A.; Spiro, T. G. *Journal of the American Chemical Society* **1986,** *108*, 1281-1285.

6. Lednev, I. K.; Karnoup, A. S.; Sparrow, M. C.; Asher, S. A. *J. Am. Chem. Soc.* **1999,** *121*, 8074-8086.

7. Huang, C. Y.; Balakrishnan, G.; Spiro, T. G. *Biochemistry* **2005,** *44*, 15734-15742.

8. Shashilov, V. A.; Lednev, I. K. *Journal of the American Chemical Society* **2008,** *130*, 309-317.

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

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

11. Austin, J. C.; Jordan, T.; Spiro, T. G. In *Biomolecular Spectroscopy;* Clark, R. J. H.; Hester, R. E. Eds.; Wiley & Sons Ltd.: New York, 1993; pp. 55-127.

12. Balakrishnan, G.; Weeks, C. L.; Ibrahim, M.; Soldatova, A. V.; Spiro, T. G. *Current Opinion in Structural Biology* **2008,** *18*, 623-629.

13. Barth, A.; Zscherp, C. *Quarterly Reviews of Biophysics* **2002,** *35*, 369-430.

14. Krimm, S.; Bandekar, J. In *Advances in Protein Chemistry;* Anfinson, C. B.; Edsall, J. T.; Richards, F. M. Eds.; Academic Press: New York, 1986; pp. 181-365.

15. Tu, A. T. In *Spectroscopy of Biological Systems;* Clark, R. J. H.; Hester, R. E. Eds.; Wiley & Sons Ltd.: New York, 1986; pp. 47-111.

16. Balakrishnan, G.; Hu, Y.; Bender, G. M.; Getahun, Z.; DeGrado, W. F.; Spiro, T. G. *Journal of the American Chemical Society* **2007,** *129*, 12801-12808.

17. Mikhonin, A. V.; Bykov, S. V.; Myshakina, N. S.; Asher, S. A. *Journal of Physical Chemistry B* **2006,** *110*, 1928-1943.

18. Haruta, N.; Kitagawa, T. *Biochemistry* **2002,** *41*, 6595-6604.

19. Lednev, I. K.; Karnoup, A. S.; Sparrow, M. C.; Asher, S. A. *Journal of the American Chemical Society* **2001,** *123*, 2388-2392.

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

21. Miura, T.; Hori-i, A.; Mototani, H.; Takeuchi, H. *Biochemistry* **1999,** *38*, 11560-11569.

22. Holtz, J. S. W.; Holtz, J. H.; Chi, Z. H.; Asher, S. A. *Biophysical Journal* **1999,** *76*, 3227-3234.

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

24. Copeland, R. A.; Spiro, T. G. *Biochemistry* **1985,** *24*, 4960-4968.

25. Asher, S. A.; Mikhonin, A. V.; Bykov, S. *J. Am. Chem. Soc.* **2004,** *126*, 8433-8440.

26. Tu, A. T. *Raman Spectroscopy in Biology: Principles and Applications*; Wiley & Sons Ltd.: New York, 1982.

27. Caswell, D. S.; Spiro, T. G. *J. Am. Chem. Soc.* **1986,** *108*, 6470-6477.

28. Wu, Q.; Balakrishnan, G.; Pevsner, A.; Spiro, T. G. *J. Phys. Chem. A* **2003,** *107*, 8047-8051.

29. Zhao, X.; Wang, D.; Spiro, T. G. *Journal of the American Chemical Society* **1998,** *120*, 8517-8518.

30. Zhao, X.; Wang, D.; Spiro, T. G. *Inorganic Chemistry* **1998,** *37*, 5414-5415.

31. Vargek, M.; Zhao, X.; Lai, Z.; McLendon, G. L.; Spiro, T. G. *Inorganic Chemistry* **1999,** *38*, 1372-1373.

32. Ming Xu, V. V. E., Wei He, Vladimir N. Uversky, Laura Fredriksen, Igor K. Lednev. *Biopolymers* **2005,** *79*, 58-61.

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

34. Rosenheck, K.; Doty, P. *PNAS* **1961,** *47*, 1775-1785.

35. Sharma, B.; Bykov, S. V.; Asher, S. A. *J. Phys. Chem. B* **2008,** *112*, 11762-11769.

36. Huang, C. Y.; Balakrishnan, G.; Spiro, T. G. *JOURNAL OF RAMAN SPECTROSCOPY* **2006,** *37*, 277-282.

37. Copeland, R. A.; Spiro, T. G. *Biochemistry* **1987,** *26*, 2134-2139.

38. Simpson, J. V.; Balakrishnan, G.; JiJi, R. D. *The Analyst* **2009,** *134*, 138-147.

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

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

41. Shashilov, V. A.; Ermolenkov, V. V.; Lednev, I. K. *Inorganic Chemistry* **2006,** *45*, 3606-3612.

42. Heyde, M. E.; Gill, D.; Kilponen, R. G.; Rimai, L. *Journal of the American Chemical Society* **1971,** *93*, 6776-6780.

43. Yamamoto, T.; Palmer, G.; Gill, D.; Salmeen, I. T.; Rimai, L. *J. Biol. Chem.* **1973,** *248*, 5211-5213.

44. Strekas, T. C.; Spiro, T. G. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1972,** *278*, 188-192.

45. Brunner, H. *Biochemical and Biophysical Research Communications* **1973,** *51*, 888-894.

46. Spiro, T. G.; Strekas, T. C. *Journal of the American Chemical Society* **1974,** *96*, 338-345.

47. Wozniak, W. T.; Spiro, T. G. *Journal of the American Chemical Society* **1973,** *95*, 3402-3404.

48. Copeland, R. A.; Dasgupta, S.; Spiro, T. G. *Journal of the American Chemical Society* **1985,** *107*, 3370-3371.

49. Williams, R. W. *Journal of Molecular Biology* **1983,** *166*, 581-603.

50. Rava, R. P.; Spiro, T. G. *Biochemistry* **1985,** *24*, 1861-1865.

51. Issei Harada, Y. S. H. M. T. S. *Journal of Raman Spectroscopy* **1975,** *4*, 91-98.

52. Yoko Sugawara, I. H., Hiroatsu Matsuura, Takehiko Shimanouchi. *Biopolymers* **1978,** *17*, 1405-1421.

53. Song, S.; Asher, S. A. *J. Am. Chem. Soc.* **1989,** *111*, 4295-4305.

54. Cheng-Yen Huang, G. B. T. G. S. *Journal of Raman Spectroscopy* **2006,** *37*, 277-282.

55. Chi, Z. H.; Asher, S. A. *Biochemistry* **1998,** *37*, 2865-2872.

56. Ming, X.; Vladimir, V. E.; Wei, H.; Vladimir, N. U.; Laura, F.; Igor, K. L. *Biopolymers* **2005,** *79*, 58-61.

57. Koichi Itoh, B. M. F., Gerald D. Fasman. *Biopolymers* **1976,** *15*, 419-455.

58. Woody, R. In *Circular Dichroism and the Conformational Analysis of Biomolecules;* Fasman, G. Ed.; Plenum Press: New York, 1996; p. 25.

59. Ming, X.; Vladimir, V. E.; Vladimir, N. U.; Igor, K. L. *Journal of Biophotonics* **2008,** *1*, 215-229.

60. Lednev, I. K.; Ermolenkov, V. V.; He, W.; Xu, M. *Analytical and Bioanalytical Chemistry* **2005,** *381*, 431-437.

61. Shashilov, V. A.; Lednev, I. K. *Journal of the American Chemical Society* **2007,** *130*, 309-317.

62. Hagen, S. J.; Hofrichter, J.; Eaton, W. A. *J. Phys. Chem. B* **1997,** *101*, 2352-2365.

63. Gilmanshin, R.; Dyer, R. B.; Callender, R. H. *Protein Sci* **1997,** *6*, 2134-2142.

64. Villegas, V.; Azuaga, A.; Catasus, L.; Reverter, D.; Mateo, P. L.; Aviles, F. X.; Serrano, L. *Biochemistry* **1995,** *34*, 15105-15110.

65. Scalley, M. L.; Yi, Q.; Gu, H.; McCormack, A.; Yates, J. R.; Baker, D. *Biochemistry* **1997,** *36*, 3373-3382.

66. Chen, E.; Lapko, V. N.; Song, P.-S.; Kliger, D. S. *Biochemistry* **1997,** *36*, 4903-4908.

67. Xiaojie, Z.; Charbel, T.; Ruopian, C.; William, R. S.; Thomas, G. S. *Journal of Raman Spectroscopy* **1999,** *30*, 773-776.

68. Bykov, S. L., Igor; Ianoul, Anatoli; Mikhonin, Aleksandr; Munro, Calum; Asher, Sanford A. *Applied Spectroscopy* **2005,** *59*, 1541-1552.

69. Kneipp, J.; Balakrishnan, G.; Spiro, T. G. *J. Phys. Chem. B* **2004,** *108*, 15919-15927.

Chapter 2: Tunable deep-UV Raman Spectrometer for UVRR spectroscopy

2.1. Introduction

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. 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. 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.[1] 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.[2] This requires a laser source for the UVRR instrument capable of generating light in what is commonly called the "deep" ultraviolet region (180-280)[3]. The use of UV radiation dictates that all UVRR spectrometers must be dispersive instruments, featuring a multi-channel detector. This is primarily because of a limitation of Fourier transform (FT) Raman detectors to operate at laser wavelengths above 850 nm, below 850 nm FT-Raman detectors are susceptible to the multiplex disadvantage.[4] The multiplex disadvantage arises in

20

interferometers when the primary source of noise is no longer the detector, i.e. in the visible region and below the primary source of noise is photon noise, noise in the magnitude of photon flux. The simultaneous measurement of all spectral internals becomes a serious disadvantage under higher noise conditions because noise is additive throughout the data, the sum of all noise is applied to all spectral regions, drowning out weaker signals in the spectra.[5]

2.2. Excitation Source

The selection of excitation wavelength, or range of wavelengths, is dependent upon the chromophores of interest and the regions in which their Raman spectra will be resonance enhanced. While a wide variety of light sources exist, the use of laser sources in Raman spectrometers has become common place since the 1960s due to their intensity and monochromatic nature. Table 2-1 lists many commonly available lasers, along with their wavelengths and power ranges. While many of these lasers have the capability of producing light in the deep UV region, many of them, such as the doubled Ar+ or Kr+, are gas based and therefore fixed wavelength sources. A tunable laser source is ideal for UVRR spectroscopy of protiens as the laser can be tuned to selectively excite either the backbone amide or aromatic modes.

Power output of the selected laser source is also a major concern due to the possibility of photo-degradation of protein samples[6] and photo-ionization of histidine residues[7] from high power beams. Commonly, UVRR spectroscopy is carried out with laser powers as low as or lower than 0.5 mW to avoid sample degradation.[7] A system which produces the necessary deep-UV wavelengths, is tunable over a wide range, and operates at a power level which does not require additional optical components to reduce

beam power is the quadrupled titanium-sapphire (Ti:sapphire) laser. Therefore, this laser

system has become the standard excitation laser source for UVRR spectrometers

dedicated to protein spectroscopy.[8-11]

Table 2-1. Commonly available lasers

Type	λ(nm)	Power
Doubled Ar+[4]	244, 257, 229	15-200 mW
Ar+ (air cooled)[4]	488.0, 514.5	5-50 mW
Ar+ (water cooled)[4]	351.1, 479.9, 476.5 488.0, 496.5, 514.5, 528.7	0.1-10 W
He-Ne[4]	632.8	5-100 mW
Kr+[4]	413.1, 647.1, 752.5	0.1-4 W
Nd:YAG[4]	1064	0.1-10 W
Doubled Nd:YLF[4]	527	3-15 W
Doubled Nd:YAG[4]	532	0.05-5 W
Diode[4]	670-865	0.01-1 W
HeAg[12]	224.3	10-50 mW
NeCu[12]	248.6	50-250 mW
Quadrupled Ti:Sapphire[13]	193-240	0.5-20 mW

The Ti:sapphire laser is a doped-insulator laser with the Ti^{+3} ion, present in the

crystal as Ti_2O_3, which acts as the lasing media with the sapphire (Al_2O_3) acting as the

crystalline host generating an infrared laser beam.[12] To generate high powers, the

Ti:sapphire laser requires pumping action by either a more powerful laser, or the more

traditional xenon lamp pumping. The laser system described here (Coherent Inc., Santa

Clara, CA), is pumped by a frequency doubled neodymium-doped yttrium lithium

fluoride (Nd^{+3}:YLF) laser. The tunable output of the Ti:sapphire laser is then directed to

the second harmonic generator as shown in Figure 2-1. The second harmonic generator

(SHG) frequency doubles the Ti:sapphire beam using a lithium triborate (LBO) crystal,

producing a new beam ($\lambda_{Ti:sapphire}/2$). For wavelengths below 210 nm, a dichroic mirror is

then used to separate the remainder of the IR laser beam from the beam generated by the

SHG. These two laser beams are then spatially overlapped before entering the third

harmonic generator (THG). Within the THG the Ti:sapphire beam and second harmonic

beam are converted to a lower wavelength ($\lambda_{SHG}/1.5$) by frequency mixing between the

two beams with a beta barium borate (BBO) crystal. Again, the output beam from the

third harmonic is separated from the remainder of the IR beam with a dichroic mirror. As
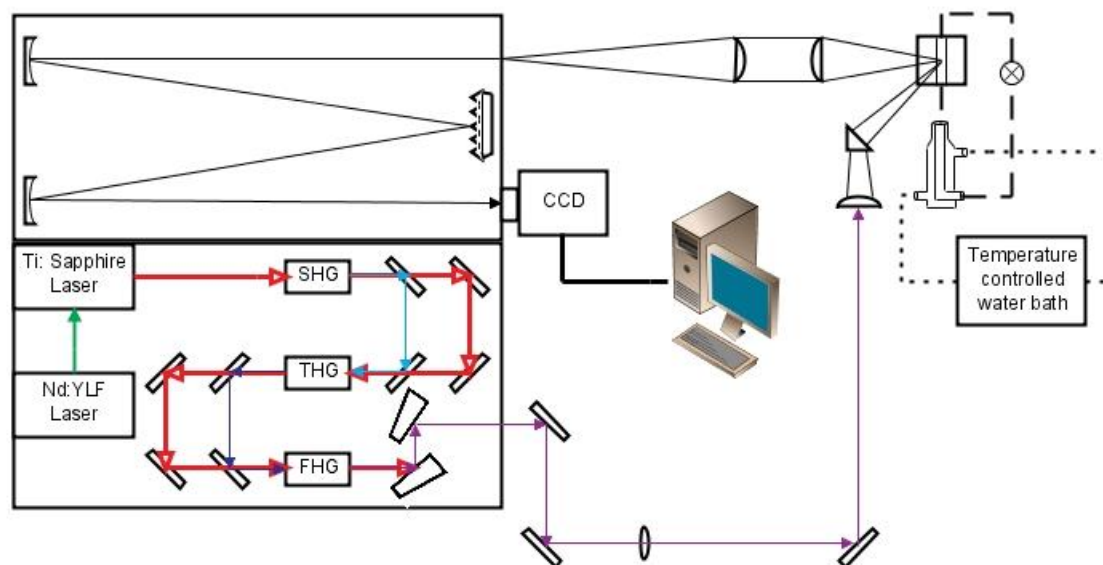
Fig 2-1 Schematic diagram of the UVRR spectrometer. A Nd:YLF laser was used as the pump source for the Ti:sapphire laser. The IR beam of the Ti:sapphire laser was then frequency doubled with a LBO (Lithium Triborate) crystal in the second harmonic generator (SHG). The SHG output is then overlapped and subject to frequency mixing in the presence of a BBO (Beta Barium Borate) crystal in the third and fourth harmonic generators. The quadrupled output beam ($\lambda_{Ti:sapphire}/4$) was directed out of the cavity with two Pellin-Broca prisms(Coherent Inc., Santa Clara, CA) . The output beam was then directed to the sample chamber using a series of mirrors and focusing optics. The wire guided jet sample chamber is shown here. Raman scattering from the sample irradiation was then collected with two plano-convex lenses before being focused on to the entrance slit of the monochromator. Once inside the monochromator, the Raman scatter was then dispersed by a 3600 groove/mm diffraction grating before being detected by a UV-phosphor coated charge coupled device (CCD) (Horiba Jobin Yvon Inc., Edison, NJ) .

in the third harmonic, the two beams are spatially overlapped in the presence of a BBO

crystal resulting in frequency mixing and a new laser beam of lower wavelength ($\lambda_{SHG}/2$

$= \lambda_{Ti:sapphire}/4$) is generated in the fourth harmonic generator (FHG).

In the case of laser excitation wavelengths above 210 nm, the output of the SHG

is frequency doubled again in the FHG with an LBO crystal. In this case there is no need

to separate and then spatially overlap the Ti:sapphire and SHG beams. In either case, the

fourth harmonic output is separeated from the IR, second, and third harmonics through

the use of a Pellin-Broca prism. A second Pellin-Broca prism is used to direct the beam

of interest out of the laser cavity. Once out of the laser cavity, mirrors are used to direct

the beam to the sample chamber.

2.3. Sample Chamber

The output of the FHG 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 design and construction of the

sample chamber can greatly affect the signal to noise ratio. Further, the choice of sample

holder will greatly affect the minimum volume needed. Two factors must be kept in mind

when designing the sample chamber. First, the absorption of materials in the deep-UV

must be considered. Even high grade quartz, the material most often used for construction

of cuvettes for absorption or fluorescence spectroscopies, absorbs greatly at 193 nm. The

absorption of light from the sample holder can greatly reduce the excitation power,

thereby reducing the Raman scatter and signal intensity. The second factor which must be

considered is the possibility of damaging the sample with high quantities of high energy
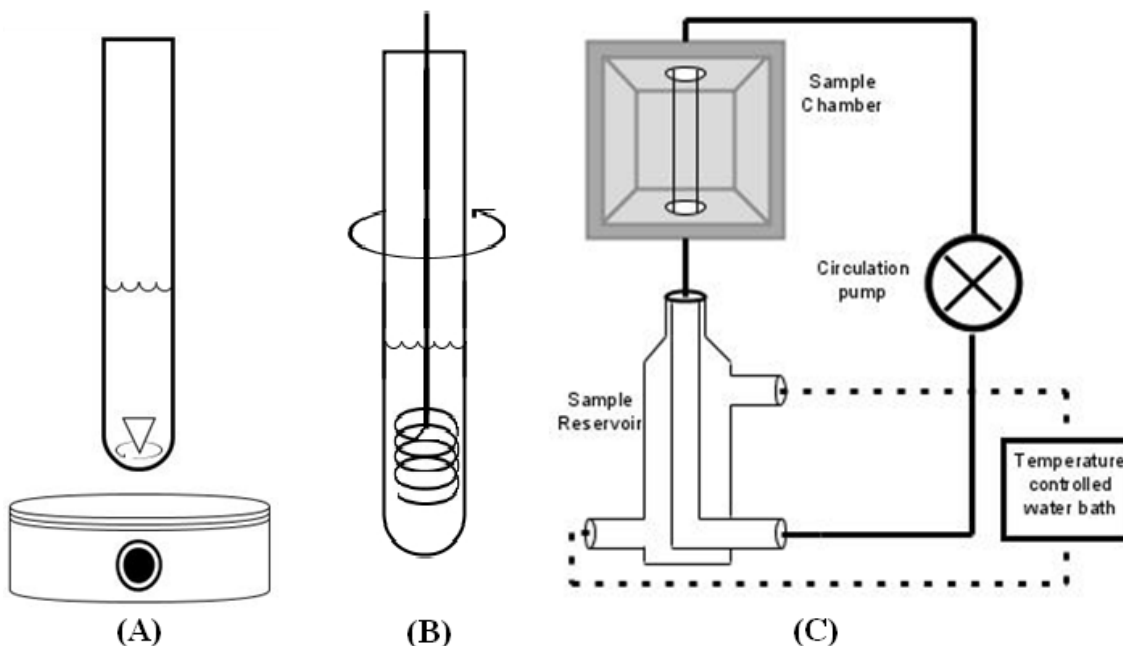
Fig 2-2 (A) Cuvette sample cell design featuring a magnetic stirrer. (B) Cuvette sample cell design featuring a helical wire stirrer. In this design, the wire is held stationary while the entire tube is rotated. (C) Diagram of open flow sample chamber. The solid lines denote the path of the sample solution, while the path of the chilled water used to regulate the temperature is noted in dotted lines. In this design, the sample solution is kept in a sample reservoir (Mid Rivers Glassblowing, Saint Charles, MO) which is temperature controlled through water circulation (Isotemp 3016D, Fisher Scientific, Pittsburgh, PA). The solution is pumped out of the bottom of the sample reservoir to the top of the sample chamber using a model 75211-10 gear pump (Cole Parmer, Vernon Hills, Illinois) where it is then allowed to flow between two wires, creating a thin film. The thin film is irradiated by the excitation beam. The sample solution is then allowed to drain into the sample reservoir.

photons through photo-ionization of the histadine residues. Reduction of the excitation beam power reduces the chances of damaging the sample, but in addition, the sample may be continuously circulated.[14] These factors have lead to the development of two main types of sample chambers, the cuvette system[10] and the free-flowing jet system.[15]

The cuvette system consists of a suprasil quartz container, the highest grade optical quartz available, which features a relatively low absorption in the deep-UV compared to more conventional materials. The solution is either stirred, or the tube itself is spun to prevent the same volume of sample from being constantly exposed to the

excitation beam. Several variations of this idea exist, some of which are highlight in

Figure 2-2. One of the most common variations involves the rotation of the quartz tube

around a stationary wire which serves to stir the solution.[16] Other variations involve the

use of a conventional magnetic stirrer situated at the bottom of the tube[17], or rotating the

tube at high speed to stir the solution.[18,19] However, these designs require higher

excitation wavelengths, typically greater than 206 nm, and are susceptible to frothing. We

have opted to both rotate the tube at high speed, and to move the tube up and down

vertically to minimize sample degradation and frothing. Cuvette based systems have the

advantage that relatively low sample volumes, between 100-300 µL, may be used.

Alternatively free-flowing jet systems eliminate the problem of absorption by the

sample cell by creating a thin film of liquid, which is directly irradiated by the excitation

beam. The sample solution is typically contained in a temperature controlled reservoir

and circulated throughout the system by either a gear pump or peristaltic pump. The

spectrometer detailed in Figure 2-1 shows a thin film being created between two wires.[8]

Many permutations of this design exist, often varying the method by which the film is

created. For example, a spray nozzle such as one used in a dye laser has been used to

create the thin film of sample solution.[20] Another method was to allow the sample to

travel a short distance under high pressure in air [11,21], during which time the sample was

irradiated. In all cases, a cell free design yields a higher signal to noise ratio compared to

the cuvette design, but the sample volume necessary for this type of sample holder are

substantially higher than the cuvette system, ranging from six to ten mL.[11]

Although the selection of sample chamber design is a critical step in the

construction of a UVRR spectrometer, with careful planning it is possible to design the

system so that the two sample chambers are interchangeable. Being able to interchange these sample chambers allow for maximum flexibility in the instrument. Once the sample has been irradiated, it will emit Raman scatter which is collected and focused on the entrance slit of the monochromator with the use of two plano-convex lenses. The monochromator diffracts the Raman scatter and directs it to the detector. The monochromter used in this system was a 1.25 m spectrometer (Horiba Jobin Yvon Inc., Edison, NJ) equipped with a 3600 grove/mm grating.

2.4. Detector/Collection Optics

The optimum position of the collection optics can be approximated by treating the point of irradiation as a point source. Once the Raman scatter has been focused on to the entrance slit of the monochromator, the diffraction grating disperses the signal. The dispersion element diffracts the collected Raman scatter, spatially separating the incident light by the energy of the light. The diffracted Raman scatter is then reflected on to the detector.

As already mentioned, nearly all UVRR spectrometers are dispersive instruments due to the lack of FT detectors which work well below 850 nm. Multi-channel detectors are usually employed in such instruments, since they eliminate the need for a second monochromator. Multi-channel detectors that have been widely utilized for UVRR instrumentation include the charged coupled device (CCD)[10,19] and the intensified photo-diode array detector (IPAD)[8,9]. The detector used in this system is a back illuminated, phosphor coated, liquid nitrogen cooled Symphony CCD camera (Horiba Jobin Yvon Inc., Edison, NJ) with a chip size of 2048 x 512.

The CCD array is a silicon based integrated circuit which converts light to an electrical charge through the photoelectric effect.[22] Structurally, a CCD is constructed from a series of treated silicon layers. The first two layers consist of *p* and *n* doped silicon, creating a *p* and *n* semiconductor junction. This is topped with a layer of silicon dioxide which acts as an insulating layer followed by a series of polycrystalline silicon electrodes.[23] For the basic type of CCD, a front-illuminated photogate, light is passed through the electrodes and insulation layer before reaching the doped silicon where electron-hole pairs are generated by the absorption of incident photons. The resulting charge is then collected in the potential wells, also known as pixels, of the CCD through diffusion to the collection site or depletion region, which exists between the doped silicon species.[24]

Employment of CCD sensors in the deep UV region is problematic due to the increased absorption of UV radiation by silicon and silicon dioxide. The response of a CCD detector rests on the ability of a photo to essentially pass through the various layers of the CCD and approach the depletion region of the *pn* junction.[24] The closer to the depletion region incident photons can travel before being absorbed, the more likely the resulting generated charge will contribute to the response of the CCD. However, as the wavelength of the photons decrease, their potential absorption by silicon and silicon dioxide increases.[23] The result is that photons which encounter the overlaying silicon dioxide layer will most likely be absorbed and fail to contribute to the signal charge generation. Photons not absorbed by the silicon dioxide layer will likely be absorbed by silicon very close to the silicon-silicon dioxide interface, making it very difficult to

collect the generated charge. These effects greatly decrease the ability of CCD detectors to respond to the collected scatter in the deep-UV.[23]

A variety of solutions have been developed to overcome the challenges of working with CCD detectors in the deep-UV region. The two most common methods employed are to coat the CCD with a UV-sensitive phosphor or switch to a backside-thinned back illuminated architecture. Each of these methods has advantages in terms of cost and quality of response.

Application of a UV-phosphor coating is the simplest, and cheapest, method to effectively detect deep-UV photons using a CCD. A standard CCD sensor is coated with a UV-sensitive phosphor coating on the front of the CCD.[24] This coating serves to absorb the incident UV light and emit at a wavelength that the conventional CCD will respond well, usually in the visible range. The most common coating material, coronene is a highly conjugated polycyclic aromatic hydrocarbon [25] which absorbs well below 400 nm and fluoresces in the green portion of the visible spectrum.[26] While this solution is simple and cost effective, consisting of only a short post manufacturing procedure, major disadvantages of this technique exist. A loss of spectral resolution of the CCD is experienced due to light scattering on the coating of the CCD. The UV-phosphor coating can be damaged through photo-degradation, eliminating the ability of the CCD to detect deep-UV light. In addition, the quantum efficiency of the CCD is dramatically reduced due to the isotropic (uniform in all directions) nature of fluorescence emission. When the coating fluoresces, only half of the photons emitted will travel to the CCD sensor, cutting the quantum efficiency of the detector in half.

The thin-back back-illuminated CCD is considered to be the most sensitive CCD option available for the detection of deep-UV radiation, greatly surpassing the UV-phosphor coated CCD design in sensitivity.[24] However, the thin-back back-illuminated CCD design is more complex to manufacture and is also the most expensive to construct. The thin-backed back-illuminated design is constructed by significantly thinning the depth of the CCD sensor doped silicon layers from ~500 μm to 45 μm or less.[27] In addition, the CCD sensor is flipped so that the incident light encounters the doped silicon layers first. By targeting the illumination on the backside of the sensor the problem of light absorption due to the polycrystalline silicon electrodes is avoided.[28] This allows for most of the incident light to be absorbed by the silicon semiconductor. Thinning the back of the CCD allows for the generation of electron-hole pairs to occur closer to the site of collection, the depletion region, increasing the likelihood of a photon contributing to the generation of charge response.[28] Due to the high cost, the thin-back back-illuminated CCD is not always the preferred detector despite its superior spectral response.

2.5. References

1. Austin, J. C.; Jordan, T.; Spiro, T. G. In *Biomolecular Spectroscopy;* Clark, R. J. H.; Hester, R. E. Eds.; Wiley & Sons Ltd.: New York, 1993; pp. 55-127.

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

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

4. McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc.: New York, 2000.

5. Plankey, F. W.; Glenn, T. H.; Hart, L. P.; Winefordner, J. D. *Analytical Chemistry* **2002,** *46*, 1000-1005.

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

7. Wu, Q.; Balakrishnan, G.; Pevsner, A.; Spiro, T. G. *J. Phys. Chem. A* **2003,** *107*, 8047-8051.

8. Balakrishnan, G.; Hu, Y.; Nielsen, S. B.; oslash; ndsted; Spiro, T. G. *Applied Spectroscopy* **2005,** *59*, 776-781.

9. Zhao, X.; Chen, R.; Tengroth, C.; Spiro, T. G. *Applied Spectroscopy* **1999,** *53*, 1200-1205.

10. Lednev, I. K.; Ermolenkov, V. V.; He, W.; Xu, M. *Analytical and Bioanalytical Chemistry* **2005,** *381*, 431-437.

11. Bykov, S. L., Igor; Ianoul, Anatoli; Mikhonin, Aleksandr; Munro, Calum; Asher, Sanford A. *Applied Spectroscopy* **2005,** *59*, 1541-1552.

12. Hollas, J. M. *Modern Spectroscopy*, 4 th ed.; John Wiley & Sons, Ltd., 2004.

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

14. Austin, J. C.; Rodgers, K. R.; Spiro, T. G.; James, F. R. a. B. L. V. In *Methods in Enzymology*; Academic Press, 1993; pp. 374-396.

15. Reider, G. A.; Traar, K. P.; Schmidt, A. J. *Appl. Opt.* **1984,** *23*, 2856-2857.

16. Rodgers, K. R.; Su, C.; Subramaniam, S.; Spiro, T. G. *Journal of the American Chemical Society* **1992,** *114*, 3697-3709.

17. Peter, J. L.; William, G. G.; Sanford, A. A. *The Journal of Chemical Physics* **1991,** *94*, 5324-5330.

18. Hideo Takeuchi, I. H. *Journal of Raman Spectroscopy* **1990,** *21*, 509-515.

19. Russell, M. P.; Vohník, S.; Thomas, G. J. **1995,** *68*, 1607-1612.

20. Sanford, A. A.; Craig, R. J.; James, M. *Review of Scientific Instruments* **1983,** *54*, 1657-1662.

21. Fodor, S. P. A.; Rava, R. P.; Hays, T. R.; Spiro, T. G. *Journal of the American Chemical Society* **1985,** *107*, 1520-1529.

22. Douglas S. Skoog, F. J. H., and Timothy A. Nieman *Principles of Instrumental Analysis*; Saunders College Publishing: New York, 1997.

23. Theuwissen, A. *Solid-State Imaging with Charge-Coupled Devices*; Kluwer Academic Publishers: Boston, 1995.

24. Li, F. a. A. N. *CCD Image Sensors in Deep-Ultraviolet: Degradation Behavior and Damage Mechanisms* 1ed.; Springer, 2005.

25. Ciesielski, A.; Cyranski, M. K.; Krygowski, T. M.; Fowler, P. W.; Lillington, M. *The Journal of Organic Chemistry* **2006,** *71*, 6840-6845.

26. Ito, T. M., Koji; Hiramatsu, Mitsuo. In *Japanese Kokai Tokkyo Koho;* Office, J. P. Ed.; HAMAMATSU PHOTONICS Japan, 1993; p. 5.

27. Jones, T. J.; Deelman, P. W.; Elliott, S. T.; Grunthaner, P. J.; Wilson, R.; Nikzad, S. "Thinned charged-coupled devices with flat focal planes for UV imaging"; Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications, 2000, San Jose, CA, USA.

28. Jansick, J. R. *Scientific Charge-Coupled Devices*, 1st ed.; SPIE Publications: Bellingham, 2001.

Chapter 3: MCR-ALS analysis of two-way UV resonance Raman spectra to resolve discrete protein secondary structural motifs[*]

The ability of ultraviolet resonance Raman (UVRR) spectroscopy to monitor a host of structurally sensitive protein vibrational modes, the amide I, II, III and S regions, makes it a potentially powerful tool for the visualization of equilibrium and non-equilibrium secondary structure changes in even the most difficult peptide samples. However, it is difficult to unambiguously resolve discrete secondary structure derived UVRR spectral signatures independently of one another as each contributes an unknown profile to each of the spectrally congested vibrational modes. This limitation is compounded by the presence of aromatic side chains, which introduce additional overlapping vibrational modes. To address this, we have exploited an often overlooked tool for alleviating this spectral overlap by utilizing the differential excitability of the vibrational modes associated α-helices and coil moieties, in the deep UV. The differences in the resonance enhancements of the various structurally associated vibrational modes yields an added dimensionality in the spectral data sets making them multi-way in nature. Through a "chemically relevant" shape constrained multivariate curve resolution-alternating least squares (MCR-ALS) analysis, we were able to deconvolute the complex amide regions in the multi-excitation UVRR spectrum of the protein myoglobin, giving

us potentially useful "pure" secondary structure derived contributions to these individual vibrational profiles.

3.1. Introduction

The applications of x-ray and more advanced NMR techniques to derive atomic resolution structures in a variety of protein structure types has truly brought forth the structural era in biochemistry and biophysics. However, even with the plethora of information gained, limitations in these techniques have become conspicuous in that many protein's secondary structure contents are not static enough within the time-scale of these techniques to be easily determined. These voids have brought about the re-emergence and advancement of classical protein secondary structure sensitive spectroscopic techniques, such as circular dichroism (CD), IR and Raman scattering derived spectroscopies. One such vibrational technique enjoying a revolution in use and interpretation is ultraviolet resonance Raman (UVRR) spectroscopy. UVRR is proving to have tremendous potential to study protein secondary structures and their dynamics in even the most difficult or problematic protein samples. Early work has established that UVRR can be used effectively to quantify relative amounts of α-helical, β-sheet and various disordered-types of secondary structures[1,2]. Recently, for instance, Huang *et al*. have added β-turn structure to the types of secondary structures that are quantifiable at excitation wavelengths below 200 nm[3]. UV excited Raman spectroscopy is unique in that selective enhancement of chromophores of interest, such as amide and or aromatic side chains, eliminates interference from the rest of the protein and thus simplifies the spectral analysis. Further, water is a weak Raman scatterer, so the UVRR spectra of proteins are nearly free of water interference, allowing structural studies to be performed in aqueous,

non-deuterated, environments. The structural sensitivity of vibrational spectroscopies (IR and Raman based) is largely derived from the polypeptide backbone amide group. The amide group, —CO–NH—, gives rise to several discrete vibrational modes, collectively termed amide modes. The position and intensity of these modes shift with changes in the ϕ and ψ dihedral angles of the polypeptide backbone[4]. There have been several good reviews on the sensitivity of the amide modes to protein secondary structure[5-9]. Deep-UV excitation (< 210 nm) results in resonance enhancement of the amide I, II, III and S modes, along with the aromatic side chains (tyrosine, phenylalanine and tryptophan)[5]. The position and intensity of the four amide modes are dependent upon the secondary structure of the protein with their relative contributions being proportional to the relative amount of each conformation. Likewise, the position and intensity of the various aromatic modes will be dependent upon their environment (hydrophobic/buried or hydrophilic/exposed) and their relative abundance.

UVRR has been applied to protein folding and unfolding studies under a host of conditions of varying pH and ionic strength[10-14]. Additionally the kinetics of structural transitions from one type to another have also been studied through the use of temperature-[4, 15-19]and pH-[20] jumps.

While changes in the position, intensity and overall spectral shape of the amide modes are dependent on secondary structure, they remain highly overlapped in the UVRR spectra of proteins. Additionally, the UVRR bands from the aromatic residues, tyrosine, tryptophan, and phenylalanine also overlap the amide regions, complicating the spectra further. Typically, individual spectra are systematically deconvoluted using nonlinear least squares (NLLS). However, these methods are

dependent on initialization parameters and constraints. When the individual Raman bands are not sufficiently resolved, the choice of initialization parameters and constraints often dictate the final solution rather than the spectroscopic data itself. Alternatively, multivariate methods could easily be applied to the afore mentioned data, greatly simplifying the analysis process. Furthermore, additional components may be inculed to account for contributions from overlapping aromatic modes or other non random spectral interferents.

With all the advantages of multivariate analysis, it is surprising that the application of chemometric methods to bilinear UVRR data is a relatively recent phenomenon [21-23]. As resonance enhancement occurs when the excitation wavelength is at or near the electronic absorption band of the analyte, the excitation wavelength may be tuned to differentially excite the various structures. The observed absorption maxima of the amide transitions in the polypeptide backbone are similar to those observed for *N*-methyacetamide (NMA), which has three electronic transitions at 165 nm ($S_3 \leftarrow S_0$), 185 nm ($S_2 \leftarrow S_0$) and 210 nm ($S_1 \leftarrow S_0$)[5]. However, the absorption maxima vary slightly with the conformation of the peptide backbone[24]. Likewise, the molar scattering ratio (excitation profile) of the backbone amide is conformation dependent[1, 3, 25, 26]. Thus, every protein UVRR spectrum will be dependent upon both the excitation wavelength and secondary structural composition of the protein. Varying the excitation wavelength or secondary structural composition would result in bilinear UVRR data that could be analyzed using a multitude of multivariate methods.

Bilinear model based chemometric techniques have been shown to be extremely useful in the investigation of mixture based chemical systems, especially

factor analysis based techniques, like multivariate curve resolution (MCR)[27].

Coupling MCR with iterative algorithms, such as alternating least squares (ALS), has

produced techniques which are flexible enough to cope with many different kinds of

data structures and chemical problems[28-32]. For example, MCR-ALS has been

successfully applied to resolve kinetic concentration profiles from reaction based

chemical sensors [33], intermediates in protein folding[34] and the solvation of solutes as a

function of solvent composition[35].

In this study, UVRR spectra of a single protein, myoglobin, collected at six

different deep UV excitation wavelengths, were used as a novel multivariate data set.

Myoglobin, is a well studied model protein with a mostly α-helical structure. Thus,

myoglobin is an attractive model system for the determination of a"pure" α-helix

basis spectrum and the associated steady state amide mode peak positions. Through

the use of a novel chemically relevant shape constraint, imposed within the MCR-

ALS algorithm, analysis of this multi-way data revealed what we believe are pure α-

helical spectral profiles for the amide I, III and S regions versus excitation

wavelength. Although a unique α-helical profile could not be resolved for the amide

II region, the resolved spectral profiles could be used as starting parameters for

UVRR analysis in the future.

3.2. Experimental

3.2.1 Sample preparation

Myoglobin was dissolved in 0.1 M phosphate buffer at pH 7 to a final

concentration of 0.2 mM. Sodium perchlorate (0.2 M) was added to the sample as an

internal intensity standard. Myoglobin and sodium perchlorate were purchased from

Sigma (St. Louis, Missouri). The average laser power was 1.5 mW for all excitation wavelengths and each spectrum was averaged for 15 minutes.

3.2.2 Raman spectroscopy

The UV Raman spectrometer used in data collection was previously described by Spiro and coworkers[36]. Specifically, the excitation source was the fourth harmonic of a tunable Ti:Sapphire laser (Photonics International), which was pumped by a frequency-doubled Nd:YLF laser. The sample solution was circulated through a wire guided cell and held at a constant 25°C, using a circulating water bath and a water jacked reservoir. The UVRR spectra of myoglobin were collected at six excitation wavelengths for a total of six spectra. (193 nm, 197 nm, 200 nm, 203 nm, 206 nm, 210 nm). All the spectra were calibrated using the Raman spectra of cyclohexane and acetone.

3.3. Data analysis

3.3.1 Data pre-processing

In Raman dispersive instruments, the mechanical grating drive follows a sinusoidal function and the dispersion along the spectral axis is linear with respect to "wavelength (nm)" and nonlinear with respect to "wavenumber (cm^{-1})". As a result, an instrument with 5 nm spectral bandbass will cover about 1308 cm^{-1} at 196 nm (193-198) excitation whereas the same instrument will cover about 1140 cm^{-1} at 210 nm excitation. Chemometric methods are based on matrix operations and require the same number of data points (equal length vectors) for spectra measured at multiple wavelengths. For this study, periodic datums were removed at higher excitation wavelengths ($\lambda_{Ex} > 193$ nm) and the spectra were aligned to within 0.5 cm^{-1} based on

their associated calibration. Alignment to within a 0.5 cm^{-1} was considered sufficient for this study as the spectral resolution ranged from 1.35 cm^{-1} at 193 nm to 1.13 cm^{-1} at 210 nm, at least twice the possible error in the alignment. Although removal of a nominal number of selected datums could potentially reduce the spectral resolution and introduce slight alignment errors, the simplicity of this method and the minimal distortion to the original spectral data led us to choose this method for our initial studies. In future studies, the effect of alternate methods, including the use of linear and spline interpolants to obtain equal number of points among different excitation wavelengths will be investigated.

3.3.2 Non-linear least squares (NLLS) analysis

For comparison, each spectrum was initially deconvoluted using a NLLS fitting algorithm. NLLS analysis was performed in the Matlab environment (Mathworks, Natick, MA) using the standard NLLS optimization function available with the optimization toolbox. The spectral profiles were constrained to have a mixture of Guassian/Lorentzian distributions. Each band was defined by four parameters: position (center), height, width, and fraction of Lorentzian content. The fitted spectral profiles from the UVRR spectrum excited at 193 nm (Figure 3-1b) were used as the initialization parameters for the constrained MCR-ALS models. This will be discussed in more detail in the next section.
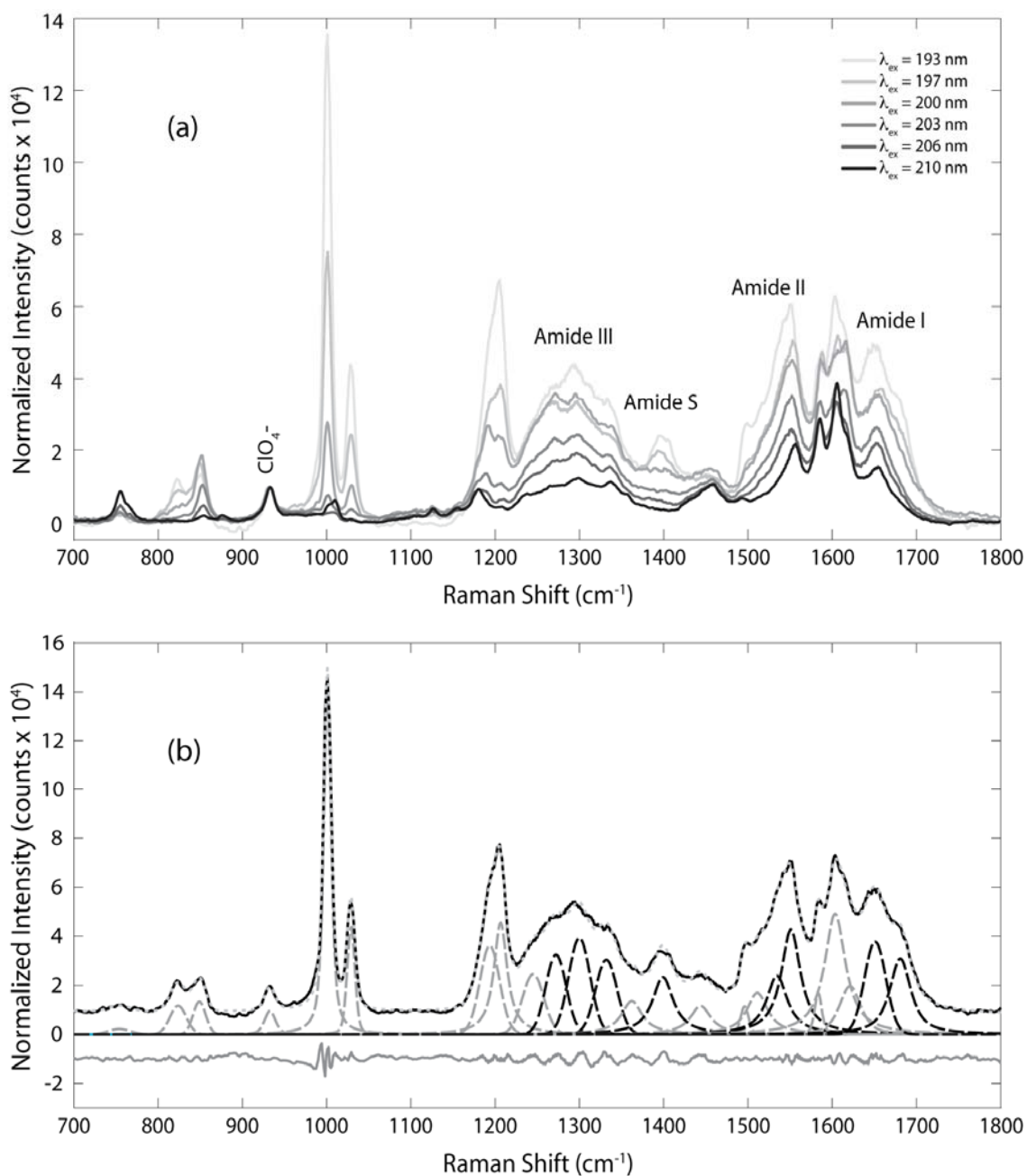
Fig. 3-1 (a) Normalized deep UVRR spectra of myoglobin, collected at excitation wavelengths from 193 nm (light gray) to 210 nm (black). Spectra were normalized to the internal intensity standard, perchlorate (ClO_4^-). (b) Nonlinear least squares fit (- - -) of the UVRR spectrum of myoglobin (——), excited at 193 nm. A series of mixed Gaussian/Lorentzian line profiles (*dashed lines*) were used to fit the entire experimental spectrum. Components denoted with black dashed lines (– –), were used as initialization profiles for the respective amide region. The residuals are denoted with in gray (——).

3.3.3 Multivariate curve resolution - alternating least squares analysis (MCR-ALS)

3.3.3.1 MCR-ALS analysis

The ALS algorithm that we have utilized was written in house and is based on that outlined by Bro and Sidiropoulos[37]. 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[38]. Given that UVRR spectra collected at multiple excitation wavelengths are bilinear in nature, the resultant data matrix ($\mathbf{X}$, $I \times J$) may be decomposed into the Raman spectral profiles ($\mathbf{A}$, $I \times N$ matrix) and excitation profiles or cross sections ($\mathbf{D}$, $J \times N$ matrix) plus an error matrix ($\mathbf{E}$) of the same dimensions as the original data matrix. The data matrix, $\mathbf{X}$, may then be reproduced according to Equation $\mathbf{1}$[33].

$$\mathbf{X} = \mathbf{A}\mathbf{D}^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

Equation $\mathbf{1}$ is solved iteratively using an ALS algorithm. This optimization procedure requires reasonable initial estimates of either the excitation ($\mathbf{D}$) or spectral ($\mathbf{A}$) profiles[39]. At each iteration of the optimization, $\mathbf{A}$ or $\mathbf{D}$ is optimized, while the opposing matrix is held constant.

As mentioned above, the spectral profiles ($\mathbf{A}$) were initialized using the deconvoluted spectral bands from NLLS analysis of the UVRR spectrum of myoglobin (λ_{ex} = 193 nm). Subsequently, the excitation profiles or relative cross-sections were calculated according to Equation $\mathbf{2}$.

$$\mathbf{D} = \mathbf{X}^{\mathrm{T}}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \qquad (2)$$

Non-negativity constraints were applied to both the spectral and excitation profiles, by setting all negative values to an infinitesimally small value. Although setting

negative values to zero or a relatively small value is not a least squares solution, it was sufficient to eliminate convergence to unreasonable solutions. Additionally, the spectral profiles were constrained to a mixed Gaussian/Lorentzian profile by minimizing $\| \mathbf{C}_n - \mathbf{d}_n\mathbf{a}_n^T \|$ subject to the vector $\mathbf{a}_n$ being a Gaussian/Lorentzian profile. $\mathbf{C}_n$ represents the two-dimensional UVRR spectrum of the nth component (Equation **3**),

$$\mathbf{C}_n = \mathbf{d}_n\mathbf{a}_n^T + \mathbf{E} \qquad (3)$$

where $\mathbf{E} = \mathbf{X}\text{-}\mathbf{DA}^T$. The pure spectral profile of the nth component was calculated according to Equation **4.**

$$\mathbf{a}_n = \mathbf{d}_n^T\mathbf{C}_n(\mathbf{d}_n^T\mathbf{d}_n)^{-1} \qquad (4)$$

Bro and Sidiropoulos showed that minimization of $\| \mathbf{C}_n - \mathbf{d}_n\mathbf{a}_n^T \|$ was equivalent to minimizing $\| \beta - \mathbf{a}_n \|$ where $\beta = (\mathbf{d}_n^T\mathbf{C}_n)/(\mathbf{d}_n^T\mathbf{d}_n)$ and $\mathbf{a}_n$ is constrained to unimodallity, non-negativity, etc.... [37] This allows the easy implementation of the Gaussian/Lorentzian peak shape constraint through the incorporation of a NLLS optimization step that minimizes $\| \beta - \mathbf{a}_n \|$ subject to $\mathbf{a}_n$ being a Gaussian/Lorentzian profile. The use of nonlinear optimization methods to enforce chemically relevant constraints has been used previously to determine reaction rate constants from three-way LC-DAD data[40].

3.3.3.2 Initialization of nested NLLS optimization

Each spectral profile was constrained to a Gaussian/Lorentzian profile and each component was described by four parameters: position, height, width, and fraction of Lorentzian. To minimize the influence of user defined parameters on the final model, each peak parameter was initialized using pre-defined or systematically

determined values. Peak width and fraction of Lorentzian content were initialized

using pre-defined generic values of 50 cm^{-1} and 0.5, respectively. The peak height

parameter was initialized as one half the maximum intensity. The peak center was

estimated by fitting the cumulative sum of the unconstrained n^{th} spectral profile (β) to

a sigmoidal function (Equation **5**, Figure 3-2)

$$f(x) = y_{min} + \left(\frac{y_{min} - y_{max}}{1 + e^{-(x-x_0)/w}} \right) \quad (5)$$

where y_{min} and y_{max} and are the minimum and maximum intensity of the curve, x is the

spectral range, x_0 is the center of the estimated curve. Thus, the peak position, x_0, was

initialized using the experimental data rather than an initial guess. This method was

more efficient that using the position of maximum spectral intensity. This was most

likely due the higher level of noise in the unconstrained spectral profiles of highly

overlapped modes, such as in the amide III region.
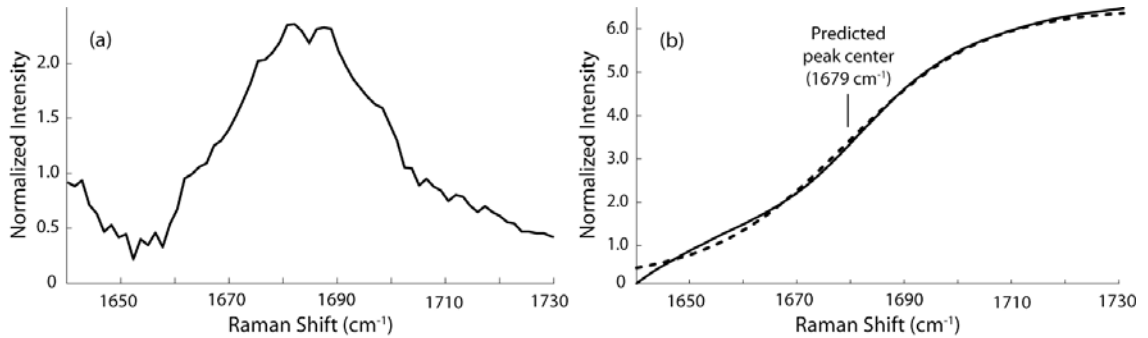
Fig. 3-2 (a) Second unconstrained component from MCR-ALS fitting of the amide I region. (b)
Cumulative sum of the second unconstrained component from MCR-ALS fitting of the amide I region (——)
and the corresponding fitted sigmoidal curve (- - -).

3.3.3.3 Convergence

The convergence criteria was based on the scaled sum of the squared

differences (*SSD*) between successive iterations according to Equation **6**, where

$\text{Vec}(\hat{\mathbf{X}}_i)$ is the vectorized form of the matrix $\hat{\mathbf{x}}_i$ at the ith iteration. *SSD* represents the difference between the current ($\hat{\mathbf{x}}_i$) and previous ($\hat{\mathbf{x}}_{i-1}$) estimates of the data matrix $\mathbf{X}$. As *SSD* becomes small, it indicates that the difference between successive estimates is decreasing. Convergence was defined as when *SSD* fell below a sufficiently small value, signaling that successive estimates were essentially identical.

$$SSD = \left(\frac{\text{Vec}(\hat{\mathbf{X}}_i)}{\left\| \text{Vec}(\hat{\mathbf{X}}_i) \right\|_F} - \frac{\text{Vec}(\hat{\mathbf{X}}_{i-1})}{\left\| \text{Vec}(\hat{\mathbf{X}}_{i-1}) \right\|_F} \right)^T \left(\frac{\text{Vec}(\hat{\mathbf{X}}_i)}{\left\| \text{Vec}(\hat{\mathbf{X}}_i) \right\|_F} - \frac{\text{Vec}(\hat{\mathbf{X}}_{i-1})}{\left\| \text{Vec}(\hat{\mathbf{X}}_{i-1}) \right\|_F} \right) \qquad (6)$$

SSD values were also calculated for the excitation and spectral profiles. Convergence was defined as the point at which all three of the calculated *SSD* values fell below the pre-defined threshold.

3.3.3.4 Model evaluation

The quality of each model was determined by its ability to accurately represent the data. Evaluation of the model was performed by two methods: visual inspection and using the sum of the squared residuals (*SSR*), where $\mathbf{AD}^T$ corresponded to the model and $\mathbf{X}$ corresponded to the experimental data (Equation **7**).

$$SSR = \text{Vec}(\mathbf{X} - \mathbf{AD}^T)^T \times \text{Vec}(\mathbf{X} - \mathbf{AD}^T) \qquad (7)$$

This operation was performed iteratively; a decrease in the *SSR* indicated an improvement in the quality of the fit of the model to the experimental data. Furthermore, the *SSR* value was used to evaluate the performance of the unconstrained, fully and partially constrained models with respect to one another.

3.3.4 Post-processing

Calculation of the relative cross sections (σ) for each component was performed using Equation **8**[41].

$$\sigma_N = \sigma_S \times \left(\frac{I_N}{I_S}\right) \times \left[\frac{(\nu_0 - \nu_S)}{(\nu_0 - \nu_N)}\right]^4 \times \left(\frac{C_S}{C_N}\right) \qquad (8)$$

The cross section of each spectral band was determined from the ratio of peak intensities (I_N/I_S), concentrations (C_S/C_N) along with a frequency term, where the letters N and S, designate sample and internal standard, respectively. The frequency term is composed of the ratio of the difference between the laser excitation frequency (ν_0) and the vibrational frequencies of the sample (ν_S) and standard (ν_N), to the fourth power. In effect, the frequency term utilizes the ratio of the selected peak position for the sample and the peak position of the internal standard. All of these terms are multiplied by the cross section of the internal standard (σ_s), which was taken from Balakrishnan et al.[36] Cross sections are typically expressed in units of millibarns/(molecule·steradian).

Although cross sections are often determined from peak area rather than peak height, peak area can be much more susceptible to error arising from inaccuracies in the estimation Lorentzian content (see Table 3-1). Inadequate baseline correction can result in over estimation of the Lorenzian content, which results in over estimation of the cross sections due to the more extended wings of Lorentzian bands. Given that the peak width and fraction of Lorentzian content should be independent of excitation wavelength, and are modelled as such with MCR-ALS, the ratio of peak intensities rather than areas was used in this study. Although this will result in differences in the relative magnitude of the cross sections, it will have no effect on the shape of the excitation profile, which should be similar to the shape of the UV absorption spectrum[25, 26, 36].

3.4. Results And Discussion

3.4.1 NLLS analysis of UVRR spectra

A common approach to spectral deconvolution is to fit each spectrum to a series of Gaussian-, Lorentzian- or a Gaussian/Lorentzian line shapes using NLLS optimization methods. These methods usually produce excellent approximations of a global vibrational spectral line shape, as can be seen in Figure 3-1b. However, the minimized solution depends upon parameters defined by the user at the outset of the optimization, including the number of components, the initial peak parameters (height, width, position, shape) and the various constraints to be applied to these peak parameters during the optimization. One means by which to alleviate this inherent pre-analysis user bias is to globally fit data sets where individual components can vary independently as a function of some designed variable, such as excitation wavelength, pH, temperature, etc… This added dimensionality limits the final solution to a global minimum, although still as a function of the initialization parameters chosen. UVRR spectra of myoglobin, collected at multiple excitation wavelengths, comfortably fulfill this strategic design principle (Figure 3-1a).

The results of NLLS fits of this data (Table 3-1) clearly show that the combination of random noise and non-random error, possibly arising from small changes in the alignment or calibration of the instrument, significantly influence the fitted parameters. In general, we find that NLLS estimates of lower intensity bands have a higher degree of error associated with them than more intense spectral features, as evidenced by Raman bands arising from non-helical secondary structure, such as the minimally present β-turn and coiled regions, (see Table 3-3 for

Table 3-1. Peak parameter estimation from NLLS analysis of UVRR spectra of myoglobin at each excitation wavelength

	Excitation λ (nm)							
	193	197	200	203	206	210	$\bar{x}$	σ
Amide III								
Center (cm^{-1})	1262	1264	1265	1265	1268	1258	1264	3
Width (cm^{-1})	55	55	55	38	55	55	52	7
% Lorentzian	0	0	79	85	54	0	36	41
Center (cm^{-1})	1294	1300	1302	1299	1303	1299	1299	3
Width (cm^{-1})	29	29	29	29	29	29	29	0
% Lorentzian	0	0	0	0	0	69	11	28
Center (cm^{-1})	1328	1330	1335	1333	1336	1334	1333	3
Width (cm^{-1})	55	38	38	38	38	38	41	7
% Lorentzian	0	0	100	100	90	100	65	50
Amide S								
Center (cm^{-1})	1397	1397	1392	1385	1384	--[a]	1391	6
Width (cm^{-1})	36	38	38	38	38	--[a]	38	1
% Lorentzian	0	36	100	100	100	--[a]	67	47
Amide II								
Center (cm^{-1})	1512	1518	1519	1515	1521	1515	1517	3
Width (cm^{-1})	37	23	31	26	43	38	33	8
% Lorentzian	100	0	28	0	0	0	21	40
Center (cm^{-1})	1551	1554	1555	1552	1552	1554	1553	2
Width (cm^{-1})	26	32	27	29	32	33	30	3
% Lorentzian	100	100	100	100	100	100	100	0
Amide I								
Center (cm^{-1})	1650	1655	1656	1653	1653	1654	1653	2
Width (cm^{-1})	49	45	48	38	34	36	42	6
% Lorentzian	82	41	13	0	9	0	24	32
Center (cm^{-1})	1683	1685	1697	1684	1682	1686	1686	5
Width (cm^{-1})	29	37	38	33	29	25	31	5
% Lorentzian	47	50	0	0	0	0	16	25

[a] The intensity of the amide S band was too low to fit a band in this region in the 210 nm excited UVRR spectrum of myoglobin.

position of the non-helically associated amide S band varied by more than 10 cm^{-1}

($\lambda_{193\ nm}$: 1397 cm^{-1}, $\lambda_{206\ nm}$: 1384 cm^{-1}). Likewise, the estimated position of the

second component in the amide I region varied by 15 cm^{-1} ($\lambda_{200\ nm}$: 1697 cm^{-1}, $\lambda_{206\ nm}$: 1682 cm^{-1}). These values are in stark contrast to the standard deviation of the predicted peak position, width and shape, for the relatively intense α-helical derived amide II band which were 2, 3 and 0 cm^{-1}, respectively.

The arbitrary influence of noise and instrumental variations on the resolved spectral features of minor secondary structural components highlights one of the primary limitations of the application of NLLS fitting of vibrational spectra, which is that each spectrum is being analyzed in isolation. An alternate fitting algorithm, MCR-ALS, is ideal as chemically relevant constraints are easily incorporated in to the ALS minimization and a series of spectra may be analyzed as a whole. However, reasonable estimates of the spectral or excitation profiles are required to initialize the MCR-ALS minimization. In this work, these initial profiles were derived from the peak parameters obtained from the NLLS fit of the UVRR spectrum of myoglobin with the lowest excitation wavelength (λ_{ex} = 193 nm).

3.4.2 Effect of constraint level on MCR-ALS model

As a proof of principle, the MCR-ALS method described in the materials and methods section was first applied to only the amide I (1640-1730 cm^{-1}) region of the preprocessed data (Figure 3-3) in an attempt to resolve what can be thought of as pure amide I features associated with the α-helical and non-helical (random coil + β-turn) secondary structures. In order to robustly evaluate the effectiveness of chemically relevant constraints, three models, an unconstrained, a fully constrained, and a partially constrained variant, were applied to the amide I region (Figure 3-3a-c). Both the partially constrained and fully constrained models incorporated the

48

aforementioned Gaussian/Lorentzian peak shape constraint in the ALS optimization.

Each spectral component, as they are optimized iteratively, represents the best global

fit of the data set under a specified set of constraints. All models were evaluated

based on the *SSR* values and the nearness to the assumption that the amide I region

should have two discrete, unimodal, amide I components corresponding to the α-

helical and non-helical portions of the protein's structure[3, 18]. The results are

summarized in Table 3-2.

Table 3-2. Effect of constraint level on the fitted peak parameters and *SSR* for the amide
I region.

Constraint Level	Height	Center (cm^{-1})	Width (cm^{-1})	% Lorentzian	*SSR*
Unconstrained[a]	4.63×10^4	1652	39	19	$1.97 \times 10^{8,b}$
	2.35×10^4	1685	36	100	
Fully constrained	4.60×10^4	1652	37	0	5.04×10^8
	2.63×10^4	1684	35	100	
Partially constrained	3.66×10^4	1652	37	0	1.95×10^8
	2.10×10^4	1683	44	73	

[a] Peak parameters for the unconstrained model were obtained by fitting the resolved unconstrained
components to single Gaussian/Lorentzian peak profiles using NLLS.

[b] The *SSR* based on the unconstrained components.

3.4.2.1 Unconstrained Model

The unconstrained model produced an excellent fit of the amide I region

(Figure 3-3a). However, the second component was clearly not unimodal, with a

minor maxima at the edge of the spectral region (~1640 cm^{-1}) and another larger

maxima at ~1685 cm^{-1}. The maxima at ~1640 cm^{-1} is likely due to a small

contribution from the neighboring aromatic bands. Resonance enhancement of

tyrosine and phenlalanine is quite strong in the deep-UV[36, 42]. Attempts in previous

studies have been made to subtract these contributions; however, subtraction does not completely remove the aromatic bands from the UVRR spectrum. [3, 43]

3.4.2.2 Fully Constrained Model

The same initialization profiles were used for the fully constrained model as used previously for the unconstrained model, with the exception that each spectral profile was constrained to have a Gaussian/Lorentzian profile. Comparing the fully constrained model (Figure 3-3b) to the unconstrained model in Figure 3-3a, it is clear that the fully constrained model results in a poorer fit of the amide I region. Furthermore, the fully constrained model resulted in the highest *SSR* value (Table 3-2) as compared to the unconstrained and partially constrained models. This indicates that the fully constrained model is too rigid and some contribution from the neighboring aromatic modes must be incorporated in any model of this region.

3.4.2.3 Partially Constrained Model

In order to account for non-random, non-unimodal spectral contributions, an unconstrained component was introduced. The partially constrained model is generated using the same initialization parameters as the fully constrained model, with the first two amide I components subjected to a Gaussian/Lorentzian shape constraint. However, to account for the non-random putatively aromatic mode contribution, a third component was introduced lacking any type of shape contraint. This unconstrained component was generated by subtacting the fitted amide I profiles from the amide I region of the UVRR spectrum of myoglobin (λ_{ex} = 193 nm). The residuals were then used to initialize the unconstrained component. This unconstrained component (Figure 3-3c) easily accounts for the influence of the

neighboring aromatic modes as well as an additional small previously unmodelled contribution in the 1680-1690 cm^{-1} region of the spectra.

MCR-ALS analysis of the amide I region from all six excitation wavelengths yielded two amide I components, 1652 and 1683 cm^{-1}, (Table 3-2, Figure 3-3c). Interestingly, the mean values of 1653 (± 2) and 1686 (± 5) cm^{-1} (Table 3-1) from the

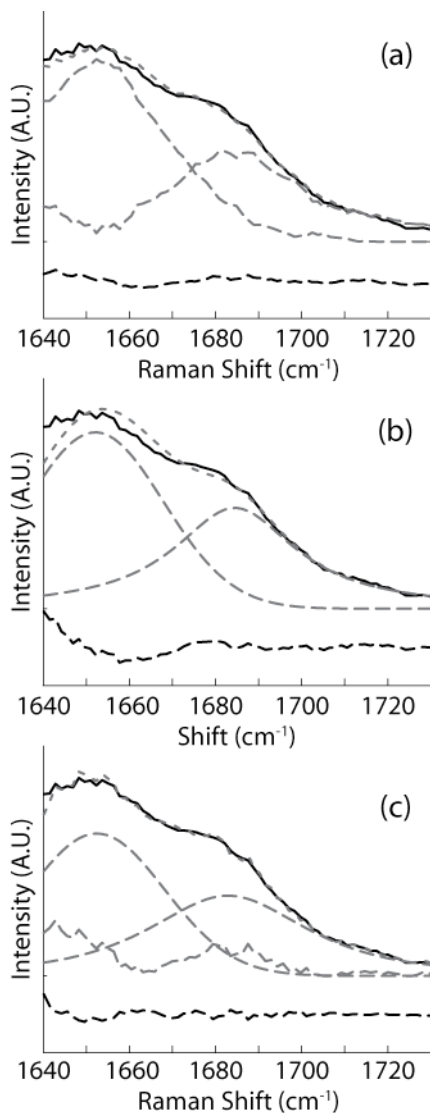**Fig. 3-3** Spectral reconstruction of the amide I region (λ_{ex} = 193 nm) using (a) an unconstrained, (b) fully constrained and (c) partially constrained model. The original spectrum (——) and residuals (– –) at 193 nm excitation are shown in black. The fitted components (– –) and predicted spectrum (- - -) are shown in gray.

NLLS analyses are quite close to predicted peak positions from MCR-ALS analysis.

3.4.3 MCR-ALS with NLLS optimization of Gaussian/Lorentzian shape constraints

The amide II, III and S regions were also analyzed using a partially

constrained model. Each component was classified as either α-helical or non-helical

according to previous studies[7, 14, 44] (Table 3-3). Two amide bands were resolved in

the amide II region, occurring at 1520 and 1550 cm^{-1}, respectively (Figure 3-4). These

values are similar to those obtained by Huang *et al.* (1525 and 1552 cm^{-1}) for α-

helices[3]. A unique non-helical component was not resolvable in the amide II region.

This is likely due to the direct overlap of the amide II bands from α-helices and non-

helical coil structures, which are both predicted to occur at 1552 cm^{-1} [3]. As with the

amide I region, the predicted peak positions from MCR-ALS compare well with the

mean values from the NLLS analyses, which are 1517 ($\pm$ 3) and 1553 ($\pm$ 2) cm^{-1}

(Table 3-1).

Table 3-3. Results from MCR-ALS fitting of the amide regions using a partially constrained model

	Center (cm^{-1})	Width (cm^{-1})	% Lorentzian	Classification
Amide III	1254	32	0	coil/turn
	1298	54	0	α-helix
	1345	40	0	α-helix
Amide S	1397	44	74	coil/turn
Amide II	1520	40	0	α-helix
	1550	38	100	α-helix, coil/turn
Amide I	1652	37	0	α-helix
	1682	44	72	coil/turn

The amide III region warrants special attention, as it is by far the most

spectrally congested, with α-helical, β-turn and disordered structures present in the

protein, this region could have up to seven different components[2, 3, 18]. It has been

shown that the amide III position is dependent upon the ψ dihedral angle[45]; Huang *et*

al. resolved amide III bands for β-turn and coil structures at 1244 and 1253 cm[-1],

respectively[3]. Thus, this region was modeled with both three and four constrained

components plus one unconstrained component. Models with >4 components

generally failed to converge to reasonable solutions indicating that there was

insufficient resolution in this data to resolve discrete β-turn and unordered amide III

profiles. It is likely that the exctiation profiles of the β-turn and unordered structures

are too overlapped to be resolved, given the limited number of excitation wavelengths

in this study.

     The four-component model (three constrained, one unconstrained) yielded

three peaks at 1254, 1298 and 1345 cm[-1] (Figure 3-4). The peaks at 1298 and 1345

cm[-1] have been assigned previously to α-helical structure[3, 18], while the contribution at

1254 cm[-1] is consistent with published unordered structure contribution assignments [3].

It has been proposed that α-helical structures have an additional lower frequency

amide III band at ~1263 cm[-14, 26]. Indeed, the main α-helical band of poly(L-lysine) at

1293 cm[-1] has a shoulder at 1276 cm[-1]. Most likely, this band is not resolvable from

the main band at 1298 cm[-1] because their excitation profiles are too similar to be

uniquely resolved.   Unlike the amide I and II regions, the predicted peak positions

are slightly shifted from the mean values of the NLLS analyses (Table 3-1). For

example, the mean value of the highest frequency component from NLLS analyses is

1333 (± 3) cm[-1], 12 cm[-1] downshifted from the position predicted by MCR-ALS.

However, previous studies report that the higher frequency amide III component
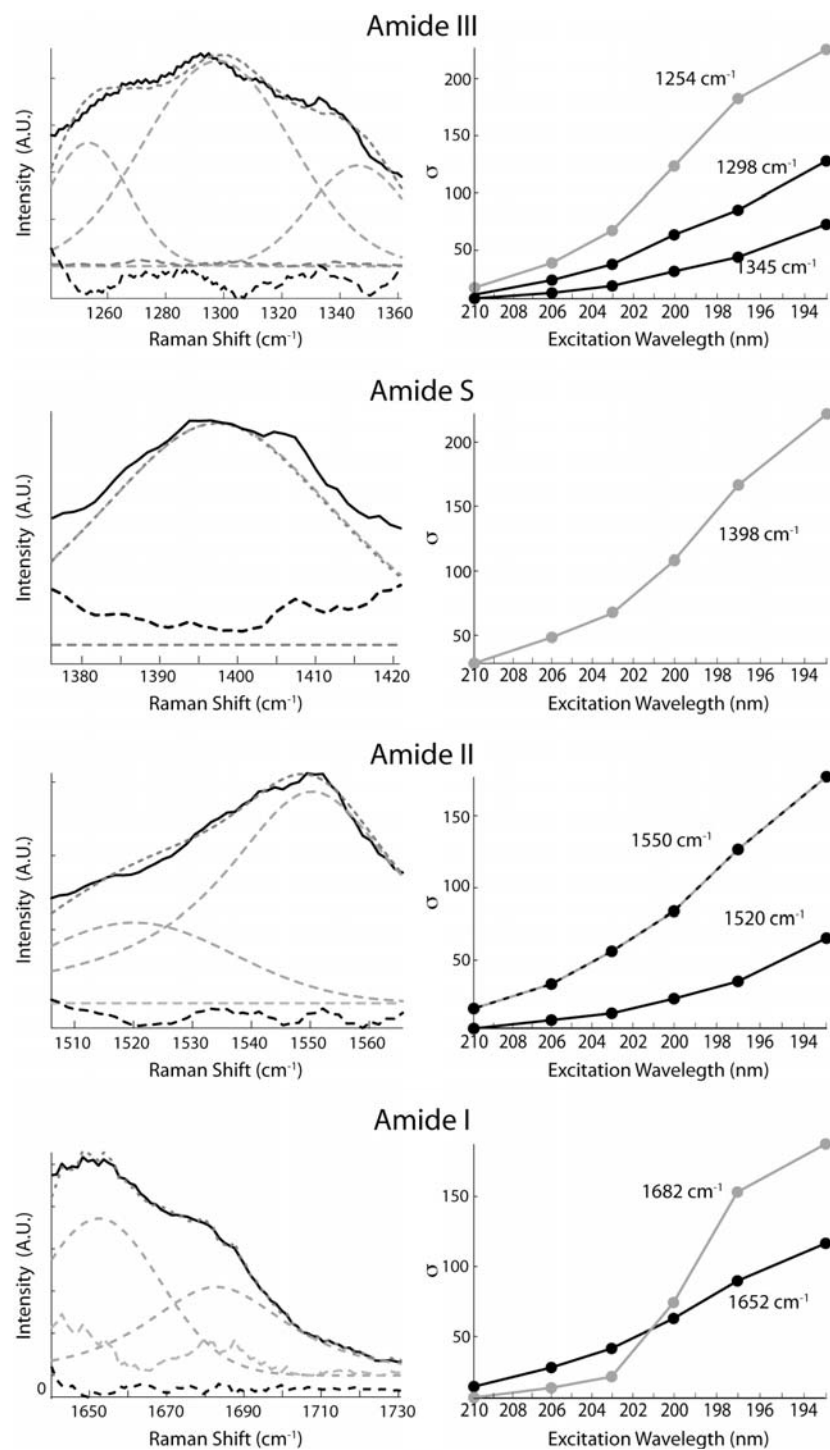
Fig. 3-4 *Column 1*: Spectral reconstruction of the amide I, II, III and S regions from the sum (- - -) of the individual components (– –) from the partially constrained model, multiplied by the cross section at 193 nm excitation. The original spectrum (—) and residuals (▬ ▬) are shown in black. *Column 2*: Estimated cross section for the α-helical (*black*) and non-helical (*gray*) portions of the protein. The component at 1550 cm⁻¹, which is thought to have contributions from both the α-helical and non-helical structures, is shown in both gray and black. Cross sections (σ) are expressed in units of millibarns/(molecule·steradian).

54

occurs at ~1340 cm^{-1} [3, 18, 44], closer to the value predicted by the MCR-ALS model.

As the amide S band is only resonance enhanced in non-helical structures[14, 46], only one amide S component was expected in this region. Therefore, the amide S region was modeled using one constrained and one unconstrained component (Figure 3-4). The constrained component was centered at 1397 cm^{-1}, which is 6 cm^{-1} higher than the mean of the NLLS analyses (Table 3-1, Figure 3-4). It should be noted that the contribution from non-helical conformations is stronger at lower excitation wavelengths, and at these wavelengths, NLLS does predict an amide S band at 1397 cm^{-1} (Table 3-1). These results suggest that simultaneous analysis of bilinear UVRR spectra using MCR-ALS with shape constraints provides a distinct advantage over sequential analysis by NLLS.

3.4.4 Pure α-helical and non-helical UVRR spectral profiles

An advantage of bilinear methods is the ability to produce estimates of the resolved pure component spectral profiles. Estimates of the pure spectral profiles for the α-helical and non-helical portions of myoglobin at each excitation wavelength are shown in Figure 3-5. The amide II band at 1550 cm^{-1} has contributions from both α-helical and non-helical structures. Therefore, the fractional intensity of the α-helix and non-helical 1550 cm^{-1} bands were estimated relative to the amide I excitation profile. Interestingly, at 193 nm, 64% of the 1550 cm^{-1} band intensity was estimated to arise from α-helical structure and 37% from non-helical structures. At 210 nm, 86% of the intensity was estimated to arise from α-helical structure and only 18% from the non-helical structures, illustrating the greater selectivity for α-helical spectral signatures at longer wavelengths and non-helical structures at shorter

Fig 3-5 Estimated pure UVRR spectral profiles of the (a) α-helical and (b) non-helical portions of myoglobin at each excitation wavelength.

wavelengths.

The estimated α-helical profiles are qualitatively similar to previous estimates of the pure spectral profiles from multi-protein studies[2, 3]. In addition, the estimated α-helical profiles are strikingly similar to the UVRR spectrum of poly(L-lysine) in its α-helical conformation at high pH and in 40% trifluoroethanol (TFE), an α-helix promoting solvent that also dehydrates the protein[18]. Likewise, the non-helical pure spectral profiles are qualitatively similar to the unordered UVRR profiles resolved

previously[2, 3], although, the non-helical amide I band is shifted about 20 cm^{-1} higher than the amide I band associated with unordered structures[2, 3].

3.4.5 Estimation of α-helical and non-helical cross sections

The resolved excitation profiles were converted to cross sections using Equation **8**. The calculated α-helical and non-helical cross sections (Figure 3-4, column 2) both increase with decreasing excitation wavelength, consitent with previous observations[25, 26]. The non-helical cross sections exhibited a steep increase from 203 to 197 nm, similar to the estimated cross sections[25] and absorption profile[24] of poly(L-lysine) in the random coil conformation. The α-helical cross sections increased more gradually from 210 to 193 nm, similar to the estimated cross sections[26] of a 21-residue, primarily alanine, peptide and the absorption profile of poly(L-lysine)[24] in their α-helical conformations. However, the α-helical corss sections calculated in this study lacked a shoulder around 205 nm, which was observed by Sharma *et al.*[26]. This is probably due to the limited number of excitation wavelengths. In addition, at 192 nm, the molar exctinction coefficient of α-helices is approximately half that of unordered or β-sheet structures [24]. This is reflected in the calculated cross sections at the lowest excitation wavlength (193 nm) of the α-helical bands at 1298 and 1652 cm^{-1} with respect to the non-helical bands at 1254 and 1682 cm^{-1}.

3.5. Conclusions

Through the use of a partially constrained MCR-ALS algorithm, we have demonstrated the successful deconvolution of the amide I, II, III and S regions of the 2-dimensional multi-excitation UVRR spectrum of myoglobin without prior

subtraction of the aromatic bands. In effect, we were able to identify "pure" excitation wavelength dependent α–helical and non-helical derived UVRR spectral features for the first time an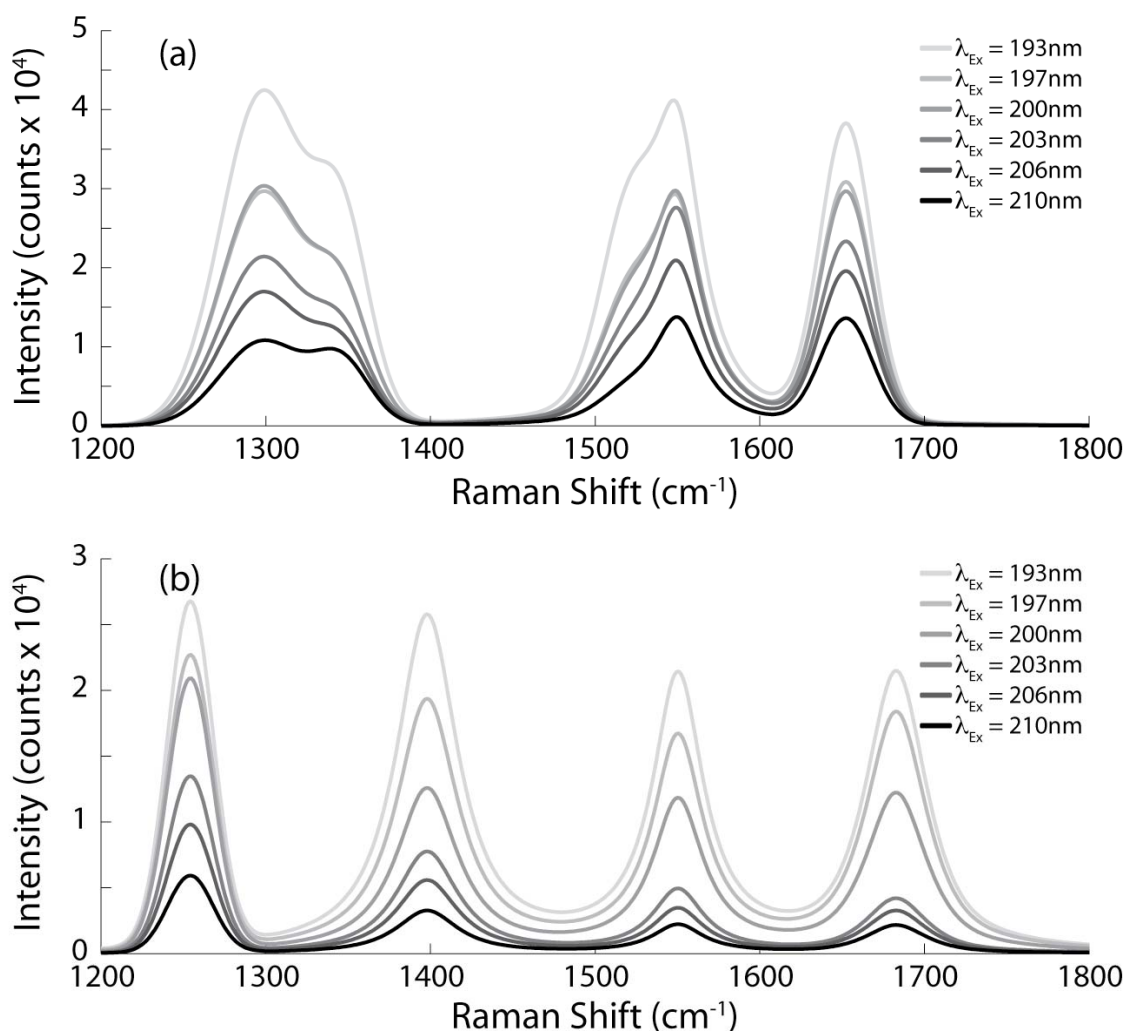d estimate the pure UVRR spectral profiles of the α-helical and non-helical portions of myoglobin at each excitation wavelength (Figure 3-5). Additionally, we managed to remove the user defined and instrumentation based biases that occur with traditional spectral fitting methods. This represents a significant advancement from standard NNLS fitting methods, which commonly require prior knowledge of peak position, shape etc… as well as rigid constraints to achieve reproducible results. The NLLS implementation of chemically relevant Gaussian/Lorentzian shape constraints merges the veteran NLLS fitting methods with the advantages of multivariate analysis. With the exception of the amide II band at 1550 cm^{-1}, each amide component was assigned to either α-helical or non-helical secondary structures. Last, this study highlights the potentially powerful tool of incorporating multiple excitation wavelengths in resonance Raman spectroscopy to aid in the resolution of highly overlapped spectral components, with minimal user input.

3.6. Acknowledgements

3.7. References

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

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

3. C. Y. Huang, G. Balakrishnan and T. G. Spiro, *J. Raman Spectrosc.*, 2006, **37**, 277- 282.

4. A. V. Mikhonin, S. V. Bykov, N. S. Myshakina and S. A. Asher, *J. Phys. Chem. B*, 2006, **110**, 1928-1943.

5. J. C. Austin, T. Jordan and T. G. Spiro, in *Biomolecular Spectroscopy*, eds. R. J. H. Clark and R. E. Hester, Wiley & Sons Ltd., New York, 1993, vol. 20, pp. 55-127.

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

7. S. Krimm and J. Bandekar, in *Advances in Protein Chemistry*, eds. C. B. Anfinson, J. T. Edsall and F. M. Richards, Academic Press, New York, 1986, vol. 38, pp. 181- 365.

8. A. T. Tu, in *Spectroscopy of Biological Systems*, eds. R. J. H. Clark and R. E. Hester, Wiley & Sons Ltd., New York, 1986, vol. 13, pp. 47-111.

9. G. Balakrishnan, C. L. Weeks, M. Ibrahim, A. V. Soldatova and T. G. Spiro, *Curr. Opin. Struct. Biol.*, 2008, **18**, 623-629.

10. R. A. Copeland and T. G. Spiro, *Biochemistry*, 1985, **24**, 4960-4968.

11. J. S. W. Holtz, J. H. Holtz, Z. H. Chi and S. A. Asher, *Biophys. J.*, 1999, **76**, 3227- 3234.

12. T. Miura, A. Hori-i, H. Mototani and H. Takeuchi, *Biochemistry*, 1999, **38**, 11560- 11569.

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

14. Y. Wang, R. Purrello, T. Jordan and T. G. Spiro, *J. Am. Chem. Soc.*, 1991, **113**, 6359- 6368.

15. G. Balakrishnan, Y. Hu, G. M. Bender, Z. Getahun, W. F. DeGrado and T. G. Spiro, *J. Am. Chem. Soc.*, 2007, **129**, 12801-12808.

16. I. K. Lednev, A. S. Karnoup, M. C. Sparrow and S. A. Asher, *J. Am. Chem. Soc.*, 1999, **121**, 8074-8086.

17. I. K. Lednev, A. S. Karnoup, M. C. Sparrow and S. A. Asher, *J. Am. Chem. Soc.*, 2001, **123**, 2388-2392.

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

19. C. Y. Huang, G. Balakrishnan and T. G. Spiro, *Biochemistry*, 2005, **44**, 15734-15742.

20. N. Haruta and T. Kitagawa, *Biochemistry*, 2002, **41**, 6595-6604.

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

22. V. A. Shashilov, V. V. Ermolenkov and I. K. Lednev, *Inorg. Chem.*, 2006, **45**, 3606-3612.

23. V. A. Shashilov, M. Xu, V. V. Ermolenkov and I. K. Lednev, *J. Quant. Spectrosc. Radiat. Transfer*, 2006, **102**, 46-61.

24. K. Rosenheck and P. Doty, *Proc. Nat. Acad. Sci. U.S.A.* , 1961, **47**, 1775-1785.

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

26. B. Sharma, S. V. Bykov and S. A. Asher, *J. Phys. Chem. B*, 2008, **112**, 11762-11769.

27. R. Tauler, *Anal. Chim. Acta*, 2007, **595**, 289-298.

28. R. Gargallo, R. Tauler and A. Izquierdo-Ridorsa, *Anal. Chem.*, 1997, **69**, 1785-1792.

29. R. Gargallo, M. Vives, R. Tauler and R. Eritja, *Biophys. J.*, 2001, **81**, 2886-2896.

30. M. Vives, R. Gargallo and R. Tauler, *Anal. Biochem.*, 2001, **291**, 1-10.

31. J.-H. Wang, P. K. Hopke, T. M. Hancewicz and S. L. Zhang, *Anal. Chim. Acta*, 2003, **476**, 93-109.

32. C. A. Holden, S. S. Hunnicutt, R. S·nchez-Ponce, J. M. Craig and S. C. Rutan, *Appl. Spectrosc.*, 2003, **57**, 483-490.

33. R. Tauler, A. K. Smilde, J. M. Henshaw, L. W. Burgess and B. R. Kowalski, *Anal. Chem.*, 1994, **66**, 3337-3344.

34. S. Navea, A. de Juan and R. Tauler, *Anal. Chem.*, 2002, **74**, 6031-6039.

35. S. Nigam, A. de Juan, R. J. Stubbs and S. C. Rutan, *Anal. Chem.*, 2000, **72**, 1956-1963.

36. G. Balakrishnan, Y. Hu, S. B. Nielsen and T. G. Spiro, *Appl. Spectrosc.*, 2005, **59**, 776-781.

37. R. Bro and N. D. Sidiropoulos, *J. Chemom.*, 1998, **12**, 223-247.

38. H. A. L. Kiers, *J. Chemom.*, 2000, **14**, 105-122.

39. R. Tauler, B. Kowalski and S. Fleming, *Anal. Chem.*, 1993, **65**, 2040-2047.

40. E. Bezemer and S. C. Rutan, *Chemom. Intell. Lab. Syst.*, 2006, **81**, 82-93.

41. S. P. A. Fodor, R. A. Copeland, C. A. Grygon and T. G. Spiro, *J. Am. Chem. Soc.*, 1989, **111**, 5509-5518.

42. X. Zhao, R. Chen, C. Tengroth and T. G. Spiro, *Appl. Spectrosc.*, 1999, **53**, 1200-1205.

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

44. S. H. Song and S. A. Asher, *J. Am. Chem. Soc.*, 1989, **111**, 4295-4305.

45. S. A. Asher, A. Ianoul, G. Mix, M. N. Boyden, A. Karnoup, M. Diem and R. Schweitzer-Stenner, *J. Am. Chem. Soc.*, 2001, **123**, 11775-11781.

46. T. Jordan and T. G. Spiro, *J. Raman Spectrosc.*, 1994, **25**, 537-543.

Chapter 4: Robust pre-processing of ultraviolet resonance Raman spectra

The application of multivariate data analysis methods to UVRR spectra is of critical importance to the determination of protein secondary structure from UVRR spectra. However, difficulties in the application of such methods are encountered due to spectral shifts arising from a combination of factors beyond the control of the researcher. In addition, increasing spectral resolution with increasing excitation wavelength can make the alignment of multi-excitation data sets problematic. Here we present the use of correlation optimized warping (COW) as a possible pre-processing method to alleviate these alignment errors. In addition, the subtraction of the water band can also be problematic for multivariate data analysis as the water band overlaps the amide I mode. As demonstrated, subtraction of the water band according to the intensity of the internal intensity standard can result in an over subtraction of the water band. The over subtraction of the water band yields negative intensity in the amide I region. Presented here is method in which the concentration of the water band is determined through a least square optimization of the 1750-1900 cm^{-1} range which is free of influence from the amide I region.

4.1. Introduction

Multivariate data analysis methods have been applied to a wide range of chemical data. A requirement of the application of multivariate methods is that all response vectors

(chromatograms, spectra, etc..) have the same length. Additionally, optimal performance of multivariate methods can only be achieved when the response vectors are well aligned. Unfortunately, due to factors often beyond the control of the researcher analytical data often does not meet these most basic requirements. For example, chromatographic techniques often suffer from drifts in retention times due to subtle experiment variations, i.e. minor changes the mobile phase composition and changes in the performance of the column.[1,2] Spectroscopic methods are also subject to these issues. Near infrared spectroscopy (NIR) is greatly influenced by temperature effects, resulting in major shifts peak maxima.[3] UVRR is no exception to these experimental difficulties. In addition to spectral shifts arising from experimental variations due to day to day re-optimization of the focusing and collection optics , the resolution of multi-excitation data increases as the excitation wavelength increases.[4] The increase in resolution results in an increase in the length of the response vector for the same spectral region, a clear violation of the most basic requirement of multivariate methods.

UVRR spectrometers utilized for protein spectroscopy are all dispersive instruments due to the necessity to operate at excitation wavelengths from 193-230 nm, far below 850 nm where the multiplex disadvantage[5] makes the use of Fourier transform detectors impractical. The collected Raman spectra on dispersive spectrometers must be calibrated, usually with a reference spectra consisting of 5 to 6 well characterized peaks.[6] Calibration must be performed for each collection session due to the possibility of thermal effects or mechanical drift of the grating, in addition to changes in the optical pathway and focal point of sample irradiation, leading to inaccuracies in the calculated Raman shifts. A calibration vector is calculated using a second or third order polynomial,

due to the nonlinear nature of Raman shift to position on a charge coupled device (CCD) detector.[7]

A variety of approaches have been proposed to correct shifts in both spectroscopic and chromatographic methods, including multivariate standardization[8,9] and warping algorithms such as dynamic time warping (DTW)[10-13] and correlation optimized warping (COW)[10-13]. These warping algorithms are of particular interest, having already been applied to chromatographic data[11,12], NMR[10], two dimensional gas chromatography coupled to time of flight mass spectroscopy[13], and NIR spectra[11] to resolve shifts in the time, wavelength, and frequency dimensions of the measured data. In this work COW has been applied to UVRR spectra to alleviate day to day spectral shifts in the data and to align multi-excitation spectra.

Water, although a weak Raman scatterer, gives rise to a major multi-component Raman signal arising from the –OH bending motion.[14] This water signal heavily overlaps the amide I, amide II, and aromatic signals, complicating the analysis of UVRR spectra. Due to the difficulty of determining direct Raman intensities, an internal intensity standard is frequently utilized to determine the concentration of the water band in a sample spectrum.[15] The relative intensity determined through a internal standard helps to account for variations in laser power and detector efficiency. However, the use of the internal intensity standard to determine the magnitude of the of the water band subtraction often results in a over subtraction, leading to negative intensity in the amide I region of the spectrum. Here we present a method for determine the concentration of the water band without the use of an internal intensity standard. Instead, a least squares

optimization is used to determine the concentration of the water band from 1750-1900 cm^{-1} region of the water band and sample spectra.

4.2. Theory

4.2.1. Correlation optimized warping

Correlation optimized warping (COW)[16] is a preprocessing method which has been used to correct subtle shifts in peak position by segmented expansion or compression (warping) of chromatograms or spectra by linear interpolation to a reference chromatogram or spectrum. For any two given spectra, one is chosen as the sample, **P**, of length L_P, and the other is designated as the reference spectra, **T**, of length L_T. The sample spectrum, **P**, is divided into segments of length m, resulting in N sample segments (Equation 1). The reference spectrum, **T**, is then divided into N segments, the length of which may be less than, greater than, or equal to m.

$$N = \frac{L_P}{m} \qquad (1)$$

Each sample segment may be warped by all integer values between 0 and t, called "the slack" which is a user designated parameter, limiting the maximum warping for each segment. A defined maximum for the compression or expansion results in a finite number of possible warpings for each segment. In cases where the lengths of the sample and reference spectra vary greatly, such as multi-excitation UVRR spectra[4], an expanded, but finite range of warping values for each segment is allowed. This new warping range is Δ-t to Δ+t, where Δ is defined by Equation 2.

$$\Delta = \frac{L_T}{N} - m \qquad (2)$$

Each segment of the sample spectrum is warped according to the corresponding segment in the reference spectrum. For the sample segment, from **P**, every possible result of the warping is determined, **P'**, with the addition or subtraction of every integer value of points from 0 to t to the segment. In order to select the optimum warping for each segment, **w**, the correlation coefficient, ρ, for each possible warping is calculated according to equation 3:

$$\rho = \frac{(\mathbf{w}_T - \overline{w}_T)^T (\mathbf{w}_{P'} - \overline{w}_{P'})}{\sigma_{w_T} \sigma_{P'}} \qquad (3)$$

The correlation coefficient in Equation 3 is calculated from the segment of the sample ($\mathbf{w}_{P'}$) and corresponding reference spectra ($\mathbf{w}_T$), mean of the segment of both the reference ($\overline{w}_T$) and sample spectra ($\overline{w}_{P'}$) and the standard deviations of the reference (σ_{w_T}) and sample ($\sigma_{P'}$) segments. The warped segment ($\mathbf{w}_{P'}$) with the highest correlation coefficient is selected as the optimum warped segment.

The warping of each segment according to the slack parameter results in a finite number of solutions, for a finite number of segments. The determination of the optimal solution of the warping of the sample to reference spectra may be performed by dynamic programming [17], a method which may solve a properly constrained form of the combinatorial optimization problem. Each segment of the spectrum is defined by end points, known as nodes. Two matrices, **F** and **U** of dimensions $((N+1)\times(L_T+1))$, a row for each node and a column for each possible node position, are used to record the optimum warping pathway.

The optimization starts with the final segment. The correlation coefficient is calculated for all possible warpings of the target segment, the highest correlation

coefficient result is stored in the elements of matrix **F**. At the same time, the matrix **U** serves to store the warped segment with the highest correlation coefficient. This process continues to the first segment of the data, the first row of the matrixes **F** and **U**. This allows for the reconstruction of the optimum warping of the data from the elements stored in matrix **U**. The end result of this segmented warping procedure is that the endpoints of the spectra are essentially fixed, however, the total warping of the spectra builds as the algorithm proceeds through each segment, allowing for significant warping toward the middle of the spectra while keeping the ends fixed.

4.2.2. Water band subtraction

Scaling the water band subtraction according to the internal intensity standard often results in over subtraction of the water band. Over subtraction of the water band frequently results in negative intensity in the amide I region. In order to determine the relative contribution of the water band in the protein spectra, the relative intensity of the water band in the 1750-1900 cm^{-1}, above the maximum Raman shift range of the amide I region, 1680 cm^{-1}, was estimated using Equation 4. The relative intensity of the water band in the sample spectra, c, was calculated by multiplying the selected region of the sample spectrum, **s**, by the pseudo inverse of the same region of the water or buffer spectrum, **w**. The calculated intensity (c) was then multiplied by the entire water or buffer spectra and directly subtracted from the sample spectrum.

$$c = [\mathbf{s}]^{\mathrm{T}}[\mathbf{w}^{+}] \ (4)$$

The c-terminal carboxyl group and the carboxyl group of the amino acid glutamic acid may contribute in this region. However, contributions from carboxyl groups are relatively weak in UVRR spectra and are typically not observed in globular proteins.

4.3. Experimental

4.3.1. Sample Preparation

Mono- and dibasic potassium phosphate, sodium perchlorate, L-tyrosine, and ovalbumin were purchased from Sigma-Aldrich (St. Louis, MO) and used without further purification.

Phosphate buffers were prepared by dissolving appropriate amounts of mono- and dibasic potassium phosphate in 18.2 MΩ RO (reverse osmosis) water for a final concentration of 10 mM. Buffers were then pH adjusted to 7.4 with small additions sodium hydroxide or hydrochloric acid. The internal intensity standard, sodium perchlorate was added to all buffers at a final concentration of 0.2 M, unless otherwise stated. All buffers were filter sterilized and sparged with argon for 20 minutes to remove oxygen.

A concentrated stock solution of tyrosine was prepared by dissolving L-tyrosine in 10 mM phosphate buffer with 0.2 M sodium perchlorate. The absorption spectrum of the stock solution was collected on a Hewlett Packard 8453 UV-Vis Spectrometer (Palo Alto, CA). The concentration of L-tyrosine was then estimated from the molar extinction coefficient at 280 nm, which is 1280 M^{-1}cm^{-1}.[18] Sample solutions at 200, 100, and 20 μM L-tyrosine were then prepared by dilution of an appropriate volume of the stock solution in buffer.

Concentrated stock solutions of ovalbumin were prepared by dissolving the protein in 10 mM phosphate buffer with 0.2 M sodium perchlorate. A small volume of the stock solution was diluted 10 fold into buffer. The absorption spectrum of the resulting solution was collected on a Hewlett Packard 8453 UV-Vis Spectrometer (Palo

Alto, CA). The concentration of ovalbumin in the stock solution was then estimated from the calculated concentration of the dilute ovalbumin solution. The molar extinction coefficient of ovalbumin at 280 nm is 32100 $M^{-1}cm^{-1}$[19]. The absorption of the dilute ovalbumin solution at 280 nm was used to calculate the concentration. Sample ovalbumin solutions at a concentration of 0.5 mg/ml were then prepared by dilution of the stock solution in to 10 mM phosphate buffer with 0.2 M perchlorate.

Poly-L-lysine hydrobromide (1000-4000 Da) was purchased from Sigma-Aldrich (St. Louis, MO) and desalted to remove bromide from the protein solution using a Bio-scale Mini P6 cartridge (Bio-Rad, Hercules, CA) as bromide absorbs the incident excitation and scattered photons greatly reducing the signal intensity. Approximately 0.5 mg of poly-L-lysine hydrobromide was dissolved in 18.2 MΩ RO water and desalted. The resulting solution was then lyophilized and re-suspended in 5 mL of 10 mM phosphate buffer with 0.2 M perchlorate at pH 4.0.

4.3.2. UVRR spectroscopy

All UVRR spectra were collected on a custom built UVRR spectrometer of analogous design to those already presented in the literature.[20,21] A tunable frequency-quadrupled titanium-sapphire laser (Coherent Inc., Santa Clara, CA), pumped by the second harmonic of a diode pumped Nd:YLF laser, was used as the excitation source. The sample was circulated through a temperature-controlled sample chamber and water-jacketed reservoir, maintained at ~7°C, of an in-house design and manufactured by Mid Rivers Glassblowing (Saint Charles, MO), using a model 75211-10 gear pump (Cole Parmer, Vernon Hills, Illinois). A thin film of the sample was created by flowing the solution through a 19 gauge needle and between two thin nitinol wires (0.005 in diameter)

(Small Parts, Inc. Miramar FL). The sample film was directly irradiated by the incident excitation beam. Unless otherwise specified the excitation wavelength was 200 nm.

The Raman scattering was collected in the 135˚ backscattering geometry and dispersed using a 1.25 m spectrometer (Horiba Jobin Yvon Inc., Edison, NJ) equipped with a 3600 grove/mm grating. The spectrometer was equipped with a back illuminated, phosphor coated, liquid nitrogen cooled Symphony CCD camera (Horiba Jobin Yvon Inc., Edison, NJ) with a chip size of 2048 x 512. The laser power at the sample chamber was kept below 0.5 mW to avoid sample degradation.[22] Each spectrum was collected over 180 minutes producing 72 individual spectra which were then averaged. Spectra were collected and exported in CSV format using Synergy software (Horiba Jobin Yvon Inc., Edison NJ).

The spectrum of cyclohexane and the peak positions reported in Ferraro and Nakamoto [23] were used to calibrate the collected UVRR spectra. New calibration spectra were collected every collection session and used to calibrate all spectra collected in each session on each day. Separate calibration spectra were collected for each excitation wavelength used in the multi-excitation dataset and used to calibrate the buffer and sample spectra for each excitation wavelength.

4.3.3. Data Analysis:

Analysis of UVRR spectra was carried out in the Matlab environment, version 7.1 (Mathworks, Natick, MA). The correlation optimized warping algorithm was used without modification and is available for download at www.models.life.ku.dk.[16]

4.4. Results and Discussion

4.4.1. Water Band Subtraction:

Although water is a weak Raman scatter, water has two primary Raman bands at ~1640 cm^{-1} and ~ 3400 cm^{-1},[20] which are associated with the –OH bending and –OH stretching vibrations.[14,24] The lower energy energy –OH bending mode at ~1640 cm^{-1} directly overlaps the amide I mode of proteins, and to a smaller extent, the amide II mode, as well as, the aromatic modes in this region (Figure 4-1A). Subtraction of this water band is typically based on the peak height or area of the internal intensity standard included in both the buffer and sample solutions. For these studies, perchlorate was employed as the internal intensity standard as perchlorate (ClO$_4^-$) has a well-defined strong feature at 932 cm^{-1} that is well resolved from the amide and aromatic modes of proteins. Subtraction of the water band based on the intensity of the perchlorate band works well at higher sample concentrations but often results in over subtraction of the water band at lower sample concentrations.

The UVRR spectra of 20, 100, and 200 μM L-tyrosine solutions were collected (Figure 4-1). The major vibrational signals associated with tyrosine are easily observable, even in with the overlap of the water band (Figure 4-1A). Two prominent features in the spectra, the peaks at 1601 cm^{-1}and 1617 cm^{-1}, arise from the C-C stretching of the aromatic ring, referred to υ_{8b} and υ_{8a} respectively. At lower energies, the C-O stretching signal, $\upsilon_{7a'}$, occurring at 1263 cm^{-1}, overlaps with the C$_{ring}$-C$_{exterior}$ stretching motion located at 1210 cm^{-1} (υ_{7a}). The C-H bending signal, υ_{9a}, is located nearby at 1180 cm^{-1}.[25] The final Raman signal which results from tyrosine is located at 852 cm^{-1}, the product of the Fermi doublet which arises from the interaction of the ring breathing motion (υ_1)and
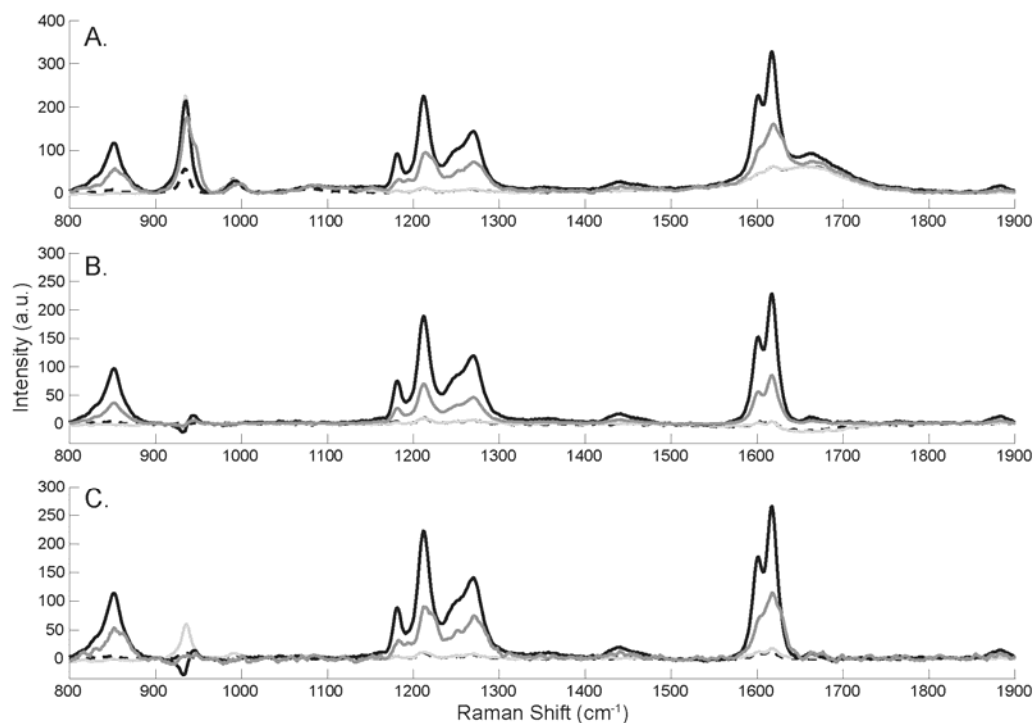
Fig. 4-1 (A) UVRR spectra of L-tyrosine, 20 µM (blue), 100 µM (red), 200 µM (green) in 10 mM phosphate buffer with an internal intensity standard, sodium perchlorate, at 0.2 M concentration, which can be observed due to the characteristic peak at 932 cm^{-1}. The UVRR spectrum of 20 µM with 0.05 mM sodium perchlorate (cyan) is also shown. (B) UVRR spectra of 20 µM (blue), 100 µM (red), and 200 µM (green) tyrosine (0.2 M perchlorate) with 20 µM tyrosine (0.05 M perchlorate)(cyan) after subtraction of the buffer spectrum based on the internal intensity standard. (C) UVRR spectra of 20 µM (blue), 100 µM (red), and 200 µM (green) tyrosine (0.2 M perchlorate) with 20 µM tyrosine (0.05 M perchlorate)(cyan) after subtraction of the buffer spectrum based on the tail end of the water band.

the out of plane ring deformation ($2\upsilon_{16a}$). The relative intensity of the water band in each

solution was then estimated from the peak area of the perchlorate band between 900 cm^{-1}

and 970 cm^{-1}. An artifact of the perchlorate subtraction, in the form of an inflection point,

can be observed due to slight differences in the shape and height of the perchlorate band

in the buffer spectrum compared to the perchlorate bands present in the L-tyrosine

spectra despite the normalization of the perchlorate peaks according to peak area (Figure

4-1B). The resultant UVRR spectra of 100 and 200 µM L-tyrosine resembles previously

reported spectra of L-tyrosine.[25] However a negative feature appears in the amide I region

of the UVRR spectra of 20 µM L-tyrosine with either 0.2 M or 0.05 M perchlorate

indicating over subtraction of the buffer spectrum (Figure 4-1B) Despite the difference in the concentration of perchlorate, subtraction of the water band based on the internal intensity standard produced similar spectra indicating that the concentration of the internal intensity standard had little or no impact on the intensity of the water band. Thus it can be concluded that the magnitude of the water band with respect to the perchlorate band is greater in the buffer spectrum than in the sample spectra. At higher sample concentrations this difference is likely to be negligible and only become problematic at lower sample concentrations when the intensity of the water band approaches that of the protein bands.

In order to address the issue of over subtraction of the water band at low analyte concentrations, a method was developed to estimate the contribution of the water band to the overall spectrum based on the intensity of the tail end of the water band from 1750-1900 cm^{-1}. Although the signal in this region is relatively low, the selectivity for the water band is quite high. It should be noted that contributions from carboxyl groups associated with glutamic acid or the *c*-terminus are located in this region. However, the CO_2^- mode is often to0 weak to be observed in the UVRR spectra of globular proteins. Therefore, for these studies it was assumed that the contribution from this mode was negligible in all spectra.

Subtraction of the buffer spectrum based on the intensity of the tail end of the water band resulted in similar spectra for 100 and 200 μM L-tyrosine (Figure 4-1C) when compared to subtraction based on the perchlorate intensity. Further, the 1601 and 1618 cm^{-1} bands of L-tyrosine are now readily visible in the UVRR spectra of 20 μM L-tyrosine.

70
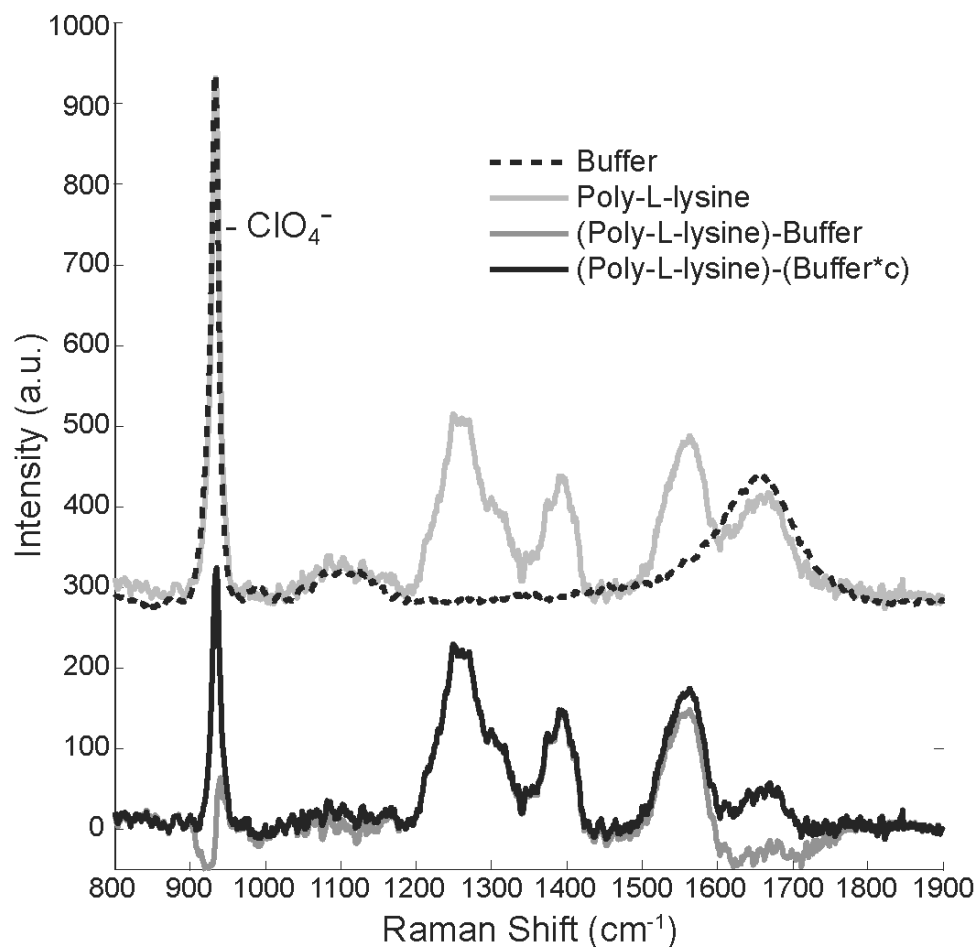
Fig. 4-2 UVRR spectra (λ_{ex} = 206 nm) of 10 mM phosphate buffer (blue), poly-L-lysine (1000-4000 Da) (green), and poly(L-lysine) after subtraction of the buffer spectrum based on the internal intensity standard, 0.2 M perchlorate, (red), and the tail end of the water band (1700-1900 cm^{-1}).

In order to determine whether this method would perform equally well on proteins, the UVRR spectrum of the homo-polypeptide poly-L-lysine was collected. Comparison of the buffer and poly-L-lysine spectra (λ_{ex}=206 nm) clearly shows the disproportionate intensity of the water band in the buffer and peptide spectra with respect to the perchlorate band (Figure 4-2). Subtraction of the buffer spectrum based on the perchlorate band results in a negative feature in the amide I region. Whereas subtraction based on the tail end of the water band resulted in a reasonable spectrum free of negative artifacts.

71

4.4.2. Correlation Optimized Warping

UVRR spectra of 1 mg/mL ovalbumin were collected on four separate days. The calibration of the instrument was expected to change slightly each day, due to a combination of effects which can include expansion of the optics due to temperature and variations in the calibration of the mechanical drive of the monochromter controlling the diffraction grating. In addition, slight variations in the day to day calibration of the spectrometer can arise from slight changes in the optical path, due to re-optimization of the focusing and collection optics. A calibration standard, usually an organic solvent, is typically used to calibrate the spectra post-collection. The UVRR spectra of 1 mg/mL ovalbumin collected on four separate days were calibrated using cyclohexane spectra collected on the same day (Figure 4-3A). The spectra are similar, however, there are obvious variations in the positions of the aromatic modes at ~1000 cm^{-1} and ~1207 cm^{-1} Correlation optimized warping was employed to correct for position variations. A segment size of 10 points with a slack parameter of 1 was employed in the warping of the collected spectra. The standard deviation of the peak maxima of the phenylalanine band at ~1000 cm^{-1} was 2.4 cm^{-1}. After warping, the standard deviation in the position of the peak maxima was decreased to 0.4 cm^{-1}, the corrected spectra are shown in Figure 4-4B.

In order to determine whether correlation optimized warping could be applied to UVRR spectra collected at multiple excitation wavelengths, the UVRR spectra of 1 mg/mL ovalbumin were collected using 197, 201, 205, and 209 nm excitation wavelengths. Cyclohexane UVRR spectra were used to calibrate the protein spectra (Figure 4-4A). Again, the spectra appear misaligned in the aromatic regions, specifically 1580-1600 cm^{-1}.

Application of COW to multi-excitation UVRR spectra in order to alleviate distortions is a challenge due to the increased resolution of the spectrometer with excitation wavelength. The increase in resolution with increasing excitation wavelength coupled with temperature effects and subtle variations in the path of the incident excitation and scattered photons results in slight variations in the observed positions of the amide and aromatic modes as the excitation wavelength changes. The use of a single spectrum as the reference spectrum for the warping algorithm would ultimately lead to progressive distortions in the warped spectra. In order to warp the spectra, without introducing distortions, the UVRR spectra were warped using the previous excitation solutions of ovalbumin. To determine how robust COW is for the treatment of multi-
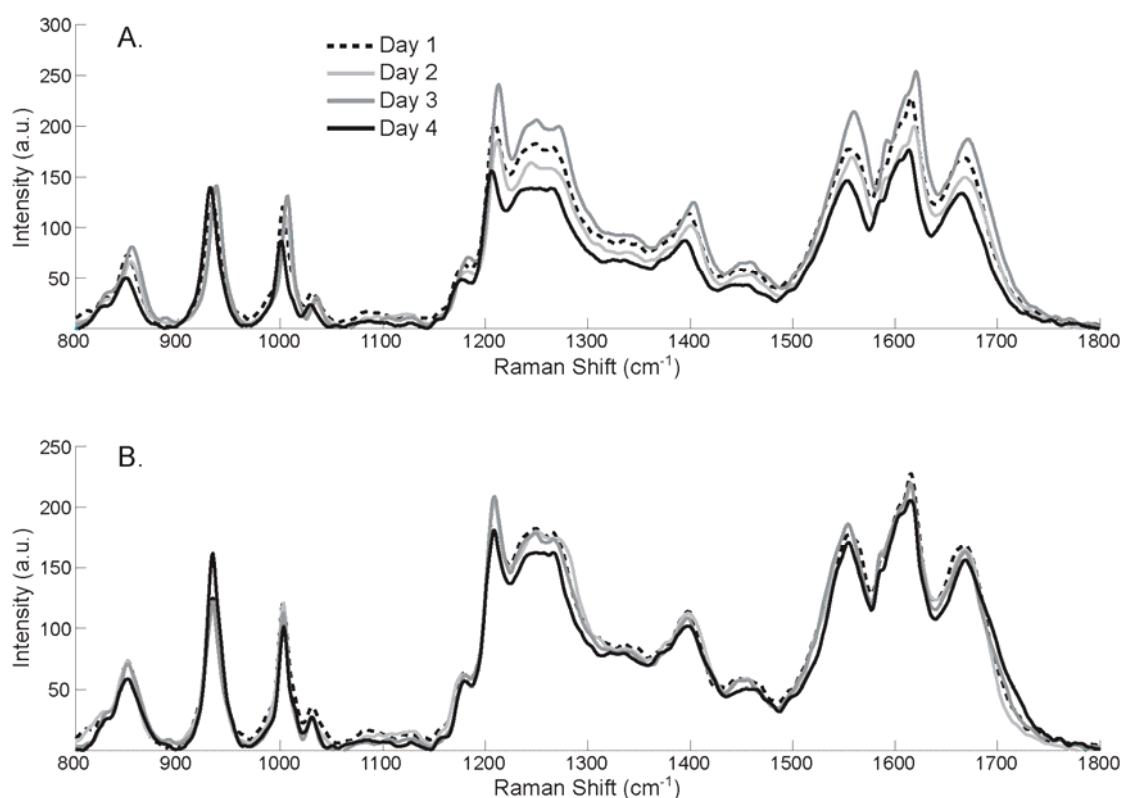
Fig. 4-3 UVRR spectra of 1 mg/ml ovalbumin excited at 200 nm over four days, pre (A) and post (B) application of correlation optimized warping. The water band has not been subtracted in these spectra.

excitation data, the warping procedure was performed stepping from the lowest excitation wavelength to the highest and vice-versa. The result of each warping pathway was found to be essentially identical. However, warping from the highest excitation wavelength to the lowest (Figure 4-4C) produce a slight upshift in the Raman shift of the maxima of the aromatic modes at ~1207 cm^{-1} and a downshift in the Fermi doublet below 900 cm^{-1}, according to the position of these features in the reference spectra, of approximately 2 cm^{-1}.
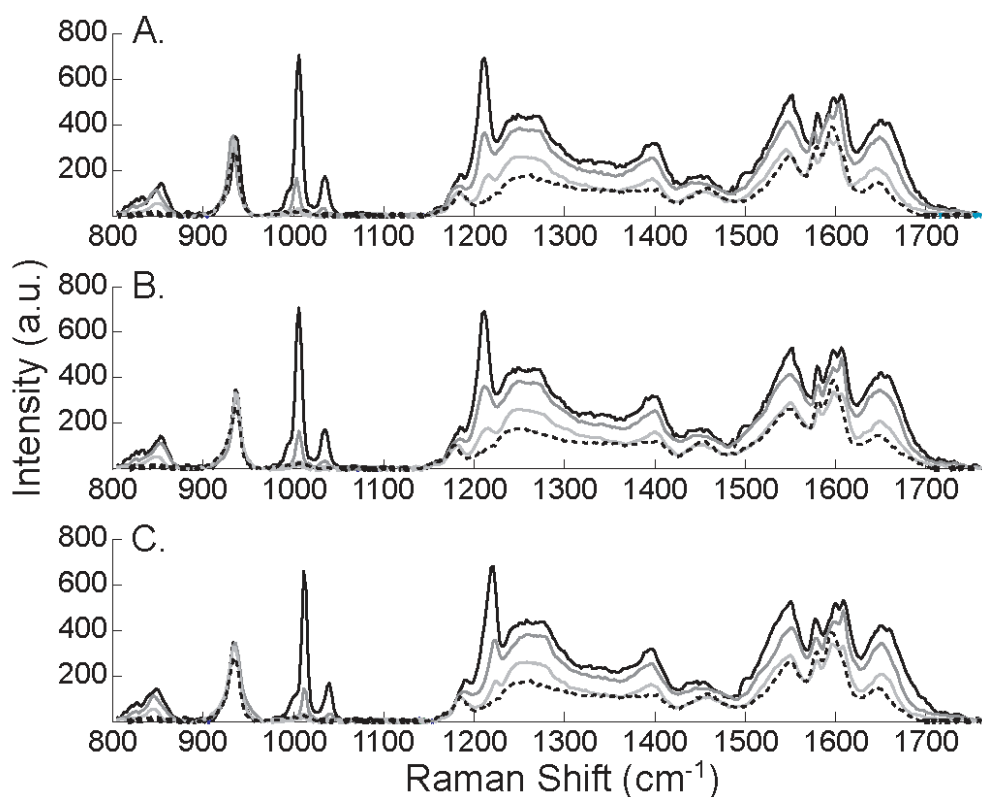

Fig. 4-4 UVRR spectra of ovalbumin (1 mg/ml) excited at 197 nm, 201 nm, 205 nm, and 209 nm, pre (A) and post (B) correlation optimized warping starting at 197 nm as the first reference spectra and (C) 209 nm as the first reference spectrum.

4.5. Conclusion

Removal of the water band from UVRR spectra has been shown to be problematic at low analyte concentrations. The analysis of the UVRR spectra of L-tyrosine at increasing concentrations indicates that the quality of the water band subtraction based on the internal intensity standard is adequate at higher analyte concentrations (100-200 μM), however at low concentration (20 μM) the water band is over subtracted, resulting in negative intensity in the amide I region. The same effect was observed for poly-L-lysine. However, it was shown that the relative intensity could be determined from the tail end of the water band from the 1750-1900 cm^{-1}. Thus enabling removal of the interfering water band without introduction of spectral artifacts at low analyte concentrations.

Slight spectral shifts occur in UVRR spectra due day to day re-optimization of the focusing and collection optics In addition, the increasing spectral resolution as excitation wavelength increases makes multi-excitation data sets especially prone to spectral misalignments. These factors greatly hinder the use of multivariate analysis methods on UVRR spectra of proteins. Through the use of COW, these spectral shifts can be eliminated.

4.6. References

1. Li, B.-Y.; Hu, Y.; Liang, Y.-Z.; Xie, P.-S.; Du, Y.-P. *Analytica Chimica Acta* **2004,** *514*, 69-77.

2. Ebrahimi, D.; Hibbert, D. B. *Journal of Chromatography A* **2008,** *1198-1199*, 181-187.

3. Rutan, S. C.; de Noord, O. E.; Andrea, R. R. *Analytical Chemistry* **1998,** *70*, 3198-3201.

4. Simpson, J. V.; Balakrishnan, G.; JiJi, R. D. *The Analyst* **2009,** *134*, 138-147.

5.   McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc.: New York, 2000.

6.   Hamaguchi, H. O. *Applied Spectroscopy Reviews* **1988,** *24*, 137 - 174.

7.   Wollman, S. T.; Bohn, P. W. *Applied Spectroscopy* **1993,** *47*, 125-126.

8.   Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. *Analytical Chemistry* **1991,** *63*, 2750-2756.

9.   Wang, Y.; Lysaght, M. J.; Kowalski, B. R. *Analytical Chemistry* **1992,** *64*, 562-564.

10.   Kim, S.; Wang, Z.; Hiremath, B. *Annals of Operations Research* **2008**.

11.   Pravdova, V.; Walczak, B.; Massart, D. L. *Analytica Chimica Acta* **2002,** *456*, 77-92.

12.   Giorgio, T.; Frans van den, B.; Claus, A. *Journal of Chemometrics* **2004,** *18*, 231-241.

13.   Zhang, D.; Huang, X.; Regnier, F. E.; Zhang, M. *Analytical Chemistry* **2008,** *80*, 2664-2671.

14.   Scherer, J. R.; Go, M. K.; Kint, S. *The Journal of Physical Chemistry* **1974,** *78*, 1304-1313.

15.   Dudik, J. M.; Johnson, C. R.; Asher, S. A. *The Journal of Chemical Physics* **1985,** *82*, 1732-1740.

16.   Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. *Journal of Chromatography A* **1998,** *805*, 17-35.

17.   Hillier, F. S.; Lieberman, G. J. *Introduction to Mathematical Programming*; McGraw-Hill, 1995.

18.   Edelhoch, H. *Biochemistry* **2002,** *6*, 1948-1954.

19.   Fasman, G., Ed. *Practical handbook of biochemistry and molecular biology*; CRC Press LLC: Boca Raton, Florida, 1989.

20.   Huang, C. Y.; Balakrishnan, G.; Spiro, T. G. *Biochemistry* **2005,** *44*, 15734-15742.

21.   Lednev, I. K.; Ermolenkov, V. V.; He, W.; Xu, M. *Analytical and Bioanalytical Chemistry* **2005,** *381*, 431-437.

22.   Wu, Q.; Balakrishnan, G.; Pevsner, A.; Spiro, T. G. *J. Phys. Chem. A* **2003,** *107*, 8047-8051.

23.   Ferraro, J. R.; Nakamoto, K. *Introductory Raman Spectroscopy*; Academic Press, Inc.: New York, 1994.

25.   Fodor, S. P. A.; Copeland, R. A.; Grygon, C. A.; Spiro, T. G. *J. Am. Chem. Soc.* **1989,** *111*, 5509-5518.

Chapter 5: Using EEM fluorescence in combination with PARAFAC analysis to simultaneously monitor quercetin in it deprotonated, aggregated, and protein bound states

Flavonoids, a group of naturally occurring polyphenols, are of great interest due to their many apparent health benefits including their anti-oxidant properties and anti-amyloidogenic properties. Selected flavonoids have been shown to reduce the toxicity of the β-amyloid peptide *in vitro*. However, the behavior of flavonoids in aqueous environments can be unpredictable due to the potential formation of concentration and pH dependent aggregates, in addition to hydrolysis products. A better understanding of the complex behavior of flavonoids is needed before flavonoid based therapies can be introduced. Therefore, we have characterized a series of aqueous quercetin solutions, a potent anti-amyloidogenic compound, which varied in both pH and concentration in the presence of BSA, using excitation-emission matrix (EEM) fluorescence. Analysis with parallel factor analysis (PARAFAC) indicates distinct concentration and pH dependent fluorophores, suggesting aggregates are present at low micro-molar concentrations and more likely to form at particular pH ranges. In addition, results indicate a quantifiable spectral change due to the interaction of quercetin with BSA in solution. Coupled with PARAFAC and EEM fluorescence, we demonstrate detection and quantification of multiple fluorescence quercetin species in solution at a wide range of pH and concentration values simultaneously, including the interaction of quercetin with BSA.

5.1. Introduction

Flavonoids are a large group of naturally occurring molecules which are utilized in plants for pigmentation, nitrogen fixation, and chemical defense along with other biological processes.[1] They are found in many fruits, flowers, and plant derived products and as a result, these compounds are abundant in the human diet.[2] They have received much attention in recent years due to the possible positive health benefits of a diet high in flavonoids. Flavonoids have been shown to have a wide range of therapeutic properties such as: anti-inflammatory, antineoplastic and cardioprotective properties.[3,4] However, the methods by which these effects are achieved are still poorly understood. While flavonoids have been shown to be potent antioxidants, and act as free-radical scavengers[5,6], they readily bind proteins in solution, which may impair their anti-oxidant activity.[7,8] In addition, many common flavonoids, such as quercetin, are oxidized under a variety of conditions, leaving an incomplete picture of the behavior of flavonoids in solution.[9-12]

The term flavonoid encompasses a wide array of compounds including flavanones, anthocyanidins, flavones, and flavonols. They have a common architecture in that they all have a diphenylpropane ($C_6C_3C_6$) skeleton featuring an oxygen containing heterocycle. Different classifications of compounds arise from variations in the side groups which populate the open sites of the skeleton, such as the number and locations of hydroxyl groups. Flavonols, such as quercetin, show evidence of a photo-induced tautomerism when excited in the UV.[13] This tautomerism occurs due to an allowed excited state intramolecular proton transfer (ESIPT) from the 3-hydroxy group to the nearby carbonyl group allowing conversion between the enolic and keto forms (Figure 5-

1).[14-16] This proton transfer is extremely sensitive to proton transfer from solvents and hydrogen bonding effects, often leading to a reduction in the photo-tautomerism process.[17-20]
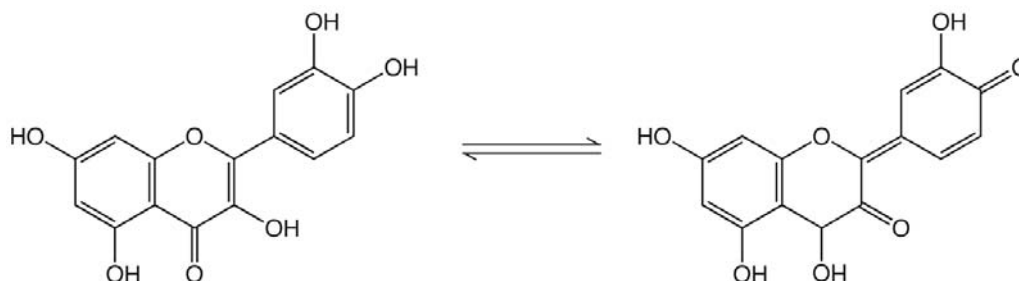

Fig. 5-1 Chemical structure of quercetin in both the enolic (left) and keto (right) forms.

While previous studies of quercetin and other flavonoids have established a number of possible health benefits from the consumption of flavonoids, understanding the methods by which these effects are achieved is a continuing process. While it is known that these molecules freely interact with proteins in solution, and have high binding affinities for some proteins *in vitro* [21,22], the nature of that interaction is still unknown. Research has shown that these compounds have an effect on various enzymes and transduction pathways, such as inhibition of DNA topoisomerases and phosphorylase kinase.[23-25] Quercetin in particular has been shown to have a high affinity for binding with human serum albumin.[26-29] Quercetin's affinity for binding to serum albumin is of particular interest as serum albumin is an abundant carrier protein, allowing transportation of endogenous and exogenous ligands to their target receptors.[30] As such, an understanding of how flavonoids, such as quercetin, bind to transport proteins is of critical interest to the therapeutic use of flavonoids. Numerous studies have examined the fluorescence response of quercetin binding to either human or bovine serum albumin.

Normally, a decrease in the fluorescence intensity of the aromatic amino acids is noted with increasing quercetin concentration due to fluorescence quenching.[31] In addition, shifts in the fluorescence emission spectra of quercetin in the presence of BSA have been reported.[32] The fluorescence of tryptophan residues in serum albumin undergoes quenching in the presence of quercetin and other flavonoids.[26] Increasing fluorescence intensity has also been observed for quercetin in both the presence and absence of BSA, with increasing pH in solution.[33]

Previous studies have utilized fluorescence spectra collected at a handful of excitation wavelengths to tease out the complex behavior of flavonoids in aqueous environments and in the presence of proteins. However, the excitation and emission spectra of the various forms of flavonoids, (enol, keto, aggregate, and protein bound) are highly overlapped, leading to complex spectra with multiple underlying components.

In order to gain a better understanding of the complex behavior of flavonoids and their interactions with proteins, methods in which the various forms can be monitored and quantified simultaneously need to be developed. Simultaneous measurement of the various forms of the flavonoid quercetin was achieved using excitation-emission matrix (EEM) fluorescence. Parallel factor analysis (PARAFAC) was then employed to resolve the pure spectral profiles of the various forms of quercetin. In addition, pH and concentration dependent behavior were also modeled. Employment of EEM fluorescence in combination with PARAFAC analysis allowed concurrent determination of concentration, pH, excitation, and emission profiles for each quercetin derived species, providing a more complete picture of the complex equilibra occurring in solution.

5.2. Theory

A major advantage of fluorescence spectroscopy is the ability to measure fluorophores at concentrations well below the capabilities of normal absorption spectroscopy. In addition, many molecules do not exhibit significant fluorescence, greatly reducing the number of spectral interferents. However, excitation and emission fluorescence spectra are often very broad in nature, leading to decreased selectivity if multiple fluorophores are present. One method which has been employed to overcome this limitation is the use of excitation-emission matrix (EEM) fluorescence spectra, i.e. collecting the emission spectra of the sample at a number of excitation wavelengths, followed by the use of advanced multivariate deconvolution techniques.[34,35]

PARAFAC is an extremely powerful chemometric method which has been used effectively in the analysis of multi-way data to determine both qualitative and quantitative information. [34-36] PARAFAC is based on the trilinear model, as shown in Equation 1.[36]

$$R_{ijk} = \sum_{n=1}^{N} \hat{X}_{i,n} \hat{Y}_{j,n} \hat{Z}_{k,n} + E_{ijk} \quad (1)$$

In Equation 1, R_{ijk} represents the measured response of the kth sample at the ith excitation and jth emission wavelengths. In this study, each kth slice of the trilinear data cube R_{ijk} represents one EEM fluorescence spectrum in a series of samples and has the dimensions of $I \times J$, where I is the number of excitation wavelengths and J is the number of emission wavelengths. The number of factors, N, chosen to model the system, is defined by the user. Determination of the correct number of factors is often based on a combination of experimental knowledge of the system and a comparison of the sum of

squared residuals (SSR) for a series of models with an increasing number of factors. The SSR may be determined by solving Equation 2 for each model .[35]

$$\sum_{i,j,k}\left(R_{i,j,k} - \sum_{n=1}^{N} \hat{X}_{i,n}\hat{Y}_{j,n}\hat{Z}_{k,n}\right)^2 \quad (2)$$

The variables $\hat{X}_{i,n}$, $\hat{Y}_{j,n}$, and $\hat{Z}_{k,n}$ represent the estimated responses of the nth factor at the ith excitation and jth emission wavelengths, in the kth sample. The difference between the measured value, R_{ijk}, and the estimated value, $\hat{R}_{ijk}$, is the sum of the product of $\hat{X}_{i,n}$, $\hat{Y}_{j,n}$, and $\hat{Z}_{k,n}$ over all N factors, and represented by an error term, E_{ijk}.

An alternating least squares (ALS) algorithm is used to estimate the three component matrixes, **X**, **Y**, and **Z**. In the ALS process, each component matrix is estimated from the data and prior estimates of the other two component matrices. Each component matrix is solved in turn, with the entire process repeating until the estimates of the component matrices are no longer changing. Initial estimates of the two component matrices are needed to begin the ALS process. Matrices of random numbers, in the proper dimensions can be used if good estimates of the pure spectral profiles are unavailable. The ability to use randomly generated profiles to initialize the ALS process is a distinct advantage of this method, as good estimates of the pure profiles are difficult to generate when complex equilibria between fluorophores are involved.

Initial estimates of the matrix **Z** are found from the initialization values for **X** and **Y**, as well as the data cube **R**. The K slices of the data cube **R** are unfolded in the XY plane into the matrix $\mathbf{R_C}$ which is of dimensions $IJ \times K$. This allows for solving

$$\mathbf{R_C} = \mathbf{C}\hat{\mathbf{Z}}^{\mathrm{T}} \quad (3)$$

such that

$$\hat{\mathbf{Z}}^{\mathrm{T}} = \mathbf{C}^{+}\mathbf{R}_{\mathrm{C}} \quad (4)$$

resulting in the calculation of $\hat{\mathbf{Z}}$. The matrix $\mathbf{C}^{+}$ is the pseudo-inverse of $\mathbf{C}$, a matrix

which is composed in a similar fashion to $\mathbf{R}_{\mathrm{C}}$ from the N columns of X and Y such that

$C_{(j-1)I+i,n} = X_{i,n}Y_{j,n}$ just as $R_{C(j-1)I+i,k} = R_{i,j,k}$.

      Subsequently the component matrix, $\mathbf{X}$, is estimated from the new estimate of $\mathbf{Z}$.

Again, the data cube $\mathbf{R}$ is unfolded, this time in the YZ plane, creating $\mathbf{R}_{\mathrm{A}}$ where

$R_{A(k-1)I+i,j} = R_{i,j,k}$. The new estimate of $\mathbf{X}$ may be found then by solving the equation

$$\mathbf{R}_{\mathrm{A}} = \mathbf{A}\mathbf{X}^{\mathrm{T}} \quad (5)$$

such that

$$\hat{\mathbf{X}}^{\mathrm{T}} = \mathbf{A}^{+}\mathbf{R}_{\mathrm{A}} \quad (6)$$

in the same way that the estimate of $\mathbf{Z}$ was found. An analogous procedure is then

performed for the estimation of the component matrix, $\mathbf{Y}$. After each iteration, in which

new estimates of each component matrix are calculated, the difference between the prior

and current estimates are computed. The algorithm proceeds iteratively, repeating the

mentioned steps till the difference between the prior and current estimates is nominal and

essentially unchanged. Typically a threshold value is used to indicate convergence.

      The critical assumption of the PARAFAC method is that the data is trilinear in

nature, meaning that the contribution of each component to each EEM can be defined as

the product of three variables, the relative excitation, emission, and concentration.

However, the assumption of trilinearity is not always fulfilled in chemical data.[37] EEM

fluorescence spectra typically contain contributions from both Raleigh and Raman

scatter, greatly complicating PARAFAC analysis as the scattering is dependent on the

excitation wavelength and therefore does not fulfill the trilinearlity constraint.[38] In order

to minimize the influence of Rayleigh and Raman scatter on the analysis, a weighted

PARAFAC algorithm was employed. [38-45]

The weighted PARAFAC algorithm is similar to the PARAFAC algorithm. The

principle difference being the introduction of a weighting matrix, which can be used to

mitigate the effects of data points which do not fit the trilinear model.[45,38] Typically, the

weighting matrix is composed of simple binary values designated by the user. The

weighting matrices were designed such that regions corresponding to Raleigh or Raman

scattering were given a weight of zero. To construct the appropriate weighting matrix for

the EEM fluorescence spectra, the EEM spectra of the buffer blank was used to identify

the pattern of Raleigh and Raman scatter; these regions were then given a weight of zero.

All other regions were assigned a weight of one. Identical weighting matrices were used

for all EEM spectra. The SSR for a weighted PARAFAC model, and the quality of the

simulation of the data, can be found by solving Equation 7. Alternatively, Equation 7 can

be minimized through the minimization of equations 8a through 8c at each iterative step

of the algorithm.

$$\sum_{i,j,k}\left[W_{i,j,k}\left(R_{i,j,k} - \sum_{n=1}^{N} \hat{X}_{i,n}\hat{Y}_{j,n}\hat{Z}_{k,n} \right)^{2} \right] \qquad (7)$$

$$\left\| \mathbf{W_A} \circ \left(\mathbf{R_A} - \mathbf{A}\hat{\mathbf{X}}^{\mathbf{T}} \right) \right\|_{\mathbf{F}}^{2} \quad (8a)$$

$$\left\| \mathbf{W_B} \circ \left(\mathbf{R_B} - \mathbf{B}\hat{\mathbf{Y}}^{\mathbf{T}} \right) \right\|_{\mathbf{F}}^{2} \quad (8b)$$

$$\left\| \mathbf{W_C} \circ \left(\mathbf{R_C} - \mathbf{C}\hat{\mathbf{Z}}^{\mathbf{T}} \right) \right\|_{\mathbf{F}}^{2} \quad (8c)$$

In the above equations, $\mathbf{W_A}$, $\mathbf{W_B}$, and $\mathbf{W_C}$ are made by performing the same unfolding

procedure as was performed to construct $\mathbf{R_A}$, $\mathbf{R_B}$, and $\mathbf{R_C}$.

5.3. Experimental

5.3.1 Materials and sample preparation:

Mono- and dibasic potassium phosphate, quercetin dihydrate, bovine serum albumin (BSA) and spectroscopic grade acetone were purchased from Sigma-Aldrich (St. Louis, MO) and used without further purification.

Phosphate buffers were prepared by dissolving appropriate amounts of mono- and dibasic potassium phosphate in 18.2 MΩ RO (reverse osmosis) water. Buffers were prepared with final pH values ranging from 6.2 to 8.0 in 0.2 pH increments at a concentration of 10 mM. Buffers were then pH adjusted with small additions of mono- or dibasic potassium phosphate as needed. All buffers were filter sterilized and sparged with argon for 20 minutes to remove oxygen.

Concentrated quercetin stock solutions were prepared by dissolving quercetin dihydrate in 10 mL of spectroscopic grade acetone. The concentration of the quercetin dihydrate stock solution was determined by a 1000 fold dilution of the stock solution in 95% ethanol. The absorption spectrum of the resulting solution was collected on a Hewlett Packard 8453 UV-Vis Spectrometer (Palo Alto, CA). The quercetin concentration of the diluted solution was estimated from the molar extinction coefficient of quercetin dihydrate at 374 nm peak (21880 $M^{-1}cm^{-1}$ in 95% ethanol)[46]. The concentration of the stock solution was then back calculated from the concentration of the dilute quercetin solution. Subsequent dilutions from this stock were made as needed.

Fresh BSA stock solutions were prepared by dissolving BSA in 10 mL of pre-prepared buffer prior to each experiment. The concentration of the BSA stock solution was determined by a 10 fold dilution of the stock solution in buffer. The absorption

spectrum of the resulting solution was then collected on a Hewlett Packard 8453 UV-Vis Spectrometer (Palo Alto, CA). The concentration of BSA was then estimated from the molar extinction coefficient of BSA at 280 nm peak (43824 $M^{-1}cm^{-1}$).[47] The concentration of BSA in the stock solution was then back calculated from this value.

Quercetin solutions at 5 and 25 μM were prepared by diluting small aliquots (<100 μL) of the quercetin/acetone stock solution up to 10 mL with the appropriate pH buffer. At pH 7.4, a series of quercetin samples were prepared at the following concentrations: 0, 0.25, 0.5, 1, 5, 10, 25, 50, 75, and 100 μM. In addition a series of quercetin sample solutions were prepared at 0, 5, 10, 15, 20, 25, 50, 75, and 100 μM at pH 7.4 with and without 5 μM BSA. Freshly prepared solutions were then degassed under vacuum for 15 minutes prior to use in spectroscopic analysis.

5.3.2 Fluorescence:

Polymethacrylate fluorescence cuvettes, designed to be used in the range of 285-800 nm, were purchased from Perfector Scientific (Atascadero, CA). All fluorescence emission spectra were measured on a Cary-Eclipse Fluorometer (Varian, Palo Alto, CA). Fluorescence emission spectra for quercetin solutions were collected from 300 to 650 nm, in 2 nm increments, at increasing excitation wavelengths from 270 to 500 nm, in 10 nm increments with an excitation and emission slit width of 5 nm. For samples containing BSA, fluorescence emission spectra were collected from 400 to 750 at 2 nm increments to include the secondary fluorescence of tryptophan, at increasing excitation wavelengths from 300 to 450 nm, in 5 nm increments with an excitation and emission slit width of 5 nm.

5.3.3 Data analysis:

EEM spectra were analyzed using weighted- PARAFAC[48,49] in the Matlab environment, version 7.1 (Mathworks, Natick, MA). The weighted PARAFAC algorithm was obtained from by Dr. Karl Booksh (University of Delaware) and used without modification.

5.4. Results and discussion

The behavior of quercetin in solution is difficult to characterize. However, previous work has established a relationship between increasing pH, over pH 6.4-8.6, and increased fluorescence intensity. In order to characterize this established behavior in solution the EEM spectra of a series of multi-concentration solution was analyzed with the PARAFAC algorithm. Figure 5-2 shows the resolved emission, excitation, and pH profiles for a 5 μM quercetin solution from pH 6.2 to pH 8.0. Four components were needed to model the fluorescent species, as well as residual scattering. As can be seen in the pH profiles, a relationship exists between the solution pH and emission intensity in the component centered at 540 nm (emission). The pH profile levels off above pH 7.6, which is the experimentally determine pKa for hydroxyl deprotonation in quercetin. The excitation profile is broad, ranging from 350-400 nm. Due to the pH dependence of this peak, we conclude that this quercetin species is likely a combination of the basic forms of the enolic and keto tautomers. An additional component was needed to model a fluorescent impurity, the source of which could not be identified definitively or eliminated. However, the impurity appeared in blank solutions, (buffer) indicating that neither quercetin nor acetone were the sources. The impurity had an excitation and
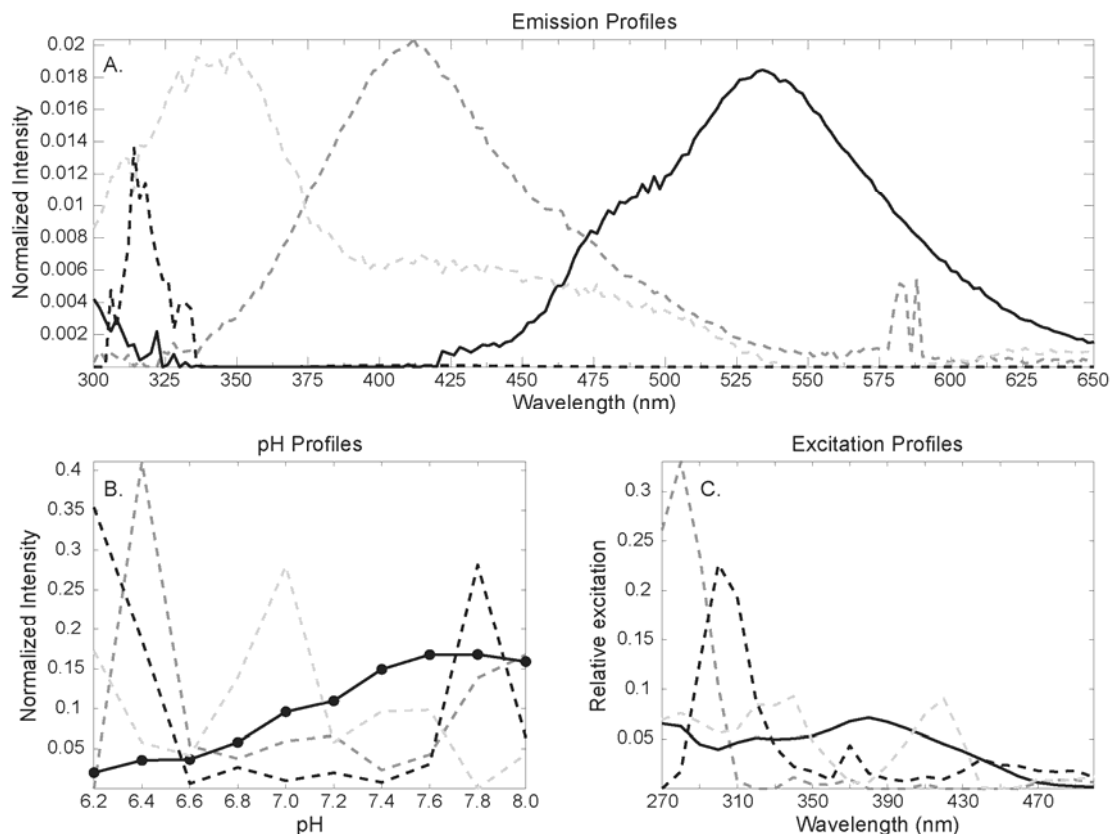
Fig. 5-2 (a) Emission, (b) pH, and (c) excitation profiles of 5 µM quercetin at pH values ranging from 6.2 to 8.0. The solid black line denotes the profiles which are believed to arise from the combination of the enolic and keto forms of quercetin at these conditions. The dark dashed grey line denotes the profiles which arise from the fluorescent contaminant present in the buffer. Both the black dashed and light grey dashed lines are factors which we believe models Raman and Rayleigh scatter which was not completely eliminated by the weighting matrix.

emission maxima of 300 and 410 nm, respectively. The impurity appeared randomly

throughout the experiments. An advantage of this method is that such impurities or

interferents can be modeled along with the fluorophores of interest. Two additional

components were need to model background fluorescence and residual scattering. The

sharp narrow peak around 310 nm in Figure 5-2a is likely residual Raman scattering as

the 310 nm corresponds to the Raman shift of water at 280 nm excitation, the excitation

maximum of this component (Figure 5-2c). Further, the pH profiles of this component

(Figure 5-2c) is random, with only two large intensity values at pH 6.8 and pH 7.8, suggesting the scattering in these samples was of greater intensity.

When the concentration of quercetin was increased from 5 µM to 25 µM, an additional component centered at 600 nm (emission) was needed to model the EEM spectra (Figure 5-3). The excitation profile of the higher wavelength component was broad and similar to that of the combination enolic/keto tautomer factor (Figure 5-3c). Additionally, the relative concentration of this component does not appear to be pH dependent (Figure 5-3b). Again the pH dependence in the 540 nm (emission) is
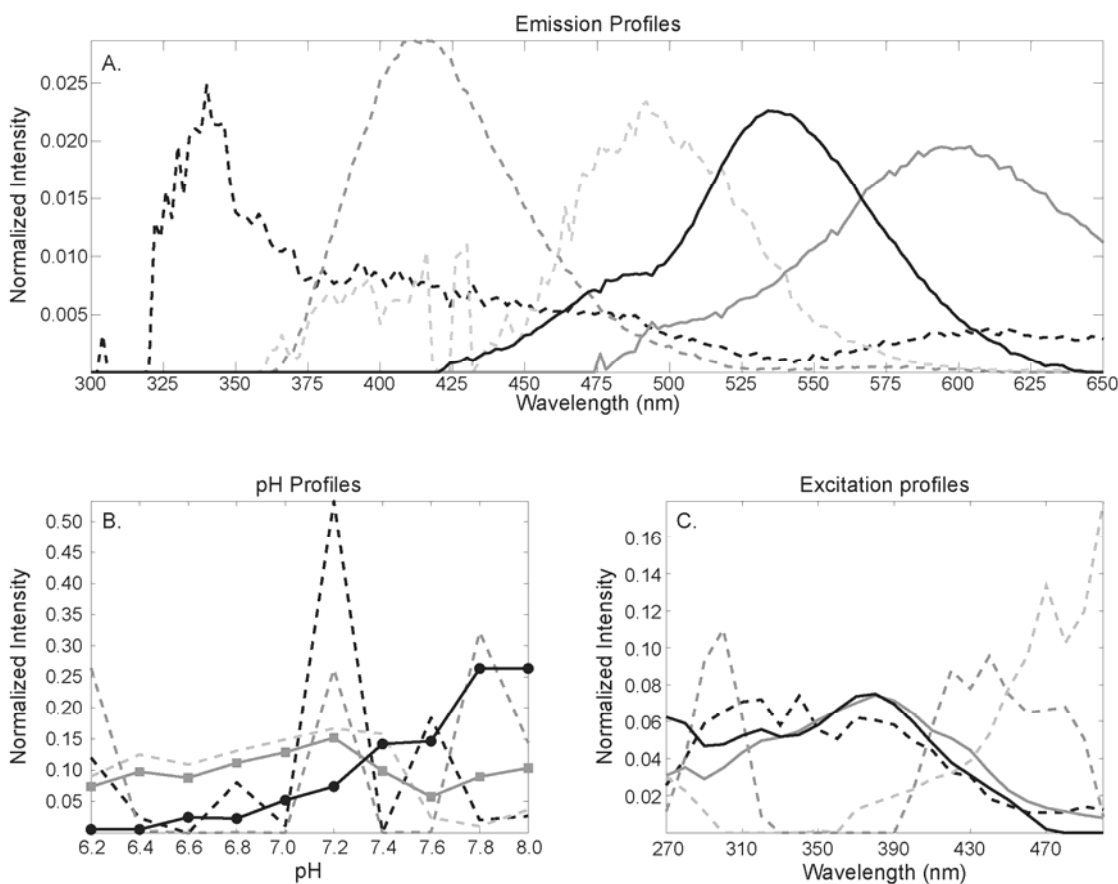

Fig 5-3 (a) Emission, (b) pH, and (c) excitation profiles of 25 µM quercetin at pH values ranging from 6.2 to 8.0. The solid black line denotes the profiles which are believed to arise from the combination of the enolic and keto forms of quercetin under these conditions. The solid grey line denotes the profiles which are believed to arise from higher order aggregates of quercetin (>dimer). The grey dashed line is believed to arise from an unknown interferent which is present in the buffer. Both the light grey dashed line and dashed black line denote factors which are believed to arise from a combination of Rayleigh and Raman scattering which was not eliminated with use of the weighting matrix.

observable, exhibiting the same behavior as in the lower concentration solution. Other

components were used in to again model the presence of an interferent (λ_{em}= 410 nm) and

Raman/Rayleigh scatter not eliminated by weighting matrix. The appearance of the

additional component centered at 600 nm (emission) at a higher concentration implies a

concentration dependence in the formation of the species.

The concentration dependence of the two quercetin species are shown in Figure 5-

4b. Fluorescence from the enolic/keto tautomer (λ_{em}= 540 nm) is detectable at 5 μM
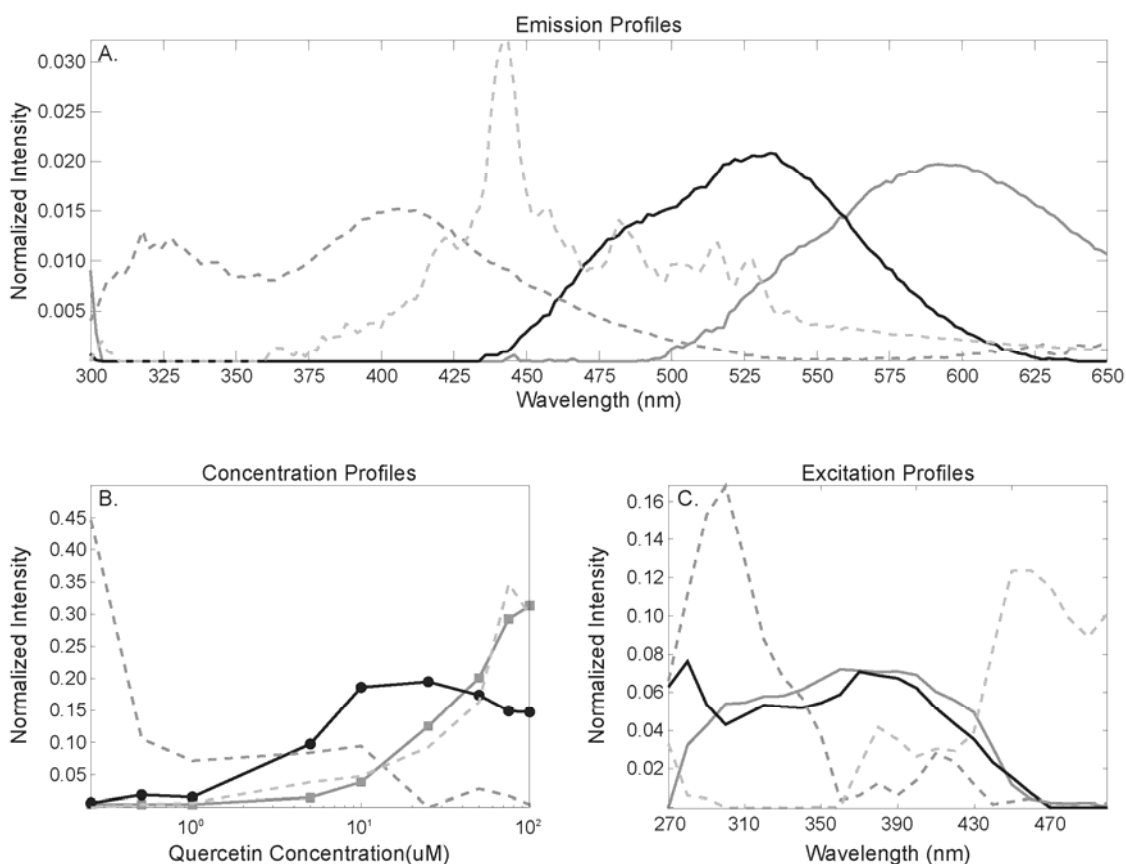

Fig 5-4 (a) Emission, (b) concentration, and (c) excitation profiles of quercetin solutions ranging from 0.25 to 100 μM at pH 7.4. The solid black line denotes the profiles which are believed to arise from the combination of the enolic and keto forms of quercetin at these conditions. The solid grey line denotes the profiles which are believed to arise from higher order aggregates of quercetin (>dimer). The dark grey dashed line is believed to arise from the combination of an unknown interferent which is present in the buffer and Rayleigh/Raman scatter which was not eliminated from the model through the use of the weighting matrix. The light grey dashed line denotes factors which are believed to arise from a combination of Rayleigh and Raman scattering which was not eliminated with use of the weighting matrix.

quercetin and reached a maximum by 10 μM. Interestingly, fluorescence from the higher wavelength component (λ_{em}= 600 nm) begins to appear at 10 μM and steadily increases with increasing quercetin concentration. This data suggest that the pH independent fluorophore is an aggregate species, of higher order than a dimers, since the enol/keto tautomerization is stabilized in the dimer form[13]. As such, the higher wavelength component (λ_{em}= 600 nm) can be assigned to an aggregated form of quercetin. An additional component was necessary in order to model a combination of the fluorescence of the interferent, due to the local emission maxima at 410 nm, and Raman scattering from water due to the emission at 310 nm.

As already mentioned, quercetin has a number of biological effects which are of interest. While the previous experiments have characterized the behavior of quercetin in solution, the behavior of quercetin in the presence of proteins is of great interest in order to better understand the biological effects of quercetin and how it travels through a biological system. BSA serves as a good model protein for studies of protein-flavonoid interaction. Figure 5-5 shows the emission, excitation, and concentration profiles of quercetin with and without BSA. An additional component appears in the presence of BSA, which has an emission maximum of approximately 535 nm and a smaller shoulder at 450 nm. The excitation profile of this component is well resolved from the excitation profiles of the enol/keto forms and the excitation profile of the higher order aggregates of quercetin with a maximum excitation at 445 nm. This component can be assigned to BSA-bound quercetin and has been observed previously [32]. Although this component was modeled in the non-BSA containing samples, the relative intensity was low and likely modeling intensity from the free enolic/keto form normally anticipated at 540 nm, and a result of

the inclusion of both BSA and non-BSA containing data in the PARAFAC model.

Interestingly, the drop off in the intensity of the enol/keto form is more pronounced than

the previous study (Figure 5-4b), which we believe to be due to an artifact of the model

arising from the spectral overlap of the BSA-quercetin complex emission and the
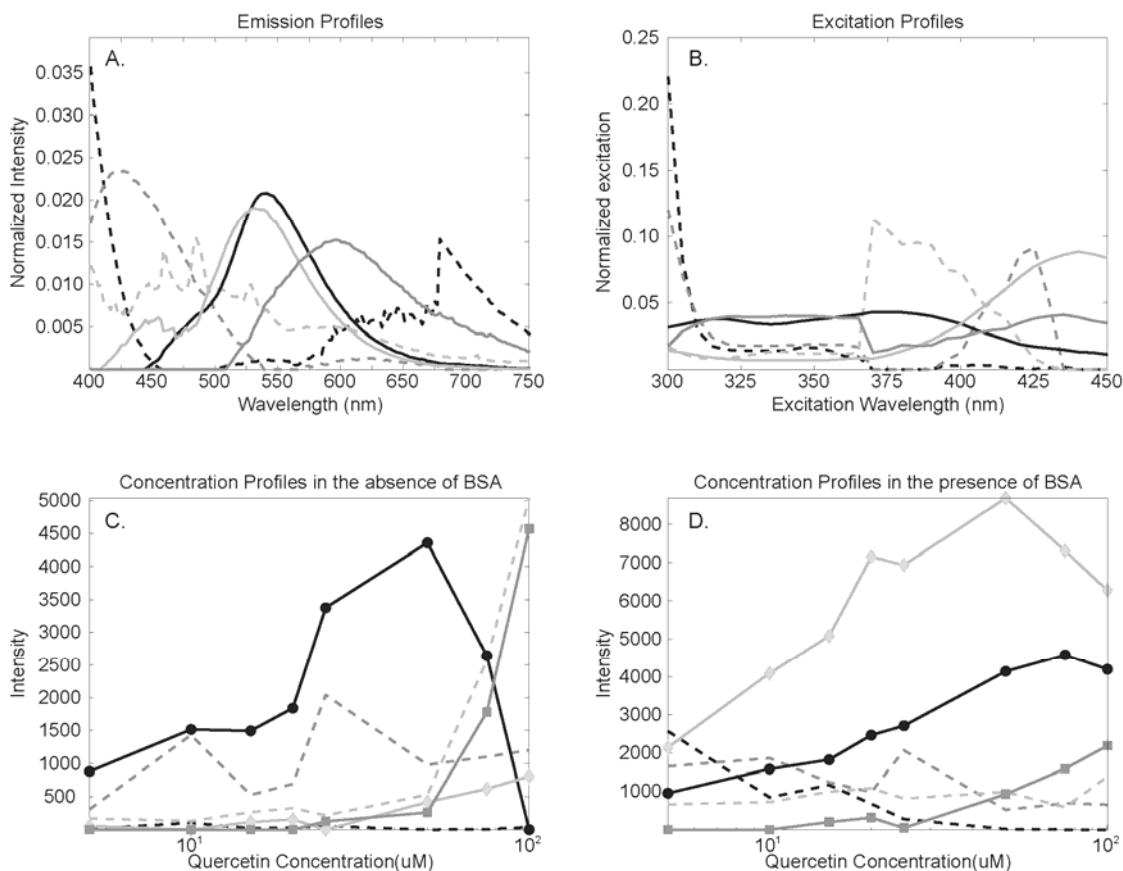
emission of the enol/keto form.

Fig 5-5 (a) Emission, (b) excitation, (c) concentration in the absence of BSA, and (d) concentration in the presence of BSA profiles of quercetin solutions ranging from 5 to 100 µM at pH 7.4, with 5 µM of BSA held constant in the appropriate sample. The solid black line denotes the profiles which are believed to arise from the combination of the enolic and keto forms of quercetin at these conditions. The solid grey line denotes the profiles which are believed to arise from higher order aggregates of quercetin (>dimmer). The solid light grey line denotes the profiles which are believed to arise from the BSA-quercetin complex. The dark grey dashed line is believed to arise from an unknown interferent which is present in the buffer. The light grey dashed line denotes factors which are believed to arise from a combination of Rayleigh and Raman scattering which was not eliminated with use of the weighting matrix. The black dashed line represents the factors arising from the intrinsic fluorescence of BSA.

5.5. Conclusion

The first component associated with quercetin exhibited a fluorescence emission maximum at 540 nm and a broad excitation profile ranging from 300-450 nm. This component exhibited a strong pH dependence, leading to higher fluorescence intensities, most likely due to an increase in quantum yield due to increased resonance structures from deprotonation of quercetin. As such, it was identified as the basic enolic/keto form of quercetin. The second component associated with quercetin had a very similar excitation profile to basic enolic/keto form of quercetin. However, it was characterized by a fluorescence emission maximum located at 600 nm. This component demonstrated no pH dependence, but it did show a strong concentration dependence. As such, this species was assigned to higher order quercetin aggregates. The third component associated with quercetin was assigned to the BSA-quercetin complex. This component exhibited a concentration dependence with the presence of BSA.

The behavior of quercetin in solution is complex and not easily determined though conventional fluorescence spectroscopy. While EEM fluorescence spectroscopy provides a wealth of data on the fluorophores of a system, relationships present in data are difficult to manually determine. However, by coupling EEM spectra to weighted PARAFAC analysis, it was possible to simultaneously determine the spectral and chemical properties of each major component in the system. In addition interferents can be modeled along with fluorophores of interest.

5.6. References

1. *The Flavonoids, Advances in Research since 1986*; Chapman and Hall: London, 1994.

2. Harborne, J. B.; Williams, C. A. *Phytochemistry* **2000,** *55*, 481-504.

3. NC Cook; Samman, S. *The Journal of Nutritional Biochemistry* **1996,** *7*, 66-76.

4. E. Middleton; C. Kandaswami; Theoharides., T. C. *Pharmacological Reviews* **2000,** *52*, 673-751.

5. Counet, C.; Collin, S. *Journal of Agricultural and Food Chemistry* **2003,** *51*, 6816-6822.

6. Brenna, O. V.; Pagliarini, E. *Journal of Agricultural and Food Chemistry* **2001,** *49*, 4841-4844.

7. Riedl, K. M.; Hagerman, A. E. *Journal of Agricultural and Food Chemistry* **2001,** *49*, 4917-4923.

8. Arts, M. J. T. J.; Haenen, G. R. M. M.; Wilms, L. C.; Beetstra, S. A. J. N.; Heijnen, C. G. M.; Voss, H.-P.; Bast, A. *Journal of Agricultural and Food Chemistry* **2002,** *50*, 1184-1187.

9. Makris, D. P.; Rossiter, J. T. *Journal of Agricultural and Food Chemistry* **2000,** *48*, 3830-3838.

10. Krishnamachari, V.; Levine, L. H.; Pare, P. W. *Journal of Agricultural and Food Chemistry* **2002,** *50*, 4357-4363.

11. Hanem M. Awad; Marelle G. Boersma; Jacques Vervoort; Rietjens, I. M. C. M. *Archives of Biochemistry and Biophysical* **2000,** *378*, 224-233.

12. Galati, G.; Moridani, M. Y.; Chan, T. S.; O'Brien, P. J. *Free Radical Biology and Medicine* **2001,** *30*, 370-382.

13. Smith, G. J.; Markham, K. R. *Journal of Photochemistry and Photobiology A: Chemistry* **1998,** *118*, 99-105.

14. McMorrow, D.; Kasha, M. *The Journal of Physical Chemistry* **1984,** *88*, 2235-2243.

15. Sengupta, P. K.; Kasha, M. *Chemical Physics Letters* **1979,** *68*, 382-385.

16. Kasha, M. *Faraday Transactions 2* **1986,** *82*, 2379.

17. Woolfe, G. J.; Thistlethwaite, P. J. *Journal of the American Chemical Society* **1981,** *103*, 6916-6923.

18. Itoh, M.; Fujiwara, Y.; Sumitani, M.; Yoshihara, K. *The Journal of Physical Chemistry* **1986,** *90*, 5672-5678.

19. Itoh, M.; Tokumura, K.; Tanimoto, Y.; Okada, Y.; Takeuchi, H.; Obi, K.; Tanaka, I. *Journal of the American Chemical Society* **1982,** *104*, 4146-4150.

20. Brewer, W. E.; Studer, S. L.; Chou, P.-T.; Orton, E. *Chemical Physics Letters* **1989,** *158*, 345-350.

21. Manach, C.; Morand, C.; Texier, O.; Favier, M.-L.; Agullo, G.; Demigne, C.; Regerat, F.; Remesy, C. *J. Nutr.* **1995,** *125*, 1911-1922.

22. Manach, C.; Texier, O.; Régérat, F.; Agullo, G.; Demigné, C.; Rémésy, C. *The Journal of Nutritional Biochemistry* **1996,** *7*, 375-380.

23. Srivastava, A. K. *Biochemical and Biophysical Research Communications* **1985,** *131*, 1-5.

24. Matter, W. F.; Brown, R. F.; Vlahos, C. J. *Biochemical and Biophysical Research Communications* **1992,** *186*, 624-631.

25. Boege, F.; Straub, T.; Kehr, A.; Boesenberg, C.; Christiansen, K.; Andersen, A.; Jakob, F.; Köhrle, J. *J. Biol. Chem.* **1996,** *271*, 2262-2270.

26. Sengupta, B.; Sengupta, P. K. *Biochemical and Biophysical Research Communications* **2002,** *299*, 400-403.

27. Bidisa, S.; Pradeep, K. S. *Biopolymers* **2003,** *72*, 427-434.

28. Dufour, C.; Dangles, O. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2005,** *1721*, 164-173.

29. Kaldas, M. I.; Walle, U. K.; van der Woude, H.; McMillan, J. M.; Walle, T. *Journal of Agricultural and Food Chemistry* **2005,** *53*, 4194-4197.

30. Carter, D. C.; Ho, J. X.; C.B. Anfinsen; Edsall, J. T.; Richards, F. M.; Eisenberg, D. S. In *Advances in Protein Chemistry*; Academic Press, 1994; pp. 153-176, 176a, 176b, 176c, 176d, 176e, 176f, 176g, 176h, 176i, 176j, 176k, 176l, 177-203.

31. Papadopoulou, A.; Green, R. J.; Frazier, R. A. *Journal of Agricultural and Food Chemistry* **2004,** *53*, 158-163.

32. Dangles, O.; Dufour, C.; Bret, S. *Perkin Transactions 2* **1999**, 737-744.

33. Gutzeit, H. O.; Henker, Y.; Kind, B.; Franz, A. *Biochemical and Biophysical Research Communications* **2004,** *318*, 490-495.

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

35. JiJi, R. D.; Andersson, G. G.; Karl, S. B. *Journal of Chemometrics* **2000,** *14*, 171-185.

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

37. Booksh, K. S.; Kowalski, B. R. *Analytica Chimica Acta* **1997,** *348*, 1-9.

38. JiJi, R. D.; Booksh, K. S. *Analytical Chemistry* **2000,** *72*, 718-725.

39. Morteza, B.; Rasmus, B.; Colin, S.; Abbas, A. *Journal of Chemometrics* **2006,** *20*, 99-105.

40. Rinnan, Å.; Booksh, K. S.; Bro, R. *Analytica Chimica Acta* **2005,** *537*, 349-358.

41. Rinnan, Å.; Andersen, C. M. *Chemometrics and Intelligent Laboratory Systems* **2005,** *76*, 91-99.

42. Andersen, C. M.; Bro, R. *Journal of Chemometrics* **2003,** *17*, 200-215.

43. Wentzell, P. D.; Nair, S. S.; Guy, R. D. *Analytical Chemistry* **2001,** *73*, 1408-1415.

44. Renée, D. J.; Greger, G. A.; Karl, S. B. *Journal of Chemometrics* **2000,** *14*, 171-185.

45. Andersson, G. G.; Dable, B. K.; Booksh, K. S. *Chemometrics and Intelligent Laboratory Systems* **1999,** *49*, 195-213.

46. Budavari, S., Ed. *The Merck Index*, 11th ed.; Merck & Co.: Rahway N.J., 1989.

47. Fasman, G., Ed. *Practical Handbook of Biochemistry and Molecular Biology*; CRC Press: Boston, 1992.

48. Harshman, R. A. *UCLA Working Papers in phonetics* **1970,** *16*, 1.

49. Carroll, J.; Chang, J.-J. *Psychometrika* **1970,** *35*, 283-319.

Chapter 6: Conclusions

Ultra violet resonance Raman (UVRR) is a powerful technique for determining secondary structure compositions. However, due to spectral overlap in the amide regions direct secondary structure determination has been problematic. The application of chemometric methods to UVRR spectra can improve the quality of the secondary structure determination, and allow all the data present in the spectra to be utilized in the secondary structure determination.

The application of chemometric methods has allowed determination of the pure secondary structure Raman spectrum of the α-helix secondary structure from a multi-excitation data set of the protein myoglobin by utilizing the unique spectral response with excitation wavelength with a powerful data decomposition method, multivariate curve resolution using the alternating least squares algorithm. This demonstrates the bilinear relationship between the pure secondary structure Raman spectra (PSSR) and spectral response, or cross section, and that a deconvolution of such a data set can result in the accurate determination of both the PSSRS and absolute Raman cross sections. This has laid the ground work for future studies which will incorporate both multi-excitation data, and varying secondary structure compositions, enabling a three dimensional data deconvolution, producing a unique solution.

Future multi-excitation multi-protein UVRR studies will face challenges in both aligning UVRR spectra and successfully subtracting the water band from the sample

spectra. Multi-excitation UVRR spectra are notoriously difficult to align and easily

become shifted due to increasing resolution as excitation wavelength increases. To aid in

the alignment of UVRR spectra, correlation optimized warping (COW) has been applied.

The internal intensity standard method of water band subtraction frequently over

subtracts the water band in case of low analyte concentration, resulting in negative

intensities in the amide I region of UVRR spectra. However, presented here is a new

method where the concentration of the water band is determined based on the far outlying

regions of the UVRR spectra (1750-1900 cm^{-1}) through a least squares optimization. The

resulting water band subtraction does not appear to over subtract the water band in all

cases, both high and low analyte cases.

Utilizing parallel factor analysis (PARAFAC) applied to excitation-emission

matrix (EEM) fluorescence spectroscopy, the behavior of quercetin, a model flavonoid

compound, has been thoroughly examined both in solution and in the presence of bovine

serum albumin (BSA). The application of PARAFAC has enabled the simultaneous

monitoring of three different quercetin species: the basic keto/enol form of quercetin,

higher order aggregates of quercetin (>dimer), and the BSA-quercetin complex. In

addition, the pH and concentration dependence of these species was successfully

detected.

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, meriting continued research.

VITA

John V. Simpson was born on June 22, 1982, in Saint Charles Missouri, a suburban community outside of Saint Louis. There he attended Saint Charles Senior High School, graduating in 2000. He then attended the University of Missouri – Rolla, now Missouri University of Science and Technology, graduating in 2004 with a Bachelor of Science degree in chemistry. After graduation, in 2005, he was accepted as a Ph.D. candidate in chemistry at the University of Missouri, and was the first graduate student to join the lab of Dr. Renee Jiji in the spring of 2005. He will receive his doctorate in Chemistry in September of 2009 and will pursue a post doctoral position at the National Corn to Ethanol Research Institute in Edwardsville Illinois.