# Using Machine Learning to Determine Fold Class and Secondary Structure Content from Raman Optical Activity and Raman Vibrational Spectroscopy

A thesis submitted to the University of Manchester for the degree of MPhil in the Faculty of Life Sciences

2012

**Myra Kinalwa-Nalule**

**Table of Contents**

**List of Appendices**

## List of Figures

## List of Tables

## List of Abbreviations

| | |
|---|---|
| AFM | Atomic Force Microscopy |
| CASP | Critical Assessment of Methods of Protein Structure Prediction |
| CID | Circular Intensity Difference |
| CD | Circular Dichroism |
| DALI | **d**istance matrix **ali**gnment |
| $DCP_I$ ROA | In-phase dual circular polarization Raman Optical Activity |
| $DCP_{II}$ ROA | Out-of-phase dual polarization Raman Optical Activity |
| ϵ-SVM regression | *epsilon* Support Vector Machines regression |
| FRET | Fluorescence Resonance Energy Transfer |
| FSSP | **f**amilies of **s**tructurally **s**imilar **p**roteins |
| ICP ROA | Incident Circularly Polarized Raman Optical Activity |
| k-NN | k-Nearest Neighbour |
| NMR | Nuclear Magnetic Resonance spectroscopy |
| PCA | Principal Component Analysis |
| PDB | Protein Data Bank |
| PLS | Partial Least Squares |
| PLS-DA | Partial Least Squares-Discriminant Analysis |
| PSA test | prostate specific antigen test |
| PPII | polyproline helix II |
| RBF | Radial Basis Function |
| RF | Random Forests |
| ROA | Raman Optical Activity |
| SVM | Support Vector Machines |
| SCOP | Structural Classification of Proteins |
| SCP ROA | Scattered Circular Polarized Raman Optical Activity |
| VCD | Vibrational Circular Dichroism |

**Abstract**

The objective of this project was to apply machine learning methods to determine protein secondary structure content and protein fold class from ROA and Raman vibrational spectral data. Raman and ROA are sensitive to biomolecular structure with the bands of each spectra corresponding to structural elements in proteins and when combined give a fingerprint of the protein. However, there are many bands of which little is known. There is a need, therefore, to find ways of extrapolating information from spectral bands and investigate which regions of the spectra contain the most useful structural information.

Support Vector Machines (SVM) classification and Random Forests (RF) trees classification were used to mine protein fold class information and Partial Least Squares (PLS) regression was used to determine secondary structure content of proteins. The classification methods were used to group proteins into α-helix, β-sheet, α/β and disordered fold classes. The PLS regression was used to determine percentage protein structural content from Raman and ROA spectral data.

The analyses were performed on spectral bin widths of $10cm^{-1}$ and on the spectral amide regions I, II and III. The full spectra and different combinations of the amide regions were also analysed. The SVM analyses, classification and regression, generally did not perform well. SVM classification models for example, had low Matthew Correlation Coefficient (MCC) values below 0.5 but this is better than a negative value which would indicate a random chance prediction.

The SVM regression analyses also showed very poor performances with average $R^2$ values below 0.5. $R^2$ is the Pearson's correlations coefficient and shows how well predicted and

observed structural content values correlate. An $R^2$ value 1 indicates a good correlation and therefore a good prediction model. The Partial Least Squares regression analyses yielded much improved results with very high accuracies. Analyses of full spectrum and the spectral amide regions produced high $R^2$ values of 0.8-0.9 for both ROA and Raman spectral data. This high accuracy was also seen in the analysis of the 850-1100 $cm^{-1}$ backbone region for both ROA and Raman spectra which indicates that this region could have an important contribution to protein structure analysis. $2^{nd}$ derivative Raman spectra PLS regression analysis showed very improved performance with high accuracy $R^2$ values of 0.81-0.97.

The Random Forest algorithm used here for classification showed good performance. The 2-dimensional plots used to visualise the classification clusters showed clear clusters in some analyses, for example tighter clustering was observed for amide I, amide I & III and amide I & II & III spectral regions than for amide II, amide III and amide II&III spectra analysis. The Random Forest algorithm also determines variable importance which showed spectral bins were crucial in the classification decisions. The ROA Random Forest analyses performed generally better than Raman Random Forest analyses. ROA Random Forest analyses showed 75% as the highest percentage of correctly classified proteins while Raman analyses reported 50% as the highest percentage.

The analyses presented in this thesis have shown that Raman and ROA vibrational spectral contains information about protein secondary structure and these data can be extracted using mathematical methods such as the machine learning techniques presented here. The machine learning methods applied in this project were used to mine information about protein secondary structure and the work presented here demonstrated that these techniques are useful and could be powerful tools in the determination protein structure from spectral data.

## Declaration

No part of this work presented in this thesis has been submitted for support for another degree at the University of Manchester or any other research or learning institute.

Parts of the work submitted in this thesis have been published in the sources listed below and the published papers are in Appendix L.

**Accurate Determination of Protein Secondary Structure Content from Raman and Raman Optical Activity Spectra.**
Kinalwa M., Blanch E. W. and Doig A. J.
Analytical Chemistry, 2010, 82, 6463-6471

**Determination of Protein Fold Class from Raman or Raman Optical Activity Spectra using Random Forest**
Kinalwa M., Blanch E. W. and Doig A. J.
Protein Science,2010, 20, 1668–1674

**Copyright Statement**

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "copyright") and he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made **only** in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publications and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of School of Faculty of Life Sciences (or the Vice-President) and the Dean of the Faculty of Life Sciences, for Faculty of Life Sciences' candidates.

**Acknowledgements**

I would like, first of all, to thank my mother, Thecla Kinalwa, who has supported me throughout this degree and without whom this opportunity would not have been possible.

Secondly, I would like to thank my supervisors Dr. Ewan Blanch and Professor Andrew Doig, who have also been my teachers, critics and mentors throughout these four years and for whose support I will eternally be grateful.

Special thanks also go to Dr. Jo Avis, my advisor, for her advice and whose support I knew I could always count on.

Last but not least, I would like to extend my gratitude to Professor L. D. Barron and Dr. L. Hecht at the Department of Chemistry of the University of Glasgow for provision of Raman and ROA spectra.

# 1. Protein Structure

## 1.1 Introduction

The main objectives of the project were to apply machine learning analytical techniques to determine protein structural content and protein fold class from Raman scattering and Raman Optical Activity (ROA) vibrational spectra. Vibrational spectroscopy is an umbrella group of methods which can be used to access protein structural information. ROA and Raman spectroscopy are discussed in Chapter 2. The analyses in this project were performed on spectra from Raman scattering and Raman Optical Activity. Raman and ROA spectroscopies are complementary to each other. Raman spectroscopy is an insightful probe into the side chains and structure of proteins. ROA is sensitive to the most chiral, and rigid, parts of the protein. It is for this reason that ROA spectra are dominated by bands from the peptide backbone and so give more direct information about secondary and tertiary structure (Barron et.al.,1992, 1992, 2003). Generally in Raman and ROA vibrational spectroscopies, laser light irradiates the sample and photons are scattered. When this happens, there is a transfer of energy from the light photons to the molecule causing vibrational excitation. This results in Stokes scattering where the frequency of the scattered light is at a lower frequency than the incident frequency. It is the intensities and the frequencies of the scattered light that are measured to give information about the structure of the protein. Raman and ROA are sensitive to biomolecular structure with the bands of each spectrum corresponding to structural elements in proteins and when combined these give a spectral fingerprint of the protein. However, there are many bands of which little is known. There is also a problem in the analysis of overlapping bands, a common accurrence in vibrational spectra. The commonly used method for analysing overlapping bands is curve fitting which finds the best

fit of component bands within a band envelope from a number of possible solutions. However the solution depends on the given number of bands and band shapes and the solution may not be the best fit over each component band (Kaoshi and Ozaki, 2003). There is a need, therefore, to find ways of extrapolating information from spectral bands and investigate which regions of the spectra contain the most useful structural information. This project investigated a range of computational methods such as Support Vector Machines (SVMs), regression and decision trees to derive information about protein structure from Raman and ROA spectra. This involved analysing how much information about the structure of proteins could be extracted from spectral data by each of these methods, and examining how much information can be obtained by these methods about the structural hierarchy of proteins.

Whilst vibrational spectroscopic methods do not usually provide information at atomic resolution, there is no size limit to the proteins that can be studied and molecules can be investigated in more physiologically relevant conditions i.e. in solution, which is vital as protein structure can be studied in a state that is as close as possible to that of the natural state of the biomolecules *in vivo* . In addition, vibrational spectra can be used to study proteins whose structures cannot be resolved by more common methods. For instance, many proteins are not easily crystallisable and therefore cannot have their structures solved by X-ray chrystallography. This project aimed to use machine learning methods to mine for structural information, which could be useful in the development of new techniques to determine protein structure from spectral data.

## 1.2 Protein structure

The 3-dimensional structure of proteins is usually a requirement for protein function. The 3-dimensional structure of a protein exhibits four hierarchical levels of structure: primary, secondary, tertiary and quarternary. The genetically determined linear composition of amino acid units linked by peptide bonds is the primary structure. The secondary structure is made up of amino acid residues linked by hydrogen bonds and constitutes mainly α-helices and β-sheets. The other elements of secondary structure include β-turns and unordered structure. β-turns are sharp turns that connect the adjacent strands in an anti-parallel β-sheet, however, these can be found elsewhere. Unordered structure is generally a catch-all term for regions that do not fall into one of the other categories. Tertiary structure arises when various elements of secondary structure pack tightly together to form the well-defined 3-dimensional structure. The tertiary structure is held together by favourable interactions between the side chains. Quaternary structure is the association of tertiary subunits to form a globular protein with multiple components. Vibrational spectroscopy mechanisms produce spectral bands that are sensitive to the different structural elements of proteins. Understanding protein structure is very important because the behaviour and function of the protein is intrinsically linked to its structure and so studying structure is an important step towards understanding biological processes.

### 1.2.1 Primary, secondary and tertiary protein structure

The backbone of a polypeptide chain comprises the amide nitrogen (NH), the alpha carbon ($C_\alpha$) and the carbonyl carbon (CO) that are contributed by each amino acid unit (Figure 1.1). The side chains comprise the "R" groups, and are bonded to the backbone. Proteins are uniquely identified by the number of amino acids they contain and the sequential order of the amino acids. The primary structure is therefore a linear composition of amino acid units

linked by peptide bonds. The peptide bond is formed by a hydrolysis reaction between the NH of one amino acid residue and the CO of the neighbouring amino acid. The carboxyl bond of the peptide bond partially delocalizes electrons creating a partial double bond character which makes the bond more rigid and stronger. This also means that the three non-hydrogen atoms that participate in the bond (the carbonyl oxygen O, the carbonyl carbon C, and the amide nitrogen N) are co-planar and that free rotation about this bond is limited. The only degrees of freedom for the peptide units are the rotations about the $C_\alpha$-C and N-$C_\alpha$ on either side of the peptide bond. The angle of rotation around the $C_\alpha$-C bond is called the psi (ψ) torsion angle and the angle around the N- $C_\alpha$ bond is called the phi (φ) torsion angle (Figure 1.1). As these are the only degrees of freedom, the possible number of conformations the polypeptide can adopt is restricted. In a polypeptide chain, the N-Cα and the $C_\alpha$-C bonds have a restricted freedom of allowed rotation. These rotations are represented by the phi ϕ and psi ψ torsion angles. The Ramachandran plot was developed by Ramachandran et. al. (1963) and is used to show the possible stable conformations which disallow steric clashes in protein structural elements.

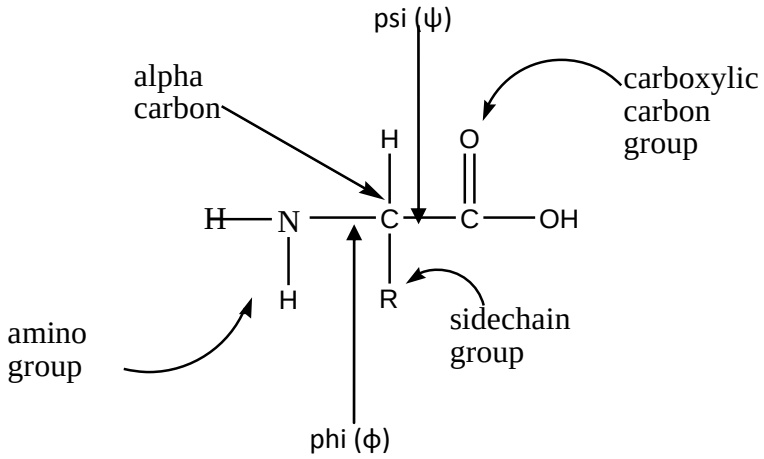



**Figure 1.1 The general structure of an amino acid.**

The amide bonds are covalent bonds that hold proteins together. Many proteins also form disulphide bridges between the side chains of cysteine residues. The rest of the stabilizing forces in a protein are provided by noncovalent interactions for example van der Waals interactions and hydrogen bonds.

Van der Waals forces are weak interactions between groups of atoms arising from fluctuations in electron distributions around the nuclei of the atoms. The fluctuating electron clouds around one atomic nucleus induce an opposite fluctuating dipole in a non-bonded neighbouring atom. The van der Waals interactions diminish as the interacting species get further apart, so only atoms that are within 5Å or less of each other can participate in such interactions. Van der Waals interactions are weak but collectively make a significant energetic contribution to holding the protein structure together (Petsko and Ringe, 2004).

Hydrogen bonds are formed when a hydrogen atom acquires a partial positive charge because it is covalently bonded to a more electronegative atom (the donor atom) and is attracted to a neighbouring atom with a partial negative charge (the acceptor atom). The charge delocalization leads to a local dipole moment that lead to hydrogen bonding. This non-covalent bond between the partially charged atoms draws the non-hydrogen atoms closer.

Most known proteins fold up into globular structures with hydrophobic residues tightly packed into a core away from contact with water creating the tertiary structure (Branden and Tooze,1999). There are three common and well recognised types of secondary structure elements namely; α-helix, β-sheets and β-turns. Secondary structure contributes to the overall stability of the protein. Helices and β-sheets consist of extensive networks of hydrogen bonds. The hydrogen bonding in these elements provide much of the stabilization that allows the formation of the hydrophobic core (Branden and Tooze, 1999). The structural assembly of more than one tertiary structural subunit forms the quaternary structure of proteins.

### 1.2.2 The helix structure

Alpha helices are formed when a stretch of consecutive residues have φ, ψ angles of approximately -60˚ and -50˚ respectively. In α-helix, the carbonyl oxygen of residue *i* forms a hydrogen bond with the amide nitrogen four residues (*i*+4) further along. All of the NH and CO groups are, therefore, joined by hydrogen bonds apart from the first NH groups and the last CO groups at the ends of the α-helix. This bonding results in a helical structure with 3.6 residues per turn and protruding side chains on the outside.

In globular proteins, alpha helices vary in length from five residues to over forty residues. The average length is about ten residues (Petsko and Ringe, 2004). An average helix is therefore 15Å in length which corresponds to the 1.5Å rise per residue along the helical axis. Helical structures can be either right-handed or left-handed. The left-handed α-helix is not allowed because of steric restrictions on the conformation of the L-amino acids. Hence, the α-helix that is observed in proteins is always right-handed. With 3.6 residues per turn there is a tendency of amino acids to vary from hydrophobic to hydrophilic for every three to four residues. This creates an α-helix that has a hydrophobic face and a hydrophilic face; such a helix is called an amphipathic helix. The most common location of this kind of α-helix is along the outside of a protein, with one side facing the interior/hydrophobic core and the other facing the solvent (Drin et.al.,2007; Hristova et.al, 1999; Segrest et.al., 1990, 1999).

There are variations of the α-helix with hydrogen bonds to residues i+5 or *i*+3, instead of the standard *i*+4, called the π helix and the $3_{10}$ helix, respectively. The $3_{10}$ helix has 3 residues per turn while the π helix has 4.4 residues per turn. Both the π helix and the $3_{10}$ helix occur rarely in proteins and usually at the ends of alpha helices. They are not energetically favourable

since the backbone atoms are too tightly packed in the $3_{10}$ helix and too loosely packed in the π helix (Petsko and Ringe, 2004).

### 1.2.3 The β-sheet structure

The second major structural element found in globular proteins is the β-sheet. In proteins, β-strands that are widely distant from each other in the protein sequence may be aligned adjacent to each other such that hydrogen bonds are formed between them. The hydrogen bonds are formed between the CO groups of one β-strand and the NH on an adjacent β-strand. β-sheets are either parallel, the strands all running in the same direction from amino to carboxyl termini, or anti-parallel, in which each successive strand runs in the opposing direction, amino terminal to carboxyl terminal followed by carboxyl terminal to amino terminal. β-sheets can also be mixed sheets with both parallel and anti-parallel sheets (Lesk, 2001). Each of the two forms has a distinctive pattern of hydrogen bonding. The anti-parallel sheets have narrowly spaced hydrogen bonds that alternate with widely spaced hydrogen bonds. Parallel β-sheets have evenly spaced hydrogen bonds. Within both types, all possible hydrogen bonds are formed, except for the two flanking β-strands of the β-sheets that each have only one neighbouring β-strand. The edge strands may make hydrogen bonds with water if they are exposed or pack against polar side chains. The sheet may curve round on itself to form a barrel structure (referred to as β-barrel structure), with the two edge strands hydrogen bonding to one another to complete the closed cylinder (Hames and Hooper,2000). Antiparallel sheets are commonly connected by β-turns while parallel sheet strands are commonly connected by α-helix that packs against a face of the β-sheet (Petsko and Ringe, 2004; Branden and Tooze, 1999).

The simplest secondary β-structure element is the β-turn which involves four residues. A β-turn is formed when the carbonyl oxygen of one residue (*i*) makes a hydrogen bond to the amide hydrogen of residue (*i*+3) reversing the chain direction. This secondary structure element is also called a reverse turn or a hairpin turn. β-turns are usually found on the surfaces of folded proteins, where they are in contact with an aqueous environment. By reversing the direction of the chain they limit the size of the molecule making it more compact. This motif is very frequent in β-sheets between two antiparallel β-strands (Finkelstein and Ptitsyn,2002).

### 1.2.4 Disordered structure

Traditionally, it had been widely accepted that disordered proteins would be degraded by proteases *in vivo*. However, Dunker *et al.* (2002) highlighted the existence of intrinsic disorder *in vivo* and mechanisms that help disordered proteins from being eliminated. Disordered regions are usually protected from proteases by steric configurations that render then inaccessible to proteases; other disordered regions have residues that are resistant to proteolytic degradation; certain disordered regions exist transiently changing binding partners constantly. Comparison of evolutionary rates between ordered and disordered regions within the same protein showed faster rates of evolution in the disordered regions. The residues within the functional regions were highly conserved. The faster rates of evolution in the disordered regions suggest that the disordered regions lack stable side chain interactions and this has been cited as evidence for the existence of disorder *in vivo by* Dunker *et.al.* (2002). The same group also found that disordered proteins were involved in molecular recognition in protein binding interactions and that they are also found at protein modification sites. Disordered regions are known to be functionally important and varied in structure. Uversky (2002) suggested the existence of the pre-molten state in which the secondary structure in the

protein molecules is less compact than in native or in the molten globule protein. The molten globule is described by Finkelstein and Ptitsyn (2002) as an intermediate of protein folding and unfolding, less compact than the native protein with secondary structure that is nearly as developed as that of the native protein but very little ordering of its side chains. Neural Networks were used by Romeno *et. al.* (1997) to predict disordered regions, showing 70% prediction accuracy on disordered regions of short, medium and long length and also showed that these regions were compositionally different from each other. Radivojac *et. al.* (2007) revealed that disordered proteins could be distinctly identified from ordered proteins using bioinformatics analysis by using the distinct sequences found in the disordered regions and their preferences for specific amino acid compositions. Intrinsically disordered proteins have been shown to play important functional roles in regulation, cell signalling and control pathways (Uversky *et.al.,* 2005; Iakouvcheva *et. al.*, 2002; Wright *et. al.*, 1999). NACP a protein reported to confer protection against Alzheimer's apo E4 allele, has been shown to be natively unfolded by Weinreib *et. al*. (1996). Intrinsically disordered proteins were grouped by Dunker *et. al.* (2002) according to their functions, into molecular recognition, molecular assembly, protein modification and entropic chain activities. This recent research into disordered structure of proteins has challenged the dominant theory that the three-dimensional structure of a protein is the requirement for the protein to carry out its functions. Like the 3-dimensional protein structure, the disordered structure appears to play an important functional role as well.

Another related secondary structure is the polyproline helix II (PPII) structure. The polyproline helix II has phi and psi angles of -75 degrees and 145 degrees on the Ramachandran plot respectively (Stapley and Creamer, 1999). The PPII structure tends to exist on the surface of the protein molecule where it can form hydrogen bonds with the

solvent but form few main hydrogen bonds with the rest of the protein (Adzhubei and Sternberg, 1993). Deber and Rath (2005) suggested PPII plays a role in protein recognition functions and in protein unfolded states.

### 1.2.5 Protein Motifs

Secondary structure elements are connected by segments such as reverse turns to form structural motifs. Here, the term motif refers to a set of contiguous secondary structure elements that either have a functional significance or define part of a domain. The simplest motif consists of two α-helices joined by a loop region, which is called the helix-turn-helix motif and is specific for DNA and calcium binding (Branden and Tooze, 1999). The secondary structure elements are connected to form motifs which are associated with specific functions. Some of these motifs are discussed here below:

**EF hand motif**

The EF hand motif is formed by two α-helices that are connected by a short loop and is associated with calcium binding in domains that regulate cellular activity, for example in the proteins calmodulin (Taylor *et.al.*,1991) and muscle protein parvalbumin (Moews and Kretsinger, 1975). The loop region provides the right conformation to bind the calcium ion. The EF motif is also found in DNA binding proteins for example catabolite gene activating protein (CAP), which binds to DNA to activate transcription (Branden and Tooze, 1999). The helix-turn-helix motif provides a scaffold that holds the substrate in the proper position to bind and release calcium (Branden and Tooze, 1999).

**The β-α-β motif**

The β-α-β motif is frequently used to connect two parallel β-sheets and is found as part of almost every protein structure that has a parallel β-sheet. The α-helix in the β-α-β motif

connects the carboxyl end of one strand with the amino group of the next β-strand and is usually oriented such that the helical axis is approximately parallel to the β-strands with the α-helix packing against the β-strands. The β-strands and the α-helix are connected by loop regions. The loop that connects the carboxyl end of the β-strand to the amino end of that α-helix is often involved in the active site of a protein (Argos *et. al.*, 1997).

α/β Structures are built up with the beta-alpha-beta motif (Reardon and Farber 1995; Petsko and Farber 1990; Farber 1993). There are three main classes of α/β proteins. In the first class, the core is made up of parallel β-strands and the α-helices that connect them are on the outside of the barrel forming the TIM barrel structure (Banner *et.al.*, 1975; Lesk, 2001). The second class consists of an open β-sheet with the parallel strands laid out adjacent to each other surrounded by α-helices on both sides. This fold is often referred to as the Rossman fold. The third class is formed from amino acid sequences with repetitive leucine regions along the sequence. These leucine-rich regions form α-helices and β-strands. The β-strands form an approximate semi-oval curved parallel sheet with all the α-helices outside the curved face (Branden and Tooze, 1999; Ptitsyn and Finkelstein, 2002).

Several secondary structural elements and motifs in a single polypeptide chain combine to form one or several compact domains. This constitutes the tertiary structure of the protein. The fundamental unit of the tertiary structure is a domain. A domain is defined as a polypeptide chain or a part of a polypeptide chain that can fold independently into a stable tertiary structure. The α-helices and β-strands are adjacent to each other in the three dimensional globular structure and are connected by loop regions (Thornton *et.al.*, 1988). Efimov (1993) divides loop regions into four types classified according to the secondary structures they connect; αα, αβ, βα and ββ. These can be further subdivided according to the positions adopted by the α-helices and β-strands and their overall structure. The number of

combinations of different motifs is quite limited in proteins suggesting that some combinations are structurally favoured (Petsko and Ringe, 2004). It is because of these structurally favoured motifs that similar domain structures occur in different proteins with different functions and different amino acid sequences.

The explosion in the amount of sequence data and the number of proteins whose structures have been resolved has led to the need to store these data. Numerous protein databases have been established with different techniques to group proteins into structural classes as a way of understanding the structure and functions of proteins.

### 1.3 Classification of proteins and protein databases

Understanding protein structure can be improved by comparing similar folds to new structures in order to assign new fold families. Amino acid sequences with high similarity infer homology and an evolutionary relationship and this can be extended to similar 3-dimensional folds (Holm *et. al*., 1992; Orengo *et. al*., 1993; Holm and Sander, 1994; Pascella and Argos, 1992). However, for proteins that lack a significant sequence similarity, having a similar fold does not always imply evolutionary relationship (Orengo *et.al*., 1994; Thorton *et.al.*, 1999; Dolittle 1994). Sequence motifs or fingerprints which characterise short regions of amino acids that relate to function or structural constraints are also used in fold comparison databases to identify new folds. Another method is the comparison of 3-dimensional models to sequences whose structure is unknown. Protein repositories allow for new experimental structures to be identified by comparative methods developed by different groups discussed below.

## CATH

The CATH criteria for protein classification uses sequence alignments of the resolved structures (Orengo *et al*., 1997; 2002a; 2002b) in the Protein Data Bank (PDB) database (Holm *et. al*., 1992). The four main levels of classification used are Class, Architecture, Topology and Homologous superfamily. Global similarities suggest evolutionary relationships and structural similarity is applied in the assignment of families if there is a lack of an evolutionary relationship (Greene *et. al*., 2007; Pearl *et. al*., 2002; Pearl *et.al*., 2003). Multidomain proteins are broken down into their constituent domain folds. Sequence alignment groups sequences into protein classes based on secondary structure composition i.e. the amount of αβ and α-β structure. Architecture defines the shape of the protein by the assembly of the secondary structure elements without taking into consideration their connections. The architecture is assigned manually with reference to the literature. The topology level groups proteins assembled with the same arrangement and number of structure elements as well as the same connecting links. At the homologous superfamily level, proteins are clustered by structural and functional similarity. CATH also classifies proteins into sequences families by grouping proteins with >35% similarity.

## PFAM

The PFAM database uses protein family profiles built from a sequence set representative of a protein family with the hidden Markov model algorithm to search sequence databases for homologues. The sequences that align against the profile with a score above a given threshold are added to the full alignment representative of the protein family. The database is based on UniProt and the recent release contains 12273 families (Finn *et. al.*, 2010). The FSSP (families of structurally similar proteins, Holm and Sander, 1996) database is a collection of 2860 entries of 3-dimensional fold classification based on structural alignments of known protein chains from the PDB database. The sequences are aligned to a family of

representatives using the DALI (distance matrix alignment) algorithms with a cut-off threshold of less than 30% sequence identity. The DALI algorithm uses the 3-dimensional coordinates of each protein to calculate the inter-residue distance matrices between structure elements of the proteins to make alignments. The resulting alignment of sequence families is then further hierarchically clustered, grouping proteins with strong sequence similarity into fold classes and those with similar topography but low evolutionary similarity in the lower level of the tree (Holm, 2010).

**Structural Classification of Proteins (SCOP) database**

SCOP (Structural Classification of Proteins, Murzin *et. al.*, 1995) is a protein structural database that was built to facilitate the understanding of, and provide access to, the growing amount of information on resolved protein structures. It provides an extensive description of the structural and evolutionary relationships of proteins whose three dimensional structures have been solved. The classification of protein structures in the database is based on evolutionary relationships and on the principles that govern their three dimensional structure. It includes proteins that are in the current PDB database. As of 2011, there are 74,140 PDB entries in the database. SCOP uses visual inspection and automatic tools to speed up the process of classification. The unit of classification is the protein domain. Small proteins are treated as a whole, while domains in large proteins are classified individually.

The classification is based on hierarchical levels that embody the evolutionary and structural relationships. Proteins are clustered together into families on the basis of one of the two criteria that imply their having a common evolutionary origin: the proteins that have residue identities of 30% and greater; the proteins with lower sequence identities but whose functions and structures are very similar, for example globins, with sequence identities of 15% or higher. Proteins classed under one superfamily are those with low sequence identity but

whose structures and in many cases functional features suggest a common evolutionary origin. Proteins with a common fold are superfamilies and families with the same secondary structures in the same arrangement with the same topological connection links. SCOP uses unique identifiers to represent each entry in the SCOP hierarchical levels and is also integrated with data about protein families from other databases like Pfam, CATH and InterPro (Conte *et.al*., 2002; Andreeva *et.al.,* 2004).

Folds in the SCOP database have been grouped into five classes: all α, proteins whose structures are essentially formed by α-helices; all β, proteins whose structures are essentially formed by β-sheets; α + β, proteins in which the α-helices and β-strands are largely segregated; α and β, for those proteins in which the α-helices and the β-strands are largely interspersed; multi-domains, for proteins with domains of different folds and for which there are no known homologues (Murzin *et. al.*, 1995).

The SCOP database was created as a tool for understanding evolutionary relationships based on the structures of proteins. The classifications also show how diverse protein structures are. Understanding protein structure is very important because the behaviour of a protein is determined by its structure and so studying structure can help us learn more about protein function and misfunction. Hegyi and Gerstein (1999) investigated the relationship between the structure and function of proteins by comparing functionally characterized enzymes in SwissProt with structurally characterised domains in SCOP. This study showed that α/β folds tend to be enzymes and all α-proteins tend to be non-enzymes.

These repositories facilitate the progress of structural biology making information on structures of proteins and complex assemblies accessible. Various methods have been used to study and determine protein structures.

## 1.4 Determination of protein structure

Different methods have been developed and used by different groups to predict or solve the structures of proteins. Commonly used structural methods like X-ray crystallography and NMR are not always applicable to all proteins, as discussed above, and so it is important to develop other resources to study protein structure. Different methods have been used in the study of protein structure. The methods used vary from laboratory-based methods to computational methods. A variation in these methods used in the determination of protein structures allows for wider number of structures to be studied, for example, methods such as X-ray crystallography require the protein under investigation to be crystallized; in cases where this is not achievable, other methods can be applied.

X-ray crystallography is the most widely used technique for structural studies and the source of most entries in the Protein data Bank (PDB) (Sussman *et. al*, 1998). In X-ray crystallography an x-ray beam irradiates the protein crystal and the regular lattice spacing of the protein crystal diffracts the x-rays. The diffracted x-rays produce reflections at specific points on an x-ray detector. The pattern of diffracted spots is then used to calculate the dimensions of the protein (Rupp, 2009). X-ray crystallography provides high level structural detail and can be used on large proteins of at least 150,000 subunit molecular weight. However, it only works with crystals whose creation is a complex process (Weber, 1997; Rhodes, 2006) and there is also the phase problem where the phase of the x-ray wave is lost in the measurement. Another problem with using x-rays is the damage that can be caused to biomolecules by the free radicals released from the excitation of atoms which can cause denaturation and cleavage of polypeptides.

Nuclear magnetic resonance spectroscopy (NMR) is also widely used to investigate the structures of proteins in solution (Blumich, 1994). It is non-invasive but it has an upper size

limit of ~60k Da. and it requires the use of radioactive labelled isotopes of atoms such as $^{2}H$, $^{13}C$, $^{15}N$ which have a magnetic moment (Sattler and Feslik, 1996). NMR involves placing active nucleii in a magnetic field which excites the nucleii causing them to behave like bar magnets. The nucleii align along or against the direction of magnetic field resonating between the spin field of the nucleii and the external magnetic field. The resonance frequency depends on the nature of the nucleus, the strength of the external magnetic field and the local environment of the atom. Different protons in a molecule in different chemical environments give rise to different chemical shifts. The chemical shift is measured as a change in frequency relative to a reference frequency. Each proton's identity can be assigned from the its respective frequency resonance.

The properties of flurophores for example tyrosine, are also used to study proteins. Fluorescence Resonance Energy Transfer (FRET) for example, can be used to study conformational changes and protein-protein interactions using fluorophore tagged proteins (Periasamy and Day, 2005). FRET involves the transfer of energy between two fluorophores provided they have overlapping absorption and flourescence profiles.

Another method which can be used with crystalline and non-crystalline samples is electron microscopy (Goodhew *et.al*., 2001). Electrons at high velocity in a vacuum are focused, using magnets, on to a specimen. The electrons are scattered by the electronic clouds around the atoms of the moelcules. Electron microscopy can be used with very large complexes, it doesn't necessarily need crystals and there is no phase problem. However the electrons that bombard the specimen and the vacuum can destroy the samples.

Atomic Force Microscopy (AFM) is another method used to study the structure of proteins (Morris *et.al*., 1999; Jena *et.al*.,2002). In AFM the protein is bound to a tip at the end of the cantilever that is moved over the surface of the specimen as it measures the forces of

interaction holding the protein structure. A picture of the sample can be built by mapping the interaction between the probe tip and the sample as in AFM imaging. One disadvantage of the AFM method is the tip can easily destroy the sample. Vibrational spectroscopies, like Raman and ROA that are discused in chapter 2 are non invasive and are applicable to a wide range of proteins and experimental conditions.

Inference of homology from sequence similarity has become routine in protein structural studies. When two protein sequences are homologous, there is a possibility be that they are both descendants from a common ancestor or they converged from different origins to have similar functional or structural constraints. The challenge is to determine accurately homologs given the different evolutionary relationships between proteins. An evaluation of different search algorithms by Pearson and Sierk (2005) showed that the percentage accuracy of the sequence searches for homolog matches varied.

Computational methods have also been developed to investigate the structures of proteins. Comparative modelling is based on sequence matches through homology. In comparative modelling the new target sequence is compared to other sequence databases. Any matched sequence is scanned against a database of known structures and a template built. The key part of comparative modelling is the quality of the sequence alignment. Loop regions are tidied up by looking for the best fit from a database. In the absence of a previously identified target, *ab initio* physics-based approaches are used to calculate energy potentials and simulations (Leach, 2001). Another related method, threading (Jones *et.al*., 1992), uses a combination of *ab initio* and comparative modelling. Threading involves a search against the database folds and susbsequently ranks the folds by calculated energies.

Critical Assessment of methods of Protein Structure Prediction (CASP) is a community-wide protein prediction exercise that aims to assess the various methods for protein fold prediction

that has run every two years since 1994 (Moult, 2005). Structural targets, gathered from structural groups, are distributed amongst theoretical groups for prediction using *ab initio*, comparative modelling and threading types of methods. The predictions are then assessed and the results published. Recent observations stipulate that challenges still faced by template-based modelling include the identification of correct template, a step that is dependent on the generation of correct alignments. Another bottleneck is the large search space for any particular protein target (Xu *et. al*., 2009). Template free model prediction using *ab initio* prediction of 3-dimensional structures is still not reliable (Ben-David, 2009).

The problems mentioned above makes it useful to investigate other methods to complement efforts in studying protein structure. Vibrational spectroscopy is another important and useful method for probing the structures of protein. Vibrational spectroscopy methods are discussed further in chapter 2. Spectral data have previously been shown to contain information about secondary structure such that the data can be clustered by Principal Component Analysis (PCA) into their structural classes e.g. α-helix, β-sheet, α/β and disordered (Barron *et al*.,2003). Our analyses looked at deriving fold class and structural content information from data obtained from the vibrational spectroscopy techniques namely: Raman spectroscopy and Raman Optical Activity (ROA). The protein assignments, using SCOP as a reference, used in the analyses were of four classes; α-helix, β-sheet, αβ and other. Using SVM regression and classification, PLS regression and Random Forest classification we have been able to extract structural fold class and quantitative structural fractions of α-helix, β-sheet, αβ and other from the Raman and ROA spectral data of proteins. The results of our analyses are discussed in chapters 5, 6, 7 of the thesis.

## 1.5 References


Adzhubei A.A., Sternberf M.J.,(1993) Left-handed polyproline II helices commonly occur in globular proteins. *Journal of Molecular Biology* **229**, 472-493

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic. Acid Research* **32**, 226-229

Argos P., Rossman M.G., Johnson J.E. (1997)A four-helical super-secondary structure. *Biochemical and Biopphysical Research* Communications **75**, 83-86

Banner D.W., Bloomer A.C.,(1975) Structure of chicken muscle triose phosphate isomerise determined crystallographically at 2.5A resolution using amino acid sequence data. *Nature* **255**, 609-614

Barron L.D., Blanch E.W., McColl I.H., Syne C.D., Hecht L., Nielsen K. (2003) Structure and behaviour of proteins, nucleic acids and viruses from vibrational Raman optical activity. *Spectroscopy* **17**, 101-126.

Barron L.D., Cooper A., Ford S.J., Hecht L., Wen Z.Q. (1992) Vibrational Raman optical activity of enzymes. *Faraday Discussions* **93**, 259-268.

Barron L.D., Wen Z.Q., Hecht L. (1992) Vibrational Raman optical activity of proteins. *Journal of American Chemical Society* **114**, 784-786

Ben-David M., Noivirt-Brik O., Paz A., Prilusky J., Sussman J.L., Levy Y. (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins* **77**, 50-65

Brandon C. and Tooze J.( 1999) *Introduction to Protein Structure.* 2nd ed., Garland Publishing:New York

Blanch E.W., McColl I.H.,Hecht L.,Nielsen K., Barron L.D. (2004) Structural characterization of proteins and viruses using Raman optical activity. *Vibrational Spectroscopy* **35**, 87-92

Blumich B.,(1994) Methods and applications of solid-state NMR. Springer-Verlag, Berlin

Davidson R.A., Deber A.R. (2005) The structure of unstructures regions in the peptides and the proteins:role of polyproline II helix in protein folding and recognition. *Biopolymers* **80**, 179-185

Doolittle R.F. (1994) Convergent evolution: The need to be explicit. *Trends in Biochemical Sciences* **214**, 149-159

Drin G., Casella J.-F., Gautier R., Boehmer T., Schwartz T.U., Antonny B. (2007) A general amphipathic α-helical motif for sensing membrane curvature. *Nature Structural and Molecular Biology* **14**, 138-146

Dunker A.K, Brown C.J., Lawson J.D., Iakoucheva L.M., Obradovic Z. (2002) Intrinsic disorder and protein function. *Biochemistry* **41**, 6573-6582

Dunker A.K., Brown C.J., Obradovic Z.,(2002) Identification and functions of usefully disordered proteins. *Advances in Protein Chemistry* **62**, 25-49

Efimov A.F. (1993) Patterns of loops regions in proteins. *Current Opinion in Structural Biology* **3**, 379-384

Farber G., Petsko G. (1990) The evolutuon of α/β barrel enzymes. *Trends in Biochemical Sciences* **15**, 228-234

Farber G. (1993) An α/β barrel full of evolutionary trouble. *Current Opinion in Structural Biology* **3**, 409-412

Finkelstein A.V., Ptitsyn B.O. (2002) Protein Physics: A course of lectures, Academic Press London, UK

Finn R.D., Mistry J. , Tate J., Coggill P., Heger A., Pollington J. E., Gavin O.L., Gunasekaran P., Ceric G, Forslund K., Holm L., Sonnhammer E. L. L, Eddy S.R.,Bateman A. (2010) The Pfam protein families database. *Nucleic Acids Research* **38**, 211-222

Goodhew P.J., Humphreys J., Beanland R. (2001) *Electron microscopy and analysis*. 3rd ed., Taylor and Francis Inc, London

Greene L.H., Lewis T.E., Addou S., Cuff A., Dallman T., Dibley M., Redfern O., Pearl F., Nambudiry R., Reid A., Sillitoe I., Yeats C., Thornton J.M., Orengo C.A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research* **35**, 291-297

Hames B.D., Hooper N.M.(2000) Instant Notes Biochemistry, 2nd ed., BIOS Scientific Publishers Limited, Oxford

Heygi H., Gerstein M. (1999) The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *Journal of Molecular Biol*ogy **288**, 147-164

Holm L., Ouzounis C., Sander C., Tuparev G., Vriend G. (1992). A database of protein structure families with common folding motifs. *Protein Science* **1**, 1691-1698

Holm M., Sander C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research* **22**, 3600-3609

Holm M., Sander C. (1996) Dali/FSSP classification of three-dimensional folds. *Nucleic Acids Research* **24**, 206-210

Holm L., Rosenstroem P. (2010) Dali Server: conservation mapping in 3D. *Nucleic Acids Research* **38**, 545-549

Hristova K., Wimley W.C., Mishra V.K., Anantharamiah G.M., Segrest J. P., White S.H. (1999) An amphipathic α-Helix at a membrane interface: a structural study using a novel x-ray diffraction method. *Journal of Molecular Biology* **290**, 99-117

Iakoucheva L.M., Brown C.J., Lawson J.D., Obradovic Z., Dunker A. K. (2002) Intrinsic disorder in cell-signalling and cancer-associated proteins. *Journal of Molecular Biology* **323**, 573-584

Jena B.P., Hoerber Heinrich J.K. (2002) Atomic Force Microscopy in Cell Biology. Academic Press, London

Jones D.T., Taylor W.R., THorton J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89

Koashi K. and Ozaki Y. (2003) Analysis of overlapping bands by an analytic geometric approach: The potential of band-stripping and the complementary matching method in extraction of band components from overlapping band. *Applied Spectroscopy* **57**, 1528-1538

Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acid Research* **30**, 264-267

Moews P.C. and Kretsinger R.H. (1975) Refinement of he structure of carp muscle calcium-binding parvalbumin by model building and difference fourier analysis. *Journal of Molecular Biology* **91**, 201-228

Morris V.J., Kirby A.R., Gunning A.P. (1999) Atomic force microscopy for biologists.Imperial College Press, London

Moult J. (2005) A decade of CASP progress; bottlenecks and prognosis in protein structure prediction. Current Opinion in Structural Biology 15, 285-289.

Murzin A.G., Brenner S. E., Hubbard T., Chothia C. (1995) SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biol*ogy **247**, 536-540

Leach A.R. (2001) Molecular modelling: Principles and applications, 2nd ed., Pearson Eduaction

Lesk A.M. (2001) Introdcution to protein structure: the structural biology of proteins, Oxford University Press Inc., New York

Orengo C.A., Flores T.P., Taylor W.R., Thorton J.M.(1993*)* Identifying and classifying protein fold families. *Protein Engineering* **6**, 485-500

Orengo C.A., Jones D.T., Thornton J.M.(1994) Protein superfamilies and domains superfolds. *Nature* **372**, 631-634

Orengo C.A., Bray J.E., Buchan D.W.A., Harrison,A., Lee D., Pearl F., Sillitoe I., Todd A.E., Thornton J.M. (2002a). The CATH protein family database: A resource for structural an functional annotation of genomes. *Proteomics* **2**, 11-21

Orengo C.A., Pearl F., Thornton J.M. (2002b). The CATH domain structure database. *Methods of Biochemical Analysis* **44**, 249-271

Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thorton J.M. (1997) CATH-a hierarchic classfication of protein domain structures. *Structure* **5**, 1093-1108

Pascarella S., Argos P. (1992) A data bank merging related protein structures and sequences *Protein Engineering* **5**,121-137

Pearl F., Lee D., Bray J.E., Buchan D.W.A., Shepherd A.J., Orengo C.A. (2002). The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Science* **11**, 233-244

Pearl F.M., Bennett C.F., Bray J.E., Harrison A.P., Martin N., Shepherd A., Sillitoe I., Thornto J., Orengo C.A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids* Research **31**, 452-455

Pearson W.R., Sierk M. (2005) The limits of protein sequence comparison. *Current Opinion in Structural Biology* **15**, 254-260

Periasamy A., Day R. N. (2005) Molecular imaging: FRET microscopy and spectroscopy, Oxford University Press, New York

Petsko G. A., Ringe D. ( 2004) Protein Structure and Function, Blackwell Publishing, Oxford

Radivojac P, Iakoucheva L.M, Oldfield C. J, Obradovic Z., Uversky V. N, A. Keith Dunker A. K. (2007) Intrinsic disorder and functional proteomics. *Biophysical Journal* **92**, 1439-1456

Ramachandran G.N., Ramakrishnan C., Sasissekharan V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95-99

Reardon D. And Farber G. (1995) The structure and evolution of α/β proteins. *Journal of the Federation of American Societies of Experimental Biology* **9**, 497-503

Rhodes G. (2006) Crystallography made crystal clear: A guide for users of macromolecular models, Academic Press

Romero P., Obradovid Z., Kissinger C. , Villafranca J.E. , and Dunker A.K. (1997 ) Identifying disordered regions in proteins from amino acid sequence. *International Conference on Neural Networks* **1**, 90-95

Rupp B. (2009) Biomolecular Crystallography: Principles. Practice and application to structural Biology, Garland Science, New York

Sattler M., Fesik S.W.,(1996) Use of deuterium labelling in NMR: overcoming a sizebale problem. *Structure* **4**, 1245-1249

Segrest J. P., Deloof H., Dohlman J. G, Brouillette C. G., Anantharamaiah G. M. (1990). Amphipathic helix motif - classes and properties. *Proteins: Structure Function Genetics* **8**, 103-117.

Segrest J. P., Jones M. K., Deloof H., Brouillette C. G.,Venkatachalapathi Y. V., Anantharamaiah G. M. (1992). The amphipathic helix in the exchangeable apolipoproteins - a review of secondary structureand function. *Journal of Lipid Research* **33**, 141-166

Stapley B.J. and Creamer T.P. (1999) A survey of left handed polyproline II helices. *Protein Science* **8**, 587-595

Sussman J. L., Lin D., Jiang J., Manning N. O., Prilusky J., Ritter O. and Abola E. E. (1998) Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallographica* **54**, 1078-1084

Taylor D.A. (1991) Structure of a recombinant calmodulin *from Drosophila melanogaster* refined at 2.2-A resolution. *Journal of Biological Chemisrtry* **266**, 21375-21380

Thorton J.M., Sibanda B.L., Edwards M.S., Barrow D.J. (1988) Analysis, design and modification of loop regions in proteins. *Bioessays* **8**, 63-69

Thornton,J.M., Orengo,C.A., Todd,A.E., Pearl,F.M. (1999). Protein folds, functions and evolution. *Journal of Molecular Biology* **293**, 333-342

Weinreb P.H., Zhen W., Poon A.W, Conway K. A., Lansbury P. T., Jr. (1996) NACP, A protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* **35**, 13709-13715

Weber P.C. (1997) Overview of protein crystallisation methods. *Methods in Enzymology* **276**, 13-22

Wright P.E. and Dyson H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure paradigm. *Journal of Molecular Biology* **293**, 321-331

Uversky V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Science* **11**, 739-756

Uversky V. N.,Oldfield C.J. and Dunker A. K. (2005) Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signalling. *Journal of Molecular Recognition* **18**, 343-384

Xu J., Peng J., Zhao F. (2009) Template-based and free modelling by RAPTOR++ in CASP8. *Proteins* **77**, 133-137

# 2. Vibrational Spectroscopy

## 2.1 Introduction

Vibrational spectroscopic analyses allow for probing of biomolecules to determine their structure. As discussed in Chapter 1, although methods like X-ray crystallography and NMR have proven to be powerful methods for studying the structure of proteins, these are not always viable in some cases. Vibrational spectroscopy methods undertake the study of protein structure in aqueous solution which is typically closer to their natural environment. Whilst the analysis of structural information from X-ray diffraction data and NMR experiments is now a routine process, the extraction of structural information from vibrational spectra is often more complicated and subjective. There is a need, therefore, to better understand and correctly identify spectral marker bands and the corresponding protein structural features they pertain to.

### 2.2 Vibrational Energies

Vibrational spectroscopy methods generally involve irradiating biological samples with laser light causing photons of light to be scattered or absorbed or the photons may not interact with any material but simply pass through it. An energy transfer occurs from the light photons to the molecules in the sample if the energy in the incident photon is the same as the difference between the energy of the molecules at ground state and the energy of the molecules in the excited state. This energy transfer from photons to molecules causes vibrational excitation in a molecule as it absorbs energy and is elavated to a higher energy state. Molecular energy can be divided into translational, vibrational and rotational energies. Translational energy is described in terms of three orthogonal vectors in Cartersian space; $x,y,z$ and hence has 3 degrees of freedom. Rotational energy is also described in three degrees of freedom. Linear molecules, for example $CO_2$, have two degrees of rotation, that is they can either rotate about

their axis or parallel to it. Hence, the $CO_2$ molecule has three translational degrees of freedom and two rotational degrees of freedom. Thus, if a molecule has N atoms, the number of vibrational degrees of freedom and, therefore, the number of vibrations possible is 3N-6 for all molecules except linear molecules which have 3N-5 possible vibrations. The types of motion that contribute to the 3N-6 or 3N-5 vibrational modes include; stretching motions between two bonded atoms, bending motion between three atoms connected by two bonds, out-of-plane deformation modes that change an otherwise planar structure into a non-planar one (Ewen and Dent, 2005; Nafie, 2011). In the case of infra red spectroscopy, if the frequency of a specific vibration is equal to the frequency of the radiation directed at the molecule, the molecule absorbs the radiation. As the molecule absorbs radiation it gains energy, which causes it to undergo transitions through different vibrational excitation levels. In Raman spectroscopy, the light photon interacts with the molecules, displacing the electron clouds around the nuclei. The displaced electrons relax back to their normal state to re-establish the stable state in the molecule and the photon is scattered off the molecule. The intensities and energies of the Raman scattered light are measured and contain structural details of the protein being studied.

A combination of fundamental frequencies, differences of fundamental frequencies, coupling interactions of two fundamental absorption frequencies, and coupling interactions between fundamental vibrations and overtones or combination bands creates a unique spectrum for each compound (Banwell, 1972).

### 2.3 Raman Spectroscopy

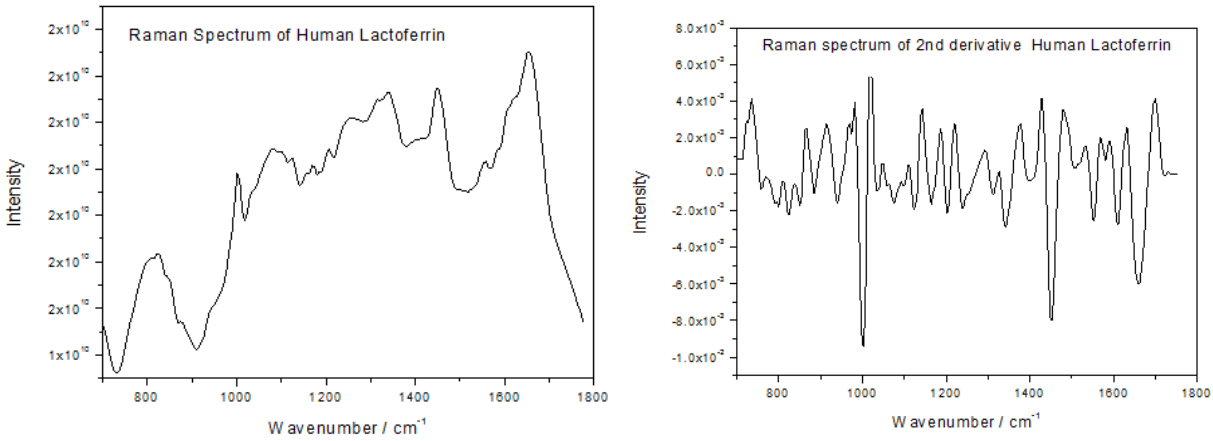
The process of inelastic scattering of light that is central to Raman spectroscopy was first observed experimentally in 1928 by Raman and Krishnan (Raman and Krishnan 1928, 1929; Raman 1928). Incident light irradiating a molecule can cause distortion in the electron

distribution of the molecule and the incident photons are then scattered. When light photons are scattered from a molecule in this way the photons are said to be elastically scattered. These scattered photons have the same frequency and, therefore, wavelength, as the incident photons. This is referred to as Rayleigh scattering. Rayleigh scattering occurs when the electron cloud relaxes without any nuclear movement. There is no change in the energy of the Rayleigh scattered photons. If the scattering process involves motion of the nucleus, energy is transferred between the molecule and photon and the photons are scattered at frequencies above or below the frequency of the incident photons. This type of scattering is referred to as inelastic and constitutes Raman scattering. Raman spectroscopy measures the difference in energy between the incident photons and the Raman scattered photons (Figure 2.1A). A molecule will only generate a Raman spectral band if the vibration changes the polarizabilty. For infra red spectroscopy, the vibration has to involve a change in dipole moment. The Raman effect is due to the interaction of the electromagnetic field of the incident radiation with a molecule (Long, 2002). When a molecule is placed in an electromagnetic field, it undergoes some distortion in which the positively charged nuclei are attracted towards the negative pole of the field and the electrons to the positive pole. This induces an electric dipole moment in the molecule and the molecule is said to be polarised. For example, large atoms such as xenon have a strong polarizability because their electron clouds, distant from the xenon nucleus, are relatively easy to distort with an applied electric field. Helium atoms, which are smaller and more compact, have a small polarizability. Polarizabilities for atoms are the same in all directions, whereas polarizabilities for molecules may vary with position about the molecule, depending on the molecule's symmetry. It is convenient to label polarizability components using Cartesian coordinates to indicate the particular direction to which a polarizability component refers. The size of the induced dipole moment, $\mu$, depends

on the strength of the incident electric field, *E*, and the ease with which the molecule's particles can be distorted (Eqn 2.1)(Banwell, 1972).

$$\mu = \alpha E \qquad (2.1)$$

where $\alpha$ is the polarizability of the molecule and *E* is the electromagnetic field of the incident radiation. Figure 2.1A shows an example of Raman spectra for human lactoferrin protein. Figure 2.1B shows the 2nd derivative Raman spectra of human lactoferrin protein. 2nd derivative Raman spectral data were used in Partial Least Squares (PLS) regression analysis discussed in chapter 6.




A B

**Figure 2.1 A. Raman spectra of human lactoferrin protein B. Second derivative Raman spectra of human lactoferrin protein**

### 2.3.1 The Raman Effect

The Raman effect arises when light interacts with the molecule and causes the atoms of the molecule to polarise. This interaction can be considered as a 'complex' that is not stable and disintegrates releasing the light as scattered radiation (Long, 2002). The scattered radiation is referred to as Stokes scatter. Raman scattering plots are often presented as a Stokes spectrum which is given as a shift of the scattered photon with respect to energy of the laser beam. The excitement of the molecule from ground state to higher energy levels leads to absorption of

energy by the molecule and loss of energy by the photons. This causes what is known as the Stokes shift in the spectrum. However, there are instances when the molecule is in a higher energy state such that when the incident beam interacts with the molecule, energy is transferred from the molecule to the photons. Subsequently, the energy of the incident laser beam is higher than the energy of the scattered light. The latter is referred to as the anti-Stokes shift. At any one time, there are more molecules in the ground state, at lower energy level, than at any higher energy state and as such spectra caused by anti-Stokes shift are very weak (Ferraro *et. al*., 2003).

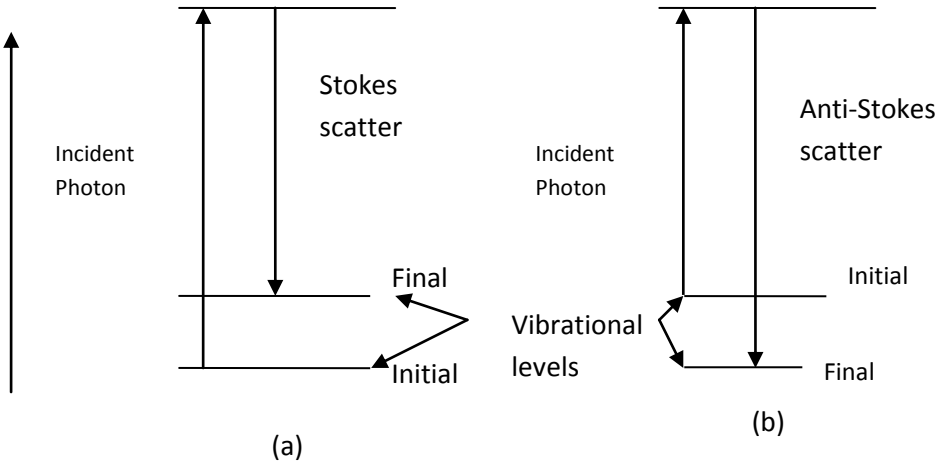



**Figure 2.2 Energy level diagram for Raman scattering; (a) Stokes Raman scattering (b) anti-Stokes Raman scattering. The energy difference between the incident and scattered photons is represented by the arrows of different lengths. The population of vibrational excited states is low under normal conditions. Here, the initial state of a scattering molecule is the ground state and the scattered photon will have lower energy than the exciting photon. This Stokes shifted scatter is what is usually observed in Raman spectroscopy. Raman scattering from vibrationally excited molecules leaves the molecules in the ground state. This anti-Stokes-shifted Raman spectrum is always weaker than the Stokes shifted Raman spectrum.**

Figure 2.2 above illustrates the energy difference between the initial and final vibrational levels of the incident and Raman scattered photons, respectively. Some of the vibrational energy is dissipated as heat but because of the low intensity of Raman scattering, the effects of the heat energy emitted can be neglected as this does not significantly affect the

temperature of the sample, in the absence of absorbance. Raman spectroscopy provides molecular vibrational spectra by means of the inelastic scattering of visible light (Barron, 2004). During the Stokes Raman scattering event, the interaction of the molecule with the incident visible photon of energy, $h\omega$, where $\omega$ is its angular frequency and $h$ is Plank's constant, can leave the molecule in an excited vibrational state of energy, $h\omega_v$, with a corresponding energy loss, and hence a shift to lower angular frequency $\omega$-$\omega_v$, of the scattered photon. Therefore, by analyzing the scattered light with a spectrometer, a complete vibrational spectrum may be obtained.

### 2.4 Raman Optical Activity (ROA)

Raman optical activity (ROA) is another vibrational spectroscopic technique used to study the structure of biological molecules (Barron *et.al*., 2000, 2006). ROA is a chirally-sensitive form of Raman spectroscopy and is therefore an insightful probe into the backbone structure of proteins. ROA measures the difference in the intensity of vibrational Raman scattering from chiral molecules in right- and left- handed circularly polarised light (Barron, 2004; Barron and Buckingham, 1971; Barron *et.al*., 1973; Hug, 2002). ROA is sensitive to the chiral parts of a molecule, so that while conventional Raman spectra tend to be dominated by bands from amino acid side chains, most bands in ROA spectra arise largely from secondary, loop and turn structures thereby providing information about the tertiary fold (Nafie, 2011).

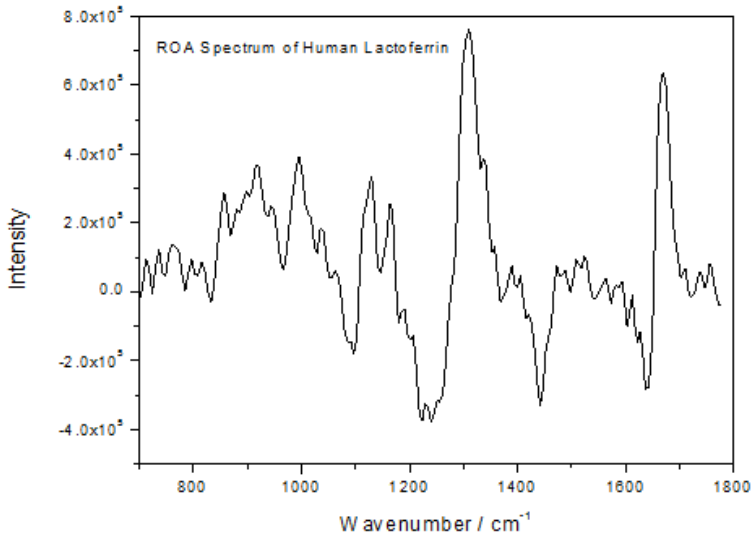



**Figure 2.3 Raman Optical Activity (ROA) spectrum of human lactoferrin protein**

Figure 2.3 shows an example of a protein ROA spectrum. ROA spectra, unlike Raman spectra, are bisignate and so have both positive and negative bands. ROA sensitivity to molecular chirality of 3-dimensional structures can give rich structural detail and the ability to study samples in aqueous solution, with no restrictions on the size of the biomolecules, making ROA ideal for studies of protein structure. The data used in this project were produced by two forms of ROA:- Incident Circularly Polarized (ICP) ROA which uses incident circularly polarized radiation, and Scattered Circular Polarized (SCP) ROA which uses scattered circular polarized radiation. ROA, using the ICP strategy, was first observed in (Barron and Atkins, 1971) and has been used as a probe into the structure and behaviour of proteins (Blanch *et.al*.,2004; Barron *et. al*.,2000;2002;2006) and proteinaceous particles including bacteriophages and virus capsids (Blanch *et. al.*,1999;2001;2002a;2002b) in aqueous solution.

### 2.4.1 Basic ROA Theory

In ROA, the quantity measured can be referred to as the dimensionless circular intensity difference (CID) (Barron, 2004). This analytical measure of chiral signal strength for ROA is given by:

$$\Delta = \frac{I^R - I^L}{I^R + I^L} \qquad (2.2)$$

where $I$ is scattered Raman intensity when incident left (L) or right (R) circularly polarized light is present and $I^R$ and $I^L$ are the right and left incident circularly polarized scattered Raman intensities. The ratio normalizes out instrument variations and the variable of time, allowing an analytical standard to be measured. ROA is measured as the difference between left and right circularly polarized Raman scattered radiation. The $\Delta$ (Eqn 2.2) is observed through a polarization modulator between the scattering molecule and the detector (Oboodi *et.al*., 1985). The original form of ROA measured experimentally was ICP right angle scattering where the scattered radiation could be polarized or depolarized depending on whether polarization analyzer is placed perpendicular or parallel to the scattering plane (Barron, 1989).

Hecht *et.al*. (1989) implemented backscattered ROA which has a higher signal to noise ratio compared to 90˚ or forward scattering. In backscattered ROA, the light is scattered by 180˚ with respect to the incident beam. Almost all protein ROA experiments are performed using the backscattering geometry.

There are several forms of ROA, two of which are mentioned above; in Incident Circular Polarization (ICP) ROA the incident laser is modulated between left and right polarised light measuring the fixed linear or unpolarized Raman intensity. The Scattered Circular Polarization (SCP) form of ROA comprises fixed polarized or unpolarized incident radiation

and measures the difference in the intensity of right and left circularly ($I_R$-$I_L$) polarised Raman scattered radiation. In-phase dual circular polarization ($DCP_I$) ROA involves switching, silmutaneously, the polarization states between right and left circular states of both the incident and scattered Raman radiation. The last form of ROA, out-of-phase dual polarization ($DCP_{II}$) ROA, the polarization states between right and left circular states are modulated conversely (Che *et.al*., 1991; Barron *et.al*., 1989; Nafie, 2011). Figure 2.4 gives an overview of the different forms of ROA.

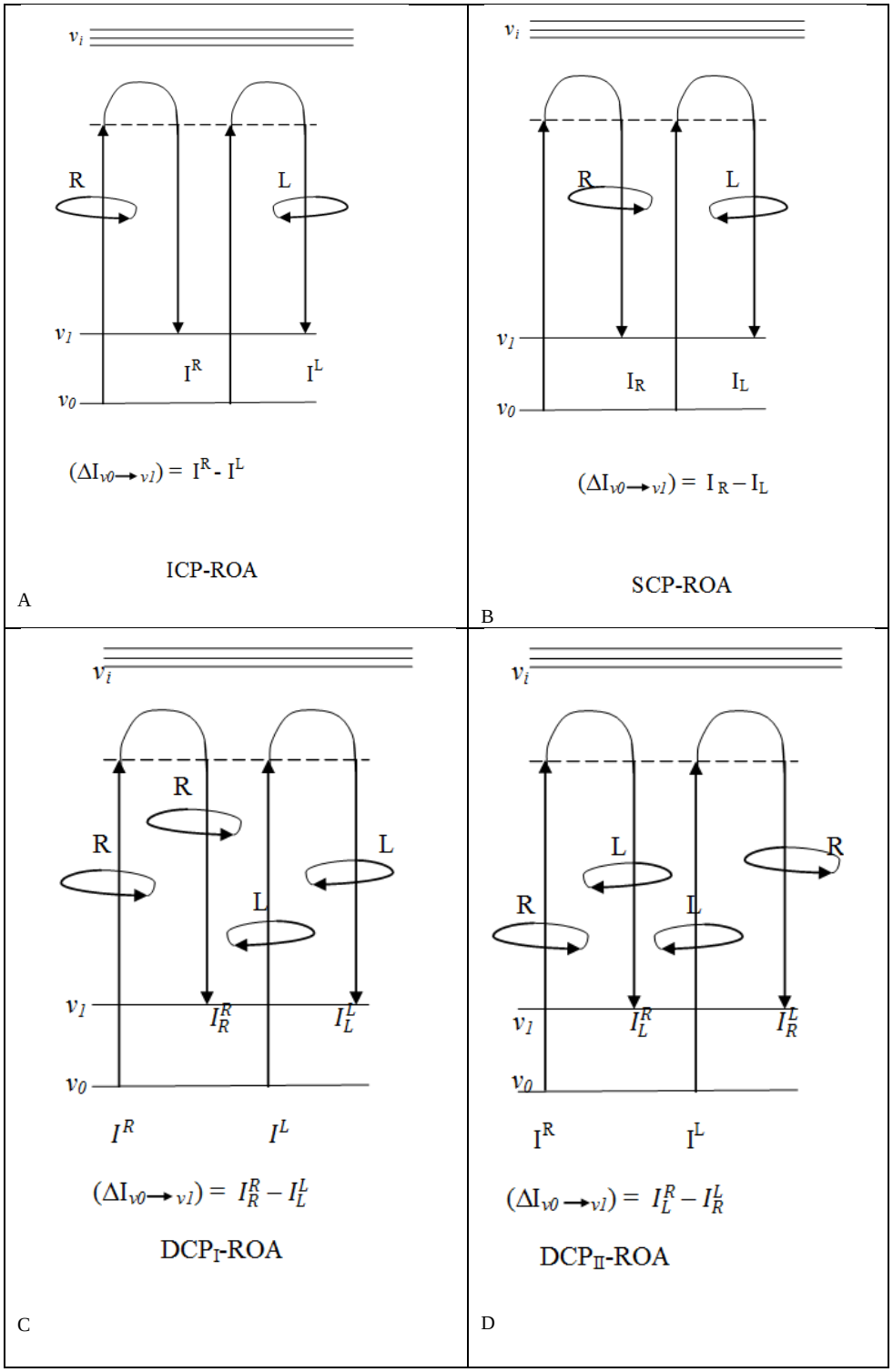



**Figure 2.4 Illustration of energy levels of the different forms of ROA as the molecule goes from low energy level ground state $v_0$ to high energy level $v_1$. $V_i$ represents the higher energy vibration levels. The subscripts refer to the state of the scattered light whilst the superscripts refer to the state of the incident light. In A, the incident laser is modulated between left and right circular polarization states and the differential Raman intensity is measured in an unpolarized state( $I^R - I^L$) . In B, SCP ROA, fixed incident light is used and the difference between left and right polarized Raman scattered light is measured ($I_R$-$I_L$). In C, $DCP_I$ ROA, both the incident light and the Raman scattered light are modulated in-phase ($I_R^R - I_L^L$). In D, $DCP_{II}$ ROA, both the incident and the scattered light are modulated out-of-phase ($I_L^R - I_R^L$). The interchanged subscripts and superscripts in the differential Raman scattered intensity equations in C and D indicate the in-phase and out-of-phase modulations of light for $DCP_I$ ROA and $DCP_{II}$ ROA respectively (Nafie, 2011).**

In ROA, contributions in the spectral intensities arise from the molecular electronic quadrupole and magnetic dipole moments interacting with the incident radiation. The effect the radiation's interaction on the electric dipole, magnetic dipole and electric-quadrupole moments of the molecules are presented as vectors in the Cartesian directions x, y, and z. Equation 2.2 above is calculated by expressing the electric and magnetic field vectors produced by the oscillating electric dipole, magnetic dipole and electric quadrupole moments induced by right- and left- circularly polarized incident light. In Cartesian tensor notation, this is expressed as a fraction of the electric field amplitude of the incident radiation and the fore-mentioned polarizability tensors that arise from the electric dipole-electric dipole polarizability $\alpha_{\alpha\beta}$, the electric dipole-magnetic dipole polarizability $G_{\alpha\beta}$, and the electric dipole- electric quadrupole polarizability $A_{\alpha\beta}$. All the different polarizability-polarizability and polarizability-optical activity tensor components are averaged out over all the angles of orientations of the scattering molecule to generate tensor products constant along axis rotations (Barron, 2004; Nafie 2011).

### 2.5 Spectroscopic Structural Band Assignments

Vibrational spectra are sensitive to biomolecular structure providing a wealth of information on side-chain and backbone conformations. The bands of each spectrum generally correspond to specific structural elements in proteins and when combined give a fingerprint of the protein. There are a number of bands, however, of which little is known. It is necessary to find ways to extrapolate information from spectral bands and investigate which regions of the spectra contain the most structural information. Vibrations of the backbone in polypeptides and proteins are usually associated with four main regions of the Raman spectrum (McColl *et. al.*, 2003; Barron et. al., 2003; Blanch et.al., 2004; Lee and Krimm, 1998; Tsuboi et.al., 2000). The extended amide III region from ~1230 – 1350 $cm^{-1}$ is often assigned to in-phase

combination of the in-plane N-H deformation with the C-N stretch. The amide II region is designated from ~1510 – 1570 $cm^1$ and arises from out of phase combination of the in plane N-H deformation and C-N stretch. It is not as prominent in Raman and ROA spectra but stronger in IR spectra. The amide I region is designated from ~ 1630 – 1700 $cm^{-1}$ and arises mostly from C=O stretch. The amide I region involves mixing between N-H and the $C_\alpha$-H deformations. The coupling between N-H and $C_\alpha$-H deformations generates a rich and informative ROA band structure (Barron *et. al.,* 2003). Table 2.1 shows a few of the sources of spectral band structural information in the amide I, II and III regions.

**Table 2.1 showing some of the secondary structure band assignments for ROA and Raman spectra**

| Structure Type | Amide Region | Raman($cm^{-1}$) | ROA ($cm^{-1}$) |
|---|---|---|---|
| α-helix | AmideIII | ~1330,~1300,~1299 | ~1340, ~1342 |
| | Amide I | | ~1665 |
| β-sheet | Amide III | ~1230-1245 | ~-1219- -1247,~+1260,~+1295 |
| | Amide II | | ~-1550 |
| | Amide I | ~1665-1680 | ~-1658,~+1677 |
| β-turns | Amide III | 1260-1295 | ~+1296,-1347,~-1376 |

## 2.6 Circular Dichroism

Another vibrational spectroscopy method related to the study of 3-dimensional protein structure is vibrational circular dichroism (VCD) (Fasman, 1996; Berova *et.al*., 2000; Norden and Roger, 1997). Circular dichroism measures a difference in absorption of right- and left-circularly polarized light by a molecule. For a molecule to emit a CD signal, it has to be chiral, for example, in proteins where the $C_\alpha$ carbon atom of amino acid residues are bonded to four different substituents except for the Glycine residue; the molecule is situated in an asymmetric environment by the 3-dimensional structure rendered by the molecule; the molecule is linked to a chiral centre (Kelly *et. al*., 2005). Figure 2.5 is a pictorial depiction of the CD differential absorption.

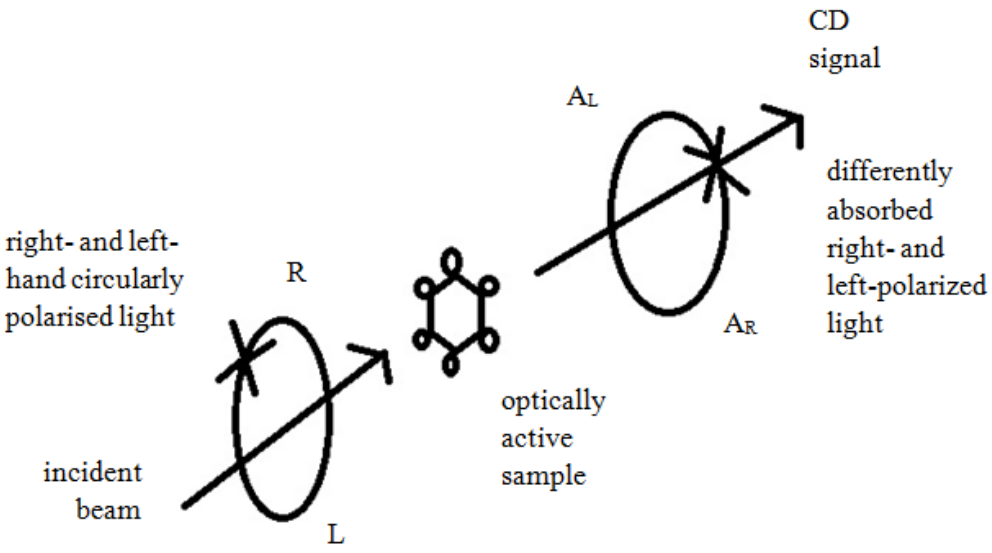



**Figure 2.5 Optically active molecules absorb different amounts of right-(R) and left-(L) circularly polarized light. $A_L$ and $A_R$ are the absorbed left- and right- polarized light.**

The differential absorbance in the CD signal is measured as an ellipticity:

Ellipticity (θ) = constant x ($A_L$-$A_R$) (in units of degrees)

where $A_L$ is the absorption of left-circularly polarized light and $A_R$ is the absorption of right-circularly polarized light.

To compare the ellipticity (θ) values from different protein samples, θ is normalized to obtain the mean molar ellipticity per residue (mrd) (Kelly and Price, 2000; Adler *et. al*., 1973):

$$\theta_{mrd} = \theta_d . \frac{M_W}{c.l.n_r} \qquad (2.3)$$

where $M_W$ is the molecular weight, $c$ is the concentration, $l$ is the spectrometer path length, $\theta_d$ is the differential absorption and $n_r$ is the number of residues in the protein.

As the linearly polarised light transverses the molecule, the properties of the two components of the circularly polarized light change as they are slowed down to different extents. The differently absorbed components result in different amplitudes. The slower component will be absorbed more and have a stronger amplitude than the faster component which is absorbed more. Each component also has a different refractive index. This causes the plane of polarization to be rotated by a small angle producing an effect called optical rotary dispersion (ORD). The superposition of the resulting radiation is not along the plane of linearly polarised light but along an ellipsoid path along which the radiation rotates. This effect of ellipticity (θ) constitutes circular dichroism.

The intensities and energy of the electronic transitions in the molecule which result from exposure to circularly polarisized light depend on the φ and ψ angles of the amide groups of the bimolecular. The observable CD spectra are therefore distinct for different secondary structure motifs in proteins and can yield valuable information about the structural properties of molecules. This is possible because different chromophores in different structural motifs absorb light in different amounts. CD spectra is sensitive to protein secondary structures and different bands from different structural elements like α-helix, β-sheet, disordered structure, β-turns can be determined (Johnson, 1990; Correa and Ramos, 2009).

The CD spectrum of a protein is the sum of its different structural element components and therefore it can be used to estimate the secondary structure composition of a protein (Saxena and Wetlaufer, 1971; Chen and Yang, 1971; Johnson 1988, 1990; Yang et.al., 1986). CD can be used for small quantities of samples, does not require crystals to be used and is widely used in structural biology in combination with other methods like NMR, X-ray crystallography and fluorescence which are discussed briefly in Chapter 1. Several groups have applied CD to the study of biological systems (Woody, 1995), protein ligand interactions and conformational changes (Greenfield, 1999), and protein folding and

unfolding (Kelly and Price, 1997; Correa and Ramos, 2009), globular proteins and nucleic acid structure (Bondesen and Schuh, 2001; Eriksson and Norden, 2001; Norden and Kurucsev, 1994).

### 2.7 Chemometrics analysis of vibrational spectroscopy

Vibrational Spectroscopic methods like these discussed above provide rich structural information which can be mined using bioinformatics computational methods. Chemometric analysis (Geladi *et.al*., 2004) of spectral data is widely applied as an investigation tool for different purposes, including classification, identification, recognition (Breitman *et. al*., 2007; Navas *et.al*., 2008; Manzano *et.al*., 2009). Among the different chemometrics tools, PCA is a powerful data-mining technique that reduces data dimensionality and provides a more interpretable representation of the data under investigation (Jollife, 2002; Bishop, 2006). Barron *et.al*. (2003) showed how proteins could be clustered in a Principal Component Analysis (PCA) plot based on their structural classes.

PCA and Partial Least Squares-Discriminant Analysis (PLS-DA) on Raman spectra were used to classify and identity adulterated milk powder samples (Almeida *et.al*., 2011). The benefits of applying combined Raman spectroscopy and PCA for identifying and analysing potential drug target polymorphs were reported by Strachan *et.al.* (2004). A similar approach (Sato-Berru *et. al*., 2006) was applied to discriminate and identify different carbon walled nanotubes with varying composite walls.

Vibrational spectroscopy used in the detection of disease in human cells (Wong *et.al*., 1991; Diem *et.al*., 1999, 2004; Kendall *et.al*., 2009). Chemometric analyses applied to infra-red spectroscopy has been used to differentiate between cancerous tissue (Romeo and Diem, 2005; Bird *et.al*., 2008; Lasch *et.al*., 2004; Baker *et.al*., 2010), cells at different stages of the

cell division cycle (Boydston-White *et.al*., 2006), drug interaction in human cell lines(Flower *et.al*., 2011).

Chemometrics methods have been used in the diagnosis of disease. A study by Thompson *et. al*. (2004) based on prostate-biopsy specimen showed that 15% of 2940 men that had prostate-specific antigen (PSA) levels of 3.1 to 4 ng per millilitre, levels considered to be normal, did have prostate cancer. Thompson *et. al*. (2004) used logistic regression to assess the risk of having prostate cancer in men aged between 62 and 99 years. The analyses looked at variables such as age in years, history of prostate cancer in the family, race and PSA levels. Their results show that application of chemometrics methods to diagnosis can give useful information. Pathologists manually examine the biopsies to determine diagnosis of the disease, however this method can be time consuming and is susceptible to variabilities in judgement (Bhargava, 2007; Stoler and Schiffman, 2001). A study of diagnostic interpretations on 887 cervical cancer biopsies by Stoler and Schiffman (2001) showed only 42.26% agreement on diagnoses by pathologists.

Vibrational spectroscopy is a non-destructive method for the analysis of cells, tissues and fluids and could be useful in the diagnosis of diseases such as cancer if combined with chemometric analysis methods. Chemometric pattern recognition methods could be a viable alternative to manual diagnosis of disease providing an accurate, less subjective and automated method of disease diagnosis. The project reported in this thesis applied the capabilities of chemometrics analyses to the extraction of protein fold class and secondary structure information from ROA and Raman vibrational spectroscopy.

Chemometrics methods applied to spectral data can be a powerful method of mining data, the analysis and extraction of quantitave estimates depending on the system under investigation. The machine learning methods used in this thesis were Random Forests, Support Vector

Machines and PLS Regression, these are discussed in detail in the following chapter 4. The results from these analyses are presented in Chapters 5-6.

## 2.8 References


Alder A.J., Greenfield N.J. and Fasman G.D. (1973) Circular dichroism and optical rotary dispersion of proteins and polypeptides. *Methods in Enzymology* **27**, 675-735

Almeida M., R., Oliveira K., Stephani R., Oliveira L.F. (2011) Fourier-transform Raman analysis of milk powder: a potential method for rapid quality screening. *Journal of Raman Spectroscopy* **42**, 1548-1552

Bailey P.M. (1973) The analysis of circular dichroism of biomolecules. *Progress in Biophysics and Molecular Biology* **27**, 1-76

Baker M., Clarke C., Demoulin D., Nicholson J., Lyng F., Byrne H., Hart C., Brown M., Clarke N., Gardner P. (2010) An investigation of the RWPE prostate derived family of cell lines using FTIR spectroscopy. *Analyst* **135**, 887-894.

Banwell C.N. (1972) Fundamentals of molecular spectroscopy, 3rd ed., McGraw-Hill Book Company:Maidenshead

Barron L.D. (2004) Molecular light scattering and optical activity, 2nd ed., Cambridge University Press: Cambridge

Barron L.D., Hecht L., Blanch E.W., Bell A.F. (2000) Solution Structure and Dynamics of Biomolecules from Raman Optical Activity. *Progress in Biophysics & Molecular Biology* **73**, 1-49

Barron, L.D., Blanch, E.W., Hecht, L. (2002) Unfolded proteins studied by Raman optical activity. *Advances in Protein Chemistry* **62**, 51–90

Barron L. D., Buckingham A. D. (1971) Rayleigh and Raman scattering from optically active molecules. *Molecular Physics* **20**, 1111-1119

Barron L.D., Bogaard M.P., Buckingham A.D. (1973) Raman scattering of circularly polarized light by optically active molecules. *Journal of American Chemical Society* **95**, 603–605.

Barron L.D., Blanch E.W., McColl I.H., Syne C.D., Hecht L., Nielsen K.,( 2003) Structure and Behaviour of Proteins, Nucleic Acids and Viruses from Vibrational Raman Optical Activity. *Spectroscopy* **17**,101–126

Barron L.D., Zhu F. , Hecht L. (2006) Raman optical activity: An incisive probe of chirality, and of biomolecular structure and behaviour. *Vibrational Spectroscopy* **42**,15–24

Barron L.D., Hecht L., Hug W., and MacIntosh M.J. (1989) Backscattered Roman Optical Activity with a CCD Detector. *Journal of the American Chemical Society* **111**, 8731-8732

Berova N., Nakanishi K., Woody R.,(2000) Circular Dichroism: Principles and applications. 2nd ed., Wiley-VCH,New York

Berru-Sato R.Y., Basiuk E.V., Saniger J.M. (2006) Application of principal component analysis to discriminate the Raman spectra of functionalized multiwalled carbon nanotubes. *Journal of Raman Spectroscopy* **37**, 1302-1306

Blanch E.W., Hecht L., Syme C.D., Volpetti V., Lomonossoff G.P., Nielsen K., Barron L.D. (2002a) Molecular structures of viruses from Raman optical activity. *Journal of General Virology* **83**,2593 – 2600

Blanch E.W., Robinson D.J., Hecht L., Syme, C.D., Nielsen K., Barron, L.D. (2002b). Solution structures of potato virus X and narcissus mosaic virus from Raman optical activity. *Journal of General Virology*. **83**, 241–246.

Blanch E.W., Hecht, L., Day L.A., Pederson D.M., and Barron, L.D. (2001) Tryptophan absolute stereochemistry in viral coat proteins from Raman optical activity. *Journal of the American Chemical Society* **123**, 4863–4864.

Blanch E.W., Gill, A.C., Rhie A.G.O., Hope J., Hecht L., Nielsen K., and Barron L.D. (2004). Raman optical activity demonstrates poly(L-proline) II helix in the N-terminal region of the ovine prion protein: implications for function and misfunction. *Journal of Molecular Biology* **343**,467–476.

Bhargava R. (2007) Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology. *Analytical and Bioanalytical Chemistry* **389**, 1155–1169

Bishop C.M. (2006) Pattern recognition and machine learning, Springer, New York

Bird B., Romeo M.J., Diem M., Bedrossian K., Laver N., Naber S.,(2008) Cytology by infrared micro-spectroscopy: automatics distinction of types in urinary cytology. *Vibrational Spectroscopy* **48**, 101-106

Breitman M., Ruiz-Moreno S., Lopez Gil A. (2007) Experimental problems in Raman spectroscopy applied to pigment identification in mixtures. *Spectrochemica Acta Part A* **68**, 1114-1119

Bondesen B.A., Schuh M.D. (2001) Circular dichroism of globular proteins. *Journal of Chemical Education* **78**, 1244-1247

Boydston-White S., Romeo M., Chernenko T., Regina A., Miljković M., Diem M.,(2006) Cell-cycle-dependent variations in FTIR micro-spectra of single proliferating HeLa cells: Principal component and artificial neural network analysis. *Biochimica et Biophysica Acta* **1758**, 908-914

Che D., Hecht L., Nafie L. A. (1991) Dual and incident circular polarization Raman optical activity backscattering of (-) –trans-pinane. *Chemical Physics Letters* **180**, 182-190

Chen Y., Yang J. T. (1971) A new approach to the calculation of secondary structures of globular proteins by the optical rotary dispersion and circular dichroism. *Biochemical and Biophysical Research Communications* **44**, 1285-1291

Correa D.H.A., Ramos C.H.I. (2009) The use of circular dichroism spectroscopy to study protein folding, form and function. *Journal of Biochemistry Research* **3**, 164-173

Eriksson M., Nordén B.,(2001) Linear and circular dichroism of drug-nucleic acid complexes. *Methods in Enzymology* **340**,68-98

Ewen S., Dent G.,(2005) Modern Raman spectroscopy; a practical approach, John Wiley & Sons, Chichester

Fasman, G.D. (1996) Circular Dichroism and the Conformational Analysis of Biomolecules, Plenum Press, New York

Ferraro J.R., Nakamoto K., Brown C. W.,(2003) Introductory Raman spectroscopy, Academic Press, San Diego

Flower K. R., Khalifa I., Bassan P., Demoulin D., Jackson E., Lockyer N.P., McGown A.T., Miles P., Vaccari L., Gardner P. (2011) Synchrotron FTIR analysis of drug treated ovarian A2780 cells: an ability to differentiate cell response to different drugs? *Analyst* **136**, 498-507

Geladi P., Sethson B, Nystrom J., Lillhonga T., Lestander T., Burger J. (2004) Chemometrics in Spectroscopy Part 2. Examples. *Spectrochimica Acta B* **59**, 1347-1357

Greenfield, N.J. (1999) Applications of circular dichroism in protein and peptide analysis *Trends in Analytical Chemistry* **18**, 236-244

Hecht L., Barron L.D., Hug W. (1989) Vibrational Raman optical activity in backscattering. *Chemical Physics Letters* **158**, 341–344.

Hug, W. (2002) Raman optical activity. In *Handbook of Vibrational Spectroscopy*, Volume 1, J.M. Chalmers and P.R. Griffiths, eds., John Wiley & Sons, Inc.,Chichester ,pp. 745–758.

Jolliffe I.T. (2002*)* Principal component analysis, 2nd ed., Springer-Verlag, New York

Johnson W. C. Jr. (1988) Secondary structure of proteins through circular dichroism spectroscopy. *Annual Reviews of Biophysics and Biophysical Chemistry* **17**,145-166

Johnson W. C. Jr. (1990) Protein secondary structure and circular dichroism: A practical guide. *Proteins: Structure, Function and Genetics* **7**, 205-214

Kelly S.M., Price N.C. (1997) The application of circular dichroism to studies of protein folding and unfolding. *Biochimica et Biophysica Acta* **1338**, 161-185

Kelly S.M., Jess J.T., Price N. C. (2005) How to study proteins by circular dichroism *Biochimica et Biophysica Acta* **1751**,119-139

Kelly S.M., Price N.C. (2000) The use of circular dichroism in the investigation of protein structure and function. *Current Protein and Peptide Science* **1**, 349 – 38

Lasch P., Haensch W., Naumann D., Diem M. (2004) Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica* **1688**, 176-186

Long D.A. (2002) The Raman effect, John Wiley & Sons, Chichester

Manzano E., Navas, Checa-Moreno R., Rodriguez-Simon L., Capitan-Vallvey L.F. (2009) Preliminary study of UV ageing process of proteinaceous paint binder by FT-IR and principal component analysis. *Talanta* **77**, 1724-1731

McColl I.H., Blanch E.W., Gill A.C., Rhie A. G.O., Ritchie M.A., Hecht L., Nielsen K., Barron L.D. (2003) A new perspective on β-sheet structures using vibrational Raman optical activity: From Poly(L-lysine) to the prion protein. *Journal of the American Chemical Society* **125**, 10019-10026

Nafie L.A. (2011) Vibrational optical activity: principles and application, John Wiley & Sons, Chichester

Lee S.H., Krimm S. (1998) General treatment of vibrations of helical molecules and application to transition dipole coupling in amide I and amide II modes of α-helical poly(L-alanine). *Chemical Physics* **230**,277–295.

Navas N., Romero-Pastor J., Manzano E., Cardell C. (2008) Benefits of applying combined diffuse reflectance FTIR spectroscopy and principal component analysis for the study of blue tempera historical painting. *Analytica Chemica Acta* **630**, 141-149

Nordén B., Kurucsev T. (1994) Analysing DNA complexes by circular and linear dichroism. *Journal of Molecular Recognition* **7**, 144-155

Oboodi M. R., Davies M.A., Gumnia U., Blackburn M. B., Diem M. (1985) Instrumnental advances in Raman optical activity. *Journal of Raman Spectroscopy* **16**, 366-372

Raman C. V. and Krishnan K.S. (1929) The production of new radiations by light scattering-Part 1. *Proceedings of the Royal Society* **122**, 23-35

Raman C.V. (1928) A new radiation. *Indian Journal of Physics* **2**,387-398

Raman C.V. and Krishnan K.S. (1928) The negative absorption of radiation. *Nature* **122**,12-13

Romeo M.J., Diem M. (2005) Infrared spectral imaging of lymph nodes: strategies for analysis and artefact reduction. *Vibrational Spectroscopy* **38**, 115-119

Rodger A., Norden B. (1997) Circular Dichroism and Linear Dichroism, Oxford University Press, Oxford

Saxena V.P., Wetlaufer D.B. (1971) A new basis for interpreting the circular dichroic spectra of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 969-972

Smith E., Dent G.(2005) Modern Raman Spectroscopy, A Practical Approach, John Wiley & Sons, New York

Strachan C.J., Pratiwi D., Gordon K.C., Rades T. (2004) Quantitative analysis of polymorphic mixtures of carbamazepine by Raman spectroscopy and principal components analysis. *Journal of Raman Spectroscopy* **35**, 347-352

Stoler M.H., Schiffman M. (2001) Interobserver reproducibility of cervical cytologic and histologic interpretations realistic estimates from the ASCUS-LSIL triage study. *Journal of the American Medical Association* **285**,1500-1505

Thompson I. M., Pauler D. K., Goodman P. J., Tangen C. M., Lucia M. S., Parnes H. L., Minasian L. M., Ford L. G., Lippman S. M., Crawford E. D., Crowley J. J., Coltman C. A. Jr (2004) Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per millilitre. *New England Journal of Medicine* **350**, 2239–2246

Tsuboi M., Suzuki M., Overman S.A., Thomas G.J.,(2000) Intensity of the polarized Raman band at the 1340-1345cm$^{-1}$ as a an indicator of protein alpha-helix orientation:application to Pf1 filamentous virus. *Biochemistry* **39**, 2677-2684

Woody R.W. (1995) Circular dichroism. *Methods in Enzymology* **246**, 34-71

Yang J.T., Wu C.S.C., Martinez H.M. (1986) Calculation of protein conformation from circular dichroism. *Methods in Enzymology* **130,** 208-269

# 3. Machine Learning

## 3.1 Introduction

There has been a huge explosion in the amount of biological data in recent years resulting from high-throughput experiments. This has cumulated in a rapid increase in the number of databanks with large quantities of data, such as GenBank and the Protein Data Bank (PDB). As a result there is a need to match the rate at which information is retrieved from these data to prevent bottlenecks in the flow from raw data to fully analysed information. One of the ways this can be achieved, in addition to wet lab experiments, is the use of software analysis tools. Bioinformatics employs mathematical, statistical and computer science methods to analyse biological data.

Some bioinformatics problems can be modelled in a machine learning framework. Machine learning methods search and exploit complex patterns in data. 'The field of pattern recognition is concerned with discovery of regularities in data through the use of computer algorithms which allows classification of data into categories.' (Bishop, 2006). The learning problem is presented as mapping $x->$ y where $x$ is a large dataset $\{x_1,...,x_N\}$ called the training set and $y$ is the set of categories into which $x$ will be classified. $Y$ represents a target vector with the class. Given a training set $(x_i, y_i)$ …. $(x_m, y_m)$, the main goal is to find the most suitable function that best generalises previously unseen data to the best predicted class, $y$. The function $t(x)$ takes as its input the new data $x$, the training set, and outputs vector $y$. The precise output of the function can be adjusted by varying the variable parameters for optimal performance. This is referred to as learning the classifier, also known as training the model. Once the model is trained it can then be used to predict new class labels, which constitute the test set. The ability to identify new examples is known as generalisation. The original

variable inputs are usually transformed into a new feature space to reduce variability and speed up the computing time, this is referred to as feature extraction. Cases where the input training data has known corresponding target vectors are known as supervised learning problems. If the aim of the application is to assign the input training data to one of the distinct number of categories, that is termed as classification. If the desired output are continuous variables, then that is termed as regression. In certain pattern recognition problems, the training set has no known target vectors. In this case, the aim is to find groups that have similarities as seen in the clustering method.

In our analyses, supervised learning was used to train the models for Support Vector Machines (SVM) regression and classification applications. Other machine learning methods employed, which will be discussed in later sections, were Random Forests classification and Partial Least Squares (PLS) regression.

### 3.2 Support Vector Machines (SVM) Classification

The aim of Support Vector Machine (SVM) classification is to find patterns in data and was introduced by Vapnik, 1995. Given a SVM classification problem with a set of data points $x$ and class labels $y$, we want to learn the mapping: $X \rightarrow Y$, where $x \in X$ is the data vector and $y \in Y$ is a class label. The goal of SVM classifiers is to establish a function that divides $X$ leaving all the points of the same class on the same side of a separating margin while maximising the minimum distance between the two classes and the hyperplane. This is referred to as generalisation, where given $x \in X$, as previously seen, the classifier finds the suitable $y \in Y$. In a 2-class classification, as used in our analyses, the classes are assigned as $y = 1$ or $y = -1$. Finding the suitable class label involves learning or training the classifier $y = f(x,\alpha)$, where $f$ is the generalisation function and $\alpha$ are the parameters that are varied to

optimise the separating function. The training data set comprises N input vectors $x_{1,\ldots},x_N$ with corresponding target values $y_{1,\ldots},y_N$. Different functions can generalise a given set of data to give a completely different set of test label predictions. It is therefore necessary to have a way of choosing the most suitable function. The Support Vector Machine algorithm approaches this problem by searching for the separating boundary with the maximum margin, which is defined as the smallest distance between the decision boundary and any of the data points as shown below in figure 3.1. The "optimal" boundary is defined as the most distant hyper plane from both sets of +1 and -1 data.

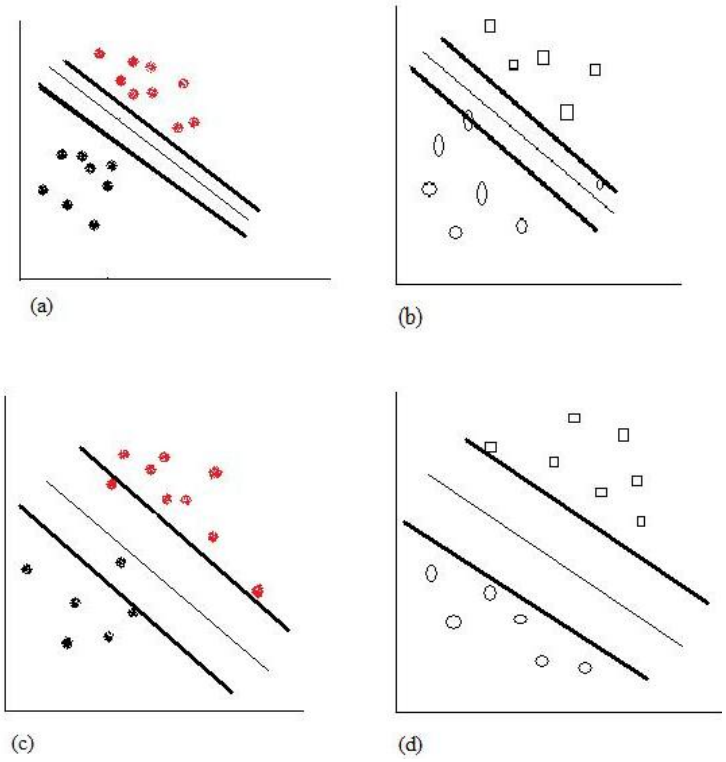



**Figure 3.1 An overview of the training procedure. The classifier is trained on training data to find the optimum parameters of the classifying model without overfitting. Over fitting makes the model a poor discriminator when used to predict on new data. (a) Training data and producing an overfitting model. (b) Using the overfitting model on the test data yields poor results. (c) Training the model on training data to improve performance (d) Using the improved model on training data yields better generalisation results. ● and ● represent test data; □and ○represent training data. (Chih-Wei, 2008)**

The simplest SVM model of linear separation of the input space is represented by Eq. (3.1) below:

$$y(x) = \quad w^T \emptyset(x) + b \tag{3.1}$$

where $w$ is a weight vector, ø represents the function that projects the data into feature space and $b$ is the intercept term. An input vector $x$ is assigned to class $y_{+1}$ if $y(x) \geq 0$ and class $y_{-1}$ if $y(x) \leq 0$. The decision boundary is therefore determined by the relation $y(x) = 0$. If we have two points $x_a$ and $x_b$ that lie on the decision boundary $y(x_a) = y(x_b) = 0$, we have $w^T(x_a - x_b) = 0$ hence the vector $w$ is orthogonal to every vector along the decision boundary and determines the orientation of the decision boundary. By geometry, the distance between the hyperplanes is $\frac{2}{||w||}$. Among all the possible hyperplanes separating the data, there is one unique one that yields a maximum margin of separation between the classes so we want to maximise $||w||$ (Burges 1998, Muller *et.al.*, 2001, Chen *et. al.*, 2007). Constraints are added to ensure data points are on either side of the separating hyperplane and not in between.

$$x_i \; w + b \geq +1 \; for \; y_i \; = +1 \tag{3.2}$$

$$x_i \; w + b \geq -1 \; for \; y_i \; = -1 \tag{3.3}$$

Equations (3.2) and (3.3) can be combined to be written as

$$y_i(w^T x_i + \; b) \geq 1 \; for \; i = 1, \ldots, m \tag{3.4}$$

where $y_i$ is 1 if $x_i$ belongs to one class and –1 if $x_i$ belongs to the other class. If the boundary classifies the vectors correctly, then $y_i(w^T x_i + \; b) \geq 0$ and it is identical to the margin as shown in figure 3.2 below. Among all the possible hyperplanes separating the data, there is one unique one that yields a maximum margin of separation between the classes, and the

capacity (the ability of a function to separate data in all possible ways) decreases with increasing margin. Maximizing the margin leads to a specific choice of decision boundary. In order to maximize the margin, parameters *w* and *b* have to be optimized.

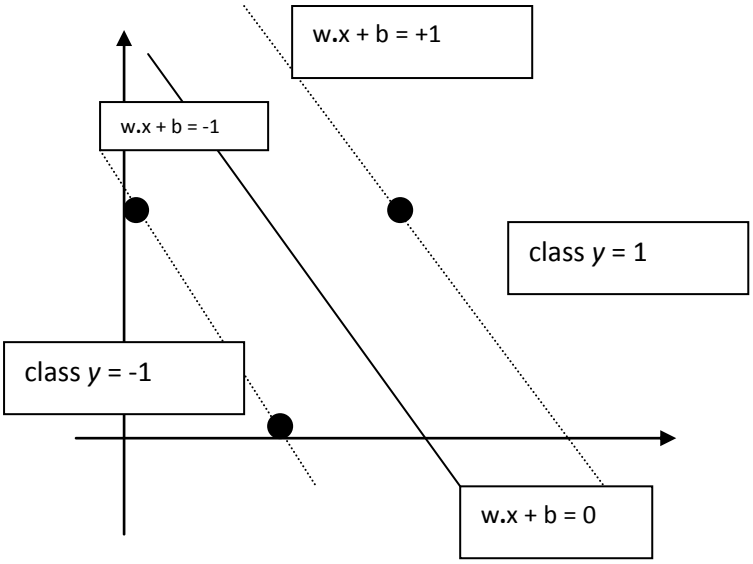



**Figure 3.2 The SVM algorithm finds the largest distance between the hyperplanes, the margin denotes the dot product. The vector w is a normal vector, it is perpendicular to the hyperplane. The algorithm chooses the *w* and *b* to maximise the distance between the parallel hyperplanes separating the data as much as possible at the same time. (Schölkopf and Oldenbourg, 1997)**

The support vectors lie closest to the decision boundary and are critical to the separation of the data points.

### 3.2.1 Soft Margin

The above discussion is applicable to the case of linearly separable sets only. If the sets are not linearly separable, a hyperplane exactly classifying the sets does not exist because of a high noise level that causes a large overlap of the classes, as explained in the previous

section. The method called *soft margin* is a solution to such cases (Chen *et. al.*, 2007). This method replaces the restriction in Eq.(3.4) with the following:-

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \ldots, m \quad (3.5)$$

where $\xi_i$ called *slack variables*, are positive variables that indicate tolerances of misclassification. If the point $x_i$ satisfies inequality in Eq. (3.4), the term $\xi_i$ is null and Eq. (3.5) reduces to Eq.(3.4). If, on the other hand, point $x_i$ does not satisfy inequality in Eq. (3.4), the term - $\xi_i$ is added to the right hand side of Eq.(3.4) to obtain inequality in Eq.(3.5). This replacement indicates that a training vector is allowed to exist in a limited region in the erroneous side along the boundary. A classifier that generalises well can be found by controlling both the classifier capacity through variations of the margin ||w|| and the sum of the slacks $\Sigma_i \xi_i$. Several variants of the soft margin SVM algorithm have been proposed for example, the C-SVM introduces the C parameter (Chen *et. al*, 2007).

$$minimise\ w, \xi\ \frac{1}{2}||w||^2 + C \sum_i \xi_i \quad (3.6)$$

where C is a parameter that controls the trade off between the requirements of a large margin and a few misclassifications. The parameter C can be regarded as a regularization parameter.

### 3.2.2 Kernel Methods

In cases where the data is not linearly separable, the data are projected into a higher dimensional space using a kernel function. The fundamental concept of a kernel method is a mapping of the vector space itself to a higher dimensional feature space. This makes it applicable to non-linear classification problem (Bishop, 2006; Tsuda and Vert, 2004). Various methods can then be applied in that space to find relations between data points. The data is represented by an *n x n* matrix of pair wise comparisons $k_{i,j} = k(x_i, x_j)$. A square matrix is a key condition for most kernel methods which can only process square matrices

which are symmetric positive definite; that is it should satisfy $t_{j,i} = t_{i,j}$ for any $1 \leq i,j \leq n$. The kernel function makes it possible for kernel methods to operate in the feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the all pairs of data in the feature space (Tsuda and Vert, 2004). This operation is often computationally cheaper than the explicit computation of the coordinates.

**Linear kernel**

The linear kernel is the simplest of the kernel functions. Given a vector set $x = (x_{1,\ldots,}x_m)$. The vectors are compared using their inner product;

$$k = (x, x') = x^T x' = \sum_{i=1}^{m} x_i x_i' \quad (3.7)$$

If the data analysed are not vectors, the data can be first represented as a vector by mapping them into a feature space using a transformation $\phi$ and then finding the pairwise inner product of the vectors hence defining the kernel function.

$$k = (x, x') = \phi(x)^T \phi(x') \quad (3.8)$$

The transformation $\phi$ does not need to be computed for every point as only for the pairwise dot products are necessary. This makes computation simpler by reducing the dimensionality.

**Gaussian Radial Basis Function kernel**

The Gaussian Radial Basis Function introduces the notion of distance in the measure of similarity. The output of the kernel is dependent on the Euclidean distance *d of* $x_j$ from $x_i$ . Given a set of input vectors $\{x_{1,\ldots,}x_N\}$ with their corresponding target values $\{y_1,\ldots,y_N\}$, the aim is to find a function *f(x)* that fits every target value exactly. This is achieved by expressing *f(x)* as a linear combination of radial basis function, one centred on every data point:

$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (3.9)$$

A support vector will be the centre of the RBF and $\sigma$ will determine the area of influence this support vector has over the data space. A larger value of σ will give a smoother decision surface and more regular decision boundary. This is because an RBF with large σ will allow a support vector to have a strong influence over a larger area. A larger σ value also increases the value α (the Lagrange multiplier) for the classifier. When one support vector influences a larger area, all other support vectors in the area will increase in $\alpha$ value to counter this influence. Hence all α -values will reach a balance at a larger magnitude. A larger σ -value will also reduce the number of support vectors. Since each support vector can cover a larger space, fewer are needed to define a boundary (Bishop ,2006).

**Polynomial kernel**

The polynomial kernel function is directional, i.e. the output depends on the direction of the two vectors in low-dimensional space. This is due to the dot product in the kernel for a two dimensional polynomial kernel. The equation below shows a polynomial kernel function:

$$K(x_i, x_j) = (x_i . x_j + 1)^d \quad (3.10)$$

where $d$ is the degree of the function

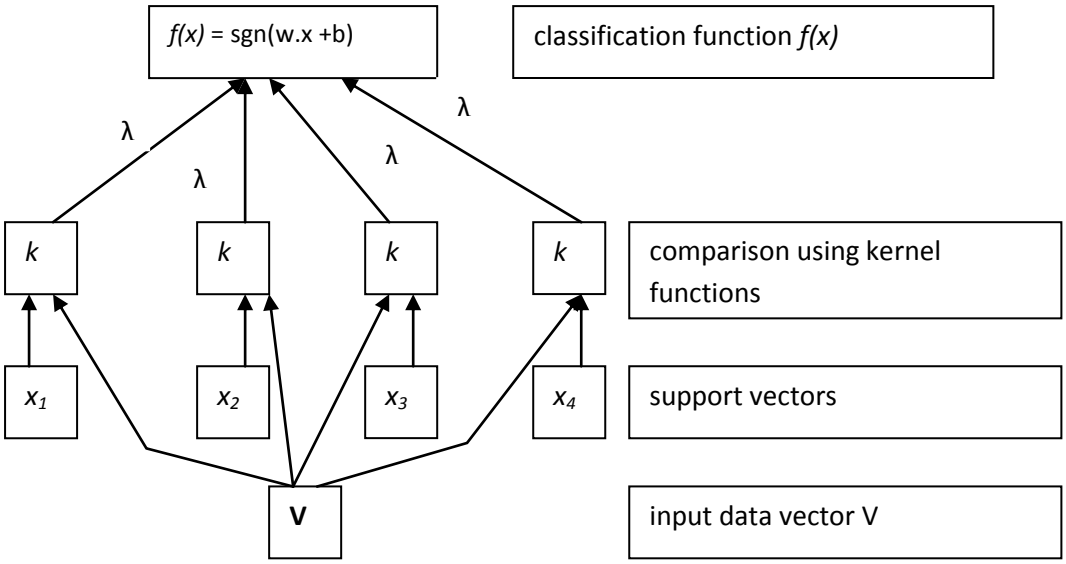



**Figure 3.3 The SVM Architecture: The diagram above shows the overall structure of SVMs. Similarity measures are applied to the input vector V in the form of kernel functions. These vary from dot product kernels for linear separation to Gaussian and sigmoid kernels for non linear separation. Only the support vectors are included in the optimal classification solution. The weights $\lambda_i$ are adjusted to find the optimal separating function f(x) (Scholkopf B., 2005).**

Figure 3.3 above is a schematic representation of the training process in SVM classification. If the training data of input space cannot be separated linearly, these data are mapped into a higher dimension space, which is called feature space. The data are then linearly separated in this space. To reduce the complexity of dimensionality, kernel methods are used for non-linear data. In the Kernel method, a nonlinear transformation is substituted by an inner product, which is defined by a kernel function. An area of intense research in SVM's is the criteria to select kernel functions (Tao Wu *et. al*., 2001). At present, selection of kernels is done by experimenting and experience. Test or training errors are also used as a measure of the efficiency of the kernel. It is important to get a function that not only generalises on the training data but also generalises unseen example well. To achieve this it may be necessary to select a kernel function with a bit of training error. This presents a trade-off between choosing low training error and the ability to generalise.

### 3.3 SVM Regression

Support Vector Machine (SVM) can be applied to regression problems by using variations of the loss function. A loss function represents a measure of the loss of accuracy that arises from the difference between the true observed value and the estimated or predicted value. SVM regression optimizes the generalization margins given for regression and defines Vapnik's loss function that allows for some errors, which are situated within a certain distance of the true value (Vapnik, 1995). This type of function is called epsilon intensive loss function or ε-insensitive loss function and defined as:

$$L(x, f(y)) = \begin{cases} 0 \\ |x - f(y)| - \varepsilon \end{cases} \quad (3.11)$$

where ε > 0 is a predefined threshold value that controls the amount of deviation allowed from the true value. The aim of the ε-insensitive loss function is to find a function *f(y)* that has the most ε deviation from the desired target for all training data and at the same time produces hyperplanes of an *n*-dimensional space i.e. is flat.

In SVM regression, the input vector ***x*** is first mapped onto an *n*-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space (Farag and Mohamed, 2004, Smola and Schoelkopf B. 2004). Using mathematical notation, the linear model (in the feature space) is given by Eq.(3.12):

$$f(y) = \langle w, y \rangle + b \; with \; w \in X, b \in \mathbb{R} \quad (3.12)$$

where w is the weight vector
$X$ is the input space
$b$ is the offset value
$\langle w, y \rangle$ is the dot product of vectors $w$ and $y$
$\mathbb{R}$ all values represented in the feature space

The SVM algorithm aims to find a function, Eq. (3.12) above, that is flat and fits all the training data points. This is achieved by minimizing $\| \omega \|^2$ , where $||\omega||^2$ is the normalised vector that represents the width of the margin. The problem can thus be represented as a convex optimisation problem i.e. the solution is subject to constraints and lies within a specified subspace:-

$$minimize \ \frac{1}{2}\|w\|^2 \tag{3. 13}$$

$$subject \begin{cases} x_i - (w.y + b) \leq \varepsilon \\ (w.y + b) - x_i \leq \varepsilon \end{cases} \tag{3.14}$$

The constraints in Eq. (3.14) set limits on the accepted solution set. However, there can be cases where the solutions fall outside this set limit. This can handled by introducing slack variables $\xi_i, \xi_i^* \ i = 1,...n$ , to measure the deviation of training samples outside the $\varepsilon$ -insensitive zone (figure 3.4) . Now Eq. (3. 13) and Eq. (3.14) are formulated as:

$$\text{minimise } \frac{1}{2}\|\omega\|^2 \ + \ C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3.15}$$

$$subject\ to \begin{cases} x_i - \langle w, y_i \rangle - b \leq \varepsilon \ + \ \xi_i \\ \langle w, y_i \rangle + b - x_i \leq \varepsilon \ + \ \xi_i^* \\ \xi_i \ \xi_i^* \qquad\qquad\qquad \geq 0 \end{cases} \tag{3.16}$$

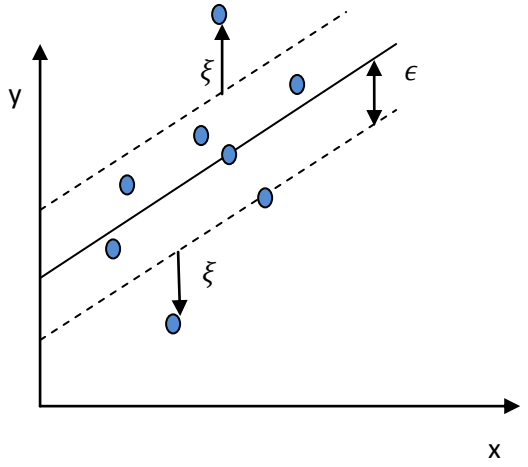



**Figure 3.4 One dimensional regression with epsilon ε insensitive band. The band marks the margins where erroneously classified data points are allowed. The points marked ξ along the dashed lines represent the data points that fall outside the 'acccepted' margin of error.**

Figure 3.5 shows the one-dimensional linear regression function with epsilon intensive – band. The slack variables $\xi_i^*$ ,$\xi_i$ measure the cost of the errors on the training points. The value of the slack variable is zero for all points that are inside the band.

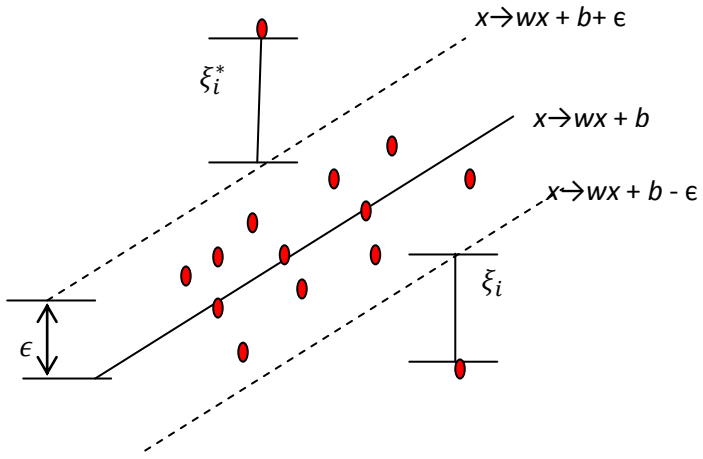



**Figure 3.5 Detailed picture of epsilon band with slack variables and selected data. ε represents the epsilon insensitive region, $\xi_i^*$ and $\xi_i$ represents the data points outside the error margin . The data points outside the dotted lines are points that have fallen outside the allowed error limits referred to as the slack variables. The terms *wx+b+ε, wx+b-ε* refer to the optimum margin of separation, *wx+b* refers to the hyperplane.**

In figure 3.5, $\xi_i$ controls the error induced by observations that are larger than the upper bound of the $\epsilon$-tube. $\xi_i^*$ controls the error for the observations small than the lower bound. The dashed line indicates the boundaries of the areas where the loss is zero i.e. target value and true observed value are close.

### 3.4 Random Forests

#### 3.4.1 Introduction

Random Forests was developed by Breiman (2001) and uses a combination of many decision trees such that each tree depends on randomly selected vectors. To grow the trees, the Random Forest algorithm generates random vectors which contain random integers. These vectors determine the way in which the training data set is split into subsamples. The integer values in each vector present the number of examples from the training data selected in every subsample. For each subsample the best splitter is used to split the node in question. This is done recursively for each vector creating an ensemble of trees which lends its name to this technique. Each subsample then becomes a classifier to which an input vector is then applied for classification by the best split. At each subsample the best splitter is determined using the Gini index. The Gini index chooses the split which best increases the information gain. The information is the probability of the possible outcomes of each split. This information is measured in units of bits and is referred to as entropy. The information gain is the difference in information before a sample is split and after the split. The split with the most information gain is the most favourable. Each tree in the ensemble is grown to maximal size repeating the splitting process at each node such that a different subset of variables is selected from the training data. The number of variables in each subset being set by the integers in the random vectors generated.

### 3.4.2 Gini Index

The criterion used in splitting the nodes in RF trees is the Gini Index rule (Breiman, 1984). The splitting decision is based upon the decrease in the impurity at the node. The impurity at the node is the proportion of samples that are false positives at that particular node. The higher the number of samples at a node belonging to one class the more pure the node. The Gini value of a split at a node into subsets is defined as:

$$\Delta imp(sp, n) = imp(n) - P_L imp(n_l) - P_R\, imp(n_r) \quad (3.17)$$

where *n* is a given node; *sp* is a split in the node

After splitting, the new nodes are referred to as the left and right child nodes.

$P_R$ is the number of variables at the given node *n* that split into the right child node $n_R$

$P_L$ is the number of variables at the given node *n* that split into the left child node $n_L$

*imp*($n_R$) is the impurity of the child node on the right $n_R$

*imp*($n_L$) is the impurity of the child node on the left $n_L$

The Gini impurity maximises the average purity of the two child nodes. The selected splits are those that decrease the Gini Index the most. The impurity value *imp* for the node *n* is:

$$imp(n) = 1 - sum(p1^2, p1^2, p1^2, p1^2 \ldots, pn^2) \quad (3.18)$$

where *p1...n* are the frequency ratios of class *1...n* in node *n*

After splitting the node *n* into $n_R$ and $n_L$ child nodes, with $N_R$ and $N_L$ number of samples respectively, the gini value $\Delta imp(sp,n)$ is calculated from the impurities of the right and left nodes $\Delta imp(n_R)$ and $\Delta imp(n_L)$ as:

$$\Delta imp(sp, n) = \frac{N_R}{N} \Delta imp(n_R) + \frac{N_L}{N} \Delta imp(n_L) \quad (3.19)$$

where N is the total number of sample in a node. The split variable that generates the smallest gini value Δ*imp(sp,n)* is chosen to split the node on.

### 3.4.3 Random Forest prediction

The trees are grown to their maximal size. Classification trees assign each case to one class only. For example, 75 trees will have 75 different assignments for each case. To combine the trees, the votes for each class are counted up and the class with the most votes is the overall 'winner'. The number of votes is the RF score. The proportion of the votes correlates to the probability of the class membership. RF allows setting of class weights to correct imbalances in the data and to boost the accuracy of the specified class. This is described in the methods chapter 4 section 4.4.3.

### 3.4.4Out-of-Bag (OOB) data

Each tree is grown out of 2/3 of the original data while the remaining 1/3 are used to test a single tree. The left out data, referred to as Out-of-bag (OOB), is used to assess the performance accuracy of the model at each tree (Breiman, 1996a; 1996b). Model results are combined to give a single prediction through a voting system for classification problems. All training data is used in the training, but only a subset at a time.

### 3.4.5 Distance Metric

RF evolves trees to their entirety without pruning. The left out data are run down each tree. If observations *i* and *j* end up in the same terminal node the similarity between *i* and *j* is increased by one. The more similar data records are, the more likely they will land in the terminal node of a tree. The distance metric can be used to construct a dissimilarity matrix

input into multidimensional scaling (MDS) (Groenen and Velden, 2004; Naud, 2006). The dissimilarity matrix represents the pairwise relationships between the variables derived from the data features. Dissimilarity matrices reduce the high dimensionality of large data as the size of the matrices is directly proportional to the number of objects ($O|N^2|$) and is independent of the data's dimension. Multidimensional Scaling (MDS) is a method used to analyse dissimilarity matrices to produce distances based on a specified number of dimensions. MDS analysis based on two dimensions, as used for the analysis presented here, produces a two-dimensional representation of the data variables and a two-dimensional map is obtained. MDS aims to arrange the data variables in a particular space with a specific number of dimensions in order to reproduce the observed distances. In this case the MDS represents the RF distances in a two-dimensional map.

As stated above, the MDS moves the data variables around in order to find the configuration that best estimates the observed distances. The MDS algorithm evaluates all the different possible configurations and uses a function minimisation with the aim of maximising the goodness-of-fit to choose the optimum configuration. The most common measure of goodness-of-fit of the approximation between the original dissimilarities and the interpoint distances produced by MDS is the stress measure.

In MDS for a set of *n*, *m*-dimensional data items, a configuration of *X* of *n* points in a reduced dimensional space *r* is sought and this should minimise the stress function. Stress is defined as the square of differences between the dissimilarities and the *N* r-dimensional objects in the ouput space.

There are variations in the calculation of stress: In metric stress, the stress is normalised by the sum of squares of the dissimilarities. In squares stress (sstress), the stress is normalised by the fourth power of the Euclidean distances. In Sammon mapping, the sum of the scaled, squared differences between the distances and the dissimilarities, is normalised by the sum of

the dissimilarities. The squared differences are scaled by the dissimilarities before summing (Ahmed *et. al.,* 2005). Sammon mapping was chosen for our data because it gave the smallest stress value. The smaller the stress function the better the model represents the input data (Jaworska *et. al.*, 2009). The actual axes on the MDS dimensional plot are arbitrary meaning that the distances remain the same whichever way the map is looked at.

### 3.5 Partial Least Squares (PLS) Regression

#### 3.5.1 Introduction

PLS Regression, introduced by Wold (1966) in the social sciences, is a widely used technique in chemo metrics and has been used by various groups in the analysis of data such as the correlation of particular gene expression levels between expression levels of the other genes in a study on expression data from *Saccharomyces cerevisae* (Datta, 2001), sensory evaluation (Martens and Naes, 1989) and analysis of neuroimage data (Mcintosh *et.al.*, 1996). Boulesteix and Strimmer (2005) used a partial least square approach to investigate the relationships between transcription factors using gene expression data.

PLS reduces dimensions of data and can handle a high number of variables (Chun, 2008). Partial Least Squares (PLS) regression is a multivariate data analysis technique which can be used to relate several responses (X) variables to explanatory (Y) variables. The method aims to identify the underlying factors, or linear combination of the Y variables, which best model the X dependent variables. PLS can deal efficiently with data sets where there are many variables that are highly correlated and involving substantial random noise. The need to predict Y response variables from X-variables in scientific data is a common problem for example in biology; estimating the rate of photosynthesis from the amount of light; estimating the rate of growth of specimen from the amount of substrate.

Given a data set with response variables Y in matrix form and a data set with predictor variables X in matrix form, PLS is a multiple linear regression technique that builds a linear regression model (Boulesteix and Strimmer, 2007).

$$Y = XB + E \tag{3.20}$$

where Y is a *p rows* by *n columns* matrix of response variables, X is a $p$ rows by *m columns* matrix of predictor variables, B is a *m* by *n* regression coefficient matrix and E is a matrix of noise terms of $p$ by $n$. PLS extract factors from predictor matrix X producing factor score matrix $T = XW$ where W is a weight matrix.

The Y matrix is decomposed as response matrix $Y = UQ$ where U is the score matrix and Q is a matrix of regression coefficients (loadings matrix) of U. The overall aim is to use the scores of X to predict the responses in the population. The factors T are used to predict the Y scores U. The predicted Y scores are then used to construct predictions for the responses in the population. The scores for X and Y are chosen such that the directions between the factors represent a high variance but are limited to the directions that yield the accurate predictions (Hoskuldsson, 2004; Tobias, 1997).

### 3.5.2 The PLS Model

The PLS model is developed from a set of P observables with M $x$-variables; $x_m$ where m = 1,...,M and P $y$-variables $y_n$ where n=1,...,N, forming two matrices X of P by M and Y of P by N dimensions (Abdi H., 2007). The PLS algorithm projects the rows, $p_i$ and columns, $t_{i,}$ of the X–matrix iteratively into a dimensional space as vectors. Each column defines an element of a vector of loadings $p_i$ and each row defines an element in the scores vector $t_{i.}$ Each column defines one coordinate axis in the in the dimensional space.

The algorithm then defines a *p* rows by *n* columns dimensional hyperplane which is defined by one axis per component. The cosine direction coefficients of these axes are the loadings $p_{i.}$ The row data *i* of the X-variable matrix is then projected down onto this hyperplane with the coordinates of the data points defined by the scores $t_i$ . The loadings are the cosine direction of each of the principal components with respect to each of the coordinate axes. The position of the hyperplane is such that this plane estimates the row vectors *i* of X-variable matrix and at the same time ensures that the coordinates of the projected data, the scores $t_i$, are related to the reponses $y_i$ as the aim of regression is to get a meaningful relationship between X and Y matrices. The principle component can be described as a line drawn between X and Y to show their correlation. A simple linear regression model between the score factors of X, $t_i$ (Eq. 3.21) and those from the Y response matrix, $u_i$ (Eq. 3.22) is then computed (Geladi and Kowalski, 1986).

$$X = t_1 p_1 + \cdots + t_i\, p_i \tag{3.21}$$

The Y response matrix is also treated in the same way decomposing into loadings *q* and score vectors *u* :-

$$Y = u_1 q_1 + \cdots + u_i\, q_i \tag{3.22}$$

The PLS algorithm extracts latent factors, *u* (as in Eq. 3.22) and *t* (as in Eq. 3.21) from the dependent variables Y and the predictor variables X respectively. These latent factors are defined as linear combinations constructed between predictor X and response variables Y, such that the original multidimensionality is reduced to a lower dimension that describes the structure in the relationships between predictor variables (X) and between these latent factors and the response variables (Y).

The latent factors *t* are extracted from predictor variables X such that the explained variance in the dependent variables Y is maximized. The latent factors *t* are also referred to as X-

scores and the latent factors *u* as Y-scores. New X-scores and Y-scores are predicted by varying the directions (loadings) in the factor space such that variation is maximised but approach as closely as possible to the Y responses. The predicted vectors of scores are removed from the matrices Y and X and steps reiterated with the remainders of Y and X until the X-and Y- scores converge. The PLS regression technique makes it possible to analyse the effects of linear combinations of several predictors on a response variable. The technique is suitable for the analysis of data where the number of predictor variables is higher than the observed responses (Carrascal *et.al*., 2009).

### 3.5.3 PLS number of components

As not all the principle components are used, PLS uses the cross validation method to choose the number of components to use (Geladi and Kowalski, 1986). Cross validation is used to test the prediction accuracy of the model. This is performed by dividing the data into *g* number of groups. One group at a time is removed and a model produced from the remaining data. The model is then used to make predictions on the group, the differences between the predicted and actual Y-values are then calculated. This is done iteratively until all the data has been removed. The sum of squares of these differences are calculated in turn for each model. The model that yields the smallest differences is then used.

## 3.6 Chemometric studies of protein vibrational spectroscopy

Random Forest (RF) clustering was used to profile tumour samples based on tissue microarray data from patients with renal cell carcinoma (Shi *et. al*., 2005) and visualised using multidimensional plots based on random forest dissimilarity. The RF clustering grouped the grade 2 data samples into two distinct groups with significantly different survival profiles: one group had a median survival time of 12 years and the other of 27 years. This

showed that Random Forest clustering could differentiate different tumour marker expressions and was sensitive to expression heterogeneity.

Another method of studying protein structure is the deconvolution of the amide bands related to different structural signatures where the abundance of each structure is proportional to the area under the curve. More recent methods include regression methods in which a relationship is established between a matrix of spectra (X) and a vector of responses (Y) that define the structural fractions as discussed in Section 3.5. The band narrowing interval Partial Least Squares (iPLS) regression method of Navea *et. al.*, 2005 was used to study structural conformations of 24 proteins from Infrared spectra of the amide regions I, II and III. iPLS regression develops local PLS models from sub-regions of the whole spectrum cutting out the noise from the other regions. The spectra were divided into intervals varying from 38.6 to 7.7-9.6 $cm^{-1}$ and then all different combinations of intervals used to build PLS models to determine which spectral ranges would improve the predictive capability of the models.

The best iPLS model for prediction of α-helix structure was obtained from the amide I band which had the highest percentage variance of 79.18% and root square mean value (RMS) of 0.12. The RMS value has been defined in Chapter 4 of this thesis.The best iPLS model for β-sheet prediction from the amide I band had a percentage variance of 83.05% and RMS value of 0.08. Navea *et. al.* (2005) concluded that the approach used for their analyses revealed that the amide III band (1350-1200$cm^{-1}$) was not relevant to the prediction of the secondary structure for the IR data.

Navea *et. al*. (2005) found the best model for the determination of β content was built from combinations of 2 intervals: 1647.1-1639.4 $cm^{-1}$ from amide I and 1569.9-1562.2$cm^{-1}$ from amide II with a percentage variation of 87.48% and RMS value 0.08. The best model for α determination was from intervals of 1668.3-1658.7 $cm^{-1}$ and 1550.6-1542.9$cm^{-1}$ showing

percentage variation of 82.19% and RMS of 0.10. These iPLS results showed the positive effect of combining two or more spectral bands in the determination of secondary structural content.

PLS regression method was used by Dosseau and Pezlot (1990) to detect the structural fractions in infrared (IR) spectral data from 13 proteins. Infrared spectra were measured for α-helix, β-sheet and undetermined structure. The analyses concentrated on the amide I and II regions of IR spectra. Less accurate determinations were obtained from models from amide I band spectra alone than with amide I and II spectra combined. The PLS models from amide I and II for α content determination had an $R^2$ value of 0.95 , β-sheet content had 0.96 and undetermined content had 0.57. The PLS models from the amide I band alone had an $R^2$ value of 0.79 for α-content determination, 0.79 for β-sheet content and 0.57 for undetermined content.

IR spectra of 18 proteins with known X-ray crystal structure were analysed (Lee *et. al.,* 1990) using factor analysis to investigate which spectral ranges within the amide regions most accurately estimate structural content. The group had high accuracies with $R^2$ values of 0.995 and 0.989 for α-helix and β-sheet respectively from normalised IR data. Analyses on $2^{nd}$ derivative IR data yielded slightly lower correlation values of 0.991 and 0.963 for α-helix and β-sheet respectively.

A study by Oberg *et. al*. (2004) involved the comparison of protein structural determination results from CD and IR of 50 proteins with known X-ray crystallography structural references. The group's studies aimed to predict fractions of the structural assignment classes of α-helix β-sheet, turns and others and to investigate the structural information in amide I and II regions. The IR amide I region, analysed by the PLS algorithm, showed root mean square (RMS) values of 8.97, 6.91 and 9.66 for α-helix, β-sheet and other respectively. The

IR amide II prediction analysis showed less accuracy with 7.53, 10.83 and 10.43 RMS values. The group's results also showed that the CD and IR methods could complement each other as analyses on IR+CD combined spectral data produced more accurate results than analyses on separate IR and CD spectral data. Their analyses also showed lower correlation coefficients values where α-helix and β-sheet were further broken down into their categories for example $3_{10}$ helix, π helix, anti-parallel β-sheet, parallel β-sheet.

A method was developed by Provencher and Glockner (1981) in which one circular dichroism spectrum at a time is analysed as a linear combination of 16 CD spectra. Provencher and Glockner (1981) reported correlation coefficient values $R^2$ and root mean square error values (in brackets) from circular dichroism for α-helix, β-sheet and other respectively as follows; 0.96(5%), 0.94(6%) and 0.49(11%).

A study using Raman spectroscopy (Bussian and Sander, 1989) obtained $R^2$ values of 0.95 for α-helix and 0.88 for β-sheet. An earlier analysis of Raman amide I (Williams, 1983) reported $R^2$ values of 0.97, 0.97 and 0.37 for α-helix, β-sheet and undetermined respectively. While the analysis of Raman amide III (Williams, 1986) showed $R^2$ values of 0.98 for α-helix, 0.97 for β-sheet and -0.08 for undetermined which are higher than our determination $R^2$ values from Raman amide III band of 0.78, 0.76, 0.75 for α-helix, β-sheet and other respectively.

### 3.7 Related Chemometrics Methods

PCA decomposes the covariance matrix of N dimensions of the input matrix X into its N eigenvectors and eigenvalues. The eigenvectors are perpendicular to each other and they provide information about patterns in the data. The eigenvectors of the covariance matrix are ordered by eigenvalues, from highest to lowest. The eigenvector with the highest eigenvalue

is the first principle component of the data set. The smaller the eigenvalue, the less significant the component and these smaller eigenvalues can be discarded. As such, from a data set of N dimensions, N eigenvectors and eigenvalues are calculated, then the first *v* eigenvectors are chosen and subsequently the final data set has on *v* dimensions thereby reducing the dimensionality of the data.

PCA is a widely used analysis method for the analysis of spectra as discussed in chapter 2 section 3. PCA software is widely available and easy to access and also has the advantage or reducing the multidimensions of the data. The property of orthogonality of eigenvectors eliminates the problem of high collinearity between spectral data points which can lead to a degradation of the prediction accuracy.

Classification tree algorithms have been developed by various groups. Among the tree-based programs are CART and QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001) and C4.5 (Quinlan and Kohavi, 1999; Quinlan, 1993). QUEST and CRUISE use a bivariate linear discriminate model where the model is fitted between all pairs of data points and the pair with the smallest estimated misclassification cost is selected. Each node is split using univariate split selection rules reported in Kim and Loh (2001). CART splits nodes recursively with the aim of making the data more homogenous until the full tree is grown. Homogeneity at a node *q* can be defined by the Gini index (Breiman *et.al.*, 1984).

The Gini index is a split criterion that favours the reduction of impurity at a node. C4.5 decision tree algorithm uses conditions of ‘if-then’ rules as the tree grows. At each node, the split criterion is based on an ‘if’ test in which, if the sample at the node satisfies given conditions, the next node is the left node otherwise, it is the right node. The direction the tree

grows at each node is determined by the result at the previous node until a leaf node is reached. The predicted class is specified by the leaf node.

k-Nearest Neighbour (k-NN) (Friedman *et.al*., 1975; Zouhal *et.al*., 1998) is a learning algorithm that classifies a sample by taking into account the class of *k* neighbouring samples that are in closest proximity. The class to which the majority in the sample belong to is returned as the predicted class for the new observation. Euclidean distance is used as a measure of similarity distance between different observations by calculating the sum of the squared distances for all the data points.

Artificial Neural Networks (Bishop, 1995; Fausett, 1994) are systems of information processing nodes whose structure and function are based on the biological nervous system. A neural network consists of a set of neurons with each neuron connected to several other nodes. Each connected neuron has an associated weight. The network classifier is learned by adjusting the weights. The inputs are multiplied by the weights then summed up, the output is subject to a threshold, the output belongs to class +1 or -1 depending on whether it is above or below the set threshold. Neural networks can be multilayered, where each layer combines the preceding classifiers. Two layer network outputs as a binary classifier. A multilayer network generalises in high dimensions using hyperplanes to bisect the feature space.

Radial Basis Function (RBF) networks use radial basis functions for classification instead of hyperplanes. An RBF network has a maximum value at the mean of a given set of vectors and the value reduces moving away from this mean centre point. This point is used as is used as a reference point for how far away a data point is from the centre of the RBF for the node. The output layer of the RBF network combines the response from the individual RBF nodes.

Artificial Neural Networks can solve complex systems of nonlinear equations though they can be computationally expensive.

Naive Bayes (Lee, 2010) is a classification technique for which the probability that an observation belongs to a class is calculated. The outcome is the class with the highest probability. The naive Bayes classifier assumes that a given class *x* and the data features are independent. Given a class *i* and data vectors $x_1,x_2,...,x_m$, the probabilistic model for a naive Bayes classifier is $P(C_i| x_1,x_2,...,x_m)$, the probability that data item $x_m$ belong to a class *i*.. The prior probability $P(C_{+1})$ can be calculated as the ratio of the number of sample in class +1 to the total number of samples. The maximum likelihood function can be written as $\Pi_{j=1}^{m} p(x_j|C)$. A new observation is classified to the class with maximum posterior probability.

Chemometrics methods have been applied to a wide number of areas. For example, the Naive Bayes classifier has been used in text classification (McCallum and Nigam 1998; Su *et.al.*, 2011) and drug target studies (Bender *et.al.*, 2004; Xia *et.al.*, 2004; Klon *et.al.*, 2004) . Reinhardt and Hubbard (1998 ) applied Neural Networks to the prediction of the subcellular location of proteins. Binary classification ANN was used by Griadeck *et.al.* (1997, 2004) to distinguish basal cell carcinorma cells from normal samples based on Raman spectra. Nijssen *et. al.* (2002) used PCA on Raman spectra to reduce data dimensions and then applied cluster analysis to identify carcinoma tissue from normal dermis and epidermis tissue. A multidimensional type of clustering was used by the group Zhu *et.al.* (2006) to class 80 ROA spectra from proteins into their structural classes. The prediction of subcellular location of proteins has been carried out using SVM analyses on protein sequences (Chou *et.al.*, 2002; Park *et.al.*, 2003; Garg *et.al.*, 2005; Matsuda *et.al.*, 2005).

Sequence search based methods have been used to predict structure. Protein fold class can be inferred by identifying proteins of similar sequence or similar structure. Structure is conserved more than sequence, hence similarity in sequence is not always an indication of similar fold class as proteins can have similar structures when sequence similarity has been lost. The methods discussed above can be applied to the investigation of protein structure in the absence of sequence similarity. There is always a need to expand ways in which protein structure can be determined. The use of chemometrics analyses on vibrational spectra has been shown to be a useful probe in areas such as cancer diagnostics, quality control and so following on from the various research that showed the capability of such studies. This thesis reports on the application of PLS regression, Random Forest Classification, SVM regression and classification to determine protein fold class and secondary structure from Raman and ROA. These analyses are discussed in the following chapters.

## 3.8 References


Abdi, H. (2007). Partial least square regression (PLS regression). In N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage. pp. 740-744

Ahmed A.A., Yi S., Hongchi S. (2005) Variants of Multidimensional Scaling for Node Localization. Proceedings of the 11th International Conference on Parallel and Distributed Systems (ICPADS'05)

Bender, A., Mussa, H. Y.,Glen, R. C., Reiling, S. (2004) Molecular similarity searching using atom environments, information-based feature selection, and a Naive Bayesian classifier. *Journal of Chemical Information and Computer Science* **44**, 170-178

Bishop C.M. (2006) Pattern Recognition and Machine Learning, 1st ed., Springer Publishing: New York

Boulesteix A. L., Strimmer K. (2005) Predicting transcription factor activities from combined analysis of microarray and chip data: A partial least squares approach. *Theoretical Biology and Medical Modeling* **2**, 23

Boulesteix A.L., Strimmer K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* **8**, 32-44

Breiman L., Friedman J.H., Olsen R.A., Stone C.J.,(1984) Classification and Regression Trees, Wadsworth, Belmont

Breiman, L. (1996a). Bagging predictors. *Machine Learning* **26**, 123–140

Breiman, L. (1996b). Out-of-bag estimation, ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Burges C.J.C.,(1998) A tutorial on Support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**, 121-167

Bussian B.M. and Sander C. (1989) How to determine protein secondary structure in solution by Raman spectroscopy: practical guide and test case DNase I. *Biochemistry* **28**, 4271-4277

Carrascal L.M., Galva´n I., Gordo O. (2009) Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* **118**, 681-690

Chen P., Lin C. and Scholkopf B.,(2007) A tutorial on υ-support vector machines. (www.csie.ntu.edu.tw/~cjlin/papers/nusvmtutorial.pdf)

Chih-Wei H., (2008) A practical guide to support vector classification. (www.csie.ntu.edu.tw/~cjlin)

Chou, K. C., Cai, Y. D.,(2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry* ***277***, 45765-45769.

Chun H. (2008) Sparse partial least squares regression for simultaneous dimension reduction, UMI Microform, Ann Arbor, MI

Datta S. (2001) Exploring relationships in gene expressions: a partial least squares approach. *Gene Expression* **9**, 257–264

Dousseau F. And Pezolet M. (1990) Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods. *Biochemistry* **29**, 8771-8779

Farag A. And Mohamed R.M. (2004) Regression using Support vector machines:Basic Foundations. (www.cvip.uofl.edu)

Fausett L.,(1994) Fundamentals of Neural Networks: architectures, algorithms and applications. Prentice Hall, Upper Saddle, NJ

Friedman, J. H., Baskett, F., Shustek, L. J.,(1975) An algorithm for finding nearest neighbors *IEEE Transactions on Information Theory* **24**, 1000- 1006

Garg, A., Bhasin, M., Raghava, G. P. (2005) Support vector machine based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of Biological Chemistry* ***280***, 14427-14432

Geladi P. and Kowalski B. (1986), Partial least squares regression: A tutorial. *Analytica Chimica Acta* **185**, 1-17

Gniadecka M., Wulf H.C., Mortensen N.N., Nielsen O.F., Christensen D.H.,(1997) Diagnosis of Basal Cell Carcinoma by Raman Spectroscopy. *Journal of Raman Spectroscopy* **28**, 125-129

Gniadecka M., Peter Alshede Philipsen P. A., Sigurdsson S, Wessel S, Nielsen O. F., Christensen D.H., Hercogova J, Rossen K, Thomsen H.K., Gniadecki R., Lars Kai Hansen L.K., Wulf H. C. (2004) Melanoma Diagnosis by Raman Spectroscopy and Neural Networks: Structure Alterations in Proteins and Lipids in Intact Cancer Tissue. *The Journal of Investigative Dermatology* **122**, 443-449

Groenen P.J.F. and van de Velden M. (2004) Multidimensional Scaling. *Econometric Institute Report.*

Hennessey J.P. and Johnson W.C. (1981) Information content in the circular dichroism of proteins. *Biochemistry* **20**,1085-1094

Höskuldsson A. (2004) PLS regression and the convariance. http://www.acc.umu.se/~tnkjtg/Chemometrics/Editorial

Huebner W. And Rahmelow K. (1996) Secondary structure determination of proteins in aqueous solution by infrared spectroscopy: A comparison of multivariate data analysis methods. *Analytical Biochemistry* **241**,5-13

Jaworska N. and Chupetlovska-Anastasovav A. (2009) A Review of MDS and its utility in Various Psychological Domains.*Tutorials in Quantitative Methods for Psychology*,**5** 1-10

Jollife I.T. (1986) Principal Components Analysis, Springer Verlag, New York

Kim H., Loh W.-Y. (2001) Classification trees with unbiased multiway splits. *Journal of the American Statistical* Association **96**, 589-604

Klon, A. E., Glick, M.,Thoma, M., Acklin, P., Davies, J. W.,(2004)Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *Journal of Medicinal Chemistry* **47**, 2743-2749

Kohavi R., Quinlan R. (1999) Decision Tree Discovery. (http://ai.stanford.edu/~ronnyk/treesHB.pdf)

Lee D.C., Haris P.I., Chapman D., Mitchell R.C. (1990) Determination of protein secondary structure using factor analysis of infrared spectra. *Biochemistry* **29**, 9185-9193

Lees J.G., Miles A.J., Janes R.W., Wallace B.A. (2006) Novel methods for secondary structure determination using low wavelength (VUV) circular dichroism spectroscopic data. *BMC Bioinformatics* **7**,507-517

Lee J.K.(2010) Statistical Bioinformatics for Biomedical and Life Science researchers, John Wiley & Sons, Hoboken, New Jersey

Loh W.-Y., Shih Y.S. (1997) Split selection methods for classification trees. *Statist. Sinica* **7**, 815-840

Matsuda, S., Vert, J. P., Saigo, H., Ueda, N., Toh, H., Akutsu, T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science* **14**, 2804-2813

Martens H., Naes T. (1989) Multivariate Calibration, Wiley, London

McCallum A., Nigam K. (1998) A comparison of event models for Naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization.*

McIntosh A.R., Bookstein F.L., Haxby J.V., Grady C.L. (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* **3**, 143-157

Müller K., Mika S. Rätsh G., Tsuda K. and Sch ölkopf B. (2001) An Introduction to Kernel-Based Learning Algorithm. IEEE *Transactions on Neural Networks*. **2**, 181-201

Naud A. (2006) An accurate MDS-based algorithm for the visualization of large multidimensional datasets. (www.phys.uni.torun.pl/˜naud)

Navea S., Tauler R., de Juan A. (2005) Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure, *Analytical Biochemistry* **336**, 231–242

Nijssen A., Schut T. C. B., Heule F, Caspers P. J., Hayes D P, Neumann M. H. A, and Puppels G. J. (2002) Discriminating Basal Cell Carcinoma from its Surrounding Tissue by Raman Spectroscopy *The Journal of Investigative Dermatology* **119**, 64-69

Orberg K.A., Ruysschaert J., Goormaghtigh E. (2004) The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *European Journal of Biochemistry* **271**, 2937-2748

Park, K. J. (2003) Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics **19***, 1656-1663

Provencher S. W. And Glockner J. (1981) Estimation of globular protein secondary structure from circular dichroism *Biochemistry* **20**,33-37

Quinlan R. (1993) C4.5 Programs for Machine Learning, Morgan Kaufmann Publishers Inc, San Mateo, CA

Reinhardt A., Hubbard T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research* **26**, 2230-2236.

Schölkopf B., Oldenbourg R. (1997) Support Vector Learning, Verlag, München

Schölkopf B. (2005) (www.ccs.neu.edu/home/vip/teach/MLcourse/lectures/aop05_scholkopf_km.pdf)

Smola A.J., Schoelkopf B. (2004) A tutorial on support vector regression. *Journal of Statistics and Computing* **14**, 199-222

Su J., Sayyad-Shirabad J., Matwin S. (2011) Large scale text classification using semi-supervised multinomial Naive Bayes. *Proceedings of the 28$^{th}$ International Conference on Machine Learning*

Tao W., Hangen H., Dewen H. (2002) On the separability of kernel functions. *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02)*

Tobias R.D. (1997) An Introduction to partial least squares regression, TS-509, SAS Institute Inc., Cary, N.C.

Tsuda K. , Vert J. (2004) Kernel methods in computational Biology,MIT Press,Cambridge

Vapnik V. (1999) The Nature of Statistical Learning Theory, Springer-Verlag, New York

Williams R. W. (1983) Estimation of protein secondary structure from the laser Raman amide I spectrum.*Journal of Molecular Biology* **166**,581-603

Williams R. W. (1986) Protein secondary structure analysis using Raman amide I and amide III spectra. *Methods in Enzymology* **130**, 311-331

Wold H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiaah P.R. (Ed.) *Multivariate Analysis*. Academic Press, New York

Xia X., Maliski E. G., Gallant P., Rogers D.(2004) Classification of kinase inhibitors using a Bayesian model. *Journal of Medicinal Chemistry* **47**, 4463-4470

Zhou, G. P., Assa-Munt, N. (2001) Some insights into protein structural class prediction. *Proteins: Structure Function Genetics* **44**, 57-59

# 4 Data and Methods

## 4.1 Datasets

A dataset of ROA and Raman spectra from proteins and viruses was compiled from previous studies published in literature (Zhu *et. al*, 2005; McColl et. al., 2003). The samples were prepared with typical concentrations of 10-100 mg/ml and data collection times of 4-24 hours. Both ICP and SCP ROA spectra were used, however for the analyses discussed in this thesis, all spectra were processed the same way (i.e. binning and scaling), as in the far-from-resonance limit ICP and SCP spectra are identical (Nafie, 1997). In cases where a protein has several spectra available, the spectrum obtained at conditions closest to physiological was retained and the others were excluded. The Raman and ROA data include an array of α-helical, β- strand and disordered poly-L-amino acids. The classes of the proteins in the data set were delimited by the known structural classes as reported by SCOP (Murzi *et.al.*, 1995) using a 3 state scheme of α-helix, β-sheet and other. For the SVM analyses, the polypeptides and proteins were split into training and validation data sets as the algorithm required. Table 4.1 is a summary table of the protein data used.

**Table 4.1 Table of protein names, pdb codes (where available), structural information. The analyses in which the proteins were used and the type of spectra are indicated by the "▪" mark in the respective column.**

| | PDB Code | Number of residues | α % SCOP | β % SCOP | O % SCOP | SVM | PLS | RF | ROA | RAMAN |
|---|---|---|---|---|---|---|---|---|---|---|
| creatine kinase (rabbit) | 1i0e | 365 | 31 | 13.2 | 54.2 | ▪ | ▪ | ▪ | ▪ | |
| filamentous bacteriophage fd | 1fdm | 50 | 50 | 0 | 50 | ▪ | ▪ | ▪ | ▪ | ▪ |
| serum albumin (human) | 1e78 | 578 | 70.2 | 0 | 29.1 | ▪ | ▪ | ▪ | ▪ | ▪ |
| filamentous bacteriophage Pf1 | 1pfi | 46 | 91.3 | 0 | 8.7 | ▪ | ▪ | ▪ | ▪ | ▪ |
| prion protein (ovine,90-230) | 1qm1 | 104 | 51 | 3.8 | 45.2 | ▪ | ▪ | ▪ | ▪ | ▪ |
| S100B (rabbit) | 1uwo | 91 | 59.3 | 0 | 40.7 | ▪ | ▪ | ▪ | ▪ | |
| S100A6 (calcyclin, rabbit) | 1k96 | 89 | 69.7 | 0 | 30.3 | ▪ | ▪ | ▪ | ▪ | ▪ |
| cowpea mosaic virus (empty protein capsid) | 1ny7 | 558 | 5.4 | 41 | 49.6 | ▪ | ▪ | ▪ | ▪ | ▪ |
| serum amyloid P component (human) | 1sac | 204 | 3.9 | 44.1 | 50.5 | ▪ | ▪ | ▪ | ▪ | |
| invertase | 2ac1 | 537 | 2.6 | 45.1 | 49.5 | ▪ | ▪ | ▪ | ▪ | ▪ |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| P.69 pertactin (Bordetella pertussis) | 1dab | 539 | 0 | 59.9 | 45.6 | ■ | ■ | ■ | ■ | |
| pepsin (porcine) | 1am5 | 324 | 9 | 41.7 | 45.7 | ■ | ■ | ■ | ■ | |
| serum amyloid protein(domain a) | 1lgn | 204 | 4.4 | 44.6 | 49.5 | ■ | ■ | ■ | ■ | |
| satellite tobacco mosaic virus | 1a34 | 147 | 0 | 39.5 | 54.4 | ■ | ■ | ■ | ■ | ■ |
| trypsin (bovine) | 1k1l | 223 | 7.2 | 32.3 | 59.2 | ■ | ■ | ■ | ■ | |
| ovomucoid (turkey) | 1m8c | 56 | 17.9 | 16.1 | 66.1 | ■ | ■ | ■ | ■ | |
| lactoferrin (human) | 1fck | 691 | 29.4 | 18.1 | 47.9 | ■ | ■ | ■ | ■ | ■ |
| MS2 virus (empty protein capsid) | 2ms2 | 129 | 16.3 | 46.5 | 34.9 | ■ | ■ | ■ | ■ | ■ |
| ribonuclease A (bovine) | 1afu | 124 | 17.7 | 37.4 | 45.2 | ■ | ■ | ■ | ■ | ■ |
| subtilisin Carlsberg (Bacillus licheniformis) | 1ndq | 269 | 30.9 | 18.2 | 50.9 | ■ | ■ | ■ | ■ | |
| ubiquitin (bovine) | 1ubq | 76 | 15.8 | 31.6 | 46.1 | ■ | ■ | ■ | ■ | |
| Bowman-Birk inhibitor (soybean) | 1pi2 | 61 | 0 | 23 | 77 | ■ | ■ | ■ | ■ | ■ |
| α-1-acid glycoprotein (bovine) | 3kq0 | 175 | 18.9 | 41.1 | 33.1 | ■ | ■ | ■ | ■ | ■ |
| ribonuclease B (bovine) | 1rbb | 124 | 17.7 | 33.1 | 46.8 | ■ | ■ | ■ | ■ | ■ |
| tobacco mosaic virus | 1vtm | 158 | 34.8 | 4.4 | 59.5 | ■ | ■ | ■ | ■ | ■ |
| α-lactalbumin (bovine) | 1f6s | 122 | 31.1 | 8.2 | 47.5 | ■ | ■ | ■ | ■ | ■ |
| filamentous bacteriophage M13 | 2cpb | 50 | 74 | 0 | 26 | ■ | ■ | ■ | ■ | ■ |
| aldolase (rabbit) | 2ald | 363 | 42.1 | 13.8 | 41.9 | ■ | ■ | ■ | ■ | ■ |
| lysozyme (hen) | 1lsc | 129 | 30.2 | 6.2 | 52.7 | ■ | ■ | ■ | ■ | ■ |
| ovomucoid (human) | 1m8c | 56 | 17.9 | 16.1 | 66.1 | ■ | ■ | ■ | ■ | ■ |
| lysozyme (human) | 1gaz | 130 | 30 | 7.7 | 53.1 | ■ | ■ | ■ | ■ | ■ |
| immunoglobulin G (human) | 1ig2 | 455 | 3.3 | 40.9 | 52.5 | ■ | ■ | ■ | ■ | ■ |
| insulin (bovine, monomeric, low PH) | 1zeh | 51 | 52.9 | 7.8 | 35.3 | ■ | ■ | ■ | ■ | |
| β-chymotrypsin (bovine) | 4cha | 239 | 7.5 | 31.8 | 58.6 | ■ | ■ | ■ | ■ | |
| amylase (Bacillus licheniformis) | 1kgu | 496 | 17.9 | 17.9 | 57.3 | ■ | ■ | ■ | ■ | ■ |
| avidin (hen) | 1rav | 124 | 0 | 50 | 42.7 | ■ | ■ | ■ | ■ | ■ |
| bovine beta lactalbumin | 1b8e | 152 | 8.6 | 44.7 | 40.8 | ■ | ■ | ■ | ■ | ■ |
| concanavalin A (jack bean) | 3cna | 237 | 0 | 40.5 | 59.5 | ■ | ■ | ■ | ■ | |
| trypsinogen (bovine) | 2tga | 223 | 7.2 | 32.3 | 59.2 | ■ | ■ | ■ | ■ | ■ |
| antifreeze protein type III (arctic flounder) | 1b7i | 66 | 6.1 | 12.1 | 74.2 | ■ | ■ | ■ | ■ | |
| metallothionein (rabbit) | 4mt2 | 62 | 0 | 0 | 100 | | ■ | ■ | ■ | ■ |
| full length prion protein (ovine,23-230) | 1y2s | 113 | 38.3 | 3.5 | 46 | ■ | | ■ | ■ | ■ |
| bovine beta lactalbumin pH 2 | 1dv9 | 162 | 9.9 | 36.4 | 51.9 | ■ | ■ | ■ | ■ | |
| lysozyme (equine) | 2eql | 129 | 31.8 | 9.3 | 51.2 | ■ | | | ■ | |
| filamentous bacteriophage Ike | 1ifl | 53 | 96.2 | 0 | 3.8 | | | | ■ | |
| lysozyme (equine, A-state) | | | | | | ■ | | | ■ | |
| β-lactoglobulin (bovine, pH=2.0) | | | | | | ■ | | | ■ | |
| β-lactoglobulin (bovine, pH=6.5) | | | | | | ■ | | | ■ | |
| β-lactoglobulin (bovine, pH=9.0) | | | | | | ■ | | | ■ | |
| ubiquitin (bovine) | | | | | | ■ | | | ■ | ■ |
| cowpea mosaic virus (with RNA 1) | | | | | | ■ | | | ■ | |
| cowpea mosaic virus (with RNA 2) | | | | | | ■ | | | ■ | |
| ovalbumin (hen) | | | | | | ■ | | | ■ | |
| α-lactalbumin (human) | | | | | | ■ | | | ■ | ■ |
| α-synuclein (human) | | | | | | | | | ■ | |
| α-lactalbumin (bovine, A-state) | | | | | | ■ | | | ■ | |
| lysozyme (hen, reduced) | | | | | | ■ | | | ■ | |
| lysozyme (human, pre-fibillar intermediate) | | | | | | ■ | | | ■ | ■ |
| prion protein (ovine, 94-233, reduced) | | | | | | ■ | | | ■ | |
| narcissus mosaic virus | | | | | | ■ | | | ■ | |
| papaya mosaic virus | | | | | | ■ | | | ■ | |
| potato virus X | | | | | | ■ | | | ■ | |
| regulatory protein 2 (Streptococcus pyogenes) | | | | | | ■ | | | ■ | |
| A-gliadin (wheat) | | | | | | ■ | | | ■ | |
| tobacco mosaic virus dimer | | | | | | ■ | | | ■ | |
| tobacco rattle virus | | | | | | ■ | | | ■ | |
| prion protein (mouse, 23-231) | | | | | | ■ | | | ■ | |
| ovomucoid (hen) | | | | | | ■ | | | ■ | ■ |
| ribonuclease B (bovine) | | | | | | ■ | | | ■ | ■ |
| β-sheet poly(L-lysine) | | | | | | ■ | | | ■ | ■ |
| ABA-1 allergen (Ascaris lumbricoides) | | | | | | ■ | | | ■ | ■ |
| α-casein (bovine) | | | | | | ■ | | | ■ | ■ |
| β-casein (bovine) | | | | | | ■ | | | ■ | ■ |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| κ-casein (bovine) | | | | | | ■ | | | ■ | |
| ω-gliadin (wheat) | | | | | | ■ | | | ■ | |
| β-synuclein (human) | | | | | | ■ | | | ■ | |
| γ-synuclein (human) | | | | | | ■ | | | ■ | |
| T-A-1 peptide (wheat glutenin subunit) | | | | | | ■ | | | ■ | ■ |
| tau46 (human P301L mutant) | | | | | | ■ | | | ■ | |
| tau46 (human) | | | | | | ■ | | | ■ | |
| OOAAAAAAAOO peptide | | | | | | ■ | | | ■ | |
| phosvitin (hen) | | | | | | ■ | | | ■ | |
| disordered poly(L-glutamic acid) | | | | | | ■ | | | ■ | |
| disordered poly(L-lysine) | | | | | | ■ | | | ■ | |
| collagen structure | | | | | | ■ | | | ■ | |
| α-1-acid glycoprotein (bovine) | | | | | | ■ | | | ■ | |
| A-gliadin in 60% MeOH | | | | | | ■ | | | ■ | |
| rF1 antigen | | | | | | ■ | | | ■ | |
| ribonuclease A (bovine, reduced) | | | | | | ■ | | | ■ | |
| disordered poly(L-alanine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-alanine) | | | | | | ■ | | | ■ | |
| β-sheet poly(L-alanine) | | | | | | ■ | | | ■ | |
| α-helical poly(benzyl) | | | | | | ■ | | | ■ | |
| disordered poly(L benzyl) | | | | | | ■ | | | ■ | |
| α-helical poly(L-glutamic acid) | | | | | | ■ | | | ■ | |
| disordered poly(L-histidine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-histidine) | | | | | | ■ | | | ■ | |
| disordered poly(L-leucine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-leucine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-lysine) | | | | | | ■ | | | ■ | |
| disordered poly(L-ornathine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-ornathine) | | | | | | ■ | | | ■ | |
| disordered poly(L-proline) | | | | | | ■ | | | ■ | |
| disordered poly(L-tryptophan) | | | | | | ■ | | | ■ | |
| α-helical poly(L-tryptophan) | | | | | | ■ | | | ■ | |
| disordered poly(L-threonine) | | | | | | ■ | | | ■ | |
| disordered poly(L-tyrosine) | | | | | | ■ | | | ■ | |
| α-helical poly(L-tyrosine) | | | | | | ■ | | | ■ | |

SVM=Support Vector Machines; PLS =Partial Least Squares regression; RF=Random Forests

## SVM analyses

In the SVM classification analyses, a set of 67 ROA spectra of proteins and 53 Raman spectra of proteins including denatured proteins and polypeptides was used. The ROA models were built from a training set of 35 proteins comprised of 9 αβ proteins, 9 α-helix proteins, 8 β- sheet proteins and 9 'other' proteins. The ROA validation set of 32 spectra had 8 αβ proteins, 5 α-helix proteins, 8 β- sheet proteins and 11 'other' proteins. The denatured proteins and polypeptides were removed for subsequent SVM regression analyses to match proteins to their SCOP references. The models from ROA spectra for SVM regression analyses were built from 25 proteins (8 αβ proteins, 7 α-helix proteins, 8 β- sheet proteins and 2 'other' proteins) and the models from Raman spectra were built from 15

proteins (4 αβ proteins, 3 α-helix proteins, 5 β- sheet proteins and 3 ‘other’ proteins). The validation set for the SVM regression models had 17 proteins for ROA :- 6 αβ proteins, 3 α-helix proteins, 5 β- sheet proteins and 3 ‘other’ proteins and 11 proteins for Raman:- 6 αβ proteins, 4 α-helix proteins, 1 β- sheet proteins.

**PLS regression and Random Forests analyses**

A dataset of 44 ROA and 24 Raman spectra were used for PLS regression and Random forest analysis. The nature of these algorithms did not need the splitting of the datasets as for the SVM analyses. These algorithms are discussed in chapter 3. The ROA dataset comprised 15 αβ proteins, 9 α-helical proteins, 14 β-sheet proteins and 4 ‘other’ proteins. The Raman dataset comprised 8 αβ proteins, 6 α-helical proteins, 7 β- sheet proteins and 3 ‘other’proteins. A table of all the proteins used in the analyses can be seen in Appendix A.

## 4.2 Dataset representation

Pattern recognition algorithms require input data to be represented as vectors. Each vector consisted of a class label (-1 or +1) or percentage fraction followed by the sparse representation of the input vector of real numbers, in this case, ROA spectral intensities. The intensities of the spectra formed the input matrix X. The class labels [1,+1] or the percentage structural fraction made up the Y response vector. The spectral data were in $(x_i, y_i)$ pairs where $x_i$ represented the wave numbers and $y_i$ the intensities, derived from each spectrum. For a variable range of intensities, the mean values of intenstities of each bin range are used as the features in the vectors. The figure 4.1 shows a simple illustration of spectral files being processed to output vectors which would constitute input matrix X.

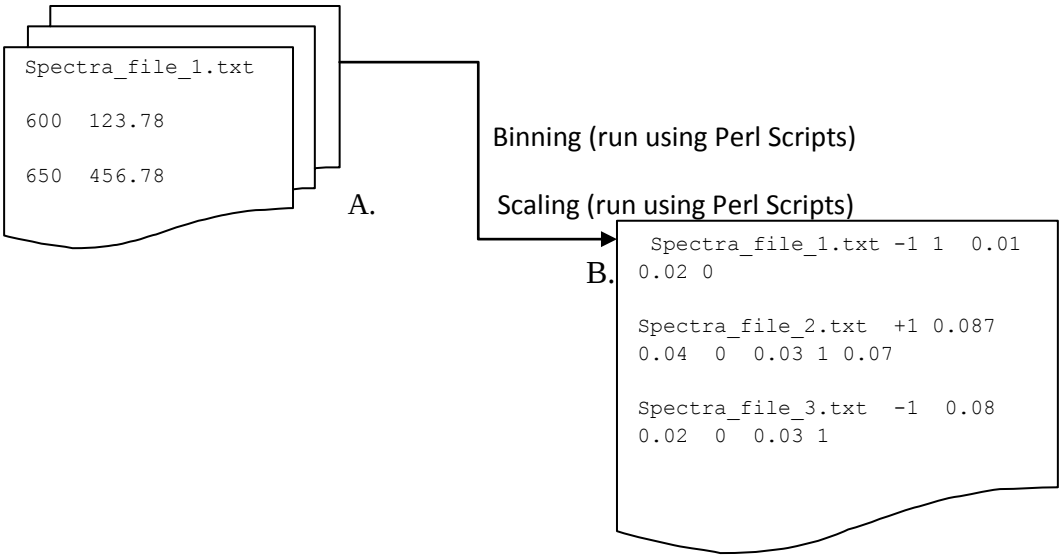



**Figure 4.1 An illustration showing how the raw spectra data were encoded in the feature vectors used in the machine learning algorithms. The spectra are contained in text files (A). Each text file has rows of wavelength and spectral intensity pairs. Each file is read in by the binning script (Appendix B) and then the scaling script (Appendix D). The scaling scripts outputs each file to a single file (B) as a single row. These rows of spectral files (B) make up the input vectors for the machine learning algorithms.**

## 4.3 Data Processing

### 4.3.1 Binning

Spectral intensities from across the whole range of the spectra were first grouped into bins of widths 10, 20 and 100 $cm^{-1}$ wavenumbers using a Perl script *binAvg.pl* (Appendix B) by averaging within 10, 20, 100 $cm^{-1}$ intervals. The binning reduces the effects of experimental noise in the spectra. The average intensities in bins of 10 $cm^{-1}$ width were produced from the full spectra from 620 to 1850 $cm^{-1}$, giving 123 bins in total. Each bin typically included 3 or 4 data points. Bin widths of 20 $cm^{-1}$ and 100 $cm^{-1}$ bin widths were also investigated giving 62 and 13 bins in total respectively. Bin widths of 20 $cm^{-1}$ typically had 7 or 8 data points while bin widths of 100 $cm^{-1}$ had 36 data points in them. Decreasing the bin width much below 10 $cm^{-1}$ was not possible, as data would not always be available for each bin. These bin intensities constituted the vectors in the input matrices used in the machine learning algorithms.

### 4.3.2 Range Selection

Further investigations involved systematically testing the amide regions I, II and III of the spectra to analyse which sections yield more structural information. Combinations of the amide regions were also analysed; I+II, I+III, II+III, I+II+III. The selection of the amide ranges was scripted in *ranges.pl* (Appendix C). Analyses were performed on the following regions; amide I (1600-1700 $cm^{-1}$), amide II (1540-1600$cm^{-1}$) and amide III (1200-1340$cm^{-1}$), a combination of these, the backbone stretch modes (850-1100 $cm^{-1}$) . Figure 4.2 illustrates the subdivisions of spectral data.

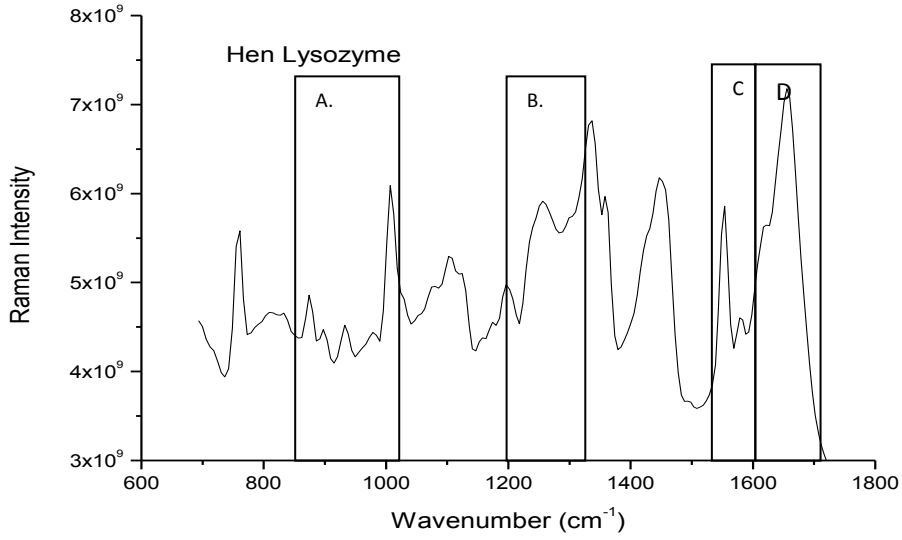



**Figure 4.2 A figure of a Raman spectrum with the subdivisions of the spectral data used in the analyses marked. The analyses were performed on the following regions; A. the backbone stretch modes (850-1100 $cm^{-1}$); B. amide III (1200-1340$cm^{-1}$); C. amide II (1540-1600$cm^{-1}$); D. amide I (1600-1700 $cm^{-1}$).**

### 4.3.3 Scaling

Scaling is important as it avoids attributes in greater numeric ranges dominating those in the smaller numeric ranges. Scaling was achieved by Perl script *svmscale.pl* (Appendix D). For ROA spectra, each attribute in the vector was linearly scaled to the [-1,+1] range. Scaling to the range of -1,+1 means that the lowest value allowed in each vector is -1 and the highest

value allowed is +1. Whilst for Raman the intensities for each spectrum were scaled to between [1,0]. The linear scaling method used is shown below:-

$$Scaling = \frac{2(z_i - mi)}{(Mi - mi)} - 1 \qquad (4.1)$$

where $z_i$ is the attribute in the vector, $Mi$ is the maximal value of the vector, $mi$ is the minimal value of the vector.

## 4.4 Training the Models

### 4.4.1 SVM Models

**Choosing the best C and γ parameters**

The method for training the model is cross validation as it prevents producing a classifier that is too specific to the training dataset. Ideally, we require a classifier that can generalise what it has learnt when it encounters novel data. In a *t*-fold cross validation, the training set is split into *t* subsets with the same distribution of positive or negative examples. Then *t*-1 subsets are arbitrarily selected to train the SVM model and the remaining set that was left out is tested using the model. This is repeated until each set is tested against the other *t*-1. The average accuracy is then reported as the cross validation accuracy.

When cross validation is used to produce a model, grid-search is performed to find the best C and γ parameters. Pairs of (C, γ) are tried and the one with the best cross validation accuracy is selected. Grid-search parallelises the parameter search such that each pair of C and γ is independent as opposed to other methods that use iterative processes. Figure 4.3 below shows a grid-search plot.

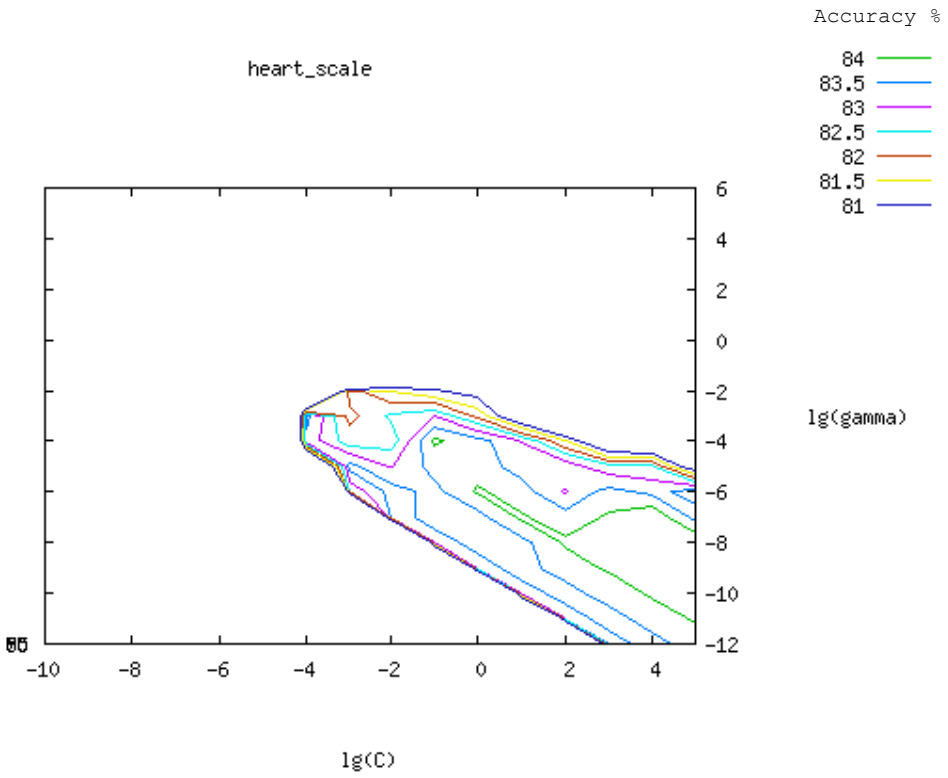



**Figure 4.3 Grid search plot using C= $2^{-10}$…$2^{4}$ and γ=$2^{-12}$..$2^{6}$. An initial search produces a coarse grid. A region with the highest accuracy is noted on the coarse grid and the C and γ ranges in this region are investigated further creating a finer grid search i.e. C= $2^{-4}$…$2^{4}$ and γ=$2^{-8}$..$2^{0}$. This grid plot was taken from LIBSVM example data heart_scale (www.csie.ntu.edu.tw/~cjlin/libsvm).**

The grid-search involves identifying a coarse grid first. After identifying a better region on the grid, a finer grid search on that grid can be conducted. After the best pair is found, the whole training set is trained again to generate the final classifier using the parameters.

While training, the model weights can be adjusted to improve accuracy. The weights are a penalty for misclassification of data. The ratios of the number of samples in the two classes were used to decide which weights to use. As an example, given a dataset with two classes 1 and -1; if 1 has 10 samples and -1 has 20 samples then it is twice as likely that samples in class 1 will be misclassified as class -1 samples. The ratio of the samples (1:-1) is 1:2. So the

penalty for misclassifying a positive (1) sample as a negative (-1) sample will be set to twice the penalty for misclassifying a negative sample as a positive sample.

The Matthews correlation coefficient (MCC) was used to analyse the perfomance of the SVM classification models. The formula of the MCC is as shown;

$$\frac{TP*TN - FP*FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad (4.2)$$

where TP is true positives, TN is true negatives, FN is false negatives and FP is false positves. The MCC is a reliable evaluation of the prediction that takes into account both sensitivity and specificity. Sensitivity is the probability that a protein belongs to a given class when in fact it is a true member of that class. Specificity is the probability that a protein does not belong to a given class when in fact it is not a member of that class. The MCC is an estimate of the measure of similarity between the true classes and the predicted classes (Baldi, 2000). The MCC is always between -1 and +1. A value of +1 indicates a perfect prediction, 0 indicates random predictions and -1 indicates false predictions. The SVM analyses results are discussed in chapter 5.

### 4.4.2 PLS Models

#### PLS Cross Validation

The PLS algorithm uses cross validation and the root mean squared error to validate all possible correlation models. 5-fold cross validation was carried out splitting the data into 5 groups. One group is iteratively removed and the remaining groups produce a model which is validated on the remaining group using the root mean squared error. The model with the least root mean square is chosen. The PLS algorithm used the maximal possible number of

components. In our analyses, the number of components was 10. Models were produced for each of the amide region data and the different combinations as well the whole spectra.

**Performance analysis of the PLS Regression models**

The PLS models' performances were reported using the Pearson's correlation coefficient ($R^2$), the root squared mean (RMS) and the determination enhancement parameter ζ, which is the ratio of RMS and the standard deviation in the observed data. The correlation coefficient $R^2$ was calculated as the squared sum of differences between the observed percentage content and the value predicted by the trained SVM regression models. A low correlation between the predicted value and the observed value of percentage structure content results in very low $R^2$ values.

$$RMS = \sqrt{\sum \frac{(f_p - f_o)^2}{N}} \quad (4.3)$$

where $f_p$ is the predicted fraction of the structural motif, $f_o$ is the observed fraction of the structural motif and N is the number of samples

$$SD = \sqrt{\sum \frac{(f_o - Avg)^2}{N-1}} \quad (4.4)$$

where $f_o$ is the observed fraction of the structural motif, *Avg* is the mean of the observed structural fractions and N is the number of samples.

The root squared mean (RMS) measures the deviation of observed values from predicted values. The determination enhancement parameter ζ is the ratio of RMS and the standard deviation in the observed data. This parameter compares the variation in the predicted variable to the variation in the observed data. It estimates the amount of structural information extracted. The higher the value of the ζ parameter, the higher the determination

accuracy. The results of the PLS regression analyses on Raman and ROA spectra are discussed in chapter 6.

### 4.4.3 Random Forest Models

The Random Forest (RF) analyses were performed using programs developed for MATLAB. Random Forests uses a combination of trees to using random samples from the input data. The number of tress to be grown was set to 500. The number of predictors sampled for splitting at each node was set, by default, to the square root of the size of the two-dimensional input matrix. The RF program in MATLAB has an option that allows setting of class weights. The weights were set roughly on ratios based on the number of samples of each class in the set e.g. for ROA [9 14 15 4] and Raman [6 7 8 3] for α, β, α/β and 'other' respectively.

**Visualisation**

RF proximity distances were fed into Multidimensional Scaling (MDS) to create 2D visualisation plots (Groenen and Velden, 2004; Naud, 2006). MDS maps the similarities between two objects ($p$ ,$q$) into predetermined MDS configuration X which consists of all the objects. The interdistances, $d_{pq}$, of the objects in X are translated from the RF proximity distances, $p_{pq}$, by a loss function. The loss function sums the differences in distances $d_{pq}$ between all the objects in X. To find the best configuration, MDS iterates over all the possible configurations while measuring the best fit between the proximity distances, $p_{pq,}$ and the distances $d_{pq}$. The best fit is the configuration where the proximity distances are as close as possible to the distances $d_{pq}$. The iterative process proceeds until the square error of differences, $p_{pq}$ - $d_{pq}$ $(X)$ is minimised.

**Performance analysis of the Random  Forest models**

The measure of accuracy of the Random Forest models was the percentage of the correctly predicted observation divided by the total number of samples. The Random Forest results are discussed in chapter 7.

## 4.5 References


Groenen P.J.F. and van de Velden M. (2004) Multidimensional Scaling. *Econometric Institute Report.*

MATLAB (http://www.mathworks.com/help/techdoc/)

Murzin A.G., Brenner S. E., Hubbard T., Chothia C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biol*ogy **247**, 536-540

Naud A. (2006) An accurate MDS-based algorithm for the visualization of large multidimensional datasets. (www.phys.uni.torun.pl/˜naud)

McColl I.H, Blanch E.W., Gill A.C., Rhie A. G.O., Ritchie M.A, Hecht L., Nielsen K. and Barron L. D. (2003) A New Perspective on β-Sheet Structures Using Vibrational Raman Optical Activity: From Poly(L-lysine) to the Prion Protein . *Journal of American Chemical Society* **125**,10019-10026

Nafie L. A. (1997) Infrared and Raman vibrational optical activity. *Annual Review of Physical Chemistry* **48**, 357–386

Zhu F., Isaacs N.W., Hecht L., Tranter G. E., Barron L.D. (2006)Raman optical activity of proteins, carbohydrates and glycoproteins. *Chirality* **18**,103-115

# 5. Support Vector Machine (SVM) Classification and Regression analyses of Raman and ROA spectra

## SVM Analyses of Raman and ROA

The aim of the SVM analyses was to investigate the determination, from Raman and ROA spectra, of the fraction of the different structural fold classes α helix, β sheet, αβ and 'other' and the relationship between spectral features and these structural motifs. Raman and ROA spectral bands arise from vibrational modes that are characteristic of the arrangement of atoms in the protein molecule as discussed in chapter 2 on vibrational spectroscopy. Each protein produces a distinct spectrum that distinguishes it uniquely from other proteins. It then follows that we can apply chemometrics methods to try to extract this structural information and group proteins into fold classes α-helix, β-sheet, αβ and other.

The SVM classification algorithm classifies samples by producing a hyperplane that best divides the data projected in feature space into their classes. SVM classification is discussed in further details in chapter 3 on machine learning methods. The classes and structural fractions of the proteins used to build the prediction SVM models were based on the SCOP classification of proteins discussed in section 1.3 of chapter 1 on protein structure. The models were generated for α-helix, β-sheet, αβ and other fold classes. The models allow for fold class prediction on unseen data so proteins whose structure is unknown can have their spectra analysed for prediction of their fold class therefore giving us more information about the protein. The measure used to test the performance rate of the models was the Matthews correlation coefficient (MCC) value, an accuracy value that indicates how close the predicted value is to the observed value. This is discussed in detail in chapter 4 on methods. The higher the MCC value the better the performance of the model.

## 5.1 Data Pre-processing

### 5.1.1 Choosing the Bin Factor

The first step was varying the bin sizes between 10, 20 and 100 $cm^{-1}$ to test for sensitivity and high performance of the models. To achieve this, the Raman and ROA spectra were averaged over 10, 20 and 100 $cm^{-1}$ . The most favourable bin size was the one that produced the most accurate predictive models. Tables 5.1 and 5.2 below show the results for bin sizes 10, 100 and 200 $cm^{-1}$ for ROA and Raman. The results showed that bin size 10 $cm^{-1}$ produced the best performing models. The ROA models for α-helix, β-sheet, αβ and other classes yielded MCC values of 0.62, 0.74, 0.61 and 0.12 respectively for bin size 10 $cm^{-1}$. ROA bin sizes of 20 $cm^{-1}$ showed poorer performances with MCC value of 0.59 for αβ prediction and MCC values of 0.00 for α-helix, β-sheet and other fold classes. ROA bin size of 100 $cm^{-1}$ performed even worse with negative MCC values -0.04, -0.10, -0.35 for the αβ, α-helix and other predictions respectively. The Raman models showed poorer performances with lower MMC values for bin size 10 $cm^{-1}$ of 0.07, -0.13, -0.14 for αβ, α-helix and β-sheet respectively compared to those for ROA for the same fold classes.

The Raman model performances produced very low MCC values however, the other prediction with 10 $cm^{-1}$ bin was the exception yielding an MCC value of 0.63 for the 'other' model which was better than the MCC value from the 'other' model derived from the averaged ROA data of 10 $cm^{-1}$ .Overall, the Raman MCC values from the models generated from data averaged over 10, 20 or 100 $cm^{-1}$ showed worse performance than the ROA models. The bin size of 10cm-1 was chosen to be used throughout all subsequent analyses because it showed the best performance accuracy out of the three bin sizes tested.

**Table 5.1 Performance accuracies for SVM classification models Bins 10, 20 and 100 $cm^{-1}$ for Raman data analyses.**

| **Raman Classification Bins** | | | | |
|---|---|---|---|---|
| | Alpha Beta | AlphaHelix | BetaSheet | Other |
| **Bin 10** $cm^{-1}$ | | | | |
| **MCC** | 0.07 | -0.13 | -0.14 | 0.63 |
| **Bin20** $cm^{-1}$ | | | | |
| **MCC** | -0.01 | -0.17 | -0.13 | 0.20 |
| **Bin100** $cm^{-1}$ | | | | |
| **MCC** | 0.12 | -0.01 | 0.00 | 0.20 |

**Table 5.2 Performance accuracies for SVM classification models Bins 10, 20 and 100 $cm^{-1}$ for ROA data analyses.**

| **ROA Classification Bins** | | | | |
|---|---|---|---|---|
| | Alpha Beta | AlphaHelix | Beta Sheet | Other |
| **Bin 10** $cm^{-1}$ | | | | |
| **MCC** | 0.61 | 0.62 | 0.74 | 0.13 |
| **Bin 20** $cm^{-1}$ | | | | |
| **MCC** | 0.59 | 0.00 | 0.00 | 0.00 |
| **Bin 100** $cm^{-1}$ | | | | |
| **MCC** | -0.04 | -0.10 | 0.53 | -0.35 |

The analyses of the variation of bin sizes (tables 5.1 and 5.2) were the preliminary step, as a separate experiment, to investigate the optimum bin size to use. The data sets were re-arranged i.e. samples that were in the test sets changed to the training set, for the subsequent analyses whose results are presented in tables 5.3 and 5.4. The investigation analyses of optimum bin sizes were done independently from the subsequent analyses of Support Vector Machine classification and regression, Random Forest classification and Partial Least Squares regression and the order of data sets was changed in between analyses in the process of data preparation and pre-processing. However, this should not change the importance of bin size $10cm^{-1}$ giving the most accurate results. The bin size of 10 $cm^{-1}$ was used in subsequent analyses for Support Vector Machine classification and regression, Random Forest classification and Partial Least Squares regression. However the MCC values reported in 5.3 and 5.4 varied because of the re-arrangement (shuffling) of data samples during script testing and data processing.

## 5.2 Results and Discussion

### 5.2.1 SVM Classification Analyses Results of ROA and Raman spectra

On the whole, the analyses for whole spectra from ROA yielded better results than the Raman whole spectra. The ROA models from the full spectra for α-helix, β-sheet, αβ prediction showed high accuracy with the most correctly predicted proteins for their respective classes. The α-helix, β-sheet and αβ models had MCC values of 0.62, 0.78 and 0.67 respectively. Poor performance was observed from the ROA other fold class prediction which reported low MCC value of 0.05. For all the fold classes, Raman SVM completely failed to predict fold class, as the MCC values were all close to zero. Bar graphs of ROA and Raman SVM classification positive and negative predictions are in Appendix E.

The SVM model performances varied across the different the different amide regions I, II and III (refer to section 4.3.3 chapter 4 for amide region ranges) and the different combinations I&II, I&III, II&III and I&II&III (Table 5.3). The Raman models for the α-helix fold class prediction generally performed the best across the Raman spectra amide combinations compared to the other 3 classes. The highest MCC values were yielded by α-helix prediction on the amide I&III Raman spectra with 0.90 and other structural class prediction on the amide III band Raman spectra with MCC value of 1 predicting all the 'other' class proteins correctly. The Raman full spectra models showed very poor performance accuracy. Table 5.3 below shows the tabulated performance analyses results for the Raman analyses using SVM classification.

**Table 5.3 SVM classification performance accuracies for amide regions and whole spectra of Raman data**

| Raman Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Alpha Helical Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 5 | 2 | 2 | 4 | 6 | 3 | 5 | 1 |
| **TN** | 17 | 14 | 18 | 15 | 18 | 18 | 17 | 15 |
| **FP** | 1(1ab) | 4(2b:2ab) | 0 | 3(2b:1ab) | 0 | 1(1b) | 1(1b) | 3(1a:1b:1o) |
| **FN** | 2 | 5 | 5 | 3 | 1 | 4 | 2 | 6 |
| **MCC** | 0.69 | 0.07 | 0.47 | 0.40 | 0.90 | 0.46 | 0.70 | 0.03 |
| **Beta Sheet Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **TN** | 20 | 20 | 21 | 18 | 18 | 19 | 15 | 21 |
| **FP** | 1(1ab) | 1(1ab) | 0 | 3(2ab:1o) | 3(2ab:1o) | 2(2o) | 6(2ab:4o) | 0 |
| **FN** | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 |
| **MCC** | -0.10 | 0.27 | 0.00 | -0.16 | -0.16 | -0.00 | -0.25 | 0.00 |
| **AlphaBeta Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 4 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| **TN** | 16 | 17 | 19 | 15 | 16 | 18 | 14 | 17 |
| **FP** | 3(2a:1b) | 2(1b:1o) | 0 | 4(3a:1o) | 2(1b:1o) | 2(2b) | 5(3a:1b:1o) | 2(1ab:1o) |
| **FN** | 2 | 5 | 4 | 4 | 4 | 4 | 4 | 5 |
| **MCC** | 0.48 | 0.08 | 0.52 | 0.12 | 0.26 | 0.30 | 0.07 | 0.08 |
| **Other Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 7 | 7 | 8 | 6 | 5 | 6 | 3 | 1 |
| **TN** | 13 | 9 | 17 | 12 | 10 | 9 | 15 | 12 |
| **FP** | 4(3b:1ab) | 8(5a:2ab:1b) | 0 | 5(3b:2ab) | 7(2a:4b:1ab) | 8(3a:2b:3ab) | 2(1b:1ab) | 5(2a:1b:2ab) |
| **FN** | 1 | 1 | 0 | 2 | 3 | 2 | 5 | 7 |
| **MCC** | 0.61 | 0.41 | 1 | 0.41 | 0.20 | 0.26 | 0.30 | -0.18 |

**TP=true positives; TN=true negatives; FP=false positives; FN=false negatives; MCC=Matthews Correlation Coefficient; a=α-helix; β=β-sheet; ab=αβ; o=other.**

As seen in the Raman models, performances for the ROA models, tabulated in Table 5.4 below, varied across the different amide combinations and fold classes with no distinct trends. The performance did show very slight improvement compared to the Raman model performances. Most of the MCC values were very close to zero. The exception was seen with the model from the combination of amide I&II&III for α-helix prediction which produced MCC value of 1. The other fold class predictions from the combination of amide I&II&III produced reasonably good results; β-sheet model had MCC value of 0.56 and the other prediction model had MCC value of 0.59. The αβ model had a low MCC value of 0.11.

Average MCC values were reported by the ROA model from amide I band spectra for β-sheet with 0.54 and the model from amide spectra combination amide I&III with MCC value of 0.59.

**Table 5.4 SVM classification performance accuracies for amide regions and whole spectra of ROA data**

| ROA Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Alpha Helical Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 1 | 1 | 3 | 1 | 3 | 3 | 5 | 3 |
| **TN** | 27 | 26 | 21 | 23 | 24 | 19 | 27 | 26 |
| **FP** | 0 | 1(1ab) | 6(4ab:2o) | 4(2ab:2o) | 3(2a:1o) | 8(5ab:3o) | 0 | 1(1o) |
| **FN** | 4 | 4 | 2 | 4 | 2 | 2 | 0 | 2 |
| **MCC** | 0.48 | 0.24 | 0.31 | 0.05 | 0.45 | 0.23 | 1 | 0.62 |
| **Beta Sheet Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 6 | 0 | 4 | 7 | 3 | 5 | 3 | 5 |
| **TN** | 20 | 25 | 24 | 19 | 24 | 22 | 24 | 24 |
| **FP** | 4(3ab:1o) | 0 | 0 | 5(3ab:1a:1o) | 0 | 2(2o) | 0 | 0 |
| **FN** | 2 | 8 | 4 | 1 | 5 | 3 | 5 | 3 |
| **MCC** | 0.54 | 0 | 0.35 | 0.60 | 0.41 | 0.42 | 0.56 | 0.78 |
| **AlphaBeta Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 3 | 2 | 0 | 2 | 5 | 1 | 2 | 6 |
| **TN** | 21 | 17 | 23 | 23 | 16 | 22 | 20 | 22 |
| **FP** | 3(2a:1b) | 7(5o:2b) | 0 | 1(1b) | 8(3a:2b:3o) | 2(1o:1b) | 3(1o:1a:1b) | 2(1b:1a) |
| **FN** | 5 | 6 | 8 | 6 | 3 | 7 | 7 | 2 |
| **MCC** | 0.28 | -0.04 | 0 | 0.13 | 0.26 | 0.06 | 0.11 | 0.67 |
| **Other Models** | | | | | | | | |
| **Amide Region** | **I** | **II** | **III** | **I&II** | **I&III** | **II&III** | **I&II&III** | **Whole** |
| **TP** | 4 | 5 | 2 | 5 | 5 | 1 | 5 | 2 |
| **TN** | 19 | 13 | 21 | 18 | 21 | 19 | 21 | 18 |
| **FP** | 2(1a:1ab) | 8(3ab:5b) | 0 | 3(1b:2ab) | 0 | 2(1b:1ab) | 0 | 3(3b) |
| **FN** | 7 | 6 | 9 | 6 | 6 | 10 | 6 | 9 |
| **MCC** | 0.33 | 0.07 | 0.34 | 0.34 | 0.59 | -0.01 | 0.59 | 0.05 |

**TP=true positives; TN=true negatives; FP=false positives; FN=false negatives; MCC=Matthews Correlation Coefficient; a=α-helix; β=β-sheet; ab=αβ; o=other**

The SVM classification technique generally predicted very poorly with very low accuracies for both Raman and ROA spectra. The full spectrum analyses from ROA spectra had the best accuracy but only in the predictions for α-helix, β-sheet and αβ classes. These analyses showed that there was no value added to the structural information derived by the SVM classification technique by looking at individual amide band regions of Raman and ROA spectra and their different combinations. The good performance from the ROA full spectra

analyses however, showed that this technique might work better for ROA than for Raman spectra. The values for the bin 10 $cm^{-1}$ for the full spectra presented in tables 5.3 and 5.4 vary from those presented in tables 5.1 and 5.2.

### 5.3 SVM Regression Analyses Results of Raman and ROA spectra

SVM regression works on the sample principle as SVM classification where a separating margin splits the input data using a kernel function which transforms the data into high dimension feature space. In SVM regression, the generalisation margin is optimised such that the function finds a boundary, the ε-tube, about the margin within which the data points that are certain distance from the actual values are not penalised. The points that fall outside this boundary are referred to as the slack variables and are penalised. The function is referred to as epsilon –intensive-loss function. SVM regression in discussed in further detail in chapter 3 on machine learning.

#### 5.3.1 SVM Regression Analyses of Raman and ROA

The SVM regression showed generally poor performance across all the models in the 4 fold classes for both Raman and ROA. Model performance was measured using the $R^2$ correlation coefficient. The lower the $R^2$ value the lower the correlation between the observed percentage content and the value predicted by the SVM regression models. Graphs of SVM regression models showing correlation plots are in Appendix F. Model performance was also described by the root squared mean deviation RMSD and the determination enhancement parameter ζ which is the ratio of RMSD and the standard deviation of the observed data. The higher the value of the ζ parameter, the higher the determination accuracy. Tables 5.5 and 5.6 show the performance accuracies of the SVM regressions for both ROA and Raman spectra analyses.

The SVM regression technique predicted the fractions of the different structural motifs very poorly. The ‘other’ models from both Raman and ROA were particularly poor and most showed over-fitting with the same values being predicted for the entire test set samples. These are marked by the null values for the $R^2$ values in the tables below.

**Table 5.5 ROA SVM regression performance accuracies for analyses using whole spectra and spectra from amide regions**

| | H | | | | E | | | | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| Amide I | 0.02 | 25.32 | 21.14 | 0.83 | 0.44 | 14.90 | 16.57 | 1.11 | 0.00 | 18.16 | 16.54 | 0.91 |
| Amide II | 0.45 | 19.94 | 21.14 | 1.06 | 0.39 | 17.68 | 16.57 | 0.94 | - | 18.26 | 16.54 | 0.91 |
| Amide III | 0.40 | 16.95 | 21.14 | 1.25 | 0.17 | 17.50 | 16.57 | 0.95 | 0.30 | 14.85 | 16.54 | 1.11 |
| Amide I & II | 0.14 | 29.24 | 21.14 | 0.72 | 0.38 | 15.58 | 16.57 | 1.06 | 0.00 | 20.55 | 16.54 | 0.80 |
| Amide II & III | 0.07 | 42.05 | 21.14 | 0.50 | 0.13 | 18.54 | 16.57 | 0.89 | 0.27 | 15.72 | 16.54 | 1.05 |
| Amide I & III | 0.67 | 12.27 | 21.14 | 1.72 | 0.43 | 14.24 | 16.57 | 1.16 | 0.10 | 16.11 | 16.54 | 0 |
| Amide I & II & III | 0.53 | 14.31 | 21.14 | 1.48 | 0.53 | 13.09 | 16.57 | 1.27 | 0.02 | 16.96 | 16.54 | 0.97 |
| Whole | 0.67 | 14.69 | 21.14 | 1.44 | 0.40 | 15.17 | 16.57 | 1.09 | 0.10 | 18.27 | 16.54 | 0.91 |

**H =Alpha Helix, E= Beta Sheet, O= Other,$R^2$ = Regression Coefficient Value, SD= Standard Deviation of SCOP Reference %, RMS=Root Mean Square**

**Table 5.6 Raman SVM regression performance accuracies for analyses using whole spectra and spectra from amide regions**

| | H | | | | E | | | | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| Amide I | 0.12 | 30.08 | 26.68 | 0.89 | 0.03 | 22.27 | 10.96 | 0.49 | - | 18.07 | 18.21 | 1.01 |
| Amide II | 0.03 | 34.11 | 26.68 | 0.78 | 0.06 | 29.67 | 10.96 | 0.37 | - | 18.07 | 18.21 | 1.01 |
| Amide III | 0.08 | 46.58 | 26.68 | 0.57 | 0.01 | 25.52 | 10.96 | 0.43 | - | 18.07 | 18.21 | 1.01 |
| Amide I & II | 0.59 | 21.05 | 26.68 | 1.27 | 0.19 | 21.21 | 10.96 | 0.52 | - | 18.07 | 18.21 | 1.01 |
| Amide II & III | 0.10 | 32.51 | 26.68 | 0.82 | 0.42 | 15.68 | 10.96 | 0.70 | 0.23 | 19.04 | 18.21 | 0.96 |
| Amide I & III | 0.51 | 20.95 | 26.68 | 1.27 | 0.14 | 19.67 | 10.96 | 0.56 | 0.42 | 14.05 | 18.21 | 1.30 |
| Amide I & II & III | 0.53 | 22.30 | 26.68 | 1.19 | 0.23 | 17.63 | 10.96 | 0.62 | 0.46 | 12.96 | 18.21 | 1.40 |
| Whole | 0.03 | 38.18 | 26.68 | 0.70 | - | 14.98 | 10.96 | 0.73 | 0.00 | 30.76 | 18.21 | 0.59 |

**H =Alpha Helix, E= Beta Sheet, O= Other,$R^2$ = Regression Coefficient Value,SD= Standard Deviation of SCOP Reference %, RMS=Root Mean Square**

### 5.3.2 Discussion

The SVM regression analyses generally showed very poor performances for both ROA and Raman with very low correlation between the observed and predicted values and low determination accuracies. Moderate performance was observed from models from amide I&II&III ROA spectra for predictions of α helix and β sheet which both had MCC value of 0.53. The α helix models for both Raman and ROA had comparably better results than other prediction accuracies for the other fold classes with a few models reporting MCC values of greater than 0.5. The ROA α helix models from amide I&III, amide I&II&III and full spectrum had MCC values of 0.67, 0.53 and 0.51 respectively. The α helix prediction models from Raman amide I&III, amide I&II&III and amide I&II spectral bands had MCC values of 0.51, 0.53 and 0.59 respectively.

The SVM analyses failed to predict with any significant accuracy with all models on all the fold classes investigated here. In the SVM classification analyses on Raman spectra (Table 5.3) a few α helix models had high MCC values of 0.69 for amide I, 0.90 for amide I&II and 0.70 for amide I&II&III. By contrast, in the ROA classification analyses (Table 5.4), the α-helix models from amide spectra reported MCC values below 0.5 except the model from amide I&II&III which had an MCC value of 1. Some ROA β sheet models showed moderate prediction accuracy; amide I with MCC of 0.54, amide I&II with MCC of 0.60, amide I&II&III with MCC of 0.56. The predictions for αβ and other motifs from both ROA and Raman amide band spectra had low accuracies except models built from amide I and amide III Raman spectral band which reported MCC values of 0.6 and 1.

The α helix and β sheet structural motifs are formed with well ordered hydrogen bonded properties. The other motif (turns and random coils) has higher variability and looser spatial arrangements. The spectral features of the former are therefore better defined than the latter whose spectral characteristics (location, intensity) may be variable. This may explain why some models had relatively high accuracy in α helix and β-sheet predictions by models from amide band spectra from Raman and ROA compared to predictions of αβ and 'other' fold classes. The more uniform spectral features of the more ordered α helix and β-sheet motifs are easier to model. The poor results from the αβ prediction models might be due to the difficulty in differentiating the combined αβ patterns from the individual motifs. Perhaps the models fail to find a consensus pattern that represents the combined motifs.

The classification analyses of full spectra from ROA spectra produced good quality models except for the 'other' model while the models from full Raman spectra had poor models in all four fold classes revealing that the SVM classification might be better applied to analyses of well defined structural motifs using full spectra from ROA than Raman spectra. The models

built from amide band spectra did not generally confer any improvements in accuracy as would have been expected as all the noise from other regions is excluded.

The SVM regression analyses of ROA and Raman did not yield very good results (Table 5.5 and Table 5.6). Slightly better accuracies (MCC values higher than 0.50) were observed from some models for α-helix prediction. This could be due to the same reasons cited above. Sreeman and Woody (2002) investigated the number of proteins needed in a protein reference set to sufficiently represent the protein fold class space. The groups had five protein reference sets of 29, 37, 42, 43 and 48 proteins depending on the circular dichroism wavelength range for their analyses. Their findings showed the larger reference set performed better than the smaller sets because of the higher structural and spectra variability covered by a larger protein reference set.

In the SVM classification analyses, the models were built from a set of 35 ROA proteins and 28 Raman proteins including denatured proteins and polypeptides. The denatured proteins and polypeptides were removed for subsequent SVM regression analyses to match proteins to their SCOP references. The models from ROA spectra for SVM regression analyses were built from 25 proteins and the models from Raman from 15 proteins. Further analyses, PLS Regression and Random Forest clustering discussed in chapters 6 and 7, used similar reference sets from the SVM regression analyses but showed very good results. In the SVM classification the ROA predictions had more models with MCC values greater than 0.5 revealing that the SVM technique might be more susceptible to the number of proteins in the training set and needs larger number of proteins to increase performance accuracy compared to the other analysis methods PLS regression and Random Forests.

## 5.4 References

Sreerama N. and Woody R. W. (2000) Estimation of the protein secondary structure from circular dichroism spectra: Comparison of CONTIN, SELCON and CDSSTR methods with an expanded reference set. *Analytical Biochemistry* **287**, 252-260

# 6. Partial Least Squares (PLS) Regression Analyses of Raman and ROA spectra

Partial Least Squares (PLS) regression finds the relation between several responses (X) variables and explanatory (Y) variables identifying the underlying factors which best model the X dependent variables. The latent factors derived are those that contribute to the highest variation in the response while maintaining the interspatial relationships between X and Y variables. PLS regression is further discussed in chapter 3. PLS regression was used to determine the fractions of the structural α helix, β sheet, αβ and 'other' motifs from Raman and ROA spectra. The PLS regression models performed better than the SVM regression models in chapter 5.

## 6.1 Methods

Analyses were done on 44 ROA spectra and 24 Raman using MATLAB PLS software and 10 principal components. The PLS models' performances were reported using the Pearson's correlation coefficient ($R^2$), the root squared mean (RMS) and the determination enhancement parameter ζ, which is the ratio of RMS and the standard deviation(SD) in the observed data. The ζ parameter estimates the amount of structural information extracted. The higher the ζ value the more significant the more accurate the prediction. The subcategories of the secondary structures were designated as: H=α-helix, E=β-sheet, O= other and $3_{10}$ helix. Plotted graphs of the PLS regression models showing the correlation between predicted and observed structural content are in Appendix G.

The amide I, II and II and backbone spectral regions were analysed in isolation and in combination to investigate the determination of secondary structure. Analyses were performed on the following regions; amide I (1600-1700 $cm^{-1}$), amide II (1540-1600$cm^{-1}$)

and amide III (1200-1340cm$^{-1}$), a combination of these, the backbone stretch modes (850-1100 cm$^{-1}$) and second derivative Raman spectra.

### 6.2 Results and Discussion

PLS Regression on ROA and Raman spectra performed well compared to SVM Regression (Tables 5.5 and 5.6 in chapter 5). High correlation was observed across all the 3 subcategories classes between the predicted response values and observed structural content values. Excellent results were reported by the analyses on whole Raman and ROA spectra. Raman analyses results showed that the most accurate predictions were obtained when the whole spectra was used: α-helix prediction had $R^2$ of 0.97, β-sheet prediction had $R^2$ of 0.88 and other prediction had $R^2$ of 0.95. The full spectra outperformed all spectral subdivisions showing that there is structural information in the regions outside of the amide and backbone regions.

Tables 6.1 and 6.2 show the PLS regression results on ROA and Raman data on the amide I, II and II regions and also the different combinations of the all the regions as well the whole spectra. High correlations were reported for all amide regions and combinations on the amide regions for ROA and Raman. ROA amide band II prediction was an exception and showed low accuracy with RMS ranges from 11-19% across the different classes. The $R^2$ values for ROA amide II models were 0.29, 0.32 and 0.33 for α-helix, β-sheet and other predictions respectively. This was also observed in the Raman analyses of the amide regions spectra where the prediction of α-helix and β-sheet structural content on amide II spectra showed low correlation values $R^2$ of 0.52 and 0.30. However, the Raman amide II model for 'other' structure performed better with an $R^2$ value of 0.70. The 'other' models generated from ROA spectra performed slightly less accurately than the Raman models for the amide I, II, I&II, II&III spectra. The ROA and Raman analyses showed the amide I and amide III spectral data

were more useful than isolated amide II data. The combined spectra also showed high prediction accuracy showing that combining spectra from different amide ranges provides useful structural information. The α-helix results generally produced higher correlation values than those from the β-sheet or other classes. The higher accuracy of the α-helix maybe be due to more variability in the backbone dihedral angles of β-sheet or 'other' structural classes.

**Table 6.1 ROA PLS regression statistics for analyses using whole spectra and spectra from amide regions**

| | H | | | | | E | | | | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ | (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| **Amide I** | 0.72 | 12.29 | 23.30 | | 1.90 | 0.81 | 7.67 | 17.81 | 2.32 | 0.44 | 10.80 | 14.58 | 1.35 |
| **Amide II** | 0.29 | 19.37 | 23.30 | | 1.20 | 0.32 | 14.52 | 17.81 | 1.23 | 0.33 | 11.76 | 14.58 | 1.24 |
| **AmideIII** | 0.84 | 9.25 | 23.30 | | 2.52 | 0.81 | 7.78 | 17.81 | 2.29 | 0.75 | 7.21 | 14.58 | 2.02 |
| **Amide I&II** | 0.84 | 1.93 | 23.30 | | 12.07 | 0.84 | 6.96 | 17.81 | 2.56 | 0.67 | 8.22 | 14.58 | 1.77 |
| **Amide II &III** | 0.87 | 8.23 | 23.30 | | 2.83 | 0.87 | 6.36 | 17.81 | 2.80 | 0.78 | 6.92 | 14.58 | 2.11 |
| **Amide I&III** | 0.89 | 7.67 | 23.30 | | 3.03 | 0.91 | 5.25 | 17.81 | 3.40 | 0.80 | 6.48 | 14.58 | 2.25 |
| **AmideI&II&III** | 0.90 | 7.42 | 23.30 | | 3.14 | 0.94 | 4.34 | 17.81 | 4.11 | 0.83 | 5.95 | 14.58 | 2.45 |
| **Whole** | 0.98 | 2.88 | 23.30 | | 8.09 | 0.98 | 2.47 | 17.81 | 7.22 | 0.96 | 2.73 | 14.58 | 5.34 |

**H=Alpha Helical; E= Beta Sheet; O=Other. R2 = correlation coefficient; RMS(δ) = Root Mean Squared deviation; SD = Standard Deviation (SCOP reference %); ζ =Ratio of SD/RMS**

**Table 6.2 Raman PLS regression statistics for analyses using whole spectra and spectra from amide regions**

| | H | | | | E | | | | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| **AmideI** | 0.88 | 9.43 | 27.64 | 2.93 | 0.68 | 10.22 | 18.18 | 1.78 | 0.76 | 8.83 | 18.22 | 2.06 |
| **AmideII** | 0.52 | 18.79 | 27.64 | 1.47 | 0.30 | 14.89 | 18.18 | 1.22 | 0.70 | 9.72 | 18.22 | 1.87 |
| **Amide III** | 0.78 | 2.17 | 27.64 | 2.17 | 0.76 | 8.71 | 18.18 | 2.09 | 0.75 | 8.88 | 18.22 | 2.05 |
| **Amide II and III** | 0.90 | 8.37 | 27.64 | 3.30 | 0.86 | 6.58 | 18.18 | 2.76 | 0.94 | 4.32 | 18.22 | 4.22 |
| **Amide I and II** | 0.93 | 7.29 | 27.64 | 3.79 | 0.86 | 6.58 | 18.18 | 2.76 | 0.82 | 7.57 | 18.22 | 2.41 |
| **Amide I and III** | 0.89 | 9.02 | 27.64 | 3.06 | 0.90 | 5.70 | 18.18 | 3.19 | 0.87 | 6.33 | 18.22 | 2.88 |
| **Amide IandIIand III** | 0.94 | 6.50 | 27.64 | 4.26 | 0.92 | 4.89 | 18.18 | 3.71 | 0.91 | 5.22 | 18.22 | 3.49 |
| **Whole** | 0.97 | 4.39 | 27.64 | 6.30 | 0.88 | 4.86 | 18.18 | 2.92 | 0.95 | 4.08 | 18.22 | 4.46 |

**H=Alpha Helical; E= Beta Sheet; O=Other. R2 = correlation coefficient; RMS(δ) = Root Mean Squared deviation; SD = Standard Deviation (SCOP reference %); ζ =Ratio of SD/RMS**

### 6.2.1 The Backbone regions

Analyses were also performed on the spectral region 850-1100cm$^{-1}$, which is designated the backbone region and is associated with backbone skeletal region of the protein (Wen et.al., 1994). It was included in our analyses to investigate the quality of prediction of the fractions of structural fold from this region. The performance analyses are shown in the Tables 6.3 and 6.4 below. The Raman backbone regions showed high $R^2$ values for α-helix, β-sheet and other

models of 0.97, 0.91 and 0.93 respectively. The ROA analyses, on the other hand, showed lower $R^2$ values of 0.67 for the α-helix model, 0.81 for the β-sheet model and 0.01 for the other model. The Raman backbone analyses gave very accurate secondary structure predictions that were comparable to the full spectra.

**Table 6.3 PLS regression results on Raman spectra in the 850-1100 $cm^{-1}$ backbone region**

| Raman | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H | | | | E | | | | O | | | |
| | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) |
| | 0.97 | 4.92 | 27.64 | 5.62 | 0.91 | 5.37 | 18.18 | 3.38 | 0.93 | 4.58 | 18.22 | 3.98 |

**H =Alpha Helix, E= Beta Sheet, O= Other,$R^2$ = Regression Coefficient Value,SD= Standard Deviation of SCOP Reference %, RMS=Root Mean Square**

**Table 6.4 PLS regression results on ROA spectra in the 850-1100 $cm^{-1}$ backbone region**

| ROA | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H | | | | E | | | | O | | | |
| | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) | $R^2$ | RMS(δ) | SD | ζ(SD/ δ) |
| | 0.67 | 13.32 | 23.30 | 1.75 | 0.81 | 7.67 | 17.81 | 2.32 | 0.01 | 10.81 | 14.58 | 1.35 |

**H =Alpha Helix, E= Beta Sheet, O= Other,R2 = Regression Coefficient Value,SD= Standard Deviation of SCOP Reference %, RMS=Root Mean Square**

### 6.2.2 2nd Derivative Raman spectra data

Very high correlation coefficients were obtained from analyses performed on 2nd derivative Raman whole spectra data. To investigate if the performance accuracy could improved by removing interference from the spectra, we analysed the used derivative of the Raman spectra. The 1st derivative measures how the intensity varies with the wave number ($cm^{-1}$). This can be calculated by measuring differences in intensities at points along the spectrum. However, this variation in intensity can be estimated from the spectral graph of a protein. The 2nd derivative makes it easier to determine the true nature of the intensity peaks by highlighting the varying spectral features arising from vibrational modes of the structural motifs, as discussed in chapter 2, while removing the background interferences in the original spectra. Table 6.5 below shows both analyses results from whole Raman spectra data and 2nd derivative Raman spectra. The $R^2$ values were slightly higher than the whole spectra with strong correlation shown by value of 0.99 for all the 3 subdivisions.

**Table 6. 5 PLS regression results on 2nd derivative Raman spectra. The performance results are compared to the results of analyses on whole Raman spectra without the 2nd derivative.**

| | H | | | | E | | | | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| **Raw** | 0.97 | 4.39 | 27.64 | 6.30 | 0.88 | 6.24 | 18.17 | 2.91 | 0.95 | 4.08 | 18.22 | 4.46 |
| **2nd derivative** | 0.99 | 1.68 | 27.64 | 16.42 | 0.99 | 0.63 | 18.17 | 28.69 | 0.99 | 1.36 | 18.22 | 13.38 |

**H =Alpha Helix, E= Beta Sheet, O= Other, $R^2$ = Regression Coefficient Value, SD= Standard Deviation of SCOP Reference %, RMS=Root Mean Square**

### 6.2.3 Alternative helix structure assignment

Lees *et. al.*, 2006 investigated various methods of determining secondary structure using circular dichroism spectroscopy data and examined how different secondary structure assignments affected performance of the prediction of PLS regression models, for example dividing the β strand motif into parallel and antiparallel. An assignment scheme of PPII helix, core β-sheet, $\beta_D$, α-helix and other produced PPII helix predictions with $R^2$ (regression coefficient) values of 0.64, the α-helix model reported $R^2$ value of 0.97, core β-sheet had an $R^2$ value of 0.86 and other had $R^2$ value of 0.86. PLS regression analysis using an assignment scheme of β-strand parallel β-strand anti-parallel , β turns, core β-sheet, PP-II helix and $3_{10}$ helix reported $R^2$ values of 0.88 for core β-sheet, 0.64 for PP-II helix prediction and 0.39 for $3_{10}$ helix prediction. A simple 3-state secondary structure assignment scheme of α-helix, β-sheet and 'other' from PLS regression analyses showed high prediction accuracies with $R^2$ values of 0.97 for the α-helix, 0.90 for β sheet and 0.78 for other. This follows the general trend of the results discussed in this chapter with α-helix performing better than the other two structural classes. However, the ROA results (Table 6.1), were as accurate for β-sheet as for α-helix secondary structure prediction.

The group concluded that the 3-state structural assignment was robust enough as most proteins rich in α-helix and β-strand structures will have less $3_{10}$ helix and β-bridge content. It is also the structural assignment that is used by most structural prediction programs today and therefore would be easily adaptable to different structural analyses problems. The splitting of

the structural classes into smaller components for example, splitting the β-sheet structural class into β-strand parallel β-strand anti-parallel, β turns, does not seem to be very beneficial in adding anymore structural information. Table 6.6 showed no significant improvements in prediction of α-helix secondary structure content therefore suggesting that the analysis using the 3-state scheme of α-helix, β-sheet and other may be sufficient to extract structural information from spectral data.

Oberg *et. al.*, 2004 performed analyses on circular dichroism spectra investigating different structural assignment schemes and if the assignments enhanced the performances of the PLS regression models. It was found that combinations of helix structures (α-helix & $3_{10}$ helix or α-helix& irregular structures) or of strand (β-strand & isolated-extended β residues) did not improve prediction accuracy compared to the α-helix or β-strand predictions on their own. The isolated-extended β residues were described as residues with φ,ψ angles that fall within the β-sheet Ramachandran range, but do not form the β-sheet structure. Oberg *et. al*.(2004) findings coincide with our results for PLS analysis using a combination of $3_{10}$ helix and α-helix for helix secondary structure assignment. The combination of helix content did not show any significant improvements in prediction accuracy of Raman or ROA helix prediction models. Table 6.6 below shows the performance accuracy when the helical content was combined for Raman and ROA spectra.

**Table 6.6 PLS Regression performance statistics on combined helical content of α-helix and $3_{10}$helix**

| Raman | | | | | ROA | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMS(δ) | SD | ζ (SD/δ) | | $R^2$ | RMS(δ) | SD | ζ (SD/δ) |
| **AmideI** | 0.88 | 9.33 | 27.55 | 2.95 | **AmideI** | 0.71 | 12.35 | 23.09 | 1.87 |
| **AmideII** | 0.49 | 19.15 | 27.55 | 1.44 | **AmideII** | 0.05 | 19.04 | 23.09 | 1.21 |
| **Amide III** | 0.81 | 11.86 | 27.55 | 2.32 | **Amide III** | 0.82 | 9.64 | 23.09 | 2.40 |
| **Amide II and III** | 0.90 | 8.62 | 27.55 | 3.19 | **Amide II and III** | 0.86 | 8.57 | 23.09 | 2.70 |
| **Amide I and II** | 0.91 | 8.02 | 27.55 | 3.43 | **Amide I and II** | 0.87 | 1.61 | 23.09 | 14.39 |
| **Amide I and III** | 0.89 | 8.82 | 27.55 | 3.12 | **Amide I and III** | 0.89 | 7.50 | 23.09 | 3.08 |
| **Amide IandIIand III** | 0.95 | 6.16 | 27.55 | 4.47 | **Amide IandIIand III** | 0.90 | 7.34 | 23.09 | 3.15 |
| **Whole** | 0.94 | 6.39 | 27.55 | 4.31 | **Whole** | 0.98 | 3.01 | 23.09 | 7.65 |

**H=Alpha Helical; E= Beta Sheet; O=Other. $R^2$ = correlation coefficient; RMS(δ) = Root Mean Squared deviation; SD = Standard Deviation (SCOP reference %); ζ =Ratio of SD/RMS.**

The results showed consistent or slightly lower ζ scores and $R^2$ correlation values for the combined helical prediction accuracies (Table 6.6) compared to predictions for α-helix alone (Tables 6.1 and 6.2). There were small increases seen in prediction accuracies only for Raman amide II, III and I&II&III. The prediction accuracies for Raman amide II, III and I&II&III for α-helix content alone and α-helix content with $3_{10}$ have been summarised in table 6.7.

**Table 6.7 Comparison of prediction accuracies for Raman amide II,III and I&II&III between Alpha helix content alone and Alpha helix content with $3_{10}$ helix content.**

| | Alpha Helix (H) | | Alpha Helix (H) + $3_{10}$ Helix | |
|---|---|---|---|---|
| | $R^2$ | ζ | $R^2$ | ζ |
| Raman amide II | 0.94 | 4.26 | 4.47 | 0.95 |
| Raman amide III | 0.78 | 2.17 | 2.32 | 0.81 |
| Raman amide I&II&III | 0.94 | 4.26 | 4.47 | 0.95 |

**H=Alpha Helical; $R^2$ = correlation coefficient; ζ =Ratio of SD/RMS; RMS = Root Mean Squared deviation; SD = Standard Deviation (SCOP reference %);**

The ROA analyses did not show improvements in the prediction accuracies in the combined helix predictions. The amide II&III combined helical content (α-helix & $3_{10}$ helix) prediction was unusual showing the same value $R^2$ of 0.87 as that for α-helix content alone. There was however, an increase in information $\zeta_{H,ROA} = 2.83 \rightarrow \zeta_{H+3_{10}H,ROA} = 14.39$. The correlation between the observed structural content and the predicted variable remained constant, but the there was more content information.

The PLS regression algorithm produced very accurate results for prediction of fractions of structural motifs from ROA and Raman spectra. The full spectra seem to contain more information as they have slightly higher performance indices compared to the amide band regions showing that the full spectral features contains more information that could be lost when only the amide regions are used.

In the analyses of the amide spectral bands, the models from the amide II band ROA and Raman showed low prediction accuracy. However, the Raman spectra models from amide II for α helix prediction had $R^2$ value of 0.52 and for other prediction had $R^2$ value of 0.70

making these the exception. This might show that there is some structural information that could be mined from the amide II region which is often left out of structural analyses of spectral data as the amide II Raman band has been reported as weak (Salzer and Siesler, 2009; Socrates, 2001; Susi and Bylert, 1988).

The analyses using the secondary structure scheme of α-helix and $3_{10}$ helix concurred with studies from Lees and Orberg groups, as discussed above, reporting that the structural assignment scheme where helical structure is distinguished into α-helix and $3_{10}$ helix did not enhance the performance of the prediction models for the α helix structural motif. Sreerama and Woody (2004) and Lees *et.al*. (2006), using multivariate regression analyses to determine secondary structure from circular dichroism (CD) spectral data and reported correlation coefficient values in the ranges of $R^2$= 0.85-0.94 for α-helix and $R^2$= 0.64-0.84 for β-sheet. The analyses discussed here reported higher correlation coefficient values for α-helix in the range of $R^2$ = 0.97-0.98 for both ROA and Raman data analysis (Tables 6.1 and 6.2). β-Sheet secondary structure analysis also gave high correlation coefficient values in the range of $R^2$= 0.88-0.98 for both Raman and ROA spectral data analysis (Tables 6.1 and 6.2).

The excellent results from PLS regression algorithm proved that there is a strong relationship between ROA and Raman spectral features and structural information which can be extracted using multivariate analytical methods like PLS regression. Unlike SVM regression, PLS regression showed very good results despite the small protein sets making it an ideal analytical method to use on smaller protein sets. The analyses presented here were also submitted for publication (Kinalwa *et.al*., 2010).

## 6.3 References


Kinalwa M., Blanch E. W., Doig A. J. (2010) Accurate Determination of Protein Secondary Structure Content from Raman and Raman Optical Activity Spectra. *Analytical Chemistry*, 82, 6463-6471

Lees J.G., Miles A.J., Janes R.W., Wallace B.A. (2006) Novel methods for secondary structure determination using low wavelength (VUV) circular dichroism spectroscopic data. *BMC Bioinformatics* **7**,507-518

Orberg K.A., Ruysschaert J., Goormaghtigh E. (2004) The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *European Journal of Biochemistry* **271**, 2937-2748

Salzer R., Siesler H. W.(2009) Infrared and Raman spectroscopic imaging, Wiley-VCH, Weinheim,Germany

Sreerama N., Woody R.W.(2004) Computation and Analysis of Protein Circular Dichroism Spectra. *Methods in Enzymology* **383**,318-351

Socrates G.(2001) Infrared and Raman characteristic group frequencies: tables and charts, 3rd ed., John Wiley & Sons, West Sussex, England

Susi H., Bylert D. M. (1988) Fourier deconvolution of the amide I Raman band of proteins as related to conformation. *Applied Spectroscopy* **42**, 819-826

Wen Z.Q. Hecht L.,Barron L.D. (2004) β-Sheet and associated turn signatures in vibrational Raman optical activity spectra of proteins. *Protein Science* **3**,435-439

# 7. Random Forest Cluster Analyses of ROA and Raman spectra

A random forest (RF) is a combination of many decision trees such that each tree depends on randomly selected vectors. The trees are grown without pruning, splitting nodes following the Gini index criteria. Error rates at each node are estimated using Out-of-Bag sampling. Each tree is grown to maximal size repeating the splitting process at each node such that a different subset of variables is randomly selected each time. A system of class voting then take place where the class a variable belongs to is defined as the class which the majority of the trees in the 'forest' have placed that variable. The random forest clustering algorithm was used for the determination of fold classes of the proteins based on their ROA and Raman spectra.

## 7.1 Model Performance Analysis

The measure of accuracy of the Random Forest models was the percentage of the correctly predicted observation over the total number of samples. Based on this measure, the ROA RF models yielded better results than those from Raman models. ROA RF models had more samples correctly predicted than the Raman RF models. Table 7.1 shows the percentages of correct prediction for both ROA and Raman RF models generated from whole and amide region spectra. Predictions on every amide region of ROA showed better performance than the predictions on ROA whole spectra. ROA RF correctly predicted only 55 % of the samples (22/44). The lowest accuracy was seen in the prediction of the Raman amide II RF model which had only 2 out of 24 sample 8% correctly predicted.

**Table 7.1 Percentage of proteins whose classes were correctly predicted. Numbers in parenthesis show the number of correctly predicted observations out of the total number of observations**

| | **Raman** | **ROA%** |
|---|---|---|
| Amide I | 54% (13/24) | 64% (28/44) |
| Amide II | 8% (2/24) | 64% (28/44) |
| Amide III | 46% (11/24) | 59% (26/44) |
| Amide I+II | 50% (12/24) | 73% (32/44) |
| Amide II+III | 38% (9/24) | 75% (33/44) |
| Amide I+III | 50% (12/24) | 61% (27/44) |
| Amide I+II+III | 38% (9/24) | 70% (31/44) |
| Whole | 38% (9/24) | 55% (22/44) |

The Raman and ROA spectra datasets in Table 7.1 above were visualised in two multi dimensional scaling (MDS) plots below. Multidimensional scaling is a technique that computes points whose interdistances are as close as possible to the proximity distances produced by the Random Forest algorithm. The measure of how the proximity distances are close to the MDS calculated distances is referred to as the stress fitness function. In these analyses, the Shepard plot was used to visualise the fitness function. The Shepard plot, shown in figure 7.1, is a scatter plot of the interpoint distances vs. the original dissimilarities. In the Shepard plot, the nearer the scatter is to 1:1 distribution the better the fit of the distances to the dissimilarities.

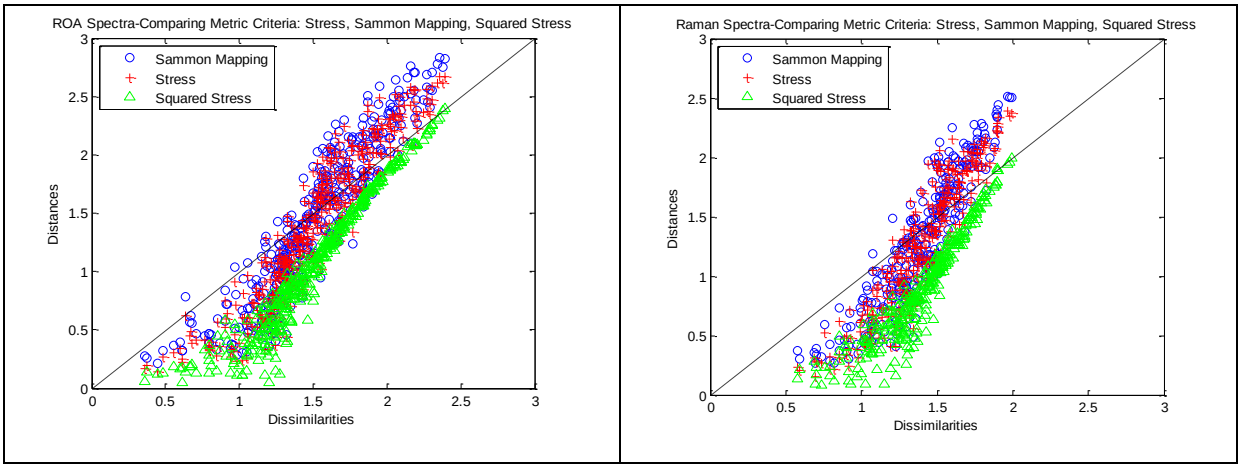



**Figure 7.1 Plots showing Shepards plots of ROA(left) and Raman (right) showing the different stress functions provided in the MATLAB software, Sammon Mapping, Metric Stress and Squared Stress. The scatter for the Squared Stress is not close to the 1:1 line except the very few points at the largest dissimilarity values. Out of the three stress measures Sammon Mapping tends the closest towards the 1:1 line which means the interpoint distances will be closely approximated from the original dissimilarities therefore preserving them.**

The MDS configuration in which the points are arranged is predetermined by the user. In our analyses, we found the two-dimensional configuration sufficient enough for visualisation purposes. The MDS technique is discussed in chapter 4 section 4.4.3 and chapter 3 section 3.4.5. The data points are coloured by their SCOP structural class: red-αβ proteins blue-α helical proteins, green-β sheet proteins, black- other proteins. The ROA whole spectra, figure 7.2 below left side, shows clearly distinct clusters of α-helical proteins, αβ proteins and β sheet proteins. The other proteins are dispersed among the β sheet proteins and αβ proteins. The Raman spectral data on the other hand does not show clear, distinct clusters unlike those observed in the ROA data. The figures below show the Shepard's plot for both ROA and Raman data.

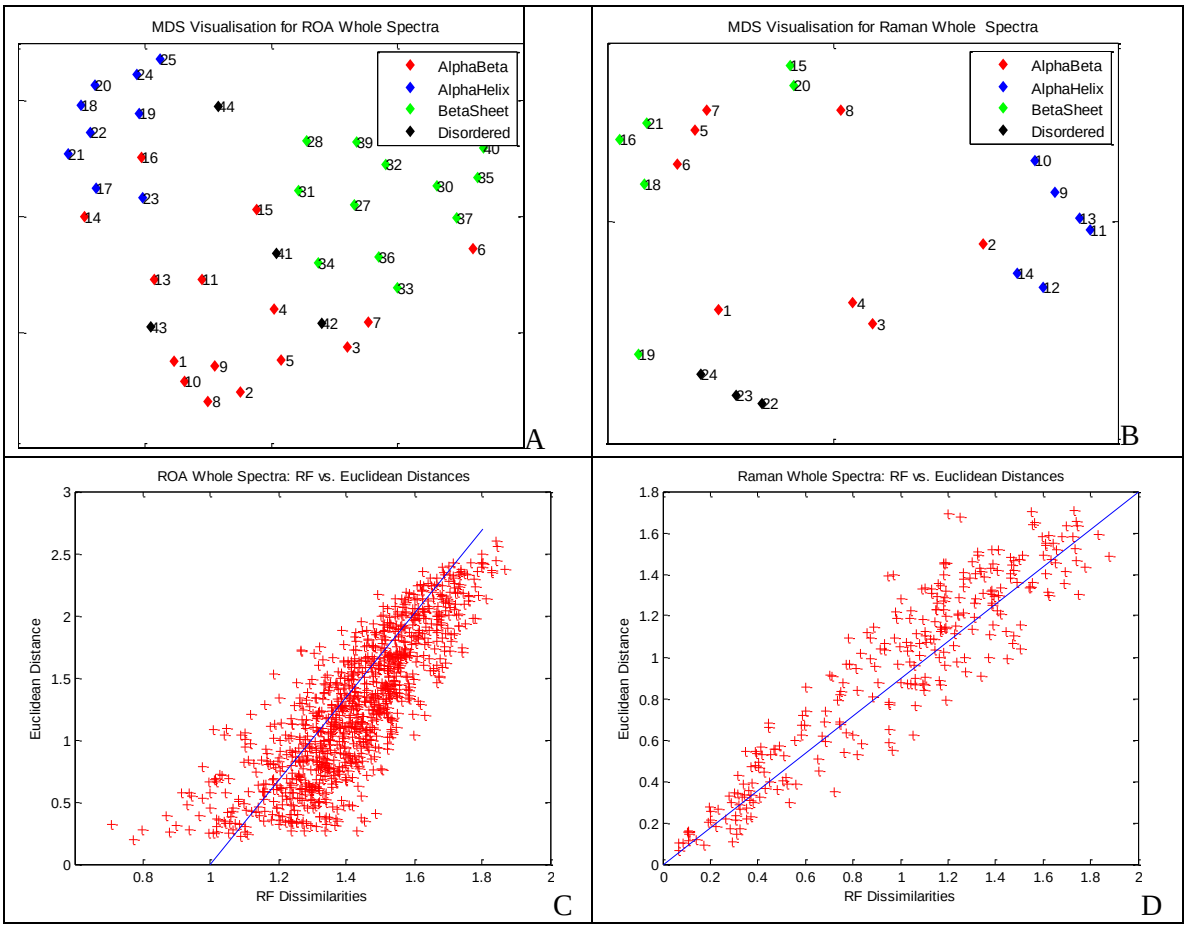



**Figure 7.2 The figures at the top show the RF multidimensional scale (MDS) plots used to visualise the clusters in the data from Raman (above right(B)) and ROA( above left (A)) whole spectra. The MDS clusters were calculated using the Sammon mapping stress function. ROA data showed clusters of the α-helical, the β-sheet and the α/β proteins. Raman data showed clusters of the α-helical and the other structural class proteins. The figures at the bottom show the plot of the Random Forest dissimilarities vs. the Euclidean Distances to show how well the two distances correlated with high Spearman's correlation values of 0.91 for Raman whole spectra (lower bottom right (D)) and 0.83 for ROA data(lower bottom left(C)) .**

## 7.2 Results and Discussion

### 7.2.1 ROA spectra Random Forest Analyses

The ROA plots for amide I, amide I & III and amide I & II & III (figures 7.3, 7.4 and 7.5) showed clear clusters of α-helix , β-sheet and α,β proteins. In figure 7.7, showing the amide I&III plot, the other proteins are scattered throughout the plot. The other protein 42 (orosmucoid) and β-sheet protein 7 (ubiquitin) were grouped nearer the α-helix cluster, while the disordered protein 43 was nearer the β-sheet cluster. The amide I & II & III plot shows proteins 41 and 42 (Bowman-Birk inhibitor and orosmucoid, respectively) clustering closer to the α-helix cluster. Orosmucoid (3kq0), an important drug binding protein whose secondary structure, solved by Schonfeld *et. al.* (2008), comprises an 8 stranded β-barrel with four loops that form the ligand binding site and α-helix structure. The secondary structure of Bowman-Birk inhibitor (1pi2) has two domains connected by a region of four antiparallel β-strands and four connecting loops (Chen *et.al.*, 1992). It is possible that RF's can distinguish the presence of ordered structure and classes these proteins closer to an ordered structure motif.

The β-sheet protein 7 (ubiquitin 1ubq) is also an outlier lying closer to the α-helix cluster. A small protein of 76 amino acid residues, ubiquitin plays an important role in the degradation of proteins by signalling to cells which proteins are to be removed. The grouping of ubiquitin could be explained by the secondary structure which contains β-sheet with anti-parallel strands and a short α helix region. Appendix J has the full list of proteins in the MDS plots for the ROA analyses.

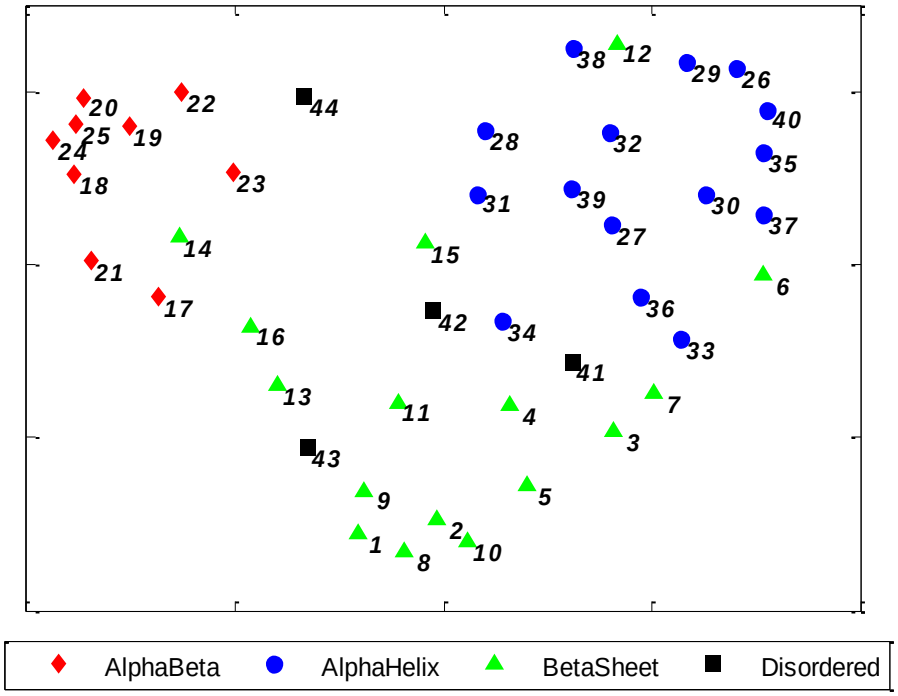



**Figure 7.3 Multidimensional Scaling Plot for ROA Amide I spectra**

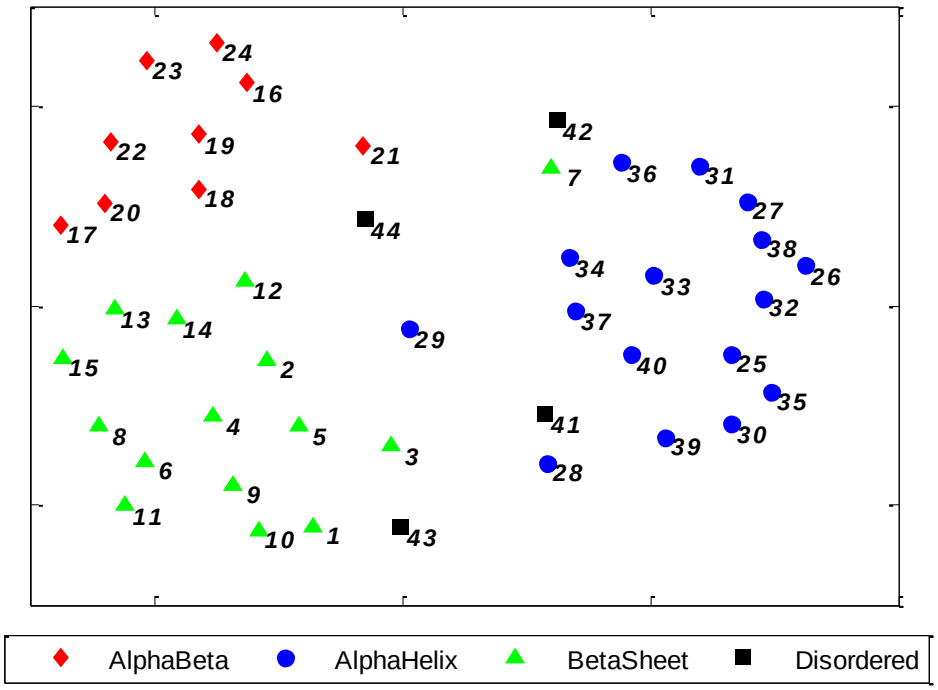



**Figure 7.4 Multidimensional Scaling Plot for ROA Amide I & III spectra**

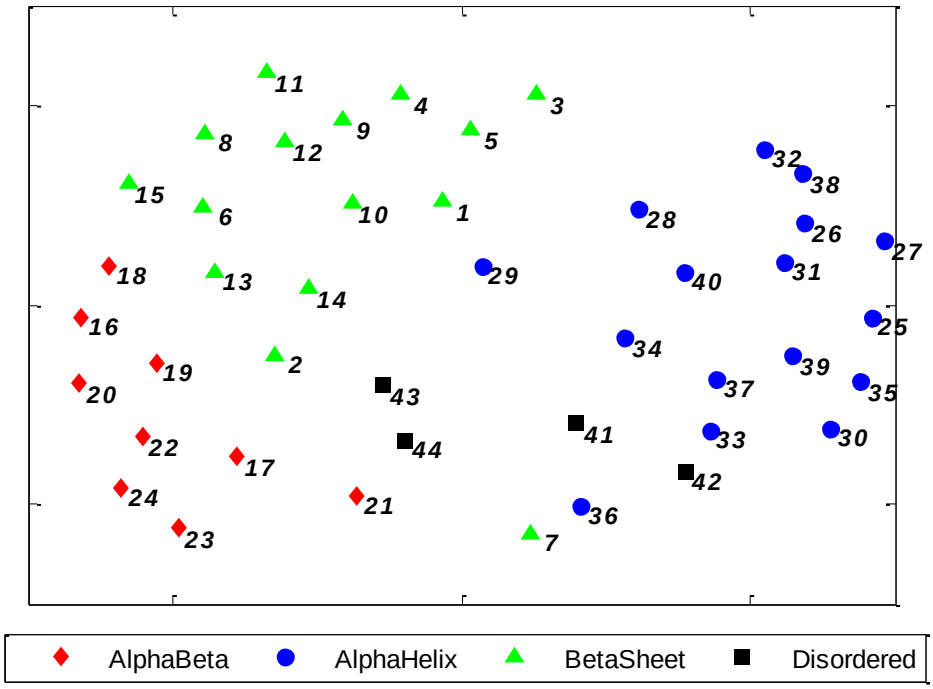



**Figure 7.5 Multidimensional Scaling Plot for ROA Amide I & II& III spectra**

The amide I&II plot, seen in figure 7.6, shows a distinct α-helix cluster, while the other subdivisions do not show any distinct groupings.

α-Helix outlier, avidin (36) lies closer to the β-sheet proteins. Avidin (1rav) has a high affinity for biotin makes it an important component in biotechnological and biomedical applications (Nardone *et. al.*, 1998). For example in immunoassays the avidin-biotin system is used to increase sensitivity and specificity. Structurally, avidin is a tetramer of identical subunits each comprising 8 stranded antiparallel β-barrel. This is probably why the RF groups it closer to the β-sheet proteins. The other proteins are seen to be scattered throughout the plot. The plots for amide II and amide III seen in figures 7.7 and 7.8 respectively, do not show any clear clusters. The α-helical proteins show some clustering for amide II&III in figure 7.9.

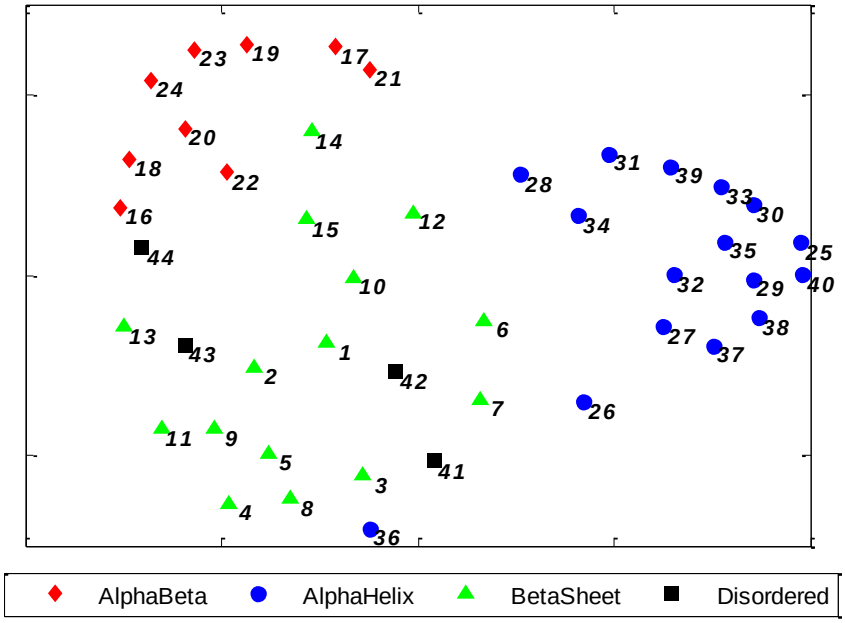



**Figure 7.6 Multidimensional Scaling Plot for ROA Amide I & II spectra**

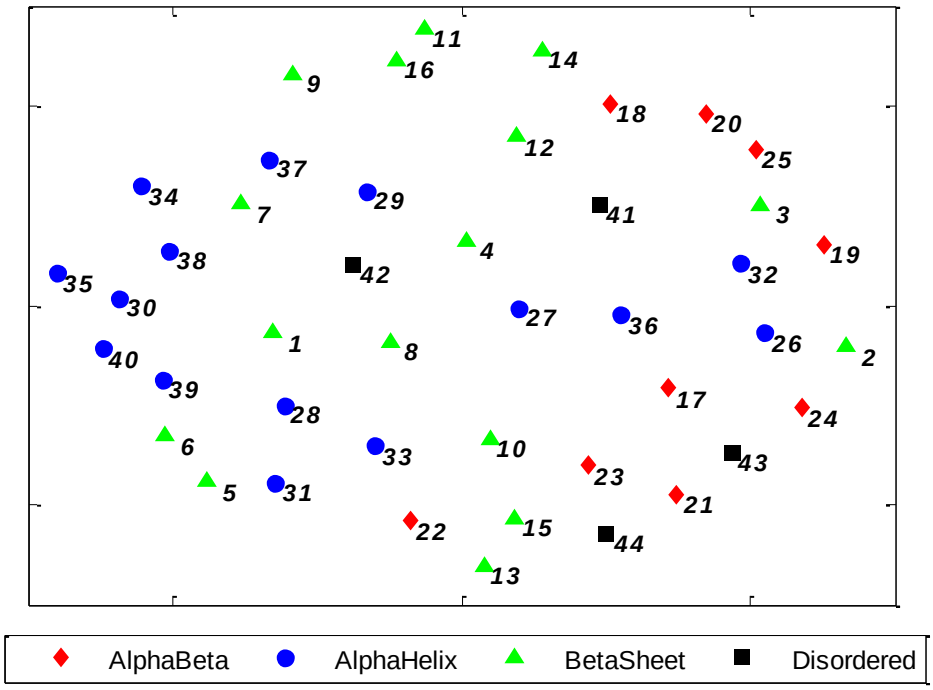



**Figure 7.7 Multidimensional Scaling Plot for ROA Amide II spectra**

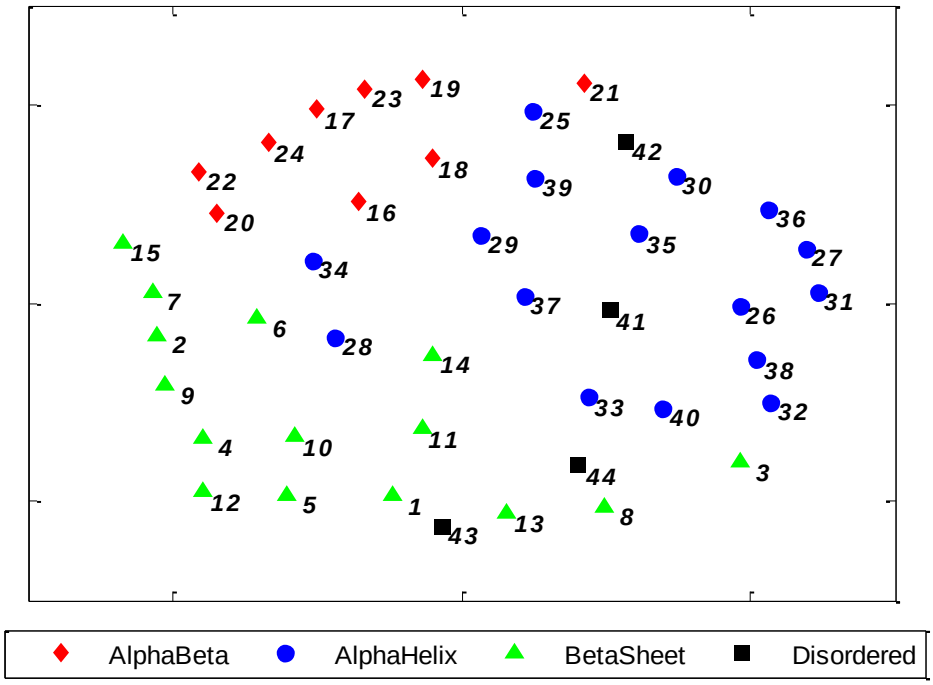



**Figure 7.8 Multidimensional Scaling Plot for ROA Amide III spectra**

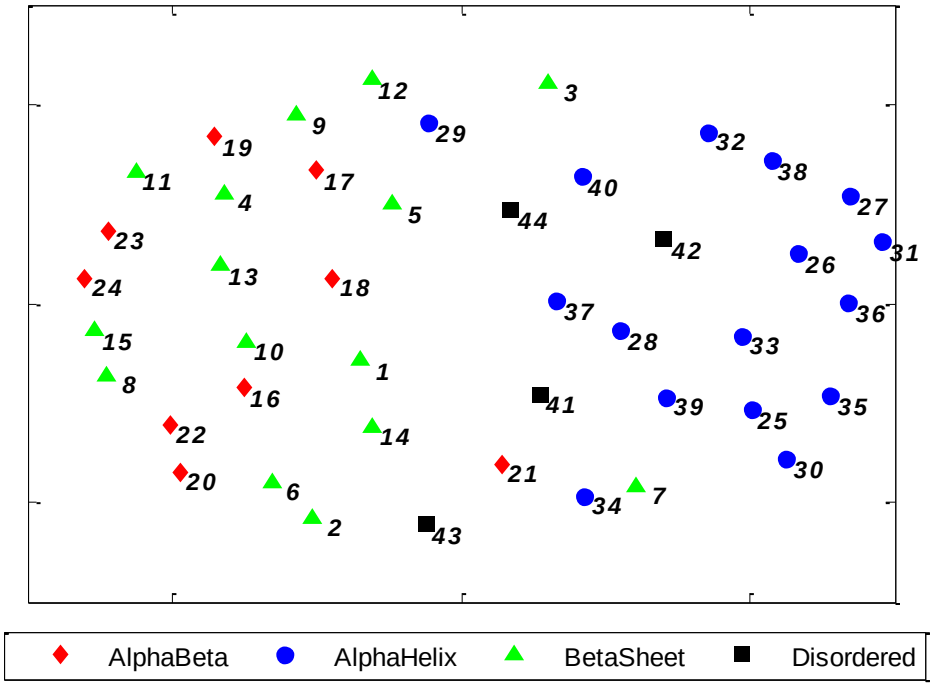



**Figure 7. 9 Multidimensional Scaling Plot for ROA Amide II & III spectra**

### 7.2.2 Raman spectra Random Forest Analyses

On the whole, the Raman MDS plots did not show any significant clustering between the fold classes, unlike those observed in the ROA MDS plots. The most distinct cluster observed in the Raman MDS plots was formed by the α,β proteins in almost all the Raman amide plots. β-Sheet protein 7 (human lysozyme) lies closer to the α,β protein cluster in the amide III plot (figure 7.10) and the amide I+II+III plot (figure 7.11). The α-helix proteins 21 (α lactalbumin) and 18 (satellite tobacco mosaic virus) in the amide II&III plot (figure 7.12) are shown to group nearer the αβ cluster of proteins. The proteins α –lactalbumin (1f6s) and human lysozyme (1gaz) are homologous in amino sequence and tertiary structure (Takano *et. al.* 2000, Chrysina *et. al.* 2000). The structure of human lysozyme has an α domain with four α-helices and a β domain with two β-sheets. The lactose synthase regulator α-lactalbumin also has α and β structural motifs with a larger α-helical subdomain and a smaller subdomain with three antiparallel β-strands. This RF classification decision shows the sensitivity of this method to the spectral features that arise from the ordered α-helix and β-sheet structural motifs. The amide II plot (figure 7.13) showed no distinct groupings at all. See Appendix K for the full key of proteins used in Raman RF analyses. The MDS plots not shown here are in Appendix I.

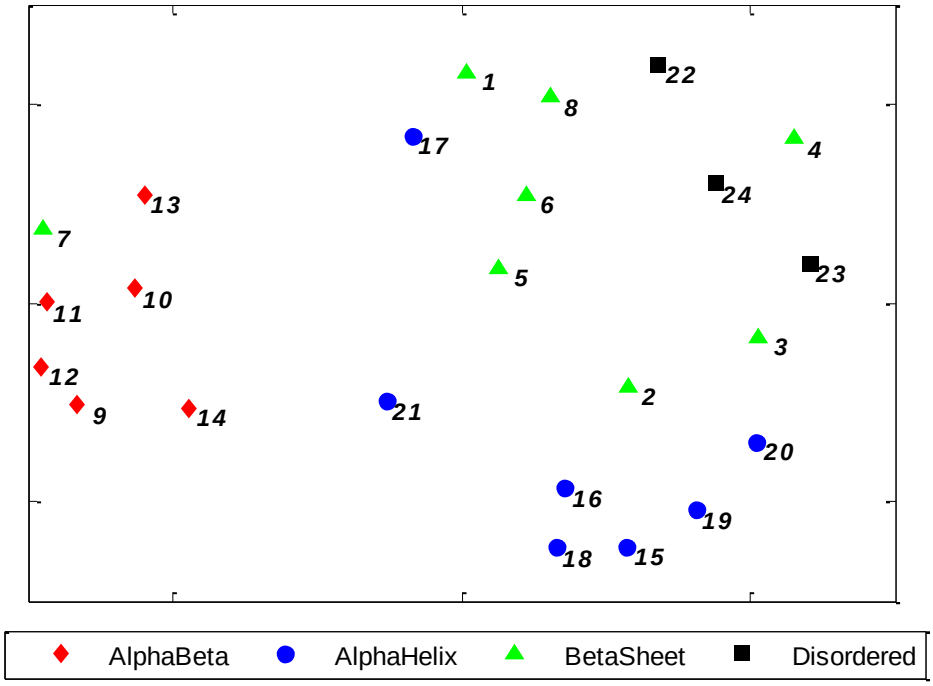



**Figure 7.10 Multidimensional Scaling Plot for Raman Amide III spectral**

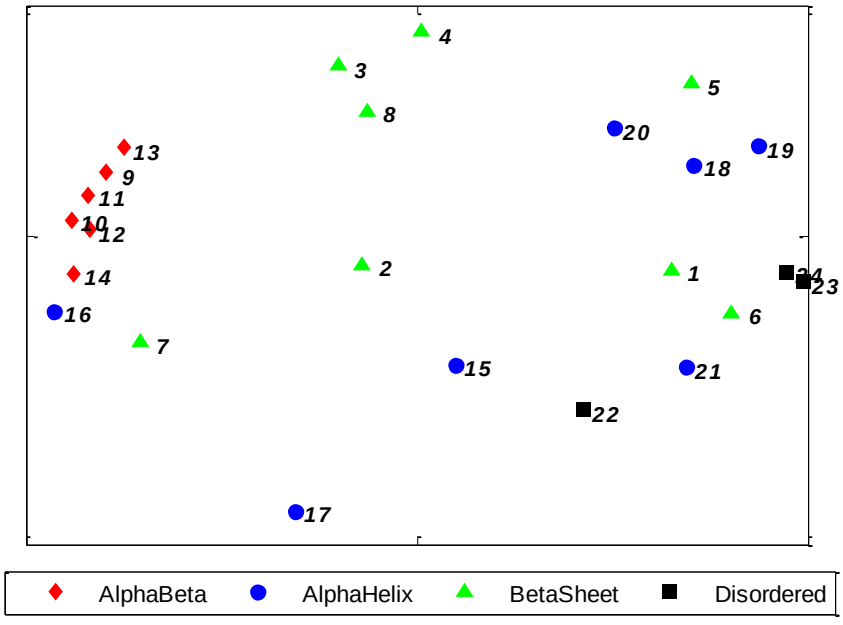



**Figure 7.11 Multidimensional Scaling Plot for Raman Amide I & II& III spectra**

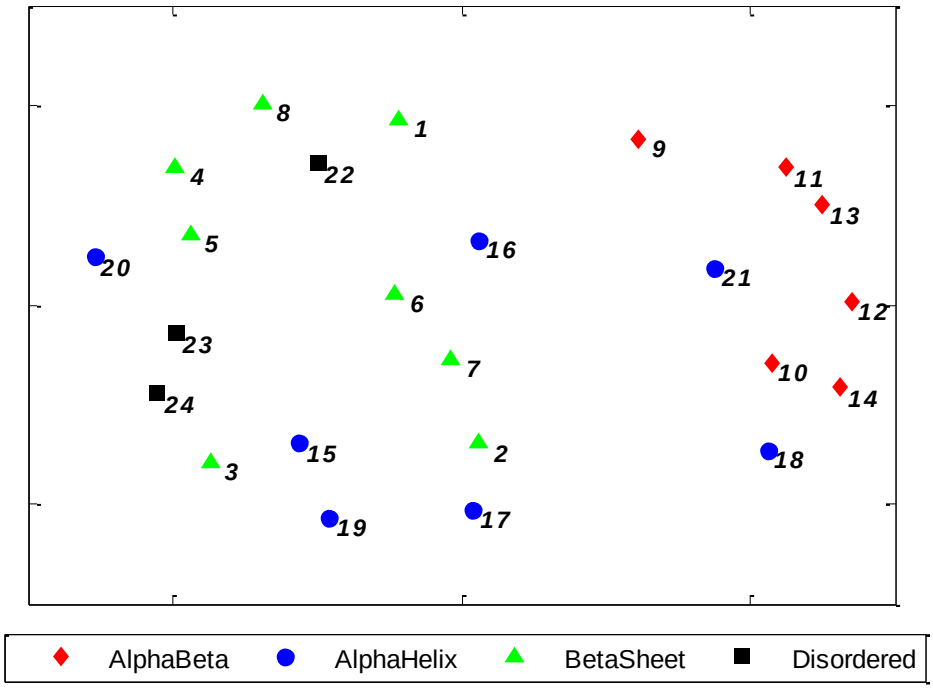



**Figure 7.12 Multidimensional Scaling Plot for Raman Amide II& III spectra**

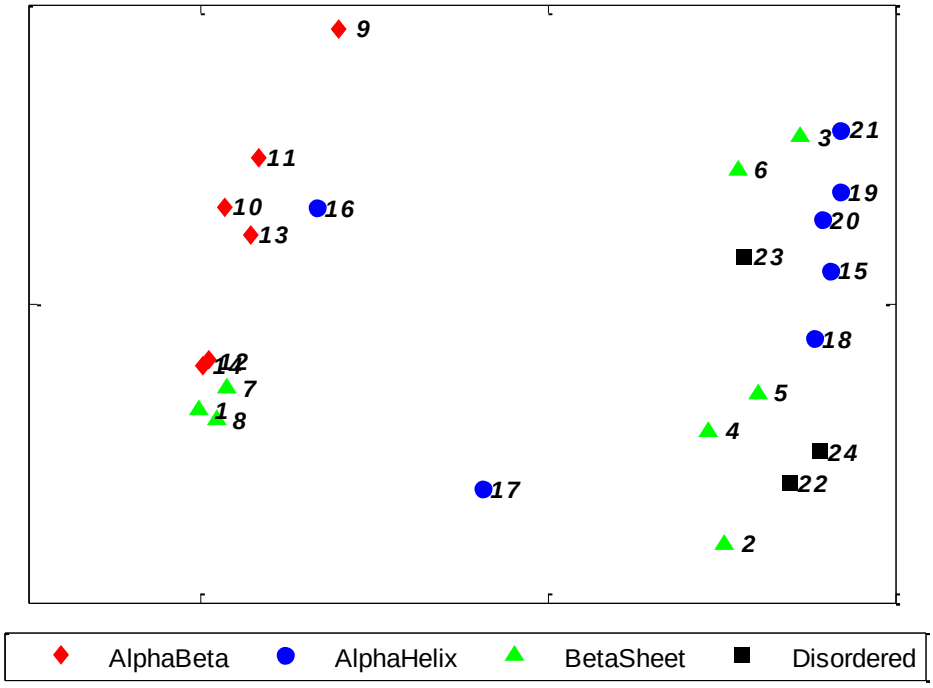



**Figure 7.13 Multidimensional Scaling Plot for Raman Amide II spectra**

### 7.2.3 Variable Importance

The Random Forest algorithm measures which bins were most important in the classification of proteins. Tables 7.2 and 7. 3 show which bins were most important in the clustering of the proteins into the α-helix, β-sheet, αβ and other fold classes. The variable importance is calculated using the difference between the proportions of correctly predicted classes and the total number of cases predicted. This figure is referred to as the margin M. Each variable is left out in iteration constituting what is termed as the out-of-bag cases. Given the known margin $M_{old}$ which was calculated before the out-of-bag cases were removed from the node, the values of the $x^{th}$ variable on the $k^{th}$ tree are permuted and a new margin $M_{new}$ is now calculated. The difference between the original margin $M_{old}$ and the margin of the permuted subset mutation $M_{new}$ represents the variable importance. The Random Forest algorithm is explained in detail in chapter 3. The bigger the difference the higher the importance attached to the given variable to the classification decision. Bar graphs of the variable importance for these analyses are shown in Appendix H. The M variables presented in table 7.2 are estimated from the bar graphs of variable importance shown in Appendix H.

**Table 7.2 Most important bins for assigning fold class for Raman Random Forest analyses**

| Region | Bin Number | Spectral range $cm^{-1}$ | M variable |
|---|---|---|---|
| Amide I | Bin 7 | 1665-1675 | 0.08 |
| | Bin 8 | 1675-1685 | 0.05 |
| | Bin 9 | 1685-1695 | 0.05 |
| Amide II | Bin 2 | 1525-1535 | 0.036 |
| Amide III | Bin 6 | 1255- 1265 | 0.06 |
| Amide I+II | Bin 14 | 1665-1675 | 0.06 |
| | Bin 15 | 1675-1685 | 0.05 |
| Amide II+III | Bin 4 | 1235-1245 | 0.024 |
| | Bin 5 | 1245-1255 | 0.024 |
| | Bin 6 | 1255- 1265 | 0.024 |
| Amide I+III | Bin 21 | 1665-1675 | 0.036 |
| | Bin 22 | 1675-1685 | 0.028 |
| Amide I+II+III | Bin 28 | 1665-1675 | 0.036 |
| | Bin 29 | 1675-1685 | 0.02 |
| Whole | Bin 37 | 965-975 | 0.012 |

*M Variables are read from Variable Importance bar graphs in Appendix H

**Table 7.3 Most important bins for assigning fold class for ROA Random Forest analyses**

| Region | Bin Number | Spectral range $cm^{-1}$ | M variable |
|---|---|---|---|
| Amide I | Bin 6 | 1655-1665 | 0.08 |
| Amide II | Bin 3 | 1535-1545 | 0.032 |
| Amide III | Bin 14 | 1335-1345 | 0.09 |
| Amide I+II | Bin 7 | 1575-1585 | 0.05 |
| | Bin 13 | 1655-1665 | 0.05 |
| | Bin 14 | 1665-1675 | 0.07 |
| Amide II+III | Bin 14 | 1335-1345 | 0.07 |
| Amide I+III | Bin 14 | 1335-1345 | 0.09 |
| | Bin 20 | 1655-1665 | 0.07 |
| Amide I+II+III | Bin 14 | 1335-1345 | 0.07 |
| | Bin 27 | 1655-1665 | 0.06 |
| | Bin 28 | 1665-1675 | 0.03 |
| Whole | Bin 70 | 1295-1305 | 0.013 |
| | Bin 105 | 1645-1655 | 0.006 |

*M Variables are read from Variable Importance bar graphs in Appendix H

Some of the bins in the Tables 7. 2 and 7.3 are already associated to structure motif markers. The 1665-1675 $cm^{-1}$ Raman variable importance bin may be due to the β-Sheet signature amide I bands that are associated to the range ~1665-1680 $cm^{-1}$ (Barron *et. al.*, 2003; Blanch *et. al.*, 2004). Amide III bands that also arise from a β-sheet motif in Raman spectra assigned to the region ~1230-1245 $cm^{-1}$ could be linked to the bin of 1235-1245 $cm^{-1}$. Undefined structure has previously been assigned to amide III band at ~1252$cm^{-1}$ falls within the 1245-1255 $cm^{-1}$ bin (Maiti *et. al.*, 2004; Barron *et. al.*, 2003; Blanch *et. al.*, 2004).

The α helix characteristic amide I ROA couplet that is negative at ~1640 $cm^{-1}$ and positive at 1665 $cm^{-1}$ could be represented by the 1645-1655 $cm^{-1}$ and 1655-1665 $cm^{-1}$ bins. A β-sheet signature in the amide I region of ROA, a couplet negative at ~1658 $cm^{-1}$ and positive at ~1677$cm^{-1}$ (Wen *et.al*, 1994, Barron *et. al.*, 2003 ) could be denoted in the 1665-1675 $cm^{-1}$ bin. An amide III positive band at ~1340 $cm^{-1}$ associated with α-helix in a hydrophobic environment would fall in the 1335-1345 $cm^{-1}$ bin (Barron *et. al.*, 2003, Blanch *et. al*, 2004). ROA studies by Wen and group (1994) on α-lactalalbumin and lysozyme attributed a positive ~1339 $cm^{-1}$ band and ~the negative ~1245 $cm^{-1}$ band to loop structure possibly from binding site helix-loop-helix motifs. Barron *et.al.* (1992) conducted ROA studies on

homologous proteins α-lactalalbumin and bovine serum albumin and suggested that the positive ROA band ~1340 $cm^{-1}$ could arise from reverse turn conformations in the proteins.

Certain bins appearing together in Tables 7.2 and 7.3 could allude to association between different bands where some bands may be boosted by other bands. For example in Table 7.2, the Raman variable importance bins 1665-1675 $cm^{-1}$ and 1675-1685 $cm^{-1}$ appear to be characterized together as important variables in amide I&II, amide I&III and amide I&II&III combinations.

The amide II band region has been used in studies to investigate structural conformations. Oberg *and* group (2004) analysed the IR and CD amide II region and noted that this region is usually left out of structural analyses due to the complexity of the dependency of the bands shapes to structural conformation. A study by Jacob *et. al.* (2009) demonstrated that the amide II band in Raman spectra on poly-alanine peptides showed a large intensity for $3_{10}$ helix and weak visibility for α helical conformation. A similar observation also reported by Gangani *et. al*. (2002) who found that VCD amide II band showed the same trend for $3_{10}$ helix and α helix signals. Navea *et. al.* (2005) used Partial Least Square analysis to analyse structural content of IR and cited the amide II region as hardly being used in the determination of structure content of proteins due to overlapping of $D_2O$ bands and other bands which makes it difficult to make structural motif assignments. The RF variable importance analyses reported bins 1525-1535 $cm^{-1}$ from Raman spectra in Table 7.2 and 1535-1545 $cm^{-1}$ from ROA spectra as important variables with values of 0.036 and 0.035 respectively (Tables 7.2 and 7.3). This could imply that they are possibly some regions in the amide II region that could contain structural information.

The Random Forest algorithm proved sensitive to structural information in ROA and Raman spectra although ROA seemed to be more sensitive to the algorithm as the ROA analyses

reported higher accuracies than Raman analyses (Kinalwa *et.al*., 2010). Our variable importance analyses showed which spectral regions were significant in the classification decision some of which have been reported in literature but also others like those from the amide II region which is usually omitted from structural studies.

## 7.3 References


Barron L.D., Blanch E. W., McColl I.H., Syme C.D., Hecht L., Nielsen K. (2003)Structure and behaviour of proteins, nucleic acids and viruses from vibrational Raman optical activity.*Spectroscopy* **17**,101-126

Barron L.D., Cooper A., Ford S.J., Hecht L., Wen S.Q. (1992) Vibrational Raman optical activity of enzymes. *Faraday Discussions* **93**, 259-268.

Blanch E. W., McColl I.H., Hecht L., Nielsen K., Barron L.D., Structural characterization of proteins and viruses using Raman optical activity (2004)*Vibrational Spectroscopy* **35**,87-92

Chen P., Rose J, Love R,Wei C.H., Wang BC (1992) Reactive sites of an anticarcinogenenic Bowman-Birk proteinase inhibitor are similar to other trypsin inhibitors. *Journal of Biological Chemistry* **267**,1990-1994

Chrysina E.D., Brew K., Acharya K.R. (2000) Crystal structures of apo- and holo-bovine alpha-lactalbumin at 2. 2-A resolution reveal an effect of calcium on inter-lobe interactions. *Journal of Biological Chemistry* **275**,37021-37029

Gangani R. A., Silva D., Yasui S. C., Kubelka J., Formaggio F., Crisma M., Toniolo C., Keiderling T.A. (2002) Discriminating 310- from α-helices: Vibrational and electronic CD and IR absorption study of related aib-containing oligopeptides. *Biopolymers* **65**, 299-243.

Jacob C.R., Luber S., Reiher M. (2009) Analysis of secondayr effects on the IR and Raman spectra of polypeptides in terms of localized vibrations. *Journal of Physical Chemistry* **113**, 6558-6573.

Kinalwa M., Blanch E. W., Doig A. J. (2010) Determination of Protein Fold Class from Raman or Raman Optical Activity Spectra using Random Forest. *Protein Science* **20**, 1668–1674

Maiti N.C., Apetri M. M., Zagorski M. G., Carey P. R., Anderson V. E. (2004) Raman spectroscopic characterization of secondary structure in natively unfolded proteins: α-Synuclein. *Journal of the American Chemical Society* **126**,2399-2408

Nardone E, Rosano C. , Santambrogio P., Curnis F., Corti A., Magni F., Siccardi A.G., Paganelli G., Losso R., Apreda B., Bolognesi MSidoli ., A., Arosio P. (1998) Biochemical characterization and crystal structure of a recombinant hen avidin and its acidic mutant expressed in *Escherichia* coli . *European Journal of Biochemistry* **256**,453-460

Navea S., Tauler R., Juan A. (2005) Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure.*Analytical Biochemistry* **336**,231-242

Oberg K.A., Ruyssychaert J., Goormaghtigh E. (2004) The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *European Journal of Biochemistry* **271**, 2937-2948

Schonfeld D.L., Ravelli R.B., Mueller U, Skerra A. (2008) The 1.8-A crystal structure of alpha 1-acid glycoprotein (Orosmucoid) solved by UV RIP reveals the broad drug-binding activity of this human plasma lipocalin. *Journal of Molecular Biology* **384**,393-405

Takano K., Yamagata Y, Yutani K.,(2000)Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry* **39**,8655-8665

Wen Z.Q., Hecht L., Barron L.D. (1994) α-Helix and associated loop signatures in vibrational Raman optical activity spectra of proteins. *Journal of American Chemical Society* **116**, 443-445

Wen Z.Q., Hecht L., Barron L.D. (1994) β-Sheet and associated turn signatures in vibrational Raman optical activity spectra of proteins. *Protein Science* **3**, 435-439

# Chapter 8

# Conclusion

The machine learning analyses in this thesis have been used to extract structural information from Raman and ROA spectral data. The analyses were applied to spectral data of 10 $cm^{-1}$ bin widths. Full spectrum data from 605 and 1785 $cm^{-1}$ were analyzed, as well as the amide I, II, and III regions in isolation and in different combinations. Partial Least Squares (PLS) regression was used on Raman and ROA spectra to determine the secondary structure contents (α-helix, β-sheet, or other) of proteins with high accuracy. Analysis was also carried out using second derivative Raman data yielding and root mean square (rms) values of 0.99, and 0.6-1.7%, respectively. The results shown here out performed similar chemometric analyses applied to data from circular cichroism (CD), a high throughput method used for the measurement of α-helix and β-sheet structure content. This regression analysis to the different spectral amide regions showed the importance of these regions in secondary structure analysis.

The Random Forest algorithm was used to determine protein fold class from Raman or ROA spectra. The Raman analysis showed poorer performance then ROA analysis. The best performance was obtained from ROA amide II and III data with 75% of the proteins correctly assigned to α-helix, β-sheet, αβ and other SCOP class. The SVM classification and SVM regression analysis showed poor accuracy compared to the previous two types of analyses. The varying performance accuracies of the methods discussed highlights the importance of the choice of machine learning method to be used to investigate the problem under study. The high accuracies obtained showed the potential of using machine learning methods to mine

and extract information on protein structure and the application of these methods in as useful tools in the determination of protein structure.

Appendix A – Summary table of proteins used in the analyses with pdb code and SCOP class.

| | Protein name | PDB code | SCOP classification |
|---|---|---|---|
| ABA-1 allergen (Ascaris lumbricoides) | | | |
| filamentous bacteriophage fd | FD bacteriophage[1] | 1ifj | coiled coil proteins |
| filamentous bacteriophage IKe | IKE virus | 1ifi | coiled coil proteins |
| filamentous bacteriophage M13 | M13 virus | 1ifl | coiled coil proteins |
| filamentous bacteriophage Pf1 | PF1 virus | 1pfi | coiled coil proteins |
| α-helical poly(L-glutamic acid) | α helix polyglutamic acid | | α helical |
| α-helical poly(L-lysine) | α helical polylysine | | α helical |
| S100A6 (calcyclin, rabbit) | | 2cnp | α helical |
| S100B (rabbit) | | | α helical |
| serum albumin (human) | human serum albumin | 1ao6 | all α proteins |
| prion protein (ovine,90-230) | helical domain of S23 | | |
| α-gliadin (wheat) | α gliadin | | unfolded |
| aldolase (rabbit) | aldolase | 1ado | α and β proteins(a/b) |
| creatine kinase (rabbit) | creatin kinase | 2crk | all α |
| α-lactalbumin (bovine) | bovine α lactalbumin | 1f6s | α and β proteins (a+b) |
| α-lactalbumin (human) | human α lactalbumin | 1b9o | α and β proteins (a+b) |
| lysozyme (equine) | equine lysozyme | 2eql | α and β proteins (a+b) |
| lysozyme (hen) | hen lysozyme | 1lse | α and β proteins (a+b) |
| lysozyme (human) | human lysozyme | 2vb1 | α and β proteins (a+b) |
| narcissus mosaic virus | narcissus mosaic virus[3] | | α and β proteins (a+b) |
| papaya mosaic virus | papaya mosaic virus[10] | | α helical proteins |
| potato virus X | pvx virus[9] | | α helical proteins |
| regulatory protein 2 (Streptococcus pyogenes) | | | |
| tobacco mosaic virus | tobacco mosaic virus[11] | 2tmv | α helical proteins |
| tobacco mosaic virus dimer | | | α helical proteins |
| tobacco rattle virus | tobacco rattle virus | | α helical proteins |
| prion protein (mouse, 23-231) | | | |
| amylase (Bacillus licheniformis) | amylase | 1bli | all β proteins |
| response regulator O2 receiver domain (Streptococcus pneumoniae) | | | |
| molten globule state ovalbumin (hen) | | | |
| ovalbumin (hen) | | 1ova | |
| ovomucoid (hen) | hen ovomucoid | | small proteins |
| ovomucoid (turkey) | turkey ovomucoid | 1m8c | α/β |
| ribonuclease A (bovine) | ribonuclease A | 1rph | α and β |
| ribonuclease B (bovine) | ribonuclease B | | α and β |
| subtilisin Carlsberg (Bacillus licheniformis) | subtilisin | 1sca | α and β proteins (a/b) |
| invertase | β proteins | 2ac1 | N/A |
| β-chymotrypsin (bovine) | α chymotrypsin | 4cha | all β proteins |
| β-lactoglobulin (bovine, pH=2.0) | bovine β lactoglobulin | 1cj5 | all β proteins |
| β-lactoglobulin (bovine, pH=6.5) | bovine β lactoglobulin(pH6.8) | 1cj5 | all β proteins |
| β-lactoglobulin (bovine, pH=9.0) | bovine β lactoglobulin(pH9) | 1cj5 | all β proteins |
| MS2 virus (empty protein capsid) | MS2 virus | 1ms2 | α and β proteins (a+b) |
| pepsin (porcine) | pepsin | 3pep | all β proteins |
| satellite tobacco mosaic virus | satellite tobacco mosaic virus | 1a34 | all β proteins |
| trypsin (bovine) | trypsin | 1k1l | all β proteins |
| trypsinogen (bovine) | trypsinogen | 1tgt | all β proteins |
| ubiquitin (bovine) | ubiquitin | 1g6j | α and β proteins (a+b) |
| serum amyloid protein(domain a) | serum amyloid proteins | 1lgn | all β proteins |
| avidin (hen) | avidin | 1rav | all β proteins |
| concanavalin A (jack bean) | concanavalin | 2cna,2ctv | all β proteins |
| cowpea mosaic virus (empty protein capsid) | cowpea mosaic virus | 2bfu | all β proteins |
| cowpea mosaic virus (with RNA 1) | same as above with RNA | 1ny7 | all β proteins |
| cowpea mosaic virus (with RNA 2) | same as above with RNA | 1ny7 | all β proteins |
| immunoglobulin G (human) | immunoglobin G | 1hzh | α and β proteins (a+b) |
| P.69 pertactin (Bordetella pertussis) | P69 pertactin | 1dab | all β proteins |
| β-sheet poly(L-lysine) | polylysine | | all β proteins |
| serum amyloid P component (human) | human serum amyloid protein | 1sac | all β proteins |

| | | | |
|---|---|---|---|
| antifreeze protein type III (arctic flounder) | antifreeze protein | | Mainly disordered |
| T-A-1 peptide (wheat glutenin subunit) | | | Mainly disordered |
| Bowman-Birk inhibitor (soybean) | bowman-birk inhibitor | 1pi2 | disordered/Mainly disordered |
| α-casein (bovine) | α casein | | Mainly disordered |
| β-casein (bovine) | β casein | | Mainly disordered |
| κ-casein (bovine) | kappa-casein | | Mainly disordered |
| ω-gliadin (wheat) | omega gliadin | | Mainly disordered |
| α-synuclein (human) | α synuclein | 1xq8 | Mainly disordered |
| β-synuclein (human) | β synuclein | | Mainly disordered |
| γ-synuclein (human) | gamma synuclein | | Mainly disordered |
| tau46 (human P301L mutant) | mutants of tau proteins | | Mainly disordered |
| tau46 (human) | tau protein | | Mainly disordered |
| metallothionein (rabbit) | rabbit metallothionein[7,8] | | All disordered |
| OOAAAAAAAOO peptide | | | All disordered |
| phosvitin (hen) | phosvitin | | All disordered |
| disordered poly(L-glutamic acid) | disordered polyglutamic acid | | All disordered |
| disordered poly(L-lysine) | disordered polylysine | | All disordered |
| collagen structure | collagen structure | | All disordered |
| α-1-acid glycoprotein (bovine) | orosomucoid[6] | | unfolded (glycoprotein) |
| A-gliadin in 60% MeOH | α gliadin | | α disordered |
| insulin (bovine, monomeric, low PH) | insulin | 2a3g | small proteins |
| α-lactalbumin (bovine, A-state) | α lactalbumin | 1f6r | α and β proteins (a+b) |
| lactoferrin (human) | lactoferrin[4,5] | 1cb6 | α and β proteins |
| rF1 antigen | lactoferrin antigen | | |
| lysozyme (equine, A-state) | equine lysozyme | 2eql | α and β proteins (a+b) |
| lysozyme (hen, reduced) | reduced lysozyme | 2vb1 | unfolded |
| lysozyme (human, pre-fibillar intermediate) | human lysozyme partially unfolded | 1gaz | disordered |
| prion protein (ovine, 94-233, reduced) | helical domain of S23 | 1qm1 | all α |
| full length prion protein (ovine,23-230) | ovine prion protein | | α and β proteins (a+b) |
| ribonuclease A (bovine, reduced) | reduced ribonuclease A | | unfolded |
| disordered poly(L-alanine) | disordered polyalanine | | all disordered |
| α-helical poly(L-alanine) | α helical polyalanine | | α helical |
| α-helical poly(L-alanine) | α helical polyalanine | | α helical |
| α-helical poly(L-alanine) | α helical polyalanine | | α helical |
| β-sheet poly(L-alanine) | β sheet polyalanine | | β sheet |
| α-helical poly(benzyl) | α helical poly benzyl | | α helical |
| α-helical poly(benzyl) | α helical benzyl | | α helical |
| disordered poly(L benzyl) | disordered poly benzyl | | all disordered |
| disordered poly(L-glutamic acid) | disordered polyglutamic acid | | all disordered |
| disordered poly(L-glutamic acid) | disordered polyglutamic acid | | all disordered |
| disordered poly(L-histidine) | disordered polyhistidine | | all disordered |
| α-helical poly(L-histidine) | α helix polyhistidine | | α helical |
| disordered poly(L-leucine) | disordered polyluecine | | all disordered |
| α-helical poly(Lleucine) | helical polyleucine | | α helical |
| α-helical poly(L-lysine) | α helical polylysine | | α helical |
| disordered poly(L-ornathine) | disordered polyornathine | | all disordered |
| α-helical poly(L-ornathine) | α helical polyornathine | | α helical |
| α-helical poly(L-ornathine) | α helical polyornathine | | α helical |
| disordered poly(L-proline) | disordered polyproline | | all disordered |
| disordered poly(L-proline) | disordered polyproline | | all disordered |
| disordered poly(L-tryptophan) | disordered polytryptophan | | all disordered |
| α-helical poly(L-tryptophan) | α helical polytrptophan | | α helical |
| disordered poly(L-threonine) | disordered polythreonine | | all disordered |
| disordered poly(L-tyrosine) | disordered polytyrosine | | all disordered |
| α-helical poly(L-tyrosine) | helical polytyrosine | | α helical |
| disordered poly(L-lysine) | disordered polylysine | | all disordered |
| disordered poly(L-lysine) | disordered polylysine | | all disordered |

## Appendix B- **The binAvg.pl script used to bin the data**

```
#!/usr/local/bin/perl

use POSIX;
use FileHandle;

use Getopt::Long;


my $label = 1;

my $result = GetOptions(

    "outfile=s" => \$outfile,

    "label=i" => \$label,


    )or die "$!\n";



#$numArgs = $#ARGV + 1;

#print "you entered $numArgs command-line arguments.\n";


#foreach $argnum(0..$#ARGV)

        #{

                #print "$ARGV[$argnum]\n";

        #}


#print @ARGV," the ARGV array\t\n";




sub midpoint

     {

         ($low,$high) = (@_);

             $middle = ($low+$high)/2;
```

```
            return $middle;
      }


sub roundoff
{
 my ( $x, $factor) = @_;
 $b = sprintf("%.0f", floor($x));
 $div =  sprintf("%.0f",floor($b/$factor));
 $value = $div*$factor;
 return $value;
}


sub getFiles#not working yet
   {
        my $argsref =  shift(@_);
        my @fileArray;
        foreach (@{$argsref})
               {
                      print  "$_\n";
                   #push(@fileArray,"$_");
            }
     #print @fileArray;
     #return (\@fileArray);

   }

 sub correctFilename
     {
            my $spectralFile =  shift(@_);
            $filenamelength = length($spectralFile);
```

```
         #get file extension length

         $filextension = $filenamelength-3;

         #get file extension txt

         $extensionfragment  = substr($spectralFile, $filextension,
$filenamelength);

         $extensionstring = ".";

         $extensionstring .=  $extensionfragment ;

         #modify last segment of $spectralFile to '.txt'

         substr( $spectralFile, $filextension, $filenamelength)= $extensionstring;

         return $spectralFile;

   }




 sub getFile

   {

        my $spectralFile =  shift(@_);

        #$spectralFile = correctFilename($spectralFile);

        #print $spectralFile;

     open (FILE, $spectralFile) or die ("Error opening $argsref: $!");

     chomp;

     while (<FILE>){

          unless ($_=~/^\s+$/){ #get rid on newlines in file

       @lines = <FILE>;

     }

     return \@lines;


   }

}


   sub receive_array

   {
```

```
   my $argsarray =  shift(@_);
   my $class = shift(@_);
   my $filename = shift (@_);
   my $vector = "$class  ";
   local @store_lines = @{$argsarray};
  ($lo_bound,$hi_bound) = split (/\s+/,$store_lines[0]);
   $round_lobound = roundoff($lo_bound,10);
   $size = scalar @store_lines;
   #print $round_lobound,"\n";
   #print $store_lines[$#store_lines],"\n";
   #print $#store_lines,"\n";
   #print $size,"\n";
   #$fh = new FileHandle ">> $outfile";
   $filenamelength = length($filename);
    #get file extension length
$filextension = $filenamelength-3;
$outfile  = substr($filename,0,$filextension);
   my $trainfile = $outfile."train.data";
   #print $trainfile;
   $fh = new FileHandle "> $trainfile";
   #foreach $line (@{$argsarray}){
          #print $line,"\n";
   #}
   for($i=$round_lobound; $i<$store_lines[$#store_lines]; $i+=20)
  {
       $lo = $i;
       #print "start is $lo\n";
        for ($x =0; $x<20; $x++)
          {
            $lo += 1;

            #print "x is $x\n";
```

```
        }
      #print "range is $i - $lo\n";
 for my $element (0..$#store_lines)
  {
   ($x,$y) = split(/\s+/,$store_lines[$element]);
   #print $y,"\n";
        if($x > $i && $x < $lo)
              {
                  $ttIntensity += $y;
                  $count_elements++;
               #print $store_lines[$element],"\n";

             }#end if

    }   #end for element loop

    $mid = midpoint($i,$lo);
    chomp($mid);
    $avgIntensity = $ttIntensity/$count_elements;
    #print "Total wavenumber in this bin is $ttIntensity\n";
    #print "The mid point of this range is $mid     $avgIntensity\n";
    #print "there are $count_elements elements in this range\n";
    #print "The average intensity in this range is $avgIntensity\n";

    $vector .= "1:$avgIntensity ";
    if (defined $fh) { print $fh $vector;}
    $count_elements = 0;
    $ttIntensity = 0;
    $avgIntensity = 0;
    $vector = ' ';
}#end for round_lobound
print $fh "\n";
```

```
          $fh->close;
  }

#$fh = new FileHandle ">> outtest.txt";

foreach $file (@ARGV){
        #print "\n";


##how to pass array to bin subroutine
#$a=getFiles(\@ARGV);
$a=getFile($file);
#$a=getFile(testtext.txt);
@b = @$a;
receive_array(\@b,$label,$file);

}
```

Appendix C- The *ranges.pl* script used to select amide regions

```
#!/usr/local/bin/perl
use POSIX;
use FileHandle;


#$spectralFile = 'ABA-1A.txt';

#@array = (100..900);

my $buildString = ' ';

#########################################################################
########################
#while ($check ne 'q') {
#    print "enter lower bound of range:";

#    chomp(my $first = <STDIN>);



#    print "enter upper bound of range:";

#    chomp(my $second = <STDIN>);

#     $buildString .= "$first".."$second.",";


#    print "Enter q to quit or press 'Enter' to continue\n";

#    chomp($check = <STDIN>);
    #print $check,"\n";
#}

#chop($buildString);




#    $temp2 = 0;
#    @temp=split(/,/,$buildString);




#    foreach $foo (@temp)
#    {

#      $ranges[$temp1][$temp2]=$foo;
#     $temp1++;
#     $temp2 += 1;

#    }
```

```
foreach $spectralFile (@ARGV){

$rangeFile = $spectralFile."range.txt";


open (FILE, "$spectralFile") or die ("Error opening $argsref: $!");
     chomp;
    while ($line = <FILE>){
        unless ($_=~/^\s+$/){ #get rid on newlines in file
            @trio = split(' ',$line);
            #print $line,

            #@lines = <FILE>;

         }
     }

print $rangeFile,"\n";

$fh = new FileHandle "> $rangeFile";

#AmideI+II
#@ranges = ([1510..1580],[1600..1700]);

#AmideI+III
#@ranges = ([1200..1340],[1600..1700]);

#Amide II+III
#@ranges = ([1200..1340],[1510..1600]);

#AmideI+II+III
#@ranges = ([1200..1340],[1510..1580],[1600..1700]);

#for single range

#AmideI
#@ranges = ([1600..1700]);

#Amide II
#@ranges = ([1510..1600]);

#Amide III
#@ranges = ([1200..1340]);

#whole range
#@range = ([500..1800]);

#Backbone region
@range = ([850..1100]);

my $label = shift(@trio);

if (defined $fh) { print $fh $label," ";}

foreach $v (@trio){
  ($index,$waveNumber,$intensity) = split(':',$v);
```

```
   #print $waveNumber,"\n";
   for (my $x = 0; $x <= $#ranges; $x++) {
        #for (my $y = 0; $y <= $#{$ranges[$x]}; $y++) {


            #print ' ';
            my $firstValue =  $ranges[$x][0];
            #print $ranges[$x][0],"\n";


           my $secondValue = $ranges[$x][$#{$ranges[$x]}];
            #print $ranges[$x][$#{$ranges[$x]}],"\n";

   if ($waveNumber >= $firstValue && $waveNumber <= $secondValue){
            print $waveNumber,"\n";
            if (defined $fh) { print $fh $waveNumber.":".$intensity,"
";}
          }


   }



 }


close FILE;
}
```

Appendix D- The svmscale.pl used to scale the data

```
#!C:/Perl/bin/perl.exe

use FileHandle;

$max = 1;
$min = -1;
$scalefile = "scale.train.data";
sub openfile {
     $outfile = shift(@_);
open (FILE, $outfile) or die ("Error opening $argsref: $!");
     chomp;
     while ($line = <FILE>)
     {
          #$line =~ s/^\d+\s+//;
             #print $line,"\n";
          @arrayline = split(/\s+/,$line);
          #$data .= $line;




     }
close FILE;
return \@arrayline

}

sub findrange {
```

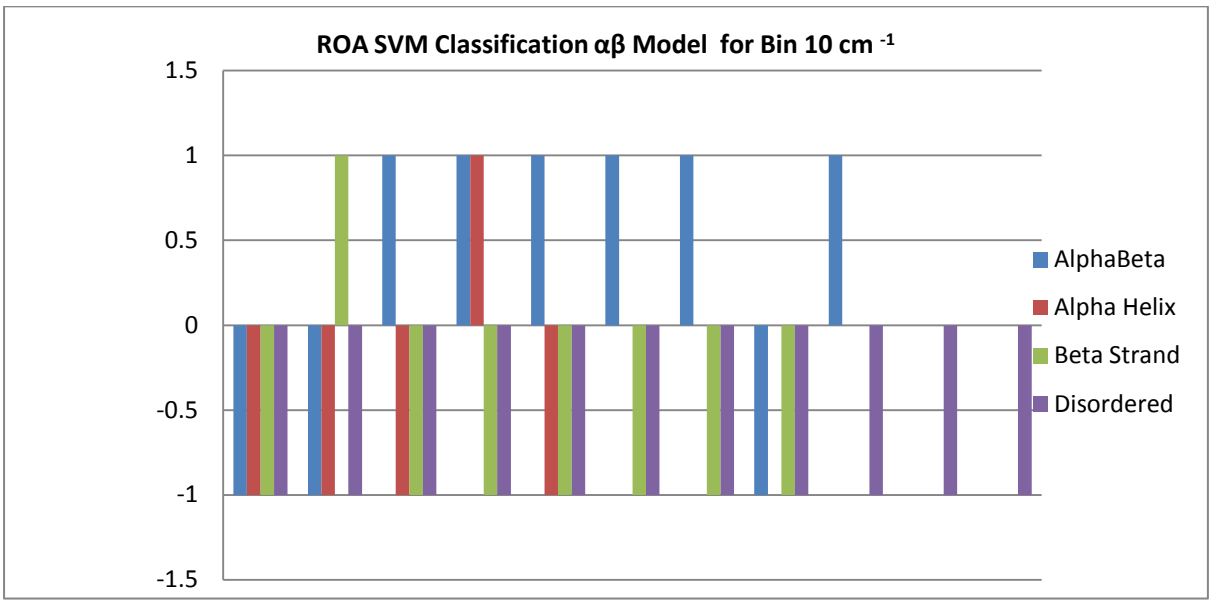
```
      (@arrayline) = @{(shift)};
         #print "@arrayline";
         shift @arrayline;
       foreach my $val (@arrayline) {
                     #print $val,"\n";
                ($f,$v) = split(/:/,$val);
                #push (@values,$v);
                      #print $f,"\n";
                if ($v < $min){
                    $min = $v
               }
                if ($v > $max){
                  $max = $v
             }

        }
      print "maximum is ", $max,"\n";
      print "minimum is  ",$min,"\n";
      return ($min,$max);
      $max = 1;
      $min = -1;
}

sub scale {
      $min = shift(@_);
      $max = shift(@_);
      $scaledfile = shift(@_);
      (@arrayline) = @{(shift)};
        my $label = '';
```

```
        my $string = '';


    #$scaledfile =~ /(\w+)\./;
    #print $1,"\n";
    #$outfile  = $1;
    $label = shift(@arrayline);
    #print $label,"\n";
     # my $scalefile = $outfile.".train.scale.data";
      #print $scalefile,"\n";
      $fh = new FileHandle ">> $scalefile";
 foreach my $val (@arrayline) {
                ($f,$v) = split(/:/,$val);
                      #print $f,":",$v,"\n";
                #push (@values,$v);
                $vscale = ((2*($v-$min))/($max-$min))-1;
                #if (defined $fh) { print $fh $f,":",$vscale,"\t";}
                     $string .= "$f:$vscale ";
                #print "$f:$vscale\n ";


       }
       $label .= " $string";
       #print $label;
       if (defined $fh) { print $fh $label;}
       print $fh "\n";
       $fh->close;
       #$label = '';
       #$string = '';

}
```

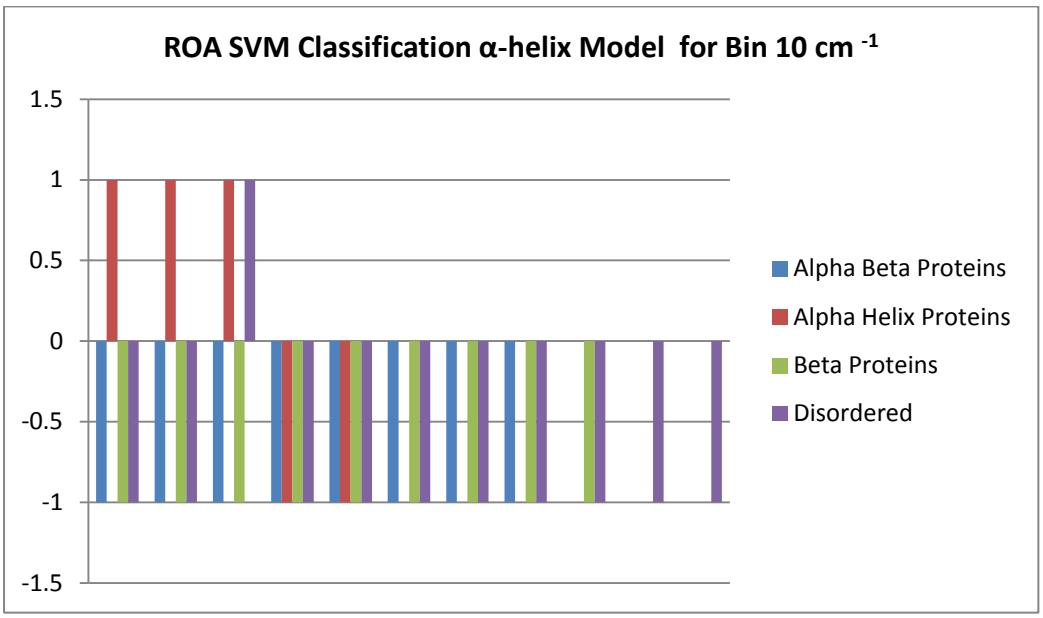
```
@array = ("PLEHs.train.data","PBAHs.train.data");
#change to $file(\@ARGV) to pass files at command line
foreach $file (@ARGV)
{
print  $file,"\n";
$array = openfile($file);
@a = @$array;

#print "@a\n";
($min,$max) = findrange(\@a);
scale($min,$max,$file,\@a);
#print "@a\n";
@a = ();
$min = '';
$max='';
}
```

Appendix E- Graphs showing predictions of SVM classification models (positives are above zero mark and negatives are below the zero mark)

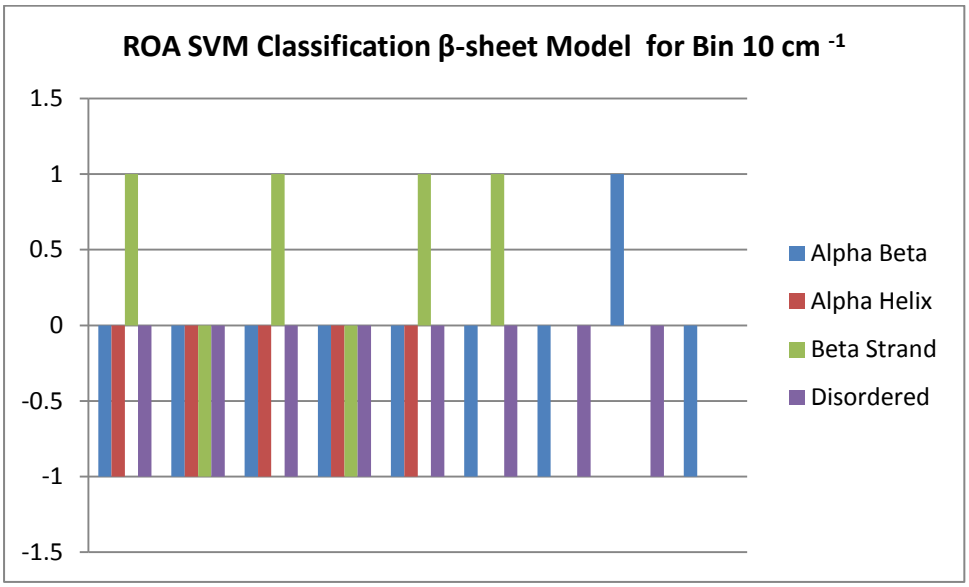
Bar Graph of ROA SVM Classification Full Spectrum αβ Model for Bin 10 cm $^{-1}$



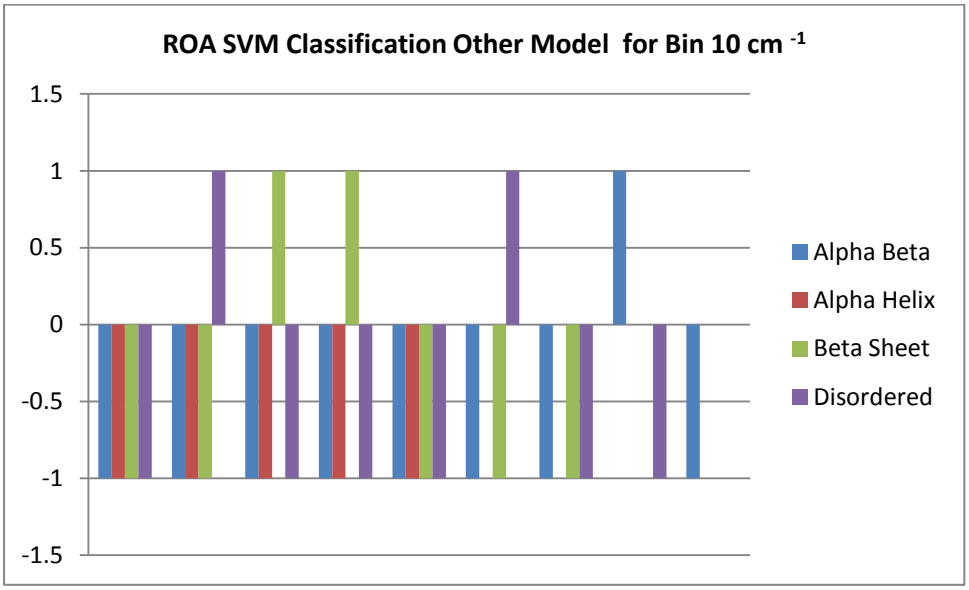

Bar Graph of ROA SVM Classification Full Spectrum α-helix Model for Bin 10 cm $^{-1}$



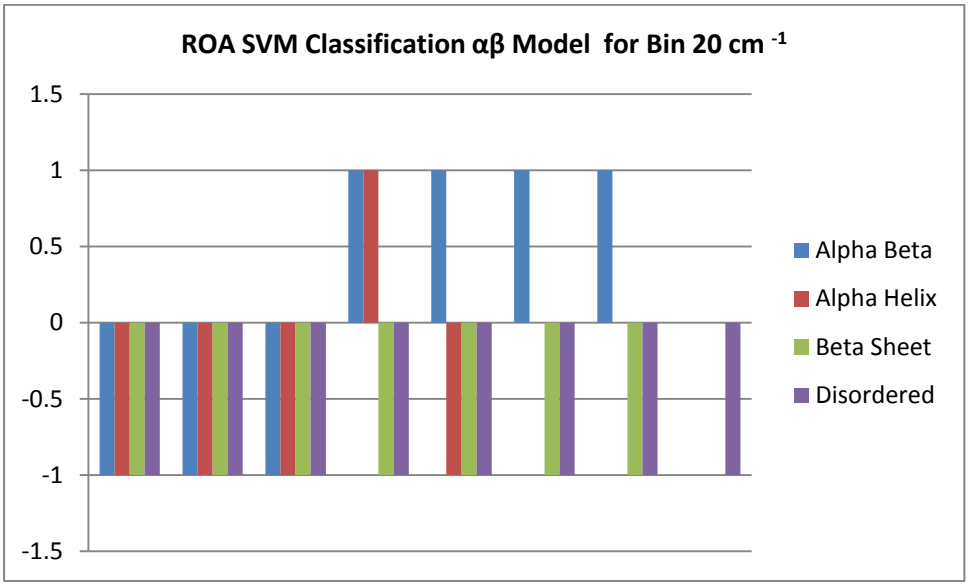
Bar Graph of ROA SVM Classification Full Spectrum β-sheet Model for Bin 10 cm $^{-1}$



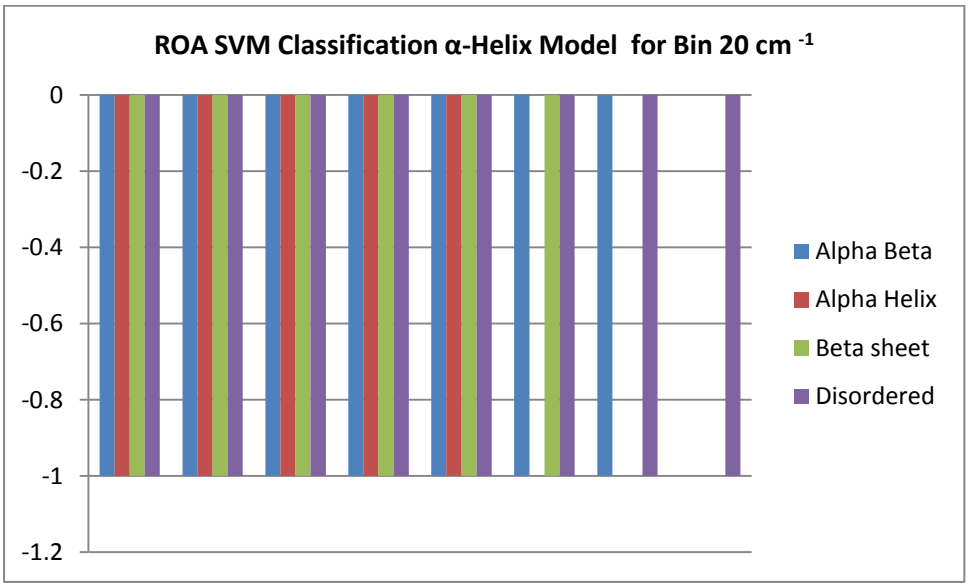

Bar Graph of ROA SVM Classification Full Spectrum Other Model for Bin 10 cm $^{-1}$



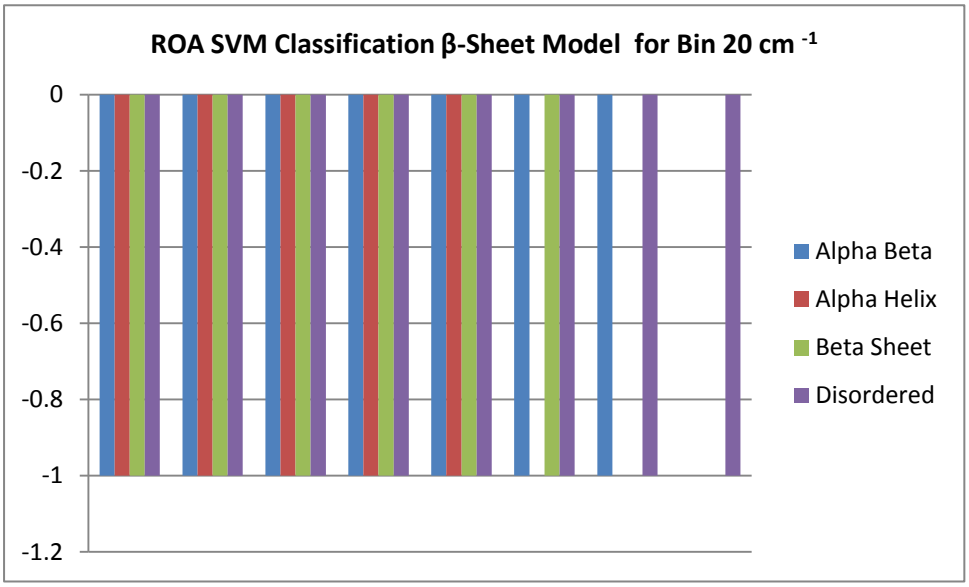
Bar Graph of ROA SVM Classification Full Spectrum αβ Model for Bin 20 $cm^{-1}$



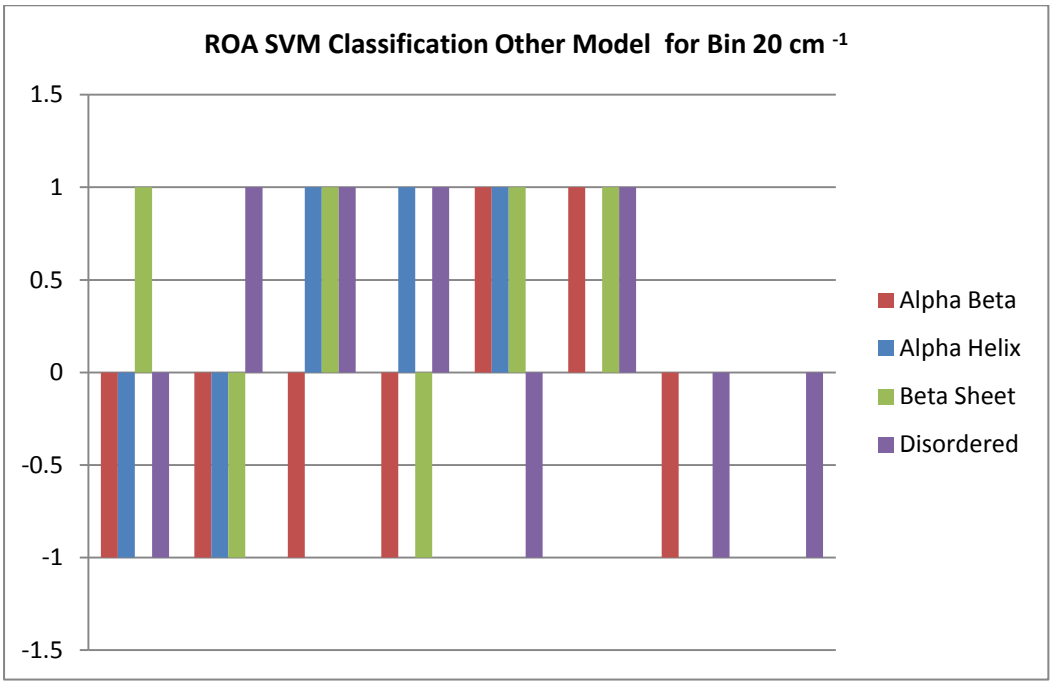

Bar Graph of ROA SVM Classification Full Spectrum α-Helix Model for Bin 20 $cm^{-1}$



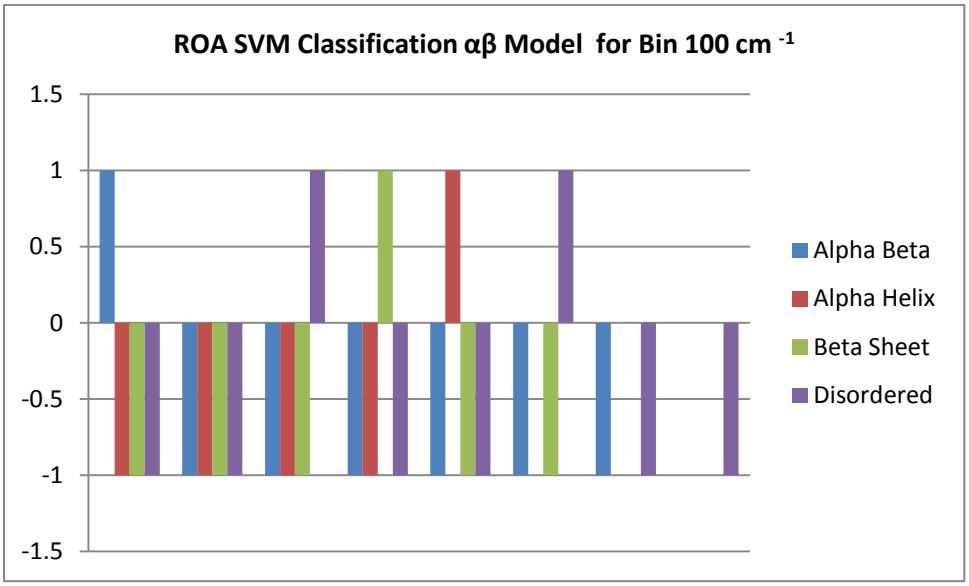
Bar Graph of ROA SVM Classification Full Spectrum β-Sheet Model for Bin 20 cm $^{-1}$



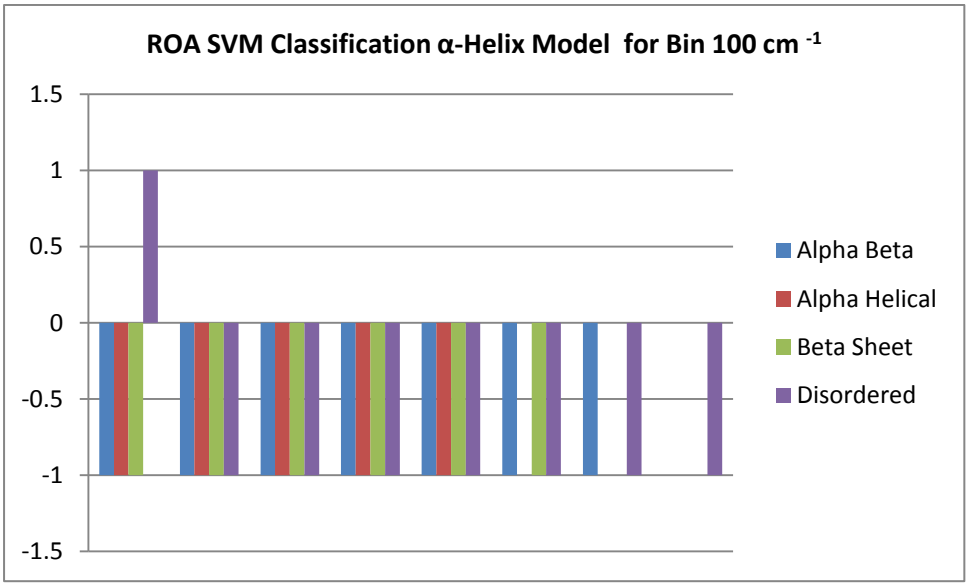

Bar Graph of ROA SVM Classification Full Spectrum Other Model for Bin 20 cm $^{-1}$



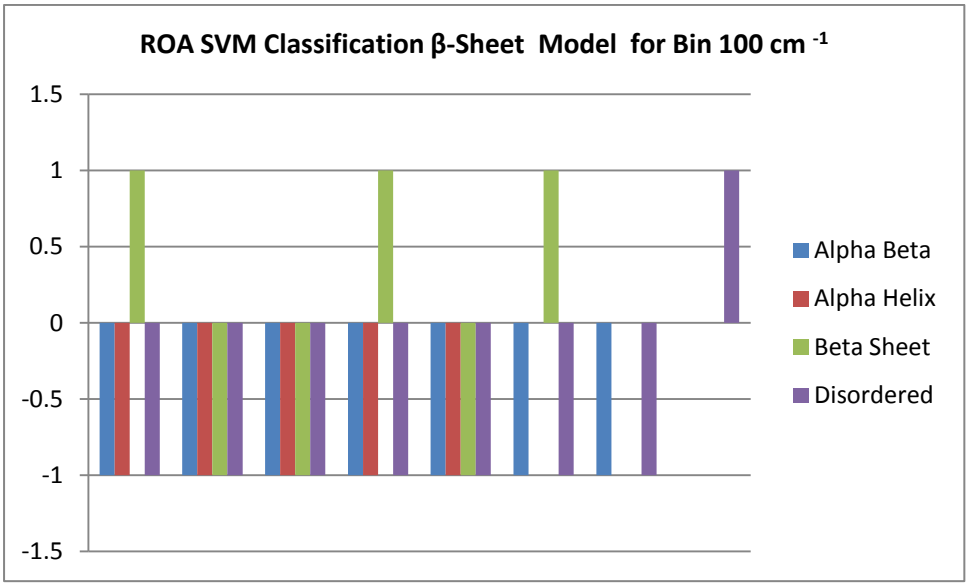
Bar Graph of ROA SVM Classification αβ Model for Bin 100 cm $^{-1}$



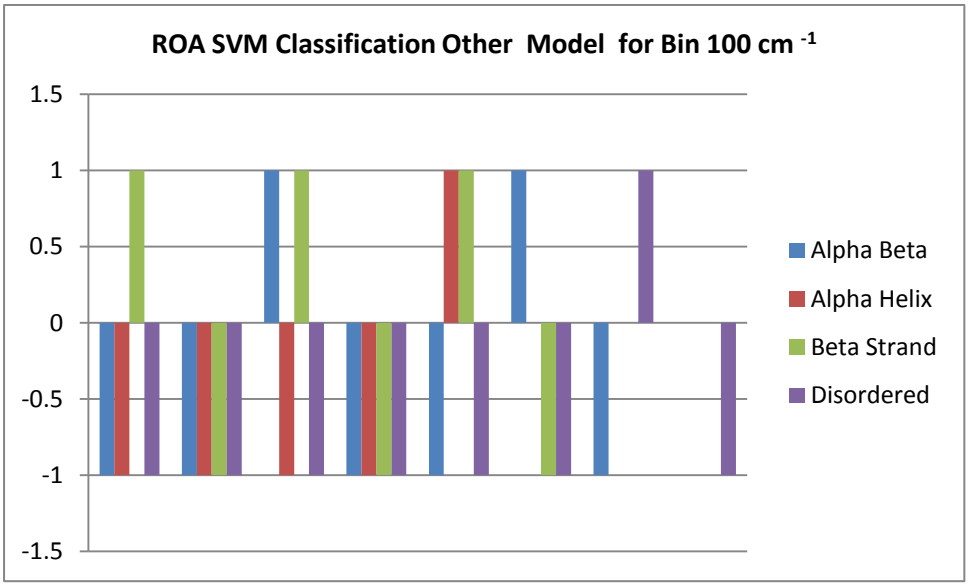

Bar Graph of ROA SVM Classification Full Spectrum α-Helix Model for Bin 100 cm $^{-1}$



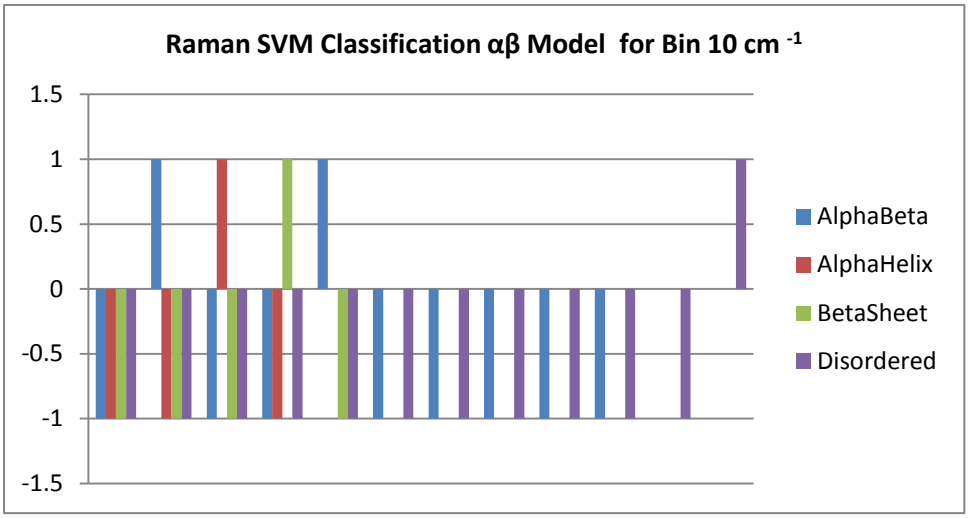
Bar Graph of ROA SVM Classification Full Spectrum β-Sheet Model for Bin 100 cm $^{-1}$



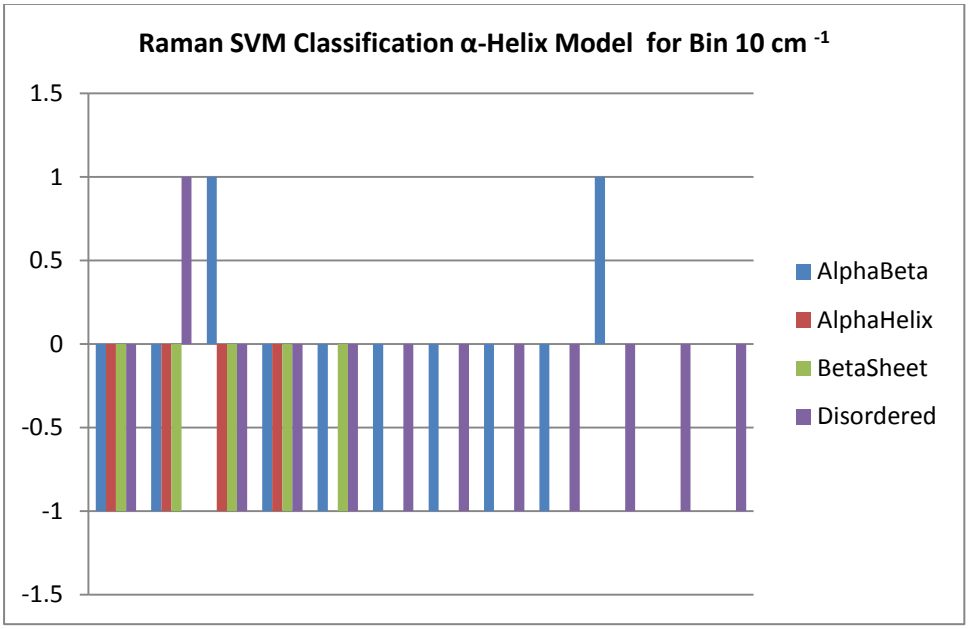

Bar Graph of ROA SVM Classification Full Spectrum Other Model for Bin 100 cm $^{-1}$



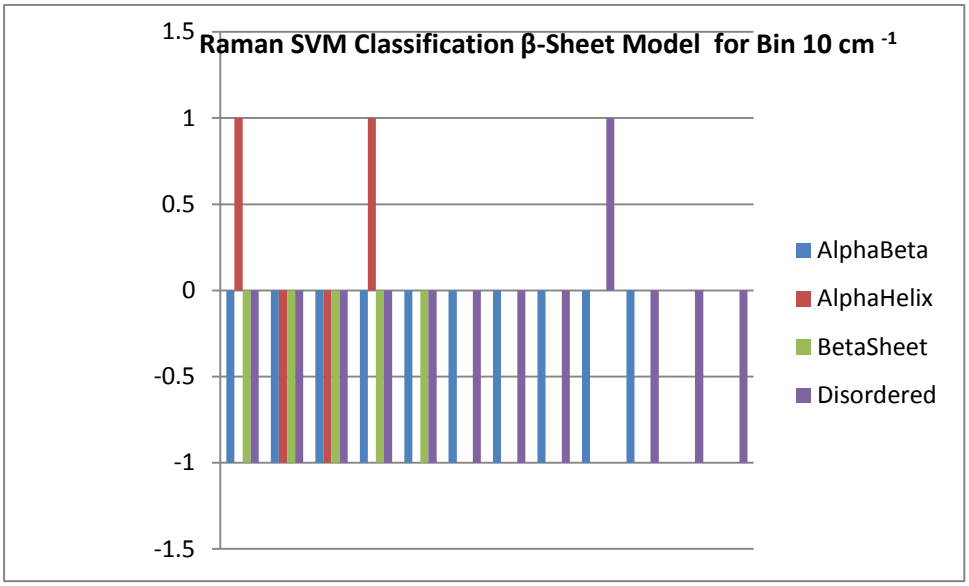
Bar Graph of Raman SVM Classification Full Spectrum αβ Model for Bin 10 cm $^{-1}$



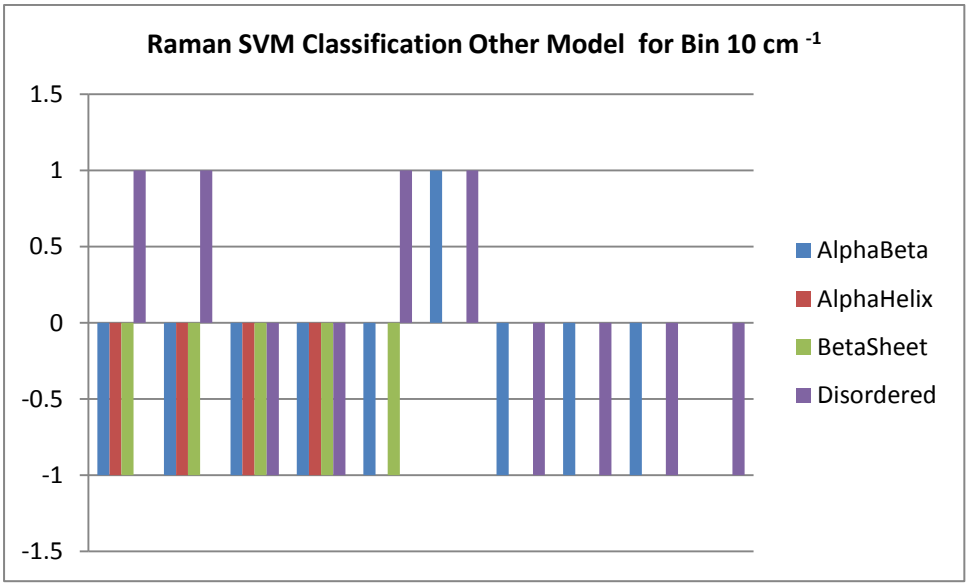

Bar Graph of Raman SVM Classification Full Spectrum α-Helix Model for Bin 10 cm $^{-1}$



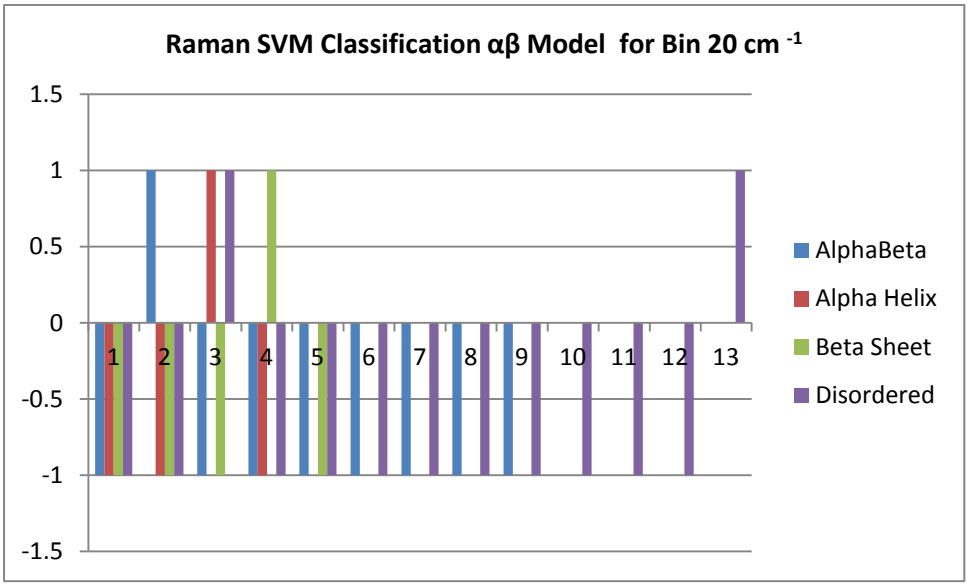
Bar Graph of Raman SVM Classification Full Spectrum β-Sheet Model for Bin 10 $cm^{-1}$



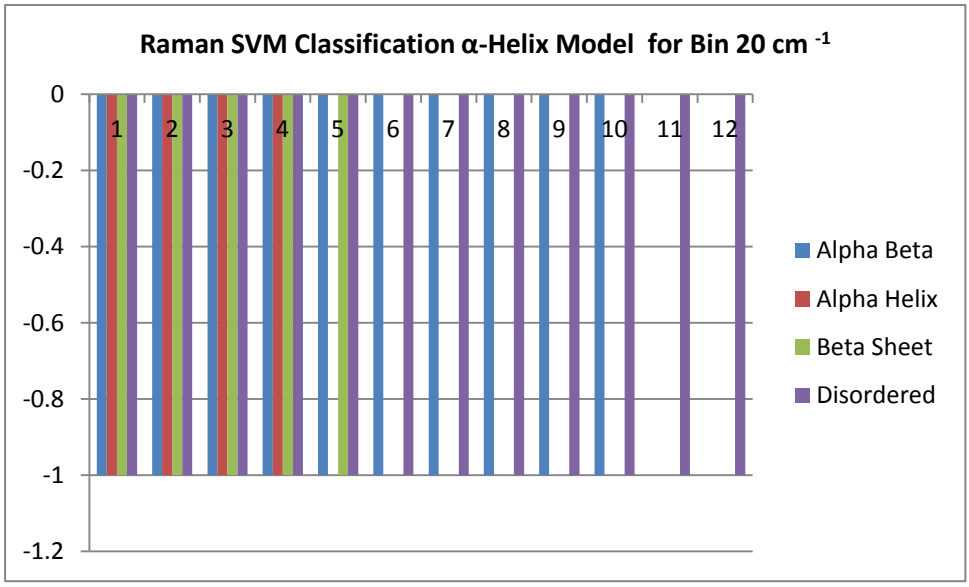

Bar Graph of Raman SVM Classification Full Spectrum Other Model for Bin 10 $cm^{-1}$



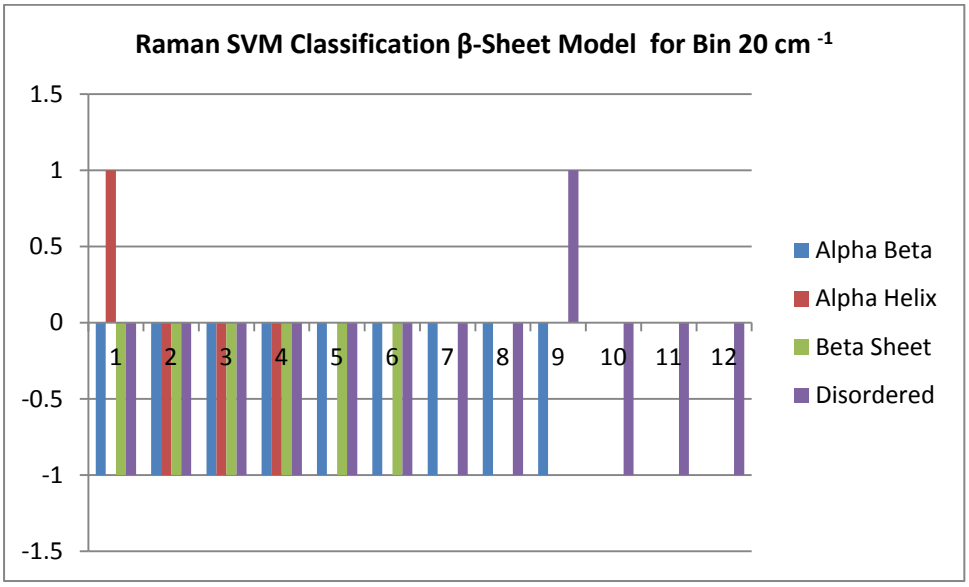
Bar Graph of Raman SVM Classification Full Spectrum αβ Model for Bin 20 cm $^{-1}$



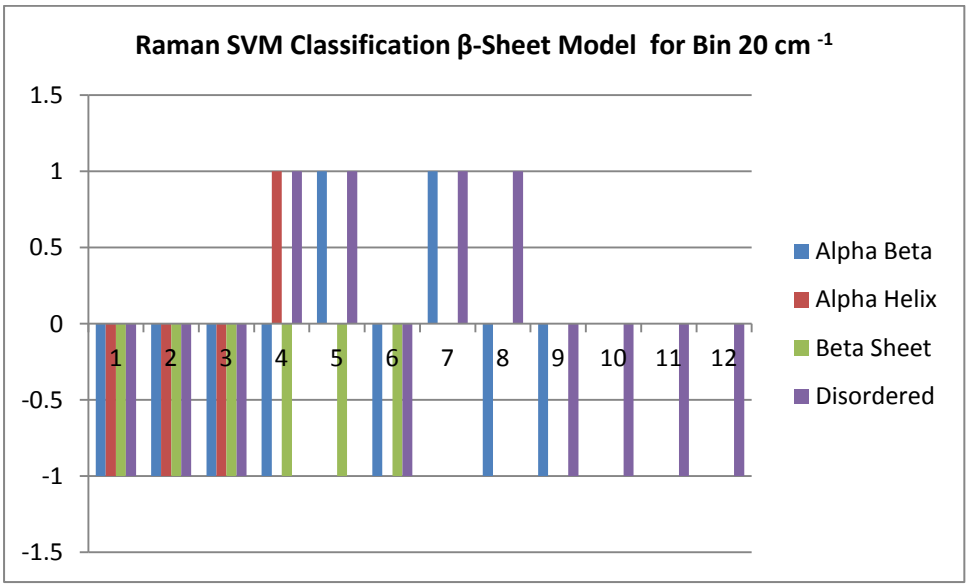

Bar Graph of Raman SVM Classification Full Spectrum α-Helix Model for Bin 20 cm $^{-1}$



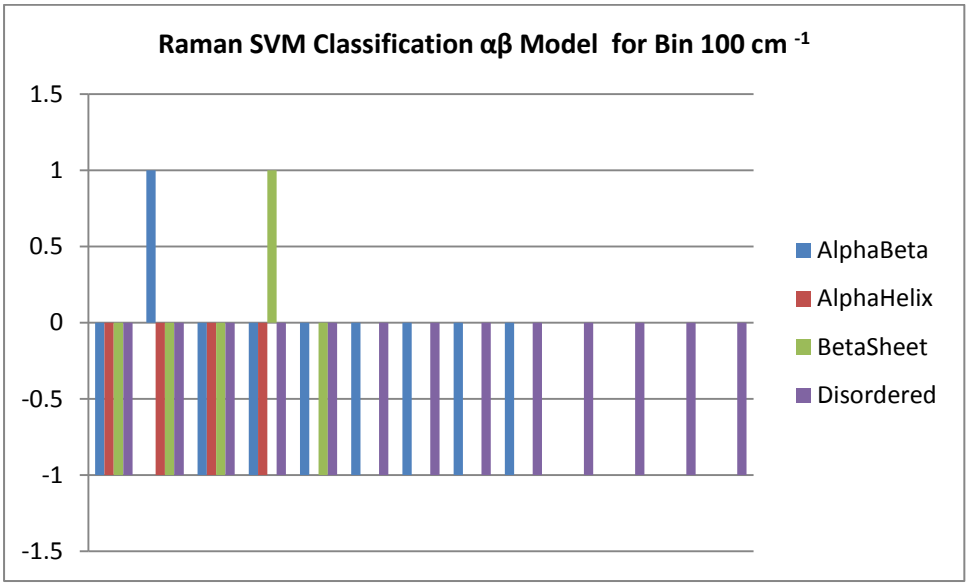
Bar Graph of Raman SVM Classification Full Spectrum β-Sheet Model for Bin 20 $cm^{-1}$



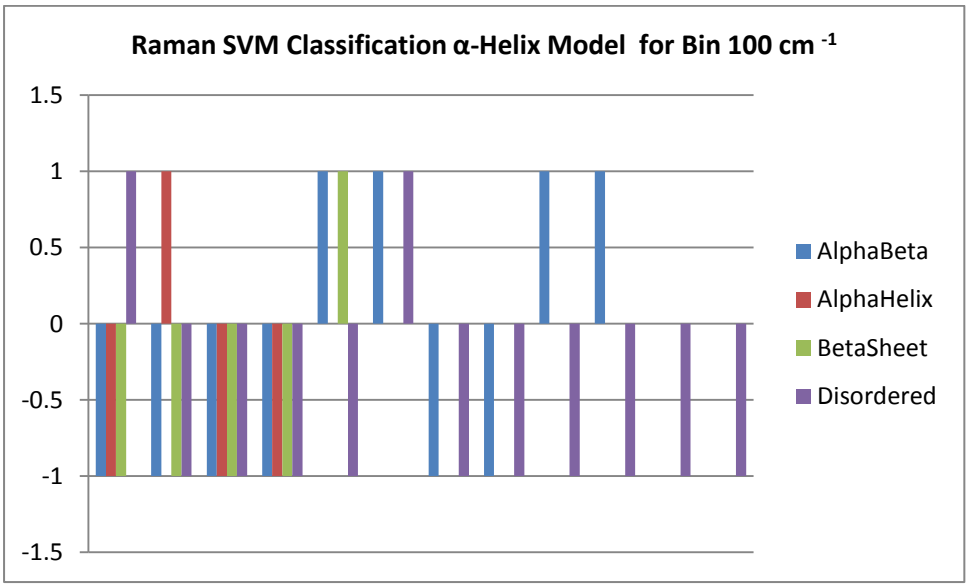

Bar Graph of Raman SVM Classification Full Spectrum β-Sheet Model for Bin 20 $cm^{-1}$



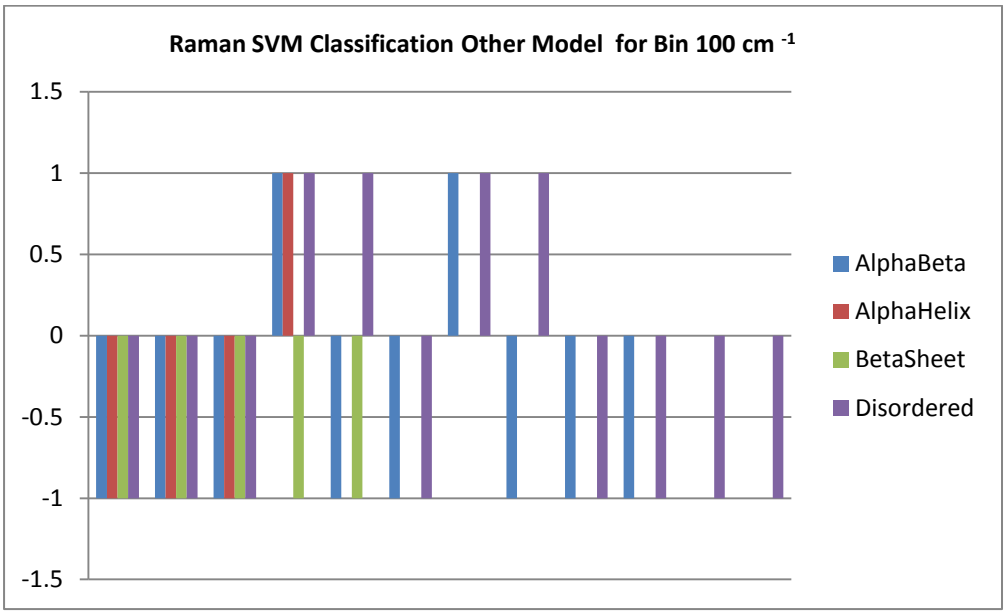
Bar Graph of Raman SVM Classification Full Spectrum αβ Model for Bin 100 cm $^{-1}$



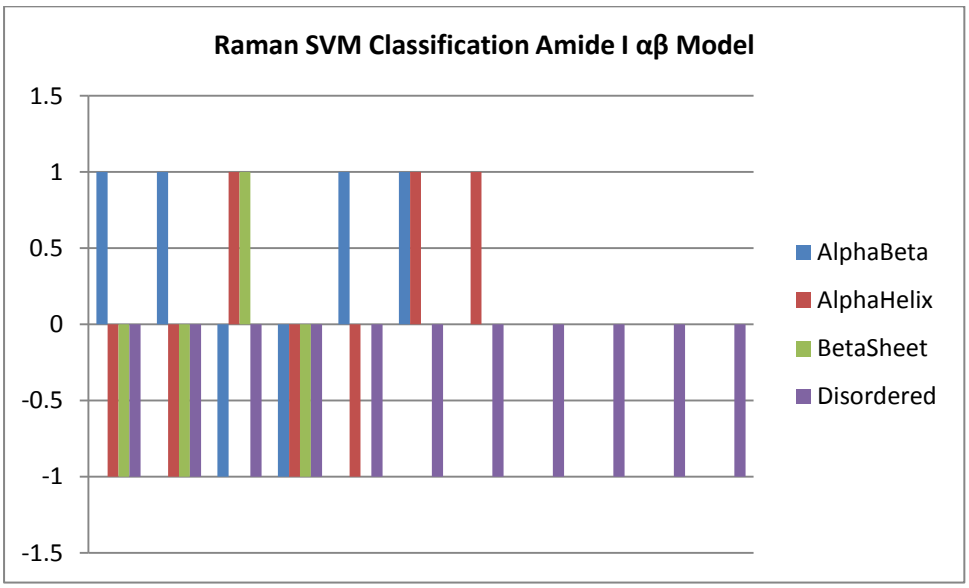

Bar Graph of Raman SVM Classification Full Spectrum α-Helix Model for Bin 100 cm $^{-1}$



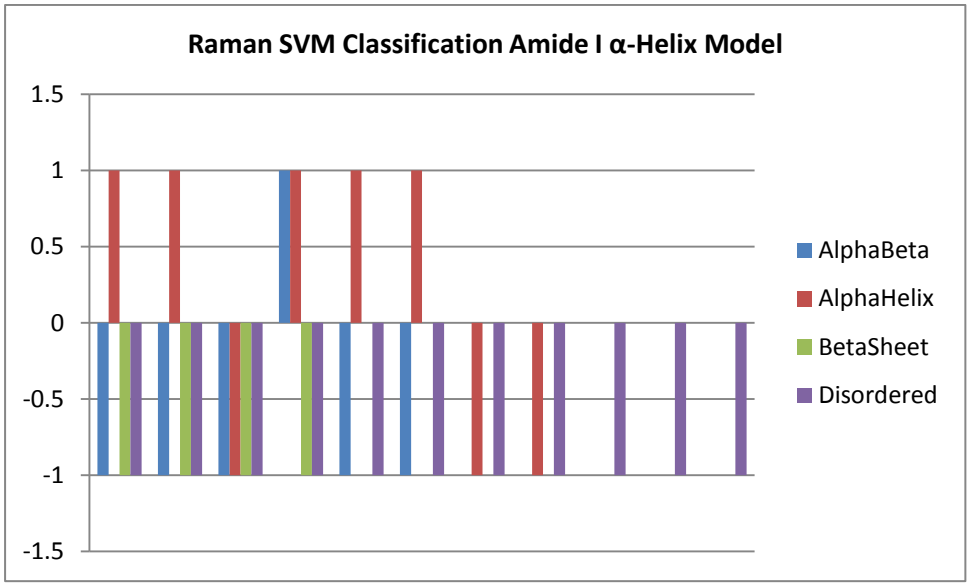
Bar Graph of Raman SVM Classification Full Spectrum Other Model for Bin 100 cm $^{-1}$



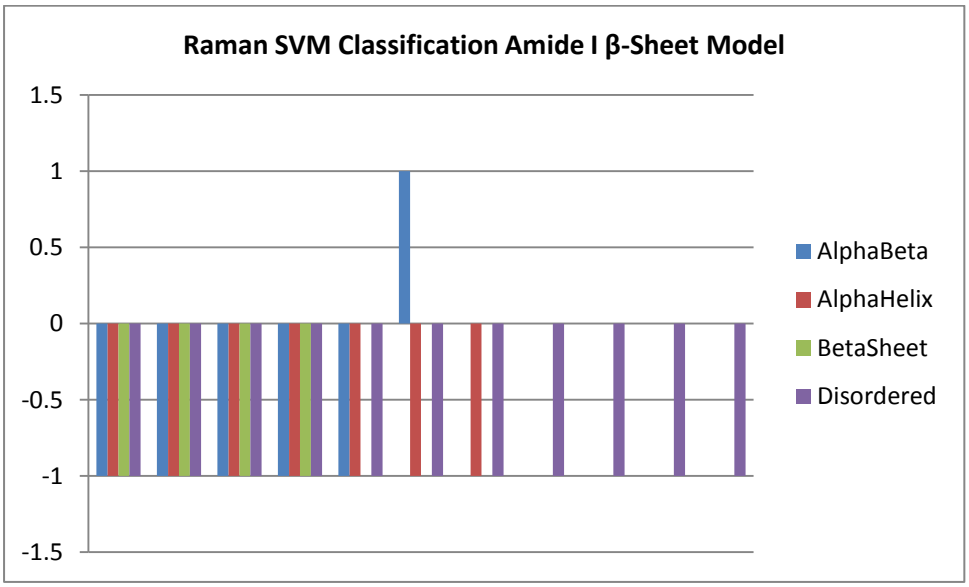

Bar Graph of Raman SVM Classification Amide I αβ Model using Bin 10 cm $^{-1}$



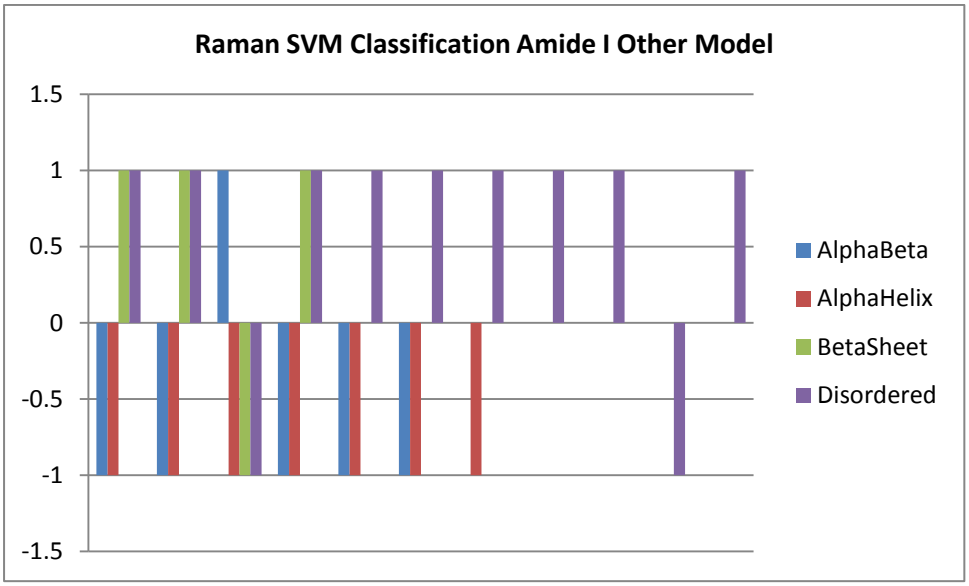
Bar Graph of Raman SVM Classification Amide I α-Helix Model using Bin 10 $cm^{-1}$



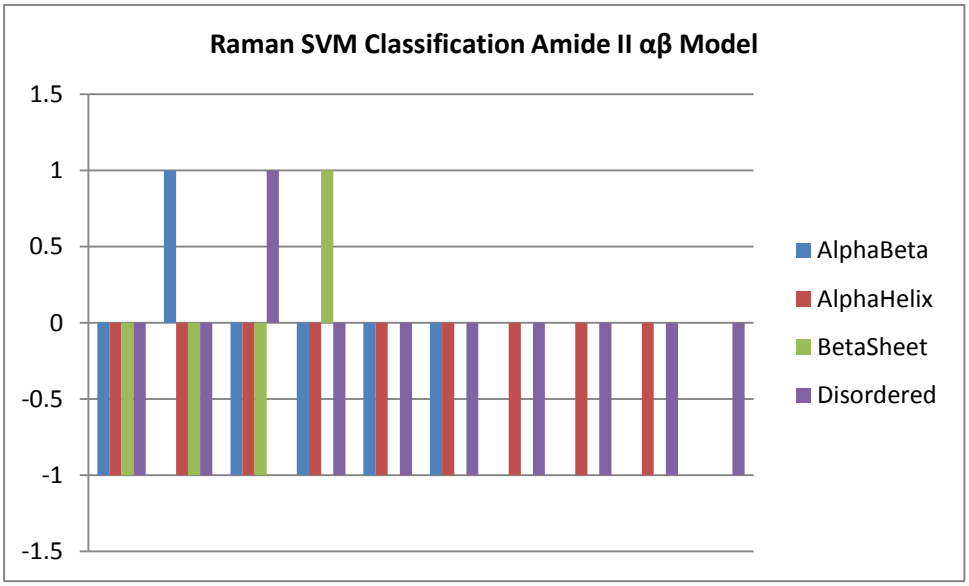

Bar Graph of Raman SVM Classification Amide I β-Sheet Model using Bin 10 $cm^{-1}$



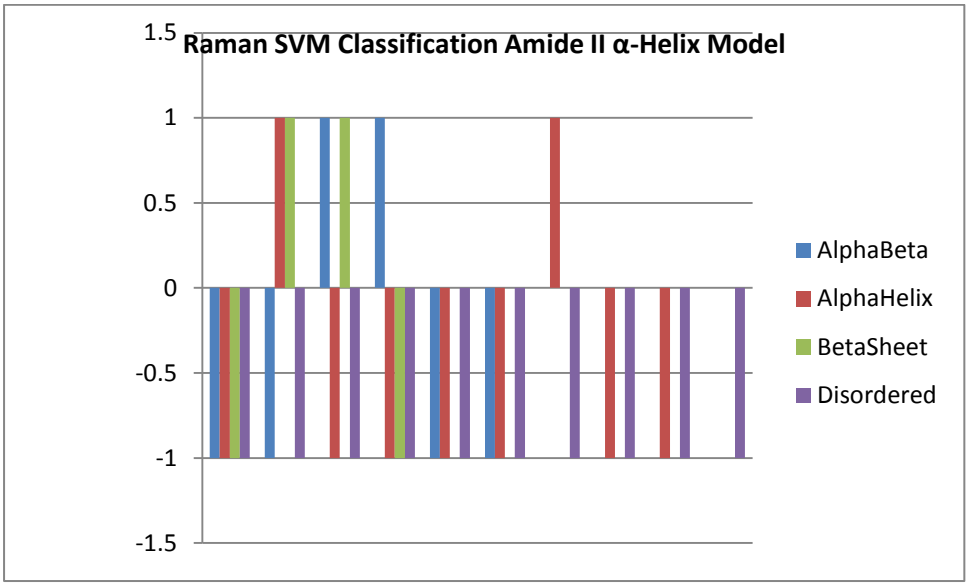
Bar Graph of Raman SVM Classification Amide I Other Model using Bin 10 $cm^{-1}$



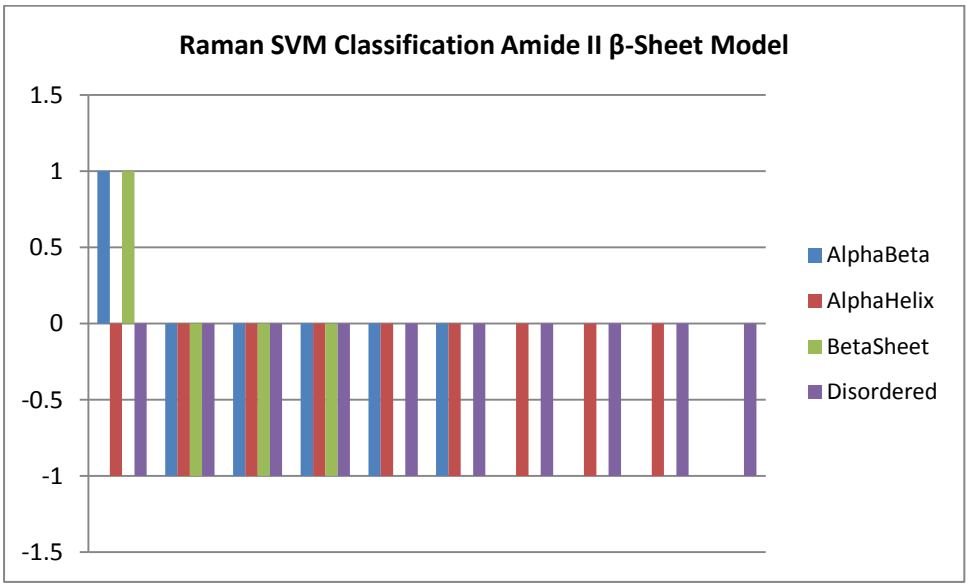

Bar Graph of Raman SVM Classification Amide II αβ Model using Bin 10 $cm^{-1}$



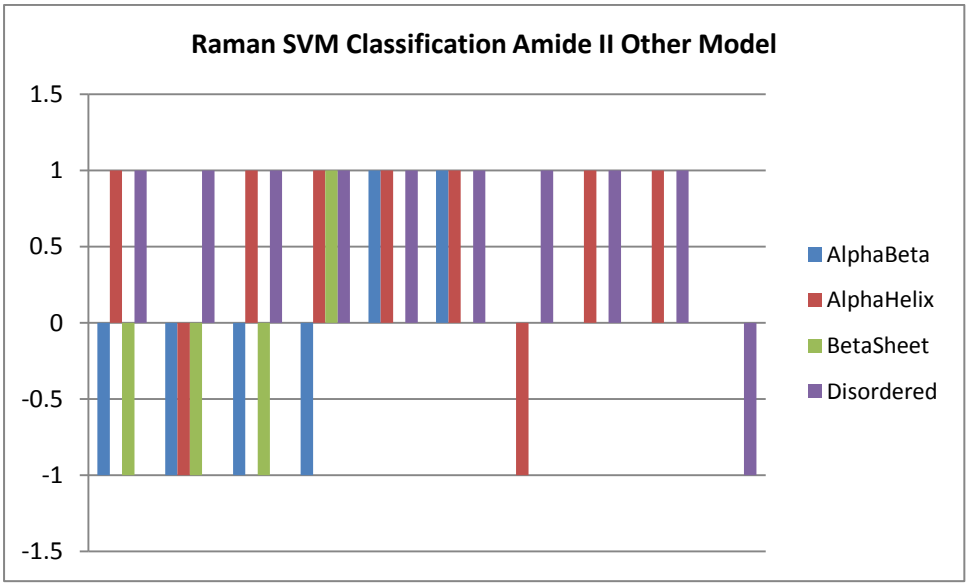
Bar Graph of Raman SVM Classification Amide II α-Helix Model using Bin 10 $cm^{-1}$



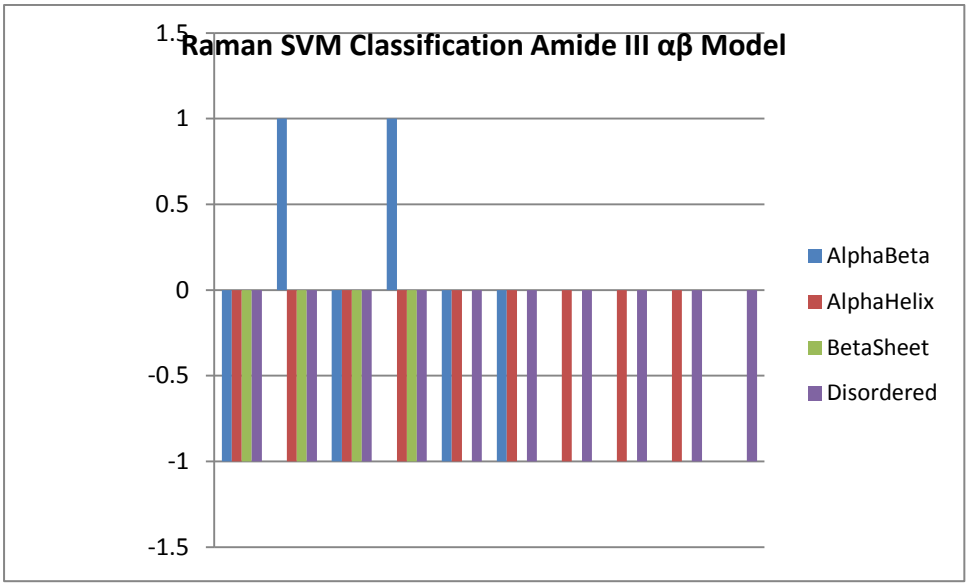

Bar Graph of Raman SVM Classification Amide II β-Sheet Model using Bin 10 $cm^{-1}$



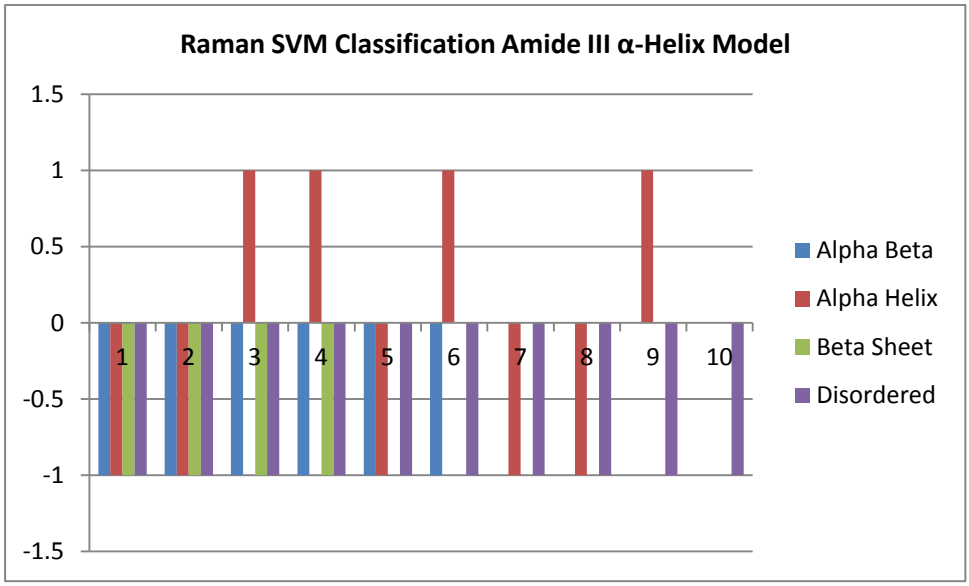
Bar Graph of Raman SVM Classification Amide II Other Model using Bin 10 $cm^{-1}$



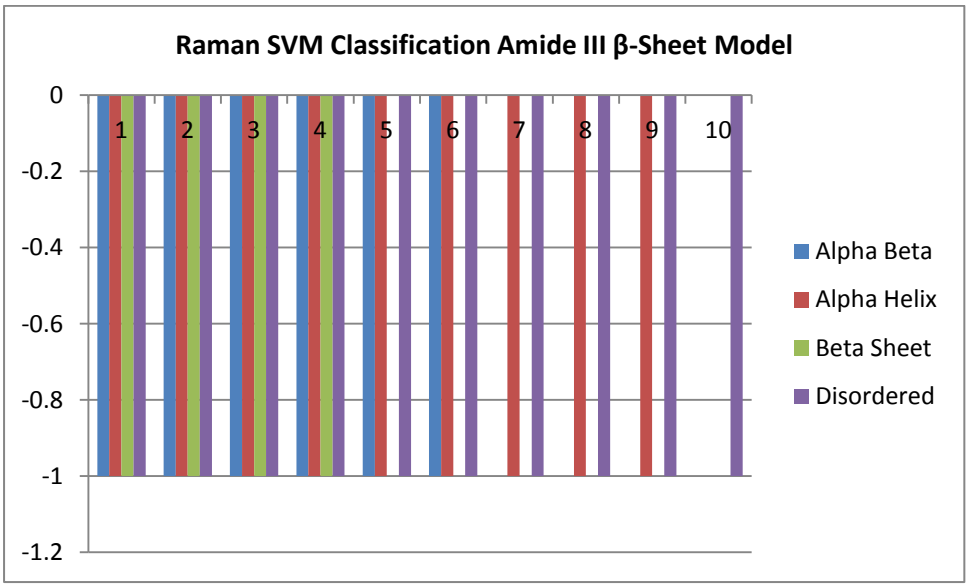

Bar Graph of Raman SVM Classification Amide III αβ Model using Bin 10 $cm^{-1}$



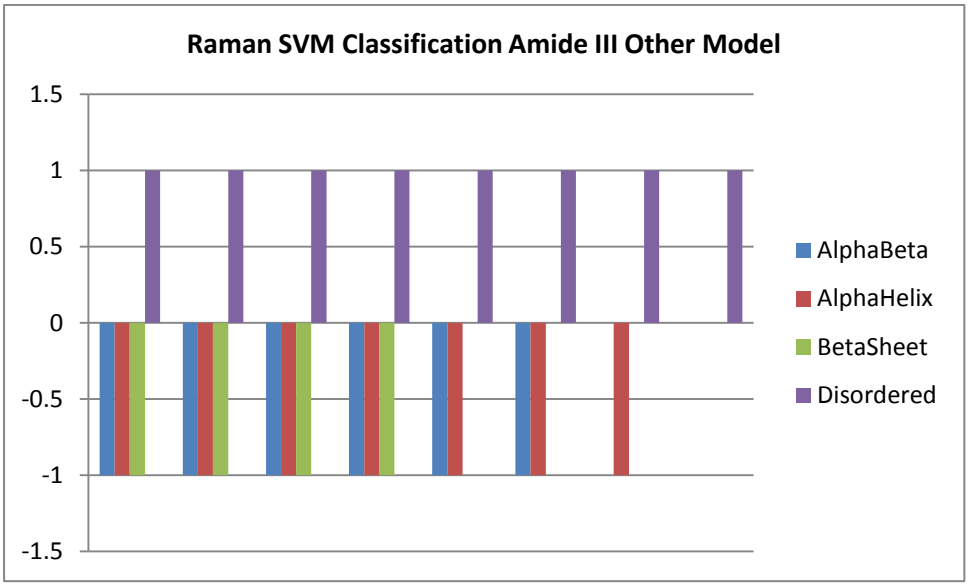
Bar Graph of Raman SVM Classification Amide III α-Helix Model using Bin 10 $cm^{-1}$



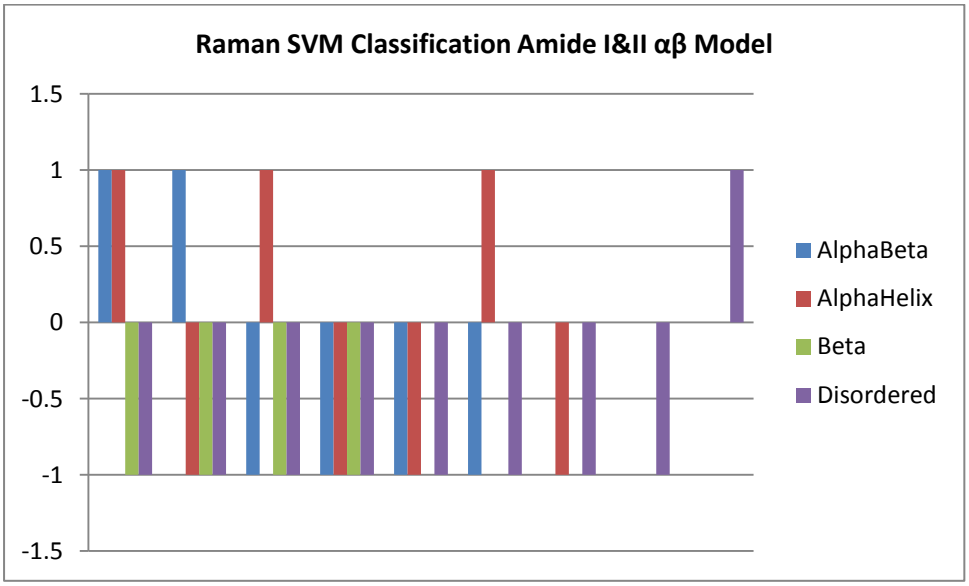

Bar Graph of Raman SVM Classification Amide III β-Sheet Model using Bin 10 $cm^{-1}$



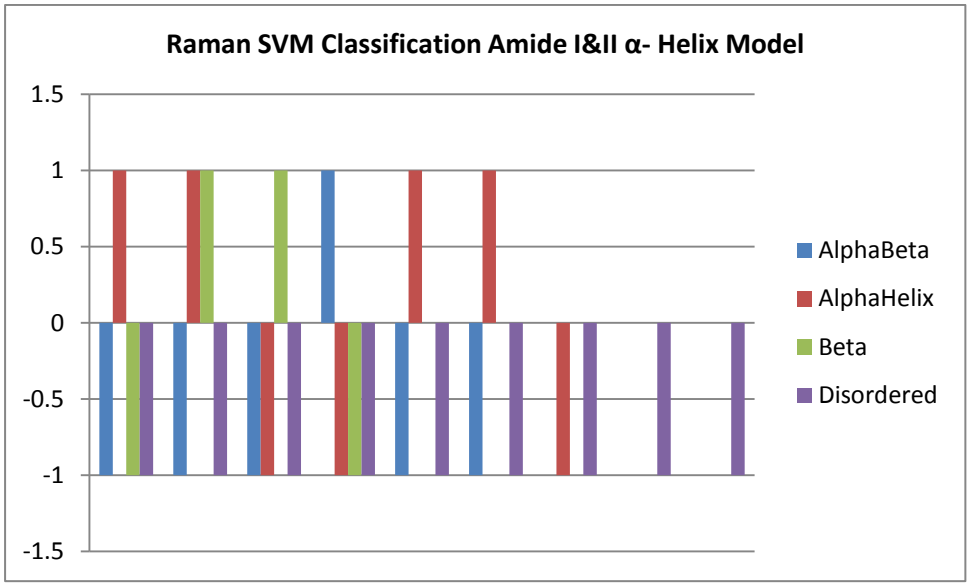
Bar Graph of Raman SVM Classification Amide III Other Model using Bin 10 cm $^{-1}$



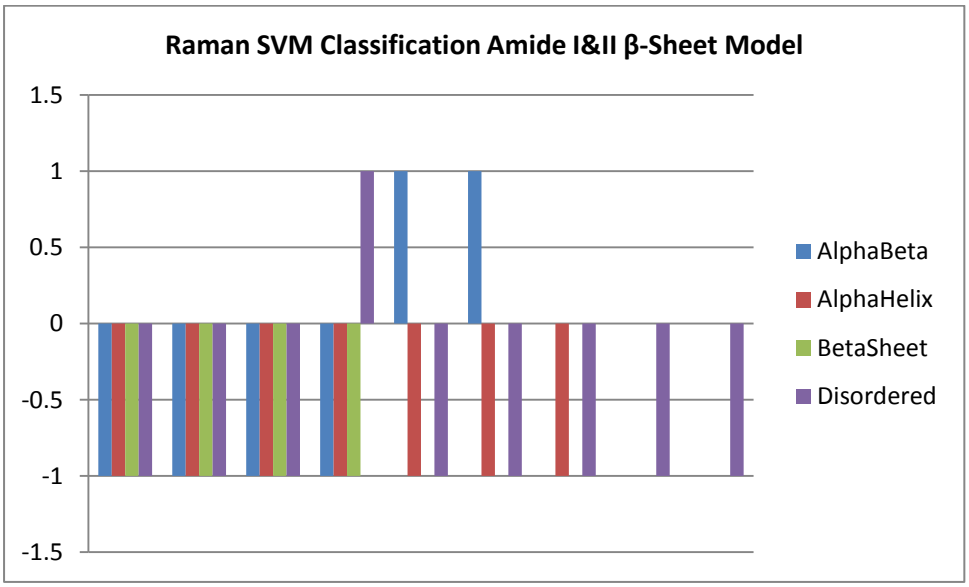

Bar Graph of Raman SVM Classification Amide I&II αβ Model using Bin 10 cm $^{-1}$



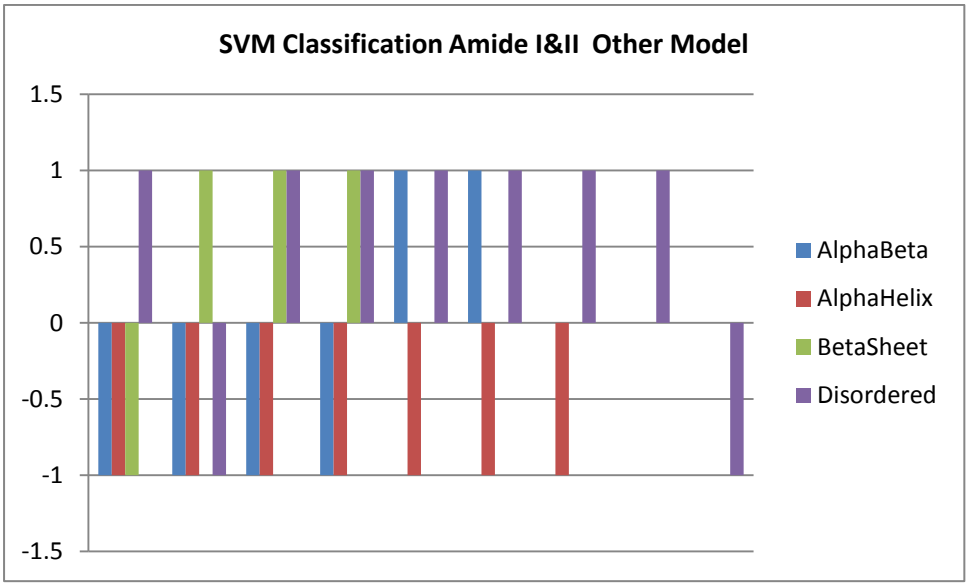
Bar Graph of Raman SVM Classification Amide I&II α- Helix Model using Bin 10 cm $^{-1}$



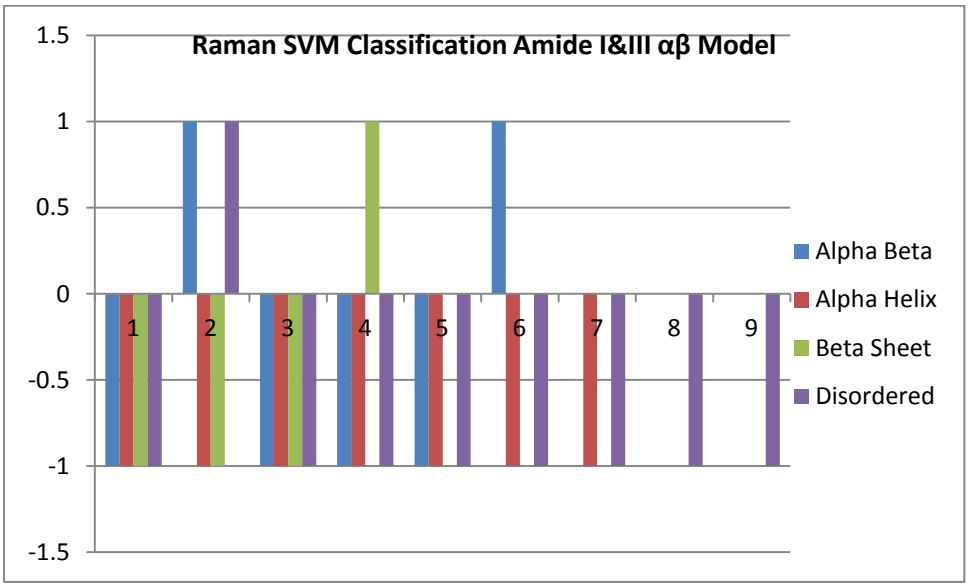

Bar Graph of Raman SVM Classification Amide I&II β-Sheet Model using Bin 10 cm $^{-1}$



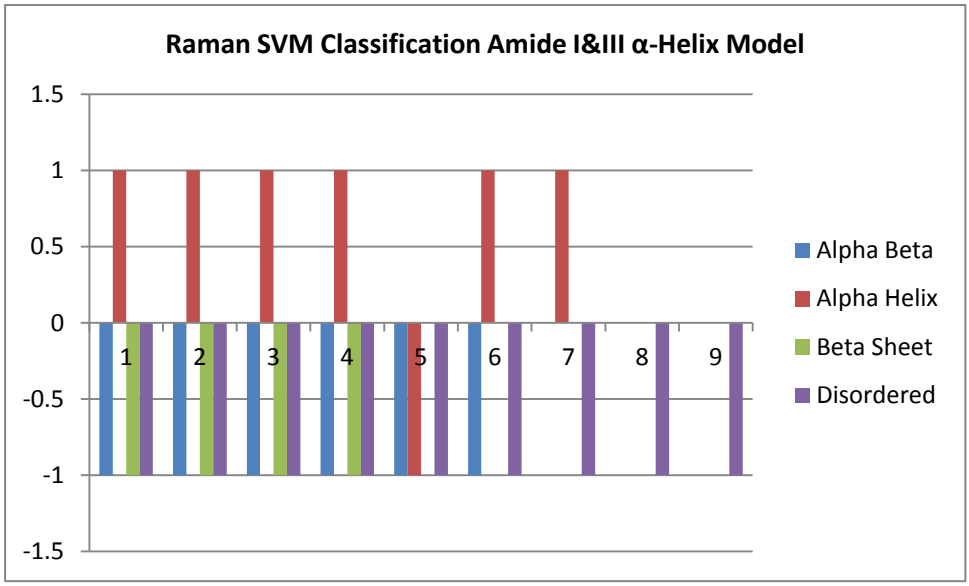
Bar Graph of Raman SVM Classification Amide I&II Other Model using Bin 10 $cm^{-1}$



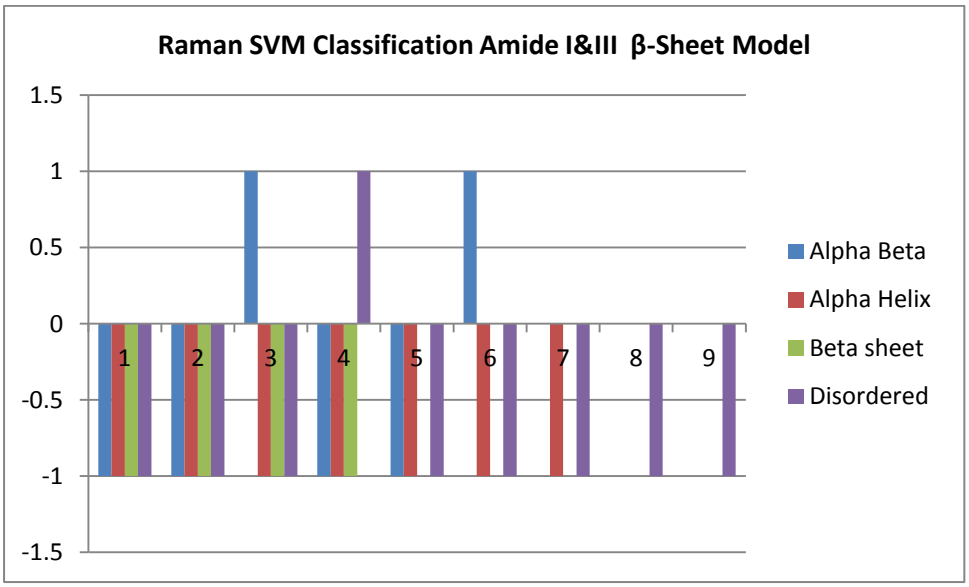

Bar Graph of Raman SVM Classification Amide I&III αβ Model using Bin 10 $cm^{-1}$



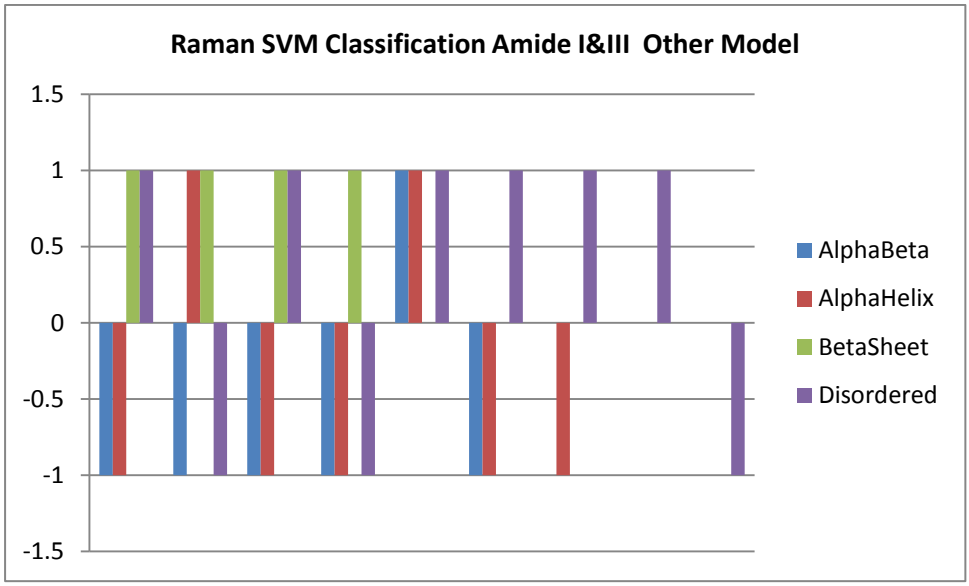
Bar Graph of Raman SVM Classification Amide I&III α-Helix Model using Bin 10 $cm^{-1}$



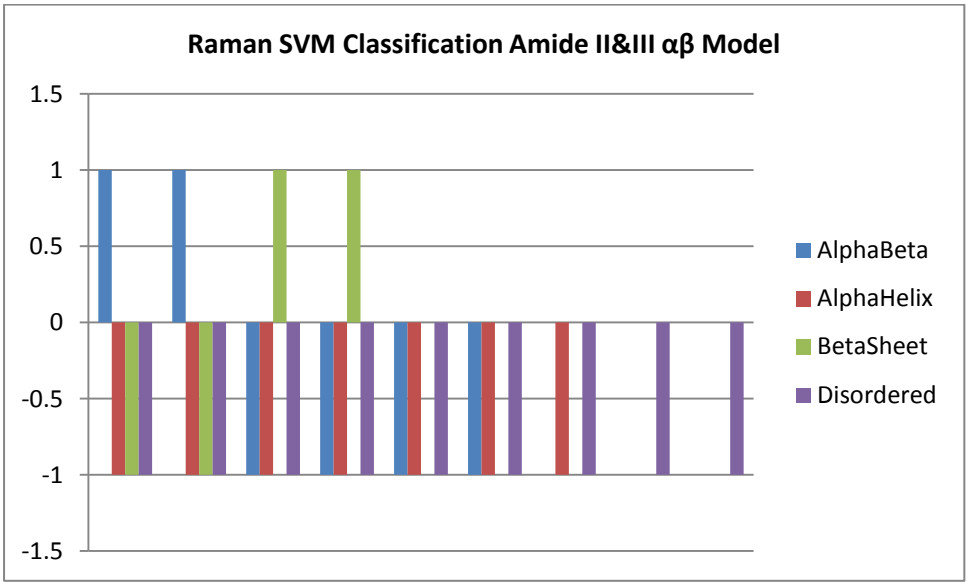

Bar Graph of Raman SVM Classification Amide I&III β-Sheet Model using Bin 10 $cm^{-1}$



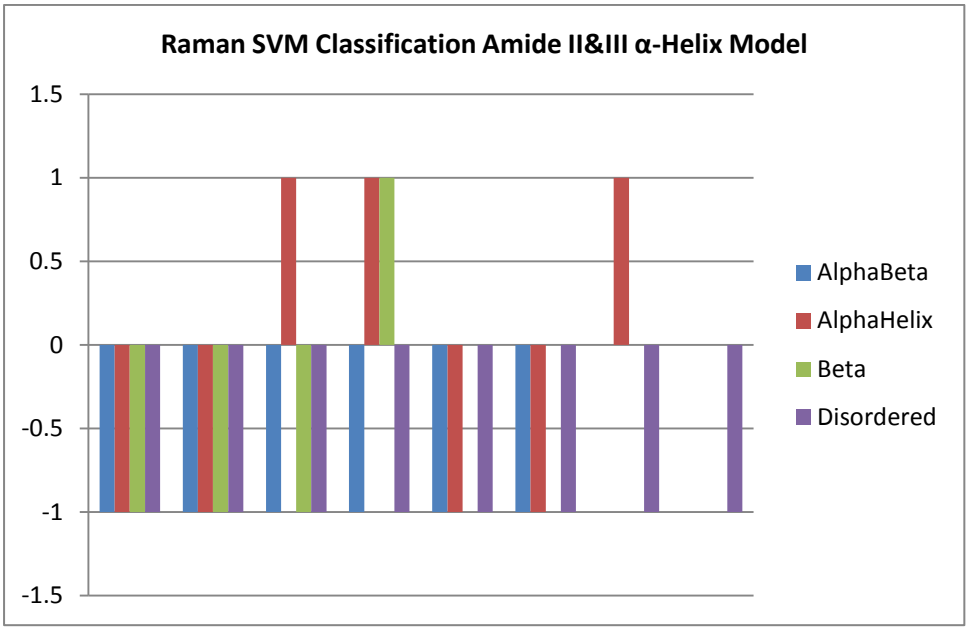
Bar Graph of Raman SVM Classification Amide I&III Other Model using Bin 10 cm $^{-1}$



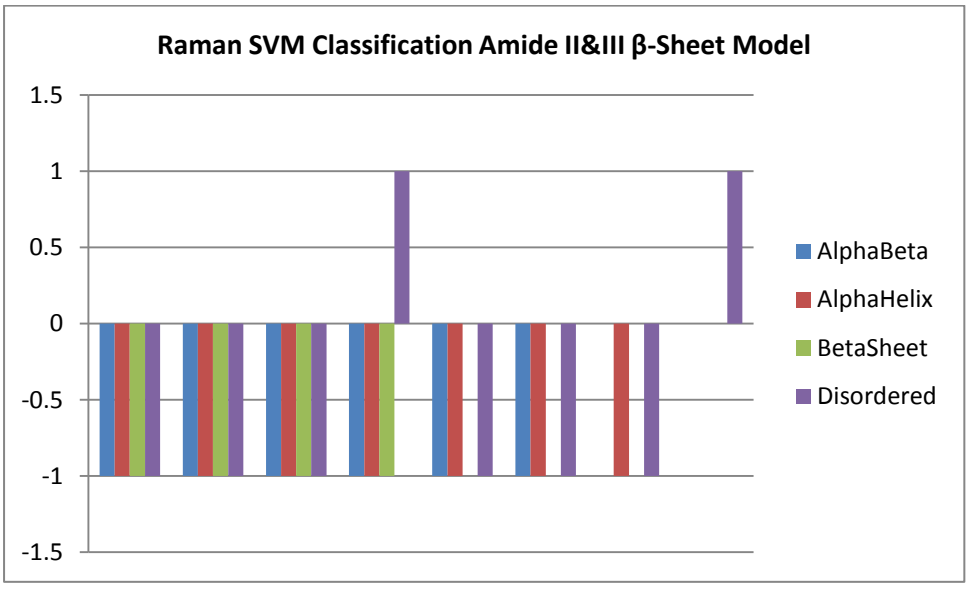

Bar Graph of Raman SVM Classification Amide II&III αβ Model using Bin 10 cm $^{-1}$



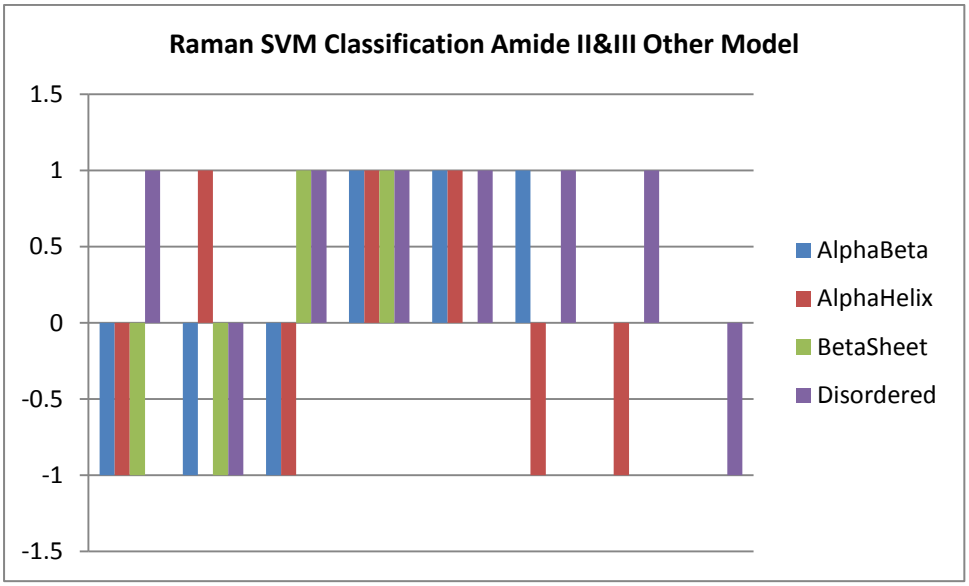
Bar Graph of Raman SVM Classification Amide II&III α-Helix Model using Bin 10 $cm^{-1}$



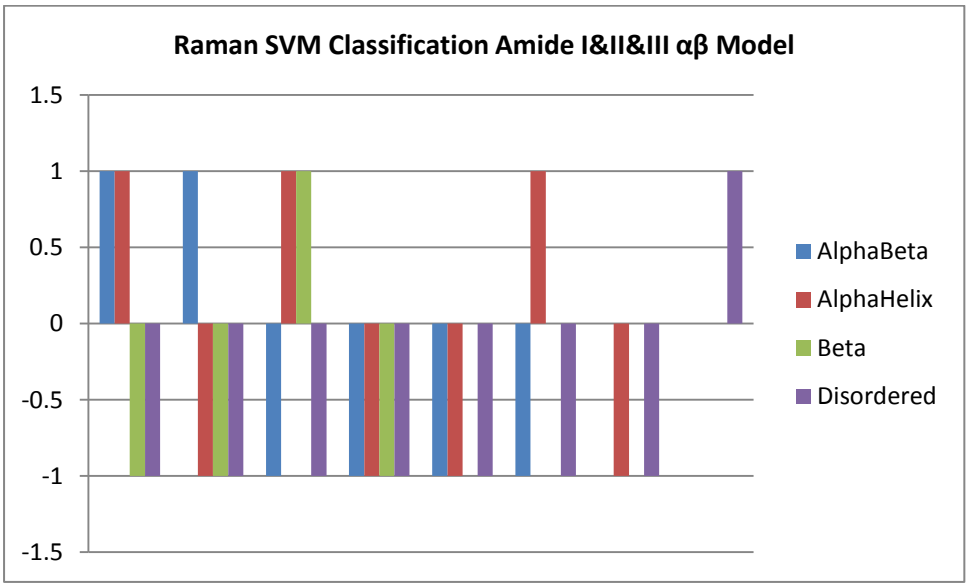

Bar Graph of Raman SVM Classification Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



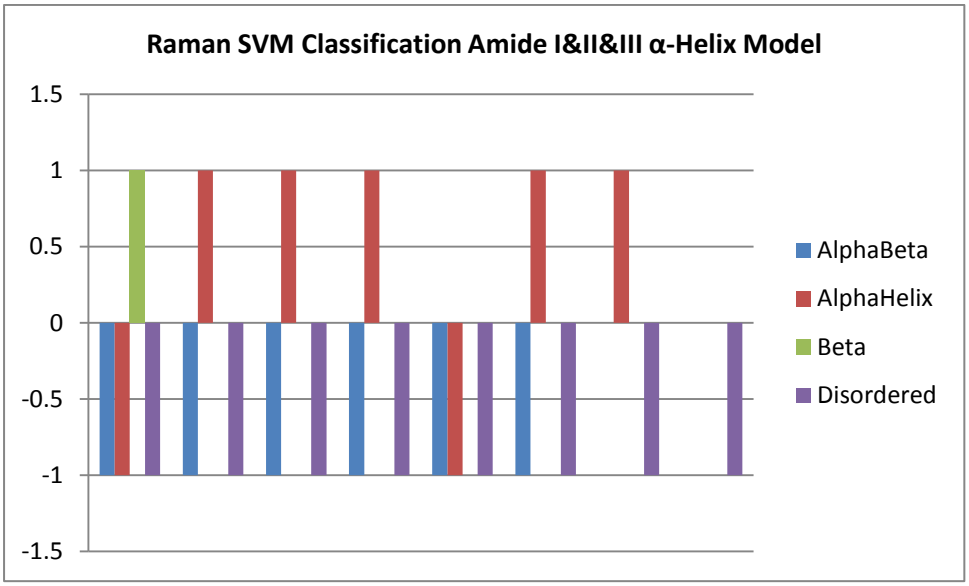
Bar Graph of Raman SVM Classification Amide II&III Other Model using Bin 10 cm $^{-1}$



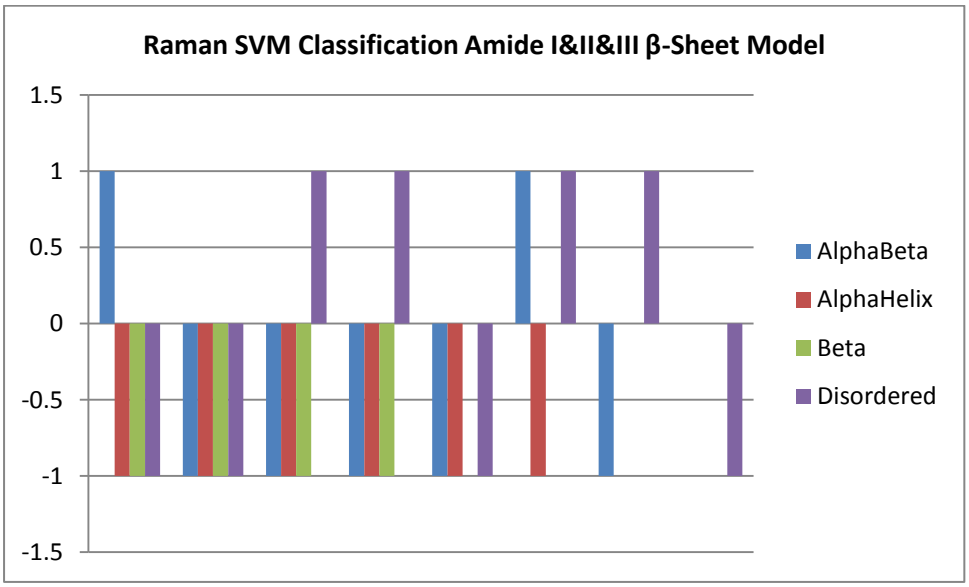

Bar Graph of Raman SVM Classification Amide I&II&III αβ Model using Bin 10 cm $^{-1}$



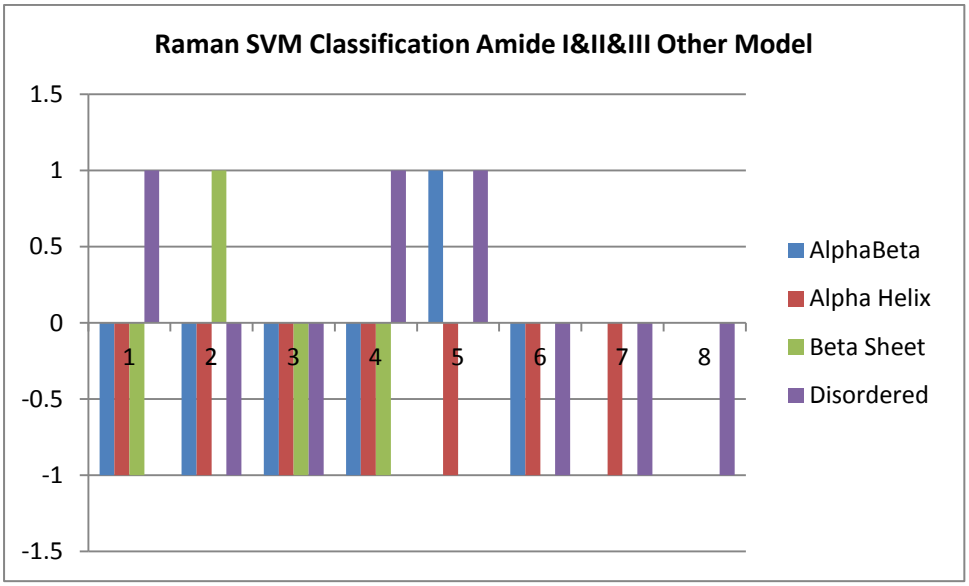
Bar Graph of Raman SVM Classification Amide I&II&III α-Helix Model using Bin 10 $cm^{-1}$



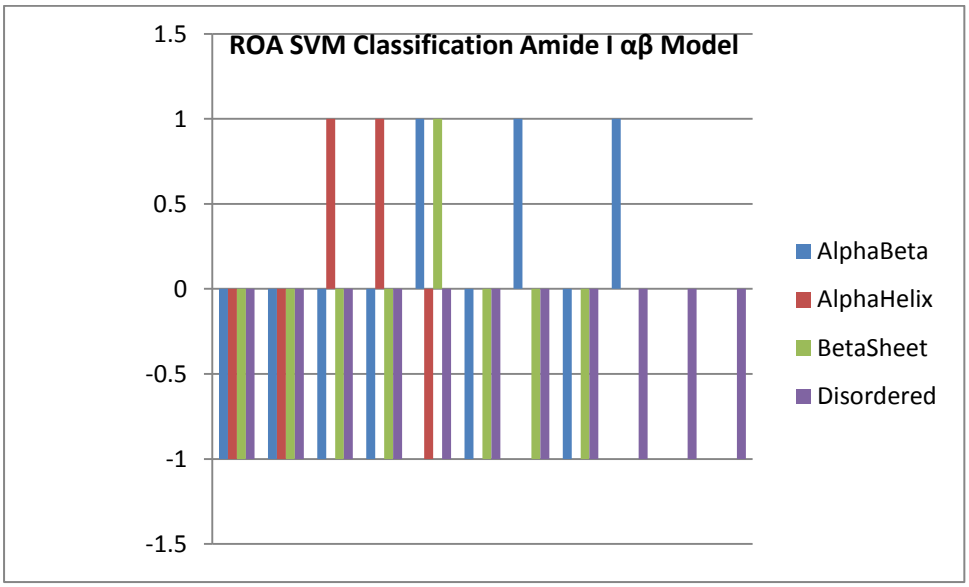

Bar Graph of Raman SVM Classification Amide I&II&III β-Sheet Model using Bin 10 $cm^{-1}$



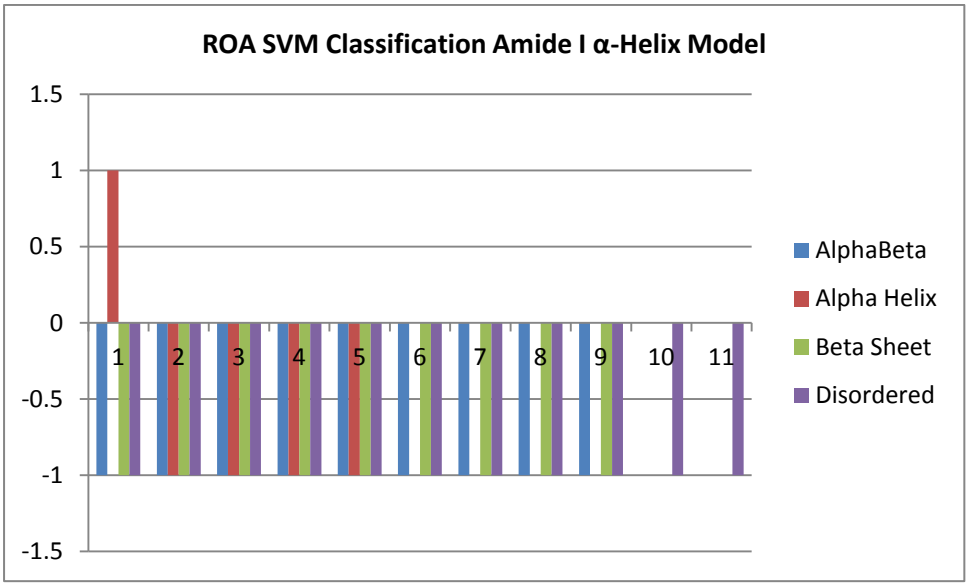
Bar Graph of Raman SVM Classification Amide I&II&III Other Model using Bin 10 $cm^{-1}$



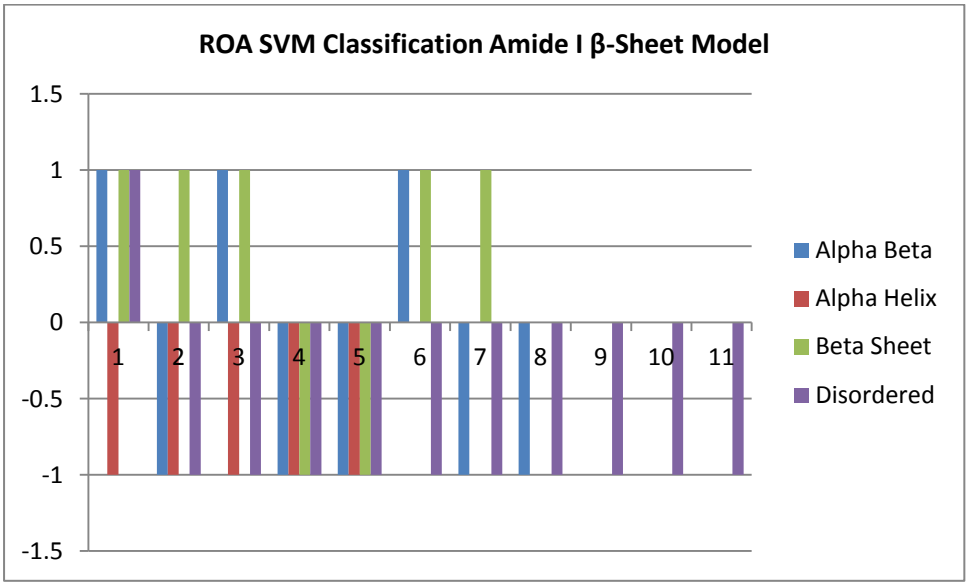

Bar Graph of ROA SVM Classification Amide I αβ Model using Bin 10 $cm^{-1}$



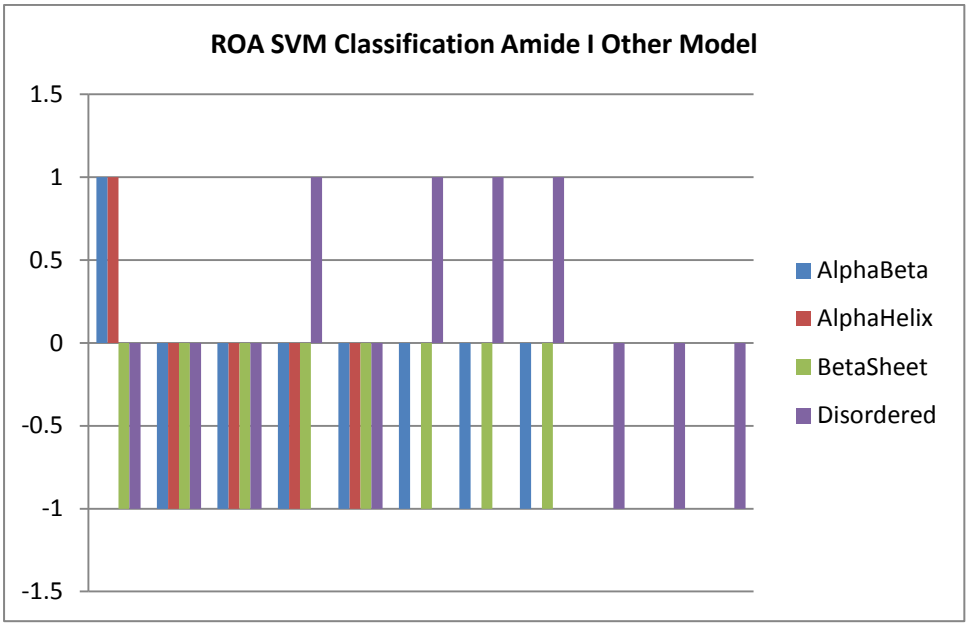
Bar Graph of ROA SVM Classification Amide I α-Helix Model using Bin 10 $cm^{-1}$



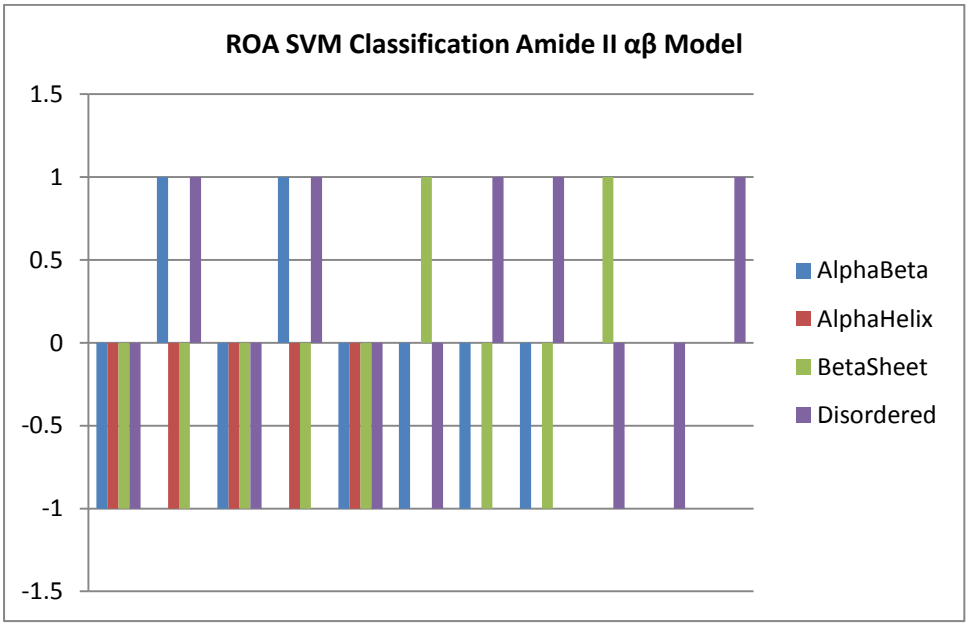

Bar Graph of ROA SVM Classification Amide I β-Sheet Model using Bin 10 $cm^{-1}$



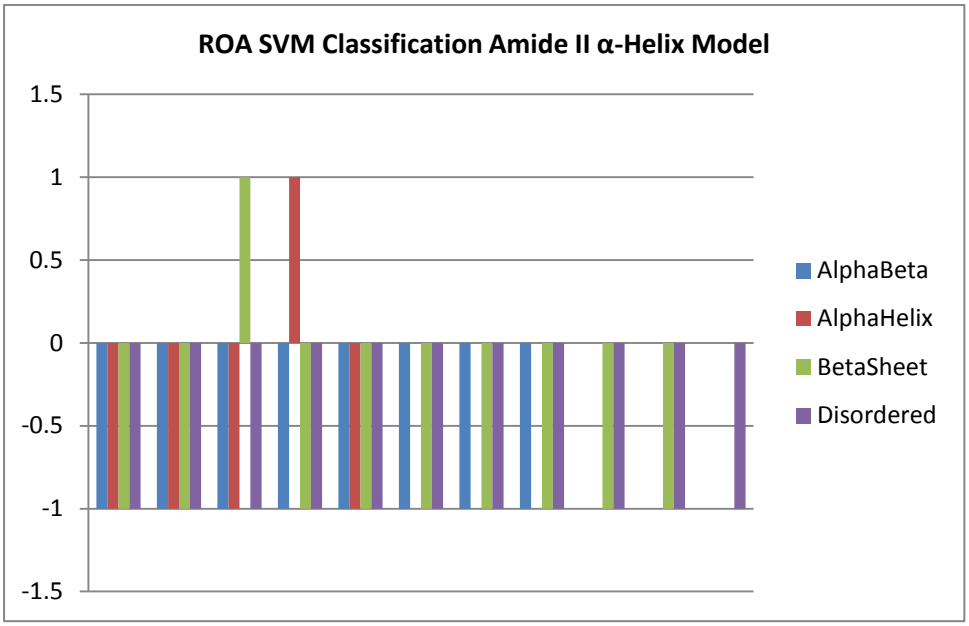
Bar Graph of ROA SVM Classification Amide I Other Model using Bin 10 $cm^{-1}$



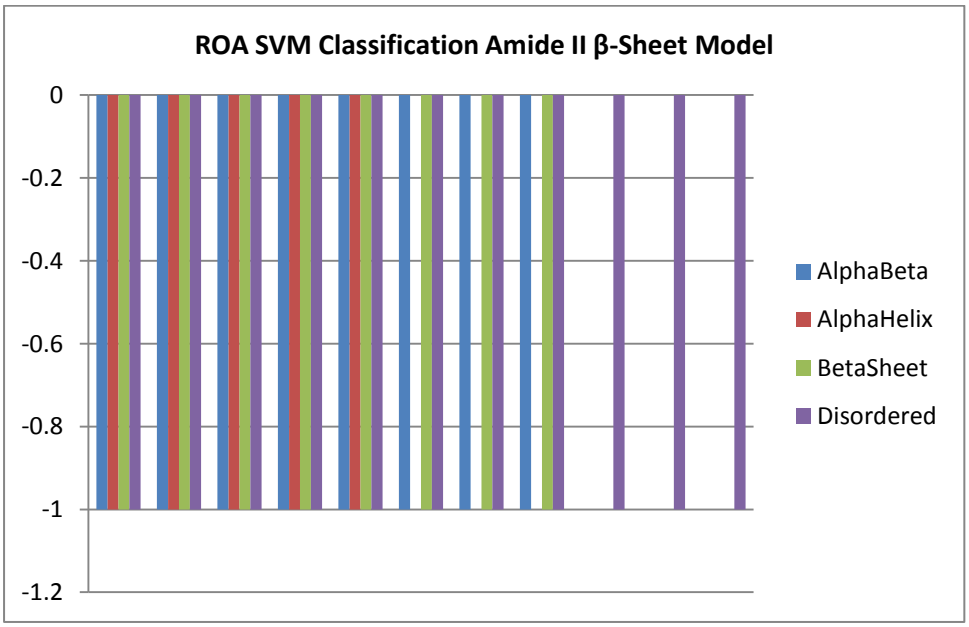

Bar Graph of ROA SVM Classification Amide II αβ Model using Bin 10 $cm^{-1}$



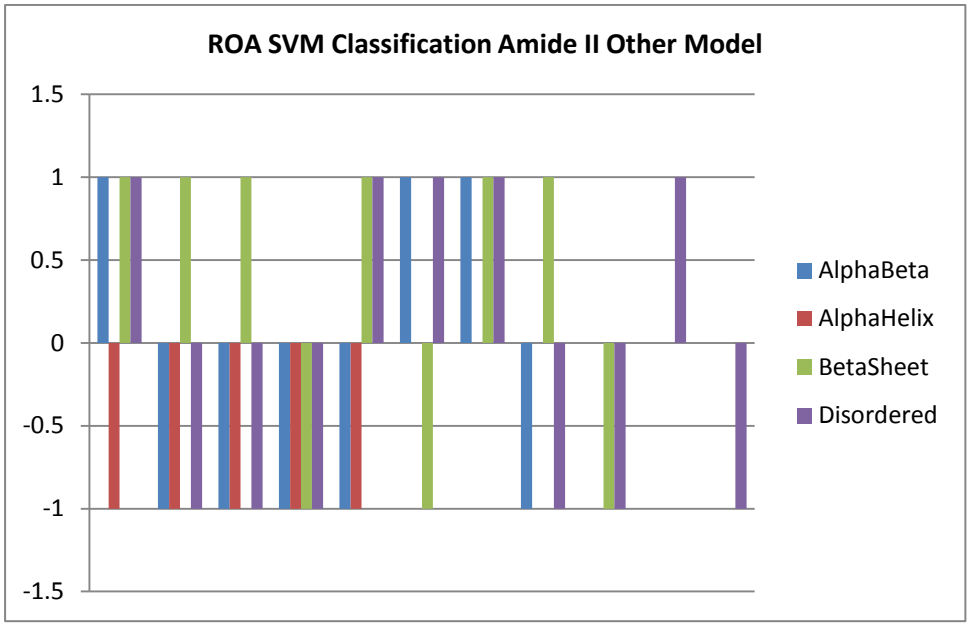
Bar Graph of ROA SVM Classification Amide II α-Helix Model using Bin 10 $cm^{-1}$



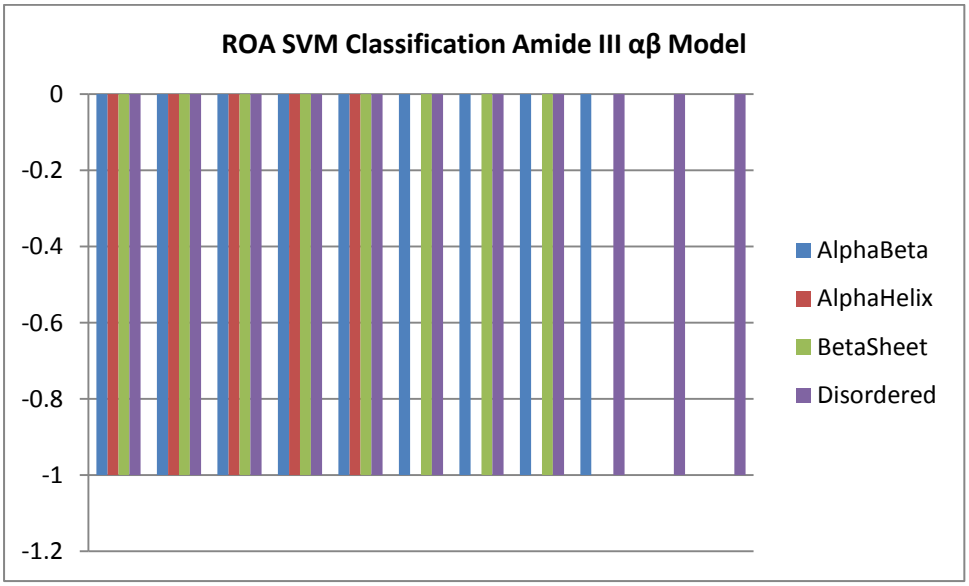

Bar Graph of ROA SVM Classification Amide II β-Sheet Model using Bin 10 $cm^{-1}$



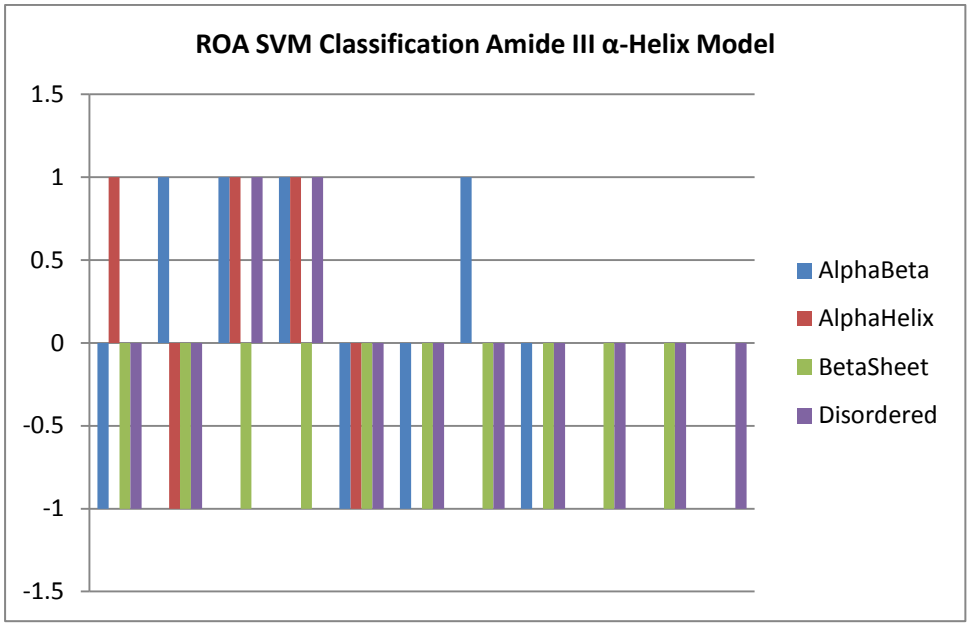
Bar Graph of ROA SVM Classification Amide II Other Model using Bin 10 $cm^{-1}$



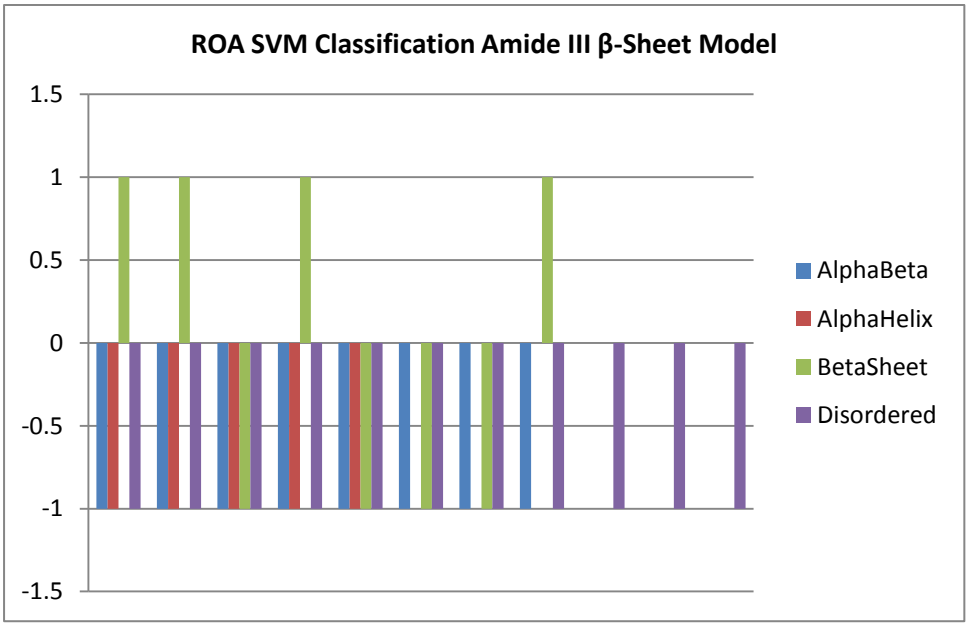

Bar Graph of ROA SVM Classification Amide III αβ Model using Bin 10 $cm^{-1}$



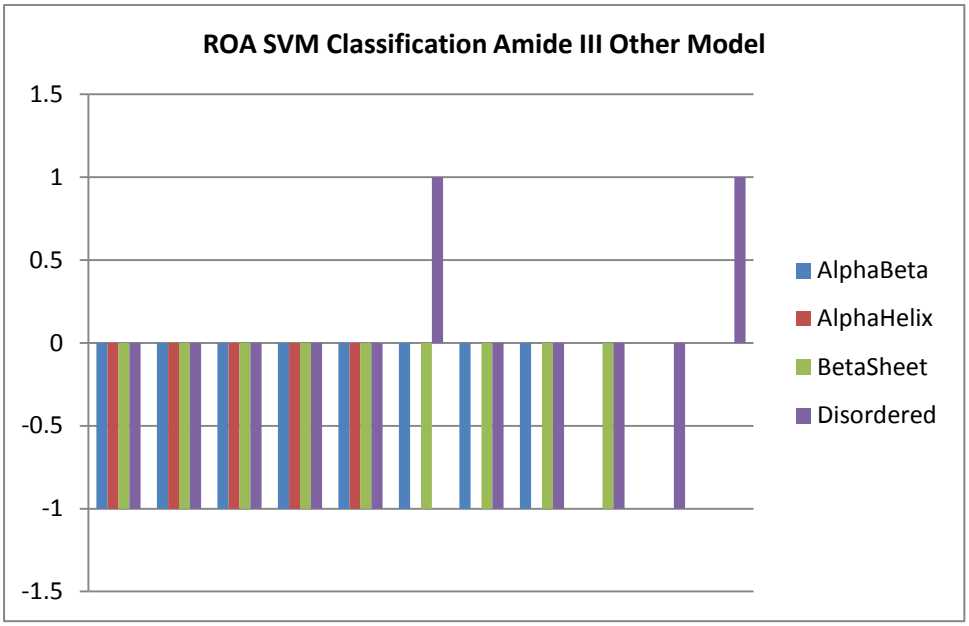
Bar Graph of ROA SVM Classification Amide III α-Helix Model using Bin 10 $cm^{-1}$



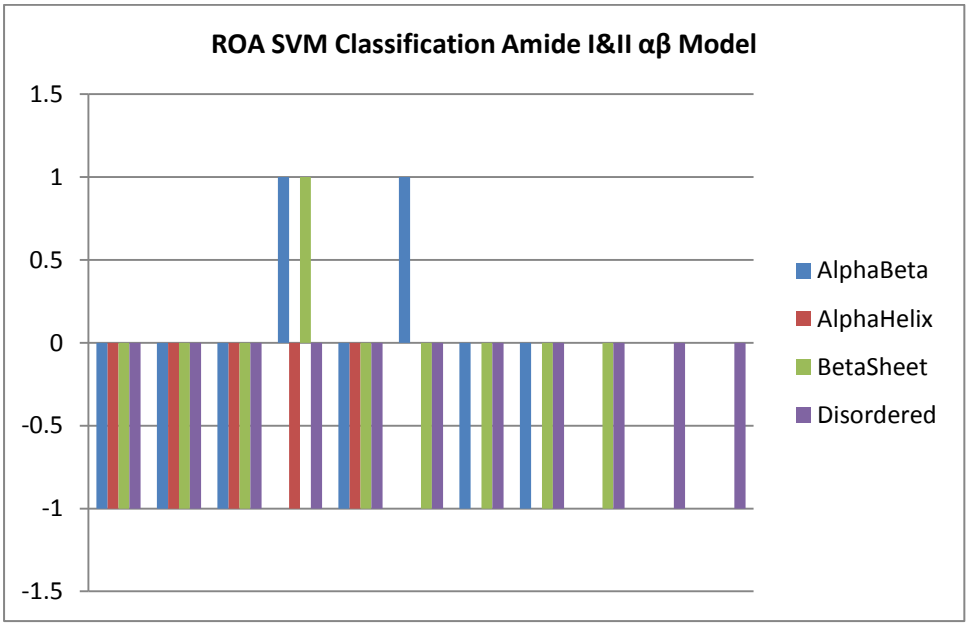

Bar Graph of ROA SVM Classification Amide III β-Sheet Model using Bin 10 $cm^{-1}$



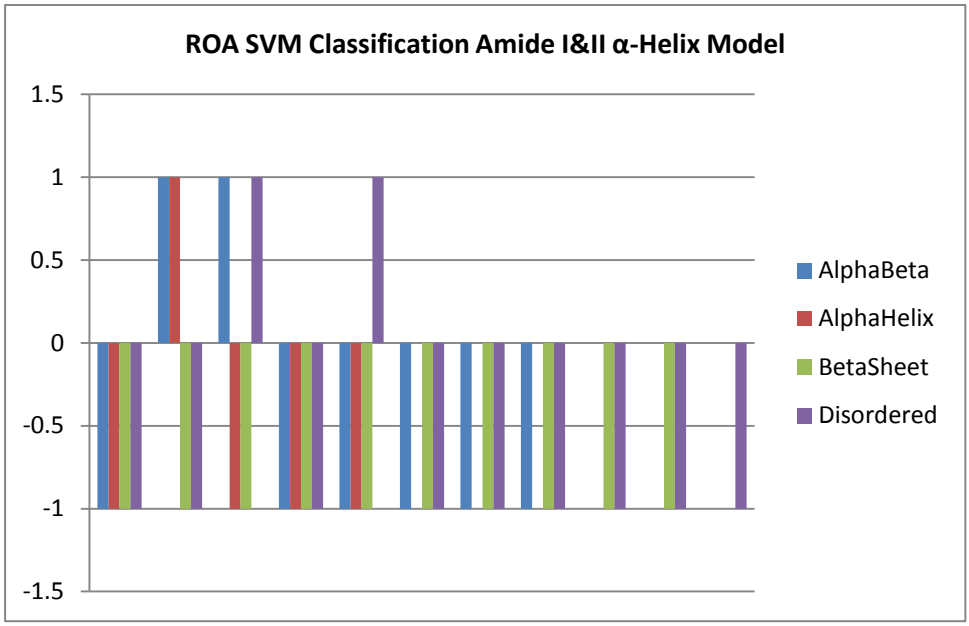
Bar Graph of ROA SVM Classification Amide III Other Model using Bin 10 $cm^{-1}$



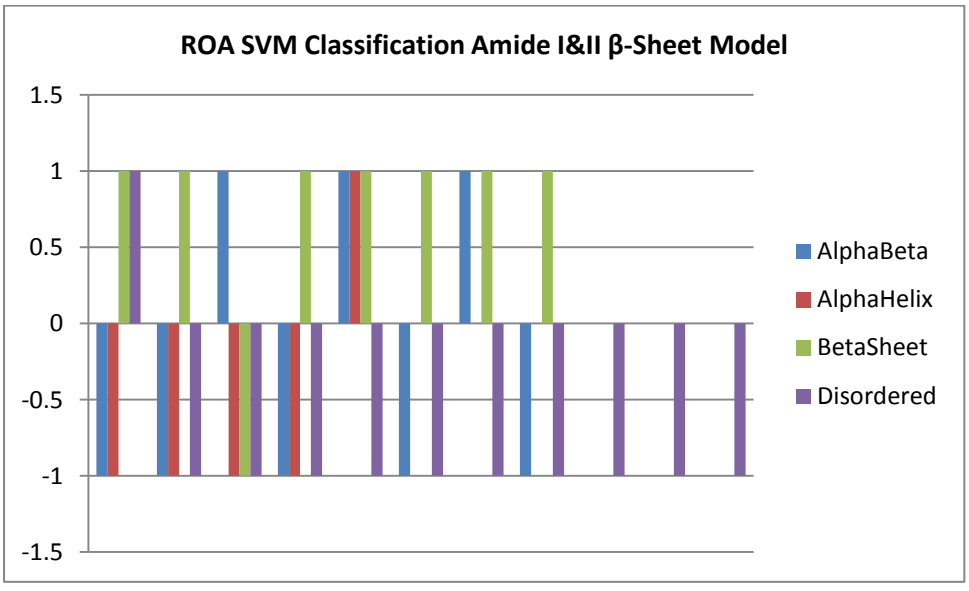

Bar Graph of ROA SVM Classification Amide I&II αβ Model using Bin 10 $cm^{-1}$



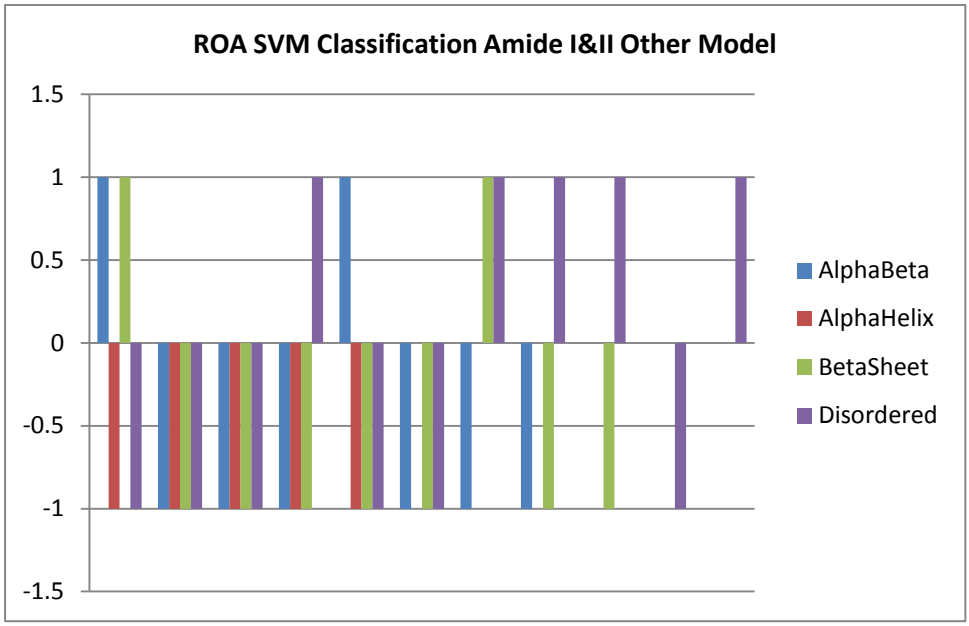
Bar Graph of ROA SVM Classification Amide I&II α-Helix Model using Bin 10 $cm^{-1}$



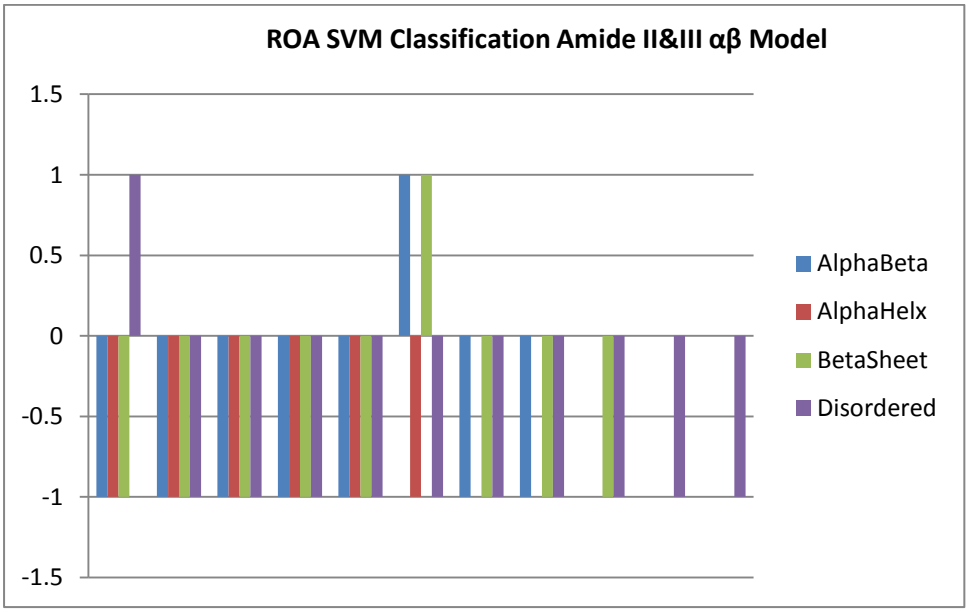

Bar Graph of ROA SVM Classification Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



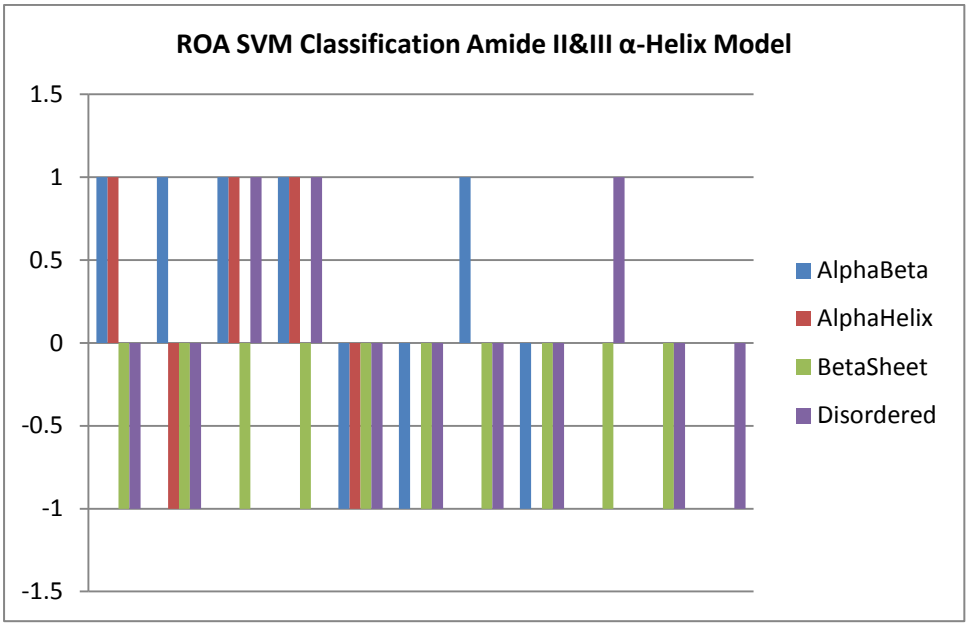
Bar Graph of ROA SVM Classification Amide I&II Other Model using Bin 10 $cm^{-1}$



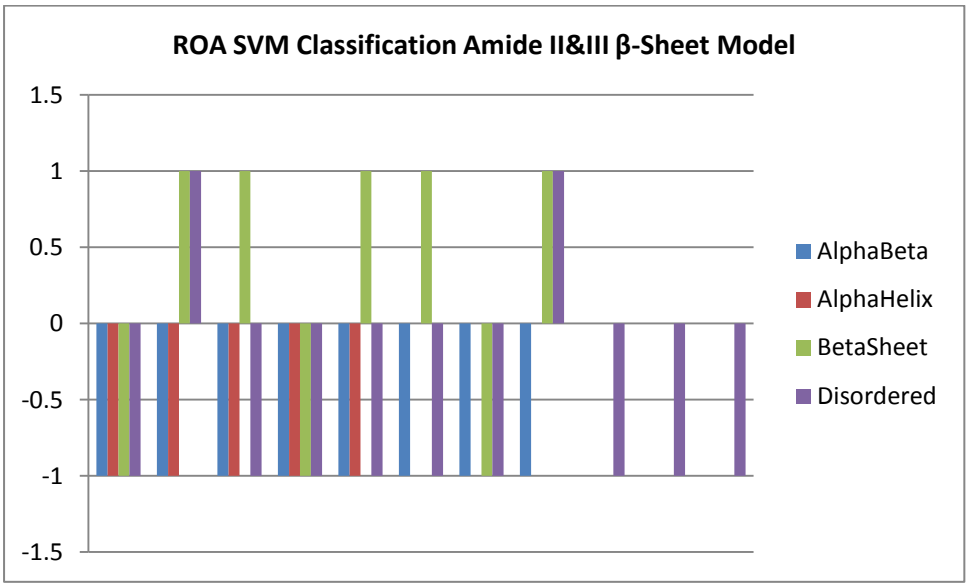

Bar Graph of ROA SVM Classification Amide II&III αβ Model using Bin 10 $cm^{-1}$



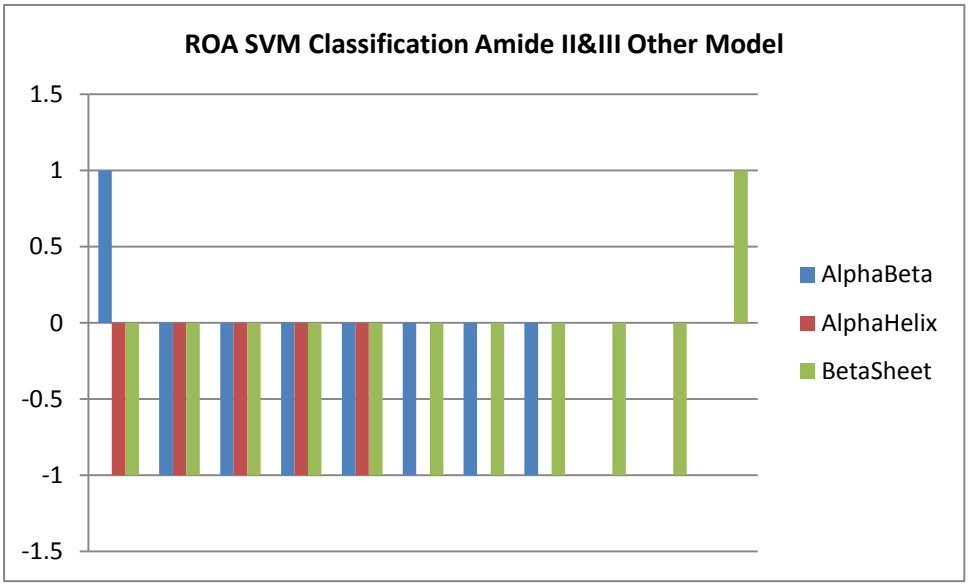
Bar Graph of ROA SVM Classification Amide II&III α-Helix Model using Bin 10 $cm^{-1}$



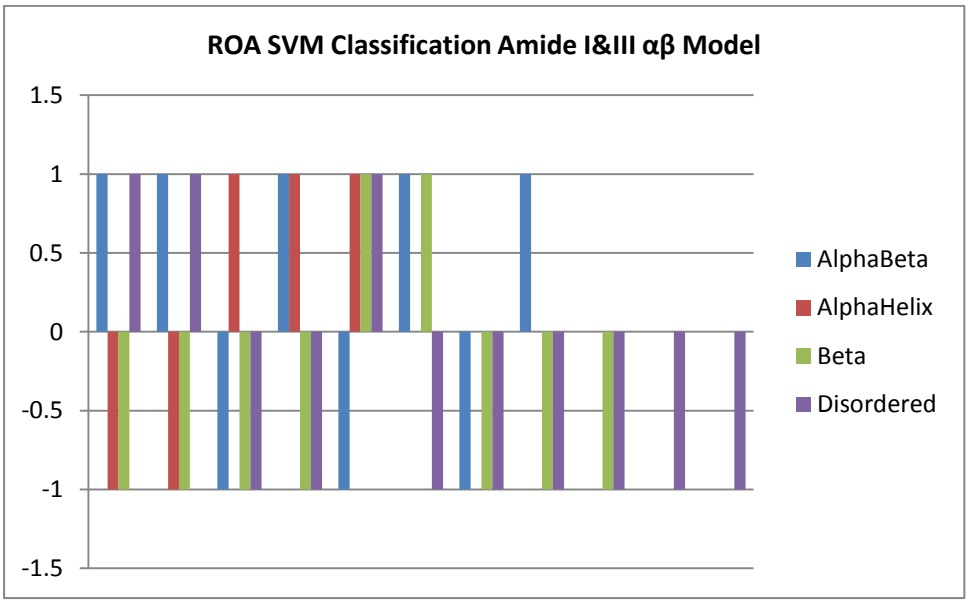

Bar Graph of ROA SVM Classification Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



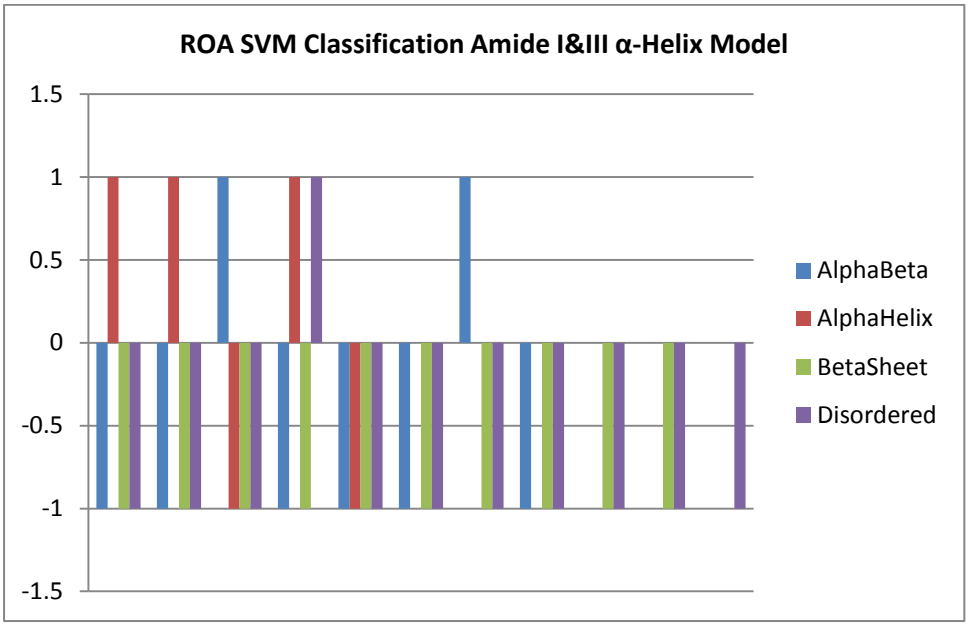
Bar Graph of ROA SVM Classification Amide II&III Other Model using Bin 10 $cm^{-1}$



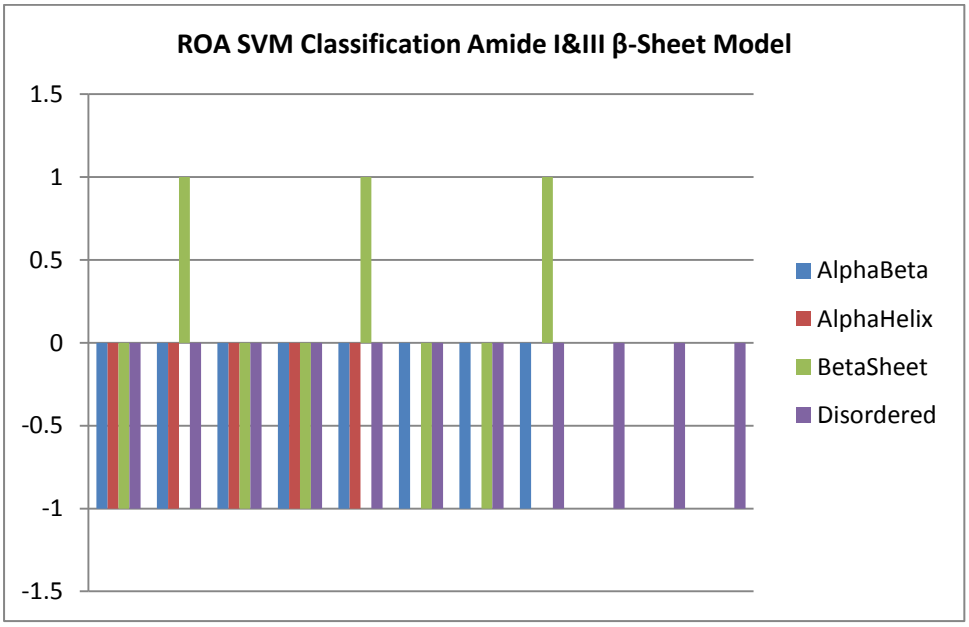

Bar Graph of ROA SVM Classification Amide I&III αβ Model using Bin 10 $cm^{-1}$



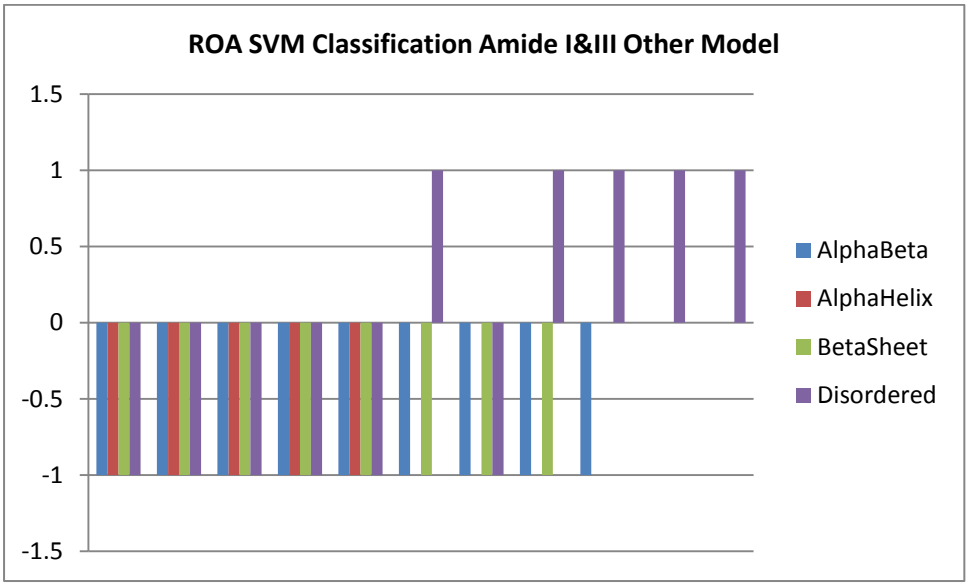
Bar Graph of ROA SVM Classification Amide I&III α-Helix Model using Bin 10 $cm^{-1}$



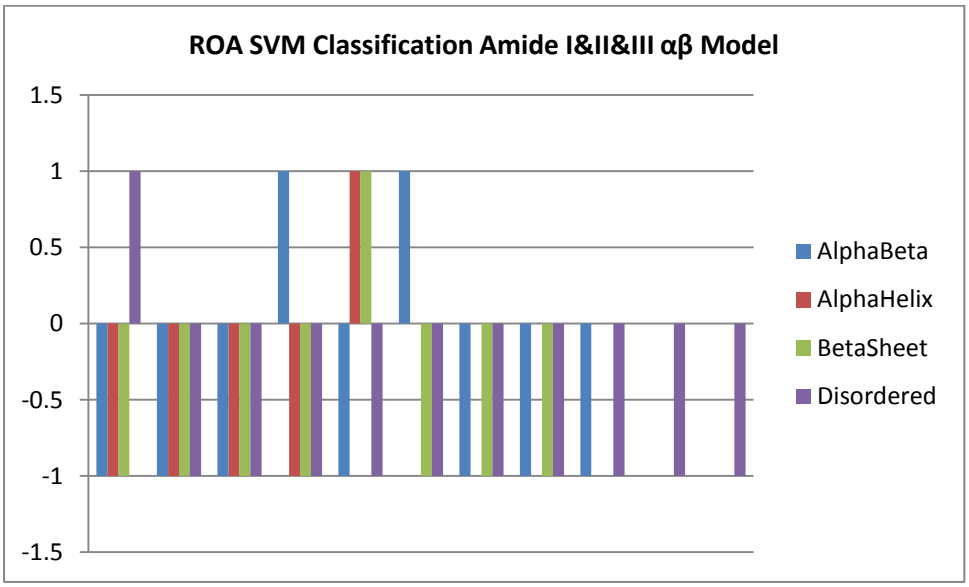

Bar Graph of ROA SVM Classification Amide I&III β-Sheet Model using Bin 10 $cm^{-1}$



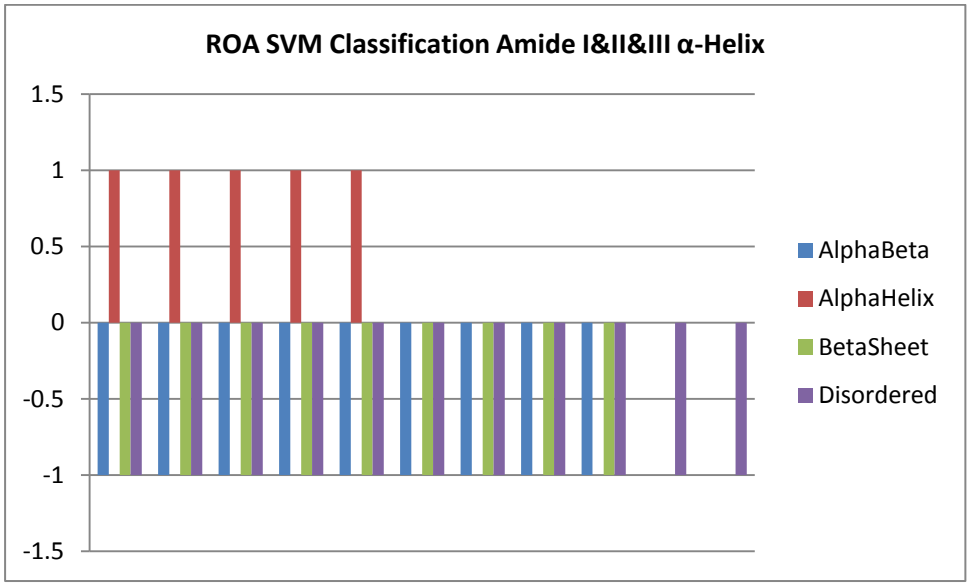
Bar Graph of ROA SVM Classification Amide I&III Other Model using Bin 10 $cm^{-1}$



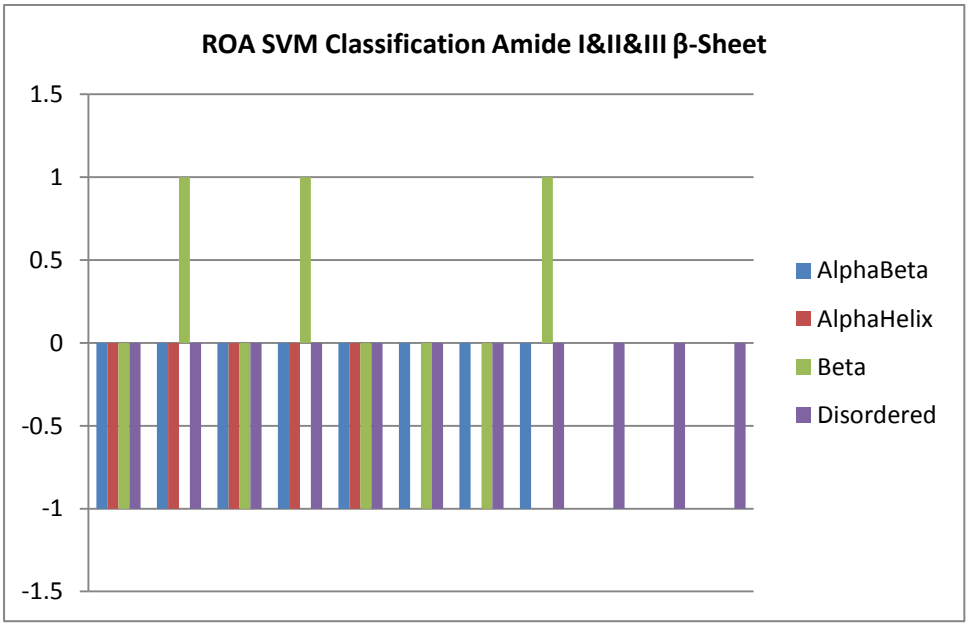

Bar Graph of ROA SVM Classification Amide I&II&III αβ Model using Bin 10 $cm^{-1}$



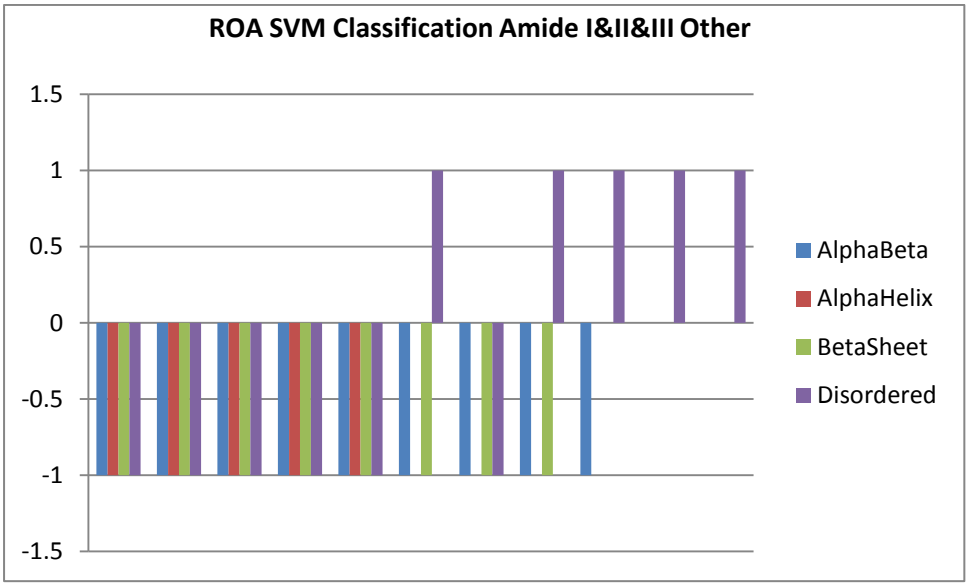
Bar Graph of ROA SVM Classification Amide I&II&III α-Helix using Bin 10 $cm^{-1}$




Bar Graph of ROA SVM Classification Amide I&II&III β-Sheet using Bin 10 $cm^{-1}$

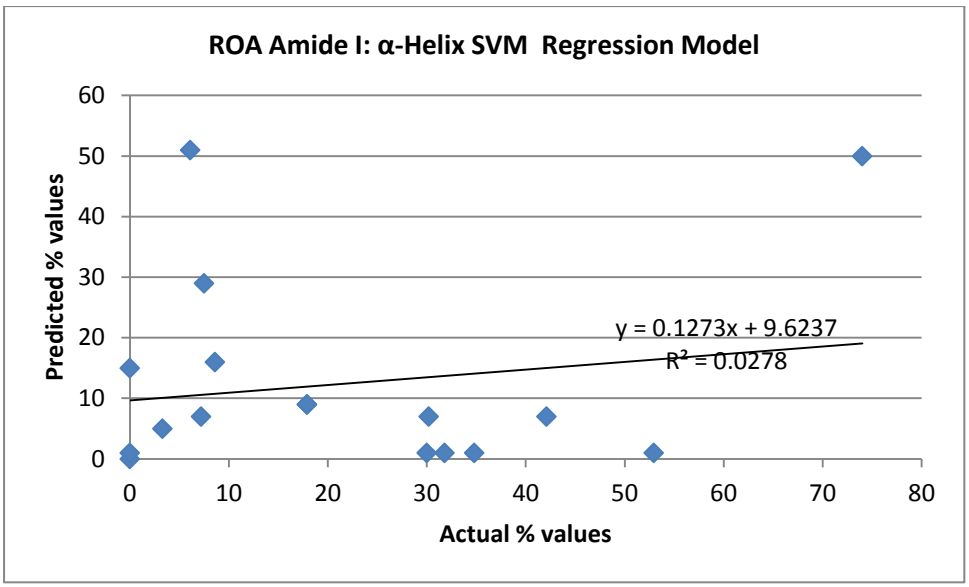


Bar Graph of ROA SVM Classification Amide I&II&III Other using Bin 10 cm $^{-1}$

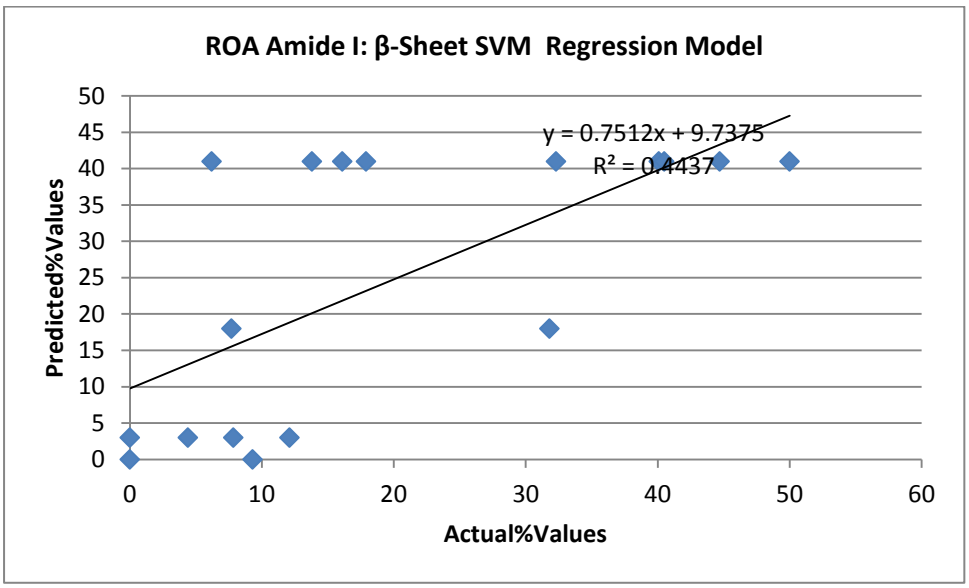


Appendix F- Graphs of SVM regression models showing correlation plots

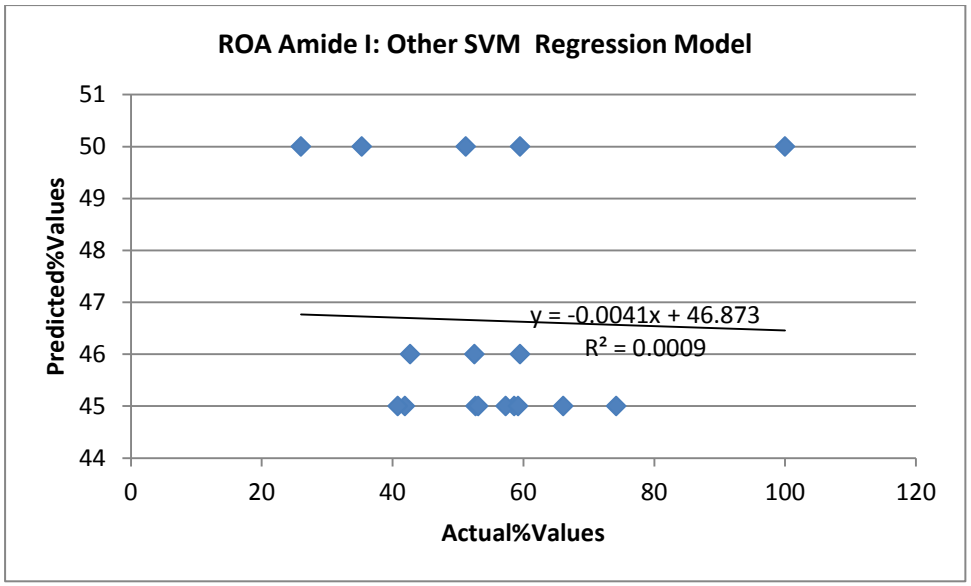
Graph of ROA SVM Regression Amide I α-Helix Model using Bin 10 $cm^{-1}$



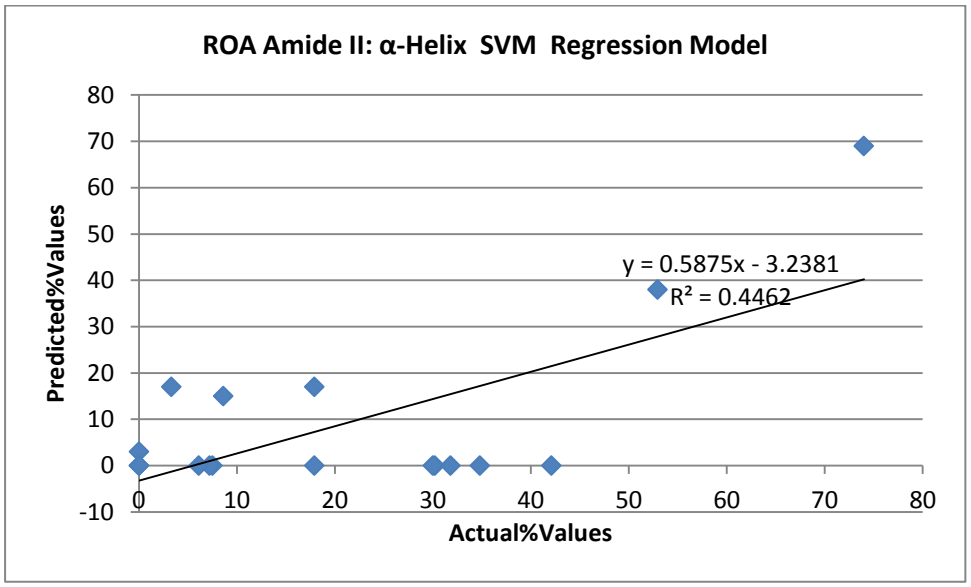

Graph of ROA SVM Regression Amide I β-Sheet Model using Bin 10 $cm^{-1}$



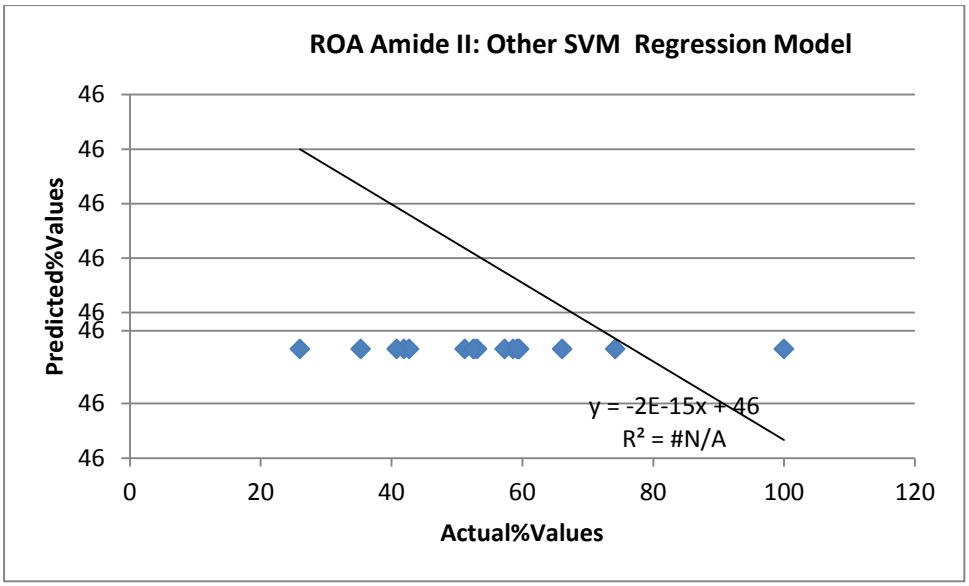
Graph of ROA SVM Regression Amide I Other Model using Bin 10 $cm^{-1}$



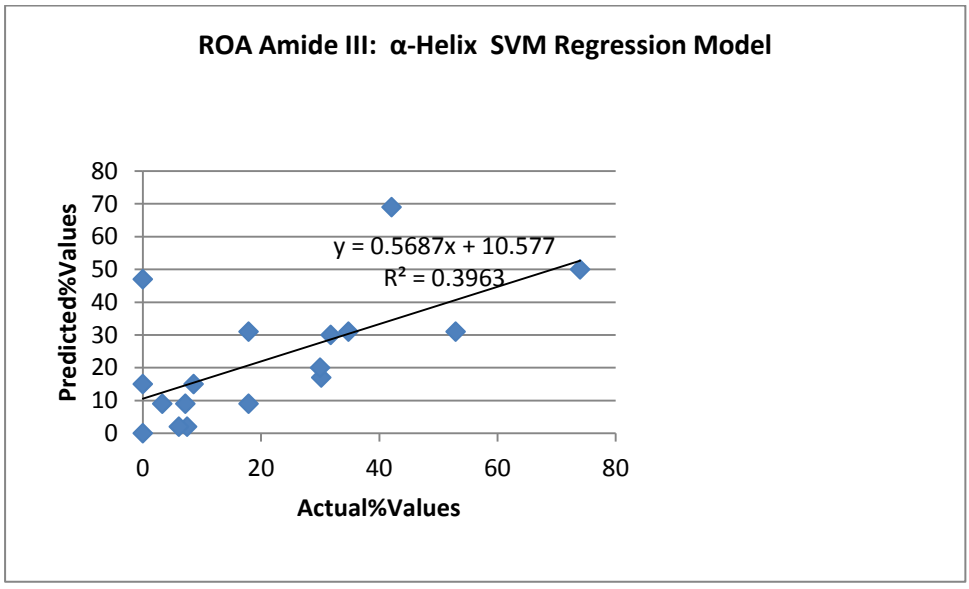

Graph of ROA SVM Regression Amide II α-Helix Model using Bin 10 $cm^{-1}$



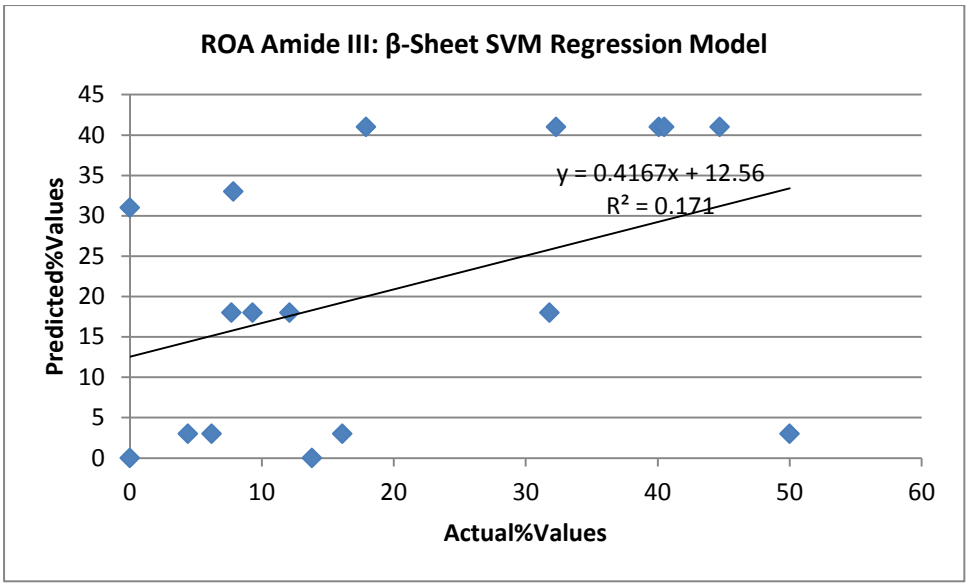
Graph of ROA SVM Regression Amide II Other Model using Bin 10 $cm^{-1}$



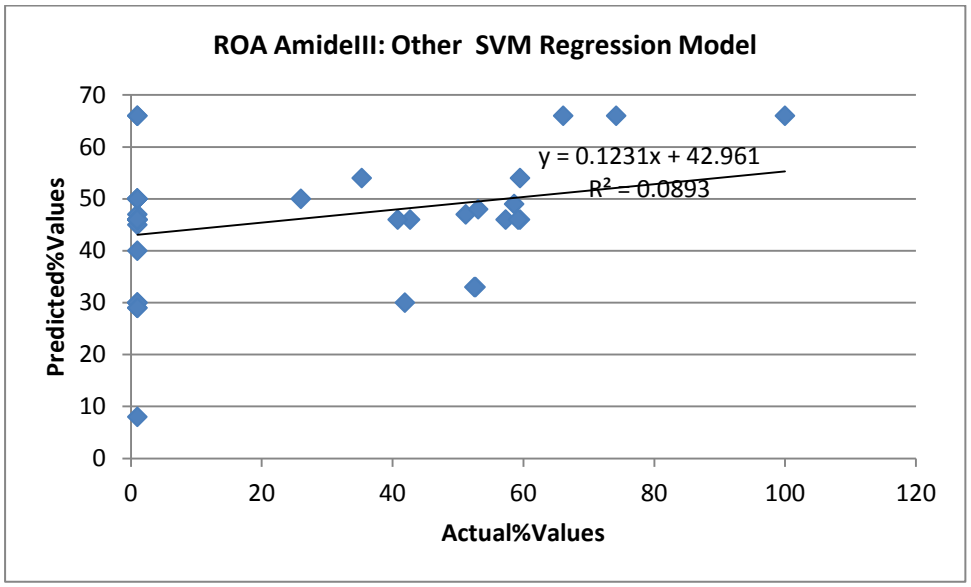

Graph of ROA SVM Regression Amide III α-Helix Model using Bin 10 $cm^{-1}$



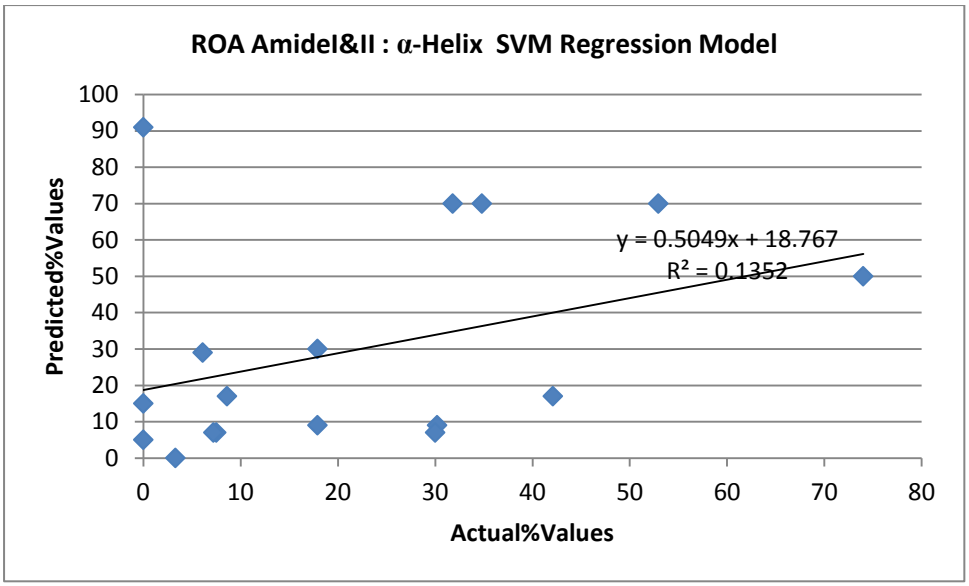
Graph of ROA SVM Regression Amide III β-Sheet Model using Bin 10 $cm^{-1}$



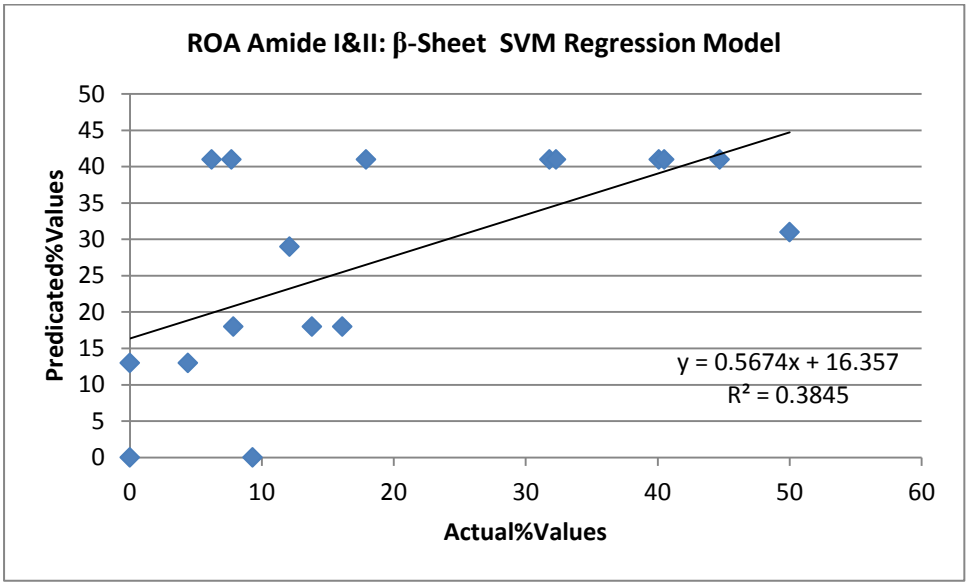

Graph of ROA SVM Regression Amide III Other Model using Bin 10 $cm^{-1}$



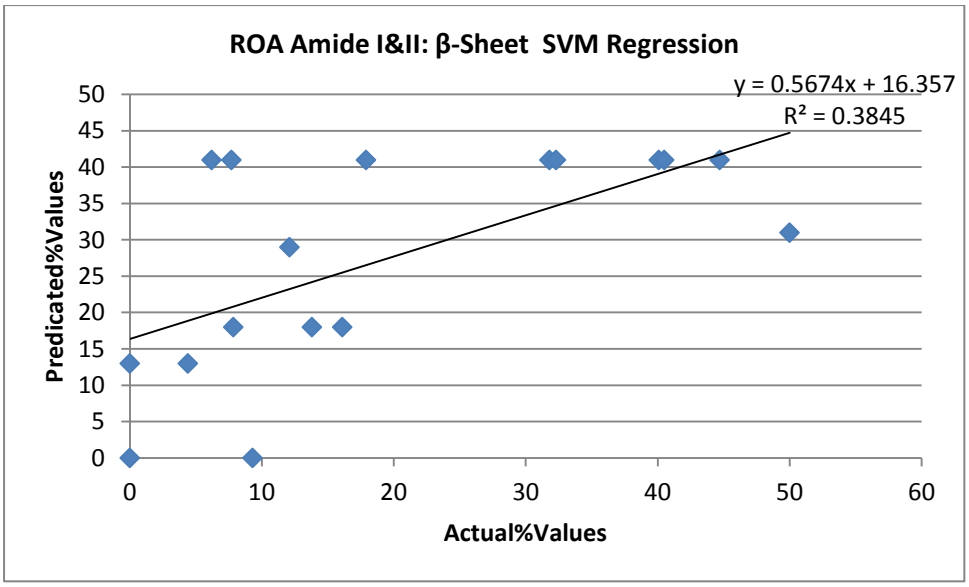
Graph of ROA SVM Regression Amide I&II α-Helix Model using Bin 10 $cm^{-1}$



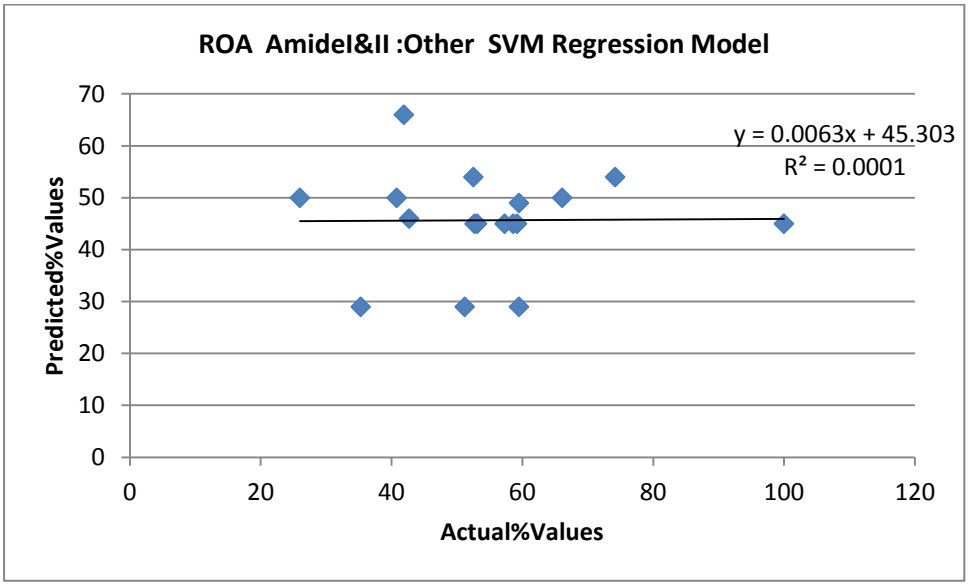

Graph of ROA SVM Regression Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



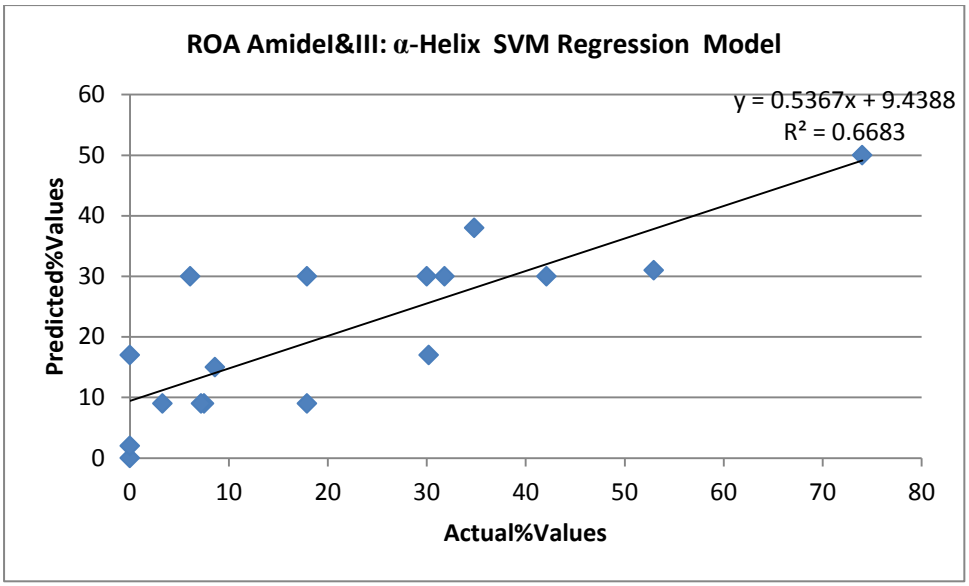
Graph of ROA SVM Regression Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



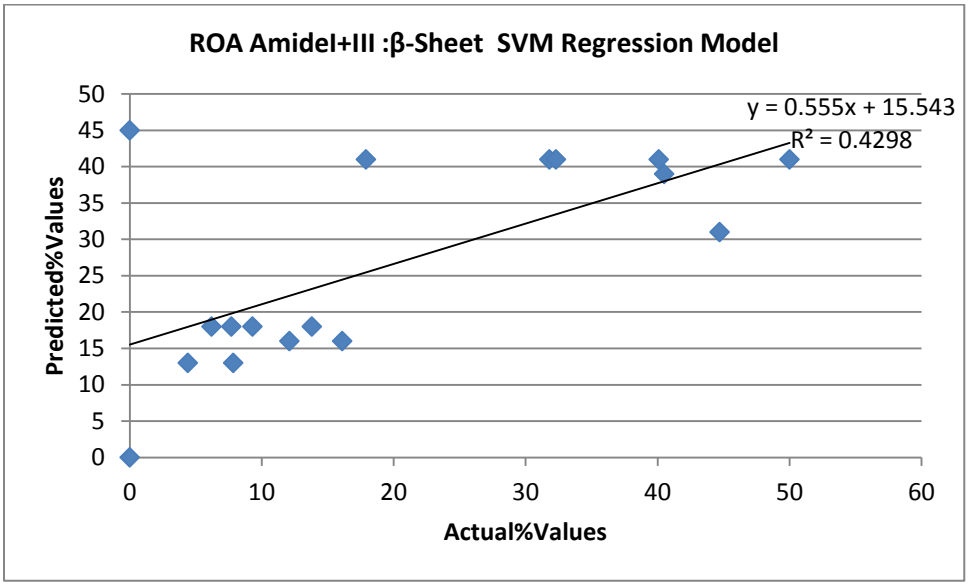

Graph of ROA SVM Regression Amide I&II Other Model using Bin 10 $cm^{-1}$



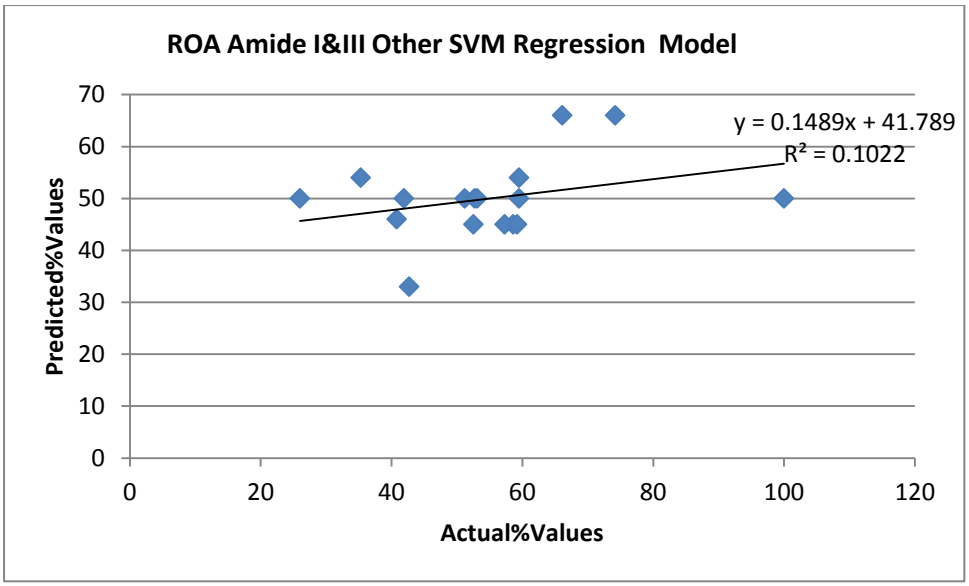
Graph of ROA SVM Regression Amide I&III α-Helix Model using Bin 10 $cm^{-1}$



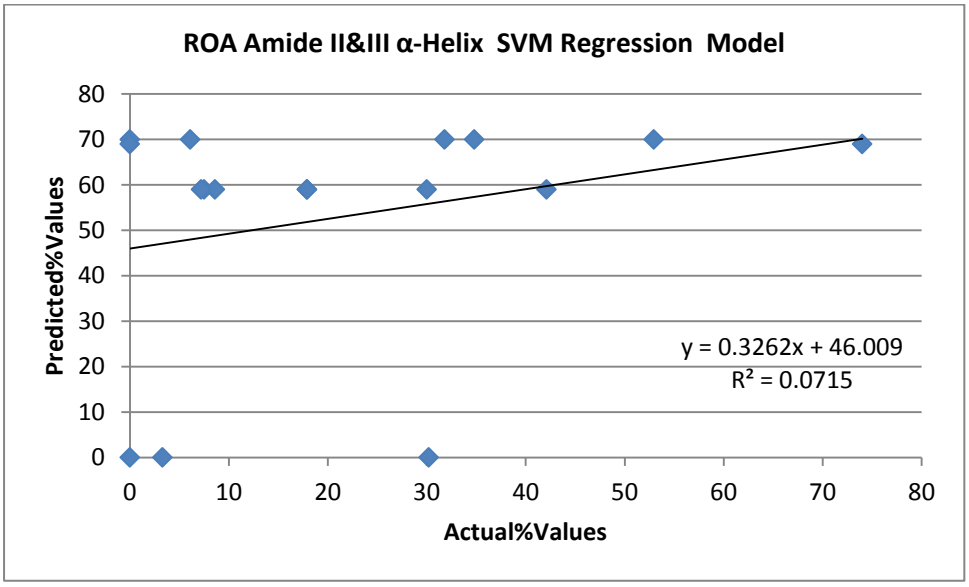

Graph of ROA SVM Regression Amide I&III β-Sheet Model using Bin 10 $cm^{-1}$



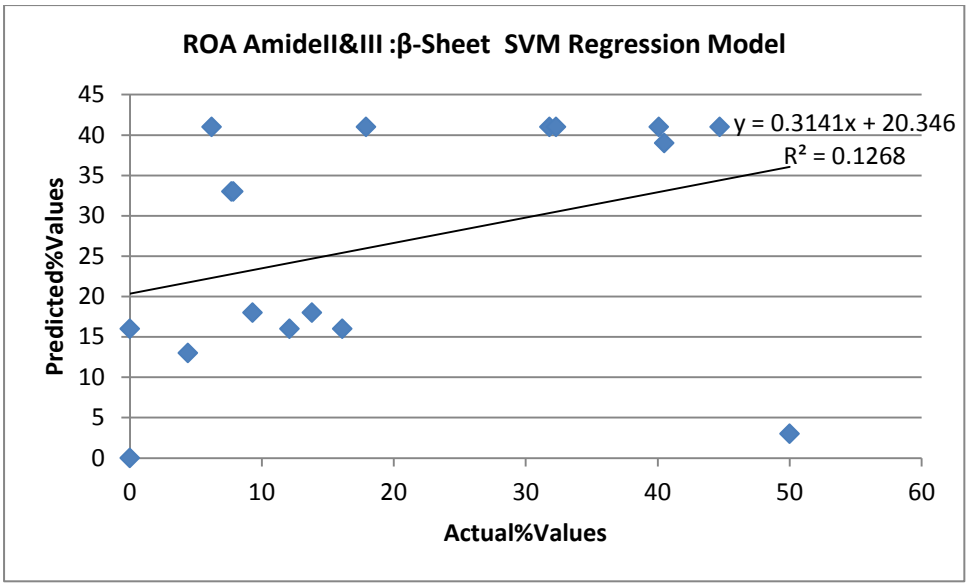
Graph of ROA SVM Regression Amide I&III Other Model using Bin 10 $cm^{-1}$



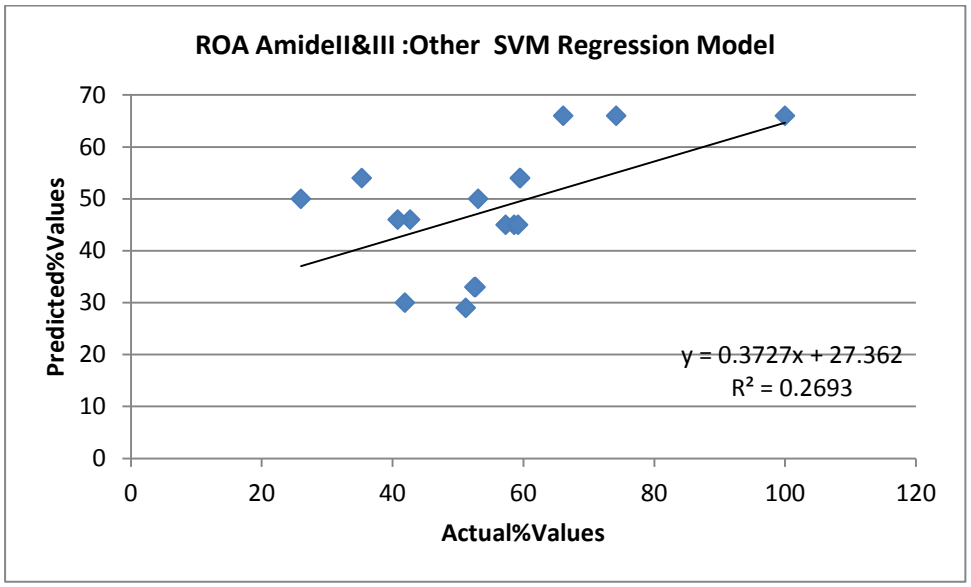

Graph of ROA SVM Regression Amide II&III α-Helix Model using Bin 10 $cm^{-1}$



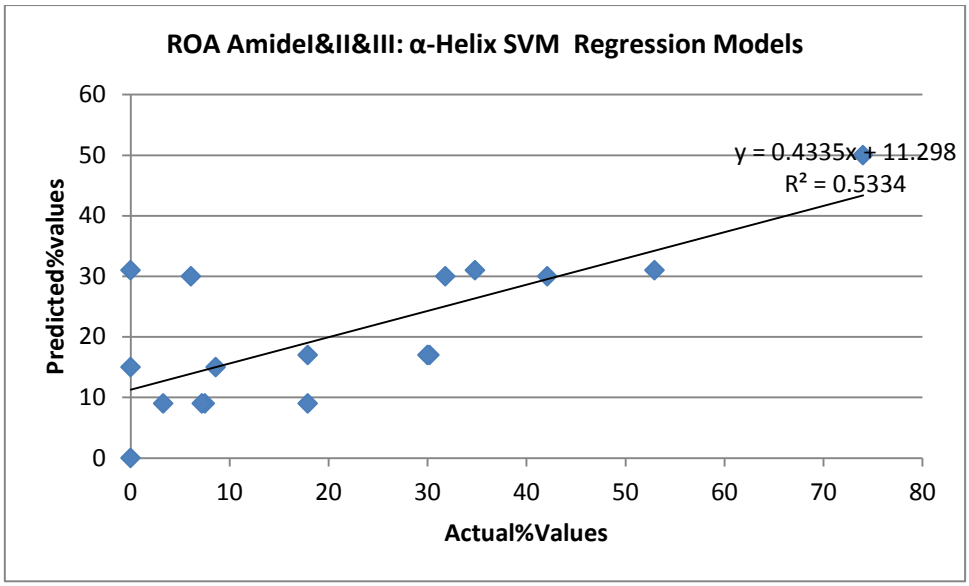
Graph of ROA SVM Regression Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



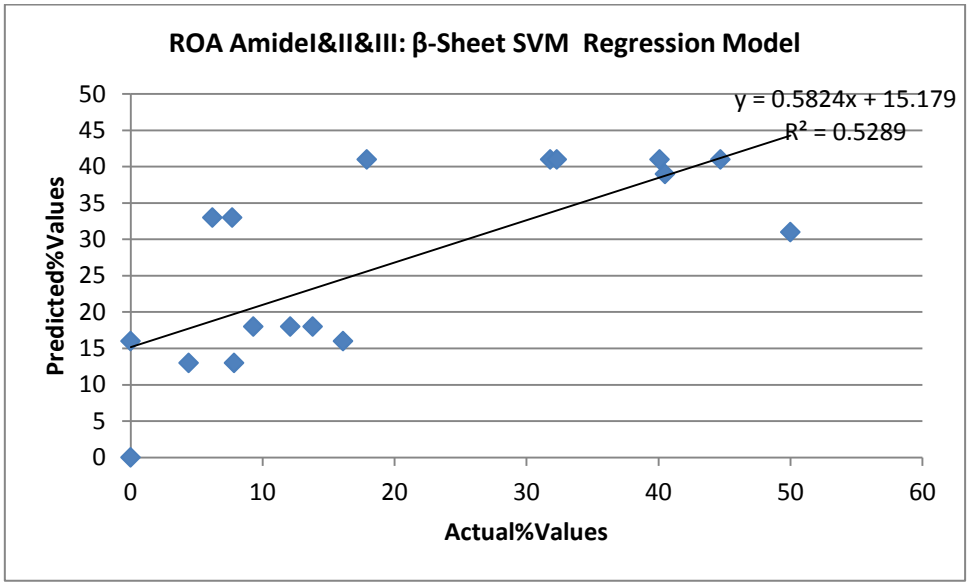

Graph of ROA SVM Regression Amide II&III Other Model using Bin 10 $cm^{-1}$



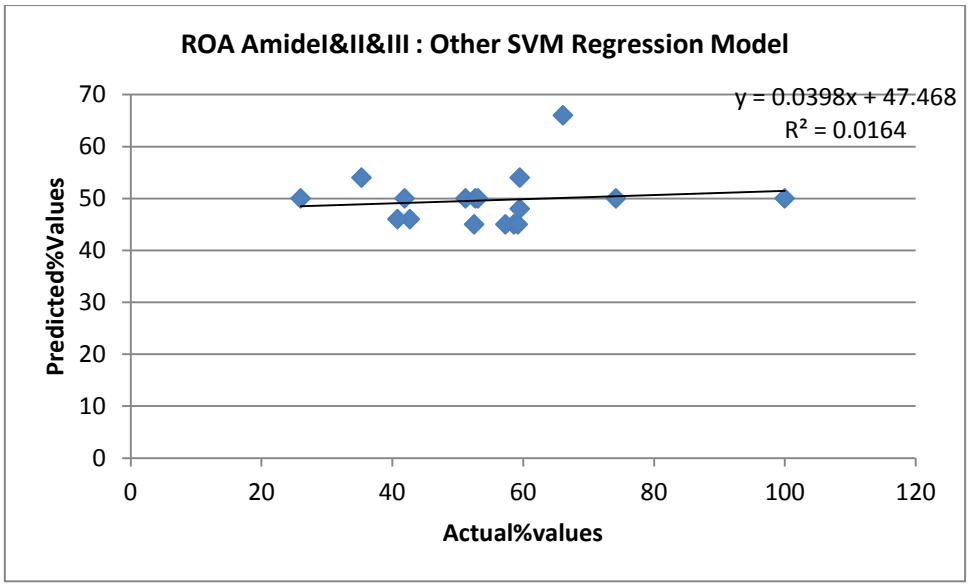
Graph of ROA SVM Regression Amide I&II&III α-Helix Model using Bin 10 $cm^{-1}$



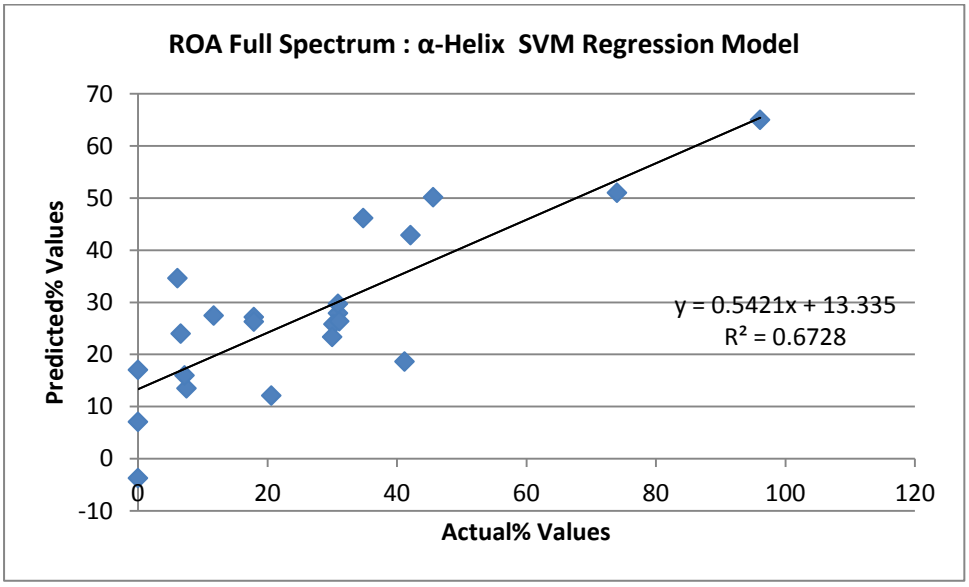

Graph of ROA SVM Regression Amide I&II&III β-Sheet Model using Bin 10 $cm^{-1}$



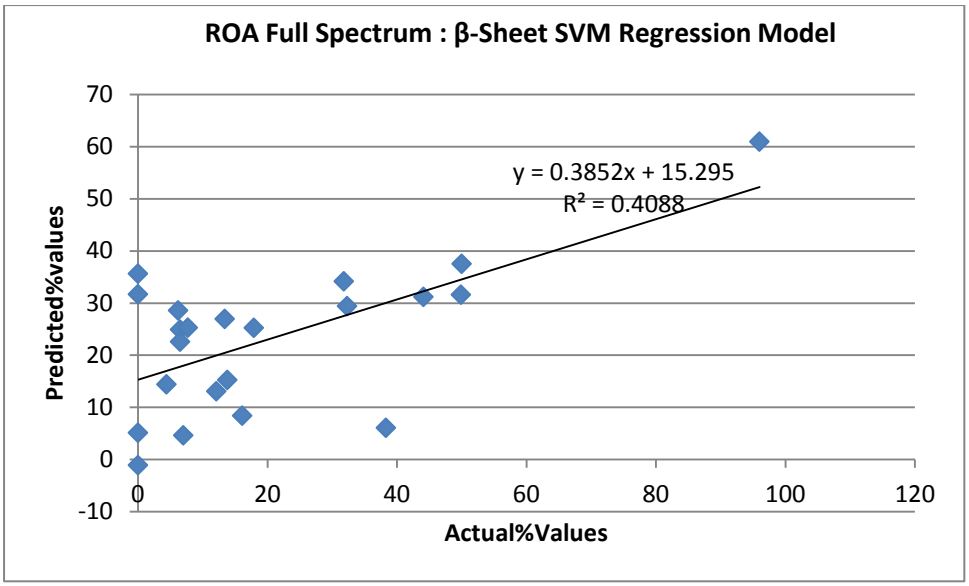
Graph of ROA SVM Regression Amide I&II&III Other Model using Bin 10 $cm^{-1}$



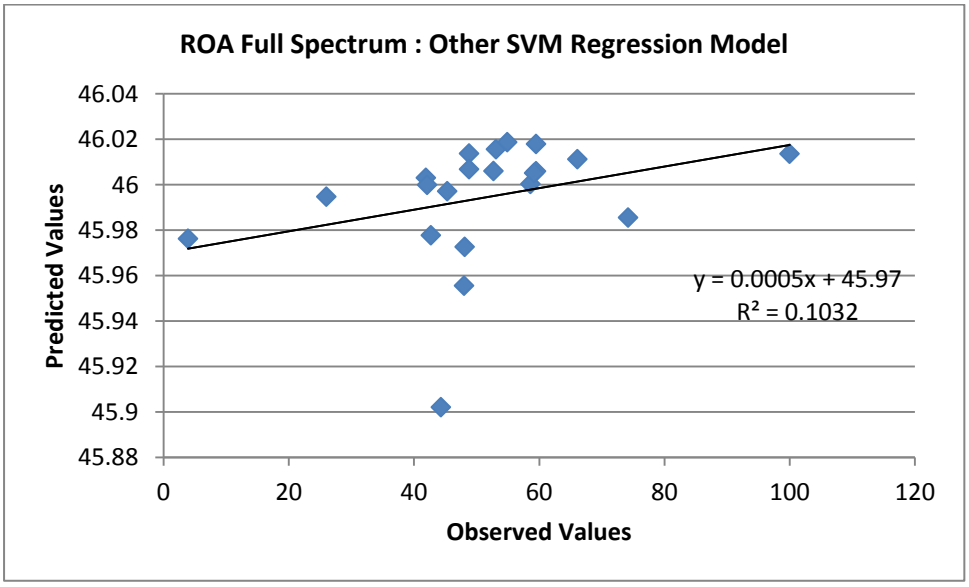

Graph of ROA SVM Regression Full Spectrum α-Helix Model using Bin 10 $cm^{-1}$



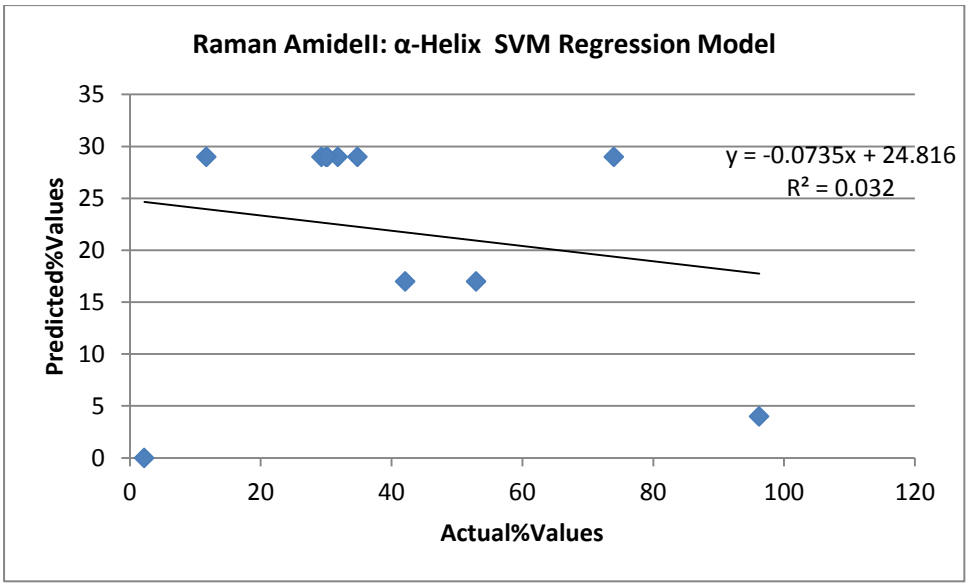
Graph of ROA SVM Regression Full Spectrum β-Sheet Model using Bin 10 $cm^{-1}$



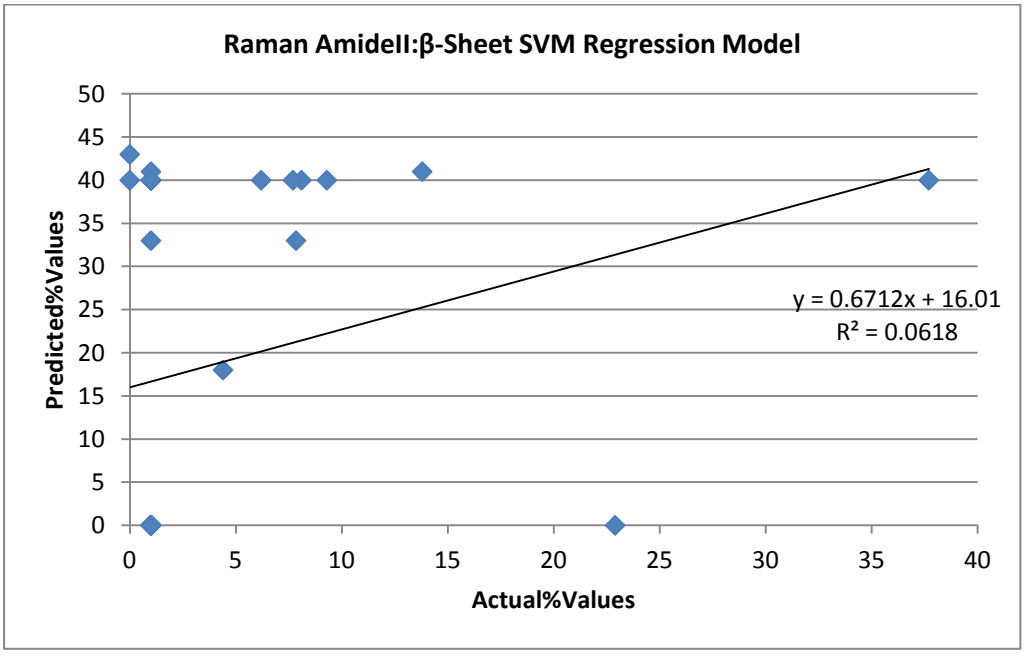

Graph of ROA SVM Regression Full Spectrum Other Model using Bin 10 $cm^{-1}$



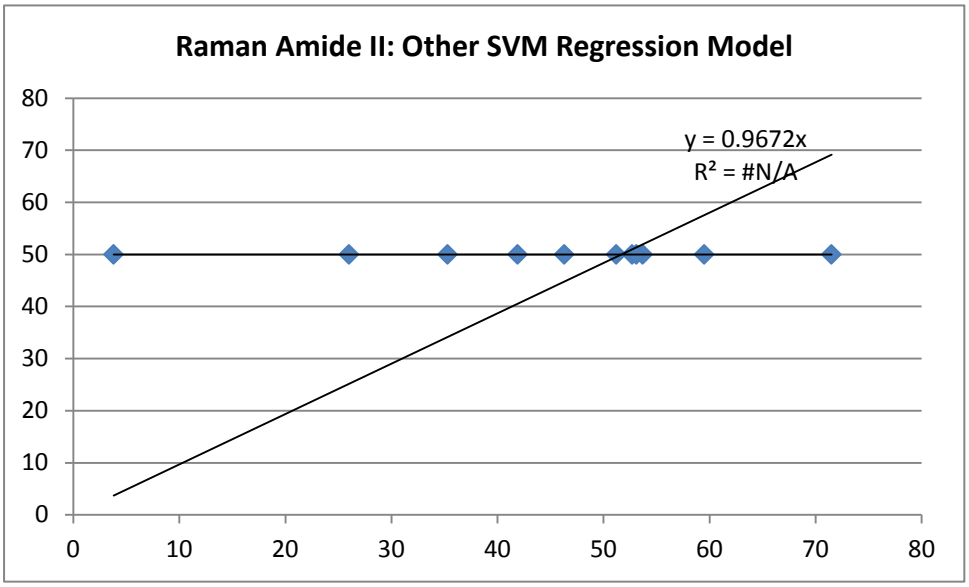
Graph of Raman SVM Regression Amide II α-Helix Model using Bin 10 $cm^{-1}$



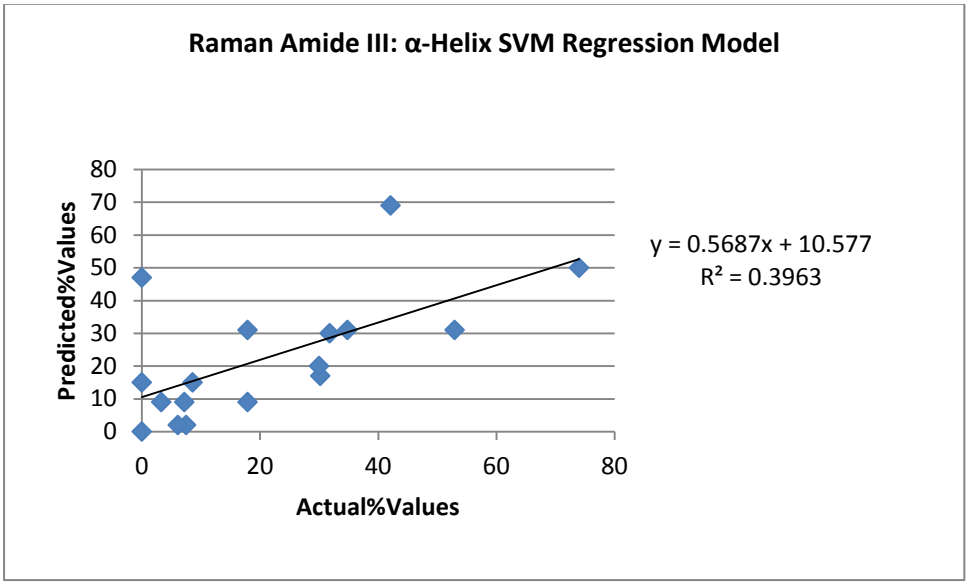

Graph of Raman SVM Regression Amide II β-Sheet Model using Bin 10 $cm^{-1}$



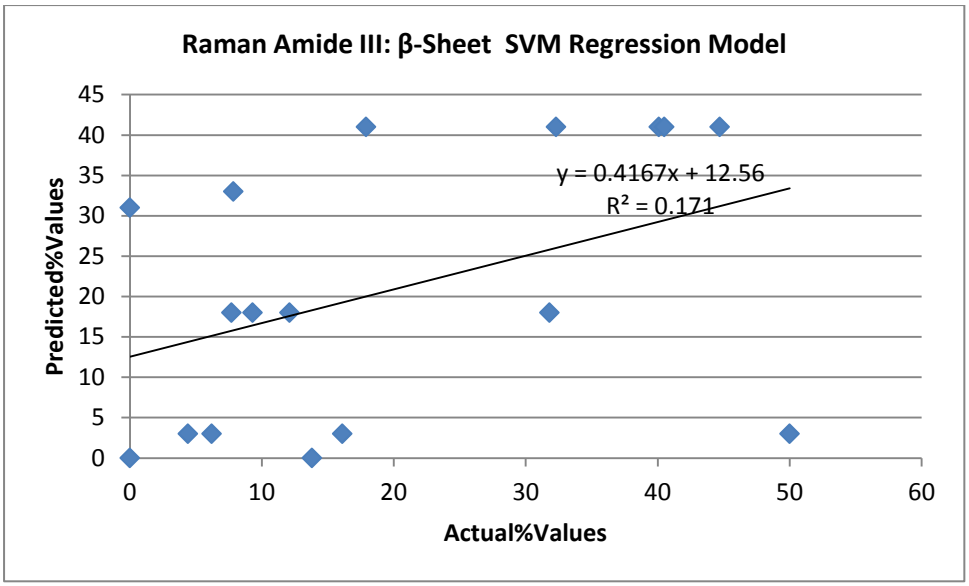
Graph of Raman SVM Regression Amide II Other Model using Bin 10 cm $^{-1}$



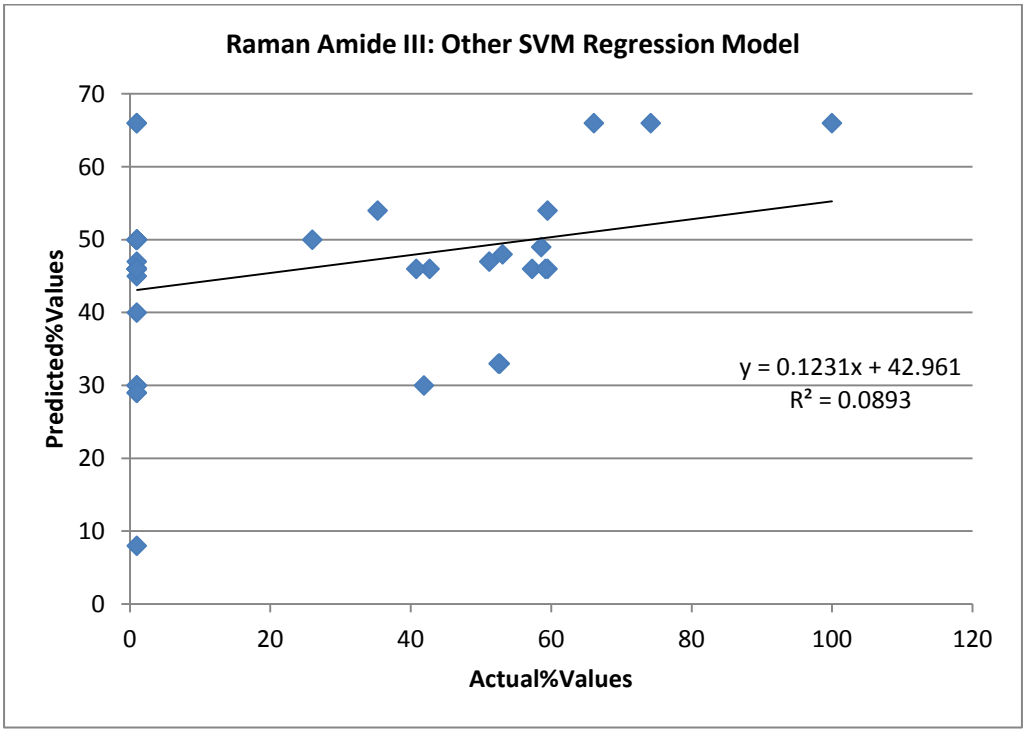

Graph of Raman SVM Regression Amide III α-Helix Model using Bin 10 cm $^{-1}$



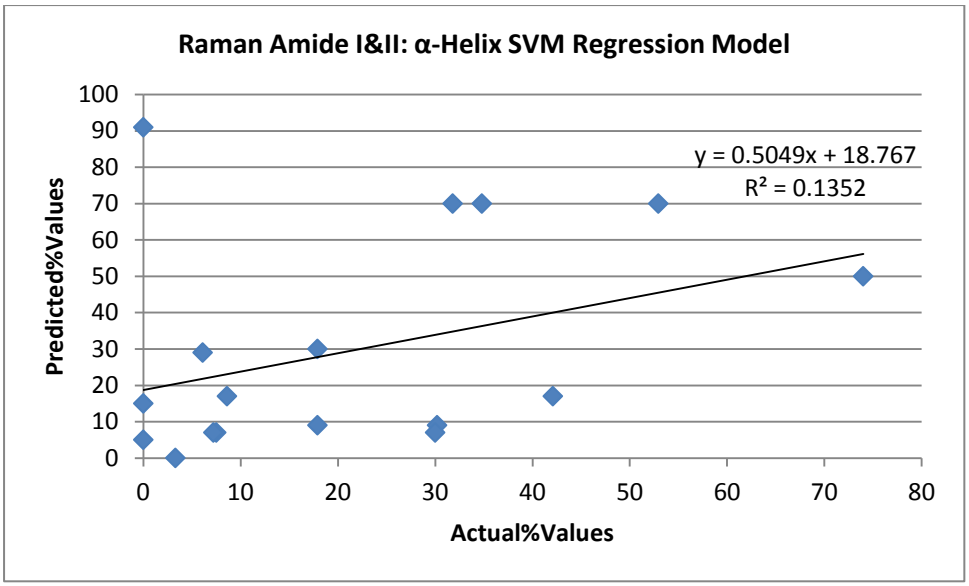
Graph of Raman SVM Regression Amide III β-Sheet Model using Bin 10 cm $^{-1}$



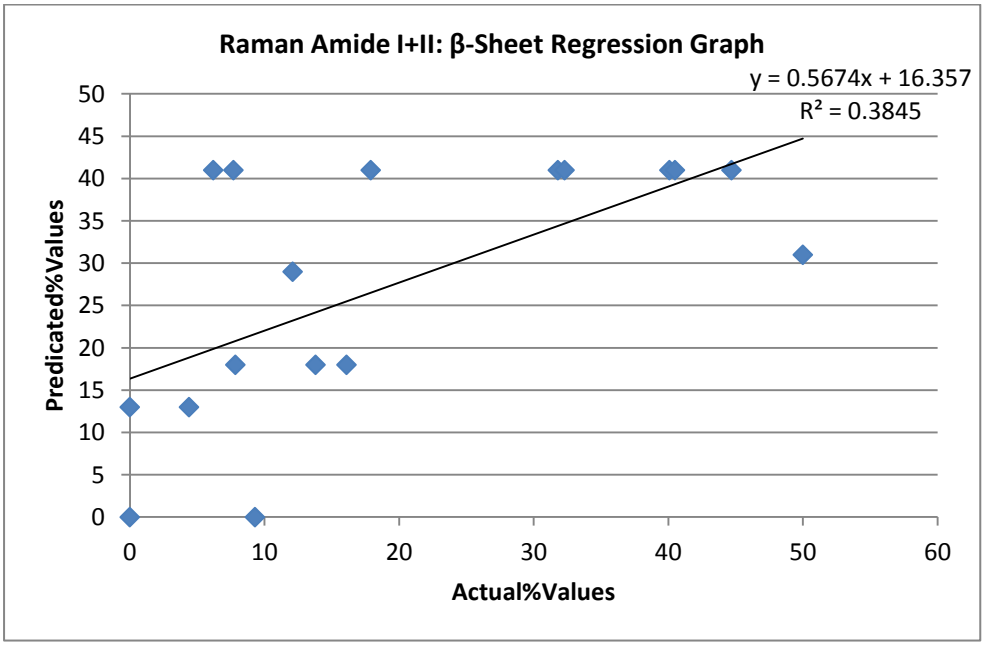

Graph of Raman SVM Regression Amide III Other Model using Bin 10 cm $^{-1}$



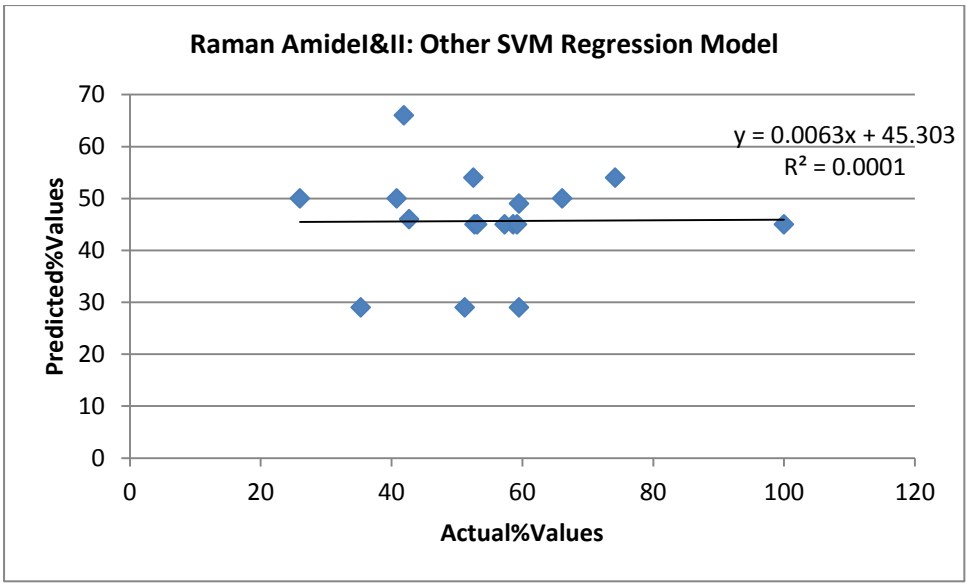
Graph of Raman SVM Regression Amide I&II α-Helix Model using Bin 10 $cm^{-1}$



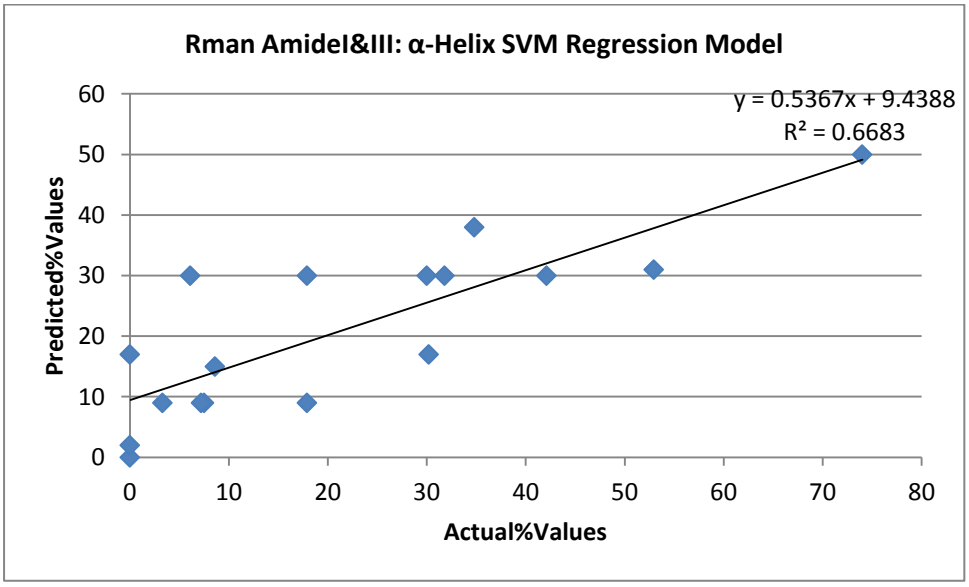

Graph of Raman SVM Regression Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



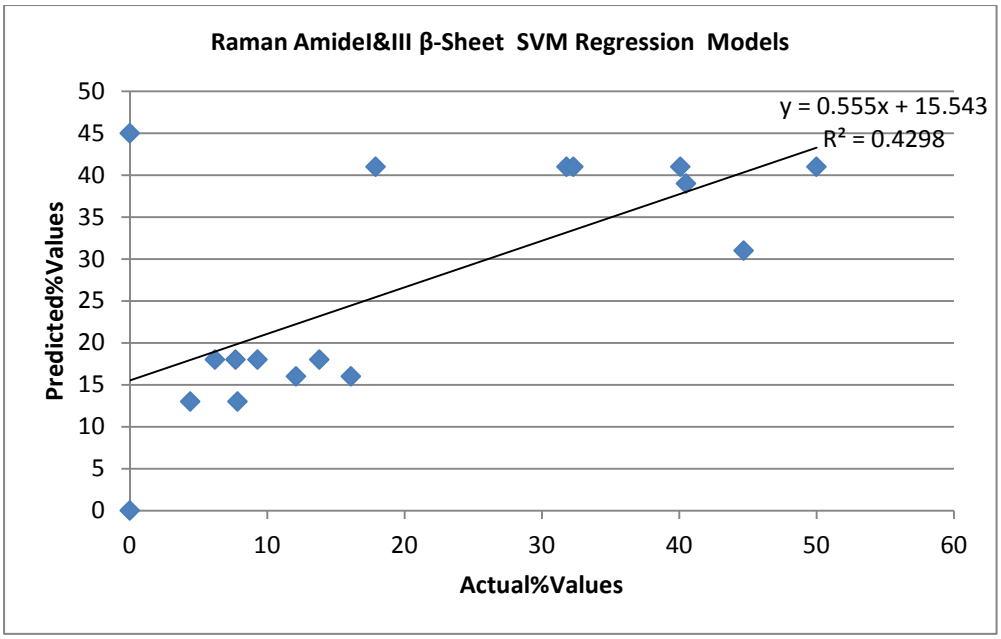
Graph of Raman SVM Regression Amide I&II Other Model using Bin 10 $cm^{-1}$



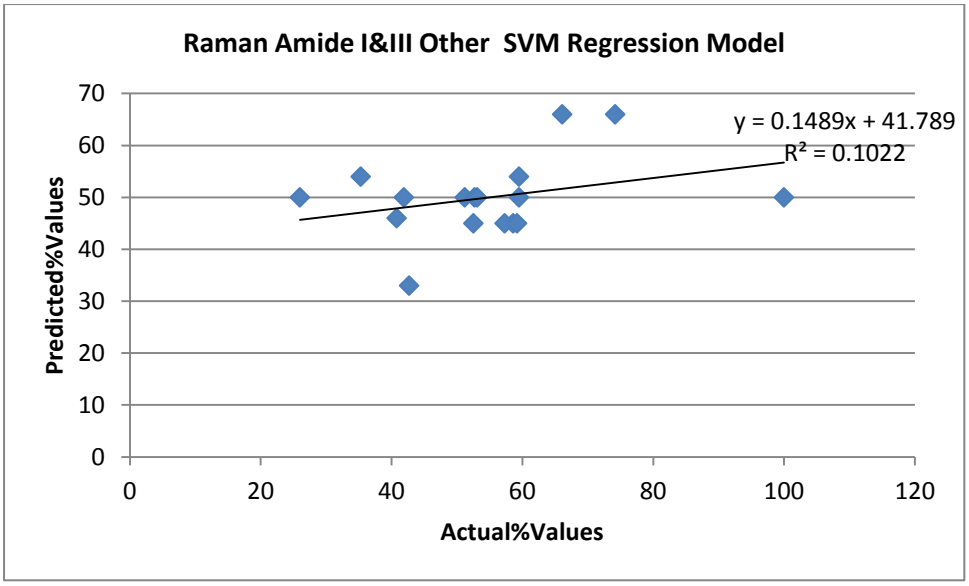

Graph of Raman SVM Regression Amide I&III α-Helix Model using Bin 10 $cm^{-1}$



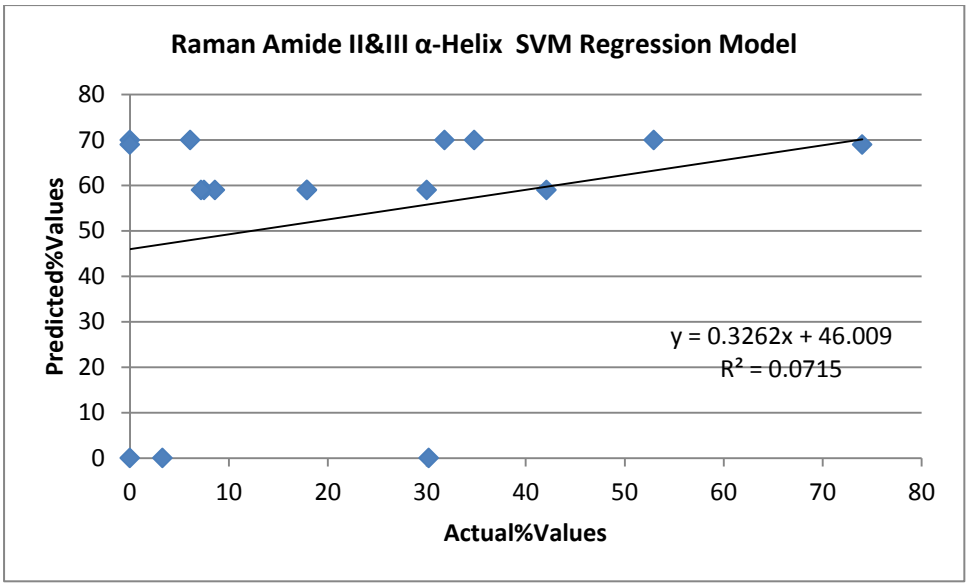
Graph of Raman SVM Regression Amide I&III β-Sheet Model using Bin 10 $cm^{-1}$



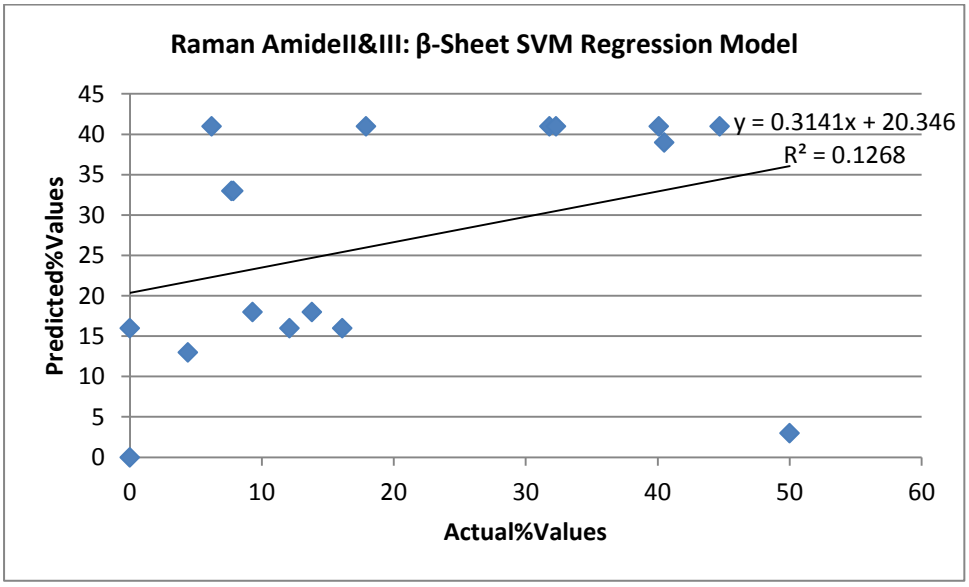

Graph of Raman SVM Regression Amide I&III Other Model using Bin 10 $cm^{-1}$



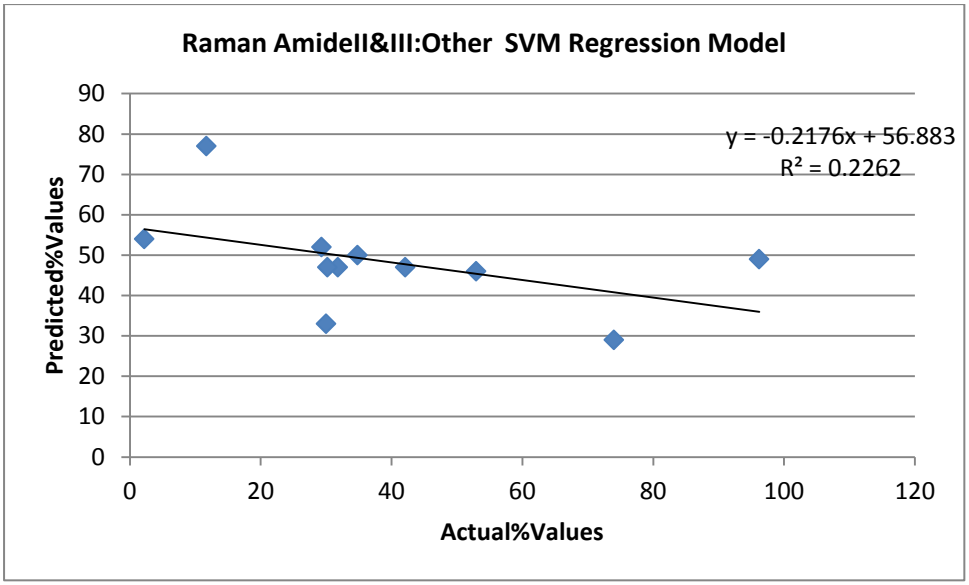
Graph of Raman SVM Regression Amide II&III α-Helix Model using Bin 10 $cm^{-1}$



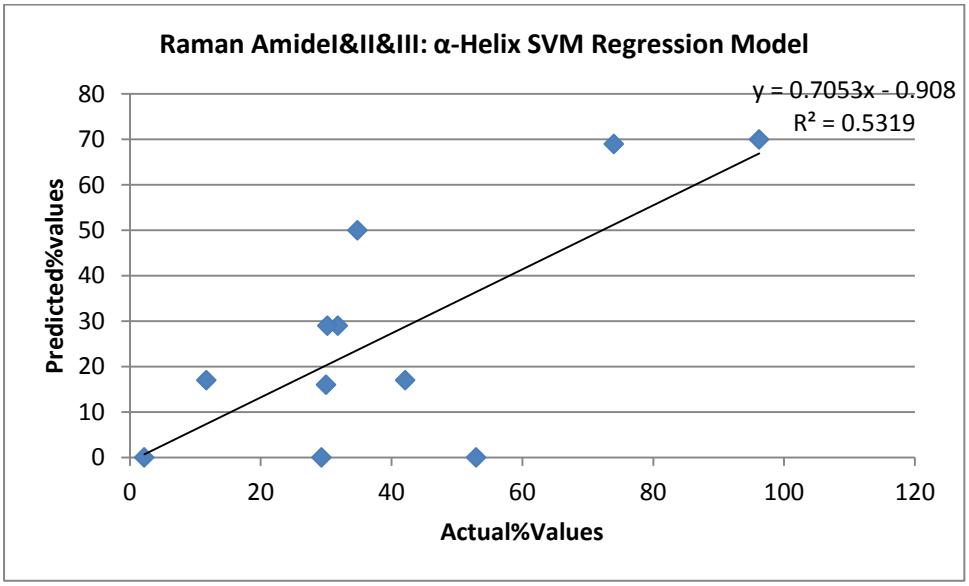

Graph of Raman SVM Regression Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



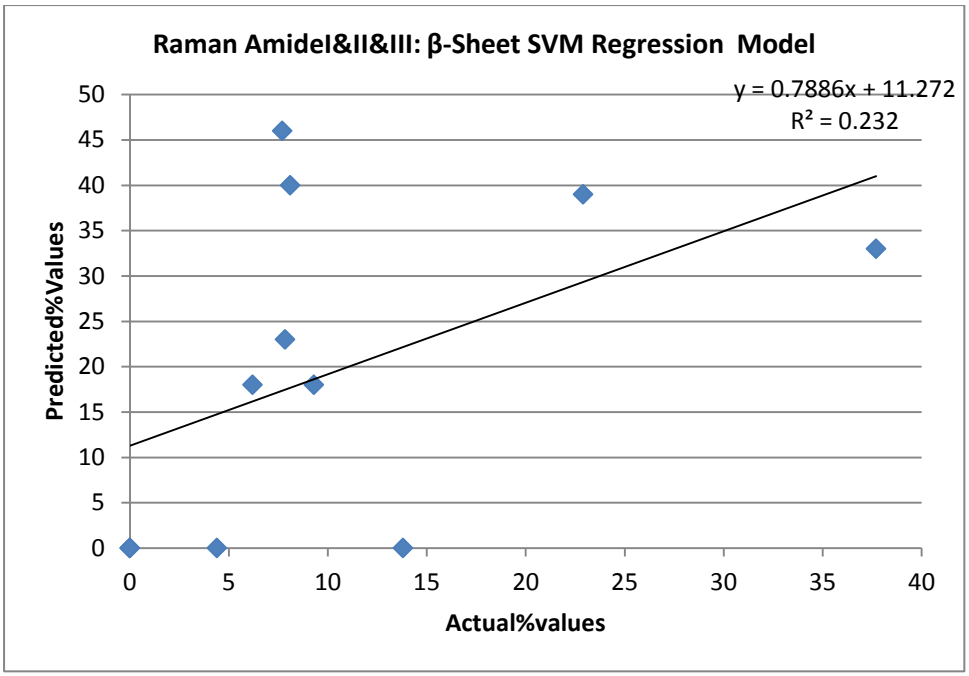
Graph of Raman SVM Regression Amide II&III Other Model using Bin 10 cm $^{-1}$



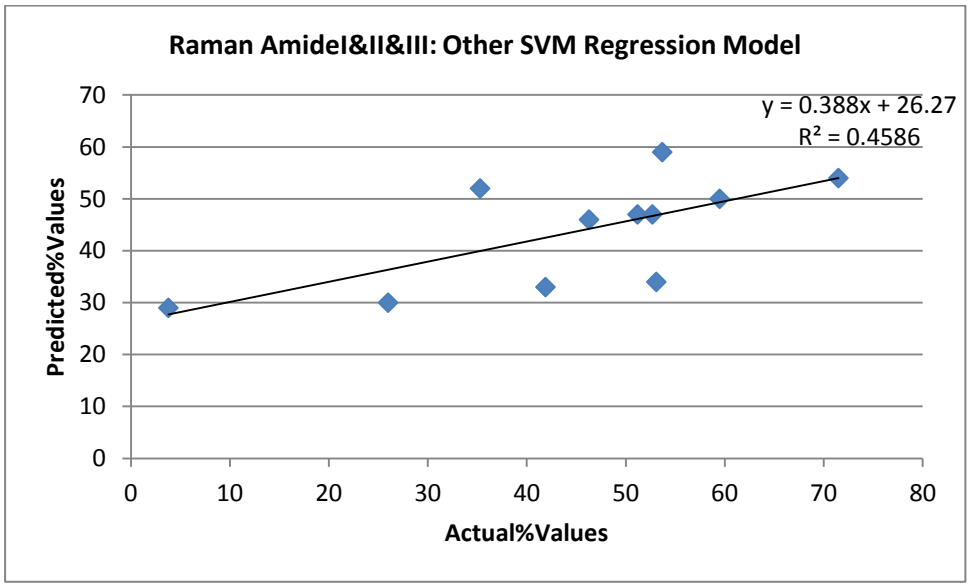

Graph of Raman SVM Regression Amide I&II&III α-Helix Model using Bin 10 cm $^{-1}$



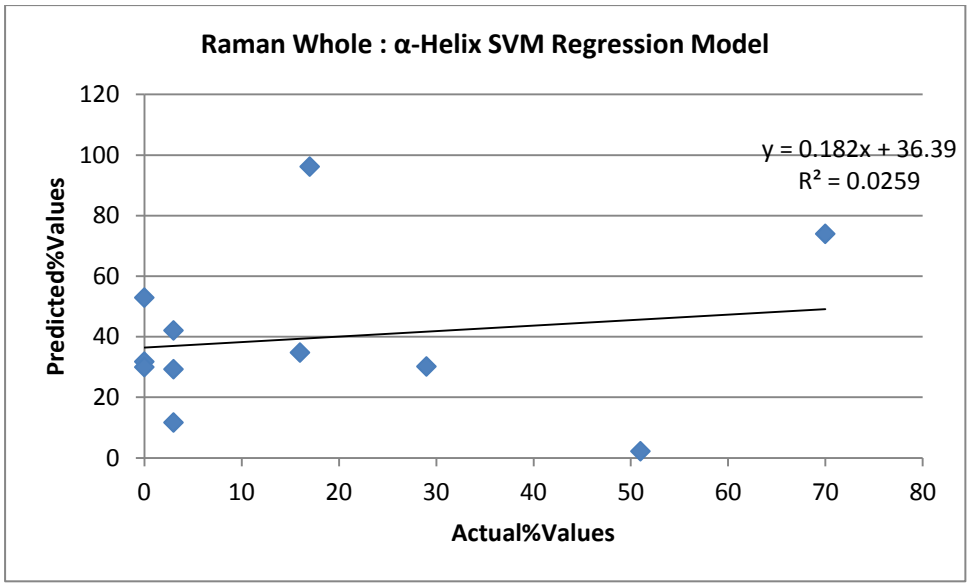
Graph of Raman SVM Regression Amide I&II&III β-Sheet Model using Bin 10 $cm^{-1}$



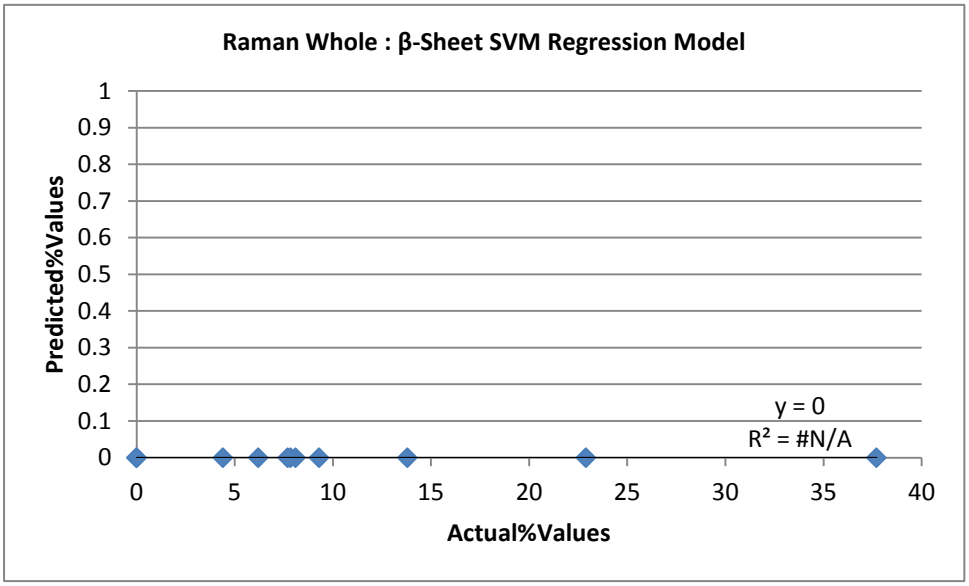

Graph of Raman SVM Regression Amide I&II&III Other Model using Bin 10 $cm^{-1}$



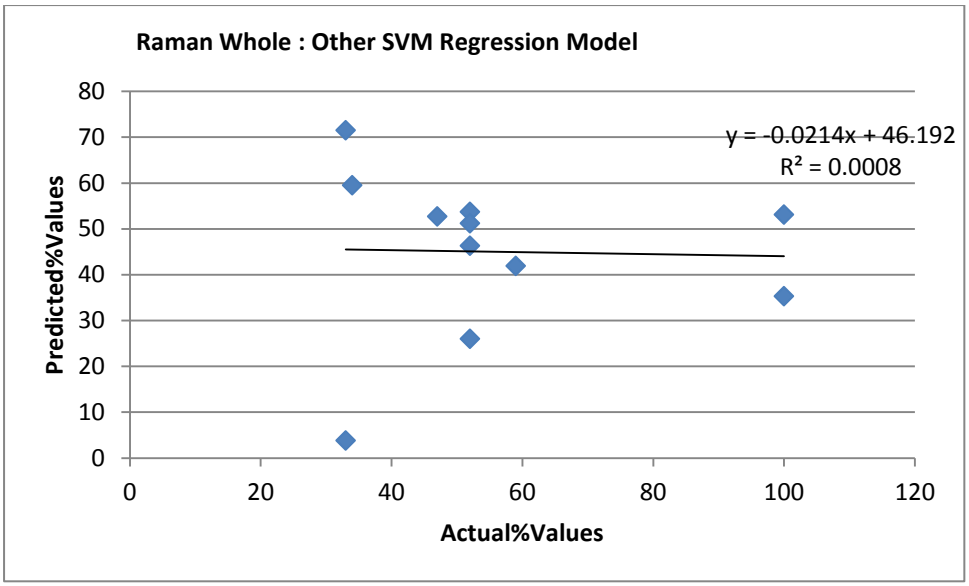
Graph of Raman SVM Regression Full Spectrum α-Helix Model using Bin 10 $cm^{-1}$




Graph of Raman SVM Regression Full Spectrum β-Sheet Model using Bin 10 $cm^{-1}$

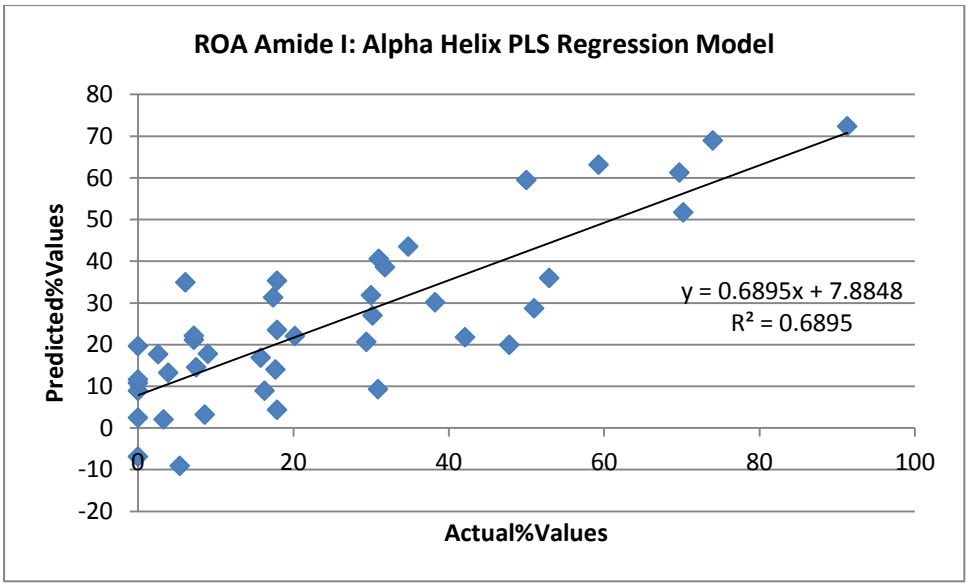


Graph of Raman SVM Regression Full Spectrum Other Model using Bin 10 $cm^{-1}$

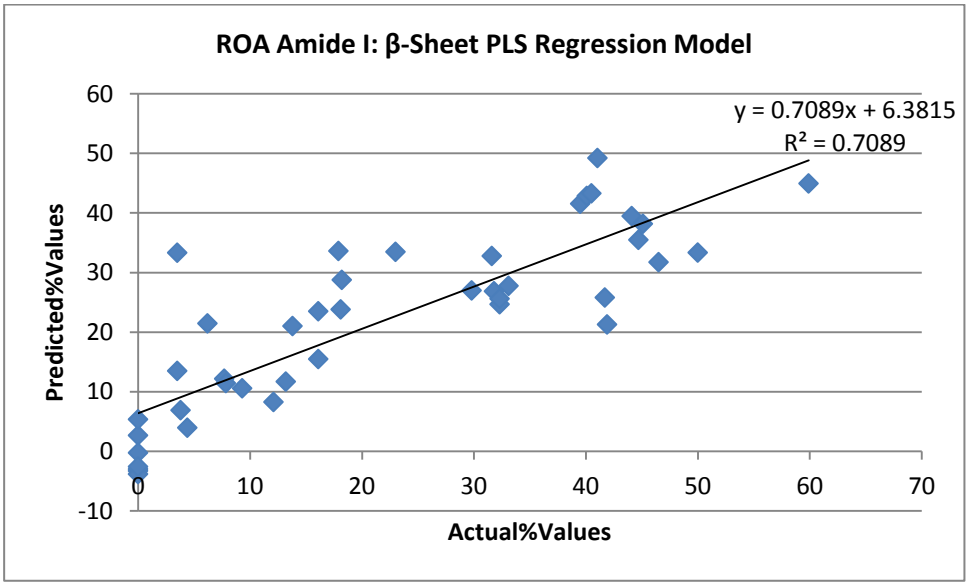


Appendix G- Graphs showing correlation of PLS regression models

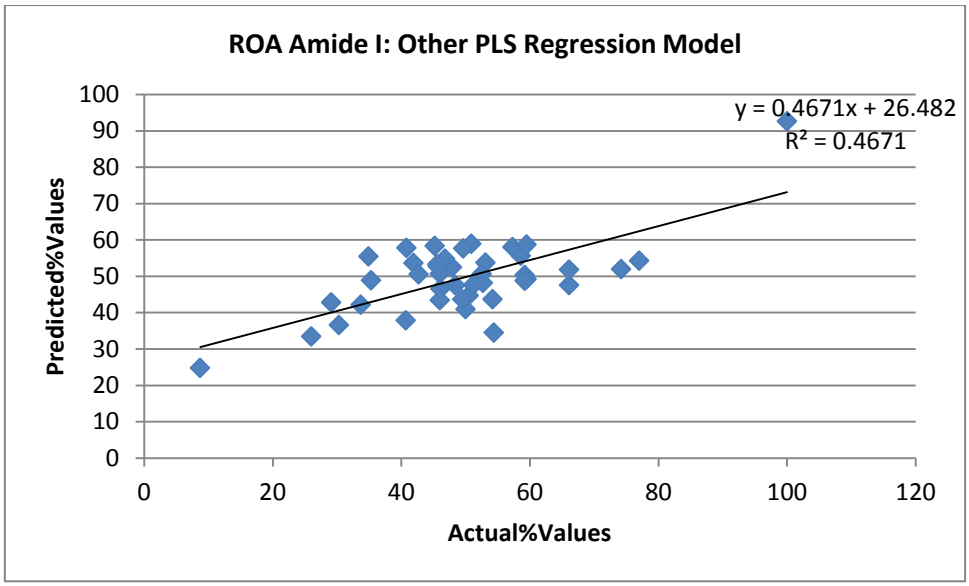
Graph of ROA PLS Regression Amide I α-Helix Model using Bin 10 $cm^{-1}$



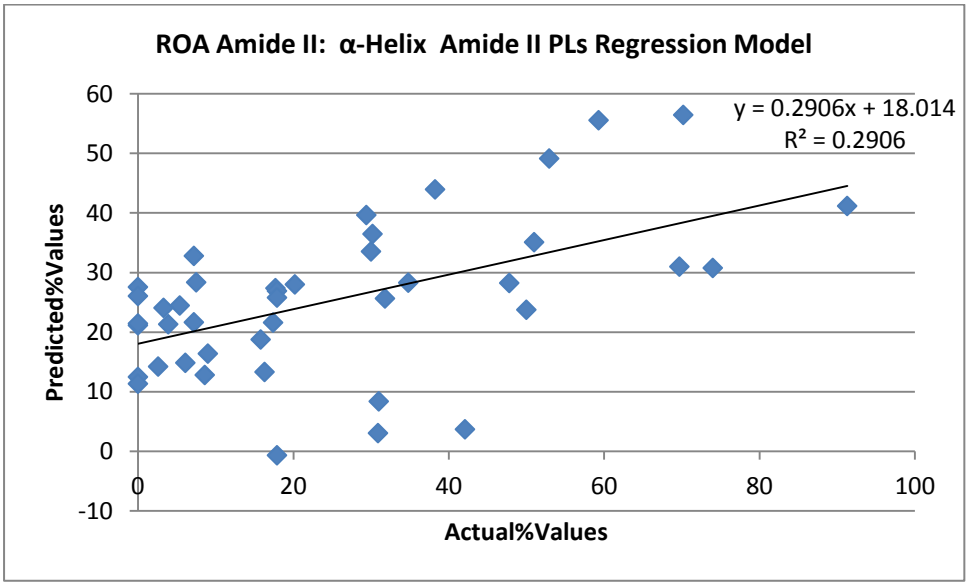

Graph of ROA PLS Regression Amide I β-Sheet Model using Bin 10 $cm^{-1}$



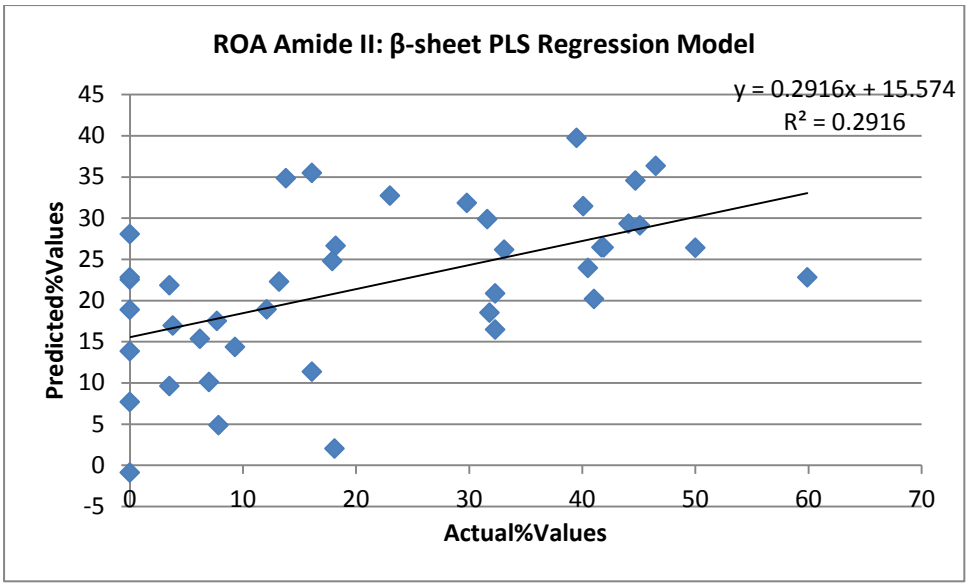
Graph of ROA PLS Regression Amide I Other Model using Bin 10 cm $^{-1}$



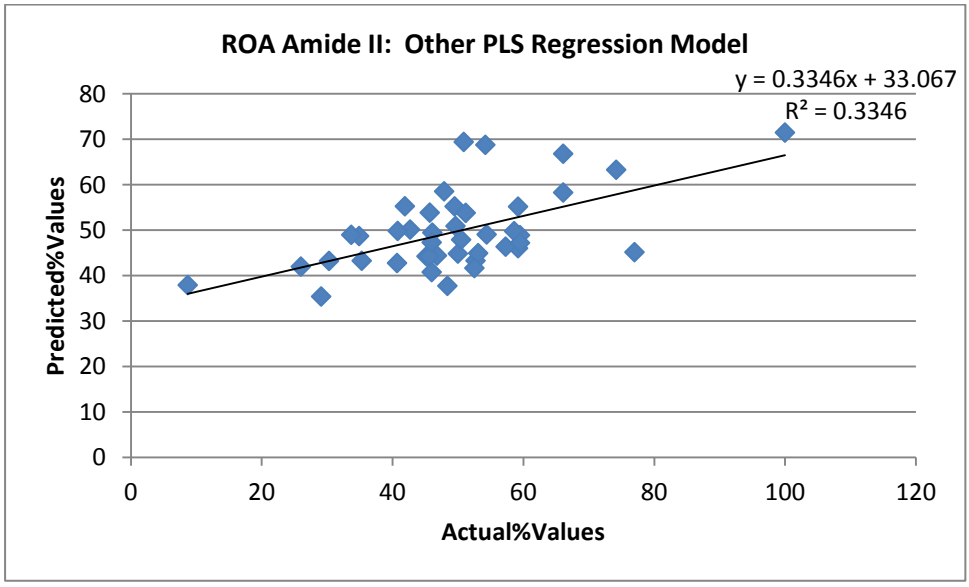

Graph of ROA PLS Regression Amide II α-Helix Model using Bin 10 cm $^{-1}$



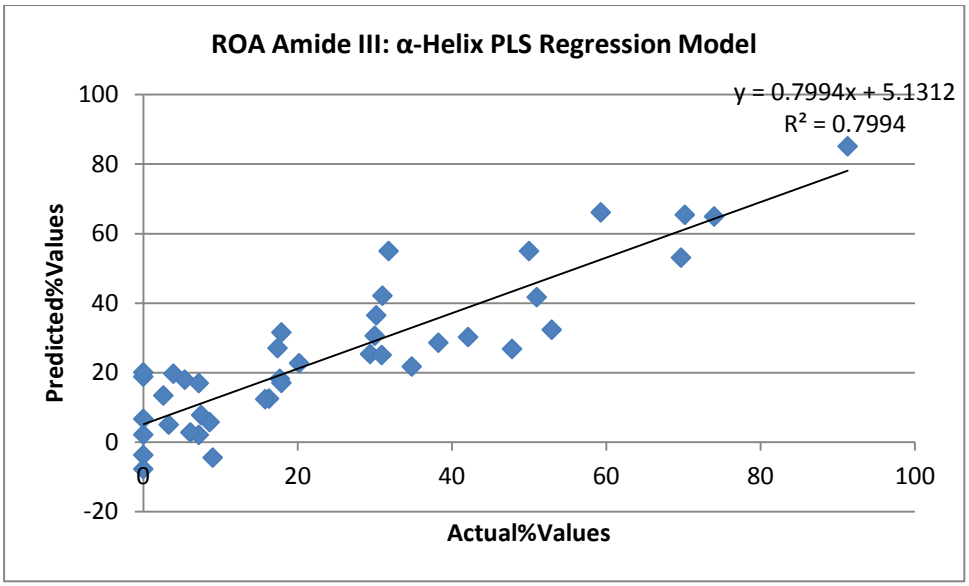
Graph of ROA PLS Regression Amide II β-sheet Model using Bin 10 $cm^{-1}$



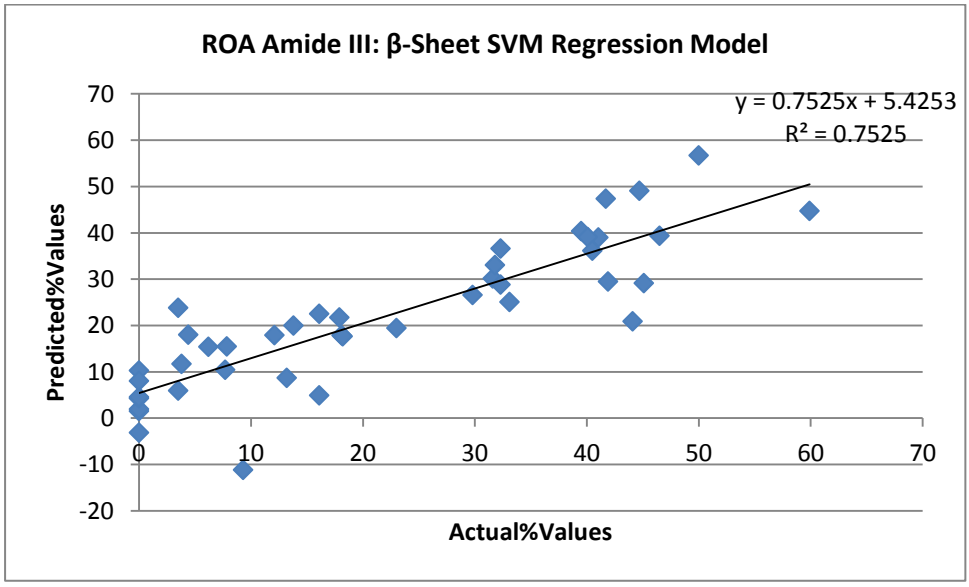

Graph of ROA PLS Regression Amide II Other Model using Bin 10 $cm^{-1}$



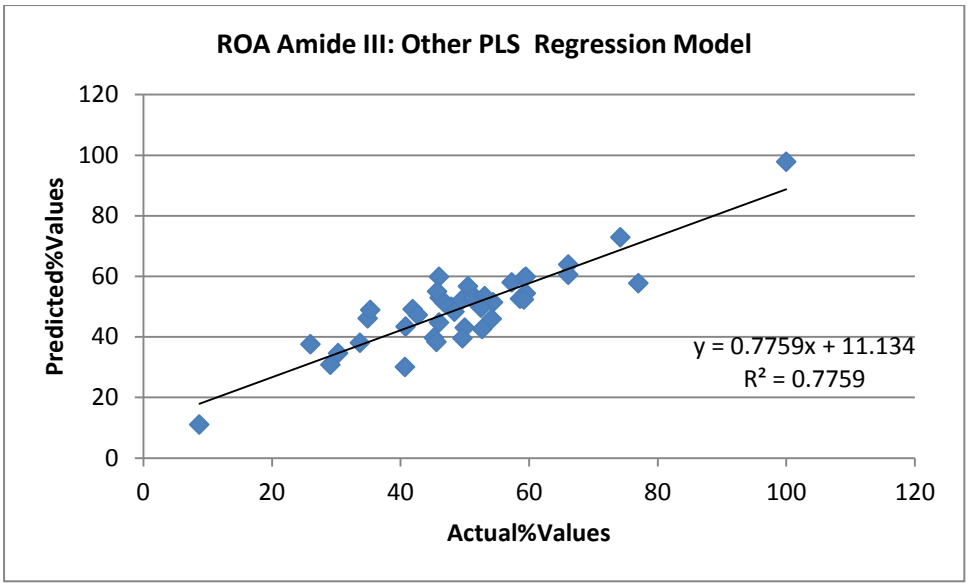
Graph of ROA PLS Regression Amide III α-Helix Model using Bin 10 $cm^{-1}$



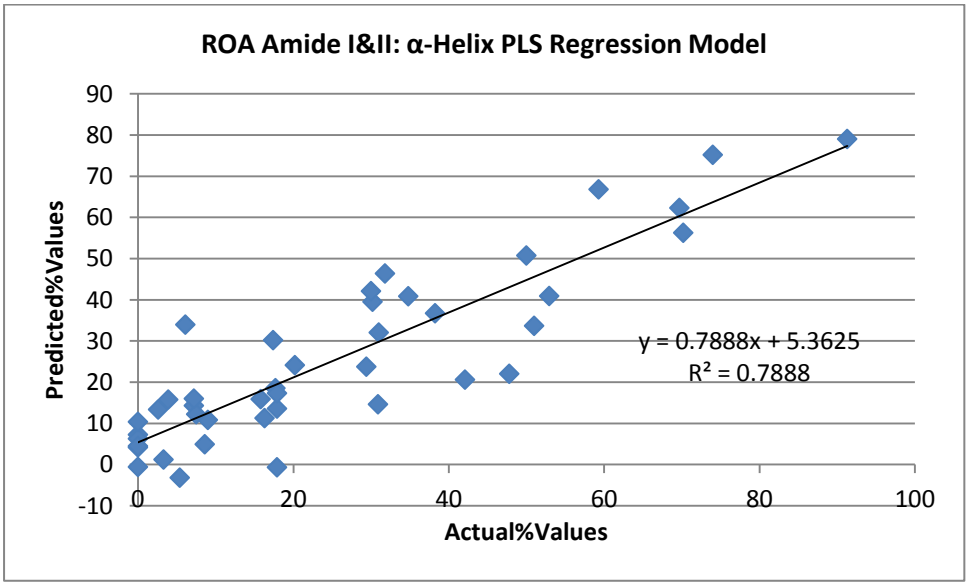

Graph of ROA PLS Regression Amide III β-Sheet Model using Bin 10 $cm^{-1}$



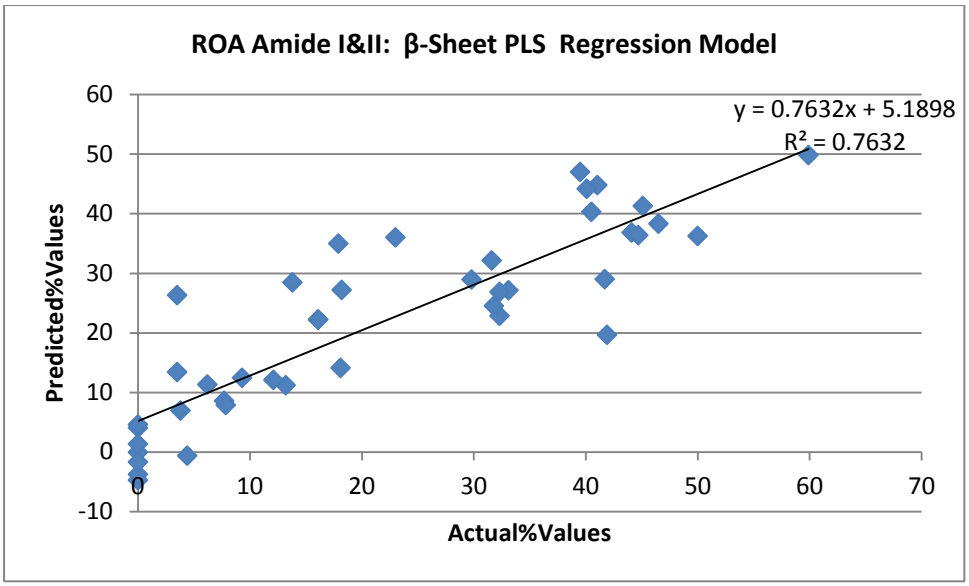
Graph of ROA PLS Regression Amide III Other Model using Bin 10 $cm^{-1}$



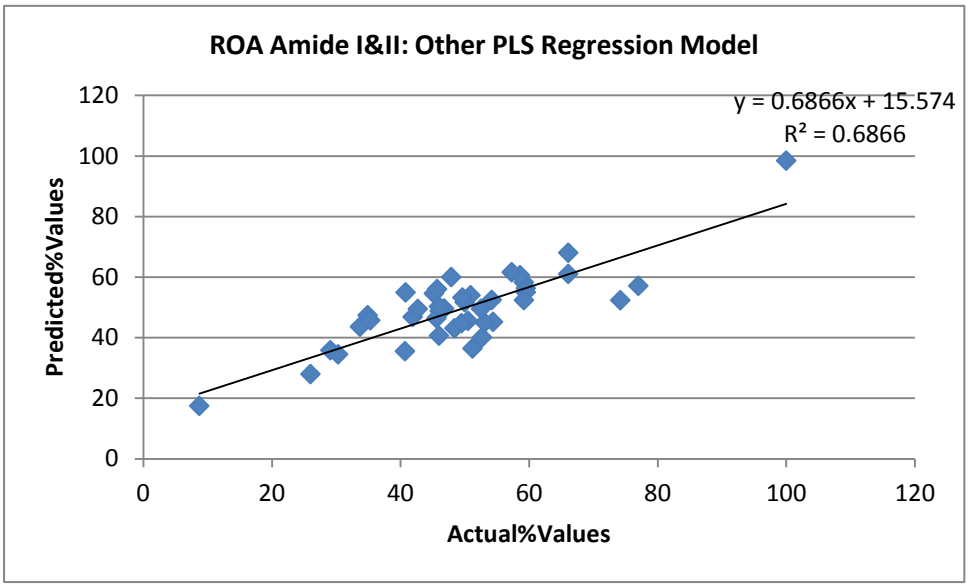

Graph of ROA PLS Regression Amide I&II α-Helix Model using Bin 10 $cm^{-1}$



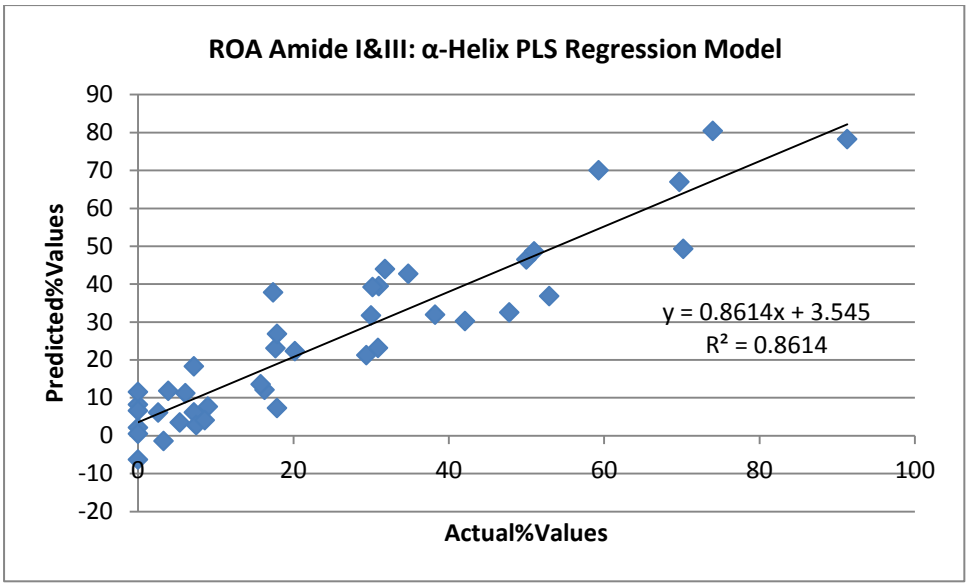
Graph of ROA PLS Regression Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



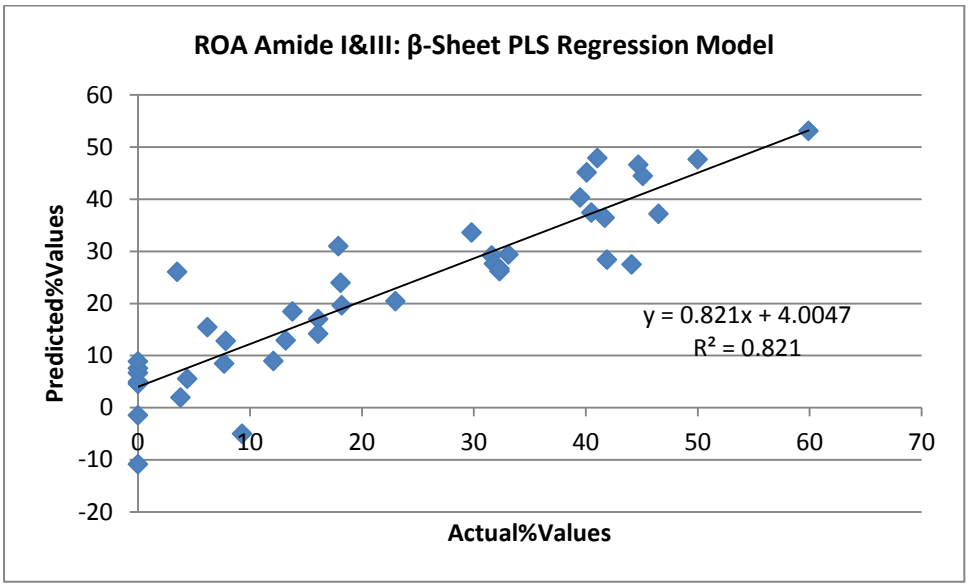

Graph of ROA PLS Regression Amide I&II Other Model using Bin 10 $cm^{-1}$



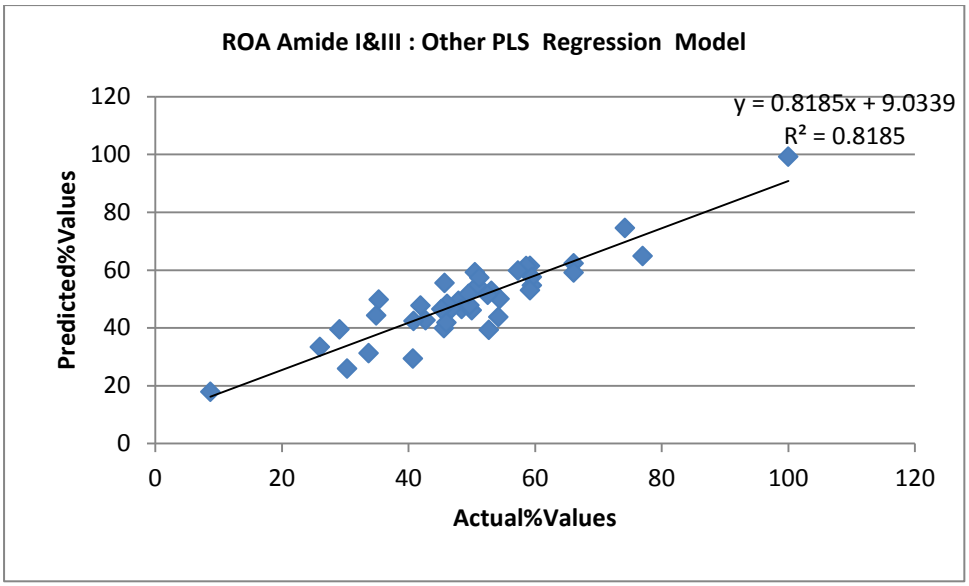
Graph of ROA PLS Regression Amide I&III α-Helix Model using Bin 10 $cm^{-1}$



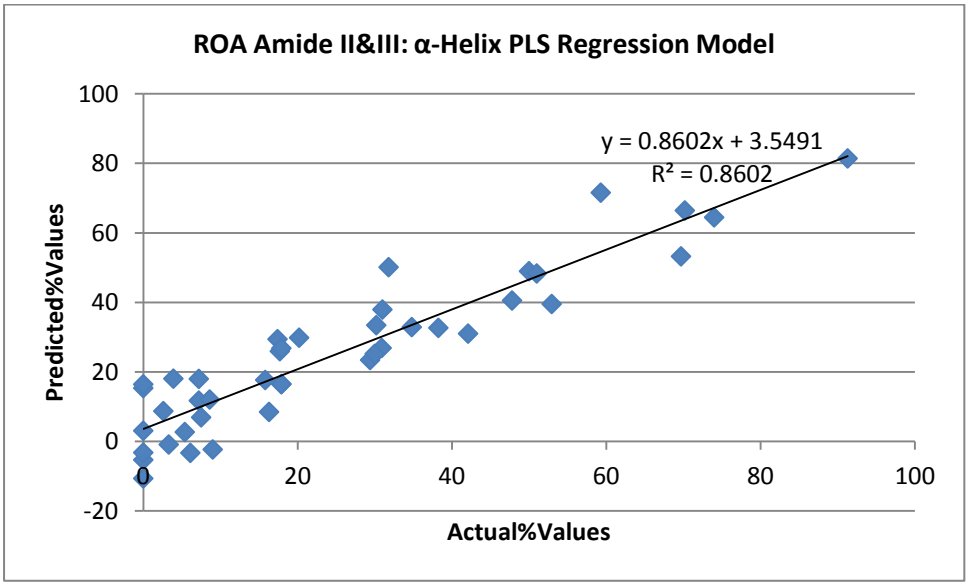

Graph of ROA PLS Regression Amide I&III β-Sheet Model using Bin 10 $cm^{-1}$



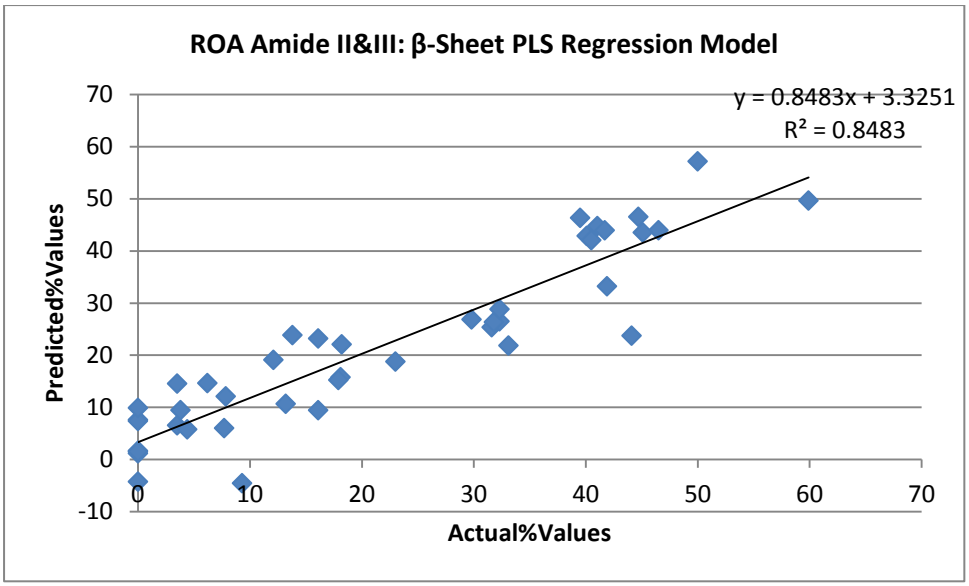
Graph of ROA PLS Regression Amide I&III Other Model using Bin 10 $cm^{-1}$



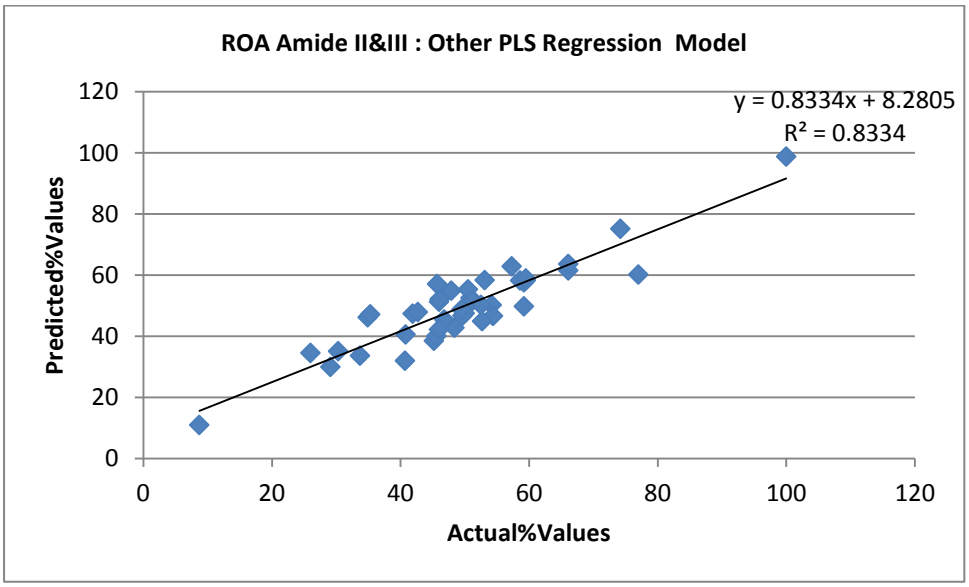

Graph of ROA PLS Regression Amide II&III α-Helix Model using Bin 10 $cm^{-1}$



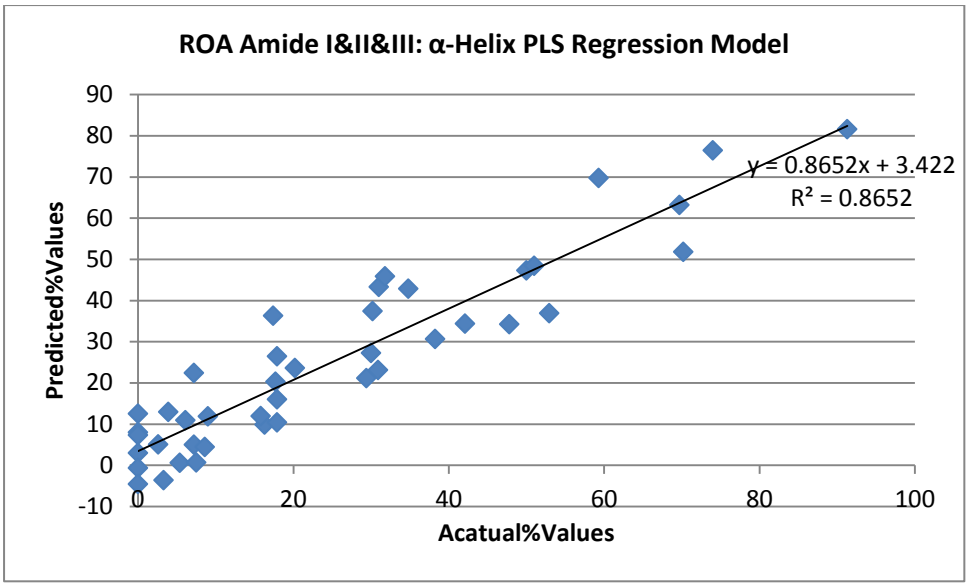
Graph of ROA PLS Regression Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



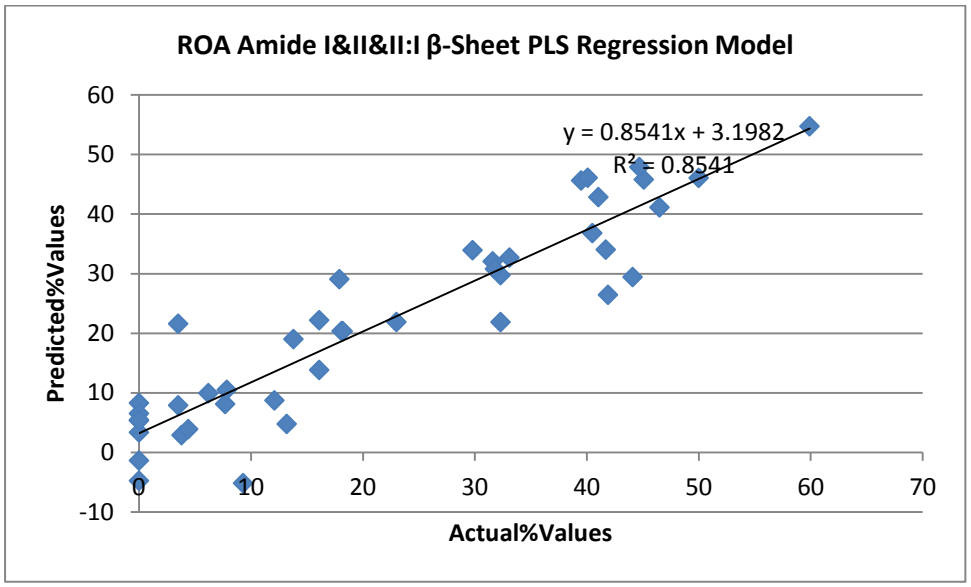

Graph of ROA PLS Regression Amide II&III Other Model using Bin 10 $cm^{-1}$



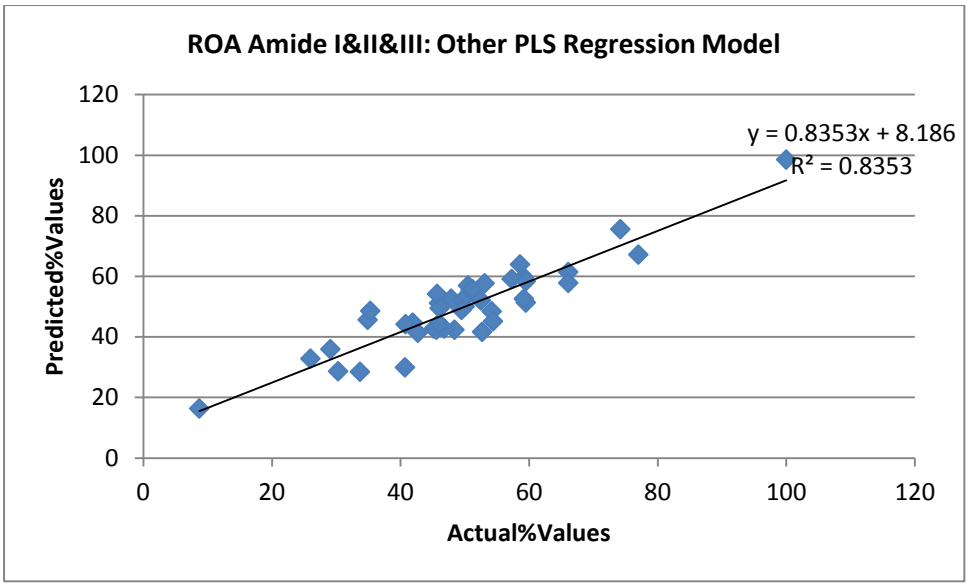
Graph of ROA PLS Regression Amide I&II&III α-Helix Model using Bin 10 $cm^{-1}$



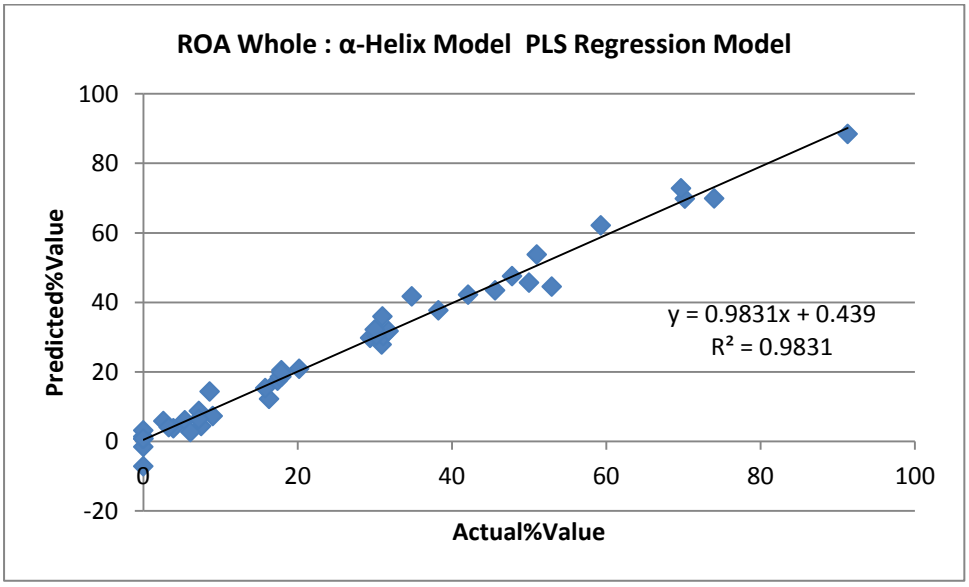

Graph of ROA PLS Regression Amide I&II&III β-Sheet Model using Bin 10 $cm^{-1}$



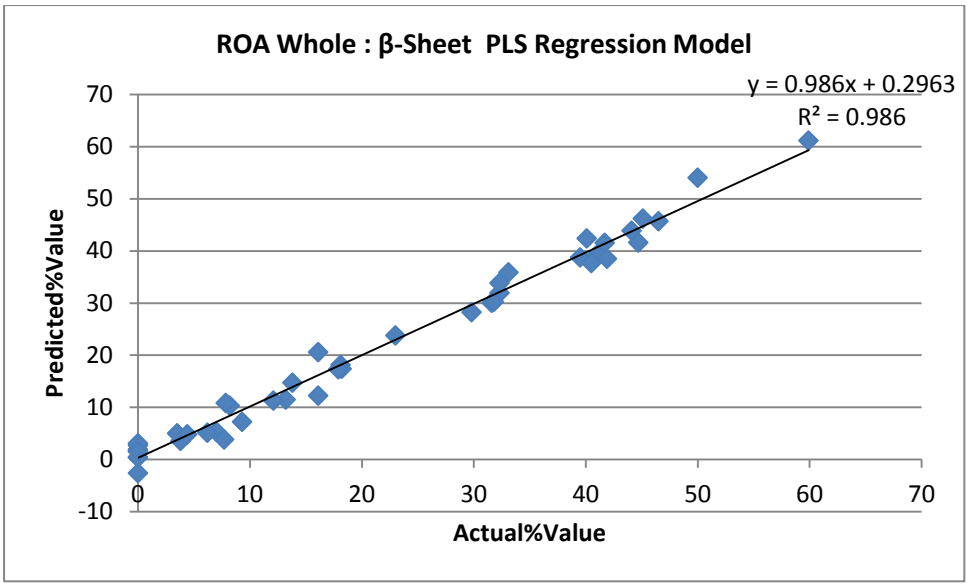
Graph of ROA PLS Regression Amide I&II&III Other Model using Bin 10 cm $^{-1}$



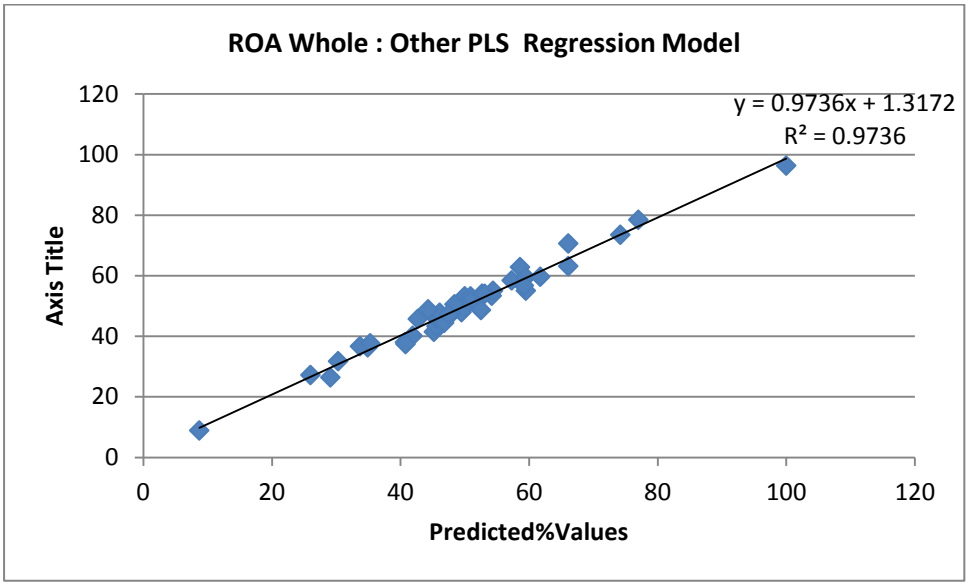

Graph of ROA PLS Regression Full Spectrum α-Helix Model using Bin 10 cm $^{-1}$



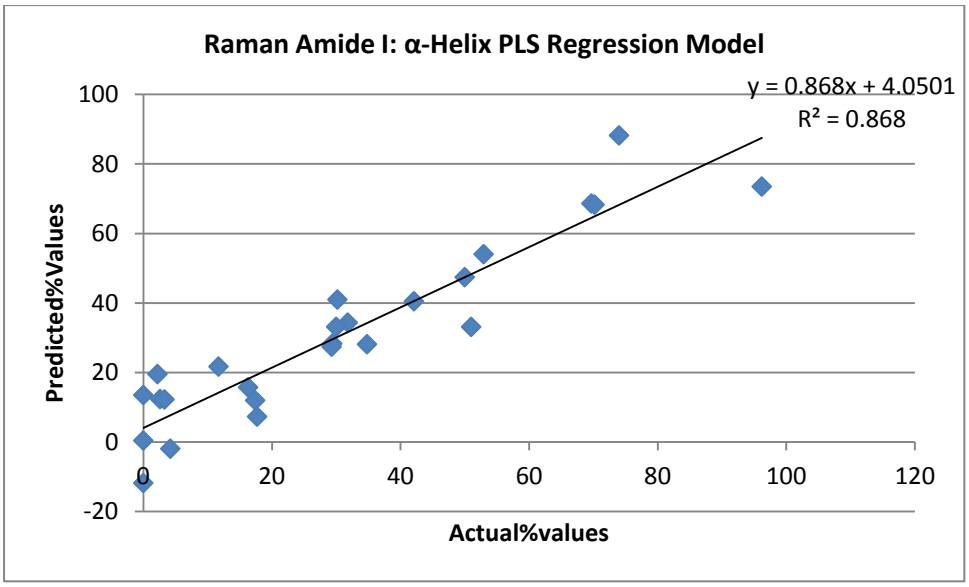
Graph of ROA PLS Regression Full Spectrum β-Sheet Model using Bin 10 cm $^{-1}$



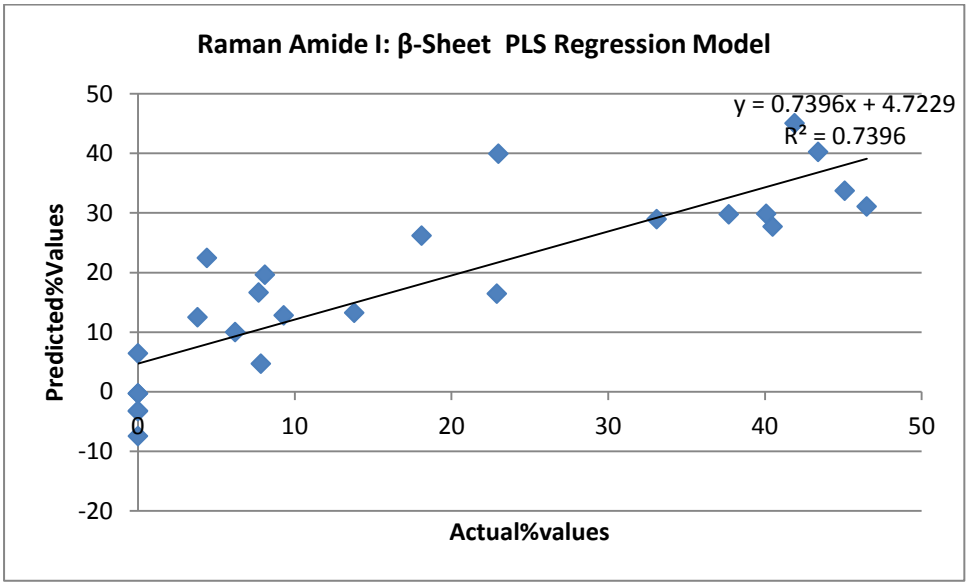

Graph of ROA PLS Regression Full Spectrum Other Model using Bin 10 cm $^{-1}$



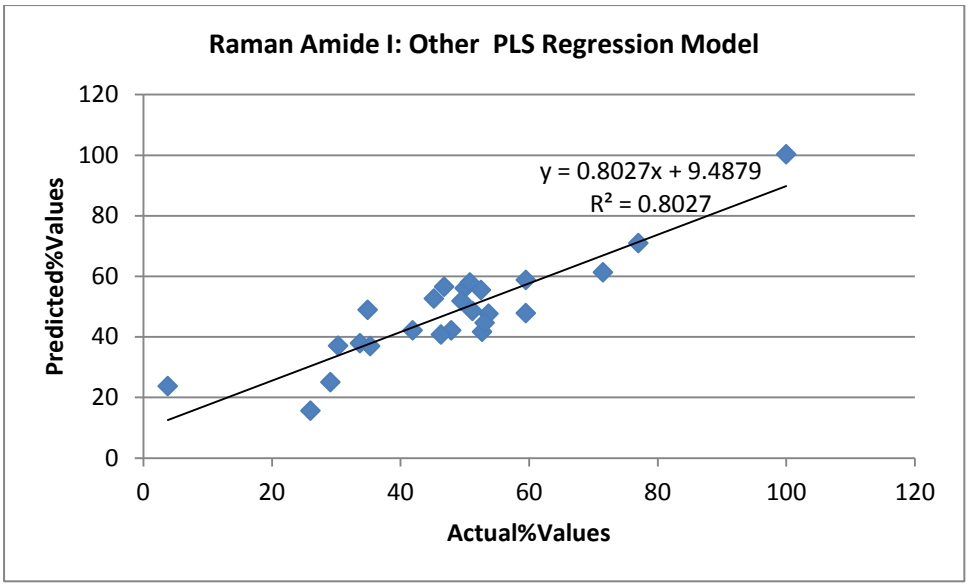
Graph of Raman PLS Regression Amide I α-Helix Model using Bin 10 $cm^{-1}$



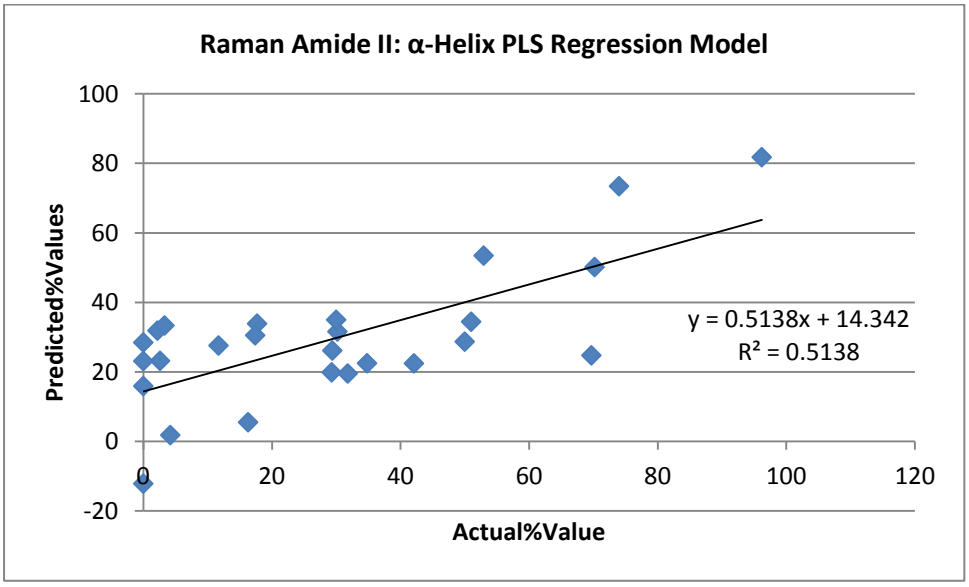

Graph of Raman PLS Regression Amide I β-Sheet Model using Bin 10 $cm^{-1}$



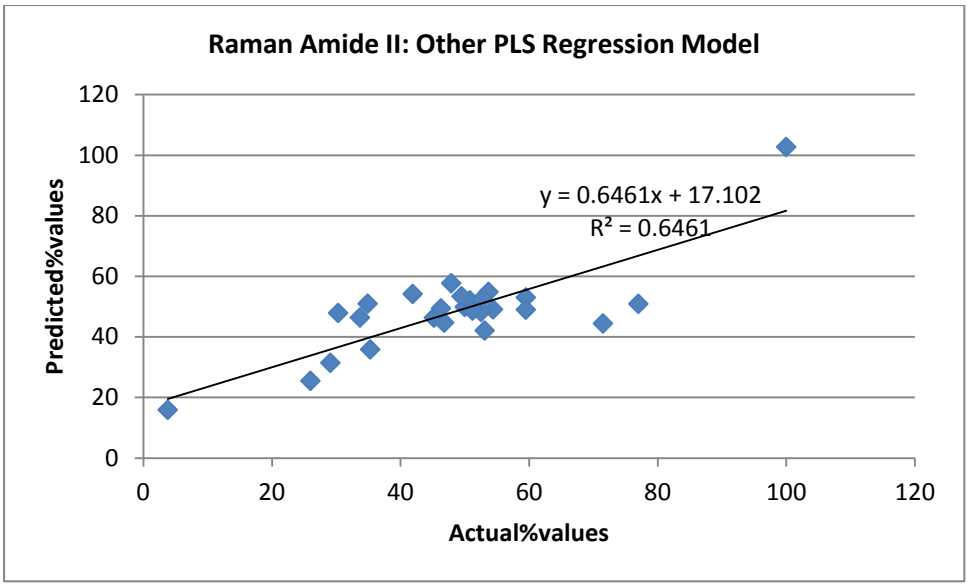
Graph of Raman PLS Regression Amide I Other Model using Bin 10 $cm^{-1}$



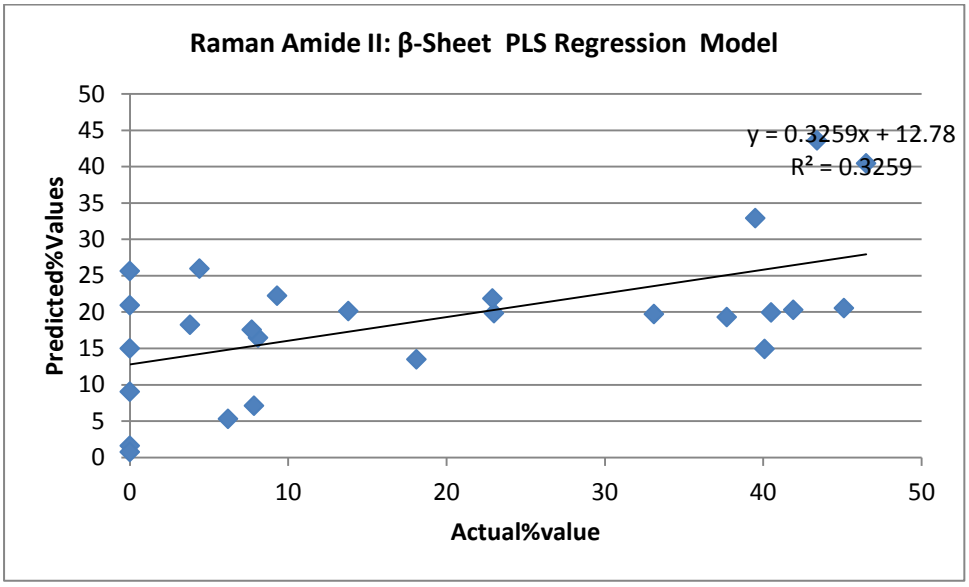

Graph of Raman PLS Regression Amide II α-Helix Model using Bin 10 $cm^{-1}$



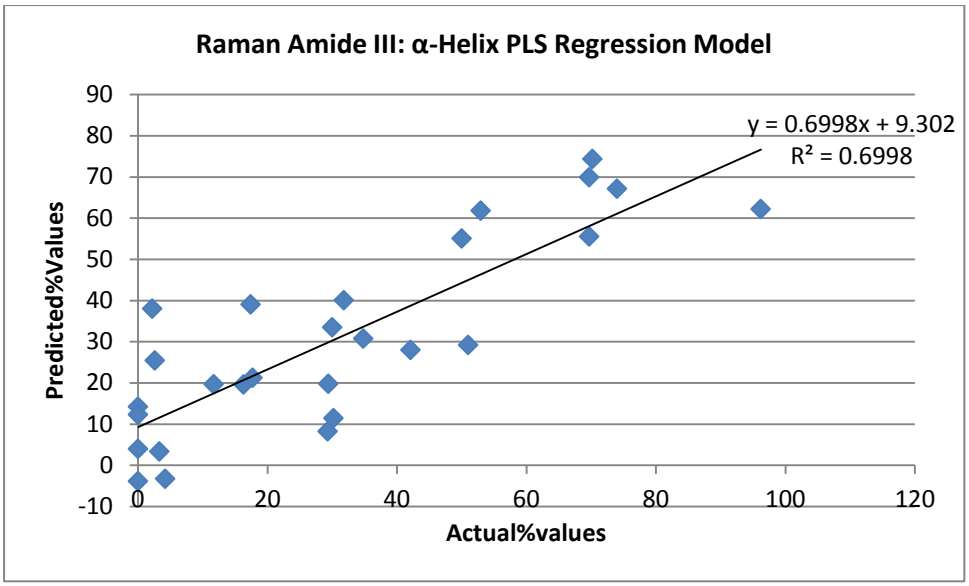
Graph of Raman PLS Regression Amide II Other Model using Bin 10 $cm^{-1}$



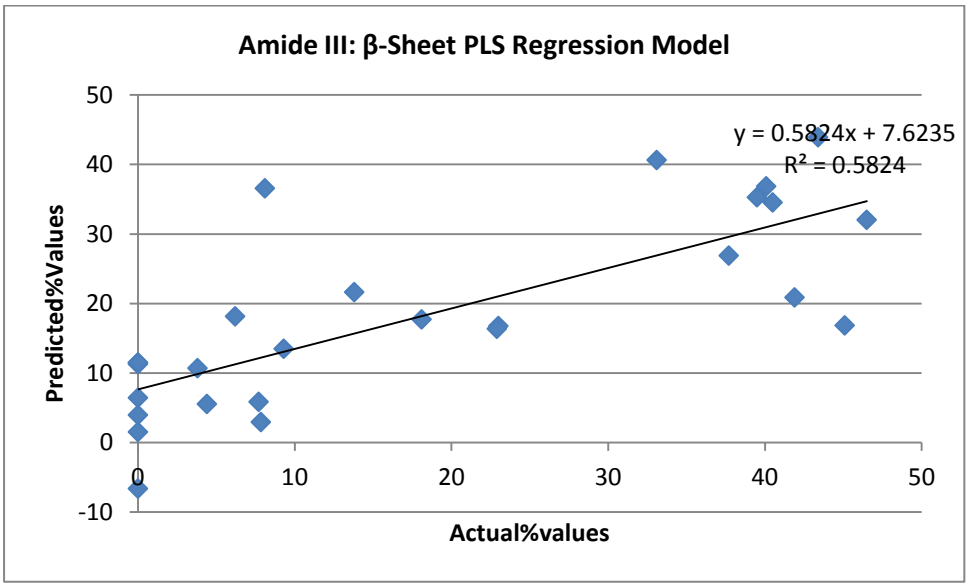

Graph of Raman PLS Regression Amide II β-Sheet Model using Bin 10 $cm^{-1}$



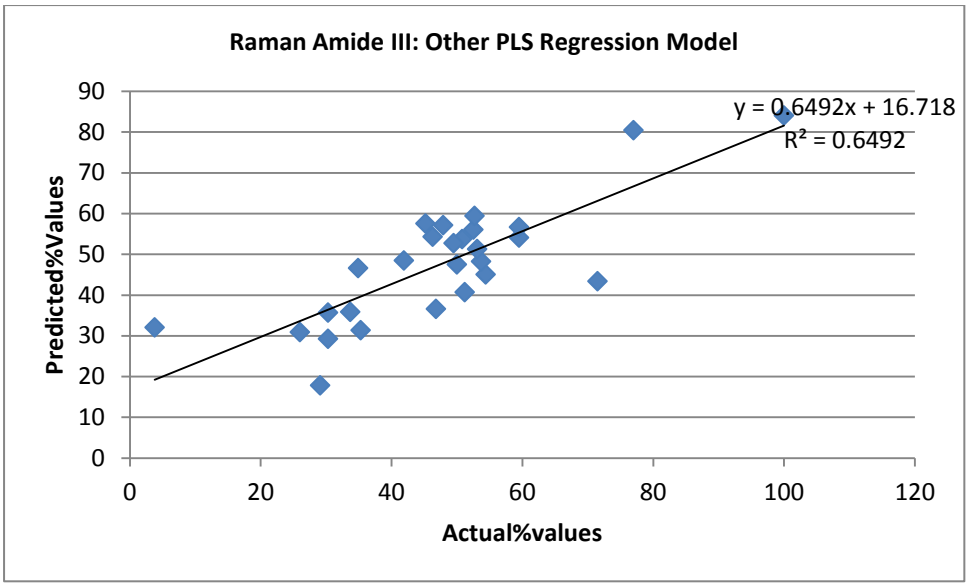
Graph of Raman PLS Regression Amide III α-Helix Model using Bin 10 $cm^{-1}$



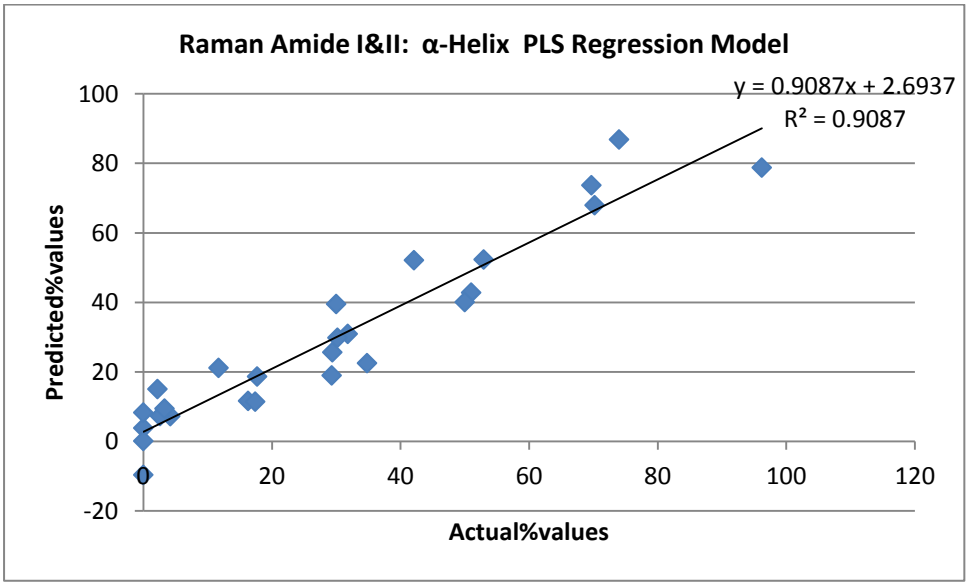

Graph of Raman PLS Regression Amide III β-sheet Model using Bin 10 $cm^{-1}$



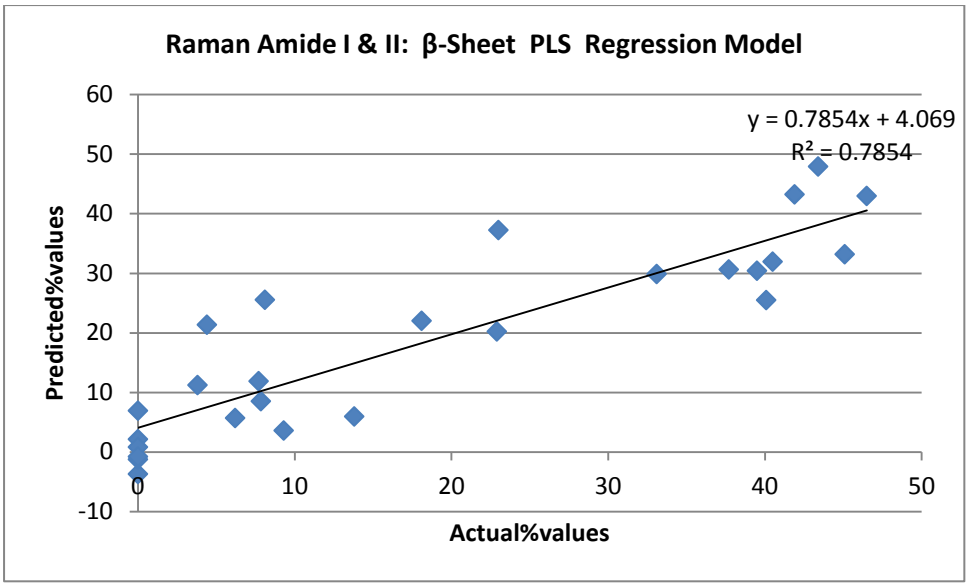
Graph of Raman PLS Regression Amide III Other Model using Bin 10 cm $^{-1}$



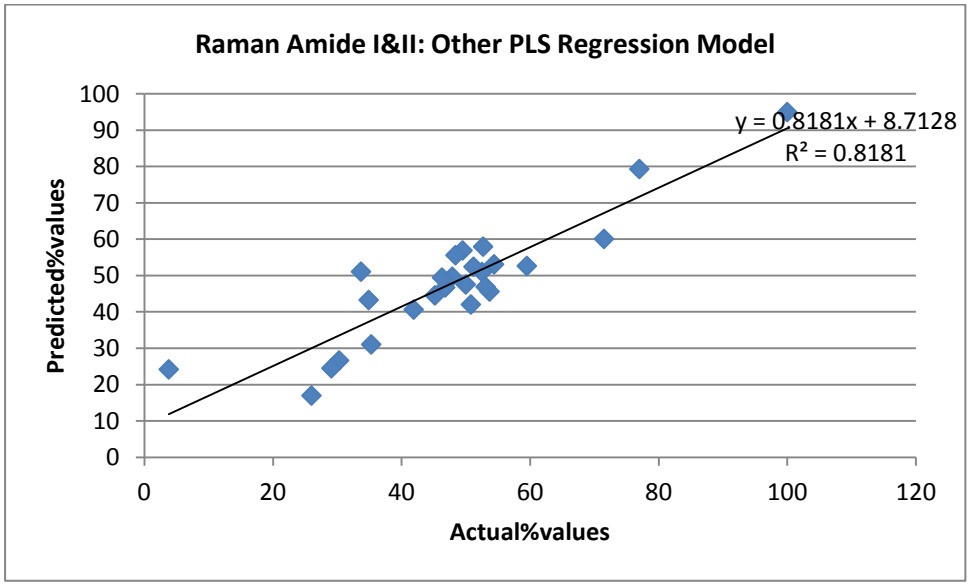

Graph of Raman PLS Regression Amide I&II α-Helix Model using Bin 10 cm $^{-1}$



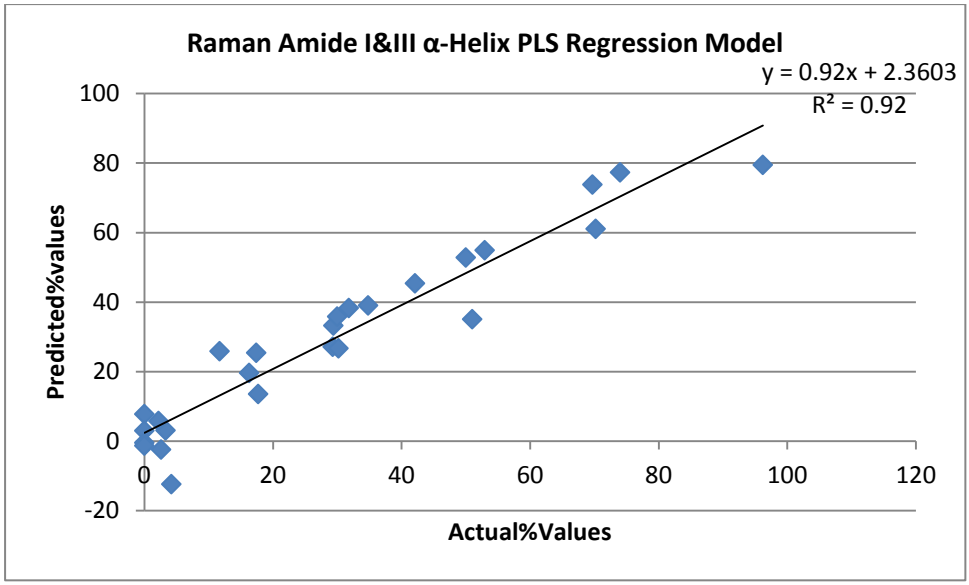
Graph of Raman PLS Regression Amide I&II β-Sheet Model using Bin 10 $cm^{-1}$



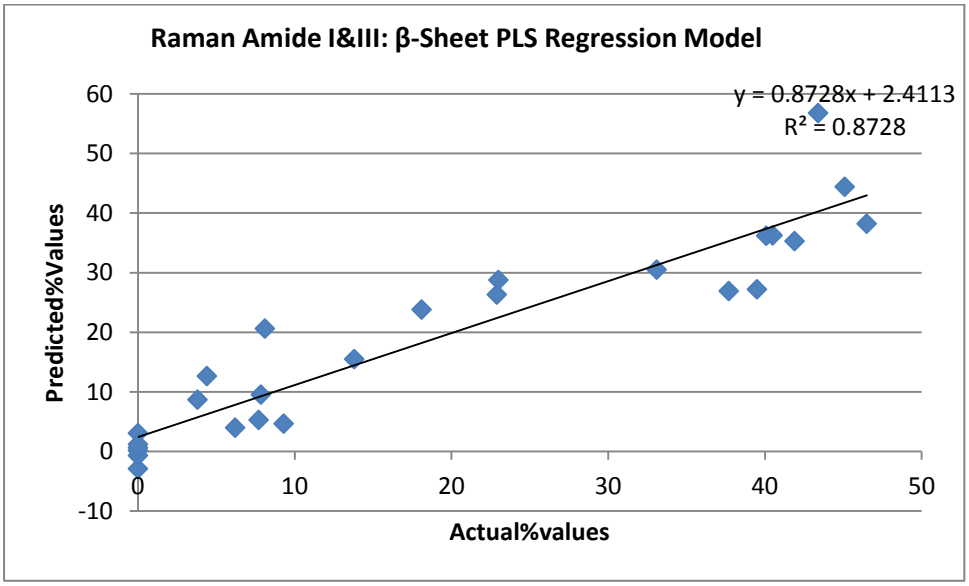

Graph of Raman PLS Regression Amide I&II Other Model using Bin 10 $cm^{-1}$



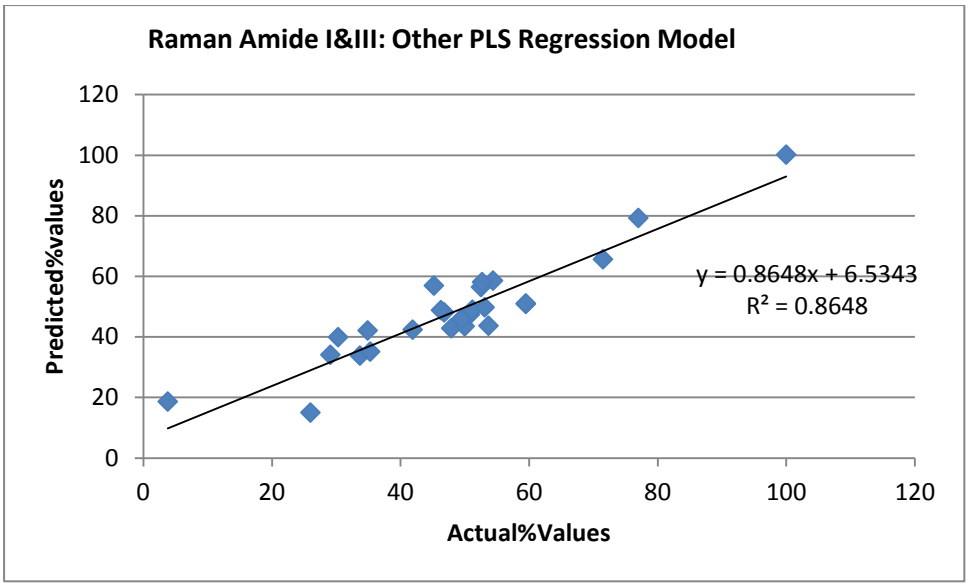
Graph of Raman PLS Regression Amide I&III α-Helix Model using Bin 10 cm $^{-1}$



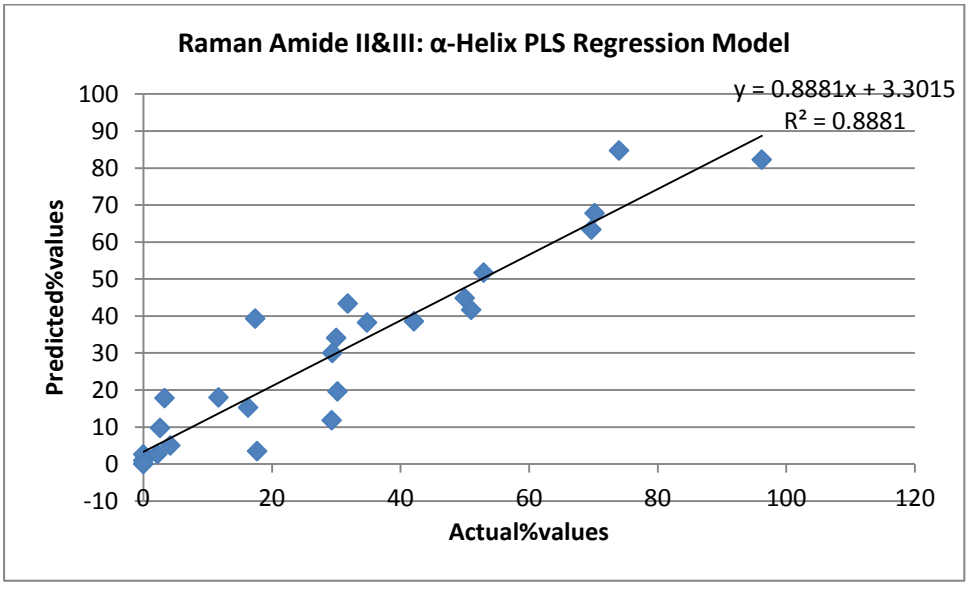

Graph of Raman PLS Regression Amide I&III β-Sheet Model using Bin 10 cm $^{-1}$



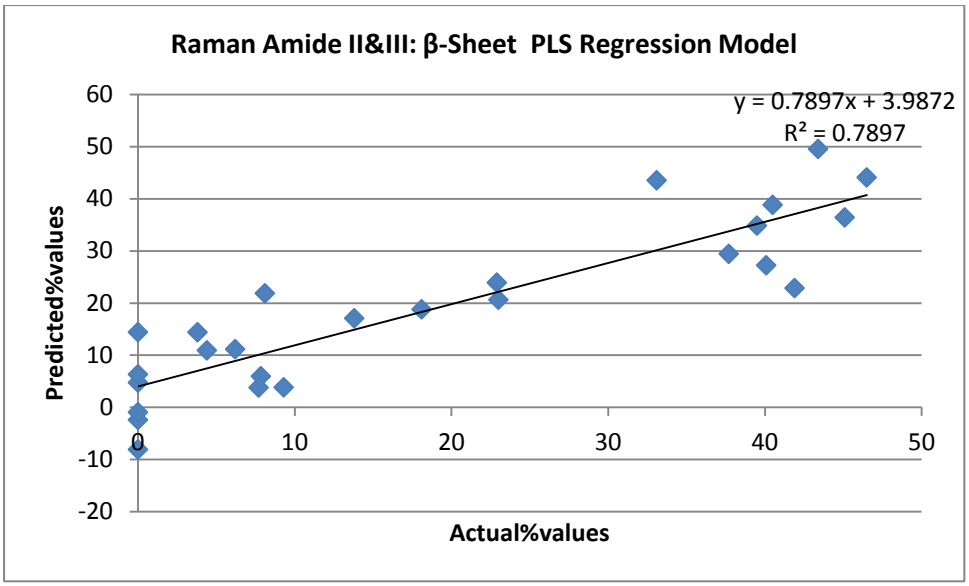
Graph of Raman PLS Regression Amide I&III Other Model using Bin 10 cm $^{-1}$



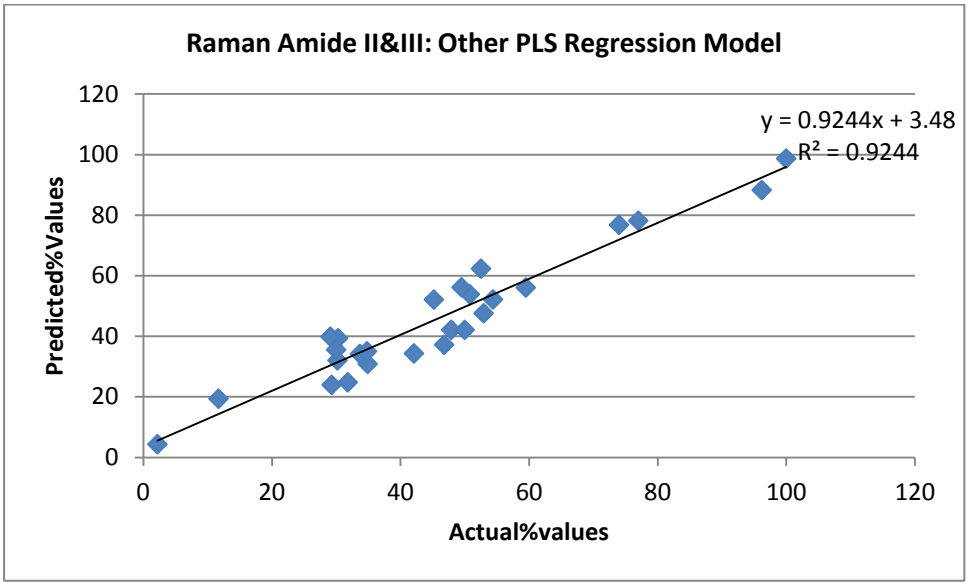

Graph of Raman PLS Regression Amide II&III α-Helix Model using Bin 10 cm $^{-1}$



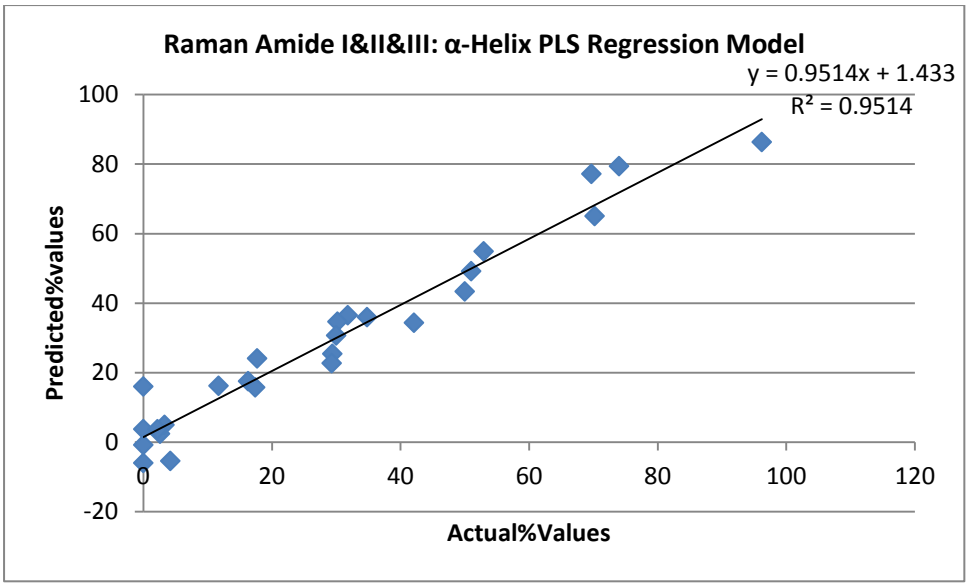
Graph of Raman PLS Regression Amide II&III β-Sheet Model using Bin 10 $cm^{-1}$



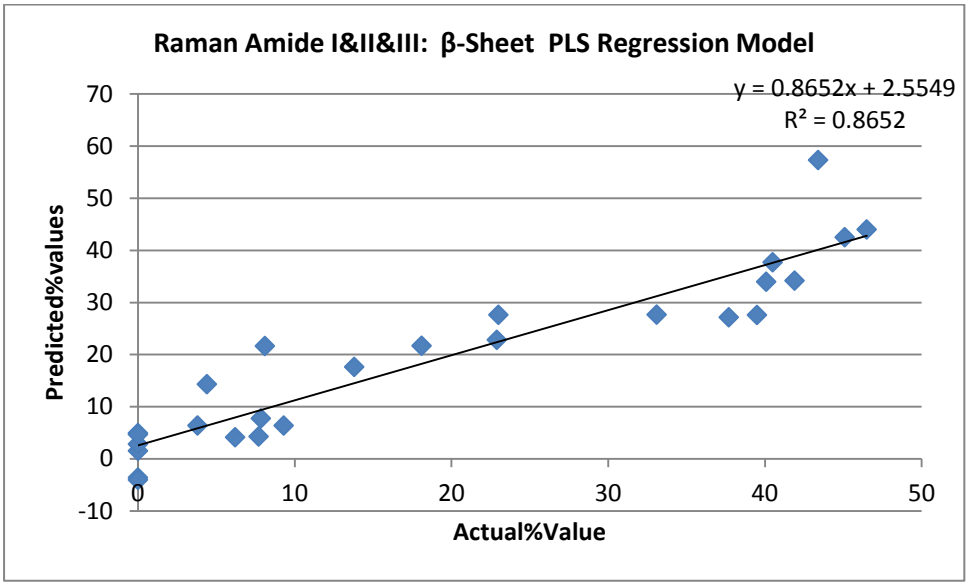

Graph of Raman PLS Regression Amide II&III Other Model using Bin 10 $cm^{-1}$



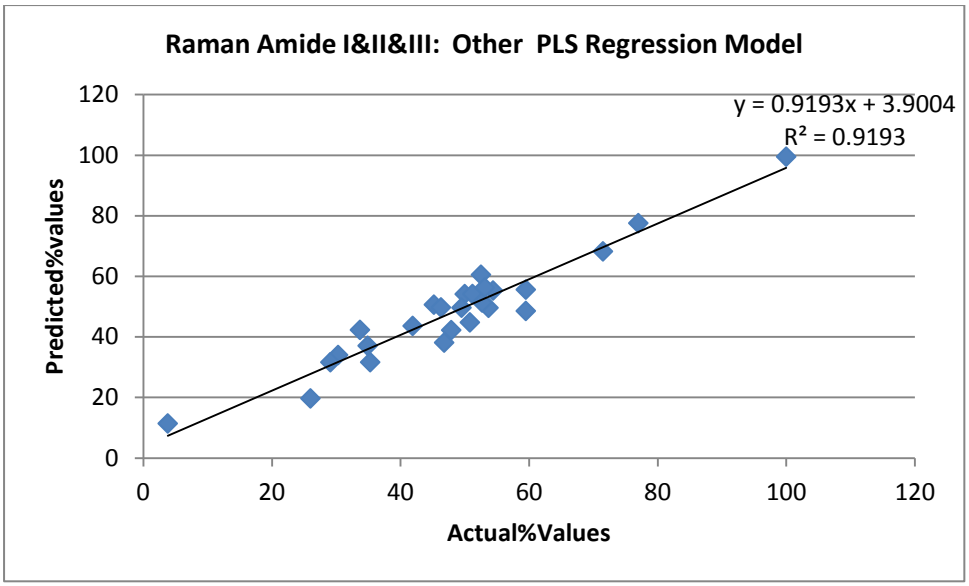
Graph of Raman PLS Regression Amide I&II&III α-Helix Model using Bin 10 cm $^{-1}$



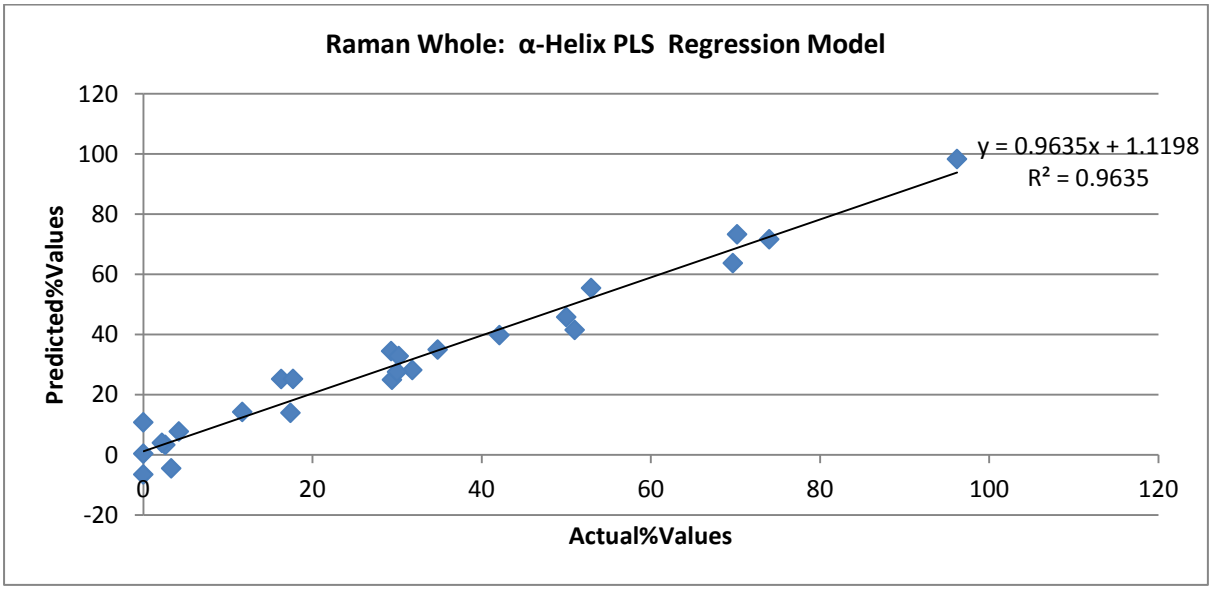

Graph of Raman PLS Regression Amide I&II&III β-Sheet Model using Bin 10 cm $^{-1}$



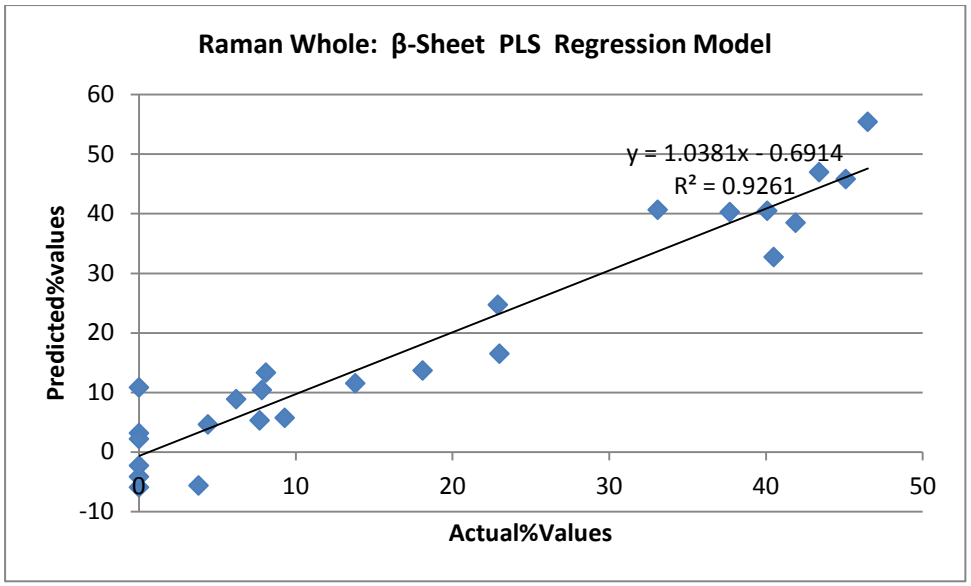
Graph of Raman PLS Regression Amide I&II&III Other Model using Bin 10 cm $^{-1}$



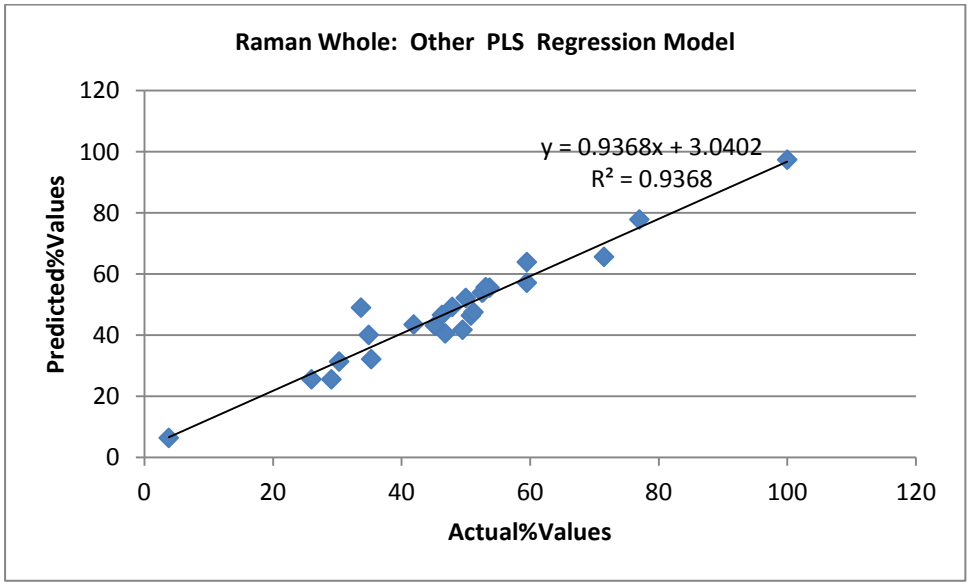

Graph of Raman PLS Regression Full Spectrum α-Helix Model using Bin 10 cm $^{-1}$



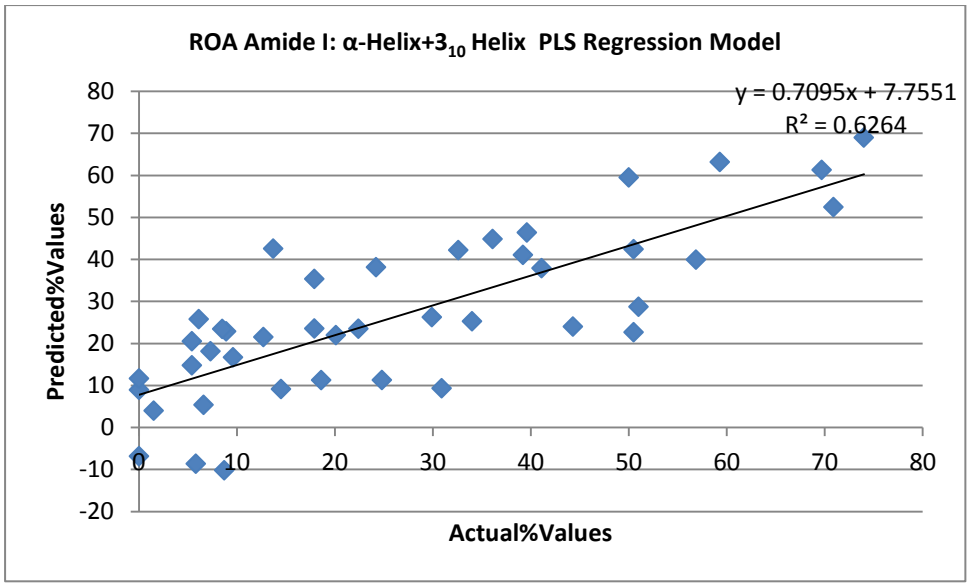
Graph of Raman PLS Regression Full Spectrum β-Sheet Model using Bin 10 cm $^{-1}$



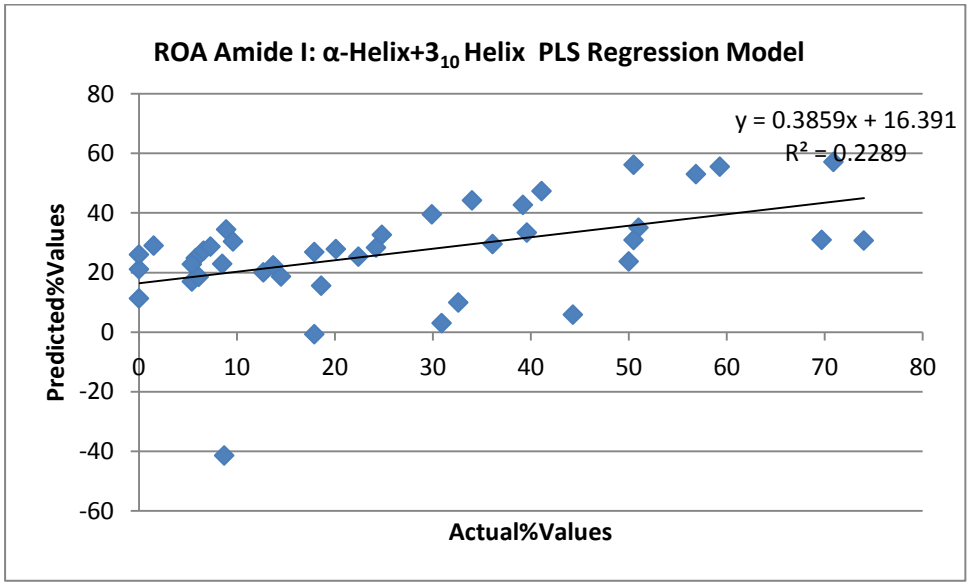

Graph of Raman PLS Regression Full Spectrum Other Model using Bin 10 cm $^{-1}$



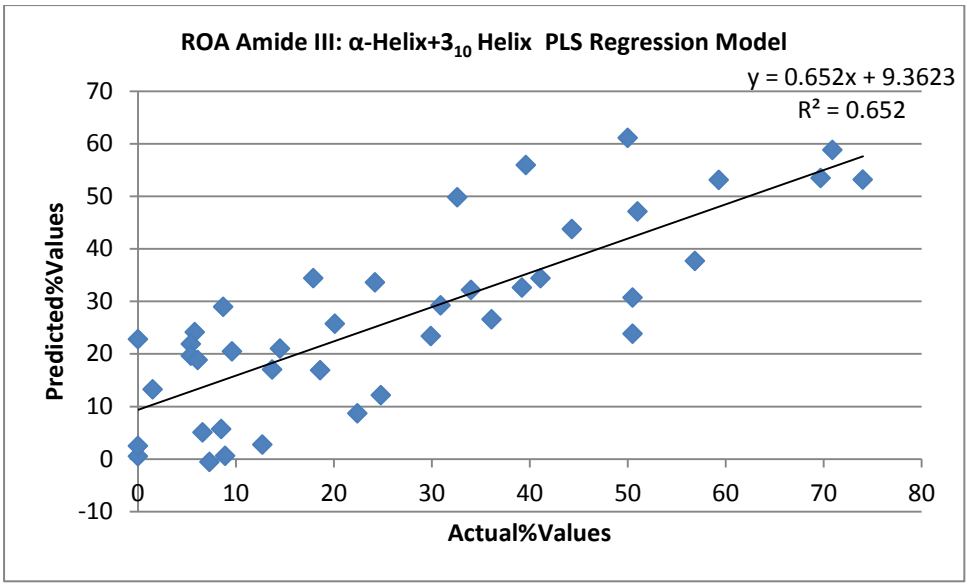
Graph of ROA PLS Regression Amide I α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



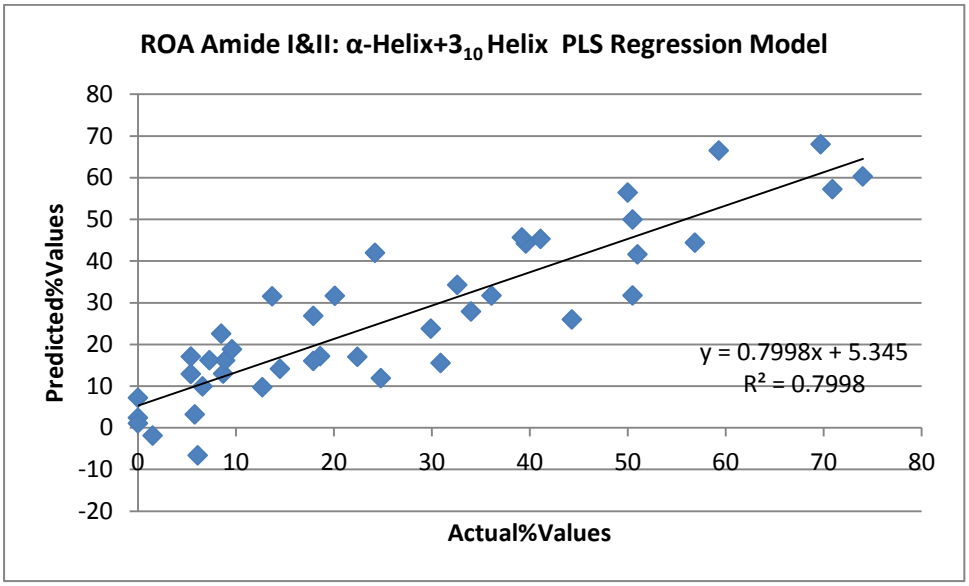

Graph of ROA PLS Regression Amide II α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



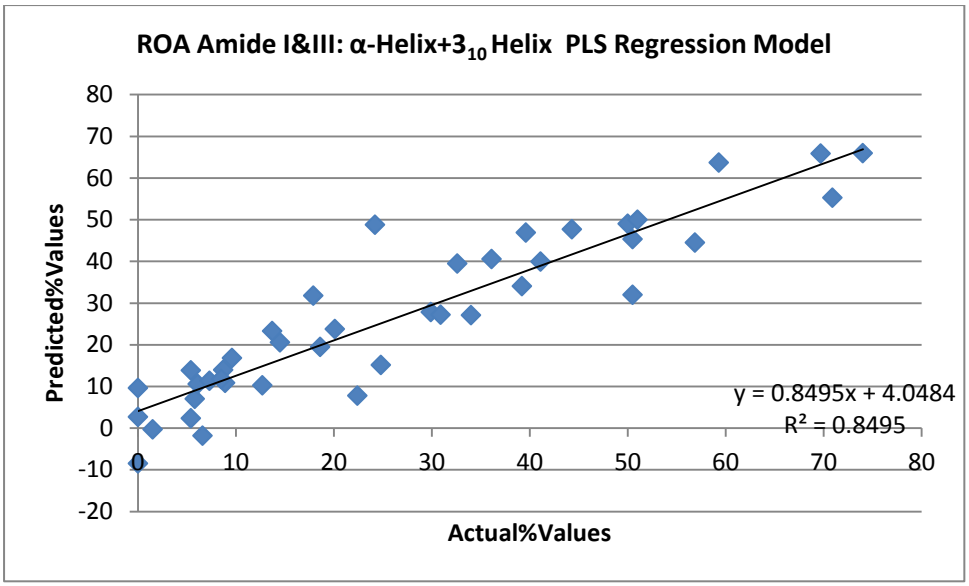
Graph of ROA PLS Regression Amide III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



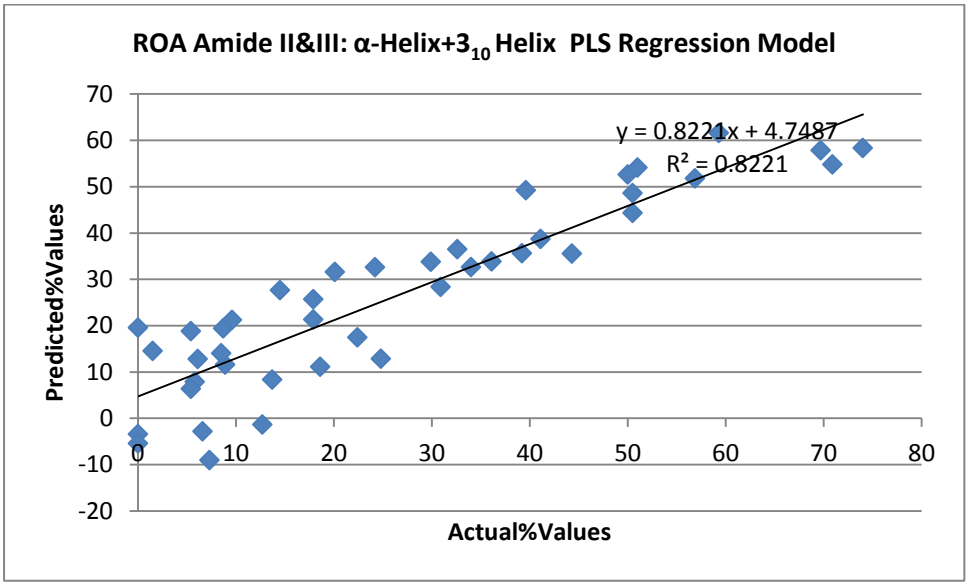

Graph of ROA PLS Regression Amide I&II α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



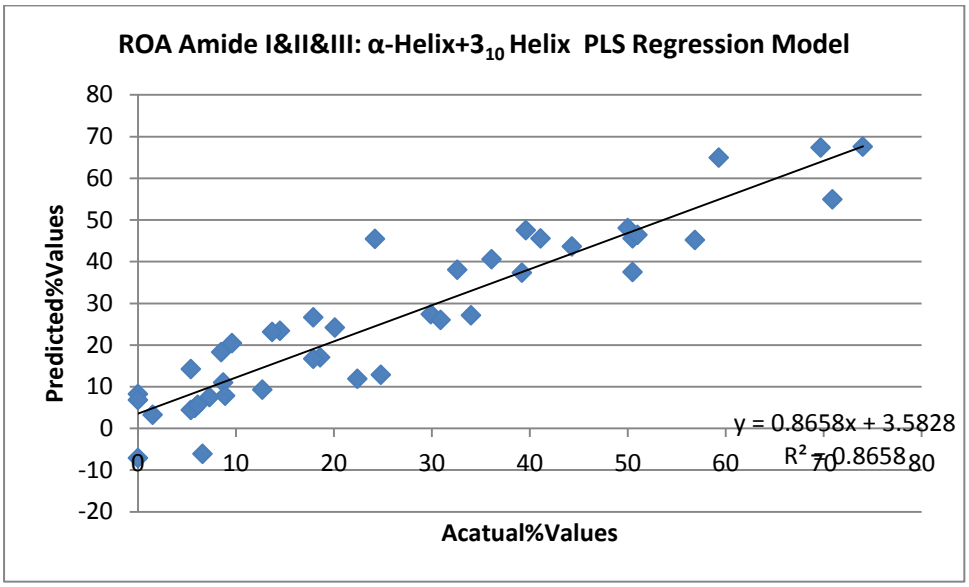
Graph of ROA PLS Regression Amide I&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



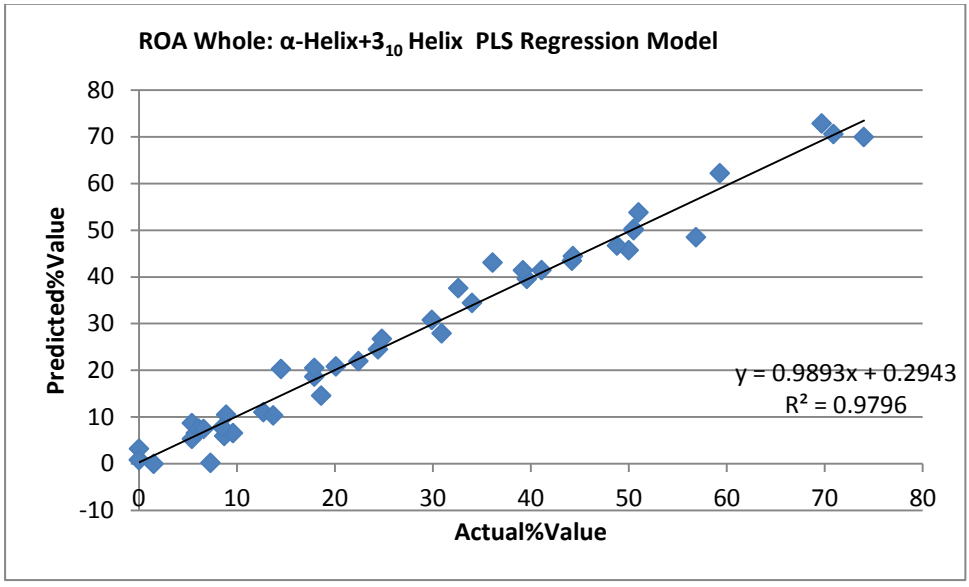

Graph of ROA PLS Regression Amide II&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



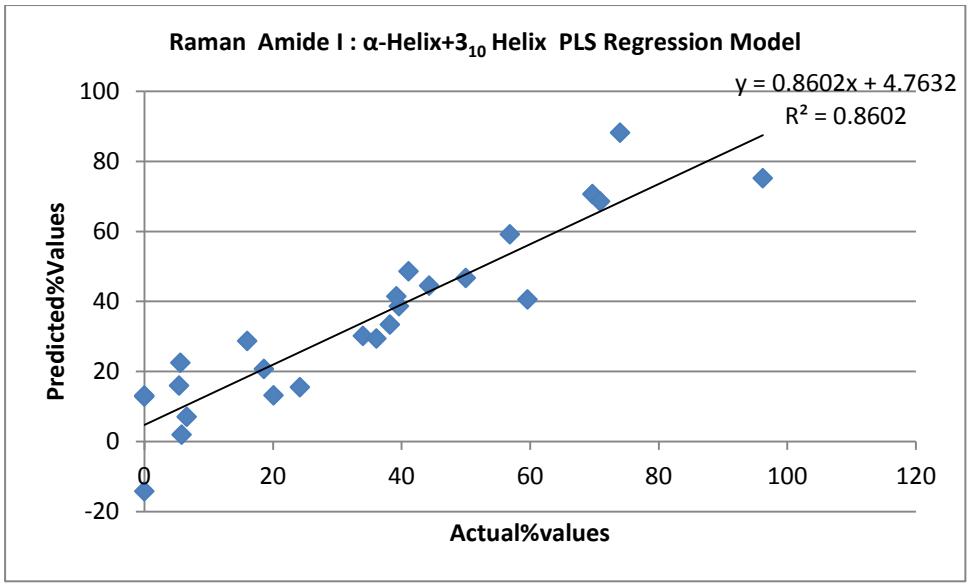
Graph of ROA PLS Regression Amide I&II&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



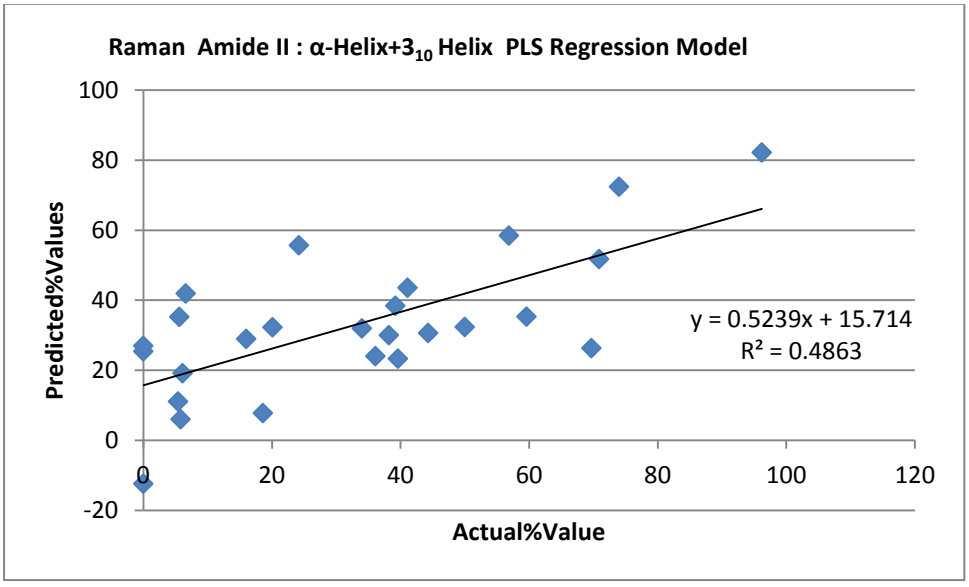

Graph of ROA PLS Regression Full Spectrum α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



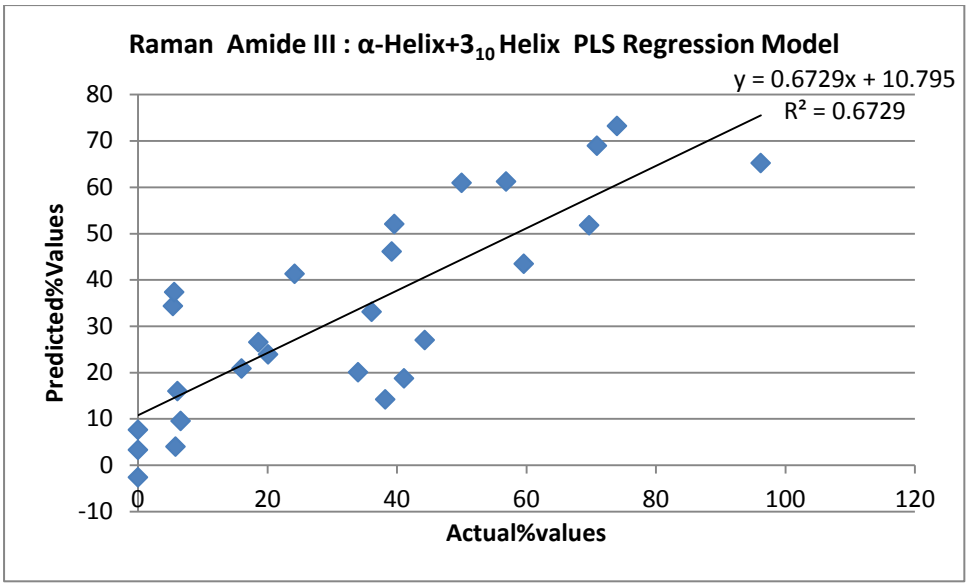
Graph of Raman PLS Regression Amide I α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



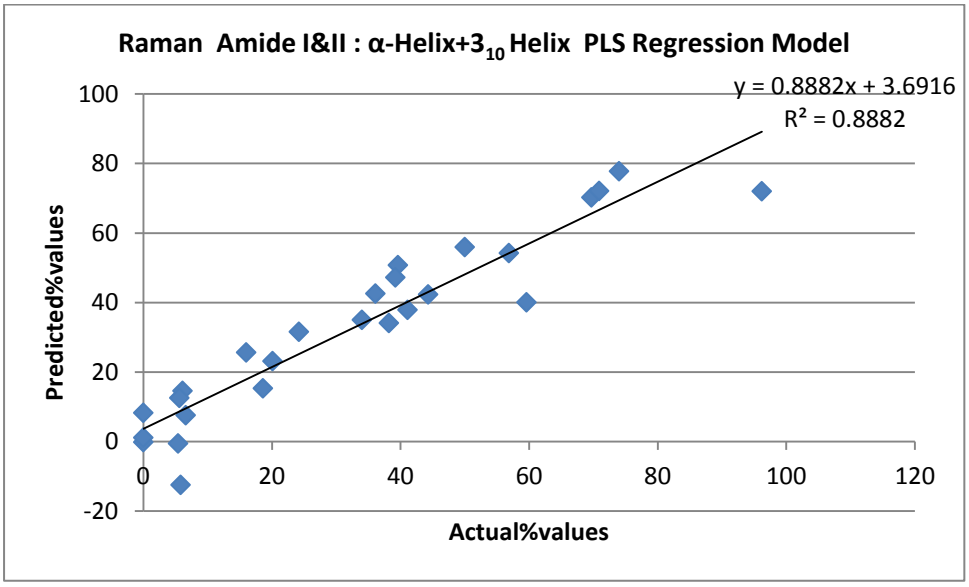

Graph of Raman PLS Regression Amide II α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



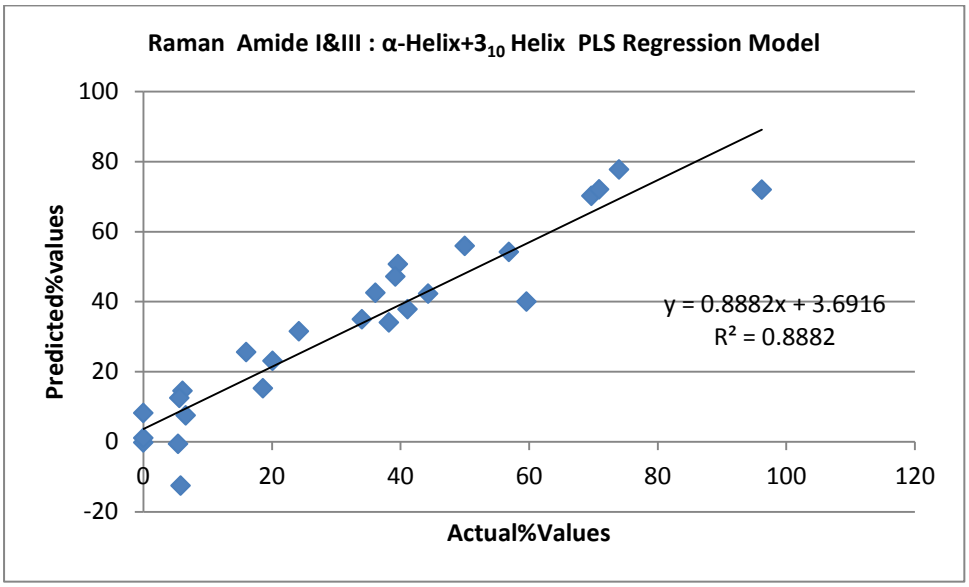
Graph of Raman PLS Regression Amide III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



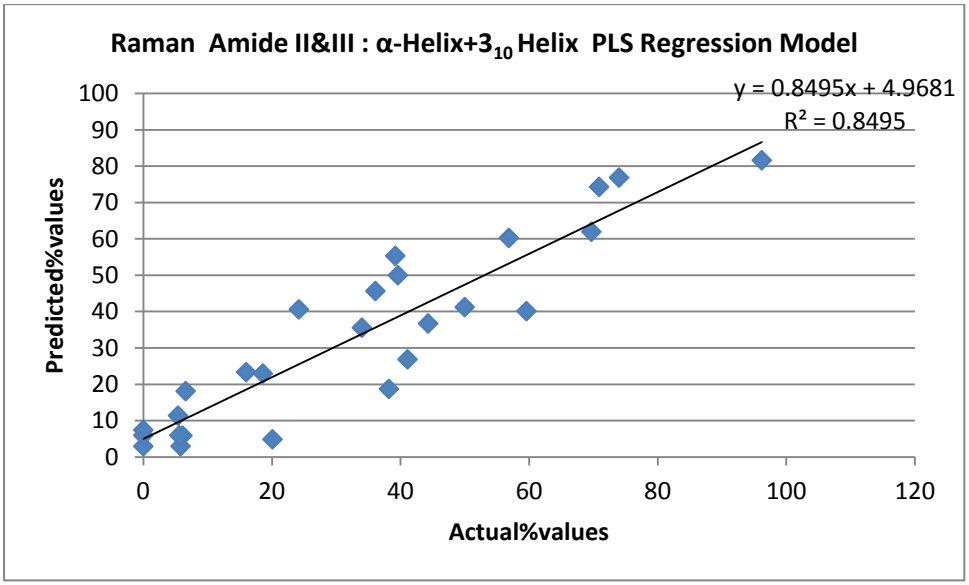

Graph of Raman PLS Regression Amide I&II α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



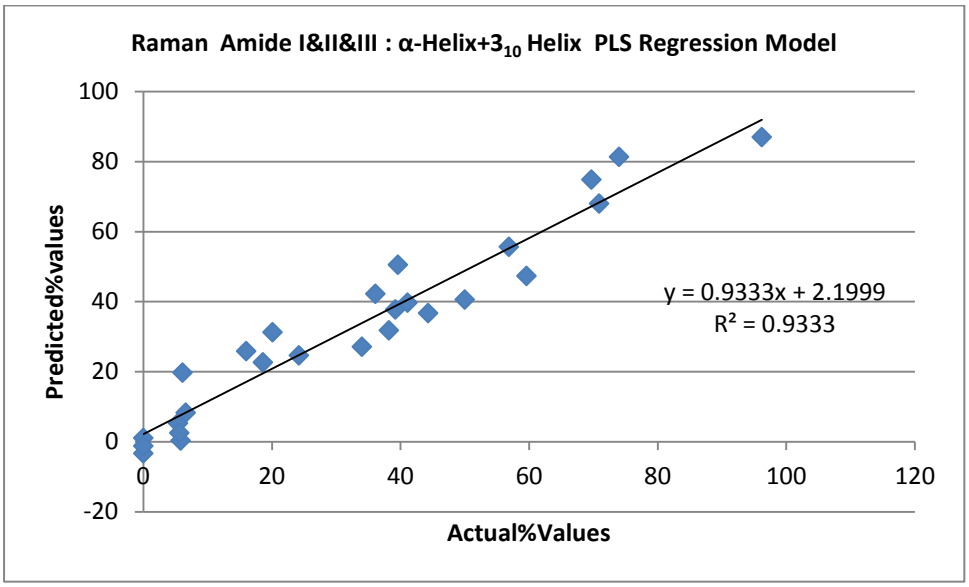
Graph of Raman PLS Regression Amide I&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



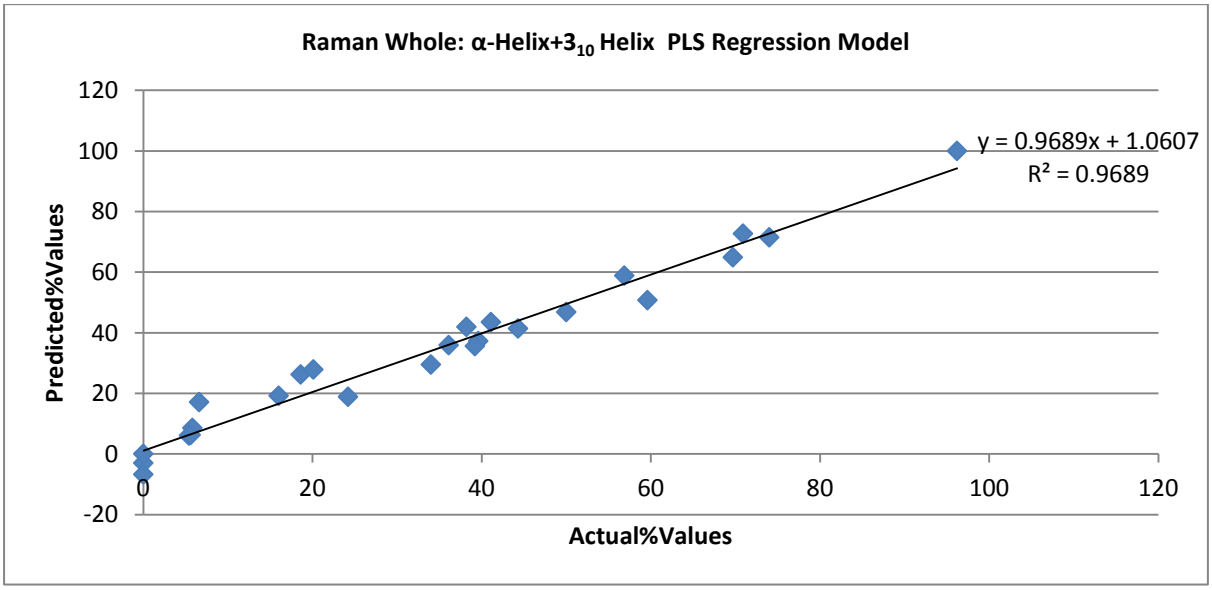

Graph of Raman PLS Regression Amide II&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



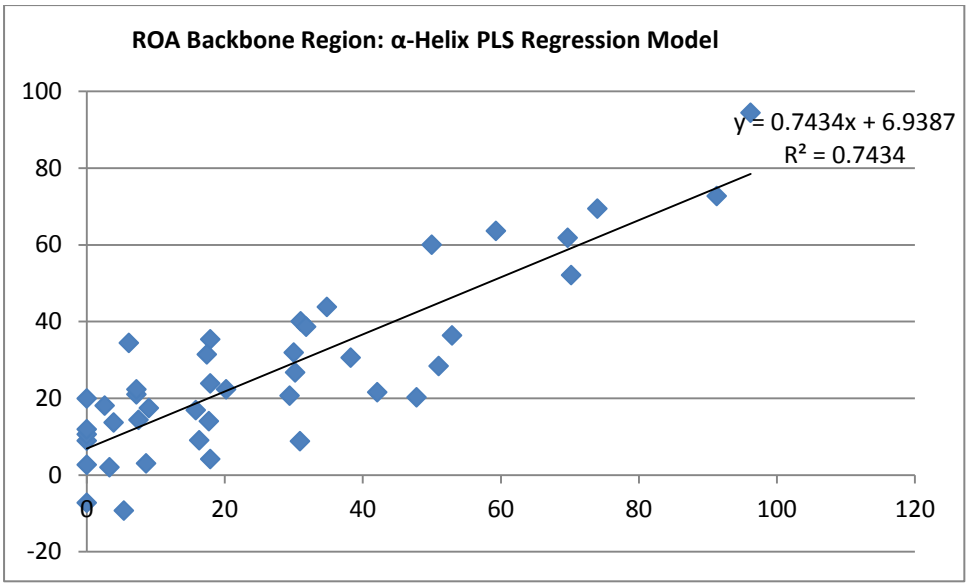
Graph of Raman PLS Regression Amide I&II&III α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



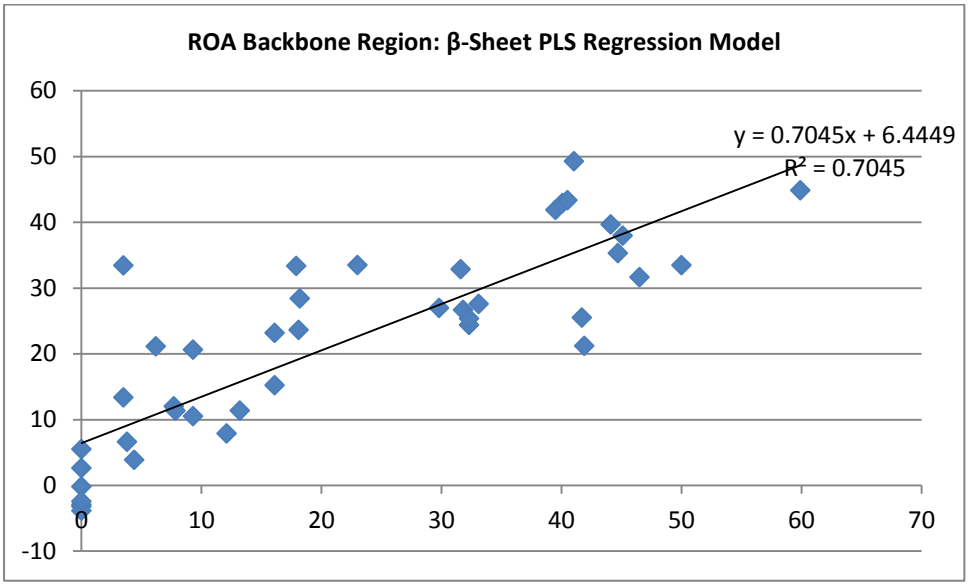

Graph of Raman PLS Regression Full Spectrum α-Helix+$3_{10}$ Helix Model using Bin 10 $cm^{-1}$



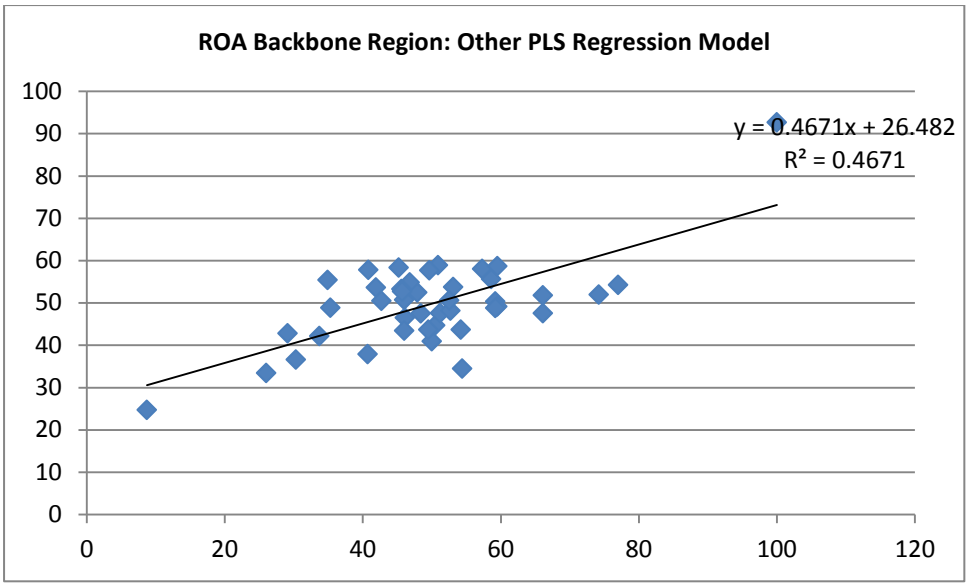
Graph of ROA PLS Regression Backbone Region α-Helix Model using Bin 10 $cm^{-1}$



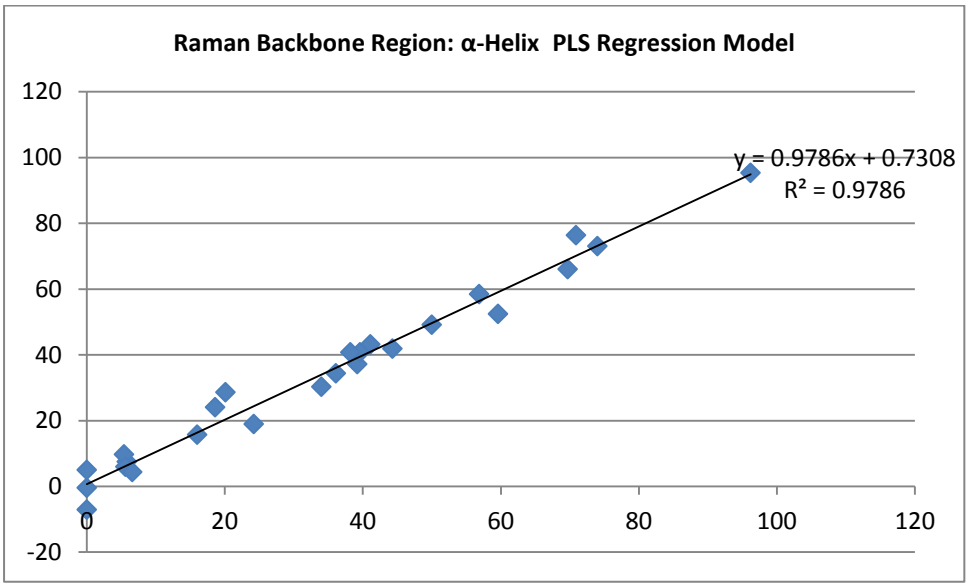

Graph of ROA PLS Regression Backbone Region β-Sheet Model using Bin 10 $cm^{-1}$



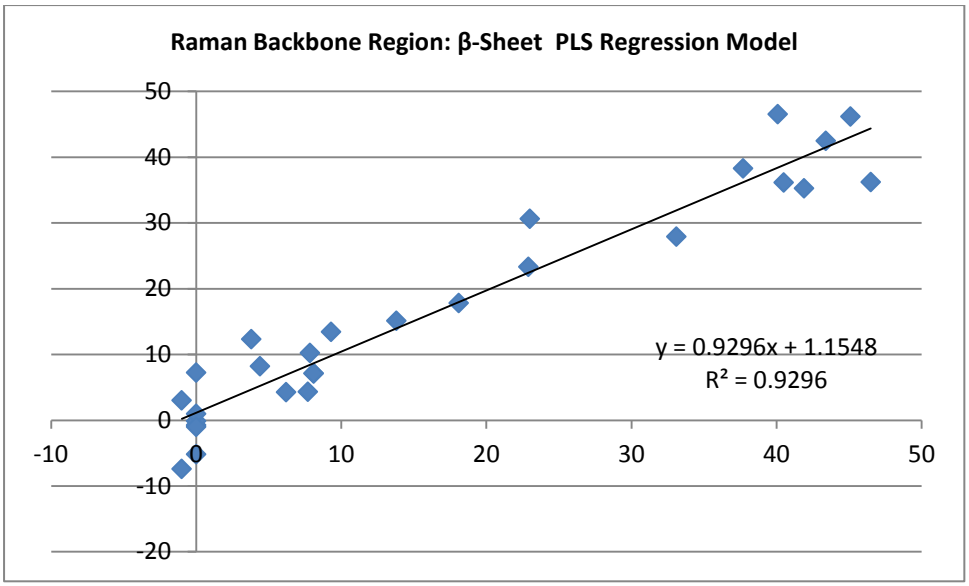
Graph of ROA PLS Regression Backbone Region Other Model using Bin 10 cm $^{-1}$



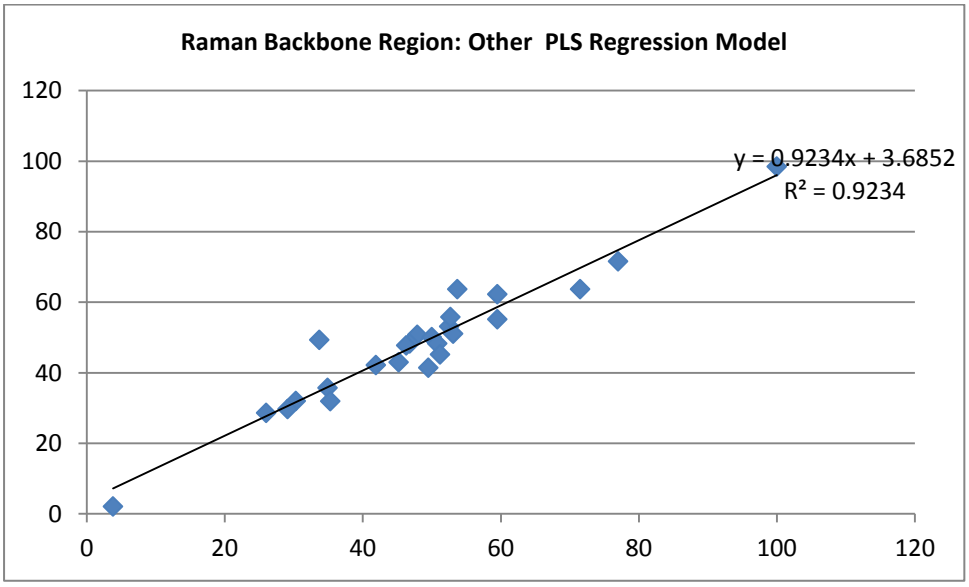

Graph of Raman PLS Regression Backbone Region α-Helix Model using Bin 10 cm $^{-1}$



Graph of Raman PLS Regression Backbone Region β-Sheet Model using Bin 10 $cm^{-1}$

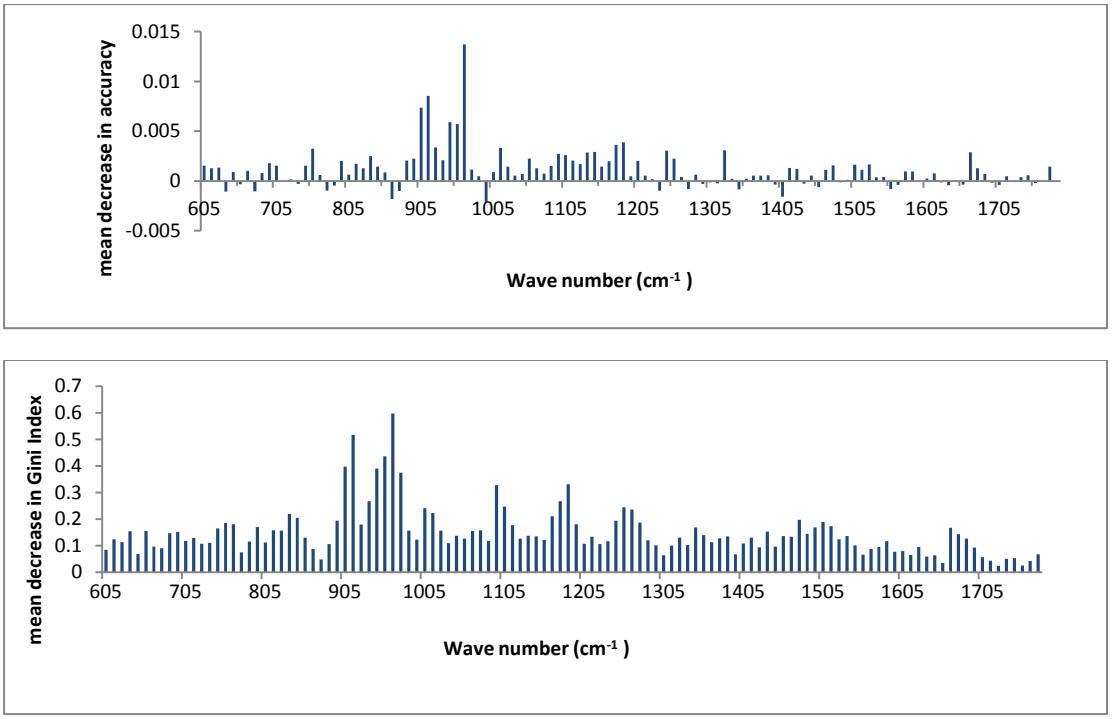



Graph of Raman PLS Regression Backbone Region Other Model using Bin 10 $cm^{-1}$



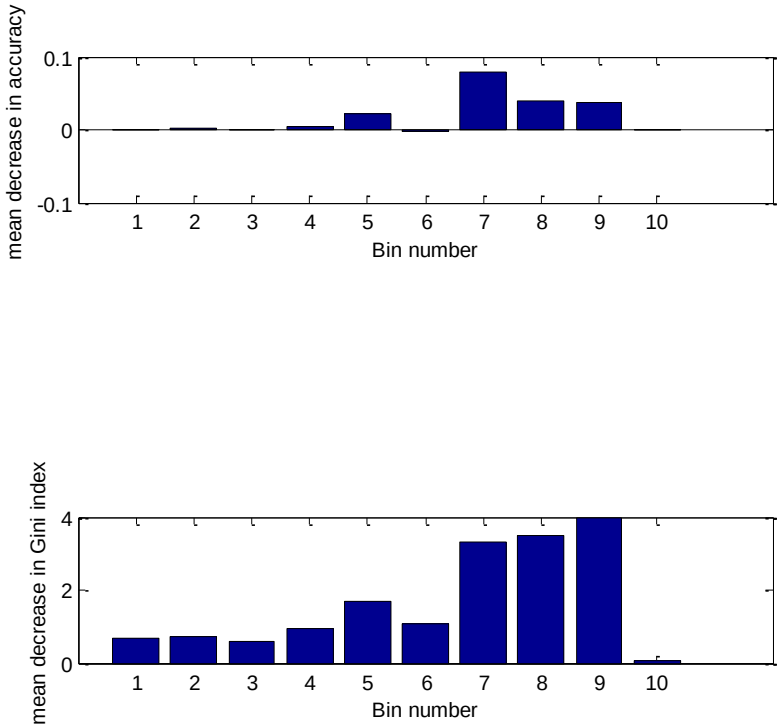
Appendix H- Bar graphs of variable importance and mean decrease in accuracy for the Random Forest analyses

Raman Whole Variable Importance (top) and Gini Index (bottom) plots for Spectral Subdivisions




Bins are 10 $cm^{-1}$ width from 605- 615 $cm^{-1}$ to 1765-1775 $cm^{-1}$

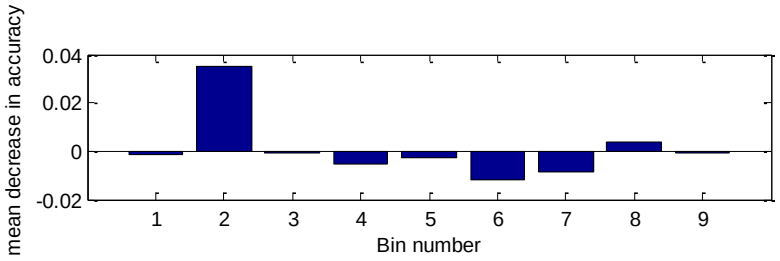
Raman Amide I variable importance plots: Mean decrease in Accuracy (top); Mean decrease in Gini index (bottom).

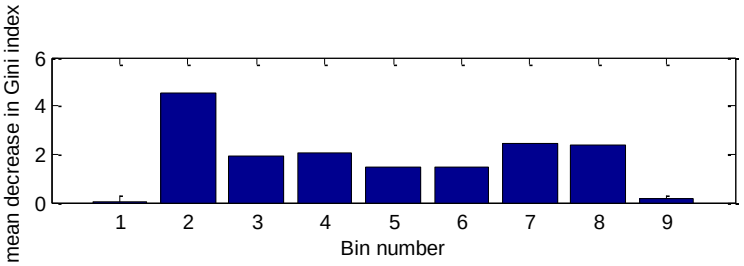



Bins are 10 $cm^{-1}$ width from 1605- 1615 $cm^{-1}$; (bin 1) to 1695-1705 $cm^{-1}$ (bin 10). Data from 1665 $cm^{-1}$ to 1695 $cm^{-1}$ is most important.

1-1605- 1615 $cm^{-1}$; 2-1615-1625$cm^{-1}$; 3-1625-1635$cm^{-1}$;4-1635-1645$cm^{-1}$; 5-1645-1655$cm^{-1}$;6-1655-1665$cm^{-1}$;7-1665-1675$cm^{-1}$;8-1675-1685$cm^{-1}$;9-1685-1695$cm^{-1}$;1695-1705$cm^{-1}$

Raman Amide II variable importance plots: Mean decrease in Accuracy (top); Mean decrease in Gini index (bottom).

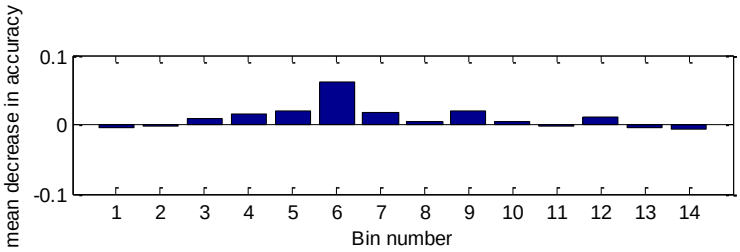


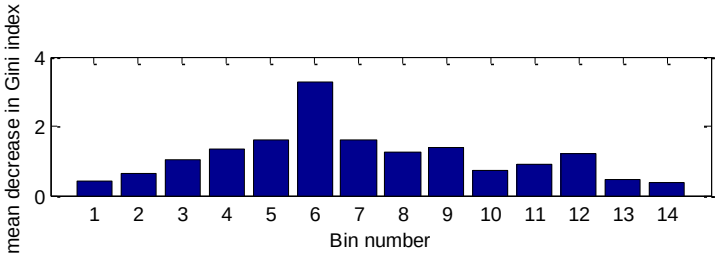




Bins are 10cm$^{-1}$ width from 1515- 1525 cm$^{-1}$; (bin 1) to 1595-1605 cm$^{-1}$ (bin 9). Data from 1525 cm$^{-1}$ to 1535 cm$^{-1}$ is most important.

1-1515- 1525 cm$^{-1}$; 2-1525-1535cm$^{-1}$; 3-1535-1545cm$^{-1}$;4-1545-1555cm$^{-1}$; 5-1555-1565cm$^{-1}$;6-1565-1575cm$^{-1}$;7-1575-1585cm$^{-1}$;8-1585-1595cm$^{-1}$;9-1595-1605cm$^{-1}$

Raman Amide III variable importance plots: Mean decrease in Accuracy (top); Mean decrease in Gini index (bottom).

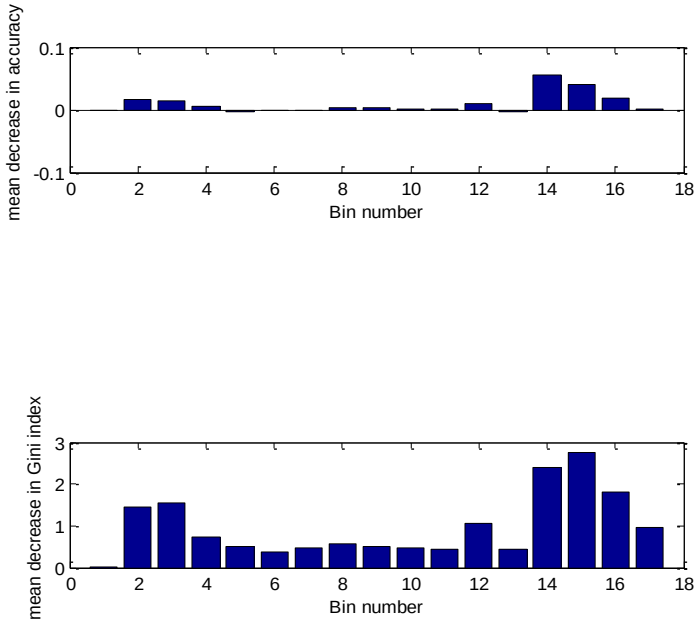



Bins are 10cm$^{-1}$ width from 1205- 1215 cm$^{-1}$; (bin 1) to 1335-1345 cm$^{-1}$ (bin 14). Data from 1255 cm$^{-1}$ to 1265 cm$^{-1}$ is most important.

1-1205- 1215 cm$^{-1}$; 2-1215- 1225 cm$^{-1}$; 3-1225-1235cm$^{-1}$; 4-1235-1245cm$^{-1}$;5-1245-1255cm$^{-1}$; 6-1255- 1265 cm$^{-1}$; 7-1265-1275cm$^{-1}$; 8-1275-1285cm$^{-1}$;9-1285-1295cm$^{-1}$; 10-1295-1305cm$^{-1}$;11-1305-1315cm$^{-1}$;12-1315-1325 cm$^{-1}$;13-1325-1335 cm$^{-1}$;14-1335-1345cm$^{-1}$

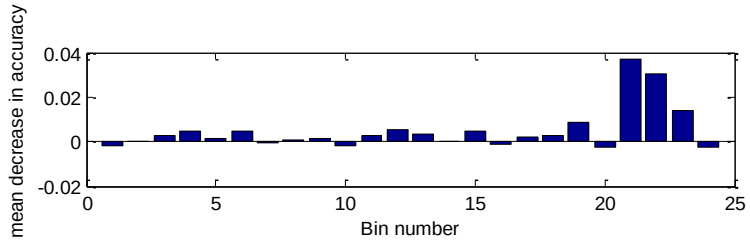
Raman Amide I+II  Variable Importance and Gini Index Plots

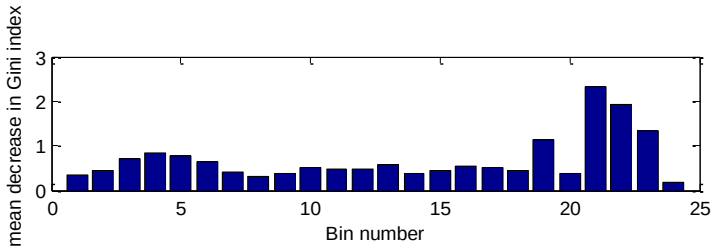



Raman Amide I+II variable importance plots: Mean decrease in Accuracy (top); Mean decrease in Gini index (bottom). Bins are 10cm$^{-1}$ width from 1515- 1525 cm$^{-1}$; (bin 1) to 1595-1605 cm$^{-1}$ (bin 9).

1-1515- 1525 cm$^{-1}$; 2-1525-1535cm$^{-1}$; 3-1535-1545cm$^{-1}$;4-1545-1555cm$^{-1}$; 5-1555-1565cm$^{-1}$;6-1565-1575cm$^{-1}$;7-1575-1585cm$^{-1}$;8-1585-1595cm$^{-1}$;9-1595-1605cm$^{-1}$;10-1605- 1615 cm$^{-1}$; 11-1615-1625cm$^{-1}$; 12-1625-1635cm$^{-1}$;13-1635-1645cm$^{-1}$; 14-1645-1655cm$^{-1}$;15-1655-1665cm$^{-1}$;16-1665-1675cm$^{-1}$;17-1675-1685cm$^{-1}$;

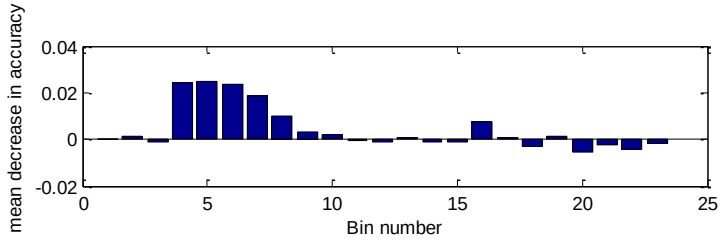
Raman Amide I+III  Variable Importance and Gini Index Plots

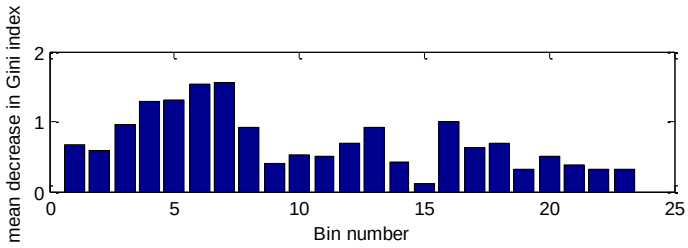



1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$;14-1605- 1615 $cm^{-1}$; 15-1615-1625$cm^{-1}$; 16-1625-1635$cm^{-1}$;18-1635-1645$cm^{-1}$; 19-1645-1655$cm^{-1}$;20-1655-1665$cm^{-1}$;21-1665-1675$cm^{-1}$;22-1675-1685$cm^{-1}$;23-1685-1695$cm^{-1}$;24-1695-1705 $cm^{-1}$;

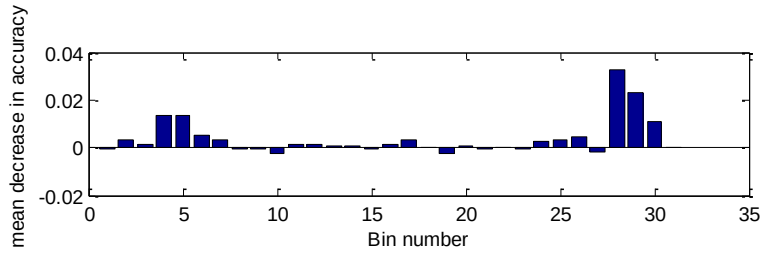
Raman Amide II+III  Variable Importance and Gini Index Plots

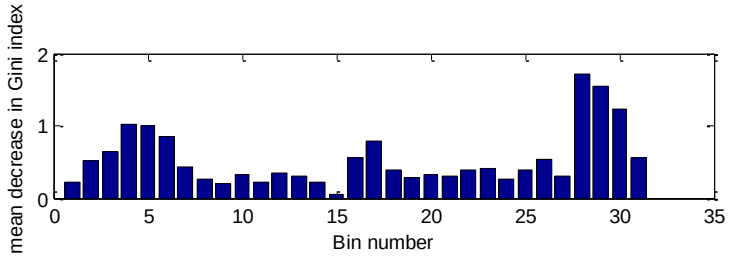



1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$; 14-1515- 1525 $cm^{-1}$; 15-1525-1535$cm^{-1}$; 16-1535-1545$cm^{-1}$;17-1545-1555$cm^{-1}$; 18-1555-1565$cm^{-1}$;19-1565-1575$cm^{-1}$;20-1575-1585$cm^{-1}$;21-1585-1595$cm^{-1}$;22-1595-1605$cm^{-1}$;23-1605-1615 $cm^{-1}$;

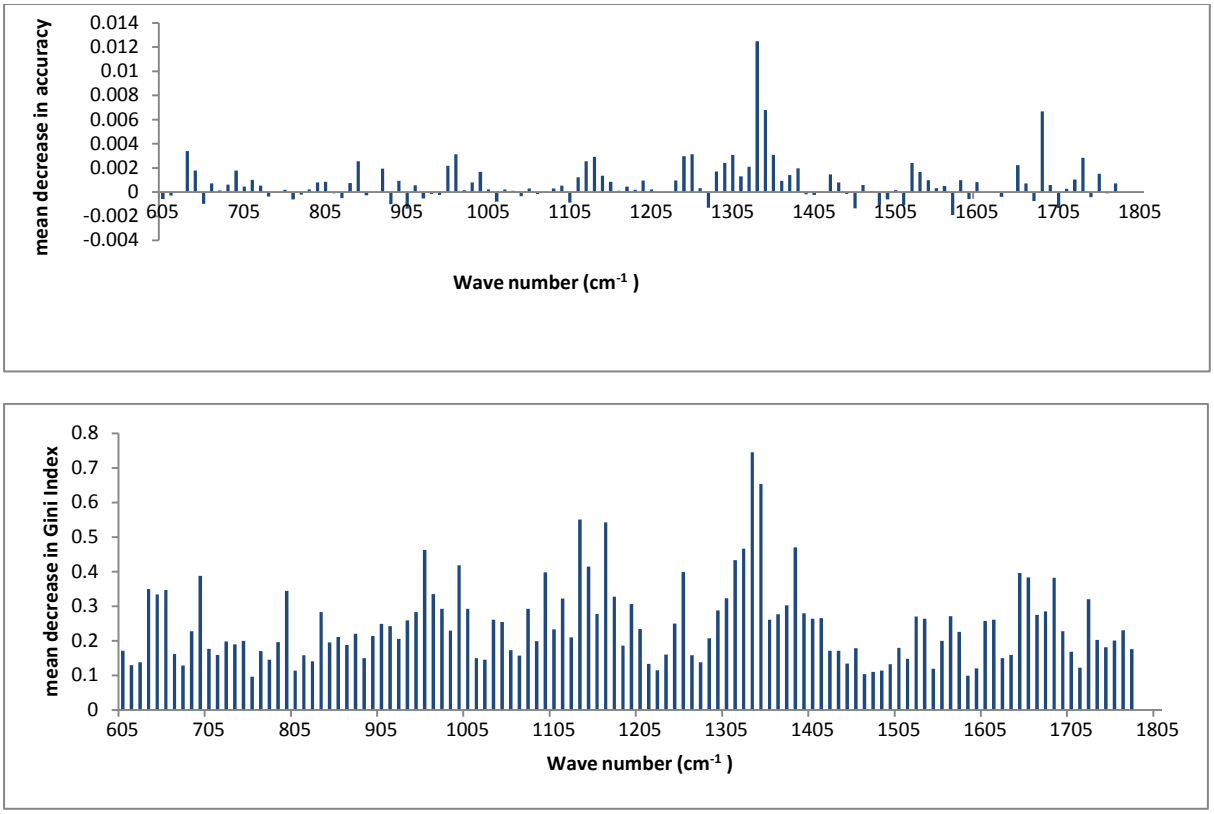
Raman Amide I+ II+III  Variable Importance and Gini Index Plots



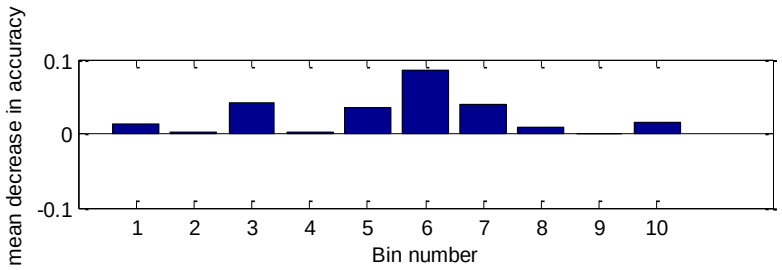

1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$; 14-1515- 1525 $cm^{-1}$; 15-1525-1535$cm^{-1}$; 16-1535-1545$cm^{-1}$;17-1545-1555$cm^{-1}$; 19-1555-1565$cm^{-1}$;20-1565-1575$cm^{-1}$;21-1605-1615$cm^{-1}$;22-1615-1625$cm^{-1}$23-1625- 1635 $cm^{-1}$; 24-1635-1645$cm^{-1}$; 25-1645-1655$cm^{-1}$;26-1655-1665$cm^{-1}$; 27-1665-1675$cm^{-1}$;28-1675-1685$cm^{-1}$;29-1685-1695$cm^{-1}$;30-1695-1705$cm^{-1}$;31-1705-1710$cm^{-1}$;

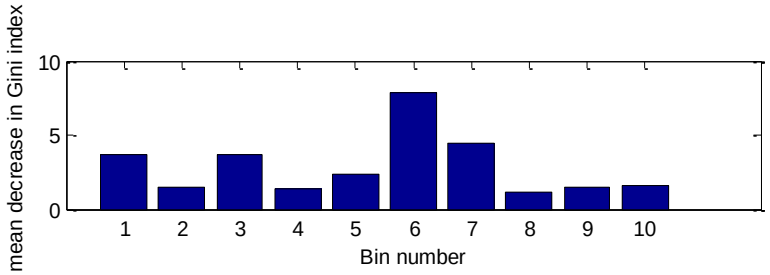
Fig. 6 ROA Whole Variable Importance (top) and Gini Index (bottom) plots for Spectral Subdivisions




Bins are 10 $cm^{-1}$ width from 605- 615 $cm^{-1}$ to 1765-1775 $cm^{-1}$

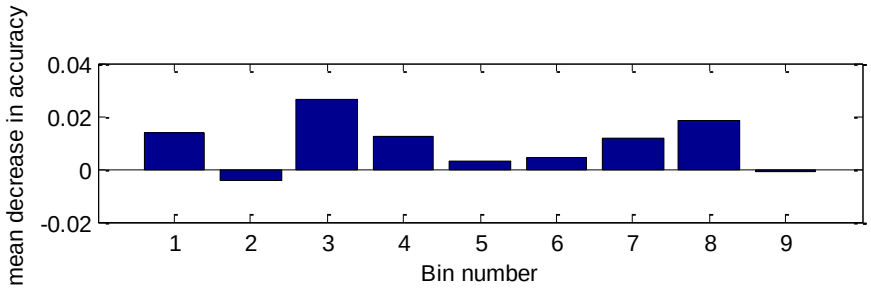
ROA Amide I Variable Importance & Gini Index Plots

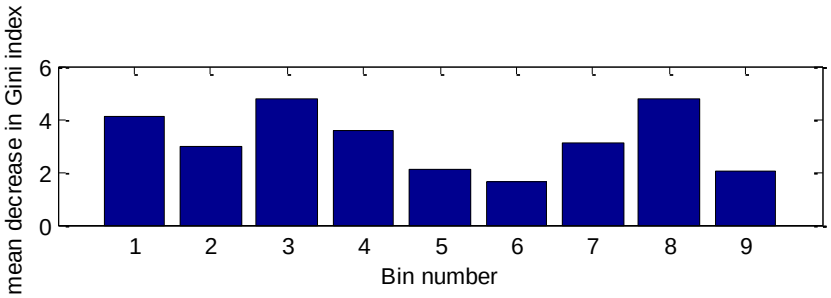



1-1605- 1615 $cm^{-1}$; 2-1615-1625$cm^{-1}$; 3-1625-1635$cm^{-1}$;4-1635-1645$cm^{-1}$; 5-1645-1655$cm^{-1}$;6-1655-1665$cm^{-1}$;7-1665-1675$cm^{-1}$;8-1675-1685$cm^{-1}$;9-1685-1695$cm^{-1}$; 10-1695-1705$cm^{-1}$;

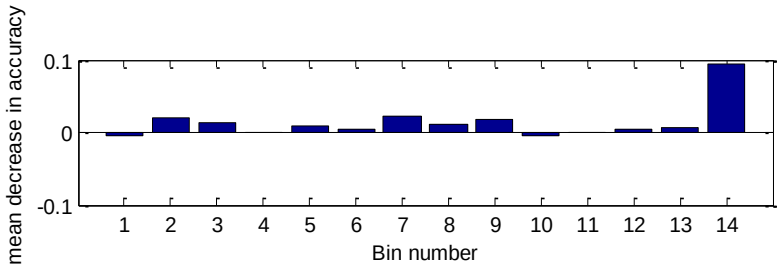
ROA Amide II Variable Importance & Gini Index Plots

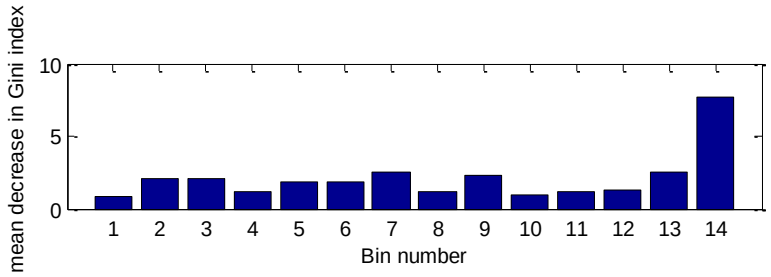



1-1515- 1525 $cm^{-1}$; 2-1525-1535$cm^{-1}$; 3-1535-1545$cm^{-1}$;4-1545-1555$cm^{-1}$; 5-1555-1565$cm^{-1}$;6-1565-1575$cm^{-1}$;7-1575-1585$cm^{-1}$;8-1585-1595$cm^{-1}$;9-1595-1605$cm^{-1}$;

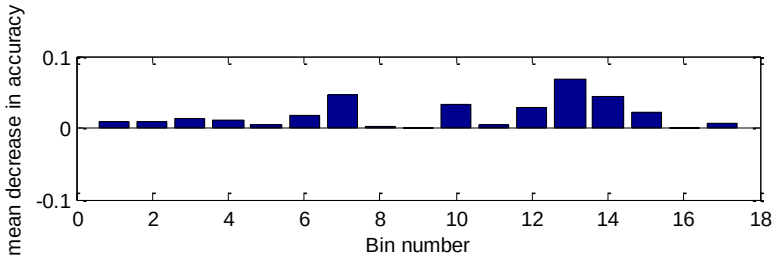
ROA Amide III Variable Importance & Gini Index Plots

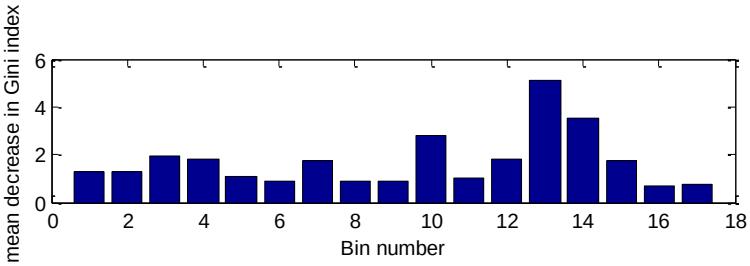



1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$;14-1335-1345$cm^{-1}$;

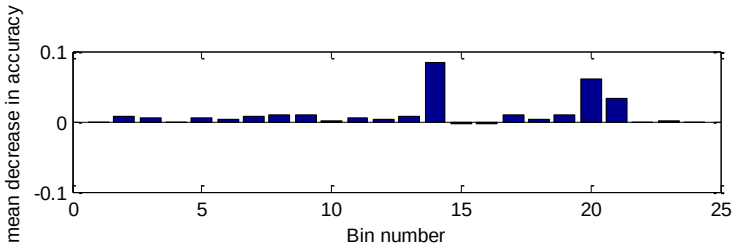
ROA Amide I+ II Variable Importance & Gini Index Plots

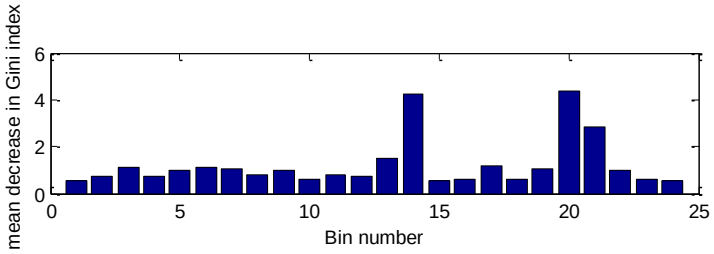



1-1515- 1525 $cm^{-1}$; 2-1525-1535$cm^{-1}$; 3-1535-1545$cm^{-1}$;4-1545-1555$cm^{-1}$; 5-1555-1565$cm^{-1}$;6-1565-1575$cm^{-1}$;7-1575-1585$cm^{-1}$;8-1585-1595$cm^{-1}$;9-1595-1605$cm^{-1}$;10-1605- 1615 $cm^{-1}$; 11-1615-1625$cm^{-1}$; 12-1625-1635$cm^{-1}$;13-1635-1645$cm^{-1}$; 14-1645-1655$cm^{-1}$;15-1655-1665$cm^{-1}$;16-1665-1675$cm^{-1}$;17-1675-1685$cm^{-1}$;

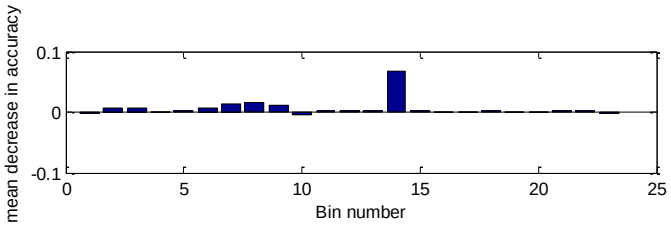
ROA Amide I+ III Variable Importance & Gini Index Plots

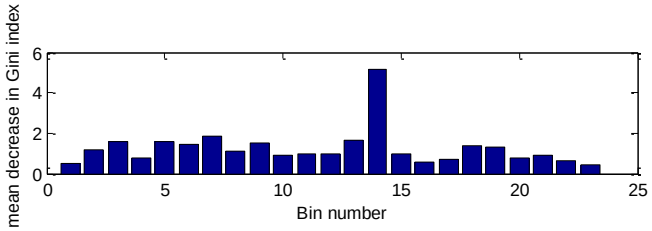



1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$;14-1605- 1615 $cm^{-1}$; 15-1615-1625$cm^{-1}$; 16-1625-1635$cm^{-1}$;18-1635-1645$cm^{-1}$; 19-1645-1655$cm^{-1}$;20-1655-1665$cm^{-1}$;21-1665-1675$cm^{-1}$;22-1675-1685$cm^{-1}$;23-1685-1695$cm^{-1}$;24-1695-1705 $cm^{-1}$;

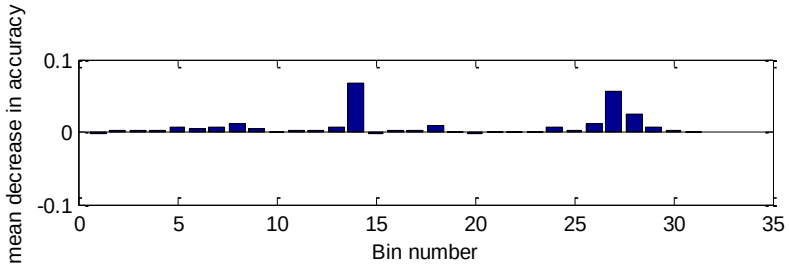
ROA Amide II+ III Variable Importance & Gini Index Plots

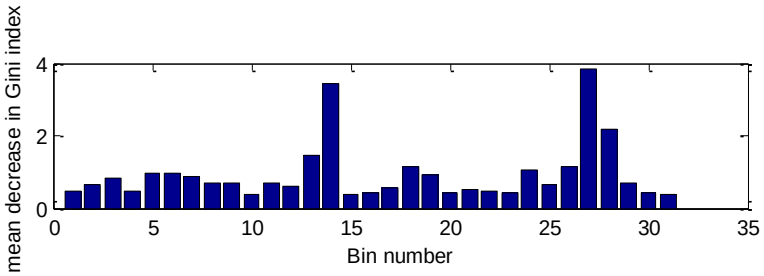



1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$; 14-1515- 1525 $cm^{-1}$; 15-1525-1535$cm^{-1}$; 16-1535-1545$cm^{-1}$;17-1545-1555$cm^{-1}$; 18-1555-1565$cm^{-1}$;19-1565-1575$cm^{-1}$;20-1575-1585$cm^{-1}$;21-1585-1595$cm^{-1}$;22-1595-1605$cm^{-1}$;23-1605-1615 $cm^{-1}$;

ROA Amide I+ II+ III Variable Importance & Gini Index Plots

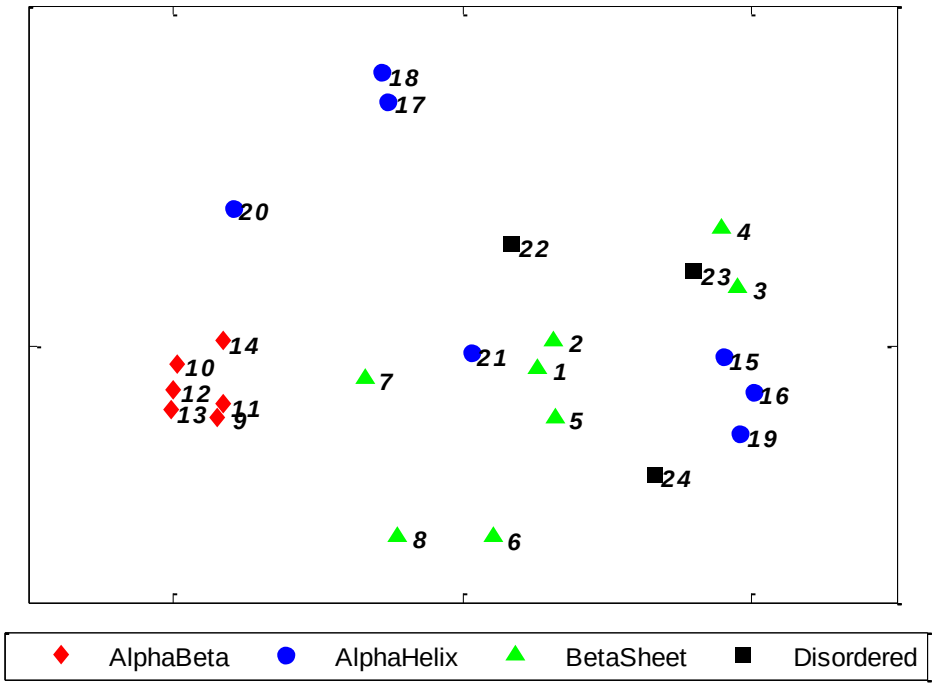


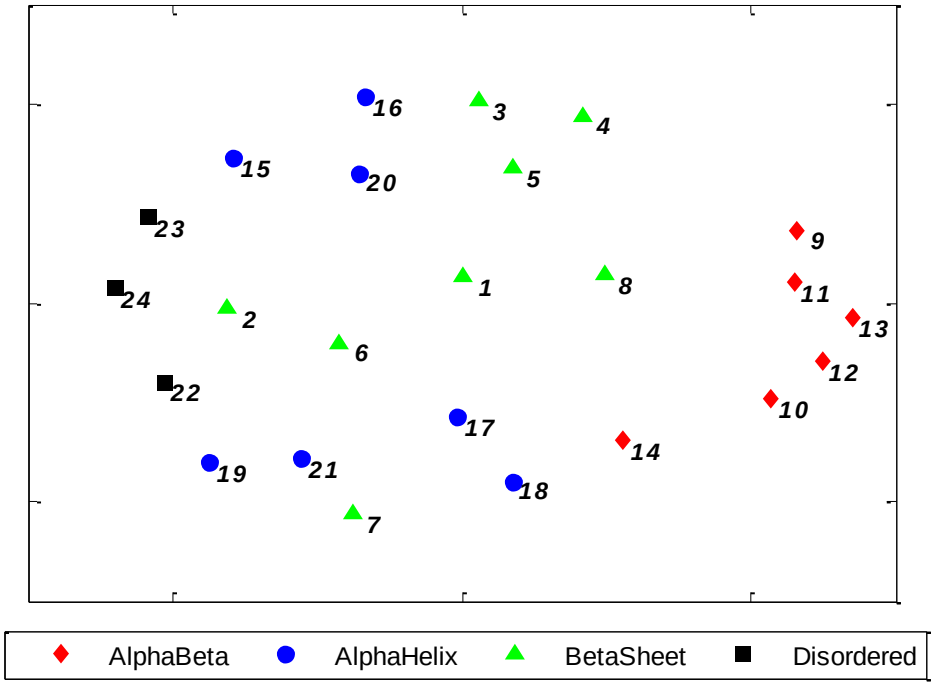

1-1205- 1215 $cm^{-1}$; 2-1215- 1225 $cm^{-1}$; 3-1225-1235$cm^{-1}$; 4-1235-1245$cm^{-1}$;5-1245-1255$cm^{-1}$; 6-1255- 1265 $cm^{-1}$; 7-1265-1275$cm^{-1}$; 8-1275-1285$cm^{-1}$;9-1285-1295$cm^{-1}$; 10-1295-1305$cm^{-1}$;11-1305-1315$cm^{-1}$;12-1315-1325$cm^{-1}$;13-1325-1335$cm^{-1}$; 14-1515- 1525 $cm^{-1}$; 15-1525-1535$cm^{-1}$; 16-1535-1545$cm^{-1}$;17-1545-1555$cm^{-1}$; 19-1555-1565$cm^{-1}$;20-1565-1575$cm^{-1}$;21-1605-1615$cm^{-1}$;22-1615-1625$cm^{-1}$23-1625- 1635 $cm^{-1}$; 24-1635-1645$cm^{-1}$; 25-1645-1655$cm^{-1}$;26-1655-1665$cm^{-1}$; 27-1665-1675$cm^{-1}$;28-1675-1685$cm^{-1}$;29-1685-1695$cm^{-1}$;30-1695-1705$cm^{-1}$;31-1705-1710$cm^{-1}$;

Appendix I- Multidimensional Scaling (MDS) plots for Random Forest analyses

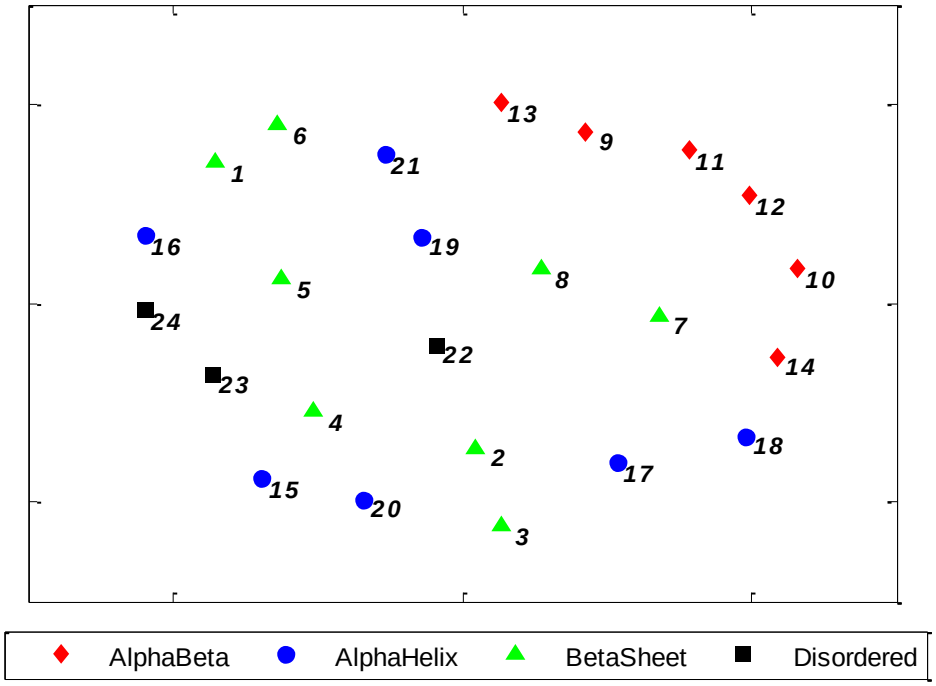
Multidimensional Scaling Plot Raman Amide I spectra




Multidimensional Scaling Plot for Raman Amide I& II spectra



Multidimensional Scaling Plot for Raman Amide I & III spectra



Appendix J-Key to proteins used in ROA MDS Plot

| | . | PDB Code |
|---|---|---|
| 1 | turkey ovomucoid | 1m8c |
| 2 | human lactofferin | 1fck |
| 3 | MS2 virus | 2ms2 |
| 4 | ribonuclease A | 1afu |
| 5 | Ribonuclease b | 1rbb |
| 6 | subtilisin | 1ndq |
| 7 | ubiquitin | 1ubq |
| 8 | aldolase | 2ald |
| 9 | hen lysosyme | 1lsc |
| 10 | *human ovomucoid | 1m8c |
| 11 | human lysozyme | 1gaz |
| 12 | insulin | 1zeh |
| 13 | α-lactalbumin | 1f6s |
| 14 | full length prion protein | 1y2s |
| 15 | equine lysozyme | 2eql |
| 16 | tobacco mosaic virus | 1vtm |
| 17 | filamentous bacteriophage M13 | 2cpb |
| 18 | creatine kinase | 1i0e |
| 19 | filamentous bacteriophage fd | 1fdm |
| 20 | human serum albumin | 1e78 |
| 21 | filamentous bacteriophage Pf1 | 1pfi |
| 22 | prion protein (90-230) | 1qm1 |
| 23 | S100Bs | 1uwo |
| 24 | S100A6 calcyclin | 1k96 |
| 25 | immunoglobin G | 1ig2 |
| 26 | cowpea mosaic virus | 1ny7 |
| 27 | serum amyloid protein | 1lgn |
| 28 | invertase | 2ac1 |
| 29 | bordetella pertussis P.69 pertactin | 1dab |
| 30 | pepsin | 1am5 |
| 31 | human serum amyloid P component | 1sac |
| 32 | satellite tobacco mosaic virus | 1a34 |
| 33 | bovine trypsin | 1k1l |
| 34 | chymotrypsin | 4cha |
| 35 | amylase | 1kgu |
| 36 | avidin | 1rav |
| 37 | bovine beta lactalbumin | 1b8e |
| 38 | concanavalin A | 3cna |
| 39 | trypsinogen | 2tga |
| 40 | bovine beta lactalbumin pH 2 | 1dv9 |
| 41 | Bowman-Birk inhibitor | 1pi2 |
| 42 | orosmucoid | 3kq0 |
| 43 | antifreeze protein | 1b7i |
| 44 | metallothionein | 4mt2 |

Appendix K- Key to the proteins used in Raman MDS plots

| | | PDB Code |
|---|---|---|
| 1 | lactoferrin | 1fck |
| 2 | MS2 Caspid | 2ms2 |
| 3 | Ovine PrP90-230 β-isoform | 1qm1 |
| 4 | ribonucleaseA | 1afu |
| 5 | insulin | 1zeh |
| 6 | hen lysozyme | 1lsc |
| 7 | human lysozyme | 1gaz |
| 8 | aldolase | 2ald |
| 9 | apoS100A6 | 1k96 |
| 10 | filamentous bacteriophage fd | 1fdm |
| 11 | human serum albumin | 1e78 |
| 12 | filamentous bacteriophage IKE | 1ifl |
| 13 | filamentous bacteriophage M13 | 2cpb |
| 14 | tobacco mosaic virus | 1vtm |
| 15 | concanavalin A | 3cna |
| 16 | human immunoglobinG | 2ig2 |
| 17 | cowpea mosaic virus | 1ny7 |
| 18 | satellite tobacco mosaic virus | 1a34 |
| 19 | invertase | 2ac1 |
| 20 | bovine β-Lactalbumin pH2 | 1dv9 |
| 21 | α-lactalbumin | 1f6s |
| 22 | orosomucoid | 3kq0 |
| 23 | Bowman-Birk protease inhibitor | 1pi2 |
| 24 | metallothionein | 4mt2 |

**Appendix L**-Published Papers