Order Number 9029972

Experimental and computational approaches to protein secondary structure

Prestrelski, Steven Joseph, Ph.D.

City University of New York, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

EXPERIMENTAL AND COMPUTATIONAL APPROACHES TO PROTEIN SECONDARY STRUCTURE

by

STEVEN J. PRESTRELSKI

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, the City University of New York.

1990

This manuscript has been read for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

3/5/90
Date

Chair of Examining Committee

3/21/90
Date

Executive Officer

Supervisory Committee

The City University of New York

Abstract

EXPERIMENTAL AND COMPUTATIONAL APPROACHES TO PROTEIN SECONDARY STRUCTURE

by

STEVEN J. PRESTRELSKI

Adviser: Doctor Michael N. Liebman

The specific objective of this thesis was to increase the usefulness of Fourier-transform infrared spectroscopy in the determination of protein conformation in solution, though study of the conformation-sensitive amide I band. This has been accomplished through a combined computational and experimental approach. First, an algorithm has been developed for description and classification of protein secondary structures, from known protein structures, which operates independent of a predetermined structure template. Independence from this template was demonstrated to provide a more comprehensive classification of protein conformation. Fourier-transform infrared spectroscopic experiments, using derivative spectroscopy and Fourier self-deconvolution, were performed on series of bovine trypsin conformers a homologous series of proteins, the serine proteases, for which high-resolution crystal structures were available. The spectra were interpreted in the light of the enhanced description of protein conformation and these experiments have resulted in new spectra-structure correlations for the amide I infrared band allowing for a more complete interpretation of the spectral information in this region. These results allow for greater extraction of conformational information from protein infrared spectra.

Acknowledgements

There are many people whose help has been invaluable in producing this dissertation. To list them all would be impossible and I would risk missing names. I would like to specifically thank; Dr. Michael Liebman, my thesis advisor, for all of the time and work involved in generating this dissertation; Dr. Thomas Kumosinski, for his initial encouragement and continuing support; Drs. Sandy Ross and Herman Wyssbrod, for their invaluable help and consistent good humor; Dr. Michael Byler, for his invaluable assistance in collecting and analyzing the spectroscopic data as well as for making that aspect of this thesis so enjoyable; Drs. William Laws and Massimo Sassiroli, for reading this thesis and for their many helpful suggestions; and Dr. Harel Weinstein, for guidance and support. I would also like to thank the many friends and colleagues whose acquaintance has made the last several years stimulating as well as enjoyable. Most of all, I would like to extend very special, heartfelt thanks my mother and father. Without their constant support and encouragement this thesis would not have been possible.

I would also like to acknowledge the partial financial support received from the National Institutes of Health, grant R01GM39750, and from a grant to Dr. Michael Liebman from ImClone Systems Inc. In addition I would like to acknowledge grants of instrument and computer time from the United States Dept. of Agriculture, Eastern Regional Research Center, and from the AMOCO Technology Company. Further, I would like to thank Dr. Richard Schultz and the members of the Department of Biochemistry, Loyola University, Stritch School of Medicine for materials and assistance in preparation of many of the compounds studied for this dissertation.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

The three-dimensional structure of biological macromolecules and their physiological function are intimately related. A long standing goal of molecular biophysics is the elucidation of this complex relationship. Proteins are complex biomolecules whose functions are responsible for a large number of biological processes. The importance of understanding structure-function relationships of proteins involved in these processes and the recently acquired ability of biotechnology to produce altered versions of natural proteins as well as de novo synthesis of designed proteins has underscored the need for a rapid, reliable and sensitive probe of molecular conformation. Such a method should allow study in aqueous media and also be suitable for use in measuring the effects of environmental factors, ligand-binding and other perturbations on molecular conformation.

Empirical, computer-based methods have been used in attempts to predict protein structure based on statistical information derived from previously determined structures or from energetic considerations. However, these computational methods have not reached a sufficient degree of sophistication to

determine protein structures from readily available information (e.g., the amino acid sequence). Thus, experimental methods remain the most important resource for obtaining protein structure information. The most widely used experimental methods for determination of protein structure and conformation have been circular dichroism (CD), nuclear magnetic resonance spectroscopy (NMR) and x-ray crystallography. While these techniques are capable of providing information concerning protein conformation, each has certain inherent limitations. The experimental determination of protein conformation has been limited by both experimental considerations and limitations in analytically defining protein conformational elements in an unambiguous, objective and encompassing manner.

Conformational analysis of polypeptides and proteins using circular dichroism spectroscopy (and its complement, optical rotatory dispersion) is limited conceptually as well as experimentally. Its present analysis entails limited structural models to describe protein conformation. Spectral analysis of proteins utilizes a basis set approach in which observed spectra are considered to be a linear combination of basis spectra. These basis spectra are derived from model homopolymers or proteins of known structure. Thus, spectral features arising from particular secondary structures are not observed directly. Further, experimental estimates of certain conformations, in particular β-strands, from CD are qualitative or, at best, semi-quantitative in nature (Brahms and Brahms, 1980; Tinoco and Williams, 1984, Johnson, 1988).

One- and two-dimensional NMR techniques show great promise for polypeptide and protein structural analysis. However, at present these methods involve extensive manipulations to produce peak assignments which often involve several resonance techniques (e.g. COSY, NOESY, etc.) and nuclear probes. These and other factors limit their effective use for protein structure

determination to low molecular weight proteins (Markely and Urich,1984). Further, it is not possible to study interactions and environmental effects under several conditions in a reasonable amount of time.

X-ray crystallography currently yields the most detailed information concerning protein conformation and structure. However, determination of protein structure with diffraction techniques is a difficult, time-consuming process, often taking years to solve a single protein structure. Its use is limited to those proteins which can form perfect, single crystals (Blundell and Johnson, 1976). Further, because the experiments are performed on crystals, diffraction studies introduce certain limitations due to the crystalline environment, which may differ from the proteins native environment. Thus, it is difficult to study the responses to environmental factors and interactions with ions and other important ligands.

Infrared spectroscopy is currently used for limited conformational analysis of proteins (Susi and Byler, 1986, Surewicz and Mantsch, 1988). Using modern, interferometric spectrometers, this method can provide high-resolution, high signal-to-noise spectra in a short time. It is simple, rapid, non-destructive and provides information from several conformation-sensitive bands which arise from vibrations localized within the peptide structure. Spectral features arising from vibrations of individual folding types are observed directly and a basis set approach, such as that used with circular dichroism spectroscopy, is not necessary. Aqueous protein solutions can be studied, but experiments are not limited to this medium. Despite the extensive information gained with this method, present analysis lacks essential correlations between spectral features and protein conformational elements. Thus, at present, much structural information afforded by the conformation-sensitive bands remains uninterpretable.

The objective of this thesis project is the further development of infrared spectroscopy as a method for the determination of protein conformation in aqueous solution. This is achieved through an integrated approach which involves spectroscopic studies in addition to development of a more advanced algorithm for protein secondary structure analysis. Specifically, this thesis focuses on: 1) development of a more encompassing analytical method for definition and classification of protein conformational elements from atomic coordinate data, 2) discerning correlations between infrared spectral features and protein conformational elements through examination of a series of closely related protein molecules for which high-resolution crystallographic structures have been determined and, 3) extension and evaluation of these and previously proposed spectra-structure correlations by studying a homologous series of proteins

CHAPTER 2

BACKGROUND AND METHODS

2.1. INFRARED SPECTROSCOPY AND PROTEIN CONFORMATION

Infrared spectroscopy is widely used as a probe of molecular structure. This method has been recognized as a tool for studying conformation in polypeptides and proteins since Elliot and Ambrose (1950a) demonstrated that polypeptides known from diffraction experiments to adopt different conformations gave rise to different absorbance maxima in their infrared spectra. Polypeptides known to be in the α or β forms (i.e., helical or extended, respectively) from x-ray diffraction studies, displayed absorbance of the C=O stretching and N-H deformation modes characteristic of peptides at different frequency. The α form exhibited absorbance maxima at 1658 cm^{-1} for the peptide C=O stretching and 1550 cm^{-1} for the peptide N-H deformation modes and the β form at 1627 cm^{-1} and 1527 cm^{-1}, for these two modes respectively. In subsequent studies, these investigators established empirical correlations for C=O stretching and N-H deformation modes for α and β polypeptides. Polypeptides adopting the α conformation exhibit absorbance maxima at

approximately 1660 cm^{-1} and 1550 cm^{-1} and polypeptides adopting the β conformation at 1630 cm^{-1} and 1525 cm^{-1} (Elliot, 1953; Elliot, 1954; Bamford et al.,1953; Ambrose and Elliot, 1951a). Upon denaturation the C=O stretching frequency becomes a broad peak at 1640 cm^{-1} (Elliot et al, 1950). Thus, random coils or unordered conformations were assigned to this frequency. These observations were extended to interpret infrared spectra of fibrous (Ambrose et al.,1951; Ambrose and Elliot, 1951b; Bradbury and Elliot, 1963; Elliot and Malcolm, 1956) and globular (Elliot et al.,1950; Elliot and Ambrose,1950b; Ambrose and Elliot, 1951b) proteins. However, Beer et al. (1959), after an extensive study, warned that the earlier extrapolations to proteins were premature and that such spectra-structure correlations may not be directly applicable when several different conformations exist, as is observed in proteins.

Miyazawa and coworkers attempted to determine the origin of the characteristic peptide vibrations in a series of experimental (Miyazawa et al., 1956; Miyazawa, 1961) and theoretical (Miyazawa et al., 1958; Miyazawa, 1960a; Miyazawa, 1962) studies. These investigators determined the chemical groups which give rise to the characteristic amide vibrations using peptide analogues. The characteristic amide modes of N-methylacetamide, a small molecule which contains the planar CONH group characteristic of peptides, are listed in Table 2-1. The amide I,II and III modes are, for the greater part, localized within the peptide CONH group while the others involve displacement of other atoms. Thus, the amide I, II and III modes are the focus of most attempts to correlate polypeptide and protein conformation with infrared spectral properties, with the amide I mode being the most studied.

The amide I mode involves predominantly stretching of the C=O bond with smaller contributions from C--N stretching and N--H bending. While the

amplitude of the N--H bending displacement is considerable, its contribution to the potential energy is small (~10%). Thus, by the potential energy criterion the amide I mode is primarily a C=O stretching motion. This displacement accounts for approximately 80% of the potential energy of the amide I vibration in N-methylacetamide (Miyazawa, 1962). The simplicity and localization of the amide I vibrational mode results in relative ease of interpretation and, in part, accounts for widespread use in conformational studies.

Initial attempts to explain shifts in the absorbance maxima of the amide I and II modes due to differences in conformation were those of Krimm (1955), who computed the frequency as a function of hydrogen bond distances and angles, and of Cannon (1956), in terms of dipole-dipole interactions between adjacent oscillators in the peptide groups of polypeptides. However, neither of these attempts proved adequate in interpreting polypeptide spectra. Miyazawa proposed a perturbation treatment for α-helical and anti-parallel β-sheet conformations which described the perturbation of a vibrational frequency for a free oscillator by interactions with adjacent oscillators as well as intra- and inter-chain hydrogen bonding (Miyazawa, 1960b; Miyazawa and Blout, 1961). Krimm proposed a modification of this treatment (Krimm, 1962) and extended it to include other conformations (Krimm and Abe, 1972; Moore and Krimm,1975). Bradbury and Elliot (1963) applied the perturbation treatment in a study of polyamides, polypeptides and fibrous proteins and concluded that it was useful in interpreting infrared spectra of these compounds. However, while this perturbation treatment proved significant in understanding the spectra of polypeptides and fibrous proteins, this treatment is not applicable to other types of peptide structures such as those which occur in globular proteins because it is not possible to extend the coefficients describing the interactions to irregular or aperiodic structures.

The inability of Miyazawa's theoretical treatment to provide a broadly applicable basis for conformational interpretation of polypeptide and protein spectra led to use of the more generalizable method of normal mode analysis. This method treats a molecule as a series of multidimensional, coupled harmonic oscillators in which atoms are represented by point masses and a force field is employed to describe the interaction of each atom with other bonded and non-bonded atoms in the molecule (Wilson, 1939; Wilson, 1941; Wilson et al., 1955). The frequency of all vibrations can be computed and compared with experimentally measured spectra. Much of the present research in this field concentrates on developing adequate force fields for macromolecular systems (Krimm and Bandekar, 1986). Normal mode analysis has provided significant information applicable to interpretation of polypeptide and protein vibrational spectra by providing theoretical frequencies for many ideal polypeptide conformations (Krimm and Bandekar, 1986). The calculated frequencies agree well with those determined experimentally; however, limitations exist in extending these results to protein spectra. Symmetry considerations which are often necessary to simplify calculations in ideal systems are rarely applicable to proteins. In addition, the approximation of side chains by point masses and use of model systems which correspond to infinite length structures, limits extension to protein molecules where individual conformations are of finite length. Normal mode analysis itself is not extendable to protein molecules because their complexity and large size makes its application computationally prohibitive at present. Currently, normal mode analysis can be successfully applied to polypeptides up to 20 residues in length, if side-chains are approximated as point masses (Naik and Krimm, 1986a,b), and 6-8 residues with full atom treatment (Krimm and Bandekar, 1986). Thus, normal mode analyses have provided guidelines for

interpretation of protein vibrational spectra but, at present, direct correlations are not possible.

Homopolypeptides of known three-dimensional structure have been used to examine the infrared spectral properties of secondary and tertiary (i.e., β-sheet) structures in proteins. These experimental studies have often been performed in conjunction with theoretical treatments. Studies have been performed to determine the effects of conformation on vibrational spectra of polypeptides of known structure including α-helices (Elliot, 1954; Bamford et al, 1953; Ambrose and Elliot, 1951a; Elliot and Malcolm; 1956; Chirgadze and Rashevskaya, 1969; Chirgadze and Brazhnikov, 1974; Suzuki et al., 1966), β-sheets (Elliot, 1954; Ambrose and Elliot, 1951a,Elliot and Malcolm, 1956; Suzuki et al., 1966;Chirgadze et al., 1973), β-turns (Bandekar et al., 1982; Bandekar and Krimm, 1979; Bandekar and Krimm, 1979; Naik and Krimm, 1984; Maxfield et al., 1981), γ-turns (Bandekar and Krimm, 1985), 3_{10}-helices (Dwivedi et al., 1984) as well as random coil polypeptides (Chirgadze et al., 1973). Other studies have used model peptide systems to examine spectral intensities of these conformations (Chirgadze and Rashevskaya, 1969; Chirgadze and Brazhnikov, 1974; Chirgadze et al., 1973), the effects of side-chains on amide vibrations (Chirgadze et al., 1975) as well as other environmental and solution effects (Chirgadze and Brazhnikov, 1973). While such studies have provided broad guidelines for structural interpretation of vibrational spectra, certain limitations exist in applying these results to the interpretation of spectra of globular proteins. Individual conformations in globular proteins are typically shorter in length and exhibit less regularity than the model homopolymers. In addition, the sequence identity of a given conformation in a protein is non-homogenous. It is not presently known in detail to what extent any or all of these factors affect the spectral properties.

Attempts have been made at discerning infrared spectra-structure correlations by studying proteins which are known from crystallographic experiments to adopt one predominant conformation for the majority of the backbone (e.g., myoglobin, α-helix, β-lactoglobulin, antiparallel β-sheet) (Timasheff et al., 1967; Susi, 1969). These studies showed that vibrational frequencies for conformations determined in studies on model systems described above were roughly transferable to proteins for the α-helix (~1652 cm^{-1}), β-sheet (~1630, 1675 cm^{-1}) and unordered (~1645 cm^{-1}) conformations. For other conformations (i.e., reverse turns, loops), the absorptions were not resolved and no assignments could be made.

The previously obtained information relating spectral characteristics to protein conformations from theoretical and experimental studies have been used to predict the structural composition of globular proteins using the conformation-sensitive amide regions, particularly the amide I region (Susi, 1969: Susi, 1972; Susi et al., 1967; Timasheff et al., 1967). However, data analysis in terms of conformation with infrared spectra from conventional (i.e. grating or dispersive) instruments is limited. These spectra typically exhibit low signal-to-noise ratios and poor wavenumber accuracy. More importantly, the component bands arising from the amide vibrations are broad and lie in close proximity so that the observed composite band is, therefore, often featureless. Thus, conformational information obtainable from such data was of a limited and qualitative nature; at best one could only discern a rough estimate of the predominant conformation in a given protein. Later efforts that made use of iterative curve-fitting techniques provided a more quantitative estimate of the proportion of β structure which were in good agreement with observations from x-ray crystallography and estimates from circular dichroism (Ruegg et al., 1975). The amide I peak arising from α-helices, however, could often not be resolved

from the composite amide I band in proteins containing significant amounts of other conformations. Thus, conformational analysis of proteins using conventional instrumentation is limited to this level primarily by the inherent broadness and overlapping of amide component bands. Increased instrumental resolution is to no avail.

The advent of interferometric or Fourier transform infrared spectrometers has led to significant advances in infrared spectroscopy, (Griffiths and deHaseth, 1986) particularly in protein conformational analysis (Parker, 1986). The interferometric spectrometers are based on the original design of Michelson which employs a system of fixed and moving mirrors whose position results in constructive or destructive interference. Spectral intensity is measured as a function of the moving mirror position and this information is termed the interferogram. The spectrum, then, is the Fourier-transform of the interferogram.

Opposed to conventional instrumentation, interferometric, or Fourier-transform spectrometers, provide several advantages. It allows an increase in the speed of data collection because the spectral contributions from all frequencies are measured simultaneously (this has been termed the multiplexing, or Fellget's, advantage). In addition the throughput of light in the interferometric configuration is not slit limited as in conventional instruments (throughput or Jacquinot's advantage). These two advantages lead to a several-fold increase in the signal-to-noise ratio (Griffiths and DeHaseth, 1986). This configuration also allows use of a laser reference system which determines the position of the moving mirror to high accuracy, greatly increasing wavelength accuracy (Connes advantage). Further, the inherent digital nature of the Fourier transform spectrometer facilitates processing spectra using solvent subtraction, optical corrections and other quantitative calculations.

The digital character of Fourier transform infrared spectroscopic data, combined with high signal-to-noise ratio, also allows use of 'resolution-enhancement' techniques. Derivative spectroscopy (Butler, 1979; Maddams and Tooke, 1982) and Fourier self-deconvolution (Kauppinen et al., 1981a) enhance the spectroscopist's ability to visualize component peaks in multicomponent, overlapping bands such as those arising from conformation-sensitive modes in protein spectra. It is important to note, however, that these techniques do not increase the actual instrumental resolution but only increase the investigators ability to visualize component bands in multicomponent, overlapped bands.

Present determination of protein conformation with Fourier transform infrared spectroscopy (FTIR) involves analysis of the amide I band in terms of component peaks disclosed by resolution enhancement techniques (Byler and Susi, 1986; Purcell and Susi, 1984; Susi et al., 1984; Yang et al., 1985). The amide I band (1700-1620 cm^{-1}) has been the most studied and appears to be the least complex of the conformation-sensitive IR bands associated with the peptide group. Application of these procedures to the infrared spectra of 20 proteins has yielded component peaks in 11 distinct frequency regions which, ideally, arise from the component secondary structures (Byler and Susi, 1986). Present experimental analyses can empirically determine protein conformation with a limited three-state (i.e. helix, extended structure, random coil) structural model (Byler and Susi, 1986). Bands associated with other conformations have not been assigned with certainty. It appears, however, that this analysis is only extendable to proteins of certain structural classes. For example, it may not adequately determine conformation quantitatively with proteins with high helix content (Byler and Susi, 1986). The information derived from these analyses make use of only a fraction of the component peaks in a deconvoluted amide I

peak. Further extraction of structural information from this band is limited by the lack of adequate spectral-structural correlations (Byler and Susi, 1986, Surewicz and Mantsch, 1988a).

Recent studies have examined the use of the conformation-sensitive amide III region (1310-1215 cm^{-1}) for protein conformation analysis (Anderle and Mendelsohn, 1987; Kaiden et al., 1987). Use of the amide III region, compared to the amide I region, has several advantages: 1) the frequency range is somewhat greater than the amide I, 2) there is no strong absorbance from water in this region, and 3) it appears to be about three times as sensitive to conformation as the amide I band (Anderle and Mendelsohn, 1987). However, correlations between features in this region with structural components appear much more complex. In addition, vibrational analysis indicates that the amide vibration responsible for this mode is often strongly coupled with other non-peptide skeletal modes (Miyazawa, 1962). Also, it sometimes appears that a single conformation can produce several discrete bands (Anderle and Mendelsohn, 1987)

Several experimental attempts have been proposed recently to produce spectral-structural correlations with the amide I band. Two of these (Olinger et al., 1986; Haris et al., 1986) use the small shift in absorbance maxima of the amide I band components upon exchange of the amide hydrogen for deuterium (approximately 5-10 cm^{-1}). It is argued that because different secondary structures exchange at different rates due to differences in hydrogen bonding characteristics, bands can be assigned to secondary structures by observing exchange kinetics in the deconvoluted amide I band. In spite of limited success, proton-deuteron exchange is not specific for conformational type alone. Solvent accessibility, in addition to secondary structural type, affects exchange rates (Englander et al., 1972). Another attempts to assign bands has involved

monitoring the effect of non-aqueous denaturing solvents on proteins (Wasacz et al., 1987). However, the effects on structure of non-aqueous solvents on protein structure are not well characterized. The investigators could only extrapolate from information based on interactions of model homopolymers with non-aqueous solvents.

In order to extract quantitative information concerning conformation from IR spectra it is also necessary to determine the molar absorptivity of the individual structural components as these are not necessarily equal (Mantsch et al., 1989). These values can, in theory, be computed from normal mode analyses (Krimm and Bandekar, 1986), but these calculations are very lengthy and not directly applicable to proteins for reasons discussed previously. Further, values of molar absorptivities determined experimentally for ideal homopolymers are not directly transferable. Several attempts have been made to determine molar absorptivities of the various protein conformations experimentally. Lack of a good experimental system has hindered progress towards this goal. However, Byler and Susi (1986) have demonstrated good correlation between the relative intensities of amide I bands assigned to helices and extended strands and the fractional composition of these conformations based on a procedure for automatic secondary structure analysis (that of Levitt and Greer, 1977). Chapman has examined the molar absorptivities of poly(l-lysine) in several different conformations by inducing conformational transitions by varying the pH. However, it is not known how the environmental differences nor the ionization of side chains in the molecule affect the absorptivity. Mantsch (1989) employed a similar strategy by inducing conformational transitions in proteins with trifluoroethanol which has been demonstrated to induce the α-helical conformation in polypeptides. As with pH induced transitions, the nature of the environment and the newly formed conformations are not well known. Although

these studies can only provide inexact estimates, they indicate that the molar absorptivities of the various conformations studied are roughly equal.

2.2. DESCRIPTION AND CLASSIFICATION OF PROTEIN CONFORMATION

The first protein structure of myoglobin was solved crystallographically by Kendrew (Kendrew et al., 1958). This structure revealed a surprising degree of complexity and lack of symmetry in the three-dimensional conformation of the protein. Since then, much attention has concentrated on describing the conformations and structures which make up proteins. Not surprisingly, the greatest portion of protein structures were found to comprise the classical conformations, the α-helix and β-pleated sheet, first proposed by Pauling (Pauling and Corey, 1951; Pauling et al., 1951) and later observed in polypeptides (Richardson, 1981). Although conformations resembling the classical secondary structures account for most of a protein's conformation, other non-classical secondary structures make up a significant portion of known globular protein structures. The next major class is reverse turns. First described by Venkatachalam (1968), reverse turns represent a conformation which allows the polypeptide chain to reverse direction by approximately 180°. These structures are found to account for the conformation of about 30% of residues in proteins of known structure (Zimmerman and Scheraga, 1977; Crawford et al., 1973), and occur mainly at the surface (Kuntz, 1972). Several types of turns exist and have been described by their constituent backbone dihedral angles and their hydrogen bonding pattern. β-Turns, of which several forms have been observed, have a hydrogen bond interaction between the i-th and i + 3 residues. γ-Turns, another class of reverse turns, have a hydrogen

bond interaction between the i-th and i + 2 residues (Smith and Pease, 1980; Rose et al., 1985)

For purposes of correlations with spectroscopic and other physicochemical data, there has been much interest in development of analytical and objective methods for defining and quantitating protein structural elements. Several such methods for automatic secondary structural analysis from atomic coordinates have been proposed. Levitt and Greer (1977) have described a template matching method which searches a structure database for conformations based on linear distance, backbone dihedral angle and hydrogen bonding parameters. This algorithm is based only on a three state representation of protein conformation (i.e.α-helix, β-strand, random coil). Kabsch and Sander (1983) have compiled a 'dictionary of protein secondary structure' based on a similar algorithm which includes hydrogen bond and geometric features. However, they have expanded the structure model to account for other conformations (e.g., reverse turns, etc.) and imperfections in the classical secondary structures.

Other algorithms for protein conformational analysis from atomic coordinates have been proposed. These algorithms employ a common approach which consist of defining a structure template based on α-carbon distances (Chou and Fasman, 1977; Crawford et al., 1973; Lewis et al., 1971; Levitt and Greer, 1977; Kolaska et al.,1980; Kuntz, 1972; Ramakrishnan and Soman, 1982; Richards et al., 1988), α-carbon torsion angles (Levitt and Greer, 1977; Kuntz, 1972; Ramakrishnan and Soman, 1982), phi-psi angles (Chou and Fasman, 1977; Crawford et al., 1973; Hohne and Kretschmer, 1985), hydrogen bonding (Chou and Fasman, 1977; Crawford et al., 1973; Hohne and Kretschmer, 1985; Kabsch and Sander, 1983) and other parameters, or some combination of these. These algorithms are methodologically similar in that templates are

constructed for each conformation of interest based on its particular parameters. The search for these conformations in proteins is essentially a pattern recognition process. Variations exist within each conformational class and somewhat arbitrary mismatch limits need to be set. This is termed the template approach.

While most conformations can be described by a series of structural parameters, the growing body of knowledge concerning protein secondary structures has shown that there exist other classes of conformations previously classified as 'random coil' (Fetrow et al., 1988). The commonly used term 'random coil' is a misnomer. Such structures are not statistically derived random structures, nor are they necessarily coils. They appear to be as highly organized and as firmly held as other structures, but have been difficult to describe with traditional secondary structure classification schemes (Richardson, 1981). Rose has identified a class of irregular conformations termed Ω-loops as these structures typically resemble the Ω symbol (Lesczynski and Rose, 1986).

The template approach has done well at locating and quantitating protein structural elements within the crystal structures. These conformations are found to comprise approximately 80% of residues in proteins of known structures (Lesczynski and Rose, 1986). This approach, however, is inherently limited in its ability to describe protein conformation in that in only considers conformations for which a template has been constructed. It does not provide any description or classification of conformations which do not match the template. This excludes a significant portion of protein conformations from consideration. These remaining conformations are classified only as random coil or unordered conformations. Further, template-based algorithms are subject to uncertainties in defining the boundaries of the regions of

conformations as well as the degree to which similar, but slightly different conformations are included in a given structure class. As such, new algorithms are necessary which can not only objectively locate and classify the known secondary structures, but incorporate other regions of conformation whose folding patterns are less regular.

2.3 METHODS

2.3.1. COMPUTATIONAL STUDIES OF PROTEIN CONFORMATION

Two geometrically-derived parameters, the Linear Distance (LD) value and the Backbone Dihedral Angle (BDA), will be used to represent the protein backbone conformation from the α-carbon coordinates. The positions of the α-carbon atoms are determined with the highest accuracy in crystallographic experiments (Blundell and Johnson, 1976). Thus, conformational descriptors which are derived from α-carbon coordinates will be the least succeptible to error. Both of these descriptors are described in detail below. The combination of these two descriptors provides a unique description of the protein backbone folding.

2.3.1.1. Linear Distance Value

The linear distance representation has been described in detail previously (Liebman et al., 1985; Liebman, 1986). Briefly, the LD value of each amino acid, I, in a protein sequence is computed by summing the series of distances from it's own α-carbon to each of four successive α-carbons as is illustrated in the following equation.

$$LD_i = \sum_{j=1,4} (d_{i,i+j}), \qquad (1)$$

where **d** is the through-space distance between the α-carbon coordinates. This value is calculated beginning at the N-terminal residue of the protein of interest. Each Linear Distance value describes the conformation of five residues. For a protein of N amino acids there are N-4 LD values. The linear distance calculation is shown schematically in Figure 3-1. A plot of the Linear Distance versus sequence position yields a detailed profile of local folding (see Figure 3-2).

2.3.1.2. Backbone Dihedral Angle

The Backbone Dihedral Angle of each amino acid, **I**, is computed as follows. For four α-carbons designated, i, i+1, i+2 and i+3, two planes are readily defined, the first containing i, i+1 and i+2 and the second by i+1, i+2 and i+3. The BDA is the dihedral angle between these two planes. This descriptor has also been termed the 'α-torsion angle' or the 'virtual bond angle' (Ramakrishnan and Soman, 1982). The backbone dihedral angle calculation is shown schematically in Figure 3-1. One BDA value describes the orientation of 4 residues. This descriptor provides a description of the handedness or direction of the polypeptide fold. As an example, a right-handed and

left-handed α-helix will yield the same linear distance values but their backbone dihedral angles will be the same magnitude but opposite in sign.

LD and BDA values have been calculated for a series of ideal periodic and aperiodic conformations (Tables 3-2 and 3-3). These two descriptors, when considered together, provide a unique representation of the conformation of the protein backbone. In addition, these values, as opposed to the actual coordinates, provide a simple, direct means of comparing conformations in proteins. As such, these values have been used in development of an algorithm which generates a library of protein substructures from crystallographic coordinates. The development and results of this algorithm are the basis for Chapter 3 of this thesis. Table 3-4 lists the comparison of LD and BDA values among the various ideal conformations.

2.3.1.3. Structural Superpositioning

Structural superpositioning provides a method for estimating the structural similarity between two sets of coordinates, $[X_1]_A$ and $[X_1]_B$. The best transformation of one molecule to another is defined as that which minimizes the root-mean-square difference (rmsd) between two sets of equivalent atomic positions as in the following equation:

$$RMSD = \left(\sum ([X_1]_B - [X_1]_{A'})^2 / n\right)^{1/2} \qquad (2)$$

This is accomplished by transforming the orthogonalized center-of-gravity coordinates of A onto the orthogonalized center of gravity coordinates of B by a

series of independent rotations and translations. The transformation is performed by first transfering the center-of-gravity of the respective coordinate sets and then applying a minimization procedure to obtain the rotation matrix that achieves the best superposition of the equivalenced atomic coordinates. This method has been shown to be useful in establishing the similarity of two molecules at low resolution. In higher resolution analysis of detailed differences between two structures the superpositioning technique is less successful because its metric is a single statistic that is averaged over the entire set of equivalenced atoms. In these studies, structural superpositions were performed according to the method first described in Cox (1967) and later modified as described in Liebman et al., (1985) to examine the distribution of deviation values as opposed to a single averaged statistic..

2.3.2. SPECTROSCOPIC DATA COLLECTION AND ANALYSIS

2.3.2.1. Data Collection

For IR spectroscopy protains were prepared as 3.0% (w/v) solutions (1.3 mM) in 20 mM acetate (pD=5.0) or 20 mM imidazole (pD=6.9) buffers made up in D_2O. pD was determined by adding 0.4 to the pH reading (Covington et al., 1968) measured using a Horiba Cardy glass electrode pH meter. The spectrum of each protein was collected at the same pH (pD in this case) as the crystal structure determination. Sufficient $CaCl_2$ (1M in D_2O) was added to make the final solution 20 mM in Ca^{++}. Solutions were placed in IR cells with CaF_2 windows and teflon spacers with a pathlength of 75 μm.

IR spectra were collected at ambient temperature using a Nicolet 740 SX FTIR system equipped with a water-cooled Globar source, a Ge-coated KBr

beam splitter and a broad range mercury/cadmium/telluride detector. All spectra were recorded at a resolution of 2 cm^{-1} by coadding 4096 double-sided interferograms which were Fourier-transformed after application of a Happ-Genzel apodization function. The spectrometer and sample chamber were continuously purged with dry nitrogen.

For each lyophilized protein sample, except when used in time course studies, three separate solutions were prepared and their spectra collected at different times, often several days apart. Further, each of these spectra was analyzed independently. All spectra were collected as soon as possible after mixing. The time required to load the cell, transfer it to the spectrometer and purge the sample chamber was approximately 30 min. Collection time for a single spectrum of 4096 coadded interferograms was also 30 min.

2.3.1.2. Solvent and Water Vapor Correction

Spectral contributions from residual H_2O vapor in the light path and from buffers were subtracted using programs provided with the Nicolet FTIR software, Version 4.3. Factors for water vapor subtraction were determined by subtracting a second derivative spectrum of water vapor from the second derivative spectrum of the sample. The subtraction factor was varied until the region from 1700-1800 cm^{-1} was featureless. Subtraction using second derivative spectra is preferable to the original absorbance spectra because sharp vapor lines, which are often invisible in original spectra, are amplified by derivatization. This results in a more reliable correction for water vapor.

2.3.1.3. Derivative Spectroscopy

Derivative spectroscopy, calculation of second, fourth and higher order derivatives of infrared spectra, allows visualization of component bands in overlapped peaks. Derivative spectra can be calculated analytically (Maddams and Tooke, 1982) or through manipulations in the Fourier domain. For an ideal Lorentzian band, the second derivative band has a bandwidth which is reduced by a factor of 2.7 (Maddams and Tooke, 1982). However, artifacts may be present due to interactions of positive and negative lobes of adjacent bands or from uncompensated water vapor bands (particularly when studying the amide I and amide II bands). Further, calculation of the second derivative enhances the noise present relative to the absorbance bands and noise features can often resemble true absorbance bands. Thus, care must be taken to avoid introduction of artifacts by collecting spectra with a sufficiently high signal-to-noise ratio and by elimination of water vapor absorbances.

2.3.1.4. Fourier Self-Deconvolution

Fourier self-deconvolution functions by performing mathematical manipulations in the Fourier domain. Upon performing the inverse Fourier transform these manipulations result in a reduction of the bandwidths of component bands in a given spectral envelope, resulting in increased visual resolution (see Kauppinen et al., 1981). The investigator must optimize two parameters to obtain maximal deconvolution of the spectral region of interest. First, an estimate must be made of the bandwidth of the component bands in the original spectrum. Values ranging from 10 to 25 cm-1 have been used for the amide I region of proteins with success (Byler and Susi, 1986; Susi and Byler,

1986; Surewicz and Mantsch, 1988). Second, a resolution enhancement factor must be estimated. The best choice of estimate for the original bandwidth depends on the actual bandwidths (which because of overlap can not be determined). Choosing too large a bandwidth results in overdeconvolution which results in negative going features and side lobes. Choosing too low a value for the original bandwidth results in less than maximal deconvolution. The maximal value for the resolution enhancement factor, the ratio of the original bandwidth to the deconvoluted bandwidth, depends on the signal-to-noise ratio of the original spectrum. Choice of too high a factor results in the introduction of periodic noise in the spectrum which can alter the intensities and positions of spectral bands as well as produce spurious peaks. Thus, the resolution enhancement factor is optimized by starting with a low value and increasing it until periodic noise appears in regions of the spectrum where no absorbances exist. Although Fourier self-deconvolution does alter the original bandshape, if one chooses these two parameters judiciously, so as to avoid periodic noise and other artifacts, the band area will be unaffected. Thus, quantitative information is, in principle, obtainable from deconvoluted spectra and the use of iterative curve-fitting techniques with resolution-enhancement methods has been demonstrated to be a powerful tool for infrared spectroscopic investigations of molecular structure (Griffiths et al., 1987) and particularly for protein secondary structure (Byler and Susi, 1986). The application of these methods to protein conformation analysis from conformation-sensitive infrared bands has allowed investigators to overcome some of the limitations inherent in studying infrared spectra of proteins and other biological macromolecules (Byler and Susi, 1986; Mantsch et al., 1986; Surewicz and Mantsch, 1988a).

2.3.1.5. Curve-Fitting

Iterative, least-squares curve fitting can be used to quantitate the information in a given spectral envelope. The achievement of a good representative fit requires the knowledge of the number of distinguishable bands and an assumption about the shape of these bands. By use of an iterative procedure that minimizes the sum of the squares of the differences between the curve-fitted spectrum and the data, appropriate initial estimates of the individual positions, heights and widths of the bands are varied to arrive at a best fit. The most important values for input to the curve-fitting routine are the number of bands and their positions. Curve-fitting is most useful when the individual band are well resolved (Pierce et al., 1990). Thus, before application of curve-fitting, an estimate of the number and positions of bands which comprise the spectral envelope and a method for resolving overlapped bands. The former is provided by derivative spectroscopy and the latter through Fourier-self deconvolution, as described above. Thus, prior to the application of curve-fitting algorithms, the derivative spectra are calculated to provide and estimate of the number and positions of bands in the spectral region of interest. Additionally, Fourier self-deconvolution is performed to resolve overlapped bands. Curve-fitting is then applied to the deconvoluted spectra as described in (Pierce et al., 1990).

Table 2-1. Characteristic Amide Vibrational Modes Associated with Peptide Group

Designation	Frequency	Description
A	3300 cm^{-1}	N-H stretch (Fermi resonance with
B	3100	2 x Amide II 2 X Amide II)
I	1650	C=O stretch
II	1560	C-N stretch, N-H bend
III	1300	C=O stretch, C-N stretch, N-H bend, O=C-N bend
IV	625	O=C-N bend
V	725	N-H bend
VI	600	C=O bend
VII	200	C-N torsion

CHAPTER 3

GENERATION OF A SUBSTRUCTURE LIBRARY FOR THE DESCRIPTION AND CLASSIFICATION OF PROTEIN CONFORMATION

3.1. INTRODUCTION

Since the first protein crystal structure was solved by Kendrew (Kendrew et al., 1958) there has been much interest in the description and classification of protein backbone conformation otherwise called secondary structure. Compared with peptide homopolymers, protein conformation is complex and often adopts seemingly unordered folding patterns. Additionally, proteins usually contain several regions of differing conformational types as well as regions of aperiodic or irregular structure. Overall, the largest portion of protein conformation exists in conformations similar to the classical secondary structures, the α-helix and β-strand, first proposed by Pauling and observed in silk fibers (Pauling and Corey, 1951; Pauling et al.,1951). Subsequently, reverse turns, proposed by Venkatachalam (1969) for short peptides, have been observed in proteins (Chou and Fasman, 1977). Additionally, Rose has introduced the nomenclature W-loop to describe a class of irregular, loop-like

conformations (Leszczynski and Rose, 1986). These four broad classes - α-helix, β-strand, reverse turns and loops- account for approximately 80% of the residues in proteins as calculated by Rose (Lesczynski and Rose, 1986). All remaining backbone folding patterns have been considered, by subtraction, to be random coil or undefined structure.

Several investigators have developed algorithms for automatic description and classification of the various protein conformations, typically for purposes of correlation with amino acid sequence or for determination of secondary structure by experimental methods, such as spectroscopy. These algorithms typically use predefined structure templates, usually based on physical or geometrical parameters, representing the various ideal secondary structures to search an atomic coordinate database for conformations similar to the template. This form of conformational searching is termed the template approach. Investigators have proposed secondary structure templates based on such parameters as α-carbon distance (Chou and Fasman, 1977; Crawford et al., 1973; Lewis et al., 1971; Levitt and Greer, 1977; Kolaska et al.,1980; Kuntz, 1972; Ramakrishnan and Soman, 1982; Richards et al., 1988), phi-psi angles (Chou and Fasman, 1977; Crawford et al., 1973; Hohne and Kretschmer, 1985), α-carbon torsion angles (Levitt and Greer, 1977; Kuntz, 1972; Ramakrishnan and Soman, 1982), hydrogen bonding patterns (Chou and Fasman, 1977; Crawford et al., 1973; Hohne and Kretschmer, 1985; Kabsch and Sander, 1983) and other parameters. This form of search for conformations similar to the template in actual proteins involves a pattern recognition of pre-established template based on one or more conformational parameters. Conformations in proteins are not uniform and thus, classification requires determination of criteria for matching which are somewhat arbitrary. The result of applying such

algorithms is an outline of the protein sequence in terms of its secondary structure elements based on these templates.

The template-based methods described above succeeded well at locating template-like conformations in proteins. However, by their nature, these methods are limited to classification among these pre-determined conformations. The limitations of the template approach are two-fold: 1) a significant portion of the protein backbone adopts non-template-like conformations which have been previously described only as 'random coil'; and 2) the classical secondary structures, or templates, in proteins may contain one or more subtypes which are structurally or functionally similar. The term 'random coil' has been described by Richardson (1981) as a double misnomer being neither random nor necessarily coils.

Described here is a method for classification of protein conformation, from the α-carbon coordinates, which operates independent of a pre-defined structure template. Two geometrical parameters, α-carbon distances and α-carbon torsion angles, are used to represent the protein conformation from atomic coordinates. As opposed to searching a coordinate database for particular types of secondary structure based on a template, this method functions by identifying regions of similar conformation among the proteins within a data set. The result of this analysis is a 'library' of protein substructures observed in the structures studied. Because the algorithm functions independent of a template, this method can describe secondary structural patterns or conformations independent of its complexity of folding. The algorithm automatically gives a description of how secondary structure fragments are related among proteins. Further, it provides a higher level of description of protein conformation because it can classify conformations previously referred to as undefined or random structures as well as sub-classify

conformation classes previously described only as a single class. Presented here is a validation of this method, a review of the results of its application and a comparison with those of template-based algorithms.

3.2. DATA

3.2.1. COORDINATE DATA

All α-carbon coordinates for proteins were obtained from the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977). A set of 14 proteins were chosen from the PDB for purposes of refinement of the comparison parameters. A list of these proteins and their reported atomic resolutions are provided in Table 3-1. Only structures with a resolution better than 2.5 Å were used. Approximately equal numbers of α, β, α+β and α/β proteins, as classified by Levitt and Chothia (1976), were chosen. Additionally, these 14 proteins were chosen such that none shows extensive structural homology to any of the others in the set. Coordinates for ideal homopolymer conformations were generated using the program Quanta (Polygen, Waltham, Massachusetts) using standard phi-psi angles for the classical secondary structures (Richardson, 1981) and reverse turns (Smith and Pease, 1980)

3.3. METHODS

3.3.1 STRUCTURE DESCRIPTORS

The algorithm uses two geometrically-derived parameters, the Linear Distance (LD) value and the Backbone Dihedral Angle (BDA) to represent the backbone conformation from the α-carbon coordinates.

3.3.1.1. Linear Distance Value

The linear distance representation has been described in detail previously (Liebman et al., 1985; Liebman, 1986). Briefly, the LD value of each amino acid, **I**, in a protein sequence is computed by summing the series of distances from it's own α-carbon to each of four successive α-carbons as is illustrated in the following equation.

$$LD_i = \sum_{j=1,4} (d_{i,i+j}), \qquad (1)$$

where **d** is the through-space distance between the α-carbon coordinates. The LD calculation is shown schematically in Figure 3-1. This value is calculated beginning at the N-terminal residue of the protein of interest. Each Linear Distance value describes the conformation of five residues. For a protein of N

amino acids there are N-4 LD values. A plot of the Linear Distance versus sequence position yields a detailed profile of local folding (Figure 3-2).

3.3.1.2. Backbone Dihedral Angle

The Backbone Dihedral Angle of each amino acid, I, is computed as follows. For four α-carbons designated, i, i+1, i+2 and i+3, two planes are readily defined, the first containing i, i+1 and i+2 and the second by i+1, i+2 and i+3. The BDA is the dihedral angle between these two planes. This descriptor has also been termed the 'α-torsion angle' or the 'virtual bond angle' (Ramakrishnan and Soman, 1982). One BDA value describes the orientation of 4 residues. This descriptor provides a description of the handedness or direction of the polypeptide fold. Calculation of the BDA is shown schematically in Figure 3-1.

LD and BDA values have been calculated for a series of ideal periodic and aperiodic conformations (Tables 3-2 and 3-3). These two descriptors, when considered together, provide a unique representation of the conformation of the protein backbone. In addition, these values, as opposed to the actual coordinates, provide a simple, direct means of comparing conformations in proteins. Table 3-4 lists the comparison of LD and BDA values among the various ideal conformations.

3.3.2. STRUCTURE MATCHING ALGORITHM

In order to compare fragments of conformation (contiguous backbone chain segments of defined length) to determine if they are equivalent we have developed a structure matching algorithm based on similarity in the LD and

BDA values. The algorithm functions as follows. First, LD and BDA values are computed for each protein in the set and stored as a separate file for each protein. All fragments of a fixed length in a protein chain are then compared with all other fragments of that same length, using their LD and BDA values. A cost function based on a linear combination of differences in the linear distance values and backbone dihedral angles is used to evaluate equivalence between any two fragments. (The cost function is described in detail below.) All fragments are considered sequentially by shifting the window of computation by one residue. Thus a chain of N residues will have $N - L_f + 1$ fragments where L_f is the fragment length in residues. All fragments which are determined by the cost function to be equivalent to the test fragment are recorded in a file. This procedure is performed for all fragments in the protein set under study. The result is a file for each fragment listing all other fragments in the data set which have equivalent conformation. The complete set of fragment lists contains multiple redundancies. To put this information into a more usable form, these lists are condensed into files which list all fragments in the protein set which are judged to be of equivalent conformation.

A substructure, by this definition, is any such conformation which is observed repeatedly in the set of proteins studied. This constitutes an internal versus an external definition of protein substructures. A protein substructure is defined in terms of a specific, repeated folding pattern observed internally within a structure database rather than by similarity to a template defined externally to that data set.

3.3.3. COST FUNCTION

A cost function has been developed for comparing fragments of conformation which includes terms for the differences in both the LD and BDA values. The following equation describes the cost function for comparing two fragments, I and J, 5 residues in length.

$$Cost_{IJ} = C1 \times (|LD_i - LD_j|) + C2 \times (\tan|BDA_i - BDA_j|) + C2 \times (\tan|BDA_{i+1} - BDA_{j+1}|) \quad (2)$$

To bring the differences in BDA values into a common range with that of differences in LD values the tangent of the BDA difference is computed. The convenient properties of the tangent function (tan θ) are its being approximately linear at small values of θ and its exponential increase at larger values of θ. Two fragments are considered to be of equivalent conformation if the value of the cost function is less than 2.0.

Fragments sizes greater than 5 residues can also be considered. To compare fragments greater than 5 residues in length, the cost function is summed over a moving window, shifting the window one residue. The maximum value for conformational equivalence for fragments longer than 5 residues is then calculated using the following equation.

$$Max = (L_f - 4) \times maxdif \times 0.75 \quad (3)$$

where L_f is the fragment length and **maxdif** is the maximum value of the cost function for each corresponding 5 residue fragments, 2.0 as stated above. The factor 0.75 is included to avoid a series of marginal equivalences in these longer fragments. For this study, a fragment length of 8 residues was used.

3.3.4. STRUCTURAL SUPERPOSITIONS

Structural superpositions of equivalenced fragments were performed using the method first described in Cox (1967) and later modified as described in Liebman et al., (1985). For fragments of equivalent length, similar atoms were equivalenced in superpositioning.

3.3.5. CALCULATIONS

The algorithms described here are coded as FORTRAN programs. All computations have been performed on a Microvax II and a VAX 11/785, operating under the VMS operating system.

3.4. RESULTS AND DISCUSSION

3.4.1. REFINEMENT OF THE COST FUNCTION PARAMETERS

The cost function described in the methods (Equation 2) section can be made more or less stringent by adjusting the values of the coefficients of the terms representing the differences in the LD and BDA values, C1 and C2 in

equation (2), respectively. The relative weighting within the cost function of the differences in the LD and BDA values can be altered similarly. The first step in computing a substructure library is to refine the values of C1 and C2 so that fragments determined to be equivalent by the cost function are truly equivalent within an empirically observable natural variability among conformationally equivalent fragments. These refinement procedure for these coefficients proceeded as follows.

A tentative substructure library was generated using the initial values C1=1.0 and C2=1.0. Two criteria were used to determine if this library was sufficient. First, all equivalent fragments in each substructure are structurally superimposed with all other fragments contained in that substructure and the root-mean-square deviation (RMSD) recorded. Table 3-5 lists in matrix form the root-mean-square (RMS) differences in atomic coordinates upon superpositioning ideal conformations of eight residues listed in tables 3-2 and 3-3, and the distribution of these values is plotted in Figure 3-3. Based on these values we chose an RMS difference value of 1.0 Å as the upper limit for conformationally equivalenced fragments. Except for very similar ideal conformations, such as parallel and antiparallel β-strands, all RMSD values fall well above 1.0 Å. For a generated substructure library the RMS deviation should be no greater than 1.0 Å for more than 95% of conformationally equivalenced fragments. The RMSD provides a single, average statistical measure of the structural similarity among two fragments. It is possible that two fragments which are similar throughout much of the structure with a single large deviation can yield an RMSD value which is acceptable by this criteria. Thus, the additional constraint that the maximum difference in any of the equivalenced coordinates be no greater than 1.5 Å was also imposed .

An obvious criterion for evaluating the suitability of a library is that the overall substructure library list for the α-helix substructure should agree with those of other investigators. Two commonly used algorithms of automatic secondary structure analysis were chosen for comparison, that of Levitt and Greer (1977), and that of Kabsch and Sander (1983). The α-helix was chosen because it has been well studied and, among globular proteins, is the most rigid and dimensionally well defined of the classical secondary structure elements (Richardson, 1981). A similar comparison was not done for the β-strand because of its well documented intrinsic variability (Richardson, 1981). In addition, as will be demonstrated, at the level at which the classification of α-helices agrees with other methods, application of this algorithm subclassifies the β-strands into several components.

If a tentative substructure library did not meet these criteria, the coefficients C1 and C2 were adjusted and a new library generated. When satisfactory values were reached each value was then altered in either direction to determine if this improved the library with respect to the two criteria. The final range of values tested was 0.6 to 1.0 for C1 and 1.0 to 2.2 for C2, at intervals of 0.1.

The values C1=0.9 and C2=2.0 were found to best satisfy the cost function criteria outlined in the methods section. Adjusting these values in either direction resulted in poorer results based on the two criteria. Increasing the value of either one or both of these coefficients resulted in a lower portion of equivalenced fragments which did not meet superpositioning criteria, as would be expected. However, increasing these values also resulted in poorer agreement with α-helix boundaries of other studies. More specifically, the helices delineated by the substructure library calculations were shorter than those of the two previous investigations suggesting that the coefficients were

too stringent. 95.9% of conformationally equivalenced fragments met structural superposition criteria. Agreement with other investigators concerning the α-helix substructure was such that disagreement of these results with any single investigator was no greater than the disagreement among the two previous investigations Table 3-7 lists the α-helices delineated by the two template based methods and the substructure library for two proteins, carboxypeptidase A and triosephosphate isomerase.

3.4.2. SUBSTRUCTURE CLASSIFICATIONS

The results of a substructure library generation for the 14 non-homologous proteins using the refined cost function are reported here. Each individual substructure can be represented by the mean values of each of the 4 LD and 5 BDA values which represent a conformational fragment of 8 residues. By comparison of these values with those of the ideal secondary structures, substructures which correspond to these ideal conformations (i.e., helices, turns, etc.) can be identified. In this manner, 113 substructure classes were generated from 2378 fragments, from the 14 non-homologous proteins using the refined cost function. The 30 most frequently occurring substructures are described in Table 3-6.

Generation of a substructure library for the 14 non-homologous proteins listed in Table 3-1 resulted in description of the conformation of 67% of the residues among these proteins. That is, 67% of the residues among the 14 proteins occur in some substructure. Unlike the template based algorithms, the degree to which the algorithm employed here can describe and classify regions of conformation is dependent on the number and nature of proteins. To further examine this point, two additional libraries were generated. A library was

generated for a series of 20 proteins which included the original 14 proteins with the addition of 6 proteins which were not homologous to any of the original 14. Addition of these 6 proteins resulted in the percentage of residues whose conformation is described rising to 77%. Additionally, a substructure library was also generated for a series of homologous proteins, the trypsin-like serine proteases. For the 10 proteins in this data set the percent description was 96%. Thus, as would be expected intuitively, both addition of proteins to the data set and homology among members of the data set results in a greater portion of the residues occurring in substructures.

Detailed output of the results can be arranged in two formats. One format contains a list of all sequence positions of a given substructure for every protein in the set. One such list is created for each substructure classification. This is a convenient format for structural comparisons among proteins. Another format lists for a given protein from the set, the positions of all the substructure classifications which have been observed in that protein. One such list is compiled for each protein. This format is useful for study of a protein in terms of its constituent substructures.

3.4.2.1. α-Helices

The most frequently occurring substructure is that which is classically defined as the right handed α-helix. There are 321 occurrences of 8 residue α-helices among the 14 non-homologous proteins. The mean LD and BDA values of this substructure are essentially equal to those for an ideal α-helical homopolymer. Several occurrences of α-helices shorter than 8 residues are also observed as parts of other substructures. Several investigators have previously assigned α-carbon distance and backbone torsion angle parameters for α-helices (Levitt

and Greer, 1977, Ramakrishnan and Soman, 1982). Along these lines, a substructure is considered α-helical if its mean LD value is 20.7 Å +/- 0.9 Å and if the mean BDA values are 49.0° +/- 10°, with the total deviation of two BDA values from the ideal not exceeding 15. 41 additional substructures contain α-helical portions. These substructures consist predominantly of α-helix with the N-terminus or C-terminus adopting varying conformations. These results agree well with other studies which show the α-helix to be the predominant backbone conformation in globular proteins (Richardson, 1981).

It is of interest to examine the substructures which contain portions of α-helices. Several examples are listed in Table 3-6 (Substructures 2, 3, 4 and 5 are examples). These substructures provide a description of the folding patterns which precede and follow helices. Whereas template based methods usually define helices with specific endpoints, this method, by virtue of its independence from a template, is able to describe and classify those folding patterns which include the regions bounding them. These results demonstrate that some portion of these regions can be separated into discrete sub-classes.

3.4.2.2. Extended strands

The extended or β-strand has ideal parameters of LD = 33.25 Å and BDA = 177.80° for the parallel β−sheet configuration and LD= 34.93 Å and BDA=178.87° for the antiparallel configuration in ideal β-sheet forming homopolymers (Table 3-2). However, as has been observed previously this conformation is highly variable in globular proteins (Richardson, 1981). Several investigators have previously assigned α-carbon distance and backbone torsion angle parameters for extended strands (Levitt and Greer, 1977, Ramakrishnan and Soman, 1982). Along these lines we use the

following parameters to categorize extended strands; mean values of the linear distance must be greater than 31.0 Å and mean values of the BDA must be between 130° and 180° or between -130° and -180°.

Several examples of β-strands 8 residues in length are observed in the substructure library for the 14 non-homologous proteins. There are 16 such substructures. Each of these substructures differs in its degree of extendedness and degree and orientation of curvature in it's folding. The observed 8 residue extended strands appear to fall into two categories. The first consists of a pure β-strand. That is, all linear distance values are > 31.0 Å and all backbone angle values are within the defined boundaries. 4 of the 16 are in this class. In the second class, all residues are in an extended conformation but there exists a kink or bend. This is characterized by a single backbone angle value outside the range flanked by those within the range. This is analogous to a kink or bend in a helix where all residues are in a helical conformation but the entire helix is not contiguous. Shorter segments of extended strand are also observed within as parts of other substructures. The same criteria are used to locate parts of substructures which are extended strand. 31 additional substructures contain portions of extended strand.

As stated previously several investigators have noted the variability of the β-strands among proteins (Richardson, 1981; Richards and Kundrot; 1988). However, these authors have treated the β-strand as a single class. An advantage of an algorithm such as the one used here is the ability to automatically sub-classify this highly variable conformational class. It does so without the necessity of providing a template *a priori*. The algorithm, in essence, identifies the 'template' as well as the classification.

To further illustrate this point, it is of interest to examine 3 classes of β-strand identified in this library. Substructures 11 (see Table 3-6), 59 and 60 are

observed only in the protein concanavalin A (2CNA). These classes of extended structure are highly extended, marked by high LD values (>35.0 Å) and BDA values close to 180°. Thus, this algorithm provides the capability to define and classify discrete classes of β-strand. This is analogous to the classification of β-sheets as β-barrel, parallel and antiparallel β-sheet.

3.4.2.3. Other substructures

In addition to substructures containing α-helices and extended strands, the algorithm recognizes substructures containing reverse turns and structures which fall into no previously defined class of structures. Substructure 25 in Table 3-6 is an example of a substructure containing a reverse turn. Substructure 27 is an example of a conformational fragment whose folding pattern does not fall into a class which has been previously defined. Indeed, the independence of this algorithm from a template results in a significantly higher degree of description of protein conformation. Another interesting class are those which contain segments of both of the two major classes of conformation, α-helices and extended strands. Substructures 19 and 25, in Table 3-6, provide two such examples. This demonstrates a not so obvious advantage of this method; it can recognize not only the predominant conformations, the α-helices and extended strands, but also the ways in which they are connected. The template methods usually define only the endpoints of such conformations.

3.5. CONCLUSION

Traditional analysis of secondary structure in proteins consists of defining a structure template from a series of ideal conformations and then searching a structure database of structures for regions which are similar, within preset limits, to one of the template structures. This results in a less than adequate description of protein conformation. Template -like structures, usually based on structures observed in model homopolymers, account for only a fraction of total conformations observed in proteins of known three-dimensional structure. The remainder of conformations which do not resemble a template structure are considered, by subtraction, to be undefined or random coil.

A method has been developed for automatic classification of protein conformation which operates independent of a predefined template. This method has the ability classify and define protein conformation based on an internal rater than external definition. That is, protein structure elements, substructures, are defined on the basis of observation in proteins as opposed to similarity to an externally defined template. Such a template-independent basis for definition and classification has several advantages. First, it has demonstrated the ability of this method to reproduce the findings of the template-based algorithms and expanded on their interpretation. Further, this method has the ability to subclassify a class of conformations previously grouped into a single class as well as define and classify conformations previously described as undefined or random conformations. The result of these advantages is an enhanced description of conformations which exist in proteins, often classifying the conformation of greater than 95% of residues.

Table 3-1. Four-letter Codes for 14 Non-homologous Protein Structures Used in this Study.

Code	Protein	Source	# AA[1]	Å[2]	Class[3]
1MBN	Myoglobin	Sperm whale	149	2.0	α
1RN3	Ribonuclease A	Bovine pancreas	124	1.45	α+β
1TIM	Triosephosphate Isomerase	Chicken muscle	247	2.5	α/β
2CAB	Carbonic Anhydrase	Human erythrocytes	261	2.0	α/β
2CNA	Concanavalin A	Jack bean	237	2.0	β
2SOD	Cu, Zn Superoxide Dismutase	Bovine erythrocytes	152	2.0	β
3CPV	Calcium Binding Parvalbumin	Carp muscle	109	1.85	α
3CYT	Cytochrome C	Tuna heart	104	1.8	α+β
3FXN	Flavidoxin	Clostridium MP	138	1.9	α/β
4PTI	Pancreatic Trypsin Inhibitor	Bovine pancreas	58	1.5	α+β
5CHA	α-Chymotrypsin	Bovine pancreas	236	1.67	β
5CPA	Carboxypeptidase A	Bovine pancreas	308	1.54	α/β
6LYZ	Lysozyme	Hen egg white	128	2.0	α+β
9PAP	Papain	Papaya	212	1.65	α+β

[1] Number of amino acid residues.
[2] Atomic resolution.
[3] As defined by Levitt and Chothia (1976).

Table 3-2. LD and BDA Values for Ideal Homopolymer Conformations

	Conformation	LD value[1]	BDA value[2]
1	α-Helix	20.71	51.40
2	2.2_7 Helix	29.08	178.24
3	3_{10} Helix	23.32	85.00
4	γ-Helix	20.20	6.75
5	Left Hand α-Helix	20.71	-51.40
6	ω-Helix	19.25	-34.72
7	π-Helix	19.30	34.26
8	Anti-Parallel β-Sheet	34.93	178.87
9	Parallel β-Sheet	33.25	177.80
10	Collagen	32.15	-98.14
11	Polyglycine	33.45	-110.18
12	Polyproline	32.28	-108.00

[1] In angstroms.
[2] In degrees.

Table 3-3. LD and BDA Values for Ideal Aperiodic Conformations[1]

Conformation	Linear Distance values				Backbone Dihedral Angle values				
Type I β-turn	33.18	26.86	20.86	27.32	177.42	-140.30	47.45	52.29	177.44
Type II β-turn	33.73	27.21	20.92	28.14	177.42	-107.86	0.06	85.12	177.43
Type III β-turn	33.18	27.47	21.66	25.94	177.42	-140.33	67.81	25.56	177.42
Type V β-turn	33.57	29.93	22.30	27.37	177.42	-128.52	-63.98	9.89	177.45
γ-turn	31.82	23.17	26.79	35.24	177.42	-13.37	23.39	177.43	177.42

[1]Eight residue fragments containing turns in residues 3-6 (3-5 for γ-turn), bounded on both sides with anti-parallel sheet.

Table 3-4. Differences in LDP and BDA Values for Ideal Conformations[1]

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.0	1.4	2.6	0.5	0.0	1.4	1.4	14.2	11.6	11.5	12.8	12.6
2	126.8	0.0	5.8	8.9	8.4	9.8	9.8	5.8	4.2	3.1	4.4	3.2
3	33.6	93.2	0.0	3.1	2.6	4.0	4.0	4.6	10.0	8.9	10.2	9.0
4	44.7	171.5	78.3	0.0	0.5	0.9	0.9	14.7	13.1	12.2	13.3	12.1
5	1.2.8	130.0	136.4	58.2	0.0	1.4	1.4	14.2	11.2	11.5	12.8	12.6
6	86.1	147.0	119.7	41.5	16.7	0.0	0.1	15.6	14.0	12.9	14.2	13.0
7	17.1	144.0	50.8	27.5	85.6	0.5	0.0	15.6	14.0	12.9	14.0	13.0
8	127.5	0.6	93.9	172.1	129.7	146.4	144.7	0.0	1.6	2.7	1.4	2.6
9	126.4	0.4	92.8	171.0	130.8	147.5	143.6	1.1	0.0	1.1	0.2	1.0
10	149.5	83.6	176.9	104.9	36.7	63.4	132.3	83.0	84.1	0.0	1.3	0.1
11	161.6	70.8	164.8	116.9	58.7	75.5	144.4	71.0	72.0	12.0	0.0	1.2
12	159.4	73.8	167.0	114.8	56.6	73.3	142.2	73.1	74.2	9.9	2.2	0.0

[1]Upper triangle: differences in LDP values for ideal conformations; lower triangle: differences in BDA values for ideal conformations

[2]Refer to Table 2 for conformations corresponding to these numbers.

Table 3-5. Root Mean Square Differences upon Superpositioning of Ideal Conformations[1]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	3.73	1.99	3.87	2.89	3.13	1.62	10.35	5.06	4.32	4.37	0.45	4.39	4.42	4.28	3.59	4.13
2		0	2.91	6.51	3.69	4.46	4.83	2.28	1.28	1.78	1.74	4.01	4.99	5.53	4.51	4.05	4.51
3			0	5.23	3.07	3.63	3.39	4.23	3.76	3.30	3.26	2.38	4.15	4.49	3.91	3.67	3.95
4				0	4.76	3.85	2.32	8.60	7.89	7.74	7.19	3.46	3.99	4.28	4.11	6.13	5.39
5					0	1.19	3.34	11.21	7.48	4.48	3.90	2.98	4.80	4.42	3.87	4.08	4.04
6						0	3.24	8.78	6.92	4.52	4.86	3.09	4.59	4.74	4.57	4.82	4.51
7							0	6.10	10.13	6.61	6.59	1.18	4.13	4.96	4.00	4.46	4.83
8								0	0.49	1.80	1.75	5.84	8.12	8.61	5.76	6.66	9.49
9									0	1.61	1.49	4.93	5.63	6.31	5.24	4.64	5.11
10										0	0.60	4.72	5.50	6.17	5.12	4.34	4.98
11											0	4.68	5.37	6.04	4.99	4.39	4.92
12												0	4.53	4.55	4.47	3.80	4.31
13													0	2.04	0.54	6.44	5.00
14														0	2.98	5.41	5.57
15															0	6.80	4.53
16																0	3.31
17																	0

[1]Numbers 1-12 correspond to conformations in table 3-2. Numbers 13-17 correspond to conformations 1-5 in table 3-3 respectively.

Table 3-6. Mean LD and BDA Values for the 30 Most Frequently Occurring Substructures Among 14 Non-homologous proteins.

#	Occurrences	Mean LDP				Mean BDA					Secondary Structure[1]
1	321	20.67	20.61	20.60	20.60	50.14	49.74	49.80	49.57	49.23	HHHHHHHH
2	17	20.30	20.57	20.92	22.57	49.46	48.33	48.28	52.76	82.73	HHHHHHHT
3	16	24.89	20.61	20.66	20.56	-107.11	53.18	47.96	50.32	49.17	UHHHHHHH
4	13	20.74	20.36	19.96	19.76	52.59	48.97	45.96	34.99	72.54	HHHHHHTT
5	12	27.18	21.14	20.48	20.66	-149.33	55.82	45.95	51.49	47.55	THHHHHHH
6	11	34.24	33.29	33.59	33.29	-152.21	-166.61	-161.36	-170.17	172.05	SSSSSSSS
7	10	25.63	20.92	20.50	20.50	106.98	48.86	50.20	47.18	47.90	THHHHHHH
8	9	22.93	20.85	20.56	20.41	72.62	53.97	46.74	52.94	44.79	THHHHHHH
9	9	34.52	34.35	33.28	33.29	-176.99	-150.41	-168.56	-150.95	171.93	SSSSSSSS
10	9	20.71	20.49	20.65	19.38	48.26	50.20	51.49	52.62	103.98	HHHHHHHT
11	8	35.29	35.27	35.51	35.49	-169.20	-174.11	-179.19	175.46	169.58	SSSSSSSS
12	7	20.50	20.75	19.45	25.63	50.90	51.66	55.05	-107.04	126.10	HHHHHHTT
13	6	33.07	34.08	32.96	33.36	177.67	-141.76	-148.84	-166.70	148.19	SSSSSSSS
14	6	20.71	20.24	19.76	23.79	52.54	45.12	39.64	76.76	124.01	HHHHHUUU
15	6	31.53	24.80	20.62	20.37	-99.96	-118.19	59.12	46.62	47.93	UUUHHHHH
16	6	20.60	20.12	22.50	26.36	49.78	48.65	43.23	94.55	161.34	HHHHHUUS
17	5	27.97	21.88	20.91	20.71	171.02	59.36	55.61	45.27	55.05	TTHHHHHH[2]
18	5	23.27	25.33	21.19	20.55	53.02	99.78	48.30	52.73	46.56	TTHHHHHH[3]
19	5	32.52	24.97	20.73	20.93	-127.50	-103.54	46.78	49.47	48.53	SUHHHHHH
20	5	32.58	33.91	33.17	32.75	-156.51	-173.55	-146.51	-151.82	171.66	SSSSSSSS
21	4	20.84	20.32	20.47	23.92	54.31	48.84	46.09	47.80	109.42	HHHHHHHT
22	3	19.99	19.93	19.61	24.26	45.10	49.01	30.08	20.07	89.90	HHHHHUUU
23	4	30.96	24.36	21.41	21.70	-138.59	-106.08	58.17	59.65	53.12	UUUHHHHH
24	4	33.32	33.33	33.04	32.78	-141.22	-153.06	-173.29	-114.60	140.66	SSSSSSSS
25	4	32.93	32.12	24.62	20.67	-158.24	-114.45	-112.01	52.87	49.60	STTHHHHH[4]
26	4	34.23	32.12	33.66	31.73	-151.65	-158.52	-163.69	-169.34	125.99	SSSSSSSS
27	4	22.73	22.40	23.44	26.90	65.71	72.05	54.31	100.96	110.39	UUUUUUUU
28	3	26.11	26.94	21.40	20.25	89.72	-145.14	59.10	46.06	51.87	TTHHHHHH
29	3	33.55	33.13	32.36	32.55	-168.40	-152.10	-165.59	-167.73	173.85	SSSSSSSS
30	3	20.24	21.80	22.36	21.28	45.59	57.04	59.13	79.14	15.53	HHHHHUUU

[1] H=α-helix, S=extended strand, T=turn, U=previously undefined.

[2] Type I or Type III turn. First residue of helix is last residue of turn.

[3] Type I turn. First residue of helix is last residue of turn.

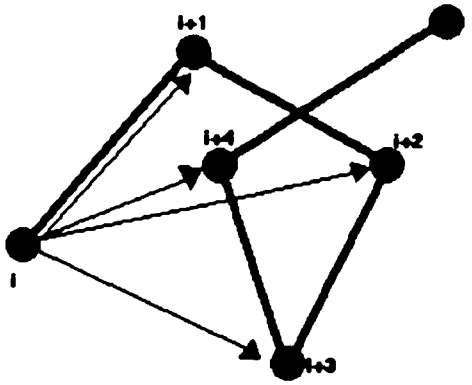
[4] Strand residue is first residue of turn, first helix residue is last residue of turn.

Table 3-7. Comparison of α-Helices as Determined by the Substructure Library and other Methods.

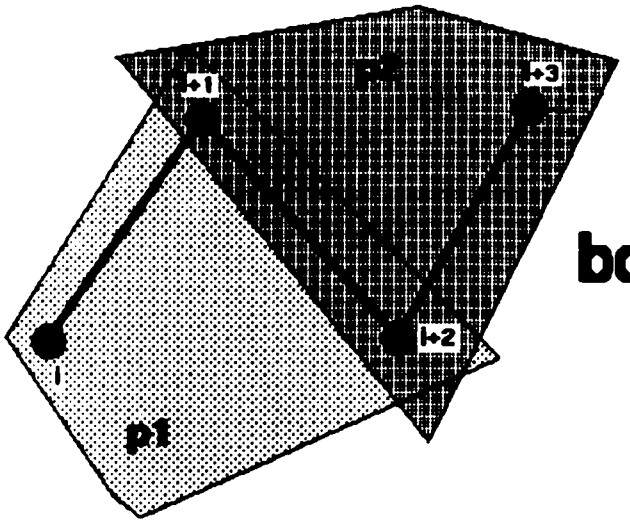
Carboxypeptidase A (5CHA)			Triose Phosphate Isomerase (1TIM)		
SSL	L + G[1]	K + S[2]	SSL	L + G	K + S
14-29	14-29	14-26	16-30	16-30	17-29
73-90	72-90	72-88	46-54	43-54	46-53
93-103	93-102	93-99	79-86	78-85	79-85
112-121	113-121	112-120	94-102	95-101	95-101
	143-147		104-119	104-118	105-118
173-187	173-186	172-185	129-136	130-135	130-135
215-232	215-233	215-228	137-153	137-153	138-152
253-262	253-261	253-260	176-195	176-195	177-195
285-307	285-305	285-304	196-203	196-203	197-202
			214-221	213-222	214-220
			237-244	232-244	238-244

[1] Levitt and Greer (1977).
[2] Kabsch and Sander (1983).

Figure 3-1. Schematic representation of the linear distance and backbone dihedral angle calculations form α-carbon coordinates..

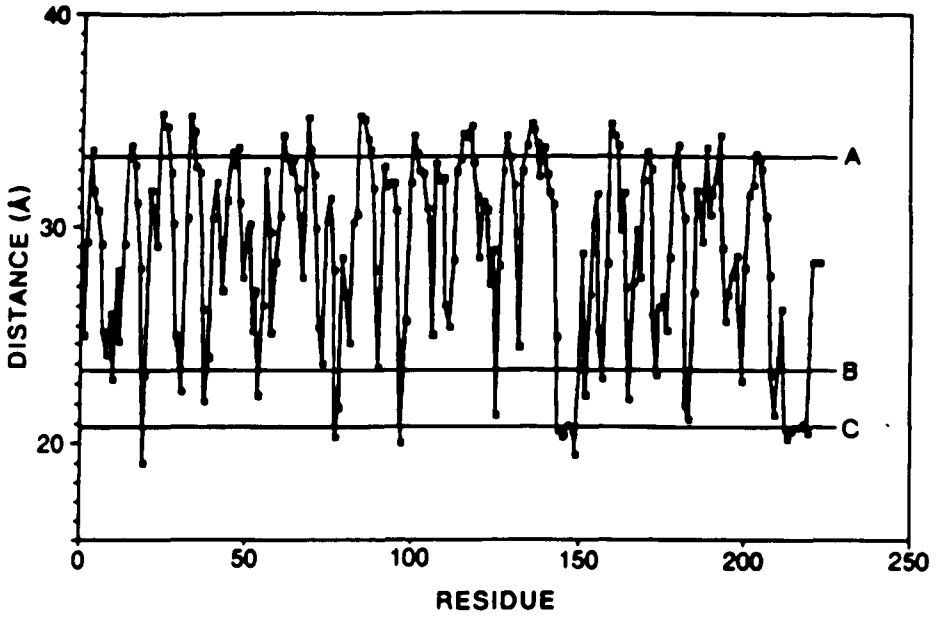


$$ldp_i = \sum_{j=1}^{4} d_{i,i+j}$$



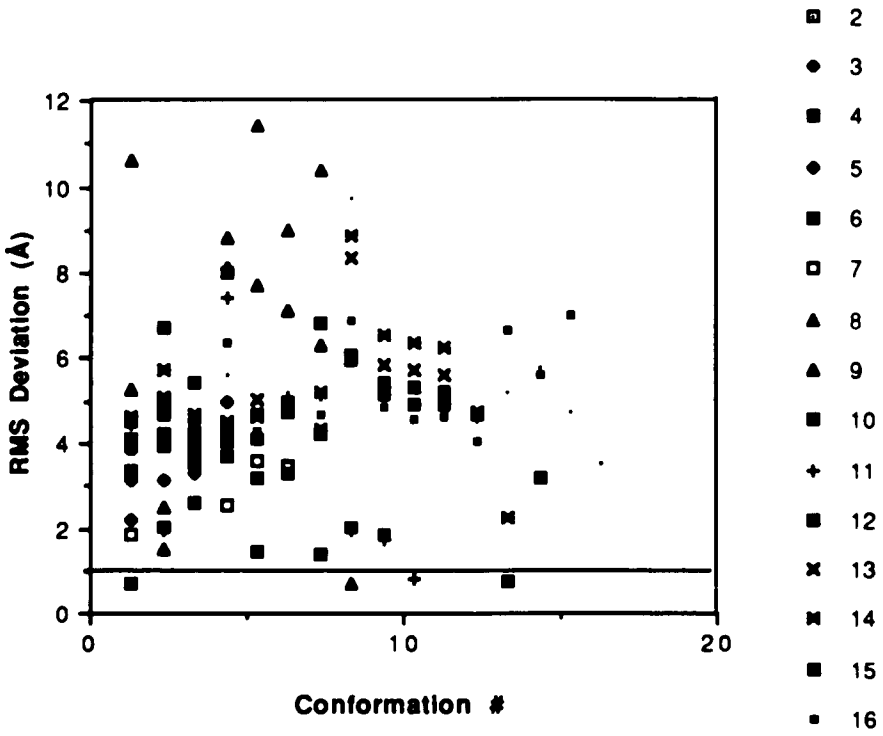
$$bda_i = \angle p1 \cdot p2$$

Figure 3-2. Linear distance plot of bovine β-trypsin at pH=5.0.[a]



[a]The linear distance value for each residue is calculated by summing the series of distances from its own α-carbon to the α-carbons of the next four residues (see Liebman, 1986). The horizontal lines in the plot correspond to the values calculated for ideal homopolymers in the A) β-sheet, B) 3_{10}-helix and the C) α-helix conformations.

Figure 3-3. Distribution of RMS deviation values upon superpositioning of ideal conformations.[1]



[1]The horizontal line through the graph corresponds to the value 1.0 Å.

CHAPTER 4

COMPARISON OF VARIOUS MOLECULAR FORMS OF BOVINE TRYPSIN: CORRELATION OF INFRARED SPECTRA WITH X-RAY CRYSTAL STRUCTURES

4.1. INTRODUCTION

The ability of biotechnology and protein engineering to produce new and novel proteins has underscored the need for a simple, reliable probe of protein conformation in a variety of conditions and environments. Currently, x-ray crystallography provides the most detailed information concerning protein structure. However, the length and complexity of the experiments and analysis as well as the difficulty in crystallizing many proteins severely limits use of this tool. Circular dichroism has been widely used as method for studying protein conformation in solution but its relative insensitivity to certain protein conformations such as β-sheets results in essentially qualitative results (Johnson, 1988).

Fourier-transform infrared spectroscopy has become recognized as a valuable tool for the examination of protein conformation in aqueous solution.

The reliability and wide applicability of FTIR has resulted in greatly expanded use in studies of protein secondary structure. It has been used to study the effects of ligand binding (Arrondo et al, 1988; Trewhella et al., 1989), pH (Casal et al., 1988; Byler and Purcell, 1989a), temperature (Casal et al., 1988; Byler and Purcell; 1989b), pressure (Wong et al., 1989) and solvent denaturation (Purcell and Susi, 1984; Wasacz et al, 1987) on conformation of globular, structural (Payne and Veis, 1988), and membrane-related proteins (Surewicz et al., 1987; Surewicz and Mantsch, 1988a; Haris et al., 1989).

Most infrared studies of protein conformation focus on absorptions in the Amide I (Amide I' in deuterated proteins) region (1700-1620 cm^{-1}) which arise primarily from stretching vibrations of the backbone C=O groups. The frequency of these vibrations has been shown to be sensitive to the molecular geometry and hydrogen bonding characteristics of the peptide backbone (Miyazawa and Blout, 1961) and specific conformational types (helix, strand, turns, etc.) give rise to discrete bands in the Amide I region (Byler and Susi,1986). Component bands in this region are typically broad and lie in close proximity to one another and the observed Amide I band often appears featureless. Thus, mathematical methods termed resolution-enhancement techniques, including derivative spectroscopy (Susi and Byler, 1983) and Fourier self-deconvolution (Byler and Susi, 1986), are necessary to resolve the individual Amide I components. The combination of these methods with the judicious application of curve-fitting techniques can reveal a wealth of information concerning the secondary structure of proteins (Byler and Susi,1986; Lee and Chapman,1986; Susi and Byler, 1986; Surewicz and Mantsch,1988).

Despite the extensive information which can be extracted using present methods, the application of FTIR to examining protein conformation is

somewhat limited by the lack of correlation between specific backbone folding types and individual component bands in the infrared spectrum. The theoretical studies of model peptide systems performed by Krimm and coworkers [for an excellent review see Krimm and Bandekar, 1986] provide guidelines for the interpretation of protein IR spectra. However, the inherent complexity, size and lack of symmetry in globular proteins often prevents direct application of these results to protein spectra. Aided by tools developed for macrostructural analysis, (e.g. Levitt and Greer, 1977; Kabsch and Sander, 1983; Liebman,1986) comparisons of IR spectra with high-resolution crystallographically determined protein structures can establish necessary spectra-structure correlations (Byler and Susi, 1986; Arrondo et al., 1988) as well as verify previous assignments which have often been based simply on empirical observations. This process can be further enhanced if additional biochemical information is available .

Trypsin is a well studied, autolytic serine protease. The process of autolysis in bovine trypsin is well defined and the sequence positions of several successive autolytic cleavages have been precisely determined (Schroeder and Shaw, 1968; Smith and Shaw, 1969). Three-dimensional coordinates of bovine trypsin in several closely related molecular states, including the active enzyme, the zymogen, in complex with various small molecule and macromolecular inhibitors, and at several pH values, have been determined by x-ray crystallography (Bernstein et al., 1977). The atomic resolution of a number of these structures is high (1.5 to 1.8 Å for proteins used in this study). Liebman (1986) has reported an extensive analysis of these structures using tools developed for analysis of secondary and tertiary structure in proteins as well as methods for discerning small conformational differences in closely related molecules. This study found that the conformational changes which

occur between different molecular states of bovine trypsin are localized within well defined regions of the molecule and, thus, the overall backbone conformation remains unchanged. This system is well suited for analysis of the spectral changes which accompany small conformational changes.

Reported here spectra-structure correlations derived from small but significant conformational changes which accompany 1) autolytic cleavage, 2) zymogen activation and 3) active site-directed inhibition in bovine trypsin and resulting changes in the Amide I' region of the infrared spectra. Additionally, our studies have enabled us for the first time, to our knowledge, to estimate the variability which results from the collection and analysis of protein FTIR spectra and comment on the effects that such variability may have on the interpretation of the spectra and subsequent correlation to protein spectra.

4.2. EXPERIMENTAL

4.2.1. MATERIALS

Bovine trypsin (3X crystallized, TRL3), and trypsinogen (1X crystallized, TG) were purchased from Worthington Biochemical Corp., Freehold, New Jersey. Diisopropylfluorophosphate (DIFP) was purchased from Sigma Chemical Co., St. Louis, MO. α-Trypsin and β-trypsin were separated from the commercial preparation and purified as described in Schroeder and Shaw (1968). Diisopropylphosphoryl-inhibited β-trypsin was prepared as described in Cunningham (1954) using purified β-trypsin. Trypsinogen was used without further purification. Purified components were stored below 0° C as lyophilized powders. All other compounds used were reagent grade.

Tryptic activity was measured using the substrate N-α-benzoyl-DL-arginine-p- nitroanilide (DL-BAPA) according to the method of Erlanger et al., (1961). Tryptic activity for purified α- and β-trypsin were in agreement with those observed by Schroeder and Shaw (1968). Trypsinogen displayed only negligible activity towards the DL-BAPA. DIP-β-trypsin showed no detectable activity towards the tryptic substrate.

Throughout this report we will refer to the amino acid sequence positions based on the ordered residue list in β-trypsin, i.e., 1 to 223 for the 223 residues in β-trypsin. Thus, the seventh residue in trypsinogen, isoleucine, assumes the number 1. The first six residues will be referred to as the N-terminal hexapeptide. The utility of this numbering has been discussed previously (Liebman, 1986).

4.2.2. SPECTROSCOPY

For IR spectroscopy proteins were prepared as 3.0% (w/v) solutions (1.3 mM) in 20 mM acetate (pD=5.0) or 20 mM imidazole (pD=6.9) buffers made up in D_2O. pD was determined by adding 0.4 to the pH reading (Covington et al., 1968) measured using a Horiba Cardy glass electrode pH meter. The spectrum of each protein was collected at the same pH (pD in this case) as the crystal structure determination. Sufficient $CaCl_2$ (1M in D_2O) was added to make the final solution 20 mM in Ca^{++}. Solutions were placed in IR cells with CaF_2 windows and teflon spacers with a pathlength of 75 μm.

IR spectra were collected at ambient temperature using a Nicolet 740 SX FTIR system equipped with a water-cooled Globar source, a Ge-coated KBr beam splitter and a broad range mercury/cadmium/telluride detector. All spectra were recorded at a resolution of 2 cm^{-1} by coadding 4096 double-sided

interferograms which were Fourier-transformed after application of a Happ-Genzel apodization function. The spectrometer and sample chamber were continuously purged with dry nitrogen. Spectral contributions from residual H_2O vapor in the light path and from buffers were subtracted using programs provided with the Nicolet FTIR software, Version 4.3. Factors for water vapor subtraction were determined by subtracting a second derivative spectrum of water vapor from the second derivative spectrum of the sample. The subtraction factor was varied until the region from 1700-1800 cm^{-1} was featureless. Subtraction using second derivative spectra is preferable to the original absorbance spectra because sharp vapor lines, which are often invisible in original spectra, are amplified by derivatization. This results in a more reliable correction for water vapor.

For each lyophilized protein sample, except when used in time course studies, three separate solutions were prepared and their spectra collected at different times, often several days apart. Further, each of these spectra was analyzed independently. All spectra were collected as soon as possible after mixing. The time required to load the cell, transfer it to the spectrometer and purge the sample chamber was approximately 30 min. Collection time for a single spectrum of 4096 coadded interferograms was also 30 min.

Second derivative spectra were calculated analytically as described by Susi and Byler (1983) with the Nicolet software except that the first derivative function was applied twice. Thus, each of the second derivative data points is calculated over five data points rather than three. Fourier self-deconvolutions were also calculated using the Nicolet software which is based on the method of Kaupinnen (1981). For these proteins deconvolution parameters of 18 cm^{-1} and 2.8 for the undeconvoluted band halfwidth and resolution enhancement factor,

k, respectively, were found to be optimal. A Lorentzian line shape function was assumed for the original spectra and a Bessel apodization function were used.

Curve fitting was performed using program ABACUS, an iterative Gauss-Newton non-linear regression program (see Byler and Susi, 1986). Deconvoluted spectra were fitted with Gaussian band profiles. Gaussian bands, as opposed to Lorentzian or a Gauss-Lorentz combination, were used to fit deconvoluted spectra because the deconvolution procedure removes the Lorentzian contribution to the band shape. The band shape in the deconvoluted spectra then takes the shape of the Bessel apodization function which is better approximated by a Gaussian profile (Griffiths et al., 1987). Initial band positions were taken directly from the second derivative spectra and no additional bands were added. Initial values for the peak heights and widths were estimated from the spectra. A non-sloping baseline was estimated from the spectrum and held constant. For the final fits, the heights, widths and positions of all bands were varied simultaneously. The relative intensity of each band (i.e., the band area as a fraction of the total Amide I area) was calculated from the final fitted band heights and widths.

4.3. RESULTS

4.3.1.β-TRYPSIN

Figure 4-1 shows the Amide I' region of the non-resolution enhanced spectra of β-trypsin at pD=5.0 as well as its second derivative and Fourier self-deconvoluted spectra. At this pD there autolytic activity is negligible within the time needed for sample preparation and data collection (R. Buono, personal communication). Table 4-1 lists the peak positions and areas relative to the

total area of the Amide I' region calculated by curve fitting the deconvoluted spectra (see Figure 4-2). The predominant feature in the deconvoluted Amide I' region is the intense peak at 1634 cm^{-1}. An additional strong peak appears at 1643 cm^{-1}. Weaker peaks are also present at 1625, 1654, 1663, 1673, 1684 and 1693 cm^{-1}. The second derivative spectrum of β-trypsin resolves two peaks at 1689 and 1695 cm^{-1} in place of the single band at 1693 cm^{-1} in the deconvoluted spectrum. Figure 4-2 also shows peaks, due to side chain vibrations (Chirgadze et al., 1975), fitted at 1590, 1601 and 1611 cm^{-1}. We routinely include these peaks in the curve fitting analysis to avoid the approximation otherwise incurred with addition of a sloping baseline parameter. These peaks are not part of the Amide I' band and, therefore, are not used in calculating the relative intensities of its components.

Examination of the ratio of the intensity of the Amide II' (i.e. deuterium exchanged) band to the intensity of the Amide II band showed that for all forms of trypsin studied, hydrogen-deuterium exchange was mostly complete. Additional spectra collected at later times showed no differences other than small downward shifts in the frequency (1-2 cm^{-1}) most likely due to exchange of a few more slowly exchanging hydrogens.

4.3.2. α-TRYPSIN

Figure 4-3 shows the curve fitted, deconvoluted Amide I' region of α-trypsin at pD=5.0. As with β-trypsin, at this pD autolytic activity is negligible. Table 4-1 lists the frequencies of the peaks and their relative areas. The overall shape of the Amide I' region as well as the positions and relative intensities of its components is similar to that of β-trypsin. This is in general agreement with biochemical studies which show α-trypsin to be an enzymatically active

species, although with reduced activity (Schroeder and Shaw, 1968). Closer inspection does show some clear distinctions between the spectra of α-trypsin and β-trypsin. The greatest difference occurs at the peak occurring about 1645 cm^{-1}. This peak becomes broader and increases in relative intensity upon conversion from β to α-trypsin. The relative intensity increases by 0.09. Concomitant with the increase in the intensity of the 1645 cm^{-1} peak are decreases in the intensities of several peaks. The band showing the greatest decrease is 1655 cm^{-1}. Its relative intensity decreases by 0.05. Additionally, the peaks at 1624, 1634, 1673 and 1683 cm^{-1} all decrease slightly, but significantly, in relative intensity.

Some of the second derivative spectra of α-trypsin appeared to have two separate peaks in the 1645 cm^{-1} region. However, the signal-to-noise ratio was insufficient to verify this. Therefore, the curve-fitting analysis was carried out considering both possible cases, using one or two peaks in this region. The inclusion of two peaks instead of one did not affect the final results to a significant degree. Thus, we chose the simpler data set for presentation and interpretation.

4.3.3. TRYPSINOGEN AND DIP-β-TRYPSIN

Table 4-1 lists the peak positions and areas of the Amide I' component bands, relative to the total Amide I' area, revealed from curve fitting the deconvoluted spectra of trypsinogen and DIP-β-trypsin (see Figures 4-4 and 4-5). The predominant feature in the deconvoluted Amide I' region of these spectra studied is the intense peak occurring near 1634 cm^{-1}. An additional strong peak appears around 1644 cm^{-1}. This band has the largest band area in trypsinogen (Figure 4-4). Weaker peaks are also present at about 1625, 1655,

1665, 1674, and 1683 cm^{-1} in the spectra of these proteins. Like β-trypsin, the second derivative spectrum of trypsinogen has two peaks at 1689 and 1695 cm^{-1} whereas the deconvoluted spectrum has only a peak at 1693 cm^{-1}. In contrast, the resolution enhanced spectra of DIP-β-trypsin display only a single peak at 1690 cm^{-1}.

4.3.4. AUTOLYSIS TIME COURSE OF β-TRYPSIN

Table 4-2 lists the peak positions and relative areas in the Amide I' region of β-trypsin, revealed through curve fitting its deconvoluted spectra, measured at various times after mixing with the pD=6.9 imidazole buffer. At pD=6.9, trypsin displays substantial autolytic activity (R. Buono, personal communication). The spectrum measured at 0.5 hr demonstrates that significant autolysis has taken place. The decreased intensity of the peak at 1655 cm^{-1} and the broadening and increased intensity of the 1645 cm^{-1} peak is indicative of substantial α-trypsin being present. Under these conditions two bands are observed in the second derivative and deconvoluted spectra at 1691 and 1696 cm^{-1} analogous to those seen only in the second derivative spectrum at pD=5.0. At 2 hr the spectrum reveals only small changes when compared to the 0.5 hr spectrum. At 5 hr the second derivative and the curve-fitted spectra (not shown) clearly demonstrate the almost complete loss of the peak at 1691 cm^{-1} while the remainder of the spectrum remains relatively unchanged. The spectrum at 20 hr reveals numerous small changes which may be indicative of further autolysis and possibly denaturation of the proteins. At this time, the 1691 peak has completely disappeared in the deconvoluted as well as the second derivative spectrum. In addition to changes in relative intensity, small shifts in frequency occur with time but these are invariably towards lower frequency and are most

likely isotopic shifts due to further deuteration of a few slowly exchanging peptide hydrogens.

4.3.5. AUTOLYSIS TIME COURSE OF α-TRYPSIN

Table 3 lists the peak positions and the relative areas of the component bands of α-trypsin at pD=6.9 measured at various times after mixing. Figure 4-6 shows the accompanying second derivative spectra. The spectrum measured at 45 min resembles closely the spectrum of α-trypsin at pD=5.0, showing an increase in relative intensity at 1645 cm^{-1} and a decrease in relative intensity at 1657 cm^{-1} compared to the same bands for β-trypsin. Deconvolution of this spectrum also resolves two peaks in the high frequency region at 1691 and 1694 cm^{-1}. At 2.5 hr the second derivative reveals a significant loss of intensity of the peak at 1691 cm^{-1} as is also demonstrated in the later stages of autolysis of β-trypsin. This is also evident in the deconvoluted spectrum which shows only a peak at 1694 cm^{-1} although the second derivative still shows a shoulder on the higher frequency side of the 1684 cm^{-1} peak. The peak at 1694 cm^{-1} increases in relative intensity somewhat but this is probably due to some residual intensity from the 1691 cm^{-1} peak which is no longer resolved. At 4 hours the second derivative shows the almost complete loss of the peak at 1691 cm^{-1}. As with β-trypsin, small shifts occur in the frequencies of some of the bands with time but these are toward lower frequency and most likely isotopic shifts due to further deuteration.

4.4. DISCUSSION

4.4.1. SPECTRAL VARIABILITY

In studying small differences among spectra of closely related molecules it is important to evaluate some measure of the reproducibility or variability within a given spectrum. For this reason, multiple data sets were collected for each protein sample under carefully controlled conditions as described in the Experimental section (Section 4-2). These data allow us to estimate the inherent variability which is present in the calculated individual band areas (i.e., relative intensities) of protein IR spectra. This evaluation is important in that it allows us to estimate a lower limit for interpretation of spectral differences measured using interferometric infrared techniques and analyzed using resolution-enhancement and curve-fitting methods. To our knowledge, no previous studies have addressed this in such detail for IR spectra of proteins.

Various sources of potential variability are present in FTIR measurements of proteins when analyzed with resolution-enhancement and curve-fitting techniques. Variability can be introduced in the sample preparation and data collection processes including small differences in the protein concentration and cell pathlength, various factors which may lead to instrument instability, the extent of hydrogen-deuterium exchange when using D_2O buffers and small differences in temperature and pH. Further, small differences in pH and temperature can result in spectral differences which lead to residual features after buffer subtraction. Analysis of measured spectra using mathematical techniques can also introduce variability. Correction of spectra for residual water vapor in the light path and for buffers is a subjective process. Additionally, noise and residual absorptions for uncorrected H_2O vapors are

amplified in the resolution enhancement techniques such as derivatization and self-deconvolution (Mantsch et al, 1989). The process of Fourier self-deconvolution can potentially introduce further variations including non-random noise and loss of detail due to smoothing induced by the apodization function. Curve-fitting can introduce additional variability because occasionally more than one set of parameters can provide a 'fit', particularly if the peaks are not well defined. The variability in the final calculated band areas, then, is the sum total of all of these contributing factors.

Table 4-1 lists the mean and standard deviation for each calculated relative intensity (band area) from three independent experiments. For the component peaks of β-trypsin, α-trypsin and DIP-β-trypsin the SD is less than 0.010 for all but the most intense peak in DIP-β-trypsin. Similar results are observed for trypsinogen except for the very broad, intense peak at 1645 cm-1 which has an SD of 0.016. This peak, being very broad is less well defined in the deconvoluted spectra (Figure 4-4). The peaks in the deconvoluted spectra of α-trypsin, β-trypsin and DIP-β-trypsin are quite well resolved. It has been our experience that when peaks are not well defined, this results in additional variability in the calculated relative intensities because more than one set of parameters may potentially give good quality fits. Thus, we estimate that if good definition of peaks is achieved in the deconvoluted IR spectra, the calculated relative intensities are sensitive to changes on the order of 0.01. Thus, these methods are capable of producing extremely reliable results. It has been estimated that circular dichroism measurements, a widely used method for conformational analysis in proteins, is sensitive to changes on the order of 0.05 (R. Buchet, unpublished results).

Stating that these methods are sensitive to changes of 0.01 in the relative intensities of Amide I' component bands does not imply changes of 1% in

percent composition of secondary structures, as the molar absorptivities of the individual conformations have not been established. Progress towards establishing these parameters has been slow due to lack of a suitable experimental system. However, recent studies (Mantsch et al., 1989; Jackson et al., 1989) and the earlier studies of Byler and Susi (1986) show that the ratios of molar absorptivities among the various conformations is probably close to 1. Thus, these methods are suitable for observing conformational changes in relative terms, particularly for one protein as it is perturbed by various factors.

The results presented here emphasize the importance of careful data collection and analysis. For studies which examine small spectral differences, it is necessary to collect spectra of sufficiently high signal-to-noise so that smoothing is not necessary. Our experience suggests that for optimal results, the number of scans which are signal averaged be high enough such that the non-absorbing region above 1700 cm-1 be essentially flat in the second derivative spectrum (after vapor correction) without additional smoothing. High signal-to-noise spectra are also important in Fourier self-deconvolution as the upper limit of the resolution enhancement factor is determined by the noise level. Thus, in order to achieve good definition of component peaks, high signal-to-noise is necessary. Further, careful subtraction of water vapor and buffer are necessary before proceeding to deconvolution because lack of or improper subtraction may lead to spurious bands in the resolution-enhanced spectra (Mantsch et al., 1989). We would also like to point out that when examining differences among closely related spectra that a series of spectra be collected in order to estimate the variability among the spectra and therefore provide confidence that small measured spectral differences arise from conformational differences and not simply variability.

4.4.2. β-TRYPSIN

Figure 4-7 displays the linear distance plot of bovine β-trypsin. Analysis of the crystal structure of β-trypsin reveals that the molecule consists primarily of the extended strand conformation (Levitt and Greer, 1977; Marquart et al., 1983; Liebman, 1986). This is also illustrated in the linear distance plot in Figure 4-7, as stretches where the value is greater than 33 Å. Twelve of the extended strands interact in antiparallel β-sheets (Marquart et al., 1983; Liebman, 1986). This agrees well with the intense band in the IR spectrum at 1634 cm^{-1} which has been previously assigned to extended strand structures (Byler and Susi, 1986; Surewicz and Mantsch, 1988a). The relative intensity of this band along with the bands occurring at 1625 and 1674 cm^{-1}, which have also been assigned to the extended strand conformation (Byler and Susi, 1986), totals to 0.41. This is in good agreement with estimates of the percentage strand structure in β-trypsin which range from 30% (Marquart et al., 1983) to 56% (Levitt and Greer, 1977). The disagreement between these two methods is due to different criteria for determining a particular conformation.

Crystal structure analysis shows three α-helices in bovine β-trypsin (Marquart et al., 1983; Liebman, 1986). A single turn of α-helix occurs at residues 144-150 and two distinct but contiguous helices occur at the carboxy terminus. Overall, Liebman (1986) estimates 9% of the residues in β-trypsin adopt the α-helical conformation. These appear in the linear distance plot as stretches of values near 20.5 Å. This agrees well with the results of Levitt and Greer (1977) and (Marquart et al., 1983), both of which estimate 9% α-helical content in β-trypsin. Bands occurring about 1655 cm^{-1} in the IR spectrum have been assigned to the α-helix (Byler and Susi,1986; Surewicz and Mantsch; 1988). β-trypsin has a peak at 1654 cm^{-1} which accounts for 0.15 of the total

Amide I' intensity. This value is in general agreement with the crystal structure value but appears somewhat high when compared with the close agreement demonstrated between the fraction of α-helix determined by an automatic secondary structure analysis algorithm (Levitt and Greer, 1977) and the relative intensity of the band occurring near 1655 cm^{-1} for 20 proteins (Byler and Susi, 1986).

A strong band occurs at 1645 cm^{-1} in the β-trypsin spectra accounting for 0.24 of the Amide 1' intensity (Table 4-1). Such a band probably arises from carbonyls which hydrogen bond to solvent molecules and has previously been assigned to 'unordered' conformations, or conformations which do not take part in intrachain hydrogen bonds (Byler and Susi, 1986, Surewicz and Mantsch, 1988a). The term 'unordered', as used in this context, is somewhat of a misnomer as conformations which hydrogen bond to solvent molecules do not necessarily lack order. Correlations between specific conformations and this band have not been discerned partly due to this ambiguity in definition. In this paper we will use the term 'irregular' to refer to conformations which are not classified as helix, extended strand, reverse turns or loops. The remainder of bands in the β-trypsin spectrum, as in most globular proteins, are tentatively attributed to reverse turns and loops.

4.4.3. AUTOLYSIS OF β-TRYPSIN

Figure 4-8 summarizes the autolysis process in bovine trypsin. The first autolytic cleavage, after that which results in activation of the zymogen, in bovine trypsin occurs between Ser-125 and Asp-126, when allowing β-trypsin to autodigest, resulting in the formation of α-trypsin. The backbone conformation in this region, in the β form of trypsin, consists of a loop bounded

by stretches of extended strand (Marquart et al., 1983; Liebman, 1986). As outlined in the Results section, the greatest difference between β and α-trypsin is the loss of intensity in the 1656 cm^{-1} band with an increase in the 1645 cm^{-1} band. Amide I' bands around 1655 cm^{-1} in proteins have been assigned to α-helices (Byler and Susi, 1986; Surewicz and Mantsch, 1988a). Examination of the crystallographically determined structure shows the α–helices in trypsin are spatially and sequentially distant from this loop (see Figure 4-7) and are presumably unaffected by this autolytic cleavage. Thus it appears from these results that certain types of loops in proteins may also absorb around 1655 cm^{-1}. Further evidence for this interpretation lies in the relative intensity of this band in the two proteins. In β-trypsin the relative intensity of the 1656 cm^{-1} band is 0.15 and 0.10 in α-trypsin. As stated above x-ray data indicate that 9% of trypsin is in the α-helical conformation. Additionally, Surewicz and Mantsch (1988 b,c), and Surewicz et al., (1988) , based on an unexpected band around 1655 cm^{-1} in several small peptides and proteins determined by circular dichroism to contain no helices, have suggested that this peak arises from an unexchanged irregular conformation or an 'atypical nonperiodic' conformation. Our results would suggest the latter because the autolytically cleaved loop appears on the surface where it is susceptible to autolytic cleavage and should therefore be available for hydrogen exchange. Thus, these results suggest caution in interpreting bands near 1655 cm^{-1} as resulting entirely from α-helices.

In previous work bands near 1644 cm^{-1} in proteins have been assigned to unordered structures (Byler and Susi ,1986; Surewicz and Mantsch, 1988a). The increased relative intensity and broadening of this band upon conversion from β to α-trypsin is consistent with this assignment as one would expect local unfolding at this part of the protein molecule upon cleavage of a peptide bond.

This is further corroborated by crystallographic experiments with trypsin which has been inhibited with diisopropylfluorophosphate (Chambers and Stroud, 1979). In this determination, crystals of this compound were determined to contain an approximately 50-50 distribution of β and α-trypsin. The electron density in the region of the autolytic cleavage is weaker indicating disordering in the cleaved, two-chain component (i.e. α-trypsin). This interpretation is consistent with previous studies which indicate broadening of bands near 1645 cm^{-1} with an increase in irregular or disordered structures (Purcell and Susi, 1984; Byler and Susi, 1986; Surewicz et al., 1987). The broadening is most likely due to a greater population of slightly different hydrogen bonding patterns to D_2O molecules by polypeptide fragments lacking well-defined repetitive structures (Surewicz and Mantsch, 1988a).

Other spectroscopic changes which occur upon conversion from β to α-trypsin are small decreases in intensity in the 1625, 1634, 1673 and 1684 cm^{-1} bands. The 1625 and 1634 cm^{-1} bands are assigned to extended strands. Byler and Susi (1986) assigned bands at 1675 +/- 4 cm^{-1} as the high frequency component of β-strands but the precise position of this component still remains unclear although the 1673 cm^{-1} band is a likely choice. The loss in intensity of the bands assigned to β-strands is probably due to disordering or disruption in hydrogen bonding of the β-strands which bound this loop. We can not assign the 1684 cm^{-1} band to a specific conformation with presently available data.

4.4.4. AUTOLYSIS OF α-TRYPSIN

A second autolytic cleavage in bovine trypsin occurs between residues Lys-170 and Asn-171 (see Figure 4-8) when α-trypsin is dissolved under conditions which allow autodigestion (Smith and Shaw, 1969). The resulting

twice-cleaved product has been designated ψ-trypsin. Unlike the conformation in the region of the first autolytic cleavage, the conformation in this region is complex and contains several types of regular secondary structure. It consists of a short stretch of extended strand bounded by a 3_{10}-loop or Type III β-turn on the N-terminal side and two consecutive standard Type II β-turns on the C-terminal side (Chambers and Stroud, 1979). Spectroscopically, the most significant changes upon conversion of α-trypsin to ψ-trypsin are loss of the peak at 1691 cm^{-1} and loss of intensity at 1633 cm^{-1} and an increase in relative intensity at 1644 cm^{-1} (see Figure 4-5, Table 4-3). After 4 hr of incubation, ψ-trypsin becomes the predominant species (Smith and Shaw, 1969). Loss of intensity in the 1634 cm^{-1} band and gain in intensity in the 1644 cm^{-1} band can be interpreted as a loss of extended strand structure either from the strand in the region of the cleavage or further disordering of extended strands resulting from the first autolytic cleavage or some combination of the two. Because of the complexity of the folding in this region, the origin of the 1691 cm^{-1} band remains unclear. In theoretical and experimental studies of model peptide systems, Krimm and Bandekar (1986) have suggested that certain types of β-turns can absorb in this high frequency region. While this is consistent with the presence of turn structures in this portion of the molecule, the lack of more definite information concerning structural changes in this complexly folded region prohibits unequivocal assignment of the 1691 cm^{-1} band to a specific conformation.

Another result which merits discussion is the observation that the spectrum of β-trypsin after 20 hr of incubation (see Table 4-2) and the spectrum of α-trypsin after 4 hr of incubation (see Table 4-3) are essentially identical, except for small differences in the position of some of the bands. This is a desired result as it is expected that after conversion to α-trypsin, the solution initially containing

β-trypsin should then convert to ψ-trypsin (Schroeder and Shaw, 1968; Smith and Shaw, 1969). These two spectra are presented simultaneously in Figure 4-9. This result illustrates the utility of FTIR spectroscopy in monitoring biochemical processes which involve relatively small changes in protein conformation.

4.4.5. TRYPSINOGEN TO TRYPSIN TRANSITION

Conversion of trypsinogen to β-trypsin occurs as a result of an autolytic cleavage of the protein chain which releases the N-terminal hexapeptide (see Figure 4-8). This cleavage results in a series of conformational changes which has been described as a disorder-to-order transition (Huber and Bode, 1978). Three regions of the trypsinogen molecule, collectively termed the activation domain, are known from crystallographic experiments to be disordered (Felhammer et al., 1977, Walter et al., 1982). These regions are outlined in Figure 4-7. These residues become ordered, although not necessarily adopting regular secondary structure, to form an active enzyme, β-trypsin. The three regions which make up the activation domain, residues 122-132, 164-175 and 193-200, account for 31 residues. In addition, first six residues in trypsinogen, the N-terminal hexapeptide, are observed to be disordered in the trypsinogen molecule (Felhammer et al., 1977). Thus, the total number of residues disordered in trypsinogen is 37. The term 'disordered' as used here refers to crystallographic disorder which means that in the crystallographic experiments no significant electron density is found for these regions. This is due to dynamic conformational flexibility in these regions or the presence of multiple stable conformations (Walter et al., 1982). The term 'disordered', thus, contrasts with the term 'irregular'. We shall use the term 'disordered' only to refer to regions

which are crystallographically disordered. The remainder of the trypsinogen molecule, which accounts for 83 % of the residues, has almost identical conformation when compared to β-trypsin using a difference linear distance technique (Liebman, 1986) and the RMS deviation is only 0.2 Å upon structural superposition of these regions (Huber and Bode, 1978).

The greatest difference between the Amide I region of the IR spectra of trypsinogen and β-trypsin occurs in the band near 1645 cm^{-1}. This band is much broader and greater in relative intensity in the trypsinogen spectra (Table 4-1). Its relative intensity is 0.41 versus 0.24 for β-trypsin, a difference of 0.17 upon conversion to the active enzyme. Amide I' infrared bands at about 1645 cm^{-1} have been previously assigned to unordered conformations in proteins (Byler and Susi, 1986; Surewicz and Mantsch, 1988a). Such an assignment is consistent with our results. Ordering of the 32 residues of the activation domain and the loss of 6 disordered residues of the N-terminal hexapeptide, a total of 17% of the residues, correlates surprisingly well with a 0.17 decrease in relative intensity of this band as well as the decreased bandwidth. Thus, our results imply that assignment of bands around 1645 cm^{-1} to unordered or disordered conformations is consistent with respect to disorder in a crystallographic sense. The presence of this peak in β-trypsin, where no regions are disordered in the crystal, suggests that other 'irregular' conformations absorb in this region as previously assumed (Surewicz and Mantsch, 1988a). It appears that is only possible to distinguish the two with other information such as that from crystallographic experiments.

The large decrease in relative intensity in the 1645 cm^{-1} band occurs concomitant with increases in the relative intensity of several bands upon conversion from trypsinogen to β-trypsin (see Table 4-1). This is the only Amide I' component band which shows a decrease in relative intensity upon this

conversion. Bands near 1625, 1634, 1654, 1674 and 1684 cm^{-1} show increases in relative intensity. These increases in intensity are presumably due to formation of the ordered conformations observed crystallographically in the activation domain of β-trypsin.

One of the newly ordered regions is also the location of the first autolytic cleavage in β-trypsin. The autolysis experiments reported here imply that this region, a loop in the β-trypsin form, absorbs around 1655 cm^{-1}. The observed increase in the relative intensity of the band around 1655 cm^{-1} and decrease of the band at 1645 cm^{-1} with the trypsinogen-trypsin transition are consistent with this assignment. In trypsinogen, where this region is disordered, the relative intensity of the band near 1655 cm^{-1} is significantly lower. While bands in the IR spectrum around 1655 cm^{-1} have been previously assigned solely to α-helices (Byler and Susi, 1986; Surewicz and Mantsch; 1988), the crystal structures of trypsinogen and trypsin show no conformational differences in the α-helical regions (Huber and Bode, 1978; Liebman, 1986).

The two other flexible loops in the activation domain of trypsinogen also become ordered in trypsin, one forming an extended strand ending in a type II β-turn and the other forming a loop which also contains a type II β-turn. Amide I' bands near 1625 and 1634 cm^{-1} have been assigned to low-frequency vibrations of extended strands (Byler and Susi, 1986; Surewicz and Mantsch, 1988a). Our results are consistent with this assignment in that increases in relative intensity are observed for bands near these frequencies on conversion of trypsinogen to β-trypsin. A high frequency vibration near 1675 cm^{-1} is also expected for extended strand structures (Byler and Susi. 1986) but because of overlap with bands thought to arise from reverse turns, unequivocal assignment of a band has yet to be determined for this mode (Surewicz and Mantsch, 1988a). Increases in relative intensity are observed for bands near 1674 and

1683 cm^{-1}. One of these presumably arises from the high frequency extended strand mode and the other from the β-turns formed. At present, specific classes of reverse turns have not been assigned to individual bands in the Amide I region. However, these data imply that type II β-turns absorb either near 1675 cm^{-1} or near 1683 cm^{-1}. This suggestion is consistent with theoretical studies of model peptide reverse turn structures reported by Krimm and Bandekar (1986) which predict that type II β-turns may have absorptions in the 1680 cm^{-1} region.

4.4.6. INHIBITION WITH DIFP

Liebman (1986) has reported that the two regions of greatest local conformational perturbation in β-trypsin upon inhibition with the active site directed protease inhibitor DIFP involve residues 122-130 and residues 190-197. Significant spectral changes also accompany this inhibition (Figures 4-1 and 4-3, Table 4-1). The spectrum of the inhibited form displays increases in the relative intensity of bands at 1644 cm^{-1} and at 1690 cm^{-1} concomitant with decreases in intensity of bands at 1655 cm^{-1} and 1674 cm^{-1}. Additionally, a band near 1695 cm^{-1} is present in the spectrum of β-trypsin but is not present in DIP-β-trypsin.

One of the structurally perturbed regions in DIP-β-trypsin, residues 122-130, is part of the previously described activation domain and is also the location of the first autolytic cleavage (Schroeder and Shaw, 1968). The loop at this region has been found to contribute to the band at 1655 cm^{-1} in the β-trypsin spectrum. As with trypsinogen, the results observed here corroborate this assignment as the relative intensity of the band near 1655 cm^{-1} is significantly lower compared with β-trypsin. Again, no perturbation is observed crystallographically in the helical regions (Huber and Bode, 1978; Liebman, 1986). This corroborates the

assignment of this loop to the 1655 cm-1 region proposed in the autolysis and zymogen activation studies discussed above.

We can assign the loop at the 122-130 region in β-trypsin to absorptions in the 1655 cm^{-1} region by analogy to two other systems. However, in the absence of specific assignments to the other perturbed regions, both in β-trypsin and DIP-β-trypsin, the complexity of the spectral changes which are observed upon inhibition with diisopropylfluorophosphate do not allow unambiguous assignment of conformations to spectral bands. Thus, these results are indeterminate. Further, some recently published results (Casal et al., 1988) have suggested that bands near 1695 cm^{-1} in protein infrared spectra are not Amide I vibrations but represent unionized carboxyl groups. Such a possibility further complicates the assignment process. The absence of this band in DIP-β-trypsin may be due to the different pD but a spectrum of β-trypsin at the higher pD (not shown) indicates that the higher pH alone does not result in loss of this band.

Table 4-4 summarizes the spectra-structure assignments for the molecules examined in this study. These assignments may prove useful for future infrared studies of protein conformation. However, future studies are necessary to determine the general applicability of these assignments to protein molecules.

4.5. CONCLUSION

The results presented here have demonstrated that FTIR spectroscopy, combined with resolution-enhancement and curve-fitting techniques, is a very sensitive probe of secondary structure in proteins. One can examine the often minute effects of ligands and environmental factors on conformation. To further increase the utility of FTIR for examining protein conformation it is necessary to

discern the relationships between the presence of specific conformational types and the pattern and intensity of bands in the conformation sensitive Amide regions. In the absence of rigorous theoretical methods such as normal mode analysis, which are presently limited by lack of computational power to the study of relatively short peptides, other methods of discerning these relationships are needed.

It has been demonstrated that comparison of spectral changes with crystallographically determined structures along with biochemical information can provide a means of ascertaining necessary spectra-structure correlations. In particular, we have shown that certain loop structures may absorb around 1655 cm^{-1}, a region previously assigned solely to α-helices. This appears to be the first assignment of an actual region of conformation to a particular amide I component band. It has also been demonstrated that in D_2O 'disordered' regions absorb near 1645 cm^{-1}. Additionally, we have provided further evidence for previous assignments of bands to extended strand and 'irregular' structures. Future progress towards spectra-structure assignments may come from application of methods exemplified here to other similar systems.

In addition to new spectra-structure assignments, the results of multiple, independent experiments and analysis of each protein sample have provided an estimate of the variability present in FTIR spectra when analyzed using resolution enhancement and curve-fitting techniques. These results estimate that if good resolution of peaks is achieved in the deconvoluted IR spectra, the calculated relative intensities are sensitive to changes on the order of 0.01. Thus, these techniques are capable of producing extremely reliable results. It has been estimated that circular dichroism measurements, a widely used method for conformational analysis in proteins, is sensitive to changes on the order of 0.05.

The results of studies which monitored the amide I region during the process of autolysis have demonstrated that resolution-enhanced FTIR spectroscopy is a sensitive probe of changes in molecular conformation which accompany biochemical processes. The combination of iterative curve-fitting and resolution-enhancement methods provide an accurate and reproducible conformational 'fingerprint'. Thus, FTIR spectroscopy has potential value as a tool for monitoring biochemical processes if such processes involve conformational changes.

Table 4-1. Peak Positions and Relative Intensities of various molecular states of Trypsin

β-Trypsin				α-Trypsin				Trypsinogen				DIP-β-Trypsin			
ν[a]	HW[b]	A[c]	σ[d]	ν	HW	A	σ	n	HW	A	s	n	HW	A	s
1625	2.5	0.08	0.002	1625	2.6	0.08	0.003	1625	2.4	0.06	0.001	1625	2.9	0.08	0.001
1634	3.1	0.23	0.001	1634	3.5	0.22	0.007	1633	3.3	0.20	0.007	1634	3.4	0.25	0.005
1643	3.9	0.24	0.003	1645	6.2	0.33	0.004	1645	8.0	0.41	0.016	1645	5.5	0.28	0.011
1654	3.4	0.15	0.004	1656	3.2	0.10	0.001	1656	2.6	0.07	0.004	1656	3.0	0.10	0.007
1664	2.9	0.12	<0.001	1664	3.4	0.12	0.001	1665	3.4	0.12	0.003	1665	3.4	0.12	0.001
1674	3.3	0.10	0.001	1674	2.9	0.08	0.003	1674	2.7	0.07	<0.001	1674	2.7	0.07	<0.001
1684	2.9	0.06	<0.001	1684	2.5	0.05	0.002	1683	2.5	0.05	<0.001	1682	2.9	0.06	<0.001
												1690	2.2	0.04	<0.001
1693	2.7	0.02	0.001	1692	3.2	0.02	0.001	1692	3.4	0.02	0.001				

[a] Frequency positions are rounded off to the nearest integer.
[b] Half-width at half height.
[c] Mean relative intensity calculated for three independent experiments.
[d] Standard deviation fo relative intensities calculated for three independent experiments.

Table 4-2. Peak Positions and Relative Intensities of β-trypsin, pD=6.9, at various times after mixing.

0.5 hours		2 hours		5 hours		20 hours	
ν[1]	A	ν	A	ν	A	ν	A
1625	.07	1625	.07	1625	.07	1625	.09
1635	.27	1634	.25	1634	.24	1633	.15
1647	.30	1646	.33	1646	.33	1643	.38
1657	.09	1657	.09	1656	.08	1655	.10
1665	.12	1665	.11	1664	.13	1664	.13
1674	.07	1674	.07	1674	.07	1674	.09
1684	.05	1684	.05	1684	.06	1684	.05
1691	.02	1691	.02				
1696	.01	1696	.01	1694	.01	1694	.01

[1]Wavenumbers in cm^{-1}

Table 4-3 Peak Positions and Relative Intensities of α-Trypsin, pD=6.9, at various times after mixing.

0.75 hours		2.5 hours		4 hours	
ν^1	A	ν	A	ν	A
1625	.07	1625	.08	1625	.08
1635	.22	1634	.21	1634	.14
1647	.33	1646	.35	1643	.40
1657	.10	1656	.09	1656	.10
1665	.12	1665	.12	1664	.13
1674	.07	1674	.08	1674	.08
1684	.05	1684	.06	1684	.06
1691	.02				
1696	.01	1696	.02	1694	.01

[1]Wavenumbers in cm^{-1}

Table 4-4. Summary of Amide I Spectra-Structure Assignments for Bovine Trypsin

Band Frequency	Conformation
1625	extended strand[a,b]
1635	extended strand[a,b]
1645	irregular[a,b], disordered[a]
1655	α-helix[b], loops[a]
1664	turns[b]
1674	extended strand[b], possibly Type II β-turn[a]
1683	turns[b], possibly Type II β-turn[a]
1689	turns[b]
1695	turns,[b] possibly carboxyl C=O[c]

[a] Proposed or corroborated from this study.
[b] Assigned in previous investigations, see Surewicz & Mantsch (1988) and references therein.
[c] See Casal et al., 1989.

Figure 4-1. Amide I region of the infrared spectrum of bovine β-trypsin at pD=5.0, 20 mM Ca^{2+} (b) second derivative spectrum (a) non-resolution enhanced IR absorption spectrum and (c) spectrum after Fourier self-deconvolution.

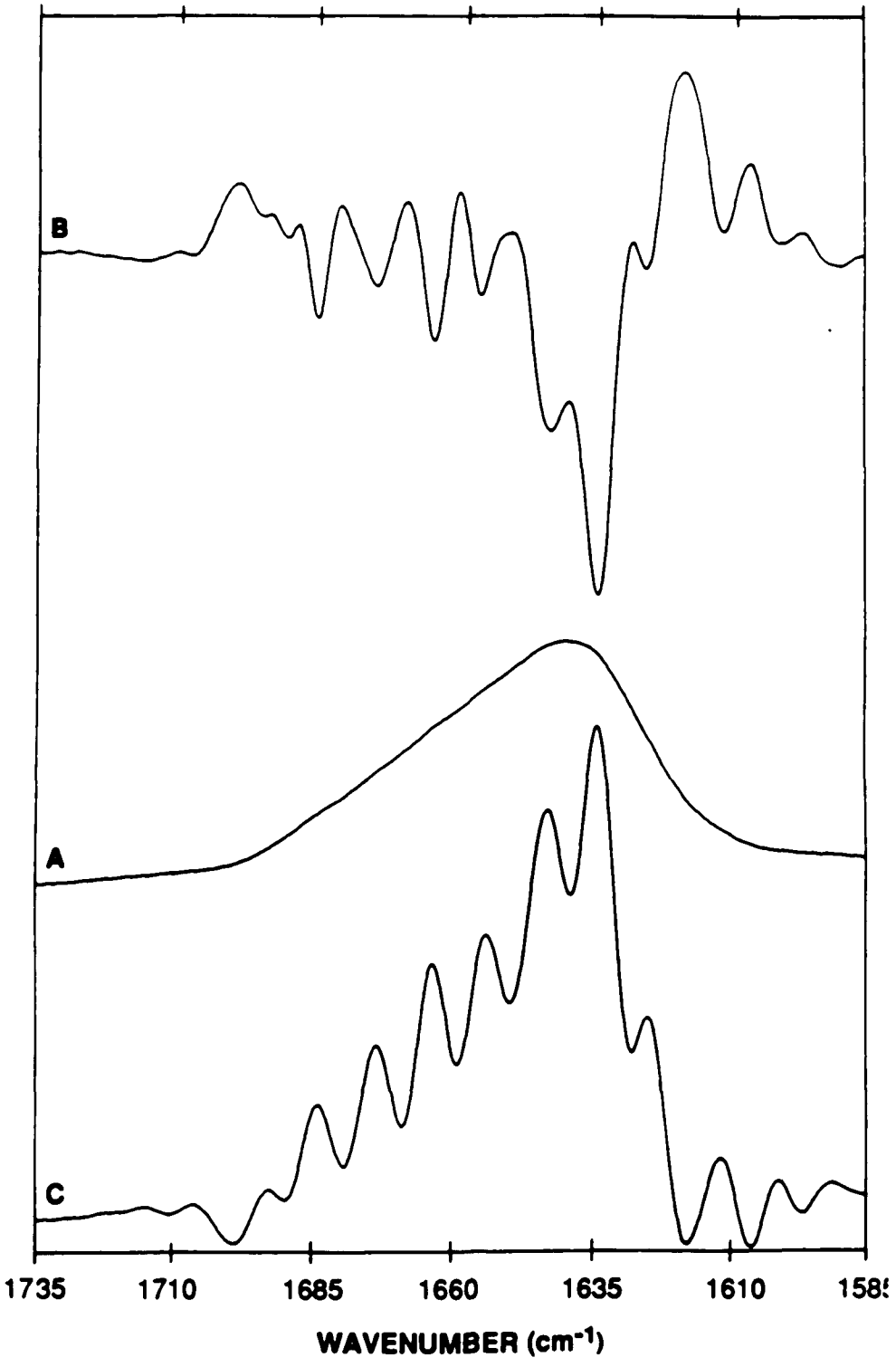


Figure 4-2. Deconvoluted infrared Amide I' band of bovine β-trypsin at pD=5.0, 20 mM Ca^{2+} (++++). Individual Gaussian components and their sum (———).

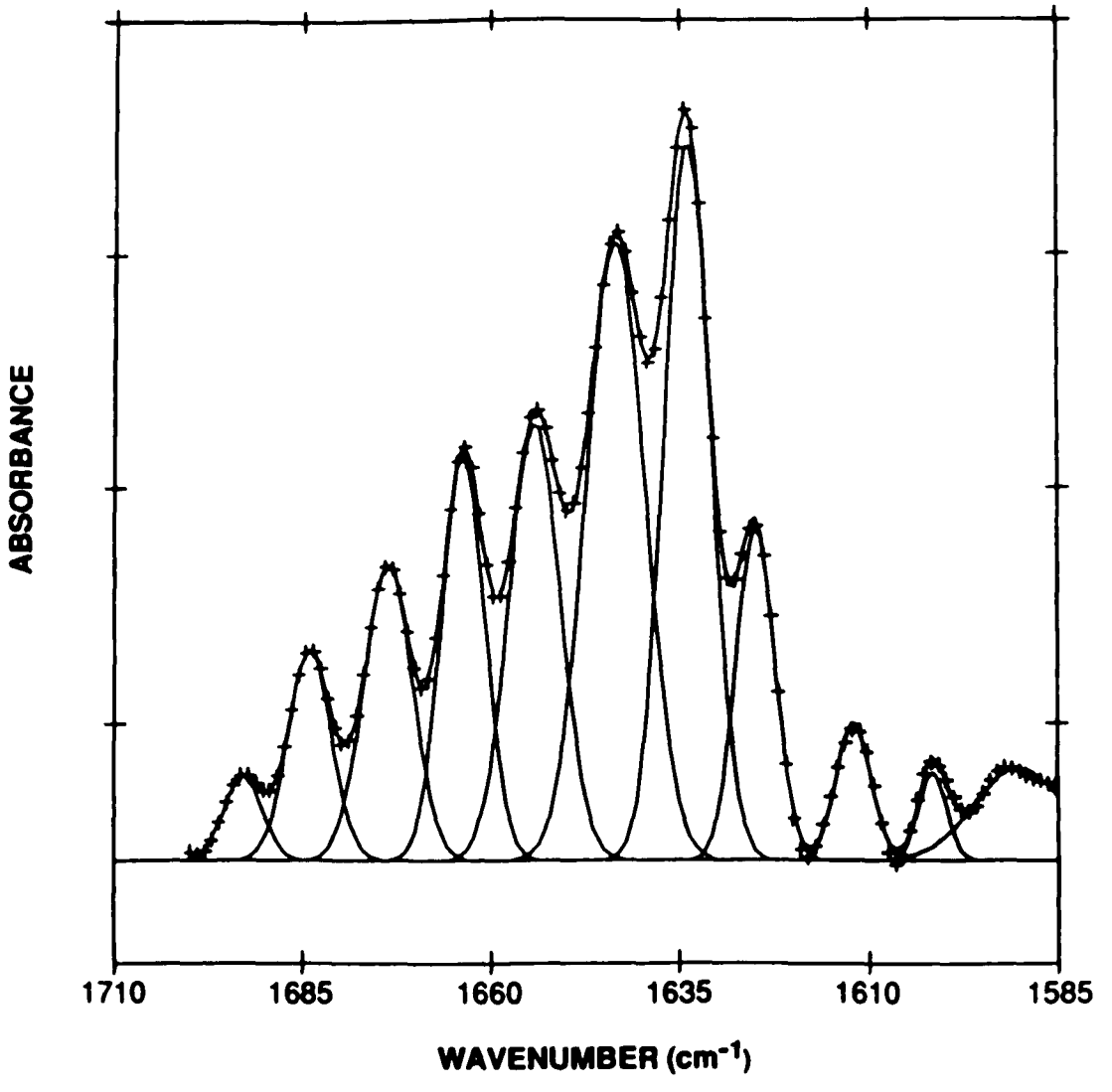


Figure 4-3. Deconvoluted infrared Amide I' band of bovine α-trypsin at pD=5.0, 20 mM Ca^{2+} (++++). Individual Gaussian components and their sum (———).

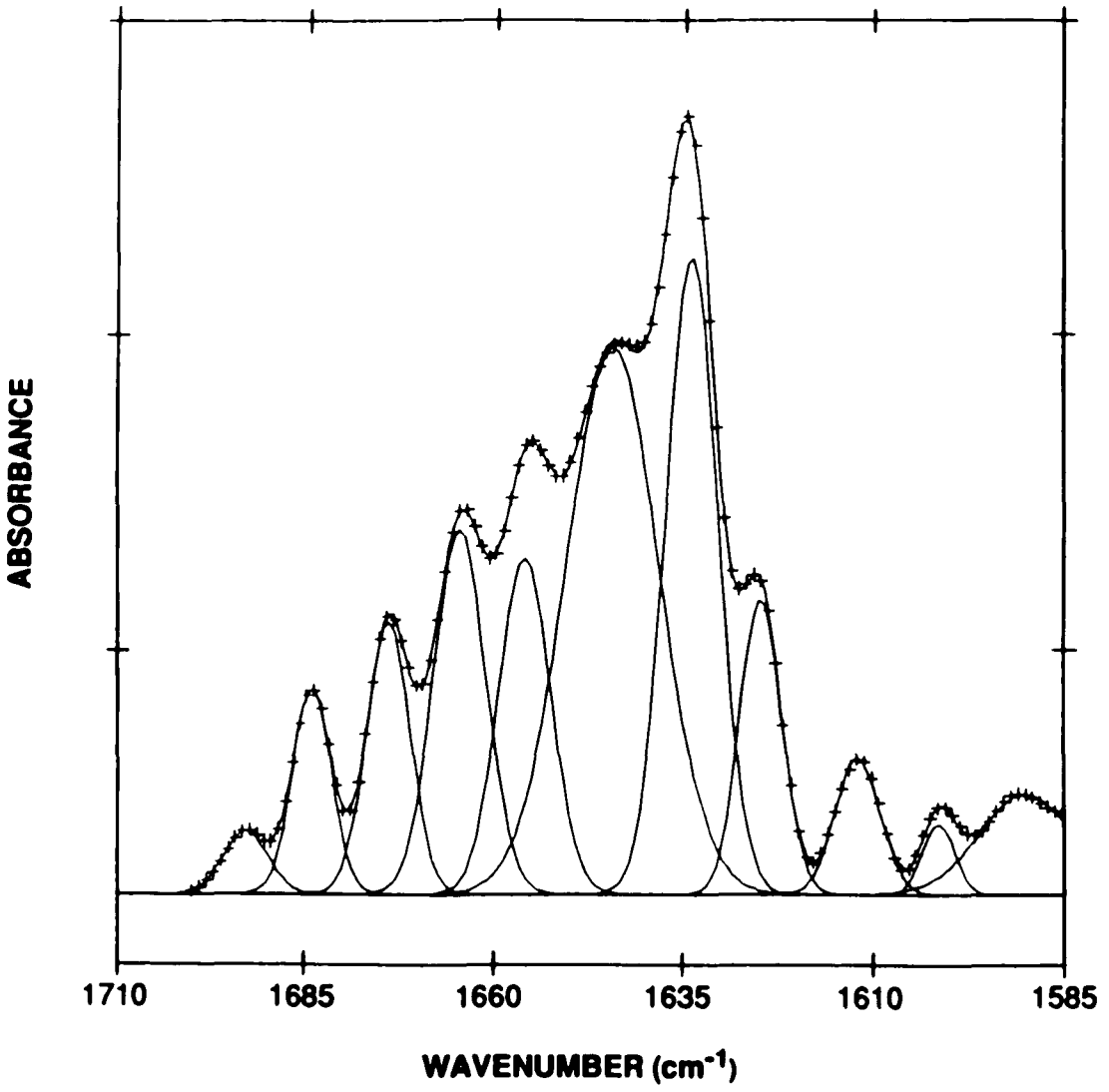


Figure 4-4. Deconvoluted infrared Amide I' band of bovine trypsinogen at pD=6.9, 20 mM Ca^{2+} (++++). Individual Gaussian components and their sum (———).

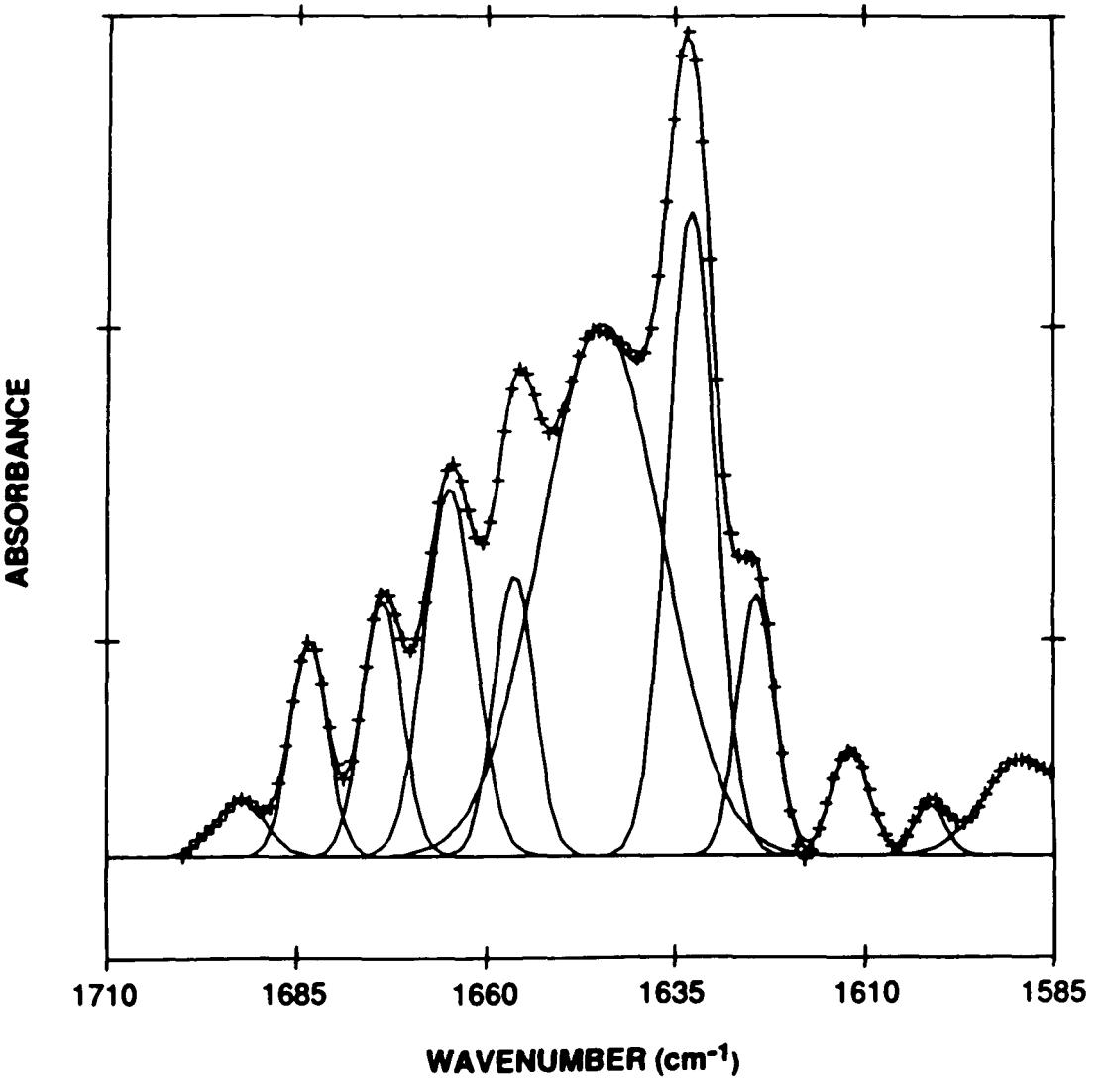


Figure 4-5. Deconvoluted infrared Amide I' band of bovine DIP-β-trypsin at pD=6.9, 20 mM Ca^{2+} (++++). Individual Gaussian components and their sum (———).

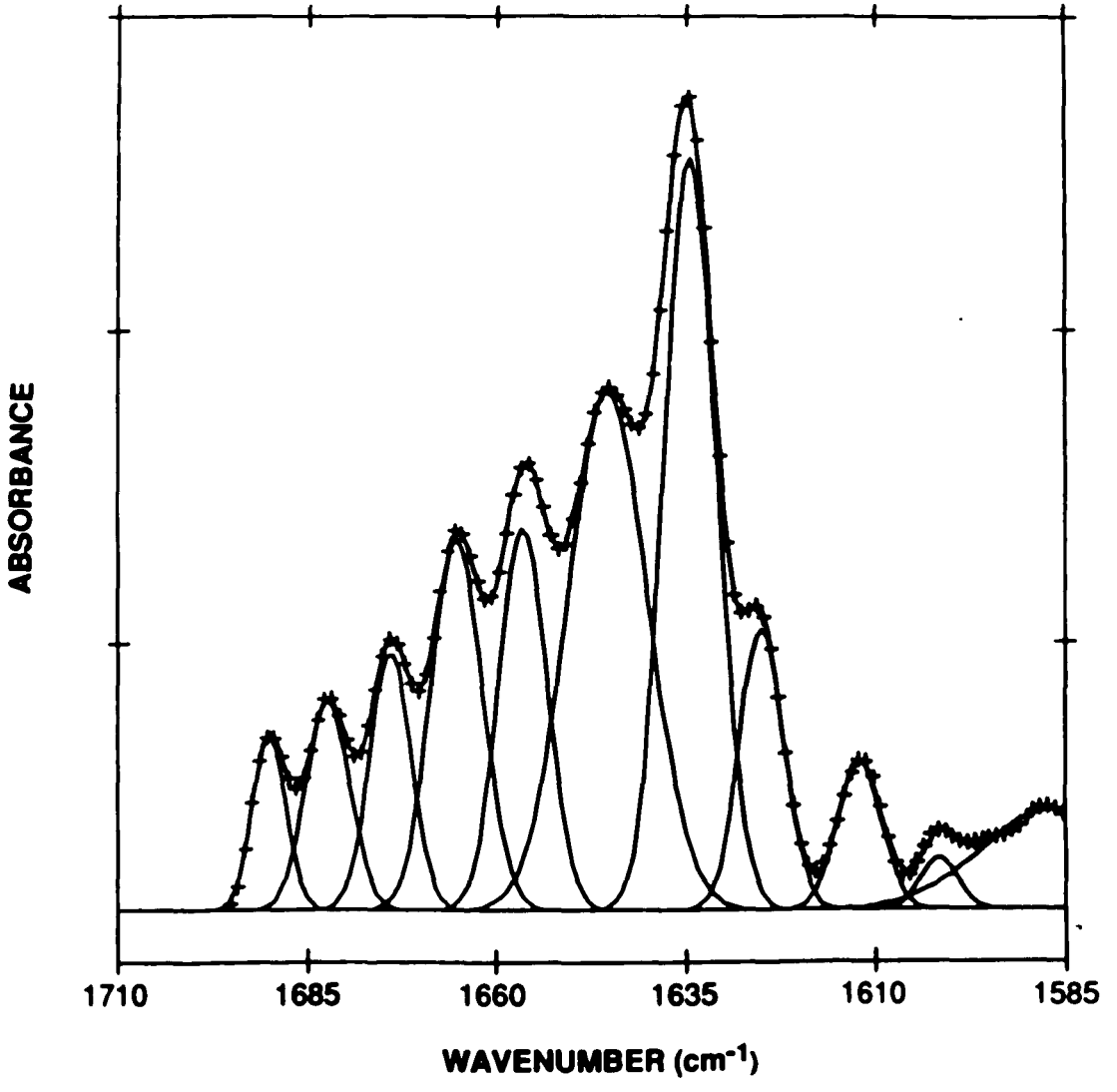
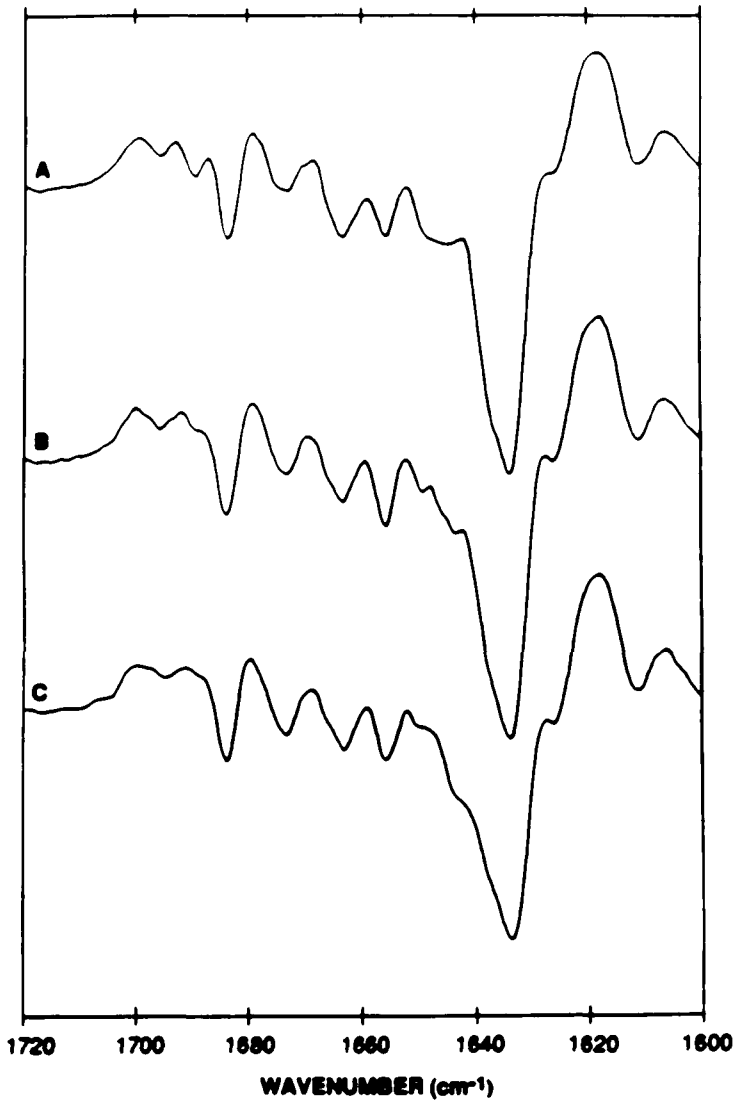
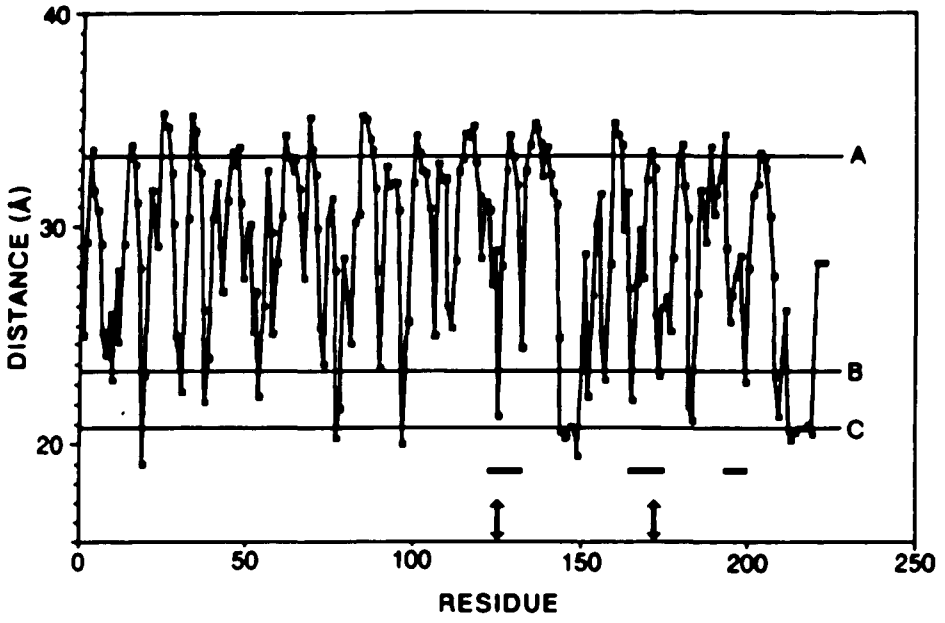


Figure 4-6. Second derivative spectra of the Amide I' region of α-trypsin at pD=6.9, 20 mM Ca^{2+}, at various times after mixing: (a) 45 min, (b) 2.5 hr and (c) 4 hr.[1]



[1]These spectra are normalized so that the peak intensity for the tyrosine band at 1515 cm^{-1} is constant.

Figure 4-7. Linear distance plot of bovine β-trypsin at pH=5.0.[a]



[a]The linear distance value for each residue is calculated by summing the series of distances from its own α-carbon to the α-carbons of the next four residues (see Liebman, 1986). The horizontal lines in the plot correspond to the values calculated for ideal homopolymers in the A) β-sheet, B) 3_{10}-helix and the C) α-helix conformations. The three bars under the plot indicate the regions in bovine trypsinogen found to be disordered in the crystal structure. The two arrows indicate the positions of the autolytic cleavages in bovine trypsin.

Figure 4-8. Sequential outline of autolytic cleavages in bovine trypsin.

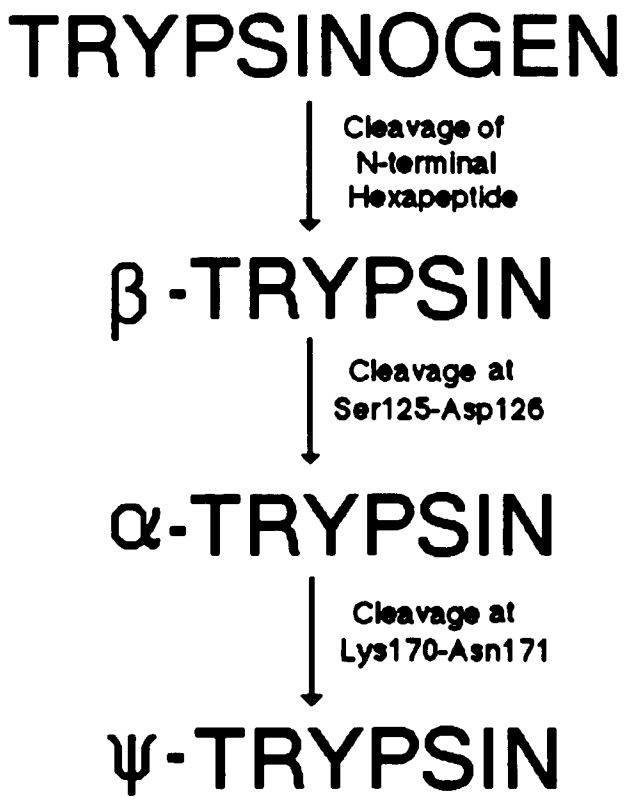
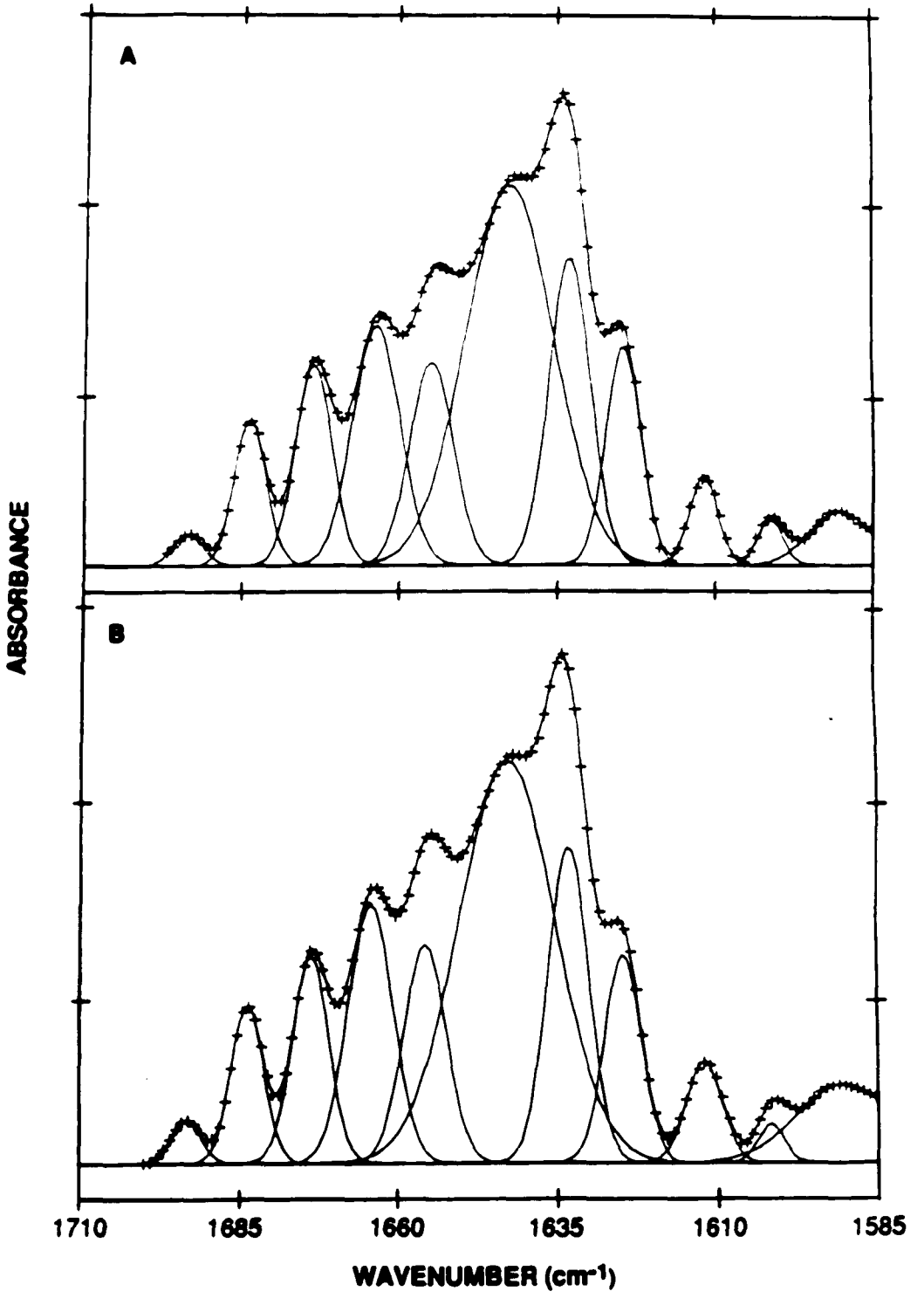


Figure 4-9. Deconvoluted spectra of a) β-trypsin after 20 hr of incubation at pD=6.9 and b) α-trypsin after 4 hr of incubation at pD=6.9 (++++). Individual Gaussian components and their sum (·———).



CHAPTER 5

A STUDY OF A SERIES OF HOMOLOGOUS PROTEINS: SPECTROSCOPIC EXAMINATION AND CONFORMATIONAL ANALYSIS OF THE TRYPSIN-LIKE SERINE PROTEASES

5.1. INTRODUCTION

Fourier transform infrared spectroscopy is a sensitive and informative tool for studying secondary structure in proteins. Absorptions arising from the conformation-sensitive amide vibrational modes provide much information concerning the secondary structure of molecules studied. Component bands in the amide I region of proteins (and other conformation-sensitive regions) have been assigned to various protein conformational elements based on empirical correlations with normal mode calculations (Krimm and Bandekar, 1986) and by studying the spectra of ideal peptide homopolymers whose conformation has been determined independently, usually by x-ray crystallography (Surewicz and Mantsch, 1988a). Because of symmetry considerations and other approximations which do not apply to protein structures as well as the effects of

tertiary structures interactions, the results of normal mode analyses and spectral studies of ideal homopolymers are not directly applicable in interpreting protein spectra. However, several spectra-structure assignments have been confirmed for proteins through quantitative correlations between spectra and results of algorithms for automatic secondary structure analysis (Byler and Susi, 1986).

While use of algorithms for automatic secondary structure analysis to discern correlations with protein infrared spectral features has been partially successful, these algorithms utilize limited models of protein conformation and thus, limit the extent to which infrared component bands can be assigned to different backbone conformations. These algorithms are generally template-based, which function by searching a protein coordinate database for folding patterns which match a predefined template using pattern recognition methods. Template-based methods of protein conformational analysis, by their nature, describe and quantitate only those conformations which resemble the template, within preset constraints. This may exclude a significant portion of specific protein folding types which are described only as 'unordered' or 'random coils' (Leszczynski and Rose, 1986). These terms are misnomers in that these conformations are as ordered and reproducible as other conformations, they are only more difficult to describe and classify (Richardson, 1981).

In Chapter 2 of this thesis an algorithm for description and classification of protein conformation which operates independent of a predefined template was described. It was demonstrated to provide a more encompassing model of protein conformation. Particularly, it has the ability to describe and classify conformations previously assigned only as random coil and to further subclassify certain classical secondary structures in proteins, such as β-strands. The result is a more complete and encompassing description of protein conformation. As a greater portion of folding patterns are classified, the results

of this algorithm provide a more complete basis for discerning correlations between protein conformational elements and conformation-sensitive spectral features.

In the previous chapter it was demonstrated that spectroscopic examination of a protein molecule whose conformation has been perturbed provides a suitable system for studying spectra-structure relationships in proteins. A similar approach is taken here except that a series of different but highly homologous proteins, the trypsin-like serine proteases is examined. Liebman (1986) has performed an in depth analysis of structural similarities and differences in the backbone folding among this group of proteins. This study, and a previous study (Greer, 1981), have outlined the structural organization present among these proteins which consists of alternating regions of structurally conserved and structurally variable folding patterns. In this chapter, the results of studies in which the substructure library calculations for a series of homologous proteins, the trypsin-like serine proteases, are used to discern correlations between protein secondary structure and amide I component bands in infrared spectra.

5.2. METHODS

5.2.1 MATERIALS

α-Chymotrypsin (3X crystallized, CDI) and chymotrypsinogen A (5X crystallized, CGC), both from bovine pancreas, were products of Worthington Biochemical Corp. (Freehold, NJ). Porcine pancreatic elastase (Type III) was purchased from Sigma Chemical Co. (St. Louis, MO). These proteins were used without further purification. The preparation of the trypsinogen and

β-trypsin samples are described in the previous chapter. All other chemicals used were reagent grade.

5.5.2. SPECTROSCOPY

All proteins were prepared as 3% solutions in D_2O buffers. Spectroscopic data were collected at the same pH value used in the crystal structure determinations. The α-Chymotrypsin solution was prepared as a 0.02M citrate buffer (pD=3.5). Chymotrypsinogen and elastase were prepared in 0.02M acetate buffer (pD=5.0). Experimental conditions for β-trypsin and trypsinogen are described in the previous chapter of this thesis.

Experimental conditions for spectroscopic data collection and the subsequent spectral analysis were the same as those described in the previous chapter with the following exceptions. Parameters used in Fourier self-deconvolution of elastase were 14 cm^{-1} and 2.8 for the undeconvoluted halfwidth and resolution enhancement factor, respectively. Similarly, values of 13 cm^{-1} and 2.4 were found to be optimal for deconvolution of both chymotrypsin and chymotrypsinogen. The optimal bandwidth parameter used in Fourier self-deconvolution depends on the original bandwidths of the component bands, which may differ among different proteins. However, as Kauppinen et al., (1981) has demonstrated, the process of Fourier self-deconvolution does not affect the component band areas which are used in these studies in measurements of the relative intensity.

5.2.3. SUBSTRUCTURE LIBRARY CALCULATIONS

Calculation of the substructure library for the 10 serine proteases in this study was performed using the algorithm and parameters described in Chapter 2 of this thesis. Atomic coordinates for all proteins studied were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977). The 10 protein structures used and their reported atomic resolutions are listed in Table 5-1.

5.3. RESULTS

5.3.1. ELASTASE

The amide I region of elastase along with the second derivative and deconvoluted spectra are shown in Figure 5-1. The second derivative spectrum reveals an intense band near 1632 cm^{-1} with other less intense peaks at 1645, 1660 1672, 1684, 1694 and 1699 cm^{-1}. The peak near 1660 cm^{-1} appears to be split into two peaks. However, this may be the result of noise as each of the features on this peak is not sufficiently above the noise level which can be evaluated by examining the region above 1700 cm^{-1} in the second derivative spectrum. The deconvoluted spectrum reveals only one peak at 1660 cm^{-1}. The deconvoluted and curve-fitted amide I region of elastase is shown in Figure 5-2, and the results of curve-fitting this spectrum are listed in Table 5-2. An additional peak at 1623 cm^{-1} is resolved in the deconvoluted spectrum of elastase. The deconvoluted spectrum also shows a strong shoulder on the high frequency side of the intense 1632 cm^{-1} peak. A shoulder is also present in this region of the second derivative spectrum.

5.3.2. CHYMOTRYPSINOGEN AND CHYMOTRYPSIN.

The second derivative spectra of chymotrypsinogen and chymotrypsin are shown in Figure 4-3. Each of these spectra reveal peaks near 1637, 1645, 1655, 1666, 1675, 1682 and 1689 cm^{-1}. The second derivative spectrum of chymotrypsinogen reveals an additional peak at 1697 cm^{-1}. The fourth derivative spectra of the two proteins are shown in Figure 4-4. These spectra reveal additional peaks at 1627 cm^{-1} in both chymotrypsinogen and chymotrypsin. The fourth derivative of chymotrypsinogen also reveals a peak at 1634 cm^{-1} whose intensity is similar to the noise level. Thus, it cannot be assumed to correspond to a real band. The deconvoluted spectrum, shown in Figure 5-5 reveals only a peak at 1637 cm^{-1}. The deconvoluted and curve-fitted spectra of chymotrypsinogen and chymotrypsin are shown in Figures 5-5 and 5-6, and the results of curve-fitting these spectra are also listed in Table 5-2.

5.3.3. SUBSTRUCTURE CALCULATIONS

The calculation of the substructure library for the 10 serine proteases listed in Table 5-1 reveals 353 substructures among these proteins, of which 296 occur within the 5 proteins for which spectroscopic data was collected. As was found for the series of non-homologous proteins studied previously (see Chapter 2), the α-helix is the most frequently occurring substructure. Although the α-helix is the most frequently occurring substructure, the extended strand is the predominant conformation among these proteins. Unlike the α-helix, the extended strands are distributed among many substructures. 13 substructures correspond to 8 residue extended strand substructures. Additionally, numerous

substructures which consist of shorter extended strand structures are present. A total of 123 substructures contain some portion of β-strand.

The description of conformations in the serine proteases is almost complete. The conformation of 96% of the residues in the 5 serine proteases studied spectroscopically is described in with this algorithm. Two proteins, β-trypsin (1TPO) and trypsinogen (2TGA), are 100% described using the substructure library. α-Chymotrypsin (5CHA) has the lowest description at 92 % of the residues.

5.4. DISCUSSION

5.4.1. ELASTASE

The amide I region of the elastase spectrum is similar in its overall shape to the trypsin conformers studied previously (Figure 5-1). The strongest peak occurs at 1632 cm^{-1} as is expected for a predominantly β-sheet protein (Byler and Susi, 1986, Surewicz and Mantsch, 1988a). An additional strong peak appears near 1645 cm^{-1}. Peaks in this region are indicative of irregular (Byler and Susi, 1986) or disordered (this thesis, Chapter 3) structures. A small peak is present at 1638 cm^{-1}. This peak has been assigned to the 3_{10} helix structure (Halloway and Mantsch, 1989; Prestrelski and Byler, unpublished results). The relative intensity of this component (Table 5-2) agrees well with the x-ray structure which shows a short stretch of 3_{10} helix in elastase (Sawyer et al.,1978).

It is interesting to note the absence of a peak near 1655 cm^{-1} in the second derivative spectrum (Figure 5-1). A peak is expected in this region because elastase is known to contain α-helices, which absorb near 1655 cm^{-1} in

deuterated proteins (Byler and Susi, 1986). The deconvoluted spectrum in Figure 5-1 clearly reveals a peak in this region. The absence of this band in the second derivative spectrum illustrates a potential source of artifacts in interpreting second derivative spectra. The second derivative of a Lorentzian band contains positive side lobes which can interact with the negative peaks of adjacent bands (Maddams and Tooke, 1982). It appears that in this case the positive lobes of adjacent, stronger bands have completely canceled the negative peak which is expected near 1655 cm^{-1}.

Because they result in greater resolution enhancement, calculation of higher order derivatives can sometimes reveal peaks not observed in lower order derivative spectra. A fourth derivative spectrum was calculated for elastase. However, the resulting decreased signal-to-noise ratio was such that any resolution enhancement was not discernable because of the added in noise.

The variability of the relative intensities is higher in elastase than found in the trypsin spectra (Table 5-2). This is most likely because the peaks in the deconvoluted spectrum are not as well defined as in the trypsin examples. Those peaks with the greatest variation in relative intensity values correspond to peaks which are closely overlapped. This results in increased variability because it is possible that more than one set of parameters can produce an adequate 'fit'.

5.4.2. CHYMOTRYPSIN AND CHYMOTRYPSINOGEN

The spectra of chymotrypsin and chymotrypsinogen are similar to the spectra of the other serine proteases studied. The strongest peak is near 1637 cm^{-1}. The low frequency peaks which are assigned to extended strands in these proteins are higher in frequency relative to the other serine proteases

studied (see Table 5-2 and table from). No peak corresponding to 3_{10} helices is present in either of these spectra, although one is expected because short stretches of this conformation are observed in the crystal structures (Wang et al., 1985; Blevins and Tulinsky, 1985). However, the high values of the component peaks assigned to extended strands, which are several fold greater in intensity, probably results in the 1638 cm^{-1} band, if present, being overlapped by the stronger peak. The peak at 1697 cm^{-1} in the spectrum of chymotrypsinogen disappears upon conversion to α-chymotrypsin. It has been suggested (Casal et al., 1989) that peaks at this high of a frequency do not arise from amide I vibrations but from unionized carboxyl groups of acidic side chains. Thus, the exact nature of this peak cannot be determined with present information. However, the low relative intensity of this band will not significantly affect the fractional intensities of the amide I' component peaks, if this is the case.

A large difference in the relative intensity of the component peak near 1646 cm^{-1} is observed upon conversion from chymotrypsinogen to α-chymotrypsin. Bands in this region have been assigned to irregular (Byler and Susi, 1986; Surewicz and Mantsch, 1988a) and disordered structures (Chapter 3 of this thesis). The increase in relative intensity upon conversion, from 0.02 in chymotrypsinogen to 0.11 in chymotrypsin, is consistent with the results of crystallographic studies (Wang, et al., 1978; Blevins and Tulinsky; 1985). The crystal structures indicate an increase in disordered conformations upon this conversion which results from cleavage of two dipeptides from the chymotrypsinogen molecule. The increase in disorder is the opposite of the disorder-order transition observed in trypsinogen-trypsin conversion described in the previous chapter. In both cases the relative intensity of the bands assigned to disordered structures is consistent with crystallographic information.

The variability in the relative intensities for chymotrypsinogen is very low, similar to values found for trypsin (see Table, 5-2). However, the variability in the chymotrypsin spectra is somewhat higher, particularly of the bands in the 1625-1640 cm^{-1} region. This most likely occurs as a result of the weak, poorly defined band present at 1628 cm^{-1} in the chymotrypsin spectrum which only appears as a shoulder on the higher frequency peak at 1637 cm^{-1}.

5.4.3. QUANTITATIVE CORRELATIONS WITH SUBSTRUCTURE LIBRARY

The use of a series of molecules whose three dimensional structures are highly homologous can be advantageous in studies of spectra-structure correlations because the high degree of structural similarity is expected to result in similar spectra. This has proven to be true for the spectra of the five trypsin-like serine proteases studied here. The high degree of similarity in the spectra highlights the differences, facilitating spectra-structure assignments. A further advantage in using such a system is that it allows the investigator to examine spectral features previously assigned to specific conformations. The conformational makeup of homologous proteins falls into two distinct categories, the structurally conserved region and the variable regions (Greer, 1981). The conformation of approximately 80 % of the residues among this series of proteins is topographically equivalent (Liebman, 1986). It is expected that bands assigned to conformations which comprise the structurally conserved regions, which are the β-stranded barrels and the α-helices in the serine proteases (Greer, 1981), should remain relatively constant in the spectra while bands assigned to variable regions are expected to vary. (If this were not in fact true, the basis for using vibrational spectroscopy for conformational studies would be invalid.) Thus previous assignments of bands to β-strands,

α-helices and 3_{10} helices can be evaluated for consistency within this homologous series.

Other investigators have previously attempted quantitative estimates from resolution-enhanced infrared spectra. The relative intensity of the bands assigned to various conformations, determined with curve-fitting techniques, is taken as an estimate of its fractional composition (Byler and Susi, 1986; Susi and Byler, 1986). This procedure operates under the assumption that the molar absorptivities of the different conformations are approximately the same for a given vibrational mode. However, due to lack of a sufficient experimental system for studying the molar absorptivities among different conformational states, this assumption has not been established for polypeptides (Surewicz and Mantsch, 1988a). Limited experimental evidence suggests that the molar absorptivities of various conformations are relatively equivalent (Mantsch et al., 1989; Jackson et al., 1989).

Table 5-3 lists the estimates of fractional composition of α-helices and extended strands from the substructure calculations and from curve-fitting the deconvoluted IR spectra of each of the proteins studied. IR estimates for fractional composition for extended strands are calculated by summing the fractional intensity (band areas) of the bands assigned to extended strands (i.e., ca. 1625, 1635 and 1675 cm^{-1} ,Byler and Susi, 1986[1]). The fractional intensity of the band near 1655 cm^{-1} has been used for the estimate of fractional composition of α-helix (Byler and Susi, 1986; Surewicz and Mantsch, 1988a). Estimates from the substructure library are determined by summing the residues

[1]The relative intensity of bands in the 1638-1640 cm^{-1} range are not included in this calculation because bands in this region have more recently been assigned to 3_{10} helices (Halloway and Mantsch (1989), Prestrelski and Byler, unpublished results.)

which adopt each conformation using the boundaries outlined in the chapter describing the substructure library calculations (see Chapter 2).

The substructure library results for the fractional composition of α-helices among the serine proteases studied are very consistent (see Table 5-3). These results indicate that each of these proteins contains approximately 8 to 9 % α-helical structure. Conversely, estimates for fractional composition of α-helix from infrared spectra in this study are quite variable, ranging from 7% in trypsinogen to 26% in chymotrypsinogen. Estimates for helical structure from the deconvoluted, curve-fitted spectra are in good agreement with substructure library information for trypsinogen and elastase, but the spectroscopic estimates are relatively high for the other proteins studied (see Table 5-3). Thus, it would appear that other conformations result in amide I vibrations which absorb near 1655 cm^{-1}. The studies of zymogen activation, autolysis and active site inhibition in trypsin, described in the previous chapter, have provided strong evidence that another conformation, the first autolytic loop in β-trypsin, absorbs in the 1655 cm^{-1} region. The high relative intensity value for the peaks near 1655 cm^{-1} in chymotrypsin and chymotrypsinogen suggest that these molecules also contain additional folding patterns which absorb in this region. Surewicz and Mantsch (1988 b,c), and Surewicz et al., (1988) , based on an unexpected band around 1655 cm^{-1} in several small peptides and proteins determined by circular dichroism to contain no helices, have suggested that this peak arises from an unexchanged irregular conformation or an 'atypical nonperiodic' conformation. It is also possible that certain types of turns absorb in the 1655 cm^{-1} region. Normal mode calculations predict a mode at 1656 cm^{-1} for turns (Krimm and Bandekar, 1986) although no assignment has been established for actual protein molecules in which observed frequencies can deviate significantly from those predictions from model systems.

Another possible explanation for the high intensity in the 1655 cm^{-1} band is that additional intensity arises from regions of irregular structure which are not deuterated as these structures are expected to absorb in the 1650-1660 cm^{-1} region (Surewicz and Mantsch, 1988a). However, the amide II to amide II' intensity ratio indicates that the protein is essentially completely deuterated. Additional experiments were performed over several days to examine deuteration effects. The spectrum collected after 48 hr indicates no significant differences in intensity of the 1655 and 1645 cm^{-1} bands.

The substructure library provides a means for examining similarities among protein structures. The loop structure in the region of the first autolytic cleavage in β-trypsin assigned to 1655 cm^{-1} is among the substructures found in the serine proteases. The library was examined to determine whether this conformation appears in other proteins. This may explain the anomalously high intensity of the 1655 cm^{-1} components in the two chymotrypsin forms. However, no matches were found for this substructure among any of the other proteins studied spectroscopically. Thus it would appear that other structures in addition to the α-helix and the loop described in β-trypsin may absorb near 1655 cm^{-1}. As stated previously, these results suggest caution in interpreting amide I' component bands around 1655 cm^{-1} as resulting entirely form α-helices.

Estimates for fractional composition of extended strands from infrared spectroscopy and the substructure calculations are in general agreement (see Table 5-3). This difference may be due in part to the somewhat more arbitrary criteria set for the conformational parameters of β-strands. Unlike the α-helix, whose conformational parameters are quite rigid (Richardson, 1981; Barlow and Thornton, 1988), the β-strand is quite variable (Richardson, 1981; Richards and Kundrot, 1988).

While the frequency position for the amide I component band assigned to α–helices (ca. 1655 cm^{-1}) appears to remain relatively constant among the proteins studied, the frequencies for those bands assigned to extended strands vary much more widely, usually within the 1620-1638 cm^{-1} region. In addition, more than one band often appears in this region for a single protein. The frequency position for the high frequency extended strand peak has not been unequivocally established because of overlap with bands thought to arise from turns. However, peaks near 1672 cm^{-1} are a likely candidate (Byler and Susi, 1986; Susi and Byler; 1988). The low frequency bands assigned to extended strands for elastase, β-trypsin and trypsinogen are near 1633 and 1624 cm^{-1} while those for chymotrypsin and chymotrypsinogen are near 1637 and 1629 cm^{-1}. One possible explanation is that the proteins differ in the degrees of deuteration. However, in all cases examination of the amide A and amide II bands, which shift greatly upon deuteration, indicate essentially complete deuteration. Additionally, further experiments were carried out over longer time periods to insure that no further changes due to deuteration occurred.

Another possible explanation is that the conformations of the strands are somewhat different which could result in somewhat different frequencies (Surewicz and Mantsch, 1988a). The variability of the β-strand conformation in globular proteins is well documented (Richardson, 1981; Richards and Kundrot, 1988). This variability results in several different types of conformations for β-strands which differ in the extendedness and curvature as was demonstrated by the substructure library calculations described in Chapter 2. Thus, the portion of the substructure library which describes and classifies β-strands was examined for possible conformational differences which may exist among the proteins studied.

Of the 123 substructures which consist entirely of or contain portions of extended strands, 92 are observed among the 5 proteins studied spectroscopically. If those substructures which contain portions of extended strands which are completely overlapped by other substructures are omitted, 32 substructures remain. These 32 substructures, therefore, span the entire set of extended strands in the 5 proteins. Each of the 5 proteins is comprised of a different set of these substructures although many of the substructures are shared.

Table 5-4 outlines the relative similarity among the extended strands which make up each protein studied as well as the similarity in the frequency positions of the low frequency (1620-1638 cm^{-1}) bands assigned to this conformation. The results of the substructure library estimates of similarity in the extended strands are consistent with those of Liebman (1986) who studied global structural similarity using an RMS superpositioning technique. The correlation between the two measures is very good. That is, the greater the degree of similarity in the conformation of the β-strands among any two proteins, as estimated from substructure library calculations, the greater the similarity in the frequencies for those low frequency bands assigned to extended strands. This provides a plausible explanation for the variability in the peak frequencies assigned to β-strands. Correlations between individual strands, or groups of strands, and either of the two low frequency bands in not possible, however because the problem is indeterminate. There are many more extended strand conformations (i.e., substructures) than there are bands. Additionally, the different geometry of the different strands is likely to result in somewhat different inter-strand interactions. Interactions between strands as well as their geometries affect the frequency (Krimm and Bandekar, 1986), although the two are intimately related. However, such a correlation provides a basis for

designing experimental and theoretical studies for further probing this question. Correlations between different types of strand conformations and particular bands in the amide I, or other conformations-sensitive vibrational modes, is potentially of great diagnostic value.

Results of recent studies (Halloway and Mantsch, 1989; Prestrelski and Byler, unpublished results) have indicated that bands absorbing in the 1638-1640 cm^{-1} region arise from absorptions of 3_{10} helices rather than extended strands as previously proposed (Byler and Susi, 1986), The x-ray structures indicate regions in elastase, chymotrypsin and chymotrypsinogen which adopt a 3_{10} helical conformation (Sawyer et al., 1978; Blevins and Tulinsky, 1985; Wang et al., 1985). Conversely, no 3_{10} helices are observed in trypsin or trypsinogen (Felhammer et al, 1977; Marquart et al., 1983). These data are consistent with the results of infrared spectroscopic studies of the serine proteases. Trypsinogen and trypsin, which contain no 3_{10} helices, do not reveal amide I component peaks in the 1638-1640 cm^{-1} region (Chapter 4). Elastase has a small peak at 1638 cm^{-1} whose relative intensity (Table 5-2) is roughly in agreement with the amount of 3_{10} helix observed crystallographically. Chymotrypsinogen and chymotrypsin have 3_{10} helices similar to elastase. However, the position of the low frequency extended strand peaks is relatively high when compared with other serine proteases, and other proteins (Byler and Susi, 1986), and it appears that these intense bands may overlap a much smaller band which may be present near 1639 cm^{-1}.

Table 5-5 summarizes the amide I' band assignments for the 5 trypsin-like serine proteases as proposed from the results of these studies.

5.5. CONCLUSION

The use of a non-template based method for describing and classifying protein conformation has resulted n a higher level of description of the backbone folding patterns. For a series of homologous proteins, use of this method has resulted in almost complete description of the protein backbone conformation, and thus, it has allowed examination of conformation-sensitive modes in IR spectra in a manner not previously possible when using template based algorithms.

The results of these studies have suggested that in addition to α-helices and certain types of loops, other conformations absorb in the 1655 cm^{-1} region. This agrees with the results of Chapter 4 which demonstrated intensity in this region arising from non-helical structures. Until recently, it was assumed that only α-helices absorb in this region. Also, the higher level description of the β-strand conformation has afforded the potential to examine the variability and the presence of several peaks which arise in infrared spectra due to absorptions of β-strands. The results indicate a strong correlation between the similarity of β-strands among proteins studied and the frequency positions of the low frequency amide I' component peaks assigned to these structures. Thus, IR spectroscopy is a potential diagnostic tool for examining the finer structure of extended strands and β-sheets in proteins. Spectroscopic results indicate that the recent new assignment of amide I peaks in the 1638-1640 cm^{-1} region to 3_{10}-helices is consistent with crystallographic data for the serine proteases.

Table 5-1. Four Letter Codes of the 10 Serine Proteases Used for Substructure Library Generation

Code	Protein	Source	Atomic Resolution[1]
1SGT	Trypsin	*Streptomyces griseus*	1.8
1TPO	β-Trypsin	Bovine pancreas	1.7
2ALP	α-Lytic Protease	Lysobacter enzymogenens	1.7
2CGA	Chymotrypsinogen A	Bovine pancreas	1.8
2PKA	Kallikrein	Porcine pancreas	2.05
2SGA	Proteinase A	*Streptomyces griseus*	1.5
2TGA	Trypsinogen	Bovine pancreas	1.5
3EST	Elastase	Porcine pancreas	1.65
3RP2	Protease II	Rat mast cell	1.9
5CHA	α-Chymotrypsin	Bovine pancreas	1.67

[1] In angstroms.

Table 5-2. Peak Positions and Relative Intensities for the Serine Proteases used in this Study

Chymotrypsin			Chymotrypsinogen			Elastase		
ν^a	A^b	σ^c	ν	A	σ	ν	A	σ
1629	0.15	0.028	1627	0.07	0.006	1623	0.18	0.031
1638	0.25	0.031	1637	0.40	0.009	1632	0.22	0.028
						1638	0.05	0.013
1646	0.11	.0.006	1647	0.02	0.002	1645	0.29	0.005
1655	0.24	.0.009	1653	0.26	0.007	1657	0.10	0.008
1667	0.11	0.005	1666	0.13	0.005	1664	0.06	0.020
1675	0.05	0.003	1675	0.05	0.002	1673	0.06	0.009
1682	0.05	0.001	1682	0.05	<0.001	1684	0.03	<0.001
1689	0.04	<0.001	1689	0.02	<0.001	1693	0.01	<0.001
			1697	0.01	<0.001	1700	0.01	<0.001

[a] Wavenumbers in cm^{-1}.
[b] Mean relative intensity calculated from three independent experiments
[c] Standard deviation in relative intensity.

Table 5-3 Comparison of Estimates of Fractional Composition of α-Helix and Extended Strand from the Substructure Library and Infrared Results[a]

	α-Helices		Extended Strands	
Protein	X-ray	FTIR	X-ray	FTIR
β-Trypsin	9	15	42	39
Trypsinogen	9	7	40	35
Elastase	8	10	37	46
Chymotrypsin	8	24	42	45
Chymotrypsinogen	8	26	45	52

[a] See text for description of estimates.

Table 5-4. Comparison of Extended Strand Substructures and Peak Frequencies Among the Serine Proteases[a]

	1TPO	2TGA	3EST	2CGA	5CHA
1TPO	0	0.85	0.63	0.50	0.56
2TGA	1	0	0.72	0.55	0.49
3EST	4	3	0	0.45	0.42
2CGA	5	6	9	0	0.72
5CHA	8	9	12	3	0

[a] Each upper triangular matrix element corresponds to the ratio of the number of extended strand residues which occur in equivalent substructures to the total number of extended strand residues. Each lower triangular element correspond to the sum of the absolute differences between the low frequency extended strand peaks.

Table 5-5. Summary of Amide I Spectra-Structure Assignments for the Serine Proteases Examined in this Study

Band Frequency	Conformation
1625	extended strand[a,b]
1635	extended strand[a,b]
1639	3_{10} helix
1645	irregular[a,b], disordered[a]
1655	α-helix[b], loops[a], other
1664	turns[b]
1674	extended strand[b], possibly Type II β-turn[a]
1683	turns[b], possibly Type II β-turn[a]
1689	turns[b]
1695	turns,[b] possibly carboxyl C=O[c]

[a] Proposed or corroborated from this study.
[b] Assigned in previous investigations, see Surewicz & Mantsch (1988) and references therein.
[c] See Casal et al., 1989.

Figure 5-1. Amide I region of the infrared spectrum of porcine pancreatic elastase at pD=5.0 (a) second derivative spectrum (b) non-resolution enhanced IR absorption spectrum and (c) spectrum after Fourier self-deconvolution.

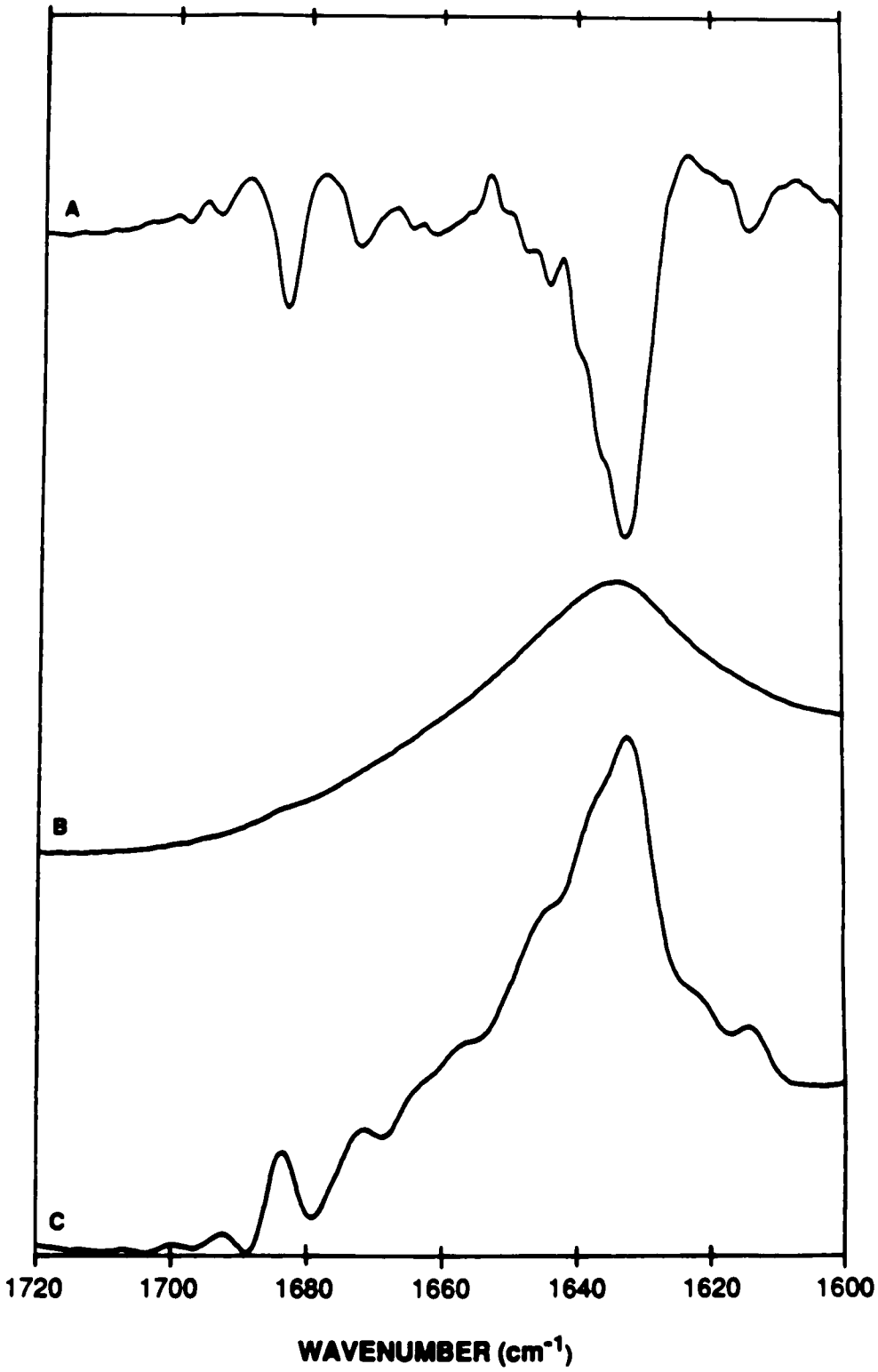


Figure 5-2. Deconvoluted infrared Amide I' band of porcine pancreatic elastase at pD=5.0 (++++). Individual Gaussian components and their sum (———).

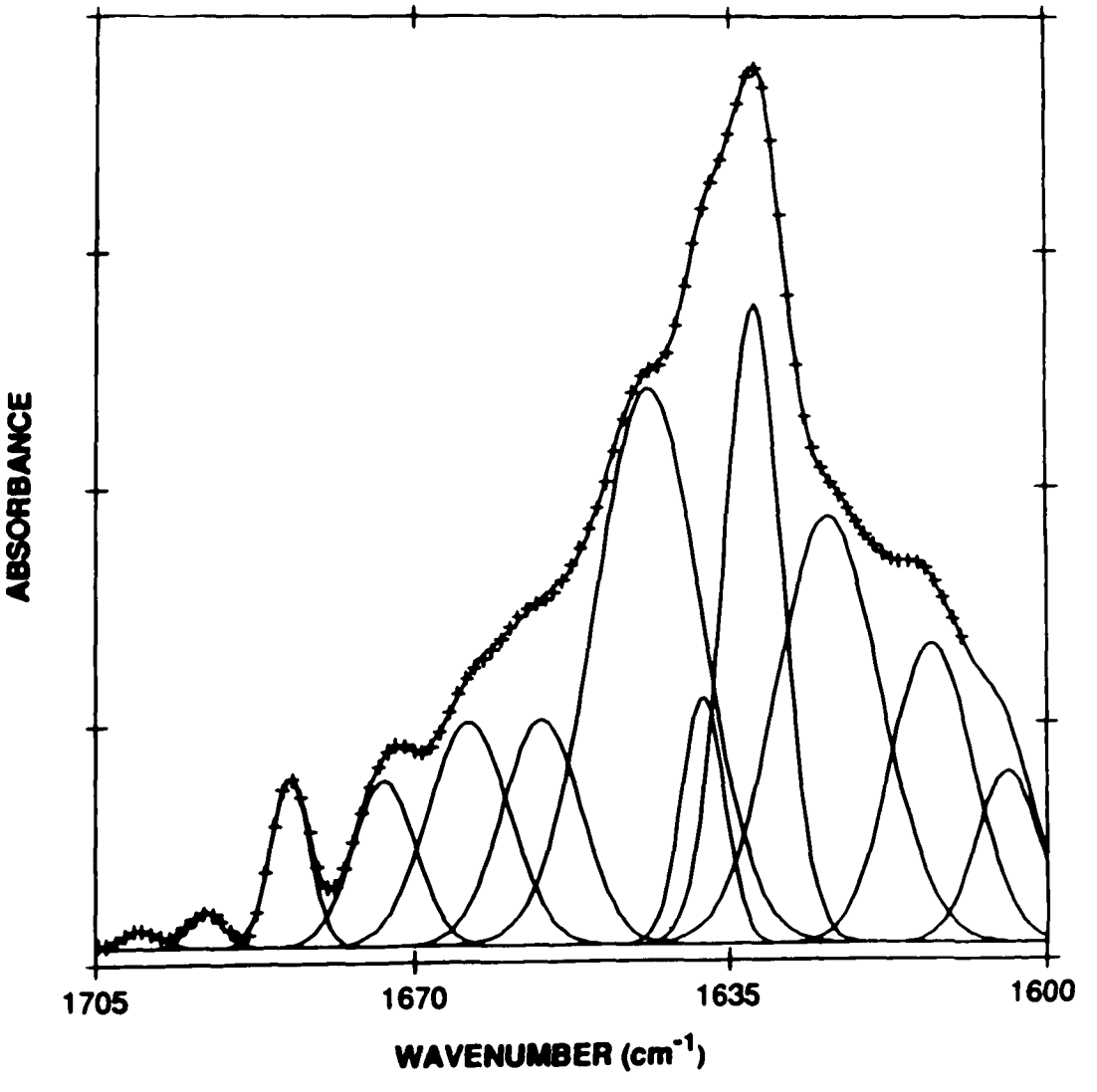
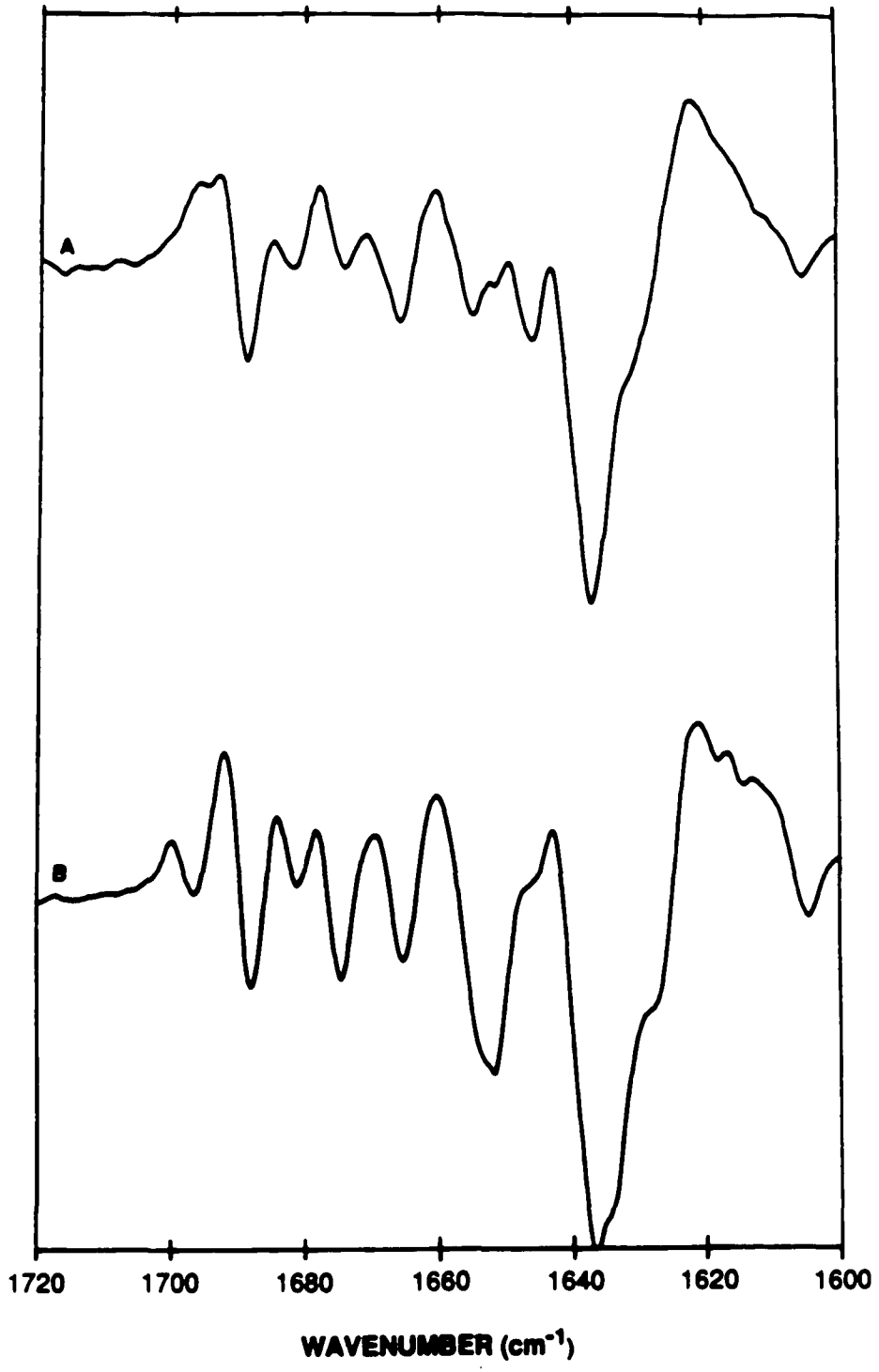
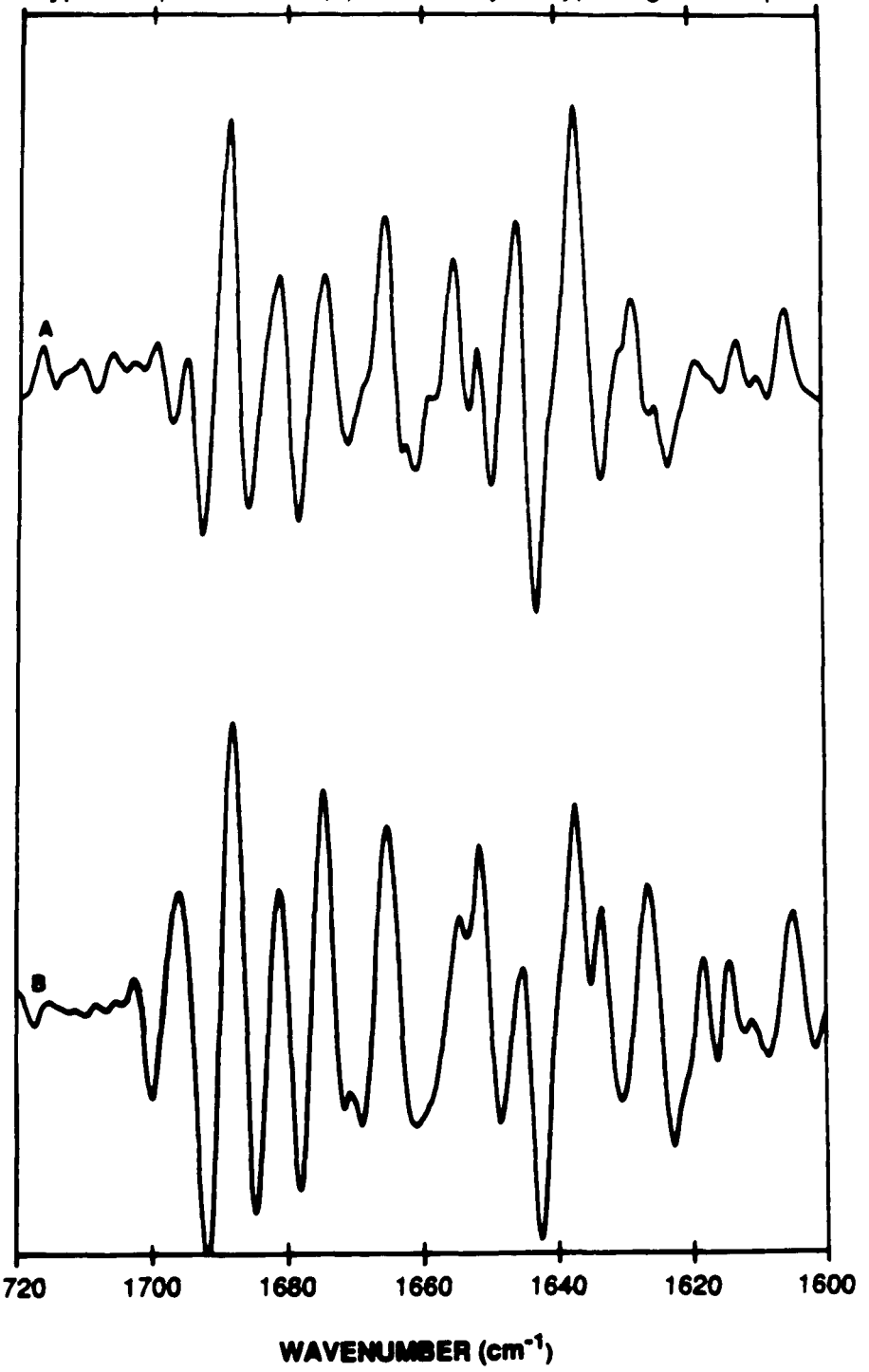


Figure 5-3. Second derivative spectra of the Amide I' region of (a) bovine α-chymotrypsin at pD=3.5 and (b) bovine chymotrypsinogen A at pD=5.0.[1]



[1]These spectra are normalized so that the peak intensity for the tyrosine band at 1515 cm^{-1} is constant.

Figure 5-4. Fourth derivative spectra of the Amide I' region of (a) bovine α-chymotrypsin at pD=3.5 and (b) bovine chymotrypsinogen A at pD=5.0.[1]



[1]These spectra are normalized so that the peak intensity for the tyrosine band at 1515 cm^{-1} is constant.

Figure 5-5. Deconvoluted infrared Amide I' band of bovine α-chymotrypsin at pD=3.5 (++++). Individual Gaussian components and their sum (———).

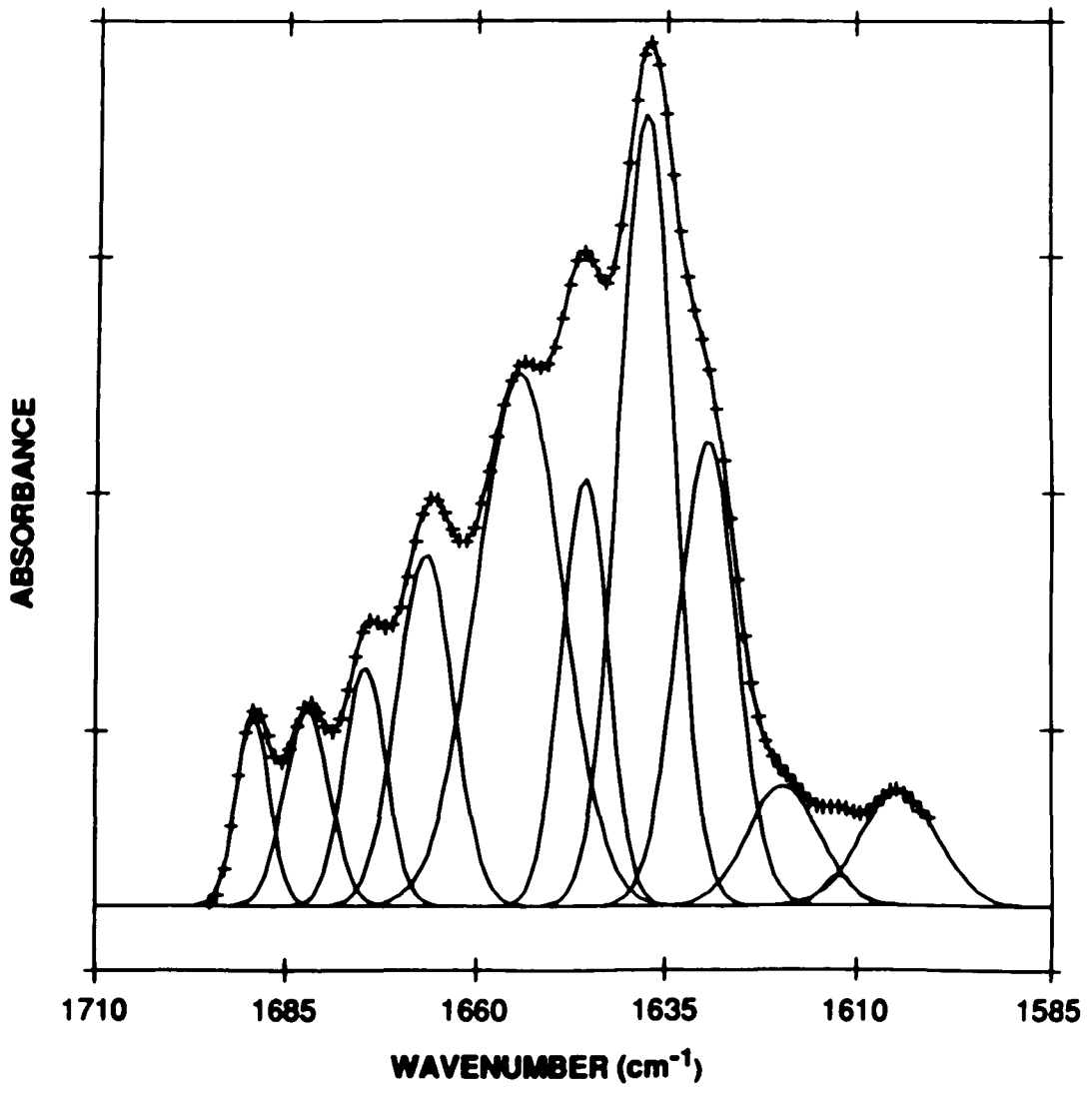
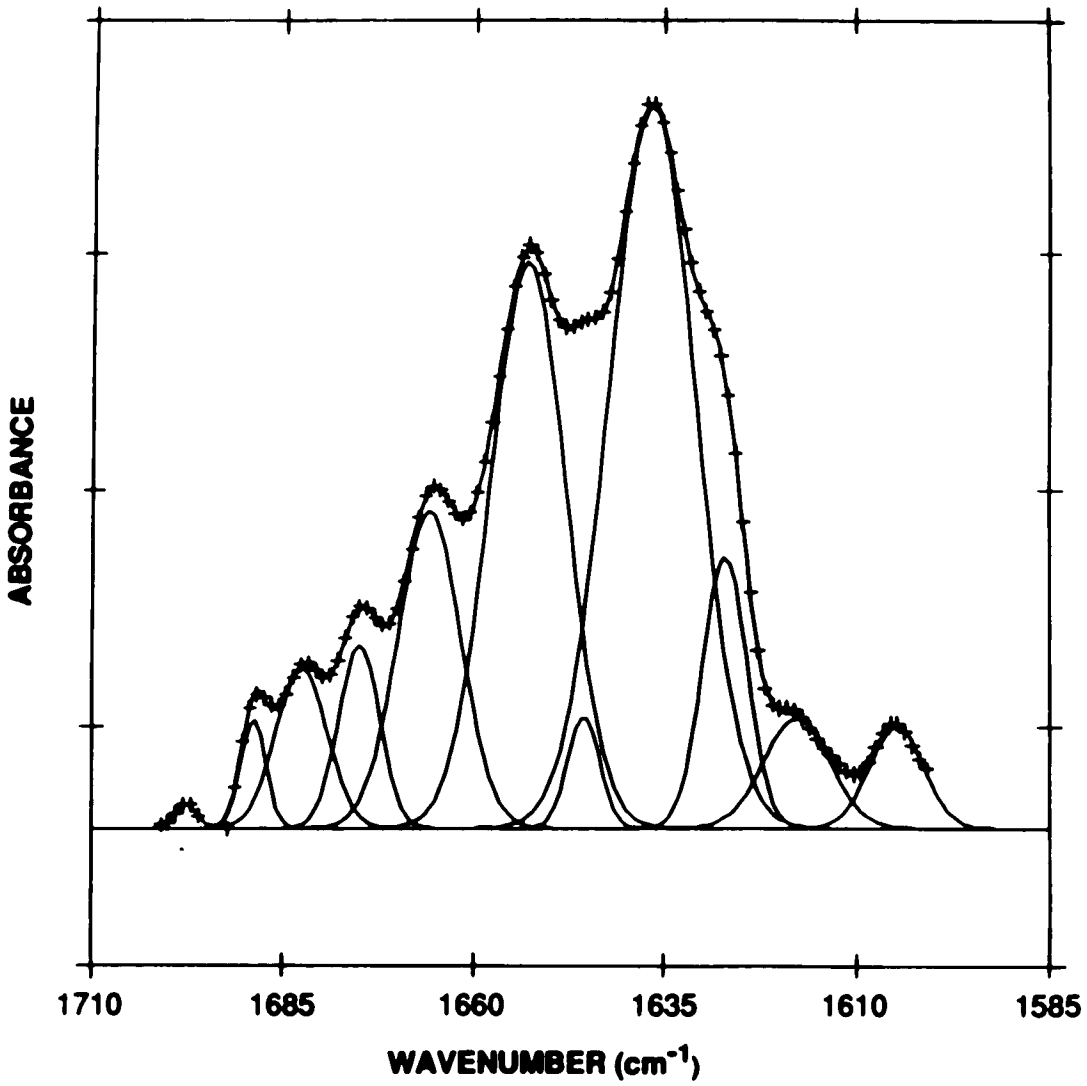


Figure 5-6. Deconvoluted infrared Amide I' band of bovine chymotrypsinogen A at pD=5.0 (++++). Individual Gaussian components and their sum (———).



CHAPTER 6

CONCLUSION

Vibrational spectroscopy is a valuable tool for studying molecular structure and conformation. The potential energies associated with several infrared active vibrational modes of proteins are sensitive to the backbone conformation of polypeptides, and thus, infrared spectroscopy has been recognized as a valuable method for studying protein folding. To extract conformational information from these conformation-sensitive vibrational modes, correlations between their component bands and protein conformational elements are essential. Proteins are large and complex and because of this, classical tools for theoretically assigning spectral bands to structures, calculation of the frequencies and intensities of the normal vibrational modes, are not currently applicable to proteins. Thus, an empirical approach is necessary to discern these essential spectra-structure assignments. Additionally, available methods for definition and classification of protein conformational elements do not account for certain folding patterns which comprise a significant portion of the backbone. These conformations also contribute to the vibrational spectra of

proteins. Therefore, it is essential to address the problem of classifying protein secondary structure in a more encompassing manner.

The specific objective of this thesis project has been the further development of infrared spectroscopy as a method for the determination of protein conformation in aqueous solution. This has been achieved through an integrated approach which involves spectroscopic studies in addition to development of a more advanced algorithm for protein secondary structure analysis. Specifically, this thesis has addressed the problems of discerning spectra-structure relationships and classification of proteins secondary structure through: 1) development of a more encompassing method for definition and classification of protein conformational elements from atomic coordinate data by development and implementation of an algorithm which classifies backbone folding patterns independent of a pre-established structure template (Chapter 3), 2) development of an approach for discerning correlations between infrared spectral features and protein conformational elements which involves examination of a series of closely related protein molecules, the bovine trypsin system, for which high-resolution crystallographic structures have been solved (Chapter 4) and, 3) extension and evaluation of these and previously proposed spectra-structure correlations by studying a homologous series of proteins utilizing both infrared spectroscopy and the new algorithm for protein conformational classification described in objective 1 (Chapter 5).

Previous spectroscopic studies have indicated that some of the infrared bands, particularly the amide I band, arising from the conformation-sensitive modes in infrared spectra provide more information concerning conformation than is presently revealed by algorithms for secondary structure analysis. These are typically template-based algorithms which describe protein

secondary structure using a limited four state model (i.e., α-helices, β-strands, turns and undefined structures) and therefore neglect a significant portion of the proteins conformation. A specific objective of this thesis project was to develop an algorithm for classification of protein conformation in a more complete and encompassing manner. This was accomplished through development of an algorithm for description and classification of protein conformation which functions independent of template.

Application of this algorithm produces a library of conformational elements, i.e. substructures, which consists of folding patterns observed repeatedly in the structure data set. This algorithm was demonstrated to provide a higher level and more comprehensive description of protein conformation. Specifically, this algorithm has the ability to classify and describe conformations which have no apparent structural regularity and also subclassify conformations previously assigned to a single classification.

Spectroscopic studies of conformational perturbation of trypsin molecules described above, interpreted in light of the information from analysis of crystallographic data, have resulted in several new spectra-structure correlations. These results provide strong evidence that certain types of loop structures may absorb in the region of 1655 cm^{-1}. Previously, Amide I' infrared bands near 1655 cm^{-1} have been interpreted as arising solely from α-helices. These new data suggest caution in interpreting this band as arising solely from α-helices. It appears that this is the first assignment of a particular conformational region in a protein to an individual component band in the amide I region.

The results of studies on trypsin conformers have also demonstrated that regions of protein molecules which are known from crystallographic experiments to be disordered absorb in the 1645 cm^{-1} region and also that type

II β-turns absorb in the region of 1672 -1685 cm^{-1}. Further, these results also corroborate assignment of the low frequency component of extended strands to bands below 1636 cm^{-1}. Previous to this study no reports have been published which examine the variability which is inherent in spectroscopic measurements and data analysis such as was performed here. The results of multiple measurements have allowed us to estimate the variability present in component band areas calculated by curve-fitting resolution-enhanced IR spectra. We estimate that this approach to data analysis and interpretation is sensitive to changes of 0.01 or less in the relative intensities of component bands in spectra whose peaks are well-resolved. In addition to discerning new spectra-structure assignments, this study has established a means for studying such spectra-structure relationships, examination of a series of closely related molecules for which high-resolution atomic coordinates have been determined. Further, spectroscopic studies of the process of autolysis have demonstrated the utility of infrared spectroscopy in studying functionally significant problems.

For the homologous series of molecules studied, the serine proteases, the use of a non-template based method for describing and classifying protein conformation results in a significantly more comprehensive level of description. For this series of homologous proteins, use of this method has resulted in almost complete description and classification of the protein backbone conformations, and thus, it has allowed examination of conformation-sensitive modes in IR spectra in a manner not previously possible when using template based algorithms.

The results of these studies have suggested that in addition to α-helices and certain types of loops, other conformations absorb in the 1655 cm^{-1} region, in agreement with results for the trypsin conformers. As stated above, until recently it was assumed that only α-helices absorb in this region. Also, the

higher level description of the β-strand conformation has afforded the potential to examine the variability and the presence of several peaks which arise in infrared spectra due to absorptions of β-strands. The results reported here have indicated a strong correlation between the similarity of β-strands among proteins studied and the similarity in the frequency positions of the peaks, although individual strand folding patterns could not be assigned to specific bands. Thus, IR spectroscopy has potential diagnostic value in studying the finer structure of extended strands and β-sheets structures. Spectroscopic results also indicate that the recent assignment of amide I peaks in the 1638-1640 cm^{-1} region to helices of the 3_{10} type is consistent with crystallographic data for the serine proteases.

The results of studies performed here have further established vibrational spectroscopy, in particular Fourier transform infrared spectroscopy, as a valuable method for probing protein conformation. Previous to these studies, conformational analysis of proteins using FTIR could reproduce those of the most widely used spectroscopic tool for studying protein conformation, circular dichroism spectroscopy. These, and other studies, have demonstrated that FTIR is a significantly more sensitive probe of protein conformation and is capable of providing more detailed information. Unlike circular dichroism, infrared spectroscopy appears to be able to distinguish α-helices from 3_{10}-helices and potentially provide information concerning the fine structure of β-strands. Spectral components arising from conformational elements are observed directly with IR spectroscopy. Thus, a basis spectra approach, such as is used in analysis of circular dichroism spectra, is avoided. A major drawback of the basis spectra approach is that there is no direct method of determining the number of features which comprise the composite spectrum.

Thus, infrared spectroscopy, at present, is clearly a more informative and direct means of studying protein conformation spectroscopically.

While the studies performed here have advanced the understanding of and ability to interpret the information within conformation-sensitive amide vibrational modes in complex protein molecules, more studies are necessary to provide a complete model of spectra-structure relationships. Several amide I component bands remain unassigned. These bands represent potential conformational information which can be extracted from spectroscopic data. Further, it is not well understood how tertiary structural interactions affect the spectra or if tertiary structural information can be extracted from examination of the absorptions of conformation-sensitive modes. As was demonstrated in both chapters 4 and 5, amide I component bands near 1655 cm^{-1}, previously assumed to arise solely from α-helices, can arise from structures other than α-helices. While this development is important in interpretation of this spectral feature, it produces a complication not previously addressed. That is, it is possible that more than one conformational element can absorb in the same spectral region. Thus, it appears that the amide I mode alone cannot provide a complete description of secondary structure. It may be necessary to include information from other vibrational modes or other methods, such as Raman spectroscopy or vibrational circular dichroism, to provide a complete representation of the vibrational nature of complex protein molecules.

The study of structure-function relationships in biological macromolecules, particularly proteins, requires a firm understanding of the molecular structure and conformation as well as the physicochemical properties exhibited by such structures. Previous methods for determination of protein conformation from

spectroscopic information have utilized limited structural models. This thesis has explored the development of a comprehensive method for classification of protein conformation and use of the results of this method as a basis for interpretation of information arising from conformation-sensitive vibrational modes in infrared spectra of proteins. The result has been a more complete understanding of the spectroscopic information derived from the conformation-sensitive vibrational modes. Additionally, it has been demonstrated that certain approaches to studying spectra-structure relationships, such as those employed here, are advantageous. In conclusion, integration of methods for discerning spectra-structure correlations with a comprehensive basis for description and classification of protein conformation provides a powerful approach to understanding relationships between a proteins spectral properties and its molecular conformation.

Bibliography

Adler, A. J., Greenfield, N. J. and Fasman, G. D.: Circular Dichroism and Optical Rotatory Dispersion of Proteins and Polypeptides: Meth. Enzymol. 27: 675-735, (1973)

Ambrose, E. J., Bamford, C. H, Elliot, A. and Hanby, W. E.: Water Soluble Silk: An α-Protein: Nature 167: 264-265, (1951).

Ambrose, E. J. and Elliot, A.: Infrared Spectroscopic Studies of Globular Protein Structure: Proc. Roy. Soc. A208: 75-90, (1951).

Ambrose, E. J. and Elliot, A.: Infrared Spectra and the Structure of Fibrous Proteins: Proc. Roy. Soc. A206: 206-219, (1951).

Ambrose, E. J. and Elliot, A.: The Structure of Synthetic Polypeptides. II. Investigations with Polarized Infrared Spectroscopy: Proc. Roy. Soc. A205: 47-60, (1951).

Anderle, G., and Mendelsohn, R.: Thermal Denaturation of Globular Proteins: Fourier Transform Infrared Studies of the Amide III Spectral Region: Biophys. J. 52: 69-74, (1987).

Arrondo, J. L. R., Young, N. M., and Mantsch, H. H.: The Solution Structure of Concanavalin A Probed by FT-IR Spectroscopy: Biochim. Biophys. Acta 952: 261-268 (1988).

Bamford, C. H., Brown. L, Elliot, A., Hanby, W. E. and Trotter, I. F.: Some New Investigations on the Structure of Synthetic Polypeptides: Proc. Roy. Soc. B141: 49-59, (1953).

Bandekar, J., Evans, D. J., Krimm, S., Leach, S. J., Lee, S., McQuie, J. R., Minasian, E., Nemethy, G., Pottle, M. S., Scheraga, H. A., Stimson, E. R. and Woody, R. W.: Conformations of cyclo(-L-Alanyl-L-Alanyl-ε-Aminocaproyl) and cyclo(-L-Alanyl-D-Alanyl-ε-Aminocaproyl): Cyclized Models for Specific Types of β-Bends: Int.J.Peptide.Protein.Res. 19: 187-205, (1982).

Bandekar, J. and Krimm, S.: Vibrational Analysis of Peptides, Polypeptides and Proteins: VII. Normal Modes and Vibrational Spectra of a Type I β-Turn Tetrapeptide, in "Peptides: Structure and Function. Proceedings of the 6th American Peptide Symposium": 241-245, (1979)

Bandekar, J. and Krimm, S.: Vibrational Analysis of Peptides, Polypeptides and Proteins: Characteristic Amide Bands of β-Turns: Proc. Natl. Acad. Soc. USA 76: 774-777, (1979).

Bandekar, J. and Krimm, S.: Vibrational Analysis of Peptides, Polypeptides and Proteins: XXX. Normal Mode Analysis of γ-Turns: Int. J. Peptide. Protein. Res. 26: 407-415, (1985).

Barlow, D. J., and Thornton, J. M. : Helix Geometry in Proteins: J. Mol. Biol. 201: 601-619, (1988).

Beer, M., Sutherland, G. B. B. M., Tanner, K. N. and Wood, D. L.: Infrared Spectra and the Structure of Proteins: Proc. Roy. Soc. A249: 147-172, (1959).

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F.,Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M.: The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures: J. Mol. Biol. 112: 535-542, (1977).

Bradbury, E. M. and Elliot, A.: Infrared Spectra and Chain Arrangement in Some Polyamides, Polypeptides and Fibrous Proteins: Polymer 4: 47-59, (1963).

Brahms, S. and Brahms, J.: Determination of Protein Secondary Structure by Vacuum Ultraviolet Circular Dichroism: J.Mol.Biol. 138: 149-178, (1980).

Blevins, R. A., and Tulinsky, A. : The Refinement and the Structure of the Dimer of α-Chymotrypsin at 1.67 Å Resolution: J. Biol. Chem. 260: 4264 (1985).

Blundell, T. L. and Johnson, L. N.: "Protein Crystallography": Academic Press, London, (1976).

Butler, W.L.: Fourth Derivative Spectra: Meth. Enzymol. 56; 501-515 (1979)

Byler, D. M., and Purcell, J. M.: Examination of Slow, Reversible Variations in Protein Secondary Structure by FTIR Spectroscopy: Spectrosc. Biol. Mol., Proc. Eur. Conf. 3rd 3, 21-24.(1989a).

Byler, D. M., and Purcell, J. M.: FTIR Examination of Thermal Denaturation and Gel Formation in Whey Proteins: Proc. SPIE Int. Soc. Opt. Eng. (Fourier Transform Spectroscopy) 1145: 415-417, (1989b).

Byler, D. M. and Susi, H.: Examination of the Secondary Structure of Proteins by Deconvolved FT-IR Spectra: Biopolymers 25: 469-487, (1986).

Cannon, C. G.: Infrared Frequency Shifts and Amide Group Interactions: J. Chem. Phys. 24: 491-492, (1956).

Casal, H. L., Kohler, U., and Mantsch, H. H.: Structural and Conformational changes of β-Lactoglobulin B: An Infrared Spectroscopic Study of the Effect of pH and Temperature: Biochim. Biophys. Acta 957: 11-20, (1988).

Chambers, J. L., and Stroud, R. M.: Difference Fourier Refinement of the Structure of DIP-Trypsin at 1.5 Å with a Minicomputer Technique: Acta Crystallogr. B. 33: 1824-1837,.(1977).

Cheam, T. S. and Krimm, S.: Infrared Intensities of Amide Modes in N-Methylacetamide and Poly (Glycine I) from ab initio Calculations of Dipole Moment Derivatives of N-Methylacetamide: J. Chem. Phys. 82: 1631-1641, (1985)

Chirgadze, Y. N. and Brazhnikov, E. V.: Intensity of the Infrared Amide I Band of Dipeptides in Heavy Water: Biopolymers 12: 2185-2188, (1973).

Chirgadze, Y. N. and Brazhnikov, E. V.: Intensities and Other Spectral Parameters of Infrared Amide Bands of Polypeptides in the α-Helical Form: Biopolymers 13: 1701-1712, (1974).

Chirgadze, Y. N., Federov, O. V. and Trushina, N. P.: Estimation of Amino Acid Residue Side-chain Absorption on the Infrared Spectra of Proteins in Heavy Water: Biopolymers 14: 679-694, (1975).

Chirgadze, Y. N. and Rashevskaya, E. P.: Intensity of the Characteristic Vibrations of the Peptide Group in the Infrared Spectrum of Poly-γ-Benzyl-Glutamate in the Helical Conformation: Biofizika (USSR) 14: 608-614, (1969).

Chirgadze, Y. N., Shestopolov, B. V. and Venyaminov, S. Y.: Intensities and Other Spectral Parameters of Infrared Amide Bands of Polypeptides in the β- and Random Forms: Biopolymers 12: 1337-1351, (1973).

Chou, P. Y., and Fasman, G. D.: β-Turns in Proteins: J. Mol. Biol. 115: 135-175, (1977).

Covington, A. K., Paabo, M., Robinson, R. A., and Bates,R. G.: Use of the Glass Electrode in Deuterium Oxide and the Relation between the Standardized pD (p_aD) Scale and the Operational pH in Heavy Water: Anal. Chem. 40: 700-706,.(1968).

Cox, J. M.: A Method for Structural Superpositioning of Proteins: J. Mol. Biol. 28: 151-156, (1967).

Crawford, J. L., Lipscomb, W. N. and Schellman, C. G.: The Reverse Turn as a Polypeptide Conformation in Globular Proteins: Proc. Natl. Acad. Sci. U.S.A. 70: 538-542, (1973).

Cunningham, L. W., Jr.: Molecular-Kinetic Properties of Diisopropyl Phosphoryl Trypsin: J. Biol. Chem. 13: 211-218, (1954).

Dwivedi, A. M., Krimm, S. and Malcolm, B. R.: Vibrational Analysis of Peptides, Polypeptides and Proteins: XXIV. Conformation of Poly(α-Aminoisobutyric Acid): Biopolymers 23: 2025-2065, (1984).

Elliot, A. and Ambrose, E. J.: Structure of Synthetic Polypeptides: Nature 165: 921-922, (1950).

Elliot, A.: The Infrared Spectra of Some Optically Active and meso-Synthetic Polypeptides: Proc. Roy. Soc. A221: 104-114, (1953).

Elliot, A.: The Infrared Spectra of Polypeptides with Small Side Chains: Proc. Roy. Soc. A226: 408-421, (1954).

Elliot, A., Ambrose, E. J. and Robinson, C.: Chain Configurations in Natured and Denatured Insulin: Evidence from Infrared Spectra: Nature 166: 194, (1950a).

Elliot, A. and Ambrose, E. .: Evidence of Chain Folding in Polypeptides and Proteins: Disc. Faraday Soc.,(Cambridge) 9: 246-251, (1950b).

Elliot, A. and Malcolm, B. R.: Structure and Properties of Synthetic Polypeptides and Silk Proteins: Trans. Faraday. Soc. 52, 528-536, (1956).

Englander, S. W., Downer, N. W., and Teitelbaum, H.: Hydrogen Exvhange: Ann. Rev. Biochem. 41: 903-924, (1972).

Erlanger, B. F., Kokowsky, N., and Cohen, W.: The Preparation and Properties of Two New Chromogenic Substrates of Trypsin: Arch. Biochem. Biophys. 95: 271-278 (1961).

Fehlhammer, H., Bode, W., and Huber, H. :Crystal Structure of Bovine Trypsinognen at 1.8 Å Resolution II. Crystallographic Refinement, Refined Crystal Structure and Comparison with Bovine Trypsin: J. Mol. Biol. 111: 415-438, (1977).

Fetrow, J. S., Zehfus, M. H. and Rose, G. D.: Protein Folding: New Twists: Biotechnology 6: 167-171, (1988).

Greer, J. : Comparative Model-building of the Serine Proteases; J. Mol. Biol. 153: 1027-1042, (1981).

Griffiths, P. R and DeHaseth, J. A.: "Fourier Transform Infrared Spectroscopy": Wiley, New York, (1986).

Griffiths, P. R., Pierce, J. A., and Hongjin, G.: "Curve Fitting and Fourier Self-Deconvolution for the Quantitative Representation of Complex Spectra" in Computer-Enhanced Spectroscopy (Meazelear, H.L.C., and Elsenhour, T.L., eds.) pp 29-54, Plenum Press, New York.(1987).

Halloway, P. W., and Mantsch, H. H.: Structure of Cytochrome b_5 in Solution by Fourier Transform Infrared Spectroscopy: Biochemistry 28: 931-935 (1989).

Haris, P. I., Coke, M., and Chapman, D.: Fourier Transform Infrared Spectroscopic Investigation of Rhodopsin Structure and its Comparison with Bacteriorhodopsin: Biochim. Biophys. Acta 995:160-167.(1989).

Haris, P. I., Lee, D. C. and Chapman, D.: A Fourier Transform Infrared Investigation of the Structural Differences Between Ribonuclease A and Ribonuclease S: Biochim. Biophys. Acta 874: 255-265, (1986).

Hohne, E., and Kretschmer, R. G.: Description of Secondary Structures in Proteins: Stud. Biophys. 108: 165-186 (1985).

Huber, R., and Bode, W. : Structural Basis of the Activation and Action of Trypsin: Acc. Chem. Res. 11: 114-122, (1978).

Jackson, M., Haris, P. I. and Chapman, D.: Conformational Transitions in Poly(l-lysine): Studies Using Fourier Transform Infrared Spectroscopy: Biochim. Biophys. Acta 998: 75-79.(1989)

Johnson, W. C.: Secondary Structure of Proteins through Circular Dichroism Spectroscopy: Annu. Rev. Biophys. Biophys. Chem. 17: 145-167, (1988).

Kabsch, W. and Sander, C.: A Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features: Biopolymers 22: 2577-2637, (1983).

Kaiden, K., Matsui, T. and Tanaka, S.: A Study of the Amide III Band by FT-IR Spectrometry of the Secondary Structure of Albumin, Myoglobin, and γ-Globulin: Applied Spectroscopy 41: 180-184, (1987).

Kauppinen, J. K., Moffat, D. J., Mantsch, H. H. and Cameron, D. G.: Fourier Self-Deconvolution: A Method for Resolving Intrisically Overlapped Bands: Applied Spectroscopy 35: 271-276, (1981).

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R.G. and Wyckoff, H.: A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-ray Analysis: Nature 181: 662-666, (1958).

Kolaskar, A. S., Ramabraham, V., and Soman, K. V.: Reversals of the Polypeptide Chain in Globular Proteins: Int. J. Pept. Protein Res. 16: 1-11, (1980).

Krimm, S.: Frequency Shift and the C=O Stretching Band in Polypeptides and Proteins: J. Chem. Phys. 23: 1371-1372, (1955).

Krimm, S.: Infrared Spectra and Chain Conformation in Proteins: J. Mol. Biol. 4: 528-540, (1962).

Krimm, S. and Abe, Y.: Intermolecular Interaction Effects in the Amide I Vibrations of β-Polypeptides: Proc. Natl. Acad. Sci., USA 69: 2788-2792, (1972).

Krimm, S. and Bandekar, J.: Vibrational Analysis of Peptides, Polypeptides and Proteins: Adv. Prot. Chem. 38: 181-364, (1986).

Kuntz, I. D.: Protein Folding: J. Am. Chem. Soc. 94: 4009-4012, (1972).

Lee, D. C. and Chapman, D.: Infrared Spectroscopic Studies of Biomembranes and Model Membranes: Bioscience Reports 6: 235-255,.(1986)

Leszczynski, J. F. and Rose, G. D.: Loops in Globular Proteins: A Novel Category of Secondary Structure: Science 234: 849-855, (1986).

Levitt, M., and Chothia, C.: Structural Patterns in Globular Proteins: Nature 261:552-558, (1976).

Levitt, M. and Greer, J.: Automatic Identification of Secondary Structure in Proteins: J. Mol. Biol. 114: 181-293, (1977).

Lewis, P. N., Momany, F. A., and Scheraga, H. A.: Folding of Polypeptide Chains in Proteins: A Proposed Mechanism for Folding: Proc. Natl. Acad. Sci. USA 68: 2293-2297, (1971).

Liebman, M. N.: Structural Organization in the Serine Proteases. I. Macromolecular Specificity in Limited Proteolysis: Enzyme 36: 115-140, (1986).

Liebman, M. N., Venanzi, C. A. and Weinstein, H.: Structural Analysis of Carboxypeptidase A and its Complexes With Inhibitors as a Basis for Modelling Enzyme Specificity: Biopolymers 24: 1721-1758, (1985).

Maddams, W. F. and Tooke, P. B.: Quantitative Conformational Studies on Poly(Vinylchloride): J. Macromol. Sci.-Chem. A17: 951-968, (1982).

Mantsch, H. H., Casal, H. L., and Jones, R. N.; "Resolution Enhancement of Infrared Spectra of Biological Systems" in Spectroscopy of Biological Systems (Clark, R.J.H. and Hester, R.E., eds.) Wiley & Sons, New York (1986).

Mantsch, H. H., Surewicz, W. K., Muga, A., Moffatt, D. J. and Casal, H. L.: Proc. SPIE Int. Soc. Opt. Eng. (Fourier Transform Spectroscopy) 1145: 580-581, (1989).

Markely,J. L. and Urich,E. L.: Detailed Analysis of Protein Structure and Function by NMR Spectroscopy: Survey of Resonance Assignments. Ann. Rev. Biophys. Bioeng. 13: 493-521, (1984).

Marquart, M., Walter, J., Diesenhofer, J., Bode, W., and Huber, R.: The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and Its Complexes with Inhibitors: Acta Crystallogr. B. 39, 480-490, (1983)

Maxfield, F. R., Bandekar, J., Krimm, S., Evans, D. J., Leach, S. J., Nemethy, G. and Scheraga, H. A.: Conformation of cyclo(-L-Alanylglycyl-ε-Aminocaproyl), A Cyclized Dipeptide Model for a β-Bend. 3. Infrared and Raman Spectroscopic Studies: Macromolecules 14: 997-1003, (1981).

Miyazawa, T.: Perturbation Treatment of Characteristic Vibrations of Polypeptide Chains in Various Configurations: J. Chem. Phys. 32: 1647-1652, (1960).

Miyazawa, T.: The Characteristic Band of Secondary Amides at 3100 cm^{-1}: J. Mol. Spec. 4: 168-172, (1960).

Miyazawa, T.: Internal Rotation and Low Frequency Spectra of Esters, Monosubstituted Amides and Polyglycine: Bull. Chem. Soc. Japan 34: 691-696, (1961).

Miyazawa, T.: "Characteristic Amide Bands and Conformations of Polypeptides", in "Polyamino Acids, Polypeptides and Proteins: Proceedings of an International Symposium Held at the University of Wisconsin, 1961", University of Wisconsin Press, Madison, (1962).

Miyazawa, T. and Blout, E. R.: The Infrared Spectra of Polypeptides in Various Conformations: Amide I and Amide II Bands: J. Am. Chem. Soc. 83: 712-719, (1961).

Miyazawa, T., Shimanouchi, T. and Mizushima, S.I.: Characteristic Infrared Bands of Monosubstituted Amides: J. Chem. Phys. 24: 408-418, (1956).

Miyazawa, T., Shimanouchi, T. and Mizushima, S. I.: Normal Vibrations of N-methylacetamide: J. Chem. Phys. 29: 611-616, (1958).

Moore, W. H. and Krimm, S.: Transition Dipole Coupling in Amide I Modes of β-Polypeptides: Proc. Natl. Acad. Sci., USA 72: 4933-4935, (1975).

Naik, V. M. and Krimm, S.: Vibrational Analysis of Peptides, Polypeptides and Proteins: XVII. Normal Modes of Pro-Leu-Glu-NH2, a Type II β-Turn: Int. J. Peptide. Protein. Res. 32: 1-24, (1984).

Naik, V. M. and Krimm, S.: Vibrational Analysis of the Structure of Gramicidin A: I. Normal Mode Analysis: Biophys. J. 49: 1131-1145, (1986a).

Naik, V. M. and Krimm, S.: Vibrational Analysis of the Structure of Gramicidin A: II. Vibrational Analysis: Biophys. J. 49: 1147-1154, (1986b).

Olinger, J. M., Hill, D. M., Jakobsen, R. J. and Brody, R. S.: Fourier Transform Infrared Studies of Ribonuclease in H_2O and ^2H_2O Solutions: Biochim. Biophys. Acta. 869: 89-98, (1986).

Parker, F.: Biochemical Applications of Fourier Transform Infrared Spectroscopy: Canadian J. Spectroscopy 31: 1-6, (1986).

Pauling, L. and Corey, R. B.: Configurations of Polypeptide Chains with Favored Orientations Around Single Bonds: Two New Pleated Sheets: Proc. Natl. Acad. Sci. U.S.A. 37: 729-740, (1951).

Pauling, L., Corey, R. B. and Branson, H. R.: The Structure of Proteins: Two Hydrogen Bonded Helical Configurations of the Polypeptide Chain: Proc. Natl. Acad. Sci. U.S.A. 37: 205-211, (1951).

Payne, K. J. and Veis, A.: Fourier Transform IR Spectroscopy of Collagen and Gelatin Solutions: Deconvolution of the Amide I Band for Conformational Studies: Biopolymers 27: 1749-1760 (1988).

Pierce, J. A., Jackson, R. S., Van Every, K. W., Griffiths, P. R., and Hongjin, G.: Combined Deconvolution and Curve Fitting for Quantitative Analysis of Unresolved Specrtral Bands: Anal. Chem. 62: 477-484 (1990).

Purcell, J. M. and Susi, H.: Solvent Denaturation of Proteins as Observed by Fourier Transform Infrared Spectroscopy: J. Biochem. Biophys. Meth. 9: 193-199, (1984).

Ramakrishnan, C. and Soman, K. V.; Identification of Secondary Structure in Globular Proteins - A New Algorithm: Int. J. Peptide Protein Res. 20: 218-237 (1982).

Richards, F. M. and Kundrot, C. E.: Identification of Structural Motifs From Protein Coordinate Data: Secondary Structure and First Level Supersecondary Structure: Proteins 3: 71-84, (1988).

Richardson, J. S.: The Anatomy and Taxonomy of Protein Structure: Adv. Prot. Chem. 34: 167-399, (1981).

Rose, G. D., Gierasch, L. M. and Smith, J. A.: Turns in Peptides and Proteins: Adv. Prot. Chem. 37: 1-109, (1985).

Ruegg, M., Metzger, V. and Susi, H.: Computer Analysis of Characteristic Infrared Bands of Globular Proteins: Biopolymers 14: 1465-1471, (1975).

Sawyer, L., Shotton, D. M., Campbell, J. W., Wendell, P. L., Muirhead, H., Watson, H. C., Diamond, R., and Ladner, R. C. : The Atomic Structure of Porcine Pancreatic Elastase at 2.5 Å Resolution: Comparisons with the Structure of α-Chymotrypsin: J. Mol. Biol. 118: 137-208, (1978).

Schroeder, D. D., and Shaw, E.: Chromatography of Trypsin and Its Derivatives: Characterization of a New Active Form of Bovine Trypsin: J. Biol. Chem. 243: 2943-2949, (1968).

Smith, J. A. and Pease, L. G.: Reverse Turns in Peptides and Proteins: Crit. Rev. Biochem. 8: 315-399, (1980).

Smith, R. L., and Shaw, E.: Psuedotrypsin: A Modified Bovine Trypsin Produced by Limited Autodigestion: J. Biol. Chem. 244: 4707-4712, (1969).

Surewicz, W. K., and Mantsch, H. H.: New Insight into Protein Secondary Structure from Resolution-enhanced Infrared Spectra: Biochim. Biophys. Acta 952: 115-130,.(1988a).

Surewicz, W. K., and Mantsch, H. H.: Solution and Membrane Structure of Enkephalins as Studied by Infrared Spectroscopy: Biochem. Biophys. Res. Commun. 150: 245 251,.(1988b).

Surewicz, W. K., and Mantsch, H. H.: Conformational Properties of Angiotensin II in Aqueous Solution and in a Lipid Environment: A Fourier Transform Infrared Spectroscopic Investigation: J. Am. Chem. Soc. 110: 4412-4414.(1988c)

Surewicz, W. K., Moscarello, M. A., and Mantsch, H. H.: Fourier Transform Infrared Spectroscopic Investigation of the Interaction between Myelin Basic Protein and Dimyristoylphosphotidylglycerol Biylayers: Biochemistry 26, 3881-3886, (1987).

Surewicz, W. K., Stepanik, T. M., Szabo, A. G., and Mantsch, H. H.: Lipid-induced Changes in the Secondary Structure of Snake Venom Cardiotoxins: J. Biol. Chem. 263: 786-790,.(1988).

Susi, H.: "Infrared Spectra of Biological Molecules and Related Systems" in Structure and Stability of Biological Macromolecules (Timasheff, S. N., and Fasman, G. eds,) pp 575-663, Marcel Dekker, New York, (1969).

Susi, H.: Infrared Spectroscopy- Conformation: Meth. Enzymol. 26: 455-472, (1972).

Susi, H. and Byler, D. M.: Protein Structure by Fourier Transform Infrared Spectroscopy: Second Derivative Spectra: Biochem. Biophys. Res. Comm. 115: 391-397, (1983).

Susi, H. and Byler, D. M.: Resolution Enhanced Fourier-transform Infrared Spectroscopy of Enzymes: Meth. Enzymol. 130: 290-311, (1986).

Susi, H., Byler, D. M. and Purcell, J. M.: Estimation of β-Structure Content by Means of Deconvolved FT-IR Spectra: J. Biochem. Biophys. Meth. 11: 235-240, (1985).

Susi, H., Timasheff, S. N. and Stevens, L.: Infrared Spectra and Protein Conformations in Aqueous Solutions I. The Amide I Band in H_2O and D_2O Solutions: J. Biol. Chem. 242: 5460-5466, (1967).

Suzuki, S., Iwashita, Y., Shimanouchi, T. and Tsuboi, M.: Infrared Spectra of Isotopic Polyglycines: Biopolymers 4: 337-350, (1966).

Timasheff, S. N., Susi, H. and Stevens, L.: Infrared Spectra and Protein Conformations in Aqueous Solutions II. Survey of Globular Proteins: J. Biol. Chem. 242: 5467-5473, (1967).

Tinoco, I., Jr. and Williams, A. L., Jr.: Differential Absorption and Differential Scattering of Circularly Polarized Light: Applications to Biological Macromolecules. Ann. Rev. Phys. Chem. 35: 329-355, (1984).

Trewhella, J., Liddle, W. K., Heidorn, D. B., and Strynadka, N.: Calmodulin and Troponin C Structures Studied by Fourier Transform Infrared Spectroscopy: Effects of Ca^{2+} and Mg^{2+} Binding: Biochemistry 28: 1294-1301.(1989).

Venkatachalam, C. M.: Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of Three-Linked Peptide Units: Biopolymers 6: 1425-1436, (1968).

Walter, J., Steigmann, W., Singh, T. P., Bartunik, H., Bode, W., and Huber, H.: On the Disordered Activation Domain in Trypsinogen: Chemical Labelling and Low-Temperature Crystallography: Acta Cryst. B38: 1462-1472, (1982)

Wang, D., Bode, W., and Huber, R.: Bovine Chymotrypsinogen A : X-ray Crystal Structure Analysis and Refinement of a New Crystal Form at 1.8 Å Resolution: J. Mol. Biol 185: 595-624, (1985).

Wasacz, F. M., Olinger, J. M. and Jakobsen, R. J.: Fourier-Transform Infrared Studies of Proteins Using Non-Aqueous Solvents: Effects of Methanol and Ethylene Glycol on Albumin and Immunoglobulin G: Biochemistry 26: 1464-1470, (1987).

Wilson, E. B., Jr.: A Method for Obtaining the Expanded Secular Equation for the Vibration Frequencies of a Molecule: J. Chem. Phys. 7: 1047-1052, (1939).

Wilson, E. B., Jr.: Some Mathematical Methods for the Study of Molecular Vibrations: J. Chem. Phys. 9: 76-84, (1941).

Wilson, E.B.,Jr., Decius, J.C. and Cross, P.C.: "Molecular Vibrations": Dover Publications, New York, (1955).

Wong, P. T. T., Saint Girons, I., Guillou, Y., Cohen, G. N., Barzu, O., and Mantsch, H. H.: Pressure-induced Changes in the Secondary Structure of the *Escherichia Coli* Methionine Repressor Protein: Biochim. Biophys. Acta 996: 260-262, (1989).

Yang, W. J., Griffiths, D. M., Byler, D. M. and Susi, H.: Protein Conformation by Infrared Spectroscopy: Resolution-enhancement by Fourier Self-deconvolution: Applied Spectroscopy 39: 282-287, (1985).

Zimmerman, S. S. and Scheraga, H. A.: Local Interactions in Bends in Proteins: Proc. Natl. Acad. Sci. U.S.A. 74: 4126, (1977).