

# Talanta

---

Talanta 43 (1996) 199–219

Quantitation of the global secondary structure of globular proteins by FTIR spectroscopy: comparison with X-ray crystallographic structure<sup>1</sup>

Thomas F. Kumosinski\*, Joseph J. Unruh

*U.S. Department of Agriculture, ARS, Eastern Region Research Center, 600 East Mermaid Lane, Philadelphia, PA 19118, USA*

# Quantitation of the global secondary structure of globular proteins by FTIR spectroscopy: comparison with X-ray crystallographic structure<sup>1</sup>

Thomas F. Kumosinski\*, Joseph J. Unruh

*U.S. Department of Agriculture, ARS, Eastern Region Research Center, 600 East Mermaid Lane, Philadelphia, PA 19118, USA*

Received 3 March 1995; revised 21 August 1995; accepted 24 August 1995

---

## Abstract

Fourier transform infrared spectroscopy (FTIR) is potentially a powerful tool for determining the global secondary structure of proteins in solution, providing the spectra are analyzed using a statistically and theoretically justified methodology. We have performed FTIR experiments on 14 globular proteins and two synthetic polypeptides whose X-ray crystal structures are known to exhibit varying types and amounts of secondary structures. Calculation of the component structural elements of the vibrational bands was accomplished using nonlinear regression analysis, by fitting both the amide I and amide II bands of the Fourier self-deconvoluted spectra, the second-derivative spectra, and the original spectra.

The methodology was theoretically justified by comparing (via nonlinear regression analysis) the global secondary structure determined after deconvolving into component bands the vibrational amide I envelopes with the calculated structure determined by first principles from Ramachandran analysis of the X-ray crystallographic structure of 14 proteins from the Brookhaven protein data bank. Justification of the nonlinear regression analysis model with respect to experimental and instrumental considerations was achieved by the decomposition of all the bands of benzene and an aqueous solution of ammonium acetate into component bands while floating the Gaussian/Lorentzian character of the line shapes. The results for benzene yield all pure Lorentzian line shapes with no Gaussian character while the ammonium acetate spectra yielded all Gaussian line shapes with no Lorentzian character. In addition, all-protein spectra yielded pure Gaussian line shapes with no Lorentzian character. Finally, the model was statistically justified by recognizing random deviation patterns in the regression analysis from all fits and by the extra sum of squares *F*-test which uses the degrees of freedom and the root mean square values as a tool to determine the optimum number of component bands required for the nonlinear regression analysis.

Results from this study demonstrate that the globular secondary structure calculated from the amide I envelope for these 14 proteins from FTIR is in excellent agreement with the values calculated from the X-ray crystallographic data using three-dimensional Ramachandran analysis, providing that the proper contribution from GLN and ASN side chains to the 1667 and 1650 cm<sup>-1</sup> component bands has been taken into account. The standard deviation of the

---

\* Corresponding author.

<sup>1</sup> Reference to a brand or firm name does not constitute endorsement by the US Department of Agriculture over others of a similar nature not mentioned.

regression analysis for the per cent helix, extended, turn and irregular conformations was found to be 3.49%, 2.07%, 3.59% and 3.20%, respectively.

*Keywords:* Fourier transform infrared spectroscopy; Globular proteins; Protein secondary structure

## 1. Introduction

Previous studies comparing the global secondary ( $2^\circ$ ) structure of globular proteins calculated from their X-ray crystal structure with those determined from Fourier self-deconvolution (FSD) FTIR spectroscopy were performed in  $D_2O$ . However,  $D_2O$  may cause increased hydrophobic interactions which could lead to spurious  $2^\circ$  structural changes in some proteins [1]. Using  $D_2O$  also results in the elimination of the amide II peptide band, which may be helpful for validation of the amide I assignments. With the use of an extremely short path length (6–12  $\mu\text{m}$ ) and accurate water vapor subtraction [2], the determination of protein secondary structure in  $H_2O$ , using both the amide I and amide II regions, is now possible. However, controversy exists among researchers concerning the deconvolution of the FTIR amide I and II envelopes into their component bands. Not only the number of bands, but also the character of the bands (whether they are of pure Gaussian, pure Lorentzian, or a combination of Gaussian and Lorentzian character) is suspect. In addition, when using FSD to ascertain the number of bands, the magnitude of the half-width at half-height of the band and the value of the resolution enhancement factor (REF) are also open to discussion. Additionally, when experiments are performed in  $H_2O$ , the amide II band, while present, is not always deconvolved. Analysts also routinely apply conservatively low values for the REF with large half-width at half-height values of 13–18  $\text{cm}^{-1}$ , in order to not overdeconvolute the spectrum, and in the next step (nonlinear regression analysis), only the smallest number of component bands are used for the nonlinear regression analysis.

The rationale for these procedures is to avoid the possibility of distorting the experimental spectrum. However, no studies analyzing the same spectrum using nonlinear regression analysis with

varying FSD parameters have as yet been performed; thus the optimum parameter limits to use without causing distortion are unknown. Using higher REFs and narrower half-widths during FSD increase the number of component bands. We will attempt to show that properly choosing to use this increased number of bands (with equal half-widths at half-height) in the nonlinear regression fit of the experimental spectrum results in much lower root mean square (RMS) values (i.e. the square root of the average of the squares of the differences between the experimental spectra and the nonlinear regression fitted spectra).

While nonlinear regression analysis using an increased number of component bands is more difficult and more time consuming (and cannot be easily performed on older microcomputers), the correct number of peaks must be obtained to insure the correct band assignments to the secondary structure of the protein. If one band is used when two are predicted from theory, then a larger amount of disordered, helical or extended structure could be calculated because of incorrect assignments. With the methodology employed herein, the researcher can use the maximum number of component bands to fit the theoretical curve to the experimental data. This in fact yields the lowest RMS value, and the amide II envelope is fit simultaneously with the amide I envelope. It is critical to allow a zero slope baseline to vary to a calculated value. It is noted that during the calculations, the use of too many bands may result in the heights of the excess bands approaching a zero or negative value. The number of bands were controlled by use of the *F*-test as well as the agreement of calculated frequencies with previously reported experimental and theoretical assignments. Thus, the statement that excessive numbers of bands always yield better fits to the data is not valid. This statement only applies to the use of polynomial curve fitting algorithms and not to nonlinear regression analyses.

In this work we measure FTIR spectra in H<sub>2</sub>O of 14 globular proteins and two polypeptides — with varying types and amounts of 2° structures — whose X-ray crystal structures are known. A more complete theoretically, experimentally, and statistically based analysis of the spectral data in H<sub>2</sub>O using FSD, second-derivative spectra, and band curve-fitting techniques is presented. The analysis allows the individual 2° structural components to be distinguished and then compared with results in D<sub>2</sub>O [3,4] and with the global 2° structural parameters calculated from the protein X-ray crystallographic data. Previous frequency assignments of structural components are also assessed and contrasted with the new information in H<sub>2</sub>O.

## 2. Methods

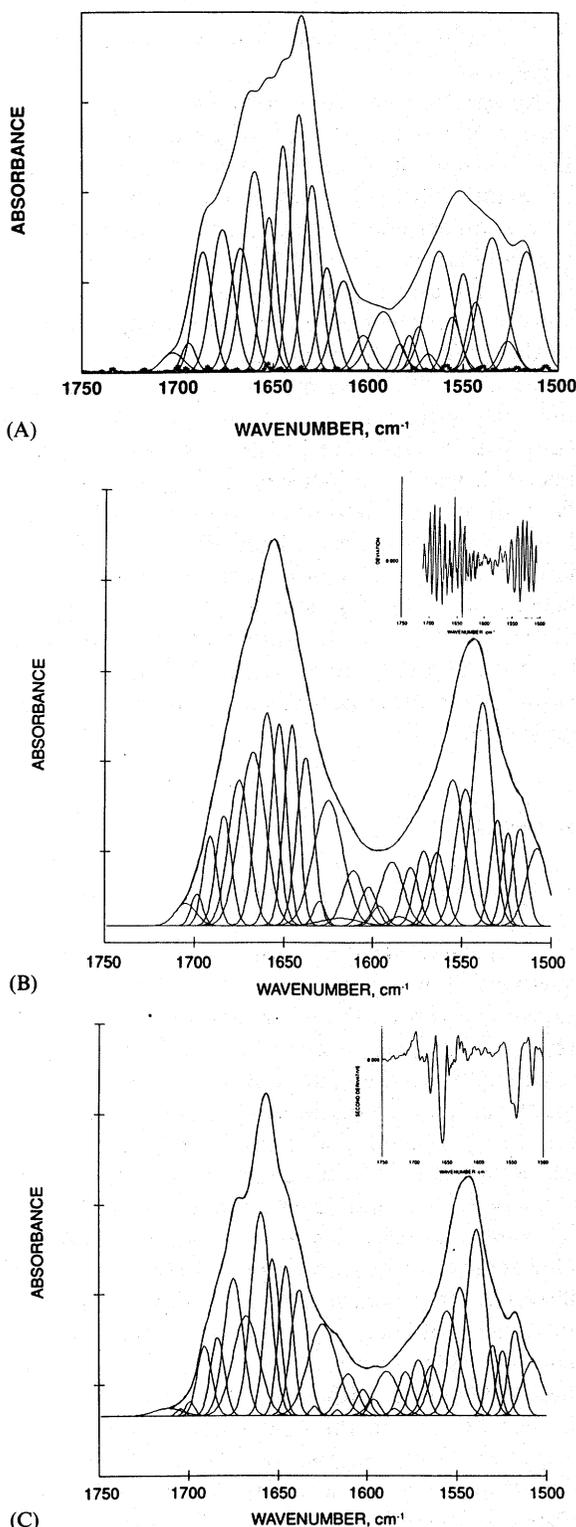
### 2.1. Infrared measurement

The individual proteins were prepared as 4 solutions (by weight) in a 20 mM imidazole buffer, pH 6.7. All protein solutions and buffers were filtered using a HVIP 0.45 μm Millipore low-protein retention filter prior to the FTIR experiments. The protein concentrations after filtration were calculated from their respective absorptivities at 280 nm. All samples were introduced into a demountable cell with CaF<sub>2</sub> windows and a 12 μm Teflon spacer. Spectra were obtained using a Nicolet 740 FTIR spectrometer equipped with the Nicolet 660 data system. Data collection was carried out following a 30 min nitrogen purge of the sample chamber. Eight data sets of 512 drift and intensity correlated interferograms were collected, coadded (net for each spectrum was 4096 double-sided interferograms with 16 384 data points), phase-corrected, apodized (Happ-Genzel function), and fast-Fourier transformed. The eight data sets were routinely checked for drift using calculated second derivatives before summing. Nominal instrument resolution was 2 cm<sup>-1</sup> with one data point every 1 cm<sup>-1</sup>. Water vapor absorption was routinely subtracted from all spectra using the second-derivative method [2].

To reduce the water vapor in the sample chamber, the FTIR spectrometer is purged with nitro-

gen which has been passed through a Balston model 75-52 FTIR purge gas generator, drying the nitrogen to a dew point of -100°F. In addition, three additional nitrogen purging lines were added to the sample compartment of the FTIR spectrometer, the net effect being a purge rate of 60 l min<sup>-1</sup>. The total amount of water vapor observed by the IR detector in the course of a typical protein solution scan is illustrated in Fig. 1A, where the water vapor spectrum is represented by crosses. For comparison purposes, the FTIR amide I and II envelopes of the FSD spectra of trypsin, the component bands obtained from the nonlinear regression fit of the amide I and II envelopes, and the resulting sum of the component bands are also illustrated (as single lines). (The use of nonlinear regression analysis for decomposition of the amide I and II envelopes into component bands will be discussed in another section of this manuscript.) As can be seen in Fig. 1A, the water vapor spectrum is significantly smaller than the actual protein amide I and II envelopes as well as the component bands derived from nonlinear regression analysis. The heights are 50–100 times smaller and the half-widths at half-height are also significantly smaller in magnitude. Thus, even if no vapor subtraction is performed on the protein FTIR spectrum, the water vapor spectrum cannot be a source of the component bands attributed to the protein. Hence, the proposition that the component bands derived from nonlinear regression analysis are useful in defining unique populations of protein vibrating molecular groups is valid.

It is important to emphasize that to obtain acceptable FTIR spectra of proteins in water, a good signal-to-noise ratio (SN) is mandatory. SN is dictated by the sensitivity, resolution and stability of the FTIR spectrometer and its computer. For our system, only a resolution of 2 cm<sup>-1</sup> is possible due to the long scanning times required for a 1 cm<sup>-1</sup> resolution spectrum, i.e. four times as long or approximately a 2 h scan. However for researchers having a quicker, more sensitive and stable instrument (such as a Nicolet 800 series spectrophotometer), a resolution of 1 cm<sup>-1</sup> should be possible to attain, although for comparison purposes, an FTIR spectrum of one protein (with corresponding buffer, background and vapor spec-



tra) was accumulated at a resolution of  $1 \text{ cm}^{-1}$ . Decomposition of this spectrum into component bands by nonlinear regression analyses yielded fractional areas for both the amide I and II envelopes in agreement to within 1% with fractional areas from the spectrum of the same protein solution collected at a resolution of  $2 \text{ cm}^{-1}$ . Also, using the  $2 \text{ cm}^{-1}$  resolution parameters, a protein solution was diluted to a concentration of 3% with buffer, and its FTIR spectrum was accumulated. Comparison of the areas of the component bands from nonlinear regression analysis agreed to within 1% with those determined using a 4% solution. These two tests add confidence to the methodology used in this study.

To ascertain whether the FTIR instrumental parameters yield line shapes that are Lorentzian, Gaussian or composites of both line shapes, FTIR spectra were obtained for benzene using KBr windows and an aqueous solution of ammonium acetate using  $\text{CaF}_2$  windows. Our regression analysis program contains line shape parameters for the amount of Gaussian character (i.e. pure Gaussian, 1; pure Lorentzian, 0), the line shape is varied along with the frequency position, height, and half-width at half-height variables. The results conclusively show that all component bands

Fig. 1. (A) FSD FTIR spectra of the amide I and II envelopes of trypsin in  $\text{H}_2\text{O}$ , represented by the solid line; component bands from nonlinear regression analysis, and the composite sum of the bands are also represented as solid lines (note: the composite sum and the FSD overlap). The crosses denote the atmospheric water vapor FTIR spectrum experienced during a typical protein accumulation if no second-derivative water vapor subtraction is performed. The crosses also represent all instrumental and phase errors (see Section 2 of text). (B) FTIR spectrum showing the amide I and amide II envelopes of lysozyme in aqueous solution. The outer envelope double line represents the original spectrum. The single line on the outer envelope and the individual component bands underneath are the results of nonlinear regression analysis, as described in the text. The inset shows a plot of connected residuals or deviations between calculated and experimental absorbances vs. frequency. (C) Fourier self-deconvolution of the FTIR spectrum of lysozyme in Fig. 1. The single line on the outer envelope and the individual component bands underneath were found by nonlinear regression analysis as described in the text. The double line represents connected experimental data. The inset shows the unsmoothed second derivative of the original spectrum.

of benzene have pure Lorentzian line shapes — the theoretical line shape for most pure condensed phase samples and gases [5]. While the aqueous ammonium acetate solution spectra yield pure Gaussian line shapes — theoretically predicted by Beer's law for the absorbance of a solid dissolved in a liquid [5]. In addition, all residual plots had a pseudorandom character, with a magnitude equivalent to an instrument residual plot.

## 2.2. Data analysis

Protein spectra (amide I and II regions) were obtained by subtracting the buffer spectra from the respective protein solution spectra in the 2000–1350  $\text{cm}^{-1}$  region. Subtractions were performed interactively using the subtraction function in the Sx software of the Nicolet 660 data system. The scaling factor (FCR) was individualized during each subtraction by adjusting the FCR parameter value until the region from 2000 to 1800  $\text{cm}^{-1}$  was as flat as possible. The subtracted region from 1800 to 1350  $\text{cm}^{-1}$  was then saved as the protein spectrum. This method for subtraction does not take into account the spectral modifications due to water–protein interactions. Previous NMR studies by Kakalis and Kumosinski [6] have shown that the water–protein interaction is small on a molar basis (water to protein), and therefore should not interfere with this type of subtraction. The protein spectra thus obtained were then used to calculate the second-derivative spectra by a simple analytical procedure that employs every data point [3]. Second-derivative spectra served as sensitive indicators for the identification of individual peak positions, and the initial number of bands to be used in subsequent nonlinear regression analysis. The unresolved spectra were subjected to FSD using an algorithm developed from that described by Kauppinen et al. [7].

FSD was undertaken with a number of resolution enhancement factors. Qualitatively, under-FSD was judged by the absence of peak positions indicated in the spectra and over-FSD by the appearance of side lobes in the flat portions of the spectra [4]. Over-FSD also resulted in excessively low baselines computed by the regression routine.

The methodology used will be illustrated for lysozyme.

All spectra were deconvolved (decomposed into their component structural elements) using a Gauss–Newton nonlinear iterative curve-fitting program ABACUS developed at this laboratory [8], which can assume Gaussian, Lorentzian shapes or a combination of both for the band envelope shape of the resolved component bands. The deconvolving curve-fitting program was applied only to the amide I and II envelopes, which consisted of at least 200 experimental points. In practice, the four parameters of each component band (Lorentzian character, height, peak frequency, and half-width at half-height) were allowed to float during the iterations, as was the baseline. Integrated areas were calculated for those peaks that correspond to conformational elements, e.g. helices, sheets, turns, and loops [9]. This procedure yielded the relative areas of the component bands, which serve to estimate the fraction of the various 2° structural elements in the protein molecule.

## 3. Results

### 3.1. Sample calculation: analysis of lysozyme

A typical FTIR spectrum of hen's egg white lysozyme showing the amide I and amide II regions is shown as the outer envelope in Fig. 1B. This spectrum can be considered to be the sum of a variety of individual bands arising from the specific structural components of the protein, such as  $\alpha$ -helix,  $\beta$ -sheets and turns. Identification of all the components of the spectrum by fitting it directly with an undefined number of peaks by nonlinear regression would be a daunting task. To alleviate this dilemma, we first examine the second derivative of the spectrum (inset, Fig. 1C) to find the number of component bands and the approximate positions of those bands, and to compare these assignments with those of Byler and Susi [3].

The next step in the analysis is to enhance the resolution of the original spectrum via an FSD algorithm. Care must be taken to choose the proper half-width and REF values used in the

of benzene have pure Lorentzian line shapes — the theoretical line shape for most pure condensed phase samples and gases [5]. While the aqueous ammonium acetate solution spectra yield pure Gaussian line shapes — theoretically predicted by Beer's law for the absorbance of a solid dissolved in a liquid [5]. In addition, all residual plots had a pseudorandom character, with a magnitude equivalent to an instrument residual plot.

## 2.2. Data analysis

Protein spectra (amide I and II regions) were obtained by subtracting the buffer spectra from the respective protein solution spectra in the 2000–1350  $\text{cm}^{-1}$  region. Subtractions were performed interactively using the subtraction function in the Sx software of the Nicolet 660 data system. The scaling factor (FCR) was individualized during each subtraction by adjusting the FCR parameter value until the region from 2000 to 1800  $\text{cm}^{-1}$  was as flat as possible. The subtracted region from 1800 to 1350  $\text{cm}^{-1}$  was then saved as the protein spectrum. This method for subtraction does not take into account the spectral modifications due to water–protein interactions. Previous NMR studies by Kakalis and Kumosinski [6] have shown that the water–protein interaction is small on a molar basis (water to protein), and therefore should not interfere with this type of subtraction. The protein spectra thus obtained were then used to calculate the second-derivative spectra by a simple analytical procedure that employs every data point [3]. Second-derivative spectra served as sensitive indicators for the identification of individual peak positions, and the initial number of bands to be used in subsequent nonlinear regression analysis. The unresolved spectra were subjected to FSD using an algorithm developed from that described by Kauppinen et al. [7].

FSD was undertaken with a number of resolution enhancement factors. Qualitatively, under-FSD was judged by the absence of peak positions indicated in the spectra and over-FSD by the appearance of side lobes in the flat portions of the spectra [4]. Over-FSD also resulted in excessively low baselines computed by the regression routine.

The methodology used will be illustrated for lysozyme.

All spectra were deconvolved (decomposed into their component structural elements) using a Gauss–Newton nonlinear iterative curve-fitting program ABACUS developed at this laboratory [8], which can assume Gaussian, Lorentzian shapes or a combination of both for the band envelope shape of the resolved component bands. The deconvolving curve-fitting program was applied only to the amide I and II envelopes, which consisted of at least 200 experimental points. In practice, the four parameters of each component band (Lorentzian character, height, peak frequency, and half-width at half-height) were allowed to float during the iterations, as was the baseline. Integrated areas were calculated for those peaks that correspond to conformational elements, e.g. helices, sheets, turns, and loops [9]. This procedure yielded the relative areas of the component bands, which serve to estimate the fraction of the various 2° structural elements in the protein molecule.

## 3. Results

### 3.1. Sample calculation: analysis of lysozyme

A typical FTIR spectrum of hen's egg white lysozyme showing the amide I and amide II regions is shown as the outer envelope in Fig. 1B. This spectrum can be considered to be the sum of a variety of individual bands arising from the specific structural components of the protein, such as  $\alpha$ -helix,  $\beta$ -sheets and turns. Identification of all the components of the spectrum by fitting it directly with an undefined number of peaks by nonlinear regression would be a daunting task. To alleviate this dilemma, we first examine the second derivative of the spectrum (inset, Fig. 1C) to find the number of component bands and the approximate positions of those bands, and to compare these assignments with those of Byler and Susi [3].

The next step in the analysis is to enhance the resolution of the original spectrum via an FSD algorithm. Care must be taken to choose the proper half-width and REF values used in the

algorithm so that the FTIR spectrum is not over or under FSD.

Initially, all amide I component band assignments of Byler and Susi [3] are used to fit the FSD spectra, along with comparable theoretical amide II bands and side-chain assignments from Krimm and Bandekar [10]. Initial half-width at half-height values are assigned as 2 or 2.5  $\text{cm}^{-1}$ , while the initial heights are assigned values higher than the FSD experimental data. The number of bands fit to the amide I and amide II envelopes is also controlled via the statistical *F*-test, which will be discussed later.

For faster convergence of the nonlinear regression analysis, the investigator may initially float only the heights, the half-widths at half-height, or the peak positions and the Gaussian/Lorentzian parameter in an alternating fashion during the process of the fitting program. Later, three of the parameters can be floated in an alternating fashion. And finally, all four parameters are floated simultaneously until the nonlinear regression analysis converges. The criteria for progress in the nonlinear regression analysis is a continually decreasing RMS value. If divergence takes place, the iteration factor that determines the step increase in the parameter values for the next iteration should be lowered. It is mandatory that the baseline value be allowed to vary during all of the above iterations. The imposition of a fixed baseline value adds error to the results, slows the convergence process, and in some cases causes calculations which ultimately diverge.

In addition, to alleviate the local minima problem, the whole process should be repeated starting with heights of the components that are lower in value than the FSD experimental data. The results of these two calculations should be in agreement. Failure of these two calculations from different starting points to agree for all component bands indicates local minima resulting from use of a lower number of component bands. It is noted that all calculations using nonlinear regression analysis should be performed in this manner.

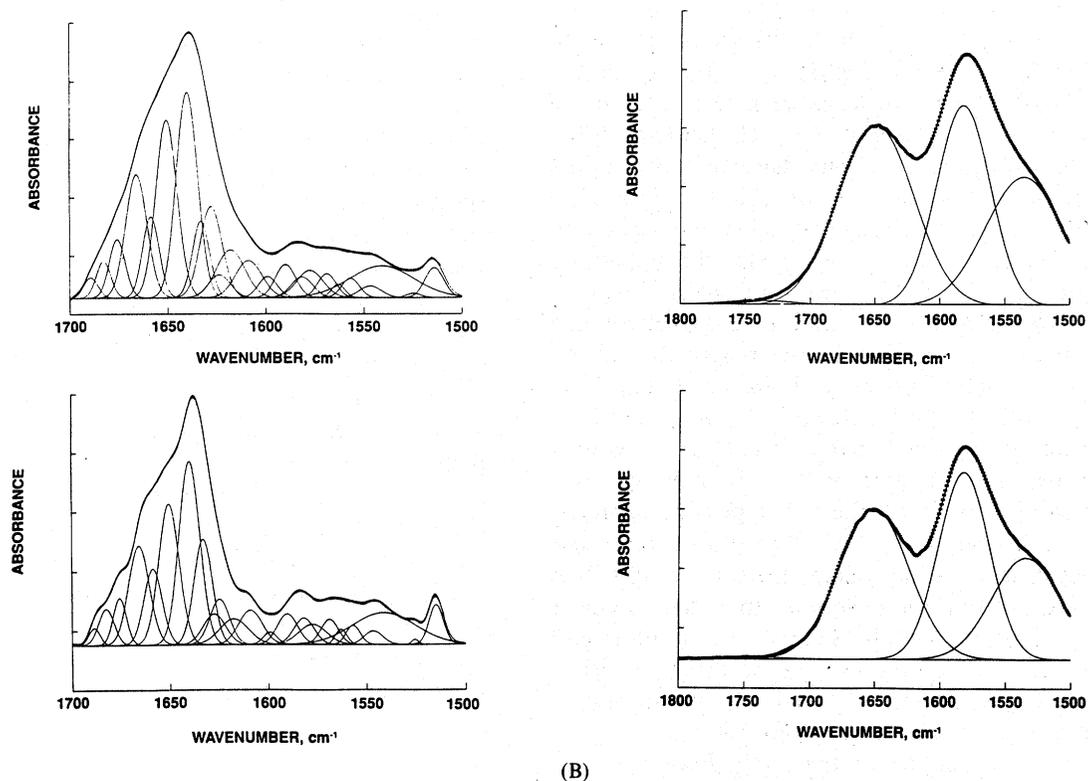
Fig. 1C shows the fit of 29 component Gaussian bands for the Fourier self-deconvoluted amide I and II FTIR envelope of lysozyme using deconvolution values of 13  $\text{cm}^{-1}$  and 2 REF. The

RMS value for this fit is 0.000125 which is well within the criteria for acceptability as stated above. This lack of Lorentzian character is in agreement with the conclusions of Byler and Susi [3] that only Gaussian component band shapes occur in FSD spectra.

The use of a high REF value results in over-FSD of the FTIR spectra, usually an unacceptable RMS value, and a baseline value far below that expected from the experimental data. Repeating the above nonlinear regression process at a variety of FSD parameters for all the proteins in our database yielded the optimum FSD parameter values of 2.3 for the REF and 9  $\text{cm}^{-1}$  for the half-width at half-height (see Fig. 4A).

After fitting the FSD spectrum of lysozyme, the next step was to use these results to fit the original spectrum shown in Fig. 1B. Here, the frequency positions from the FSD FTIR envelope of Fig. 1C were initially not allowed to iterate. Only the heights, half-widths at half-height, and the fraction of Lorentzian/Gaussian character of the band shape were allowed to alternately vary. For lysozyme as for all proteins in our database, the fraction of Lorentzian character vanished. Even when the peak positions were allowed to vary, only Gaussian band shapes were found to be present. A final convergent deconvoluted original spectrum of lysozyme is presented in Fig. 1B. A plot of the difference between the experimental and the fit absorbances vs. frequency is presented in the inset of Fig. 1B. This deviation plot is seen to be pseudorandom in shape, which further adds validity to the analysis. Furthermore, the fractional areas of the component bands of the amide I envelope in Fig. 1B were in excellent agreement with those determined from the FSD spectra in Fig. 1C. These findings (random deviation plot and agreement of the original and FSD FTIR spectral fits) for lysozyme were used as criteria of acceptability for all the proteins analyzed in the presented database.

Since the lack of Lorentzian character in the original FTIR spectrum has not been reported, we initiated further studies on other systems to aid in the validation of this conclusion. Recently, we inherited the previously reported [3] FTIR spectra of proteins in  $\text{D}_2\text{O}$ . Fig. 2A shows the FSD



(A) (B)

Fig. 2. (A) FTIR spectrum showing the amide I and II envelopes of ribonuclease in  $D_2O$  from the previously published work of Byler and Susi [3]. The outer envelope double line represents connected experimental data. The solid line on the outer envelope is the sum of component bands (solid line) calculated from nonlinear regression described in the text. The upper spectrum is the original FTIR spectrum; the lower spectrum is the Fourier self-deconvoluted spectrum. (B) FTIR spectra showing the amide I and II envelopes of polyaspartic acid under environmental conditions where the polymer adopts a disordered conformation. Same representation as in (A). The upper spectrum is the original FTIR spectrum; the lower spectrum is the Fourier self-deconvoluted spectrum.

(lower) and original (upper) FTIR spectra of ribonuclease A in  $D_2O$  [3]. In both cases a small amide II envelope due to incomplete exchange of the  $D_2O$  still exists. Using the above methodology for fitting both the amide I and II envelopes for both the FSD and the original spectra using Gaussian component bands was extremely successful, as seen in Fig. 2A. Ten amide I component bands for the FSD and original spectra with both fractional areas yielding equivalent global secondary structure result. No Lorentzian character bands were found by the nonlinear regression fitting analysis of either the FSD or the original spectra. Other FTIR spectra in  $D_2O$ , i.e. lysozyme, cytochrome C, concanavalin A, etc., yielded the same results.

Finally, FTIR studies of polyaspartic acid and polylysine at pH 7 in 0.08 M KCl and  $H_2O$  were initiated to further test the above results. At pH 7 in 0.08 M KCl, both polymers exist in a nonperiodic or random conformation. The above methodology presented for lysozyme was carried out on each of the spectra for each of the two polymers. For the FSD spectra, a half-width at half-height of  $9\text{ cm}^{-1}$  and an REF of 2.3 was used. The nonlinear regression analysis was started with 29 component bands, and the Lorentzian/Gaussian character of the bands was permitted to vary. For polyaspartic acid, only three major bands at  $1648\text{ cm}^{-1}$ ,  $1581\text{ cm}^{-1}$  and  $1534\text{ cm}^{-1}$  with fractional area of 40%, 32% and 28%, respectively, were present. The nonlinear

regression program caused the excess bands to converge to the lower boundary of zero for the height, then these bands were removed from the calculation. The program's result also had all the bands with pure Gaussian character. The results of these calculations for the FSD and original FTIR spectra of polyaspartic acid are presented in Fig. 2B (lower spectra are FSD; the upper are original). Excellent fits for both spectra can be observed, and the fractional areas are equivalent for both spectra. Results for polylysine yielded only two major bands at  $1646\text{ cm}^{-1}$  and  $1536\text{ cm}^{-1}$  with fractional areas of 70% and 30%, respectively. A small band at  $1745\text{ cm}^{-1}$  with a less than 0.4% fraction of the total area was also present for both of these polypeptide samples. The total study was also repeated at a lower salt concentration, which yielded the same results. It is apparent from these studies that the amide I assignment of  $1646\text{--}1648\text{ cm}^{-1}$  for a random coil or irregular conformation is appropriate [3]. Also, the  $1581\text{ cm}^{-1}$  band is most likely due to unprotonated side chain carboxyl groups.

Thus, the validity of the methodology for deconvolving FTIR spectra into component bands using nonlinear regression analysis is consistent for proteins in  $\text{H}_2\text{O}$ , proteins in  $\text{D}_2\text{O}$ , and polypeptides in  $\text{H}_2\text{O}$ . Even the number of component bands decreased markedly from 29 bands observed for lysozyme to 2 or 3 for the two polypeptides in irregular conformations. Also, no Lorentzian character is observable in any of the FTIR spectra.

Further validation of the calculated components of the amide I and II bands can be obtained by mathematical comparison of the second-derivative FTIR spectrum obtained from the original spectrum with the calculated second derivative obtained from the model fit. The result of such a fit to a second-derivative spectrum of lysozyme is shown in Fig. 3, where the irregular line represents the second-derivative spectrum of the fitted model and the smooth line that of the experimental data. The inset of this figure shows a pseudorandom plot of the connected residuals which further establishes the reliability of this methodology for quantitatively resolving FTIR spectra of proteins into component bands.

### 3.2. Number of parameters for the nonlinear regression analysis

To evaluate the effects of fitting a protein with too few component bands, let us examine the spectrum of lysozyme. Fig. 4A shows the fit of the Fourier self-deconvoluted amide I and II envelopes of lysozyme with 28 component Gaussian bands. The fit is excellent with no discernible difference between the resulting theoretical and experimental curves. However, when the band at  $1659\text{ cm}^{-1}$  is removed and only 27 peaks ( $N - 1$ ) are utilized, nonlinear regression analysis yields a poor fit, with some component bands becoming much broader than others (see Fig. 4B). Also, it can be seen that the fit in the amide II envelope is affected even though no band was removed from that range of frequencies. In addition, as seen in Table 1, the RMS for 27 peaks is six times larger than for 28 peaks, i.e. 0.00157 versus 0.000251, respectively. If additional bands are removed

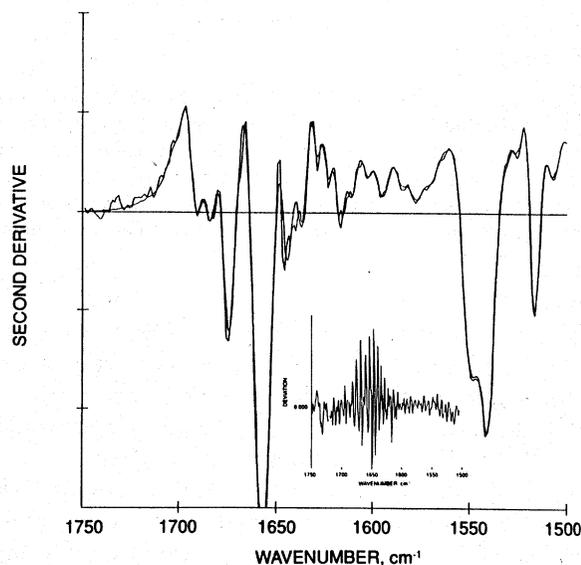


Fig. 3. Second-derivative FTIR spectrum of the amide I and II bands of lysozyme in aqueous solution. The single smooth line represents connected experimental data. The irregular line is the second derivative of the composite sum of the individual component bands from the nonlinear regression analysis as described in the text. The inset shows a plot of connected second-derivative residuals between calculated (composite sum of nonlinear regression analysis) and experimental second derivatives vs. frequency.

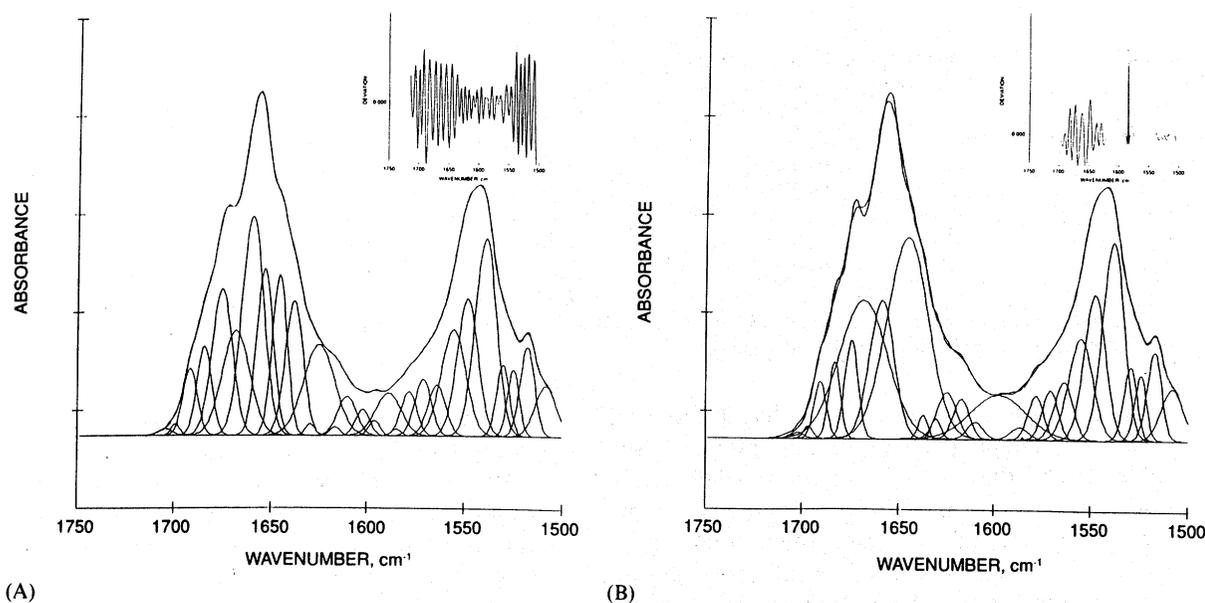


Fig. 4. Best fit for Fourier self-deconvoluted lysozyme FTIR spectrum using nonlinear regression analysis, a half-width at half-height ( $W$ ) of  $9\text{ cm}^{-1}$ , and a resolution enhancement factor, (REF) of 2.5. The light line represents connected experimental points; the dark line represents the theoretical sum of component bands shown by the light lines. (A) For 28 peaks; (B) for 27 peaks. Inserts in each show plots of connected residuals between spectra and composite nonlinear fit.

( $N-2$  and  $N-3$ ), the RMS continues to increase but at a smaller rate (see Table 1).

The above comparison is also observed for all the proteins reported here but is only illustrated in Table 1 for the fits of trypsin, elastase, and myoglobin. In all cases the fits after removing peaks ( $N-1$ ,  $N-2$ ,  $N-3$ ) in the area of  $1659\text{ cm}^{-1}$  are extremely poor. Removing the peaks results in some extremely broad peaks in the amide I region, which in turn influences the fit of the amide II region. It should be noted that FSD causes the component bands to have almost equal half-widths at half-heights, and in the work of Byler and Susi [3], the appearance of overly broad component bands in the nonlinear regression analysis results were considered unacceptable. The RMS values of the unacceptably poor fits are on the order of three to six times higher than the best fits (see Table 1). Despite the improved fits (better RMS values), the number of peaks which exist under the amide I or II envelope should be based upon more theoretically and statistically sound concepts to prove the above hypothesis.

Rusling et al. [11] have successfully used the sum of squares principle and the extra sum of squares  $F$ -test to determine the confidence level of whether an electrochemically active compound binds to a detergent micelle according to a monophasic or biphasic mechanism. In our study, Eq. (1) (Eq. (14) of Rusling et al. (11)) relates the RMS and the degrees of freedom (which is defined as the number of data points ( $n$ ) minus

$$F_{(p_2 - p_1, n - p_2)} = \frac{(S_1 - S_2)/(p_2 - p_1)}{S_2/(n - p_2)} \quad (1)$$

the number of parameters (bands,  $p$ ) in the nonlinear regression analysis) to an  $F$ -ratio, which in turn can be used in conjunction with standard  $F$ -test tables to obtain the confidence level with which the model with the smaller number of degrees of freedom fits the data better.

In the third column of Table 1, the probabilities ( $\rho$ ) are calculated for  $N$  number of bands from the RMS values in the  $N$  and  $N-1$  columns. In all cases, it is noted that the probability of  $N$  bands yields a confidence level of over 99%. Therefore using  $N$  bands is well-determined on a

Table 1  
Influence of number of Gaussian peaks on RMS values

Protein	$N$	$\text{RMS}_N \rho^a$ (%)	$\text{RMS}_{N-1}$	$\text{RMS}_{N-2}$	$\text{RMS}_{N-3}$
Lysozyme	28	0.000251, >99	0.00157	0.00176	0.00187
Trypsin	25	0.000362, >99	0.00140	0.00222	0.00281
Elastase	24	0.000624, >99	0.00174	0.00191	0.00197
Myoglobin	17	0.000306, >99	0.00154	0.00161	0.00166
Papain	26	0.000759, >99	0.00201	0.00213	0.00236

<sup>a</sup> Probability confidence level with which  $\text{RMS}_N$  can be considered significantly smaller than  $\text{RMS}_{N-1}$  after correction for degrees of freedom in each nonlinear regression fit analysis.

statistical basis. This calculation has been used for the remaining proteins in our database (see Table 2 for the list of proteins) which gives a statistical confidence for the number of component bands found for each protein. Such statistical tests should be utilized by all investigators when performing nonlinear regression analysis. The use of this statistical method clearly adds confidence to the choice in the number of component bands.

The above discussion deals with the error in choosing too few bands in the nonlinear regression analysis to fit the amide I and II envelopes. The opposite consideration would be choosing too many bands. One could argue that choosing an infinite number of infinitely narrow bands would lead to an exact fit. For an integral in calculus, this would be true, but in FTIR there are certain theoretical and realistic constraints available which control the maximum number of bands chosen. Theoretically, the number and position of bands should correlate with the calculated theoretical values of Krimm and Bandekar [10]. Realistically, the number of bands and their positions should approximate the number of bands observed in the second-derivative and FSD spectra. The approximate number of bands indicated by the second-derivative and FSD spectra will vary with the parameters chosen to carry out those analyses (e.g. the degree of smoothing chosen for the second-derivative spectra, or the REF and half-width at half-height chosen for the FSD spectra), so these values should not limit the number of bands in the nonlinear regression analysis, but should represent a good number to begin the regression analysis. The above theoretical and

realistic constraints should be considered when adding bands to the regression analysis.

An additional theoretical and realistic concern is that there can be no negative bands in fits for vibrational spectra. A feature of the ABACUS software we use for our nonlinear regression analysis is our ability to set a lower boundary for the heights of the bands at zero. So when we add too many bands, we have found the unnecessary bands ( $N+1$ ,  $N+2$ , ...) approach zero height. When this occurs, they are removed from the analysis and no additional bands are added.

### 3.3. Secondary structure assignment and theoretical justification

Controversy also exists in the literature concerning the assignment of the frequency of peaks to unique protein secondary structure. Studies attempting to resolve this issue have ranged from the theoretical to the experimental, as well as combinations of both.

Because of discrepancies with peak assignments, we have tentatively assigned our secondary structural elements to agree with the assignments of the following investigators: Byler and Farrell [4], Dong et al. [2], Dousseau et al. [12], Dousseau and Pezole [9], Krimm and Bandekar [10], Kuzmosinski and Farrell [13], noting of course, that these assignments can change in the future. We shall assign the  $\alpha$ -helix structure to 1650–1660  $\text{cm}^{-1}$ , the irregular or disordered structure to 1642–1648  $\text{cm}^{-1}$ , the strand or extended structure to 1624–1638  $\text{cm}^{-1}$ , and the turn conformations to all frequencies above 1670  $\text{cm}^{-1}$  up to 1695  $\text{cm}^{-1}$ . No attempt will be made to distinguish the

Table 2  
Percent influence of side chain — glutamine and asparagine

Protein	1651 cm <sup>-1</sup> (C-N)	1667 cm <sup>-1</sup> (C=O)	% Glutamine and asparagine
Hemoglobin	19.8 ± 0.7	7.2 ± 0.3	5.4
Myoglobin	31.2 ± 0.6	5.5 ± 0.4	3.4
Cytochrome C	18.3 ± 2.2	6.1 ± 0.4	3.6
Lysozyme	11.4 ± 0.3	14.6 ± 0.4	8.6
Ribonuclease	6.4 ± 0.5	13.6 ± 1.3	8.9
Papain	12.3 ± 1.2	22.8 ± 0.8	8.3
Pancreatic trypsin inhibitor	4.0 ± 0.4	5.1 ± 1.3	5.4
α-Chymotrypsin	11.6 ± 1.2	11.5 ± 0.3	7.0
Trypsin	8.4 ± 0.5	9.3 ± 1.1	7.9
Elastase	8.0 ± 0.5	10.0 ± 0.3	8.7
Carbonic anhydrase	11.3 ± 0.6	8.6 ± 0.5	6.1
β-Lactoglobulin	9.7 ± 0.3	10.0 ± 0.4	6.3
Concanavalin A	11.0 ± 0.9	8.3 ± 0.6	5.5

type of turns assigned to frequency ranges in this study.

To add a theoretical basis for the nonlinear regression model and the tentative frequency assignments that were chosen, we turn our attention to the theoretical work of Torii and Tasumi [14–16]. In a series of papers they have developed a model for the calculation of amide I envelopes from the X-ray crystallographic data of proteins. Their model consists of assigning one oscillator with a transition dipole to each peptide group. Coupling between these oscillators is then introduced through a transition dipole coupling mechanism. While this method allows for faster computational times, contributions from disulfide bonds and side chain–side chain and side chain–backbone are nonexistent. In one paper by Torii and Tasumi [14], the theoretical FTIR amide I spectrum of lysozyme was calculated using the X-ray crystallographic data. In their calculation, they utilized a Gaussian envelope with a half-width at half-height of 3.0 cm<sup>-1</sup> for each peptide oscillator. Their calculations were used to qualitatively compare theoretical spectra of several proteins with their experimental counterparts in D<sub>2</sub>O, so the force constants used were optimized to agree with D<sub>2</sub>O and not H<sub>2</sub>O results. For all the above reasons we cannot exactly compare our experimental results for lysozyme with their theoretical spectrum. We can, however, deconvolve their spectrum for lysozyme into the component

Gaussian peaks using nonlinear regression analysis and compare the number of peaks and fractional areas with our experimental FTIR spectrum. Here, it must be stressed that the theoretical spectrum must contain at least but not less than the same number of bands in our experimentally analyzed spectrum.

Fig. 5 shows the deconvolved theoretical FTIR spectra from Torii and Tasumi [14] with the best fit of the sum of 14 Gaussian bands for the amide I region. Attempts to use less than 14 bands resulted in poor fits, while the addition of more

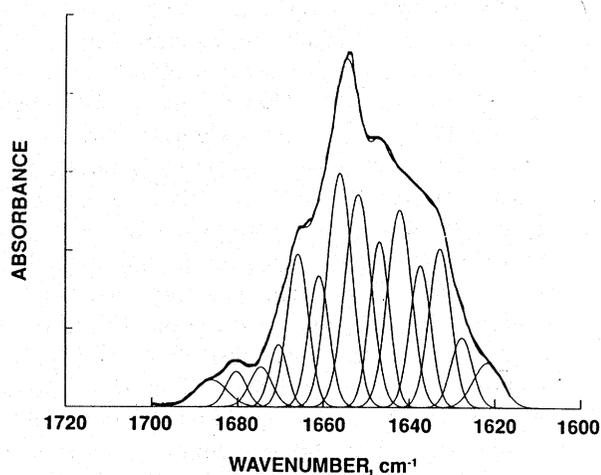


Fig. 5. Best fits by nonlinear regression analysis using the amide I band of lysozyme. (From theoretical calculations of Torii and Tasumi [14].)

peaks caused the height of the extra bands to approach a zero or negative value. Calculation of the percentage of extended, helix, turn and irregular structure from the fractional areas of Fig. 5, and using the above assignments, yields values of 25.2%, 26.0%, 36.3% and 12.5%, respectively. Inspection of Tables 3, 4 and 5 yields experimental values of 28.9%, 23.4%, 32.6% and 15.3% for the percentage of extended, helix, turn and irregular conformation of lysozyme, respectively. The agreement between the theoretical calculations and the experimental data is acceptable because of the assumptions in the calculation of Torii and Tasumi [14]. Other proteins which were common to both studies (myoglobin, concanavalin A, trypsin, and ribonuclease) also yielded comparable results well within a 5% deviation. It also should be noted that the model of Torii and Tasumi [14–16] does not take into account any water interactions or internal dynamic motion of the structure which would occur for a protein in solution.

#### 4. Discussion

##### 4.1. Contribution of asparagine (ASN) and glutamine (GLN) side chains

Recently, Venyaminov et al. [19] performed FTIR experiments on amino acids in water and deconvolved the spectra into component bands to ascertain the influence of side chains on the overall amide I envelope. They found that large bands due to the side chains of arginine (ARG), ASN and GLN exist within the amide I envelope. In subsequent articles [17,18], they attempted to eliminate side-chain contributions to the amide I and II envelopes by subtracting, on a molar basis, the amino acid side-chain absorptions from the amide I and II regions of the protein absorptions. They assumed that the absorptivity of the side chains of free amino acids and those in proteins were equivalent. This assumption may not be valid, when the dipolar coupling and energy changes are considered for the side chain which now is part of a compact globular protein structure. Finally, the DuPont curve analyzer used in

Table 3  
Percent extended structure

Protein	FTIR <sup>a</sup> result	R <sup>b</sup> result
Hemoglobin	7.7	8.2
Myoglobin	10.4	5.9
Cytochrome C	10.5	11.6
Lysozyme	29.6	30.5
Ribonuclease	41.3	37.5
Papain	22.2	19.8
Pancreatic trypsin inhibitor	31.8	29.5
$\alpha$ -Chymotrypsin	38.8	37.9
Trypsin	40.6	39.6
Elastase	36.6	33.6
Carbonic anhydrase	41.7	36.0
$\beta$ -Lactoglobulin	44.5	44.2
Concanavalin A	44.2	44.5
Oxytocin	0	0
Polylysine	0	0
Polyaspartic	0	0

<sup>a</sup> Average error,  $\pm 1.6 \text{ cm}^{-1}$ .

<sup>b</sup> Ramachandran.

the latter study has a limited range of applicability to FTIR experimental data.

Basing our starting point on their results, we obtained spectra of ARG, polyARG, ASN, polyASN, GLN, and polyGLN in order to analyze the results for side-chain contributions. We fit the spectra using our nonlinear regression methodology. For the free amino acids we found some differences in the frequencies for the proposed contributions. These differences are probably due to the fact that Venyaminov et al. [19] subtracted the spectrum of alanine to arrive at the possible side-chain contributions of the amino acids. The frequency and the ratio of height to half-width at half-height need to be the same for spectral subtractions to be valid; this is not the case when subtracting alanine from ARG, ASN, or GLN spectra. Our methodology permits the identification of the side-chain contributions without the subtraction of the alanine spectrum. We found that the possible contribution for ASN and GLN is the same at  $1668 \text{ cm}^{-1}$  with a small band at  $1650 \text{ cm}^{-1}$ , while the possible contribution for ARG is at  $1679 \text{ cm}^{-1}$ . Further differences are noted for the polyamino acids. Here the influence of dipolar coupling is observed for the side-chain

Table 4  
Percent helix structure

Protein	$\alpha$		3/10, Bent strand	
	FTIR <sup>a</sup> result	R <sup>b</sup> result	FTIR <sup>a</sup> result	R <sup>b</sup> result
Hemoglobin	76.7	81.6	1.2	—
Myoglobin	76.4	78.1	1.9	—
Cytochrome C	46.2	44.6	2.2	4.0
Lysozyme	25.6	27.1	4.2	5.5
Ribonuclease	10.0	18.0	10.3	8.8
Papain	15.2	18.8	17.4	15.6
Pancreatic trypsin inhibitor	4.2	12.0	0	6.9 BE <sup>c</sup>
$\alpha$ -Chymotrypsin	11.9	12.4	3.5	6.8
Trypsin	17.4	10.4	0	8.3
Elastase	14.1	9.7	0	8.2
Carbonic anhydrase	14.2	14.2	1.7	7.1 BE <sup>c</sup>
$\beta$ -Lactoglobulin	15.8	5.0	3.1	9.2 BE <sup>c</sup>
Concanavalin A	13.6	1.7	2.2	8.8 BE <sup>c</sup>
Oxytocin	0	0	0	0
Polylysine	0	0	0	0
Polyaspartic	0	0	0	0

<sup>a</sup> Average error  $\alpha$ -helix,  $\pm 0.8$  cm<sup>-1</sup>.

<sup>b</sup> Ramachandran.

<sup>c</sup> BE, bend strand.

contributions of ASN and GLN but not for ARG. The contributions of ASN and GLN side chains relative to the amide I carbonyl band remains relatively the same while the ARG side-chain contribution relative to the amide I carbonyl decreases dramatically. Despite these differences we attempted to correlate the appropriate side-chain contribution to band area with their percent content in the proteins in our data base. We only found significant contribution and correlation for ASN and GLN combined. While the contribution of ARG appears to be negligible, it should not be discounted until further analysis is carried out on other proteins which have a higher ARG content than our proteins; in the latter case the results might correlate well.

Table 2 lists the area percent for the 1651 cm<sup>-1</sup> and 1667 cm<sup>-1</sup> bands in the first and second columns, respectively, for the indicated proteins in our database. Column 3 lists the combined percentage content of GLN and ASN residues present in the respective proteins. As can be seen, even though the global secondary structures of these proteins change with increasing row num-

bers (i.e. ranging from almost pure helix to turn, and then to strand structures), the percentage areas of the 1651 and 1667 cm<sup>-1</sup> bands are relatively invariant, with the exception of myoglobin, hemoglobin and cytochrome C. Higher percent values of the 1651 cm<sup>-1</sup> band for these first three helical proteins occurs as a result of the overlap of the E band of the  $\alpha$ -helix structure. If the 1651 cm<sup>-1</sup> bands of the three helical proteins are not considered, then a reasonable correlation exists between the experimental bands and the percentage of ASN and GLN residues. The results of the linear regression analysis of the area percent of the 1651 and 1667 cm<sup>-1</sup> bands versus the percent GLN and ASN, where the percent areas of the 1651 cm<sup>-1</sup> for the first three predominately helical proteins (hemoglobin, myoglobin and cytochrome C) are eliminated from this analysis, yielded an intercept value of 2.83 (standard error, 2.06;  $\sigma = 0.184$ ) and a slope of 0.948 (standard error 0.292;  $\sigma = 0.00388$ ). Thus, it appears that the intercept value could statistically have a zero value, with the slope having a value near unity. While the analysis suggests a correlation, it should

Table 5  
Percent non-periodic structure

Protein	Turn or twisted strand		Irregular	
	FTIR <sup>a</sup> result	R <sup>b</sup> result	FTIR <sup>a</sup> result	R <sup>b</sup> result
Hemoglobin	4.3	7.5	10.1	10.8
Myoglobin	2.3	2.1	8.8	14.6
Cytochrome C	37.9	35.8	3.1	3.1
Lysozyme	28.1	26.6	12.5	9.6
Ribonuclease	24.2	26.0	14.2	9.7
Papain	43.3	39.6	15.7	6.2
Pancreatic trypsin inhibitor	60.5	49.6	3.5	8.9
$\alpha$ -Chymotrypsin	25.2	22.8	20.6	20.1
Trypsin	26.3	28.8	15.7	16.7
Elastase	31.8	33.6	17.4	16.0
Carbonic anhydrase	22.3	26.0	20.1	16.3
$\beta$ -Lactoglobulin	17.3	21.2	19.3	20.4
Concanavalin A	22.2	22.7	17.7	22.3
Oxytocin	100.0	100.0	0	0
Polylysine	0	0	100.0	100.0
Polyaspartic	0	0	100.0	100.0

<sup>a</sup> Average error: turn or twisted strand,  $\pm 1.4 \text{ cm}^{-1}$ ; loop  $\pm 0.8 \text{ cm}^{-1}$ .

<sup>b</sup> R, Ramachandran.

be viewed as an assumption. Only when the analysis of at least 50 proteins yields similar results would this assumption be considered proven. In addition, new studies of deaminated proteins in  $\text{H}_2\text{O}$  could help resolve this issue.

For this report, we will assign these bands to GLN and ASN side-chain modes. Most likely the  $1651 \text{ cm}^{-1}$  band is assigned to a C-N deformation and the  $1667 \text{ cm}^{-1}$  band is assigned to C=O stretch. The true fraction of GLN and ASN should be subtracted from the fractional areas of these bands and any excess area arising should be assigned to the appropriate global secondary structure. If the percentage of area is less than the percentage of GLN and ASN residues present, then the experimental areas should be subtracted from the amide I envelope and all the remaining bands should be normalized.

#### 4.2. Ramachandran analysis

Of paramount importance to predicting the global secondary structure of proteins (by correctly interpreting the assignments of the FTIR amide I bands) is the correct calculation of the

secondary structure from the results of X-ray crystallography. Not only is the amount of  $\alpha$ -helix, turn and extended conformation important, but the length of the helix and extended conformation is important as well. In addition, whether or not internal backbone hydrogen bonding exists may also be relevant descriptors for correlation with the percent areas of the component bands of the amide I region. Until recently, researchers in this field used the values provided in the Brookhaven Protein Data Bank. However, no quantitative algorithm has been used for subdividing the periodic structure. The values depend on the definitions of conformation adopted by each crystallographer. The definitions can, also, change over a period of time. What is needed, is an algorithm consistent with FTIR results to be used on all X-ray crystallographic structures. To date, no consensus in the scientific community for the appropriate algorithm has been found.

Kalnin et al. [18] have recently subdivided both the helices and sheets into hydrogen-bonded and non-hydrogen-bonded conformations, which along with the turn and all other conformations form a basis of six instead of four conformations.

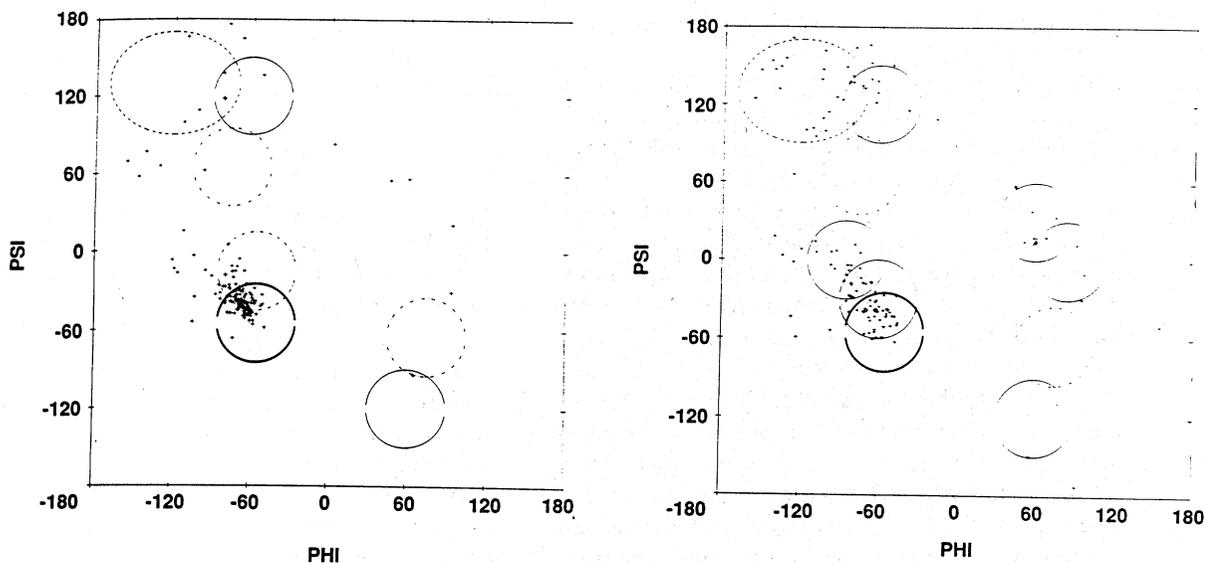


Fig. 6. Ramachandran plot from X-ray crystallographic structure of (A) myoglobin, and (B) lysozyme.

Their calculations, however, are correlated with FTIR results using factor analysis instead of non-linear regression analysis. They normalize the structures in their analyses (i.e. the total conformation of an individual protein adds up to 100%) and show reasonable correlation with FTIR results. Other researchers using factor analysis (e.g. Pancoska et al. [20,21], Pancoska and Keiderling [22]) do not achieve the same good correlations — they use only four conformations and do not normalize structures to 100%.

An algorithm developed recently in Liebman's laboratory [23,24] shows good agreement with FTIR results when deconvolving the amide I into component bands. However, these experiments were performed only for serine proteases in  $D_2O$  and were correlated with helix and extended structures. While this method appears to be promising, it is still in its infancy and more correlations between experimental results and calculated fractions of conformations must be reported, and may ultimately be correlated with the Gaussian components described above. In this study, we use the traditional Ramachandran plot calculated from the X-ray crystallographic structure of proteins in conjunction with the secondary conformations reported in the protein data banks. We strongly stress at this time that the adoption of Ramachan-

dran analysis does not in any way imply that the transitional dipolar coupling mechanisms of Krimm and Bandekar [10] or Torii and Tasumi [14–16] do not apply here to the vibrational spectroscopy of proteins. We believe their mechanisms are valid.

Here, we use the Ramachandran plot as a method for corroborating the reported fractional values. If discrepancies exist, we use other molecular modeling techniques available in the Sybyl Molecular Modeling software (such as inspection by ribbon) in conjunction with the Ramachandran analysis. The latter analysis, of course, does not take into account the required minimum number of sequential residues to sustain a periodic conformation. In Figs. 6A and 6B the Ramachandran plots are given for myoglobin and lysozyme, respectively. Fig. 6A shows a template Ramachandran with theoretical curves for predominately helical proteins. The broken lines in the upper left are for a  $\beta$  sheet structure. The solid circle next to it is defined as the second position for a  $\beta$  turn type II. Directly under this circle is the theoretical envelope for a one-residue inverse "a" ( $\gamma$ ) turn. Directly under the inverse "a" turn is the area which defines the 3/10 helix; the double line area represents the  $\alpha$ -helix region. At the lower right hand region is an area representative

of a type II  $\beta$  turn, and above it is the area common for an "a" ( $\gamma$ ) turn.

Fig. 6B uses a template plot which is more consistent with proteins containing a large amount of turn conformation. Here the broken lines and double lines represent the sheet and  $\alpha$ -helix regions, respectively, as in Fig. 6A. Adjacent to the sheet region is the second position for the type II  $\beta$  turn and directly under that region is the envelope for the one-center inverse "a" turn. The third position for a type I and II  $\beta$  turn region is below the inverse "a" turn. Directly above the double-line acceptable region for an  $\alpha$ -helix is the envelope for the second position of the type III  $\beta$  turn. In the lower right hand quadrant and proceeding upward are the acceptable regions for the conformations of the second position of a type II  $\beta$  turn, an "a" turn, the third position of type I and II  $\beta$  turn, and the second position of type III  $\beta$  turn, respectively. Also shown in Fig. 6B (as + symbols) are the calculated dihedral angles  $\phi$  and  $\psi$ , from the Sybyl modeling program analysis of the X-ray crystallographic structure of lysozyme (6LYZ).

A three-dimensional Ramachandran plot, where the residue number is plotted on the third axis, can also be calculated using the Sybyl molecular modeling software. With this plot, adjacent residues within a periodic structure as well as those residues which are part of a turn conformation can easily be determined. With all the above analyses, the global secondary structure calculated from the crystallographic data for each of the proteins studied in the FTIR was calculated. The results along with the corresponding experimental values are presented in Tables 3, 4 and 5 for strand, helix, and turn with irregular, respectively. The results for each conformation will be discussed in the following sections of this report.

#### 4.3. Extend (strand) content

Table 3 shows a comparison between the experimentally determined global  $2^\circ$  structure of the extended conformation (column 2, headed FTIR) along with values calculated by the composite Ramachandran analysis (column 3, headed R). The amount of extended conformation was deter-

mined from FTIR amide I results by summing the component bands from 1638 to 1623  $\text{cm}^{-1}$ . Bands at frequencies lower than 1623  $\text{cm}^{-1}$  and closer to 1615  $\text{cm}^{-1}$  are presumed to be caused by unprotected side-chain carboxyl groups [10]. Note that we assumed the bands (1638–1623  $\text{cm}^{-1}$ ) quantify not only the amount of sheet conformation but also extended or strand non-hydrogen-bonded structures, as did Prestrelski et al. [23–24]. For a proper comparison, we sum all the points in the upper left hand quadrant of the Ramachandran plot for  $\phi$  values above 130°, as well as those within the sheet region. Connectivity for any extended structure was determined using a three-dimensional Ramachandran plot (not shown).

As shown in Table 3, comparison between the experimental (FTIR) and calculated (R) extended conformation is excellent. Only in carbonic anhydrase, a high helical protein, is the deviation greater than 4%. Quantitative correlation of the amount of extended structure determined via FTIR and Ramachandran analyses of the X-ray crystal structure for the 16 proteins in this database was accomplished using linear regression analysis and polynomial curve fitting analysis. Using the *F*-test, it was found for the helix, turn and irregular structure as well as the extended structure that linear regression analysis was the most statistically significant methodology. For the extended structure, linear regression analysis yielded an intercept of  $0.65 \pm 0.95$  ( $\sigma = 0.49$ ) and a slope of  $1.03 \pm 0.03$  ( $\sigma = 0.0001$ ). The standard deviation for the regression analysis was 2.07%. In addition, this good correlation for the extended structures is supported by the fact that all proteins except for carbonic anhydrase lie within the 0.95 confidence level in the linear regression analysis. For carbonic anhydrase, if the excess amount of white extended structure (calculated from Ramachandran analysis in Table 4, last two columns) minus the amount from FTIR at 1667–1669  $\text{cm}^{-1}$  (i.e.  $7.1 - 1.7 = 5.4$ ) is added to the amount of extended structure from Ramachandran analysis in the last column of Table 3 (i.e.  $36.0 + 5.4 = 41.4$ ), then good agreement is obtained with the experimental amount of extended structure via FTIR, 41.7. Thus, the need for better mathematical algorithms for the calculation of an

global secondary substructure from the X-ray crystal structure is imperative for correlating with FTIR global secondary structure.

We compared our results for trypsin,  $\alpha$ -chymotrypsin and elastase with those reported by Liebman's group — calculated and experimentally determined values in D<sub>2</sub>O [23,24]. Values calculated using the Liebman algorithms were 42, 42 and 37. Their experimental results were 39, 45, and 46, while our experimental values were 41, 39 and 37, (all values respective to trypsin,  $\alpha$ -chymotrypsin, and elastase). While both our experimental values for trypsin and  $\alpha$ -chymotrypsin agree equally with the calculated values, our elastase value of 37 is more in agreement with the calculated value of 37 than the experimental value of 46% in D<sub>2</sub>O [23,24]. This adds further support for our methodology and the supposition that the amount of GLN and ASN side-chain residues must be subtracted from the 1667 and 1651 cm<sup>-1</sup> component amide I bands.

#### 4.4. Helix content

Table 4 lists the calculated and experimental results for the  $\alpha$ -helix (columns 1–3), 3/10 helix (columns 5, 6). Linear regression analysis of the amount of helical structure determined via FTIR versus the amount calculated from three-dimensional Ramachandran analysis of the X-ray structure for these 16 proteins yielded an intercept of  $2.13 \pm 1.81$  ( $\sigma = 0.26$ ) and a slope of  $0.92 \pm 0.06$  ( $\sigma = 0.0001$ ). The standard deviation of the regression analysis was 5.6%, which is significantly larger than the error calculated for the amount of extended structure, especially for the low content helical proteins. In fact, five proteins did not fall within the 95% confidence region. Close inspection of these differences may provide a rationale for the change. Not counting the high helical proteins (hemoglobin and myoglobin) we observe that the largest differences occur for ribonuclease, the serine protease system (i.e. pancreatic trypsin inhibitor (PTI), and trypsin), concanavalin A (CON A) and  $\beta$ -lactoglobulin. The last two proteins in this series contain significant amounts of  $\beta$ -barrel structures. Such structures appear as antiparallel  $\beta$ -sheets which are highly bent. The

Ramachandran plots of these proteins also yield points in the lower left quadrant which is normally considered a forbidden region. The Ramachandran plots for the serine proteases also contain some  $\phi$ ,  $\psi$  angles in this region. Hence, the discrepancy in the experimental and calculated helical content for concanavalin A,  $\beta$ -lactoglobulin and perhaps the serine proteases may be caused by  $\beta$ -barrels which could have bands at  $1658 \pm 2$  cm<sup>-1</sup>. More experimental and theoretical studies must be performed before this hypothesis can be concluded. But since ribonuclease contains no  $\beta$ -barrel, inspection of the work of Kalnin et al. [18] may provide an answer to the discrepancy. In their study, the  $\alpha$ -helix was subdivided into an ordered and unordered class as was the sheet structure. They calculated values of 13% and 10% for their ordered and unordered  $\alpha$ -helix and found experimental values of 11% and 8%, respectively. Upon inspection of the ribbon structure, a major distortion of the helical region of ribonuclease is found. In addition, a value of 10.3% has been obtained from the excess area of the 1667 cm<sup>-1</sup> band which we have assigned as a 3/10 helix, in accordance with the results of Krimm and Bandekar [10]. If the  $\alpha$ -helix and 3/10 helix values are summed they add up to more acceptable values. However, the Ramachandran plot for ribonuclease shows 11 residues within the type III or 3/10 helix region which calculates to a theoretical value of 8.8% for these possible conformations. It should also be stressed that Kalnin et al. [18] calculates an ordered  $\alpha$ -helix conformation of 27% for lysozyme which agrees well with our experimental and theoretical values of 25.6% and 27.1%. Therefore, we believe that the discrepancy for the ribonuclease helical structure is the result of its structure, which our Ramachandran analysis could not adequately calculate.

The discrepancy between the amount of helix determined from Ramachandran analysis and FTIR experiments for trypsin can be easily explained by closer inspection of Table 4. Ramachandran analysis of trypsin yields 10.4%  $\alpha$ -helix, and 8.3% 3/10 helix content where the experimentally determined values from FTIR are 17.4% and 0%, respectively. As can be seen in the Ramachandran plot for myoglobin (Fig. 6A),

there is a large overlap between the acceptable regions for the  $\alpha$  and 3/10 helices. This is due to the small differences between their respective  $\phi$  and  $\psi$  values. For this reason it is not possible by this method to precisely determine the differences between these two structures. Moreover, a protein in solution does not adopt a static structure but an average dynamic structure. The dynamic structure, due to thermal fluctuations, causes internal motions within the solvated protein configuration. Under these circumstances, the  $\phi$ ,  $\psi$  angles for a 3/10 helix may easily change to those for an  $\alpha$ -helix. Only intense molecular dynamics calculations of a protein in water can attempt to approximate the dynamic motions of a protein. However, for this study it will be assumed that the 3/10 helix calculated via Ramachandran analysis closely approximates an  $\alpha$ -helix. Therefore, we now change the value of the  $\alpha$ -helix calculated from Ramachandran analysis to 18.7% (= 10.4% + 8.3%) which more closely agrees with the experimental value of 17.4%.

It should be stated that the 1676  $\text{cm}^{-1}$  band was also summed along with the excess area for the 1651  $\text{cm}^{-1}$  band, and the 1658  $\text{cm}^{-1}$  band for obtaining the total  $\alpha$ -helix content of hemoglobin and myoglobin. The 1676  $\text{cm}^{-1}$  band represented approximately 17% of the total helical structure. The areas of the 1658  $\text{cm}^{-1}$  and 1651  $\text{cm}^{-1}$  bands summed to 63%, and a value of 63% was also calculated as the amount of unordered helix content in myoglobin by Kalnin et al. [18]. Moreover, close inspection of the ribboned structures of hemoglobin and myoglobin reveal highly distorted helical segments which could not be observed using Ramachandran analysis. The 1676  $\text{cm}^{-1}$  band, assigned by Krimm and Bandekar [10] as a turn may be reflective of a type III  $\beta$ -turn. Such a turn would have  $\phi$ ,  $\psi$  values seriously overlapping the  $\alpha$ -helical region of a Ramachandran plot (see Fig. 6B). Nevertheless, for high helical proteins (i.e. above 55%) it may be more prudent for investigators to utilize UV circular dichroism analysis, because as Torii and Tasumi [14–16] have recently reported, a serious overlap of E and A bands for  $\alpha$ -helices with varying lengths occurs, thus resulting in theoretical amide I envelopes which contain bands well

below the 1650  $\text{cm}^{-1}$  region. With lower helical proteins, FTIR correlates much better than circular dichroism since the turn conformation can be more easily determined.

Finally, the excess areas in the 1667  $\text{cm}^{-1}$  band above the GLN and ASN side-chain contribution is shown in Table 4 as a 3/10 helix or bent strand, i.e. a possible “a” turn. These small values cannot be easily correlated with Ramachandran analysis and no firm assignments are made. Excellent correlation between three-dimensional Ramachandran analysis and FTIR determination of the  $\alpha$ -helix at 1659  $\text{cm}^{-1}$  and the 3/10 helix at 1667  $\text{cm}^{-1}$  is obtained for papain (see Table 4). However, we do not observe any large amounts of 3/10 helix in lysozyme, especially in the 1638  $\text{cm}^{-1}$  region where Prestrelski et al. [23,24] have concluded that such 3/10 helical bands exist. While we have not performed any experiments on  $\alpha$ -lactalbumin, we still concur with the assignment of Krimm and Bandekar [10] that the 3/10 helix is in the range of 1665  $\text{cm}^{-1}$  rather than in the low range of 1638–1640  $\text{cm}^{-1}$  as reported by Prestrelski et al. [23,24]. Perhaps this discrepancy may be explained by the fact that their experiments were performed in  $\text{D}_2\text{O}$  rather than in  $\text{H}_2\text{O}$ .

With the above changes in the  $\alpha$ -helix percentages, we can now recalculate the correlation (via linear regression analysis) between the experimentally determined FTIR values and the Ramachandran calculated amount of  $\alpha$ -helix.

#### 4.5. Turn and irregular content

Table 5 shows the turn and irregular content determined experimentally from analysis of the FTIR amide I band and calculated from the three-dimensional Ramachandran analysis of X-ray crystallographic structures of the 16 listed proteins. The turn content was determined from the sum of all amide I bands from 1670 to 1694  $\text{cm}^{-1}$ . The irregular content was calculated from the normalized area of the  $1646 \pm 2 \text{ cm}^{-1}$  band. The irregular theoretical structure was calculated as all other structure not defined by this analysis. Good agreement between the experimental and theoretical values is observed with the standard deviations from the regression analysis of 3.6%

and 3.2% calculated for the turn and irregular content, respectively. Linear regression analysis of the FTIR values vs. the Ramachandran values yielded an intercept of  $0.97 \pm 1.39$  ( $\sigma = 0.49$ ) and a slope of  $1.00 \pm 0.04$  ( $\sigma = 0.0001$ ) calculated for the turn conformation, and an intercept of  $-0.75 \pm 1.02$  ( $\sigma = 0.98$ ) and a slope of  $1.01 \pm 0.03$  ( $\sigma = 0.0001$ ) calculated for the irregular or "all other" conformation. These values for the turn and irregular conformations are well within the standard deviations observed in the strand and  $\alpha$ -helix content, i.e. 2.1% and 5.5%. However, it should be noted that in the case of cytochrome C and  $\beta$ -lactoglobulin,  $\phi$  and  $\psi$  values exist in the upper right hand region of the Ramachandran plot. Although this region has been considered forbidden, closer inspection shows that these  $\phi$  and  $\psi$  values are a result of twisted sheets. Hence, since no other proteins in this database exhibited  $\phi$  and  $\psi$  values in this region, we have concluded that the  $1676 \text{ cm}^{-1}$  band may also be assigned to a twisted strand. However, more studies must be performed before a definite assignment can be made.

## 5. Conclusions

Calculation of the component  $2^\circ$  structural elements of the vibrational bands, i.e. approximately 25 Gaussian bands, was accomplished by fitting both the amide I and II bands using nonlinear regression analysis of the Fourier self-deconvoluted spectrum, the second-derivative spectrum, and the original spectrum. Fixed frequencies initially used in the original spectral analysis were obtained from both the FSD spectrum and the second-derivative analyses. The criterion for the acceptance of any analysis was that the fractional areas calculated from all three methods were in agreement (within experimental error calculated from the nonlinear regression analysis). Results clearly show that  $2^\circ$  structural conformations determined in water were in better agreement with global  $2^\circ$  structural analysis of X-ray structures than the previously reported values determined in  $\text{D}_2\text{O}$ . Also, with  $\text{H}_2\text{O}$ , the  $2^\circ$  structural elements can be calculated from the amide II envelope to

be used for validation of amide I assignments. In addition, the resolution of amide I spectra in water is greater than that in  $\text{D}_2\text{O}$  as observed in comparable second-derivative spectra and lower half-width values of 9 instead of 13–18  $\text{cm}^{-1}$  for the  $\text{H}_2\text{O}$  and  $\text{D}_2\text{O}$  spectra, respectively. The deterioration of resolution of FTIR spectra in  $\text{D}_2\text{O}$  may result primarily from the incomplete exchange of protein protons to deuterons. These results lay the foundation for the study of conformational changes in proteins induced by ligands, cosolutes or perhaps structural changes from site-directed mutagenesis [7], and in other system applications where for example  $\text{D}_2\text{O}$  may obscure water-protein interactions.

In this study, we have presented a method for analyzing the FTIR of proteins in water and determining their global secondary structure. Analysis of 16 proteins whose X-ray crystallographic structures are known showed agreement of secondary structure content to within 2–6% between the experimental FTIR and the X-ray coordinate calculations. The bands which are assigned to these structures are shown in Table 6 along with their tentative structural assignments. These assignments differ from previous assignments for only the GLN and ASN side-chain contributions and the  $1675 \text{ cm}^{-1}$  band which is now assigned to a turn rather than to an extended conformation. The high frequency  $\beta$  structure band at  $1675 \text{ cm}^{-1}$  reported by Byler and Susi [3] was not observed in this study. In our data base we have tentatively assigned this frequency as a turn conformation. This discrepancy may have arisen from the fact that Byler and Susi [3] did not consider turn conformations in their study, nor did they account for GLN or ASP side-chain contributions. In this study we tentatively assign the  $1676 \text{ cm}^{-1}$  band to the "a" turn conformation. While these results are quite promising, it must be stressed that only after a database of at least 50 proteins is obtained can any definite conclusions be reached. It is hoped that this study will inspire other investigators to adopt this methodology and add more information to increase this database above 16 proteins.

There are recent reports in the literature on the potential of FTIR to evaluate the global  $2^\circ$  struc-

ture of proteins which discuss the limitations of "band-narrowing" methodologies used to analyze FTIR spectra of proteins [25,26]. Our methodology addresses some of these limitations.

(1) We have eliminated atmospheric water completely with a thorough dry nitrogen purge and an accurate vapor spectra subtraction.

(2) We have considered the influence of side-chain absorptions, and recommend compensating for ASN and GLN side-chain absorptions as described above.

(3) We obtain a large number of interferograms to insure a high signal-to-noise ratio (S/N). A high S/N is necessary to ensure accurate vapor and buffer subtractions. A high S/N also reduces the influence of the instrumental noise.

(4) The ABACUS nonlinear regression program eliminates the operator influence on band shape — the relative Gaussian/Lorentzian contribution is determined by the fit.

(5) We statistically choose the number of bands used to fit the spectral data with the use of the extra-sum-of-squares *F*-test and the ABACUS program's ability to lower-bound all bands at zero height.

(6) Although we only use the amide I results for assignments, we fit both the amide I and II envelopes simultaneously. This improves the amide I fit compared to fitting the arbitrarily terminated amide I envelope alone.

(7) We use and analyze the original, the FSD, and the second-derivative spectra, accepting results only when all three regression fits statistically agree. This addresses the question of unique fits.

(8) We fit the spectra from two different starting values — small heights and large heights — requiring both analyses to come to the same answer. This eliminates the possibility of local minima being reached and ensures a unique fit.

(9) In general our methodology eliminates as much operator influence on the nonlinear regression fitting process as possible.

(10) Spectra are obtained in H<sub>2</sub>O. This is the natural environment of proteins and the environment in which most X-ray crystal structures are analyzed. This eliminates any possible conformational distortions which may occur with the use of D<sub>2</sub>O.

(11) We reanalyze each protein's X-ray crystal structure data using three-dimensional Ramachandran analysis. This ensures a consistent 2° structural X-ray analysis procedure from which to compare the protein structures obtained by FTIR.

(12) Finally, we recommend the use of amino acid secondary-sequenced-based prediction algorithms in conjunction with FTIR to analyze proteins with unknown structure.

## References

- [1] S.N. Timasheff, in H. Peeters (Ed.), *Protides of the Biological Fluids*, 20th Colloquium, 1973, pp. 511–519.
- [2] A. Dong, P. Huang and W.S. Caughey, *Biochemistry*, 29 (1990) 3303.
- [3] D.M. Byler and H. Susi, *Biopolymers*, 25 (1986) 469.
- [4] D.M. Byler and H.M. Farrell, Jr., *J. Dairy Sci.*, 72 (1989) 1719.
- [5] P.R. Griffiths and J.A. de Haseth, *Fourier Transform Infrared Spectrometry*, Wiley, New York, 1986, pp. 102–103.
- [6] L.T. Kakalis and T.F. Kumosinski, *Biophys. Chem.*, 43 (1992) 39.
- [7] J.K. Kauppinen, D.J. Moffatt, H.H. Mantasch and D.G. Careron, *Appl. Spectrosc.*, 35 (1981) 271.
- [8] W. Damert, *Quantum Chem. Program Exchange Bull.*, 14(4) (1994) 61. (Note: Version F.1 is available on the Internet from QCPE (their program number 652) or directly from ERRC. Persons desiring the program may use FTP, connect to "ishtar.arserrc.gov", and use the account "anonymous" which requires no password. They would then select directory "abacusf" and download all files. Installation instructions and user documentation are also provided.)
- [9] R. Dousseau and M. Pezolet, *Biochemistry*, 29 (1990) 8771.
- [10] S. Krimm and J. Bandekar, *Adv. Protein Chem.*, 38 (1986) 181.
- [11] J.F. Rusling, C.-N. Shi and T.F. Kumosinski, *Anal. Chem.*, 60 (1988) 1260.
- [12] F. Dousseau, M. Therrien and M. Pezplet, *Appl. Spectrosc.*, 43 (1989) 538.
- [13] T.F. Kumosinski, H.M. Farrell, Jr., *J. Protein Chem.*, 10 (1991) 3.
- [14] H. Torii and M. Tasumi, *J. Chem. Phys.*, 96 (1992) 3379.
- [15] H. Torii and M. Tasumi, *J. Chem. Phys.*, 97 (1992) 86.
- [16] H. Torii and M. Tasumi, *J. Chem. Phys.*, 97 (1992) 92.
- [17] S.Yu. Venyaminov and N.N. Kalnin, *Biopolymers*, 30 (1990) 1259.
- [18] N.N. Kalnin, I.A. Baikalov and S.Yu. Venyaminov, *Biopolymers*, 30 (1990) 1273.
- [19] S. Yu. Venyaminov and N.N. Kalnin, *Biopolymers*, 30 (1990) 1243.

- [20] P. Pancoska, S.C. Yasui and T.A. Keiderling, *Biochemistry*, 30 (1991) 5089.
- [21] P. Pancoska, L. Wang and T.A. Keiderling, *Protein Sci.*, 2 (1993) 411.
- [22] P. Pancoska and T.A. Keiderling, *Biochemistry*, 30 (1991) 6885.
- [23] S.J. Prestrelski, A.L. Williams and M.N. Liebman, *Proteins, Structure, Function and Genetics*, 14 (1992) 430.
- [24] S.J. Prestrelski, A.L. Williams and M.N. Liebman, *Proteins, Structure, Function and Genetics*, 14 (1992) 440.
- [25] W.K. Surewicz, H.H. Mantsch and D. Chapman, *Biochemistry*, 32 (1993) 389.
- [26] E. Goormaghtigh, V. Cabiaux and J.M. Ruyschaert, in H.J. Hilderson and G.B. Ralston (Eds), *Subcellular Biochemistry*, Vol. 23, *Physicochemical Methods in the Study of Biomembranes*, 1994, p. 405-450.