

## Global-Secondary-Structure Analysis of Proteins in Solution

### Resolution-Enhanced Deconvolution Fourier Transform Infrared Spectroscopy in Water

Thomas F. Kumosinski and Joseph J. Unruh

Eastern Regional Research Center, Agricultural Research Service,  
U.S. Department of Agriculture, 600 East Mermaid Lane,  
Philadelphia, PA 19118

Previous studies comparing the global secondary ( $2^\circ$ ) structure of proteins by Fourier deconvolution FTIR were performed in  $D_2O$ .  $D_2O$  however increases hydrophobic interactions leading to spurious  $2^\circ$  structural changes. We have now performed FTIR experiments in  $H_2O$  on globular proteins with varying amounts and types of  $2^\circ$  structure. A method has been developed to increase the sensitivity of the analysis of these FTIR spectra. Calculation of the component  $2^\circ$  structural vibrational bands was accomplished by fitting both amide I and amide II envelopes by nonlinear regression analysis. The method entails fitting of: Fourier deconvoluted spectra, second derivative spectra, and refits of the component bands to the original spectra. Criteria for acceptance of the analysis was that the fractional areas from all three methods were in agreement. Results show good agreement with known X-ray crystallographic structures, and allow prediction of  $2^\circ$  structures for non-crystallizable proteins.

During the past several decades, controversy has existed within the literature concerning the methodology appropriate for analyzing experimental results to obtain the global secondary structure of proteins in solution; whether the experimental method used was circular dichroism (CD), Fourier transform infrared (FTIR) or vibrational circular dichroism (VCD). Traditional CD experiments were analyzed using a sum of Gaussian bands ( $I$ ). In that publication, the criterion for the correct number of bands was established when the theoretical bands not only fit the CD data, but also when they were mathematically transformed (via a Krönig-Cramers transformation) to theoretical

optical rotatory dispersion (ORD) curves, with these curves agreeing with experimentally determined ORD results. However, this type of analysis necessitated performing ORD as well as CD experiments, which became cumbersome and costly for most investigators. The next method for analyzing CD results utilized factor analysis, and hypothesized that the CD spectra for an unknown protein was equal to a linear combination at each wavelength of pure secondary structural elements — such as  $\alpha$ -helix, random,  $\beta$ -sheet, etc. To develop a basis set for analyses, model polypeptides known to adopt to almost pure structure ( $\alpha$ -helix or  $\beta$ -sheet, etc.) were analyzed. However, this basis set was short lived because the fits to unknown protein data were poor.

With the increase in the number of protein structures determined from X-ray crystallography, many scientists developed new basis sets using the calculated secondary structure from the X-ray crystal structure of proteins. At first, factor analysis seemed appropriate due to the extremely low signal-to-noise and high error of the CD experiment. However, the questions of which class of proteins should be used and how the secondary structure should be calculated from the X-ray crystallographic structure became a never ending stumbling block for proteins with low  $\alpha$ -helix and high  $\beta$ -sheet or turn conformations. It should be noted that many investigators who used basis set methodology have reported only the results of their calculations. They usually did not show a plot of the theoretical versus the experimental spectrum. If a good fit between the experimental and theoretical curve is not achieved, the basis set used is inappropriate for analysis and a new basis set should be sought. The question is whether such a basis set does, in fact, really exist.

When FTIR instruments with extremely good precision, accuracy and signal-to-noise were developed, current investigations of the IR spectra of proteins were possible. Unfortunately, factor analysis of the spectra rather than deconvoluting the spectra into its component bands was practiced.

Examining theoretical principles, other research groups (such as Mantsch (2) and Susi (3,4)) noted that the amide I band was a sum of badly overlapped Gaussian or Lorentzian bands. They adapted a methodology using calculated second derivative spectra, Fourier deconvolution algorithms, and nonlinear regression analysis for deconvoluting the amide I envelope into individual component bands. However, controversy exists to this day concerning the number of bands, the fraction of Lorentzian character, and the choice of parameter values for Fourier deconvolution.

In this paper we now present FTIR experiments in H<sub>2</sub>O on thirteen globular proteins with varying types and amounts of 2° structures. Analysis of the spectral data using Fourier deconvolution, second derivative, and band curve-fitting techniques allows the individual 2° structural components to be distinguished and compared (3,5) with the known X-ray crystallographic data.

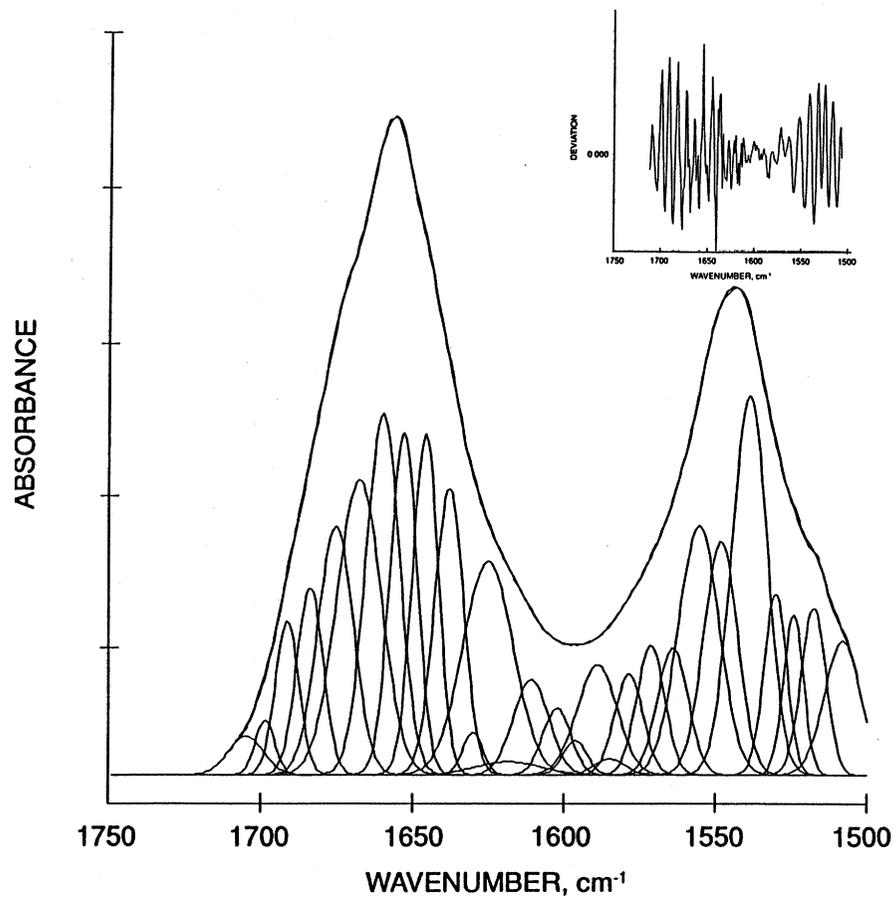
## Methods

**Infrared Measurement.** The individual proteins were prepared as 4% solutions in 20 mM, pH = 6.7 imidazole buffer. All samples were introduced into a demountable cell with CaF<sub>2</sub> windows separated by a 12  $\mu$ m Teflon spacer. Spectra were obtained using a Nicolet 740 FTIR spectrometer equipped with the Nicolet 660 data system. Following nitrogen purge of the sample chamber to reduce water vapor to a minimum, data collection was carried out. Each spectrum consisted of 4096 double-sided interferograms, co-added, phase-corrected, apodized (Happ-Genzel function), and fast-Fourier transformed. 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 (6-7).

**Data Analysis.** Difference spectra obtained by subtraction of buffer absorption from the respective protein solution absorptions were used to calculate second-derivative spectra by a simple analytical procedure that used every data point (3). Second-derivative spectra served as sensitive indicators for identifying individual peak positions used in subsequent processing. The unresolved spectra were subjected to Fourier deconvolution (FD) using an algorithm developed from the one described by Kauppinen et al. (2). The deconvolution was undertaken with a number of resolution enhancement factors. Qualitatively, under-FD was judged by the absence of peak position indications in the spectra and over-FD by the appearance of side lobes and deconvolved noise in the flat portions of the spectra (5,8). 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 developed at this laboratory, which assumes Gaussian band envelopes for the resolved components. In practice, the three parameters of each band (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, such as helices, sheets, turns, and loops (9). The areas serve to estimate the fraction of the various secondary elements in the protein molecule. Note the terms deconvolution, deconvolving and deconvolved will refer to the nonlinear regression fitting procedure.

**Sample Calculation: FTIR Analysis of Lysozyme.** A typical FTIR spectrum of hen's egg white lysozyme showing just the amide I and amide II regions is in Figure 1 (outer envelope). The spectrum is considered to be a sum of the variety of individual absorption bands arising from the specific structural components of the protein — such as  $\alpha$ -helix,  $\beta$ -sheets and turns. Fitting it directly with an undefined number of Gaussian bands by nonlinear regression would be a daunting task. To alleviate this dilemma, we first examine the



**Figure 1.** FTIR spectrum showing amide I and amide II bands of lysozyme in aqueous solution. Outer envelope double line is connected original spectrum. Line on outer envelope and individual component peaks underneath are the results of nonlinear regression analysis as described in text. Inset shows plot of residuals (connected by line) of the differences between the calculated and experimental absorbances vs. frequency.

second derivative of the spectrum (inset, Figure 2) to determine the number of component bands and the approximate positions of these bands.

The next step in the analysis is to enhance the resolution of the original spectrum via the FD algorithm developed by Kauppinen et al. (2). Care must be taken to choose the proper values for the band width and resolution enhancement factor used in this algorithm, so that the FTIR spectrum is not over- or under-deconvoluted. As the deconvolution procedure progresses, analysis of the FD spectrum by nonlinear regression analysis is used in an iterative fashion to determine the proper FD parameters.

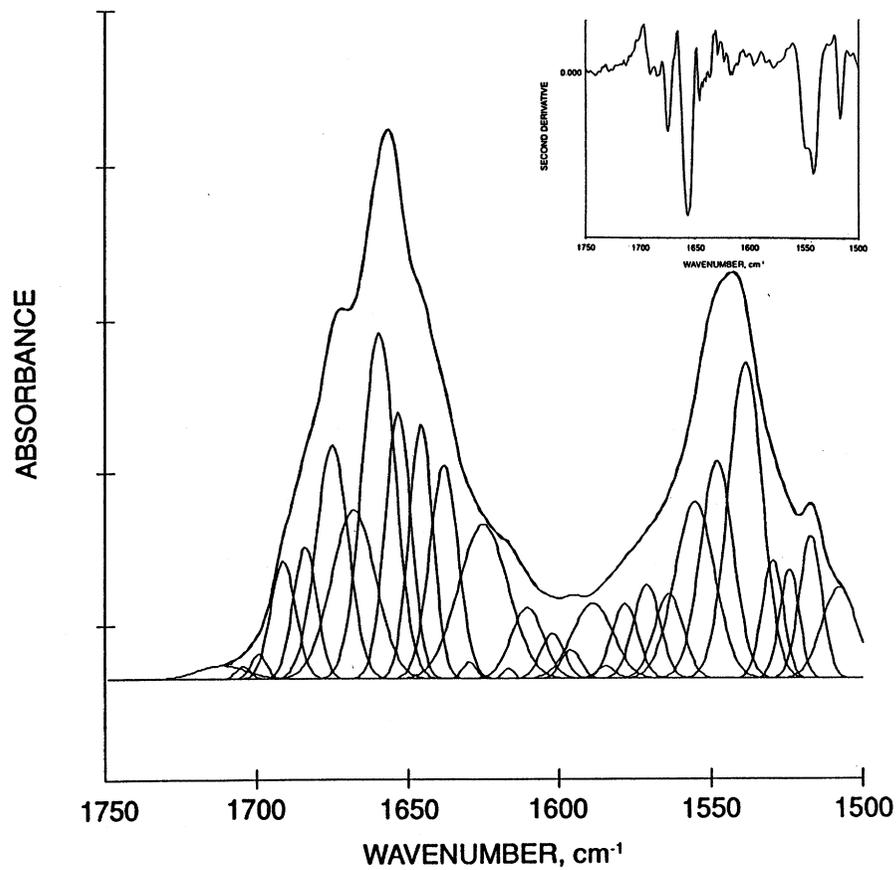
Quantitative criteria to insure correct deconvolution are: (1) correlation of all band assignments with the second derivative peaks; (2) agreement of calculated and experimental baselines; (3) a root-mean-square value of the fit  $\leq$  instrumental noise; (4) a successful fit to the original spectrum of the model using fixed frequencies found by fitting the FD spectrum. In practice, attainment of these criteria may require several cycles of FD and regression, until an optimal fit is achieved. Criterion 4 involves using the results of the regression analysis of the FD spectrum (Figure 2) to provide the number of bands and their frequencies, which are then fixed in a model to perform a nonlinear regression analysis of the original spectrum.

The final fit to the lysozyme FTIR spectrum is shown in Figure 1 with its 28 component peaks. The inset (Figure 1) shows the residuals of the regression are reasonably random, indicating the model is a reliable fit to the data. Calculated relative areas under the component bands of the original spectrum are in good agreement with those calculated from results of the regression analysis of the FD spectrum.

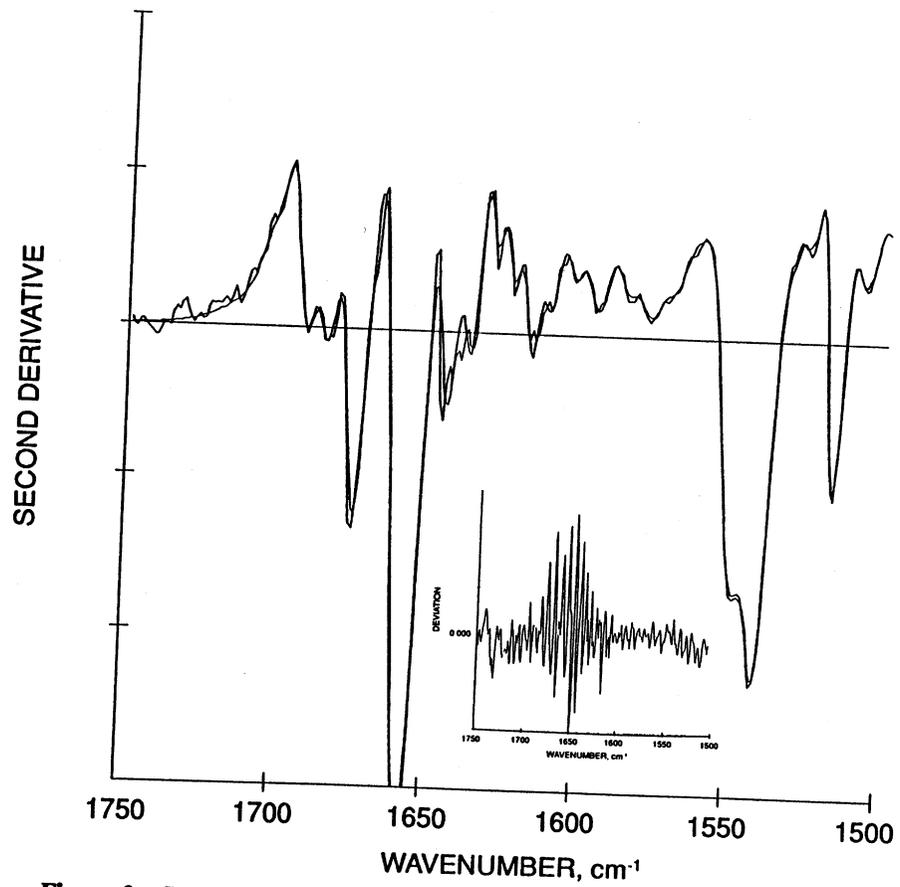
Further validation of the calculated components of the amide I and II bands can be obtained by mathematical comparison of the second derivative FTIR spectrum obtained from the original with the calculated second derivative obtained from the model fit. The results of such a comparison for lysozyme are shown in Figure 3, where the jagged line represents the fitted model and the smooth line the experimental data. The inset of this figure shows a reasonable pseudo-random residual plot which further establishes the reliability of this methodology for quantitatively resolving FTIR results of proteins into their individual component absorption bands.

## Results

**Rational for Deconvoluting into Component Gaussian Peaks.** We want to note some problems in reported protein 2° structure results. First, some of the groups still using factor analysis have reported protein structures within their database which do not add up to 100% structure (10-11); others, however, do (12). Second, the same calculated 2° structure from X-ray crystallography was used whether the experimental method was CD, FTIR or VCD. This assumption may have serious problems since the theoretical parameters



**Figure 2.** Fourier deconvolution of FTIR spectrum of lysozyme in Fig. 1. Lines on outer envelope and individual component peaks underneath were found by regression analysis as described in text. Double line is connected experimental data. Insert shows second derivative of original spectrum.



**Figure 3.** Second derivative FTIR spectrum of amide I and II bands of lysozyme in aqueous solution. Smooth line is connected experimental data. Jagged line on outer envelope is the results of final regression analysis as described in text. Insert shows plot of connected residuals between calculated and experimental second derivative results vs. frequency.

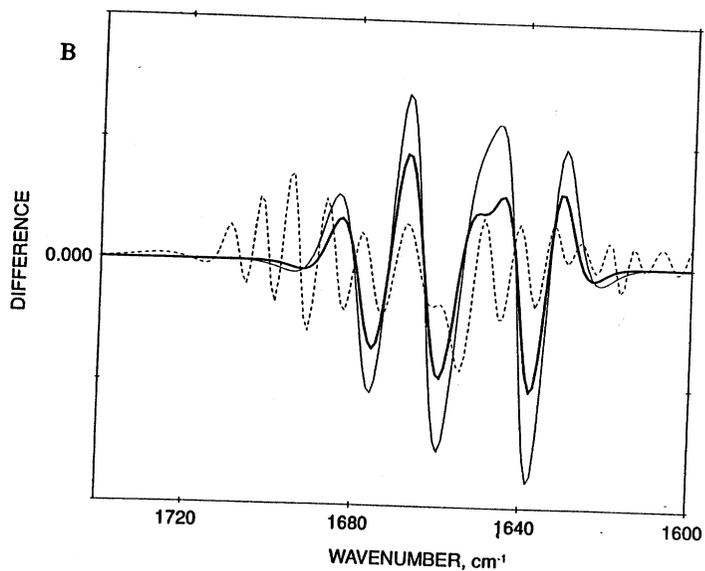
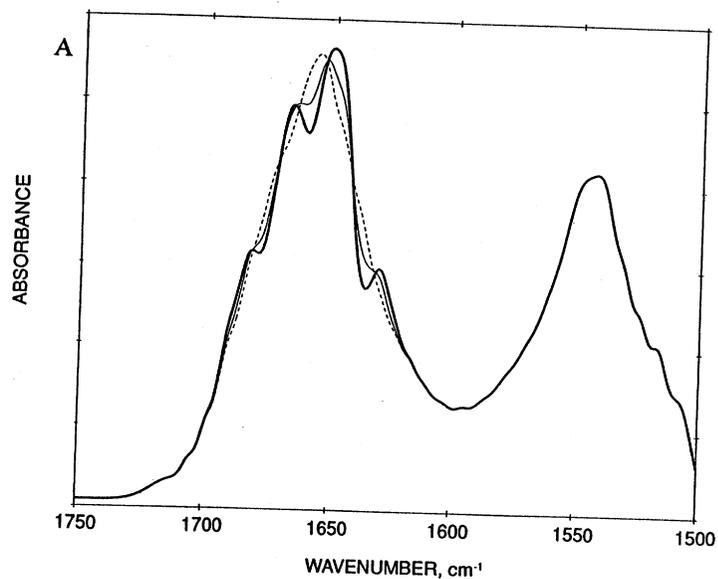
measured by these methods all differ: CD measures  $\text{Im} \langle r r \times P \rangle$ , for electronic transitions; VCD measures  $\text{Im} \text{ too}$ , but for vibrational transitions; and FTIR measures the transition dipole  $\langle r \rangle$ , for vibrational transitions. Third, changes in the shape and position of the spectra, for which factor analysis relies, may not reflect any structural changes in a protein. To illustrate these problems we shall use the results obtained from the analysis of the lysozyme spectrum.

Using the component Gaussian bands calculated in the analysis of the original FTIR spectrum (Figure 1) of lysozyme, we have calculated a theoretical FTIR spectrum (see Figure 4A, dashed line). We also calculated an altered spectra by changing the shape of the 1676, 1659 and 1638  $\text{cm}^{-1}$  bands (solid line in Figure 5A). The heights of the respective bands were divided by a factor of 1.2 while the half-width at half-height was multiplied by 1.2. This results in a different overall shape of the calculated amide I curve, while the resulting fractional areas of these bands remain constant. In this study, we shall refer to this modified theoretical spectrum as envelope 1.2 and the exact theoretical spectrum of lysozyme as the exact envelope.

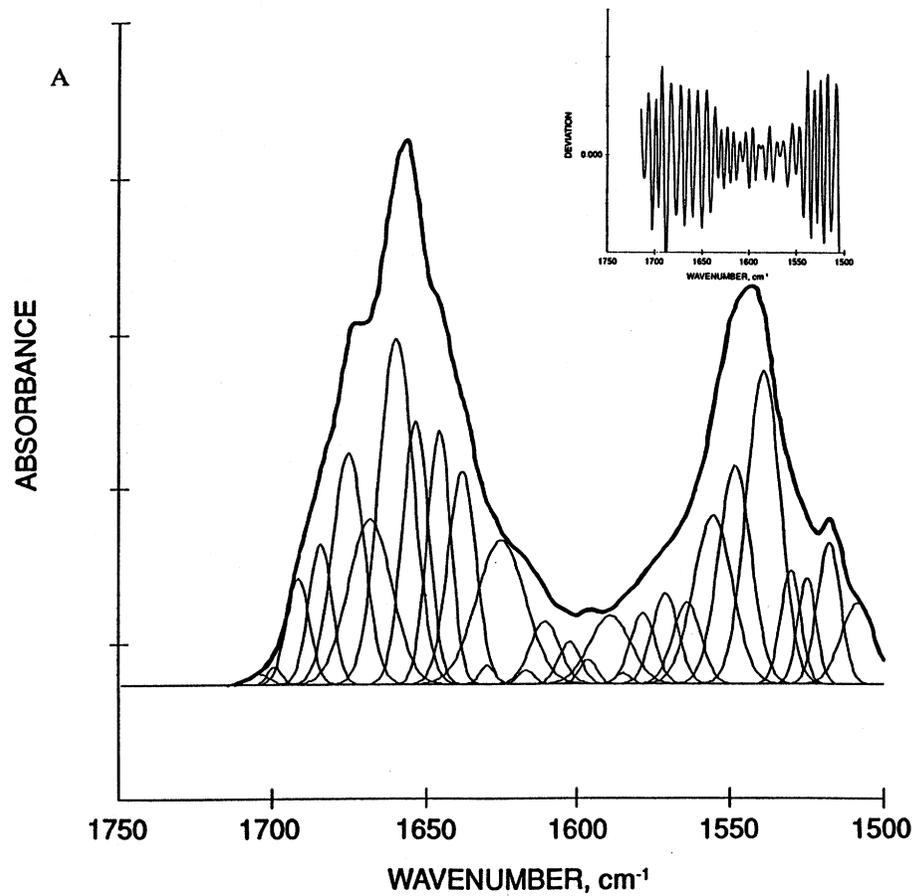
In addition, an envelope of 1.5 was also constructed and is presented in Figure 4A (double solid line). The dashed line and the solid lines in Figure 4A represent the exact and the 1.2 envelope, respectively. It can be seen in Figure 4A that as the factor is increased from 1.0 (for the exact), to 1.2 and 1.5, dramatic changes occur in the amide I envelope. Here, with a factor of 1.0 only one band appears in the amide I region where with factors of 1.2 and 1.5 four bands appear, three of which are near the affected 1676, 1659 and 1638  $\text{cm}^{-1}$  bands. The resolution of these three bands increases with an increasing factor. This overall shape change however, does not affect the calculated areas of the peaks. The effect on the area was corroborated by fitting the theoretical curves to a sum of 29 peaks via a nonlinear regression analysis. The calculated areas of the absorption bands at 1676, 1659, and 1638  $\text{cm}^{-1}$  remained constant. Thus factor analysis would reveal changes in conformation while regression analysis would not.

Since FTIR has an extremely large signal-to-noise ratio and yields very precise spectra, it is capable of seeing very small conformational changes in proteins as a function of varying environmental conditions. It has been suggested (5) that the difference spectra between FTIR second derivative spectra can at certain frequencies reflect small  $2^\circ$  structural changes due to changes in environmental conditions. However, care must be taken using this methodology for several reasons. The first reason being that the areas of the two amide I envelopes must be exactly the same and it is never clear where the amide I envelope ends at the low end of the frequency range. The other reason will be discussed below using the results presented in Figure 4.

Figure 4B shows the calculated second derivative spectrum of the exact envelope of lysozyme as a dashed line. The double and single line in Figure 4B represent the difference between the calculated second derivative spectra of the exact envelope and the 1.2 as well as the 1.5 envelope, respectively. It can



**Figure 4.** Theoretical spectrum from analysis of lysozyme results in Figure 1A. dashed line, exact curve from component bands of Figure 1; solid line, height decreased and width-at-half-height increased by a factor of 1.2 for 1676, 1959 and 1638  $\text{cm}^{-1}$  bands; double line, same as single line but with a factor of 1.5. B. dashed line, calculated second derivative of exact curve in Figure 4A dashed line; solid line, difference between calculated second derivative of 4A dashed line and 4B dashed line; double line, difference between calculated second derivative of 4A single line and 4B dashed line.



**Figure 5.** Best fit for Fourier deconvoluted lysozyme FTIR spectrum using non-linear regression analysis, half width at half height,  $W$ , of  $9 \text{ cm}^{-1}$  and a resolution enhancement factor of 2.5: single lines are experimental points, double lines are theoretical sum of component peaks shown in single line. A. for 29 peaks and B. for 28 peaks. Insets are plots of residuals between individual fits and experimental values.

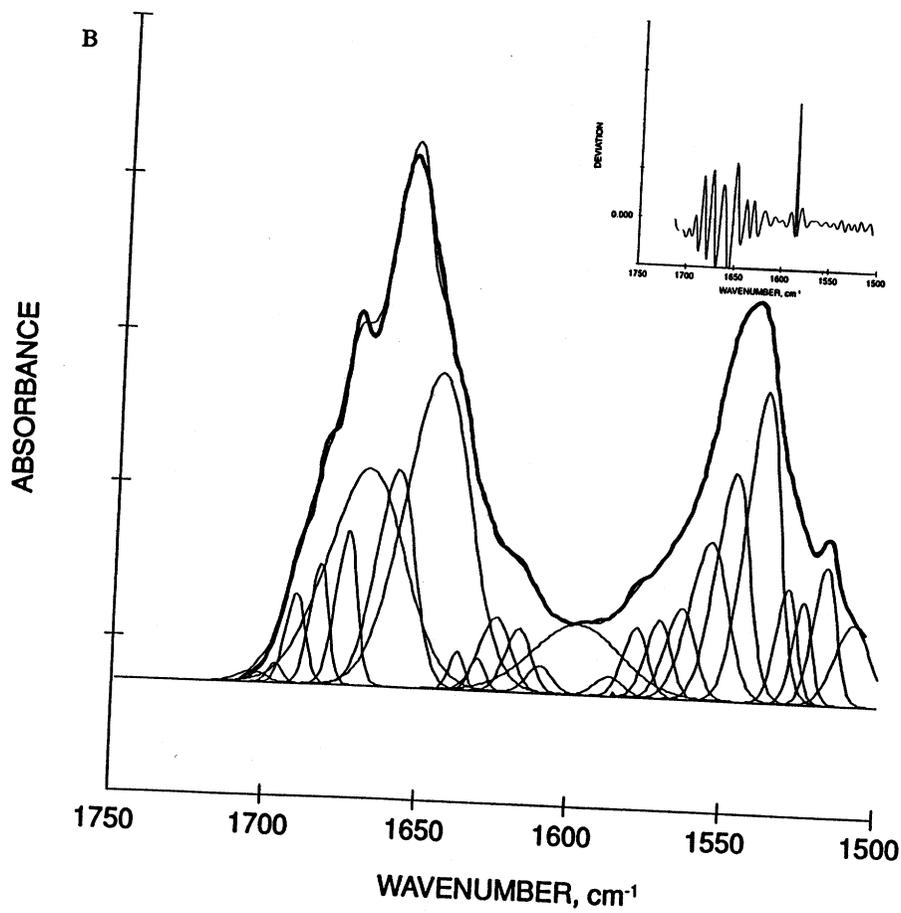


Figure 5. Continued.

easily be seen that the difference spectra in the amide I region (double and single line and in Figure 4B) are larger in magnitude than the exact calculated second derivative envelope of lysozyme. Such large differences may be easily interpreted as substantial conformation changes but, in reality, no change in area of the 1676, 1659, and 1638  $\text{cm}^{-1}$  bands has occurred — only the distribution of their Gaussian envelopes has changed. Therefore, second derivative difference spectra are not valid methods to study protein conformational changes.

The only sure methodology is to deconvolve the FTIR spectrum into its component Gaussian bands. But to insure accurate deconvolving, two important aspects of the methodology must be considered: 1, using the proper number of bands in the calculations; and 2, using both the amide I and II envelopes in the calculations. Using too few bands and not considering the amide II envelope could lead to serious errors and misinterpretations.

**Rational for the Parameters of the Non Linear Regression.** Controversy exists among researchers concerning the deconvolving of the FTIR amide I and II into their component bands. The determination of the number of bands and the character of the bands (whether they are of pure Gaussian or a combination of Gaussian and Lorentzian character) requires careful consideration. And, when using FD to ascertain the number of bands, the magnitude of applied FD variables must be justified.

Traditional FTIR experiments were performed in  $\text{D}_2\text{O}$  where no amide II envelope exists, and only the amide I envelope was deconvolved into its component bands. So, when experiments were eventually performed in  $\text{H}_2\text{O}$ , the amide II band while present was not always deconvolved. Analysts then also routinely applied conservatively low values for the resolution enhancement factor (REF) with a large half-width at half-height value of 13  $\text{cm}^{-1}$ , in order to avoid overdeconvolution of the spectrum. And then in the next step, used only the smallest number of component peaks for the nonlinear regression analysis.

Their rational was to avoid the possibility of distorting the experimental spectrum. However, no studies analyzing the same spectrum using nonlinear regression analysis with varying FD parameters have as yet been performed, thus the parameter limits before causing distortion are not known. Using higher REFs and narrower half-widths during FD increases the number of component bands. 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.

While nonlinear regression analysis using an increased number of component bands is more difficult and more time consuming (and cannot be easily performed on microcomputers), the correct number of peaks must be obtained to insure the correct assignments of the 2° structure of the protein spectra. If one band is used when two really exist, then a larger amount of disordered, helical or extended structure could be calculated due to an incorrect assignment. With our methodology we use the maximum number of component bands to fit the theoretical curve to the experimental data, that yields the

lowest root-mean-square value. We also fit the amide II along with the amide I envelope, and we allow a zero slope baseline to float to a calculated value. However, we still will maintain a low REF for FD of the spectra of 2.5 with a half-width at half-height of  $9 \text{ cm}^{-1}$  for all spectra. Furthermore, we only use pure Gaussian component bands for fits to FD, original and calculated second derivative amide I and II spectra. Attempts to use a constant fraction of Lorentzian character and optimize the value of the fraction of Lorentzian character were unsuccessful, resulting in non convergence of the Gauss-Newton nonlinear regression program. It should also be noted that during our calculations, the use of too many bands resulted in the heights of the excess bands approaching a zero or negative value. Thus, the excuse that excess bands 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 analysis.

The effects of choosing too few component bands are discussed as follows. Figure 5A shows the fit of the Fourier deconvoluted amide I and II envelopes of lysozyme with 29 component Gaussian bands. The fit is excellent with no discernable difference between the resulted theoretical and experimental curves. However, when the band at  $1659 \text{ cm}^{-1}$  is removed and only 28 peaks are utilized, non linear regression analysis yields a poor fit with some component bands becoming much broader than others (see Figure 5B). Also, it can be seen that the fit to the amide II bands is affected even though no band was removed from that range of frequencies. In addition, as seen in Table II, the RMS for 28 peaks is six times larger than for 29 peaks, i.e. 0.00157 verses 0.000251, respectively.

This behavior is also seen in the fits of 25 peaks verses 24 peaks for trypsin, 24 peaks verses 23 peaks for elastase and 17 peaks verses 16 peaks for myoglobin. In all cases, the fits using one less peak in the range of  $1659 \text{ cm}^{-1}$  are extremely poor, with some extremely broad peaks in the amide I region which in turn influences the fit of the amide II region. Also the RMS of the poor fits are on the order of a factor of 3 to 6 times higher than the best fits (see Table I). It should be noted, that FD causes the component bands to have almost equal half-width at half-heights. As described by Byler and Susi (3), the apperance of broad component bands in the results of nonlinear regression analysis was considered unacceptable. Despite the improved fits, the number of peaks which exist under the amide I or II envelopes should be based upon more theoretical concepts to prove the above hypothesis.

**Table I. Influence of number of Gaussian peaks on RMS**

Protein	N	RMS	N <sub>2</sub>	RMS <sub>2</sub>
Lysozyme	29	0.000251	28	0.00157
Trypsin	25	0.000362	24	0.00140
Elastase	24	0.000624	23	0.00174
Myoglobin	17	0.000306	16	0.00154

**Table II. % Extended content of globular proteins**

	<b>FTIR</b>	<b>Ps</b>	<b>R</b>	<b>BS</b>	<b>XBS</b>
Hemoglobin	7.7	24.5	8.2	25	0
Myoglobin	10.4	10.5	5.9	24	0
Cytochrome C	10.5	19.4	11.6	34	10
Lysozyme	29.6	25.6	30.5	21	19,16
Ribonuclease	41.3	23.4	37.5	50	46,40
Papain	22.2	39.6	19.8	32	29
PTI	31.8	25.9	29.5	52	50
$\alpha$ -Chymotrypsin	38.8	34.2	37.9	51	49,34
Trypsin	40.6	38.1	39.6	55	56
Elastase	36.6	43.8	33.6	45	47,52
Carbonic Anhydrase	41.7	26.7	36.0	49	45,40
$\beta$ -Lactoglobulin	44.5	18.5	44.2	50	-
CON A	44.2	32.9	44.5	60	60,51
Oxytocin	0	0	0	-	-

**FTIR** = Average error  $\pm 1.6 \text{ cm}^{-1}$ ; **Ps** = Modeling predicted % sheet structure (26); **R** = Ramachandran; **BS**, **XBS** = Byler, D.M. and Susi, H. (4), Data and X-ray respectively.

In recent articles by Torii and Tasumi (13-15), the theoretical FTIR amide I spectrum was calculated using the three dimensional structure of lysozyme from X-ray crystallography and a Gaussian envelope of each peptide oscillator with a half-width at half-height of  $3.0 \text{ cm}^{-1}$ . Their calculations were used to qualitatively compare theoretical spectra of several proteins with their experimental counterparts in  $\text{D}_2\text{O}$ , so the force constants used were optimized to agree with  $\text{D}_2\text{O}$  and not  $\text{H}_2\text{O}$  results. For this reason we cannot directly 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 with our experimental FD spectrum. Here, it must be emphasized that the theoretical spectrum must contain at least but not less than the same number of bands in our experimentally analyzed spectrum.

Figure 6A shows deconvolved theoretical FTIR spectra from Torii and Tasumi (13) with the best fit of the sum of 14 Gaussian bands. Attempts to use less resulted in poor fits while the addition of more peaks caused the height of the extra bands to approach a zero or negative value. The experimental FD FTIR spectrum of lysozyme using an REF of 3.8 (a value considered much too large by most investigators) is shown in Figure 6B. Here, the amide I region is fit to the sum of 14 Gaussian peaks with success. A low RMS value (to within 0.1%) and a pseudo random deviation pattern was obtained with the 14 band fit. No additional bands could be successfully added to the 14 band fit. In addition, none of these 14 bands had extremely broad half-widths at half-height just as in the calculated bands of Figure 1 which used only 10 bands for the amide I region (comparing same region, 1690-1620  $\text{cm}^{-1}$  only 10 component bands).

Attempts to fit the 14 bands to experimental data using a universally more accepted REF value of 2.5 ended with some bands, especially those with the highest and lowest frequency, to become unacceptably broad (see Figure 6C).

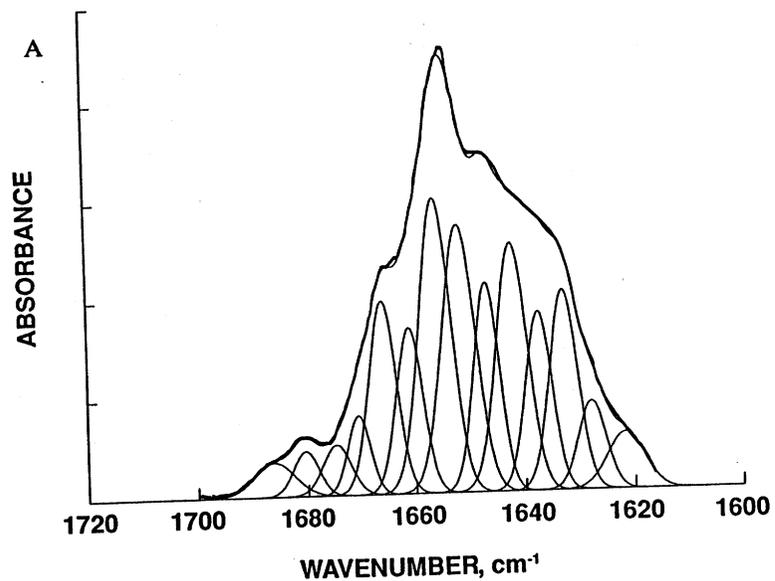


Figure 6. Best fits by non-linear regression analysis using amide I band of lysozyme. A: from theoretical calculations of Torii and Tasumi (13-15). B: Fourier deconvoluted lysozyme spectra using a resolution enhancement factor of 3.8. C: same as 6B with a factor of 2.5. *Continued on next page.*

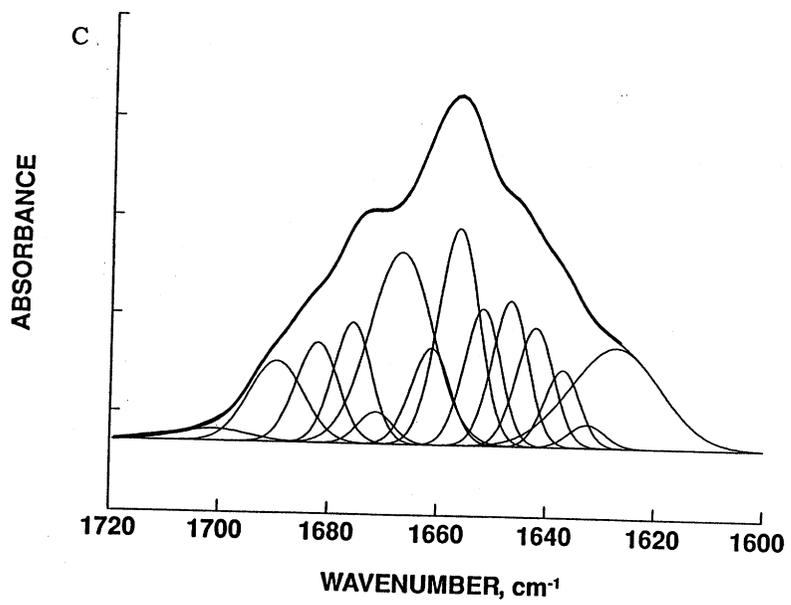
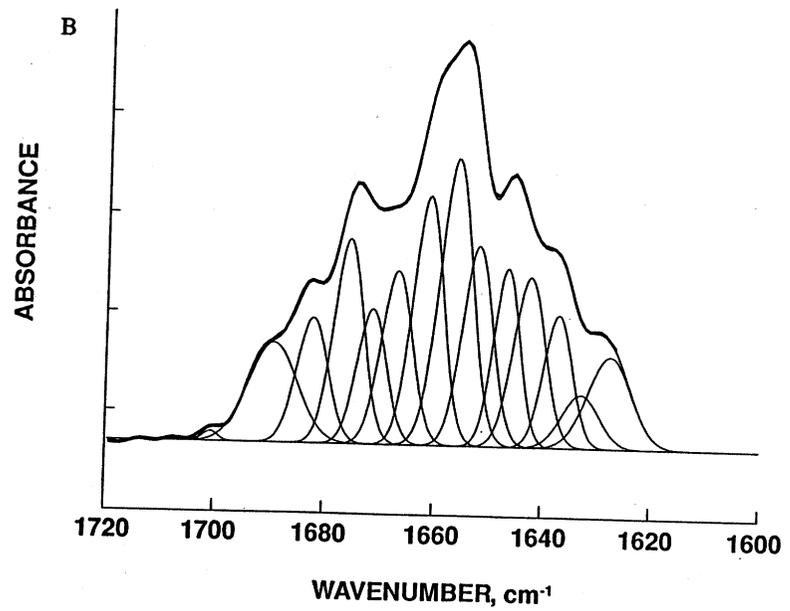


Figure 6. Continued.

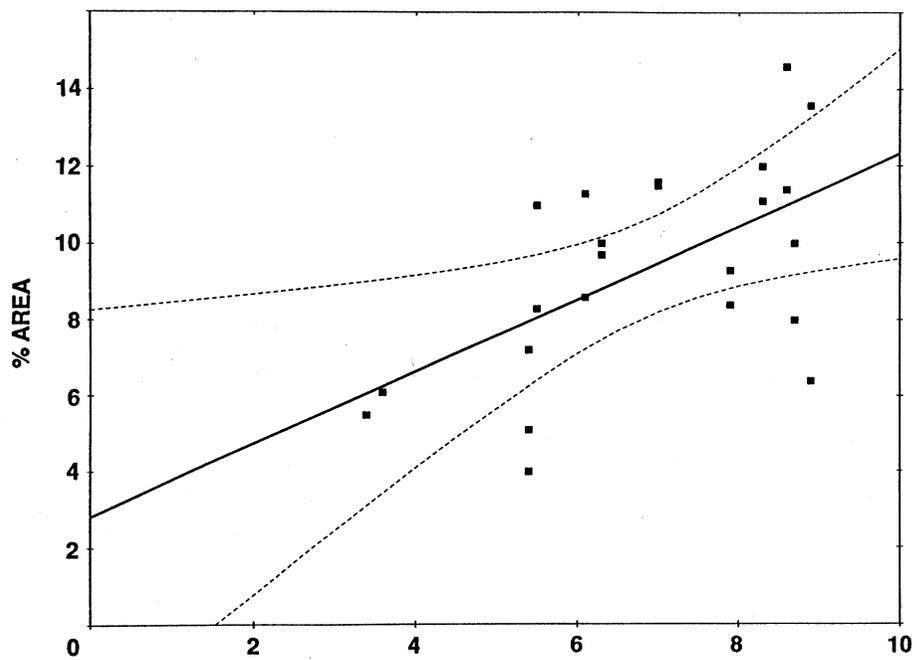
However, if the amide II band is fit simultaneously with the amide I (using REF of 2.5 and fitting 29 bands), inordinately large values of the half-width at half-height are not present (see Figure 2). Hence, it is extremely important to simultaneously deconvolve the amide I and II envelopes if the FD spectra is generated using lower REF values (such as 2-3). Calculations excluding the amide II envelope may lead to large errors in the estimated secondary structures assigned to a protein.

Even though the calculated spectra of Torii and Tasumi are based on D<sub>2</sub>O force constants, we will attempt to compare the 2° structure calculated from the theoretical with the experimental spectra for all the reported proteins in a separate paper. It also, should be noted that the experimental spectra analyzed with 29 peaks (Figure 2), only has 10 peaks in the amide I region. The four extra peaks determined from the theoretical spectra are well within the strand and turn region and could easily be summed to obtain the total turn and strand structure.

## Discussion

**Contribution of ASN and GLN Side Chain.** Recently, Venyaminov and Kalnin (16) 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 asparagine (ASN) and glutamine (GLN) exist within the amide I envelope. In subsequent articles (12,17), they attempted to eliminate all side chain contributions to the amide I and II by subtracting, on a molar basis, the amino acids from the amide I and II envelopes of proteins. They assumed the absorptivity of amino acids and proteins were equivalent, and they found a band at 1668 cm<sup>-1</sup> which is invariant to changes in environmental conditions for both ASN and GLN. They have assigned this band as the C=O stretch of the GLN and ASN side chains. In addition, they show in their figures a small broad band at 1650 cm<sup>-1</sup> in all of their GLN and ASN studies, which also maintained the same frequency as the solvent or pH were varied. Since this band also exists in their side chain analysis of arginine, we expect that this band may be a C-N deformation of GLN and ASN. Now we will attempt to ascertain if any of the amide I component Gaussian bands calculated for the FTIR of the 14 proteins correlate with the amount of GLN and ASN present in these proteins.

Figure 7 is a plot of the area % of the 1651 and 1667 cm<sup>-1</sup> bands verses the % GLN and ASN for the proteins listed in Table II. The % areas of the 1651 cm<sup>-1</sup> for the first three predominately helical proteins, i.e. hemoglobin, myoglobin and cytochrome C, are eliminated from this analysis. The best fit line and the 95% confidence lines (dashed lines) were calculated and are also shown in Figure 7. The results of the linear regression analysis yields an intercept value of 2.83 (S.E. = 2.06,  $\sigma$  = 0.184) and a slope of 0.948 (S.E. = 0.292,  $\sigma$  = 0.00388). Thus, it appears the intercept value could statistically



**Figure 7.** Linear regression analysis for area percent of 1667 and 1651  $\text{cm}^{-1}$  bands versus present GLN and ASN in Protein for the 14 proteins listed in Tables II-IV. Filled squares: data points; dashed lines: confidence curves at level 0.95; double line: best straight line using linear regression analysis with no weighted points.

have a zero value, and the slope a value of near unity. While the analysis suggests a correlation, it should be viewed as an assumption. Only when the analysis of at least 50 proteins yields similar results would this assumption be considered proven.

For this report, we will assign the 1667 and 1651  $\text{cm}^{-1}$  bands to GLN and ASN side chain modes. Most likely 1651  $\text{cm}^{-1}$  is a C-N deformation and 1667  $\text{cm}^{-1}$  is a 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 percentages of areas are less than the percent 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 renormalized to unity.

**Ramachandan Analysis.** Of paramount importance to obtaining 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 the amount of  $\alpha$ -helix, turn and extended conformation is important, but the length of the helix and extended conformation as well as whether internal backbone hydrogen bonding exists may be relevant descriptors for correlation with the % areas of the component bands in the amide I region. Until recently, researchers in this field used the values provided in the Brookhaven Protein Data Bank. The values depended 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. (12) has recently subdivided both the helices and sheets into hydrogen bonded and non hydrogen bonded conformations, which along with the turn and all other conformations, forms a basis of six instead of four conformations. Their calculations, however, are correlated with FTIR results using factor analysis instead of nonlinear regression analysis. In their analysis normalized structures (i.e. the total conformation of an individual protein adds up to 100%) show reasonable correlation with FTIR results. The same good correlations were not obtained by the use of factorial analysis (11,18) in which structure normalization was ignored and only four conformations were considered.

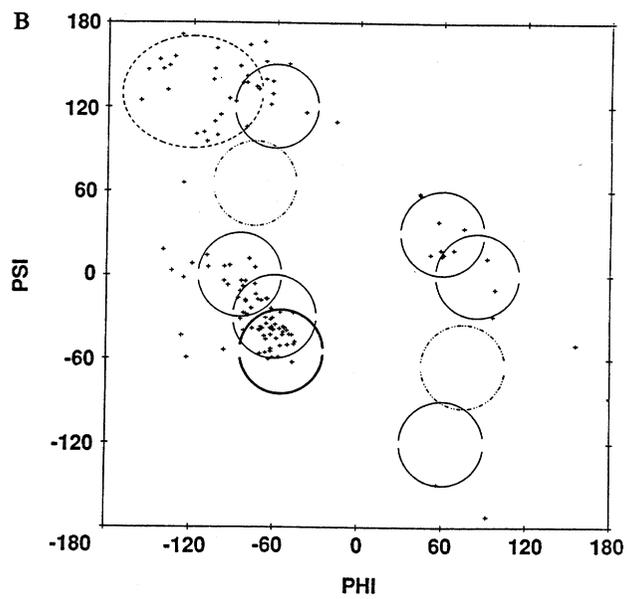
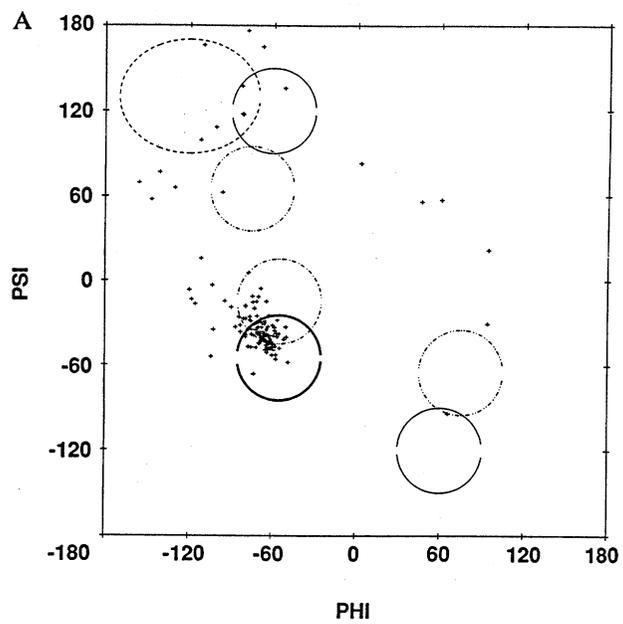
An algorithm developed recently in Liebman's laboratory (19,20) shows good agreement with FTIR results when deconvolving the amide I into component bands (19,20). However, these experiments were performed only for serine proteases in  $\text{D}_2\text{O}$ . While this method appears to be promising it is still in its infancy, more correlations between experimental (in  $\text{H}_2\text{O}$ ) and calculated conformations must be reported.

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 that the choice of Ramachandran analysis does not in any way imply that the transitional dipolar coupling mechanism of Krimm et al. (21,22) and Torii et al. (13-15) may not apply to the vibrational spectroscopy of proteins. We believe that this mechanism is correct.

Here, we use the Ramachandran plot only as a method for correlating reported fractional values. If discrepancies exist we shall use other molecular modeling techniques available in the Sybyl Molecular Modeling software (such as inspection by ribbon, hydrogen bonding) 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 Figures 8A and 8B the Ramachandran plots are given for myoglobin and lysozyme, respectively. Figure 8A shows a template Ramachandran with theoretical curves for predominately helical proteins. The dashed 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.

Figure 8B uses a template plot which is more consistent with proteins containing a large amount of turn conformation. Here the dashed lines and double lines represent the sheet and  $\alpha$ -helix region, respectively, as in Figure 8A. Adjacent to the sheet region in 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 lined 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 positive 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. Shown in Figure 8B (as "+") are the calculated phi, phi results from the Sybyl modeling program for 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 were calculated, the results along with the corresponding experimental values are presented in Tables II, III, and IV for strand, helix, and turn plus irregular, respectively. The results for each conformation will be discussed in separate sections for the remainder of this report.



**Figure 8.** Ramachandran plot from X-ray crystallographic structure of A: myoglobin and B: lysozyme.

**Table III. % Helix content of globular proteins**

	$\alpha$				3/10, Bent Strand	
	FTIR	P <sub>H</sub>	R	BS	FTIR	R
Hemoglobin	76.7	49.7	81.6	74	1.2	-
Myoglobin	76.4	73.9	78.1	76	1.9	-
Cytochrome C	46.2	36.9	44.6	51	2.2	4.0 BE
Lysozyme	25.6	20.9	27.1	41	4.2	5.5 BE
Ribonuclease	10.0	29.8	18.0	21	10.3	8.8
Papain	15.2	11.3	18.8	27	17.4	15.6
PTI	4.2	8.6	12.0	10	0	6.9 BE
$\alpha$ -Chymotrypsin	11.9	16.7	12.4	12	3.5	6.8
Trypsin	17.4	12.1	10.4	16	0	8.3
Elastase	14.1	10.0	9.7		0	8.2
Carbonic Anhydrase	14.2	22.4	14.2	13	1.7	7.1
$\beta$ -Lactoglobulin	15.8	63.0	5.0	10	3.1	9.2 BE
CON A	13.6	16.9	1.7	-	2.2	8.8 BE
Oxytocin	0	0	0	-	-	-

FTIR = Average error  $\alpha$  helix  $\pm 0.8 \text{ cm}^{-1}$ ; P<sub>H</sub> = Modeling predicted %  $\alpha$ -helix (26); R = Ramachandran; BE = Bent strand; BS = Byler, D.M. and Susi, H. (4).

**Table IV. % Non-periodic content of globular proteins**

	Turn or Twisted Strand			Irregular		
	FTIR	P <sub>T</sub>	R	FTIR	P <sub>I</sub>	R
Hemoglobin	4.3	7.5	7.5	10.1	18.4	10.8
Myoglobin	2.3	6.5	2.1	8.8	9.2	14.6
Cytochrome C	37.9	20.4	35.8	3.1	23.3	3.1
Lysozyme	28.1	42.6	26.6	12.5	10.9	9.6
Ribonuclease	24.2	40.3	26.0	14.2	6.5	9.7
Papain	43.3	35.4	39.6	2.0	13.7	6.2
PTI	60.5	56.9	49.6	3.5	8.6	8.9
$\alpha$ -Chymotrypsin	25.2	31.1	22.8	20.6	18.0	20.1
Trypsin	26.3	34.5	28.8	15.7	15.2	16.7
Elastase	31.8	26.7	33.6	17.4	19.6	16.0
Carbonic Anhydrase	22.3	22.4	26.0	20.1	28.6	16.3
$\beta$ -Lactoglobulin	17.3	18.7	21.2	19.3	14.8	20.4
CON A	22.2	17.3	22.7	17.7	32.9	22.3
Oxytocin	100.0	100.0	100.0	0	0	0

FTIR = Average error: turn or twisted strand  $\pm 1.4 \text{ cm}^{-1}$ , loop  $\pm 0.8 \text{ cm}^{-1}$ ;  
P<sub>T</sub> = Modeling predicted % turn or twisted strand (26); R = Ramachandran;  
P<sub>I</sub> = Modeling predicted % irregular.

**Extend (Strand) Content.** Table II shows a comparison between the experimentally determined global 2° structure of the extended conformation (column 2, heading FTIR) along with values calculated by the composite Ramachandran analysis (column 4, heading R). The amount of extended conformation was determined from FTIR amide I results by summing the component bands from 1638 to 1623  $\text{cm}^{-1}$ . Bands at frequencies lower than 1623  $\text{cm}^{-1}$  were closer to 1615  $\text{cm}^{-1}$  and were presumed to be caused by unprotected side chain carboxyl groups (21). Note that we assumed the bands (1638-1623  $\text{cm}^{-1}$ ) quantitate not only the amount of sheet conformation but also extended or strand non hydrogen bonded structures as did Prestrelski et al. (18-19). For proper comparison, we summed all 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 three dimensional Ramachandran (not shown).

For comparison, the fifth and six columns of Table II, labeled BS and XBS are the experimentally determined (FTIR) and calculated (X-ray crystallographic) conformation results of Byler and Susi (3) where the FTIR results were determined in D<sub>2</sub>O.

As seen in Table II, comparison between the experimental (FTIR) and calculated (R) extended conformation is excellent. Only in myoglobin, a high helical protein, is the deviation greater than 4%. In fact, the average deviation between the experimental (FTIR) and calculated (R) strand structure for these 14 proteins is 1.8%. This value is lower than we had hoped, since Mantsch (23) had recently reported that FTIR cannot determine the absolute value of a conformation to within 2% and the average deviation for the 12 proteins in the Byler and Susi report (BS vs R) is 9.2%. Inspection of their results in Table II shows that in almost all instances their values are much higher than those in this study. The discrepancy could either be due to the use of D<sub>2</sub>O which will not exchange 100% of all the protein protons and may cause an increase in hydrophobic interactions (24), or to the use of two few bands which leads to serious mis-assignments.

Next, we compared our FTIR results for % extended in trypsin,  $\alpha$ -chymotrypsin and elastase with those reported by Liebman's group (19-20). We obtained FTIR values of 40.6, 38.8 and 36.6% for trypsin,  $\alpha$ -chymotrypsin and elastase, respectively. Liebman reported FTIR values (determined in D<sub>2</sub>O) of 39, 45 and 46, respectively, and calculated values (using their own algorithms) of 42, 42 and 47, respectively. While all experimental values for trypsin and  $\alpha$ -chymotrypsin agree equally with their calculated values, our elastase value of 36.6% agrees far better with their calculated value of 37 than their experimental (in D<sub>2</sub>O) value of 46%. This adds further support for our methodology (spectra in H<sub>2</sub>O) and the supposition that the amount of GLN and ASN side chain residues must be subtracted for the 1667 and 1651 cm<sup>-1</sup> component amide I bands.

Shown in column 3 of Table II with a heading of P<sub>S</sub> is the predicted amount of sheet structure using a secondary structure sequence based prediction algorithm by Garnier et al. (25). Other algorithms were attempted but this method yielded the most comparable results. In Tables III and IV, the P<sub>H</sub>, P<sub>T</sub>, and P<sub>I</sub> for the amount of  $\alpha$ -helix, turn and irregular conformation calculated by this algorithm are also presented.

**Helical Content.** Table III lists the calculated and experimental results for the  $\alpha$ -helix (column 1-3), 3/10 helix (column 5, 6), and those reported by Byler and Susi (3) — BS (column 4). Here the calculated average deviation between our experimental (from the 1659 and 1651 cm<sup>-1</sup> bands) and calculated % helix is twice as high (3.6%) as for the extended structure (1.8%), but is still

acceptable. 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 proteases (i.e. Pancreatic Trypsin Inhibitor (PTI), trypsin, and  $\alpha$ -chymotrypsin), concanavalin A (CONA) 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 also yield points in the lower left quadrant which is normally considered a forbidden region. The Ramachandran plots for the serine proteases all contain some  $\phi$ ,  $\psi$  angles in the region. Hence, the discrepancy in the experimental and helical constant for concanavalin A,  $\beta$ -lactoglobulin and perhaps the serine proteases may be caused by  $\beta$ -barrel which could have bands at  $1658 \pm 2 \text{ cm}^{-1}$ . More experimental and theoretical studies must be performed before this hypothesis can be concluded.

But ribonuclease contains no  $\beta$ -barrel and inspection of the work by Kalnin, et al. (12) may provide an answer. In this study, the  $\alpha$ -helix was subdivided into an ordered and unordered class as was the sheet structure. They calculate values of 13% and 10% for their ordered and unordered  $\alpha$ -helix conformation and find 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  $1667 \text{ cm}^{-1}$  which we have assigned as a 3/10 helix, in accordance with the results of Krimm and Bandekar (21). 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. (12) 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 a result of its structure which our Ramachandran analysis could not adequately calculate.

It should be stated at this time that the  $1676 \text{ cm}^{-1}$  band was also summed along with the excess area for the  $1651 \text{ cm}^{-1}$  as well as the  $1658 \text{ cm}^{-1}$  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%. A value of 63% was also calculated as the amount of unordered helix content in myoglobin by Kalnin et al. (12). 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 (21) as a turn may be reflective of a type III  $\beta$ -turn. Such a turn would have  $\phi$ ,  $\psi$  values overlapping the  $\alpha$ -helical region of a Ramachandran plot (see Figure 8B). 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 (13-15) 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. But 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 III as a 3/10 helix or bent strand i.e. a possible "a" ( $\gamma$ ) turn. These small values cannot be easily correlated with Ramachandran analysis and no firm assignments are made. 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. (20,21) 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 (21) 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. (19-20). Perhaps this discrepancy can be explained by the fact that their experiments were performed in  $\text{D}_2\text{O}$  rather than  $\text{H}_2\text{O}$ .

**Turn and Irregular Content.** Table IV 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-crystallographic structure of the 14 listed proteins. The turn content was determined from the sum of all amide I bands from  $1670\text{ cm}^{-1}$  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 agreements between the experimental and theoretical values were observed with average calculated deviations of 2.9% and 2.6% for the turn and irregular content, respectively. These values are well within the deviations observed in the strand and  $\alpha$ -helix content i.e. 1.8% and 3.6%. 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 the result of twisted sheets. Since no other proteins in this database exhibited phi and psi values in this region, we have concluded that the  $1676\text{ cm}^{-1}$  may also be assigned to a twisted strand. However, more studies must be performed before a definite assignment can be made.

## Conclusions

Calculation of the component 2<sup>o</sup> structural elements of the vibrational bands, i.e., approximately 25 Gaussian bands, was accomplished by fitting both the

**Table V. Secondary structure assignments [this study]**

---

1681 - 1695 cm <sup>-1</sup>	Turn I, I'
1673 - 1679 cm <sup>-1</sup>	Turn II, II', III, III', twisted sheet
1667 - 1669 cm <sup>-1</sup>	GLN (C=O) & ASN (C-O) side chain, 3/10 helix, bent strand
1657 - 1661 cm <sup>-1</sup>	$\alpha$ -Helix (A Band)
1651 - 1653 cm <sup>-1</sup>	GLN (N-H) & ASN (C-N) side chain, $\alpha$ -helix (E Band)
1643 - 1648 cm <sup>-1</sup>	Disordered, irregular, gentle loop
1622 - 1638 cm <sup>-1</sup>	Extended strand, rippled and pleated sheets

---

amide I and II bands using nonlinear regression analysis of: the Fourier deconvoluted spectra, the second derivative spectra, and the original spectrum. Fixed frequencies used in the original spectra analysis were obtained from both the FD spectra and 2nd derivative analyses. The criterion for acceptance of any analysis was that the fractional areas calculated from all three methods were in agreement. Results clearly show that 2° structural conformations determined in water were in better agreement with global 2° structure analysis of X-ray structures than the previously reported values determined in D<sub>2</sub>O. Also with H<sub>2</sub>O the types of turns can be correlated with the X-ray structure, and 2° structure elements can be calculated from the amide II band to be used for validation of amide I assignments. In addition, resolution of amide I spectra in H<sub>2</sub>O is greater than that in D<sub>2</sub>O. The deterioration of resolution of FTIR spectra in D<sub>2</sub>O results primarily from incomplete exchange of protein protons to deuterons. The results lay the foundation for the study of conformational changes in proteins induced by ligands, cosolutes or perhaps structural changes from site directed mutagenesis (9).

In this study, we have presented a method for analyzing the FTIR of proteins in water and determining their global 2° structure. Analysis of 14 proteins whose X-ray crystallographic structures are known, showed agreement between experimental and theoretical 2° structure content to within 4%. The bands which are assigned to these structures are shown in Table 5 along with their tentative structural assignments. While the results are excellent, 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 others investigators to adopt this methodology and add more information to increase this database above 14 proteins. In this laboratory, we too will continue to add to this database.

#### **Acknowledgments**

Reference to a brand or firm name does not constitute an endorsement by the U.S. Department of Agriculture over others of a similar nature not mentioned.

## Literature Cited

1. Townend, R.; Kumosinski, T. F.; Timasheff, S. N. *J. Biol. Chem.*, **1967**, *242*, 4538-4545.
2. Kauppinen, J. K.; Moffatt, D. J.; Mantsch, H. H., Cameron, D. G. *Appl. Spec.*, **1981**, *35*, 271-276.
3. Byler, D. M.; Susi, H. *Biopolymers*, **1986**, *25*, 469-487.
4. Susi, H.; Byler, D. M. *Methods. Enzymol.*, **1986**, *130*, 290-311.
5. Byler, D. M.; Farrell, Jr., H. M. *J. Dairy Science*, **1989**, *72*, 1719-1723.
6. Birke, S. S.; Dien, M. *Biochemistry*, **1992**, *31*, 450-455.
7. Cantor, C. R.; Schimmel, P. R. In *Biophysical Chemistry Part III: Techniques for the Study of Biological Structure and Function*, Freeman, W. H., Ed., **1980**, pp. 687-791.
8. Dong, A.; Huang, P.; Caughey, W. S. *Biochemistry*, **1990**, *29*, 3303-3308.
9. Dousseau, F.; Pezolet, M. *Biochemistry*, **1990**, *29*, 8771-8779.
10. Pancoska, P.; Yasui, S. C.; Keiderling, T. A. *Biochemistry*, **1991**, *30*, 5089-5103.
11. Pancoska, P.; Keiderling, T. A. *Biochemistry*, **1991**, *30*, 6885-6895.
12. Kalnin, N. N.; Baikalov, I. A.; Venyaminov, S. Y. *Biopolymers* **1990**, *30*, 1273-1280.
13. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *96*, 3379-3387.
14. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *97*, 86-91.
15. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *97*, 92-98.
16. Venyaminov, S. Y.; Kalnin, N. N. *Biopolymers*, **1990**, *30*, 1243-1258.
17. Venyaminov, S. Y.; Kalnin, N. N. *Biopolymers*, **1990**, *30*, 1259-1271.
18. Pancoska, P.; Wang, L.; Keiderling, T. A. *Protein Science*, **1993**, *2*, (in press).
19. Prestrelski, S. J.; Williams, A. L.; Liebman, M. N. *Proteins, Structure, Function and Genetics*, **1992**, *14*, 430-439.
20. Prestrelski, S. J.; Byler, D. M.; Liebman, M. N. *Proteins, Structure, Function and Genetics*, **1992**, *14*, 440-450.
21. Krimm, S.; Bandekar, J. *Adv. Protein Chem.*, **1986**, *38*, 181-364.
22. Krimm, S.; Bandekar, J. *Biopolymers*, **1980**, *19*, 1-29.
23. Surewicz, W. K., and Mantasch, H. H., and Chapman, D. *Biochemistry*, **1993**, *32*, 389-394.
24. Timasheff, S. N. In *Protides of the Biological Fluids, 20th Colloquium*, Peters, H. ed., **1973**, p. 511-519.
25. Garnier, J.; Osguthorpe, D.; Robson, B. *J. Mol. Biol.*, **1978**, *120*, 97-120.