A Distinct Utility of the Amide III Infrared Band for Secondary Structure Estimation of Aqueous Protein Solutions Using Partial Least Squares Methods[†]

Shuowei Cai and Bal Ram Singh*

Department of Chemistry and Biochemistry, University of Massachusetts Dartmouth, Dartmouth, Massachusetts 02747

Received June 13, 2003; Revised Manuscript Received December 5, 2003

ABSTRACT: Fourier transform infrared spectroscopy is becoming an increasingly important method to study protein secondary structure. The amide I region of the protein infrared spectrum is the widely used region, whereas the amide III region has been comparatively neglected due to its low signal. Since there is no water interference in the amide III region and, more importantly, the different secondary structures of proteins have more resolved differences in their amide III spectra, it is quite promising to use the amide III region to determine protein secondary structure. In our current study, a partial least squares (PLS) method was used to predict protein secondary structures from the protein IR spectra. The IR spectra of aqueous solutions of 16 different proteins of known crystal structure have been recorded, and the amide I, the amide III, and the amide I combined with the amide III region of these proteins were used to set up the calibration set for the PLS algorithm. Our results correlate quite well with the data from X-ray studies, and the prediction from the amide III region is better than that from amide I or combined amide I and amide III regions.

The three-dimensional structures of biological macromolecules and their physiological functions are intimately related. Proteins are thought to be the "second part of the genetic code", and they play a pivotal role in living cells. Their functions are responsible for a large number of biological processes, including the facilitation of metabolism, communication, transport, and maintenance of structural integrity. The diversity of the functions of proteins is only matched by the diversity of protein structure. The importance of understanding the structure−function relationship of proteins has underscored the need for a rapid, reliable, and sensitive probe of the molecular conformation of proteins. Such a method should allow study in aqueous media and also be suitable for use in measuring the effects of environmental factors, ligand binding, and other perturbations on molecular conformation. A variety of techniques have been applied to the elucidation of the three-dimensional structure of proteins, ranging from computer-aided prediction based on the sequence of the constituent amino acids and energetic consideration (*1*) to precise methods for identification of their molecular coordinates, such as NMR spectroscopy (*2*) and X-ray diffraction (*3*).

Fourier transform infrared (FTIR)[1] spectroscopy has emerged as a useful tool for the characterization of protein secondary structure with a precision lying between that of the purely predictive and the molecular coordinate approaches (*4*). One of the strengths of infrared spectroscopy is that it is amenable to a variety of sample forms including solid films or powder, solutions, liquid crystals, and so forth. Protein crystals are not necessary nor are external molecular probes required, which would supply information only about the surrounding microenvironment. Infrared spectroscopy not only provides information about protein structure in the native environment, but it can also provide insight into conformational alternations associated with changing environmental conditions, such as pH, temperature, pressure, and solvent. This advantage renders FTIR spectroscopy especially useful for probing membrane-associated proteins that are difficult to be probed by other spectroscopic techniques.

Methods currently being used to extract information on protein secondary structure from infrared spectra are based on empirical correlation between the frequencies of certain vibrational modes and types of secondary structure of polypeptide chains such as α-helix, β-sheet, β-turn, and random coil. The mode most often used and by far best characterized in this respect is the so-called amide I mode. It represents primarily the C═O stretching vibrations of amide groups and gives rise to infrared band(s) in the region between 1600 and 1700 cm^{-1}. Due to the strong absorption of water between 1640 and 1650 cm^{-1}, most structure determinations by amide I mode are performed in D_2O solutions. However, uncertainty in the NH/ND exchange process may cause a certain degree of ambiguity (*5*). Also, serious overlapping of the random coil and the α-helix bands in the amide I region makes it difficult to accurately predict α-helix contents in proteins. Attempts have also been made to exploit other vibrational modes, particularly the amide II and amide III bands. Unfortunately, even though the intensity of the amide II region is relatively strong, it is not very sensitive to the secondary structure changes of proteins.

[†] This study was supported in part by a grant (NS33740) from the National Institute of Neurological Disorders and Strokes, National Institutes of Health. B.R.S. is a Henry Dreyfus Teacher−Scholar.

* To whom correspondence should be addressed. Telephone: 508-999-8588. Fax: 508-999-8451. E-mail: bsingh@umassd.edu.

[1] Abbreviations: FTIR, Fourier transform infrared spectroscopy; PLS, partial least squares; CLS, classic least squares; ATR, attenuated total reflectance; SEP, standard error of prediction; PCC, Pearson correlation coefficient.

10.1021/bi030149y CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/10/2004

Furthermore, the amide II bands are strongly overlapped by bands originating from amino acid side chain vibrations (*6*). On the other hand, the amide III bands, which are predominantly due to the in-phase combination of N−H in-plane bending and C−N stretching vibrations, are highly sensitive to the secondary structure folding (*7*, *8*). In addition, there is no H_2O interference in this region. Therefore, even though the signal of the amide III bands is ∼5−10-fold weaker than that of the amide I bands, the amide III region is still very promising to estimate protein secondary structure content.

In recent years, several researchers have used the amide III region to study protein structures (*9*−*15*). Since both the amide I and amide III regions are broad bands, it is not possible to resolve individual bands corresponding to different secondary structure elements from the original spectra. Even though spectral resolution in the amide III region is higher than that in the amide I region, the resolution enhancement methods, such as second derivatization, are still needed to appropriately curve fit the spectra for estimating secondary structures of proteins (*7*−*15*). The most common method used to solve this problem is to employ resolution enhancement or band-narrowing methods (*4*). However, there is a certain degree of subjectivity associated with methods based on the band-narrowing approach such as the initial choice of input parameters and the assignment of secondary structures in boundary frequency regions.

Chemometric methods have been developed to resolve the individual species in a multicomponents system (*16*). Since such a method is based on full spectral analysis, it involves the use of a calibration matrix of the IR spectra of proteins with known X-ray structure and therefore avoids the need to deconvolve the spectra and assign the bands. Thus, such a method can avoid most of the subjectivity associated with the band assignment process. The PLS method is one of these methods, which is commonly used in analytical chemistry (*17*). Previously, the PLS method has been applied to amide I and amide II regions of protein IR spectra to predict protein secondary structure (*18*). In this paper, we describe application of the PLS method in the amide III region of protein IR spectra to predict the protein secondary structures. We used 16 different proteins, whose structures have already been resolved by X-ray crystallography, to set up the calibration matrix, and we employed the amide I, amide III, and a combination of amide I and amide III as the spectral region for estimating protein secondary structure. The results indicate better than 90% prediction accuracy for α-helix, *β*-sheet, and *β*-turns and 74% prediction accuracy for random coils.

MATERIALS AND METHODS

Proteins and Infrared Spectroscopic Measurement. The following proteins were purchased from Sigma Chemical Co. (St. Louis, MO) and used without further purification: alcohol dehydrogenase, carbonic anhydrase, α-chymotrypsinogen, α-chymotrypsin, concanavalin A, cytochrome *c*, hemoglobin, immunoglobulin G, lactate dehydrogenase, lysozyme, ovalbumin, myoglobin, papain, ribonuclease A, trypsin, and trypsin inhibitor. These proteins were chosen because their secondary structures have already been known from X-ray crystallography. The protein solutions were prepared in 20 mM sodium phosphate buffer, pH 7.2, except for α-chymotrypsinogen (pH 4.5, 20 mM sodium acetate buffer), α-chymotrypsin (pH 3.8, water with trace hydrochloric acid), ribonuclease A (pH 3.8, water with trace hydrochloric acid), concanavalin A (pH 5.0, 20 mM sodium acetate buffer), lysozyme (pH 4.5, 20 mM sodium acetate buffer), and trypsin (pH 8.0, 20 mM sodium phosphate buffer). The pH values were chosen in order to match the X-ray crystallography condition and to compare the structure of proteins between solution and crystal. We used pH 7.2 conditions for recording of other proteins whose crystal structures were solved in the pH range of 6.7−7.8. It is recognized that the conditions for crystal structures (salts, organic solvents, etc.) were not identical to the solution conditions used in this study for recording IR spectra. However, assuming the ordered crystal structures of proteins as true representatives of the structure in aqueous solution, even at slightly different pHs, we believe it is appropriate to use X-ray crystallographic data to predict secondary structures of proteins based on IR spectral recordings. Similar approaches have been successfully used in the past for using X-ray crystallographic data for estimating protein secondary structure from IR spectra (*19*). The concentration of these proteins used in our experiments was 1 mg/mL.

A Nicolet Model 8210 FTIR spectrometer, equipped with a zinc selenide attenuated total reflectance (ATR) accessory and DTGS detector, was used for spectral recordings at room temperature. The spectrometer was purged with CO_2-free dry air for at least 24 h before spectra were recorded. For each spectrum, a 256-scan interferogram was collected at a resolution of 4 cm^{-1}. In every case, the single beam spectrum of the buffer and the protein solutions were divided by the background single beam spectrum and then converted to the absorbance spectra. To obtain the protein spectra, the buffer spectra were subtracted. The following criteria were used to judge the water subtraction: a flat baseline between 2000 and 1700 cm^{-1} and no negative lobe between this range (*20*, *21*). All spectra were smoothed with a nine-point Savitsky−Golay function to remove the possible noise before further data analysis.

Partial Least Squares Method (PLS). The basis of all quantitative analysis using spectroscopy is Beer's law. The multicomponent analyses are based on the additivity of Beer's law; i.e., the absorbance at a specific wavenumber (wavelength) is the sum of the absorbance of all sample components that absorb at that wavenumber (wavelength). If the absorbance at different wavenumbers is collected, Beer's law can be written in the form:

$$\mathbf{A} = \sum(\mathbf{KLC} + \mathbf{E_a}) \tag{1}$$

where **A** = the vector of absorbances, **K** = the matrix of absorptivities, **L** = the vector of path lengths, **C** = the vector of concentration, and $\mathbf{E_a}$ = the matrix of the spectral error.

The classic least squares method (CLS) belongs to this kind of method and is most widely used. For the CLS method, it is assumed that protein spectra are linear combinations of *l* pure structure spectra, i.e., α-helix, *β*-sheet, *β*-turn, and random coil. Supposing there are *m* calibration proteins measured at *n* wavenumbers, then

$$\mathbf{A} = \mathbf{CK} + \mathbf{E_a} \tag{2}$$

where **A** is an $m*n$ matrix of the spectra of the m calibration proteins, **C** is an $m*l$ matrix of the concentration, and **K** is an $l*n$ matrix in which rows are the pure structure spectra.

The **K** matrix can be obtained from the calibration set as follows:

$$\mathbf{K} = \mathbf{A}\mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathbf{C}^{\mathrm{T}})^{-1} \qquad (3)$$

where $\mathbf{C}^{\mathrm{T}}$ is the transpose of matrix **C** and superscript -1 stands for the inverse of a matrix.

The **K** matrix contains the absorptivity coefficients for each of the l components at the n selected wavenumbers.

Once the **K** matrix is known, the unknown concentration can be calculated as follows:

$$\mathbf{c} = (\mathbf{K}\mathbf{K}^{\mathrm{T}})^{-1}\mathbf{K}\mathbf{A} \qquad (4)$$

where **A** is the spectrum of the protein to be analyzed. Since the resolving **K** matrix is the central part of CLS, CLS is also called the **K** matrix.

The CLS method is used widely because the mathematical steps are straightforward, and many standards and wavenumbers can be used in the calibration to obtain an averaging effect (therefore, the CLS method actually is a full-spectrum method). A major disadvantage of the CLS method is that all interfering chemical components in the spectral region of interest need to be known and included in the calibration. This is not an easy task for complex spectra like those of proteins. In addition, the two matrix inversions required by this method are a major source of errors.

The partial least squares method (PLS) is a factor analysis method, which has many of the full-spectrum advantages of the CLS method (*17*). In the PLS method, the calibration spectra can be represented as follows:

$$\mathbf{A} = \mathbf{T}\mathbf{B} + \mathbf{E_a} \qquad (5)$$

where **B** is an $h*n$ matrix with the rows of **B** being the new PLS basis set of h full-spectrum vectors, called loading vectors or loading spectra, **T** is an $m*h$ matrix of intensities (or scores) in the new coordinate system of the h PLS loading vectors (also called factors) for the calibration spectra, and $\mathbf{E_a}$ is an $m*n$ matrix of spectral residuals not fitted by the PLS model.

The analogy between the PLS model and the CLS model is quite clear since both equations involve the decomposition of **A** into the product of two smaller matrices. However, rather than basis vectors being the pure component spectra in the CLS, they are the loading vectors generated by the PLS algorithm. The intensities in the new coordinate system are no longer the concentrations that they were in CLS; instead, they are linearly related to the concentrations. The new basis set of full-spectrum loading vectors is composed of linear combinations of the original calibration spectra. The amount (i.e., intensities) of each of the loading vectors that are required to reconstruct each calibration spectrum is the score. In general, only a small number of the full-spectrum basis vector is required to represent the calibration spectra (**A**). Therefore, the PLS algorithm reduces the number of intensities (n) of each spectrum in the spectra matrix **A** to a small number of intensities (h) in the new coordinate system of the loading vectors. The data compression step also reduces the noise, as noise is distributed throughout all loading vectors, while the true spectral variation is generally concentrated in the early loading vectors.

The **c** vector of size m, containing the concentration, can be related to the spectral intensities (**T**) in the new coordinate system by solving the following set of equations:

$$\mathbf{c} = \mathbf{T}\mathbf{v} + \mathbf{e_c} \qquad (6)$$

where **v** is the $h*1$ vector of coefficients relating the scores (intensities) to the conformation fractions, **T** is the matrix of scores (intensities) from the PLS spectral decomposition, and $\mathbf{e_c}$ is the vector of concentration errors. During calibration, the least squares solution of eq 6 is

$$\mathbf{v} = [\mathbf{T}^{\mathrm{T}}\mathbf{T}]^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{c} \qquad (7)$$

During prediction, the unknown concentration is obtained by solving the equation:

$$\mathbf{c} = \mathbf{t}\mathbf{v} \qquad (8)$$

where **t** is the vector of size h of the intensities of the PLS loading data in the new coordinate system for the spectrum of the unknown sample.

The advantages of the PLS method are that it is insensitive to the presence of impurities and makes full use of all the spectral data. In addition, this method eliminates the problem of calculating the two matrix inversions in the CLS method. Since columns of **T** are orthogonal in the PLS algorithm, the least squares solution to **v** involves a trivial inversion of the diagonal ($\mathbf{T}^{\mathrm{T}}\mathbf{T}$) matrix.

The cross-validation was used to determine the optimum number of factors for the PLS algorithm. The standard error of prediction (SEP) was used to evaluate the prediction accuracy (*17, 18*):

$$\mathrm{SEP} = [(\sum (x_i - y_i)^2/(N - 1)]^{1/2}$$

where x_i is the result from the PLS method, y_i is the result from X-ray studies, and N is the number of proteins used.

To carry out the cross-validation (*17*), one of the samples is left out of the calibration set. The rest of the samples are used to perform the decomposition for one factor (loading factor) and to calculate the calibration matrix. The calibration matrix is used to predict the concentration of the sample left out. Subsequently, the next sample is left out, and the above prediction is repeated, until all samples are predicted. The SEP for this factor is calculated. The above procedure is repeated using two factors. The process is continued in this fashion up to the maximum number of factors (most of the time, the maximum number of factors is half of the sample numbers in the calibration set).

For this study, we used the PLSplus program created by Galactic Industries Corp. (Salem, NH) to calibrate and predict the secondary structure of proteins.

The spectral regions we used are amide I (1700−1600 cm^{-1}), amide III (1350−1200 cm^{-1}), and the combination of amide I and amide III. Before calibration and prediction, the spectra were normalized to a total intensity of one in each region. This step helped to remove the protein concentration effect. To predict the secondary structure of unknown proteins, a series of calibration spectra whose secondary structures are already known must be set up. After these

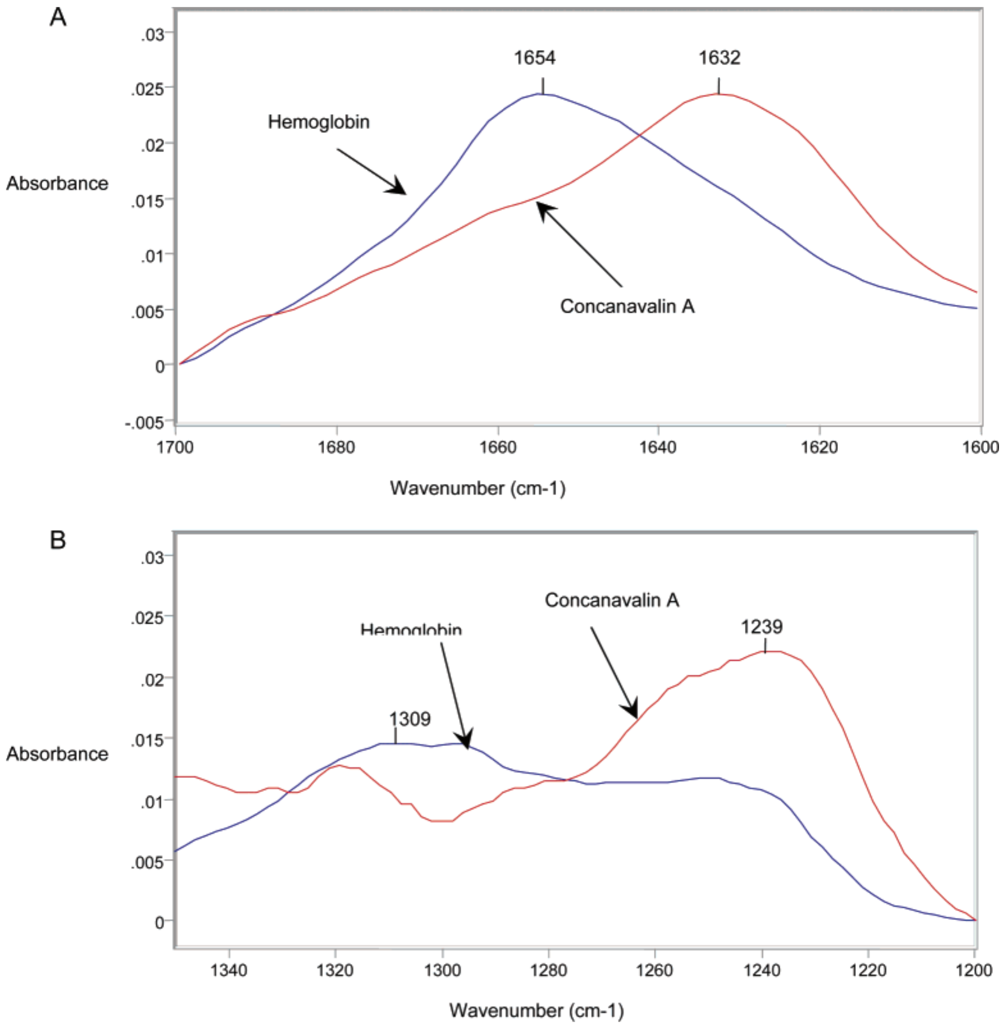

FIGURE 1: Comparison of FTIR spectra of the α-helix protein (hemoglobin) and β-sheet protein (concanavalin A). Panels: (A) amide I region and (B) amide III region.

spectra are loaded into a calibration file and the contents of each structure are input, the calibration file is ready to be used to predict the protein with an unknown structure. To predict the secondary structure of a protein, the protein is eliminated from the calibration file, and its structure is predicted by using the other 15 proteins as the calibration set.

RESULTS

Amide I and amide III regions of the infrared spectra of proteins are the two regions that are very sensitive to their secondary structure. For example, Figure 1 shows the spectra of aqueous solutions of hemoglobin and concanavalin A, which are two proteins with completely different conformations. Hemoglobin is a high α-helix protein (86% helix) with no β-sheet, and concanavalin A is a high β-sheet protein (65% β-sheet) with 3% α-helix. These significant differences in shape and frequency render the amide I and amide III regions particularly useful for predicting the conformation of proteins in an aqueous solution from their infrared spectra.

The amide III region was first analyzed using the CLS method. By doing so, we resolved the pure spectra of different secondary structures in protein on the basis of the assumption that amide III infrared bands are only contributed by four types of protein structures (Figure 2). The maximum absorbance of pure spectra for α-helix is around 1300 cm^{-1}, the maximum absorbance for β-sheet is around 1235 cm^{-1}, and the β-turn bands are located around 1260−1280 cm^{-1}, while the random coil is located around 1240−1260 cm^{-1}. Those pure spectra of four types of protein structure are approximately matched to the results of an earlier study using resolution enhancement analysis of the amide III band (*22*), suggesting that the most contribution of amide III infrared bands is from protein amide bonds, especially for α-helix and β-sheet. The pure spectra of β-turn and random coil resolved from the CLS method showed a more complicated pattern. The β-turn spectrum showed a peak at 1265 cm^{-1} and a shoulder at 1235 cm^{-1}, with some negative absorption at 1320−1330 cm^{-1} (Figure 2). The pure spectrum of random coil has shown two bands at 1245 and 1285 cm^{-1}, with significant negative absorption beyond 1290 cm^{-1} (Figure 2). The β-sheet, on the other hand, showed some positive peaks above 1290 cm^{-1} (Figure 2). Both β-turn and random coil structures are not expected to have any spectral contribution above 1300 cm^{-1}. Therefore, the peaks (both positive and negative) seen in β-turn and random coil spectra are likely to arise from contributions of other vibrational modes. The protein amide III region not only involves vibrational modes of the peptide group but also includes the side chain vibrations as well as other nondefined modes of vibration.

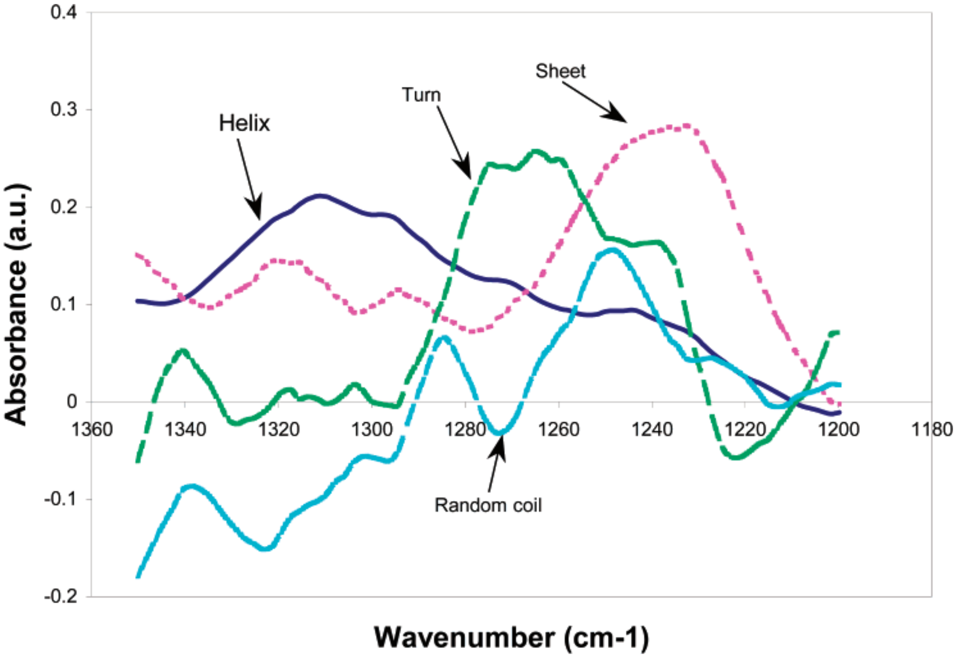

FIGURE 2: Infrared spectra calculated by the classical least squares method for the α-helix, β-sheet, β-turn, and random coil.

Since vibrational contributions other than those from protein secondary structures are not defined in the calibration file, the CLS algorithm mixes nondefined spectral contributions to those structures in the defined calibration file on the basis of least squares minimization. This approach obviously adds error in the prediction of the secondary structure content. Thus the complicated features of the amide III region make it very difficult to use CLS for resolving the protein secondary structures. That is the reason for the complicated spectral features shown in random coil and β-turn spectra (Figure 2). Those observations also suggested that the side chain and other interference have more effect on the random coil and β-turn spectral region.

To alleviate the limitations of the CLS method described above, we employed the PLS analysis to obtain accurate prediction of the secondary structure of proteins. The PLS algorithm is used to analyze both the amide I and amide III infrared bands of proteins. To determine the optimum number of factors used in the PLS algorithm, a cross-validation calculation for all the samples in the training set has been performed. The prediction depends on the number of loading factors as shown in Figure 3. The number of factors (components) involved in different secondary structures is higher in amide III alone or in the combined regions of amide I and III compared to the amide I region alone, suggesting that the amide III bands of the protein are more complex than the amide I bands. The number of factors corresponding to the minimum SEP is chosen for the construction of the calibration model.

The PLS algorithm was applied to amide I and amide III spectra to predict the protein's secondary structure. The results are listed in Table 1 and are compared with data from the X-ray studies (*23*).

We also performed the statistical tests to assess correlation between secondary structure estimation based on IR spectra and secondary structure estimated from X-ray crystallography. Protein secondary structures estimated from the PLS method were compared with data from X-ray crystallography by computing the Pearson correlation coefficients (PCC) (*25*) and SEP:

$$\mathrm{PCC} = \left(N\sum x_i y_i - \sum x_i \sum y_i\right) / \left\{\left[N\sum x_i^{2} - \left(\sum x_i\right)^2\right]^{1/2}\left[N\sum y_i^{2} - \left(\sum y_i\right)^2\right]^{1/2}\right\}$$

where x_i is the result from the PLS method and y_i is the result from X-ray studies. The statistical test results are listed in Table 2. For a perfect correlation, a PCC of 1.0 and a SEP of 0 are expected. For totally noncorrelated spectra, the PCC will be 0.

As can be derived from Table 2, β-turn and random coil structures predicted from amide III spectra were better correlated with X-ray crystallographic structures than those obtained from either amide I or amide I plus amide III spectra. The correlation analysis between the PLS result and X-ray data also showed that the α-helix and β-sheets were highly correlated with the X-ray data (PCC above 0.900). The β-turn or the random coil, especially the random coil (PCC was 0.771 when using amide III spectra and only 0.351 when using amide I spectra), on the other hand, were not well correlated.

For α-helix, the Pearson correlation coefficient (PCC) between X-ray data and amide III IR spectra was 0.950, which was reduced to 0.925 when amide I region was used alone. When amide I plus amide III was used, the PCC increased to 0.958. For β-sheets, the PCC for combined regions of amide III and amide I was the highest observed at 0.970, while it was reduced to 0.963 and 0.950 when using amide III spectra and amide I spectra, respectively.

DISCUSSION

The band-narrowing techniques and curve-fitting procedures have been applied to extract quantitative information of protein secondary structures from the infrared spectra. However, the lack of uniqueness in band assignments and elements of subjectivity in the initial choice of input

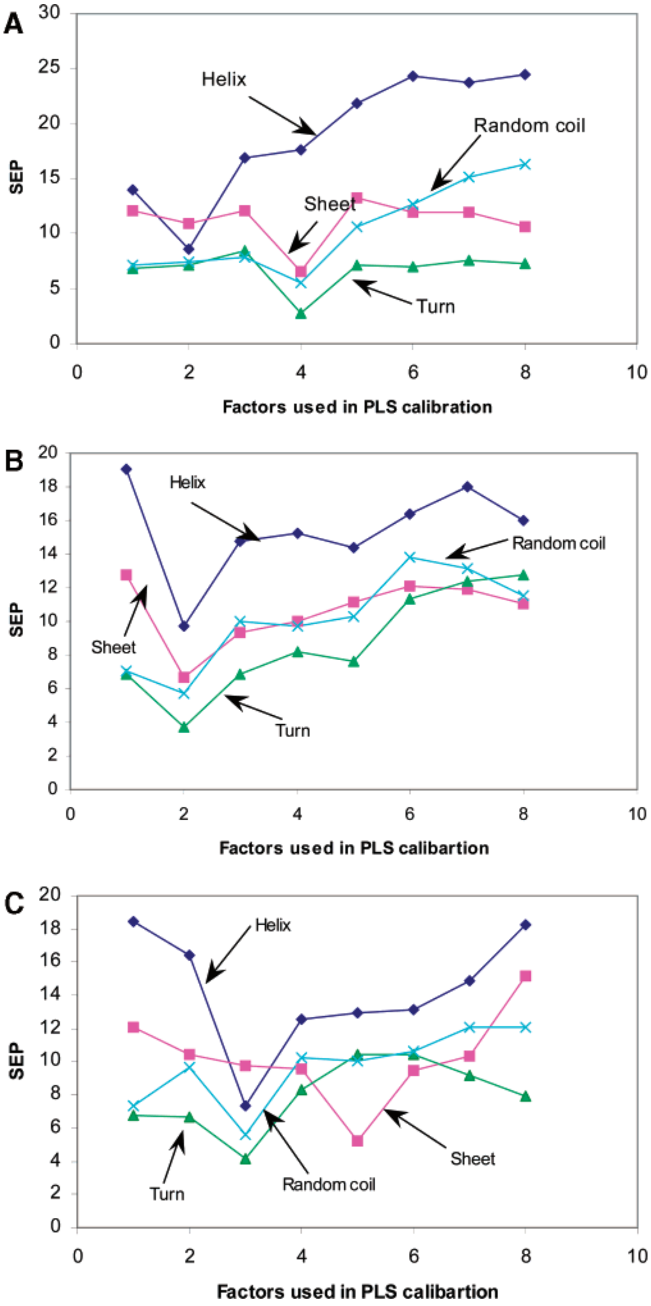

FIGURE 3: Dependence of α-helix, β-sheet, β-turn, and random coil SEP values for PLS from the number of factors used. Panels: (A) using the amide III region; (B) using the amide I region; (C) using the combined regions of amide I and amide III.

parameters for the deconvolution and curve-fitting step raise doubts about the general validity of this procedure (*26*). Chemometric methods can avoid spectral deconvolution and band assignments. Since the chemometric methods are based on the pattern recognition, the pretreatment of the data can be kept to the minimum. Therefore, it is possible to set up a spectral database for proteins of known structure, which can be used to predict the unknown structures of proteins. These methods were first applied to circular dichroism (CD) spectroscopy (*27*, *28*), but they are equally applicable to vibrational spectroscopy (*18*, *29*, *30*). The PLS algorithm belongs to these chemometric methods.

Our results indicate that the amide III region gives a better prediction for four types of secondary structures of protein than either the amide I region or amide I plus amide III region (Table 2 and Figure 4). This is because different secondary structures reflect a larger spectral difference in the amide III region than in the amide I region (see Figure 1). According to eq 1, if two spectra are very similar, the equation will become mortal, and the error of the prediction will become larger. For some special cases, if the two spectra are the same, no prediction can be obtained. Therefore, the greater the difference present in the spectra of the species in the system, the less the possibility to get the mortal equation and, thus, the greater the accuracy of the result obtained. The α-helix and random coil are overlapped in the amide I region, which makes it a higher risk to be collinear. Therefore, most studies did not differentiate those two structures when using chemomertic-based methods in the amide I region (*18*, *29*, *31*). When we separate the two structures in the current study, the random coil prediction becomes significantly deviated from the X-ray study with just the PCC of 0.385 (Table 2). Since the amide I region of protein IR spectra has a higher risk to get collinear spectra for different secondary structures, the amide III region is a better region for use in the PLS algorithm.

The interference from the side chain of the peptide is a major concern in the use of the amide III region. By analysis of the loading factors (components) for each type of secondary structure of a protein, the amide I region clearly shows a simpler spectral character, with the minimum SEP when the factor reaches 2. This suggests that only two factors contributed to each type of structure in the amide I region; adding more factors will only add more “noise” vectors in the calibration model, leading to a higher SEP. In the amide III region, on the other hand, the SEP reaches its minimum at factors of 2, 4, 4, and 4 for α-helix, β-sheet, β-turn, and random coil, respectively. This means that there are structures other than the protein secondary structure contributing to this spectral region. When amide I and amide III spectral regions were combined, the SEP reaches a minimum at factors of 3, 5, 3, and 3 for α-helix, β-sheet, β-turn, and random coil, respectively. The complicated feature of the combined region is apparently from the amide III region. Those complicated spectral features are due to both side chain contribution and the other less well defined vibration mode. Due to the complicated feature of the amide region of a protein, it will be very difficult to use the CLS method to resolve the protein structure. We have analyzed the protein secondary structure from amide III bands using the CLS method and compared the results obtained from the PLS algorithm. As seen from Table 3, when amide III bands are used, the PLS algorithm showed better prediction results than the CLS algorithm. This point is supported by the data presented in Figure 2. While the maximum absorbance of α-helix and β-sheet is close to the reported values from the curve-fitting method (*22*), the β-turn and random coil structures are different from the reported values (*22*). This is due to the unknown species existing in the calibration set. The PLS method can tolerate some unknown species in the system by including those as the loading vectors in the calibration set. This advantage makes the PLS method less sensitive to those less well defined vibration modes than both the CLS and curve-fitting methods. As shown in Table 3, the prediction for β-turn and random coil from amide III bands is much better when the PLS algorithm is used.

Table 1: Comparison of Protein Secondary Structures Determined by PLS and X-ray Crystallography

protein	α-helix	*β*-sheet	*β*-turn	random coil	method
ovalbumin	35	35	18	12	X-ray (*24*)
	46	29	16	9	PLS amide III
	28	33	22	17	PLS amide I
	36	31	19	14	PLS amides I and III
alcohol dehydrogenase	29	40	19	12	X-ray (*23*)
	22	43	19	16	PLS amide III
	31	39	16	13	PLS amide I
	32	40	14	14	PLS amides I and III
carbonic anhydrase	16	45	25	14	X-ray (*23*)
	12	50	24	14	PLS amide III
	9	52	24	15	PLS amide I
	9	52	22	16	PLS amides I and III
α-chymotrypsin	11	50	25	15	X-ray (*23*)
	20	41	21	18	PLS amide III
	4	60	23	13	PLS amide I
	13	50	21	17	PLS amides I and III
chymotrypsinogen	12	49	23	16	X-ray (*23*)
	4	56	25	15	PLS amide III
	14	52	20	15	PLS amide I
	4	55	20	21	PLS amides I and III
concanavalin A	3	65	22	10	X-ray (*23*)
	6	67	18	9	PLS amide III
	3	66	23	8	PLS amide I
	4	67	24	5	PLS amides I and III
cytochrome *c*	49	11	18	22	X-ray (*23*)
	60	10	18	12	PLS amide III
	64	7	13	16	PLS amide I
	56	12	17	15	PLS amides I and III
hemoglobin	86	0	8	6	X-ray (*23*)
	68	10	10	12	PLS amide III
	67	9	10	14	PLS amide I
	66	10	10	14	PLS amides I and III
IgG	3	67	18	12	X-ray (*23*)
	1	65	21	12	PLS amide III
	15	56	19	10	PLS amide I
	10	60	19	10	PLS amides I and III
lactate dehydrogenase	42	26	5	27	X-ray (*23*)
	46	27	1	26	PLS amide III
	60	14	11	15	PLS amide I
	54	17	10	19	PLS amides I and III
lysozyme	45	19	23	13	X-ray (*23*)
	50	17	23	10	PLS amide III
	43	22	18	17	PLS amide I
	47	20	19	13	PLS amides I and III
myoglobin	85	0	8	7	X-ray (*23*)
	79	0	10	11	PLS amide III
	75	7	6	12	PLS amide I
	87	0	5	9	PLS amides I and III
papain	28	29	18	25	X-ray (*23*)
	27	38	19	21	PLS amide III
	32	32	20	16	PLS amide I
	30	31	22	18	PLS amides I and III
ribonuclease A	23	46	21	10	X-ray (*23*)
	31	39	20	10	PLS amide III
	27	41	15	17	PLS amide I
	25	40	20	15	PLS amides I and III
trypsin inhibitor	26	45	16	13	X-ray (*23*)
	21	51	15	13	PLS amide III
	18	48	20	14	PLS amide I
	23	43	20	14	PLS amides I and III
trypsin	9	56	24	11	X-ray (*23*)
	10	53	25	12	PLS amide III
	11	52	23	14	PLS amide I
	7	55	26	12	PLS amides I and III

Table 2: Correlation Analysis of Protein Secondary Structures between the PLS Method and X-ray Data

statistical parameter	α-helix	*β*-sheet	*β*-turn	random coil	spectral region
PCC	0.950	0.963	0.935	0.771	amide III
SEP	8.3	5.8	2.3	3.9	amide III
PCC	0.925	0.950	0.828	0.351	amide I
SEP	10.0	6.8	3.6	5.7	amide I
PCC	0.958	0.970	0.874	0.622	amides I and III
SEP	7.5	5.2	3.1	4.8	amides I and III

Przybycien et al. (*31*) have tried to decompose the protein structure into the baseline background (including the residues from buffer and water), the vibrational peaks in the frequency range analyzed that are not correlated with protein secondary structure, and the amide I and/or amide III bands only contributed to protein secondary structure. By using the nonlinear and multiple linear regressions, they isolated the "ideal" reference spectra in the amide I and amide III regions, corresponding only to protein secondary structures. By analyzing those ideal reference spectra of proteins with known X-ray structures, they could predict the structure of unknown proteins. While this method is better than the traditional CLS method, it may over- or underestimate the contributions from noise by setting the function minimum. The PLS method, on the other hand, uses the loading factors to correct the unknown species in the calibration set. When the cross-validation is performed, the optimized number of loading factors can be determined for each structural type of protein. This step will avoid any possibility of either overestimation or underestimation of the spectral contributions from nondefined vibrational modes.

The prediction accuracy of PLS depends on the size of the data and the proteins in the sample set. We have compared the results from this paper with our preliminary results from a set of nine proteins (Table 4) (*32*). The prediction accuracy for random coil and *β*-turn improved in our current study. The prediction of α-helix and *β*-sheet is similar to the earlier study despite the increased sample set. The PLS algorithm has encountered difficulties in those cases where the spectral properties of the unknown protein lie outside the properties of the spectra within the calibration set. We have three proteins (lactate dehydrogenase, cytochrome *c*, and papain) in this sample set, where the random coil is over 20% from X-ray crystallography, whereas only one protein's random coil is over 20% (cytochrome *c*) in the previous study (*32*). This suggests that the proteins in the sample set should cover most types of structure in a wide range. This is difficult to achieve for certain types of structure, such as random coil, since this structure is very low in most proteins.

The ATR sampling technique also raises some concern, since binding of proteins on the ATR crystal may change the protein structure. However, several researchers (*13*, *14*, *33*, *34*) suggest that binding of the protein on the crystal does not change protein structure significantly. The comparison of protein infrared spectra collected from ATR and transmission also has been studied (*35*, *36*), and no significant differences have been observed between these two sampling techniques (*35*, *36*). The result from our current work also supports this statement.

Two assumptions were made for the development of the PLS method: first, the secondary structure of proteins in

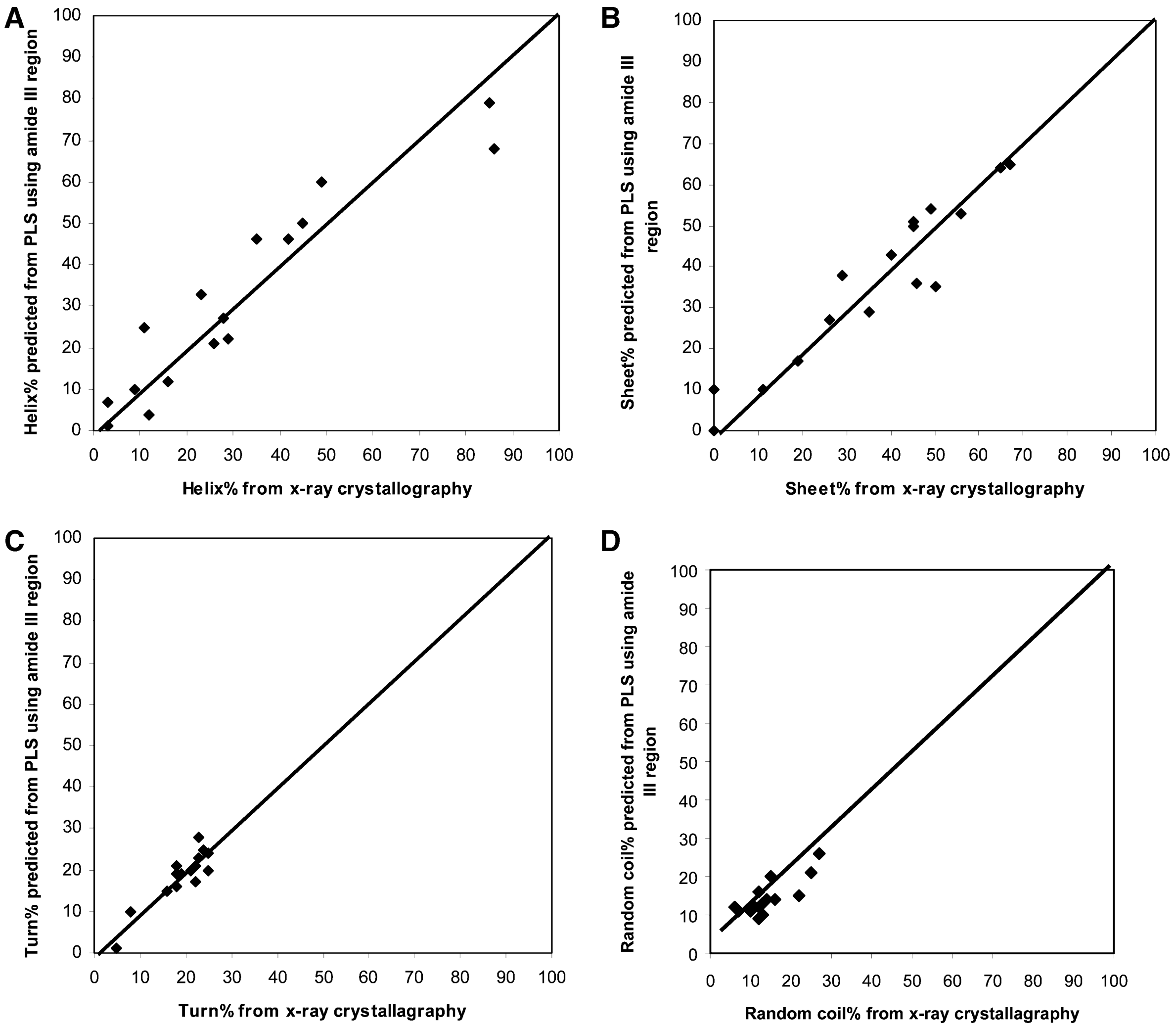

FIGURE 4: Correlation of secondary structure determined by the PLS method from amide III with values determined from X-ray crystallography (*20*). Panels: (A) α-helix; (B) *β*-sheet; (C) *β*-turn; (D) random coil.

Table 3: Comparison of Prediction Accuracy in the Amide III Region Using Different Algorithms

statistical parameter	prediction algorithm	α-helix	*β*-sheet	*β*-turn	random coil
PCC	PLS	0.950	0.963	0.935	0.771
SEP		8.3	5.8	2.3	3.9
PCC	CLS	0.916	0.940	0.751	0.402
SEP		11.4	7.8	6.3	13.0

Table 4: Effect of Sample Size on the Prediction Accuracy in the Amide III Region

statistical parameter	sample size	α-helix	*β*-sheet	*β*-turn	random coil
PCC	16 (this study)	0.950	0.963	0.935	0.771
SEP		8.3	5.8	2.3	3.9
PCC	9 (*29*)	0.943	0.968	0.821	0.412
SEP		11.0	6.8	4.0	5.1

solution is considered to be the same as that determined by X-ray diffraction in the crystalline form; second, the normalization procedure assumes equal absorptivity of amide bands irrespective of secondary structure variation.

The most definitive comparisons of solution versus crystal structure are provided by multidimensional NMR studies, which can yield solution structures for small proteins (<50 kDa) (*37*), comparable to those from X-ray diffraction. One review (*37*) concluded that, in most cases, the solution and crystal structures are very similar, allowing for some differences in the local conformation and dynamics of surface residues. Thus, the practice of using crystal structures to calibrate spectroscopic methods for determining solution structures is generally satisfactory, but anomalous cases may be encountered (*10*).

When the spectra in the amide I and amide III regions were normalized so that the sum of the absorbances in each region is equal to unity, the assumption was made that the integrated area in each region (amide I or amide III) is nearly constant for different proteins. Therefore, it was also assumed that there are equal molar absorptivities of amide bands regardless of the type of secondary structure (as in the case of the band-narrowing analysis). The correlation study done by Susi and Byler (*4*) suggested that this may be a reasonable assumption. However, studies on poly(L-lysine) have shown that the molar absorptivities of different secondary structures are significantly different (*38*).

The combination of amide I and amide III regions showed a better prediction for α-helix and *β*-sheet than using the amide III alone but a worse prediction for *β*-turn and random coil. This is possibly due to the fact that the amide I region signal is much higher than the amide III signal, and the amide I region is not very sensitive to *β*-turn and random coil, especially so for the latter (Table 2).

CONCLUDING REMARKS

On the basis of all our experimental results, we believe that the PLS method is a very promising method for predicting the protein secondary structure; it will make prediction more objective than curve-fitting-based methods. The amide III region showed a much better prediction than the amide I region. Although there is a concern that there is more contribution of less well defined structures in the amide III region, the PLS method can tolerate those unknown species.

REFERENCES

1. Fasman, G. D. (1989) Protein conformational prediction, *Trends Biochem. Sci. 14*, 295−299.
2. Braun, W. (1987) Distance geometry and related methods for protein structure determination from NMR data, *Q. Rev. Biophys. 19*, 115−157.
3. Wlodawer, A., Bott, R., and Sjolin, L. (1982) The refined crystal structure of ribonuclease A at 2.0 A resolution, *J. Biol. Chem. 257*, 1325−1332.
4. Susi, H., and Byler, D. M. (1986) Resolution-enhanced Fourier transform infrared spectroscopy of enzymes, *Methods Enzymol. 130*, 290−311.
5. Englander, S. W., Downer, N. W., and Teitelbaum, H. (1972) Hydrogen exchange, *Annu. Rev. Biochem. 41*, 903−924.
6. Chirgadze, Y. N., Fedorov, O. V., and Trushina, N. P. (1975) Estimation of amino acid residue side-chain absorption in the infrared spectra of protein solutions in heavy water, *Biopolymers 14*, 679−694.
7. Anderle, G., and Mendelsohn, R. (1987) Thermal denaturation of globular proteins. Fourier transform-infrared studies of the amide III spectral region, *Biophys. J. 52*, 69−74.
8. Kaiden, K., Matsui, T., and Tanaka, S. (1987) A study of the amide III band by FT-IR spectroscopy of the secondary structure of albumin, myoglobin, and γ-globin, *Appl. Spectrosc. 41*, 861−865.
9. Singh, B. R., Fuller, M. P., and Schiavo, G. (1990) Molecular structure of tetanus neurotoxin as revealed by Fourier transform infrared and circular dichroic spectroscopy, *Biophys. Chem. 46*, 155−166.
10. Singh, B. R., Fu, F.-N., and Ledoux, D. N. (1994) Crystal and solution structures of superantigenic staphylococcal enterotoxins compared, *Nat. Struct. Biol. 1*, 358−360.
11. Griebenow, K., and Klibanov, A. M. (1995) Lyophilization-induced reversible changes in the secondary structure of proteins, *Proc. Natl. Acad. Sci. U.S.A. 92*, 10966−10976.
12. Costantino, H. R., Griebenow, K., Mishra, P., Langer, R., and Klibanov, A. M. (1995) Fourier transform infrared spectroscopic investigation of protein stability in the lyophilized form, *Biochim. Biophys. Acta 1253*, 69−74.
13. Griebenow, K., and Klibanov, A. M. (1996) On protein denaturation in acqueous-organic mixtures but not in pure organic solvents, *J. Am. Chem. Soc. 118*, 11695−11700.
14. Bramanti, E., Benedetti, E., Sagripanti, A., Papineschi, F., and Benedetti, E. (1997) Determination of secondary structure of normal fibrin from human peripheral blood, *Biopolymers 41*, 545−553.
15. Fu, F.-N., DeOliveira, D. B., Trumble, W., Sarkar, H. K., and Singh, B. R. (1994) Secondary structure estimation of proteins using the amide III region of Fourier transform infrared spectroscopy: application to analyze calcium-binding-induced structural changes in calsequestrin, *Appl. Spectrosc. 48*, 1432−1441.
16. Lober, A., and Kowalski, B. R. (1988) Effect of interferences and calibration design on accuracy: implications for sensor and sample selection, *J. Chemom. 2*, 67−79.
17. Haaland, D. M., and Thomas, E. V. (1988) Partial least-squares methods for spectral analysis. 1. Relation to other quantitative calibration methods and extraction of qualitative information, *Anal. Chem. 60*, 1193−1202.
18. Dousseau, F., and Pezolet, M. (1990) Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods, *Biochemistry 29*, 8771−8779.
19. Dong A., Huang, P., and Caughey, W. S. (1990) Protein structures in water from second-derivative amide I infrared spectra, *Biochemistry 29*, 3303−3308.
20. Haris, P. I., Lee, D. C., and Chapman, D. A. (1986) Fourier transform infrared investigation of the structural differences between ribonuclease A and ribonuclease S, *Biochim. Biophys. Acta 847*, 255−265.
21. Mitchell, R. C., Haris, P. I., Fallowfield, C., Keeling, D. J., and Chapman, D. (1988) Fourier transform infrared spectroscopic studies on gastric H^+/K^+-ATPase, *Biochim. Biophys. Acta 941*, 31−38.
22. Cai, S., and Singh, B. R. (1999) Identification of beta-turn and random coil amide III infrared bands for secondary structure estimation of proteins, *Biophys. Chem. 80*, 7−20.
23. Levitt, M., and Greer, J. (1977) Automatic identification of secondary structure in globular proteins, *J. Mol. Biol. 114*, 181−239.
24. Stein P. E., Leslie, A. G., Finch, J. T., and Carrell, R. W. (1991) Crystal structure of uncleaved ovalbumin at 1.95 Å resolution, *J. Mol. Biol. 221*, 941−959.
25. Kalnin, N. N., Baikalov, I. A., and Venyaminov, S. Y. (1990) Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. III. Estimation of the protein secondary structure, *Biopolymers 30*, 1273−1280.
26. Surewicz, W. K., Mantsch, H. H., and Chapman, D. (1993) Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment, *Biochemistry 32*, 389−394.
27. Yang, J. T., and Wu, C. S. (1986) Martinez HM. Calculation of protein conformation from circular dichroism, *Methods Enzymol. 130*, 208−269.
28. Manavalan, P., and Johnson, W. C., Jr. (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra, *Anal. Biochem. 167*, 76−85.
29. Lee, D. C., Haris, P. I., Chapman, D., and Mitchell, R. C. (1990) Determination of protein secondary structure using factor analysis of infrared spectra, *Biochemistry 29*, 9185−9193.
30. Compton, L. A., and Johnson, W. C. (1986) Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication, *Anal. Biochem. 155*, 155−167.
31. Vedantham, G., Sparks, H. G., Sane, S. U., Tzannis, S., and Przybycien, T. M. (2000) A holistic approach for protein secondary structure estimation from infrared spectra in H_2O solutions, *Anal. Biochem. 285*, 33−49.
32. Cai, S., and Singh, B. R. (2000) Determination of the secondary structure of proteins from amide I and amide III infrared bands using partial least-square method, in *Infrared analysis of peptides and proteins* (Singh, B. R., Ed.) pp 117−129, American Chemical Society, Washington, DC.
33. Jakobsen, R. J., and Wasacz, F. M. (1990) Infrared spectra-structure correlations and adsorption behavior for helix proteins, *Appl. Spectrosc. 44*, 1478−1482.
34. Boncheva, M., and Vogel, H. (1997) Formation of stable polypeptide monolayers at interfaces: controlling molecular conformation and orientation, *Biophys. J. 73*, 1056−1072.
35. Singh, B. R., and Fuller, M. P. (1991) FT-IR in combination with the attenuated total reflectance technique: A very sensitive method for the structural analysis of polypeptides, *Appl. Spectrosc. 45*, 1017−1021.
36. Oberg, K. A., and Fink, A. L. (1998) A new attenuated total reflectance Fourier transform infrared spectroscopy method for the study of proteins in solution, *Anal. Biochem. 256*, 92−106.
37. Wagner, G. (1997) An account of NMR in structural biology, *Nat. Struct. Biol. 4*, 841−844.
38. Jackson, M., Haris, P. I., and Chapman, D. (1989) Fourier transform infrared spectroscopic studies of lipids, polypeptides and proteins, *J. Mol. Struct. 214*, 329−355.

BI030149Y