Sokolovsky, M., Riordan, J. F., & Vallee, B. L. (1966) *Biochemistry 5*, 3582–3589.
Stubbe, J. A. (1989) *Annu. Rev. Biochem. 58*, 257–285.
van der Ouderaa, F. J., Buytenhek, M., Nugteren, D. H., & van Dorp, D. A. (1977) *Biochim. Biophys. Acta 487*, 315–331.
Vickery, L., Nozawa, T., & Sauer, K. (1976) *J. Am. Chem. Soc. 98*, 343–350.

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†

Françoise Dousseau and Michel Pézolet*
Centre de Recherche en Sciences et Ingénierie des Macromolécules, Département de Chimie, Université Laval, Cité Universitaire, Québec, Canada G1K 7P4

Received March 14, 1990; Revised Manuscript Received June 4, 1990

ABSTRACT: A method for estimating protein secondary structure from infrared spectra has been developed. The infrared spectra of H_2O solutions of 13 proteins of known crystal structure have been recorded and corrected for the spectral contribution of water in the amide I and II region by using the algorithm of Dousseau et al. [Dousseau, F., Therrien, M., & Pézolet, M. (1989) *Appl. Spectrosc. 43*, 538–542]. This calibration set of proteins has been analyzed by using either a classical least-squares (CLS) method or the partial least-squares (PLS) method. The pure-structure spectra calculated by the classical least-squares method are in good agreement with spectra of poly(L-lysine) in the α-helix, β-sheet, and undefined conformations. The results show that the best agreement between the secondary structure determined by X-ray crystallography and that predicted by infrared spectroscopy is obtained when both the amide I and II bands are used to generate the calibration set, when the PLS method is used, and when it is assumed that the secondary structure of proteins is composed of only four types of structure: ordered and disordered α-helices, β-sheet, and undefined conformation. Attempts to include turns in the secondary structure estimation have led to a loss of accuracy. The standard deviation of the difference between X-ray and infrared secondary structure estimates with this method is 4.8% for the α-helix, 3.7% for the β-sheet, and 5.1% for the undefined structure, whereas the regression coefficients are 0.95, 0.96, and 0.56, respectively. The spectra of the calibration proteins were also recorded in ^2H_2O solution. After correction for the contribution of the combination band of ^2H_2O in the amide I′ band region, the spectra were analyzed with PLS, but the results were not as good as for the spectra obtained in H_2O, especially for the α-helical conformation.

The complete tertiary structure of proteins is currently accessible only by X-ray crystallography and a few closely related diffraction techniques. All these techniques require that the molecule can form a well-ordered crystalline array, which is only attainable for a small fraction of proteins. Indeed, some proteins are disordered by their very nature and cannot be studied by high-resolution diffraction techniques. Two-dimensional NMR[1] spectroscopy offers a viable alternative to diffraction techniques but, so far, it has been limited to low molecular weight proteins. Therefore, methods for the determination of the secondary structure of proteins in solution by spectroscopic techniques like circular dichroism (CD) and vibrational spectroscopy have been developed and provide a valuable average picture of the structure of proteins in an environment that is close to their native one.

The marked sensitivity of infrared amide bands to the conformation of the peptide backbone of proteins is well established. In their pioneering paper, Elliot and Ambrose (1950) were the first to demonstrate the existence of empirical correlations between the amide I and amide II infrared bands of polypeptides and their conformation. Later, Miyazawa and Blout (1961) and Krimm and Bandekar (1986 and earlier papers) have refined these observations by making detailed vibrational analyses of the amide bands in order to establish correlations between the frequency of these bands and types of secondary structure of polypeptides such as α-helix, β-sheet, turns, and undefined structure.

Since the polypeptide backbone of globular proteins is normally folded in more than one conformation, the amide bands of these proteins result from the superimposition of bands corresponding to the different types of structure. Byler and Susi (1986) and Surewicz and Mantsch (1988) have extracted information from the infrared spectra of proteins in ^2H_2O solution by resolution enhancement of the broad amide I′ band. These band-narrowing methods allow the decomposition of the amide I′ band into its underlying components. Although this methodology provides a basis for the qualitative estimation of the secondary structure of proteins, it is not a routine procedure since there are still difficulties associated with obtaining artifact-free resolution-enhanced spectra and with the band-fitting procedure. On the other hand, Eckert et al. (1977) have calculated from the amide I′ bands of proteins of known conformational composition the charac-

† This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (M.P.) and the Fonds FCAR of the Province of Québec (M.P.).

[1] Abbreviations: CD, circular dichroism; CLS, classical least-squares method; PLS, partial least-squares method; NMR, nuclear magnetic resonance; FTIR, Fourier transform infrared; TES, *N*-[tris(hydroxymethyl)methyl]-2-aminoethanesulfonic acid.

0006-2960/90/0429-8771$02.50/0

teristic reference spectra of the α-helix, β-sheet, and undefined conformations and used these basis spectra to estimate the conformation of globular proteins from their amide I′ bands. More recently, Lee et al. (1989) have proposed to analyze directly the amide I band by using a factor analysis coupled with multiple linear regression, while Dong et al. (1990) have used the second derivative of the amide I band to predict the conformation of proteins.

Raman spectroscopy has also been used to extract quantitative information on the secondary structure of proteins (Lippert et al., 1976; Pézolet et al., 1976; Thomas & Agard, 1984). Williams (Williams & Dunker, 1981; Williams, 1983, 1986) has analyzed the Raman amide I and III bands of proteins as a linear combination of a set of spectra of proteins of known structures. Recently, Bussian and Sander (1989) have modified the Williams method by using a new simplified water subtraction procedure and a novel numerical procedure. On the other hand, Berjot et al. (1987) have also proposed a method based on the principle of the calculation of Raman spectral intensities corresponding to pure classes of secondary conformation giving rise to peptide backbone amide I vibrations.

The majority of infrared investigations of proteins in solution have been carried out in ^2H_2O to overcome the strong interfering water absorption in the amide I region. However, the improved sensitivity and extensive data manipulation of modern FTIR spectrophotometers now permit reliable spectral subtraction, so that algorithms have recently been developed to subtract quantitatively the water spectrum from the reflection (Powell et al., 1986) and transmission (Dousseau et al., 1989) spectra of proteins. These water subtraction methods allow the use of both the amide I and II bands for the determination of the secondary structure content of proteins. In the current study, several methods for the determination of the secondary structure content of proteins from their water-corrected infrared spectra have been investigated. First, the amide I and II bands of aqueous solutions of proteins have been analyzed by a least-squares method similar to that used by Berjot et al. (1987) for Raman spectroscopy. In this method, the protein structures are described as ordered and disordered α-helices, β-sheets, and undefined structure, according to the classification of Levitt and Greer (1977). Therefore, parallel and antiparallel β-sheets are not distinguished, and turns are included in the undefined structure. Second, the same water-corrected spectra have been analyzed without calculating the spectra of the pure classes of structure, by using the partial least-squares multivariate statistical method (PLS), whose application to spectral analysis has recently been reviewed by Haaland and Thomas (1988). Finally, for comparison, PLS was also applied to the amide I′ band of the spectra of proteins in ^2H_2O solutions.

Materials and Methods

Infrared Spectroscopy. Spectra were recorded with a Bomem DA3-02 Fourier transform spectrophotometer with a narrow-band mercury–cadmium–telluride detector and a germanium-coated KBr beam splitter. Typically, 1000 interferograms were routinely recorded with a maximal optical retardation of 0.5 cm, coadded, triangularly apodized, and Fourier transformed to yield a resolution of 2 cm^{-1}.

In order to subtract quantitatively the water spectrum from those of aqueous solutions, a homemade closed cell of constant thickness was used (Dousseau et al., 1989). The two windows were separated by a 6-μm-thick spacer to keep the absorbance of the water-bending mode band at 1645 cm^{-1} below unity. The windows were held tightly in a copper jacket whose temperature was maintained at 20.0 ± 0.1 °C by a thermoelectric device similar to the one described elsewhere for Raman spectroscopy (Pézolet et al., 1983).

Table I: Sources of Proteins and Buffers in Which They Were Dissolved

protein	buffer[a]	source[b]
adenylate kinase	G	chicken muscle
α-chymotrypsin	A	bovine pancreas[c]
chymotrypsinogen A	B	bovine pancreas
concanavalin A	A	jack bean
cytochrome *c*	E	horse heart
elastase	B	porcine pancreas
lactate dehydrogenase	F	rabbit muscle
lysozyme	A	egg white[c]
myoglobin	E	horse skeletal muscle
papain	E	papaya latex
ribonuclease A	C	type XII-A, bovine pancreas
triosephosphate isomerase	G	rabbit muscle
trypsin inhibitor	D	soya bean
tropomyosin	I	rabbit muscle
polylysine[d]	H	

[a] Buffer A, 0.05 M cacodylate, pH 5.0; buffer B, 0.05 M cacodylate, pH 7.0; buffer C, 0.05 M cacodylate and 1 M NaCl, pH 5.0; buffer D, 0.05 M cacodylate and 1 M NaCl, pH 7.0; buffer E, 0.01 M phosphate, pH 6.8; buffer F, 0.02 M phosphate and 0.08 M NaCl, pH 6.8; buffer G, 0.1 M phosphate, pH 7.0; buffer H, H_2O, pH 12; buffer I, 0.01 M TES, 0.1 M KCl, and 0.02 M $MgCl_2$. [b] All sources are from Sigma except as noted. [c] Calbiochem. [d] MW 40 000.

Proteins. Proteins and buffers used for the current study are listed in Table I. They were all used without further purification except for lactate dehydrogenase and tropomyosin, which were dialyzed in order to remove excess salts. Solutions were prepared by dissolving solid proteins in proper buffers to obtain final protein concentrations of 5–7% by weight. The ^2H_2O solutions of proteins were prepared at least 3 h before recording the spectra. The completeness of the hydrogen–deuterium exchange was verified by the reproducibility of successive spectra in the amide regions. The p^{2}H was adjusted with NaO^2H or ^2HCl.

Data Analysis. (*a*) *Water Spectrum Subtraction.* The numerical method for the water spectrum subtraction uses the combination band of water at approximately 2125 cm^{-1} as an internal intensity standard for the determination of the scaling factor (Dousseau et al., 1989). The algorithm starts by multiplying the water spectrum by 0.7. The scaled spectrum is then subtracted from the protein spectrum, a second-order polynomial function is passed through the difference spectrum in the 1750–2650-cm^{-1} region by a least-squares method, and the sum of the squared deviations (χ^2) is determined. The scaling factor is then increased by an increment of 0.1, and the operation is repeated until a minimum of χ^2 is reached. Then the scaling factor is decreased by 5 times the increment, which is also reduced by a factor of 2. The whole process is repeated until the increment falls below 10^{-5}.

(*b*) *^2H_2O Spectrum Subtraction.* Spectra of proteins in ^2H_2O solutions had to be corrected for the contribution of the combination band of ^2H_2O that underlies the amide I′ band. It was found that the best correction criterion is to subtract the ^2H_2O spectrum from that of a protein solution until a flat baseline is obtained between 1730 and 2100 cm^{-1}. The algorithm starts by multiplying the ^2H_2O spectrum by a scaling factor of 0.7. The scaled spectrum is then subtracted from the protein spectrum, and a straight baseline is passed through the difference spectrum in the 1730–2100-cm^{-1} region by a least-squares method. The scaling factor is then increased by an increment of 0.1 and the operation is repeated. When the ^2H_2O spectrum is oversubtracted, the slope of the baseline changes sign. Then, the increment is set to 0.01 and the scaling

factor is decreased until the slope of the baseline changes sign again. The whole process is finally repeated with an increment of 0.001.

(*c*) *Baseline Correction and Normalization.* For the determination of the secondary structure content of proteins of H_2O solutions from the amide I and II region, straight baselines were subtracted between 1480 and 1720 cm^{-1}. The baseline-subtracted spectra were then normalized to a total intensity of one. When only the amide I, or the amide I′, region was considered, a flat baseline was subtracted between 1600 and 1720 cm^{-1} and spectra were also normalized.

(*d*) *Classical Least-Squares Method (CLS).* Since the classical and partial least-squares methods have recently been clearly reviewed by Haaland and Thomas (1988), only a short summary of these methods will be presented in the following two sections.

For the CLS method, it is assumed that protein spectra are linear combinations of l pure-structure spectra, i.e., α-helix, β-sheet, and undefined structure:

$$A = CK + E_A \tag{1}$$

where A is an $m \times n$ matrix of the spectra of the m calibration proteins, C is an $m \times l$ matrix of the conformation fractions of the calibration proteins, i.e., fractions of residues with the lth conformation in the proteins, K is an $l \times n$ matrix in which the rows are the pure-structure spectra, and E_A is the $m \times n$ matrix of spectral errors.

The classical least-squares solution to eq 1 during calibration is

$$\hat{K} = [C'C]^{-1}C'A \tag{2}$$

where C' is the transposed matrix of C. During prediction, the least-squares solution for the vector of unknown conformation fractions, c, is

$$\hat{\mathbf{c}} = [\hat{K}\hat{K}']^{-1}\hat{K}a \tag{3}$$

where a is the spectrum of the protein to be analyzed.

(*e*) *Partial Least-Squares Method (PLS).* PLS is a factor analysis method that has many of the full-spectrum advantages of the CLS method. For this model, the calibration spectra can be represented as follows (Haaland & Thomas, 1988), based on the Lindberg et al. (1983) spectral decomposition notation:

$$A = TB + E_A \tag{4}$$

where B is an $h \times n$ matrix in which the rows are the new basis set of h full-spectrum vectors, often called loading vectors or loading spectra, T is an $m \times h$ matrix of intensities (or scores) in the new coordinate system of the h PLS loading vectors for the calibration spectra, and E_A is an $m \times n$ matrix of spectral residuals not fit by the PLS model.

As shown in eq 4, A is decomposed in two smaller matrices with PLS, as for CLS. However, the basis vectors are not the pure-component spectra but loading vectors generated by PLS. The intensities in the new coordinate system are no longer the conformation fractions of the calibration proteins, but they are linearly related to these fractions. The new basis set of full-spectrum loading vectors is composed of linear combinations of the original calibration spectra. The amounts (i.e. intensities) of each of the loading vectors that are required to reconstruct each calibration spectrum are the scores. Since the rank of A is normally smaller than the number of calibration proteins (m), PLS reduces the number of intensities (n) of each spectrum in the spectral matrix A to a small number of intensities (h) in the new coordinate system of the loading vectors. This data compression step also reduces the noise, since noise is distributed throughout all loading vectors.

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

$$\mathbf{c} = T\mathbf{v} + \mathbf{e}_c \tag{5}$$

where **v** is a vector of size h containing the coefficients relating scores to the conformation fractions and $\mathbf{e}_c$ is the vector of conformation fractions errors. During calibration, the least-squares solution of eq 4 is

$$\hat{\mathbf{v}} = [T'T]^{-1}T'\mathbf{c} \tag{6}$$

where T' is the transposed matrix of T. During prediction, the unknown fraction, c, for each conformation is obtained by solving the following equation:

$$\hat{c} = \mathbf{t}'\hat{\mathbf{v}} \tag{7}$$

where $\mathbf{t}'$ is the vector of size h of the intensities, in the new coordinate system, of the PLS loading vector for the spectrum of the protein of unknown conformation. This method eliminates the problem of calculating the two matrix inversions of eqs 2 and 3. Since columns of T are orthogonal in PLS, the least-squares solution of **v** in eq 6 involves the trivial inversion of the diagonal $[T'T]$ matrix.

For this study, we have used the PLS algorithms for calibration and prediction developed by Haaland and Thomas (1988) and implemented by Galactic Industries Corporation (Salem, NH).

(*f*) *Accuracy of the Method.* The standard deviation on the difference between X-ray and infrared estimates of the secondary structure content of the calibration set of proteins was calculated from the following equation:

$$\sigma = \{[\sum_k (c_k^x - c_k^i)^2/n] - [\sum_k (c_k^x - c_k^i)/n]^2\}^{1/2}$$

where c_k^x and c_k^i are, for the kth protein of the calibration set, the secondary structure contents estimated from X-ray and infrared results for each class of structure, respectively, and n is the number of proteins.

Results

The amide I and II region of the infrared spectra of proteins is very sensitive to their secondary structure. For example, Figure 1 shows the spectra of aqueous solutions of myoglobin and concanavalin A, which are two proteins of completely different conformation. In the spectrum of myoglobin, whose conformation is 77% α-helical with no β-sheet, the amide I and II bands are narrow and symmetrical and are centered at 1660 and 1550 cm^{-1}, respectively. On the other hand, 64% of the residues of concanavalin A are arranged in β-sheets. In this case, the amide I and II bands are much broader and are located at about 1630 and 1540 cm^{-1}, respectively. In addition, the amide I band is asymmetrical toward high frequencies and a weaker component is clearly observed around 1690 cm^{-1}. These significant differences in shape and frequency render the amide I and II region particularly useful to predict the conformation of proteins in aqueous solutions from their infrared spectra.

Several methods of analysis of the amide bands have been used in the current study. First, the amide I and II region has been analyzed by the classical least-squares method as a linear combination of bands due to only three classes of structure, i.e., α-helix, β-sheet, and undefined structure. No distinction was, therefore, made between parallel and antiparallel β-sheets, and turns were included in the undefined structure. The first step of this method was the calculation of the pure-structure

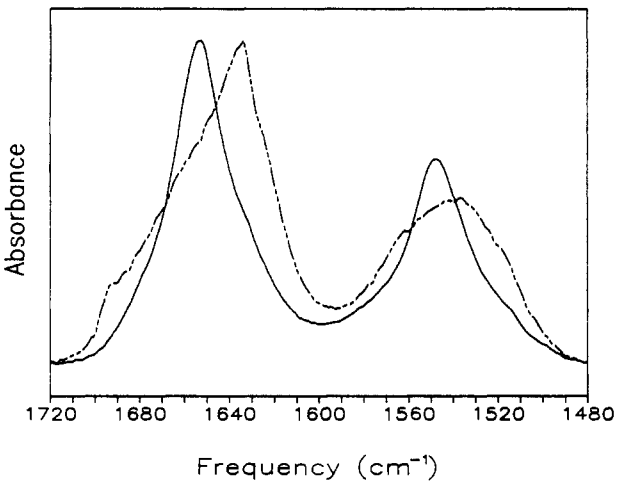

FIGURE 1: Water-corrected infrared spectra of myoglobin (—) and concanavalin A (- - -) in the amide I and II region.

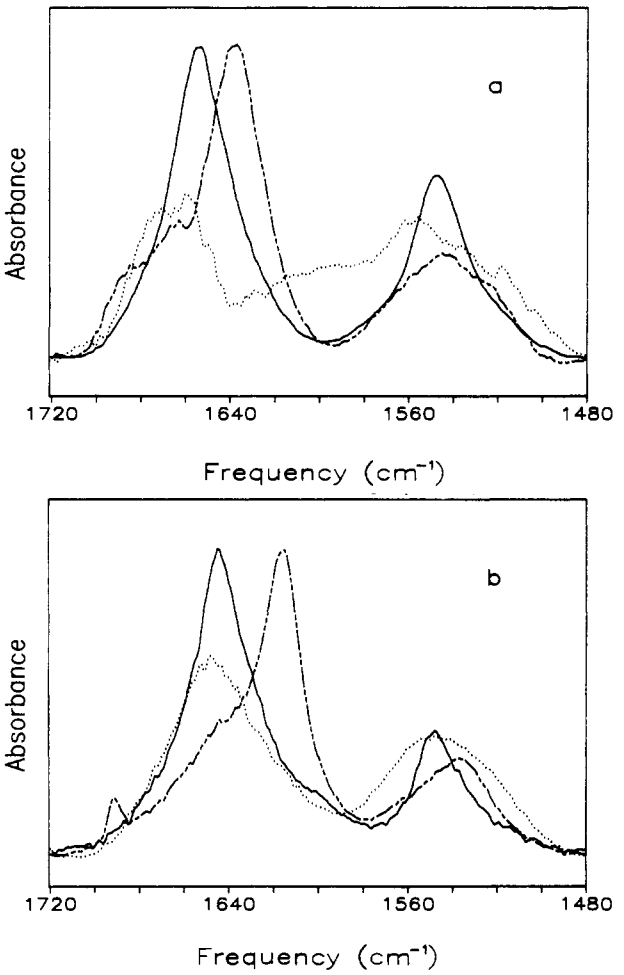

FIGURE 2: (a) Infrared spectra calculated by the classical least-squares method for the α-helix (—), β-sheet (- - -), and undefined (···) conformations. (b) Infrared spectra of polylysine in the α-helix (—), β-sheet (- - -), and undefined (···) conformations.

spectra from a set of 13 calibration proteins of known conformation from X-ray diffraction. The secondary structure estimates from X-ray diffraction were calculated from the data of Levitt and Greer (1977), except for myoglobin, for which the results of Kabsch and Sanders (1983) were used. Figure 2 shows the calculated spectra for each type of structure and, for comparison, those of polylysine in the α-helix, β-sheet, and random coil conformations. As seen in this figure, the calculated pure-structure spectra are qualitatively in very good

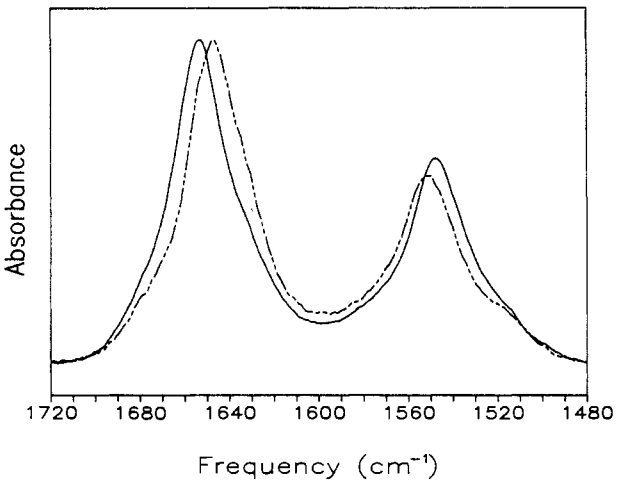

FIGURE 3: Amide I and II infrared bands of myoglobin (—) and tropomyosin (- - -).

agreement with those of polylysine. For the undefined structure, however, the amide bands in the spectrum of polylysine are much more symmetric than those in the calculated spectrum. This could be partly due to the fact that, because of the regular peptide sequence of polylysine, the random coil conformation of this polypeptide is more ordered than the undefined conformation in proteins with nonrepetitive sequences. In addition, it should be kept in mind that the undefined conformation includes the turns and any other regular conformation that is different from α-helix and β-sheet.

The second step of this method was to use the calculated pure-structure spectra to estimate the secondary structure content of proteins. In order to test the method, each protein was in turn eliminated from the calibration set, and the remaining proteins were used to predict its secondary structure content. The calculated secondary structure contents are presented in Table II (method 1), in comparison with the X-ray estimates (method X), while the average standard deviations of the differences, or errors, between infrared and X-ray diffraction estimates and the correlation coefficients between the two estimates are given in Table III. As seen in these tables, the highest disagreement between the secondary structure content estimated from infrared and X-ray data is for the α-helix. Indeed, the average absolute standard deviation is 11.7% for the prediction of the α-helix content, while it is 6.6% and 6.7% for the β-sheet and undefined structure contents, respectively. This result is not surprising if the spectra of two proteins with a large amount of α-helix, such as myoglobin and tropomyosin, are compared. As seen in Figure 3, the spectra of these proteins, which are approximately 80% and 90% α-helical, respectively, display symmetric and narrow amide I and amide II bands characteristic of the α-helix, but the maximum absorbance of the amide bands is approximately 10 cm^{-1} lower in frequency in the spectrum of tropomyosin, suggesting stronger hydrogen bonds in long strands of α-helices as those encountered in tropomyosin. Therefore, as done previously for the analysis of the amide I band in the Raman spectra of proteins (Williams & Dunker, 1981; Williams, 1983, 1986; Berjot et al., 1987; Bussian & Sander, 1989) two types of α-helices instead of one were introduced in the calculation to account for the difference in strength of intramolecular hydrogen bonding.

As suggested by Williams (Williams & Dunker, 1981), the first two and the last two residues of an α-helix segment were considered to form disordered helix while the remaining residues form the ordered helix. As a model for the ordered helix, the spectrum of polylysine in the α-helix conformation was

Table II: Comparison of Secondary Structure Content (%) As Estimated from Infrared and X-ray Data[a,b]

protein		method[c] X	1	2	3	4	5	6	protein		method[c] X	1	2	3	4	5	6
adenylate kinase	H_t	63	48	56	49	54	66	45	lysozyme	H_t	46	61	47	50	46	40	34
	H_o	43		41		40	42	28		H_o	24		19		27	25	22
	H_d	20		15		14	24	17		H_d	22		28		19	15	12
	S	18	30	21	21	25	20	29		S	19	13	28	−4	24	29	19
	U	19	22	23	21	21	15	26		U	35	26	24	28	28	31	43
α-chymotrypsin	H_t	8	14	14	13	12	14	−11	myoglobin	H_t	77	72	84	73	75	83	51
	H_o	3		6		4	3	−12		H_o	53		60		48	54	32
	H_d	5		8		8	11	1		H_d	24		24		27	29	19
	S	55	50	49	52	52	49	62		S	0	2	−9	−1	−2	−5	11
	U	37	36	36	36	36	37	45		U	23	26	26	28	26	22	29
chymotrypsinogen A	H_t	12	21	18	13	16	2	29	papain	H_t	25	0	0	23	21	25	11
	H_o	7		7		8	0	19		H_o	13		−6		2	2	−8
	H_d	5		11		8	2	10		H_d	12		6		19	23	19
	S	49	48	51	53	51	57	42		S	29	40	41	31	32	19	43
	U	39	32	31	36	33	41	28		U	43	57	58	47	47	57	47
concanavalin A	H_t	3	0	13	−5	−6	25	37	ribonuclease A	H_t	23	30	27	24	22	20	36
	H_o	0		14		−6	13	24		H_o	11		12		19	1	16
	H_d	3		−1		0	12	13		H_d	12		15		13	19	20
	S	64	62	49	60	63	43	52		S	46	36	40	40	40	40	40
	U	33	38	38	44	43	32	25		U	31	35	33	36	38	41	27
cytochrome *c*	H_t	46	46	41	45	41	54	60	triosephosphate isomerase	H_t	52	45	50	52	53	30	45
	H_o	27		21		23	34	41		H_o	36		35		30	21	28
	H_d	19		20		18	20	19		H_d	16		15		23	9	17
	S	15	13	17	19	21	20	10		S	24	30	24	23	23	30	29
	U	39	40	40	34	36	15	30		U	24	26	27	27	27	35	26
elastase	H_t	10	10	11	14	16	11	43	trypsin inhibitor	H_t	26	14	22	22	22	29	44
	H_o	5		6		15	8	34		H_o	12		17		14	21	27
	H_d	5		5		1	3	9		H_d	14		5		8	8	17
	S	46	54	52	46	46	52	29		S	45	49	40	49	46	45	36
	U	44	34	35	39	37	37	24		U	29	37	38	29	31	26	25
lactate dehydrogenase	H_t	42	56	51	46	47	48	29									
	H_o	21		30		30	25	23									
	H_d	21		21		17	13	6									
	S	26	20	25	26	24	31	27									
	U	32	26	25	27	29	32	40									

[a] Abbreviations: H_t, total helix; H_o, ordered helix; H_d, disordered helix; S, parallel and antiparallel β-sheets; U, undefined structure including turns. [b] The infrared estimates for each protein were calculated after eliminating the protein from the calibration set. [c] Method X, X-ray data from Levitt and Greer (1977); method 1, CLS with H_t, amide I and II regions; method 2, CLS with H_o and H_d, amide I and II regions; method 3, PLS with H_t, amide I and II regions; method 5, PLS with H_o and H_d, amide I region only; method 6, PLS with H_o and H_d, amide I′ region only.

Table III: Comparison of Correlation Coefficients (r) and Standard Deviations (σ) for Infrared Estimates of Secondary Structure[a]

method	α-helix σ (%)	α-helix r	β-sheet σ (%)	β-sheet r	undetermined σ (%)	undetermined r	turns σ (%)	turns r
CLS, amide I and II								
H_t	11.7	0.75	6.6	0.86	6.7	0.44		
$H_o + H_d$	9.6	0.82	7.2	0.84	7.2	0.40		
PLS, amide I and II								
H_t	5.1	0.94	6.9	0.88	5.0	0.57		
H_t + turns	10.8	0.78	4.4	0.94	6.4	0.03	6.3	0.5
$H_o + H_d$	4.9	0.95	3.7	0.96	5.1	0.56		
$H_o + H_d$ + turns	6.6	0.93	5.0	0.92	5.2	0.06	5.9	0.37
PLS, amide I								
H_t + turns	5.9	0.82	6.4	0.87	3.5	0.48	8.6	0.15
$H_o + H_d$	10.5	0.79	8.2	0.79	6.3	0.58		
$H_o + H_d$ + turns	9.6	0.85	11.6	0.66	3.5	0.10	9.4	0.15
PLS, amide I′								
$H_o + H_d$	20.5	0.23	9.6	0.73	10.5	0.12		

[a] The abbreviations used are the same as in Table II. The standard deviations were calculated from differences between infrared and X-ray estimates expressed in percentage.

included in the calibration set. As seen in Table II (method 2) and Table III, the estimation of the α-helix content is significantly improved with this four-structure method, but predictions for the β-sheet and the undefined structure are not as good. This result is essentially due to the mathematical method used. Indeed, in the CLS method, two matrix inversions are necessary, one for the calculation of the pure-structure spectra and the other for the prediction itself (see eqs 2 and 3). Matrix inversion is a major source of errors in the calculations, and the addition of one dimension to the matrices increases these errors. In addition, another disadvantage of CLS is that all interfering components in the spectral region of interest need to be known and included in the calibration.

In order to improve the analysis of the infrared data, the partial least-squares method was also used. The results obtained with PLS [Table II (methods 3 and 4) and Table III] for the amide I and II region are much better for all cases investigated and especially when two types of helices are introduced in the calculation. For the latter case, the average standard deviation is 4.9% and 3.7% for the α-helix and β-sheet conformations, respectively, which corresponds to a significant

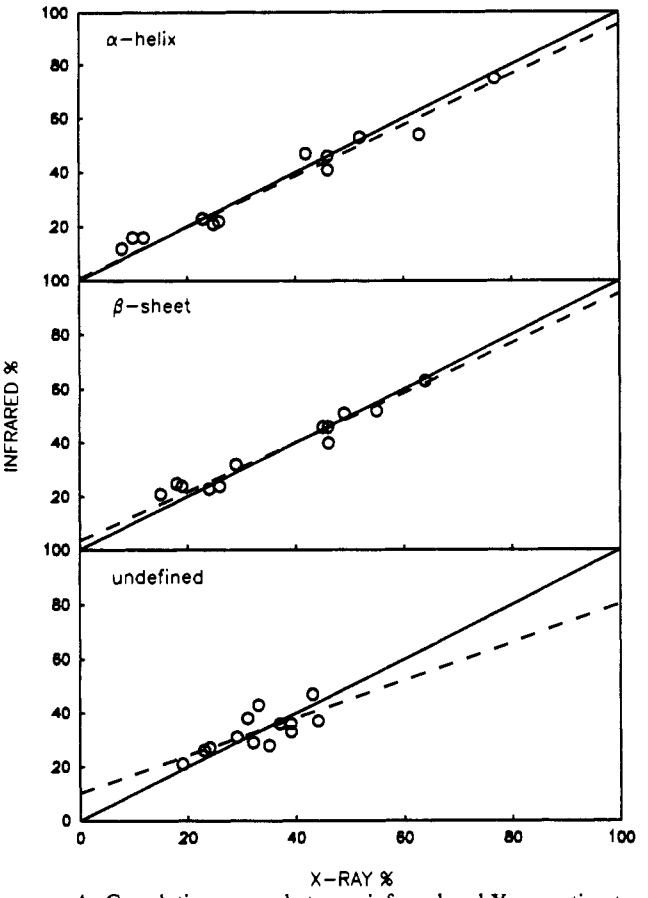

FIGURE 4: Correlation curves between infrared and X-ray estimates of the percentages of α-helix, β-sheet, and undefined conformation of the calibration proteins.

improvement over the results obtained with CLS. Figure 4 shows the correlation curves between the X-ray and infrared estimates for the secondary structure content. As seen, the correlation is very good for the α-helix and β-sheet content, the slopes of the correlation lines being 0.96 and 0.92, with ordinates of 0.4% and 3.3%, respectively. For the undefined structure, the correlation coefficient is much lower, 0.56, which is partly due to the narrow range of undefined structure content of the calibration proteins.

Attempts have also been made to improve the results either by adding turns to the types of secondary structure or by using only the amide I band as suggested by Lee et al. (1989). As seen in Table III, most of the infrared estimates are less accurate when turns are introduced in the calculation, most likely because this type of structure is not well defined in proteins and is best accounted for in the undefined category. Similarly, less accurate estimates were also obtained when only the amide I band was used to predict the secondary structure of proteins (Table II, method 5), even though this band is more sensitive to the conformation of proteins than the amide II feature. We believe that this is due to the fact that the amide I and amide II bands overlap in the 1600-cm^{-1} region and also because more data points are used in the calculation when both the amide I and amide II bands are analyzed.

Finally, since most infrared studies of proteins that have been published so far have been carried out in ^2H_2O and also because in some cases it is difficult to subtract quantitatively the water spectrum from those of proteins, e.g., for highly charged proteins or for lipid–protein complexes, spectra of the calibration proteins in solution in ^2H_2O were also recorded. After the correction for the combination band of ^2H_2O, the amide I′ band between 1600 and 1720 cm^{-1} was analyzed with PLS. As seen in Table II (method 6) and Table III, the highest differences between the infrared and X-ray estimates for the secondary structure content were obtained with this method of analysis.

Discussion

The above results show clearly that the best method to estimate the secondary structure content of proteins in solution from their infrared spectra is to consider the amide I and II region between 1480 and 1720 cm^{-1} after normalization of the integrated intensity between these limits and to analyze the data with the partial least-squares method, taking into account ordered and disordered helices, including only one type of β-sheet, and including turns in the undefined structure.

Two main assumptions were made for the development of such a secondary structure prediction method: first, the secondary structure of the proteins in solution is considered to be the same as that determined by X-ray diffraction for the crystalline form, and second, the integrated molar extinction coefficient of the amide I and amide II bands is assumed to be constant for each type of structure.

The first assumption is generally accepted since hydrated crystals are used for X-ray structure determination. In addition, Yu and Jo (1973) have shown that the Raman spectra of globular proteins like ribonuclease A and carboxypeptidase A in solution and in crystals are quite similar. Minor differences are observed for some bands due to the side chains of the proteins, suggesting that their environment is not exactly the same in the two states. However, the bands due to the amide I and III backbone vibrations are identical for the proteins in the crystalline state and in solution. Therefore, the secondary structure of proteins in crystals depicts well that of proteins in solution.

By normalizing the spectra in the amide I and II region so that the sum of the absorbances between 1480 and 1720 cm^{-1} is equal to unity, the assumption that the integrated molar absorptivities are nearly constant for each type of structure was made. In order to verify this hypothesis, the spectra of polylysine in the α-helix and β-sheet conformations were recorded by varying the temperature of the solution without opening the sample cell. The results have shown that the integrated intensities of the amide I and II bands are almost the same for the two types of structure. Therefore, the normalization of the spectra seems well justified. On the other hand, the determination of infrared molar absorptivities for the amide bands of the different types of structure is very difficult, since it has to rely on protein concentrations that are determined by ultraviolet absorption spectroscopy. In fact, for each calibration protein, several ultraviolet molar absorptivities are found in the literature, so that concentration errors can be introduced in the calculation. In addition, the use of infrared molar absorptivities would have eliminated the determination of the secondary content of proteins of unknown ultraviolet molar absorptivities. Finally, the good agreement with the X-ray data obtained here indicates that the above assumption is reasonably valid.

The choice of the calibration proteins is one of the most important steps in the development of a method for the estimation of the secondary structure content of proteins. The conformation of the chosen proteins must be well-known and must also span a broad percentage range for each type of structure. The calibration proteins chosen in this paper have also been used for the development of structure prediction methods using either circular dichroism (Chang et al., 1978; Provencher & Glöckner, 1981) or Raman spectroscopy (Williams & Dunker, 1981; Berjot et al., 1987). Their sec-

ondary structure has been well described by Levitt and Greer (1977), and the percentage range covered for each type of structure represents well the range expected for most globular proteins. Polylysine in the α-helix conformation has also been used as a calibration protein for the ordered helix, which has improved considerably the accuracy of the infrared estimates.

The distinction made between ordered and disordered helices is first based on an experimental result. Indeed, highly α-helical proteins, such as myoglobin and tropomyosin, have infrared amide bands that are similar in shape but whose maximum absorbances are at different frequencies (Figure 3). The major difference between these two proteins is that tropomyosin is composed of long strands of extended helices while myoglobin contains small regions of α-helices separated by regions containing predominantly turns or undefined conformation. Van Wart and Scheraga (1978) have suggested that the vibrational frequencies of the amide groups at the end of helical segments are different from those in the center. They have separated end-helical amides into two classes: $\alpha(-,+)$, when the amide NH groups is hydrogen-bonded to a third neighbor amide group while the amide CO is not hydrogen-bonded, and $\alpha(+,-)$, when the CO group is hydrogen-bonded to a third neighbor amide in the helix while the NH group is not hydrogen-bonded. Since these end-groups are most likely hydrogen-bonded to water, their characteristic amide frequencies are significantly different from those of the center $\alpha(+,+)$ amide groups. Differences in the frequency of the amide I and II vibrations between helical homopolypeptides and helices in real proteins have been recognized for some time and have been assigned to differences in symmetry and restrictions of the crystal selection rules (Lord, 1971; Chen & Lord, 1974). However, proteins with long helical segments may contain short regions of helix that vibrationally resemble homopolypeptides. Miyazawa (1960) has also shown that the intrachain hydrogen-bonding interation between each pair of third-neighboring residues in the α-helix is important in determining the frequeny of the amide I band. Finally, Williams (Williams & Dunker, 1981; Williams, 1983) and Berjot et al. (1987) have also observed that two types of helices are necessary to obtain good estimation of the α-helical content of proteins by Raman spectroscopy. The results of this study also emphasize the need for introducing ordered and disordered helices in the analysis of the conformation of proteins by vibrational spectroscopy.

In the spectral analysis, no distinction was made between parallel and antiparallel β-sheets. The differences between these two conformations is essentially the length of the hydrogen bonds, which are longer in the parallel pleated sheet conformation. Since the set of calibration proteins is composed of globular proteins in which part of the peptide chain could be hydrogen-bonded with a neighboring chain in a parallel pleated sheet and with another neighbor in an antiparallel pleated sheet, the distinction between the protein residues that are involved in parallel and antiparallel β-sheet is very difficult to make. Turns were included in the undefined conformation because of the uncertainty of the position of their infrared bands. According to Krimm and Bandekar (1986), broad frequency distributions are predicted for the amide bands of turns since, even for a given type of β-turn, the frequency of the amide bands varies with the dihedral angles. Thus, while α-helix and β-sheet components of a protein are expected to give relatively constant amide I frequencies, the same is not likely to be true for the β-turn component. In fact, infrared spectroscopy is not accurate enough to allow the determination of two types of β-sheet and of turns in proteins in solution. Our results show that, by restricting the number of conformations, a better accuracy is obtained for the estimation of the secondary structure content of proteins.

Our results show also that the mathematical method used for the calculation of the secondary structure of proteins is also very important. The two methods chosen here, i.e., CLS and PLS, are full-spectrum methods, since all the spectral frequencies are used. Techniques using CLS (Antoon et al., 1977) have been successfully applied to the quantitative analysis of multicomponent mixtures, even in cases where there is complete overlap of the infrared spectral features. The inclusion of all the data in the spectral region of interest also significantly improves the precision and accuracy of the results. One advantage of CLS is that the least-squares estimates of the pure-structure spectra provide useful information about the average vibrational frequency distribution of the amide groups assigned to classes of structure and on the significance of a class of structure. However, with this method, all interfering components in the spectral region of interest need to be known and included in the calibration, which is not an easy task for complex spectra such as those of proteins. In addition, the two matrix inversions required by this method are a major source of errors. Our results show that PLS is much more efficient than CLS for determination of the secondary structure content of proteins. The major advantage of PLS is that the number of conformations that is analyzed may be smaller than the total number of conformations in the proteins if the calibration proteins contain variable amounts of all structure types. In fact, the analysis can be performed one conformation at a time. In addition, with PLS only one matrix inversion is necessary, and PLS generates loading vectors that are linear combinations of calibration spectra. The resulting data compression distributes the noise throughout all loading vectors, while the true spectral variation is generally concentrated in the early loading vectors. Therefore, the interference due to the noise in the spectra, such as residual water vapor bands, is less important for the structure prediction in PLS than in CLS.

Even though it is well-known that the amide I band is very sensitive to the conformation of proteins, the amide II band has hardly ever been used to determine the structure content of proteins. The reason is that most studies were undertaken with ^2H_2O solutions, which shift the amide II band under the band due to the bending mode of ^2H_2O. In addition, several bands due to the absorption of the side chains of proteins underlie the amide II region, such as the bands due to the ionized carbonyl group of aspartic acid at about 1584 cm^{-1} or the ionized side-chain group of arginine at about 1577 and 1600 cm^{-1}. A shoulder on the amide II band of proteins, assigned to the absorption of the nonionized form of the tyrosine residues (Chirgadze et al., 1975), is also often observed. Surprisingly, despite the side-chain absorptions, the best prediction for the structure content of proteins is obtained when the amide II band is included in the spectral region used. In fact, the side-chain contribution seems to be minimized since a straight baseline is drawn between 1480 and 1720 cm^{-1}. If a flat baseline is subtracted between the same limits, the standard deviation between the helix content estimated by infrared spectroscopy and X-ray diffraction increases from 4.9% to 8.5%.

The method proposed in this paper allows the determination of the secondary structure content of proteins in H_2O solutions. This is now feasible because of the development of an algorithm for the water spectrum subtraction (Dousseau et al., 1989; Powell et al., 1986). However, the subtraction of the

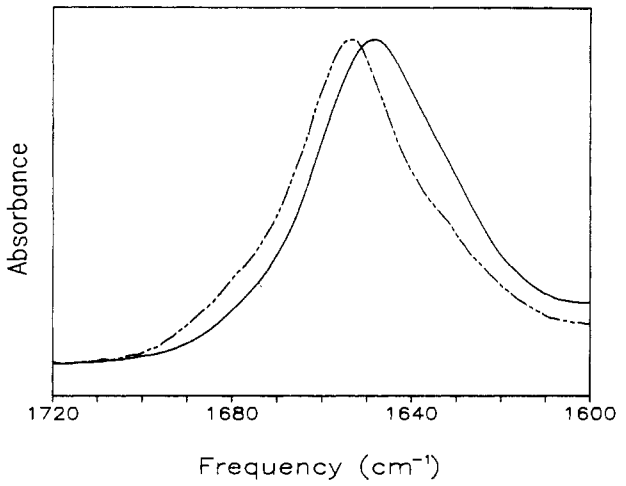

FIGURE 5: Infrared spectra of myoglobin in H_2O (- - -) and ^2H_2O (—) solutions.

water absorption from the spectra of lipid–protein complexes is very difficult with this algorithm, since the water spectrum is modified in the presence of phospholipids. Lafleur et al. (1989) have established that the profile of the Raman band due to the O–H stretching modes depends strongly on the extent of the domains in which the water molecules are trapped. Since in lipid dispersions part of the water molecules are trapped between bilayers, similar effects are expected for the infrared band due to the bending mode of water. In fact, we have found that in the presence of polar lipids, or even for highly charged proteins, the amide I band is never well corrected for the water contribution. In such cases, proteins have to be studied in ^2H_2O solutions.

The PLS method was, therefore, applied to the spectra of proteins in ^2H_2O solutions. As seen in Table II (method 6), infrared estimates from the amide I′ band are those showing the greatest differences with those obtained by X-ray diffraction. The reduction of the spectral region may account for part of this lack of accuracy. Indeed, even with proteins in H_2O solutions, when the spectral region used for the prediction is reduced, i.e., when the amide I band alone is used, the results are not as good as those obtained from both the amide I and II bands. In addition, the results show that the estimation for the α-helix content in proteins from the amide I′ band is worse than the estimation for the other types of structure. Eckert et al. (1977) have also encountered the same difficulties for the estimation of the secondary structure of proteins from their infrared amide I′ band. They have obtained a relatively good estimation for the β-sheet structure, as we do, but did not succeed in predicting accurately the α-helix content. The fact that the infrared amide I band for the α-helix is essentially due to two vibrational modes might be at the origin of this problem. A small downshift of the amide I frequency occurs when the amide hydrogen is exchanged by a deuterium, since the amide I modes have a small N–H in-plane bend contribution (Miyazawa, 1958; Rey-Lafon et al., 1973). On deuteration, the frequency shift of the two infrared-active amide I modes may not be necessary equal, which would results in a broadening of the amide I′ band. This broadening is clearly seen in Figure 5, where the amide I and I′ bands of myoglobin are compared. Sen and Keiderling (1984) have also observed that in vibrational circular dichroism the amide I and I′ bands of α-helical polypeptides have different shapes. This broadening of the amide I′ band can lead to a loss of precision in the estimation of the secondary structure content of proteins. Finally, the H–D exchange may never be entirely complete without some denaturation of the protein in ^2H_2O solutions. Indeed, some hydrogen atoms remain unattainable.

At this point it is interesting to compare frequency-based methods for the analysis of infrared amide bands (Dong et al., 1990; Surewicz & Mantsch, 1988; Byler & Susi, 1986) with that proposed in this paper. In order to compare these results with those of Table III, we have calculated the standard deviations on the differences between infrared and X-ray estimates found in these papers. Such calculations were not made for the method of Surewicz and Mantsch (1988), since it has not yet been applied to a large enough number of proteins. The standard deviations obtained by Dong et al. (1990) for the α-helix and β-sheet prediction are 5.2% and 5.9%, respectively, which are slightly higher than the values of 4.9% and 3.7%, respetively, obtained by our method (Table III). On the other hand, the standard deviations found by Byler and Susi (1986) are 2.4% for the α-helix and 3.1% for the β-sheet, which are better than our results. However, it should be kept in mind that these frequency-based methods involve major assumptions that may influence significantly the final results. For example, they rely on peak assignments of either second derivative or deconvolved spectra, which is not always an easy task for complex spectra such as those of proteins. In addition, to use these methods, one has to guess either the frequency limits of each band in the case of second derivative spectra or the band shapes in the case of curve-fitted deconvolved spectra. The method described in this paper, which is more a pattern recognition method than a frequency-based method, does not require any of these assumptions and should, therefore, be easier to apply to proteins of unknown structure.

The results presented in this paper demonstrate clearly that infrared spectroscopy offers a good alternative to the widely used circular dichroism. As a matter of fact, Provencher and Glöckner (1981) have obtained root-mean-square deviations of their estimates from X-ray values of 5% and 6% for the α-helix and β-sheet structures, respectively. The accuracy of our estimation for the α-helix content of proteins (4.9%) is comparable to that of CD. On the other hand, our results for the determination of the β-sheet content (3.7%) represent a substantial improvement over existing CD methods.

The estimation of the secondary structure content of proteins in H_2O solutions by vibrational spectroscopy was first developed for Raman spectorscopy, as the water contribution in the amide I region is much weaker in Raman than in infrared spectra. The latest results obtained by Williams (1986), using Raman spectroscopy and an unconstrained least-squares approach, show that the accuracy for the determination of α-helix and β-sheet structures if 4% and 3%, respectively, which is slightly better than the present infrared results. On the other hand, it is often very difficult to obtain Raman spectra of proteins with a good signal-to-noise ratio because of interfering luminescence background. Therefore, infrared spectroscopy offers a method of choice, since sample luminescence does not interfere in the spectra as it does for Raman spectroscopy.

Supplementary Material Available

Two tables giving spectra of the calibration proteins in the 1480–1720-cm^{-1} region, in which the spectra are water corrected, the baseline is subtracted, and the intensity is normalized to 1 (10 pages). Ordering information is given on any current masthead page.

References

Antoon, M. K., Koenig, J. H., & Koenig, J. L. (1977) *Appl. Spectrosc. 31*, 518–524.

Berjot, M., Marx, J., & Alix, A. J. P. (1987) *J. Raman Spectrosc. 18*, 289–300.
Bussian, B. M., & Sander, C. (1989) *Biochemistry 28*, 4271–4277.
Byler, D. M., & Susi, H. (1986) *Biopolymers 25*, 469–487.
Chang, T. C., Wu, C. S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13–31.
Chen, M. C., & Lord, R. C. (1974) *J. Am. Chem. Soc. 96*, 4750–4752.
Chirgadze, Y. N., Fedorov, O. V., & Trushina, N. P. (1975) *Biopolymers 14*, 679–694.
Dong, A., Huang, P., & Caughey, W. S. (1990) *Biochemistry 29*, 3303–3308.
Dousseau, F., Therrien, M., & Pézolet, M. (1989) *Appl. Spectrosc. 43*, 538–542.
Eckert, K., Grosse, R., Malur, J., & Repke, K. R. H. (1977) *Biopolymers 16*, 2549–2563.
Elliot, A., & Ambrose, E. J. (1950) *Nature 165*, 921–922.
Haaland, D. M., & Thomas, E. V. (1988) *Anal. Chem. 60*, 1193–1202.
Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.
Krimm, S., & Bandekar, J. (1986) *Adv. Protein Chem. 38*. 181–364.
Lafleur, M., Pigeon, M., Pézolet, M., & Caillé, J.-P. (1989) *J. Chem. Phys. 93*, 1522–1526.
Lee, D. C., Haris, P. I., Chapman, D., & Mitchell, R. C. (1989) in *Spectroscopy of Biological Molecules: State of the Art* (Bertoluzza, A., Fagnano, C., & Monti, P., Eds.) pp 57–58, Società Editrice Escullapio, Bologna, Italy.
Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–293.
Lindberg, W., Persson, J.-A., & Wold, S. (1983) *Anal. Chem. 554*, 643–648.
Lippert, J. L., Tyminsky, D., & Desmeules, P. J. (1976) *J. Am. Chem. Soc. 98*, 7075–7080.
Lord, R. C. (1971) *Pure Appl. Chem. Suppl. 7*, 179–191.
Miyazawa, T. (1958) *J. Chem. Phys. 29*, 246–248.
Miyazawa, T. (1960) *J. Chem. Phys. 32*, 1647–1652.
Miyazawa, T., & Blout, E. R. (1961) *J. Am. Chem. Soc. 83*, 712–719.
Pézolet, M., Pigeon-Gosselin, M., & Coulombe, L. (1976) *Biochim. Biophys. Acta 453*, 502–512.
Pézolet, M., Boulé, B., & Bourque, D. (1983) *Rev. Sci. Instrum. 54*, 1364–1367.
Powell, J. R., Wasacz, F. M., & Jakobsen, R. J. (1986) *Appl. Spectrosc. 40*, 339–344.
Provencher, S. W., & Glöckner, J. (1981) *Biochemisry 20*, 33–37.
Rey-Lafon, M., Forel, M. T., & Garrigou-Lagrange, C. (1973) *Spectrochim. Acta Part A 29*, 471–486.
Sen, A. C., & Keiderling, T. A. (1984) *Biopolymers 23*, 1519–1532.
Surewicz, W. K., & Mantch, H. H. (1988) *Biochim. Biophys. Acta 952*, 115–130.
Thomas, G. J., Jr., & Agard, D. A. (1984) *Biophys. J. 46*, 763–768.
Van Wart, H. E., & Scheraga, H. A. (1978) *Methods Enzymol. 49*, 67–149.
Williams, R. W. (1983) *J. Mol. Biol. 166*, 581–603.
Williams, R. W. (1986) *Methods Enzymol. 130*, 311–331.
Williams, R. W., & Dunker, A. K. (1981) *J. Mol. Biol. 152*, 783–813.
Yu, N. T., & Jo, B. H. (1973) *J. Am. Chem. Soc. 95*, 5033–5037.

Nucleotide Sequence and DNA Recognition Elements of *alc*, the Structural Gene Which Encodes Allantoicase, a Purine Catabolic Enzyme of *Neurospora crassa*[†,‡]

Hakjoo Lee, Ying-Hui Fu, and George A. Marzluf*
Department of Biochemistry, The Ohio State University, Columbus, Ohio 43210
Received April 25, 1990; Revised Manuscript Received June 20, 1990

ABSTRACT: The nitrogen regulatory circuit of *Neurospora crassa* contains structural genes that encode nitrogen catabolic enzymes which are subject to complex genetic and metabolic regulation. This set of genes is controlled by nitrogen limitation, by specific induction, and by the action of *nit*-2, a major positive-acting regulatory gene, and *nmr*, a negative-acting control gene. The complete nucleotide sequence of *alc*, the gene that encodes allantoicase, a purine catabolic enzyme, is presented. The *alc* gene contains a single intron, is transcribed from two initiation sites situated approximately 50 nb upstream of the translation start site, and encodes a protein comprised of 354 amino acids. Mobility shift and DNA footprint experiments identified a single binding site for the NIT2 regulatory protein in the *alc* promoter region. The binding site contains a 10 nucleotide base pair symmetrical sequence which is flanked by two possible core binding sequences, TATCT and TATCG. Mutant NIT2/β-gal fusion proteins with amino acid substitutions in a putative zinc-finger motif were shown to be completely deficient in the ability to bind to the *alc* promoter DNA fragment.

Neurospora crassa utilizes purines as secondary nitrogen sources, when primary sources such as glutamine and ammonia are limiting. Expression of the purine catabolic genes of *N. crassa* is controlled by uric acid induction, a pathway-specific regulation (Reinert & Marzluf, 1975). The purine catabolic genes and other structural genes of the nitrogen regulatory circuit of *N. crassa* are also controlled by nitrogen repression. Two regulatory genes, *nit*-2, a major positive control gene, and *nmr*, a negative control gene, together mediate nitrogen repression.

[†] This research was supported by Public Health Service Grant GM-23367 from the National Institutes of Health.
[‡] The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02927.
* Corresponding author.

0006-2960/90/0429-8779$02.50/0