Removal of Spectral Noise in the Quantitation of Protein Structure Through Infrared Band Decomposition

IZASKUN ECHABE,[1] JOSÉ ANTONIO ENCINAR,[2] JOSÉ LUIS R. ARRONDO[1]

[1] Departamento de Bioquímica, Univ. País Vasco, P.O. Box 644, 48080 Bilbao, Spain

[2] Departamento de Neuroquímica, Universidad de Alicante, Campus de S. Juan, Apdo. 374, 03080 Alicante, Spain

Received 20 January 1997; revised 13 March 1997; accepted 29 May 1997

ABSTRACT: The underlying noise in the infrared spectra of proteins may introduce artifacts in the quantitation of proteins by curve-fitting of the amide I band. Smoothing methods are able to reduce the noise but can introduce alterations in band shape that affect the information contained in the spectrum. Three methods to remove noise—Savitzky–Golay, Fourier filtering, and maximum entropy—have been used to ascertain their influence on the quantitative information when applied to protein bands. Use of artificial curves shows that whereas Savitzky–Golay and Fourier smoothing are able to reduce the noise, they distort the band shape. Maximum entropy is more efficient in reducing the noise in artificial curves with added noise, and provided a narrowest bandwidth below 12 cm^{-1}, no band-shape distortion is obtained. Using the smoothing in natural spectra, the presence of spurious bands in the initial parameters coming from artifacts introduced by deconvolution or derivation is reduced. Moreover, the dispersion in the percentage area values in a series of similar spectra is also decreased below 2%, a value that discriminates the effect of ligand binding to proteins. The maximum entropy method is proposed as a tool to improve the quantification of protein structure by infrared spectroscopy. © 1997 John Wiley & Sons, Inc. Biospectroscopy **3:** 469–475, 1997

Keywords: infrared spectroscopy; noise; smoothing; curve fitting; protein-structure quantitation

INTRODUCTION

Knowledge of protein structure is helped by spectroscopic techniques that supply information on the molecular structure and dynamics, in addition to techniques giving three-dimensional information, such as X-ray or NMR. Infrared spectroscopy is acknowledged to provide quantitative information on protein structure.[1] The infrared bands arising from the peptide bond, the so-called amide bands, are sensitive to the backbone conformation (i.e., dihedral angles) and to the hydrogen bonding of the protein. The amide I band located between 1700 and 1600 cm^{-1}, and arising mainly from peptidic C=O stretching vibrations, is most extensively used in structural studies.

Quantitative methods involve extraction of the information contained in the overlapping component bands that constitute the amide I band envelope. To separate these contributions, several mathematical techniques have been adopted, namely band narrowing and curve-fitting procedures. After component resolution, each band is associated to a structural feature of the protein. Derivation and deconvolution transform the absorption bands into narrower line shapes, thereby resolving the overlapping components. Curve-fitting is the process of regenerating a measured spectrum by mathematically coadding a predefined number of bands with known peak positions,

Correspondence to: J. L. R. Arrondo.

Biospectroscopy, Vol. 3, 469–475 (1997)

obtained from deconvolved and/or derivative spectra. These initial parameters (band positions) together with an estimation of bandwidth and line shape are iterated until the theoretically generated and the measured spectra coincide. Curve-fitting can be performed on the original amide I or in the mathematically narrowed spectrum. However, deconvolution and derivation changes the line shape and can affect the information contained in the amide I.[2] Because of the shapelessness of amide I, the solution obtained by curve-fitting the original amide I may not be unique, and restrictions to the possible solutions must be applied.[3] The most important solution is that the number and position of bands obtained after iteration must be the same as in the initial assumption.

Obtention of the number and position of bands from deconvolution is affected by the underlying noise present in the spectrum, although not always visible. Thus, it is easy to introduce artifacts in the initial values that can be interpreted as real components. This obstacle has produced a debate on the methodology used to quantify protein structure from an infrared spectrum.[4,5] To improve the signal-to-noise ratio of spectroscopic data, smoothing methods are commonly in use, such as Fourier filtering[6] or the procedure developed by Savitzky and Golay,[7] which is a polynomial method based on the least-squares criterion. These smoothing procedures, however, often induce distorsions in line shape, thus affecting the quantitative information contained in the amide I band of proteins. In the present work, we applied a novel smoothing method based on maximum entropy[8] together with the Savitzky–Golay and Fourier methods to artificial curves mimicking amide I bands with and without added noise to establish the effect of the routines on band shape. Then, we applied the values obtained to real protein bands to evaluate the ability of smoothing in removing the underlying spectral noise and in eliminating the dispersion of values in protein structure quantitation.

MATERIALS AND METHODS

Obtention of Artificial Curves and Protein Spectra

Artificial curves in the interval 1900–1400 cm^{-1}, similar to the ones obtained in natural protein spectra, were constructed using GRAMS (Galactic Inc., Salem, NH). Seven component bands, similar to a curve-fitting of a protein and centered at 1684, 1675, 1655, 1636, 1625, and 1612 cm^{-1}, were added to generate a noise-free amide I. In the study of the efficiency of smoothing methods in removal of noise, a random noise generated by the Spectra-Calc software (Galactic Inc.) was added to give signal-to-noise ratios of 1000 : 1 and 10000 : 1.

Cytochrome *bf*, the protein used in the study, was a gift from Dr. Rich and the spectra were obtained in a Nicolet Magna 550 spectrometer. Typically, 304 scans were collected at 4-cm^{-1} resolution using the Series and Rapid Scan software running under OMNIC (Nicolet Corp., Madison, WI). The interferograms were reprocessed with 1 level zero filling. Measurements were carried out in D_2O buffer using a thermostated cell with a continuous heating of 1°C/min. Ten spectra in the temperature range where no structural changes are produced were used in the study. Buffer contribution was subtracted and the resulting spectra had a signal-to-noise ratio better than 1500 : 1.

Smoothing Methods

Savitzky–Golay polynomial smoothing and Fourier filtering were applied using SpectraCalc or Grams, respectively. The maximum entropy smoothing was from Spectrum Squares Associates (Ithaca, NY) running under SpectraCalc. Different combinations of polynomial degree and number of points were used for the Savitzky–Golay method (i.e., a degree ranging from 2 to 4 and a number of points from 11 to 16). Fourier filtering was applied truncating the interferogram from a 20% up to a 70%. A truncation of 70% is the default for Fourier derivation.[9] In the maximum entropy method, noise was assumed to have a normal distribution, initial band shape was assumed to be either Gaussian (as recommended by Spectrum Squares for components of unknown band shape) or Lorentzian; the minimum bandwidth ranged from 5 to 20 cm^{-1}, with higher bandwidths totally distorting the spectral band shape.

The smoothing methods were applied to the artificial and protein curves. Changes in band shape after smoothing were investigated in noise-free spectra by subtracting the smoothed minus the original curve. If no difference in band shape is produced, a straight line is obtained. The efficiency of smoothing in removing noise is studied similarly (i.e. after smoothing); the subtraction of the original curve should give the noise introduced.

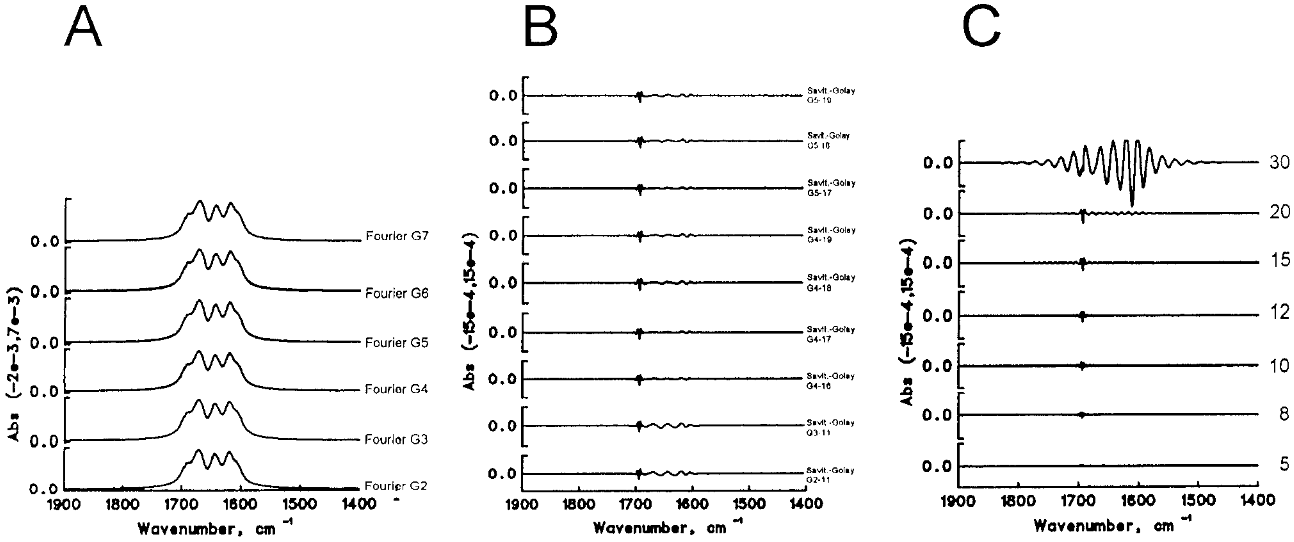

Figure 1. Difference spectra obtained by subtracting a smoothed curve from the noise-free artificial curve constructed. (A) Result obtained using Fourier smoothing with a breaking point ranging from 20% to 70% of the interferogram. (B) Result obtained using a Savitzky–Golay polynomial function with an exponent ranging from 2 to 5 and with the number of points from 11 to 19. (C) Maximum entropy smoothing with bandwidths from 5 to 30.

Curve-Fitting of the Spectra

Decomposition of the spectra was carried out on the original spectra and after smoothing of the curves using different conditions. The curve-fitting method has been described previously.[10] Briefly, the number and position of the bands is obtained from the deconvolved spectra, and the band shape was initially set at 10% Lorentzian and left to vary in the iteration procedure. However, it was verified that the components of the spectra smoothed with the maximum entropy method were Gaussian after iteration, and in this case, if the band shape is set to Gaussian, a better fit and less dispersion in the value of the areas are obtained in the software used.

To check the variability in the results produced by the smoothing methods, for each component band, the mean value for the 10 spectra selected is obtained and the sum (S) of the standard deviations of all the component bands is used as a parameter of goodness in eliminating the variability of the data.

RESULTS AND DISCUSSION

Preservation of Band Shape

The effect of the smoothing methods on the spectral band shape is studied by constructing an artificial curve free of noise similar to the protein am-

Table I. Signal-to-Noise Ratio of the Artificial Spectrum, With Added Noise, and After Subjecting Them to the Smoothing Functions

Original	0.1% Noise	Savitzky–Golay		Fourier		Maximum Entropy
∞	1200	2–11[a]	5300	0.2[b]	2500	∞
		3–11[a]	5300	0.3[b]	3500	
		4–11[a]	4400	0.6[b]	5300	
		5–17[a]	4400	0.7[b]	7300	
		5–19[a]	5700			

[a] Refers to the polynomial function and the number of points used in smoothing.
[b] Expresses the fraction of interferogram truncated upon smoothing.

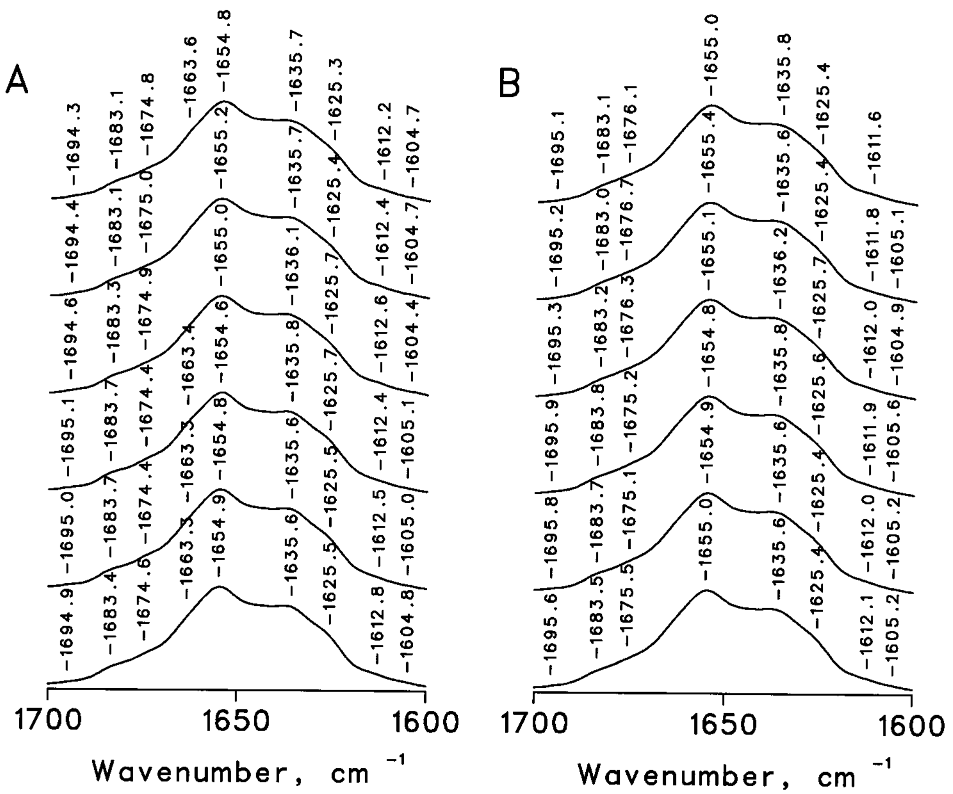

Figure 2. Band position of a series of deconvolved spectra of cytochrome *b f* before (A) and after (B) smoothing using a maximum entropy function with a width of 12 cm^{-1}.

ide I. It is then subjected to the different methods of noise removal, and finally, the original is subtracted from the smoothed curve and changes in band shape are detected as nonstraight lines. Savitzky–Golay, Fourier filtering and maximum entropy methods have been considered as noise-removal methods. Because the spectrum is noiseless, the result of a smoothing method nonperturbing the band shape should be a straight line. Figure 1 shows the result of applying the three different smoothing techniques considered using different parameters. It is clear that after Fourier smoothing, a difference spectrum, and not a straight line, is produced irrespective of the parameters used, indicating that this smoothing approach produces changes in the spectral band shape. Also, Savitzky–Golay smoothing produces differences in band shape, although less pronounced than in Fourier filtering. The maximum entropy algorithm hardly affects band shape if a line width below 12 cm^{-1} is used. Band-shape preservation using a line width below 12 cm^{-1} is observed in all variants of the maximum entropy algorithm used. The spike observed at 1700 cm^{-1} in the different smoothing techniques is an artifact that only originates in the artificial curves and is not present in natural spectra.

Noise Removal

The efficiency of smoothing in removing noise has been studied by adding to the artificial curves a random noise with final signal-to-noise ratios of 1000 : 1 and 10000 : 1. After smoothing of the curves, the signal-to-noise ratio is measured again and the results are shown in Table I. The best function in removing noise is maximum entropy, which even restores the original noise-free curves. The presence of noise did not change the effect of the smoothing functions on band shape.

Studies with artificial curves show that even if Fourier filtering or Savitzky–Golay smoothing remove noise, their distorsion of the band envelope is higher than those produced by maximum entropy methods. In the latter, the line-width parameter for the narrower component of a band similar to the amide I is 12 cm^{-1}.

Effect of Maximum Entropy Methods on the Decomposition Values of the Amide I Components of the Infrared Spectra

The critical step in curve-fitting the amide I band implies knowledge of the number and position of its components. This step is usually accomplished by using resolution-enhanced spectra and is affected by the presence of noise in the spectrum, so that an extra (noisy) component in the amide I band creates an artifactual result. Besides, small variations in band position produce an error that can imply a large number of amino acids in the protein. In fact, in a protein with a molecular mass of around 100 kd, a 5% error, which is common in curve-fitting,[5] involves ≈ 50 amino acids. To improve the dispersion in percentage area of the components produced by these variations in band position, a procedure has been proposed consisting of measuring the spectra in a narrow temperature range where protein conformation is not affected and comparing the band position values obtained.[11] However, even in these spectral series, spurious bands or changes in band position cannot be definitely excluded. Figure 2 shows the band positions of a series of 10 deconvolved spectra of a protein before and after being treated with a maximum entropy smoothing. Before smoothing, it is unclear whether the band around 1667 cm^{-1} is a true band or a noise-produced artifact because it is present in some spectra but not in the others. After smoothing, this band disappears from all spectra, suggesting that it was a product of noise. The indication can be corroborated by the results obtained with other samples of the same protein, where this band is absent, and also by comparing the results of the curve-fitting in H_2O and D_2O that must be coherent. A similar result has been obtained with other proteins. As long as the signal-to-noise ratio was high enough so as not to introduce incertitudes, maximum entropy smoothing did not change the number or position of the bands obtained.

Effect of Smoothing on the Decomposition Values of the Amide I Band

Spectral smoothing not only removes noise from the spectra but also reconstructs the band envelope, avoiding small imperfections in band shape that can introduce a dispersion in percentage area values and increase error in the measurements.

A series of 10 spectra corresponding to the protein measured in an interval not involving conformational changes has been smoothed and the results of the fitting compared with the "crude" spectra. Figure 3 shows the different values obtained for the original components of the amide I and after smoothing by the maximum entropy method. Looking at the regression lines, it can be appreciated that in the most important bands corresponding to α-helix and β-sheet, the scattering of experimental points is smaller in the treated spectrum. In fact, the dispersion of the data lowers from 4–5% to 1–2%. The example shown corresponds to a high-quality spectrum with a signal-

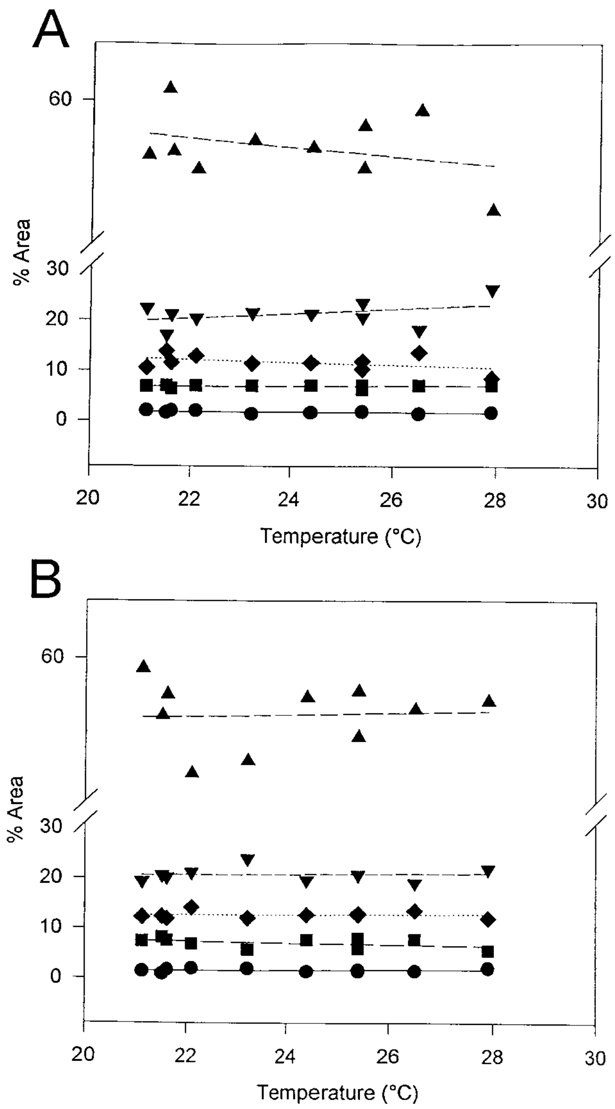

Figure 3. Values corresponding to percentage area of the components around 1657 (▲), 1632 (▼), 1620 (◆), 1678 (■), and 1687 (●) cm^{-1} obtained after curve-fitting the original (A) and the smoothed (B) spectra. The regression lines are also represented.

Table II. Values of Percentage Area and Standard Deviation Corresponding to the Curve-Fitting of the Original Amide Band and After Being Submitted to Smoothing

Method	1683 cm^{-1}	1675 cm^{-1}	1654 cm^{-1}	1635 cm^{-1}	1625 cm^{-1}	S
Original	1.8 ± 0.2	6.9 ± 0.2	58.3 ± 1.2	21.3 ± 2.6	11.6 ± 1.6	5.82
M. Entropy[a]	1.9 ± 0.7	6.9 ± 0.7	58.0 ± 1.1	20.7 ± 1.4	12.6 ± 0.7	4.6
Savitzky–Golay[b]	1.8 ± 0.7	7.0 ± 0.8	58.6 ± 1.1	19.9 ± 2.1	12.6 ± 1.4	5.98
Fourier[c]	2.3 ± 1.1	5.6 ± 1.7	60.3 ± 1.2	18.7 ± 1.9	13.1 ± 1.1	6.95

The parameter S represents the sum of the standard deviations. The best values have been represented for each smoothing method.

[a] Using a minimum bandwidth of 12 cm^{-1}.

[b] With a polynomial degree of 5 and 17 points.

[c] Truncating the interferogram at 20%.

to-noise ratio of 3000 : 1, which gives good quality fittings. In spectra with poorer signal-to-noise ratios, the improvement can be higher (data not shown). In the case presented, the dispersion is decreased by 50%. Reducing the error to 2% in the relative proportion of structural components in a protein makes small conformational variations, such as produced by ligand binding, amenable to study by infrared spectroscopy.

An overall view of the improvement after smoothing can be obtained by looking at the standard deviation of the components fitted. These deviations are added and its sum (S) will indicate a better overall fitting if the value is lower. Table II shows these values for the different smoothing methods used. In all cases, the lowest value of S is obtained using the maximum entropy smoothing.

Protein quantitation from the infrared spectra can also be obtained through factor analysis methods using the spectra of proteins with known structure to derive the "pure" structure spectrum.[12,13] The procedure described here can also be of help in this protein analysis method, because it is also dependent on the spectral band shape.

CONCLUSIONS

Amide I bands of protein spectra can have an underlying noise that will affect the number and position of component bands and the percentage area values. Introduction of a spurious band (i.e., noise) leads to a nonexistent protein structure or to an error in the assignment of the bands. The identification of a nonartifactual band is helped by obtaining the spectra in both H_2O and D_2O, by measuring different samples, and by removing the noise by a method that should not change the band shape and consequently the quantitative information contained in the amide I band. The use of artificial curves without and with added noise shows that Fourier filtering and Savitzky–Golay smoothing can remove the noise but change the band shape. Maximum entropy smoothing removes more efficiently the noise of an artificial amide I, and with a bandwidth for the narrower component below 12 cm^{-1} (as in proteins), no significant change in band shape is appreciated. Looking at original protein spectra, the efficiency of the maximum entropy smoothing in removing the spureous bands introduced by the noise is demonstrated. Moreover, a reduction in data dispersion is obtained, even in spectra of high signal-to-noise ratios. The results presented show that maximum entropy smoothing can be a tool to improve quantification of protein structure by infrared spectroscopy.

I.E. is supported by a predoctoral fellowship from the Basque Governement and J.A.E. is recipient of a predoctoral fellowship from the "Ministerio de Educación y Ciencia" of Spain. This work has been supported in part by grants PB91/0441 from DGICYT and EB-219/95 from University of Basque Country.

REFERENCES

1. J. L. R. Arrondo, A. Muga, J. Castresana, and F. M. Goñi, "Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy," *Prog. Biophys. Mol. Biol.,* **59,** 23–56 (1993).
2. D. G. Cameron and D. J. Moffat, "Deconvolution, derivation, and smoothing of spectra using Fourier transforms," *J. Test. Eval.,* **12,** 78–85 (1984).
3. J. Castresana, A. Muga, and J. L. R. Arrondo, "The structure of proteins in aqueous solutions. An assessment of triose phosphate isomerase structure

by Fourier-transform infrared spectroscopy," *Biochem. Biophys. Res. Commun.,* **152,** 69–75 (1988).

4. M. Jackson and H. H. Mantsch, "The use and misuse of FTIR spectroscopy in the determination of protein structure," *Crit. Rev. Biochem. Mol. Biol.,* **30,** 95–120 (1995).
5. W. K. Surewicz, H. H. Mantsch, and D. Chapman, "Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assesment," *Biochemistry,* **32,** 389–394 (1993).
6. W. A. Cramer, S. E. Martinez, P. N. Furbacher, D. Huang, and J. L. Smith, "The cytochrome b_6f complex," *Curr. Opin. Struct. Biol.,* **4,** 536–544 (1994).
7. A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.,* **36,** 1627 (1964).
8. L. K. DeNoyer and J. G. Dodd, "Maximum likelihood smoothing of noisy data," *Am. Lab.,* **3,** 21–27 (1990).
9. D. J. Moffat and H. H. Mantsch, "Fourier resolution enhancement of infrared spectral data," *Methods Enzymol.,* **210,** 192–200 (1992).
10. S. Bañuelos, J. L. R. Arrondo, F. M. Goñi, and G. Pifat, "Surface-core relationships in human low density lipoprotein as studied by infrared spectroscopy," *J. Biol. Chem.,* **270,** 9192–9196 (1995).
11. J. L. R. Arrondo, J. Castresana, J. M. Valpuesta, and F. M. Goñi, "The structure and thermal denaturation of crystalline and non-crystalline cytochrome oxidase as studied by infrared spectroscopy," *Biochemistry* **33,** 11650–11655 (1994).
12. D. C. Lee, P. I. Haris, D. Chapman, and R. C. Mitchell, "Determination of protein secondary structure using factor analysis of infrared spectra," *Biochemistry,* **29,** 502–512 (1990).
13. S. Corbalán-García, J. A. Teruel, J. Villalaín, and J. C. Gómez-Fernández, "Extensive proteolytic digestion of the (Ca^{2+} + Mg^{2+})-ATPase from sarcoplasmic reticulum leads to a highly hydrophobic proteinaceous residue with a mainly α-helical structure," *Biochemistry,* **33,** 8247–8254 (1994).