Anal. Chem. **2002,** *74,* 3386−3391

Single-Pass Attenuated Total Reflection Fourier Transform Infrared Spectroscopy for the Prediction of Protein Secondary Structure

Brandye M. Smith, Lisa Oswald, and Stefan Franzen*

Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695-8204

Principal component regression (PCR) was applied to a spectral library of proteins in H_2O solution acquired by single-pass attenuated total reflectance (ATR) Fourier transform infrared (FT-IR) spectroscopy. PCR was used to predict the secondary structure content, principally α-helical and the *β*-sheet content, of proteins within a spectral library. Quantitation of protein secondary structure content was performed as a proof of principle that use of single-pass ATR-FT-IR is an appropriate method for protein secondary structure analysis. The ATR-FT-IR method permits acquisition of the entire spectral range from 700 to 3900 cm^{-1} without significant interference from water bands. An "inside model space" bootstrap and a genetic algorithm (GA) were used to improve prediction results. Specifically, the bootstrap was utilized to increase the number of replicates for adequate training and validation of the PCR model. The GA was used to optimize PCR parameters, particularly wavenumber selection. The use of the bootstrap allowed for adequate representation of variability in the amide A, amide B, and C−H stretching regions due to differing levels of sample hydration. Implementation of the bootstrap improved the robustness of the PCR models significantly; however, the use of a GA only slightly improved prediction results. Two spectral libraries are presented where one was better suited for *β*-sheet content prediction and the other for α-helix content prediction. The GA-optimized PCR method for α-helix content prediction utilized 120 wavenumbers within the amide I, II, A, B, and IV and the C−H stretching regions and 18 factors. For *β*-sheet content predictions, 580 wavenumbers within the amide I, II, A, and B and the C−H stretching regions and 18 factors were used. The validation results using these two methods yielded an average absolute error of 1.7% for α-helix content prediction and an average absolute error of 2.3% for *β*-sheet content prediction. After the PCR models were developed and validated, they were used to predict the α-helix and *β*-sheet content of two unknowns, casein and immunoglobulin G.

The substantial number of protein gene sequences determined from the completion of the human genome project provides a motive for the development of rapid protein secondary structure determination methods.[1] The importance of protein secondary structure prediction methods lies with the determination of protein function for the study of both biological pathways and the mechanism of disease.[1,2] Since a mere 3% of the determined protein gene sequences have known secondary structure,[1] there is a tremendous need for methods that rapidly classify proteins and that monitor protein interactions. Protein secondary structure refers to the organization of amino acid residues in a polypeptide chain and is predominantly composed of α-helical and *β*-sheet motifs.[3,4] Several experimental methods exist for protein secondary structure determination such as circular dichroism (CD), Fourier transform infrared (FT-IR) spectroscopy, nuclear magnetic resonance (NMR), Raman spectroscopy, and X-ray diffraction. The overwhelming majority of three-dimensional coordinates currently available in the protein data bank (PDB) were elucidated from either NMR or X-ray diffraction. The techniques of NMR and X-ray diffraction used to determine the coordinates of three-dimensional structure have not yet been applied to a large part of the proteome.[1,5] Therefore, CD, Raman spectroscopy, and FT-IR spectroscopy are routinely used to rapidly classify proteins according to secondary structure motifs.[1,5] Using these three key spectroscopic methods, spectral-based correlations of proteins with known secondary structure are used to construct calibration models for secondary structure prediction of proteins with unknown three-dimensional coordinates.

This paper is the second in a two-part study, where the first paper focused on the advantages of the single-pass attenuated total reflectance (ATR)-FT-IR technique for protein analysis. The method of single-pass ATR-FT-IR is valuable since it is a rapid technique that does not require protein exchange into D_2O or sample cell assembly. Protein in H_2O solution that is exposed to a N_2 environment is brought into contact with a germanium internal reflective element (IRE). Since the protein is in an N_2 environment, it slowly dehydrates into a concentrated gel state. Rapid scanning can continuously monitor the changes from a fully hydrated state to a concentrated gel state. Concentrating the

* Corresponding author: (telephone) 919 515-8915; (fax) 919 515-8909; (e-mail) Stefan_Franzen@ncsu.edu.

[1] Pelton, J. T.; McLean, L. R. *Analy. Biochem.* **2000**, *277*, 167−176.
[2] Maggio, E. T.; Ramnarayan, K. *Drug Discovery Today* **2001**, *6*, 996−1004.
[3] Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*, 2nd ed.; Worth Publishers: New York, 1993.
[4] Holde, K. E. v.; Johnson, W. C.; Ho, P. S. *Principles of Physical Biochemistry*; Prentice Hall: Upper Saddle River, NJ, 1998.
[5] Keiderling, T. In *Infrared and Raman Spectroscopy of Biological Materials*; Yan, B., Gremlich, H., Eds.; Marcel Dekker: New York, 2001; Vol. 24.

10.1021/ac020104n CCC: $22.00 © 2002 American Chemical Society
Published on Web 06/07/2002

protein to a gel state yields spectral enhancement such that protein amide bands can be observed simultaneously without performing H_2O subtraction.

All proteins have nine characteristic absorption amide bands, labeled amide A, B, and I–VII, in the mid-infrared that can be interpreted in terms of structure.[6,7] FT-IR has been successfully used for secondary structural analysis largely based on the examination of amide I, which results from C═O stretching.[6–13] Correlations between amide I and II band frequencies, principally amide I, and secondary structure are well established for proteins in H_2O.[6,7,10,14,15] Amide I bands occur at ~1650 cm^{-1} for primarily α-helical structures, whereas, the amide I band for primarily β-sheet proteins is shifted to a lower frequency of ~1630 cm^{-1}.[6,7,10,14,15] In addition, primarily β-sheet proteins have a weak component at 1675–1690 cm^{-1}. Amide II bands occur at 1550 cm^{-1} for α-helical proteins and 1530 cm^{-1} for β-sheets,[6,10,14,15] but the secondary structure correlations in this region are not well understood. Few correlations exist for the amide III, IV, A, and B regions. Rather than the technique of frequency assignment for the determination of protein secondary structure, multivariate techniques such as multiple linear regression (MLR), partial least squares (PLS), and principal component regression (PCR) have been used to yield a more quantitative assessment.[16–19] In this study, the focus was the development of a multivariate calibration model for the prediction of protein secondary structure.

The development of such a model involves training and validation phases. The development of a representative spectral library, one that includes all anticipated sources of signal variability, is crucial since the only variability that can be recognized is that included in the model. Once an adequate spectral library with sufficient variability is obtained, the prediction power of the model will be suitable for secondary structure determination of unknown proteins. Casein and immunoglobulin (IgG) have unidentified secondary structure content that will be predicted upon completion of the training and validation phases.

PCR has been applied to a library of single-pass ATR-FT-IR protein spectra in H_2O solution to predict α-helical and β-sheet content. Due to the small number of protein spectra, a bootstrap method was applied to enlarge the data set. Once an optimum training and test set were constructed, PCR was applied. GA optimization was performed on the models to improve the accuracy and robustness of the protein secondary structure prediction. The GA optimized wavenumber selection and the number of principal component factors included in the model. A data set that was not involved in the construction of the multivariate regression model, a validation set, was used to validate the model. Upon finding an ideal model, the secondary structure content was determined for the proteins casein and IgG. Thus, this study provides an accurate and rapid methodology for the prediction of α-helical and β-sheet content of single-pass ATR-FT-IR protein spectra.

EXPERIMENTAL SECTION

The proteins listed in Table 1 were prepared without further purification to a final concentration of ~3 mM in H_2O. In the experimental apparatus, the Ge crystal is at the focus of a Cassagranian objective in a UMA500 microscope (Digilab). A 10–20-μL sample was injected onto the Teflon block using a Wheaton pipet. The protein spectra were recorded at ambient temperature and averaged over 64 scans on a Digilab FTS 6000 FT-IR spectrometer equipped with a liquid nitrogen-cooled MCT detector. The protein spectra were recorded with a resolution of 2 cm^{-1}. Background spectra were obtained subsequently. Blowing a steady stream of N_2 gas over the Teflon block gently dehydrated the protein samples. Spectra were recorded continuously after the sample was deposited onto the Teflon block until a concentrated protein gel had formed onto the Ge IRE. The Ge IRE was rinsed with H_2O and allowed to dry prior to loading a subsequent protein sample. Throughout this study, the protein samples are referred to as being in the hydrated state, the intermediate state, and the gel state. As the names imply, the hydrated state refers to the protein sample when it is first deposited into the Teflon reservoir. The gel state refers to dehydrated sample, and intermediate states are observed during the 15–30 min required to form the gel state. Spectral enhancement is seen upon formation of the intermediate and gel states. Once the peak intensities no longer increased, the gel state was achieved and numerous replicate spectra were acquired.

All spectral data were acquired using the software package Win-IR-Pro v2.97 (Digilab). The spectral range of 600–4200 cm^{-1} was used for protein analysis. Data analysis was performed using the software package Igor-Pro v3.12. There was no need for H_2O subtraction since all protein spectra used were in a gel state.

METHODS

Once the protein spectra were acquired via single-pass ATR-FT-IR, they were water vapor-subtracted, baseline-corrected, and consolidated into a protein library. The numbers of spectra for each protein in the spectral library are given in Table 1. Since the protein spectra were acquired continuously, any change in spectra due to denaturation could be observed in real time. Denaturation was not a consistent problem with any protein other than chymotrypsinogen. None of the spectra for chymotrypsinogen were included in the spectral library.

Suitable protein spectra were transferred to the software package MATLAB v. 5.3 in order to construct principal component regression models for the prediction of protein secondary structure. Since the number of spectra varied, and in some instances were low in number, the protein spectra were bootstrapped. Large

[6] Susi, H. *Methods Enzymol.* **1986**, *26*, 22.
[7] Susi, H.; Byler, D. *Methods in Enzmology* **1986**, *130*, 290–311.
[8] Jencks, W. *Methods Enzymol.* **1986**, *6*, 914–929.
[9] Miyazawa, T. *J. Chem. Phys.* **1960**, *32*, 1647–1652.
[10] Krimm, S. *J. Mol. Biol.* **1962**, *4*, 528–540.
[11] Krimm, S.; Abe, Y. *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *69*, 2788–2792.
[12] Miyazawa, T.; Shimanouchi, T.; Mizushima, S. *J. Chem. Phys.* **1956**, *24*, 408–418.
[13] Baello, B.; Pancoska, P.; Keiderling, T. *Analy. Biochem*. **2000**, *280*, 46–57.
[14] Susi, H.; Byler, D.; Purcell, J. *J. Biochem. Biophys. Methods* **1985**, *11*, 235–240.
[15] Parker, F. S. *Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine*; Plenum Press: New York, 1971.
[16] Lee, D. C.; Haris, P. I.; Chapman, D.; Mitchell, R. C. *Biochemistry* **1990**, *29*, 9185–9193.
[17] Douseeau, F.; Pezolet, M. *Biochemistry* **1990**, *29*, 8771–8779.
[18] Baumruk, V.; Pancoska, P.; Keiderling, T. *J. Mol. Biol.* **1996**, *259*, 774–791.
[19] Vedantham, G.; Sparks, H. G.; Sane, S. U.; Tzannis, S.; Przybycien, T. M. *Analy. Biochem*. **2000**, *285*, 33–49.

Table 1. Proteins Used in This Study

protein	company	catalog no.	no. of spectra in library	PDB ID no.
α-casein	Sigma	C-6780	8	n/a[a]
caspase	NCSU Biochemistry Department		6	1CP3
α-chymotrypsin	ICN	100461	16	5CHA
chymotrypsinogen	ICN	100477	did not use	1CHG
concanavalin A	ICN	150710	8	1APN
concanavalin A	Sigma	C-7275	20	1APN
cytochrome *c*	Aldrich	10,520−1	8	1CCR
elastase	Sigma	E−1250	6	3EST
glutathione reductase	Sigma	G-6004	7	3GRS
hemoglobin	Sigma	H-2500	27	1A3N
IgG	Fluka	56834	3	n/a
lactalbumin	Sigma	L-5385	30	1HFX
lactoglobulin	Sigma	L-2506	30	1BEB
lysozyme	Sigma	L-6876	33	1LYZ
myoglobin	Sigma	M-1882	33	1MBS
myosin	University of California at San Diego		4	1B7T
papain	ICN	100921	30	1PPD
pepsin	Sigma	P-6887	35	4PEP
ribonuclease A	ICN	193980	11	7RSA
ribonuclease A	Sigma	R-4875	16	7RSA
ribonuclease B	Sigma	R-7884	30	1RBB
subtilisin	Sigma	P-5380	8	1SBT
trypsin	ICN	153571	13	1TPO
trypsin inhibitor	ICN	100612	18	4PTI
trypsinogen	Sigma	T-1143	10	1TGN

[a] n/a, not available.

characteristic data sets are required for multivariate analysis, particularly for the development of multivariate regression models. Small data sets that yield sparsely populated principal component clusters in multivariate space can be expanded by the application of the bootstrap resampling method to yield more densely populated principal component clusters. In this study, a parametric bootstrap technique was used to enlarge small data sets for a better estimation of the protein secondary structure content. The particular bootstrap method used was developed by Gemperline and Smith[20] and resamples from the "inside model space". The terminology "inside model space", first adopted by Van Der Voet et al.,[21] refers to the multidimensional space defined by a principal component model that employs the *k* largest principal components. The complementary space defined by the residuals of the *k* factor principal component model is referred to as the outside model space.

First developed by Efron in 1979, the bootstrap is a method for obtaining estimates of statistical parameters and of the uncertainty in these statistical parameters[22,23] based upon resampling from an empirical distribution.[24−30] In this study, the application of the bootstrap resampling procedure is reported for improving the robustness of the PCR model.

The inside model space bootstrapping method involved the resampling of the column-mode eigenvectors, **U**. The following steps were performed to implement this unique method. First, the original data set was decomposed via the truncated singular value decomposition function:

$$\mathbf{A}_{(n\times m)} = \mathbf{U}_{(n\times k)} \times \mathbf{S}_{(k\times k)} \times \mathbf{V^T}_{(k\times m)} \qquad (1)$$

where **A** represents the spectral data, **U** represents the column-mode eigenvectors, **S** represents the diagonal matrix of the principal component scores, and $\mathbf{V^T}$ represents the row-mode eigenvectors, *n* is the number of spectra, *m* is the number of wavenumbers, and *k* is the number of factors used in the model. The complete matrix of column-mode eigenvectors, **U**, was repeated sequentially until *N*, the number of bootstrap samples, was satisfied. The order of factors in each column of **U** was then randomized independently of all other columns to produce **U***. Finally, a new bootstrapped matrix, **A***, of spectra was generated from eq 2. This process was repeated *J* times, *J* was equal to either

$$\mathbf{A^*}_{(N\times m)} = \mathbf{U^*}_{(N\times k)} \times \mathbf{S}_{(k\times k)} \times \mathbf{V^T}_{(k\times m)} \qquad (2)$$

5 or 10 in this study, and the resulting spectra were averaged together. To compensate for the effect of averaging, the averaged data were multiplied by the square root of *J*. The bootstrap method used does not introduce any new sample variability, since the bootstrapped data have the same sample distribution as the

[20] Smith, B.; Gemperline, P. *J. Chemom.* **2002**, *16*, 241−246.

[21] Voet, H. V. D.; Coenegracht, M. J.; Hemel, J. B. *Anal. Chim. Acta* **1987**, *192*, 63−75.

[22] Efron, B. *Ann. Stat.* **1979**, *7*, 1−26.

[23] Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.

[24] Park, D.; Willemain, T. *Comput. Stat. Data Anal.* **1999**, *31*, 187−202.

[25] Meinrath, G. *Chemom. Intell. Lab, Syst.* **2000**, *51*, 175−187.

[26] Milan, L.; Whittaker, J. *J. R. Stat. Soc. (Appl. Stat.)* **1995**, *44*, 31−49.

[27] Hinkley, D. V. *J. R. Stat. Soc., Ser. B (Methodological)* **1988**, *50*, 321−337.

[28] Furusjo, E.; Danielsson, L. *J. Chemom.* **2000**, *14*, 483−499.

[29] Efron, B. *J. R. Stat. Soc., Ser. B (Methodological)* **1992**, *54*, 83−127.

[30] Conlin, A. K.; Martin, E. B.; Morris, A. J. *J. Chemom.* **2000**, *14*, 725−736.

original data set. Thus, the spectral line shapes are preserved in the bootstrapped data set and the only significant difference was the peak intensities.

PCR is a multivariate technique that predicts protein secondary structure from the line shapes of particular regions within the protein mid-infrared spectrum, particularly the amide I and II regions. Since the spectroscopy–structure correlation is based upon line shapes, protein concentration is not of particular interest. In the current application, PCR models were developed to establish a correlation between spectral data to that of known protein secondary structure content, particularly α-helical and β-sheet content. There is a training phase, a testing phase, and a validation phase. PCR begins with decomposing the training set via the singular value decomposition (SVD) function as seen in eq 1. Upon computing the regression model, the regression vector, **b**, is computed by the following equation,

$$\mathbf{b} = \mathbf{V}_{(m\times k)} \times \mathbf{S}^{-1}{}_{(k\times k)} \times \mathbf{U}^{\mathrm{T}}{}_{(k\times n)} \times \mathbf{c}_{\mathrm{std}(n\times c)} \qquad (3)$$

where $\mathbf{c}_{\mathrm{std}}$ is the matrix of known α-helix and β-sheet content. Using the regression vector, the predicted α-helix and β-sheet content are computed for the training and test sets.

$$\mathbf{c}_{\mathrm{pred}} = \mathbf{A}_{\mathrm{std}(n\times m)} \times \mathbf{b}_{(m\times c)} \qquad (4)$$

Once the model has been optimized, the α-helix and β-sheet content of the validation set is calculated to test the robustness of the model. If an adequate validation is achieved, then one can be assured that a global model has been obtained. Thus, the model can be used to predict the content of unknowns via eq 5.[31]

$$\mathbf{c}_{\mathrm{pred}} = \mathbf{A}_{\mathrm{unk}} \times \mathbf{b} \qquad (5)$$

In this study, the genetic algorithm (GA) was used to optimize both wavenumber selection and the number of principal component factors to be included in the multivariate regression model. The GA is a popular optimization technique that employs a probabilistic, nonlocal search heuristic that was inspired by Darwin's theory of natural selection.[32,33] The GA manipulates binary strings known as chromosomes, which contain genes that encode experimental parameters. An initial population of random binary strings is produced giving an $n \times m$ data matrix, where n is the number of individuals in the initial population and m is the length of each chromosome. The multivariate regression models specified by these chromosomes are constructed, tested, and ranked according to the desired figure of merit. The "best" individuals have the greatest possibility of surviving in the GA. The chromosomes of the best individuals are recombined to produce offspring chromosomes with even better genetic material. Mutations are allowed to occur in the population at a very low rate. The mutations generally result in models that are worse; however, occasionally mutations can produce a change that results in a better model that is then incorporated in the evolutionary process of producing a new population. A number of different groups have reported the use of genetic algorithms as a tool for wavelength selection[32–38] and the determination of principal component factors to be used[39] in multivariate calibration.

In the GA method optimization procedure, one chromosome contained sufficient information to completely specify the parameters needed for calibration. Each chromosome contained two types of genes. The first gene represented the wavenumbers to be used in the principal component regression model. The wavenumber selection gene was a $1 \times m$ vector of randomly generated numbers between 0 and 1, where m represents the maximum number of wavenumbers to be included in the PCR model. Random numbers were rounded to either the ceiling or the floor (1 or 0). A bit position equal to 1 signified including this representative wavenumber, whereas, a 0 signified omitting the corresponding wavenumber. Using this procedure, one would expect an average of 50% of the wavenumbers bit positions being coded as 1. One also wants to avoid generating a chromosome that codes for the inclusion of too few wavenumbers to be included in the PCR model. Thus, features were added to offset the probability of having 50% of the wavenumbers as well as require a minimum of 1/5 of the wavenumbers to be included in the PCR model. The second gene contained 18 bits and encoded the number of principal components to be used for building the principal component model. The gene contained 18 bits to signify 1–18 principal components. Only one bit was allowed to be true, and the number of principal components was based upon the position of the true bit. A maximum of 18 factors was employed to prevent overtraining of the principal component model.

The training and test sets were used for monitoring the GA optimization. The figure of merit was the average of the standard error of calibration (SEC) and the standard error of prediction (SEP). The best 50% of the chromosomes were selected as parents to produce a new population. The remaining chromosomes were discarded. The discarded chromosomes were replaced with a new set, equal in number, of randomly generated chromosomes.

An offspring population of chromosomes was produced by recombining chromosomes from the parents at a given number of random crossover points and by introducing a random mutation rate of 5%. The crossover sites were randomly generated from the bit positions that were involved in wavenumber selection. After selection of the random crossover points, the resulting pieces of chromosomes were randomly shuffled and recombined to generate the offspring population. The offspring chromosomes were then translated into their respective wavenumber regions and number of principal component factors, and the resulting principal component regression models were computed. The training set was reclassified using the new models, and the offspring were ranked to find the individuals producing the best

[31] Gemperline, P., personal communication, 1997.

[32] Leardi, R.; Boggia, R.; Terile, M. *J. Chemom.* **1992**, *6*, 267.

[33] Cong, P.; Li, T. *Anal. Chim. Acta* **1994**, *293*, 191.

[34] Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. *Anal. Chim. Acta* **1994**, *286*, 135.

[35] Bangalore, A. S.; Shaffer, R. E.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1996**, *68*, 4200–4212.

[36] Arcos, M. J.; Ortiz, M. C.; Villahoz, B.; Sarabia, L. *Anal. Chim. Acta* **1997**, *339*, 63–77.

[37] Jouan-Rimbaud, D.; Massart, D. L.; Leardi, R.; Noord, O. E. D. *Anal. Chem.* **1995**, *67*, 4295–4301.

[38] Leardi, R. *J. Chemom.* **1994**, *8*, 65–79.

[39] Depczynski, U.; Frost, V. J.; Molt, K. *Anal. Chim. Acta* **2000**, *420*, 217–227.

SEC and SEP. Fifty percent of the fittest offspring were used for the next generation in addition to randomly generated chromosomes, as described earlier. The process of producing new generations of chromosomes was repeated until the number of given, in this case 20, generations had been satisfied. When the GA optimization routine was completed, the last generation of chromosomes was returned to the user.

RESULTS AND DISCUSSION

The purpose of this study was the development of a multivariate regression model for α-helix content and *β*-sheet content prediction of single-pass ATR-FT-IR data. The development of an accurate model focused on two issues. The first issue was the development of representative training, test, and validation sets. Each of two spectral libraries constructed in this study, denoted spectral library 1, for *β*-sheet prediction, and spectral library 2, for α-helical prediction, consisted of a training set, a test set, an independent validation set and an additional data set of unknowns. The second focus was on which wavenumber regions to include in the multivariate regression model.

In spectral library 1, the proteins caspase, chymotrypsin, concanavalin, cytochrome *c*, elastase, glutathione reductase, lactalbumin, lysozyme, myoglobin, myosin, papain, ribonuclease A, subtilisin, trypsin inhibitor, trypsin, and trypsinogen were included in both the training and test sets. The validation set within spectral library 1 contained the proteins hemoglobin, lactoglobulin, pepsin, and ribonuclease B. The construction of the training and test sets involved bootstrapping the original protein spectra 10 times to make a total of 100 spectra. The bootstrapped data set was then split into odd-numbered bootstrapped spectra (training set) and even-numbered spectra (test set). This enabled an equal number of representative protein spectra to be included in the training set since the number of original protein spectra varied (Table 1). The validation set was also bootstrapped to make a total of 100 spectra; however, the odd-numbered bootstrapped spectra were kept while the even-numbered spectra were discarded. It is common to use bootstrapped data as test and validation sets but not for the training set. The rationale for the use of bootstrapped data in the training set was that mid-infrared spectra of proteins in H_2O solution are only moderately reproducible in the amide A, amide B, and C−H stretching regions due to the varying levels of hydration observed by the single-pass ATR-FT-IR method. Thus, this type of variability must be introduced to the calibration model for adequate prediction. This was achieved by performing the bootstrap, since the resulting bootstrapped data set had a greater population density of the varying levels of hydration in the amide A and B regions. The GA was applied to optimize the number of factors as well as the wavenumbers in the region from 600 to 4200 cm^{-1} with the exception of the CO_2 region. The resulting PCR parameters yielded reasonable training and test set predictions; however, the validation results needed improvement. Further optimization of the calibration model was performed where only specific regions were used. Surprisingly, the best regions were 1475−1750 and 2400−3900 cm^{-1}. This wavenumber region included amide I, amide II, amide A, and amide B and the C−H stretching region. In the literature, primarily the amide I and II regions are used in multivariate models to predict secondary structure content.[17,18,40,41] The GA was applied to optimize the PCR parameters in the wavenumber regions 1475−1750 and 2400−3900 cm^{-1}. The validation results from the GA-optimized PCR parameters are given in Table 2 and a plot of a sample validation spectrum at the 580 GA optimized wavenumbers is given in Figure 1.

Table 2. The Two PCR Methods That Yielded the Best Validation Results

	α-Helix Prediction	
library	spectral library 2	
regions	amide I, amide II, amide IV, amide A, amide B, and C−H stretching region	
factors	18	
wavenumbers	120	
validation set	actual α-helix	pred α-helix content using PCR
cytochrome *c*	39.6	38.5
lactoglobulin	10.3	10.4
pepsin	11.0	10.5
ribonuclease B	17.7	12.8
	β-Sheet Prediction	
library	spectral library 1	
regions	amide I, amide II, amide A, amide B, and C−H stretching region	
factors	18	
wavenumbers	580	
validation set	actual *β*-sheet content	pred *β*-sheet content using PCR
hemoglobin	0.0	0.1
lactoglobulin	42.9	42.2
	43.6	44.0
robonuclease B	33.1	41.2

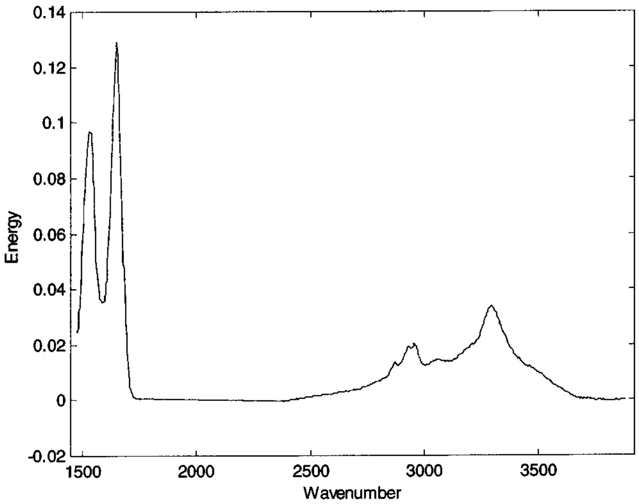

Figure 1. Single-pass ATR-FT-IR spectrum of the protein hemoglobin at the GA-optimized wavenumbers.

The prediction results from the above model resulted in good *β*-sheet predictions, but the α-helix content predictions needed improvement. To enhance the PCR model for α-helix prediction,

[40] Wi, S.; Pancoska, P.; Keiderling, T. A. *Biospectroscopy* **1998**, *4*, 93−106.

[41] Pribic, R.; van Stokkum, I. H. M.; Chapman, D.; Haris, P. I.; Bloemendal, M. *Analy. Biochem.* **1993**, *32*, 366−378.

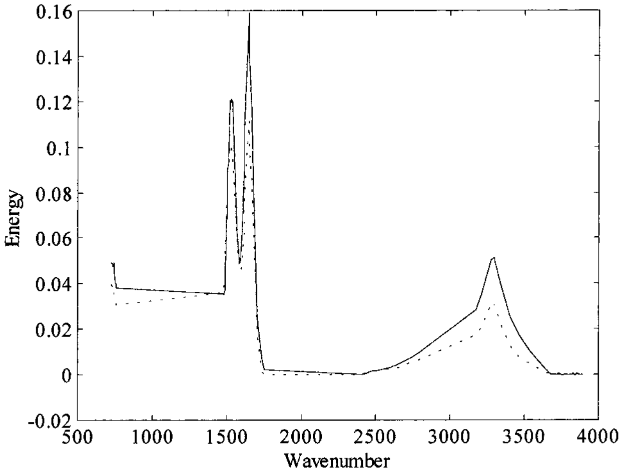

Figure 2. Single-pass ATR-FT-IR spectra of the proteins lactoglobulin (dashed) and pepsin (solid) at the GA-optimized wavenumbers.

the training, test, and validation sets were modified and the inclusion of different amide regions was investigated. In spectral library 2, the proteins caspase, chymotrypsin, concanavalin, elastase, glutathione reductase, lactalalbumin, lysozyme, myoglobin, myosin, papain, ribonuclease A, subtilisin, trypsin inhibitor, and trypsin were included in both the training and test sets. The validation set within spectral library 2 contained the proteins cytochrome *c*, lactoglobulin, pepsin, and ribonuclease B. The training, test, and validation sets were bootstrapped as indicated earlier. It was determined that inclusion of the amide IV region, 720−760 cm^{-1}, in addition to the amide I, amide II, amide A, amide B, and C−H stretching regions, yielded better α helical predictions. GA optimization within the amide I, II, IV, A, and B and C−H stretching regions resulted in a model using 120 wavenumbers and 18 factors. These results were the best for α-helix content prediction and are given in Table 2. A plot of two sample validation spectra at the 120 GA-optimized wavenumbers is given in Figure 2.

Spectral library 1 along with the use of the amide I, amide II, amide A, amide B and C−H stretching regions yielded the best *β*-sheet content predictions. The average absolute error of calibration for this model was 2.0% for *β*-sheet content, and the average absolute error of prediction for the test set was 2.3% for *β*-sheet content. The average absolute error for the validation results was 2.3%. Spectral library 2 along with the use of the amide I, II, and IV regions yielded the best α-helix content predictions. The average absolute error of calibration for this model was 1.9% for α-helix content and the average absolute error of prediction for the test set was 1.9% for α-helix content. The average absolute error for the validation results was 1.7%. Two proteins with unknown secondary structure content, casein and IgG, were bootstrapped to produce a total of 50 spectra. The secondary structure content of casein and IgG were predicted using the first model for *β*-sheet content prediction and the second model for α-helix content prediction. The α-helix content of these proteins was determined to be 24.2 and 8.4%, whereas the *β*-sheet content was determined to be −2.7 and 29.3%, respectively.

CONCLUSIONS

This study establishes the proof of principle that the spectra of proteins in H_2O solution acquired by single-pass ATR-FT-IR can provide estimates of α-helical and *β*-sheet content comparable to studies that use other FT-IR methods. Multivariate models have been applied primarily to amide I and II regions in transmission FT-IR and multipass ATR-FT-IR protein spectra.[17,18,40,41] Spectral enhancement of these and other amide bands occurs when using the single-pass ATR-FT-IR method since the protein sample is continuously monitored by FT-IR in various hydration states.[42] The single-pass ATR-FT-IR method permits the inclusion of more amide regions in the calibration model but raises the issue of protein denaturation. This study produced better α-helical and *β*-sheet predictions than those in previous studies.[7,16−19,40] The resulting α-helical and *β*-sheet predictions, comparable to methods given in the literature on transmission FT-IR and multipass ATR-FT-IR spectra, substantiate that protein denaturation does not occur on the time scale required for secondary structure determination by single-pass ATR-FT-IR. The improved predictions are likely due to the inclusion of a greater number of spectral regions, which can be acquired using the single-pass ATR-FT-IR technique.[42] The multivariate analysis methods of the two spectral libraries combined with the single-pass ATR-FT-IR technique suggest future work that can be based upon existing techniques that sort proteins into classes.[43] The methods developed in this study can then be used to determine secondary structure with greater accuracy in each class. The fact that single-pass ATR-FT-IR has the potential to be automated suggests that spectral libraries sorted by class may become an important tool for proteomic analysis.

SUPPORTING INFORMATION AVAILABLE

Detailed training, test, and validation set predictions. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review February 15, 2002. Accepted May 1, 2002.

AC020104N

[42] Smith, B. M.; Franzen, S. *Anal. Chem.*, in press.
[43] Sreerama, N.; Woody, R. W. *J. Mol. Biol.* **1994**, *242*, 497−507.