

# Partial Least Square Analysis of Lysozyme Near-Infrared Spectra

SHIH-YAO B. HU,<sup>1</sup> AMY LILLQUIST,<sup>1</sup>  
MARK A. ARNOLD,<sup>\*1</sup> AND JOHN M. WIENCEK<sup>2</sup>

<sup>1</sup>*Department of Chemistry*  
and <sup>2</sup>*Department of Chemical and Biochemical Engineering,*  
*University of Iowa, Iowa City, IA 52242,*  
*E-mail: mark-arnold@uiowa.edu*

Received November 1, 1998; Revised April 1, 1999;

Accepted April 15, 1999

## Abstract

Proteins possess strong absorption features in the combination range (5000–4000 cm<sup>-1</sup>) of the near infrared (NIR) spectrum. These features can be used for quantitative analysis. Partial least squares (PLS) regression was used to analyze NIR spectra of lysozyme with the leave-one-out, full cross-validation method. A strategy for spectral range optimization with cross-validation PLS calibration was presented. A five-factor PLS model based on the spectral range between 4720 and 4540 cm<sup>-1</sup> provided the best calibration model for lysozyme in aqueous solutions. For 47 samples ranging from 0.01 to 10 mg/mL, the root mean square error of prediction was 0.076 mg/mL. This result was compared with values reported in the literature for protein measurements by NIR absorption spectroscopy in human serum and animal cell culture supernatants.

**Index Entries:** Near infrared; Fourier transform infrared; calibration; protein; partial least squares; lysozyme; limit of detection.

## Introduction

Near infrared (NIR) spectroscopy is a valuable method for measuring concentrations of biochemical species in aqueous solutions. The procedure involves passing a selected band of NIR radiation through the sample and extracting the desired chemical information from the resulting spectra. A particularly attractive feature of this method is the ability to operate in a non-destructive and reagent-less manner. These attributes motivate the

\*Author to whom all correspondence and reprint requests should be addressed.

development of NIR sensing schemes for various industrial (1), environmental (2), biomedical (3), and biotechnological (4) processes.

Recently, Harthum et al. (5) assessed the use of NIR spectroscopy to monitor soluble protein levels during the cultivation of animal cells engineered to produce recombinant proteins. They report a standard error of prediction (SEP) of 0.274  $\mu\text{g}/\text{mL}$ , which is nearly four orders of magnitude lower than SEP values reported by others for NIR measurements of protein in human serum samples (6,7). The difference between these values is significant. If NIR spectroscopy is capable of measuring protein in the microgram/milliliter region, then NIR spectroscopic sensing of proteins could be envisioned for monitoring protein production in recombinant processes, as suggested by Harthum et al. (5), and for measuring protein in urine, which is an important clinical measurement for the diagnosis of renal failure (8).

Major differences between the work reported by Harthum et al. (5) and the earlier reports for the measurement of protein in serum include the complexity of the sample matrix and the normal protein levels. Undiluted serum is considerably more complex in terms of chemical composition compared to cultivation fluid. Such an increase in matrix complexity might adversely affect the detection limit of the method. In addition, protein levels in serum are relatively high, and the uncertainty in the corresponding reference method generally limits the uncertainty in the NIR calibration models. These differences raise the question, what is the limit of detection for protein under ideal conditions of limited chemical complexity in the sample matrix? To address this question, we collected NIR spectra from a series of lysozyme standard solutions with concentrations ranging from 0.01 to 10  $\text{mg}/\text{mL}$ . We then assessed the ability to accurately measure these protein levels on the basis of partial least squares (PLS) multivariate calibration models. Under our spectrometer conditions, lysozyme protein measurements were limited to levels  $>0.08 \text{ mg}/\text{mL}$ . This estimate of limit of detection is lower than those reported for the analysis of human serum but substantially higher than that suggested by Harthum et al. (5).

## Materials and Methods

Combination band spectra were collected with a Nicolet 550 Fourier transform spectrometer (Nicolet Instrument, Madison, WI) equipped with a 75-W tungsten-halogen lamp, calcium fluoride beam splitter, and cryogenically cooled indium antimonide (InSb) detector. The incident light was restricted to the spectral range of 5000–4000  $\text{cm}^{-1}$  (2.00–2.50  $\mu\text{m}$ ) by using a multilayer optical interference filter (Barr Association, Westford, MA). The aperture was set to 55 and the gain was set to 8 for all measurements. Samples were placed in an aluminum-jacketed sample cell (Wilma Glass, Buena, NJ) with sapphire windows (Meller Optics, Providence, RI). The path length of the cell was 1.5 mm. Sample temperature was maintained at  $20.0 \pm 0.1^\circ\text{C}$  with a programmable circulation bath Model 9110 (VWR Scientific, Pittsburgh, PA). Sample temperature was estimated from the out-

put of an external temperature probe that was positioned in the jacket of the sample cell.

Hen egg-white lysozyme solutions (Sigma, St. Louis, MO) were prepared by dissolving approx 2 g of lysozyme in 15 mL of deionized water. Solutions were placed in a dialysis tube (MWCO 3500; Spectrum Medical, Houston, TX) and dialyzed first against two sequential 900-mL vol of deionized water followed by two sequential 900-mL vol of 0.1 M sodium acetate buffer, pH 4.6. The final concentration of the lysozyme stock solutions ranged between 40 and 60 mg/mL. All solutions were passed through disposable 0.2-mm sterile syringe filters before use. The lysozyme stock solution was then mixed with the appropriate quantity of acetate buffer to obtain the desirable final concentration. The acetate buffers were prepared from acetic acid (glacial, A.C.S. Plus grade) and sodium acetate trihydrate (high-performance liquid chromatography grade), both purchased from Fisher (Pittsburgh, PA). Standard lysozyme concentrations were established from absorbance measurements taken at 280 nm with a Shimadzu UV-VIS Scanning Spectrometer Model UV-2101 PC (Shimadzu, Kyoto, Japan). The molar absorptivity used to compute concentrations was  $2.635 \text{ (mg/mL)}^{-1} \cdot \text{cm}^{-1}$ , as reported by Sophianopoulos et al. (9). The UV reference method typically provides an accuracy of 1% or less for lysozyme (10,11). The standard deviation (SD) among replicates is typically <1% in our experimental setup. This means the error should not exceed 2% in the worst situations. A total of 47 samples were prepared with lysozyme concentrations between 0.01 and 10.0 mg/mL.

Spectra were collected as 128 coadded, double-sided interferograms. Interferograms were triangularly apodized and Fourier transformed to produce single-beam spectra with  $1.9\text{-cm}^{-1}$  point spacing. Three spectra were collected for each sample, and the sampling order was random with respect to lysozyme concentration to minimize temporal correlations. Computer algorithms were implemented in FORTRAN 77 on a Silicon Graphics workstation. Multiple linear computations were performed with subroutines from the IMSL software package (IMSL, Houston, TX).

## Results and Discussion

The NIR spectrum of lysozyme was similar to that reported for albumin. Figure 1 presents absorbance spectra for three levels of lysozyme. Inspection of these spectra reveals two major absorption bands centered at 4590 and  $4375 \text{ cm}^{-1}$  with two smaller features appearing as shoulders located at approx 4420 and  $4265 \text{ cm}^{-1}$ . The magnitudes of these absorption bands vary according to lysozyme concentration. The spectra in Fig. 1 also demonstrate baseline variations typically observed in NIR spectra of aqueous solutions. Spectral baselines frequently exhibit positive or negative slopes and/or significant curvature, particularly at low protein concentrations where the extent of analyte absorption approaches the magnitude of such baseline variations. PLS regression effectively accounts for baseline variations of this nature by incorporating this information into some of the model factors.

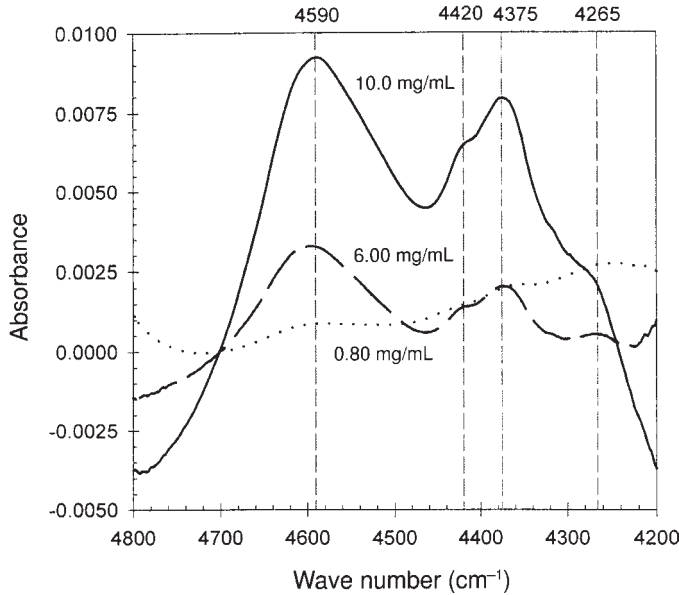


Fig. 1. Absorbance spectra for the indicated concentration of lysozyme.

The leave-one-out cross-validation method (12) was used to establish the best number of PLS factors, or rank, for quantifying lysozyme. Typically, all spectra for a given sample were removed from the data set, a PLS calibration model was generated with spectra from the remaining samples, and this model was used to predict the lysozyme concentration in the removed sample from the replicate spectra. The removed spectra were then returned to the data set, after which all spectra for the next sample were removed and the process was repeated. This procedure was repeated until each sample had been left out once. The predicted residual error sum of square (PRESS) was computed according to the following equation:

$$\text{PRESS} = \sum_{i=1}^N (C'_i - C_i)^2 \quad (1)$$

in which  $N$  is the total number of spectra,  $C'_i$  is the predicted concentrations, and  $C_i$  is the actual concentrations. PRESS values were compared for different numbers of factors from which the optimum rank was established (13).

Spectral range is a critical processing parameter that must be optimized for a given analyte. It would be impractical to test all the possible combinations of sample arrangement, spectral range, and PLS factors as a strategy for identifying a minimum PRESS. An iteration approach is necessary under such conditions. In the present study, first the model rank (number of PLS factors) was optimized by using the aforementioned cross-validation method with the full spectral range of 4800–4200  $\text{cm}^{-1}$ . Second, this number of factors was used for range optimization. Third, the model rank

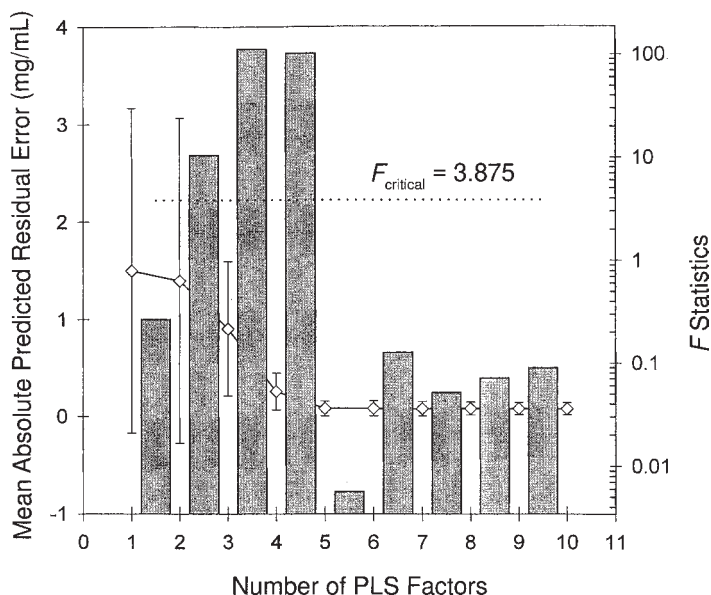


Fig. 2. Effect of PLS factor number (4800–4200  $\text{cm}^{-1}$ ). Diamonds represent the MAPRE values at each number of factors, and the SDs associated with the MAPRE are represented by the error bars. The bars between the number of factors are the  $F$  statistics associated with the increase in the number of factors.

was again optimized by using the new spectral range. If the new rank matched that found in the first step, the optimization was completed. Otherwise, the new number of factors was used to repeat step two. Steps two and three were iterated until the model rank and spectral range converged to fixed values. In addition, to characterize statistically the effect of model rank in step 1, a mean absolute predicted residual error (MAPRE) along with the standard error associated with the mean were used instead of PRESS. MAPRE is defined as follows:

$$\text{MAPRE} = \frac{1}{N} \sum_{i=1}^N \text{ABS}(C'_i - C_i) \quad (2)$$

Just like PRESS, a “good” calibration model should have a low MAPRE because absolute values of residuals are used and it is always a positive number.

Figure 2 shows MAPRE (diamonds) as a function of the number of PLS factors with the full range (4800–4200  $\text{cm}^{-1}$ ). Initially, MAPRE decreased rapidly with the number of factors but leveled off after five factors. The  $F$  statistics (the bars corresponding to the scale of the right axis in Fig. 2) show the effect of changing from one factor to the next (e.g., the bar between factors 1 and 2 represents the effect of changing the model rank from one to two). For a change to be significant, an  $F$  value  $>3.875$  is required for 140 spectra with a 95% confidence interval. Increasing the number of fac-

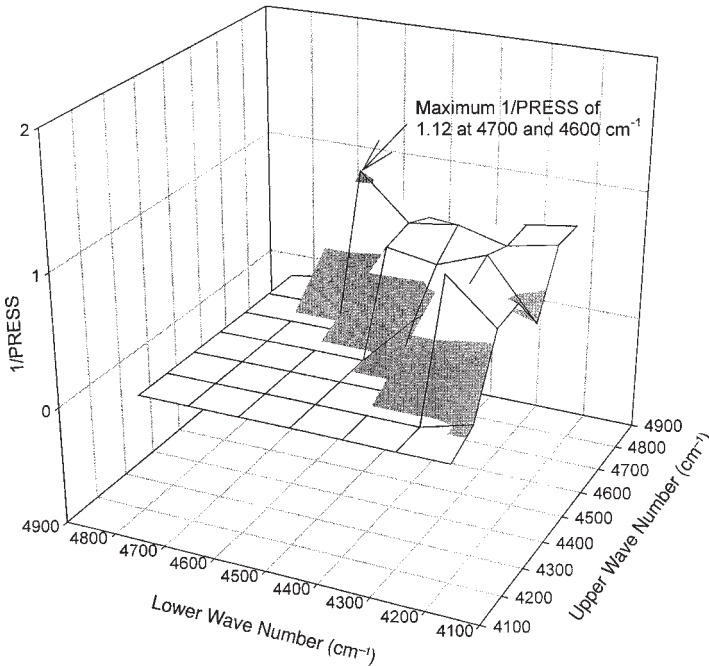


Fig. 3. Surface plot showing results from the first iteration in range optimization with 1/PRESS plotted as a function of spectral range in the lysozyme PLS analysis. The range that gives the highest 1/PRESS value is returned to the optimization algorithm for further refinement.

tors from one to two does not improve model performance because of the large SDs in both cases. However, increasing the model rank to 3, 4, and 5 does improve the calibration significantly, as indicated by  $F$  values  $>3.875$ . Increasing the number of factors beyond five does not improve the calibration performance.

The ideal spectral range was then established with a five-factor PLS model. The range between 4800 and 4200  $\text{cm}^{-1}$  was first evenly divided into 100 wave number intervals and the cross-validation analysis was applied. The region resulting in the lowest value of PRESS was identified and subsequently refined into smaller intervals. This process is illustrated in Figs. 3 and 4. In Fig. 3, 1/PRESS is plotted against the upper and the lower limits of the spectral region used in the PLS analysis. The best region should have the highest 1/PRESS value. In this case, the region between 4700 and 4600  $\text{cm}^{-1}$  gave the best results. This particular region is identified by an open circle in Fig. 4 on the grid line among other tested combinations of range (solid circles). The region surrounding the open circle, as indicated by the gray area, was further tested with 50 wave number increments to narrow down the range further. The procedure was repeated using 25 and then 10 wave number increments. Using this algorithm, it was found that when five factors are used, the region between 4720 and 4540  $\text{cm}^{-1}$  gives the highest 1/PRESS of 1.24  $\text{mg/mL}$ .

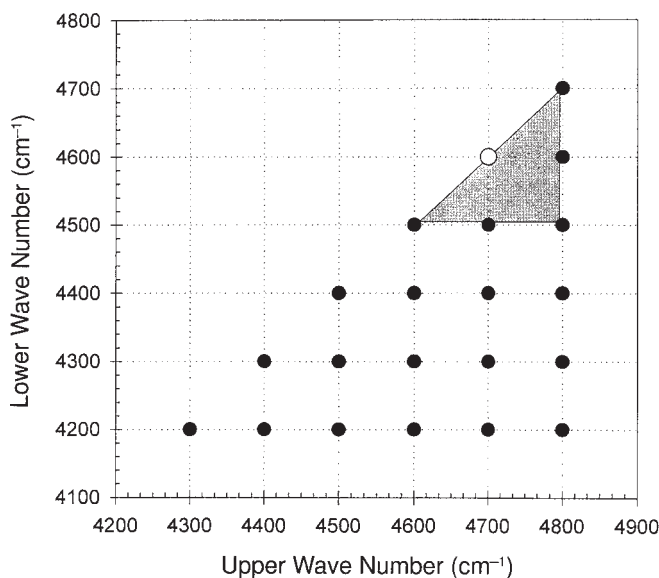


Fig. 4. The first iteration in spectral range optimization in the lysozyme PLS analysis. The range that gives the highest 1/PRESS (○, see Fig. 3 for details) and its surrounding area (gray triangle) is subsequently refined in smaller intervals of 50, 25, and finally 10 wave numbers.

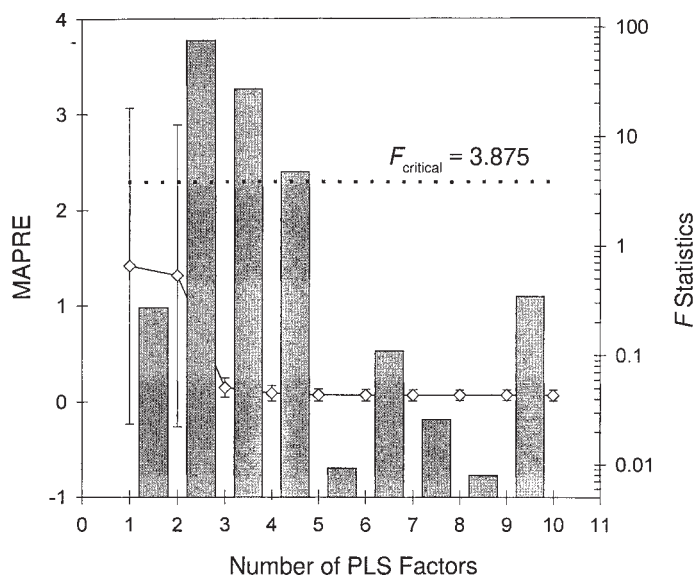


Fig. 5. Effect of number of PLS factors when using the 4720–4540  $\text{cm}^{-1}$  spectral range. Refer to Fig. 2 for details.

Once the spectral range was optimized, the effect of the number of PLS factors was reexamined, as shown in Fig. 5. As can be seen, after the spectral range was narrowed down, the improvements made by increasing the

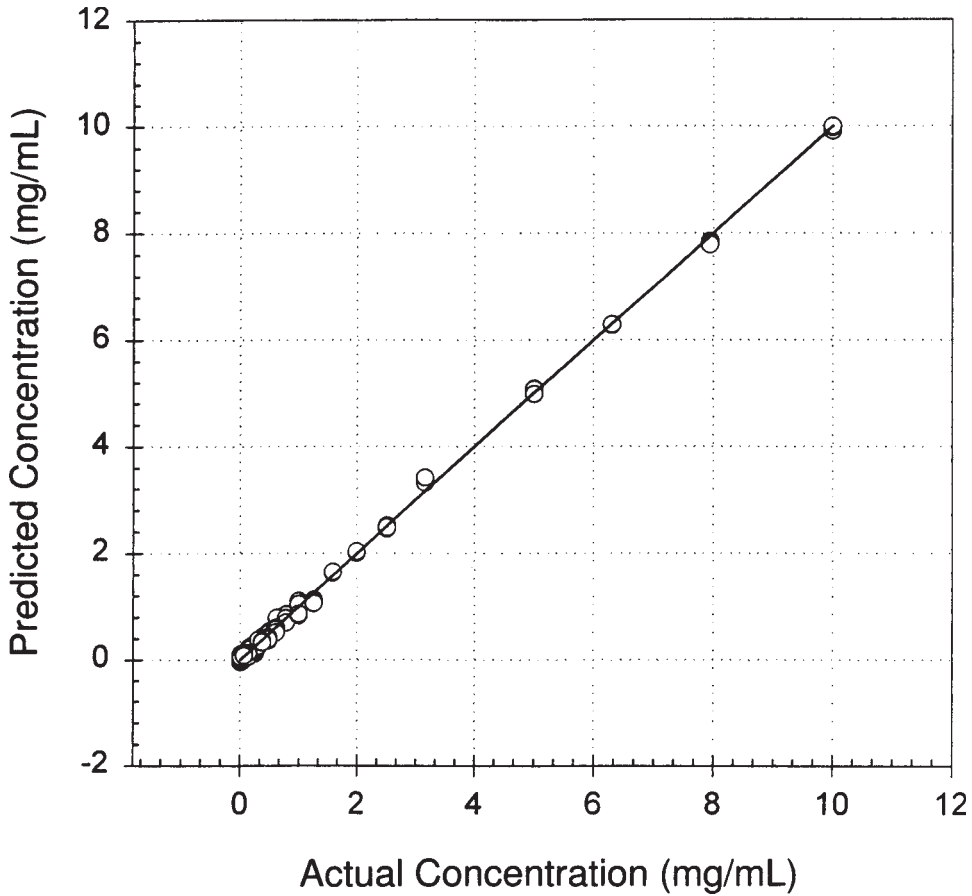


Fig. 6. PLS concentration correlation plot for the lysozyme model based on five factors and the 4720–4540  $\text{cm}^{-1}$  spectral range. (○) How the predicted concentrations compare with the actual sample concentrations. The solid line is a visual aid representing the ideal calibration and is not the regression line.

number of factors from three to five is less significant as compared to Fig. 2. Although MAPRE from three and four factors are quite close to that from five factors, the larger standard errors associated with the three- and four-factor models make them inferior to the model with five factors. This finding is confirmed by the  $F$  statistic, since the  $F$  values between factors 2 and 3, 3 and 4, and 4 and 5 are  $>3.875$ . Subsequent increases in the number of factors did not significantly improve the quality of calibration. Thus, we conclude that a calibration model based on five PLS factors and a spectral range between 4720 and 4540  $\text{cm}^{-1}$  provides the best results. This spectral range corresponds to the main absorption feature around 4590  $\text{cm}^{-1}$ . It is also close to the range of 4670–4560  $\text{cm}^{-1}$  obtained in a separate study by Hu and Arnold (14) in which a univariate calibration model was used.

The full cross-validation analysis was repeated with the optimized rank and spectral range; Figure 6 presents the results. The open circles

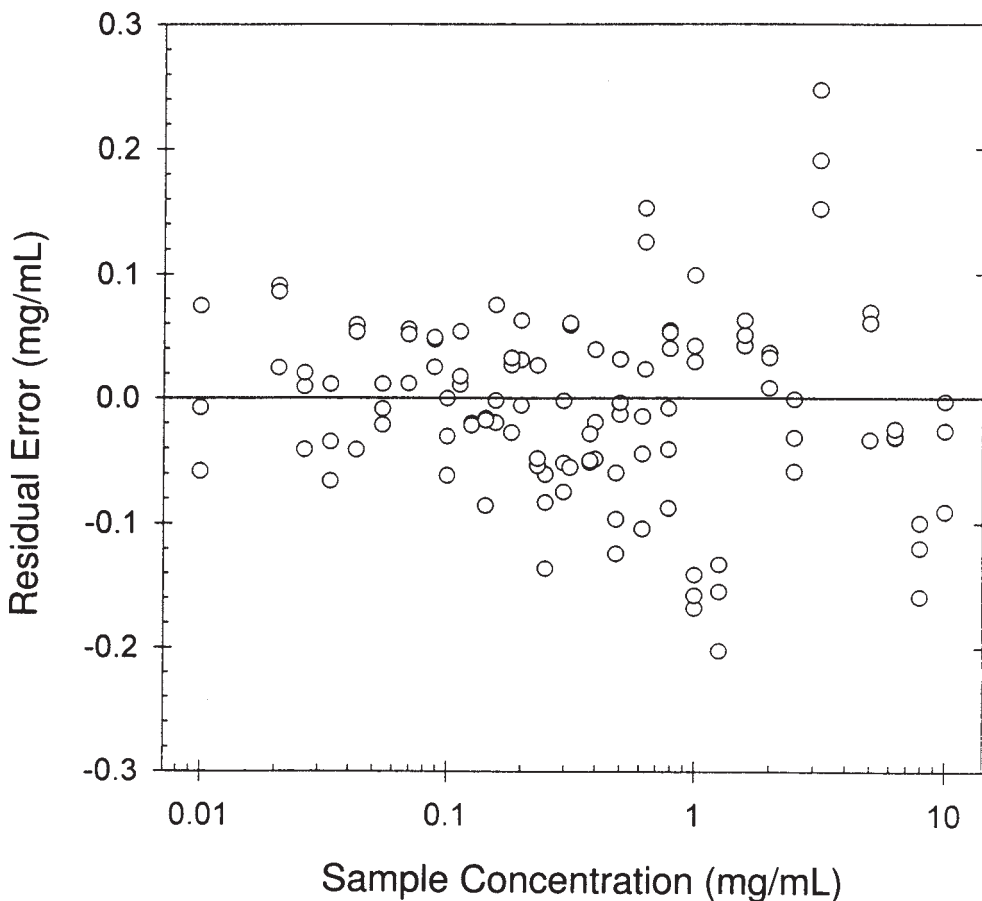


Fig. 7. Residual error from the lysozyme PLS analysis. (○) Residual errors of prediction across the sample concentration range. The solid line indicates the ideal calibration (zero residual).

show how the predicted concentrations compare with the actual sample concentrations. The solid line is provided as a visual guide to show the ideal result and is not the regression line. All the correlation points fall close to the ideal unit line. The root mean square error of prediction (RMSEP), calculated as follows,

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^N (C'_i - C_i)^2}{N}} \quad (3)$$

is 0.076 mg/mL for this calibration model. Figure 7 plots the residual errors. In all cases, the residual error is <0.26 mg/mL, and it is <0.1 mg/mL for more than 85% of the observations. This error is not limited by the reference method (<1% variance) but corresponds to spectral variations. Overall, PLS

regression of NIR spectra can be confidently used to analyze lysozyme aqueous solutions.

The measurement sensitivity and limit of detection can be enhanced by increasing the optical path length. In practice, larger paths result in lower optical throughputs, which adversely affect spectral quality. An optimal path length exists where the signal-to-noise ratio is maximized. This optimum depends on the properties of the spectrometer used to collect the spectra.

The limit of detection ( $<0.1 \mu\text{g}/\text{mL}$ ) and SEP ( $0.275 \mu\text{g}/\text{mL}$ ) reported by Harthum et al. (5) are significantly lower than the values found in our investigation. Several important points are noteworthy. First, the Harthum et al. (5) experiments covered a wider spectral region that extended from 10,000 to  $4000 \text{ cm}^{-1}$  ( $1.0\text{--}2.5 \mu\text{m}$ ). In addition, they used a water background spectrum for their reference spectrum as opposed to a spectrum of protein-free medium, which would have been a better match for the matrix. Under such conditions, all NIR active components in the sample will contribute to their spectral analysis and will increase the likelihood of calibration models based on concentration correlations within the data set. Indeed, our examination of NIR spectra reveals poor quality protein information over the  $10,000\text{--}5000 \text{ cm}^{-1}$  spectral range. Furthermore, Harthum et al. (5) report that the full spectral range is best for their protein models, which is inconsistent for models based on protein information but is consistent with a model based on concentration correlations within the data set. They report that a nine-factor model is required when using the full spectral range. They show that their SEP values level off after three to five factors, as we observed (*see* Figs. 2 and 5). However, their SEP values decrease sharply again after five factors. This type of response is likely a sign of overmodeling by incorporating nonprotein-specific spectral information into the model. The dependency on nonprotein information may also explain why their model failed in crosswise predictions (5). Several other factors favor our findings. First, the only analyte in our system was protein so there was no possible interference from other chemical species. As such, our analysis should, if anything, overpredict the capability of NIR spectroscopy for measuring proteins in aqueous samples. Second, the optical path length used by Harthum et al. (5) is simply too short for meaningful analytical data. Their path length was only 0.2 mm (compared with 1.5 mm used in the current study) and thus provided less protein absorption. Third, the Harthum et al. (5) used a PbS detector, which is typically much noisier than the liquid nitrogen-cooled InSb detector used in the current study. The combination of a shorter optical path and noisier detector suggests a lower signal-to-noise ratio in the Harthum et al. (5) data set.

All these factors combined indicate that the limit of detection stated by Harthum et al. (5) does not accurately reflect the true capabilities of NIR spectroscopy. We conclude that a more realistic expectation for the NIR analysis of aqueous solutions for protein measurements is a detection limit in the milligram/milliliter range instead of the milligram/milliliter range.

## Acknowledgments

We greatly appreciate the help of Dr. Jason Bermeister and Chris Eddy on the instrumentation and analysis. Preliminary work on this subject by Dr. Mark Riley and Carolyn Green proved the feasibility of protein NIR analysis. This work was supported by grants from NASA (NA6-8-1352) and the Iowa Space Grant Consortium. The program for the PLS analysis was provided by Prof. Gray Small at the Center for Intelligent Chemical Instrumentation in the Department of Chemistry at Ohio University.

## References

1. Zilberman, I., Bigman, J., and Sela, I. (1996), *Hydrocarbon Process.* **75**, 91, 92.
2. Han, L. and Rundquist, D. C. (1997), *Remote Sensing Environ.* **62**, 253–260.
3. Flewelling, R. (1995), in *The Biomedical Engineering Handbook*, Bronzino, J. D., ed., CRC Press, Boca Raton, FL, pp. 1346–1356.
4. Hall, J. W., McNeil, B., Rollins, M. J., Draper, I., Thompson, B. G., and Macaloney, G. (1996), *Appl. Spectros.* **50**, 102–108.
5. Harthum, S., Matischak, K., and Friedl, P. (1997), *Analyt. Biochem.* **251**, 73–78.
6. Hazen, K. H., Arnold, M. A., and Small, G. W. (1998), *Analytica Chimica Acta* **371**, 255–267.
7. Hall, J. W. and Pollard, A. (1993), *Clin. Biochem.* **26**, 483–490.
8. Ruggenenti, P., Gaspari, F., and Remuzzi, G. (1998), *BMJ* **316**, 504–512.
9. Sophianopoulos, A. J., Rhodes, C. K., Holcomb, D. N., and Van Holde, K. E. (1962), *J. Biol. Chem.* **237**, 1107–1112.
10. Gill, S. C. and von Hippel, P. H. (1989), *Analyt. Biochem.* **182**, 319–326.
11. Pace, C. N., Vadjos, F., Fee, L., Grimsky, G., and Gary, T. (1995), *Protein Sci.* **4**, 2411–2423.
12. Martens, H. and Maes, T. (1989), *Multivariate Calibration*, John Wiley & Sons, Chichester, United Kingdom, pp. 237–266.
13. Kramer, R. (1998), in *Chemometric Techniques for Quantitative Analysis*, Marcel Dekker, New York, pp. 99–110.
14. Hu, S. B. and Arnold, M. A. (1998), Paper presented at the analytical chemistry poster session, ACS annual meeting, Dallas, TX.

