Anal. Chem. **2002,** *74,* 3392−3399

Multivariate Calibration Models for Lysozyme from Near-Infrared Transmission Spectra in Scattering Solutions of Monodisperse Microspheres

Carolyn E. Green,[†] John M. Wiencek,[‡] and Mark A. Arnold*,[†]

Department of Chemistry, Department of Chemical and Biochemical Engineering, and Optical Science and Technology Center, University of Iowa, Iowa City, Iowa 52242

The ability to quantify lysozyme is demonstrated for a series of aqueous samples with different degrees of scattering. Near-infrared spectra are collected for two sets of lysozyme/scattering solutions. In both sets of samples, the solutions are composed of lysozyme dissolved in acetate buffer with suspended monodisperse latex microspheres of polystyrene. The diameter of the microspheres is 6.4 μm for the first set and 0.6 μm for the second. For each set, the amount of microspheres range from 0.005 to 0.998 wt %, the lysozyme concentrations range from 0.834 to 28.6 mg/mL, and solution compositions are designed to minimize correlations between the concentration of lysozyme and percentage of microspheres. Near-infrared spectra are collected individually for each set of solutions. Single-beam spectra are collected over the combination spectral range (5000−4000 cm^{-1}, 2.0−2.5 μm) by transmitting the incident radiation through a 1.5-mm-thick sample that is maintained at 21 °C. Partial least-squares calibration models are evaluated individually for each data set both with and without wavelength optimization. Results indicate that models from raw, nonmodified, single-beam spectra are incapable of extracting lysozyme concentration from these highly scattering solutions. Accurate concentration measurements are possible, however, by implementing either a multiplicative scatter correction to the single-beam spectra or by taking the ratio of these single-beam spectra to an appropriate reference spectrum. In addition, digital Fourier filtering of these spectra enhances model performance. The best calibration model in the presence of 6.4-μm microspheres is obtained from multiplicative scatter corrected single-beam spectra over the 4550−4190-cm^{-1} spectral range. The mean percent error of prediction (MPEP) and standard error of prediction (SEP) for this model are 2.2% and 0.28 mg/mL, respectively. Likewise, the multiplicative scatter corrected spectra with wavelength optimization provided the best calibration model for the 0.6-μm data set. In this case, the MPEP and SEP are 2.3% and 0.44 mg/mL, respectively. In addition, the ability to predict lysozyme concentrations is evaluated for the situation where the degree of scattering is greater in the predication samples compared to the calibration samples. Differences in the prediction ability are noted between the 6.4- and 0.6-μm data sets.

Success of the Human Genome Project has created a demand to identify and characterize thousands of new proteins.[1] Tertiary structure is an important characteristic for understanding both protein function and regulation. X-ray diffraction is currently the best method for establishing tertiary structure. In many cases, X-ray diffraction analysis is hampered by the inability to grow protein crystals of sufficient quality and size. Large arrays of crystallization liquors are typically used to screen a multitude of possible physical and chemical conditions for growing protein crystals.[2] The objective is to find the chemical and physical conditions that are suitable for growing diffraction-quality crystals for individual proteins.

The rate of crystallization affects the packing density, and hence quality, of protein crystals.[3] Slower crystallization promotes greater packing densities and higher order crystals, which ultimately produces superior spatial resolution. The utility of growing protein crystals in microgravity is currently a subject of debate.[4−6] Several observations suggest that crystals formed under microgravity conditions are more highly ordered compared to their ground-based counterparts.[4−6] One possibility is that the lack of gravitationally induced convective forces promotes slower crystallization by limiting mass transport to diffusion. The kinetics of protein crystal growth can also be controlled by actively adjusting the degree of supersaturation of the protein within the crystallization liquor.[7] Solution temperature is one parameter that can be used to control the degree of supersaturation. We recently

* To whom correspondence should be addressed: (e-mail) mark-arnold@uiowa.edu.

[†] Department of Chemistry and Optical Science and Technology Center.

[‡] Department of Chemical and Biochemical Engineering and Optical Science and Technology Center.

(1) Heinemann, U.; Frevert, J.; Hofman, K.-P.; Illing, G.; Maurer, C.; Oschkinat, H.; Saenger, W. *Prog. Biophys. Mol. Biol.* **2000**, *73*, 347−362.

(2) Stevens, R. C. *Curr. Opin. Struct. Biol.* **2000**, *10*, 558−563.

(3) Schall, C. A.; Riley, J. S.; Li, E.; Arnold, E.; Wiencek, J. M. *J. Cryst. Growth* **1996**, *165*, 299−307.

(4) Garcia-Ruiz, J. M.; Otalora, F. *J. Cryst. Growth* **1997**, *182*, 155−167.

(5) Ng, J. D.; Lorber, B.; Giege, R.; Koszelak, S.; Day, J.; Greenwood, A.; McPherson, A. *Acta Crystallogr.* **1997**, *D53*, 724−733.

(6) Broutin, I.; Ries-Kault, M.; Ducruix, A. *J. Cryst. Growth* **1997**, *181*, 97−108.

(7) Jones, W. F.; Wiencek, J. M.; Darcy, P. A. *J. Cryst. Growth* **2001**, *232*, 221−228.

10.1021/ac010976+ CCC: $22.00
Published on Web 06/12/2002

modeled the feasibility of using a feedback-regulated temperature controller to control crystal growth rates.[8] In this system, the concentration of protein within the crystallization liquor is monitored continuously and the rate of crystal growth is computed on the basis of mass balance. The temperature is adjusted in real time to maintain a targeted rate of crystal growth.

Near-infrared spectroscopy is being evaluated as a means for collecting the required concentration of soluble protein during the above-mentioned temperature-controlled crystallization process. The nondestructive nature of near-infrared spectroscopy is an important advantage for this approach. In addition, near-infrared spectroscopy provides suitable limits of detection[9] and computational methods are available to provide accurate measurements under variable-temperature conditions.[8,10]

The effect of light scattering on the accuracy of near-infrared spectroscopic measurements is a critical question given the nature of the crystallization process. Scattering processes at the solution/crystal interface will alter the path of incident photons, which can adversely impact the analytical measurement. As a first step toward understanding the impact of light scattering on analytical measurement of protein with near-infrared spectroscopy, we have examined the effect of different levels of monodisperse polystyrene microspheres on the ability to measure the concentration of lysozyme. Lysozyme is selected because of its common use as a model protein for crystallization processes. In this examination, the impact of microspheres with diameters of 0.6- and 6.4-μm are determined for near-infrared spectroscopic measurements over the combination spectral range of the near-infrared spectrum (5000−4000 cm^{-1}, 2.0−2.5 μm).

EXPERIMENTAL SECTION

Apparatus. Spectra were collected with a Nicolet Magna 550 Fourier transform infrared (FT-IR) spectrometer system (Nicolet Analytical Instruments, Madison, WI). The spectrometer was configured with a 25-W tungsten−halogen source, CaF_2 beam splitter, and liquid nitrogen-cooled InSb detector. A K band-pass filter (Barr Associates, Westford, MA) was used to isolate the combination spectral range. A Wilmad sample cell (Wilmad, Buena, NJ) with a path length of 1.5 mm and sapphire windows was used for sample containment. Solution temperature was maintained at 21.0 ± 0.1 °C with a Polystat constant-temperature controller refrigerated bath (Cole Parmer, Vernon Hills, IL).

Reagents. Monodispersed polystyrene latex microspheres were purchased from Sigma Chemical Co. (St. Louis, MO). The 0.6- and 6.4-μm micorspheres were supplied as suspensions in water with a reported density for polystyrene of 1.05 g/mL. Lysozyme was obtained from Sigma Chemical Co. and was supplied as a lyophilized powder with a manufacturer's stated purity of 95% protein. Sodium acetate salt and glacial acetic acid were purchased from common providers. All solutions were prepared with an 18.2 MΩ water obtained by passing house-distilled water through an Elgestat Maxima purification unit (US Filter, Lowell, MA).

Procedures. Standard solutions were prepared by diluting a stock lysozyme solution with an appropriate amount of a pH 4.6 acetate buffer and adding known amounts of microspheres. Two sets of standard solutions were prepared. The first consisted of 89 solutions composed of lysozyme and 0.6-μm microspheres. The second was composed of 89 samples of lysozyme and 6.4-μm microspheres. In both cases, the stock lysozyme solution was prepared by dialyzing lyophilized lysozyme against water followed by a second dialysis against a pH 4.6 acetate buffer.[11] This dialysis procedure was done to remove low molecular weight contaminants within the commercial lysozyme preparation. The final concentration of lysozyme in the stock solution was determined by a absorbance measurement at 280 nm according to the procedure established by Darcy and Wiencek.[11] The concentration of lysozyme in the stock solutions used to prepare the 0.6- and 6.4-μm solutions was 79.12 and 69.61 mg/mL, respectively.

The composition of each sample solution was determined by weight-based dilutions. The final concentrations of lysozyme and microspheres were designed to minimize the correlation between the amounts of each component between samples. A uniform design protocol was used according to the optimization results tabulated by Fang and Wang.[12] Of the tabulated designs, the 89-sample design was used in order to provide sufficient samples for the partial least-squares (PLS) factor analysis. The concentrations of lysozyme and microspheres were obtained by implementing this design over the noted concentration ranges. The efficiency of this approach is exemplified by the magnitude of the r^2 values for the concentration correlations between lysozyme and microspheres. r^2 values are 0.0002 and 0.000 03 for lysozyme/microsphere correlations in the 0.6- and 6.4-μm microspheres data sets, respectively. These r^2 values indicate that only 0.02 and 0.003% of the variance in the concentration of lysozyme is correlated with the variance in the amount of scattering particles.

Single-beam spectra were collected in random order with respect to both concentration of lysozyme and weight percent of microspheres. The 0.6- and 6.4-μm data sets were collected in two separate sessions. Initially, solutions were placed in a water bath maintained at 21 °C. Subsequently, a small volume of the solution was placed in the thermostated sample cell, which was maintained at 21 °C. The solution thickness was 1.5 mm, and each sample was allowed to equilibrate in the light path for 5 min. Spectra were collected in triplicate without removing the sample from the spectrometer. Buffer spectra were collected in the same manner and were obtained after every fifth sample. The collection of triplicate spectra without removing the sample from the instrument provides replicates in case a spectrum is inadvertently lost or contaminated. In addition, such replicates allow us to obtain 100% lines that are valuable in computing the root-mean-square (rms) noise, which is inversely related to the instrumental SNR. It is important to note, however, that back-to-back spectra only incorporate short-term instrument variations and do not provide information pertaining to long-term instrument stability.

Each spectrum was collected as 128 coadded, double-sided, 8K interferograms with no zero-filling. Interferograms were processed by using a triangular apodization and Mertz phase correction. This process produced single-beam spectra with a 1.928-cm^{-1} point spacing and 4-cm^{-1} resolution. All single-beam spectra were transferred to an Iris Indigo computer (Silicon

(8) Hu, S. B.; Arnold, M. A.; Wiencek, J. M. *Anal. Chem.* **2000**, *72*, 696−702.
(9) Hu, S. B.; Lillquist, A.; Arnold, M. A.; Wiencek, J. M. *Appl. Biochem. Biotechnol.* **2000**, *87*, 153−163.
(10) Olesburg, J. T.; Arnold. M. A.; Hu, S. B.; Wiencek, J. M. *Anal. Chem.* **2000**, *72*, 4985−4990.
(11) Darcy, P. A.; Wiencek, J. M. *J. Cryst. Growth* **1999**, *196*, 243−249.
(12) Fang, K. T.; Wang, Y. *Number-Theoretic Methods in Statistics*; Chapman and Hall: New York, 1994.

Graphics, Inc., Mountain View, CA) for further processing. All computer software used for the spectral processing was obtained from Gary W. Small from the Center for Intelligent Chemical Instrumentation in the Department of Chemistry at Ohio University, Athens, OH. All algorithms were implemented in Fortran 77. PLS regression and Fourier filtering routines involved subroutines obtained from the IMSL software package (IMSL Inc., Houston, TX).

PLS regression analysis was used to generate calibration models for lysozyme. Calibration models were established and evaluated from an analysis of calibration and prediction subsets of the spectral data. The full set of 89 samples was split into calibration and prediction sets. Twenty samples were randomly selected for prediction purposes, and all the corresponding spectra were assigned to the prediction set. Twenty samples were selected to leave 69 samples for training purposes, thereby permitting up to 11 factors in the PLS models.[13] Calibration models were established with all spectra associated with these remaining 69 samples. To identify the ideal spectral range and number of factors, this calibration set was split into training and monitoring sets. All spectra associated with 20 randomly selected samples were taken from the calibration set to serve as a monitoring set, and the remaining spectra served as the training set. These groups of monitoring and training spectra were used in the procedure described below to optimize the spectral range and number of PLS factors. This process of randomly splitting the calibration data and optimizing the model parameters was performed a total of three times in order to minimize adverse effects caused by any outlying spectra or samples. After the ideal calibration parameters were established, the training and monitoring data sets were recombined and the final calibration model was established by performing the PLS regression with the optimized parameters on the full calibration data set. Validity of the final model was judged by the accuracy of model predictions from spectra in the prediction data set. The standard error of prediction (SEP) and mean percent error of prediction (MPEP) were used to quantify the performance of the final model. Equations used to compute SEC, and SEP are provided elsewhere.[14]

Spectral range and number of model factors were optimized by procedures detailed elsewhere.[15] Briefly, a modified grid search was used to find the optimum spectral range. For each tested spectral range, PLS calibration models were generated from spectra in the training set, and model performance was judged by computing the standard error of monitoring (SEM), which is a measure of the accuracy of predicting lysozyme concentrations from spectra in the monitoring set. This grid search procedure was implemented iteratively as the number of factors was incremented systematically from 1 to 11. As described in detail elsewhere,[16] SEM values initially decrease as additional factors incorporate relevant spectral information into the PLS analysis and the SEM increases when the system is overmodeled. The best combination of spectral range and number of factors was determined by comparing SEM values after each iteration. SEM values were compared by using the *F*-test calculation at the 95% confidence levels to determine the significance of SEM values for subsequent iterations. Optimal conditions corresponded to parameters for the first iteration for which the SEM is not significantly different compared to the next iteration.

Digital Fourier filtering and multiplicative scatter correction (MSC) were explored as preprocessing schemes to enhance model performance. Fourier filtering is effective in removing broad baseline variations and high-frequency noise from either single-beam or absorbance spectra, thereby enhancing the single-to-noise ratio of the measurement.[17] In this work, the Fourier filtering was implemented with a Gaussian-shaped band-pass function that was defined by the position and width of the Gaussian-shaped filter in digital frequency units. Details of this filtering strategy and its implementation for processing near-infrared spectra are provided elsewhere.[17,23] MSC is a technique developed by Norris to account for the nonspectral impact of scattering particles on a set of spectra.[18] Scattering causes an apparent offset in the absorbance, which can be independent of the chemical composition of the sample. Such offsets complicate multivariate algorithms for calibration.[19] The MSC algorithm uses linear regression of spectral variables for the average spectrum and simultaneously corrects for both multiplicative and additive scatter effects.[19]

RESULTS AND DISCUSSION

Spectra. Representative absorbance spectra are presented in Figure 1 for comparison. Spectra are provided for samples containing only lysozyme and samples with lysozyme and either 0.6- or 6.4-μm polystyrene microspheres. The spectrum without microspheres reveals the near-infrared spectrum of lysozyme with absorption bands centered at 4595 and 4368 cm^{-1}. The effect of the polystyrene is evident in the spectra with microspheres. The narrow absorption bands centered at 4680, 4600, and 4580 cm^{-1} correspond to polystyrene superimposed on the protein absorbance. An additional polystyrene absorption band at 4350 cm^{-1} is evident in solutions with higher levels of microspheres.

Established models for light scattering depend on the relative size of the particle and the wavelength of light. In these experiments, the wavelength of light ranges from 2.0 to 2.5 μm while the particle diameters are 0.6 and 6.4 μm, respectively. For the 6.4-μm data set, a Mie scattering model is most appropriate because the wavelength is smaller than the particle diameter.[1] On the other hand, the wavelength is longer than the particle diameter for the 0.6-μm microspheres, which means that a Debye scattering model is more appropriate.[20]

The effect of microspheres on the intensity of the single-beam spectra is illustrated by the data plotted in Figure 2. This plot

(13) American Society for Testing and Materials. E 1655, Standard Practices for Infrared, Multivariate, Quantitative Analysis. In *Annual Book of ASTM Standards*; ASTM: West Conshohocken, PA, 1995; Vol. 03.06, pp 1−24.

(14) Arnold, M. A.; Burmeister, J. J.; Small, G. W. *Anal. Chem.* **1998**, *70*, 1773−1781.

(15) Riley, M. R.; Rhiel, M.; Zhou, X.; Arnold, M. A.; Murhammer, D. W. *Biotechnol. Bioeng.* **1997**, *55*, 11−15.

(16) Shaffer, R. E.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1996**, *68*, 2663−2675.

(17) Arnold, M. A.; Small, G. W. *Anal. Chem.* **1990**, *62*, 1457−1464.

(18) Norris, K. H. In *Food Research and Data Analysis*; Martens, H., and Rtasswurm, H., Jr., Eds.; Applied Science Publishers: Essex, U.K., 1983; p 95.

(19) Isaksson, T.; Naes, T. *Appl. Spectrosc.* **1998**, *42*, 1273−1284.

(20) Ingle, J. D.; Crouch, S. R. *Spectrochemical Methods*; Prentice Hall: Upper Saddle River, NJ, 1988; Chapter 3.

(21) Hazen, K. H.; Arnold, M. A.; Small, G. W. *Appl. Spectrosc.* **1998**, *52*, 1597−1605.

(22) Hazen, K. H.; Arnold, M. A.; Small, G. W. *Appl. Spectrosc.* **1994**, *48*, 477−483.

(23) Pan, S.; Chung, H.; Arnold, M. A.; Small, G. W. *Anal. Chem.* **1996**, *68*, 1124−1135.

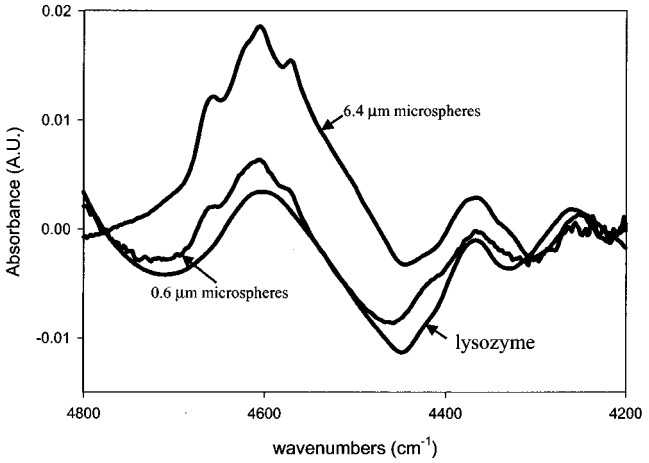

Figure 1. Absorbance spectra for solutions containing lysozyme alone and lysozyme with polystyrene microspheres. Indicated spectra correspond to solutions composed of 14 mg/mL lysozyme, a mixture of 15 mg/mL lysozyme and 0.6 wt % of the 0.6-μm diameter microspheres, and a mixture of 14 mg/mL lysozyme and 0.6 wt % of the 6.4-μm microspheres.

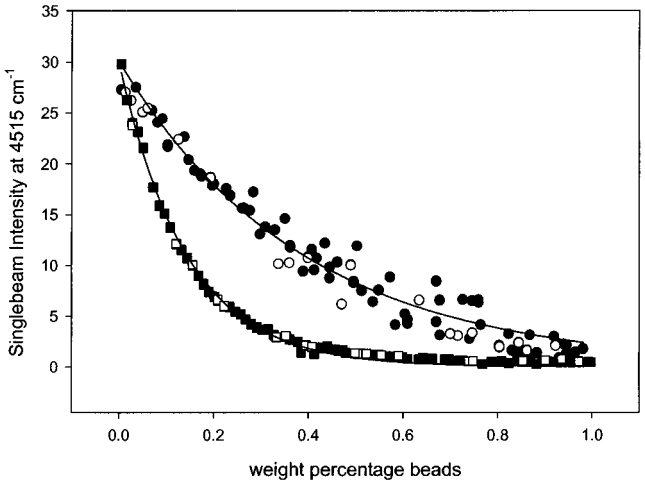

Figure 2. Effect of weight percent microspheres on the detected intensity at 4515 cm^{-1} for solutions with 0.6- (circles) and 6.4-μm microspheres (squares). Open and closed symbols indicate samples used in the calibration and prediction data sets, respectively. Solid lines indicate fits to single-exponential functions.

shows the maximum single-beam intensity (at 4515 cm^{-1}) as a function of weight percent microspheres. For both the 0.6- and 6.4-μm microspheres, the attenuation is exponential with microsphere concentration, which is consistent for both Mie and Debye scattering. In both models, the intensity of light transmitted through the sample should follow an exponential decay according to the following expression:

$$I = I_0 e^{-\gamma l}$$

where I is the recorded intensity in the presence of scattering particles, I_0 is the intensity in the absence of particles, l is the optical path length and γ is a sensitivity coefficient, which corresponds to the combination of the scattering and absorption coefficients for the sample. Results of fitting each data set in Figure 2 to a single-exponential function are indicated as solid lines in

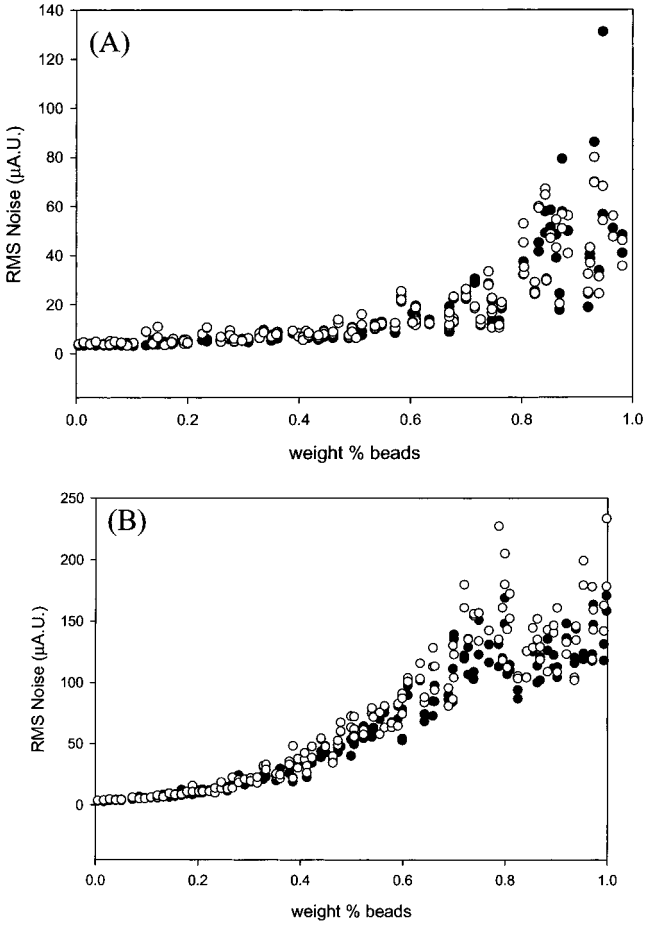

Figure 3. Effect of microsphere amount on the rms noise on 100% lines corresponding to solutions composed of 0.6- (A) and 6.4-μm microspheres (B). Open and closed symbols correspond to noise measurements over the 4600−4500- and 4500−4400-cm^{-1} spectral ranges, respectively.

the figure. These exponential fits reveal I_0 values of 29.5 ± 0.4 and 29.7 ± 0.1 for the 0.6- and 6.4-μm data sets, respectively. Assuming the path length of light is equivalent to the thickness of the sample cell (1.5 mm), the sensitivity coefficients are 258 ± 5 and 693 ± 5 mm^{-1} for these data sets, respectively.

A clear difference between these two data sets is the higher degree of variance in the 0.6-μm intensities. Two factors contribute to the variability in the 0.6-μm data. First, the stock solution of the smaller microspheres was difficult to handle when solutions were being prepared, which likely contributed greatly to the variability in these samples. Second, a thin film of residue from the scattering particles developed on the windows of the sample holder when spectra were being collected for the 0.6-μm data set. Attempts to remove this film by rinsing with solvent between samples were generally unsuccessful. Because the spectra were collected randomly with respect to the weight percent of microspheres, the impact of this effect should be uniform across the data set.

The radiant power at the detector impacts spectral quality by influencing the overall signal-to-noise ratio of the measurement.[21] The impact of scattering microspheres on spectral quality is illustrated in the Figure 3. This plot shows the root-mean-square (rms) noise on 100% lines measured as a function of weight percent microspheres in solution.[21] Rms noise values were

Table 1. Lysozyme Calibration Models from Spectra with 6.4-μm-Diameter Microspheres

spectral type	spectral range (cm^{-1})	rank[a]	SEC (mg/mL)	SEP (mg/mL)	MPEP (%)
single beam	4800−4200	6	5.9	6.5	68
single beam	4800−4500	2	6.9	6.6	71
MSC single beam	4800−4200	3	0.36	0.36	2.8
MSC single beam	4550−4190	3	0.26	0.28	2.2
absorbance	4800−4200	5	0.39	0.40	3.7
absorbance	4610−4300	4	0.39	0.43	3.1
absorbance Fourier filtered (0.020, 0.005*f*)	4610−4350	3	0.31	0.34	2.6
absorbance Fourier filtered (0.017, 0.006*f*)	4480−4270	2	0.38	0.39	2.6

[a] Rank is the number of latent variables or factors used in the model.

computed by first taking the ratios of replicate single-beam spectra, converting to microabsorbance units (μAU), and then fitting the resulting points to a second-order polynomial for a series of 100-cm^{-1} segments (i.e., 5000−4900, 4900−4800, 4800−4700 cm^{-1}, etc.). More details of this process are provided elsewhere.[21] The average rms noise about the 4600−4500- and 4500−4400-cm^{-1} spectral segments are plotted in the Figure 3. As expected, the rms noise increases as the number of particles increases and the degree of scattering increases. The measured noise levels are inversely related to the detected radiant power as is evident by comparing the plots in Figures 2 and 3. Correspondingly, the quality of the spectra degrades as the amount of microspheres increases. Over all samples, the average rms noise values for the 4600−4500- and 4500−4400-cm^{-1} spectral ranges are 52.9 and 60.7 μAU for the 6.4-μm data set and 25.9 and 34.5 μAU for the 0.6-μm data set, respectively. Higher noise levels for the 6.4-μm data set are consistent with the lower radiant powers as indicated in Figure 2.

Calibration Models from the 6.4-μm Data Set. Calibration models were evaluated from both single-beam and absorbance spectra. Models from single-beam spectra were generated with and without implementing the MSC algorithm. Absorbance spectra were generated by dividing each single-beam spectrum by the preceding reference buffer spectrum and computing the logarithm. Absorbance spectra were used directly and after Fourier filtering. For both single-beam and absorbance spectra, models were examined over the full 4800−4200-cm^{-1} spectral range and over the wavelength-optimized spectral range. This 4800−4200-cm^{-1} range excludes the extremely noisy regions (5000−4800 and 4200−4000 cm^{-1}) that correspond to strong light attenuation by water absorption.[22] Results for each model are summarized in Table 1.

Poor calibration models result when generated from raw single-beam spectra over the full spectral range. As noted in Table 1, this model required six factors to provide a MPEP of 68%. The wavelength optimization algorithm provides no improvement in performance with a MPEP of 71%. In both cases, the SEP is similar to the standard deviation of the prediction (SDP) data set (i.e., SEP = 6.5 mg/mL and SDP = 7.74 mg/mL). This degree of similarity indicates a complete failure in the model to predict lysozyme concentrations from raw single-beam spectra.

Information pertaining to the concentration of lysozyme can be accurately extracted from single-beam spectra after implementing the MSC algorithm. As indicated in Table 1, the MSC single-beam spectra resulted in calibration models with considerably fewer factors and lower prediction errors compared to models from untreated single-beam spectra. When the full spectral range is used, the SEP and MPEP are 0.36 mg/mL and 2.8%, respectively, for an optimized three-factor model. When an optimized wavelength range of 4550−4190 cm^{-1} is used, values for the SEP and MPEP decrease to 0.28 mg/mL and 2.2%, respectively.

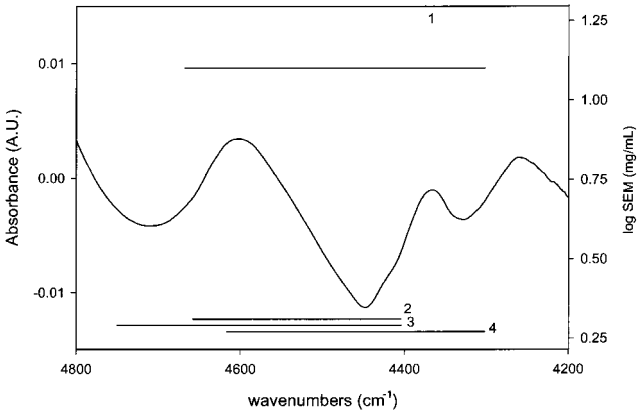

Figure 4. Absorbance spectrum of lysozyme (14.5 mg/mL) superimposed on lines that indicate the optimized spectral range for the indicated number of factors used in the PLS calibration model detailed as entry 6 in Table 1. Spectral range lines are positioned according to the logarithm of the SEM for the corresponding calibration model.

Likewise, functional calibration models are obtained from absorbance spectra. Over the full spectral range, models based on absorbance spectra greatly outperform those based on single-beam spectra. Five PLS factors were necessary when the full spectral range was used. SEC and SEP values for the resulting calibration model are 0.39 and 0.40 mg/mL, respectively, and the MPEP is 3.7%. These values are slightly higher than those obtained from the MSC single-beam spectra.

Compared to models from absorbance spectra over the full spectral range, wavelength optimization results in fewer model factors, or latent variables, and essentially equivalent analytical performance. By using the optimized spectral range of 4610−4300 cm^{-1}, a four-factor PLS model provides SEC and SEP values of 0.39 and 0.43 mg/mL, respectively, and a MPEP of 3.1%. Progress of the wavelength optimization algorithm is illustrated in Figure 4 for this 6.4-μm data set. This figure presents the absorbance spectrum of lysozyme over the full 4800−4200-cm^{-1} spectral range. Superimposed on this figure is a series of lines, which denote the optimized spectral range that was identified for the indicated

number of factors. These lines are positioned on the plot according to the logarithm of the corresponding SEM. A log scale is used to clearly separate the similar SEM values as the ideal set of conditions is reached. This figure clearly shows a reduction in prediction errors with an increase in the number of factors from one to four. A majority of the drop in SEM is accounted for with the first two factors. In addition, this figure indicates that the optimum spectral range depends on the number of factors used in the model. A relatively broad range is needed with only one factor and the prediction errors are large. The prediction error drops significantly with the addition of a second factor, and the spectral range narrows slightly toward the central wavelengths. Additional factors continue to lower the prediction errors, and the optimized spectral range varies within the major lysozyme absorption features. No statistically significant reduction in prediction error is realized beyond the fourth factor. The final optimum range incorporates information from both of the major protein absorption bands.

Digital Fourier filtering can reduce the size of the model further. This preprocessing step is known to enhance model performance by reducing the impact of baseline variations and high-frequency spectral noise.[23,24] Two digital Fourier filters were explored. The mean and standard deviation were 0.020*f* and 0.005*f* for the first filter and 0.017*f* and 0.006*f* for the second filter, where the units correspond to digital frequency units. This first set of filter parameters was taken from our previous work for measuring glucose in aqueous solutions from near-infrared spectra.[16] The wavelength optimization was repeated with the filtered spectra, and the best spectral range was 4610−4350 cm^{-1}. A three-factor PLS model was deemed optimal for the filtered spectra, and the resulting SEC, SEP, and MPEP are 0.31 mM, 0.34 mM, and 2.6%, respectively. The second set of filter parameters was obtained from an optimization routine that ranks filter parameters on the basis of the lowest SEM. Details of this optimization scheme are published elsewhere.[25] Wavelength optimization was again performed with these filtered spectra, and the 4480−4270-cm^{-1} spectral range coupled with only two PLS factors was ideal. Values of SEC, SEP, and MPEP are 0.38 mg/mL, 0.39 mg/mL, and 2.6%, respectively.

The above results are consistent with our earlier findings for the measurement of glucose in various biological matrixes. The Fourier filtering step effectively removes spectral variance in the form of baseline shifts, sloping baselines, and high-frequency noise. Compared to unfiltered spectra, fewer PLS factors are needed to account for spectral variance that is not associated with the concentration of analyte. Models with fewer factors are generally considered more robust, and the elimination of non-analyte-specific spectral variance by Fourier filtering enhances the overall utility of the method.

Inspection of the values summarized in Table 1 indicates that the model prediction errors are similar for all cases, except when raw single-beam spectra are used. SEP values range from 0.28 to 0.43 mg/mL for models based on processed spectra. These results demonstrate that some type of spectral processing is necessary in order to successfully extract lysozyme concentration information from the spectral data. By converting single beams to absorbance spectra, sufficient amounts of nonanalytical spectral variance are removed to permit successful quantification. Further improvements are realized by optimizing the wavelength range and by implementing either Fourier filtering or MSC. These additional steps reduce the number of latent variables, or factors, which provides a more robust model.

An example of model performance is provided in the concentration correlation plot presented in Figure 5. These data correspond to the three-factor model with MSC single-beam spectra over the optimized 4550−4190-cm^{-1} spectral range (see Table 1). Results are presented for both the calibration and prediction points. In both cases, model predictions closely follow the actual concentrations of lysozyme. A linear regression analysis indicates values for the correlation slope, *y*-intercept, and r^2 of 0.990 ± 0.002, 0.014 ± 0.004 mg/mL, and 0.9990, respectively, where these uncertainty values correspond to standard errors. Residual plots are shown in the Figure 5 insets, where the residuals of lysozyme predictions are plotted as a function of the lysozyme concentration and weight percent of microspheres. In both cases, lysozyme residuals are evenly scattered about the zero line, which indicates no systematic bias related to either of these parameters.

Calibration Models from the 0.6-μm Data Set. The same series of calibration models described above for the 6.4-μm data set was used to characterize models for measuring lysozyme in the presence of the 0.6-μm microspheres. Results are tabulated in Table 2.

The same pattern described above for the 6.4-μm data set is evident with the 0.6-μm data set. Essentially no analytical information is extractable from the raw single-beam spectra. The SDP for this data set is 8.40 mg/mL, which is similar to the prediction errors with raw single-beam spectra. Again, functional models are realized only after implementing either the MSC algorithm or converting the single-beam spectra to absorbance spectra. Again, wavelength optimization and Fourier filters reduce the number of model factors by removing nonanalytical spectral variance.

Compared to the 6.4-μm data set, prediction errors are larger for the 0.6-μm data. These larger errors are likely caused by the greater degree of variation within the 0.6-μm data, as indicated in Figure 2. Larger errors correspond to more scatter in the corresponding concentration correlation plots. An example of a concentration correlation plot for the 0.6-μm data is presented in Figure 6. Values plotted in this figure correspond to both the calibration and prediction data used for the model based on absorbance spectra with an optimized spectral range (entry six in Table 2). Both the calibration and predict points fall along the unity line. A linear regression analysis of these plotted points indicates values for the correlation slope, *y*-intercept, and r^2 of 0.995 ± 0.005, 0.05 ± 0.08 mg/mL, and 0.9986, respectively. The insets show residual plots as a function of lysozyme concentration and amount of microspheres in the samples. Again, no systematic bias is evident from these residual analyses.

Model Extrapolation. An experiment was carried out to determine the ability of PLS calibration models to predict lysozyme concentrations under the conditions where the level of scattering particles in the prediction samples is greater than those in the calibration samples. Under these conditions, the model must extrapolate beyond the conditions established by the calibration

(24) Marquardt, L. A.; Arnold, M. A.; Small, G. W. *Anal. Chem*. **1993**, *65*, 3271−3278.
(25) Chung, H.; Arnold, M. A.; Rhiel, M.; Murhammer, D. *App. Biochem. Biotechnol.* **1995**, *50*, 109−125.

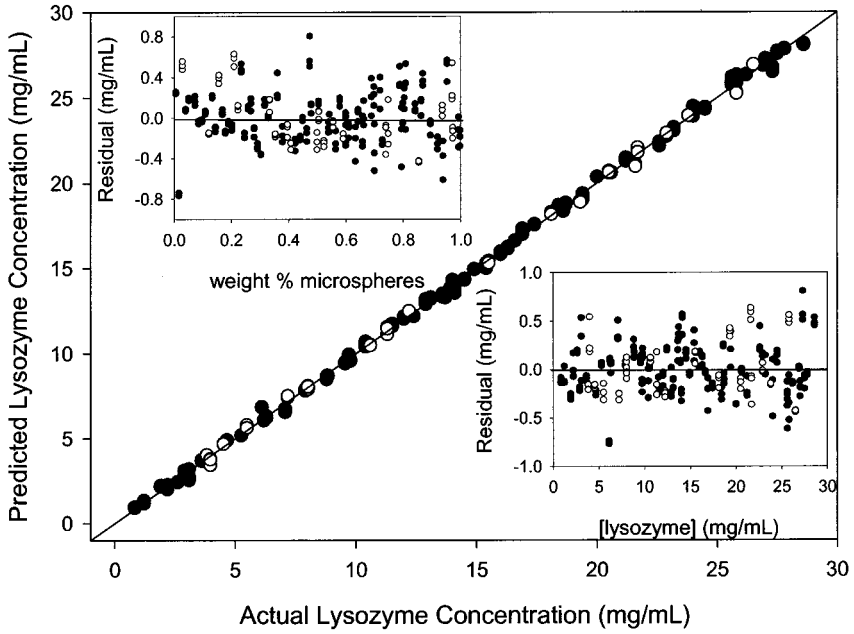

Figure 5. Concentration correlation plot for the three-factor PLS calibration model generated from MSC single-beam spectra taken with 6.4-μm microspheres over an optimized spectral range of 4550−4190 cm^{-1}. Open and closed symbols refer to calibration and prediction data, respectively, and the solid line represents the ideal unity line. Model residuals are plotted in the insets as functions of weight percent microspheres (upper left) and concentration of lysozyme (lower right).

Table 2. Lysozyme Calibration Models from Spectra with 0.6-μm-Diameter Microspheres

spectral type	spectral range (cm^{-1})	rank[a]	SEC (mg/mL)	SEP (mg/mL)	MPEP (%)
single beam	4800−4200	3	5.1	5.4	44
single beam	4690−4600	2	5.3	5.6	48
MSC single beam	4800−4200	4	0.69	0.52	2.7
MSC single beam	4700−4110	3	0.71	0.44	2.3
absorbance	4800−4200	6	0.68	0.49	2.6
absorbance	4740−4500	4	0.61	0.46	3.2
absorbance Fourier filtered (0.020, 0.005*f*)	4700−4590	2	0.67	0.56	4.9
absorbance Fourier filtered (0.020, 0.004*f*)	4700−4590	2	0.80	0.69	5.1

[a] Rank is the number of latent variables or factors used in the model.

training set. In this experiment, all spectra for samples with the 20 highest percentages of microspheres were used for prediction and the remaining spectra were used for calibration purposes. For both the 6.4- and 0.6-μm data sets, the calibration data consisted of spectra from samples with 0.834−28.6 mg/mL concentrations of lysozyme and 0.005−0.74% microspheres. In comparison, the prediction data included spectra from samples with 3.08−27.0 mg/mL concentrations of lysozyme and 0.75−0.998% microspheres. Successful measurements in the prediction data require the ability to extrapolate beyond the magnitude of scattering caused by microspheres within the calibration data set.

All the spectral processing schemes noted above and listed in Tables 1 and 2 were used to build models, and the ability to extrapolate was determined by comparing SEP values. Overall, SEP values were consistently higher for the 0.6-μm data set compared to the 6.4-μm data set. For example, SEP values are 4.13 and 0.58 mg/mL for the 0.6- and 6.4-μm data sets, respectively, for calibration models based on MSC single-beam spectra with spectral range optimization. In fact, the difference between the SEP and SEC of all models based on MSC or Fourier filtered spectra is significantly greater for the 0.6-μm data set. For the 6.4-μm data, the difference between SEP and SEC is approximately 0.2−0.3 mg/mL, whereas this difference is ∼3.5 mg/mL for the 0.6-μm data set.

The above results indicate that, despite higher rms noise levels, the calibration model from the 6.4-μm data is better able to predict lysozyme concentrations in samples with more particles than accounted for in the calibration model. Inspection of the data presented Figure 2 explains this finding. For both data sets, calibration models were generated from samples with 0.005−0.74% microspheres and predictions were made for samples with 0.75−0.998% microspheres. Although the spectra within the prediction data set correspond to samples with the same percentage of microspheres, the prediction spectra for the 6.4-μm data set are better represented in the calibration data compared to those in the 0.6-μm data set. As illustrated in Figure 2 for the 6.4-μm data, single-beam intensities for the prediction spectra are similar to those from spectra generated from samples with 0.6−0.75% microspheres. For the 0.6-μm data set, on the other hand, spectra in the prediction data set are not adequately represented in the

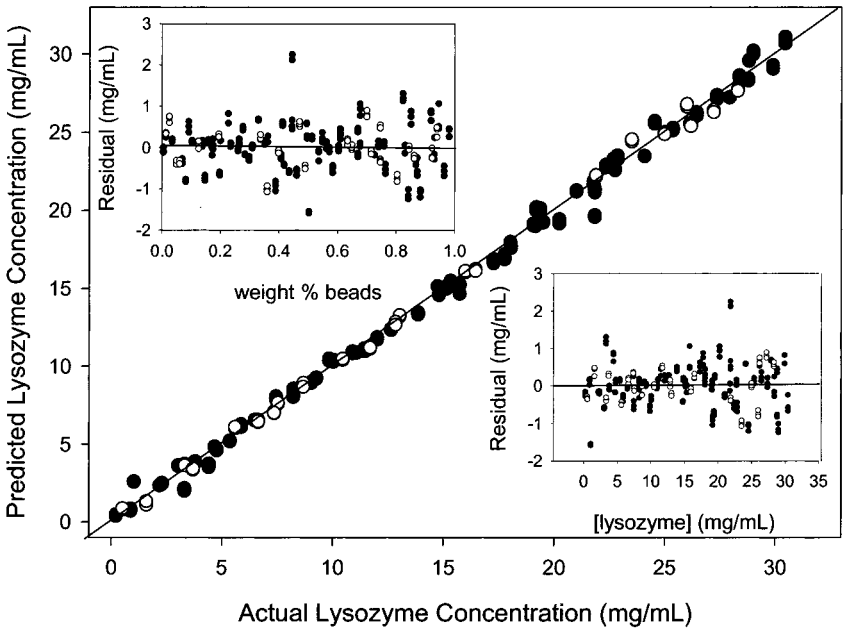

Figure 6. Concentration correlation plot for the four-factor PLS calibration model generated from absorbance spectra taken with 0.6-μm microspheres over an optimized spectral range of 4740−4500 cm^{-1}. Open and closed symbols refer to calibration and prediction data, respectively, and the solid line represents the ideal unity line. Model residuals are plotted in the insets as functions of weight percent microspheres (upper left) and concentration of lysozyme (lower right).

calibration data, as illustrated by major differences in the single-beam intensities shown in Figure 2. The poor match between spectra in the calibration and prediction data sets is responsible for this inability of these PLS models to successfully predict lysozyme concentrations in the presence of higher levels of microspheres.

CONCLUSIONS

Lysozyme concentrations can be accurately predicted from samples with a wide degree of scattering provided that the degree of scattering is adequately represented within the calibration model. PLS analysis of near-infrared spectra collected over the combination spectral region is able to predict protein concentrations with prediction errors on the order of 0.3 mg/mL in the presence of highly scattering media. Accurate calibration models are demonstrated for the situations where the diameter of the scattering microspheres is larger and smaller than the wavelength of the probing radiation. The best models are generated with single-beam spectra after MSC. Similar analytical performance is obtained from raw and Fourier filtered absorbance spectra. The results from this work indicate that, although variations of the scattering properties of the sample drastically alter sample transmission properties, acceptable noninvasive analytical measurements are possible if calibration standards are suitably designed to account for such scattering variations.

ACKNOWLEDGMENT

This research was funded through a grant from the Microgravity Science & Applications Division of the National Aeronautics and Space Administration (NAG8-1352).

Received for review September 5, 2001. Accepted May 5, 2002.

AC010976+