2000

Effect of Sample Complexity on Quantification of Analytes in Aqueous Samples by Near-Infrared Spectroscopy

Mark R. Riley
University of Nebraska-Lincoln, mriley3@unl.edu

Mark A. Arnold
University of Iowa, Iowa City, Iowa, mark-arnold@uiowa.edu

David W. Murhammer
University of Iowa, Iowa City, Iowa, david-murhammer@uiowa.edu

Effect of Sample Complexity on Quantification of Analytes in Aqueous Samples by Near-Infrared Spectroscopy

MARK R. RILEY,* MARK A. ARNOLD, and DAVID W. MURHAMMER

Department of Chemical and Biochemical Engineering (M.R.R., D.W.M) and Department of Chemistry (M.A.A.), University of Iowa, Iowa City, Iowa 52242

This study was undertaken to quantitate the impact of increasing sample complexity on near-infrared spectroscopic (NIRS) measurements of small molecules in aqueous solutions with varying numbers of components. Samples with 2, 6, or 10 varying components were investigated. Within the 10-component samples, three analytes were quantified with errors below 6% and seven of the analytes were quantified with errors below 10%. An increase in the number of varying components can substantially increase the error associated with measurement. A comparison of measurement errors across sample sets, as gauged by the standard error of prediction (SEP), reveals that an increase in the number of varying components from 2 to 6 increases the SEP by approximately 50%. An increase from 2 to 10 varying components increases the SEP by approximately 340%. While there appear to be no substantial correlations between the presence of a specific analyte and the errors associated with quantification of another analyte, several analytes do display a small degree of sensitivity to varying concentrations of certain background components. The analysis also demonstrates that calibrations containing an overestimation of the numbers of varying components can substantially increase measurement errors and so calibrations must be constructed with an accurate understanding of the number of varying components that are likely to be encountered.

Index Headings: **Near-infrared spectroscopy; NIRS; Aqueous solution; Quantification of analytes; Glucose.**

INTRODUCTION

Near-infrared spectroscopic (NIRS) techniques can provide noninvasive quantification of multiple analytes in aqueous materials.[1–9] Such capabilities are particularly useful for monitoring cell culture bioreactors that contain a liquid medium consisting of a large number of components.[2,8] Cell culture media for insect or mammalian cells typically contain 20 or more compounds such as carbohydrates, amino acids, and small organic species present at concentrations greater than 1 mM.

Many investigations to evaluate the feasibility of employing NIRS for bioreactor monitoring have focused on quantifying a limited number of components, usually numbering two to five.[1–9] Such measurement schemes typically focus on the primary cellular metabolites, glucose and glutamine, and cellular wastes, ammonia, and lactate.[10–14] Bioreactor operating conditions such as the flow rate of fresh culture medium (with high glucose and glutamine levels and no ammonia or lactate) are adjusted on the basis of the measurements so as to match the prevailing metabolic demand for nutrients and to minimize exposure to wastes.

Received 13 January 1999; accepted 10 September 1999.

* Author to whom correspondence should be sent. Present address: Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ 85721-0038.

Recent evidence[15] suggests that there may be a substantial need for monitoring a greater number of components in bioreactors for animal cells. Simpson et al.[15] have demonstrated that depletion of any single amino acid in a hybridoma cell cultivation can lead to apoptotic cell death. On the basis of this result, monitoring the glucose and glutamine concentrations alone may not suffice for maintaining a healthy culture. Quantification of potentially 10–20 analytes may be required. NIRS can provide such concentration information for a large number of analytes and requires only a short period of measurement time. To date, there have been few documented studies on the ability to quantify such a large number of components in a cell cultivation or on the effect of a large number of varying components on measurements. To the best of our knowledge, information on the impact of increasing sample complexity for identical instrumental conditions, sample preparation, and spectral quality has not been formally reported in the literature.

Samples that contain a large number of components present at high concentrations are likely to produce a greater degree of measurement error than would samples with fewer components. In the near-infrared, absorbance bands for amino acids and carbohydrates are broad and demonstrate maxima at similar positions (see Fig. 1). When a sample includes many of these analytes, NIR spectra contain substantial overlap in absorbance bands resulting from the similarly structured chemical species. If significant overlap exists between analytes, spectroscopic quantification methods can yield substantial errors.

The purpose of this work is to provide insight into: (1) the ability of NIRS to provide accurate, simultaneous quantification of a large number of components; (2) the effect of an increase in the number of varying components on measurement error; (3) the correlation between measurement error and the concentration of background constituents; and (4) the impact of over- or under-estimating the number of varying components in a calibration on the resultant measurement error. Some of these issues have been previously addressed to a small degree, primarily through qualitative comparisons. The study detailed here provides quantitative information on the above issues.

We have evaluated the effect of sample complexity through use of samples with increasing numbers of varying components present at concentrations ranging from as much as 0 to 60 mM. Samples were prepared by identical procedures. Each sample set contained random and uncorrelated concentrations of amino acids and carbohydrates typically present in most types of culture media

0003-7028/00/5402-0255$2.00/0

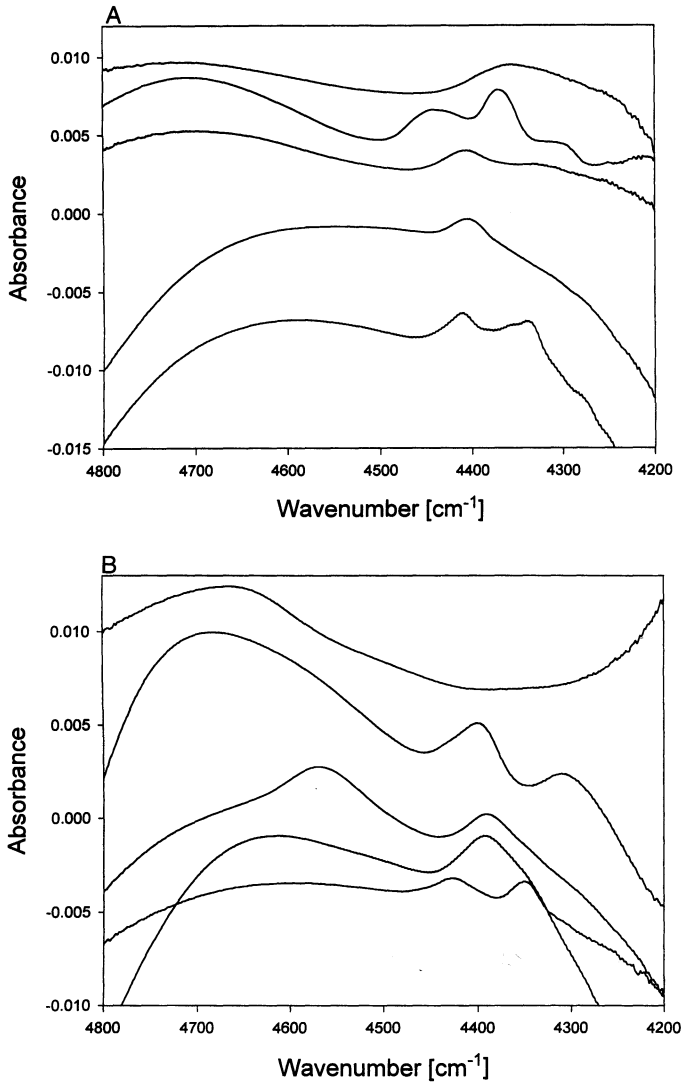


FIG. 1. NIR spectra of the 10 components quantified in this study. Listing from absorbances at the top of each figure at 4500 cm^{-1} and downwards: (**A**) aspartate, alanine, serine, cysteine, and leucine; (**B**) ammonia, glucose, glutamine, glutamate, and lactate. The concentration of each analyte is 25 mM.

for insect or animal cell cultivation. The first set of samples contained glucose and glutamine dissolved in an aqueous buffer, which are termed the 2-component samples. The second set contained alanine, cysteine, glucose, glutamate, glutamine, and leucine dissolved in the same buffer material. These are termed the 6-component samples. The third set contained alanine, ammonia, aspartate, cysteine, glucose, glutamate, glutamine, lactate, leucine, and serine. These are termed the 10-component samples. The concentration of each component varied over the same levels across sample sets. These concentration ranges are typically wider than those observed in typical cell culture media; however, the range encountered depends greatly on the culture medium formulation. Some culture media contain roughly 55–60 mM glucose, while others contain 10–22 mM glucose. The former typically do not reach total exhaustion of the available glucose; however, the latter can be limited by complete cellular consumption of all available glucose. To provide information for all concentration ranges, we have employed samples with wide concentration ranges from 0 to slightly more than that obtained in a culture media. Selection of components for each set of samples was based upon relative importance to the cellular metabolism. For example, glucose and glutamine are the primary nutrients for most animal and insect cell cultivations. The analytes present in the 6-component samples vary in concentration over the course of an insect cell cultivation to a larger degree than do the four additional analytes present in the 10-component samples.

Spectra were collected on the same spectrometer with the same sample cell temperature, instrumental conditions, and data processing procedures. Collections were made within a span of approximately six weeks. Both the root-mean-squared (rms) noise levels and the signal-to-noise ratios (SNR) were nearly the same for all samples, which suggests that differences observed in quantification are likely due to differences in the analyte environment caused by the presence or absence of other chemical species. A central focus for this study involves a comparison of the quantification of glucose and glutamine, which are present in all samples, in conditions of varying complexity and number of varying analytes.

MATERIALS AND METHODS

Samples were prepared by weighing random, known amounts of each analyte into an aqueous buffer solution containing 0.35 g/L $NaHCO_3$ and 1.013 g/L NaH_2PO_4 in deionized water, adjusted to pH = 6.35. These conditions are similar to the buffer conditions of Sf-900 II growth medium (Gibco, Grand Island, NY), commonly used to cultivate Sf-9 insect cells.

Spectra were collected on a Nicolet 550 Fourier transform infrared (FT-IR) spectrometer (Nicolet Analytical Instruments, Madison, WI) equipped with a 50 W tungsten–halogen lamp, calcium fluoride beamsplitter, and liquid nitrogen-cooled indium antinomide (InSb) detector. The optical sample cell was maintained at 27 °C, as in an insect cell bioreactor. Single-beam spectra were collected with a resolution of 2 cm^{-1} with 128 coadded scans from 5000 to 4000 cm^{-1} with a 1.5 mm optical pathlength. An interference filter (Barr and Associates, Cambridge, MA) was employed to isolate this spectral region. A background spectrum of pure buffer was collected after every fourth sample. Spectra of each set of samples were collected consecutively. All collections were completed within a six-week span. Additional details of the experimental setup may be found elsewhere.[16,17]

Triplicate spectra were collected consecutively for each sample, which were randomly divided into calibration, monitoring, and prediction data sets. Replicate spectra were always incorporated into the same sample set. For each analyte, the samples with the highest and lowest concentration were always placed in the calibration set. Partial least-squares (PLS) regression was applied to develop unique calibration models for each analyte. Data processing was performed on a Silicon Graphics Indigo computer operating spectral processing software provided by Professor Gary Small at the University of Ohio. A large number of calibration models were developed by varying the spectral range and number of PLS factors. Prediction errors for these different data sets were com-

TABLE I. Summary of quantification results for glucose and glutamine in samples containing only two varying species. Results represent an average of three rounds of modeling in which samples were divided into separate calibration, monitoring, and prediction sets. The monitoring set was used to determine model parameters (spectral range and number of PLS factors). This monitoring set was then added to the calibration set and used to predict a separate data set.

Analyte	Conc. range (mM)	Spectral range (cm^{-1})	No. of PLS factors	SEC (mM)	SEM (mM)	SEP (mM)	MPE (%)
Glucose	0–60	4600–4200	8	0.42	0.38	0.37	1.42
Glutamine	0–40	4590–4300	9	0.19	0.31	0.19	2.21

TABLE II. Summary of quantification results for the 6 analytes of varying concentrations in the 6-component samples.

Analyte	Conc. range (mM)	Spectral range (cm^{-1})	No. of PLS factors	SEC (mM)	SEM (mM)	SEP (mM)	MPE (%)
Alanine	0–40	4770–4230	13	0.26	0.18	0.17	1.07
Cysteine	0–20	4650–4250	16	0.16	0.33	0.38	4.58
Glucose	0–60	4590–4300	14	0.42	0.50	0.47	2.30
Glutamate	0–40	4780–4300	13	0.36	0.43	0.38	1.83
Glutamine	0–40	4750–4250	16	0.19	0.40	0.32	1.61
Leucine	0–20	4640–4300	11	0.14	0.24	0.23	4.48

puted as standard error of calibration (SEC), standard error of monitoring (SEM), and standard error of prediction (SEP).

To select calibration model parameter including the spectral range and number of PLS factors to be applied for each calibration set, one fourth of the calibration samples were temporarily removed from the set and employed as an internal validation, or monitoring set. Optimal conditions were determined for the remaining calibration samples used to predict concentrations in the monitoring set (SEM). Once such conditions were determined, the monitoring samples were returned to the calibration set, and the same calibration parameters were used to establish PLS models. By employing a monitoring set, one ensures that model parameters are independent of the samples present in the prediction data set, which should produce more robust calibration models that are less dependent on the specific samples employed. Each analysis was performed in triplicate so as to reduce any sample-dependent effects. For each round of processing, calibration, monitoring, and prediction, samples were randomly selected.

In an effort to thoroughly investigate the many possible combinations of calibration parameters, a C-shell computer script was written to systematically develop calibration models with varying numbers of PLS factors and with varying spectral ranges within the 5000–4000 cm^{-1} region. The script followed a modified grid search that provided the unbiased evaluation of many calibration data sets in a short period of time. For each model size (as defined by the total number of PLS factors), the script searched for spectral ranges within the 5000–4000 cm^{-1} region that contained significant analyte information, as judged by a low SEM. Initially, SEM values were calculated for 100 cm^{-1} wide regions at 100 cm^{-1} intervals beginning with 4100–4000 cm^{-1} and progressing to 5000–4900 cm^{-1}. The region that yielded the minimum SEM was judged to contain significant analytical information. The upper and lower values of this range were increased and decreased by a predetermined amount (50 cm^{-1} for the first iteration, then 20 cm^{-1}, 10 cm^{-1}, and finally 5 cm^{-1}), yielding a combination of eight new spectral ranges focused around the region that contains analyte dependent information. Corresponding SEM values were calculated and the region with the lowest SEM was selected as the optimum spectral range. This process was repeated four times with each iteration stepping through decreasing deviations in the spectral range. The number of PLS factors was then incremented and another spectral range search was implemented. For all models, the spectral loadings were evaluated to ensure that adequate spectral information was available so that calibrations were not "over-modeled". SEP values were then determined for the conditions found to be optimal for the monitoring set.

RESULTS AND DISCUSSION

Infrared spectra of three separate sets of samples were collected with varying components numbering 2, 6, or 10 (not including the concomitant changes in the water concentration). Figure 1 presents absorbance spectra of each analyte at a concentration of 25 mM. Standard errors for analyte quantification can be reasonably compared across conditions only if the levels of both signal intensity and noise are equivalent. Root-mean-squared noise levels were calculated by comparison of spectra to a first-order polynomial in the 5000–4900 cm^{-1} region, and signal-to-noise ratios were calculated for the peak to peak from 5000 to 4000 cm^{-1}. Calculations of the rms noise were performed on Omnic software from Nicolet Analytical Instruments (Madison, WI). With the use of software, rms noise is calculated as the deviation from a linear relationship and so influenced the selection of the spectral range for rms noise calculation as a region where similar linear relationships could be found. The region of 5000–4900 cm^{-1} contains few analyte absorbances and so is reasonably consistent across samples. Additional calculations for SNR were performed on spectral processing software provided by Professor Small at the University of Ohio. Values of 0.031 (rms noise) and 1100 (SNR) for the 2-component samples, 0.025 (rms noise) and 1200 (SNR) for the 6-component samples, and 0.023 (rms noise) and 1200 (SNR) for the 10-component samples were obtained. The generally close agreement between root-mean-squared noise levels and between the signal-to-noise ratios lends credence to comparisons between sample sets.

Glucose and Glutamine Measurements. Glucose and glutamine are present in all three sets of samples with concentrations ranging from 0–60 mM and 0–40 mM, respectively. SEPs for these analytes are particularly low for the 2-component samples both in terms of the absolute error and the mean percent error (Table I). These errors are in general agreement with previously reported measurement levels for such conditions.[4]

A comparison of the levels of prediction error for glucose and glutamine in the 2-, 6-, and 10-component samples reveals that an increase in the number of varying components has a substantial effect on SEP (Tables I–III). For these comparisons the calibration and prediction

TABLE III. Summary of quantification results for the 10 analytes of varying concentrations in the 10-component samples.

Analyte	Conc. range (mM)	Spectral range (cm^{-1})	No. of PLS factors	SEC (mM)	SEM (mM)	SEP (mM)	MPE (%)
Alanine	0–40	4730–4240	15	0.49	0.91	0.78	5.15
Ammonia	0–25	4700–4320	17	0.46	0.98	0.91	8.15
Aspartate	0–25	4720–4250	15	0.86	1.03	1.18	13.01
Cysteine	0–20	4780–4200	19	1.07	1.71	1.38	17.83
Glucose	0–60	4710–4240	20	0.83	1.57	1.14	5.71
Glutamate	0–40	4590–4250	18	0.67	1.04	1.08	9.82
Glutamine	0–40	4800–4270	19	0.40	0.97	0.72	6.99
Lactate	0–25	4560–4280	12	0.44	0.34	0.30	3.45
Leucine	0–20	4730–4290	16	0.36	0.50	0.49	8.43
Serine	0–25	4730–4240	18	0.72	1.28	1.14	11.22

sets contain the same number of varying components. SEPs in the 6-component samples increase by a factor of 1.3 for glucose and 1.7 for glutamine compared to the 2-component samples. The mean percent error for glucose measurement increases by a factor of 1.6, but that for glutamine decreases by a factor of 0.7. Measurement errors in the 10-component samples increase for glucose by a factor of 3 and for glutamine by a factor of 3.8 compared to the 2-component samples. The mean percent error for glucose measurement increases by a factor of 2.4 and that for glutamine increases by a factor of 4.3. In general, these results demonstrate that an increasing number of components present at relatively high concentrations increases the difficulty in quantifying a specific analyte through spectroscopic techniques.

Absolute errors and mean percent errors change by distinct amounts because the impact of additional components is concentration dependent. Low-concentration samples have the largest effect on the mean percent error, whereas high concentration samples have a greater contribution to the absolute error. Some of these variations could be attributed to the increasing numbers of PLS factors utilized with increasing numbers of sample components. Likely causes for glucose errors increasing by different amounts than those of glutamine are that the location of the spectral information provided by each analyte is distinct, analytes can demonstrate altered sensitivities to the sample environment, and concentration ranges for these analytes are dissimilar.

Note that, in constructing calibrations for each analyte, the number of PLS factors employed increases with increasing number of varying components. Any relationship of SEC, SEM, or SEP observed across sample sets must be analyzed only in relative terms. Increasing the number of PLS factors typically decreases the model SEC unless the data set has been over-modeled. For all models, we have analyzed the spectral loadings so as to ensure that over-modeling of the calibration set did not occur.

Measurement of Alanine, Cysteine, Glucose, Glutamate, Glutamine, and Leucine. The analytes present in the 6-component samples are also present in the 10-component samples over the same concentration ranges. For these analytes, SEPs are consistently lower in the 6-component samples (Table II) than in the 10-component samples (Table III). Comparing the prediction errors obtained for analytes in the 6-component samples to those in the 10-component samples, alanine errors increase substantially from 0.17 mM to 0.78 mM. Similarly, cysteine errors increase from 0.38 to 1.38 mM. In the 6-component samples, the average mean percent error is 2.6% and the average absolute error is 0.32 mM, whereas these averages increase by approximately a factor of 3 to 8.9% and 0.93 mM, respectively, in the 10-component samples.

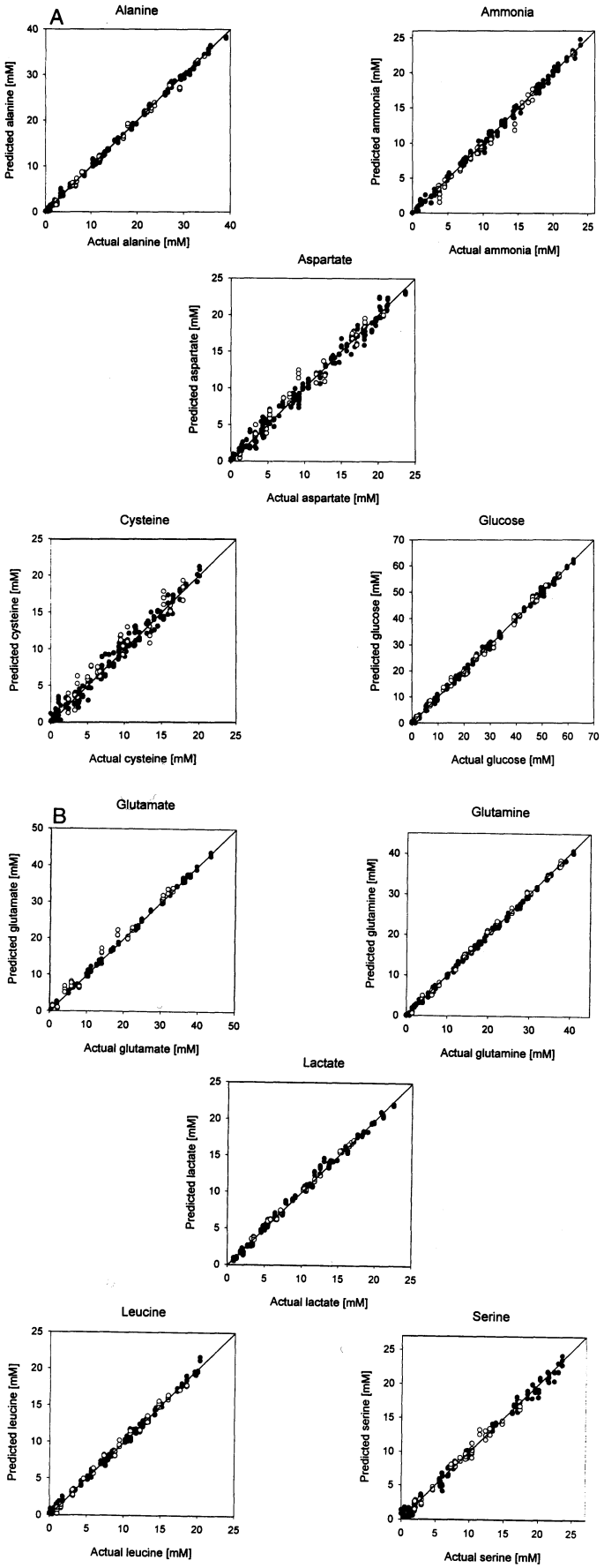
Sample complexity also impacts PLS calibration model parameters. The number of PLS factors required to generate satisfactory calibration models increases with the number of varying components. On the average, the optimum spectral range as determined by a sequential grid search increases with the sample complexity. For the 6-component samples, the average ideal spectral range encompasses 425 cm^{-1}, but these same 6 analytes require an average spectral range of 475 cm^{-1} in the 10-component samples. This increase is most likely due to the large degree of overlap in analyte information within the 5000–4000 cm^{-1} region. An increase in the optimal spectral range increases the amount of information required by PLS to make accurate predictions.

Despite the overall increase in SEP in the 10-component samples, some of the analytes may be quantified with a fairly low level of error. Predictions of the concentrations of all analytes in the 10-component samples are presented in Fig. 2. The SEP for leucine measurement is 0.49 mM, which represents roughly a doubling in the SEP compared to the 0.23 mM SEP in the 6-component samples. Glucose, lactate, and alanine all have mean percent errors below 6%. These three analytes appear to have spectral characteristics that are somewhat distinct from those provided by the other components present in these samples (Fig. 1). Alanine has four distinct absorbance maxima centered at 4710, 4440, 4380, and 4310 cm^{-1}. Glucose has three distinct maxima including a broad peak at 4700 cm^{-1}. Lactate has two distinct absorbance features centered at 4420 and 4370 cm^{-1}; however, absorbance maxima at these positions are quite common. Features in the NIR are broad, and some degree of overlap exists between nearly all absorbance maxima for these structurally similar compounds.

In the 10-component samples, analytes with the three largest percent errors (cysteine, serine, and aspartate) require the widest spectral ranges. Cysteine has an absolute error of 1.38 mM and a mean percent error of 17.8%, and these predictions display a large degree of scatter throughout the concentration range. An absorbance spectrum of pure cysteine has only one distinct feature, which is apparent at 4400 cm^{-1}, a common position for absorbance features for these analytes. Aspartate and serine have large SEPs, and both calibration and prediction samples display scatter. Spectra of aspartate and serine present several absorbance features; however, these features are broad and generally not distinct from those presented by the other sample components.

On the basis of the types and location of spectral features, ammonia might be anticipated to be difficult to quantify by NIR spectroscopy in the 5000–4000 cm^{-1} region. The spectrum of ammonia (Fig. 1B) provides only a broad absorbance feature centered around 4650 cm^{-1}. This appears to not be a hindrance to the quantification of ammonia as both the SEP and the mean percent error are near the average level for analytes present in the 10-component samples.

Correlation Analysis. One of the requirements of the



application of Beer's law to correlate analyte concentrations with light absorbance is that the chemical species present in the sample not substantially interact with each other. If a sizeable interaction exists, a correlation may become apparent between the error for measurement of one analyte with the concentration of one or more of the background constituents. To evaluate the effect of chemical species concentrations on the measurement of each analyte in the 10-component samples, we performed a correlation analysis between analyte prediction error for each sample and the concentration of additional components in each sample. All analytes, except glutamine, demonstrate a positive correlation between that analyte's concentration and the measurement error. Correlation coefficients are approximately 0.1 for all analytes. Glutamine yields only a slightly negative correlation that might be attributed to more accurate measurement at high concentrations.

A correlation analysis was performed between analyte prediction error for each sample and the concentration of each other chemical species present in the sample. The average correlation coefficient from this analysis was 0.11. The strongest correlation appeared between the alanine prediction error and the glutamate concentration, resulting in a correlation coefficient of 0.33, as the largest errors in alanine measurement appear in samples with high glutamate concentrations.

Five sets of correlations between glutamate errors and the concentrations of specific chemical species yielded correlation coefficients greater than 0.22. These conditions were between the glutamate error and serine (0.29), alanine (0.26), glutamine (0.27), leucine (0.22), and the total concentration of added chemical species (0.26). Only two other conditions [the aforementioned alanine–glutamate correlation and that between serine error and the lactate concentration (0.28)] yielded correlation coefficients greater than 0.21. These results for glutamate are unusual and suggest that glutamate measurements are quite sensitive to the sample environment.

To provide a basis for comparison, we compared the correlation coefficients described above to correlation coefficients obtained between analyte prediction errors and a series of randomly selected chemical species concentrations. Two thousand sets of random concentrations were compared to the actual prediction errors, and the largest correlation coefficient obtained was 0.26. Therefore, any correlations between prediction errors and the actual component concentrations that are larger than 0.26 represent some relationship that is substantially stronger than the largest correlation that could be achieved entirely due to chance. This influence could be due to interactions between chemical species in solution or due to overlap in spectral information.

Cross-Sample Analysis. The literature on PLS analysis recommends that proper calibration models must

←

FIG. 2. Concentration correlation plots for the analytes quantified in the 10-component samples. (**A**) Alanine, ammonia, aspartate, cysteine and glucose; (**B**) glutamate, glutamine, lactate, leucine, and serine. Open symbols represent prediction samples; closed symbols represent calibration samples.

TABLE IV. Glucose models employing a cross-study analysis.

Calibration set	Prediction set	Spectral range (cm^{-1})	No. of PLS factors	SEC (mM)	SEM (mM)	SEP (mM)	MPE (%)
2 Comp.	6 Comp.	4580–4260	5	0.63	7.57	8.50	77.24
2 Comp.	10 Comp.	4580–4200	7	0.46	8.37	13.02	158.9
6 Comp.	2 Comp.	4500–4300	9	0.64	0.78	0.74	2.70
6 Comp.	10 Comp.	4690–4200	9	1.23	3.06	4.81	98.48
10 Comp.	2 Comp.	4700–4210	15	1.41	1.53	1.83	8.10
10 Comp.	6 Comp.	4620–4200	14	1.40	1.40	1.78	18.42

contain a greater degree of sample complexity than the prediction samples and that the calibration analyte concentrations must span the range of concentrations present in the prediction samples.[18] When calibration models are being developed to be applied to a biological process, the identity and magnitude of concentration changes of background components are often not known *a priori.* The usual strategy is either to ignore changes in the concentrations of most components (thereby assuming that these changes are small) or to account for as many components as possible. Neither strategy can be used with any certainty without some understanding of the ramifications of developing calibrations with too few or too many varying components.

To address this issue, we performed a cross-sample analysis by employing mismatched calibration and prediction data sets. For example, the 10-component samples were used to generate calibrations to predict the glucose concentration in the 2-component samples and in the 6-component samples. This analysis was performed only for glucose and glutamine as these components are present in all three sets of samples. All calibration sets contained both samples with a concentration of 0 and samples with the maximum concentration evaluated. Model parameters were determined by using a monitoring set composed of samples removed from the experimental data set that was to serve as the prediction set. These monitoring samples remained separate from the prediction set for SEP calculation. The results for this cross-sample analysis for the six possible combinations of calibration and prediction data sets are summarized in Tables IV and V for glucose and glutamine, respectively.

Calibrations consisting of the lower complexity samples yield very poor predictions of the higher complexity samples. Glucose absolute errors are 4.8–13.0 mM (Table IV), whereas glutamine errors are somewhat lower at 1.4–2.8 mM (Table V). Mean percent errors are around 100% for glucose models, which indicates that the low-concentration samples are predicted with very poor accuracy. For glutamine models, the mean percent errors with the use of lower complexity calibrations are roughly the same magnitude as those for higher complexity calibrations of low-concentration samples. In some cases, the glutamine mean percent errors are moderate (around 20%), whereas the absolute errors are above 2.8 mM. Apparently, use of lower complexity calibrations for glutamine measurement yields smaller errors for low-concentration samples than does the use of lower complexity calibrations for glucose measurement. Clearly, a lack of accounting of several varying components compromises prediction capability.

TABLE V. Glutamine models employing a cross-study analysis.

Calibration set	Prediction set	Spectral range (cm^{-1})	No. of PLS factors	SEC (mM)	SEM (mM)	SEP (mM)	MPE (%)
2 Comp.	6 Comp.	4740–4490	12	0.15	1.19	1.40	13.04
2 Comp.	10 Comp.	4760–4500	17	0.10	2.25	2.89	21.29
6 Comp.	2 Comp.	4800–4250	13	0.41	0.44	0.67	7.77
6 Comp.	10 Comp.	4750–4360	9	0.80	1.88	2.88	19.87
10 Comp.	2 Comp.	4700–4200	18	0.63	0.76	1.01	16.25
10 Comp.	6 Comp.	4770–4200	13	1.00	1.65	1.47	15.32

Conversely, calibrations with greater complexity could be used in some cases to generate reasonable predictions of samples with a lower degree of complexity. Calibrations consisting of the 6-component samples could be used to predict analyte concentrations in the 2-component samples with moderate errors of 0.73 mM for glucose and 0.67 for glutamine. These levels of error were substantially the smallest obtained for the cross-samples analyses. The 10-component samples used as a calibration set to predict analyte concentrations in the 2- and 6-component studies yield errors all greater than 1.0 mM. Apparently, PLS recognizes greater similarity between the 6-component samples and the 2-component samples than between the 6-component samples and the 10-component samples even though there is a difference of four components each.

Figure 3 presents glucose concentration predictions resulting from four of these cross-study analyses. Figures 3a and 3b display calibration and prediction samples for a calibration set consisting of the 10-component samples and the 2- and 6-component samples employed as prediction sets, respectively. In both of these figures, calibration models with a high degree of complexity are applied to prediction samples with fewer varying components. Predicted glucose concentrations and the known concentrations are in fairly good agreement, and predictions have little bias. Calibration and prediction samples here have similar amounts of scatter, which is reflected in the generally similar levels of error.

Figures 3c and 3d, on the other hand, demonstrate the possible problems resulting from use of a calibration data set which does not incorporate all of the chemical species variations present in the prediction samples. Figures 3c and 3d display results of the 2-component samples used as a calibration data set for the 10- and 6-component samples, respectively. In both circumstances, the appropriate trends in glucose concentrations are obtained; however, there is a substantial degree of scatter. For a number of samples, the error is greater than the actual concentration; some low-concentration glucose samples yield negative glucose concentration predictions. The absolute error in glucose predictions is roughly the same at low and

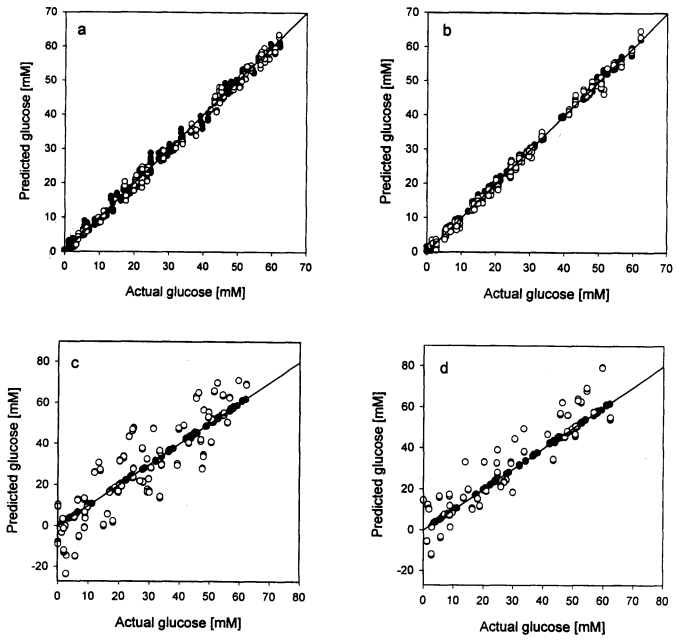


FIG. 3. Concentration correlation plots for cross-study analyses of glucose: (**a**) 10-component samples for calibration and 2-component samples for prediction; (**b**) 10-component samples for calibration and 6-component samples for prediction; (**c**) 2-component samples for calibration and 10-component samples for prediction; (**d**) 2-component samples for calibration and 6-component samples for prediction. Open symbols represent prediction samples; closed symbols represent calibration samples.

high glucose concentrations. Clearly, the presence of eight or four additional varying components not present in the calibration samples but present in the prediction samples can have a sizeable influence on concentration predictions.

As demonstrated here, calibration models must incorporate similar or slightly greater levels of sample variations in order to yield accurate analyte predictions. Errors presented for the cross-sample analyses are always higher than those obtained when both the calibration and prediction data sets have the same number of varying components. Incorporation of a greater number of varying components than are present in the prediction samples can degrade measurement accuracy, as can be observed by comparing the glucose and glutamine results in Tables I–V. Incorporation of six varying components into a calibration when actually 2-component variations are required can increase prediction error by a factor of 2 to 3. This cross-sample analysis suggests that when spectroscopic calibrations are being prepared it is important not only to include all major variations in calibration samples but also to avoid inclusion of variations beyond those which are present in the prediction samples.

The compounds evaluated here have comparable molar absorptivities, and so at similar concentrations will yield similar influences on the resultant collected spectra. The concentration of each component likely would influence the impact of over- or underestimation of the number of varying components. An error in inclusion of a component with a small or negligible concentration would likely have a smaller effect than would the neglect of a component with a high concentration. The analysis and discussion presented here focus on the situation where analyte concentrations are of the same order of magnitude. Future discussions should address the effect of concentrations.

CONCLUSION

An increase in the number of varying components can at times substantially increase the error associated with NIRS measurements of analytes in aqueous solutions. An increase in the number of varying compounds from 6 to 10 on the average increases error by a factor of 3. Even with 10 varying components, however, it is possible to quantify some analytes with low levels of error, which may be useful for industrial monitoring requirements. There appear to be no substantial correlations between the presence of any of these analytes and the errors associated with quantification; however, glutamate is sensitive to changes in the concentrations of glutamine and serine, and serine is somewhat sensitive to changes in the lactate concentration. When generating spectroscopic calibrations to be applied to samples with a large number of components, application of greater numbers of varying components than are actually changing in the samples to be quantified may significantly degrade measurement capabilities.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Martin Rhiel for development of the C-shell computer script for optimization of spectral ranges and numbers of PLS factors and Professor Gary Small from the Center for Intelligent Chemical Instrumentation in the Department of Chemistry at Ohio University for providing the spectral processing software. Financial support for this work was provided by American Cyanamid Corporation and from NASA (#NAG 9-824).

1. A. G. Cavinato, D. M. Mayes, Z. Ge, and J. B. Callis, Anal. Chem. **62,** 1977 (1990).
2. T. Yano and M. Harata, J. Ferment. Bioeng. **77,** 659 (1994).
3. X. Zhou, H. Chung, M. A. Arnold, M. Rhiel, and D. W. Murhammer, "Selective Measurement for Glutamine and Asparagine in Aqueous Media by Near-Infrared Spectroscopy", in *Biosensor and Chemical Sensor Technology: Process Monitoring and Control,* K. R. Rogers, A. Mulchandani, and W. Zhou, Eds., American Chemical Society Symposium Series 613 (ACS Books, Washington, D.C., 1995), Chap. 12, pp. 116–132.
4. H. Chung, M. A. Arnold, M. Rhiel, and D. W. Murhammer, Appl. Spectrosc. **50,** 270 (1996).
5. J. W. Hall, B. McNeil, M. J. Rollins, I. Draper, B. G. Thompson, and G. Macaloney, Appl. Spectrosc. **50,** 102 (1996).
6. P. Fayolle, D. Picque, B. Perret, E. Latrille, and G. Corrieu, Appl. Spectrosc. **50,** 1325 (1996).
7. R. P. Kasprow, A. J. Lange, and D. J. Kirwan, Biotech. Prog. **14,** 318 (1998).
8. M. J. McShane and G. L. Cote, Appl. Spectrosc. **52,** 1073 (1998).
9. D. L. Doak and J. A. Phillips, Biotech. Prog. **15,** 529 (1999).
10. W. M. Miller, C. R. Wilke, and H. W. Blanch, Biotech. Bioeng. **33,** 477 (1989).
11. S. F. White, A. P. F. Turner, U. Bilitewski, J. Bradley, and R. D. Schmid, Biosensors Bioelect. **10,** 543 (1995).
12. L. Xie and D. I. C. Wang, Biotech. Bioeng. **43,** 1175 (1994).
13. N. Vriezen, B. Romein, K. Luyben, and J. P. van Dijken, Biotech. Bioeng. **54,** 272 (1997).
14. S. S. Ozturk, J. C. Thrift, J. D. Blackie, and D. Naveh, Biotech. Bioeng. **53,** 372 (1997).
15. N. H. Simpson, R. P. Singh, A. Perani, C. Goldenzon, and M. Al-Rubeai, Biotech. Bioeng. **59,** 90 (1998).
16. M. R. Riley, M. Rhiel, X. Zhou, M. A. Arnold, and D. W. Murhammer, Biotech. Bioeng. **55,** 11 (1997).
17. M. R. Riley, M. A. Arnold, D. W. Murhammer, E. L. Walls, and N. DelaCruz, Biotech. Prog. **14,** 527 (1998).
18. T. Naes and T. Isaksson, Appl. Spectrosc. **43,** 328 (1989).