



ELSEVIER

Analytica Chimica Acta 384 (1999) 333–343

ANALYTICA
CHIMICA
ACTA

Evaluation of nonlinear model building strategies for the determination of glucose in biological matrices by near-infrared spectroscopy

Qing Ding^a, Gary W. Small^{a,*}, Mark A. Arnold^b

^aCenter for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Ohio University, Athens, OH 45701, USA

^bDepartment of Chemistry, Optical Science and Technology Center, University of Iowa, Iowa City, IA 52242, USA

Received 7 April 1998; received in revised form 8 September 1998; accepted 18 September 1998

Abstract

Nonlinear model building techniques are applied to near-infrared spectra to predict glucose concentrations in samples containing an aqueous matrix of varied concentrations of bovine serum albumin (BSA) and triacetin. The triacetin is used to model triglycerides in human blood, and the BSA is used to model blood proteins. The non-linear model building techniques included in this study are quadratic partial least-squares regression (QPLS), stepwise QPLS, and PLS followed by artificial neural networks (PLS-ANN). The optimal models obtained for glucose provide standard errors of prediction of 0.53 mM, 0.54 mM, and 0.48 mM for the QPLS, stepwise QPLS and PLS-ANN models, respectively, over the clinically relevant concentration range of 1–20 mM. These results indicate significant improvement in prediction performance relative to that obtained with linear PLS models. This improvement is confirmed through the use of *F*-tests at the 95% confidence level. The significant quadratic terms included in the stepwise QPLS models also confirm that nonlinear information exists in the data set studied. This suggests that there is a need to develop suitable nonlinear model building strategies for noninvasive blood glucose determinations. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Near-infrared; Glucose; Neural network; Nonlinear modeling; Partial least-squares

1. Introduction

The development of near-infrared (near-IR)-based techniques for the noninvasive determination of glucose concentrations in blood has attracted significant interest in recent years [1–4]. The motivation for this measurement is to free diabetics from the need to use invasive methods for monitoring their blood glucose levels. The concept is to extract glucose information from the spectra collected through transmission or

reflectance measurements of a vascular region of the human body. Since the glucose signal is very small and overlapped with the signals arising from other matrix constituents such as water, protein, and fat, the development of appropriate data analysis strategies is essential to the successful implementation of non-invasive glucose sensors.

Our approach to the development of methodology for potential application to this problem is to identify the physical and chemical parameters which affect the glucose measurement by increasing the matrix complexity systematically. Through this approach, data handling methodologies can be developed step by step

*Corresponding author. Tel.: +1-740-593-1748; fax: +1-740-593-0148; e-mail: small@ohiou.edu

to overcome potential problems in the noninvasive measurement of blood glucose. We began with a feasibility study to determine glucose at physiologically relevant concentrations in an aqueous matrix [5]. Our research has demonstrated the capability to measure glucose concentrations accurately over the near-IR spectral region of 5000–4000 cm^{-1} while the sample matrix complexity increased from the phosphate buffer to phosphate buffer with a constant concentration of bovine serum albumin (BSA) [6], undiluted bovine plasma [7], and variable matrices of triacetin and BSA [8]. In the latter study, the calibration models built by coupling digital filtering with partial least-squares (PLS) regression provided standard errors of prediction (SEP) of 0.5 and 0.2 mM in the triacetin and BSA matrices, respectively.

In these previous studies, the model building strategies for the determination of glucose concentrations have primarily been based on linear PLS regression, which has been demonstrated as a useful technique to extract glucose information from near-IR spectra with increased matrix complexity. However, when the relationship between absorbance and concentration deviates from the Beer–Lambert law, linear multivariate calibration techniques such as PLS regression or principal component regression (PCR) are not optimal for modeling the nonlinear relationship [9,10]. Gemperline et al. [10] attributed the sources of nonlinear response to factors such as nonlinear detector response, stray light, and baseline shift, and to shifts of position and width of absorption bands arising from changes in sample temperature and solvent composition. In the actual noninvasive measurement of blood glucose, because of the extremely small glucose signal, light scattering from human blood and tissue, the temperature-sensitive water background, and spectral interferences from blood constituents such as protein and triglycerides, calibration models built with linear PLS regression may not be ideal to extract the glucose information. With mid-infrared spectra, Bhandare et al. [11] reported improved results in determining glucose in whole blood by combining PLS with an artificial neural network (ANN) as compared to the use of either PLS or PCR alone.

In this paper, the use of nonlinear PLS regression (i.e., quadratic PLS (QPLS) and stepwise QPLS) and ANNs to incorporate nonlinear information into the glucose calibrations is studied. These nonlinear cali-

bration models are evaluated and compared to linear calibration models based on PLS regression. Samples of glucose in an aqueous matrix of BSA and triacetin (termed GTB) are used in this investigation.

2. Experimental

2.1. Apparatus and reagents

The near-IR spectra used in this work were collected with a Digilab FTS-60A Fourier transform spectrometer (Bio-Rad, Cambridge, MA). The spectrometer was configured with a 100 W tungsten–halogen source, CaF_2 beamsplitter, and liquid nitrogen-cooled InSb detector. The near-IR range of 5000–4000 cm^{-1} served as the focus of the analysis. A K-band optical interference filter (Barr Associates, Westford, MA) was used to isolate this spectral region. Transmission measurements were performed with samples held in an Infrasil quartz cell with 2 mm path length. Sample temperatures were controlled to the range of 37–38°C through the use of a water jacketed cell holder and refrigerated circulator. Temperatures of the sample solutions were monitored during the spectral collection by use of a type-T thermocouple and digital thermocouple meter (Omega Engineering, Stamford, CT).

The GTB data set was constructed by use of samples prepared in 0.1 M phosphate buffer at pH 7.4, to which 0.044% 5-fluorouracil was added as a preservative. In this data set, BSA and triacetin were used to model blood proteins and triglycerides, respectively. Proteins and triglycerides represent two common interferences in the spectroscopic determination of blood glucose. Separate stock solutions of 50 mM glucose (ACS reagent, Fisher Scientific, Fair Lawn, NJ), 190 g/l BSA (Cohn fraction V powder, product no. A4503, Sigma Chemical Co., St. Louis, MO), and 35 g/l triacetin (99%, Sigma Chemical Co.) were prepared in phosphate buffer. Individual samples were made by mixing appropriate volumes of the stock solutions and diluting to 50 ml with the phosphate buffer. A factorial design was employed to vary the concentrations of the three components to minimize correlations among the component concentrations. With this factorial design, the data set consisted of ten levels of glucose (1, 3, 5, 7, 9, 11, 13, 15, 17,

19 mM), four levels of triacetin (1.4, 2.1, 2.8, 3.5 g/l), and four levels of BSA (49.3, 64.4, 79.8, 96.7 g/l). This produced a total of $10 \times 4 \times 4 = 160$ samples.

2.2. Procedures

For each GTB sample, single-sided interferograms containing 16384 points were collected, and 256 co-added scans were used. Samples were run in a random order with respect to concentration to minimize the chance of correlations between glucose, BSA, or triacetin concentrations and any time-dependent data artifacts. The importance of minimizing such correlations in the near-IR-based analysis of glucose has recently been reported [12].

Three replicate interferograms were collected for each sample. At the beginning of each data collection session and after every five samples, interferograms of phosphate buffer were acquired for subsequent use in the calculation of spectra in absorbance units. The collected interferograms were Fourier processed to single-beam spectra by use of software resident on the Bio-Rad SPC-3200 computer controlling the spectrometer. Triangular apodization and Mertz phase correction were employed. This produced single-beam spectra with a point spacing of 1.9 cm^{-1} .

The single-beam spectra were transferred to a Silicon Graphics 4D/460 computer (Silicon Graphics, Mountain View, CA), where the remainder of the data analysis was performed. This computer operated under Irix (version 5.2). The software used for QPLS and stepwise QPLS regressions was implemented in FORTRAN 77. Multiple linear regression computations were performed with subroutines from the IMSL software package (IMSL, Houston, TX). The software used for PLS-ANN was implemented in Matlab (version 4.2c, The MathWorks, Inc., Natick, MA).

3. Theory

3.1. Overview of linear PLS and QPLS regression

Linear PLS regression has been widely used in near-IR spectroscopy because of its ability to extract analyte information from multicomponent samples, even if the absorption bands are relatively broad and overlapped with bands arising from interferences. As

implemented in this work, linear PLS regression consists of two steps [13,14]. The first step is data reduction. In the context of this specific application, the mean-centered spectral absorbance data matrix (n (spectra) \times p (spectral points)) is decomposed into a scores matrix by taking into consideration the covariance between the independent and dependent variables. The absorbance values in the spectral matrix form the independent variables, and the vector of analyte (glucose) concentrations corresponding to the spectra in the calibration set forms the dependent variable. The dimensionality of the resulting scores matrix ($n \times h$) is typically significantly less than that of the original data matrix, where h is the number of latent variables (PLS factors) to be optimized by the procedure described below. The collinearity of the original data matrix is also removed since the resulting PLS scores are orthogonal. In the second step, the scores matrix is regressed against the analyte concentrations to build a multivariate regression model of the form

$$c_i = b_0 + b_1x_{i,1} + \dots + b_hx_{i,h} \quad (1)$$

where c_i is the predicted analyte concentration for the i th spectrum, the x_i values are the computed PLS scores for spectrum i , and the b terms are regression coefficients.

Although linear PLS regression has proven to be an effective model building technique for multivariate calibration, researchers have worked to develop more generalized and potentially more powerful nonlinear PLS regression methods, beginning with the relatively simple QPLS regression [15–17]. The concept of QPLS regression is to model the curved relationships between the dependent variables and the resulting scores matrix obtained from the original spectral matrix. In practice, with only one dependent variable, the first step of QPLS regression is the same as that of linear PLS regression, i.e., to decompose the spectral data matrix ($n \times p$) into the orthogonal and lower dimensional PLS scores matrix ($n \times h$). The difference between QPLS and linear PLS regression is found in the second step. QPLS regression models have the form

$$c_i = b_0 + b_1x_{i,1} + \dots + b_hx_{i,h} + b_{h+1}x_{i,1}^2 + \dots + b_{2h}x_{i,h}^2 \quad (2)$$

where the terms in Eq. (2) are the same as those in Eq. (1). Note that the quadratic polynomial terms of the PLS scores are included in the QPLS regression model to help explain curved relationships between the analyte concentrations and PLS scores.

QPLS regression is seen to be more generalized and powerful than linear PLS regression by considering that linear PLS regression is a special case of QPLS regression. However, QPLS regression models are more easily overfit than those computed with linear PLS regression. Motivated by this concern, stepwise QPLS regression was also studied. The difference between stepwise and standard QPLS regression is that in the second step, Eq. (2), stepwise QPLS regression only includes the terms in the model which are statistically significant. The stepwise regression procedure operates by adding terms to the model one at a time. At each step, the term is added that accounts for the greatest amount of unexplained variance in the dependent variable. If no variable meets a minimum level of significance as determined by a statistical *F*-test, the procedure terminates.

3.2. Overview of artificial neural networks

ANNs are simplified mathematical models which imitate the known activity of biological neural networks. ANNs have become popular in solving chemical problems because of their power and flexibility to explain nonlinearity [18,19]. However, ANNs also have the drawbacks of a long training time, models that are easily overfit, and models that are hard to interpret. In this paper, a multilayer feedforward neural network with one hidden layer and full connectivity between the adjacent layers was used to explain the nonlinear information in the glucose calibration models. A schematic of this network architecture is presented in Fig. 1. Note that the only output node is the predicted glucose concentration. The input layer is used to distribute the input variables to the next layer for processing. The weights represent the strength of the connection between the nodes in the adjacent layer.

The operation of the network consists of two phases. The first one is the forward propagation of activation to predict the output results (i.e., glucose concentrations). Then, by comparing predicted and known glucose concentrations, the errors in prediction

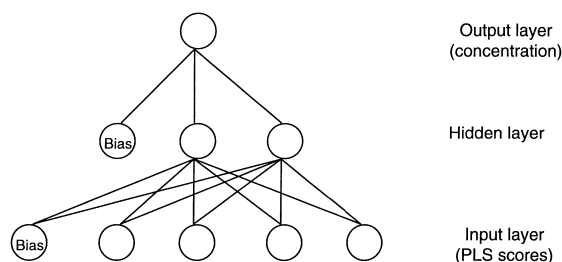


Fig. 1. Schematic architecture of the PLS-ANN model used in this study. The PLS scores were used as inputs to the networks. The output is the predicted glucose concentration corresponding to the spectrum that produced the input scores.

are used in conjunction with a numerical optimization algorithm to adjust the weights to minimize the error. One such iteration is termed an epoch. The number of epochs needed for training an ANN is a parameter which can be optimized for different applications to obtain the minimal training error without overfitting the models. The net input of a given node can be calculated by the dot product of the weight vector of the node with its input vector

$$v_j = \sum_{i=1}^n w_{ji}x_i + b_j \quad (3)$$

where v_j is the net input for node j , the x_i are the input values, the w_{ji} values are the weights associated with those inputs, b_j is the offset or bias to the node j , and n is the number of connections or synapses for the node.

The output of a given node can be determined by passing the net input v to a transfer function. In this paper, a hyperbolic tangent transfer function was used for the hidden layer, and a linear transfer function for the output layer. The Levenberg–Marquardt learning rule was selected to adjust the weights and minimize the training error of the computed neural networks [20,21].

In the context of a quantitative spectral analysis, the absorbance values can be used as the inputs for the first layer of the networks [22]. However, researchers have found that the network training is faster when PLS or principal component analysis (PCA) are employed as a preprocessing technique to generate orthogonal inputs to the ANN [9,10]. In our research, we first attempted to use the original absorbance values as the inputs to the ANN, but the networks did not converge to a minimum training error. Subsequently, the PLS

scores computed from the absorbance spectra were used as inputs to the ANN. This will be termed the PLS-ANN method. The choice of PLS over PCA was motivated by a desire to compare the PLS-ANN technique with QPLS regression. The optimization procedure for the selection of the number of nodes in the hidden layer and the number of PLS factors (input nodes) will be detailed below.

4. Results and discussion

4.1. Data set characteristics

Fig. 2 shows absorbance spectra for glucose at a concentration of 50 mM (A), BSA at 190 g/l (B), and triacetin at 35 g/l (C). The glucose C–H combination band centered at 4400 cm^{-1} has been found useful for extracting information on glucose [5]. As evident in the figure, this glucose band is severely overlapped with the BSA absorption band centered at approximately 4373 cm^{-1} and the triacetin band centered near 4446 cm^{-1} . Assuming there is a linear relationship between absorbance and concentration, the absorbance values corresponding to the largest concentrations of glucose, BSA, and triacetin in the GTB data set can be estimated [23]. The absorbance of glucose at 4400 cm^{-1} for 19 mM is estimated as $927\text{ }\mu\text{AU}$

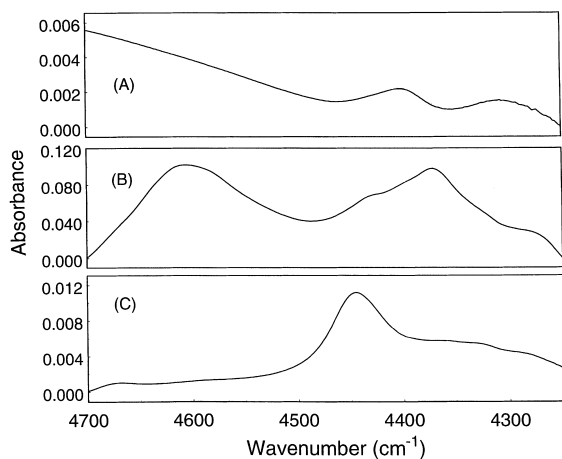


Fig. 2. Near-IR absorbance spectra of (A) glucose at 50 mM, (B) BSA at 190 g/l, (C) triacetin at 35 g/l in phosphate buffer over the region of $4700\text{--}4250\text{ cm}^{-1}$. Note that the three spectra are not on the same y-axis scale.

(microabsorbance units), while the absorbance for 94.9 g/l BSA at 4373 cm^{-1} is approximately $84561\text{ }\mu\text{AU}$, and the absorbance for 3.5 g/l triacetin at 4446 cm^{-1} is about $3450\text{ }\mu\text{AU}$.

The glucose signal in the GTB spectra is so small that it is overwhelmed by interference from BSA and triacetin. This can be observed in the spectrum of a GTB sample consisting of 19 mM glucose, 79.83 g/l BSA, and 2.80 g/l triacetin (Fig. 3). It is clear from the figure that the spectra of the GTB samples are dominated by the protein features. The glucose information cannot visually be observed. Also, nonlinear information may exist in this data set because of nonlinear baseline variation, scattering by protein molecules, and the shift of position and width of the absorption bands arising from slight changes in temperature.

In order to apply and evaluate the different model building strategies, 75% of the whole data set was randomly partitioned into a calibration set (120 samples, 360 spectra). The remaining 25% of the data set was used as a prediction set. The calibration set was further randomly divided into a calibration subset (96 samples, 288 spectra) and monitoring set (24 samples, 72 spectra). The calibration models were built with the use of only the calibration set. The prediction set was only used as an independent validation set to evaluate the performance of the optimal calibration models built with the calibration set. All three replicate spectra were carried together into the same subset.

4.2. Optimization Results

4.2.1. QPLS and stepwise QPLS regression

Analogous to the optimization of linear PLS calibration models, the spectral range used in the computation of the PLS scores and the number of PLS factors employed in the calibration model are critical parameters in the QPLS and stepwise QPLS methods. To select the optimal calibration model (i.e., the best combination of spectral range and number of PLS factors), the effect of changing the spectral range was studied by varying the size and location of the range in a systematic manner. The number of PLS factors used with QPLS was varied with each selected spectral range. For a given range size, the spectral window was shifted across the region of $4750\text{--}4200\text{ cm}^{-1}$ in steps of 25 cm^{-1} . The range size varied from 200 cm^{-1} to 500 cm^{-1} in increments of 25 cm^{-1} . The contiguous

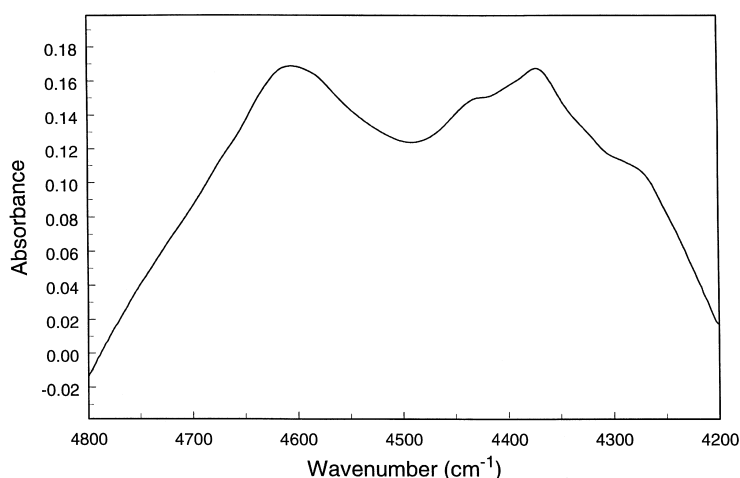


Fig. 3. Near-IR absorbance spectrum of a GTB mixture of 19 mM glucose, 2.80 g/l triacetin, and 79.83 g/l BSA.

spectral points within each selected range were used in the computation of the PLS scores. For each spectral range, the number of PLS factors employed in the calibration model was varied from 10 to 20.

For each selected spectral range and number of PLS factors, a calibration model was built with the use of the calibration subset and evaluated by using the model to predict the glucose concentrations corresponding to the spectra in the monitoring set. The optimal calibration models were determined on the basis of the best prediction results with the monitoring set. The statistic used to make these comparisons was the standard error of monitoring (SEM), computed as

$$\text{SEM} = \frac{\sqrt{\sum_{i=1}^{n_m} (c_i - \hat{c}_i)^2}}{n_m} \quad (4)$$

where n_m is the number of spectra in the monitoring set, c_i is the actual analyte concentration of spectrum i , and \hat{c}_i is the analyte concentration predicted by the model. Once the optimal spectral range and number of PLS factors were selected, the full calibration set (i.e., the calibration subset + monitoring set) was employed to build a final model. The prediction set was then used to test the performance of the final model. The prediction set was thus an independent validation data set which was not used at any stage during the model optimization.

With the above optimization procedure, the optimal calibration model for QPLS based on the spectral

range 4675–4275 cm^{-1} and 17 PLS factor (plus 17 quadratic terms) provided an R^2 of 99.96%, a standard error of calibration (SEC) of 0.39 mM, and a standard error of prediction (SEP) of 0.54 mM. The SEC was calculated by

$$\text{SEC} = \sqrt{\frac{\sum_{i=1}^{n_c} (c_i - \hat{c}_i)^2}{n_c - h - 1}} \quad (5)$$

where n_c is the number of spectra in the calibration set, h is the number of PLS factors used in the calibration model, and the other terms are the same as in Eq. (4). The SEP was calculated by using an equation analogous to Eq. (4), where n_p , the number of spectra in the prediction set, was substituted for n_m .

Concentration correlation plots of the optimal QPLS model are provided in Fig. 4(A) and (B). Fig. 4 (A) plots the results for spectra in the calibration set. The prediction set results are plotted in Fig. 4 (B). Excellent correlation between the predicted and actual glucose concentrations is observed.

The optimization procedure used for the selection of the stepwise QPLS models was the same as that used for the QPLS models. As noted previously, the difference between the QPLS and stepwise QPLS models is that the stepwise models only include the terms that have been judged statistically significant. The stepwise QPLS models could improve the efficiency of use of the calibration models, and the prediction performance is potentially made more

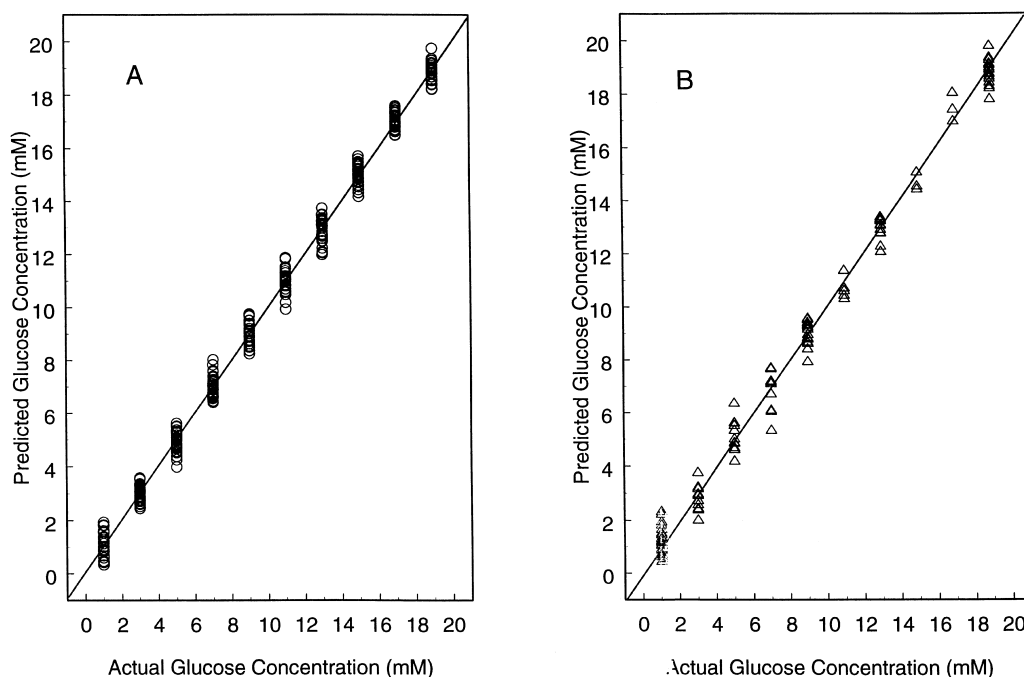


Fig. 4. Concentration correlation plots for GTB absorbance spectra with QPLS regression (A) for the calibration data set, (B) for the prediction data set. The calibration model was based on the spectral range of $4675\text{--}4275\text{ cm}^{-1}$ and 17 linear plus 17 quadratic PLS scores.

stable than that of the QPLS models by removal of statistically insignificant terms. Also, stepwise QPLS can be used to test the significance of the nonlinear terms in the models. Therefore, it could be a viable tool to determine the existence of nonlinear relationships that correspond to the functional form of the model (quadratic in this case).

The significance levels for entering and deleting variables for stepwise QPLS regression were both set at 95%. The optimal stepwise QPLS calibration model based on the spectral range of $4700\text{--}4300\text{ cm}^{-1}$ and 21 significant terms (18 linear terms + 3 quadratic terms) provided an R^2 of 99.59%, an SEC of 0.37 mM, and an SEP of 0.54 mM. The 21 significant terms included in the optimal calibration model and their corresponding partial F -values are listed in Table 1. Note that the first 18 terms are linear PLS scores and terms 20, 23, 27 are quadratic terms corresponding to linear scores 2, 5 and 9. Besides, selection of all 18 linear terms, several quadratic terms are also selected as significant contributors to the model. This suggests that nonlinear relationships exist in the data set, and it is useful to employ the nonlinear model building strategy.

Interestingly, the first several PLS factors did not have the largest partial F -values. The most plausible explanation for this is that because the glucose signal is so small relative to the spectral baseline variation, the first few PLS factors are used to explain the baseline variations rather than to explain correlation with the glucose concentrations. This is also one of the reasons why there is a need for many PLS factors for this data set. Glucose concentration correlation plots for the optimal stepwise QPLS calibration model are shown in Fig. 5(A) and (B) for the calibration and prediction sets, respectively. Again, excellent correlation between predicted and actual glucose concentration is noted.

4.2.2. PLS-ANN

Previous studies with PLS-ANN often consider PLS as a preprocessor to the ANN to accelerate the training of the network [9]. From the point of view of nonlinear PLS regression, however, PLS-ANN is not much different from QPLS and stepwise QPLS regression. The ANN is used to model the relationships (both linear and nonlinear) between the glucose concentra-

Table 1
Partial *F*-values for terms selected for the optimal stepwise QPLS model

Terms ^a	Partial <i>F</i>
1	256.366
2	1800.93
3	3039.03
4	4610.81
5	9084.19
6	29799.2
7	24330.3
8	268.571
9	1499.07
10	498.513
11	615.053
12	514.108
13	168.845
14	93.6452
15	108.755
16	56.4546
17	75.2988
18	39.7713
20	11.0635
23	9.46546
27	4.25729

^a Terms 1–18 are linear terms, while terms 20, 23, and 27 are quadratic terms corresponding to linear terms 2, 5, and 9, respectively.

tions and PLS scores. However, the ANN is not limited to a specific functional shape like the quadratic polynomial.

The optimization procedure used to compute the PLS-ANN models was similar to that employed in the QPLS and stepwise QPLS regression studies. For a given range size, the spectral range was changed over the region of 4750–4250 cm^{-1} in steps of 25 cm^{-1} . The range size was varied from 325 to 425 cm^{-1} in increments of 25 cm^{-1} . For each selected spectral range, the number of PLS factors employed was varied from 14 to 18. Fewer levels of the variables were used in this optimization than in the QPLS and stepwise QPLS studies because of the significant computational time required to train each network.

In addition to optimizing the spectral range and the number of PLS factors used as inputs to the networks, the number of hidden units was varied from 1 to 5. For each combination of spectral range, number of PLS factors, and number of hidden units, three replicate networks were trained based on three random sets of initial weights. This increased the chance of finding the lowest training error for each network configuration.

The number of training epochs was studied in order to design an appropriate training protocol. On the basis of our experiences with the Levenberg–Marquardt

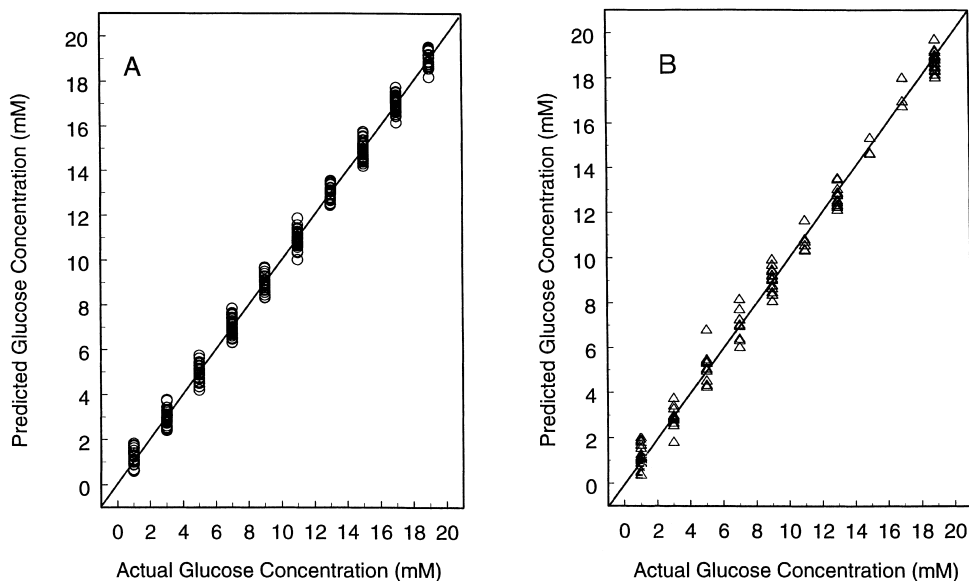


Fig. 5. Concentration correlation plots for GTB absorbance spectra with stepwise QPLS regression (A) for the calibration set, (B) for the prediction set. The calibration model was based on the spectral range of 4700–4300 cm^{-1} and 18 linear plus three quadratic PLS scores.

learning algorithm, the networks trained much faster than with standard steepest descent (backpropagation) methods. With steepest descent learning rules including momentum and adaptive learning rate parameters, networks trained with several thousand epochs did not converge. By contrast, with the Levenberg–Marquardt learning rule, networks almost always converged within approximately 100 epochs.

The effect of the number of epochs was studied with the use of a spectral range of 4720–4295 cm^{-1} , because this range was found to be optimal with linear PLS models. For this study, the number of PLS factors was varied from 12 to 18, and the number of hidden units was varied from 1 to 5. The number of epochs was changed from 5 to 200 with increments of five epochs. For each training increment of five epochs, the current network was applied to predict the concentrations corresponding to the spectra in the monitoring set, and the SEM was computed. The optimal number of epochs in each case was taken as that which produced the lowest SEM value.

The results obtained through this procedure indicated that the optimal number of epochs is influenced by the number of PLS factors and the number of hidden units used. There was no unique number of epochs that could be selected for all cases. However, there was a general trend that the networks almost always converged within 100 epochs for all different combinations of the number of PLS factors and the number of hidden units. In addition, networks trained with 100 epochs exhibited little or no evidence of overfitting. Therefore, subsequent networks were trained for a fixed limit of 100 epochs.

As with the optimization of the QPLS and stepwise QPLS models, the configurations of the PLS-ANN models were optimized by use of the calibration subset and monitoring set. The optimal configurations were determined on the basis of the best prediction results

with the monitoring set. In comparing network configurations, the lowest SEM value among the three replicate networks was used. Once the optimal architecture was determined, the full calibration set was used, and three replicate networks were trained with different random initial weights. Among these three replicates, the network with the lowest training error was applied to predict the concentrations corresponding to the spectra in the prediction set. As before, the prediction set was used only as an independent validation set to test the performance of the optimal network configurations. The SEC and SEP were estimated by the same equations used with the QPLS regressions.

The optimal PLS-ANN architecture was selected with the spectral range of 4700–4325 cm^{-1} , 17 PLS factors, and four hidden units. This PLS-ANN model provided an SEC of 0.36 mM and an SEP of 0.48 mM. The glucose concentration correlation plots for this model are shown in Fig. 6(A) and (B) for the calibration and prediction sets, respectively. The predicted glucose concentrations are again highly correlated with the actual glucose concentrations.

4.2.3. Comparison of results

The best results with the optimal models for linear PLS, QPLS, stepwise QPLS and PLS-ANN models are summarized in Table 2. The results with linear PLS were obtained from previous work in our laboratory [23]. The number of quadratic terms included in the QPLS and stepwise QPLS regressions are listed in the nonlinear term column, along with the number of hidden units used in the PLS-ANN configuration. The results from QPLS, stepwise QPLS, and PLS-ANN were compared with those obtained with the linear PLS calibration models. The statistical *F*-test was used at the 95% confidence level to test the significance of the improvement with these nonlinear model building techniques. The results indicated that all three non-

Table 2
Results comparison for different model building strategies

Calibration models	Spectral range (cm^{-1})	No. of PLS factors	Nonlinear terms	SEC (mM)	SEP (mM)
Linear PLS	4719–4294	16	None	0.41	0.66
QPLS	4675–4275	17	17	0.39	0.53
Stepwise QPLS	4700–4300	18	3	0.37	0.54
PLS-ANN	4700–4325	17	4 ^a	0.35	0.48

^a Number of hidden units.

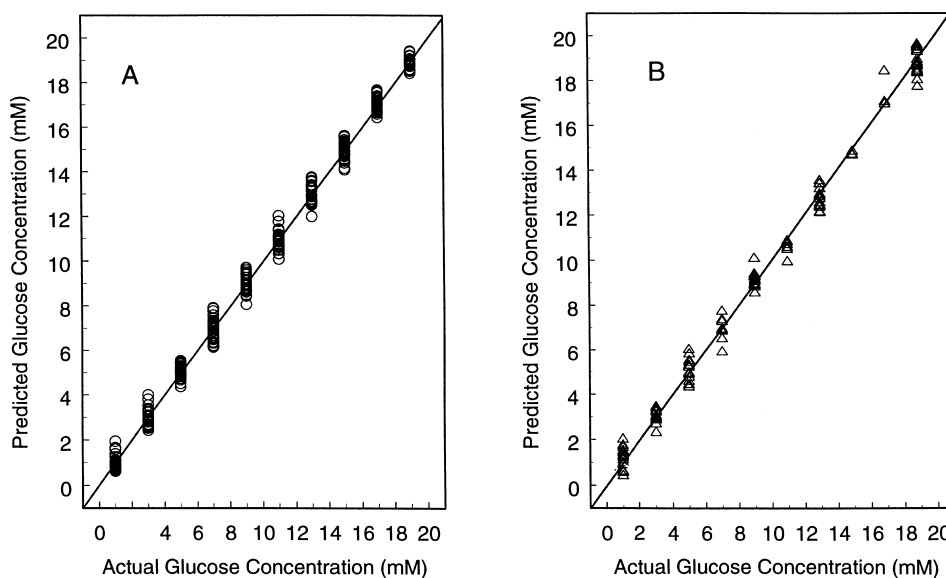


Fig. 6. Concentration correlation plots for GTB absorbance spectra with PLS-ANN (A) for the calibration set, (B) for the prediction set. The PLS-ANN configuration was based on 17 input nodes of 17 linear PLS scores computed over the spectral range of 4700–4325 cm^{-1} and four hidden units.

linear models provided significant improvement in the prediction results relative to those obtained with linear PLS regression. However, prediction results among the nonlinear models were not significantly different.

5. Conclusions

The significant quadratic terms included in the stepwise QPLS models and the significant improvement in prediction ability obtained with the nonlinear models relative to the linear PLS models suggest that nonlinear information exists in the GTB data set. Thus, it can be concluded that nonlinear model building strategies are needed to explain the nonlinear relationships between glucose concentrations and the PLS scores computed from the spectral data. Because the stepwise QPLS models always have fewer terms than the corresponding QPLS models, stepwise QPLS regression is preferred to QPLS regression.

The PLS-ANN models are more flexible than the stepwise QPLS models because the nonlinear information explained by the PLS-ANN method is not

limited to a fixed nonlinear shape such as that encoded by a quadratic polynomial. However, the optimization of the PLS-ANN configurations is more complicated than that for the stepwise QPLS models, and the iterative training of the PLS-ANN models is also slower. The prediction results with the PLS-ANN models were not significantly better than those obtained with the stepwise QPLS models. This may be because there is only mild nonlinear information in the data set or because the optimization of the PLS-ANN configurations needs to be improved. On the basis of the results obtained with the GTB data set, it is reasonable to expect that suitable nonlinear model building strategies will help in the noninvasive determination of blood glucose.

Acknowledgements

This research was supported by the National Institutes of Health under grant DK45126. Mutua Mattu and Ndumiso Cingo are thanked for their assistance in collecting the spectral data used in this work. The Department of the Army is acknowledged for providing the Silicon Graphics 4D/460 computer.

References

- [1] I. Amato, *Science* 258 (1992) 892.
- [2] D.M. Haaland, M.R. Robinson, G.W. Koepp, E.V. Thomas, R.P. Eaton, *Appl. Spectrosc.* 46 (1992) 1575.
- [3] R. Marbach, Th. Koschinsky, F.A. Gries, H.M. Heise, *Appl. Spectrosc.* 47 (1993) 875.
- [4] M.A. Arnold, in: G.R. Kost (Ed.), *Handbook of Clinical Laboratory Automation, Robotics, and Knowledge Optimization*, Wiley, New York, 1996, pp. 631–647.
- [5] M.A. Arnold, G.W. Small, *Anal. Chem.* 62 (1990) 1457.
- [6] L.A. Marquardt, M.A. Arnold, G.W. Small, *Anal. Chem.* 65 (1993) 3271.
- [7] G.W. Small, M.A. Arnold, L.A. Marquardt, *Anal. Chem.* 65 (1993) 3279.
- [8] S. Pan, H. Chung, M.A. Arnold, G.W. Small, *Anal. Chem.* 68 (1996) 1124.
- [9] C. Borggaard, H.H. Thodberg, *Anal. Chem.* 64 (1992) 545.
- [10] P.J. Gemperline, J.R. Long, V.G. Gregoriou, *Anal. Chem.* 63 (1991) 2313.
- [11] P. Bhandare, Y. Mendelson, R.A. Peura, G. Janatsch, J.D. Karuse-Jarres, R. Marbach, H.M. Heise, *Appl. Spectrosc.* 47 (1993) 1214.
- [12] M.A. Arnold, J.J. Burnmeister, G.W. Small, *Anal. Chem.* 70 (1998) 1773.
- [13] H. Martens, T. Næs, *Multivariate Calibration*, Wiley, New York, 1989, Chap. 7.
- [14] P. Geladi, B.R. Kowalski, *Anal. Chim. Acta* 185 (1986) 1.
- [15] S. Wold, N. Kettaneh-Wold, B. Skagerberg, *Chemom. Intell. Lab. Syst.* 7 (1989) 53.
- [16] S. Wold, *Chemom. Intell. Lab. Syst.* 14 (1992) 71.
- [17] I. Frank, *Chemom. Intell. Lab. Syst.* 27 (1995) 1.
- [18] B.J. Wythoff, *Chemom. Intell. Lab. Syst.* 18 (1993) 115.
- [19] J.R.M. Smits, W.J. Melssen, L.M.C. Buydens, G. Kateman, *Chemom. Intell. Lab. Syst.* 22 (1994) 165.
- [20] H. Demulth, M. Beale, *Neural Network Toolbox User's Guide*, The MathWorks, Natick, MA, 1994.
- [21] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1987, Chap. 14.
- [22] J.R. Long, V.G. Gregoriou, P.J. Gemperline, *Anal. Chem.* 62 (1990) 1791.
- [23] A.S. Bangalore, R.E. Shaffer, G.W. Small, *Anal. Chem.* 68 (1996) 4200.