

Genetic Algorithm-Based Protocol for Coupling Digital Filtering and Partial Least-Squares Regression: Application to the Near-Infrared Analysis of Glucose in Biological Matrices

Ronald E. Shaffer[†] and Gary W. Small*

Center for Intelligent Chemical Instrumentation, Department of Chemistry, Ohio University, Athens, Ohio 45701

Mark A. Arnold

Department of Chemistry, University of Iowa, Iowa City, Iowa 52242

A multivariate calibration procedure is described that is based on the use of a genetic algorithm (GA) to guide the coupling of bandpass digital filtering and partial least-squares (PLS) regression. The measurement of glucose in three different biological matrices with near-infrared spectroscopy is employed to develop this protocol. The GA is employed to optimize the position and width of the bandpass digital filter, the spectral range for PLS regression, and the number of PLS factors used in building the calibration model. The optimization of these variables is difficult because the values of the variables employ different units, resulting in a tendency for local optima to occur on the response surface of the optimization. Two issues are found to be critical to the success of the optimization: the configuration of the GA and the development of an appropriate fitness function. An integer representation for the GA is employed to overcome the difficulty in optimizing variables that are dissimilar, and the optimal GA configuration is found through experimental design methods. Three fitness function calculations are compared for their ability to lead the GA to better calibration models. A fitness function based on the combination of the mean-squared error in the calibration set data, the mean-squared error in the monitoring set data, and the number of PLS factors raised to a weighting factor is found to perform best. Multiple random drawings of the calibration and monitoring sets are also found to improve the optimization performance. Using this fitness function and three random drawings of the calibration and monitoring sets, the GA found calibration models that required fewer PLS factors yet had similar or better prediction abilities compared to calibration models found through an optimization protocol based on a grid search method.

Partial least-squares (PLS) regression has become an increasingly popular tool for use in developing multivariate calibration models in analytical chemistry. An application in which PLS regression has been particularly successful is near-infrared

spectroscopy.¹ Near-IR spectra of complex samples often contain analyte information obscured by overlapping spectral bands or spectral artifacts such as noise and baseline variation. PLS regression is often able to extract relevant analyte information from such spectra, thereby allowing useful calibration models to be constructed.

In previous work in our laboratory, a protocol was established for coupling PLS regression with bandpass digital filtering.² This work was motivated by the desire to compute multivariate calibration models for glucose in biological matrices using Fourier transform near-infrared (FT-near-IR) absorbance spectra. It was demonstrated that the use of digital filtering as a spectral preprocessing technique could improve the results of PLS regression by removing spectral artifacts such as noise and baseline variation prior to the model building step. The key to the success of this strategy was the definition of optimal values for five important variables: the position and width of the bandpass filter, the starting and ending points of the spectral range submitted to the PLS regression, and the number of terms (latent variables) employed in the calibration model. In previous work, this optimization was performed by fixing the spectral range and model size and then employing a grid search to vary the filter position and width.

The drawback of this approach is that it fails to take into account interaction effects present among the five variables. Because bandpass digital filtering and PLS regression both attempt to extract analyte information from the spectra, these steps need to work together rather than independently to realize any synergistic effects. However, a grid search involving all the variables would be prohibitively large and require extensive computation time. If the combination of bandpass digital filtering and PLS regression is to be established as a general-purpose tool, an efficient protocol for finding the optimal values of these variables must be established.

This paper describes the application of genetic algorithms (GAs) to this optimization problem. GAs are flexible numerical optimization techniques based on the concepts of genetics and

(1) *Near Infra-Red Spectroscopy. Bridging the Gap Between Data Analysis and NIR Applications*; Hildrum, K. I., Isaksson, T., Næs, T., Tandberg, A., Eds.; Ellis Horwood: New York, 1992.

(2) Small, G. W.; Arnold, M. A.; Marquardt, L. A. *Anal. Chem.* **1993**, *65*, 3279–3289.

[†] Present address: Environmental Chemistry and Sensor Chemistry Section, Code 6116, Chemical Dynamics and Diagnostics Branch, Naval Research Laboratory, Washington, DC 20375-5342.

natural selection.³⁻⁹ The measurement of glucose in three different biological matrices by FT-near-IR spectroscopy is used to explore this methodology. Issues addressed in this work include the development of a GA configuration that can optimize dissimilar variables and the design of an optimal fitness function for use in monitoring the progress of the optimization.

EXPERIMENTAL SECTION

Methods. A focus of research in our laboratories has been the development of a clinical glucose sensor based on near-IR spectroscopy.^{2,10-13} As part of these research efforts, data sets have been collected with glucose in a variety of biological matrices. In the work presented here, three data sets involving the analysis of glucose were employed: (1) glucose in a matrix of triacetin and bovine serum albumin (BSA) (GTB data set), (2) glucose in a human serum matrix (serum data set), and (3) glucose in a bovine blood matrix (blood data set). The GTB data set was collected at the Center for Intelligent Chemical Instrumentation at Ohio University, while the serum and blood data sets were collected at the University of Iowa.

For the data sets collected at the University of Iowa, FT-near-IR spectra were collected over the 5000–4000 cm^{-1} region of the near-IR spectrum by placing an optical interference filter in the optical path of the spectrometer. The presence of useful C–H combination bands of glucose in this spectral range has been discussed previously.¹⁰⁻¹⁵ Double-sided interferograms with 16 384 points were collected, based on 256 coadded scans. Interferograms were triangularly apodized, corrected with Mertz phase correction, and Fourier transformed. The point spacing in the transformed spectra was $\sim 1.9 \text{ cm}^{-1}$, corresponding to 519 spectral points. Part of the data collection procedure involved the collection of background spectra for use in computing absorbance spectra. Because no true background was available, the background spectra in each data set consisted of spectra of 0.1 M, pH 7.4 phosphate buffer. Background spectra were collected routinely throughout the data collection period. Each data set contained samples with glucose concentrations within the clinically relevant range of 1–20 mM. Two or three replicate spectra were collected for each sample. Additional details of the data collection for these data sets have been reported previously.¹³

Unlike the serum and blood data sets, the GTB data set featured a systematically controlled synthetic sample matrix. The sample matrix consisted of glucose, BSA, and triacetin (glyceryl triacetate) in pH 7.4, 0.1 M phosphate buffer. The BSA was employed to model proteins present in blood, while triacetin was

employed to model triglycerides. A full factorial experimental design was employed to vary the concentrations of the three components in such a way as to minimize correlations among the component concentrations. The factorial design consisted of glucose at 10 levels (1, 3, 5, 7, 9, 11, 13, 15, 17, and 19 mM), BSA at four levels (49.3, 64.4, 79.8, and 94.7 g/L), and triacetin at four levels (1.4, 2.1, 2.8, and 3.5 g/L). These concentration ranges span the levels commonly found in human clinical samples. The factorial design based on a three-component system with 10, 4, and 4 levels required $10 \times 4 \times 4 = 160$ samples.

Stock solutions of glucose, BSA, and triacetin were prepared by dilution of the pure reagent with the phosphate buffer. Reagent grade chemicals were obtained from common suppliers and used throughout the sample preparation. The BSA used was a Cohn fraction V powder obtained from Sigma Chemical Co. (St. Louis, MO; product no. A 4503). The buffer was prepared with reagent grade water obtained by passing house-distilled water through a Milli-Q Plus water purification system (Millipore Corp., Bedford, MA). The 160 samples were prepared by mixing appropriate volumes of the stock solutions and diluting to 50 mL with the phosphate buffer. Class A volumetric ware was used throughout the sample preparation.

Spectra were collected with a Digilab FTS-60A Fourier transform spectrometer (Bio-Rad, Cambridge, MA). The instrument configuration consisted of a 100 W tungsten–halogen lamp, a CaF_2 beam splitter, and InSb detector. The incident light was restricted to the 5000–4000 cm^{-1} range by use of a K-band interference filter (Barr Associates, Westford, MA). All spectra were collected while the sample was placed in an Infrasil quartz transmission cell with a 2 mm path length. Sample temperatures were controlled to the range of 37–38 °C, and samples were run in a random order. Single-sided interferograms of 16 384 points were collected, based on 256 coadded scans, and Fourier processed with triangular apodization and Mertz phase correction. The point spacing in the transformed spectra was $\sim 1.9 \text{ cm}^{-1}$. Three replicate spectra were collected for each sample, resulting in $160 \times 3 = 480$ spectra. Background spectra of the phosphate buffer were collected periodically and were used in computing absorbance spectra of the 160 samples.

Data Analysis. The spectra comprising the three data sets were transferred to a Silicon Graphics 4D/460 computer operating under the Irix operating system (version 5.2, Silicon Graphics, Inc., Mountain View, CA). All data analysis calculations were performed on this system with software written in FORTRAN-77. Fourier filtering and multiple linear regression were performed with the aid of subroutines from the IMSL software package (IMSL, Inc., Houston, TX).

The protocol developed previously for coupling digital filtering and PLS regression requires that each data set be divided into calibration, monitoring, and prediction subsets.² Each data set was divided in the following manner. First, 80% of the samples were randomly placed in the full calibration set. The remaining 20% of the samples were placed in the prediction set. Second, for GA optimizations where static calibration and monitoring sets were employed, the full calibration set was further divided into smaller calibration (80%) and monitoring (20%) subsets. For GA optimizations where random drawings of the calibration and monitoring sets were performed, the full calibration set was used as the input to the GA optimization. In all cases, replicate spectra of a given sample were kept together in the given data set. The

(3) Lucasius, C. B.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 1–33.

(4) Lucasius, C. B.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 99–145.

(5) Hibbert, D. A. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277–293.

(6) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(7) Lucasius, C. B.; Beckers, L. M.; Kateman, G. *Anal. Chim. Acta* **1994**, *286*, 135–153.

(8) Vankeerberghen, P.; Smeyers-Verbeke, J.; Leardi, R.; Karr C. L.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 73–87.

(9) Shaffer, R. E.; Small, G. W. *Chemom. Intell. Lab. Syst.* in press.

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

(11) Marquardt, L. A.; Arnold, M. A.; Small, G. W. *Anal. Chem.* **1993**, *65*, 3271–3278.

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

(13) Hazen K. H. Ph.D. Dissertation, University of Iowa, Iowa City, IA, 1995.

(14) Martens, H.; Næs. T. *Multivariate Calibration*; Wiley: New York, 1989; Chapter 3.

(15) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1–17.

Table 1. Summary of Calibration and Prediction Sets

| data set | calibration ^a | monitoring | prediction | total |
|----------|--------------------------|------------|------------|-----------|
| GTB | 96 (288) | 24 (72) | 40 (120) | 160 (480) |
| serum | 152 (456) | 38 (113) | 48 (145) | 238 (714) |
| blood | 20 (60) | 5 (15) | 10 (30) | 35 (115) |

^a The first number in each column refers to the number of samples, and the number in parentheses denotes the total number of replicate spectra.

calibration and monitoring sets were employed in the optimization of the digital filtering and PLS regression parameters, while the prediction data were withheld for use as an independent test set. Table 1 summarizes the numbers of samples and spectra in the calibration, monitoring, and prediction subsets for each data set.

RESULTS AND DISCUSSION

Overview of Data Analysis Strategy. PLS regression models have 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 spectrum i , the x_j terms are PLS factor scores derived from the spectral data matrix and concentrations, and the b terms are regression coefficients.^{14,15} The calibration model is built through multiple linear regression using the scores matrix and the matrix of known analyte concentrations as the inputs.

Previous work in our laboratories has shown that PLS regression has difficulty with complex near-IR absorbance spectra where there is significant baseline variation and noise components are present that correlate with analyte concentration. To eliminate these spectral artifacts, we have developed a protocol for using bandpass digital filtering techniques prior to PLS regression.² The digital filtering techniques employed in this work were implemented through Fourier filtering and have been discussed in detail in previous work.² The basic assumption of the filtering step is that the analyte and nonanalyte information can be separated by decomposing the data into its underlying frequency harmonics.^{16,17} Baseline artifacts will be described primarily by low-frequency harmonics, while the noise constituents will be dominated by high-frequency information. The analyte information will be concentrated in some middle range of frequencies. By tuning the position and width of the passband region of the filter, the filter can be made selective for analyte information. Employing a properly positioned filter, baseline artifacts and some noise features are removed from the spectra. The filtered spectra can then be used as input to PLS regression in place of the raw data. Previous work has shown that the use of digital filtering prior to PLS regression is an effective general data analysis approach for near-IR spectroscopy applications.^{2,11-13}

The key to the successful implementation of this approach was the development of a protocol for combining the filtering and PLS regression steps. This translates into finding the optimal values for five control variables: the position and width of the filter bandpass, the starting and ending points of the spectral range

submitted to the PLS regression procedure, and the number of PLS factors (latent variables) used in constructing the calibration model.

Optimization of Digital Filtering and PLS Regression Parameters. In previous work, the filter bandpass parameters were found through large-scale grid searches using a fixed spectral range and number of PLS factors. This approach does not, however, take into account interaction effects present among the bandpass filter characteristics, the spectral range, and the number of PLS factors. For example, the optimal bandpass filter position and width would be different if a spectral range with less baseline variation were chosen. Also, in previous work it was noted that bandpass digital filtering reduces the number of PLS factors required to explain the spectral variation. Based on this observation, a correlation might exist between the width of the filter bandpass and the required number of PLS factors. While no attempt has been made to prove statistically the existence of such correlations, empirical evidence suggests they do exist.

Thus, an optimization technique is needed to find the optimal settings for the five control variables described above. The simplest optimization approach would be a grid search involving all the variables. A grid search consists of trying all possible combinations of the given variable settings. In this application, the number of possible combinations is extremely large and would require extensive computation time. A more efficient strategy would be to employ a numerical optimization technique such as simulated annealing,¹⁸ simplex optimization,¹⁹ or GAs.³⁻⁹ Of these techniques, GAs have generated much attention recently and have been the subject of recent work in our laboratory in a pattern recognition application.⁹ GAs are simple to use, are flexible, and have been shown to work well with difficult optimization problems. Based on these characteristics and our experience, GAs were an ideal choice for this application.

Overview of Genetic Algorithms. GAs are numerical optimization methods based on the concepts of genetics and natural selection. In a GA, the variables being optimized are considered genes in a chromosome. A group of these chromosomes defines a population. Through the use of genetic operators such as natural selection, mutation, and recombination, this population of chromosomes is allowed to evolve over a number of generations. The evolution process is guided by a search for chromosomes with the greatest "fitness", i.e., chromosomes corresponding to optimal settings for the experimental variables under study. GAs belong to the general category of optimization methods that do not require derivatives of the response surface. GAs have been the topic of several review articles in the chemistry literature and are discussed in detail there.³⁻⁶ For completeness, a brief description of the four steps involved in a simple GA will be given here.

The first step in the GA involves the creation of an initial population of chromosomes. The chromosome in this application is a particular combination of variable settings (i.e., starting point in the spectral range, ending point in the spectral range, bandpass filter position, bandpass filter width, and number of PLS factors). As depicted in Figure 1, the chromosome based on these five genes is represented numerically as a vector of five integers, each specifying a value for one of the five variables under study. The

(16) Oppenheim A. V.; Schaffer, R. W. *Discrete-Time Signal Processing*; Prentice Hall: Englewood Cliffs, NJ, 1989.

(17) Horlick, G. *Anal. Chem.* **1972**, *44*, 943-947.

(18) *Adaption of Simulated Annealing to Chemical Optimization Problems*; Kalivas, J. H., Ed.; Elsevier: Amsterdam, 1995.

(19) Walters, F. H.; Parker, L. R., Jr.; Morgan, S. L.; Deming, S. N. *Sequential Simplex Optimization*; CRC Press: Boca Raton, FL, 1991.

| | | | | | |
|-------------|-------------------------------|--------------------------------|--------------------------|--------------------------|----|
| Parent 1 | 4700 cm ⁻¹ (31) | 4300 cm ⁻¹ (111) | 0.0240 <i>f</i> (241) | 0.0053 <i>f</i> (54) | 12 |
| Parent 2 | 4650 cm ⁻¹ (41) | 4495 cm ⁻¹ (72) | 0.0315 <i>f</i> (316) | 0.0122 <i>f</i> (123) | 10 |
| Child 1 | 4650 cm ⁻¹ (41) | 4495 cm ⁻¹ (72) | 0.0240 <i>f</i> (241) | 0.0053 <i>f</i> (54) | 12 |
| Child 2 | 4700 cm ⁻¹ (31) | 4300 cm ⁻¹ (111) | 0.0315 <i>f</i> (316) | 0.0122 <i>f</i> (123) | 10 |

Figure 1. Parent and child chromosomes. From left to right, the five genes in each chromosome correspond to (1) the starting wavenumber of the spectral range used in the PLS regression, (2) the ending wavenumber of that range, (3) the bandpass filter position in units of digital frequency (f), (4) the filter width (f), and (5) the number of PLS factors used in the calibration model. The wavenumber variables are mapped onto the range of 4850–4250 cm⁻¹ in 5 cm⁻¹ intervals, producing the integer sequence numbers listed in parentheses. The computer representation of the chromosome employs these integers. The filter position and width variables are similarly mapped onto the range of 0.0–0.1 f in intervals of 0.0001 f . The dark vertical line indicates a crossover point on the chromosome. In the one-point crossover method of recombination, the genes up to the crossover point are swapped between the two parents, forming the two child chromosomes. The child chromosomes are then passed to the next generation of the optimization.

population size in this work was 50 chromosomes, corresponding to 50 different combinations of variable settings. The initial population of chromosomes was created by perturbing an input chromosome through the use of a mutation operator (see below). In this work, the initial variable settings were chosen randomly.

The second step is termed the evaluation step. Each chromosome in the population must be evaluated for its fitness. Fitness in this application refers to a numerical measure of performance for the calibration model developed using the variable settings from a particular chromosome. The goal of the GA is to find the chromosome that has the highest possible fitness value. The calculation of the fitness value will be discussed in detail below.

The exploitation or natural selection step is the third step in the GA. Exploitation involves selecting the chromosomes with the largest fitness values to survive to the next generation. The chromosomes with low fitness values are removed. In this work, exploitation was performed by employing the binary tournament method.²⁰ In this method, two chromosomes are selected at random from the population. A copy of the chromosome with the larger fitness value of the two is placed in a separate group, called the mating subset, and then both chromosomes are placed back into the population. This step is performed until the size of the mating subset equals the population size. To ensure that the best chromosome is always included in the next generation, an elitist operator was employed in this work.⁶ The elitist operator places the best chromosome from the current population unchanged into the next generation.

The final step, exploration, involves the genetic operators of mutation and recombination. These operators are designed to

increase the diversity of information in the population (mutation) and to exchange information between chromosomes (recombination). As depicted in Figure 1, pairs of parent chromosomes are selected from the mating subset and recombined to form two new child chromosomes. The probability that the two parents will undergo recombination is governed by the user-specified recombination probability (P_r). The two methods of recombination used in this work, one-point and two-point crossover, will be discussed below. Mutation involves a random perturbation of the chromosome. The probability (P_m) that mutation will occur is usually kept low to ensure that good chromosomes are not destroyed. The recombination and mutation operators create a new generation of child chromosomes. These chromosomes need to be evaluated, and steps 2–4 are repeated. This process continues for a fixed number of generations. All optimization runs employed in this work were based on 50 generations.

GAs offer a generational improvement in the fitness of the chromosomes in the population. They have been described as being a combination of hill-climbing and stochastic optimization methods. The exploitation step and the elitist operator give the GA hill-climbing properties by ensuring that chromosomes with high fitness will propagate in future generations. Recombination and mutation allow the population of chromosomes to explore the response surface and move toward the globally optimal region.

Critical Issues. In this application, two issues are paramount for determining the success of GAs. Very little work has been done in the chemistry literature on applications where the variables being optimized are so different from each other and on such different numerical scales. One of the attributes of GAs that is exploited in this work is the ability to tailor the representation of the variables to suit the application. Thus, representations and GA configurations for this application must be developed. The second critical issue is the calculation of an appropriate fitness function. In previous work in our laboratories using a grid search optimization method for coupling digital filtering and PLS regression, an objective (fitness) function compatible with PLS regression was developed. This objective function introduced the concept of including prediction information in the optimization through the use of a subset of the calibration set termed the monitoring set. That work did not incorporate the spectral range or the number of PLS factors into the optimization, however. Thus, a fitness function appropriate for the optimization of the spectral range and the number of PLS factors in addition to the digital filter characteristics is necessary.

Representation of the Variables. The feature that separates GAs from other optimization methods is that they optimize on a representation or coding of the variables. The determination of which representation to choose is one of the most critical decisions the GA user makes. The conventional GA representation is to convert the settings for each variable to a binary string. The binary strings for each variable are concatenated to form one long binary string. Although the theory of this approach is well established, it does not take into account that the variables being optimized in this application are very different from each other. An alternative approach employed here is to use the natural state of the variable settings as their representation.²¹ In this application, three of the five variables are naturally integers (starting and ending points of the spectral range and the number of PLS

(20) Goldberg D. E.; Deb, K. A. In *Proceedings of the First Workshop on the Foundations of Genetic Algorithms*; Rawlings, G. J. E., Ed.; Morgan Kaufman: San Mateo, CA, 1991; pp 69–93.

(21) Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991; Chapter 4.

factors). The variable settings for the filter characteristics are on a linear digital frequency scale of 0–0.5 f but can be easily mapped onto an integer range by use of a fixed resolution for conversion of the floating point number to an integer.

In an integer representation, the chromosome is simply a vector of integers.^{21–23} In this application, there are five genes present on the chromosome. Recombination is performed through the one-point or two-point crossover methods. Two parents from the mating subset are randomly selected, along with a crossover point or location along the chromosome. In the one-point method, the genes along the chromosome up to the crossover point are swapped between the two parents, thereby forming two new child chromosomes. This process is depicted in Figure 1, where the crossover point is indicated by the dark vertical line. In the two-point crossover method, two crossover points are selected, and the genes between the crossover points are exchanged. Mutation is performed by adding a Gaussian-distributed random deviate that has been scaled to a step size to the integer value for a gene. Because the variables are on different scales, each gene was given a unique mutation operator by having a different step size for each gene. The magnitude of the step size was determined by the integer range on which that variable was mapped. A gene with a larger range (i.e., more possible values) requires a larger step size.

Fitness Function Calculation. The choice of the fitness (objective) function is critical to the success of GAs. The goal of the fitness function is to encode the performance of the chromosome numerically. Stated differently, it is a measure of how well a particular combination of variable settings perform. In this application, performance should reflect the ability of the calibration model computed by use of the settings specified by the chromosome to predict analyte concentrations from data that were not used to develop the model (i.e., prediction). If the fitness calculation does not measure chromosome performance accurately, then the advantages gained by the exploitation and exploration steps are lost. The fitness function calculation should be sensitive enough to allow the GA to find the optimal variable settings. Many researchers have developed objective functions for feature selection that are based on computing the orthogonality of the input spectral data matrix.^{7,24} These functions are specific for multiple linear regression and are not relevant for application to PLS regression. The most appropriate objective function for PLS regression is to compute a calibration model at each evaluation and use the statistics from the regression to produce a measure of how well the model performed. Thus, at each fitness evaluation, the variable settings associated with the chromosome being evaluated are employed to build a calibration model. First, the absorbance spectra are filtered using a bandpass digital filter with the position and width specified by the corresponding genes on the chromosome. Using the filtered spectra, PLS regression is used to build a calibration model based on the spectral range and number of PLS factors specified by the corresponding genes on the chromosome.

Previous work in our laboratories established the use of the mean-squared error in calibration (MSE) and the mean-squared error in monitoring (MSME) as measures of model fitness.²

MSME was used in an attempt to include prediction performance into the measure of calibration model performance. To accomplish this, the calibration set was randomly subdivided into calibration and monitoring subsets. The filtered spectra in the calibration subset were used to generate the calibration model using PLS regression. The resulting model was used to predict the analyte concentrations corresponding to the filtered spectra in the monitoring subset. These calibration and prediction results were employed in an objective function to measure calibration model performance. The function used was

$$1 / \left[\sum_{i=1}^{n_c} (c_i - \hat{c}_i)^2 / (n_c - h - 1) + \sum_{i=1}^{n_m} (m_i - \hat{m}_i)^2 / n_m \right] \quad (2)$$

where n_c is the number of spectra in the calibration set, n_m is the number of spectra in the monitoring set, h is the number of PLS factors used in the calibration model, c_i is the measured glucose concentration of spectrum i in the calibration set, \hat{c}_i is the glucose concentration predicted by the model, m_i is the measured glucose concentration of spectrum i in the monitoring set, and \hat{m}_i is the glucose concentration predicted by the model. The denominator in this equation represents a summation of MSE and MSME. MSE and MSME of 0.0 represent a perfect fit of the calibration data to the model and a perfect fit between actual concentrations and the concentrations predicted by the model for the monitoring set data. In the work that resulted in eq 2, objective functions employing MSE as the lone criterion were also studied.² The results showed that the objective function shown in eq 2 was more robust. Optimizations employing an objective function based on MSE led to calibration models that performed very well in calibration but did not perform well in prediction.

In previous work, the objective function shown in eq 2 was used in conjunction with a grid search to find the optimal bandpass filter characteristics.² In that work, no attempt was made to include the spectral range or the number of PLS factors into the optimization. The number of PLS factors and the spectral range were kept constant throughout the optimization. To gain maximum benefit from coupling digital filtering and PLS regression, these variables must be included in the optimization. When eq 2 was employed as the fitness function in our initial GA studies, the GA converged to calibration models that employed more terms (i.e., a large number of PLS factors) than were necessary. These additional terms provided a small increase in the fitness score. Although the calibration error was reduced by including these additional terms, the errors in the monitoring set were nearly the same. Using a fitness function based on eq 2, the calibration models with the additional terms would be judged superior. The information contained in the extra PLS factors may not be useful for predicting analyte concentrations from new data, however. Calibration models with unnecessary terms do not follow the concept of parsimonious data modeling defined by Seasholtz and Kowalski.²⁵ Their work concluded that, of two models that perform equally well in calibration, the one with the smaller number of terms will tend to have better predictive abilities given new data. Because the eventual goal of the optimization is to find calibration models that can predict analyte concentrations given new data, the concepts of parsimonious data modeling need to

(22) Bramlette, M. F. *Proceedings of the Fourth International Conference on Genetic Algorithms*; Morgan Kaufman: San Mateo, CA, 1991; pp 100–107.

(23) Ichikawa, Y.; Ishisi, Y. *1993 IEEE International Conference on Neural Networks, Volume 2*; IEEE Press: San Francisco, CA, 1993; pp 1104–1109.

(24) Kalivas, J. H.; Roberts N.; Sutter, J. M. *Anal. Chem.* **1989**, *61*, 2024–2030.

(25) Seasholtz, M. B.; Kowalski, B. *Anal. Chim. Acta* **1993**, *277*, 165–177.

be included in the calculation of the fitness function. In this work, three fitness functions were studied that included a term in the fitness calculation for minimizing the number of PLS factors used in the calibration model. The functions studied were

$$(\text{MSE} + \text{MSME} + h^w)^{-1} \quad (3)$$

$$(\text{MSE} + h^w)^{-1} \quad (4)$$

$$(\text{MSME} + h^w)^{-1} \quad (5)$$

where MSE and MSME are defined above, h is the number of PLS factors, and w is a weighting factor. The weighting factor is critical because it places h on the same scale as the error terms. Because it dictates the influence the number of PLS factors is given in the optimization, the choice of a proper w is critical for this application. A protocol for selecting the optimal weighting factor will be discussed later. For purpose of simplicity, the fitness functions shown in eqs 3–5 will be termed fitness functions 1, 2, and 3, respectively.

Although previous work has shown that an objective function employing just MSE was not as robust as an objective function that included both MSE and MSME, it was deemed necessary to study this issue again.² In the previous study, it was hypothesized that an objective function based on MSE led to optimized bandpass filters that filtered the data such that random data values were introduced that correlated with analyte concentration. These random values were encoded into the PLS factors and led to the poor predictive ability of those models. It was concluded that these calibration models were not robust. However, by introducing a term to reduce the number of PLS factors into the fitness function, it was hypothesized that the chance of random data values being introduced by filtering would be reduced.

Fitness function 3 was not studied in previous work because it was believed that an objective function based on MSME alone would be very susceptible to spectral artifacts that were present in the monitoring set. Because the monitoring set in the current application contains only 20% of the total calibration data, decisions regarding the best bandpass filter characteristics would necessarily be based on only a few samples. In an attempt to overcome this limitation and thereby increase the viability of fitness function 3, several mechanisms were studied for varying the selection of the calibration and monitoring sets.

Selection of Calibration and Monitoring Sets. The purpose of the digital filtering and PLS calculations is to compute a set of “true” spectral components that allow the analyte absorbance to be extracted. The digital filtering step removes baseline and noise artifacts, while the PLS computation decomposes the filtered spectra into the required set of underlying components. The success of the filtering and PLS calculations is based on (1) the degree to which the set of spectra that comprise the calibration and monitoring sets is reflective of the “global” set of spectra likely to be encountered and (2) the degree to which optimal values are found for the five variables being studied in this work. Unfortunately, sufficient data are almost never available to allow truly global calibration and monitoring sets to be constructed. Thus, a concern exists as to whether artifacts associated with a particular set of calibration and monitoring spectra will have an undue effect on the computed model. The presence of artifacts in the monitoring set spectra is particularly significant, given the

facts that this set contains only a fraction (e.g., 20%) of the total calibration data and that MSME is given significant weight in the calculation of fitness functions 1 and 3.

Furthermore, if spectra containing artifacts are used in the optimization of the five control variables being studied here, a further skewing of the calibration model may result. In the present work, if the calibration and monitoring spectra contain artifacts not present in the rest of the data, the GA will select chromosomes to account for these spectral artifacts. These chromosomes may not be globally representative of the data set and may result in models that do not perform well for data not used in the optimization. In this case, a lack of universality in the calibration and monitoring sets prevents the “true” spectral components from being uncovered, thereby resulting in a flawed calibration model.

In previous work using the objective function shown in eq 1, the calibration and monitoring sets were chosen once at the beginning of the analysis before any optimization was performed. The calibration and monitoring sets were kept constant throughout the data analysis. As described above, the disadvantage of these “static” calibration and monitoring sets is that the chance exists that relatively few spectra can have an unduly large influence on the optimization. A solution to this problem is to redraw the calibration and monitoring sets randomly at each fitness evaluation in the optimization. This procedure reduces the impact of spectra with artifacts because, during a single generation of the optimization, each sample in the full calibration set has the opportunity to appear in both the calibration and monitoring subsets. Thus, the contribution of any individual sample to the optimization is lessened, and the chance for the optimization variables to be skewed is decreased.

A potential drawback of this use of dynamic calibration and monitoring sets is that sampling error is introduced into the computed fitness scores. The fitness scores for identical chromosomes will change as different calibration and monitoring sets are selected. Thus, a chromosome that is exceptionally fit in one generation may not produce a similar fitness score in the next generation due to the differences in the calibration and monitoring sets. Also, since the number of PLS factors is one of the optimization variables and is therefore fixed for a given function evaluation, the change in the calibration set may cause the PLS calculation to decompose the spectra differently. Thus, p PLS factors may be optimal for one calibration set, while q factors may be optimal for another set. It can be argued that these sources of variability may slow the optimization, possibly preventing the optimal values for the variables to be found. Thus, a tradeoff is apparent between allowing individual spectra to have too much weight in the calculations versus introducing variability in the fitness evaluation.

One solution to this problem is to draw several calibration and monitoring sets at each fitness evaluation and compute the mean fitness score for the group. This procedure has the twin advantages of averaging out the sources of variability noted above while being even more resistant to spectral artifacts than a single random drawing. The disadvantage to this approach is that it increases the number of calculations that need to be performed. For example, if three random drawings are performed, three calibration models must be computed. This would be more computationally intensive than the use of either static calibration and monitoring sets or a single random drawing.

Table 2. Settings in the Experimental Design Study of the Genetic Algorithm Configurations

| variable | level 1 | level 2 |
|----------------------|---------------------|---------------------|
| P_m | 0.1 | 0.2 |
| population size | 50 | 100 |
| P_r | 0.8 | 0.9 |
| recombination method | one-point crossover | two-point crossover |
| step size | 0.1 | 0.2 |

Table 3. Optimal Genetic Algorithm Configurations

| variable | GTB | serum | blood |
|----------------------|---------------------|---------------------|---------------------|
| P_m | 0.1 | 0.2 | 0.2 |
| population size | 50 | 50 | 50 |
| P_r | 0.9 | 0.9 | 0.9 |
| recombination method | one-point crossover | two-point crossover | one-point crossover |
| step size | 0.1 | 0.2 | 0.2 |

GA Configuration. The configuration of the GA is critical for a successful optimization. Although the representation of the variables is the most critical factor in configuring a GA, there are several other parameters that dictate the performance of the optimization. The GA configuration variables that are important in this GA implementation are the mutation rate (P_m), recombination rate (P_r), population size, method of recombination, and the step size for mutation and initialization. To determine the optimal configuration variable settings, a separate experimental design study was performed for each data set. In this work, a two-level fractional factorial experimental design was employed.²⁶ In this design, two variable settings (levels) were chosen in advance for each variable. A two-level fractional factorial design for the five variables consists of $2^{5-1} = 16$ combinations of variable settings. The variable settings used in the experimental design study were based on literature references and our experience in using GAs for other optimization problems. The variable settings used are shown in Table 2. For each experiment, a GA optimization of the filter position and width, spectral range, and the number of PLS factors was performed using the fitness function shown in eq 3 with a static calibration and monitoring set. The weighting factors were 2, 0.33, and 0.33 for the serum, GTB, and blood data sets, respectively. The choice of the weighting factor during the experimental design does affect the decisions made concerning the optimal GA configuration. The products of the optimization were a list of the best chromosomes containing the optimized variables and their associated fitness scores. The response employed in the experimental design was the fitness score for the best chromosome. Replication was performed by starting the GA optimization in three different locations. The same three starting points were employed for all experiments. The mean response for each variable setting was computed. The configuration variable settings with the largest mean response were judged the optimal settings. The optimal GA configuration for each data set is shown in Table 3.

GA Optimization Protocol. To compare fitness function models and methods of selecting the calibration and monitoring sets, protocols for choosing the optimal weighting factor (w) for use in computing the fitness function and an unbiased approach

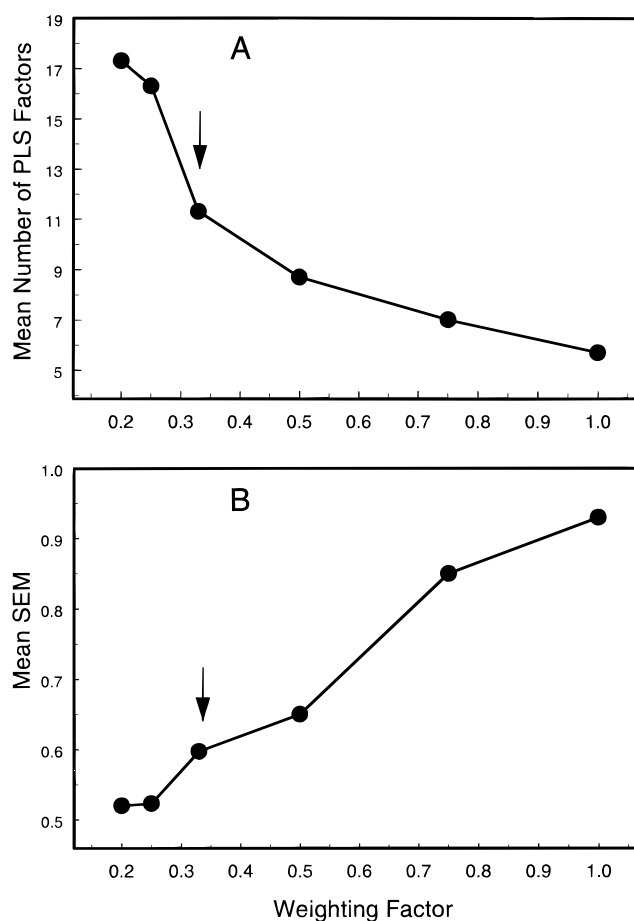


Figure 2. Graphical representation of the protocol for choosing the optimal weighting factor used in the fitness calculation for the GTB data set and fitness function 1. (A) Plot of the mean number of PLS factors versus weighting factor. (B) Plot of mean SEM versus weighting factor. The optimal weighting factor strikes a balance between a low mean SEM and a low mean number of PLS factors. As illustrated by the arrow, the optimal weighting factor in this example is 0.33.

for selecting the best calibration models from the GA optimization were needed.

A simple protocol was developed for choosing the best w . The weighting factor places the number of PLS factors on the same scale as the error terms. The protocol allows the user to find a balance between giving too much weight to h , thereby causing the resulting models to have too few factors to explain the concentration variance, and not giving enough weight to h , thereby causing the resulting models to no longer be parsimonious. The optimal w allows the GA to find chromosomes that lead to calibration models that feature good predictive abilities as well as having a minimum number of PLS factors. Figure 2 illustrates the protocol through an example from the GTB data set and fitness function 1 using a static calibration and monitoring set. Figure 2A shows the mean h for the five best calibration models selected by the GA versus weighting factor, while Figure 2B similarly plots the mean standard error in concentration for the monitoring set spectra (SEM) versus w . The five best calibration models were selected on the basis of fitness score. In this example, a weighting factor of 0.33 offers the best compromise of low SEM and low h . SEM was employed instead of the standard error in concentration for the prediction set spectra (SEP) in the selection of w because it was felt that the decision regarding the best fitness function

(26) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978; Chapter 12.

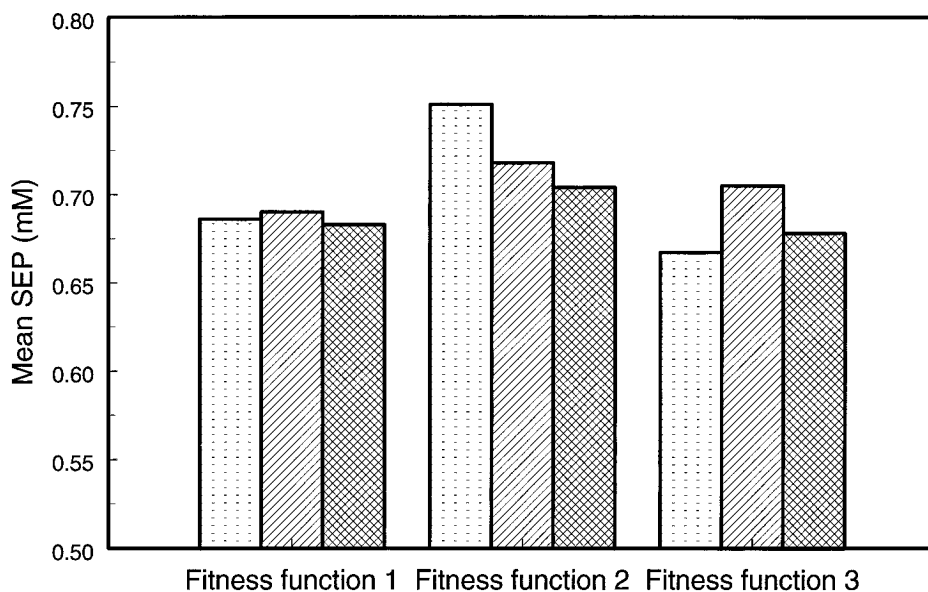


Figure 3. Bar plot indicating the mean SEP for each fitness function and method of selecting the calibration and monitoring sets for the GTB data set. From left to right within each group, the bars correspond to static, one random drawing, and three random drawings, respectively, of the calibration and monitoring sets.

and method for selecting the calibration and monitoring sets may be biased if the choice of w were based on prediction results. This protocol was employed to select the optimal weighting factor for each data set, fitness function calculation, and method of selecting the calibration and monitoring sets. The optimal w ranged between 0.2 and 0.33, 0.167 and 0.33, and 1.25 and 2.0 for the GTB, blood, and serum data sets, respectively. A significantly different w was needed for the serum data set because the glucose concentration values input to the optimization for this data set were in the standard clinical units of mg/dL, while for the GTB and blood data sets the glucose concentrations were in units of mM.

To compare the fitness function calculations and the methods for selecting the calibration and monitoring sets, an optimization protocol was developed to select the best calibration models on the basis of fitness score. Each optimization was replicated three times using different starting points for the optimization. The starting points were chosen once and used in all subsequent optimizations. Three different starting points were selected for each data set. From previous research, it was found that the optimal filter position and width typically lie in the range of 0.0–0.1 f .^{2,11–13} Thus, for this study, the filter position and width settings in the chromosome were restricted to this range. For conversion to an integer, the filter position and width were mapped onto an integer range. This mapping is depicted in the example chromosomes in Figure 1. The resolution used in the mapping was 0.0001 f . Thus, 1001 possible settings for both filter position and width were possible. The spectral range was restricted to the 4850–4250 cm^{-1} region. All previous studies in our laboratories involving the FT-near-IR analysis of glucose have used this region. The starting and ending points in the spectral range were mapped onto an integer range using intervals of 5 cm^{-1} . This mapping procedure is also shown in Figure 1. There were 121 possible values for each spectral range variable. For cases in which the starting point in the spectral range came after the ending point, the chromosome was assigned a fitness score of zero. The number of PLS factors was restricted to a maximum of 20. Any variable setting outside the acceptable range was

replaced with the closest possible allowable setting. The mapping procedure resulted in an optimization problem with $1001 \times 1001 \times 121 \times 121 \times 20 = 2.93 \times 10^{11}$ possible combinations of variable settings.

The outputs of the GA optimization were the variable settings associated with the 10 chromosomes with the highest fitness scores. Replication was performed three times, and the overall five chromosomes with the highest fitness scores among the replicates were selected. Five calibration models were computed using the filter positions and widths, spectral ranges, and numbers of PLS factors specified by these chromosomes. These calibration models were computed with the full calibration set (i.e., calibration set data + monitoring set data). The five calibration models were then used to predict the glucose concentrations for the spectra in the prediction set. Values for SEP were computed on the basis of each of the five sets of prediction results, and the mean of these SEP values was employed as the criterion for judging GA optimization performance. The ability of the GA to find calibration models with good predictive performance (low SEP) is directly related to how suitable the fitness function and the method of selecting the calibration and monitoring sets are for this application.

GA Optimization Results. Figure 3 depicts the optimization results for the GTB data set. This figure is a bar plot showing the mean SEP for the top five calibration models for each fitness function and method of selecting the calibration and monitoring sets. From left to right within each group, the bars correspond to static, one random drawing, and three random drawings of the calibration and monitoring sets, respectively. The results shown in this figure indicate that each of the three fitness functions and three methods of selecting the calibration and monitoring sets can lead the GA to good models. The difference in mean SEPs between the best and the worst combination of fitness function and method of selecting the calibration and monitoring sets was <0.1 mM.

Figure 4 depicts the optimization results for the serum data set. To remain consistent with the other figures, the units of the glucose concentration have been converted from mg/dL to mM.

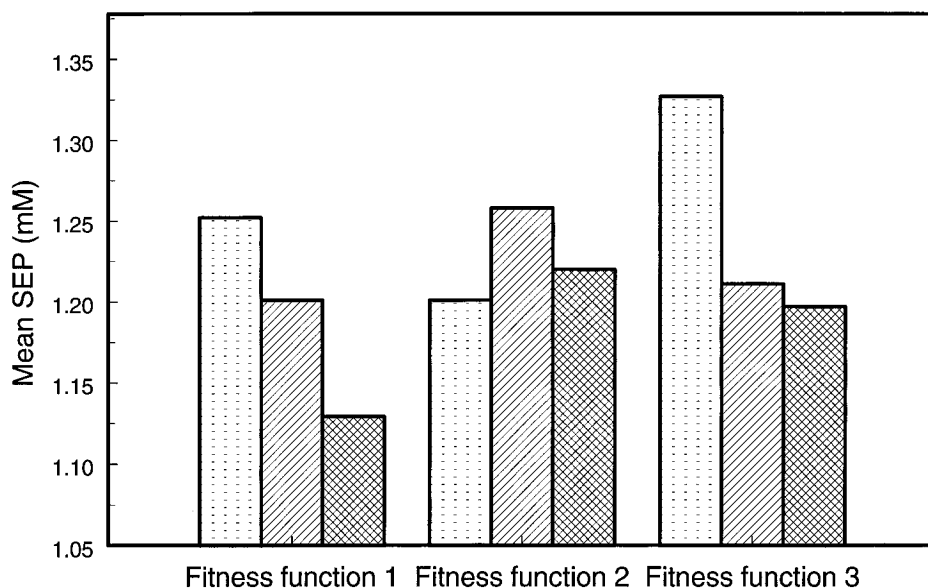


Figure 4. Bar plot indicating the mean SEP for each fitness function and method of selecting the calibration and monitoring sets for the serum data set. From left to right within each group, the bars correspond to static, one random drawing, and three random drawings, respectively, of the calibration and monitoring sets.

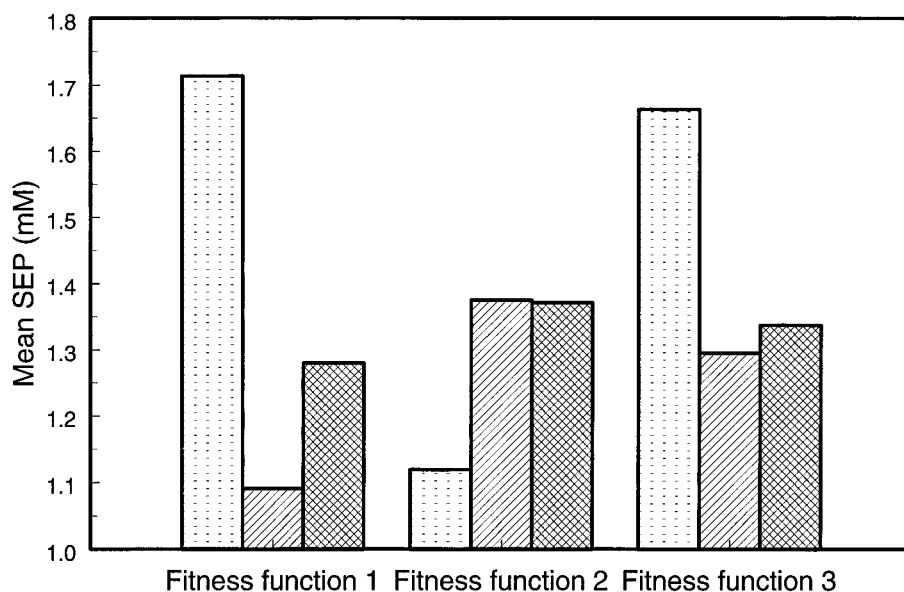


Figure 5. Bar plot indicating the mean SEP for each fitness function and method of selecting the calibration and monitoring sets for the blood data set. From left to right within each group, the bars correspond to static, one random drawing, and three random drawings, respectively, of the calibration and monitoring sets.

The arrangement of the bars is identical to that in Figure 3. The results shown in this figure point to a different conclusion regarding the method of selecting the calibration and monitoring sets than the results from the GTB data set. For fitness functions 1 and 3, random drawings of the calibration and monitoring sets improve the GA search performance. This trend was not observed for fitness function 2, however. The observation that fitness function 2 does not realize the benefits of random drawings was expected. Fitness function 2 employs only the calibration results and the number of PLS factors. Because the calibration set is composed of four times the number of samples as the monitoring set, the stability of the calibration results is much greater than the corresponding results based on the monitoring set. These results do not establish one fitness function as being statistically better, however. Again, the differences in mean SEP for the best and worst optimizations are small (~ 0.2 mM). The calibration

models for the serum data set do not calibrate and predict as well as the calibration models for the GTB data set due to the presence of a more challenging sample matrix.

Figure 5 is a plot analogous to Figures 3 and 4, and depicts the GA optimization results for the blood data set. The results shown in this figure point to a conclusion similar to that in Figure 4. For fitness functions 1 and 3, random drawings of the calibration and monitoring sets allow the GA to find better calibration models. This trend was not observed for fitness function 2. The difference in optimization performance was significant for fitness functions 1 and 3 due to the small size of the blood data set. As seen in Table 1, the monitoring set for the blood data set consists of only five samples. These results demonstrate the need for random drawings of the calibration and monitoring sets for data sets composed of small numbers of samples. The difference in mean SEPs for the top five calibration

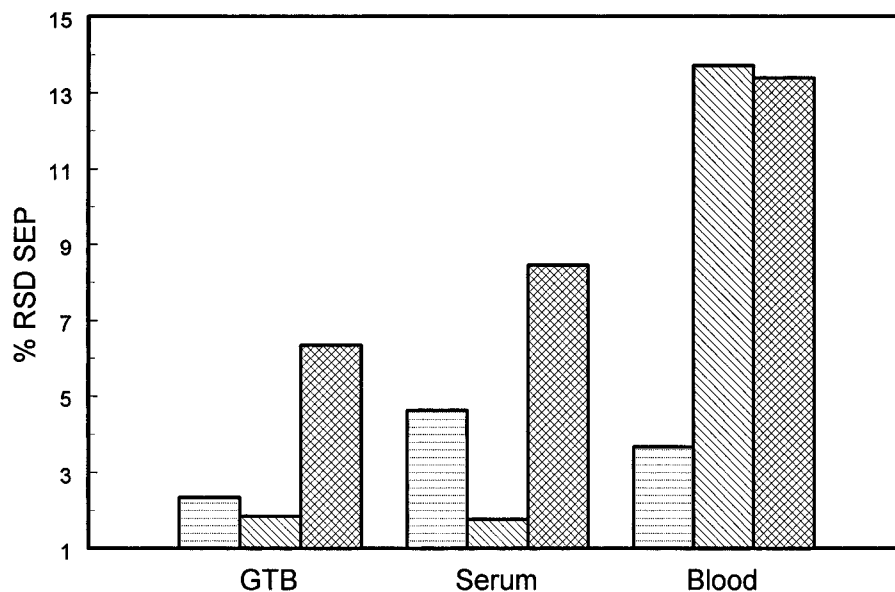


Figure 6. Bar plot indicating the %RSD computed from the top five SEP values for each fitness function and data set. From left to right within each group, the bars correspond to fitness functions 1, 2, and 3, respectively.

models found using fitness function 1 and a static monitoring set and the top five models found using fitness function 1 and one random drawing of the calibration and monitoring sets was considerably larger (~ 0.6 mM) than similar differences observed with the other data sets.

An inspection of Figure 5 also indicates that, for fitness functions 1 and 3, one random drawing produces a model with a lower SEP than the corresponding model based on three random drawings. This is counterintuitive, as the use of additional random drawings would be expected to reduce further the chance of individual spectra having too large of an influence on the calibration model. To investigate this result, additional runs were made with five random drawings. For both fitness functions, the SEP values produced by the models based on five random drawings decreased relative to the values derived from three random drawings. For fitness function 1, the resulting SEP value was virtually indistinguishable from that produced by one drawing. For fitness function 3, the lowest SEP of all was produced by the model based on five random drawings. These results suggest that there is sampling variation associated with the small number of samples in the blood data set and that, for this case, it may take more than three random drawings for the SEP to stabilize.

The results from the blood data set again produced no statistically valid choice regarding the optimal fitness function. The calibration models for this data set predict worse than the calibration models for the serum and GTB data sets. This is due to the combination of a challenging background matrix and the small number of samples chosen. When the full data set was employed, the prediction results improved dramatically.¹³

For two of the three data sets, the fitness function calculation with the lowest mean SEP for the top five calibration models was fitness function 1 using random drawings of the calibration and monitoring sets. These results, however, are not statistically significant. Because there was not enough evidence to determine the best fitness function, additional measures were employed. Instead of using the mean SEP as the performance criterion, the standard deviation of the SEP values for the top five calibration models was computed. The hypothesis was that the best models found by the GA should all achieve good prediction results. If

the GA were able to find good calibration models consistently, the standard deviation of the SEPs would be small. Similarly, if the standard deviation were large, then the GA was producing inconsistent results. If the GA performed consistently, fewer replications of the optimization would need to be performed to find the optimal regions of the response surface. Because the GA search is guided by the fitness function, the standard deviation of the SEP values is a valid tool for comparing the optimization performance of GAs employing different fitness functions. Figure 6 shows a bar plot depicting the percent relative standard deviation (%RSD) of the SEP values for the top five calibration models for each data set and fitness function. The results displayed in this figure employed three random drawings of the calibration and monitoring sets. Plots employing the other methods of selecting the calibration and monitoring sets look similar and lead to the same conclusion. From left to right in Figure 6, the bars within each group correspond to fitness functions 1, 2, and 3, respectively. The results clearly show that the stability of the best five calibration models is data set dependent. For the two large data sets, the choice of fitness function is not critical. Any of the three fitness functions will lead the GA to consistently select calibration models with good predictive performance. However, for smaller data sets, it is apparent that fitness function 1 outperforms the other fitness functions significantly. Thus, fitness function 1 was employed for all subsequent optimizations.

The results shown in Figures 3–6 demonstrate that the stability gained by using two measures from the regression model (MSE and MSME) computed during the fitness evaluation is important. The conclusion that a fitness function calculation employing both MSE and MSME was more robust and clearly defined the optimal parameter settings has been noted previously.² Fitness functions using both measures are less susceptible to random data values being correlated to analyte concentration. Multiple random drawings of the calibration and monitoring sets clearly benefitted GA optimizations involving fitness function 3. However, fitness function 3 still lacked the stability of fitness function 1 for smaller data sets. As expected, random drawings of the calibration sets did not improve the results of the GA optimization involving fitness function 2.

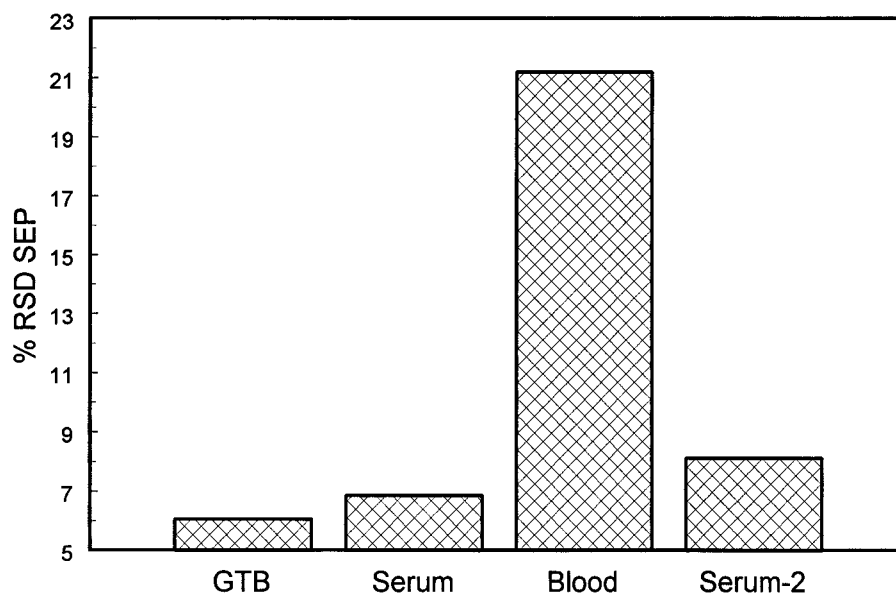


Figure 7. Bar plot indicating for each data set the %RSD computed from the top SEP values corresponding to the GA optimizations based on the use of each of the five static calibration/monitoring subsets.

Limitations of a Static Calibration and Monitoring Set.

Figures 3–5 discussed above illustrate that data sets with fewer samples require a fitness function that employs random drawings of the calibration and monitoring sets. There was a noticeable difference in the results depending on the size of the data set involved. The limitation of the static method was caused by having a monitoring set composed of only 20% of the total calibration data. Because fitness function 1 gives equal weight to the calibration and monitoring errors, any spectral artifacts present in the monitoring set will have a large impact on the computed fitness score.

To lend further support to the conclusions regarding the static monitoring sets, a fourth data set was created by use of a subset of the serum data set. This smaller data set was created by randomly selecting 40 samples (119 spectra) from the 238 samples (714 spectra) in the serum data set. From this pool of data, the samples were split into a full calibration set and a prediction set. The full calibration set contained 32 samples (96 spectra), and the prediction set contained eight samples (23 spectra). The full calibration set was further subdivided into a smaller calibration subset comprised of 26 samples (78 spectra) and a monitoring subset comprised of six samples (18 spectra). This smaller data set will be termed serum-2.

Using the four data sets, new experiments were performed to demonstrate the limitations of the static calibration and monitoring sets for smaller data sets. For each data set, the full calibration set was used to select four additional random calibration and monitoring subsets. Using these new static calibration and monitoring sets, the GA optimizations were repeated. These new GA optimizations employed fitness function 1 and the same w , GA configuration, and starting points as the original optimizations described by Figures 3–5. The only difference in the new optimizations was the distribution of which samples were in the calibration and monitoring sets. The hypothesis in this experiment was that, for the larger data sets (GTB and serum), each optimization would produce calibration models that predict equally well, while for the smaller data sets (blood and serum-2), the best calibration models would offer a wide range of prediction results. Stated differently, changing the mix of the samples in the

calibration and monitoring sets would be expected to have a larger impact on the results for the two smaller data sets than for the two larger data sets.

The GA optimization protocol was employed to select the top five calibration models for each of the five static calibration and monitoring sets. These five sets of five models were then applied to the prediction set, and the lowest SEP based on each calibration/monitoring subset was determined. The %RSD of these five SEP values was then computed for each data set. Figure 7 shows a bar plot depicting the results from this experiment. From left to right, the bars on the x -axis correspond to the GTB, serum, blood, and serum-2 data sets, respectively. The results shown in this figure confirm the hypothesis that the use of a static calibration and monitoring set has a larger impact on search performance for smaller data sets than for larger ones. The results were more dramatic than expected for the blood data set. The SEPs for these optimizations ranged from a low of 0.962 mM to a high of 1.713 mM. This result is very interesting. If we had chosen to use the calibration and monitoring set that led the GA to select the calibration model that achieved a prediction performance of 0.962 mM, the conclusions from the initial study would have been different. This result illustrates the potential danger of using an optimization technique when the data employed in the optimization are not globally representative. While the results from the serum-2 data set are not as unstable as those from the blood data set, the %RSD is still 18.6% greater than that from the serum data set and 34.0% greater than that from the GTB data set.

The experiments described above indicate clearly that random drawings of the calibration and monitoring sets should be performed. The next issue to address is how many drawings need to be made. As noted previously, the results displayed in Figures 3–5 show enough sampling variability to prevent clear conclusions on this matter from being drawn. To study this issue in more detail, an additional experiment was performed. An average-performing chromosome was selected from previous optimizations for each data set. The fitness score of this chromosome was computed 500 different times using randomly drawn calibration and monitoring sets. The hypothesis for this experiment was that

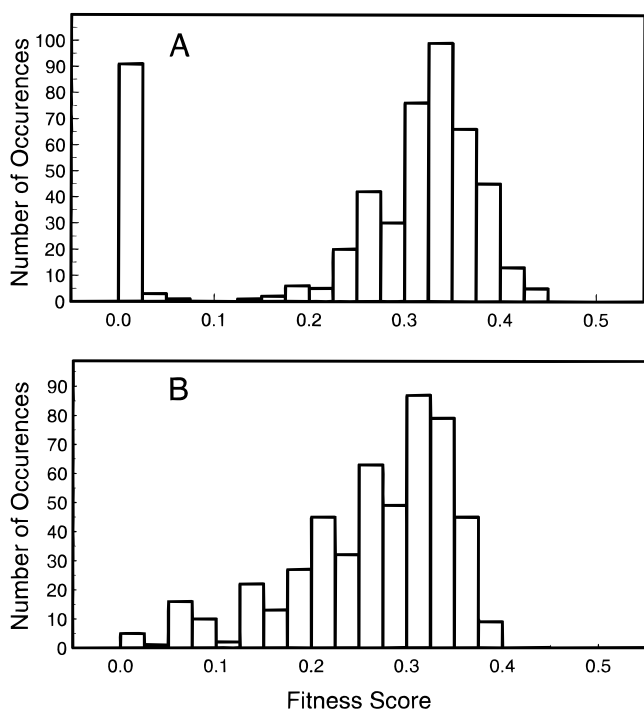


Figure 8. (A) Histogram showing the distribution of the fitness scores for a single random drawing of the calibration and monitoring sets for the blood data set. (B) Histogram showing the distribution of the mean fitness scores for three random drawings of the calibration and monitoring sets for the blood data set.

Table 4. Fitness Score Statistics

| data set | high | low | mean | SD | %RSD |
|----------|-----------|-----------|-----------|------------|-------|
| GTB | 0.373 | 0.270 | 0.343 | 0.016 2 | 4.72 |
| serum | 0.000 698 | 0.000 330 | 0.000 537 | 0.000 0876 | 16.31 |
| blood | 0.435 | 0.002 14 | 0.265 | 0.133 | 50.28 |
| serum-2 | 0.000 693 | 0.000 144 | 0.000 354 | 0.000 111 | 29.14 |

a wider range of fitness scores would be obtained for the two smaller data sets than for the two larger data sets. For an ideal data set, the fitness score would be similar no matter which samples were placed in the calibration and monitoring sets. Table 4 shows the results from this experiment. Using the %RSD fitness score as the criterion, the hypothesis was confirmed. The blood and serum-2 data sets show significantly more variation in fitness score than the GTB or serum data sets.

To study this issue further, at each fitness evaluation, three random drawings were performed, and the mean fitness was reported. For this case, the ranges of fitness scores observed in Table 4 should decrease. Using three random drawings instead of a single drawing, the %RSD fitness score decreased from 50.28% to 31.06% and from 29.14% to 20.52% for the blood and serum-2 data sets, respectively. This point is illustrated graphically in Figure 8. Parts A and B are histograms of the fitness scores for the blood data set using a single random drawing of the calibration and monitoring sets and three random drawings, respectively. The lower histogram has a much tighter distribution of fitness scores. This experiment yields enough evidence to conclude that multiple random drawings are better than a single drawing.

The decision concerning the correct number of drawings should be based on two considerations. Data sets with large numbers of samples or in which the spectral variation within the samples is small may need fewer drawings than small data sets

or those in which the spectral variation is large. Second, computation time should be considered. Each random drawing increases the computation time required for the optimization. Based on the Central Limit Theorem of statistics, the benefit to random drawings increases as the square root of the number of random drawings.^{26,27} Thus, a point of diminishing returns is reached. For most data sets, three random drawings should sufficiently balance these concerns. For data sets with characteristics similar to blood and serum-2, one or two additional random drawings may be beneficial.

Comparisons of GA Procedure and Grid Searches. Previous work in our laboratories has established baseline calibration and prediction results for comparison purposes using the GTB and serum data sets.^{13,28} These results were obtained by use of grid searches to find the optimal filter parameters for a small number of spectral ranges and model sizes. Using a GA optimization procedure that consisted of fitness function 1, three random drawings of the calibration and monitoring sets, weighting factors of 2 and 0.33 for serum and GTB, respectively, the optimal variable settings for the filter position and width, spectral range, and number of PLS factors for the serum and GTB data sets were found. Table 5 lists the optimized filter characteristics, spectral ranges, and number of PLS factors and their corresponding calibration and prediction results for the previous work performed by use of a series of grid searches and the GA approach. As demonstrated by the results shown in the table, the GA optimization approach performed very well. For the GTB data set, the GA was able to find calibration models that required fewer PLS factors than the grid search optimization method and resulted in approximately 12% and 28% reductions in the values of SEC and SEP, respectively. For the serum data set, the GA and grid search optimization procedures found calibration models that performed nearly the same.

CONCLUSIONS

The results presented in this paper demonstrate that GAs are a viable approach for the joint optimization of bandpass filter position and width, spectral range, and the number of PLS factors. This approach allows these variables to be optimized together, thereby providing the most efficient coupling of digital filtering and PLS regression. To implement the GA technique successfully, however, the user must study the configuration of the GA in seeking optimal values for several key parameters that govern the optimization. A second critical issue is the choice of fitness function used to monitor the progress of the GA. The fitness function found to be most useful here appears to be generally applicable across a range of data sets, although the weighting term used to control the calibration model size is data set dependent.

Finally, the question arises as to whether the results obtained through the use of the GA are sufficiently better than those obtained through a conventional grid search to warrant the effort invested in implementing and configuring the GA. For the two data sets analyzed here through the use of both the grid search and GA procedures, the GA produced a significant improvement in one case and essentially the same results in the other case. In our opinion, however, the key advantage of the GA technique lies not so much in its potential for lowering prediction errors or reducing the size of the calibration model but rather in its ability

(27) Güell, O. A.; Holcombe, J. A. *Anal. Chem.* **1990**, *62*, 529A–542A.

(28) Cingo, N. A.; Small, G. W., unpublished work.

Table 5. Comparison of Genetic Algorithm and Grid Search Methods

| data set | filter position ^a | filter width ^a | start spectral range ^b | end spectral range ^b | no. of PLS factors | SEC ^c | SEP ^c |
|-------------------|------------------------------|---------------------------|-----------------------------------|---------------------------------|--------------------|------------------|------------------|
| Grid Search | | | | | | | |
| GTB | 0.0021 | 0.0491 | 4459 | 4309 | 11 | 0.684 | 0.908 |
| serum | 0.039 | 0.014 | 4850 | 4250 | 13 | 1.51 | 1.20 |
| Genetic Algorithm | | | | | | | |
| GTB | 0.0733 | 0.0251 | 4600 | 4295 | 9 | 0.602 | 0.656 |
| Serum | 0.0476 | 0.0286 | 4795 | 4300 | 14 | 1.49 | 1.18 |

^a In units of digital frequency (f). ^b in units of cm^{-1} . ^c In units of mM.

to implement the calibration model optimization in a largely unbiased manner. In any grid search, the necessity to limit the number of computations dictates that the user must choose precise limits for the search space of the variables. In addition, decisions must be made regarding which variables to study together versus which variables to optimize independently. Together, these requirements dictate that a small search space is explored, thus making the grid search a highly biased procedure. By contrast, no assumptions regarding the independence of the variables are made in the GA approach, and the search space is so large that the procedure can be considered unbiased. In our opinion, this advantage alone justifies the use of the GA procedure.

ACKNOWLEDGMENT

This research was supported by the National Institutes of Health under Grant 1-R01-DK45126-01A1. Mutua Mattu, Ndumiso Cingo, and Qing Ding are thanked for collecting the spectra in the GTB data set. The Department of the Army is acknowledged for providing the Silicon Graphics 4D/460 computer system.

Received for review January 17, 1996. Accepted May 10, 1996.[⊗]

AC960049G

[⊗] Abstract published in *Advance ACS Abstracts*, June 15, 1996.