

Articles

Genetic Algorithm-Based Wavelength Selection for the Near-Infrared Determination of Glucose in Biological Matrixes: Initialization Strategies and Effects of Spectral Resolution

Qing Ding and Gary W. Small*

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

Mark A. Arnold

Department of Chemistry, Iowa Advanced Technology Laboratories, University of Iowa, Iowa City, Iowa 52242

An improved genetic algorithm (GA)-based wavelength selection procedure is developed to optimize both the near-infrared wavelengths used and the number of latent variables employed in building partial least-squares (PLS) calibration models. This GA-based wavelength selection algorithm is applied to the determination of glucose in two different biological matrixes. With random selection of a small number of initial wavelengths, a dramatic reduction in the number of wavelengths required for building the PLS calibration models is observed. The fitness function used to guide the GA, the method of recombination used, and the effect of spectral resolution on the wavelength selection are also studied. In the resolution study, the original data with a point spacing of 2 cm⁻¹ are deresolved to 4-, 8-, and 16-cm⁻¹ point spacings by truncating the collected interferograms before applying the Fourier processing step. The use of lower resolution spectra is found to reduce further the number of final wavelengths selected by the GA, and the performance of the optimal calibration models obtained with the original spectra is maintained with the lower resolution spectra of both 4- and 8-cm⁻¹ point spacing. Degradation in performance is observed with the spectra computed with a point spacing of 16 cm⁻¹, however.

Multivariate calibration models are widely used in near-infrared (near-IR) spectroscopy to allow analyte spectral information to be extracted from overlapping spectral bands arising from the sample matrix. Wavelength selection methods are feature (variable) selection techniques that allow calibration models to be constructed with a subset of spectral points instead of with full spectra. This allows the wavelengths representing relevant spectral information to be selected, while points dominated by noise or other extraneous sources of variation are not included in the calibration model. The resulting calibration models may exhibit improved

performance relative to those based on full spectra.^{1,2}

Various wavelength selection algorithms and criteria for use with multivariate calibration have been reported.^{3–17} Partial least-squares (PLS) regression is widely used for processing full spectra because of its ability to extract analyte information from the many sources of variance within the spectral data matrix. Wavelength selection methods have traditionally not been used with PLS regression models because of this ability to decompose the data matrix in a manner biased toward the isolation of analyte-dependent information. However, recent studies have indicated that the performance of PLS models can be improved through wavelength selection.^{13–17} A mathematical justification of the theory that wavelength selection can enhance the performance of PLS models was also reported recently.¹⁷

- (1) Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472–1479.
- (2) Rossi, D. T.; Pardue, H. L. *Anal. Chim. Acta* **1985**, *175*, 153–161.
- (3) Kalivas, J. H.; Roberts, N.; Sutter, J. M. *Anal. Chem.* **1989**, *61*, 2024–2030.
- (4) Liang, Y.; Xie, Y.; Yu, R. *Anal. Chim. Acta* **1989**, *222*, 347–357.
- (5) Sasaki, K.; Kawata, S.; Minami, S. *Appl. Spectrosc.* **1986**, *40*, 185–190.
- (6) Salamin, P. A.; Bartels, H.; Forster, P. *Chemom. Intell. Lab. Syst.* **1991**, *11*, 57–62.
- (7) Brown, P. J. *J. Chemom.* **1993**, *7*, 255–265.
- (8) Brown, P. J. *J. Chemom.* **1992**, *6*, 151–161.
- (9) Lucasius, C. B.; Kateman, G. *TrAC, Trends Anal. Chem.* **1991**, *10*, 254–261.
- (10) Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. *Anal. Chim. Acta* **1994**, *286*, 135–153.
- (11) Jouan-Rimbaud, D.; Massart, D.; Leardi, R.; Noord, O. D. *Anal. Chem.* **1995**, *67*, 4295–4301.
- (12) Hörchner, U.; Kalivas, J. H. *Anal. Chim. Acta* **1995**, *311*, 1–13.
- (13) Rimbaud, D. J.; Walczak, B.; Massart, D.; Last, I. R.; Prebble, K. A. *Anal. Chim. Acta* **1995**, *304*, 285–295.
- (14) Navaroo-Vailoslada, F.; Perez-Arribas, L. V.; Leon-Gonzalez, M. E.; Polo-Diez, L. M. *Anal. Chim. Acta* **1995**, *313*, 93–101.
- (15) Bangalore, A. S.; Shaffer, R. E.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1996**, *68*, 4200–4212.
- (16) McShane, M. J.; Cote, G. L.; Spiegelman, C. H. *Appl. Spectrosc.* **1997**, *51*, 1559.
- (17) Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L. *Anal. Chem.* **1998**, *70*, 35–44.

To search for the optimal set of wavelengths, numerical optimization techniques such as genetic algorithms (GAs)^{9–11} and simulated annealing (SA)¹² have been employed. These optimization methods are efficient algorithms for interrogating a large search space in which many combinations of wavelengths are possible.

One of the research interests of our laboratories is to develop techniques based on near-IR spectroscopy for the measurement of glucose in various biological matrixes.^{15,18–25} As part of this work, a GA-based wavelength selection procedure for use with PLS regression has been successfully implemented to accomplish simultaneous optimization of the wavelengths selected and the number of latent variables employed in building a calibration model.¹⁵ In this paper, an enhanced GA-based wavelength selection procedure is developed through the investigation of strategies for initializing the GA in an optimal manner, further exploration of the configuration of the GA, and evaluation of the impact of spectral resolution on the wavelength optimization.

EXPERIMENTAL SECTION

Instrumentation and Reagents. Two data sets were employed for this research. One focused on the analysis of glucose in human serum samples (serum data set) and was collected at the University of Iowa. The other focused on the analysis of glucose in an aqueous matrix of bovine serum albumin (BSA) and triacetin (GTB data set) and was collected at Ohio University. The BSA and triacetin were used for modeling proteins and triglycerides, respectively, in human blood. These two data sets were used in the previous study.¹⁵

Spectra in the serum data set were collected with a Nicolet 740 Fourier transform spectrometer (Nicolet Instrument Corp., Madison, WI) configured with a 250-W tungsten–halogen source, CaF₂ beam splitter, and liquid nitrogen-cooled InSb detector. The near-IR spectral region of 5000–4000 cm⁻¹ was used. A K-band interference filter (Barr Associates, Westford, MA) was used to isolate this spectral region. The samples were placed in an Infracil quartz cell with 2.5-mm path length. The temperature of the samples was controlled to 37.0 ± 0.2 °C by use of a water-jacketed cell holder.

The GTB data set was collected with a Digilab FTS-60A Fourier transform spectrometer (Bio-Rad, Cambridge, MA) configured with a 100-W tungsten–halogen source, CaF₂ beam splitter, and InSb detector cooled with liquid nitrogen. The data were also collected over the spectral region of 5000–4000 cm⁻¹, which was isolated by a K-band interference filter (Barr Associates). An Infracil quartz cell with a path length of 2 mm was used, and the

temperature was controlled to 37–38 °C with a water-jacketed cell holder.

Human serum samples were collected from patients at the University of Iowa Hospitals and Clinics. The glucose levels in the samples were determined with a conventional clinical glucose analyzer by the hospital clinical chemistry laboratory. The precision of such measurements is typically in the range of 0.3 mM.²⁶ All serum samples were frozen until just before the collection of the spectra. The procedures used in collecting and handling the serum samples complied with approved ethical and safety standards at the University of Iowa. The 235 samples used for this study spanned a range of 3.2–31.9 mM glucose concentration.

The GTB samples were prepared in a pH 7.4, 0.1 M phosphate buffer solution. 5-Fluorouracil (0.044% w/w) was added as a preservative. Regent-grade glucose, sodium phosphate salts, 5-fluorouracil, triacetin, and BSA were purchased from common suppliers. The reagent-grade water used for the preparation of the sample solutions was obtained from a Mili-Q Plus water purification system (Millipore, Inc., Bedford, MA). A factorial design was adopted for the concentration levels of glucose, BSA, and triacetin to minimize the correlation among these components. The data set consisted of samples prepared from all combinations of 10 levels of glucose (1, 3, 5, 7, 9, 11, 13, 15, 17, and 19 mM), four levels of BSA (50, 65, 80, and 95 g/L), and four levels of triacetin (1.4, 2.1, 2.8, and 3.5 g/L). A total of 160 (10 × 4 × 4) samples were prepared for the GTB data set.

Procedures. Double-sided interferograms of 16 384 points were collected for the serum data set. Two to four replicate interferograms were collected for each serum sample based on 256 coadded scans. The single-beam spectra were computed from the collected interferograms with software resident on the Nicolet 620 computer controlling the spectrometer. Triangular apodization and Mertz phase correction were used in Fourier processing the interferograms. The resulting spectra had a nominal point spacing of 2 cm⁻¹. The samples were measured in a randomized order with respect to glucose concentration. Spectra of a pH 7.3, 0.1 M phosphate buffer were acquired periodically for use as background spectra in computing spectra in absorbance units.

The procedure for collecting the GTB spectra was similar to that used for the serum data. However, single-side interferograms of 16 384 points were collected based on 256 coadded scans. Again, single-beam spectra with a nominal point spacing of 2 cm⁻¹ were computed from the collected interferograms, and triangle apodization and Mertz phase correction were employed. The software used for the Fourier processing was resident on the Bio-Rad SPC-3200 computer controlling the spectrometer. Three replicate interferograms were collected for each sample. The data collection was also randomized with respect to glucose concentrations, and interferograms of phosphate buffer were collected periodically for use in computing spectra in absorbance units.

All data analysis was performed with a Silicon Graphics Indigo² R10000 workstation (Silicon Graphics, Mountain View, CA) operating under Irix (version 6.2). The software used for the data analysis was written in Fortran 77. Subroutines used for multiple linear regression computations were obtained from the IMSL software package (IMSL, Inc, Houston, TX).

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

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

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

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

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

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

(24) Mattu, M. J.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1997**, *69*, 4695–4702.

(25) Ding, Q.; Small, G. W. In *Fourier Transform Spectroscopy: 11th International Conference*; de Haseth, J. A., Ed.; American Institute of Physics: Woodbury, NY, 1998; pp 264–267.

(26) Burmeister, J. J.; Arnold, M. A. *Anal. Lett.* **1995**, *28*, 581–592.

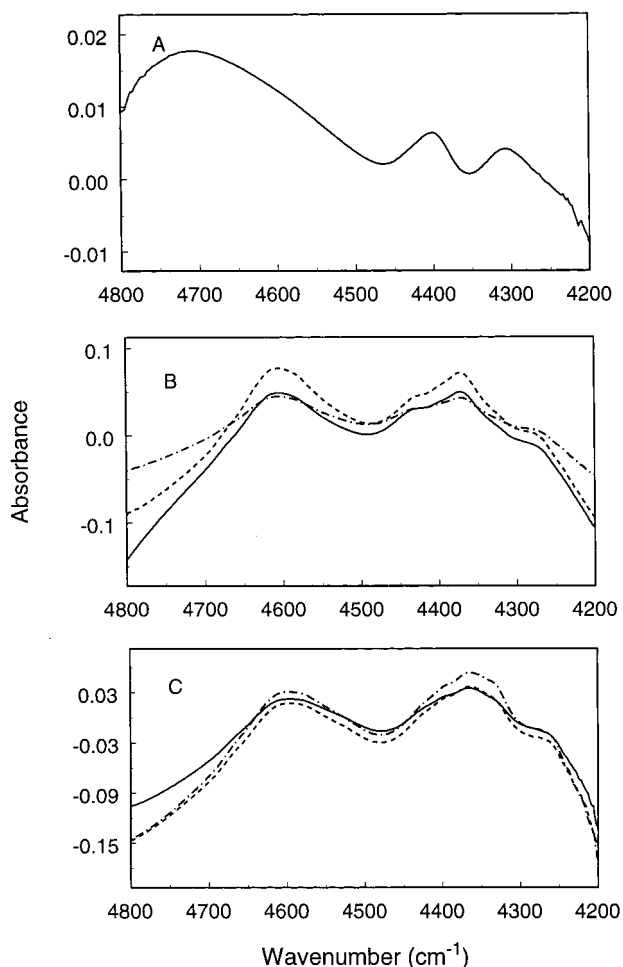


Figure 1. Near-IR absorbance spectra of (A) glucose at 96 mM, (B) three GTB samples with the glucose concentration at 19 mM, and (C) three human serum samples from three patients. The glucose concentration for each serum sample was 17.0 mM.

Spectra at reduced resolution were obtained by truncating the interferograms appropriately before the Fourier processing step. Through this procedure, spectra with nominal point spacings of 4, 8, and 16 cm^{-1} were obtained for both data sets. The Fourier processing calculations for the spectra at reduced resolution were performed with original software implemented on the Silicon Graphics system. Triangular apodization and Mertz phase correction were also used in the computation of these spectra.

RESULTS AND DISCUSSION

The characteristics of the data sets employed in this study have been detailed in the previous paper.¹⁵ Overall, these two data sets have relatively low signal-to-noise ratios in terms of the analyte (glucose) absorption features, and the glucose spectral information is overwhelmed by the spectral features arising from the other constituents of the sample matrix. A typical near-IR absorbance spectrum of glucose in water at a concentration of 96 mM is shown in Figure 1A over the range of 4800–4200 cm^{-1} . In this region, glucose absorption bands are located near 4700, 4400, and 4300 cm^{-1} . The C–H combination band centered near 4400 cm^{-1} was found most useful in modeling glucose concentrations in a previous study.¹⁸

Figure 1B shows the spectra of three GTB samples that have the same glucose concentration of 19 mM. The three samples

consisted of 79.8 g/L BSA/3.5 g/L triacetin, 95.0 g/L BSA/1.4 g/L triacetin, and 49.4 g/L BSA/2.1 g/L triacetin, respectively. Figure 1C shows the spectra of three serum samples from three different patients who happened to have the same blood glucose concentration of 17.0 mM. The other constituents analyzed in these samples were 59 g/L total protein, 131 mg/dL cholesterol, 0.65 g/L triglyceride, and 24 mg/dL urea; 71 g/L total protein, 187 mg/dL cholesterol, 1.85 g/L triglyceride, and 22 mg/dL urea; and 79 g/L protein, 294 mg/dL cholesterol, 3.33 g/L triglyceride, and 20 mg/dL urea. In Figure 1B and C, no glucose absorption features can be observed in the spectra of the GTB and serum samples. Instead, the other components (especially the proteins¹⁵) produce the dominant spectral features in the near-IR region. Also, although the samples have the same glucose concentrations, there are significant variations in the spectra because the samples have different concentrations of the other components in the matrix. The complexity of these data sets illustrates the challenge of determining glucose in biological matrixes by near-IR spectroscopy and prompts the need for suitable data analysis methods for use in extracting the glucose information from the spectra.

A GA-based wavelength selection method has been implemented in our laboratories to allow joint optimization of the wavelengths used and the number of PLS factors employed to build optimal calibration models.¹⁵ Three near-IR data sets were used in that work, corresponding to the measurement of glucose in the same serum and GTB data sets used here and a data set focusing on the determination of methyl isobutyl ketone (MIBK) in water. Of the three data sets, the MIBK measurement represented the simplest determination, while the GTB and serum data sets, respectively, were progressively more challenging.

In the previous study, although a significant reduction in the number of wavelengths used to build the calibration models was realized relative to the use of full spectra, several hundred wavelengths were still used in the optimal models built for the two glucose data sets. The purpose of the work reported here was to explore ways to improve the original GA-based wavelength selection procedure in an effort to decrease the number of wavelengths selected in building calibration models for challenging measurements such as the glucose determination.

Overview of GA-Based Wavelength Selection. GAs are efficient numerical optimization methods based on the principles of genetics and natural selection. Since efficient optimization is one of the key requirements in implementing wavelength selection, GA-based methods have become popular for selecting subsets of wavelengths for use in building multivariate calibration models.^{9–11,15} Shaffer and Small have described the basic steps of implementing a GA-based optimization.²⁷ These concepts will be summarized briefly here.

In a GA, the collection of variables whose values are to be optimized is termed a chromosome, and the individual variables are called genes. A chromosome represents a candidate solution to the optimization problem. In pursuit of the optimal chromosome, a GA operates simultaneously on a group of chromosomes called a population. The first population is generated by perturbing an initial chromosome that is either generated randomly or supplied by the user.

(27) Shaffer, R. E.; Small, G. W. *Anal. Chem.* **1997**, *69*, 236A–242A.

After the first population is formed, the fitness of each of the individual chromosomes in the population is evaluated on the basis of a user-defined objective function (fitness function). To implement a GA successfully, the fitness function must be selected to encode the degree to which the settings of the variables in the chromosome are optimal.

The chromosomes with the best fitness values are selected to generate a new set of child chromosomes through the methods of recombination and mutation. The recombination approach that we have typically employed is termed single-point crossover. Given two selected parent chromosomes, a gene location on the chromosome is chosen randomly, and the values of all the genes up to that point are interchanged between the two parents to form two new child chromosomes. Mutation is applied to the child chromosomes and involves altering the gene values on a gene-by-gene basis. Whether or not mutation occurs for a given gene is governed by a user-specified mutation probability. Recombination and mutation introduce diversity into the child chromosomes while preserving the information carried by the parents.

The new population formed with the child chromosomes replaces the original, and the chromosomes with the best fitness values in the new population are again selected to reproduce through recombination and mutation. This procedure is an iterative evolutionary process in search of the chromosome with the highest fitness value. The formation of each new population represents one iteration of the algorithm and is termed a generation. The algorithm terminates after a fixed number of generations or when a chromosome with a user-specified level of fitness is found.

In our implementation of the GA, the values to be optimized were which of the individual spectral points to use as input variables in the PLS calculation and the number of the resulting PLS factors to use in constructing the calibration model. The chromosome consisted of a binary gene for each spectral point and an integer gene to store the number of PLS factors. The binary genes stored values of 1 or 0 indicating whether the corresponding spectral point was included in the PLS calculation. For the data with a nominal spectral point spacing of 2 cm^{-1} , there were 519 resolution elements between 5000 and 4000 cm^{-1} . Thus, in this case, the chromosome consisted of $519 + 1 = 520$ genes. The order of the 519 genes was the same as that of the corresponding points in the spectrum (e.g., the first and second genes corresponded to 5000 and 4998 cm^{-1} , respectively).

The initial population was formed by randomly perturbing an initial starting chromosome. The perturbation of each binary gene involves changing the value of the gene from 1 to 0, or vice versa, according to an initial probability set by the user. The perturbation of the number of PLS factors was performed by scaling a Gaussian-distributed random deviate with a step size and adding the scaled value to the previous number of PLS factors used. These same perturbation steps were also used to mutate genes in child chromosomes formed through the recombination process.

To implement the GA-based wavelength selection, the data sets were partitioned randomly into a calibration set and a prediction set, as shown in Table 1. The spectra in the calibration set were used during the GA calculations, while those in the prediction set were withheld entirely from the optimization and used subsequently to assess the performance of the optimized calibra-

Table 1. Data Set Partitioning

data set	no. of samples (spectra)	
	serum	GTB
calibration set	188 (561)	120 (360)
prediction set	47 (140)	40 (120)
total	235 (701)	160 (480)

tion models. The replicate spectra of the samples were allocated together into the corresponding data subsets.

During the optimization, for each calculation of the fitness function, the calibration set was further divided randomly three times to produce three calibration subsets (80% of the calibration samples) and three monitoring sets (20% of the calibration samples). As before, the replicate spectra of the samples were allocated together into the data subsets. For the chromosome being evaluated, a calibration model was computed with the spectra in each calibration subset, and each resulting model was used to predict the glucose concentrations for the spectra in the corresponding monitoring set. This allowed the predictive ability of the model to be made part of the fitness evaluation. The use of multiple calibration/monitoring sets and repartitioning the data before each fitness calculation helped to keep individual samples from having undue influence on the fitness value.

The fitness function used in two previous studies^{15,22} was

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

where MSE is the mean squared error in concentration of spectra in the calibration subset, MSME is the mean squared error in concentration of spectra in the monitoring set, h is the number of PLS factors employed in the calibration model, and w is a weighting factor that controls the influence of h on the fitness value. The incorporation of h into the fitness function allows a joint optimization of the selected wavelengths and the model size without the requirement for a separate optimization of h at each evaluation of the fitness function. If desired, the final model produced by the optimization can be evaluated further to ensure that the value of h is optimal. Procedures based on randomization tests²⁸ and the statistical F -test²⁹ are commonly used for this purpose.

In practice, the fitness value is taken as the mean of eq 1, computed across the three calibration subset/monitoring set combinations. This fitness function was used as the starting point for the work reported here. As detailed below, further investigation of this function was performed in subsequent studies. For use with eq 1, a value of $w = 0.45$ was optimized for the GTB data set through a procedure described previously to balance the predictive performance provided by the calibration models and the model sizes.²² A value of $w = 2.0$ was found to be optimal previously for the serum data set and was also used in this research.

For the current research, the GA configuration was similar to that used in the previous work.¹⁵ The single-point crossover method of recombination was employed in the initial experiments.

(28) van der Voet, H. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313–323.

(29) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193–1202.

As detailed below, further investigation of this parameter was subsequently performed. The other GA parameters optimized previously were adopted in the current work without further study. These included a mutation probability of 0.001, a recombination probability of 0.9, and the use of 100 generations in the optimization. The population size was 100 and 150 chromosomes for the serum and GTB data sets, respectively.

Investigation of Factors Influencing the Number of Wavelengths Selected. Previous work has suggested that the choice of an initial chromosome has a critical effect on the final wavelengths selected by a GA. In the previous study,¹⁵ all wavelengths in a specified spectral range were used in the initial chromosome. The spectral ranges selected were those that included the glucose absorption bands and had proven useful previously for building calibration models. Therefore, in the previous work that employed contiguous sections of the spectrum, the spectral ranges used for the GA initialization were relatively broad. With this approach, a total of 292 wavelengths were selected by the GA in building the optimal calibration model for the serum data set, and 150 wavelengths were needed for the optimal calibration model built with the GTB data set.

In the current research, our goal was to decrease the number of wavelengths required to build an effective calibration model. We hypothesized that one strategy for reducing the number of final wavelengths might be to reduce the number of wavelengths selected in the initial chromosome. In the previous study,¹⁵ all the wavelengths over the spectral range of 4850–4250 cm^{-1} (312 spectral points) were used for the initial chromosome for the serum data set. Interestingly, 286 out of the 312 wavelengths in this range remained in the final 292 wavelengths selected by the GA to build the optimal calibration model. The GA appeared not to be efficient at deleting wavelengths from the initial chromosome.

To investigate this phenomenon, the initial chromosome was set to a narrower spectral range of 4460–4420 cm^{-1} (22 spectral points). Three GA runs were performed through the use of three different seeds for the random number generator. The same three seeds were used for all the experiments. On the basis of the computed fitness values, the 10 best chromosomes were saved from each GA run. The overall top five chromosomes from the three GA runs were then selected to build the optimal calibration models. Using the full calibration set of spectra, the five calibration models were constructed on the basis of the specified wavelengths and the selected numbers of PLS factors. The resulting models were then applied to predict the glucose concentrations corresponding to the spectra in the independent prediction set. The standard error of prediction (SEP) was computed to characterize the prediction results.

Of the five models tested, the model that produced the lowest SEP will be used for comparison with the model produced previously with the initialization range of 4850–4250 cm^{-1} . The number of wavelengths selected by the GA was decreased from 292 to 91 with the initialization of the narrow range. This represents a reduction of more than a factor of 3 in the number of wavelengths selected. Furthermore, the prediction results are better for the calibration model built with fewer wavelengths (SEP = 1.31 mM vs 1.44 mM for the model based on 292 spectral points). This suggests that the initialization of a narrow spectral

range helps to decrease the number of wavelengths in the optimal sets selected by the GA.

Figure 2A displays the 91 wavelengths present in the best chromosome produced with the initial range of 4460–4420 cm^{-1} . Similar to the case with the initialization of the broad range of 4850–4250 cm^{-1} , 19 of the 22 spectral points in the range of 4460–4420 cm^{-1} are also retained in the optimal set of wavelengths. This suggests that the initialization of all wavelengths in a specified range may not be a wise strategy for the GA-based wavelength selection. The high correlation between the spectral information encoded in adjacent spectral points in the specified range might prevent the GA from removing the wavelengths within that range. Since the model typically performs no worse with the additional wavelengths included, there is no driving force to remove them once they have been added. For example, while the fitness function specified in eq 1 applies a penalty to calibration models constructed with a large number of PLS factors, there is no similar penalty applied to models on the basis of the number of wavelengths used. Furthermore, since the single-point crossover recombination procedure swaps entire sections of the parent chromosomes, adjacent wavelengths are naturally carried along into the child chromosomes.

On the basis of these observations, three modifications to the optimization were investigated. First, a procedure was investigated in which a random set of wavelengths was initialized within a specified range. Second, modifications to the fitness function of eq 1 were studied. Third, an alternate recombination method was evaluated.

Random Initialization of Wavelengths. A procedure was evaluated in which approximately 10% of the wavelengths in different ranges were randomly selected as the initial wavelengths. The optimal results obtained with this procedure with the serum data set are reported in Table 2. Over each range listed in Table 2, the random selection of initial wavelengths significantly decreased the numbers of wavelengths required in the optimal calibration models compared to the number of wavelengths selected with the initialization of the broad range of 4850–4250 cm^{-1} . The numbers of wavelengths present in the optimal chromosomes with the initialization ranges of 4800–4200 and 4700–4300 cm^{-1} are smaller than those produced with the initialization ranges of 5000–4000 and 4900–4100 cm^{-1} . Also, the SEP values in Table 2 demonstrate that improved model performance is obtained when the 4800–4200- and 4700–4300- cm^{-1} initialization ranges are used. This observation is consistent with the fact that the glucose absorption bands are located within the range of 4800–4200 cm^{-1} and that the 5000–4800- and 4200–4000- cm^{-1} regions exhibit increased spectral noise.¹⁵ On the basis of these results, random initialization over the 4800–4200- cm^{-1} range was adopted for all further work with the serum data set.

The procedure of random selection of initial wavelengths was also applied to the GTB data set. Because the optimal calibration models previously built for the GTB data set were based on spectral ranges around 4700–4300 cm^{-1} , this range was used for the random selection of the initial wavelengths for the GTB data set. As expected, with random selection of the initial wavelengths, the number of wavelengths present in the optimal chromosome dramatically decreased. Comparison of the optimal results obtained with the initialization of all wavelengths in the spectral

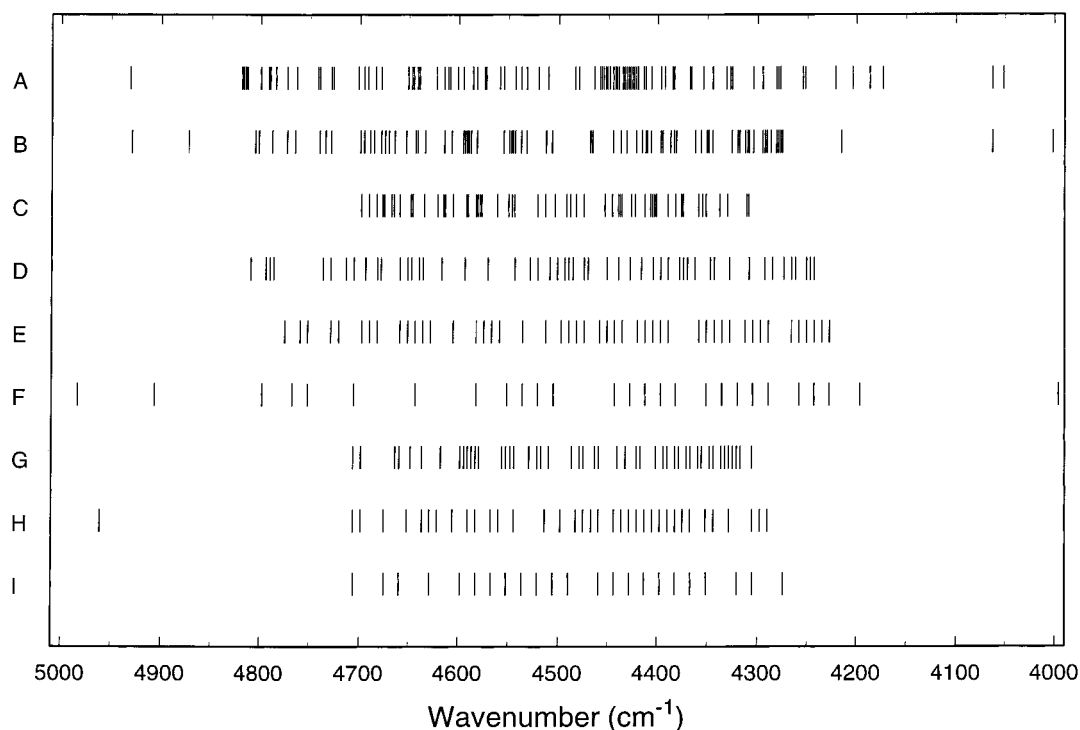


Figure 2. Spectral points present in the best chromosome selected by the GA. (A) serum data set, 2-cm⁻¹ point spacing, all wavelengths in the range of 4460–4420 cm⁻¹ selected in the initial chromosome; (B) serum data set, 2-cm⁻¹ point spacing, random selection of initial wavelengths from the range of 4800–4200 cm⁻¹; (C) GTB data set, 2-cm⁻¹ point spacing, random selection of initial wavelengths from the range of 4700–4300 cm⁻¹; (D–F) serum data set, random selection of initial wavelengths from 4800 to 4200 cm⁻¹, point spacings of (D) 4, (E) 8, and (F) 16 cm⁻¹; (G–I) GTB data set, random selection of initial wavelengths from 4700 to 4300 cm⁻¹, point spacings of (G) 4, (H) 8, and (I) 16 cm⁻¹.

Table 2. Effect of Initial Range with Random Selection for the Serum Data Set

initial range (cm ⁻¹)	no. of wavelengths		no. of PLS factors	SEC ^a (mM)	SEP ^b (mM)
	initial	selected			
5000–4000	50	97	24	1.35	1.55
4900–4100	42	91	21	1.36	1.54
4800–4200	29	77	19	1.26	1.35
4700–4300	18	72	21	1.35	1.37

^a Standard error of calibration computed from the residuals of the calibration model. ^b Standard error of prediction.

range of 4675–4375 cm⁻¹ (156 spectral points initially selected) and random selection of 10% of the wavelengths in the range of 4700–4300 cm⁻¹ (23 points initially selected) reveals a reduction in selected wavelengths from 150 to 55. The SEP values produced by the corresponding models are effectively identical (SEP = 0.63 mM for the model based on 150 wavelengths vs 0.61 mM for the model based on 55 points). A factor of 3 reduction in the number of wavelengths selected in the optimal chromosome was again obtained with no degradation in model performance.

Parts B and C of Figure 2 display the optimal sets of wavelengths selected by the GA with the random initialization procedure for the serum and GTB data sets, respectively. Panels A and B of Figure 3 are correlation plots of predicted vs actual or measured glucose concentrations obtained from these optimal calibration models for the GTB and serum data sets, respectively. Even with many fewer wavelengths in the calibration models, good correlations between predicted and actual glucose concentrations are observed in both plots.

Evaluation of Fitness Function Modifications. Modifications to eq 1 were investigated in an effort to limit the number of wavelengths used in building the calibration models. By adding the number of wavelengths, p , used in computing the PLS factors to the denominator of the equation, models based on fewer wavelengths were given an increased fitness score. Experiments were performed in which the number of wavelengths was used directly in the equation and in which this value was weighted in various ways (analogous to the weighting of h in eq 1). The modified fitness functions were tested both with and without the random initialization procedure described in the previous section. The results of these experiments indicated clearly that the random initialization procedure was the key step in limiting the number of wavelengths used in the optimal calibration model. When used in conjunction with random initialization, the modified fitness functions worked well. Without random initialization of the starting wavelengths, however, the modified fitness functions did not perform significantly better than eq 1. On the basis of these results and given that the inclusion of an additional term in the fitness function raises additional concerns about appropriate balancing of the contributions of MSE, MSME, h , and p , it was decided to retain eq 1 as the fitness function for subsequent work.

Evaluation of Uniform Crossover Recombination. Since the single-point crossover recombination method interchanges contiguous sections of the parent chromosomes in the creation of the new child chromosomes, it was hypothesized that this procedure contributes to the carry-along of adjacent spectral points that contribute potentially redundant information. To address this issue, optimizations were also performed with the uniform

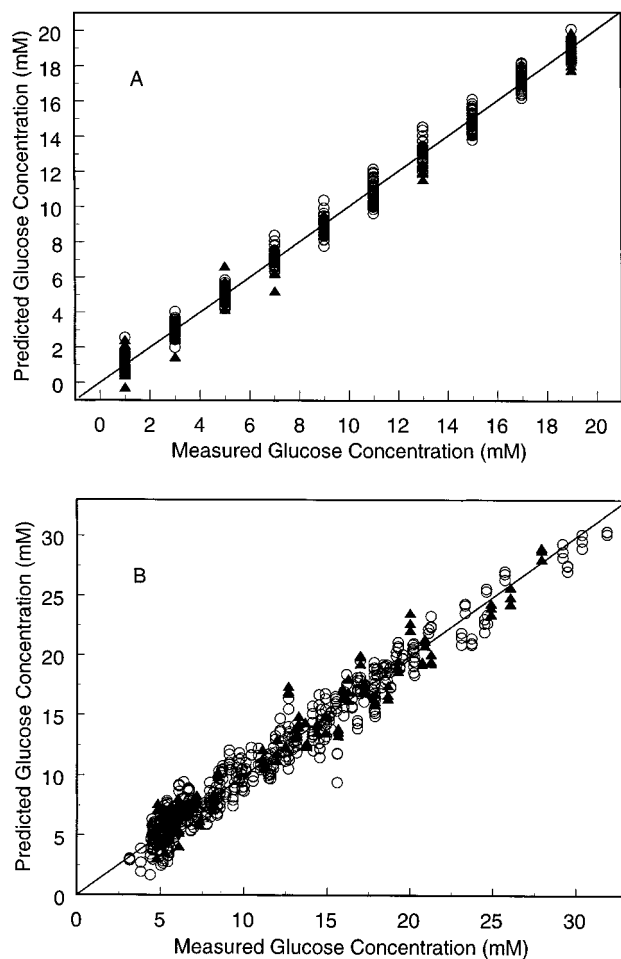


Figure 3. Glucose concentration correlation plots for (A) GTB data set, 2-cm^{-1} point spacing, random selection of the initial wavelengths, and (B) serum data set, 2-cm^{-1} point spacing, random selection of the initial wavelengths. The open circles and solid triangles denote the spectra in the calibration and prediction sets, respectively.

crossover method of recombination. With this method, the interchange is performed on a gene-by-gene basis. The decision of whether to interchange a given gene between the parents is random but is governed by a uniform crossover probability (e.g., a 30% probability that genes will be exchanged).

This recombination method was compared to the single-point crossover approach in a series of trials. However, as in the experiments with the modified fitness functions, the key to reducing the number of wavelengths needed in the calibration model was again the random initialization procedure. Given that the uniform crossover method adds an additional GA configuration parameter (i.e., the crossover probability), it was decided to retain the single-point crossover technique in subsequent work.

Effect of Spectral Resolution. The use of lower resolution spectra is analogous to the selection of wavelengths at equidistant locations in the original spectra. For this reason, we investigated the effect of spectral resolution on the GA-based wavelength selection procedure. Also, spectral resolution is an important experimental parameter in near-IR spectroscopy because the required resolution affects the complexity of a dedicated instrument that might be used to implement an analysis such as the measurement of glucose in a biological sample. Preliminary

Table 3. Optimal Results with the Serum Data Sets of Different Resolutions

point spacing (cm^{-1})	no. of wavelengths selected	no. of PLS factors	SEC ^a (mM)	SEP ^b (mM)
2	77	19	1.26	1.35
4	52	19	1.30	1.32
8	48	15	1.50	1.38
16	27	17	1.54	1.54

^a Standard error of calibration computed from the residuals of the calibration model. ^b Standard error of prediction.

Table 4. Optimal Results with the GTB Data Sets of Different Resolutions

point spacing (cm^{-1})	no. of wavelengths selected	no. of PLS factors	SEC ^a (mM)	SEP ^b (mM)
2	55	13	0.53	0.61
4	48	13	0.57	0.61
8	37	13	0.63	0.68
16	23	13	0.67	0.84

^a Standard error of calibration computed from the residuals of the calibration model. ^b Standard error of prediction.

results obtained from the study of the effect of spectral resolution on wavelength selection with the serum data set have been reported.²⁵

The same procedure of random selection of initial wavelengths with a limited spectral range described in the previous section was used for the resolution study. In addition, eq 1 was used as the fitness function, and the single-point crossover method of recombination was employed.

The initialization probability was set at 10% for the original serum and GTB data sets at the 2-cm^{-1} spectral point spacing. However, to maintain approximately the same number of initial wavelengths selected with the different resolutions, the initialization probability was set at 20%, 40%, and 80% for the data sets based on spectra with 4-, 8-, and 16-cm^{-1} point spacings. Tables 3 and 4 summarize the optimal results obtained with the serum and GTB data sets of different resolutions, respectively. Also, the mean SEP values and the corresponding 95% confidence limits computed from the prediction results with the top five calibration models are displayed in panels A and B of Figure 4 for the GTB and serum data sets of the different resolutions, respectively. As before, those five models are based on the top five chromosomes from the three replicate GA runs.

For both the serum and GTB data sets, as the spectral resolution decreases, the number of wavelengths in the optimal chromosome also decreases. This is expected because the total number of wavelengths available to be selected is reduced. Encouragingly, as the resolution decreases from 2 to 8-cm^{-1} , the overall prediction results (both optimal results and mean SEP values) for both data sets are maintained, while the number of wavelengths in the optimal chromosome is reduced from 77 to 48 for the serum data set and from 55 to 37 for the GTB data set. As the resolution further decreases to 16-cm^{-1} , however, the prediction results become significantly worse for both data sets. Parts D–F of Figure 2 display the wavelengths present in the

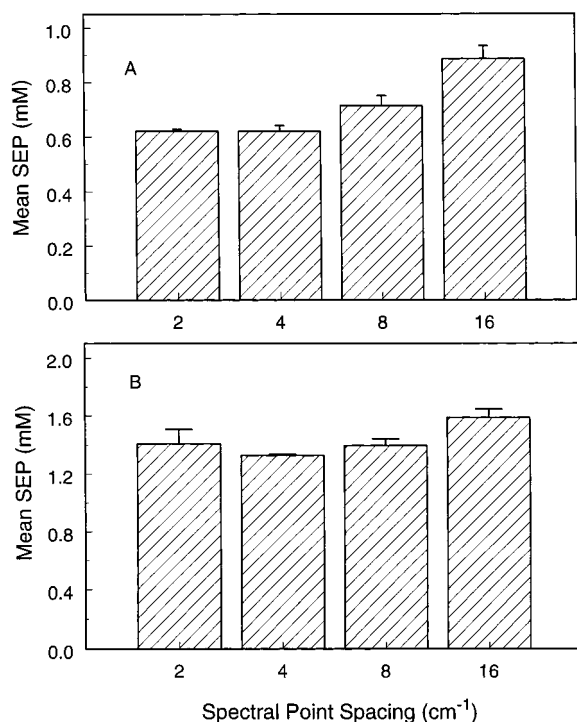


Figure 4. The mean SEP values and the corresponding upper 95% confidence limits (error bars) from the top five models generated for the data sets of different spectral resolutions. (A) GTB data set. (B) serum data set.

optimal chromosomes for the different resolutions (4, 8, and 16- cm^{-1} point spacings, respectively) for the serum data set. Parts G–I of Figure 2 present the corresponding selected wavelengths for the GTB data sets with point spacings of 4, 8, and 16 cm^{-1} , respectively.

Although these wavelengths were selected in building the optimal calibration models for the data sets of different resolutions, they do not represent a unique collection of wavelengths which are required to build a good calibration model. They are displayed to show the rough pattern of the distribution of the wavelengths selected by the GA-based procedure. It is clear from an inspection

of parts A–I of Figure 2, however, that through all the optimal sets of wavelengths selected, spectral points relevant to glucose information (e.g., close to the 4400- and 4300- cm^{-1} glucose bands) were selected for use in building the glucose calibration models.

CONCLUSIONS

The choice of initial wavelengths is critical for GA-based wavelength selection. The choice of a starting chromosome not only affects the efficiency of the GA to search for the optimal set of wavelengths but also has a crucial effect on the number of wavelengths present in the optimal chromosome. With random selection of a relatively small number of initial wavelengths, the number of wavelengths selected by the GA in building the optimal calibration models has been dramatically reduced. With the use of lower spectral resolution, the GA optimization efficiency can be further improved and the number of wavelengths in the optimal chromosome can be further decreased. For the glucose analysis, the prediction results based on the optimal calibration models with data sets of lower resolution (i.e., 4- and 8- cm^{-1} point spacing) are not significantly different from those obtained with data sets of the original 2- cm^{-1} point spacing. Reduced performance was obtained with spectra at 16- cm^{-1} point spacing, however. This indicates the feasibility of use of lower resolution spectra for wavelength selection and also for other data analysis methodologies applied to the near-IR measurement of glucose in biological matrixes.

ACKNOWLEDGMENT

This research was supported entirely by the National Institutes of Health under Grant DK45126. Mutua Mattu, Ndumiso Cingo, and Kevin Hazen are thanked for their assistance in collecting the spectral data used in this research. Ronald Feld is thanked for his help in obtaining the glucose levels in the human serum samples. Ronald Shaffer and Arjun Bangalore are acknowledged for writing the original version of the GA software.

Received for review April 27, 1998. Accepted August 23, 1998.

AC980451Q