Wavelength Selection for Multivariate Calibration Using a Genetic Algorithm: A Novel Initialization Strategy

Héctor C. Goicoechea[†] and Alejandro C. Olivieri*[,‡]

Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, Santa Fe (3000), Argentina, and Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario (S2002LRK), Argentina

Received March 28, 2002

Genetic algorithms and other procedures mimicking natural processes are being increasingly used for variable selection, to improve the predictive ability of partial least-squares multivariate calibration. Two issues are critical for the success of genetic algorithms: initialization (setting the first candidates for solving the problem at hand) and overfitting (the tendency to produce excellent results when training, but poor predictions toward fresh samples). A new procedure is presented for sensor selection problems, involving iterative reinitialization based on a statistical analysis of the included sensors. It is shown to give excellent results without the requirement of preparing independent test data sets. Monte Carlo simulations using a theoretical three-component example illustrate how partial least-squares regression greatly benefits from variable selection when the analyte of interest is diluted, and how the new initialization method compares with other strategies. The new genetic algorithm was applied to five experimental data sets. The target parameters were the concentrations of diluted analytes in four pharmaceutical mixtures studied by UV−visible spectrophotometry and the octane number in gasolines analyzed by near-infrared spectroscopy.

1. INTRODUCTION

Multivariate calibration techniques such as partial least-squares (PLS) are widely used to extract relevant information from different types of spectral data, to predict analyte concentrations or properties in samples with complex background components.[1] Due to its efficiency in accomplishing the latter task, it was initially believed that PLS yielded equally good results whether the calibration models were constructed using full spectra or a restricted subset of spectral data. However, it has been shown both experimentally and theoretically that there are reasons to expect an improvement in performance after suitable sensor selection.[2−4]

Ideally, sensor selection trains the PLS model with spectral points containing information which is related to variations in the concentration of a given analyte, discarding those dominated by (1) spectral sources of variability which are irrelevant as to the analytical problem at hand, such as noise, temperature changes, or presence of interferents and (2) signal intensities which are nor linearly related to the target parameter.[5] In practice, wavelength selection seeks for the best combination of data points which will optimize a certain objective function. Examples of functions that measure the quality of the calibration model are the cross-validation variance,[6] the degree of departure of net analyte regression plots from linearity,[7] the root-mean-square predicted error for a test set of samples, the correlation coefficient, etc. Objective functions which are a combination of parameters such as prediction error, number of latent variables, and number of included sensors have also been employed.[8]

Many different selection procedures have been discussed in the literature, including simulated annealing,[9] artificial neural networks,[10] genetic algorithms (GA),[11] and a variety of other techniques.[12] In some cases, extensive searches have been conducted using moving window strategies in order to find the best figure of merit.[7,13] These procedures do not allow for the study of multiple wavelength regions, since a comprehensive search of this type would be prohibitively time-consuming. For spectra composed of hundreds or thousands of data points, a comprehensive search of all combinations of sensors is impractical. Several algorithms, however, make such a search unnecessary. An example is the genetic algorithm, a numerical optimization technique which mimics the process of natural selection.[14,15]

In a GA, a probabilistic approach inspired in natural selection mechanisms is applied, employing binary strings (the chromosomes) containing genes which encode the experimental variables.[16] An initial population is produced in the form of an $M \times N$ binary matrix, where M is the number of variables (either individual wavelengths or wavelength intervals) and N is the number of chromosomes. The PLS models specified by these chromosomes are built and ranked according to a given figure of merit (the objective function to be minimized or maximized). The chromosomes with the best figures of merit are allowed to survive, mutate, and recombine to produce offsprings. After a number of generations in which only a group of selected chromosomes is retained, the information encoded in a small number of

* Corresponding author phone: 54-341-4372704; e-mail: aolivier@fbioyf.unr.edu.ar.
[†] Universidad del Litoral.
[‡] Universidad de Rosario.

10.1021/ci0255228 CCC: $22.00
Published on Web 08/21/2002

top chromosomes is subsequently checked against an independent test set of samples. The best chromosome is then employed for model build and prediction.[8]

The selection of the first population of chromosomes as candidates for the best fit to the objective function (the initialization process) has been found to be a critical step for the success of the GA in sensor selection.[8] It has been suggested that the high correlation between the information encoded in adjacent data points may prevent the GA from removing unwanted sensors, especially when single-point crossover is employed for chromosome recombination.[8] In this latter scheme, a random point is first selected along a pair of parent chromosomes. The entire genetic information encoded in one of the parents up to the selected point is then transferred to the offspring, while the remaining genes are taken from the other parent (an alternative information transfer operates in creating a second offspring). To reduce the number of wavelengths employed for regression, a procedure has been explored in which the initial chromosomes contained a given percentage (10−20%) of all possible wavelengths, selected at random from the full spectral calibration range. A provision was also made for avoiding an unnecessary increase in the number of wavelengths, by including an appropriate term in the objective function.[8] This procedure was not as successful as could be anticipated on the basis of the known ability of GA to find global minima.

One conclusion drawn from previous studies is that the best initialization method should include those spectral regions where the analyte of interest is known to respond.[8] However, this may not be useful in cases in which (1) the target parameter is not the concentration of a given analyte but a property of the system which depends in a rather complicated way on the concentrations of several components, as when predicting octane number in gasolines or viscosity in oils, or (2) the response of the pure analyte is known but is seriously affected by interaction with background constituents. In any case, initialization procedures requiring previous knowledge of adequate spectral regions conspire against the automatic selection of sensors.

The risk of overfitting is another important issue regarding the use of GA in sensor selection.[11] Chance predictions, along with the tendency of GA to carry along entire chromosome regions through recombination, make these tools prone to overfit the training set of spectra and to show a poor behavior toward new samples. For these reasons, it is common practice to check a number of top chromosomes against an independent test set of samples (not previously used for training) after stopping the algorithm.[11] This obviously requires the preparation of additional samples than those employed in the calibration set, which may constitute a problem when samples are valuable, such as those of natural origin.

There is an alternative strategy which does not require an independent test set. It iteratively reinitializes the GA according to the results provided by a statistical study of repeated randomly initialized GA calculations. We will distinguish between the usual implementation of a GA, herein designed as randomly initialized genetic algorithm (RIGA) from the new, iteratively reinitialized genetic algorithm (IRGA). The performance of both methods is illustrated by Monte Carlo simulations on theoretical examples. A number of experimental cases where PLS models greatly benefit from wavelength selection is also presented.

2. GENETIC ALGORITHMS

The RIGA was implemented starting with a population of 20 chromosomes, initialized with 50% of all wavelengths selected at random from the full spectral range. The choice of 50% of sensors for initialization allows a better comparison with the results provided by IRGA (see below). The main results obtained with the simulated data sets do not significantly change if 10−20% of all sensors are initially included in RIGA.

Each gene was chosen to encode regions of three consecutive data points. The single crossover scheme with 50% probability was employed for recombination (the alternative multiple crossover procedure gave similar results), and a probability of 0.05 was applied to mutations after offsprings were produced. The algorithm was stopped after 50 generations. The setup of the algorithm is similar to that recently described for classification purposes.

To define the objective function to be optimized, the calibration set was randomly divided into two subsets, one used for calibration (including 70% of the calibration samples) and the other one for monitoring (the remaining samples), and the process was repeated three times using different random seeds for partitioning the calibration set.[8] The root-mean-square error of prediction (RMSEP) values were then calculated for a number of factors ranging from 1 to a certain maximum (estimated from full spectral cross-validation on the complete calibration data set). In each of these three calibration/prediction procedures, the minimum error ($RMSEP_{min}$) was first considered, but the selected RMSEP was the one which was not higher than the product $F_{\alpha,I\mathrm{cal}-A,I\mathrm{mon}-A} \times RMSEP_{min}$ ($F_{\alpha,I\mathrm{cal}-A,I\mathrm{mon}-A}$ is the statistical F ratio computed for $\alpha = 0.05$ with $I_{cal} - A$ and $I_{mon} - A$ degrees of freedom, I_{cal} and I_{mon} are the number of calibration and monitoring samples in the subsets, and A the number of spectral factors). This procedure mimics the usual selection of factors by cross-validation[18] and avoids the inclusion of a factor-depending term into the objective function. The three selected RMSEP values were then averaged, defining the figure of merit to be minimized by the algorithm. After stopping the RIGA, the top five chromosomes were submitted to PLS model building, and the final selection was made according to the results provided by an independent test set of samples, different than those employed during training. Figure 1A summarizes RIGA in the form of its flow sheet.

The IRGA method, on the other hand, involves the following steps:

(1) Initialize the GA with 50% of all sensors, selected at random, and setting the number of factors as equal to the optimum value for full spectral cross-validation on the complete calibration data set. The initial percentage of included sensors was selected to give all sensors equal initial probability of being included for regression. Within the framework of IRGA, there is no need of selecting smaller initial percentages in order to reduce the final number of sensors, because the reduction is automatically performed.

(2) Run the GA aiming at minimizing the mean RMSEP for the three monitoring subsets, as described above.

(3) Repeat steps (1) and (2) a number of times (20 in the present work), registering a statistical histogram of the inclusion of a given sensor range in the final top chromo-

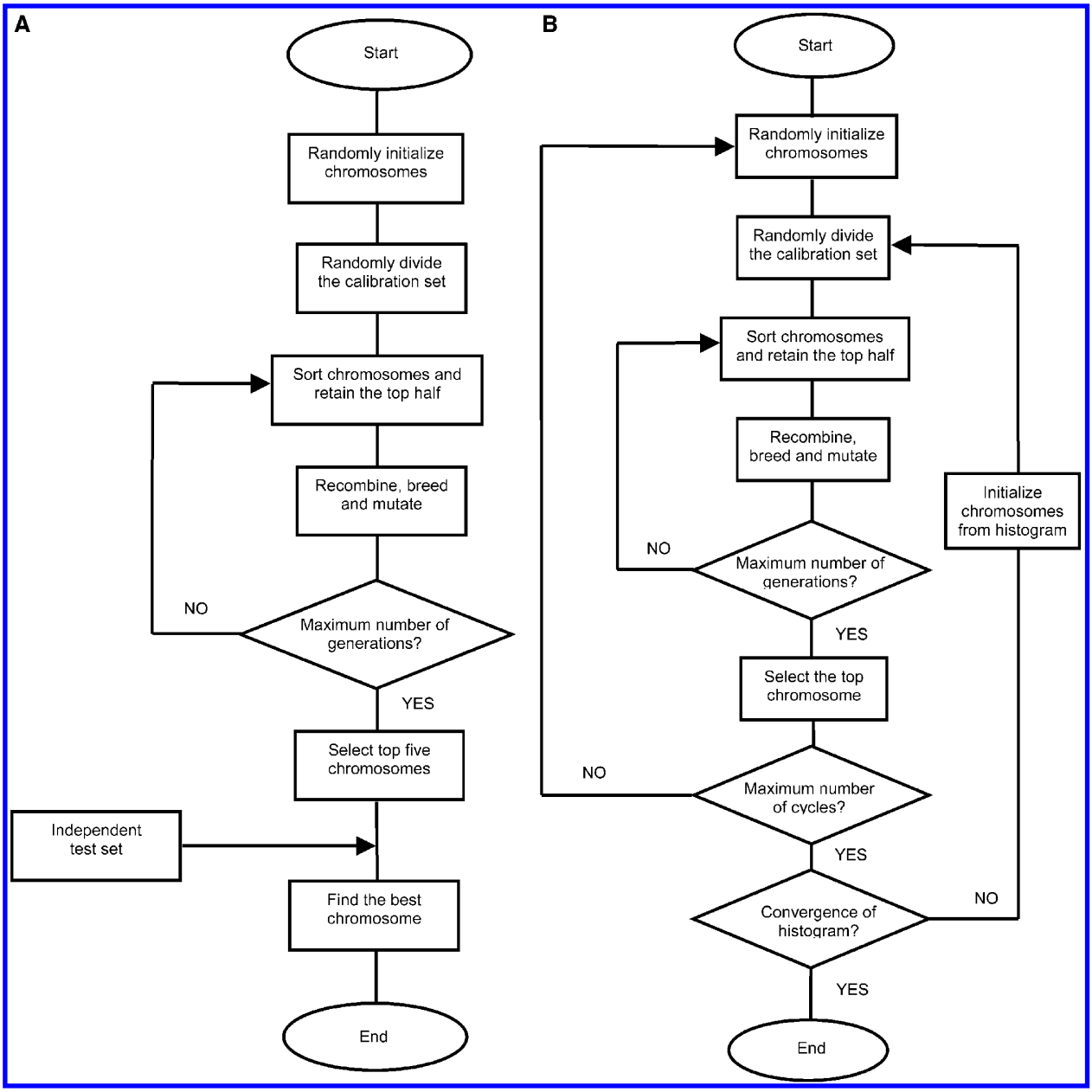

Figure 1. Flow sheets illustrating genetic algorithms for sensor selection: (A) randomly initialized genetic algorithm (RIGA) and (B) iteratively reinitialized genetic algorithm (IRGA).

some. Each time the calculation is repeated, the random partitioning of the calibration set (see above) is performed again.

(4) Select the sensor ranges included in the histogram a certain percentage of times above a threshold, for example, 70%.

(5) Perform cross-validation using the sensors obtained in step (4) in order to re-estimate the number of factors.

(6) Repeat steps (2)−(5), reinitializing the GA with the sensors obtained in step (4) and setting the maximum number of factors as obtained in step (5). Continue until the sensors selected in step (4) stabilize.

(7) Use the wavelengths selected in step (4) for PLS model building and prediction.

The flow sheet summarizing IRGA is shown in Figure 1B. Both versions of the genetic algorithm were implemented with suitable in-house routines written in Matlab 5.3,[19] which are available from the authors on request. Typical computer times ranged between 10 and 20 min for the complete iteration cycle, running on a Pentium III 550 microcomputer.

3. DATA SETS

Simulated Data Sets. Simulated data sets were constructed in the following way. Three analytes (A1, A2, and A3) are present in all samples, A1 being the analyte of interest. The pure spectra are shown in Figure 2 and are composed of Gaussian-shaped peaks: two for A1, centered at sensors 15 and 85, within a full spectral range from 1 to 150, and single Gaussians for A2 and A3, centered at sensors 45 and 55, respectively. The peak positions have been selected so that the interferents (A2 and A3) appear in the middle of the two A1 peaks, allowing for a region which is mainly dominated by noise (sensors 100−150). This will be used to test whether the GA is able to discard noisy regions as well as to select both spectral ranges where A1 is known

Table 1. Experimental Data Sets Employed in the Present Work

	data set			
	1	2	3	4
reference	21	21	22	23
type	decongestant tablets	decongestant tablets	opthalmic solutions	injectables
analyte of interest	phenylpropanolamine	diphenhydramine	dexamethasone	vitamin B12
composition of a typical unknown sample	phenylpropanolamine: 25 mg paracetamol: 500 mg	diphenhydramine: 25 mg paracetamol: 500 mg	dexamethasone: 5 mg mL^{-1} chlorpheniramine: 100 mg mL^{-1}	vitamin B12: 10 mg L^{-1} vitamin B6: 250 mg L^{-1}
calibration concn range	1.0−2.0 mg L^{-1}	1.0−2.0 mg L^{-1}	5.0−15.0 mg L^{-1}	1.0−5.0 mg L^{-1}
calibration set[a]	16	16	30	16
calibration design	four-level full factorial	four-level full factorial	two central composites	four-level full factorial
test set[a]	10	10	27	10

[a] Number of samples.

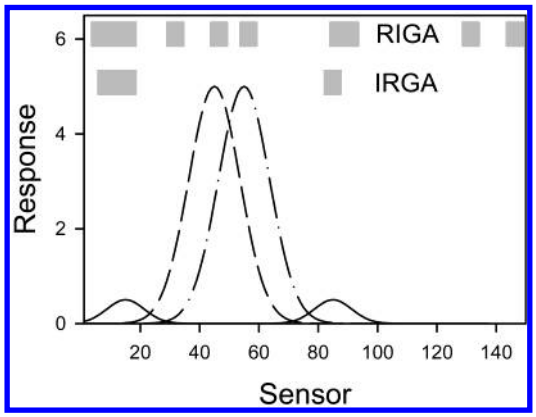

Figure 2. Theoretical spectra for the three analytes employed in the Monte Carlo simulations: (—) A1, (− − −) A2, and (−·−) A3. The concentrations are as follows: $c_1 = 0.5$, $c_{2,3} = 5$. The wavelengths selected by RIGA and IRGA are indicated at the top of the plot.

to respond. This latter point is relevant when comparing the results with methods which use extensive searches in regions of varying size, such as the so-called moving window strategy, in which the inclusion of discontinuous spectral regions is prohibitively time-consuming.

Calibration spectra were created starting from the theoretical noiseless pure spectra at unit concentration. Eight series of 15 calibration spectra were generated using central composite designs. In all cases, the concentrations for the interferents A2 and A3 ranged from 0 to 5, whereas A1 ranged between 0 and eight different maximum values (selected from 0.25 to 5). This corresponds to extreme values of 20 and 1 for the ratio ($\bar{c}_{2,3}/\bar{c}_1$), with $\bar{c}$ indicating the mean calibration concentration of a given component (i.e, $\bar{c}_1$ for A1, $\bar{c}_2$ for A2, and $\bar{c}_3$ for A3). Spectra for eight test sets of 10 samples were also produced for prediction, with random concentrations of all three analytes (within each of the calibration ranges). Additional test sets of 10 samples with random concentrations were required by one of the genetic algorithms under consideration. Linearity between response and concentration was assumed in all cases. Random numbers obtained from a Gaussian distribution with standard deviations of 0.01 and 0.1 units were added to all nominal concentrations and spectra, respectively. Using a Monte Carlo approach,[20] this calibration/prediction process was repeated 1000 times using the full spectral range and the wavelengths provided by the sensor selection procedures, to study the average RMSEP values for the test sets. All data were mean-centered as a preprocessing step.

All calculations were performed with in-house routines written in Matlab 5.3.

Experimental Data Sets. Five experimental data sets have been chosen for illustrating the performance of IRGA. In these cases, the lack of a sufficient number of available samples precluded the application of RIGA. The first four sets involve highly diluted components in pharmaceutical mixtures studied by UV−visible spectrophotometry, where PLS has been found to improve with wavelength selection.[21−23] Specific details can be found in Table 1. In these previous works, the selection was carried out after qualitative analysis of the pure component spectra[21] or by producing a three-dimensional plot of the cross-validation variance as a function of the first sensor and window width within a moving window strategy.[22,23] The examples are taken from a project aimed at developing simple spectrophotometric methods for the simultaneous determination of active principles in pharmaceuticals. To be applicable to quality control programs, prediction errors should be sufficiently small (i.e., below ca. 5%),[24] and this requires suitable wavelength working ranges in order to obtain the desired accuracy with PLS models.

The fifth set consists of 34 near-infrared (NIR) spectra of gasoline samples collected in a local distillery, in the range 4020−9996 cm^{-1} each 12 cm^{-1} (499 data points) using a Bran+Luebbe Infraprover II FT NIR spectrophotometer. The corresponding octane numbers were determined by the reference method for research octane number of spark-ignition engine fuel[25] and span the range from 90.9 to 98.2. The set was randomly divided into a 24-sample calibration and a 10-sample test set.

As with the simulated data sets, both spectra and concentrations were mean-centered.

4. COMPARISON OF RESULTS

Monte Carlo Simulations. Table 2 shows the results for the theoretical three-component example, obtained using PLS in the full spectral range (1−150). The quoted RMSEP values are averages of a large number of calibration/prediction processes (typically 1000). As can be seen in Figure 3, where the RMSEPs are plotted as a function of the ratio ($\bar{c}_{2,3}/\bar{c}_1$), the errors considerably increase for decreasing concentrations of A1, reaching a maximum of ca. 0.05 when ($\bar{c}_{2,3}/\bar{c}_1$) is about 10. This implies that the challenge posed by a progressive dilution of A1 effectively leads to an increasingly poor behavior of PLS if full spectra are employed for regression.

Before applying GA-based sensor selection, cross-validation performed on a number of simulated calibration sets

Table 2. Results Obtained When Applying Monte Carlo Simulations to the Theoretical Example in Different Spectral Ranges

	no. of factors, no. of sensors, and average RMSEP[a]								
	full range			RIGA selection			IRGA selection		
$(\bar{c}_{2,3}/\bar{c}_1)$	*A*	*J*	RMSEP	*A*	RMSEP	*J*	*A*	*J*	RMSEP
1	3	150	0.028	3	0.028	45	1	18	0.028
2	3	150	0.030	3	0.031	48	1	21	0.028
4	3	150	0.034	3	0.032	45	1	21	0.028
6	3	150	0.038	3	0.035	45	1	18	0.029
8	3	150	0.040	3	0.036	51	1	18	0.030
10	3	150	0.048	3	0.041	54	1	18	0.030
15	3	150	0.052	3	0.046	48	1	21	0.032
20	3	150	0.053	3	0.048	45	1	18	0.033

[a] *A* = number of factors estimated by cross-validation on 10 different calibration sets of samples, *J* = number of sensors, RMSEP = average root-mean-square prediction error.

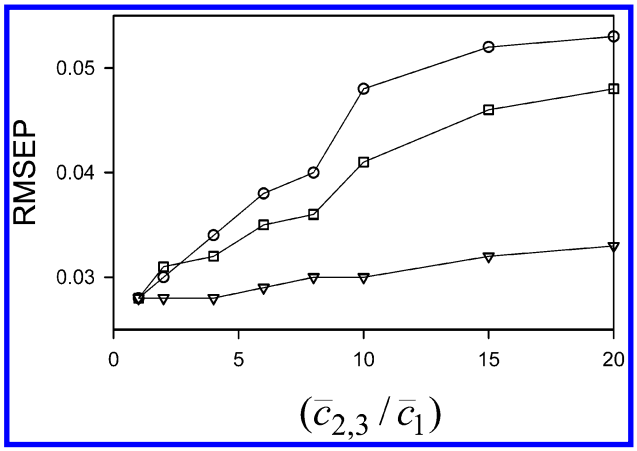

Figure 3. Average Monte Carlo RMSEP calculated for theoretical test sets of samples as a function of the concentration ratio $(\bar{c}_{2,3}/\bar{c}_1)$, using the following: circles, the full spectral range, squares, the sensors provided by RIGA, and triangles, those found by IRGA.

established that the optimum number of factors when using the full spectral 1−150 range is three (in accordance with the known number of components). Application or RIGA leads to the results shown in both Table 2 and Figure 3. The comparison with the use of full spectra indicates that PLS shows better predictive ability when RIGA sensor selection is applied, especially as the concentration of A1 decreases relative to those for A2,3. As an example, the wavelengths selected in the case $(\bar{c}_{2,3}/\bar{c}_1) = 10$ are shown at the top of Figure 2. It should be noticed that the cross-validation numbers of factors (Table 2) using the sensors suggested by the best chromosomes are still equal to three in all cases. Close examination of the structure of the top chromosomes indicates that they include wavelengths which are known to be unresponsive for A1 (and dominated either by noise or by signals from A2 and A3). Although RIGA effectively lowers the prediction error for the three monitoring subsets created from the original calibration data, it does so by including wavelengths which subsequently spoil, to a certain extent, the prediction ability on new test sets. It has been suggested that inclusion of noisy regions may be needed to correct for baseline drifts,[2] which however is not the case in the simulated experiments, showing that RIGA has a tendency to carry along useful as well as useless information from one generation to the next. In any case, PLS is seen to improve its performance after selection of wavelengths with RIGA, and lower RMSEP are obtained for the test sets using considerably less wavelengths as compared to full spectral models.

IRGA was then applied to the theoretical examples, starting again with a maximum number of factors of 3, as estimated from full spectral cross-validation. This number was then reduced to one in successive cycles, indicating that regions where only A1 responds were being retained. The final results clearly improve, as judged both from the lower RMSEP values (see Table 2 and Figure 3), in comparison with the use of both full spectra and RIGA selection. IRGA selects both regions which are sensitive to analyte A1 [see Figure 2 for the case in which $(\bar{c}_{2,3}/\bar{c}_1) = 10$], a result which cannot be obtained by comprehensive moving window searches. Furthermore, only a single latent PLS variable is required to successfully build the predictive model when IRGA is applied (Table 2), implying that appropriate sensor selection automatically leads to models requiring less spectral factors, since regions where interferents respond have been eliminated. Noisy regions are also automatically discarded.

It should be noticed that the RMSEP errors in all the simulated cases may be expected to have a lower bound given by the standard deviation in the predicted concentrations, which in PLS is approximately given by[26]

$$s_1 = [(I^{-1} + h)(s_c^2 + ||\mathbf{b}_1||^2 s_R^2) + ||\mathbf{b}_1||^2 s_R^2]^{1/2} \quad (1)$$

where s_1 is the standard deviation in predicted concentrations, I is the number of calibration samples, $\mathbf{b}_1$ is the vector of final regression coefficients calculated by the PLS model for analyte A1, s_R^2 is the variance in the analytical response, and s_c^2 is the concentration variance, and h is the sample leverage, which gives the sample position in the calibration space, and is defined by

$$h = ||\mathbf{R}^{+\mathrm{T}}\mathbf{r}||^2 \quad (2)$$

In eq 2, **R** is the calibration data matrix, **r** is the sample spectrum, the superscprit "+" indicates pseudoiverse, and $||\cdot||$ is the Euclidean norm of a vector. Inserting the appropriate values for the full spectral range 1−150, eq 1 gives an average value of s_1 of 0.03 for all theoretical cases studied. As can be seen in Table 2 and Figure 3 in the case of low concentrations of A1, only when wavelength regions are appropriately selected with IRGA do the average RMSEP values approach this limit.

An example of the progression of sensor inclusion statistics with repeating IRGA cycles is displayed in Figure 4A−D, for the case in which $(\bar{c}_{2,3}/\bar{c}_1) = 10$. From the initial random distribution of wavelengths, the histogram obtained after the first IRGA cycle (Figure 4A) clearly contains hints that sensor ranges near the maxima for A1 (located at sensors 15 and 85) are included more frequently than unresponsive regions. However, wavelengths dominated either by interferents or noise are also added. When the IRGA steps outlined above are followed, successive histograms show that the useful regions near sensors 15 and 85 are persistently included. Moreover, the shapes of the histograms are similar after three and four cycles. IRGA may be stopped when the sensors included a certain percent of times above a certain threshold, for example, 70%, stabilize. For the particular example examined in Figure 4, i.e., when the concentration

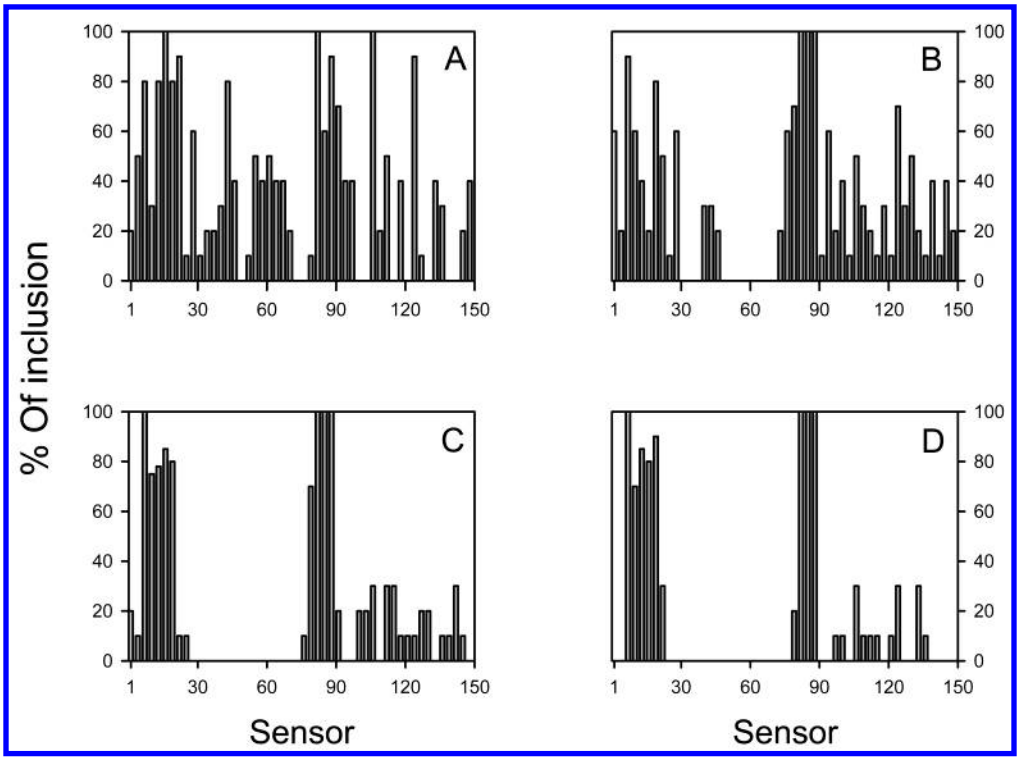

Figure 4. (A)−(D) Histograms representing the percent of inclusion of chromosome-encoded ranges of three sensors for a theoretical example, after 1−4 IRGA cycles, respectively, when the concentration ratio of the major components to the minor one is 10.

ratio of A1 with respect to A2,3 is 1:10, this takes place after four cycles (implying that the sensor ranges 7−19 and 82−88 should be selected for PLS regression). This procedure includes a minimal set of sensor ranges which are useful in predicting new samples, automatically excluding unresponsive wavelength regions without the requirement of producing an independent test sample set for chromosome selection. Further, it gives IRGA additional flexibility to those which are intrinsic to RIGA, since the proportion of sensors included in the histogram to be retained for prediction may be considered as an additional tunable parameter.

It is apparent that the higher probability of a given sensor (or group of sensors) to be included in the top chromosome after a number of generations have passed is a measure of its relevance for prediction. A high probability that a given gene encoding a sensor is selected in the top chromosome is associated with a model displaying good predictive ability toward future samples. It should be stressed that all simulations were produced by assuming linearity between response and concentration, indicating that sensor selection is also an important previous step to multivariate calibration, even when linearity is strictly fulfilled.

Experimental Examples. In the four pharmaceutical examples chosen to illustrate IRGA, suitable calibration sets of samples were employed, with the concentration ranges for the analytes of interest given in Table 1. The mean calibration concentrations for all components were selected according to their relative proportions in the pharmaceutical samples.[21−23] In all cases, test sets of samples were available, with concentrations of all analytes different than those used for calibration (but within the corresponding calibration ranges). Figure 5 shows the spectra for all pure components, in relative concentrations which are typical of real samples. As can be seen, the minor constituents present a severe challenge to multivariate regression techniques, as judged both from the relative intensities of the absorption bands and the high spectral overlap.

Table 3. Results Obtained on Test Sets for the Experimental Examples Using the Full Spectral Ranges and Those Provided by Sensor Selection with IRGA

	data set			
parameter[a]	1	2	3	4
	No Sensor Selection			
spectral range/nm	205−300	205−300	200−350	215−365
no. of factors	6	8	6	5
RMSEP/mg L^{-1}	0.30	0.13	1.80	0.46
REP/%	20	8.5	18	15
	Sensor Selection with IRGA			
spectral range/nm	211−220	211−225	241−260	346−360
no. of factors	2	3	3	3
RMSEP/mg L^{-1}	0.06	0.06	0.18	0.13
REP/%	4.0	4.0	1.8	4.3

[a] RMSEP = root-mean-square prediction error, REP = relative error of prediction = RMSEP × 100/$\bar{c}$.

Sensor selection is critical in order to achieve prediction errors below ca. 5% in the test set of samples. Indeed, Table 3 shows that the use of the full spectral ranges leads to unsuitable prediction errors, which are as large as 20% for data set no. 1. Furthermore, the number of optimum factors estimated by cross-validation is significantly larger than the known number of components in these systems of rather simple composition. Wavelength regions selected by IRGA for quantitating the minor constituents are shown in Table 3 and also pictorially in Figure 5. In all cases IRGA selects the regions where not only the analyte of interest absorbs but also where overlapping with the major component peaks is minimal (this is apparent in Figure 5A,C). The number of optimum factors employed to construct the models decreases to a value which is reasonably compatible with the known number of sample components (Table 3). More importantly,

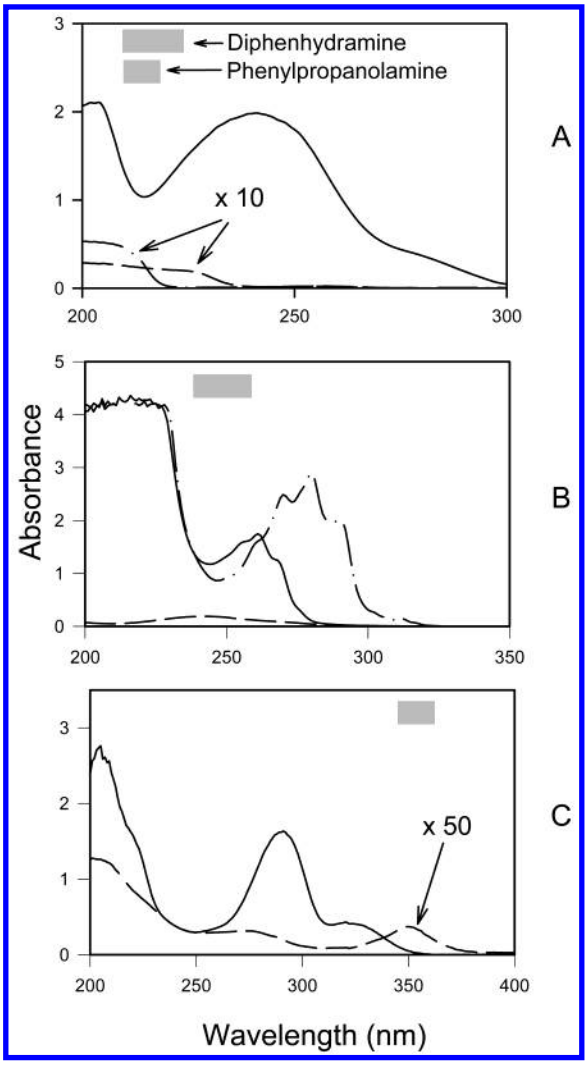

Figure 5. Spectra for solutions of the pure analytes in the studied experimental samples, in relative concentrations which are typical of real samples: (A) (—) paracetamol, 38.0 mg mL^{-1}, (– – –) diphenhydramine, 2.0 mg mL^{-1}, and (–·–) phenylpropanolamine, 2.0 mg mL^{-1}; (B) (—) chlorpheniramine, 100.0 mg mL^{-1}, (– – –) dexamethasone-21-phosphate, 5.0 mg mL^{-1}, and (–·–) naphazoline, 100.0 mg mL^{-1}; (C) (—) vitamin B12, 2.0 mg mL^{-1}, and (– – –) vitamin B6, 50.0 mg mL^{-1}. For clarity, the spectra for diphenhydramine and phenylpropanolamine in (A) have been increased 10 times and that for vitamin B12 in (C) has been increased 50 times. The wavelengths selected by IRGA for predicting the minor analytes are shown at the top of each plot.

however, all RMSEP values are now below 5%, indicating that spectrophotometry/multivariate calibration, if coupled to adequate sensor selection, is feasible for quality control of the studied pharmaceutical forms.

In the case of data set no. 5, used to model octane number in gasolines, the corresponding calibration NIR spectra are shown in Figure 6. As has been previously reported, NIR spectroscopy is sensitive to octane number through hydrocarbon vibrational overtones and combination bands.[27−29] The wavenumbers selected by IRGA (see Figure 6) are near those recently found to be useful for octane number prediction by ridge regression analysis, i.e., 5520 and 8400 cm^{-1}.[29] Application of PLS to our calibration set using the full spectral range (4020−9996 cm^{-1}) with eight latent factors (as selected by cross-validation) yields an RMSEP of 0.9 units on the independent 10 sample test set. In comparison, the use of IRGA selected sensors leads to an RMSEP of 0.3

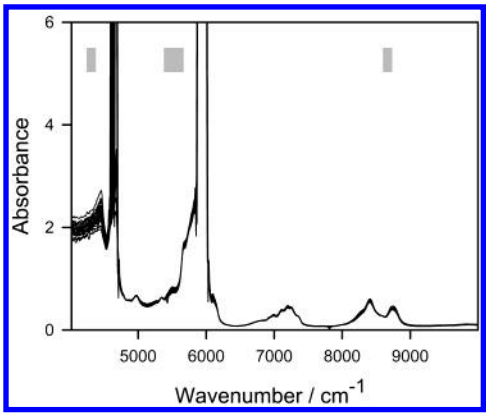

Figure 6. NIR spectra of the 24 gasoline samples employed for calibrating a PLS model for predicting the octane number. The sensors selected by IRGA are shown at the top.

units, with a considerable reduction in the optimum number of PLS factors (which cross-validation estimated as 2). It should be noticed that the reference method for research octane number determination has an average error of ca. 0.3 units.[25] This example illustrates how IRGA allows one to select appropriate sensors in a case where spectral regions which are responsive to the target parameter are not known a priori, opening the possibility for an automatic variable selection procedure. In this case, the successful combination of NIR, multivariate calibration, and sensor selection may allow the replacement of expensive and time-consuming techniques by simple and rapid spectroscopic methods for routine industrial applications.

5. CONCLUSIONS

Genetic algorithms with different initialization procedures were discussed as regards the selection of variables to be employed in multivariate calibration techniques such as partial least-squares regression. Common initialization strategies were overviewed, and a new, iterative method was presented which does not require the use of independent test sample sets. Theoretical examples and Monte Carlo simulations have been employed in order to illustrate the benefits of sensor selection in achieving lower prediction errors toward new test samples. Experimental examples taken from a pharmaceutical quality control program and data from a local distillery concerning the determination of octane number in gasolines have been used to show the performance of the discussed techniques in connection with real problems.

ACKNOWLEDGMENT

Financial support from CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Universidad Nacional de Rosario, Agencia Nacional de Promoción Científica y Tecnológica (Project PICT99 No. 06-06078), and Universidad Nacional del Litoral (Project CAI+D 17-1-41) is gratefully acknowledged. A.C.O. is a fellow of the John Simon Guggenheim Memorial Foundation (2001−2002).

REFERENCES AND NOTES

(1) Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, 1989.
(2) Bangalore, A. S.; Shaffer, R. E.; Small, G. W.; Arnold, M. A. Genetic Algorithm-Based Method for Selecting Wavelengths and Model Size for Use with Partial Least-Squares Regression: Application to Near-Infrared Spectroscopy. *Anal. Chem*. **1996**, *68*, 4200−4212.

(3) McShane, M. J.; Cote, G. L.; Spiegelman, C. H. Variable Selection in Multivariate Calibration of a Spectroscopic Glucose Sensor. *Appl. Spectrosc*. **1997**, *51*, 1559−1564.
(4) Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L. Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm. *Anal. Chem*. **1998**, *70*, 35−44.
(5) Thomson, M. L. Selection of Variables in Multiple Regression. *Int. Stat. Rev*. **1978**, *46*, 1−19.
(6) Collado, M. S.; Mantovani, V. E.; Goicoechea, H. C.; Olivieri, A. C. Simultaneous Spectrophotometric-Multivariate Calibration Determination of Several Components of Ophthalmic Solutions: Phenylephrine, Chloramphenicol, Antipyrine, Methylparaben and Thimerosal. *Talanta* **2000**, *52*, 909−920.
(7) Goicoechea, H. C.; Olivieri, A. C. Wavelength Selection by Net Analyte Signals Calculated with the Multivariate Factor-Based Hybrid Linear Analysis (HLA). A Theoretical and Experimental Comparison with Partial Least-Squares (PLS). *Analyst* **1999**, *124*, 725−731.
(8) Ding, Q.; Small, G. W.; Arnold, M. A. Genetic Algorithm-Based Wavelength Selection for the Near-Infrared Determination of Glucose in Biological Matrixes: Initialization Strategies and Effects of Spectral Resolution. *Anal. Chem*. **1998**, *70*, 4472−4479.
(9) Swierenga, H.; Wülfert, F.; de Noord, O. E.; de Weijer, A. P.; Smilde, A. K.; Buydens, L. M. C. Development of robust calibration models in near infrared spectrometric applications. *Anal. Chim. Acta* **2000**, *411*, 121−135.
(10) Todeschini, R.; Galvani, D.; Vilchez, J. L.; del Olmo, M.; Navas, N. Kohonen Artificial Neural Networks as a Tool for Wavelength Selection in Multicomponent Spectrofluorimetric PLS Modelling: Application to Phenol, *o*-Cresol, *m*-Cresol and *p*-Cresol Mixtures. *Trends Anal. Chem*. **1999**, *18*, 93−98.
(11) Leardi, R. Genetic Algorithms in Chemometrics and Chemistry: a Review. *J. Chemom*. **2001**, *15*, 559−569.
(12) Ugulino Araújo, M. C.; Bezerra Saldanha, T. C.; Harrop Galvão, R. K.; Yoneyama, T.; Caldas Chame, H.; Visani, V. The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis. *Chemom. Intell. Lab. Syst*. **2001**, *57*, 65−73.
(13) Ferré, J.; Rius, F. X. Detection and Correction of Biased Results of Individual Analytes in Multicomponent Spectroscopic Analysis. *Anal. Chem*. **1998**, *70*, 1999−2007.
(14) Davis, L. *Genetic Algorithm and Simulated Annealing*; Pitman: London, 1987.
(15) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1988.
(16) Shaffer, R. E.; Small, G. W. Learning Optimization from Nature. Genetic Algorithms and Simulated Annealing. *Anal. Chem*. **1997**, *69*, 236A-242A.
(17) Kemsley, E. K. A Genetic Algorithm (GA) Approach to the Calculation of Canonical Variates. *Trends. Anal. Chem*. **1998**, *17*, 24−34.
(18) Haaland, D. M.; Thomas, E. V. Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Anal. Chem*. **1988**, *60*, 1193−1202.
(19) MATLAB 5.3, The MathWorks Inc.; Natick, Massachusetts, U.S.A., 1999.
(20) Meier, P. C.; Zund, E. *Statistical Methods in Analytical Chemistry*; John Wiley & Sons: New York, 1993; pp 145−150.
(21) Goicoechea, H. C.; Olivieri, A. C. Simultaneous Multivariate Spectrophotometric Analysis of Paracetamol and Minor Components (Diphenhydramine of Phenylpropanolamine) in Tablet Preparations. *J. Pharm. Biomed. Anal*. **1999**, *20*, 255−261.
(22) Goicoechea, H. C.; Collado, M. S.; Satuf, M. L.; Complementary use of Partial Least-Squares and Artificial Neural Networks for the non-Linear Spectrophotometric Analysis of Pharmaceutical Samples. Olivieri, A. C. *Anal. Bioanal. Chem.* In press.
(23) Damiani, P. C.; Nepote, J. A.; Olivieri, A. C. Chemometrics Assisted Spectroscopic Determination of Vitamin B6, Vitamin B12 and Dexamethasone in Injectables *J. Pharm. Biomed. Anal*. Submitted for publication.
(24) Green, J. M. A practical guide to analytical method validation. *Anal. Chem*. **1996**, *68*, 305A−309A.
(25) ASTM Method D 2699-99, *Annual Book of ASTM Standards*; ASTM: West Conshohocken, PA, Vol. 05.05.
(26) Faber, K.; Kowalski, B. R. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J. Chemom*. **1997**, *11*, 181−238.
(27) Kelly, J. J.; Barlow, C. H.; Jinguji, T. M.; Callis, J. B. Prediction of Gasoline Octane Number from Near-Infrared Spectral Features in the Range 660−1215 nm. *Anal. Chem*. **1989**, *61*, 313−320.
(28) Bohács, G.; Ovádi, Z.; Salgó, A. Prediction of Gasoline Properties with Near Infrared Spectroscopy. *J. Near Infrared Spectrosc*. **1998**, *6*, 341−348.
(29) Chung, H.; Lee, H.; Jun, C.-H. Determination of Research Octane Number using NIR Spectral Data and Ridge Regression. *Bull. Korean Chem. Soc*. **2001**, *22*, 37−42.

CI0255228