Pharmaceutical Research (© 2007)
DOI: 10.1007/s11095-007-9260-1

Research Paper

Molecular Factor Computing for Predictive Spectroscopy

Bin Dai,[1] Aaron Urbas,[1] Craig C. Douglas,[2] and Robert A. Lodder[3,4]

Received November 13, 2006; accepted February 1, 2007

Purpose. The concept of molecular factor computing (MFC)-based predictive spectroscopy was demonstrated here with quantitative analysis of ethanol-in-water mixtures in a MFC-based prototype instrument.

Methods. Molecular computing of vectors for transformation matrices enabled spectra to be represented in a desired coordinate system. New coordinate systems were selected to reduce the dimensionality of the spectral hyperspace and simplify the mechanical/electrical/computational construction of a new MFC spectrometer employing transmission MFC filters. A library search algorithm was developed to calculate the chemical constituents of the MFC filters. The prototype instrument was used to collect data from 39 ethanol-in-water mixtures (range 0–14%). For each sample, four different voltage outputs from the detector (forming two factor scores) were measured by using four different MFC filters. Twenty samples were used to calibrate the instrument and build a multivariate linear regression prediction model, and the remaining samples were used to validate the predictive ability of the model.

Results. In engineering simulations, four MFC filters gave an adequate calibration model ($r^2 = 0.995$, RMSEC = 0.229%, RMSECV = 0.339%, $p = 0.05$ by f test). This result is slightly better than a corresponding PCR calibration model based on corrected transmission spectra ($r^2 = 0.993$, RMSEC = 0.359%, RMSECV = 0.551%, $p = 0.05$ by f test). The first actual MFC prototype gave an RMSECV = 0.735%.

Conclusion. MFC was a viable alternative to conventional spectrometry with the potential to be more simply implemented and more rapid and accurate.

KEY WORDS: chemometrics; genetic algorithm; multivariate analysis; near infrared spectroscopy (NIR); optical computing.

INTRODUCTION

Near infrared spectroscopy (NIR) has become an important process analytical method for simultaneous multicomponent chemical analysis. NIR has found many applications in process environments and in measurements in the biotechnology and pharmaceutical industries (1–6), where NIR spectroscopy provides online, nondestructive and noninvasive sensing. In September of 2004, the US FDA released a Guidance for Industry, PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance (7,8). This guidance is designed to facilitate innovation in, process development and quality assurance. Process Analytical Technology (PAT) will help in better design, monitor and control of pharmaceutical manufacturing process by integrating multivariate modeling, sensors design and process optimization with the goal of ensuring final product quality (9).

Industrial environments are usually less friendly to analytical instrumentation than research laboratories. Filter instruments are usually much more stable and rugged than their dispersive or interferometric counterparts, making them ideally suited for the harsh conditions found in industrial environments (10,11).

Multivariate calibration is a well-established tool in chemometrics for analysis of NIR, UV-Visible, and Raman spectra. Conventional measurement of chemical or physical properties from spectra is carried out by constructing a predictive model (12–14). Two of the most commonly used methods to construct a predictive model are partial least squares (PLS) and principal component regression (PCR). In a conventional spectrometer with typical chemometrics, data collection and processing of raw data can be time consuming and computationally expensive, especially when spatial relationships (image data) are required. Methods for selecting small but highly relevant variables to represent the original data in a reduced coordinate space and methods for integrated sensing and processing (ISP) are therefore receiving much attention (14,15).

ISP aims to design and optimize sensing systems that integrate the traditionally independent units of sensing,

[1] Department of Chemistry, University of Kentucky, Lexington, USA.
[2] Department of Computer Science, University of Kentucky, Lexington, USA.
[3] Department of Pharmaceutical Sciences, University of Kentucky, Lexington, Kentucky 40506-0055, USA.
[4] To whom correspondence should be addressed. (e-mail: Lodder@uky.edu)

signal processing, communication and targeting. By employing ISP, computational complexity within traditional sensing system has been substantially reduced through determining efficient low-dimensional representations of those sensing problems that were originally posed in high-dimensional settings by traditional sensing architecture. Successful ISP is expected to yield entirely new ways of designing and operating sensor systems (16).

One approach currently being investigated to simplify both instrumentation and computational analysis involves optical pattern encoding (17). This technique involves tailoring the optical spectrum of filters to encode high level information about the samples in sensing stage. Theoretical treatment of this methodology can be found in the literature (18,19). Myrick *et al*. have demonstrated some practical applications of this methodology in UV-visible and NIR spectroscopy (20–28). Encoding applications are based on the fabrication of thin film solid-state optical filters, termed multivariate optical elements (MOEs). MOEs are designed to replicate the multivariate regression pattern by transmitting and reflecting weighted optical signals over a broad wavelength band.

Recent publications from our laboratory have offered an alternative approach for spectral encoding (29,30). Molecular absorption filters can be used as mathematical factors in spectral encoding to generate a factor-analytic optical calibration in a high-throughput spectrometer, which we term molecular factor computing (MFC). The molecules in the filter effectively compute the calibration function by weighting the signals received at each wavelength over a broad range of wavelengths (see Fig. 1). Given a set of training spectra collected at all available wavelengths (see Fig. 1 left), it is possible to rationally select molecular filter materials to perform a factor analysis procedure like principal component analysis (PCA) (see Fig. 1 right). PCA is designed to maximize the signals from the spectral regions with the most variability by most heavily weighting them (loadings line in Fig. 1 left). However, PC loadings heavily weight wavelength information in the positive and negative direction, which is difficult to implement optically with molecular absorbance filters. MFC uses absolute values in MFs, so two filters are required for each PC, one for the positive loadings (MF1) and one for the negative loadings (MF2). The filter materials are selected by examining their spectra. The transmission spectrum (%T) of the filter material should be as similar as possible to the absolute value of the loadings spectrum being targeted. Bandpass filters are selected to ignore regions of the spectrum where there is no difference between the training spectra, as extra photons in those regions simply saturate the detector or add noise without providing any additional signal. The MF filters do not have to be featureless in the areas away from their peaks in Fig. 1 as long as bandpass filters (or prisms or gratings) are used to wipe out the transmission peaks in undesired areas.

One or more molecular filters are used in an MFC-based spectrometer to produce detector signals correlated to desired sample information. Advantages of this new approach over conventional spectroscopy include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost.

This report describes the instrumentation and application of a molecular factor computing-based spectrometer. Such a spectrometer may be particularly useful in applications where real-time video analyses of remote sensing data are required. In such cases, molecular filters placed in front of near-IR cameras would produce images in which the intensities were proportional to the factor scores, without the need for additional computation. Ethanol in water mixtures were selected as training and validation samples to design molecular filters that would test the concept of MFC-based spectroscopy. Ethanol is used in liquid pharmaceuticals to enhance solubility, for example. Ethanol is also sometimes abused in the general population. Sensing alcohol in the environment is necessary in such an application to evaluate the effectiveness of pharmacotherapy or other therapies for alcohol abuse.

MATERIALS AND METHODS

Traditional NIR Training Spectra Collection. The ethanol was reagent grade, obtained from AAPER (Shelbyville,

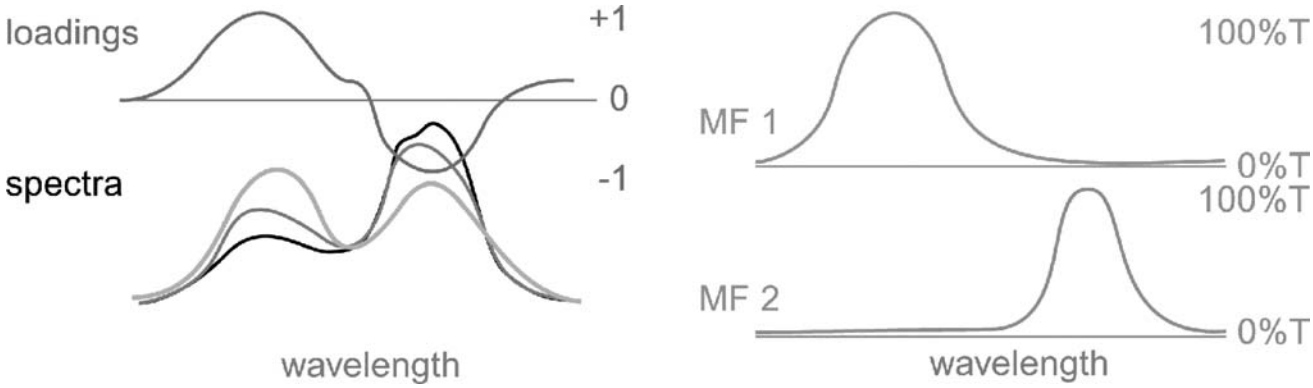

Fig. 1. A set of training spectra collected over multiple wavelengths yields factor loadings that are greatest where the spectra have greatest variability (*left*). MFC rationally selects molecular filter materials to match the factor loadings and perform a factor analysis procedure like principal component analysis in the the spectrometer (*right*).

KY). The water was distilled in house. Quantitative mixtures of alcohol and water were prepared by volume using grade-A volumetric flasks and burettes. Twenty samples were prepared with ethanol concentrations ranging from 0 to 14%. Sample solutions were placed in a quartz cuvette with a path length of 1 mm. Conventional near-infrared transmission spectra for comparison with MFC were collected using a dispersive spectrometer (Ocean Optics NIR256 temperature-regulated NIR, Dunedin, FL) over the wavelength range of 900–2,500 nm to acquire a total of 256 data points per spectrum. Data analysis was limited to 1,400–2,200 nm to avoid the short wavelength region, which was not used for the MFC tests. As a result, the selected transmission spectra included 128 data points. The sample temperature remained constant at 25°C during the data collection period. Each recorded spectrum was the average of ten scans, with the total integration time ca. one second. The transmission spectra are shown in Fig. 2a. To calculate the required composition of MFC filters for ethanol determinations, full spectra of ethanol/water mixtures over the wavelength range of interest must be available. For maximum accuracy, these spectra must represent the optical characteristics of the MFC spectrometer, not the conventional instrument. As a result, the transmission spectra of the dispersive spectrometer were convolved with the transmission spectra of a 1,400-nm long pass filter, the emission spectrum of the tungsten NIR source, and the response curve of the InGaAs photodiode in the prototype instrument to give a corrected representation of the MFC instrument response. These corrected transmission spectra were used as training spectra for MFC filters selection and multivariate analysis. The corrected spectra are presented in Fig. 2b.

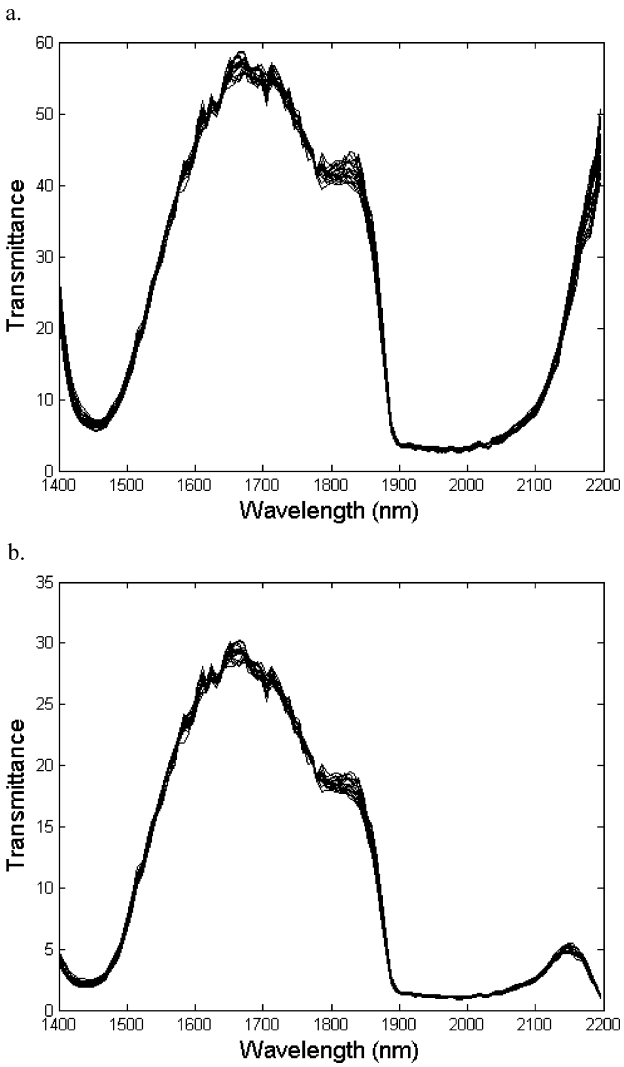

Fig. 2. (**a**) Raw, uncorrected transmission spectra of 20 ethanol / water mixtures acquired on a conventional dispersive NIR spectrometer. (**b**) Corrected spectral response function. These data are based on the transmission spectra in Fig. 2a, convolved with following radiometric vectors: radiance spectrum of tungsten light source, the transmission spectrum of 1,400 nm long pass filter, and the response curve of the InGaAs photodiode.

MFC-based High Throughput NIR Spectrometer. A graphic representation of the instrumental setup is given in Fig. 3. A 12 V, 100 W tungsten-halogen broadband source (model 621, McPherson Inc., Chelmsford, MA) with 1,400-nm long pass filter (Thorlabs, Newton, NJ) was used as the source of broadband NIR light. The tungsten-halogen light source has more intense radiation in the shorter NIR wavelength region. To avoid saturating the detector with short wavelength NIR radiation that contains little chemical information about the samples, the 1,400 nm long pass filter was used to block the short wavelength radiation. The source beam was modulated with an optical chopper (Model SR540, Stanford Research Systems Inc., Sunnyvale, CA) at a frequency of 280 Hz. The light beam was focused onto an InGaAs photodiode (Fermionics Opto-Technology, Simi Valley, CA) through a convex lens after passing through the molecular filter cuvette and sample cuvette. A step-indexed sliding cuvette tray was constructed in-house that permitted manual selection of cuvettes in the beam path. All cuvettes used for holding the liquid MFC filters were 2 mm path length optical glass. The sample cuvette had 1 mm path length. A two-factor spectrum from a sample consisted of four data points because the positive and negative factor loadings were represented by separate molecular filter mixtures. Thirty-nine ethanol-in-water mixtures were scanned with the MFC-based spectrometer. Twenty samples were used to calibrate the instrument and build a multivariate linear regression prediction model, and the remaining samples were used to validate the predictive ability of the model. To avoid possible false responses due to instrument drift, samples were measured in a random order. The sample temperature was held constant at 25°C during the data collection period. A 3-s integration was employed at each MFC filter.

Data Analysis. All data analysis was carried out using Matlab 7.0 (Mathworks, Inc., Natick, MA). The PLS toolbox v3.51 for Matlab (Eigenvector Research, Inc. Wenatchee, WA) was used for multivariate analysis. A genetic algorithm and direct search toolbox for Matlab were used to perform the NIR library search to generate combinations of liquids for use as MFC filters.

Theory. As illustrated in Fig. 3, using the MFC approach, traditional bulky multi-channel wavelength selection devices such as gratings and moving mirrors are replaced with simple MFC filters. Only a light source, detector and MFC filters are needed to construct a minimal MFC-based spectrometer. The weighted combination of spectral res-

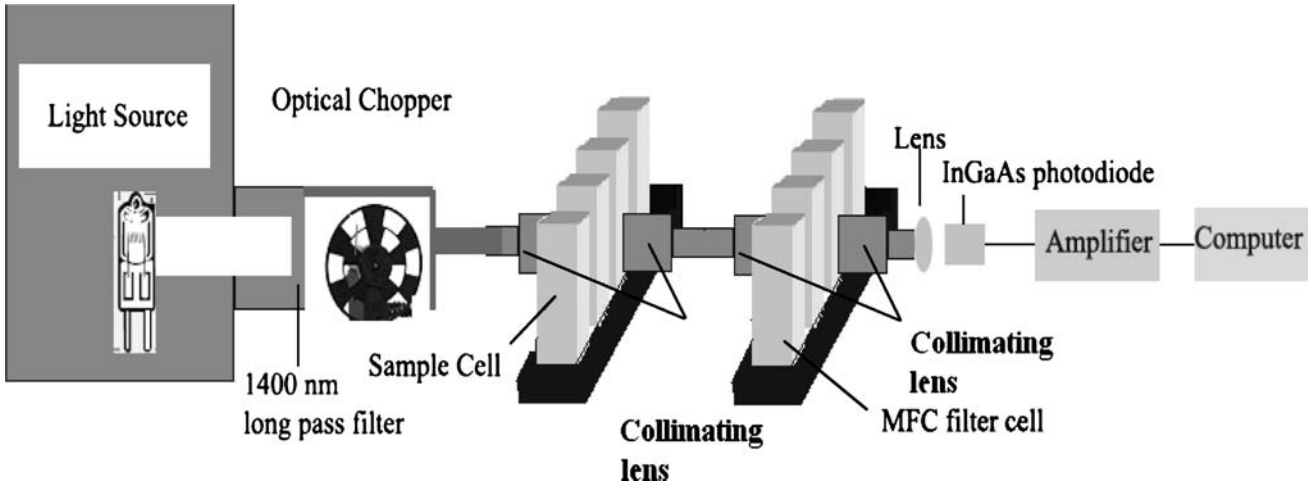

Fig. 3. A graphical representation of the MFC-based high throughput spectrometer.

ponses from the filters is designed to match the regression vector from transmission spectra-based factor methods like PCR or PLS calibration. Because a multivariate regression vector can be positive or negative while all transmission spectra of MFC filters are positive, two distinct MFC filters are employed to represent accurately the multivariate regression vector. Depending on the complexity of the regression vector and availability of MFC filter materials, an exact match of the regression vector to the filter might be very difficult. Fortunately, an exact match is not absolutely necessary, for reasons that are addressed in MFC filters selection. For each MFC filter, the signal produced at the detector is a dot product of the filter transmission spectrum and the sample transmission spectrum, with a signal offset $\boldsymbol{v}_{Offset}$ in practice (24).

$$\boldsymbol{v}_{out} = G \times \vec{s} \cdot \vec{f} + v_{offset} \tag{1}$$

$\boldsymbol{v}_{out}$ is the output voltage, $\boldsymbol{G}$ is the constant amplifier gain, , $\boldsymbol{f}$ represents the MFC filter spectrum vector, and $\boldsymbol{s}$ represents the corrected sample spectrum vector.

For m samples and n filters, $\boldsymbol{V}_{out}$ (m by n) is output voltage matrix.

$$V_{out} = G \times SF^T + V_{offset} \tag{2}$$

where $\boldsymbol{F}$ (n by k) is the transmission spectra matrix of MFC filters, and $\boldsymbol{S}$ (m by k) is transmission spectra matrix of samples.

The vector of concentration values, $\boldsymbol{Y}$ (m by *1*), of the training samples are predicted by multivariate linear regression (MLR) according to Eq. 3:

$$\hat{Y} = V_{out}C + E = G \times SF^T C + Offset \tag{3}$$

where $\boldsymbol{C}$ (n by *1*) are the regression coefficients, $\boldsymbol{E}$ is a scalar, m is the number of training samples, and n is the number of MFC filters.

After MFC filters were selected and the regression coefficients $\boldsymbol{R}$ obtained,

$$R = F^T C \tag{4}$$

this $\boldsymbol{R}$ works in a similar fashion to PCR loadings.

$$\hat{y}_i = G \times S_i F^T C + offset = G \times S_i R + offset \tag{5}$$

For m training samples, the root-mean-square error of calibration (RMSEC) (27) is

$$RMSEC = \left[\sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{m}\right]^{1/2} = \left[\sum_{i=1}^{m} \frac{(G \times S_i R + offset - y_i)^2}{m}\right]^{1/2} \tag{6}$$

The minimum RMSEC is reached by searching a NIR spectral library to select the best molecules for MFC filters. $\boldsymbol{G}$ and ***offset*** are the parameters adjusted after the MFC filters have been chosen. While one could select MFC filter molecules to match a regression vector that provides a fixed RMSEC specified a priori, searching the NIR library to find a combination of MFC filters that minimizes the RMSEC is usually more desirable. A perfect spectral match may require a large number of different filters molecules or filter molecules that are not available in the library.

Spectral Region Selection. Theoretically, the MFC-based spectrometer approach should function in any spectral region where molecular filters are available. For this research, the NIR spectral region was used because NIR spectrometry is a widely employed PAT and ethanol has a significant absorbance between two water absorbance bands in the NIR region between 1,400 and 2,200 nm.

Radiometric Correction. The multivariate prediction of analyte concentration using MFC is inherently radiometric in nature. Radiometric measurement is based on a detector response that is directly related to sample transmission instead of absorbance. Of course, sample concentration is linearly related to absorbance when Beer's law holds and transmission is logarithmically related to sample concentration. In a low absorbance regime, transmission relates to concentration approximately linearly, however, in a higher absorbance regime, the nonlinear relationship between concentration and transmission predominates. To model both regimes in transmission mode, extra principal components or latent variables have to be used in a linear multivariate calibration model (31).

Before using transmission spectra to perform a library search for MFC filter constituent selection, the transmission spectra were corrected for unique optical characteristics of the MFC spectrometer. Data provided by manufacturers' test sheets were used to form the correction factors. In the experimental MFC system, the radiometric correction was performed by convolving the transmission spectra with the emission spectrum of the source lamp, the transmission spectra of the 1,400 nm long pass filter, and the response curve of the InGaAs photodiode in the prototype instrument. Thus, the corrected transmission spectra represented an unbiased detector response as a function of wavelength. The corrected spectra in Fig. 2b revealed that the transmission of the spectrometer is not completely cut off at 1,400 nm. The transmission of the actual 1,400-nm long pass filter employed was approximately 25% at 1,400 nm. However, the transmission was much lower at shorter wavelengths and was less than 1% at 1,370 nm. Because the variation of the sample spectra from 1,370 to 1,400 nm was small, the effects of the slightly wider bandpass on prediction of sample composition were negligible.

MFC Filter Selection. The chemicals chosen as MFC filters were found by searching a library of near-IR transmission spectra containing 1,923 compounds (Wiley). The library consisted of two spectra of each compound collected over slightly overlapping regions, 952–1,587 nm and 1,388–2,630 nm. Because the coverage of the MFC system is 1,400–2,200 nm, only the spectra from 1,388–2,630 nm were used in the library search. Molecular factor scores were calculated from the product of the transmission spectra from the NIR spectral library and the corrected transmission spectra of ethanol / water mixtures:

$$U_{m\times l} = S_{m\times k} L^{T}_{l\times k} \tag{7}$$

where $\boldsymbol{U}$ is the score matrix, $\boldsymbol{L}$ is the transmission spectra of the NIR library, $\boldsymbol{S}$ is the corrected transmission spectra of training samples, l is number of compounds in the library ($l = 1{,}923$), m is number of training spectra ($m = 20$), and k is the number of wavelength values in the spectra ($k = 117$). A modified genetic algorithm (32) was used to search the score space to find four filters that yielded a predictive model with the lowest root mean square error of cross validation (RMSECV). The RMSECV function was used as the fitness function of the genetic algorithm. A genetic algorithm (GA) is a search procedure employed in computing to find actual or approximate solutions to optimization and search problems. Genetic algorithms are classified as global search heuristics. Genetic algorithms form a specific class of evolutionary algorithms that are based on methods motivated by evolutionary biology such as inheritance, mutation, selection, and crossover (also termed recombination). Genetic algorithms are executed as a computer simulation in which a population of conceptual symbols (termed chromosomes, or the genotype or the genome) of possible solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves in the direction of superior solutions. Usually, solutions are symbolized in binary, but other symbol encodings are also feasible. The evolution usually begins from a population of randomly generated individuals and occurs in generations. In each generation, the fitness of each individual in the population is assessed, multiple individuals are randomly selected from the existing population (based on their fitness), and adapted (recombined and perhaps mutated) to create a new population. The new population is then employed in the subsequent iteration of the algorithm. A typical genetic algorithm needs two items to be specified: (a) a genetic representation of the solution domain, and (b) a fitness function to assess the solution domain. A fitness function is a specific form of objective function that quantifies the optimality of a solution (i.e., a chromosome) in a genetic algorithm in order that that individual chromosome may be ranked against every one of the other chromosomes. Optimal chromosomes, or at least chromosomes that are more optimal, are permitted to breed and combine their datasets by numerous techniques, leading to a new generation that will (with luck) be improved. An ideal fitness function connects closely with the algorithm's aim, and still can be computed rapidly. Speed of calculation is vital, because a conventional genetic algorithm must be iterated lots of times in order to yield a practical result for a nontrivial problem.

The genetic algorithm library search was performed 50 consecutive times. Due to the indefinite nature of the genetic algorithm, each time the search routine produced a somewhat different MFC filter combination, but roughly the same RMSECV. Four common chemicals were selected as molecular filters: water, methanol, ethanesulfonic acid, and 2, 2-diethoxypropane. The transmission spectra of these chemicals are shown in Fig. 4. These four MFC filters gave an adequate calibration model ($r^2 = 0.995$, RMSEC = 0.229%, RMSECV = 0.339%, $p = 0.05$ by f test). This result is slightly better than a corresponding PCR calibration model based on corrected transmission spectra ($r^2 = 0.993$, RMSEC = 0.359%, RMSECV = 0.551%, $p = 0.05$ by f test). The PCR regression vector and simulated regression vector based on MFC filters are both presented in Fig. 5. It is evident in Fig. 5 that these two regression vectors do not match. The search for a regression vector by genetic algorithm is intended to reach a minimum on a multidimensional response surface. The PCR regression vector is one of many such minima, and it can be visualized as a point in a reduced orthogonal p-factor

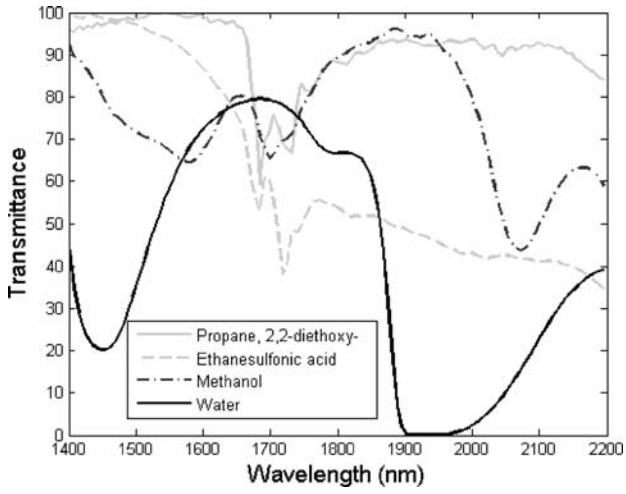

Fig. 4. The transmission spectra of the selected MFC filters.

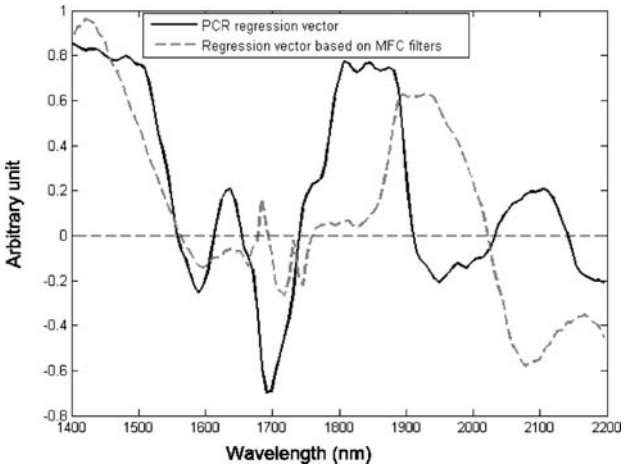

Fig. 5. The regression vectors versus wavelength. The *solid line* shows the PCR regression vector, and the *dashed line* shows the regression vector based on MLR calibration of the MFC filter.

space that describes a linear relationship between the spectra and concentration. Due to the stochastic nature of the genetic algorithm, it is possible to obtain several essentially equivalent solutions to the optimization problem. Therefore, it is not surprising that the regression vector generated by MFs did not match a predefined PCR regression vector. Such a pattern match is unnecessary. Indeed, the fact that multiple solutions exist makes it easier to find molecular filters that are stable and compatible with other molecules in the filter system.

RESULTS AND DISCUSSION

Analysis of Ethanol in Water Mixtures. Multivariate Analysis of Absorbance and Transmission Spectra. In order to compare the results of MFC measurements with the results of multivariate analysis using a conventional spectrometer, PCR was performed on the same training set used for selection of MFC filters. First, the PCR calibration was performed with corrected transmission spectra. The optimum predictive model was defined as the model with lowest RMSECV by leave-one-out cross validation. Four principal components were required to build a calibration model with optimum predictive ability. Theoretically, two principal components should be sufficient to model the ethanol/water mixtures. The extra principal components were included due to the nonlinear response between the transmission spectra and concentration. The RMSEC was 0.359%, corresponding to 2.56% error relative to the range of the calibration set. The four PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.551%, or 3.93% relative to the range of the calibration set. Next, a PCR calibration was carried out on absorbance spectra, which were calculated from original transmission spectra (using $A = 1/\log T$). Three principal components were required to build an optimum calibration model. The RMSEC was 0.309%, corresponding to 2.20% error relative to the range of the calibration set. The three PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.494%, or 3.53% relative to the mean of the calibration set. Compared to the PCR model based on corrected transmission spectra, the PCR model based on absorbance spectra required fewer principal components and had a slightly lower RMSEC and RMSECV. The better performance of the model based on absorbance spectra is expected because of the linear response between absorbance and concentration.

Expectation from Simulation. As described in the MFC filters selection section, the simulation study for the MFC filters predicted a RMSEC of 0.229% and RMSECV of 0.339% with corrected transmission spectra. The result showed that the PCR model is not necessarily the best model. MFC filters outperform the traditional scanning PCR model in terms of RMSECV. The simulation result in Fig. 6 shows a plot of the predicted ethanol concentrations versus the actual ethanol concentrations using a MLR model based on four MFC filters and a four-component PCR model based on corrected transmission spectra.

Determination of Ethanol with the MFC Approach. The voltage output from the detector was recorded for each of 39 samples through each MFC filter. The samples were split into two groups for cross validation, and 20 samples were used to calibrate the MFC-based instrument, while the other 19 samples were used as the validation dataset. The 20 calibration samples were different from those samples used as training samples for the MFC filters selection, but were prepared at the same nominal concentrations. Additional calibration was necessary because the correction factors used with the training spectra were all obtained from manufacturer's test datasheets and set-ups, and might be different once assembled in the prototype instrument. The optimal correlation between detector output voltage and ethanol concentration were obtained by following equation.

$$\widehat{Y} = \begin{bmatrix} v_{out[1,1]} v_{out[1,2]} v_{out[1,3]} v_{out[1,4]} \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ v_{out[m,1]} v_{out[m,2]} v_{out[m,3]} v_{out[m,4]} \end{bmatrix} \begin{bmatrix} -28,567 \\ -14,368 \\ 27,997 \\ 21,164 \end{bmatrix} - 34 \quad (8)$$

Where $\widehat{Y}$ was the predicted ethanol concentration, and v_{out} was the voltage output of each sample for each MFC filter. The RMSEC of the model was 0.748%, and the RMSEP by data splitting was 0.735%. Fig. 7 shows a plot of predicted ethanol concentrations versus actual ethanol concentrations of all 39 samples.

The estimated RMSEP (0.735%) of the MFC-based measurement was not as good as the RMSEP (0.339%) predicted by the simulation. Still, the actual MFC result shows that the MFC instrument is able to produce a useful numerical concentration result. The difference between the simulated instrument and the actual instrument results was due to several factors:

1. Sampling noise arose from the positioning of molecular filter cuvettes and/or sample cuvettes that did not exist in the simulation.
2. The transmission spectra in the NIR library were obtained with a path length of 2.5 mm, while cuvettes with a path length of 2 mm were used as MFC filters in the prototype instrument. The difference in the profile of transmission spectra due to the different path length likely increased prediction error.

3. Instrumental limitations prevented obtaining the exact transmission spectrum of the 1,400 nm long pass filter, the emission spectrum of the light source, and the detector response curve in the prototype MFC-based instrument to correct the training transmission spectra. Alternative correction factors were obtained from manufacturers' datasheets. Better results might be expected if each individual optical component in the prototype instrument were carefully calibrated.
4. Although an optical chopper and lock-in-amplifier were used to reduce noise and thermal drift, the MFC-based prototype instrument was shown to have a significant instrument drift. Simple studies with the light source (e.g. 1,400 nm long pass filter in place but without MFC chemicals or sample cell present) exhibited signal drift as high as 4% relative over 20 min, which was roughly the time required to scan all 39 samples in the MFC instrument. This significant drift could contribute to the high RMSEP. Future studies will utilize a double-beam design to eliminate this drift.
5. Using the genetic algorithm-based MFC filters selection algorithm, only the predictive ability of the MLR model was considered in the fitness function. The sensitivity of each individual MFC filter to changes in ethanol concentration was not taken into account. PCR is a regression method based on orthogonal principal components that maximize variance. However, MLR only aims to minimize the sum of the squared errors, and variance maximization for dependent variables is not taken into account. Therefore, the genetic algorithm-based MFC filters selection could select MFC filters with high prediction ability but low sensitivity, which results a hypothetical low RMSEP in the simulation study that is difficult to achieve with real, physical filters. (A new search algorithm that takes both prediction and sensitivity into account is currently being investigated.)

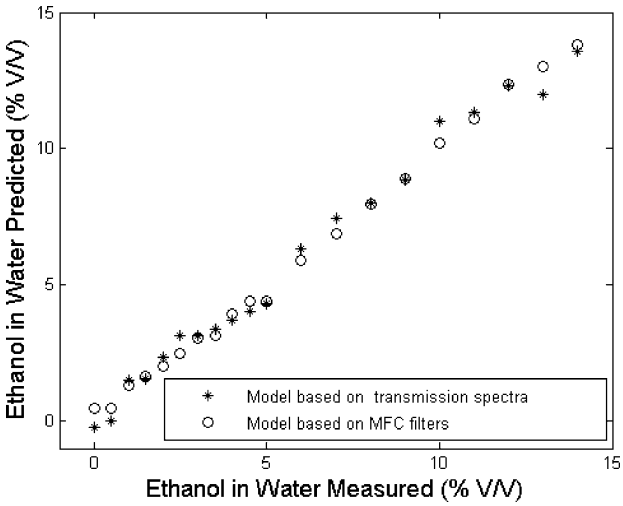

Fig. 6. A plot of the predicted ethanol concentrations versus the actual ethanol concentrations using a MLR model based on 4 simulated MFC filters and a PCR model based on corrected transmission spectra. *Stars*: PCR model based on corrected transmission spectra, RMSEC = 0.359%, RMSECV = 0.551%. *Circles*: MLR model based on four simulated MFC filters, RMSEC = 0.229%, RMSECV = 0.339%.

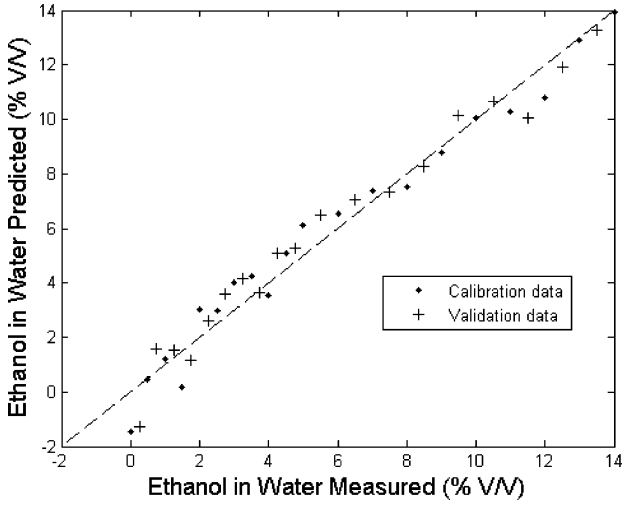

Fig. 7. A plot of the predicted ethanol concentrations versus the actual ethanol concentrations of all 39 samples. *Diamonds*: calibration samples, $r^2 = 0.968$, RMSEC = 0.748%. *Crosses*: validation samples, RMSEP = 0.735%. Significant at $p = 0.05$ by f test.

In addition to the multivariate regression model for ethanol concentration, an estimate of the detection limit for binary mixtures of ethanol and water was also calculated. The estimate was based on an extension of the BEST metric for sub-cluster detection with sample populations that has been described previously (33,34). The experimental MFC data were then analyzed to estimate the limits of detection of each component in binary mixtures of two components. This was performed by translating the sample population mean of 1% ethanol in water sample towards pure water sample population's mean until the two clusters could not be differentiated using the BEST subcluster detection algorithm. The estimate of the detection limit for ethanol in water determined by this procedure is 0.26%. The dynamic range for ethanol detection by MFC was a factor of 57. The extended BEST metric provided lower errors than traditional regression approaches because it took both changes in sample cluster location as well as scale into account. However, to achieve its better results the extended BEST requires multiple replicates of the same sample, which can be impractical in real-life remote sensing applications.

In order to assess of the long-term stability of molecular filters, the molecular filters were directly exposed to the near-IR light beam for 10 h. For each of those four molecular filters, the signal was continually monitored and variations in signal level of 4% were observed in this study (the same range as the variation in the light source intensity). The molecular filters were also sealed in cuvettes over two-month period, and there appeared no visible degradation of these molecular filters over that time. It is worth noting that, for some other molecular filters that were not used in this study, severe degradation of MFs can be observed. Thus, it is necessary to compile a spectral library using only stable molecules for MFC.

The susceptibility of MFC-based spectroscopic measurement to complex matrix interference in samples is not well understood. Theoretically, the MFC-based instrument should be able to precisely measure the specific chemical species of interest as long as the potential interferences were introduced and modeled in the training set. Future research will include determination of ethanol containing other alcohols as interferences that are not in the training set to evaluate the susceptibility of MFC to this sort of interference.

CONCLUSION

A prototype MFC-based spectrometer was designed, constructed, and tested for the analysis of ethanol-in-water mixtures. The concept of molecular factor computing was demonstrated. The results obtained from an MFC-based measurement were compared to PCR calibration based on conventional scanning spectrometry. Although the actual results from MFC-based prediction in the first prototype were slightly worse than from conventional PCR prediction, the MFC simulation study suggested that a better prediction model could be built based on MFC. A double-beam MFC instrument under construction may achieve the superior results predicted by the simulation. Advantages of the MFC approach over conventional spectroscopy include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost. The high optical throughput of an MFC system could offer improved analytical ability in systems with a weak signal.

Problems with reproducibility in positioning of filter cuvettes and samples cuvettes increased measurement noise in the MFC-based prototype spectrometer. The effect will be reduced by using aperture control and through better design of slides for holding filters and samples.

A new library search algorithm should be developed to select the optimal MFC filters. Prediction ability and sensitivity of MFC filters both should be taken into account in the fitness function of genetic algorithm-based searches.

The number of potential filter materials is huge. Solutions and solid-state mixtures could both be used as molecular filters. The use of organic solvents as MFC filters introduces some ruggedness problems for process analysis. To simplify the instrument and improve the system stability, solid-state MFC filters constructed from materials such as polymers may offer a good alternative to liquid filters (10).

MFC offers users a simpler ISP instrument with significant reduction of computational complexity and processing time at the cost of some experimental flexibility. In other words, MFC-based instruments are not general-purpose research tools. Instead, the MFC approach is for practical measurement in the real world where fast results are needed and achieved by integrating the processing into the sensing stage.

In addition to applications of this technique as a process analytical technology (PAT), MFC-based remote NIR imaging for real-time surveillance has gained interest. A MFC-based NIR imaging system for remote ethanol sensing is currently under construction in our laboratory. The range of possible applications is likely to expand when imaging systems are available.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation through CNS-0540178, the Kentucky Science and Education Fund, and by the National Institutes of Health through N01AA 33003 and T32 HL072743.

REFERENCES

1. R. J. Dempsey, D. G. Davis, R. G. Buice, and R. A. Lodder. Biological and medical applications of near-infrared spectroscopy. *Appl. Spectrosc.* **50**:18A–34A (1996).
2. J. K. Drennen and R. A. Lodder. Nondestructive near-infrared analysis of intact tablets for determination of degradation products. *J. Pharm. Sci.* **79**:622–627 (1990).
3. A. S. El-Hagrasy, H. R. Morris, F. D'Amico, R. A. Lodder, and J. K. Drennen, 3rd. Near-infrared spectroscopy and imaging for the monitoring of powder blend homogeneity. *J. Pharm. Sci.* **90**:1298–1307 (2001).
4. A. Urbas, M. W. Manning, A. Daugherty, L. A. Cassis, and R. A. Lodder. Near-infrared spectrometry of abdominal aortic aneurysm in the ApoE−/− mouse. *Anal. Chem.* **75**:3318–3323 (2003).
5. T. D. Ridder, S. P. Hendee, and C. D. Brown. Noninvasive alcohol testing using diffuse reflectance near-infrared spectroscopy. *Appl. Spectrosc.* **59**:181–189 (2005).
6. J. C. Soto, C. P. Meza, W. Caraballo, C. Conde, T. Li, K. R. Morris, and R. J. Romanach. On line non-destructive determination of drug content in moving tablets using near infrared spectroscopy. *Journal of Process Analytical Technology* **2**(5):8–14 (2005).
7. Guidance for Industry PAT—A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation (CDER), and Research (CDER),Center for Veterinary Medicine (CVM), Office of Regulatory Affairs (ORA), September 2004.
8. A. S. Hussain. Process analytical technology: a first step in a jouney towards the desired state. *Journal of Process Analytical Technology* **2**(1)8–13 (2005).
9. S. R. Byrn, J. K. Liang, S. Bates, and A. W. Newman. PAT—process understanding and control of active pharmaceutical ingredients. *Journal of Process Analytical Technology* **3**(6): 14–19 (2006).
10. M. R. Fischer and G. M. Hieftje. Near-IR multiplex bandpass spectrometer utilizing polymer filters. *Appl. Spectrosc.* **50**:1246–1252 (1996).
11. A. Fong and M. G. Hieftje. Near-IR multiplex bandpass spectrometer using liquid molecular filters. *Appl. Spectrosc.* **49**:493–498 (1995).
12. K. R. Beebe and B. R. Kowalski. Introduction to multivariate calibration & analysis. *Anal. Chem.* **59**:1007A–1017A (1987).
13. H. Martens and M. Martens. Multivariate analysis of quality an introduction. Wiley, Chicester (2001).
14. H. Martens and T. Naes. Multivariate calibration. Chapman and Hall, London (1989).
15. R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets. *J. Chemom.* **14**:643–655 (2000).

16. R. Leardi. Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data Handl. Sci. Technol.* **23**:169–196 (2003).
17. C. Schwartz. Integrated Sensing and Processing http://www.darpa.mil/dso/thrust/math/isp.htm.
18. O. Soyemi, D. Eastwood, L. Zhang, H. Li, J. Karunamuni, P. Gemperline, R. A. Synowicki, and M. L. Myrick. Design and testing of a multivariate optical element: the first demonstration of multivariate optical computing for predictive spectroscopy. *Anal. Chem.* **73**:1069–1079 (2001).
19. S. E. Bialkowski. Species discrimination and quantitative estimation using incoherent linear optical signal processing of emission signals. *Anal. Chem.* **58**:2561–2563 (1986).
20. A. M. C. Prakash, C. M. Stellman, and K. S. Booksh. Optical regression: a method for improving quantitative precision of multivariate prediction with single channel spectrometers. *Chemometr. Intell. Lab. Syst.* **46**:265–274 (1999).
21. F. G. Haibach, A. E. Greer, M. V. Schiza, R. J. Priore, O. O. Soysmi, and M. L. Myrick. On-line reoptimization of filter designs for multivariate optical elements. *Appl. Opt.* **42**:1833–1838 (2003).
22. F. G. Haibach and M. L. Myrick. Precision in multivariate optical computing. *Appl. Opt.* **43**:2130–2140 (2004).
23. M. L. Myrick, O. Soyemi, J. Karunamuni, D. Eastwood, H. Li, L. Zhang, A. E. Greer, and P. Gemperline. A single-element all-optical approach to chemometric prediction. *Vibr. Spectrosc.* **28**:73–81 (2002).
24. M. L. Myrick, O. Soyemi, H. Li, L. Zhang, and D. Eastwood. Spectral tolerance determination for multivariate optical element design. *Fresenius' J. Anal. Chem.* **369**:351–355 (2001).
25. M. L. Myrick, O. O. Soyemi, F. Haibach, L. Zhang, A. Greer, H. Li, R. Priore, M. V. Schiza, and J. R. Farr. Application of multivariate optical computing to near-infrared imaging. *Proc. SPIE Int. Soc. Opt. Eng.* **4577**:148–157 (2002).
26. M. L. Myrick, O. O. Soyemi, M. V. Schiza, J. R. Farr, F. Haibach, A. Greer, H. Li, and R. Priore. Application of multivariate optical computing to simple near-infrared point measurements. *Proc. SPIE Int. Soc. Opt. Eng.* **4574**:208–215 (2002).
27. O. O. Soyemi, F. G. Haibach, P. J. Gemperline, and M. L. Myrick. Nonlinear optimization algorithm for multivariate optical element design. *Appl. Spectrosc.* **56**:477–487 (2002).
28. O. O. Soyemi, F. G. Haibach, P. J. Gemperline, and M. L. Myrick. Design of angle-tolerant multivariate optical elements for chemical imaging. *Appl. Opt.* **41**:1936–1941 (2002).
29. L. A. Cassis, B. Dai, A. Urbas, and R. A. Lodder. *In vivo* applications of a molecular computing-based high-throughput NIR spectrometer. *Proc. SPIE-Int. Soc. Opt. Eng.* **5329**:239–253 (2004).
30. L. A. Cassis, A. Urbas, and R. A. Lodder. Hyperspectral integrated computational imaging. *Anal. Bioanal. Chem.* **382**: 868–872 (2005).
31. P. Geladi and B. Kowalski. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185**:1–17 (1986).
32. E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M.-H. Tsou, C.-F. Horng, A. B. E. S. Iversen, M. Liao, C.-M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene expression predictors of breast cancer outcomes. *Lancet* **361**:1590–1596 (2003).
33. R. A. Lodder and G. A. Hieftje. Detection of subpopulations in near-infrared reflectance analysis. *Appl. Spectrosc.* **42**:1500–1512 (1988).
34. Y. Zou, *et al.* Making your best case—near-IR spectral identification of soil. *Anal. Chem.* **65**:A434–A439 (1993).