Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1040 (2004) 273–282
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

High-throughput automated post-processing of separation data

Jonathan G. Shackman [a], Christopher J. Watson [a], Robert T. Kennedy [a,b,*]

[a] *Department of Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109-1055, USA*
[b] *Department of Pharmacology, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109-1055, USA*

Received 22 December 2003; received in revised form 5 April 2004; accepted 7 April 2004

Abstract

The development of an efficient method for high-throughput analysis of multiple electropherograms or chromatograms collected in series is presented. The method, encoded in a computer program designated "Cutter", utilizes batch processing for determining chromatographic figures of merit (CFOM) including peak centroid times, heights, areas, signal-to-noise ratios (S/N), variance (σ^2), skew, excess, and plate number (N) across a set of separations collected serially. The software was validated using simulated data with varying S/N, skew, and excess. The accuracy of the analysis was comparable to or improved over commercial software with area calculation relative errors (RE) below 5% for simulated data with S/N = 5. File sets containing 1300 electropherograms were analyzed in 5 min, representing a nearly 200-fold reduction in analysis time from other methods. Incorporated within the program is a novel method for automated peak deconvolution using an Empirically Transformed Gaussian function. Area measurements of deconvoluted peaks were within 3% of the true value of a simulated data set with S/N = 5 and resolution (R_s) = 1 for equivalent peaks, and within 10% when the ratio of the overlapped peak heights was 10:1.

Keywords: Empirically transformed Gaussian; Software; Bioinformatics

1. Introduction

Breakthroughs in sample injection and high-sensitivity detection have enabled significant improvements in the speed of capillary electrophoresis (CE) separations (for review see reference [1]). High-resolution separations with over 300,000 theoretical plates have been achieved in less than 10 s [2,3]. Sub-second separations have been reported with some degradation of resolution [3–5]. Chromatographic separation speeds have also increased dramatically with the advent of small particles in liquid chromatography (LC; for review see reference [6]) and advanced injection methods coupled with open tubular columns in gas chromatography (GC) [7]. These high-speed separations may facilitate many new applications, such as high-throughput screening, bed-side clinical assays, and chemical monitoring. In chemical monitoring a rapid separation is coupled on-line to sampling and derivatization chemistry (if necessary) so that changes in the concentration of analyte(s) can be tracked over time by serial separations. Applications include monitoring industrial processes [8,9], chemical reactions [10,11], separations (i.e. two-dimensional separations) [12,13], neurotransmitters in vivo [14,15], and hormone secretion from cells [16]. With the improved instrumentation for fast separations, it is necessary to develop methods that can efficiently analyze high-speed separations data.

Chemical monitoring by separations can generate large data sets with special requirements for analysis. For example, monitoring a process for 8 h by electrophoresis at 10 s temporal resolution would generate nearly 3000 electropherograms. These electropherograms may contain several peaks of interest for quantification that change over wide concentration ranges. For method development, ensuring good data quality, or diagnosis of problems it would also be of interest to determine chromatographic figures of merit (CFOM) on the resulting peaks. Furthermore, with a high-speed separation compromises in the separation quality may result in overlapping peaks that require deconvolution to produce accurate peak characterization. Useful data reporting would include quantitatively plotting analyte concentration as a function of time and tabulating peak parameters such as peak centroid location, skew, efficiency, signal-to-noise ratio (S/N), and plate number (N).

* Corresponding author. Tel.: +1-734-615-4363; fax: +1-734-615-6462.
E-mail address: rtkenn@umich.edu (R.T. Kennedy).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.04.004

Conventional data analysis software is not well-suited for meeting the analysis requirements of large sets of electropherograms (or chromatograms). Most programs employ serial analysis wherein the user loads one file containing a single electropherogram, analyzes it, obtains a result, and repeats for all files obtained from a given experiment. Although programs may allow for scripting of common actions, this is still a serial mode of data analysis and the scripts can be limited in utility. This methodology is relatively slow, often requiring a few minutes to analyze a file, meaning that data analysis can become the rate limiting step in the case of high-throughput separations that are collected in a few seconds. The one-at-a-time method also fails to produce comprehensive reports across a data set requiring many single reports to be combined to perform useful file-to-file correlations. Another frequent limitation of commercial software is difficulty in tracking peaks that change from undetectable (i.e. "blanks") to detectable during the course of analysis. For example, when monitoring a reaction, the initial electropherograms may not have a peak corresponding to the product; however, the product signal, known to occur at a given migration time, will increase as the reaction proceeds. Many programs do not provide the option of defining a value at the point where the product will appear, thus giving no zero point.

A few programs with multi-chromatogram analysis capability have been reported such as CHAS (a FORTRAN program), which loaded a set of chromatograms into a master library from which single files were extracted for analysis [17]. More recently, MICHROM has been developed for the analysis and optimization of chromatographic data [18]. Although the program can load up to 50 chromatograms, the analysis time or mode of operation (serial or parallel) was not reported. Some exceptions to serial methodology are commercially available such as the GRAMS (Thermo Galactic, Salem, NH) software package for spectroscopy data processing. Although a chromatography add-on module is available, it is intended mainly for the construction of calibration curves and presentation of data instead of extracting CFOM.

In this work, we describe a new method for efficient analysis of large batches of separation data using a LabVIEW-based program. The program, "Cutter", utilizes a bulk flow paradigm for data analysis and generates comprehensive reports including relevant CFOM across a given data set. Additionally, it incorporates a novel automated means of peak deconvolution based upon a form of the Empirically Transformed Gaussian function [19–21].

2. Experimental

2.1. Cutter program construction

The data analysis program Cutter was written in LabVIEW 6.1 and follows a modular design (National Instruments, Austin, TX). A general outline of its operation is given in Fig. 1. After all data is loaded it may be treated by optional data manipulations (e.g., smoothing or baseline subtraction) before peaks are located and CFOM calculated. The end product is a plot of CFOM for all peaks from all files, with the ability to perform statistical analyses upon these results. The program was run on a 400 MHz PC computer unless otherwise noted. Cutter is available as a free download at: http://www.umich.edu/~rtkgroup.

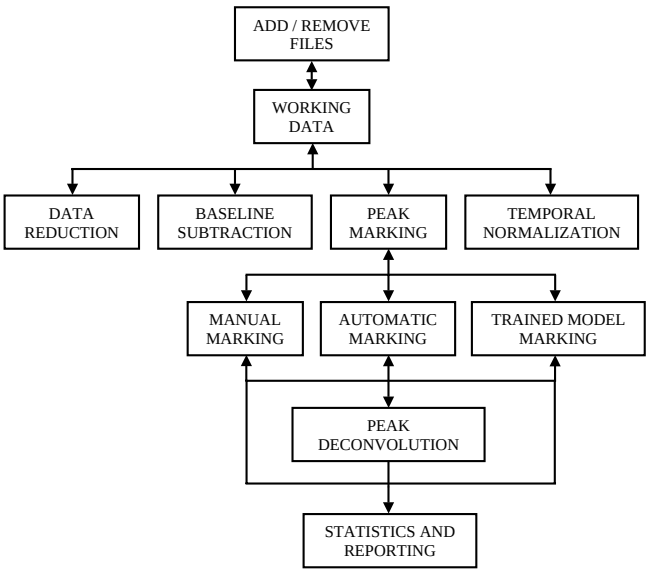

Fig. 1. Flow chart outline of the operation of the Cutter high-throughput analysis program.

2.1.1. Building the data set

In the program, a batch of files are loaded as a 3-dimensional array, with the first two dimensions being the abscissa and ordinate data of each file (i.e. time and intensity) and the third being each file loaded. In some instances, temporal normalization of the data files was used. In this case, the program aligned individual electropherograms about a common peak. File starting points were then equalized, and points equivalent to the minima of the data were added to the file ends to match the longest file loaded, yielding equivalent file sizes. With this manipulation, migration time deviation statistics of the common peak were lost, however, it was useful for simplifying peak marking in some cases.

2.1.2. Baseline subtraction and filtering

Baseline subtraction was accomplished either by fitting an n-order polynomial to the entire data set and then subtracting the value of this polynomial from the signal at each time point or by fitting x-number of n-order polynomials to consecutive blocks of data (i.e. portions of the electropherograms) before subtraction, which allowed for extrication of the baseline from peak components of the data. For this manipulation, the subtracted data ordinate values could

either optionally be centered about the average of the noise or be constrained to only positive values by determining the minimum value of the ordinate data and setting it to zero. Solving of n-order polynomials was typically accomplished using a singular value decomposition (SVD) [22], though other algorithms could be utilized (e.g., Givens or Householder [22]). Generally, for CE data, a zero order line was needed, although higher orders could be used. For data exhibiting a baseline not easily fit by a polynomial (e.g. a highly irregular baseline or one containing a matrix interferent), a selected data file could be used as a "blank" for subtraction from all files. In order to subtract baselines from individual peaks, versus the entire electropherogram or chromatogram, a simple line was fitted to the start and end points of the peaks. Although several methods were available in Cutter for data smoothing, all work discussed herein was performed in the absence of filtering of the raw data.

2.1.3. Peak marking and quantitative analysis

Several methods of peak marking (i.e. defining the beginning and end of peaks within the data) were incorporated into Cutter for testing. In manual mode, the user located a desired peak between two cursors in an overlaid plot of all the electropherograms or chromatograms. This process could be repeated for all peaks of interest. Successful use of this mode was aided by temporal normalization because it required that the migration time reproducibility be sufficient to avoid substantial marking errors. In automated modes, peaks were detected using a first derivative algorithm [23]. In this method, first derivatives, which could be filtered using a Bessel algorithm, were calculated for all files. Peaks were located by determining the time at which the derivative signal crossed a threshold. The threshold was defined by the user as a multiple, typically 10 or greater, of the standard deviation of the derivative in a baseline region. For a peak to be marked, the derivative was required to pass though a positive threshold twice and a negative threshold twice. A minimum peak width, typically at least nine points, was defined to exclude noise spikes. The algorithm also allowed for cross correlation of peaks found across all files, meaning that a peak found within a given region of one file could be marked in all files, which was useful for marking "blank" peaks or in cases of low S/N. Finally, peaks could be determined for all files by using a single file as a model with either manual or automatic peak marking [24]. For example, a standard test mixture containing only analytes of interest could be used to target peaks for marking from more complicated (e.g. in vivo) experiments.

Once the peak locations were determined, the program calculated the following figures of merit: centroid time, height, area, S/N [25], variance (σ^2), skew, excess [26], and N based both on direct calculation of σ^2 and by the width at half height [27]. Heights and areas were calculated both as a simple maxima and integration, respectively, or by fitting and subtracting a linear polynomial from the start and end points of the individual peaks prior to calculation.

2.1.4. Deconvolution of overlapped peaks

Overlapped peaks were resolved using the following form of the empirically transformed Gaussian (ETG) equation [19]:

$$h(t) = \frac{H''}{\{1 + \lambda_l \exp[k_l(t_l - t)]\}^{\alpha} + \{1 + \lambda_t \exp[k_t(t - t_t)]\}^{\beta} - 1} \quad (1)$$

where λ_l and α are leading edge parameters, λ_t and β are trailing edge parameters, k are σ^{-1} for a symmetrical peak, and t_l and t_t are roughly the half width times of the leading and trailing edges, respectively. H'' is given by:

$$H'' = 2H\exp(0.5) \quad (2)$$

where H is the peak maxima.

The ETG function was chosen because it has been thoroughly characterized [19,20] and shown to be a highly accurate model for a wide variety of peak shapes. Furthermore, ETG converges rapidly and can be solved with limited a priori knowledge of the peak. (It is recognized however that more robust equations have been developed for extreme peak asymmetries [21,28].) Using initial values found during the peak marking routine for H'', t_l, t_t, k_l, k_t, and a unity value for the remaining parameters (λ_l, λ_t, α, and β), the equation was solved using the Levenberg–Marquardt method [29]. H'' was considered to be a true value based on peak maxima and the other eight parameters (t, k, λ, α, β) were solved. For the deconvolution of n-peaks, Eq. (1) was summed across the number of peaks (e.g., a three peak system would have 24 parameters determined). The speed, accuracy, and precision of convergence were enhanced by first sampling only 25% of a data set (e.g. for 100 points, only every fourth point was used) and solving for the parameters. These solved parameters were then used as initial guesses for a 50% sampling, followed by 75% sampling, and finally the results of the 75% were used as initial guesses for the full data set. Each sampling could be by-passed, as was necessary when peaks were undersampled (e.g. a peak with less than 10 data points); alternatively, in the case of large data sets, the 100% sampling itself could be bypassed.

2.2. Data sets analyzed

2.2.1. Simulated data

In order to validate Cutter, a set of data with known characteristics was constructed as recommended previously [30]. Using a simplified form of Eq. (1):

$$h(t) = \frac{H''}{1 + \exp[k_l(t - t_l)] + \exp[k_t(t - t_t)]} \quad (3)$$

H'', k_l, k_t, t_l, and t_t were manually manipulated until Gaussian-based peaks with various skew (−0.3, 0, and 0.3) and excess (−0.3, 0, and 0.3) were obtained. The signal generated by Eq. (3) was considered "noiseless". To simulate white noise, LabVIEW was used to generate random

numbers with a Gaussian distribution with an average value of 0 and a standard deviation of 1, which were multiplied by an empirically determined constant to create the desired S/N ratio (S/N = 5, 10, 100, and 1000), and added as a series to the noiseless data sets. These operations yielded a total data set of 45 peaks with 1000 points per file.

Height and area calculations were performed for all simulated data within Cutter and Origin 6.0 (Microcal Software Inc., Northampton, MA). Percent relative errors (RE) of the data were calculated using the deviation of the area from the noiseless signal. For deconvolution testing, simulated equivalent peaks with zero skew and excess were added together with a temporal offset to create a set of overlapped peaks with various R_S (0.50, 0.75, 1.00, 1.50, and 2.00). A second set maintaining $R_S = 1.00$ at varying S/N was constructed while varying the ratio of the two peak sizes from 1.3 to 20. A third set was created at varying S/N while maintaining $R_S = 1$, the first peak with zero skew and excess, and the second peak with variable (−0.3, 0.0, or 0.3) skew and excess.

2.2.2. Microfluidic electrophoresis data

Data from a microfluidic device performing CE immunoassays (CE-IA) for insulin were obtained as discussed before [16]. Briefly, 0–1500 nM insulin, 150 nM fluorescein isothiocyanate-labeled insulin (FITC-insulin), and 75 nM monoclonal antibody to insulin were mixed on-line within a fluidic network microfabricated in glass. The sample and immunoreagents were electroosmotically driven along a heated channel where they were allowed to react. The reaction stream was then sampled via a flow-gated injection interface every 10 s and separated by CE within 5 s using an electric field of 500 V/cm. Laser-induced fluorescence (LIF) detection was performed via an epi-fluorescent microscope using the 488 nm line of a 20 mW Ar^+ laser and photon counting photomultiplier tube (PMT) sampling at 100 Hz. Instrument control and data collection was via a computer and LabVIEW controlled data acquisition board.

2.2.3. In vivo capillary electrophoresis data

Amino acids were measured in vivo by microdialysis coupled on-line to CE-LIF as described elsewhere [2]. Briefly, microdialysis probes were implanted into the striatum of ovariectomized female Sprague–Dawley rats. After recovery from surgery, the animals were placed in a cage fitted with a swivel system and the dialysis probe was connected to a perfusion pump and the CE-LIF system. Probes were perfused at 1 μl/min. Dialysate was derivatized on-line with a solution composed of 10 mM *o*-phthaldehyde (OPA), 40 mM β-mercaptoethanol (BME), 36 mM borate, 0.81 mM hydroxypropyl-β-cyclodextrin (HPBCD), and 10% methanol (v/v) at pH 9.5. 200 ms electrokinetic injections of derivatized dialysate were controlled by a flow-gated interface. Capillary electrophoresis was carried out in a 9 cm long, 10 μm inner diameter, 150 μm outer diameter fused-silica capillary with an electric field of 2.2 kV/cm. Fluorescence was induced with the 351 nm laser line from an Ar^+ laser and detected off-column using a sheath-flow cuvette [31] via a PMT and current-amplifier. Instrument control and data collection was via a computer and LabVIEW controlled data acquisition board. After stable electropherograms were recorded, an artificial cerebral spinal fluid solution containing 75 mM K^+ was perfused through the probe for 10 min to evoke stimulation of neurotransmitter release.

3. Results and discussion

3.1. Validation

Initial experiments were aimed at validating the peak marking capabilities and peak parameter calculations of the new program as previously recommended [30]. The heights and areas of Gaussian-based peaks with skew and excess varied from −0.3 to 0.3 and S/N varied from 5 to noiseless (Fig. 2a) were calculated using Cutter and compared to a commercial program (Origin) using automated peak

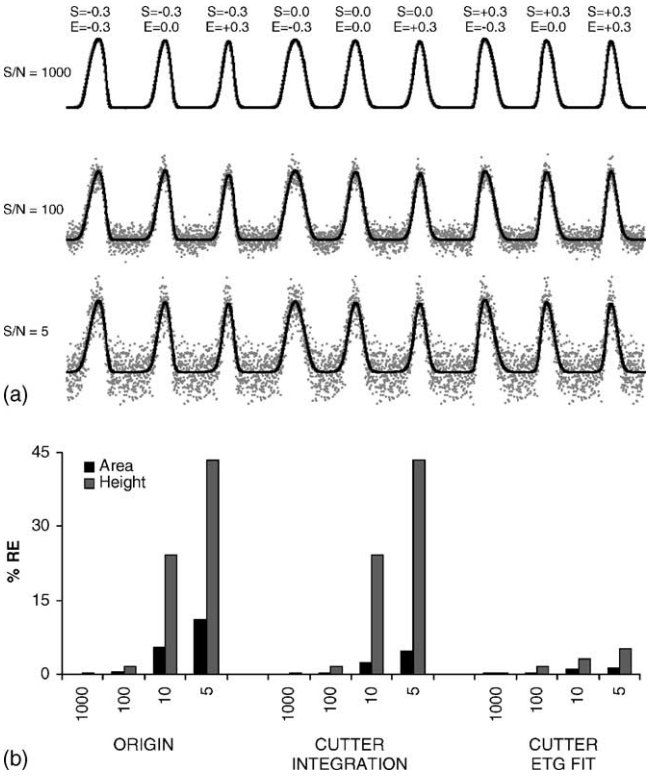

Fig. 2. Simulated data sets used to validate Cutter. (a) Peaks (grey dots) are Gaussian-based with various degrees of skew (S) and excess (E). The different S/N values were obtained by adding increasing amounts of white noise to the noiseless signal. Results of fitting an ETG function to the simulated data are shown as black lines. (b) Average %RE of height and area calculations across all peaks vs. S/N. Results compare the error obtained using Origin, Cutter with simple numeric integration, and Cutter with fitting of ETG function to the peaks.

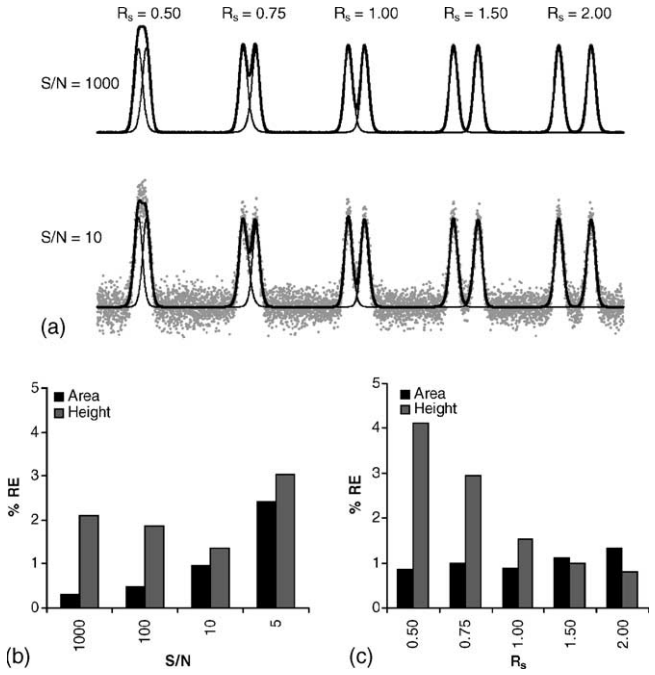

Fig. 3. Deconvolution via ETG of two equivalent Gaussian peaks as R_S and S/N are varied. (a) Deconvolution of data (grey dots) via ETG fit at S/N = 1000 (top) and S/N = 10 (bottom). Individual peaks are shown as dashed lines and their sum resulted in the calculated signal (heavy lines). (b) Average %RE of height and area calculations across all peaks vs. S/N. (c) Average %RE of height and area calculations across all S/N vs. R_S.

detection. Using height and area values obtained from the noiseless signal as true values, the average RE across the various skews and excesses obtained by both Origin and Cutter for height calculations were 43.4, 24.1, 1.5, and 0.1% at S/N = 5, 10, 100, and 1000, respectively. Peak area RE for Origin were 11.0, 5.5, 0.4, and 0.0% at S/N = 5, 10, 100, and 1000, respectively; while Cutter obtained RE of 4.5, 2.4, 0.2, and 0.0% (Fig. 2b). The errors in height tend to be greater than those in area because the error is affected by imprecision in both baseline and height determinations, whereas the area calculations are effectively "smoothed". The error was influenced more by the S/N than the extent of deviation from a true Gaussian in accordance with previous studies [32]. For example, with skew and excess equal to zero, the area RE given by Cutter for S/N = 5, 10, 100, and 1000 were 5.0, 1.1, 0.1, and 0.0%, respectively, while with skew and excess equal to 0.3 the area RE were 4.5, 0.3, 0.1, and 0.0%. Based upon these results, it was concluded that the peak marking method and peak parameter calculations of Cutter were valid for a variety of asymmetries and S/N. Further studies would be required to determine the limits under more extreme conditions, such as extreme baseline drift, skew, or non-Gaussian noise.

3.2. Deconvolution

We next investigated the use of the ETG function to deconvolute peaks. Deconvolution was expected to improve: (1) reproducibility for peak parameter calculation at low S/N by effectively smoothing the data [23], and (2) improve accuracy and reproducibility of peak analysis for overlapping

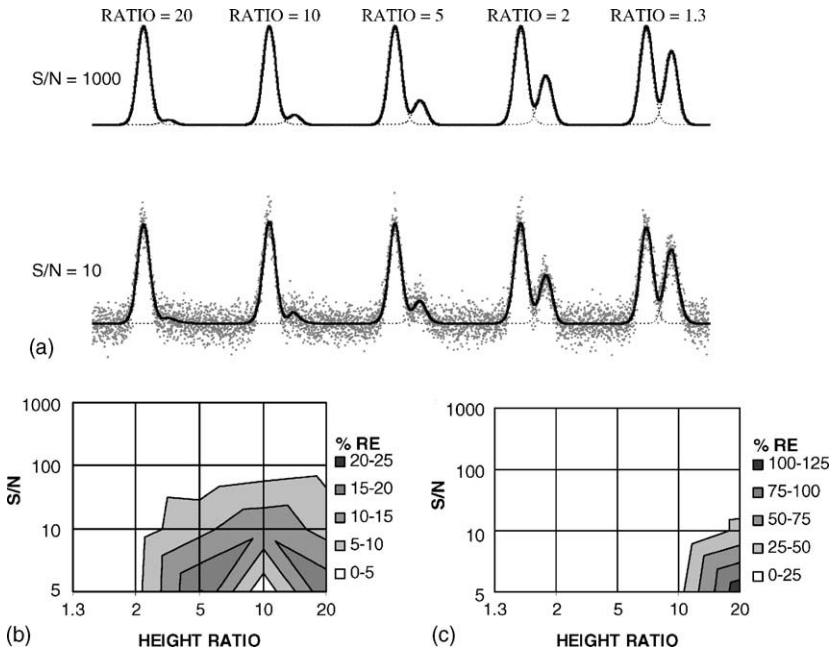

Fig. 4. Deconvolution via ETG of two Gaussian peaks as the ratio of heights and S/N are varied. (a) Deconvolution of data (grey dots) via ETG fit at S/N = 1000 (top) and S/N = 10 (bottom). Individual peaks are shown as dashed lines and their sum resulted in the calculated signal (black lines). Height (b) and area (c) %RE of the smaller peak plotted as height ratio vs. S/N.

zones. While the ETG function has been successfully applied to fitting of both simulated and real data for single resolved peaks [19–21], it has not been applied to deconvolution of overlapped peaks.

In this work, the form of ETG used collapses exponent factors of the original form into single variables, as was necessary for faster convergence and to avoid computer buffer under- and overflows resulting in a failed fit. Initial values of the parameters were automatically determined based upon peak shape without user intervention. The algorithm was first tested by fitting Eq. (1) to simulated single peaks of varying asymmetries and noise levels shown in Fig. 2a. The height calculation RE at S/N = 5, 10, 100, and 1000 after fitting of the ETG to the symmetrical Gaussian was 4.3, 2.83, 2.14, and 0.12%, respectively, while area RE was 0.8, 0.1, 0.3, and 0.3%. The average RE across all the different asymmetries for height calculations after fitting of the ETG at S/N = 5, 10, 100, and 1000 were 5.2, 3.0, 1.5 and 0.1%, respectively, and 1.2, 1.1, 0.3, and 0.3% for area measurements (Fig. 2b). These results are a vast improvement over results without deconvolution (see comparisons in Fig. 2b) and confirm the utility of the ETG function for peak fitting [19,20].

The program was then tested with simulated data containing two equivalent overlapped Gaussian peaks as the resolution was varied from 0.5 to 2.0 at different levels of S/N. Example results at S/N of 10 and 1000, solved via summation of Eq. (1) over the 2 peaks, are given in Fig. 3a. RE was calculated based upon either the height or area of a single Gaussian without noise versus the individual deconvoluted measurements. As expected, lower S/N gave the largest margin of error, although even at $R_s = 0.5$ with S/N = 5 the average RE of the two peaks was only 3.0 and 2.4% for heights and areas, respectively. The degree of overlap appeared to have a slightly greater contribution to the error than the S/N (Fig. 3b and c).

The effectiveness of deconvolution can be altered when one of the overlapping peaks has a much greater magnitude than the other. To test the effect of varying peak size ratio, a simulated pair of Gaussian peaks with a constant $R_s = 1$ and peak height ratios of 20, 10, 5.0, 2.0, and 1.3 at different S/N were examined (Fig. 4a). RE as a function of peak height ratio and S/N are shown in the contour plots of Fig. 4 for both peak height (Fig. 4b) and peak area measurements (Fig. 4c) of the smaller peak, which illustrate the range of conditions wherein a given error will be obtained. The smaller peak was more dramatically affected than the larger peak, especially as the S/N decreased and the peak ratio increased. Nevertheless, even with a peak ratio of 20 and S/N of 10 the height RE for the smaller peak was 8.0%.

Peak deconvolution was also tested with data sets at varying S/N where skew and excess of one peak was held

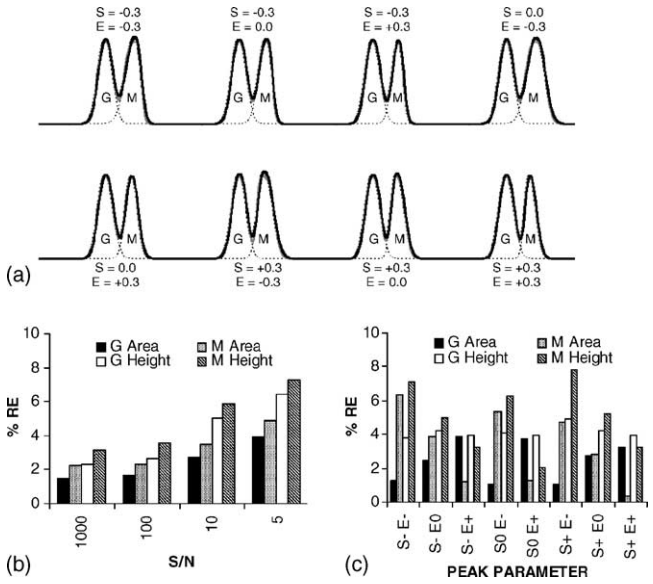

Fig. 5. Deconvolution via ETG of Gaussian (G) and modified Gaussian (M) peaks as skew (S), excess (E), and S/N are varied. (a) Deconvolution of data (grey dots) via ETG fit at S/N = 1000. Individual peaks are shown as dashed lines and their sum resulted in the calculated signal (black lines). (b) Average %RE of height and area calculations across all peaks vs. S/N. %RE are reported for the G and M peaks at each set of peak parameters. (c) Average %RE of height and area calculations across all S/N vs. skew and excess parameters. −0.3, 0.0, and 0.3 have been abbreviated as (−), (0), and (+), respectively.

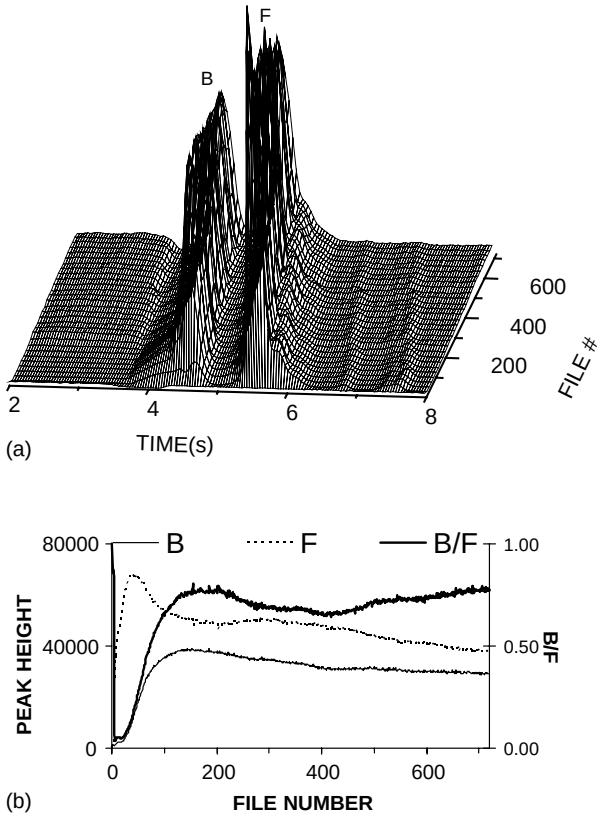

Fig. 6. Cutter results for a CE-IA data set. (a) 720 electropherograms attained at a 100 Hz sampling rate every 10 s; peaks are marked as bound (B) and free (F) analyte. (b) Analysis results obtained within 5 min plotting peaks heights for B and F and the ratio of the two peaks (B/F).

constant at zero while the second peak had variable (−0.3, 0.0, and 0.3) skew and excess and R_s was held constant at 1 (Fig. 5a). Average RE for both peaks were under 10% at all S/N, whether for area or height calculations. Error appeared to be most heavily influenced by S/N instead of the degree of deviation from Gaussian (Fig. 5b and c) in accordance with previous studies [32]; although, those peaks with an excess of −0.3 tended to exhibit slightly higher RE than those with no excess or positive excess components.

3.3. CE immunoassay analysis

After initial characterization of the method on simulated data, tests were performed on real data sets. To test the ability of the method to analyze large files, batches of CE-based immunoassay data were analyzed. A set of 720 immunoassay electropherograms, collected every 10 s using a 100 Hz sampling rate on a microchip CE system [16], are shown in Fig. 6a. The two most intense peaks correspond to free fluorescent tracer (F) and fluorescent tracer bound to antibody (B). Analysis of the total data set, which included marking each peak and determining CFOM, was completed with automated peak marking by Cutter in <5 min. A portion of the data report, a plot of peak height for the B and F peaks as well as for the peak ratio is illustrated in Fig. 6b. Serial analysis of this data set, even with automatic peak marking, required ∼12 h.

At the beginning of the data set the peak heights had not stabilized because the fluidic device was not filled with sample thus causing the initial fluctuation in peak height. These data illustrate the utility of marking peaks in "blank" electropherograms, accomplished with automatic peak marking and file-to-file peak correlation, as a means of providing zero readings at the appropriate points. Without this function, the peaks would not be marked and tracking of a single peak corresponding to bound or free would be difficult.

As a test of the program for larger data sets, the data set was duplicated five times to generate a total of 3600 electropherograms. This data set mimics the data load produced for a series of injections made every 10 s for 10 h. Analysis was completed in ∼20 min on a 400 MHz computer; however, analyzing the same set using a 2.4 GHz computer decreased the analysis time to <5 min.

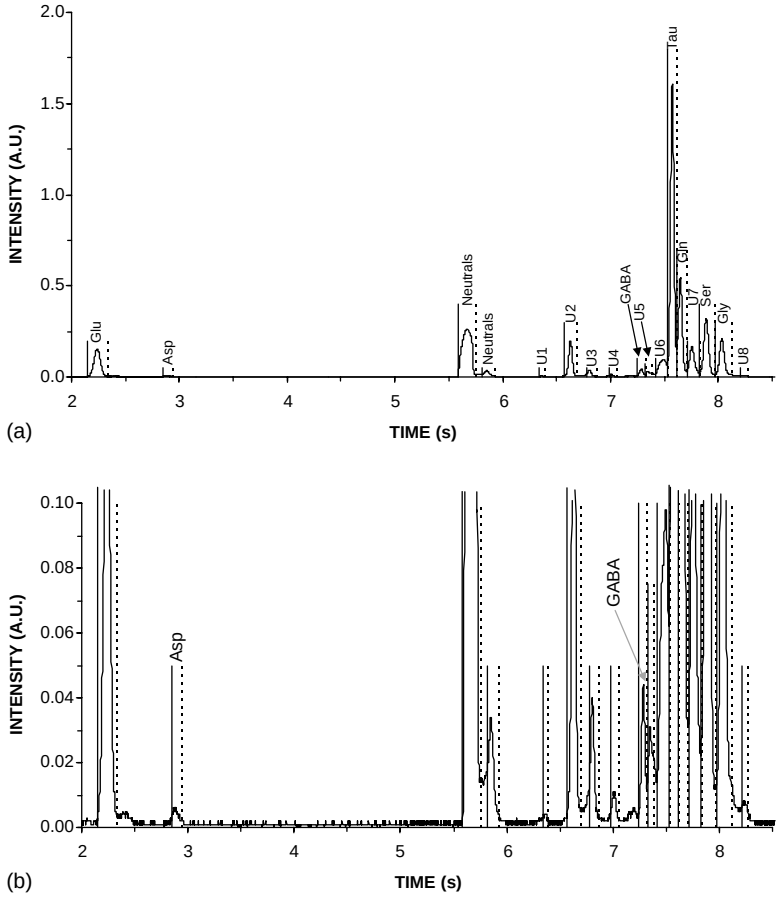

Fig. 7. Automatic peak marking for CE data monitoring in vivo neurotransmitters. (a) Peaks automatically detected in an in vivo electropherogram during a K^+ stimulation (full scale). Peaks were identified as: (1) Glu; (2) Asp; (3) neutral species; (4) GABA; (5) Tau; (6) Gln; (7) Ser; and (8) Gly. Unidentified peaks are denoted as U. Peak starts are marked by solid lines and peak ends by dashed lines. (b) Magnified scale of the same data to emphasize low level peak detection.

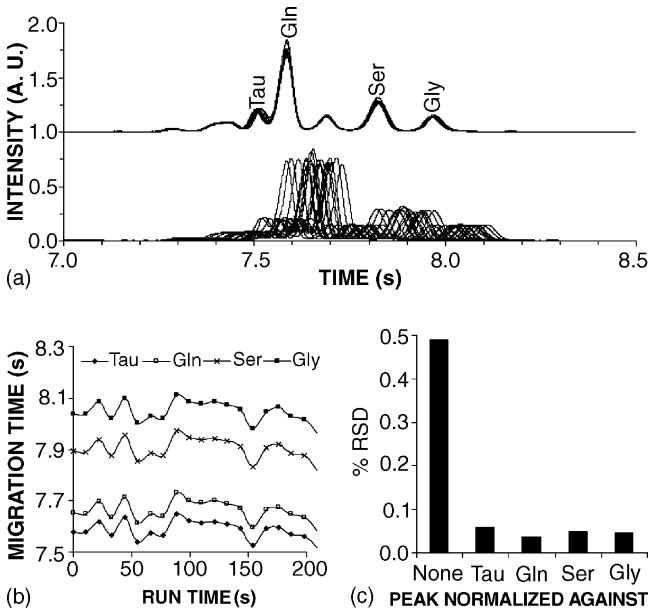

Fig. 8. Effect of temporal normalization upon migration time precision. (a) Overlaid plots of 20 basal in vivo electropherograms. The lower traces are the original data and the upper have been temporally normalized about the highest peak (Gln). (b) Actual migration times for the four major peaks prior to normalization. (c) Reduction in % R.S.D. of the migration times resulting from temporal normalization. % R.S.D. was calculated excluding the peak normalized against.

3.4. CE in vivo neurotransmitter analysis

Immunoassays typically present relatively simple electropherograms with just two peaks of interest that are well-resolved. To test the efficacy of Cutter with more complex data sets, the program was used to analyze electropherograms collected while monitoring neurotransmitter levels of live rats when neuronal secretion was stimulated by passing 75 mM K^+ through the dialysis probe. The separation was optimized for seven amines: glutamate (Glu), aspartate (Asp), γ-aminobutyric acid (GABA), glutamine (Gln), taurine (Tau), glycine (Gly), and serine (Ser) [2]. A representative electropherogram collected in vivo is illustrated in Fig. 7 with the locations of automatically marked peaks. (Peaks were identified from previous experiments based on migration times and spiking of samples.) In collection of this data electropherograms were overlapped, i.e. 9 s after one injection a second injection was made. As a result, late migrating peaks for Glu and Asp appear at the beginning of the following electropherograms as shown in Fig. 7. Overlapping injections, possible when slower migrating compounds will not co-elute with faster compounds in the following electropherogram as in this case, allow the temporal resolution of the monitoring experiment to be improved. Overlapping injections can cause minor complications for peak analysis as discussed below.

The effect of temporal normalization on migration time deviations was examined. Using 20 basal electropherograms

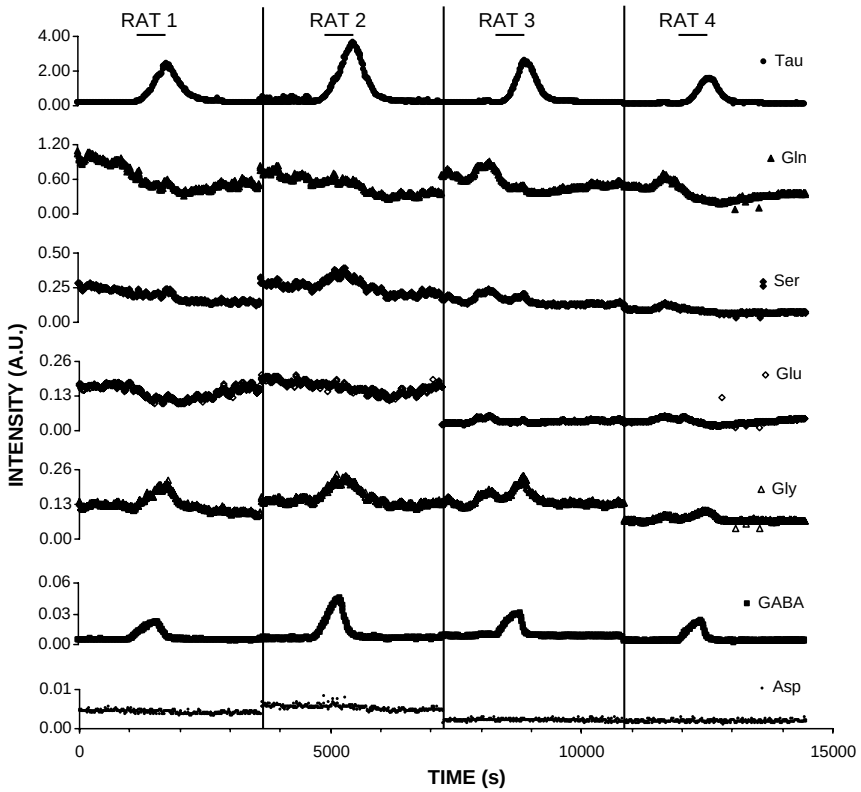

Fig. 9. Peak heights of seven amino acids collected from four rats during the course of K^+ stimulations (marked as horizontal lines). The total data set included 1300 electropherograms.

collected in vivo, the reproducibility of peak migration times prior to normalization was compared to that after normalizing to various peaks (Fig. 8). Non-normalized migration times shown in Fig. 8b displayed no apparent pattern to the deviations from file-to-file; however, peak shifts were correlated within a file. Although the migration time standard deviations of Tau, Gln, Ser, and Gly were low (∼40 ms) for the non-normalized data, the average relative standard deviations (R.S.D.) were improved approximately 10-fold with normalization (Fig. 8c). The improvement was independent of the peak used for normalization suggesting that a systemic error in the instrument, such as irreproducible file acquisition start times or the inherent mechanical deviation of the solenoid-controlled injection system, rather than variation in actual electrophoretic velocity is the dominant source of variability in migration time. Temporal normalization corrects for this error and enables simplified manual peak marking. For example, with the high degree of reproducibility shown in the normalized plot of Fig. 8a, it is possible to set fixed times as the start and end of each peak instead of marking peaks each electropherogram individually. As analyte migration times are further removed from the peak chosen for normalization, the errors in migration time are expected to be larger. In such a case, successive normalizations are required.

To test the analysis time for more complex data, a set of 1300 electropherograms (containing $\sim 1.2 \times 10^6$ data points), acquired during the course of four in vivo experiments were analyzed by Cutter. Peak heights for the seven identified amines as a function of time during the K^+ treatment are presented in Fig. 9. Peak marking was accomplished by first examining the Glu and Asp peak region separately from the rest of the data set. Electropherograms were first normalized to Glu, which was automatically marked. The low S/N for Asp at some points required that its boundaries be manually set, showing a limitation of the first derivative method of peak marking. (Manual set points were used across the entire data set.) The region of electropherograms containing the other peaks was then excised and the data was temporally normalized against Gln. The remaining peaks (Tau, Gln, Ser, and Gly) were automatically marked, except GABA, which was manually added due to its low S/N at basal levels. (The separate analysis for the acidic amino acids versus the others was necessary because of the use of overlapping injections (see above). In particular, temporal normalization was not effective since the Glu and Asp peaks in a given file were actually injected with the prior electropherograms.) CFOMs were calculated for this set within 5 min on a 2.4 GHz computer. The results were equivalent to those obtained with serial file analysis that required ∼20 h to complete.

3.5. Deconvolution of real data sets

To investigate the utility of deconvolution for increasing the accuracy of measurements on real data sets, standard electropherograms of Ser and Gly were analyzed (Fig. 10).

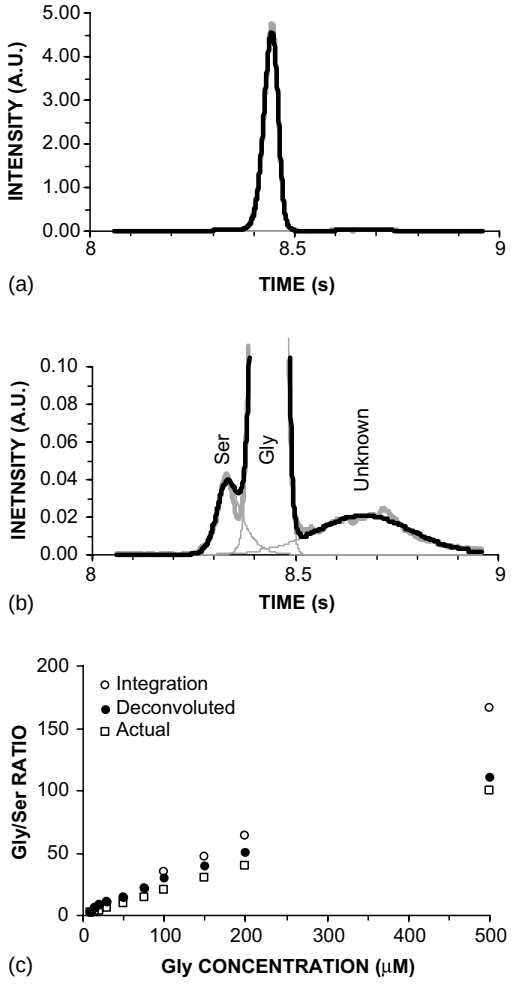

Fig. 10. Effect of deconvolution upon Ser and Gly concentration measurement accuracy. (a) Full scale plot of deconvolution of raw data (heavy grey line) via ETG fit at a Gly/Ser ratio of 100. The individual peaks, corresponding to Ser, Gly, and an unknown contaminant, are shown as thin grey lines and their sum resulted in the calculated signal (black line). (b) Magnified scale plot to emphasize the lower S/N peaks. (c) Plot depicting Gly/Ser ratios calculated from simple area integration (○), area after deconvolution (●), and the actual concentrations (□) vs. the actual Gly concentration.

For these experiments, Ser was held constant at 5 μM while Gly was varied from 10 to 500 μM. As Gly concentration is increased, the overlap of the peaks increases so that eventually the peaks are highly overlapped. Fig. 10a and b illustrate the appearance of the peaks at a 100:1 ratio of Gly to Ser that result in extensive overlap. Examination of Fig. 10b illustrates that a third, unknown peak also overlaps the Gly/Ser zones. Overlaid on these electropherograms are the deconvoluted peak shapes. As the ratio of Gly is increased, simple numeric integration of the peak areas resulted in ratios of Gly/Ser that were falsely high as illustrated in Fig. 10c. This effect was attributed to generating lower than true values for Ser and higher than true values for Gly. Deconvolution of the three peaks allowed for more accurate calculations of the areas, resulting in Gly/Ser ratios closer to that predicted

from actual concentrations (Fig. 10c). These results illustrate that the use of ETG deconvolution can improve the accuracy of peak characterization for moderately overlapped peaks.

4. Conclusion

Advances in data analysis methods are required to keep pace with the large data output and new applications of rapid separation techniques. Although many powerful applications are available for single file analyses, they are inefficient in terms of bulk data processing that might be produced from diverse applications of serial data analysis. The approach used here, encoded in Cutter, allows for the rapid analysis of large volumes of data to better match the time scale of the actual acquisition. The method is especially well-suited for long-term monitoring experiments wherein multiple compounds are continuously quantified over long periods. The new program has also successfully demonstrated the applicability of ETG to automated deconvolution of overlapped peaks. The use of deconvolution will be especially useful in chemical monitoring applications where the separation is well-characterized, that is the detected compounds are known yet overlap occurs. Further studies of the algorithms would be required to confidently apply them to cases where the electropherograms exhibit other anomalies such as severely distorted peaks or baseline spiking and drift; however, it is anticipated that the programs described here are most likely to be applied in high-throughput applications that typically have highly reliable separations conditions. Future work will also explore alternative peak detection algorithms that are better suited to more accurately characterize low S/N peaks and large dynamic ranges of peaks, such as second derivative methods [23], histogram methods [33], or Fourier analyses [34].

Acknowledgements

This research was supported by National Institutes of Health grants NS38476 and DK46960. We thank M. Bonner Denton (University of Arizona) and Daniel A. Gilmore (Spectral Instruments, Tucson, AZ) for LabVIEW assistance. J.G.S. acknowledges support from an Eastman Analytical Chemistry Fellowship.

References

[1] R.T. Kennedy, Chem. Rev. 99 (1999) 3081.
[2] M.T. Bowser, R.T. Kennedy, Electrophoresis 22 (2001) 3668.
[3] A.W. Moore, J.W. Jorgenson, Anal. Chem. 65 (1993) 3550.
[4] S.C. Jacobson, C.T. Culbertson, J.E. Daler, J.M. Ramsey, Anal. Chem. 70 (1998) 3476.
[5] M.L. Plenert, J.B. Shear, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 3853.
[6] K.K. Unger, M. Huber, K. Walhagen, T.P. Hennessy, M.T.W. Hearn, Anal. Chem. 74 (2002) 200A.
[7] T. Veriotti, R. Sacks, Anal. Chem. 73 (2001) 4395.
[8] C.A. Knaak, A.M.J. Hawrylechko, PSIT 1 (1998) 300.
[9] S.N. Walford, J. Chromatogr. A 956 (2002) 187.
[10] S.C. Jacobson, R. Hergenröder, A.W. Moore, J.M. Ramsey, Anal. Chem. 66 (1994) 4127.
[11] E.E. Jameson, J.M. Cunliffe, R.R. Neubig, R.K. Sunahara, R.T. Kennedy, Anal. Chem. 75 (2003) 4297.
[12] T.F. Hooker, J.W. Jorgenson, Anal. Chem. 69 (1997) 4134.
[13] R. Sacks, C. Coutant, T. Veriotti, A. Grall, J. High Resol. Chromatogr. 23 (2000) 225.
[14] M.W. Lada, T.W. Vickroy, R.T. Kennedy, J. Neurochem. 70 (1998) 617.
[15] P.S. Loupe, X. Zhou, M.I. Davies, S.R. Schroeder, R.E. Tessel, S.M. Lunte, Pharmacol. Biochem. Behav. 74 (2002) 61.
[16] M.G. Roper, J.G. Shackman, G.M. Dahlgren, R.T. Kennedy, Anal. Chem. 75 (2003) 4711.
[17] R.J. Marshall, A.J. Bleasby, R. Turner, E.H. Cooper, Chemom. Intell. Lab. Sys. 1 (1987) 285.
[18] J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, J.J. Baeza-Baeza, J. Chromatogr. 348 (1997) 187.
[19] J.W. Li, Anal. Chem. 69 (1997) 4452.
[20] J.W. Li, J. Chromatogr. A 952 (2002) 63.
[21] R.D. Caballero, M.C. García-Alvarez-Coque, J.J. Baeza-Baeza, J. Chromatogr. A 954 (2002) 59.
[22] G.H. Golub, Matrix Computations, third edition, Johns Hopkins University Press, Baltimore, 1996.
[23] A. Felinger, Data Analysis and Signal Processing in Chromatography, Elsevier, Amsterdam, 1998.
[24] H. Du, M.J. Stillman, Anal. Chim. Acta 354 (1997) 65.
[25] J.D. Ingle, S.R. Crouch, in: Spectrochemical Analysis, Prentice Hall, Upper Saddle River, NJ, 1988 (Chapter 5).
[26] E.E. Grushka, M.N. Myers, P.D. Schettler, J.C. Giddings, Anal. Chem. 41 (1969) 889.
[27] J.C. Giddings, in: Unified Separation Science, Wiley, New York, 1991 (Chapter 5).
[28] T.L. Pap, Z. Pápai, J. Chromatogr. A 930 (2001) 53.
[29] D.W. Marquardt, J. Soc. Ind. Appl. Math. 11 (1963) 431.
[30] A. Felinger, G. Guiochon, J. Chromatogr. A 913 (2001) 221.
[31] Y.F. Cheng, N.J. Dovichi, Science 242 (1988) 562.
[32] J.W. Li, Anal. Chim. Acta 388 (1999) 187.
[33] K.H. Jarman, D.S. Daly, K.K. Anderson, K.L. Wahl, Chemom. Intell. Lab. Sys. 69 (2003) 61.
[34] A. Felinger, E. Vigh, A. Gelencsér, J. Chromatogr. A 839 (1999) 129.