Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Chromatography A, 996 (2003) 141–155

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis

Kevin J. Johnson[a], Bob W. Wright[b], Kristin H. Jarman[b], Robert E. Synovec[a,]*

[a]*Center for Process Analytical Chemistry, Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA*
[b]*Pacific Northwest National Laboratory, Battelle Boulevard, P.O. Box 999, Richland, WA 99352, USA*

Received 10 January 2003; received in revised form 1 April 2003; accepted 2 April 2003

Abstract

A rapid retention time alignment algorithm was developed as a preprocessing utility to be used prior to chemometric analysis of large datasets of diesel fuel profiles obtained using gas chromatography (GC). Retention time variation from chromatogram-to-chromatogram has been a significant impediment against the use of chemometric techniques in the analysis of chromatographic data due to the inability of current chemometric techniques to correctly model information that shifts from variable to variable within a dataset. The alignment algorithm developed is shown to increase the efficacy of pattern recognition methods applied to diesel fuel chromatograms by retaining chemical selectivity while reducing chromatogram-to-chromatogram retention time variations and to do so on a time scale that makes analysis of large sets of chromatographic data practical. Two sets of diesel fuel gas chromatograms were studied using the novel alignment algorithm followed by principal component analysis (PCA). In the first study, retention times for corresponding chromatographic peaks in 60 chromatograms varied by as much as 300 ms between chromatograms before alignment. In the second study of 42 chromatograms, the retention time shifting exhibited was on the order of 10 s between corresponding chromatographic peaks, and required a coarse retention time correction prior to alignment with the algorithm. In both cases, an increase in retention time precision afforded by the algorithm was clearly visible in plots of overlaid chromatograms before and then after applying the retention time alignment algorithm. Using the alignment algorithm, the standard deviation for corresponding peak retention times following alignment was 17 ms throughout a given chromatogram, corresponding to a relative standard deviation of 0.003% at an average retention time of 8 min. This level of retention time precision is a 5-fold improvement over the retention time precision initially provided by a state-of-the-art GC instrument equipped with electronic pressure control and was critical to the performance of the chemometric analysis. This increase in retention time precision does not come at the expense of chemical selectivity, since the PCA results suggest that essentially all of the chemical selectivity is preserved. Cluster resolution between dissimilar groups of diesel fuel chromatograms in a two-dimensional scores space generated with PCA is shown to substantially increase after alignment. The alignment method is robust against missing or extra peaks relative to a target chromatogram used in the alignment, and operates at high speed, requiring roughly 1 s of computation time per GC chromatogram.

Keywords: Peak matching algorithm; Retention times; Diesel fuel; Principal component analysis

*Corresponding author. Tel.: +1-206-685-2328; fax: +1-206-685-8665.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).

0021-9673/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0021-9673(03)00616-2

1. Introduction

Gas chromatography (GC) is a powerful tool in the quantitative and qualitative analysis of complex mixtures. A typical high-speed gas chromatogram of a complex chemical mixture contains a large amount of chemical information, provided in a relatively short amount of time. Analysis of chromatographic profiles, generally with the goal of making a classification of one sort or another, is known as ‘‘fingerprinting’’. A large-scale fingerprinting project, such as quality control of a process stream over a long period of time by chromatographic analysis of samples, may involve the systematic comparison of hundreds or thousands of chromatographic profiles acquired many days apart, possibly from different GC instruments.

Chromatographic data analysis has traditionally centered either on resolving the peaks of interest and quantifying them through a peak area calibration or visual comparison of more complex profiles for subjective pattern matching. Using these methods for analysis, retention time precision was less crucial for success. Within the past two decades, however, it has become entirely feasible to perform complex mathematical manipulations on entire chromatograms via the use of inexpensive desktop computers. This has opened the door for the application of chemometric analysis techniques to the analysis of chromatographic data. The issue of run-to-run retention time variation, however, has been a significant impediment in the development of chemometric analysis techniques for chromatographic data. Chromatographic data exhibits uncertainty along the time axis due to unavoidable variations in instrument parameters. In GC, for example, temperature and flow fluctuations from run-to-run, matrix effects from sample-to-sample, and stationary phase composition changes from column-to-column (as well as column degradation over time) all lead to variations in retention time for a given analyte over a series of many chromatographic runs. Analogous factors lead to the same phenomena in other separation methods.

Multivariate chemometric analysis techniques are generally constructed to interpret chemical variation as changes from sample-to-sample in the magnitude of the detector signal at corresponding variables (i.e., retention times). Therefore, chemometric techniques are inherently sensitive to retention time precision. For example, two chromatograms offset from each other in retention time, but identical in every other respect would tend to be incorrectly interpreted by most chemometric analysis techniques as being chemically different. To accurately ascertain differences in chemical composition between samples from their chromatograms, there must be a one-to-one correspondence between the variables of the two objects being compared. That is, the chromatographic peak for a given analyte must have the same retention time in all chromatograms analyzed. In particular, Malmquist and Danielsson provide a discussion on the sensitivity of principal component analysis (PCA) to retention time shifting in chromatographic data [1]. Bahowick and Synovec report on the effect of retention time precision on quantification of liquid chromatographic data using classic least-squares regression [2] and Poe and Rutan report on its effect on quantification of liquid chromatographic data with diode-array fluorescence detection by generalized rank annihilation method (GRAM) [3]. Fraga et al. [4] and Prazen et al. [5] each discussed the effect of retention time precision on quantification of two-dimensional GC data by GRAM.

The impetus behind the development of a mathematical retention time alignment algorithm as a means of enabling chemometric analysis of chromatographic data lies in the relative cost-effectiveness of computer time over instrument time. With the current trend of rapidly increasing computational power at decreasing cost, it becomes financially advantageous to shift some of the burden of chemical analysis onto computational algorithms. This tendency has been evidenced by the rapid growth of the field of chemometrics over the past two decades. Furthermore, with the utilization of a retention time alignment utility, it becomes more feasible to include chromatograms acquired from different instruments, or even chromatograms acquired under different chromatographic conditions, vastly increasing the potential of chemometric analysis of chromatographic data for long-term, large-scale analysis projects.

Retention time shifting of chromatographic peaks occurs due to phenomena related to the instrument itself as well as to the chemical interactions between different samples and the instrument. For example, an offset in which every point in the sample chromatogram is shifted the same amount relative to the target chromatogram, although rare, can occur in some situations, and is most likely due to an injection-timing problem. More commonly, a stretch (or shrink) along one chromatogram's retention time axis relative to another occurs when a parameter is globally different between two chromatographic runs, such as flow-rate, temperature, or temperature program rate. A third mode of instrument-induced shifting is due to more rapid fluctuations in these same parameters during a chromatographic run that are irreproducible from run-to-run. Chemically caused shifting arises from selectivity changes as the stationary phase of the column degrades over time as well as from matrix effects and non-linear chromatographic conditions.

While instrument-related sources of shifting have been lessened for GC through the development of electronic control systems (such as auto-injectors and electronic pneumatic control) and chemically related sources of shifting can be minimized by carefully adjusting the conditions of the chromatographic separation and replacing columns before significant degradation occurs, there generally remains an appreciable amount of subtle run-to-run retention time variation in large sets of chromatographic data taken over longer time spans that cannot be corrected by means of a simple stretch or shift, i.e., the third mode of shifting just described. Correction of this subtle run-to-run retention time variation requires the development of retention time alignment algorithms such as the one reported herein.

The main issues to address in providing a useful retention time alignment algorithm are the following. First, the algorithm must preserve chemical selectivity differences between samples of differing composition while minimizing run-to-run retention time differences in the analysis of the samples. Second, the algorithm must provide retention time precision that is significantly better than that initially provided by the instrumentation. Third, the alignment algorithm must be fast in order to rapidly deal with a large number of data sets in a short period of time, thus providing high sample throughput.

The mathematical correction involved in retention time alignment involves some algorithmic means of adjusting the retention times of chromatographic peaks such that corresponding peaks from different chromatograms appear at the same retention time. Mayfield and Bertsch developed a method for the rapid analysis of jet fuel chromatograms in which the profiles were reduced to a peak area table [6]. With this method, the retention time corresponding to each detected peak is adjusted for retention time variation through the use of a standardization algorithm using approximately 10 ‘‘internal marker peaks.’’ Unfortunately, the peak area table method becomes less accurate as chromatograms become more complex and overlapped, leading to ambiguity in peak assignment and assessment of peak overlap. The complex profiles resulting from many overlapped chromatographic peaks may confound retention time assignments and area calculations and provide unreliable values for each, making algorithms dependent on peak area calculations less useful for fingerprinting applications. Thus, in the case of complex chromatograms, it is advisable to correct retention times throughout the entire chromatographic profile, rather than the retention times of individual peaks.

Towards this end, Andersson and Hämäläinen created a simplex-optimized chromatographic profile alignment utility in which the retention time shifting is modeled by a linear equation containing a constant term for offset and a slope reflecting the degree of stretch or shrink relative to the target chromatogram [7]. Optimization of these terms for best alignment was based on correlation between target and sample chromatograms in two retention time windows. Malmquist and Danielsson's alignment algorithm involves four rounds of iterative shifting to optimize sample-to-target correlation [1]. First, shifting of the entire profile to correct for a scalar offset in retention times between the sample and target, then of a smaller subset of the largest peaks for coarse peak-to-peak retention time shifting. Fine-tuning is then accomplished with iterative shifting of a slightly larger subset containing more peaks, and a fourth step that involves non-integer shifting of a smaller subset of peaks. The non-integer shifts required for each peak in the subset are retained and used to construct a time displacement function that will correct the sample chromatogram's retention times. Nielsen et al. developed a method known as correlation optimized warping, or COW [8]. In this method, the chromatographic profiles to be aligned are divided up into a series of retention time ‘‘windows’’ of equal length. Alignment occurs through the systematic, iterative stretching and shrinking of the sample chromatogram’s retention time windows so as to find a combination of stretches and shrinks that optimizes sample-to-target correlation. Unfortunately, retention time alignment utilities based on correlation optimization through iterative stretching and shrinking of the chromatogram can be quite time-consuming for large data sets of long chromatograms. Further, complex correction schemes can also be unnecessary, provided the chromatograms have been collected using properly operated, state of the art GC instrumentation.

In the work reported herein, a simple and fast peak-matching algorithm will be shown to provide significant retention time correction for effective chemometric analysis of diesel fuel gas chromatograms that exhibit modes of retention time variation uncorrectable by global shift functions such as linear scaling, or ‘‘rubber band’’ stretch of the retention time axis. The chromatographic alignment utility described is shown to successfully address the issues of preserving chemical selectivity, providing high sample throughput, and enhancing retention time precision for the analysis of large data sets of diesel fuel chromatograms acquired with state-of-the-art, electronically controlled gas chromatograph instruments. The alignment algorithm is then successfully applied to the pattern recognition of diesel fuel samples. Another key benefit of the work is that additional knowledge is gained regarding the nature of within-run retention time variation that is observed using the electronic pressure-controlled GC instrument. The work reported herein focused on the analysis of one type of sample. However, it is expected that this alignment approach should be promising for use with other types of chromatographic data in which the mode of retention time shifting is similarly dominated by small, stochastic shifts influenced by varying instrument parameters that, if left uncorrected, raise havoc with the subsequent chemometric data analysis.

2. Algorithm

The fundamental task in retention time alignment is to locate matching features between the sample and target chromatograms. The retention time axis of the sample chromatogram is then altered so that these matched features occur at the same time in both axes. The mathematical correction involved in retention time alignment involves some algorithmic means of adjusting the retention times of chromatographic peaks such that corresponding peaks from different chromatograms appear at the same retention time. The chromatogram to be aligned is generally referred to as the ‘‘sample’’ chromatogram and the chromatogram to which it is aligned is referred to as the ‘‘target’’ chromatogram.

In the peak-matching retention time alignment approach developed here, the chromatographic peaks themselves are the corresponding features between the sample and target chromatograms. The alignment process requires corresponding peaks from chromatogram-to-chromatogram to not be offset (i.e., misaligned) in retention time by more than the typical distance between adjacent peaks. If this requirement is not met, a course alignment would need to be applied prior to the peak-matching alignment (as will be demonstrated in the second study). A target-chromatogram peak is matched to the sample-chromatogram peak that has the nearest retention time to it, within a given threshold distance along the retention time axis. The regions between peaks in the sample chromatogram are stretched or shrunk by interpolating more or fewer points in such a way as to force the peaks in the second chromatogram to occur at the same retention times as the target peaks with which they have been matched. A flow chart describing the alignment process is shown in Fig. 1.

The peak matching retention time alignment process is preceded by a baseline subtraction to correct for baseline offset and drift from run-to-run. A baseline offset is calculated as a best-fit line through a set of predefined baseline points. In this case, the baseline offset was calculated for and subtracted from each chromatogram using the signal from the first and last 2 s of that chromatogram. While it may appear that the center portions of the diesel fuel chromatograms suffer from severe baseline drift, this is not the case. An examination of the baseline at the beginning and end of the chromatograms demonstrates very small drift, if any, between the two

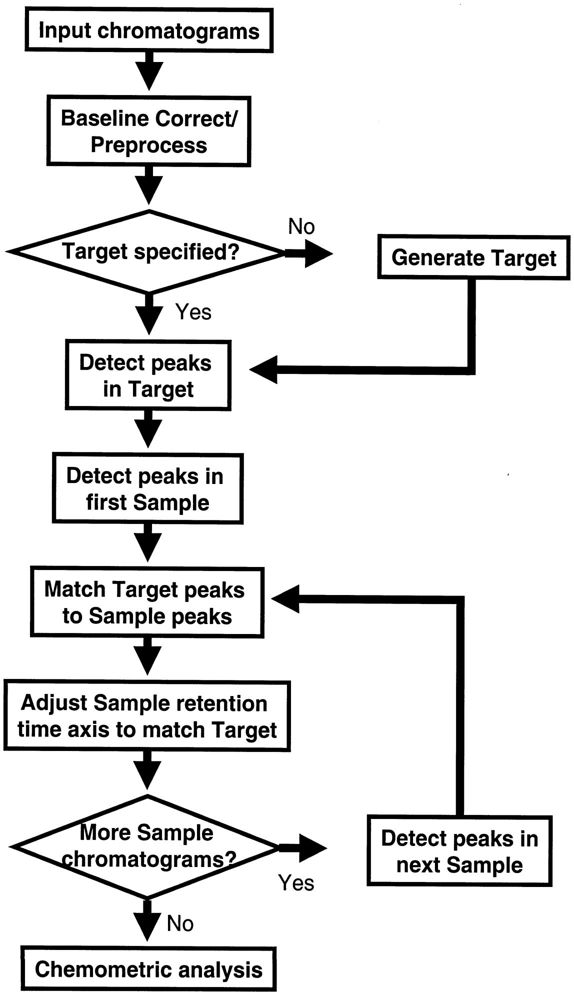
Input chromatograms
Baseline Correct/ Preprocess
Target specified?
No
Generate Target
Yes
Detect peaks in Target
Detect peaks in first Sample
Match Target peaks to Sample peaks
Adjust Sample retention time axis to match Target
More Sample chromatograms?
Yes
Detect peaks in next Sample
No
Chemometric analysis

Fig. 1. A flow chart diagram of the peak matching alignment algorithm.

regions. The apparent baseline increase in the middle is due to the fact that these chromatograms are comprised of a great many overlapped peaks, with only the most prominent components being observed as independent peaks in the central portion. Regardless, the baseline correction described was applied in order to remove a chemically unimportant source of variation in the data prior to chemometric analysis, and not necessarily to enhance the performance of the peak-matching retention time correction algorithm. This portion of the alignment program is modular, and an alternative method of baseline correction could be substituted quite easily, if desired. Additionally, extra preprocessing steps following baseline correction could be included, such as a peak area normalization to account for injection volume variation.

An alignment target chromatogram must be chosen carefully. It is important for a chromatographic peak in the target chromatogram to be located as close as possible to the center of the distribution of peak positions for that peak in the set of chromatograms to be aligned. It is also desirable for the target to contain, as nearly as possible, all of the peaks contained by the rest of the chromatograms in the data set to be retention time aligned. If a target with peaks on the edge of the distributions of peak locations (i.e., a chromatogram that is significantly shifted relative to the others in the set) is chosen, the likelihood that a peak mismatch will occur increases. Thus, due to the risk inherent in automatically choosing one of the chromatograms as a target, an alignment target is generated rather than arbitrarily being selected from the data set. The alignment target is generated as a trimmed mean of all of the chromatograms. A trimmed mean, for the purposes of this discussion, is a mean that disregards a certain number of extreme chromatograms in a set. For this work, a mean that discarded the highest and lowest tenth of the data was chosen. In this way, extreme outliers in the chromatographic data set would have little effect on the generated alignment target.

In the first step of the alignment process (Fig. 1), chromatographic peaks are identified in both the sample and target chromatograms and a list of their retention times is generated. The peaks are located automatically by finding zero crossings in an estimate of the chromatogram's first derivative. The algorithm operates on one chromatographic profile at a time and functions by stepping through each point in the chromatogram and calculating an estimated derivative as the difference between signal strength of the current point and the point that preceded it. Theoretically, the maximum difference observed in the baseline noise will be approximately four times the standard deviation of this noise. When this difference increases beyond a threshold set to a value of five times the standard deviation of the baseline noise, the algorithm has detected the leading edge of a chromatographic peak, and begins looking for a zero crossing in the estimated derivative. The next retention time at which the estimated derivative is equal to zero is calculated with interpolation, rounded to the nearest integer time point in the chromatogram, and saved to a list of peak retention times. Location of peaks with small signal-to-noise ratios is problematic due to the possibility of either a peak falling below the peak-finding threshold or an increased incidence of detection of false zero crossings. Finally, due to the manner in which the derivative of the chromatogram was estimated, the performance of this approach to peak detection is dependent on the sampling rate of the chromatogram. As the sampling rate increases, this particular peak finding algorithm will detect fewer of the smaller peaks in a complex chromatogram. At the sampling rate of 20 Hz utilized in this study, this method was found to adequately detect the peaks of each chromatogram. No special detection method is employed for ‘‘shoulders’’ of larger peaks. Accordingly, some are detected and some are not, depending on their size relative to the parent peak. An alternative peak finding method, such as locating peaks as local maxima within windows of a certain size and above a certain threshold, could be used. This portion of the algorithm is completely modular.

Next, the sample and target chromatograms are compared by stepping through each of the peaks in the target chromatogram and finding the sample chromatogram peak that most closely matches it in retention time. If the closest match is within a selected distance from the retention time of the target peak, then the peaks are matched. If there is no match within this selected peak matching window width, then the peak is assumed to not be present in the sample and is not used in retention time alignment, thus allowing alignment of chromatograms with different numbers of peaks. Selection of the optimum peak matching window width is critical and occurs during the step in Fig. 1 in which sample peaks are matched to target peaks. The retention time axis of the sample chromatogram is then adjusted through interpolation so that matched peaks have the same retention times. This is accomplished by interpolating more or fewer data points between chromatographic peaks in order to expand or constrict the retention time axis, respectively. A Matlab implementation of this alignment algorithm is available at http://synoveclab.chem.washington.edu.

A common metric for determining alignment quality of similar chromatographic profiles is the correlation coefficient between them [2,8,9]. The Pearson correlation coefficient between two chromatograms, X and Y with N points each, is calculated as:

$$\text{Coefficient} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \tag{1}$$

This correlation coefficient indicates the degree to which the two chromatograms are linearly related. Two chromatograms that are identical to each other have a correlation coefficient of one. Retention time differences between chromatograms, as well as actual chemical differences (i.e., changes in peak height) will lead to correlation coefficients of less than one. Thus, an improvement in chromatographic alignment will lead to a relative increase in correlation between chromatograms, but will not necessarily result in a perfect correlation. Prior to calculating the correlation coefficient between two chromatograms, a Wallis filter was temporarily applied to each chromatogram in order to minimize the effect of varying peak heights between the two chromatograms, in the same manner as Nielsen et al. [8] The effect of the Wallis filter is to adjust the mean and variance of local regions of a chromatogram to arbitrary values. The width of this local area is dictated by the width of the filter. The filter used in this work had a width of 41 data points (20 points/s), meaning that the data 1 s preceding and following each point are used to scale that point. Chromatograms were normalized to maximum intensity prior to application of the filter. The local mean was set to zero and the variance to one.

To evaluate the impact of retention time alignment with the peak matching alignment algorithm, two sets of diesel fuels were studied with prior and post-alignment analysis by PCA. PCA is a common method for reducing multivariate data sets to their lowest dimensionality [9]. PCA functions by projecting the original multivariate data onto a set of orthogonal axes defining a subspace of the original multivariate data space that maximally describes the variation contained within that multivariate data. The axes are known as principal components and are arranged in descending order of the amount of variation in the original data they explain. Each principal component has an associated loadings and scores vector. The loadings vector defines the direction of the principal component axis in the original measurement space, and thus describes how strongly each original measurement contributes to that principal component. The scores are the projection of the set of original measurements for each sample onto that principal component, and give information about the relative difference between samples in the reduced measurement space of that principal component.

3. Experimental

Two sets of diesel fuel GC data were examined. The first set of chromatograms was generated from the analysis of 20 diesel fuel samples that were acquired from eight commercial fuel terminals, as shown in Table 1. Samples acquired from the same fuel terminal were collected from different storage tanks or fuel delivery racks at the terminal, or from incoming fuel shipments from refineries to the terminal. Each sample was subjected to three replicate temperature-programmed GC runs, generating a total of 60 chromatograms. The second set of chromatograms consisted of data generated from the analysis of 21 diesel fuel samples, each collected from one of six different fuel terminals. Two replicate chromatograms were generated for each sample, the second replicate being acquired 1 day after the first, for a total of 42 chromatograms. The two data sets were unrelated, and shared no common fuel source terminals.

GC analyses were performed using Agilent Model 6890 gas chromatographs with flame ionization detection (FID). Fused-silica capillary columns 30-m×100-μm I.D. 100% methylpolysiloxane were used. Split injections at 500:1 were made of the diesel fuels samples. A routine temperature program was used. The chromatograms were imported from Chemstation (Agilent Technologies) into Matlab 6.1 (The Mathworks) where the alignment and subsequent chemometric analyses were performed. Each chromatogram was loaded into a Matlab workspace as a vector composed of the time series of FID detector readings over the duration of that particular GC run. Each chromatographic run was 17 min long with FID readings acquired at a rate of 20 Hz (a 50-ms sampling interval), yielding 20 400 points per chromatogram. For either data set, the chromatograms to be examined were stacked into a matrix of which each row consisted of a separate chromatogram. This matrix was submitted to the retention time alignment algorithm, aligning the chromatograms to an internally generated target as was previously described. Alignment of 60 chromatograms took approximately 1 min and 15 s, or about 1.25 s per chromatogram, utilizing a standard IBM clone personal computer equipped with a 1.2-GHz processor, 512 megabytes of RAM, and the Microsoft Windows 2000 operating system. By contrast, alignment by the COW utility (available at

Table 1
Diesel fuel samples and their terminals of origin

Sample	Chromatograms	Origin	No. of peaks (SD)
1	1–3	Terminal A	318 (26)
2	4–6	Terminal B	308 (12)
3	7–9	Terminal B	301 (5)
4	10–12	Terminal B	308 (4)
5	13–15	Terminal B	303 (3)
6	16–18	Terminal B	304 (18)
7	19–21	Terminal C	304 (23)
8	22–24	Terminal C	304 (8)
9	25–27	Terminal D	302 (5)
10	28–30	Terminal D	280 (19)
11	31–33	Terminal E	324 (7)
12	34–36	Terminal E	309 (13)
13	37–39	Terminal F	310 (21)
14	40–42	Terminal G	302 (15)
15	43–45	Terminal G	328 (13)
16	46–48	Terminal H	300 (13)
17	49–51	Terminal H	319 (13)
18	52–54	Terminal H	326 (12)
19	55–57	Terminal F	315 (18)
20	58–60	Terminal F	319 (18)

Samples from the same terminal were collected from different storage tanks or fuel delivery racks except for samples 10, 14, 15, and 16, which were collected from incoming fuel shipments to their respective terminals. Samples 14 and 16 were collected from the same oil tanker on the same day, but at two different terminal locations. Each sample was analyzed in triplicate by temperature programmed GC as indicated by chromatogram number. The mean number of peaks detected per chromatogram as well as the standard deviation are reported in the last column.

http://www.ibt.dtu.dk/mycology/cow/cow.htm) of the same 60 chromatograms took approximately 1 h and 53 min, or 113 s per chromatogram. Thus, the alignment method reported here executed roughly 100 times faster than the COW utility.

4. Results and discussion

In the first study, 20 diesel fuel samples with three replicate GC runs each were analyzed by PCA, both with and without implementation of the alignment algorithm. (See Table 1 for more details regarding the samples examined.) A typical temperature programmed gas chromatogram is shown in Fig. 2. An attempt to analyze the raw data by PCA is shown in Fig. 3. The replicate chromatograms of each sample are labeled with the number of that sample according to Table 1. Although the replicates should be chemically the same as well as distinct from those of other samples from different sources, this is not seen in the scores plot. Rather, intra-replicate precision is shown to be poor enough to obscure actual chemical differences between samples. Although the chromatograms appear aligned when overlaid and viewed in their entirety, closer inspection reveals significant retention time shifting between the chromatograms. In order to characterize the retention time variation present in the data set, a subset of the 60 chromatograms were subjected to manual peak identification and matching. To minimize the possibility of incorrectly matched peaks, the chromatograms chosen for this subset were from three replicate runs each of five diesel fuel samples originating from the same source (samples 2–6, terminal B in Table 1). Since the 15 chromatograms chosen for this subset represent samples coming from the same source, there is a minimum of chemical variation present in the data, making reliable manual identification of matching peaks in the data set possible. This subset of 15 replicate chromatograms is depicted in Fig. 4A, where the retention time shifting is clearly visible. In Fig. 4B, a different view of the same data is presented as a plot of peak position versus chromatogram number. In this plot, the location of each peak is marked with a black bar. The standard deviation of the peak position multiplied by four is graphically represented as 4σ (denoted as 4s in Figs. 4B and 5D), along with the nominal peak-to-peak distance, *D*, and the prospective peak-matching window width, *W*.

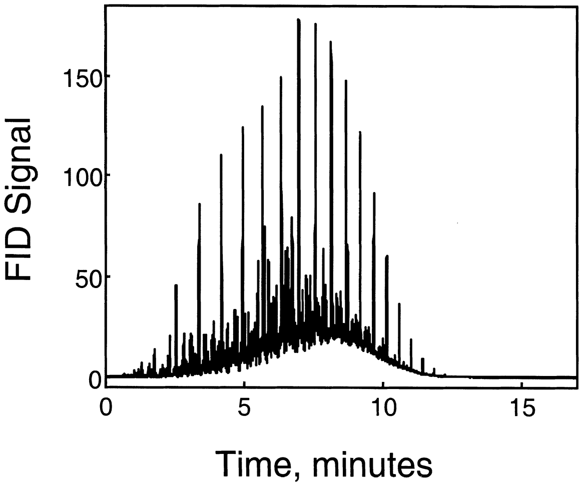

Fig. 2. A typical temperature-programmed gas chromatogram of a diesel fuel sample.

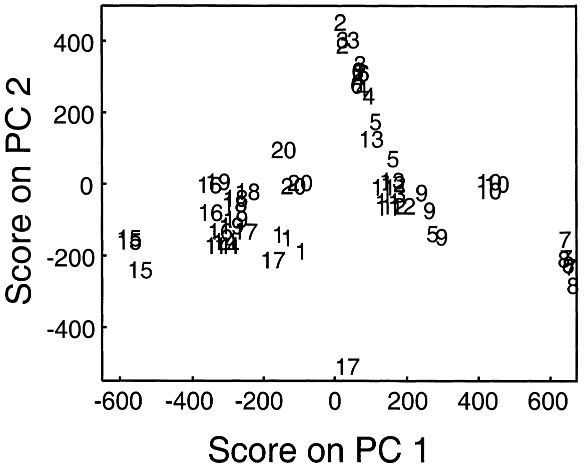

Fig. 3. Scores plot from PCA of raw diesel fuel chromatograms. Replicates are labeled by sample number according to Table 1. Retention time variation has led to relatively poor precision among replicate chromatograms in this scores plot.

An examination of these parameters, 4σ, *D*, and *W*, in this data subset of 15 replicate chromatograms from samples 2–6 is provided in Fig. 5. Manual peak matching was accomplished by plotting an overlay, similar to that shown in Fig. 4B, of the 15 chromatograms, locating a particular peak in each of the

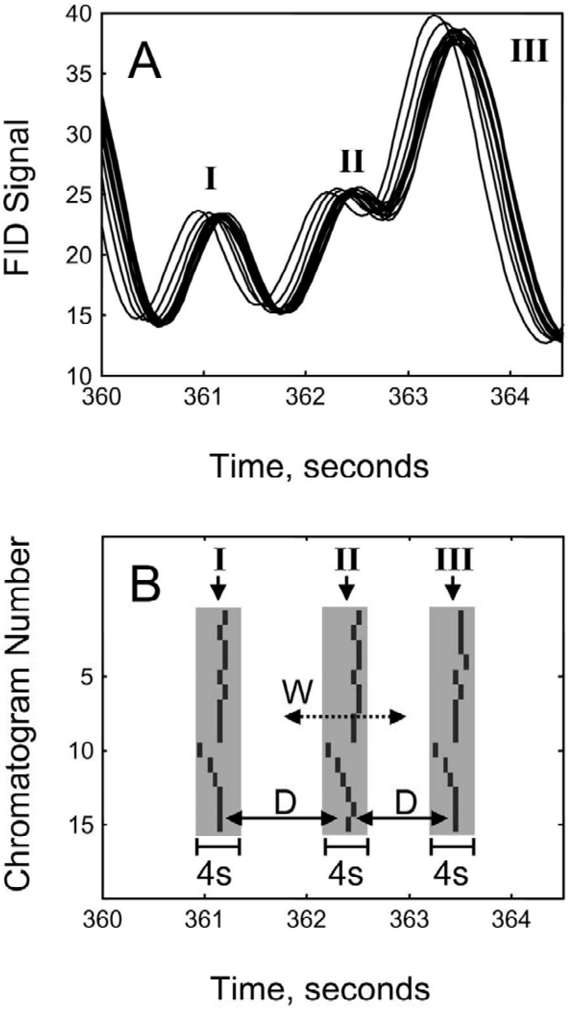

Fig. 4. (A) Although the chromatograms appear aligned when viewed in their entirety, the enlarged view of three peaks from samples 2–6 (Table 1, all terminal B) demonstrates a significant amount of retention time variation is present in the chromatographic data set. (B) A second view of the same region marks the location of each of the three peaks in each chromatogram with a black bar. Quantities of interest are the standard deviation of peak location across a set of chromatograms, σ, the distance between adjacent peaks, *D*, and the peak-matching window width to use, *W*. Note, 4σ is represented as 4s in Figs. 4B and 5D.

chromatograms, and recording that peak's retention time in a table. The result of this process is a table in which each row corresponds to a different chromatogram and each column corresponds to a different chromatographic peak. Three hundred fifty-seven (357) peaks present in all 15 chromatograms were manually located and tabulated for each chromatogram. Fig. 5A is a plot of the observed deviations from a trimmed-mean generated target chromatogram for each peak in three of the 15 chromatograms examined. A complex shifting pattern is observed, despite the electronic pressure and temperature control afforded by the Agilent 6890 GC. The shifting patterns observed generally do not follow a simple function, such as a ‘‘rubber band’’ type linear stretch or shrink of the time axis, and accordingly, retention time correction algorithms that are based upon fitting function parameters will not be successful at correcting this type of retention time variation. The standard deviation, σ, of each of the detected peaks across the set of all 15 chromatograms is shown in Fig. 5B. For most regions of the time axis, this value is on the order of 75 ms. At a nominal retention time of 8 min, the nominal standard deviation in retention time of 75 ms corresponds to a relative standard deviation percentage of 0.016%. It is noteworthy to point out that even with this impeccable retention time precision provided by the electronic pressure controlled GC, it was not sufficient to result in acceptable pattern recognition performance as demonstrated in Fig. 3. Typical peak-to-peak distances between adjacent chromatographic peaks in the data set were also examined. In order to do this, the automatic peak-finding algorithm was applied to each of the chromatograms, and a list of each of the distances between adjacent peaks was compiled. A histogram of adjacent peak-to-peak distances is shown in Fig. 5C. In these 15 chromatograms, the approximate range of retention times for each peak (taken as four times the standard deviation, 4σ) was on the order of 300 ms, while peak-to-peak distances of partially or fully resolved peaks discernable to the peak find step of the alignment algorithm were distributed around roughly 1.5 s. Note that at a peak-to-peak distance *D* of 500 ms, the chromatographic resolution is 0.25 since the average peak width is about 2 s. Thus, occurrences of peak-to-peak distances less than 500 ms dropped to a mean of less than one occurrence per chromatogram, primarily a consequence that the peak find portion of the alignment algorithm generally interpreted overlapped peaks with a resolution of 0.25 or less as if they were one peak. Indeed, for the purpose of pattern recognition applications it is appropriate that peaks that are essentially unresolved

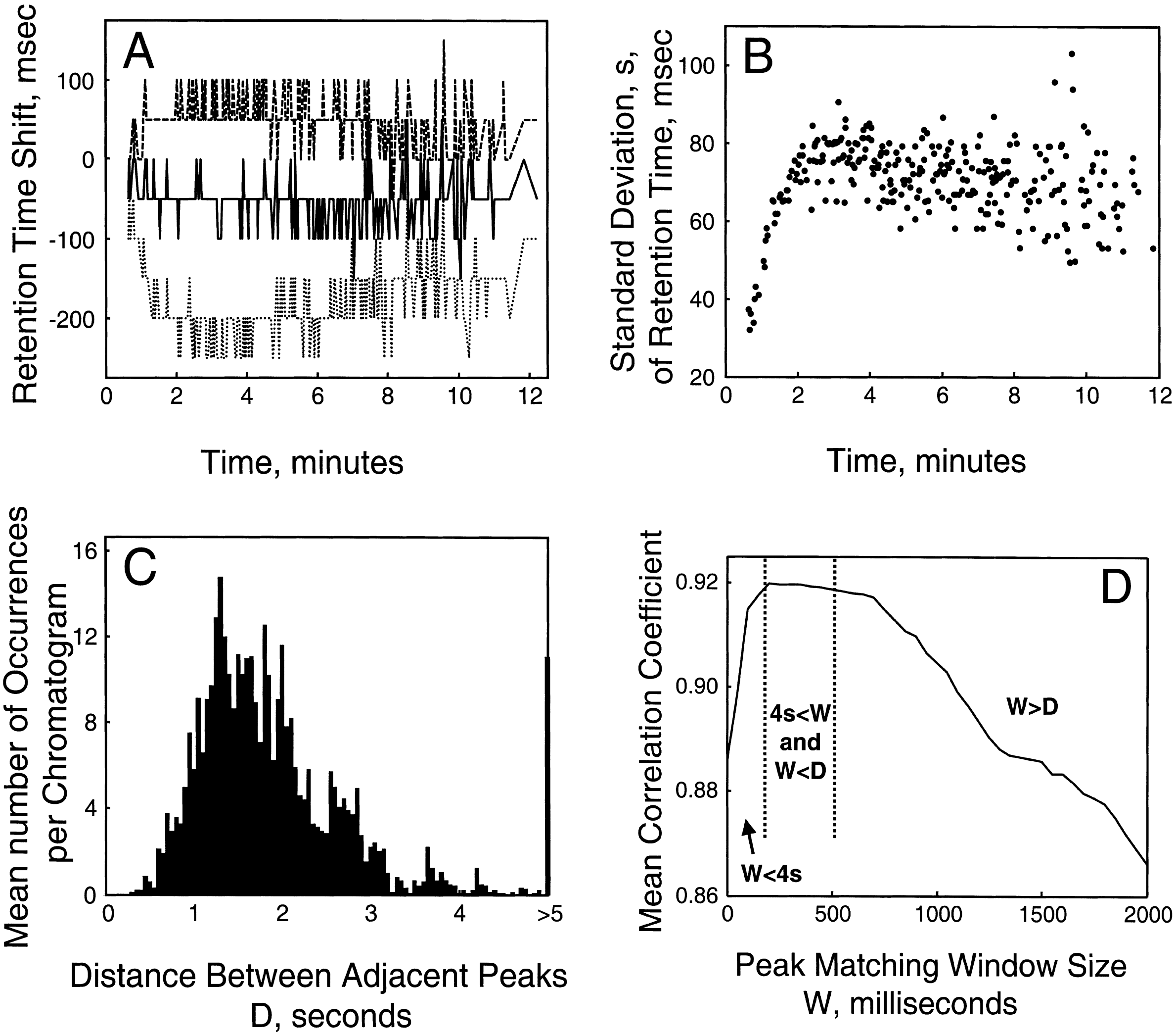

Fig. 5. Observed retention time shifts in 15 replicate diesel fuel chromatograms (samples 2–6, Table 1, all from terminal B). Three hundred and fifty-seven peaks present in all 15 chromatograms were manually located and tabulated for each chromatogram. (A) A plot of shift observed relative to the target chromatogram versus retention time for three of the 15 chromatograms. (B) The standard deviation of the retention time of each peak across the set of 15 chromatograms plotted versus the mean retention time value for that peak. (C) A histogram of peak-to-peak distances in the set of 15 diesel fuel chromatograms. Peak-to-peak distances are plotted on the x-axis versus mean number of occurrences per chromatogram of that distance on the y-axis. (D) Optimization of peak matching window width in the set of 15 diesel fuel chromatograms. Mean correlation of the set of 15 diesel fuel chromatograms with the target chromatogram is graphed versus peak matching window width used in the alignment. Correlation is optimized for a window width of 300 ms.

be treated as one peak. This indicates that retention time alignment by automated peak matching is appropriate for this data, as the variations in retention time (300 ms or less) are significantly less than the discernable peak-to-peak distances (500 ms or more).

An optimum peak matching window width was calculated by successive alignment and correlation calculations while varying *W*, the peak matching window width. The peak-matching alignment algorithm was used to align the subset of 15 chromatograms with a peak match window width varying from 0 to 2 s in steps of 50 ms. After each alignment, a correlation coefficient (utilizing a Wallis filter) was calculated between each chromatogram and the alignment target. The mean correlation coefficient for each peak window width used is plotted in Fig. 5D. The optimum peak-matching window width occurred at a *W* of 300 ms, which concurs with the degree of shifting observed in Fig. 5B. In the plot in Fig. 5D, three distinct regions are visible. In the first region, the peak matching window width, *W*, is less than the magnitude of the retention time variation (expressed by four times the standard deviation of the distribution of retention times for a typical peak across the set of chromatograms). In the second region, the peak matching window width is of roughly the same magnitude or slightly more than that of the retention time variation, but it is less than the typical peak-to-peak distance, *D*, observed in the chromatograms. In the third, the peak matching window width is greater than both the retention time variation and the peak-to-peak distance. From the graph, it is apparent that the peak-matching window width that provides the optimum alignment is located in the second region of the graph in Fig. 5D and is greater than the magnitude of the retention time variation, but less than the typical distance between adjacent peaks observed in the chromatogram.

Next, the entire data set of 60 chromatograms was analyzed. In the entire data set, we again see a subtle, yet significant, degree of shifting in a representative enlarged region shown in Fig. 6A. For clarity, a subset of 18 of the 60 chromatograms is shown. The chromatograms from sample groups 3, 4, and 5 are drawn with a solid line and the chromatograms from sample groups 16, 17, and 18 are drawn with a dashed line. From Table 1, we see that samples 3, 4, and 5 came from a common terminal (terminal B), as did samples 16, 17, and 18 (terminal H). While prior to alignment, the chromatograms in Fig. 6A do indicate some chemical differences between samples, the underlying similarities between samples acquired from the same terminal are masked by retention time variation. A peak-matching window width optimization was performed, this time using the entire set of 60 chromatograms and a target generated, as before, as a trimmed mean of the chromatograms to be aligned. The result, shown in Fig. 6B, is very similar to the optimization using the 15 chromatogram subset shown in Fig. 5D. Following alignment with the optimum peak matching window, the aligned profiles in Fig. 6C demonstrate the increase in retention time precision afforded by the peak matching alignment algorithm relative to the unaligned profiles shown in Fig. 6A. The aligned profiles much more clearly depict the chemical similarities between samples acquired from the same terminal as well as the differences between samples from different fuel terminals. Additionally, chromatographic peaks such as the one located to the far left of the interval shown in Fig. 6C are properly aligned, even though they do not occur in every chromatogram. As can be seen in Table 1, the number of peaks found in each chromatogram varied significantly, as well as the number of peaks found from sample-to-sample. Although a typical chromatogram had roughly 300 detected peaks, the difference in the number of peaks detected between pairs of chromatograms in the data set was as much as 76 peaks, or about 25%. This indicates that situations similar to those seen in Fig. 6C occur with some regularity. This capability of handling chromatograms with different numbers of peaks is possible both because the peak matching portion of the algorithm does not require every peak to be matched and also because the chromatograms are aligned to an internally generated target, rather than to any one chromatogram from the data set.

Finally, a scores plot resulting from PCA of the aligned data is shown in Fig. 6D. As with Fig. 3, each chromatogram is again represented by the sample group number with three replicates of each sample. In Fig. 6D, the improvement in retention time precision provided by the alignment prior to PCA has resulted in a substantial improvement in the scores plot relative to Fig. 3, which was calculated from unaligned chromatograms. A comparison between these two scores plots reveals a relative decrease in intra-replicate variance as an increase in the proximity of same-sample replicates on the scores plot of the aligned data relative to the unaligned data, as well as further demonstrating the sensitivity of PCA to retention time precision. Furthermore, the clustering of the samples in Fig. 6D is consistent with the sample origin information provided in Table 1. Again, it is important to note that the peak matching alignment algorithm was able to

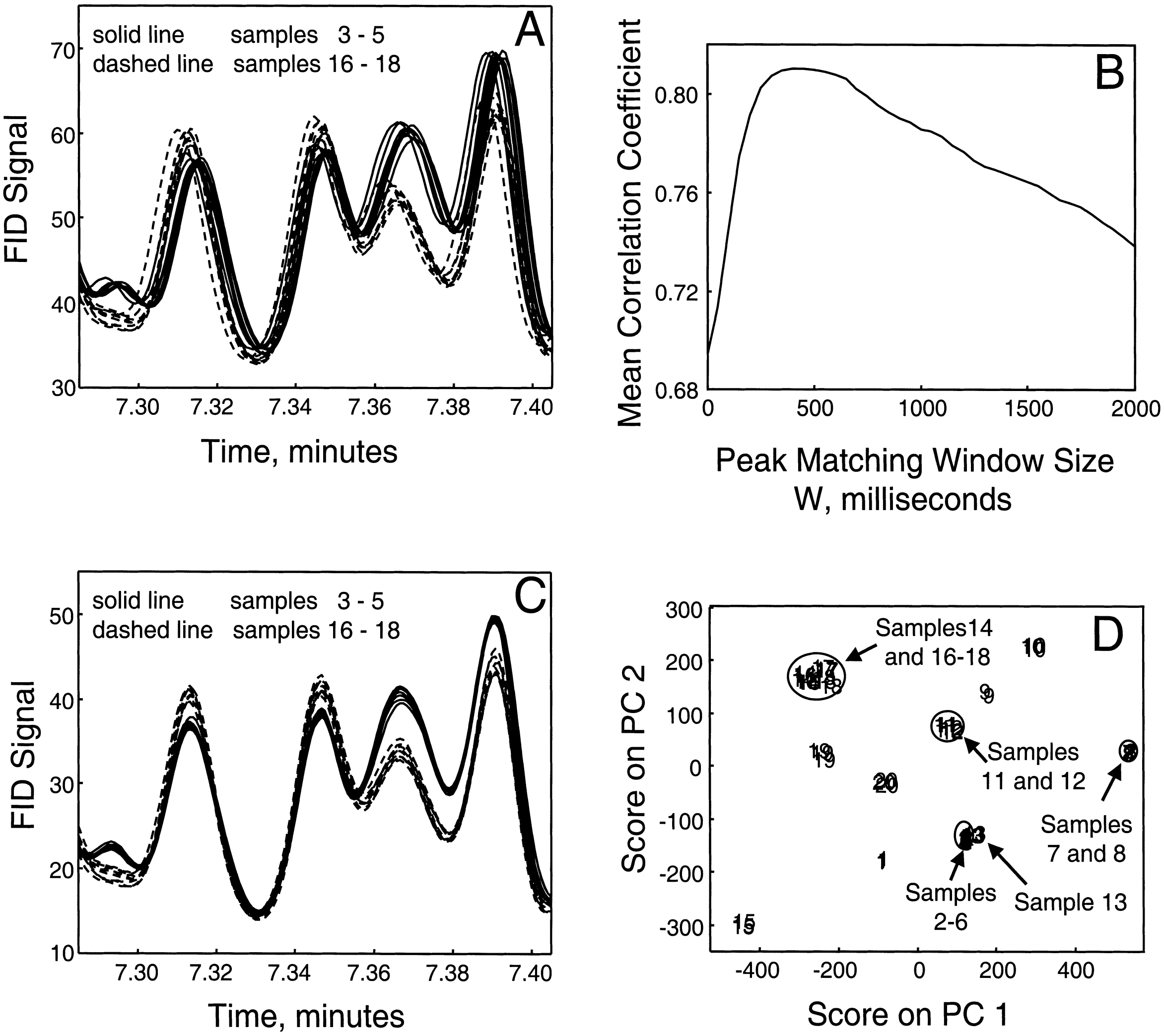

Fig. 6. Retention time alignment in the full set of 60 chromatograms (Study 1). (A) An expanded view prior to alignment of 18 of the 60 chromatograms, representing six of the 20 diesel fuel samples for clarity. Samples 3, 4, and 5 (solid line) are chemically similar to each other and chemically different from samples 16, 17, and 18 (dashed line). (B) A plot depicting the optimization of peak matching window width in the set of 60 diesel fuel chromatograms. Mean correlation of the set of 60 diesel fuel chromatograms with the target chromatogram is graphed versus peak matching window width used in the alignment. Correlation is optimized for a window width of 450 ms. (C) The same region shown in (A), after alignment using the optimized peak-matching window width. (D) A scores plot from PCA of all 60 diesel fuel chromatograms following alignment with the optimized peak matching window width. Samples are labeled according to Table 1. The improvement in retention time precision provided by the alignment has resulted in a corresponding increase in precision and information content relative to the scores plot in Fig. 3.

effectively enhance this PCA analysis despite the significant variations in the number of peaks detected in each chromatogram seen in Table 1. A corresponding PCA analysis of the same chromatograms aligned via the COW algorithm yielded essentially identical results, indicating that, in this example, the nearly 100-fold decrease in alignment time with the algorithm reported here came at no cost to the quality of the aligned chromatograms.

In cases where peak misalignments are on the order of or greater than peak-to-peak distances, peak matching alignment will not initially function properly. These chromatograms must be coarsely aligned first so that the general assumptions made in the peak matching algorithm are satisfied. In order to demonstrate this, a second set of 42 chromatograms, unrelated to the first set of 60 chromatograms, was examined. In this data set, large shifts in retention time were observed from chromatogram to chromatogram. The large shifting observed was facilitated by application of a slightly different set of operating conditions for the GC instrument. Fig. 7A displays an overlay of the second set of 42 chromatograms. These chromatograms demonstrate a greater degree of retention time shifting than seen in the previous data set studied, and provide a greater chromatographic alignment challenge. Comparing the single chromatogram in Fig. 2 to the overlay of 42 chromatograms shown in Fig. 7A, it can be observed that there are two distinct groupings of chromatograms in Fig. 7A: one group shifted about 10 s from the other along the retention time axis. A scores plot of the first two principal components taken from PCA of the mean-centered unaligned chromatograms shown in Fig. 7B demonstrates a clear separation between chromatograms from the first and second replicate sets on the first principal component, regardless of which terminal the chromatogram represents. No clear grouping according to sample terminal origin is evident, and different replicates from the same diesel sample appear as being chemically different due to retention time shifting.

The large, 10-s shift was corrected by locating and matching alkane peaks, using one replicate group to construct the alignment target. Following this coarse alignment, the chromatograms were further aligned using the peak-matching algorithm. An overlay of the shifts the alignment algorithm required to align the chromatograms displayed in Fig. 7A is plotted in Fig. 7C. Consistent with what is seen in Fig. 7A, two distinct groupings of chromatograms are visible: one set that required very little shift correction and one set that required shift correction on the order of several seconds. Fig. 7D shows a scores plot from PCA of the chromatograms after alignment. Compared to the scores plot in Fig. 7B, this plot demonstrates a greatly reduced separation between chromatograms of different replicate sets from the same terminal, although a slight demarcation between the two groups is still visible. Additionally, in this plot, groupings between chromatograms of samples originating from the same terminals are observed. Thus, the alignment utility has again substantially increased retention time precision while simultaneously preserving chemical selectivity in the subsequent multivariate analysis.

5. Conclusions

The peak-matching alignment methodology employed herein has been shown to effectively enhance a fingerprinting-style analysis of diesel fuel chromatograms. The clearest potential limitation of this technique is that it assumes the closest peak is the correct match. Thus, the technique relies upon the magnitude of the run-to-run shifting being less than the typical distance between peaks. The threshold for peak matching indicates the largest shift that will be allowed. Thus, theoretically, the peak matching window width should be set to a value just large enough to correct for shifting present, but not any larger in order to minimize the chances of a mismatch occurring. In cases where retention time shifting is severe, i.e., the retention time variation is greater than the typical adjacent peak-to-peak distance, it becomes necessary to either use a different alignment technique, or coarsely align the chromatograms prior to alignment by peak matching. This can be accomplished by utilizing retention indices, biomarker landmarks, or similar techniques to locate matching features between the chromatograms to provide at least a crude approximation to alignment. In cases where only minor retention time variation is present in a chromatographic data set, the peak matching alignment algorithm has been shown to provide capable retention time correction and selectivity preservation, at a substantial decrease in computation time. The work that has been described here is aimed at rapidly correcting subtle, yet significant, retention time shifting between chromatograms in order to expose subtle chemical differences between samples, rather than the alignment of greatly dissimilar chromatographic profiles. While it is like-

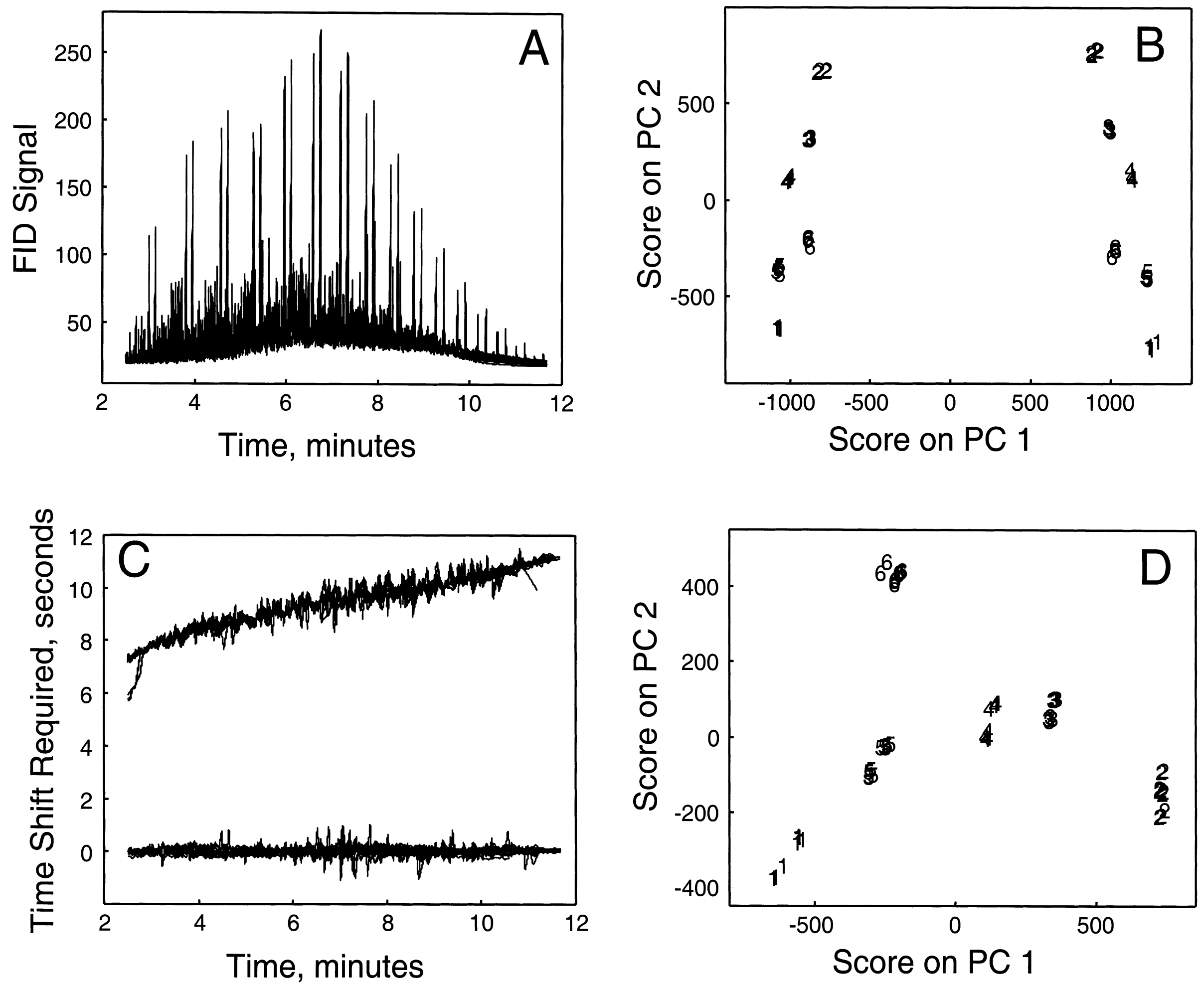

Fig. 7. An example where the magnitude of retention time shifting exceeds typical adjacent peak distances in a set of chromatograms (Study 2). (A) An overlay of 42 diesel fuel chromatograms, representing 21 diesel fuel samples from six different terminals. For each sample two replicate chromatograms were generated, the second 1 day after the first. Two distinct groupings of chromatograms shifted in retention time relative to each other offset by about 10 s are easily visible in the taller alkane peaks. (B) A scores plot from PCA of these diesel fuel chromatograms. Two groups of replicates for each sample are clearly visible, reflecting retention time shifting observed in (A). (C) An overlay of the retention time correction shifts required for alignment of the 42 chromatograms. The large, 10-s shift was corrected by locating and matching alkane peaks. Following this coarse alignment, the chromatograms were further aligned using the peak-matching algorithm. Again, two distinct groupings of chromatograms are visible: one set that required very little shifting and one set that required shifting on the order of several seconds. This indicates that one group displayed was used to construct the alignment target. (D) A scores plot from PCA of the chromatograms after alignment. After alignment, replicates from the same terminal are more closely grouped than those in (B), although a slight demarcation between the two groups is still visible.

ly, as in any chromatographic alignment technique, that alignment quality will suffer as the sample and target chromatograms become less similar to each other (i.e., having fewer and fewer chromatographic peaks in common), it is also true that the utility of using chemometric pattern matching techniques to probe for subtle chemical differences will likewise suffer. Although the alignment algorithm was demonstrated here for GC data, the same principles could be applied to other separation techniques such as liquid chromatography, capillary electrophoresis, and so on.

Acknowledgements

This work was supported by the Internal Revenue Service under a Related Services Agreement with the US Department of Energy under Contract DE-AC06-76RLO1830. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as the official Internal Revenue Service position, policy, or decision unless so designated by other documentation.

References

[1] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71.

[2] T.J. Bahowick, R.E. Synovec, Anal. Chem. 67 (1995) 631.

[3] R.B. Poe, S.C. Rutan, Anal. Chim. Acta 283 (1993) 845.

[4] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Anal. Chem. 73 (2001) 675.

[5] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[6] H.T. Mayfield, W. Bertsch, Comput. Appl. Lab 1 (1983) 130.

[7] R. Andersson, M.D. Hämäläinen, Chemom. Intell. Lab. Syst. 22 (1994) 49.

[8] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.

[9] E.R. Malinowski, D.G. Howery, in: Factor Analysis in Chemistry, Wiley, New York, 1991.