Talanta 83 (2011) 738–743

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Real-time target selection optimization to enhance alignment of gas chromatograms

Thomas I. Dearing[a], Jeremy S. Nadeau[a], Brian G. Rohrback[b], L. Scott Ramos[b], Robert E. Synovec[a,*]

[a] Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA
[b] Infometrix, Inc., 10634 E. Riverside Dr., Suite 250, Bothell, WA 98011, USA

ARTICLE INFO

Article history:
Received 27 July 2010
Received in revised form 14 October 2010
Accepted 18 October 2010
Available online 27 October 2010

Keywords:
Target selection
Alignment
Window
GC data
Real time analysis

ABSTRACT

An improved method for real-time selection of the target for the alignment of gas chromatographic data is described. Further outlined is a simple method to determine the accuracy of the alignment procedure. The target selection method proposed uses a moving window of aligned chromatograms to generate a target, herein referred to as the window target method (WTM). The WTM was initially tested using a series of 100 simulated chromatograms, and additionally evaluated using a series of 55 diesel fuel gas chromatograms obtained with four fuel samples. The WTM was evaluated via a comparison to a related method (the nearest neighbor method (NNM)). The results using the WTM with simulated chromatograms showed a significant improvement in the correlation coefficient and the accuracy of alignment when compared to the alignments performed using the NNM. A significant improvement in real-time alignment accuracy, as assessed by a correlation coefficient metric, was achieved with the WTM (starting at ~1.0 and declining to only ~0.985 for the 100th sample), relative to the NNM (starting at ~1.0 and declining to ~0.4 for the 100th sample) for the simulated chromatogram study. The results determined when using the WTM with the diesel fuels also showed an improvement in correlation coefficient and accuracy of the within-class alignments as compared to the results obtained from the NNM. In practice, the WTM could be applied to the real-time analysis of process and feedstock industrial streams to enable real-time decision making from the more precisely aligned chromatographic data.

Published by Elsevier B.V.

1. Introduction

The use of alignment for the preprocessing of chromatographic data is becoming increasingly prevalent and important. Alignment of chromatographic data is generally useful and often provides significant improvement in sample analysis for retention time (or Kovat Index) based analyses. For chemometric analysis, alignment is also an essential first step to remove retention-time shifts that are a normal part of any chromatographic experiment, in order to optimize the subsequent chemometric data analysis. In the context of this report, the alignment procedure can be thought of as occurring in two stages, the real-time selection of a target followed by the real-time application of the alignment algorithm. Once the appropriate target chromatogram has been selected, a given sample chromatogram of interest is aligned to the chosen target chromatogram. The alignment can be accomplished using various alignment software approaches (COW, piecewise alignment, etc). In this report the COW algorithm was employed [1,2], as implemented in the LineUp 3.0 software (as per Section 2). Alignment algorithms are generally designed to optimize the cumulative correlation coefficient between the target and sample chromatograms [1–8]. Whilst considerable research focus has been devoted to the advancement of the alignment algorithms, the methods by which alignment targets are selected has received much less attention, in particular for automated target selection, such as would be required for real-time analysis of chromatographic data such as for on-line process control and/or remote monitoring applications.

Selection of an appropriate target is of the upmost importance for the overall alignment procedure, as it has a significant influence upon the ability for chemometric algorithms to optimally glean useful information from the final aligned data. Selecting an optimum target allows the alignment algorithm to run quickly whilst avoiding the introduction of artifacts into the aligned data [5,7,9]. An optimum target should ideally be representative of all the chromatograms within the data set in terms of general features such as retention times, peak locations, peak shapes and peak areas. Furthermore for on-line, real-time applications these general features may not be known ahead of time, and thus, any target selection method must be adaptive and robust to account for the possibility of variation.

Typically the implementation of an alignment procedure takes place in an off-line capacity, commonly when all of the chromato-

* Corresponding author.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).

0039-9140/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.talanta.2010.10.026

graphic data for a particular application (system) has been collected [1–5,7–9]. This is an acceptable approach when dealing with closed finite systems, when the amount of time needed to collect all of the data is a fixed quantity. This is not the case for continual process systems (i.e., industrial, laboratory kinetic studies, environmental monitoring, etc) where it may be necessary to make real-time measurements. The data that requires alignment is being continually collected and thus must be aligned in real-time, and chemometric analysis in real-time, if decisions are demanded to be made in real-time. This means that the parameters for alignment, particularly target selection must also be determined in real-time. Real-time alignment coupled with chemometric data analysis facilitates the final goal of real-time decision-making, for example, allowing an operator to make decisions regarding product quality, feedstock formulating, or instrument status with regard to maintenance.

One of the most common off-line methods for alignment target selection uses principal components analysis (PCA) and a mahalanobis distance calculation to determine the target [9,10]. The *a priori* library of chromatograms is processed using PCA to produce the sample scores, and the mahalanobis distance of each of the sample scores is calculated. The target is generally determined to be the sample that has the smallest mahalanobis distance. The major advantage of this method is that, by using PCA and the mahalanobis distance, the sample which best represents the entire data set is selected as the target. However, this method requires collection of the entire data set before PCA can be employed, making it an unfeasible method for on-line continual processes or time-dependent systems where collecting all of the data prior to alignment is not possible. In these situations targets must be collected (i.e., determined and updated) as the system under study progresses. Furthermore, an appropriate real-time method for target selection must be sufficiently robust to handle variations in a process as data collection proceeds. To this end, Zhu et al. recently reported a method that allows for new target selection as a process continues [11], referred to in this report as the nearest neighbor method (NNM). The NNM target selection functions by aligning the nth chromatogram to the aligned form of the $n-1$th chromatogram. This method of target selection showed that an adaptive method for selecting targets could be successfully applied to an evolving process. This allows minor variation in a procedure or reaction process to be minimized in a final aligned system. However, this approach relies upon the quality of the previous chromatogram, an erroneous sample or process change will have an adverse effect on the alignment of future chromatograms.

In this report, we describe an adaptation and extension, building from the NNM target selection approach, which we refer to as the window target method (WTM). The window refers to the collection of aligned samples used to produce the target, and the target is created from the average of the aligned samples in the window. With the WTM, the window employed contains a number of previously aligned chromatograms from which a target will be generated. By increasing the number of chromatograms from which a target is generated, this has the effect of decreasing the chromatogram-to-chromatogram variation, and thus increases the overall accuracy of the real-time target selection and subsequent alignment procedure. Additionally, in this report we discuss methods by which these target selection procedures are judged. A series of comparisons are performed using the correlation coefficients of the respective target chromatogram to the aligned chromatogram and the first chromatogram collected to the aligned chromatogram. The results from this study will demonstrate that by calculating the correlation coefficient to the respective target, as defined by the NNM and WTM respective procedures, the resulting correlation coefficients can be misleading. Whereas, we shall demonstrate that the correlation coefficients determined from the first aligned chromatogram(s) relative to the newly aligned nth chromatogram give a more reliable indication for monitoring the accuracy of the alignment procedure and of the overall alignment process. Simulated chromatographic data as well as GC data from diesel fuel samples are used to develop and study the NNM and WTM target selection procedures.

2. Experimental

There were two sources of data used for the evaluation of the methods in this report: simulated and real. The simulated data study was produced using in-house algorithms, designed and run in MatLab 7.5. Alignment of all chromatograms was performed using LineUp 3.0 (Infometrix Inc., Bothell, WA 98011, USA). Alignment of the simulated data was performed using a slack of 3 and a segment size of 12. The real data was aligned using a slack of 2 and a segment size of 250. A total of 100 chromatograms were simulated. Each contained four analyte peaks, whereby each peak was Gaussian-shaped. Variation from one chromatogram to the next was introduced in two forms: a random shift in the peak-to-peak retention times greater than a single peak width and a uniform shift in the retention time applied to all peaks in the chromatogram. Please note that the use of a smaller segment size with the simulated data is possible because the simulated data contained significantly fewer data points, this reason also accounts for the application of a smaller slack value when aligning the real data.

The real data was a series of 55 chromatograms of four different diesel fuel samples, labeled generically as i (15 replicates), ii (15 replicates), iii (13 replicates), iv (12 replicates), that were collected on an Agilent 6890 GC with a 7683 auto-injector and FID detector (Agilent Technologies Inc., Santa Clara, CA, USA). The column used was a 10 m length by 100 μm inside diameter with a 0.1 μm film thickness (RTX) stationary phase (Restek, Bellefonte, PA, USA).

The two target selection algorithms were written and developed in MatLAB 7.5. Fig. 1 illustrates the procedural operation of the nearest neighbor method (NNM) [11]. This algorithm proceeds by collecting the first two chromatograms. The first chromatogram collected is the designated initial target; the second chromatogram is then aligned to this, forming a newly aligned chromatogram. The newly aligned chromatogram is then the target for the next chromatogram collected, and so on. Therefore, with the NNM, target selection functioned by aligning the nth chromatogram to the aligned form of the $n-1$th chromatogram.

The procedure for the window target method (WTM) is outlined in Fig. 2. It proceeds in a similar manner to the NNM, however, instead of beginning by collecting two samples the window method collects a larger number of chromatograms. These chromatograms form the first window from which the initial target will be selected. By selecting a larger group of samples (chromatograms), outlier impact and irrelevant variation can be minimized, and hence reduce the leverage upon the target selection process. From the samples in the window a mean chromatogram is calculated and used as a target for alignment. Following this the next chromatogram is collected and aligned to the target. The newly aligned chromatogram is moved into the window whilst simultaneously removing the oldest chromatogram in the window of chromatograms. This method avoids the issue of trying to classify the samples with a method like PCA and then removing the furthest sample from the pool. It has been shown in other reports about alignment that the classification of samples fails as the retention time imprecision increases [4,5]. The number of samples to include within the window was determined prior to the start of the alignment and target selection procedures. This occurred by performing the alignment procedure multiple times with increasing numbers of samples contained within the target window. The optimum window size was determined to be the number that resulted in the best overall alignment of the most dissimilar samples.

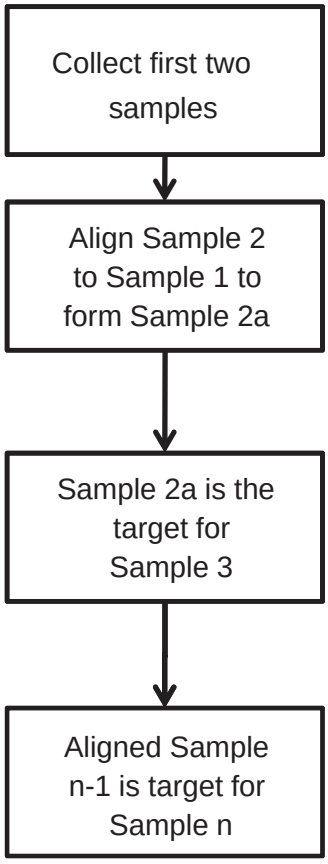

Fig. 1. Flowchart depicting the algorithm used to select targets using the nearest neighbor method (NNM).

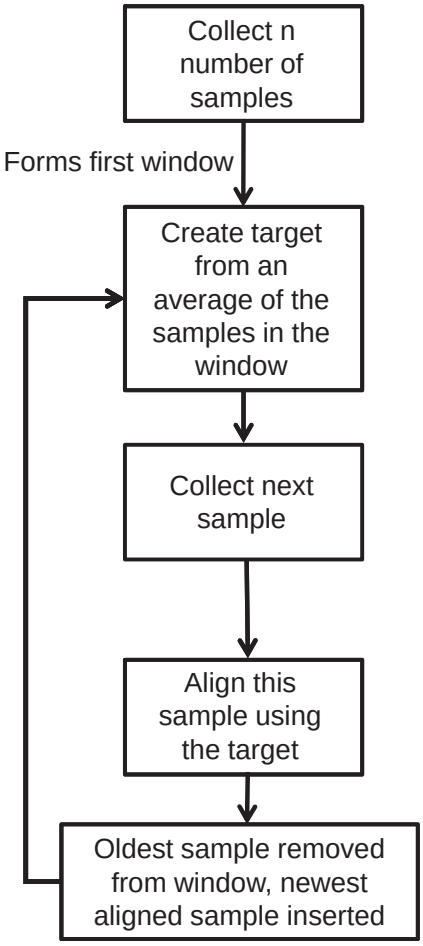

Fig. 2. Flowchart depicting the algorithm used to select a target using the window target method (WTM).

The final quality "metric" of the alignment procedure was determined using two approaches, in order to assess the performance and robustness of the NNM and the WTM in relation to each other. For one method, the correlation coefficient of the subsequent alignment to the respective target was determined and utilized as a metric. For the other method, and as it turned out much more illuminating in terms of gaining understanding, the correlation coefficient of the subsequent alignment to the first sample collected was also used as a metric.

3. Results and discussion

3.1. Simulated data study

The simulated data consisted of 100 chromatograms. Each chromatogram contained four resolved Gaussian peaks shown in Fig. 3A. There were two forms of variation in the data. The random normally distributed variation gave the data a realistic run-to-run variation seen in experimental data. The process shift (i.e., a drift) introduced gave the chromatograms a realistic type of variation typically seen in an on-line process over a number of days, weeks, or months. Because very limited effect has been seen from peak height in published results [1,2,4–7], chemical and random variations in the peak height were not included in the simulated data. Fig. 3B shows the variation (run-to-run retention time shifting) incorporated into the chromatograms. The simulated data spans a relatively narrow range of retention time shifting (~1 peak width at base), which is acceptable for samples run over a relatively short period of time. For longer periods of time, e.g., for separations collected over weeks to months or even years, more retention time variation could be observed, and this would warrant future investigation.

The number of samples (chromatograms) to be included within the target window (window size) was optimized with the results presented in Fig. 4. This procedure occurs by iteratively increasing the number of samples within the target window and comparing the resulting alignments of the first chromatogram to the last fifteen chromatograms (of the 100 chromatogram data set). The window size that results in the best correlation is determined to be the optimum. Fig. 4 shows the change in the average correlation coefficient of the final 15 simulated samples relative to the first simulated sample. The final 15 samples were utilized because they exhibited the largest amount of misalignment relative to the first few samples, in order to demonstrate the accuracy of the alignment. As the number of samples is increased the average correlation also increases until a window size of 11 samples. After reaching a window size of 11 samples, inclusion of more samples caused an overall decrease in the average correlation of the last 15 samples to the first sample. This may be due to the addition of irrelevant variation or samples that vary significantly when compared to the other samples within the window. Also highlighted in Fig. 4 is the result when the window size was one sample wide, this is the result for the NNM. This result shows that by increasing the number of samples in the window from one (i.e., which is applying the NNM) to just two samples, there is a doubling of the correlation coefficient. The inclusion of an extra sample allows for an additional minimization of the sample-to-sample leverage. To a great extent, Fig. 4 indicates that a window of only two samples is needed to apply the WTM with good success and by going to 11 samples, whilst optimum, results

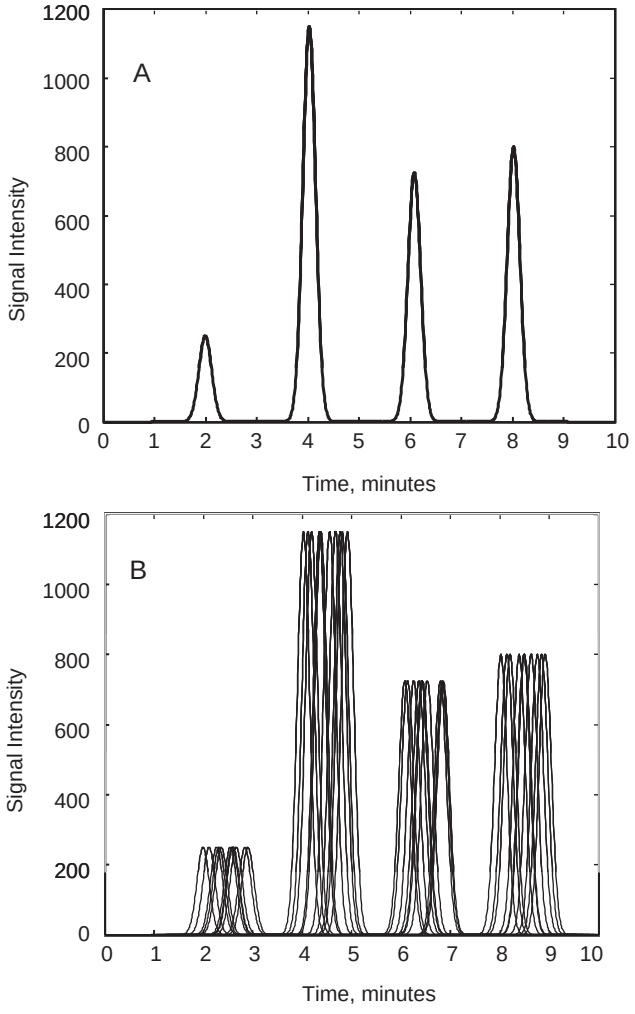

Fig. 3. (A) One simulated chromatogram, displaying the Gaussian peak shapes and height differences between each of the peaks. (B) A series of ten chromatograms (overlay) highlighting the chromatogram-to-chromatogram shifting as well as the peak-to-peak movement.

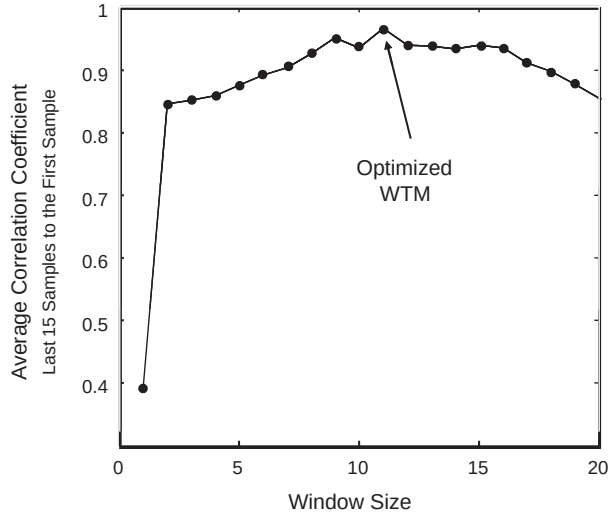

Fig. 4. Optimization of the window size for the simulated chromatograms (as per Fig. 3) using the average correlation of the last fifteen samples to the first sample with varying window size. The optimum for the WTM was determined to be 11 samples, although a significant improvement was observed by going from one sample to just two samples in the window.

in a modest improvement over the large gain from going from the one sample (NNM) to just two samples (WTM).

Optimized alignment was then performed on the 100 simulated chromatograms using each of the adaptive target selection methods. The correlations to the target for the NNM are shown in Fig. 5A. This figure shows that the correlation to target is excellent with all values being approximately 1.0. However, what appears to be good target selection and alignment, does not truly show accuracy in the alignment. To do this assessment, the correlation of each aligned sample to the first sample collected must be calculated, in order to see how well the chromatograms align more collectively. When examined via correlation to the first sample collected (also in Fig. 5A for the NNM), this more rigorous approach to assess correlation accuracy, showed a significant decline in correlation with the NNM as the process progresses and more samples are collected (starting at a correlation coefficient of ~1.0 and declining to ~0.4). This is due to an accumulation of small amounts of misalignment as new samples are aligned to a target that is not sufficiently robust. Comparisons of the correlations in Fig. 5A show that calculating the correlation to the target via the NNM may give the impression that the alignment process being observed is more accurate than it truly is and that calculating the correlation to the first sample collected gives the user a better metric for determining the long term accuracy and overall performance of the alignment procedure. Hence, the NNM does not appear to be sufficiently robust for this data set.

Examination of the correlations calculated for the same data set when using the WTM tells a different story. As with the NNM study (and Fig. 5A), the correlation coefficient was calculated for a given chromatogram in relation to the target, as well as to the first sample collected. The results are presented in Fig. 5B. In this case, the decline in the correlation coefficient for the WTM when compared to the first sample collected is essentially insignificant (dropping from about 1.0 to 0.985), as compared with the NNM (dropping from about 1.0 to 0.4). This comparison of the NNM and WTM is made side by side in Fig. 5C, where the correlations to the first target for each of the target selection methods are plotted. This figure clearly illustrates the better performance of the WTM compared to the NNM for the simulated data set.

3.2. Real data study

A series of 55 chromatograms of four different diesel fuels (i, ii, iii and iv) were collected over a period of four days. A typical chromatogram from one of the fuels is shown in Fig. 6. As with the simulated data, the number of samples to be included in the window was optimized. The procedure used was similar to that outlined using the simulated data. The correlation of a 13 sample subset to the first sample was determined to be optimum using the same method as presented in Fig. 4. However, unlike the simulated data study, as the number of samples contained within the window increased the average correlation to the first sample continued to increase until a plateau was reached. After a sample size of 13 there was no significant change in the average correlation. Since this is real data and has real sources of noise and variation by adding more samples into the window, this in turn means the target generated will contain more of this variation and thus be a more representative target for the alignment. However, the additional improvement means collection of a far greater set of data before alignment can be performed. In some situations where real time analysis is being employed this would be an unacceptable position.

Fig. 7 shows the correlation values calculated of the *n*th chromatogram in relation to the first chromatogram for the NNM and the WTM for the alignment of the diesel samples. The four separate types of fuels (i–iv) are evident, based upon the correlation of each category to the initial sample. When comparing the alignments for type (i) samples there is a declination observed in the corre-

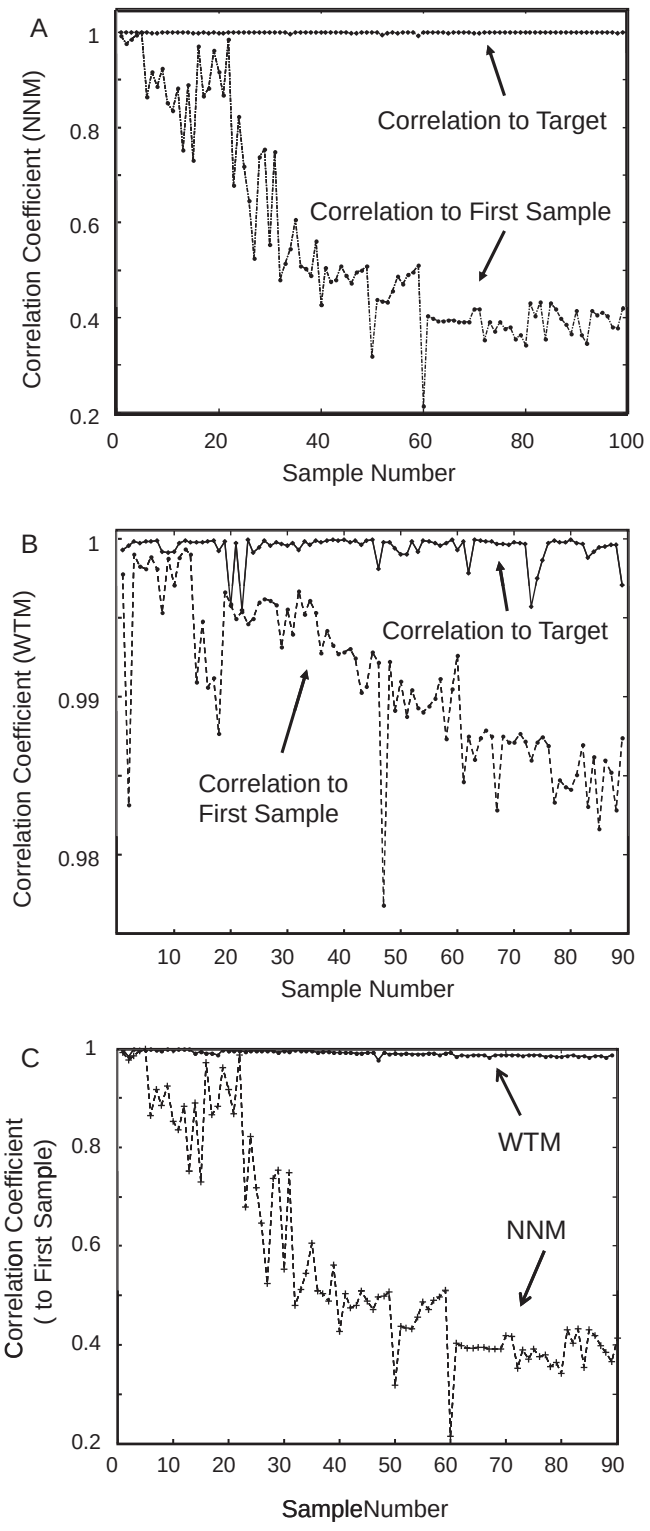

Fig. 5. (A) Using the NNM with the alignment of the 100 simulated chromatograms, correlation to the respective target (♦), i.e., the *n*th sample to the $n-1$th sample, and in contrast, correlation to the first sample collected (●), i.e., the *n*th sample relative to the first sample. (B) Using the WTM with the alignment of the 100 simulated chromatograms, correlation to the respective target (♦), and in contrast, correlation of the *n*th sample to the first sample collected (●). The correlation to the first sample for the WTM remains much higher than for the NNM. (C) Putting the key comparisons together, correlation relative to the first sample for the NNM (+) and for the WTM (●) using the 100 simulated chromatograms.

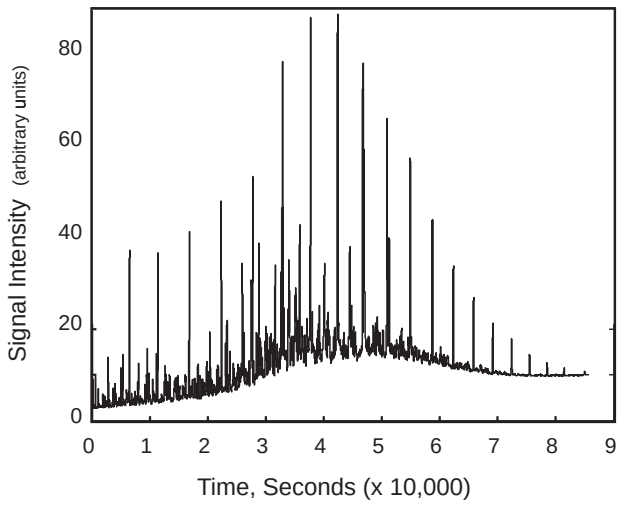

Fig. 6. A representative chromatogram from one of the four diesel fuel samples. These chromatograms have significantly more peaks than the simulated chromatograms.

lation coefficients for the samples aligned with the NNM. We note that the WTM produced a more accurate alignment of the chromatograms. However, the use of the window means that there is a slower response to a change of sample type. This delay is due to the time needed for the window to contain enough samples of the new sample type. As a result, the target generated by the WTM is likely to be more representative of the new sample type. By using the WTM with a suitable window, an erroneous sample will have a significantly reduced effect on the aligned data. The NNM reacts more promptly to the change in sample type but will also be more significantly affected by outlying samples. Furthermore, aligning to an erroneous sample using NNM will cause the incorrect alignment information to be incorporated in subsequent alignments. When analyzing the correlation coefficients of types (ii–iv) there are similar results as observed with type (i), namely, there is a declination in the correlation coefficient of the samples aligned using the NNM. The samples aligned using the WTM show greater accuracy in the correlation coefficients and thus the alignment for each sample type.

The accuracy of the alignment procedure is a concept that was apparent with both the simulated and real data types. The WTM

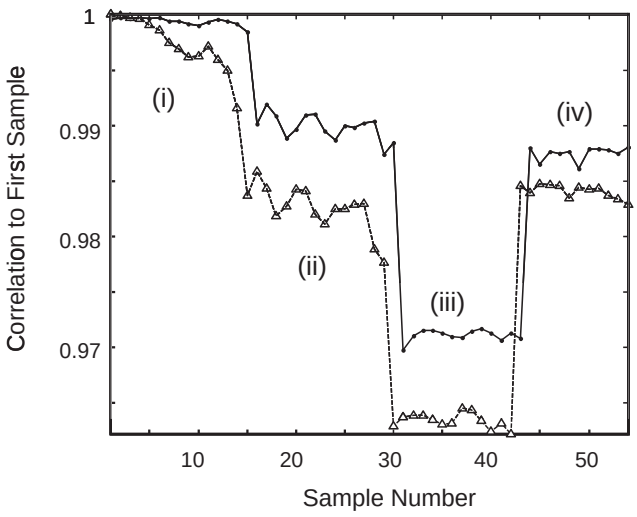

Fig. 7. Correlation to the first sample of the diesel GC chromatograms aligned using the NNM (△) and WTM (●) for the four diesel fuels i, ii, iii and iv.

produced a more accurate series of alignments than the NNM. This is a great advantage when using this method for the real-time analysis of chromatographic data. Furthermore, use of the WTM allows for a more robust system that can minimize the effect of outlying erroneous samples. However, using the WTM means that a user-selected number of samples must be collected before the alignment can proceed. In this regard, the approach in Fig. 4 demonstrates that a window size of two can cause a significant increase in the correlation coefficient compared to the NNM.

4. Conclusions

When applied with the simulated 100 chromatogram data set, the WTM showed a marked improvement of over twice the correlation coefficient and the accuracy of the alignment procedure, when compared to the NNM. In this regard, determining the correlation between the newly aligned sample and the first aligned sample highlighted the accuracy of the alignment procedure. This would not be observed if the correlation was determined between a newly aligned sample and the target (as defined by the method protocols in Figs. 1 and 2). It was also noted that by expanding the sample window from one sample (NNM) to two samples there was a significant improvement in alignment accuracy. When the WTM was applied to real GC data there was again an improvement in the correlation coefficient and the accuracy of within-group alignment (for the four diesel fuels).

Acknowledgements

T.I. Dearing and J.S. Nadeau are grateful for the funding from the Washington Technology Center and the Center for Process Analytical Chemistry.

References

[1] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[2] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[3] P.H.C. Eilers, Anal. Chem. 76 (2004) 404.
[4] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101.
[5] K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1141 (2007) 106.
[6] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, J. Chromatogr. A 1216 (2009) 4020.
[7] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemom. 20 (2006) 484.
[8] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y.V. Heyden, J. Chromatogr. A 1118 (2006) 199.
[9] M. Daszykowski, B. Walczak, J. Chromatogr. A 1176 (2007) 1.
[10] M. Fransson, S. Folestad, Chemom. Intell. Lab. Syst. 84 (2006) 56.
[11] L.F. Zhu, R.G. Brereton, D.R. Thompson, P.L. Hopkins, R.E.A. Escott, Anal. Chim. Acta 584 (2007) 370.