Journal of Chromatography A, 1218 (2011) 2584–2592

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Analysis of petroleum compositional similarity using multiway principal components analysis (MPCA) with comprehensive two-dimensional gas chromatographic data

G. Todd Ventura[a,b,*], Gregory J. Hall[c], Robert K. Nelson[a], Glenn S. Frysinger[c], Bhavani Raghuraman[d], Andrew E. Pomerantz[d], Oliver C. Mullins[d], Christopher M. Reddy[a]

[a] Woods Hole Oceanographic Institution, Department of Marine Chemistry and Geochemistry, Woods Hole, MA, USA
[b] University of Oxford, Department of Geosciences, Oxford, UK
[c] Department of Science,USA Coast Guard Academy, New London, CT, USA
[d] Schlumberger Doll Research, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 17 January 2011
Received in revised form 1 March 2011
Accepted 2 March 2011
Available online 11 March 2011

Keywords:
Multiway principal components analysis
MPCA
Oil
Fingerprinting
Comprehensive two-dimensional gas chromatography
Multivariate
Biomarker
Hydrocarbons

ABSTRACT

The accurate establishment of oil similarity is a longstanding problem in petroleum geochemistry and a necessary component for resolving the architecture of an oil reservoir. Past limitations have included the excessive reliance on a relatively small number of biomarkers to characterize such complex fluids as crude oils. Here we use multiway principal components analysis (MPCA) on large numbers of specific chemical components resolved with comprehensive two-dimensional gas chromatography-flame ionization detection (GC × GC–FID) to determine the molecular relatedness of eight different maltene fractions of crude oils. MPCA works such that every compound eluting within the same first and second dimension retention time is quantitatively compared with what elutes at that same retention times within the other maltene fractions. Each maltene fraction and corresponding MPCA analysis contains upwards of 3500 quantified components. Reservoir analysis included crude oil sample pairs from around the world that were collected sequentially at depth within a single well, collected from multiple depths in the same well, and from different depths and different wells but thought to be intersected by the same permeable strata. Furthermore, three different regions of each GC × GC–FID chromatograms were analysed to evaluate the effectiveness of MPCA to resolve compositional changes related to the source of the oil generating sediments and its exposure to biological and/or physical weathering processes. Compositional and instrumental artefacts introduced during sampling and processing were also quantitatively evaluated. We demonstrate that MPCA can resolve multi-molecular differences between oil samples as well as provide insight into the overall molecular relatedness between various crude oils.

1. Introduction

The accurate establishment of oil similarity is a longstanding problem in petroleum geochemistry and in the determination of reservoir architecture, which is perhaps the largest technical uncertainty in oil. Crude oils within a subsurface reservoir can be probed using geochemical methods. However, difficulties ensue because petroleum is a complex fluid that can comprise more than 10,000 different compounds [1–5]. In addition, tracing petroleum accumulations in the environment back to their origin is complicated due to physical (e.g. evaporation, emulsification, natural dispersion, dissolution and sorption), chemical (photodegradation) and biological (mainly microbial degradation) weathering processes [6]. It can thus be difficult to establish relationships between oils that are differentially weathered. Subsequently, the accurate measurement of oil similarity is needed for the assessment of reservoir connectivity [7], production allocation [8], and environmental forensics [9].

Although a vast array of methods can be used to compare fluid composition one of the most simple and common approaches depends on the comparison of integrated peak areas of two or more chemical compounds detected by gas chromatography–mass spectrometry (GC–MS) or flame ionization detection (GC–FID). This typically involves the comparison of biomarker ratios to assess the thermal history, depositional environment, and the type of organic matter that characterizes the source rock of a given petroleum sample [8]. By comparing biomarker ratios, the fingerprints of various

* Corresponding author at: IUPUI Department of Earth Sciences 723 West Michigan Street, SL118, Indianapolis, Indiana 46202, USA. Tel.: +1 508 289 2316.
E-mail addresses: Todd.Ventura@earth.ox.ac.uk, gtventura@yahoo.com (G.T. Ventura).

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.03.004

oils are generated and used to make inferences about compositional similarity between samples [8–12].

The limitation to this approach is that relatively few analytes are typically quantified in comparison to the number of compounds present in oil, and of these many are present in very low mass fraction and/or may be sensitive to specific types of overprinting. For this reason a variety of ratios are compared with one another. For example, the use of a variety of ratios may prevent a false declaration of two oils being dissimilar and therefore occupying separate compartments when these two oils are actually from the same compartment but have been biodegraded slightly differently. The accuracy of this approach is limited with respect to what compounds are chosen for analysis and how well the compound ratios that are compared accurately interpret the various chemical characteristics and history of the petroleum. Furthermore, analytical problems arise because oils are at least three orders of magnitude more complex than what is resolvable by GC–MS [1–5]. As such, many oil modifying processes can go undetected due to these myopic, comparative approaches.

Comprehensive two-dimensional gas chromatography (GC × GC) has vastly expanded peak resolving capacity relative to traditional gas chromatography [13–22]. However, the accurate integration of the larger number of GC × GC peaks has been cumbersome, and new data reduction methods are still evolving. Currently, the vast majority of compounds in these complex mixtures are still often ignored, and a tremendous amount of chemical information is unrealized. One potential method for overcoming this problem is to employ classification techniques for the binning of GC × GC data (i.e. [23,24]). Classifications can be configured to group analytes of similar compound classes across specific retention index ranges [24]. Such classification is difficult with traditional gas chromatography because compounds from different classes can have very similar retention times.

These simple data reduction methods are the focus of a larger array of statistical applications for chemical data referred to as chemometrics [25]. The field of chemometrics began in the 1970s [26] and is defined as the development and use of mathematical techniques to extract useful information out of data acquired through chemical analysis [27–29]. One chemometric technique widely used in petroleum geochemistry is principal components analysis (PCA). This unsupervised classification algorithm models inherent variations by decomposing a data matrix (the data set of many entire sample profiles) into its scores and loadings matrices, which are based upon the eigenvectors that model the greatest variance inherent in that data set [30]. Scores that cluster near to each other in principal component space are chemically more similar to each other than those with more distant scores. These geochemical data sets may comprise of quantified or normalized peak integrations or ratios of specific analytes for various samples.

More recently, multiway principal components analysis (MPCA) with GC × GC data has been demonstrated to hold promise as an exploratory data analysis technique [31–35]. GC × GC data sets produced from flame-ionization detection (FID) can be regarded as three-way [36,37] in which the individual GC × GC chromatograms are treated as objects (or samples) that describe a large number of variables (or peaks within the chromatogram). When all of the second-dimension chromatograms are stacked on top of each other, each data element is then indexed by first- and second-dimension retention axes, by sample number, and by FID response. MPCA enables the classification of these samples so that profiles of compounds within the GC × GC chromatogram can be discovered, which differentiate groups of samples. This decomposition of a multi-way array is developed as the product of a score vector and a loading array, where the score vectors have the same properties as those of ordinary two-way PCA [36]. As such, MPCA can be used to determine the molecular relatedness of complex chemical mixtures.

This multivariate technique requires highly reproducible and accurate retention time alignment for analytes in both GC dimensions, which has been particularly difficult to achieve, and various post-data processing methods have been developed to overcome the problem [38]. For example, Fraga et al. [39] developed an algorithm to align sub-regions of 2D separations along the primary time axis by interpolating the data, calculating the singular value decomposition, and interactively shifting a sample sub-region along a target chromatogram until a minimum percent residual variance was obtained. This algorithm was later adjusted to align both dimensions [40]. Alternative methods include correlation-optimized shifting algorithms based on the inner-product correlation for local subregions [41]; windowed rank minimization alignment with interpolative stretching between windows using set anchor points [42]; as well as affine transformations that match peak patterns between a peak template and target peak pattern [43]. Other approaches include dynamic time warping and correlation optimized warping (COW), which work for a broad range of chromatograms [44–46]. However, it may be possible to omit these transformations with modern GCs and strict protocols for data acquisition.

In this study, we test the effectiveness of MPCA to distinguish compositional similarities between various maltene fractions of crude oils without the use of the above post-data processing techniques. Six oil samples were analysed by GC × GC–FID (Table 1). Two sample pairs were collected *in situ* by a Modular Formation Dynamics Tester (MDT^{TM}) tool [47–49]. The first pair, labelled F-1 and F-2 was collected within the same well at the same depth and should constitute the same crude oil. The F-1 sample was injected three times to test the reproducibility of sample injection and data processing, which can yield statistical artefacts due to poor data registration caused by variations in GC conditions and matrix effects. The second pair, labelled PER-1 and PER-2, was collected within two different wells at a depth that was intersected by a potentially permeable sedimentary layer vertically offset by 658 feet. The chemical relatedness of these two sample pairs was previously assessed with different GC × GC fingerprinting techniques by Ventura et al. [24] that utilized compound classes and retention index ranges to group analytes for comparative analyses. Two additional crude oil samples were also analysed. One sample, labelled sample B, was obtained from an unknown location close to the F-1 pair and is thus from the same oil field. The second sample, collected from the North Slope, Prudhoe Bay, Alaska is analysed as an outlier. Both samples pairs and sample B were contaminated with varying amounts of olefin-based drilling fluids [20]. In this way, the samples selected for MPCA analysis are designed to test injection replication, differences stemming from sampling, and the ability to determine compositional similarity between spatially distant samples (Table 1).

Additionally, MPCA was performed on three sections of the GC × GC–FID chromatograms (Fig. 1). These areas were chosen in order to focus the analysis within discrete sections of the GC × GC chromatogram where specific compound classes of petroleum hydrocarbons elute. Region I represents the entire GC × GC–FID chromatogram. However, the oils analysed here are dominated by paraffins that are easily altered by many post-expulsion processes. The molecular differences with respect to these compounds can overload the statistical variance associated with the descending order of magnitude of the eigenvalues for the correlation matrix. Subsequently, the method validation also incorporated two separate regions of the GC × GC–FID chromatogram for analysis. Region II spans the chromatographic area of low to medium molecular weight aromatic hydrocarbons. Region III covers an area where sterane and hopane biomarkers elute. In this way, the comparison

Table 1
Experimental design.

Sample name	Sample location	Sample extraction	Experimental objective
F-1a-c	Same depth, same well as F-2	DFA	Injection replicate test
F-2	Same depth, same well as F-1	DFA	Sampling replicate test
B	Different depth, same well as F-1 and F-2	DFA	Stratigraphic drill hole variability test
PER-1[a]	Different depth, different well, same stratigraphic unit	DFA	Oil connectivity test
PER-2[a]	Different depth, different well, same stratigraphic unit	DFA	Oil connectivity test
Exxon, North Slope	Prudhoe Bay, Alaska	Surface sample	Outlier

[a] Samples labelled P-1 and P-2 in Ventura et al. [24].

of chromatograms regions were used to assess variations in actual total mass for each sample due to sample preparation, statistical problems potentially arising from retention time drifts and misalignments, contamination, and source related differences between the crude oils.

2. Methods

2.1. Oil sample acquisition

Four of the five samples were acquired from an open hole well using the MDT™ tool [47–49], which has a stout tube that presses against the borehole wall to establish hydraulic communication with the permeable zone of interest. A rubber packer around the probe creates a hydraulic seal against the mud cake formed by the drilling mud, reducing the potential for contamination by borehole fluids. The MDT™ tool has a pump that draws down the fluid from the formation into a sample bottle. The fluid can subsequently be analysed in the laboratory.

2.2. Sample preparation

The four downhole samples were flashed to remove volatile components (<*n*-C_5 alkanes). The asphaltenes of each were removed by precipitation with *n*-heptane prior to GC analysis and the resulting maltene fractions were collected with vacuum filtration by passing 40 ml heptane/g of oil through a 0.5 μm precombusted Millipore GF/F fibreglass filter. Most of the *n*-heptane was removed by rotary evaporation. The sample was then transfered to a vial and further blown down under a continuous stream of N_2.

2.3. GC × GC–FID analysis

The GC × GC–FID system employed a dual stage cryogenic modulator (Leco, Saint Joseph, Michigan) installed in an Agilent 7890A gas chromatograph configured with a 7683 series split/splitless auto-injector and two capillary columns. Each sample was injected in splitless mode and the purge vent was opened at 0.5 min. The inlet temperature was 300 °C. The first-dimension column was a nonpolar Restek Rtx-1 Crossbond (20 m × 0.25 mm i.d., 0.25 μm film thickness) that was held at 60 °C for 12 min and then ramped to 315 °C at 1.5 °C min^{-1}. The thermal modulator cold jet gas was dry N_2, chilled with liquid N_2. The thermal modulator hot jet air was heated to 60 °C above the temperature of the main GC oven. The hot jet was pulsed for 0.4 s every 10.0 s with a 5.6 s cooling period between stages. Second-dimension separations were performed with a 50% phenyl polysilphenylene-siloxane column (SGE BPX50, 1 m × 0.10 mm i.d., 0.1 μm film thickness) that was held at 70 °C for 12 min and then ramped to 325 °C at 1.5 °C min^{-1}. The carrier gas was H_2 at a constant flow rate of 1 ml min^{-1}. The FID detector signal was sampled at a rate of 200 data points s^{-1}.

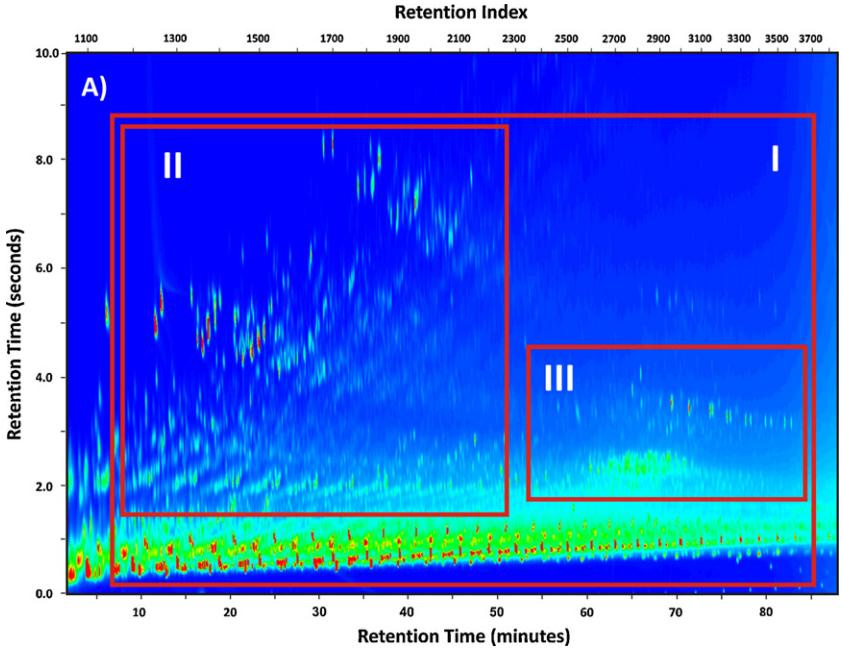

Fig. 1. GC × GC–FID chromatogram of the North Slope sample. Red boxes indicate the temporal ranges for the three Regions (I, II, and III) used for MPCA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

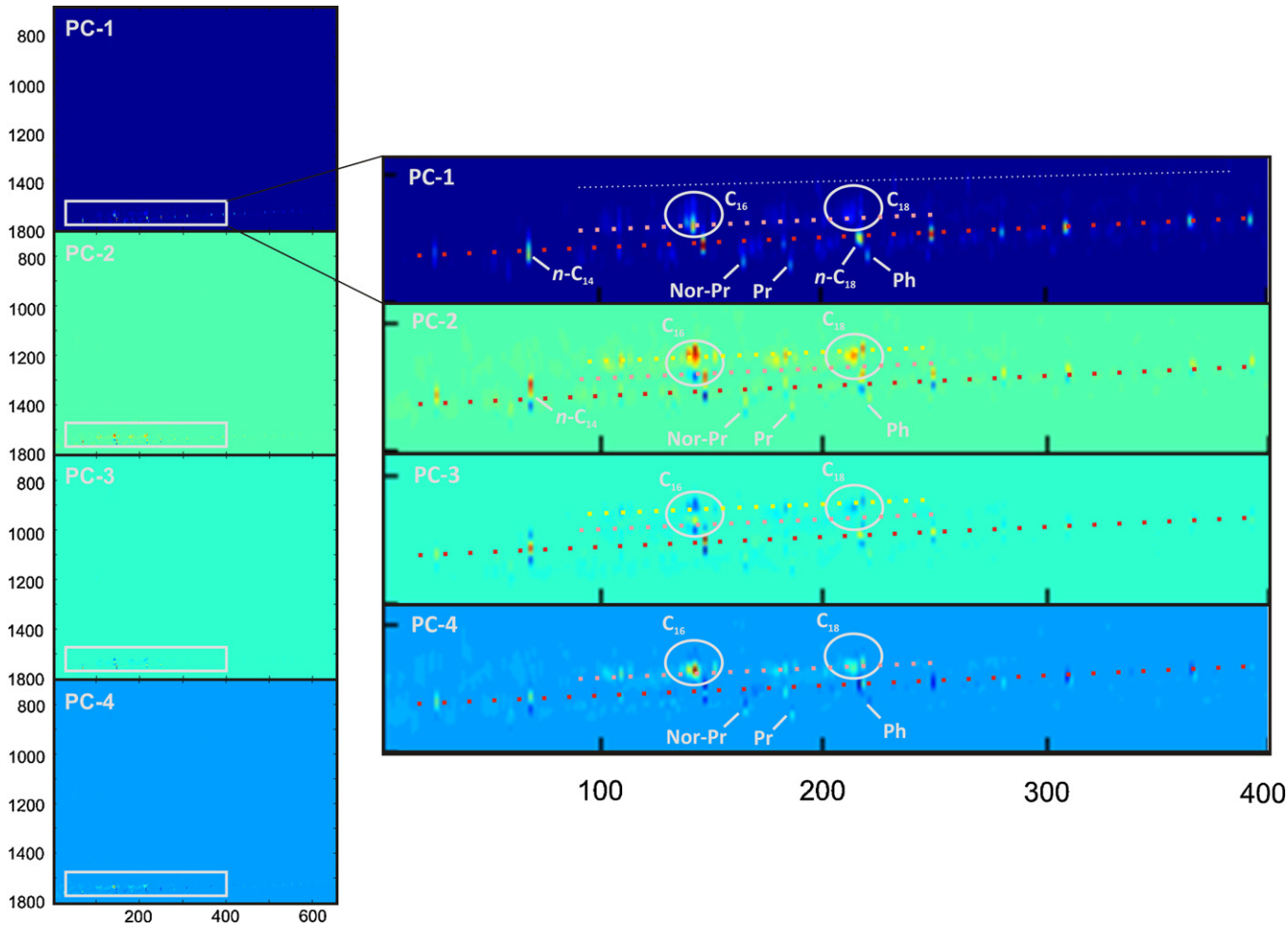

Fig. 2. Factor loading plots for PC-1, PC-2, PC-3, and PC-4 of Region I of the GC × GC chromatogram. Nor-Pr, Pr, and Ph are the acyclic isoprenoids, norpristane, pristane, and phytane, respectively. Red, pink, yellow, and white lines indicate the elution of *n*-alkanes, alkenes, alkadienes, and *n*-alkylcyclohexanes respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2.4. GC × GC–ToFMS analysis

The GC × GC time of flight mass spectrometry (-ToFMS) system employed a dual stage cryogenic modulator (Leco, Saint Joseph, Michigan) installed in an Agilent 6890 N gas chromatograph. Each extract was injected in splitless mode and the purge vent was opened at 0.5 min. The inlet temperature was 300 °C. The first-dimension column was a nonpolar Restek Rtx-5 Crossbond (15 m × 0.18 mm i.d., 0.2 μm film thickness) that was held at 50 °C for 5 min and then ramped to 300 °C at 3 °C min^{-1}. Compounds eluting from the first-dimension column were cryogenically modulated on deactivated fused silica (0.5 m × 0.11 mm i.d.). The modulator cold jet gas was dry N_2, chilled with liquid N_2. The thermal modulator hot jet air was heated to 60 °C above the temperature of the main GC oven. The hot jet was pulsed for 1 s every 8 s with a 3 s cooling period between stages. Second-dimension separations were performed with a 50% phenyl polysilphenylene-siloxane column (SGE BPX50, 0.70 m × 0.10 mm i.d., 0.1 μm film thickness) that was held at 70 °C for 5 min and then ramped to 320 °C at 3 °C min^{-1}. The carrier gas was He at a constant flow rate of 1.1 ml min^{-1}. The ToFMS detector signal was sampled at 50 spectra s^{-1}. The transfer line from the second oven to the ToFMS was deactivated fused silica (0.5 m × 0.18 mm i.d.), which was held at a constant temperature of 280 °C. The ToF source temperature was 230 °C and the detector was set to 1575 V.

2.5. GC × GC data processing

GC × GC–FID and GC × GC–ToFMS data acquisition were performed using ChromaToF© software. Individual peaks were automatically detected on the basis of a 50:1 signal to noise ratio. Biomarkers were identified using GC × GC–ToFMS and then quantified with GC × GC–FID. Specific compounds of various compound classes were identified using standards from Aldrich, US National Institute of Standards and Technology (NIST) and Chiron (Trondheim, Norway). Sample F-1 was divided into four aliquotes, which were sequentially analysed by GC × GC–FID. These F-1 sample replicates were treated as separate samples for the statistical analysis (Table 1). GC × GC–FID data files were exported as csv (comma separated variable) files and loaded into Noesys Transform version 2.4. Data files were baseline subtracted and normalized to the peak integration area of the recalcitrant 17α(H), 21β(H)-hopane biomarker. Three different sections of the chromatographic area were isolated from the data matrix (Fig. 1) and exported as csv files into Matlab™. MPCA was performed using the Matlab™ PLS_Toolbox 5.5 from Eigenvector Research Incorporated. The data was mean-centered, which translates the axes of the coordinate system to the center of gravity, or centroid, of the data [37].

MPCA is an unfold method in that the two-way data for each sample is unfolded row-wise and PCA is performed on the unfolded data. Two-way loadings presented herein are refolded from the

loadings that are computed by the analysis on the unfolded data. MPCA models were validated through removal of outliers and appropriate selection of principle components (PCs). Outliers samples, such as the North Slope, Prudhoe Bay crude oil that produced residuals placing it outside a 95% confidence interval were removed from analysis with respect to the model and the model was then re-fit after the exclusion. PCs were included until they did not increase the variance captured by more than 1%. For each analysis, the optimal number of PCs was determined by the percent variance encoded in each PC. Never more than four PC were used for any interpretative measurement. However, several MPCA calculated PCs having low variance were associated with factor loadings or scores plots containing geochemically interpretative data. In these cases the low variance was assumed to be focused on the chemically meaningful data and the results were included.

3. Results and discussion

Two attempts were made at analysing the oil samples. In both cases the samples were randomized and then injected in a single sequence to reduce systematic, temporal variations in the retention time offset of analytes. The initial sample sequence was analysed with a Leco Pegasus III GC × GC–ToFMS system, which uses an Agilent 6890N gas chromatograph. However, the exported total ion current (TIC) chromatograms had retention time offsets that progressively became more severe. A second attempt with GC × GC–FID data using an Agilent 7890A gas chromatograph was more successful. The improved electronic pneumatics controls, digital electronics, and faster oven cool-down of this gas chromatograph greatly improved the precision in data registration of the sample sequence.

MPCA generates factor loadings and scores plots. Factor loadings are the calculated variances associated with each variable in a sample set for a given principal component. When applied to GC × GC data, the *x*- and *y*-axis of a factor loading plot is the same as the first and second dimension of a GC × GC chromatogram. The *z*-axis indicates the positions in the data contributing to the variance between analytes. The *z*-axis thus allows for peak assignment of variance. Peaks with yellow to red colors have positive loadings. Blue peaks have negative loadings. Scores plots display the projection of the data onto these loadings, and similarities in these scores group similar samples. As such, scores plots can be used to assess the statistical similarity of various GC × GC amenable complex mixtures (i.e. petroleum samples). Taken together, the two plots not only enable a similarity estimation for comparing different complex chemical mixtures, but also provide the ability to determine which compounds vary and how such compounds vary between various chemical mixtures. This information in turn can be used to interpret the chemical meaning of differences in samples [50].

3.1. Region I – the entire GC × GC–FID chromatogram

The Region I data set contained 8 × ~3500 (objects or samples × variables or peaks). Dominant factor loadings within Region I on PC-1, the first principle component, are observed for the low to middle molecular weight range (*n*-C_{13} to *n*-C_{22}) of *n*-alkanes (Fig. 2). Preferentially high loadings are observed for *n*-C_{14}, *n*-C_{16}, *n*-C_{18}, *n*-C_{19}, *n*-C_{21}, *n*-C_{23}, and *n*-C_{24}. The lack of a systematic order (i.e. even over odd or odd over even) to this homologous series indicates the high factor loadings are not due to some samples having a carbon numbered preference, but instead the function of differences in the source, preservation and weathering of the oils. To a lesser extent, positive loadings are also observed for branched alkanes, alkenes, the acyclic isoprenoids nor-pristane, pristane, and phytane, as well as for the homologous series of *n*-alkylcyclohexanes. The presence of alkene isomers ranging from C_{15} to C_{19} had been previously documented Reddy et al. [20] as being contaminants derived from drilling fluids dissolved within an oil sample during extraction with the MDT™ tool. All but the North Slope, Prudhoe Bay, Alaska sample contains some level of these contaminants [24] and their high factor loadings within PC-1 was expected. Dominant loadings on PC-2 include C_{13}–C_{24} *n*-alkanes and C_{15}–C_{18} alkadienes. PC-3 has positive loadings for *n*-alkanes spanning C_{13}–C_{20} and negative loadings for C_{16} and C_{18} alkadienes. PC-4 produces high loadings for C_{16} and C_{18} alkenes of sample PER-1. However, this PC also includes a broad range of low molecular weight branched and cyclic paraffins that derive from the North Slope oil sample. Negative PC-4 loadings are present for lower molecular weight *n*-alkanes from *n*-C_{16} to *n*-C_{23} on the PER-1 sample. Reasons for different components appear in the same PC is subject of future work.

The PC-1, PC-2, PC-3, and PC-4 factor loading plots of Region I suggest that the fundamental chemical differences between the oil samples was not only differences in paraffin distributions, but also the degree of contamination received from drilling fluids (Fig. 2). This is significant because alkene peaks are difficult to quantitatively separate from crude oil peaks in standard GC analysis and the quantitative removal of these contaminants is not possible with GC–FID or GC–MS due to the co-elution with *n*-alkanes, as well as the similar occurrence of many fragment ions [20]. The presence of alkadienes with similar carbon number ranges as the alkenes indicates that these compounds are also likely component of the drilling fluids, which had escaped detection in prior investigations [20,24].

Oil samples with scores clustering near to one another in principal component space are chemically similar. Within Region I 72.03%, 17.55%, 6.76%, and 2.79% of molecular variance between the oil samples is associated with PC-1, PC-2, PC-3, and PC-4, respectively (Fig. 3). The F-1 and F-2 sample pair and injection replicates cluster close to one another with high positive scores on PC-1 and low negative scores on PC-2, PC-3, and PC-4. The close association of these samples indicates these oils have similar distributions of *n*-alkanes, acyclic isoprenoids, and unsaturated alkanes and potentially also similar distributions for all other compounds. Sample B has similar scores as the F-1 and F-2 sample pair for PC-1, PC-2, and PC-4 and higher positive score for PC-3. The PER-1 and PER-2 sample pair contain very different scores for PC-1 and PC-2 and similar scores for PC-3 and PC-4. All MDT samples contain high relative abundances of alkenes. However, these compounds constitute over 10% and 8% of the mass of resolvable paraffins for the PER-1 and PER-2 sample pair [24]. The differences in the high concentrations of unsaturated alkanes may be reflected in the different PC scores. The North Slope sample outlier is separated from all of the other samples.

3.2. Region II – aromatic compounds

The Region II data set contained 8 × ~1000 (objects or samples × variables or peaks). Within Region II the dominant loadings on PC-1 are obtained from compound classes alkylbenzenes, naphthalenes, benzothiophenes, fluorenes, phenanthrenes, and dibenzothiophenes (Fig. 4). Linear and substituted alkylbenzenes display progressively higher PC-1 factor loadings with increasing carbon number of the alkyl chain. Specific analytes with high PC-1 factor loadings are also observed for monomethyl- to tetramethyl-substituted naphthalenes and monomethyl- to trimethyl-substituted phenanthrenes. PC-2 contains dominant loadings for monomethyl and dimethyl-substituted naphthalenes as well as for monomethyl and dimethyl-substituted benzothiophenes. PC-3 contains highfactor loadings for linear and substituted alkylbenzenes, monomethyl-naphthalenes, two isomers of dimethylnaphtalenes, and three isomers of trimethyl-

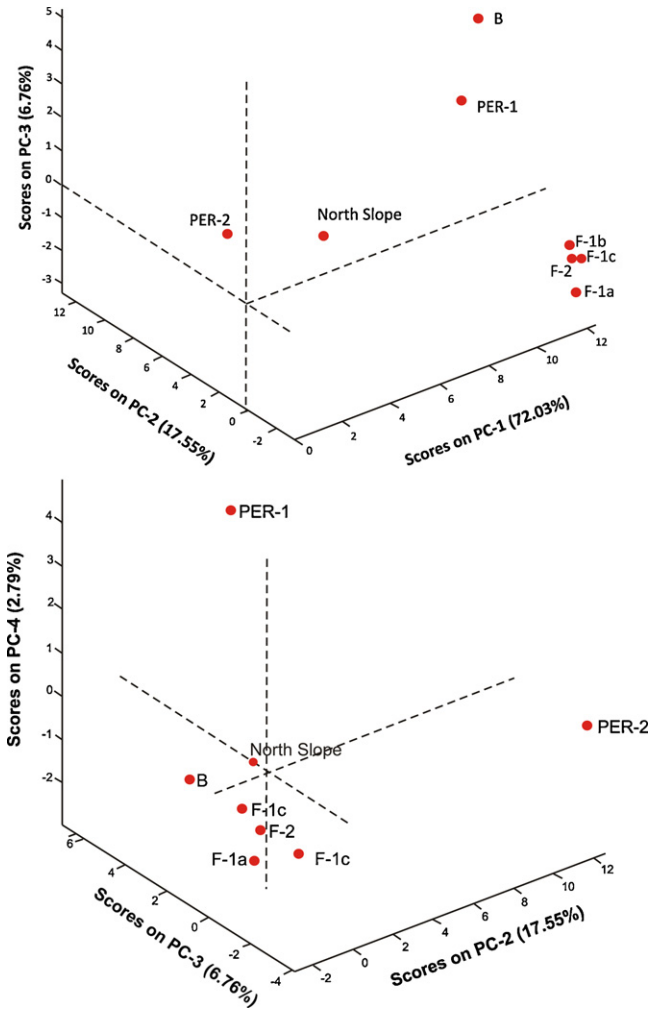

Fig. 3. Scores plots for PC-1, PC-2, PC-3, and PC-4 of the Region I of the GC × GC chromatogram.

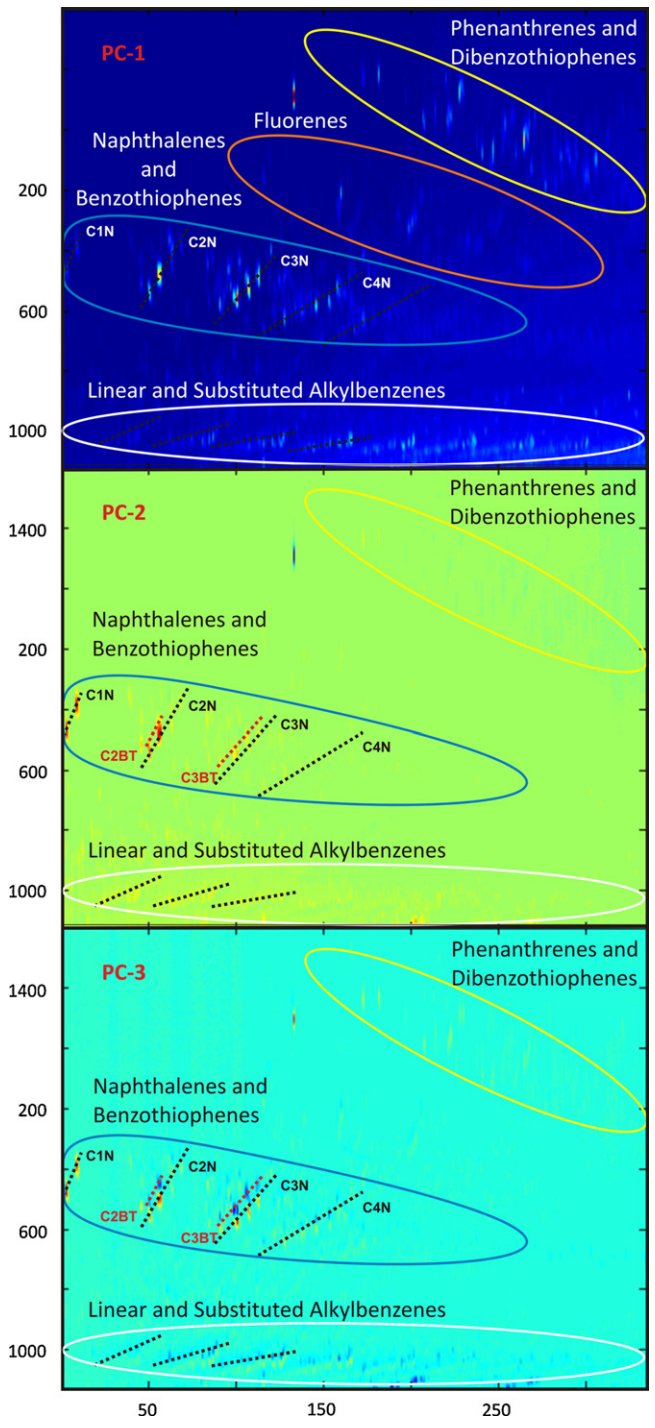

Fig. 4. Region II factor loading plots for PC-1, PC-2, and PC-3 representing the elution of low to medium molecular weight aromatic compounds. C1N–C4N compound series refer to monomethyl- to tetramethyl substituted naphthalenes. C2BT and C3BT are di- and trimethyl substituted benzothiophenes.

naphthalenes. PC-3 also contains negative loadings for several linear alkylbenzenes, as well as monomethyl- and dimethyl-substituted benzothiophene isomers.

The scores plot of Region II indicates 94.74%, 2.64%, and 1.45% of variance is associated with PC-1, PC-2, and PC-3, respectively (Fig. 5). As with Region I, the F-1 and F-2 sample pair and injection replicated cluster close to one another with high positive PC-1 scores. However for Region II these samples have near zero scores for PCs 2 and 3. Sample B contains similarly high positive PC-1 scores and near zero scores for PC-2, but has more negative scores on PC-2. For Region II, the PER-1 and PER-2 sample pair is separated from all of the other samples and clusters closely together with positive low PC-1 and PC-3 scores and negative scores on PC-2. The North Slope outlier sample is clearly separated from all of the MDT samples with positive low scores on PC-1 and positive high scores on PC-3 (Fig. 5).

Although Region II contains compounds that are prone to biochemical and physical weathering processes, the compounds are not as volatile and prone to loss or contamination by sample extraction and preparation or biodegradaton when compared with the analysis of Region I. MPCA of Region II therefore provides a more accurate picture of the compositional similarity of the various oil samples. The MPCA results of Region II indicate systematic differences exist between the general abundance of different compound classes as well as differences in specific isomers within a compound class for the various oil samples. The associated scores plot indicates the oil samples with more similar geographic locations also have a greater degree of molecular similarity.

3.3. Region III – Sterane and hopane biomarkers

The Region III data set contained 8 × ~150 (objects or samples × variables or peaks). Problems with data registration was more significant for the Region III MPCA. Second dimension retention time offsets were identified by the systematic production

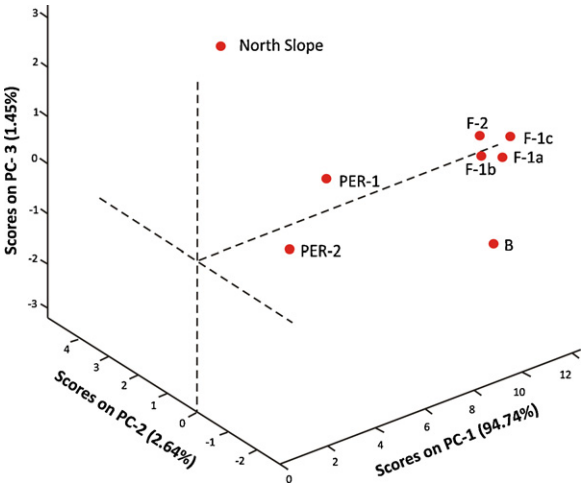

Fig. 5. Scores plot for PC-1, PC-2, and PC-3 of the Region II of the GC × GC chromatogram.

of peak shadows with low or high factor loading for each of the measured analytes (Fig. 6). Misalignments were most prominent for the suite of hopane biomarkers. However, even with these offsets various compositional differences are identifiable. Within the biomarker elution area, high positive factor loadings on PC-1 were observed for the C_{27} 13β(H),17α(H)-diacholestane 20S and R and the C_{29} 24-ethyl-13β(H),17α(H)-diacholestane 20S and R and the C_{27} 5α(H),14β(H),17β(H)-cholestane 20R, C_{24} tetracyclic terpenoid (TT), C_{29} 17α(H),21β(H)-norhopane and C_{30} 17α(H),21β(H)-hopane (H). No variation from PC-1 to PC-2 was observed for the hopanes. However, relatively higher positive factor loadings are observed for the steranes C_{27} 13β(H),17α(H)-diasterane 20S and R and C_{27} 5α(H),14β(H),17β(H)-cholestane 20R. Additionally, high positive factor loadings were also observed for the C_{28} 28,30 bisnorhopane (BNH), which was influenced by the high abundance of this compound in the North Slope oil sample.

The scores plots of Region III indicate 98.14% and 0.83% of variance is associated with PC-1 and PC-2, respectively (Fig. 7). Although the MPCA of Region III is affected by poorer data registration, the scores plot for this region produces clusters that relate

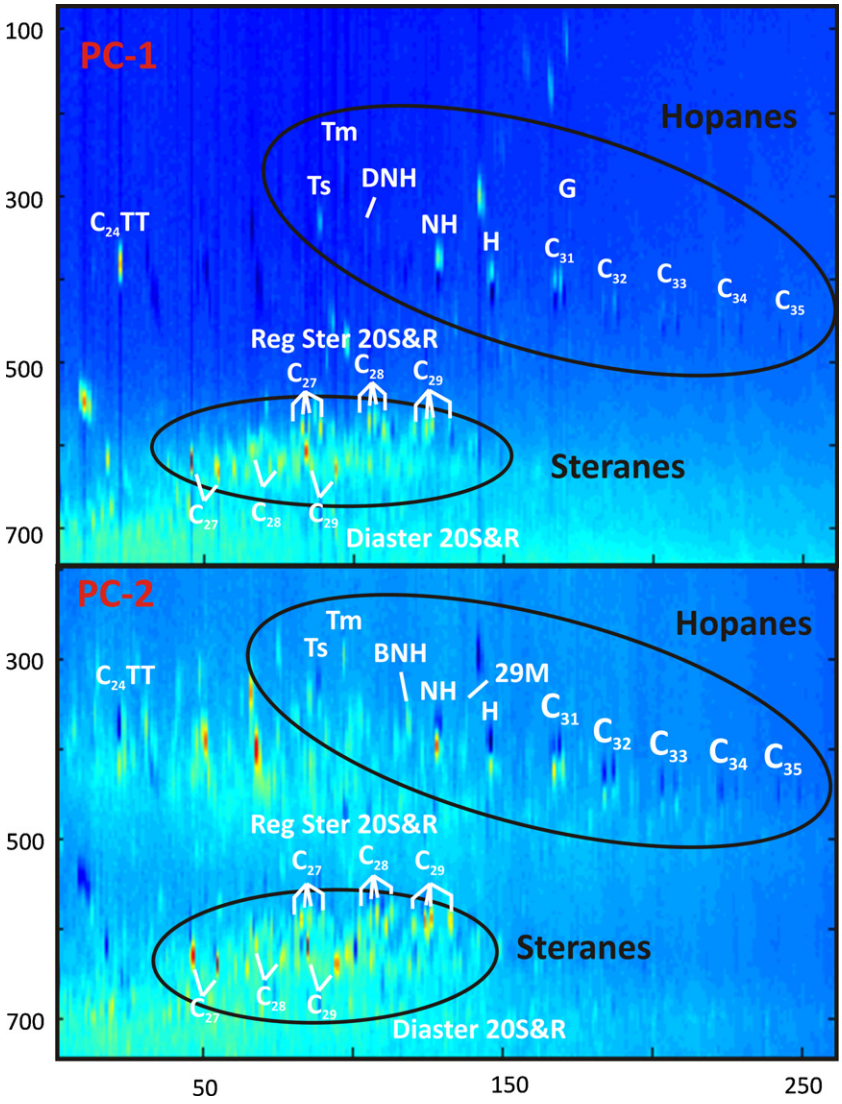

Fig. 6. Factor loading plots for PC-1 and PC-2 of Region III of the GC × GC chromatogram representing the elution region of sterane and hopane biomarkers. C24 TT is the C_{24} tetracyclic terpane, Ts, Tm, BNH, NH, 29 M, and H are C_{27} 17α-22,29,30-Trisnorhopane, C_{27} 18α-22,29,30-Trisnorhopane, C_{28} bisnorhopane, norhopane, C_{29} normoretane, and hopane, respectively. 22S and R homohopane isomers are labelled C_{31}–C_{35}. G denotes gammacerane. Reg Ster and Diaster are regular and diasteranes.

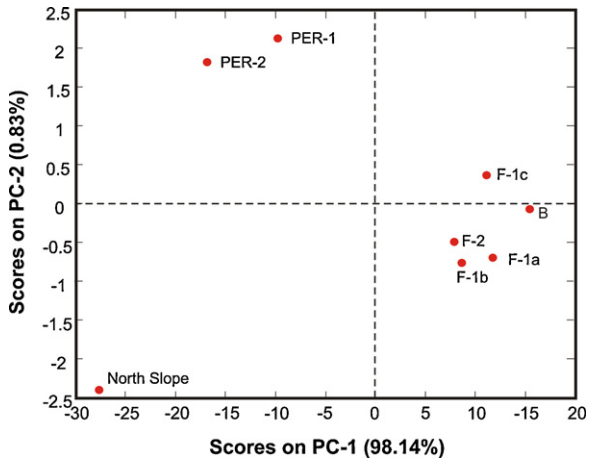

Fig. 7. Scores plot for PC-1 and PC-2 of the Region III of the GC × GC chromatogram.

to the expected genetic similarities of each oil sample. For example, all of the F-1 replicates and the F-2 sample cluster together. Sample B, which was collected from an unknown location in the same oil field, also groups with the F-1 and F-2 sample pairs and replicates, suggesting these oils formed from the same source rock. The same situation is observed for the PER-1 and PER-2 sample pair. All of the sample pairs are clearly separated from one another as well as from the North Slope outlier sample. Such clearly delineated groupings with the reduced data registration could not be a random result suggesting the MPCA techniques was still effective.

3.4. Comparison with other oil fingerprinting methods

Different MPCA results are observed based on which region of the GC × GC chromatogram that was analysed. Region I effectively demonstrated that subtle compositional differences of paraffins and drilling fluid contaminants could be detected by using MPCA. However, the potential to discriminate genetic similarities between sample B, F-1, and F-2 and between the PER-1 and PER-2 sample pairs was not possible because of the variable contributions of alkenes. This contrasts with a parallel analysis performed by using biomarker ratios and classification schemes that compared, specific compound class contributions, as well as the hypothesis testing of potential differences between compound classes across 1st dimension retention index ranges [24]. For that study, it was possible to either ignore the presence of drilling fluid contaminants or completely quantitatively remove their influence. The PER-1 and PER-2 samples were demonstrated to have nearly identical biomarker ratios and concentrations of many compound classes such as paraffins. However, this limitation observed with the Region I MPCA analysis was largely overcome without the additional use of retention time alignment algorithms by analysing other regions of the chromatogram and in this regard the sensitivity of the analysis was likely improved. For example, the close clustering of the MPCA scores plot of Region III (Fig. 7) is in contrast to the similarity calculated for the same compound classes in Ventura et al. [24].

3.5. Interpretative value

As noted by Mispelaar et al. [51] multivariate-analysis (MVA) techniques cannot distinguish between informative variables such as GC × GC peaks describing differences between samples and 'uninformative variables', which are peaks that do not describe relevant differences. In the event that these can be distinguished, variations between individual compounds or compound classes can be interpreted as representing source specific differences from the oil reservoir. For example, the identification of elevated factor loadings of methyldibenzothiophene isomers may be useful in that differences in these compounds have been linked to oils derived from carbonate versus siliciclastic source rocks [52]. In this respect MPCA can be used to generate inferences about the origin of the oils. Additionally, the chromatographic regions chosen for analysis can limit the type of phenomenon examined. In this respect, oil modifying processes such as water-washing and biodegradation, which are more prominent from Regions I and II can be separated from factors that causing different source rock specific biomarker compositions of Region III.

However, ultimately one would like to know with what certainty various oils are the same or different. In other words, one could ascertain the probability that two or more oil samples are related and therefore come from the same compartment. Such hypothesis testing is possible with PCA and subsequently also MPCA. To facilitate this form of decision making, a greatly increased number of samples is required to define a distribution of a particular class. For this particular study, a class is a specific sample source or location that would then be sampled multiple times to determine its chemical variability. Afterwards a *t*-test can be performed to determine if the distributions are different.

4. Conclusion

Within this experiment we demonstrate that MPCA can be used to effectively resolve the molecular differences between very similar oil samples as well as be used to group crude oils based on the degree of their molecular similarity. Compositional and instrumental artefacts introduced during sampling and analytical processing were quantitatively evaluated. Contaminants such as alkenes and alkadienes introduced during sampling were easily identified within factor loading plots and their high relative abundance did not impede the determination of the primary compositional relatedness of the oil samples. These constituents can impede the determination of the primary composition and relatedness of the oil samples that were not contaminated if the degree to which these contaminants occur is large. In such cases other chromatographic areas should be analysed and integrated into the process of model validation. Unique molecular differences between the various samples were also readily identified.

High-resolution techniques such as GC × GC are necessary to elucidate minute differences in oil composition. However, the large data sets that are associated with these types of analysis require other novel processing approaches. The data mining power of MPCA enables the simultaneous comparison of thousands of GC amenable compounds as well as a simple and effective method to differentiate minor chemical differences between oils that cannot be achieved with GC–MS. This method expands upon current statistical applications relying on the comparison of a few common analytes, such as biomarker ratios, to establish the chemical similarity of oil. MPCA with GC × GC data should be capable of dissevering petrochemical changes associated with such processes as reservoir connectivity and the physical or biological oil weathering processes. MPCA can be used to screen a vast array of fluid contaminants and be used to discriminate between source dependent and weathering related processes that uniquely impact the molecular composition of crude oils.

Acknowledgements

The above represents only the opinions of the author and do not represent the position of the U.S. Coast Guard or the United States of America. This study was supported by the U.S. National Science

Foundation (IIS-0430835), U.S. Department of Energy (DE-FG02-06ER15775) and The Seaver Institute.

References

[1] B.T. Brooks, C.E. Bood, S.S. Kurtz Jr., L. Schmerling, The Chemistry of Petroleum Hydrocarbons, vol. 1, Reinhold Publishing, New York, 1957.
[2] B.P. Tissot, D.H. Welte, Petroleum Formation and Occurrence, Springer-verlag, Berlin, 1984.
[3] A.G. Marshall, C.L. Hendrickson, G.S. Jackson, Mass Spectrom. Rev. 17 (1998) 1.
[4] C.A. Hughey, R.P. Rogders, A.G. Marshall, Anal. Chem. 74 (2002) 4145.
[5] O.C. Mullins, R.P. Rodgers, P. Weinheber, G.C. Klein, L. Venkataramanan, A. Ballard, A.G. Marshal, Energy Fuels 20 (2006) 2448.
[6] G. Thouand, P. Bauda, J. Oudot, G. Kirsch, C. Sutton, J.F. Vidalie, Can. J. Microbiol. 45 (1999) 106.
[7] A.E. Pomerantz, G.T. Ventura, A.M. McKenna, J.A. Cañas, R.K. Nelson, C.M. Reddy, R.P. Rodgers, A.G. Marshall, K.E. Peters, O.C. Mullins, Org. Geochem. 41 (2010) 812.
[8] K.E. Peters, C.C. Walters, J.M. Moldowan, The Biomarker Guide. Biomarkers and Isotopes in the Environment and Human History, vol. 1, Cambridge University Press, 2005.
[9] Z. Wang, S.A. Stout, Oil Spill Environmental Forensics. Fingerprinting and Source Identification, Elsevier, 2007.
[10] H.L. ten Haven, E. Lafargue, M. Kotarba, Org. Geochem. 20 (1993) 935.
[11] P.D. Boehm, G.D. Douglas, W.A. Burnds, P.J. Mankiewicz, D.S. Page, E. Bence, Mar. Pollut. Bull. 34 (1997) 599.
[12] Z. Wang, M.F. Fingas, Mar. Pollut. Bull. 47 (2003) 423.
[13] J. Blomberg, P.J. Schoenmakers, J. Beens, J.R. Tijssen, J. High Resolut. Chromatogr. 20 (1997) 539.
[14] R.B. Gaines, G.S. Frysinger, M.S. Hendrick-Smith, D. Stuart, Environ. Sci. Technol. 33 (1999) 2106.
[15] R.B. Gaines, G.S. Frysinger, C.M. Reddy, R.K. Nelson, in: Z. Wang, S. Stout (Eds.), Spill Oil Fingerprinting and Source Identification, Academic Press, New York, 2006, p. 169.
[16] G.S. Frysinger, R.B. Gaines, High Resolut. Chromatogr. 23 (2000) 197.
[17] G.S. Frysinger, R.B. Gaines, L. Xu, C.M. Reddy, Environ. Sci. Technol. 37 (2003) 1653.
[18] J. Beens, J. Blomberg, P.J. Schoenmakers, J. High Resolut. Chromatogr. 23 (2000) 182.
[19] C.M. Reddy, T.I. Eglinton, A. Hounshell, H.K. White, L. Xu, R.B. Gaines, G.S. Frysinger, Environ. Sci. Technol. 36 (2002) 4754.
[20] C.M. Reddy, R.K. Nelson, S.P. Sylva, L. Xu, E.A. Peacock, B. Raghuraman, O.C. Mullins, J. Chromatogr. A 1148 (2007) 100.
[21] G.T. Ventura, F. Kenig, C.M. Reddy, R.K. Nelson, S. Schieber, G. Frysinger, R. Gaines, PNAS 104 (2007) 14261.
[22] G.T. Ventura, F. Kenig, C.M. Reddy, R.K. Nelson, G. Frysinger, B. Van Mooy, R. Gaines, Org. Geochem. 39 (2008) 846.
[23] S.S. Betancourt, G.T. Ventura, D. Pomerantz, O. Viloria, F.X. Dubost, J. Purcell, R.K. Nelson, R.P. Rodgers, C.M. Reddy, A.G. Marshall, O.C. Mullins, Energy Fuels 23 (2008) 1178.
[24] G.T. Ventura, B. Raghuraman, R.K. Nelson, G. Lambertus, O.C. Mullins, C.M. Reddy, Org. Geochem. 41 (2010) 1026.
[25] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons Inc., 2003.
[26] B.R.J. Kowalski, J. Chem. Info. Comp. Sci. 15 (1975) 201.
[27] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
[28] R. Kramer, Chemometric Techniques for Quantitative Analysis, Marcel Dekker, New York, 1998.
[29] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley-Interscience, New York, 1998.
[30] J.E. Jackson, J. Qual. Technol. 13 (1981).
[31] V.G.V. Mispelaar, A.C. Tas, A.K. Smilde, P.J. Schoenmakers, A.C.V. Asten, J. Chromatogr. A 1019 (2003) 15.
[32] V.G.V. Mispelaar, H.-G. Janssen, A.C. Tas, P.J. Schoenmakers, J. Chromatogr. A 1071 (2005) 229.
[33] K.M. Pierce, J.L. Hope, J.C. Hoggard, R.E. Synovec, Talanta 70 (2006) 797.
[34] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 78 (2006) 2700.
[35] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341.
[36] S. Wold, P. Geladi, K. Esbensen, J. Öhman, J. Chemometr. 1 (1987) 41.
[37] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch., PLS_Toolbox 4.0 for use with MATLAB™, Eigenvector Research, Inc., 3905 West Eaglerock Drive, Wenatchee, WA 98801, USA, 2006.
[38] K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1141 (2007) 106.
[39] C.G. Fraga, B.J. Prazen, R.E. Synovec, J. High Resolut. Chromatogr. 23 (2000) 215.
[40] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833.
[41] A.E. Sinha, K.J. Johnson, B.J. Prazen, S.V. Lucas, C.G. Fraga, R.E. Synovec, J. Chromatogr. A 98 (2003) 195.
[42] K.J. Johnson, B.J. Prazen, D.C. Young, R.E. Synovec, J. Sep. Sci. 27 (2004) 410.
[43] M.T. Ni, S.E. Reichenbach, A. Visvanathan, J. TerMaat, E.B. Ledford, J. Chromatogr. A 1086 (2005) 165.
[44] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[45] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231.
[46] J.H. Christensen, G. Tomasi, A.B. Hansen, Environ. Sci. Technol. 39 (2005) 255.
[47] G. Fujisawa, O.C. Mullins, in: O.C. Mullins, E.Y. Sheu, A. Hammami, A.G. Marshall (Eds.), Asphaltenes, Heavy Oils and Petroleomics, Springer, New York, 2007, p. 589.
[48] O.C. Mullins, G.T. Ventura, R.K. Nelson, S.S. Betancourt, B. Raghuraman, C.M. Reddy, Energy Fuels 22 (2008) 496.
[49] O.C. Mullins, The Physics of Reservoir Fluids, Discovery Through Downhole Fluid Analysis, Schlumberger Press, Houston, TX, 2008.
[50] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Environ. Forensics 7 (2006) 77.
[51] V.G.V. Mispelaar, A.K. Smilde, O.E.D. Noord, J. Blomberg, P.J. Schoenmakers, J. Chromatogr. A 1096 (2005) 156.
[52] W.B. Hughes, in: J.G. Palacas (Ed.), Petroleum Geochemistry, Source Rock Potential of Carbonate Rocks, AAPG Studies in Geology, vol. 18, 1984, p. 181.