# Combination of Statistical Methods and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for More Comprehensive, Molecular-Level Interpretations of Petroleum Samples

**ARTICLE** *in* ANALYTICAL CHEMISTRY · DECEMBER 2009

Impact Factor: 5.64 · DOI: 10.1021/ac901748c · Source: PubMed

*Anal. Chem.* **2010,** *82,* 211–218

# Combination of Statistical Methods and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for More Comprehensive, Molecular-Level Interpretations of Petroleum Samples

**Manhoi Hur,[†,⊥] Injoon Yeo,[‡] Eunsuk Park,[‡] Young Hwan Kim,[‡] Jongshin Yoo,[‡] Eunkyoung Kim,[§] Myoung-han No,[§] Jaesuk Koh,[§] and Sunghwan Kim*[,∥]**

*Department of Bioinformatics, Korea University, Seoul, Korea, Mass Spectrometry Group, Korean Basic Science Institute, Mass Spectrometry Team, Ochang, Korea, SK Energy Institute of Technology, Daejeon, Korea, and Department of Chemistry, Kyungpook National University, Daegu, Korea*

**Complex petroleum mass spectra obtained by Fourier-transform ion cyclotron resonance mass spectrometry (FTICR MS) were successfully interpreted at the molecular level by applying principle component analysis (PCA) and hierarchical clustering analysis (HCA). A total of 40 mass spectra were obtained from 20 crude oil samples using both positive and negative atmospheric pressure photoionization (APPI). Approximately 400 000 peaks were identified at the molecular level. Conventional data analyses would have been impractical with so much data. However, PCA grouped samples into score plots based on their molecular composition. In this way, the overall compositional difference between samples could be easily displayed and identified by comparing score and loading plots. HCA was also performed to group and compare samples based on selected peaks that had been grouped by PCA. Subsequent heat map analyses revealed detailed compositional differences among grouped samples. This study demonstrates a promising new approach for studying multiple, complex petroleum samples at the molecular level.**

Since its introduction by Comisarow and Marshall,[1] Fourier-transform ion cyclotron resonance mass spectrometry (FTICR MS) has become a powerful tool for studying natural organic mixtures at the molecular level. FTICR MS has been widely applied to investigate metabolites,[2] vegetable oils,[3] wine,[4] explosives,[5] coal extracts,[6] and humic materials,[7,8] and it has proven especially useful for determining the chemical composition of petroleum, a technique known as petroleomics.[9,10] Broadband FTICR MS spectra of petroleum are usually very complex, with peaks appearing over a wide dynamic range. However, the ultrahigh mass resolution and high mass accuracy of FTICR MS makes it possible to identify individual organic molecules.[11] Knowing the major elemental constituents of a sample allows molecules in even the most complex natural mixtures, such as crude oil, to be identified solely on the basis of measured $m/z$ values.

Many studies have been devoted to developing new instrumentation,[12] ionization methods,[13,14] and analytical methods[8,15,16] to improve the application of FTICR MS to petroleum samples. Given the extreme complexity of petroleum samples, one of the key research areas is the development of improved data interpretation methods. A single FTICR MS spectrum of petroleum routinely contains 5000−15 000 peaks, and as many as 50 000 peaks have been observed in a single spectrum.[12] This level of complexity poses a clear analytical challenge. Obtaining and evaluating specific information on selected peaks among the more than 10 000 other peaks is analogous to the colloquial needle in

* Corresponding author. Phone: 82-53-950-5333. Fax: 82-53-950-6330. E-mail: sunghwank@knu.ac.kr.
† Korea University.
⊥ Current address: BNF Technology Inc., Daejeon, Korea.
‡ Korean Basic Science Institute.
§ SK Energy Institute of Technology.
∥ Kyungpook National University.

(1) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489–490.
(2) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26*, 51–78.
(3) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *J. Agric. Food Chem.* **2004**, *52*, 5322–5328.
(4) Cooper, H. J.; Marshall, A. G. *J. Agric. Food Chem.* **2001**, *49*, 5710–5718.
(5) Wu, Z.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2002**, *74*, 1879–1883.
(6) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2004**, *18*, 1424–1428.
(7) Kim, S.; Kaplan, L. A.; Hatcher, P. G. *Limnol. Oceanogr.* **2006**, *51*, 1054–1063.
(8) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336–5344.
(9) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53–59.
(10) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18090–18095.
(11) Kim, S.; Rodgers, R. P.; Marshall, A. G. *Int. J. Mass Spectrom.* **2006**, *251*, 260–265.
(12) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. *Anal. Chem.* **2008**, *80*, 3985–3990.
(13) Schaub, T. M.; Hendrickson, C. L.; Qian, K. N.; Quinn, J. P.; Marshall, A. G. *Anal. Chem.* **2003**, *75*, 2172–2176.
(14) Purcell, J. M.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1682–1689.
(15) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2001**, *73*, 4676–4681.
(16) Kim, S.; Rodgers, R. P.; Blakney, G. T.; Hendrickson, C. L.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 263–268.

10.1021/ac901748c 
Published on Web 12/07/2009

a haystack. Reducing and visualizing the complex, ultrahigh resolution mass spectra inherent to the FTICR MS technique is not a trivial task.

Kendrick mass defect and van Krevelen diagram analyses have been successfully applied to ultrahigh resolution mass spectra[8,15] and have allowed data interpretation and spectral comparisons of complex data sets. In Kendrick mass defect analyses, peaks are sorted by their homologous relatives across the range of masses. Compounds whose compositions differ by specific masses associated with a given structural unit (e.g., $CH_2$, COOH, $CH_2O$) can be discerned in two-dimensional plots wherein structurally related peaks plot out on horizontal or diagonal straight lines.[15,17] In van Krevelen diagram analyses, the ratio between the number of elements in molecular formulas and the observed relative abundance of those formulas is plotted.[8] In this way, any changes or differences at the molecular level can be visualized. However, despite these efforts, the extraction of information out of such complex mass spectra, especially when multiple samples must be analyzed and compared, remains difficult. For example, a van Krevelen diagram of petroleum spectra displays only one sample at a time. A detailed comparison of 20 samples would require the generation and manual inspection of 20 individual van Krevelen diagrams.

Statistical analyses have been successfully applied to process and extract information from large data sets such as those obtained with DNA microarrays,[18,19] organic thin films,[20] proteomics,[21,22] and humics.[23,24] The application of similar statistical methods in petroleomics could provide a viable means of data analysis. However, despite its potential, no reports to date have cited statistical methods for interpreting complex petroleum mass spectra. The current study describes the application of principal component analysis (PCA) and hierarchical clustering analysis (HCA), combined with heat map analyses, to study complex petroleomic data. To the best of our knowledge, this is the first study to apply statistical methods to study petroleum data obtained by FTICR MS.

## EXPERIMENTAL SECTION

Twenty crude oil samples (labeled X01 through X20) and their bulk properties (sulfur, nitrogen, vanadium, and total acid number, TAN) were provided by the SK Energy Corporation. Twenty oil samples were selected for this study, ten with high sulfur content and ten with low sulfur content. Samples were prepared by diluting crude oil samples to 1 mg/mL with a 50:50 v/v solution of toluene/methanol. HPLC grade methanol and toluene were purchased from Merck (Gibbstown, NJ) and used without further purification. Prepared samples were directly injected with a syringe pump (Harvard, Holliston, MA) at a flow rate of 200 $\mu$L/h. Analyses were performed with a 15 T FTICR mass spectrometer at the Korean Basic Science Institute (KBSI, Ochang-eup, Korea). The atmospheric pressure photoionization (APPI) source was obtained from Bruker Daltonics (Billerica, MA). The Apex hybrid Qq-FT instrument was equipped with a Bruker Apollo II dual source. Nitrogen was used as the drying and nebulizing gas. For the APPI analyses, the nebulizing temperature was set to 200 °C with a 3.0 L/min flow rate, and the drying gas temperature was set to 200 °C with a 2.0 L/min flow rate; the skimmer voltage was set to 13.0 V to minimize in-source fragmentation. Ionized samples were stored in an argon-filled collision cell for 1 s and transferred to the ICR cell with a 2 ms time-of-flight window. Both sidekick and gated trapping approaches were utilized. A sidekick voltage of 20 V was used to initially trap the ions. After transferring the ions to the ICR trap, the trap voltage was raised to 3 V and ramped down to 1.5 V for detection. At least 100 scans were accumulated and averaged to improve the signal-to-noise ratio of the resulting spectra. For each spectrum, at least $2 \times 10^6$ data points were recorded. An average resolving power of more than 300 000 at $m/z$ ~400 was routinely observed.

Spectral interpretation and statistical computations were performed with R version 2.9.0 and ChemBrowser software[25] (BNF Technology Inc., South Korea). Initially, the threshold for peak picking was a signal-to-noise ratio higher than 4.5. An automated peak-picking algorithm was later implemented for more reliable and faster results.[26] After peak picking, elemental formulas were calculated and assigned on the basis of $m/z$ values within a 1 ppm error range. Normal conditions for petroleum data ($C_cH_hN_nO_oS_s$, $c$ unlimited, $h$ unlimited, $0 \le n \le 5$, $0 \le o \le 10$, $0 \le s \le 2$)[11] were used in these calculations. Typically, more than 98% of the observed peaks could be successfully assigned with elemental formulas. A more detailed description of the statistical analyses is provided below in the Results and Discussion section.

## RESULTS AND DISCUSSION

**Data Preprocessing for Statistical Analyses.** Twenty petroleum samples were analyzed by FTICR MS utilizing both positive and negative mode APPI. Each sample was analyzed in triplicate for statistical significance. Therefore, a total of 60 mass spectra from 20 oil samples were obtained in each ionization mode. To ensure identical instrumental conditions, each full set of 60 spectra was obtained within 20 h over the course of two consecutive days. Triplicate spectra were later combined into a single spectrum, resulting in 20 mass spectra and an equal number of peak lists for each ionization mode. Data obtained in this manner were fairly consistent and showed standard deviations of less than 5% of the initial values (refer to Table 1).

Each peak list contained 10 000−15 000 peaks. This implies that at least 200 000 peaks were processed for each ionization mode. The relative abundance of individual peaks was normalized


(17) Kendrick, E. *Anal. Chem.* **1963**, *35*, 2146–2154.
(18) Komura, D.; Nakamura, H.; Tsutsumi, S.; Aburatani, H.; Ihara, S. *Bioinformatics* **2005**, *21*, 439–444.
(19) Lee, C. Y.; Harbers, G. M.; Grainger, D. W.; Gamble, L. J.; Castner, D. G. *J. Am. Chem. Soc.* **2007**, *129*, 9429–9438.
(20) Cheng, F.; Gamble, L. J.; Grainger, D. W.; Castner, D. G. *Anal. Chem.* **2007**, *79*, 8781–8788.
(21) America, A. H. P.; Cordewener, J. H. G.; van Geffen, M. H. A.; Lommen, A.; Vissers, J. P. C.; Bino, R. J.; Hall, R. D. *Proteomics* **2006**, *6*, 641–653.
(22) Polpitiya, A. D.; Qian, W. J.; Jaitly, N.; Petyuk, V. A.; Adkins, J. N.; Camp, D. G.; Anderson, G. A.; Smith, R. D. *Bioinformatics* **2008**, *24*, 1556–1558.
(23) Kujawinski, E. B.; Longnecker, K.; Blough, N. V.; Vecchio, R. D.; Finlay, L.; Kitner, J. B.; Giovannoni, S. J. *Geochim. Cosmochim. Acta* **2009**, *73*, 4384–4399.
(24) Sleighter, R. L.; Hatcher, P. G. 2009, in preparation.
(25) Hur, M.; Shin, S.; Seo, H.; Yeo, I.; Park, E. S.; Kim, E.; No, M. H.; Kim, Y. H.; Kim, S. Interpretation of Crude Oil High Resolution Spectra obtained by ESI and APPI FT-ICR Mass Spectrometry using Principal Components Analysis. *Proceeding of the 57th ASMS Conference on Mass Spectrometry and Allied Topics*, Philadelphia, PA, May 31−June 4, 2009.
(26) Hur, M.; Oh, H. B.; Kim, S. *Bull. Korean Chem. Soc.* **2009**, *30*, 2665–2668.

**Table 1. Example of a Merged Table Used in PCA of FTICR Spectra of Petroleum[a]**

| *m/z* | X01 | X02 | X03 | X19 | X20 |
|---|---|---|---|---|---|
| 366.27917 | 0.0118 (±0.0004) | 0 | 0.0079 (±0.0006) | 0.0041 (±0.0006) | 0 |
| 366.29173 | 0.0058 (±0.0003) | 0.0042 (±0.0010) | 0.0082 (±0.0003) | 0.0056 (±0.0010) | 0.0062 (±0.0002) |
| 366.31554 | 0.0279 (±0.0012) | 0.0042 (±0.0005) | 0.0098 (±0.0007) | 0.0069 (±0.0007) | 0.0042 (±0.0004) |
| 366.31919 | 0.0067 (±0.0008) | 0 | 0.0044 (±0.0005) | 0.0045 (±0.0004) | 0 |
| 366.32796 | 0.1145 (±0.0037) | 0.0812 (±0.0004) | 0.0853 (±0.0021) | 0.0932 (±0.0010) | 0.0851 (±0.0002) |

[a] Standard deviations obtained from triplicate experiments are listed in parentheses.

to the summed relative abundance of each peak list. Assignments of molecular formulas were performed for more than 95% of the total number of peaks. The existence of appropriate isotope peaks was checked to ensure the correctness of the assignment, although isotope peaks for low abundance compounds were too small to allow this checking procedure. Peaks that were not possible to assign with formulas were filtered out and not considered in subsequent analysis steps.

For each ionization mode, the 20 resulting peak lists with elemental formula assignments were merged into a single table for statistical analysis. An example of a merged table is given in Table 1. Peaks with the same elemental formulas, but found in different samples, were merged into a single line of the table, and normalized abundances are listed. Data merging was performed on the basis of assigned elemental formulas, which were equivalent to theoretical mass values. A zero value in Table 1 indicates that the peak was not observed in that particular sample.

**PCA of Positive Mode APPI Data.** PCA was performed with the data in the merged table obtained from positive mode APPI spectra. A detailed description of the PCA method can be found in the literature.[27,28] Briefly, PCA reduces the number of variables through a linear combination of the original variables, e.g., peaks observed in a mass spectrum. The new variables that result are called principal components (PC) and are representative of the variance, or essential differences, among the objects in question, e.g., oil samples. PCs are then ordered by their contribution to the variance of the data set. Thus, the first PC explains more of the variation in the data set than the second one. Each linear combination contains the original variables and a set of coefficients, also known as loadings. By the same token, each object can be described by a linear combination of PCs with a new set of coefficients, called scores. A score plot, which shows the distribution of objects, can then be generated. Likewise, a loading plot displays the distribution of variables. Both score and loading plots are typically used to display the results of a PCA analysis.

Score and loading plots, generated from the oil samples studied herein, are shown in Figure 1a,b, respectively. The score plot was constructed with PC1 and PC2 variables from each sample and shows the distribution of oil samples. On the basis of these data, the 20 samples were roughly divided into three groups. Group I contained samples X02, X03, X13, X17, X18, X19, and X20. Group II consisted of X04, X05, X06, X07, X08, X09, and X10, and Group III contained samples X1, X12, and X16. Loading values for PC1 and PC2 are plotted, respectively, along the abscissa and ordinate (Figure 1b). Each triangle in Figure 1b represents the elemental composition (or theoretical *m/z*) observed in the samples. Each

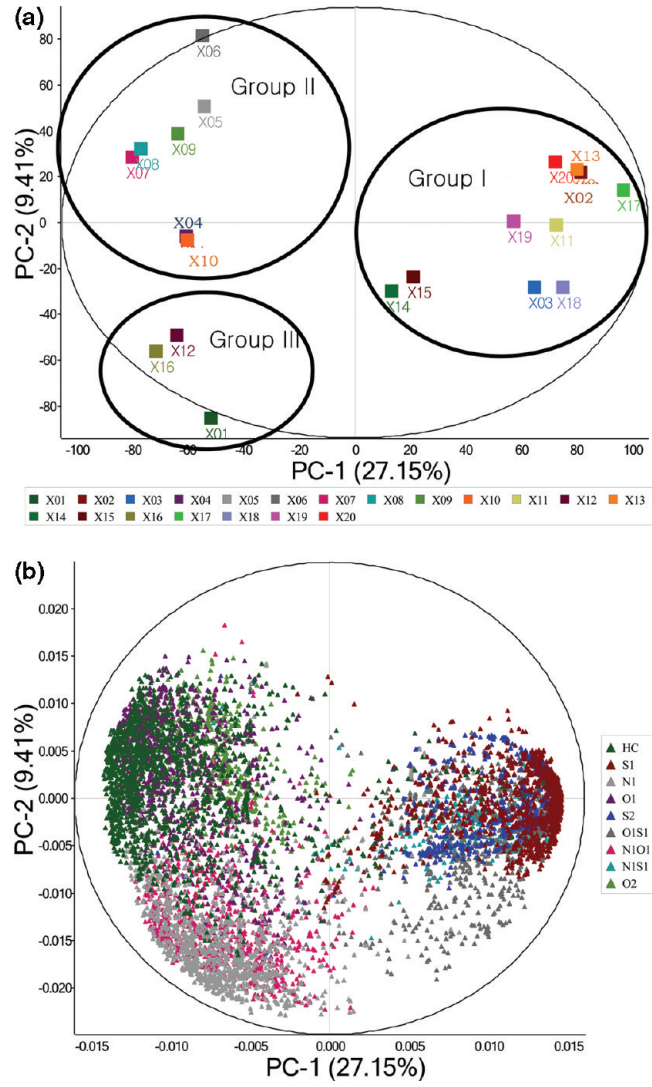


**Figure 1.** Score (a) and loading (b) plots are shown for PCA of 20 APPI positive mode spectra.

class of compounds was color coded so that the class distribution could be easily recognized and evaluated. For example, $S_1$ class compounds are marked in red brown, and hydrocarbon class compounds are shown in dark green. Sulfur-containing classes such as $S_1$, $S_2$, and $N_1S_1$ are located primarily on the right-hand side of the loading plot. This implies that the samples in Group I were rich in sulfur. Through the same comparison, Group II samples were shown to be rich in hydrocarbons (HC) and $O_1$- and $O_2$-class compounds. These results demonstrate the utility of PCA to describe the similarities and/or differences between complex spectra.

(27) Giri, N. C. *Multivariate statistical analysis*; Marcel Dekker: New York, 2004; p 558.
(28) Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*; Prentice Hall: Upper Saddle River, N.J., 2007; p 773.

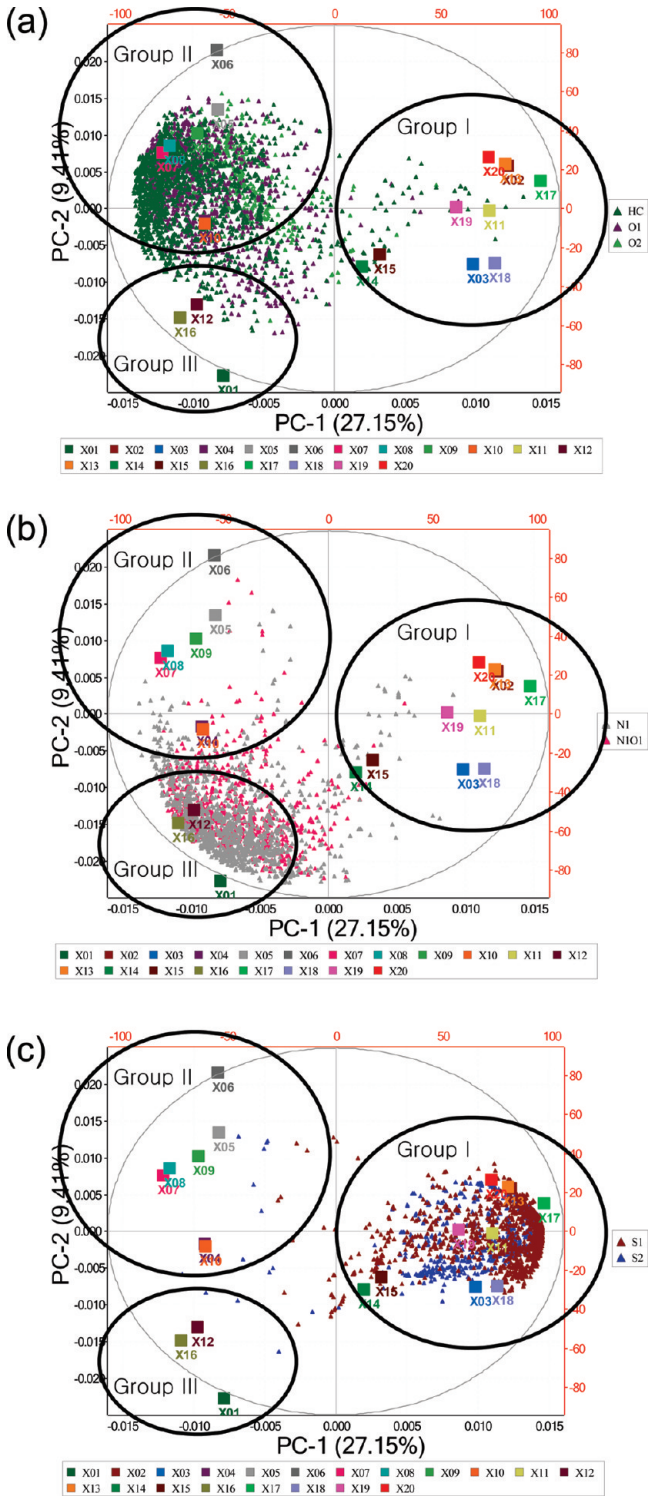


**Figure 2.** Biplots show the relationships between the score and loading plots displayed in Figure 1 for (a) HC, $O_1$, and $O_2$, (b) $N_1$ and $N_1O_1$, and (c) $S_1$ and $S_2$ class compounds.

A dual score-loading plot, or biplot, which is essentially an overlap of score and loading plots, can be employed for simpler interpretation. Figure 2 shows three biplots for HC, $O_1$, and $O_2$ (Figure 2a), $N_1$ and $N_1O_1$ (Figure 2b), and $S_1$ and $S_2$ (Figure 2c) class compounds. To generate Figure 2, PCA was performed with all compound classes. However, in this case, other classes were hidden to reduce the complexity of the plot. Figure 2 helps visualize the relationships between the samples and observed peaks. The bulk properties of the crude oils are listed in Table 2. A comparison of Figure 2 with Table 2 shows a good correlation between the PCA results and the bulk properties. It is interesting to note that the Group I crude oils had higher sulfur and vanadium content than the rest of the samples (refer to Table 2), while the Groups II and III crude oil samples generally contained low levels of vanadium and sulfur. Only a few peaks were observed to have vanadium in their elemental composition, and no peaks were identified with sulfur and vanadium atoms together. Therefore, the comparison between separation of samples displayed in Figure 2 (e.g., Group I vs Groups II and III) and their properties implies that the correlational relationship exists between vanadium content and sulfur containing molecules. Correlational relationship between bulk sulfur and vanadium contents in crude oil samples were previously reported.[29] In addition, the Group III crude oils exhibited a high nitrogen content. Samples X14 and X15, located near the Group III cluster, were high in both sulfur and nitrogen.

**Table 2. Bulk Properties of Samples Used in This Study**

| sample number | sulfur (mg/L) | nitrogen (ppm) | TAN[a] (mg KOH/g) | vanadium (ppm) |
|---|---|---|---|---|
| X01 | 0.61 | 3861 | 2.05 | 17.5 |
| X02 | 2.87 | 1645 | 0.29 | 56.3 |
| X03 | 4.5 | –[b] | 3.5 | 105.6 |
| X04 | 0.13 | 949 | 0.58 | 0 |
| X05 | 0.09 | – | 0.54 | – |
| X06 | 0.08 | – | 0.87 | – |
| X07 | 0.08 | 2865 | 2.34 | 0 |
| X08 | 0.11 | 1884 | 4.26 | 0 |
| X09 | 0.2 | 3231 | 1.46 | 1.3 |
| X10 | 0.13 | 1654 | 0.25 | 1.7 |
| X11 | 4.79 | 2136 | 0.27 | 54.6 |
| X12 | 0.18 | 3532 | 0.79 | 3.7 |
| X13 | 2.71 | 1350 | 0.18 | 28.2 |
| X14 | 1.85 | 3536 | 0.25 | 308.7 |
| X15 | 2.01 | 4011 | 0.18 | 329.8 |
| X16 | 0.25 | 4405 | 3.15 | 0 |
| X17 | 3.77 | 1620 | 0.3 | 43.8 |
| X18 | 3.57 | 4019 | 0.45 | 93.5 |
| X19 | 3.53 | 3123 | 0.47 | 108.3 |
| X20 | 1.91 | 900 | 0.38 | 11.5 |

[a] TAN, total acid number. [b] –, data not available.

**HCA and Heatmap Analysis of Positive Mode APPI Data.** Clustering analyses can be used to generate "clusters" of related objects. In HCA, a series of partitioning calculations are performed to group objects into clusters. Ward linkage[30,31] and Kendall rank correlations[32,33] were used in this study. In a ward linkage, clusters are formed at each step to minimize the sum of squares error, calculated by the following equation:

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (1)$$

where $X$ represents the value of an object and $\bar{X}$ is the estimated mean value of the cluster.

(29) Barwise, A. J. G. *Energy Fuels* **1990**, *4*, 647–652.
(30) Ward, J. H. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
(31) Yu, S.; Vooren, S. V.; Tranchevent, L.-C.; Moor, B. D.; Moreau, Y. *Bioinformatics* **2008**, *24*, i119–i225.
(32) Kendall, M. G. *Biometrika* **1938**, *30*, 81–93.
(33) Baumgartner, R.; Somorjai, R.; Summers, R.; Richter, W. *Magn. Reson. Imaging* **1999**, *17*, 1525–32.

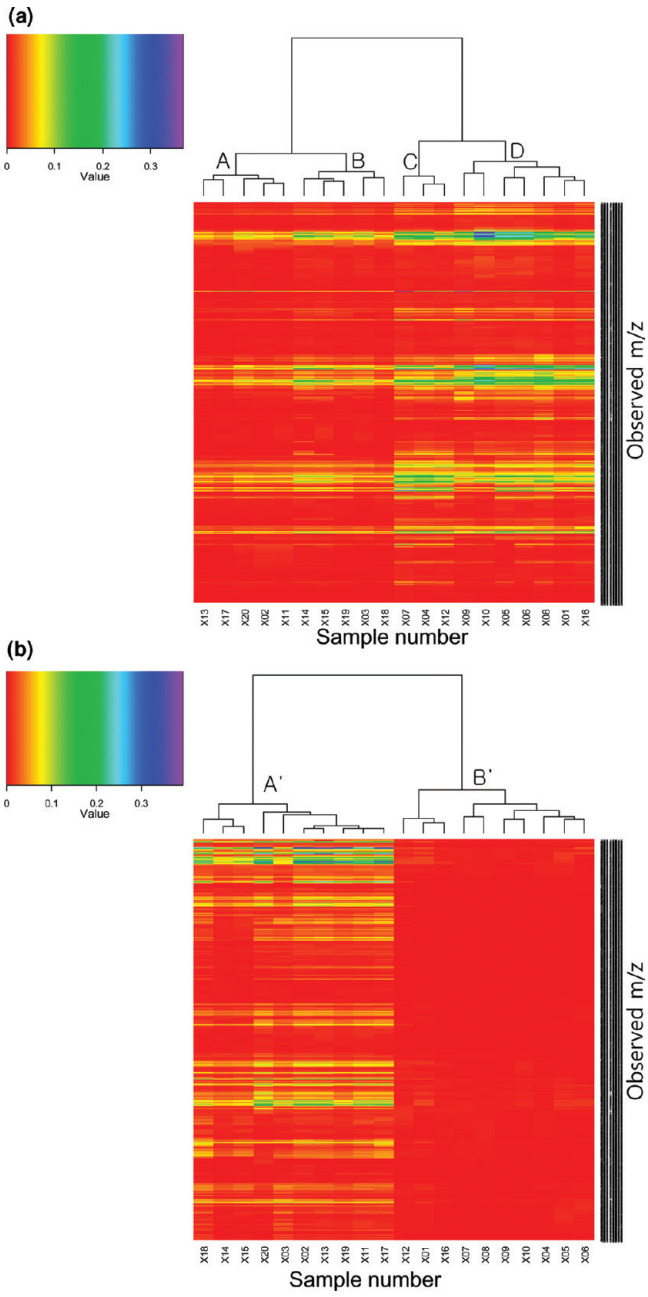


**Figure 3.** Heatmaps and dendrograms show the HCA results of the peaks presented in (a) Figure 2a and (b) Figure 2c. List of observed *m/z* in the ordinate of heatmap are provided in the supporting information.

The following equation determines the Kendall tau ($\tau$):

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \qquad (2)$$

where $n_c$ and $n_d$ are the numbers of concordant and discordant values, respectively. Two clusters in perfect positive correlation will exhibit $\tau = 1$. A perfect negative correlation results in $\tau = -1$.

For the oil samples analyzed here, HCA was used to reveal relationships between samples based on selected groups of peaks. Each of the data points presented in Figure 2a,c were selected for HCA. The resulting cluster patterns are given as dendrogram

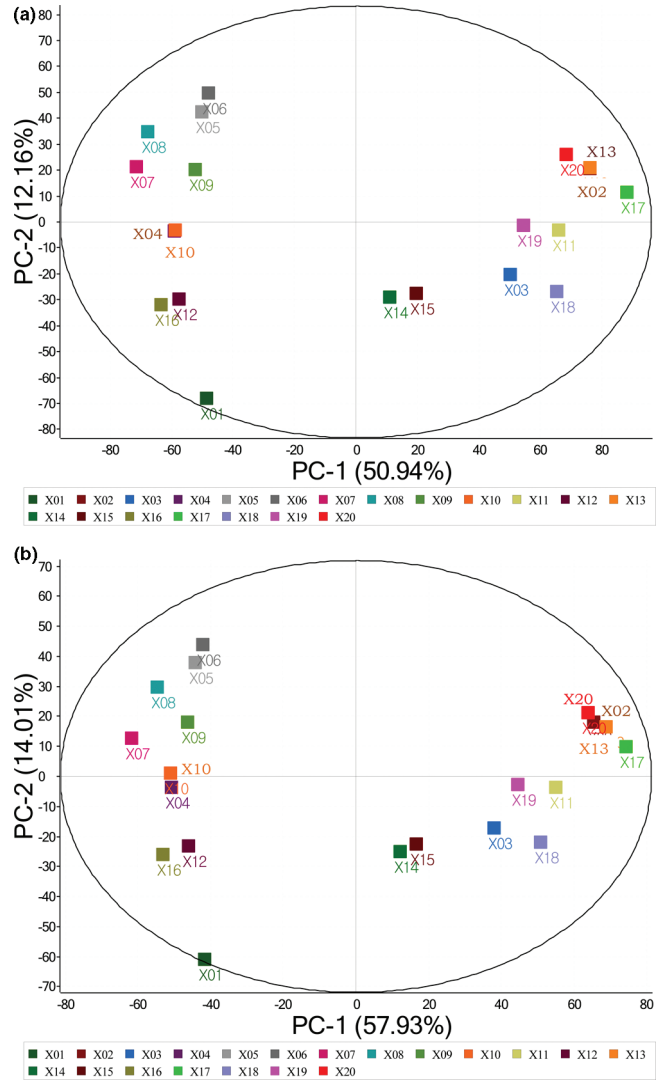


**Figure 4.** Score plots generated from selected peaks observed in at least (a) 5 and (b) 10 samples out of the 20 that were analyzed.

plots and heat maps in Figure 3. In Figure 3a, the crude oil samples were segregated largely into two clusters. A comparison of these clusters with the physical and chemical properties of the bulk oils in Table 2 reveals that samples in subcluster A and B contained a higher percentage of vanadium. For example, subcluster C was composed of samples X07, 04, and 12. These three samples contained 0, 0, and 3.7 ppm vanadium, respectively. In contrast, all of the samples in subcluster B (X14, 15, 19, 03, and 18) contained more than 90 ppm vanadium. The A cluster contained samples with moderate vanadium content ranging from 11.5 to 54.6 ppm.

A heat map generated by HCA of the peaks in Figure 2c, which represent the sulfur-containing compounds $S_1$, $S_2$, and $N_1S_1$, is shown in Figure 3b. A clear distinction was observed in relative abundance patterns between the two groups. Samples in subcluster A′ contained more sulfur than samples in subcluster B′. These examples show that a combination of HCA and PCA can be used to study detailed distributions of selected peaks in FTICR MS spectra.

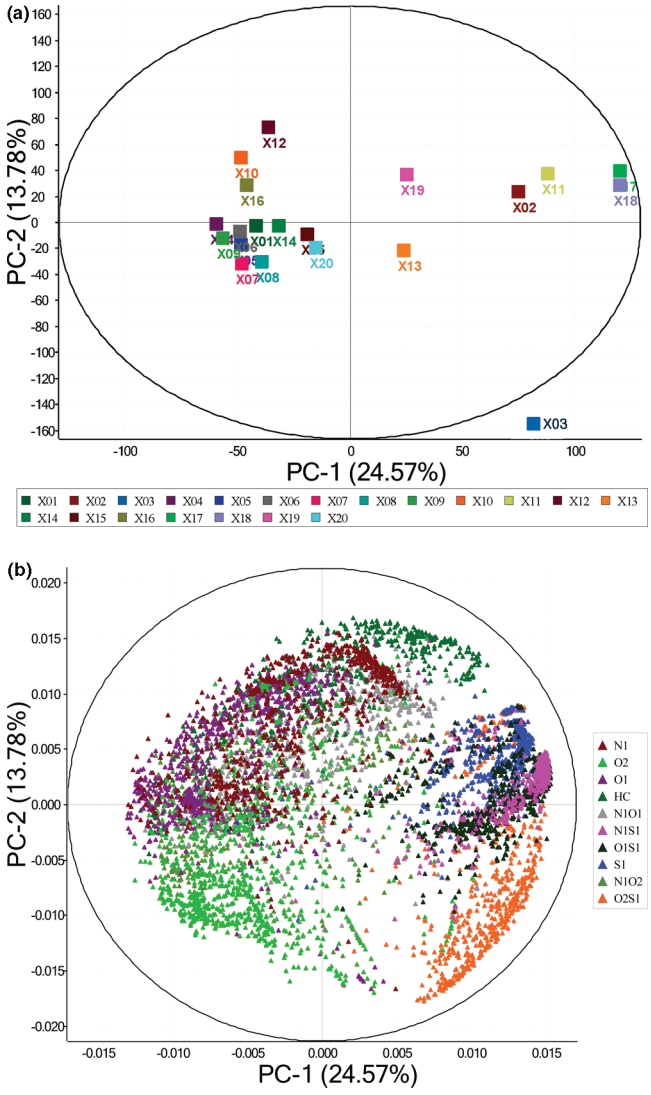


**Figure 5.** Score (a) and loading (b) plots are shown for PCA of 20 APPI negative mode spectra.

**Effect of Zero Values on PCA Results.** Appropriate treatment of missing values may be important in multivariate analyses.[34] In this study, the relative abundance of peaks not observed in samples was set to zero when merging data. Therefore, it is possible that the results of PCA analyses were determined by the number of zero values rather than by the variance in relative abundance among samples. To test this possibility, PCA was performed only with peaks that were observed in at least 5 or 10 out of the 20 samples; the corresponding score plots are shown in Figure 4a,b, respectively. Both of the score plots in Figure 4, as well as the distribution of classes in the loading plots (not shown), were similar to those shown in Figure 1, in which PCA was performed with all of the observed peaks. Therefore, the PCA results presented herein were not unduly influenced by zero values and were a function of the variance in the observed relative abundance.

**PCA of Negative Mode Data.** PCA was performed on negative mode APPI data after the generation of a merged table.

(34) Pedreschi, R.; Hertog, M.; Carpentier, S. C.; Lammertyn, J.; Robben, J.; Noben, J. P.; Panis, B.; Swennen, R.; Nicolai, B. M. *Proteomics* **2008**, *8*, 1371–1383.

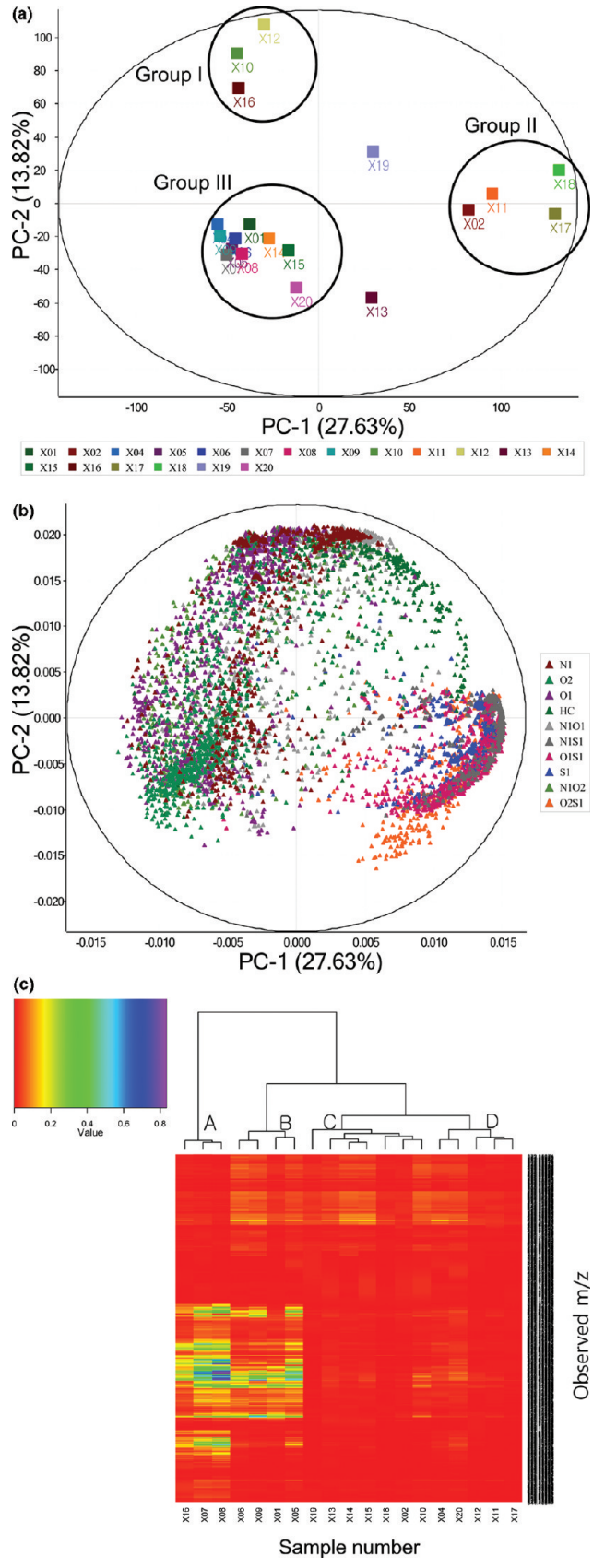


**Figure 6.** Score (a) and loading (b) plots generated from PCA data of 19 of the 20 APPI negative mode spectra, and (c) heatmap and dendrogram based on $O_2$ and $O_1$ class of compounds. Sample XO3 was excluded (refer to the text for more detailed information). List of observed *m/z* in the ordinate of heatmap are provided in the supporting information.

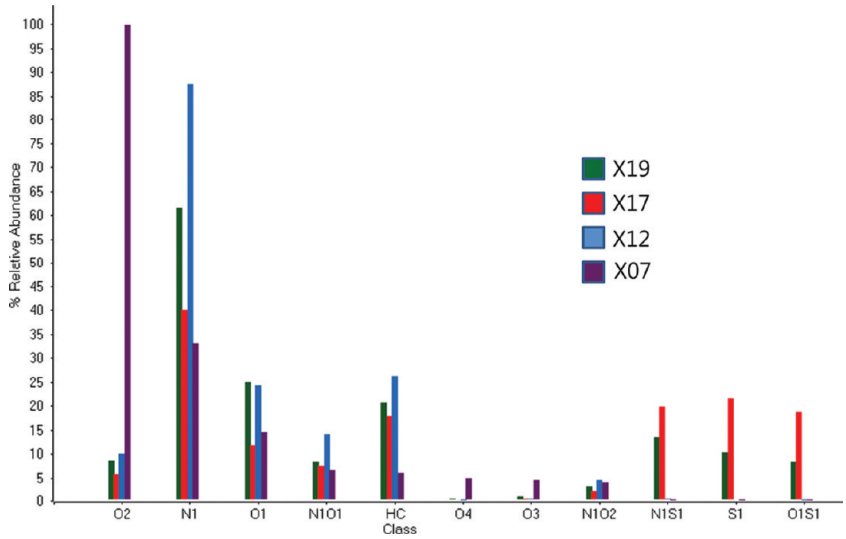


**Figure 7.** Class distribution of negative mode APPI spectra, shown for samples representing each group in the PCA score plot.

The resulting score and loading plots for all 20 samples are shown in Figure 5a,b, respectively. The separation between groups in the score plot was not as distinct as that observed for positive mode data. Note that sample X03, which was the only unconventional sample (bitumen extracted from oil sands), was well separated from all of the other samples. In Figure 5b, each peak was color-coded by class, as was done for the positive mode data. The plot shows that $S_xO_y$ ($x \geq 1$ and $y \geq 1$) compounds were uniquely abundant in sample X03. These results demonstrate that unique sample components can be identified with PCA.

The marginal separation among samples shown in Figure 5a is most likely due to the uniqueness of sample X03; by comparison, all of the other samples appeared tightly grouped. PCA was, therefore, repeated with the exclusion of X03. The resulting score and loading plots are shown in Figure 6a,b. The grouping and separation among groups in Figure 6 were markedly better than in Figure 5. In Figure 6b, peaks were largely divided into three groups. The sulfur-containing classes, e.g., $N_1S_1$ and $S_1O_1$, are primarily on the right side of the loading plot, with nitrogen-containing compounds, e.g., N and NO, at the top. Oxygen-containing $O_2$ classes are seen on the left side of the loading plot.

A comparison of Figures 1 and 6 shows differences in grouping characteristics between crude oils subjected to positive and negative mode APPI. These differences may arise from the different sensitivities exhibited by each of these ionization techniques toward different classes of compounds. For example, $O_2$ class compounds were one of the major products in negative mode analyses while HC or $S_1$ class compounds were the major products in positive mode analyses.

$O_2$ and $O_1$ class compounds, at the left-hand side of the loading plots, were selected for further analysis by HCA (refer to Figure 6c). The samples were divided into three clusters. All of the samples in A and B clusters exhibited very high TAN values. For example, three samples in cluster 1 has TAN values 2.34, 4.26, and 3.15 (refer to Table 2). In contrast, TAN values were generally low for the samples in the cluster C and D. The TAN is a measure of the amount of potassium hydroxide required to neutralize the acids in 1 g of oil and indicates the acidity of the crude. Therefore, the $O_1$ and $O_2$ class compounds observed in negative mode APPI were found to be related to sample's acidity. It is important to note that the clustering at the top of the dendrogram did not always agree with our observation of the heatmap. For example, the clusters B, C, and D are grouped at the top of the dendrogram in Figure 6c. The distribution of peak intensity in the heatmap shows that A and B clusters are likely to share more common features. However, the clusterings at the lower level were fairly consistent, regardless of peak selection and methods used for clustering.

Four samples (X07, X12, X17, and X19) were selected for class distribution studies, shown in Figure 6. Sample X17 was representative of Group I and contained an abundance $S_1$, $N_1S_1$, and $O_1S_1$ compounds. Sample X12 (Group II) was rich in $N_1$ and $N_1O_1$ compounds, and sample X07 (Group III) contained an abundance of $O_2$ compounds. Sample X19 was located between Groups I and II in Figure 5. The class distribution in Figure 7 shows significant amounts of both nitrogen- and sulfur-containing compounds in X19. The class distribution, therefore, matched well with the PCA results.

## CONCLUSIONS

Statistical analyses were applied to high-resolution mass spectra obtained from 20 oil samples by FTICR MS. PCA and HCA were shown to be useful for molecular-level identification of compounds based on similarities and/or differences between groups of peaks in complex mass spectra. PCA was used to examine the overall distribution of peaks and samples. HCA was employed to reveal relationships between samples based on selected groups of peaks. Overall, it was demonstrated that statistical interpretation represents a promising new approach for studies involving complex petroleum spectra.

PCA and HCA were chosen because they are two of the most frequently used statistical methods in these types of scientific studies. However, several other techniques could also be employed. For example, during the course of the current study, it was observed that the correlation method could be useful for establishing a relationship between high-resolution mass spectra and chemical or physical properties of petroleum samples. Research is currently underway using this methodology.

## ACKNOWLEDGMENT

The authors thank Dr. Hyunsik Kim, Dr. Myungchul Choi, and Mr. Seungyoung Kim for helpful discussions on APPI analyses of petroleum. M.H. wishes to thank Dr. Hojoon Seo and Ms. Somi Shin for their help with the data analysis software. This research was supported by the Converging Research Center Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0081904).

## SUPPORTING INFORMATION AVAILABLE

Lists of observed *m/z* in the ordinate of heatmaps presented in Figure 3a,b and Figure 6. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 4, 2009. Accepted November 16, 2009.

AC901748C