Definition and Characterization of Petroleum Compositional Families Using Principal Component Analysis of Gasoline and Saturate Fraction Compositional Ratios[1]

K.G. Osadetz [2], N. Pasadakis [3], and M. Obermajer [1]

Osadetz, K.G., Pasadakis, N., and Obermajer, M. (2002): Definition and characterization of petroleum compositional families using principal component analysis of gasoline and saturate fraction compositional ratios; *in* Summary of Investigations 2002, Volume 1, Saskatchewan Geological Survey, Sask. Industry and Resources, Misc. Rep. 2002-4.1, p3-14.

Abstract

Gasoline range hydrocarbons (GRH) and >210°C boiling point saturate fraction hydrocarbons (SFH) carry information regarding petroleum systems often obscured by compositional diversity and the simultaneous effect of multiple processes on single compounds. Multivariate statistical methods enhance the analysis and interpretation of these fractions, especially in conjunction with independent information. In the Williston Basin, the composition of these fractions follow polycyclic terpane and sterane biomarker traits. However, only Family A oils, from Ordovician sources, are sufficiently distinctive to be classified using derived variables from principal component (PC) compositional analysis. Families B, C, and D oils, from Middle Devonian–Tournaisian source rocks, have characteristic compositions that corroborate biomarker-based families. However, compositional variation within individual families overlaps interfamilial variations such that they are insufficient to be an independent classification tool. Consideration of additional derived variables using multiple models assists the identification of families B, C, and D.

Keywords: Williston Basin petroleum systems, light hydrocarbon, saturate hydrocarbon, principal component analysis.

1. Introduction

Petroleum systems are defined by linking petroleum to source kerogen through petroleum-petroleum and petroleum-source rock correlations. The petroleum-petroleum correlation defines compositionally related "families" originating from a "single source rock" by considering processes that introduce compositional variation such as thermal maturation, migration, water washing, and biodegradation (Dow, 1974; Magoon and Dow, 1994). Petroleum-source rock correlation links a petroleum family to a stratigraphic unit or facies containing the source kerogen (Curiale, 1994). These two analyses characterize the petroleum system and assist in the identification of undiscovered resources (DeMaison and Murris, 1984). Petroleum systems have also been proposed where source rock descriptions are absent or deficient (e.g. Lenheer and Zumberge, 1987; Osadetz *et al.*, 1994).

Biomarkers are generally structurally complicated compounds that occur in low abundance and have high molecular weight, such as polycyclic alkanes. These are usually the preferred basis of petroleum family definition because the biochemical, sedimentological, and physical processes that singly or competitively affect their compositional variations are understood (Moldowan *et al.*, 1985; Peters and Moldowan, 1993). Several shortcomings arise, however, from the use of biomarkers alone to correlate petroleum facies. For example, petroleum composition and compositional data obtained from analytical protocols are commonly ignored (e.g. Osadetz *et al.*, 1992). Additionally, reliance is placed on thermally non-persistent compounds whose relative abundance can result from or be affected by competitive processes that obscure origin (e.g. Peters and Moldowan, 1991). The effects of important processes, especially mixing, might not be discernable where absolute concentrations vary between mixed components (Jiang and Li, in press). Biomarker techniques neither objectively define criteria for separating families, nor describe the internal compositional variations within a single family. While pitfalls can be avoided by careful study, even consideration of many compositional characteristics in large sample sets (e.g. Osadetz *et al.*, 1992) may miss important petroleum system features (Jiang and Li, in press). Determination of whether abundant simple compounds exhibit characteristics and variations consistent with an interpretation based on less abundant but more complicated compounds is therefore desirable.

---

[1] GSC Contribution Number 2002011.
[2] Geological Survey of Canada, 3303 - 33rd Street NW, Calgary, AB T2L 2A7.
[3] Mineral Resources Engineering Department, Technical University of Crete, Chania 73100 Greece.

This paper discusses improvements in petroleum-petroleum correlations by using multivariate statistical analysis of easily obtained, thermally persistent organic compounds. We have used this exploratory statistical approach to analyze petroleum systems previously defined in the Williston Basin using multiple biomarker studies. Specifically, we performed principal component analysis (PCA) on compositional ratios of hydrocarbons of the gasoline (iC_5 to nC_8) range hydrocarbons (GRH) and saturate fraction hydrocarbons (SFH) having greater than 210 °C boiling point from petroleum families in the Williston Basin (Osadetz *et al.*, 1992). Recent developments in the analysis of Williston Basin petroleum systems make this expedient (Jiang and Li, in press; Jarvie, 2001). PCA is also useful as an example of statistical evaluation and characterization of compositional variation. The findings indicate that PCA maximizes the ability to characterize familial relations among SFH and GRH.

2. Previous Study of Williston Basin Petroleum Systems

Protracted and episodic subsidence of the Williston Basin, a sub-circular epicratonic basin approximately 800 km in diameter (Figure 1), resulted in preservation of up to 5 km of Phanerozoic strata overlying eroded Archean Superior and Hearne/Wyoming cratons (Gerhard *et al.*, 1982; Ahern and Mrkvicka, 1984; Leclair *et al.*, 1997). Petroleum occurs throughout the Phanerozoic succession (Figure 2) in a variety of traps associated with Precambrian basement structure (Clement, 1987; Gerhard *et al.*, 1987; LeFever *et al.*, 1987) and stratigraphic relationships (Podruski *et al.*, 1988).

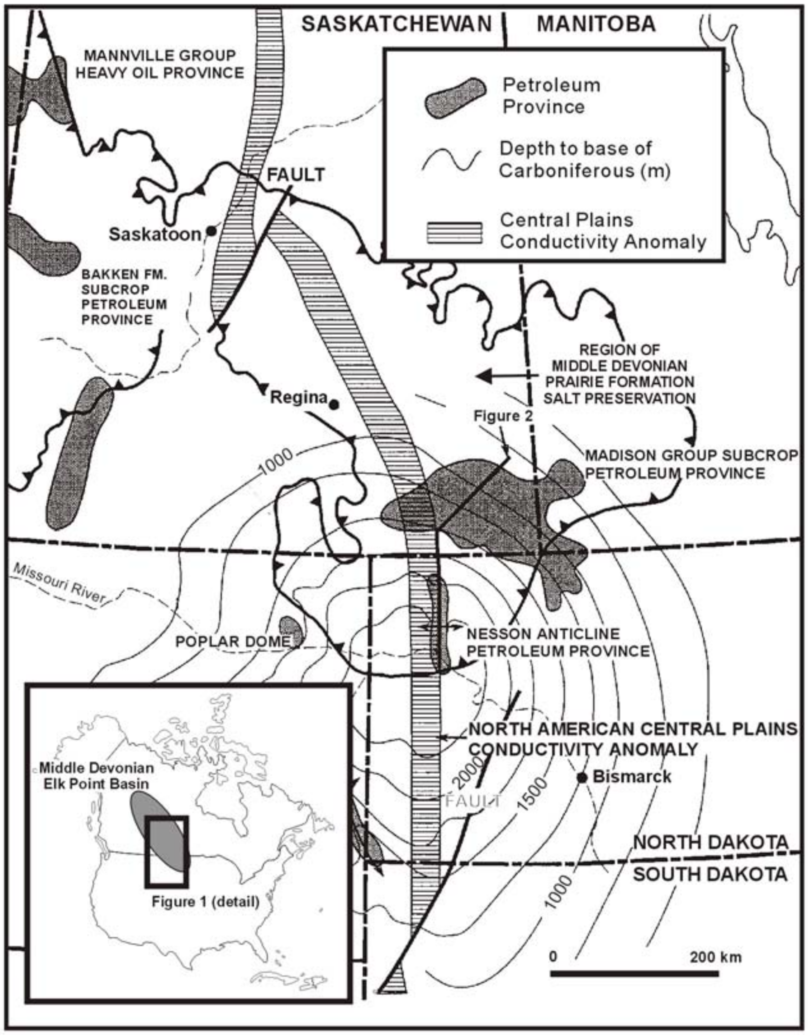

Figure 1 - Location map showing the main geological and geophysical elements of the Williston Basin and environs. The region of anomalous subsidence that is the Williston Basin proper (Ahern and Mrkvicka, 1984) is generally coincident with the 1 km depth contour on Carboniferous strata. The region of preserved Middle Devonian Prairie Formation salt deposited in the Elk Point Basin is illustrated. Inset shows location of the Williston Basin and the extent of the Elk Point Basin. Samples from petroleum pools entrapped at the subcrop of the Upper Paleozoic succession in southeastern Saskatchewan and southwestern Manitoba, as well as American samples, constitute the sample set for this study (following Burrus** et al.**, 1996b).

Williams (1974) first classified Williston Basin petroleum families and Dow (1974) defined the petroleum system concept used here. They recognized three oil types, based on stable isotopic and gasoline range hydrocarbon compositions. Type1 oil, occurring predominantly in Ordovician and Silurian reservoirs, was attributed to Middle Ordovician Winnipeg shale sources. Type II oil, found in Upper Devonian, Mississippian and Mesozoic reservoirs, was inferred to have Fammenian-Tournaisian Bakken Formation source rocks. Type III oil, restricted to Pennsylvanian reservoirs, was correlated to Pennsylvanian Tyler Formation source rocks. The isotopic compositions of carbon and sulphur in oils indicated similar family divisions and, together with other data, showed some pools to be water washed and biodegraded (Bailey *et al.*, 1973a, 1973b; Thode, 1981; Thompson, 1983). Using source rock samples from the Bakken Formation and 16 crude oil samples from Paleozoic reservoirs ranging in age from Cambrian to Mississippian, Zumberge (1983) and Leenheer

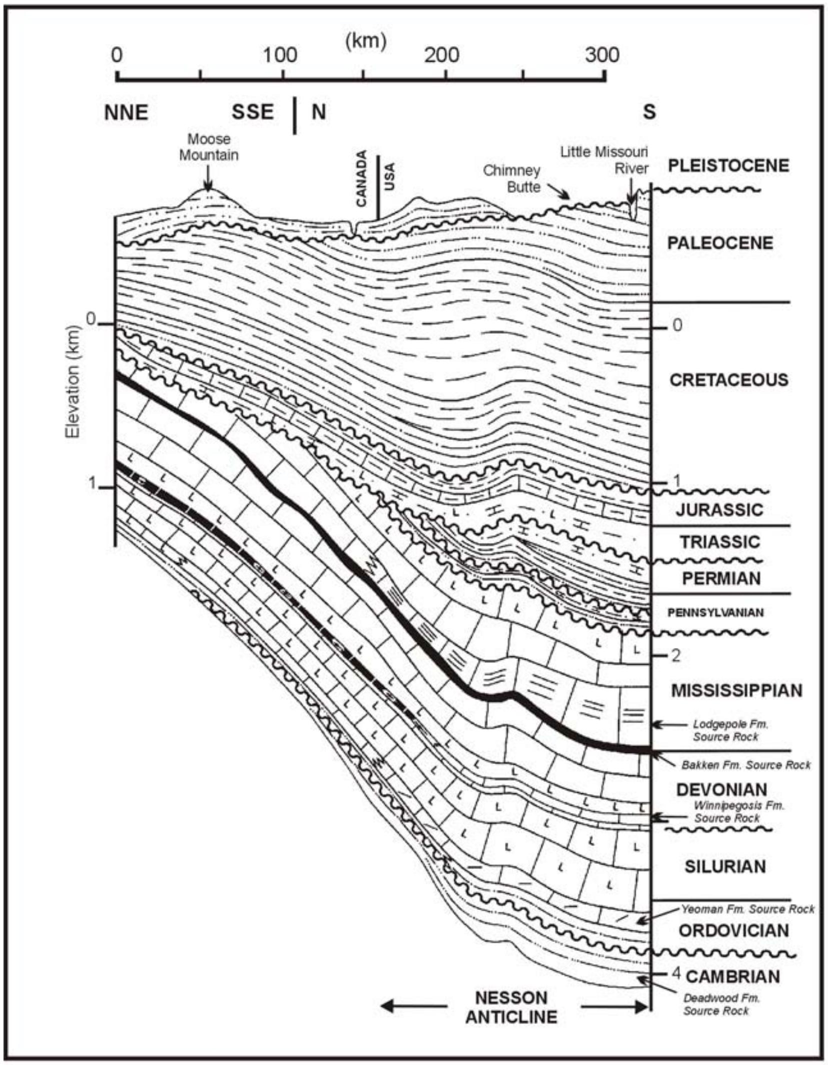

Figure 2 - Section of the Williston Basin illustrating the Phanerozoic stratigraphic succession and the stratigraphic position of the four effective Paleozoic petroleum source rocks for the four oil families A, B, C, and D (Osadetz* et al.*, 1992) analyzed in this paper (following Burrus* et al.*, 1996b).

and Zumberge (1987) further organized the petroleum families into five groups based on biomarkers and stable carbon isotope compositions.

Brooks *et al.* (1987) outlined a new petroleum-systems model for the Canadian Williston Basin that was modified and correlated to source rocks by Osadetz *et al.* (1992, 1994) and extended into the U.S.A. by Price and LeFever (1994), Osadetz *et al.* (1995), and Obermajer *et al.* (2000). Petroleum was subdivided into four compositional families using combinations of terpane, sterane, and select n- and iso-alkane characteristics (Osadetz *et al.*, 1992).

Family A oil is common to Upper Ordovician reservoirs and has distinctive n-alkane compositions and low acyclic isoprenoid/ pentacyclic terpane ratios. These compositions and ratios match solvent extracts from kukersites of the Upper Ordovician Bighorn Group (Osadetz *et al.*, 1992; Osadetz and Snowdon, 1995), but not extracts from shales of the Winnipeg Formation as initially suggested (Dow, 1974; Williams, 1974).

Family B oils are similar to Type II oils and they have high relative diasterane abundance without a prominence in the pentacyclic terpanes. They are found mainly in the Bakken Formation, but may occur in reservoirs as young as Early Cretaceous. The revised petroleum system model relates Family B oils to a source within the Bakken shale (Osadetz *et al.*, 1992, 1994; Osadetz and Snowdon, 1995)

Family C oils also are similar to Type II oils, but differ from Family B oils in that they have low relative diasterane abundance and a prominence of C_{35} pentacyclic terpane. They are found mainly in Mississippian and Mesozoic strata and are sourced from bituminous carbonates of the Mississippian (Tournaisian) Lodgepole Formation. Both Family B and C oils have been identified in the American Williston Basin (Price and LeFever, 1994; Obermajer *et al.*, 2000).

Family D oils have low tricyclic/pentacyclic terpane ratios, but lack the GRH and n-alkane characteristics of Family A oils. Family D oil likely correlates to Groups 3, 4, and 5 of Leenheer and Zumberge (1987). Family D oils can be subdivided into D1 and D2 oils based on n-alkane/acyclic isoprenoid ratios and stratigraphic occurrence (Osadetz *et al.*, 1992). Family D1 oil is found mainly in younger Devonian reservoirs and, as yet, has an undefined source. Thin organic-rich beds in Winnipegosis platform carbonates, the Birdbear Formation, and other Upper Devonian rocks have been suggested as potential source rocks (Osadetz *et al.*,1992; Osadetz and Snowdon, 1995; Obermajer *et al.*, 1999). Family D2 oil occurs in pinnacle reefs of the Middle Devonian Winnipegosis Formation. It is sourced from the Brightholme Member that was deposited in a basinal setting between the pinnacle reefs of the Winnipegosis Formation.

Oils having similar molecular compositions to D2 oils have been found in the Upper Cambrian Deadwood Formation (Leenheer and Zumberge, 1987; Fowler *et al.*, 1998), Silurian pools of the Nesson Anticline (Downey, 1996), and new discoveries in the Middle Ordovician Winnipeg Formation. They have, however, very different isotopic compositions of carbon and sulphur, suggesting that a still-undescribed petroleum system exists in Paleozoic strata (Osadetz *et al.*, 1998, 2000). Indeed, Jarvie and Inden (1997), Jarvie and Walker (1997), and Jarvie (2001), have indicated that many more petroleum systems and sub-systems are present in the Williston Basin than have previously been recognized. The degree and consequences of mixing of petroleum families is also under debate. Such mixing could greatly impair the fundamentals of biomarker analysis (Burrus *et al.*, 1996a, 1996b; Jiang *et al.*, 2001; Li and Jiang, 2001; Jiang and Li, in press;), and rigorous examination of a wider range of organic compounds for petroleum-system identification within the Williston Basin is likely to be helpful.

3. Analytical Techniques

Hydrocarbons in the gasoline range (iC5 to nC8) were analyzed by gas chromatography on whole oil samples. Saturated hydrocarbons that have boiling points above 210 °C were measured using open column liquid chromatography. Complete analytical results and techniques are described in Obermajer *et al.* (2000). Gas chromatograms of typical samples of GRH and SFH of the four petroleum families defined by biomarkers are illustrated in Figure 3. The gas chromatography mass-spectrometry (GCMS) analyses used to identify more complex organic components was performed and reported separately (Osadetz *et al.*, 1992; Obermajer *et al.*, 2000).

4. Statistical Techniques

a) Exploratory Multivariate Statistics

Principal Component Analysis (PCA) is a statistical method used to “discover” patterns and relations within a data set (Pearson, 1901; Hotelling, 1933). In this study, we use PCA to determine commonality among GRH and SFH measured in 171 samples of petroleum from the Williston Basin. This multivariate statistical method derives a set of variables known as principal components. Each principal component attempts to account for the largest possible portion of the original total variance of the data. Successive principal components explain progressively less of the original variance. PCA is useful in situations where experiments or observations result in large amounts of data that may obscure relationships among variables. In this work, we use PCA for variable reduction and classification purposes. Multivariate statistical modeling has been shown by Odden and Kvalheim (2000) to be efficient for the discrimination of petroleum sources. In the Canadian Williston Basin, Snowdon and Osadetz (1988) employed correspondence and factor analysis to characterize and classify gasoline range components. Petroleum systems are well suited to such exploration by PCA because molecular compositions result from a complex interaction of biological, environmental, geological, and physical processes working competitively and simultaneously.

b) Principal Component Analysis Inputs and Models

Samples and compositional data used in this analysis are essentially similar to those discussed by Obermajer *et al.* (2000), an augmentation of the set discussed by Osadetz *et al.* (1992). We used samples from four previously defined compositional families (A to D of Osadetz *et al.*, 1992), but excluded D1 oils because their petroleum-source rock correlation is uncertain (Downey, 1996; Obermajer *et al.*, 2000). All variables used in PCA models were standardized by subtracting the mean and dividing by the standard deviation, a common practice to ensure an equal sample weight.

Although PCA is commonly used to reveal variables that influence relationships within a dataset, selecting the original variables prior to analysis is also possible and is aided by using existing models of system behaviour and experience consistent with specific hydrocarbon fractions and experiments. In this study, the number of original variables was reduced by using sequential PCA models or independent information. In effect, we chose those variables most likely to: a) represent interesting data patterns, b) permit the interpretation of variable loadings, and c) maximize the classification potential of the PCA. PCA also permits identification of compositional characteristics associated with specific chemical and geological processes, the most important of which, for petroleum system analysis, is source rock composition. PCA calculations were performed using the “UNSCRAMBLER” computer program. To help interpret the results of PCA, we used the previously performed classification based on biological markers (Osadetz *et al.*, 1992).

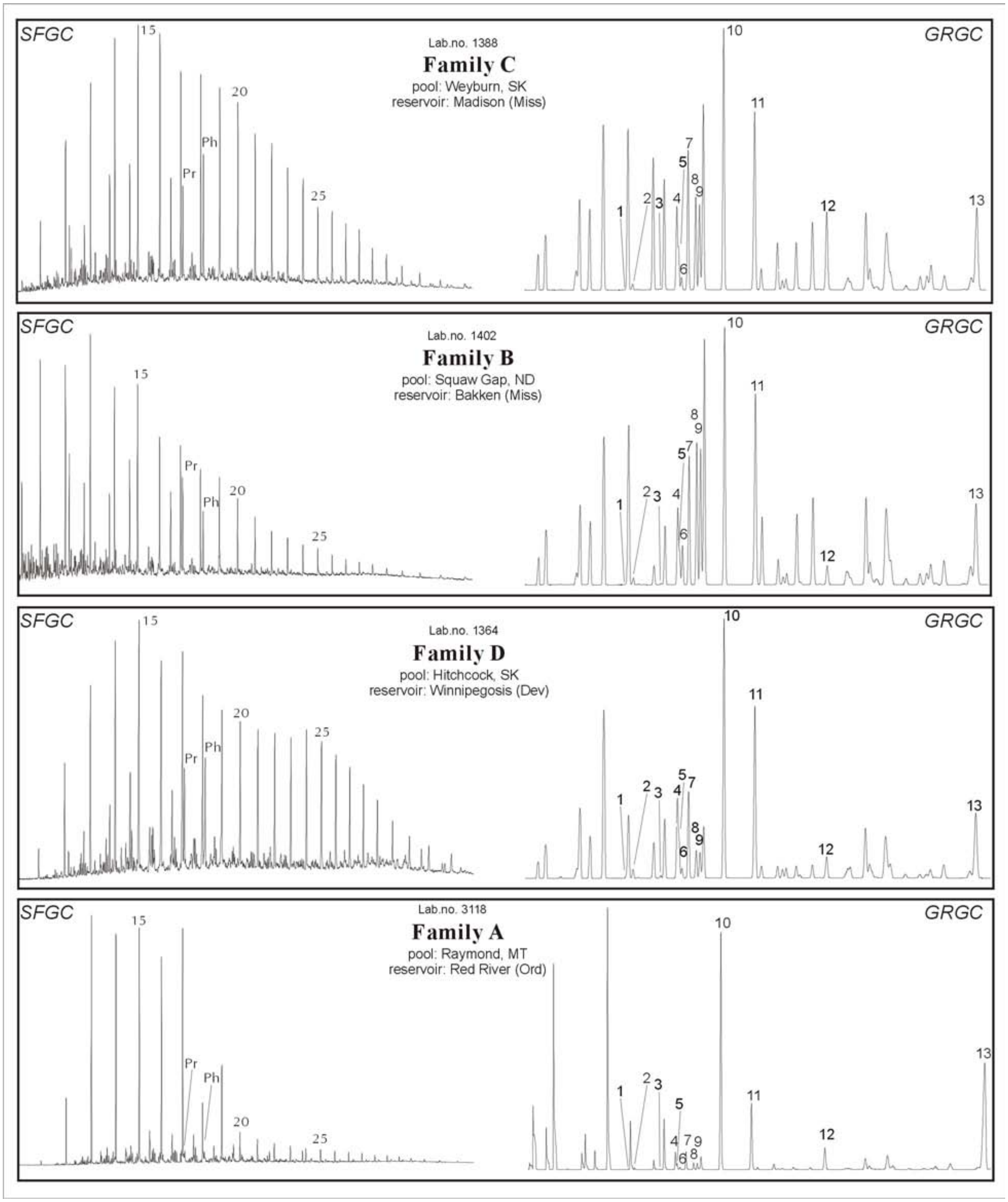

Figure 3 - Representative +210°C saturate fraction and gasoline range (i-C_5H_{12} to n-C_8H_{18}) gas chromatograms showing compositional variations in the analyzed Williston Basin oils. Examples of each of the four compositional families that are defined by independent analysis of terpane and sterane biological marker compounds (Osadetz** et al., **1992; Obermajer** et al., **2000) are illustrated. Specific compounds identified include: Pr-pristane, Ph-phytane, 15, 20, 25 - C_{15}, C_{20}, C_{25} normal alkanes; 1) 2,2-dimethylpentane; 2) 2,4-dimethylpentane; 3) 3,3-dimethylpentane; 4) 2-methylhexane; 5) 2,3-dimethylpentane; 6) 1,1-dimethylcyclopentane; 7) 3-methylhexane; 8) 1c3-dimethylcyclopentane; 9) 1t3-dimethylcyclopentane; 10) heptane; 11) methylcyclohexane; 12) toluene; and 13) octane.

c) Principal Component Analysis of Petroleum Compositional Families

Two independent models were used to explore hydrocarbon compositions for identification of petroleum systems: 1) Gasoline Range Ratios Model (GRRM) and 2) Saturate Fraction Ratios Model (SFRM).

Model 1: Gasoline Range Ratios Model

The Gasoline Range Ratios Model (GRRM) uses original variables composed of compositional ratios commonly employed in the interpretation of GRH. Beginning with all standard compositional ratios used to describe gasoline range hydrocarbons, successive analyses of principal components indicated the strongest discrimination of the four oil families is provided by only four compositional ratios. The four variables are: Paraffin Index I (Isoheptane Value); the branched to total C7 compound ratio; the 2,4-dimethylpentane to 2,3-dimethylpentane ratio; and the K1 value. The four variables were defined by Thompson (1983) and Mango (1987, 1990), who attributed specific interpretive significance to them.

Sample Scores

Sample scores describe a position in principal component space, and each original variable has loadings that describe their contribution to each principal component. The sample scores of the first two principal components of the GRRM and the respective loading diagrams are presented in Figure 4. Several linear gradients are observed in the data by comparing the sample scores of the first two principal components. When the samples are associated with their biomarker-based oil families, all four petroleum families exhibit distinctive gradients defined by the first two principal components of the GRRM. These gradients indicate that both distinctive family characteristics and linear compositional variations of the original variables exist within each family. Sample scores of Family A oils exhibit the most coherent grouping and are characterized by a tight positive gradient defined by positive PC1 scores and negative PC2 scores. There is considerable overlap of Family B with families C and D, and a slight overlap of Family D with Family A. The strong gradients of families B and C are characterized by negative PC1 scores and mildly negative to strongly positive PC2 scores. Family C commonly has more positive PC2 scores for a given value of PC1 than Family B, but not enough to be uniquely distinguishable. Family D samples are generally positive for both PC1 and PC2. The GRRM provides a clearer characterization of Family D oils than other methods.

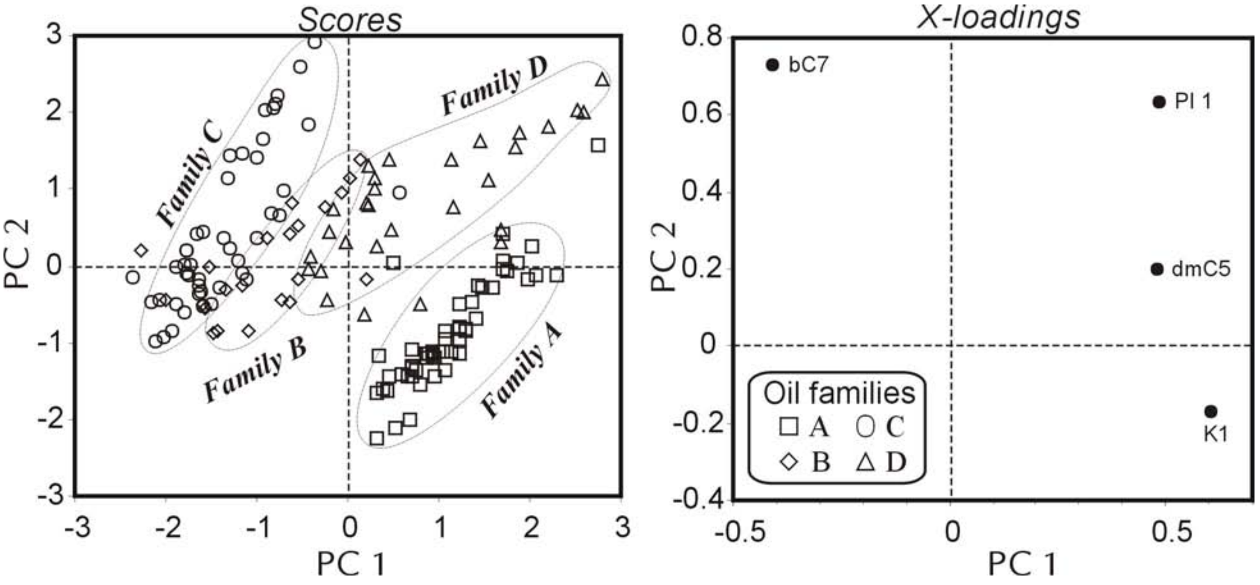

Figure 4 - Sample scores (left) and original variable loadings (right) for the first two principal components resulting from the Gasoline Range Ratio Model (GRRM) compositional principal component analysis of selected Williston Basin petroleum oils. Sample symbols indicate compositional families determined by independent analyses of terpane and sterane biological marker compositions (Osadetz** et al.**, 1992; Obermajer** et al.**, 2000). Families A, B, C, and D are generally characterized by principal components controlled by the relative loadings of gasoline range ratios that include K1 factor (Mango, 1987, 1990), the ratio of branched to total C7 compounds (bC7), and Paraffin Index 1 or Isoheptane Value (PI1). Family A is clearly distinguished, but Family B overlaps the fields of families C and D. Variations of PC2 are strongly affected by Paraffin Index 1 loadings, which are interpreted to be characterized thermal maturity variations, except that the gradient in Family C sample scores distinguishes between the aromatic enriched and aromatic impoverished compositions. This variation in Family C scores is interpreted to be a mixing response (see text).

Variable Loadings

Variable loadings may help to understand the role and importance of the original variables. The 2,4-dimethylpentane to 2,3-dimethylpentane ratio (dmC5) has positive loadings of three principal components, whereas the K1 value has positive PC1 and negative PC2 and PC3 loadings. These variable loadings control the gradient that separates independently defined oil families. The Paraffin Index I is characterized by high positive PC1 and PC2 loadings and mildly negative PC3 loadings. This variable loading is the principal control on the internal variation within families. The branched to total C7 compound ratio has negative PC1 and PC3 loadings, but a positive PC2.

Interpretation and Analysis

The GRRM confirms and improves the familial classification defined by biomarker analysis (Osadetz *et al.*, 1992). In particular, the GRRM provides an improved characterization of Family D samples in PC1 vs. PC2 space. There is a strong gradient among the four families that is controlled by the loadings of both K1 factor and the branched to total C7 compound ratio. This demonstrates that K1 is primarily a source indicator (ten Haven, 1996; Obermajer *et al.*, 2000), contrary to the initial interpretation of this parameter (Mango, 1987).

The discrimination of Family B from Family C in PC1 vs. PC2 space is fairly clear, and the distinction between them is further improved by the consideration of PC3. However consideration of PC3 does not effectively separate families A and D. The variations along the gradient of Family C samples distinguishes aromatic-enriched samples from those that are aromatic impoverished. Family C samples that overlap the range of Family B in the GRRM have fewer aromatic compounds in the GRH, whereas those that are distinctive compared to the range of Family B scores are enriched in benzene and toluene. This variation in Family C is probably due to mixing of Lodgepole-sourced and Bakken-sourced oils rather than to water washing as previously interpreted (Thompson, 1988). This analysis also suggests that a discrimination diagram plotting Paraffin Index 1 vs. the 2,4-dimethylpentane to 2,3-dimethylpentane ratio would distinguish families B, C, and D, which previous studies found the most difficult to classify.

An important feature of the GRR model is the sub-parallel orientation of the four familial gradients indicating linear variations within each (Figure 4). The similarity of these gradients suggests the influence of a common process on the internal linear variation of sample scores within each family. This process is inferred to be related to thermal maturity because the observed gradients are strongly controlled by the Paraffin Index 1, which is widely accepted to reflect thermal maturation (Thompson, 1988). A most interesting feature of this model is the almost orthogonal relationship between the loadings of Paraffin Index 1 and the K1 factor. This suggests the weak effective impact of thermal maturity on the K1 factor, as reported by Schaefer and Littke (1988). We therefore conclude that K1 is a source indicator independent of thermal maturity effects (ten Haven;1996) and that source classification is improved by combining the K1 factor with the branched to total C7 compound ratio. This is apparent when the relatively effective classification of Family C oils in the GRRM is compared to the less effective results of semi-quantitative methods (Obermajer *et al.*, 2000). The classification of compositional families defined through biomarkers is evidently distinct in the gasoline range, supporting its general usefulness.

Model 2: Saturate Fraction Ratios Model

The original variables used in the Saturate Fraction Ratios Model (SFRM) include the five compositional ratios and factors Pr/Ph, Pr/nC17, Ph/nC18, and the carbon preference indices for both lighter (nC14 to nC20) and heavier (nC22 to nC30) alkanes of the SFH. Ratios of the original peak areas are standardized, as in previous models, without additional modification. This model was constructed to test the discriminating power of SFGC ratios that were ascribed significance by the biomarker-based classification scheme (Osadetz *et al.*, 1992). The sample scores of the first two principal components of the SFRM and the respective loading diagrams are presented in Figure 5.

Sample Scores

In general, sample scores illustrate two gradients (Figure 5). A tight negative gradient is displayed by samples with PC1 scores always less than -1.0, and the other more dispersed negative gradient by samples with commonly positive PC1 scores. When the samples are compared to biomarker-based oil families, Family A is again clearly distinguished by consistently negative PC1 scores and a linear variation of PC2 with PC1. Family A PC2 values range between -2.6 and 1.2. The other three families are distributed along the more dispersed negative gradient of PC1 scores that are generally between -1.0 and 3.0. Among these, Family B samples have the generally highest PC2 scores for a given value of PC1. This commonly distinguishes Family B samples from Family C samples that generally have lower PC2 scores. As in previous models, the fields of PC1 and PC2 in Family D overlap those of other families, particularly B and C, effectively obscuring their separation. In detail, however, Family D samples

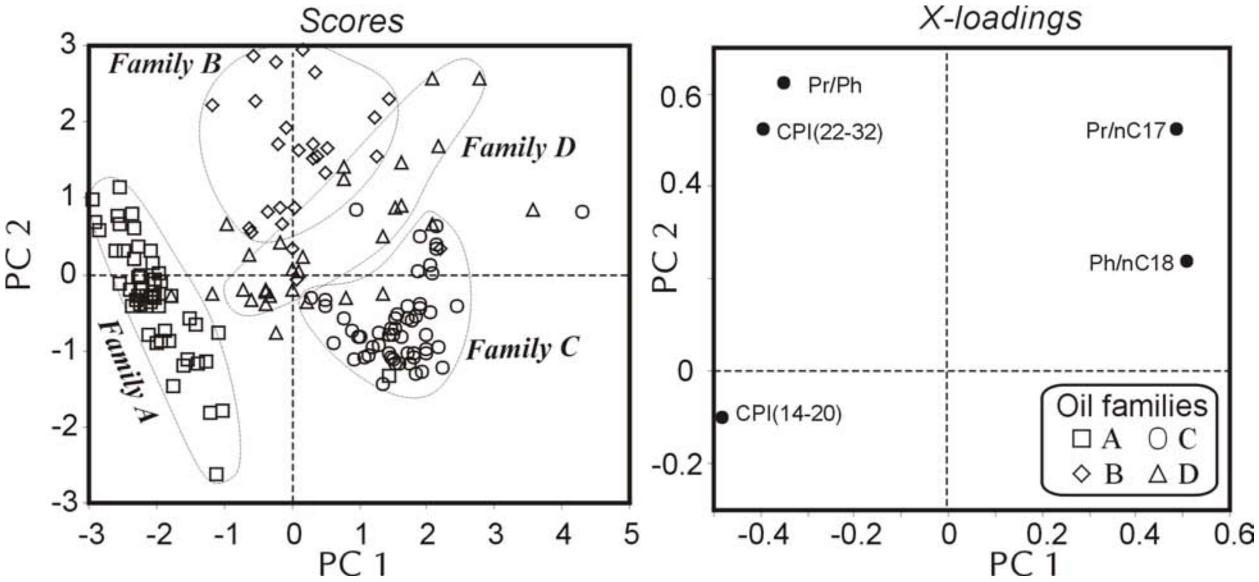

Figure 5 - Sample scores (left) and original variable loadings (right) for the first two principal components resulting from the Saturate Fraction Ratio Model (SFRM) compositional principal component analysis of selected Williston Basin oils. Sample symbols follow compositional families determined by independent analyses of terpane and sterane biological marker compositions (Osadetz* et al.*, 1992; Obermajer* et al.*, 2000). Family A is effectively characterized by the first two principal components. Other families exhibit characteristic sample scores, but they are not distinguished clearly from one another. Families B and C have relatively distinctive sample scores, considering the indications in other models for overlapping scores that are attributed to mixing.

appear to fall along a positively correlated gradient in PC1 vs. PC2 space, in sharp contrast to the three other families, all of which tend toward lower PC2 scores and increasing PC1 scores.

Variable Loadings

The variable loadings of the SFRM indicate a strong discrimination between data types. The isoprenoid/n-alkane ratios have strongly positive PC1 and PC2 loadings. The carbon preference index (CPI) and the Pr/Ph ratio are negative in PC1 loadings and positive in PC2 loadings. In contrast, the negative PC2 loading of the CPI for the light n-alkanes distinguishes it from both the CPI of the heavy n-alkanes and the Pr/Ph ratio. Somewhat unexplained is the lack of discriminating power of the Pr/nC17 and Ph/nC18 ratios with respect to families B and C. Osadetz *et al.* (1992) proposed that a cross plot of these two figures distinguished Family B and C oils as effectively as any of the biomarker parameters. Obermajer *et al.* (2000), Jarvie (2001), and Jiang and Li (in press) supported this inference. Jiang and Li (in press) used the variation of Pr/Ph to identify the potential degree of mixing between pristine Family C and Family B oils. The fact that the loadings of these two variables are not a clear discriminator in this model suggests that, whereas these ratios may be locally significant for distinguishing between these two specific families, they are not universally useful in discriminating across the entire sample set.

Interpretation and Analysis

The SFRM model has strong potential for classifying families A, B, and C. For example, the distinctive gradient of Family A cannot be attributed only to differences in thermal maturity as it is also associated with variable loadings such as Pr/Ph, commonly attributed to source rock depositional environments. More noticeable are two gradient trends between the Ordovician-sourced and Devono-Carboniferous-sourced oils, which appear to be controlled by loadings of Pr/nC17 and the CPI for the light n-alkanes. This variation is a general discriminant of source rock stratigraphic age, although Family D internal variations indicate that additional factors control these derived variables. The orthogonal gradient of Family D samples suggests that the processes affecting their composition are distinct from the attributed source rock and thermal maturity variations in this and previous models. Also important is the distinction between Family B and Family C samples that is controlled by the loadings of the Pr/Ph ratio and the heavy n-alkane CPI. Both of these are commonly interpreted as indications of source rock depositional environment (Peters and Moldowan, 1993). Compared to the continuous and overlapping variations of sample scores in previous models, this model illustrates distinct and characteristic sample scores for families B and C. If overlap of Family C and B scores in some models is from mixing of end members, then a more pronounced overlap of scores than observed in the SFRM might be expected.

5. Geochemical Significance of Principal Component Analyses

Origins and variations in the composition of GRH and SFH are critical to understanding the Williston Basin petroleum systems. Although thermally fragile compounds such as polycyclic alkanes are in low abundance, they are widely employed as the preferred basis of petroleum family definition (Moldowan *et al.*, 1985; Peters and Moldowan, 1993). It is desirable to determine whether classifications based on more abundant, simpler hydrocarbons are consistent with those based on less abundant components that may be destroyed during catagenesis. Ideally, the most easily obtained, most abundant and most thermally persistent components should be used for petroleum family definition.

This work demonstrates that GRH and SFH retain important information regarding petroleum origins, composition, and the petroleum system. Principal component analysis has helped to identify relations among petroleum constituents that may otherwise be obscured by compositional diversity. PCA indicates multiple processes simultaneously affect GRH and SFH compositional variables. These processes may complicate interpretation and reduce the usefulness of compositional variable for classification. However, familial characteristics and common processes can be discerned from reduced variable sets, possibly with superior results to those obtained using less abundant, more complicated compounds. In addition, the results of PCA suggest that the common interpretation of simple compounds must be performed carefully. PCA also permits of samples to be scored using factor loadings.

The best classifications of oil families result from GRRM with SFRM producing less distinctive relationships. Only Family A oils from Ordovician kukersite sources have GRH and SFH compositions sufficiently distinctive to allow classification using only principal component compositional models. Families B, C, and D oils from Middle Devonian–Tournaisian sources have distinctive GRH and SFH characteristics, but compositional overlap among families is sufficient to reduce confidence of definitive classification.

The impact of PCA for benchmark definition of petroleum systems is significant. Revisions of these benchmark petroleum systems are themselves important, but less so than the recognition that standard biomarker ratio analyses may be inadequate to define petroleum systems. The degree to which biomarker analysis cannot define petroleum is unknown because current interpretive techniques neglect or find semi-quantitative confirmation of the biomarker-based classification in GRH and SFH (Osadetz *et al.*, 1992; Obermajer *et al.*, 2000). The hydrocarbon fractions analyzed in this study can be more effective at illustrating mixing, even where they are not effective for classification (Jiang and Li, in press). This study also suggests that certain standard compositional component and ratio interpretations should be re-evaluated. Like other studies (ten Haven, 1996; Obermajer *et al.*, 2000), it indicates that "K" value is controlled by source, independent of thermal maturity (Mango, 1987, 1990; Schaefer and Littke, 1988). Additionally, the variation of gasoline range aromatic hydrocarbons is primarily controlled by the mixing of two oils from different sources rather than by water washing as initially suggested (Thompson, 1983).

6. Conclusions

1) Gasoline range and >210°C boiling point saturate fraction hydrocarbons carry petroleum system information affected by multiple processes that simultaneously complicate compositional traits and limit their independent value for classification.
2) Principal components of these two fractions, interpreted separately or in combination, enhance the interpretation of petroleum systems, especially when combined with information from biological marker compounds.
3) In Williston Basin samples, GRH and SFH have polycyclic terpane and sterane biomarker compositions previously interpreted as indicative of source rock and thermal maturity characteristics, but herein related to a more complicated set of processes.
4) Only Family A oils, from Ordovician sources, have sufficiently distinctive compositions to be classified using derived variables from principal component analysis. Families B, C, and D oils, from Middle Devonian–Tournaisian sources, have overlapping hydrocarbon compositions that preclude their use as a primary classification tool. Consideration of multiple models, pair-wise models, and additional derived variables can minimize ambiguities. In particular, pair-wise discrimination using multiple models assists the identification of families B, C, and D.
5) Both principal component models and individual biomarker-defined families exhibit linear and non-linear compositional variations in Williston Basin oils due to:
   - kerogen composition,
   - compositional mixing,
   - source rock depositional environment, and
   - thermal maturity.

6) Despite their lack of diagnostic capability for familial classification, GRH and SFH may be better indicators of certain processes, such as compositional mixing, than more complicated compounds present in lower abundance. Therefore, the classification and analysis of petroleum systems benefit from a combination of biological marker analysis and principal component analysis of GRH and SFH.

7. References

Ahern, J.L. and Mrkvicka, S.R. (1984): A mechanical and thermal model for the evolution of the Williston Basin; Tect., v3, p79-102.

Bailey, N.J.L., Jobson, A.M., and Rogers, M.A. (1973a): Bacterial degradation of crude oil: Comparison of field and experimental data; Chem. Geol., v11, p203-221.

Bailey, N.J.L., Krouse, H.R., Evans, C.R., and Rogers, M.A. (1973b): Alteration of crude oils by waters and bacteria-evidence from geochemical and isotope studies; Amer. Assoc. Petrol. Geol. Bull., v57, p1276-1290.

Brooks, P.W., Snowdon, L.R., and Osadetz, K.G. (1987): Families of oils in southeastern Saskatchewan; *in* Carlson, C.G. and Christopher, J.E. (eds.), Proceedings of the Fifth International Williston Basin Symposium, Sask. Geol. Soc., Spec. Publ. No. 9, p253-264.

Burrus, J., Osadetz, K.G., Wolf, S., Doligez, B., Visser, K., and Dearborn, D. (1996a): A two-dimensional regional basin model of Williston Basin hydrocarbon systems; Amer. Assoc. Petrol. Geol. Bull., v80, p265-291.

Burrus, J., Osadetz, K.G., Wolf, S., and Visser, K. (1996b): Physical and numerical modeling constraints on oil expulsion and accumulation in the Bakken and Lodgepole petroleum systems of the Williston Basin (Canada-USA); Bull. Can. Petrol. Geol., v44, p429-445.

Clement, J.H. (1987): Cedar Creek: A significant paleotectonic feature of the Williston Basin; *in* Longman, M.W. (ed.), Williston Basin: Anatomy of a Cratonic Oil Province, Rocky Mtn. Assoc. Geol., Denver, p323-336.

Curiale, J.A. (1994): Correlation of oils and source rocks–a conceptual and historical prospective; *in* Magoon, L.B. and Dow, W.G. (eds.), The Petroleum System–From Source to Trap, Amer. Assoc. Petrol. Geol., Mem. 60, p251-260.

Demaison, G. and Murris, R.J. (eds.) (1984): Petroleum Geochemistry and Basin Evaluation; Amer. Assoc. Petrol. Geol., Mem. 35, 426p.

Dow, W.G. (1974): Application of oil-correlation and source-rock data to exploration in Williston Basin; Amer. Assoc. Petrol. Geol. Bull., v58, p1253-1262.

Downey, M. (1996): Variations of carbon and sulphur isotope ratios in oils from the Williston Basin and their relation to thermal maturity, oil-oil and oil-source correlations; unpubl. B.Sc. thesis, Univ. Calgary, 52p.

Fowler, M.G., Stasiuk, L.D., Li, M., Obermajer, M., Osadetz, K.G., and Idiz, E. (1998): Reexamination of the Red River petroleum system, southeastern Saskatchewan, Canada; *in* Christopher, J.E., Gilboy, C.F., Paterson D.F., and Bend S.L. (eds.), Proceedings of the Eighth International Williston Basin Symposium, Sask. Geol. Soc., Spec. Publ. No. 13, p11-13.

Gerhard, L.C., Anderson, S.B., and LeFever, J.A. (1987): Structural History of the Nesson Anticline, North Dakota; *in* Longman, M.W. (ed.), Williston Basin: Anatomy of a Cratonic Oil Province, Rocky Mtn. Assoc. Geol., Denver, p337-353.

Gerhard, L.C., Anderson, S.B., LeFever, J.A., and Carlson, C.G. (1982): Geological development, origin, and energy mineral resources of Williston Basin, North Dakota; Amer. Assoc. Petrol. Geol. Bull., v66, p989-1020.

Hotelling, H. (1933): Analysis of a complex of statistical variables into principal components; J. Educ. Psychol., v24, p417-441 and p498.

Jarvie, D.M. (2001): Williston Basin petroleum systems: Inferences from oil geochemistry and geology; The Mtn. Geol., v38, p19-41.

Jarvie, D.M. and Inden, R.F. (1997): Re-evaluation of Williston Basin potential Paleozoic source rocks and petroleum systems; Annual Meeting, Amer. Assoc. Petrol. Geol., Prog. Abstr., p55.

Jarvie, D.M. and Walker, P.R. (1997): Correlation of oils and source rocks in the Williston Basin using classical correlation tools and thermal extraction very high resolution C7 gas chromatography; *in* Proceedings of the 18th International Meeting on Organic Geochemistry, Maastricht, The Netherlands, Sept. 22-26, p51-52.

Jiang, C. and Li, M. (in press): Bakken/Madison petroleum systems in Canadian Williston Basin; Part 3: Geochemical evidence for significant Bakken-derived oils in Madison Group reservoirs, Org. Geochem.

Jiang, C., Li M., Osadetz, K.G., Snowdon, L.R., Obermajer, M., and Fowler, M.G. (2001): Bakken/Madison petroleum systems in the Canadian Williston Basin; Part 2: Molecular markers diagnostic of Bakken and Lodgepole source rocks, Org. Geochem., p1037-1054.

Leclair, A.D., Lucas, S.B., Broome, H.J., Viljoen, D.W., and Weber, W. (1997): Regional mapping of Precambrian basement beneath Phanerozoic cover in southeastern Trans-Hudson Orogen, Manitoba and Saskatchewan; Can. J. Earth Sci., v34, p618-634.

Leenheer, M.J. and Zumberge, J.E. (1987): Correlation and thermal maturity of Williston Basin crude oils and Bakken source rocks using terpane biomarkers; *in* Longman, M.W. (ed.), Williston Basin: Anatomy of a Cratonic Oil Province; Rocky Mtn. Assoc. Geol., Denver, p287-298.

LeFever J.A., LeFever R.D., and Anderson, S.B. (1987): Structural evolution of the central and southern portions of the Nesson Anticline, North Dakota; *in* Carlson, C.G. and Christopher, J.E. (eds.), Proceedings of the Fifth International Williston Basin Symposium, Sask. Geol. Soc., Spec. Publ. No. 9, p147-156.

Li, M. and Jiang, C. (2001): Bakken/Madison petroleum systems in the Canadian Williston Basin; Part 1: C21-C26 20-n-alkylpregnanes and their triaromatic analogs as indicators for Upper Devonian-Mississippian epicontinental black shale derived oils?; Org. Geochem., p667-675.

Magoon, L.B. and Dow, W.G. (1994): The petroleum system; *in* Magoon, L.B. and Dow, W.G. (eds.), The Petroleum System–From Source to Trap; Amer. Assoc. Petrol. Geol., Mem. 60, p3-24.

Moldowan, J.M., Seifert, W.K., and Gallegos, E.J. (1985): Relationship between petroleum composition and depositional environment of petroleum source rocks; Amer. Assoc. Petrol. Geol. Bull., v69, p1255-1278.

Mango, F.D. (1987): An invariance in the isoheptanes of petroleum; Sci., v237, p514-517.

__________ (1990): The origin of light cycloalkanes in petroleum; Geochim. Cosmochim. Acta, v54, p23-27.

Obermajer, M., Osadetz, K.G., Fowler, M.G., and Snowdon, L.R. (1999): Geochemistry and familial association of crude oils from the Birdbear Formation in southeastern Saskatchewan, Williston Basin; Bull. Can. Petrol. Geol., v47, p255-269.

__________ (2000): Light hydrocarbon (gasoline range) parameter refinement of biomarker-based oil-oil correlation studies: An example from Williston Basin; Org. Geochem., v31, p959-976.

Odden, W. and Kvalheim, O.M. (2000): Application of multivariate modeling to detect hydrocarbon components for optimal discrimination between two source rock types; Appl. Geochem., v15, p611-627.

Osadetz, K.G., Brooks, P.W., and Snowdon, L.R. (1992): Oil families and their sources in Canadian Williston Basin, (southeastern Saskatchewan and southwestern Manitoba); Bull. Can. Petrol. Geol., v40, p254-273.

Osadetz, K.G., Kohn, B.P., O'Sullivan, P., Feinstein, S., Hannigan, P.K., Everitt, R.A., Gilboy, C.F., Bezys, R.K., and Stasiuk, L.D. (1998): Thermotectonics of the Williston Basin and environs: Variations in heat flow and hydrocarbon generation; *in* Christopher, J.E., Gilboy, C.F., Paterson, D.F., and Bend, S.L. (eds.), Proceedings of the Eighth International Williston Basin Symposium, Sask. Geol. Soc., Spec. Publ. No. 13, p147-165.

Osadetz, K., Obermajer, M., Li, M., Snowdon L., Fowler, M., Hansen, W., LeFever, J., Diehl, P., Burke, R., Rygh, M., and Carroll, K. (2000): Oils in American Williston Basin; Amer. Assoc. Petrol. Geol./Soc. Econ. Paleont. Mineral., Annual Convention, New Orleans, April 2000, prog. abstr.

Osadetz, K.G. and Snowdon, L.R. (1995): Significant Paleozoic petroleum source rocks, their distribution, richness and thermal maturity in Canadian Williston Basin (southeastern Saskatchewan and southwestern Manitoba); Geol. Surv. Can., Bull. 487, 60p.

Osadetz, K.G., Snowdon, L.R., and Brooks, P.W. (1994): Oil families in Canadian Williston Basin southwestern Saskatchewan; Bull. Can. Petrol. Geol., v42, p155-177.

Osadetz, K.G., Snowdon, L.R., and Rygh, M.E. (1995): Oil families and their sources in American Williston Basin; *in* Hunter, L.D. and Schalla, R.A. (eds.), Proceedings of the Seventh International Williston Basin Symposium, Sask. Geol. Soc., Spec. Publ. No. 12, p451-452.

Pearson, K. (1901): On lines and planes of closets fit to systems of points in the space; Phil. Mag., v2, p559-572.

Peters, K.E. and Moldowan, J.M. (1991): Effects of source, thermal maturity, and biodegradation on the distribution and isomerization of homohopanes in petroleum; Org. Geochem., v17, p47-61.

__________ (1993): The Biomarker Guide; Prentice-Hall, Englewood Cliffs, 363p.

Podruski, J.A., Barclay, J.E., Hamblin, A.P., Lee, P.J., Osadetz, K.G., Procter, R.M., and Taylor, G.C. (1988): Conventional Oil Resources of Western Canada (Light and Medium), Part 1: Resource Endowment; Geol. Surv. Can., Pap. 87-26, 149p.

Price, L.C. and LeFever, J.A. (1994): Dysfunctionalism in the Williston Basin: The Mid-Madison/Bakken Petroleum System; Bull. Can. Petrol. Geol., v42, p187-218.

Schaefer, R.G. and Littke, R. (1988): Maturity-related compositional changes in the low-molecular-weight hydrocarbon fraction of Toarcian shales; *in* Mattavelli, L. and Novelli, L. (eds.), Advances in Organic Geochemistry 1987, Org. Geochem., v13, p887-892.

Snowdon, L.R. and Osadetz, K.G. (1988): Geological processes interpreted from gasoline range analyses of oils from southeast Saskatchewan and Manitoba; *in* Current Research, Part D, Geol. Surv. Can., Pap. 88-1D, p33-40.

ten Haven, H.L. (1996): Applications and limitations of Mango's light hydrocarbon parameters in petroleum correlation studies; Org. Geochem., v24, p957-976.

Thode, H.G. (1981): Sulphur isotope ratios in petroleum research and exploration: Williston Basin; Amer. Assoc. Petrol. Geol., Bull., v65, p1527-1535.

Thompson, K.F.M. (1983): Classification and thermal history of petroleum based light hydrocarbons; Geochim. Cosmochim. Acta, v47, p303-316.

__________ (1988): Gas-condensate migration and oil fractionation in deltaic systems; Marine Petrol. Geol., v5, p237-246.

Williams, J.A. (1974): Characterization of oil types in Williston Basin; Amer. Assoc. Petrol. Geol. Bull., v58, p1243-1252.

Zumberge, J.E. (1983): Tricyclic diterpane distributions in the correlation of Paleozoic crude oils from the Williston Basin; *in* Bjoroy, M. (ed.), Advances in Organic Geochemistry 1981, John Wiley & Sons Ltd., New York, p738-745.