Reliable Peak Selection for Multisample Analysis with Comprehensive Two-Dimensional Chromatography

Reliable Features for Comparative Analysis with Comprehensive Two-Dimensional Chromatography

Stephen E. Reichenbach,[*,†] Xue Tian,[†] Akwasi A. Boateng,[‡] Charles A. Mullen,[‡] Chiara Cordero,[*,¶] and Qingping Tao[*,§]

University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA, Sustainable Biofuels and Co-Products Research Unit, USDA-ARS, Eastern Regional Research Center, Wyndmoor PA 19038-8598, USA, Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy, and GC Image, LLC, PO Box 57403, Lincoln NE 68505-7403, USA

E-mail: reich@cse.unl.edu; chiara.cordero@unito.it; qtao@gcimage.com

Abstract

Comprehensive two-dimensional chromatography is a powerful technology for analyzing the patterns of constituent compounds in complex samples, but matching chromatographic features across large sample sets is difficult. Various methods have been described for pairwise peak matching between two chromatograms, but, for most features, the pairwise matches are incomplete or inconsistent across many chromatograms. This paper describes a new, automated method for selecting chromatographic peaks that reliably correspond across many chromatograms. Reliably corresponding peaks can be used both for directly comparing relative compositions and for aligning chromatographic data for comprehensive comparative analyses of large numbers of samples. The Consistent Cliques Method (CCM) for selecting reliable features for a set of patterns represents all pairwise feature matches in a graph, finds the maximal cliques, and then combines cliques with shared vertices to extract reliable features. For comprehensive two-dimensional chromatography, CCM takes pairwise peak matches between chromatograms and then determines the peaks that are consistently matched across many chromatograms. The parameters of CCM are the minimum number of chromatograms with complete pairwise peak matches and the desired number of reliable peaks. A particular threshold for the minimum number of chromatograms with complete pairwise matches ensures that all matches for reliable peaks are conflict-free. Experimental results with samples of complex bio-oils analyzed by comprehensive two-dimensional gas chromatography (GCxGC) coupled with mass spectrometry (GCxGC-MS) indicate that CCM provides a good foundation for comparative analysis of complex chemical mixtures.

[*]To whom correspondence should be addressed
[†]University of Nebraska – Lincoln
[‡]USDA-ARS
[¶]Università degli Studi di Torino
[§]GC Image

Introduction

Feature matching is the problem of establishing correspondences among attributes of different objects. In some pattern analysis problems, features or attributes are explicitly labeled so the correspondences are known and feature matching is not required. For example, if fish are to be classified on the basis of physical attributes such as weight and length, the values of those attributes are labeled such that one value is known to be the weight and another value is known to be the length. Then, weights are compared to weights and lengths are compared to lengths.

In other pattern analysis problems, features or attributes are not labeled and correspondences must be inferred. For example, a classic problem requiring feature matching is image alignment.[1] Features such as edges or corners are detected in each image, but the correspondences, *e.g.*, which corner in one image matches to which corner in another image, are unknown. Feature matching establishes such correspondences and is important for a variety of tasks including analyzing, aligning, and comparing patterns.

The motivating application for this research is the analysis of large numbers of complex multidimensional patterns in data produced by comprehensive two-dimensional gas chromatography (GCxGC), comprehensive two-dimensional liquid chromatography (LCxLC), or other comprehensive two-dimensional separation technologies. GCxGC separates complex mixtures using two columns interfaced by a modulator and connected to a detector.[2,3] If the chromatographic separation is fully effective, each compound induces a brief, isolated peak in the two-dimensional data. The GCxGC chromatogram of a complex mixture will exhibit hundreds or thousands of peaks, each of which is a characteristic feature of the data from that sample. Figure 1 illustrates the most relevant region of a GCxGC chromatogram of a complex bio-oil in which the x-axis is the retention time (RT) in the first chromatographic column, the y-axis is the retention time in the second chromatographic column, and the color indicates the relative value of the signal. In this image, the value at each pixel is the total intensity count (TIC) of the mass spectrum at the corresponding retention times and the pseudocolor is determined by a conventional cold-hot color map (in which the color progression blue, cyan, green, yellow, and red indicates increasing value), with automated value mapping to accentuate smaller peaks.[4]

A fundamental problem in GCxGC data analysis is to identify the compound that induces each peak — in other words, the labeling of each peak with its chemical identity. If the peak for a known compound is identified in each chromatogram, *e.g.*, by its retention times and/or spectrum, then that feature of the sample data can be labeled and compared directly across samples. However, even when the chemical identity for a peak cannot be established definitively, as is common for peaks in complex mixtures, comparative analysis requires that peaks be uniformly labeled (*e.g.*, with an identification number) across samples so that peaks resulting from the same compound in different samples are recognized as such (even if the compound identity is unknown). Therefore, comprehensive comparative analyses of complex samples by two-dimensional chromatography requires feature matching.

Figure 2 illustrates the problem of matching peak features for uniform labeling. In the top sub-image of a chro-

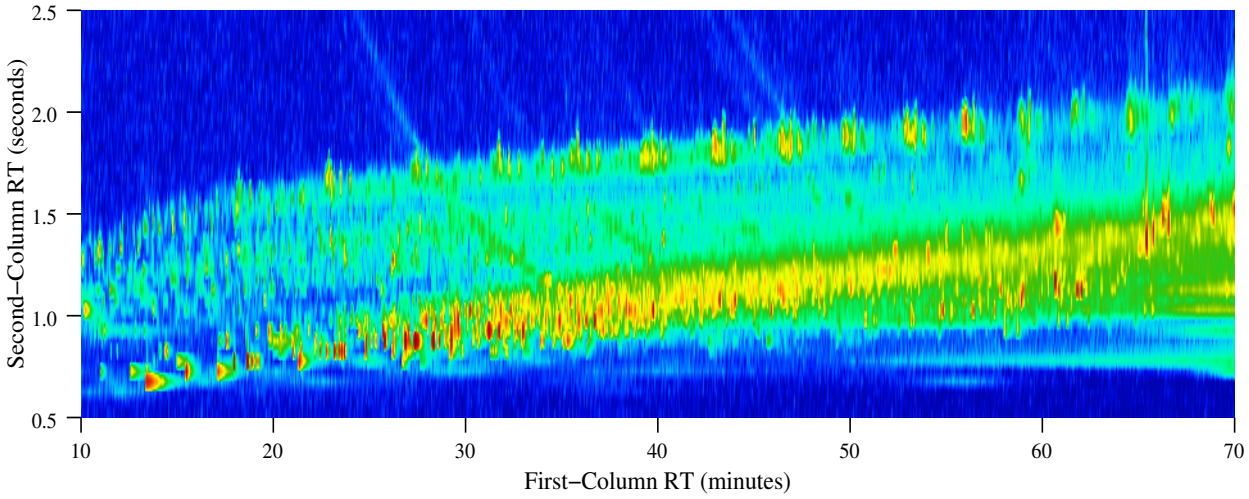

Figure 1: A pseudocolorized image of the total intensity count (TIC) for a GCxGC-MS chromatogram from a complex bio-oil. Each compound induces a two-dimensional peak in the data array output by the detector. Only a subregion is shown.

matogram, there are nineteen overlaid semi-transparent bubbles (some cyan and some red) indicating the locations and intensities (by bubble area) of detected peaks. In the bottom sub-image, there are only thirteen detected peaks. Some of the peaks in the top sub-image can be matched to peaks in the bottom sub-image and vice versa. For example, the twelve prominent peaks with cyan-colored bubbles in each image can be matched to the twelve peaks with cyan-colored bubbles in the other image (even if the retention times and mass spectra of the peaks do not provide definitive compound identifications). However, matching of the other peaks with red-colored bubbles is not definitive, because of differences in the numbers of detected peaks, their retention times, and/or their mass spectra. Such differences may be due to chromatographic variations, instrument noise, and/or compositional differences between samples.

Various researchers have proposed alternative methods for pairwise peak matching.[5] CCM uses pairwise peak matches to determine reliable peaks, but does not itself perform pairwise matching, so any method for pairwise peak matching could be used. Here, the template matching method[6–8] is used to match the pattern of peaks observed in one chromatogram (referred to as the template) to the peaks detected in another chromatogram (referred to as the target). Template matching uses both the retention-times pattern and spectral matching criteria. Template matching returns zero or one matching target peak for each template peak and each matched target peak is the best match for the matching template peak, subject to user-specified constraints and consistency with other peak matches. Alternative peak matching algorithms and/or different parameters might potentially improve pairwise matching performance, but unmatched and mismatched peaks are inevitable for large sets of complex chromatograms.

This paper considers the problem of finding reliable peaks that match not just for pairs of chromatograms, but across large sample sets of up to hundreds of chromatographic patterns. Figure 3 shows a graph that illustrates pairwise matchings between peaks in three chromatograms. In the graph, each peak is represented by a *vertex* labeled

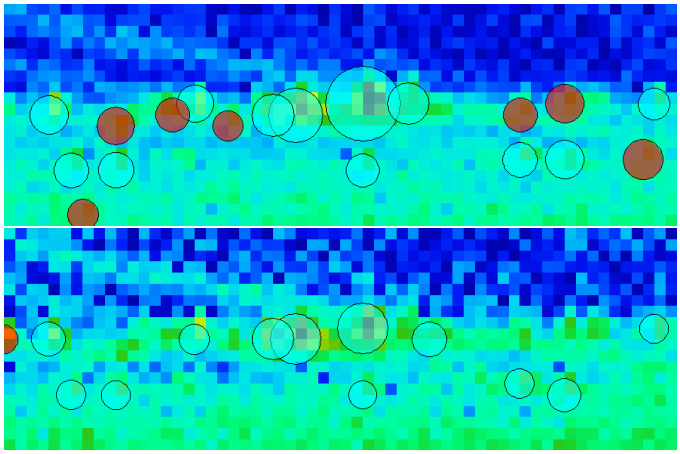

Figure 2: Small subregions of two chromatograms from different bio-oil samples. The data points in this figure are shown as rectangles to clearly illustrate the granularity of the modulation and detector sampling. The peaks marked by cyan-colored bubbles are definitively matched in each direction, but the peaks marked by red-colored bubbles are not definitively matched.

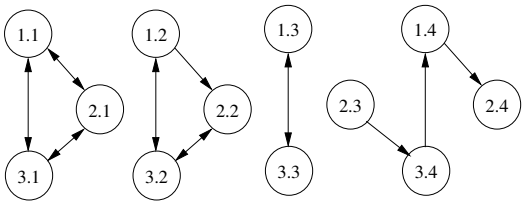

Figure 3: Example pairwise matchings, shown by arrows, between four peak features in each of three chromatographic patterns — Peaks 1.1 to 1.4 in Pattern 1, Peaks 2.1 to 2.4 in Pattern 2, and Peaks 3.1 to 3.4 in Pattern 3.

as <*chromatogramID*>.<*peakID*> and each pairwise template-to-target peak matching is indicated by a directed *edge*. The matchings for Peaks 1.1, 2.1, and 3.1 are reliable over every pair of chromatograms. For Peaks 1.2, 2.2, and 3.2, there are pairwise matchings between Chromatograms 1 and 3 and between Chromatograms 2 and 3, but there is only partial matching between Chromatograms 1 and 2 because Peak 2.2 failed to match Peak 1.2. The matchings for Peaks 1.3 and 3.3 are incomplete because no peak in Chromatogram 2 is matched. The matchings for 2.3, 1.4, 2.4, and 3.4 are conflicting.

Several possible approaches have been suggested to find peaks that match across multiple chromatograms. One approach is to determine reliable peaks by hand.[9,10] Such an approach can achieve better success than automated methods (by having fewer errors), but is extremely tedious, potentially requiring days of manual labor.[10] Another approach is to designate a reference chromatogram and match all peaks in other chromatograms to it.[11–15] However, in many applications there is no true reference chromatogram and this approach could yield different results depending on the arbitrary selection of a reference chromatogram. Even if there is a natural choice for the reference chromatogram, that chromatogram may not exhibit peaks that could be reliably matched across many other chromatograms to provide important chemical information. Another approach is to proceed sequentially through the chromatograms, progressively

modifying the set of peaks,[16,17] but such methods can yield different results depending on the ordering and some sets of chromatograms have no natural ordering.

The approach taken in this work, as described in the next section, is automated and does not bias the result by the selection of a reference chromatogram nor by the ordering of the chromatograms. This new approach automatically considers all pairwise matches in an unbiased fashion to find peaks that can be matched reliably across the set of chromatograms.

Once a set of reliably matched peaks is identified, they can be used for direct comparison or for other tasks such as alignment. For example, in comparing chromatograms, the signal intensity within a peak is indicative of the amount of the compound and so is an important quantitative measurement for comparison. Unfortunately, comprehensive peak-based comparisons across large sets of complex chromatograms are intractable because peak matches often are unreliable.[5] However, reliable peaks can be used to align chromatograms for comprehensive region-based comparisons. In experiments presented here, reliable peaks in fifteen chromatograms from bio-oil samples, with three samples produced by each of five different catalysts, were used both to compare chemical compositions and to align the chromatograms for more comprehensive region-based comparative analyses.

Methods and Theory

Initial Algorithm and Concepts

Our Initial Algorithm for finding reliable peaks, shown in Figure 4, selects peaks that have pairwise matches across every chromatogram by finding cliques that are as large as the number of chromatograms. A *clique* is a subset of the vertices of a graph such that every vertex in the subset is connected to every other vertex in the subset; that is, every pair of peaks in a clique are pairwise matched with one another. A clique with its edges is a *complete subgraph* of the matching graph built in Step 3, so each clique reported in Step 4 contains a set of peaks that are pairwise matched across all chromatograms. Peaks that don't match pairwise across all chromatograms are not reported as reliable features. For example, in the graph of Figure 3, the peaks {1.1, 2.1, 3.1} form a clique across all three chromatographic patterns, so that peak feature is reported as reliable; but the peaks {1.2, 2.2, 3.2} do not form a clique across all three chromatographic patterns (because Peak 2.2 does not match Peak 1.2), so that peak feature is not reported as reliable. The Supporting Information describes some of the properties of these matching graphs in more detail.

Unfortunately, this Initial Algorithm has two problems. First, this approach does not find peak features that are in most chromatograms but which are undetected in at least one chromatogram. If the goal is alignment, then this issue may not be a problem as long as enough reliable peaks are identified; but, if the goal is comparison, then some pertinent

Initial Algorithm:

1. Detect the peaks in each chromatogram.
2. Perform all pairwise peak matchings between chromatograms.
3. Build a graph, as in Figure 3, in which each peak from Step 1 is a vertex and each matching of a template peak in one chromatogram to a target peak in another chromatogram from Step 2 is a directed edge.
4. Find and report cliques in the graph from Step 3 that contain a peak from each chromatogram.

Figure 4: The Initial Algorithm.

peak features may not be selected for comparison. The second issue is more serious: this approach does not scale well for large sample sets. If the peak matching is relatively reliable but imperfect (as most real-world phenomena are), then as the number of chromatograms increases, the likelihood of a failed match or inconsistency — even for highly reliable features — grows exponentially.

To illustrate this second problem, note that the number of possible features with a matched peak in each chromatogram is limited by the number of peaks in the chromatogram with the fewest detected peaks. On the other hand, the number of pairwise matches required for a feature with a peak from each chromatogram is equal to $n(n-1)$, where n is the number of chromatograms, because a clique must have n matching peaks and each peak must match to a peak in each of the other $(n-1)$ chromatograms. So, while the number of possible features is fixed or diminishes as new chromatograms are acquired, the requirements for complete consistency for any feature increases exponentially with larger numbers of chromatograms.

For example, if pairwise peak matches are 99.9% reliable, then for a set of ten chromatograms, more than 91% of such peaks are expected to be matched across all chromatograms ($0.999^{10(10-1)} = 0.913890$); but for a set of 100 chromatograms, fewer than one in 20,000 of such peaks are expected to be matched across all chromatograms ($0.999^{100(100-1)} = 0.000499$). In this example of 100 chromatograms, if the chromatogram with the fewest peaks contains only a few thousand peaks, then it is likely that no reliable peaks would be found. For peaks with lower pairwise-matching reliability, the problem is apparent for even smaller sets of chromatograms. As described in the next subsection, the solution for these issues is to allow the user to relax the requirement for complete pairwise matching across all chromatograms.

Consistent Cliques Method

The CCM, developed in this paper, is shown in Figure 5. CCM is the same as the Initial Algorithm through Step 3, but CCM then selects peaks that are consistently matched over some, but perhaps not all, chromatograms. Step 2 finds cliques that contain peaks from at least s chromatograms, but which do not necessarily contain peaks from

CCM Algorithm:

1. Build a graph from pairwise peak mathings by performing Steps 1–3 of the Initial Algorithm.
2. Find maximal cliques in the graph from Step 1 that contain peaks in at least s chromatograms, where s is a parameter of the algorithm.
3. Combine the maximal cliques from Step 2 that share common peaks.
4. Report the combined cliques from Step 3 in order from largest to smallest until the number of reliable peak features requested by the user are reported or all combined cliques are reported.

Figure 5: The CCM algorithm.

all chromatograms. The user can reduce the minimum size of the maximal cliques, s in Step 2, thereby yielding additional but less reliable peaks with pairwise matches in fewer chromatograms. The percentage of chromatograms that have peaks in the clique can be regarded as a *measure of the reliability* of a peak feature with respect to a set of chromatograms. For example, if a peak is matched consistently across twelve of fifteen chromatograms, it can be said to be 80% reliable.

Cliques that are smaller than the number of chromatograms may share common peaks, *i.e.*, peaks that should be regarded as one common feature could form more than one maximal clique. So, Step 3 combines those cliques sharing common peaks. The resulting combined cliques may be of different sizes, so, in Step 4, if the user asks for a number of reliable features that is less than the number of combined cliques, then only the most reliable features with peaks in the largest number of chromatograms are reported, up to the number of requested features.

For example, given the graph in Figure 3, CCM Step 2 detects the four maximal cliques of size $s = 2$ or larger: {1.1, 2.1, 3.1}, {1.2, 3.2}, {2.2, 3.2}, and {1.3, 3.3}. In Step 3, cliques {1.2, 3.2} and {2.2, 3.2} are combined to form {1.2, 2.2, 3.2}, because they have Peak 3.2 in common. In this way, CCM finds the peak feature that was missed by the Initial Algorithm. If the user asks for two features, then only {1.1, 2.1, 3.1} and {1.2, 2.2, 3.2} are reported; but if the user asks for more than two features, then {1.3, 3.3} also is reported.

Setting $s > \lceil 2n/3 \rceil$, where n is the number of chromatograms, ensures that sets of feature cliques that share common peaks are conflict-free; that is, the union of such cliques has no more than one peak in each chromatogram. The proof of this is provided in the Supporting Information. If there are no conflicts, then feature cliques that share common peaks in Step 3 can be combined by a simple union. The minimum size of the maximal feature cliques can be fixed to the smallest value that ensures conflict-free results:

$$s_n = \lceil (2n+1)/3 \rceil. \quad (1)$$

The user still is provided parametric control of the number of desired features, in Step 4, to constrain the relative

reliability of the reported features.

With $s = s_n$ in Step 2 and the union of cliques in Step 3, CCM selects peaks that are consistently matched across more than two-thirds of the chromatograms. The threshold s can be set to smaller values, but then it might be necessary to deal with conflicts between cliques that share common peaks in Step 3. Three possible alternative methods for dealing with such conflicts in the combining of cliques that share common peaks are: (a) eliminate from the combination those peaks in the chromatograms for which there is a conflict, (b) eliminate from the combination all cliques for which there is a conflict, and (c) do not report a combination of cliques for which there is a conflict.

Experimental Procedures

Experimental Samples

The experiments reported here analyzed upgraded pyrolysis oils from the presscake of pennycress seeds. Fast pyrolysis oils from biomass materials with high protein content generally are more stable and partially deoxygenated compared with those from mostly lignocellulosic biomass (*e.g.*, wood, grasses) due to nucleophilic substitution of nitrogen for oxygen. However, in order for these products to be used as transportation fuels or petroleum refinery feedstocks, the pyrolysis oils still must be upgraded to reduce their heteroatom (N, O, S) content. Because the compositions of these proteinaceous pyrolysis oils differ greatly from those from lignocellulosic feedstocks, their behavior in various upgrading steps will be different. Therefore, research on Sustainable Biofuels and Coproducts at the Eastern Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, is studying hydrogenation of fast pyrolysis oils from the presscake of pennycress seeds as a model for hydrotreating proteinaceous fast pyrolysis oils.[18]

Pennycress makes a good bioenergy crop because it is a winter crop and therefore can be an additional crop that does not replace a food crop. The presscake is the material remaining after mechanical pressing to remove most of the vegetable oil (which is used for biodiesel or green diesel production). The presscake is pyrolyzed to produce the bio-oil.

The bio-oil samples for analysis were produced from a batch hydrogenation over five alternative precious metal catalysts using a Parr reactor, with standard conditions of 2000 psi H_2 at 300° C to hydrogenate the bio-oil. The main goal of the hydrogenation is deoxygenation and denitrogenation to remove O and N (as water and ammonia) and produce hydrocarbons. These hydrogenation conditions are not extreme enough to completely accomplish this, but partial heteroatom removal can improve the material for blending with petroleum for processing in a refinery or for use as a feedstock for renewable chemicals. Current research involves higher pressure hydrogenations.

The goals for chemical analysis include characterizing the chemical transformations that occur, comparing these products with those from hydrogenation of lignocellulosic pyrolysis oils, and comparing the selectivity of the catalysts

for some of the individual reactions in the complex system. To this end, three experimental replicates (*i.e.*, from different samples under the same conditions) for each of the five catalysts were analyzed by GCxGC-MS, as described in the next subsection.

Chromatography and Mass Spectrometry

The fifteen samples described in the previous subsection were analyzed by GCxGC-MS with a Shimadzu (Kyoto, JP) GCMS-QP2010S GC-MS system and a Zoex (Houston TX, USA) ZX-2 LN2 cooled-loop GCxGC thermal modulation system. Samples were 15 wt% in methanol. The flow rate was 3 mL/min He, with a 30:1 split ratio, 250° C injector temperature, and 40 psi initial pressure. The first-dimension separation was performed on a Restek (Bellefonte PA, USA) Rtx-1701 column (14% cyanopropylphenyl and 86% dimethyl polysiloxane, 60 m length, 0.25 mm internal diameter, and 0.25 μm film thickness) and the second-dimension separation on a Restek Rtx-1 column (100% polyimethylpolysiloxane, 2 m length, 0.25 mm internal diameter, and 0.25 μm film thickness). The temperature program was 45° C for 4 min, increased by 3° C/min to 280° C, then held for 20 min. The modulation cycle was 4 seconds. The total run time was 102 min and data was acquired from 9.5 to 70 min. The mass spectrometer used 70 e-V electron impact (EI) ionization and acquired data over the mass-to-charge (m/z) range 35–400, at a sampling rate of 20 spectra per second.

Data Processing

It is important to place the determination of reliable peaks and their use in the larger context of data processing. The fifteen chromatograms, pictured in the Supporting Information, were processed by GC Image (Lincoln NE, USA) GCxGC Edition Software, R2.3a0. Processing was fully automated with the following operations.

1. **Chromatogram processing.** In each chromatogram, the data was shifted as necessary to align the first datapoint relative to the modulation start-time.[8] Then, the baseline was corrected so that the peaks rise above a near-zero-mean baseline.[19] Then, the blob-peaks were detected using the Drain Algorithm[7] which performs true two-dimensional peak detection.[20]

2. **Template construction.** From each chromatogram, a template[6,7] was constructed to record the retention times and normalized mass spectrum of each detected peak. Each template peak also was given a mass-spectral matching constraint written in CLIC,[21] which generally required that the NIST match factor[22] for matched peaks be at least 700 (although lower match factors were allowed for peaks if no nearby peak exhibited a similar spectrum).

3. **Template matching.** The template from each chromatogram generated in Processing Step 2 was matched[8] with

each of the other fourteen processed chromatograms.

4. **Reliable peaks selection.** The pairwise matches from Processing Step 3 were used to determine which peaks are matched reliably across the set of fifteen chromatograms. The results of this operation are discussed in greater detail in the following section. A reliable-peak feature was created from each clique by averaging the retention times and mass spectra of the peaks in the clique. The reliable-peak features then were collected in a *reliable-peaks template*.

5. **Composite chromatogram construction.** The reliable-peaks template determined in Processing Step 4 was matched to each of the fifteen chromatograms. Then, each chromatogram was geometrically transformed by the inverse of the affine transformation matching the template to the chromatogram. In this way, each chromatogram was aligned with the template. Then, the aligned chromatograms were added together to create a composite chromatogram.[23]

6. **Feature template construction.** The feature template was constructed by adding a peak-region object for each peak detected in the composite chromatogram (from Processing Step 5) to the reliable-peaks template (from Processing Step 4). Each peak-region object was delineated by the retention-times outline of a detected peak and therefore is called a peak-region feature.[5] Each peak-region feature is described by the name of the compound from the mass spectral library that has the best NIST match factor with the mass spectrum of the peak (pending more definitive identifications for features of interest).

7. **Chromatogram analysis.** Each chromatogram was analyzed by matching the reliable peaks of the feature template to the detected peaks of the chromatogram, aligning the template and its peak-region objects according to the geometric transformation of the matching, and then characterizing the detector response for each matched peak and each peak-region object.[5,23]

8. **Comparative feature analysis.** Once the detector response in each chromatogram has been characterized for each peak feature and each peak-region feature, various multivariate statistical operations can be used to compare samples, identify potential marker compounds, build sample classifiers, discover trends, and/or perform other multi-sample analyses. Some results of the comparative analysis of the bio-oil samples are presented in the next section and in the Supporting Information.

The computation time for the CCM algorithm to select the reliable peaks in Processing Step 4 is relatively small. For the fifteen samples presented here, executing the CCM program required less than 30 seconds on a desktop personal computer (specifically, a Parallels Desktop 6® virtual machine running Microsoft Windows 7® on a Mac Pro® with two 2.8GHz Quad Core Intel Xeon® CPUs and 6GB 667 MHz RAM) operating on data stored on a networked file

Table 1: Percentage Peak-Matching Rates by Catalyst.

Target	Template Catalyst					
Catalyst	1	2	3	4	5	Average
1	43.9	43.0	39.2	43.1	42.2	42.3
2	38.9	41.7	34.0	37.2	43.0	38.9
3	41.5	40.0	43.8	45.6	41.0	42.4
4	44.3	42.2	44.2	42.8	44.2	43.5
5	41.1	46.2	37.6	41.9	45.9	42.6
Average	41.9	42.6	39.8	42.1	43.2	41.9

server. The total time required for the entire processing sequence was nearly 4.5 hours. Processing Step 3 was the most time consuming, requiring more than 2.5 hours to perform the 210 pairwise matchings between chromatograms, and Processing Step 5 required nearly 1.5 hours to match the reliable peaks template to every chromatogram and build the composite chromatogram. Much of the time was required for reading and writing the large data files across the network, with each data-file read requiring about 30 seconds and each data-file write requiring one to two minutes. A system with local RAID and/or SSD storage would operate more quickly. Undoubtedly, the speed of the software could be improved, but even so the total time for data processing is only a fraction of the time required for the chromatography.

Results and Discussion

Reliable Peaks

The average number of peaks detected in each bio-oil chromatogram was 567, with a range of 436 to 699 over the fifteen chromatograms. The average number of peaks in the chromatograms for each of the catalysts did not vary greatly, with 571, 515, 608, 587, and 552 peaks respectively for Catalysts 1 to 5.

Table 1 shows the average peak-matching rates between chromatograms for the five catalysts. The overall average peak-matching rate between pairs of chromatograms was 41.9%, which is a fairly low rate, reflecting compositional differences; the large dynamic range of peak intensities, including many faint peaks; and peak crowding. The average peak-matching rate between chromatograms for the same catalyst (six pairwise matches each) was higher than the overall average, at 43.6%. By comparison, the average peak-matching between chromatograms for different catalysts (nine pairwise matches each) was 41.5%. Even if matching could be improved by better tuning of the parameters or by using another peak-matching method, these numbers suggest that many peak features cannot be matched reliably across large sets of such complex chromatograms. It is for this reason that peak-region features, rather than peaks, are a more robust foundation for comprehensive comparative analysis.

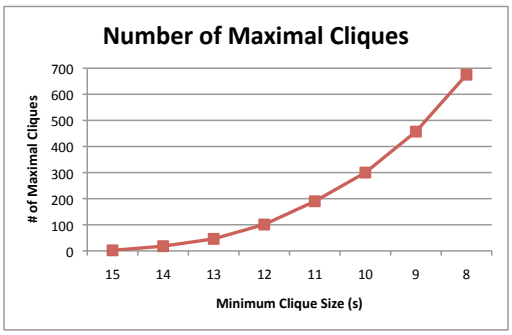

Figure 6: The graph shows the number of maximal cliques of peaks in the bio-oil chromatograms as a function of the mimimum clique size.

The Initial Algorithm for selecting reliable peaks found only two maximal cliques of size fifteen, with a peak in each chromatogram. Two peak features is not sufficient to effectively characterize or compare the chromatograms nor even to determine a fully parameterized affine transformation for alignment. This example illustrates the need for a more flexible method for selecting reliable peaks and the motivation for CCM.

Figure 6 illustrates the number of maximal cliques as a function of the minimum clique size s. As just noted, there are only two cliques with a peak in every one of the fifteen chromatograms. However, as the minimum size is decreased, the number of maximal cliques that are sufficiently large increases. With the threshold for the clique size set as $s_{15} = 11$, which is the smallest threshold that guarantees that cliques that share a common peak are conflict-free, there are 190 maximal cliques. With the threshold for the clique size set as $s = 8$, which provides cliques that have complete pairwise matchings for more than half the chromatograms, there are 675 maximal cliques. It is possible to have more cliques than characteristic peaks for a set of chromatograms because cliques may share common peaks.

After combining maximal cliques of size $s_{15} = 11$ or larger that share peaks, there are 80 combined cliques. The user can take a subset of these combined cliques to get only the more reliable features. For example, as shown in Figure 7, 29 of the combined cliques have a peak in each of the fifteen chromatograms, 45 of the combined cliques are size fourteen or larger, *etc.* After combining maximal cliques of size $s = 8$ or larger that share common peaks, there are 201 combined cliques. In this case, none of those combined cliques exhibit any conflict. As shown in Figure 7, combining maximal cliques of size $s = 8$ or larger yields 65 combined cliques with a peak in each of the fifteen patterns, 96 are size fourteen or larger, *etc.* When all of the combined cliques for $s = 8$ are allowed, there are 201 peak features, but that still is only 35% of the average number of peaks in each chromatogram. The problem that so many peaks in complex chromatograms are difficult to match appears to make comprehensive peak matching intractable and is the motivation for using peak features for alignment and then using peak-region features for comprehensive comparative analysis,[5] as demonstrated in the following subsection.

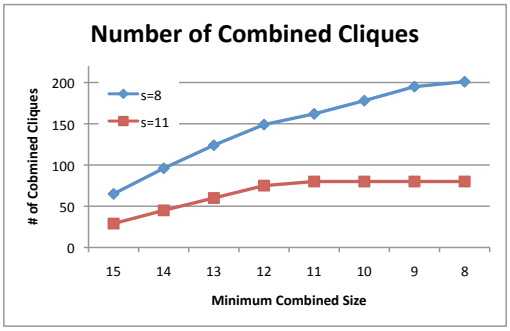

Figure 7: The graph shows the number of combined cliques for both $s = 8$ and $s_{15} = 11$ as a function the minimum combined size.

Comparative Analysis

The goal of this paper is to demonstrate a new method for selecting peaks that are reliably matched across many chromatograms. Therefore, the purpose of the comparative analysis presented here is to demonstrate the utility of the selected reliable peaks for comparative analyses. Extracting and applying all of the information available in the analysis in order to develop knowledge about the performance of the different catalysts in the pyrolysis of the bio-oils is beyond the scope of this paper and will be addressed in other publications.[24]

Reliable peaks can be used to compare chromatograms directly with respect to those features or to align the chromatograms for more comprehensive analysis with peak-region features. Figure 8 shows the composite chromatogram (from Processing Step 5) for the fifteen aligned bio-oil chromatograms. The overlay in Figure 8 shows the feature template (from Processing Step 6) with reliable peaks shown with white circles and peak regions for the peaks detected in the composite chromatogram shown with red dotted outlines. There are 660 peak-region features, which is within the range of the number of peaks detected in the individual chromatograms. Note that this is more than three times the number of reliable peaks and so provides a much more comprehensive basis for characterizing and comparing samples.

Figure 9 shows the mean percent-response across all chromatograms for each of the reliable peaks in the feature template. The percent-response of a peak is computed as its summed TIC divided by the total of the summed TICs of all peaks in the chromatogram (not just the reliable peaks). For each reliable peak, there is a circle in Figure 9 positioned according to the average retention times for the peak over all chromatograms in which it is found. The area for each circle is determined by the average percent-response over all chromatograms (including zero for chromatograms in which the peak is not matched). Such a peak *fingerprint* can be computed for each catalyst, and, as shown in the Supporting Information, the peak fingerprints of different catalysts can be compared.

Unfortunately, a fundamental concern in these comparisons is that the peak fingerprints are not comprehensive — that is, there are many peaks that are not reliably matched across chromatograms and therefore are not compared. In

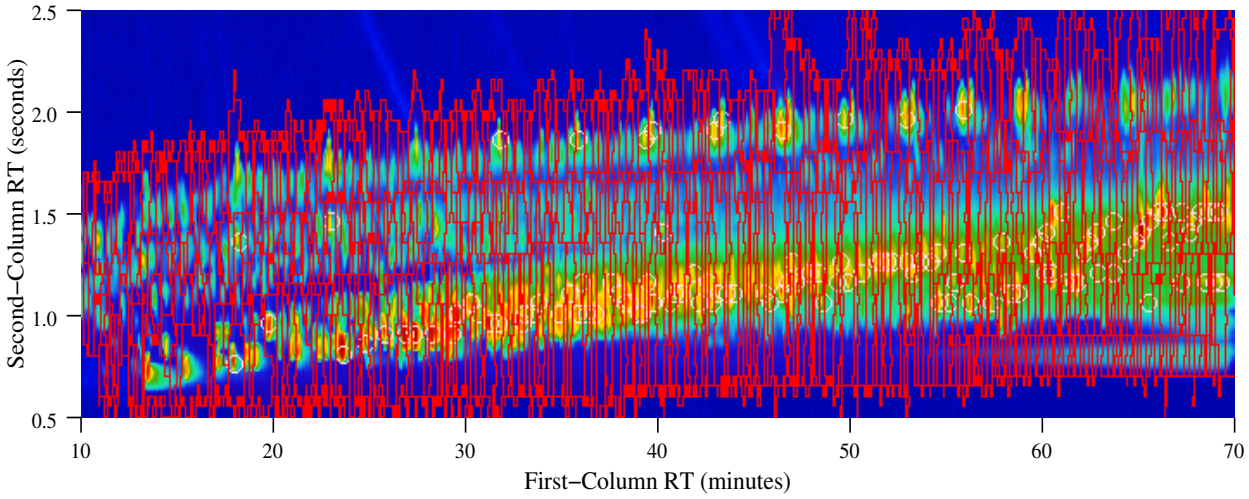

Figure 8: The pseudocolorized image of the total intensity count (TIC) for the composite chromatogram from fifteen aligned chromatograms. The overlay shows the reliable peaks (with white circles) and the peak-region features (with red dotted outlines).

this example, there are few reliable peaks in the band of alkylbenzenes between the lower horizontal band of amides and phenols and the upper horizontal band of aliphatic hydrocarbons. A related concern is that peak matching errors can mask true differences or appear as false differences.

Peak-region features offer a more robust basis for comprehensive comparisons. Figure 10 shows the mean percent-responses using the peak-region features derived from the composite chromatogram shown in Figure 8. The comparison of this figure to Figure 9 clearly shows that the peak-region features provide far more comprehensive fingerprints. For example, the characterizations of the aliphatic hydrocarbons and alkylbenzenes (at the top of the chromatogram) are more complete.

The Supporting Information presents comparative analyses of peak-region fingerprints, evaluations of statistical significance using Fisher's Ratio, comparative analyses using statistically significant chemical markers, and the results of cross-validation experiments supporting the significance of the chemical markers.

Conclusion

The CCM is an algorithm for selecting features that are reliably matched across many patterns. CCM can be used with applications that involve complex data with unlabeled features, such as comprehensive two-dimensional chromatography with unlabeled peaks. Comparative multi-sample analyses with comprehensive two-dimensional chromatography require peaks that are reliably matched (*i.e.*, deemed to result from the same compound) across many samples. CCM overcomes problems with previous approaches: CCM is fully automated, it is not dependent on the selection of a reference pattern, and the result does not depend on the ordering of the patterns.

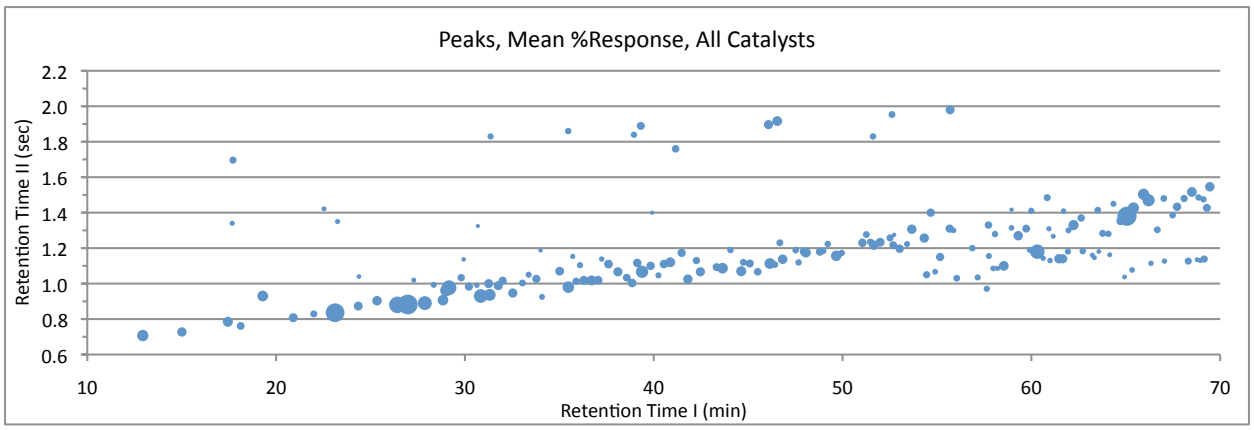

Figure 9: Each circle indicates the retention times and mean percent-response for one of the reliable peaks. The (x,y) position of each circle indicates the average first and second dimension retention times and the area of each circle indicates the mean TIC percent-response across all chromatograms (on a scale relative to the largest percent-response). The percent-response is the ratio of the response for the peak to the total responses for all peaks.

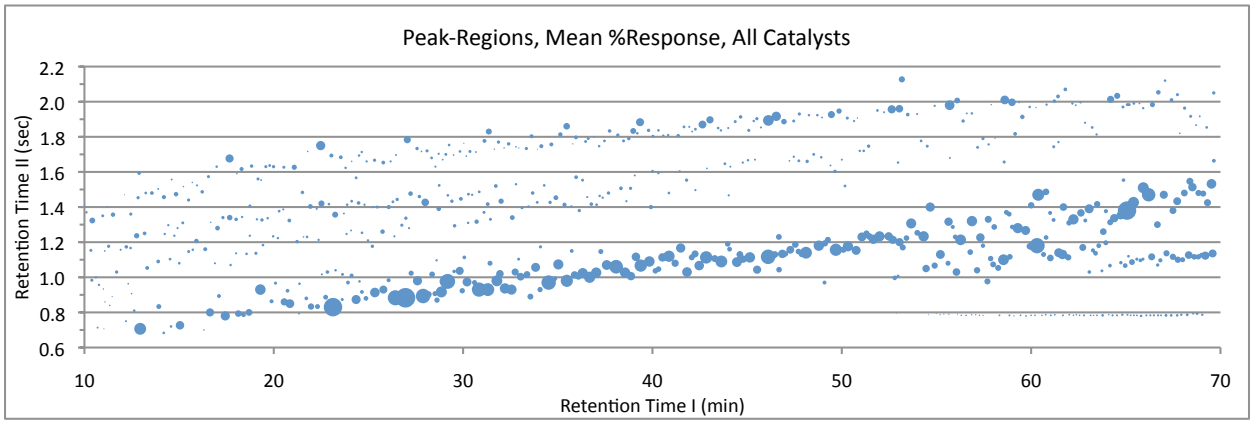

Figure 10: Each circle indicates the retention times and mean percent-response for one of the peak-region features. The percent-response is the ratio of the response for the peak-region to the total responses for all peak-regions.

Here, CCM was demonstrated with fifteen chromatograms of complex bio-oil samples with nearly 600 peaks detected, on average, for each sample. Only two peaks matched consistently across every one of the 210 possible pairwise matchings, but CCM identified more than 200 peaks that were matched across more than half of the chromatograms. The reliable peaks were used to directly compare the characteristics of the samples and to align the chromatograms for truly comprehensive comparisons with peak-region features.

Future work on CCM might relax the constraint that, in each pairwise matching, each feature is matched only with its best match. However, allowing more than one potential match in pairwise matchings could significantly increase computational complexity. Taken in its general form, the problem of whether a graph contains a clique larger than a given size is NP-complete, meaning that even moderately sized problems can be intractable. Moreover, such an approach could produce conflicting feature sets. In its current form, CCM can be performed rapidly and if the minimum size of the clique is set to more than 2/3 of the number of samples, then the feature sets will be conflict-free.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation underAward Number IIP-1013180 and by the Nebraska Center for Energy Sciences Research.

Supporting Information Available

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org

References

(1) Horaud, R.; Skordas, T. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 1168–1180.

(2) Bushey, M. M.; Jorgenson, J. W. *Anal. Chem.* **1990**, *62*, 161–167.

(3) Liu, Z. Y.; Phillips, J. B. *J. Chromatogr. Sci.* **1991**, *29*, 227–331.

(4) Visvanathan, A.; Reichenbach, S. E.; Tao, Q. *J. Electron. Imaging* **2007**, *16*, 033004.

(5) Reichenbach, S. E.; Tian, X.; Cordero, C.; Tao, Q. *J. Chromatogr. A* **2012**, *1226*, 140–148.

(6) Reichenbach, S. E.; Carr, P. W.; Stoll, D. R.; Tao, Q. *J. Chromatogr. A* **2009**, *1216*, 3458–3466.

(7) Reichenbach, S. E.; Ni, M.; Kottapalli, V.; Visvanathan, A. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 107–120.

(8) Reichenbach, S. E. In *Comprehensive Two Dimensional Gas Chromatography*; Ramos, L., Ed.; Elsevier: Oxford UK, 2009; Chapter 4, pp 77–106.

(9) van Mispelaar, V. G. Chromametrics. Ph.D. thesis, University of Amsterdam, 2005.

(10) Koek, M. M.; van der Kloet, F. M.; Kleemann, R.; Kooistra, T.; Verheij, E. R.; Hankemeier, T. *Metabolomics* **2011**, *7*, 1–14.

(11) Shellie, R. A.; Welthagen, W.; Zrostliková, J.; Spranger, J.; Ristow, M.; Fiehn, O.; Zimmermann, R. *J. Chromatogr. A* **2005**, *1086*, 83–90.

(12) Wardlaw, G. D.; Arey, J. S.; Reddy, C. M.; Nelson, R. K.; Ventura, G. T.; Valentine, D. L. *Environ. Sci. Technol.* **2008**, *42*, 7166–7173.

(13) Oh, C.; Huang, X.; Regnier, F. E.; Buck, C.; Zhang, X. *J. Chromatogr. A* **2008**, *1179*, 205–215.

(14) Gaquerel, E.; Weinhold, A.; Baldwin, I. T. *Plant Physiol.* **2009**, *149*, 1408–1423.

(15) Li, X.; Xu, Z.; Lu, X.; Yang, X.; Yin, P.; Kong, H.; Yu, Y.; Xu, G. *Anal. Chim. Acta* **2009**, *633*, 257–262.

(16) Cordero, C.; Liberto, E.; Bicchi, C.; Rubiolo, P.; Schieberle, P.; Reichenbach, S. E.; Tao, Q. *J. Chromatogr. A* **2010**, *1217*, 5848–5858.

(17) Castillo, S.; Mattila, I.; Miettinen, J.; Orešiǩ, M.; Hyötyläinen, T. *Anal. Chem.* **2011**, *83*, 3058–3067.

(18) Mullen, C. A.; Boateng, A. A. *BioEnergy Res.* **2011**, *4*, 303–311.

(19) Reichenbach, S. E.; Ni, M.; Zhang, D.; Ledford, E. B., Jr. *J. Chromatogr. A* **2003**, *985*, 47–56.

(20) Latha, I.; Reichenbach, S. E.; Tao, Q. *J. Chromatogr. A* **2011**, *1218*, 6792–6798.

(21) Reichenbach, S. E.; Kottapalli, V.; Ni, M.; Visvanathan, A. *J. Chromatogr. A* **2004**, *1071*, 263–269.

(22) Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770–781.

(23) Reichenbach, S. E.; Tian, X.; Tao, Q.; Ledford, E. B., Jr.; Wu, Z.; Fiehn, O. *Talanta* **2011**, *83*, 1279–1288.

(24) Mullen, C. A.; Boateng, A. A.; Reichenbach, S. E. Hydrotreating of Fast Pyrolysis Oils from the Proteinaceous Biomass of Pennycress Seeds. Submitted, 2012.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Supporting Information for

Reliable Features for Comparative Analysis with Comprehensive Two-Dimensional Chromatography

Stephen E. Reichenbach,[*,†] Xue Tian,[†] Akwasi A. Boateng,[‡] Charles A. Mullen,[‡] Chiara Cordero,[*,¶] and Qingping Tao[*,§]

University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA, Sustainable Biofuels and Co-Products Research Unit, USDA-ARS, Eastern Regional Research Center, Wyndmoor PA 19038-8598, USA, Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy, and GC Image, LLC, PO Box 57403, Lincoln NE 68505-7403, USA

E-mail: reich@cse.unl.edu; chiara.cordero@unito.it; qtao@gcimage.com

Matching Graphs and Cliques

The matching graphs for sets of chromatographic peaks have two important properties. First, the set of peaks in each chromatogram is disjoint from the sets of peaks in other chromatograms. Second, there are no matches between peaks in the same chromatogram. From these properties, the graph constructed in Initial Algorithm Step 3 is *multi-partite*, meaning that it can be partitioned into multiple partite sets (or parts) that are *disjoint* (no shared vertices between sets) and each of which is *independent* (no edges between vertices in the same set).

Cliques typically are considered for undirected graphs (in which edges do not have directions), but can be considered for directed graphs (*digraphs*).[1] Here, the matching is bidirectional, so, to ensure consistency, vertices in the cliques must be bidirectionally connected.

In a graph, a *maximal clique* is a clique for which no vertex outside the clique is connected with every vertex in the clique; that is, the clique is not a subset of a larger clique. A clique is a *maximum clique* if the graph has no larger clique. If a clique in a matching graph has a vertex from each pattern, as is required in Initial Algorithm Step 4, it

[*]To whom correspondence should be addressed
[†]University of Nebraska – Lincoln
[‡]USDA-ARS
[¶]Università degli Studi di Torino
[§]GC Image

is necessarily a maximum clique because no vertex in the clique can be connected to any vertex outside the clique (because there are no matches between peaks in the same chromatogram). So, the Initial Algorithm selects peaks that can be bidirectionally matched across every pair of chromatograms.

Ensuring Conflict-Free Cliques

For five patterns, the threshold to ensure conflict-free cliques with shared vertices is $s_5 = \lceil (2 \cdot 5 + 1)/3 \rceil = 4$. Consider two disjoint cliques from five patterns, with each clique having fewer vertices than this threshold: C_1 = {1.1, 2.1, 3.1} and C_2 = {3.2, 4.1, 5.1}. These cliques have a conflict in Pattern 3, but share no common vertex. (If they did share a common vertex, there could not be a conflict because the common peak could not match two different peaks in the same pattern.) Suppose that there is another clique for the same five patterns: C_3 = {1.1, 2.1, 4.1, 5.1}. Clique C_3 shares common vertices with both C_1 and C_2, but the union of the three cliques contains a conflict in Pattern 3, with Peak 3.1 from Clique C_1 and Peak 3.2 from Clique C_2.

If two conflicting cliques each must have at least s_n vertices, then they must have at least $2s_n - n$ conflicting vertices and no more than $2(n - s_n)$ vertices in one clique and not in the other. (Conflicting cliques cannot share common vertices, *i.e.*, are disjoint, because a shared vertex cannot match two different peaks in the same pattern.) Because the number of vertices not in conflict is less than the threshold, *i.e.*, $2(n - s_n) < s_n$, a third clique (or union of other cliques) that shares non-conflicting vertices with each of the two conflicting cliques also must contain a vertex in a pattern for which the first two cliques conflict. However, no vertex in a pattern for which there is a conflict in the first two cliques can match with the vertices of both those two cliques. Therefore, there cannot be conflicts between cliques of size $s > \lceil 2n/3 \rceil$ that share common vertices.

Bio-oil Chromatograms

Relevant regions of the fifteen chromatograms are shown in Figures 1 through 5, with one figure for each of the five catalysts and three chromatograms in each figure for each of the experimental replicates. The retention-times delimiting the regions are the same as those shown in Figures 1 and 8. The quality of the chromatograms is generally good, but there is some crowding in both dimensions, with analytes eluting in about a two-second range in the second-column and first-column peaks having a peak-width standard deviation approximately equal to 4 s or one modulation. Clearly, comprehensive multi-sample comparative analysis of such complex comprehensive two-dimensional gas chromatography (GCxGC) chromatograms is a difficult challenge.

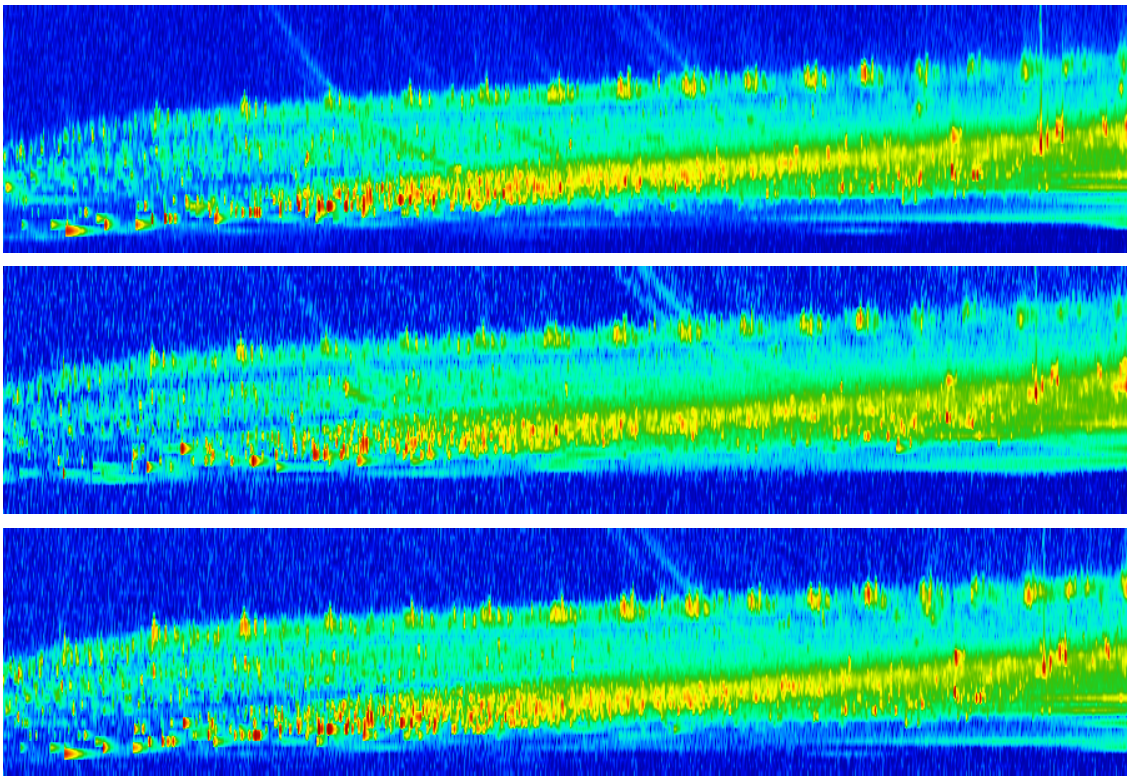

Figure 1: Relevant regions of GCxGC chromatograms from bio-oils produced with Catalyst 1, with three experimental replicates (*i.e.*, different samples).

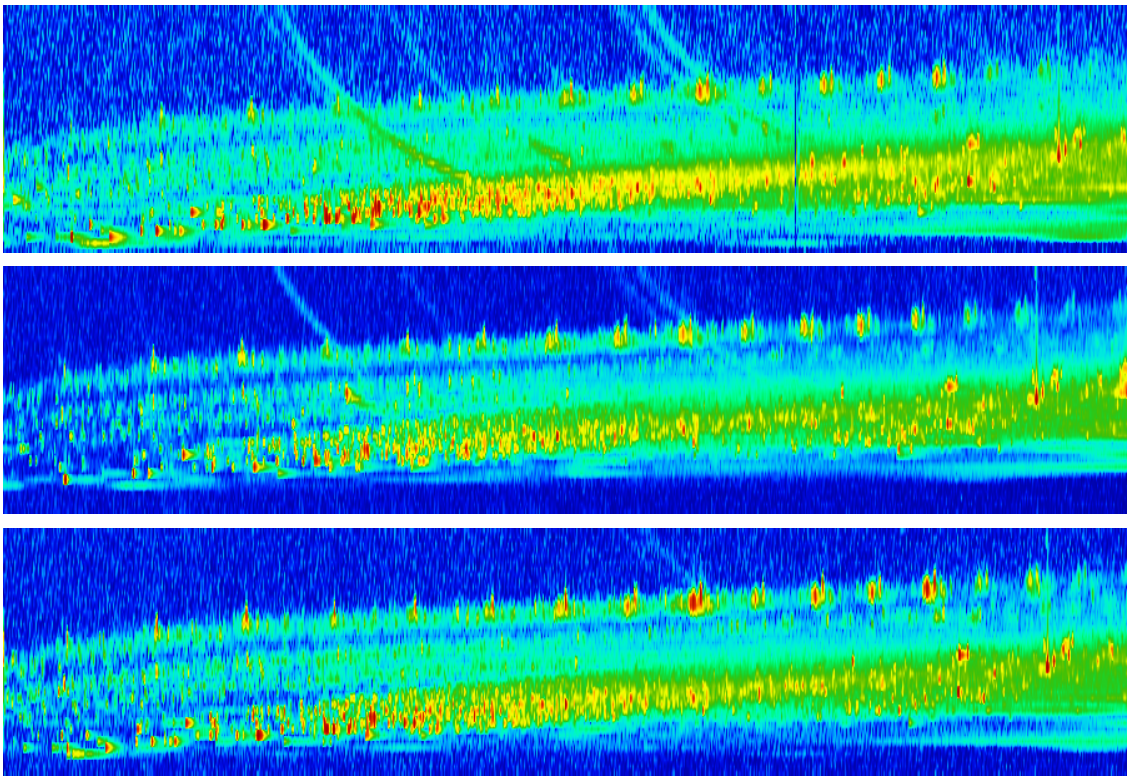

Figure 2: Relevant regions of GCxGC chromatograms from bio-oils produced with Catalyst 2, with three experimental replicates (*i.e.*, different samples).

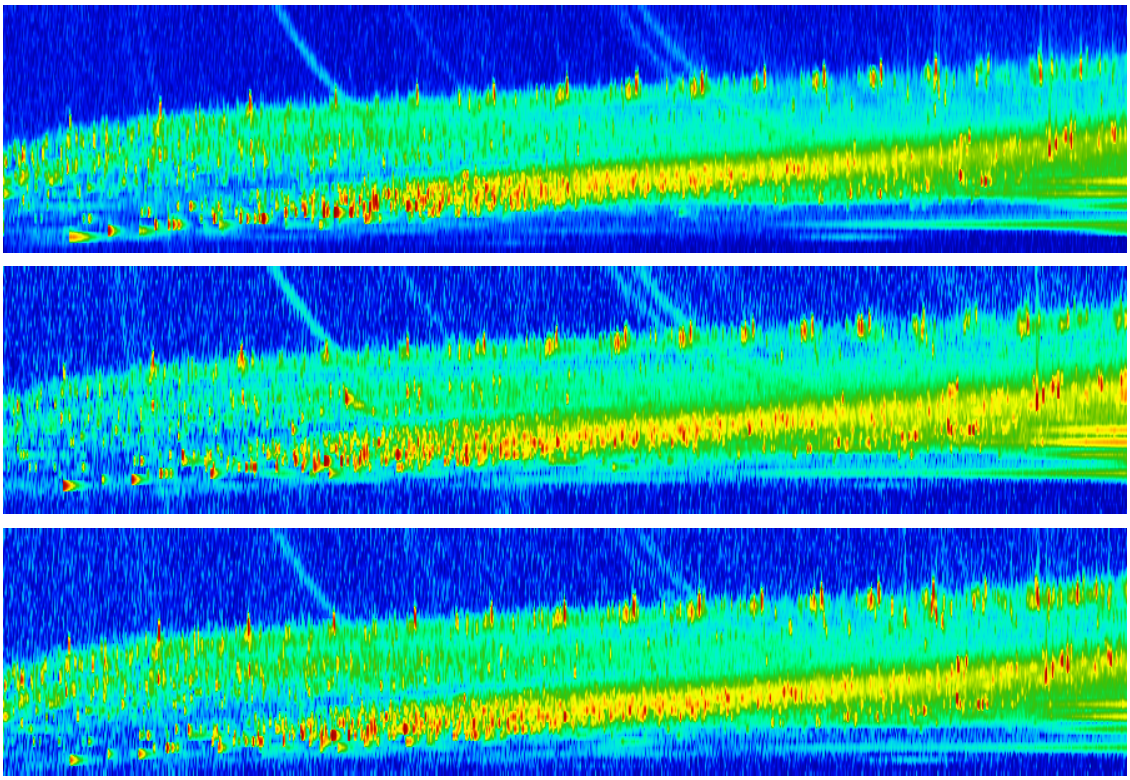

Figure 3: Relevant regions of GCxGC chromatograms from bio-oils produced with Catalyst 3, with three experimental replicates (*i.e.*, different samples).

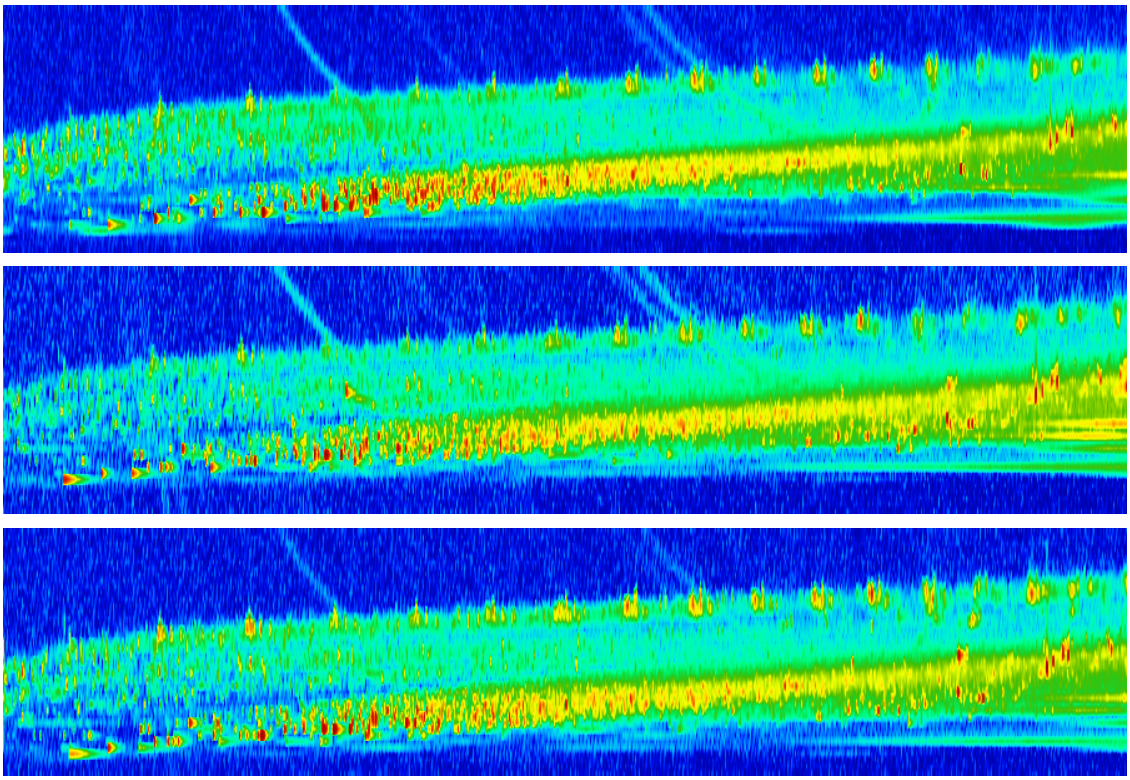

Figure 4: Relevant regions of GCxGC chromatograms from bio-oils produced with Catalyst 4, with three experimental replicates (*i.e.*, different samples).

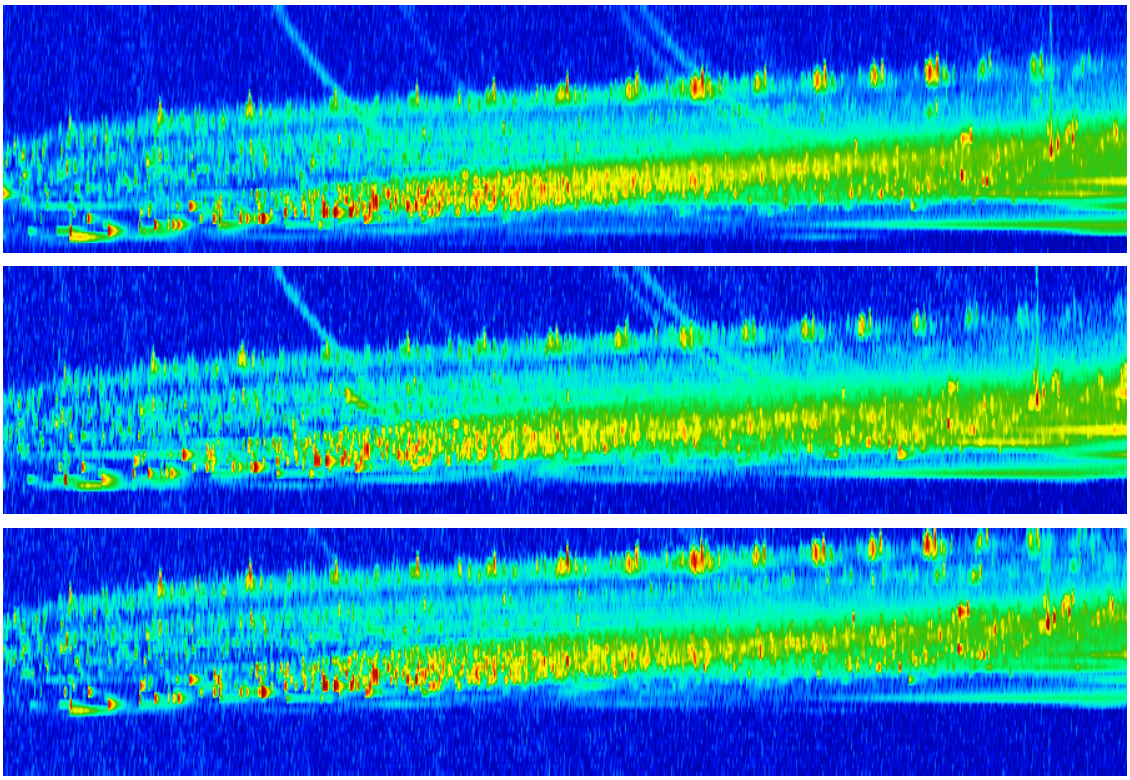

Figure 5: Relevant regions of GCxGC chromatograms from bio-oils produced with Catalyst 5, with three experimental replicates (*i.e.*, different samples).

Feature Analysis

As described in the main text, a peak fingerprint can be created for a set of chromatograms characterizing the mean percent-response for each reliable peak. Then, these peak fingerprints can be compared, e.g., to visualize the compositional differences. For example, Figure 6 shows the difference between the mean percent-responses of peaks in chromatograms with Catalyst 3 and the mean percent-responses of peaks in chromatograms with Catalyst 4. In Figure 6, blue circles indicate that the mean peak percent-response is greater with Catalyst 3 than with Catalyst 4 and red circles indicate that the mean peak percent-response is greater with Catalyst 4 than with Catalyst 3. An experienced chromatographer who is familiar with the chemistry of the samples can use such plots in assessing which catalyst is more effective. In this case, Catalyst 3 produced the higher abundance of both alkyl and aromatic hydrocarbons (peaks with the middle and largest second-column retention times) compared with Catalyst 4.

As noted in the main text, peak-region features provide a basis for more comprehensive comparisons. Figure 7 shows the differences between Catalysts 3 and 4 using the peak-region features derived from the composite chromatogram shown in Figure 8. The comparison of this figure to Figure 6 clearly shows that the peak-region features provide far more comprehensive fingerprints.

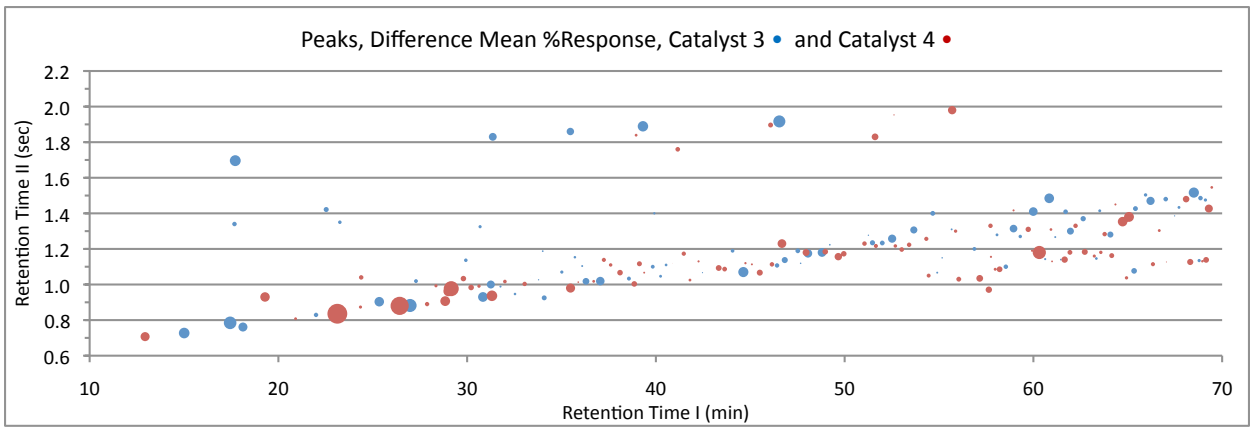

Figure 6: Each circle indicates the retention times and difference between the mean percent-responses for peaks with Catalyst 3 and Catalyst 4. A blue circle indicates that the peak percent-response with Catalyst 3 is larger and a red circle indicates that the peak percent-response with Catalyst 4 is larger. The scale for circle areas is relative to the magnitude of the largest difference.

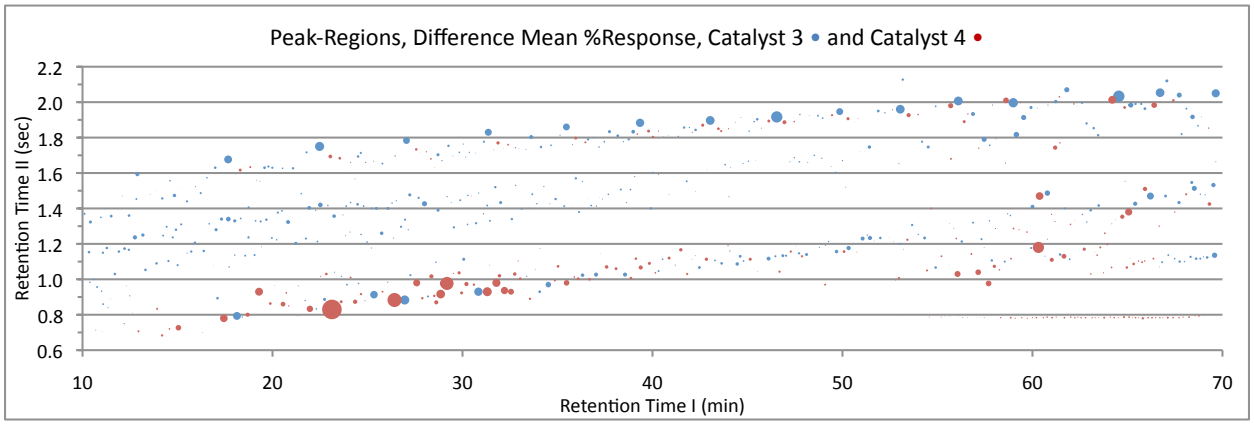

Figure 7: Each circle indicates the retention times and difference between the mean percent-responses for peak-regions with Catalyst 3 and Catalyst 4. A blue circle indicates that the peak-region percent-response with Catalyst 3 is larger and a red circle indicates that the peak-region percent-response with Catalyst 4 is larger. The scale for circle areas is relative to the magnitude of the largest difference.

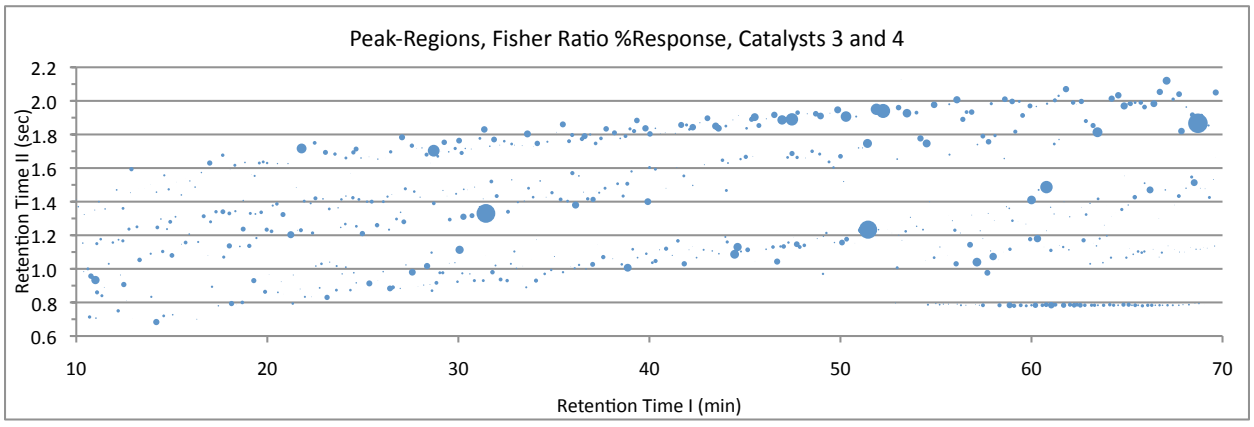

Figure 8: Each circle indicates the retention times and the Fisher's Ratio between the mean percent-responses for peak-regions with Catalyst 3 and Catalyst 4. The scale for circle areas is relative to the magnitude of the largest Fisher's Ratio.

Table 1: Percent of peak-region features and percent correct sample discrimination as a function of Fisher's Ratio threshold.

	Fisher's Ratio Threshold				
	0.00	1.00	2.00	4.00	8.00
% Features	100.00	49.88	35.98	21.07	10.27
% Correct	68.83	93.13	96.74	99.08	99.95

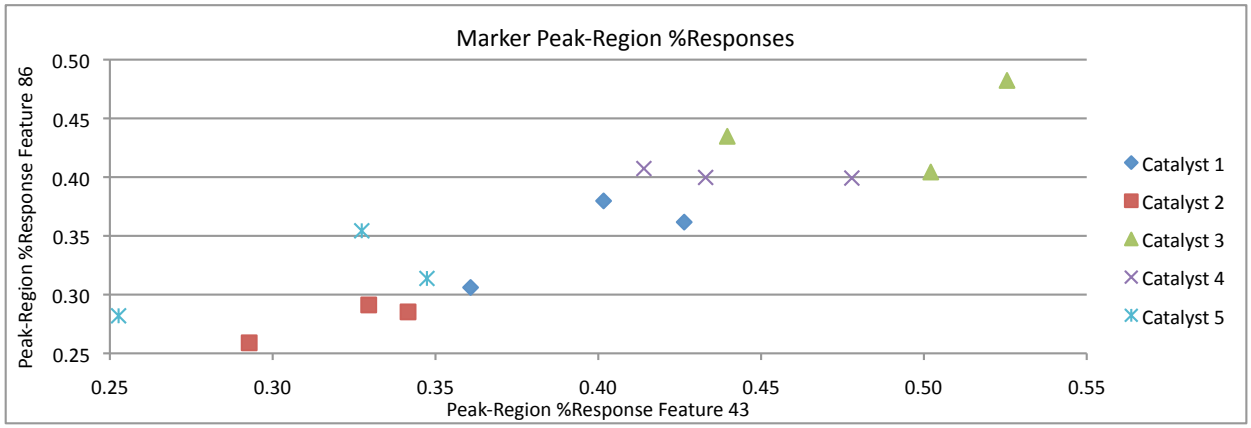

Figure 9: The percent-responses for two marker peaks for the fifteen bio-oil chromatograms.

In many comparative analyses, it is important to evaluate the statistical significance of the observed differences. Fisher's Ratio is the squared difference between class means (the between-class scatter) relative to the sum of the class variances (the within-class scatter):[2]

$$\text{Fisher's Ratio} \quad = \quad \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \tag{1}$$

where μ_i and σ_i^2 are the mean and variance for class i. In this example, each class comprises the chromatograms from one of the five different catalysts. Figure 8 shows the Fisher's Ratio for peak-region features in chromatograms with Catalysts 3 and 4. The Fisher's Ratios for the features ranged from 0.0 to 13.0, with an average of 0.5. The distribution of peak-region features by Fisher's Ratio threshold is shown in Table 1, *e.g.*, nearly half of the peak-region features have a Fisher's Ratio of at least 1.00. The Fisher's Ratios can be used to identify potential chemical markers that are especially discriminating with respect to the different catalysts.

A feature with a large Fisher's ratio is a promising chemical marker of the differences between sample classes. Figure 9 shows the percent-response values in each of the fifteen chromatograms for two peak-region features that are good multi-class linear discriminants (*i.e.*, have large Fisher's ratios) and have higher than average percent-responses. For these features, chromatograms with Catalyst 3 have the largest average percent-response and chromatograms with Catalysts 2 and 5 have the smallest average percent-responses. It is notable that the chemical similarities for Catalysts

1 and 4 and for Catalysts 2 and 5 (each pair uses the same metal) are evident in the plot in Figure 9.

Cross-validation experiments support the discriminatory significance of the peak-region features with larger Fisher's Ratios. In these experiments, each one of the three sets of samples for each catalyst was taken in turn as the testing set and the other two sets of samples for each catalyst were taken in turn as the training set. Then, for each pair of catalysts, each peak-region feature was used to discriminate the pair of test samples with respect to the catalysts. In this, the training-set chromatograms (for which the catalysts are taken to be known) are used to distinguish the catalysts of the test chromatograms (for which the catalysts are taken to be unknown). Random guessing would be expected to distinguish the correct catalysts at a 50% rate.

Overall, individual peak-region features were able to distinguish the samples correctly in 68.83% of the tests. For the 49.88% of peak-region features with Fisher's Ratio of 1.00 or greater, 93.13% of samples were distinguished correctly; and, for the 10.27% of peak-region features with Fisher's Ratio of 8.00 or greater, 99.95% of samples were distinguished correctly. The success rates for intermediate Fisher's Ratio thresholds are shown in Table 1. These experiments indicate that even with very small sample sets, many of the peak-region features are useful for discriminating differences with respect to catalysts.

References

(1) Meeusen, W.; Cuyvers, L. *J. Comput. Appl. Math.* **1975**, *1*, 185–203.

(2) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; John Wiley and Sons: New York NY, 2000.