RELIABLE PEAK SELECTION USING COMPUTER VISION-BASED CLUSTERING FOR MORE EFFICIENT ALIGNMENT OF MANY TWO-DIMENSIONAL CHROMATOGRAMS

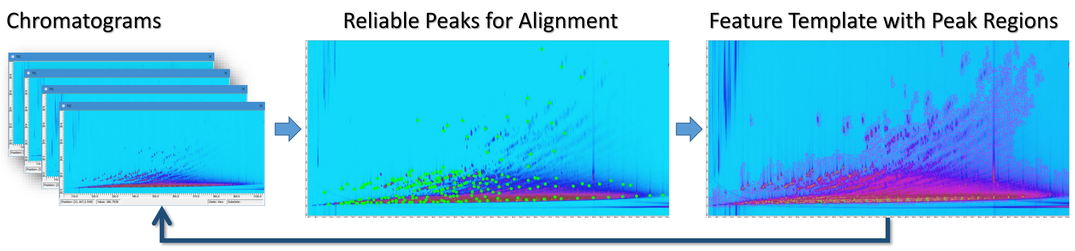
Chase Heble;[1] Daniel Geschwender;[1] Qingping Tao;[1] Stephen E. Reichenbach;[1,2]
[1] GC Image, LLC, Lincoln NE; [2] University of Nebraska, Lincoln NE

Introduction

Comprehensive two-dimensional chromatography, such as GCxGC and LCxLC, is a powerful technology for analyzing the patterns of constituent compounds in complex samples. However, matching chromatographic features for comparative analysis across large data sets is challenging and time consuming. Efficient chromatographic feature alignment and matching methods are necessary.

Investigator Workflow

Our Investigator framework **[1]** applies peak matching techniques to align multiple chromatograms automatically, and then converts peaks into peak-region features that comprehensively capture the pattern of peaks across all sample chromatograms.



To find reliable peaks that match across many chromatograms,

- A pairwise peak matching is performed on the entire data set.
- From the pairwise matching, a comprehensive set of reliable peaks is created, which correspond across all or most chromatograms.

However, this approach is limited by the combinatorial nature of the pairwise matching and can be time consuming for analyzing hundreds of chromatograms. To overcome these drawbacks, we evaluated various cluster-based methods for reliable peak selection.

Method – Cluster-based Reliable Peak Selection

To limit the number of pairwise comparisons required, samples can be grouped into clusters. Pairwise matching will be performed within-cluster. The within-cluster results can then have between-cluster pairwise matching performed to get a global reliable peak selection.

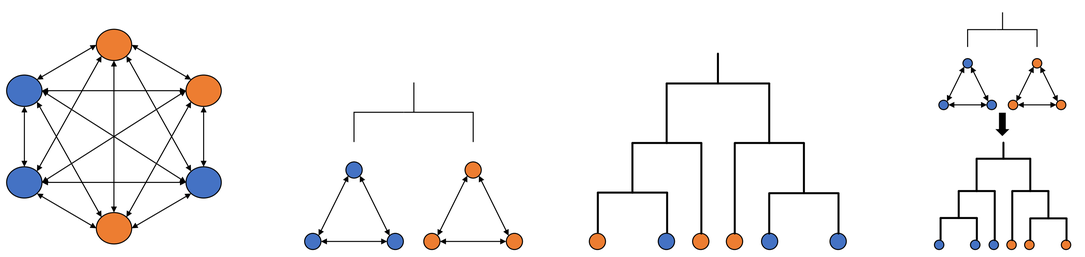
Technique 1 - Class-based Clustering using manual classification labeling to make a single cluster for each class. This utilizes class composite images which have been proved useful for between class analysis **[2]**.

Technique 2 - Hierarchical Clustering using computer vision-based similarity metrics that group chromatograms with similar features.

- Enhanced Correlation Coefficient (ECC), a measure of image similarity from computer vision, is calculated between all the chromatograms.
- The ECC score matrix can be quickly calculated and passed to a hierarchical cluster algorithm (HCA) to generate a tree of the chromatogram set.
- The hierarchical cluster algorithm (HCA) tree is then flattened based on appropriate cluster sizes and number of clusters.

Technique 3 - Class-Based ECC HCA using an alteration of the ECC HCA tree by introducing the classification clusters as the first step in the algorithm.

To evaluate the three techniques, brute-force pairwise matching was also performed to obtain a baseline.



Evaluation – Sample Sets

Three different sample sets are used to evaluate the performance of the above techniques.

Chocolate - Samples of 9 different dark chocolates **[3]**
- Instrumentation: Agilent 7890 GC/J&X SSM 1800 coupled with JEOL AccuTOF

Diesel - Samples of diesel fuel from 4 difference brands **[4]**.
- Instrumentation: Shimadzu 2010 Ultra GC/MS with Zoex ZX2 thermal modulator

Wine - Samples of Cabernet and Merlot wines from wineries in Brazil **[5]**.
- Instrumentation: LECO Pegasus GCxGC-TOFMS

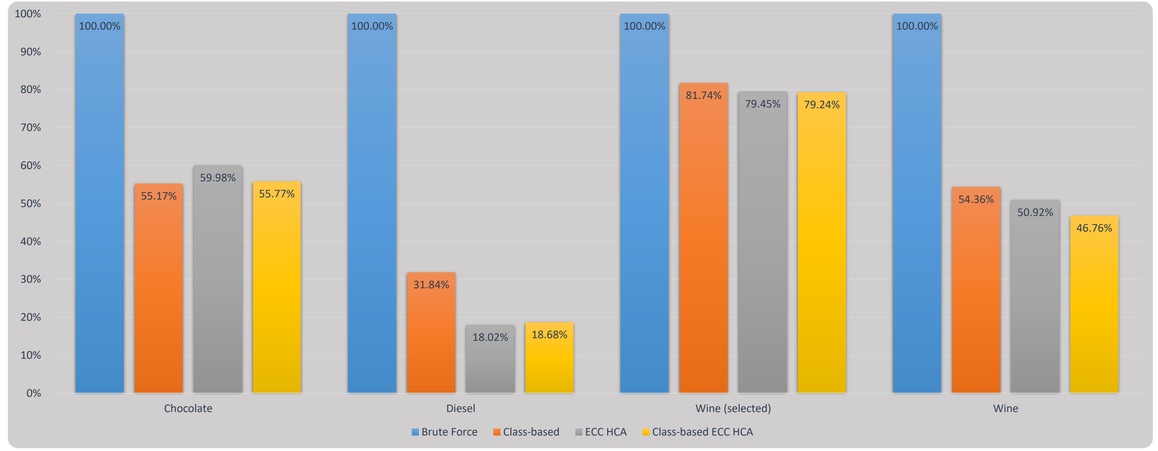
Evaluation – Performance

The three data sets have varying number of chromatograms, number of peaks, file size, and number of classes as shown in the table below.

Data Set	# of Chromatograms	Average # of Peaks per Chromatogram	Average File Size (MB)	Number of Classes
Chocolate	40	928	23.68	10
Diesel	80	2300	528.56	4
Wine (selected)	75	181	937.30	2
Wine	149	181	937.30	2

(A subset of the wine samples was used as an additional data set to evaluate the method.)

Performance of each technique was evaluated on a Win 10 PC with Intel Core i7 CPU.



The values are shown as the runtime relative to the brute-force pairwise runtime. Lower is better.

The performance chart shown depicts a substantial decrease in runtime when performing the cluster-based peak selection.

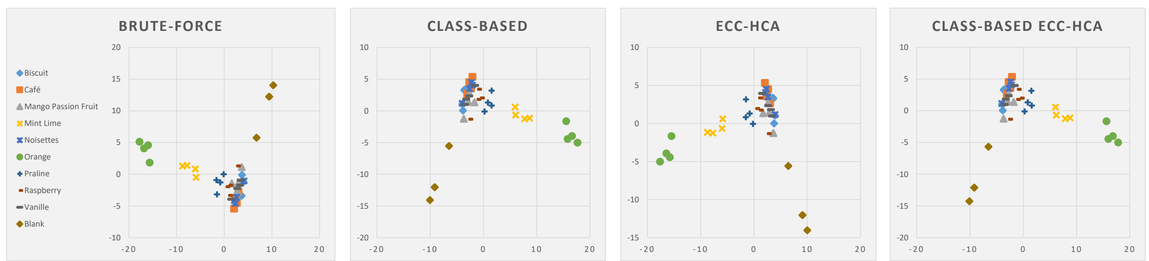
- Class-based clustering improved throughout, performing similar to the HCA clustering methods for some of the data sets.
- Between the data sets, the performance improvements of the clustering scaled with the number of chromatograms (seen in comparing the wine datasets) and the number of features in the samples (seen in the diesel dataset).
- The file size of the chromatograms had a negative effect on the cluster performance, where the IO had an increased impact on the runtime. This was apparent in the Wine data set, where there were the average peak count per sample was relatively small while having about 1GB data file size each.

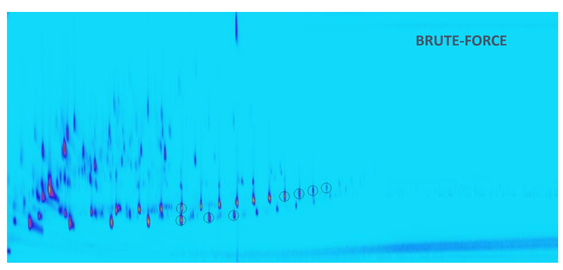
Evaluation – Quality

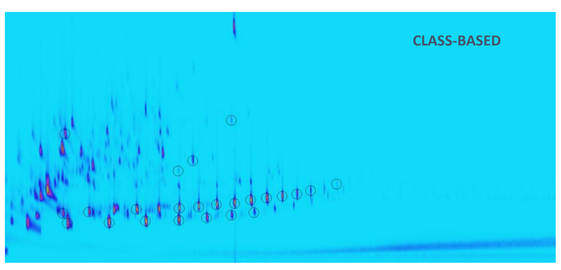
The results of each techniques were verified with the following evaluations:

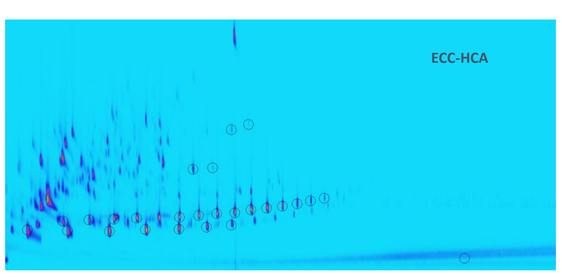
- The reliable peak set is viewed to ensure a minimum set of reliable peaks are selected.
- The alignment is evaluated by checking the peak-regions in the feature template.
- The PCA chart is generated using the Percent Response of the feature peak-regions. Using the Brute-force results as a baseline, the class separation is compared.

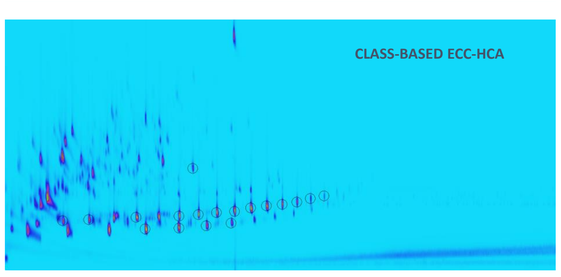
Example – Chocolate:
- The PCA charts show equivalent class separation.
- Similar and more reliable peaks are found with clusters as shown on the cumulative chromatograms.











GC Image
Software for Multidimensional Chromatography

University of Nebraska Lincoln

References

1. S. Reichenbach, X. Tian, Q. Tao, E. Ledford, Z. Wu, O. Fiehn. "Informatics for Cross-Sample Analysis with Comprehensive Two-Dimensional Gas Chromatography and High-Resolution Mass Spectrometry (GCxGC-HRMS)". Talanta, 83(4):1279-1288, 2011.
2. A. Caratti, S. Squara, C. Bicchi, Q. Tao, D. Geschwender, S. Reichenbach, F. Ferrero, G. Borreani, C. Cordero. "Augmented Visualization by Computer Vision and Chromatographic Fingerprinting on Comprehensive Two-dimensional Gas Chromatographic Patterns: Unraveling Diagnostic Signatures in Food Volatilome." Journal of Chromatography A, 1699: 464010, 2023.
3. B. Weggler, L. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S. Reichenbach, P. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J. Focant. "Benchmark GCxGC Data, Chocolate". https://doi.org/10.7910/DVN/AKT6BH, Harvard Dataverse, 2020.
4. Z. Wu, J. Coleman, Q. Tao. "Distinguishing Commercial Diesel Fuel Brands Using Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry". The International Symposium on Capillary Chromatography (ISCC), Riva del Garda, Italy, May 2018.
5. S. Reichenbach, C. Zini, K. Nicolli, J. Welke, C. Cordero, and Q. Tao. "Benchmarking Machine Learning Methods for Comprehensive Chemical Fingerprinting and Pattern Recognition". Journal of Chromatography A, 1595:158-167, 2019
6. Data processes and screenshots for this publication are from an alpha version of GC Image v2022r2 GCxGC-HRMS (Visit www.gcimage.com for current v2022 releases).