Documentation for the Arey-Gros code to perform peak-tracking across GC×GC chromatograms as implemented in Matlab

Version 1.0.0

J. Samuel Arey and Jonas Gros, 2021.

Please cite the following article when publishing any results obtained by use of this software:

- Wardlaw, G. D., Arey, J. S., Reddy, C. M., Nelson, R. K., Ventura, G. T., Valentine, D. L., "Disentangling oil weathering at a marine seep using GC×GC: Broad metabolic specificity accompanies subsurface petroleum biodegradation", *Environmental Science & Technology*, 42, 19, 7166-7173, 2008.

1 The purpose of the algorithm

The Matlab code is designed to track corresponding peaks across two chromatograms obtained with comprehensive two-dimensional gas chromatography (GC×GC) coupled to a univariate detector, for example a flame ionization detector (FID). The peak tracking is based on matching retention times of detected peaks within a search window. The code takes *peak tables* as input, where we define "peak tables" as tables of integrated peaks listing their first- and second- dimension retention times and peak volume.

For robust results, we advise to include the following steps prior to using the current algorithm:

- Apply a carefully chosen baseline correction (e.g., https://github.com/jsarey/GCxGC-baseline-correction)[1–4]
- Perform peak delineation and integration to generate peak tables (e.g., using the inversed watershed algorithm implemented in the GC Image software)[5,6]
- Align peak tables (e.g., https://github.com/jsarey/GCxGC-alignment)[7]
- Correct peak tables for any evaporative loss during sample concentration
- Normalize[e.g.,8] peak tables

2 What the Matlab code does

It applies an updated version of the peak-tracking algorithm of Wardlaw et al. (2008).[9] The code has been modified to enable peaks to be labeled as "disappeared", as can happen for example during biodegradation of oil. One further update[10] was applied to the Wardlaw et al. codes.

The code matches peaks between a template and a target chromatogram. In other words, it finds peaks common to two peak lists based on retention-time matching. The different criteria have been described in detail in Wardlaw et al. (2008). Briefly, a search oval is defined by the user, and the algorithm looks for unique corresponding peaks within the search oval. Multiple peaks in the search oval are not retained for peak matching. Absence of any peak in the target peak list within the search oval of a peak in the template peak list may be considered a disappeared peak (option offered by the code). Except for disappeared peak, a positive peak matching corresponds to the unique presence of a peak of the target peak list within the search oval of a peak of the template chromatogram, and vice versa. Peaks that are within the search oval of each other within the template peak list are discarded.

3 Organization of the model file directory. Where to find what.

The model code is organized as follows. From the base directory of the program two folders are present, called `users/`, and `model_code/`.

These two folder names should not be changed.

The user should only need to operate from within the folder called `users/`. Normally, nothing should be changed or adjusted in the folder called `model_code/`.

Within the folder called `users/`, the organization of folders and files is user-defined. The user can define directory paths with the following two model variables in the file `main.m`:

A) `input_path`. This variable indicates the directory path location of the input files. The input path variable is set in the file called `main.m`, and it assumes that `main.m` is located in the directory `users/`. The `input_path` variable also assumes that the indicated directory exists.

B) `output_path`. This variable indicates the directory path location of the output files.

Note: both `input_path` and `output_path` should be relative paths, starting with `users/`, which itself should be situated within the program base directory.

Note 2: The operating system must allow Matlab to write files within the `output_path` directory. For example, on windows computers, **do not** locate the base directory within the `C:\Program Files` folder.

4 Steps for use of the algorithm

4.1 Prepare input files

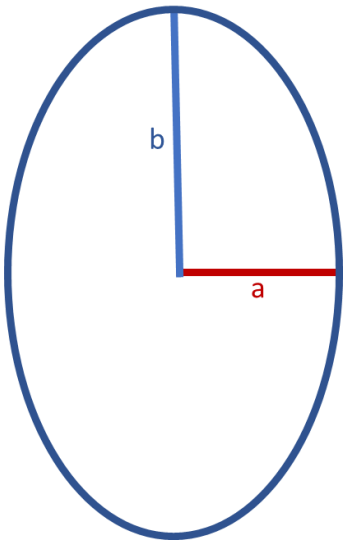
The model requires peak tables. The required structure of these files is:

The peak tables should have three columns corresponding to: (1) first dimension retention time (min); (2) second dimension retention time (s); (3) peak volume (integrated FID signal). In the file, the columns should be separated using commas as separators. Each line in the file corresponds to an individual peak. The algorithm expects the file to be a text file (e.g., csv file).

4.2 Adjust parameters in **`main.m`**

Adjust the parameter settings that appear in `main.m`. This file can be read and modified from within Matlab or using a generic text editor. This is the only Matlab file that you need to adjust. Most of these parameters are self-explanatory and/or discussed above.

The search oval is defined as:



where a is the search distance along the first dimension (min) and is set with the variable `search_dist`. The ratio a/b (min/s) is defined with the variable `x_to_y_ratio`. Together, these two variables define the size of the search oval.

Set `match_disappeared_peaks` to 1 to enable peaks to disappear (an absence of a peak in the target peak list within the search oval of a peak in the template peak list is considered as a 'disappeared peak' or 'absent peak'). If `match_disappeared_peaks` is set to 0, then these peaks are considered as non-matched and not included in the list of matched peaks.

5 Name and contents of the output file

The tracked peaks are saved in the file `Output_file_name`. The files contains one line per tracked peak. The first and second columns are the first and second dimension retention times, respectively. They are provided in units of min and s, and the retention times listed are for the selected template peak table. The third and fourth columns provide the peak volumes in the template and target chromatograms, as defined in main.m. Disappeared peaks (if enabled by choosing `match_disappeared_peaks = 1`) are listed as having peak volumes of "-9999".

Additionally, the minimum peak volume in each peak table is displayed in the Matlab main window. This value might be taken as a proxy for detection limit. Disappeared peaks might be assigned a peak volume of half of this value in further data processing, might be counted as zero, or any different choice made by the user.

6 References

(1) Reichenbach, S. E.; Ni, M.; Zhang, D.; Ledford, E. B. Image Background Removal in Comprehensive Two-Dimensional Gas Chromatography. *Journal of Chromatography A* **2003**, *985* (1), 47–56. https://doi.org/10.1016/S0021-9673(02)01498-X.

(2) Eilers, P. H. C. Parametric Time Warping. *Anal. Chem.* **2004**, *76* (2), 404–411. https://doi.org/10.1021/ac034800e.

(3) Gros, J.; Eilers, P. H. C.; Arey, J. S. *Gros-Eilers-Arey Code to Perform Baseline Correction of GC×GC Chromatograms*; https://github.com/jsarey/GCxGC-baseline-correction, 2015.

(4) Gros, J.; Reddy, C. M.; Aeppli, C.; Nelson, R. K.; Carmichael, C. A.; Arey, J. S. Resolving Biodegradation Patterns of Persistent Saturated Hydrocarbons in Weathered Oil Samples from the *Deepwater Horizon* Disaster. *Environ. Sci. Technol.* **2014**, *48* (3), 1628–1637. https://doi.org/10.1021/es4042836.

(5) Reichenbach, S. E.; Ni, M.; Kottapalli, V.; Visvanathan, A. Information Technologies for Comprehensive Two-Dimensional Gas Chromatography. *Chemometrics and Intelligent Laboratory Systems* **2004**, *71* (2), 107–120. https://doi.org/10.1016/j.chemolab.2003.12.009.

(6) Samanipour, S.; Dimitriou-Christidis, P.; Gros, J.; Grange, A.; Samuel Arey, J. Analyte Quantification with Comprehensive Two-Dimensional Gas Chromatography: Assessment of Methods for Baseline Correction, Peak Delineation, and Matrix Effect Elimination for Real Samples. *Journal of Chromatography A* **2015**, *1375*, 123–139. https://doi.org/10.1016/j.chroma.2014.11.049.

(7) Gros, J.; Nabi, D.; Dimitriou-Christidis, P.; Rutler, R.; Arey, J. S. Robust Algorithm for Aligning Two-Dimensional Chromatograms. *Anal. Chem.* **2012**, *84* (21), 9033–9040. https://doi.org/10.1021/ac301367s.

(8) Prince, R. C.; Elmendorf, D. L.; Lute, J. R.; Hsu, C. S.; Haith, C. E.; Senius, J. D.; Dechert, G. J.; Douglas, G. S.; Butler, E. L. 17α(H),21β(H)-Hopane as a Conserved Internal Marker for Estimating the Biodegradation of Crude Oil. *Environ. Sci. Technol.* **1994**, *28* (1), 142–145. https://doi.org/10.1021/es00050a019.

(9) Wardlaw, G. D.; Arey, J. S.; Reddy, C. M.; Nelson, R. K.; Ventura, G. T.; Valentine, D. L. Disentangling Oil Weathering at a Marine Seep Using GC×GC: Broad Metabolic Specificity Accompanies Subsurface Petroleum Biodegradation. *Environ. Sci. Technol.* **2008**, *42* (19), 7166–7173. https://doi.org/10.1021/es8013908.

(10) Gros, J. Investigating the Fate of Petroleum Fluids Released in the Marine Environment with Comprehensive Two-Dimensional Gas Chromatography and Transport Models, EPFL, Lausanne, 2016.

Contacts:

For questions, problems, or bug reports, feel free to contact Jonas Gros (gros.jonas@gmail.com) or J. Samuel Arey (arey@alum.mit.edu)