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

Version 1.0.0

Jonas Gros and J. Samuel Arey, 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 (or *search oval*). 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. Such peak tables can be generated with most GC×GC softwares (e.g., GC Image using the inversed watershed algorithm[1,2]) and formatted by the user according to the needs of the current algorithm (as described in section 4.1).

Depending on the context of the study, different additional steps may be considered by the user such as baseline correction (e.g., https://github.com/jsarey/GCxGC-baseline-correction),[3–6] chromatogram alignment (e.g., https://github.com/jsarey/GCxGC-alignment),[7] or normalization.[e.g.,8]

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" in a second peak list relative to a reference peak list, as can happen for example during biodegradation of oil (or as a result of other processes). The code matches peaks between a first (reference) and a second chromatogram. In other words, it finds peaks common to two peak lists based on retention-time matching. The different criteria are described below.

The algorithm assigns each considered peak table as either a “template” peak list or as a “target” peak list, according to the procedure described below.

Step 1. The algorithm assigns the reference peak table as the “template” peak list, and it assigns the second peak table as the “target” peak list.

Step 2. For each candidate peak in the template peak list, the algorithm evaluates all other peaks in both the template peak list and the target peak list, using a search oval centered on the two-dimensional retention time coordinates of the candidate template peak.

The algorithm then applies the following criteria:

(a) If the algorithm finds multiple target peaks within the search oval radius of the candidate template peak, then the algorithm rejects all of these target peaks and also the candidate template peak that originated the search oval. For this case, it is interpreted that the algorithm lacks sufficient resolution (represented by the search oval) to effectively distinguish among multiple target peaks.

(b) If the algorithm finds any additional template peak(s) within the search oval originated by the candidate template peak, then it rejects all of these template peaks, including the template peak that originated the search oval, and it also rejects any target peaks that fall within the same search oval. Analogous to case (a), in case (b) it is interpreted that the algorithm lacks sufficient resolution (represented by the search oval) to effectively distinguish among multiple template peaks.

(c) If the algorithm finds no target peaks within the search oval radius of the candidate peak, then either:

    (c1) the algorithm accepts the template peak for this step and labels it as ‘disappeared’ (absent) peak in the target chromatogram. Case (c1) represents the interpretation that the target peak had disappeared due to some process (biodegradation, evaporation, dissolution, or any other process relevant to the specific data set). This differs from the criterion described by case (c2) below.

    (c2) the algorithm rejects the template peak. Case (c2) imposes the strict criterion that a tracked peak must be rejected if it does not appear within the peak lists of both the template and target samples.

    If chosen by the user (see section 4.2), the algorithm uses criteria (c1) at step 2 when the reference peak list is selected as the template, and criteria (c2) when the reference peak list is selected as the target chromatogram (in step 3). Alternatively, the user can decide that the algorithm always applies the criterion (c2), which may be more relevant to the context of some studies. (The criterion (c2) was applied by Wardlaw et al.[9]).

(d) If the algorithm finds only a single target peak within the search oval and it finds no other template peaks within the search oval, then this target peak is considered a tentative match with the template peak that originated the search oval, and the target peak and template peak both pass step 2.

In the evaluation of cases (a)-(d), the term "reject" only indicates that the rejected peaks are flagged for non-acceptance. The algorithm does not remove any rejected peaks from the peak tables. Also, the algorithm includes the presence of all previously rejected peaks, when evaluating the acceptance criteria for any given peak. These features ensure that the status of any individual peak does not artefactually bias the results obtained for any other peak. They also ensure that the results of the peak-tracking algorithm are not dependent on the order in which the peaks are evaluated.

Step 3. The algorithm repeats the search procedure described in step 2, except that the assignments of the template and target peak tables are swapped. The algorithm now assigns the second peak table as the template peak list, and it assigns the first (reference) peak table as the target peak list. The algorithm considers that candidate peaks are successfully matched only if they met the acceptance criteria in step 2 and also created the same matching pairs both before and after the swap.

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

The peak-tracking algorithm assumes that all of the chromatograms whose peaks will be "tracked" were collected using an identical GC×GC instrument program and column plumbing (or otherwise were aligned).

4.1 Prepare input files

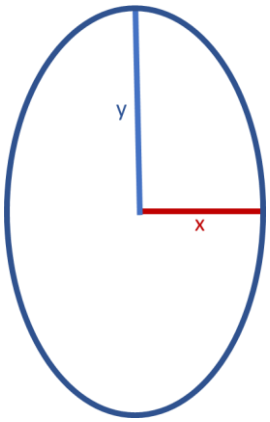
The model requires peak tables. The required structure of these files is:

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

(note: the input dimensions of the search oval depend of the units used for retention times (e.g., minutes versus seconds). We advise the users to follow the suggested choice of units described in this documentation to ensure consistent results.)

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' when the template peak list is the reference peak list). 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 file contains one row per 'tracked peak', i.e. that each row represents a peaks that is considered to appear in both peak tables. 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 reference peak table. The third and fourth columns provide the peak volumes in the reference and second peak tables, 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.; 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.

(2) 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.

(3) 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.

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

(5) 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.

(6) 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.

(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.

Contacts:

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