Documentation for the Zushi-Gros-Arey code to align GC×GC–MS chromatograms as implemented in Matlab (ZAG alignment code)

Version 1.0.0

Jonas Gros, Yasuyuki Zushi, and J. Samuel Arey, 2017.

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

Zushi, Y., Gros, J., Tao, Q., Reichenbach, S. E., Hashimoto, S., Arey, J. S. Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry. *J. Chromatogr. A,* **2017**, *28*, 121-129.

Gros, J., Nabi, D., Dimitriou-Christidis, P., Rutler, R., Arey, J. S. Robust algorithm for aligning two-dimensional chromatograms. *Anal. Chem.* **2012**, *84*, 9033-9040.

1 The purpose of the algorithm

The Matlab code is designed to correct the small run-to-run shifts in retention times in GC×GC chromatograms coupled to a mass spectrum (MS) detector. For data acquired with a univariate detector, please use the similar algorithms available at: https://github.com/jsarey/GCxGC-alignment.

2 The tested validity of the algorithm

The algorithm was tested for chromatograms resulting from comprehensive two-dimensional gas chromatography (GC×GC) coupled to a high-resolution time-of-flight mass spectrometry detector (HRTOFMS). It is believed that the algorithm should be applicable to any two-dimensional separation coupled to a multivariate detector, e.g, LC×LC-MS, LC×LC-DAD, and 2-DE-MS. However, some careful evaluation is advisable for very different systems. (Refer to Zushi et al. 2017 for the systems tested.)

3 What the Matlab code does

It takes a target chromatogram and aligns it to a reference chromatogram, thereby generating an aligned chromatogram. The retention times of peaks in the aligned chromatogram should be closer to the retention times of peaks in the reference chromatogram than is the case between the target and reference chromatograms.

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

The model code is organized as follows. From the base directory 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 `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/`. `users/` should be situated in 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.

5 Steps for use of the algorithm

5.1 Prepare input files

The model requires the reference chromatogram, the target chromatogram, and the positions of the alignment points in both the target and reference chromatograms. The required structure of these files is described in turn below:

A) The reference and target chromatograms.
(variables `Reference_chromatogram_file` and `Target_chromatogram_file`)
By default, these chromatograms should be cdf files. The cdf file follow the Network Common Data Form (netCDF) file format, used to store array-oriented scientific data. For more information see Unidata website. The cdf file format does not intrinsically depend on application-specific formats. Most of the MS instruments from various suppliers support the conversion of their analytical output to the cdf file format. If an instrument does not provide the option to export a cdf file, several softwares, e.g. GC Image, can convert the instrumental output to the cdf file format.

B) The positions of the alignment points in the reference and target chromatograms.
(variables `Reference_alignment_pts_file` and `Target_alignment_pts_file`)

These csv files should contain two columns with first and second dimension retention positions of alignment points, respectively, in units of pixels. By default, the convention retained here is that the first pixel in the chromatogram has the position (1,1) in pixel units. When importing peak positions from an external program, the user is advised to check for consistency.

However, the user can alternatively provide inputs in time units (minutes for the first dimension and seconds for the second dimension). If doing so, the user has to set `units` to `‘time’` in `main.m` (instead of the default value of `‘pixel’`). If the user does so, all his inputs will be assumed to be in units of time.

Note: the order of the alignment points in the two files should correspond.

How to select good alignment points?

- How to identify corresponding peaks? The user needs to be sure that the peaks selected correspond to the same analyte in the reference and target chromatograms. This can be ascertained by comparison of MS spectra.
- How many? The number of alignment points required to achieve good alignment may vary for widely different chromatograms. In practice, we found that a minimum of ~10 alignment points is sufficient for the alignment.
- Where? We advise that alignment points are chosen in each part of the chromatogram that exhibit shifting trends different from neighboring regions, and that most of the part of the chromatogram that is of interest to the user is within the convex hull of the alignment points chosen.
- And finally, the user can also proceed in an iterative way: in case the result of the alignment is not satisfactory enough, the set of alignment points can be modified, and the alignment code run again.

What is the position of a peak?

The definition retained here is that the position of a peak corresponds to the position of the pixel of this peak that has the maximum signal intensity value.

The use of other definitions for the position of a peak is possible but was not tested.

5.2 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, for normal use of the alignment code. Most of these parameters are self-explanatory and/or discussed above.

Three of these parameters are discussed below.

`Typical_peak_width`: this variable contains the first dimension (first element) and second dimension (second element) typical size of a peak, in units of pixels (by default). It corresponds approximately to the number of pixels corresponding to two standard deviations of a peak (assumed Gaussian), for a typical peak. The determination of the typical peak width parameter is somewhat arbitrary, but the algorithm results are not very sensitive to this parameter. Therefore, a visual determination by inspection of one-dimensional slices of a typical peak (ideally close to the center of the chromatogram) should suffice.

If the user chooses to set `units` to `'time'` in `main.m` (instead of the default value of `'pixel'`), then `Typical_peak_width` has to be provided in units of time (minutes for the first dimension and seconds for the second dimension). If the user does so, all his inputs will be assumed to be in units of time.

`Model_choice`: this variable enables the user to select among the two models investigated by Zushi et al. 2017. The default (`'normal'`) is advised for most cases. A second method called `'DualSibson'` may perform better under particular circumstances. Refer to the publication for more details.

6 Name and contents of the output file

The aligned chromatogram is saved in a cdf file. The name of the file is the same as the name of the target chromatogram file, with the string "_ALIGNED" appended.

7 Figures displayed

After the completion of the alignment, two figures are displayed:

A) A three-panel figure entitled "Total Ion Chromatograms (TICs)": the total ion chromatograms (TICs) of the reference, the target, and the aligned chromatograms. Each of them is overlaid with the positions of the alignment points as black circles.

B) A two-panel figure entitled "Difference chromatograms": pixel-by-pixel difference chromatograms between the TICs of the reference and target chromatograms (first panel), and between the TICs of the reference and aligned chromatograms (second panel). Pixels that appear blue have a larger signal intensity value in the reference chromatogram compared to the other chromatogram. Pixels that appear red have a larger signal intensity value in the other chromatogram with respect to the reference chromatogram. Pixels that appear white have about the same signal intensity value in both chromatograms.
If the chromatograms are somehow normalized, especially for chromatograms sharing high compositional similarity, a better alignment should be highlighted by a whiter difference chromatogram (less red and blue regions).
(If your chromatograms correspond to samples having very different compositions or if they are not normalized, please consider only the "Total Ion Chromatograms (TICs)" figure.)
("Difference chromatograms" are defined in the article "*Tracking the weathering of an oil spill with comprehensive two-dimensional gas chromatography*" by R. K. Nelson et al., *Environmental Forensics*, 2006.)

8 References

- Gros, J., Nabi, D., Dimitriou-Christidis, P., Rutler, R., Arey, J. S. "Robust algorithm for aligning two-dimensional chromatograms". *Anal. Chem.* **2012**, *84*, 9033-9040.
- Swarthout, R. F., Gros, J., Arey, J. S., Nelson, R. K., Valentine, D. L., Reddy, C. M. "Comprehensive Two-Dimensional Gas Chromatography to Assess Petroleum Product Weathering", chapter in the

book “Hydrocarbon and Lipid Microbiology Protocols”, McGenity, T. J., Timmis, K. N., Nogales Fernández, B. (Eds.), Springer Protocols Handbooks, Springer: Berlin, **2017**.

- Zushi, Y., Gros, J., Tao, Q., Reichenbach, S. E., Hashimoto, S., Arey, J. S. “Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry”. Journal of Chromatography A 2017, vol 28, p 121-129.

9 Acknowledgements

We warmly thank Shunji Hashimoto for agreeing to perform beta-testing during the development of the codes.

Contacts:

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