User's Guide

GCxGC-NMF-Classification_v1.1.2

Version 1.1.2

2018.1

Please cite the following article when users use this source code.
Zushi Y. and Hashimoto S., Direct classification of GC × GC-analyzed complex mixtures using non-negative matrix factorization based feature extraction, *Anal. Chem.* 2018, *In Press*.

Description:

This R source code, which is available from the GitHub repository "GCxGC-NMF-Classification", is developed for classifying GCxGC data based on a NMF algorithm. When univariate GCxGC data, such as GCxGC-FID or TIC of GCxGC-MS, are provided as the input data, these data are automatically classified by the NMF algorithm according to the similarity of their 2D chromatogram patterns. As the subsequent analysis, unknown samples of the GCxGC data are rapidly classified based on the ready-made NMF class.

Author (Contact):

Yasuyuki ZUSHI (yasuyuki.zushi@gmail.com)

Requirement:

Hardware computer (≧16GB RAM)
Software R freely available from https://cran.r-project.org/
R package "NMF"*
R package "JPEG"*
*These packages are installed along with the code execution.

How to use:

1. Download the zip file of "GCxGC-NMF-Classification_v1.1.2", unzip it, and place the folder where you prefer.
2. NMF dataset: Open the "GCxGC-NMF-Classification_v1.1.2" folder, and place your GCxGC data as csv format in the folder of "Input" > "NMF_Input_Dataset" for making

NMF-classes. Be sure that all the GCxGC data in the “NMF_Input_Dataset” folder are used for the NMF, therefore, any of unnecessary file must not be placed in the folder.

Unknown sample dataset: Also, place your GCxGC data of unknown samples in the folder of “Unknown_Samples_to_Classify” to classify them into an NMF-class. Be sure that any of unnecessary file must not be placed in the folder.

Both the “NMF dataset” and “Unknown sample dataset” should be taken by same conditions on GC run (run time, modulation period) and detector (data rate).

3. Start R software, then open the code “Just_Run_Me.r” included in the folder as a R script file.
4. In the R script, rewrite the folder location path that is fit for your own environment. Rename an output file name, if you prefer.
5. If you have not install R package “NMF” and “JPEG”, required to install it as an initial setting. As a default setting, the package installation will be proceeded. Next time to run the code, you can skip this process.
6. Change default parameters, such as data sampling rate, masking range, picture resolution, and the number of ranks, for classification according to the reference 1) and comments in “Just_Run_Me”.
7. After setting all the parameter, click [Edit]=>[Run all], then output files will be produced in the “Output” folder. To complete the process for the test dataset that initially included in the folder, it takes several tens of seconds.

   You can skip the process of NMF calculation as long as you have the respective NMF result, then directly run the code for the unknown sample classification. This will be useful after the first time classification.
8. In the “Output” folder, extracted chromatogram features (GCxGC_chromatogram_Rank[X].png), NMF coefficients for each GCxGC sample (Sample_Coefficients.png), list of classified result for each GCxGC sample (A file by named by yourself, as a default, GCxGC_NMF_result.csv), a picture of all the GCxGC chromatogram as NMF input with assigned NMF-class (NMFinput_Append_withImageSimilarity.png), a picture of unknown’s chromatograms with assigned NMF-class determined by cosine distance (Unknowns_Append_withImageSimilarity.png) and folder named “NMFinput_Unknowns_eachpic”. In the folder of “NMFinput_Unknowns_eachpic”, all the NMF input and unknown’s GCxGC chromatograms with classification result, sample list used for NMF and unknown’s classification (“sampleIDlist.csv”, “unknownIDlist.csv”), are

included. Refer the ID number in the pictures of chromatogram to the produced "sampleIDlist.csv" and "unknownIDlist.csv".

9. If you want to perform other designed classification analysis, replace all the "Input" dataset for which you intend to classify. Then, re-start it from the step 3.

Supplementary:

RT shift correction for 2D chromatogram, which enhance the accuracy of the classification, is available from https://github.com/jsarey/GCxGC-alignment (Matlab free source code for univariate GCxGC data), https://github.com/GCxGC/GCxGC-MS-alignment (Matlab free source code for multivariate GCxGC data). Both of them are implemented in shareware of GC Image (≧R 2.6) http://www.gcimage.com/. See Reference 2) and 3).

Reference:

1) Zushi Y. and Hashimoto S., Direct classification of GC × GC-analyzed complex mixtures using non-negative matrix factorization based feature extraction, *Anal. Chem.* 2018, *In Press*.
2) 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.
3) 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.