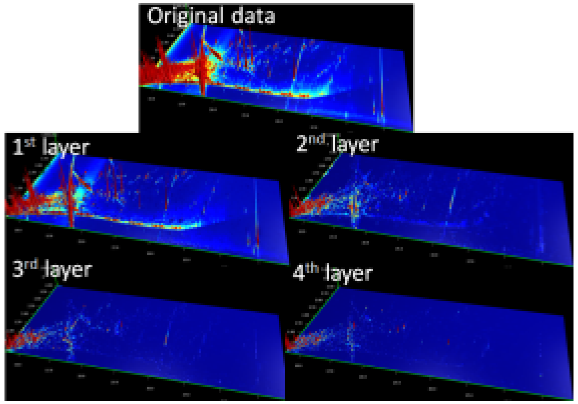


12th June, 2020

Operation manual for GUI-based NMFwithDBcreator

Developer
National Institute for Environmental Studies
Guest Scientist
Yasuyuki ZUSHI

Framework for NMFwithDBcreator

- NMFwithDBcreator involves a four step process and consists of the following three executable files:
- *I_NMFdeconvolution_expand.exe*
- *II_IV_LibrarySearch.exe*
- *III_DBCreator+.exe*
- Each of the files is executed in the corresponding step; then, the database file is created as the final output.
- Detailed explanations are given on the following pages.

Requirements for NMFwithDBcreator

- OS: Windows 10
- At least 10 GB RAM and a 64-bit OS are recommended for large amounts of data.
- Installation of “R” statistical software is required. Freely downloadable from:

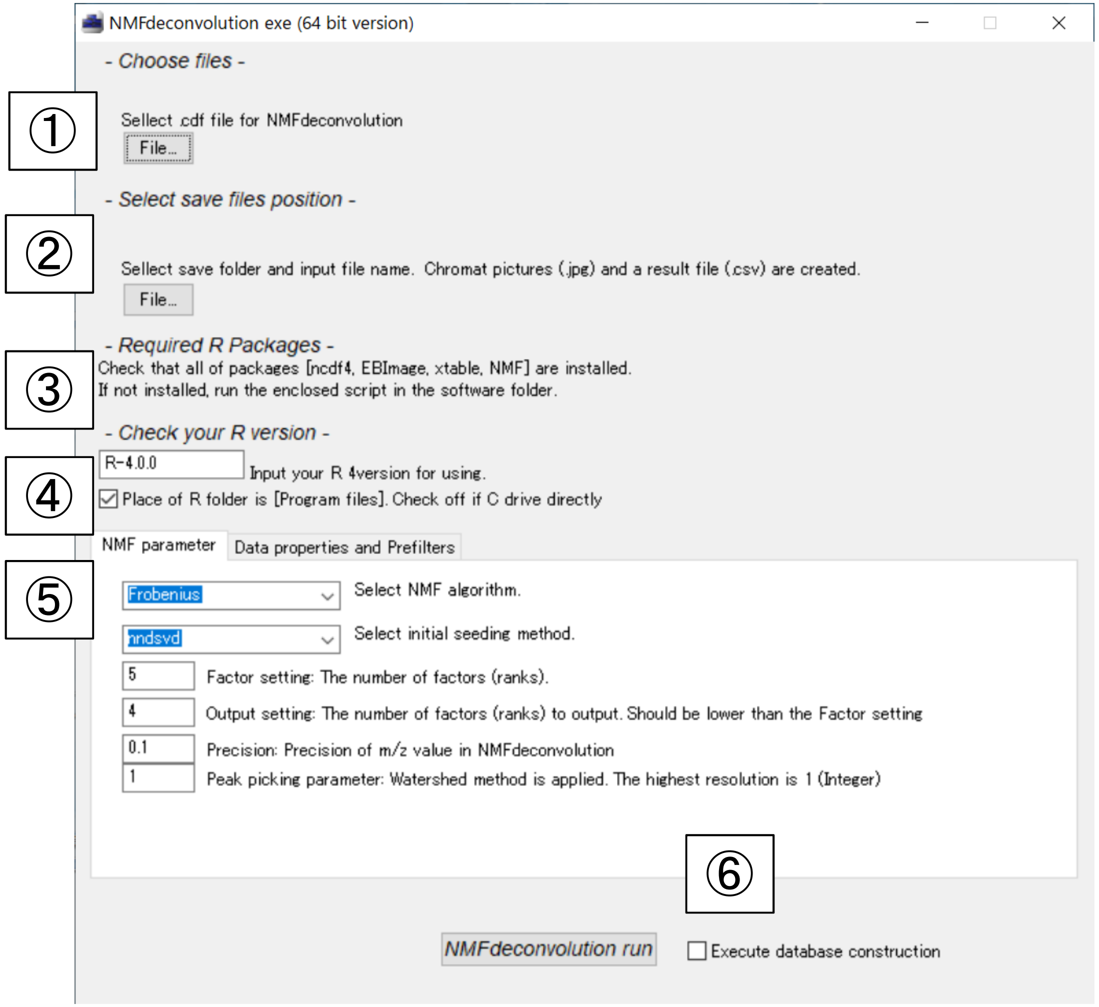
  https://cran.ism.ac.jp/
- netCDF data from GCxGC-HRTOFMS have been confirmed to work in this software.
- Do not touch the computer keyboard during a few seconds, because this tool automatically inserts the required information – there is no need for you to do anything. Please save and close important files before starting the tool.
- Demonstration data are included in this software folder. Also, it is available via the following link (Japanese site with English translation):

  https://www.nies.go.jp/analysis/downloads.html
- NIST MS Search software is required for “*II_IV_LibrarySearch.exe*” and “*III_DBCreator+.exe*”.

I_NMFdeconvolution_expand.exe

Step 1:
Spectral deconvolution is executed. Then, all the generated peaks, including the original peak with RTs and MS spectra, are listed in a csv file (optional). In addition, a file with the format for a library batch search in the NIST library is generated.

I_NMFdeconvolution_expand.exe
How to Execute (1)



NMFdeconvolution exe (64 bit version)

- Choose files -

Sellect .cdf file for NMFdeconvolution

File...

- Select save files position -

Sellect save folder and input file name. Chromat pictures (.jpg) and a result file (.csv) are created.

File...

- Required R Packages -

Check that all of packages [ncdf4, EBImage, xtable, NMF] are installed.
If not installed, run the enclosed script in the software folder.

- Check your R version -

R-4.0.0 Input your R 4version for using.

☑ Place of R folder is [Program files]. Check off if C drive directly

NMF parameter | Data properties and Prefilters

Frobenius — Select NMF algorithm.

nndsvd — Select initial seeding method.

5 — Factor setting: The number of factors (ranks).

4 — Output setting: The number of factors (ranks) to output. Should be lower than the Factor setting

0.1 — Precision: Precision of m/z value in NMFdeconvolution

1 — Peak picking parameter: Watershed method is applied. The highest resolution is 1 (Integer)

NMFdeconvolution run

☐ Execute database construction

① Choose your measurement data file (.cdf).

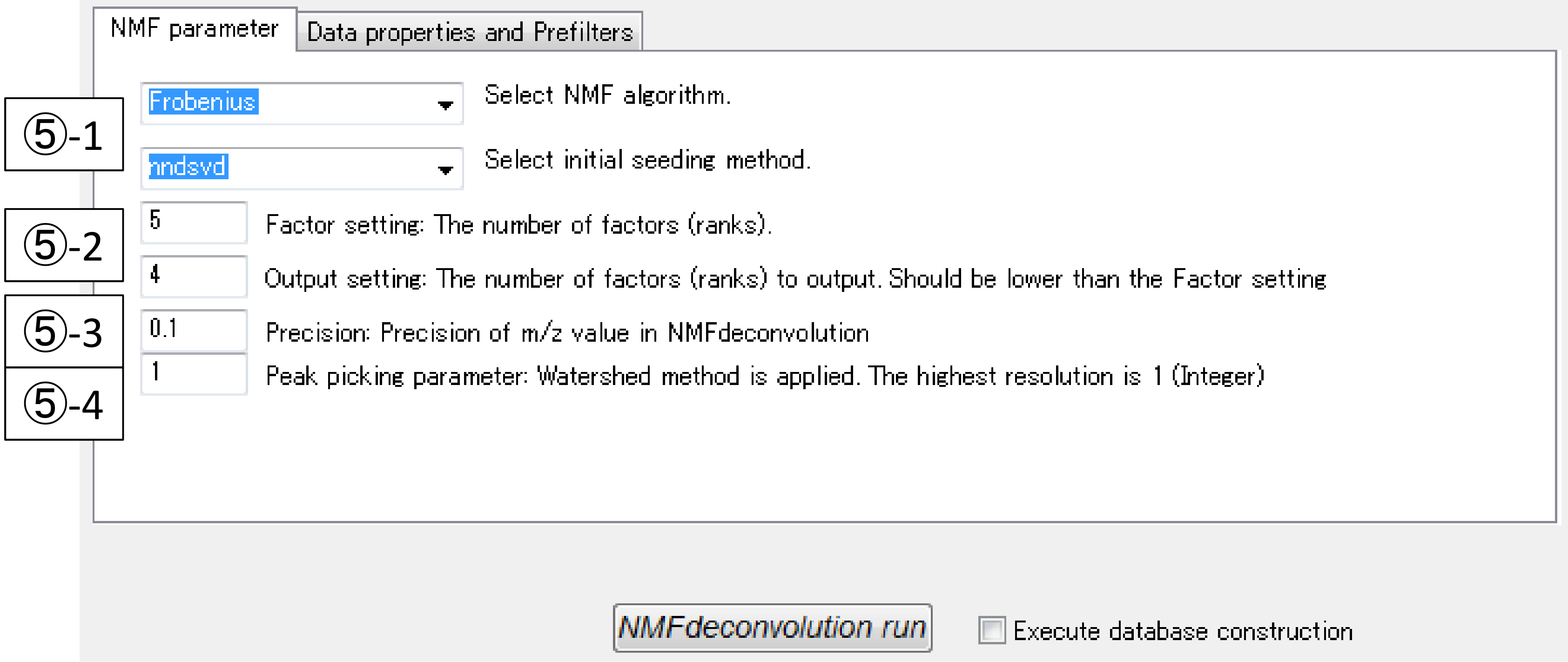
② Set an output folder and file name.

③ You need to install the required program packages. If not, install them by running a script in the provided file “Run_me_for_package_instllation.r”.

④ Set file path to run “R”. You need to know where and which version of “R” is installed on your computer. Is it located directly on the C drive or in “Program Files”?

⑤ and ⑥ See next two pages.

I_NMFdeconvolution_expand.exe
How to Execute (2)



Frobenius — Select NMF algorithm.

nndsvd — Select initial seeding method.

5 — Factor setting: The number of factors (ranks).

4 — Output setting: The number of factors (ranks) to output. Should be lower than the Factor setting

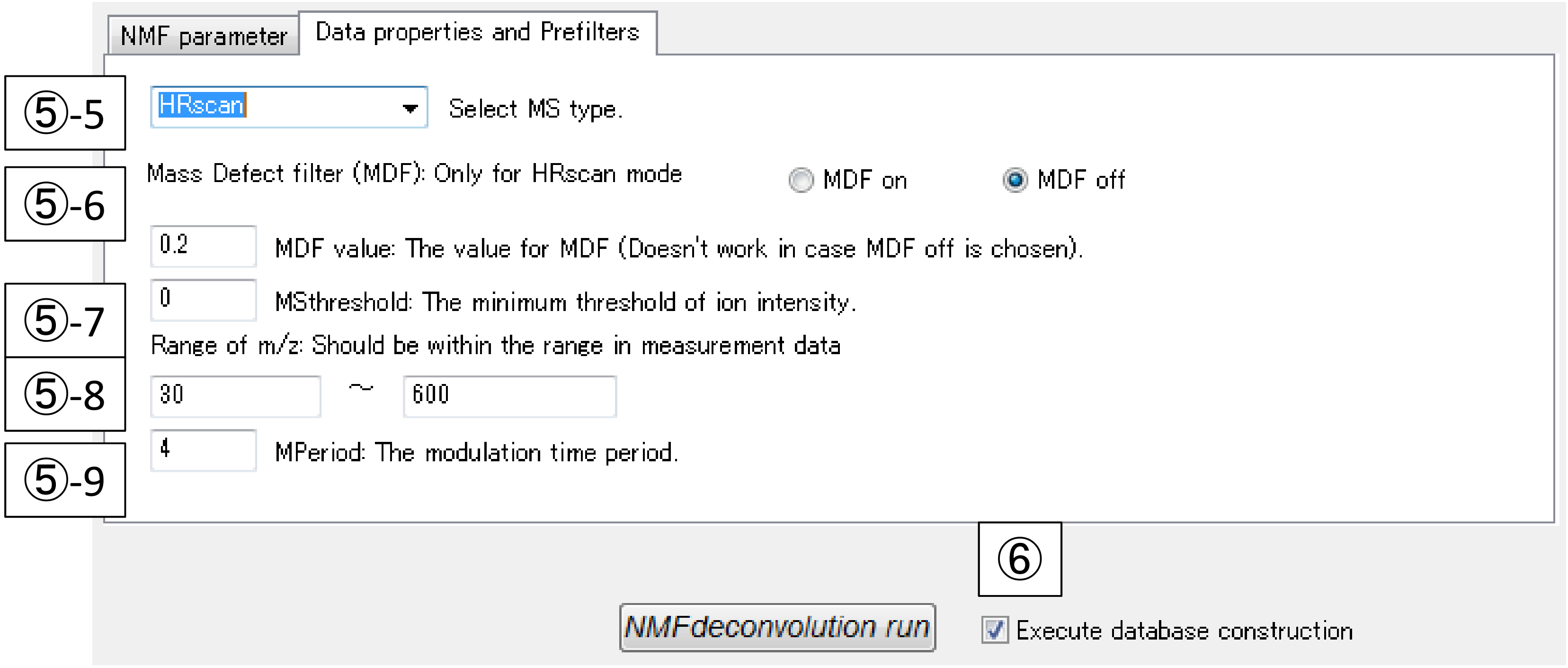
0.1 — Precision: Precision of m/z value in NMFdeconvolution

1 — Peak picking parameter: Watershed method is applied. The highest resolution is 1 (Integer)

☐ Execute database construction

- ⑤-1 Select NMF algorithm and initial seeding method. Frobenius and nndsvd method are the default settings.
- ⑤-2 Input the number of factors (ranks). “Output setting” describes the number of factors to output. Factors with a higher rank are chosen in descending order.
- ⑤-3 Set precision of m/z, used as the variance in NMF deconvolution: m/z = 100.1, 100.2, 100.3, … in the case of 0.1. Lower values have higher calculation cost.
- ⑤-4 Peak range is defined based on this parameter. Basically the value one should be chosen. If you want to expand the range for peaks, choose a larger value (must be an integer).

I_NMFdeconvolution_expand.exe
How to Execute (3)



HRscan ▾ Select MS type.

Mass Defect filter (MDF): Only for HRscan mode ○ MDF on ● MDF off

0.2 MDF value: The value for MDF (Doesn't work in case MDF off is chosen).

0 MSthreshold: The minimum threshold of ion intensity.

Range of m/z: Should be within the range in measurement data

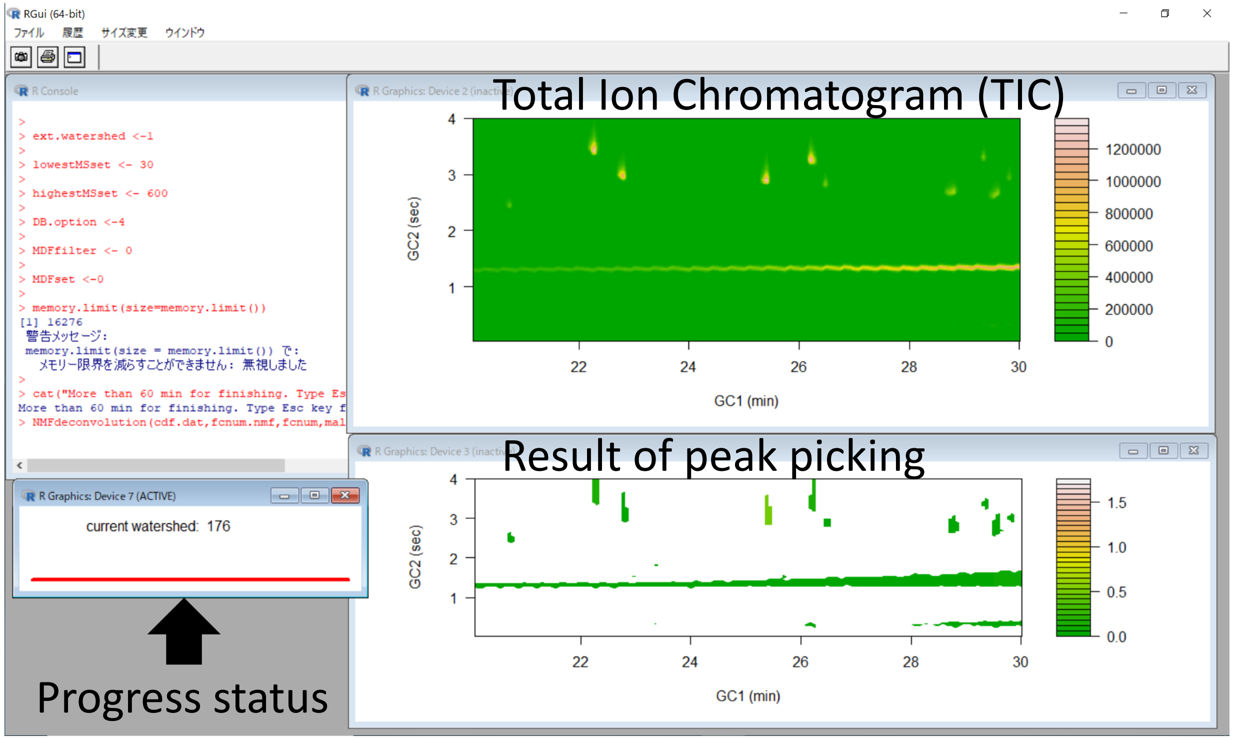
30 ~ 600

4 MPeriod: The modulation time period.

☑ Execute database construction

- ⑤-5 Select "HRscan" for data with high mass resolution. Select "scan" for nominal mass data. This parameter is related to MDF only. Usually, "HRscan" should be selected.
- ⑤-6 MDF setting as a pre-filter for the data. The next column describes the MDF value. See Hashimoto et al., 2013, J. Chromatogr. A 183-189 for details of MDF processing.
- ⑤-7 Threshold for ion intensity values is adjusted here.
- ⑤-8 Describes the range of m/z in NMF deconvolution. This range should be within the range of the measurement data.
- ⑤-9 Describes the modulation period (sec) in the measurement by GCxGC.
- ⑥ If checked, files of peak lists and NIST Search are generated in database construction.

Screenshot of Process



```
>
> ext.watershed <-1
>
> lowestMSset <- 30
>
> highestMSset <- 600
>
> DB.option <-4
>
> MDFfilter <- 0
>
> MDFset <-0
>
> memory.limit(size=memory.limit())
[1] 16276
 警告メッセージ:
 memory.limit(size = memory.limit()) で:
   メモリー限界を減らすことができません: 無視しました
>
> cat("More than 60 min for finishing. Type Es
More than 60 min for finishing. Type Esc key f
> NMFdeconvolution(cdf.dat,fcnum.nmf,fcnum,mal
```

```
current watershed:  176
```

Total Ion Chromatogram (TIC)

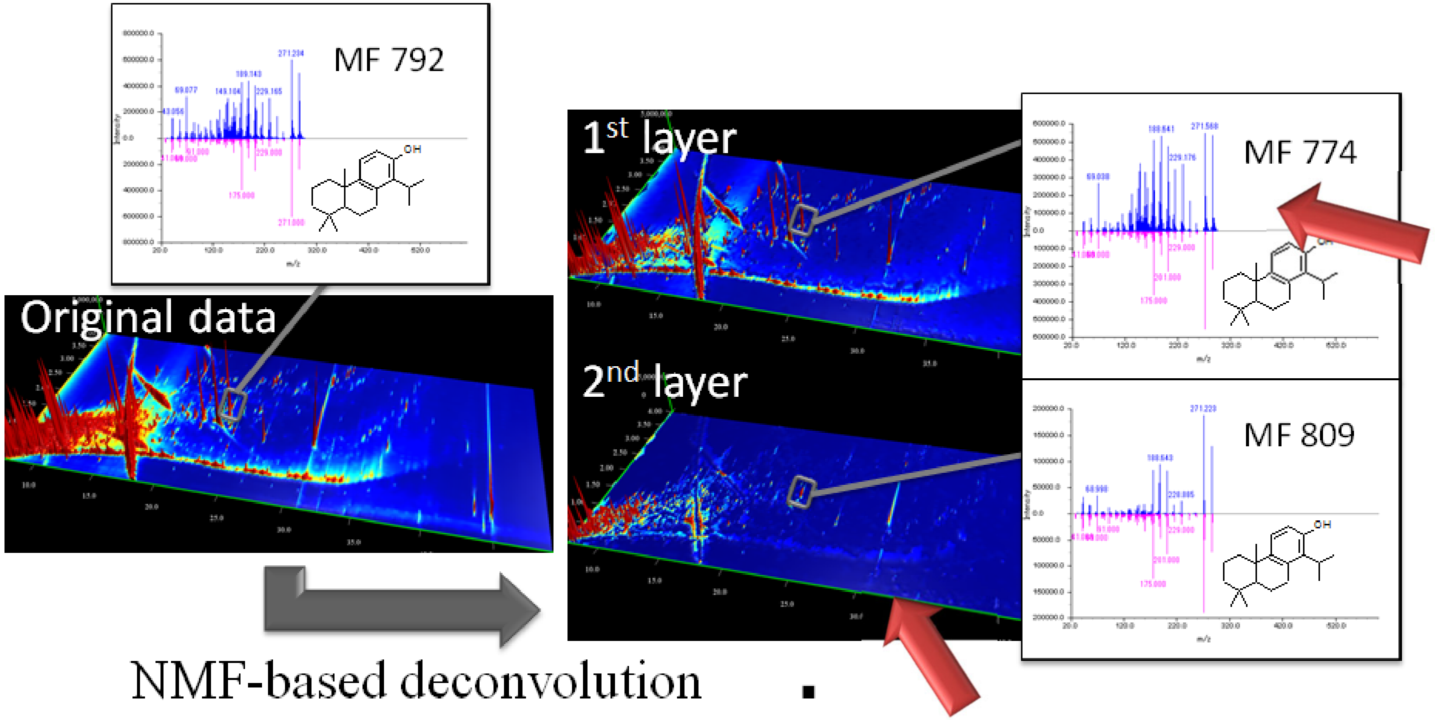
Result of peak picking

Progress status

- After clicking button “NMFdeconvolution run” with correct configuration, “R” is activated and starts to load the code according to the configuration.
- After loading all the code, peak deconvolution for each peak begins.
- Typically, it takes 1 to 5 seconds for a peak, and an hour for an entire chromatogram of size 100 MB to be deconvoluted.
- The calculation cost increases according to the precision of m/z. It takes an hour to generate three deconvoluted layers with precision m/z = 1, an hour to generate a layer with precision m/z = 0.1, and 2~3 hours to generate a layer with precision m/z = 0.05.
- A CDF file is generated in the save folder as the output of the deconvolution.
- In the following process, files of peak lists and NIST Search are generated if ⑥ is checked.

Output File (1)

The generated CDF file can be opened with the software GCImage.
Other several free tools, such as R, is available.
Easily, the following site provides a web-application to visualize GCxGC data.
GCxGC Mixture Touch:
http://www.mixture-platform.net/Mixture_Touch_open/



NMF-based deconvolution

$$Y \approx WH$$

$$\min_{W,H \geq 0} [D(Y, WH) + R(W, H)]$$

Deconvoluted spectra are stored in each layer.

*Higher intensity in peak top within the deconvoluted peak is stored in upper layer.

The layer is output according to the “output setting”.

Output File (2)

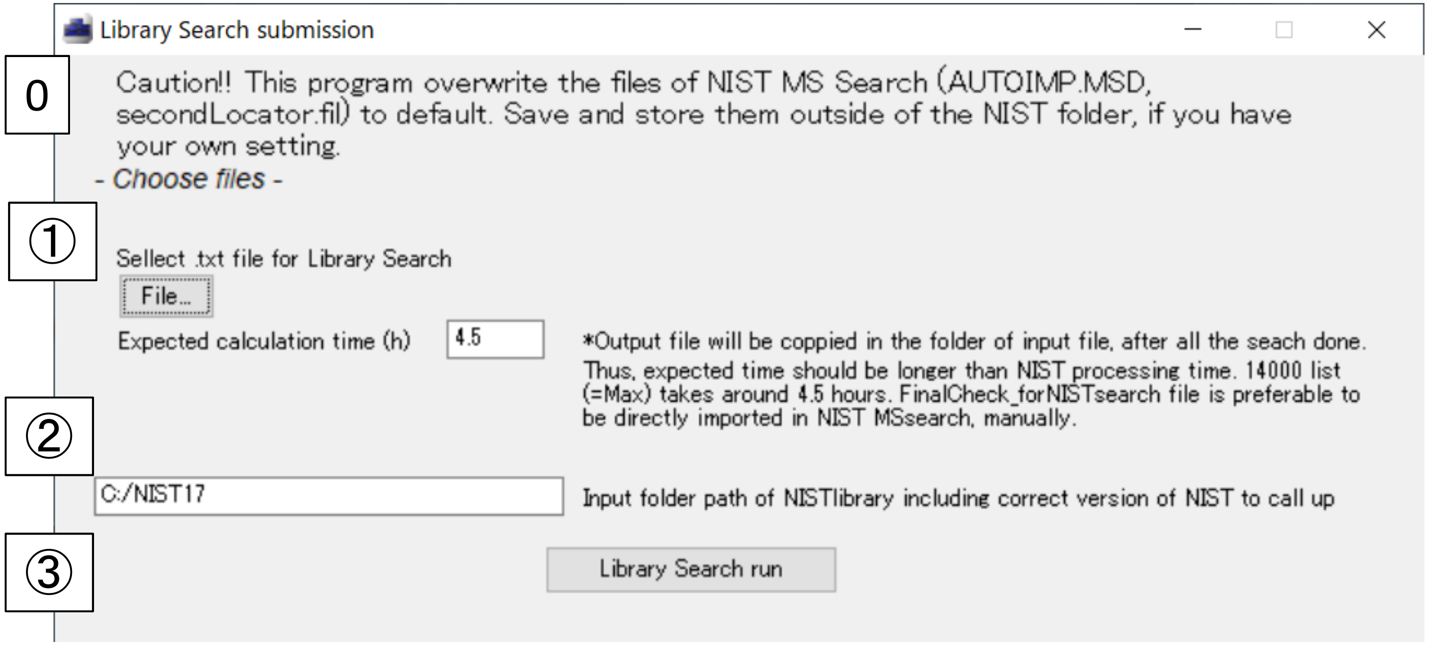
- A number of output files are generated depending on the setting for “output setting”.
- The name of the original cdf file is set to “**filename**_*layer0.cdf*”.
- The name of the deconvoluted file for the first layer is set to “**filename**_ *layer1.cdf*”.
- If ⑥ is checked, files of peak lists and NIST Search are generated following the peak picking process, after the deconvolution process. This process takes a few minutes.
- Both
  “*MSpeaklist_*__filename__*_layer0.csv*” and
  “*MSpeaklist_forNISTsearch_*__filename__*_layer0.txt*”
  are generated in the process.
- Furthermore,
  “*Combined_MSpeaklist_*__filename__*.csv*” and
  “*Combined_MSpeaklist_forNISTsearch_*__filename__*.txt*” form the combined list of information on peaks in each layer.

II_IV_LibrarySearch.exe

Step 2 and Step 4:
The file for batch search in NIST library is loaded and the result is output.
Step 4 is provided merely as a final check of extracted peaks after the database construction in step 3.

＊Step 4 using this exe file is not absolutely essential. The simplest way is to manually open and load the file for batch search in the NIST MS Search software.

II_IV_LibrarySearch.exe
How to Execute (1)



Caution!! This program overwrite the files of NIST MS Search (AUTOIMP.MSD, secondLocator.fil) to default. Save and store them outside of the NIST folder, if you have your own setting.

- *Choose files* -

Sellect .txt file for Library Search

Expected calculation time (h) 4.5

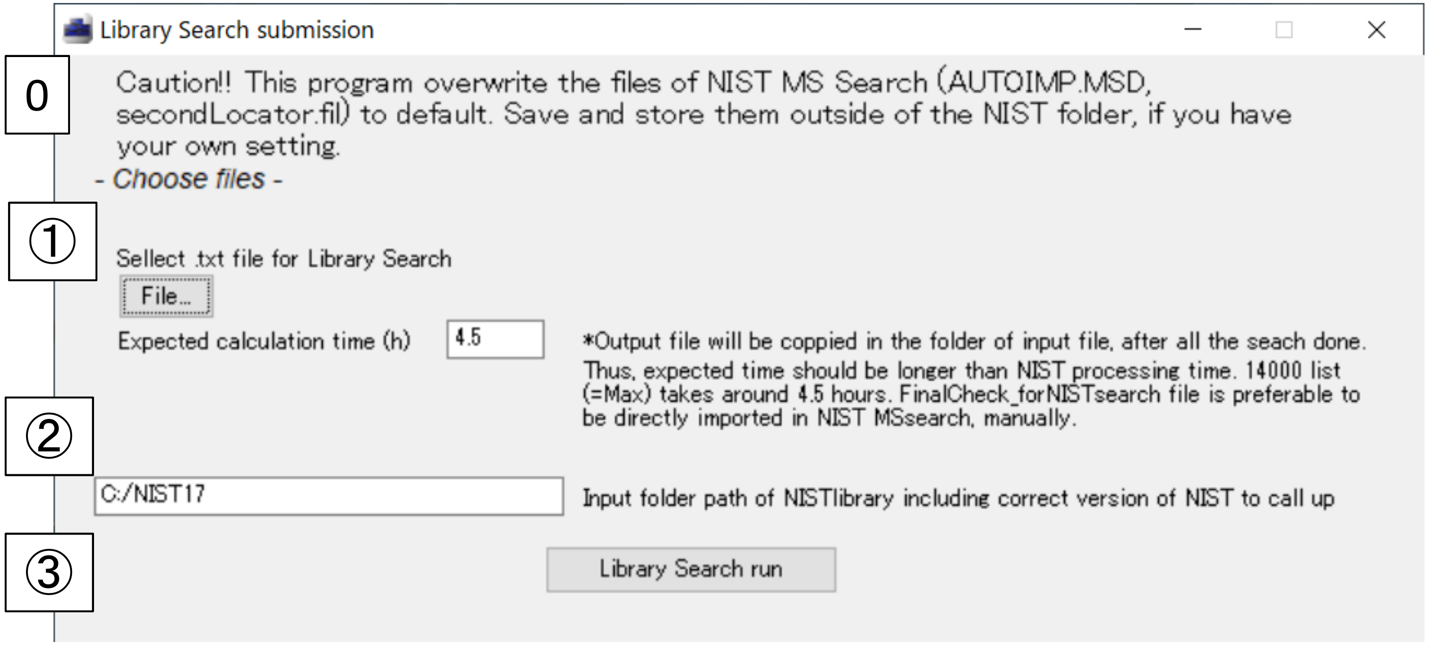
*Output file will be coppied in the folder of input file, after all the seach done. Thus, expected time should be longer than NIST processing time. 14000 list (=Max) takes around 4.5 hours. FinalCheck_forNISTsearch file is preferable to be directly imported in NIST MSsearch, manually.

C:/NIST17

Input folder path of NISTlibrary including correct version of NIST to call up

- 0 The following two files are automatically placed in the folder for the NIST library program, prior to starting the search:
  “*AUTOIMP.MSD*” and “*secondLocator.fil*”
- Caution!! Original files are stored in the location “*C:/NIST*version*/MSSEARCH*”; please move the original files to a different location, if you have your own search setting.
- ① Select a file to execute batch search in the NIST library. The file name should be ”*Combined_MSpeaklist_forNISTsearch_*filename*.txt*”, if the user has not changed it.

II_IV_LibrarySearch.exe
How to Execute (2)



Caution!! This program overwrite the files of NIST MS Search (AUTOIMP.MSD, secondLocator.fil) to default. Save and store them outside of the NIST folder, if you have your own setting.

- *Choose files* -

Sellect .txt file for Library Search

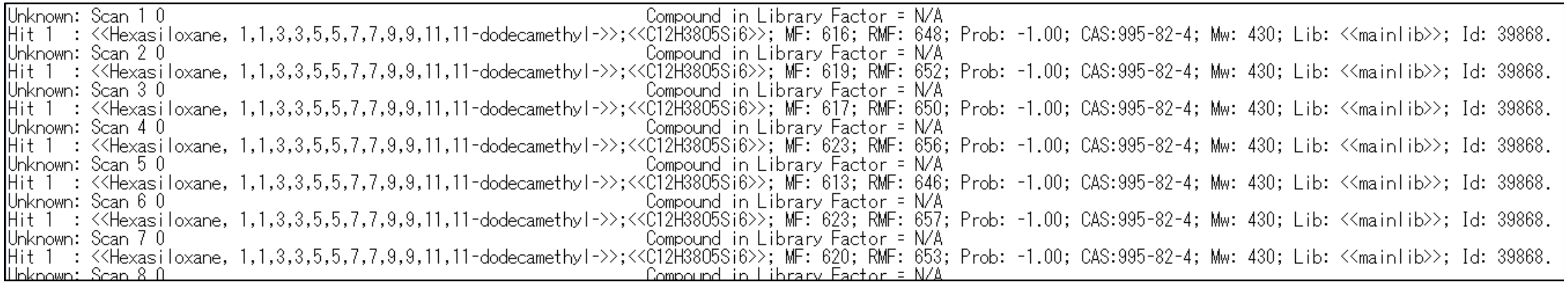
Expected calculation time (h) 4.5

*Output file will be coppied in the folder of input file, after all the seach done. Thus, expected time should be longer than NIST processing time. 14000 list (=Max) takes around 4.5 hours. FinalCheck_forNISTsearch file is preferable to be directly imported in NIST MSsearch, manually.

C:/NIST17 Input folder path of NISTlibrary including correct version of NIST to call up

- ② Input an estimated required time to complete NIST Search. Owing to the process of automatically moving an output file after completion of NIST Search, the required time should be estimated with plenty of margin; e.g., it generally takes 4 hours for 14000 entries, and thus 4.5 hours is recommended as input.
- ③ Input the folder path of the NIST library to use in the search. The path depends on which version of the NIST library is used.
- All parameter settings in the NIST library should be completed before execution. For the parameter settings, run the NIST MS Search software and select “library search option”. The recommended settings are “similarity” with “simple search” in the search mode, and one for “Number of hits to print”.

Output File



```
Unknown: Scan 1 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 616; RMF: 648; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 2 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 619; RMF: 652; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 3 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 617; RMF: 650; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 4 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 623; RMF: 656; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 5 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 613; RMF: 646; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 6 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 623; RMF: 657; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 7 0                                                    Compound in Library Factor = N/A
Hit 1  : <<Hexasiloxane, 1,1,3,3,5,5,7,7,9,9,11,11-dodecamethyl->>;<<C12H38O5Si6>>; MF: 620; RMF: 653; Prob: -1.00; CAS:995-82-4; Mw: 430; Lib: <<mainlib>>; Id: 39868.
Unknown: Scan 8 0                                                    Compound in Library Factor = N/A
```

An example of the output

- A text file describing the results of the NIST Search of all peaks is the output.
- Search results with a number matching the setting “Number of hits to print” are assigned to a peak.
- In the case of the upper figure, “Number of hits to print” is set as one, and therefore, the search result is recorded in the output file.
- The output file name is set to: “Combined_MSpeaklist_forNISTsearch_**filename**.txtOutputLibrarySearch.txt”.

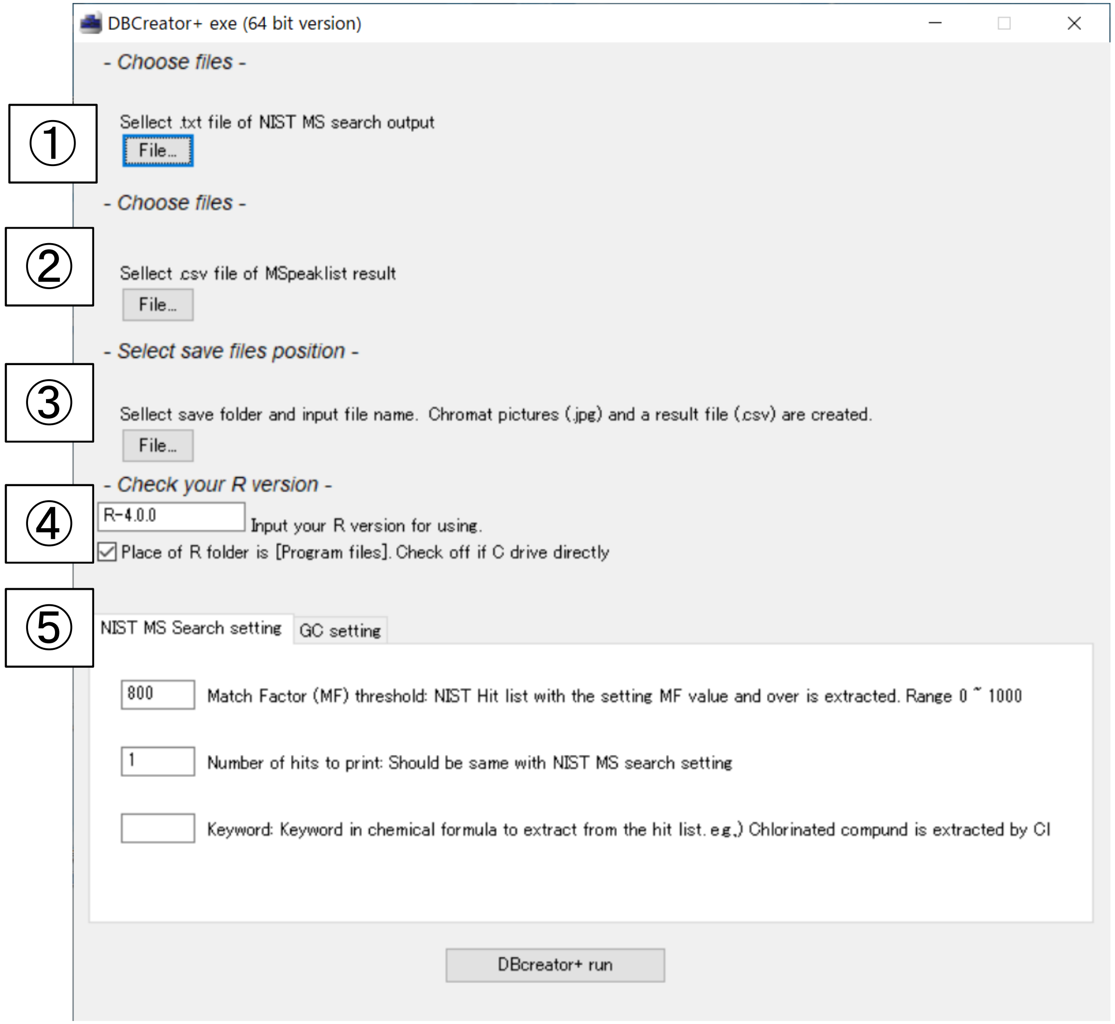
III_DBCreator+.exe

Step 3:

The list of all peaks obtained by deconvolution is extracted under certain conditions in this step.

Through this process, the database file and the file for the final check of spectra in the NIST library are generated simultaneously.

III_DBCreator+.exe
How to Execute (1)



- Choose files -

Sellect .txt file of NIST MS search output

- Choose files -

Sellect .csv file of MSpeaklist result

- Select save files position -

Sellect save folder and input file name. Chromat pictures (.jpg) and a result file (.csv) are created.

- Check your R version -

R-4.0.0 Input your R version for using.

☑ Place of R folder is [Program files]. Check off if C drive directly

800 Match Factor (MF) threshold: NIST Hit list with the setting MF value and over is extracted. Range 0 ~ 1000

1 Number of hits to print: Should be same with NIST MS search setting

Keyword: Keyword in chemical formula to extract from the hit list. e.g.) Chlorinated compund is extracted by Cl

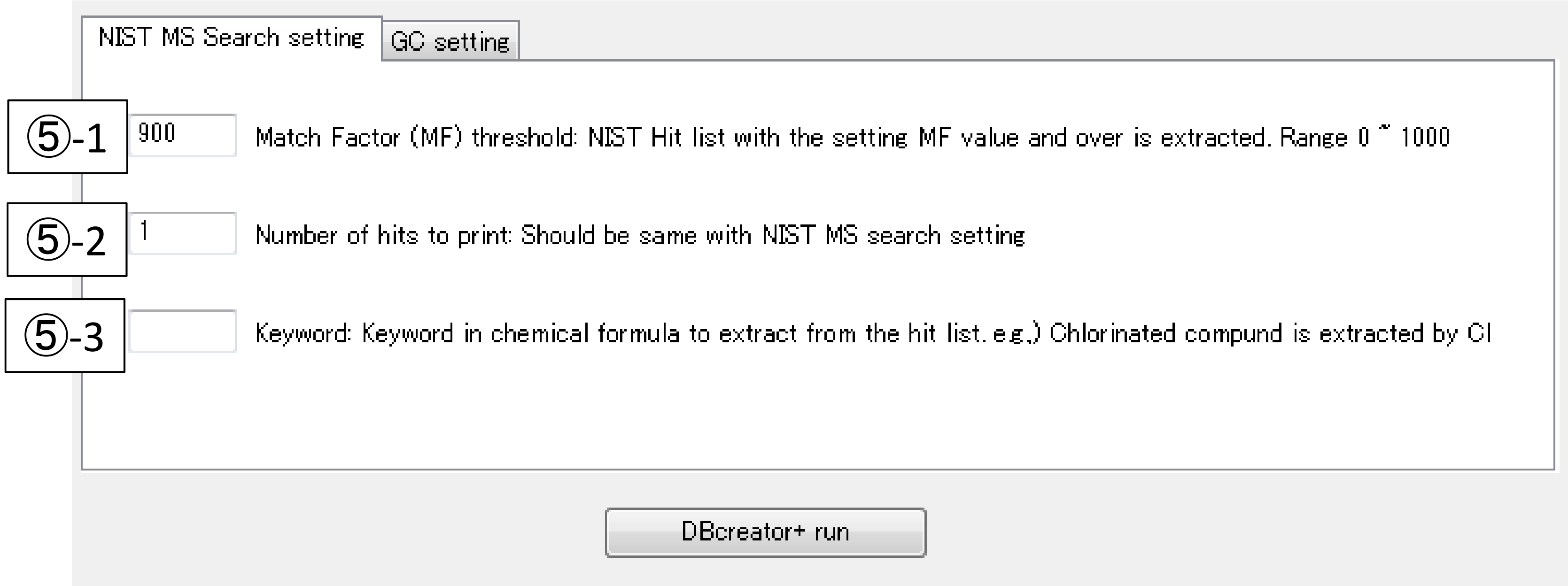
① Select a txt file for the NIST search result. The name of the file is "Combined_MSpeaklist_forNISTsearch_**filename**.txtOutputLibrarySearch.txt", if no changes have been made to the file name in the previous step.

② Select a csv file with the peak list; default name is "Combined_MSpeaklist_**filename**.csv".

③ Choose save file location and input save file name.

④ Select version and location of software R.

III_DBCreator+.exe
How to Execute (2)



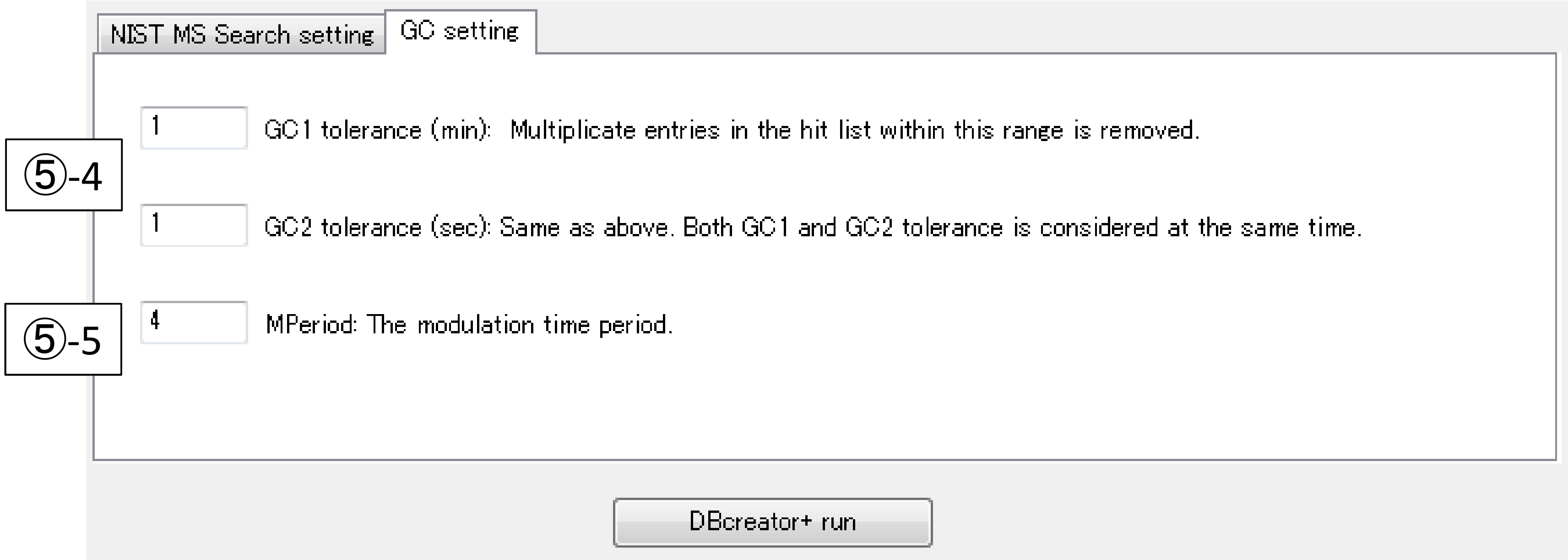
⑤-1 900 Match Factor (MF) threshold: NIST Hit list with the setting MF value and over is extracted. Range 0 ~ 1000

⑤-2 1 Number of hits to print: Should be same with NIST MS search setting

⑤-3 Keyword: Keyword in chemical formula to extract from the hit list. e.g.) Chlorinated compund is extracted by Cl

- ⑤-1 Lower threshold value of MF to extract from the peak list.
- ⑤-2 "Number of hits to print", which is selected in "Library Search Option" in NIST MS Search.
- ⑤-3 Entries containing this keyword in the assigned formula are extracted from the list. If "Cl" is input, entries with "Cl" in their formula and at the same time, greater than the MF threshold, are extracted.

III_DBCreator+.exe
How to Execute (3)



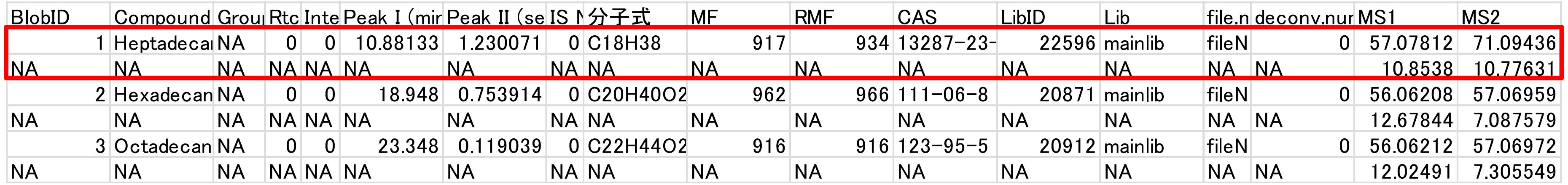
1 GC1 tolerance (min): Multiplicate entries in the hit list within this range is removed.

1 GC2 tolerance (sec): Same as above. Both GC1 and GC2 tolerance is considered at the same time.

4 MPeriod: The modulation time period.

- ⑤-4 Input tolerance of retention time in GC1 and GC2.
- If 1 min and 1 sec are chosen, entries in the same assignment within this tolerance are regard as duplicates. Thus, the entries are deleted except for the entry with the highest MF.
- ⑤-5 Input modulation time period applied in the measurement.

Output File



BlobID	Compound	Grou	Rtc	Inte	Peak I (mir	Peak II (se	IS	分子式	MF	RMF	CAS	LibID	Lib	file.n	deconv.nur	MS1	MS2
1	Heptadeca	NA	0	0	10.88133	1.230071	0	C18H38	917	934	13287-23-	22596	mainlib	fileN	0	57.07812	71.09436
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	10.8538	10.77631
2	Hexadecan	NA	0	0	18.948	0.753914	0	C20H40O2	962	966	111-06-8	20871	mainlib	fileN	0	56.06208	57.06959
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	12.67844	7.087579
3	Octadecan	NA	0	0	23.348	0.119039	0	C22H44O2	916	916	123-95-5	20912	mainlib	fileN	0	56.06212	57.06972
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	12.02491	7.305549

…

Part of the list in database file

Information on a peak is summarized across two rows

The following three output files are generated.

1. The list of extracted peaks based on certain conditions. This file consists of meta data such as compound name, RT, formula. The file name is set to “SimplePeaklist_condition**Keyword**_MF_**Mfvaluefilename**.csv”.
2. This file comprises a compound database that is used by the TSEN program. It includes MS spectra in addition to file (1), and the file name is “Database_condition**Keyword**_MF_**MFvaluefilename**.csv”.
3. This file displays the compound spectra in the final list of the database. The file name is set to “FinalCheck_forNISTsearchDatabase_condition**Keyword**_MF_**MFvaluefilename**.csv.txt”. The MS spectra in the final list are visually confirmed point-by-point by loading this file in NIST MS Search.

Precautions

- About the source file

- Keep the source files “1NMFdeconvolution.r”, “2MSpeaklist.r”, “3FinalDBcreate.r” and “4DBfinalCheckInNIST.r” in the same directory as the .exe files “I_NMFdeconvolution_expand_en.exe”, “II_IV_LibrarySearch.exe”, and “III_DBCreator+_en.exe”.

- If the software freezes, you can exit by clicking the icon at the bottom right of the screen.

- License: Artistic License 2.0

- Disclaimer: The developer is not responsible for any hardware or software damages, data losses or false inferences caused by the use of this software.