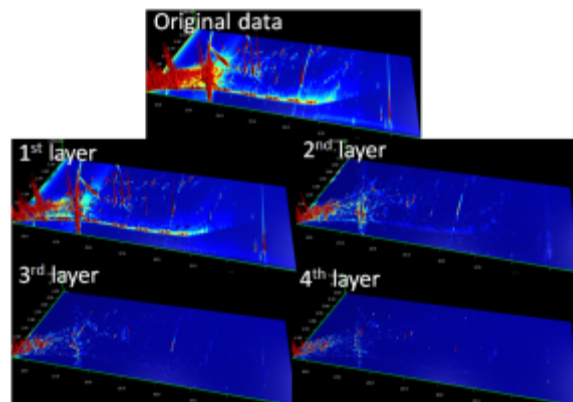


2020.06.12



# NMFwithDBcreatorツール(GUI version)の 利用マニュアル

作成  
国立環境研究所  
客員研究員  
頭士 泰之

## NMFwithDBcreatorの構成について

- NMFwithDBcreatorは全部で下記の3つのexeファイルと4つのステップから成ります。
- *I\_NMFdeconvolution\_expand.exe*
- *II\_IV\_LibrarySearch.exe*
- *III\_DBCreator+.exe*
- 各Stepごとに順次対応するexeファイルを実行し、最終的なデータベースファイルを得ます。
- 詳細について、次ページ以降で説明します。

# NMFwithDBcreatorの動作環境について

- Windows 10のOSで動作確認済みです。
- 他社OSでは正しく動作しません。
- データ解析ソフト「R」のインストールが必要です。下記サイトからフリーダウンロード可能です。
- <https://cran.ism.ac.jp/>
- GCxGC-HRTOFMSのnetCDFデータに対して動作を確認済みです。
- NISTライブラリサーチには、別途NIST MS Searchソフトウェアが必要です。そのため「II\_IV\_LibrarySearch.exe」と「III\_DBCreator+.exe」については、NIST MS Searchが無い場合、動作対象外です。
- 本ツールはAutoKey技術を利用しており、PCはプログラム実行中、自動キー操作されます。本プログラムによりR画面にスクリプトが読み込まれているタイミングでは、PC画面に触れないようにしてください。スクリプトが乱れて実行エラーとなる可能性があります。大事なファイル等も保存して、閉じてから本ツールを実行してください。

# NMFwithDBcreatorの動作環境について

- 配布フォルダ中の「I\_NMFdeconvolution\_expand.exe」では大規模計算が実行されます。このため大容量の物理メモリ(推奨は32GB以上)が必要です。
- メモリ使用量の関係上、基本的には64bitOS環境下のみでしか動作しません。
- GCxGC-HRTOFMSのデモ用データは、同梱されているものを利用可能です。もしくは下記サイトからダウンロード可能です。
- <https://www.nies.go.jp/analysis/downloads.html>

## *I\_NMFdeconvolution\_expand.exe*

ステップ1の実行ファイルです。  
対象のGCxGCデータに対し、デコンボリューションを行ないます。  
その後、生成したピーク情報をデータベース用にまとめNISTライブラリのバッチ検索用ファイルを生成します(オプション)。

# I\_NMFdeconvolution\_expand.exe 起動画面と操作方法について

The screenshot shows the NMFdeconvolution exe (64 bit version) window. It has a title bar with standard Windows window controls. The interface is divided into several sections:

- Choose files -**: Contains a label "Select .cdf file for NMFdeconvolution" and a "File..." button. This is annotated with a circled 1.
- Select save files position -**: Contains a label "Select save folder and input file name. Chromat pictures (.jpg) and a result file (.csv) are created." and a "File..." button. This is annotated with a circled 2.
- Required R Packages -**: Contains text: "Check that all of packages [ncdf4, EBImage, xtable, NMF] are installed. If not installed, run the enclosed script in the software folder." This is annotated with a circled 3.
- Check your R version -**: Contains a text input field with "R-4.0.0" and a label "Input your R 4version for using." This is annotated with a circled 4.
- NMF parameter**: A tabbed interface with "Data properties and Prefilters" selected. It contains:
  - A dropdown menu for "Select NMF algorithm." with "Frobenius" selected. This is annotated with a circled 5.
  - A dropdown menu for "Select initial seeding method." with "nndsvd" selected.
  - A text input field for "Factor setting: The number of factors (ranks)." with "5" entered.
  - A text input field for "Output setting: The number of factors (ranks) to output. Should be lower than the Factor setting" with "4" entered.
  - A text input field for "Precision: Precision of m/z value in NMFdeconvolution" with "0.1" entered.
  - A text input field for "Peak picking parameter: Watershed method is applied. The highest resolution is 1 (Integer)" with "1" entered.
- Buttons**: At the bottom, there is an "NMFdeconvolution run" button and a checkbox labeled "Execute database construction". The "run" button is annotated with a circled 6.

- ① 解析するデータ(CDF)を選びます。
- ② 解析結果を出力するフォルダとファイル名を設定します。
- ③ 初回の場合、必要なパッケージインストールのため、同梱のRスクリプト「Run\_me\_for\_package\_instllation.r」を事前に実行しておいて下さい。
- ④ 「R」を呼び出すためのパス設定をします。インストールされたRのバージョンを入力し、「R」フォルダがCドライブ直下にあるか、それ以外かを指定してください。
- ⑤&⑥ 次ページをご参照ください。

# *I\_NMFdeconvolution\_expand.exe*

## 起動画面と操作方法について

The screenshot shows the 'NMF parameter' tab of the *I\_NMFdeconvolution\_expand.exe* application. The interface is divided into two main sections: 'NMF parameter' and 'Data properties and Prefilters'. The 'NMF parameter' section contains the following settings:

- ⑤-1** Select NMF algorithm: Frobenius (dropdown menu)
- ⑤-2** Select initial seeding method: nndsvd (dropdown menu)
- ⑤-3** Factor setting: The number of factors (ranks): 5 (input field)
- ⑤-4** Output setting: The number of factors (ranks) to output. Should be lower than the Factor setting: 4 (input field)
- Precision: Precision of m/z value in NMFdeconvolution: 0.1 (input field)
- Peak picking parameter: Watershed method is applied. The highest resolution is 1 (Integer): 1 (input field)

At the bottom of the interface, there is a button labeled 'NMFdeconvolution run' and a checkbox labeled 'Execute database construction'.

- ⑤-1 NMFのアルゴリズムと初期値発生法を指定します。デフォルトはFrobenius, nndsvdとしています。
- ⑤-2 因子数(ランク数)を設定します。この内上位いくつを出力するか指定します。2つとも「0」を入力するとデコンボリューションは行われず、オリジナルデータのPeak pickingのみ行われピーク情報が出力されます(デコンボリューションの有無の比較用)。
- ⑤-3 NMFデコンボリューションで変数として扱われるm/zの、値刻み幅を設定します(0.1の場合、m/z 100.1, 100.2のようになる)。小さな値程、計算負荷がかかります。
- ⑤-4 ピークを抽出する際の、ピーク範囲の取り方を決めます。基本的には1を用います。範囲を広げたい場合大きな値(整数)とします。

# *I\_NMFdeconvolution\_expand.exe*

## 起動画面と操作方法について

The screenshot shows the 'NMF parameter' tab of the software interface. It contains several input fields and checkboxes for configuring the deconvolution process. Numbered callouts are placed over the interface to guide the user:

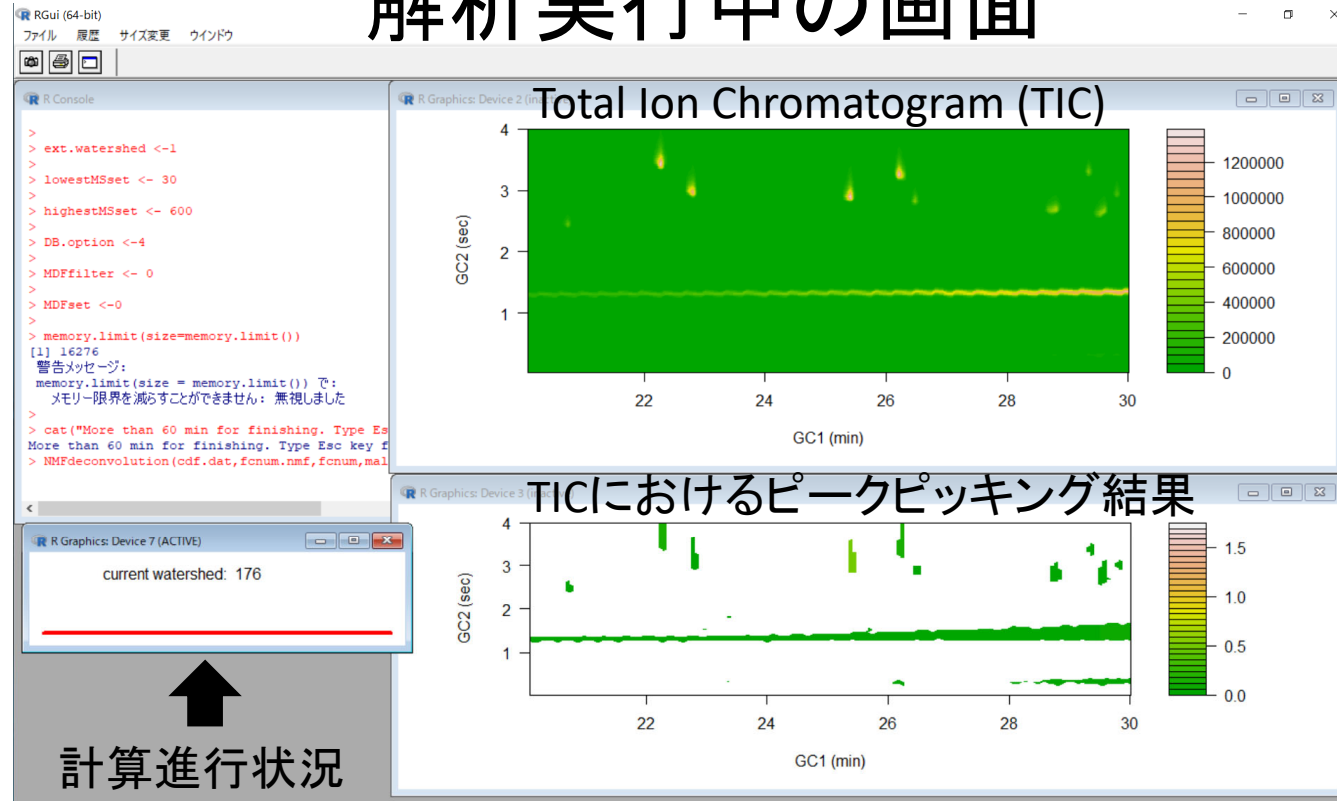
- ⑤-5** points to the 'MS type' dropdown menu, which is currently set to 'HRscan'.
- ⑤-6** points to the 'MDF value' input field, which is set to '0.2'.
- ⑤-7** points to the 'MStreshold' input field, which is set to '0'.
- ⑤-8** points to the 'Range of m/z' input fields, which are set to '30' and '600'.
- ⑤-9** points to the 'MPeriod' input field, which is set to '4'.
- ⑥** points to the 'Execute database construction' checkbox, which is checked.

At the bottom of the interface, there is a button labeled 'NMFdeconvolution run' and a checkbox labeled 'Execute database construction'.

- ⑤-5 質量データを高分解能として扱う場合「HRscan」、整数質量として扱う場合は「scan」を選択します。単にMDFに関わる設定で、基本的にはHRscanを選択してください。
- ⑤-6 データ前処理としてMDFを利用するかどうかを設定します。下の欄でMDFの値を設定してください。MDFについては [Hashimoto et al., J. Chromatogr. A, 2013, 183-189.](#) を参照してください。
- ⑤-7 イオン強度の閾値カットオフを設定できます。
- ⑤-8 NMFデコンボリューションで用いるm/zの範囲を設定します。装置の測定範囲内で設定してください。
- ⑤-9 GCxGCの折り返し時間であるモジュレーション時間を設定してください。
- ⑥ デコンボリューションに続きデータベース作成プロセスを行う場合チェックを付けて下さい。



# 解析実行中の画面



- 正しく設定が行われると、「R」が起動し、設定に従ったコードが読み込まれます。
- 「R」に全てコードが渡されると、各ピークのデコンボリューションが始まります。
- 100MB程度のデータサイズでは1ピークのデコンボリューションに1～5秒程度かかり、1つのレイヤの出力に1時間程度かかります。m/zの値を小さく設定すると計算時間が長くなります。1の設定では3つのレイヤに1時間、0.1では1つのレイヤに1時間、0.05の設定では1つのレイヤに2～3時間かかります。
- 解析終了後、保存先に設定したフォルダに、デコンボリューションされたcdfファイルが各レイヤごとに作成・保存されます。
- その後、各レイヤの全ピークの情報がリストにまとめられ、csvファイルとして生成されます。

# 出力ファイルについて

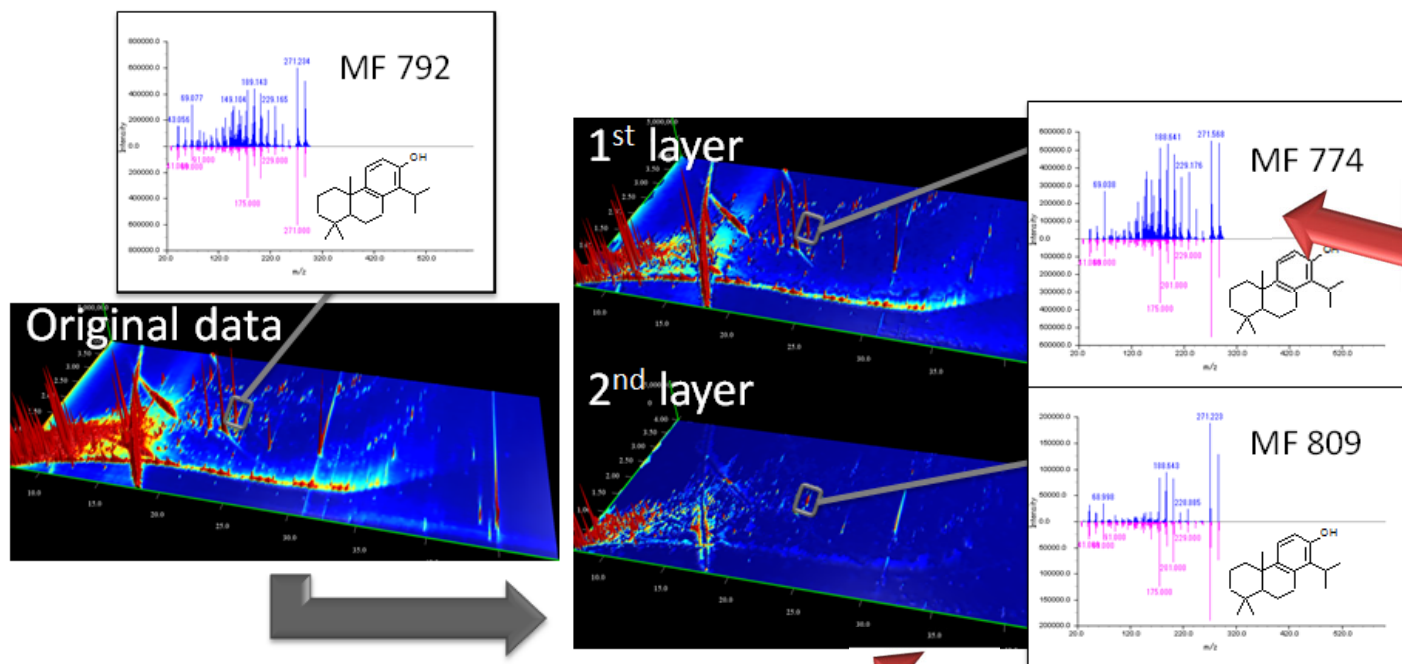
生成されるCDFファイルはGCIImageなどで閲覧できます。

Rなどでも解析できるツールがあります。

手軽な方法として、下記のWebサイトでGCxGCデータを閲覧・解析できます。

GCxGC Mixture Touch:

[http://www.mixture-platform.net/Mixture\\_Touch\\_open/](http://www.mixture-platform.net/Mixture_Touch_open/)



各レイヤに、分離後のスペクトルが保存されています。

ピークトップのTIC値が高いものが上位レイヤに保存されます。

NMF-based deconvolution

$$Y \approx WH$$

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

指定した数だけレイヤのファイルが出力されます。

# 出力ファイルについて

- 前ページの出力ファイルについて、デコンボリューションではOutput settingで指定した数のファイル(レイヤ)が出力されます。オリジナルデータは
- 「ファイル名\_layer0.cdf」
- として保存され、デコンボリューション後のレイヤ1は「ファイル名\_layer1.cdf」のようになります。
- Database Constructionを選択実行した場合、デコンボリューション処理に引き続いて、各レイヤに対するピークピッキングが行われ、全ピークの情報抽出が行われます。このプロセスは数分程度で終わります。
- 出力として、csv形式の
- 「MSpeaklist\_ファイル名\_layer0.csv」とtxt形式の
- 「MSpeaklist\_forNISTsearch\_ファイル名\_layer0.txt」
- の2種のファイルが生成されます。
- さらに上記の各レイヤ出力ファイルを纏めた
- 「Combined\_MSpeaklist\_ファイル名.csv」と
- 「Combined\_MSpeaklist\_forNISTsearch\_ファイル名.txt」
- の2つも生成されます。

## *II\_IV\_LibrarySearch.exe*

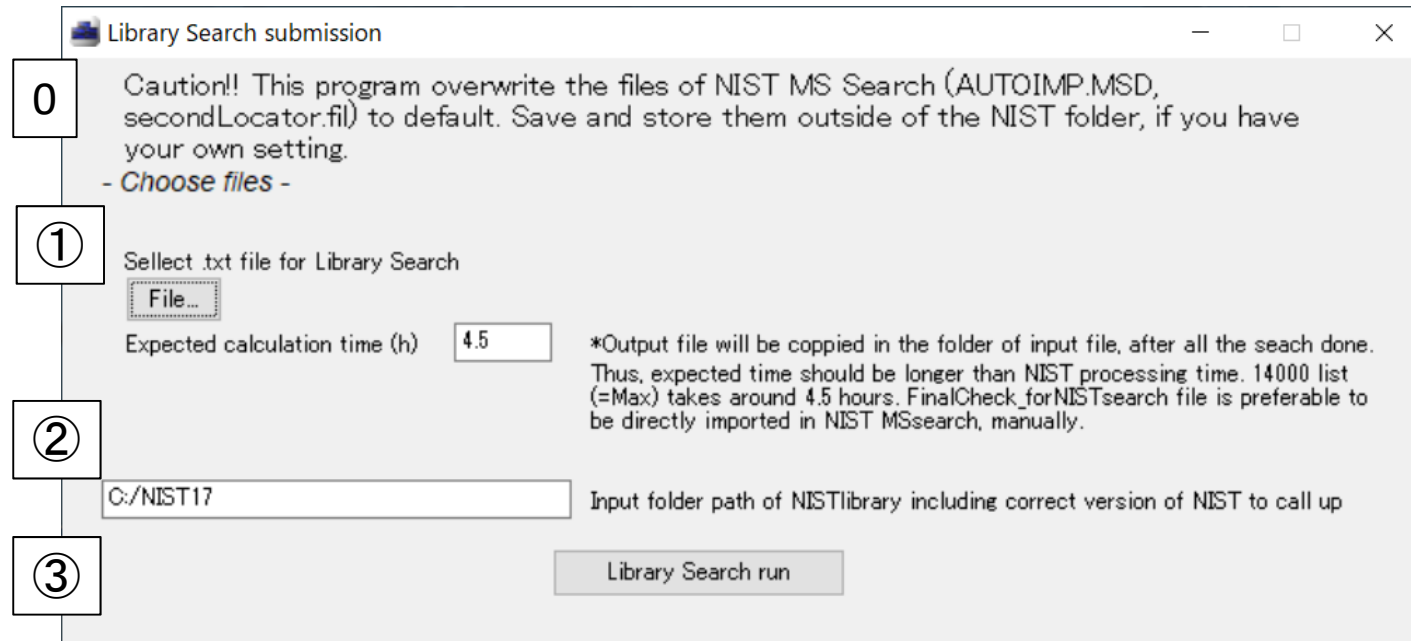
ステップ2とステップ4の実行ファイルです。

NISTライブラリ用のバッチ検索ファイルを読み込み、一括検索を行い、結果の出力ファイルを得ます。

ステップ4は、ステップ3のデータベース構築が終わった後、その内容のスペクトルをチェックするために実行します。

\* ただしステップ4については、(プログラム実行で結果の出力ファイルを得る必要はないため)直接マニュアルでNISTライブラリからバッチ検索ファイルを開き、実行することを推奨します。

## II\_IV\_LibrarySearch.exe 起動画面と操作方法について



- 0 本プログラム実行により、NISTプログラムフォルダの2つのファイル(AUTOIMP.MSD, secondLocator.fil)がデフォルトセッティングに直されます。
- あらかじめ同じファイルがNISTプログラムフォルダ内にあります。元ファイルが重要な場合、別の場所に保存しておいて下さい。
- ① NISTライブラリでバッチ検索するファイルを選択します。ファイル名は「Combined\_MSpeaklist\_forNISTsearch\_ファイル名.txt」です。

## II\_IV\_LibrarySearch.exe 起動画面と操作方法について

The screenshot shows the 'Library Search submission' window. On the left, a vertical line with four numbered boxes (0, 1, 2, 3) indicates the sequence of steps. Step 0 points to a caution message. Step 1 points to the 'File...' button. Step 2 points to the 'Expected calculation time (h)' input field. Step 3 points to the 'Library Search run' button.

Library Search submission

0 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 -

1 Select .txt file for Library Search  
File...

2 Expected calculation time (h) 4.5  
\*Output file will be copied in the folder of input file, after all the search 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.

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

- ② NISTライブラリサーチにかかる推定時間を入力してください。結果ファイルのコピー処理の関係上、長めに見積もってください。14000リストのサーチに4時間と想定して、4.5時間程度を見積もることを推奨します。
- ③ 検索に使用するNISTライブラリのフォルダパスを入力してください。バージョンによってパス名が異なります。
- またNISTライブラリの検索方法もあらかじめ設定しておいてください。NIST MS Searchソフトを立ち上げLibrary search option で設定します。推奨は「similarity」と「simple search」の検索モード、Number of hits to printは1です。

# 出力ファイルについて

```
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
```

## サーチ結果の出力例

- 出力ファイルとして、全ピークのNISTサーチ結果が入力されたテキストファイルが得られます。
- 1つのピークに対して、Number of hits to print数に応じた数の検索結果が割り当てられています。
- 上記の出力例の場合、Number of hits to printは1に設定されており、Hit1のみ出力されています。
- ファイル名は「Combined\_MSpeaklist\_forNISTsearch\_ファイル名.txtOutputLibrarySearch.txt」となります。

## III\_DBCreator+.exe

ステップ3の実行ファイルです。

デコンボリューション後に得られた全ピークのリスト出力ファイルと、そのNISTサーチの出力ファイルを読み込んで、任意の条件でピークを抽出し、データベースファイルとして出力します。抽出されたピークのスペクトルと検索結果を確認しやすいよう、それらピークについてのNISTライブラリサーチ用のファイルも同時に生成されます。



# III\_DBCreator+.exe

## 起動画面と操作方法について

The screenshot shows the DBCreator+ application window (64 bit version) with the following sections and steps:

- Step 1:** Select .txt file of NIST MS search output. A "File..." button is highlighted.
- Step 2:** Select .csv file of MSpeaklist result. A "File..." button is highlighted.
- Step 3:** Select save folder and input file name. Chromat pictures (.jpg) and a result file (.csv) are created. A "File..." button is highlighted.
- Step 4:** 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
- Step 5:** NIST MS Search setting GC setting
  - Match Factor (MF) threshold: NIST Hit list with the setting MF value and over is extracted. Range 0 ~ 1000. Value: 800
  - Number of hits to print: Should be same with NIST MS search setting. Value: 1
  - Keyword: Keyword in chemical formula to extract from the hit list. e.g.) Chlorinated compound is extracted by Cl

At the bottom right, there is a button labeled "DBCreator+ run".

① NISTサーチ結果のtxtファイルを選択します。ファイル名は手動変更がなければ「Combined\_MSpeaklist\_forNISTs\_earch\_ファイル名.txtOutputLibrarySearch.txt」です。

② ピークリストのcsvファイルを選択します。ファイル名は手動変更がなければ「Combined\_MSpeaklist\_ファイル名.csv」

③ 保存する場所とファイル名を入力します。

④ 使用するRのバージョン等を選択します。

### III\_DBCreator+.exe

## 起動画面と操作方法について

NIST MS Search setting GC setting

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 compound is extracted by Cl

DBcreator+ run

- ⑤-1 NISTサーチしたピークリストの内、Match Factor(MF)が900(上記例の場合)以上のものののみ抽出します。
- ⑤-2 NISTサーチを実行した際の設定値です。この値はNISTサーチでの設定値と一致させてください。
- ⑤-3 NISTサーチしたピークリストの内、組成式内に含まれる単語を条件抽出できます。「Cl」とした場合、組成式内に「Cl」を含みかつMFが設定値以上のものののみ抽出されます。

### III\_DBCreator+.exe

## 起動画面と操作方法について

NIST MS Search setting GC setting

⑤-4 1 GC1 tolerance (min): Multiply 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.

⑤-5 4 MPeriod: The modulation time period.

DBcreator+ run

- ⑤-4 GC1とGC2におけるRTの許容幅を設定します。
- 各々1minと1secとした場合、1次元方向に前後1minかつ2次元方向に前後1secで溶出した同じアサインの化合物リストはダブリとみなされ、1つ以外は削除されます。MFの低いリストが削除されます。
- ⑤-5 GCxGCの分析時に設定したModulation時間を入力します。

# 出力ファイルについて

BlobID	Compound	Group	Rtc	Inte	Peak I (min)	Peak II (sec)	IS	分子式	MF	RMF	CAS	LibID	Lib	file.n	deconv.nur	MS1	MS2
1	Heptadecan	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

## 2の化合物データベースの1部抜粋

2行に1つのピークについてのデータが収められている。

- 下記3つの出力ファイルが生成されます。
- 1 条件に合致した抽出リストが出力されます。化合物名、RT、組成式などメタデータを含むシンプルな出力です。ファイル名は「SimplePeaklist\_conditionKeyword\_MF\_MFvalueファイル名.csv」です。
- 2 TSENなどに利用するフォーマットの化合物データベースとして、1の結果にMSスペクトル情報を含んだ「Database\_conditionKeyword\_MF\_MFvalueファイル名.csv」が得られます。
- 3 抽出された化合物リストのスペクトルをNISTライブラリで一括表示確認するためのファイル「FinalCheck\_forNISTsearchDatabase\_conditionKeyword\_MF\_MFvalueファイル名.csv.txt」が得られます。NIST MS Searchソフトウェアで本ファイルを読みこむことにより、目視で抽出リストのMSスペクトルとヒットしたリファレンスMSスペクトルを一つずつ確認できます。

# 予想されるエラーと対処について

Q: 「NMFdeconvolution run」ボタンを押した後、NMFdeconvolution.exeがフリーズして解析が始まりません。

A: データ解析ソフト「R」のリンク先が誤って入力された可能性があります。→タスクバー右下にNMFdeconvolutionのアイコンが表示されているので、右クリックして一度Exitしてください。再度解析を際に、R.exeのファイルパスを確かめ、NMFdeconvolution.exeに正しく情報入力してください。「Library search run」や「DBcreator+ run」についても同様です。詳しくは本マニュアルの「起動画面と操作方法について」を参照してください。

Q: 同じ設定でNMFdeconvolutionツールを実行しているのに、解析がうまく始まるときとエラーで終わるときがあります。

A: 自動キー入力安定していない事が原因と考えられます。PCのその他処理信号と重複によるものかもしれませんので、うまく始まらなかった際には再度試してください。

# 注意事項について

- ソースファイルについて
- 配布フォルダ内にソースファイル「1NMFdeconvolution.r」「2MSpeaklist.r」「3FinalDBcreate.r」「4DBfinalCheckInNIST.r」があります。
- これらは全てexeファイルと同じフォルダ(もしくは同じ階層)に収めたままにしてください。
- License: Artistic License 2.0
- 免責: 本ツールを利用したことにより生じた、いかなる不利益も補償されません。