

PAH-Finder User Instruction

(Version 1.0)

Zixuan Zhang, Mingliang Fang*

Department of Environmental Science and Engineering, Fudan University, Shanghai
200433, China

* Author to whom correspondence should be addressed:

Dr. Mingliang Fang

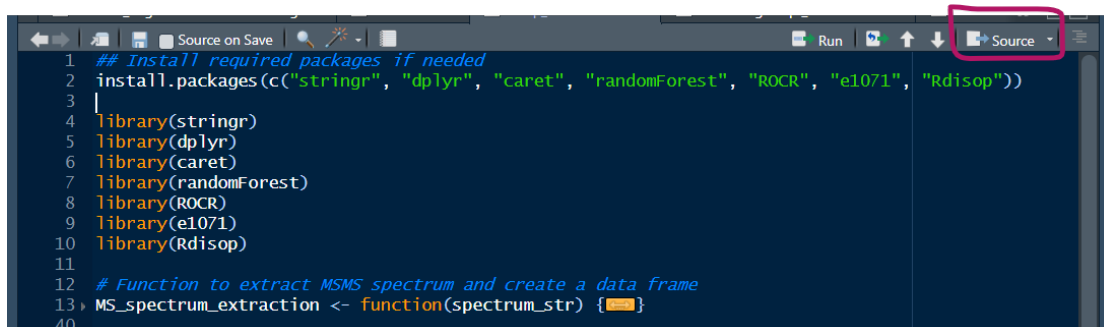
Email: mlfang@fudan.edu.cn

PAH-Finder is a computational workflow designed to identify PAH (Polycyclic Aromatic Hydrocarbons) derivatives from GC-HRMS (Gas Chromatography-High Resolution Mass Spectrometry) data. This workflow leverages machine learning to analyze mass spectral patterns, identify doubly charged fragments, and predict molecular formulas, facilitating the robust and efficient annotation of PAH compounds. This user manual provides step-by-step instructions on how to set up and use PAH-Finder, along with detailed explanations of its features and capabilities.

PAH-Finder setup

Install R and RStudio

- Download and install R studio following the instruction on the RStudio website (<https://www.rstudio.com/>).
- Download the R script “Setup_PAH-Finder.R” from the Github (<https://github.com/Barnett-Cheung/PAH-Finder>) then save them in the folder for data processing. Click “Source” to load the required packages and define the functions needed. Note: if you already have the required packages, you can ignore the line 2 (Figure 1).



```
1 ## Install required packages if needed
2 install.packages(c("stringr", "dplyr", "caret", "randomForest", "ROCR", "e1071", "Rdisop"))
3
4 library(stringr)
5 library(dplyr)
6 library(caret)
7 library(randomForest)
8 library(ROCR)
9 library(e1071)
10 library(Rdisop)
11
12 # Function to extract MSMS spectrum and create a data frame
13 MS_spectrum_extraction <- function(spectrum_str) {
14
15 }
16
17 }
```

Figure 1 Setup_PAH-Finder

Download the R script, “Demo_PAH-Finder.R”, the random forest model, “RFmodel_PAH-Finder.rds”, demo csv file “Demo_MSMS.csv” from (<https://github.com/Barnett-Cheung/PAH-Finder>) then save them in the folder for data processing. Demo MSMS file should be identical as Figure 2.

[illegible]

Figure 2 Demo file

Below is the step-by-step explanation of code:

Preparation

In line 15, specify the working directory. Ensure your data is in a CSV file with a column named MSMS containing the MS/MS spectra. Then load the data.

```
14 ## Set the working directory
15 setwd("C:/Users/46355/Desktop/PAH-Finder_Demo")
16
17 demo <- read.csv("Demo_PAH-Finder.csv")
```

Figure 3 Load the data

Extract MS/MS Spectra

```
19 # Process each MSMS spectrum and store in a list  
20 List_MS_spectra <- lapply(demo$MSMS, MS_spectrum_extraction)
```

Figure 4 Extract the spectra

After this step, each of your spectrum in the list should look like Figure 5.

mz	intensity	group
126.0459	24.3188	fragment
139.0538	171.8360	fragment
139.5556	44.4881	fragment
252.0926	11.1213	fragment
278.1077	999.0000	fragment
279.1121	156.3196	fragment
280.1154	20.2784	fragment

Figure 5 Example data frame of spectrum

Note: Input for PAH-Finder should be the purified MS/MS spectra. In the demo file, the format of MS spectrum is the fragment-intensity pair. Relative intensity (normalized to base peak) is used in PAH-Finder. In every spectrum, the intensity of base peak is 999. If you need help in this step, please feel free to contact me.

PAH suspect list generation

Code in Figure 6 can be run directly to prepare for machine learning prediction. Please see original manuscript for detailed discussion of each ML feature.

```
22 molecular_ion_mz <- as.numeric(unlist(lapply(List_MS_spectra, Estimate_molecular_ion_mz)))
23
24 # Calculate each machine learning features respectively
25 # Rename and combine into a dataframe for prediction
26 RIR_1 <- Calculate_region_intensity_ratio(List_MS_spectra, molecular_ion_mz, 52.5)[, 2]
27
28 RIR_2_3 <- Calculate_region_intensity_ratio(List_MS_spectra, molecular_ion_mz, 35)[, c(1,3)]
29
30 RIR_4 <- Calculate_region_intensity_ratio(List_MS_spectra, molecular_ion_mz, 10.5)[, 10]
31
32 Spectral_features <- Calculate_spectral_features(List_MS_spectra, molecular_ion_mz)
33
34 df_ML_features <- cbind(RIR_1, RIR_2_3, RIR_4, Spectral_features)
35
36 colnames(df_ML_features) <- c("RIR1", "RIR2", "RIR3", "RIR4",
37                               "Skewness", "Entropy", "BasePeakRatio")
```

Figure 6 Estimate the m/z of molecular ion and generate ML feature of each spectrum

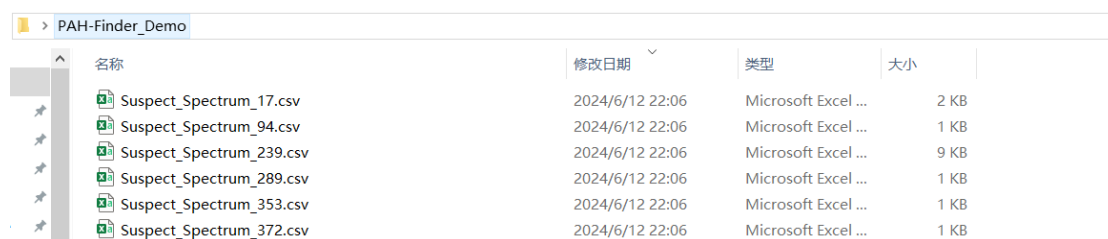
Combine positive prediction from machine learning and doubly charged filter to narrow down the suspect list of PAHs (Figure 7).

```
39 RFmodel <- readRDS(file="RFmodel_PAH-Finder.rds")
40
41 predictions_prob <- predict(RFmodel, newdata = df_ML_features, type = "prob")
42 # If you need the predictions to be binary classes (0 or 1) rather than probabilities,
43 # you can apply a threshold, here we use 0.5 as threshold to distinguish PAH from background
44 binary_predictions <- ifelse(predictions_prob > 0.5, 1, 0)
45
46 ML_prediction <- as.numeric(binary_predictions[, 2])
47
48 DB_frag_check <- unlist(lapply(List_MS_spectra, Identify_doubly_charged_fragments))
49
50 Suspect_indices <- which(ML_prediction == 1 & DB_frag_check == T)
51
52 Suspect_list <- List_MS_spectra[Suspect_indices]
```

Figure 7 Combine ML and DB filter to obtain a suspect list

Output and store the list of suspect spectra in the folder (Figure 8).

```
54 # Iterate through the Suspect_list and write each element to a CSV file
55 for (i in seq_along(Suspect_list)) {
56   # Create a unique file name for each element
57   file_name <- paste0("SuspectSpectrum_", Suspect_indices[i], ".csv")
58
59   # Write the element to a CSV file
60   write.csv(Suspect_list[[i]], file = file_name, row.names = FALSE)
61 }
```



名称	修改日期	类型	大小
Suspect_Spectrum_17.csv	2024/6/12 22:06	Microsoft Excel ...	2 KB
Suspect_Spectrum_94.csv	2024/6/12 22:06	Microsoft Excel ...	1 KB
Suspect_Spectrum_239.csv	2024/6/12 22:06	Microsoft Excel ...	9 KB
Suspect_Spectrum_289.csv	2024/6/12 22:06	Microsoft Excel ...	1 KB
Suspect_Spectrum_353.csv	2024/6/12 22:06	Microsoft Excel ...	1 KB
Suspect_Spectrum_372.csv	2024/6/12 22:06	Microsoft Excel ...	1 KB

Figure 8 Output the suspect list in the folder

Molecular formula validation

For each suspect spectrum, formula prediction is necessary for validation. Molecular ion peak of PAHs usually shows up in spectrum, which can be used to predict formula with elemental restrictions (**Figure 9**). In the given example, “C₂₂H₁₄” best explains the spectrum.

```

76 # Example usage
77 mass <- 278.1077
78 decomposeMass(mass, ppm = 10, minElements = "C10H8", maxElements = "C28H20")
79
80 # Isotope pattern: masses and their corresponding intensities
81 masses <- c(278.1077, 279.1121)
82 intensities <- c(999, 156.3196)
83
84 # Using decomposeIsotopes to predict formula with isotope information
85 decomposeIsotopes(masses, intensities, ppm = 10,
86                   minElements = "C10H8", maxElements = "C28H20")
87
88 ## In above output, C22H14 is the most reasonable formula that explain the spectrum

```

Figure 9 Formula validation with *Rdisop*

Troubleshooting

1. Ensure all required packages are installed and loaded.
2. Verify the working directory and file paths are correctly set.
3. Check the formatting of input data (e.g., correct column names and data types).

For any troubleshooting, preprocessing and data analysis assistance using PAH-Finder, please contact:

Email: zixuanzhang245@gmail.com

I am more than happy to provide help in using this workflow for academic research.