SulfurFinder User Manual

(Version 1.0, Feb. 12, 2025)

Brian Low[1], Tingting Zhao[1], Xingfang Li,[2,]* and Tao Huan[1,]*

1. Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, British Columbia, V6T 1Z1, Canada

2. Division of Analytical and Environmental Toxicology, Department of Laboratory Medicine and Pathology, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta T6G 2G3, Canada

Corresponding Authors

Tao Huan − Department of Chemistry, University of British Columbia, Vancouver, BC, Canada, V6T 1Z1. https://orcid.org/0000-0001-6295-2435. Email: thuan@chem.ubc.ca

Xing-Fang Li − Division of Analytical and Environmental Toxicology, Department of Laboratory Medicine and Pathology, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada, T6G 2G3. https://orcid.org/0000-0003-1844-7700. Email: xingfang.li@ualberta.ca

SulfurFinder is an R program for the accurate recognition of sulfur (S)-containing compounds from high-resolution mass spectrometry (MS) analysis. It contains three key modules, including 1) data cleaning, 2) S recognition, and 3) S number prediction. The R script “SulfurFinder.R” is the main program. Six machine learning (ML) models are used by SulfurFinder. The R script, ML models, and demo data are publicly available on GitHub: https://github.com/HuanLab/SulfurFinder.

Preparation

1) Software installation

    a. Install the R language (https://www.r-project.org/) (ver. 4.1.3 was used to write the program).

    b. Install RStudio (https://posit.co/) (2022.07.2 Build 576 was used to write the program).

    c. Install the following required R packages. If these packages are not already installed, the following R code can be used: install.packages("package_name"). The version used to write the program is indicated in parenthesis.

        i. "xcms" (ver. 3.16.1)

        ii. "foreach" (ver. 1.5.2)

        iii. "doParallel" (ver. 1.0.17)

        iv. "ranger" (ver. 0.14.1)

        v. "xlsx" (ver. 0.6.5)

2) Download the following R script and ML models from GitHub: https://github.com/HuanLab/SulfurFinder.

    a. "SulfurFinder.R"

    b. "M+2_S_recog.RDS"

    c. "M+3_S_recog.RDS"

    d. "M+4_S_recog.RDS"

    e. "M+2_S_number.RDS"

    f. "M+3_S_number.RDS"

    g. "M+4_S_number.RDS"

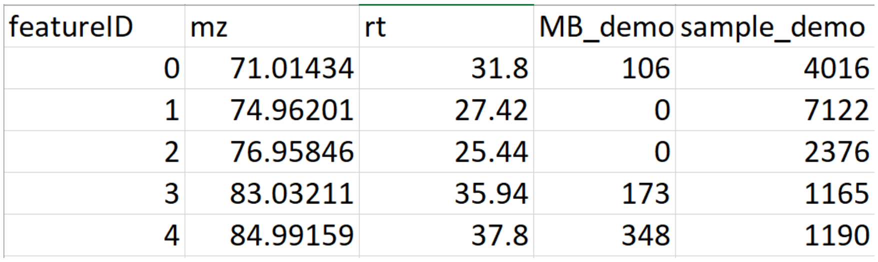
3) Prepare feature table in .xlsx format. Note that the feature table should be on the first sheet on the Excel spreadsheet. The first row of the feature table are the labels. The labels of the first three columns should match the screenshot below. The feature table should have at least 4 columns:

    a. “featureID" are identifiers for each feature. The identifiers should be unique.

    b. “mz” are the *m/z* for each feature

    c. “rt” are the retention times (in seconds) for each feature.

    d. Column 4 and everything after that are the sample intensities. The sample labels should be unique. If blanks were run, the “MB” label should be used (see below).



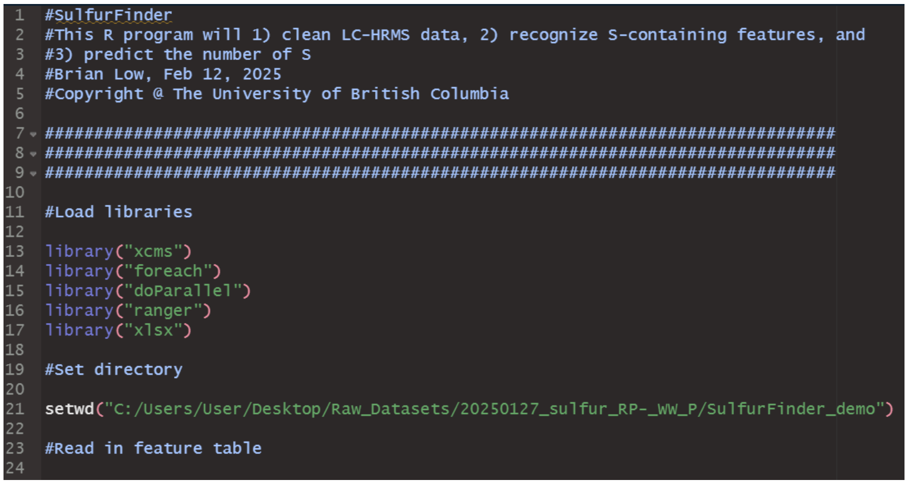
featureID	mz	rt	MB_demo	sample_demo
0	71.01434	31.8	106	4016
1	74.96201	27.42	0	7122
2	76.95846	25.44	0	2376
3	83.03211	35.94	173	1165
4	84.99159	37.8	348	1190

4) Prepare LC-HRMS raw data files in .mzML format. Note that the raw data file names need to be the same as the sample label names in the feature table.

5) Create a folder. Put the feature table, raw data, and the six ML models in it.

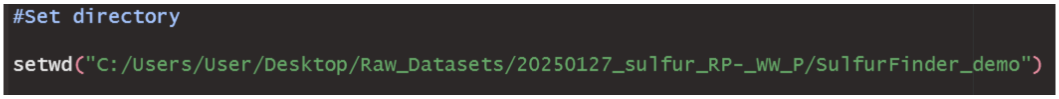
Main

1) Open the “SulfurFinder.R” script in RStudio.



```
1  #SulfurFinder
2  #This R program will 1) clean LC-HRMS data, 2) recognize S-containing features, and
3  #3) predict the number of S
4  #Brian Low, Feb 12, 2025
5  #Copyright @ The University of British Columbia
6
7  ##################################################################################
8  ##################################################################################
9  ##################################################################################
10
11 #Load libraries
12
13 library("xcms")
14 library("foreach")
15 library("doParallel")
16 library("ranger")
17 library("xlsx")
18
19 #Set directory
20
21 setwd("C:/Users/User/Desktop/Raw_Datasets/20250127_sulfur_RP-_WW_P/SulfurFinder_demo")
22
23 #Read in feature table
24
```

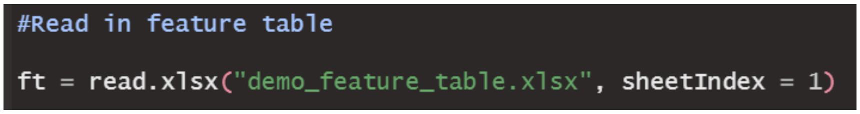
2) On line 21, set the working directory to where the folder previously created in **Preparation** is.



```
#Set directory

setwd("C:/Users/User/Desktop/Raw_Datasets/20250127_sulfur_RP-_WW_P/SulfurFinder_demo")
```

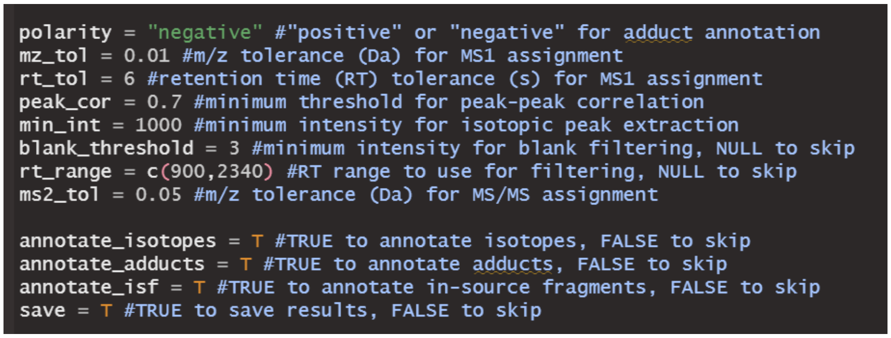
3) Read in the feature table using the following code: read.xlsx(“feature_table_name.xlsx”, sheetIndex = 1). The feature table should be the first sheet in the Excel spreadsheet.



```
#Read in feature table

ft = read.xlsx("demo_feature_table.xlsx", sheetIndex = 1)
```

4) Set parameters in lines 31-43. The function for each parameter is detailed below.

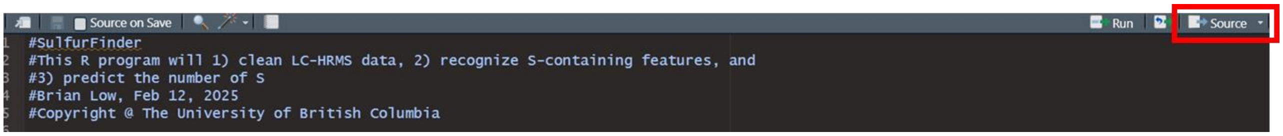


```
polarity = "negative" #"positive" or "negative" for adduct annotation
mz_tol = 0.01 #m/z tolerance (Da) for MS1 assignment
rt_tol = 6 #retention time (RT) tolerance (s) for MS1 assignment
peak_cor = 0.7 #minimum threshold for peak-peak correlation
min_int = 1000 #minimum intensity for isotopic peak extraction
blank_threshold = 3 #minimum intensity for blank filtering, NULL to skip
rt_range = c(900,2340) #RT range to use for filtering, NULL to skip
ms2_tol = 0.05 #m/z tolerance (Da) for MS/MS assignment

annotate_isotopes = T #TRUE to annotate isotopes, FALSE to skip
annotate_adducts = T #TRUE to annotate adducts, FALSE to skip
annotate_isf = T #TRUE to annotate in-source fragments, FALSE to skip
save = T #TRUE to save results, FALSE to skip
```

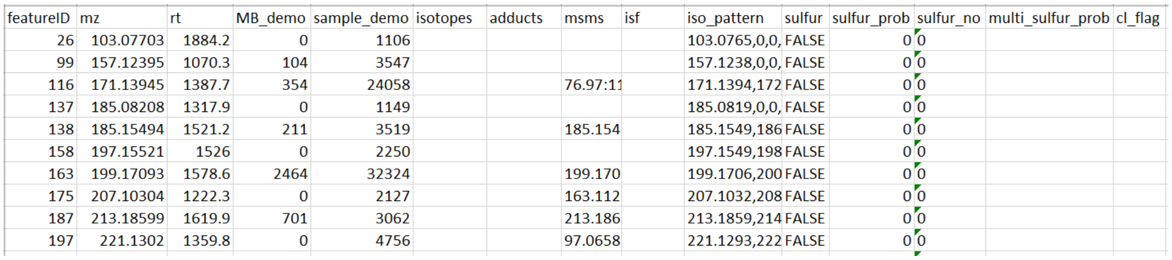
Parameter	Function
polarity	Character, if data was collected in positive ionization mode, set “positive. If data was collected in negative ionization mode, set “negative”. Default: “positive”
mz_tol	Numeric, *m/z* tolerance (Da) for MS1 assignment of isotopes, adducts, and in-source fragments. Default: 0.01 Da
rt_tol	Numeric, retention time (s) tolerance for MS1 assignment of isotopes, adducts, and in-source fragments. Default: 6 s
peak_cor	Numeric, minimum peak-peak correlation for isotope, adduct, and in-source fragments assignment. Default: 0.7
min_int	Numeric, minimum M isotopic peak intensity for MS1 isotope pattern averaging. Default: 1000
blank_threshold	Numeric, features with an average sample intensity less than blank_threshold times the average blank intensity will be removed. If no blank filtering is needed, set NULL. Default: 3
rt_range	Numeric, retention time range (s) between [rt_1, rt_2] will be considered for downstream analysis. If no retention time filtering is needed, set NULL. Default: NULL
ms2_tol	Numeric, *m/z* tolerance (Da) for MS/MS matching used for the annotation of in-source fragments. Default: 0.05 Da
annotate_isotopes	Logical, TRUE or FALSE. If TRUE, natural isotopes in the feature table will be annotated. Default: TRUE
annotate_adducts	Logical, TRUE or FALSE. If TRUE, adducts in the feature table will be annotated. Default: TRUE
annotate_isf	Logical, TRUE or FALSE. If TRUE, in-source fragments in the feature table will be annotated. Set FALSE if no MS/MS was collected. Default: TRUE
save	Logical, TRUE or FALSE. If TRUE, SulfurFinder results will be saved in the directory. Default: TRUE

5) Run the script by clicking Source.



```
#SulfurFinder
#This R program will 1) clean LC-HRMS data, 2) recognize S-containing features, and
#3) predict the number of S
#Brian Low, Feb 12, 2025
#Copyright @ The University of British Columbia
```

6) If save was TRUE, the SulfurFinder results named “SulfurFinder_results.xlsx” will be saved in the directory. Depending on the parameters used, up to 10 additional columns will be added to the feature table.



featureID	mz	rt	MB_demo	sample_demo	isotopes	adducts	msms	isf	iso_pattern	sulfur	sulfur_prob	sulfur_no	multi_sulfur_prob	cl_flag
26	103.07703	1884.2	0	1106					103.0765,0,0,	FALSE	0	0		
99	157.12395	1070.3	104	3547					157.1238,0,0,	FALSE	0	0		
116	171.13945	1387.7	354	24058			76.97:11		171.1394,172	FALSE	0	0		
137	185.08208	1317.9	0	1149					185.0819,0,0,	FALSE	0	0		
138	185.15494	1521.2	211	3519			185.154		185.1549,186	FALSE	0	0		
158	197.15521	1526	0	2250					197.1549,198	FALSE	0	0		
163	199.17093	1578.6	2464	32324			199.170		199.1706,200	FALSE	0	0		
175	207.10304	1222.3	0	2127			163.112		207.1032,208	FALSE	0	0		
187	213.18599	1619.9	701	3062			213.186		213.1859,214	FALSE	0	0		
197	221.1302	1359.8	0	4756			97.0658		221.1293,222	FALSE	0	0		

Label	**Description**
isotopes	Natural isotope annotation result. If a feature was annotated to be an isotope of another feature, there will be a result here. For example, if a feature has the result 100[M+1], then that feature was annotated to be the M+1 isotopic peak of featureID 100.
adducts	Adduct annotation result. If a feature was annotated to be an adduct of another feature, there will be a result here. For example, if a feature has the result 100[M+Na]+, then that feature was annotated to be the Na adduct of featureID 100.
msms	Experimental MS/MS spectrum of the feature.
isf	In-source fragment (ISF) annotation result. If the feature was annotated to be an ISF of another feature, there will be a result here. For example, if a feature has the result 100[L1], then that feature was annotated to be a Level 1 ISF of featureID 100.
iso_pattern	Experimental isotope pattern of the feature.
sulfur	S recognition. TRUE if feature contains S.
sulfur_prob	Probability of feature containing S.
sulfur_no	Number of S. 1 if feature contains 1 S, 2 if feature contains ≥2 S.
multi_sulfur_prob	Probability of feature containing ≥2 S.

cl_flag	“Cl flag” if feature was flagged to contain Cl based on M+2 isotopic peak intensity (≥25%).