Avir.R User Instruction

(Version 1.0, Jun 08, 2023)

Zixuan Zhang, Tao Huan*

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

* Author to whom correspondence should be addressed:

Dr. Tao Huan

Tel: (+1)-604-822-4891

E-mail: thuan@chem.ubc.ca

Website: https://huan.chem.ubc.ca/

- Avir stands for Alignment and Integration Evaluator. It is a support vector machine (SVM)-based R program to predict the quality of metabolic peak integration in liquid chromatography-mass spectrometry-based metabolomics data.

- Avir.R script is freely available for non-commercial use.

- The instructions are given below:

Preparation

1) Software installation

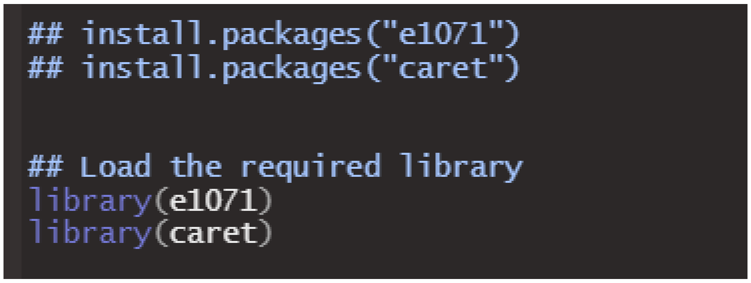
   Download and install R studio following the instruction on the RStudio website (https://www.rstudio.com/).

2) R package installation

   If the R package "e1071" and "caret" are not installed. Please run the following code then load these two packages:

   install.packages("e1071")

   install.packages("caret")



```
## install.packages("e1071")
## install.packages("caret")


## Load the required library
library(e1071)
library(caret)
```

Figure 1

3) Data preparation

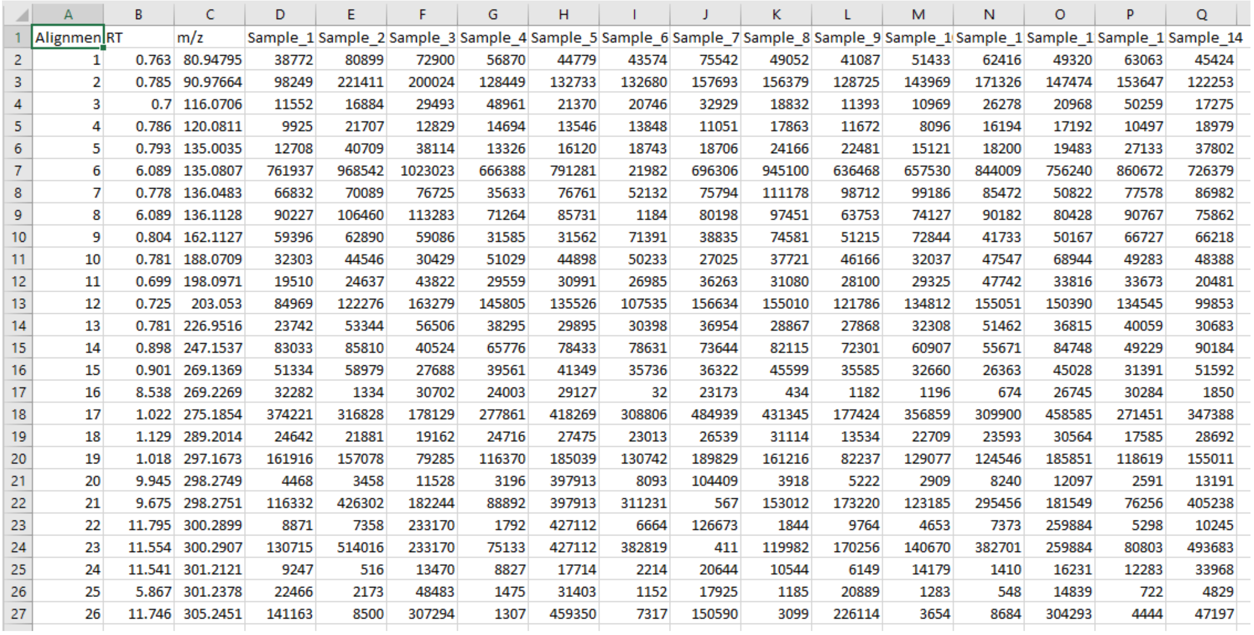
a. Download the R script, “Avir.R”, the SVM model, “Avir.rds”, from (https://github.com/HuanLab/AVIR.R) then save them in the folder for data processing.
Within the same folder, prepare two .csv files, one for sample metabolite-intensities in peak area and the other for sample metabolite intensities in peak height. The content in each column should be prepared as follows (**Figure 2**).

Column 1: alignment ID

Column 2: retention time

Column 3: m/z value

Column 4 to the last column: MS signal intensities of real samples



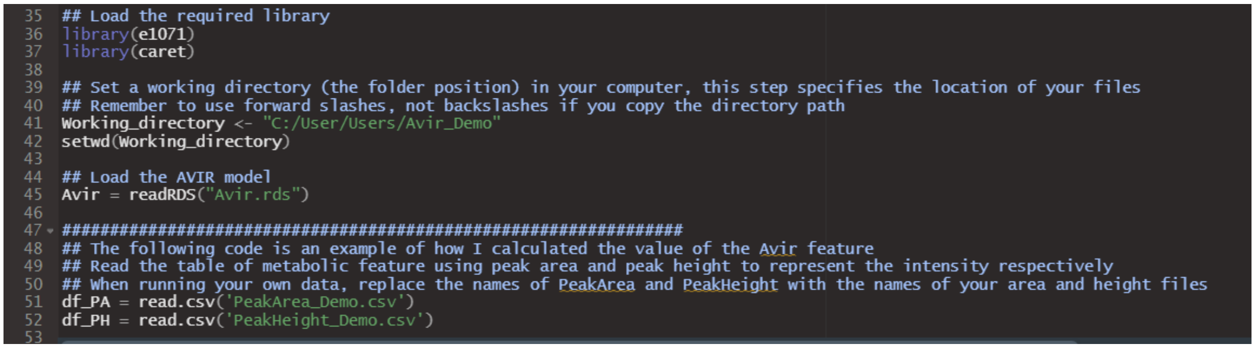
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Alignmen	RT	m/z	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	Sample_7	Sample_8	Sample_9	Sample_1	Sample_1	Sample_1	Sample_1	Sample_14
2	1	0.763	80.94795	38772	80899	72900	56870	44779	43574	75542	49052	41087	51433	62416	49320	63063	45424
3	2	0.785	90.97664	98249	221411	200024	128449	132733	132680	157693	156379	128725	143969	171326	147474	153647	122253
4	3	0.7	116.0706	11552	16884	29493	48961	21370	20746	32929	18832	11393	10969	26278	20968	50259	17275
5	4	0.786	120.0811	9925	21707	12829	14694	13546	13848	11051	17863	11672	8096	16194	17192	10497	18979
6	5	0.793	135.0035	12708	40709	38114	13326	16120	18743	18706	24166	22481	15121	18200	19483	27133	37802
7	6	6.089	135.0807	761937	968542	1023023	666388	791281	21982	696306	945100	636468	657530	844009	756240	860672	726379
8	7	0.778	136.0483	66832	70089	76725	35633	76761	52132	75794	111178	98712	99186	85472	50822	77578	86982
9	8	6.089	136.1128	90227	106460	113283	71264	85731	1184	80198	97451	63753	74127	90182	80428	90767	75862
10	9	0.804	162.1127	59396	62890	59086	31585	31562	71391	38835	74581	51215	72844	41733	50167	66727	66218
11	10	0.781	188.0709	32303	44546	30429	51029	44898	50233	27025	37721	46166	32037	47547	68944	49283	48388
12	11	0.699	198.0971	19510	24637	43822	29559	30991	26985	36263	31080	28100	29325	47742	33816	33673	20481
13	12	0.725	203.053	84969	122276	163279	145805	135526	107535	156634	155010	121786	134812	155051	150390	134545	99853
14	13	0.781	226.9516	23742	53344	56506	38295	29895	30398	36954	28867	27868	32308	51462	36815	40059	30683
15	14	0.898	247.1537	83033	85810	40524	65776	78433	78631	73644	82115	72301	60907	55671	84748	49229	90184
16	15	0.901	269.1369	51334	58979	27688	39561	41349	35736	36322	45599	35585	32660	26363	45028	31391	51592
17	16	8.538	269.2269	32282	1334	30702	24003	29127	32	23173	434	1182	1196	674	26745	30284	1850
18	17	1.022	275.1854	374221	316828	178129	277861	418269	308806	484939	431345	177424	356859	309900	458585	271451	347388
19	18	1.129	289.2014	24642	21881	19162	24716	27475	23013	26539	31114	13534	22709	23593	30564	17585	28692
20	19	1.018	297.1673	161916	157078	79285	116370	185039	130742	189829	161216	82237	129077	124546	185851	118619	155011
21	20	9.945	298.2749	4468	3458	11528	3196	397913	8093	104409	3918	5222	2909	8240	12097	2591	13191
22	21	9.675	298.2751	116332	426302	182244	88892	397913	311231	567	153012	173220	123185	295456	181549	76256	405238
23	22	11.795	300.2899	8871	7358	233170	1792	427112	6664	126673	1844	9764	4653	7373	259884	5298	10245
24	23	11.554	300.2907	130715	514016	233170	75133	427112	382819	411	119982	170256	140670	382701	259884	80803	493683
25	24	11.541	301.2121	9247	516	13470	8827	17714	2214	20644	10544	6149	14179	1410	16231	12283	33968
26	25	5.867	301.2378	22466	2173	48483	1475	31403	1152	17925	1185	20889	1283	548	14839	722	4829
27	26	11.746	305.2451	141163	8500	307294	1307	459350	7317	150590	3099	226114	3654	8684	304293	4444	47197

Figure 2

Note: Demo files can be found in “PeakArea_Demo.csv” from (https://github.com/HuanLab/AVIR.R).

Main:

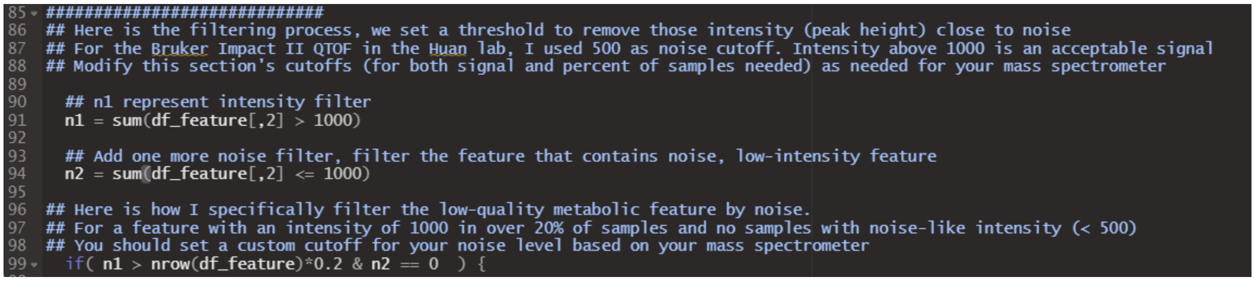
1) Set up the working directory and specify the input files (code line 41).
   For demonstration, here we put the folder in the desktop named "Avir_Demo". Specify the name of sample metabolite-intensity in peak area and peak height (code lines 51 and 52).



```r
35  ## Load the required library
36  library(e1071)
37  library(caret)
38
39  ## Set a working directory (the folder position) in your computer, this step specifies the location of your files
40  ## Remember to use forward slashes, not backslashes if you copy the directory path
41  Working_directory <- "C:/User/Users/Avir_Demo"
42  setwd(Working_directory)
43
44  ## Load the AVIR model
45  Avir = readRDS("Avir.rds")
46
47  #################################################################
48  ## The following code is an example of how I calculated the value of the Avir feature
49  ## Read the table of metabolic feature using peak area and peak height to represent the intensity respectively
50  ## When running your own data, replace the names of PeakArea and PeakHeight with the names of your area and height files
51  df_PA = read.csv('PeakArea_Demo.csv')
52  df_PH = read.csv('PeakHeight_Demo.csv')
53
```

Figure 3

2) Specify the intensity threshold and reproducibility filter for high-quality metabolic features.



```r
#############################
## Here is the filtering process, we set a threshold to remove those intensity (peak height) close to noise
## For the Bruker Impact II QTOF in the Huan lab, I used 500 as noise cutoff. Intensity above 1000 is an acceptable signal
## Modify this section's cutoffs (for both signal and percent of samples needed) as needed for your mass spectrometer

  ## n1 represent intensity filter
  n1 = sum(df_feature[,2] > 1000)

  ## Add one more noise filter, filter the feature that contains noise, low-intensity feature
  n2 = sum(df_feature[,2] <= 1000)

## Here is how I specifically filter the low-quality metabolic feature by noise.
## For a feature with an intensity of 1000 in over 20% of samples and no samples with noise-like intensity (< 500)
## You should set a custom cutoff for your noise level based on your mass spectrometer
  if( n1 > nrow(df_feature)*0.2 & n2 == 0  ) {
```

Figure 4

To enhance the performance and prediction accuracy of the SVM model, it's essential to set appropriate intensity thresholds. These thresholds enable the exclusion of low-quality, noise-like peaks so that Avir only works on high-quality metabolic features that are more likely to be real metabolites.

Step 1: Set the intensity threshold
Lines 91 and 94 of the R script control the intensity threshold. For Bruker's Impact II QTOF mass spectrometer, we recommend setting this value to 1000 counts. Metabolic features that don't reach this threshold are considered low quality and excluded from further analysis. For the Impact II QTOF, a setting of 1000 counts is recommended. Any feature below this level will be classified as noise and subsequently discarded.

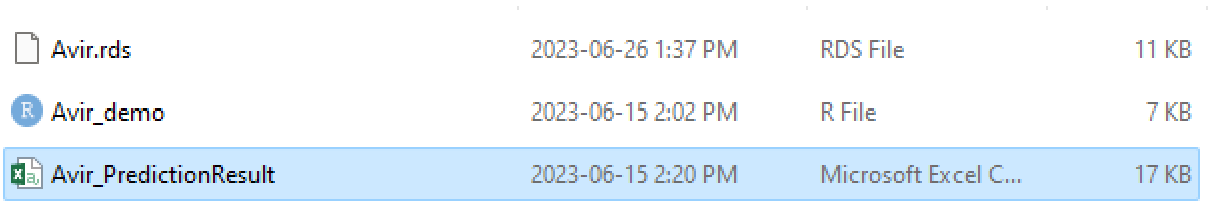
Step 2: Apply reproducibility filter
Line 99 controls the reproducibility level. To ensure the reliability of our predictions, metabolic features with low reproducibility are also filtered out. A feature will be considered high quality if at least 20% of the samples show an intensity above the set threshold of 1000 counts.

Please note that these settings are dependent on the specific experimental design, and may need adjustment depending on your research aim. The reproducibility filter can be customized by the user to fit their specific needs. By carefully adjusting these parameters, you can increase the quality of features included in the model, and potentially improve prediction accuracy.

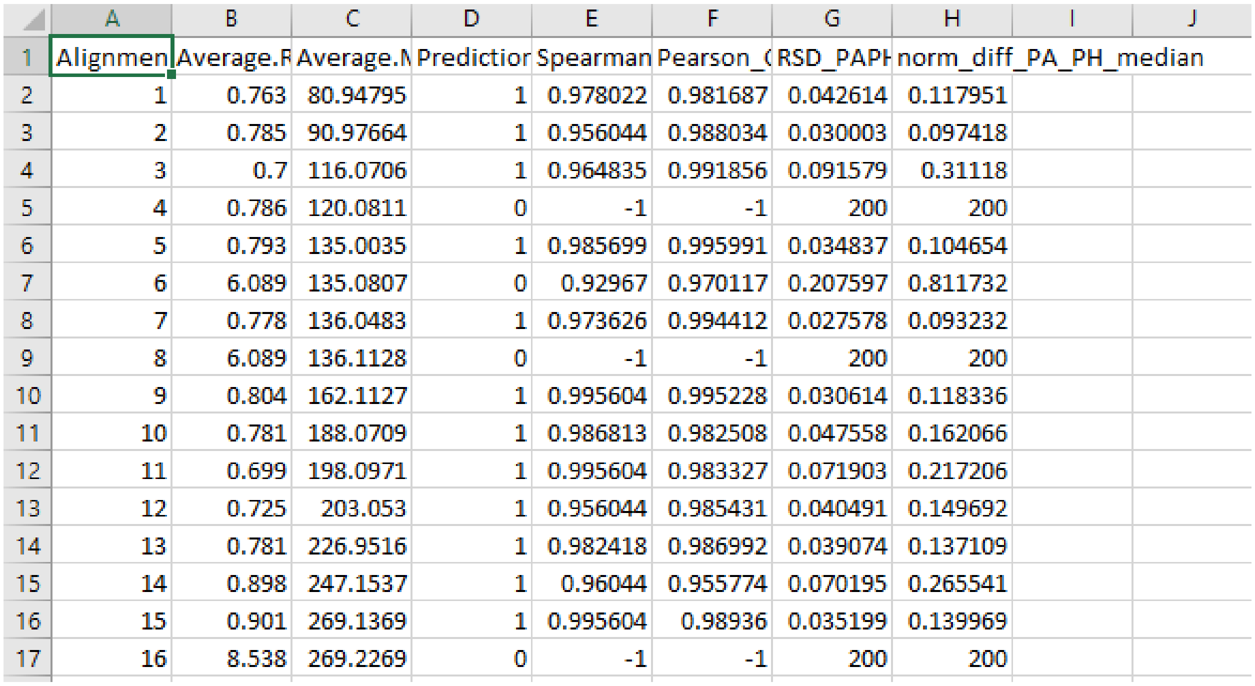
3) Run the R script by clicking “Source” on the top right of R studio panel.

4) Check the output prediction by Avir.
The csv file named “Avir_PredictionResult.csv” contains the prediction outcome. The first three columns of the table represent the information of Alignment ID, retention time and m/z. The four column “Prediction” represent the output of Avir. 1 means TRUE, there is no computational variation in the metabolic feature. and 0 means FALSE, there is

high computational variation in the metabolic feature. The rest of four columns are the statistical properties that is used for Avir prediction.



Avir.rds	2023-06-26 1:37 PM	RDS File	11 KB
Avir_demo	2023-06-15 2:02 PM	R File	7 KB
Avir_PredictionResult	2023-06-15 2:20 PM	Microsoft Excel C...	17 KB



	A	B	C	D	E	F	G	H	I	J
1	Alignmen	Average.R	Average.M	Predictior	Spearman	Pearson_(	RSD_PAPH	norm_diff_PA_PH_median		
2	1	0.763	80.94795	1	0.978022	0.981687	0.042614	0.117951		
3	2	0.785	90.97664	1	0.956044	0.988034	0.030003	0.097418		
4	3	0.7	116.0706	1	0.964835	0.991856	0.091579	0.31118		
5	4	0.786	120.0811	0	-1	-1	200	200		
6	5	0.793	135.0035	1	0.985699	0.995991	0.034837	0.104654		
7	6	6.089	135.0807	0	0.92967	0.970117	0.207597	0.811732		
8	7	0.778	136.0483	1	0.973626	0.994412	0.027578	0.093232		
9	8	6.089	136.1128	0	-1	-1	200	200		
10	9	0.804	162.1127	1	0.995604	0.995228	0.030614	0.118336		
11	10	0.781	188.0709	1	0.986813	0.982508	0.047558	0.162066		
12	11	0.699	198.0971	1	0.995604	0.983327	0.071903	0.217206		
13	12	0.725	203.053	1	0.956044	0.985431	0.040491	0.149692		
14	13	0.781	226.9516	1	0.982418	0.986992	0.039074	0.137109		
15	14	0.898	247.1537	1	0.96044	0.955774	0.070195	0.265541		
16	15	0.901	269.1369	1	0.995604	0.98936	0.035199	0.139969		
17	16	8.538	269.2269	0	-1	-1	200	200		

Figure 5

Note: For slow-quality metabolic features that are not worth running machine learning prediction, extreme coefficients and values are assigned. These features can be easily recognized from the outcome as '-1' is assigned to Spearman correlation and Pearson correlation (one kind of low-quality metabolic features is assigned -1, and the other kind of metabolic features that mostly contains noise is assigned 0.). In addition, RSD of PA/PH and the normalized range of PA/PH are assigned as assigned as 200. In **Figure 5**, we can see metabolic feature #4, #8 and #16 (Alignment ID, the first column) are in low quality and thus assigned extreme values.

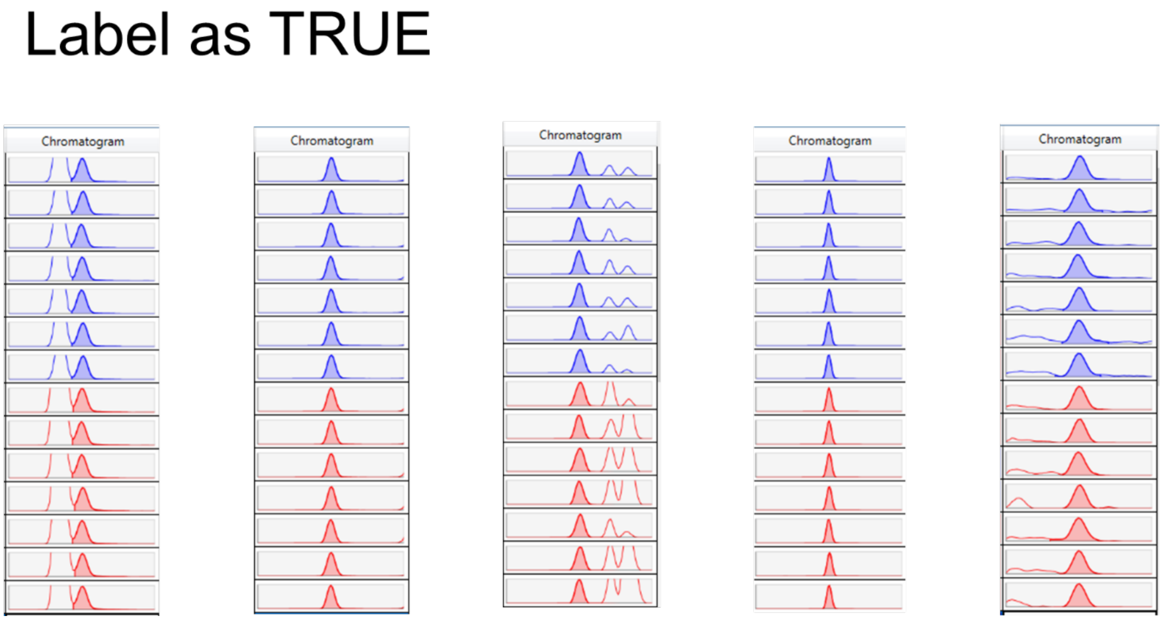
SVM Model Development

This section talks about how users can train their own SVM model in R for Avir applications.

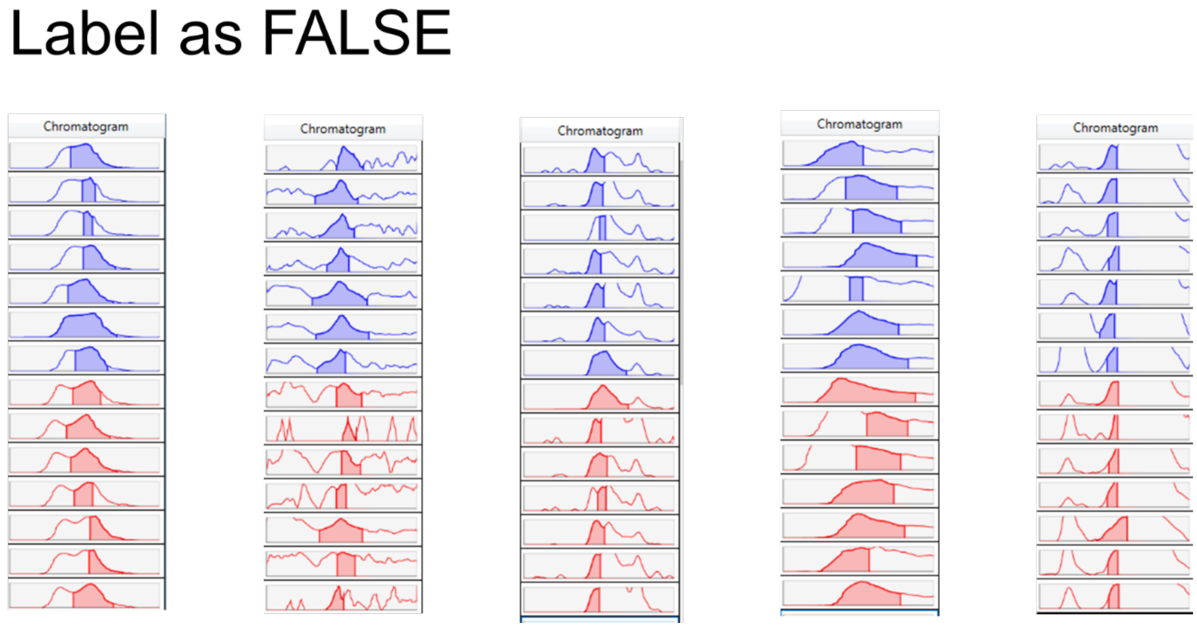
Training data labeling

Here we show the examples of labeling metabolic features, you can label the metabolomics data from your LC-MS platform to prepare the training dataset and test dataset.

The following figure shows five examples of metabolic features with no computational variation. These features are labeled as TRUE in the training data.



The following figure shows five examples of metabolic features with high computational variation. These features are labeled as FALSE in the training data.



Model Training

After collecting enough training data (it would better that size of training data is larger than 500), you can follow guidance below to generate your SVM model for prediction:

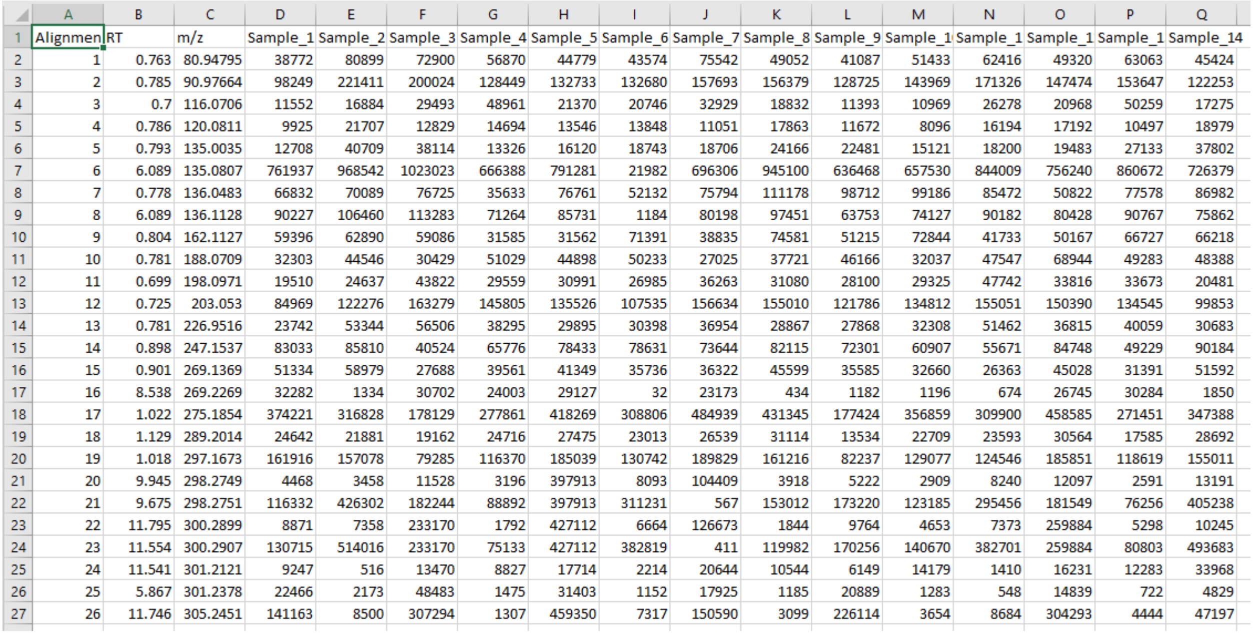
a. Download the R script, “ModelGeneration.R”, from (https://github.com/HuanLab/AVIR.R) then save them in the folder for data processing.
Within the same folder, prepare three .csv files, one for sample metabolite-intensities in peak area,one for sample metabolite intensities in peak height and one for the labeling results of metabolic features. The content in each column of sample metabolite-instensities should be prepared as **Figure 2**.

Column 1: alignment ID

Column 2: retention time

Column 3: m/z value

Column 4 to the last column: MS signal intensities of real samples



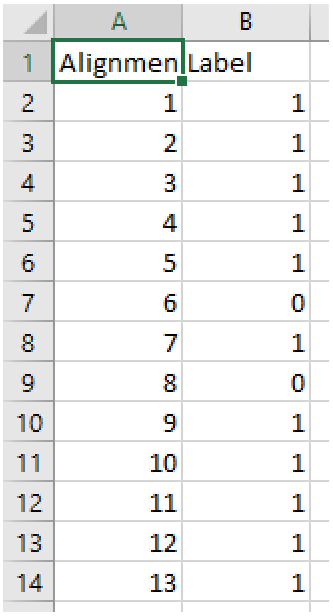
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Alignmen	RT	m/z	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	Sample_7	Sample_8	Sample_9	Sample_1	Sample_1	Sample_1	Sample_1	Sample_14
2	1	0.763	80.94795	38772	80899	72900	56870	44779	43574	75542	49052	41087	51433	62416	49320	63063	45424
3	2	0.785	90.97664	98249	221411	200024	128449	132733	132680	157693	156379	128725	143969	171326	147474	153647	122253
4	3	0.7	116.0706	11552	16884	29493	48961	21370	20746	32929	18832	11393	10969	26278	20968	50259	17275
5	4	0.786	120.0811	9925	21707	12829	14694	13546	13848	11051	17863	11672	8096	16194	17192	10497	18979
6	5	0.793	135.0035	12708	40709	38114	13326	16120	18743	18706	24166	22481	15121	18200	19483	27133	37802
7	6	6.089	135.0807	761937	968542	1023023	666388	791281	21982	696306	945100	636468	657530	844009	756240	860672	726379
8	7	0.778	136.0483	66832	70089	76725	35633	76761	52132	75794	111178	98712	99186	85472	50822	77578	86982
9	8	6.089	136.1128	90227	106460	113283	71264	85731	1184	80198	97451	63753	74127	90182	80428	90767	75862
10	9	0.804	162.1127	59396	62890	59086	31585	31562	71391	38835	74581	51215	72844	41733	50167	66727	66218
11	10	0.781	188.0709	32303	44546	30429	51029	44898	50233	27025	37721	46166	32037	47547	68944	49283	48388
12	11	0.699	198.0971	19510	24637	43822	29559	30991	26985	36263	31080	28100	29325	47742	33816	33673	20481
13	12	0.725	203.053	84969	122276	163279	145805	135526	107535	156634	155010	121786	134812	155051	150390	134545	99853
14	13	0.781	226.9516	23742	53344	56506	38295	29895	30398	36954	28867	27868	32308	51462	36815	40059	30683
15	14	0.898	247.1537	83033	85810	40524	65776	78433	78631	73644	82115	72301	60907	55671	84748	49229	90184
16	15	0.901	269.1369	51334	58979	27688	39561	41349	35736	36322	45599	35585	32660	26363	45028	31391	51592
17	16	8.538	269.2269	32282	1334	30702	24003	29127	32	23173	434	1182	1196	674	26745	30284	1850
18	17	1.022	275.1854	374221	316828	178129	277861	418269	308806	484939	431345	177424	356859	309900	458585	271451	347388
19	18	1.129	289.2014	24642	21881	19162	24716	27475	23013	26539	31114	13534	22709	23593	30564	17585	28692
20	19	1.018	297.1673	161916	157078	79285	116370	185039	130742	189829	161216	82237	129077	124546	185851	118619	155011
21	20	9.945	298.2749	4468	3458	11528	3196	397913	8093	104409	3918	5222	2909	8240	12097	2591	13191
22	21	9.675	298.2751	116332	426302	182244	88892	397913	311231	567	153012	173220	123185	295456	181549	76256	405238
23	22	11.795	300.2899	8871	7358	233170	1792	427112	6664	126673	1844	9764	4653	7373	259884	5298	10245
24	23	11.554	300.2907	130715	514016	233170	75133	427112	382819	411	119982	170256	140670	382701	259884	80803	493683
25	24	11.541	301.2121	9247	516	13470	8827	17714	2214	20644	10544	6149	14179	1410	16231	12283	33968
26	25	5.867	301.2378	22466	2173	48483	1475	31403	1152	17925	1185	20889	1283	548	14839	722	4829
27	26	11.746	305.2451	141163	8500	307294	1307	459350	7317	150590	3099	226114	3654	8684	304293	4444	47197

Figure 2

The Label.csv" file associate the True/False labels with the Alignment ID of metabolic feature. The content in each column should be prepared as follows (**Figure 6**):

Column 1: alignment ID

Column 2: label ('1' represents true and '0' represents false)



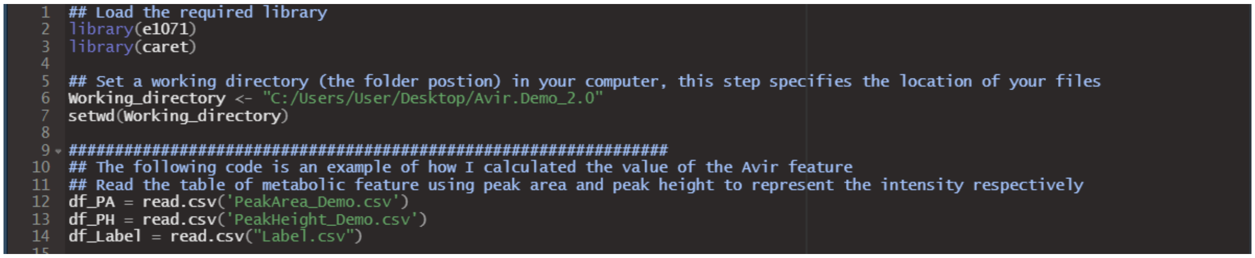
	A	B
1	Alignmen	Label
2	1	1
3	2	1
4	3	1
5	4	1
6	5	1
7	6	0
8	7	1
9	8	0
10	9	1
11	10	1
12	11	1
13	12	1
14	13	1

Figure 6

Note: Demo files can be found in “PeakArea_Demo.csv”, “PeakHeight_Demo.csv, “Label.csv” from (https://github.com/HuanLab/AVIR.R).

b. Set up the working directory and specify the input files (code line 6).

For demonstration, here we put the folder in the desktop named “Avir_Demo_2.0”. Specify the name of sample metabolite-intensity in peak area and peak height, and the file name of label. (code lines 12-14).



```
 1  ## Load the required library
 2  library(e1071)
 3  library(caret)
 4
 5  ## Set a working directory (the folder postion) in your computer, this step specifies the location of your files
 6  Working_directory <- "C:/Users/User/Desktop/Avir.Demo_2.0"
 7  setwd(Working_directory)
 8
 9  #################################################################
10  ## The following code is an example of how I calculated the value of the Avir feature
11  ## Read the table of metabolic feature using peak area and peak height to represent the intensity respectively
12  df_PA = read.csv('PeakArea_Demo.csv')
13  df_PH = read.csv('PeakHeight_Demo.csv')
14  df_Label = read.csv("Label.csv")
```

Figure 7

c. Run the R script by clicking “Source” on the top right of R studio panel.

d. Check the output SVM model.

Your own SVM model will be be named “SVM.rds” and saved in “rds” format in the folder. Then you can load the model for further applications.