基于非靶向代谢组学的化学成分差异性分析

Yandong Yin

2019-09-06

目录

表格

插图

寄言

借此为代谢组学贡献绵薄之力

简介

液相色谱-质谱联用技术（LC-MS）在代谢组学中有着非常广泛的应用，代谢组学借助于 LC-MS，可以广泛的筛查样品中存在的代谢物（小分子化合物）及其含量，具有高通量、高灵敏度、动态范围广等特点，可以实现优质的代谢组学分析。根据实验目的，一般我们可以采用靶向代谢组学方法进行特定目标代谢物的高准确度定量分析，如 MRM/SRM/PRM 等技术，也可以采用非靶向代谢组学筛查样品中包含的代谢物及其差异，从而实现广谱的代谢物筛查，根据质谱中数据采集方式的不同，分为数据依赖型采集（DDA）和数据非依赖型采集（DIA）两类 (Figure 1)。

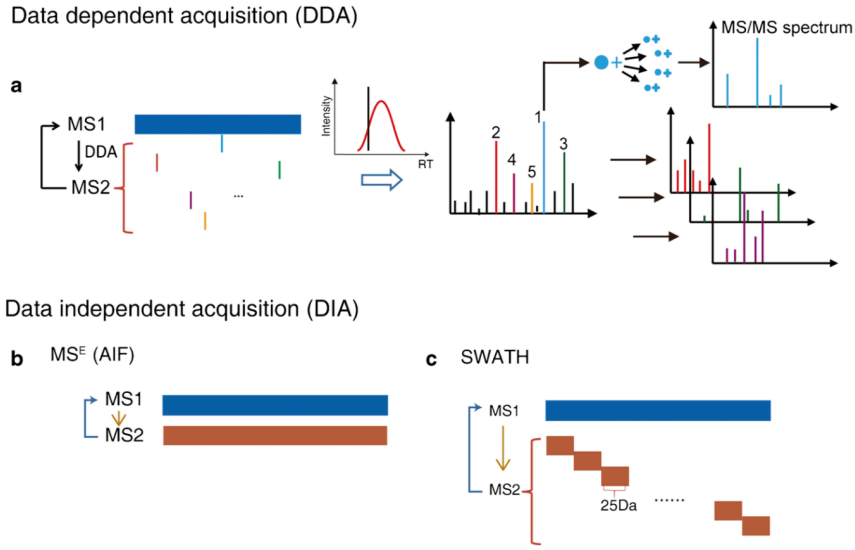


图 1: DDA and DIA MS techniques(Wang et al., 2019)

其中，DDA 技术根据一级质谱扫描时扫描带的离子的强度信息，选择强

度较高的部分离子分多次进行碎裂，每次得到于所选择单个目标离子对应的二级谱图，从而实现借助于二级谱图的代谢物鉴定。而 DIA 技术则是每次选择全部（MSE/AIF）或者部分（SWATH）母离子进行碎裂，理论上可以得到所有母离子的二级信息，然而从数据采集和记录本身无法直接获取每个母离子对应的二级谱图，因而数据分析较为复杂。综合而言，DDA 技术作为传统的非靶向 LC-MS/MS 技术，虽然受限于采集到的二级谱图覆盖范围（一般在 30%-60%），但受益于其数据分析的便利性，在代谢组学中有着广泛的应用，而 DIA 技术，虽然理论上可以获取所有母离子的二级谱图（实际覆盖范围在 90% 左右），但数据分析较为复杂，数据分析软件仍然有一定的局限性，因而在代谢组学中应用依然较少。在此，我们主要借助于 R 和 xcms 探讨 DDA 技术的数据分析原理和代谢物鉴定的方法，并介绍常用软件和代谢物鉴定数据库，以及代谢物差异性分析，如有任何错漏或者疑问，欢迎联系本人（ydrick@gmail.com[1]）

[1] mailto:ydrick@gmail.com

第一章　R 语言及环境配置

在此，我们主要基于 R 语言，并利用 XCMS 等关于代谢组学数据处理 R 包进行示例说明，以此来介绍非靶向代谢组学中 LC-MS 数据特点和数据处理方法，因而首先介绍环境配置。

1.1　R 语言环境安装

1.1.1　R 语言安装

R 语言可于官方网站[1]下载系统需要的版本，一般选择最新稳定版本。

- Windows 和 MacOS 系统请直接下载安装文件。
- Linux 系统可以使用系统的软件管理程序自动安装或者下载系统对应的二进制文件进行安装（参见官方介绍[2]，也可以通过下载源码包，自行编译安装。
- Linux 通过系统软件管理程序安装代码如下（如果需要安装最新版可能需要安装新的软件源，具体请搜索查找，在此不再列出）：

```
# Ubuntu
sudo apt-get install r-base
# CentOS/Fedora/Redhat
# (for CentOS/Redhat to add EPEL repositories)
```

[1] https://cran.r-project.org

[2] https://cran.r-project.org/bin/linux/

```
sudo yum install epel-release
sudo yum install R
```

1.1.2 Rstudio 安装

Rstudio 是目前市面上最常用的 R IDE，利用 Rstudio 可以方便的进行 R 包开发和数据分析，因而推荐使用 Rstudio 作为主要 IDE，Rstudio 可以从官方网站[3]选择系统对应的版本进行下载，且有免费版本使用，一般免费版本已经足够支持基本的开发和数据分析了

如果习惯使用 VIM、Emacs 或者 Sublime Text，亦可以找到 R 相关的插件，以便进行开发和调试，在此不再赘述。

1.1.3 Rstudio 中设置 R 包安装源

用 Rstudio 进行 R 包安装时，默认会使用 Rstudio 源 (https://cran.rstudio.com/) ，在部分地区和网络下载速度会比较慢，因而可以选择最近的 CRAN 源进行安装。

选择 Tools - Global Options, 在打开的对话框中找到 Packages，点击 Change 进行选择（Figure 1.1)

1.1.4 Linux/MacOS 中安装 R 缺失库文件

R 中有许多 R 包是调用 C/C++ 开发的，以实现复杂步骤的快速运算，提高运行效率，对这类包或者部分其他包，在 linux 或 MacOS 中进行安装时，会提示无法找到某个 lib，这种情况是系统中缺乏对应依赖的库文件缺失造成的，因而需要安装对应的库文件，然后再安装该 R 包即可。具体对于每个提示的缺失的 lib，可以搜索对应的 Error 提示信息和系统信息（如 Ubuntu）进行搜索，即可找到缺失的库文件应该通过安装哪个 lib，在此建议使用 Google 进行搜索，如果使用百度，请在搜索时使用 “R 语言” 而非 “R”。

[3] https://www.rstudio.com/products/rstudio/download/

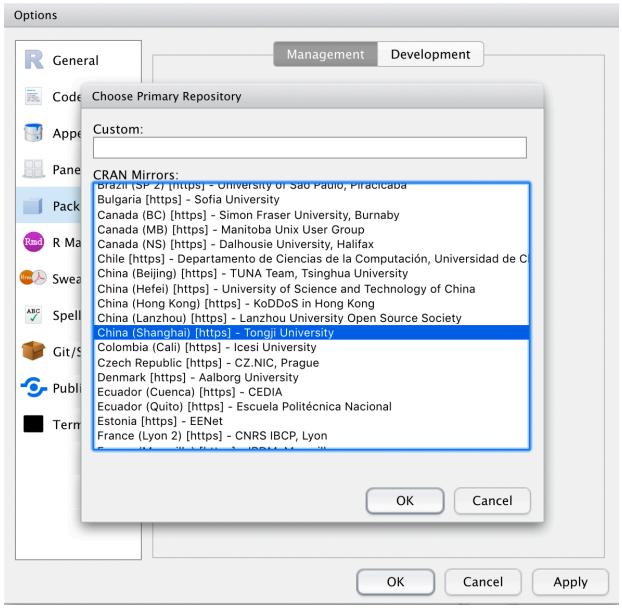


图 1.1: Change CRAN source in Rstudio

1.2 质谱数据分析相关 R 包安装

1.2.1 安装 Bioconductor

Bioconductor[4]提供了 R 语言的生物信息软件包，主要用于生物数据的注释、分析、统计、以及可视化等，最新版本的安装可以参考官方说明[5]

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install()
```

新版本的 BiocManager 包中 install 方法可以安装 Bioconductor 中的 R 包，亦可以安装 CRAN 中的 R 包，因而可以统一使用该方法进行 R 包安装。

如果是旧版本的 R，或者安装旧版本的 Biocunductor，请按照一下操作

[4] https://www.bioconductor.org

[5] https://www.bioconductor.org/install/

```
source("https://bioconductor.org/biocLite.R")
biocLite("BiocInstaller")
BiocInstaller::biocLite()
```

1.2.2 安装 XCMS 及 CAMERA 包

```
library(BiocManager)
install(c('xcms', 'CAMERA'))
```

1.2.3 安装需要用到的辅助包

```
library(BiocManager)
install(c("RColorBrewer", "faahKO", "MSnbase"))
```

1.3 代谢物鉴定软件

MS-DIAL[6]是 Windows 下免费的数据处理和代谢物鉴定软件，该软件使用 C# 编写，因而只支持 Windows 用户使用。该软件新版本自带 MoNA 谱图数据库，同时支持自定义谱图库导入，因而对常规质谱数据处理和代谢物鉴定有很大的帮助。

1.4 数据转换软件

目前仪器采集的原始数据，在利用很多厂商提供的软件之外的数据处理软件进行数据处理前需要将之转换为通用数据格式，如 mz(X)ML/ABF/netCDF 等。xcms 和 MZmine 2 等支持 mz(X)ML，

[6] http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/

MS-DIAL 需要转为 ABF。一般，我们可以使用 ProteoWizard[7]套件中的 msConvert 进行数据转换，将数据转换为文本格式，如 mzXML，MGF 等，如果使用 MS-DIAL 进行数据处理，则需要使用 Reifycs Abf Converter[8]将数据转为 Abf(Analysis Base File) 格式

1.5 数据准备

在此，进行质谱数据处理的样例数据，放在 GitHub 的 yddream/MSAnalysis[9]Repository 的 data 文件夹里，可以克隆本 Repository，直接运行 Rmd 文件，或者单独下载进行分析。

[7] http://proteowizard.sourceforge.net

[8] https://www.reifycs.com/AbfConverter/

[9] https://github.com/yddream/MSAnalysis

第二章　DDA-MS 技术概要

2.1　DDA 技术介绍

第三章　基于 XCMS 的峰检测

3.1　LC-MS 中峰检测简介

在 LC-MS 中，每一个色谱峰（peak）代表一个代谢物，峰检测 (peak detection 或 peak spotting) 主要目的是检测样品中存在的代谢物，并将多个样品中属于同一个代谢物的色谱峰归类到一起（称之为峰分组（peak grouping），分组之后的代表多个样品中同一个代谢物峰组称之为一个 feature）。

目前，最常用的峰检测算法是 centWave，该算法灵敏度高，可以自动对峰宽在指定范围内利用小波分析进行判断，从而判定峰的位置和起止点，但该方法在峰的起止点的判定上由于其算法的局限性，并不能做到非常准确的判断。

峰分组是在单个样品峰检测的基础上，考虑保留时间（RT）漂移等因素，对属于同一个代谢物的峰进行分组，在峰分组之前需要峰对齐（peak alignment），在 xcms 中最常用的峰对齐方法是 Obi-Warp，该方法利用采集到的样品数据中的 mz，intensity 和 RT 信息进行全局三维对齐，对齐之后的 RT 用于峰分组。一般 xcms 中常用峰密度分布（group.density）方法进行分组。

在 LC-MS 的峰检测过程中，但样品的峰检测、跨样品的 RT 校正（峰对齐）和分组是非常重要且必不可少的过程，缺少任何一步，都会对数据处理质量造成影响，而在某些特定条件下，Obi-Warp 方法会失效，因而在数据处理时要注意看 xcms 反馈的提示信息，并绘制 RT 漂移的曲线方便检查数据是否进行了正确的 RT 校正。

3.2 原始数据转换

MS 采集的数据，不同的仪器厂商有不同的数据记录方式，而利用 XCMS 等数据处理软件进行处理之前，需要将原始数据转换为通用的软件支持的格式，xcms 支持 xml、mzData、mzXML、mzML、netCDF 等数据格式，因而在利用 xcms 进行数据处理之前，需要对于质谱采集的原始数据进行格式转换，一般可使用 ProteoWizard 等，具体可以参考 GNPS 的文档介绍[1]，在此不再赘述。目前 ProteoWizard 数据转换支持情况可参见 Figure 3.1(GNPS, 2019)

Vendor	Instrument Software	File Format	Recommended Converter	Notes
AB Sciex	Analyst	.wiff	MSConvert	verified
Agilent	MassHunter	.d	MSConvert	verified (with issues with scan number export)
Bruker	DataAnalysis/Compass	.d	CompassXport	This conversion is through the DataAnalysis software and is detailed here
ThermoFisher	Xcalibur	.raw/.RAW	MSConvert	verified
Waters	MassLynx	.raw	MSConvert is for full scan/DDA datasets. Symphony is for other modes such as MSe/SONAR/HDMSe/HD-DDA	detailed instructions coming soon!
Shimadzu		.lcd	MSConvert	

图 3.1: Converting data to MZML/MZXML using ProteoWizard(Wang et al., 2019)

3.3 利用 XCMS 进行峰检测

xcms 峰检测有两种方法，一种是在 xcms3 中新近出现的 findChromPeaks 方法，该方法有很多新的功能和特性，并方便借用 xcms 内部的绘图方法，但目前其输出数据格式与其他的常用软件（如 CAMERA）兼容性问题；另一种是传统的 xcmsSet 方法，由于历史积累和相关软件更新，该方法目前使用仍较为广泛。以下分别介绍 xcms 中峰检测的 findChromPeaks 和 xcmsSet 方法

[1] https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/

3.3.1 xcms3 中峰检测的新方法

xcms3 的官方文档中有非常详细的数据处理方法的介绍，其中 *LCMS data preprocessing and analysis with xcms*[2]中介绍了如何利用 xcms3 的新方法进行一级质谱数据处理。

3.3.1.1 数据导入

```
require(xcms)
require(RColorBrewer)
## Get the full path to the data files(mzxml)
files <- list.files('Data', pattern = '(?i)mzxml$',
                    full.names = TRUE, recursive = TRUE)
## Create a phenodata data.frame
pd <- data.frame(sample_name = sub(basename(files),
                                   pattern = ".mzXML",
                                   replacement = "",
                                   fixed = TRUE),
         sample_group = c(rep("grp1", 2), rep("grp2", 1)),
         stringsAsFactors = FALSE)
rawData <- readMSData(files = files,
                      pdata = new("NAnnotatedDataFrame", pd),
                mode = "onDisk")
pd
```

```
##   sample_name sample_group
## 1    DDAdata1         grp1
## 2    DDAdata2         grp1
## 3    DDAdata3         grp2
```

- 注：readMSData 是 MSnbase 包中读取质谱数据的方法，返回结果为一个属于 'OnDiskMSnExp' 类的对象。

[2] https://bioconductor.org/packages/release/bioc/vignettes/xcms/inst/doc/xcms.html#6_correspondence

3.3.1.2 利用 centWave 算法进行峰检测

```
# xcms 3 new methods
cwp <- CentWaveParam(peakwidth = c(5, 50),
                     noise = 1000, snthresh = 10)
xdata <- findChromPeaks(rawData, param = cwp)
head(xdata@msFeatureData$chromPeaks)
```

```
##           mz mzmin mzmax    rt rtmin rtmax  into  intb
## CP0001 702.2 702.2 702.2 45.65 43.14 48.05  8450  8446
##         maxo   sn sample
## CP0001  5619 5618      1
##  [ reached getOption("max.print") -- omitted 5 rows ]
```

下图展示检测到的一个峰的示例（Figure @ref(fig:pd_showPeaks)）和在样品中的 EIC 信息（Figure @ref(fig:pd_showEIC)）

```
## Define the rt and m/z range of the peak area
rtr <- c(375, 400)
mzr <- c(132.07604, 132.07639)
## extract the chromatogram
chrRaw <- chromatogram(xdata, mz = mzr, rt = rtr)
cls <- paste0(brewer.pal(3, "Set1")[1:2], "60")
names(cls) <- c("grp1", "grp2")
plot(chrRaw, col = cls[chrRaw$sample_group], peakBg = cls[chrRaw$sample_group]
```

```
library(magrittr)
# plot EIC/XIC
xdata %>%
    filterRt(rt = rtr) %>%
    filterMz(mz = mzr) %>%
    plot(type = "XIC")
```

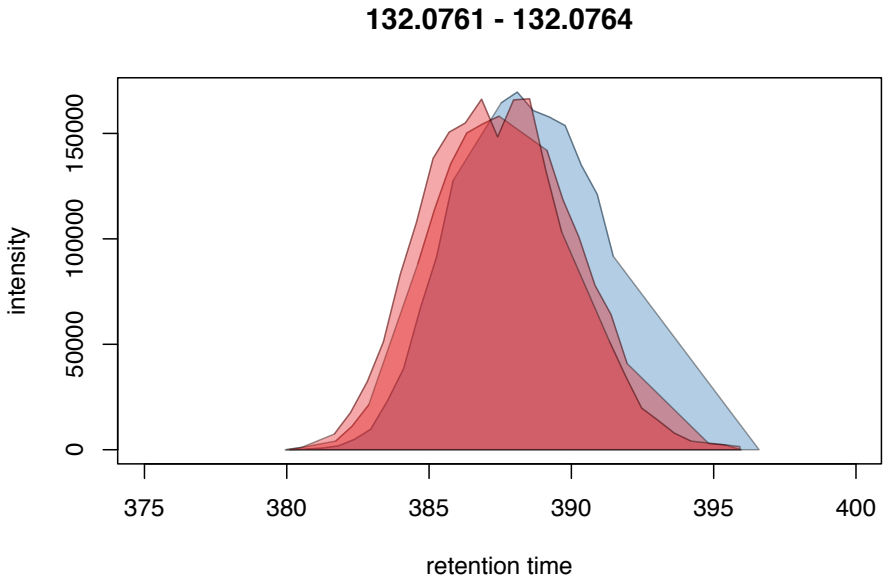


图 3.2: Extracted ion chromatogram for one peak

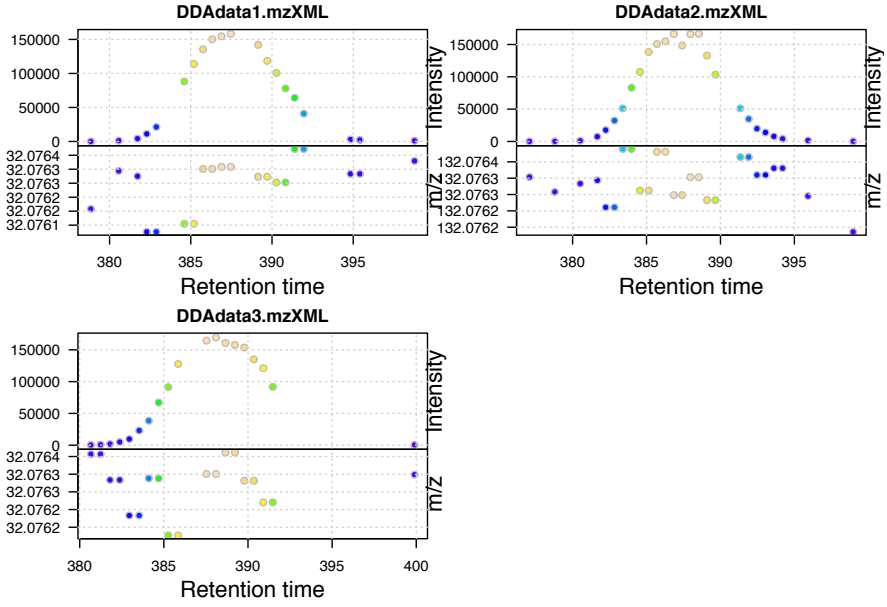


图 3.3: Visualization of the raw MS data for one peak

3.3.1.3 峰对齐及分组 (peak alignment & grouping)

同一代谢物在不同样品的流出时间会略有差异，因而对于峰检测的结果需要进行对齐 (alignment) 并将不同样品中的统一代谢物分到各自可以表征该代谢物的峰组 (peak group) 中 (代谢组学中一般成为 feature)，从而进一步比较不同样品间统一代谢物的含量。通常我们可以使用'obiwarp' 算法进行 peak alignment，然后利用'density' 算法进行 peak grouping.

```
## Correspondence: group peaks across samples.
pdp <- PeakDensityParam(sampleGroups = xdata$sample_group,
            minFraction = 1)
xdata <- groupChromPeaks(xdata, param = pdp)
## Now the retention time correction.
pgp <- PeakGroupsParam(minFraction = 1)
## Get the peak groups that would be used for alignment.
# (grouping with 'peakgroup' method)
xdata <- adjustRtime(xdata, param = pgp)
# otherwise, one can also use 'Obi-warp' method for alignment
xdata <- adjustRtime(xdata, param = ObiwarpParam(binSize = 0.1))
## Grouping with RT corrected peaks
pdp <- PeakDensityParam(sampleGroups = xdata$sample_group,
                        minFraction = 0.4, bw = 20)
xdata <- groupChromPeaks(xdata, param = pdp)
```

峰补齐 (filling gaps) 对于峰检测过程中会有部分 feature 在某些样品中未检出对应代谢物峰的清醒，xcms 可以根据已检出 feature 的信息，在相应样品中强行提取 EIC 信息，从而计算该代谢物在该样品中的含量信息，我们一般称之为 filling gaps

```
xdata <- fillChromPeaks(xdata)
```

3.3.2 xcms 中传统的峰检测方法

```
require(xcms)
## Get the full path to the data files(mzxml)
files <- list.files('Data', pattern = '(?i)mzxml$',
                    full.names = TRUE, recursive = TRUE)
xdata <- xcmsSet(files, method = 'centWave',
                 peakwidth = c(5, 50), noise= 1000, snthr = 10)
# peak alignment & groups
# peakgroups method
# xdata <- xcms::group(xdata, minfrac = 1)
# xdata<- retcor(xdata, method = 'peakgroups', plottype = 'deviation')
# or obiwarp method
xdata<- retcor(xdata, method = 'obiwarp', plottype = 'deviation',
               profStep = 0.1)
```

```
## center sample:  DDAdata3
## Processing: DDAdata1  DDAdata2
```

```
xdata<- xcms::group(xdata, bw = 20, minfrac = 1)
# fill gaps
xdata <-fillPeaks(xdata)
# saveRDS(xdata, file = "Data/xset.Rda")
```

补充说明 findChromPeaks 是 xcms 中用于进行峰检测的新方法，输入值为'OnDiskMSnExp' 对象和峰检测参数对象，CentWaveParam 用以创建'CentWaveParam' 对象，该对象设置使用'centWave' 算法进行峰检测时所需要的参数，其中比较常用的参数如下：- ppm – 峰检测时 MS1 的 m/z tolerance，以 ppm 为单位 - peakwidth – 长度为 2 的向量，设置峰检测时峰宽范围，事实上该参数对应的每个峰可以跨越多少个质谱检

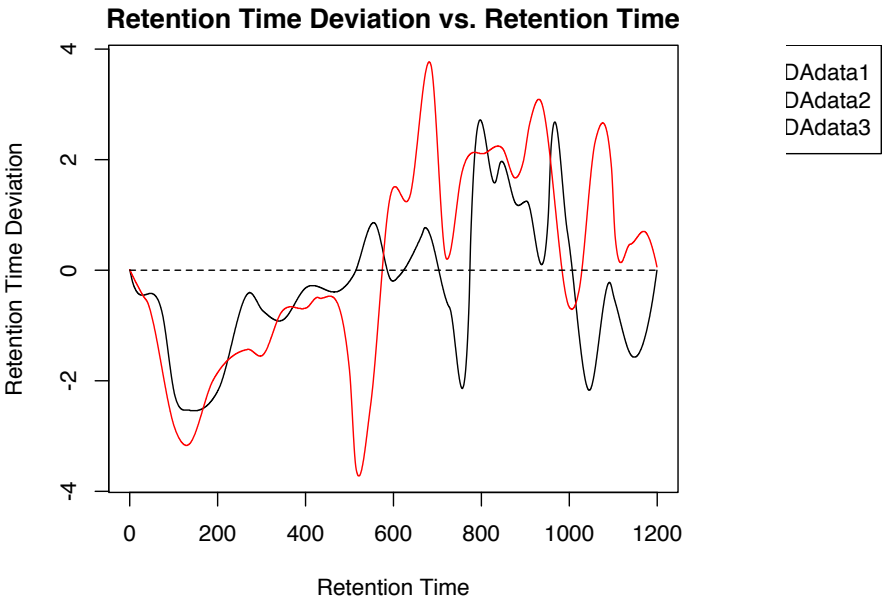


图 3.4: RT correction plot

测的 scan，而非多少秒 - snthresh – 峰检测时信噪比要求 - return.type – 返回数据类型，可以根据要求返回'XCMSnExp' 类数据（默认）、传统的'xcmsSet' 类以及'list'

xcmsSet 是 xcms 中可以用于峰检测的传统方法，使用 method 参数设置峰检测算法，另外根据不同的峰检测算法可以设置该算法需要的参数，详情请参考 xcms 官方文档。其中'centWave' 算法所对应的参数与'CentWaveParam' 设置基本一致。

峰检测时除了可以使用 centWave 算法外，还可以使用其他算法，如'centWaveWithPredIsoROIs'、'massifquant'、'matchedFilter'、'MSW' 等，分别对应'CentWavePredIsoParam'、'MassifquantParam'、'MatchedFilterParam'、'MSWParam' 参数设置，详情请参考 xcms 官方文档或者使用 R help

3.4 利用 CAMERA 进行峰注释

借助 LC-MS 技术检测得到的代谢物峰包含了大量的同位素峰和加合物峰信息，对数据的进一步造成一定干扰，因而做好同位素和加合物峰的注释对代谢组学数据分析具有重要的辅助作用。常见的注释工具有 CAMERA 和 RAMCluster 等，在这里主要介绍基于 CAMERA 的代谢物峰注释基本操作和其中的一些注意事项。

3.4.1 CAMERA 包介绍

CAMERA 是一个 Bioconductor R 包，主要用于 LC-MS 数据中代谢物峰的注释，从而标记峰与峰之间的同位素和加合物关系，具体原理请参考 Carsten Kuhl 等人 2012 年发表于分析化学 (Analytical Chemistory) 上的文章 (Kuhl et al., 2012)。该 R 包具体相关信息可以参考 Bioconductor 网站[3]，关于利用 CAMERA 进行注释的详细方法和例子可以参考最新文档[4]，

3.4.1.1 注释前的预备工作

峰检测完成后则可以开始 CAMERA 注释的操作了。在正式注释同位素和加合物峰之前，需要对峰检测数据进行进一步的处理，创建 CAMERA 注释对象，并进行分组并检验分组结果

```
require(CAMERA)
#Create an xsAnnotate object
xa <- xsAnnotate(xdata)
#Group after RT value of the xcms grouped peak
xag <- groupFWHM(xa, perfwhm=0.6)
```

```
## Start grouping after retention time.
## Created 107 pseudospectra.
```

[3] https://bioconductor.org/packages/release/bioc/html/CAMERA.html

[4] https://bioconductor.org/packages/release/bioc/vignettes/CAMERA/inst/doc/CAMERA.pdf

```
#Verify grouping
xac <- groupCorr(xag)
```

```
## Start grouping after correlation.
## Generating EIC's ..
##
## Calculating peak correlations in 107 Groups...
##  % finished: 10  20  30  40  50  60  70  80  90  100
##
## Calculating graph cross linking in 107 Groups...
##  % finished: 10  20  30  40  50  60  70  80  90  100
## New number of ps-groups:  166
## xsAnnotate has now 166 groups, instead of 107
```

3.4.1.2 同位素峰注释

```
#Annotate isotopes, could be done before groupCorr
xac.isotope <- findIsotopes(xac)
```

```
## Generating peak matrix!
## Run isotope peak annotation
##  % finished: 10  20  30  40  50  60  70  80  90  100
## Found isotopes: 85
```

3.4.1.3 加合物峰注释

在同位素峰注释完成之后，才可以进行加合物峰的注释。

```
#Annotate adducts
xac.addu <- CAMERA::findAdducts(xac.isotope, polarity="positive")
```

```
## Generating peak matrix for peak annotation!
```

```
##
## Calculating possible adducts in 166 Groups...
##  % finished: 10  20  30  40  50  60  70  80  90  100
```

在这里需要注意的是，加合物的注释可以自己给定加合物注释的规则(rules)。事实上，CAMERA 本身内置了非常多的加合物形式的规则，但对于不同的 LC 体系，产生的加合物形式会有所不同，因而，对于特定的实验来说，最好根据大家对实验中采用的 LC 体系的了解，自己制定针对性的注释规则。对于 CAMERA 中的加合物注释规则，大家可以使用以下代码提取，在自己制定规则的时候，可以用做参考。

```
# list all rule files (for positive/negative modes, primary/extende rules
files <- list.files(system.file('rules', package = "CAMERA"), full.names
# show head lines of sample rule
head(read.csv(files[4]))
```

```
##      name nmol charge massdiff oidscore quasi ips
## 1  [M+H]+    1      1    1.007        1     1   1
## 2 [M+Na]+    1      1   22.989        8     1   1
##  [ reached 'max' / getOption("max.print") -- omitted 2 rows ]
```

对于以上每一列表示的具体含义，可以参考使用文档 ‘Create rule table’ 章节（Bioconductor v3.9 在 14 页，Section 6[5]）

如果需要使用自定义的规则，请参考以下代码：

```
my.rules <- read.csv(files[4])
xac.addu <- CAMERA::findAdducts(xac.isotope, rules = my.rules, polarity='
```

```
## Generating peak matrix for peak annotation!
## Found and use user-defined ruleset!
## Calculating possible adducts in 166 Groups...
##  % finished: 10  20  30  40  50  60  70  80  90  100
```

[5] https://bioconductor.org/packages/release/bioc/vignettes/CAMERA/inst/doc/CAMERA.pdf

注：polarity 请根据自己实验采集数据时的离子模式正确设置，在不设置自定义规则的时候，CAMERA 会调用内置的与该离子模式相同的规则（包含 primary 和 extended）进行注释。

3.4.1.4 注释结果表格输出

```
#Get final peaktable and store on harddrive
res.anno <- CAMERA::getPeaklist(xac.addu)
saveRDS(res.anno, file = "Data/AnnoRes.Rda")
# write.csv(res.anno,file="Result.csv")
knitr::include_graphics("Figures/Fig_CameraResExample.png")
```

	mz	mzmin	mzmax	rt	rtmin	rtmax	npeaks	grp1	grp2	DDAdata1	DDAdata2	DDAdata3	isotopes	adduct
23	119.08869	119.08859	119.08878	250.0900	249.860	250.320	2	2	0	25481.443	13174.968	10191.065	[2][M+1]+	
24	121.05023	121.05021	121.05041	114.0050	113.264	114.542	3	2	1	377525.612	390875.089	381991.686	[3][M]+	
25	122.05311	122.05308	122.05344	113.9700	113.264	114.568	3	2	1	21633.115	24040.768	23048.281	[3][M+1]+	
26	122.08056	122.08054	122.08057	250.1170	249.860	250.320	3	2	1	282178.059	277576.183	302236.345	[4][M]+	[M+H]+ 121.072
27	123.05496	123.05484	123.05500	79.2390	78.779	80.127	3	2	1	398050.806	440850.890	437804.789	[5][M]+	
28	123.08397	123.08374	123.08400	250.1170	249.860	250.320	3	2	1	14258.256	12570.398	12159.480	[4][M+1]+	
29	124.05747	124.05694	124.05753	79.4960	78.779	80.380	3	2	1	24410.468	28131.603	32761.744	[5][M+1]+	

Figure 1展示了 CAMERA 注释的结果，对于 adduct 列，我们看到的最后部分的数字表示该加合物形式对应的 M 的质荷比，方括号内显示的是 M 的加合物形式，方括号后面的 '+' 号则表示该加合物形式的电荷数量。而同位素注释结果中最前面的方括号表示该同位素所对应的 id，id 相同的峰为同一个 M 峰的不同同位素峰，第二个方括号中的 M+x 表示同位素峰中的同位素情况，方括号外面的内容则表示电荷情况。pcgroup 列则是在预处理时 CAMERA 生成的 group 信息

3.4.2 小结

峰检测是质谱数据处理的基础，利用峰检测可以获取样品中检测到的信号的基本信息，如 mz、RT、峰面积等，这些信息是代谢物鉴定和差异分析的基础，利用 xcms 进行峰检测，并借助于 CAMERA 进行峰注释，可以有效的检测到样品中含有的代谢物信息，并去除部分冗余。除了 xcms，还有很多同样优秀的质谱数据处理分析的软件，如 OpenMS，MZmine 2，MS-DIAL 等，在此暂不做详细介绍，如有兴趣请参考相应软件的官方说明。

第四章　二级谱图提取

4.1　DDA 数据里的二级谱图

DDA 采集模式在采集二级 (MSMS or MS2) 谱图时，会按照母离子丰度从高代低依次采集，然后记录采集谱图时目标的母离子的 m/z 信息和采集时间，但无法记录采集时该 MS1 离子的强度 (intensity)，因而，可以根据记录的母离子 m/z 和采集时间，与对峰检测中获取的峰进行对应，从而获取代表采集到的代谢物的二级谱图信息，进而通过谱图比对的方式进行代谢物鉴定，同时利用峰检测结果进行定量。甚至可以只用二级谱图的信息，直接通过谱图比对，进行代谢物鉴定。

要获取二级谱图信息，最简单的方法是利用 ProteoWizard 中的 msConvert 直接进行数据转换，只提取 MS2 信息生成 MGF 格式数据，再读取 MGF 文本信息，从而在程序里获得 MS2 数据，进行后续处理；另一种方法是直接在 xcms 中直接读取转换好的 mzXML 文件，并从中得到二级谱图信息，或者根据一级峰检测结果，从二级质谱数据中提取与之对应的谱图数据。

4.2　利用 ProteoWizard 进行二级谱图提取

MSConvert 转换数据到 MGF 时，设置如 Figure 4.1所示。转换完之后，即可得到对应的 MGF 文件 (Figure 4.2)

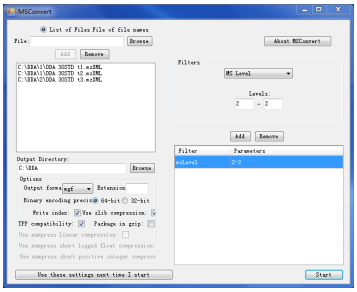


图 4.1: Converting mzXML to MGF with msConvert

```
BEGIN IONS
TITLE=DDA 30STD t1.22.22. File:"", NativeID:"scan=22"
RTINSECONDS=5.293
PEPMASS=188.175061974837
43.02180368 42.0
55.05989628 42.0
58.06416287 63.0
72.04667737 170.0
72.08140525 958.0
72.1113498 22.0
84.08266907 84.0
99.0902388 43.0
100.0770446 818.0
112.1043353 42.0
112.1132971 84.0
114.0875596 21.0
117.1011064 42.0
128.0432609 21.0
129.0332423 21.0
171.1546199 104.0
188.0703937 21.0
188.1806795 63.0
END IONS
```

图 4.2: Example of MGF data record

4.2.1 谱图提取

- 利用 R 读取 MGF 文件，并获取 MS2 谱图信息。

```
source("Code/ReadMGF.R")
files <- list.files(path = "Data", pattern = "(?i).MGF$",
                    recursive = TRUE, full.names = TRUE)
mgf.data <- ReadMGF(files[1])
spec.info <- plyr::ldply(mgf.data, `[[`, "info")
head(spec.info)
```

```
##       mz    rt
## 1 188.2 5.293
## 2 237.2 5.343
## 3 247.2 5.393
## 4 309.0 5.443
```

```
## 5 340.1 5.493
## 6 367.1 5.543
```

```
spec.all <- lapply(mgf.data, `[[`, "spec")
plot(spec.all[[2682]], type = "h", col = "red")
abline(h=0)
```

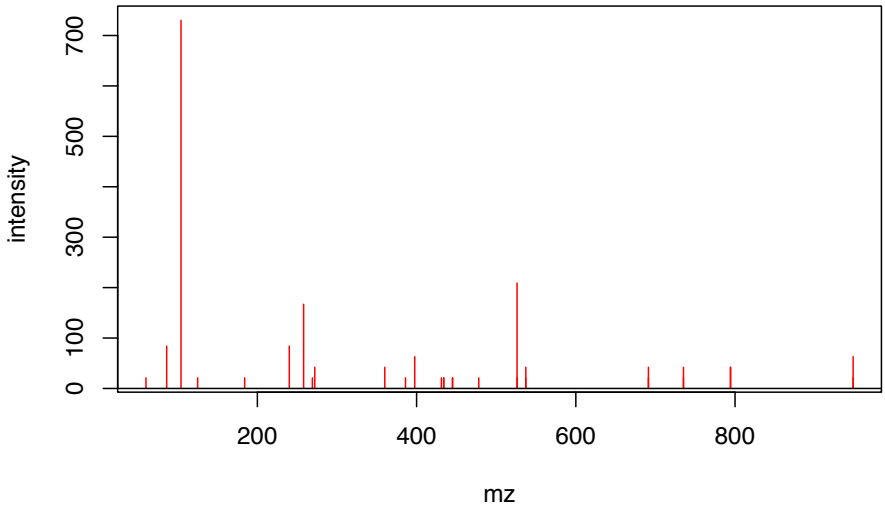


图 4.3: Example of extracted MS2 spectrum

- 对谱图信息进行去噪音处理仪器记录的原始 MS2 谱图信息含有很多噪音 (white noise)，不同的仪器有不同的噪音水平，可以根据记录的强度信息大致做出评判，一般在我们使用的这组数据中，采用 30 作为噪音的基准值 (Figure 4.4)，对于低于该水平的二级碎片，要清除掉。另外，在记录的 MS2 谱图里面，也会包含有高于母离子 m/z 的碎片出现，这些碎片也要清除掉 (Figure **??**)。

```
frag.all <- do.call(rbind, spec.all)
plot(density(frag.all[frag.all[, 2] <= 200, 2]),
     main = "Density of intensities (<=200)")
```

```
spec.denoise <- lapply(seq_along(spec.all), function(idx) {
  spec <- spec.all[[idx]]
```

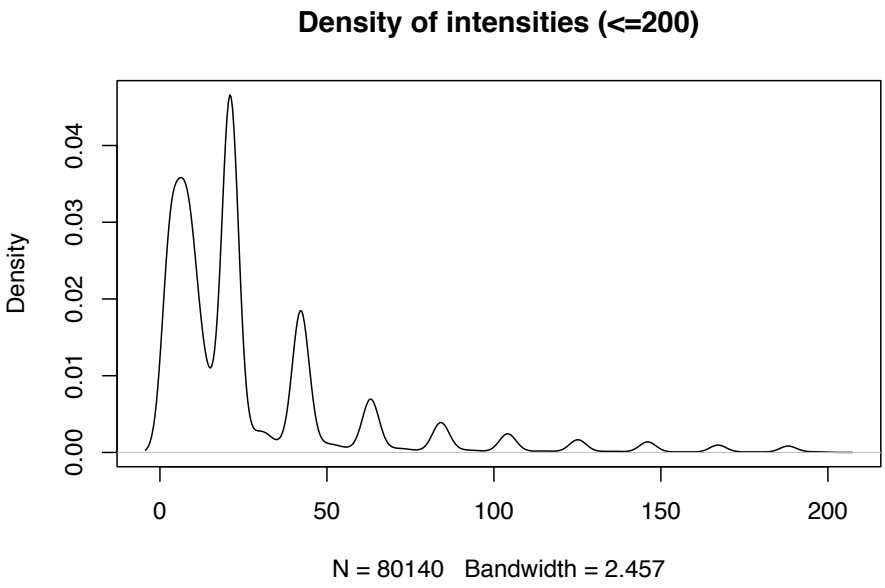
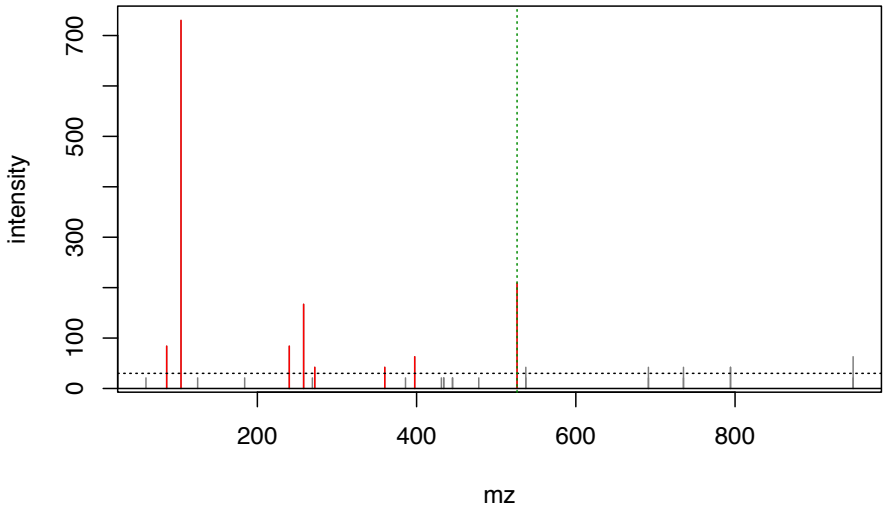


图 4.4: Density of MS2 spectra intensities

```
  spec <- spec[spec[, 2] >= 30, , drop = FALSE]
  mz.precursor <- spec.info[idx, "mz"]
  spec <- spec[spec[, 1] <= mz.precursor, , drop = FALSE]
})
plot(spec.all[[2682]], type = "h", col = "gray50")
lines(spec.denoise[[2682]], type = "h", col = "red")
abline(h=0)
abline(h=30, lty = 3)
abline(v=spec.info[2682, "mz"], lty = 3, col = "green4")
```

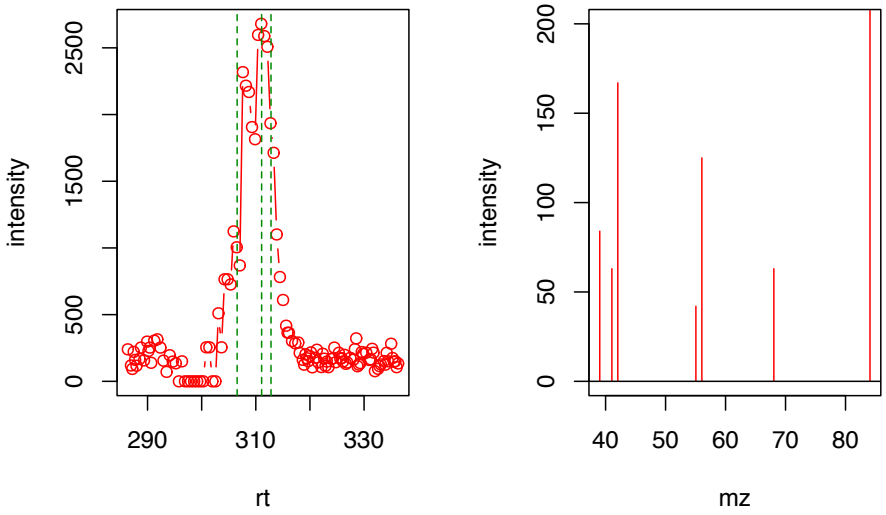


与峰检测结果进行结合根据提取出来的 MS2 谱图的 m/z 和 rt 信息，与之前得到的峰检测结果进行匹配，将匹配上的，保留为与检测到的峰对应的 MS2 谱图。对同一个代谢物，有可能打到多张 MS2 谱图，此时，要选择其中某一张谱图最为标准或者用合适的方法合并这两张谱图，从而得到在样品中采集到的二级谱图信息

```
pk.info <- readRDS("Data/AnnoRes.Rda")
mz.pk <- pk.info[, "mz"]
mz.spec <- spec.info[, "mz"]
rt.spec <- spec.info[, "rt"]
idx.spec <- apply(pk.info, 1, function(dr) {
  mz <- as.numeric(dr["mz"])
  rt <- as.numeric(dr["rt"])
  idx <- which(mz.spec >= mz - 0.01 & mz.spec <= mz + 0.01 &
                 rt.spec >= rt - 5 & rt.spec <= rt + 5)
})

require(xcms)
xset <- readRDS("Data/xset.Rda")
eic <- getEIC(xset, groupidx = 7, rtrange = 50,
                    sampleidx=sampnames(xset)[1], rt = "raw")
par(mfrow=c(1,2))
```

```
plot(eic@eic$DDAdata1[[1]], type = "b", col = "red")
# points(eic@eic$DDAdata1[[1]], col = "gray")
abline(v=rt.spec[idx.spec[[7]]], col = "green4", lty = 2)
plot(spec.denoise[[idx.spec[[7]][2]]], type = "h", col = "red", ylim = c(0, 20
abline(h=0)
```



利用 XCMS 进行二级谱图提取

4.3 小结

第五章　代谢物鉴定

代谢组学最重要的目标之一，就是获取样品中有哪些代谢物，并知道该代谢物的含量信息。因而代谢物的鉴定，成为在峰检测之后最为重要的工作，只有在获取准确的代谢物鉴定结果的基础上，才可以准确的做接下来的生物信息学的各项分析。

5.1　代谢物鉴定方法介绍

通过与标准谱图库里的 MS2 谱图进行比对，从而获取采集到的代谢物的信息是最常用且有效的方法，对于少量的数据，可以用在线的代谢物鉴定工具进行处理即可，如 METLIN，MassBank，MoNA 等。另外也可以通过理论谱图进行鉴定，如 CFM-ID，MetFrag 等。

对于植物代谢物，GNPS[1]谱图库（GNPS Public Spectral Libraries[2]）含有大量的自然产物信息，GNPS 网站支持批量通过谱图比对进行代谢物鉴定，也是非常不错的代谢物鉴定工具。除此之外，GNPS 包含了很多植物代谢组学和自然产物相关的数据处理分析的工具，如有兴趣，可以自行去官方网站了解学习，在此不多赘述。

特别要注意的是，在代谢物鉴定的时候，一定要选择与自己数据采集时相对应的仪器平台的标准谱图进行比较。这是因为，在 LC-MS 中采集到的 MS2 谱图，不同的仪器平台会有很大的差异，Q-TOF 和 Obitrap 的谱图在大多数情况下显著不同，同时，即便同一平台，不同仪器之间的谱图也会略有差异，但仍可用于代谢物鉴定，只是对鉴定的可信度略有

[1] https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp

[2] https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp

影响。而在 MassBank/MoNA 谱图库中，包含了来自所有仪器平台的谱图信息，因而，在利用该谱图库进行代谢物鉴定时一定要非常注意。另外，METLIN 数据库很多时候会屏蔽部分 IP 地址，因而国内很多地方可能无法使用，对我们会造成不便，目前美国的 IP 地址基本可以访问。

5.2 常用鉴定工具

5.2.1 利用 MS-DIAL 进行 DDA 数据处理并进行代谢物鉴定

MS-DIAL 软件的官方文档[3]有较详细的说明，

5.3 自行进行代谢物鉴定

由于很多代谢物谱图库包含了大量冗余信息，包含很多与自己数据采集时仪器平台不一致的代谢物谱图，同时对鉴定结果不易批量处理，因而，自行进行代谢物鉴定也是非常不错的途径，且可以自行定制代谢物鉴定的比对方法，考虑更多的因素到代谢物鉴定当中来，通过脚本进行自动化代谢物鉴定也是非常常见的方式。

对于利用谱图比对的方式进行代谢物鉴定，首先需要有谱图相似度的比对方法，并对相似度作出正确评判，目前最常用的相似度判断方法为点积 (Dot Product)，可以根据不同的目的和谱图库特性选择是正向 (Forward) 还是反向 (Reverse) 匹配进行计算。一般对于利用标准品采集获取标准谱图库，我们认为标准谱图库是精准的，可以严格作为判定标准，因而将标准谱图库做正向匹配，获得 Forward Dot Product，从而进行代谢物鉴定。

[3] https://mtbinfo-team.github.io/mtbinfo.github.io/MS-DIAL/tutorial

第六章　基本差异分析

6.1　代谢物差异分析简介

6.2　Fold change

6.3　基本统计检验

6.3.1　Students' t-Test

6.3.2　Wilcoxon Test

6.3.3　小结

第七章　主成分分析（PCA）

7.1　PCA 原理及简介

7.2　利用 R 进行 PCA 分析

7.3　结果解读

第八章　PLS-DA 和 OPLS-DA

8.1　PLS-DA

8.2　OPLS-DA

附录 A　余音绕梁

呐，到这里朕的书差不多写完了，但还有几句话要交待，所以开个附录，再啰嗦几句，各位客官稍安勿躁、扶稳坐好。

参考文献

GNPS (2019). Mass spectrometry file conversion. https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/. Accessed September 2, 2019.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289.

Wang, R., Yin, Y., and Zhu, Z.-J. (2019). Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology. *Analytical and bioanalytical chemistry*, 411(19):4349–4357.