Meta-Workflow

Miao YU

2020-06-30

Contents

Preface

This is an online handout for mass spectrometry based metabolomics data analysis. It would cover a full reproducible metabolomics workflow for data analysis and important topics related to metabolomics. Here is a list of topics:

- Sample collection
- Pretreatment
- Principles of metabolomics data analysis
- Software selection
- Batch correction
- Annotation
- Omics analysis
- Exposome

This is a book written in **Bookdown**. You could contribute it by a pull request in Github.

R and **Rstudio** are the softwares needed in this workflow.

Chapter 1

Introduction

Information in living organism commuicates along the Central Dogma in different scales from individual, population, community to ecosystem. Metabolomics (i.e., the profiling and quantitation of metabolites) is a relatively new field of “omics” studies. Different from other omics studies, metabolomics always focused on small moleculars with much lower mass than polypeptide with single or doubled charged ions. Here is a demo of the position of metabolomics in “omics” studies(?).

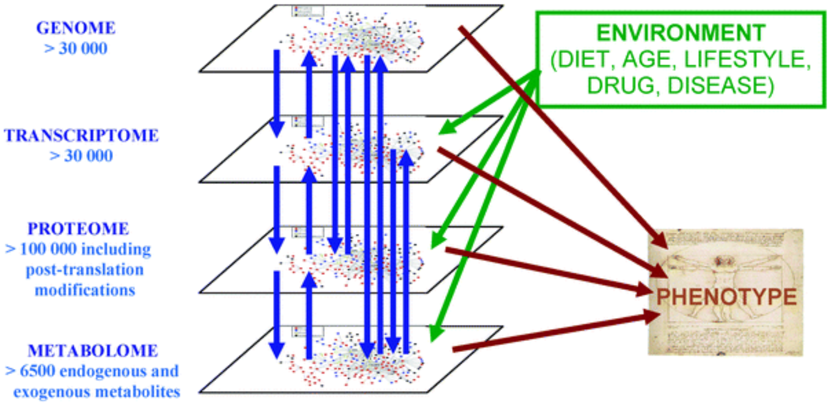

GENOME
> 30 000
TRANSCRIPTOME
> 30 000
PROTEOME
> 100 000 including post-translation modifications
METABOLOME
> 6500 endogenous and exogenous metabolites
ENVIRONMENT
(DIET, AGE, LIFESTYLE, DRUG, DISEASE)
PHENOTYPE

Figure 1.1: The complex interactions of functional levels in biological systems.

Metabolomics studies always employ GC-MS(?), GC*GC-MS(?), LC-MS(?), LC-MS/MS(?) or NMR(??) to measure metabolites. However, this workflow will only cover mass spectrometry based metabolomics or XC-MS based research.

1.1 History

1.1.1 History of Mass Spectrometry

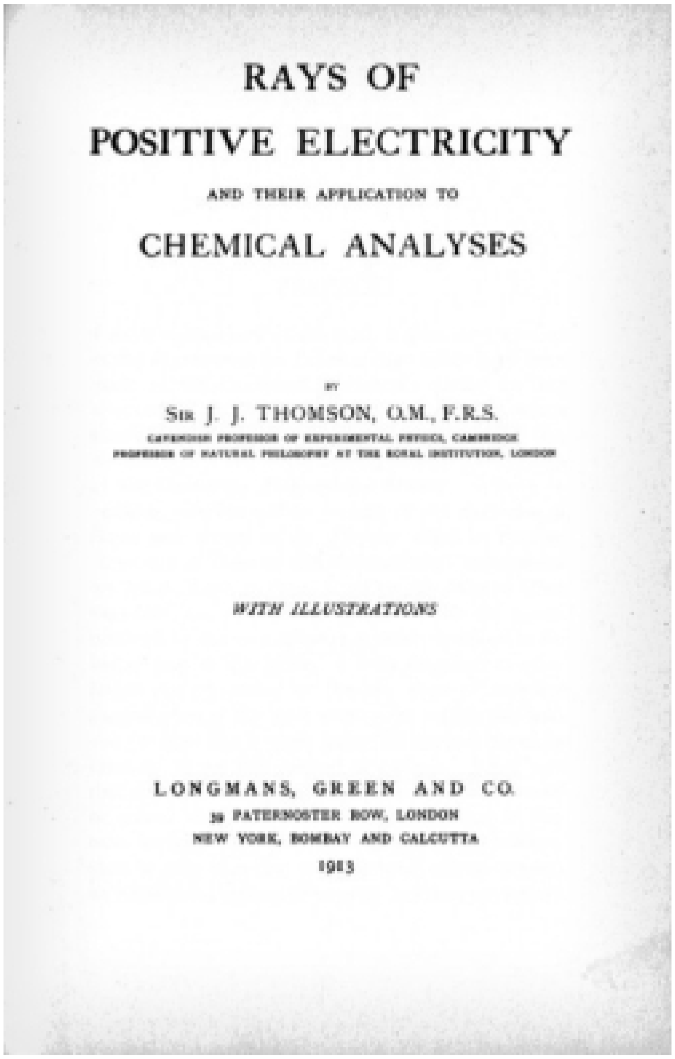
- 1913, Sir Joseph John Thomson “Rays of Positive Electricity and Their Application to Chemical Analyses.”

- Petroleum industry bring mass spectrometry from physics to chemistry
- The first commercial mass spectrometer is from Consolidated Engineering Corp to analysis simple gas mixtures from petroleum
- In World War II, U.S. use mass spectrometer to separate and enrich isotopes of uranium in Manhattan Project
- U.S. also use mass spectrometer for organic compounds during wartime and extend the application of mass spectrometer
- 1946, TOF, William E. Stephens
- 1970s, quadrupole mass analyzer
- 1970s, R. Graham Cooks developed mass-analyzed ion kinetic energy spectrometry, or MIKES to make MRM analysis for multi-stage mass sepctrometry
- 1980s, MALDI rescue TOF and mass spectrometry move into biological application
- 1990s, Orbitrap mass spectrometry
- 2010s, Aperture Coding mass spectrometry

1.1.2 History of Metabolomcis

According to this book section(**?**):

- 2000-1500 BC some traditional Chinese doctors who began to evaluate the glucose level in urine of diabetic patients using ants
- 300 BC ancient Egypt and Greece that traditionally determine the urine taste to diagnose human diseases
- 1913 Joseph John Thomson and Francis William Aston mass spectrometry
- 1946 Felix Bloch and Edward Purcell Nuclear magnetic resonance
- late 1960s chromatographic separation technique



RAYS OF

POSITIVE ELECTRICITY

AND THEIR APPLICATION TO

CHEMICAL ANALYSES

BY

SIR J. J. THOMSON, O.M., F.R.S.

CAVENDISH PROFESSOR OF EXPERIMENTAL PHYSICS, CAMBRIDGE
PROFESSOR OF NATURAL PHILOSOPHY AT THE ROYAL INSTITUTION, LONDON

WITH ILLUSTRATIONS

LONGMANS, GREEN AND CO.
39 PATERNOSTER ROW, LONDON
NEW YORK, BOMBAY AND CALCUTTA
1913

Figure 1.2: Sir Joseph John Thomson ”Rays of Positive Electricity and Their Application to Chemical Analyses.”

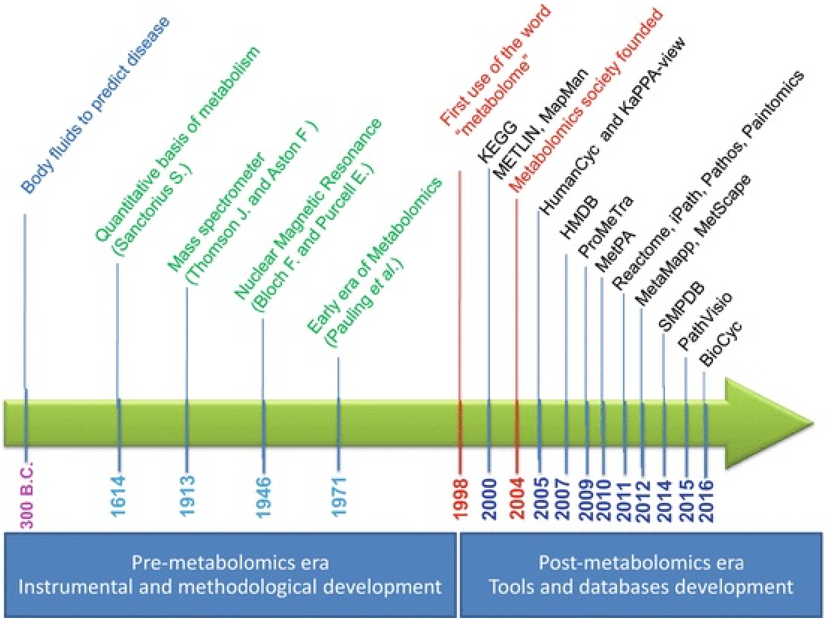

Body fluids to predict disease
Quantitative basis of metabolism (Sanctorius S.)
Mass spectrometer (Thomson J. and Aston F.)
Nuclear Magnetic Resonance (Bloch F. and Purcell E.)
Early era of Metabolomics (Pauling et al.)
First use of the word "metabolome"
KEGG
METLIN, MapMan
Metabolomics society founded
HumanCyc and KaPPA-view
HMDB
ProMeTra
MetPA
Reactome, iPath, Pathos, Paintomics
MetaMapp, MetScape
SMPDB
PathVisio
BioCyc
300 B.C.
1614
1913
1946
1971
1998
2000
2004
2005
2007
2009
2010
2011
2012
2014
2015
2016
Pre-metabolomics era
Instrumental and methodological development
Post-metabolomics era
Tools and databases development

Figure 1.3: Metabolomics timeline during pre- and post-metabolomics era

- 1971 Pauling's research team "Quantitative Analysis of Urine Vapor and Breath by Gas–Liquid Partition Chromatography"
- Willmitzer and his research team pioneer group in metabolomics which suggested the promotion of the metabolomics field and its potential applications from agriculture to medicine and other related areas in the biological sciences
- 2007 Human Metabolome Project consists of databases of approximately 2500 metabolites, 1200 drugs, and 3500 food components
- post-metabolomics era high-throughput analytical techniques

1.1.3 Defination

Metabolomics is actually a comprehensive analysis with identification and quantification of both known and unknown compounds in an unbiased way. Metabolic fingerprinting is working on fast classification of samples based on metabolite data without quantifying or identification of the metabolites. Metabolite profiling always need a pre-defined metabolites list to be quantification(**?**). However, targeted and untargeted metabolomics are also used in publicaitons. A similar concept called non-targeted analysis/screen is actually describe the similar studies or workflow.

1.2 Reviews and tutorials

Some nice reviews and tutorials related to this workflow could be found in those papers or directly online:

1.2.1 Workflow

Those papers are recommended(**????**) for general metabolomics related topics. For targeted metaabolomics, you could check those reviews(**??????**).

1.2.2 Data analysis

You could firstly read those papers(**?????**) to get the concepts and issues for data analysis in metabolomics. Then this paper(**?**) could be treated as a step-by-step tutorial.

- For annotation, this paper(**?**) is an well organized review.
- For database used in metabolomics, you could check this review(**?**).
- For metabolomics software, check this series of reviews for each year(**???**).
- For open sourced software, this review(**?**) could be a good start.
- For DIA or DDA metabolomics, check those papers(**??**).

Here is the slides for metabolomics data analysis workshop and I have made presentations twice in UWaterloo and UC Irvine.

- Introduction
- Statistical Analysis
- Batch Correction
- Annotation
- Demo

1.2.3 Application

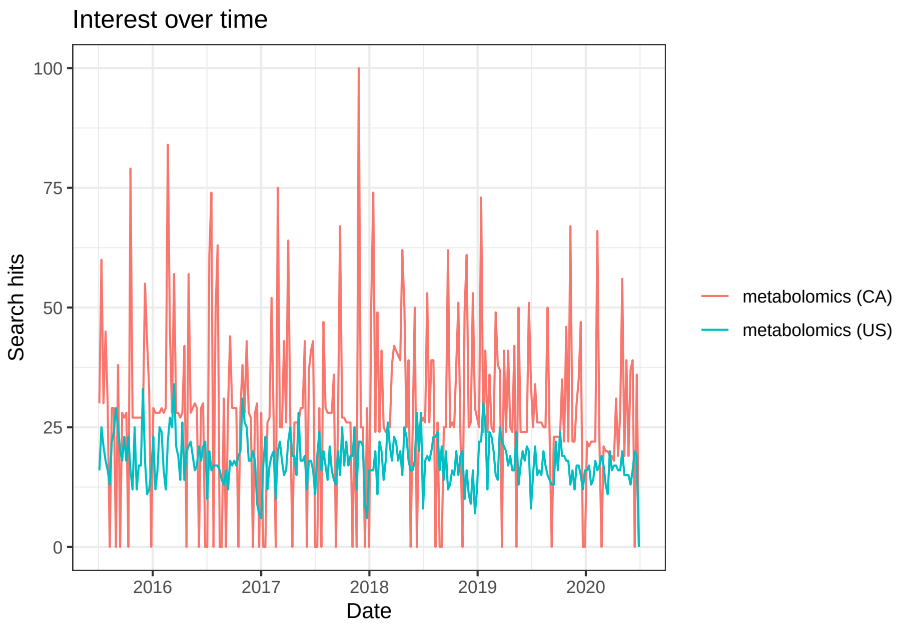
- For environmental research related metabolomics or exposome, check those papers(**??**).
- For food chemistry, check this(**?**) and this paper for livestock(**?**) and those one for nutrients(**??**)
- For disease related metabolomics such as oncology(**??**), ophthalmology(**?**), Cardiovascular(**?**) and chronic kidney disease(**?**), check those papers. This paper(**?**) cover the metabolomics realted clinic research.
- Check this piece(**?**) for drug discovery and precision medicine
- The object could be plant(**??**), microbial and mammalian(**?**), brain(**?**), human gut microbiota(**?**).
- For single cell metabolomics analysis, check here(**??**).

1.2.4 Challenge

- High throughput Metabolomics related issues could be found here (**?**). - Cohort size - Temporal resolution - Spatial resolution
- Quantitative Metabolomics related issues could be found here(**??**).
- For quality control issues, check here(**??**).

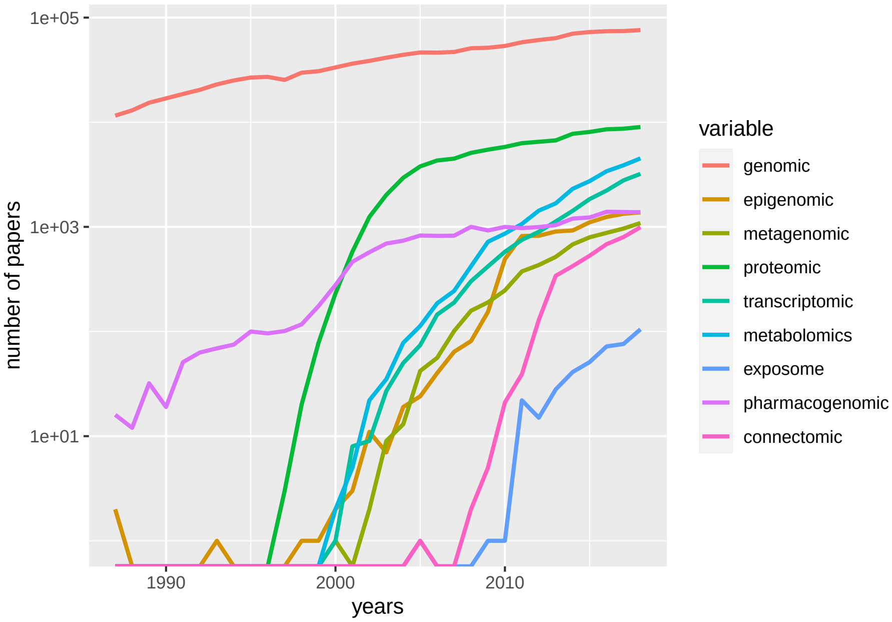
1.3 Trends in Metabolomics

```r
library(gtrendsR)
res <- gtrends(c("metabolomics", "metabolomics"), geo = c("CA","US"))
plot(res)
```



```
library(rentrez)
papers_by_year <- function(years, search_term){
    return(sapply(years, function(y) entrez_search(db="pubmed",term=search_term, mindate=y, maxda
}
years <- 1987:2018
total_papers <- papers_by_year(years, "")
omics <- c("genomic", "epigenomic",  "metagenomic", "proteomic", "transcriptomic","metabolomics",
trend_data <- sapply(omics, function(t) papers_by_year(years, t))
trend_props <- trend_data/total_papers
library(reshape)
library(ggplot2)
trend_df <- melt(data.frame(years, trend_data), id.vars="years")
p <- ggplot(trend_df, aes(years, value, colour=variable))
p + geom_line(size=1) + scale_y_log10("number of papers")
```

```
## Warning: Transformation introduced infinite values in
## continuous y-axis
```


1e+05
1e+03
1e+01
number of papers
1990
2000
2010
years
variable
genomic
epigenomic
metagenomic
proteomic
transcriptomic
metabolomics
exposome
pharmacogenomic
connectomic

1.4 Workflow

Chapter 2

Experimental design(DoE)

Before you perform any metabolomics studies, a clean and meaningful experimental design is the best start. Depending on different research purposes, experiental design can be classified into homogeneity and heterogeneity study. Technique such as isotope labeled media will not be discussed in this chapter while this paper(**?**) could be a good start.

2.1 Homogeneity study

In homogeneity study, the research purpose is about method validation in most cases. Pooled sample made from multiple samples or technical replicates from same population will be used. Variances within the samples should be attibuted to factors other than the samples themselves. For example, one wants to test if sample injection order will affect the intensities of the unknown peaks. One pooled sample or technical replicates samples could be used here and the variances of the intensities should not be designed from heterogenety samples.

Another experimental design for homogeneity study will use biological replicates to find the common feasures from a group of samples. Biological replicates mean samples from same populatio with same biological process. For example, we wanted to know metabolites profiles of a certain species and we could collected lots of the individual samples from the same species. Then only the peaks/compounds appeared in all samples will be used to describe the metabolites profiles of this species. Technical replicates could also be used with biological replicates.

2.2 Heterogeneity study

In heterogeneity study, the research purpose is to find the differents among samples. To get the heterogeneity, you need at least a baseline to perform the comparision. Such baseline could be generated by random process or control samples or background knowledge. For example, outlier detection can be performed to find abnormal samples in unsupervised manners. Distribution or spatial analysis could be used to find geological relationship of known and unknown compounds. Temporal trend of metabolites profile could be found by time series or cohort studies. Clinical trial or random control trial is also an important class of heterogeneity studies. In this cases, you need at least two groups: treated group and control group. Also you could treat this group infomation as the one primary variable or primary variables to be explored for certain research purposes. In the following discussion about experimental design, we will use random control trail as model to discuss important issues.

2.3 Sample size

Supporsing we have control and treated groups, the numbers of samples in each group should be carefully calculated.For each metabolite, such comparision could be treated as one t-test. You need to perform a Power analysis to get the numbers. For example, we have two groups of samples with 10 samples in each group. Then we set the power at 0.9, which means one minus Type II error probability, the standard deviation at 1 and the significance level(Type 1 error probability) at 0.05. Then we get the meanful delta between the two groups should be higher than 1.53367 under this experiment design. Also we could set the delta to get the minimized numbers of the samples in each group. To get those data such as the standard deviation or delta for power analysis, you need to perform prelimitary or pilot experiments.

```
power.t.test(n=10,sd=1,sig.level = 0.05,power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 1.53367
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
power.t.test(delta = 5,sd=1,sig.level = 0.05,power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 2.328877
##           delta = 5
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

However, since sometimes we could not perform preliminary experiment, we could directly compute the power based on false discovery rate control. If the power is lower than certain value, say 0.8, we just exclude this peak as significant features. In this review (**?**), author suggest to estimate an average α according to this equation (**?**) and then use normal way to calculate the sample numbers:

$$\alpha_{ave} \leq (1-\beta_{ave}) \cdot q \frac{1}{1+(1-q) \cdot m_0/m_1}$$

Other study (**?**) show a method based on simulation to estimate the sample size. They used BY correction to limit the influences from correlations. However, the nature of omics study make the power analysis hard to use one numbers for all metabolites and all the methods are trying to find a balance to represent more peaks with least samples(save money).

If there are other co-factors, a linear model or randomizing would be applied to eliminated their influences. You need to record the values of those co-factors for further data analysis. Common co-factors in metabolomics studies are age, gender, location, etc.

If you need data correction, some background or calibration samples are required. However, control samples could also be used for data correction in certain DoE.

Another important factors are instrumentals. High-resolution mass spectrum is always preferred. As shown in Lukas's study (**?**):

> the most effective mass resolving powers for profiling analyses of metabolite rich biofluids on the Orbitrap Elite were around 60000–120000 fwhm to retrieve the highest amount of information. The region between 400–800 m/z was influenced the most by resolution.

However, elimination of peaks with high RSD% within group were always omited by most study. Based on pre-experiment, you could get a description of RSD% distribution and set cut-off to use stable peaks for further data analysis. To my knowledge, 50% is suitable considering the batch effects.

2.4 Software

- MetSizeR GUI Tool for Estimating Sample Sizes for metabolomics Experiments.

Chapter 3

Pretreatment

Pretreatment will affect the results of metabolomics and cover the sample treatment from crude samples to injection vials. Sample pretreatment try to retain more interesting compounds while remove unrelated compounds. For metabolomics studies, we might not know 'interesting' compounds in advance and the unrelated compounds are highly depended on research purpose. For example, Gel Permeation Chromatograph(GPC), Florisil, Alumina, Silica gel could be used to remove lipid while alcohols and strong acid/base could make protein denaturation to release more compounds. However, if we are interested in lipid or protein, such pretreatment methods should be changed. In general, sample quenching, extraction methods, derivatization, and storage should be optimized in pretreatment.

3.1 Quenching

Quenching solvent is always used to stop stop enzymatic activity.

In this review(**?**), authors said:

> A classical approach, which works well for many analytes, is boiling ethanol. Although the boiling solvent raises concerns about thermal degradation, it reliably denatures enzymes. In contrast, cold organic solvent may not fully denature enzymes or may do so too slowly such that some metabolic reactions continue, interconverting metabolites during the quenching process.

This review(**?**) summarised the urease-dependent metabolome sample preparation and found:

> activities of urease and endogenous urinary enzymes and metabolite contaminants from the urease preparations introduce artefacts into metabolite profiles, thus leading to misinterpretation.

3.2 Extraction

According to this research(?):

> The total metabolome concentration is approximately 300 mM, whereas the protein concentration is approximately 7 mM., which implies that most cellular metabolites are in free form.

Dmitri et.al(?) thought the most orthogonal methods to methanol-based precipitation were ion-exchange solid-phase extraction and liquid-liquid extraction using methyl-tertbutyl ether.

Tissue samples need to first be pulverized into fine powders.

Feces collected with 95% ethanol or FOBT would be more reproducible and stable.

In this review(?), authors said:

> In our experience, for both cell and tissue specimens, 40:40:20 acetonitrile:methanol:water with 0.1 M formic acid (and subsequent neutralization with ammonium bicarbonate) is generally an effective solvent system for both quenching and extraction, including for ATP and other high-energy phosphorylated compounds. We typically use approximately 1 mL of solvent mix to extract 25 mg of biological specimen. ...Thus, although drying is acceptable for most metabolites, care must be taken with redox-active species.

(?) nano LC-MS could be used to analysis small numbers of cells.

For plant like soybeans(?), ammonium acetate/methanol could be selected as extraction strategies compared with water/methanol and sodium phosphate/methanol.

3.3 Derivatization

Derivatization is always used in GC-based metabolomics study. This paper(?) compared sequential derivatization methods and found different compounds would show different fluctuations during oximation or silylation process.

3.4 Storage

Samples should be stored after sample collection or sample pretreatment. -80°C or -20°C is always preferred to store samples. Dry ice should be used during sample pretreatment. However, comprehensive investigation of storage influnces found the metabolites profile will change after one day storage at -80°C. Rapid analysis of samples should be considered to capture more accurate information in the samples.

Chapter 4

Instrumental analysis

To get more infomation in the samples, full scan is perferred on GC/LC-MS. Each scan would generat a mass spectrum to cover the setting mass range. If you narrow down your mass range and keep the same scan time, each mass would gain the collection time and you would get a higher sensitivity. However, if you expand your scan range, the sensitivity for each mass would decrease. You could also extend the collection time for each scan. However, it would affect the seperation process.

Full scan is performed synchronously with the seperation process. For a better seperation on chromotograph, each peak should have at least 10 point to get a nice peak shape. If you want to seperate two peaks with a retention time differences of 10s. Assuming the half peak width is 5s, you need to collect 10 mass spectrum within 10s. So the drwell time for each scan is 1s. If you use a high resolution column and the half peak width is 1s, you need to finish a scan within 0.2s. As we talked above, shorter drwell time would decrease the sensitivity. Thus there is a trade-off between seperation and sensitivity. If you use UPLC, the seperation could be finished within 20 min while you need to calculate if you mass spectrumetry could still show a good sensitivity.

4.1 Column and gradient selection

For GC, higher temperature could release compounds with higher boiling point. For LC, gradient and functional groups of stationary phase would be more important than temperature. Polarity of samples and column should match. More polar solvent could release polar compounds. Normal-phase column will not retain non-polar compounds while reversed-phase will elute polar column in the very beginning. To cover a wide polarity range or logP value compounds, normal phase column should match with non-polar to polar gradient to get a better seperation of polar compounds while reverse phase column should match with polar to non-polar gradient to elute compounds. If you use a inappropricate order of gradient, you compounds would not be seperated well. If you have no idea about column and gradient selection, literature's condiation might help.

4.2 Pooled QC samples

Pooled QC samples are unique and very important for metabolomics study. Every 10 or 20 samples, a pooled sample from all samples and blank sample in one study should be injected as quality control samples. Pooled QC samples contain the changes during the instrumental analysis and blank samples could tell where the variances come from. Meanwhile the cap of sequence should old the column with pooled QC samples. The injection sequence should be randomized. Those papers(**??**) should be read for details.

4.3 Mass resolution

For metabolomics, high resolution mass spectrum should be used to make identification of compounds easier. The Mass Resolving Power is very important for annotation and high resolution mass spectrum should be calibrated in real time. The region between 400–800 m/z was influenced the most by resolution(**?**). Orbitrap Fusion's performance was evaluated here(**?**).

Chapter 5

Raw data pretreatment

Raw data from the instruments such as LC-MS or GC-MS were hard to be analyzed. To make it clear, the structure of those data could be summarised as:

- Indexed scan with time-stamp
- Each scan contains a full scan mass spectra

Commen formats for open source mass spectrum data are mzxml, mzml or CDF. However, **masscomp** might shrink the data size(**?**).

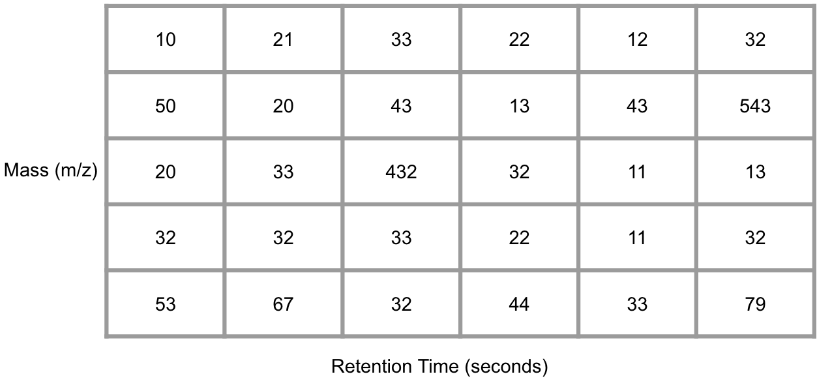
5.1 Data visualization

You could use msxpertsuite for MS data visualization. It is biological mass spectrometry data visualization and mining with full JavaScript ability(**?**).

5.2 Peak extraction

GC/LC-MS data are usually be shown as a matrix with column standing for retention times and row standing for masses after bin them into small cell.

Conversation from the mass-retention time matrix into a vector with selected MS peaks at certain retention time is the basic idea of Peak extraction. You could EIC for each mass to charge ratio and use the change of trace slope to determine whether there is a peak or not. Then we could make integration for this peak and get peak area and retention time.



	10	21	33	22	12	32
	50	20	43	13	43	543
Mass (m/z)	20	33	432	32	11	13
	32	32	33	22	11	32
	53	67	32	44	33	79

Retention Time (seconds)

Figure 5.1: Demo of GC/LC-MS data

```r
intensity <- c(10,10,10,10,10,14,19,25,30,33,26,21,16,12,11,10,9,10,11,10)
time <- c(1:20)
plot(intensity~time, type = 'o', main = 'EIC')
```

However, due to the accuracy of instrument, the detected mass to charge ratio would have some shift and EIC would fail if different scan get the intensity from different mass to charge ratio.

In the `matchedfilter` algorithm(**?**), they solve this issue by bin the data in m/z dimension. The adjacent chromatographic slices could be combined to find a clean signal fitting fixed second-derivative Gaussian with full width at half-maximum (fwhm) of 30s to find peaks with about 1.5-4 times the signal peak width. The the integration is performed on the fitted area.

The `Centwave` algorithm(**?**) based on detection of regions of interest(ROI) and the following Continuous Wavelet Transform (CWT) is preferred for high-resolution mass spectrum. ROI means a regine with stable mass for a certain time. When we find the ROIs, the peak shape is evaluated and ROI could be extended if needed. This algotithm use `prefilter` to accelerate the processing speed. `prefilter` with 3 and 100 means the ROI should contain 3 scan with intensity above 100. Centwave use a peak width range which should be checked on pool QC. Another important parameter is `ppm`. It is the maximum allowed deviation between scans when locating regions of interest (ROIs), which is different from vendor number and you need to extend them larger than the company claimed. For `profparam`, it's used for fill peaks or align peaks instead of peak picking. `snthr` is the cutoff of signal to noise ratio.

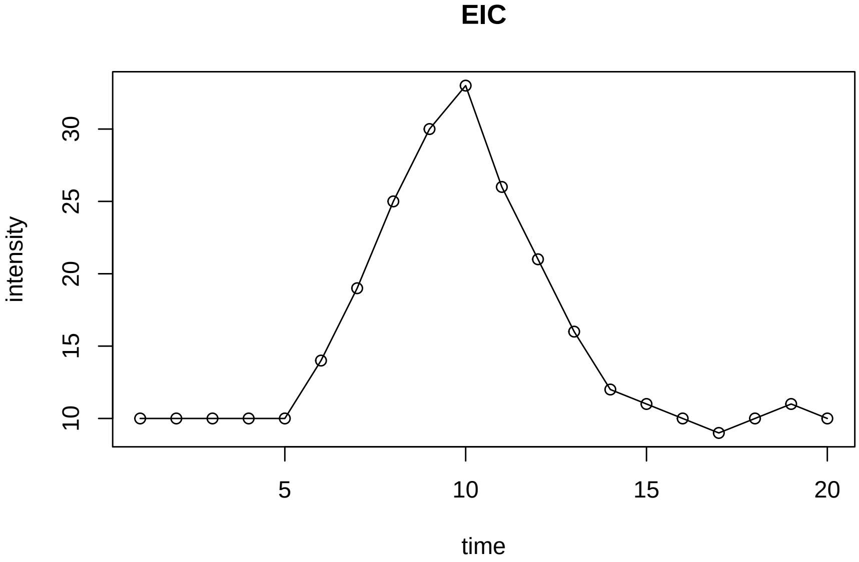


Figure 5.2: Demo of EIC with peak

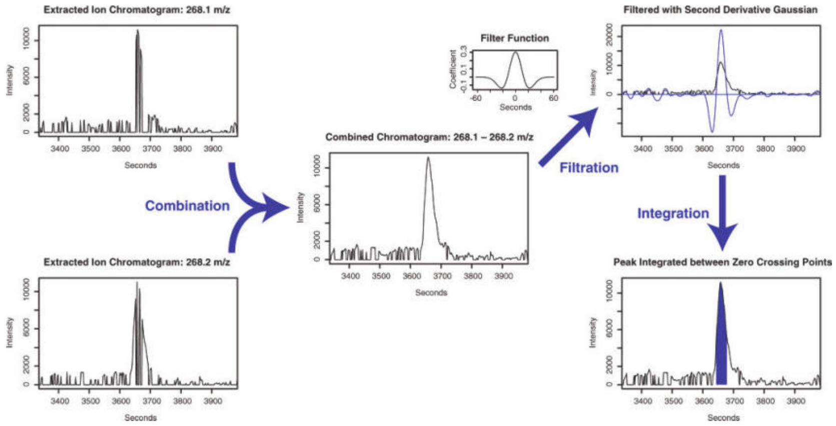

Extracted Ion Chromatogram: 268.1 m/z
Extracted Ion Chromatogram: 268.2 m/z
Combination
Combined Chromatogram: 268.1 – 268.2 m/z
Filter Function
Coefficient
Seconds
Filtration
Filtered with Second Derivative Gaussian
Integration
Peak Integrated between Zero Crossing Points
Intensity
Seconds
3400
3500
3600
3700
3800
3900

Figure 5.3: Demo of matchedfilter

5.3 Retention Time Correction

For single file, we could get peaks. However, we should make the peaks align across samples for subsquece analysis and retention time corrections should be performed. The basic idea behind retention time correction is that use the high quality grouped peaks to make a new retention time. You might choose `obiwarp`(for dramatic shifts) or loess regression(fast) method to get the corrected retention time for all of the samples. Remember the original retention times might be changed and you might need cross-correct the data. After the correction, you could group the peaks again for a better cross-sample peaks list. However, if you directly use `obiwarp`, you don't have to group peaks before correction.

(**?**) show a matlab based shift correction methods.

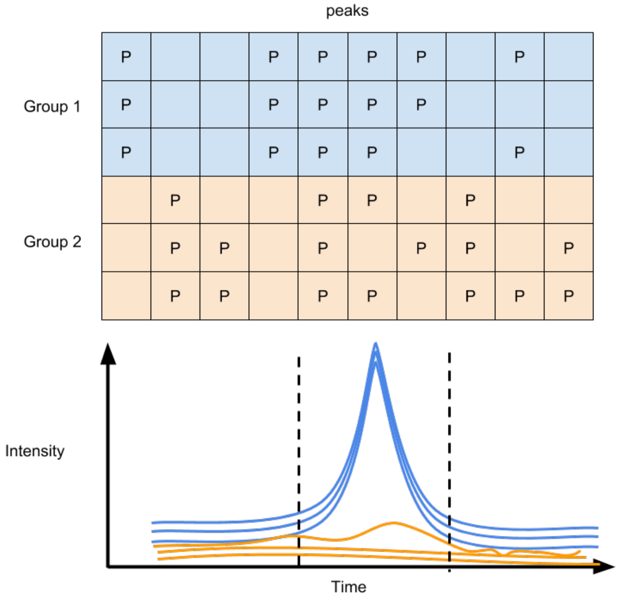
5.4 Filling missing values

Too many zeros or NA in peaks list are problematic for statistics. Then we usually need to integreate the area exsiting a peak. `xcms 3` could use profile matrix to fill the blank. They also have function to impute the NA data by replace missing values with a proportion of the row minimum or random numbers based on the row minimum. It depends on the user to select imputation methods as well as control the minimum fraction of featuers appeared in single group.

With many groups of samples, you will get another data matrix with column standing for peaks at cerntain retention time and row standing for samples after the Raw data pretreatment.

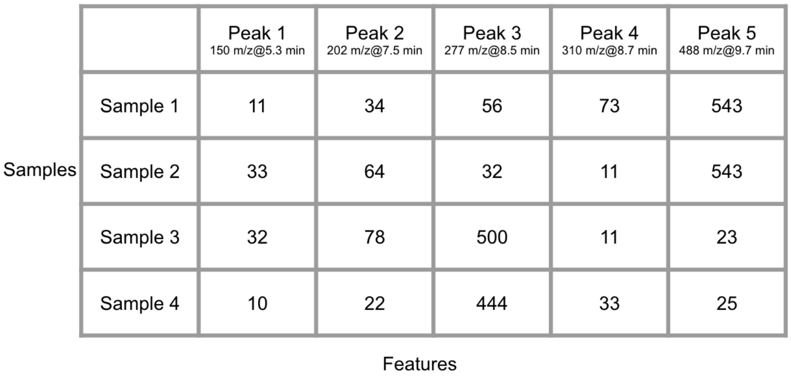
5.5 Spectral deconvolution

Without fracmental infomation about certain compound, the peak extraction would suffer influnces from other compounds. At the same retention time, co-elute compounds might share similar mass. Hard electron ionization methods such as electron impact ionization (EI), APPI suffer this issue. So it would be hard to distighuish the co-elute peaks' origin and deconvolution method(**?**) could be used to seperate different groups according to the similar chromatogragh beheviors. Another computational tool **eRah** could be a better solution for the whole process(**?**). Also the **ADAD-GC3.0** could also be helpful for such issue(**?**).



Group 1	P			P	P	P	P		P	
	P			P	P	P	P			
	P			P	P	P			P	
Group 2		P			P	P		P		
		P	P		P		P	P		P
		P	P		P	P		P	P	P

Figure 5.4: Peak filling of GC/LC-MS data



Samples	Peak 1 150 m/z@5.3 min	Peak 2 202 m/z@7.5 min	Peak 3 277 m/z@8.5 min	Peak 4 310 m/z@8.7 min	Peak 5 488 m/z@9.7 min
Sample 1	11	34	56	73	543
Sample 2	33	64	32	11	543
Sample 3	32	78	500	11	23
Sample 4	10	22	444	33	25

Features

Figure 5.5: Demo of many GC/LC-MS data

5.6 Dynamic Range

Another issue is the Dynamic Range. For metabolomics, peaks could be below the detection limit or over the detection limit. Such Dynamic range issues might raise the loss of information.

5.6.1 Non-detects

Some of the data were limited by the detect of limitation. Thus we need some methods to impute the data if we don't want to lose information by deleting the NA or 0.

Two major imputation way could be used. The first way is use model-free method such as half the minimum of the values across the data, 0, 1, mean/median across the data(`enviGCMS` package could do this via `getimputation` function). The second way is use model-based method such as linear model, random forest, KNN, PCA. Try `simputation` package for various imputation methods. As mentioned before, you could also use `imputeRowMin` or `imputeRowMinRand` within `xcms` package to perform imputation.

Tobit regression is preferred for censored data. Also you might choose maximum likelihood estimation(Estimation of mean and standard deviation by MLE. Creating 10 complete samples. Pool the results from 10 individual analyses).

```r
x <- rnorm(1000,1)
x[x<0] <- 0
y <- x*10+1
library(AER)
tfit <- tobit(y ~ x, left = 0)
summary(tfit)
```

```
##
## Call:
## tobit(formula = y ~ x, left = 0)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           1000              0           1000              0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0000     0.4310    2.32   0.0203 *
## x            10.0000     0.3162   31.62   <2e-16 ***
## Log(scale)    2.1501     0.0000     Inf   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 8.586
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 1
## Log-likelihood: -3069 on 3 Df
## Wald-statistic:  1000 on 1 Df, p-value: < 2.22e-16
```

According to Ronald Hites's simulation(**?**), measurements below the LOD (even missing measurements) with the LOD/2 or with the $LOD/\sqrt{2}$ causes little bias and "Any time you have a % non-detected >20%, for whatever reason, it is unlikely that the data set can give useful results."

5.6.2 Over Detection Limit

CorrectOverloadedPeaks could be used to correct the Peaks Exceeding the Detection Limit issue(**?**).

5.7 RSD/fold change Filter

Some peaks need to be rule out due to high RSD% and small fold changes compared with blank samples.

5.8 Power Analysis Filter

As shown in [Exprimental design(DoE)], the power analysis in metabolomics is ad-hoc since you don't know too much before you perform the experiment. However, we could perform power analysis after the experiment done. That is, we just rule out the peaks with a lower power in exsit Exprimental design.

5.9 Software

5.9.1 Peak picking

- ProteoWizard Toolkit provides a set of open-source, cross-platform software libraries and tools (**?**). Msconvert is one tool in this toolkit.
- xcms LC/MS and GC/MS Data Analysis(**?**)

- apLCMS Generate peaks list (?)
- x13cms global tracking of isotopic labels in untargeted metabolomics (?)
- FTMSVisualization is a suite of tools for visualizing complex mixture FT-MS data(?)
- MZmine is an open-source software for mass-spectrometry data processing, with the main focus on LC-MS data(?)
- MS-DAIL is a universal program for untargeted metabolomics- and lipidomics supporting any type of chromatography/mass spectrometry methods (GC/MS, GC-MS/MS, LC/MS, and LC-MS/MS etc.) (?)
- OpenMS is an open-source software C++ library for LC/MS data management and analyses(?)
- MZmatch is a Java collection of small commandline tools specific for metabolomics MS data analysis (??)
- iMet-Q is an automated tool with friendly user interfaces for quantifying metabolites in full-scan liquid chromatography-mass spectrometry (LC-MS) data.(?)
- MAVEN is an open source cross platform metabolomics data analyser.(?)

5.9.2 For MS/MS

- MS-DAIL for data independent MS/MS deconvolution of comprehensive metabolome analysis.(?)
- decoMS2 An Untargeted Metabolomic Workflow to Improve Structural Characterization of Metabolites(?)
- msPurity Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics(?)
- ULSA Deconvolution algorithm and a universal library search algorithm (ULSA) for the analysis of complex spectra generated via data-independent acquisition based on Matlab (?)

5.9.3 Improved Peak picking

- IPO A Tool for automated Optimization of XCMS Parameters(?).
- Warpgroup is used for chromatogram subregion detection, consensus integration bound determination and accurate missing value integration(?)
- xMSanalyzer improved Peak picking for xcms and apLCMS(?)

- ms-flo A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing(**?**)

5.10 Tips

- Convert XCMSnExp object into xcmsSet object

```
xcmsSetdemo <- as(XCMSnExpdemo,'xcmsSet')
```

- Split xcmsSet with multiple groups object into a list with single group object

```
list <- split(xcmsSetdemo,xcmsSetdemo@phenoData$sample_group)
# re-group the peaks with parallel computation
list2 <- BiocParallel::bplapply(list,group)
```

- Combine single group xcmsSet objects into one xcmsSet

```
xcmsSetdemoagain <- Reduce('c',list2)
# for more higher order function in R, check here: http://www.johnmyleswhite.com/notebook/2010/09
```

- Use **warpgroup** to get peaks with better quality

```
library(doParallel)
cl = makeCluster(detectCores() - 1)
registerDoParallel(cl)

xseteiclist = lapply(xcmsSetdemo@filepaths, xcmsRaw, profstep=0)
xwarpgroup = group.warpgroup(xcmsSetdemo, xseteiclist, rt.max.drift = 20, ppm.max.drift = 3, rt.a
```

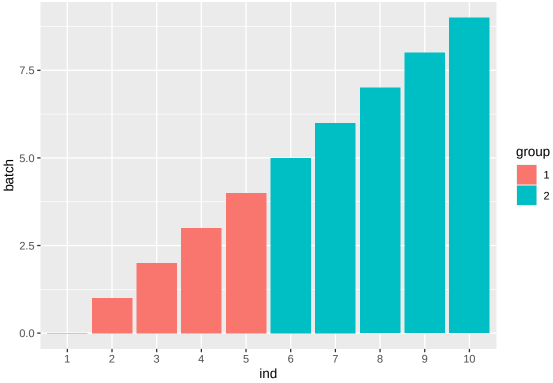
Chapter 6

Peaks normalization

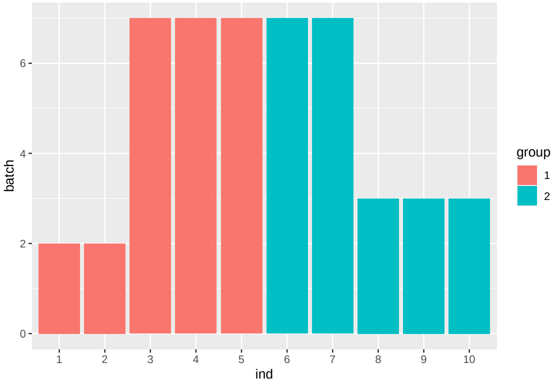
6.1 Batch effects classification

Variances among the samples across all the extracted peaks might be affected by factors other than the experiment design. There are three types of those batch effects: Monotone, Block and Mixed.

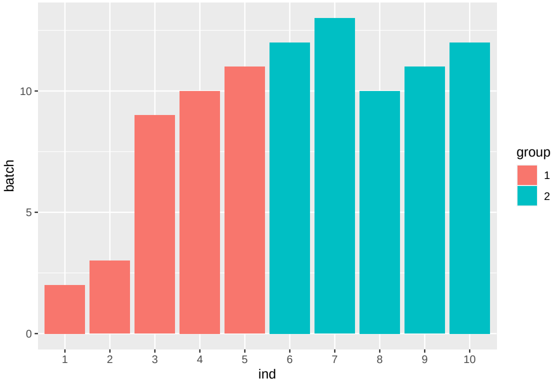
- Monotone would increase/decrease with the injection order or batchs.



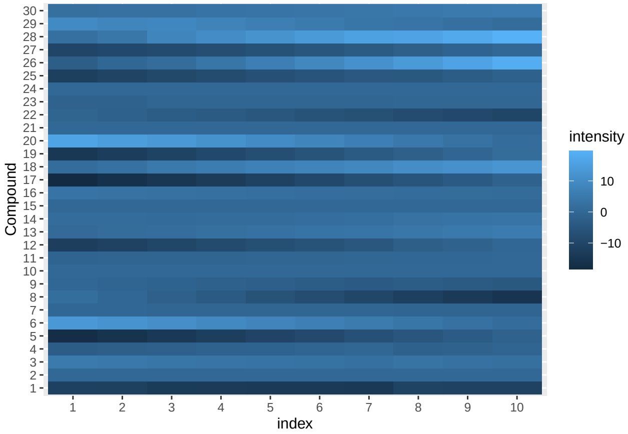
- Block would be system shift among different batchs.



- Mixed would be the combination of monotone and block batch effects.



Meanwhile, different compounds would suffer different type of batch effects. In this case, the normalization or batch correction should be done peak by peak.



6.2 Batch effects visulization

Any correction might introduce bias. We need to make sure there are patterns which different from our experimental design. Pooled QC samples should be clustered on PCA score plot.

6.3 Source of batch effects

- Different Operators & Dates & Sequences
- Different Instrumental Condition such as different instrumental parameters, poor quality control, sample contamination during the analysis, Column (Pooled QC) and sample matrix effects (ions suppression or/and enhancement)
- Unknown Unknowns

6.4 Avoid batch effects by DoE

You could avoid batch effects from experimental design. Cap the sequence with Pooled QC and Randomized samples sequence. Some internal standards/Instrumental QC might Help to find the source of batch effects while it's not practical for every compounds in non-targeted analysis.

Batch effects might not change the conclusion when the effect size is relatively small. Here is a simulation:

```r
set.seed(30)
# real peaks
group <- factor(c(rep(1,5),rep(2,5)))
con <- c(rnorm(5,5),rnorm(5,8))
re <- t.test(con~group)
# real peaks
group <- factor(c(rep(1,5),rep(2,5)))
con <- c(rnorm(5,5),rnorm(5,8))
batch <- seq(0,5,length.out = 10)
ins <- batch+con
re <- t.test(ins~group)

index <- sample(10)
ins <- batch+con[index]
re <- t.test(ins~group[index])
```

Randomization could not guarantee the results. Here is a simulation.

```
# real peaks
group <- factor(c(rep(1,5),rep(2,5)))
con <- c(rnorm(5,5),rnorm(5,8))
batch <- seq(5,0,length.out = 10)
ins <- batch+con
re <- t.test(ins~group)
```

6.5 post hoc data normalization

To make the samples comparable, normailization across the samples are always needed when the experiment part is done. Batch effect should have patterns, otherwise just noise. Correction is possible by data analysis/randomized experimental design. There are more than 20 methods to make normalization. We could devided those methods into two category: unsupervised and supervised.

Unsupervised methods only consider the normalization peaks intensity distribution across the samples. For example, quantile calibration try to make the intensity distribution among the samples similar. Such methods are preferred to explore the inner structures of the samples. Internal standards or pool QC samples also belong to this category. However, it's hard to take a few peaks standing for all peaks extracted.

Supervised methods will use the group information or batch information in experimental design to normalize the data. A linear model is always used to model the unwanted variances and remove them for further analysis.

Since the real batch effects are always unknown, it's hard to make validation for different normalization methods. Wu et.al preferred to make comparision between new methods and conventional methods(**?**). Li et.al developed NOREVA to make comparision among 25 correction method(**?**). Another idea is use spiked-in samples to validate the methods(**?**), which might be good for targeted analysis instead of non-targeted analysis.

Relative log abundance (RLA) plots(**?**) and heatmap often used to show the variances among the samples.

(**?**) some methods for batch correction in excel

6.5.1 Unsupervised methods

6.5.1.1 Distribution of intensity

Intensity collects from LC/GC-MS always showed a right-skewed distribution. Log transformation is often necessary for further statistical analysis.

6.5.1.2 Centering

For peak p of sample s in batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = I_{p,s,b} - mean(I_{p,b}) + median(I_{p,qc})$$

If no quality control samples used, the corrected abundance I would be:

$$\hat{I}_{p,s,b} = I_{p,s,b} - mean(I_{p,b})$$

6.5.1.3 Scaling

For peak p of sample s in certain batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{std_{p,b}} * std_{p,qc,b} + mean(I_{p,qc,b})$$

If no quality control samples used, the corrected abundance I would be:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{std_{p,b}}$$

6.5.1.4 Pareto Scaling

For peak p of sample s in certain batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{Sqrt(std_{p,b})} * Sqrt(std_{p,qc,b}) + mean(I_{p,qc,b})$$

If no quality control samples used, the corrected abundance I would be:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{Sqrt(std_{p,b})}$$

6.5.1.5 Range Scaling

For peak p of sample s in certain batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{max(I_{p,b}) - min(I_{p,b})} * (max(I_{p,qc,b}) - min(I_{p,qc,b})) + mean(I_{p,qc,b})$$

If no quality control samples used, the corrected abundance I would be:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{max(I_{p,b}) - min(I_{p,b})}$$

6.5.1.6 Level scaling

For peak p of sample s in certain batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{mean(I_{p,b})} * mean(I_{p,qc,b}) + mean(I_{p,qc,b})$$

If no quality control samples used, the corrected abundance I would be:

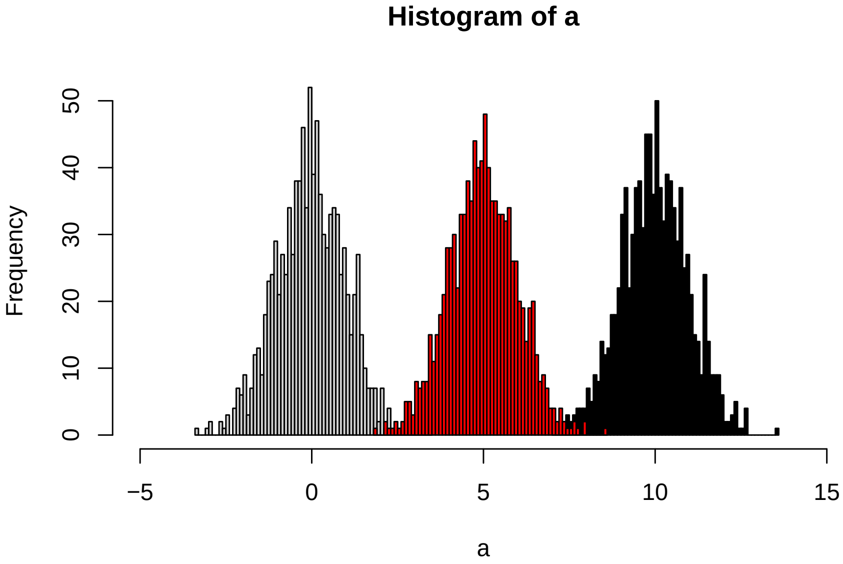
$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} - mean(I_{p,b})}{mean(I_{p,b})}$$

6.5.1.7 Quantile

The idea of quantile calibration is that alignment of the intensities in certain samples according to quantiles in each sample.

Here is the demo:

```r
set.seed(42)
a <- rnorm(1000)
# b sufferred batch effect with a bias of 10
b <- rnorm(1000,10)
hist(a,xlim=c(-5,15),breaks = 50)
hist(b,col = 'black', breaks = 50, add=T)
# quantile normalized
cor <- (a[order(a)]+b[order(b)])/2
# reorder
cor <- cor[order(order(a))]
hist(cor,col = 'red', breaks = 50, add=T)
```

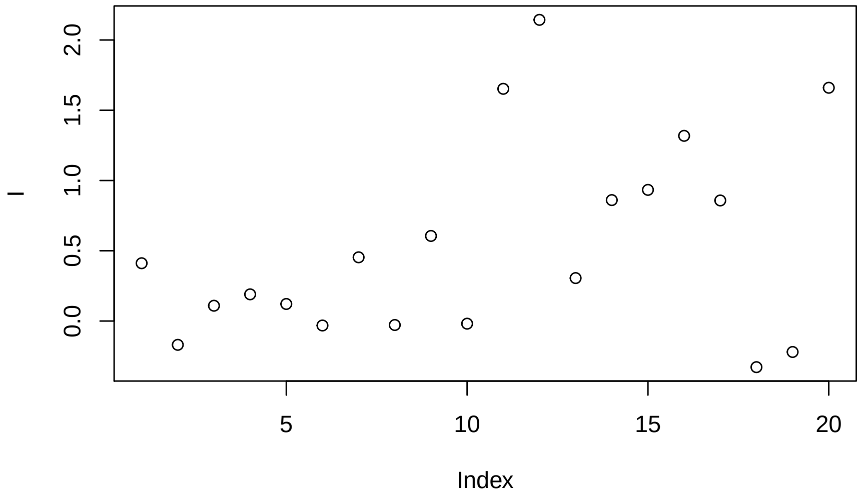


6.5.1.8 Ratio based calibraton

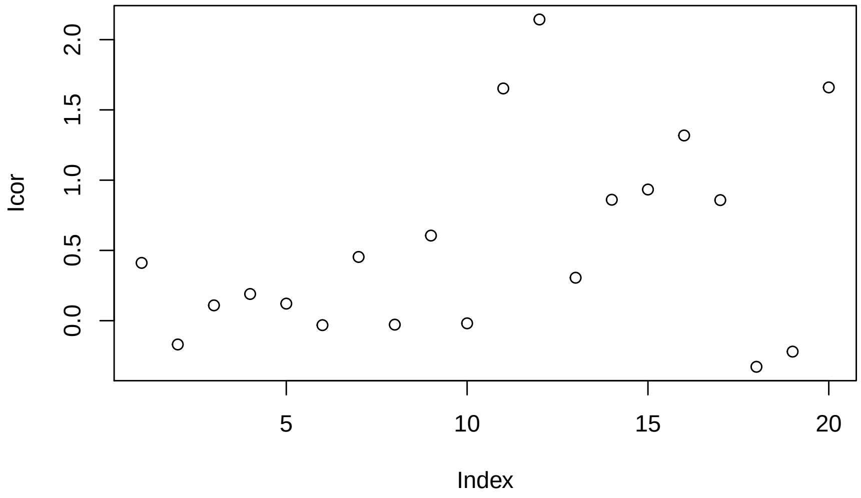
This method calibrates samples by the ratio between qc samples in all samples and in certain batch.For peak p of sample s in certain batch b, the corrected abundance I is:

$$\hat{I}_{p,s,b} = \frac{I_{p,s,b} * median(I_{p,qc})}{mean_{p,qc,b}}$$

```r
set.seed(42)
# raw data
I = c(rnorm(10,mean = 0, sd = 0.3),rnorm(10,mean = 1, sd = 0.5))
# batch
B = c(rep(0,10),rep(1,10))
# qc
Iqc = c(rnorm(1,mean = 0, sd = 0.3),rnorm(1,mean = 1, sd = 0.5))
# corrected data
Icor = I * median(c(rep(Iqc[1],10),rep(Iqc[2],10)))/mean(c(rep(Iqc[1],10),rep(Iqc[2],10)))
# plot the result
plot(I)
```



```
plot(Icor)
```



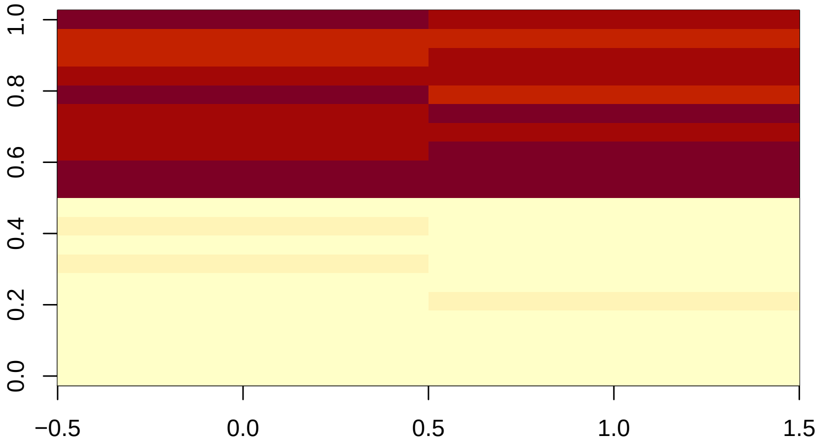
6.5.1.9 Linear Normalizer

This method initially scales each sample so that the sum of all peak abundances equals one. In this study, by multiplying the median sum of all peak abundances across all samples,we got the corrected data.

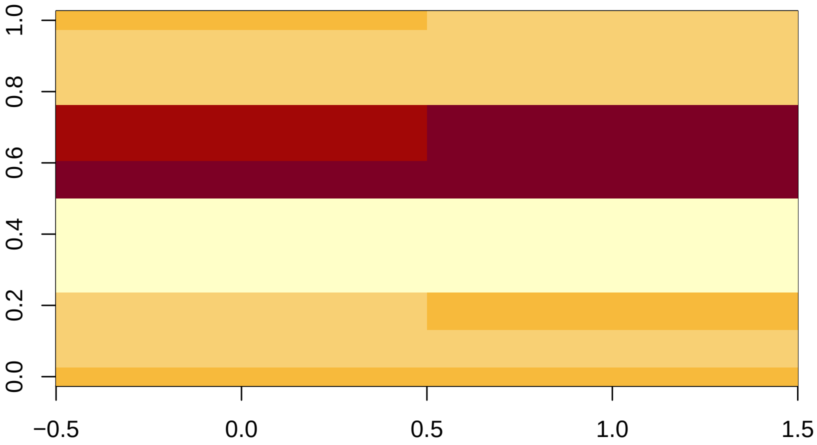
```
set.seed(42)
# raw data
peaksa <- c(rnorm(10,mean = 10, sd = 0.3),rnorm(10,mean = 20, sd = 0.5))
peaksb <- c(rnorm(10,mean = 10, sd = 0.3),rnorm(10,mean = 20, sd = 0.5))

df <- rbind(peaksa,peaksb)
dfcor <- df/apply(df,2,sum)* sum(apply(df,2,median))

image(df)
```



```
image(dfcor)
```



6.5.1.10 Internal standards

$$\hat{I}_{p,s} = \frac{I_{p,s} * median(I_{IS})}{I_{IS,s}}$$

Some methods also use pooled calibration samples and multiple internal standard strategy to correct the data(**?**). Also some methods only use QC samples to handle the data(**?**).

6.5.2 Supervised methods

6.5.2.1 Regression calibration

Considering the batch effect of injection order, regress the data by a linear model to get the calibration.

6.5.2.2 Batch Normalizer

Use the total abundance scale and then fit with the regression line(**?**).

6.5.2.3 Surrogate Variable Analysis(SVA)

We have a data matrix(M*N) with M stands for indentity peaks from one sample and N stand for individual samples. For one sample, $X = (x_{i1}, ..., x_{in})^T$ stands for the normalized intensities of peaks. We use $Y = (y_i, ..., y_m)^T$ stands for the group infomation of our data. Then we could build such modles:

$$x_{ij} = \mu_i + f_i(y_i) + e_{ij}$$

μ_i stands for the baseline of the peak intensities in a normal state. Then we have:

$$f_i(y_i) = E(x_{ij}|y_j) - \mu_i$$

stands for the biological variations caused by the our group, for example, whether treated by pollutions or not.

However, considering the batch effects, the real model could be:

$$x_{ij} = \mu_i + f_i(y_i) + \sum_{l=1}^{L} \gamma_{li} p_{lj} + e^*_{ij}$$

γ_{li} stands for the peak-specific coefficient for potentical factor l. p_{lj} stands for the potential factors across the samples. Actually, the error item e_{ij} in real sample could always be decomposed as $e_{ij} = \sum_{l=1}^{L} \gamma_{li} p_{lj} + e^*_{ij}$ with e^*_{ij} standing for the real random error in certain sample for certain peak.

We could not get the potential factors directly. Since we don't care the details of the unknown factors, we could estimate orthogonal vectors h_k standing for such potential factors. Thus we have:

$$x_{ij} = \mu_i + f_i(y_i) + \sum_{l=1}^{L} \gamma_{li} p_{lj} + e^*_{ij} = \mu_i + f_i(y_i) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + e_{ij}$$

Here is the details of the algorithm:

> The algorithm is decomposed into two parts: detection of unmodeled factors and construction of surrogate variables

6.5.2.3.1 Detection of unmodeled factors

- Estimate $\hat{\mu}_i$ and f_i by fitting the model $x_{ij} = \mu_i + f_i(y_i) + e_{ij}$ and get the residual $r_{ij} = x_{ij} - \hat{\mu}_i - \hat{f}_i(y_i)$. Then we have the residual matrix R.

- Perform the singular value decompositon(SVD) of the residual matrix $R = UDV^T$
- Let d_l be the lth eigenvalue of the diagonal matrix D for $l = 1, ..., n$. Set df as the freedom of the model $\hat{\mu}_i + \hat{f}_i(y_i)$. We could build a statistic T_k as:

$$T_k = \frac{d_k^2}{\sum_{l=1}^{n-df} d_l^2}$$

to show the variance explained by the kth eigenvalue.

- Permute each row of R to remove the structure in the matrix and get R^*.
- Fit the model $r_{ij}^* = \mu_i^* + f_i^*(y_i) + e_{ij}^*$ and get $r_{ij}^0 = r_{ij}^* - \hat{\mu}_i^* - \hat{f}_i^*(y_i)$ as a null matrix R_0
- Perform the singular value decompositon(SVD) of the residual matrix $R_0 = U_0 D_0 V_0^T$
- Compute the null statistic:

$$T_k^0 = \frac{d_{0k}^2}{\sum_{l=1}^{n-df} d_{0l}^2}$$

- Repeat permuting the row B times to get the null statistics T_k^{0b}
- Get the p-value for eigengene:

$$p_k = \frac{\#T_k^{0b} \geq T_k; b = 1, ..., B}{B}$$

- For a significance level α, treat k as a significant signature of residual R if $p_k \leq \alpha$

6.5.2.3.2 Construction of surrogate variables

- Estimate $\hat{\mu}_i$ and f_i by fitting the model $x_{ij} = \mu_i + f_i(y_i) + e_{ij}$ and get the residual $r_{ij} = x_{ij} - \hat{\mu}_i - \hat{f}_i(y_i)$. Then we have the residual matrix R.
- Perform the singular value decompositon(SVD) of the residual matrix $R = UDV^T$. Let $e_k = (e_{k1}, ..., e_{kn})^T$ be the kth column of V

- Set $\hat{K}$ as the significant eigenvalues found by the first step.
- Regress each e_k on x_i, get the p-value for the association.
- Set π_0 as the proportion of the peak intensitity x_i not associate with e_k and find the numbers $\hat{m} = [1 - \hat{\pi}_0 \times m]$ and the indices of the peaks associated with the eigenvalues
- Form the matrix $\hat{m}_1 \times N$, this matrixX_r stand for the potiential variables. As was done for R, get the eigengents of X_r and denote these by e^r_j
- Let $j^* = argmax_{1 \leq j \leq n} cor(e_k, e^r_j)$ and set $\hat{h}_k = e^r_j$. Set the estimate of the surrogate variable to be the eigenvalue of the reduced matrix most correlated with the corresponding residual eigenvalue. Since the reduced matrix is enriched for peaks associated with this residual eigenvalue, this is a principled choice for the estimated surrogate variable that allows for correlation with the primary variable.
- Employ the $\mu_i + f_i(y_i) + \sum_{k=1}^{K} \gamma_{ki} \hat{h}_{kj} + e_{ij}$ as te estimate of the ideal model $\mu_i + f_i(y_i) + \sum_{k=1}^{K} \gamma_{ki} h_{kj} + e_{ij}$

This method could found the potentical unwanted variables for the data. SVA were introduced by Jeff Leek(**???**) and EigenMS package implement SVA with modifications including analysis of data with missing values that are typical in LC-MS experiments(**?**).

6.5.2.4 RUV (Remove Unwanted Variation)

This method's performance is similar to SVA. Instead find surrogate variable from the whole dataset. RUA use control or pool QC to find the unwanted variances and remove them to find the peaks related to experimental design. However, we could also empirically estimate the control peaks by linear mixed model. RUA-random(**?**) furthor use linear mixed model to estimate the variances of random error. This method could be used with suitable control, which is commen in metabolomics DoE.

6.5.2.5 RRmix

RRmix also use a latent factor models correct the data(**?**). This method could be treated as linear mixed model version SVA. No control samples are required and the unwanted variances could be removed by factor analysis. This method might be the best choise to remove the unwanted variables with commen experiment design.

6.6 Method to validate the normalization

6.7 Software

- BatchCorrMetabolomics is for improved batch correction in untargeted MS-based metabolomics
- MetNorm show Statistical Methods for Normalizing Metabolomics Data.
- BatchQC could be used to make batch effect simulation.
- Noreva could make online batch correction.

Chapter 7

Annotation

When you get the peaks table or features table, annotation of the peaks would help you. Check this review(**?**) for a detailed notes on annotation. They proposed five levels regarding currently computational annotation strategies.

- Level 1: Peak Grouping: MS Psedospectra extraction based on peak shape similarity and peak abundance correlation
- Level 2: Peak Annotation: Adducts, Neutral losses, isotopes, and other mass relationships based on mass distances
- Level 3: Biochemical knowledge based on putative identification, potential biochemical reaction and related statistical analysis
- Level 4: Use and intergration of tandem MS data based on data dependant/independent acquistion mode or **in silico** predction
- Level 5: Retention time prediction based on library-available retention index or quantitative structure-retnetion relationships (QSRR) models.

Most of the softwares are at level 1 or 2. If we only have compounds structure, we could guess ions under different ionization method. If we have mass spectrum, we could match the mass spectral by a similarity analysis to the database. In metabolomics, we only have mass spectrum or mass-to-charge ratios. Single mass-to-charge ratio is not enough for identification. That's the one bottleneck for annotation. So prediction is always performed on MS/MS data.

7.1 Issues in annotation

The major issue in annotation is the redundancy peaks from same metabolite. Unlike genomcis, peaks or featuers from peak selection are not independant with each other. Adducts, in-source fragments and isotopes would lead to missannotation. A commen solution is that use known adducts, neutral losses, molecular multimers or multipley charged ions to compare mass distances.

Another issue is about the MS/MS database. Only 10% of known metabolites in databases have experimental spectral data. Thus **in silico** prediction are required. Some works try to fill the gap between experimental data, theoretical values(from chemical database like chemspider) and prediction together. Here is a nice review about MS/MS prediction(**?**).

7.2 Peak misidentification

- Isomer

Use seperation methods such as chromatography, ion mobility MS, MS/MS. Reversed-phase ion-pairing chromatography and HILIC is useful.Chemical derivatization is another options.

- Interfering compounds

20ppm is the least resolution and accuracy for HRMS.

- In-source degradation products

7.3 Annotation v.s. identification

According to the defination from the Chemical Analysis Working Group of the Metabolomics Standards Intitvative(**??**). Four levels of confidence could be assigned to identification:

- Level 1 ‘identified metabolites’
- Level 2 ‘Putatively annotated compounds’
- Level 3 ‘Putatively characterised compound classes’
- Level 4 ‘Unknown’

In practice, data analysis based annotation could reach level 2. For level 1, we need at extra methods such as MS/MS, retention time, accurate mass, 2D NMR spectra, and so on to confirm the compounds. However, standards are always required for solid proof.

7.4 Cheminformatics

- RDKit Open-Source Cheminformatics Software
- cdk The Chemistry Development Kit (CDK) is a scientific, LGPL-ed library for bio- and cheminformatics and computational chemistry written in Java.
- Open Babel Open Babel is a chemical toolbox designed to speak the many languages of chemical data.

7.5 MS Database for annotation

7.5.1 MS/MS

- MoNA Platform to collect all other open source database
- MassBank
- GNPS use inner correlationship in the data and make network analysis at peaks' level instand of annotated compounds to annotate the data(**?**).
- ReSpect: phytochemicals
- Metlin is another useful online application for annotation(**?**).
- LipidBlast: in silico prediction
- MZcloud
- NIST: Not free

7.5.2 MS

- Fiehn Lab
- NIST: No free
- Spectral Database for Organic Compounds, SDBS
- MINE is an open access database of computationally predicted enzyme promiscuity products for untargeted metabolomics
- Database of free solution mobilities for 276 metabolites for capillary zone electrophoresis / mass spectrometry. (**?**)

7.6 Compounds Database

- PubChem is an open chemistry database at the National Institutes of Health (NIH).
- Chemspider is a free chemical structure database providing fast text and structure search access to over 67 million structures from hundreds of data sources.
- ChEBI is a freely available dictionary of molecular entities focused on ‘small’ chemical compounds.
- RefMet A Reference list of Metabolite names.
- CAS Largest substance database

7.7 Software

7.7.1 Adducts list

You could find adducts list here from commonMZ project.

7.7.2 Molgen

molgen generating all structures (connectivity isomers, constitutions) that correspond to a given molecular formula, with optional further restrictions, e.g. presence or absence of particular substructures.

7.7.3 Isotope

Isotope pattern prediction

7.7.4 mfFinder

mfFinder predict formula based on accurate mass

7.7.5 CAMERA

Common annotation for xcms workflow(**?**).

7.7.6 RAMClustR

The software could be found here(**?**). The package included a vignette as usages. Use the following code to read:

```
vignette('RAMClustR',package = 'RAMClustR')
```

7.7.7 pmd

Paired Mass Distance(PMD) analysis for GC/LC-MS based nontarget analysis to find independant peaks(**?**)

7.7.8 nontarget

nontarget Isotope & adduct peak grouping, homologue series detection

7.7.9 xMSannotator

The software could be found here(**?**).

7.7.10 CFM-ID

CFM-ID use Metlin's data to make prediction(**?**).

7.7.11 MINE

MINE is an open access database of computationally predicted enzyme promiscuity products for untargeted metabolomics. The annotation would be accurate for general compounds database.

7.7.12 InterpretMSSpectrum

This package is for annotate and interpret deconvoluted mass spectra (mass*intensity pairs) from high resolution mass spectrometry devices. You could use this package to find molecular ions for GC-MS.

7.7.13 For Ident

For-ident could give a score for identification with the help of logD(relative retention time) and/or MS/MS.

7.7.14 Retip

Retip Retention Time Prediction for Compound Annotation in Untargeted Metabolomics (?)

7.7.15 Binner

Binner Deep annotation of untargeted LC-MS metabolomics data (?)

7.7.16 mzmatch

Use the following code to install this package:

```
source("http://bioconductor.org/biocLite.R")
biocLite(c("xcms", "multtest", "mzR"))
install.packages(c("rJava", "XML", "snow", "caTools",
   "bitops", "ptw", "gplots", "tcltk2"))
source ("http://puma.ibls.gla.ac.uk/mzmatch.R/install_mzmatch.R")
```

7.7.17 mz.unity

You could find source code here(?) and it's for detecting and exploring complex relationships in accurate-mass mass spectrometry data.

7.7.18 MAIT

You could find source code here(?).

7.7.19 ProbMetab

Provides probability ranking to candidate compounds assigned to masses, with the prior assumption of connected sample and additional previous and spectral information modeled by the user. You could find source code here(?).

7.7.20 RAMSI

You could find paper here(?).

7.7.21 Sirius

Sirius is a new java-based software framework(?) for discovering a landscape of de-novo identification of metabolites using single and tandem mass spectrometry. It could be used with CSI:FingerID.

7.7.22 MI-Pack

You could find python software here(?)

7.7.23 Plantmat

excel library based pridiction for plant metabolites(?).

7.7.24 MetFamily

Shiny app for MS and MS/MS data annotation(?).

7.7.25 Lipidmatch

in silico: in silico lipid mass spectrum search(?).

7.7.26 MolFind

JAVA based MolFind could make annotation for unknown chemical structure by prediction based on RI, ECOM50, drift time and CID spectra(?).

7.7.27 MetFusion

Java based integration of compound identification strategies. You could access the application here(?).

7.7.28 iMet

This online application is a network-based computation method for annotation(?).

7.7.29 Metscape

Metscape based on Debiased Sparse Partial Correlation (DSPC) algorithm(**?**) to make annotation.

7.7.30 MetFrag

MetFrag could be used to make **in silico** prediction/match of MS/MS data(**?**).

7.7.31 LipidFrag

LipidFrag could be used to make **in silico** prediction/match of lipid related MS/MS data(**?**).

7.7.32 MycompoundID

MycompoundID could be used to search known and unknown metabolites(**?**) online.

7.7.33 magma

magma could predict and match MS/MS files.

7.7.34 MetExpert

MetExpert is an expert system to assist users with limited expertise in informatics to interpret GCMS data for metabolite identification without querying spectral databases(**?**)

7.7.35 MS-DIA

- decoMS2 requires two different collision energies, low (usually 0V) and high, in each precursor range to solve the mathematical equations.(**?**)
- MS-DIAL data independent MS/MS deconvolution for comprehensive metabolome analysis.(**?**)
- MetDIA Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition(**?**)
- DIA-Umpire comprehensive computational framework for data-independent acquisition proteomics(**?**)

- MetaboDIA quantitative metabolomics analysis using DIA-MS(**?**)
- SWATHtoMRM Development of High-Coverage Targeted Metabolomics Method Using SWATH Technology for Biomarker Discovery(**?**)

Chapter 8

Omics analysis

When you get the filtered ions, the next step is making annotations for them. Such annotations would be helpful for omics studies. Omics analysis try to combine the information from other 'omics' to answer one sepcific question. Since we have got the annotations, Omics analysis could be performed.Upload the data obtained from the **xcms** to other tools or databases.

You will get an updated database list here

Right now, it is hard to connect different omics databases such as gene, protein and metabolites together for a whole scope of certain biological process. However, you might select few metabolites across those databases and find something interesting.

8.1 From Bottom-up to Top-down

Bottom-up analysis mean the model for each metabolite. In this case, we could find out which metabolite will be affected by our experiment design. However, take care of multiple comparision issue.

$$metabolite = f(control/treatment, co - variables)$$

Top-down analysis mean the model for output. In this case, we could evaluate the contribution of each metabolites. You need variable selection to make a better model.

$$control/treatment = f(metabolite1, metabolite2, ..., metaboliteN, co{-}varuables)$$

For omics study, you might need to intergrate dataset from different sources.

$$control/treatment = f(metabolites, proteins, genes, miRNA, co - varuables)$$

8.2 Pathway analysis

Pathwat analysis maps annotated data into known pathway and make statistical analysis to find the influenced pathway or the compounds with high inflences on certain pathway.

8.2.1 Pathway Database

- SMPDB (The Small Molecule Pathway Database) is an interactive, visual database containing more than 618 small molecule pathways found in humans. More than 70% of these pathways (>433) are not found in any other pathway database. The pathways include metabolic, drug, and disease pathways.
- KEGG (Kyoto Encyclopedia of Genes and Genomes) is one of the most complete and widely used databases containing metabolic pathways (495 reference pathways) from a wide variety of organisms (>4,700). These pathways are hyperlinked to metabolite and protein/enzyme information. Currently KEGG has >17,000 compounds (from animals, plants and bacteria), 10,000 drugs (including different salt forms and drug carriers) and nearly 11,000 glycan structures.
- BioCyc is a collection of 14558 Pathway/Genome Databases (PGDBs), plus software tools for exploring them.
- Reactome is an open-source, open access, manually curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic and clinical research, genome analysis, modeling, systems biology and education.
- WikiPathway is a database of biological pathways maintained by and for the scientific community.

8.2.2 Pathway software

- Pathway Commons online tools for pathway analysis
- RaMP could make pathway analysis for batch search

- metabox could make pathway analysis
- impala is used for pathway enrichment analysis

8.3 Network analysis

Mummichog could make pathway and network analysis without annotation.

MSS: sequential feature screening procedure to select important sub-network and identify the optimal matching for metabolimics data (**?**)

8.4 Omics integration

- Blast finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
- The Omics Discovery Index (OmicsDI) provides a knowledge discovery framework across heterogeneous omics data (genomics, proteomics, transcriptomics and metabolomics).
- Omics Data Integration Project

Chapter 9

Common analysis methods for metabolomics

The general purposes for metabolomics study are strongly associated with research goal. However, since metabolomics are usually performed in a non-targeted mode, statistical analysis methods are always started with the exploratory analysis. The basic target for an exploratory analysis is:

- Find the relationship among variables
- Find the relationship among samples/group of samples.

This is basically unsurpvised analysis.

However, sometimes we have group information which could be used to find biomarkers or correlationship between variables and groups or continous variables. This type of data need supervised methods to process. Before we talk the details of algorithms, let's cover some basic statistical concepts.

9.1 Basic Statistical Analysis

Statistic is used to describe certain property or variables among the samples. It could be designed for certain purpose to extract signal and remove noise. Statistical models and inference are both based on statistic instead of the data.

$$Statistic = f(sample_1, sample_2, ..., sample_n)$$

Null Hypothesis Significance Testing (NHST) is often used to make statistical inference. P value is the probability of certain statistics happens under H0 (pre-defined distribution).

For omics studies, you should realise **Multiple Comparision** issue when you perform a lot of(more than 20) comparisions or tests at the same time. **False Discovery Rate(FDR) control** is required for multiple tests to make sure the results are not false positive. You could use Benjamini-Hochberg method to adjust raw p values or directly use Storey Q value to make FDR control.

NHST is famous for the failure of p-value interpretation as well as multiple comparision issues. **Bayesian Hypothesis Testing** could be an options to cover some drawbacks of NHST. Bayesian Hypothesis Testing use Bayes factor to show the differences between null hypothesis and any other hypothesis.

$$Bayes\ factor = \frac{p(D|Ha)}{p(D|H0)} = \frac{posterior\ odds}{prior\ odds}$$

Statistical model use statistics to make prediction/explanation. Most of the statistical model need to be tuned for parpameters to show a better performance. Statistical model is build on real data and could be diagnosed by other general statistics such as R^2, $ROCcurve$. When the models are built or compared, model selection could be preformed.

$$Target = g(Statistic) = g(f(sample_1, sample_2, ..., sample_n))$$

Bias-Variance Tradeoff is an important concept regarding statistical models. Certain models could be overfitted(small Bias, large variance) or underfitted(large Bias, small variance) when the parameters of models are not well selected.

$$E[(y - \hat{f})^2] = \sigma^2 + Var[\hat{f}] + Bias[\hat{f}]$$

Cross validation could be used to find the best model based on training-testing strategy such as Jacknife, bootstraping resampling and n-fold cross validation.

Regularization for models could also be used to find the model with best prediction performance. Rigid regression, LASSO or other general regularization could be employed to build a robust models.

For supervised models, linear model and tree based model are two basic categories. **Linear model** could be useful to tell the independant or correlated relationship of variables and the influnces on the predicted variables. **Tree based model**, on the other hand, try to build a hierarchical structure for the variables such as bagging, random forest or boosting. Linear model could be treated as special case of tree based model with single layer. Other models like Support Vector Machine (SVM), Artificial Neural Network (ANN) or Deep Learning are also make various assumptions on the data. However, if you final target is prediction, you could try any of those models or even weighted combine their prediciton to make meta-prediction.

9.2 Differences analysis

After we get corrected peaks across samples, the next step is to find the differences between two groups. Actually, you could perform ANOVA or Kruskal-Wallis Test for comparison among more than two groups. The basic idea behind statistic analysis is to find the meaningful differences between groups and extract such ions or peak groups.

So how to find the differences? In most metabolomics software, such task is completed by a t-test and report p-value and fold changes. If you only compare two groups on one peaks, that's OK. However, if you compare two groups on thousands of peaks, statistic textbook would tell you to notice the false positive. For one comparasion, the confidence level is 0.05, which means 5% chances to get false positive result. For two comparasions, such chances would be $1-0.95^2$. For 10 comparasions, such chances would be $1-0.95^{10} = 0.4012631$. For 100 comparasions, such chances would be $1-0.95^{100} = 0.9940795$. You would almost certainly to make mistakes for your results.

In statistics, the false discovery rate(FDR) control is always mentioned in omics studies for mutiple tests. I suggested using q-values to control FDR. If q-value is less than 0.05, we should expect a lower than 5% chances we make the wrong selections for all of the comparisions showed lower q-values in the whole dataset. Also we could use local false discovery rate, which showed the FDR for certain peaks. However, such values are hard to be estimated accurately.

Karin Ortmayr thought fold change might be better than p-values to find the differences(**?**).

9.2.1 T-test or ANOVA

If one peak show significant differences among two groups or multiple groups, T-test or ANOVA could be used to find such peaks. However, when multiple hypothesis testings are performed, the probability of false positive would increase. In this case, false discovery rate(FDR) control is required. Q value or adjusted p value could be used in this situation. At certain confidence interval, we could find peaks with significant differences after FDR control.

9.2.2 LIMMA

Linear Models for MicroArray Data(LIMMA) model could also be used for high-dimensional data like metabolomics. They use a moderated t-statistic to make estimation of the effects called Empirical Bayes Statistics for Differential Expression. It is a hierarchical model to shrink the t-statistic for each peak to all the peaks. Such estimation is more robust. In LIMMA, we could add the known batch effect variable as a covariance in the model. LIMMA is different from t-test or ANOVA while we could still use p value and FDR control on LIMMA results.

9.2.3 Bayesian mixture model

Another way to make difference analysis is based on Bayesian mixture model without p value. Such model would not use hypothesis testing and directly generate the posterior estimation of parameters. A posterior probability could be used to check whether certain peaks could be related to different condition. If we want to make comparison between classical model like LIMMA and Bayesian mixture model. We need to use simulation to find the cutoff.

9.3 PCA

In most cases, PCA is used as an exploratory data analysis(EDA) method. In most of those most cases, PCA is just served as visualization method. I mean, when I need to visualize some high-dimension data, I would use PCA.

So, the basic idea behind PCA is compression. When you have 100 samples with concentrations of certain compound, you could plot the concentrations with samples' ID. However, if you have 100 compounds to be analyzed, it would by hard to show the relationship between the samples. Actually, you need to show a matrix with sample and compounds (100 * 100 with the concentrations filled into the matrix) in an informal way.

The PCA would say: OK, guys, I could convert your data into only 100 * 2 matrix with the loss of information minimized. Yeah, that is what the mathematical guys or computer programmer do. You just run the command of PCA. The new two "compounds" might have the cor-relationship between the original 100 compounds and retain the variances between them. After such projection, you would see the compressed relationship between the 100 samples. If some samples' data are similar, they would be projected together in new two "compounds" plot. That is why PCA could be used for cluster and the new "compounds" could be referred as principal components(PCs).

However, you might ask why only two new compounds could finished such task. I have to say, two PCs are just good for visualization. In most cases, we need to collect PCs standing for more than 80% variances in our data if you want to recovery the data with PCs. If each compound have no relationship between each other, the PCs are still those 100 compounds. So you have found a property of the PCs: PCs are orthogonal between each other.

Another issue is how to find the relationship between the compounds. We could use PCA to find the relationship between samples. However, we could also extract the influences of the compounds on certain PCs. You might find many compounds showed the same loading on the first PC. That means the concentrations pattern between the compounds are looked similar. So PCA could also be used to explore the relationship between the compounds.

OK, next time you might recall PCA when you need it instead of other paper showed them.

Besides, there are some other usage of PCA. Loadings are actually correlation coefficients between peaks and their PC scores. Yamamoto et.al.(**?**) used t-test on this correlation coefficient and thought the peaks with statistically significant correlation to the PC score have biological meanings for further study such as annotation. However, such analysis works better when few PCs could explain most of the variances in the datasets.

9.4 Cluster Analysis

After we got a lot of samples and analyzed the concentrations of many compounds in them, we may ask about the relationship between the samples. You might have the sampling information such as the date and the position and you could use boxplot or violin plot to explore the relationships among those categorical variables. However, you could also use the data to find some potential relationship.

But how? if two samples' data were almost the same, we might think those samples were from the same potential group. On the other hand, how do we define the "same" in the data?

Cluster analysis told us that just define a "distances" to measure the similarity between samples. Mathematically, such distances would be shown in many different manners such as the sum of the absolute values of the differences between samples.

For example, we analyzed the amounts of compound A, B and C in two samples and get the results:

Compounds(ng)	A	B	C
Sample 1	10	13	21
Sample 2	54	23	16

The distance could be:

$$distance = |10 - 54| + |13 - 23| + |21 - 16| = 59$$

Also you could use the sum of squares or other way to stand for the similarity. After you defined a "distance", you could get the distances between all of pairs for your samples. If two samples' distance was the smallest, put them together as one group. Then calculate the distances again to combine the small group into big group until all of the samples were include in one group. Then draw a dendrogram for those process.

The following issue is that how to cluster samples? You might set a cut-off and directly get the group from the dendrogram. However, sometimes you were ordered to cluster the samples into certain numbers of groups such as three. In such situation, you need K means cluster analysis.

The basic idea behind the K means is that generate three virtual samples and calculate the distances between those three virtual samples and all of the other samples. There would be three values for each samples. Choose the smallest values and class that sample into this group. Then your samples were classified into three groups. You need to calculate the center of those three groups and get three new virtual samples. Repeat such process until the group members unchanged and you get your samples classified.

OK, the basic idea behind the cluster analysis could be summarized as define the distances, set your cut-off and find the group. By this way, you might show potential relationships among samples.

9.5 PLSDA

PLS-DA, OPLS-DA and HPSO-OPLS-DA(**?**) could be used.

Partial least squares discriminant analysis(PLSDA) was first used in the 1990s. However, Partial least squares(PLS) was proposed in the 1960s by Hermann Wold. Principal components analysis produces the weight matrix reflecting the covariance structure between the variables, while partial least squares produces the weight matrix reflecting the covariance structure between the variables and classes. After rotation by weight matrix, the new variables would contain relationship with classes.

The classification performance of PLSDA is identical to linear discriminant analysis(LDA) if class sizes are balanced, or the columns are adjusted according to the mean of the class mean. If the number of variables exceeds the number of samples, LDA can be performed on the principal components. Quadratic discriminant analysis(QDA) could model nonlinearity relationship between variables while PLSDA is better for collinear variables. However, as a classifier, there is little advantage for PLSDA. The advantages of PLSDA is that this modle could show relationship between variables, which is not the goal of regular classifier.

Different algorithms(**?**) for PLSDA would show different score, while PCA always show the same score with fixed algorithm. For PCA, both new variables and classes are orthognal. However, for PLS(Wold), only new classes are orthognal. For PLS(Martens), only new variables are orthognal. This paper show the details of using such methods(**?**).

Sparse PLS discriminant analysis(sPLS-DA) make a L1 penal on the variable selection to remove the influnces from unrelated variables, which make sense for high-throughput omics data(**?**).

For o-PLS-DA, s-plot could be used to find features.(**?**)

9.6 Software

- MetaboAnalystR (**?**)
- caret could employ more than 200 statistical models in a general framework to build/select models. You could also show the variable importance for some of the models.
- caretEnsemble Functions for creating ensembles of caret models
- pROC Tools for visualizing, smoothing and comparing receiver operating characteristic (ROC curves). (Partial) area under the curve (AUC) can be compared with statistical tests based on U-statistics or bootstrap. Confidence intervals can be computed for (p)AUC or ROC curves.
- gWQS Fits Weighted Quantile Sum (WQS) regressions for continuous, binomial, multinomial and count outcomes.

Chapter 10

Workflow

10.1 Platform for metabolomics

Here is a list for related open source projects

10.1.1 XCMS online

XCMS online is hosted by Scripps Institute. If your datasets are not large, XCMS online would be the best option for you. Recently they updated the online version to support more functions for systems biology. They use metlin and iso metlin to annotate the MS/MS data. Pathway analysis is also supported. Besides, to accelerate the process, xcms online employed stream (windows only). You could use stream to connect your instrument workstation to their server and process the data along with the data acquisition automate. They also developed apps for xcms online, but I think apps for slack would be even cooler to control the data processing.

10.1.2 PRIMe

PRIMe is from RIKEN and UC Davis. They update their database frequently(**?**). It supports mzML and major MS vendor formats. They defined own file format ABF and eco-system for omics studies. The software are updated almost everyday. You could use MS-DIAL for untargeted analysis and MRMOROBS for targeted analysis. For annotation, they developed MS-FINDER and statistic tools with excel. This platform could replaced the dear software from company and well prepared for MS/MS data analysis and lipidomics. They are open source, work on Windows and also could run within mathmamtics. However, they don't cover pathway analysis. Another feature is they always show the most recently spectral records from public repositories. You could always get the updated MSP spectra files for your own data analysis.

If you make GC-MS based metabolomics, this paper(**?**) could be nice start.

10.1.3 OpenMS

OpenMS is another good platform for mass spectrum data analysis developed with C++. You could use them as plugin of KNIME. I suggest anyone who want to be a data scientist to get familiar with platform like KNIME because they supplied various API for different programme language, which is easy to use and show every steps for others. Also TOPPView in OpenMS could be the best software to visualize the MS data. You could always use the metabolomics workflow to train starter about details in data processing. pyOpenMS and OpenSWATH are also used in this platform. If you want to turn into industry, this platform fit you best because you might get a clear idea about solution and workflow.

10.1.4 MZmine 2

MZmine 2 has three version developed on Java platform and the lastest version is included into MSDK. Similar function could be found from MZmine 2 as shown in XCMS online. However, MZmine 2 do not have pathway analysis. You could use metaboanalyst for that purpose. Actually, you could go into MSDK to find similar function supplied by ProteoSuite and Openchrom. If you are a experienced coder for Java, you should start here.

10.1.5 XCMS

xcms is different from xcms online while they might share the same code. I used it almost every data to run local metabolomics data analysis. Recently, they will change their version to xcms 3 with major update for object class. Their data format would integrate into the MSnbase package and the parameters would be easy to set up for each step. Normally, I will use msconvert-IPO-xcms-xMSannotator-metaboanalyst as workflow to process the offline data. It could accelerate the process by parallel processing. However, if you are not familiar with R, you would better to choose some software above.

10.1.6 Emory MaHPIC

This platform is composed by several R packages from Emory University including apLCMS to collect the data, xMSanalyzer to handle automated pipeline for large-scale, non-targeted metabolomics data, xMSannotator for annotation of LC-MS data and Mummichog for pathway and network analysis for high-throughput metabolomics. This platform would be preferred by someone from environmental science to study exposome. I always use xMSannotator to annotate the LC-MS data.

10.1.7 DIA data analysis

Skyline is a freely-available and open source Windows client application for building Selected Reaction Monitoring (SRM) / Multiple Reaction Monitoring (MRM), Parallel Reaction Monitoring (PRM - Targeted MS/MS), Data Independent Acquisition (DIA/SWATH) and targeted DDA with MS1 quantitative methods and analyzing the resulting mass spectrometer data.

MSstats is an R-based/Bioconductor package for statistical relative quantification of peptides and proteins in mass spectrometry-based proteomic experiments. It is applicable to multiple types of sample preparation, including label-free workflows, workflows that use stable isotope labeled reference proteins and peptides, and work-flows that use fractionation. It is applicable to targeted Selected Reactin Monitoring(SRM), Data-Dependent Acquisiton(DDA or shotgun), and Data-Independent Acquisition(DIA or SWATH-MS). This github page is for sharing source and testing.

MS-DAIL is also an option for DIA.

10.1.8 Others

- MAVEN from Princeton University
- MAIT based on xcms and you could find source code here(**?**).
- metabolomics is a CRAN package for analysis of metabolomics data.
- LipidFinder A computational workflow for discovery of new lipid molecular species
- enviGCMS from environmental non-targeted analysis.
- pySM provides a reference implementation of our pipeline for False Discovery Rate-controlled metabolite annotation of high-resolution imaging mass spectrometry data.
- MetabolomeExpress a public place to process, interpret and share GC/MS metabolomics datasets.
- PhenoMeNal is an easy-to-use, cloud-based metabolomic research environment.
- MetAlign&MSClust

- MetaboliteDetector is a QT4 based software package for the analysis of GC/MS based metabolomics data.
- autoGCMSDataAnal is a Matlab based comprehensive data analysis strategy for GC-MS-based untargeted metabolomics(**?**).

10.2 Data sharing

See this paper(**?**):

- MetaboLights EU based
- The Metabolomics Workbench US based
- MetabolomeXchange search engine
- W4M(**?**)

10.3 Contest

- CASMI predict smail molecular contest

10.4 Demo

Chapter 11

Demo

11.1 Project Setup

I suggest building your data analysis projects in RStudio(Click File - New project - New dictionary - Empty project). Then assign a name for your project. I also recommend the following tips if you are familiar with it.

- Use git/github to make version control of your code and sync your project online.
- Don't use your name for your project because other peoples might cooperate with you and someone might check your data when you publish your papers. Each project should be a work for one paper or one chapter in your thesis.
- Use **workflow** document(txt or doc) in your project to record all of the steps and code you performed for this project. Treat this document as digital version of your experiment notebook
- Use **data** folder in your project folder for the raw data and the results you get in data analysis
- Use **figure** folder in your project folder for the figure
- Use **munuscript** folder in your project folder for the manuscript (you could write paper in rstudio with the help of template in Rmarkdown)
- Just double click [**yourprojectname**]**.Rproj** to start your project

11.2 Data input

xcms does not support all of the raw files formate from every mass spectrometry. You need to convert your raw data into some open-source data format such as mzData, mzXML or CDF files. The tool is **MSconvert** from **ProteoWizard**.

11.3 Optimization

IPO package could be used to optimaze the parameters for XCMS. Try the following code and here we employ 5 NIST 1950 samples and 6 matrix blank samples as demodata from `rmwf` package.

```r
library(IPO)
library(xcms)
library(rmwf)
path <- system.file("extdata/data", package = "rmwf")
files <- list.files(path,recursive = T,full.names = T)

# BiocManager::install("IPO")
library(IPO)
library(xcms)
peakpickingParameters <- getDefaultXcmsSetStartingParams('centWave')
# Demo data
path <- system.file("extdata/data", package = "rmwf")
files <- list.files(path,recursive = T,full.names = T)
# Uncomment this line to use your own data(suggested 3-5 pooled QC samples)
# path <- 'path/to/your/files'
# change to 5 for obitrap
peakpickingParameters$ppm <- 10
resultPeakpicking <-
  optimizeXcmsSet(files = files[c(7,9,11)],
                  params = peakpickingParameters,
                  plot = F,
                  subdir = NULL)

optimizedXcmsSetObject <- resultPeakpicking$best_settings$xset
retcorGroupParameters <- getDefaultRetGroupStartingParams()
resultRetcorGroup <-
  optimizeRetGroup(xset = optimizedXcmsSetObject,
                   params = retcorGroupParameters,
                   plot = F,
                   subdir = NULL)
para <- c(resultPeakpicking$best_settings$parameters,
          resultRetcorGroup$best_settings)
```

```r
save(para,file = 'para.RData')
sessionInfo()
```

11.4 Wrap function

Here we would use the optimized parameters for peak picking, retention time correction and peaks filling.

```r
load('para.RData')
library(xcms)
library(stringr)
getrtmz <- function(path,index = NULL){
  peakwidth <- c(para$min_peakwidth,para$max_peakwidth)
  ppm <- para$ppm
  noise <- para$noise
  snthresh <- para$snthresh
  mzdiff <- para$mzdiff
  prefilter <- c(para$prefilter,para$value_of_prefilter)
  integrate <- para$integrate
  profStep <- para$profStep
  center <- para$center
  response <- para$response
  gapInit <- para$gapInit
  gapExtend <- para$gapExtend
  factorDiag <- para$factorDiag
  factorGap <- para$factorGap
  localAlignment <- para$localAlignment
  bw <- para$bw
  mzwid <- para$mzwid
  minfrac <- para$minfrac
  minsamp <- para$minsamp
  max <- para$max
  distFunc <- para$distFunc
  plottype <- para$plottype
  retcorMethod <- para$retcorMethod
  mzCenterFun <- para$mzCenterFun
  fitgauss <- para$fitgauss
  verboseColumns <- para$verbose.columns
  files <- list.files(path,full.names = T,recursive = T)
  if(!is.null(index)){
    files <- files[index]
  }
  xset <- xcmsSet(files,
                  method = "centWave",
                  peakwidth       = peakwidth,
                  ppm             = ppm,
                  noise           = noise,
                  snthresh        = snthresh,
                  mzdiff          = mzdiff,
                  prefilter       = prefilter,
                  mzCenterFun     = mzCenterFun,
                  integrate       = integrate,
                  fitgauss        = fitgauss,
                  verbose.columns = verboseColumns)
  xset <- retcor(
    xset,
    method         = retcorMethod,
    plottype       = plottype,
    distFunc       = distFunc,
    profStep       = profStep,
    center         = center,
    response       = response,
    gapInit        = gapInit,
    gapExtend      = gapExtend,
    factorDiag     = factorDiag,
    factorGap      = factorGap,
    localAlignment = localAlignment)
  xset <- group(
    xset,
    method  = "density",
    bw      = bw,
    mzwid   = mzwid,
    minfrac = minfrac,
    minsamp = minsamp,
    max     = max)

  xset <- fillPeaks(xset)
  return(xset)
}
```

11.4.1 Peak picking

The first step to process the MS data is that find the peaks against the noises. In **xcms**, all of related staffs are handled by *xcmsSet* function.

For any functions in **xcms** or **R**, you could get their documents by type `?` before certain function. Another geek way is input the name of the function in the console of Rstudio and press F1 for help.

```
?xcmsSet
```

In the document of *xcmsset*, we could set the sample classes, profmethod, profparam, polarity,etc. In the online version, such configurations are shown in certain windows. In the local analysis environment, such parameters are setup by yourselves. However, I think the default configurations could satisfied most of the analysis because related information should have been recorded in your Raw data and **xcms** could find them. All you need to do is that show the data dictionary for *xcmsSet*.

If your data have many groups such as control and treated group, just put them in separate subfolder of the data folder and *xcmsSet* would read them as separated groups.

11.4.2 Data correction

Reasons of data correction might come from many aspects such as the unstable instrument and pollution on column. In **xcms**, the most important correction is retention time correction. Remember the original retention time might changed and use another object to save the new object.

11.4.3 Peaks filling

After the retention time correction, filling the missing peaks could be done by *fillpeaks*. Peaks filling could avoid two many NAs for false statistical analysis. The algorithm could use the baseline signal to impute the data.

11.5 Peaks list

Then we could extract the peaks list from `xcmsSet` objects.

```
library(enviGCMS)
# get the xcmsset object
path <- system.file("extdata/data", package = "rmwf")
# use your own data
# path <- 'path/to/your/file'
srm <- getrtmz(path)
# back up the xcmsset object, xcmsEIC object and peak list
mzrt <- enviGCMS::getmzrt(srm, name = 'srm', eic = T, type = 'mapo')
```

11.6 Peaks filtering

After we get the peaks list, it's nessary to filter the peaks list to retain the peaks with high quality for further analysis. Normally, we could use the relative standard deviation of blank and pooled QC samples to control the peaks list.

```
load('srmmzrt.RDS')
# get the peak intensity, m/z, retention time and group information as list
mzrt <- enviGCMS::getmzrt(srm)
data(mzrt)
# get the mean and rsd for each group
mzrtm <- enviGCMS::getdoe(mzrt)
gm <- mzrtm$groupmean
gr <- mzrtm$grouprsd
# find the blank group and pool QC group, demo data only have matrix blank
srm <- grepl('NIST',colnames(gm))
blk <- grepl('Matrix',colnames(gm))
# pqc <- grepl('pool',colnames(gm))
# filter by pool QC and blank's group mean intensity(pool QC should larger than three
# in demo data, use sample average intensity for each peak
sum(indexmean <- apply(gm,1,function(x) all(x[srm]>= 3*x[blk])))
# filter by pool qc rsd%, return numbers and index
# in demo data, use sample average intensity for each peak
# mean rsd analysis
# filter by pool qc rsd%, return numbers and index, here we use SRM samples
rsdcf <- 30
sum(indexrsd <- apply(gr,1,function(x) ifelse(is.na(x[srm]),T,x[srm]<rsdcf)))
# overlap with rsd% and mean filter
sum(index <- indexmean&indexrsd)
# new list, update group and remove pool qc/blk
# qcindex <- grepl('k',mzrt$group$class) | grepl('q',mzrt$group$class)
# mzrtfilter <- mzrtfilter <- enviGCMS::getfilter(mzrt,rowindex = index,colindex = !qc
mzrtfilter <- mzrtfilter <- enviGCMS::getfilter(mzrt,rowindex = index, name = 'lif', t
```

11.7 Normalization (Optional)

Normailizaiton is nesscery if you data are collected at different batch or runned in differen instrument parameters.

```
# visulize the batch effect
mzrtsim::rlaplot(mzrt$data,lv = mzrt$group$class)
mzrtsim::ridgesplot(mzrt$data,lv = mzrt$group$class)
# get the simulation data and test on NOREVA
```

```
sim <- mzrtsim::simmzrt(mzrt$data)
mzrtsim::simdata(sim)
# correct the batch effect by sva
mzrtcor <- mzrtsim::svacor(mzrt$data,lv = mzrt$group$class)
# visulize the batch effect correction
li <- mzrtsim::limmaplot(mzrtcor,lv = mzrt$group$class)
# return the corrected data
mzrt$data <- mzrtcor$dataCorrected
```

11.8 Statistic analysis

Here we could use `caret` package to perform statistical analysis.

```
library(caret)
## Spliting data
trainIndex <- createDataPartition(mzrtfilter$data, p = .8,
                                  list = FALSE,
                                  times = 1)
## Get the training and testing datasets
Train <- data[ trainIndex,]
Test  <- data[-trainIndex,]
## Set the cross validation method
fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,
                           ## repeated ten times
                           repeats = 10)
# extra papameters for GBM
gbmGrid <-  expand.grid(interaction.depth = c(1, 5, 9),
                        n.trees = (1:30)*50,
                        shrinkage = 0.1,
                        n.minobsinnode = 20)

set.seed(825)
gbmFit <- train(mzrtfilter$group ~ ., data = training,
                 method = "gbm",
                 trControl = fitControl,
                 verbose = FALSE,
                 ## Now specify the exact models
                 ## to evaluate:
                 tuneGrid = gbmGrid)
# show the fitting process
plot(gbmFit)
```

```
# ANOVA analysis for model selection
anova(fit1,fit2)
# find the important variables
Imp <- varImp(fit)
plot(Imp)
```

11.9 Annotation

Here I use `xMSannotator` package to make annotation with HMDB as reference database.

```
library(xMSannotator)
num_nodes = 10
data("adduct_weights")
negsub <- getrtmz('D:/metademo/data/oq/')
anno <- xsetplus::fanno(negsub,outloc = 'D:/metademo/anno',mode = 'neg')
```

11.10 Pathway Analysis

We could output the files for online pathway analysis with mummichog algorithm.

```
# get the file
xsetplus::mumdata(neg,lv = mzrt$group$class)
# http://mummichog.org/index.html
# mummichog1 -f 'test.txt' -o myResult
```

11.11 MetaboAnalyst

Actully, after you perform data correction, you have got the data matrix for statistic analysis. You might choose **MetaboAnalyst** online or offline to make furthor analysis, which supplied more statistical choices than xcms.

The input data format for **MetaboAnalyst** should be rows for peaks and colomns for samples. You could also add groups infomation if possible. Use the following code to get the data for analysis.

```
# get the csv file for Metaboanalyst.ca
enviGCMS::getupload(neg,name = 'metabo')
```

11.12 Summary

This is the offline metaboliomics data process workflow. For each study, details would be different and F1 is always your best friend.

Enjoy yourself in data mining!

Chapter 12

Exposome

Nature or nurture debate has a similar paradigm in environmental study: is the ecological system and human health risk dominated by heredity or environment? Twins and siblings study(**??**) show that both heritability and environmental factors could explain the phenotypic variance among population. The contribution of environment among different disease functional domain such as hematological and endocrine could achieve almost half of the total variances (**?**). However, besides those epidemiology proof, little is known about the influences of overall environmental exposure process at molecular level. Conventional exposure study always investigate one or several specific compounds and their environmental fate or toxicology endpoint. Exposome, on the other hand, tries to access multiple exposure factors from biological or environmental samples as much as possible without a predefined compounds list. Those endogenous and exogenous molecules can reveal the exposure process in details. Exposome could not only help to investigate the comprehensive molecules level changes, but also the interactions among molecules in an non-targeted design. By following annotation of captured compounds, exposome can discover exposure markers for certain type of pollution, as well as biomarkers for certain exposure process and discuss related physiological process.

According to CDC, The exposome can be defined as the measure of all the exposures of an individual in a lifetime and how those exposures relate to health. Exposomics is the study of the exposome and relies on the application of internal and external exposure assessment methods.

- Internal exposure relies on fields of study such as genomics, metabolomics, lipidomics, transcriptomics and proteomics.

- External exposure assessment relies on measuring environmental stressors.

12.1 Internal exposure

- HMDB is a freely available electronic database containing detailed information about small molecule metabolites found in the human body.

- Lipid Maps

- GMDB a multistage tandem mass spectral database using a variety of structurally defined glycans.

- KEGG is a collection of small molecules, biopolymers, and other chemical substances that are relevant to biological systems.

- Virtual Metabolic Human Database integrating human and gut microbiome metabolism with nutrition and disease.

12.2 External exposure

12.2.1 Environmental fate of compounds

12.2.1.1 QSPR

- Chemicalize is a powerful online platform for chemical calculations, search, and text processing.

- QSPR molecular descriptor generate tools list

- Spark uses computational algorithms based on fundamental chemical structure theory to estimate a wide variety of reactivity parameters strictly from molecular structure.

- OPERA OPERA models for predicting physicochemical properties and environmental fate endpoints(**?**).

LogP is important for analytical chemistry. Mannhold (**?**) report a comprehensive comparison of logP algorithms. Later, Rajarshi Guha make a comparison with logP algorithms with CDK based on logPstar dataset. Commercial software such as Spark, ACS Labs and ChemAxon might always claim a better performance on in-house dataset compared with public software like KowWIN within EPI Suite. However, we should be careful to evaluate the influnce of logP accuracy on the metabolites or unknown compounds.

12.2.1.2 Fate

- Wania Group developed software tools to address various aspects of organic contaminant fate and behaviour.
- Trent University release models to predict environmental fate for pollutions such as Level 3.
- EAWAG-BBD could provide information on microbial enzyme-catalyzed reactions that are important for biotechnology.

12.2.2 Exposure study database

- The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research.
- Environmental Health Criteria (EHC) Monographs
- CTD is a robust, publicly available database that aims to advance understanding about how environmental exposures affect human health.
- ODMOA facilitates and coordinates the collection, access to, and use of public health data in order to monitor and improve population health. This data is better for general public health research for Massachusetts.
- The Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population.
- CompTox compounds, exposure and toxicity database. Here is related data.
- T3DB is a unique bioinformatics resource that combines detailed toxin data with comprehensive toxin target information.
- FooDB is the world's largest and most comprehensive resource on food constituents, chemistry and biology.
- Phenol explorer is the first comprehensive database on polyphenol content in foods.
- Drugbank is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.
- LMDB is a freely available electronic database containing detailed information about small molecule metabolites found in different livestock species.
- HPV High Production Volume Information System