BIOF3 组学数据分析

01 质谱结果表与实验设计

导出日期：2026年5月12日

01 质谱结果表与实验设计

这一章的目标是让读者在动手做差异蛋白分析之前，先搞清楚"手上拿到的是什么格式的表、每一列代表什么、实验设计怎么对应到分析里"。

质谱数据处理的起点

蛋白质组学的原始数据是质谱仪产出的 `.raw` / `.mzML` 文件。这些文件需要经过数据库搜索（DDA）或者谱图库匹配（DIA）才能得到蛋白鉴定和定量结果。常用的处理工具：

工具	适用	输出
MaxQuant	DDA	`proteinGroups.txt` 、 `evidence.txt`
FragPipe	DDA	`combined_protein.tsv`
DIA-NN	DIA	`report.pg_matrix.tsv`
Spectronaut	DIA	自定义导出表

BioF3 的蛋白组教程**从这些工具的输出表开始**，不涉及原始质谱数据处理。

MaxQuant proteinGroups.txt 的关键列

MaxQuant 是目前最常用的 DDA 处理工具。它的 `proteinGroups.txt` 是一张宽表，每行一个蛋白组（protein group），关键列：

列名	含义
`Protein IDs`	UniProt accession，多个用分号分隔
`Gene names`	基因名（用于可视化和富集）
`LFQ intensity XXX`	每个样本的 Label-Free Quantification 强度
`iBAQ XXX`	另一种定量方式（绝对丰度估计）
`Razor + unique peptides`	用于该蛋白定量的肽段数
`Reverse`	是否匹配到反向库（应过滤掉）
`Potential contaminant`	是否是常见污染物（应过滤掉）
`Only identified by site`	是否仅通过修饰位点鉴定（通常过滤掉）

读入后第一件事：过滤掉 `Reverse == "+"` 和 `Potential contaminant == "+"` 的行。这些不是真正的样本蛋白。

实验设计表

和 bulk RNA-seq 一样，蛋白组分析也需要一张样本表把"哪个列对应哪个条件"说清楚。DEP 包要求的格式：

```
label      condition  replicate
Sample1    Control    1
Sample2    Control    2
Sample3    Control    3
Sample4    Treatment  1
Sample5    Treatment  2
Sample6    Treatment  3
```

- `label` 要和 `proteinGroups.txt` 里 `LFQ intensity` 列名的后缀对得上
- `condition` 是分组变量（后面做差异分析的对比就基于它）
- `replicate` 是生物学重复编号

缺失值：蛋白组的核心难题

和转录组最大的区别：蛋白组的缺失值非常多。一个蛋白在某些样本里完全没有信号（LFQ = 0 或 NA），原因可能是：

- **MNAR（Missing Not At Random）**：蛋白丰度太低，低于检测限 → 这种缺失和真实丰度有关
- **MCAR（Missing Completely At Random）**：随机的技术波动 → 和丰度无关

两种缺失的处理策略不同：MNAR 通常用"从分布左尾采样"（MinProb）来插补；MCAR 可以用 KNN 或均值插补。DEP 默认用 MinProb，适合大多数蛋白组场景。

从 proteinGroups.txt 到 DEP 对象

DEP 包把上面这些步骤封装成几个函数：

```
library(DEP)

# 1. 读入 MaxQuant 结果
data <- read.delim("proteinGroups.txt")

# 2. 过滤污染和反向库
data <- data[data$Reverse != "+", ]
data <- data[data$Potential.contaminant != "+", ]

# 3. 确保蛋白名唯一
data_unique <- make_unique(data, "Gene.names", "Protein.IDs", delim = ";")

# 4. 构建 SummarizedExperiment
lfq_cols <- grep("LFQ.intensity.", colnames(data_unique))
data_se <- make_se(data_unique, lfq_cols, experimental_design)
```

`make_se` 之后得到的是一个标准的 Bioconductor `SummarizedExperiment` 对象，后续的过滤、归一化、插补、差异分析都在这个对象上操作。

下一步

- 02 DEP 差异蛋白分析

- 03 功能富集与 Reactome 通路

参考资源

- MaxQuant 官方文档
- DEP Bioconductor 文档
- Perseus 教程（MaxQuant 配套可视化）
- DIA-NN 文档

扫码关注微信公众号【生信F3】

获取文章完整内容，分享生物信息学最新知识。