BIOF3 组学数据分析

蛋白质组学实践教程

导出日期：2026年6月27日

蛋白质组学实践教程

转录组告诉你哪些基因被转录，蛋白质组告诉你哪些蛋白真的被翻译、它们的丰度如何、以及是否被修饰。质谱是目前最常用的蛋白组测量手段：样本酶解后肽段上机，仪器给出质荷比和强度，数据库比对后得到蛋白身份和定量值。

做过转录组分析的读者上手蛋白组往往会遇到两个反差：数据维度小（几千个蛋白 vs 几万个基因）、缺失值多（某些蛋白在部分样本根本没测到）。因此思路上 bulk RNA-seq 的很多经验可以复用，但预处理和统计模型要调整。

本专栏假设原始质谱数据已经由 MaxQuant、FragPipe、DIA-NN、Spectronaut 等工具处理好，起点是一份带强度或 LFQ 值的蛋白表。

一个项目大致是什么样子

给你 20 个样本的 MaxQuant `proteinGroups.txt` 或 DIA-NN `report.pg_matrix.tsv`，从这里到一张能写进论文的差异蛋白结果，大致要走：

步骤	典型产物	常用工具
结果表清洗	去除污染、反向库、仅 1 肽识别的蛋白	R，dplyr
缺失值评估	每个样本缺失率分布	naniar、VIM
缺失值插补	完整的强度矩阵	MinProb、KNN、DEP
归一化	按中位数 / 中位线对齐	DEP、limma 的 normalizeBetweenArrays
批次效应检查	PCA、样本相关性	ggplot2、ComplexHeatmap
差异蛋白分析	差异蛋白表	limma、DEP、MSstats
功能富集	GO、Reactome、pathway 结果	clusterProfiler、ReactomePA
相互作用与网络	蛋白-蛋白互作图	STRING、Cytoscape

MaxQuant 和 FragPipe 是 DDA 的两条主流选择，DIA-NN 和 Spectronaut 面向 DIA 实验。处理好后的结果表在字段名上有差异，但后续的差异分析和富集分析可以共用。

常见工具栈

差异分析这一段，`limma` 仍然是最稳的选择，因为它对样本量小、噪音大、存在技术重复的实验设计非常友好。`DEP` 在 limma 之上包了一层适合蛋白组的数据流程，适合初学者。

阶段	工具	运行环境
原始数据处理	MaxQuant、FragPipe、DIA-NN、Spectronaut	GUI / bash
数据读取	MSnbase、QFeatures、DEP	R
缺失值处理	DEP、imputeLCMD	R
差异分析	limma、DEP、MSstats	R
功能注释	clusterProfiler、ReactomePA	R
蛋白互作	STRING 网页、STRINGdb、Cytoscape	R / 网页
可视化	ggplot2、ComplexHeatmap、EnhancedVolcano	R

推荐公开数据集

蛋白组的公开数据比转录组少，但以下几个入口能覆盖大多数教学需求：

数据集	类型	适合	入口
PRIDE 数据集	各种质谱项目	找到文献配套原始/结果	PRIDE
MassIVE	各种质谱项目	北美研究为主	MassIVE
DEP 包 `UbiLength`	8 样本泛素化实验	差异分析最小入门数据	DEP Bioconductor
CPTAC	癌症蛋白组 + 表型	多组学整合练习	CPTAC

`DEP` 包自带的 `UbiLength` 数据是 HeLa 细胞不同处理时间的泛素化组实验，8 个样本、大约 2700 个蛋白，规模合适教学，下面的最小例子也基于它。

最小可跑的例子

这段代码基于 `DEP` 包自带的真实数据。安装好包后能直接运行完成："读入 → 过滤 → 归一化 → 插补 → 差异分析 → 可视化" 一条链路：

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

library(DEP)

# 自带真实数据：UbiLength MaxQuant 结果
data <- UbiLength
experimental_design <- UbiLength_ExpDesign

# 清洗：去除污染蛋白、反向库匹配、仅 1 肽鉴定的蛋白
data_unique <- make_unique(data, "Gene.names", "Protein.IDs", delim = ";")
data_se <- make_se(data_unique, grep("LFQ.", colnames(data_unique)), experimental_design)

# 过滤：每组至少两个样本里被测到
data_filt <- filter_missval(data_se, thr = 0)

# 归一化 + 缺失值插补
data_norm <- normalize_vsn(data_filt)
data_imp  <- impute(data_norm, fun = "MinProb", q = 0.01)

# 差异分析（limma 后端）
data_diff <- test_diff(data_imp, type = "control", control = "Ctrl")
dep <- add_rejections(data_diff, alpha = 0.05, lfc = 1)

# 火山图
plot_volcano(dep, contrast = "Ubi6_vs_Ctrl", label_size = 2, add_names = TRUE)
```

跑完后得到的是一张火山图，横轴是 log2 fold change，纵轴是 `-log10(p)`。被标注出来的蛋白就是这组对比下显著差异的候选，可以再去做富集分析或文献查证。

真实项目里有几件事比这段复杂：样本分组更多、协变量要纳入模型、DIA 数据的插补策略不同、要输出 MSstats 报告等等，但起步的框架就是这样。

专栏模块规划

模块	主题	状态
01	质谱结果表格式与实验设计	已上线
02	DEP 差异蛋白分析	已上线
03	功能富集与 Reactome 通路	已上线
04	可视化：火山图、热图与蛋白互作	已上线
05	limma 直接建模（不用 DEP 包装）	规划中
06	DIA 数据的处理差异	规划中
07	STRING 与蛋白互作网络	规划中
08	转录组与蛋白组联合分析	规划中

目前 01-04 已上线，每个模块带一份可直接跑的 R 脚本。

推荐前置知识

- 编程基础：R、Python、Bash 学习路径
- R 数据整理与 ggplot2 可视化
- bulk RNA-seq 实践教程（差异分析思路有很多共通点）

常见坑

坑 1：忽略缺失值的非随机性

蛋白组的缺失不是随机的——低丰度蛋白更容易缺失（MNAR）。如果直接删除含缺失的行，会系统性丢失低丰度蛋白，后续差异分析只能看到高丰度蛋白的变化。正确做法是评估缺失模式后选择合适的插补策略（MinProb 适合 MNAR，KNN 适合 MCAR）。

坑 2：batch effect 在质谱中比 RNA-seq 更隐蔽

质谱的批次效应来源很多：上样量、色谱柱老化、仪器维护前后性能差异。而且不像 RNA-seq 有明确的 batch 标签，质谱的"batch"可能是连续上机的顺序效应。建议用 PCA 标注上机日期/顺序，发现聚类按 batch 而不是按条件分的就要做校正。

坑 3：LFQ intensity 和 iBAQ 混用

LFQ 适合样本间同一蛋白的相对比较（差异分析），iBAQ 适合同一样本内不同蛋白的丰度排序。用 iBAQ 做差异分析或者用 LFQ 做蛋白丰度排名都是错误的。选定量方式时要明确分析目标。

坑 4：DDA 和 DIA 结果直接合并

DDA（数据依赖采集）和 DIA（数据非依赖采集）的定量原理不同，缺失模式也不同。DDA 缺失多为 MNAR，DIA 缺失更接近 MCAR。两种采集模式的结果不能简单 `rbind` 后当作同一批数据分析。

下一步

接着深入：

- 01 质谱结果表与实验设计 — 理解数据格式和缺失值来源
- 02 DEP 差异蛋白分析 — 动手跑第一个差异分析流程

横向延伸：

- bulk RNA-seq 实践教程 — 差异分析的统计框架和蛋白组相通
- 多组学整合实践教程 — 蛋白组与转录组的联合分析

参考资源

- DEP Bioconductor 文档
- limma 用户手册
- MaxQuant 官方文档
- FragPipe
- DIA-NN
- PRIDE Archive
- STRING 数据库

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。