BIOF3 组学数据分析

06 SNF 相似性网络与亚型识别

导出日期：2026年5月12日

06 SNF 相似性网络与亚型识别

SNF（Similarity Network Fusion）的思路和 MOFA 完全不同：它不做矩阵分解，而是为每一层组学数据分别构建样本相似性网络，然后把这些网络融合成一个统一的网络。融合后的网络可以用谱聚类（spectral clustering）来识别亚型。

算法直觉

1. 对每一层组学数据，计算样本间的距离矩阵（通常用欧氏距离）
2. 把距离矩阵转换成相似性矩阵（用 scaled exponential kernel）
3. 构建 K 近邻图（KNN graph）
4. 迭代融合：每一层的网络向其他层"扩散"信息，最终收敛到一个融合网络

融合的核心思想是：如果两个样本在多层数据中都相似，它们在融合网络中的连接会被加强；如果只在一层相似，连接会被削弱。

基本流程

```
library(SNFtool)

# 输入：每层数据矩阵，行是样本，列是特征
rna_t  <- t(rna_final)
meth_t <- t(meth_final)
prot_t <- t(prot_final)

# 参数
K <- 20          # 近邻数
alpha <- 0.5     # 超参数
iter <- 20       # 迭代次数

# 1. 计算距离矩阵
dist_rna  <- dist2(as.matrix(rna_t), as.matrix(rna_t))
dist_meth <- dist2(as.matrix(meth_t), as.matrix(meth_t))
dist_prot <- dist2(as.matrix(prot_t), as.matrix(prot_t))

# 2. 构建相似性网络
W_rna  <- affinityMatrix(dist_rna, K, alpha)
W_meth <- affinityMatrix(dist_meth, K, alpha)
W_prot <- affinityMatrix(dist_prot, K, alpha)

# 3. 融合网络
W_fused <- SNF(list(W_rna, W_meth, W_prot), K, iter)
```

谱聚类确定亚型

```
# 用 eigen-gap 方法估计最优聚类数
C_best <- estimateNumberOfClustersGivenGraph(W_fused, NUMC = 2:8)
cat("推荐聚类数:", C_best$`Eigen-gap best`, "\n")

# 谱聚类
clusters <- spectralClustering(W_fused, K = C_best$`Eigen-gap best`)
table(clusters)

# 可视化融合网络
library(pheatmap)
pheatmap(W_fused,
         annotation_row = data.frame(Cluster = factor(clusters)),
         show_rownames = FALSE, show_colnames = FALSE,
         main = "Fused similarity network")
```

与 Consensus Clustering 比较

SNF 和 consensus clustering（如 `ConsensusClusterPlus`）都能做亚型发现，但思路不同：

方面	SNF	Consensus Clustering
输入	多层数据分别构建网络	通常用拼接后的单矩阵
整合方式	网络融合（中期整合）	重采样 + 聚类稳定性（早期整合）
对噪声层的鲁棒性	较好，噪声层贡献会被削弱	较差，噪声层会污染拼接矩阵
缺失值	不支持	不支持
聚类数选择	eigen-gap	CDF / delta area

实际项目中可以两种都跑，看结果是否一致。如果一致，结论更可信。

```
library(ConsensusClusterPlus)

# 拼接矩阵
combined <- rbind(rna_final, meth_final, prot_final)

# Consensus clustering
cc_results <- ConsensusClusterPlus(combined,
                                   maxK = 8,
                                   reps = 500,
                                   pItem = 0.8,
                                   pFeature = 1,
                                   clusterAlg = "hc",
                                   distance = "pearson",
                                   seed = 42,
                                   plot = "pdf")
```

亚型与临床表型关联

```
# 把聚类结果和临床信息合并
cluster_df <- data.frame(
  sample  = rownames(rna_t),
  cluster = factor(clusters),
  subtype = col_data$subtype,
  stage   = col_data$stage
)

# Fisher 检验：聚类 vs 已知亚型
fisher.test(table(cluster_df$cluster, cluster_df$subtype))

# 生存分析
library(survival)
library(survminer)

surv_obj <- Surv(col_data$os_time, col_data$os_event)
fit <- survfit(surv_obj ~ cluster_df$cluster)
ggsurvplot(fit, data = cluster_df, pval = TRUE,
           palette = "jco", risk.table = TRUE)
```

参数敏感性

SNF 对 K（近邻数）比较敏感。建议在 K = 10-30 范围内扫描，看聚类结果是否稳定。alpha 通常固定为 0.5，迭代次数 20 次足够收敛。

参考资源

- SNFtool CRAN
- SNF 原始论文 (Wang et al. 2014)
- ConsensusClusterPlus Bioconductor
- spectral clustering 原理