使用TCseq包分析基因表达的时间趋势并划分聚类群

上一篇介绍了如何使用[Mfuzz包](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247486207&idx=1&sn=7e4b7e062f912220e62f087aeb6e0f2e&chksm=97f5bae7a08233f19837ce2c2688a50b38e117a984d7d5df2a40579bd051f3fad229d1da9f9f&token=1154049339&lang=zh_CN#rd)在具有时间序列特点的转录组、蛋白质组数据中分析基因或蛋白表达的时间趋势，并将具有相似表达模式的基因或蛋白划分聚类。事实上，能够实现类似功能（时间趋势分析、聚类以及可视化作图等）的R包还有很多，本篇继续带来另一个R包的教程，TCseq包。

除了趋势分析和聚类，TCseq包也可以进行差异分析，貌似是调用的[edgeR包](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247484213&idx=1&sn=2e7bfc5df3ef4ecd34514807494e17d5&chksm=97f5b32da0823a3b04f2469546606b21cce432088a659131a8baa494efcdfb126e006d69e45f&token=66128012&lang=zh_CN&scene=21#wechat_redirect)的方法？不过本篇不涉及TCseq的差异分析，大家有兴趣可自行摸索。本篇主要通过一个涉及时间序列的蛋白质组学数据集，简单演示如何在R语言中使用TCseq包分析蛋白质表达的时间趋势，并根据时间表达模式的相似性实现聚类的过程。

# 使用TCseq包分析基因表达的时间趋势并划分聚类群的简单演示

下文中所使用的示例数据和R代码的百度盘链接（提取码，xijb）：

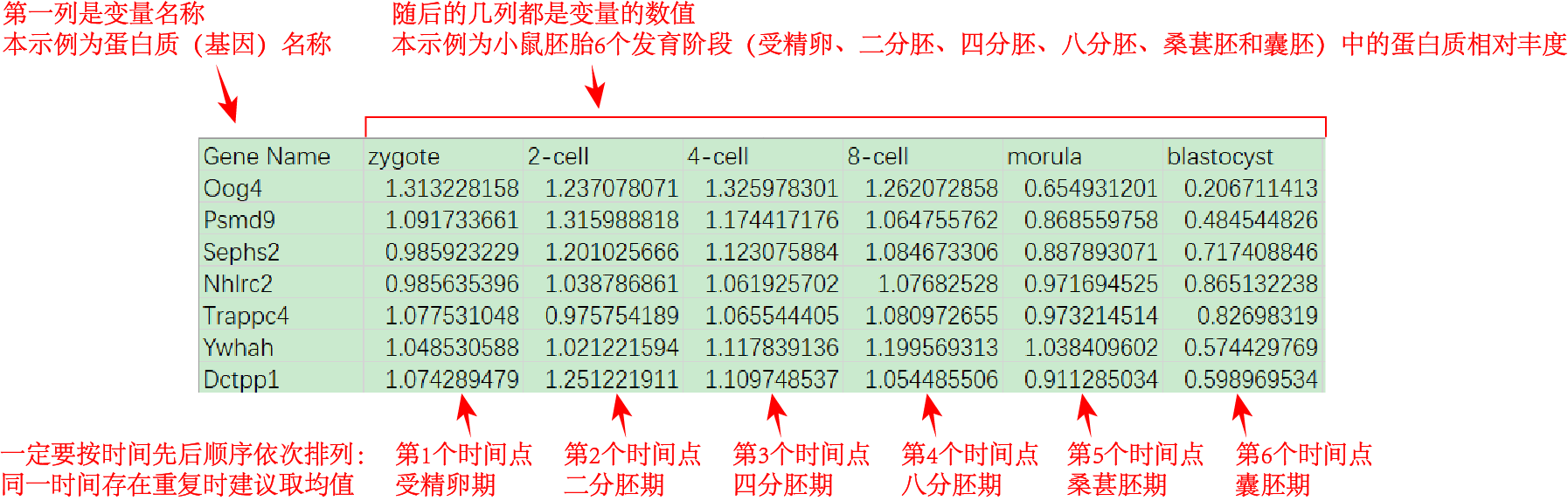
<https://pan.baidu.com/s/1o_MltUDq7_mGFznAIVEx9g>

若百度盘失效，也可在GitHub的备份中获取：

<https://github.com/lyao222lll/sheng-xin-xiao-bai-yu-2021>

## 示例数据集概要

网盘附件“protein\_exp.txt”是一个蛋白质表达矩阵，基于蛋白质谱的方法获得的小鼠胚胎着床前发育过程中的蛋白质组，包括受精卵（Zygote）、二分胚（2-cell）、四分胚（4-cell）、八分胚（8-cell）、桑葚胚（Morula）和囊胚（Blastocyst）共6个发育阶段。表格第一列为蛋白质名称，随后几列依次为这些蛋白质在小鼠胚胎着床前发育的6个阶段中的相对丰度数值。



## 使用TCseq包执行时间序列的聚类分析

为了阐明与小鼠胚胎发育有关的功能蛋白质，或者寻找在胚胎特定阶段发挥重要功能的关键蛋白质，我们首先期望分析蛋白质丰度随胚胎发育阶段的时间趋势，并根据蛋白质丰度的不同时间动力学模式对蛋白质划分功能群。在这里，就可以根据所有蛋白质在每个阶段的丰度信息，通过TCseq包对这些蛋白质执行时间序列的聚类。

TCseq包可使用bioconductor安装。加载TCseq包，将上述数据表读取到R中，转换为矩阵类型后，直接作为聚类函数timeclust()的输入。timeclust()是一个整合函数，可执行数据标准化、聚类等多步操作，将上述输入数据中具有相似的时间表达特征的蛋白聚在一类。

#使用 Bioconductor 安装 TCseq 包

#install.packages('BiocManager')

#BiocManager::install('TCseq')

#加载 TCseq 包

library(TCseq)

#以本示例的蛋白表达矩阵为例，行为基因或蛋白名称，列为时间样本

#每一列是独立的时间单位，按时间顺序提前排列好，若存在生物学重复（即一个时间点对应多个样本时）建议提前取均值

protein <- read.delim('protein\_exp.txt', row.names = 1, check.names = FALSE)

protein <- as.matrix(protein)

#聚类，详情 ?timeclust

#algo 用于指定聚类方法，例如基于 fuzzy c-means 的算法进行聚类，此时需要设定随机数种子，以避免再次运行时获得不同的结果

#k 用于指定期望获得的聚类群数量，例如这里预设为 10

#standardize 用于 z-score 标准化变量

set.seed(123)

cluster\_num <- 10

tcseq\_cluster <- timeclust(protein, algo = 'cm', k = cluster\_num, standardize = TRUE)

#作图，详情 ?timeclustplot

#颜色、线宽、坐标轴、字体等细节可以在函数中调整，具体参数详见函数帮助

p <- timeclustplot(tcseq\_cluster, value = 'z-score', cols = 3,

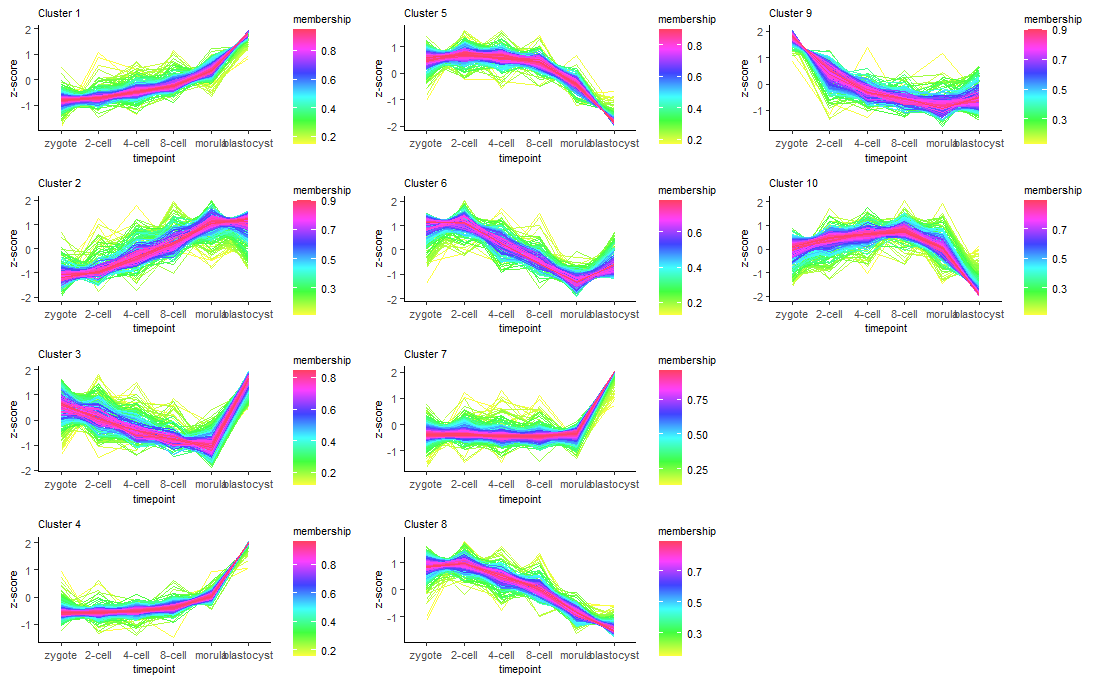
axis.line.size = 0.6, axis.title.size = 8, axis.text.size = 8,

title.size = 8, legend.title.size = 8, legend.text.size = 8)

#上述获得了 10 组聚类群

#如果绘制单个的聚类群，例如 claster 2，直接在作图结果中输入下标选取

p[2]



如上示例中，基于[模糊c均值聚类](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247484616&idx=2&sn=8e650c5ab700203fa296d2716c060ac9&chksm=97f5b4d0a0823dc69abb9f2f1d5b2dd7cdbf09316fcc3157f6436ff705ff63e89e8261c4613a&token=66128012&lang=zh_CN&scene=21#wechat_redirect)（timeclust()参数algo='cm'）的原理对蛋白质表达值的时间序列进行了聚类。timeclust()还提供了其它的聚类算法，如[层次聚类](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247484593&idx=1&sn=d8160e2918b0aa4dd90cb10c2f82d266&chksm=97f5b4a9a0823dbf52ebacc1c1390c2047dfafcf217ae9afe34583dfe4a999621d6d274c4af2&token=66128012&lang=zh_CN&scene=21#wechat_redirect)（参数algo='hc'）、[k均值划分](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247484607&idx=1&sn=c69c7161f3b6e31433dc52afc68a7831&chksm=97f5b4a7a0823db1adfd991e10127ab965777d153d9aaeb9ce5a0d99d872f47f2e6bfd7d805e&token=66128012&lang=zh_CN&scene=21#wechat_redirect)（参数algo='km'）、[围绕中心点划分](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247484607&idx=1&sn=c69c7161f3b6e31433dc52afc68a7831&chksm=97f5b4a7a0823db1adfd991e10127ab965777d153d9aaeb9ce5a0d99d872f47f2e6bfd7d805e&token=66128012&lang=zh_CN&scene=21#wechat_redirect)（参数algo='pam'）等，您也可以尝试。根据预先指定的聚类数量，最终获得了10组不同动力学模式的聚类群（蛋白群）。对于每个聚类群中的蛋白质，它们具有相似的时间表达特征；而不同聚类群的蛋白质之间的动力学模式则差异明显。

在获得了聚类结果后，即可从图中识别一些重要的或者感兴趣的蛋白集合，比方说某些聚类群的蛋白质出现了预期的随时间增加而增加或减少的趋势，在特定时间点出现了相对更高或更低的表达，或者观察到明显的拐点等。并继续对这些感兴趣的蛋白质进行功能分析（如基因集富集分析，蛋白网络分析等），以及建立和细胞或生物体的表型特征的联系等，讨论它们的生物学意义。

## 提取各聚类群的样本或变量名称和数值

当然，讨论蛋白质的功能不是本篇的内容，后续的分析需要做哪些，您自己根据实际情况来。在这之前，一个有待解决的问题是，如何获得各聚类群中，都包含哪些蛋白呢？

接下来继续在上述已获得的聚类结果中，提取10个聚类群中包含的蛋白质集合。

#查看每个蛋白所属的聚类群，展示前几个为例

head(tcseq\_cluster@cluster)

#统计每个聚类群中各自包含的蛋白数量

table(tcseq\_cluster@cluster)

#上述聚类过程中，通过计算 membership 值判断蛋白质所属的聚类群，以最大的 membership 值为准

#查看本次计算的各蛋白的 membership 值，展示前几个为例

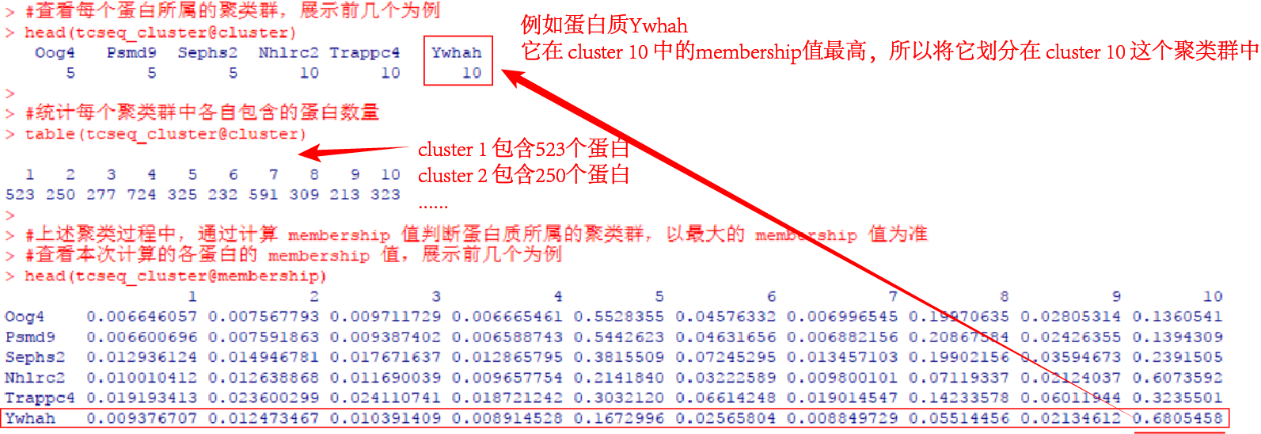
head(tcseq\_cluster@membership)

#上述聚类过程中，我们在聚类函数 timeclust() 中指定了对蛋白表达值的 z-score 标准化

#如果您想查看标准化后的表达值（也即绘制曲线图用的那个值，而非原始的蛋白表达值），展示前几个为例

head(tcseq\_cluster@data)

#啥？您若问原始的蛋白表达值在哪里看？一开始的输入数据就是啊......



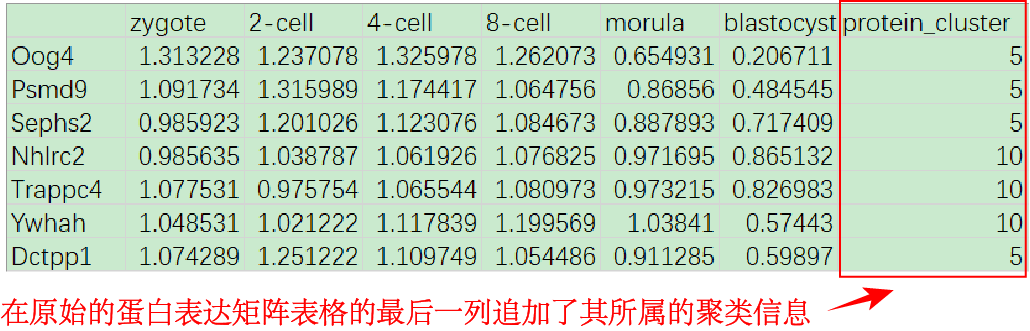
#最后，提取所有蛋白所属的聚类群，并和它们的原始表达值整合在一起

protein\_cluster <- tcseq\_cluster@cluster

protein\_cluster <- cbind(protein[names(protein\_cluster), ], protein\_cluster)

head(protein\_cluster)

write.table(protein\_cluster, 'protein\_cluster.txt', sep = '\t', col.names = NA, quote = FALSE)



这样，就将蛋白名称、蛋白表达值以及其所属的聚类群对应起来了。如果根据上文的折线图挑选出了感兴趣的时间表达特征的聚类群，就可以在该表中进一步将这些聚类群中的蛋白质信息提取出来。再往后，分析这些蛋白的功能等，不再多说。

# 其它一些常见问题

例如“聚类群数量是手动分配的，我聚为几类最佳？”，“其它类型的数据能用TCseq包分析吗？”等问题，可以参考前文“[使用Mfuzz包分析基因表达的时间趋势并划分聚类群](https://mp.weixin.qq.com/s?__biz=MzIxNzc1Mzk3NQ==&mid=2247486207&idx=1&sn=7e4b7e062f912220e62f087aeb6e0f2e&chksm=97f5bae7a08233f19837ce2c2688a50b38e117a984d7d5df2a40579bd051f3fad229d1da9f9f&token=1154049339&lang=zh_CN#rd)”的提示，它们的工作过程差不多。