软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(9):2857−2868 [doi: 10.13328/j.cnki.jos.005512]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

PUseqClust:一种 RNA-seq 数据聚类分析方法*

石险峰[1], 刘学军[1], 张 礼[2]

[1](南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

[2](南京林业大学 信息科学技术学院,江苏 南京 210037)

通讯作者: 刘学军, E-mail:xuejun.liu@nuaa.edu.cn

摘 要: 基因的聚类分析是基因表达数据分析研究的重要技术,它按照表达谱相近原则将基因表达数据归类,探究未知的基因功能.近年来,RNA-seq 技术广泛应用于测量基因表达水平,产生了大量的读段数据,为基因表达聚类分析提供了充分条件.由于读段非均匀分布的特性,对读段计数一般采用负二项分布进行建模.现有的负二项分布算法和传统的聚类算法对于聚类分析都是直接对读段计数进行建模,没有充分考虑实验本身存在的各种噪声,以及基因表达水平测量的不确定性,或者对聚类中心的不确定性考虑不够.基于 PGSeq 模型,模拟读段的随机产生过程,采用拉普拉斯方法考虑多条件多重复基因表达水平之间的相关性,获得了基因表达水平的不确定性,联合混合 t 分布聚类模型,提出 PUseqClust(propagating uncertainty into RNA-seq clustering)框架进行 RNA-seq 读段数据的聚类分析.实验结果表明,该方法相比其他方法获得了更具生物意义的聚类结果.

关键词: RNA-seq;聚类分析;负二项分布;拉普拉斯方法;混合 t 分布

中图法分类号: TP311

中文引用格式: 石险峰,刘学军,张礼.PUseqClust:一种 RNA-seq 数据聚类分析方法.软件学报,2019,30(9):2857−2868. http://www.jos. org.cn/1000-9825/5512.htm

英文引用格式: Shi XF, Liu XJ, Zhang L. PUseqClust: A clustering analysis method for RNA-seq data. Ruan Jian Xue Bao/ Journal of Software, 2019,30(9):2857−2868 (in Chinese). http://www.jos.org.cn/1000-9825/5512.htm

PUseqClust: A Clustering Analysis Method for RNA-Seq Data

SHI Xian-Feng[1], LIU Xue-Jun[1], ZHANG Li[2]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

[2](College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China)

Abstract: Clustering analysis is an important technique for gene expression data analysis. It groups the data according to similar gene expression patterns to explore the unknown gene functions. In recent years, RNA-seq technology has been widely adopted to measure gene expression. It produces a large number of read data, which provide possibilities for clustering analysis of gene expression. In this area, read counts are popularly modeled by the negative binomial distribution to reduce the impact of the non-uniform read distribution, while most existing clustering methods process directly read counts. They donot fully consider the various noise existing in the data, and the uncertainty of gene expression measurements. Some methods also ignore the variability of clustering centers. This study proposes PUseqClust (propagating uncertainty into RNA-Seq clustering) framework for clustering of RNA-seq data. This framework first uses PGSeq to model the stochastic process of read generation. Laplace method is next used to consider correlation between expressions under various conditions and replicates to obtain the uncertainty of expression estimation. Finally, the method adopts the student's t mixture model to perform gene expression clustering. Results show that the proposed methods obtained more biologically relevant clustering results.

* 基金项目: 国家自然科学基金(61170152); 航空基金(20151452021)

Foundation item: National Natural Science Foundation of China (61170152); Aeronautical Science Foundation of China (20151452021)

收稿时间: 2017-01-03; 修改时间: 2017-06-27, 2017-09-17; 采用时间: 2017-11-07

Key words: RNA-seq; clustering analysis; negative binomial distribution; Laplace method; mixture student's t distribution

第二代测序技术又称下一代高通量 DNA 测序技术(NGS),与基因芯片相比,具有通量高、速度快、成本低的优点.该技术的提出,使基因组学、基因表达学和表观遗传学产生了革命性的改变,为人们解决存在的生物问题和全面透彻地分析物种的基因组、转录组提供了重要工具[1].其中,基于深度测序的 RNA-seq 测序技术被广泛应用于转录组研究.RNA-seq 可以对某一特定物种的转录组和基因组序列进行全面分析,并获得特定生理或者病理状态下的所有信息,极大地提高了测序速度并且降低了成本.其主要思想是:通过序列对比,将读段(read)定位到参考基因组或者转录组上,获得量化的表达值[2,3].随着测序技术的不断发展,生物数据库中存储了海量的 RNA-seq 读段数据,通过读段计数,可量化基因表达水平.而如何处理与分析这些基因表达值,并从中获得有用的生物学意义,已经成为当前热门的研究方向.

目前,许多 RNA-seq 测序实验除了对某一特定条件下的物种进行分析,还在多条件多重复实验下进行测序,如不同的温度[4]、不同的组织[5]、不同的时间点[6]等条件.在自然界环境下,外界的条件会快速并且随机改变,所以为了适应这些变化,基因在不同条件下会表现出不同的表达水平[7].通过分析这些不同表达模式的数据,我们可以获得客观数据中所包含的生物意义.由于生物的性状是多个基因共同调控的结果,而聚类分析根据不同的表达模式,将基因聚到不同的类簇.通常假设功能相关的基因会聚到同一类簇,由此可揭示基因的未知生物功能.因此,聚类分析对于处理这些多条件多重复性的 RNA-seq 数据显得尤为重要.但是当前大部分聚类分析主要应用于基因芯片数据,应用到 RNA-seq 数据的聚类研究相对较少.本文旨在提出一种适用于多条件多重复性的 RNA-seq 数据的聚类分析框架,能够有效处理 RNA-seq 数据存在的各种噪声和偏差,获得更具生物学意义的聚类结果.

RNA-seq 数据的聚类分析研究工作虽已有一定进展,但仍存在不足.如文献[5]使用了 *K*-means 算法对 4 个不同组织下的表达模式进行了简单聚类并做后续分析,但实际上这些传统算法,如分层算法、*K*-means 算法、自组织映射(SOM)算法,都是基于启发式的算法,难以比较各自的优劣,并且没有一个特定的原则可以确定最优的类簇个数,给聚类分析增加了不少困难[8−10].一些基于模型的聚类算法应用到 RNA-seq 数据上,并获得了较好的聚类结果.这些方法中,大部分使用泊松分布对 RNA-seq 数据进行建模,以减少读段非均匀分布的影响[11−13].然而实验结果表明,RNA-seq 数据与泊松模型相比较具有更高的差异性,即数据的方差大于数据的均值,这将导致过离散现象[14].所以越来越多的模型采用负二项分布模拟读段数据,以便更好地处理 RNA-seq 数据的过离散现象.MBclust[15]采用了负二项式分布模拟读段计数进行聚类,通过估计数据的散度系数,获得了相比 *K*-means 等传统方法更好的聚类结果.

现有的这些聚类方法直接采用读段计数对基因表达水平进行聚类分析,这类方法存在一定不足.除了实验中 cDNA 文库的制备、cDNA 中 CG 碱基含量过高等因素造成了读段在参考序列上非均匀分布外,RNA-seq 读段非均匀分布另外一个主要原因是真核生物普遍存在选择性剪切(AS)现象[16],即:一个基因的多个共享外显子,每次选择不同的外显子进行组合形成多种异构体(isoform),导致由较多异构体共享的外显子上往往读段分布较多,而较少异构体共享的外显子上读段分布较少.现有的聚类方法中较少考虑到读段的这种非均匀分布特性,从而降低了聚类的准确性.此外,直接采用读段计数进行聚类分析的方法需要处理技术性、生物性等多种不确定性,如果仅采用一个模型,难以全面模拟各种不确定性.比如,MBclust 方法采用负二项分布模拟了读段计数的过离散特性并且考虑了生物重复性影响,但是将聚类中心固定为一个数据点,忽略了聚类中心的不确定性,降低了聚类结果的可靠性.

本文主要针对现有方法的不足,采用两步方法对 RNA-seq 数据进行聚类分析:首先,充分考虑读段数据固有的各种噪声和偏差,计算出基因的表达水平及其技术性不确定性;然后,在考虑到基因表达水平生物性不确定性情况下,对所获得的基因表达水平进行聚类分析.以往的研究结果表明:通过适当的统计分析处理相关噪声数据,对于获得有生物学意义的分析结果具有重要意义[17,18].通过概率模型,实验产生的技术性不确定性能够被融入模型中参与估计,这使得这些方法对于噪声更加鲁棒[19].在基因芯片数据分析中已证明:如果在表达水平的后

续分析中考虑其不确定性,能够获得更具生物意义的结果[20,21].先前的工作中,我们设计了 PGSeq[22]模型对 RNA-seq 数据进行基因表达水平计算,该模型考虑了读段数据中由各种噪声引起的偏差,选择性剪接是导致读段非均匀分布和多源映射问题的一个重要原因.相比其他聚类算法,我们在基因表达水平估计步骤中,着重考虑了这些问题对表达水平计算准确性的影响,故对造成读段计数非均匀分布的各种原因考虑较为完善.在 PGSeq 模型的基础上,本文提出了 PUseqClust(propagating uncertainty into RNA-Seq clustering)聚类框架,考虑了基因在不同条件、不同重复样本下的表达水平的相关性,采用多维拉普拉斯方法获得基因表达水平的不确定性,并将计算结果传递到混合 t 分布聚类模型[23]中,实现对 RNA-seq 数据的聚类分析.混合 t 分布聚类模型使用了具有鲁棒性的 t 分布模型,并且考虑了生物性重复实验数据的不确定性.本文在模拟数据集和 3 个真实数据集上,验证了 PUseqClust 的聚类性能.

1 方 法

图 1 显示了 PUseqClust 方法的流程图.输入数据为 RNA-seq 的原始读段序列,通常以 FASTA 或者 FASTQ 两种格式存储[24],经过多步处理,获得最终聚类分析结果.首先进行数据预处理,选择转录组序列作为参考序列,利用 Bowtie2[25]进行读段定位,获得读段计数,然后,利用 PGSeq 方法模拟基因读段的随机产生过程;其次,采用多维拉普拉斯方法[26]获得基因表达水平及其不确定性;再次,采用一元方差分析(one-way ANOVA)获得差异基因,为聚类分析进行简单过滤,为聚类分析过滤掉没有明显变化模式的无表达或稳定表达基因;最后,将基因表达水平及其不确定性传递到混合 t 分布聚类模型中,进行聚类分析,获得聚类结果.

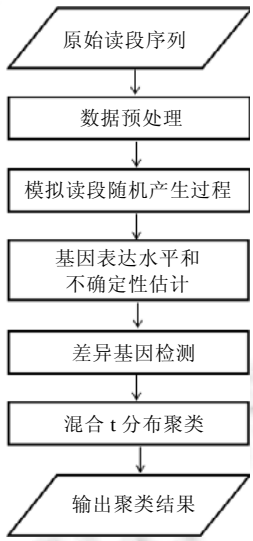


Fig.1 The flowchart of PUseqClust method

图 1 PUseqClust 方法流程图

1.1 模拟读段随机产生过程

PGSeq 方法[22]利用泊松-伽马分布模拟每个外显子读段计数的分布,获得了较为准确的基因以及异构体表达水平.我们采用 PGSeq 模型对外显子读段的随机产生过程进行了模拟,以便获得基因表达水平重要参数的随机特性.设在条件 c 的第 r 个重复样本上基因 g 的外显子 i 读段计数为 y_{gicr},由于数据都是逐个处理每个基因,因此小标 g 在后续大部分公式中被省略.y_{icr} 表示包含外显子 i 的剪接异构体的归一化读段计数之和,$y_{icr}=w_{cr}l_i\sum_k M_{ik}t_{icrk}$.其中,

- w_{cr} 表示在条件 c 的第 r 个重复样本的读段总数;

- l_i 表示外显子 i 的长度;
- M_{ik} 取值为 0 或 1,M_{ik} 取值为 1 时,表示剪接异构体 k 包含外显子 i;
- t_{icrk} 表示在条件 c 的第 r 个重复样本上剪接异构体 k 包含的外显子 i 的读段计数,并假设其服从泊松分布:t_{icrk}~$Pois(\beta_i\alpha_{crk})$,其中:$\beta_i$ 表示外显子 i 的偏差特性,为解决过离散问题,模型假设 β_i 服从 *Gamma* 分布 β_i~$Ga(a,b)$,a 表示形状参数,b 表示尺度参数;α_{crk} 表示在条件 c 的第 r 个重复样本上剪接异构体 k 的表达水平比率.

因此,基因的读段计数 $y_{icr} \sim Pois\left(w_{cr}l_i\beta_i\sum_k M_{ik}\alpha_{crk}\right)$.假设 y_{icr} 独立同分布,利用最大似然法求出模型参数 $\{\alpha_{crk}\}$,a,b 值.对数似然函数为

$$L(\{\alpha_{crk}\},a,b)=\log\prod_i\prod_{cr}p(y_{icr})=\sum_i\log\int\prod_{cr}p\left(y_{icr}\mid w_{cr}l_i\sum_k M_{ik}\alpha_{crk}\beta_i\right)p(\beta_i\mid a,b)\mathrm{d}\beta_i \tag{1}$$

1.2 基因表达水平及不确定性计算

得到模型参数的解后,在条件 c 的第 r 重复样本上剪接异构体 k 的表达水平 t_{icrk} 可以写成:

$$P(t_{icrk})=\int\mathrm{d}\beta_i P(t_{icrk}\mid\hat{\alpha}_{crk}\beta_i)P(\beta_i\mid\hat{a},\hat{b})\sim NB\left(\hat{a},\frac{\hat{\alpha}_{crk}}{\hat{\alpha}_{crk}+\hat{b}}\right) \tag{2}$$

其中,*NB* 表示负二项分布.由此推出在条件 c 的第 r 个重复样本上剪接异构体 k 表达水平的期望为

$$\langle t_{icrk}\rangle=\frac{\hat{a}}{\hat{b}}\hat{\alpha}_{crk} \tag{3}$$

对于同一个基因,由于 a,b 为各条件共享,则在不同条件下,这个基因异构体表达水平期望的随机性由 α_{crk} 的随机性决定.

考虑到各个条件重复实验下 α_{crk} 之间的相关性,我们首先采用多维拉普拉斯方法从公式(1)近似获得向量 $\alpha=\{\alpha_{crk}\}$ 的分布.对公式(1)进行泰勒展开,如下:

$$L(\alpha)\approx L(\hat{\alpha})-\frac{1}{2}(\alpha-\hat{\alpha})^T A(\alpha-\hat{\alpha})+\ldots \tag{4}$$

其中,

$$A_{ij}=-L''(\hat{\alpha})=-\frac{\partial}{\partial\alpha_i\alpha_j}L(\alpha)|_{\alpha=\hat{\alpha}} \tag{5}$$

$\hat{\alpha}$ 为最大似然函数估计值,得到多元 α 的近似高斯分布为 $N(\hat{\alpha},A^{-1})$.则对应剪接异构体 k 的表达水平 t_{crk} 为 α 的边缘正态分布 $N(u_{crk},\sigma^2_{crk})$.

假设基因的表达水平由对应的剪接异构体表达水平之和表示,即 $s_{cr}=\sum_k t_{crk}$,得到的基因表达水平服从高斯分布 $N(u_{gcr},\sigma^2_{gcr})$.均值和方差如下:

$$u_{gcr}=\frac{a}{b}\sum_k\hat{u}_{crk} \tag{6}$$

$$\sigma^2_{gcr}=\left(\frac{a}{b}\right)^2\sum_k\hat{\sigma}^2_{crk} \tag{7}$$

从基因表达水平服从的高斯分布 $N(u_{gcr},\sigma^2_{gcr})$ 中采样出 20 000 个正样本,计算得到对数刻度上的均值和方差进行后续的聚类分析.

1.3 差异基因检测

我们采用一元方差分析进行差异基因检测.方差分析(analysis of variance,简称 ANOVA)是数据分析中一种常见的统计模型,探索因变量与自变量之间关系,用于多个样本均数差别的统计假设检验.一元方差分析即为探索一个自变量对于因变量的观察值影响.零假设 H_0 为基因未发生差异表达,即对于 n 个 k 维输入样本表达均值

都相等.为了计算统计显著性,若 H_0 成立,当总偏差平方和 SST 固定不变时,检验统计量取为

$$F=\frac{SSA/(k-1)}{SSE(n-k)}\sim F(k-1,n-k) \tag{8}$$

其中,*SSE* 为组内偏差和,*SSA* 为组间偏差平方和.对于给定显著水平 *a*,查找 *F* 分布临界值 *F* 函数 F_a,得到 *p* 值:若 *p* 值大于 *a*,则接受零假设;否则,拒绝零假设.

对于采样得到的基因表达水平,我们对于每个基因在各个条件的重复样本下分别提取 100 个正样本数据,对这些数据进行一元方差分析,设置显著水平阈值,识别显著差异表达的基因.

1.4 聚类分析

对于显著差异基因,我们采用先前提出的基于不确定性的混合 t 分布聚类模型 PUMA-CLUSTII[23]进行聚类分析.PUMA-CLUSII 采用具有鲁棒性的混合学生 t 分布进行聚类,由于这是一种重尾分布,能更好地适应异常值,并且模型考虑了生物性重复不确定性及技术性不确定性,并能自动优化到最优类簇个数.

假设在条件 *c* 的第 *r* 个重复样本上基因 *g* 的表达水平 $\hat{x}_{gcr}$ 服从高斯分布 $N(u_{gcr},\sigma^2_{gcr})$,其中,$x_{gcr}$ 表示真实基因表达水平,σ^2_{gcr} 表示基因 *g* 技术性不确定性.真实基因表达水平 x_{gcr} 在同一条件上假设服从高斯分布 $N\left(w_{gc},\frac{1}{\eta_g}\right)$,$w_{gc}$ 表示条件 *c* 下基因均值,η_g 表示基因表达水平生物上的不确定性.对于每一个基因引入一个隐藏变量 u_g,则 t 分布可以写成高斯分布和 Gamma 分布的卷积形式:

$$St(w_g\mid\mu_k,\Sigma_k,v_k)=\int_0^\infty N\left(w_g\mid\mu_k,\frac{\Sigma_k}{u_g}\right)Ga\left(u_g\mid\frac{v_k}{2},\frac{v_k}{2}\right)\mathrm{d}u \tag{9}$$

其中,在第 *k* 类簇上,μ_k 和 Σ_k 分别表示均值和协方差,v_k 表示自由度.则均值 w_g 的概率为

$$P(w_g)=\sum_{k=1}^{K}\pi_k St(w_g\mid\mu_k,\Sigma_k,v_k) \tag{10}$$

假设对于每个基因所有条件下 η_g 共享,且服从 Gamma 分布:

$$\eta_g|z_{gk}=1\sim Ga(\alpha_k,\beta_k) \tag{11}$$

此时,模型隐藏变量 $h_g=(x_g,w_g,\eta_g,z_g,u_g)$,模型参数 $\theta=(\{\mu_k\},\{\Sigma_k\},\{v_k\},\{\alpha_k\},\{\beta_k\},\{\pi_k\})$.

PUMA-CLUSTII 方法利用最大似然法和变分 EM 算法对模型进行求解,该方法采用了 MMLP 准则[27]自动确定最优聚类数.

2 数据集

由于无法获得真实数据集中的准确基因表达模式的聚类结果,我们分别使用了模拟数据集和 3 个真实数据集对所提出的聚类方法进行验证.

2.1 模拟数据集

本文研究的聚类分析方法根据基因在多个不同实验条件下的表达模式进行聚类,由于真实的 RNA-seq 数据集缺少大规模已知聚类结果的基因表达水平数据,故模拟数据模拟生成已知的特定基因表达模式,作为真实聚类标签.我们的聚类方法对具有相似表达模式的基因进行聚类,不针对现实中哪一种具体的生物学波动模式,模拟数据生成方法见文献[28]以及我们先前的工作[20,23].本文在模拟数据中人为定义了 6 种不同的用数学函数描述的表达模式,且增加了一种噪声数据来验证算法的鲁棒性.本文利用 PGSeq 模型生成模拟数据集[22],对我们所提出的聚类方法进行初步评价.模拟数据集共包含 7 个条件,每个条件下包含 3 个重复实验数据,将各个条件下的基因差异表达水平看做已知类簇的基因表达模式,这样的模拟数据集可以验证聚类算法的准确性.

本文选取人类的 700 个基因作为模拟数据,并分成 7 组,其中,最后一组作为纯随机噪声组,用于模拟实际数据中难以聚到任一类簇中的基因表达模式.对于给定一个基因,外显子 *i* 的偏差特性 β_i 的分布可从 MAQC 数据集[29]中的 HBR 样本实验数据中获得,即得到 *Gamma* 分布参数 *a* 和参数 *b* 的值.对于给定基因 *g*,模拟数据产生

过程如下:

(1) 对于外显子 i,从伽马分布 $Gamma(a,b)$中采样得到β_i;

(2) 变量 A 的值从均匀分布 $U(0,5)$中采样到.设 C 为总条件个数,p 为基因 g 所属的数据组,然后按照如下方法获得 $\log\alpha_{gck}$ 的值.

- 对于第 1 组~第 4 组:

$$\log\alpha_{gck}=A\sin(2\pi\times c/C-\pi\times p/2) \tag{12}$$

- 对于第 5 组:

$$\log\alpha_{gck}=A(c\times 2/C-1) \tag{13}$$

- 对于第 6 组:

$$\log\alpha_{gck}=A(-(c-1)\times 2/C+1) \tag{14}$$

- 对于第 7 组:

$$\log\alpha_{gck}=A \tag{15}$$

由此生成具有不同模式的 $\log\alpha_{gck}$,然后从高斯分布 $N(\log\alpha_{gck},\log(\alpha_{gck})/10)$重新采样获得含有噪声的 α^*_{gck} .在每个条件下,从高斯分布 $N(\alpha^*_{gck},\alpha^*_{gck}/20)$ 重复采样 3 次,获得带生物性噪声的重复α_{gcrk};

(3) 从泊松分布 $Pois(w_{cr}l_iM_{ik}\beta_i\alpha_{gcrk})$中采样获得剪接异构体 k 在条件 c 的第 r 重复样本上的外显子 i 的读段计数 x_{icrk}.其中,w_{cr} 为在条件 c 的第 r 重复样本上的读段总数;l_i 为外显子 i 的长度;M_{ik} 为外显子 i 与剪接异构体 k 之间的关系,值为 1 表示剪接异构体 k 包含外显子 i;

(4) 外显子 i 在条件 c 的第 r 重复样本上读段计数 y_{icr} 为包含外显子 i 的所有剪接异构体读段计数 x_{icrk} 之和,基因 g 在条件 c 的第 r 重复样本上读段计数 z_{gcr} 为所有外显子的读段计数 y_{icr} 之和.最终生成的对数基因读段计数归一化后数据如图 2 所示.由图 2 可以看出:模拟数据中前 6 组基因分别具有不同的表达模式,而第 7 组基因没有一致的表达模式,其代表了数据中存在的随机噪声,用以测试聚类方法的鲁棒性.

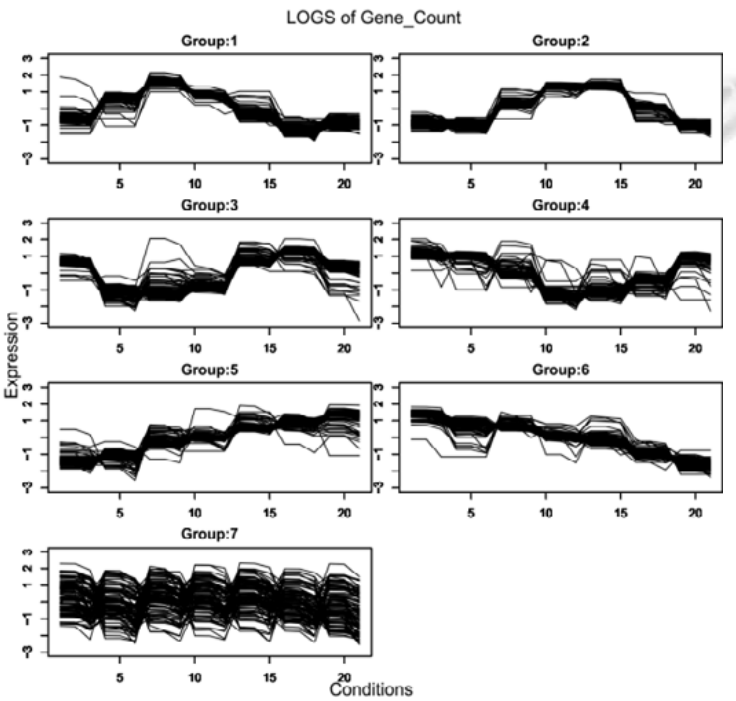


Fig.2 The gene expression patterns of the seven clusters in the simulated data

图 2 模拟数据的 7 类基因表达模式

2.2 真实数据集

(1) Zhao 等人[30]采用人类激活 T 细胞数据集对 RNA-seq 和基因芯片在基因表达水平计算方面进行了对比研究.利用 Illumina HiSeq 2000 platform 得到在 0hr,2hr,4hr,6hr,24hr,72hr 这 6 个时间点上的 T 细胞 RNA 数据,其中每个时间点上包含两个重复性实验.本文采用一元方差分析检测出该数据集的 1 963 个差异基因,并对这些基因进行聚类分析,验证我们所提出的聚类算法的性能;

(2) Äijö 等人[31]对人类辅助 T 17(Th17)细胞进行表达水平测量,利用 Illumina HiSeq 2000 platform 得到了 0.5hr,1hr,2hr,4hr,6hr,12hr,24hr,48hr,72hr 这 9 个时间点上 Th17 细胞 RNA 数据,其中,每个时间点上包含 3 个重复性实验.与前一个数据集处理相同,采用一元方差分析获得了 2 060 个差异基因进行后续聚类分析;

(3) GEO accession 为 GSE90053 的数据集为人类多功能干细胞(pluripotent stem cell,简称 PSC)数据集,利用 Illumina HiSeq platform 2500 得到了 0d,2d,8d,10d,11d,14d,21d,28d 这 8 个时间点上 PSC RNA 数据集,每个时间点上包含 3 个重复性实验.最终得到 1 448 个差异基因进行后续聚类分析.

3 实验结果与讨论

为了验证本文提出的方法在聚类分析方面的性能,我们使用 PUseqClust 方法和其他聚类分析方法在模拟数据集和真实数据集进行聚类分析,并对比了分析结果.本文采用了两种已有的聚类方法进行对比实验.一种是没有考虑表达不确定性的标准高斯混合聚类方法 Mclust[32];另一种是基于负二项分布的聚类方法 MBclust[15].

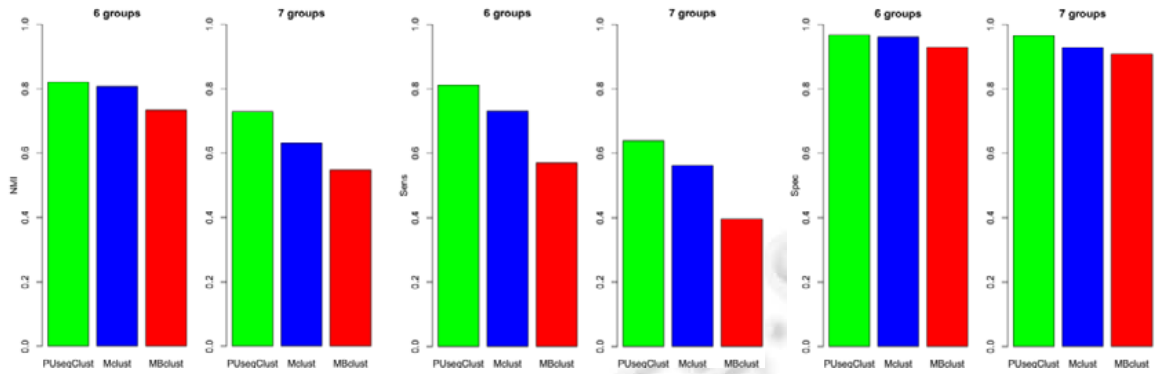
这两种方法均采用 RPKM 计算基因表达水平.Mclust 和 MBclust 方法分别包含在 R 软件包 Mclust 和 MBCluster.Seq 中.

3.1 模拟数据集实验

对于模拟数据集,我们首先在无随机噪声的前 6 组数据上进行聚类分析,然后加入第 7 组噪声数据,对聚类方法的鲁棒性进行对比.在模拟数据集上,由于各个基因所属类簇已知,所以本文在标准互信息 NMI(normalized mutual information)、敏感度和特异度这 3 个方面验证各种聚类方法的性能.

灵敏度代表所有成对基因(属于同一类簇的基因)被划分到同一类簇的的概率,灵敏度越高,代表算法对来自同一类簇基因的识别度越高.特异度代表所有非成对基因(不属于同一类簇的基因)被划分为不同类簇的概率,特异度越高,代表算法对非同一类簇基因的识别度越高[15].互信息是信息论的一种信息度量,评估一个随机变量中所包含的另外的一种随机变量的信息量,或者认为是一个随机量由于已知另一个随机变量而减少的不确定性.这里我们将其作为量化真实聚类划分和已知聚类划分的共有信息量的程度.我们利用文献[33]中的方法计算互信息的值,并且将互信息归一化至 0~1 之间的标准互信息.NMI 值越接近 1,表示真实聚类划分和已知聚类划分的相关性越高,即聚类精度越高;反之,越接近零聚类精度越低.敏感度表示来自同一类簇的成对基因被分到同一类簇的概率,特异度表示来自不同类簇的成对基因被分到不同类簇的概率.这两个值越高,表示聚类性能越好.

图 3 显示了模拟数据集中无噪声组和包含噪声组两种情况下的 NMI 值、敏感度和特异度.当无噪声组时,对于 NMI 值和敏感度,PUseqClust 方法优于 Mclust 方法,并且 PUseqClust 和 Mclust 都获得了比 MBclust 更准确的聚类结果;当含有噪声组时,各种方法的性能都有所下降,但 PUseqClust 仍然获得了较为准确的聚类结果.对于特异度,图中显示:PUseqClust 无论在含有噪声组或者不含有噪声组里,都比其余两种方法性能优越.实验结果证明:PUseqClust 方法采用两步方法进行聚类分析,考虑了技术上和生物上表达水平的不确定性,对噪声的处理能力更强,因而获得了较准确的聚类结果.



(a) NMI 值 (b) 敏感度 (c) 特异度

Fig.3 The NMI, sensitivity, specificity of different methods on the simulated data

图 3 模拟数据集上各种聚类方法对应 NMI 值、敏感度、特异度

3.2 真实数据集实验

对于真实数据集,因真实的类簇个数和聚类划分未知,PUseqClust 方法采用了 MMLP 准则[27]自动确定最优类簇个数,MClust 使用贝叶斯信息准则 BIC(Bayesian information criterion)值[34]计算最优类簇个数,而 MBclust 方法的实现软件没有提供最优类簇个数的确定方法,因此,我们只考察 PUseqClust 和 MClust 获得的最优类簇个数.本文利用 DAVID[35]对各种方法获得的聚类结果进行 GO(go ontology)注释富集分析[36].GO 注释富集分析通过寻找基因所属的特定 GO 富集功能目录直接评估聚类结果的生物意义,其分析并找出在统计上显著富集的 GO 功能目录,通过将差异基因做 GO 富集分析,可以把基因按照不同的功能进行分类,达到对基因进行注释和分类的目的.对于某个类簇的基因,GO 功能目录越多,意味着这个类簇内的基因在生物意义上相关程度越高.聚类分析结果应该尽可能增加相关程度相对较高的类簇数量,减少相关程度相对较低的类簇数量.富集计算结果在 DAVID 上是以改进的 Fisher exact test 评估,采用 p 值表示基因聚类结果的生物意义.GO 生物过程水平 5 (biological process level 5)设置阈值计数水平为 5、p 值为 0.05 时,含有 GO 富集功能目录的类簇被认为是 GO 富集类簇,因此,我们设定阈值相关计数水平为 5、p 值为 0.05,然后获得所有的 GO 富集功能目录个数,以便评价整个聚类方法的性能.

在 T 细胞数据集的分析中,PUseqClust 和 Mclust 方法获得最优类簇个数分别为 17,18.Th17 细胞数据集 PUseqClust 和 Mclust 方法获得最优聚类个数分别为 15,8.PSC 数据集中分别为 20,21.

图 4~图 6 分别显示了 T 细胞数据集、Th17 细胞数据集和 PSC 数据集上各种方法在 GO 富集功能目录指定数量范围内的类簇个数,得到聚类结果在各个层次的类簇数量.

从图 4~图 6 中可以看出:相关程度相对较低的层次,即 GO 富集功能目录个数小于 5 的范围,在最优类簇聚类时,T 细胞数据集、Th17 细胞数据集和 PSC 数据集中,PUseqClust 方法中类簇数量为(4,2,4),Mclust 和 MBclust 中类簇数量分别为(6,5,8)和(6,2,8);在非最优类簇聚类时,T 细胞数据集和 Th17 细胞数据集中,PUseqClust 方法中类簇数量为(2,0,4),Mclust 和 MBclust 中类簇数量分别为(2,1,8)和(4,1,7).由此可知,无论在最优类簇聚类还是非最优类簇聚类时,PUseqClust 方法都拥有最少的相关程度相对较低的类簇个数.

从图 4~图 6 同样可以看出:在高层次水平上,即 GO 富集功能目录个数大于 50 的范围内,PUseqClust 方法也能获得不少于 Mclust 和 MBclust 的类簇个数.所以我们认为:PUseqClust 方法与 Mclust 和 MBclust 相比较,聚类的性能更具竞争优势.

图 7 分别显示了 3 个真实数据集各个方法所有类簇 GO 富集功能目录个数的总和,用以检测不同聚类算法获得的聚类结果的生物学意义.从图中可以明显看出:PUseqClust 方法获得了比 Mclust 和 MBclust 显著多的 GO 富集功能目录,因而获得了最具生物学意义的聚类结果.

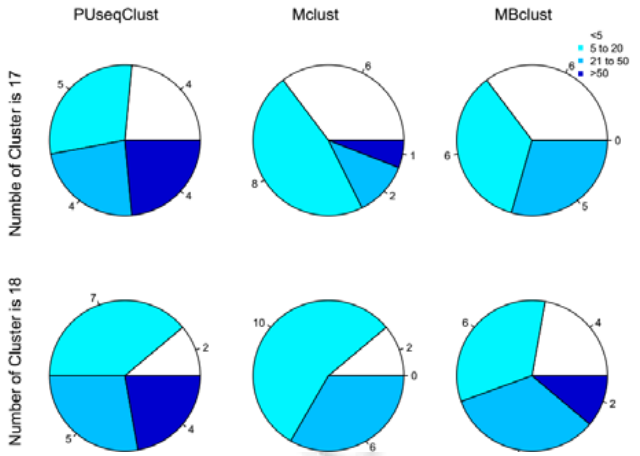


Fig.4 The cluster numbers under the specified catalogue numbers of GO enrichment function on the T cell data

图 4 T 细胞数据集各种方法在指定 GO 富集功能目录数量范围内的类簇个数

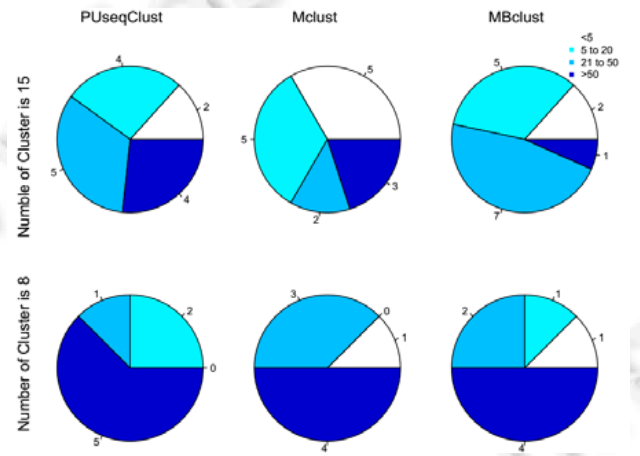


Fig.5 The cluster numbers under the specified catalogue numbers of GO enrichment function on the Th17 cell data

图 5 Th17 细胞数据集各种方法在指定 GO 富集功能目录数量范围内的类簇个数

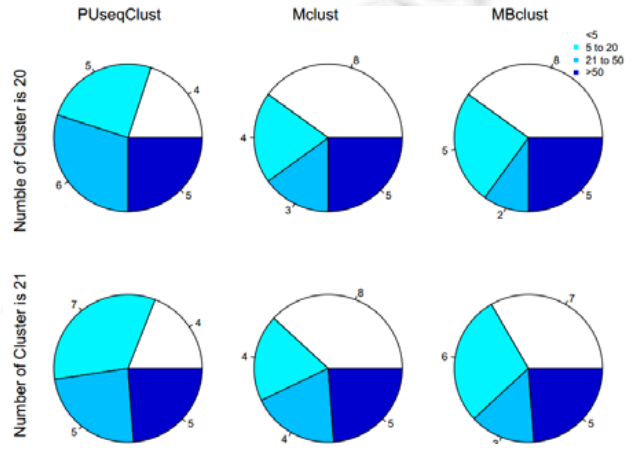
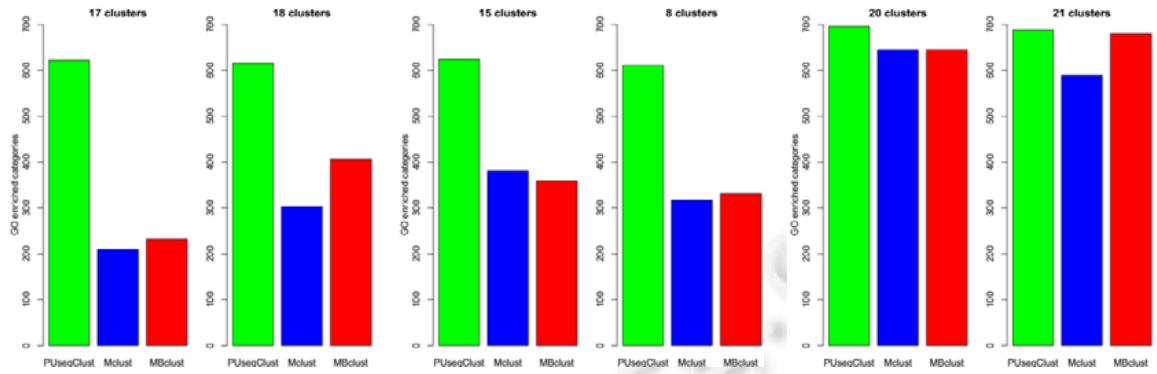


Fig.6 The cluster numbers under the specified catalogue numbers of GO enrichment function on the PSC data

图 6 PSC 数据集各种方法在指定 GO 富集功能目录数量范围内的类簇个数



(a) T 细胞数据集 (b) Th17 细胞数据集 (c) PSC 数据集

Fig.7 The total catalogue numbers of GO enrichment function

图 7 GO 富集功能目录个数总和

4 总 结

本文针对 RNA-Seq 读段数据,在 PGSeq 模型的基础上,提出了 PUseqClust 聚类框架.该聚类方法利用 PGSeq 降低读段非均匀分布的影响,利用拉普拉斯反复考虑基因表达水平在不同条件下的不同重复样本之间的相关性得到基因的对数表达水平及其不确定性,增加了模型的健壮性,并将不确定性传递到混合 t 分布聚类模型进行聚类分析,加强了对噪声的鲁棒性.

本文采用模拟数据集和真实数据集验证所提方法的性能,并与同样基于模型的聚类方法 Mclust 和 MBclust 方法进行了对比.实验结果表明:本文方法在模拟数据集上显示了更为准确的聚类性能,在真实数据集上获得了更具生物意义的聚类结果.

本文的工作表明,由于 RNA-seq 读段数据本身具有多种技术性噪声,而且 RNA-Seq 实验大都采用了生物性重复实验,导致数据中存在一定的生物噪声,这些对后续的数据处理形成了严峻挑战;同时,聚类分析本身也要考虑聚类过程中聚类中心的不确定度.因此,仅采用一个模型往往不能很好地对多种噪声和不确定性进行模拟.我们通过采用多步分析,充分考虑各个步骤中的各种噪声和不确定性,获得了较好的分析结果.

作者注 本文是我们于 2017 年 1 月 3 日投到《软件学报》的论文,该文是南京航空航天大学刘学军老师指导的 2018 年 4 月毕业研究生石险峰(本文第一作者)的硕士学位论文《RNA-Seq 数据差异表达及聚类分析研究》工作成果的一部分.特此说明.

References:

[1] Metzker ML. Sequencing technologies—The next generation. Nature Reviews Genetics, 2010,11(1):1−13.

[2] Marguerat S, Wilhelm BT, Bähler J. Next-Generation sequencing: applications beyond genomes. Biochemical Society Transactions, 2008,36(Pt 5):1091−1096.

[3] Marioni JC, Mason CE, Mane SM, *et al*. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research, 2008,18(9):1509−1517.

[4] Yang J, Chen X, Zhu C, *et al*. Using RNA-seq to profile gene expression of spikelet development in response to temperature and nitrogen during meiosis in rice (Oryza sativa L.). Plos One, 2015,10(12):386−398.

[5] Li P, Ponnala L, Gandotra N, *et al*. The developmental dynamics of the maize leaf transcriptome. Nature Genetics, 2010,42(12): 1060−1067.

[6] Sanavia T, Finotello F, Camillo BD. FunPat: Function-based pattern analysis on RNA-seq time series data. Bmc Genomics, 2015, 16(Suppl .):1−13.

[7] Sultan M, Schulz MH, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008,321(5891):956−960.

[8] Eisen MB, Spellman PT, Brown PO, *et al*. Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc. of the National Academy of Sciences of the United States of America, 1998,95(25):14863−14868.

[9] Tavazoie S. Systematic determination of genetic network architecture. Nature Genetics, 1999,22(3):281−285.

[10] Tamayo P, Slonim D, Mesirov J, *et al*. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. of the National Academy of Sciences of the United States of America, 1999,96(6): 2907−2912.

[11] Wang N, Wang Y, Han H, *et al*. A bi-Poisson model for clustering gene expression profiles by RNA-seq. Briefings in Bioinformatics, 2013,15(4):534−541.

[12] Ye M, Wang Z, Wang Y, *et al*. A multi-Poisson dynamic mixture model to cluster developmental patterns of gene expression by RNA-seq. Briefings in Bioinformatics, 2015,16(2):205−215.

[13] Witten DM. Witten DM. Classification and clustering of sequencing data using a Poisson model. Annals of Applied Statistics, 2012, 5(4):2493−2518.

[14] Di Y, Schafer DW, Cumbie JS, *et al*. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Statistical Applications in Genetics & Molecular Biology, 2011,10(1):24−24.

[15] Si Y, Liu P, Li P, *et al*. Model-Based clustering for RNA-seq data. Bioinformatics, 2014,30(2):197−205.

[16] Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: Towards a cellular code. Nature Reviews Molecular Cell Biology, 2005,6(6):386−398.

[17] Slonim DK. From patterns to pathways: Gene expression data analysis comes of age. Nature Genetics, 2002,32(Suppl):502−508.

[18] Quackenbush J. Computational genetics: Computational analysis of microarray data. Nature Reviews Genetics, 2001,2(6):418−427.

[19] Lin KK, Chudova D, Hatfield GW, *et al*. Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. Proc. of the National Academy of Sciences, 2004,101(45):15955−15960.

[20] Liu X, Lin KK, Andersen B, *et al*. Including probe-level uncertainty in model-based gene expression clustering. BMC Bioinformatics, 2007,8(1):1−19.

[21] Liu X, Milo M, Lawrence ND, *et al*. Probe-Level measurement error improves accuracy in detecting differential gene expression. Bioinformatics, 2006,22(17):2107−2113.

[22] Liu X, Zhang L, Chen S. Modeling exon-specific bias distribution improves the analysis of RNA-Seq data. Plos One, 2015,10(10): 386−398.

[23] Liu X, Rattray M. Including probe-level measurement error in robust mixture clustering of replicated microarray gene expression. Statistical Applications in Genetics & Molecular Biology, 2010,9(9):1−25.

[24] Deorowicz S, Grabowski S. Compression of DNA sequence reads in FASTQ format. Bioinformatics, 2011,27(6):860−862.

[25] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods, 2012,9(4):357−359.

[26] MacKay, Davidj C. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

[27] Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2002,24(3):381−396.

[28] Yeung KY, Medvedovic M, Bumgarner RE. Clustering gene-expression data with repeated measurements. Genome Biology, 2003, 4(5):R34.

[29] MAQC Consortium. The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 2010,28(8):827−838.

[30] Zhao S, Fungleung WP, Bittner A, *et al*. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. Plos One, 2014,9(1):e78644.

[31] Äijö T, Butty V, Chen Z, *et al*. Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. Bioinformatics, 2014,30(12):113−120.

[32] Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. Journal of Classification, 1999,16(2):297−306.

[33] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2002,3(3):583−617.

[34] Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978,6(2):15−18.

[35] Dennis G, Sherman BT, Hosack DA, *et al*. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology, 2003. P3−P3.

[36] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 2009,4(1):44.

石险峰(1992－),男,安徽安庆人,硕士,主要研究领域为生物信息学.

刘学军(1976－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为机器学习,生物信息学.

张礼(1985－),男,博士,讲师,CCF 专业会员,主要研究领域为机器学习,生物信息学.