论文

基于概率话题模型的微生物菌群结构研究

王侠林^{1†}, 左赞^{2†}, 周兰平³, 朱林^{2,3*}, 范红^{2*}, 孔祥阳^{1*}, 贺建峰^{1*}

1. 昆明理工大学信息工程与自动化学院, 昆明 650500;

2. 昆明理工大学附属医院, 云南省第一人民医院消化科, 昆明 650532;

3. 昆明市延安医院, 昆明医科大学附属延安医院急诊科, 昆明 650051

† 同等贡献

* 联系人, E-mail: jfenghe@qq.com; zhulinkin@foxmail.com; kmfanhong@163.com; kxy2772@qq.com

收稿日期: 2017-07-11; 接受日期: 2017-08-20; 网络版发表日期: 2017-11-03

国家自然科学基金(批准号: 81260077, 81560107, 11265007)资助

摘要 微生物菌群结构的异质性在影响宿主健康与疾病等方面有着十分重要的作用. 对于菌群结构的时间与空间尺度异质性研究主要有非监督学习算法以及监督学习算法. 由于菌群数据特性与文本数据特性之间的相似性, 本文采用非监督学习的LDA概率话题模型对菌群结构的时间异质性进行研究, 并与系统聚类和K-Means聚类这两种方法进行比较. 采用LDA模型折叠Gibbs抽样的蒙特卡洛算法对两种数据源北平顶猴(*Macaca leonina*)阴道菌群(MVB)和轻微型肝性脑病(MHE)菌群的时间异质性OTUs数据集进行了分析. 用LDA模型分别将MVB和MHE数据源中的27个样本和77个样本的OTUs数据集分为6个Topic和4个Topic. 这与系统聚类和K-Means聚类划分成的簇数目(分别为5, 3与4, 3)有所不同. 此外, 实验表明结合MVB样本间生理数据-pH和MHE中样本 α 多样性, pH和 α 值的分类相似性更能与LDA模型的样本分类特性保持一致. 因此, LDA在样本的聚集程度上更能精确地对OTUs数据集进行分类. 更为重要的是, LDA模型还可以鉴定出每个Topic中具有代表性的OTUs. 与系统聚类和K-Means聚类方法相比较, LDA模型不仅能更为有效地量化菌群结构的异质性, 还能鉴定出相对应影响异质性的OTUs.

关键词 LDA模型, Gibbs抽样, 蒙特卡洛算法, 系统聚类, K-Means聚类

宿主的微生态系统是由宿主与微生物菌群长期共同进化而形成的, 并对宿主的发育、生理、生殖、免疫、代谢和进化具有十分重要的作用^[1,2]. 通过对微生物菌群结构进行比较分析, 可为挖掘微生物菌群功能信息、优化和调节微生物菌群结构提供可靠的依据. 微生物菌群的研究是近年来国际研究的热点, 对营养物质代谢、人体自身发育、免疫及疾病的产生

等方面都起到极其重要的作用^[3,4].

在菌群数据研究中, 菌群数据的特性主要表现为序列数据量大, 高数据通量, OTU (operational taxonomic unit, 即操作分类单元, 通过一定的距离度量方法计算两两不同序列之间的距离度量或相似性, 继而设置特定的分类阈值, 获得同一阈值下的距离矩阵, 进行聚类操作, 形成不同的分类单元)数量多和分布稀疏等

引用格式: 王侠林, 左赞, 周兰平, 等. 基于概率话题模型的微生物菌群结构研究. 中国科学: 生命科学, 2017, 47: 1220–1234
Wang X L, Zuo Z, Zhou L P, et al. A study of microbial flora structure based on probability topic model (in Chinese). Sci Sin Vitae, 2017, in press, doi: 10.1360/N052017-00105

特性^[5].

目前,在微生物菌群结构的比较研究中常用的方法主要分为两大类,即非监督学习算法(unsupervised learning)和监督学习算法(supervised learning)^[6,7].对于监督学习,它是对具有概念标记的训练例进行学习,以尽可能正确地对训练集之外的示例概念标记进行预测.而非监督学习,它是对没有概念标记的训练例进行学习,以发现训练例中潜在隐藏的结构性知识.由于本文采用的数据集无标签,并且是挖掘潜在的数据信息,因此采用非监督学习算法.其典型的基于原型的目标函数聚类方法如K-Means聚类法^[8,9],具有快速、简单和可伸缩性等优点,但需要事先根据样本空间分布指定分类的数目,对异常值敏感.还有系统聚类法(hierarchical clustering methods)^[10]采用贪婪算法,根据类与类之间的距离计算得到其结果.但对于大样本计算量大,复杂度高,且异常值也能产生很大影响,最终结果只是局域最优,不一定是全局最优.因此,对于这两种方法,比较适合处理样本量小,数据类型不复杂,有明确目标函数的数据集.近年一些学者也提出了一些聚类算法,如邻居信息、最大子图、亲密度等方法^[11~13]可以归为密度和网格的聚类算法.其中,密度的聚类算法在于不受某个特定目标函数的限制,能够有效发现传统聚类无法发现的非凸面形状簇,可以用来处理任意形状的簇,也可以用来过滤“噪声”中的孤立点^[14,15];而网格的聚类算法主要是对空间数据进行处理,操作方法主要是将空间数据离散化,具有处理速度快,处理时间独立于数据对象数,且仅依赖于量化空间中的每一维的单元数^[16,17].本文使用的数据来源于北平顶猴(*Macaca leonina*)阴道菌群(MVB)和轻微型肝性脑病(MHE)菌群时间序列数据,具有一定的稀疏性.而密度和网格的聚类算法(邻居信息、最大子图和亲密度等方法)进行菌群的分析没有涉及时间序列特性^[18,19].因此,本文拟采用LDA(latent Dirichlet allocation)概率话题模型^[20,21]对菌群数据进行分析.LDA概率话题模型是基于主题算法的一种新的非监督学习算法.它是对所有的指标数据信息进行权重评定,对于处理时间序列数据,控制稀疏性模型,边界点检测^[22]等具有明显的优越性,其所用数据相对完整,且不需要事先指定分类.

LDA模型已逐步在生物医学研究领域中得到运用.Holmes等人^[23]利用LDA模型应用于宏基因组学的

研究.LDA概率主题模型能够有效地提高宏基因组读取的合并性能^[24],以及通过概率话题模型引入多项式混合建模(Dirichlet multinomial mixture, DMM)对肥胖和瘦弱双胞胎姐妹的肠道菌群进行分析,以此证明肥胖可能与微生物群落结构相关^[25].另外,也可以通过LDA模型对miRNAs和mRNAs进行功能验证^[23].根据文献查询目前还没有检索到运用LDA概率话题模型对菌群数据进行分析的文献.

由于菌群数据的特性与文本分析的变化模式相类似,本文利用LDA概率主题模型对时间动态变化的两种微生物群落MVB和MHE进行结构分析,并与系统聚类法和K-Means聚类这两种方法进行比较^[26,27],结果发现,LDA模型不仅能更为有效地量化菌群结构的异质性,还能鉴定出影响菌群结构异质性的OTUs.

1 材料与方法

1.1 LDA——概率话题模型

LDA模型——也称为“潜在狄利克雷分布模型”是2003年提出的基于概率模型的主题模型算法^[28,29].它涉及贝叶斯理论、Dirichlet分布、多项分布、图模型、变分推断、EM算法、Gibbs抽样等知识,是一种基于概率模型的聚类算法.

如图1所示,在S个样本中有T个OTUs.随机变量 $m_{s,t}$ 代表一个OTU在第s个样本中的索引,K是产生的K个主题(topic),其中 $s \in \{1, \dots, S\}$, $t \in \{1, \dots, T_s\}$. T_s 代表OTUs在第s个样本产生的总次数.随机变量 $z_{s,t}$ 代表第s个样本中隐藏的OTUs, $z_{s,t}$ 和 $m_{s,t}$ 与主题参数 $\theta_{s,k}$ 和 $\phi_{k,n}$ 分别都具有多项式分布 ϕ_k 和 ϕ_d .

$$\begin{aligned}\phi_k &\sim \text{Dirichlet}_T(\alpha), k = 1 \dots K \\ \phi_d &\sim \text{Dirichlet}_k(\beta), s = 1 \dots S.\end{aligned}\quad (1)$$

每个参数相应的超参数 α 和 β 有一个Dirichlet先验分布.参数 $\theta = \{\theta_{s,k}\}$ 用于模块与样本关联, $\Phi = \{\phi_{k,n}\}$ 分配OTU在模块 $Z = \{z_{s,t}\}$ 中表达的概率.因此OTU可以通过估计参数 θ 和 Φ 来识别.本课题组的模型与LDA模型紧密相关,一个OTU和一个样本分别对应LDA模型中的一个单词(words)和一个文档(documents).在主题模型中,主题(topic)表示一个概念、一个方面,表现为一组相关的单词,是这些单词的条件概率. $\theta_{s,k}$ 和

图1 阴影和非阴影的变量分别代表已知和隐藏变量

$\phi_{k,n}$ 分别与第 s ($1 < s < S$) 个样本和第 k ($1 < k < K$) 个 OTUs 呈多项式分布关系, α 和 β 是它们的超参数. 每个 OTU 变量 $m_{s,t}$ 由 k 个隐藏模块产生并由 $z_{s,t}$ 选择. T_s 代表 OTUs 在第 s 个样本中总共表达的次数.

由于 LDA 模型最重要的是对其参数的估计, 因此本文采用近似估计算法来对模型参数进行估计. 最大期望(expectation-maximization, EM)算法^[30]通过寻找参数后验概率分布来估计隐含变量的参数, 它是一个估计参数的标准方法. 但是, 当应用到概率话题模型的时候, 这种方法容易受到局部最大和计算效率低的影响. 因此, 本文采用一种折叠 Gibbs 抽样法^[31]来估计这些参数, 该方法利用蒙特卡洛算法对参数的后验概率进行抽样. 折叠 Gibbs 抽样法通过对隐含变量 $z_{s,t}$ 和 $m_{s,t}$ 联合抽样, 对每个状态样本中的 $\theta_{s,k}$ 和 $\phi_{k,n}$ 边缘化. 对于一个 OTU 变量 t 的第 s 个样本, $z_{s,t}$ 和 $m_{s,t}$ 的抽样可以表示为一个条件概率:

$$P(z_{s,t} = k \mid m_{s,t} = n, Z_{-(s,t)}, s) \propto \frac{C_{kn}^{KT} + \beta}{\sum_{n'=1}^T C_{kn'}^{KT} + T\beta} \times \frac{C_{sk}^{SK} + \alpha}{\sum_{s'=1}^S C_{s'k}^{SK} + S\alpha}, \quad (2)$$

其中, $Z_{-(s,t)}$ 是除了第 s 个条件类型外的 OTUs 目前的分配状态. C_{sk}^{SK} 是排除现在状态外, 样本 s 分配给第 k 个 OTUs 的次数. C_{kn}^{KT} 是排除现在状态外, 表达类型 n 分配给第 k 个 OTUs 的次数. k 是 OTUs 的隐含变量个数, s 是条件类型的总个数.

根据贝叶斯公式, 得出 α 和 β 的后验分布 $p(w|\alpha, \beta)$:

$$p(w|\alpha, \beta) = \int p(\theta_{s,k}|\alpha) \left(\prod_{n=1}^N \sum_{Z_n} p(Z_n|\theta_{s,k}) p(w_n|Z_n, \beta) \right) d\theta_{s,k} \\ = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^k \theta_{s,k}^{\alpha_i-1} \right) \\ \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_{s,k}^i \in \beta_{ij})^{w_n^j} \right) d\theta_{s,k}. \quad (3)$$

经过足够的抽样迭代达到稳定收敛时, 基于抽样产生的模块分配值来预测参数 $\theta_{s,k}$ 和 $\phi_{k,n}$:

$$\hat{\theta}_{s,k} = \frac{C_{sk}^{SK} + \alpha}{\sum_{s'=1}^S C_{s'k}^{SK} + S\alpha}, \quad \hat{\phi}_{k,n} = \frac{C_{kn}^{KT} + \beta}{\sum_{n'=1}^T C_{kn'}^{KT} + T\beta}, \quad (4)$$

其中 C_{sk}^{SK} 和 C_{kn}^{KT} 来自除了现在状态外所有数据分配结果算出来的, 它与以上的抽样过程不同.

1.2 3种方法分析流程

见图2流程, 本文利用 R 软件包^[32]采用 LDA 概率话题模型、系统聚类 and K-Means 聚类 3 种方法对两种数据源的 OTUs 数据集进行了分析比较. 进行数据的准备阶段, 对菌群数据集进行中心标准化处理. LDA 模型只需丰度大于 0, 因为 LDA 模型可以自动进行预处理, 而系统聚类和 K-Means 聚类则需要进行中心标准化预处理. 综合利用 LDA、系统聚类和 K-Means 聚类 3 种不同方法分别对 MVB 和 MHE 两套数据源的 OTUs 数据集进行分析. 详细步骤为: (i) 利用 R 软件包中的 Cell-tree^[32]构建 LDA 模型. 通过 Compute.Backbone.tree 的 mst.tree 生成最小生成树(minimum spanning tree, MST). 利用 Compute.Backbone.tree 的 b.tree 和 b.tree2 构建骨干树(Backbone.tree-b.tree)和(Backbone.tree-b.tree2), 以此展现 LDA 模型对样本间相似性的有效区分以及每个 Topic 中样本的聚集程度情况. 最后对 LDA 模型结果进行分析. (ii) 利用 R 软件包中的 Cluster 进行系统聚类分析, 在进行聚类分析时, 需选择结果比较好的距离测度和聚类方法. 距离测度分别选择了欧式距离测度、平方欧氏距离测度和余弦距离测度; 聚类方法分别选择了组之间的链接(between-groups linkage)、组内的链接(within-groups linkage)、中位数聚类和 Ward 等不同方法. 经过实验分析结果对比发现, OTUs 数据集在

图2 方法流程示意图(网络版彩图)

基于平方欧氏距离测度的Ward方法下, 分类结果最好. (iii) 利用R软件包中的K-Means函数进行K-Means聚类分析, 首先对K值进行确定, 主要采用轮廓系数法, 通过枚举法将 k 从2到一个固定值如10进行数次重复运行K-Means (避免局部最优解), 并计算当前 k 的平均轮廓系数, 最后选取轮廓系数最大的 k 值作为最终的集群数目. 最后, 对3种方法分析结果进行比较.

1.3 实验数据源

本文采用两套数据源的OTUs数据集进行分析计算. 第一套MVB菌群OTUs数据集, 包含两个成年雌性北平顶猴个体(PMA与PMB)的阴道菌群以及pH的时间序列数据, 共27个样本. 其中, PMA与PMB分别含有13和14个时间点数据^[33]. 第二套MHE菌群OTUs数

数据集, 来源于云南省第一人民医院消化科所采集的77个样本, 总共分为3组C, MHE和N. C (cirrhotics without MHE) 没有MHE的肝硬化, 含有25个样本; M或MHE (minimal hepatic encephalopathy) 轻微型肝性脑病, 含有26个样本; N (normal family members matched to MHE patients) 与MHE患者匹配的正常家族成员, 含有26个样本^[34], 以及元数据 α 多样性.

2 分析结果

本文采用图2所示的3种方法对MVB和MHE两种数据源的OTUs数据集进行分析. 实验结果包括2个部分: (i) LDA模型对两种数据源MVB和MHE的OTUs数据集的分析结果; (ii) 系统聚类与K-Means聚类方法对MVB和MHE的OTUs数据集的分析结果.

2.1 LDA模型对两种数据源MVB和MHE的分析结果

(1) LDA模型对MVB的分析. LDA模型将MVB的OTUs数据集的27个样本分成了6个Topic (图3), 它反映OTUs在不同Topic中占的权重值, 权值越大, 颜色越红; 反之, 越暗; 每个Topic用不同的颜色表示; 与之相对应的是每个Topic中的样本以及样本所对应的重

要OTUs.

图3展现的是不同Topic对应样本的情况, 和与之相对应的对样本影响最大的OTUs, 颜色越深, 影响越大, 反之, 越小. 其中, OTU 295 (*Selenomonas* sp.), OTU 246 (*Prevotella* sp.), OTU 637 (*Mobiluncus* sp.), OTU 77 (*Peptoniphilus* sp.)对Topic 1中的样本PMA08和PMB11影响最大; OTU 1424 (*Atopobium vaginae* sp.), OTU 609

图3 不同Topic中的样本和影响样本最大的OTUs热图
反映OTUs在不同Topic中占的权重值, 权值越大, 颜色越红; 反之, 越暗

(*Sneathia* sp.), OTU 134 (*Clostridiales* sp.)等对 Topic 2 的样本 PMA01, PMA02, PMA04, PMA13, PMB01, PMB03, PMB08 和 PMB09 影响最大; OTU 223 (*Porphyromonas* sp.), OTU 2964 (*Bacteroides* sp.)菌对 Topic 3 的样本 PMA03, PMA14 和 PMA16 影响最大; OTU 214 (*Porphyromonas* sp.), OTU 346 (*Fusobacterium* sp.), OTU 383 (*Flavobacteriaceae* sp.)等菌对 Topic 4 的样本 PMB04, PMB05 和 PMB12 等影响最大; OTU 606 (*Fusobacterium* sp.), OTU 16 (*Peptoniphilus* sp.), OTU 230 (*Porphyromonas* sp.)等菌对 Topic 5 的样本 PMA05, PMB06 和 PMB16 影响最大; OTU 2 (*Lactobacillus johnsonii* sp.), OTU 118 (*Streptococcus* sp.), OTU 156 (*Finnegoldia* sp.)菌对 Topic 6 的样本 PMA09, PMA10, PMA11, PMA12 和 PMB02 影响最大.

其中个体 PMA 主要集中在 Topic 1, Topic 2, Topic 3 和 Topic 6, 个体 PMB 则在 Topic 1, Topic 2, Topic 4 和 Topic 5, PMA 和 PMB 都共有 Topic 1, Topic 2, 且 LDA 模型将 MVB 的 27 个样本分成了 6 个 Topic. 因此, 上述结果表明, PMA 和 PMB 之间存在相互效应, 且 LDA 模型能够区分样本间的相似性. 更为重要的是, 从中发现 LDA 模型还能提取相应影响样本最大的 OTUs 菌(图 3 右半部分). 这些的代表性 OTUs 可能是北平顶猴核心的 OTUs.

得到 LDA 模型后, 构建 LDA 主题模型 MST 和 Backbone.tree. 如图 4, 反映 LDA 模型对 MVB OTUs 数据集的分类结果以及分类聚集程度.

图 4A 可以看出, MST 将 MVB 数据集分为 6 个 Topic, 且每个 Topic 用 6 种不同的颜色表示, 以此证明 LDA 能够区分样本间的相似性. 通过 Compute.Backbone.tree 主干树的 b.tree 生成骨干树, 反映这 6 个 Topic 的聚集程度, 从中可以看出, 出现了两相轨迹(图 4B). 为使聚集效果更好, 在此基础上添加能够动态地调整距离阈值的尺度因子(width.scale.factor). 最终发现, 图 4C 中骨干树由分支结构变成了线性结构, 证明了尺度因子的变化, 会导致骨干树形状的变化. 图 4C 表明, 距离阈值尺度因子在一定范围内值越高, 骨干树的聚集效果会更好. 也就是说, 图 4C 的聚集程度比图 4B 的聚集程度效果好.

(2) LDA 模型对 MHE 的分析. LDA 模型将数据源 MHE 中的 3 组 OTUs 数据集 C, MHE 和 N 的 77 个菌群样本总共分成 4 个 Topic (图 5). MHE 的样本主要集中在 Topic

1, N 的样本主要集中在 Topic 2, C 的样本几乎均匀分布在 Topic 1 和 Topic 2. 从中发现, 有两个家庭在 MHE 中, OTU 151 (*Eubacteriaceae* sp.)和 OTU 301 (*Sutterellaceae* sp.)具有损失. 而当($P=0.003$)时 MHE 组的菌 Clostridiales 明显高于 N 组和 C 组. 但当($P\leq 0.001$)时, MHE 组显著增加的有 3 种菌属 *Odoribacter*, *Flavonifractor*, *Coprobaecillus* 和损失的有 4 种菌属真杆菌, *Lachnospira*, *Parasutterella* 和 *Erysipelotrichaceae* 相比于 N 组. 从而发现, MHE 和 C 组与 MHE 和 N 组之间存在显著性的差异. 以此说明, LDA 模型能够有效地区分样本间的相似性. 更为重要的是, LDA 模型还能鉴别出影响样本最大的 OTUs, 颜色越深, 影响越大, 反之, 越小. 图 5 中 OTUs 菌就是影响每个样本中最大的 OTUs. 其中, OTU 6 (*Vulgatus* sp.), OTU 8 (*Fragi* sp.), OTU 16 (*Aerofaciens* sp.), OTU 20 (*Hydrogenotrophica* sp.)等菌对 Topic 1 的样本 C001, C005, M002, M003, M004, N002 等影响最大. OTU 9 (*Dispar* sp.), OTU 17 (*Breve*.), OTU 31 (*Anginosus* sp.), OTU 33 (*Prausnitzii* sp.)等菌对 Topic 2 的样本 N001, N007, N008, M001, M009 等影响最大. OTU 23 (*Fragilis* sp.), OTU 50 (*Nellenica* sp.), OTU 74 (*Pulmonis* sp.), OTU 95 (*Bicirculans* sp.), OTU 144 (*Gnavus* sp.)等菌对 Topic 3 的样本 C012, C013, C016, N020, N021, M005, M008 影响最大. OTU 26 (*Comes ATCC* sp.), OTU 45 (*Parainfluenzae* sp.), OTU 110 (*Piriformis* sp.)等菌对 Topic 4 的样本 N003, N011, N025, C028, M014 影响最大. 这些代表性 OTUs 可能是引起人体患轻微型肝性脑病以及肝硬化的核心 OTUs, 且 LDA 模型能够区分样本间的相似性.

同数据集 MVB 分析一样, 得到 LDA 模型后, 构建 LDA 主题模型的 MST 和骨干树(图 6). 可以看出, MST 将 MHE 数据集分为 4 个 Topic, 且每个 Topic 用 4 种不同的颜色表示, 以此证明 LDA 能够区分样本间的相似性, 且还能显示采样时间 Group 的变化过程(图 6A). 通过 Compute.Backbone.tree 主干树的 b.tree 生成骨干树, 反映这 4 个 Topic 的聚集程度, 从中可以看出, 出现了两相轨迹(图 6B). 为使聚集效果更好, 在此基础上添加能够动态地调整距离阈值的尺度因子(width.scale.factor). 最终发现, 图 6C 中骨干树由分支结构变成了线性结构, 证明了尺度因子的变化, 会导致骨干树形状的变化. 图 6C 表明, 距离阈值尺度因子在一定范围内值越高, 骨干树的聚集效果会更好. 也就是说, 图 6C 的聚集程

图 4 OTUs的Topic和Group MST和Backbone.tree

A: MST, 左边图根据单元节点的主题分布着色, 右边图以采样时间(以天为单位)着色. 每个点代表OTU单元, Topic中点的颜色表示6个不同的主题, Group中表示不同的采样时间; B: Backbone.tree-b.tree, 分别显示Topic和Group骨干树的两相轨迹, 且具有OTU单元小分支, 沿着树轨迹的排序遵循期望的时间序列. C: Backbone.tree-b.tree2, 较大的节点表示骨干OTU, 而较小的节点表示仅与它们连接的骨干略微变化的OTU. OTU节点根据C分布主题和采样时间着色

度比图5B的聚集程度效果好.

另外, 从表1可以看出, MVB和MHE两套OTUs数

据集的模型参数 P -value值都远远小于0.01, 说明样本的分类极为显著, 分类结果合理. 从自由度(degrees of

图 5 不同Topic中的样本和影响样本最大的OTUs热图
反映OTUs在不同Topic中占的权重值, 权值越大, 颜色越红; 反之, 越暗

freedom, df)可以看出, Topic 1~Topic 6呈递减趋势, 自由度越大方差越大, 而方差表示的是数据的离散程度, 以此说明, Topic 6的样本最集中且聚集效果比较好。而离差平方和(dispersion)反映这种制约因素引起的差异大小, 差异越大, 表明越不稳定, 反之, 越稳定。表1说明MVB和MHE菌群中的Topic 1很不稳定。BF表示样本的对数值。

2.2 系统聚类和K-Means聚类对MVB和MHE的分析结果

(1) 系统聚类对MVB和MHE的分析。图7所示, 系

统聚类对MVB的OTUs数据集进行分析, 计算27个样本的距离矩阵, 采用基于Ward的方法创建样本聚类树和冰柱图。图7A系统聚类树状图, 不同的颜色分别表示不同的簇类; 图7B样本冰柱图; 横轴表示样本被聚类的对象, 纵轴簇数表示被聚成几类。

如图7, 系统聚类将27个样本分为5类簇(Cluster)。5种不同的颜色表示5个不同的类簇(图7A): 第一簇类, PMA01~02, PMA04, PMA11~13, PMB01~03, PMB08~09, PMB16; 第二簇类, PMA05, PMB04~05, PMB12~15; 第三簇类, PMA03, PMA14, PMA16, PMB06; 第四簇类, PMA09~10; 第五簇类, PMA08,

图6 OTUs的Topic和Group MST和Backbone.tree

A: MST; B: Backbone.tree-b.tree; C: Backbone.tree-b.tree2

PMB11. 结果表明, 系统聚类能够区分MVB OTUs数据集样本间的相似性. 图7B冰柱图, 通过高度的不同进行划分. 首先, 从最高冰柱进行划分, 它将所有观测对象分为该柱的左右两类, 以此再寻找稍矮的下一条冰柱就可以将最高冰柱所分2类中的1类分为2类, 因此共3类, 后面依次类推. 由于冰柱图与树状图一一对应. 因此, 通过高度的变化将其冰柱图划分为5类簇最合理.

针对MHE的OTUs数据集采用相同的方法和步

骤, 由于得出的结果图与图7相类似, 所以就不再呈现. 系统聚类将MHE OTUs数据集的77个样本总共分为3类簇(Cluster). 第一簇类, C001, C005, C010~13, C016, C021, C026~27, C029, H003, H018, H020~23, M004~05, M010~11, M016, M018, M020, M023, M025; 第二簇类, C002, C004, C007~09, C011, C017~20, C023~25, H001, H002, H004~07, H009~10, H012~14, H016~17, H019, H024, H027, H029, M002~03, M006~09, M012~13, M017, M021~22, M024, M029; 第三簇类, C003, C028,

图 7 系统聚类图(网络版彩图)

A: 系统聚类树状图(dendrogram using Ward Linkage); B: 样本冰柱图(number of clusters); 横轴: 表示样本被聚类的对象, 纵轴: 簇数表示被聚成几类

表 1 LDA模型对MVB和MHE数据集分析的相关参数^{a)}

数据源	Topic ID	BF	dispersion	P-value	df
数据源1: MVB 菌群	Topic 1	1969.868	0.9435815	9.89×10^{-7}	7506
	Topic 2	2947.596	0.6882322	8.90×10^{-6}	7319
	Topic 3	3966.936	0.6222541	7.80×10^{-7}	6686
	Topic 4	5061.663	0.5565405	6.50×10^{-8}	6325
	Topic 5	6178.564	0.5038149	4.80×10^{-7}	6074
	Topic 6	5601.776	0.4899395	3.70×10^{-6}	5558
数据源2: MHE 菌群	Topic 1	1032.963524	6.89255361	1.52×10^{-178}	15230
	Topic 2	2097.390383	6.461883559	1.45×10^{-163}	14971
	Topic 3	7317.924405	5.700615746	1.37×10^{-156}	14668
	Topic 4	9273.892022	5.511625506	9.3683×10^{-102}	14361

a) BF: 样本的值; P-value: 即P值, 判定假设检验结果的一个参数

表 2 K-Means聚类的分析结果

Cluster	样本							
1	PMA09	PMA10	PMB03	PMB08	PMB09			
2	PMA02	PMA05	PMB04	PMB05	PMB12	PMB13	PMB14	PMB15
3	PMA03	PMA08	PMA14	PMA16	PMB06	PMB11	PMB16	
4	PMA01	PMA04	PMA11	PMA12	PMA13	PMB01	PMB02	

H008, H011, H025, M001, M014, M019, M027. 以此说明, 系统聚类能够区分MHE OTUs数据集样本间的相似性.

(2) K-Means 聚类对 MVB 和 MHE 的分析结果. K-Means聚类将MVB的OTUs数据集中的27个样本分成4簇类(Cluster)(图8).

从图8可以看出, Cluster 1含有5个样本, Cluster 2含有8个样本, Cluster 3含有7个样本, Cluster 4含有7个样本. 另外, 从轮廓图上看, Cluster 1和Cluster 4的聚类结果比较好, Cluster 2和Cluster 3大部分样本表现不佳, 有相当部分的点落在0.8以下.

表2可以看出, K-Means 聚类将样本分为4个Cluster类. Cluster 1包含的样本有PMA09~10, PMB03和PMB08~09; Cluster 2包含的样本有PMA02, PMA05, PMB04~05和PMB12~15; Cluster 3包含的样本有PMA03, PMA08, PMA14, PMA16, PMB06, PMB11和PMB16; Cluster 4包含的样本有PMA01, PMA04, PMA11~13和PMB01~02. 以此说明, K-Means聚类能

够区分MHE OTUs数据集样本间的相似性. 其中, k 值的选取主要采用轮廓系数法, 通过枚举将 k 从2到一个固定值如10进行数次重复运行K-Means (避免局部最优解), 并计算当前 k 的平均轮廓系数, 最后选取轮廓系数最大的 k 值作为最终的集群数目. 因此, 通过计算, 本文最终选取的 k 值为4.

针对MHE的OTUs数据集采用相同的方法和步骤, 由于得出的结果图与图8相类似, 所以就不再呈现. K-Means聚类将MHE OTUs数据集的77个样本总共分为3类簇(Cluster). 第一簇类, C001, C005, C010~12, C026~27, C029, H003, H018, H020~23, M004~05, M010~11, M016, M018, M020, M023, M025, M006~09; 第二簇类, C013, C016, C021, C002, C004, C007~09, C011, C023~25, H001~02, H004~07, H009~10, H012, H024, H027, H029, M002~03, M012~13, M017, M021~22, M024, M029; 第三簇类, H013~14, H016~17, H019, C017~20, C003, C028, H008, H011, H025, M001, M014, M019, M027. 以此说明, K-Means聚类能够区分

图 8 K-Means聚类结果

横轴表示: 轮廓值; 纵轴表示: 分类(Cluster)

MHE OTUs数据集样本间的相似性.

3 讨论

本文利用LDA模型对微生物菌群结构的时间异质性进行量化比较分析. 研究结果表明, LDA模型分别将MVB和MHE数据源中的27个样本和77个样本的OTUs数据集分为6个Topic和4个Topic. 并通过构建MST和骨干树分别展现了LDA模型对样本间相似性的有效区分以及每个Topic中样本的聚集程度情况. 还能鉴定出影响菌群结构异质性的OTUs. 此外, 通过整合LDA模型分析结果与MVB菌群的生理数据-pH和MHE菌群的 α 多样性. 本课题组发现, 结合MVB中样本间生理数据-pH和MHE中样本 α 多样性, 都为6个和4个Cluster. 而系统聚类、K-Means聚类与pH的相似性存在较大差异. 从表3可以看出, LDA模型的阈值输入达到最大. 这是因为LDA模型在Celltree包中自动进行预处理, 不需提前处理, 只需每列的丰度不全等于0, 而K-Means和系统聚类需要进行中心标准化预处理, 从而导致原始数据有一定的丢失. 从准确度上看, LDA模型的准确度达到96.8%和95.8%, 而K-Means和系统聚类最高才达到94.20%和93.80%, 以此说明, LDA模型相比于上述两种方法准确度更精确. 从分类

数目上看, LDA模型将MVB和MHE菌群分别分为6个和4个Topic, 而K-Means和系统聚类分别为4, 3和5, 3个Cluster. 如果单一的从分类数目上看, 并不能说明LDA模型能更精确地对OTUs数据集进行分类, 但是, 根据样本生理数据-pH和 α 多样性的相似度可以进行说明. 从表3可知, LDA模型的分类与MVB菌群生理数据的pH相似度达到86.95%, 而K-Means和系统聚类只为49.85%和66.76%; 同理, 与MHE菌群的 α 多样性相似度达到91.64%, 相比于K-Means聚类和系统聚类也出现显著性的差异. 这说明, LDA模型能更精确地对OTUs数据集进行分类,

即LDA模型能更好地区分样本间的相似性. 从变化率上看, LDA模型、系统聚类和K-Means聚类3种方法变化率方向基本保持一致, 但LDA模型计算结果变化趋于平缓, 而系统聚类、K-Means聚类所得结果变化波动较大. 更为重要的是, LDA模型还能找出对每个样本影响最大的OTUs菌, 例如, OTU 295 (*Selenomonas* sp.), OTU 246 (*Prevotella* sp.), OTU 6 (*Vulgatus* sp.), OTU 8 (*Fragi* sp.)等菌, 而系统聚类和K-Means聚类不能, 只能对其样本进行分类. 因此, 综合上述情况, 对微生物菌群结构的时间异质性进行量化比较分析, LDA模型优于系统聚类和K-Means聚类方法.

表 3 LDA模型、K-Means和系统聚类的结果比较

数据源	方法	LDA	K-means	系统聚类
数据源1: MVB 菌群		MVB	MVB	MVB
	阈值输入大小	9872	7857	8716
	准确度	96.80%	93.30%	92.60%
	Topic (数目)	6	4	5
	与生理数据pH的相似度	86.95%	49.85%	66.76%
	是否能找出重要的 OTU 菌群	是	否	否
	预处理	否	是	是
数据源2: MHE 菌群	变化率	趋于平稳	波动较大	波动较大
		MHE	MHE	MHE
	阈值输入大小	24622	16563	18382
	准确度	95.80%	94.20%	93.80%
	Topic(数目)	4	3	3
	与 α 多样性相似度	91.64%	78.35%	86.92%
	是否能找出重要的 OTU 菌群	是	否	否
	预处理	否	是	是
	变化率	趋于平稳	波动较大	波动较大

参考文献

1 Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 2012, 486: 207–214

2 Methé B A, Nelson K E, Pop M, et al. A framework for human microbiome research. *Nature*, 2012, 486: 215–221

3 Costello E K, Lauber C L, Hamady M, et al. Bacterial community variation in human body habitats across space and time. *Science*, 2009, 326: 1694–1697

4 Nasidze I, Li J, Quinque D, et al. Global diversity in the human salivary microbiome. *Genome Res*, 2009, 19: 636–643

5 Davey J W, Hohenlohe P A, Etter P D, et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, 2011, 12: 499–510

6 Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn*, 2001, 42: 177–196

7 Fritzke B. Growing cell structures—A self-organizing network for unsupervised and supervised learning. *Neur Networks*, 1994, 7: 1441–1460

8 Elmoufidi A, El Fahssi K, Jai-Andaloussi S, et al. Automatically density based breast segmentation for mammograms by using dynamic K-Means algorithm and Seed Based Region Growing. *Instrumentation and Measurement Technology Conference. IEEE, Pisa, Italy*, 2015. 533–538

9 Kowsari K. Comparison three methods of clustering: K-Means, spectral clustering and hierarchical clustering. *Comput Sci*, 2013, arxiv: 1312.6117 cs.LG.

10 Carlsson G, Mémoli F, Ribeiro A, et al. Hierarchical Clustering methods and algorithms for asymmetric networks. *Asilomar Conference on Signals, Systems and Computers, Asilomar Hotel & Conference Grounds*, 2013, 118: 1773–1777

11 Eustace J, Wang X, Cui Y. Overlapping community detection using neighborhood ratio matrix. *Phys A Statist Mech Appl*, 2015, 421: 510–521

12 Eustace J, Wang X, Cui Y. Community detection using local neighborhood in complex networks. *Phys A Statist Mech Appl*, 2015, 436: 665–677

13 Cui Y, Wang X, Eustace J. Detecting community structure via the maximal sub-graphs and belonging degrees in complex networks. *Phys A Statist Mech Appl*, 2014, 416: 198–207

14 Cui Y, Wang X, Li J. Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient. *Phys A Statist Mech Appl*, 2014, 405: 85–91

- 15 Wang X, Li J. Detecting communities by the core-vertex and intimate degree in complex networks. *Phys A Statist Mech Appl*, 2013, 392: 2555–2563
- 16 Li J, Wang X, Cui Y. Uncovering the overlapping community structure of complex networks by maximal cliques. *Phys A Statist Mech Appl*, 2014, 415: 398–406
- 17 Li J, Wang X, Eustace J. Detecting overlapping communities by seed community in weighted complex networks. *Phys A Statist Mech Appl*, 2013, 392: 6125–6134
- 18 Eustace J, Wang X, Li J. Approximating web communities using subspace decomposition. *Knowledge-Based Syst*, 2014, 70: 118–127
- 19 Cui Y, Wang X. Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks. *Phys A Statist Mech Appl*, 2014, 407: 7–14
- 20 Xu Z, Raschid L. Probabilistic Financial Community Models with Latent Dirichlet Allocation for Financial Supply Chains. *International Workshop on Data Science for Macro-Modeling*, 2016. 1–6
- 21 Xu Z, Raschid L. Probabilistic Financial Community Models with Latent Dirichlet Allocation for Financial Supply Chains. *International Workshop on Data Science for Macro-Modeling ACM*, 2016. 8
- 22 胡秀丽. 基于VSM和LDA模型相结合的微博话题漂移检测. *兰州理工大学学报*, 2015, 41: 104–109
- 23 Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE*, 2012, 7: e30126
- 24 Jeon H B, Lee S Y. Language model adaptation based on topic probability of latent Dirichlet allocation. *ETRI J*, 2016, in press, doi: 10.4218/etrij.16.0115.0499
- 25 Zhang R, Cheng Z, Guan J, et al. Exploiting Topic modeling to boost metagenomic reads binning. *BMC Bioinformatics*, 2015, 16: 1–10
- 26 Traore M, Duviella E, Lecoeuche S. Comparison of two prognosis methods based on Neuro Fuzzy Inference System and Clustering Neural Network. *IFAC Proc Volumes*, 2015, 42: 1468–1473
- 27 Brusco M J, Shireman E, Steinley D. A comparison of latent class, K-means, and K-median methods for clustering dichotomous data. *Psychol Methods*, 2017, 22: 563–580
- 28 Moro S, Cortez P, Rita P. Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Exp Syst Appl*, 2015, 42: 1314–1324
- 29 Vila J P, Schniter P. Expectation-maximization gaussian-mixture approximate message passing. *IEEE Trans Signal Process*, 2012, 61: 4658–4672 [1207.3107](#)
- 30 Lawrence C E, Altschul S F, Boguski M S, et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 1993, 262: 208–214
- 31 Finkel, Rose J, Grenager, et al. Incorporating non-local information into information extraction systems by Gibbs sampling. *Meeting on Association for Computational Linguistics*, 2005. 363–370
- 32 duVerle D A, Yotsukura S, Nomura S, et al. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinf*, 2016, 17: 363
- 33 Zhu L, Lei A H, Zheng H Y, et al. Longitudinal analysis reveals characteristically high proportions of bacterial vaginosis-associated bacteria and temporal variability of vaginal microbiota in northern pig-tailed macaques (*Macaca leonina*). *Zool Res*, 2015, 36: 285–298
- 34 Zhang Z, Zhai H, Geng J, et al. Large-scale survey of gut microbiota associated with MHE via 16S rRNA-based pyrosequencing. *Am J Gastroenterol*, 2013, 108: 1601–1611

Microbial flora structure based on probability topic model

WANG XiaLin¹, ZUO Zan², ZHOU LanPing³, ZHU Lin^{2,3}, FAN Hong², KONG XiangYang¹ & HE JianFeng¹

1 College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China;

2 Department of Gastroenterology, Affiliated Hospital of Kunming University of Science

and Technology, The First People's Hospital of Yunnan Province, Kunming 650532, China;

3 Emergency Department, Yan'an Hospital of Kunming City and Yan'an Hospital Affiliated to Kunming Medical University, Kunming 650051, China;

The heterogeneity of microbial flora structures plays an important role in the health and disease of the host. With respect to the temporal and spatial heterogeneity of the flora structure, both unsupervised and supervised learning algorithms have been developed. Because of the similarity of the characteristics of the flora data and the text data, in this paper, we investigate the temporal heterogeneity of the flora structure by applying the latent Dirichlet allocation (LDA) probability topic model for unsupervised learning. We then use system and *K*-Means clustering to compare these two methods. Two kinds of data sources of Beipingding monkey vaginal flora (MVB) and minimal hepatic encephalopathy (MHE) bacteria heterogeneity operational taxonomic unit (OTUs) data sets are analyzed by the Monte Carlo LDA model with the folding Gibbs sampling. We used the LDA model to divide the 27 samples and 77 sample OTUs in the MVB and MHE data sources, respectively, into six topics and four topics, which differ from the number of clusters (5, 3, and 4, 3) divided by system and *K*-Means clustering. In addition, experimental results show that the classification similarity of sample diversity, pH value with the physiological data-pH in MVB samples and the similarity of α value in MHE, the classification similarity of the pH and the α is consistent with the classification characteristics of LDA model. As such, the LDA model classifies the OTUs data sets more accurately with respect to the degree of aggregation of the samples. More importantly, the LDA model can also identify representative OTUs in each topic. Compared with the system clustering and *K*-Means clustering methods, the LDA model can not only quantify the heterogeneity of the flora structure more effectively, but also identify the corresponding heterogeneity of the OTUs.

LDA model, Gibbs sampling, Monte Carlo algorithm, system clustering, *K*-Means clustering

doi: [10.1360/N052017-00105](https://doi.org/10.1360/N052017-00105)