微生物学通报
Microbiology China
tongbao@im.ac.cn
http://journals.im.ac.cn/wswxtbcn

Dec. 20, 2020, 47(12): 4328−4337
DOI: 10.13344/j.microbiol.china.200098

生物实验室

肥胖人群肠道菌群特征分析及机器学习模型

吴桐　王鸿超*　陆文伟　赵建新　张灏　陈卫

江南大学食品学院　江苏 无锡　214122

摘　要:【背景】肠道菌群与人体健康之间的关系吸引了越来越多的关注，成为目前热门的研究热点。【目的】基于美国肠道计划公开数据库，对肥胖和健康人群肠道菌群进行比较分析，解析肥胖人群肠道菌群特征，并基于肠道菌群建立机器学习模型预测人群肥胖的状态，为基于肠道菌群干预肥胖提供理论基础。【方法】从公开数据库中获取美国肠道计划中的肠道菌数据，经过筛选得到 1 655 个健康(18.5<BMI<25)和 898 个肥胖(BMI>30)成年人的肠道菌群数据。针对 α 多样性，进行了 Wilcox 秩和检验分析并通过 Logsitic 回归判定 α 多样性与肥胖之间的关系；对 Unweighted UniFrac、Weighted UniFrac 和 Bray-Curtis 三种 β 多样性距离进行主成分分析(principal component analysis，PCA)，探索肥胖与健康人群在肠道菌群组成上的差异；对于物种差异，进行 Wilcox 秩和检验探索差异菌属；通过 PICRUSt 分析预测可能的代谢通路，同时与肠道菌群进行相关性分析。利用 Scikit-Learn 软件包基于属水平的肠道菌群数据建立肥胖分类机器学习模型，并进行网络搜索确定最佳模型参数。【结果】经过 Wilcox 秩和检验，发现肥胖人群的 α 多样性都较健康人群显著下降，Logistic 回归表明 α 多样性与人体肥胖状态有相关性。经过基于 Weighted UniFrac、Unweighted UniFrac 和 Bray-Curtis 三种距离的 PCA，肥胖和健康人群的肠道菌群结构上无明显差异；在门水平上，肥胖人群中的 *Firmicutes* 和 *Bacteroidetes* 比值较低，在属水平上共发现 57 个在两组之间具有显著性差异的属，其中肥胖人群中的 *Ruminococcus* 相对丰度较高，而 *Prevotella*、*Akkermansia* 和 *Methanobacteriales* 的相对丰度较低；PICRUSt 预测的代谢通路有 63 个代谢通路在两组之间具有显著差异；梯度提升回归树对于基于肠道菌群预测肥胖人群效果最好，受试曲线下与坐标轴围成的面积(area under curve，AUC)值可以达到 0.769，测试集精度可以达到 0.725。【结论】基于大规模的肠道菌群数据揭示了肥胖人群肠道菌群的特征，将机器学习运用到肥胖预测上面，为精准膳食、精准医疗提供新的研究思路和理论基础。

关键词: 肠道菌群，肥胖，代谢通路，机器学习

Foundation item: National Key Research and Development Program of China (2019YFF0217601)
***Corresponding author:** Tel: 86-510-85197096; E-mail: hcwang@jiangnan.edu.cn
Received: 11-02-2020; **Accepted:** 09-04-2020; **Published online:** 26-05-2020
基金项目：国家重点研发计划(2019YFF0217601)
*通信作者：Tel：0510-85197096; E-mail：hcwang@jiangnan.edu.cn
收稿日期：2020-02-11；接受日期：2020-04-09；网络首发日期：2020-05-26

Characteristics of gut microbiota of obese people and machine learning model

WU Tong WANG Hong-Chao* LU Wen-Wei ZHAO Jian-Xin ZHANG Hao CHEN Wei

School of Food Science and Technology, Jianngnan University, Wuxi, Jiangsu 214122, China

Abstract: [Background] The relationship between gut microbiota and human health has attracted much attention and became a popular research area. **[Objective]** To explore the feature of gut microbiota of obese people based on the American Gut Project. To provide a theoretical basis for the intervention of obesity based on gut microbiota by constructing machine learning models to predict the status of people obesity. **[Methods]** Total of 1 665 normal samples (18.5<BMI<25) and 898 (BMI>30) obese samples were downloaded from the website of the American Gut Project (AGP). The Wilcox rank-sum analysis was performed to explore the alteration of alpha-diversity between the obese and normal group. In addition, the logistic regression was performed to explore the correlation between alpha-diversity of gut microbiota and obese. For beta-analysis, we performed the principal component analysis (PCA) to explore the difference in the structure of gut microbiota between obese and normal groups. For the phylogenetic profiles, we performed the Wilcox rank-sum analysis to detect any significantly different taxa between the two groups. The PICRUSt analysis was used to predict the pathway based on the 16S rRNA gene sequences. Then, the Wilcox rank-sum analysis was used to detect the significantly different pathway between the two groups. To find the correlation between these significantly different pathways and genus, we performed the correlation analysis. Finally, we used the Scikit-Learn packages in python to construct the machine learning model and used the AUC value as the standard to justify the performance of each model. **[Results]** The decreasing trend of alpha-diversity in the obese population compared to the healthy population was observed after the Wilcox rank-sum analysis. In addition, the correlation between the alpha-diversity and the statues of obese was confirmed using the logistics regression. As for the beta-diversity, we did not observe the significant difference of the structure of gut microbiota after PCA based on three beta-diversity distance matrix including Weighted UniFrac, Unweighted UniFrac and Bray-Curtis. For the phylum, the high relative abundance of *Bacteroidetes* and the low relative abundance of *Firmicutes* was observed in the obese group. Besides, a total of 57 genera was significantly different between the two groups after the Wilcox rank-sum analysis. The genus of *Ruminococcus* increased in the obese groups, but the genus of *Prevotella*, *Akkermansia* and *Methanobacteriales* decreased in the obese group. All the pathway which predicted by the PICRUSt analysis were performed the Wilcox-rank-sum analysis between two groups and a total of 63 significantly different pathways was observed. The gradient boosted regression tree (GBDT) had the best performance with the AUC value (0.769) and test precise (0.725) among other models. **[Conclusion]** This study revealed the feature of gut microbiota of obese population based on a large-scale data sets. Besides, this study also constructed the machine learning models based on gut microbiota to predict the status of obese, which provide the new idea and theory basis of personalized medicine and diet.

Keywords: Gut microbiota, Obese, Metabolic pathway, Machine learning

随着肥胖的发生率越来越高,肥胖不再仅仅是个人健康问题,同时也是一个严峻的社会问题。有研究估计，全球的超重人群[身体质量指数(body mass index，BMI)在 25.0−30.0 之间]有 10 亿多人，肥胖人群(BMI>30.0)有 3 亿多人[1]，而这一数字随着人们生活的水平提高将会越来越高。尤其在发展中国家，肥胖和超重人群的增长更为迅速[2]。在中国，有 26.9% (25.7−28.1)男性和 31.1% (29.7−32.5)

女性的体重超重[3]。肥胖不仅对人们的生活造成诸多不便之处，同时也会增加罹患其他疾病的风险，如心脑血管疾病、Ⅱ型糖尿病、冠心病等[3]。

肠道菌群对于人体的健康具有不可忽视的作用，近年来吸引了越来越多的研究者关注。一般来讲，正常人体肠道内微生物的数量大概在10^{11}−10^{12}之间[4-5]。人出生时就开始从周围环境中获得肠道微生物[6-7]，之后这些微生物便定殖在人体肠道中，与人体机能的正常运转息息相关。肠道中微生物的数量是人体细胞总和的 10 倍左右[5]。近年来，国内外许多研究已经表明肠道菌群与宿主的消化、营养、代谢和免疫等方面之间存在一定联系，肠道菌群的紊乱与很多疾病之间存在关联。比如，有大量的研究表明肠道菌群的紊乱与肠易激综合征[8-9]、炎症性肠炎[10-11]、结肠癌[12-13]、肥胖[14]和Ⅱ型糖尿病[15]之间具有一定的相关性。

肥胖与肠道菌群之间潜在的联系近年来引起了许多研究者的关注。许多研究表明，肠道菌群的紊乱可能是造成肥胖的一个重要原因[6-7,16]。有研究发现肥胖人群肠道中的 *Firmicutes*/*Bacteroidetes* 比值较低[17-18]，然而有些研究却发现相反的结果，如 Mai 等[19]的研究并未发现 *Firmicutes*/*Bacteroidetes* 比值与 BMI 之间存在着关联。因此，关于 *Firmicutes* 和 *Bacteroidetes* 在肥胖人群肠道菌群中丰度的变化还需进一步研究。在属水平上，Schwiertz 等[17]报道了在肥胖人群肠道中的 *Methanobrevibacter* 相对丰度较健康人群有所降低。

本研究从美国肠道计划的公开数据库中选取健康和肥胖的成年人肠道菌群样本。在样本量较大的基础上，从 α 多样性、β 多样性、物种差异以及代谢功能等多方面系统地解析肥胖人群肠道菌群的特征，并基于肠道菌群数据建立了肥胖机器学习分类模型，为以后进一步了解肥胖与肠道菌群的关系提供基础。同时，通过更为深入地认识肥胖人群肠道菌群特征以及建立机器学习模型，以期为基于肠道菌群来干预肥胖提供新的理论和方法。

1 材料与方法

1.1 肠道数据来源

所用的数据来源于美国肠道计划的公开数据集[20]，从中筛选出最终有效测序序列在 1 250 以上的肠道菌群样本。之后，再从中根据 BMI 选取 1 655 个健康人群样本和 898 个肥胖人群样本。健康人群样本的定义为：BMI 在 18.5−25.0 之间，一年内无抗生素药物服用史、无炎症性肠炎和糖尿病病史；肥胖人群的定义为 BMI 在 30 以上。α 多样性指数(Chao1 指数、Observed otus 指数、PD whole tree 指数和 Shannon 指数)，β 多样性(Unweighted 和 Weighted UniFrac 距离)和 OTU 表均源于美国肠道计划基于 QIIME 分析平台得来。

1.2 数据分析方法

数据分析主要基于 R 3.4.4 平台，利用 Wilcox 秩和检验对肥胖与健康人群的 α 多样性、属水平物种、预测代谢通路进行差异分析。Logistic 回归用来检验 α 多样性与人体肥胖状态的关联。主成分分析(principal component analysis，PCA)用来比较分析基于 3 种不同 β 多样性距离在两组之间的差异，其中 Bray-Curtis 距离是利用 Vegan 软件包[21]基于属水平的 OTU 表计算得来。预测的代谢通路由 PICRUSt 软件[22]结合 KEGG 数据库[23]进行预测注释，并通过 R 语言中 Psych 软件包中的 corr.test 数进行相关性分析。

1.3 机器学习模型建立方法

机器学习模型使用基于 Python 的 Scikit-Learn 机器学习平台[24]而建立。选用核支持向量机、随机森林、梯度提升回归树和 Back propagation (BP) 神经网络 4 种不同的机器学习算法进行模型的建立，并使用网格搜索方法确定最佳参数。由于每个模型的参数种类繁多，取值范围大，将每个参数都考虑在内显然是不现实的，因此只选择对于每个模型中影响最重要的参数进行网络搜索来确定最佳模型的参数。

影响核支持向量机性能最重要的参数是 C 和

Gamma，其中 C 是惩罚函数，其值决定了模型对于误差的容忍程度。C 值越大，则模型对于误差的容忍度越差，模型容易过拟合；反之，模型容易欠拟合。Gamma 值是径向基函数(radial basis function, RBF)核自带的一个参数，主要隐含地决定了新特征的分布，Gamma 值越大，支持向量越少，反之则支持的向量越多，而支持向量的数量会影响模型的复杂程度。

随机森林是由许多个决策树集合而来的，因此影响随机森林的重要参数有决策树的数目(n_estimators)和单颗决策树可以使用的最大特征数目(max_features)。一般对于随机森林而言，决策树的数目总是越多越好，但是过多的决策树会增加模型的复杂程度及造成计算开销过大。同样地，单颗决策树可以使用的最大特征越大，模型的效果越好，但是会造成决策树的多样性较低和计算开销过大。

梯度提升回归树需要调整的参数是决策树的数目(n_estimators)和学习率(learn_rate)。与随机森林不同，梯度提升回归树的性能并不强依赖于决策树的数目，决策树的数目不是越大越好。学习率是指梯度提升回归树中每颗决策树对前一决策树误差的矫正程度，学习率越大，模型矫正效果越好，但同时也会增加计算开销。

BP 神经网络选择三层结构，其最主要的参数就是隐含层的节点数目。主要参考以下几个经验公式确定：

$$M=\sqrt{m+n}+a \tag{1}$$

$$M=\log_2 n \tag{2}$$

$$M=\sqrt{mn} \tag{3}$$

其中，M 为隐含层的节点数，k 为训练网络的样本数，n 是输入层的节点数，m 为输出层的节点数，a 为取值在[0,10]之间的常数。

最终确定了每个模型的超参数空间。其中，SVM 的超参数空间为：C=[0.1,2]，Gamma=[0.1,2]；对于随机森林而言，n_estimators 的取值分别为 100、1 000、10 000，max_features 的取值范围分别为 auto、$\log_2$、sqrt；对于梯度提升回归树，n_estimators 的取值与随机森林相同，learn_rate 的取值范围为[0.1,1]，步长为 0.1。神经网络的隐含层取值分别为 11、46、64。

本研究通过使用 Scikit-Learn 中将肠道菌群的数据划分为两部分，其中 70%的样本用于模型的训练，30%样本用于模型的验证，并采用受试曲线下与坐标轴围成的面积(area under curve，AUC)值作为评判模型好坏的标准[25]。

2 结果与分析

2.1 α 多样性分析

本研究中采用了 Observed otus、Chao1、Shannon 和 PD whole tree 四种 α 多样性指数。其中，Observed otus 和 Chao1 主要用来表示某一群落中的物种丰富度，Shannon 指数则反映群落中的物种的稳定性，PD whole tree 主要反映物种进化上的多样性。通过 Wilcox 秩和检验分析，4 种 α 多样性指数在肥胖人群肠道中都显著降低(P<0.01)，表明肥胖人群中肠道微生物的丰富度和稳定性都显著低于健康人群(图 1A)。

为了探究 α 多样性指数是否与肥胖存在关联，利用 Logistic 回归分析 4 种 α 多样性与肥胖的关联性(图 1B)。结果表明，PD whole tree 与人体的肥胖情况存在着强烈的相关性，表明肠道菌群多样性的变化与肥胖存在关联。

2.2 β 多样性分析

本研究中使用了 3 种不同的 β 多样性距离：Unweighted UniFrac、Weighted UniFrac 和 Bray-Curtis 距离。其中 UniFrac 距离的计算需要各个 OTU 的系统进化树，通过计算进化树各物种的系统发育关系来计算样本间的距离。Unweighted UnFrac 和 Weighted UniFrac 距离的差别在于有无考虑不同环境样本的相对丰度。Bray-Curtis 距离主要基于 OTU 表的计数统计，从而比较两个群落微生物组成上的差异。通过 PCA 分析，我们发

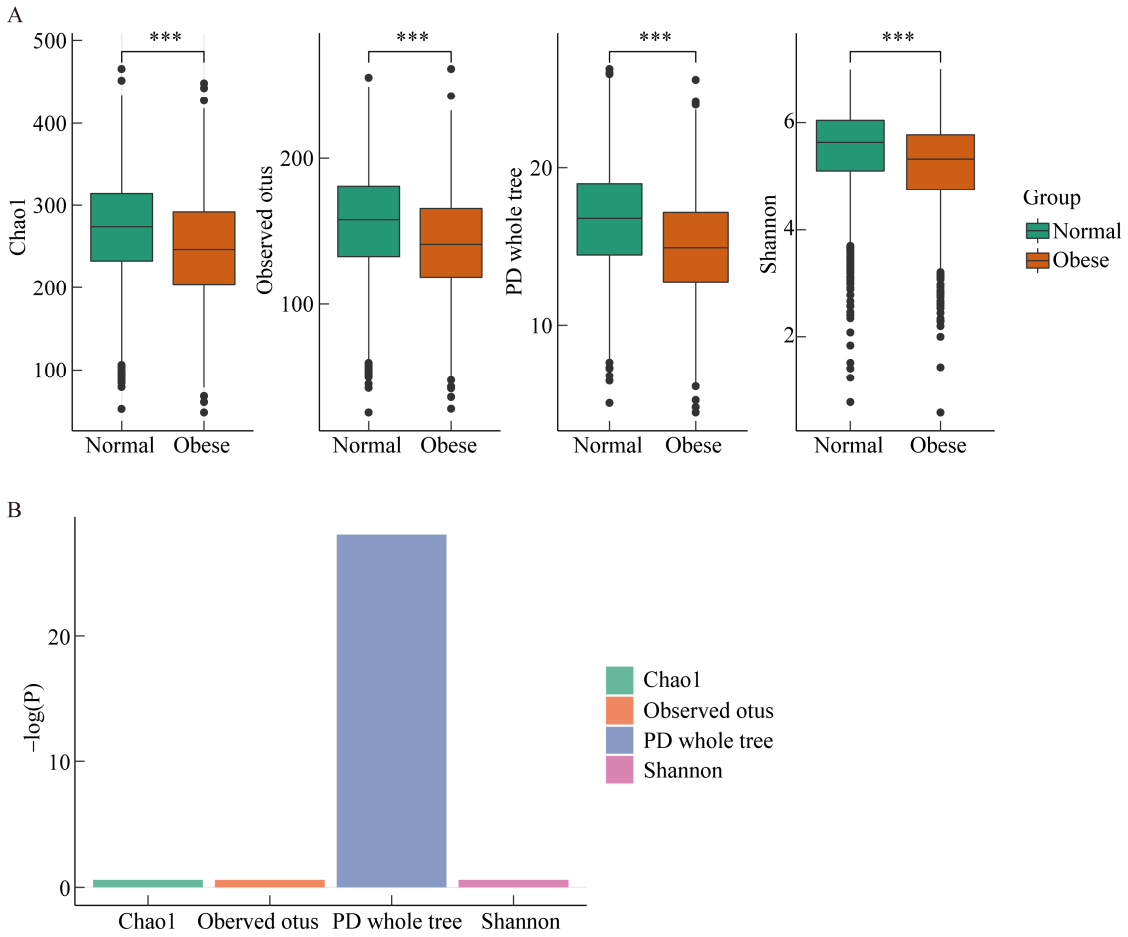

图 1　α 多样性对比(A)及 Logistic 回归分析(B)结果

Figure 1　The comparison of alpha-diversity (A) and the Logistic regression (B)

现肥胖人群和健康人群的肠道菌群在组成上没有显著的差异，说明两组人群的肠道菌群在结构上相似(图 2)。

2.3　肥胖和健康人群肠道菌群物种差异

门水平上，肥胖人群的肠道菌群中 *Firmicutes* 门的相对丰度小于健康人群，*Bacteroidetes* 门的相对丰度大于健康人群。分别计算两组人群中的 *Firmicutes*/*Bacteroidetes* 比值并进行 Wilcox 秩和检验，发现肥胖人群肠道中的 *Firmicutes*/*Bacteroidetes* 比值小于健康人群(图 3A)。

属水平上，通过 Wilcox 秩和检验，并通过 FDR 对 *P* 值进行矫正，从共 2 038 个属中找出 57 个具有显著差异(P<0.01)的菌属，将其中相对丰度排名前 10 的菌属通过 R 语言中的 ggplot2 可视化，如图 3B 所示，可以看出 *Bacteroides* 属(肥胖组：28.71%；健康组：24.27%)，*Blautia* 属(肥胖组：3.01%；健康组：2.37%)和 *Parabacteroides* 属(肥胖组：2.71%；健康组：2.17%)在肥胖人群肠道中的相对丰度较高，而 *Prevotella* 属(肥胖组：3.83%；健康组：5.71%)和 *Faecalibacterium* 属(肥胖组：6.53%；健康组：7.21%)的相对丰度在肥胖人群肠道中较低。除此之外，有可能作为下一代益生菌的 *Akkermansia* 属在肥胖人群肠道中的含量也较低，其在肥胖人群肠道中的平均相对丰度为 1.72%，而在健康人群肠道中的平均相对丰度为 1.94%；*Methanobrevibacter* 属在肥胖人群肠道中的相对丰度也较低，仅为 0.036%，而其在健康人群肠道中平均相对丰度为 0.049%。

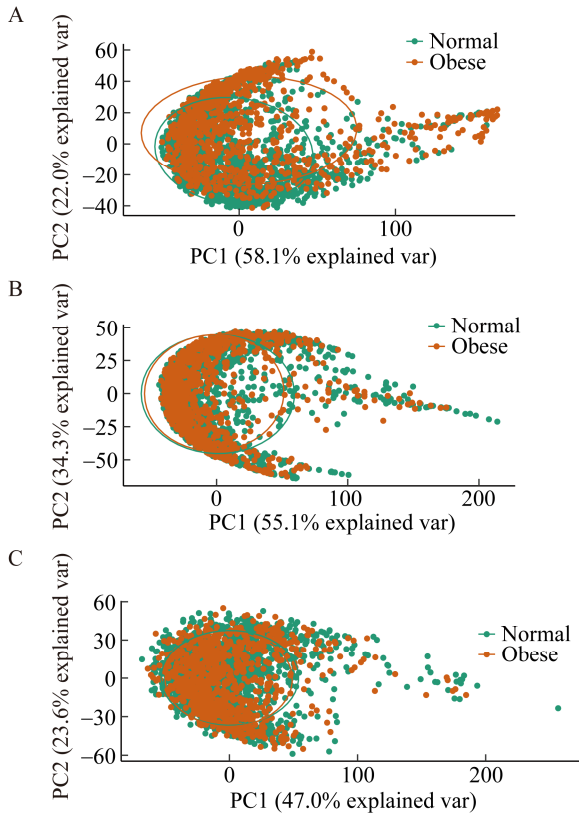

图 2　三种不同 β 多样性距离的 PCA 分析结果

Figure 2 The PCA analysis result based on three beta-diversity distance

注：A：基于 Bray-Curits 距离的 PCA 分析；B：基于 Weighted UniFrac 距离的 PCA 分析；C：基于 Unweighted UniFrac 距离的 PCA 分析.

Note: A: The PCA analysis result based on the Bray-Curtis distance; B: The PCA analysis result based on the Weighted UniFrac distance; C: The PCA analysis result based on the Unweighted UniFrac distance.

2.4　肥胖和健康人群肠道菌群代谢通路差异

利用 PICRUSt 分析软件基于属水平的 OTU 表进行代谢通路的预测分析，并使用第三层级进行注释。经过 Wilcox 秩和检验并通过 FDR 矫正 *P* 值，得到 67 个具有极显著差异的代谢通路 (P<0.01)。之后，将这 67 个的代谢通路与 2.3 节具有显著差异的菌属使用斯皮尔曼相关性系数进行相关性分析(图 4)，其中红色代表代谢通路和菌属之间存在正相关，蓝色代表负相关，白色部分则代表没有显著的相关性。经过相关性分析发现，*Akkermansia* 属与 Fluorobenzoate degradation 、 Steroid biosynthesis 、 Caffeine metabolism 和 Fatty acid elongation in miochaondiras 等代谢通路具有较强的相关性。同样，*Methanobrevibacter* 属与 Bile secretion 和 Various types of N-glycan biosynthesis 等代谢通路具有较强的相关性。

2.5　基于肠道菌群的肥胖人群预测模型

首先使用了 Scikit-Learn 软件包中的核支持向量机、随机森林、梯度提升回归树和 BP 神经网络 4 种机器学习算法的默认参数建立基线模型作为后续网络搜索的依据。发现其中梯度提升回归树的模型性能较好，其 AUC 值达到 0.769，测试集精度达到 0.725。

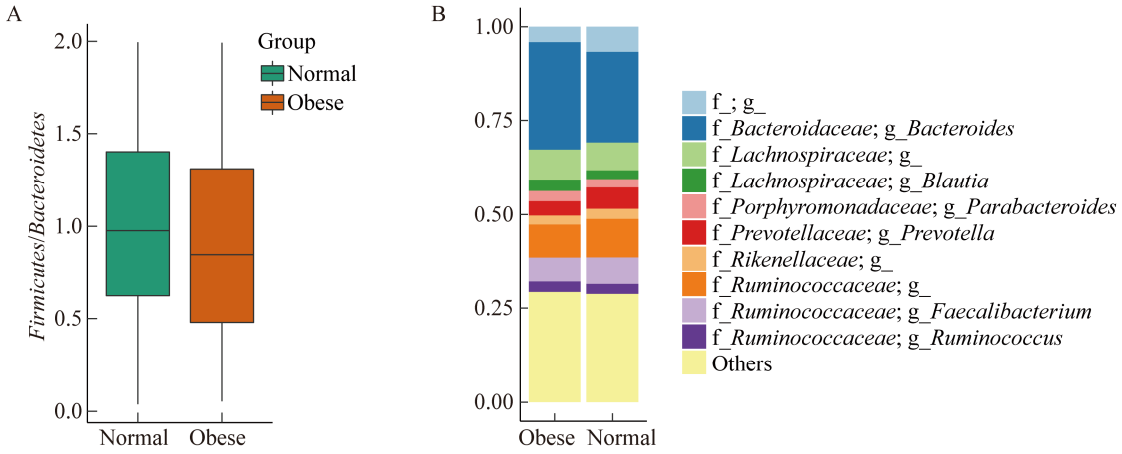

图 3　肥胖与健康人群在的 *Firmicutes*/*Bacteroidetes* 值(A)和属水平物种分布图(B)

Figure 3 The *Firmicutes*/*Bacteroidetes* ration and genus bar plot between the obese (A) and normal people (B)

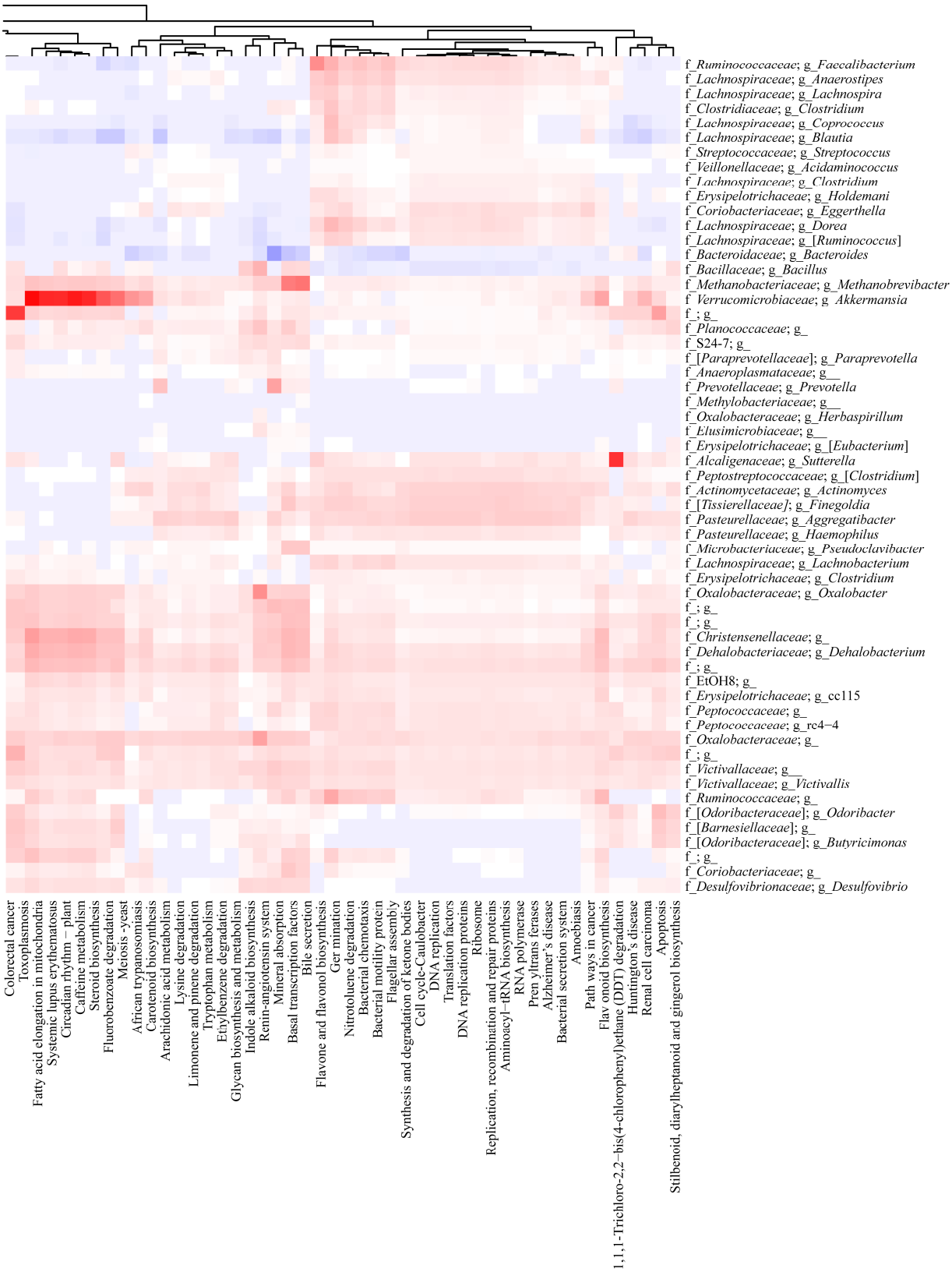

图 4　差异代谢通路与代谢菌属之间的相关性

Figure 4　The correlation between the significantly different pathway and genus

通过 Scikit-Learn 中的网格搜索，我们将 4 种模型的超参数空间中所有的参数组合进行遍历，并使用五折交叉验证保证结果的可靠性，同时对模型调整参数前后的测试集精度和 AUC 值进行比较(图 5)。结果表明，梯度提升回归树的最佳性能是其默认参数，即 Learning_rate 为 1，n_estimators 为 100；随机森林的 AUC 值由 0.639 上升到 0.754，测试集精度也由 0.668 提高到 0.698，其对应的最佳参数组合为，n_estimators 为 10 000，max_features 为 Log_2。核支持向量机经过网格搜索后，发现其最佳参数为其默认参数，即 C=1，Gamma=1，其 AUC 值为 0.707，测试集精度为 0.656。BP 神经网络隐含层的最佳节点数为 11，AUC 值由 0.615 上升到 0.641，测试集精度由 0.632 上升到 0.64。

如图 6 所示，4 种模型的受试者工作特征(receiver operating characteristic，ROC)都将随机猜测的 ROC 曲线包含在内，证明了基于肠道菌群用来预测人体肥胖状况的可行性。同时，可以看出梯度提升回归树和随机森林的 AUC 曲线都将 BP 神经网络和核支持向量机的 ROC 曲线包含在内，说明这两种基于决策树的模型效果好于 BP 神经网络和核支持向量机。梯度提升回归树和随机森林的 ROC 曲线有交叉，两者的 AUC 值也相差不大，难以从 ROC 曲线上判断两个模型的优劣。但是结合测试集的精度，梯度提升回归树的测试集精度(0.725)大于随机森林的测试集精度(0.698)。由此得出结论，梯度提升回归的模型性能在本研究中优于其他 3 种模型，可以更为有效地基于肠道菌群预测肥胖。

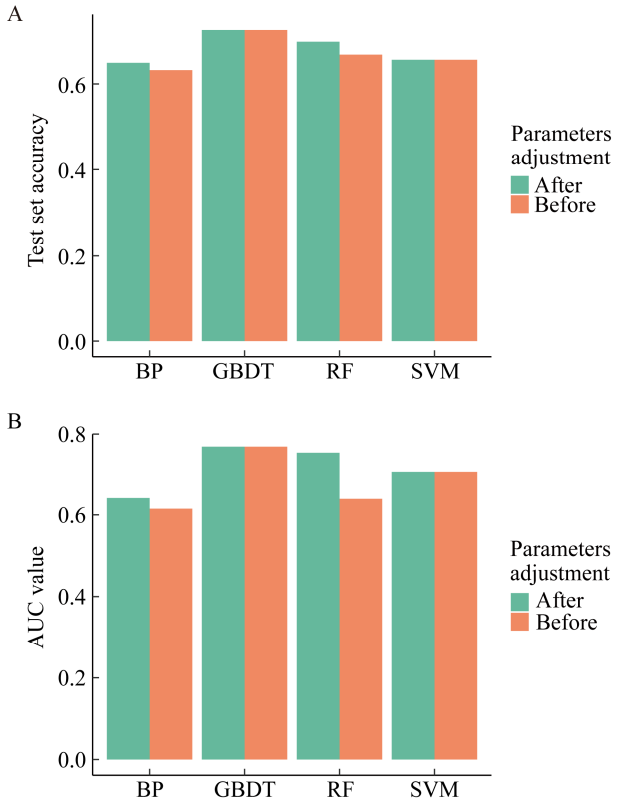

图 5　网络搜索前后的测试集精度(A)和 AUC 值(B)

Figure 5　The test set precision (A) and AUC value (B) before and after grid searching

注：BP：BP 神经网络；GBDT：梯度提升回归树；RF：随机森林；SVM：核支持向量机.

Note: BP: BP neural networks; GBDT: Gradient boosting decision tree; RF: Random forest; SVM: Supported vector machine.

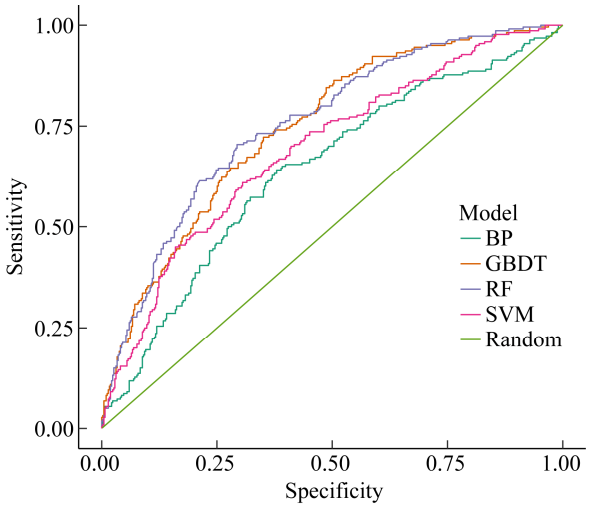

图 6　经过网络搜索后的模型的 ROC 曲线

Figure 6　The ROC curve of all the models after the grid searching

注：BP：BP 神经网络；GBDT：梯度提升回归树；RF：随机森林；SVM：核支持向量机；Random：随机预测.

Note: BP: BP neural networks; GBDT: Gradient boosting decision tree; RF: Random forest; SVM: Supported vector machine; Random: Random prediction.

3 讨论与结论

肥胖作为常见的现象严重影响着人们的健康生活。肠道菌群作为人体一种新的“器官”，吸引了国内外许多研究者的注意，很多研究都报道了肥胖与肠道菌群之间的关系。先前有研究报道，肥胖人群肠道菌群中 *Firmicutes* 门的含量较低，*Bacteroidetes* 门的含量较高[17-18]，与本研究中的报道一致。然而，有关肥胖人群肠道 *Firmicutes* 门和 *Bacteroidetes* 门的相对含量与健康人群相比的结论并不一致，有些研究甚至报道了相反的结果[7]。然而这些研究普遍存在着样本量较低的问题，说服力有限。

基于 3 种不同 β 多样性距离的 PCA 分析表明，肥胖人群和健康人群的肠道菌群在结构上无明显差异。但是，α 多样性指标以及后续的物种差异性分析仍旧表明肥胖人群与健康人群的肠道菌群存在着一定程度的差异。尤其是 *Akkermansia* 属对于维持人体健康具有重要的作用，有作为下一代益生菌的可能。在本次研究中，我们不仅观测到了 *Akkermansia* 属的相对丰度在肥胖人群肠道中较低。同时，基于 PICRUSt 预测分析代谢通路，我们对肠道菌群影响肥胖的机制进行了初步探索。通过构建具有显著差异的菌属和代谢通路之间的相关性分析，发现了 *Akkermansia* 属与许多与代谢功能有关的代谢通路具有很强的正相关性。这一发现为探索 *Akkermansia* 属影响人体健康的机制提供了启发。作为同样和许多代谢通路具有较强相关性的 *Methanobrevibacter* 属，本研究中发现 *Methanobrevibacter* 属在肥胖人群肠道中的含量较低，这一结果与文献[17,26-27]报道一致。

基于分析结果，我们尝试使用机器学习的方法构建肥胖预测模型。通过系统地网络参数搜索，我们发现基于肠道菌群的梯度提升回归树具有良好的性能。一般而言，AUC 值大于 0.75 时模型的性能就较好，而在本次研究中梯度提升回归树的模型 AUC 值则达到了 0.769，测试精度达到 0.725。这表明基于肠道菌群可以预测人体健康肥胖状况，揭示了肠道菌群的又一功能。不仅如此，这一模型的建立也为基于肠道菌群干预肥胖提供了重要参考。本文虽然针对肥胖进行了机器学习模型的研究，但是基于肠道菌群同样可以用来预测其他疾病。如 Ren 等[28]通过建立随机森林机器学习模型发现肠道菌群可以作为诊断早期肝癌的工具，其模型的 AUC 值可达到 0.806。Shah 等[29]从公开数据收集到 509 个样本(79 个结肠瘤样本、195 个结肠癌样本和 235 个对照样本)的肠道菌群数据，采用随机森林的方法建立起机器学习模型用于结肠瘤和结肠癌的分类，其中最佳模型的 AUC 值可达 0.913。Loomba 等[30]则针对非酒精性脂肪肝基于肠道菌群建立随机森林模型进行预测，其模型 AUC 值达到了 0.936。Eck 等[31]通过建立一系列 AUC 值在 0.85 以上的机器学习模型，确定基于肠道菌群可以作为诊断炎症性肠炎的可靠工具。He 等[32]基于广东肠道计划，对于炎症性肠炎、糖尿病、结肠癌等与代谢相关的疾病使用随机森林建立机器学习模型，他们发现地域会影响基于肠道菌群预测疾病模型的精度，地域范围越小则模型精度更高，从而揭示了地域可能是影响肠道菌群疾病预测模型准确度的重要因素。上述一系列研究，不仅表明了肠道菌群在疾病预测方面具有巨大的潜力，同时也为日后的精准膳食、精准医疗提供了扎实的理论基础。

总之，我们通过系统分析探究了肥胖人群和健康人群肠道菌群的差异，同时基于肥胖人群和健康人群肠道菌群的差异，利用机器学习的方法建立了肠道菌群肥胖预测模型并取得了良好的效果，为基于肠道菌群精准膳食、精准医疗提供了新思路。

REFERENCES

[1] Abelson P, Kennedy D. The obesity epidemic[J]. Science, 2004, 304(5676): 1413

[2] Misra A, Vikram NK. Insulin resistance syndrome (metabolic syndrome) and obesity in Asian Indians: evidence and implications[J]. Nutrition, 2004, 20(5): 482-491

[3] Gu DF, Reynolds K, Wu XG, et al. Prevalence of the metabolic syndrome and overweight among adults in China[J]. The Lancet, 2005, 365(9468): 1398-1405
[4] Arumugam M, Raes J, Pelletier E, et al. Erratum: Enterotypes of the human gut microbiome[J]. Nature, 2011, 474(7353): 666
[5] Qin JJ, Li RQ, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing[J]. Nature, 2010, 464(7285): 59-65
[6] Reinhardt C, Reigstad CS, Bäckhed F. Intestinal microbiota during infancy and its implications for obesity[J]. Journal of Pediatric Gastroenterology and Nutrition, 2009, 48(3): 249-256
[7] Ley RE, Turnbaugh PJ, Klein S, et al. Human gut microbes associated with obesity[J]. Nature, 2006, 444(7122): 1022-1023
[8] Ghoshal UC, Shukla R, Ghoshal U, et al. The gut microbiota and irritable bowel syndrome: friend or foe?[J]. International Journal of Inflammation, 2012, 2012: 151085
[9] Jeffery IB, Claesson MJ, O'Toole PW, et al. Categorization of the gut microbiota: enterotypes or gradients?[J]. Nature Reviews Microbiology, 2012, 10(9): 591-592
[10] Manichanh C, Borruel N, Casellas F, et al. The gut microbiota in IBD[J]. Nature Reviews Gastroenterology & Hepatology, 2012, 9(10): 599-608
[11] Li QR, Wang CY, Tang C, et al. Molecular-phylogenetic characterization of the microbiota in ulcerated and non-ulcerated regions in the patients with crohn's disease[J]. PLoS One, 2012, 7(4): e34939
[12] Arthur JC, Perez-Chanona E, Mühlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota[J]. Science, 2012, 338(6103): 120-123
[13] Castellarin M, Warren RL, Freeman JD, et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma[J]. Genome Research, 2012, 22(2): 299-306
[14] Clarke SF, Murphy EF, Nilaweera K, et al. The gut microbiota and its relationship to diet and obesity[J]. Gut Microbes, 2012, 3(3): 186-202
[15] Qin JJ, Li YR, Cai ZM, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60
[16] Angelakis E, Armougom F, Million M, et al. The relationship between gut microbiota and weight gain in humans[J]. Future Microbiology, 2012, 7(1): 91-109
[17] Schwiertz A, Taras D, Schäfer K, et al. Microbiota and SCFA in lean and overweight healthy subjects[J]. Obesity, 2010, 18(1): 190-195
[18] Collado MC, Isolauri E, Laitinen K, et al. Distinct composition of gut microbiota during pregnancy in overweight and normal-weight women[J]. The American Journal of Clinical Nutrition, 2008, 88(4): 894-899
[19] Mai V, McCrary QM, Sinha R, et al. Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers[J]. Nutrition Journal, 2009, 8: 49
[20] McDonald D, Hyde E, Debelius JW, et al. American gut: an open platform for citizen science microbiome research[J]. mSystems, 2018, 3(3): e00031-18
[21] Dixon P. VEGAN, a package of R functions for community ecology[J]. Journal of Vegetation Science, 2003, 14(6): 927-930
[22] Langille MGI, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences[J]. Nature Biotechnology, 2013, 31(9): 814-821
[23] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Research, 2000, 28(1): 27-30
[24] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830
[25] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(7): 1145-1159
[26] Armougom F, Henry M, Vialettes B, et al. Monitoring bacterial community of human gut microbiota reveals an increase in *Lactobacillus* in obese patients and *Methanogens* in anorexic patients[J]. PLoS One, 2009, 4(9): e7125
[27] Million M, Maraninchi M, Henry M, et al. Obesity-associated gut microbiota is enriched in *Lactobacillus reuteri* and depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*[J]. International Journal of Obesity, 2012, 36(6): 817-825
[28] Ren ZG, Li A, Jiang JW, et al. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma[J]. Gut, 2019, 68(6): 1014-1023
[29] Shah MS, DeSantis TZ, Weinmaier T, et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer[J]. Gut, 2018, 67(5): 882-891
[30] Loomba R, Seguritan V, Li WZ, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease[J]. Cell Metabolism, 2017, 25(5): 1054-1062.e5
[31] Eck A, de Groot EFJ, de Meij TGJ, et al. Robust microbiota-based diagnostics for inflammatory bowel disease[J]. Journal of Clinical Microbiology, 2017, 55(6): 1720-1732
[32] He Y, Wu W, Zheng HM, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models[J]. Nature Medicine, 2018, 24(10): 1532-1535