第 8 卷 第 5 期
2019 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 8 No. 5
Sep. 2019

引文格式：
李倩莹, 蔡云鹏, 张凯. 基于网络嵌入方法的肠道微生物组大数据网络分析［J］. 集成技术, 2019, 8(5): 34-48.
Li QY, Cai YP, Zhang K. Inferring gut microbial interaction network from microbiome data using network embedding algorithm［J］. Journal of Integration Technology, 2019, 8(5): 34-48.

基于网络嵌入方法的肠道微生物组大数据网络分析

李倩莹[1,2,3] 蔡云鹏[1,3] 张 凯[1,3]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学 北京 100049)
[3](健康大数据智能分析技术国家地方联合工程实验室 深圳 518055)

摘 要 厘清菌群群落与环境的相互关系及其潜在的驱动机理是肠道微生物研究的一项关键任务。通过微生物组高通量测序和大数据分析辨识微生物组分及功能是目前微生物群落分析的主要方法。现有人体肠道微生物的研究主要侧重于描述肠道菌群多样性和组成特征，缺少更深层次的菌群内部互利共生关系及其生态演替的探索。如何由微生物组数据从分子网络角度来研究肠道菌群分布的关联模式是目前亟待解决的问题。该文使用机器学习领域的网络嵌入方法改进传统生物网络结构学习技术过于依赖节点间的个体相关关系的弊端，更准确地捕捉微生物网络关联的异构性、隐变量和不均衡性等特征。通过对生成的模块与环境变量以及关键代谢物的进行相关性分析，证实了新的网络模块挖掘方法可以更好地提取肠道菌群结构中之前较少被认识到的特征模块，从而更好地评估菌群与菌群之间、菌群与环境之间的制约关系以及菌群代谢功能之间的潜在耦合机制。该研究中描述的方法不仅给肠道微生物群落结构的解析提供了新视角，还可以拓展应用到其他环境微生物领域的研究，通过数据的多阶信息更好地反映群落结构的驱动过程。

关键词 生物网络；网络嵌入；聚类；相关性分析
中图分类号 TP 399 **文献标志码** A doi:10.12146/j.issn.2095-3135.20190704001

Inferring Gut Microbial Interaction Network from Microbiome Data Using Network Embedding Algorithm

LI Qianying[1,2,3] CAI Yunpeng[1,3] ZHANG Kai[1,3]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology*, *Shenzhen* 518055, *China*)

Abstract Identifying the relationship between the gut microbial community and the host environment,

收稿日期：2019-07-04 **修回日期**：2019-08-12
基金项目：国家自然科学基金联合基金项目(U1801265)；深圳市经贸委“创新链＋产业链”融合专项扶持计划项目(20170502171625936)
作者简介：李倩莹，硕士研究生，研究方向为生物信息学；蔡云鹏(通讯作者)，研究员，研究方向为生物信息学，E-mail：yp.cai@siat.ac.cn；张凯，助理研究员，研究方向为环境微生物学。

as well as the driving mechanism, are the key tasks in gut microbial research. Microbiome high-throughput gene sequencing and big data analysis are currently the mostly used techniques for investigation microbial communities. Existing studies on human gut microbiota data mainly focus on the community diversity and composition, while methods for deep exploration of the ecological and functional relationships among bacteria species are still lacking. An urgent task is therefore raised on developing computational methods to explore the interaction pattern between gut microbial components from in the sense of molecular network from microbiome sequencing data. In this paper, we adopt the network embedding method proposed in machine learning as a remediation to the drawbacks of traditional biological network learning technology which were solely dependent on the direct correlation between nodes, with stronger power in capturing the heterogeneity, hidden variables and imbalance features in microbial network interactions. By analyzing the correlation between the created function modules with the new approach and the environmental variables as well as key metabolite components, it is confirmed that the derived functional modules managed to identify biological-relevant feature that can be less recognized with previous approaches, which are helpful for further modeling of the potential coupling mechanisms among the biological systems. The method described in this article not only provides a new perspective for the analysis of gut microbial community structure, but also can be extended to other environmental microbiology research and reflect the driving process of community structure through multi-level information of data.

Keywords　microbial network; network embedding; clustering; correlation analysis

1　引　言

人体肠道内含有超过 100 万亿个微生物细胞，是人体细胞数量的 10 倍，这些微生物所携带的基因约为人类基因组的 150 倍[1]。研究[2-3]表明，肠道菌群与宿主能量代谢、脂质积累和免疫密切相关。其中，微生物菌群发酵底物产生短链脂肪酸、有机酸及其他小分子化合物，在维持机体健康和诱发疾病方面具有重要作用。

研究微生物群落有效且通用的方法是原核生物核糖体小亚基 rRNA（16S rRNA）基因（16S rDNA）的序列分析[4]，利用 16S 保守区的扩增子测序获得菌群结构系统发育水平上的分类学信息，从而分析微生物的物种组成和多样性。然而，受到包括宿主的地域、年龄、生理状况、饮食习惯等[5]诸多因素的影响，肠道微生物群落组成在个体之间差异很大。此外，微生物群落不仅是独立个体的集合，更是交互、交叉、重组和共同进化的相互关联的生态群落的复合体[6]，对菌群之间相互作用的研究具有重要的理论意义。目前，肠道微生物的相关研究较少涉及到菌群-菌群之间的网络分析。与基于系统发育的肠道菌群的多样性分析相比，网络分析在微生物之间的“生态位”和共进化模式的挖掘上具有较强的应用潜力[7]。通过引入网络拓扑算法来揭示菌群数据的模块化结构[8]，开发基于网络嵌入的聚类新方法和新模型有助于更好地预测菌群社区的变化和扰动的影响[9]。

网络作为刻画大数据关系的有力工具，逐渐被应用于分析各种高通量生物数据。利用共存数据或丰度数据预测微生物关联网络是一个经典的计算科学中网络关系推理的问题[10-11]。目前，网络技术广泛应用于基因组学相关研究[12-13]，并逐渐开始应用在生态学研究中[14]。现有两种网络

推理方法：预测两个物种之间成对关系的方法和两个物种间更为复杂关系的方法。传统的微生物菌群成对关系的分析方法仅仅考虑了菌株与菌株一阶信息上的差异，而网络分析将同时考虑二阶信息上的差异，这为解释复杂肠道环境中微生物与宿主之间分子网络的生物意义提供了新视角。

微生物组学大数据存在着很多特点，如：(1)异构性，微生物群落往往受环境影响，动态的环境变量会增加样本异质性，进而对网络结构产生影响，导致不同数据子集上得到的网络结构不一致；(2)隐关系，具体以“生态位”的概念为例，生态位相同或竞争的群落关系不一定直接反映在一阶相关性上；(3)隐变量，以微生物-代谢物网络为例，微生物通过代谢物相互联系，而代谢物为隐变量，难以观测；(4)不均衡性，以微生物群落结构为例，优势菌株丰度高但种类少，而稀有菌株丰度低但种类多。以上导致微生物组学分析过程中可能存在假阳性比例过高，构建的网络关系重复性较差等问题，这影响网络分析结果的准确性和可靠性。本文引入目前应用于部分生物网络研究中的新兴网络技术——网络嵌入方法[15]，试图找到网络中节点间更复杂的相互作用，来优化生物数据中相关性的解析。通过网络结构中的连接关系，得到网络中顶点的向量表示，作为基本特征应用到聚类分析上。

2 数据及处理

2.1 数据来源及处理

肠道菌群在微生物研究中有重要地位且肠道菌群间相互作用仍有很多探究空间，本文考虑在肠道微生物菌群上开展研究。

(1)数据来源：本文使用 Eva-Maria 等[16]研究中的公开数据(ENA；BioProject No.：PRJEB21169)，原文在缺氧条件下将一种具有悠久传统和良好抗炎活性的草药产品柳树皮提取物与人粪便悬浮液一起培养，分别在培养 0.5 h、4 h 和 24 h 后取样，利用液相色谱-质谱代谢组分析和 16S rRNA 微生物组测序得到目标实验数据。

(2)数据概况：使用数据来源论文中的数据集，整理成一个格式化的 Biom 文件，每行为各个操作分类单元(Operational Txonomic Unit，OTU)，每列为各个样本，共有 43 387 行，72 列。表格内数据为 OTU 在样本中的相对含量，值域为 0～1。

(3)数据处理：将数据中相对含量之和小于 0.1% 的 OTU 数据删除，以减少后续研究的噪声。

(4)数据环境标签选取：选取样本培养时间点(time point)为标签，共有 3 个取值，分别是 0.5 h、4 h、24 h。

(5)代谢物标签选取：选取 Eva-Maria 等[16]研究中的 58 种主要代谢物，得到每一种代谢物在不同环境标签下的丰度。当某种环境下某种代谢物未被检测到时，其丰度置为 0。

2.2 评估方法

本文针对数据进行无监督聚类，然后使用轮廓系数评估聚类结果。其中，轮廓系数是一种评估聚类好坏的评价指标，结合了内聚度和分离度两种因素，可以用来评估不同算法在相同数据集上的效果，或同种算法在不同运行方法下聚类结果的变化。

轮廓系数方法将待分类数据进行聚类分出 n 个簇，对于簇中每一个样本 i，计算方法如下。

(1)簇内不相似度：计算样本 i 到簇内其他样本的平均距离 a_i，即为该样本 i 的簇内不相似度。a_i 越小，说明样本 a_i 越被聚类至该簇。聚类 C 中所有样本的 a_i 均值被称为簇 C 的簇不相似度。

(2)簇间不相似度：计算样本 i 到所有其他簇的平均距离 $b_{ij}(j=1,2,\cdots,n)$，取其中最小值 $b_i=\min(b_{i1},b_{i2},\cdots,b_{in})$，$b_i$ 即为样本 i 与簇 C_j 的不相似度。b_i 越大，说明样本 i 越不属于其他簇。

(3)样本 i 的轮廓系数：

$$S(i)=\frac{b(i)-a(i)}{\max\{b(i),a(i)\}} \tag{1}$$

轮廓系数值域介于[−1,1]，越趋近于 1 表示内聚度和分离度都相对较优。

3　研究方法

本文主要实验流程如图 1 所示。

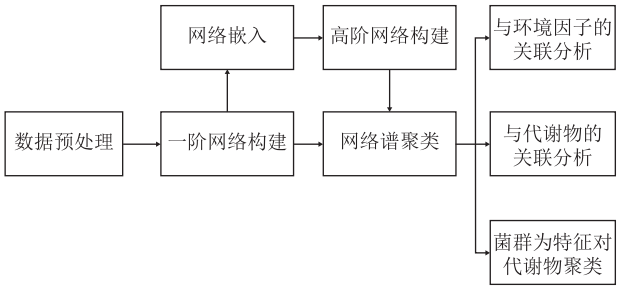

图 1　技术路线图

Fig. 1　Technical framework

3.1　基于皮尔森相关系数构建网络

本文首先定义肠道菌群网络节点，利用皮尔森(Pearson)相关系数来计算每个节点的相关性，从而构建肠道菌群相关性网络。网络中的每一个节点代表一个菌种，连接节点之间的每一条边代表各 OTU 菌群之间的相关性。其中，Pearson 相关系数是两个变量的协方差除以它们的标准偏差的乘积。给定两个随机变量(X,Y)，其中，$X=[x_1,x_2,\cdots,x_n]$，$Y=[y_1,y_2,\cdots,y_n]$。则随机变量 X 和 Y 的 Pearson 相关系数公式为：

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2}$$

其中，$\bar{x}$ 和 $\bar{y}$ 分别是 X 和 Y 的均值。Pearson 相关系数是−1～1 的实数。其中，1 表示强正相关；−1 表示强负相关；0 表示不相关。

3.2　网络谱聚类

谱聚类将菌群网络相关性网络对应的相关性矩阵作为输入，其中，相关性低的两个点之间的边权重值较低，相关性高的两个点之间的边权重值较高。通过对共存网络进行切割，使切图后不同的子图间边权重和尽可能地低，而子图内的边权重和尽可能地高，从而达到聚类的目的。通过谱聚类分析将网络进行分类，比 K 均值(K-means)聚类等传统算法更加便捷，并且可以通过标准线性代数方法有效地求解[16]，有助于挖掘菌群之间的关系以及同一类菌群与环境的关系[17]。

3.3　网络嵌入与二阶网络构建

在向量空间中使用图节点表示同时保留其属性的方法广受学界的关注[18-20]，在节点分类等多种任务中展示出较好的应用性。通常，用于解决图形问题的模型运用在原始图形邻接矩阵上或其衍生的向量空间上。这里可以将嵌入解释为描述图形数据的表示。因此，嵌入可以提供对网络属性的深入了解。嵌入作为模型的特征输入，并根据训练数据学习参数，减少了对直接应用于图表的复杂分类模型的需要。图形嵌入以两种方式使用：(1)在向量空间中表示整个图形；(2)在向量空间中表示每个单独的节点。本文使用后一种定义。

首先介绍关于网络嵌入相关的概念和定义。

定义 1(图)：一个图 $G(V, E)$由节点集合 $V=\{v_1,\cdots,v_n\}$和边集合 $E=\{e_{ij}\}_{i,j=1}^{n}$ 组成。图 G 的邻接矩阵 $\boldsymbol{S}$ 包含每条边的权重值 S_{ij}，如果节点 v_i 和 v_j 之间没有关联，则 $S_{ij}=0$。对于无向带权图，$S_{ij}=S_{ji}$，$\forall i,j\in\{1,\cdots,n\}$。边权 S_{ij} 通常被视为节点 v_i 和 v_j 之间的相似性的度量。其中，边权越高，表示两个节点越相似。

定义 2(一阶相似度)：网络的一阶相似度是指两个节点之间的局部成对相似度。对于由边(u, v)连接的两个节点，该边边权 w_{uv} 表示两个节点的一阶相似度。如果 u 和 v 之间没有边连接，则一阶相似度为 0。一阶相似度通常表示现实网络中两个节点的相似性。

定义 3(二阶相似度)：网络的二阶相似度表

示两个节点邻居网络结构的相似性。数学表达式为，若 $p_u=(w_{u,1},\cdots,w_{u,|V|})$表示节点 u 和所有其他节点的一阶相似度，则节点 u 和节点 v 之间的二阶相似度由 p_u 和 p_v 之间的相似性决定。若 u 和 v 没有相同的邻居节点，则 u 和 v 之间的二阶相似性为 0。

定义 4(网络嵌入)：给定一个大型网络 $G=(V, E)$，大规模信息网络嵌入的问题旨在将每个节点 $v\in V$ 映射到低维空间 R^d 中，即学习一个函数 $f_G:V\to R^d$，其中 $d\ll|V|$[19]。

3.3.1 DeepWalk

DeepWalk 是一种学习网络节点表示的新方法，使用随机游走(Random Walk)方法从截断的随机游走序列中得到网络的局部信息，学习到节点的向量表示。Skip-Gram(邻帧图)是 Word2vec 方法中使用单词来预测上下文的一个模型，通过最大化窗口内单词之间的共现概率来学习向量表示。DeepWalk 扩展 Skip-Gram 模型使用节点来预测上下文，并且不考虑句子中节点出现的顺序，具有相同上下文节点的表示相似节点(两个节点同时出现在一个序列中的频率越高，表明两个节点的相似度越高)，规定了节点相似性度量，即上下文的相似程度[21]。

3.3.2 Node2vecetor

Node2Vector 在 DeepWalk 的基础上改变了节点游走的方式[22]，提出了网络节点间游走(采样)的两种方式：Breadth-First Sampling(BFS)和 Depth-First Sampling(DFS)，原理分别对应广度搜索和深度搜索。Node2vecetor 采用两种游走方式结合的手段提取网络中一阶和二阶相似度信息，引入 p 和 q 参数分别作为 BFS、DFS 游走方法的权值。节点游走定义如下，π_{vx} 表示当前节点 v 到节点 x 的转移概率，$\pi_{vx}=\alpha_{pq}(t,x)\cdot\omega_{vx}$。其中，$\omega_{vx}$ 为节点 v 到节点 x 边上的权重；$\alpha_{pq}(t,x)$为当前节点 v 的上一次访问节点 t 转移规则，定义如公式(3)所示。

$$\alpha_{pq}(t,x)=\begin{cases}\dfrac{1}{p}, \text{if } d_{tx}=0\\ 1, \text{if } d_{tx}=1\\ \dfrac{1}{q}, \text{if } d_{tx}=2\end{cases} \tag{3}$$

其中，d_{tx} 表示 t 和 x 之间的最短路径。通过计算 t 能转移到不同类别节点的概率得到相应的节点概率向量，再使用 Word2vec 方法对节点 v 序列进行训练。本文通过网络嵌入方法中二阶相似度找到网络节点的隐关系，故在 Node2vec 方法中可以通过调高 p 值，来调高提取二阶相似度的 BFS 采样方法的比例。

3.3.3 LINE

LINE 方法定义了两个规则函数，分别用于一阶相似度和二阶相似度的计算，然后再最小化两个函数的组合。一阶相似函数与图因子分解函数类似，都是使两点的嵌入向量的邻接矩阵和点积更接近。二者的不同点在于图因子分解函数直接计算两者差值，而 LINE 方法为每个顶点定义两个联合概率分布：一个使用嵌入联合概率 p，如公式(4)所示；另一个使用邻接矩阵，对应联合概率 p 为公式(5)。然后，LINE 最小化这两个分布的相对熵，也称为 Kullback-Leibler(KL)散度，如公式(6)所示。两个分布和目标函数如公式(7)所示。

$$p_1(v_i,v_j)=\frac{1}{1+\exp\left(-\left\langle Y_i,Y_j\right\rangle\right)} \tag{4}$$

$$\hat{p}_1(v_i,v_j)=\frac{W_{ij}}{\sum\limits_{(i,j)\in E}w_{ij}} \tag{5}$$

$$O_1=KL(\hat{P}_1,P_1) \tag{6}$$

$$O_1=-\sum_{(i,j)\in E}W_{ij}\log p_1(v_i,v_j) \tag{7}$$

其中，Y 为节点 v 的向量表示；W_{ij} 为节点 v 之间的边权重；E 为节点的集合。

LINE 方法类似定义了二阶邻接概率分布和目标函数[19]。

3.3.4 二阶网络的构建

原始菌群网络经过网络嵌入后，得到各个菌群节点的嵌入向量。通过计算所有菌群嵌入向量之间的 Pearson 相关系数，得到相关性系数矩阵，再基于此相关性矩阵即可建立二阶网络。在简单一阶网络的基础上构建的二阶网络只关注节点间的直接关联，而本文基于网络嵌入后的向量构建的二阶网络，不仅可获取节点间关联的局部信息，还可利用网络嵌入提取全局网络潜在信息。同时，在数据量更复杂、大量节点不直接关联、局部信息不足以刻画整个网络时，网络嵌入能够从网络中提取更多的潜在信息和全局信息，对异构性和不均衡性的网络分析具有更好的鲁棒性。

4　结果和讨论

4.1　菌群网络模块与环境因子的相关性分析

4.1.1 菌群网络模块的划分和菌群与培养时间的关联

本文对菌群网络划分使用谱聚类。谱聚类的最优分类数计算通过控制分类数大小，找到合适的分类模块大小，尽可能使与环境变量相关性一致性强的 OTU 被分到一个类别内。具体做法为：将 OTU 按照与环境变量相关性大小进行排序，取相关性较强的 OTU 作为“种子”。这里将与环境相关性 $|r|>0.3$ 的 OTU 定为“种子”。如图 2 中红色部分 OTU 所示，观察聚类结果中“种子”单元的分布，再选择合适的分类数，使“种子”单元聚集在一个或少数几个模块里。如图 2 中的模块 5 和模块 7，在其他模块基本不出现，使得非种子单元主要分布在其他模块，少量在种子模块。最终针对本实验所用数据，当分类数为 8 时划分更好，存在种子单元模块。故本文谱聚类的分类数都定为 8。

本文针对菌群 OTU 在各个样本下的丰度数据集分别构建一阶网络和高阶嵌入网络。其中，1st-order 代表一阶网络的构建结果；DeepWalk 嵌入网络、Node2vec 嵌入网络和 LINE 嵌入网络分别代表 3 种算法对应的高阶嵌入网络。针对上述 4 种网络进行谱聚类，谱聚类的结果包含了每一个 OTU 的分类信息，其中相同类别的 OTU 构成了一个模块，最终每种网络得到 8 个模块。

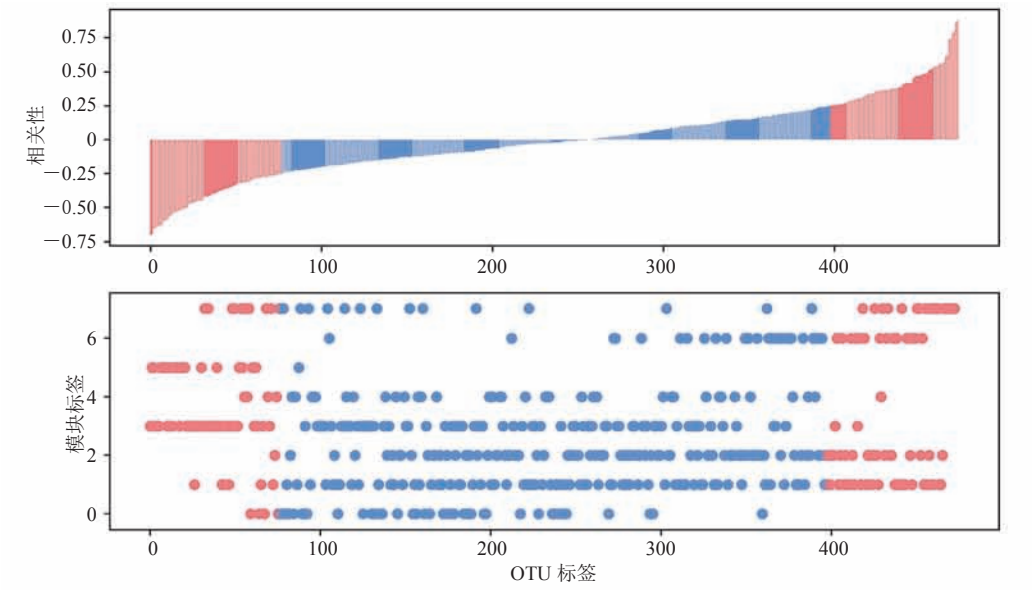

图 2　OTU 与环境因子相关性及对 OTU 谱聚类后在模块中的分布图

Fig. 2　OTU correlation with environment and distribution of OTU in spectral clustering modules

进一步将模块与培养时间进行相关性分析，查看模块内菌群丰度和与时间的关联情况，绘制如图3所示的相关性图。

图3为4种网络方法得到的总共32个模块结果(按照模块与培养时间的相关性高低进行排序)。其中，小提琴图代表模块内的OTU与培养时间的相关性，每个小提琴图对应的三角标号的纵坐标大小是网络模块整体与培养时间的相关性。

如图3所示，菌群网络模块整体比OTU具有更好的相关性结果。在4种网络方法的32个模块中，有28个模块的相关性分析数值大于模块内OTU相关性分析数值，其余4个模块相关性与OTU相关性的平均值无显著差异。这说明在评估菌群与培养时间相关性上，网络模块整体分析方法无论是一阶还是二阶都比OTU直接计算的相关性整体均值效果更好。此外，OTU相关性矩阵数值处于弱相关性区间($|r|<0.3$)内；模块散点值处于强相关性区间($|r|>0.6$)，个别模块整体相关性的绝对值超过OTU极值的绝对值。大部分OTU与时间呈现弱相关，而网络模块整体与时间呈强相关。这表明肠道菌群的共存网络之间具有更紧密的聚集程度，网络模块分析具有明显提升菌群和环境因子相关分析的作用，有效地降低了肠道菌群“功能冗余”对生态位和共存模式的干扰，有利于挖掘肠道微生物结构随时间演替过程中的生物学意义。

4.1.2 关键代谢物与培养时间的关联

代谢物作为肠道微生物和宿主的中介，其含量水平可以影响宿主健康。有效揭示代谢物与菌群相关性和代谢物与环境的相关性是探究肠道微生物与宿主环境的潜在驱动机制的先决条件。本文数据来源的论文[16]检测了超过300种代谢物浓度在培养周期中的变化情况，本文进一步将关键代谢物浓度与培养时间进行相关性分析，查看58种代谢物与时间的相关性，绘制结果见图4。

图4中列出了58种代谢物与时间的相关性分布，呈现出正相关、负相关与不相关三种类群。表明培养周期内存在典型的物质循环的动态过程。其中，乙酰水杨酸(acetylsalicin)、二氢槲皮素(dihydroquercetin)和二氢咖啡酸(dihydrocaffeic acid)等代谢物与时间的耦合性最强，暗示了强烈的生物合成或降解作用。通过代谢物随时间变化的特征分析，可以有效地筛选具有生物学意义的代谢物，为后续相关性分析提供重要参考价值。

4.2 关键代谢物与菌群模块相关性分析

4.2.1 代谢物与OTU及菌群模块关联分析

本文使用58种代谢物的丰度数据与所有OTU的丰度数据，将两组数据进行相关性分

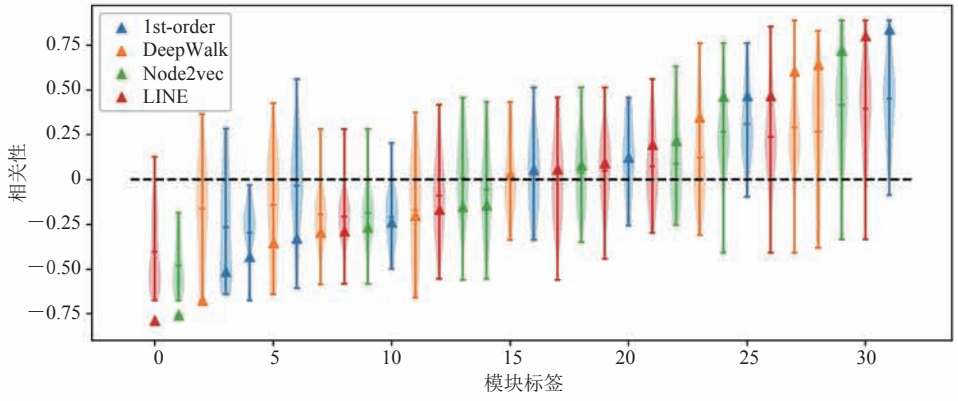

图3 OTU与时间相关性分析及模块与时间相关性分析结果对比

Fig. 3 Comparision of OTU correlation with time and module correlation with time

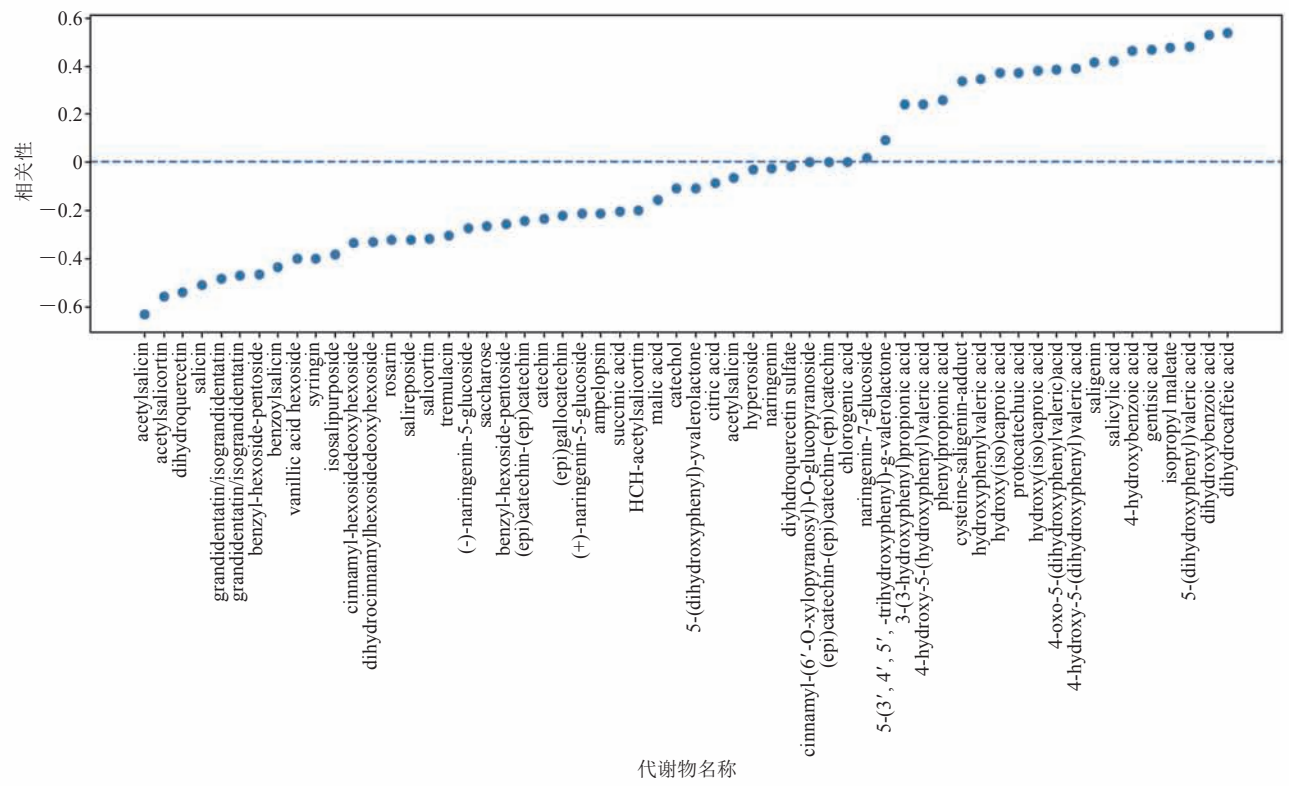

图 4　主要代谢物和培养时间相关性分布

Fig. 4　Major metabolite and culture time correlation distribution

析，得到 OTU 与代谢物的相关性矩阵和模块与代谢物的相关性矩阵。

在 4.1.2 节结果的基础上，选取与时间负相关性最大、相关性最接近 0 和正相关性最大的 3 种代谢物，分别计算其与 OTU、模块的相关性矩阵，绘制结果如图 5 所示。

小提琴图是一个模块内的 OTU 与代谢物丰度的相关性分布图，小提琴图对应的三角形标号的纵坐标值是网络模块整体和代谢物丰度的相关性值。图中展示了 4 种网络方法得到的总共 32 个模块结果(按照模块相关性大小排序)。

如图 5 所示，网络模块整体与代谢物的相关性分析结果优于 OTU 与代谢物的相关性分析。在强正相关、弱正相关、强负相关、弱负相关和不相关五种代谢物分布模式下，网络模块的方法均可以显著提高菌群和代谢物相关性的挖掘。研究表明，肠道代谢物的合成/降解过程通常由多菌株分工、分步骤实现。与传统一阶网络方法相比，高阶嵌入网络的方法可以有效降低“非功能菌群”的内在异质性对参与代谢物通路的“功能菌群”的干扰，不仅凸显出网络中单个模块和节点在代谢过程中的重要性，更能清晰地展示菌群与代谢物网络的层次。

4.2.2　一阶网络和二阶网络部分模块杰卡德相似度分析

一阶网络方法主要关注两个菌群节点之间的直接关联；高阶网络方法除了关注菌群节点之间的直接关联外，同时还关注具有隐关系的菌群节点，即不直接相连但有相似的邻居的菌群节点。为了比较两种方法得到的聚类模块在组成上的异同，进而研究一阶方法和高阶方法得到的模块组成之间的关系，本文选择计算一阶网络模块和高阶网络模块之间的杰卡德(Jaccard)相似度。四种网络方法共得到 32 个模块，选取部分一致性显

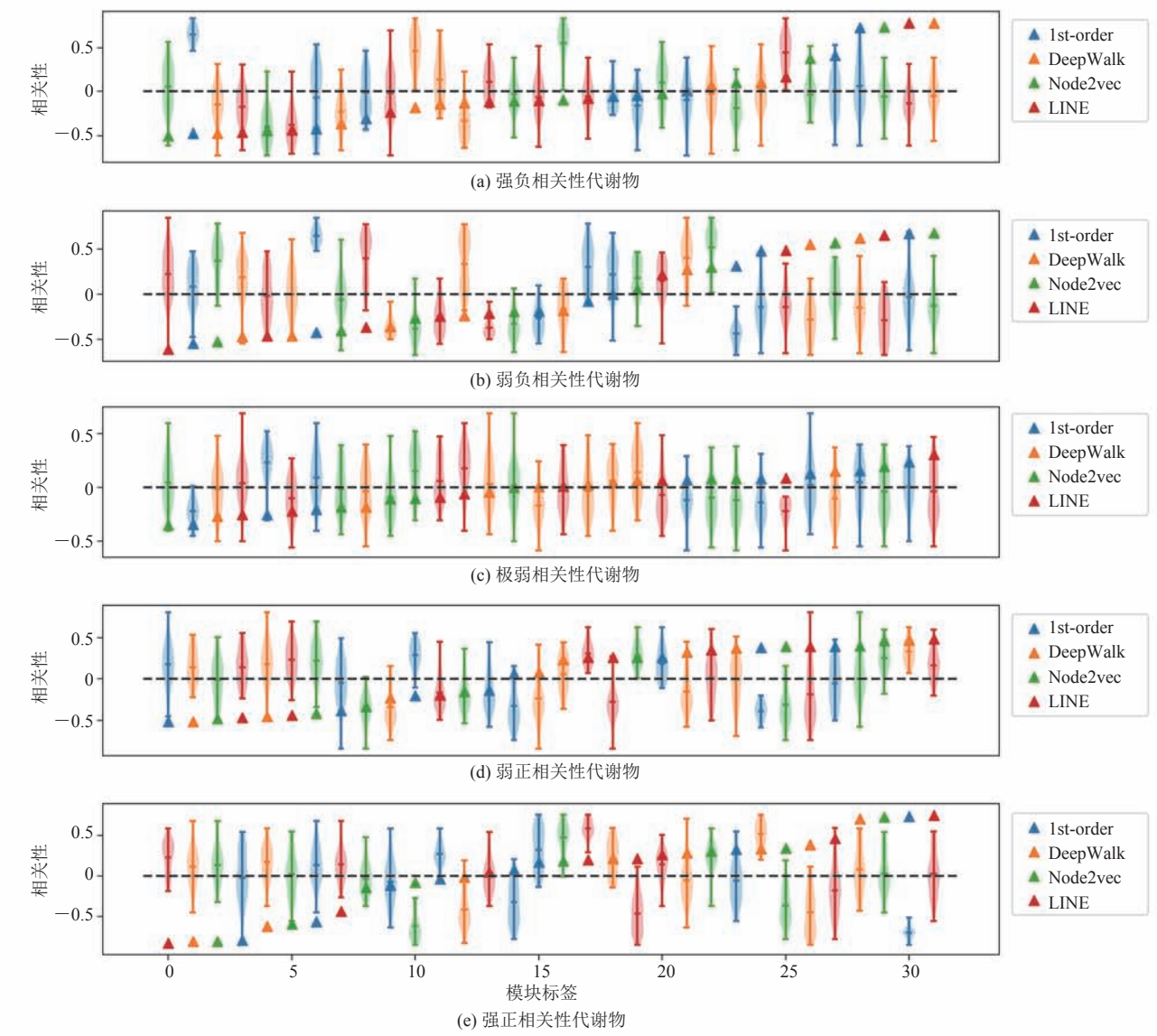

图 5 OTU 与代谢物相关性分析及模块与代谢物相关性分析结果对比

Fig. 5 Comparision of OTU correlation with metabolite and module correlation with metabolite

著的模块。这里选取与 4.2.1 节中代谢物相关性绝对值大于 0.45 的网络模块，查看这些模块之间的 Jaccard 相似度大小情况。以下是一阶网络方法和其他 3 个高阶方法作 Jaccard 相似性得到的结果。计算发现，极弱代谢物和模块的相关性绝对值不超过 0.45，这里仅列出强相关性代谢物的网络模块的结果。

(1) 强负相关性

图 6 标出了相关性图中每种网络方法的模块标签，用于表 1 中 Jaccard 系数比较的对照。

从表 1 中可以看出：①该 Jaccard 系数矩阵中，一阶方法的 3 号模块和 Node2vec 的 3 号模块 Jaccard 值最大为 0.595 4，重合度较高。但从表 2 可知，两个模块与强负相关性代谢物的相关性分别为 0.732 9 和−0.501 6，呈现相反的相关性数值。这说明两个模块重合度虽较高，但相关性一致性相反，与代谢物响应上存在一定差异。②其余的 Jaccard 系数均小于 0.5，模

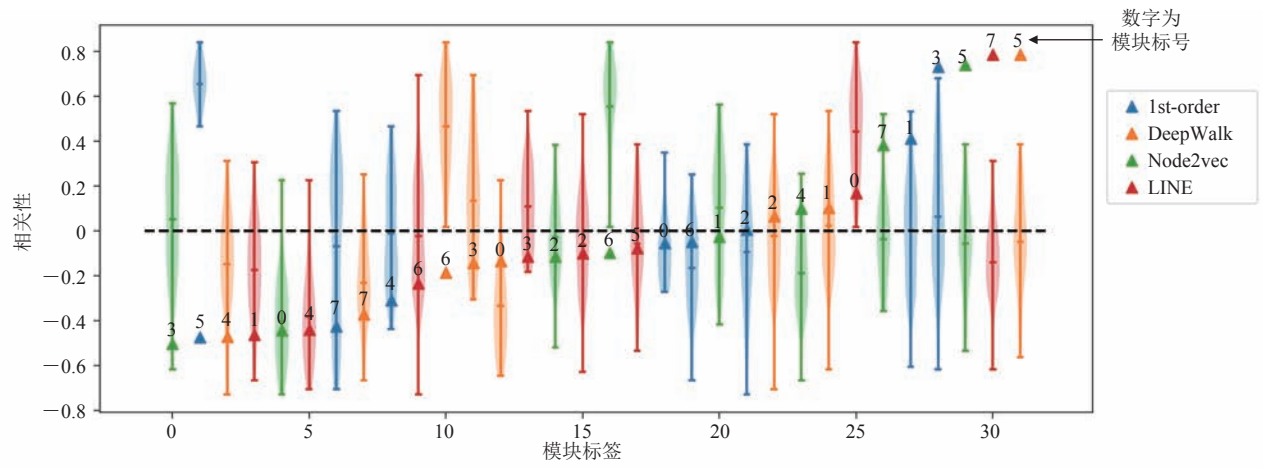

图 6　OTU 和模块与强负相关性代谢物相关性分析图

Fig. 6　OTU and module correlation analysis with strong negative correlation metabolites

表 1　一阶与二阶网络模块 Jaccard 系数矩阵

Table 1　Jaccard coefficient matrix of first-order and second-order network module

	Jaccard 值					
	DeepWlak 4 号	DeepWlak 5 号	Node2vec 3 号	Node2vec 5 号	LINE 1 号	LINE 7 号
一阶网络 3 号	0.287 1	0.022 2	0.595 4	0	0.005 9	0.410 3
一阶网络 5 号	0	0	0	0	0	0

表 2　网络模块与强负相关性代谢物相关性

Table 2　Correlation of network modules and strong negative correlation metabolites

	一阶网络方法		DeepWlak		Node2vec		LINE	
	3 号模块	5 号模块	4 号模块	5 号模块	3 号模块	5 号模块	1 号模块	7 号模块
强负相关代谢物	0.732 9***	−0.472 6	−0.471 8	−0.785 9***	−0.501 6	0.741 4***	−0.463 4	0.785 5***

注：*** $P<0.001$

块重合度较低，说明不同网络的相关性较大，模块之间呈现互补关系。三种二阶网络在一阶网络的基础上进行了补充，有助于寻找一阶网络中潜在的隐关系。

(2)强正相关性

图 7 标出了相关性图中每种网络方法的模块标签，用于表 3 中 Jaccard 系数比较的对照。

从表 3 可以看出：①该 Jaccard 系数矩阵最大值为 0.679 2，结合表 4 可知，对应的一阶网络 2 号模块相关性为−0.564 3，DeepWalk 网络 5 号模块相关性为−0.809 3，两个模块重合度较高。但是 DeepWalk 计算出的相关性绝对值明显大于一阶网络的结果，说明 DeepWalk 的模块组成更优。类似地，一阶网络 2 号模块与 Node2vec 网络 5 号模块 Jaccard 系数为 0.537 6，两者相关性分别为−0.564 3 和−0.806 6，Node2vec 模块相关性绝对值更大；一阶网络 3 号模块和 LINE 方法 7 号模块 Jaccard 系数为 0.410 2，两者相关性分别为−0.790 8 和−0.825 3，LINE 模块的相关性绝对值更大，两个数值也验证了在重合度较高的模块上，二阶网络比一阶网络表现更好。②其余的 Jaccard 数值低于 0.2，一阶方法 5 号模块与代谢物相关性较大为 0.737 3，但在其他二阶网络相关性较大的模块中找不

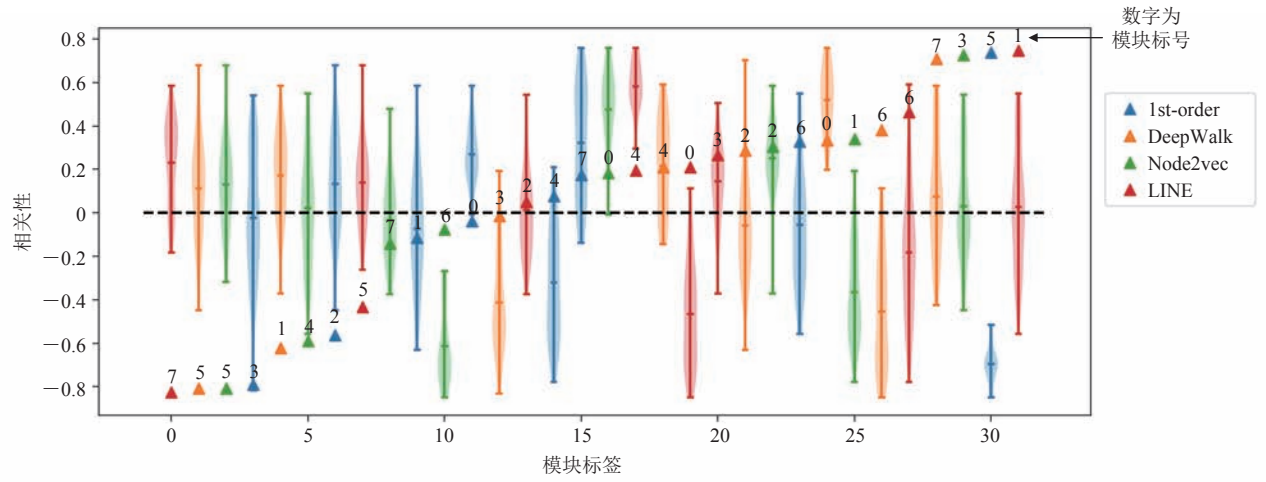

图 7 OTU 和模块与强正相关性代谢物相关性分析图

Fig. 7 OTU and module correlation analysis with strong positive correlation metabolites

表 3 一阶与二阶网络模块 Jaccard 系数矩阵

Table 3 Jaccard coefficient matrix of first-order and second-order network module

	Jaccard 值							
	DeepWlak 1 号	DeepWlak 5 号	DeepWlak 7 号	Node2vec 3 号	Node2vec 4 号	Node2vec 5 号	LINE 1 号	LINE 7 号
一阶网络 2 号	0.005 6	0.679 2	0.081 0	0.079 8	0.122 3	0.537 6	0.179 9	0.006 3
一阶网络 3 号	0.187 1	0.022 2	0	0.595 4	0	0	0.005 9	0.410 2
一阶网络 5 号	0	0	0	0	0	0	0	0

表 4 网络模块与强正相关性代谢物相关性表

Table 4 Correlation of network modules and strong positive correlation metabolites

	一阶网络方法			DeepWlak		
	2 号模块	3 号模块	5 号模块	1 号模块	5 号模块	7 号模块
强正相关代谢物	−0.564 3*	−0.790 8***	0.737 3***	−0.621 7***	−0.809 3***	0.708 1***
	Node2vec			LINE		
	3 号模块	4 号模块	5 号模块	1 号模块	7 号模块	
强正相关代谢物	0.726 4***	− 0.588 9*	− 0.806 6***	0.745 3***	− 0.825 3***	

注：*** $P<0.001$，* $P<0.05$

到与之重合的，说明该模块和其他二阶网络模块具有互补性。

综合以上结果可以看出，一阶网络方法和二阶网络方法在 Jaccard 相似度较大的模块上，后者的相关性绝对值大于前者，说明二阶网络在模块组成上比一阶网络更优。同时，一阶网络方法和二阶网络方法能够分别找到和彼此不重合但一致性高的模块，部分相关性较大的模块之间存在互补关系，二阶网络能在一阶网络的基础上找到其中的隐关系。

4.3 使用菌群网络模块特征对代谢组进行聚类分析

本文得到的网络模块可以作为一种新的生物标志物，可对环境中的代谢物作聚类。具体地，计算出菌群模块与关键代谢物的相关性矩阵，以模块内 OTU 丰度和作为特征，对选取的代谢物集合进行聚类。分别选取菌群模块和 OTU 菌群作为特征对相关性矩阵进行聚类，比较两种特征提取方法对代谢物聚类结果的影响。其中，采用

轮廓系数作为评估聚类质量的指标。

代谢物集合选择以下 4 种集合：(1)58 种代谢物；(2)58 中代谢物中具有共代谢特征的 36 种代谢物；(3)另一类具有共代谢特征的 22 种代谢物；(4)58 种代谢物加上随时间变化最显著的代谢物共 118 种代谢物。特征集合选择以下 7 种集合：(1)所有单个 OTU；(2)一阶网络模块；(3)DeepWalk 嵌入网络模块；(4)Node2vec 嵌入网络模块；(5)LINE 嵌入网络模块；(6)三个高阶嵌入网络模块集合，记为二阶网络集合；(7)一阶加三个高阶嵌入网络模块集合，记为一阶和二阶网络集合。

本文对 4 种代谢物集合分别在以上 7 种特征集合下进行 *K*-means 聚类。在聚类依次选取 2～16 类下，计算轮廓系数并绘制轮廓系数曲线图，结果如图 8 所示。框图中的一条曲线表示基于 7 种特征集合的一种特征集合，对框图对应的代谢物集合聚类得到的轮廓系数值随聚类数变化的情况。为了研究网络模块和 OTU 分别对应的轮廓系数(Silhouette Coefficient，SC)差值随聚类数的变化情况。计算每一个聚类数(2～16 类)下，对应的 6 种网络模块轮廓曲线的 SC 值减去 OTU 对应的 SC 值的差值，通过计算 6 个差值的平均值，来绘制网络模块和 OTU 轮廓系数差值随聚类数变化的曲线，结果如图 9 所示。

如图 9 所示，4 种代谢物集合中，网络模块的 SC 曲线上的轮廓值都大于 OTU 的结果，说明网络模块对代谢物集合聚类效果优于 OTU 作为特征集合，网络模块能够更好地对代谢物集合进行归类。同时，从图 9 可以看出，22 种和 36 种代谢物的 SC 曲线趋势不同，前者网络模块和 OTU 的 SC 平均差值随聚类数的增加而增加，后

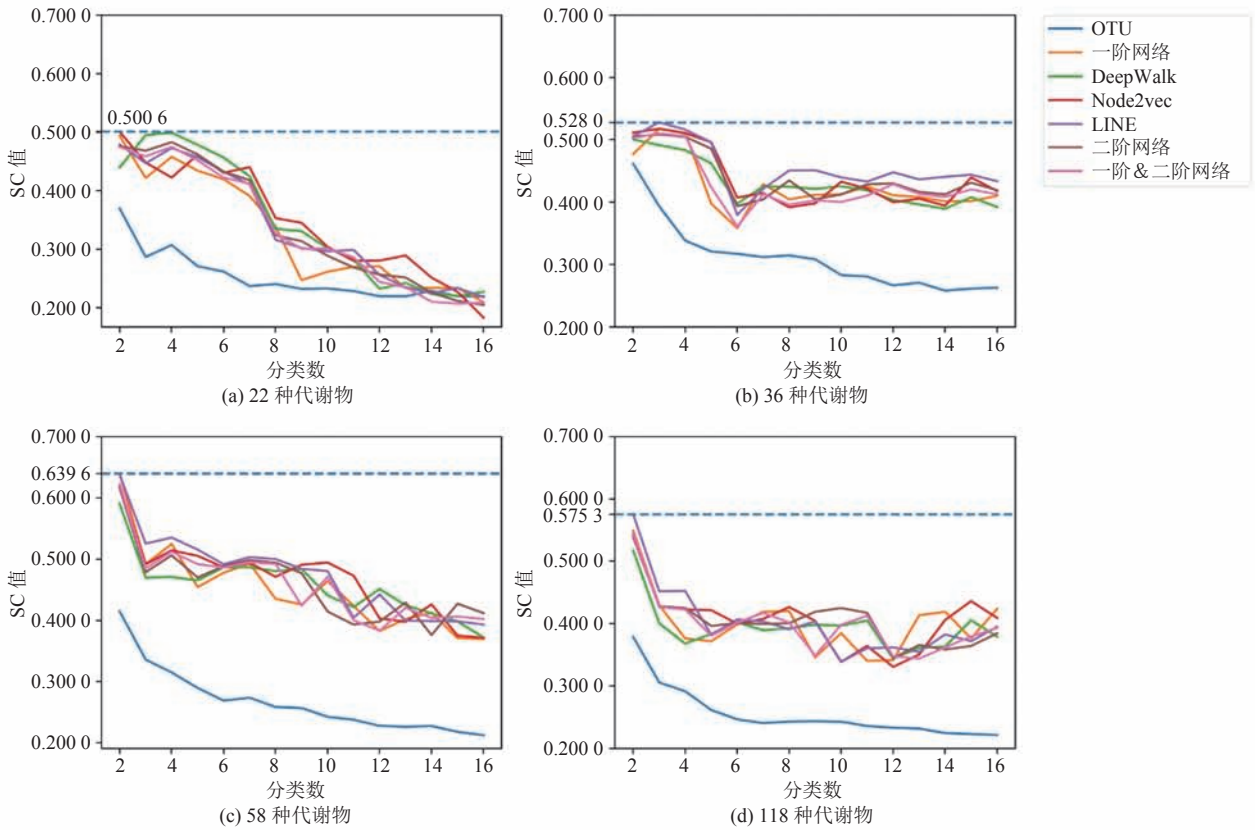

图 8　轮廓系数曲线

Fig. 8　Silhouette coefficient curve

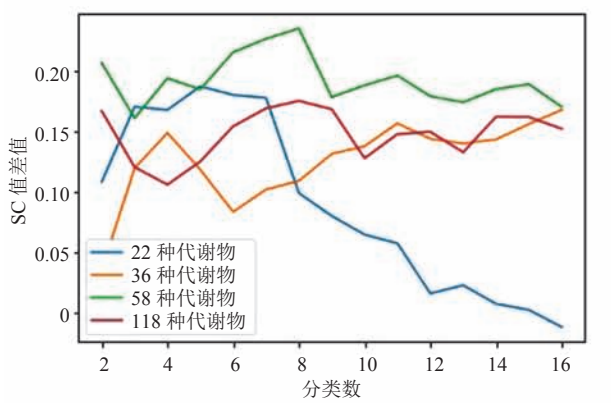

图 9 网络模块和 OTU 轮廓系数差值曲线

Fig. 9 Silhouette coefficient difference curve of network module and OTU

者随聚类数的增加而减少，并且部分曲线 SC 最大值取在聚类为 3 或 4 上。从图 10 可以看出，大部分 22 类代谢物与时间成正相关，而大部分 36 类代谢物与时间成负相关性关系。

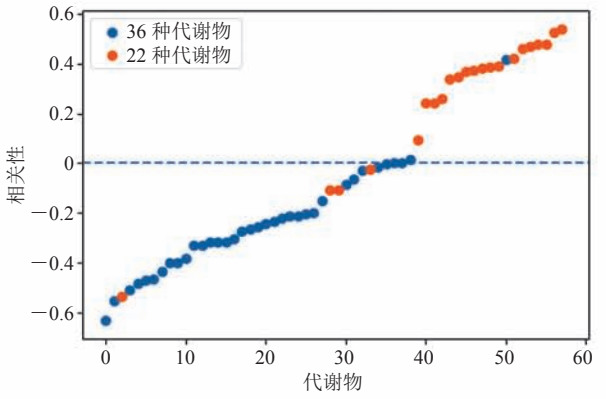

图 10 58 种代谢物两分类中相关性分布

Fig. 10 Correlation distribution in two classifications of 58 metabolites

原 58 类种代谢物中，可以分成图 10 中的两类。图 10 中的两类 SC 值最大，但在这两类中继续划分，细分为 3 或 4 类的最大 SC 值比两类的更大，且与最大 SC 值对应的方法是网络模块方法。这说明网络模块能够发现内在联系紧密的代谢物集合，在隐变量的洞察上比单个 OTU 更强。在 58 种代谢物的基础上，加入其他随时间变化显著的代谢物，获得一个含 118 种代谢物的新集合。该集合显示，网络模块和 OTU 的 SC 平均差值随聚类数改变波动不大。扩大代谢物集合后，网络模块对代谢物集合聚类的效果比 OTU 作为特征的聚类好。这说明应用到更大的代谢物集合上，网络模块也能够更好地对代谢物进行归类。4 种代谢物集合上，不同网络方法之间以及网络方法的联合 SC 曲线差异不大。但是在 58 种代谢物集合和 118 种代谢物集合上，聚类数为 2～5 时，LINE 方法略优于其他网络方法，也略优于网络方法的集合。这说明网络方法之间在对代谢物归类上差异不显著，使用一阶网络方法和二阶网络方法效果相差不大，但在代谢物集合较大、聚类数不超过 5 时，LINE 方法效果较好一些。后续需要使用网络方法得到的模块对代谢物集合聚类时，一阶和二阶网络方法都可以使用。

5 与国内外相似研究的对比分析

现阶段对肠道微生物功能研究，主要集中在与宿主的关联中[23]，Vasudevan 等[24]、Monika 等[25]引入宏基因组学，Kordalewska 和 Markuszewski[24]引入代谢组学分子研究方法，探究肠道微生物群对各种心血管疾病、肠易激综合征、糖尿病和癌症的影响，以及 Dutton 和 Turnbaugh[27]研究将人类营养健康和肠道微生物群的代谢功能关联。而现阶段网络嵌入方法才刚开始应用到生物医学数据分析，还没有进行彻底的探索[28]，但在很多生物网络分析中都有应用。药物数据分析中，Zong 等[29]使用 DeepWalk 算法在药物扩散张量成像网络中预测网络节点；多组学数据分析中，Cho 等[30]使用基于重启随机游走算法的扩散成分分析模型找到生物网络节点低维度表示；临床数据分析中，Wang 等[31]使用一种扩展 TransR 和 LINE 的方法在医学知识网络中向患者推荐适当的药物。与上述肠道微

生物研究不同的是，本文重点关注肠道菌群内部之间的关联，构建肠道菌群高阶网络，并引入网络嵌入方法，找到肠道菌群网络中的潜在功能模块。当然，本文方法还存在不足之处，未来工作重点是探究与验证模块的具体代谢功能。

6 结　论

本文使用网络方法对菌群及其代谢物进行分析，探究菌群模块分析的可行性和功能性，同时提出了使用网络嵌入技术应用到肠道微生物来解决菌群数据中可能存在的异构性、隐关系、隐变量和不均衡性等问题。首先，应用肠道微生物数据构建生物网络并对网络进行聚类得到肠道菌群模块，再将其与环境因子关联分析发现，模块化分析比单独分析肠道微生物效果更好，证明了菌群模块的适用性。然后，将菌群模块与其代谢物做关联分析，进而证明菌群模块具有功能一致性，并且发现在模块重合度较高时，二阶网络方法得到的模块与代谢物相关性比一阶网络方法更强。同时，部分相关性较大的模块之间有着互补的关系，二阶网络能在一阶网络的基础上找到其中的隐关系。最后，将模块作为特征，对选取的代谢物集合进行聚类，发现网络模块作为特征比单个 OTU 作为特征对代谢物集合具有更好的聚类效果，表明网络模块与代谢物的关联分析能够更好地对代谢物进行归类，并且在代谢物集合样本数增大时仍能得到这个结果。本文强调了网络理论应用到生物数据的可能性和重要性。在未来的研究中，将进一步研究生物网络结构，优化生物网络学习策略从中挖掘出菌群网络模块更多的信息，增强该方法在其他生物网络群落应用中的可移植性。

参 考 文 献

[1] Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography [J]. Nature, 2012, 486(7402): 222-227.

[2] Shen J, Obin MS, Zhao L. The gut microbiota, obesity and insulin resistance [J]. Molecular Aspects of Medicine, 2013, 34(1): 39-58.

[3] Turnbaugh PJ, Ley RE, Mahowald MA, et al. An obesity-associated gut microbiome with increased capacity for energy harvest [J]. Nature, 2006, 444(7122): 1027-1031.

[4] Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting [J]. Nature Methods, 2009, 6(4): 283-289.

[5] Muyzer G, Dewaal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA [J]. Applied and Environmental Microbiology, 1993, 59(3): 695-700.

[6] Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network perspective [J]. Trends in Microbiology, 2017, 25(3): 217-228.

[7] Steele JA, Countway PD, Xia L, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages [J]. The ISME Journal, 2011, 5(9): 1414-1425.

[8] Pedersen HK, Forslund SK, Gudmundsdottir V, et al. A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links [J]. Nature Protocols, 2018, 13(12): 2781.

[9] Faust K, Raes J. Microbial interactions: from networks to models [J]. Nature Reviews Microbiology, 2012, 10(8): 538-546.

[10] De Smet R, Marchal K. Advantages and limitations of current network inference methods [J]. Nature Reviews Microbiology, 2010, 8(10): 717-729.

[11] Veiga DFT, Dutta B, Balázsi G. Network inference and network response identification: moving genome-scale data to the next level of biological discovery [J]. Molecular BioSystems, 2010, 6(3): 469-480.

[12] Bonneau R, Facciotti MT, Reiss DJ, et al. A

predictive model for transcriptional control of physiology in a free living cell [J]. Cell, 2007, 131(7): 1354-1365.

[13] Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored [J]. Nucleic Acids Research, 2011, 39: D561-D568.

[14] Milns I, Smith CMBA. Revealing ecological networks using Bayesian network inference algorithms [J]. Ecology, 2010, 91(7): 1892-1899.

[15] Peng C, Xiao W, Jian P, et al. A survey on network embedding [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, doi: 10.1109/TKDE.2018.2849727.

[16] Eva-Maria PW, Kaisa K, Christine ME, et al. A combined LC-MS metabolomics- and 16S rRNA sequencing platform to assess interactions between herbal medicinal products and human gut bacteria *in vitro*: a pilot study on willow bark extract [J]. Frontiers in Pharmacology, 2017, 8: 893-906.

[17] Luxburg UV. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.

[18] Ahmed A, Shervashidze N, Narayanamurthy S, et al. Distributed large-scale natural graph factorization [C] // Proceedings of the 22nd International Conference on World Wide Web, 2013.

[19] Tang J, Qu M, Wang M, et al. LINE: large-scale information network embedding [C] // Proceedings of the 24th International Conference on World Wide Web, 2015.

[20] Wang D, Peng C, Zhu W. Structural deep network embedding [C] // ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2016.

[21] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations [J]. 2014, doi: 10.1145/2623330.2623732.

[22] Grover A, Leskovec J. Node2vec: scalable feature learning for networks [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 855-864.

[23] Yadav M, Verma MK, Chauhan NS. A review of metabolic potential of human gut microbiome in human nutrition [J]. Archives of Microbiology, 2017, 200(2): 203-217.

[24] Vasudevan D, Andiappan R, Muthuirulan P, et al. Elevated levels of circulating DNA in cardiovascular disease patients: metagenomic profiling of microbiome in the circulation [J]. PLoS One, 2014, 9(8): e105221.

[25] Monika MKV, Ahmed V, Chauhan NS. Human gut microbiome: an imperative element for human survival [J]. Microbiology and Molecular Biology Reviews, 2016, 68(4): 686-691.

[26] Kordalewska M, Markuszewski MJ. Metabolomics in cardiovascular diseases [J]. Journal of Pharmaceutical and Biomedical Analysis, 2015, 113: 121-136.

[27] Dutton RJ, Turnbaugh PJ. Taking a metagenomic view of human nutrition [J]. Current Opinion in Clinical Nutrition and Metabolic Care, 2012, 15(5): 448-454.

[28] Su C, Tong J, Zhu Y, et al. Network embedding in biomedical data science [J]. Briefings in Bioinformatics, 2018, 5(3): 1-16.

[29] Zong N, Kim H, Ngo V, et al. Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations [J]. Bioinformatics, 2017, 33(15): 2337-2344.

[30] Cho H, Berger B, Jian P. Diffusion component analysis: unraveling functional topology in biological networks [C] // International Conference on Research in Computational Molecular Biology, 2015.

[31] Wang M, Liu M, Liu J, et al. Safe medicine recommendation via medical knowledge graph embedding [J]. arXiv: 1710.05980, 2017.