**原核微生物群落随机性和确定性装配过程的计算方法**

**Calculation Method for Stochastic and Deterministic Assembly Processes of Prokaryotic Communities**

赵维1, 2，王兴彪1, 2，侍浏洋1, 2，朱婉瑜1, 2，马磊1, 2，王敬敬1, 2，徐松1, 2，杨榕1, 2，张小霞1, 2，韩一凡1, 2，黄志勇1, 2, \*

1中国科学院天津工业生物技术研究所 天津市工业生物系统与过程工程重点实验室，天津；2国家合成生物技术创新中心，天津

\*通讯作者邮箱: huang\_zy@tib.cas.cn

**摘要：**原理：微生物群落生态学的一个主要目标是了解构成跨时空物种丰度模式的过程(Stegen等，2012；Zhou等，2014；Dini-Andreote等，2015)。确定性和随机性两种类型的过程会影响群落的聚集。确定性过程(其中非生物和生物因素决定物种的存在与否和相对丰度)与生态选择相关(Vellend，2010)，随机过程包括不可预测的扰动、概率性的散布和随机的出生-死亡事件等，这些变化不是由环境决定的适应性结果(Chase和Myers，2011)。目的：判定群落间变化由确定性过程主导还是随机性过程主导。方法：使用零模型量化群落的绝对系统发育距离与随机系统发育距离的偏离度，偏离程度越大，群落受确定性因素的影响越大，偏离度越小，群落受随机性因素的影响越大。结果：该方法可以使用NTI指示单个群落内共存的分类单元相比偶然预期的关系更为紧密还是分散，使用*β*NTI指示两两群落间的变化受确定性或随机性因素影响的大小。结论：该方法可用以评估不同时空尺度下随机性和确定性过程对微生物群落组装的影响。

**关键词：**随机性过程，确定性过程，群落组装，群落周转，时空变化

**软件**

1. R version 3.4.3

**实验步骤**

1. 参照《Extraction and 16S rRNA Sequence Analysis of Microbiomes Associated with Rice Roots》(DOI:10.21769/BioProtoc.2884) (Edwards等，2018)中的方法对微生物群落进行16S rRNA序列分析获得样品的群落组成和系统发育关系树，对应的两个文件分别为OTU表格out\_table.txt，和树文件rep\_set.tre。
2. 系统发育群落组成的计算参数

为了表征每个样本内的系统发育群落组成(空间和时间上的唯一点)，使用“mntd”和‘ses.mntd’对平均最近分类距离(MNTD)和最近分类指数(NTI)进行量化(Webb等，2002)。在R软件包“picante”中使用“mntd”和“ses.mntd”能完成上述两个指数的计算。

* 1. MNTD

MNTD是mean-nearest-taxon-distance(平均最近分类距离)的缩写，MNTD可以查找样本中每个OTU之间的系统发育距离及与其最近亲属之间的系统发育距离，然后在这些系统发育距离上获取丰度加权平均值。MNTD计算一个群落内OTU之间的系统发育距离，可以提供有关群落内OTU之间生态(生态位)相似程度的信息。计算公式如下：

MNTD=min()

*f*ik：群落k中OTUi的相对丰度；

nk：群落k中OTU的数量；

∆ikjk：OTUi与同样在群落k中的所有其他OTUj之间的最小系统发育距离。

* 1. NTI

NTI是nearest-taxon-index（最近分类单元指数）的缩写，是对样本中每个分类单元到最近分类单元的系统发育距离的标准化度量，量化终端聚类的程度。NTI是“ses.mntd”输出的相反数。为了评估系统发育群落结构的非随机程度，将OTU及其相对丰度在系统发育的各个尖端进行随机分配(null.model=‘taxa.labels’in‘ses.mntd’)。NTI量化观察到的MNTD与零分布均值的标准偏差数(999个随机)。计算公式如下：

NTI=(- MNTDobs)/sd(MNTDnull)

对于单个群落，NTI大于+2表示共存的分类单元比偶然预期的关系更为紧密(系统发生聚类)；NTI小于-2表示共存的分类单元比偶然预期的亲缘关系更远(系统发生过度分散)。在所有的群落中获得的平均NTI与零值有显著差异，表示存在平均聚类(NTI>0)或过度分散(NTI<0)(Kembel，2009)。

* 1. *β*MNTD

系统发育随时间或空间的变化(系统发育β多样性)使用Beta平均最近分类距离(*β*MNTD)和Beta最近分类指数(*β*NTI)来量化，它们分别是MNTD和NTI的组间(between-assemblage)相似物(Webb等，2009；Kembel等，2011)。*β*MNTD是发生在两个不同群落中的最近亲属间的丰度加权平均系统发育距离。这种度量量化了群落装配之间的系统发育距离; *β*MNTD被使用是因为它强调短的系统发育距离的系统发育周转率,其中大力支持系统发育信号的假设。其计算公式如下：

*β*MNTD=0.5[min()+min()]

min()：群落k中的OTUi与群落m中所有OTUj之间的最小系统发育距离 (其他变量请参见上文)。

* 1. *β*NTI

为了推断出对特定群落装配之间观察到的周转率起主要作用的生态因素，鉴于随机性生态过程的主导地位，使用零模型方法生成一个预期水平的*β*MNTD。为了量化观察到的*β*MNTD值和零*β*MNTD值分布之间偏差的大小和方向，使用了β-最邻近分类指数(*β*NTI)。通过将整个系统发育过程中的OTU随机化并重新计算*β*MNTD 999次，可以得到零*β*MNTD分布。*β*NTI是观察到的*β*MNTD与零*β*MNTD分布均值的标准偏差数。

*β*NTI=(*β*MNTDobs-)/sd(*β*MNTDnull)

对于成对的群落比较，*β*NTI<-2或>+2分别表示小于或大于预期的系统发育周转。在所有成对比较中获得的平均*β*NTI与零有显著差异，表示大于(*β*NTI>0)或小于(*β*NTI<0)预期平均周转(Webb等，2009；Kembel等，2011)。

**结果与分析**

1、R中计算实例

上述参数的计算可以在R中实现，现以两个来自海洋环境的原核微生物群落的数据为例子，计算上述参数。样品分别命名为S1和S2,其OTU种类和数量及其每个OTU之间的系统发育树文件见附件(http://doi.org/10.5281/zenodo.3980071)。OTU文件命名：out\_table.txt；树文件命名：rep\_set.tre。

library(picante)

library(ape)

comm<-read.table("out\_table.txt",sep='\t',row.names=1,header=T)#读入OUT文件

phy<-read.tree("rep\_set.tre")#读入树文件

prune\_tree<-prune.sample(comm,phy)#使树文件和OTU表文件对齐

phydist<-cophenetic(prune\_tree)#计算每个OTU之间距离矩阵

mntd<-ses.mntd(comm,phydist,null.model="taxa.labels",abundance.weighted=T, runs=999)#计算MNTD

该步骤计算结果如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ntaxa | mntd.obs | mntd.rand.mean | mntd.rand.sd | mntd.obs.rank | mntd.obs.z | mntd.obs.p | runs |
| S1 | 498 | 0.13913 | 0.1712452 | 0.02944938 | 108 | -1.090523 | 0.108 | 999 |
| S2 | 586 | 0.1432301 | 0.1591859 | 0.01302431 | 90 | -1.225076 | 0.09 | 999 |

其中，

ntaxa Number of taxa in community，群落中OTU的数目；

mntd.obs Observed MNTD in community，观测到的群落MNTD值；

mntd.rand.mean Mean MNTD in null communities，随机群落的MNTD均值；

mntd.rand.sd Standard deviation of MNTD in null communities，随机群落MNTD值的标准差；

mntd.obs.rank Rank of observed MNTD vs. null communities，观察到的群落的MNTD值与随机群落MNTD值的排序；

mntd.obs.z Standardized effect size of MNTD vs. null communities (=(mntd.obs-mntd.rand.mean)/mntd.rand.sd, equivalent to -NTI) 观察到的群落的MNTD与随机群落的MNTD值的标准化数值大小；

mntd.obs.p P-value (quantile) of observed MNTD vs. null communities (= mntd.obs.rank/runs + 1) 观察到的群落的MNTD与随机群落的MNTD值比较的p值；

runs Number of randomizations，随机化的次数。

comdist.resultS12<-comdistnt(comm,phydist)#计算*β*MNTD值

该步骤计算结果如下：

|  |  |
| --- | --- |
|  | S1 |
| S2 | 0.06070395 |

#以下为*β*MNTDnull的计算过程

f<-function(){

g<-randomizeMatrix(comm,null.model=c("frequency","richness","independentswap","trialswap"),iterations=1000)#针对两组样方实现随机群落构建

fc<-comdist.result<-comdistnt(g,phydist)#针对随机构建的两个样方进行群落间*β*MNTD的计算

}

mt<-mean(replicate(999, f()))#上述f计算过程重复进行999次，得到*β*MNTD的平均值

mt

mt.sd<-sd(replicate(999, f()))#计算上述999次计算得到的*β*MNTD值的标准差

mt.sd

*β*NTI=(comdist.resultS12-mt)/mt.sd#如果样品对较多，该步骤需在计算完每组样品对的*β*MNTD、mt和mt.sd值后，手动计算得出*β*NTI值

最终，得到的*β*NTI值为4.93，说明S1和S2两个样方的群落之间的变化主要由确定性过程影响。

2、多组样品计算结果示例

针对多组样品多个样方的*β*NTI计算按照上述实例1分别计算两两样方之间的*β*NTI值，并按照设计的样方分组分析*β*NTI在不同分组中的变化范围。本例展示来自渤海湾6个不同站位表层海水样品在不同月份（5、8、10月）的*β*NTI统计值。结果如下：



图1 6个站位在不同月份的β-NTI值分布

图1结果显示来自5月份和10月份的海水样品间的β-NTI值均主要集中在-2~2区间内，说明在这两个月份内各站位的微生物群落结构变化主要受随机性因素的影响，8月份的海水样品中有5对样品间的β-NTI值大于2，说明某些站位间的微生物群落结构变化主要受到决定性因素的影响。另外，月份间（5月-8月、5月-10月以及8月-10月）样品的β-NTI值（除了一对样品）均大于2，说明不同站位微生物群落结构随着时间的变化主要受决定性因素的影响。

**失败经验**

1、OTU文件导入R后需要使样品名在行，OTU名在列，如果不是该格式后面命令会提示错误，可使用t()命令进行转置。

2、关于样品采样深度对该方法计算结果的影响，目前尚无系统性的报道，根据已有报道基于该方法进行的研究显示Stegen等(Stegen et al. 2012)研究随机和确定性过程对不同深度的沉积物样品的相对影响时，对样品的序列取样深度为500条；Wang等(Wang et al. 2013)对不同生态系统细菌群落系统发育多样性受随机性和决定性因素影响的研究中，每个样品的序列取样深度为1000条；Yu等(Yu et al. 2018)研究空间尺度对种植小麦土壤中微生物群落随机性和决定性相对作用的影响时，对每个样品的序列取样深度为20,005条。可见，随着测序技术越来越成熟，用于进行群落系统发育分析的取样深度也越来越大。另外，关于物种数对系统发育多样性的影响，以MNTD和NTI为例，其他系统发育多样性指数类似：MNTDobs是观测到的MNTD，MNTDnull是随机打乱进化树之后得到的MNTD值，这个过程重复多次（~999次），可得到MNTDnull的分布。可近似转化为均值为0，标准差为1的标准正态分布。而这个分布其实是所有物种施加影响结果的集合。NTI和*β*NTI计算过程已经排除了物种数量对系统发育的影响。

3、关于使用零模型(null model)方法解释原核微生物群落构建：零模型用于验证基于自然实验的生态学假说，它以计算机技术为基础，从已经获得的实验数据出发，构建受控的生态随机模型，最后用标准的统计检验方法验证理论假说。在解释原核微生物群落构建机制中，零模型是群落数据或者进化树数据进行随机化的方式。运行零模型主要是为了解，在随机状况下，物种系统发育关系的统计分布，从而推断真实的群落内物种的系统发育关系是否随机，以便了解环境过滤或者随机过程的相对重要性。所以在探讨这些生态学过程的相对重要性，计算NTI和*β*NTI的时候，必须要考虑零模型。目前，该方法使用基于randomizeMatrix函数提供的零模型构建主要包括物种内群落数据矩阵丰度(维持物种出现频率)的随机化(frequency)、样本内群落数据矩阵丰度(维持样本物种丰富度)随机化(richness)、利用独立交换算法(Gotelli 2000)对群落数据矩阵进行随机化(保持物种发生频率和样本物种丰富度)(independentswap)、使用试交换算法(Miklos and Podani 2004)随机化群落数据矩阵(保持物种发生频率和样本物种丰富度)(trialswap)。对于生态零模型，Gotelli指出“它是基于生态数据的随机化或者基于已知分布、设想分布的随机样本基础上的格局生成模型。……与已知数据相比，一些数据的性质在零模型中保持不变，另外一些则允许随机变化，这样就产生了一个新的格局总体。”从中可以看出零模型的构建是从实际所得数据出发的，并非真正的实验。对于原核微生物群落而言，环境中的微生物种类只能尽最大的可能通过高通量、高分辨率的分析方法获得，但是仍然无法完全获知其中的全部微生物组成，因此，零模型的构建也只能依据已知样品已获得的物种信息进行预测，对于未能获得的物种信息则尚不能纳入到零模型的构建中。但是，针对样方之间的比较，我们更关注设计样方间的相对变化规律，因此，取样深度的一致性是分析对比的前提。

**致谢**

感谢天津市科技计划项目(18ZXSZSF00100,19YFZCSF00750)；国家自然科学基金项目(41977200)的支持。感谢Campbell O Webb，Steven W Kembel，James C Stegen等人先前的研究工作。

**参考文献**

1. Chase, J. M. and Myers, J. A. (2011). Disentangling the importance of ecological niches from stochastic processes across scales. *Philos Trans R Soc Lond B Biol Sci* 366(1576): 2351-2363.
2. Dini-Andreote, F., Stegen, J. C., van Elsas, J. D. and Salles, J. F. (2015). Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc Natl Acad Sci U S A* 112(11): E1326-1332.
3. Edwards, J., Santos-Medellín, C. and Sundaresan, V. (2018). Extraction and 16S rRNA Sequence Analysis of Microbiomes Associated with Rice Roots. *Bio-protocol* 8(12): e2884.
4. Gotelli, N. J. (2000) Null Model Analysis of Species Co-Occurrence Patterns. *Ecology* 81: 2606-2621.
5. Kembel, S. W. (2009). Disentangling niche and neutral influences on community assembly: assessing the performance of community phylogenetic structure tests. *Ecol Lett* 12(9): 949-960.
6. Kembel, S. W., Eisen, J. A., Pollard, K. S. and Green, J. L. (2011). The phylogenetic diversity of metagenomes. *PLoS One* 6(8): e23214.
7. Miklos, I. & J. Podani (2004) Randomization of presence-absence matrices: Comments and new algorithms. *Ecology* 86-92.
8. Stegen, J. C., Lin, X., Konopka, A. E. and Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J* 6(9): 1653-1664.
9. Vellend, M. (2010). Conceptual synthesis in community ecology. *Q Rev Biol* 85(2): 183-206.
10. Wang, J., J. Shen, Y. Wu, C. Tu, J. Soininen, J. C. Stegen, J. He, X. Liu, L. Zhang & E. Zhang (2013) Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *Isme J* 7: 1310-21.
11. Webb, C. O., Ackerly, D. D. and Kembel, S. W. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* 24(18): 2098-2100.
12. Webb, C. O., Ackerly, D. D., McPeek, M. A. and Donoghue, M. J. (2002). Phylogenies and community ecology. *Annu Rev Ecol Syst* 33: 475-505.
13. Yu, S., Y. Li, X. Xiang, R. Sun, T. Yang, D. He, K. Zhang, Y. Ni, Y. G. Zhu & J. M. Adams (2018) Spatial scale affects the relative role of stochasticity versus determinism in soil bacterial communities in wheat fields across the North China Plain. *Microbiome* 6: 27.
14. Zhou, J., Deng, Y., Zhang, P., Xue, K., Liang, Y., Van Nostrand, J. D., Yang, Y., He, Z., Wu, L., Stahl, D. A., Hazen, T. C., Tiedje, J. M. and Arkin, A. P. (2014). Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. *Proc Natl Acad Sci U S A* 111(9): E836-845.