bio-101 www.bio-protocol.org/exxxxxxx DOI:10.21769/BioProtoc.xxxxxxxx

Dynamic Meta-Storms 算法：基于物种水平的生物分类学和系统发育信息对宏基因组进行全面比较

张玉凤 [1, 2, #]，荆功超 [2, #]，陈俞竹 [1]，徐健 [2]，苏晓泉 [1, 2, *, $]

[1] 青岛大学，青岛市，山东省；[2] 单细胞中心，中国科学院青岛生物能源与过程研究所，青岛市，山东省；

[$] 现工作单位：计算机科学技术学院，青岛大学，青岛市，山东省

*通讯作者邮箱: suxq@qdu.edu.cn

#共同第一作者/同等贡献

摘要：精确、全面地计算宏基因组样本之间的距离对于理解微生物组的 β 多样性起到至关重要的作用。目前针对宏基因组的距离计算方法往往忽略了物种之间的进化关系，抑或是舍弃掉了无法定位到系统发育树叶子结点的未知物种，从而导致对 β 多样性进行错误地解读。为解决以上问题，我们提出了 Dynamic Meta-Storms（以下简称 DMS）算法，可基于生物分类和系统发育信息，在物种水平上对宏基因组样本进行全面地比较。在计算两个微生物组的距离时，对于其中能够精确注释到物种（species）水平的成分，DMS 算法将这些物种映射到系统发育树的叶子结点上进行比较；而对于只能注释到属或更高水平分类信息的成分，DMS 算法将其动态、合理地放置到系统发育树的虚拟中间节点上，从而最大限度地利用宏基因组的信息来计算样本之间的距离。同时，得益于并行计算技术的优化，DMS 通量大且消耗资源少，在单个计算节点上计算 10 万个宏基因组样本的距离矩阵，仅用 6.4 个小时即可完成。最新版本的 DMS 软件可在 https://github.com/qibebt-bioinfo/dynamic-meta-storms 下载。输入多个宏基因组样本的物种组成后，DMS 便可计算出样本之间距离矩阵(distance matrix)，用于后续的 β 多样性分析。目前 DMS 软件已经内置了与 MetaPhlAn2 兼容的生物分类信息和系统发育树，同时也支持用户自定义的生物分类信息和系统发育树。

关键词: 宏基因组分析工具，Dynamic Meta-Storms (DMS)，生物信息算法，鸟枪宏基因组，距离矩阵，β 多样性

仪器设备

1. DMS 仅需要具有约 2 GB RAM（内存）的标准计算机即可支持用户的操作。为了获得最佳性能，我们建议您使用以下规格的计算机：

   内存：4 GB +

   CPU：4 核+

软件

1. DMS 软件(Jing 等， 2019)可运行于 Linux、Mac OS、Windows 10 内置 Linux 子系统等类 Unix 操作系统中。Dynamic Meta-Storms 依赖于 OpenMP 库来实现并行计算。常见版本的 Linux 系统已经安装了 OpenMP，无需额外配置。在 Mac OS 中，需要安装支持 OpenMP 的编译器，建议使用 homebrew 软件包管理器，通过以下命令来配置 OpenMP 库

   ```
   brew install gcc
   ```

   DMS 软件已经内置了与 MetaPhlAn2(Segata 等， 2012; Truong 等， 2015)兼容的生物分类信息和系统发育树，其中包含了 MetaPhlAn2 的默认数据库中所有的细菌物种。我们建议用 MetaPhlAn2 对宏基因组序列进行预处理，将得到的物种名称和丰度结果作为 DMS 的输入来计算距离矩阵（详见实验步骤 2.1）。DMS 内置的分类信息和系统发育树同样也适用于由 mOTUs、Karken 等其他软件得出来的宏基因组物种信息。同时，DMS 也支持用户自定义的生物分类信息和系统发育树（详见实验步骤 3.2）。

实验步骤

1. 安装 DMS

   我们建议选择步骤 1a 中自动安装的方式来配置 DMS 软件。但如果自动安装程序失败，可以按照步骤 1b 中的步骤手动安装 DMS 软件。

   a. 自动安装（首选方案）

      1) 下载安装包

      ```
      git clone https://github.com/qibebt-bioinfo/dynamic-meta-storms.git
      ```

2) 安装

运行以下安装命令：

```
cd dynamic-meta-storms
source install.sh
```

按照上述步骤操作，该软件包可以在 1 分钟内安装到计算机上。

示例数据集在安装包内“example”文件夹下，可以查看“example/Readme”中的内容来获取演示运行的详细信息，或直接运行：

```
sh Readme
```

来演示示例数据集的距离矩阵的计算。

b. 手动安装（备用方案）

1) 下载安装包

```
git clone https://github.com/qibebt-bioinfo/dynamic-meta-storms.git
```

2) 配置环境变量

将以下内容写入环境变量配置文件（一般默认的文件是“~/.bashrc”）

```
export DynamicMetaStorms="Specify the full path to DMS package here"
export PATH=$PATH:$DynamicMetaStorms/bin/
```

并启用环境变量

```
source ~/.bashrc
```

3) 编译源代码

```
cd dynamic-meta-storms
make
```

2. 从宏基因组序列获得物种丰度信息

a. 用 MetaPhlAn2 获得宏基因组的物种组成及相对丰度信息（profiling）

以单个宏基因组序列文件“sample_1.fasta”为例：

```
metaphlan2.py sample_1.fasta --input_type fasta --tax_lev s --ingore_viruses --ignore_eukaryotes --ignore_archaea > profiled_sample_1.sp.txt
```

得到的输出文件“profiled_sample_1.sp.txt”格式如下（**表 1**）

表 1. 单个宏基因组样本的物种组成及相对丰度信息文件

#Species	Abundance
s__Rothia_aeria	22.78
s__Actinomyces_naeslundii	13.9
s__Lautropia_mirabilis	12.49
s__Corynebacterium_matruchotii	11.27
s__Corynebacterium_durum	10.36
s__Streptococcus_sanguinis	8.13
s__Actinomyces_oris	6.24
s__Actinomyces_massiliensis	5.67
s__Cardiobacterium_hominis	4.83
s__Porphyromonas_sp_oral_taxon_279	4.33

其中，第一列为物种名称，第二列为相对丰度数据。如果已经由其他软件获得了该格式的物种信息或者步骤 2b 中的相对丰度表（如软件自带的示例数据集“example/dataset1.sp.abd”）那么可以忽略这个步骤。但我们强烈建议所有的样本都用相同的软件和参数来处理宏基因组序列。

b. 将多个样本的物种信息文件合并生成物种相对丰度表

将多个步骤 2a 中生成的样本的物种信息文件路径，汇总成一个列表文件（如 samples.list.txt），格式如下（**表 2**）

表 2. 样本的物种信息文件路径列表文件

Sample_1	profiled_sample_1.sp.txt
Sample_2	profiled_sample_2.sp.txt
Sample_3	profiled_sample_3.sp.txt
Sample_4	profiled_sample_4.sp.txt
Sample_5	profiled_sample_5.sp.txt

其中第一列是样本 ID，第二列是物种信息文件的路径。然后运行 DMS 的以下命令：

```
MS-single-to-table -l samples.list.txt -o samples.sp.abd
```

得到的输出文件“sample.sp.abd”格式如下（**表 3**）

表 3. 物种相对丰度表

SampleID	*Sample_1*	*Sample_2*	*Sample_3*
s__Rothia_aeria	*22.78*	*16.32*	*7.65*
s__Actinomyces_naeslundii	*13.9*	*1.74*	*9.32*
s__Lautropia_mirabilis	*12.49*	*21.18*	*5.83*
s__Corynebacterium_matruchotii	*11.27*	*1.22*	*0*
s__Corynebacterium_durum	*10.36*	*7.41*	*11.38*
s__Streptococcus_sanguinis	*8.13*	*14.25*	*5.8*
s__Actinomyces_oris	*6.24*	*17.15*	*18.46*
s__Actinomyces_massiliensis	*5.67*	*18.32*	*17.07*
s__Cardiobacterium_hominis	*4.83*	*2.41*	*10.95*
s__Porphyromonas_sp_oral_taxon_279	*4.33*	*0*	*13.54*

其中第一行是样本 ID，第一列是物种名称，表中的数值是某物种在某样本中的相对丰度。如果已经获得了以上格式的物种相对丰度表（如软件自带的示例数据集“example/dataset1.sp.abd”），那么可以忽略这个步骤。

3. 使用 DMS 计算距离矩阵

a. 基于 DMS 内置系统发育树和物种分类信息计算距离矩阵：

MS-comp-taxa-dynamic -T samples.sp.abd -o samples.sp.dist

其中“-T”参数指定的输入文件“samples.sp.abd”为**表 3** 格式的物种相对丰度表。由“-o”参数指定的输出文件“samples.sp.dist”即为输入的所有样本之间的 DMS 距离矩阵，格式如下（**表 4**）

表 4. 样本之间的 DMS 距离矩阵

	Sample_1	*Sample_2*	*Sample_3*
Sample_1	*0*	*0.071088*	*0.070619*

Sample_2	*0.071088*	*0*	*0.088648*
Sample_3	*0.070619*	*0.088648*	*0*

其中第一行和第一列是样本 ID，表中的数值是样本之间的 DMS 距离，在 0-1 之间。数值越大，表示距离越远。

b. 基于用户自定义的系统发育树和物种分类信息计算距离矩阵

1) 自定义系统发育树和物种分类信息并生成新的 DMS 库

自定义 DMS 库需要 newick 格式的二叉系统发育树（如“tree.newick”），树的叶子结点为物种名称，并提供所有叶子结点物种从并提供从 Kingdom 到 Species 水平的完整的生物分类信息（如“tree.taxonomy”），格式如下（**表 5**）

表 5. 完整的生物分类信息文件

Kingdom	*Phylum*	*Class*	*Order*	*Family*	*Genus*	*Species*
k__Archaea	*p__Euryarchaeota*	*c__Methanopyri*	*o__Methanopyrales*	*f__Methanopyraceae*	*g__Methanopyrus*	*s__Methanopyrus_kandleri*
k__Archaea	*p__Euryarchaeota*	*c__Methanobacteria*	*o__Methanobacteriales*	*f__Methanothermaceae*	*g__Methanothermus*	*s__Methanothermus_fervidus*
k__Archaea	*p__Euryarchaeota*	*c__Methanobacteria*	*o__Methanobacteriales*	*f__Methanobacteriaceae*	*g__Methanothermobacter*	*s__Methanothermobacter_thermautotrophicus*
k__Archaea	*p__Euryarchaeota*	*c__Methanobacteria*	*o__Methanobacteriales*	*f__Methanobacteriaceae*	*g__Methanothermobacter*	*s__Methanothermobacter_marburgensis*

k__Archaea	*p__Euryarchaeota*	*c__Methanobacteria*	*o__Methanobacteriales*	*f__Methanobacteriaceae*	*g__Methanobacterium*	*s__Methanobacterium_formicicum*

利用以上文件，生成自定义的 DMS 库：

MS-make-ref -i tree.newick -r tree.taxonomy -o tree.dms

输出的“tree.dms”即为用户自定义的 DMS 库。

2) 根据自定义的系统发育树和物种分类信息计算样本之间的 DMS 距离矩阵：

MS-comp-taxa-dynamic -D tree.dms -T samples.sp.abd -o samples.sp.dist

4. DMS 软件包中工具的汇总介绍

a. 基本工具

1) MS-comp-taxa-dynamic： 计算宏基因组间的 DMS 距离矩阵，示例用法见 3.1。可以运行

MS-comp-taxa-dynamic -h

了解详细的参数信息。

2) MS-comp-taxa：计算宏基因组间普通的 meta-storms 距离矩阵(Su 等，2012; Su 等， 2014) ，该方法忽略了宏基因组中未注释到物种水平的成分。用法与 3.1 中 MS-comp-taxa-dynamic 类似，可以运行

MS-comp-taxa –h

了解详细的参数信息。

b. 高级工具

1) MS-single-to-table：将输出的多个单样本文件（**表 1**）合并到同一张相对丰度表中（**表 3**），示例用法见步骤 2b。可以运行

MS-single-to-table –h

了解详细的参数信息。

2) MS-table-to-single：将相对丰度表拆分为多个单样本输出文件，是 MS-single-to-table 相反的操作。可以运行

MS-table-to-single –h

了解详细的参数信息。

3) MS-make-ref：根据自定义系统发育树和生物分类信息生成自定义 DMS 库，示例用法见步骤 3b。可以运行

```
MS-make-ref -h
```

了解详细的参数信息。

计算方法

1. 采用 Meta-Storms 算法计算可以识别的物种间的距离

在比较宏基因组样本时，对于可以直接识别的物种，DMS 使用 Meta-Storms 算法(Su 等， 2012; Su 等， 2014)基于物种水平的系统发育树计算两者之间的距离。具体来说，Meta-Storms 算法计算出两个样本在叶子结点（物种）上共享的丰度（**公式 1**），并根据系统发育树将叶子节点上剩余的丰度加权后作为其公共父节点的丰度（**公式 2**）。最终，通过遍历树的所有节点得到样本间整体相似度（**图 1**）。

$$sim(x) = min\ (abd(x, m1), abd(x, m2)) \tag{1}$$

$$abd(x,m) = \begin{cases} abd(x,m), if\ x\ is\ a\ tip\ node \\ or \\ (1 - len(x, x.left) * \big(abd(x.left, m) - sim(x.left)\big) + \\ (1 - len(x, x.right) * \big(abd(x.right, m) - sim(x.right)\big), \\ if\ x\ is\ an\ internal\ node \end{cases} \tag{2}$$

如**图 1** 所示，有两个宏基因组样本 S1 和 S2，它们的系统发育树是典型的二叉树，其中有三个物种 X，Y，Z 及每个物种的比例。第一步是根据 S1 和 S2 在 X 和 Y 上共享的丰度获得相似性。然后，将 X 和 Y 的剩余丰度乘以 1-Dist 作为它们的共同祖先的丰度（**图 1B**），并继续在 X 和 Y 的祖先结点与叶子结点 Z 上进行比较（**图 1C**）。最后，通过在根结点的比较，得到这两个样本的总体相似度为 92.8%（**图 1D**）。

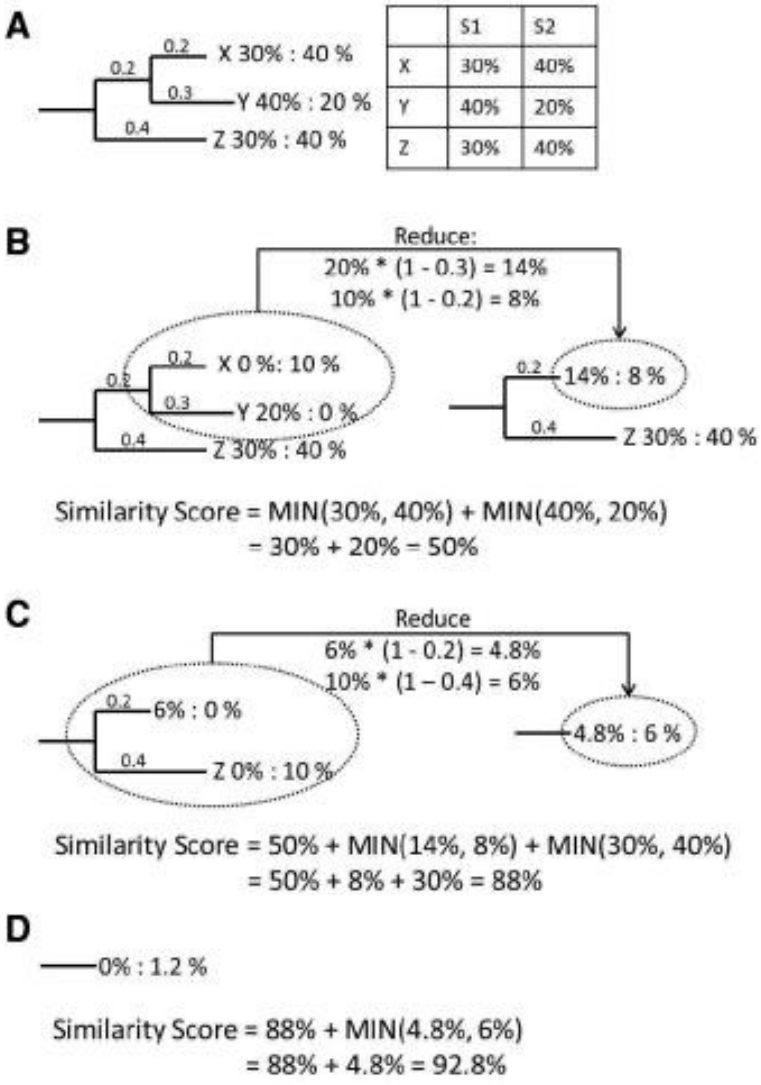

图 1. Meta-Storms 基于系统发育树来计算微生物组之间的相似度和距离

2. 采用虚拟结点映射计算未识别的群落成分

对于不能映射到物种水平（species level）但有更高层级（比如属、科、目等）分类信息的群落成分，它们不能直接映射到任何特定的叶结点或内部父节点，例如，在**图 2** 右侧的分类信息（taxonomy）中，属 A 的成分（g__A，即物种水平的 s__A_unclassified）包含的四个物种在**图 2** 左侧系统发育树（phylogeny）四个不同的分支上（s__A_sp1，s__A_sp2，s__A_sp3 和 s__A_sp4），但该系统发育树中并没有明确的中间结点来代表属 A（g__A）。例如以上四个物种在系统发育树中的共同父节点为结点 *b*，但该结点 *b* 下也同时包含了属 B 的物种 s__B_sp5 和 s__B_sp6。

为解决此问题，DMS 将未能识别到物种水平群落成分，根据其更高层级分类信息，

动态地插入到系统发育树中适当位置的虚拟结点来计算。该虚拟结点仅包含同一分类单元下的所有子分支。例如，在**图 2** 中，将属 A 映射到仅包含 s__A_sp1，s__A_sp2，s__A_sp3 和 s__A_sp4 的虚拟结点 *b'*（不包括其他属的物种）。在计算以上四个叶子结点上的相似度时不将其添加到总相似度中，而是当计算虚拟结点时，将其平均值作为虚拟节点的相似度加到总结果中。

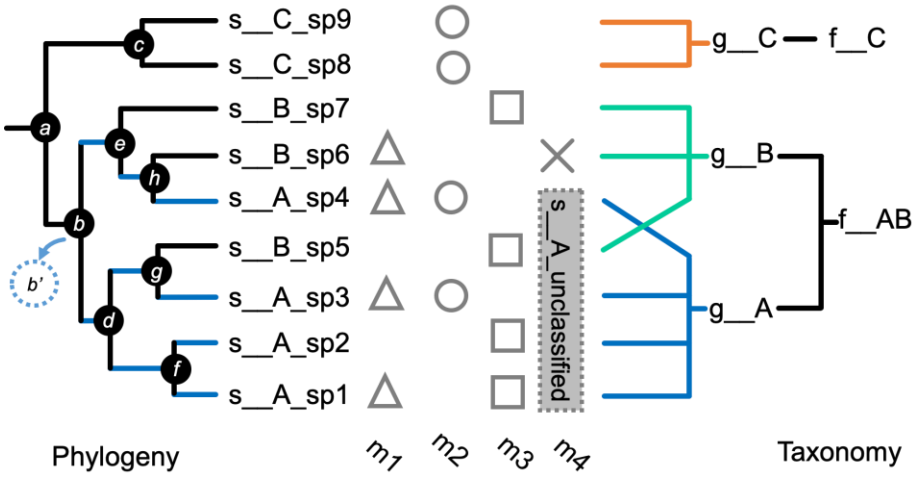

图 2. 对于缺少物种水平分类信息的群落成分，DMS 根据其更高层级分类信息，动态地插入到系统发育树中适当位置的虚拟结点来计算其对距离的贡献程度

结果与分析

为了验证 DMS 算法的准确性、可靠性与计算性能，本工作采用三个宏基因组数据集（**表 6**）对 DMS 算法进行测试，并与 Weighted UniFrac 和 Bray-Curtis 距离进行比较。

表 6. 测试数据集

数据集	样本量	来源
Synthetic Dataset 1	40	基于 48 个细菌物种的模拟合成样本
Real Dataset 1	2,355	Human Microbiome Project Phase 1 (Peterson 等， 2009)
Synthetic Dataset 2	100,000	基于 3,688 个细菌物种的模拟合成样本

以上测试中所有的数据集均可在 DMS 软件下载页面的“Supplementary”部分中下

载。

结果 1：基于模拟数据的算法验证

根据图 3a 的样本组成结构模拟合成的样本组成结构，模拟 4 组宏基因组样本（每组 10 个，共 40 个；Synthetic Dataset 1）。预期的样本距离如图 **1b** 所示，即 m1 与 m2 之间的距离大于 m1 与 m3 之间的距离（Case I）且 m1 与 m2 之间的距离大于 m1 与 m4 之间的距离（Case II）。 分别利用 DMS，Weighted UniFrac 和 Bray-Curtis 算法计算其距离矩阵，并进行主坐标分析（PCoA）来获得其 β 多样性分布。

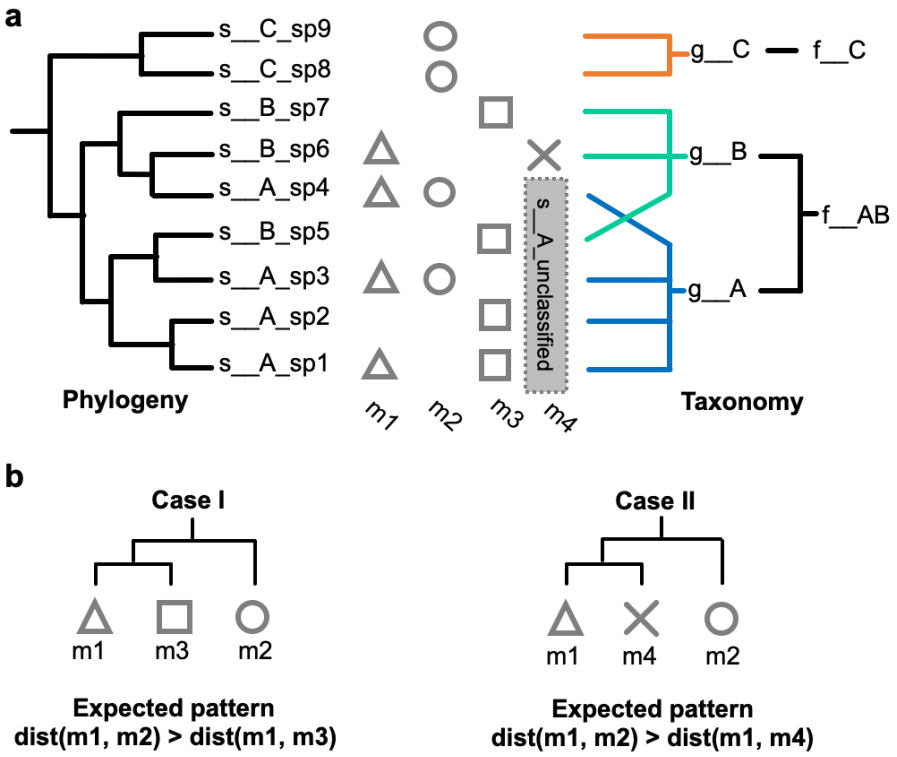

图 3. 模拟样本的组成结构（a）与预期的样本距离（b）

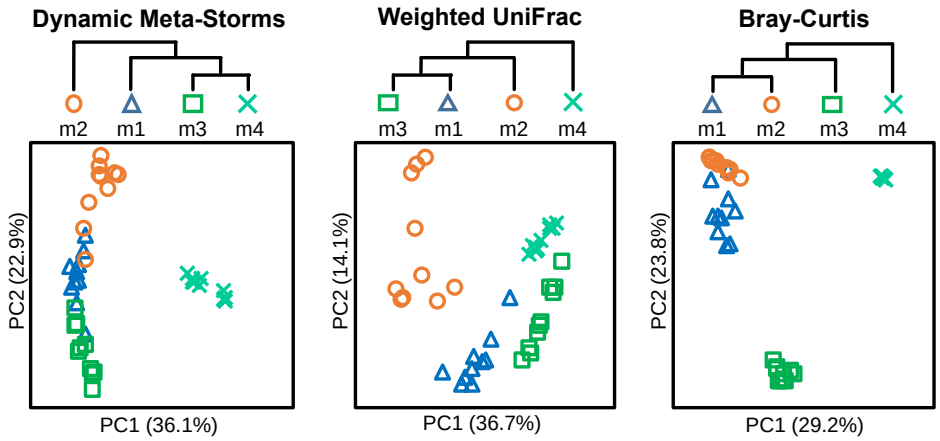

图 4. 基于 DMS，Weighted UniFrac 和 Bray-Curtis 距离的 PCoA 结果

图 4 的结果显示，只有 DMS 距离的结果与预期相符，Bray-Curtis 距离由于没有考虑物种间的进化距离，在 Case I 和 II 均出现错误；Weighted UniFrac 距离由于忽略了未注释到系统发育树叶子结点（即 species 水平）的成分，在 Case II 出现错误。

结果 2：基于真实数据的算法测试

将人类微生物组计划（HMP）第 1 阶段(Peterson 等， 2009)中的 2,355 个真实的人体宏基因组样本作为测试数据集（来自肠道、口腔、皮肤和生殖道四个身体部位；Real Dataset 1）分别利用 DMS，Weighted UniFrac 和 Bray-Curtis 算法计算其距离矩阵，Anosim 检验（1,000 次重复）结果显示 DMS 距离能够更明显地区分样本来源的身体部位（R = 0.965, *P* = 0.01；图 5）。

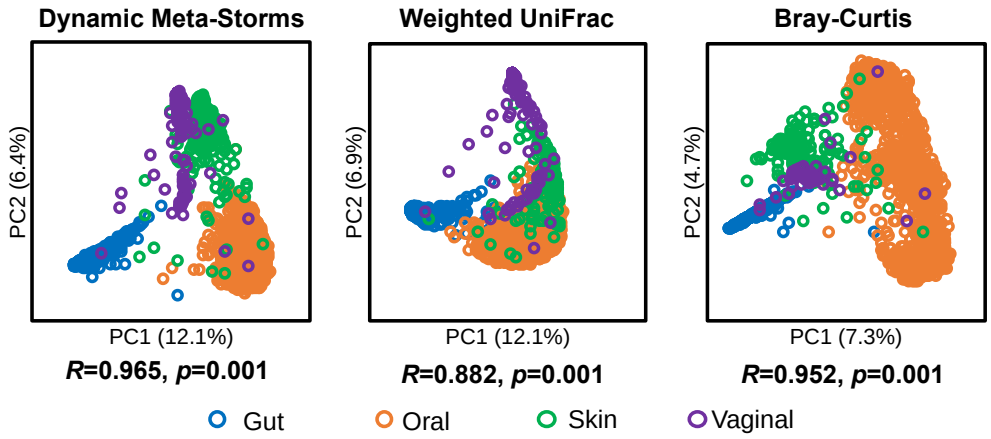

图 5. 根据 DMS，Weighted UniFrac 和 Bray-Curtis 距离进行 PCoA 分析和 Anosim 检验

结果 3：算法运行效率测试

随机选取不同数目（从 10,000 到 100,000）的宏基因组模拟样本（Synthetic Dataset 2），利用 DMS 算法计算其距离矩阵，并将 Striped UniFrac 算法(Mcdonald 等，2018)作为参照，比较两者的总体运行时间和 RAM（内存）的最大使用值。所有测试均在同一个具有 80 个线程的非共享计算节点上完成。图 6 中的结果显示，DMS 计算 10 万个宏基因组样本的距离矩阵，仅用 6.4 个小时即可完成，比参照方法快 20%，且节省了 40%以上的内存消耗。随着宏基因组样本的数量迅速增长，此功能将发挥更大的作用。

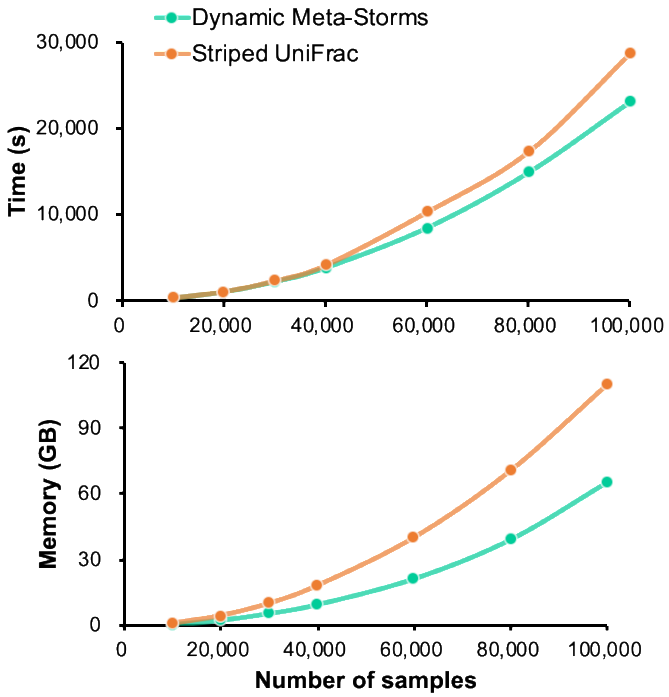

图 6. DMS 与 Striped UniFrac 计算不同数量样本的距离矩阵所消耗的运行时间和内存使用量的对比

失败经验

问题 1.

安装提示：“make: g++: command not found”

问题原因：没有安装 DMS 所需要的 g++编译器。

解决方法：根据不同的操作系统，利用相应的命令安装 g++，常见的操作系统：

Ubuntu 系统：sudo apt-get install g++

CentOS 系统：sudo yum install g++

MacOS 系统：brew install gcc

问题 2.

运行提示：“MS-comp-taxa-dynamic: command not found”

问题原因：环境变量设置失败。

解决方法：请参考实验步骤 1.2.2 中手动配置环境变量的方法将 DMS 所需要的环境变量添加到配置文件中。

问题 3.

运行提示：“Error: Please set the environment variable "DynamicMetaStorms" to the directory”

问题原因：DMS 的库文件没有配置到环境变量中。

解决方法：请参考实验步骤 1.2.2 中手动配置环境变量的方法将 DMS 的库文件配置到环境变量中。

问题 4.

运行提示：“Error: Cannot open file: XXX”。

问题原因：输入了错误的输入/输出文件路径。

解决方案：请检查正确的输入文件路径（可在输入时用 Tab 键自动补全），并确保用户在输出路径下有足够的写权限。

问题 5.

运行提示：“Argument #X Error : Arguments must start with -”。

问题原因：运行命令中所有参数选项名称必须以“-”开头。

解决方法：请检查第 X 个参数并更正。

问题 6.

运行提示：“Error: Features: s__XXXX does not have XX Samples”

问题原因：输入的物种丰度表中，物种“s__XXXX”所在行的丰度信息列数与样本数不统一。

解决方法：请按照**表 3** 的格式，检查样本数量（第一行）与“s__XXXX”所在行的丰度信息个数是否一致。如果某样本中不包含该物种，则丰度标为 0。如果输入的是该文件的转置格式（即每一行表示一个样本，每一列表示一个物种），请在运行 MS-comp-taxa 和 MS-comp-taxa-dynamic 时增加参数“-R F”。

致谢

本项工作得到了国家自然科学基金委员会 31771463 和 32070086，中国科学院 KFZD-SW-219-5，山东省自然科学基金会 ZR2017ZB0421 和 ZR201807060158，中国博士后科学基金会 2018M630807 的资助。

参考文献

1. Jing, G., Zhang, Y., Yang, M., Liu, L., Xu, J. and Su, X. (2019). Dynamic Meta-Storms enables comprehensive taxonomic and phylogenetic comparison of shotgun metagenomes at the species level. *Bioinformatics*(7): 7. https://doi.org/10.1093/bioinformatics/btz910
2. Mcdonald, D., Vázquez-Baeza, Y., Koslicki, D., Mcclelland, J., Reeve, N., Xu, Z., Gonzalez, A. and Knight, R. (2018). Striped UniFrac: enabling microbiome analysis at unprecedented scale. *Nature Methods* 15. https://doi.org/10.1038/s41592-018-0187-8
3. Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P.,

Zakhari, S., Read, J., Watson, B., Guyer, M. and Grp, N. H. W. (2009). The NIH Human Microbiome Project. *Genome Research* 19(12): 2317-2323. https://doi.org/10.1101/gr.096651.109

4. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9(8): 811-+. https://doi.org/10.1038/Nmeth.2066
5. Su, X., Wang, X., Jing, G. and Ning, K. (2014). GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU. *Bioinformatics*(7): 1031-1033. https://doi.org/10.1093/bioinformatics/btt736
6. Su, X., Xu, J. and Ning, K. (2012). Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics* 28(19): 2493. https://doi.org/10.1093/bioinformatics/bts470
7. Su, X. Q., Wang, X. T., Jing, G. C. and Ning, K. (2014). GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU. *Bioinformatics* 30(7): 1031-1033
8. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12(10): 902-903. https://doi.org/10.1038/nmeth.3589

