目录大纲

生物科学 101计划

01 **基因组学概述**

研究基因组序列、结构和功能的科学。

02 **基因组组装与注释**

基因组组装算法、基因预测技术和基因组注释方法。

03 **序列变异检测**

单碱基替换、短插入缺失和结构变异的检测技术。

04 **宏基因组学**

微生物组学概念、数据分析方法，以及在健康和环境领域的应用。

宏基因组学

01 微生物组介绍

02 微生物组高通量测序

03 微生物组测序数据和基本分析流程

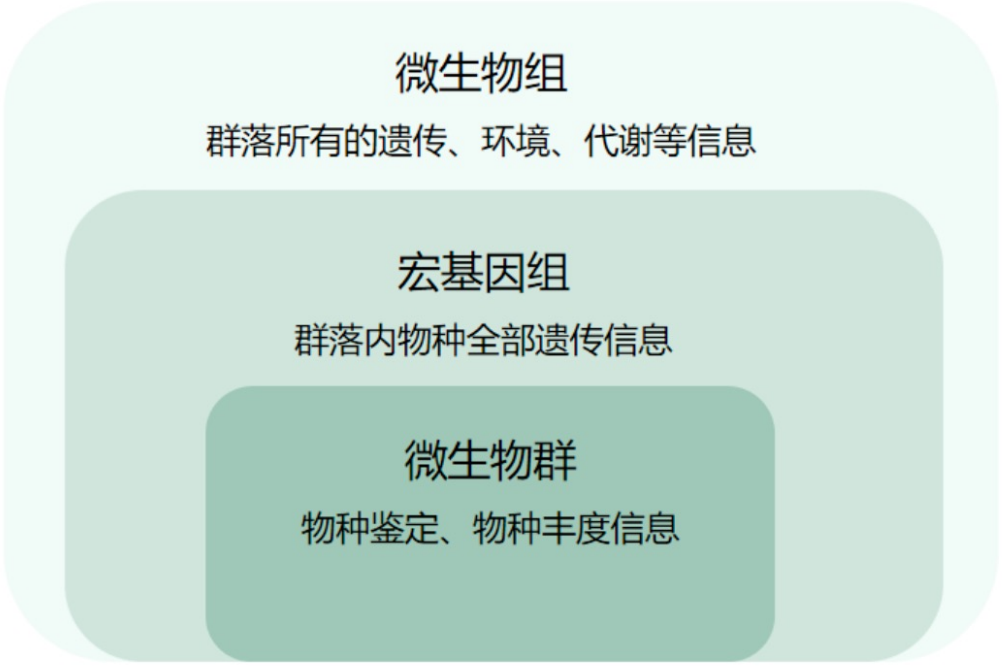
04 微生物组大数据与人工智能

宏基因组学

基本概念

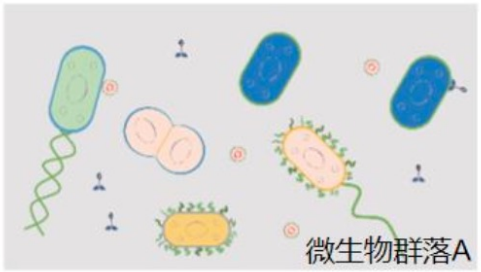
微生物群、宏基因组和微生物组

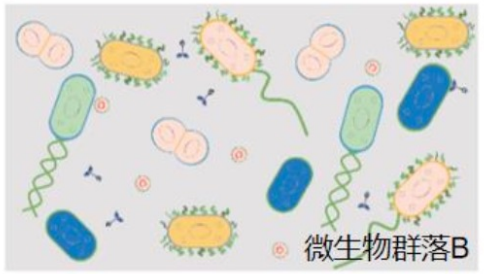
微生物组的研究范围最广，包含了微生物研究中的各种信息



微生物群

微生物群（microbiota）既包括植物体上共生或病理的微生物生态群体，也包括在土壤、水体和空气等环境中自由生存的细菌、古菌、原生动物、真菌和病毒，在宿主的免疫、代谢和激素等方面非常重要

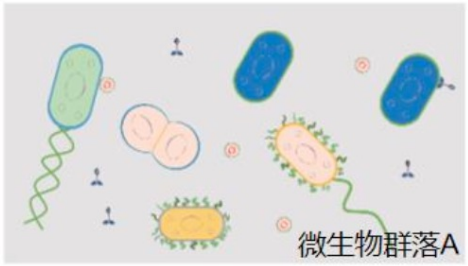


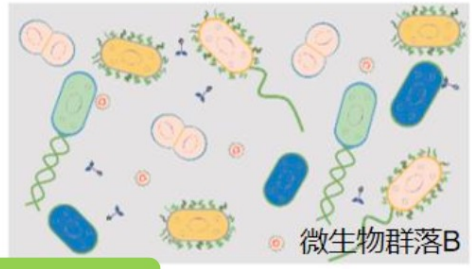


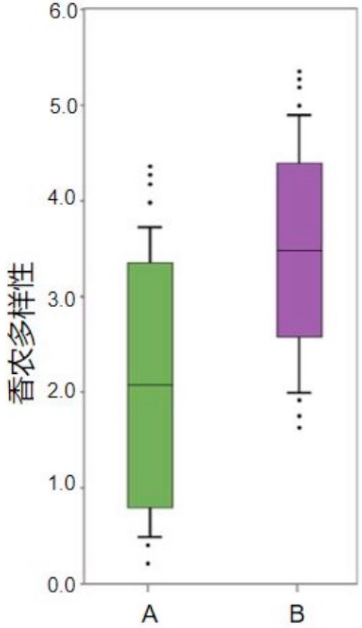
对于不同微生物群之间的差异度量主要是基于微生物群内微生物的组成和分布进行的，这是形成一个特定群落的基础，也是群落之间差异的来源。为了度量这种差异，多种度量指标被提出，其中最常用的有α-多样性、β-多样性和γ-多样性

微生物群

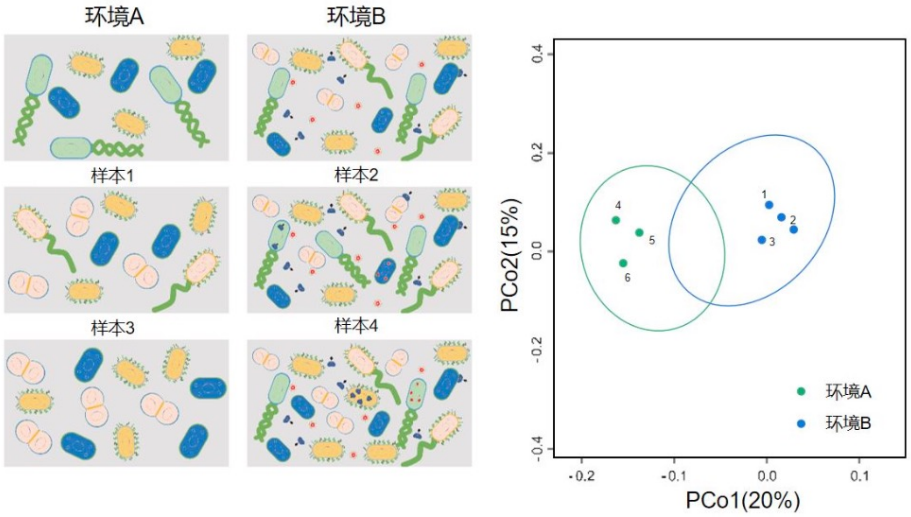
α-多样性表征一个群落内物种的个数（species richness，丰富度）和每个物种的数量及分布（evenness，均匀度）。







β-多样性是一种评估不同微生物组样本之间的特征差异的方法，其衡量的是不同样本或群落之间的物种多样性差异。



MOOC

一个高Alpha多样性的肠道菌群被认为是更健康、更稳定的生态系统，因为它功能更冗余，抵抗干扰的能力更强。

宏基因组学

宏基因组

- 宏基因组又称元基因组或微生物环境基因组（metagenome），其定义是“生境中全部微生物遗传物质的总和”。它直接将包含了可培养的和未可培养的微生物的微生物群落的所有遗传物质作为研究对象，广义来说其包括环境基因组、生态基因组学和群体基因组学。

- 传统的微生物学和微生物基因测序依赖于单克隆的培养，早期环境基因测序通过克隆16s rRNA基因等特定基因来确定自然样品中的生物多样性，但是此方法将会漏掉大量未被培养的微生物。因此，近期研究常采用鸟枪法或PCR直接测序方法来获得样品群体中所有成员无偏好的基因，这类方法可以展现从前无法发现的微生物多样性。

扩增子测序

1．基础概念

（1）16S rDNA（或16S rRNA）：编码原核生物核糖体小亚基的基因，长度约为1500bp，其分子大小适中，突变率小，是细菌系统分类学研究中最常用和最有用的标志。

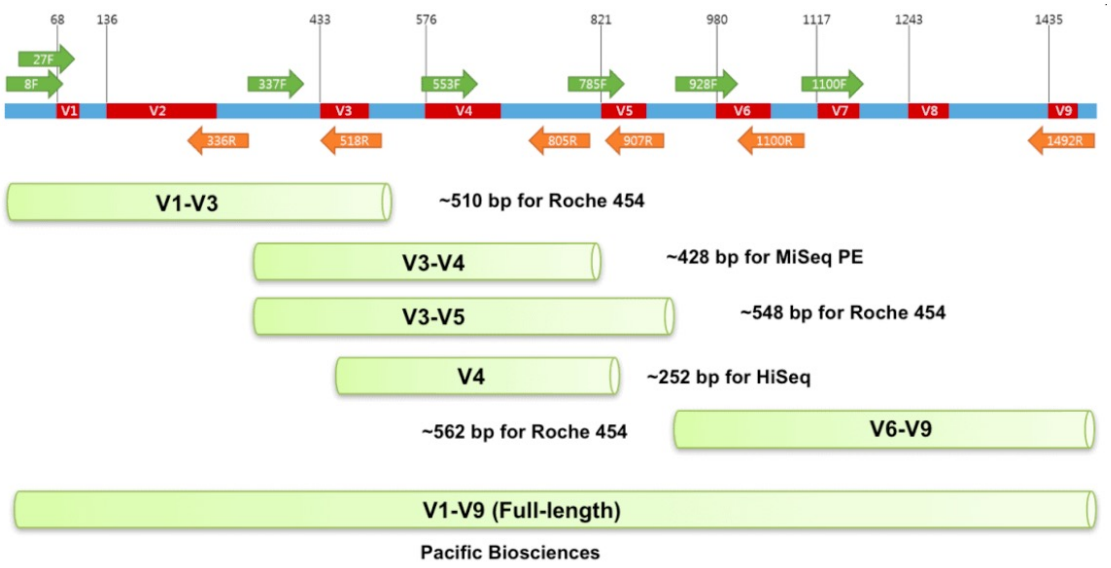
（2）操作分类单元（operational taxonomic units，OTU）：提取样品的总基因组DNA，利用16S rRNA或ITS的通用引物进行PCR扩增。不同的 16S rRNA序列的相似性高于97%就可以把它定义为一个OTU，每个OTU对应于一个不同的16S rRNA序列，也就是每个OTU对应于一个不同的细菌（微生物）种。通过OTU分析，可以知道样品中的微生物多样性和不同微生物的丰度。

（3）测序区段：由于16s rDNA较长，只能对其中经常变化的区域，也就是可变区进行测序。16s rDNA包含9个可变区，分别是V1~V9。研究中，一般对V3-V4双可变区域进行扩增和测序，也偶尔会对V1-V3区进行扩增和测序。

多拷贝，有保守区域和可变区域

扩增子测序

以16s rDNA扩增进行测序分析主要用于微生物群落多样性和构成的分析，而目前的生物信息学分析也可以基于16s rDNA的测序，对微生物群落的基因构成和代谢途径进行预测分析，大大拓展了我们对于环境微生物的微生态认知。



扩增子测序

测序过程

（1）提取样品DNA：DNA可以来自土壤、粪便、空气或水体等。

（2）质检和纯化：一般16s rDNA扩增子测序对DNA的总量要求并不高，总量>100ng，浓度>10ng/uL大多可以满足要求。如果是来自和寄主共生的环境，如昆虫的肠道微生物，提取其DNA时可能混合了寄主本身的大量DNA，因此对DNA的总量要求会有所提高。

（3）测序：对完成PCR后的产物进行测序。目前，可以采用多种不同的测序仪，如罗氏的454、Illumina的MiSeq、Life的PGM或Pacbio的RSII三代测序仪进行16s rDNA测序。

（4）数据分析：

①聚类统计，同源聚类获得OTUs；

②样本构成丰度分析，稀释曲线、Rank-Abundance曲线

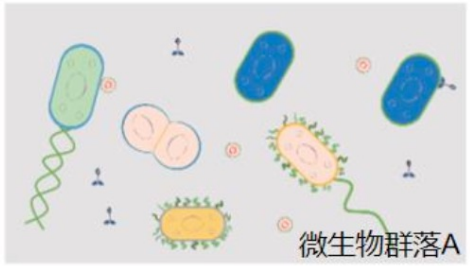
③多样性分析，PCoA、NMDS（非度量多维尺度分析）、PCA、LDA

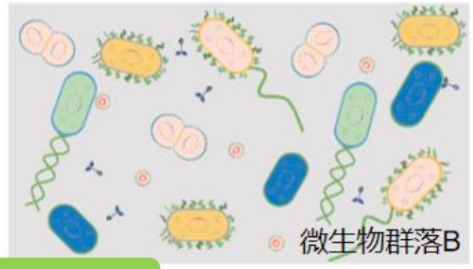
④差异性菌群分析，功能上的差异。

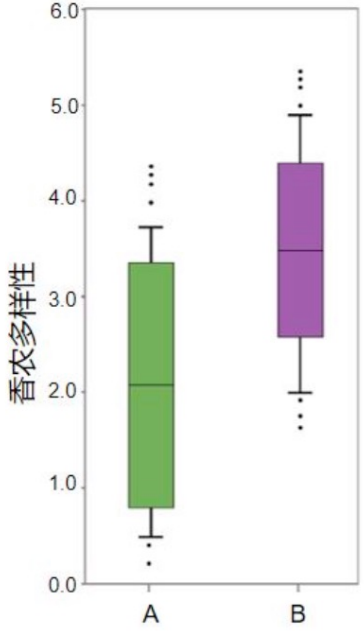
⑤环境因子分析，RDA（Redundancy analysis）、CCA（canonical correspondence analysis）

微生物群

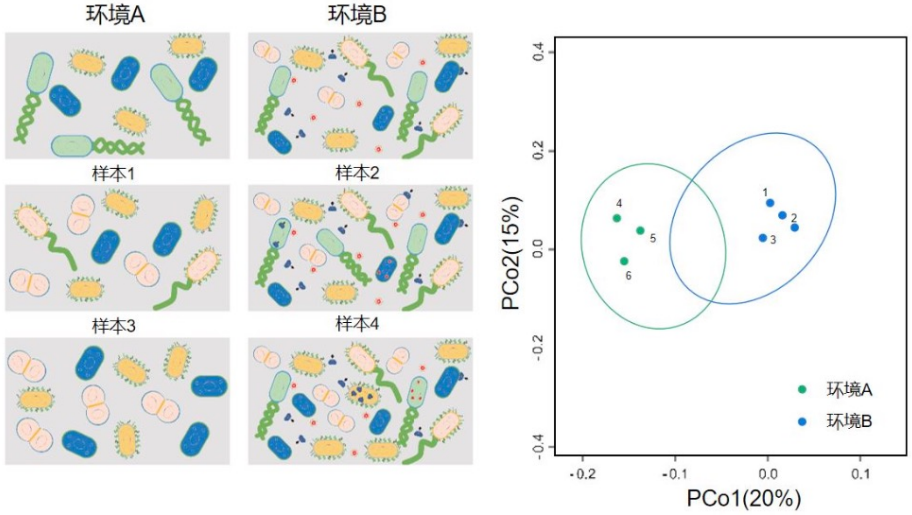
α-多样性表征一个群落内物种的个数（species richness，丰富度）和每个物种的数量及分布（evenness，均匀度）。







β-多样性是一种评估不同微生物组样本之间的特征差异的方法，其衡量的是不同样本或群落之间的物种多样性差异。



MOOC

扩增子测序的缺陷是什么？

宏基因组学

宏基因组测序

- 不同于传统的先培养微生物再提取DNA的做法，宏基因组直接收集能够代表特定生物环境生物多样性的样品；然后利用各种理化方法破碎微生物，使其释放DNA，再利用密度梯度离心等方法进行分离纯化。
- 接着，对DNA 进行酶切或者超声打断处理，并将其与合适的载体DNA 进行连接，构建重组体。
- 将带有宏基因组DNA的载体通过转化方式转入模式微生物，建立各自的无性繁殖系。
- 最后，对宏基因组文库的DNA 进行分析。

宏基因组测序有缺陷吗？

宏基因组测序

测序过程

(1) 样品总DNA的提取及基因或基因组DNA的富集

(2) 宏基因组文库的建立

(3) 宏基因组文库的筛选

鸟枪法宏基因组测序的拓展研究

2. 靶向探索“微型宏基因组”

在提取和测序DNA前，可以将复杂的微生物群落分为较小的亚组。荧光激活细胞分选（FACS）是一种复杂但更灵活、更精确随机或非随机地生成微型元数据的方法。例如，使用FACS从森林土壤中回收了一些未经培养的巨型病毒基因组，这些病毒基因组不能通过土壤样本在同一深度鸟枪测序方法中进行组装，并支持将复杂群落细分为低多样性的微观宏观基因组以恢复稀有成员的观点，并且使用传统的大规模宏基因组学方法可能会忽略这些稀有成员。

怎么精准靶向？

宏基因组学

微生物组（microbiome）包括微生物（细菌、古细菌、低等或高等真核生物和病毒）的基因组，以及其周围环境，也就是说微生物组既包括微生物物种，又包括各个物种的基因组以及相关环境因素和代谢产物。微生物组是结合了宏基因组学、代谢组学、宏转录组学、以及宏蛋白组学等和临床/环境数据的集合。

微生物组学研究内容：

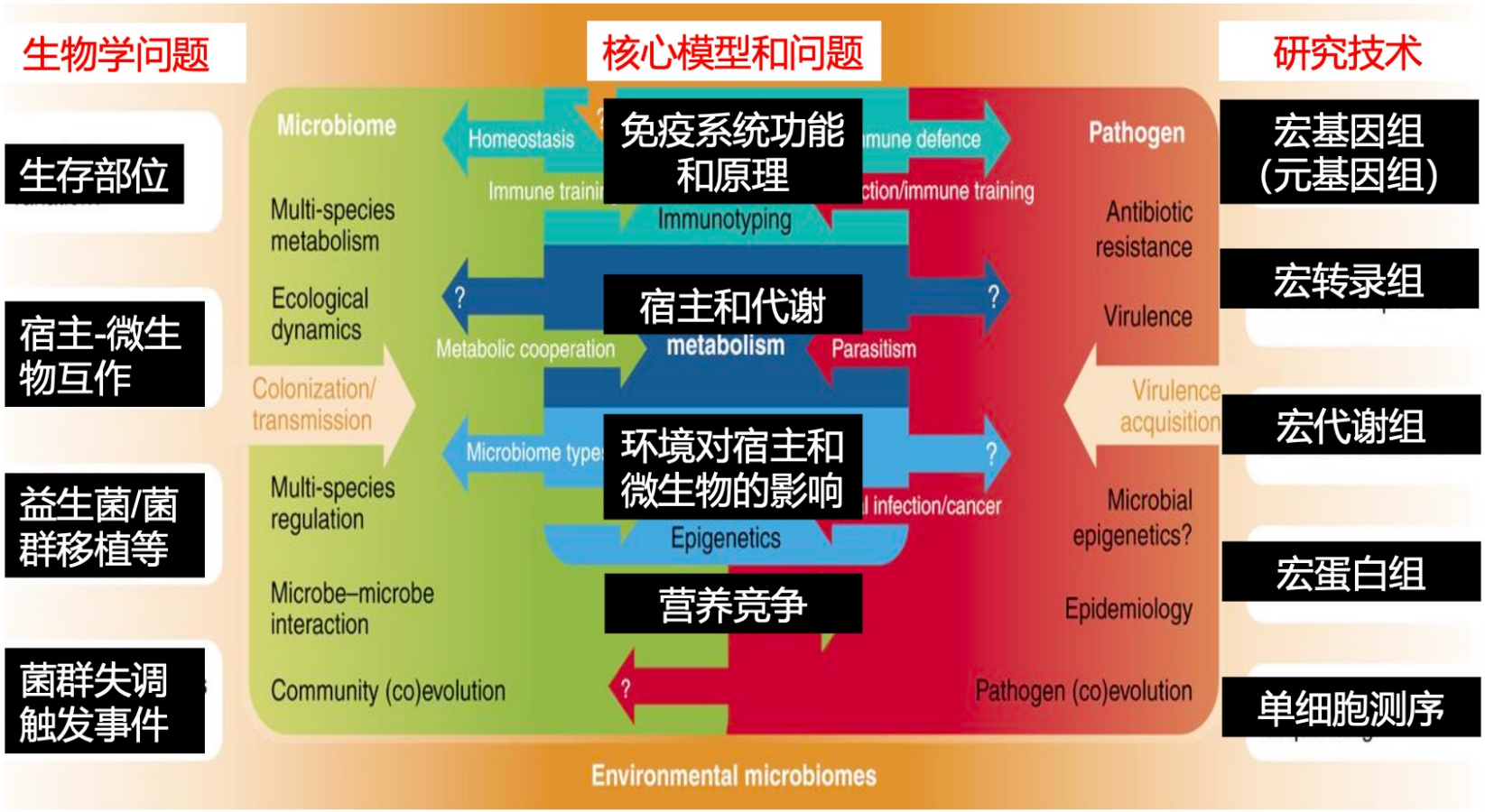
①微生物培养，这是了解微生物形态结构和生理功能最直接的方法，但是微生物培养一般费时费力，且许多微生物是不可培养的，基于高通量测序可以解决这些问题。

②微生物测序，高通量测序技术的进步极大地促进了有关微生物的研究，基于高通量技术的微生物研究平台主要包括增子测序技术和宏基因组测序技术等。

③多组学研究，基因测序方法难以鉴定微生物中的关键功能分子，单独使用时无法回答何种成员微生物通过何种方式影响宿主等关键问题。而微生物组学与代谢组学等多组学联用的优势逐渐突出，其关联分析在宿主生理、疾病病理、药物药理等领域已取得众多进展，具有良好应用前景。

宏基因组学

微生物组数据分析挖掘



Huttenhower, et al. *Molecular Systems Biology*, 2013

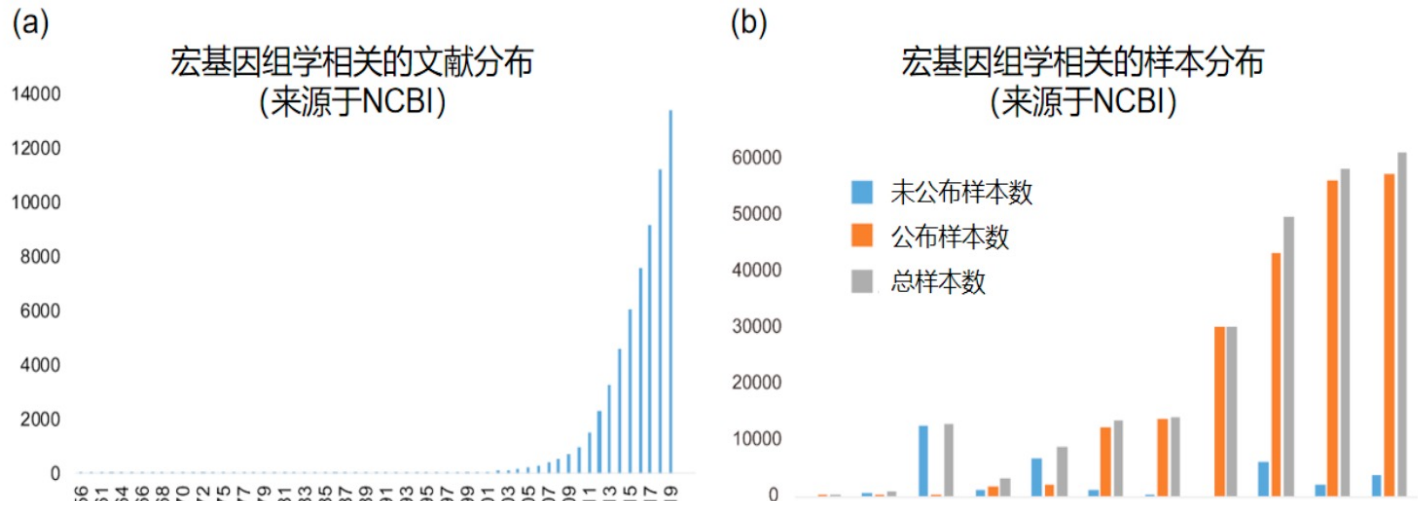
宏基因组学

微生物组测序数据和基本分析流程

微生物组数据积累和整合

微生物组学数据的积累大大促进了微生物群落的研究，在过去十年间微生物相关的论文数量呈现指数增长，微生物组数据量每年以>100TB的速度增长。国际上已经建立了许多宏基因组相关的数据库，比如MG-RAST（http://metagenomics.anl.gov/）和CAMERA（http://camera.calit2.net/）等。

其中，NCBI的Sequence Read Archive （SRA） （http://www.ncbi.nlm.nih.gov/sra）、MG-RAST以及CAMERA2中公开的宏基因组项目超过10，000个，包含超过1PB的数据。



微生物组数据积累和整合

- 按照菌群来源的生存环境（biome）而组织起来的微生物群落样本和相关测序数据，是依据生存环境本体的组织架构，通过层级结构组织起来的。例如：截止2019年底，EBI MGnify的生存环境本体组织架构包括491个本体[26]，而人体大肠排泄物菌群的本体定位是“root > Host-associated > Human > Digestive system > Large intestine > Fecal”。这种本体结构非常有利于样本的分类。然而，目前这种本体的层级组织结构并非完全是树状的，而是具有一个本体属于多个本体的直接子本体的特征，例如“Fecal”就有多达5个以上的上一级本体信息。因此，每个微生物组数据的相关生存环境本体都有可能具有多标签（multi-label）。

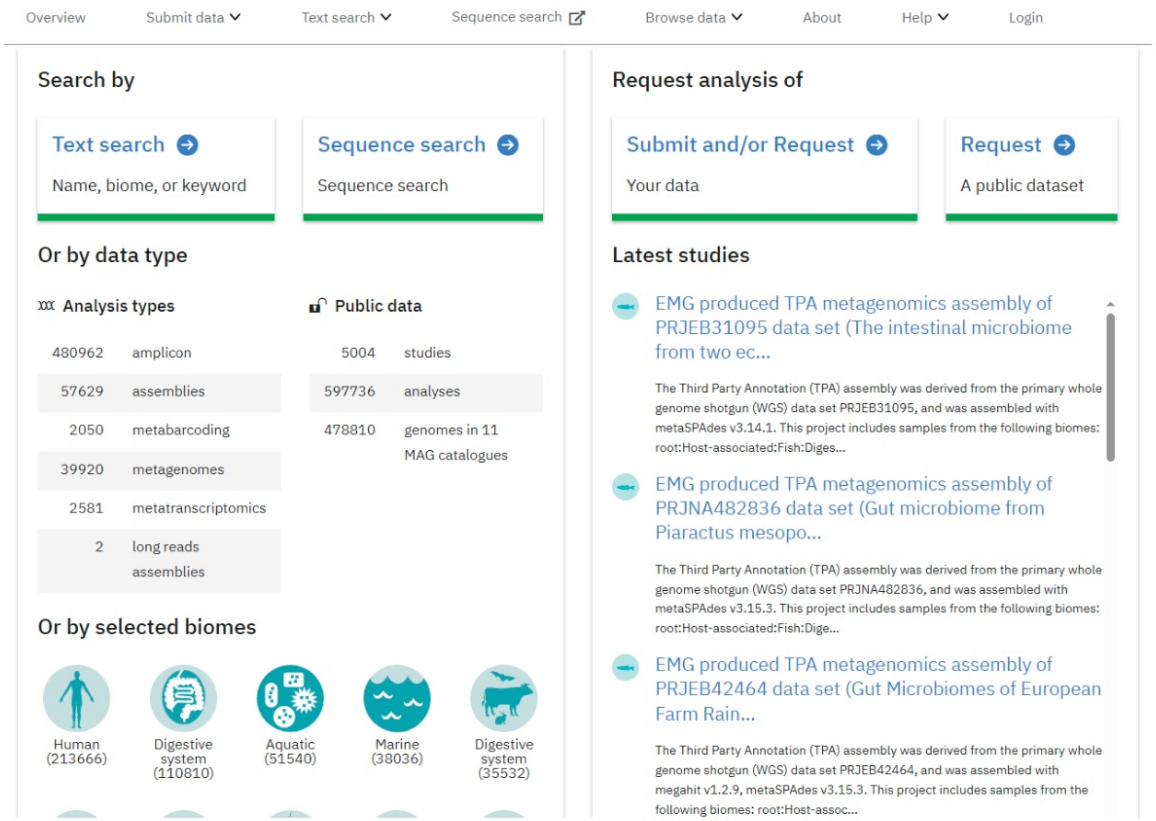
- 从一方面来说，微生物组数据的多标签属性，不利于样本的简单分类，造成了样本分类和比较方面的瓶颈。

- 另一方面，微生物组数据的多标签属性符合大数据研究的特征，利用机器学习或者深度学习等方法来处理将有望获得较好的结果。

宏基因组学

微生物组数据积累和整合

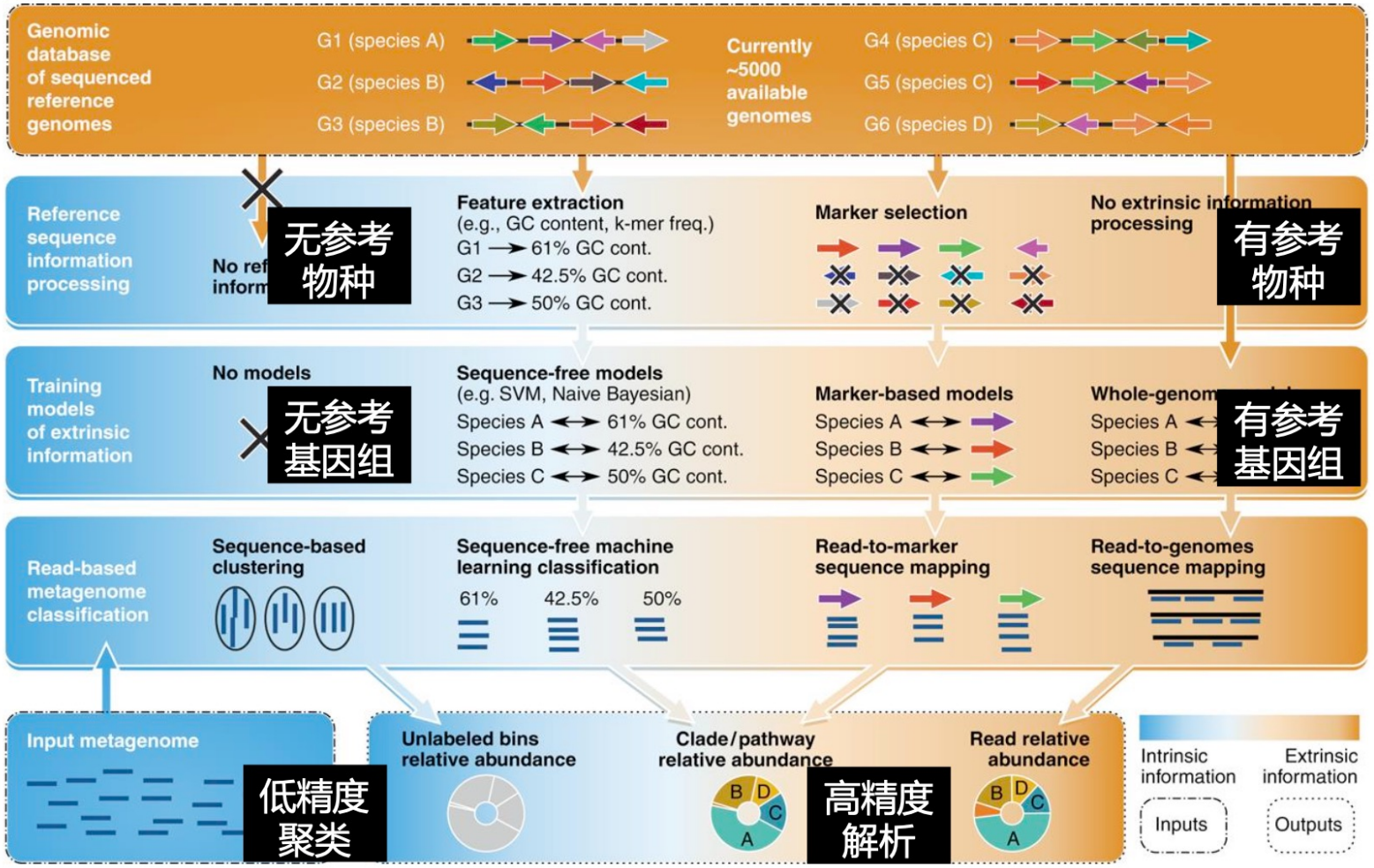
https://www.ebi.ac.uk/metagenomics



代表性的微生物组数据库（MGnify）

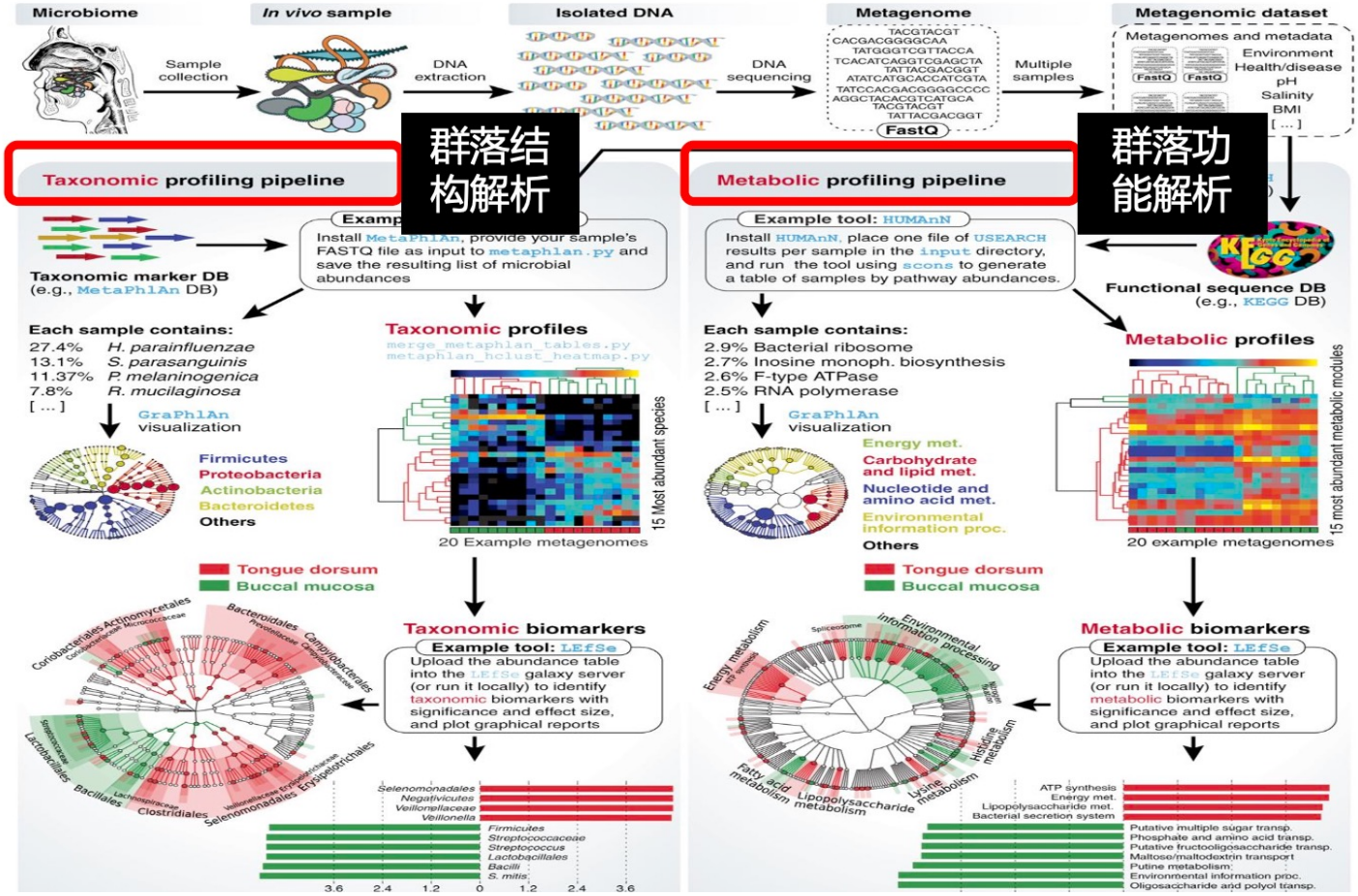
宏基因组学

微生物组数据分析挖掘



MOOC

微生物组数据分析挖掘


Microbiome
In vivo sample
Isolated DNA
Metagenome
Metagenomic dataset
Sample collection
DNA extraction
DNA sequencing
Multiple samples
Metagenomes and metadata
Environment
Health/disease
pH
Salinity
BMI
FastQ
群落结构解析
群落功能解析
Taxonomic profiling pipeline
Metabolic profiling pipeline
Taxonomic marker DB
(e.g., MetaPhlAn DB)
Example tool: MetaPhlAn
Install MetaPhlAn, provide your sample's FASTQ file as input to metaphlan.py and save the resulting list of microbial abundances
Example tool: HUMAnN
Install HUMAnN, place one file of USEARCH results per sample in the input directory, and run the tool using scons to generate a table of samples by pathway abundances.
Functional sequence DB
(e.g., KEGG DB)
Each sample contains:
27.4% H. parainfluenzae
13.1% S. parasanguinis
11.37% P. melaninogenica
7.8% R. mucilaginosa
[...]
GraPhlAn visualization
Firmicutes
Proteobacteria
Actinobacteria
Bacteroidetes
Others
Taxonomic profiles
merge_metaphlan_tables.py
metaphlan_hclust_heatmap.py
15 Most abundant species
20 Example metagenomes
Each sample contains:
2.9% Bacterial ribosome
2.7% Inosine monoph. biosynthesis
2.6% F-type ATPase
2.5% RNA polymerase
[...]
GraPhlAn visualization
Energy met.
Carbohydrate and lipid met.
Nucleotide and amino acid met.
Environmental information proc.
Others
Metabolic profiles
15 most abundant metabolic modules
20 example metagenomes
Tongue dorsum
Buccal mucosa
Taxonomic biomarkers
Example tool: LEfSe
Upload the abundance table into the LEfSe galaxy server (or run it locally) to identify taxonomic biomarkers with significance and effect size, and plot graphical reports
Metabolic biomarkers
Example tool: LEfSe
Upload the abundance table into the LEfSe galaxy server (or run it locally) to identify metabolic biomarkers with significance and effect size, and plot graphical reports
Selenomonadales
Negativicutes
Veillonellaceae
Veillonella
Firmicutes
Streptococcaceae
Streptococcus
Lactobacillales
Bacilli
S. mitis
ATP synthesis
Energy met.
Lipopolysaccharide met.
Bacterial secretion system
Putative multiple sugar transp.
Phosphate and amino acid transp.
Putative fructooligosaccharide transp.
Maltose/maltodextrin transport
Putine metabolism
Environmental information proc.
Oligosaccharide and polyol transp.

宏基因组学

微生物组数据分析挖掘

- 随着海量微生物组数据的积累，涌现了大量的微生物组数据库，以及大量的微生物组数据分析方法和软件。

- 其中主流的微生物组数据库包括EBI MGnify，QIITA等通用微生物组数据库，以及针对抗性基因挖掘的CARD数据库、针对生合成代谢基因簇挖掘的antiSMASH等。

- 微生物组数据常用的分析方法和软件包括：针对测序数据质量控制的FastQC，针对微生物组测序数据分析（从测序数据到物种结构）的QIIME 2.0和MetaPhlAn，针对微生物组的功能谱分析的 HUMAnN 2.0，针对微生物组溯源分析的SourceTracker，针对微生物组功能基因挖掘的DeepARG、antiSMASH等方法。

宏基因组学

微生物组数据分析挖掘

名称	简介	网址	参考文献
Trimmomatic	一种用于Illumina NGS数据低质量、引物和接头序列去除工具。		[23]
	MetaPhlAn用于从宏基因组中分析微生物群落的组成。	https://huttenhower.sph.harvard.edu/metaphlan2	[24]
HUMAnN2	HUMAnN 2.0可以高效、准确地分析一个群落中微生物路径存在/缺失和丰度。	https://huttenhower.sph.harvard.edu/humann2	[25]
MEGAHIT	超快、省内存的宏基因组组装软件。	https://github.com/voutcn/megahit	[26]
CD-HIT	构建非冗余基因集。	http://weizhongli-lab.org/cd-hit	[27]

代表性的宏基因组学数据分析方法和软件

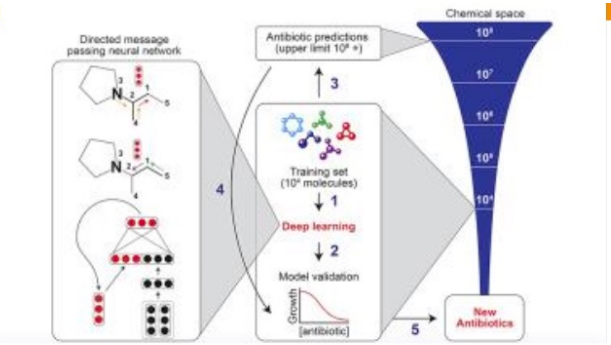
微生物组数据分析挖掘

表 1.2 常用分析工具

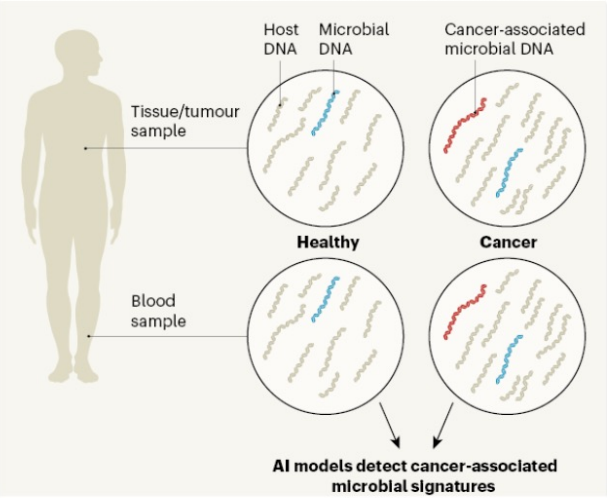
软件（平台）	分析数据对象	分析结果
MOCAT	宏基因组	物种结构、丰度和功能分类，以及物种之间的比较
MEGAN	16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较
MetaPhlAn	宏基因组	物种结构、丰度
PICRUSt	宏基因组，16S rRNA	物种结构和功能分类
antiSMASH	宏基因组	BGC 分析
CARMA	16S rRNA	物种结构和功能分类
Sort-ITEMS	16S rRNA	物种结构和功能分类
QIIME	16S rRNA	物种结构、丰度和功能分类
MG-RAST	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较
CAMERA	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较
IBDsite	宏基因组，16S rRNA	物种结构、丰度和功能分类，以及物种之间的比较

宏基因组学

微生物组大数据与人工智能



A Deep Learning Approach to Antibiotic Discovery, Cell, 2020



AI finds microbial signatures in tumours and blood across cancer types, Nature, 2020

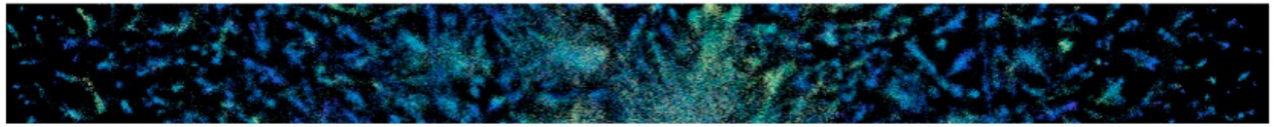
AI赋能的微生物组大数据挖掘

人工智能发掘微生物组数据特征
人工智能挖掘重要功能基因
人工智能解构时空动态变化模式
人工智能预测表型
人工智能预测疾病发生发展
……

宏基因组学

https://github.com/facebookresearch/esm

Evolutionary Scale Modeling



Update April 2023: Code for the two simultaneous preprints on protein design is now released! Code for "Language models generalize beyond natural proteins" is under examples/lm-design/. Code for "A high-level programming language for generative protein design" is under examples/protein-programming-language/.

This repository contains code and pre-trained weights for **Transformer protein language models** from the Meta Fundamental AI Research Protein Team (FAIR), including our state-of-the-art **ESM-2** and **ESMFold**, as well as **MSA Transformer**, **ESM-1v** for predicting variant effects and **ESM-IF1** for inverse folding. Transformer protein language models were introduced in the 2019 preprint of the paper "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". ESM-2 outperforms all tested single-sequence protein language models across a range of structure prediction tasks. ESMFold harnesses the ESM-2 language model to generate accurate structure predictions end to end directly from the sequence of a protein.

In November 2022, we released `v0` of the ESM Metagenomic Atlas, an open atlas of 617 million predicted metagenomic protein structures. The Atlas was updated in March 2023 in collaboration with EBI. The new `v2023_02` adds another 150 million predicted structures to the Atlas, as well as pre-computed ESM2 embeddings. Bulk download, blog post and the resources provided on the Atlas website are documented on this README.

代表性的微生物组大模型（基于宏基因组学数据）

宏基因组学

Available Models and Datasets

Pre-trained Models

Shorthand	`esm.pretrained.`	#layers	#params	Dataset	Embedding Dim	Mod
ESM-2	`esm2_t48_15B_UR50D`	48	15B	UR50/D 2021_04	5120	https:// esm/m
	`esm2_t36_3B_UR50D`	36	3B	UR50/D 2021_04	2560	https:// esm/m
	`esm2_t33_650M_UR50D`	33	650M	UR50/D 2021_04	1280	https:// esm/m
	`esm2_t30_150M_UR50D`	30	150M	UR50/D 2021_04	640	https:// esm/m
	`esm2_t12_35M_UR50D`	12	35M	UR50/D 2021_04	480	https:// esm/m
	`esm2_t6_8M_UR50D`	6	8M	UR50/D 2021_04	320	https:// esm/m
ESMFold	`esmfold_v1`	48 (+36)	690M (+3B)	UR50/D 2021_04	-	https:// esm/m
	`esmfold_v0`	48 (+36)	690M (+3B)	UR50/D 2021_04	-	https:// esm/m
	`esmfold_structure_module_only_*`	0 (+various)	various	UR50/D 2021_04	-	https:// esm/m

代表性的微生物组大模型（基于宏基因组学数据）