基于深度学习的分泌蛋白原核表达优化 初赛数据说明

第三届 Bio-OS AI 开源大赛筹备组

本压缩包包含以下文件

2025_bio-os_data

|– Readme.pdf 本数据说明文件

|– Reference.txt 数据来源

|–dataset 有标数据，共 52158 蛋白条目。

Ec.tsv 大肠杆菌有标数据，共 18780 蛋白条目。

Human.tsv 人有标数据，共 13421 蛋白条目。

mouse.tsv 鼠有标数据，共 13253 蛋白条目。

Pic.tsv 毕赤酵母有标数据，共 320 蛋白条目。

Sac.tsv 酿酒酵母有标数据，共 6384 蛋白条目。

|

└– Tests.xlsx 无标数据，共 505 蛋白条目。

dataset 文件夹下包含的有标数据供参赛者训练、验证模型使用，每个 tsv 文件中的每个蛋白条目包含以下信息：

Entry	该蛋白序列对应的 Uniport id
Reviewed	是否经过审阅（属于 Swiss-Prot 子库）
Entry Name	该蛋白序列的 UniProt 标识符
Protein names	蛋白名称
Gene Names	基因名称
Organism	该蛋白来自的生物体，如人、鼠、大肠杆菌
Length	该蛋白序列长度
Subcellular location [CC]	该蛋白的亚细胞定位
RefSeq_id	该蛋白对应的 NCBI 数据库的标识符，包含 cds 信息
	[注]当为“embl”时，序列来自 EMBL 数据库，下行为其标识符
Genome_id	该蛋白序列 cds 信息来自的基因组的 NCBI 标识符
	[注]当 RefSeq_id 为“embl”时，为 EMBL 数据库标识符
RefSeq_nn	与蛋白序列对应的 coding sequence（cds）核酸序列信息
RefSeq_aa	与核酸序列对应的蛋白序列信息，*对应终止密码子

Tests.xlsx 文件中包含以下内容，空白区域需要参赛者进行回答，回答格式请与 dataset 中 RefSeq_nn 格式保持一致（一个氨基酸对应三个核酸，且包括终止密码子）。

id	测试序列序号（按照蛋白序列长度排序，无数据集索引含义）
RefSeq_aa	未知其他蛋白信息的蛋白序列
Homo sapiens (Human)	若将该蛋白序列在人内表达，则其对应的核酸序列为？

Mus musculus (Mouse)	若将该蛋白序列在鼠内表达，则其对应的核酸序列为？
Escherichia coli	若将该蛋白序列在大肠杆菌内表达，则其对应的核酸序列为？
Saccharomyces cerevisiae	若将该蛋白序列在酿酒酵母内表达，则其对应的核酸序列为？
Pichia angusta	若将该蛋白序列在毕赤酵母内表达，则其对应的核酸序列为？

［注］本压缩包中所有提供的数据仅限于本次比赛使用。请尤其注意遵守原始数据发布者的使用条款和引用要求。为了防止测试数据泄露，初赛期间，请参赛者尽量避免使用任何形式的预训练模型，若使用请务必注明。参赛者可以使用额外的数据集进行训练，但需在提交的报告中进行明确说明（包含但不限于数据来源，与测试序列的相似性等信息），同时提供未使用额外数据的版本。