**基于DNA序列信息预测基因表达**

**一、背景与意义**

基因表达是细胞功能和生物体表型形成的核心过程，其调控机制复杂且精细。DNA序列中的遗传信息通过转录过程形成mRNA，再由mRNA翻译成蛋白质执行生物学功能。在此过程中，mRNA的半衰期，即其在细胞内稳定存在的时间，其受到DNA序列当中5端、三端、GC含量等特征影响，DNA碱基序列及半衰期相关特征对基因表达的动态调控具有重要影响。

深度学习通过构建复杂的神经网络模型，能够处理和分析大规模的生物数据，包括基因序列、转录因子结合位点、表观遗传标记等 。基于深度学习xAI 等方法，可以有效找出影响预测结果的重要特征，这对于寻找DNA序列当中对于基因表达有重要影响的碱基片段提供了有效帮助。

综上所述，通过深度学习方法分析DNA序列特征和mRNA半衰期数据，可以有效预测基因表达水平，结合xAI的方法进而揭示基因表达调控的复杂模式，为理解细胞功能、疾病机制以及指导精准医疗提供了新的视角。

## **二、题目**

* **所属领域：**基于深度学习算法的基因表达预测
* **任务描述：**本次竞赛分为三个子任务。

**子任务1：使用深度学习算法预测基因表达。**

根据给出的DNA序列和半衰期数据，设计并实现一个深度学习模型，并预测给定数据集基因表达水平。

\*注：其中不可将测试集数据用于模型训练。

**子任务2：完成模型可解释性分析**

对模型进行可解释性分析，对于测试集，识别出在启动子的DNA序列中与高表达水平相关的碱基（片段）。

**子任务3：增加额外数据特征，提升模型性能。**

除了本赛题当中给出的DNA序列和半衰期数据，请参赛选手自由寻找其它数据特征，如转录因子结合位点、表观遗传标记等，使其能够进一步提升模型预测的准确性。

\*注：（1）其中不可将测试集数据用于模型训练。（2）额外数据特征不允许包含该基因本身的表达值数据。

赛题讲解：<http://www.aisccc.cn/course/course-details?id=117>

## **三、数据来源**

1. 使用GM12878细胞系，基于在hg19坐标系当中处理后的数据集，其中包含DNA序列信息，mRNA半衰期特征等。

数据集文件包含训练集（train.h5）, 验证集（ valid.h5）, 测试集（test.h5） 三份文件。每份文件为h5格式，每份文件包含gene\_id（基因id）, halflife （半衰期）, promoter（启动子）,label（标签） 四个键值。其含义如下：

|  |  |  |
| --- | --- | --- |
| **key** | **comment** | **example** |
| gene\_id | 基因id (ENSID) | ENSG00000000003 |
| halflife | 标准化后的半衰期数据，每条数据为8维向量数据。 | array([ 1.11, 1.02, 0.87,... ]) |
| promoter | 经过one-hot 处理后的DNA序列数据，每条数据为（20000，4）维度的数据 | array([[False,False,False,True],...) |
| label | 预测目标，0为低表达，1为高表达 | 1 |

* gene\_id (基因id)：数据的每条数据为通用的ENSID ，为Ensembl数据库使用的ID标识符，用于唯一标识不同的分子特征，例如基因、转录本、外显子和蛋白质等。
* halflife（半衰期）：数据的每条数据以UTR5LEN、CDSLEN、INTRONLEN、UTR3LEN、UTR5GC、CDSGC、UTR3GC、ORFEXONDENSITY 顺序排列，每条数据经过标准化处理。半衰期具体数据含义如下：

|  |  |  |
| --- | --- | --- |
| **Feature** | **example** | **comment** |
| UTR5LEN | 1.11 | 5端长度 |
| CDSLEN | 1.03 | 序列长度（Coding Sequence） |
| INTRONLEN | 0.87 | 内显子长度 |
| UTR3LEN | 0.88 | 3端长度 |
| UTR5GC | -0.24 | 5端GC含量 |
| CDSGC | 0.32 | 序列GC含量 |
| UTR3GC | 0.05 | 3端GC含量 |
| ORFEXONDENSITY | 0.24 | 开放阅读框外显子连接密度 |

* promoter (启动子)每条数据为DNA序列数据，每一条数据为长度20000bp的DNA序列（TSS ±10000bp），其中的每一个碱基（ATCG）为one-hot 处理后的结果。碱基one-hot对应编号为{'A':0, 'C':1, 'G':2, 'T':3}。
* label (标签)为预测目标，包含0，1两类数值，其中0表示该条数据对应的基因为低表达基因，1表示该条数据对应的基因为高表达基因。

数据详情及文件下载链接: <http://www.aisccc.cn/database/data-details?id=121>

## **四、评价标准**

**1. 准确性指标：**测试集上的AUC，F1-Score。

**2. 性能指标：**模型参数数量、训练时间、推理时间等。

**3. 创新性：**包含可解释性，模型设计、特征工程等原创性工作，评价标准强调新颖的方法对于现有技术的显著改进以及对于最终效果的提升。

## **五、提交内容**

**1. 赛题设计文档：**包含选定赛题的题目、背景、实现原理、具体实现内容、结果等信息；

**2.完整的、可运行的代码：**若涉及模型，需提供训练好的模型，若涉及训练数据，需要一并提供数据集或网盘等下载链接。