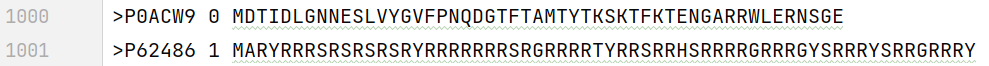
**实验7 蛋白质功能预测实验**

**1. 实验数据**

**(1) 训练集**

该训练数据集包含2000条蛋白质序列，存放于文件“ProSeqs\_Train.txt”。文件中每一行若以空格隔开，内容分别是序列ID、类标号以及蛋白质的氨基酸序列。截取文件中第1000和1001行内容，如下图所示：



其中“P62486”为序列ID，1表示该蛋白质具有某种生物功能（反之，0表示蛋白质不具有这种生物功能），而“MARYRRRSRSRSRSRYRRRRRRRSRGRRRRTYRRSRRHSRRRRGRRRGYSRRRYSRRGRRRY”为蛋白质P62486的氨基酸序列。

自然界中，天然蛋白质主要由20种标准的氨基酸构成，分别对应于字母表AA20={'A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V'}中的英文字母。若在蛋白质序列中存在非标准的氨基酸，即存在AA20以外的英文字母，则统一记为'X'。

**(2) 测试集**

另有200条未知标签的测试数据，保存在“ProSeqs\_Test.txt”文件中。每一行仅给定序列ID及其氨基酸序列，而每条序列的类别(1: 具有某种功能，0: 不具有该功能)未知，待建模识别。

**2. 实验目的**

本次实验目标如下：

**(1)** 灵活设计样本的特征向量，具备一定的特征工程能力。

**(2)** 利用机器学习分类算法，基于训练集构建分类器模型；

**(3)** 进而将构建好的分类器模型应用于测试集，给出全体未知标签样本的分类结果，即预测氨基酸序列为功能蛋白(1)或非功能蛋白(0)。

**3. 设计思路**

**(1)** 具体采用的机器学习算法不限，有条件也可以尝试深度学习模型，以预测效果最佳为目标，追求F1 score越高越好；可以尝试多种学习模型的集成。

**(2)** 需从每条序列中提取特征，表示为固定长度的向量。特征向量设计思路不限，可尝试如下：

(a) 氨基酸组分，即每种氨基酸在蛋白质序列中的比例，可以表示为一个20维的向量。

(b) 可以间隔取子序列，计算各个子序列的氨基酸组分。比如将某条蛋白质序列记为字符串pseq，则可分别计算子串pseq[ : : 2]和pseq[1: :2]的氨基酸组分。依次类推，可以尝试计算各种子串pseq[i: :k] (k=2,3,4,…; i=0,1,…,k-1)的氨基酸组分。此类特征统称为k-space氨基酸组分。

(c) 序列上连续的两个氨基酸称为二肽，且二肽的可能种类共计20\*20=400. 我们还可以考虑二肽组分。依次类推，三肽组分？…

**(3)** 针对设计好的特征向量，可能是高维的，则可尝试进行特征选择、数据变换、降维等特征工程相关预处理，实现方法不限。

**(4)** *拓展探索*：有资源有条件的同学甚至可以考虑预训练的蛋白质或自然语言模型对氨基酸序列进行编码表示，然后开展适当的特征工程，获得更高质量的蛋白质功能分类预测模型。

**4. 实验要求**

**(1)** 将预测结果保存在名为“preds.txt”的文本文件中，内容为200行， 每一行只有0 或者1，代表你的算法对测试数据的预测结果。预测数据顺序须与测试集“ProSeqs\_Test.txt”中的样本顺序保持一致。

**(2)** 将结果文件“preds.txt”以附件形式提交至学习通，将全部源代码文件压缩打包为 “src.zip”以第二个附件提交至学习通。另请注意：只需提交结果文件和源代码压缩文件，且不要对结果文件“preds.txt”进行压缩，也无需提交本次实验报告文件。

**(3)** 本次实验成绩评定采用竞赛机制，采用每位同学预测结果的F1 score作为本次实验成绩。