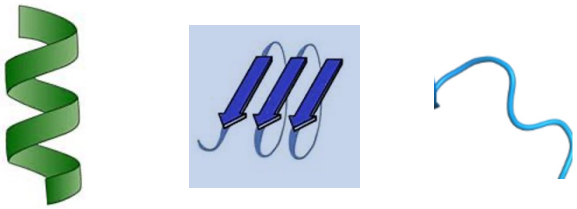
作业 1: 蛋白质二级结构预测

问题背景



Helix **Sheet** **Loop**

蛋白质的二级结构描述了其氨基酸链的局部空间排列，包括氢键形成的 α-螺旋 (Helix) 和 β-折叠 (Sheet)，以及非规则的 loop 区三种模式。

- 输入：蛋白质的氨基酸序列（字母表 20 的字符串）
- 输出：二级结构的类别（3 分类问题）:Helix, Sheet, 或者 Loop

数据集

共 3000 个蛋白质序列，可以做交叉验证。已提交到课程网站，asignment1_data.tar。

文件格式为 pickle（https://blog.csdn.net/Hardworking666/article/details/112754839），每个文件为一个样本，包含两个 key:

seq:作为输入的氨基酸序列

ssp:作为 label 的二级结构

seq 和 ssp 长度相同，每个位置一一对应。

训练和测试

- Pytorch data loader 构建数据流: Dataset and DataLoader (https://pytorch.org/tutorials/beginner/basics/data_tutorial.html)

- 构建训练 pipeline (模型，loss, 优化器等 https://pytorch.org/tutorials/beginner/introyt/trainingyt.html)
- 定义模型 (模型层数或大小不作要求)
    - MLP
    - 1D CNN (https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html) 或者 1D ResNet
- 定义 loss
    - 交叉墒 loss (https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html)
- 用 matplotlib 画出训练过程中 loss 曲线的变化
    - 可以先把每一个 step 的 loss 存入文件
    - Matplotlib 画图 https://machinelearningmastery.com/plotting-the-training-and-validation-loss-curves-for-the-transformer-model
- 计算在测试集上 Q3 accuracy,对比 MLP 和 CNN 的性能

提交

- 代码
- 文档：包括训练 loss 曲线图，预测性能分析，以及简要的实现说明（字数不作要求，可多可少）