当前课题：膜蛋白药物作用位点预测

1问题定义：药物小分子名字信息从Drugbank中提取。蛋白质序列中筛选膜蛋白，膜蛋白序列中有交互作用位点，交互的分子是药物分子的位点是正样本，其余是负样本。

2数据来源PDB和Drugbank

3

数据总量：

0.9-CDhit去冗余后：

0.3-CDhit去冗余后：

正负样本数

4处理步骤

4.1数据处理：

读取数据

分别读取3个数据文件：【seqdicts.pickle】、【labeldicts.pickle】、【seq\_3cutoff.pickle】

【seqdicts.pickle】中存放着的是所有序列的氨基酸信息，由蛋白质名字和氨基酸序列组成

【labeldicts.pickle】中存放着的是所有序列的标签信息，由蛋白质名字和标签序列组成

【seq\_3cutoff.pickle】中存放着的是经过去冗余后的蛋白质名字。

标识正负样本

窗口大小取值【31】，左右各15位氨基酸。

正样本是中间位置为作用点的序列，其余为负样本。

负样本中，去除掉左右有空格的样本，去除包含正样本位点的样本。

准备数据后存储（分为十折）

随机种子设置为【6】

分别对正负样本进行十折分配。

因为负样本通常是正样本的N倍，所以在每个十折中，会有1份正样本和N份负样本。

我们对它进行保存。

在

训练网络：

5当前结果