Project概述

宋文豪 2019012303 生命学院 范逍宇 2019013273 计算机系

问题

已有蛋白质和小分子亲和力预测算法未考虑细胞膜的限制条件，导致以下问题：

1. 亲和力靠前的结合位点可能位于跨膜区域，导致预测结果无意义，因为小分子药物只结合膜蛋白的胞外区域或者胞内区域，进而发挥作用。
2. 细胞膜可能影响小分子与蛋白质的结合，导致亲和力预测结果和实际不符

解决方案

引入膜蛋白结构信息和细胞膜信息训练模型。

1. 引入膜蛋白结构信息，可以筛选出空间上有效的结合位点，包括胞外区域与胞内区域。
2. 引入细胞膜信息，可以利用已有小分子与膜蛋白亲和力数据和结合位点数据，尝试寻找最优亲和力位点与细胞膜相对关系的pattern，建立模型。

目的

针对膜蛋白，得到具有更优生物学意义的亲和数据预测模型

1. 筛选出对胞内区或胞外区亲和性强的小分子
2. 结合细胞膜信息，如细胞膜位置和细胞膜对小分子亲和力的影响，获得更优的亲和力预测模型

数据库的建立

第一阶段只考虑一类跨膜蛋白，即 G蛋白偶联受体，因为这种蛋白具有良好的成药性，可以认为重要程度较高。目前的思路是使用 Bindbd 等数据库，对于其中 G 蛋白家族的每一种蛋白质，得到其跨膜片段信息和与其他小分子的结合信息，其中与小分子的结合信息包括小分子的种类，Ki，Kd 值等反映亲和力的信息。最后进行必要的数据清洗工作，即可完成第一阶段数据集的构建。

在完成第一阶段的迭代之后，再考虑其他跨膜蛋白。

算法

蛋白质的嵌入表示

目的是将蛋白质序列表示成特征向量，便于下一步的处理。事实上，这不是我们工作的研究方向，我们只是想得到一个比较好的蛋白质的嵌入表示，因此我们可以直接使用在大数据集上预训练的蛋白质表示学习模型，例如 Pro-Trans, ESM, MSA Transformer, ProtBERT-BFD, GearNet 等模型，然后根据我们的预测任务对模型进行微调，即可得到高质量的蛋白质的嵌入表示。

化合物小分子的嵌入表示

目的是将化合物小分子表示成特征向量，便于下一步的处理。同样的，我们可以使用在较大的化合物数据库上预训练的小分子表示学习模型，例如 GCN, AttentiveFP, GEM 等模型，然后根据我们的预测任务对模型进行微调，即可得到高质量的小分子的嵌入表示。

引入跨膜信息

我们之所以要引入蛋白质的跨膜信息，是基于这样的假设，即细胞膜与跨膜蛋白的相对空间位置会影响小分子与跨膜蛋白的结合位点、结合构象或亲和程度等指标。目前来看，这种假设有其现实依据。因此，蛋白质的跨膜情况是对亲和性预测有帮助的信息，我们预期加入这部分信息之后，模型可以对蛋白质和小分子的亲和性进行更精确的预测，这一点可以在实现时通过消融实验的方式验证。

具体地讲，跨膜蛋白被细胞膜分成几个片段，例如以下片段表示蛋白质的跨膜片段

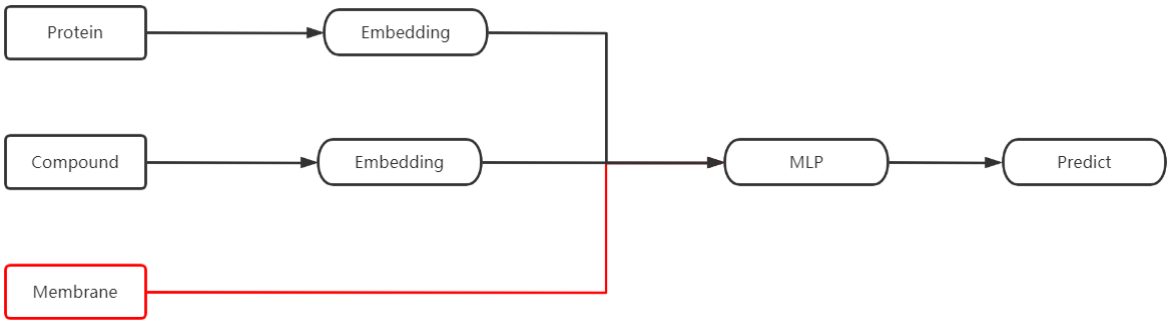
```
1(117-143), 2(153-175), 3(190-213), 4(230-251), 5(274-295, 6(344-365), 7(378-399)
```

将上述跨膜片段以某种方式进行编码，与上面的蛋白质和小分子的嵌入表示进行某种方式的特征融合，进而对蛋白质和小分子的亲和性进行预测。

我们预期模型会学到蛋白质与细胞膜的空间位置关系，这一点可以通过观察蛋白质与小分子的结合位点与跨膜片段的位置关系来验证，而蛋白质和小分子的结合位点可以通过观察蛋白质嵌入模型中蛋白质序列的注意力分数来大致判断（如果蛋白质嵌入模型是基于 Transformer 的话），可以认为注意力高的地方就是模型认为的可能的结合位点，当然这是后期的模型评估工作。

事实上，也可以显式地引入细胞膜的限制，例如如果我们假设蛋白质和小分子的结合位点不能在两层磷脂膜之间，那么我们可以直接 mask 掉蛋白质跨膜的片段，使用同样的流程训练模型，观察预测精度的变化来验证这种显式的约束是否有用。

Pipeline


Protein
Embedding
Compound
Embedding
MLP
Predict
Membrane