蛋白质-化合物亲和度与接触预测

1. 蛋白质分子结构表示
2. 一维氨基酸序列
3. 化学指纹
4. 二维原子坐标（？）
5. 二维图结构：原子作为顶点，化学键作为边
6. 三维转二维拓扑结构：使用数学工具将三维图结构转化为二维拓扑，可以用TopologyNet提供的方法
7. 三维转二维接触图：按照距离划分某两个原子是否接触，然后转为二维。接触图预测可以用RaptorX
8. 三维几何结构：原子坐标，体素（？）。三维几何预测可以用RaptorX。手性。
9. 物理化学信息：溶解度、解离常数等
10. 化合物分子结构表示
11. SMILES表示法
12. 化学指纹
13. 二维图结构。用RDKit工具计算
14. 三维原子坐标。用RDKit工具计算
15. 知识获取的难度

在蛋白质-化合物的物化性质预测中，难度在于蛋白质结构的复杂性、多样性与多变性。这要求模型能够有效获取高维知识，而且泛化能力足够强，稳健性足够强。

复杂性：要想从一维序列预测高维结构，考虑到蛋白质的折叠过程，难度很大。相对而言，二维、三维结构可能更容易获取；但是由于数据量变大了，学习可能更慢。

多样性：非同源蛋白质结构差异很大，从某一类蛋白质提取出的特征可能不适用于非同源的。而且蛋白质的三维结构哪怕是对一维序列的突变也很敏感。

多变性：蛋白质构象受到溶剂、温度等影响，需要足够稳健的方法才能识别出一大堆构象中的相同特征。

1. 论文的工作

该论文证实了不同表示方法各有优劣，且结合起来更加有效，并达到了SOTA。论文提出了连接concatenation与交叉作用cross interaction两种结合方法。我的评价：不同表示方法的优劣符合预期，而且根据复现的数据，这个模型根本没炼完，再来几百个epoch的话估计图结构就超过序列结构了。目前可以扩展的点包括：

1. 不同表示在亲和度与接触这两个任务上的表现。考虑到论文猜想接触图表示更难学，那么使用其他表示法可能更好/难学，效果更好/差
2. 在该论文的框架下加入其它表示，仍然使用连接与交叉作用。考虑到论文认为多个模态能够互通有无，那么可以验证加入更多表示法能否有所提升
3. 或许也可以比较一下其他表示中真实数据与工具生成数据对训练效果的影响。

关于预训练：根据复现结果，在图神经网络上做预训练会导致不收敛，我又看到补充材料里作者说小数据集100轮，大数据集练200轮就过拟合了，所以他们大数据集就练了1轮（评论：？真的练了吗）。

1. 数据集

目前的主要问题是我还没搞清楚论文提供的数据集的内容。目前计划是继续看一下，然后想办法把其中的蛋白质和化合物的序列表示提取出来。之后要么在已有的公开数据集里搜索是否已经有3维结构，要么用工具生成。但如果无法提取的话就要想其他办法了。

用到的数据集根据类型分为train、validate、test、unique compound、unique protein、both unique，其中protein包括序列与接触图、compound包括\_ver（？）与邻接矩阵，相互作用是\_inter。标签是\_inter\_exist（接触）与IC50（亲和度）。

论文中没有详细介绍序列数据的格式，查看发现是(591, 25, 40)数组，array([[16, 23, 9, 15, 10, 19, 9, 4, 20, 14, 10, 10, 5, 10, 19, 10,

14, 19, 19, 6, 18, 4, 4, 19, 4, 11, 4, 19, 14, 10, 18, 18,

4, 4, 18, 4, 18, 11, 10, 7],

……

[ 5, 18, 7, 7, 15, 5, 13, 14, 10, 17, 22, 19, 15, 14, 11, 8,

17, 10, 13, 4, 15, 16, 10, 11, 17, 18, 15, 7, 23, 23, 13, 14,

11, 5, 19, 14, 0, 0, 0, 0],

[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0],

[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0]], dtype=int64)

所以这个格式是什么？该数组在网络中首先被embedding，然后按行、按列过一遍RNN。

貌似pdb里也是这么存储的-\_-!，但这样一来我就不好逆向了，依然行不通。

--更新

已经打算重新生成数据，因此不能再使用现有的模型，要重新写一个。从PDBBind网站下载2020版数据集（论文使用的是2019版）。该数据集包括了氨基酸序列、物理化学信息、原子坐标（今后可以考虑不使用真实的坐标），且体量上与原论文相当，均为~5000对蛋白质-化合物，已经不需要再加其他数据集

化合物（配体？）编码（我要改网络结构，拆分每种表示）：

1d：由于可能有环，化合物无法使用序列表示。最直观的序列表示法是SMILES文本。资料表明一般都是用指纹或者mol2vec/graph2vec向量。此处我打算直接使用SMILES。【完成】

2d：可以是接触图或图像或邻接矩阵（对应GCN）。选择了邻接矩阵【完成】

拓扑：可以是拓扑编码

3d：待定

蛋白质编码：

一维：氨基酸序列编码可以使用物化信息编成1+1维矩阵。但此处有一个问题，氨基酸序列太长怎么办，如果一刀切的话明显会丢失信息。对化合物同理。目前直接使用氨基酸序列，感觉没太大区别。【完成】

二维：接触图编码不可以使用RaptorX提供的服务，上传氨基酸序列然后下载接触图。可以自己跑。更改：由于没有时间，所以放弃了使用人工智能生产的数据实验的部分，因此RaptorX与各种其他模型都不再可用。现更改为分子动力学模拟+conan方案，实测发现模拟耗时远大于能够负担的量，因此放弃该方案，转而寻找现成数据集。

拓扑：可以使用TopologyNet，输入氨基酸序列，输出拓扑编码

三维：使用zernike描述符https://github.com/biocryst/biozernike，该描述符能够描述蛋白质的几何信息。【完成】

假的原子坐标可以使用ESMFold，网站提供的API据说有限制https://esmatlas.com/about#api，也可以下载esmfold v1自己跑。

亲和度标签选择-logKd/Ki，PDBBind只有这个指标，论文使用的IC50在数据集里不全【完成】

接触标签待定，已知可以选择接触面积、最短接触距离和氢键数，都需要通过软件计算

这些数据都生成完之后，统计并拆分出4类，然后每一类都打乱拆分成train、test……

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset\Category | Length | Protein | Ligand | Unique protein | Unique ligand |
| Train | 4532 | 3174 | 3460 |  |  |
| Validate | 147 | 119 | 120 |  |  |
| Unique Protein | 176 | 570 | 564 | 344 | 453 |
| Unique ligand | 291 |
| Both unique | 170 |
| sum | 5317 | 3863 | 4144 |  |  |

1. 模型架构

（1）连接

序列+图+原子坐标，三个模型的向量接到一起，加一个CNN

（2）交叉作用

使用某种矩阵运算将三个向量合到一起，但还是不知道计算公式

由于打算重新写模型，可以模块化+多进程

外围工程包括：

1. 命令行参数读取【已完成】
2. 数据获取，未完成
3. 日志【已完成】
4. 训练器，未完成
5. 性能检测，未完成

模型包括：

1. 框架，未完成
2. 损失函数，未完成
3. 评价指标【已完成】
4. 超参数，未完成
5. 部件集合，未完成
6. 数据加载器【已完成】
7. 训练/验证函数，未完成

训练任务包括：

1. 纯1d
2. 纯2d/接触图
3. 纯拓扑
4. 纯3d
5. 四个拼起来average/concat/cross interact
6. ……（如果有时间的话再说）

desiderata【拉】迫切需要之物

A Multi-modal Framework for Protein-ligand Affinity and Contact Prediction

蛋白质-配体亲和度与接触的多模态预测框架

https://github.com/openmm/pdbfixer