蛋白质超家族分类预测

1 背景介绍

在生物物种的领域，根据不同物种的生理特征的相似度，被被划分为“种、属、科”等不同的层级。类似的，在蛋白质中，也根据结构序列相似度以及蛋白质的功能进行了层级的分类。蛋白质结构分类 (SCOP) 数据库[1]就是基于这样的一种目的来对蛋白质进行了一种划分与聚合。如图1[2]所示，SCOP 的层级分类从高到低不断细分。这种层级分类可以更好的指导我们了解蛋白质之间的演化发展关系。

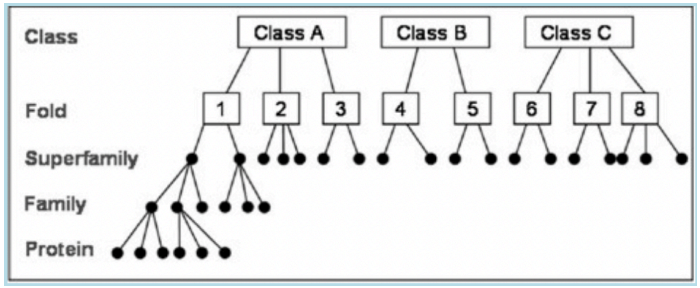

图 1: SCOP 层级分类

然而在另一方面，AlphaFold2的提出使得大量的新的蛋白质的结构被预测出来，但是这些蛋白质还没有对应的 SCOP 分类。于是考虑如何获取 AF2 新预测的蛋白质的 SCOP 分类就成为了我们下一步要解决的问题。

2 面临的挑战

现有的 SCOP 分类方法分为基于序列的方法和基于结构的方法，很多都依赖于数据库扫描以计算成对相似性。

对于基于序列的比对方法，通常速度比较快，在较短时间内就能够对比大量的蛋白质序列。然而因为远端同源蛋白质之间的序列并不相似，基于序列的同源性搜索技术依然存在一定的局限性，使得 PDB 中的很大一部分仍未被分类。

另一方面，蛋白质结构为发现远端同源蛋白质提供了更可靠的证据。尽管已经开发出许多启发式算法，但是基于结构的方法依然是非常耗时的。使用这些方法进行数据库扫描仍然不适用于一些对于时间效率要求比较高的任务。

总的来说，基于序列的算法虽然速度很快，但是结果往往并不准确；基于结构的算法虽然能够准确的预测出远程同源蛋白质，但是对于庞大的蛋白质数据库来说，效率却过于低下了。因此，我们有必要开发一种基于深度学习算法的，无须进行序列/结构对齐的分类算法。

[1] https://en.wikipedia.org/wiki/Structural$_C lassification_of_Proteins_database$

[2] https://slideplayer.com/slide/9700467/

3 问题描述

假设给定多个 AF2 预测的蛋白质结构数据，参赛者需要分析该结构数据[3][4]，利用深度学习方法将其分类到某一个 SCOP superfamily 中。参赛者的程序需要对每一个蛋白质输出 SCOP superfamily 的预测。例如，如果某个给定的蛋白质结构属于 $TIM\beta/\alpha-barrel$[5]（SCOP 编号：b.34.1），参赛者的模型需要输出 b.34.1。

3.1 输入输出描述

3.1.1 输入

我们将提供一个文件路径作为程序运行，该文件内有 N 行，每一行为需要预测的蛋白质结构文件的路径。每个蛋白质结构的原子数量 $L \leq 1024$。

3.1.2 输出

输出 N 行，分别为 N 个蛋白质的超家族分类预测结果。

3.2 样例

- 输入文件名示例: `/path/to/input`
- 文件内容示例:

```
/path/to/first/protein
/path/to/second/protein
```

- 输出结果示例:

```
a.4.5
g.82.1
```

3.3 评分指标

对于每个蛋白质 p_i，预测的超家族（Superfamily）结果 S_i, 我们会将对应超家族的每个代表蛋白质[6] $s \in S_i$ 与当前蛋白质结构做 `TM-align` [7], 取最高的 `TM-score` 作为分数。

N 个蛋白质的分数取平均值作为最终成绩。

$$score_i = Max(TM_{align}(p_i, s)),\ (s \in S_i)$$

$$score_{final} = \frac{\Sigma score_i}{N}$$

[3] Graphein

[4] Biopython

[5] The SCOP hierarchy

[6] 我们使用 SCOP 数据库公布的聚类结果作为参考，同一个中超家族（Superfamily）的剩余蛋白序列相似度小于 40%

[7] TM-align: A protein structure alignment algorithm using TM-score rotation matrix, https://zhanggroup.org/TM-align/

4 数据下载链接

4.1 AFDB

AlphaFold Protein Structure Database：https://alphafold.ebi.ac.uk/download

4.2 SCOPe

PDB-style files for SCOPe domains：https://scop.berkeley.edu/astral/pdbstyle/ver=2.08

4.3 PDB

RCSC PDB:
https://www.rcsb.org/

5 相关工作

- Graphein:

  https://graphein.ai/modules/graphein.protein.html

- ContactLib-ATT: a structure-based search engine for homologous proteins:

  https://pubmed.ncbi.nlm.nih.gov/35947567/

- TM-align:

  https://zhanggroup.org/TM-align/

6 提交方式

- 本题采用 docker 提交，关于 docker 提交的文档可以参考本文件。
- 对于测试 docker 镜像是否正确提交，以及测试运行结果，可以参考该仓库

  https://github.com/xwxztq/CBC2023_ProteinSuperfamilies