BT1501 计算生物学 2024-2025-2

Project 1. 冠状病毒的主蛋白酶序列和结构分析

冠状病毒是人类健康的一大威胁。冠装病毒完成复制周期需要主蛋白酶（main protease）来切割连在一起的多蛋白（polyproteins）。

1）请使用 NCBI 数据库找到所有基因组已经测序的冠状病毒科的病毒（Taxon: 11118）。请提交 Excel 文件 1, 给出病毒的 Name 和 Taxon ID。

2）对每种已测序的冠状病毒，获取它们的主蛋白酶的蛋白序列。请提交 Excel 文件 2，给出每种病毒主蛋白酶的序列（fasta 格式，header 请注明蛋白的 accession（如果有的话），病毒 Name 和 Taxon ID）。也可以直接提交 fasta 序列（所有 fasta 放入一个文件）。因为一些测序的冠状病毒未必做了翻译和注释，所以可能需要自己从哪些未做注释的物种中把主蛋白酶找出来（可以使用 tBLASTn，或者 nhmmer）。

3）对所有找到的主蛋白酶进行聚类处理。可以使用 cdhit，按 85% seq id 来聚类，总共可以分多少类？请提交聚类结果文件 3。

4）这些主蛋白酶中，哪些物种的主蛋白酶有结构？总共有多少个结构？这些结构的分辨率（resolution）和 R-free 的分布是怎样的？它们是由多少个课题组解析的？解析主蛋白酶数量位居前五的课题组分别是哪些？有多少个结构是主蛋白酶和药物分子结合的复合物（注意排除溶剂分子）？请提交文件 4，内容是主蛋白酶 accession+病毒 Name+Taxon ID + PDB IDs。

5）选择一个高分辨率的你认为具有代表性的主蛋白酶的结构，使用结构可视化软件（如 UCSF ChimeraX, PyMOL）进行观察。找出构成配体结合口袋的残基，把配体结合口袋存成图片文件，图片中请标注 PDB ID 以及口袋残基。请提交图片文件 5。

6）对 2）中得到的主蛋白酶序列进行多序列比对（可以使用 mafft）。在多序列比对结果中，把 5）中找到的残基所在列标注出来。请尝试根据这些标注的列对主蛋白酶序列进行重新聚类，可以使用 cdhit，按 85% seq id 来聚类。请提交聚类结果文件 6。得到的结果和 3）一样吗？这样聚类对药物开发有什么好处？

7）应对未来 COVID-X 爆发的策略之一开发广谱抗冠状病毒的药物，主蛋白酶是冠状病毒内部的蛋白，和冠状病毒表面的刺突蛋白（spike）相比，受到的进化压力要小得多，所以序列上相当保守，这为开发广谱抗冠状病毒药物提供了良好条件。如果你是相关决策人，根据以上的调查研究，你认为需要开发多少类冠状病毒主蛋白酶抑制剂就可以解决未来 COVID-X 爆发的问题。请提交文件 7 来阐述理由。