天然产物的靶点活性预测技术报告

廖一岩 周俊廷

September 9, 2024

Abstract

本报告意在介绍本队伍在天然产物靶点活性预测比赛中所采取的方法以及取得的阶段性成果。我们基于现有数据库，通过自设计的过滤器（filter）筛选天然产物的潜在靶点，获得了一批有用的、高致信的靶标分子结合靶点，并以此为 Baseline，复现了多项已有的蛋白质-ligand 活性预测方法，这些方法涵盖 Molecular Docking 以及 Deeplearning Methods，同时用我们的 Baseline 对这些方法做了初步的评估，并对实验结果优秀的工作进行加权投票，最终获得本次提交的蛋白质。我们的研究为天然产物靶点预测提供了新的思路和工具，相关的代码地址：https://github.com/zjtPKU/drug-port-prediction

1 介绍

天然产物是指来源于生物体的化学物质，包括植物、动物和微生物等。这些化合物具有丰富的化学结构和生物活性，许多现代药物的发现和开发都依赖于天然产物的提取或改造。天然产物在药物研发中具有重要意义，因为其独特的化学特性和生物活性常常使其成为新药开发的优质先导物。

免疫细胞，包括 T 细胞、B 细胞和自然杀伤细胞等，是参与免疫应答的关键细胞。它们在保护机体免受病原体侵害、识别和清除癌变细胞等方面发挥着至关重要的作用。靶向调控这些免疫细胞功能的特定分子或蛋白质，能够为治疗癌症、自身免疫疾病和感染性疾病提供潜在的治疗策略。发现和调节这些免疫细胞靶点有助于开发有效的治疗手段。

本次大赛的核心任务是利用先进的计算技术（如人工智能和计算生物学），预测天然药物分子与免疫细胞靶点的结合能力和选择性。参赛团队需要设计合理的算法和工作流程，对数百种天然药物分子与数千个免疫细胞靶点的相互作用进行评估，从而预测这些分子对靶点的结合力和选择性。

我们的研究基于现有数据库，并通过自设计的过滤器筛选天然产物的潜在靶点。我们获得了一批高致信性的靶标分子，并以此为基准，复现了多项已有的蛋白质-配体活性预测方法，包括分子对接和深度学习方法。通过对这些方法的初步评估，并对实验结果优秀的工作进行加权投票，我们最终提交了研究结果。这项研究为天然产物靶点预测提供了新的思路和工具。

2 工作内容

2.1 使用爬虫对于 Chembl 数据库已知靶点数据进行爬取

ChEMBL是一个手工管理的具有类似药物特性的生物活性分子数据库。它汇集了化学，生物活性和基因组数据，以帮助基因组信息转化为有效的新药。我们通过 smiles 获取天然产物分子的

chembl 序列号（通过相似性查找脚本 similarity_search.sh），然后利用 selenium 对网页可以搜索到的由实验验证的、已知的配体进行爬取，并清洗掉非人源的蛋白质，从而得到了高置信度的已知靶点蛋白质。

2.2 使用 SwissTargetPrediction 对天然产物靶点进行预测 [1]

SwissTargetPrediction是一个网络工具，自 2014 年以来在线，旨在预测最可能的蛋白质目标的小分子。我们使用其对于目标预测的 70 个天然产物的 smiles 进行靶点预测，并进行 uniprot 号提取与清洗(swiss_clean.py)、目标数据集映射(swiss_mapping.py)和输出格式整理(swiss_target_submission_formatted.py)。

2.3 基于机器学习的蛋白质-配体结合能力预测

蛋白质-配体结合能力预测是药物发现和设计中的关键环节。传统的实验方法虽然精确，但通常耗时且成本高昂。近年来，基于机器学习的方法已成为预测蛋白质-配体结合能力的一个重要工具。这些方法通过利用大量实验数据训练模型，能够在短时间内提供高效的预测结果，极大地加速了寻找靶点和药物研发的进程。

为此，我们选用了本领域的 SOTA——PLAPT：Protein-Ligand Binding Affinity Prediction Using Pretrained Transformers[2] 作为主要工具来完成蛋白质-配体结合能力的预测任务。

References

[1] A. Daina, O. Michielin, and V. Zoete. Swisstargetprediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Research*, 47(W1):W357–W364, 2019.

[2] Tyler Rose, Nicolò Monti, Navvye Anand, and Tianyu Shen. Plapt: Protein-ligand binding affinity prediction using pretrained transformers. *bioRxiv*, 2024.