

BT1051 计算生物学 2024-2025-2

Project 2. PurZ/PurZ0/PurA 同源序列搜索及 profile hmm 的建立

BT1051 | Project 2 | ddl 2025.4.1

Z 基因组，即 ZTCG 基因组，在基因组中 Z 完全取代了 A。目前已知几百种噬菌体具有 Z 基因组。Z 基因组的生物合成通路在 2021 年被两个研究团队发现 (Science, 2021, 372(6541):512-516. Science. 2021, 372(6541):516-520)。在 Z 基因组生物合成通路中，PurZ 是一个关键蛋白，它是嘌呤核苷合成通路中关键蛋白 PurA 的同源物。PurZ 和 PurA 最关键的区分，在 PurA 中参与催化的 Asp13 (按 E coli PurA 的序列编号) 被 Ser 所取代。后续工作中，一种 PurZ 的变体 PurZ0 被发现 (Nat Microbiol. 2023 Jul;8(7):1330-1338.)，它是 PurA 到 PurZ 进化的中间体。

- (1) 从文献/UniProt/PDB 出发，找出五条 PurZ 序列，五条 PurZ0 序列，五条 PurA 序列。分别提交名为 PurZ.fasta, PurZ0.fasta 和 PurA.fasta 的文件。
- (2) 请尝试在云桌面上本地安装 BLAST。按顺序记录下来你安装过程中的每一个命令并提交这个命令集合文件。下载 nr 库，从 PurZ0.fasta 中选一条 PurZ0 序列作为查询序列，使用本地安装的 BLASTp 搜索 nr 库来获取库中所有 PurZ0 序列 (注意不要在结果中混入 PurA 或 PurZ，阐述你如何实现这一点)。你可以找到多少条？分别来自什么物种？提交一个 Excel 文件，表头为 NCBI Accession, 物种, coverage (alignment length/length of query), seqid (aka pident), E-value, bit score, fasta 序列。为了获取 Excel 的内容，你选择了 BLAST 的哪种输出格式？
- (3) 下载 IMG/VR v4，使用本地 BLAST 的 makeblastdb 命令为 IMG/VR v4 建立索引。以 (2) 中用到的那条 PurZ0 序列作为查询序列，使用 BLASTp 搜索 IMG/VR v4 库，来获取库中所有的 PurZ0 序列 (注意不要在结果中混入 PurA 或 PurZ)。你可以找到多少条？分别来自什么物种？提交一个 Excel 文件，表头为 NCBI Accession, 物种, fasta 序列, E-value, bit score。根据你找到的 PurZ0 的序列条数，简要评价 NCBI nr 和 IMG/VR v4。
- (4) 对于 (2) 和 (3) 中找到的所有的 PurZ0 序列，使用 mafft 工具分别做多序列比对。在云桌面本地安装 hmmer3，使用 hmmer3 中的 hmmbuild 工具，为两个多序列比对生成 profile hmm 并使用 Skyline (<http://skylign.org/>) 来获得这两个 hmm 文件的可视化结果。提交名为 PurZ0_nr_hmm 和 PurZ0_imgvr4_hmm 可视化结果图片。比较这两个图片，你可以得出什么结论？
- (5) 对于 (2) 中找到的所有的 PurZ0 序列，先使用 cdhit 工具以 80% seqid 进行压缩，使用压缩后的序列文件重复 (4) 的所有操作。提交名为 PurZ0_nr_hmm_cdhit0.8 可视化结果图片。比较 PurZ0_nr_hmm 和 PurZ0_nr_hmm_cdhit0.8，你有什么结论？
- (6) 在 PurA.fasta 中选一条 PurA 作为查询序列，使用本地安装的 BLASTp 搜索 nr 库，使用输出结果的前 300 条，用 cdhit 以 80% seqid 进行压缩，用压缩后的序列文件重复 (4) 的所有操作。提交名为 PurA_nr_hmm_cdhit0.8 的可视化结果图片。对

PurZ0_nr_hmm_cdhit0.8 和 PurA_nr_hmm_cdhit0.8, 分别标出底物结合口袋残基的位置, 比较两者的异同。提交标记好的图片。

- (7) 对于 IMG/VR v4 数据库, 使用 jackhmmer 进行搜索其中的 PurZ0, 和(3)中 BLASTp 的结果进行比较。提交搜索命令和搜索结果。
- (8) 还有哪些方法可以用来搜索 IMG/VR v4 数据库中的 PurZ0? 尝试选取一种, 并提交搜索命令和搜索结果。这种方法和 BLASTp 以及 jackhmmer 有什么不同?