

序列比对与搜索 (Sequence Alignment and Search)



1. 检查序列比空中位罚分的影响。

请选择你感兴趣的两个蛋白质，它们的序列有 30%到 50%的相似性而且在序列比对中有空位。最简单找到这两个蛋白质的方法就是用一个你感兴趣的蛋白质对 UniProt/SwissProt 数据库做 BLAST 搜索 (在 [UNIPROT](#) 或者 [NCBI BLAST](#)), 根据最大的相似性分数选择第 2 个蛋白质，使之相似性在 30%到 50%之间。请检查查询的比对结果，确保它含有一些空位。

使用 Smith-Waterman 算法比对你选中的两个蛋白质 (可以用 EBI 的 [SSEARCH](#) 或者 Expasy 的 [SIM alignment](#) 工具)。使用不同的空位罚分 (罚分值取 4,8,16,32 和 64) 做序列比对。保证空位罚分(gap penalty)与空位拓展罚分 (gap extension penalty) 的比值是常数。

检查比对结果并描述增加空位罚分与重新安排空位对比对结果的影响。如果空位或者错配 (gaps or mismatches) 破坏了已知的功能或者结构模体 ([MyHits](#), [InterPro](#) or [Conserved Domains](#)), 指出哪一个比对具有最高的打分(bit score)或者质量。哪一个比对给出了最有生物功能意义的结果? 画出你的比对结果来支持你的结论。

2. 比较 Smith-Waterman, 无空位和空位的 BLAST 搜索。

通过查看在 [PROSITE](#) 数据库里面不同的模体， 找一个具有 50-100 个成员的蛋白质家族。如果你不能快速找到一个，你可以使用 pyruvate dehydrogenase E1 component 家族的一个成员作为你的黄金标准，它的酶分类号是 EC 1.2.4.1 (在 [Uniprot](#) 数据库里面搜索"name:1.2.4.1 AND reviewed:yes")。你可以使用阿尔法子单元(51 个例子在 UniProt/SwissProt 数据库里- "name:1.2.4.1 AND reviewed:yes AND name:alpha") 或者贝塔子单元(53 个例子)。你可以选择家族的任何一个成员作为你的查询序列。不要使用细菌的 pyruvate dehydrogenases 序列，它仅有一个具有 800 个氨基酸的子单元。你应该打印并保存你从 UniProt 数据库得到的“黄金标准” 家族列表。

选择家族的一个序列作为查询序列并使用 EBI [BLAST](#) 和 [SSEARCH](#) 进行无空位的 BLAST 搜索(UnGAPPED BLAST :请设置 "gap align" 参数为 "no"), 有空位的 BLAST (请保证"gap align" 参数设置为 "yes")和标准的 Smith-Waterman 算法搜索

UniProt/SwissProt 数据库。务必在每一次查询后你收集至少 100 条序列(或者 2 倍的家族成员数目)。同时保证查询过滤设置为 OFF (缺省值是 ON)。

现在,使用 Receiver-Operator Characteristic (ROC) 曲线来比较三种搜索结果。对于每一次搜索,在输出结果列表中每隔 10 个结果画一条线,统计在线上的所有真阳性和假阳性数目(累积的)。记住黄金标准决定了一个序列是否是真阳性还是假阳性。一直继续直到你已经收集了至少 50 个假阳性序列 (ROC-50 curve)。最后,对每一次搜索在一个二维的图上面用真阳性数对假阳性数进行作图。把三类曲线画在同一张图上面可能有助于你解释结果。

三类搜索具有完全一致的曲线形状吗?

有曲线比其它的要高一些么?如果是这样,哪一种搜索对你的蛋白质家族而言是最好的?

请画出 ROC 曲线来支持你的结论,最好给出 ROC 曲线下的面积值。

相关链接:

UNIPROT: <http://www.uniprot.org/>

SSEARCH:

<http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=fasta&program=ssearch&context=protein>

NCBI BLAST: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

MyHits: <http://myhits.isb-sib.ch/>

InterPro : <http://www.ebi.ac.uk/interpro/>

Conserved Domains: <http://www.ncbi.nlm.nih.gov/cdd>

SIM alignment: <http://web.expasy.org/sim/>

该题目来源于:

<http://biochem218.stanford.edu/Homework%2004.html>