“Linux 生物信息基础”课程

小组集体练习、讨论、交流

总 结 报 告

组：4　　次：4　　组长：陈奕晗　　执笔：陈奕晗

1. 时间：2021 年 4 月 15 日，15:00 ~ 17:00

2. 地点：王克桢 348

3. 人员：陈奕晗、邹济平、朱瑾煜、高培翔

4. 方式：线下讨论

5. 主题：

1) BLAST 数据库相似性搜索

2) 本小组综合课题选题方向

6. 内容

6.1 BLAST 数据库相似性搜索

6.1.1 NCBI BLAST

（1）搜索方式:

blastp 蛋白—蛋白

blastn 核酸—核酸

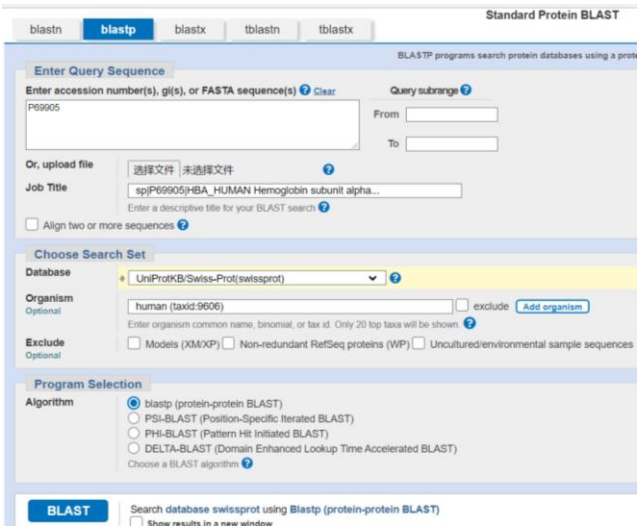
blastx 核酸—蛋白

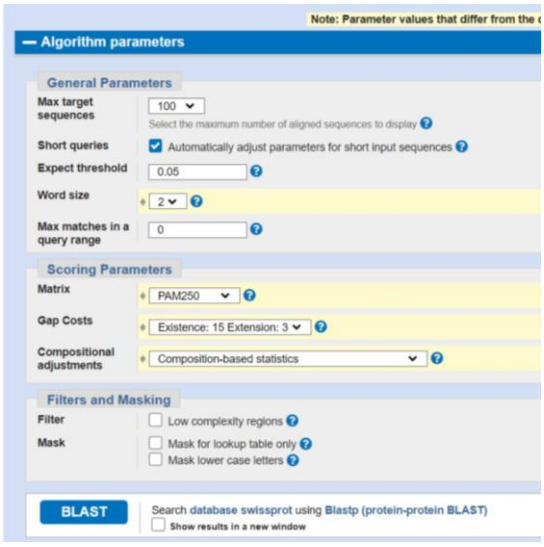
tblastn 蛋白—核酸

tblastx 编码核酸—编码核酸

（2）例题

1. 以人血红蛋白 alpha 亚基（HBA_HUMAN）为检测序列，用 BlastP 搜索 Swiss-Prot 数据库中 12 个珠蛋白

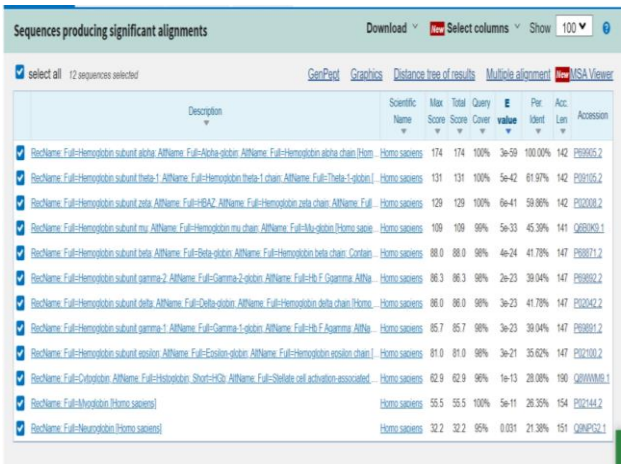




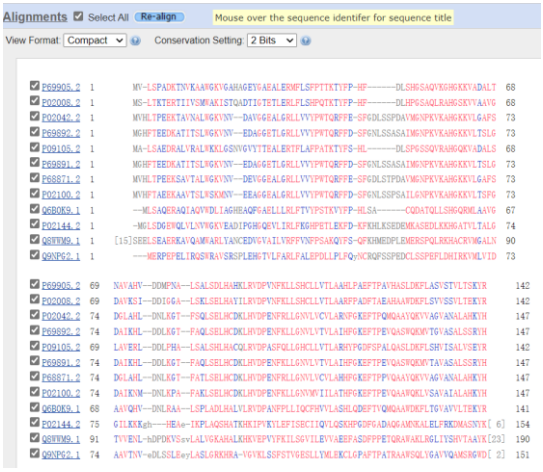
参数设置：

目标序列:	P69905
选择搜索的数据库:	UniProtKB/Swiss-Prot(swissprot)
选择搜索物种：	human(taxid:9606)
program selection (Algorithm)：	blastp (protein-protein BLAST)
Max target sequences:	100
word size:	2 （wordsize 越小，灵敏度越高）
Matrix:	PAM250 （PAM250 的灵敏度比 BLOSUM62 等其他选项都高）
Compositional adjustments:	Composition-based statistics

搜索结果

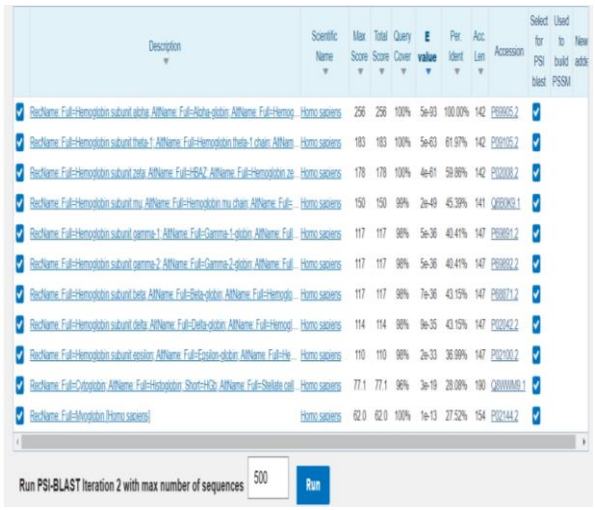


点击 Multiple alignment 可调用多序列比对程序

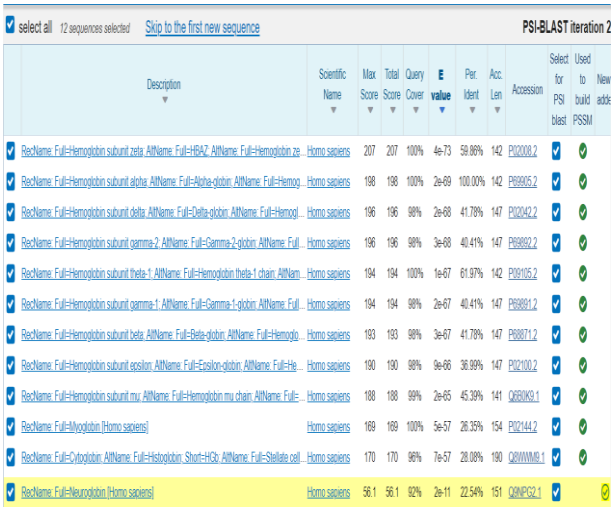


2. 以人血红蛋白 alpha 亚基（HBA_HUMAN）为检测序列，用 PSI-Blast 搜索 Swiss-Prot 数据库，找出人珠蛋白家族成员脑红蛋白（Neuroglobin）。

第一次搜索



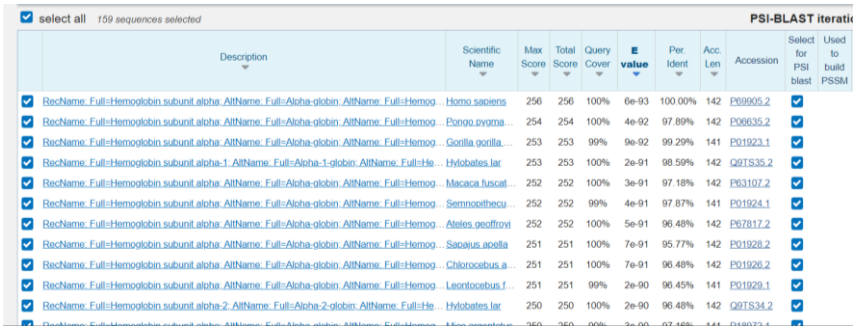
二次迭代搜索找到脑红蛋白 Neuroglobin



PSI-blast 为一种序列相似性搜索算法，可多次迭代得到更多结果

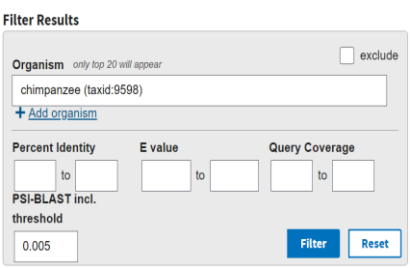
3. 以人血红蛋白 alpha 亚基（HBA_HUMAN）为检测序列，搜索 Swiss-Prot 数据库，找出灵长目动物（Primates）中所有 alpha 珠蛋白、找出黑猩猩（Chimpanzee）中所有珠蛋白、找出与 HBA_HUMAN 相同位点高于 95%（Identity>95%）的序列。

1.灵长目所有 alpha 珠蛋白：

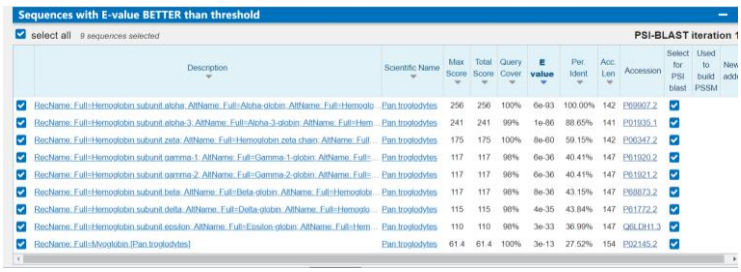


2.黑猩猩珠蛋白：

更改 Organism 为 chimpanzee

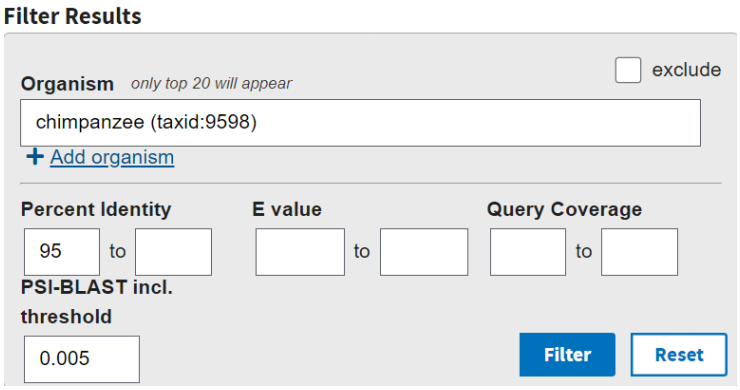


结果：

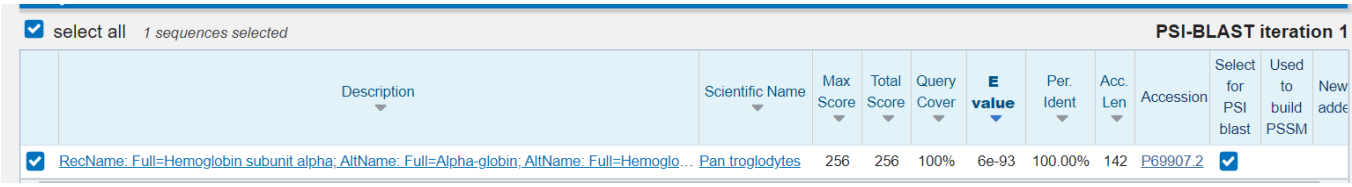


3.Identity>95%的序列

在 Filter Results 中的 Percent Identity 更改为“95 to　”筛选出一个序列。



结果：



6.1.2 linux 上的 BLAST 使用

1.　.ncbirc

cp /rd1/home/public/blast/.ncbirc .（在根目录下建立文件.ncbirc 进行数据库配置）

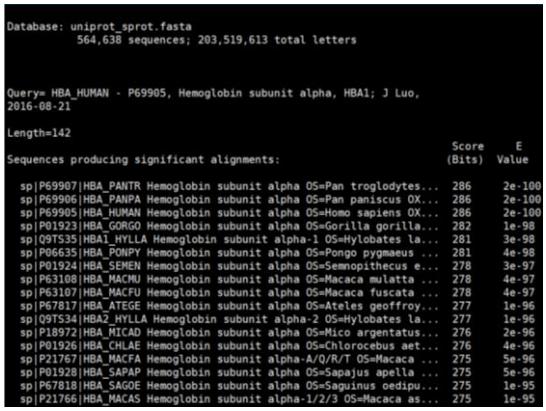
.ncbirc 文件内容：

Specifies the path where BLAST databases are installed

[BLAST]

BLASTDB=/rd1/home/blastdb

2.blastp 命令行使用



表格方式输出的一些参数

-evalue　　指定 E 值过滤假阳性（默认为 10，可选用 0.1 等更小的值）

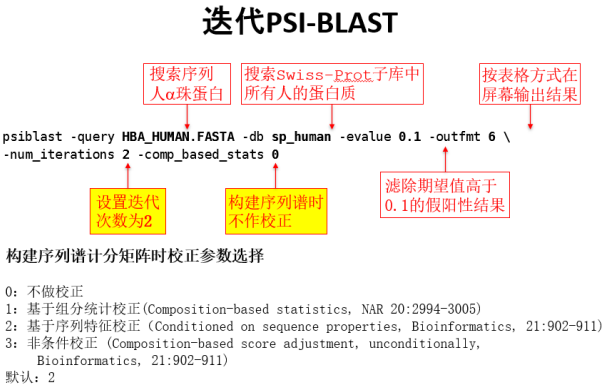
-outfmt　　-outfmt 6 只输出条目构成的表格，不显示详细比对情况

　　-outfmt 7 输出表格加一些表头信息

-word_size　设置种子序列字长

-matrix　　设置积分矩（如 PAM250）

3.迭代 PSI-BLAST



构建序列谱计分矩阵时校正参数选择

0：不做校正
1：基于组分统计校正(Composition-based statistics, NAR 20:2994-3005)
2：基于序列特征校正 (Conditioned on sequence properties, Bioinformatics, 21:902-911)
3：非条件校正 (Composition-based score adjustment, unconditionally,
　Bioinformatics, 21:902-911)
默认：2

-num_iterations　设置迭代次数

-comp_based_stats　矫正参数

结果：

```
# BLASTDB=/rd1/home/leb0/MyBlastDB
leb4a@bbt:~$ cd BLAST
leb4a@bbt:~/BLAST$ cd input
leb4a@bbt:~/BLAST/input$ blastp -query HBA_HUMAN.FASTA -db uniprot_sprot | less
leb4a@bbt:~/BLAST/input$ psiblast -query HBA_HUMAN.FASTA -db sp_human -evalue 0.1 -outfmt 6 \
> -num_iterations 2 -comp_based_stats 0
HBA_HUMAN       sp|P69905|HBA_HUMAN     100.000 142     0       0       1       142     1       142     1.13e-101       286
HBA_HUMAN       sp|P09105|HBAT_HUMAN    61.972  142     54      0       1       142     1       142     3.73e-59        179
HBA_HUMAN       sp|P02008|HBAZ_HUMAN    59.859  142     57      0       1       142     1       142     2.04e-56        172
HBA_HUMAN       sp|Q6B0K9|HBM_HUMAN     45.390  141     77      0       2       142     1       141     5.07e-40        130
HBA_HUMAN       sp|P68871|HBB_HUMAN     42.069  145     76      2       3       141     4       146     1.96e-33        114
HBA_HUMAN       sp|P02042|HBD_HUMAN     43.448  145     74      3       3       141     4       146     5.58e-33        113
HBA_HUMAN       sp|P69891|HBG1_HUMAN    42.553  141     73      3       7       141     8       146     3.18e-32        111
HBA_HUMAN       sp|P69892|HBG2_HUMAN    42.553  141     73      3       7       141     8       146     4.50e-32        110
HBA_HUMAN       sp|P02100|HBE_HUMAN     39.716  141     77      3       7       141     8       146     7.57e-28        100
HBA_HUMAN       sp|Q8WWM9|CYGB_HUMAN    28.859  149     97      2       3       142     19      167     6.14e-16        70.5
HBA_HUMAN       sp|P02144|MYG_HUMAN     26.351  148     103     1       1       142     1       148     3.21e-09        52.0
HBA_HUMAN       sp|P69905|HBA_HUMAN     100.000 142     0       0       1       142     1       142     9.46e-69        203
HBA_HUMAN       sp|P02008|HBAZ_HUMAN    59.859  142     57      0       1       142     1       142     7.76e-68        201
HBA_HUMAN       sp|P09105|HBAT_HUMAN    61.972  142     54      0       1       142     1       142     6.07e-65        194
HBA_HUMAN       sp|P68871|HBB_HUMAN     42.069  145     76      2       3       141     4       146     2.04e-64        192
HBA_HUMAN       sp|P02042|HBD_HUMAN     42.069  145     76      2       3       141     4       146     2.90e-64        192
HBA_HUMAN       sp|P69892|HBG2_HUMAN    39.583  144     79      2       4       141     5       146     1.37e-62        188
HBA_HUMAN       sp|P69891|HBG1_HUMAN    39.583  144     79      2       4       141     5       146     1.95e-62        187
HBA_HUMAN       sp|Q6B0K9|HBM_HUMAN     45.390  141     77      0       2       142     1       141     9.25e-62        186
HBA_HUMAN       sp|P02100|HBE_HUMAN     36.806  144     83      2       4       141     5       146     1.12e-61        186
HBA_HUMAN       sp|P02144|MYG_HUMAN     26.351  148     103     1       1       142     1       148     5.14e-57        174
HBA_HUMAN       sp|Q8WWM9|CYGB_HUMAN    28.859  149     97      2       3       142     19      167     3.70e-55        171
HBA_HUMAN       sp|Q9NPG2|NGB_HUMAN     22.535  142     99      3       7       138     5       145     3.94e-11        57.0
leb4a@bbt:~/BLAST/input$
```

最后一行即为第二次搜索得到序列

4.从 fasta 文件建立可用于检索的数据库文件（包括索引）,得到一个自定义搜索库

makeblastdb -dbtype<nucl/prot> -in input_file [-input_type fasta] -out dbname

6.2 本小组综合课题选题方向

目前倾向于做植物开花关键调控因子的分子演化。

7. 问题：无

8. 建议：