

尽管 blast 有了升级版，但是很多老服务器都更不上时代，而自己安装软件又特别繁琐，所以干脆又熟悉了一下 blast 的用法

1，

### 格式化数据库

输入：formatdb -i databasename -p F -o T

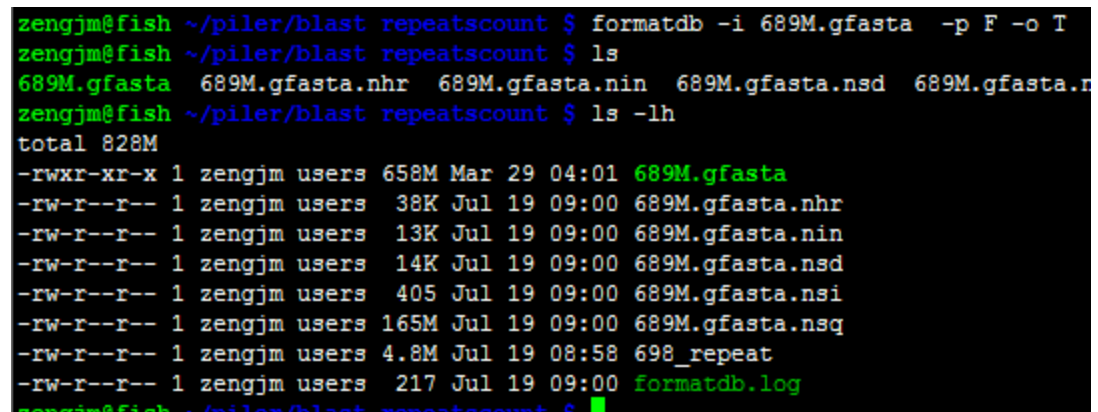
databasename 表示自己选择的数据库（最好使用绝对路径）

-i input file 参数用于指定需要格式的数据库

-p type of file 用于指定文件类型，T 为蛋白质，F 为核酸，默认为 T

-o parse options 用于指定是否解析序列 ID 并创建索引 T 为创建，F 为不创建，默认为 F。如果不用 T，会提示[NULL\_Caption] WARNING: "inputseq": Could not find index files for database "databasename"

可以输入 formatdb - 来获取相关参数的解释和帮助。



```
zengjm@fish ~/piler/blast repeatscount $ formatdb -i 689M.gfasta -p F -o T
zengjm@fish ~/piler/blast repeatscount $ ls
689M.gfasta 689M.gfasta.nhr 689M.gfasta.nin 689M.gfasta.nsd 689M.gfasta.nsi
zengjm@fish ~/piler/blast repeatscount $ ls -lh
total 828M
-rwxr-xr-x 1 zengjm users 658M Mar 29 04:01 689M.gfasta
-rw-r--r-- 1 zengjm users 38K Jul 19 09:00 689M.gfasta.nhr
-rw-r--r-- 1 zengjm users 13K Jul 19 09:00 689M.gfasta.nin
-rw-r--r-- 1 zengjm users 14K Jul 19 09:00 689M.gfasta.nsd
-rw-r--r-- 1 zengjm users 405 Jul 19 09:00 689M.gfasta.nsi
-rw-r--r-- 1 zengjm users 165M Jul 19 09:00 689M.gfasta.nsq
-rw-r--r-- 1 zengjm users 4.8M Jul 19 08:58 698_repeat
-rw-r--r-- 1 zengjm users 217 Jul 19 09:00 formatdb.log
```

对于我这个 658M 的基因组，只需要等待一分钟即可，当然，取决于你的服务器配置

2， 输入：blastall -p blastn -d databasename -i inputfile -o outputfile

-p program name 为需要使用的程序名

blastn 为核酸序列对比搜索

blastp 为蛋白质序列对比搜索

blastx 为用被翻译的核酸序列在蛋白质数据库中搜索

tblastn 为用蛋白质序列在 [核酸序列翻译后数据库] 中搜索

tblastx 为用翻译后的核酸序列 在 核酸序列翻译后数据库中搜索

可以输入 blastall - 来获取相关参数的解释和帮助。

-d databasename 指定所使用的数据库名称

-i inputfile 待搜索的序列文件（最好使用绝对路径）

-o outputfile 指定保存结果的文件（最好使用绝对路径）

除了这几个必须的参数外，还有几个重要的参数，主要是输出文件的格式控制，e-value 控制，cpu 控制。

-a: 是指计算时要用的 CPU 个数，我的机器有两个 CPU，所以用 -a 2，这样可以并行化进行计算，提高速度，当然你的计算机就一个 CPU，可以不用这个参数，系统默认值为 1，就是一个 CPU

-F: 是 filter 的简写，blastall 程序中有对简单的重复序列和低复杂度的一些 repeats 过滤调，默认是 T（注意以后的有几种参数就两个选项，T/F T 就是 true，真，你可以理解为打开该功能； F 就是 false，假，理解为关闭该功能）

-m: 有九种格式，我一般选择 -m 8 这个表格化的格式

-e: 是 Expectation value，期望值，默认是 10，一般得话设置成 e-10 就可以了！

```
21677 zengjm 20 0 1185m 545m 167m R 100 0.2 2:05.14 blastall -p blastn -d 689M.gfasta -i 698_repeat -o repeatscount_blast.result -m 8 -a 10 -e 0.01
```

可以查看内存消耗

```
zengjm@fish ~/piller/blast repeatscount $ blastall -p blastn -d 689M.gfasta -i 698_repeat -o repeatscount_blast.result -m 8 -a 10 -e 0.01
[blastall] WARNING: R=39: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Please
filtering options
[blastall] WARNING: R=49: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Plea
filtering options
[blastall] WARNING: R=128: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=353: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=371: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=667: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=669: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=683: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=760: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=830: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=856: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. Pl
filtering options
[blastall] WARNING: R=1216: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. F
filtering options
[blastall] WARNING: R=1268: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. F
filtering options
[blastall] WARNING: R=1270: Could not calculate ungapped Karlin-Altschul parameters due to an invalid query sequence or its translation. F
filtering options
```

弹出很多警告，说我的查询序列很多不合法

[illegible]

我看了看，我的查询序列都是自己找到的重复序列，看起来没啥子不合法的，反正

都是重复序列

不到半个小时就就跑完啦

结果如下,

```
zengjm@fish ~/piller/blast repeatscount $ head repeatscount_blast.result
R=4      2823      99.06      320        3        0        3      322      120271      119952      2e-173      611
R=4      2823      97.46      315        7        1        2      315      21414      21728      5e-156      553
R=4      2823      97.14      315        9        0        2      316      328988      328674      5e-156      553
R=4      2823      98.56      277        2        2        40     315      12593      12868      2e-140      502
R=4      2823      93.65      315        7        1        1      315      171699      172000      1e-134      482
R=4      2823      96.64      268        9        0       49     316      194346      194613      5e-128      460
R=4      2823      94.38      267        2        1       49     315      58346      58093      7e-118      426
R=4      2823      91.96      311       11        2        5      315      352387      352091      7e-118      426
R=4      2823      97.93      241        3        2       49     288      515778      515539      1e-116      422
R=4      2823      91.00      311       10        3        5      315      82510      82218      4e-110      400
```

整理结果可以看到

有 278 条序列没有对比到基因组上面。我查看了一下，都是短序列，估计是太短了

的原因

```
zengjm@fish ~/piler/blast repeatscount $ wc repeat_*
26943 323316 1535551 repeat_best_blast
26943 26943 204856 repeat_blast.ID
278 278 2100 repeat_not_blast.ID
```

在我的群里面共享了所有的代码及帖子内容，欢迎加群 201161227，生信菜鸟团！

<http://www.bio-info-trainee.com/?p=1>

线下交流-生物信息学

同时欢迎下载使用我的手机安卓 APP

<http://www.cutt.com/app/down/840375>