《Linux 生物信息基础》小组讨论总结报告

第 4 组，第 7 次讨论

组长：陈奕晗

执笔：高培翔

1 时间

2021 年 5 月 14 日，15:00 ~ 18:30

2 地点

泊星地咖啡厅

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

5.1 隐马氏模型软件包 HMMER 复习

5.2 Conda复习

5.3 MySQL复习

6 内容

6.1 隐马氏模型软件包 HMMER 复习

6.1.1 phmmer

`phmmer`使用一条蛋白质序列检索一个蛋白质序列库，类似于blastp，但使用了不同于blast的算法。HMMER的作者表示phmmer具有比blastp更好的性能。

示例：使用人HBA蛋白检索人类蛋白数据库，命令为`phmmer HBA_HUMAN.fasta sp_human.fasta > phmmer_HBA.txt`，生成一个txt文件，使用`less phmmer_HBA.txt`查看该文件，得到以下结果：

```
# phmmer :: search a protein sequence against a protein database
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             HBA_HUMAN.fasta
# target sequence database:        sp_human.fasta
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       sp|P69905|HBA_HUMAN  [L=142]
Description: Hemoglobin subunit alpha OS=Homo sapiens OX=9606 GN=HBA1 PE=1 SV=2
Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence             Description
    ------- ------ -----    ------- ------ -----   ---- --  --------             -----------
    5.6e-93  308.4   3.5    6.3e-93  308.2   3.5    1.0  1  sp|P69905|HBA_HUMAN   Hemoglobin subunit alpha OS=Homo sapien
    2.1e-54  183.5   0.7    2.4e-54  183.4   0.7    1.0  1  sp|P09105|HBAT_HUMAN  Hemoglobin subunit theta-1 OS=Homo sapi
      4e-52  176.2   0.3    4.4e-52  176.1   0.3    1.0  1  sp|P02008|HBAZ_HUMAN  Hemoglobin subunit zeta OS=Homo sapiens
    5.6e-37  127.2   0.1    6.1e-37  127.1   0.1    1.0  1  sp|Q6B0K9|HBM_HUMAN   Hemoglobin subunit mu OS=Homo sapiens O
    1.9e-30  106.0   0.2    2.2e-30  105.9   0.2    1.0  1  sp|P68871|HBB_HUMAN   Hemoglobin subunit beta OS=Homo sapiens
    7.3e-30  104.2   0.1    8.3e-30  104.0   0.1    1.0  1  sp|P02042|HBD_HUMAN   Hemoglobin subunit delta OS=Homo sapien
    8.8e-30  103.9   0.1    1.1e-29  103.6   0.1    1.1  1  sp|P69891|HBG1_HUMAN  Hemoglobin subunit gamma-1 OS=Homo sapi
    1.2e-29  103.4   0.1    1.5e-29  103.2   0.1    1.1  1  sp|P69892|HBG2_HUMAN  Hemoglobin subunit gamma-2 OS=Homo sapi
    6.5e-26   91.4   0.2    8.1e-26   91.1   0.2    1.1  1  sp|P02100|HBE_HUMAN   Hemoglobin subunit epsilon OS=Homo sapi
    1.1e-14   55.0   0.0    1.5e-14   54.6   0.0    1.2  1  sp|Q8WWM9|CYGB_HUMAN  Cytoglobin OS=Homo sapiens OX=9606 GN=C
    6.4e-08   33.2   0.1    7.2e-08   33.0   0.1    1.1  1  sp|P02144|MYG_HUMAN   Myoglobin OS=Homo sapiens OX=9606 GN=MB
  ------ inclusion threshold ------
      0.022   15.2   0.0      0.034   14.6   0.0    1.3  1  sp|Q9NPG2|NGB_HUMAN   Neuroglobin OS=Homo sapiens OX=9606 GN=


Domain annotation for each sequence (and alignments):
>> sp|P69905|HBA_HUMAN  Hemoglobin subunit alpha OS=Homo sapiens OX=9606 GN=HBA1 PE=1 SV=2
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  308.2   3.5   3.7e-96   6.3e-93       1     142 []       1     142 []       1     142 [] 1.00

  Alignments for each domain:
  == domain 1  score: 308.2 bits;  conditional E-value: 3.7e-96
  sp|P69905|HBA_HUMAN   1 mvlspadktnvkaawgkvgahageygaealermflsfpttktyfphfdlshgsaqvkghgkkvadaltnavahvddmpnalsalsdlhah 90
phmmer_HBA.txt
```

可以看到，获得12个相似结果，前11个蛋白与输入的HBA序列相同或相似度较高，E值很小；第12位是neuroglobin，与HBA相似度较低且E值比较大。这与前几次课程使用blastp获得的结果相似。

6.1.2 jackhmmer

`jackhmmer`使用一条蛋白序列迭代式地检索一个蛋白质序列库，类似于PSI-BLAST，但使用了不同于blast的算法。HMMER的作者表示jackhmmer具有比PSI-BLAST更好的性能。

示例：使用人HBA蛋白检索人类蛋白数据库，以E值0.01作为筛选阈值，最多迭代5次，命令为`jackhmmer -N 5 -E 0.01 HBA_HUMAN.fasta sp_human.fasta > jackhmmer_HBA.txt`，使用`less jackhmmer_HBA.txt`查看结果如下：

```
@@
@@ Round:                  3
@@ Included in MSA:        13 subsequences (query + 12 subseqs from 12 targets)
@@ Model size:             142 positions
@@

Scores for complete sequences (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence             Description
    ------- ------ -----    ------- ------ -----   ---- --  --------             -----------
      8e-63  211.0   0.4    8.8e-63  210.9   0.4    1.0  1  sp|P69905|HBA_HUMAN    Hemoglobin subunit alpha OS=Homo sapie
    3.4e-61  205.8   0.1    3.7e-61  205.6   0.1    1.0  1  sp|P02008|HBAZ_HUMAN   Hemoglobin subunit zeta OS=Homo sapien
      1e-59  200.9   0.0    1.2e-59  200.7   0.0    1.0  1  sp|P69891|HBG1_HUMAN   Hemoglobin subunit gamma-1 OS=Homo sap
    1.5e-59  200.4   0.0    1.7e-59  200.2   0.0    1.0  1  sp|P69892|HBG2_HUMAN   Hemoglobin subunit gamma-2 OS=Homo sap
    3.8e-58  195.9   0.2    4.5e-58  195.6   0.2    1.0  1  sp|P02100|HBE_HUMAN    Hemoglobin subunit epsilon OS=Homo sap
    4.6e-58  195.6   0.0    5.5e-58  195.4   0.0    1.0  1  sp|P02042|HBD_HUMAN    Hemoglobin subunit delta OS=Homo sapie
    5.1e-58  195.4   0.0    6.1e-58  195.2   0.0    1.0  1  sp|P68871|HBB_HUMAN    Hemoglobin subunit beta OS=Homo sapien
    2.4e-57  193.3   0.1    2.7e-57  193.1   0.1    1.0  1  sp|P09105|HBAT_HUMAN   Hemoglobin subunit theta-1 OS=Homo sap
      1e-55  188.0   0.8    1.1e-55  187.9   0.8    1.0  1  sp|Q6B0K9|HBM_HUMAN    Hemoglobin subunit mu OS=Homo sapiens
    7.2e-51  172.3   0.3    8.3e-51  172.1   0.3    1.0  1  sp|Q8WWM9|CYGB_HUMAN   Cytoglobin OS=Homo sapiens OX=9606 GN=
    7.5e-51  172.2   0.0    8.3e-51  172.1   0.0    1.0  1  sp|P02144|MYG_HUMAN    Myoglobin OS=Homo sapiens OX=9606 GN=M
    2.7e-39  134.8   0.0    3.2e-39  134.5   0.0    1.0  1  sp|Q9NPG2|NGB_HUMAN    Neuroglobin OS=Homo sapiens OX=9606 GN


Domain annotation for each sequence (and alignments):
>> sp|P69905|HBA_HUMAN  Hemoglobin subunit alpha OS=Homo sapiens OX=9606 GN=HBA1 PE=1 SV=2
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  210.9   0.4   5.2e-66   8.8e-63       1     142 []       1     142 []       1     142 [] 1.00
```

可以发现，迭代3次后结果已经收敛，包含12个检索结果，均具有较高的得分和很小的E值。这与前几次课程使用PSI-BLAST的结果相似。

6.1.3 hmmpress

`hmmpress`用于生成一个结构域模型数据库，里面包含多个hmm格式的模型文件。`hmmpress`的实质是对包含多个hmm的文本文件进行压缩，压缩后生成的文件可被`hmmscan`识别并用作检索的库。

`hmmpress`需要输入一个包含多个hmm的文本文件，可以使用`cat`创建，例如`cat globins4.hmm fn3.hmm Pkinase.hmm > minifam`，然后处理这个文本文件，使用命令`hmmpress minifam`，将会生成四个形如minifam.h3*格式的文件，它们可以作为`hmmscan`的被检索库。

```
(base) <leb4d@bbt> 15:51:37 ~/0424/HMMER
$ cat globins4.hmm fn3.hmm Pkinase.hmm > minifam
(base) <leb4d@bbt> 15:52:07 ~/0424/HMMER
$ hmmpress minifam
Working...    done.
Pressed and indexed 3 HMMs (3 names and 2 accessions).
Models pressed into binary file:   minifam.h3m
SSI index for binary model file:   minifam.h3i
Profiles (MSV part) pressed into:  minifam.h3f
Profiles (remainder) pressed into: minifam.h3p
```

6.1.4 hmmscan

`hmmerscan`使用一条蛋白质序列检索一个结构域模型数据库，以查看该蛋白质包含哪些已知的结构域。结构域模型数据库已在上一步由`hmmpress`创建。使用命令`hmmscan 7LESS_DROME minifam > 7LESS_scan.txt`检索果蝇sevenless蛋白中含有globin，fn3，Pkinase中的哪几个结构域及其位置。使用`less 7LESS_scan.txt`查看结果：

```
# hmmscan :: search sequence(s) against a profile database
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             7LESS_DROME
# target HMM database:             minifam
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       7LESS_DROME  [L=2554]
Accession:   P13368
Description: RecName: Full=Protein sevenless;          EC=2.7.10.1;
Scores for complete sequence (score includes all domains):
   --- full sequence ---    --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model    Description
    ------- ------ -----    ------- ------ -----   ---- --  -------- -----------
    5.6e-57  178.0   0.4    3.5e-16   47.2   0.9    9.4  9  fn3      Fibronectin type III domain
    1.1e-43  137.2   0.0    1.7e-43  136.5   0.0    1.3  1  Pkinase  Protein kinase domain


Domain annotation for each model (and alignments):
>> fn3  Fibronectin type III domain
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   -1.3   0.0      0.33       0.5      61      74 ..     396     409 ..     395     411 .. 0.85
   2 !   40.7   0.0   2.6e-14   3.8e-14       2      84 ..     439     520 ..     437     521 .. 0.95
   3 !   14.4   0.0   4.1e-06   6.1e-06      13      85 ..     836     913 ..     826     914 .. 0.73
   4 !    5.1   0.0    0.0032    0.0048      10      36 ..    1209    1235 ..    1203    1259 .. 0.82
   5 !   24.3   0.0   3.4e-09     5e-09      14      80 ..    1313    1380 ..    1304    1386 .. 0.82
   6 ?    0.0   0.0      0.13      0.19      58      72 ..    1754    1768 ..    1739    1769 .. 0.89
   7 !   47.2   0.9   2.3e-16   3.5e-16       1      85 [.    1799    1890 ..    1799    1891 .. 0.91
   8 !   17.8   0.0   3.7e-07   5.5e-07       6      74 ..    1904    1966 ..    1901    1976 .. 0.90
   9 !   12.8   0.0   1.3e-05     2e-05       1      86 []    1993    2107 ..    1993    2107 .. 0.89

  Alignments for each domain:
  == domain 1  score: -1.3 bits;  conditional E-value: 0.33
                 EES--TT-EEEEEE CS
          fn3  61 ltgLepgteYefrV 74
                 l+ L p+t+Y+fr
```

可以发现，sevenless蛋白含有9个fn3结构域，1个Pkinase结构域，没有globin结构域。

6.2 Conda

Conda是一个开源的软件包管理系统和环境管理系统。很多软件可以通过Conda进行非常简便的安装。Conda会自动地软件的依赖关系，自动安装被依赖的软件，大大简化很多软件的安装流程，降低出现异常的概率。Conda的另一功能是设置虚拟的环境，用户可以在自定义的环境中安装特定的软件，并能够对环境本身进行自由的配置。

Conda最常用的命名如下：

`conda create --name envname [pkgname]`创建名为envname的虚拟环境[，并安装名为pkgname的包]。

`conda install -n envname pkgname=x.x`在环境envname内安装版本为x.x的pkgname软件

`conda activate envname`激活环境envname。只有当一个环境被激活后，安装在该环境下的软件才能使用。

`conda list -n envname`查看环境envname下安装的所有软件

`conda search pkgname`搜索名为pkgname的软件包

需要注意，conda可以从不同的channel获得软件包，其中生物学软件主要通过名为bioconda的channel分发。我们需要将bioconda加入到conda的频道列表中。此外，为了获得更快的速度，我们可以添加bioconda设在清华大学的镜像，加快下载。使用的命令为：

`conda config --add channels bioconda`

`conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/bioconda/`

在`conda install`和`conda search`命令中都可以使用-c参数指定来源的channel。

6.3 MySQL

本节课学习了一些MySQL最基本的操作，包括：

`musql -u usrname -p`登录数据库，此后提示符变为`mysql>`

`show databases;`查看当前用户拥有的所有数据库

`use dbname`进入名为dbname的数据库

`show tables;`查看当前数据库下的所有数据表

7 收获

1. 学习了几个新的HMMER工具包命令；
2. 学习了conda的使用；

3. 学习了一些基础的MySQL操作。