**真核生物基因组的基因分析和预测**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年级： | 2015 | 专业： | 生物信息 | 评分： |  |
| 学号： | 1530416014 | 姓名： | 赵飞洋 | 签名： |  |

**一、软硬件平台：**

1. 硬件平台：（硬件配置）

(1).CPU: 2.7 GHz Intel Core i7

(2).内存: 4 GB 1333 MHz DDR3

(3).硬盘:intel 545s 512G 固态

2. 系统平台：（操作系统及其版本号）

(1).Mac OS Sierra 10.12.6

3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）

(1).R语言3.3.2

4.数据库资源

（1）pubmed：<https://www.ncbi.nlm.nih.gov/pubmed/>

**二、实验方法：**

1、基因组核酸序列的获取:

(1)由任课教师提供【AP-genome-draft】。

2、创建本地 BLAST 数据库:

使用 makeblastdb 程序，对上述 FASTA 格式的基因组序列进行处理，建立本 地 BLAST 数据库。

3、已知蛋白序列的下载

根据基因组的物种来源，从 UniProt 数据库下载近缘物种已知蛋白序列， 以 FASTA 格式保存。

4、使用小型机上安装的blast

5、同源基因搜索:

5.1、使用 tblastn 程序，把已知蛋白质序列和基因组草图序列建立的本地 BLAST 数据库进行比对，注意参数设置(如 e-value 设为 0.00001，建议输出格式 6 或 7。  
5.2、编程处理 BLAST 比对结果文件，排除冗余项:

(1)不同物种的同一种蛋白在基因组上的匹配位置存在的重叠问题;

(2)同一物种的同一蛋白家族的不同成员白在基因组上的匹配位置存在的重叠 问题;

(3)同一个蛋白在基因组上的不同位置的高相似区域问题——是家族成员问题， 还是冗余匹配;

5.3、把去除冗余后的结果转成 GFF3 格式。

7、从头预测:

7.1、从网上搜索、下载并安装基因预测相关软件【至少 1 个】;

7.2、使用该软件对基因组序列进行基因预测分析，一方面保存预测基因编码的 多肽，另一方面将基因结构信息输出成 GFF3 格式;  
7.3、使用 blastp 程序对该预测基因与已知蛋白序列进行比对，以此来鉴别预测 的基因。

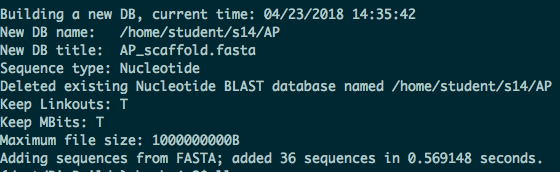
**三、试验结果**

1、基因组核酸序列的获取:

(1)由任课教师提供【AP-genome-draft】。

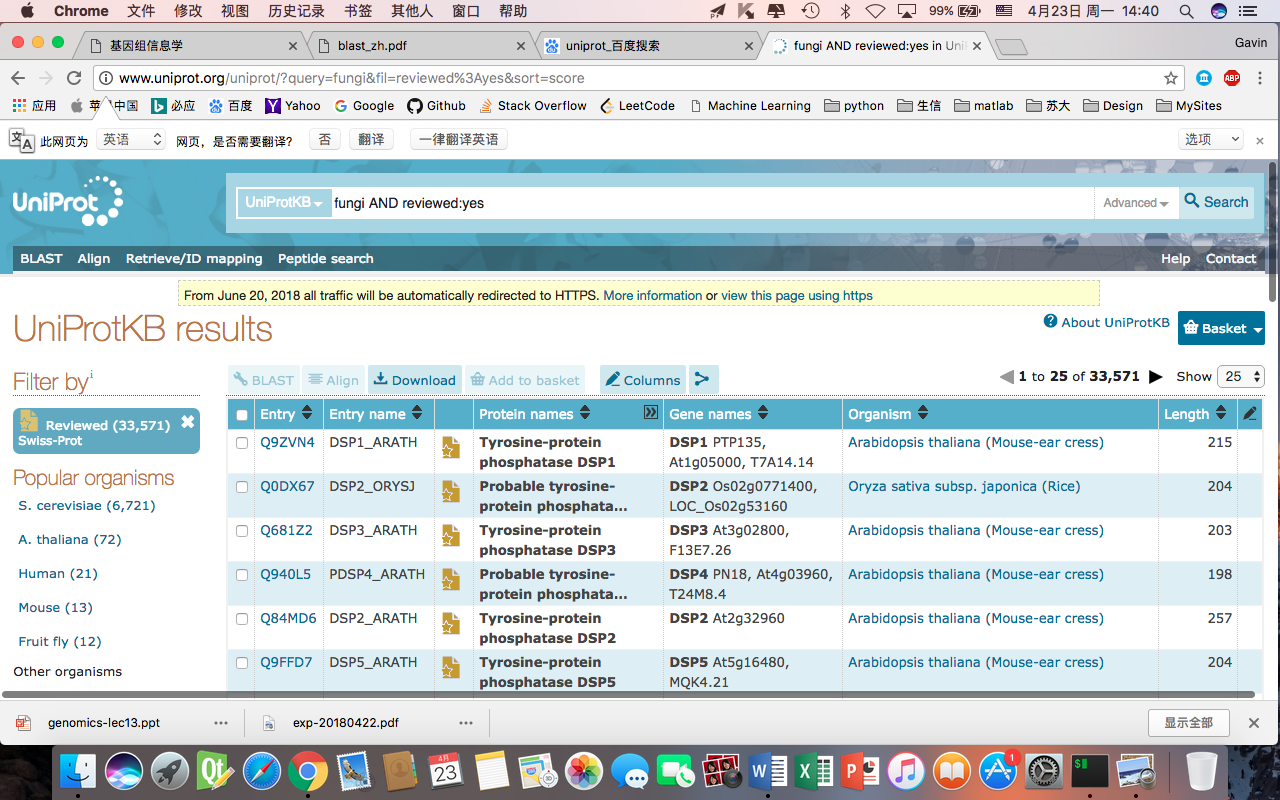
2、创建本地 BLAST 数据库:

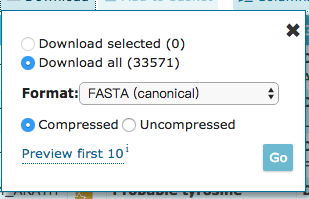
makeblastdb -in AP\_scaffold.fasta -dbtype nucl -parse\_seqids -out AP



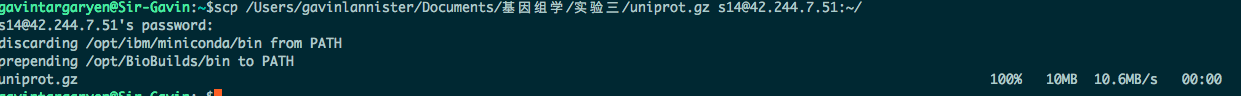
3、已知蛋白序列的下载

根据基因组的物种来源，从 UniProt 数据库下载近缘物种已知蛋白序列， 以 FASTA 格式保存。





4、使用小型机上安装的blast



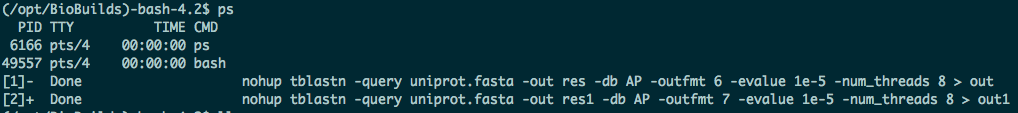
上传近缘物种序列至小型机

5、同源基因搜索:

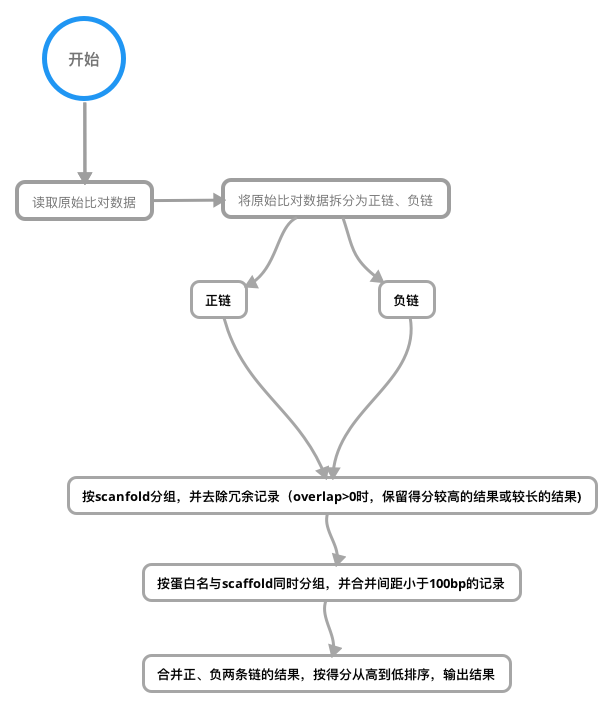
5.1、使用 tblastn 程序，把已知蛋白质序列和基因组草图序列建立的本地 BLAST 数据库进行比对，注意参数设置(如 e-value 设为 0.00001，建议输出格式 6 或 7。

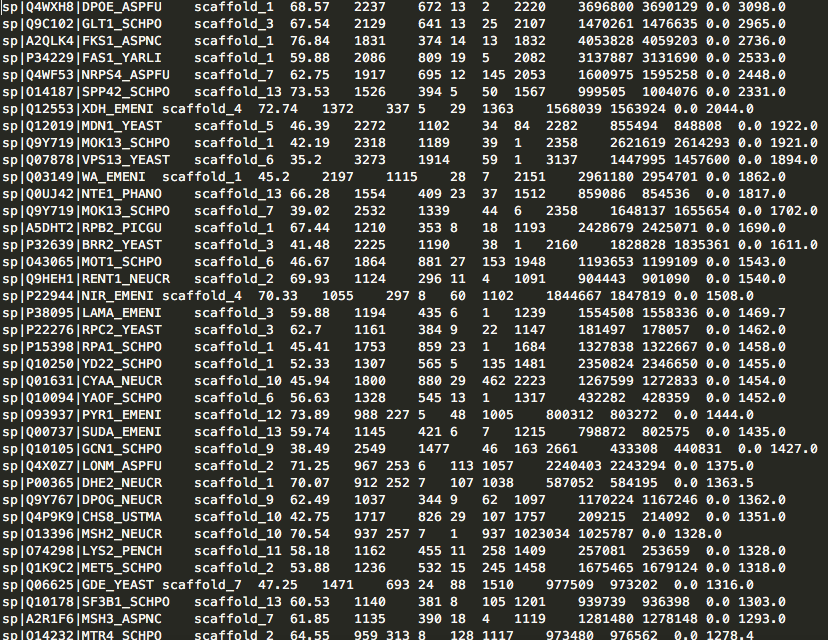
nohup tblastn -query uniprot.fasta -out res -db AP -outfmt 6 -evalue 1e-5 -num\_threads 8 > out &

nohup tblastn -query uniprot.fasta -out res1 -db AP -outfmt 7 -evalue 1e-5 -num\_threads 8 > out1 &



5.2、编程处理 BLAST 比对结果文件，排除冗余项:



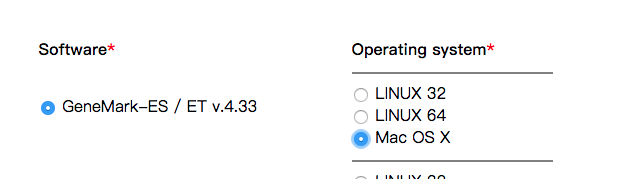


去除冗余后的结果，保留了4268条记录，处理脚本名为operate.py

7、从头预测:

7.1、从网上搜索、下载并安装基因预测相关软件【至少 1 个】;

a．下载GeneMark



b．移动key至工作目录下



c．安装perl依赖

sudo cpan YAML

sudo cpan Hash::Merge

sudo cpan Logger::Simple

sudo cpan Parallel::ForkManager

sudo cpan Getopt::Long

sudo cpan File::Spec

sudo cpan File::Path

sudo cpan Data::Dumper











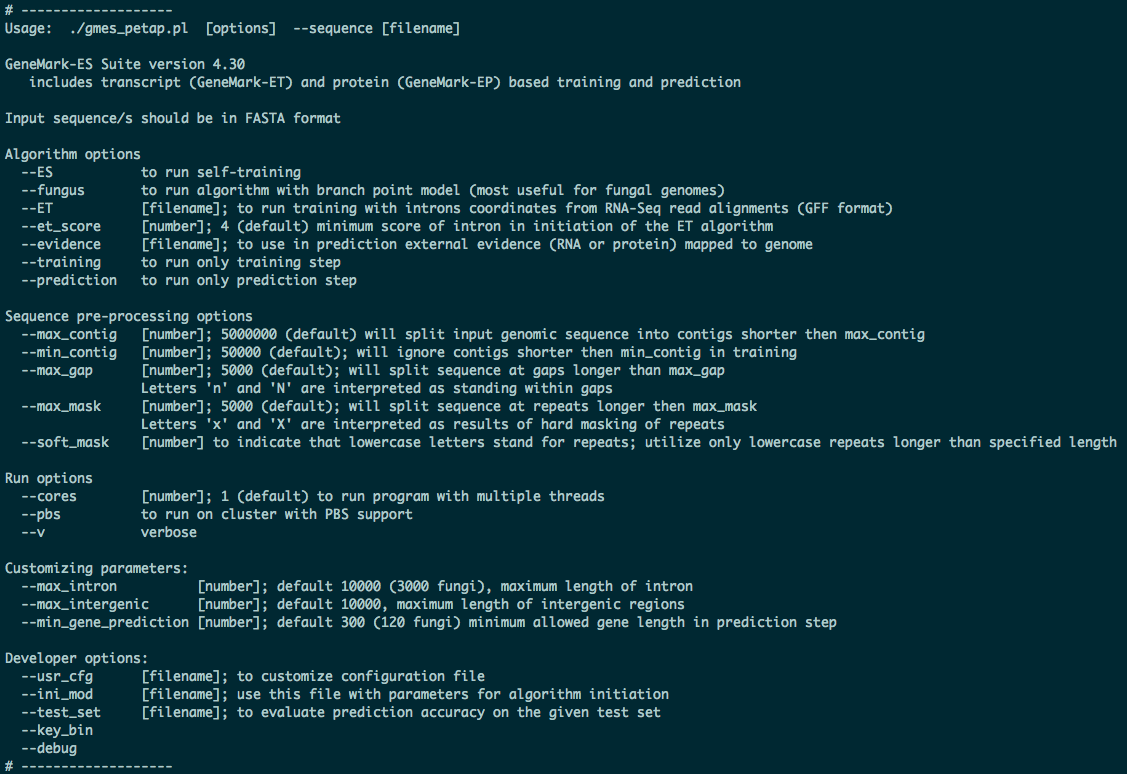






d．测试是否成功

./gmes\_petap.pl



e.发现mac版缺少—predict\_with参数，由于此工具为perl脚本工具，故在mac上试用Linux版本脚本，但执行命令依然有问题，为无法运行二进制文件

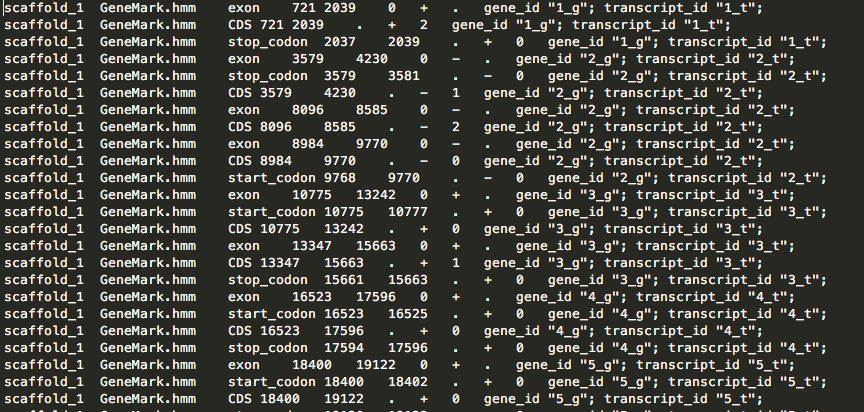


f.故将原来下载的mac版文件夹中的Gibbs3、gmhmme3、probuild三个二进制文件替换至linux版本文件夹下，运行成功

7.2、使用该软件对基因组序列进行基因预测分析

./gmes\_petap.pl --prediction --fungus --predict\_with ./heu\_dir/heu\_05\_gcode\_1\_gc\_38.mod --sequence ./AP\_scaffold.fasta





./get\_sequence\_from\_GTF.pl genemark.gtf AP\_scaffold.fasta



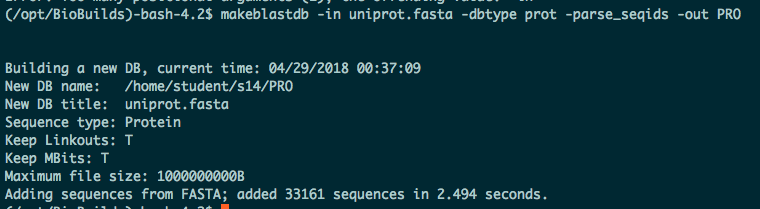




7.3、使用 blastp 程序对该预测基因与已知蛋白序列进行比对，以此来鉴别预测 的基因。

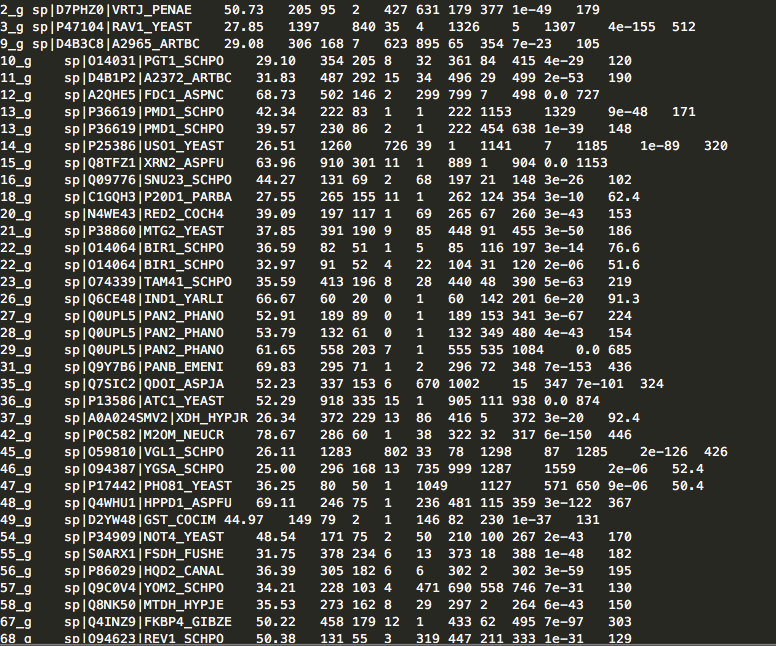
1. 建立已知蛋白数据库

makeblastdb ‐in uniprot.fasta ‐dbtype prot ‐parse\_seqids ‐out PRO



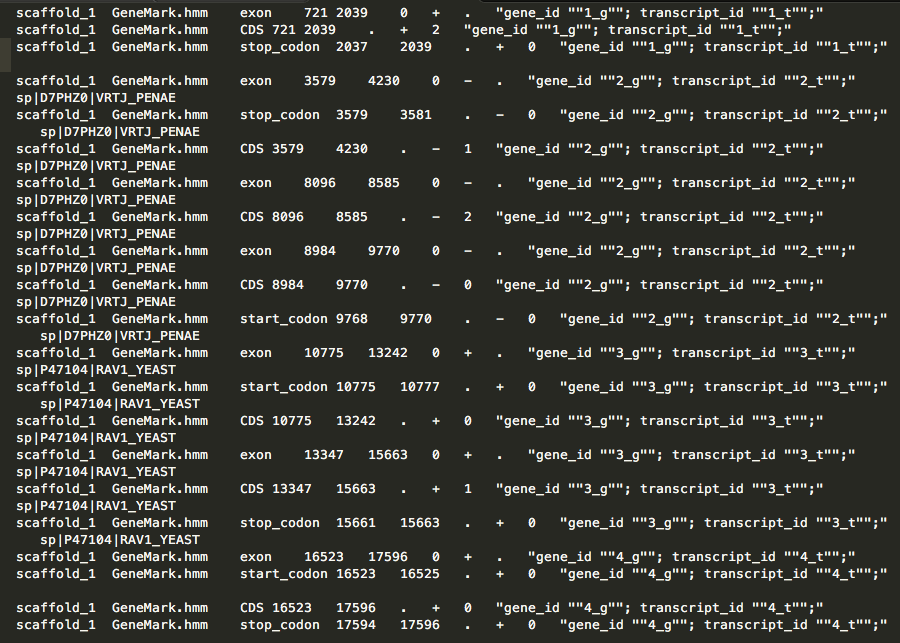
1. 使用blastp进行比对

nohup blastp -query prot\_seq.faa -out prores -db PRO -outfmt 6 -evalue 1e-5 -max\_target\_seqs 1 -num\_threads 8 > out &



blastp结果

c．编程在gtf中加入比对的query名，代码为genemark.py



**三、实验讨论：**

1.对blast比对结果文件进行去除冗余处理时，我分了三步，首先区分正、负链，之后将正、负链结果进行合并多外显子、去除冗余，这里出现了一个问题。如果先进行多外显子的合并，由于原数据中有大量冗余结果，一旦合并，会产生非常长的链，之后进行去冗余操作，这些长链会将本不该被去除的区域消除。如下图所示

合并前

合并后

由图中可知，原本下面的两段序列不会被消除，但是如果先进行合并处理，下面两条序列就会被当作冗余处理。所以先进行去冗余操作，再合并多外显子。

2.不同物种的同一蛋白在核酸序列相同位置匹配，说明此蛋白保守性较好；而不同蛋白在核算序列同一位置匹配上，则可能是此区域保守性较差，或者只是这些蛋白重复性较高。理论上讲，核算序列的某一位置应该只有一个蛋白与之匹配。

3.结果文件：

a．同源基因搜索

res：blast结果format-6

res1:blast结果format-7

result：去冗余并合并多外显子结果

b．从头预测

genemark.gtf、prot\_seq.faa、nuc\_seq.fna：从头预测原始结果

prores：blastp结果

proteinblast：添加query名的结果

c．脚本

operate.py：去冗余、合并多外显子

genemark.py:添加query名至blastp结果