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】。

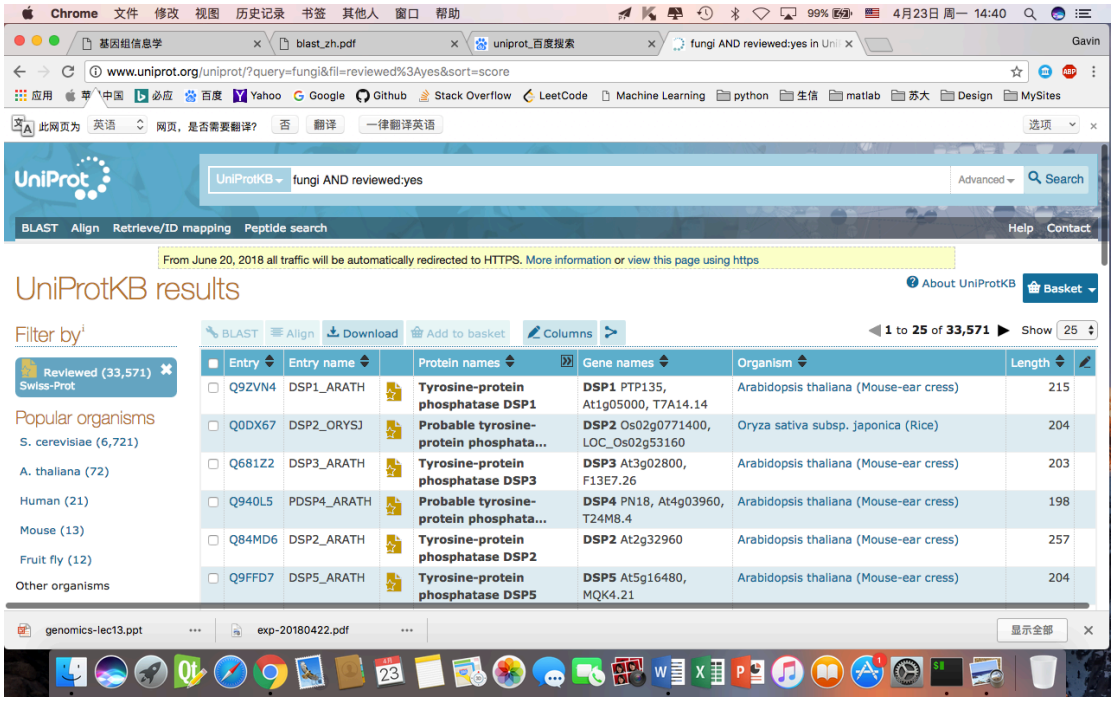
2、创建本地 BLAST 数据库:

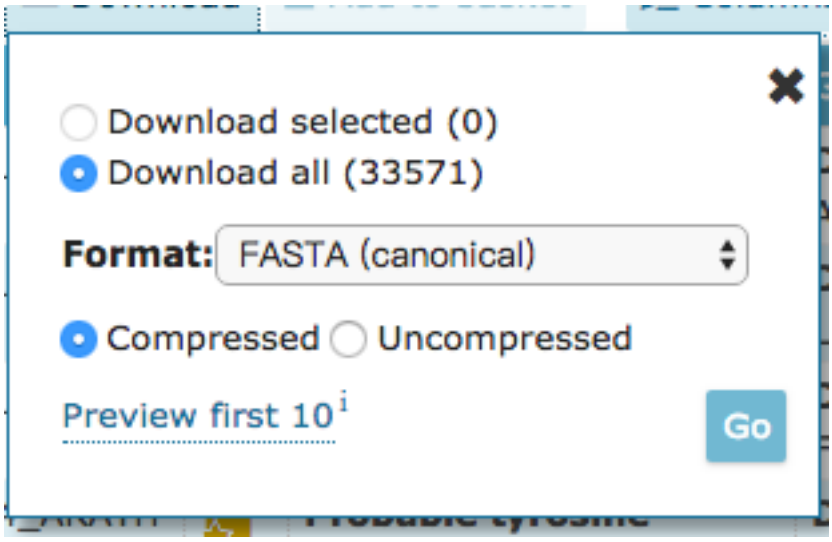
```
makeblastdb -in AP_scaffold.fasta -dbtype nucl -parse_seqids -out AP
```

```
Building a new DB, current time: 04/23/2018 14:35:42
New DB name:   /home/student/s14/AP
New DB title:  AP_scaffold.fasta
Sequence type: Nucleotide
Deleted existing Nucleotide BLAST database named /home/student/s14/AP
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 36 sequences in 0.569148 seconds.
```

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

```
gavintargaryen@Sir-Gavin:~$scp /Users/gavinlannister/Documents/基因组学/实验三/uniprot.gz s14@42.244.7.51:~/
s14@42.244.7.51's password:
discarding /opt/ibm/miniconda/bin from PATH
prepending /opt/BioBuilds/bin to PATH
uniprot.gz                                    100%   10MB  10.6MB/s   00:00
```

上传近缘物种序列至小型机

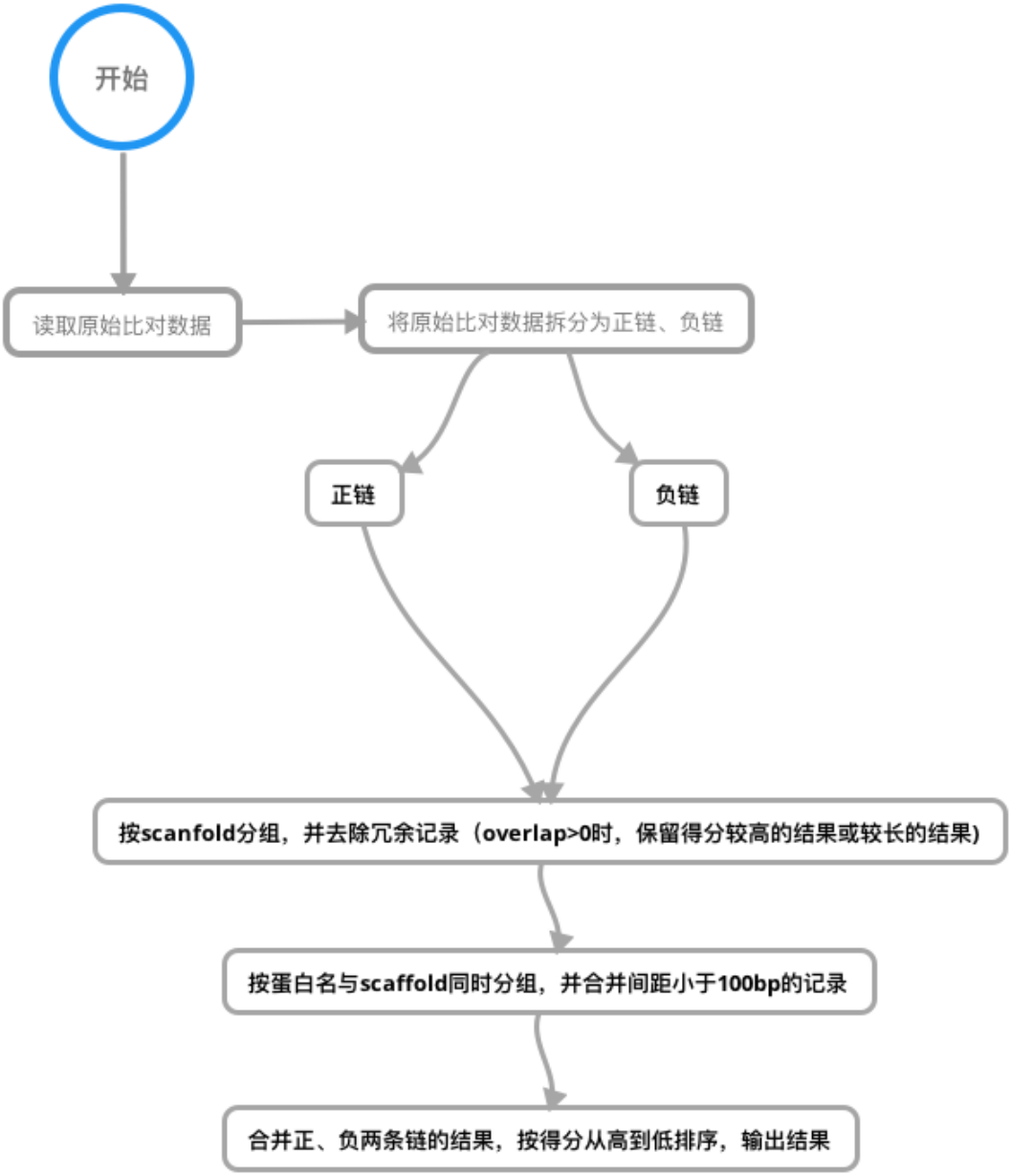
5、同源基因搜索:

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:


开始
读取原始比对数据
将原始比对数据拆分为正链、负链
正链
负链
按scanfold分组，并去除冗余记录（overlap>0时，保留得分较高的结果或较长的结果）
按蛋白名与scaffold同时分组，并合并间距小于100bp的记录
合并正、负两条链的结果，按得分从高到低排序，输出结果

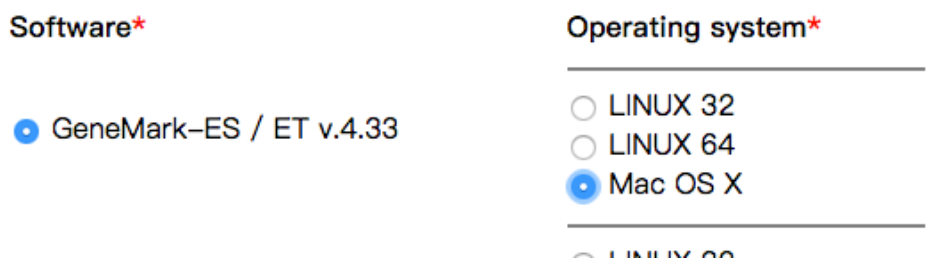
```
sp|Q4WXH8|DPOE_ASPFU    scaffold_1  68.57   2237    672 13  2   2220    3696800 3690129 0.0 3098.0
sp|Q9C102|GLT1_SCHPO    scaffold_3  67.54   2129    641 13  25  2107    1470261 1476635 0.0 2965.0
sp|A2QLK4|FKS1_ASPNC    scaffold_1  76.84   1831    374 14  13  1832    4053828 4059203 0.0 2736.0
sp|P34229|FAS1_YARLI    scaffold_1  59.88   2086    809 19  5   2082    3137887 3131690 0.0 2533.0
sp|Q4WF53|NRPS4_ASPFU   scaffold_7  62.75   1917    695 12  145 2053    1600975 1595258 0.0 2448.0
sp|O14187|SPP42_SCHPO   scaffold_13 73.53   1526    394 5   50  1567    999505  1004076 0.0 2331.0
sp|Q12553|XDH_EMENI scaffold_4  72.74   1372    337 5   29  1363    1568039 1563924 0.0 2044.0
sp|Q12019|MDN1_YEAST    scaffold_5  46.39   2272    1102    34  84  2282    855494  848808  0.0 1922.0
sp|Q9Y719|MOK13_SCHPO   scaffold_1  42.19   2318    1189    39  1   2358    2621619 2614293 0.0 1921.0
sp|Q07878|VPS13_YEAST   scaffold_6  35.2    3273    1914    59  1   3137    1447995 1457600 0.0 1894.0
sp|Q03149|WA_EMENI  scaffold_1  45.2    2197    1115    28  7   2151    2961180 2954701 0.0 1862.0
sp|Q0UJ42|NTE1_PHANO    scaffold_13 66.28   1554    409 23  37  1512    859086  854536  0.0 1817.0
sp|Q9Y719|MOK13_SCHPO   scaffold_7  39.02   2532    1339    44  6   2358    1648137 1655654 0.0 1702.0
sp|A5DHT2|RPB2_PICGU    scaffold_1  67.44   1210    353 8   18  1193    2428679 2425071 0.0 1690.0
sp|P32639|BRR2_YEAST    scaffold_3  41.48   2225    1190    38  1   2160    1828828 1835361 0.0 1611.0
sp|O43065|MOT1_SCHPO    scaffold_6  46.67   1864    881 27  153 1948    1193653 1199109 0.0 1543.0
sp|Q9HEH1|RENT1_NEUCR   scaffold_2  69.93   1124    296 11  4   1091    904443  901090  0.0 1540.0
sp|P22944|NIR_EMENI scaffold_4  70.33   1055    297 8   60  1102    1844667 1847819 0.0 1508.0
sp|P38095|LAMA_EMENI    scaffold_3  59.88   1194    435 6   1   1239    1554508 1558336 0.0 1469.7
sp|P22276|RPC2_YEAST    scaffold_3  62.7    1161    384 9   22  1147    181497  178057  0.0 1462.0
sp|P15398|RPA1_SCHPO    scaffold_1  45.41   1753    859 23  1   1684    1327838 1322667 0.0 1458.0
sp|Q10250|YD22_SCHPO    scaffold_1  52.33   1307    565 5   135 1481    2350824 2346650 0.0 1455.0
sp|Q01631|CYAA_NEUCR    scaffold_10 45.94   1800    880 29  462 2223    1267599 1272833 0.0 1454.0
sp|Q10094|YAOF_SCHPO    scaffold_6  56.63   1328    545 13  1   1317    432282  428359  0.0 1452.0
sp|O93937|PYR1_EMENI    scaffold_12 73.89   988 227 5   48  1005    800312  803272  0.0 1444.0
sp|Q00737|SUDA_EMENI    scaffold_13 59.74   1145    421 6   7   1215    798872  802575  0.0 1435.0
sp|Q10105|GCN1_SCHPO    scaffold_9  38.49   2549    1477    46  163 2661    433308  440831  0.0 1427.0
sp|Q4X0Z7|LONM_ASPFU    scaffold_2  71.25   967 253 6   113 1057    2240403 2243294 0.0 1375.0
sp|P00365|DHE2_NEUCR    scaffold_1  70.07   912 252 7   107 1038    587052  584195  0.0 1363.5
sp|Q9Y767|DPOG_NEUCR    scaffold_9  62.49   1037    344 9   62  1097    1170224 1167246 0.0 1362.0
sp|Q4P9K9|CHS8_USTMA    scaffold_10 42.75   1717    826 29  107 1757    209215  214092  0.0 1351.0
sp|O13396|MSH2_NEUCR    scaffold_10 70.54   937 257 7   1   937 1023034 1025787 0.0 1328.0
sp|O74298|LYS2_PENCH    scaffold_11 58.18   1162    455 11  258 1409    257081  253659  0.0 1328.0
sp|Q1K9C2|MET5_SCHPO    scaffold_2  53.88   1236    532 15  245 1458    1675465 1679124 0.0 1318.0
sp|Q06625|GDE_YEAST scaffold_7  47.25   1471    693 24  88  1510    977509  973202  0.0 1316.0
sp|Q10178|SF3B1_SCHPO   scaffold_13 60.53   1140    381 8   105 1201    939739  936398  0.0 1303.0
sp|A2R1F6|MSH3_ASPNC    scaffold_7  61.85   1135    390 18  4   1119    1281480 1278148 0.0 1293.0
sp|O14232|MTR4_SCHPO    scaffold_2  64.55   959 313 8   128 1117    973480  976562  0.0 1278.4
```

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

```
gavintargaryen@bogon:~/Documents/基因组学/实验三/gm_et_macosx$mv gm_key_64 .gm_key
```

c. 安装 perl 依赖

sudo cpan YAML

sudo cpan Hash::Merge

sudo cpan Logger::Simple

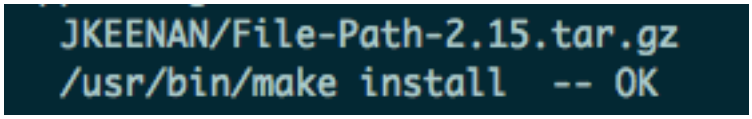
sudo cpan Parallel::ForkManager

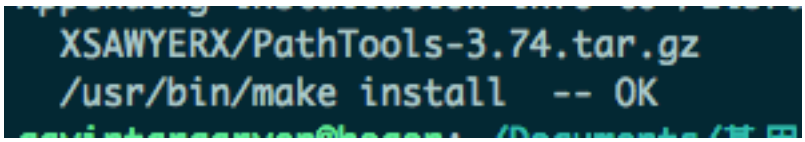
sudo cpan Getopt::Long

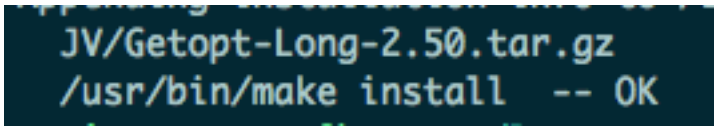
sudo cpan File::Spec

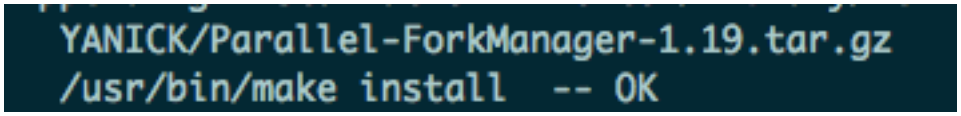
sudo cpan File::Path

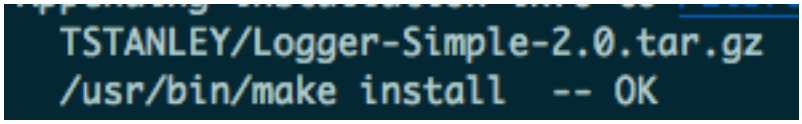
sudo cpan Data::Dumper











```
REHSACK/Hash-Merge-0.300.tar.gz
/usr/bin/make install  -- OK
```

```
TINITA/YAML-1.24.tar.gz
/usr/bin/make install  -- OK
```

```
SMUELLER/Data-Dumper-2.161.tar.gz
/usr/bin/make install  -- OK
```

d．测试是否成功

./gmes_petap.pl

```
# --------------------
Usage:  ./gmes_petap.pl  [options]  --sequence [filename]

GeneMark-ES Suite version 4.30
   includes transcript (GeneMark-ET) and protein (GeneMark-EP) based training and prediction

Input sequence/s should be in FASTA format

Algorithm options
  --ES           to run self-training
  --fungus       to run algorithm with branch point model (most useful for fungal genomes)
  --ET           [filename]; to run training with introns coordinates from RNA-Seq read alignments (GFF format)
  --et_score     [number]; 4 (default) minimum score of intron in initiation of the ET algorithm
  --evidence     [filename]; to use in prediction external evidence (RNA or protein) mapped to genome
  --training     to run only training step
  --prediction   to run only prediction step

Sequence pre-processing options
  --max_contig   [number]; 5000000 (default) will split input genomic sequence into contigs shorter then max_contig
  --min_contig   [number]; 50000 (default); will ignore contigs shorter then min_contig in training
  --max_gap      [number]; 5000 (default); will split sequence at gaps longer than max_gap
                 Letters 'n' and 'N' are interpreted as standing within gaps
  --max_mask     [number]; 5000 (default); will split sequence at repeats longer then max_mask
                 Letters 'x' and 'X' are interpreted as results of hard masking of repeats
  --soft_mask    [number] to indicate that lowercase letters stand for repeats; utilize only lowercase repeats longer than specified length

Run options
  --cores        [number]; 1 (default) to run program with multiple threads
  --pbs          to run on cluster with PBS support
  --v            verbose

Customizing parameters:
  --max_intron          [number]; default 10000 (3000 fungi), maximum length of intron
  --max_intergenic      [number]; default 10000, maximum length of intergenic regions
  --min_gene_prediction [number]; default 300 (120 fungi) minimum allowed gene length in prediction step

Developer options:
  --usr_cfg      [filename]; to customize configuration file
  --ini_mod      [filename]; use this file with parameters for algorithm initiation
  --test_set     [filename]; to evaluate prediction accuracy on the given test set
  --key_bin
  --debug
# --------------------
```

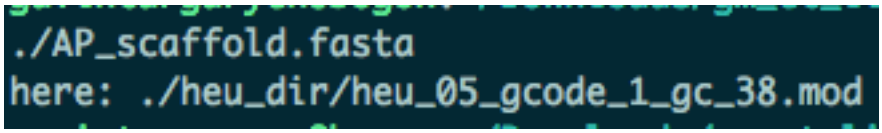
e.发现 mac 版缺少—predict_with 参数，由于此工具为 perl 脚本工具，故在 mac 上试用 Linux 版本脚本，但执行命令依然有问题，为无法运行二进制文件

```
cannot execute binary file
```

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

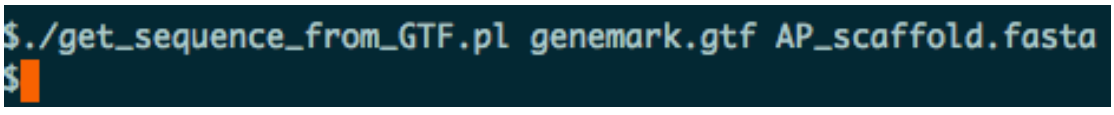
7.2、使用该软件对基因组序列进行基因预测分析

./gmes_petap.pl --prediction --fungus
--predict_with ./heu_dir/heu_05_gcode_1_gc_38.mod
--sequence ./AP_scaffold.fasta




```
scaffold_1  GeneMark.hmm    exon     721 2039    0   +   .   gene_id "1_g"; transcript_id "1_t";
scaffold_1  GeneMark.hmm    CDS 721 2039    .   +   2   gene_id "1_g"; transcript_id "1_t";
scaffold_1  GeneMark.hmm    stop_codon  2037    2039    .   +   0   gene_id "1_g"; transcript_id "1_t";
scaffold_1  GeneMark.hmm    exon    3579    4230    0   -   .   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    stop_codon  3579    3581    .   -   0   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    CDS 3579    4230    .   -   1   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    exon    8096    8585    0   -   .   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    CDS 8096    8585    .   -   2   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    exon    8984    9770    0   -   .   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    CDS 8984    9770    .   -   0   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    start_codon 9768    9770    .   -   0   gene_id "2_g"; transcript_id "2_t";
scaffold_1  GeneMark.hmm    exon    10775   13242   0   +   .   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    start_codon 10775   10777   .   +   0   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    CDS 10775   13242   .   +   0   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    exon    13347   15663   0   +   .   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    CDS 13347   15663   .   +   1   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    stop_codon  15661   15663   .   +   0   gene_id "3_g"; transcript_id "3_t";
scaffold_1  GeneMark.hmm    exon    16523   17596   0   +   .   gene_id "4_g"; transcript_id "4_t";
scaffold_1  GeneMark.hmm    start_codon 16523   16525   .   +   0   gene_id "4_g"; transcript_id "4_t";
scaffold_1  GeneMark.hmm    CDS 16523   17596   .   +   0   gene_id "4_g"; transcript_id "4_t";
scaffold_1  GeneMark.hmm    stop_codon  17594   17596   .   +   0   gene_id "4_g"; transcript_id "4_t";
scaffold_1  GeneMark.hmm    exon    18400   19122   0   +   .   gene_id "5_g"; transcript_id "5_t";
scaffold_1  GeneMark.hmm    start_codon 18400   18402   .   +   0   gene_id "5_g"; transcript_id "5_t";
scaffold_1  GeneMark.hmm    CDS 18400   19122   .   +   0   gene_id "5_g"; transcript_id "5_t";
```

./get_sequence_from_GTF.pl genemark.gtf AP_scaffold.fasta



```
>1_g
LHNELLELGWTGAGKSLDDLETSTWFDIWGQEADNCRDNLSDDLIAFLERAQMPRAGEEHSLFFYVYGLAHPKRLWDTFEWRFEEPDKHRYVTLFLANLGPSHPDGL
AFDQKTNKAIMQMSIHDASITLNGRTPWLPLEVILSAWLDMVDVGKVRAVDDSVEANEKFDPWICCHWNQGMVQETVKAFNALLDSIESRMEDQGFVVTDSKPPLLP
DATLEAAGIPNGFARSFFTQPRRPGFRYIAPDISIPSPGSFLQQPFSSIMHDEQDEDPDEDELIIEPILLFASSSHVSLANEEDKYHPFPWPYNQLLSFPAGLYLTD
SERSAGHEFEDGVKFVLPYGVGGKGFARTSDGLHIGDPRDEQDARCEDRLADLYQQGWNPFIEMHEVRLVKVLESWIEMVERGDWKVGPEGVEGSIDAFKEAETPEG
WERFVVPMGW

>2_g
MAHLLHRSVDSETAALILQLQRDDIEAALCQGDAHENVALQLQMQQLNLVETAHNDYTIAQSISRAVFDDRDLLTEELSAEQQAMRDRQLAQRLPQGADIPLDDSTL
TDNSEDDVNDRELSITCVACTNTFPWFDILETPCSHHYCVECLSELFDLSMKDETLYPPRCCRQTIPLDDAKLVLHPKLVRDFQQKSIELDAKDRTYCYDPRCSTFI
PAEHITEDVADCPDCGKRTCVICKAAAHRGDCPRDEDLQQLLQAAEHAVFVGPAQIIPAPFARPINSFLIQNAPVPRPAARPAAQAAPDLPAPAVRAVPAPAVRAVP
VPAAPVARPAAPAAPVPVPPARPAHAQAVGVVPVPAARPPTPAVRAVPAPAALATRPAVPTARAVSVPAVRPIAQDIVTNRQQRVVAQIRADLQRNHECDHERWTLE
NTLSGTIMPLDEIHNISTWAHAQDTPLHMHLDGAQLWETVVAGAGSLQEFTSCFDSVSLWFTKGLGAPIGSIIVGNAKFIQRARMIRKLLGGGLRQLGVIAGPAEVA
IEQVFLGGQLATAHRYARRLAKSWEDFGGKLQNPTETHMVWLDLEPAGVAGDTFADLAKCHGVLTMRRRLQRRLVLHYQICQDALQSIERLFYFVLNGRKCSTSVSS

>3_g
MQAILPGRPQAQLQAVSTGAFEGQTVICYISGNALIICNGPDSLLQTIYHDDLLPPSALVAVSYDHRTAKIATASSDNVYIYVLREEIKGQLRWFLDMTFALASGSI
RTLSWGTDDELLVASDKLLLLSTQDNSIRWTRSLPSPVSHATFSPSAALLATTSQYDRLVKVWRRLSFESAIFDYAYLTHPDIVTHLEWRKDSADQDVLYTICLDGR
LRVWKTALPHQIDVLGLYADLDLTKAVRPLVPSNRRYASLVPSSVFSDAVEHAVNTPTLPAERHAIEYMKELAKRNPDIVVVTDDQGHMSAWGLENVGCKTRTSNTE
HAEFFHATHVENLFFDFSKHAQIDEDNARIINFALPGQPGDIALLVHHFDGRLQWYQGRVHHVFDPSPRKHRMRNLASWTGHSSSIKKLIRTANGSAVISRTDTDEG
VVWKHRAKFLARKSKVKLQEHIHRTVLLQDGDFVVLLHHQSISLWDARLPNAKEIGRCPYSVSGKPLCLVLLPAPEDAGHLSYLATVTADMHGVVWELDLASPSPTI
TEFCAFQLDRIENMSYFLPVDPAGSSPVVSGLFDTFAQDVAVSWTTTGILQTWTAKVDRLNKSVTWLNISTTDTTIQDPSLGSGTSIRKAALVDKSRSTLTIWDTKS
GRLDYEETFENQTIQDLDWASTPDNQSILAVGFPHAVFVYTQLRYDYVTERPSWGRIKQINVRNLSPHPIGDSVWLGSGNLIIGAGNQLYLAGNEVDPQKDLSPDLQ
GSTPHKSSTRIQDVVRRLNGPLPVFHPQFISQCILAGKMDIVHRILLNLQKTLKFYTEGDDLDSFQGLDIEDFITLDMQEESTAVTEEVAQSLVSLLSEKSVPQLSS
TEQLSLADTIECVGMAEKHRRSIDENAARYLLFWRETVMRSRQSSSSAAVTWREMLWAYHSTSQDILVDLVTRQNKGQMTWSHARDTGVFVWLTDREALLQQFEAVA
RSAYTSTELRDPVNCSLHYLALRKKNVLTGLWRMATWSREQAATMRLLKNDFTDPRWRTAANKNAYALMGKRRFEYAAAFFLLADNLDSAVSVLSNQLGDVQLAIAV
ARVYGGDDSPSLKKFLTERVLPNAALDGNRCQATWTYWMLGERSSAVRALVSPLHSLLDPPGSPPDSLQAKSFRNDDPALVVLYQQLREKSLQTLRGALMISGNEEW
EFITKTATLLLRMGCDILALDLVRHWEFLHPPPVAKQEEQEEEVEESVAELPRRRLSRTTTDDFDPRKLLRRRSSLVVADLPEGRPVVHETGAPSMLDGWSAPSSQG
QQPSSLLDQWTVPASSAAKPAPSMLEQWSMPERPKQQANPAPLMLDNWATPSQPASKPAASLLDDWIAPPTKSKTPEIQLKTSAPSMLDNWSSVPQPASKPTPSLLD
DWATPIAKPKTEKAEVVSSTPSMLDQWRSPSPSAPKAVANPPSSMLDEWTAPVAKPSIESTKAAPLSLLDQWTVPPVSSKPVPTMPDQLKEPIVTSSESAVEGSQFL
ANPTMPTSMPNSANDDAKLGETMVIIQDHGASKSEPNGNESQNEEDPKAETQKSSKTGKMGDEKAEEYKLPKGLKQPPPEAFKEPDPNSLLDAFGF
```

```
>1_g
CTGCACAACGAACTTCTCGAGCTCGGATGGACAGGTGCTGGTAAATCGCTGGATGACTTGGAGACGAGTACTTGGTTCGACATTTGGGGTCAAGAAGCGGATAATTG
TCGAGATAATCTTTCAGATGACCTGATAGCTTTCCTCGAGCGAGCTCAAATGCCCAGAGCAGGTGAAGAACATTCACTATTCTTTTACGTATATGGGTTGGCGCACC
CAAAGCGATTGTGGGATACCTTCGAATGGCGATTTGAAGAACCAGATAAGCATCGTTATGTGACGCTCTTCCTTGCAAATCTCGGACCGAGCCACCCGGACGGACTT
GCTTTTGACCAAAAGACAAACAAAGCTATCATGCAAATGTCTATTCATGATGCATCCATAACATTGAATGGCCGCACTCCCTGGCTTCCGTTGGAAGTCATTTTGAG
TGCCTGGCTGGACATGGTCGATGTTGGCAAGGTTAGAGCCGTGGACGACTCTGTCGAAGCAAACGAGAAATTTGACCCATGGATATGCTGCCACTGGAACCAAGGCA
TGGTACAAGAGACAGTCAAGGCGTTCAATGCGCTCCTCGATAGCATTGAATCTCGGATGGAAGATCAAGGGTTCGTGGTGACTGACTCCAAACCACCGCTACTTCCC
GATGCAACTCTGGAAGCTGCAGGCATCCCTAATGGTTTTGCCCGGTCATTTTTTACTCAACCCCGCCGTCCCGGATTCAGATACATTGCACCAGATATCTCCATTCC
CAGTCCAGGCTCGTTCCTTCAACAACCCTTTTCGTCCATAATGCACGACGAACAAGATGAGGACCCCGACGAAGATGAGCTCATCATCGAACCTATCCTTTTATTTG
CATCTTCAAGCCACGTTAGTCTGGCCAATGAAGAAGACAAATACCATCCTTTCCCTTGGCCTTATAACCAACTGCTTTCTTTCCCTGCGGGCCTCTACCTCACTGAT
TCCGAAAGATCAGCGGGTCATGAATTTGAAGATGGCGTCAAATTTGTTCTGCCCTATGGTGTTGGCGGCAAAGGCTTTGCTCGTACCAGTGATGGACTGCACATTGG
TGATCCTCGAGATGAACAAGATGCACGATGCGAAGACAGACTTGCGGATCTTTATCAGCAAGGCTGGAATCCTTTTATCGAGATGCACGAGGTCAGACTGGTCAAAG
TTCTGGAAAGCTGGATAGAAATGGTTGAGAGAGGCGACTGGAAAGTCGGCCCAGAGGGGGTTGAGGGCAGCATTGATGCTTTCAAAGAGGCTGAAACA
CCGGAAGGTTGGGAGAGGTTTGTTGTACCCATGGGTTGGTGA

>2_g
ATGGCTCACCTTCTCCACCGGTCTGTAGACTCTGAGACAGCTGCTCTCATTCTTCAGCTGCAGCGCGACGACATCGAGGCGGCGCTTTGCCAGGGCGATGCACATGA
AAATGTCGCTCTCCAACTGCAAATGCAACAGCTCAACCTTGTCGAAACTGCGCACAACGATTACACCATCGCTCAGAGCATTTCGCGCGCAGTTTTCGACGACCGTG
ACTTGCTGACAGAGGAACTGTCTGCCGAGCAGCAGGCGATGAGAGACCGCCAGCTAGCACAACGACTCCCTCAGGGCGCTGACATCCCGCTTGACGACTCGACCCTC
ACCGACAACAGCGAAGACGACGTTAACGATCGCGAGTTGTCCATCACTTGTGTGGCCTGTACCAATACCTTTCCTTGGTTCGACATTCTCGAAACTCCCTGCAGTCA
TCACTACTGCGTCGAGTGTCTCAGCGAGCTTTTCGACTTATCCATGAAGGATGAGACCTTGTACCCACCTCGCTGCTGTCGTCAAACAATCCCTCTGGATGACGCCA
AGCTCGTGCTTCATCCCAAACTTGTCAGAGATTTCCAGCAGAAGTCCATTGAGCTTGATGCTAAGGACAGAACCTACTGCTATGATCCTCGTTGTTCGACATTCATC
CCTGCCGAACACATTACAGAAGATGTCGCTGACTGTCCTGATTGTGGCAAGAGGACTTGTGTTATCTGTAAAGCTGCAGCTCACCGCGGTGATTGCCCTCGTGACGA
AGATCTTCAGCAGCTTCTCCAGGCGGCCGAGCATGCCGTCTTTGTCGGACCTGCTCAGATTATTCCTGCTCCTTTTGCTCGCCCTATTAACAGCTTTCTCATCCAGA
ACGCCCCAGTCCCTCGCCCTGCCGCTCGTCCCGCCGCTCAGGCTGCCCCCGACCTCCCAGCCCCAGCTGTTCGTGCTGTACCCGCTCCAGCTGTCCGCGCTGTGCCC
GTCCCAGCGGCCCCAGTCGCTCGTCCCGCCGCGCCGGCTGCTCCTGTCCCAGTTCCACCTGCTCGCCCTGCTCACGCTCAAGCTGTCGGTGTTGTGCCCGTGCCAGC
TGCTCGCCCTCCCACCCCAGCCGTTCGCGCCGTACCTGCTCCAGCTGCACTTGCCACTCGTCCTGCAGTCCCGACTGCTCGTGCGGTGTCTGTCCCGGCCGTGCGTC
CCATCGCACAAGATATCGTCACTAACCGCCAGCAACGCGTGGTCGCTCAGATCAGAGCTGATCTCCAACGCAACCACGAATGCGATCATGAGAGATGGACTCTCGAG
AACACATTGTCTGGGACGATCATGCCACTGGACGAGATTCATAACATCTCAACTTGGGCCCATGCCCAAGATACCCCACTCCATATGCACTTAGACGGCGCTCAATT
GTGGGAAACCGTCGTAGCCGGAGCTGGTTCTCTTCAAGAATTTACCTCTTGCTTCGACAGTGTGTCTTTGTGGTTTACTAAAGGTCTCGGAGCACCCATTGGTTCGA
TTATAGTTGGCAATGCGAAGTTCATCCAACGAGCCCGCATGATACGAAAACTTCTTGGGGGCGGCCTGCGTCAATTGGGCGTGATAGCAGGTCCAGCAGAGGTTGCT
ATTGAGCAGGTCTTTCTCGGCGGACAATTGGCTACAGCTCATCGCTATGCTCGGAGGCTTGCCAAATCCTGGGAGGATTTTGGTGGCAAGCTGCAAAACCCGACAGA
GACCCACATGGTTTGGCTTGATCTGGAGCCTGCAGGCGTCGCCGGTGATACTTTTGCAGACCTGGCCAAGTGTCATGGTGTGCTGACAATGAGGAGAAGGCTTCAGC
GGCGCCTTGTCCTACATTATCAGATCTGCCAGGATGCCCTGCAATCAATAGAGAGGCTCTTTTATTTTGTGTTGAATGGGAGAAAGTGTTCAACAAGCGTCTCAAGC
```

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。

a. 建立已知蛋白数据库

```
makeblastdb -in uniprot.fasta -dbtype prot -parse_seqids -out PRO
```

```
(/opt/BioBuilds)-bash-4.2$ makeblastdb -in uniprot.fasta -dbtype prot -parse_seqids -out PRO


Building a new DB, current time: 04/29/2018 00:37:09
New DB name:   /home/student/s14/PRO
New DB title:  uniprot.fasta
Sequence type: Protein
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 33161 sequences in 2.494 seconds.
```

b. 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 &
```

```
2_g sp|D7PHZ0|VRTJ_PENAE    50.73   205 95  2   427 631 179 377 1e-49   179
3_g sp|P47104|RAV1_YEAST    27.85   1397    840 35  4   1326    5   1307    4e-155  512
9_g sp|D4B3C8|A2965_ARTBC   29.08   306 168 7   623 895 65  354 7e-23   105
10_g    sp|O14031|PGT1_SCHPO    29.10   354 205 8   32  361 84  415 4e-29   120
11_g    sp|D4B1P2|A2372_ARTBC   31.83   487 292 15  34  496 29  499 2e-53   190
12_g    sp|A2QHE5|FDC1_ASPNC    68.73   502 146 2   299 799 7   498 0.0 727
13_g    sp|P36619|PMD1_SCHPO    42.34   222 83  1   1   222 1153    1329    9e-48   171
13_g    sp|P36619|PMD1_SCHPO    39.57   230 86  2   1   222 454 638 1e-39   148
14_g    sp|P25386|USO1_YEAST    26.51   1260    726 39  1   1141    7   1185    1e-89   320
15_g    sp|Q8TFZ1|XRN2_ASPFU    63.96   910 301 11  1   889 1   904 0.0 1153
16_g    sp|Q09776|SNU23_SCHPO   44.27   131 69  2   68  197 21  148 3e-26   102
18_g    sp|C1GQH3|P20D1_PARBA   27.55   265 155 11  1   262 124 354 3e-10   62.4
20_g    sp|N4WE43|RED2_COCH4    39.09   197 117 1   69  265 67  260 3e-43   153
21_g    sp|P38860|MTG2_YEAST    37.85   391 190 9   85  448 91  455 3e-50   186
22_g    sp|O14064|BIR1_SCHPO    36.59   82  51  1   5   85  116 197 3e-14   76.6
22_g    sp|O14064|BIR1_SCHPO    32.97   91  52  4   22  104 31  120 2e-06   51.6
23_g    sp|O74339|TAM41_SCHPO   35.59   413 196 8   28  440 48  390 5e-63   219
26_g    sp|Q6CE48|IND1_YARLI    66.67   60  20  0   1   60  142 201 6e-20   91.3
27_g    sp|Q0UPL5|PAN2_PHANO    52.91   189 89  0   1   189 153 341 3e-67   224
28_g    sp|Q0UPL5|PAN2_PHANO    53.79   132 61  0   1   132 349 480 4e-43   154
29_g    sp|Q0UPL5|PAN2_PHANO    61.65   558 203 7   1   555 535 1084    0.0 685
31_g    sp|Q9Y7B6|PANB_EMENI    69.83   295 71  1   2   296 72  348 7e-153  436
35_g    sp|Q7SIC2|QDOI_ASPJA    52.23   337 153 6   670 1002    15  347 7e-101  324
36_g    sp|P13586|ATC1_YEAST    52.29   918 335 15  1   905 111 938 0.0 874
37_g    sp|A0A024SMV2|XDH_HYPJR 26.34   372 229 13  86  416 5   372 3e-20   92.4
42_g    sp|P0C582|M2OM_NEUCR    78.67   286 60  1   38  322 32  317 6e-150  446
45_g    sp|O59810|VGL1_SCHPO    26.11   1283    802 33  78  1298    87  1285    2e-126  426
46_g    sp|O94387|YGSA_SCHPO    25.00   296 168 13  735 999 1287    1559    2e-06   52.4
47_g    sp|P17442|PHO81_YEAST   36.25   80  50  1   1049    1127    571 650 9e-06   50.4
48_g    sp|Q4WHU1|HPPD1_ASPFU   69.11   246 75  1   236 481 115 359 3e-122  367
49_g    sp|D2YW48|GST_COCIM 44.97   149 79  2   1   146 82  230 1e-37   131
54_g    sp|P34909|NOT4_YEAST    48.54   171 75  2   50  210 100 267 2e-43   170
55_g    sp|S0ARX1|FSDH_FUSHE    31.75   378 234 6   13  373 18  388 1e-48   182
56_g    sp|P86029|HQD2_CANAL    36.39   305 182 6   6   302 2   302 3e-59   195
57_g    sp|Q9C0V4|YOM2_SCHPO    34.21   228 103 4   471 690 558 746 7e-31   130
58_g    sp|Q8NK50|MTDH_HYPJE    35.53   273 162 8   29  297 2   264 6e-43   150
67_g    sp|Q4INZ9|FKBP4_GIBZE   50.22   458 179 12  1   433 62  495 7e-97   303
68_g    sp|O94623|REV1_SCHPO    50.38   131 55  3   319 447 211 333 1e-31   129
```

blastp 结果

c．编程在 gtf 中加入比对的 query 名，代码为 genemark.py

```
scaffold_1  GeneMark.hmm    exon    721 2039    0   +   .   "gene_id ""1_g""; transcript_id ""1_t"";"
scaffold_1  GeneMark.hmm    CDS 721 2039    .   +   2   "gene_id ""1_g""; transcript_id ""1_t"";"
scaffold_1  GeneMark.hmm    stop_codon  2037    2039    .   +   0   "gene_id ""1_g""; transcript_id ""1_t"";"

scaffold_1  GeneMark.hmm    exon    3579    4230    0   -   .   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    stop_codon  3579    3581    .   -   0   "gene_id ""2_g""; transcript_id ""2_t"";"
   sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    CDS 3579    4230    .   -   1   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    exon    8096    8585    0   -   .   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    CDS 8096    8585    .   -   2   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    exon    8984    9770    0   -   .   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    CDS 8984    9770    .   -   0   "gene_id ""2_g""; transcript_id ""2_t"";"
sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    start_codon 9768    9770    .   -   0   "gene_id ""2_g""; transcript_id ""2_t"";"
   sp|D7PHZ0|VRTJ_PENAE
scaffold_1  GeneMark.hmm    exon    10775   13242   0   +   .   "gene_id ""3_g""; transcript_id ""3_t"";"
sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    start_codon 10775   10777   .   +   0   "gene_id ""3_g""; transcript_id ""3_t"";"
   sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    CDS 10775   13242   .   +   0   "gene_id ""3_g""; transcript_id ""3_t"";"
sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    exon    13347   15663   0   +   .   "gene_id ""3_g""; transcript_id ""3_t"";"
sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    CDS 13347   15663   .   +   1   "gene_id ""3_g""; transcript_id ""3_t"";"
sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    stop_codon  15661   15663   .   +   0   "gene_id ""3_g""; transcript_id ""3_t"";"
   sp|P47104|RAV1_YEAST
scaffold_1  GeneMark.hmm    exon    16523   17596   0   +   .   "gene_id ""4_g""; transcript_id ""4_t"";"
scaffold_1  GeneMark.hmm    start_codon 16523   16525   .   +   0   "gene_id ""4_g""; transcript_id ""4_t"";"

scaffold_1  GeneMark.hmm    CDS 16523   17596   .   +   0   "gene_id ""4_g""; transcript_id ""4_t"";"
scaffold_1  GeneMark.hmm    stop_codon  17594   17596   .   +   0   "gene_id ""4_g""; transcript_id ""4_t"";"
```

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