一、数据提取
用 hmmer 的 hmmsearch -E 1e-8 Ndr.hmm $ARGV[1] 做隐马查询。在物种蛋白组中找到所要查找的序列。

二、建树前的序列处理。
1、将所有物种的序列整理到一个文件中。linux 中的命令是 cat *.fasta> All.fasta,然后在 vi 中进行替换
vi 中的替换
%s/>/\r>/g
%s/\n\n\n/\r\r/g
2、用 hmmsearch -E 1e-8 Ndr.hmm InputSequence.fasta >outsearch.txt 处理序列文件,
3、运行结果 outsearch.txt 中提取序列（即从中搜索含有'jgi|ENS|gi|FBpp|LOC|ZK1|AT[2|5]G|Y48G'的项，输出要求的列）：
egrep 'jgi|ENS|gi|FBpp|LOC|ZK1|AT[2|5]G|Y48G' outsearch.txt | awk '{print $1,$3}' >output.txt
4、上述结果在日记本中再进行替代，删除。(删除无用信息)。然后用 tree_build 进行序列整合。
在结果中剔除- 等无用序列，就可以用来建树。

三，建树
1、用 clastx 做多序列比对。删除不合适序列。修齐首尾有 gap 的序列。
2、
用 mega 将 aln 文件转换为 meg 文件，然后打开 meg 文件，建 nj 树。看看结果。
用 mrbayes 建树必须先转换文件使其成为 nex 格式（用 mega 完成）。然后在文件后面加上命令：

```
begin mrbayes;
outgroup Alipes;
outgroup There;
log start replace;
charset 1st=1-576/3;
charset 2nd=2-576/3;
charset 3rd=3-576/3;
partition by_codon=3: 1st,2nd,3rd;
set partition= by_codon;
lset applyto=(1)st=6 rates=invgamma;
lset applyto=(2)nd=6 rates=propinv;
lset applyto=(3)rd=6 rates=gamma;
mcmc ngen=500000 printfreq=1000 samplefreq=100 nchains=4 temp=0.2
savebrlens=yes;
sump burnin=1250
sumt burnin=1250 contype=allcompat;
log stop;
end;
```

正式运行前必须修正 ngen，逐渐扩大直至得到合适的结果。

```
begin mrbayes;
outgroup Dict238670;
log start replace;
mcmc ngen=2000000 samplefreq=100 printfreq=100 nchains=4;
sump burnin=5000
sumt burnin=5000 contype=all compat;
```

log stop;
end;
25%*ngen=samplefreq*burnin
输入 exe File.nex 运行程序。
注：为了让程序能够在 linux 后台运行，在命令行输入:nohup mb -i file.nex > output.txt &
由于服务器程序有问题。所以需要在本机自己进行建树。输入的命令为：
exe File.nex
sump burnin=250
sumt burnin=250

用 Phyml 建 ML 树的时候先要在 clastx 中 alignment 时保存一份 phylip 格式的文档。然后在 phyml 中打开文档，用 prottest 选择合适的模型参数（在 cmd 中用 java -jar ProtTest.jar 命令打开）。即可以产生一个 txt 文档。

四、建 exon、intron 模型。
1、在 pfam 上获得 domain 信息。
2、用代码提取所要求序列的 exon，intron 信息。包括：1.Exon_Info.pl（.GTF->.INFO），2.Info_select.pl(.INFO,.SEL->.EXON)，3.Protein_Exon_Info.pl(.EXON,.SEL->.EIN)
3、在上一步产生的 ein 文件中手动添加 domain 信息。用 4.exon_domain_draw.pl 画图。
操作过程：perl 1.Exon_info.pl Homo.gtf perl 2.Info_select.pl Homo.gtf.info Info_seq.sel
perl 3.Protein_Exon_Info.pl Homo.gtf.info.exon Info_seq.sel perl 4.exon_domain_draw.pl Homo.gtf.info.exon.ein

4、对于没有 gff 的数据用 blat 进行分析。下载 DNA、mRNA、蛋白质数据。运行以下命令：
blat DNA_seq.fasta mRNA.fasta -ou=axt out1.blat
blat DNA_seq.fasta Protein.fasta -t=dnax -q=prot -out=axt out2.blat

可以收集数据的网站：ncbi、ensembl、jgi、ucsc、megg、pfam。