2020.12 基因组学实习报告

小组四人：

18生信二班肖高泓201842160518（集群号：Agswag）

18生信二班黄程201841660825（集群号：hc）

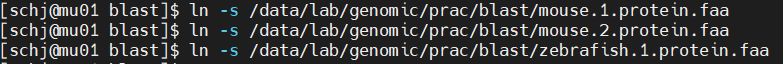
18生信二班吴选魁201841635206（集群号：schj）

18生信二班陈东南201841660824（集群号：cdn）

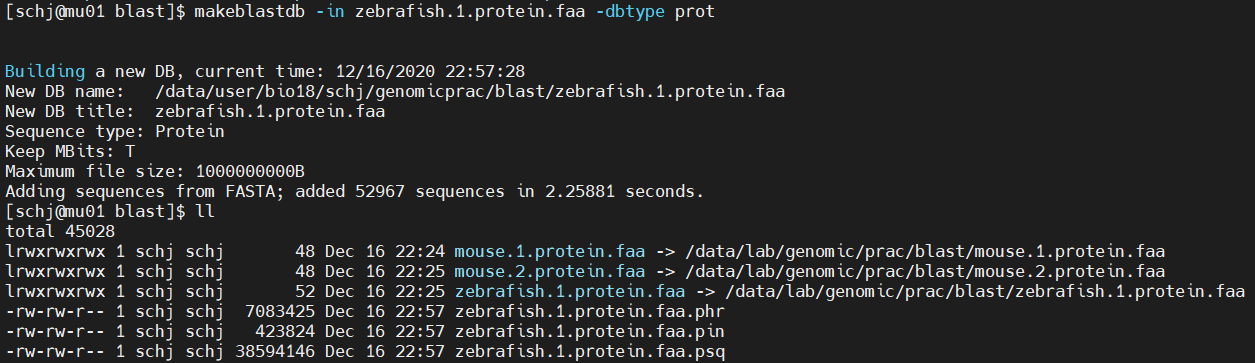
1.本地Blast

1.1 准备数据

建立数据的软链接

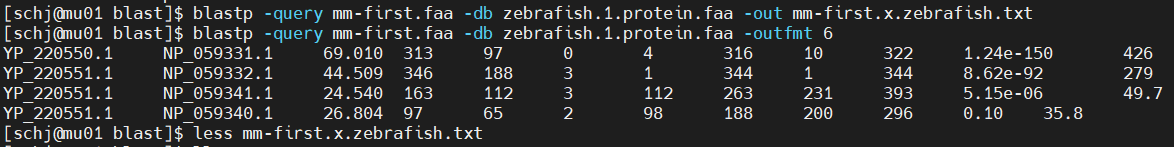


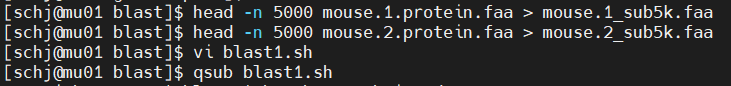
1.2 建索引



（BLAST+中makeblastdb参数详解-in后接输入文件，要格式化的fasta序列，-dbtype后接序列类型，nucl为核酸，prot为蛋白）

1.3 运行blastp





（blastp-query输入要比对的文件路径，-db格式化后的数据库名称，-out输出文件名，-outfmt:

0 = pairwise,

1 = query-anchored showing identities,

2 = query-anchored no identities,

3 = flat query-anchored,show identities,

4 = flat query-anchored,no identities,

5 = XML Blast output,输出的信息最全,

6 = tabular,表格的格式,

7 = tabular with comment lines,

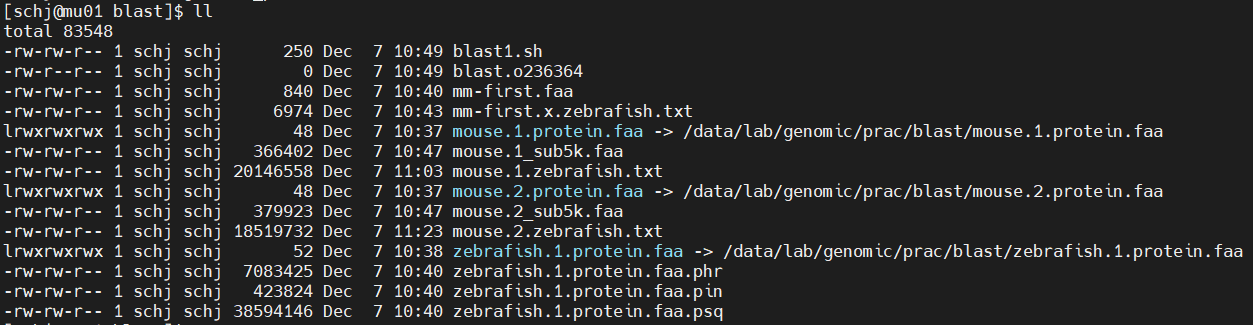
8 = Text ASN.1,

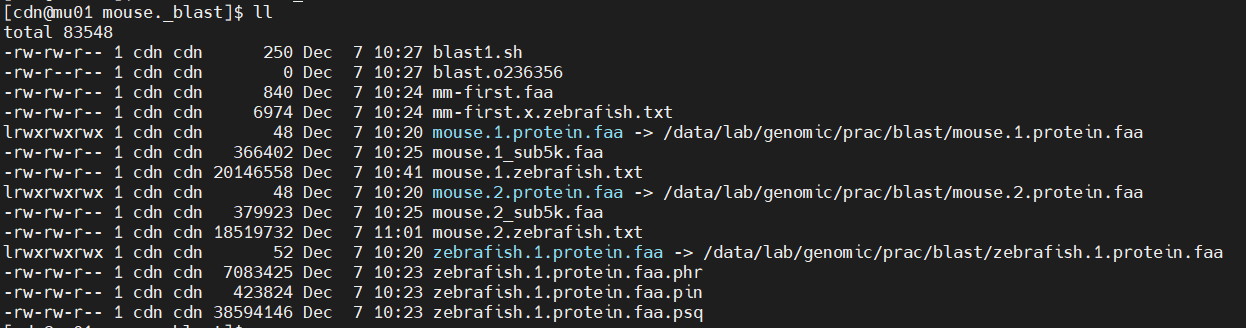
9 = Binary ASN.1,

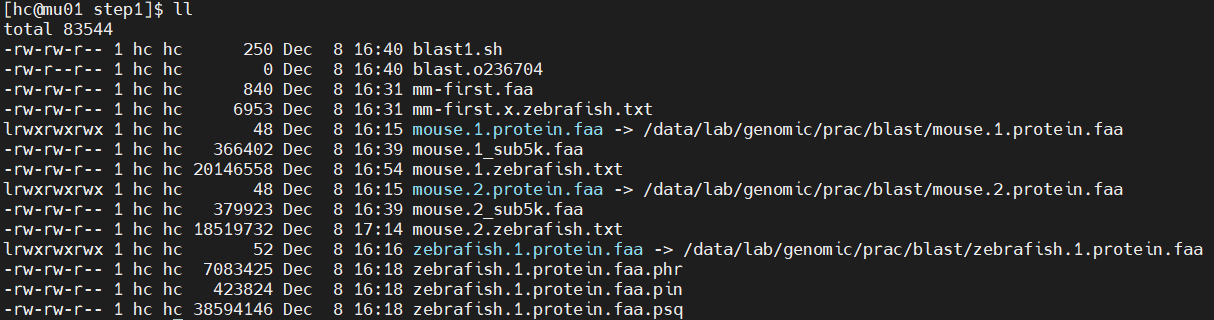
10 = Comma-separated values）

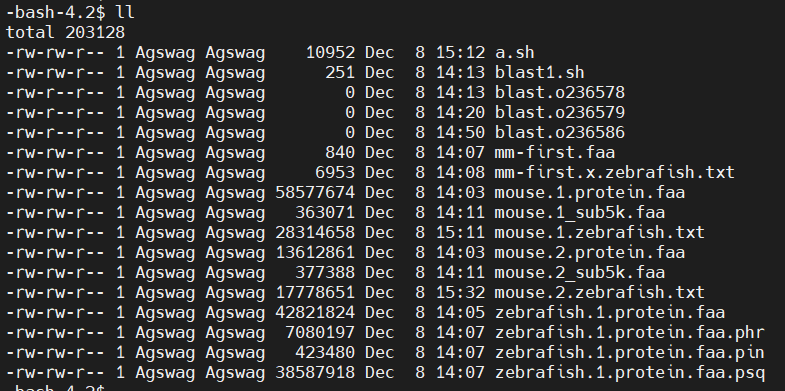
取2条序列试运行，再取前5000行，用于后续分析

成员结果如下：

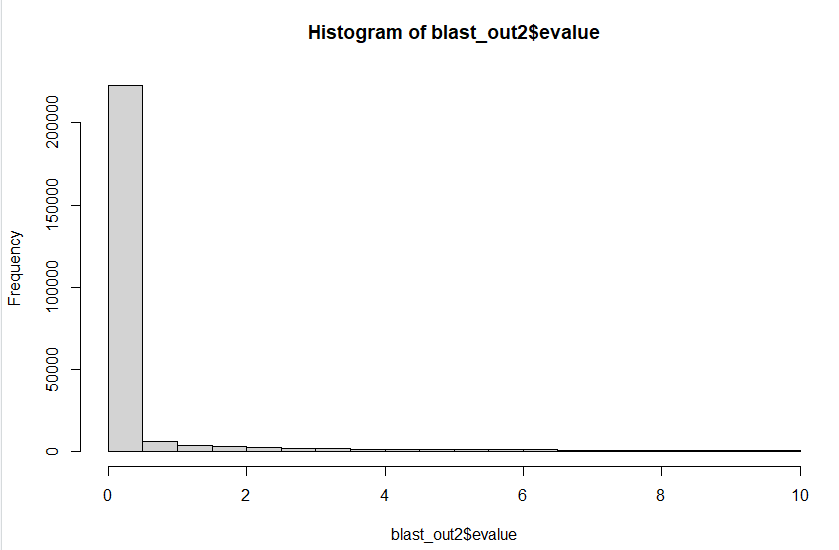
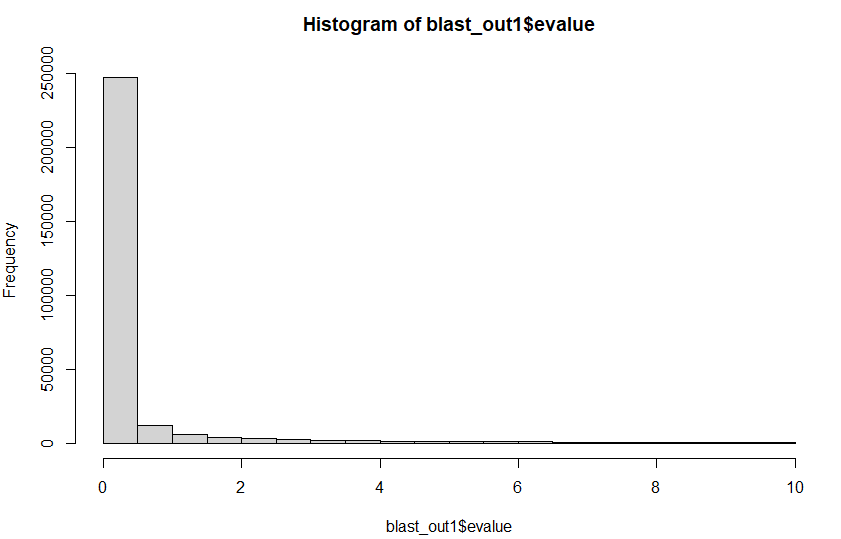
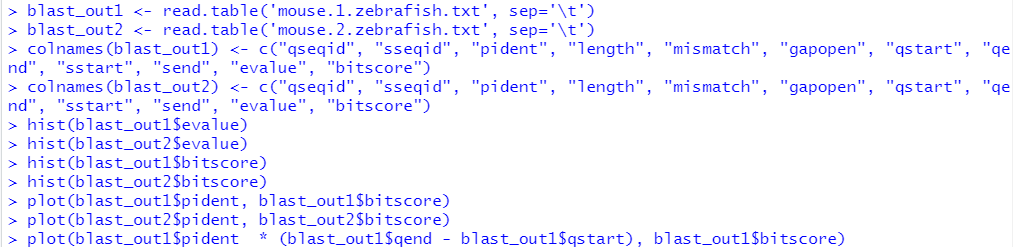


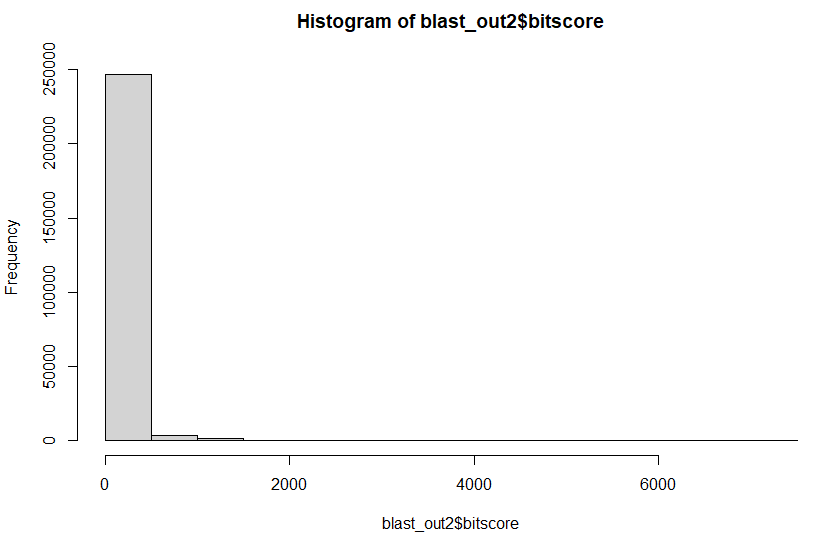
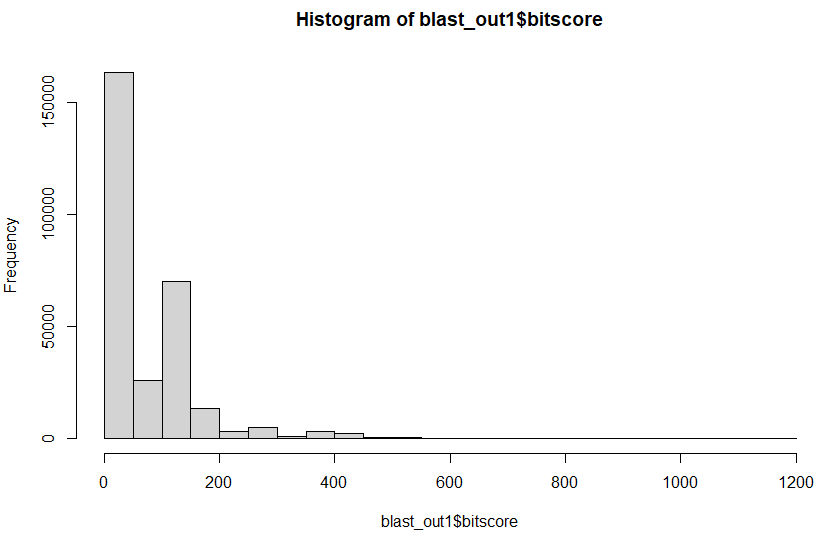


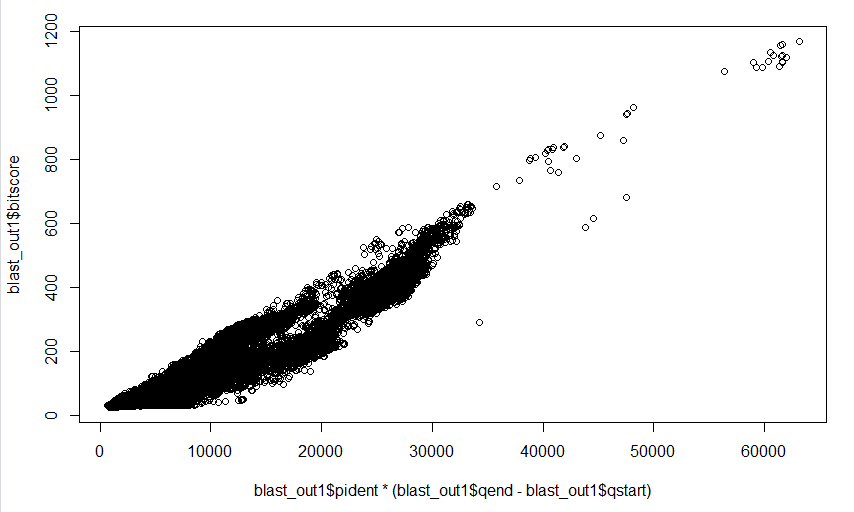
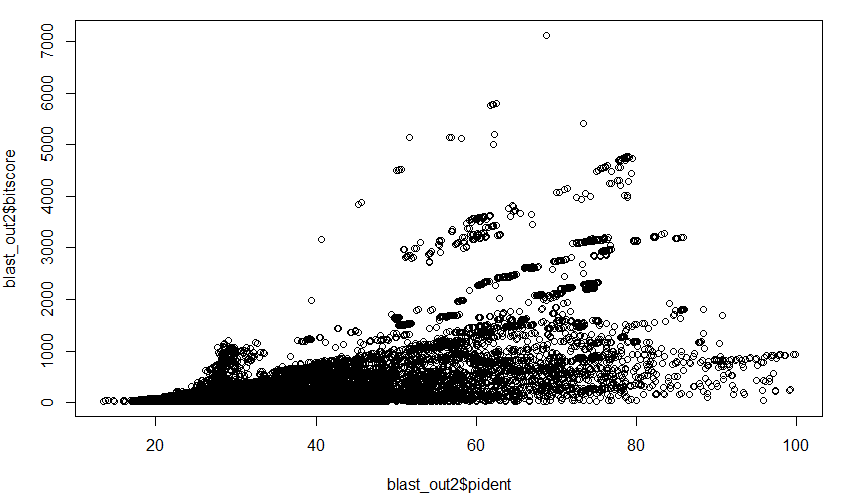
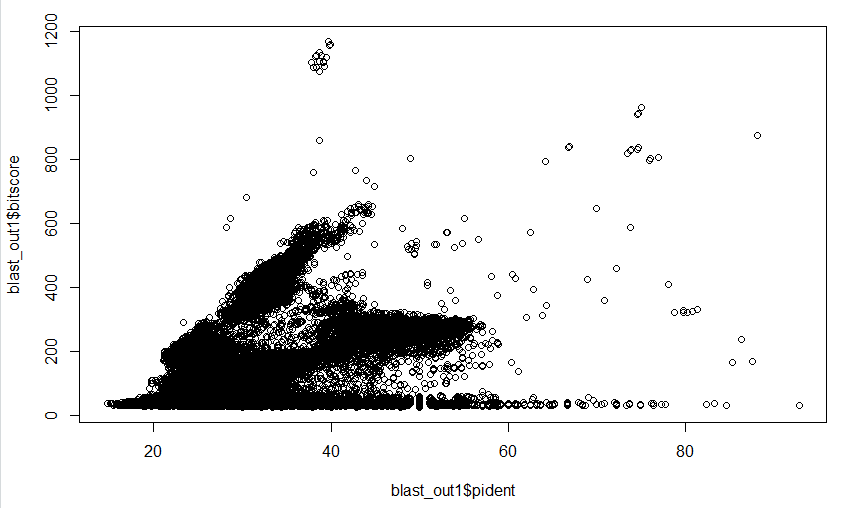




1.4 Visualizing BLAST score distributions in RStudio



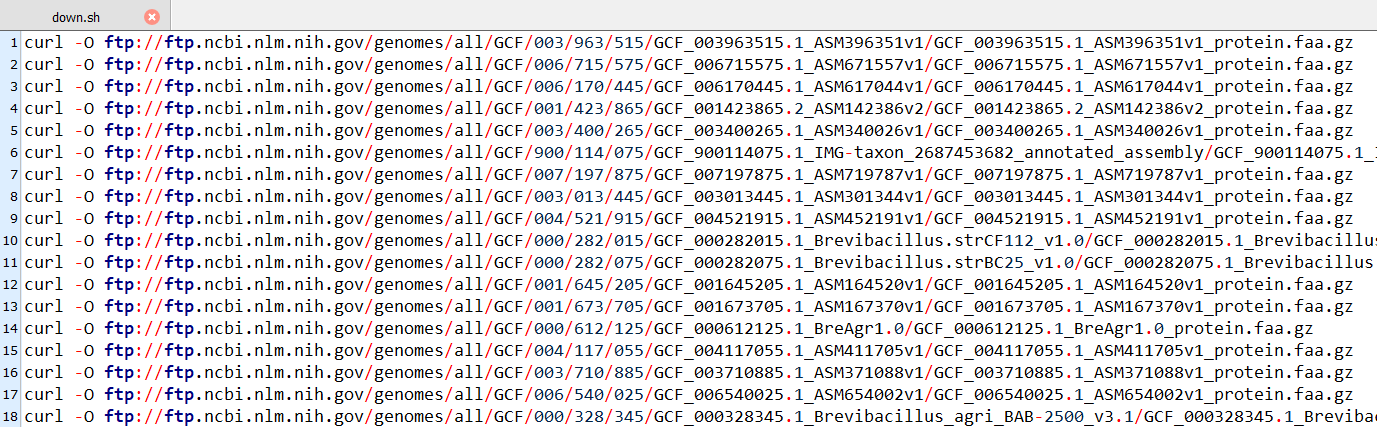




2.根据blast结果对蛋白序列进行聚类 -- 构建基因家族

2.1 数据准备：

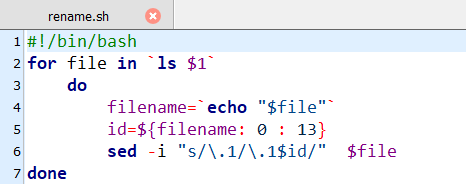
将下载命令整合到down.sh



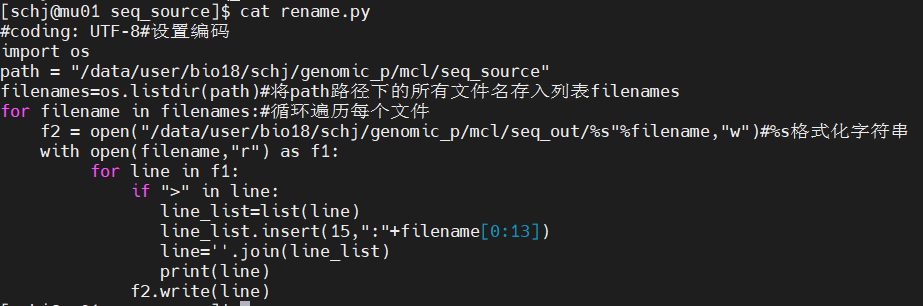
2.2 对蛋白序列进行两两比对

改序列名，在各自序列名后面添加GCF编号：

方法一rename.sh



方法二rename.py



将所有蛋白序列合并到一个文件all\_pro.faa，建索引makeblastdb -in all\_pro.faa -dbtype prot

使用更快的diamond比对，也是两步建库和比对

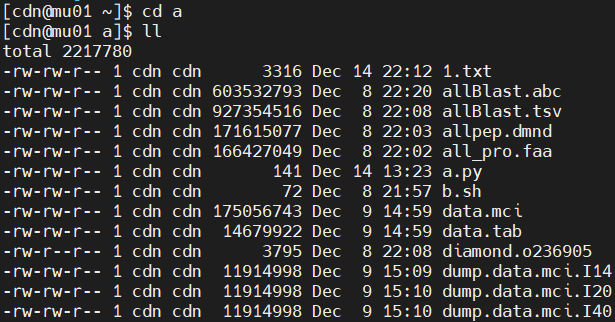
2.3 提取每个hit的score值，构建一个表征两条序列的相似性的特征值

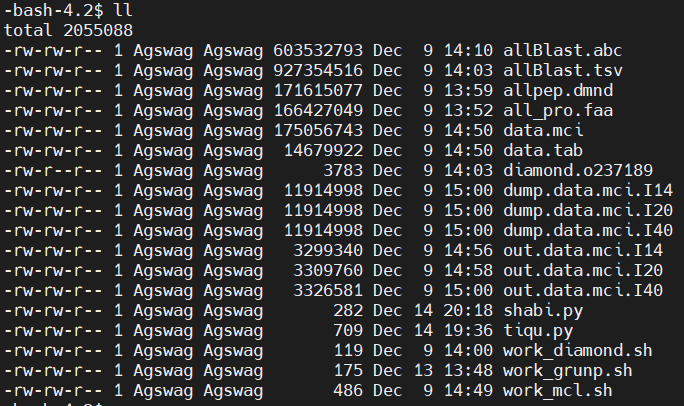
cut -f 1,2,12 allBlast.tsv > allBlast.abc

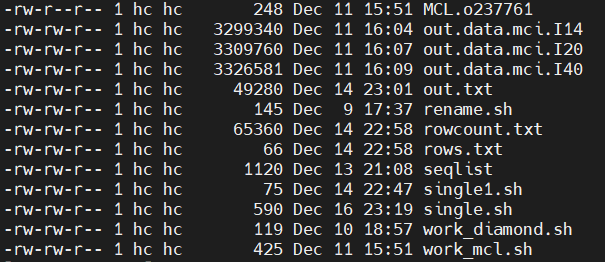
（cut-c ：以字符为单位进行分割。char）

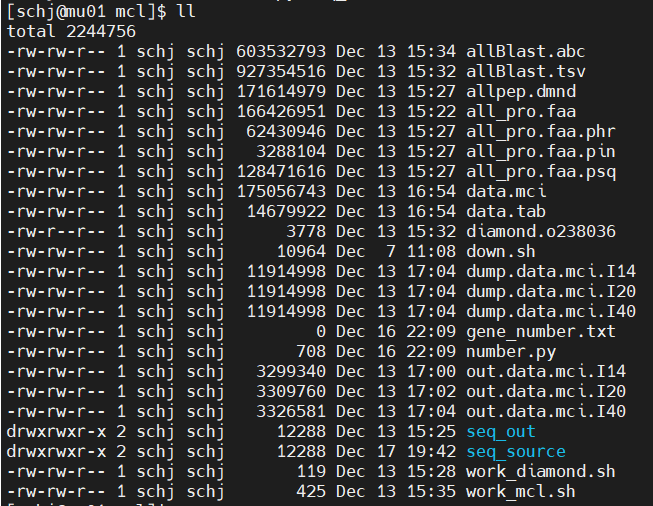
2.4 用mcl进行聚类

成员结果如下：









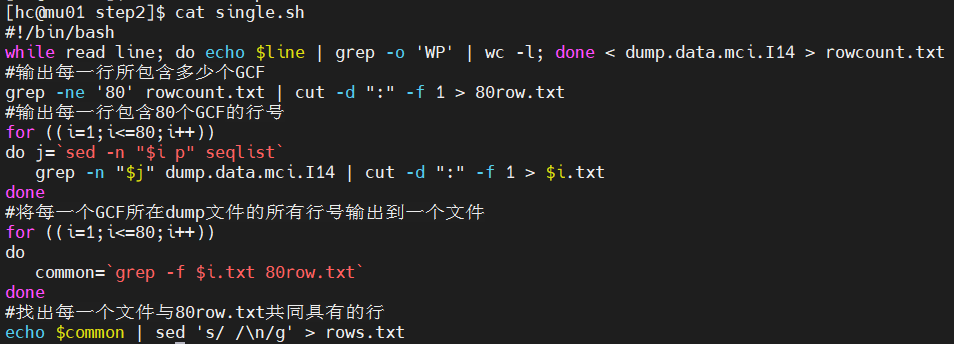
3.构建物种进化树（自行完成）

3.1 提取单拷贝基因家族的基因序列

基于dumpI40，统计满足条件的行（即单拷贝基因家族），每个菌株均出现且仅出现一次（如果统计出有100个单拷贝基因家族，就得到一个只有100行的dump新文件）

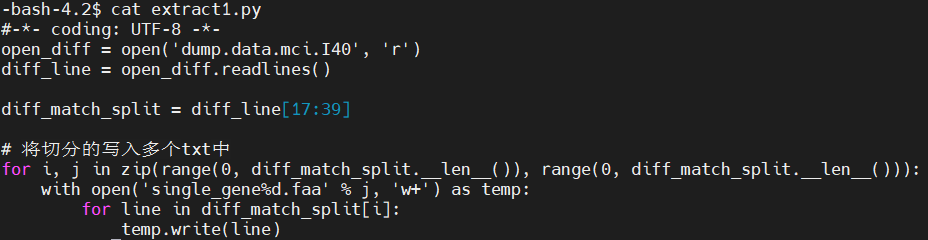
提取单拷贝基因家族行号

方法一single.sh



****

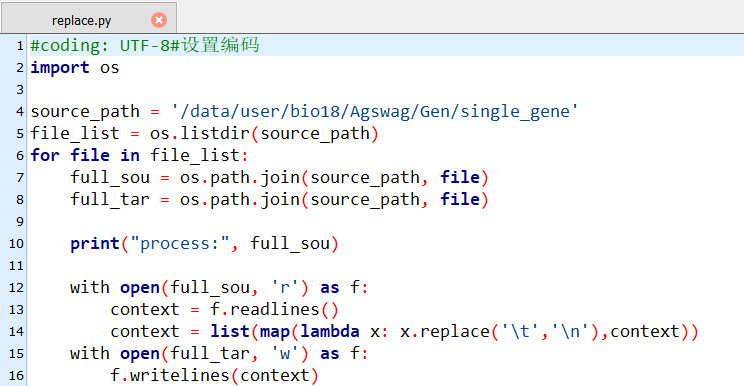
方法二extract1.py

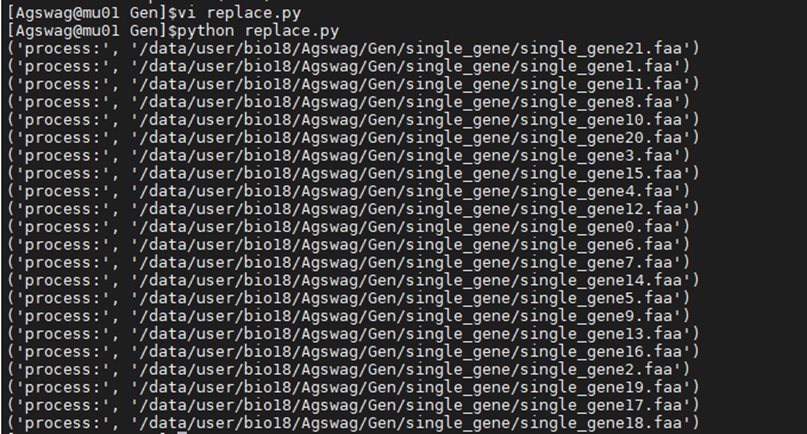


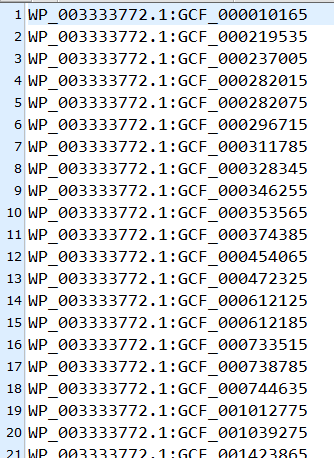
基于上一步的dump新文件，对于每一个单拷贝基因家族（即每一行），在all\_pro.faa中搜索所有相应的菌株ID及其序列，输出获得文件（如果有100个基因家族，则获得100个文件）

提取单拷贝基因家族序列

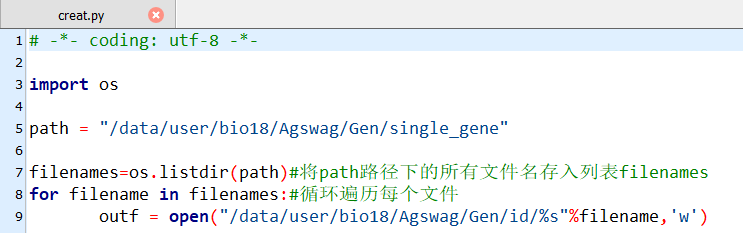
replace.py

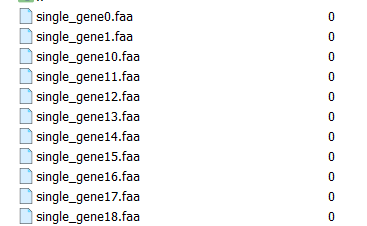


****

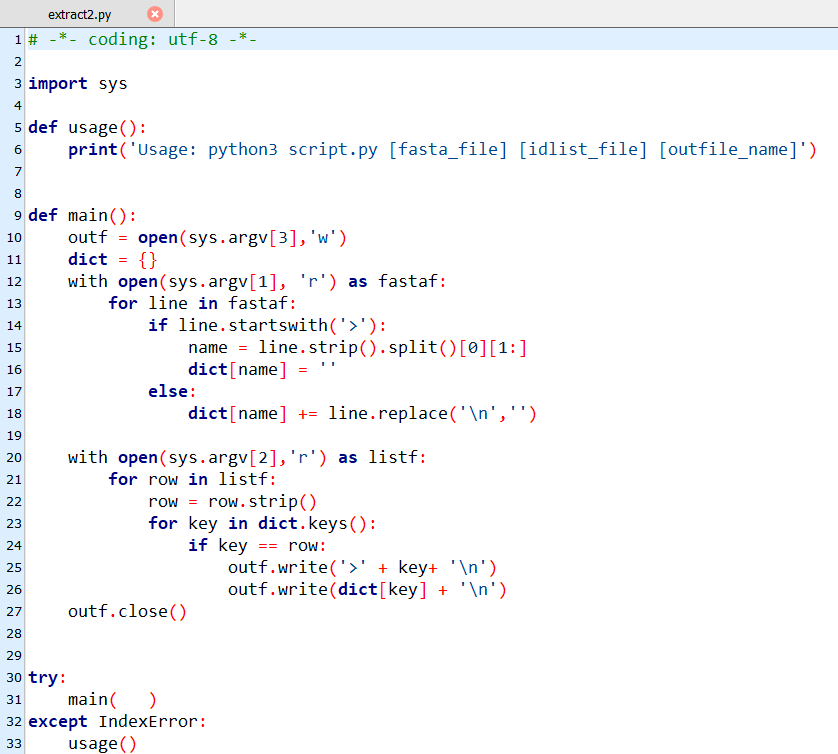
****

create.py#创建single\_gene下的同名空文件

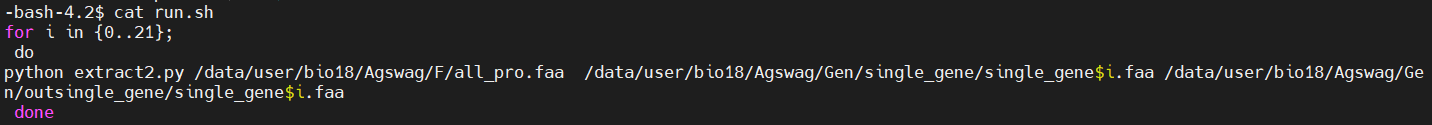


****

extract2.py#提取单个文件基因序列



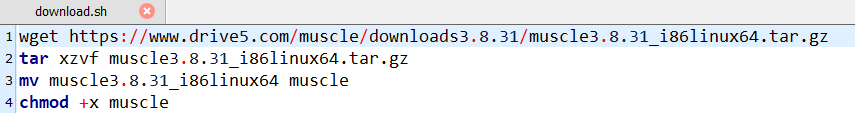
run.sh

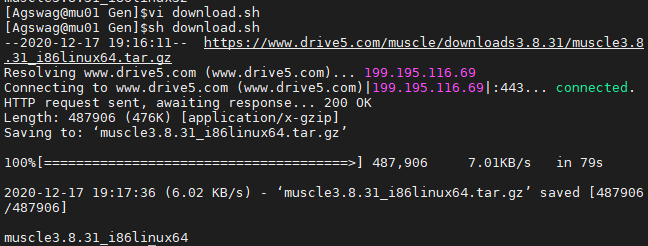


3.2 多序列比对

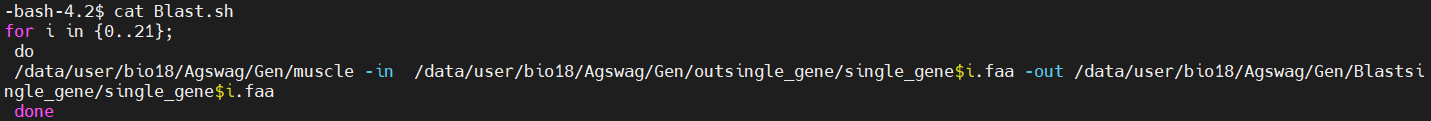
基于3.1输出的每一个文件，以muscle进行多序列比对（用shell或其他语言写到1个循环里即可），获得一个muscle输出文件

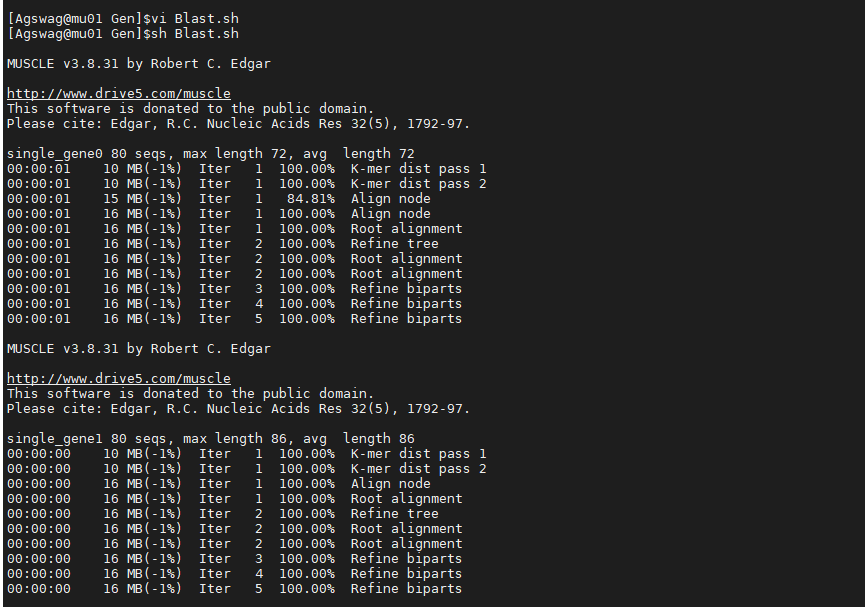
download.sh

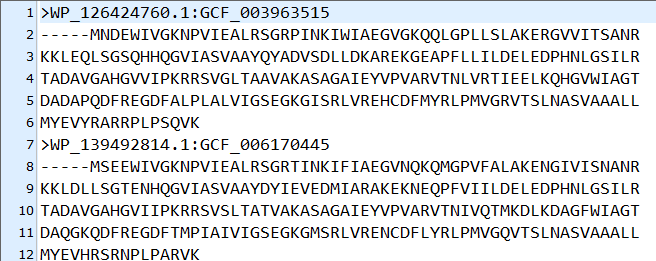
****

****

Blast.sh



****

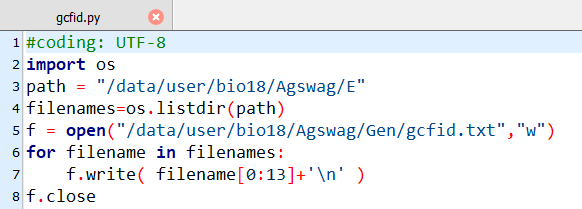


3.3 比对后序列合并

对于每个菌株的ID（GCF号），都需要在每一个muscle比对文件中搜索对应的序列，将序列逐个粘贴成1条完整的序列（如果我们有80个菌株，那么合并的文件应该有160行（fasta格式，一行ID、一行序列））



gcfid.py #得到80个GCF号

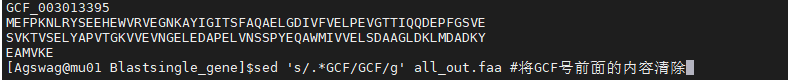




**将得到的22个比对后文件合并到all\_put.faa文件**

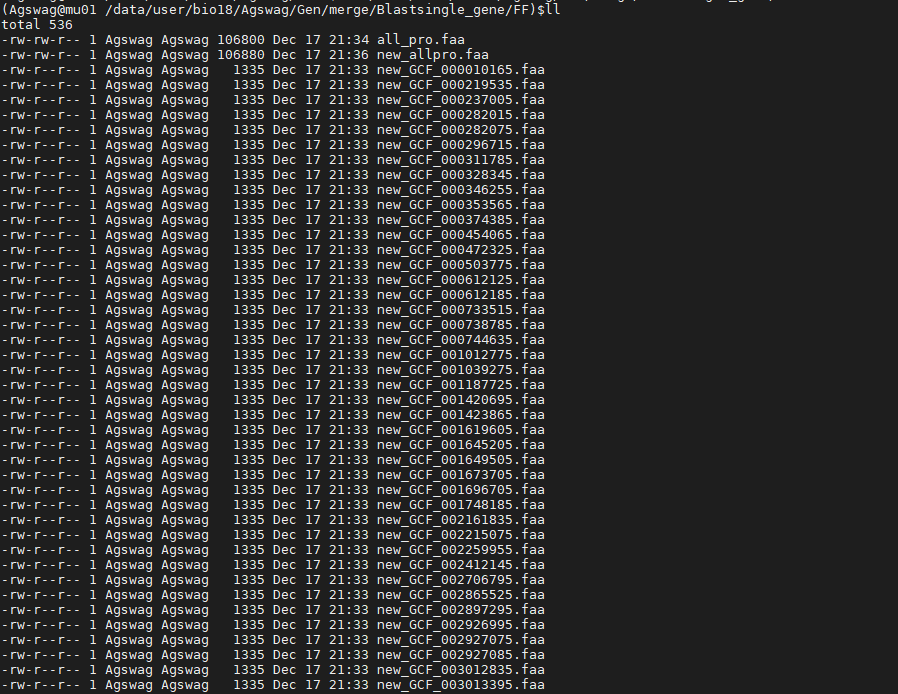
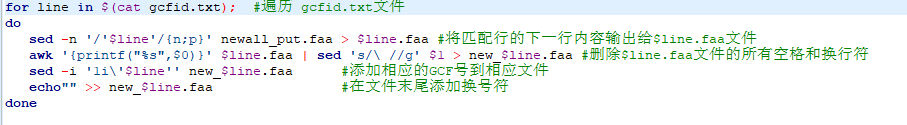


**#将GCF号前面的内容清除并输出到 newall\_put.faa 文件**

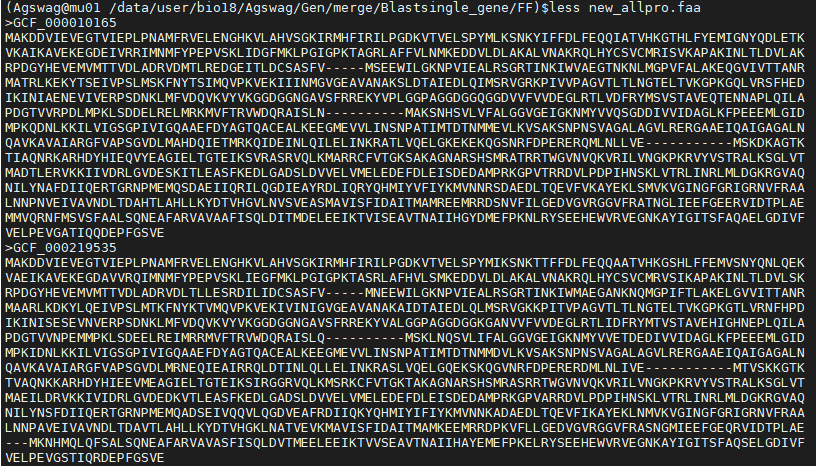




extract3.sh# 得到每个GCF号的相应序列并输出到相应的GCF号文件中

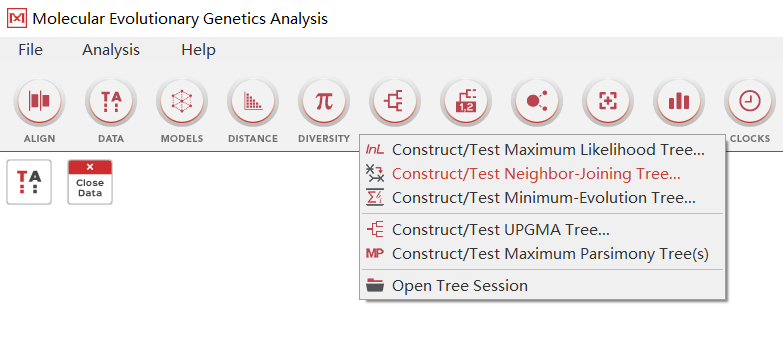


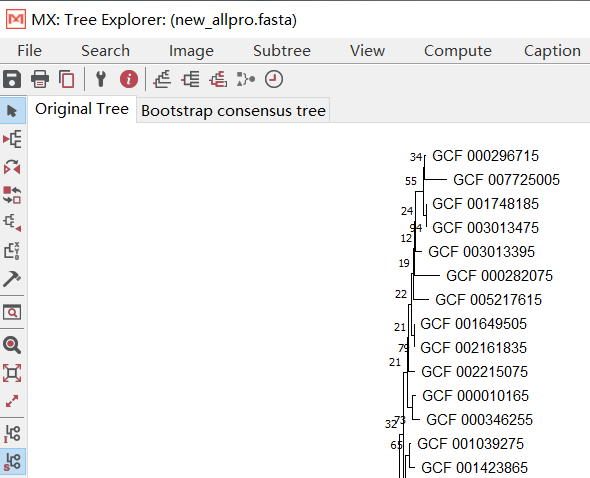
#将80个新的GCF序号文件合并

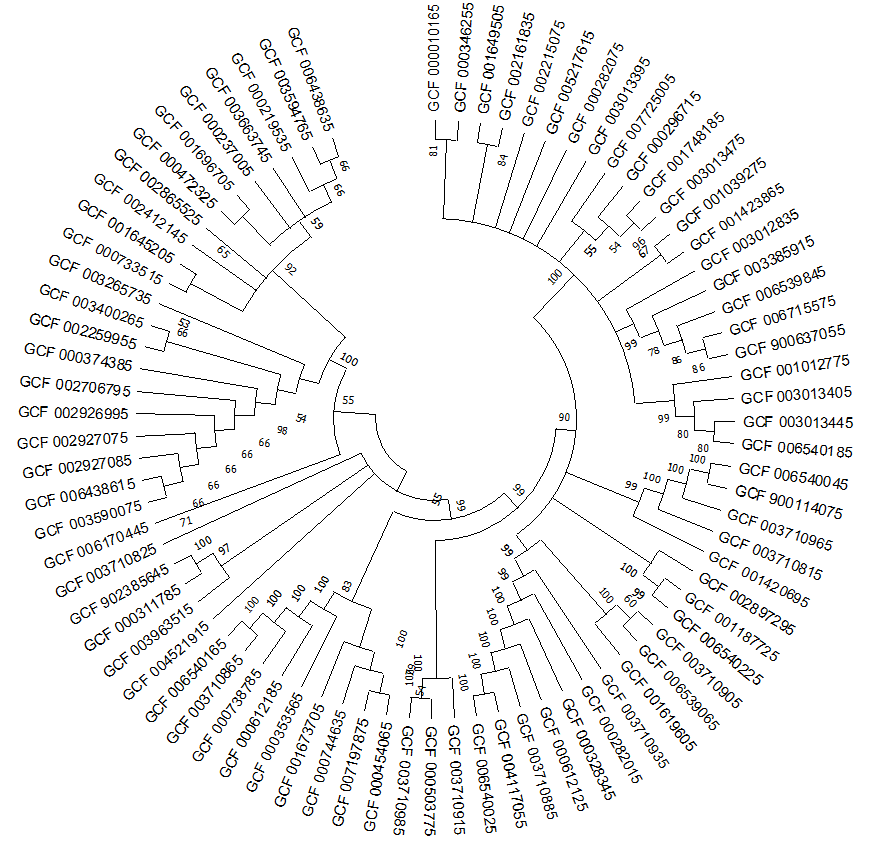


3.4 构建进化树

用MEGA构建出图，使用常见算法Neighbor-Joining（速度快）







**总结：**

分子进化树(以分子数据为依据构建的进化树)不仅精确地反映物种间或群体间在进化过程中发生的极微细的遗传变异(小至一个氨基酸或一个核昔酸差异),而且借助化石提供的大分子类群的分化年代能定量地估计出物种间或群体间的分化年代，这对进化论的研究而言无疑是一场革命。

序列比较是生物信息学中最频繁也是最有价值的工作。要知道一个序列(结构)与另一个序列(结构)或者与一批序列(结构)之间的差异，唯一的途径就是序列(结构)的比较分析。序列水平上的比较反映的是字符串之间的差异，能够发现碱基序列或者氨基酸序列的保守模式。但是，在分子生物学中，比较是多方面的，除了核酸或蛋白质序列的比较，也可以是结构的比较等。事实上，相差很大的序列可以形成具有相同功能的分子。而结构水平上的比较更能反映功能上的差异，能够发现与功能紧密相关的结构域。结构比较方面的工作都是围绕蛋白质及RNA展开的。

构建进化树的方法包括两种:一类是序列类似性比较，主要是基于氨基酸相对突变率矩阵（常用PAM250）计算不同序列差异性积分作为它们的差异性量度（序列进化树)﹔另一类在难以通过序列比较构建序列进化树的情况下，通过蛋白质结构比较包括刚体结构叠合和多结构特征比较等方法建立结构进化树。

该实验背后的生物学问题，简单说就是如何获得物种间更合理的演化关系。上述传统的仅基于物种序列的系统进化树太粗糙，不能只简单地根据序列的物理相似性来确定物种同源关系。生命是复杂的，基因通常是以网络、家族的形式来协调变化的，根据基因家族构建的psudo分子序列来构建进化树，理论上更接近真实的物种演化关系。