安装

https://www.megasoftware.net/，下载windows的GUI版本，要使用CC（命令行）版本--配置好环境变量即可。然后如果觉得windows配置不好，也可以安装linux版本（服务器），这里我选择ubantu CC（在官网中你可以直接下载能使用的二进制文件，也可以使用*.deb文件进行安装）。

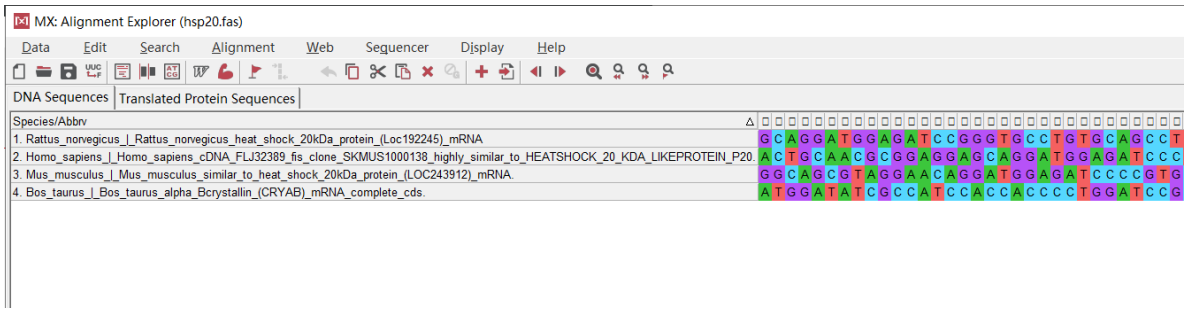
使用

> *分子进化的研究是核酸及氨基酸序列，究竟选择哪个？序列的选取要遵循以下原则：1）如果DNA序列的两两间的一致度≥70%，选用DNA序列。因为，如果DNA序列都如此相似，它的蛋白质会相似到看不出区别，这对构建系统发生树是不利的。所以这种情况下应该选用DNA序列，而不选蛋白质序列。2）如果DNA序列的两两间的一致度≤70%，DNA序列和蛋白质序列都可以选用。*

首先需要一个fasta文件，这在官网示例点击hsp20.meg,有一个四个物种没有比对好的fas文件（就是**fasta**文件）。

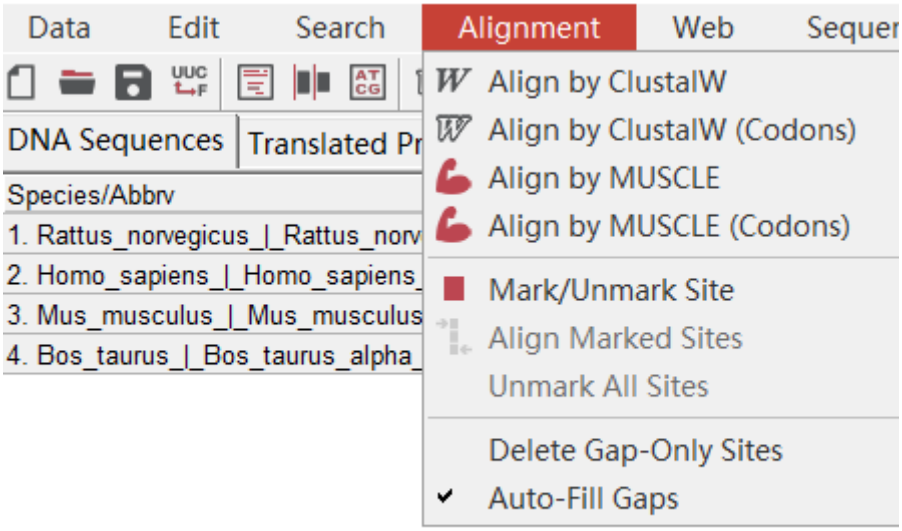
由于需要使用图片，这里我使用pdf文件分享，懒得把图片保存在云床上或者放在网站中了，虽然pdf也是放网站里，但是只需要操作一遍呀。（果然有图片什么的最烦了！）

pdf路径：pdf，看后面内容直接使用pdf吧。



可以看到一个DNA seq跟翻译的蛋白序列，然后还有个Display的栏可以更改序列的查看方式，比如换成没有背景颜色等等。具体可以自行试试。

这里我们点击Alignment栏，可以看到



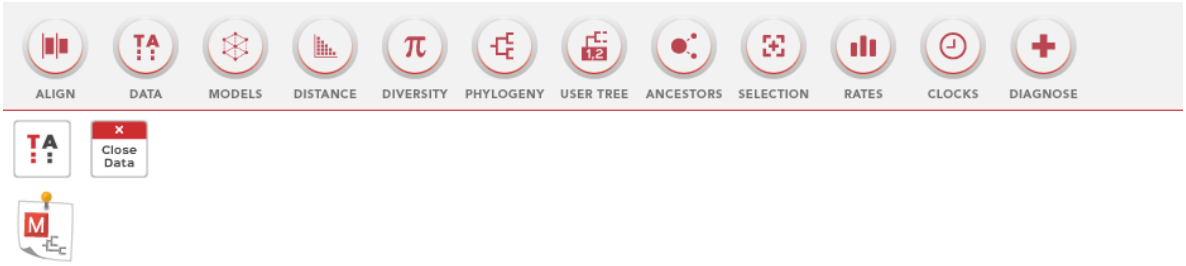
有两种比对方法：ClustalW跟MUSCLE（貌似还有一个叫T-coffee）

- ClustalW是现在用的最广和最经典的多序列比对，是目前使用最广泛的多序列比对程序。（而且也可以用于双序列比对）**它采用的是一种渐进的比对方法(progressive methods)，先将多个序列两两比对构建距离矩阵，反映序列之间两两关系；然后根据距离矩阵计算产生系统进化指导树，对关系密切的序列进行加权；然后从最紧密的两条序列开始，逐步引入临近的序列并不断重新构建比对，直到所有序列都被加入为止。**
- Muscle的速度比较快，比clustalw的速度快几个数量级,而且序列数越多速度的差别越大。不过只能用于多序列，之所以比clustalw快一方面是因为没有进行两两序列比对。

当然，也可以单独安装这几款软件进行使用。（或者使用在线工具之类的，比如 https://www.ebi.ac.uk/Tools/msa/clustalw2/）

对于示例文件这种小文件，我们使用ClustalW是不错的选择。T-coffee相比似乎更慢更精确。（不过当文件相似度>80%时，三个程序精确度都在90%以上）

> *比对结束后可以保存好比对后的fas文件，或者保存成meg文件之类的，或者每种格式都保存一次。毕竟运行一次多序列比对文件大挺耗时。*
>
> *fas文件保存的是比对后文件，meg文件可以进行下一步的进化树分析。*



这里我们就可以点击PHYLOGENY进行进化树分析了，有多种方法根据距离矩阵构建进化树，之后就可以看进化树了。

linux CC

首先在GUI界面选择

之后即可进行设置生成*.mao文件了。

参数深入理解

方法名	方法名
ML，Maximum likelihood	最大似然法
NJ，Neighbor-Joining	邻接法
MP，Maximum parsimony	最大简约法
ME，Minimum Evolution	最小进化法
Bayesian	贝叶斯推断
UPGMA	不常用

进化树分析目前相对常用的方法是NJ，一篇综述（Hall BG. Mol Biol Evol 2005, 22(3):792-802）认为贝叶斯的方法最好，其次是ML，然后是MP。

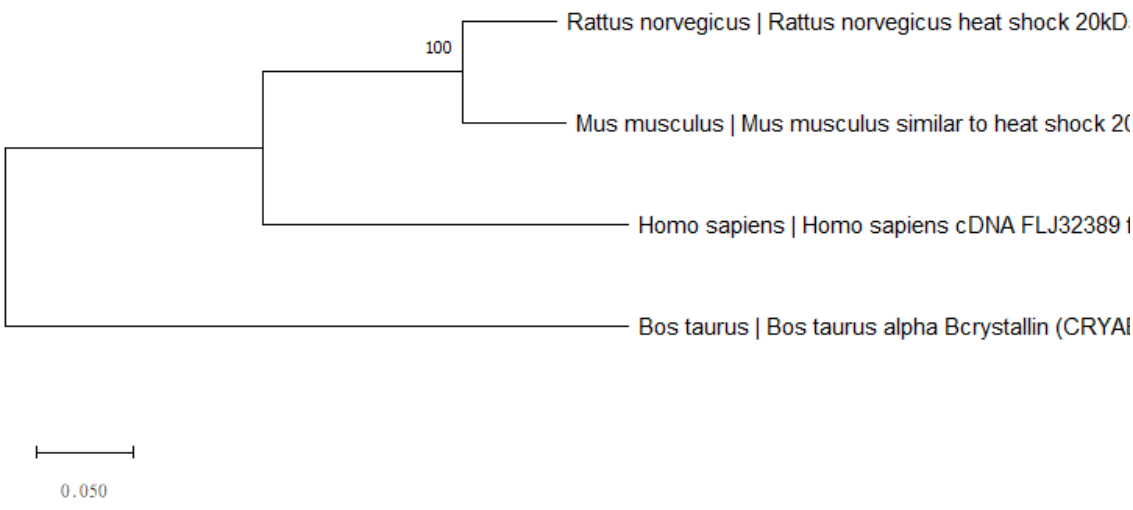
MX: Analysis Preferences

Phylogeny Reconstruction

Option		Setting
ANALYSIS		
Scope	→	*All Selected Taxa*
Statistical Method	→	*Neighbor-joining*
PHYLOGENY TEST		
Test of Phylogeny	→	Bootstrap method
No. of Bootstrap Replications	→	*500*
SUBSTITUTION MODEL		
Substitutions Type	→	*Nucleotide*
Genetic Code Table	→	Not Applicable
Model/Method	→	*Maximum Composite Likelihood*
Fixed Transition/Transversion Ratio	→	Not Applicable
Substitutions to Include	→	*d: Transitions + Transversions*
RATES AND PATTERNS		
Rates among Sites	→	*Uniform Rates*
Gamma Parameter	→	Not Applicable
Pattern among Lineages	→	*Same (Homogeneous)*
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	*Pairwise deletion*
Site Coverage Cutoff (%)	→	Not Applicable
Select Codon Positions	→	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites
SYSTEM RESOURCE USAGE		
Number of Threads	→	*4*

Help | Cancel | OK

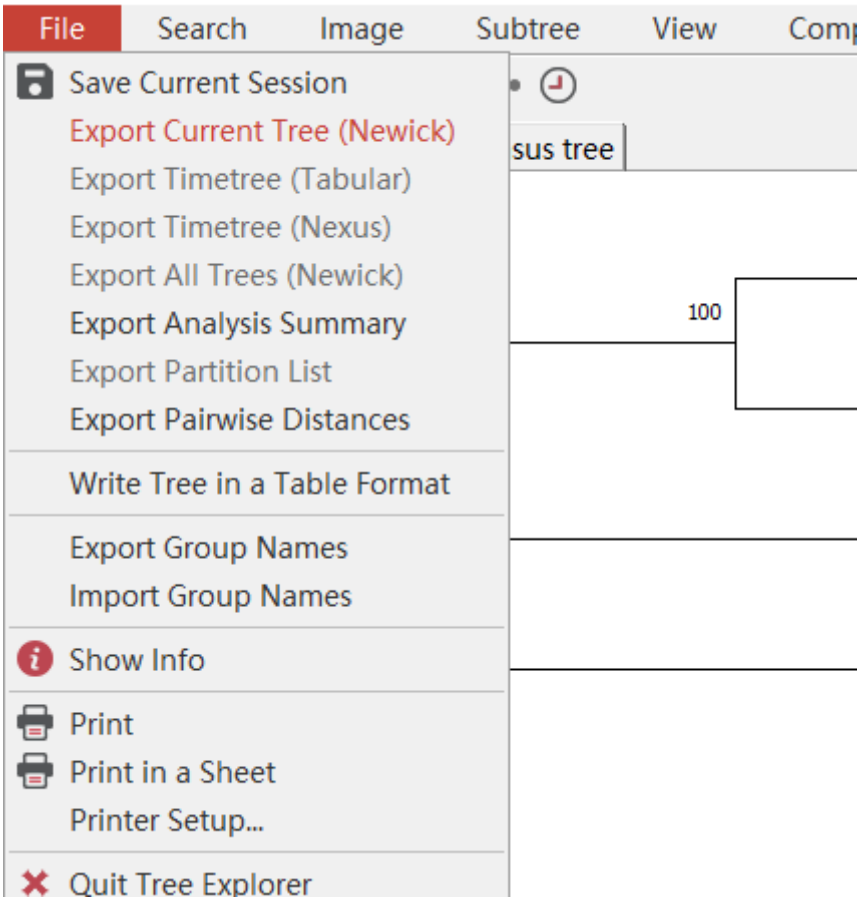
之后test选择Bootstrap method ,点击OK。得到



其中100为bootstrap值,大于70的表示这个节点比较可靠。

通过工具栏可以对树的形状进行调整。可以保存成pdf等。pdf可以用AI美化。

记得保存树文件Newick,方便美化。（需要的话把bootstrap值跟branch length保存下来，branch的值应该是0.1等小数值）



进化树美化

可以选择iTol或者R包ggtree,由于我是个gg控，所以先选择ggtree来玩玩呗。

序列比对分析

1.序列相似性比较和序列同源性分析

序列相似性比较：将待研究序列与DNA或蛋白质序列库进行比较，用于找出与此序列相似的已知序列。完成这一步只需要两两序列比对的算法。例如:BLAST、FASTA。
序列同源性分析：将待研究序列与一组与之同源，但来自不同物种的序列进行多序列比较，以确定该序列与其他序列间的同源性大小。完成这一步需要多序列比对算法。例如:Clustal。

2.序列同源性分析(多序列比对)的意义

- 用于描述一组序列之间的相似性关系，以便了解一个基因家族的基本特征，寻找motif,保守区域。(motif:是蛋白质分子具有特定功能的或者作为一个独立结构域一部分相近的二级结构聚合体)
- 用于描述一个同源基因之间的亲缘关系的远近，应用到分子进化分析中。即是进化分析。
- 其他：构建profile、打分矩阵。