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）。

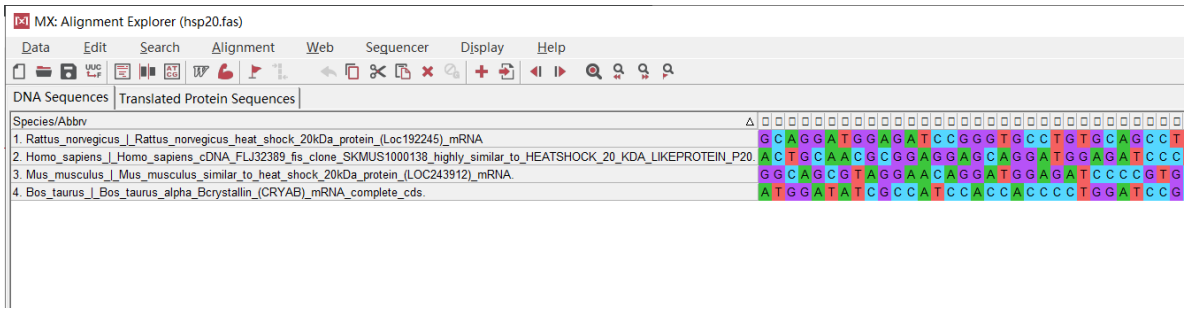
使用

> *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）。

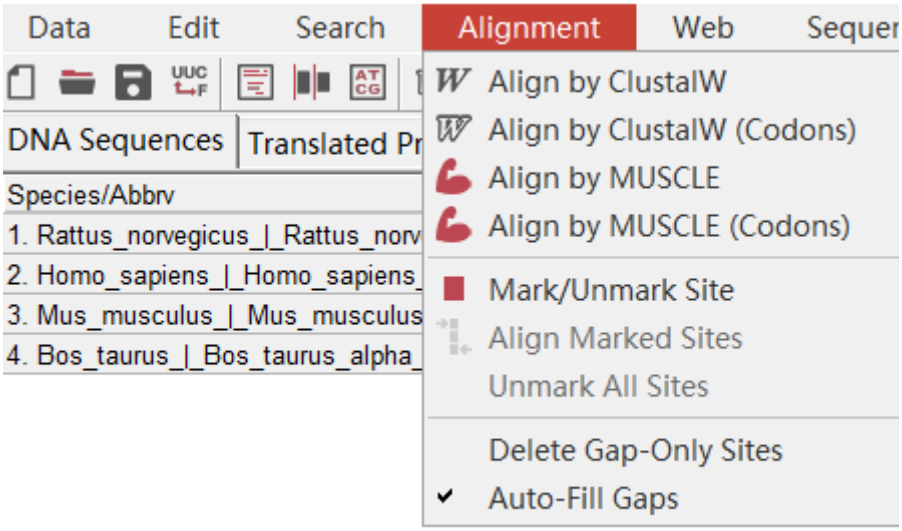
由于需要使用图片，这里我使用pdf文件分享，懒得把图片保存在云床上或者放在网站中了，虽然pdf也是放网站里，但是只需要操作一遍呀。（果然有图片什么的最烦了！）

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

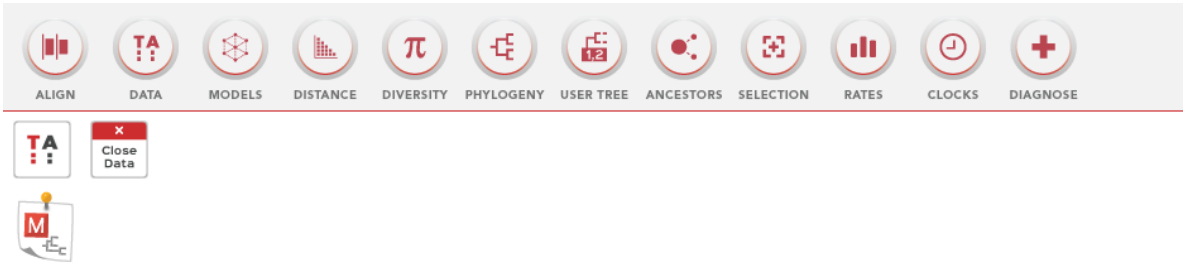


有两种比对方法：ClustalW跟MUSCLE（貌似还有一个叫T-coffee）

- 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。

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）

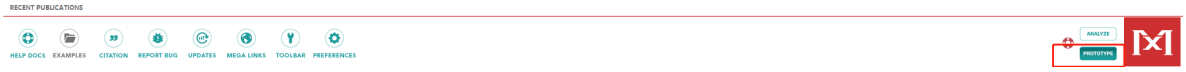
> *比对结束后可以保存好比对后的fas文件，或者保存成meg文件之类的都行，或者每种格式都保存一次。毕竟运行一次多序列比对文件大挺耗时。*
>
> *fas文件保存的是比对后文件，meg文件可以进行下一步的进化树分析。我一般还保存好比对好的fasta文件，方便后续其他分析。*



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。

比如先试下序列比对

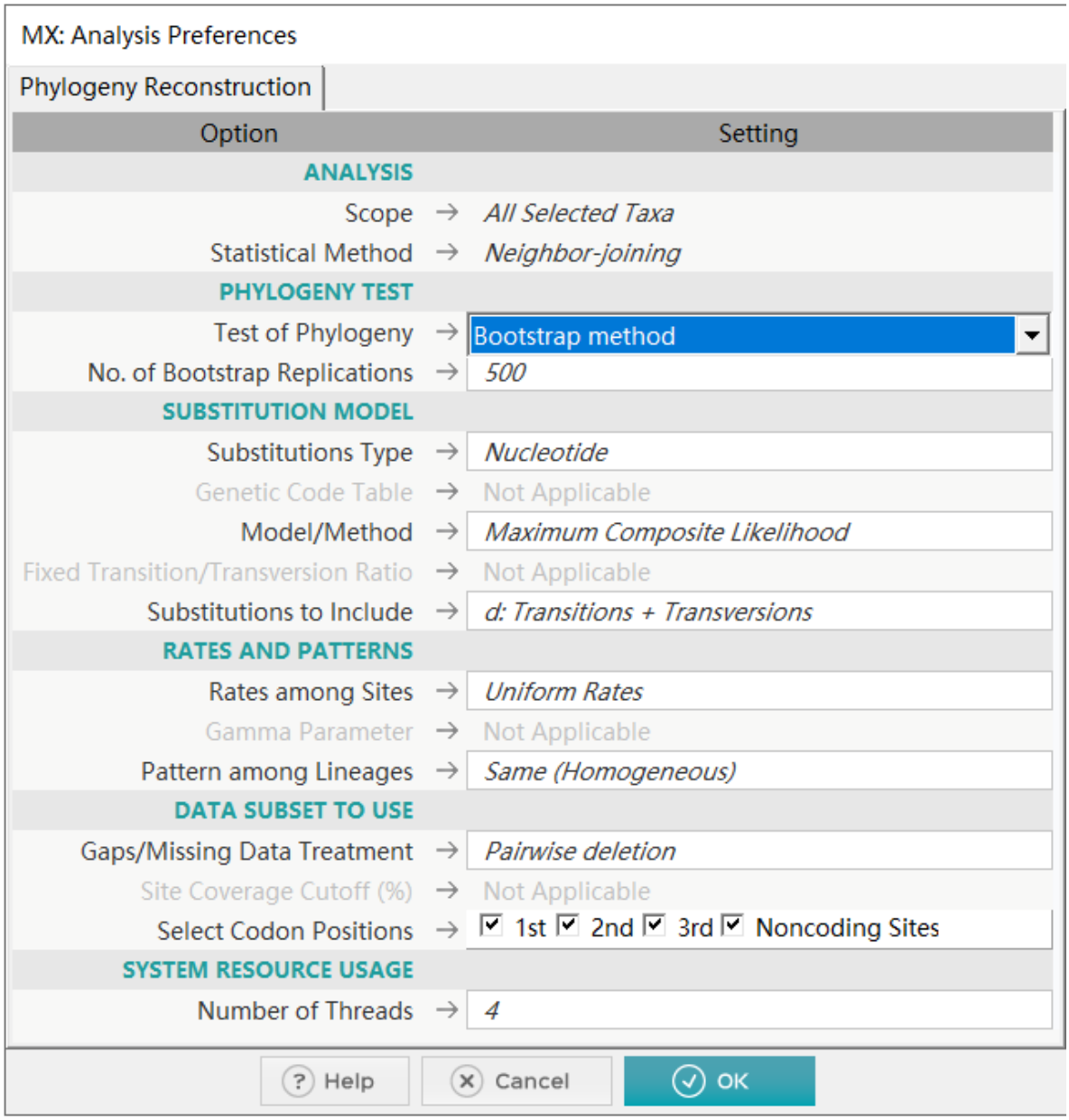
```
megacc -a *.mao -d *.fasta -f fatsa -o ./
```

-f保存成meg文件的时候不知道为什么用windows的megax一直打开失败，虽然你可以直接在linux建树，但是-f fasta可以直接megax打开建树，毕竟建树时间不长，所以我喜欢这个工作模式。

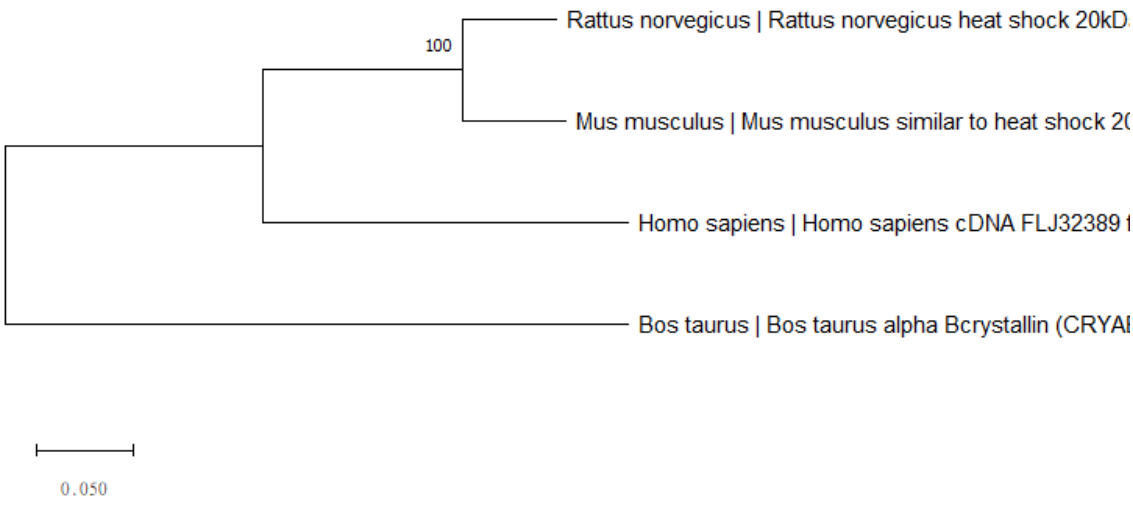
参数深入理解

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。



之后test选择Bootstrap method ,点击OK。得到

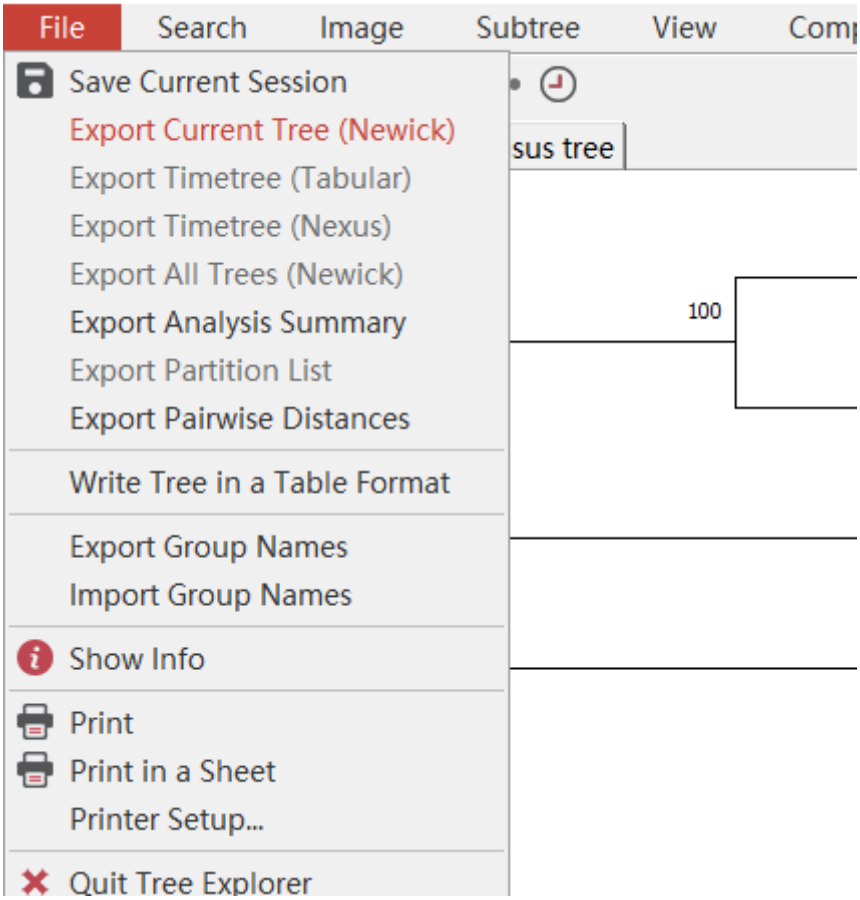


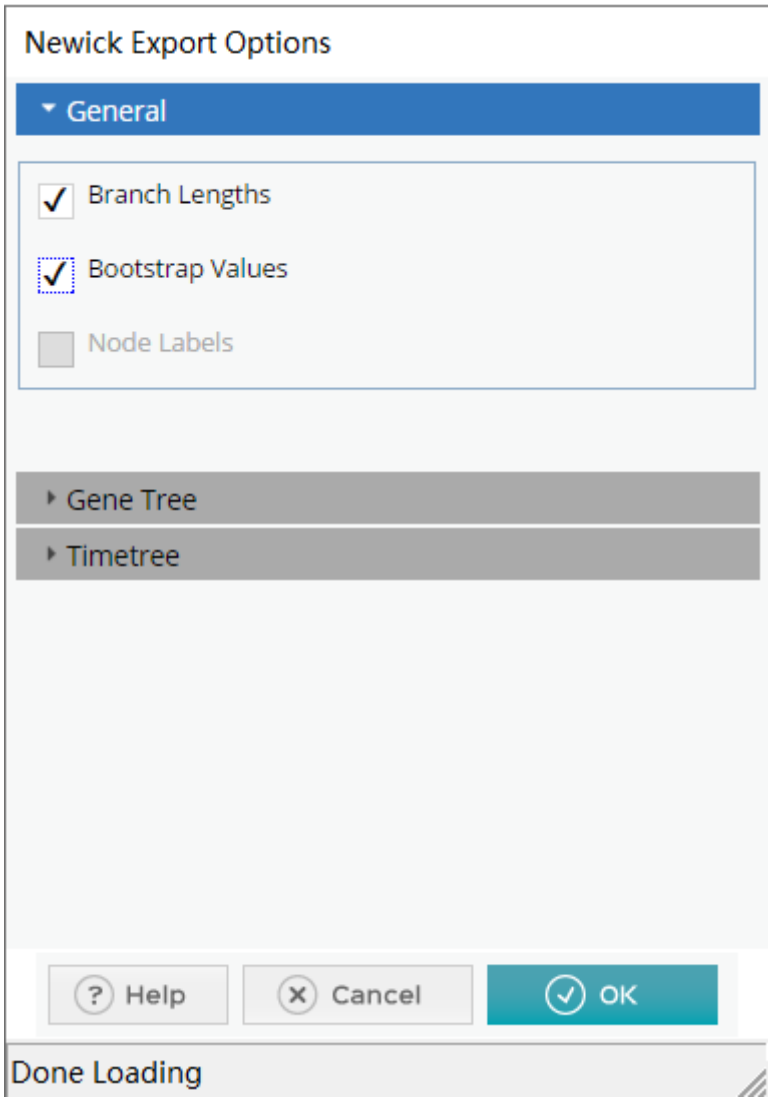
其中100为bootstrap值,大于70的表示这个节点比较可靠。

> *boostrap值含义:即自展值，可用来检验所计算的进化树分支可信度。Bootstrap几乎是构建系统进化树一个必须的选项。一般Bootstrap的值>70(或者70%)，则认为构建的进化树较为可靠。如果Bootstrap的值太低，则有可能进化树的拓扑结构有错误，进化树是不可靠的。*

通过工具栏可以对树的形状进行调整。（右键树的分枝等也能修改颜色）可以保存成pdf等。pdf可以用AI美化。（original tree是按照比例尺画的，Boostrap consensus tree则展示了树的关系图，不过original tree似乎会截断太长的树，我也不知道怎么调，所以最好选择其他画树工具。）

记得保存树文件Newick,方便美化。（最好把bootstrap值跟branch length保存下来）





保存下来的newick可以用iTOL或者ggtree等美化。iTOL使用可以直接自己尝试。

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。

多序列比对其他工具

当然，也可以安装**MUSCLE**等软件进行使用。（或者使用在线工具之类的，比如 https://www.ebi.ac.uk/Tools/msa/clustalw2/）

T-coffee目前还没用过。

R中多序列比对

msa: an R package for multiple sequence alignment文章介绍了开发的msa包，这个包可以在bioconductor获得。【ps:因为依赖Biocstrings,所以安装起来还挺麻烦，不过遇到错误慢慢解决吧，报错原因大部分是因为网络或者依赖的包版本不对之类的】

而这个包封装了msaPrettyPrint函数绘图，十分不错。当然你也可以使用mega打开fasta文件进行查看，截图并AI美化。或者选择试试R 包ggmsa。

用mega的hsa序列试下，官网下的示例文件的那条hsa.meg我命名为了hsa.fasta(个人习惯比对好的以后缀名.fas表示)

```
library(msa)
fasta <- readDNAStringSet("./hsp20.fasta")
fasta
```

```
> fasta
DNAStringSet object of length 4:
    width seq                                                 names
[1]  1310 GCAGGATGGAGATCCGGGTGCCT...CCAATAAATGCACTTGAGATTT Rattus_norvegicus...
[2]  1457 ACTGCAACGCGGAGGAGCAGGAT...AATAAATGTGCTAGAGCTCTGC Homo_sapiens | Ho...
[3]  1309 GGCAGCGTAGGAACAGGATGGAG...CAATAAATGCACTTGAGATTTG Mus_musculus | Mu...
[4]   632 ATGGATATCGCCATCCACCACCC...ACGGATTCTCTAGAAATATCCT Bos_taurus | Bos ...
>
```

很漂亮。

```
fas <- msa(fasta,method = "Muscle")
fas
print(fas,show="complete") #print全部比对好的序列
```

msa函数目前有三种比对方法，可以根据需要选择相应方法。

msaprettyprint函数的使用

不论是在linux还是windows，你都需要这个pdflatex程序。

我们需要Texlive这软件，安装很麻烦[1]，或者说用apt下载。

最后我选择使用R包tinytex安装。当然你弄过R markdown基本就是有这程序了。

devtools的安装就不在这儿说了，也非必要要用devtools下载。

```
devtools::install_github('yihui/tinytex')
tinytex::install_tinytex()
```

linux一般会帮你下在~/.TinyTeX目录下，**使用前请配好环境变量**，msa才能使用。

windows可以自己下个CTEX之类的，或者用tinytex下就行，使用时仍旧先配好环境变量。

在Linux下使用这函数你完全可以按照官方的说明文档使用。

在windows我遇到了找不到文件fasta文件及找不到texshade.sty文件的报错。虽然一开始我是通过修改msaprettyprint函数生成的tex文件解决了，但是稍微阅读了下源代码，发现比较便捷的方式处理找不到文件的错误。接下来记录下这两种解决方式。

官方文档代码：

```
library(msa)
mySequenceFile <- system.file("examples", "exampleAA.fasta",
package="msa")
mySequences <- readAAStringSet(mySequenceFile)
myFirstAlignment <- msa(mySequences)
msaPrettyPrint(myFirstAlignment, output="pdf", showNames="none",
showLogo="none", askForOverwrite=FALSE, verbose=FALSE)
```

最后这函数在我的当前目录下生成了myFirstAlignment.tex文件之后报错了。

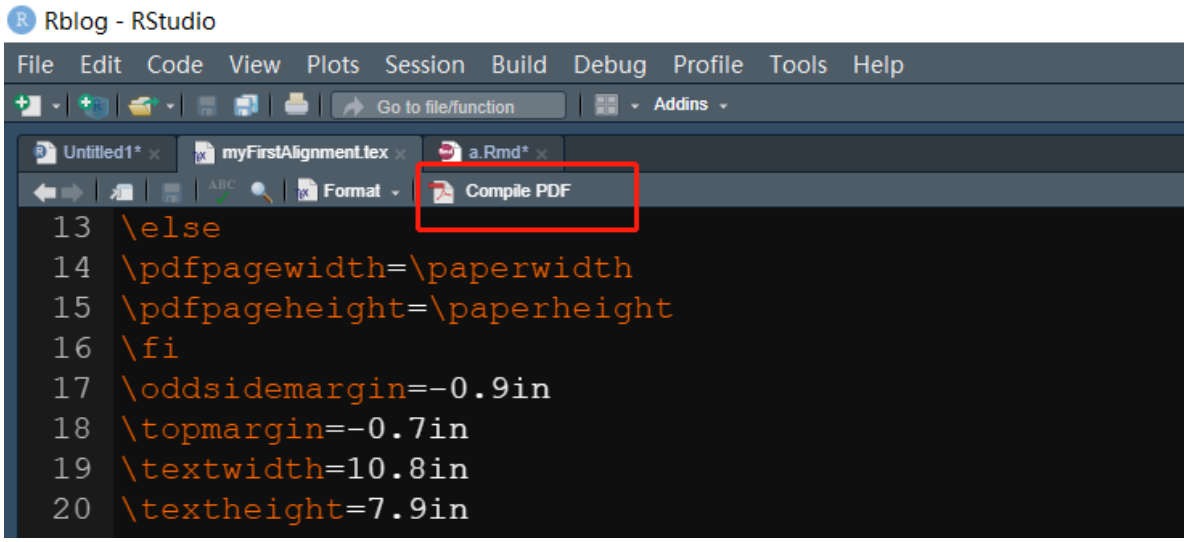
```
> msaPrettyPrint(myFirstAlignment, output="pdf", showNames="none",
+ showLogo="none", askForOverwrite=FALSE, verbose=FALSE)
Error in file(con, "r") : 无法打开链结
此外: Warning message:
In file(con, "r") :
  无法打开文件'myFirstAlignment.log': No such file or directory
>
```

通过打开tex文件，

```
\begin{texshade}{C:/Users/ADMINI~1/AppData/Local/Temp/RtmpIZbhkm
/seq2ea810e2109b.fasta}
```

会发现这一栏这个路径我们在windows是打不开的，这里需要改成正确的能打开的路径。

修改之后使用Rstudio的Compile PDF也发生报错，



找不到texshade.sty文件，而这个文件所在路径在：

```
system.file("tex", "texshade.sty", package="msa")
```

无奈看下源码，之后使用

```
tools::texi2dvi("./myFirstAlignment.tex", quiet = F, pdf =T,
texinputs = system.file("tex", package = "msa"), clean = TRUE, index
= FALSE)
```

就解决了。**ps:麻烦**

阅读源码及稍微看下帮助文档，

```
msaPrettyPrint(myFirstAlignment, output="pdf", showNames="none",
showLogo="none", askForOverwrite=FALSE, verbose=T,alFile =
"a.fasta")
```

运行上述命令即可完成编译。

`alFile = "a.fasta"`在当前目录指定生成fasta文件，`verbose = T`即是运行`tools::texi2dvi`函数,所以在windows使用时你不是utf-8编码估计应该都需要加入这两个参数。

---

1. https://stone-zeng.github.io/2018-05-13-install-texlive-ubuntu/ ↩