**背景：**

多序列比对（MSA）通常是三个或更多个相似长度的生物序列（蛋白质或核酸）的比对。从输出中，可以推断出同源性，并且可以研究序列之间的进化关系。[1]

在代表性冠状病毒系统发育关系分析中，作者使用clustalX进行多序列比对。[2]

**方法：**

1.工具选择

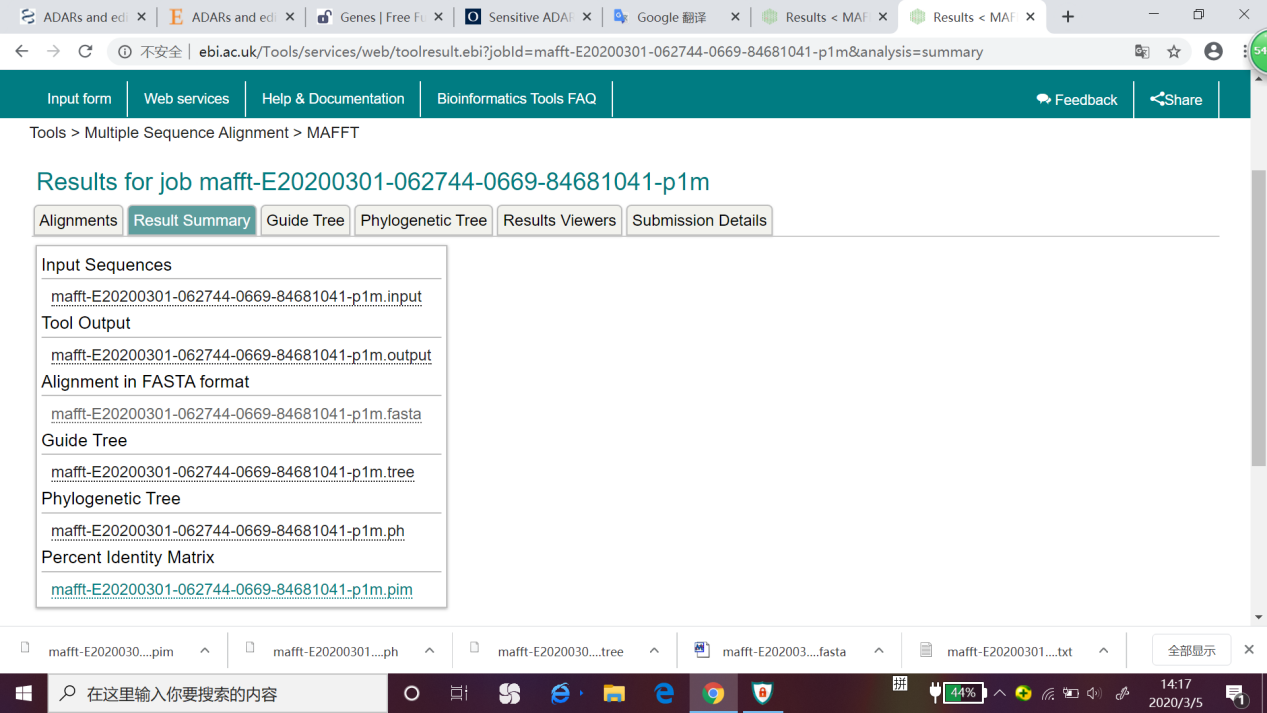
访问clustal官网，链接到官网推荐的在线比对工具EBI网络服务器，该服务器由欧洲分子生物学实验室提供支持（https://www.ebi.ac.uk/Tools/msa/）。由于原先提供的clustalW2服务已下线，对DNA的多序列比对，EBI推荐使用MUSCLE（准确的多序列比对工具，尤其擅长蛋白质比对，适用于中型比对）或MAFFT（使用快速傅里叶变换的多序列比对工具，适用于中至大型比对）。故选择MAFFT方法进行冠状病毒核酸序列比对。（MUSCLE由于运行中的未知错误，未能得到可用结果）

2.序列输入

从NCBI官网上获取了代表性冠状病毒的全基因组序列以及新型冠状病毒（2019-nCoV或SARS-CoV-2）多条全基因组序列，将以上序列按其格式要求输入到MAFFT中，选择序列类型为DNA，并选择输出格式为Pearson/FASTA。

3.结果获取

多序列比对结果储存到该服务器提供的临时站点页面中，从页面中下载多序列比对结果及引导树、系统树等。



将.fasta格式的多序列比对结果输入到MEGA7中以最大似然法（ML）构建系统发育树，选择进行步长检验并接受所有默认参数。

将构建完成的ML树导入word 2016进行序列名称修改和着色。

**结果：**

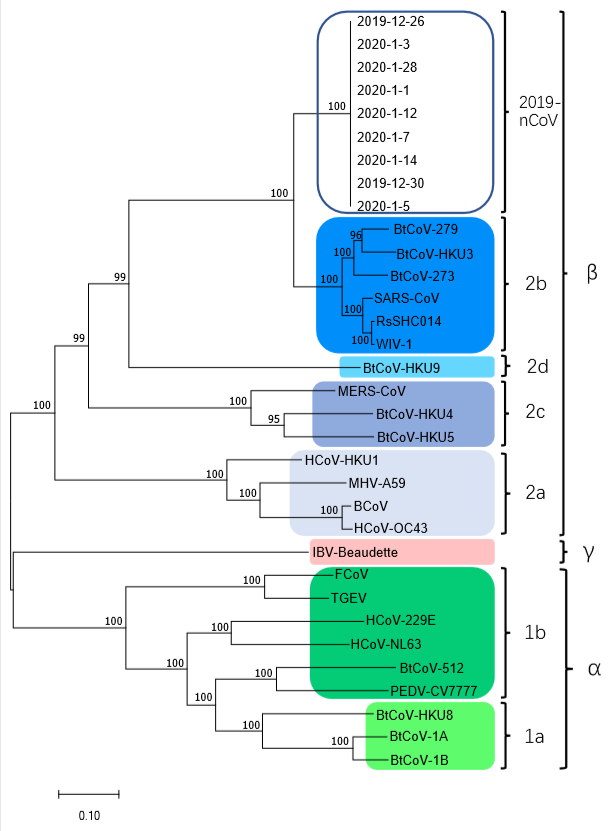
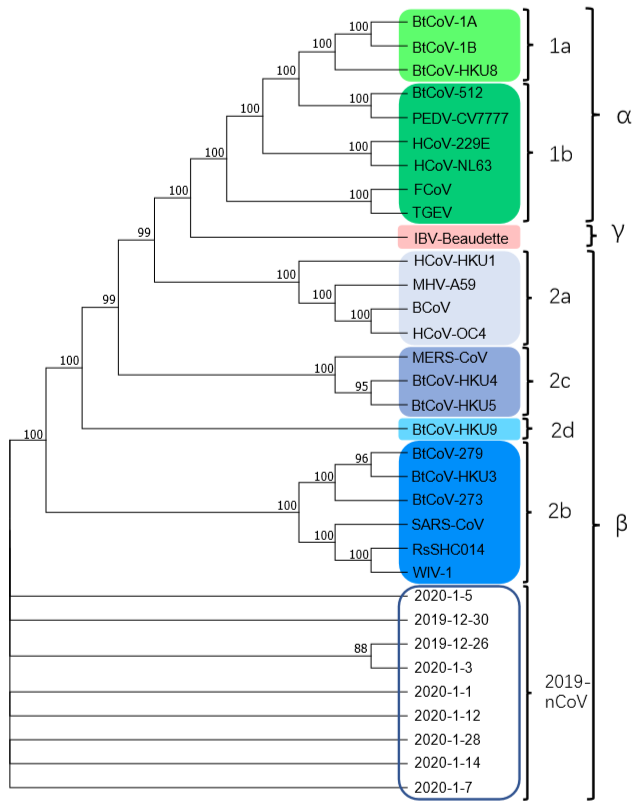
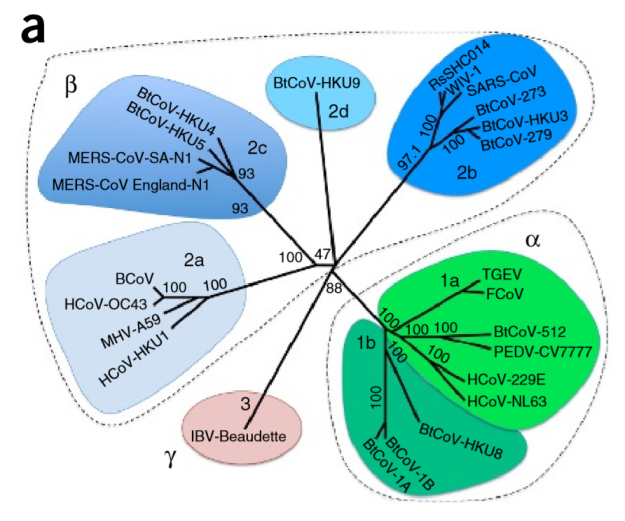
a.b.

图1.a.最大似然法系统发育树（original tree）。b.最大似然法系统发育树（bootstrap consensus tree）蓝色方框内表示新型冠状病毒全基因组序列，其他序列（即所有彩底色方框内的序列）为代表性冠状病毒全基因组序列[2]。2019-nCoV序列与2b簇处于同一分支内，两者有高度相似性。



参考文献[2]的原图，在MEGA中做的放射状系统树图视觉效果一般，这里只展示了默认形状的系统树。

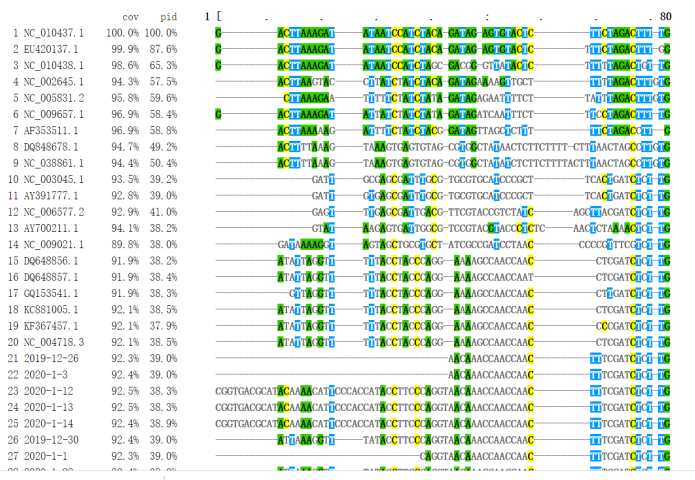
**References:**

[1]. Madeira, F., et al., The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res, 2019. 47(W1): p. W636-W641.

[2]. Menachery, V.D., et al., A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med, 2015. 21(12): p. 1508-13.

一些疑问以及可能不足需改进的地方：

1.多序列比对的原始界面（下左截图，来自于在线工具MView，似乎与Boxshade、ESpript、TeXshade类似）很长，应该如何更好地展示（着色/特定位点的序列比对展示）？以及其他图象的调整美化



2.由于代表性冠状病毒序列从参考文献[2]中寻找，是否需要与其使用相同的软件（clustalX）？有资料显示MAFFT的准确性和比对速度都优于clustal（https://blog.csdn.net/flyfrommath/article/details/52872548?ops\_request\_misc=%7B%22request%5Fid%22%3A%22158295376319724835846489%22%2C%22scm%22%3A%2220140713.130056874..%22%7D&request\_id=158295376319724835846489&biz\_id=0&utm\_source=distribute.pc\_search\_result.none-task）

3.序列比对的运行结果中包含的信息很多，哪些需要给予更多关注？

4.序列比对工具的使用中大部分保留了默认的参数设置，是否有些需要调整？

5.任务要求探究RsSHC014重组病毒与2019-nCoV是否密切同源，尝试进行两者全基因组序列的成对序列比对（Pairwise Sequence Alignment），得到一个txt文件，保存到文件夹“双序列比对（output）”中。但不太懂不同方法间的主要区别，以及如何可视化或描述