“实用生物信息技术”课程小组讨论总结报告

组：G2　　次：4　　组长：边汉青　　执笔：边汉青

1. 时间

2026.4.30

2. 方式

线上

3. 主题

对于构建系统发育树内容的复习和总结与讨论。

4. 内容

A 系统发育树的构建,以人的珠蛋白家族 12 个蛋白质序列为例。

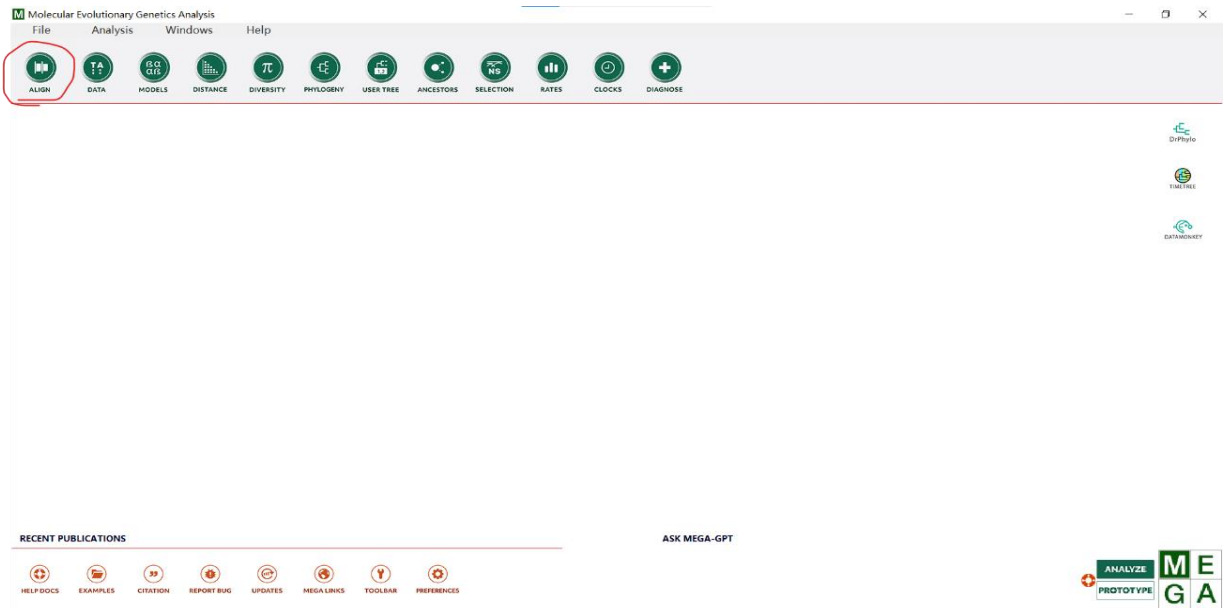
B 以人、小鼠、大鼠珠蛋白系统发生树构建，找出“先有物种、后有基因”和“先有基因、后有物种”的实例。

C 7 个代表性脊椎动物 alpha 血红蛋白系统发生树构建。

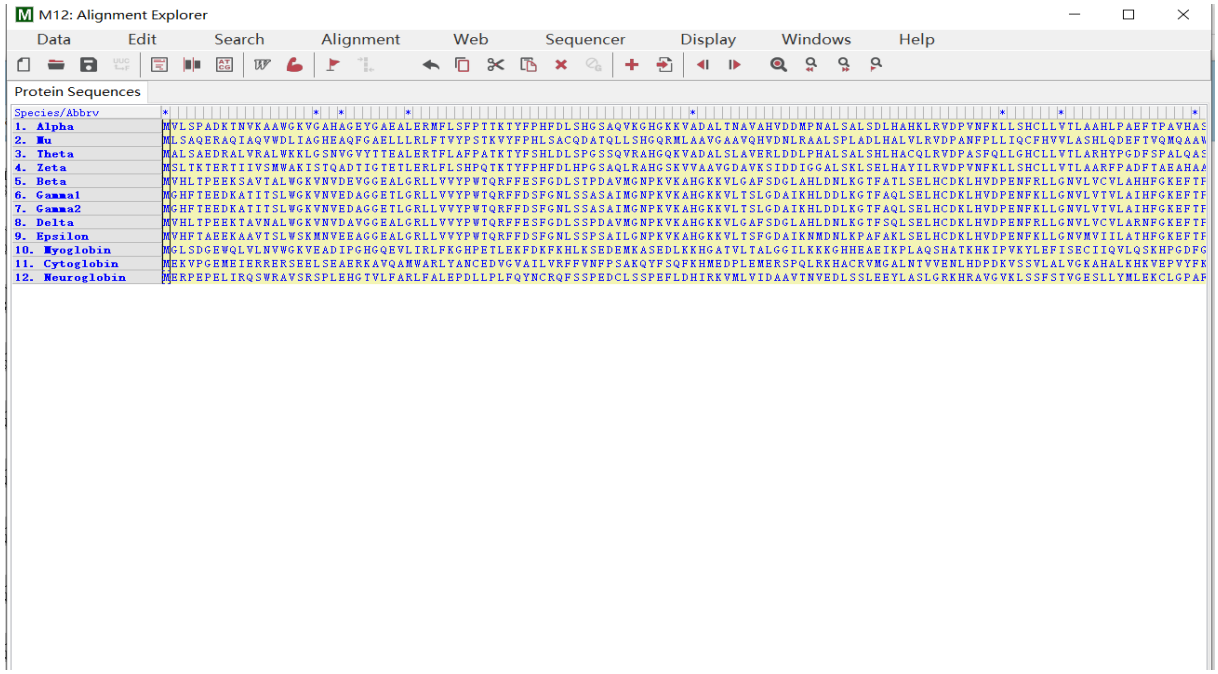
D 7 个代表性植物 18S rRNA 系统发生树构建

A 系统发育树的构建,以人的珠蛋白家族 12 个蛋白质序列为例

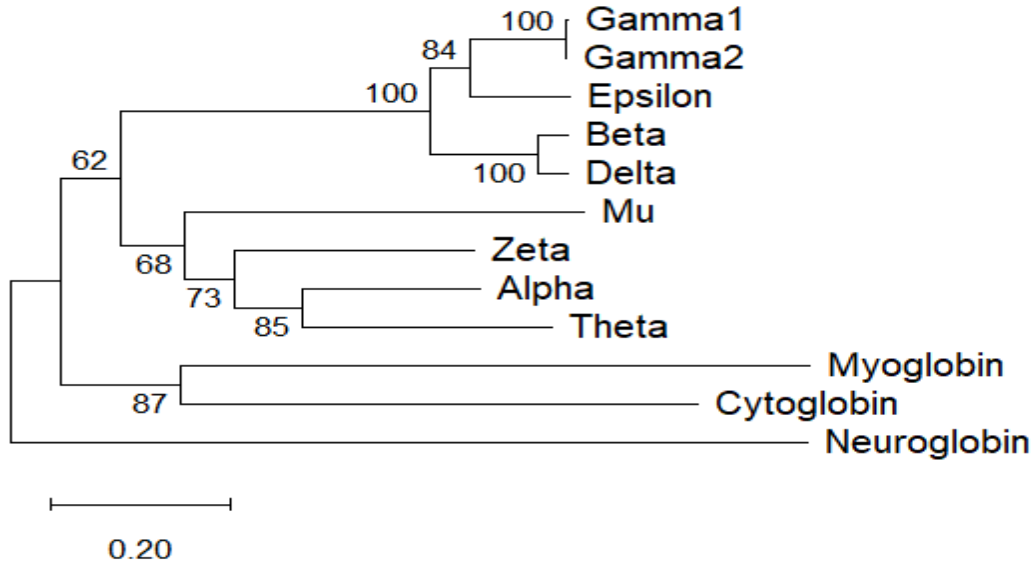
1）点开 MEGA 选择 ALIGN



2）将蛋白序列复制输入其中，之后选择 alignment 中的 align by clustalw。



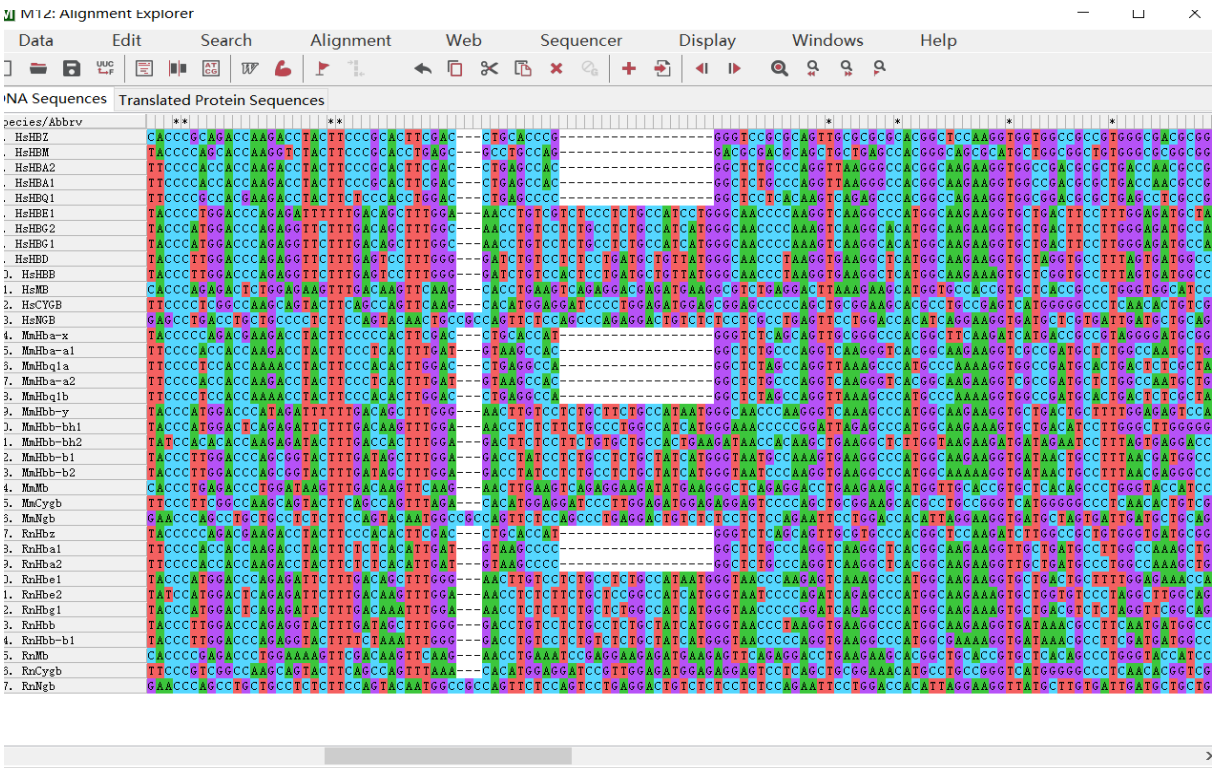
3）查看比对结果，点击 data 中的 Phylogenetic Analysis 和主界面的 Phylogeny，下拉菜单中选择邻接法 Construct/Test Neighbor-Joining Tree，在弹出窗口中将系统发生树稳定性测试 Test of Phylogeny 选项改为自举法 Bootstrap Method，将自举法重复次数改为 100，便可得到系统发育树结果。



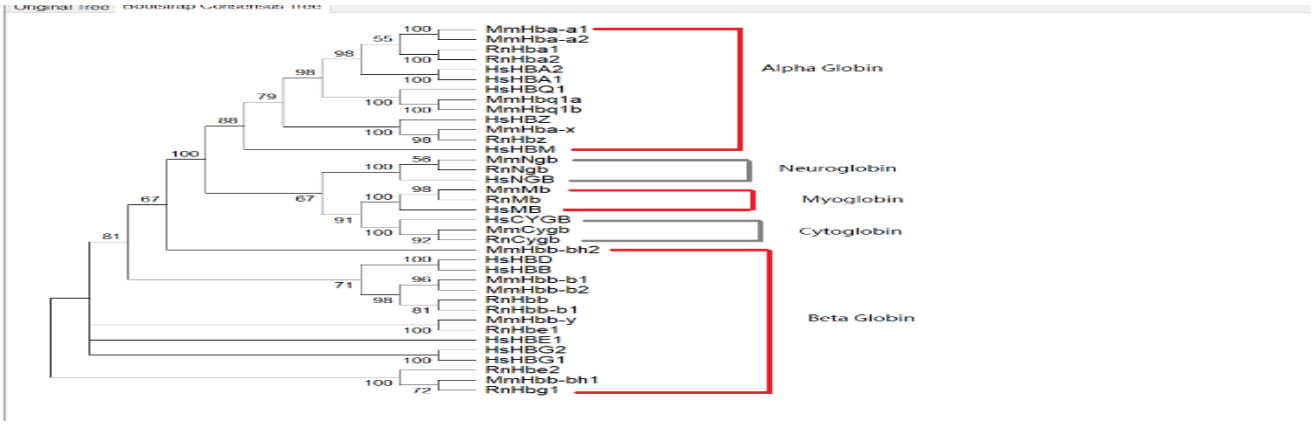
4）结果分析，数字代表分支可信度，越高代表可信度越高；枝长越短，表示序列相似性的可能性更高；拓扑结果中各物种或基因的未知代表亲缘关系的远近，聚在同一支最末端，亲缘关系高。

B 以人、小鼠、大鼠珠蛋白系统发生树构建，找出“先有物种、后有基因”和“先有基因、后有物种”的实例。

1)将人、小鼠、大鼠 37 个珠蛋白基因编码区 CDS 序列复制，打开 MEGA12 将序列先进行比对。



2）在 Data 中，选择系统发生分析 Phylogenetic Analysis 选项，选择比对蛋白质编码序列 Protein-Coding neucleiotide sequence data，主菜单中点击系统发生 Phylogeny 按钮，选择邻接法 Construct/Test Neighbor-Joining Tree，系统发生树稳定性测试 Test of Phylogeny 选项改为自举法 Bootstrap Method，将自举法重复次数改为 100，在替换类型 Substitution Type 中选择氨基酸 Amino Acid，其它参数采用默认值。

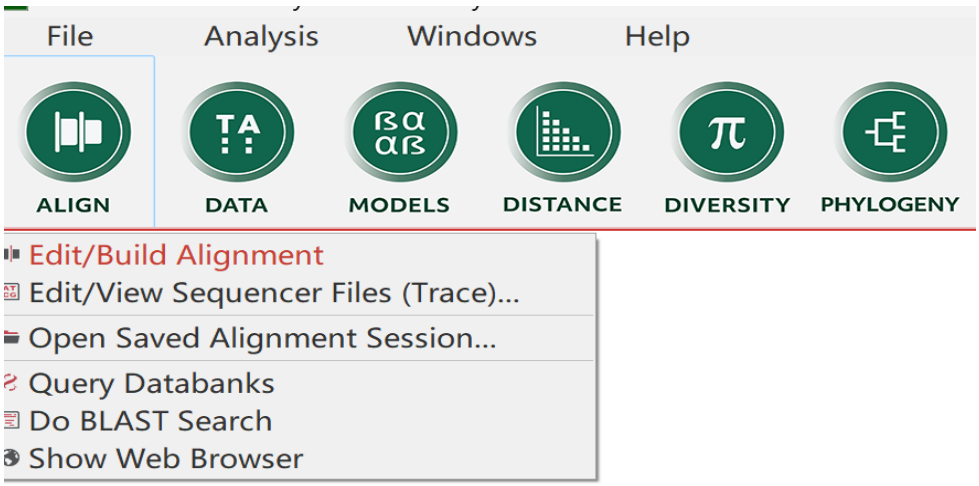


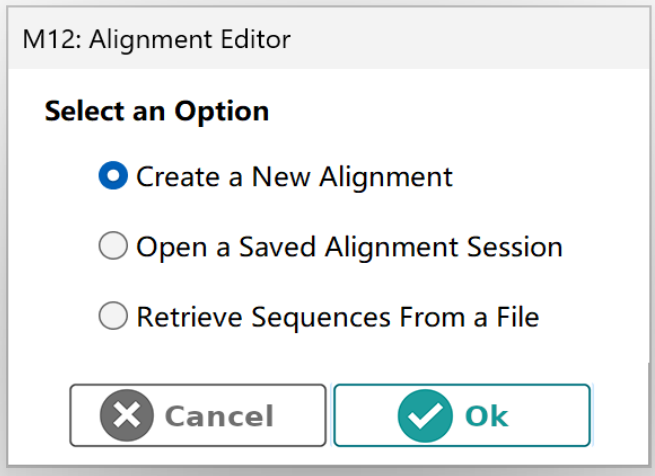
3）结果分析：

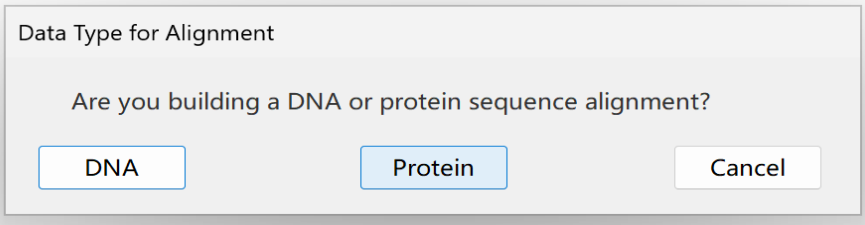
结果表明，37 个基因总体可以分为 5 个大类，即 α-珠蛋白、β-珠蛋 白、肌红蛋白、胞红蛋白和脑红蛋白。在各个大类中，包含人、小鼠、大鼠每个物种，说明 “先有基因、后有物种”。

C 7 个代表性脊椎动物 alpha 血红蛋白系统发生树构建。

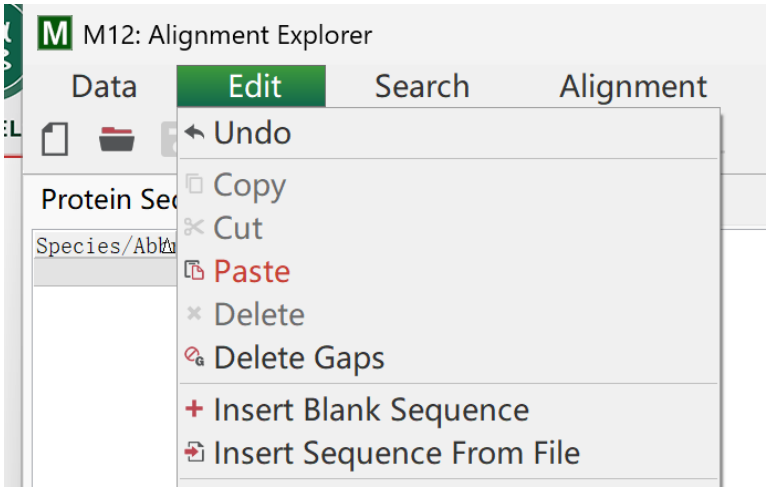
从 UniProt 数据库中找到上述 7 个代表性脊椎动物 alpha 血红蛋白，下载 FASTA 格式序列 7HBA.FAS。打开 MEGA12 系统发生树构建软件，点击主菜单中序列比对 Align 图标，在下拉菜单中选择创建 Edit/Build Alignment，在弹出会话窗口 Alignment Editor 中选择创建一个新的比对 Creat a New Aligment，在数据类型 Data Type 弹出窗口中选择蛋白质 Protein。



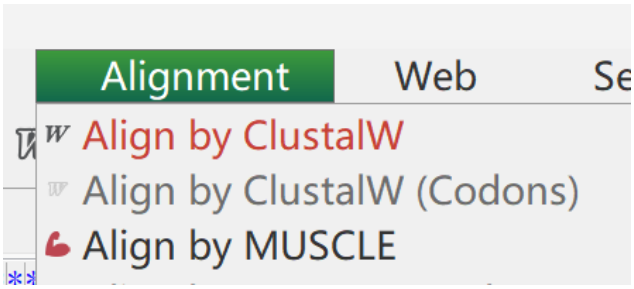


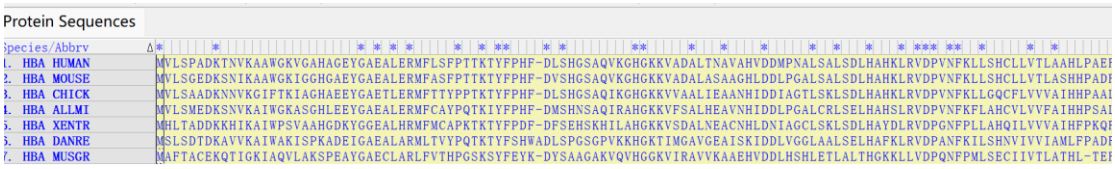


1)删除空序列 1. Sequence，点击主菜单中编辑按钮 Edit，在下拉菜单中选择粘贴 Paste，将 7 个 alpha 血红蛋白序列粘贴到编辑窗口中。

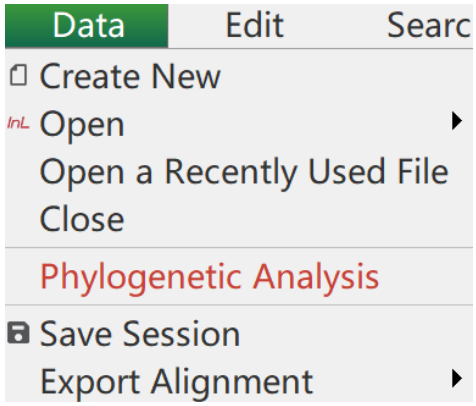


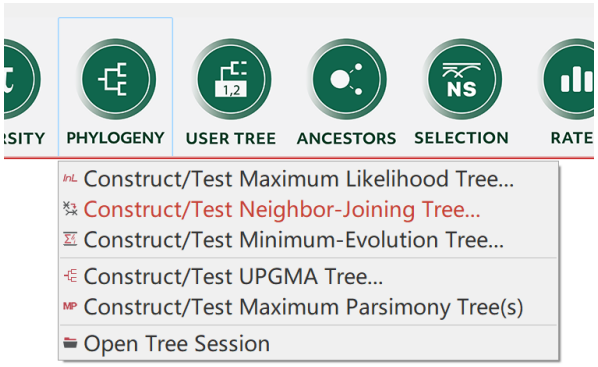
2）点击主菜单中序列比对 Alignment 按钮，在下拉菜单中选择 Align by ClustalW，在弹出会话窗口中点击 OK，选择比对所有 12 条序列， 在 ClustlW 参数选择弹出窗口中点击 OK，选择 MEGA12 给定的默认参数，查看比对结果

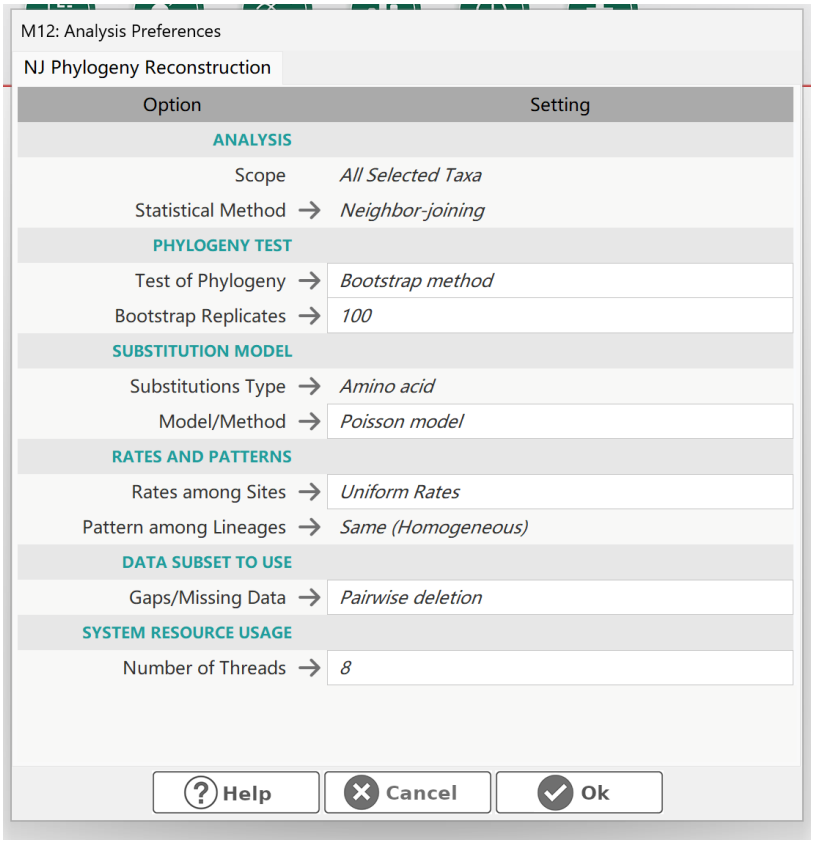




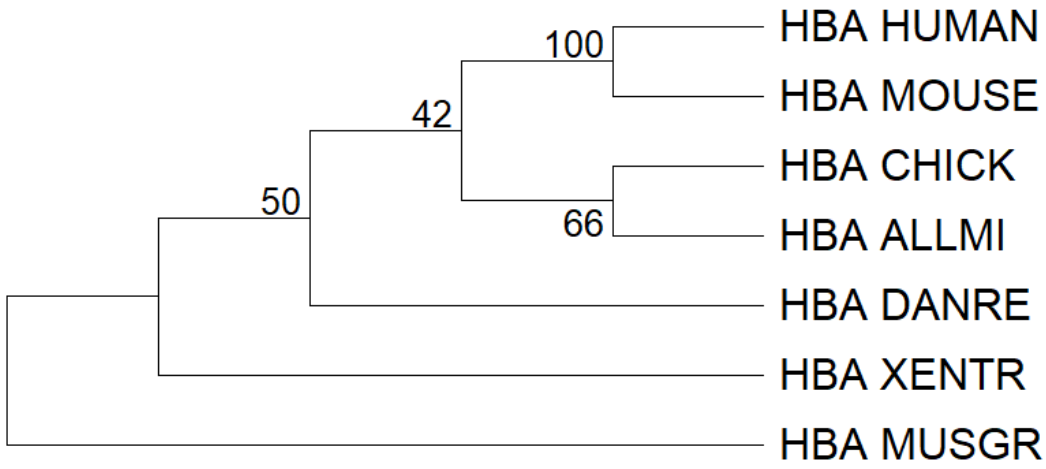
3）点击主菜单中数据 Data 按钮，在下拉菜单中选择系统发生分析 Phylogenetic Analysis。在主菜单中点击系统发生 Phylogeny 按钮，在下拉菜单中选择邻接法 Construct/Test Neighbor-Joining Tree，在弹出窗口中将系统发生树稳定性测试 Test of Phylogeny 选项改为自举法 Bootstrap Method，将自举法重复次数改为 100，其它参数采用默认值.







4）分析所构建的系统发生树是否能反映 7 个脊椎动物间的系统发生关系



该系统发生树中只有人与小鼠的哺乳类分支达到 100%的可靠支持，因此这两者的 HBA 亲缘关系反映正确；鸡与鳄的分支支持率仅 50–66，方向可能正确但统计上不可靠。而两栖类爪蟾的位置支持率低至 42，这个值非常低，表明该分支在统计上完全没有得到支持；斑马鱼与 MUSGR 作为鱼类应该共同最早分化，但支持率却很低，可能是因为该发育树分支代表的是 *HBA* 基因序列的差异程度，而不是物种间的演化亲疏，所以它不能用于推断物种间的系统发生关系。斑马鱼的 HBA 和鳄鱼的 HBA 可能因为相似的生活环境发生了相似的突变，导致序列反而更相似，而 HBA 在某些鱼类如 MUSGR 中可能经历过基因突变、丢失或功能转换等，导致序列差异反而大。

D 7 个代表性植物 18S rRNA 系统发生树构建

1）从 NCBI RefSeq 参考序列数据库中找到 6 个代表性植物 18S rRNA 序列，下载 FASTA 格式序列 18S rRNA；

rRNA 即 ribosomal RNA，核糖体 RNA，18S 是指沉降系数（Svedberg unit，单位：S）为 18S 的 rRNA 片段，数值越大，代表分子或颗粒沉降越快，整体分子量和体积也越大。

18S rRNA 是真核生物核糖体小亚基的核心 RNA，是真核生物物种鉴定和系统发育分析的常用分子标记；16S rRNA 是原核生物（细菌或古菌）核糖体小亚基的对应 RNA，广泛用于细菌、古菌的物种鉴定和微生物群落分析；而 28S rRNA 则是真核生物核糖体大亚基的主要 RNA，常与 18S rRNA 搭配使用，共同用于更高阶的进化分析。

S 是沉降系数，不是质量单位，不能直接相加，以下是真核和原核生物亚基的核糖体结构组成与沉降系数：

结构	组成	沉降系数
原核小亚基	16S rRNA + 核糖体蛋白	30S
原核大亚基	23S rRNA + 5S rRNA + 核糖体蛋白	50S
完整原核核糖体	30S + 50S	70S

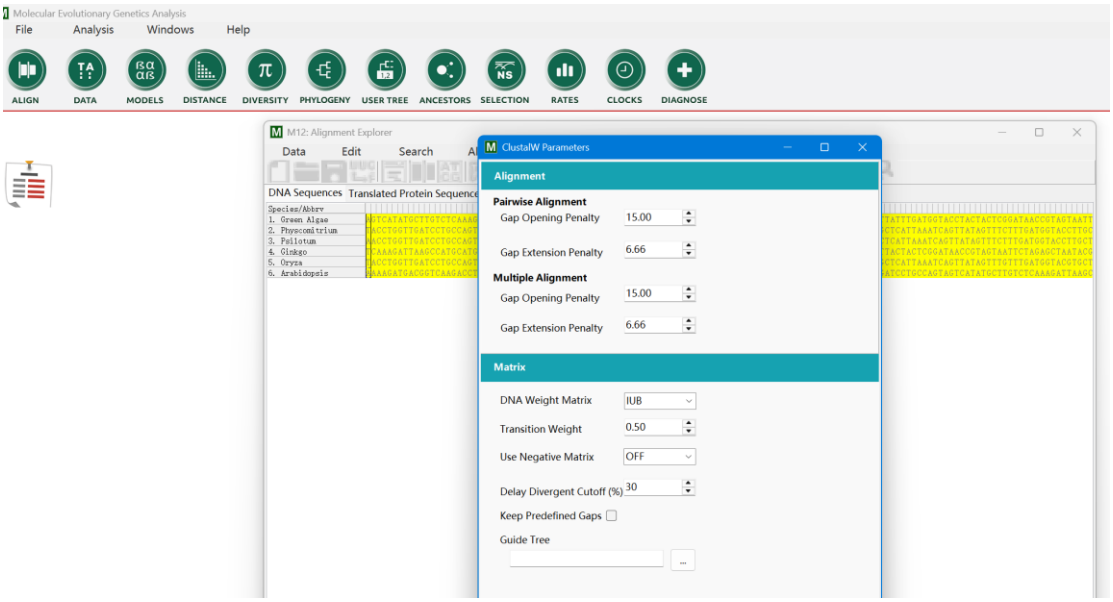
结构	组成	沉降系数
真核小亚基	18S rRNA + 核糖体蛋白	40S
真核大亚基	28S rRNA + 5.8S rRNA + 5S rRNA + 核糖体蛋白	60S
完整真核核糖体	40S + 60S	80S

```
>Green_Algae (LuZao) | Chlamydomonas reinhardtii | strain SAG 53.72
AGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTATAAACTGCTTTATACT
GTGAAACTGCGAATGGCTCATTAAATCAGTTATAGTTTATTTGATGGTACCTACTACTCG
GATAACCGTAGTAATTCTAGAGCTAATACGTGCGTAAATCCCGACTTCTGGAAGGGACGT
ATTTATTAGATAAAAGGCCAGCCGGGCTTTGCCCGACCTGCGGTGAATCATGATAACTTC
ACGAATCGCATGGCCTTGCGCCGGCGATGTTTCATTCAAATTTCTGCCCTATCAACTTTC
GATGGTAGGATAGAGGCCTACCATGGTGGTAACGGGTGACGGAGGATTAGGGTTCGATTC
CGGAGAGGGAGCCTGAGAGATGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTA
CCCAATCCCAACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCATTTCATGTCTG
GTAATTGGAATGAGTACAATCTAAATCCCTTAACGAGGATCCATTGGAGGGCAAGTCTGG
TGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTTAAGTTGTTGCAGTTAAAA
AGCTCGTAGTTGGATTTCGGGTGGGTCTTAGCGGTCCGCCTCTGGTGTGTACTGCTAGGG
CCTATCTTTCTGCCGGGGACGGGCTCCTGGGTTTAATCGCCTGGGACTCGGAGTCGGCGA
GGTTACTTTGAGTAAATTAGAGTGTTCAAAGCAAGCCTACGCTCTGAATACATTAGCATG
GAATAACACGATAGGACTCTGGCCTATCTTGTTGGTCTGTAGGACCGGAGTAATGATTAA
GAGGGACAGTCGGGGGCATTCGTATTTCATTGTCAGAGGTGAAATTCTTGGATTTATGAA
AGACGAACTTCTGCGAAAGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTG
GGGGCTCGAAGACGATTAGATACCGTCGTAGTCTCAACCATAAACGATGCCGACTAGGGA
TTGGCAGATGTTTCATTGATGACTCTGCCAGCACCTTATGAGAAATCAAAGTTTTTGGGT
TCCGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCA
GGCGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCAGACACG
GGAAGGATTGACAGATTGAGAGCTCTTTCTTGATTCTGTGGGTGGTGGTGCATGGCCGTT
CTTAGTTGGTGGGTTGCCTTGTCAGGTTGATTCCGGTAACGAACGAGACCTCAGCCTGCT
AAATAGTCACGACTGCTTTTTGCAGTTGGCCGACTTCTTAGAGGGACTATTGTCGTGTAG
GCAATGGAAGTATGAGGCAATAACAGGTCTGTGATGCCCTTAGATGTTCTGGGCCGCACG
CGCGCTACACTGACGCATTCAACGAGCCTATCCTTGGCCGAGAGGCCCGGGTAATCTTTG
AAACTGCGTCGTGATGGGGATAGATTATTGCAATTATTAGTCTTCAACGAGGAATGCCTA
GTAAGCGCGAGTCATCAGCTCGCGTTGATTACGTCCCTGCCCTTTGTACACACCGCCCGT
CGCTCCTACCGATTGGGTGTGCTGGTGAAGTGTTCGGATTGGCTTCAGGTGATGGCAACA
TCGCCTGTTGCTGAGAAGTTCATTAAACCCTCCCACCTAGAGGAAGGAGAAGTCGTAACA
AGGTTTC
>Physcomitrium patens | Moss (Taixian)
```

2）打开 MEGA12 系统发生树构建软件，点击主菜单中序列比对 Align 图标，在下拉菜单中选择创建 Edit/Build Alignment，在弹出会话窗口 Alignment Editor 中选择创建一个新的比对 Creat a New Aligment，在数据类型 Data Type 弹出窗口中选择 DNA；

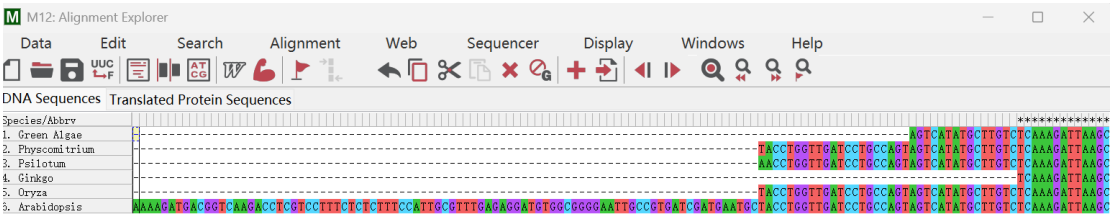
3）删除空序列 1. Sequence，点击主菜单中编辑按钮 Edit，在下拉菜单中选择粘贴 Paste，或者直接 Ctrl+V 粘贴，将 6 个 18S rRNA 序列粘贴到编辑窗口中；

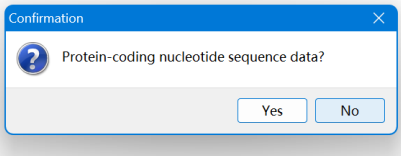
4）点击主菜单中序列比对 Alignment 按钮，在下拉菜单中选择 Align by ClustalW，在弹出会话窗口中点击 OK，选择比对所有 7 条序列，在 ClustlW 参数选择弹出窗口中点击 OK，选择 MEGA12 给定的默认参数，查看比对结果；



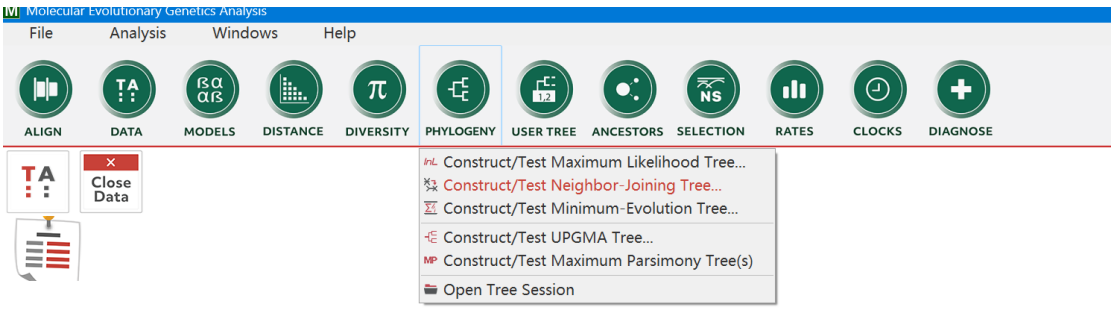
Matrix，矩阵

5）点击主菜单中数据 Data 按钮，在下拉菜单中选择系统发生分析 Phylogenetic Analysis；注意在是否是蛋白质编码核酸序列处选择 NO；





6）在主菜单中点击系统发生 Phylogeny 按钮，在下拉菜单中选择邻接法 Construct/Test Neighbor-Joining Tree；



Construct/Test Maximum Likelihood Tree 构建/检验最大似然树，缩写：ML 法，基于概率模型，计算每个进化树出现的概率，选择可能性最大的拓扑结构。特点：准确性最高、最严谨，适合大多数现

代系统发育分析，但计算速度慢，对计算机性能要求高。适用：序列数据量适中、需要高精度分析的课题。类似这样段落描述

Construct/Test Neighbor-Joining Tree 构建/检验邻接树，缩写：NJ 法，基于序列间的遗传距离矩阵，通过不断合并距离最近的节点构建进化树。特点：运算速度极快、对大数据友好，是入门级系统发育分析的常用方法，但准确性不如 ML 法。适用：快速初步分析、数据量大或作为课题前期筛选的建树方法。

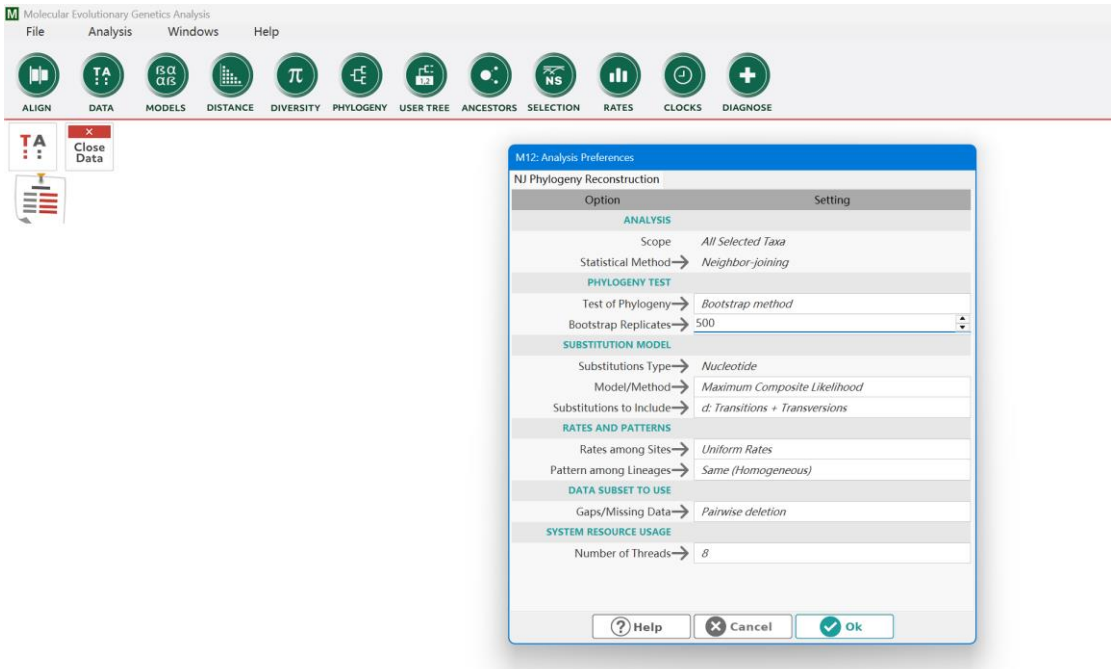
Construct/Test Minimum-Evolution Tree 构建/检验最小进化树，缩写：ME 法，基于距离矩阵，目标是构建总进化距离最小的拓扑结构，是邻接法的改进版本。特点：比 NJ 法结果更稳定，计算速度也较快。适用：距离类分析中，追求比 NJ 法更可靠结果的场景。

Construct/Test UPGMA Tree 构建/检验非加权组平均法或算术平均非加权成组配对法(Unweighted Pair Group Method with Arithmetic mean)树，缩写：UPGMA 法，基于序列距离矩阵，假设进化速率恒定，通过组间平均距离聚类构建有根树。特点：方法简单直观，但“进化速率恒定”的假设过于理想化，现代研究中使用较少。适用：教学演示或已知进化速率近似恒定的特殊场景。

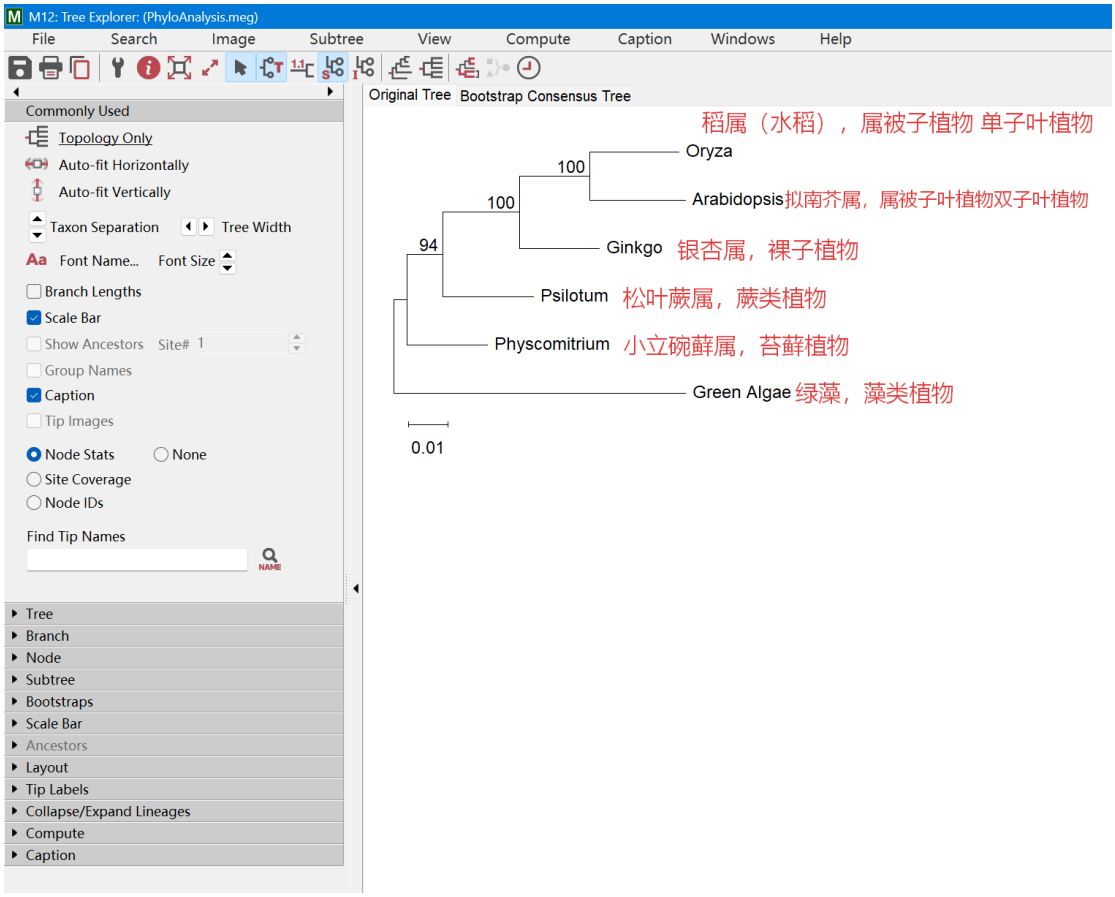
Construct/Test Maximum Parsimony Tree(s)构建/检验最大简约树，缩写：MP 法，以碱基替换次数最少为原则，选择进化步骤最少的拓扑结构，遵循“奥卡姆剃刀”原则。特点：不依赖进化模型，对序列差异小、位点信息多的数据友好，但对趋同进化敏感，大数据计算缓慢。适用：近缘物种、序列保守性高的系统发育分析。

Open Tree Session 打开树分析会话，功能：打开之前保存的 MEGA 树分析工程文件。特点：可直接恢复之前的建树参数和结果，方便后续调整、美化或重新运行分析。适用：需要对已构建的进化树进行二次编辑、调整拓扑结构的场景。

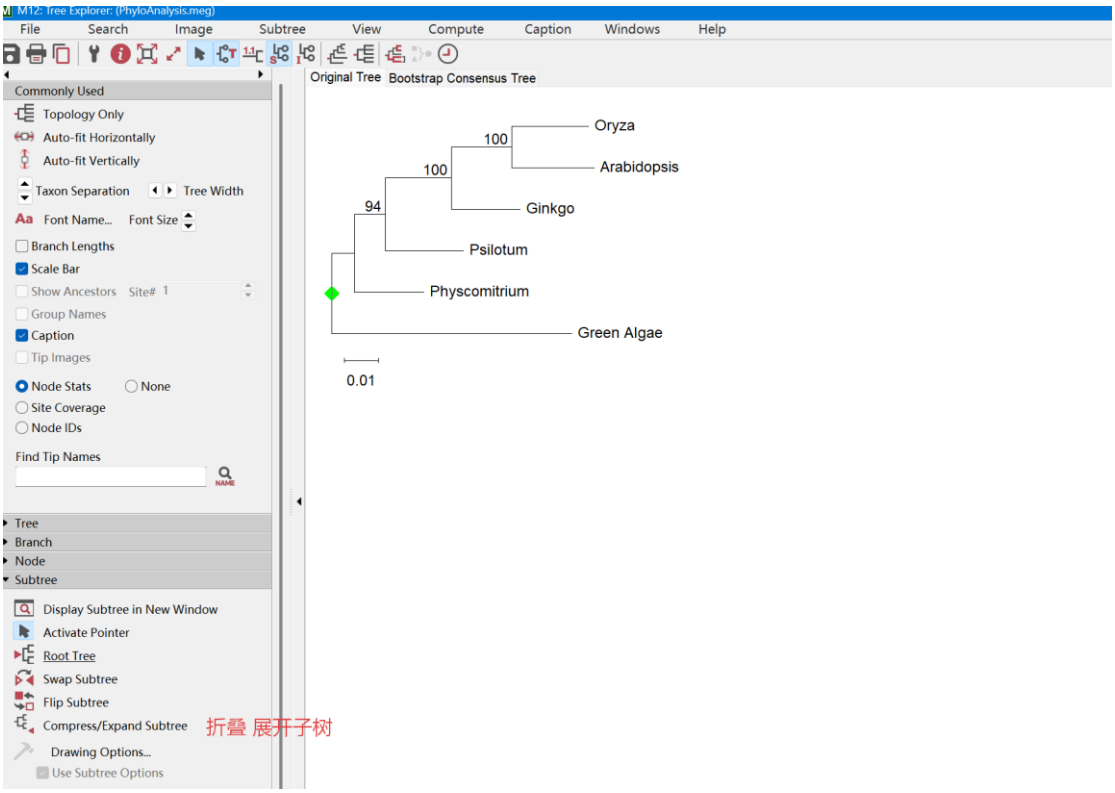
7）在弹出窗口中将系统发生树稳定性测试 Test of Phylogeny 选项改为自举法 Bootstrap Method，将自举法重复次数设置为 500，其它参数采用默认值



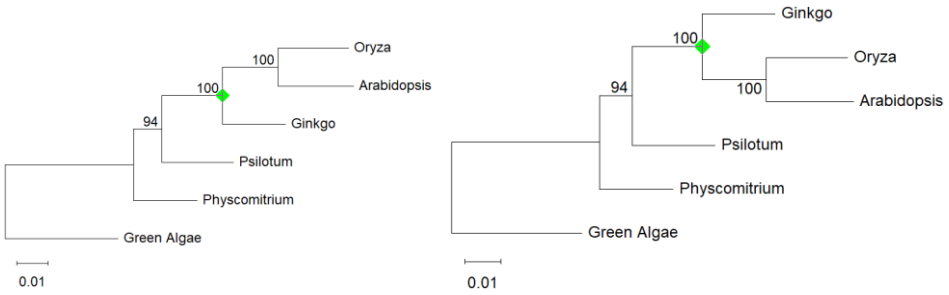
7）查看所构建的系统发生树，利用 Subtree 调整树的显示方式；



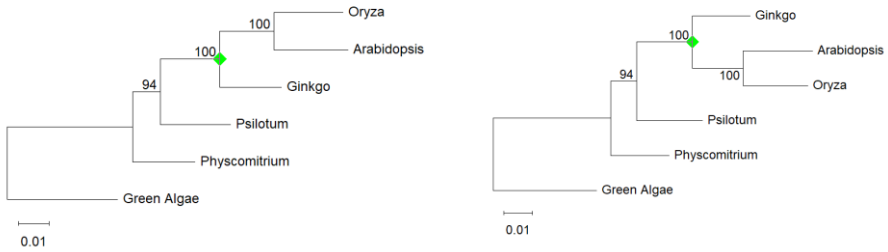
利用 Subtree 调整树的显示方式；可通过 Root Tree 确定 Green Algae 为根；



点击 Swap Subtree 可看到

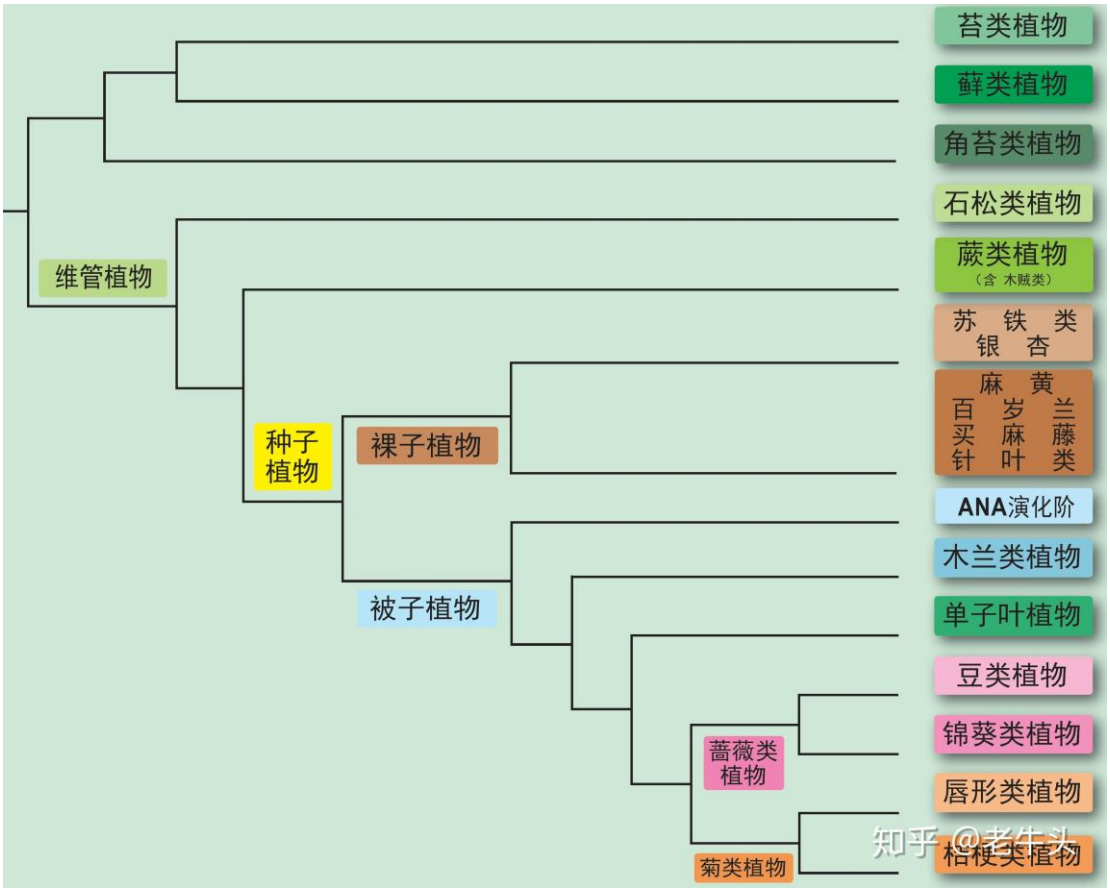


点击 Flip Subtree 可看到



Swap Subtree 是交换位置，Flip Subtree 是翻转

搜索植物分类网站，参考绿色植物分类谱系分析所构建的系统发生树能否反映 6 个代表性植物的系统发生关系



这株系统发生树能够清晰、准确地反映 6 个代表性植物的系统发生关系，其拓扑结构与绿色植物的演化谱系高度吻合。树中，绿藻作为所有陆生植物的祖先类群，被置于最基部；随后依次分化出苔藓植物（小立碗藓）、早期维管植物（松叶蕨），再到种子植物中的裸子植物（银杏），最终被子植物分

支内，单子叶植物（水稻）与双子叶植物（拟南芥）聚为姐妹群。关键节点的 Bootstrap 支持度高达 94%–100%，证明了各分支关系的稳定性与可靠性，整体树型完整呈现了从低等到高等、从非维管到维管、从孢子植物到种子植物的演化路径，是该类群系统发生关系的可靠体现。

G2A 边汉青个人总结

通过近段时间的学习，主要有了以下收获：

1.掌握了系统发育树的一些概念并回忆了之前所学的一些生物学知识。

系统发生树，是一种对历史进化、演化的假设，可以随着研究的深入而无限逼近真实的演化历史；单系类群（monophyly）：是拥有一个共同祖先、且这一个共同祖先的后裔全部包括在内的所有分类单元组成的类群；并系类群（paraphyly）：具有一个共同祖先但不包括所有后裔的类群，即并系类群包含了共同祖先的部分（而非全部）后代；多系类群（polyphyly）：一个分类群当中的成员，在系统树上分别位于相隔着其它分枝的分支上，即该类群不包括所有成员的最近共同祖先。系统发育树包括基因树（gene tree）和物种树（species tree），基因树（gene tree）：是根据 DNA 或蛋白质序列构建的系统树；物种树（species tree）是表达生物类群演化路径的系统树。

2.初步掌握了使用 MEGA 构建系统发育树的方法。

在选择以蛋白质序列进行建树时，alignment 选择 Align by ClustalW；在选择以 CDS 序列进行建树时，alignment 选择 Align by ClustalW (Codons)；在选择以 DNA 序列进行建树时选择 Align by ClustalW。这是由于 CDS 为编码区，考虑到密码子的简并性，为确保生物准确性，故而采取 Align by ClustalW (Codons)模式。

建树的方法一般采用邻接法 Neighbor-Joining，对近缘序列来说较优，但也有文章认为有综述认为贝叶斯的方法最好，其次是 ML（Maximum Likelihood），然后是 MP (Maximum Parsimony)。一般来说，用两种不同的方法构建进化树，如果所得到的进化树类似，则结果较为可靠。系统树的可靠性评估一般采用自举法（Boot strapping）。

蛋白质序列一般作为关于亲缘性较远物种的系统发育树的构建，核算或者 CDS 序列一般作为亲缘性较近物种的系统发育树的比较。

3.系统发育树的意义

1)确定物种起源，判断亲缘的远近和进化的方向。

2)适应性演化的研究，判断某一群体是否为独立分支。

3)基因功能的研究，判断已知物种的某一基因，在其他物种中的分化和发展以及功能的演化。

4)疾病的传播路径研究，判断疾病的传播路径，评估转移和传播的方式。

5)肿瘤内基因突变的演化，判断肿瘤内部。

6)学习系统发育树的构建，初步掌握了系统发育树的构建步骤和相关软件与方法。

G2B 刘奇个人总结

使用 MEGA 软件完成系统发育树的完整构建流程主要如下：

序列准备与比对

下载 FASTA 格式序列，序列命名只能用字母、数字和下划线，不能有空格或括号，否则 MEGA 报错。是 DNA 序列就选“DNA”。是蛋白质序列就选“protein”，通常亲缘关系近（属内或科内）用 DNA；亲缘关系远（纲级以上）用蛋白质，或者两者都做，对比结果是否一致。将序列导入 MEGA 后，选择“Align by ClustalW”进行比对，参数默认即可。比对完成后，手动检查两端是否对齐，两端多余区域应裁剪（Edit→Delete）。比对好的文件另存为.meg 格式。

建树方法

MEGA 中最常用的两种方法：一种是 NJ 法（邻接法），基于遗传距离矩阵的聚类方法，优点是速度快，适合大数据集或初步探索远缘序列，树的构建相对准确，计算速度快，只得到一棵树，可以分析较多序列，运行速度优于最大简约法。 缺点：序列上所有位点等同对待，且所分析的序列进化距离不能太大；另一种是 ML 法（最大似然法），它会尝试所有可能的树形，计算每棵树产生当前 DNA 数据的概率，最后选出概率最高的那棵。优点是准确率最高，抗伪像能力强；缺点是计算极慢，且必须先选对核苷酸替代模型。

建树步骤与参数设置

无论用哪种方法，操作路径都是：Phylogeny → Construct NJ Tree / Construct ML Tree。关键参数：bootstrap 次数越高越准确，但计算时间也越长。ML 法建树前，先运行“Find Best DNA Model”，记下推荐模型后在设置中手动选择。

结果解读与导出

树生成后，bootstrap 值显示在分支节点上。判断标准：≥70%的节点才认为可信，低于 70 的应在论文中标注或折叠。导出树时选择“Export Current Tree”保存为.nwk 格式，便于后续用 TBtools 等其他软件美化。建议构建发育树时添加外群，建树前在序列中加入已知最远亲缘的物种，以此来确定树根位置，来判断谁最早分化。如果大部分节点<50，说明序列差异太大或信息位点不足，解决方案包括：换更保守的基因、增加物种数、或改用 ML 法。

G2C 高倩个人总结

过去两周的学习主要围绕系统发育树构建展开，从基础理论到 MEGA 软件实操，逐步掌握从序列比对、数据分析到进化树构建。

构建系统进化树可分为三个步骤，分别是序列比对 ALIGN，然后是数据分析 Phylogenetic Analysis，最后是构建树 Phylogeny。在序列比对与数据准备环节，我延续了前期的多序列比对基础，针对不同课题序列，用 MEGA 完成了序列导入、格式转换和比对校正，为后续建树打好基础；同时结合课堂学习，理解了不同序列类型的适用场景：亲缘关系近的物种优先选择核酸序列，而关系较远的类群则更适合用蛋白序列构建系统发育树，这为后续课题中双斑萤叶甲、桃柱螟相关序列的选择提供了思路。之后重点学习了 MEGA 软件的进化树构建流程，学习了邻接法（NJ）、最大似然法（ML）、最大简约法（MP）、UPGMA 等多种建树方法的原理与操作，理解了不同方法的优缺点与适用场景。比如 NJ 法运算快适合快速分析，ML 法准确性高是现代研究的主流方法。

课堂上老师强调建树的核心原则：数据可靠、方法可选、参数可调、结果可信，通过调整 Bootstrap 重复次数、替换模型等参数，提升了对树结构可信度的判断能力。在树的可视化与解读部分，我学习了 MEGA 中 Swap Subtree（交换子树）、Flip Subtree（翻转子树）、Compress/Expand Subtree（折叠/展开子树）等常用操作，学会了调整树的拓扑结构和展示方式；也深入理解了 Bootstrap 值的含义，它代表分支支持度，数值越高说明分支关系越稳定。课堂上关于“先有基因还是先有物种”的讨论，理解了物种树和基因树的区别，让我明白了基因复制、基因丢失事件对系统发育树解读的影响，也学会了结合人、小鼠、大鼠等模式生物的系统发育实例，分析基因进化与物种进化的关系。

目前仍存在一些不足：对不同建树方法的参数设置逻辑理解还不够透彻，部分复杂树的解读和应用场景分析不够深入，后续会通过完成课堂练习、强化 MEGA 实操，把本学期所学的序列比对、BLAST、系统发育树构建等内容，和自己的课题（双斑萤叶甲致死基因挖掘、桃柱螟中肠微生物分析）结合起来。