“实用生物信息技术”课程小组讨论总结报告

组：G6　　次：4　　组长：冯睿桥　　执笔：　冯睿桥、凌子涵

1. 时间

2026 年 4 月 24 日、4 月 26 日、5 月 1 日

2. 方式

线上线下交流与讨论

3. 主题

系统发育树课后练习及近期学习内容交流与互助

4. 内容

1、 系统发育树构建基础概念

2、 基础问题讨论

3、 人珠蛋白基因家族系统发生树实例

4、 人、小鼠和大鼠三个物种珠蛋白家族系统发生树实例

5、 血红蛋白 alpha 亚基系统发生树实例

1. 系统发育树构建基础概念

(1) 分子演化和系统发生树基本概念：关注的是 DNA、RNA 或蛋白质序列随时间发生的变化，例如突变、插入缺失、基因重复等。这些变化在不同物种之间积累，形成差异。系统发生树则是把这些差异转化为进化关系的图形表达，本质上表现了哪些物种更接近，它们的共同祖先在哪里。

(2) 物种分化和分子演化：物种分化（speciation）是宏观层面的事件，而分子演化是微观层面的过程。

(3) 分支图和系统树：分支图（cladogram）只表示拓扑结构（谁和谁更近），不体现进化距离。系统树（phylogram）的分支长度有意义，通常代表进化变化量或时间。

(4) 物种树和基因树：物种树反映物种真实的进化历史。基因树（gene tree）是某个基因的进化历史。

(5) 有根数和无根树：有根树（rooted tree）有一个明确的共同祖先，能体现进化方向（时间方向）。无根树（unrooted tree）只表示相对关系，没有时间起点。

(6) 二叉树与多岐树：二叉树（binary tree）每个内部节点分成两个分支。多岐树（polytomy）一个节点分出多个分支。

(7) 外部节点和内部节点：外部节点（leaf/tip）代表现存物种或序列，内部节点（internal node）代表假想的共同祖先。

(8) 内部节点和根节点：内部节点是任意祖先节点。根节点（root）是最早的共同祖先，是整棵树的起点。根节点是特殊的内部节点，它定义了进化方向。

(9) 系统发生树稳定性检验：构建的系统树并不一定可靠，因此需要统计方法评估其稳定性。

2. 基础问题讨论

(1) 构建系统发生树的基本步骤

首先需要明确分析对象，并从数据库中收集具有同源关系的 DNA、RNA 或蛋白质序列，因为只有同源序列之间的差异才具有进化意义。随后需要进行多序列比对，通过将各序列中来源于共同祖先的位置对应排列，为后续分析建立基础。比对完成后，通常还需要对结果进行人工校正，并去除难以准确比对的区域，以减少噪声对树构建的影响。在获得可靠的比对结果后，需要根据数据特点选择适合的进化模型与建树方法。之后利用邻接法、最大简约法、最大似然法或贝叶斯法等方法正式构建系统树。树构建完成后，还需进一步利用 Bootstrap 等统计方法对各分支的稳定性进行检验，并结合外群确定根节点位置。最终再依据分支关系、分支长度以及支持率，对物种或基因之间的进化关系进行生物学解释。

(2) 构建系统发生树时选择核苷酸序列或氨基酸序列的原则

选择核苷酸序列还是氨基酸序列，主要取决于研究对象的进化距离与研究目的。对于亲缘关系较近的物种，通常优先选择核苷酸序列。因为核酸包含的信息量更大，同义突变也能提供进化信息。例如不同品种、近缘种之间的分析，核苷酸序

列能够提供更高分辨率。但对于远缘物种，核苷酸序列容易发生“饱和”（saturation），即同一位点多次突变，真实历史被掩盖。此时更适合使用氨基酸序列。由于蛋白质只有 20 种氨基酸，且功能约束较强，变化速度相对较慢，更适合分析远距离进化关系。

(3) 利用自举法（Bootstrap）检验系统发生树稳定性的原理

ootstrap 是一种用于评价系统发生树可靠性的统计学方法，其核心思想是通过重复抽样来检验树结构是否稳定。在实际分析中，研究者首先获得一个多序列比对矩阵，然后从所有比对位点中进行“有放回随机抽样”，重新生成一个新的数据集。由于是有放回抽样，因此某些位点可能被重复选中，而另一些位点则可能不被选中。利用重新生成的数据集再次构建系统发生树，并不断重复这一过程，通常重复数百次甚至上千次。最后统计某一分支在所有重复建树中出现的频率，该频率即为 Bootstrap 支持率。例如某一分支在 1000 次重复中出现 950 次，则其 Bootstrap 值为 95%。Bootstrap 值越高，说明该分支在数据扰动下仍能稳定出现，其可信度越高。一般认为支持率高于 70%的分支具有较好的可靠性，而低于 50%的分支则通常被认为缺乏统计支持。

(4) 确定无根树根节点的方法

无根树只能表示各序列之间的相对亲缘关系，而无法体现进化方向，因此需要通过定根来确定共同祖先的位置。最常用的方法是外群法，即在研究对象之外选择一个已知亲缘关系较远、但仍具有同源性的序列作为外群。由于外群通常较早从共同祖先中分离，因此外群连接的位置即可以视为整棵树的根节点。除外群法之外，还可以采用分子钟法和中点定根法。分子钟法假设各分支进化速率大致恒定，通过时间尺度推断根节点位置；中点定根法则将根放置于树中最长路径的中点，其本质也是基于近似恒定进化速率的假设。不过相比之下，外群法通常具有更明确的生物学依据，因此应用最为广泛。

(5) 如何通过所构建的系统发生树判断“先有物种”还是“先有基因”

如果某个基因在物种分化之前已经发生复制，那么不同物种中往往会保留该基因的多个旁系同源拷贝。在这种情况下，系统树中的聚类关系通常表现为不同物种中的同类基因优先聚在一起，而不是同一物种中的不同基因聚在一起。

(6) 不同建树方法的基本原理和特点

系统发生树的构建方法主要包括距离法、最大简约法、最大似然法和贝叶斯法等。距离法首先根据序列差异计算遗传距离矩阵，再依据距离逐步聚类形成系统树。代表方法包括 UPGMA 和 Neighbor-Joining（NJ）法。其中 UPGMA 假设所有分支进化速率一致，因此适用于满足分子钟假设的数据；NJ 法则不要求严格分子钟，因此应用更加广泛。距离法计算速度快，适合大规模数据分析，但由于原始序列信息被压缩为距离值，可能导致部分信息丢失。

最大简约法（Maximum Parsimony）则基于“最少进化原则”，认为最合理的系统树应当对应最少的突变次数。该方法不依赖复杂进化模型，原理较为直观，但

容易受到长枝吸引现象的影响，即快速进化的分支可能被错误聚类，因此在复杂数据中稳定性较差。

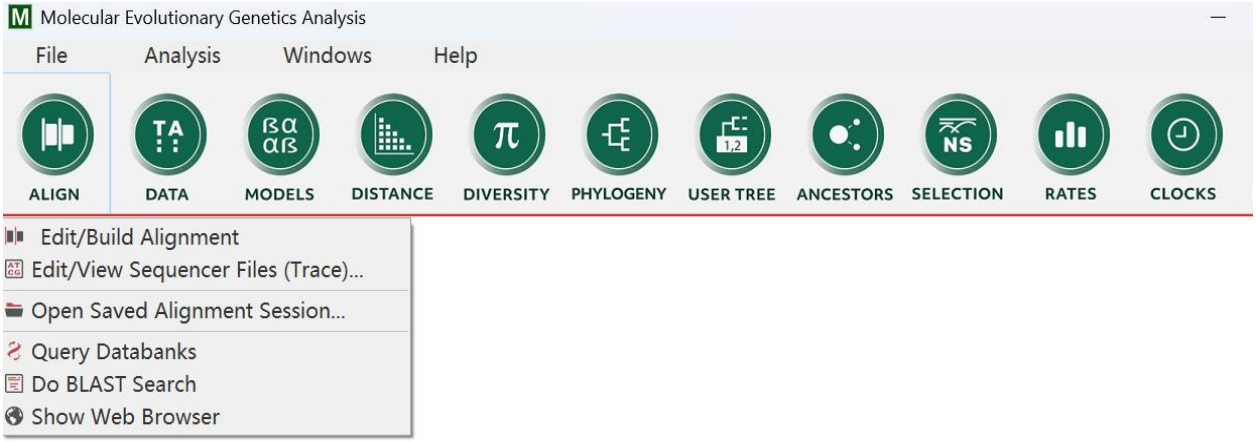
最大似然法（Maximum Likelihood）是在给定进化模型条件下，寻找最有可能产生当前观测数据的系统树。它能够综合考虑不同位点、不同替换类型及不同进化速率，因此统计学基础更加严格，通常具有较高准确性。但由于需要大量计算，其计算复杂度也明显高于距离法和简约法。

贝叶斯法则是在最大似然基础上进一步结合贝叶斯统计思想，通过计算后验概率来评估系统树，并通常采用马尔可夫链蒙特卡洛（MCMC）方法搜索最优树空间。该方法能够直接给出各分支的后验概率，统计解释较为清晰，在现代系统发育研究中应用越来越广泛，但对计算资源要求也较高。

3. 人珠蛋白基因家族系统发生树实例

以人珠蛋白基因家族 12 个成员蛋白质序列，用 MEGA 邻接法构建系统发生树；选择不同氨基酸替换模型（Substitution Model），比较所构建的系统发生树的拓扑结构和稳定性值（Bootstrap value），说明不同替换模型对结果的影响。

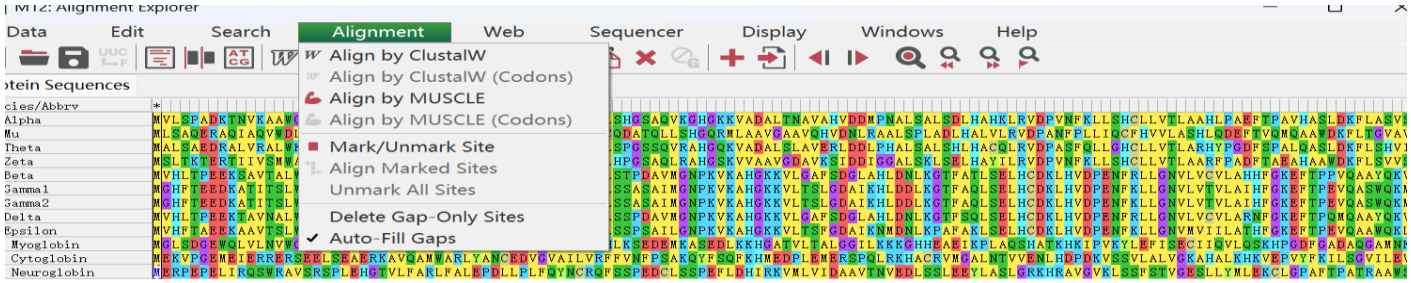
（1）在 ABC 中文主页课程内容 5 的课堂实例中复制人珠蛋白基因家族 12 个成员蛋白质序列。打开 MEGA12 系统发生树构建软件，点击主菜单中序列比对 Align 图标，在下拉菜单中选择创建 Edit/Build Alignment, 在弹出会话窗口 Alignment Editor 中选择创建一个新的比对 Creat a New Aligment, 在数据类型 Data Type 弹出窗口中选择蛋白质 Protein

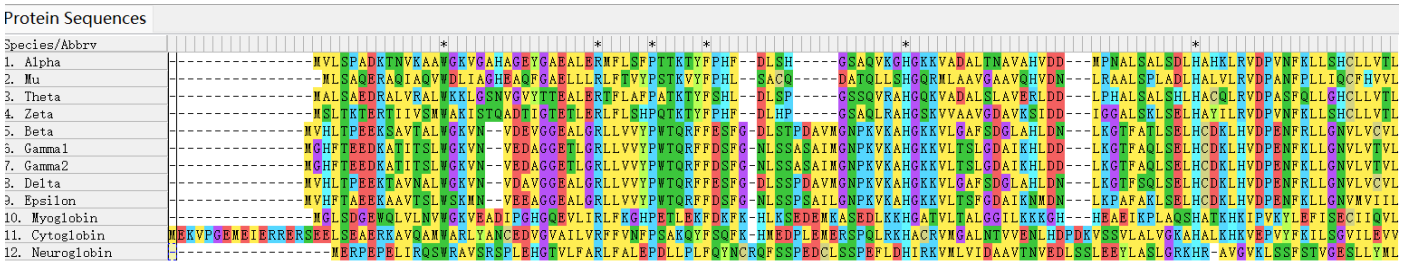


RECENT PUBLICATIONS

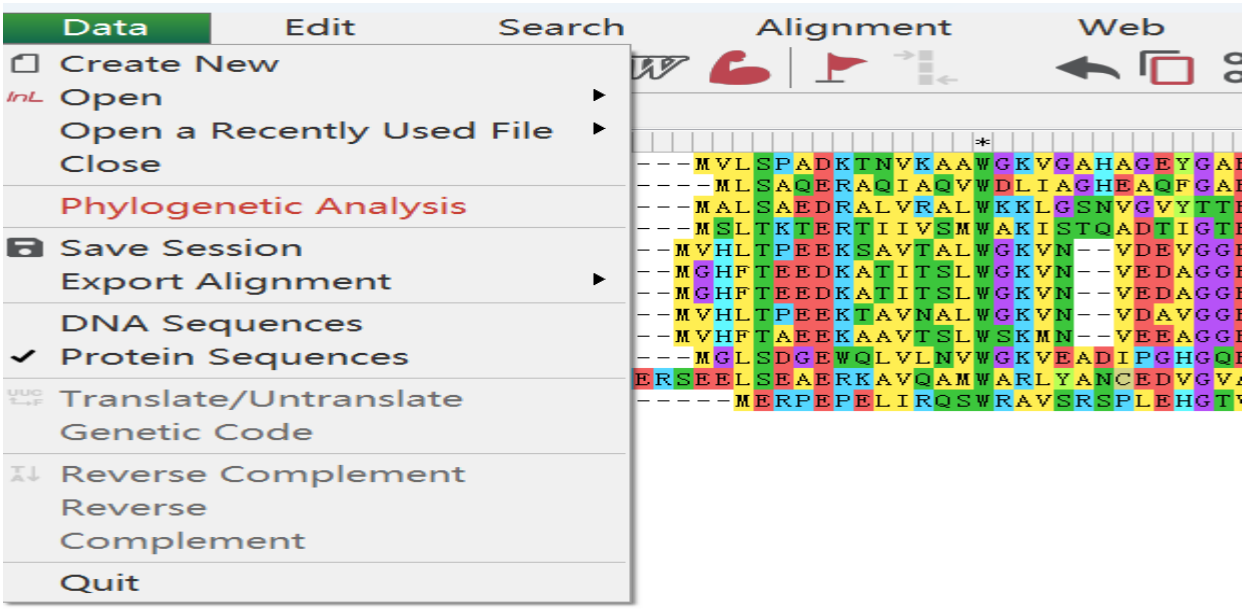
（2）删除空序列 1. Sequence, 点击主菜单中编辑按钮 Edit，在下拉菜单中选择粘贴 Paste, 将 12 个珠蛋白序列粘贴到编辑窗口中，点击主菜单中序列比对 Alignment 按钮，在下

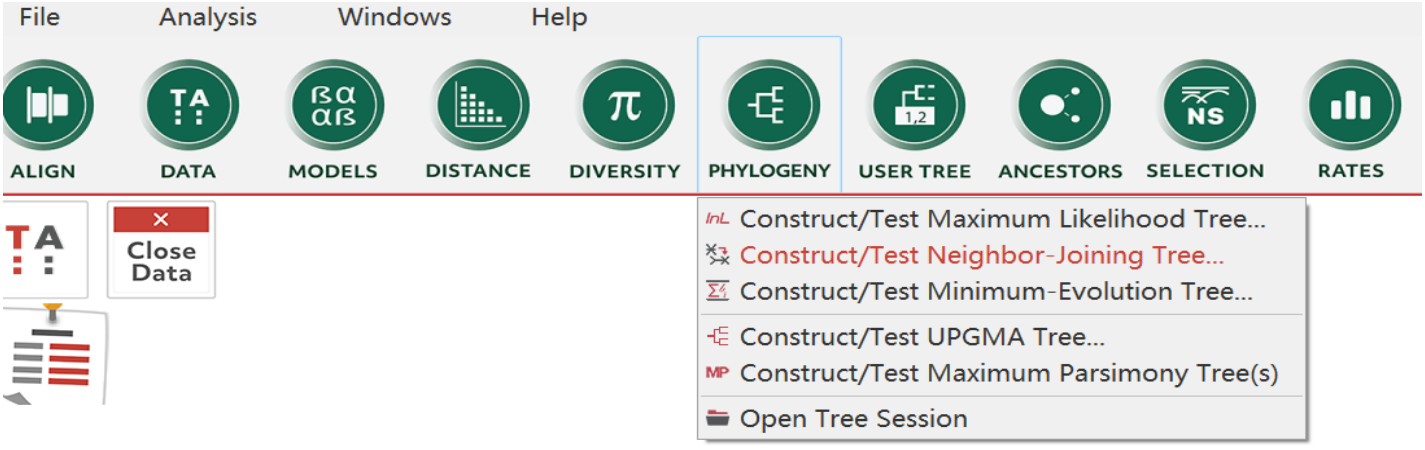
拉菜单中选择 Align by ClustalW, 在弹出会话窗口中点击 OK, 选择比对所有 12 条序列， 在 ClustlW 参数选择弹出窗口中点击 OK，选择 MEGA12 给定的默认参数，查看比对结果

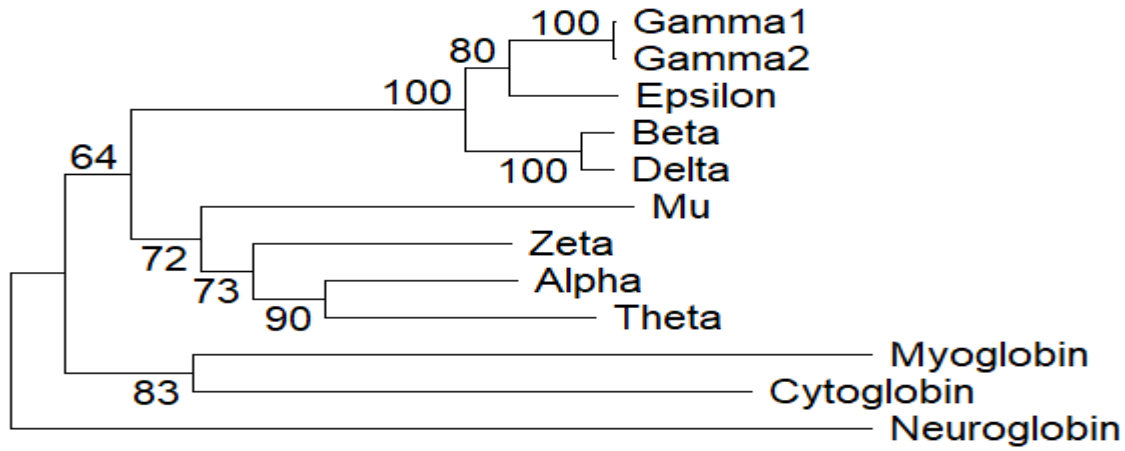




（3）点击主菜单中数据 Data 按钮，在下拉菜单中选择系统发生分析 Phylogenetic Analysis 选项，在主菜单中点击系统发生 Phylogeny 按钮，在下拉菜单中选择邻接法 Construct/Test Neighbor-Joining Tree, 在弹出窗口中将系统发生树稳定性测试 Test of Phylogeny 选项改为自举法 Bootstrap Method, 将自举法重复次数改为 100, 其它参数采用默认值。







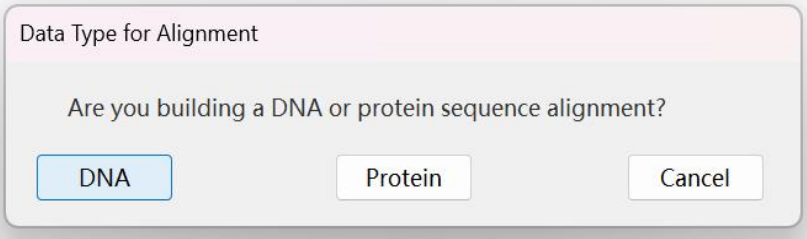
数字代表分支可信度，越高代表可信度越高；枝长越短，表示序列相似性的可能性更高，在同一支最末端，亲缘关系高。

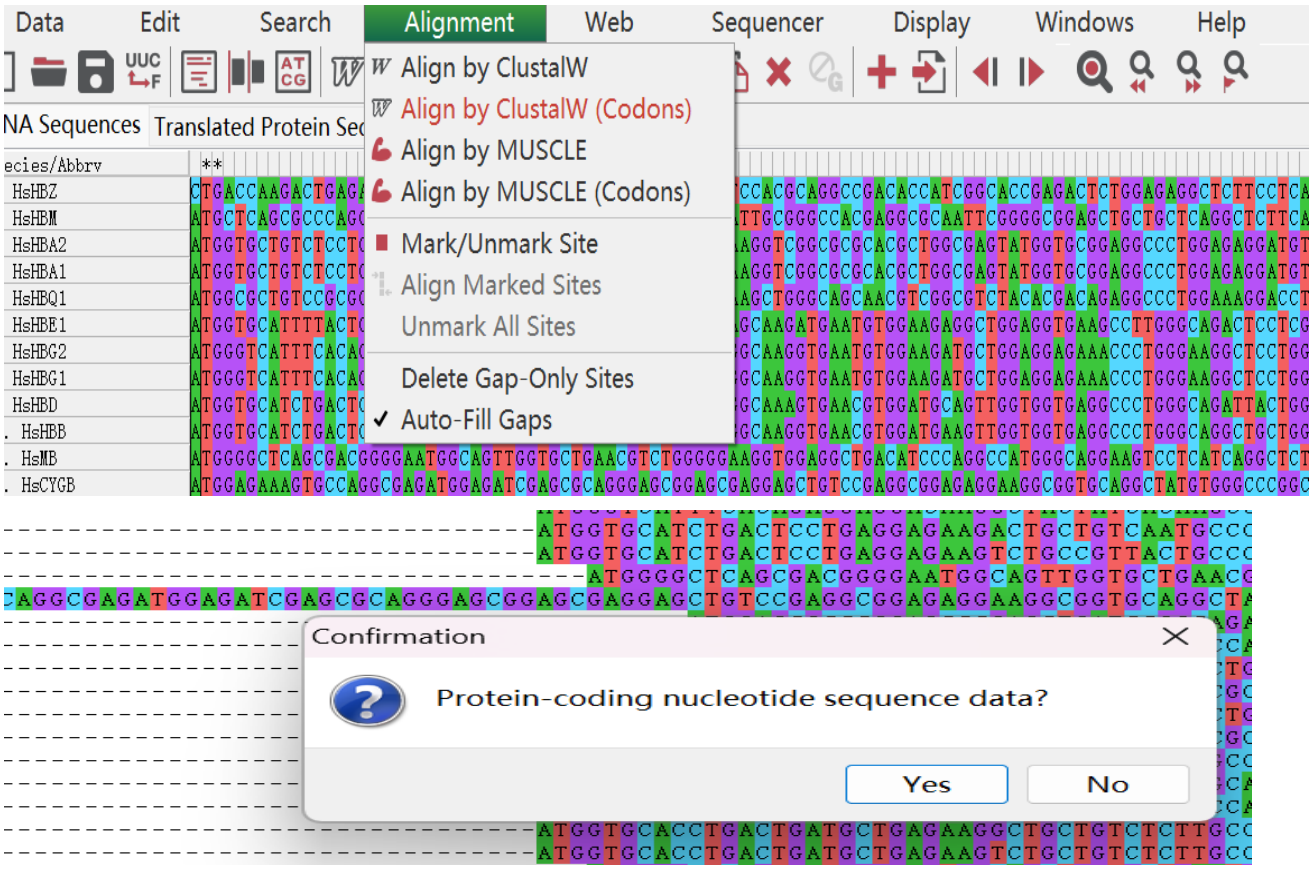
4. 人、小鼠和大鼠三个物种珠蛋白家族系统发生树实例

以人、小鼠和大鼠三个物种珠蛋白家族 37 个成员编码区序列构建系统发生树。

（1） 复制人、小鼠、大鼠 37 个珠蛋白基因编码区 CDS 序列。主要步骤与 3 类似。

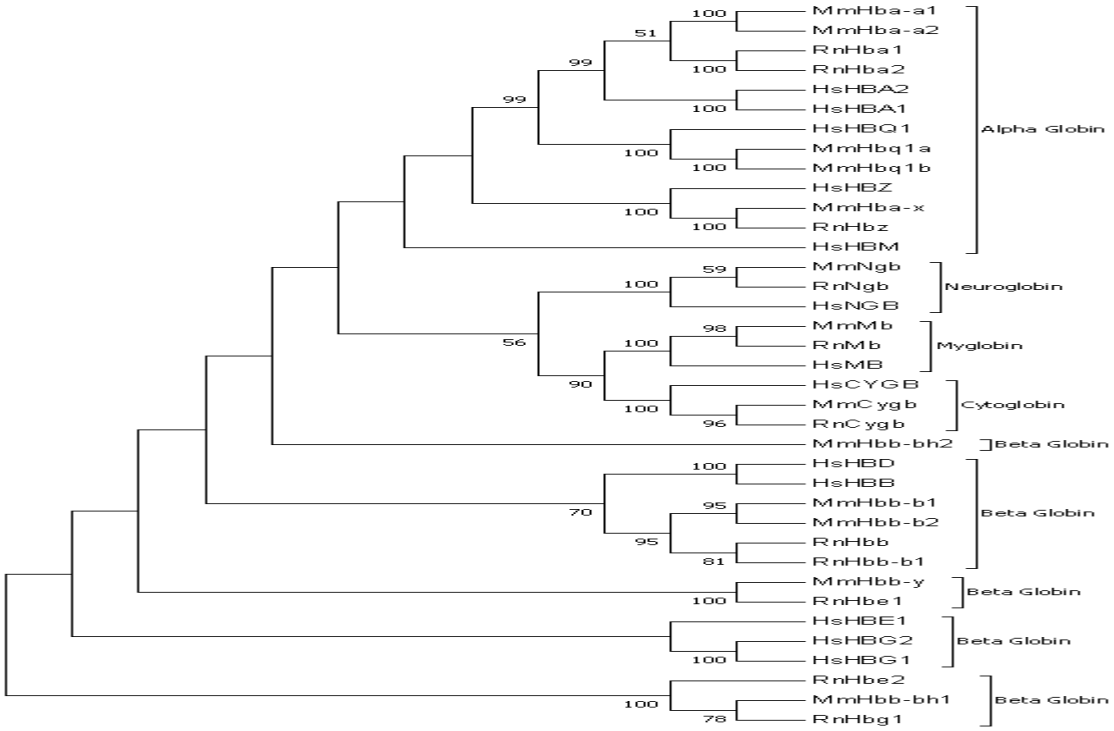
（2） 在数据类型 Data Type 弹出窗口中选择 DNA。比对 Alignment 按钮，在下拉菜单中选择 Align by ClustalW (Codons)。系统发生分析 Phylogenetic Analysis 选项，在弹出会话窗口中点击 Yes，选择比对蛋白质编码序列 Protein-Coding neucleiotide sequence data。在下拉菜单中选择邻接法 Construct/Test Neighbor-Joining Tree，在弹出窗口中替换类型 Substitution Type 中选择氨基酸 Amino Acid。





Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Scope	*All Selected Taxa*
Statistical Method →	*Neighbor-joining*
PHYLOGENY TEST	
Test of Phylogeny →	*Bootstrap method*
Bootstrap Replicates →	*100*
SUBSTITUTION MODEL	
Substitutions Type →	Amino acid
Genetic Code Table →	Nucleotide / Syn-Nonsynonymous / Amino acid
Model/Method →	
Substitutions to Include →	*All*
RATES AND PATTERNS	
Rates among Sites →	*Uniform Rates*
Pattern among Lineages →	*Same (Homogeneous)*
DATA SUBset TO USE	
Gaps/Missing Data →	*Pairwise deletion*
SYSTEM RESOURCE USAGE	
Number of Threads →	*8*



3.血红蛋白 alpha 亚基系统发生树实例

从脊椎动物中选取若干代表性物种，根据传统分类学知识，描述它们之间的系统发生关系，用 MEGA 软件中 User Tree/Display Newick trees 绘制系统发生树。

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

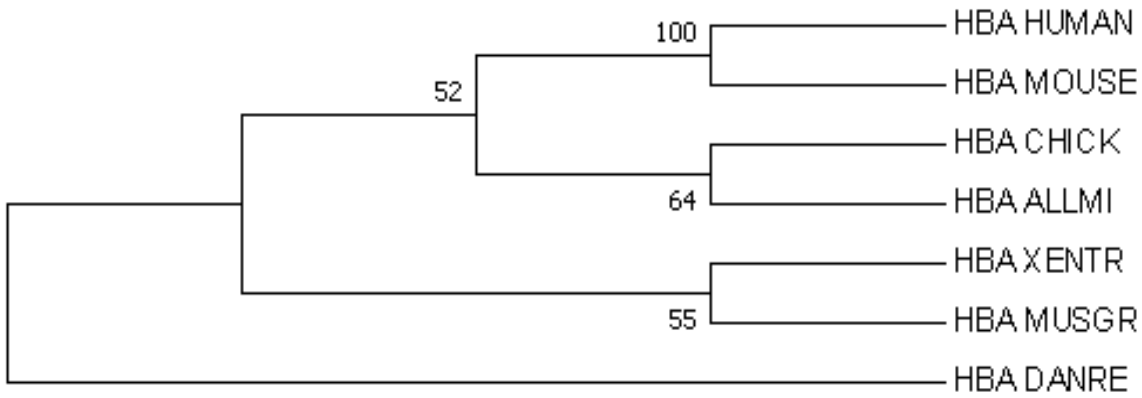
Creat a New Aligment, 在数据类型 Data Type 弹出窗口中选择蛋白质 Protein

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

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

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。



人和小鼠的 HBA 序列最先聚在一起，自展值是 100（满分），说明它们的亲缘关系非常近，这也符合生物学常识 —— 都属于哺乳动物，分化时间晚，基因序列相似度极高。人 + 小鼠的分支，和鸡 + 鳄龟的分支汇合，自展值 52。这个数值偏低，说明 “哺乳动物和鸟类 / 爬行动物的分化节点” 的统计支持度不算很高，但整体趋势是：哺乳动物先和羊膜动物（鸟类 + 爬行动物）聚成一支，再和两栖类分开。非洲爪蟾（XENTR）和 MUSGR 聚在一起，自展值 55，支持度中等。两栖类作为脊椎动物中 “从水生到陆生” 的过渡类群，和羊膜动物（哺乳 + 鸟 + 爬行）的分化时间更早。斑马鱼作为硬骨鱼，在树的最外侧，和所有陆生脊椎动物的分支都分开了，说明鱼类是这些物种中分化最早的，是整个树的 “外类群”，用来给树定根，体现了脊椎动物从水生到陆生的进化顺序。