“实用生物信息技术”课程小组讨论总结报告

组：G1　　次：R5　　组长：薛林蕾　　执笔：吴鸿珍，江浩燊，薛林蕾

1. 时间

2026.5.17

2. 方式

线上讨论

3. 主题

系统发生分析复习总结

4. 内容

一、 基本概念

1．分子演化和系统发生

分子演化：研究 DNA、RNA 或蛋白质序列在进化过程中的变化规律（如突变、选择、漂变）。

系统发生：研究物种或基因之间的进化关系，通常以树状图表示。

2． 序列相似性（Similarity）和序列同源性（Homology）

相似性：两序列在比对中相同或保守替代的位点比例，是一个定量指标（如 80%相似）。

同源性：指两个序列源自共同祖先，是定性概念（要么同源，要么不同源），不应说“%同源性”

3．直系同源（Ortholog）和旁系同源（Paralog）

直系同源：不同物种中由共同祖先基因垂直遗传而来的基因，通常功能相似。

旁系同源：同一基因组中由基因复制产生的同源基因，功能可能发生分化。

4．核苷酸替换模型和氨基酸替换模型

核苷酸替换模型：描述不同核苷酸之间替换的概率（如 JC69、K80、GTR）。

氨基酸替换模型：描述不同氨基酸之间替换的概率（如 Dayhoff、WAG、LG），一般基于经验数据或物理化学性质。

5.进化分支树（Cladogram）和系统发生树（Phylogram）

进化分支图：只表现分支顺序（拓扑结构），枝长无实际进化意义。

系统发生树：枝长代表遗传变化量（如替换数或时间），体现进化距离。

6.基因树和物种树

基因树：基于某一基因序列重建的进化树，可能与物种历史不完全一致（因不完全谱系分选、基因重复/丢失）。

物种树：反映物种形成事件的真实物种进化历史。

7.无根树和有根树

无根树：只显示节点间亲疏关系，不确定进化方向。

有根树：有唯一根节点，表示所有序列的共同祖先和进化方向。

8.分支和节点

分支：连接两个节点的边，代表一个进化谱系。

节点：分支的交点，代表分类单元或祖先。

9.内部节点和外部节点

内部节点：代表假设的祖先序列（未直接观测）。

外部节点：即叶节点，代表现存或已知的分类单元（序列）。

10.根节点和叶节点

根节点：有根树中最祖先的节点，代表整个树的共同祖先。

叶节点：树的末端，代表现存物种或序列。

11.距离法和位点法

距离法：先计算序列对间遗传距离，再基于距离矩阵建树（如邻接法 NJ、UPGMA）。

位点法：基于每个位点上的字符状态重建树，利用信息位点（如最大简约、最大似然、贝叶斯）。

12.最大简约法和最大似然法

最大简约法（MP）：选择需要最少替代次数（突变步数）的树。

最大似然法（ML）：在给定替换模型下，选择使观测数据出现概率最大的树。

二、参阅 ABC 网站中有关资料，查阅相关文献，回答以下问题

1）构建系统发生树的基本步骤

构建分子系统发生树（phylogenetic tree）的典型流程大致包括以下几个步骤。

1.确定问题与选择合适的基因 / 区段

明确要研究的是“物种之间的亲缘关系”还是“同源基因之间的进化关系”，选择具有代表性的分子标记，例如 rRNA、线粒体基因、叶绿体基因、核基因或同源蛋白序列等，要求各序列之间具有可比性并来源可靠。

2.序列收集与整理

从数据库（如 NCBI、Ensembl、UniProt 等）检索目标物种或基因的序列，去除明显不完整或标注错误的条目，对名称、ID 做好规范，必要时进行冗余序列筛除（过于相同的重复序列可删除）。

3.多序列比对（MSA）

使用 ClustalW、MAFFT、MUSCLE 等软件，对同源 DNA 或蛋白序列进行多序列比对，得到各位点一一对应的 alignment。 比对后应人工检查明显错误的对齐，特别是编码序列是否保持密码子框一致、内含子/高变区是否对齐合理。

4.比对结果修剪（选择保守区）

对于差异较大的区域，可使用 Gblocks 等工具或手工去除比对质量差、存在大量插缺或高度不确定的区段，保留保守且可信度较高的位点用于构树，从而提高系统树的稳定性与可靠性。

5.选择替代模型和建树方法

根据数据类型（核苷酸 / 氨基酸）和问题类型，选择合适的分子进化模型（如 JC69、K2P、GTR、JTT、WAG 等），并选用对应的建树方法，如距离法（Neighbor-Joining）、最大简约法（Maximum Parsimony）、最大似然法（Maximum Likelihood）、贝叶斯推断（Bayesian inference）等。 对于 ML/BI 通常需要先做模型选择（如用 ModelTest、MEGA 模块、MrModelTest 等）。

6.构建系统发生树并评估置信度

使用 MEGA、RAxML、IQ-TREE、MrBayes 等软件，根据上述模型与方法计算系统发生树拓扑结构，并通过自举法（bootstrap）或贝叶斯后验概率等指标评估各节点的置信度，绘制带支持度的系统发生树。

7.树的根定与可视化美化

根据需要选择适当的外群（outgroup）或其他方法确定树根，将无根树（unrooted tree）转化为有根树（rooted tree），并利用 iTOL、FigTree 等工具进行树形结构、分支颜色、注释信息的美化与导出，以便解释和发表。

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

选择用 DNA 序列还是蛋白质序列构树，主要取决于序列间相似度、研究层级以及比对难易程度等因素。

1.序列相似度中等到较低（远缘关系）时，优先考虑蛋白序列

蛋白氨基酸序列相较于核苷酸更为保守，更能反映远缘同源关系，当 DNA 序列相似度较低（如低于约 70%）时，多重比对结果通常不稳定，不同比对可能导致不同拓扑，此时采用蛋白序列更容易获得可靠的 alignment 和系统树。

2.序列相似度很高（近缘关系）或需要分辨最近分化事件时，DNA 序列更合适

对于近缘物种或同一物种内部品系，蛋白序列往往完全相同或差异极少，而密码子沉默突变等信息会丢失；此时使用核苷酸序列可以提供更多变异位点，提高分辨率。

3.编码序列比对时要保持阅读框与密码子结构

对编码区 DNA 进行比对时，插缺位点的长度应该是 3 的整数倍，新空位应以 3 bp 为单位出现，否则提示比对可能不合理，需要人工调整；这是使用核苷酸序列构树时必须注意的问题。

4.功能区域和分子标记的选择

编码蛋白质的基因（如单拷贝核基因、线粒体/叶绿体蛋白编码基因）和 rRNA 基因常用于系统分析，前者更适合编码区蛋白序列构树，后者多用 16S/18S rRNA 的严格比对以及二级结构信息辅助。 内含子、非编码区、高度可变区因比对难度大、插缺多，除非针对特定群体遗传研究，一般谨慎用于深度系统发育分析。

5.综合原则

远缘 / 跨大类群：优先蛋白序列；

近缘 / 种内或属内：可优先 DNA 序列；

若蛋白完全一致，可使用对应 DNA 序列增加信息量；

始终优先选择同源、可比、注释可靠且比对质量高的序列区域。

3）利用自举法（Bootstrap）检验系统发生树稳定性的原理

Bootstrap 自举法是评估系统发生树节点可靠性的常用统计方法，其基本思想是通过对比对矩阵做“重采样建树”，看某个分支在重复建树中出现的频率，从而给出该节点的支持度。

具体原理为：

在原始多序列比对基础上进行列重采样

将已经对齐的序列看作一个由 N 个比对位点组成的矩阵，自举法从这 N 个位点中“有放回”地随机抽取 N 次，形成一个长度同样为 N 的新比对数据集。由于是有放回抽样，某些位点可能被重复抽到，某些则可能未被抽到。

对每一个自举数据集构建一棵树

对上述每个 resampled alignment 使用同样的建树方法（如 NJ 或 ML）构建系统发生树。重复上述重采样和建树过程 B 次（常见为 500–1000 次）。

统计每个分支在所有自举树中出现的频率

对原始树上的每个内部分支（clade），统计在 B 棵自举树中该 clade 出现的次数。出现频率 = 出现次数 / B，通常以百分数形式显示在树的相应节点上，即 bootstrap 支持率（如 85%、99% 等）。

解释支持度

一般认为 bootstrap 值 ≥ 70% 的分支具有较可靠的统计支持，≥ 90% 则非常稳健；低于 50% 的分支通常支持较弱，在绘图时可不显示其数值或不特别强调。 需要注意的是，bootstrap 只是对给定数据和方法下树拓扑的内部一致性检验，并不能完全消除模型不当、比对错误等系统性偏差。

4）确定无根树根节点的方法

系统发育分析中很多算法（如 NJ、部分 ML 方法）首先生成的是无根树，即只表示相对拓扑关系而不指明进化方向，要解释“谁先谁后”需要对树进行定根（reroot）。

常用的定根方法包括：

外群法（Outgroup rooting）——最常用、最直观

选择一个或多个与研究对象（内群）有共同祖先，但在系统上已明确偏离的物种 / 序列作为外群（outgroup），将外群挂在树的一端，将外群分支与内群分支相连的节点视为根，从而确定树的方向。 合理的外群应与内群有适当的系统距离，既不会太近（难以分辨），也不能太远（易产生长枝吸引）。

中点定根法（Midpoint rooting）

当没有合适外群或外群不明确时，可采用中点定根：在无根树中寻找距离最远的两端点，其路径中点作为根，将树“放置”在这一点，从而使根到所有叶子的路径总长差异最小。这种方法隐含假设各分支演化速率近似均一（分子钟假设），在速率严重不均时可能偏差较大。

已知信息辅助定根

若已有独立证据（如化石记录、已发表的高置信度系统树或 Biogeography 信息）指出某一分支更早分化，也可以人工指定该分支作为根或靠近根的位置，然后用软件（如 iTOL、FigTree 等）进行 reroot 操作。

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”

这个问题往往出现在“基因家族演化”与“物种分化”交织的背景下，核心思路是比较基因树（gene tree）和物种树（species tree）的拓扑关系，判断基因复制事件发生在物种分化之前还是之后。

1.构建物种系统树

基于多个保守标记或全基因组信息构建物种树，作为物种间亲缘关系和分化顺序的参照框架。

2.构建目标基因家族的基因树

对该家族在不同物种中的同源序列（包括正交和旁系同源）进行比对并建树，得到基因树拓扑结构。

3.比较基因树拓扑与物种树拓扑

若基因树上同源基因的分支模式基本与物种分化顺序相一致，且每个物种一般只有一个拷贝，说明该基因拷贝在物种分化前可能只经历过少量复制事件，更多是“物种先分化，基因随物种一起分化”（先有物种层面的分歧）。

若在物种分化之前就已经发生基因复制（如在同一物种中存在 paralogs，它们各自又在不同物种中形成对应 clade），表现为“基因树的某些复制事件节点深于物种分化节点”，说明是“先有基因复制，再有物种分化”。

相反，如果基因树显示某些复制事件发生在物种分化之后（同一物种内多个拷贝互为最近亲，而不同物种间拷贝间距较远），则可以推断在某些谱系内有后期基因扩张，是“先有物种分化，再有该基因在某些谱系内的扩增”。

通过对比基因树和物种树，解析哪些节点代表“物种分化”，哪些节点代表“基因复制 (gene duplication)”或“基因丢失 (gene loss)”，即可在一定程度上回答“先有物种还是先有基因”的问题。

6）不同建树方法的基本原理和特点

常见的分子系统发育树建树方法大致可分为距离法、最简约法、最大似然法和贝叶斯方法等几大类，它们的原理和特点如下。

距离法（Distance methods，以邻接法 NJ 为代表）

原理：先将多序列比对转化为两两序列间的遗传距离矩阵（如 p-distance、K2P、Poisson、JTT 等），再使用层次聚类或最小进化等准则构建能最好解释该距离矩阵的树（NJ、UPGMA 等）。

优点：计算速度快，适用于大规模数据；实现简单、使用广泛。

缺点：将序列信息压缩成一个距离值，丢失位点水平信息；对于进化速率高度不均、长枝吸引等情况容易受到影响，精度有限。

最大简约法（Maximum Parsimony, MP）

原理：不事先假设复杂的替代模型，而是在所有可能的树拓扑中，寻找使得解释所观测的性状（各位点碱基/氨基酸状态）所需变异次数最少的那棵树，即“最简约”的树。

优点：概念直观，不依赖复杂的进化模型，适用于少数物种、较高相似度序列。

缺点：对同位点的平行突变和回复突变敏感；当序列长度有限且同形性变异较多时，容易给出错误拓扑；计算复杂度随物种数上升很快。

最大似然法（Maximum Likelihood, ML）

原理：在给定进化模型下，计算不同树拓扑产生观测序列数据的似然值，选择似然值最大的树作为最佳系统树；同时可估计分支长度、模型参数等。

优点：有坚实的统计学基础，可处理复杂的替代模型（碱基频率、替代率矩阵、位点速率异质性等），对多种数据类型适用，结果通常较为可靠，是当前常用的主流方法之一。

缺点：计算量较大，尤其是物种数和位点数较多、模型较复杂时计算成本高；需要选择恰当的模型并进行检验。

贝叶斯推断法（Bayesian inference, BI）

原理：在指定先验分布和进化模型的前提下，利用马尔可夫链蒙特卡洛（MCMC）方法估计树拓扑和参数的后验概率分布，得到一组采样树，再根据后验概率筛选出代表性系统树并给出节点 posterior probability。

优点：可以直接对树和参数进行联合统计推断，后验概率具有直观统计意义；在某些情况下重建准确性优于 ML；可灵活处理复杂模型与不确定性。

缺点：实现较复杂，对 MCMC 收敛和设置敏感；计算时间长，参数不当可能导致结果不稳定。

邻接法（NJ）与 ML/BI 的使用建议

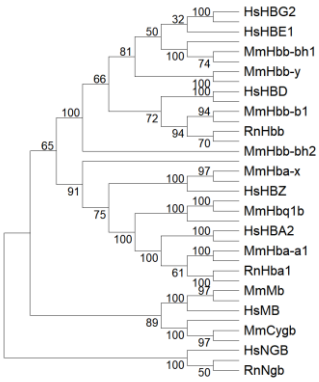
实践中常见的经验是：对于大规模或初步探索性分析，可以先用 NJ 快速构树；对于关键数据集和重要结论，建议使用 ML 或 BI，在合适模型下得到更可靠的树，并用 bootstrap 或后验概率评估支持度。

总体而言，各建树方法各有优劣，实际应用中往往需要结合数据特征、计算资源和研究目的综合选择，必要时可采用多种方法构树进行交叉验证，以增强结论的可信度。

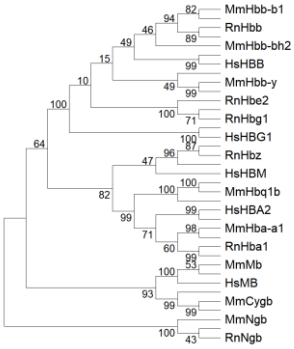
三、 人、小鼠和大鼠三个物种珠蛋白家族系统发生树实例

1) 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，采用邻接法、最大简约法和最大似然法构建系统发育树，选择适当的替换模型和参数， 比较采用不同方法、不同模型和不同参数时所构建的系统发生树的拓扑结构和稳定性值。

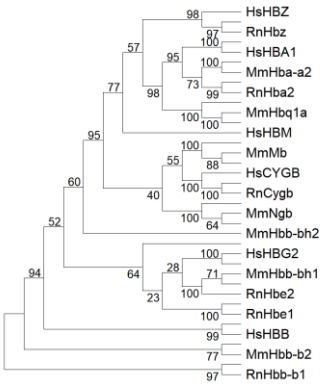
邻接法：



最大似然法



最大简约法



拓扑结构一致性：NJ 法和 ML 法构建的系统发育树在核心亚家族划分上高度一致，符合珠蛋白家族的进化关系；MP 法拓扑偏差较大，不适合该数据集的分析。

稳定性对比：NJ 法的关键节点稳定性最高，ML 法次之（部分节点受模型影响），MP 法整体稳定性较差。

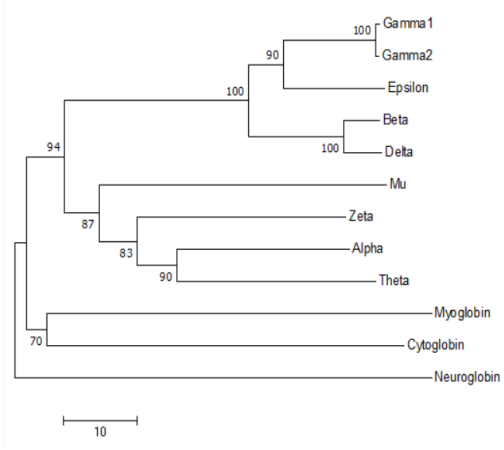
2) 根据上述人、小鼠和大鼠三个物种珠蛋白家族 37 个成员编码区序列系统发生树，参阅相关文献，说明珠蛋白基因家族的起源和演化。

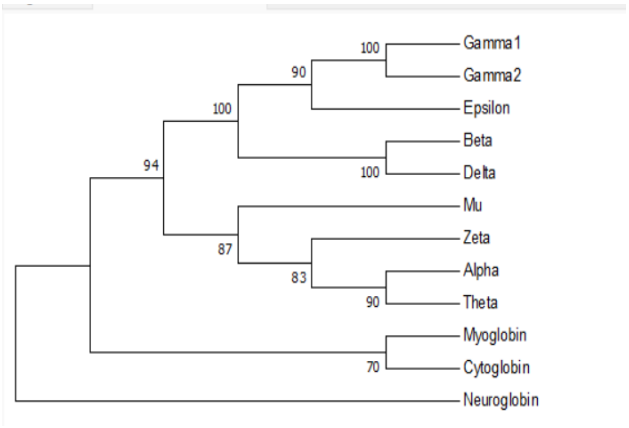
珠蛋白基因家族的演化是一个从单基因祖先出发，通过多次串联复制、全基因组复制和物种特异性复制事件，逐步分化为功能特化的亚家族的过程。系统发育树清晰展示了：非传统珠蛋白（Mb、CYGB、NGB）作为古老分支，在脊椎动物中高度保守；α / β 珠蛋白亚家族通过多次复制事件扩张，形成了发育阶段特异性的基因簇；灵长类和啮齿类的 β - 珠蛋白亚家族发生了独立的适应性演化，产生了物种特异性的基因拷贝。

三、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。

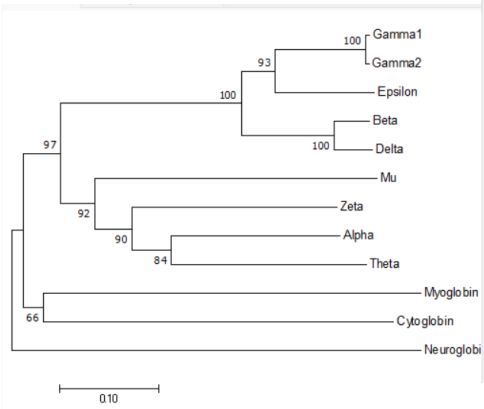
model　　original tree　　consensus tree

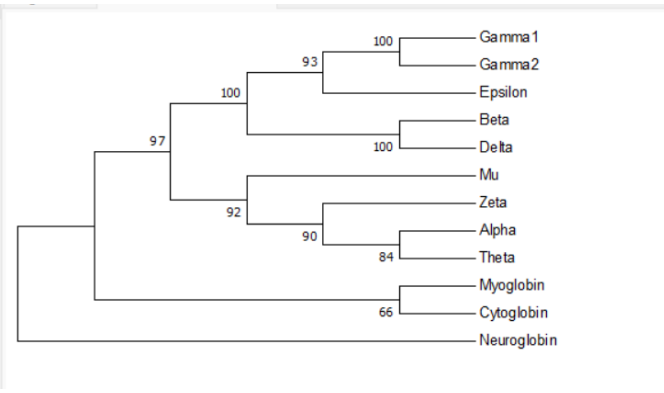
Non of distance



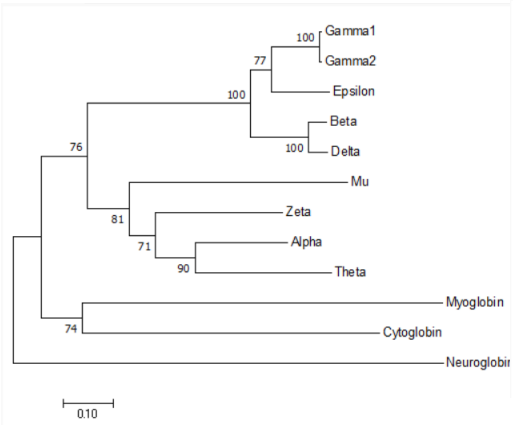


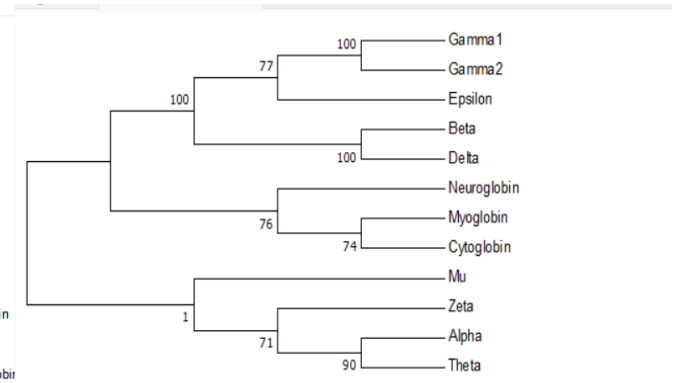
p-distance



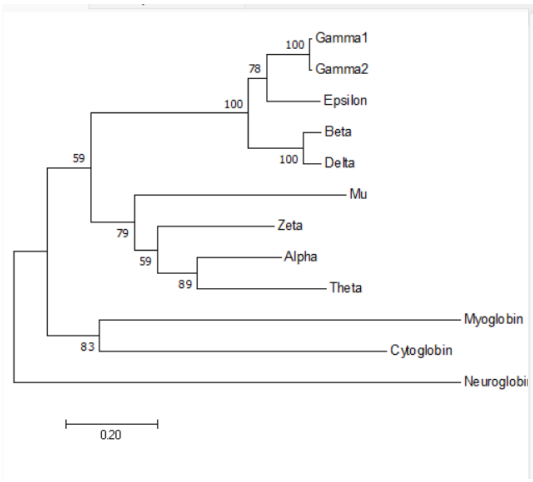


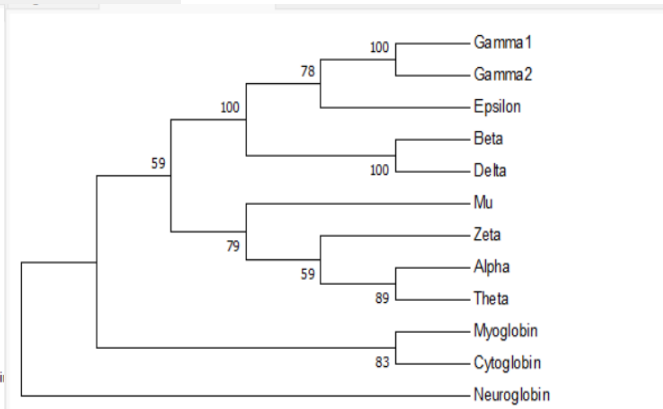
Poisson model



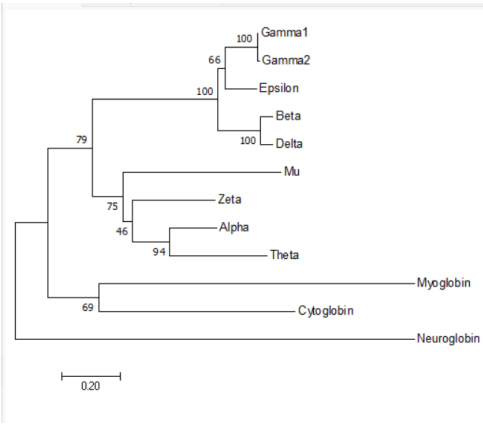


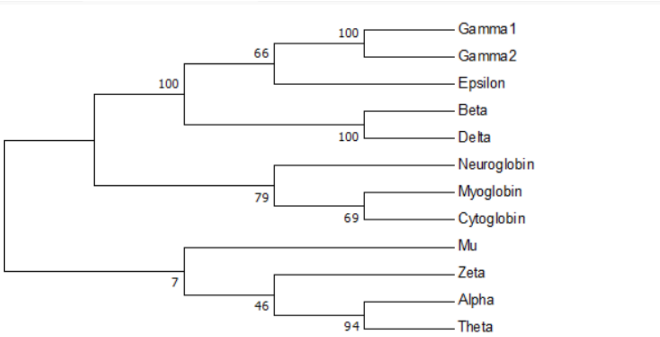
Equal input model



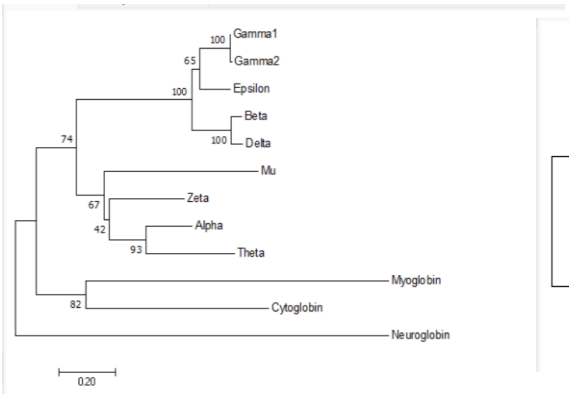


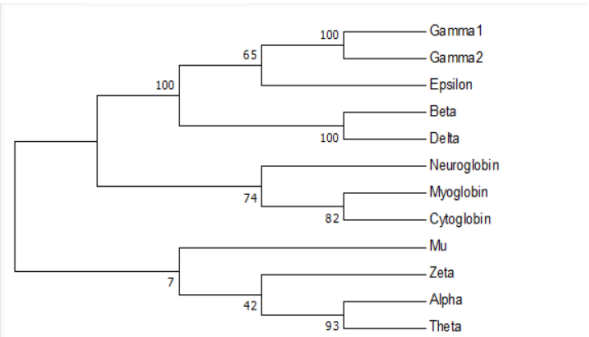
Dayhoff model





JTT model





1.No. of differences (差异个数)最直观的模型。直接数一下两条序列在相同位置上，氨基酸有多少个不一样。不考虑进化历史，也不区分“一种氨基酸变成另一种”的难易程度。如果两条序列差异很大，可能会发生多次替换（比如 A→B→C），直接数差异数会严重低估真实的进化距离。适用于差异极小的序列。

2. p-distance (p-距离)其实就是把上面差异个数换算成了比例。与差异个数一样，它也没有校正多次替换的情况。随着进化时间变长，p-距离会趋于饱和（不再线性增长），不适合差异较大的序列。适用亲缘关系非常近的序列（例如同一物种的不同个体）。

3. Poisson model (泊松模型)是第一个基于校正的模型。它假设在任何位点上，氨基酸替换发生的概率很小，且服从泊松分布。校正了可能发生的“多次替换”（包括平行替换、回返替换），因此比前两个模型更准确。

在现实中，不同氨基酸之间替换的难易程度差别很大（例如理化性质相近的更容易替换），这个假设过于简化。

4. Equal input model (等输入模型)也叫均衡模型。它假设每个氨基酸被替换成其他任何氨基酸的概率是相等的，但这个概率取决于目标氨基酸在整体序列中的出现频率。比 Poisson 模型略微精细，因为考虑到了不同氨基酸在自然界中存在的丰度不同。

5. Dayhoff model (Dayhoff 模型)是第一个也是最经典的经验模型。它不靠假设，而是靠统计大量真实数据得出的结论。有些替换（如异亮氨酸 ↔ 缬氨酸，都很疏水）非常容易发生；有些替换（如精氨酸 ↔ 脯氨酸）极其罕见。

6. Jones-Taylor-Thornton (JTT) model (JTT 模型)是对 Dayhoff 模型的重大升级版。JTT 至今仍是蛋白质系统发育分析中最常用、最经典的模型之一。如果你不确定选哪个，JTT 通常是一个很好的默认选择。

7.结果分析：用不同模型得到的进化树拓扑结构是一样的，但是稳定性值有所不同，Equal input model 稳定性值最低，说明结果不太可靠。JTT 模型的稳定性值普遍较高，结果最为可靠，这可能是因为 JTT 模型基于实际的数据库得出结论，能够更精确地描述氨基酸替换的**实际速率矩阵**。对于珠蛋白这种有丰富进化信息、但已出现一定饱和的家族，JTT 最为合适。

5. 问题

1.何理解“基因树”和“物种树”之间可能不一致？在利用系统发生树推断进化历史时应注意哪些问题？

在分子系统发育分析中，需要区分两类不同“树”：一类是基于多个基因或全基因组信息推断的物种树（species tree），反映的是不同物种之间的分化历史；另一类是针对某一基因家族或若干同源序列构建的基因树（gene tree），反映的是该基因在不同物种和不同基因拷贝之间的复制、丢失与分化历史。两者在拓扑结构上并不总是一致，这种“不一致”本身就是进化历史的重要信息来源，而不是纯粹的“错误”。

2.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，为什么邻接法系统发生树的拓扑结构和稳定性值最好？

算法适配数据特征：NJ 法基于遗传距离构建树，对珠蛋白家族序列的替换速率异质性（如非传统珠蛋白保守、β- 珠蛋白演化活跃）鲁棒性更强，避免了 MP 法的长枝吸引和 ML 法对模型参数的过度依赖，拓扑结构更稳定。

自展统计更可靠：NJ 法的距离计算对序列噪声和重采样波动的容忍度更高，核心分支的自展值普遍较高，能更清晰地反映同源基因的真实聚类关系。

3. 如果两棵树来自不同模型或不同基因如何判断哪棵更稳定？

可以比较两棵树上对应内部节点的自举值（如用 1000 次自举），选择平均自举值更高，且低支持率（<70%）的节点更少的系统发生树。