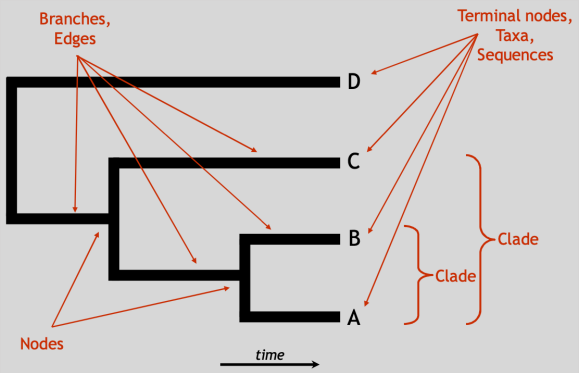
Lab4 系统发育

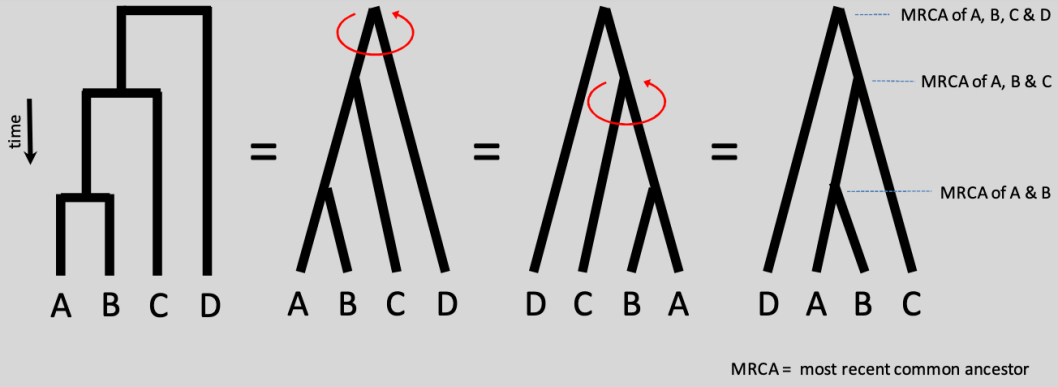
今天，我们将使用上一个实验室生成的多序列比对（MSA）来推断序列之间的进化关系。 系统发生分析产生的分支图可以清楚地说明序列之间的关系，而这些关系从诸如BLAST或MSA之类的分析中并不明显。 系统发育树显然可用于专注于进化关系和趋异模式的进化和比较研究，但对于生成有关分子和生化研究的基因或蛋白质功能的假设，它们也变得越来越重要。。

系统发育学是一个广阔的领域，通过介绍一些非常基本的工具和思想，我们将在本实验中简单介绍一下这个方法。大多数系统发育工具分为两个主要类别：基于距离的方法和基于特征的方法。我们将使用其中一种方法。在此实验室中，您的目标不仅是要了解如何构建系统发育树，还要更好地认识到这些不是像生物信息学方法那样经常采用的剪切和粘贴分析方法。通常，不同的方法可以得出不同的结论。因此，你需要使用多种方法和参数值，并最终利用头脑中萌发的生物学直觉来生成最佳树和最强分析。有关如何解释此类分析的信息，请参见Box7。

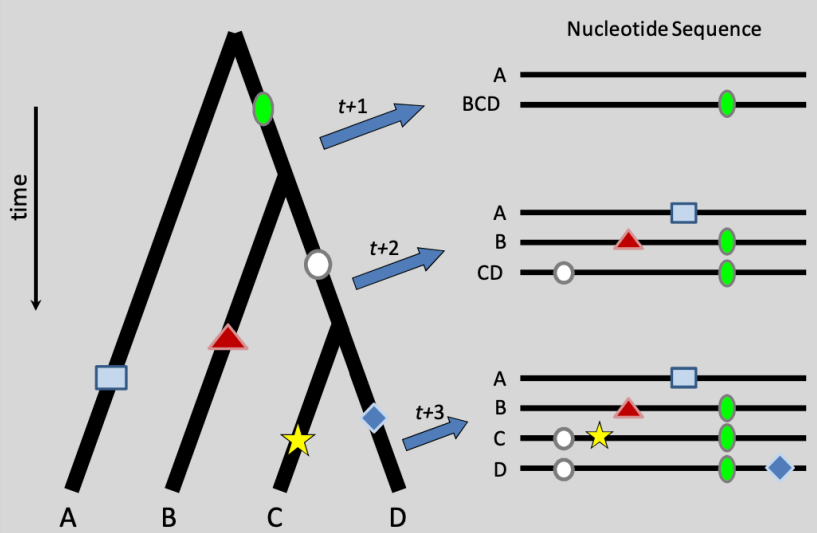
Box1. Anatomy of Phylogenetic Trees
系统发育树的结构



系统发育树可以以许多不同的形状和方向呈现。 要认识到的重要一点是，确定两个序列之间进化距离的唯一方法是找到共同祖先之前必须回溯多长时间。 因此，在右边的树中，即使序列A在页面上以字母的形式在物理上更接近于D而不是在物理上接近于C，但实际上它们与C的关系更为密切，因为它们共享了更近的共同祖先。 A-D之间的进化关系也可以使用Newick格式表示如下（（（（A，B），C），D）：括号的嵌套遵循上述树的分支。



Box2. The growth of Phylogenetic Trees
系统发育树的建成



随着生物的进化和多样化，它们的世系会累积突变（以彩色形状表示）。 这些突变将被传播给所有后代和后代，因此这种突变发生在该组病史的早期，例如 在所有三个后代谱系中都发现了绿色椭圆形突变，它发生在序列B，C和D的祖先中。 以后可能发生的突变（例如蓝色方块突变）可能会出现在较少的血统中。 核苷酸序列上突变的位置是完全任意的，仅表示每个时间点有多少个独特序列，以及这些序列之间的突变分布。

Neighbour-Joining

Box3. Distance-Based Methods
基于距离的方法

尽管这些方法的工作方式之间存在非常大的差异，但是所有基于距离的方法都通过先在所有序列之间进行成对比较并生成每对之间的遗传距离来生成“距离矩阵”，从而进行系统进化重建。在最简单的情况下，遗传距离只是序列之间错配的数量，尽管大多数距离矩阵都使用更复杂的距离度量。在最简单的情况下使用成对距离矩阵来识别两个最相关的序列会形成了树的前两个分支。然后重新制作这些成对距离矩阵，但是这次在最后一步中确定的两个最相似的序列由树中的单个节点表示。现在，确定下一个最相似的序列对（或序列-节点，或节点-节点），并重复该过程，直到表折叠为单个节点为止。然后使用序列与节点之间的所有距离绘制一棵树。

基于距离的方法的优点是计算速度非常快且构造简单。对于大多数数据集而言，通过“Neighbour-Joining”或“Minimum Evolution”之类的方法生成的树也相当可靠。然而，所有基于距离的方法也都遭受一个或多个相当强大的基础假设。其中最重要的是所谓的“附加距离”，其详细信息超出了本课程的范围（如果您有兴趣，请参阅“更多阅读材料”以获取更多详细信息）。只需说一下，如果你的序列子集的进化速度快于其他子集，就可以违反该假设。违反此假设可能会严重损害你的分析。

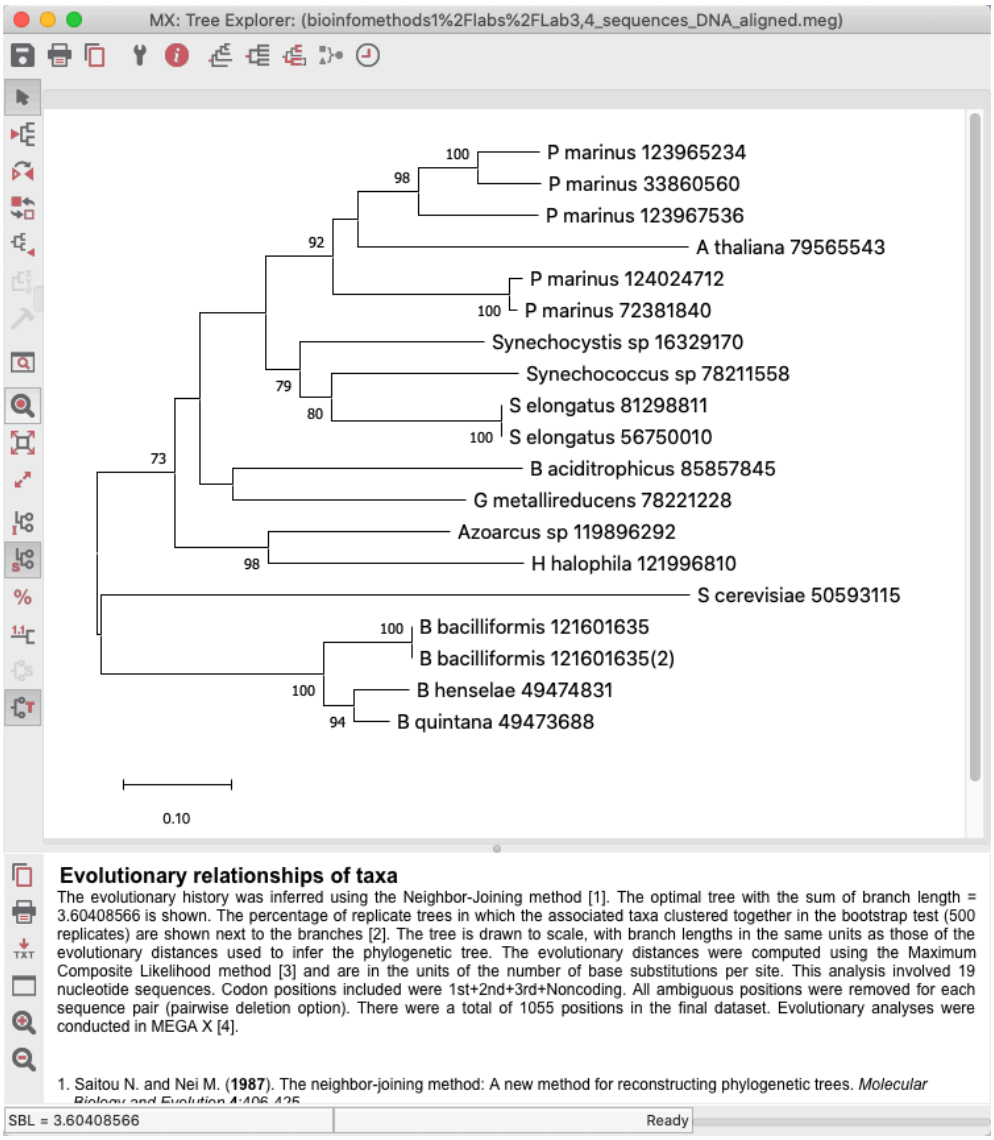
我们将从在MEGA X中建立邻近树开始。在你执行MSA时，上一个实验室介绍了MEGA。 MEGA（分子进化遗传分析）是一种非常易于使用但功能强大的应用程序，主要用于系统发育研究。邻近法（Neighbour-Joining）是快速且相当可靠的，因此在继续进行更严格的程序之前，可以通过有关一般树的拓扑/距离的假设来构建一棵优秀的入门树。

1. 打开MEGA X，然后将课程下载的比对DNA文件从FASTA格式转换为MEGA格式。(本github中的“ bioinfomethods1-labs-Lab3,4_sequences_dna_aligned.fas”文件)。
    a. 选择**File / Convert File Format to MEGA**…
    b. 要找到文件，您需要单击**Data file to convert**框右侧的非常小的文件夹图标。
    c. 以FASTA格式找到比对的核苷酸序列文件。 如果无法自动识别格式就手动选择fasta format。
    d. 选择**OK**
    e. 假设文件转换正确，请使用参考性名称进行保存。 您会注意到该文件附加了.meg文件扩展名。 这些简单文件的MEGA和FASTA格式非常相似，主要区别在于MEGA格式可以在不同字段中存储更多信息。
    f. MEGA格式有着以下的规则：输入文件的序列名称不能带有空格或以下任何内容：, ; : ‘ “ ! ? > < [ ] ~ @ # ^ &。同时圆括号必须成对使用。第一行一般要有**#MEGA**，第二行一般有**!Title:xxx**。

2. 打开新文件。
    a. 单击**File > Open a File / Session**（确保关闭了“Text File Editor”和“Format Converter Window”窗口）
    b. 找到并打开扩展名为**.meg**的新文件。
    c. 指定此文件包含**Nucleotide Sequences**，以及任何其他要求的信息应选择（protein coding sequence = Y，select genetic code = standard）。
    d. 单击“ TA”图标（主界面里的小图框）以打开数据浏览器。 这对于可视化和选择要分析的数据和区域很有用。
    e. 注意：如果文件中包含非法字符，MEGA会告诉你文件位于哪一行。

3. 返回主窗口，然后选择“**Analysis**”>“**Phylogeny**”>“**Construct / Test Neighbor-Joining Tree**”。这次我们将使用默认参数，但要保留默认参数，但要确保默认参数之一：“**Phylogeny Test / Test of Phylogeny**”下选择“**Bootstrap method**”。 单击“**Compute**”。

4. 您的输出应该是新窗口中格式正确的系统树。注意树中每个节点或分支上方的值。这些称为引导程序值，是对每个节点的统计置信度的度量。任何引导得分> 70通常都被认为是相当可靠的。
    a. 在“**Tree Explorer**”窗口中，转到“**View > Options**”（或单击扳手图标），然后选择**Branches**选项，然后展开“**Statistic / Frequency**”部分。
    b. 选择**Hide values lower than**的框并输入70％



5. 你可以在MEGA中非常简单地更改树的显示方式。
   a. 在“Tree Explorer”窗口中，转到**View > Tree Branch Style**，然后选择一些不同的格式，例如圆形，径向，传统笔直…
   b. 检查具有这些不同格式的序列的相对分支顺序

6. MEGA Tree Explorer非常强大。 您可以通过“**Options**”（View > Options）和“**Subtree**”菜单无休止地操作树。 将树恢复为传统 / 矩形格式，然后尝试使用某些控件选项。大多数“**Subtree**”选项也可以在窗口左侧作为图标使用。 请注意，这些更改都是可逆的，因此可以放心地玩这些游戏！

Box4. Rooting Phylogenetic Trees
带有根的系统树

系统发育树的根正好听起来像它应该是的样子–在树的根部有着最远的时间点。 尽管这是一个简单的想法，但是确定树的根实际上可能非常困难。 根树有两种主要确定方法：

1. 中点生根涉及将根放置在实际上是树的重心的位置。 这是MEGA和许多其他程序中使用的默认方法。 中点根非常容易做到（单击“视图”>“中点上的根”），但是这依据的假设是所有序列都以大约相同的速率进化。如果不是这种情况，那么中点生根可能是不合适的。

2. 利用外群进行组外生根涉及在分析中包含一个比其他序列更加发散的序列，然后确保它在树的最底部分支。 如果你有可靠的先验信息可以选择一个好的外群，则外群生根非常可靠。 不幸的是，许多研究并不了解此信息。
如果不确定树的生根，则始终可以显示一棵无根的树，该树从中心点辐射出的树没有方向代表时间（上面显示的放射状树）。 许多人发现这些树更难解释，但是它们提供了相同的信息，而没有增加关于哪个序列首先分支的假设。

7. 让我们看看改变树的根部如何影响我们的结论。 左键单击树的内部边缘之一（该边缘将包含在绿色框中），然后右键单击以选择“Place Root”。 该树将重新排列，以使您选择的边现在位于树的底部。
    a. 你可以通过选择“**View / Root on Midpoint**”来返回中点生根。

8. 使用“**View / Tree Branch Style > Radiation**”来绘制一棵无根发育树。

9. MEGA X还为每个分析生成一个自动注释。 选择标题以查看有关特定系统发育的详细说明（默认是在树下显示该说明）。 尽管您可能不想在出版物中“按原样”使用此标题，但此功能非常清楚地提供了描述分析的所有必要信息。

10. 现在，让我们用蛋白质序列制作一棵树。 你可以使“Tree Explorer”窗口保持打开状态，但是可以通过单击主窗口中的“**Close Data**”图标来关闭当前数据文件。使用相应的**corresdonding aligned protein sequence file**（本github中的“ bioinfomethods1-labs-Lab3,4_sequences_prot_aligned.fas”文件）：你可能是在上一个实验中自己生成了此比对，但是为了保持一致，请使用提供的版本）。 您需要先将比对的蛋白质序列转换为MEGA格式。

11. 使用邻近法建树并删除较低的校准值，同时将其和之间的核苷酸树进行比对。

12. 现在，让我们玩一些树构建参数，看看它们如何影响分析。 再次打开比对的核苷酸序列文件（请注意，您可以通过单击“**Close Data**”图标，然后选择“**File > Open a Recently Used File**”来完成此操作），然后选择“**Phylogeny > Bootstrap Test of Phylogeny > Neighbor-joining**”
    a. 在参数窗口中，修改一些参数并重建树。 请记住，你无法破坏任何事物，因此请尽可能尝试各种选择。 确定一棵树是否比另一棵树更好的一种很好的方法是查看引导程序分数。 尝试优化参数，以便为尽可能多的节点提供最强的节点支持（较高的引导程序分数）。 请特别注意以下内容：

i. **Gaps-Missing Data / Pairwise Deletion**: 如果你的序列中包含许多插入和删除（indels），则此参数很有用。完全删除（默认设置）会从分析中以任何顺序删除所有带有indel的比对列。成对删除仅从成对比较的分析中删除具有插入缺失的比对列。 尽管完全删除是最保守的方法，但有时indel的数目如此之多，以至于你会丢失大量信息。

ii. **Model / Nucleotide**: 这些替代模型就像先前实验室中讨论的替代矩阵一样。 尝试使用不同的模型，看看它们是否会影响您的结论。 有关某些模型的说明，请参见box 5。

iii. **Rates among sites / Different** (Gamma Distributed)

iv. **Gamma Parameter** / try from 0.1 - 2.0: 这控制了整个序列进化速率的变化。 例如，一个区域可能是高度保守的（进化非常缓慢），而另一区域根本不保守（进化速度要快得多）。较低的伽玛参数用于具有较高速率变化的序列。

Box5. Substitution Models

迭代模型

如前所述，迭代模型用于建模DNA或蛋白质序列在进化时间内的变化。虽然像PAM和BLOSUM这样的矩阵可用于建模蛋白质序列进化，但其他模型可用于DNA序列。以下是模型及其假设的一小部分：

1. Jukes-Cantor：所有核苷酸的频率相等；任何一个位点突变为另一个的频率均无偏差（相等的替代率）。

2. Felsenstein-81：所有核苷酸的频率不相等；平等的替代率。

3. Kimura 2-Parameter: 所有核苷酸的频率相等；转换（嘌呤到嘌呤，或嘧啶到嘧啶）和颠换（嘌呤到嘧啶，反之亦然）有不同取代率。转换发生的频率大约是颠换的两倍。

4. Taijima-Nei: 所有核苷酸的频率不相等；相等的转换频率；可变的过渡频率。

5. Tamura 3-Parameter: 所有核苷酸的频率相等；过渡和颠换的不同替代率；G + C含量偏差。

6. Hasegawa-Kishino-Yano (HKY): 所有核苷酸的频率不相等；过渡和颠换有不同替代率。

您如何选择要使用的模型？ 最重要的是，您需要查看数据。 如果核苷酸处于不同的频率，以及适当的过渡-转化率，这应该使您有所了解。 如果您真的很认真地考虑正确执行此操作，请使用名为jModelTest 2的程序，该程序使用似然法来帮助您确定最佳的替代模型和gamma参数。 该程序超出了本课程的范围，但是可以作为Java应用程序访问，网址为https://code.google.com/p/jmodeltest2/（Darriba等，2012）。

Box6. Character-Based Methods
基于特征的方法
同样，有许多基于特征的系统发生方法，它们以非常不同的方式起作用。 所有这些方法实际上都比较了MSA中每个比对列中每个残基（核苷酸或氨基酸）的状态。 他们试图确定解释数据中观察到的关系所需的最可能或最简单的解释。 他们通常通过查看所有可能的解释（换句话说，所有可能的树），并根据方法中使用的特定标准，确定能最好地解释数据的单个树或树集来做到这一点。

最大似然实际上描述了一种统计框架，在这种情况下该统计框架适用于系统发育重建。 它基本上遍历所有可能的树结构，并询问给你的特定数据集特定树的可能性。 因此，例如，非常相似的序列很有可能会非常靠近地分支（在树的尖端附近），而不是很远地分支（在树的根部附近）。

基于特征的方法（例如最大似然）通常是非常复杂的方法，可以对统计框架中的演化变化进行逼真的建模。 不幸的是，它们也可能更难以运行（或至少正常运行），并且也许最重要的是，由于它们有效地检查了每种可能的树结构，因此它们的计算量很大。 实际上，这意味着它们不能应用于非常大的数据集。

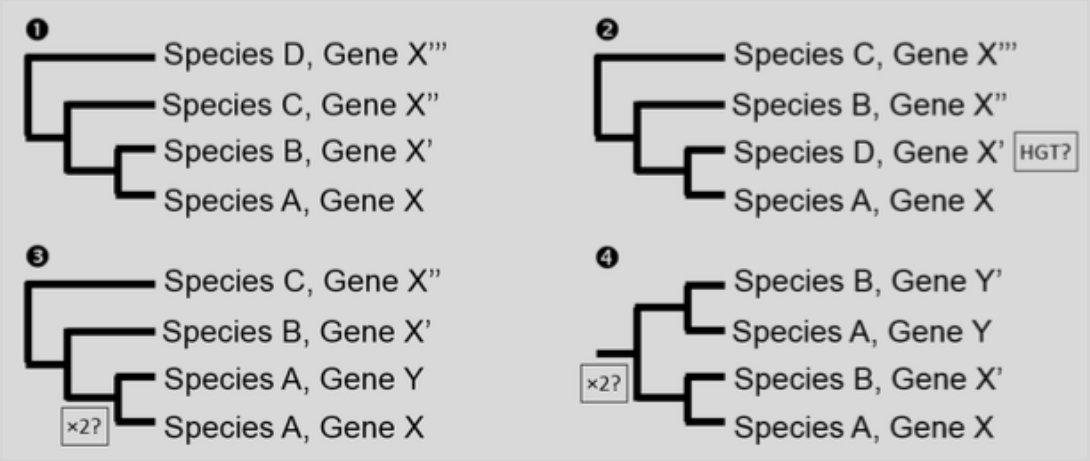
现在，我们来制作一个最大似然（ML）树。 如上所述，机器学习是最强大的系统发育方法之一，但不幸的是，它不像邻域连接那样容易执行。 MAGA免费提供了许多用于ML分析的良好应用程序。 我们将使用MEGA中的实现，尽管你可能还想熟悉可通过PHYLIP（系统发育推断包）获得的ML工具，它是一组功能强大且全面的免费系统发育应用程序。 PHYLIP可通过命令界面在大多数计算机平台上运行，但也可以通过许多公共可用的Web界面使用（请参见下面的“在何处获取？”）。

1. 启动MAGA并加载序列
2. 在“**Analysis > Phylogeny > Construct / Test Maximum Likelihood Tree…**”下将**Phylogeny Test** / Test of Phylogeny设置为Bootstrap，进行500次复制，并将所有其他设置保留为默认值（有关默认值，请参见附录2；但是增加线程数可能会加快分析速度）。 单击”**OK**”开始分析。
3. 等待的时间比使用“**Neighbour-Jointing**”分析的时间长得多。 如Box中所述，最大似然算法中使用了更多的计算。 进度条将显示您的分析完成的程度。 分析完成后，“树查看器”将打开并显示结果树。
4. 你可以像以前那样对树进行各种设置。
5. 支持系统发育分析的最强方法之一是使用至少两种独立的方法来执行它。 例如，如果使用邻居连接和ML获得相同的基本拓扑，则您很有理由相信您的分析是正确的。 几乎每个分子进化杂志都要求使用多种方法进行所有系统发育分析。

Box7. Interpreting Phylogenetic Analyses

如何解释系统发育分析

系统发育分析对于确定目标基因的进化历史非常有力。请考虑以下四种情况，并假定节点都具有良好的（高级别）引导程序支持。基因X，X’，X”，X”’是彼此的直系同源物，而基因Y和Y’与这些基因同源。假设物种A，B，C和D沿字母的距离越远，它们之间的距离就越远。在场景中，该基因落入其预期进化枝之内，即基因树反映了物种树。这里没有惊喜！在方案2中，相对于物种树，该基因不会落入我们期望的位置。该基因可能是通过水平转移事件HGT？来与该基因分组的物种（或密切相关的物种）产生联系。在方案3中，该基因来自的物种中存在一个旁系同源物，但其他物种中没有任何旁系同源物。假设已经对其他物种的基因组进行了测序，并且我们的E值阈值不太严格，这可能表明我们感兴趣的基因在物种中发生了部分或整个基因组复制事件，表示为×2？。在方案4中，该基因的旁系同源物在其他物种中也具有同系物。同样，在良好的基因组覆盖范围和适当的E值截止值的警告下，这可能表明复制事件发生在两个物种的祖先中，由×2？表示。



在何处获取?

MEGA X软件：https://www.megasoftware.net/

PHYLIP：https://evolution.genetics.washington.edu/phylip.html

Web-based ML analysis: http://bar.utoronto.ca/webphylip/