实验 6 演化树构建

（4 学时）

一、实验背景

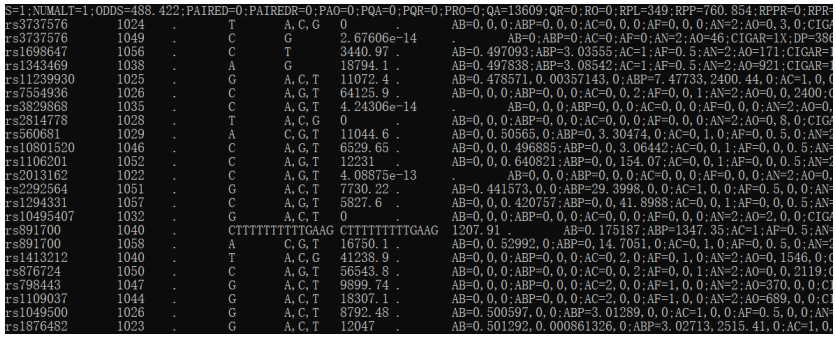
演化树，也被称为"生命之树"，表征地球生命的演化历程。它向我们展示，哪些生物之间的遗传关系更密切（比如，狗和狼，人类和黑猩猩），或者哪些生物之间的遗传关系更疏远。在本节中，我们将学习根据给定物种的部分 DNA 序列或变异位点来构建演化树的方法。

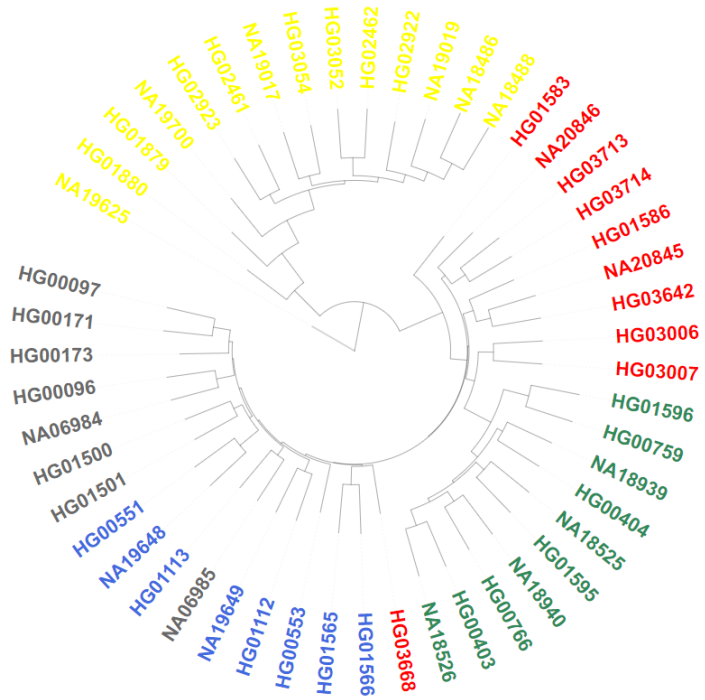
二、教学目标

如何通过构建演化树推断演化关系？如，我与同学、不同地理位置的人群、人类与其他物种等。我们将从自己的数据出发，学习在不同层次上整合数据集构建具有不同生物学意义和功能的演化树。

本节课程，我们将个体变异位点与千人基因组数据集或其他多个样本的数据集进行整合，通过以下实验流程，构建演化树，并完成可视化和可靠性评价：

数据准备、序列比对、计算推断演化树、树的评价、可视化展示

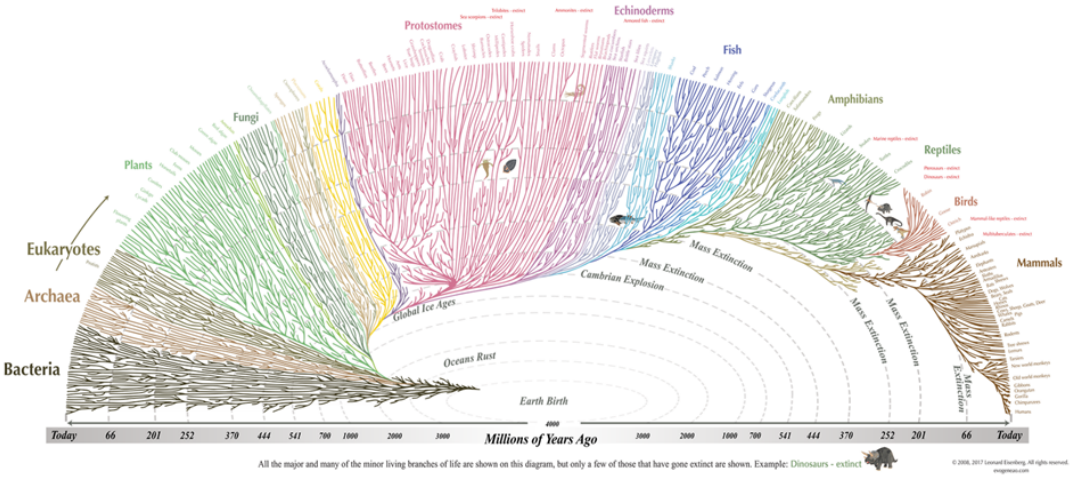




三、实验原理

第一部分 什么是演化树？

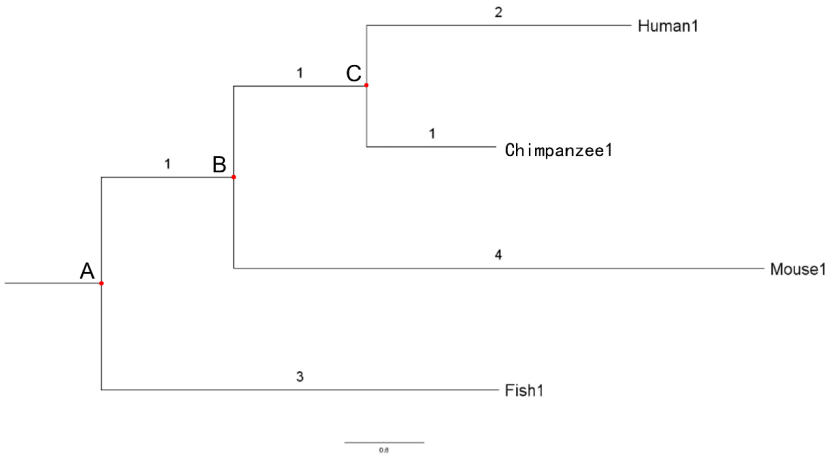
来自形态学、生物化学和基因组学的证据表明，地球上所有生物都有系统发生关系，我们可以用一棵巨大的演化树来表示这些关系。



在演化树当中，我们所研究的特定物种的部分 DNA 序列被表示为演化树当中的不同要素，包括叶子节点（只有一条边的节点/没有子节点的节点），内部节点（有多条边的节点/有子节点的节点）和分支，以此来描述序列之间的演化关系。多数情况下，构建演化树所用的 DNA 序列是来自不同生物体的基因序列，能用于推断生物体的演化历程。

解构一棵最简单的演化树

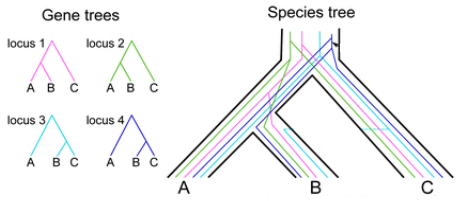
下图中有 4 条序列，分别标注为 Human1、Chimpanzee1、Mouse1 和 Fish1，依次来自于人类、黑猩猩、小鼠和鱼，代表 4 个将葡萄糖转化为能量的同源或等效基因。由这 4 个基因推断产生的演化树揭示了这 4 个基因是如何演化的：



首先，由共同的祖先基因 A 分裂或演化为 2 个不同的基因，一个是今天的 Fish1 基因，另一个代表现存老鼠、黑猩猩和人类的共同祖先基因 B。此后，基因 B 演化产生 2 个分支，一个是现存的 Mouse1 基因，另一个是现存黑猩猩和人类的共同祖先基因 C。最后，基因 C 演化分歧产生现存的 Chimpanzee1 和 Human1。分支长度显示了 4 个基因的相对演化距离，例如，在共同祖先 C 发生演化分歧后，Human1 基因的演化距离是 Chimpanzee1 基因的 2 倍。如果要衡量 2 个基因序列之间的演化距离，则可以穿越一个序列到达另一个序列，将所经历的分支长度相加。例如，Fish1 和 Human1 之间的演化距离 distance(Fish1,Human1)=3+1+1+2=7。

基因树与物种树

基于单个同源基因差异构建的演化树应称为基因树（gene tree），由多个基因构建的能代表多个物种演化关系的树，则称为物种树（species tree）。以上这样一棵由 4 个基因序列推断所得的基因树只显示了基因之间的演化关系，不一定能代表这 4 个物种的演化关系。由于演化发生在生物种群水平，而非个体水平，但基因分化的发生通常会早于新物种的种群分离，因此，当只考虑一个基因的时候，个体可能表现出与其他物种（非该个体所属的物种）的成员更近。基因分化事件常常在物种形成前或在物种形成后发生，并不与物种分歧时间完全重叠。因此，在研究物种演化分歧时，对构树数据的选择尤为重要。



a. Based on the figure above, which gene tree(s) disagree with the species tree?

b. What factor(s) might result in this discrepancy between gene and species trees?

c. What term describes this discrepancy?

第二部分 演化树构建原理与常用方法

构建演化树可选择的方法众多，主要可分为两类：基于离散特征的方法，和基于数据的方法。

基于离散特征的方法，总体思路是通过搜索各种可能的树，从中选择一个最能解释物种、不同个体或不同基因之间演化关系的演化树。这一类方法可实现算法众多，比较常见的有最大简约法、最大似然法和贝叶斯法。

基于数据的方法，其核心是最小演化原理，通过构造一个距离矩阵，用来表示两个物种之间的演化距离，再根据距离矩阵，采用聚类法对研究序列进行分类。这一类方法包含非加权组平均法（UPGMA）、邻接法（NJ）、距离变换法等。本实验选择邻接法进行构树。

邻接法核心算法

邻接法是最早的基于距离计算和推断演化树的算法，该方法计算速度快，准确度高，但也有局限性，更适用于参与构树的物种、个体或基因数量较少的情况，当参与计算的个体或序列数量过多时，该方法计算量大增，只能推断得到近似的最优树，无法找到精确的最小演化树。

接下来我们举例说明采用邻接法构树的过程：

假设有以下 5 组同源序列，每条序列有 50 个碱基，每两个序列之间的错配碱基只考虑替换，暂时忽略插入、缺失这一类变异类型：

S1: GTGCTGCACGGCTCAGTATAGCATTTA

CCCTTCCATCTTCAGATCCTGAA

S2: ACGCTGCACGGCTCAGTGCGGTGCTTA

CCCTCCCATCTTCAGATCCTGAA

S3: GTGCTGCACGGCTCGGCGCAGCATTTAC

CCTCCCATCTTCAGATCCTATC

S4: GTATCACACGACTCAGCGCAGCATTTGC

CCTCCCGTCTTCAGATCCTAAA

S5: GTATCACATAGCTCAGCGCAGCATTTG

CCCTCCCGTCTTCAGATCTAAAA

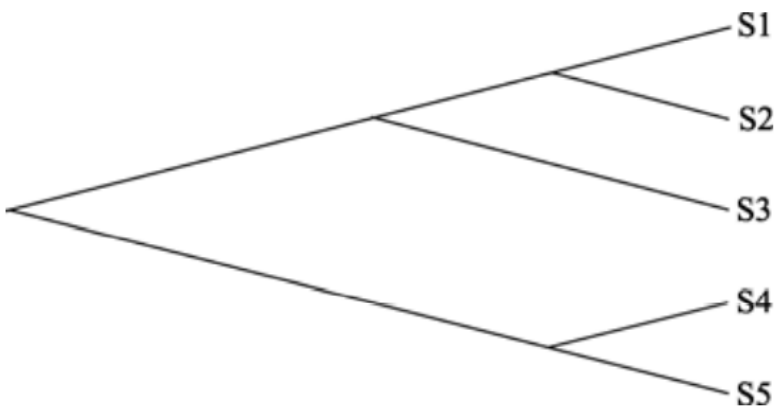
通过计算错配碱基数，可得到如下距离矩阵：

序列 Sequence	*S1*	*S2*	*S3*	*S4*
S2	9			
S3	8	11		
S4	12	15	10	
S5	15	18	13	5

通过依次将距离小的序列进行聚类合并，最终得到以下距离矩阵：

序列 Sequence	*S12*	*S3*
S3	9.5	
S45	15	11.5

根据该矩阵可以得到以下拓扑结构：



即是其中一棵推断树。

第三部分 本实验构建演化树的流程

数据准备

构建演化树的目的是为了描述不同物种之间、不同个体之间或不同基因之间的演化关系，我们需要根据研究的问题，选择相应的数据，通过下载、筛选、整合等步骤，将同源 DNA 序列数据按特定格式进行整理，备后续分析使用。本实验中，我们需从千人数据集中挑选本实验关注的变异位点，与个体变异位点数据进行整合。

序列比对

为保证序列的同源性和所得演化树的可靠性，需要对数据进行比对和校正。在本实验中，我们将对数据进行格式转换和比对处理。

计算推断演化树

进行演化树计算和推断的方法众多，在本实验中，我们选择邻接法（Neighbor-Joining，简称 NJ），采用 Phylip 构树软件，从所有可能的演化树中选择所有演化分支长度总和最小的那棵树作为实验结果。

树的评价

演化树评价的目的，是判断所构建演化树的置信度，常见方法有自举检验法及刀切法。

自举法（Bootstrap），即放回式抽样统计法。在选取序列片段长度一定的条件下，从原始序列中随机选取碱基组成，使得在所有参与构树的样本中总是有一部分重复的碱基，又有一部分新增的碱基，按照这样的方法将获得不同的数据集，构建得到多个演化树，将不同的演化树进行比较，所有树当中，不同分支都将进行置信度评估，置信度高的分支被认为是可靠的分支，含可靠分支最多的树被认为是最优的树。

在本实验中，我们可通过优化设置 Phylip 构树软件的自举法参数来优化构树结果。

演化树可视化与图形优化

目前有很多软件包可进行演化树可视化展示，我们选择用户评价较好的 iTOL 作为本实验教学的可视化工具。

iTOL 可快捷实现在线数据导入、数据集注释、分支上色和树形结构快速选择等功能。可将 Phylip 计算推断的拓扑树转换为形式多样、视觉美观的树形结构。

〖**教学重点**〗在本实验环节中，学生需理解演化树的意义和功能，根据自己感兴趣的问题选择合适的变异位点数据集，掌握使用 Phylip 软件构树的关键步骤和参数选择依据，构建演化树，并对分析结果进行可视化优化、生物学意义解释和可靠性评价。

四、软件安装与数据准备

phylip 软件 已安装

五、实验步骤

第一部分 数据格式转换

将变异检测实验课程所得 VCF 文件转换为 phylip 格式：

```
$ python ./06_tree/vcf2phylip.py -i ./06_tree/Merge.51samples.SNP.vcf
$ less Merge.51samples.SNP.min4.phy
```

第二部分 计算与推断演化树

应用安装的以下4个程序进行构树：

seqboot、dnadist、neighbor、consense
注意：
四个子程序要按照列出的顺序来运行；
每个子程序的默认输出名字是“outfile”，因此在进行下一步分析前，需要对上一步的 outfile 进行重命名，以防造成混淆。
这四个子程序都是交互式的，即需要用户一步步的输入/修改参数，为了流程式地运行子程序，我们提前把待输入/修改的参数写在.par 文件 中。

1、进入工作目录，建立本次工作目录及复制相关文件

```
$ cd /home/LY && mkdir 06_tree
$ cp /home/50samples-HG01880-01.vcf /home/merge.py /home/vcf2phylip.py ./06_tree/
```

2、合并示例样本和自己个人样本的 vcf： python merge.py 自己样本 vcf 样本名称 >输出文件名

```
$ python ./06_tree/merge.py ./03_SNPCalling/test.clean.SNP.vcf test >./06_tree/Merge.51samples.SNP.vcf
```

3、进入目录生成随机数据集。

```
$ cd 06_tree
$ cp ../Merge.51samples.SNP.min4.phy ./
$ /home/ecoli/2024fuda/06_Tree/phylip-3.697/exe/seqboot #然后输入下列参数

Merge.51samples.SNP.min4.phy  #本程序的输入文件（本行的注释内容不要写到脚本中，否则会报错）
R #选择 bootstrap
100 #设置 bootstrap 的值，即重复的 replicate 的数目，通常使用 1000 或者 100，注意此处设定好后，后续两步的 M 值也为 1000 或者 100
Y #yes 确认以上设定的参数
9 #设定随机参数，输入奇数值。


#重命名文件
$ mv ./outfile ./seqboot.out
```

4、计算距离矩阵。

```
$ /home/ecoli/2024fuda/06_Tree/phylip-3.697/exe/dnadist
seqboot.out #本程序的输入文件 （本行的注释内容不要写到脚本中，否则会报错）
T  #选择设定 Transition/transversion 的比值
```

```
2.3628  #比值大小
M  #修改 M 值
D  #修改 M 值
100  #设定 M 值大小
2  #将软件运行情况显示出来
Y  #确认以上设定的参数
```

#重命名文件

```
$ mv ./outfile ./dnadist.out
```

5、设置 NJ 法输入参数

```
$ /home/ecoli/2024fuda/06_Tree/phylip-3.697/exe/neighbor
```

```
dnadist.out  #本程序的输入文件 （本行的注释内容不要写到脚本中，否则会报错）
M
100  #设定 M 值大小
9  #设定随机数，输入奇数值
Y  #确认以上设定的参数
```

重命名输出数据文件

```
$ mv ./outfile ./nei.out && mv ./outtree ./nei.tree
```

6、从多重树中选择最优树。

```
$ /home/ecoli/2024fuda/06_Tree/phylip-3.697/exe/consense

nei.tree  #本程序的输入文件 （本行的注释内容不要写到脚本中，否则会报错）
Y #确认以上设定的参数
```

运行以上脚本，重命名输出数据文件

```
$ mv ./outfile ./cons.out && mv ./outtree ./constree
```

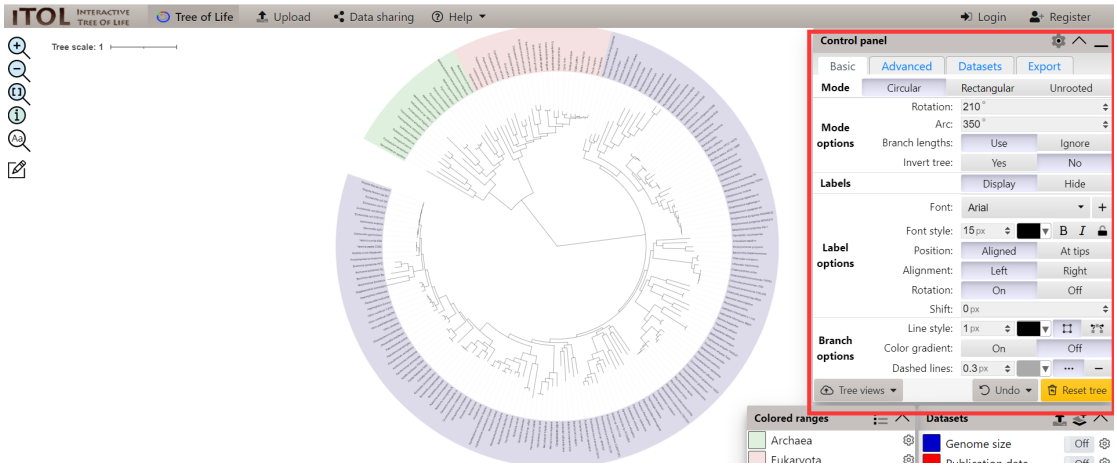
至此，采用 Phylip 软件构树的分析过程就已完成。下一步演化树可视化展示。

六、实验后处理和预期结果

采用 iTOL 进行可视化与图形优化

iTOL 在线工具画图界面网址：

https://itol.embl.de/itol.cgi



使用右侧工具栏，可快捷完成拓扑数据导入、数据集注释添加和色块选择。可根据个人喜好，自主调节演化树类型和配色方案。