分子系统发育分析
Molecular Phylogenetics

杨茜
北京大学生命科学学院
2011-01-09

提纲

1 系统发育的基本概念

2 系统发育树的构建方法

3 系统发育分析的软件

4 案例：分析*NADH1*序列

5 参考和推荐书目

系统发育的相关概念

系统发育学	研究物种进化和系统分类
系统发育分析	推断或评估这些进化和分类关系
系统发育树	通过系统发育分析所推断出来的进化关系一般用分支图来描述

系统发育 (Phylogeny)

系统发育分析是研究物种进化和系统分类的一种方法，其常用一种类似树状分支的图形来概括各种（类）生物之间的亲缘关系，这种树状分支的图形称为系统发育树。

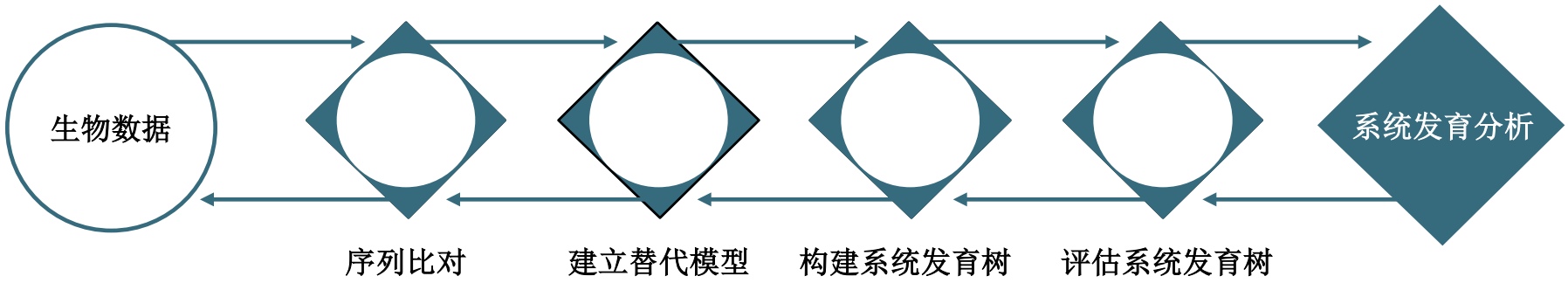
系统发育树描述了同一谱系的进化关系，包括了分子进化、物种进化以及分子进化和物种进化的综合。通过系统发育树，我们可以找到亲缘关系最近的物种或者基因，探索基因的功能，追溯基因的起源。

在现代系统发育学研究中，研究的重点已经不再是生物的形态学特征或者其他特性，而主要是序列信息。

虽然已有了大量的用于系统发育的算法、过程和计算机程序，但是这些方法的可靠性和实用性还是依赖于数据的结构和大小。

系统发育分析一般没有实验基础，因为系统发育的发生过程都是已经完成的历史，只能去推断或者评估，而无法重现。

系统发育分析的概念性步骤



选择序列时的注意事项

1. 序列有指定的来源并且正确无误
2. 序列是同源的，即所有的序列都起源于同一祖先序列
3. 样本序列之间的差异包含了足以解决感兴趣的问题的信息位点。
4. 样本序列是随机进化的。
5. 序列中的每一个位点的进化都是独立的。

序列比对

建立一个序列比对的基本步骤包括：选择合适的比对程序；然后从比对结果中提取数据。至于如何提取有效数据，取决于所选择的建树程序如何处理容易引起歧义的比对区域和插入/删除序列（即所谓的indel或者gap）。

分析DNA序列的方法基本上仍然是通过碱基和密码子的替代来考察序列的差异；这个方法同样应用于对蛋白质序列的分析，但是由于氨基酸的生物化学多样性，我们必须引入更多的参数。

从比对中提取数据

如果比对中出现可变长度，我们通常会根据比对的不确定性程度和处理indel状态的原则这两个标准对比对结果进行取舍，从中选择所需的数据；其中针对indel状态的处理方法取决于建树方法以及从比对结果中发掘出的信息，最极端的方法是把包括空位在内的所有indel位点从比对中清除出去，在分析时不加考虑，这个方法的好处是可以把序列的变化包容在取代模型中，而不需要特别的模型来处理indel状态，但是它的缺点也很明显：indel区域的信息完全被忽略了。

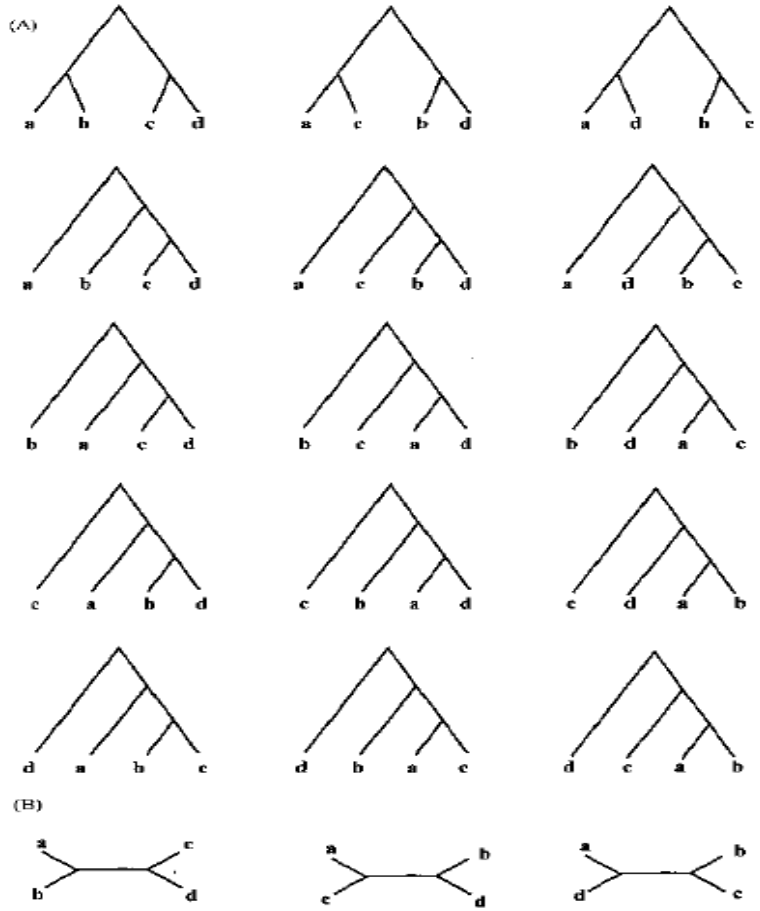
系统发育树的种类

1. 有根树和无根树
2. 基因树和物种树
3. 期望树和现实树

有根树和无根树

- 有根树是具有方向的树。包含唯一的根节点，将其作为树中所有物种的最近共同祖先。
- 无根树是没有方向的，其中线段的两个演化方向都有可能。
- 如果类群数(m)为4，就有15种可能的有根树拓扑结构和3种无根树拓扑结构。
- 可能的拓扑结构随m的增加而迅速增加，这些拓扑结构中只有一种是真实树。

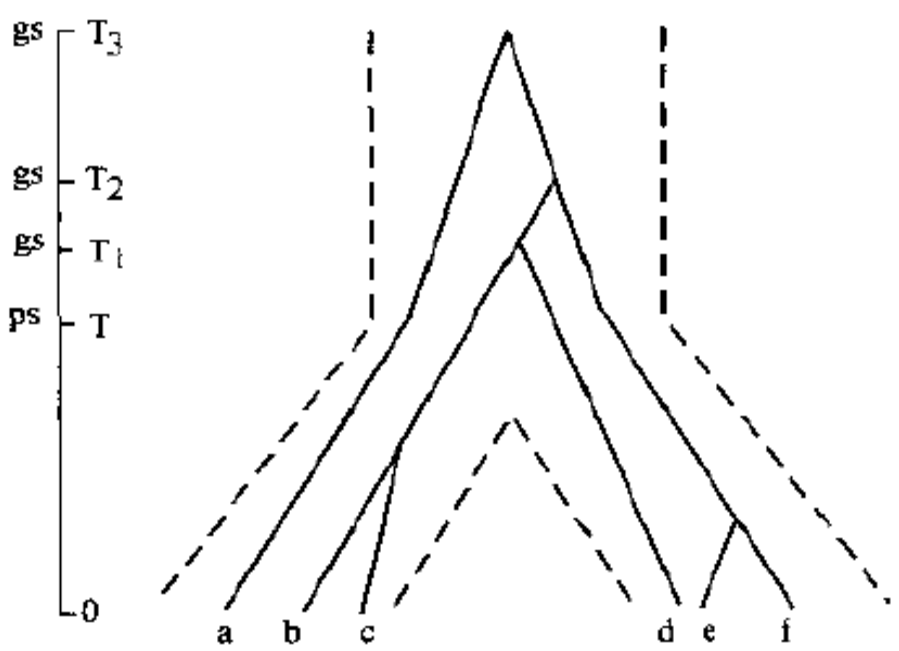
#species	Unrooted	Rooted
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$

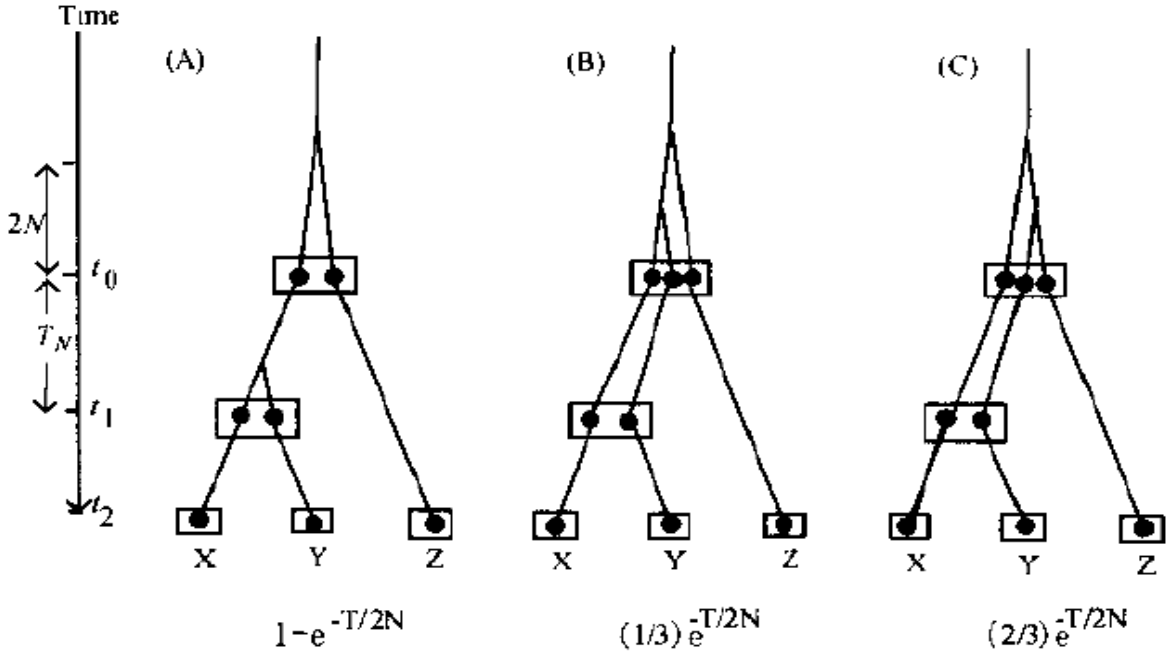


物种树和基因树

代表一个物种或群体进化历史的系统发育树被称为物种树。根据基因构建的树称为基因树。

基因树可能不同于物种树。

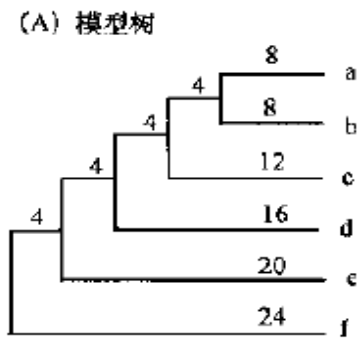


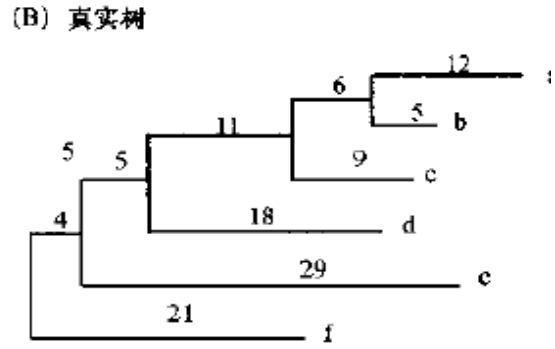


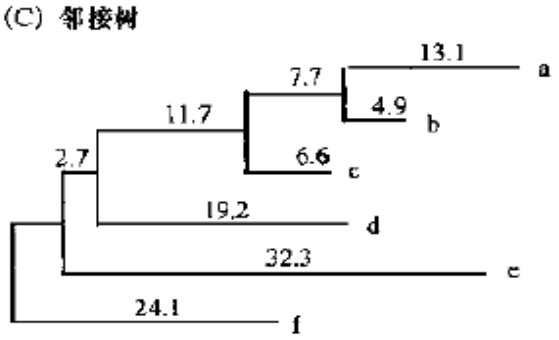
期望树与真实树

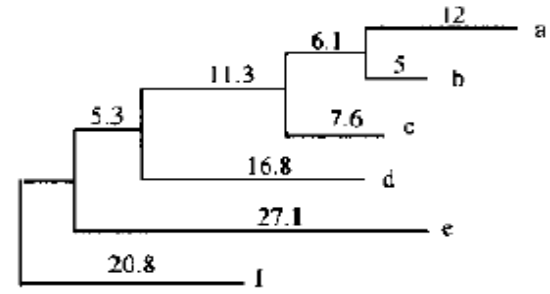
一个用无限长的序列或每一分支的替代树构建成的树称为期望树。而建立在实际替代数基础上的树称为真实树。

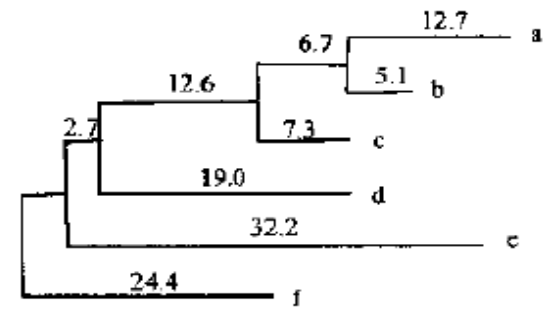
要注意的是，期望树和真实树通常不同于由所观察到的序列数据重建的树，即重建树或推论树。由于基因的进化改变受限于随机误差和某些自然选择因素，即使由很多基因构建的树也可能不同于真实树。

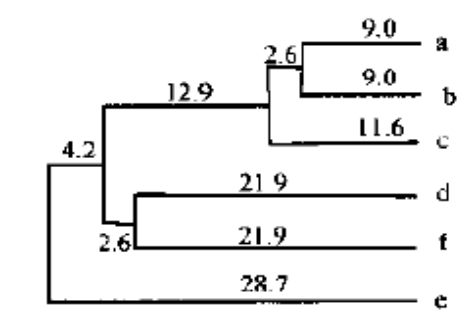


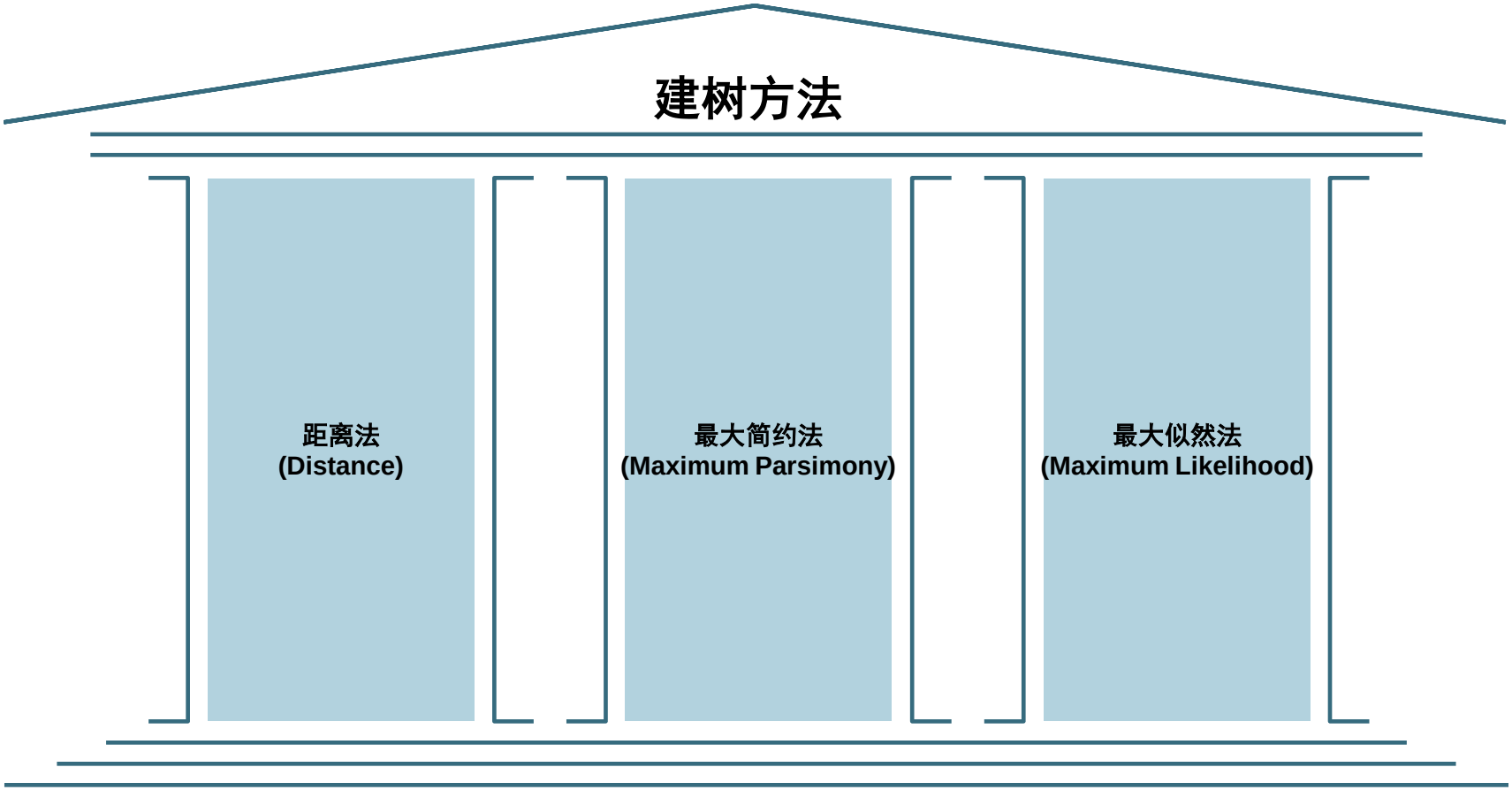











建树方法
距离法
(Distance)
最大简约法
(Maximum Parsimony)
最大似然法
(Maximum Likelihood)

距离法 (Distance)

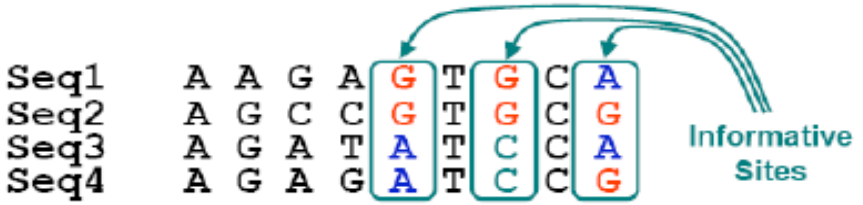
运用距离法或距离矩阵法时，系统发育树的构建基于所有类群间的进化距离值的关系。根据所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度

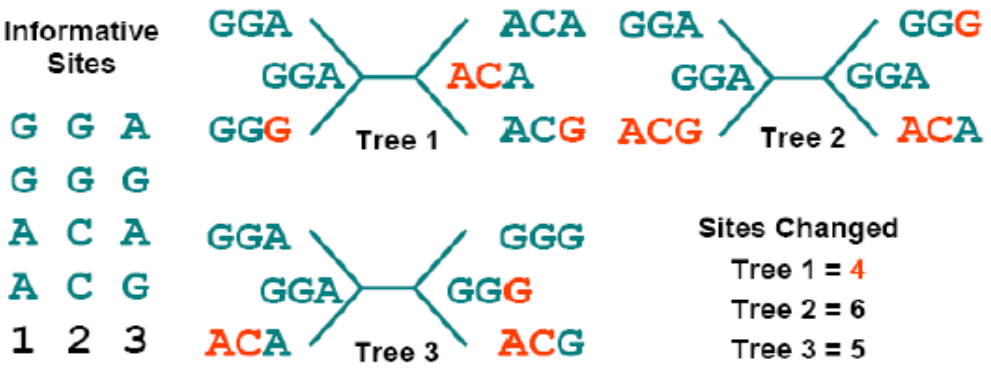
- 使用算术平均的不加权的组对法 (UPGMA)
- 最小进化法 (Minimal Evolution)
- 邻接法 (Neighbor-joining)

最大简约法 (Maximum Parsimony)

最大简约法根据序列的多重比对结果，对所有可能正确的拓扑结构进行计算并挑选出所需替代数最少的拓扑结构作为最优树，即能够利用最少的步骤去解释多重比对中的碱基差异。理论基础是解释一个过程最好的理论是所需假设数目最少的那一个。前提是要选择信息位点。

- 加权MP法 （转换和颠换）
- 不加权MP法

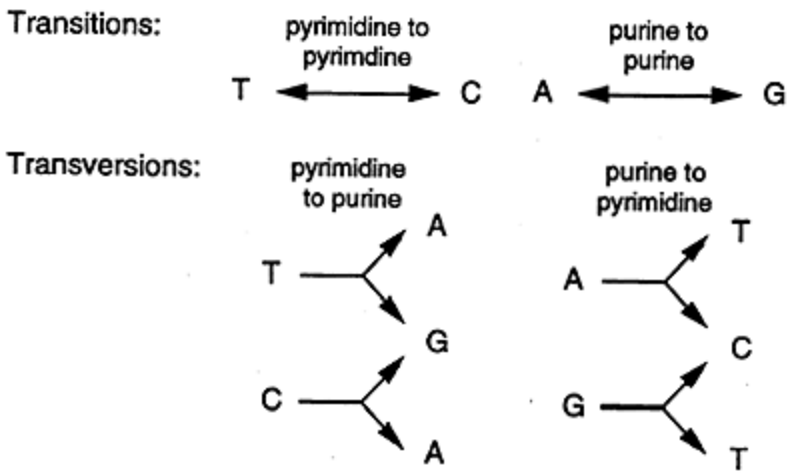




最大似然法 (Maximum Likelihood)

最大似然法以一个特定的替代模型分析一组序列数据的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，使所获得的每一个拓扑结构的似然率均为最大，挑选似然率最大的拓扑结构作为最优树。

建树过程费时，计算量大，每个步骤都要考虑内部节点的所有可能性。前提是要选择合理并正确的替代模型。



核苷酸替换模型

JC69

$$\begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}$$

K80

$$\begin{bmatrix} . & \alpha & \beta & \beta \\ \alpha & . & \beta & \beta \\ \beta & \beta & . & \alpha \\ \beta & \beta & \alpha & . \end{bmatrix}$$

HKY85

$$\begin{bmatrix} . & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & . & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & . & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & . \end{bmatrix}$$

系统发育树可靠性的检验

自展法 (bootstrap)是对所比较序列上的替换位点作多次随机取样，根据每次取样的数据可以得到新的树形图，相同的组合出现在某一个节点上的次数占总取样次数的百分比就是该节点的bootstrap值。

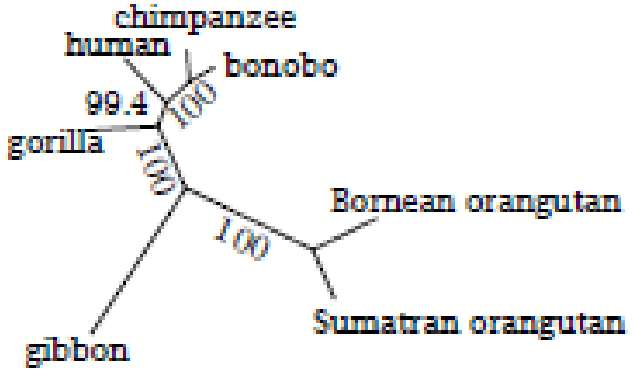
>95%统计学上有意义，>70%可信，<50%不可信。

Original alignment

site	1	2	3	4	5	6	7	8	9
10									
human	N	E	N	L	F	A	S	F	I	A
chimpanzee	N	E	N	L	F	A	S	F	A	A
bonobo	N	E	N	L	F	A	S	F	A	A
gorilla	N	E	N	L	F	A	S	F	I	A
orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
gibbon	N	E	N	L	F	T	S	F	A	T

One bootstrap pseudo-sample

site	2	4	1	9	5	8	9	1	3	7
human	E	L	N	I	F	F	I	N	N	S
chimpanzee	E	L	N	A	F	F	A	N	N	S
bonobo	E	L	N	A	F	F	A	N	N	S
gorilla	E	L	N	I	F	F	I	N	N	S
orangutan	E	L	N	T	F	F	T	N	D	P
Sumatran	E	L	N	I	F	F	I	N	S	P
gibbon	E	L	N	A	F	F	A	N	N	S



(Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791)

不同建树方法的优缺点

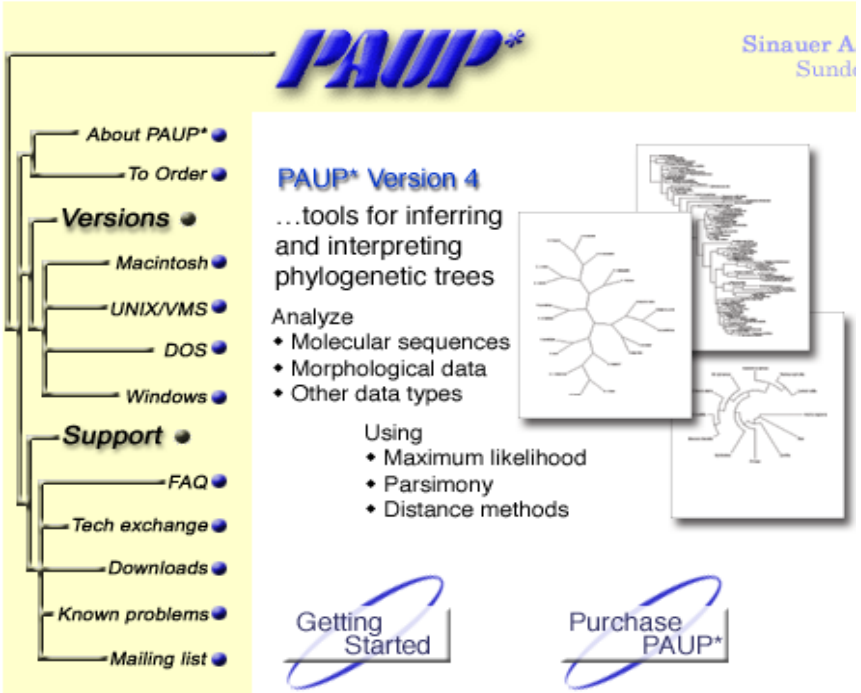
名称	基本特征	适用范围	优点	缺点
邻接法	不需要分子钟假设，是基于最小进化原理，进行类的合并时，不仅要求待合并的类是相近的，而且要求待合并的类远离其他的类。	远缘序列，进化距离不大，信息位点少的短序列	假设少，树的构建相对准确，，计算速度快，只得一颗树，可以分析较多的序列，运行速度优于最大简约法	序列上的所有位点等同对待，且所分析的序列的进化距离不能太大
最大简约法	基于进化过程中碱基替代数目最少这一假说，不需要替代模型，对所有可能的拓扑结构进行计算，并计算出所需替代数最小的那个拓扑结构，作为最优树	近缘序列 物种序列的数目≤12	善于分析某些特殊的分子数据如插入、缺失等序列有用。	只适于序列数目N≤12。存在较多回复突变或平行突变时，结果较差。变异大的序列会出现长枝吸引而导致建树错误。
最大似然法	依赖于某一个特定的替代模型来分析给定的一组序列数据，使得获得的每一个拓扑结构的似然率都为最大值，然后再挑出其中似然率最大的拓扑结构作为最优树。	特定的替代的模型，远缘序列	很好的统计学基础，大样本时似然法可以获得参数统计的最小方差，在进化模型确定的情况下，ML法是与进化事实吻合最好的建树算法	所有可能的系统发育树都计算似然函数，计算量大，耗时时间长。依赖于合适的替代模型，

常用的建树软件

1. MEGA (Molecular Evolutionary Genetics Analysis)
2. PHYLIP (PHYLogeny Inference Package)
3. PAUP (Phylogenetic Analysis Using Parsimony)
4. PAML (Phylogenetic Analysis by Maximum Likelihood)

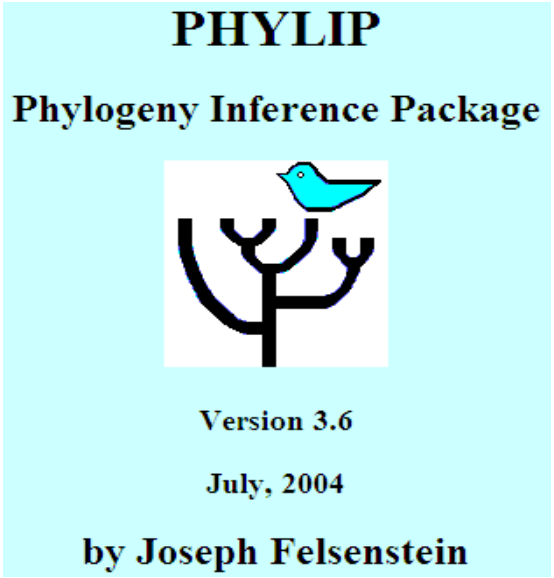
PAUP (Phylogenetic Analysis Using Parsimony)

PAUP是为系统发育分析提供一个简单的，带有菜单界面的、拥有多种功能（包括进化树图）的程序。PAUP 4.0可以针对核苷酸数据进行与距离方法和ML方法相关的分析功能。



PHYLIP (PHYLogeny Inference Package)

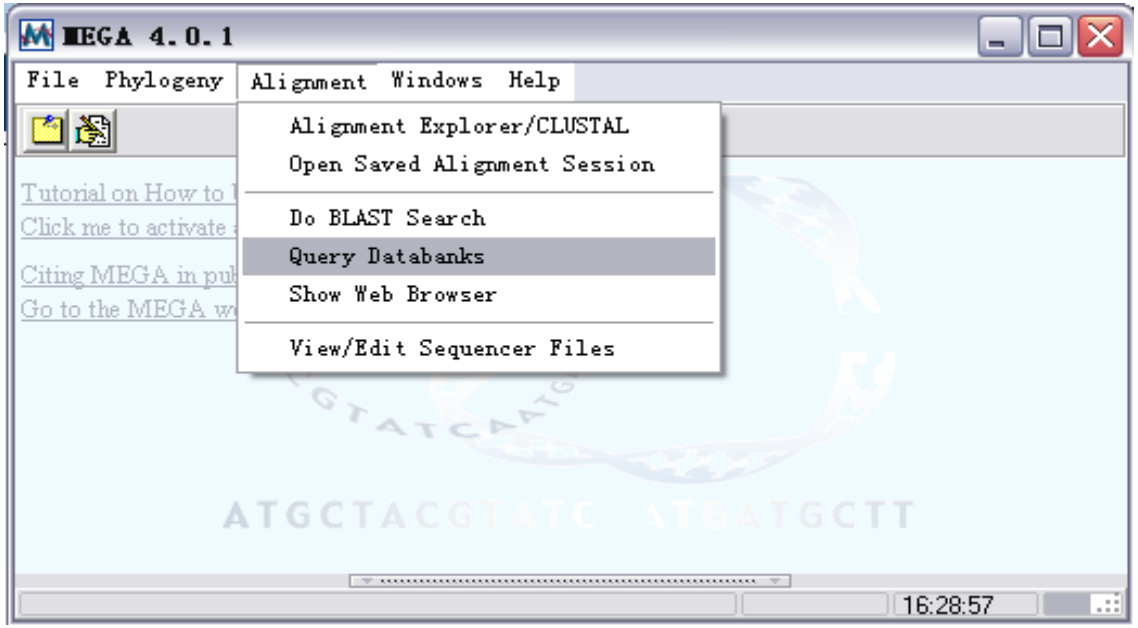
PHYLIP是一个包含了大约30个程序的软件包，这些程序基本上囊括了系统发育的所有方面。PHYLIP是最广泛使用的系统发育程序。PHYLIP是一个命令行程序，没有PAUP那样的鼠标点击的界面。软件的文档写得非常好，容易理解，命令行界面也很简明。



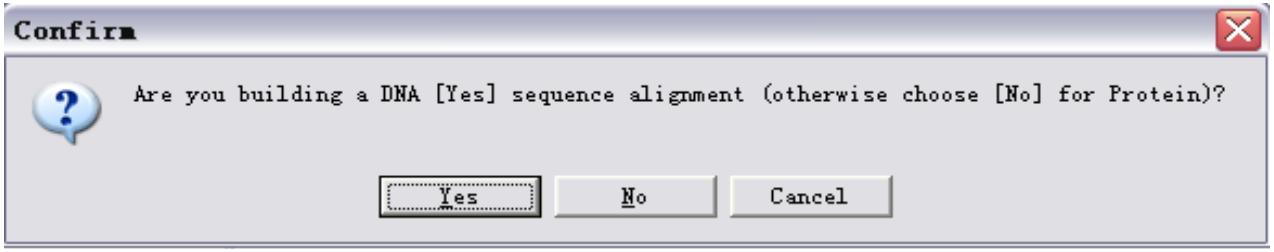
MEGA (Molecular Evolutionary Genetics Analysis)

MEGA是一个关于序列分析以及比较统计的软件包，其中包括有距离建树方法和MP建树方法。针对核苷酸数据建树，MEGA的效果不如PAUP或者PHYLIP。进化树图形很简单。虽然MEGA可以通过密码子数据和氨基酸数据建立距离进化树，但是使用的取代模型太简单，对于绝大多数数据而言，不能产生可靠的进化树。

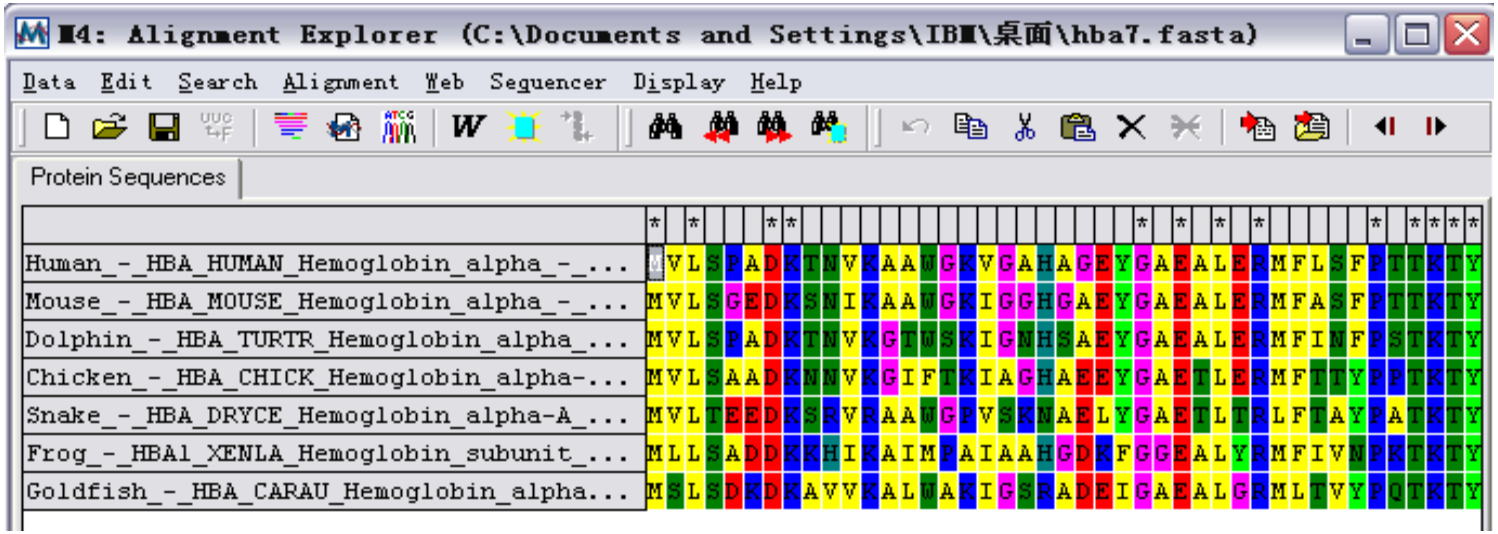
使用MEGA构建系统发育树
通过这个界面我们可以看到有四个条目，一个是MEGA的使用指南；二是打开数据文件；三是发表文章时要注明引用 MEGA；四是MEGA 的网站。我们可以通过使用指南熟悉MEGA的使用



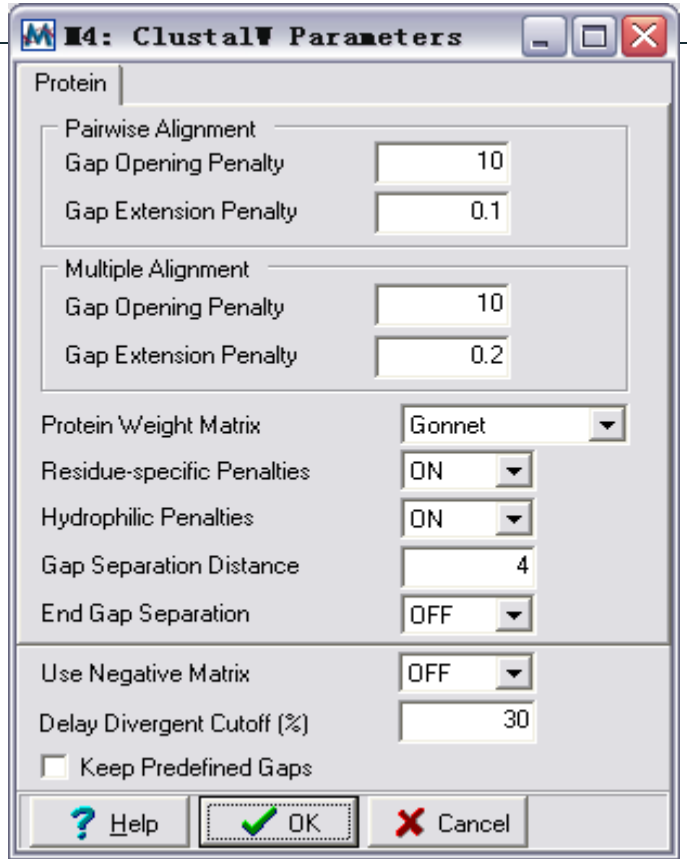
可以利用已有的数据，也可以利用MEGA直接通过NCBI直接查找所感兴趣的序列。



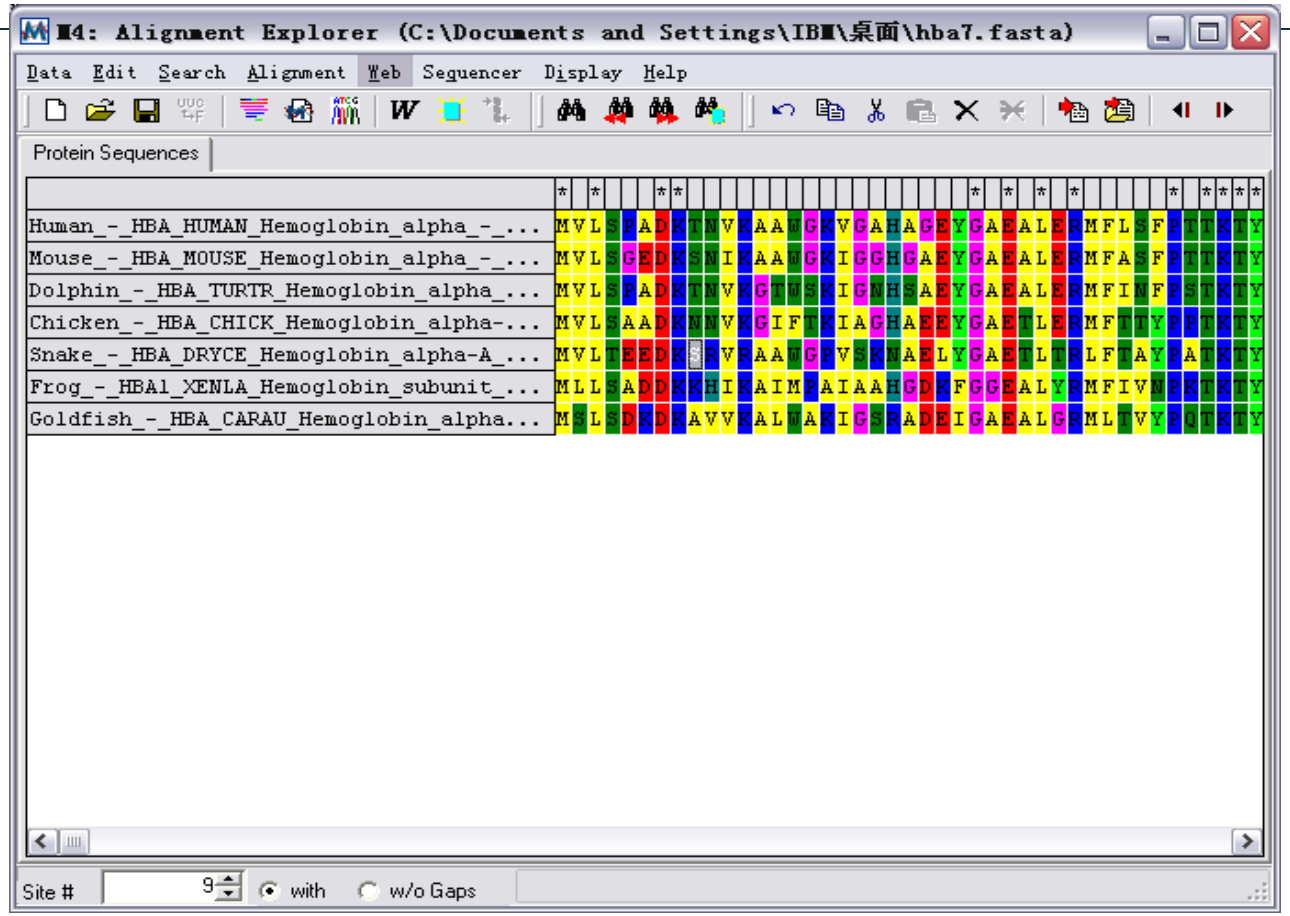
以七个物种的血红蛋白α亚基为例



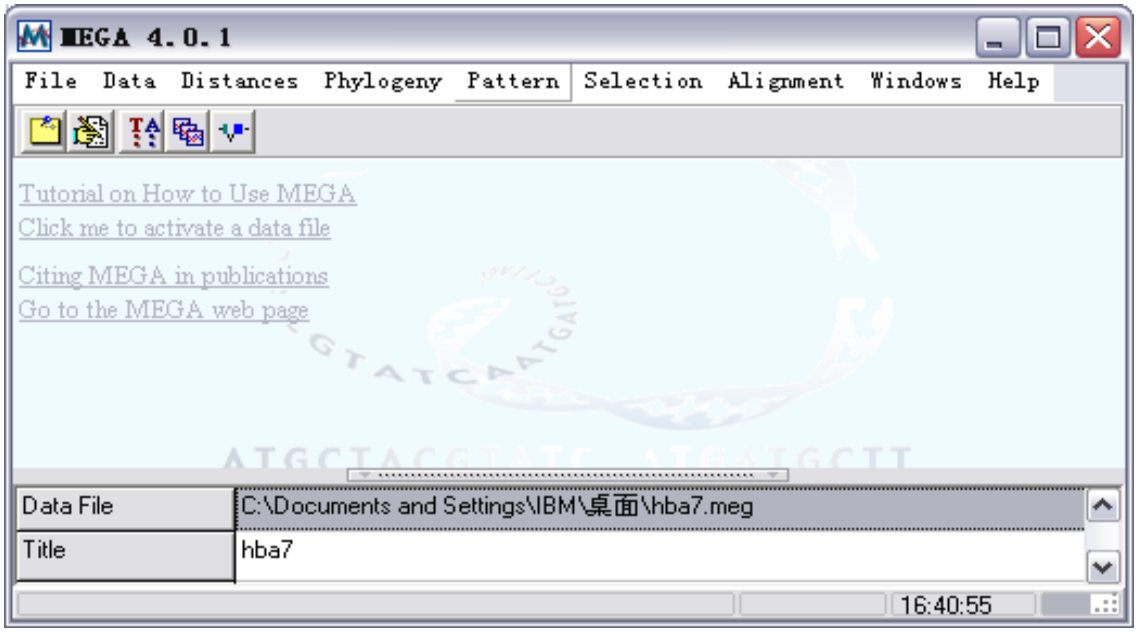
将我们的数据转化为FASTA格式， 导入Alignment explorer、 在Alignment explorer下还可以对序列进行编辑和插入、进行BLAST搜索等。



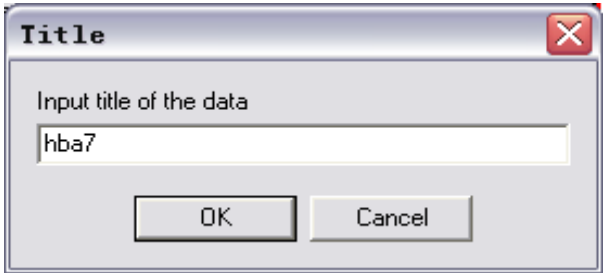
数据导入后将序列全部选中也可选一部分进行多序列比对，参数为默认值

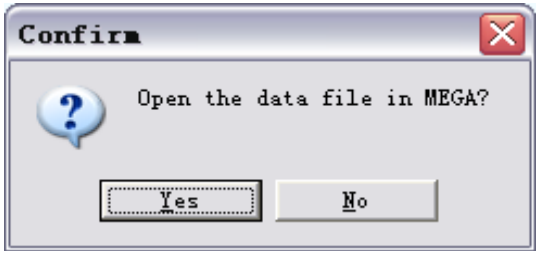


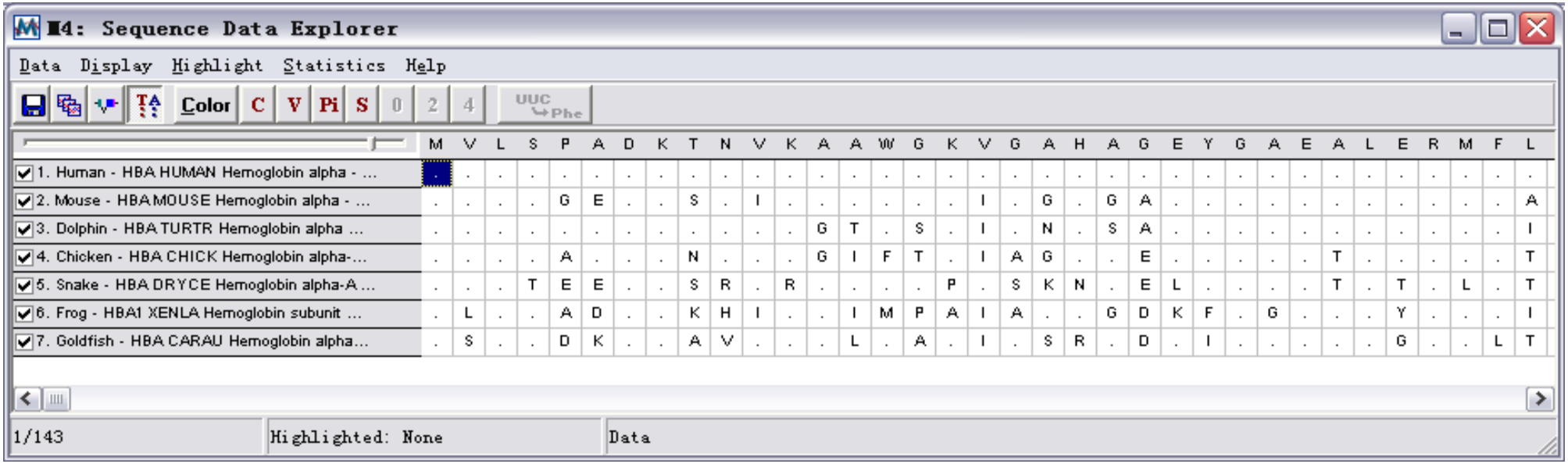
Data->Export alignment->MEGA format

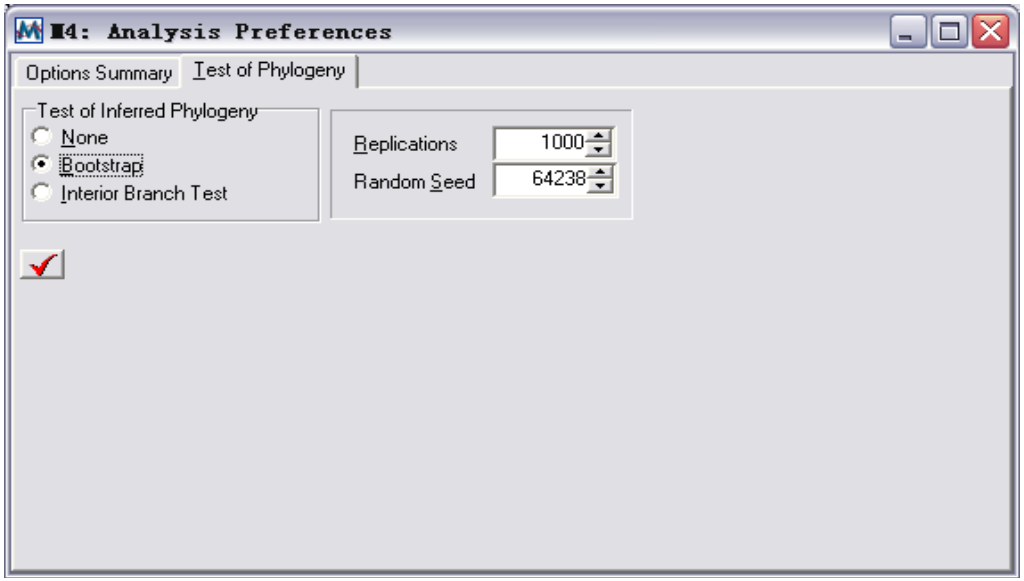


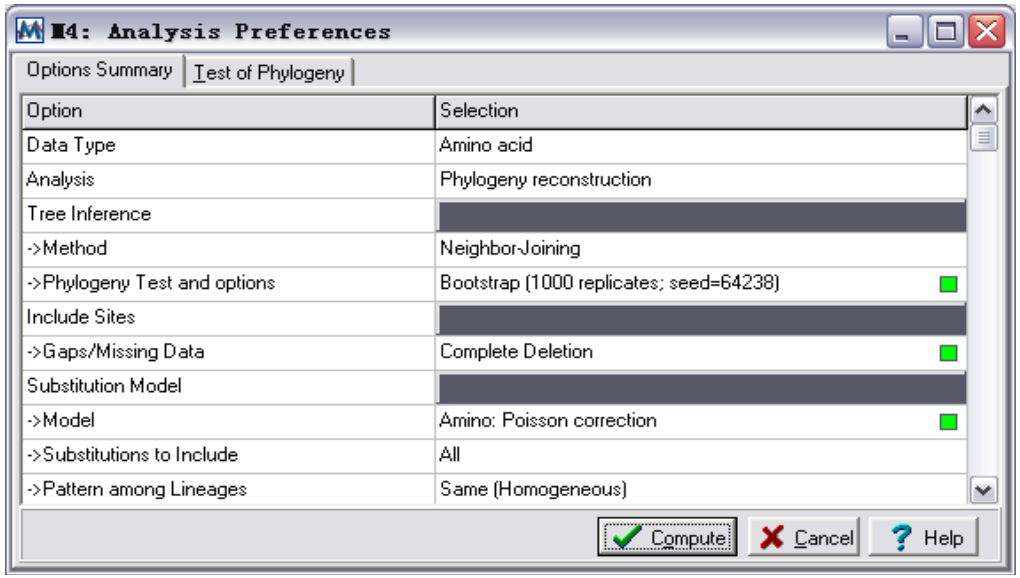
Phylogeny— Bootstrap Test Phylogeny— NJ 法


Title
Input title of the data
hba7
OK
Cancel

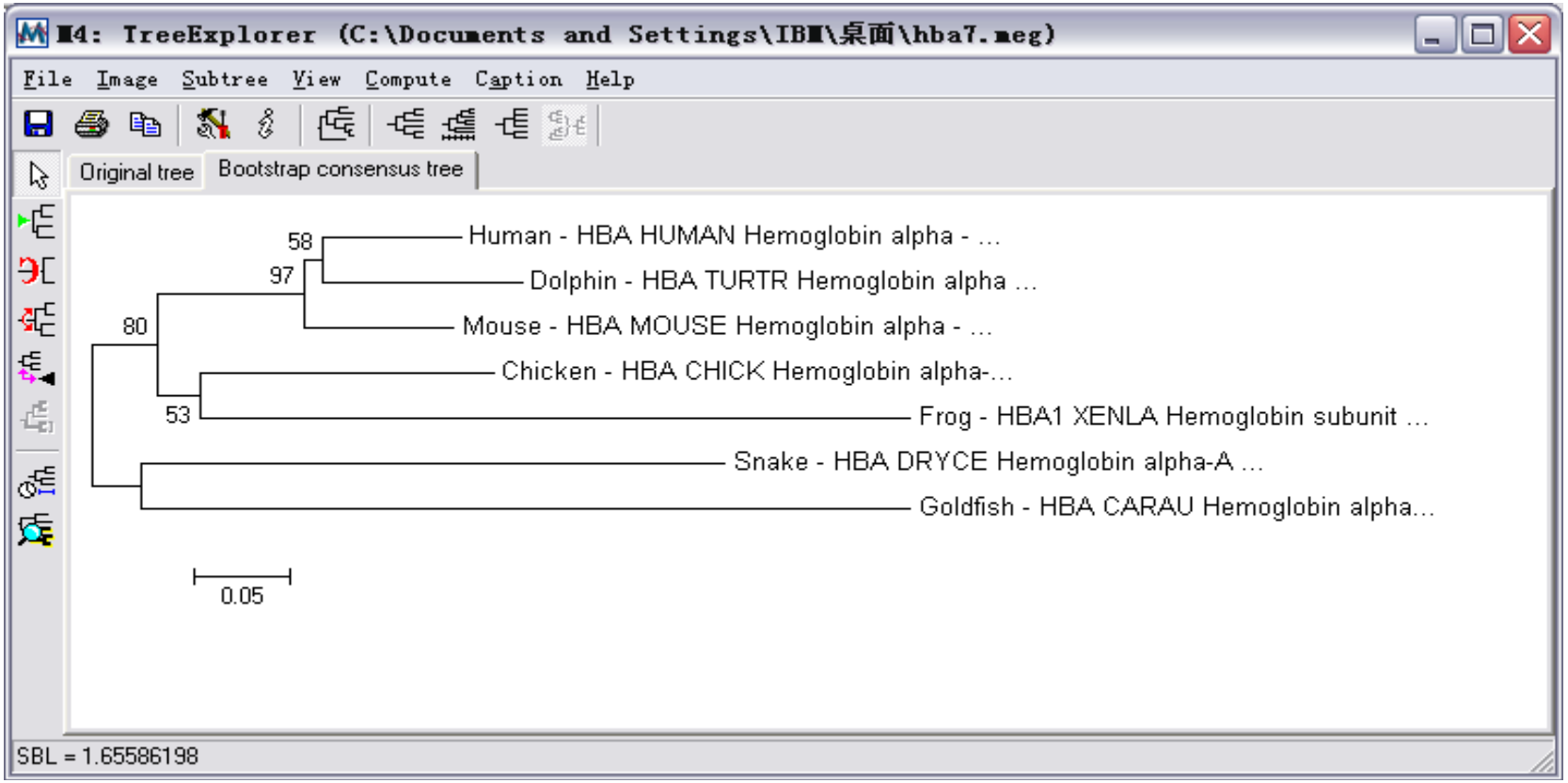

Confirm
Open the data file in MEGA?
Yes
No


M4: Sequence Data Explorer
Data Display Highlight Statistics Help
Color C V Pi S 0 2 4
UUC Phe
1. Human - HBA HUMAN Hemoglobin alpha - ...
2. Mouse - HBA MOUSE Hemoglobin alpha - ...
3. Dolphin - HBA TURTR Hemoglobin alpha ...
4. Chicken - HBA CHICK Hemoglobin alpha-...
5. Snake - HBA DRYCE Hemoglobin alpha-A ...
6. Frog - HBA1 XENLA Hemoglobin subunit ...
7. Goldfish - HBA CARAU Hemoglobin alpha...
1/143
Highlighted: None
Data





可以根据需要选择不同的模型，修改参数



结果将会出现两种树。一种为original tree,另一种为consense tree ,一般我们选择后者，即一致树。

PAML (Phylogenetic Analysis by Maximum Likelihood)

PAML是一个共享软件包，可以建立ML模型，模拟实验，进行基于ML进化树的分析，能够进行进化树评估以及数据和进化树的统计。对于密码子数据和氨基酸数据，提供了最详细的和最灵活的参数指定和评估方案。对于核苷酸数据替换模型的范围同PAUP的一样广泛，可能包括了所有值得考虑的模型。

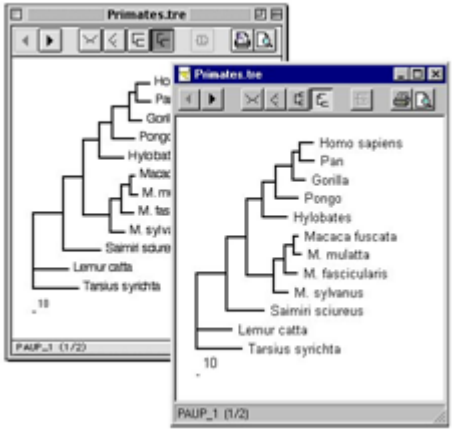
Phylogenetic Analysis by Maximum Likelihood (PAML)

Ziheng Yang

TreeView

这个软件可以读取标准的NEXUS和PHYLIP格式的系统发育树文件，允许用户重新定义树根和其它一些简单的节点，系统发育树可以打印或者保存在一个文件中以备处理。

TreeView



Phylogenetic Analysis of *NADH1* Sequences from 8 Primate Species

Common name	Binomial	Genbank accession #
Human	*Homo sapiens*	X93334
common chimpanzee	*Pan troglodytes*	X93335
pygmy chimpanzee	*Pan paniscus*	D38116
Gorilla	*Gorilla gorilla*	X93347
Bornean orang-utan	*Pongo pygmaeus*	D38115
Sumatran orangutan	*Pongo abelii*	X97707
Gibbon	*Hylobates lar*	X99256
hamadryas baboon	*Papio hamadryas*	Y18001

Summary of data and method

In this study, DNA sequences from the mitochondrial NADH 1 dehydrogenase genes were used to analyse the phylogenetic relationship among 8 primate species.

We retrieved sequences from GenBank, used clustalW to align sequences, and used programs in the phylip package to reconstruct phylogenetic trees and perform bootstrap analysis. Phylogeny reconstruction was carried out under distance method based on Jukes-Cantor substitution model and parsimony methods.

We further demonstrated that how one partitioned the data by codon position showed different topologies from the full length sequences.

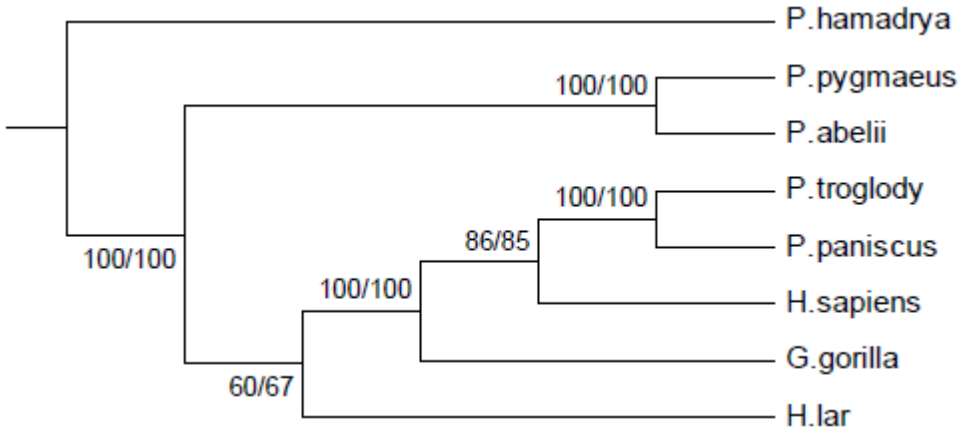


Figure 1. Phylogenetic tree based on full length sequence

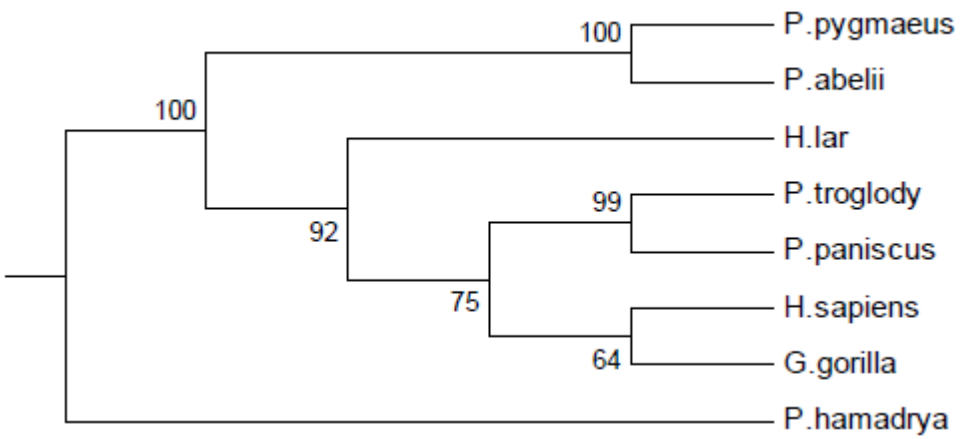


Figure 2. Phylogenetic tree based on codon position 1 under NJ method

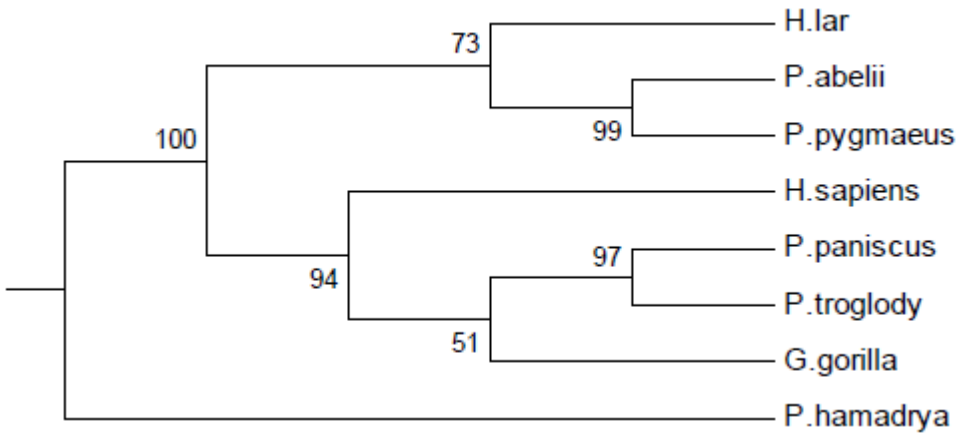


Figure 3. Phylogenetic tree based on codon position 2 under NJ method

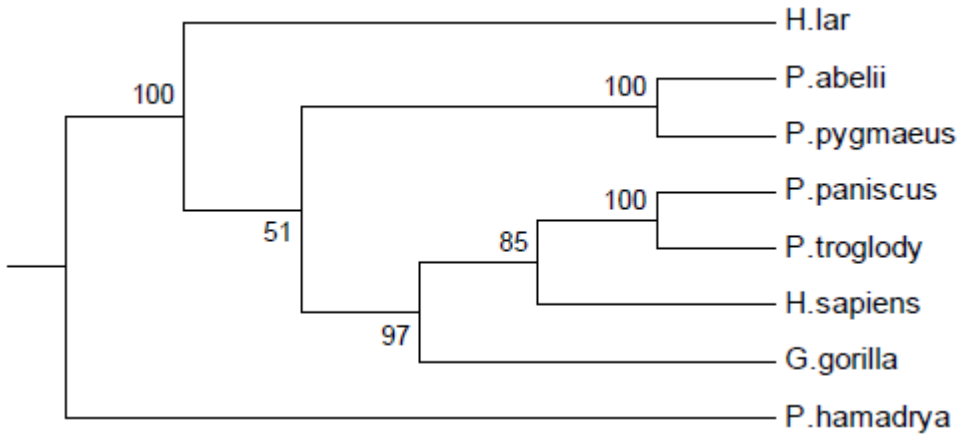


Figure 4. Phylogenetic tree based on codon position 3 under NJ method

Discussion

Phylogenetic tree of *NADH1* gene from 8 species obtained by using distance method and parsimony method are identical. It is mainly caused by the low sequence divergence and limited sequence number. This suggest that when the extent of sequences divergence is not very high and a substantial number of sequences are used, NJ and MP generally give the same or similar topologies. In this case, a large number of nucleotides or more different genes should be used to establish the tree.

While the topologies were different among the phylogenetic trees reconstructed using 3 codon positions. Strategies that partitioned the *NADH1* gene by the third codon position performed better than other partition strategies which tree was correlated with the full length phylogenetic tree.

In summary, our results indicated that the 957-bp portion of the mitochondrial *NADH 1* dehydrogenase gene was useful for inferring a phylogeny for the 8 primate species included in the study. The results should be evaluated with more molecular, cytogenetic, and morphological data to derive a phylogenetic evolutionary history.

注意事项

1. 到目前为止，在进行系统发育分析中，最重要的因素不是采用的建树方法，而是输入数据的质量。数据选择和序列比对都非常重要。因为即使是最复杂的系统发育分析方法都不能校正输入数据的错误。

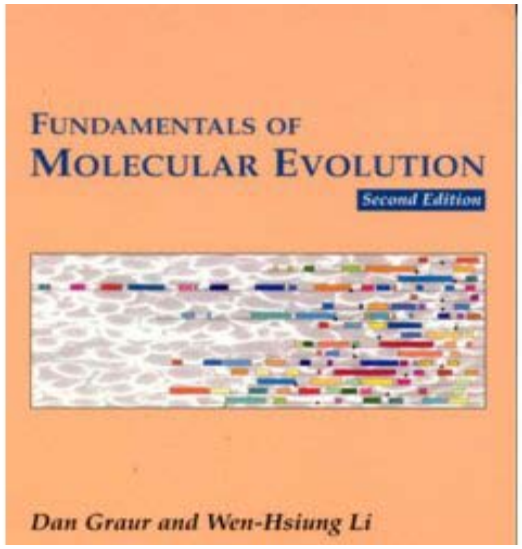
2. 从尽可能多的角度观察数据。使用三种主要方法的每一个，然后比较它们所建立的进化树的一致性。

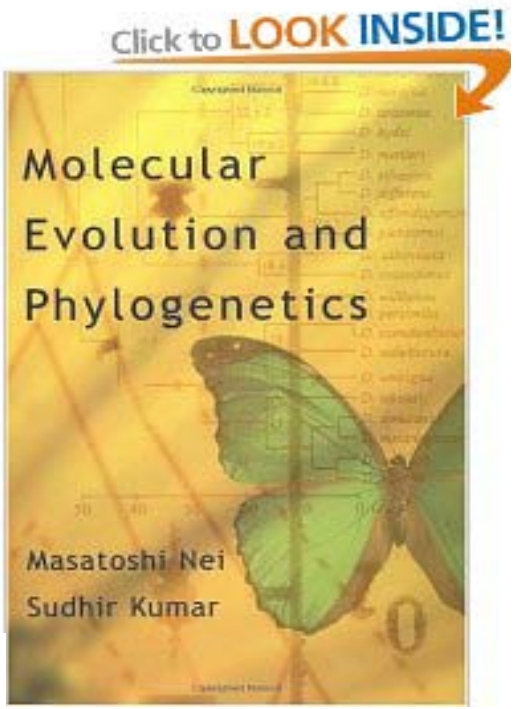
3. 外类群对于分析的影响是相当的。使用无可争议的同源物种作为外类群，这个外类群要足够近，以提供足够的信息，但又不能太近以至于和树中的种类相混。

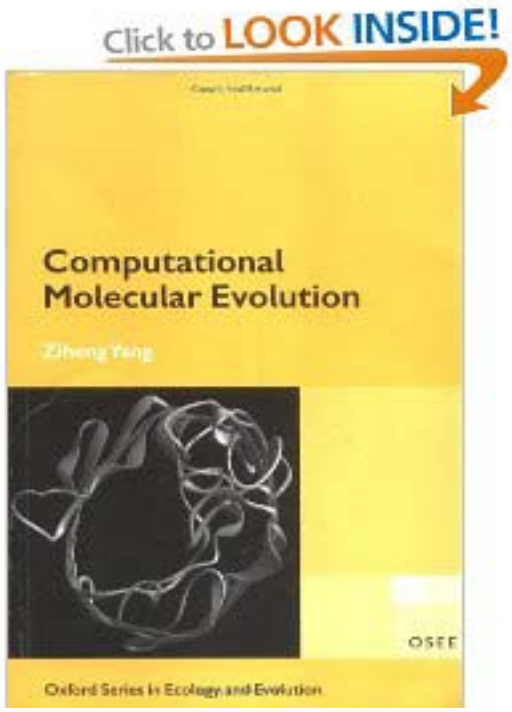
4. 有时候程序可以给出不同的进化树，仅仅是因为序列出现在输入文件的顺序不同。

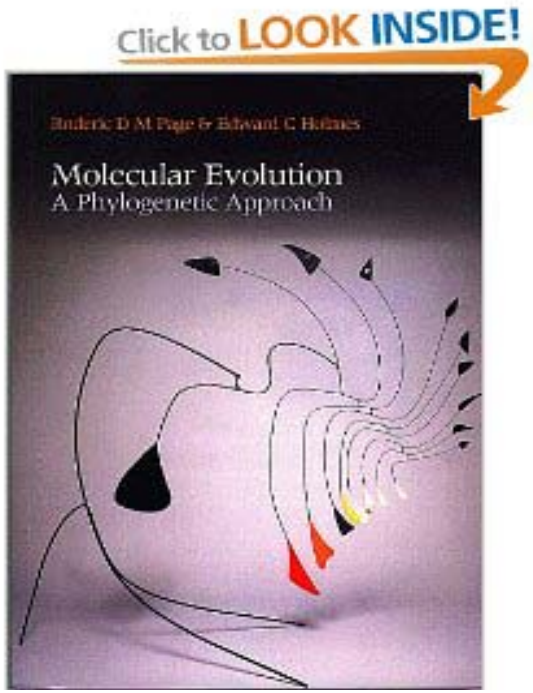
推荐书目

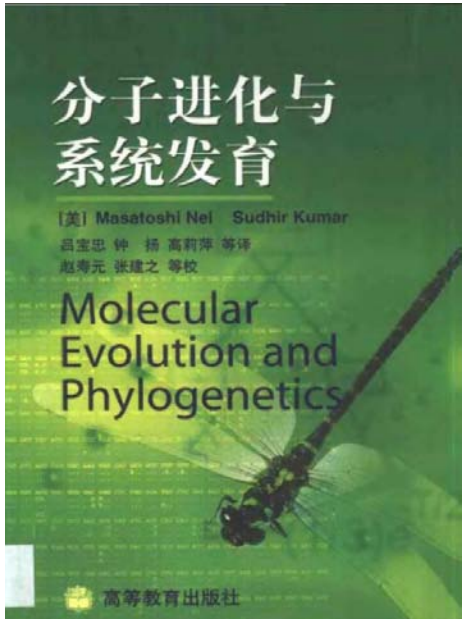
- **Fundamentals of Molecular Evolution, Second Edition.** Dan Graur, University of Houston, and Wen-Hsiung Li, University of Chicago
- **Molecular Evolution and Phylogenetics.** Masatoshi Nei and Sudhir Kumar. Oxford University Press, Oxford. 2000
- **Molecular Evolution: A Phylogenetic Approach.** Roderick D.M. Page, Edward C. Holmes, January 1991, Wiley-Blackwell
- **Computational Molecular Evolution.** Ziheng Yang, October 2006, Oxford University Press

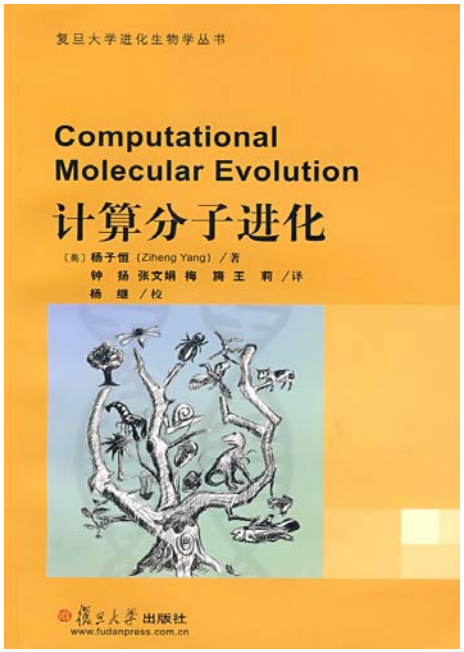

FUNDAMENTALS OF
MOLECULAR EVOLUTION
Second Edition
Dan Graur and Wen-Hsiung Li


Click to LOOK INSIDE!
Molecular
Evolution and
Phylogenetics
Masatoshi Nei
Sudhir Kumar


Click to LOOK INSIDE!
Computational
Molecular Evolution
Ziheng Yang
OSEE
Oxford Series in Ecology and Evolution


Click to LOOK INSIDE!
Molecular Evolution
A Phylogenetic Approach


分子进化与
系统发育
[美] Masatoshi Nei Sudhir Kumar
Molecular
Evolution and
Phylogenetics
高等教育出版社


复旦大学进化生物学丛书
Computational
Molecular Evolution
计算分子进化
复旦大学出版社
www.fudanpress.com.cn

谢谢!