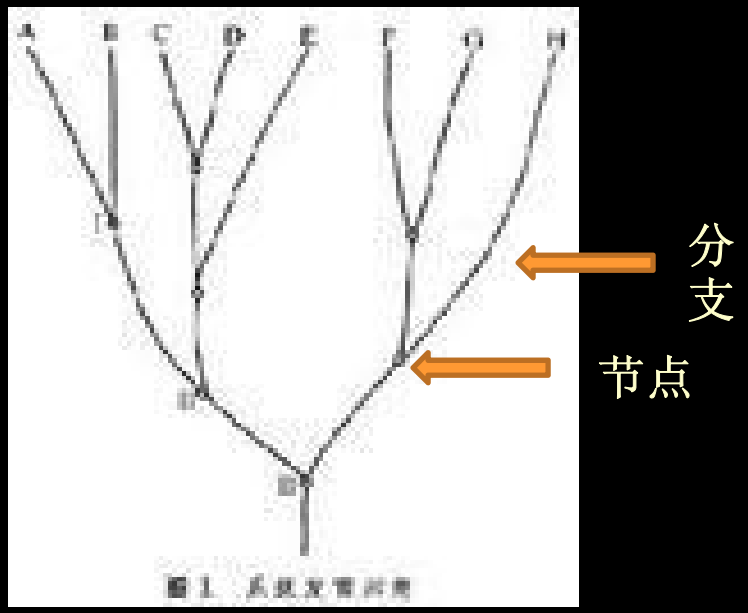
分子系统发育分析

2008级生物信息学一班C队
张学才

系统发育分析相关基本知识

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成为系统发育树。



系统发育分析相关基本知识

- 系统进化树可以分为有根树和无根树，有根树是有方向的树，具有一个唯一的根节点，代表树中所有物种的共同祖先。

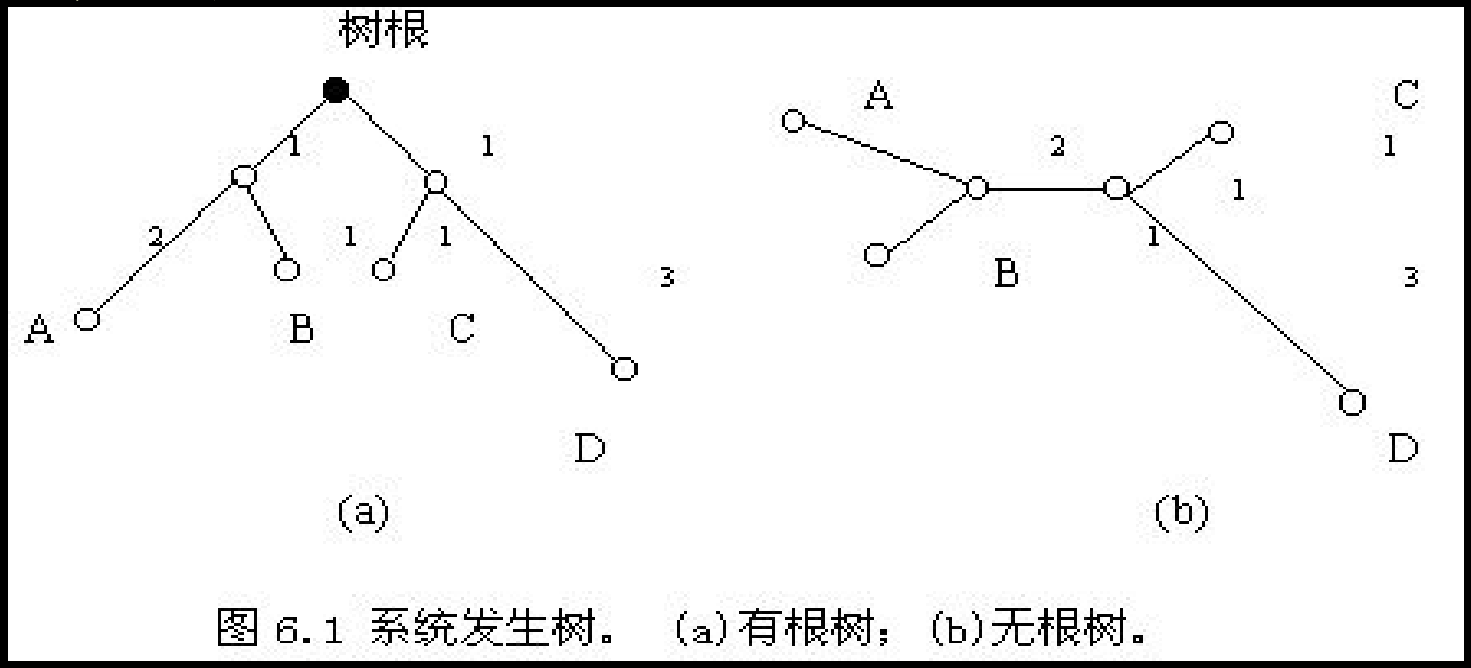
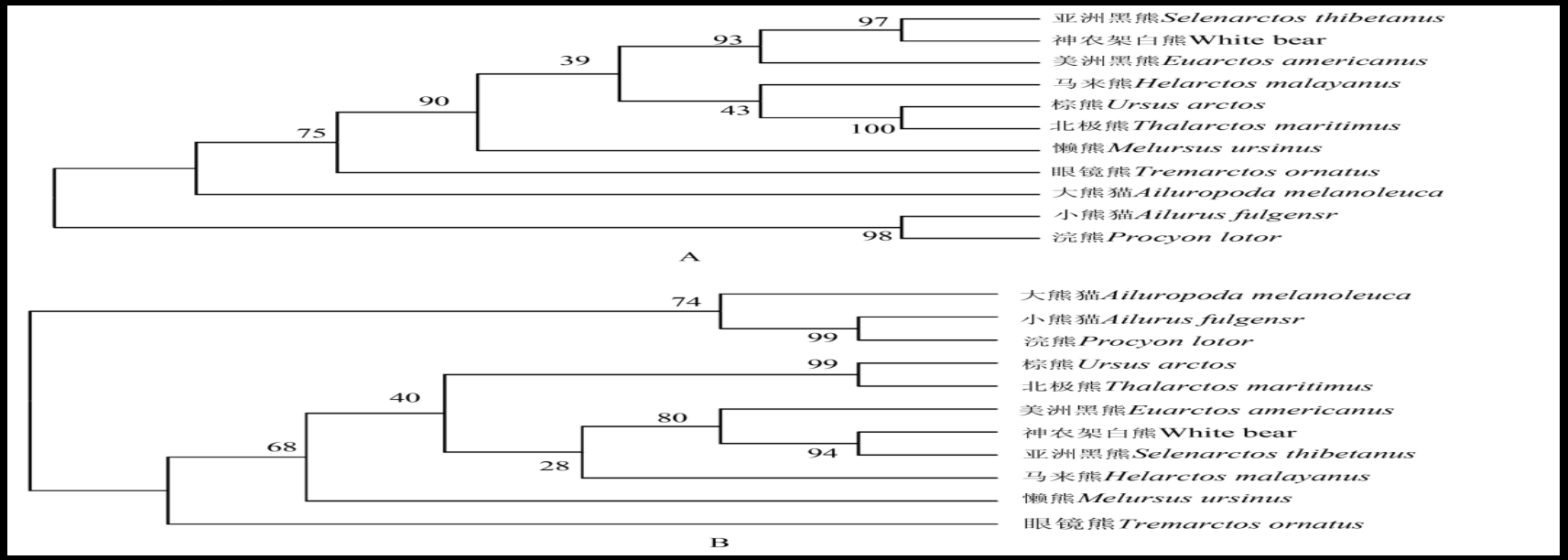


图 6.1 系统发生树。 (a)有根树；(b)无根树。

系统发育分析相关基本知识

分子系统发育树通过比较生物大分子序列差异的数值构建的系统发育树，常用的生物大分子为蛋白质序列和核酸序列。



基于Cyt *b*基因序列构建的分子系统树(重复次数为1 000次)
图中数字为自举置信水平(BCL)值; A: NJ树; B: MP树。

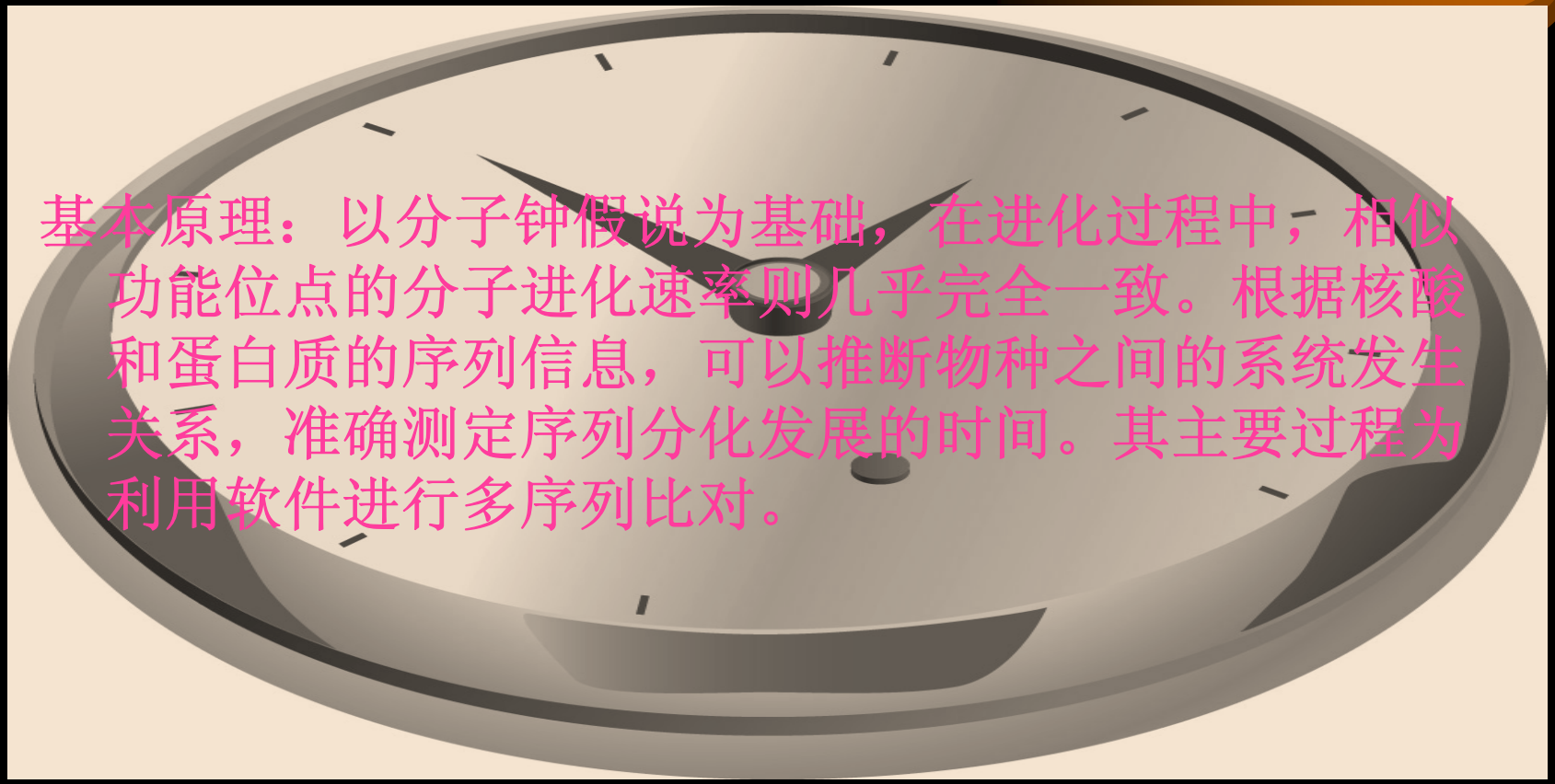
分子系统发生分析的主要步骤

1. 分析分子序列或特征数据

2. 构建系统发育树

3. 检验结果

分子序列或特征数据的分析



基本原理：以分子钟假说为基础，在进化过程中，相似功能位点的分子进化速率则几乎完全一致。根据核酸和蛋白质的序列信息，可以推断物种之间的系统发生关系，准确测定序列分化发展的时间。其主要过程为利用软件进行多序列比对。

系统发生树的构造

按照某种方法，算出代表序列两两之间的差异度，基于这些差异度，绘制系统发生树。

分类	名称	简称
Distance Matrix methods(DM)	平均连接聚类法	UPGMA
	最小进化法	ME
	邻接法	NJ
characters	最大简约法	MP
	最大似然法	ML
	进化简约法	EP

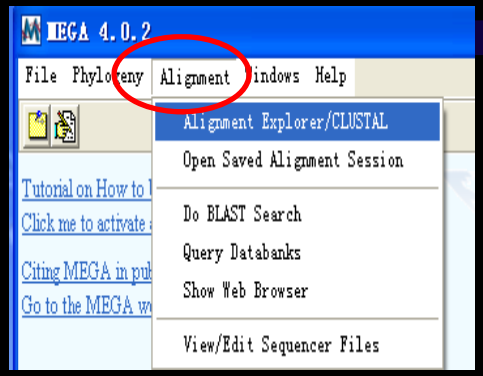
常用方法的基本信息

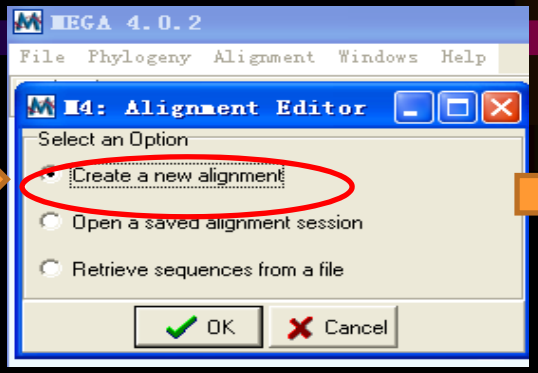
名称	基本特征	适用范围	优点	缺点
邻接法	不需要分子钟假设，是基于最小进化原理，进行类的合并时，不仅要求待合并的类是相近的，而且要求待合并的类远离其他的类。	远缘序列，进化距离不大，信息位点少的短序列	假设少，树的构建相对准确，，计算速度快，只得一颗树，可以分析较多的序列，运行速度优于最大简约法	序列上的所有位点等同对待，且所分析的序列的进化距离不能太大
最大简约法	基于进化过程中碱基替代数目最少这一假说，不需要替代模型，对所有可能的拓扑结构进行计算，并计算出所需替代数最小的那个拓扑结构，作为最优树	近缘序列 物种序列的数目≤12	善于分析某些特殊的分子数据如插入、缺失等序列有用。	只适于序列数目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，

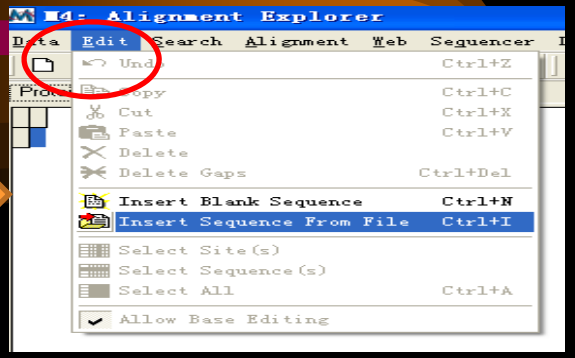
分子进化与系统发育分析软件

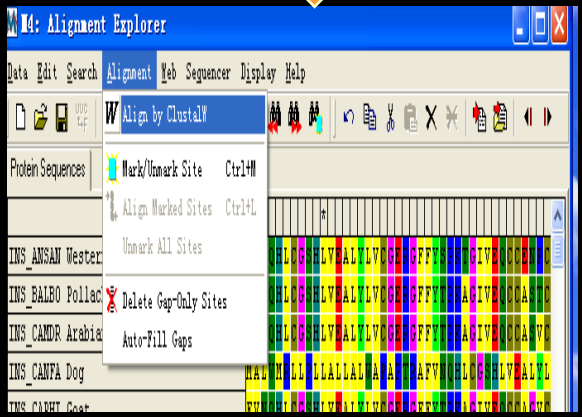
软件名称	网址	说 明
PHYLIP	http://evolution.genetics.washinton.edu/phylip/software.html	目前发布最广，用户最多的通用系统树构建软件，由美国华盛顿大学Felsenstein开发，可免费下载，适用绝大多数操作系统
PAUP	ftp://onyx.si.edu/paup	国际上最通用的系统树构建软件之一，美国simthsonion institute开发，仅适用Apple-Macintosh和UNIX操作系统
Tree of Life	http://phylogeny.arizona.edu/tree/program/program.html	美国University of Arizona建立的系统发育方面网站
MEGA	http://bioinfo.weizmann.ac.il/databases/info/mega.sof	美国宾西法尼亚州立大学MasatoshiNei开发的分子进化遗传学软件
MOLPHY	ftp://ftpsunmh.ism.ac.jp/pub/molphy	日本国立统计数理研究所开发，最大似然法构树
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	英国University college London开发，最大似然法构树和分子进化模型
PUZZLE	ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle	应用quarter puzzling方法(一种最大简约法)构建系统树
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	英国University of Glasgow开发
phylogeny	http://www.ebi.ac.uk/biocat/phyl	欧洲生物信息研究所(EBI)的系统发育分析软件

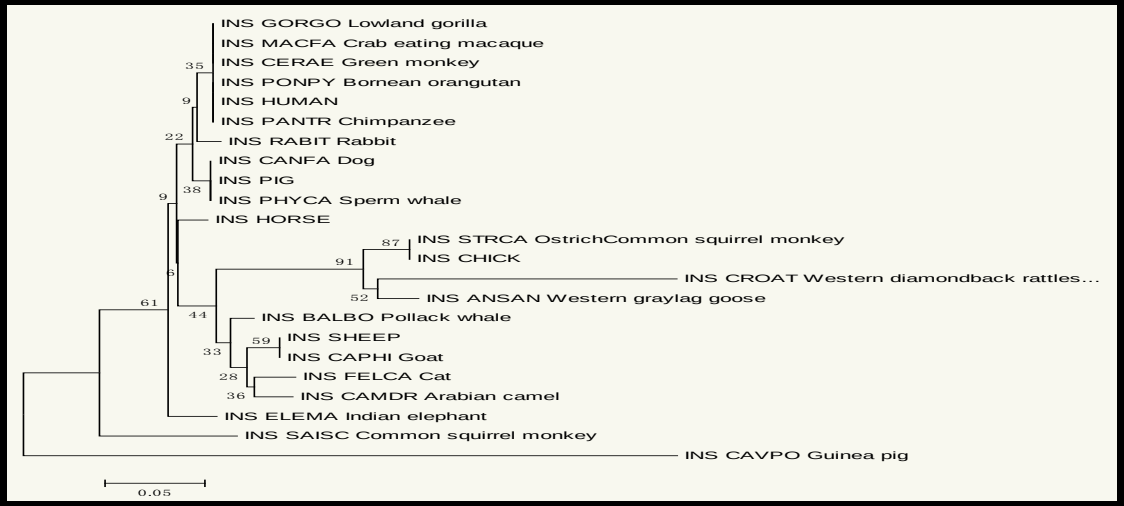
MEGA软件











- 胰岛素蛋白NJ法的系统发育树

PHLIP 软件-1

PHLIP软件-2

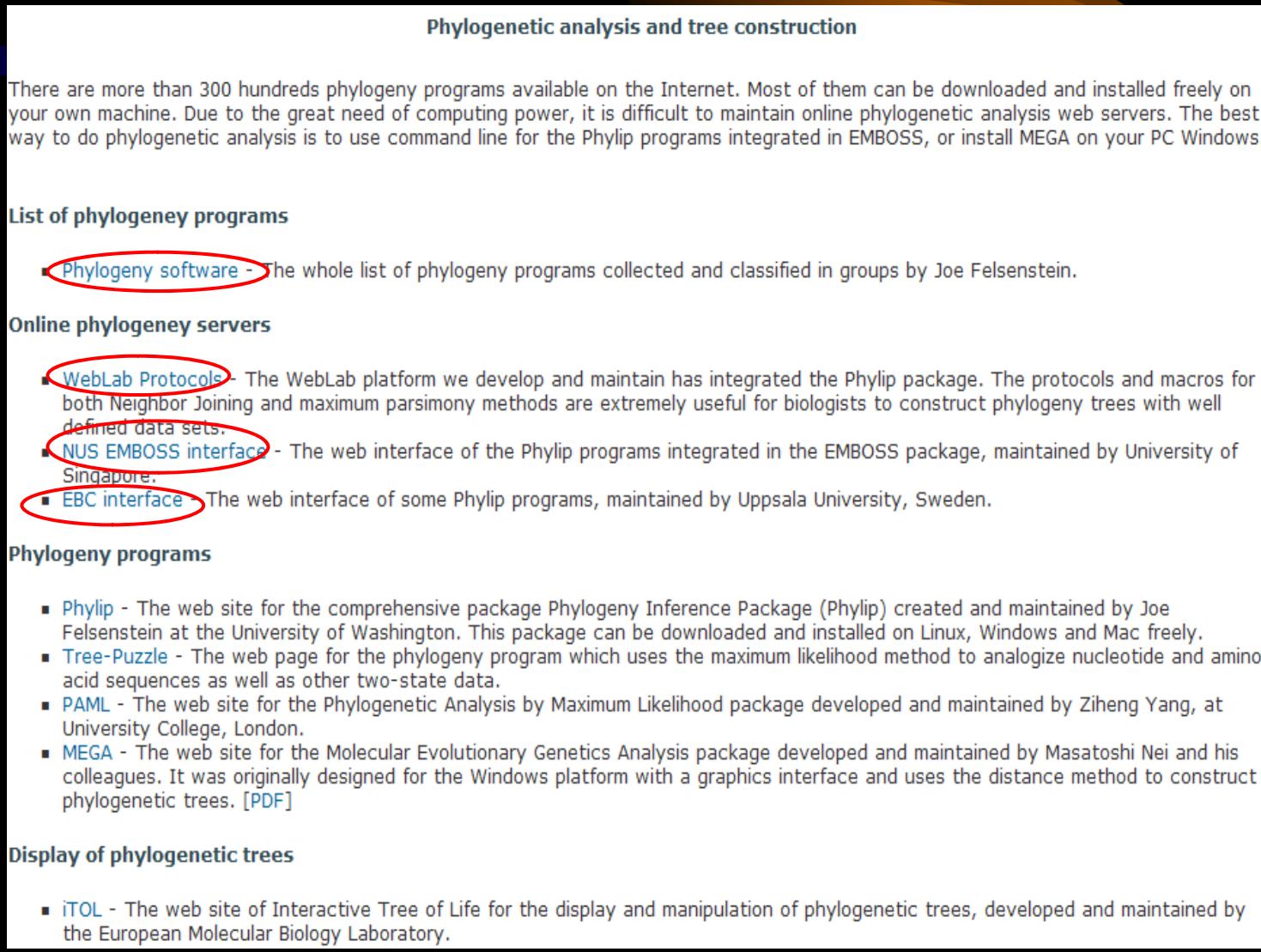
网络资源的利用

提供网上应用软件下载的链接：

- Phylogeny software

提供网上应用程序的网站：

- Weblab
- EMBOSS
- EBC
- ExPASy



Phylogenetic analysis and tree construction

There are more than 300 hundreds phylogeny programs available on the Internet. Most of them can be downloaded and installed freely on your own machine. Due to the great need of computing power, it is difficult to maintain online phylogenetic analysis web servers. The best way to do phylogenetic analysis is to use command line for the Phylip programs integrated in EMBOSS, or install MEGA on your PC Windows

List of phylogeney programs

- Phylogeny software - The whole list of phylogeny programs collected and classified in groups by Joe Felsenstein.

Online phylogeney servers

- WebLab Protocols - The WebLab platform we develop and maintain has integrated the Phylip package. The protocols and macros for both Neighbor Joining and maximum parsimony methods are extremely useful for biologists to construct phylogeny trees with well defined data sets.
- NUS EMBOSS interface - The web interface of the Phylip programs integrated in the EMBOSS package, maintained by University of Singapore.
- EBC interface - The web interface of some Phylip programs, maintained by Uppsala University, Sweden.

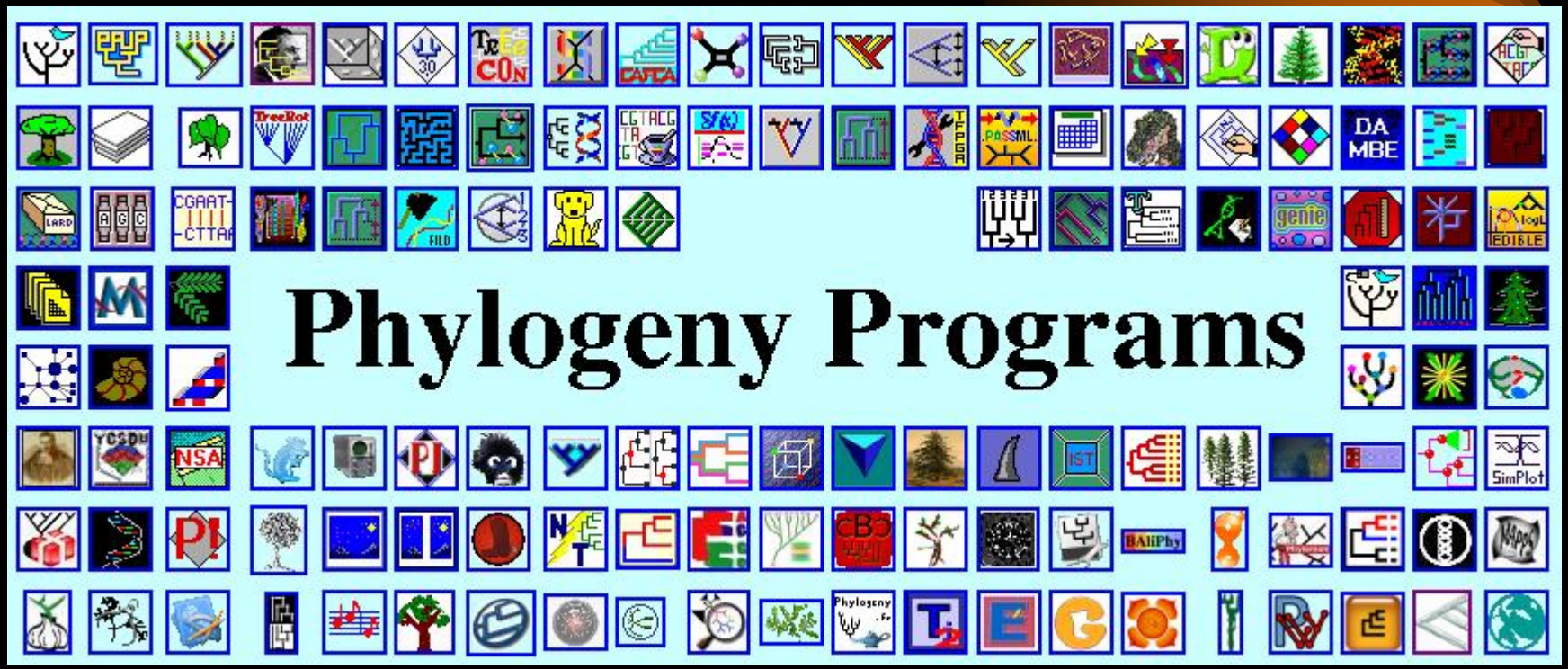
Phylogeny programs

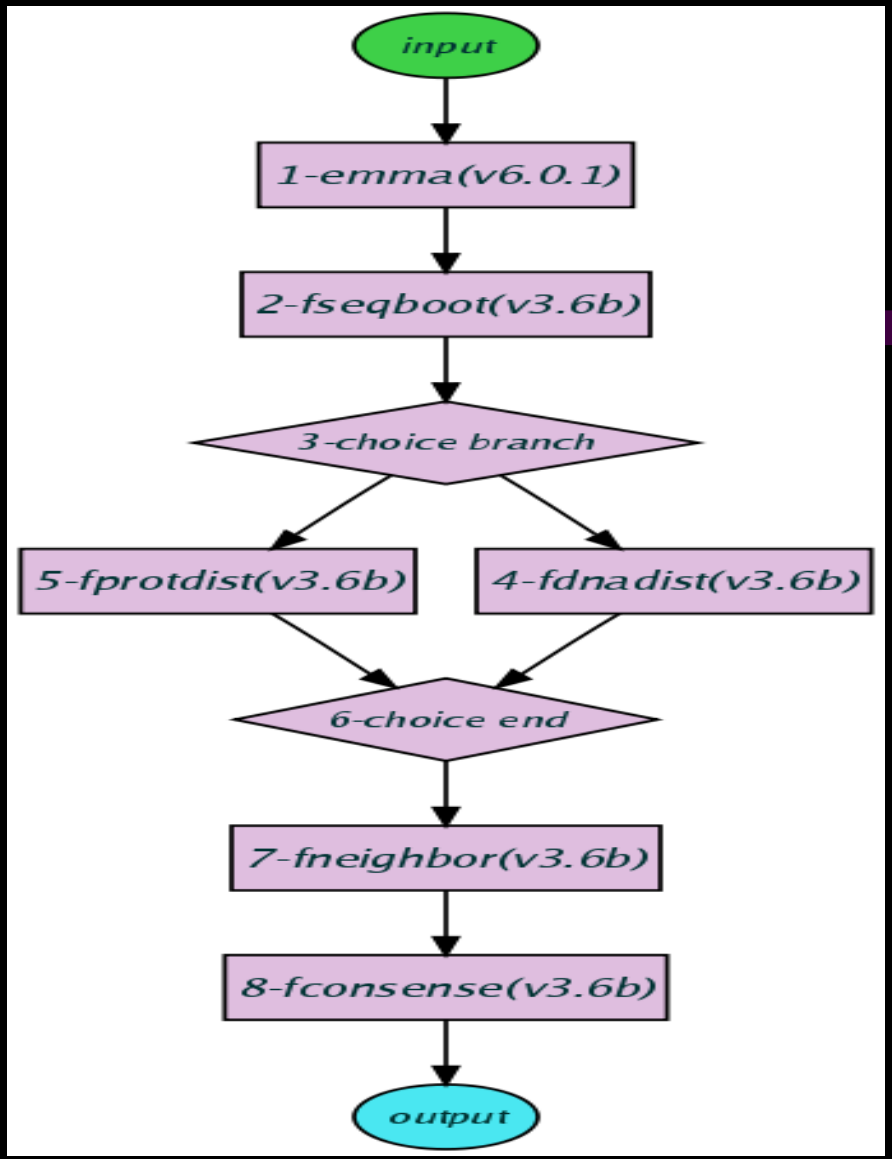
- Phylip - The web site for the comprehensive package Phylogeny Inference Package (Phylip) created and maintained by Joe Felsenstein at the University of Washington. This package can be downloaded and installed on Linux, Windows and Mac freely.
- Tree-Puzzle - The web page for the phylogeny program which uses the maximum likelihood method to analogize nucleotide and amino acid sequences as well as other two-state data.
- PAML - The web site for the Phylogenetic Analysis by Maximum Likelihood package developed and maintained by Ziheng Yang, at University College, London.
- MEGA - The web site for the Molecular Evolutionary Genetics Analysis package developed and maintained by Masatoshi Nei and his colleagues. It was originally designed for the Windows platform with a graphics interface and uses the distance method to construct phylogenetic trees. [PDF]

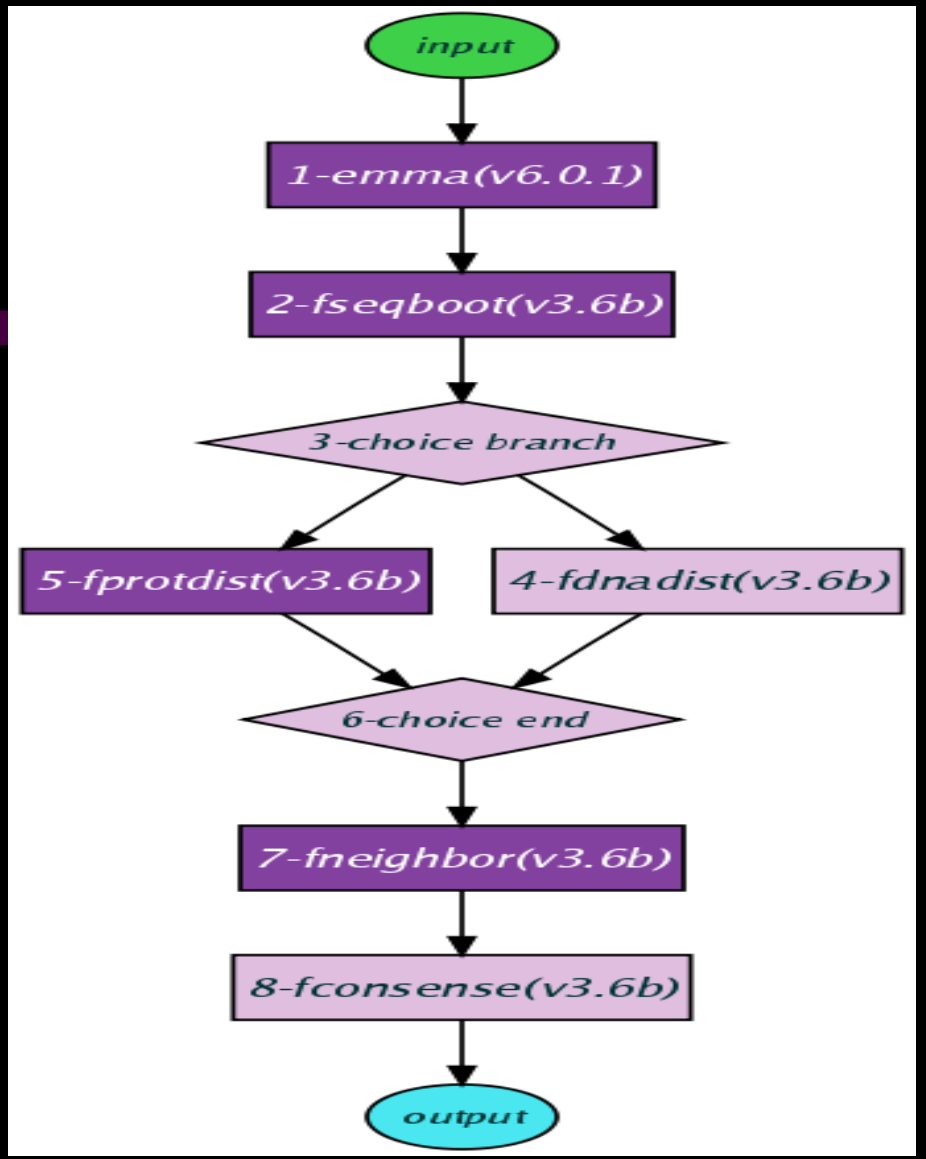
Display of phylogenetic trees

- iTOL - The web site of Interactive Tree of Life for the display and manipulation of phylogenetic trees, developed and maintained by the European Molecular Biology Laboratory.

网络资源的利用





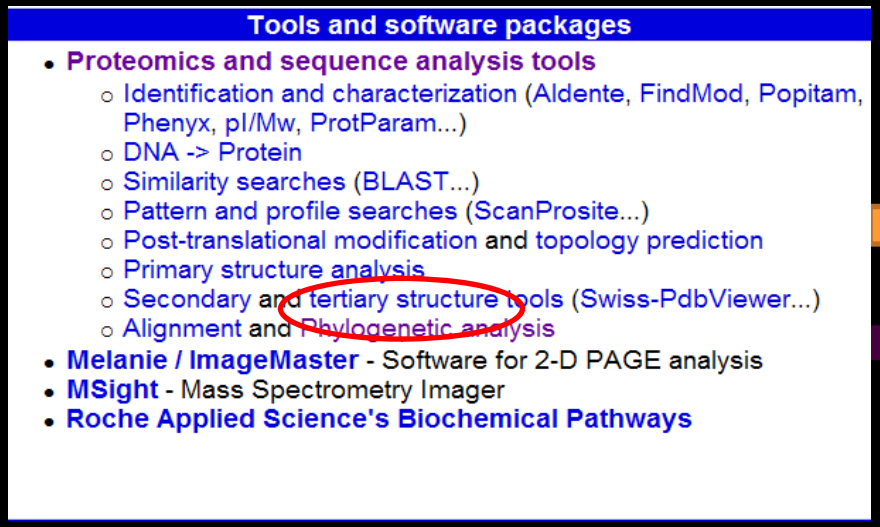


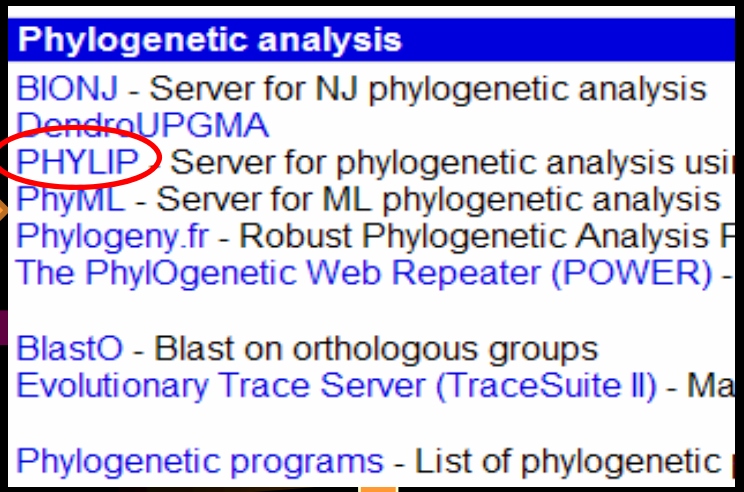
Weblab中可利用邻近法和最大简约法构建系统发育树

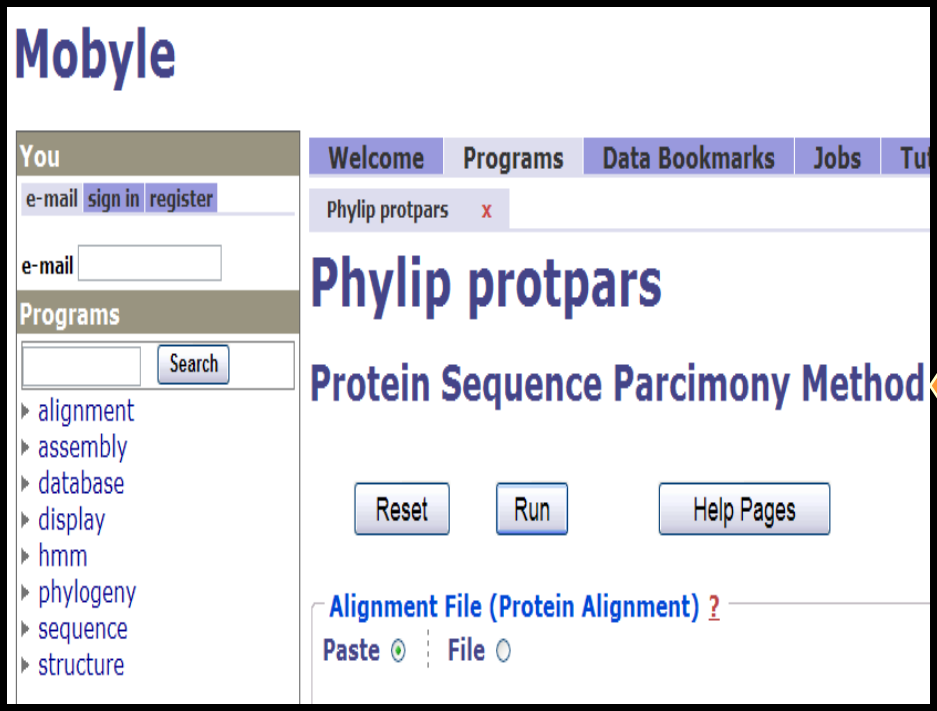
运行结果

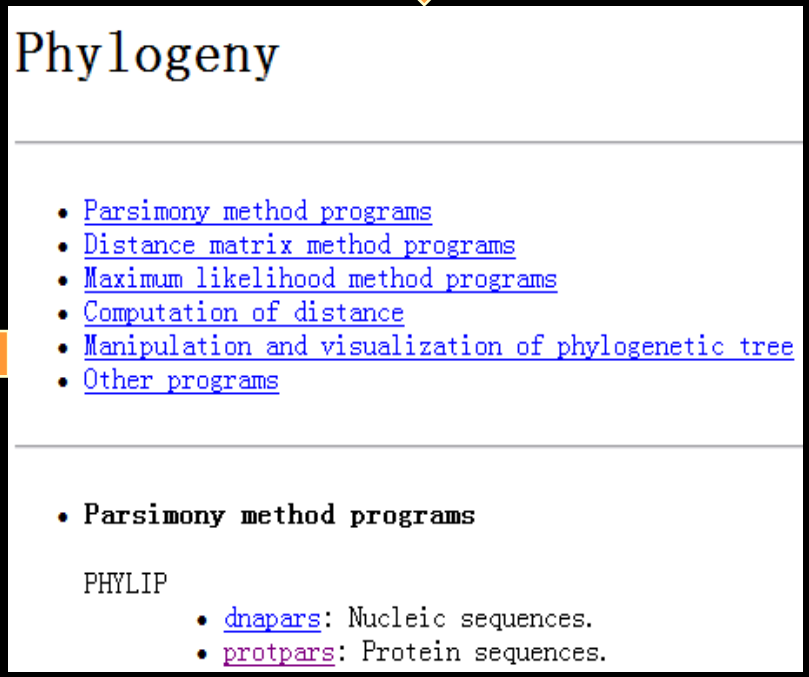
ID	PROGRAM	STATUS	INPUT FILE	OUTPUT FILE	START TIME	RUN TIME	ERROR MESSAGE
86532	emma(v6.0.1) (Multiple alignment program - interface to ClustalW program)	Finished	1235638072133.fasta	7hba.dendrogram 7hba.msf	2009-02-26 16:47 PM	1.0s	
86539	fseqboot(v3.6b) (Bootstrapped sequences algorithm)	Finished	7hba.msf	7hba.seqboot	2009-02-26 17:03 PM	1.0s	
86547	fprotdist(v3.6b) (Protein distance algorithm)	Finished	7hba.seqboot	7hba.dmatrix	2009-02-26 19:15 PM	9.0s	
86548	fneighbor (v3.6b) (Phylogenies from distance matrix by N-J orUPGMA method)	Finished	7hba.dmatrix	7hba.treefile 7hba.neighbor	2009-02-26 19:24 PM	1.0s	

```
Data set # 100:

 Neighbor-joining method

 Negative branch lengths allowed


  +---Mouse
  !
  ! +-----Dolphin
  ! !
  4-5       +-----Chicken
  ! !     +-2
  ! !     ! +----------------------Snake
  ! +----3
  !       !    +---------------------Frog
  !       +---1
  !            +---------------------------Goldfish
  !
  +--Human


remember: this is an unrooted tree!

Between        And             Length
-------        ---             ------
   4          Mouse           0.07115
   4             5            0.02287
   5          Dolphin         0.08562
   5             3            0.08438
   3             2            0.02744
   2          Chicken         0.09995
   2          Snake           0.38149
   3             1            0.06541
   1          Frog            0.35471
   1          Goldfish        0.47121
   4          Human           0.04677
```









- ExPASy中利用PHYLIP构建系统发育树

可信度检验

常用的三种方法：

- 1. The bootstrap
- 2. Delete-half-jackknifing
- 3. Permuting species within characters

分子系统发育分析需要注意的问题

特征分子的选择：既可以用核酸序列又可以用蛋白序列，用核酸序列还是蛋白序列主要取决于序列的性质和研究的目的。

- 对于具有很近亲缘关系的生物来说，选择核酸序列研究要比选择蛋白序列更快的推断出结果
- 在大多数情况下，以蛋白质为基础的发生树比以DNA为基础的发生树更恰当。

① 蛋白质序列含有更多相对保守的序列。
② 蛋白质序列的比对比DNA序列的比对更灵敏。

分子系统发育分析需要注意的问题

- 直系同源：同源的基因是由于共同的祖先基因进化而产生的；
- 旁系同源：同源的基因是由于基因复制产生的。
- 用于分子进化分析中的序列，必须是直系同源的，才能真实反映进化过程

分子系统发育分析需要注意的问题

系统发生树的可靠性

- 系统发生的推断分析中，很难准确地建立一个发生树，一定要根据序列信息的特点及目的选择适当的方法与分析软件。
- 用不同的方法分析同一组数据，如果能够产生相似的系统发生树，这样的树可以认为是可靠的

- 谢谢！