

浅谈系统发育分析

A Brief Introduction to Phylogenetics

许一娜 (Xu Yina) 郑亚风 (Zheng Yafeng)
程焕义 (Cheng Huanyi) 郑婉婧 (Zheng
Wanjing)

2012-6-8

outline

- Phylogenetics basics
- Phylogenetic tree construction methods and programs

Phylogenetics basics

Ontogeny:(also ontogenesis or morphogenesis) is the origin and the development of an organism.

Phylogeny : a reconstruction of the evolutionary history of a collection of organisms.



From <http://tolweb.org/tree/>

SYSTEMATICS

Classification:

study on “products” of evolution

Evolution:

Studies on “processes”, and forces driving processes

Phylogeny:

Studies on “**relationship**” of taxa and evolutionary events

Modified from Stuessy, 1990. Plant taxonomy, p.8. Columbia UP, NY.

Phylogenetics is the study of evolutionary relation among groups of organisms (e.g. species, populations), which is discovered through molecular sequencing data and morphological data matrices.

-----wikipedia

Phylogenetics is the study of the evolutionary history of living organisms using treelike diagrams to represent pedigrees of these organisms.

-----Jin Xiong

Major assumptions

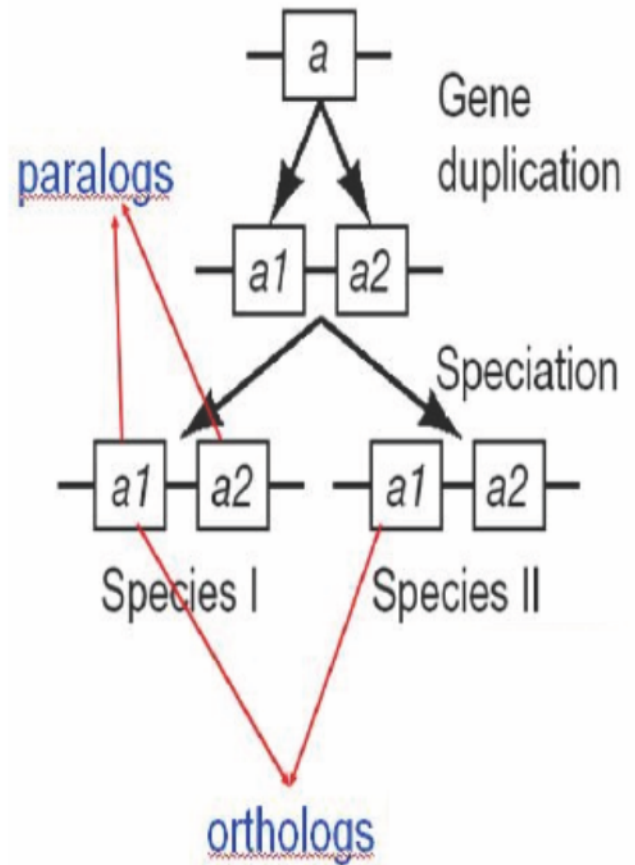
- The molecular sequences used in phylogenetic construction are homologous (they share a common arigin and subsequently diverged through time).
- Each position in a sequence evolved independently.

Orthologs and paralogs

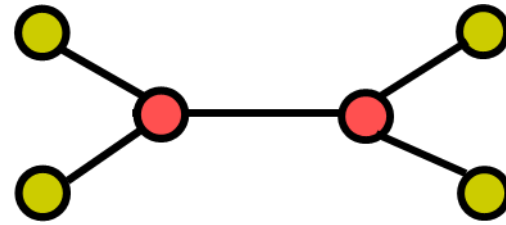
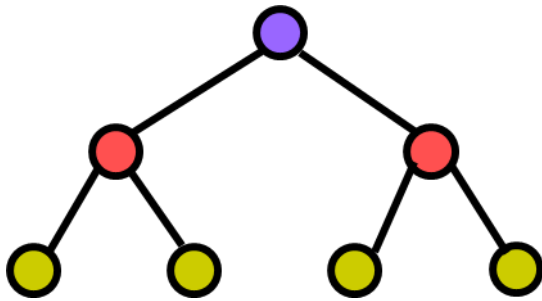
Orthologs(直系同源): 由于共同的祖先基因进化而产生的

Paralogs (并系同源) : 由于基因复制产生的

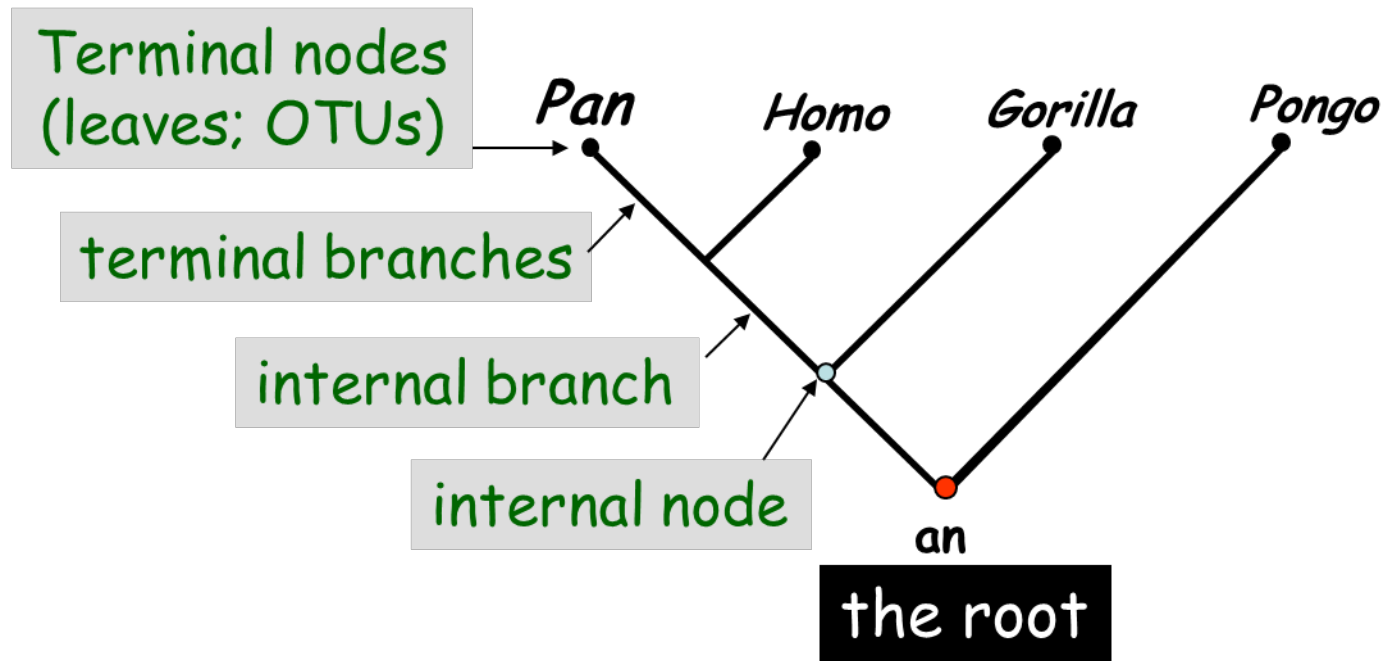
用于分子进化分析中的序列必须是直系同源的，才能真实反映进化过程



Rooted trees and unrooted trees



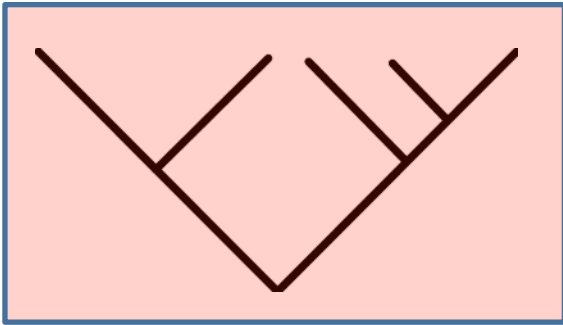
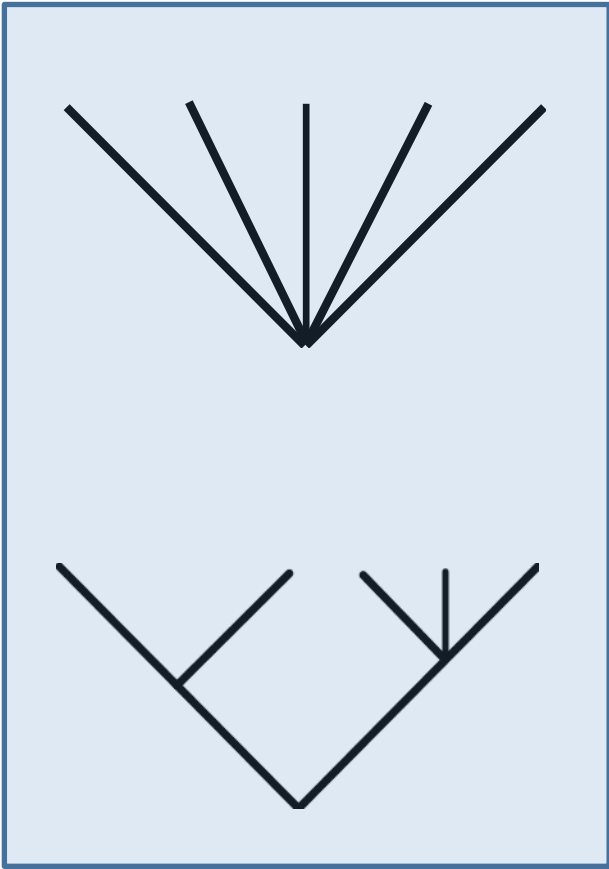
A phylogenetic tree without definition of a root is unrooted.



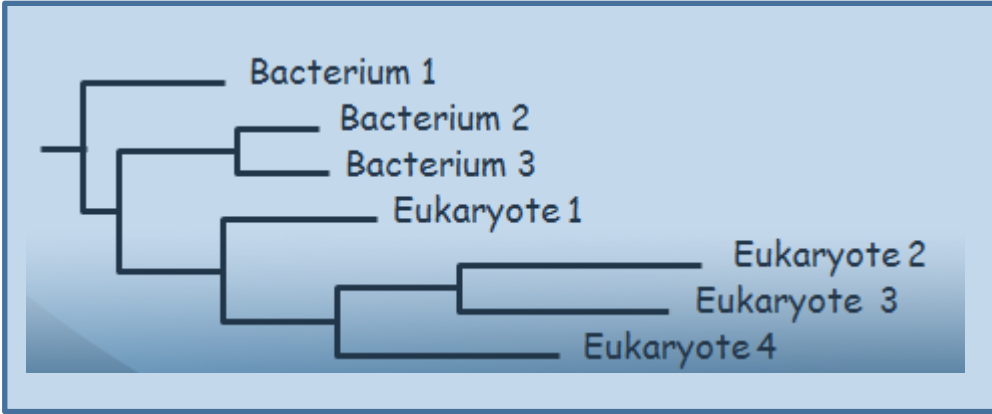
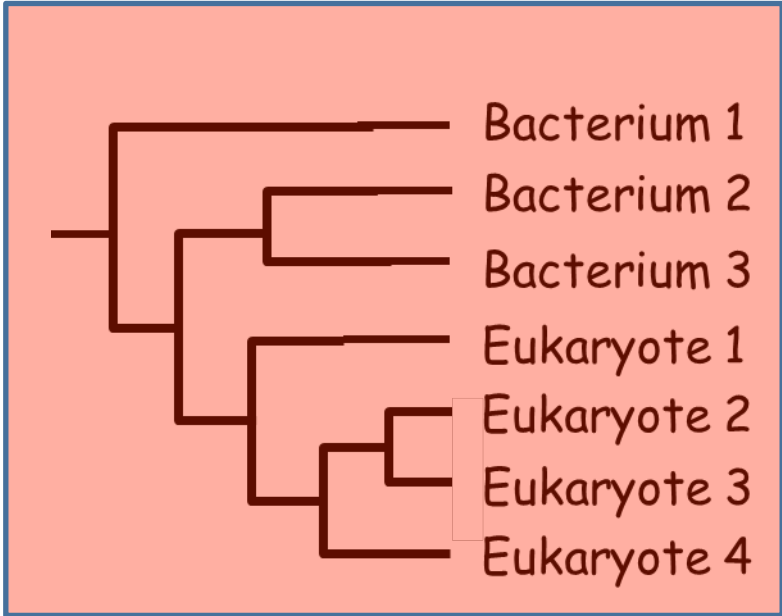
The leaves of a phylogenetic tree usually represent the present day taxa;

The internal nodes represent hypothesized ancestors.

Polytomy and Dichotomy



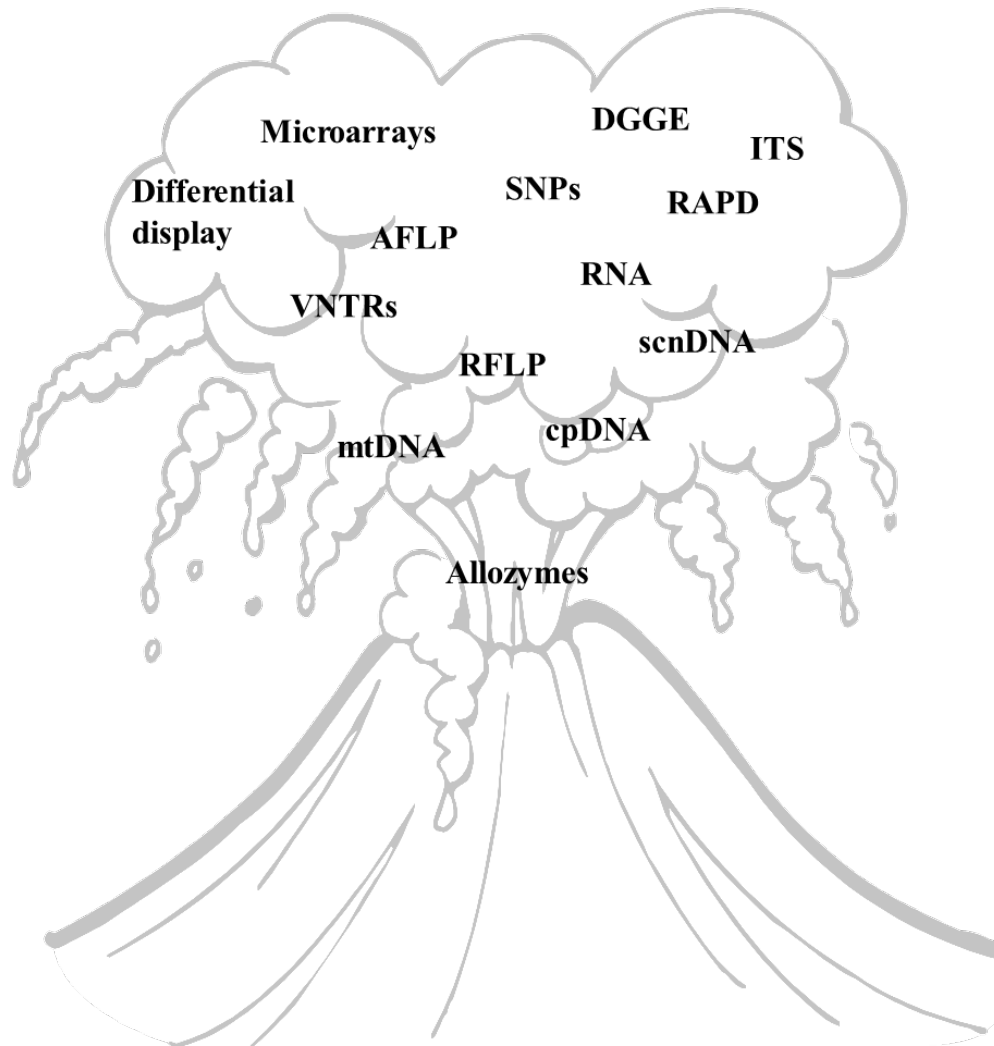
Cladograms and Phylograms



Phylogenetic tree construction methods and programs

- Choosing molecular marker
- Sequence alignment
- Models of evolution
- Tree building
- Tree evaluation

Choosing molecular marker



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



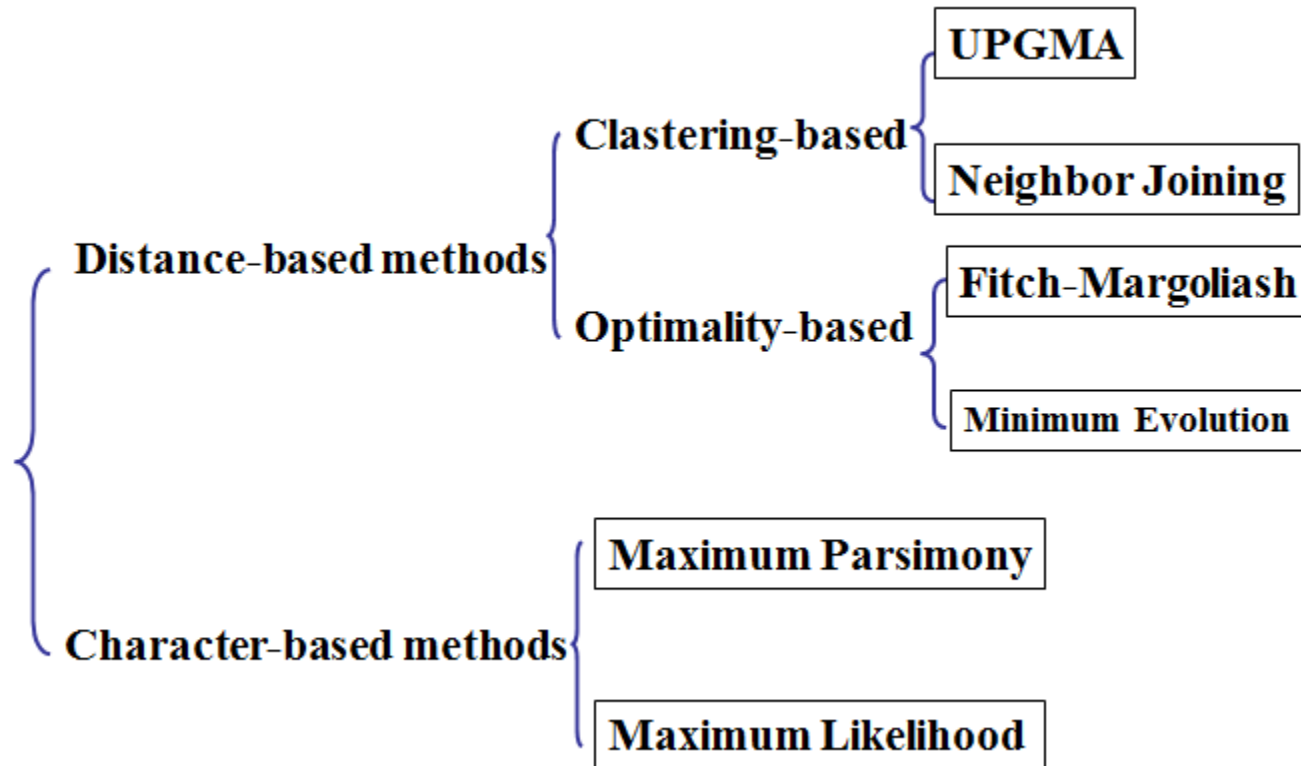
Sequence alignment

- **Clustalw**:一种渐进的多序列比对方法，先将多个序列两两比对构建距离矩阵，反应序列之间两两关系；然后根据距离矩阵计算产生系统进化指导树，对关系密切的序列进行加权；然后从最紧密的两条序列开始，逐步引入临近的序列并不断重新构建比对，直到所有序列都被加入为止。
- **CLUSTALX**—是**CLUSTAL**多重序列比对程序的Windows版本。**Clustal X**为进行多重序列和轮廓比对和分析结果提供一个整体的环境。
- **Bioedit**，可以进行人工手动的调整。

Models

- Jukes-Cantor :单参数
- Kimura : 双参数
- F81 model
- HKY85 model
- T92 model
- TN93 model
- GTR: Generalised time-reversible
- PAM
- JTT

Tree building



Distance-based methods:

- All sequences involved are homologous (orthology).
- Tree branches are additive.

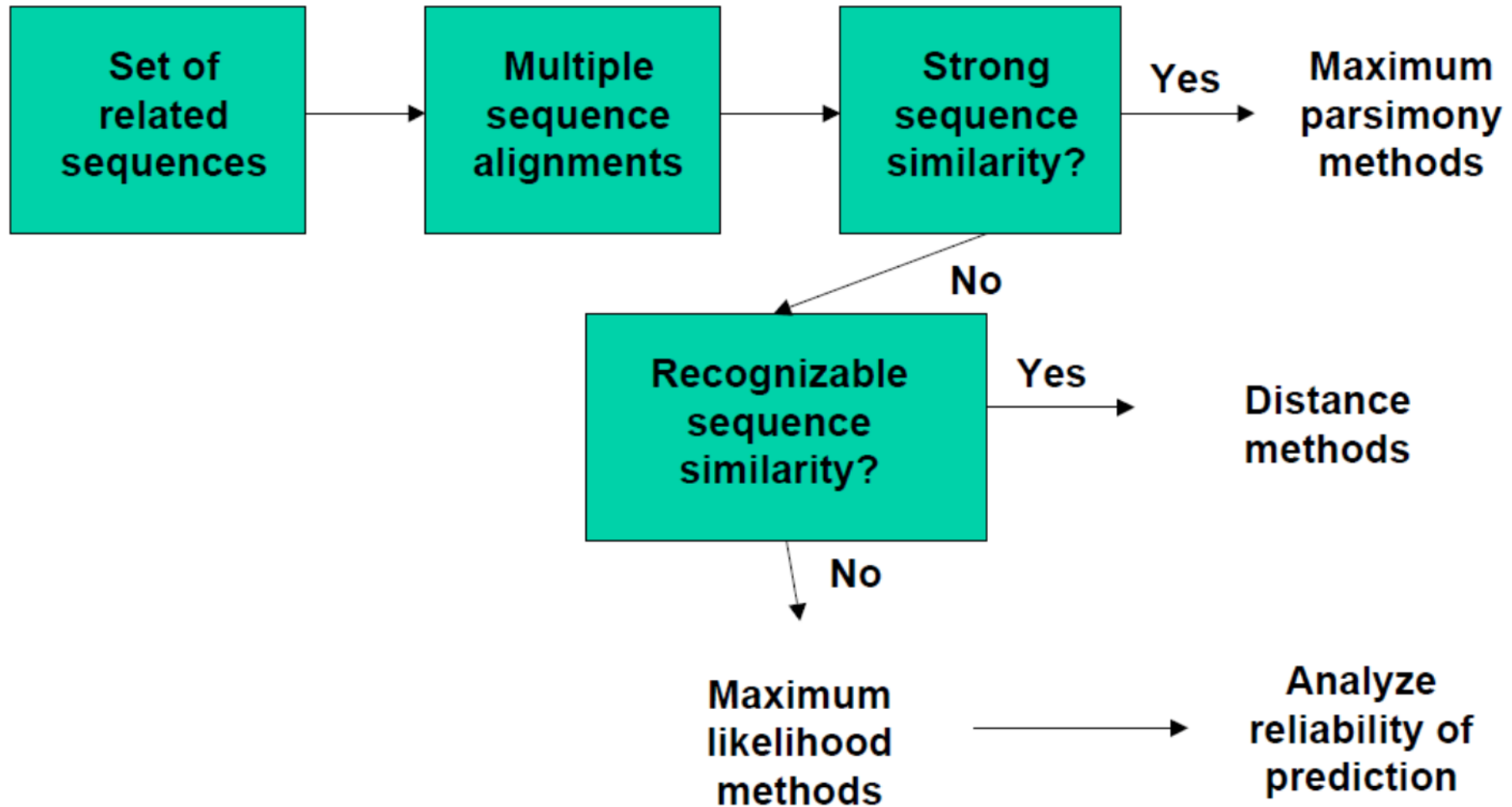
Character-based methods:

- Characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved.
- Each character evolves independently and is treated as an individual evolution unit.

常用方法

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 $N \leq 12$ 。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，

Methods for phylogenetic trees construction.



Tree evaluation

- The bootstrap
- Delete-half-jackknifing



- ☺ High bootstrap values don't mean that your tree is the true tree!
- ☺ Alignment and evolutionary assumptions are key!

常用软件

软件名称	网址	说明
PHYLIP	http://evolution.genetics.washington.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	scavotto@sinauer.com 或 ftp://onvx.si.edu/paup	国际上最通用的系统树构建软件之一，美国simthsonian institute开发，仅适用Apple-Macintosh和UNIX操作系统
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，图形化、集成的进化分析工具，不包括ML
MOLPHY	ftp://sunmh.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

软件名称	网址	说明
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, 进化树显示工具
phylogeny	http://www.ebi.ac.uk/biocat/phylogenv.html	欧洲生物信息研究所(EBI)的系统发育分析软件
PHYML	http://atgc.lirmm.fr/phyml/	快速的ML建树工具
MrBayes	http://mrbayes.csit.fsu.edu/	基于贝叶斯方法的建树工具
MAC5	http://www.agapow.net/software/mac5/	基于贝叶斯方法的建树工具
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

Example(MEGA)

The image displays a screenshot of the MEGA 5.0.5 software interface, showing multiple instances of the Alignment Explorer window and the Analysis Preferences dialog box. The Alignment Explorer windows show protein sequences for various species, including human, chimpanzee, gorilla, orangutan, gibbon, rhesus monkey, pig-tailed macaque, baboon, green monkey, talapoin, squirrel monkey, tamarin, spider monkey, woolly monkey, and leaf monkey. The Analysis Preferences dialog box is open, showing the Phylogeny Test section with Bootstrap method selected. The Substitution Model section shows the P0 model selected. The dialog box also includes sections for Analysis, Scope, Statistical Method, Rates and Patterns, and Data Subject to Use.

MEGA 5.0.5

Files Analysis Help

Align Data Models Distance Diversity Phylogeny User Tree Ancestors

Select Rates Clocks

Construct/ Test Maximum Likelihood Tree

MEGA Analysis Preferences

Options Summary

Option	Selection
Analysis	Phylogeny/Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replicates	None
Substitution Model	Bootstrap method
	Uniform branch test
Substitution Type	Amino acid
Model/Method	P0 model
Rates and Patterns	
Fewer among Sites	Uniform rates
Gamma Parameter	Not Applicable
Pattern among Lineages	Same (Homogeneous)
Data Subject to Use	
Gap/Missing Data Treatment	Complex deletion
Job/Log Name/Case/ID	Not Applicable

Buttons: [OK] [Cancel] [Help]

选择邻接法建树

选择Bootstrap检验



Thank you!