Molecular Phylogenetics

Li Jing
Peking University
Nov.30, 2008

Contents

Introduction

Phylogenetics:

The study of the evolutionary history of living organisms.

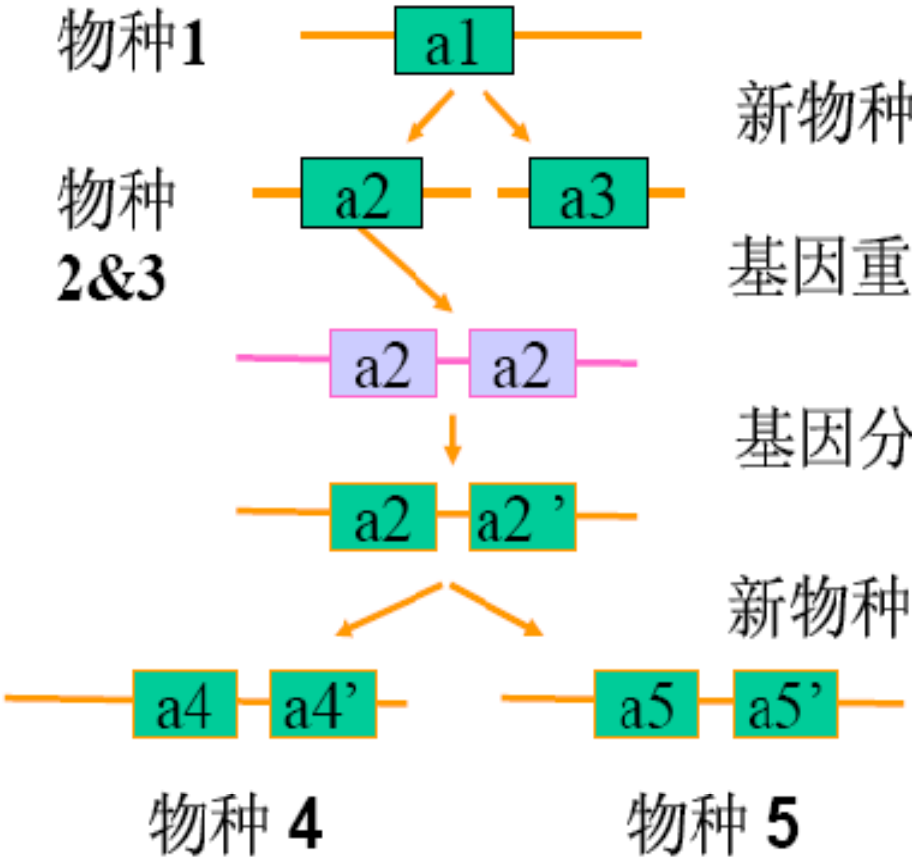
Two major assumptions:

1. Molecular sequences used in phylogenetics construction are homologous.
2. Each position in a sequence evolved independently.

同源性有两种：

直系同源**orthology**：随着新物种产生而“产生”的同源基因（纵向）

并系同源 **paralogy**：由基因的重复而产生的同源基因（横向）



直系同源基因：

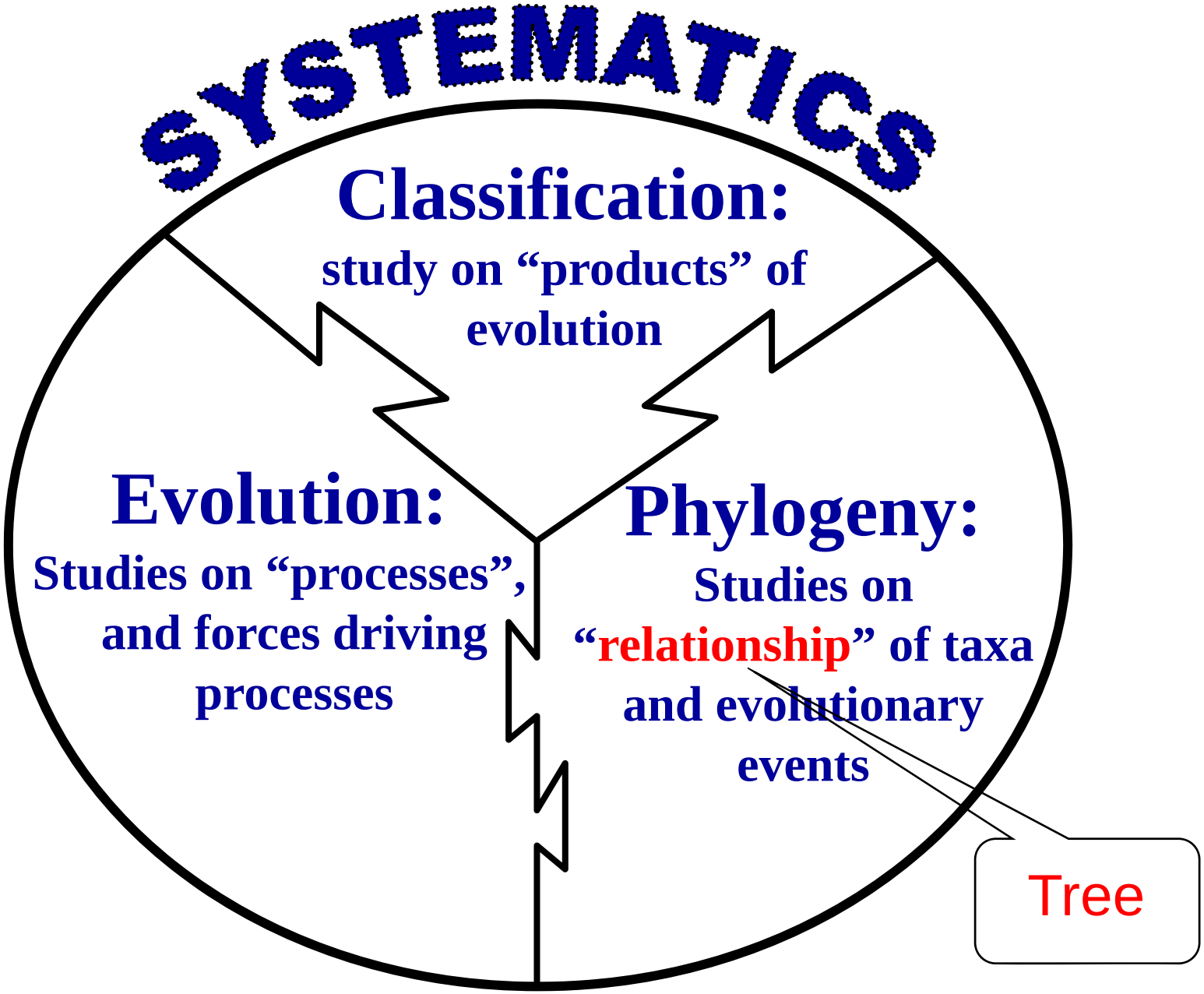
a1, a2, a3, a4, a5

a2’, a4’, a5’

并系同源基因：

a1, a2, a3, a4, a5 与

a2’, a4’, a5’ 互为并系同源



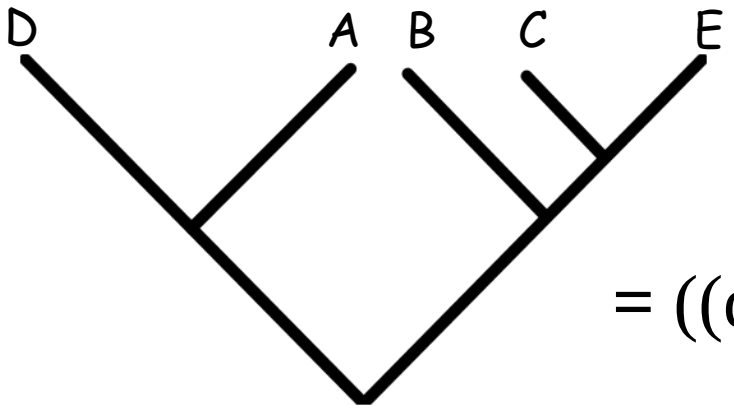
Modified from Stuessy, 1990. Plant taxonomy, p.8. Columbia UP, NY.

Study of phylogeny

A phylogeny is a reconstruction of the evolutionary history of a collection of organisms.

- **A phylogeny is a type of pedigree.**
- **Showing relationships between taxa (ex. species), not individuals.**
- **Reconstructs pattern of events leading to the distribution and diversity of life**
- **Often shown as a network or tree**

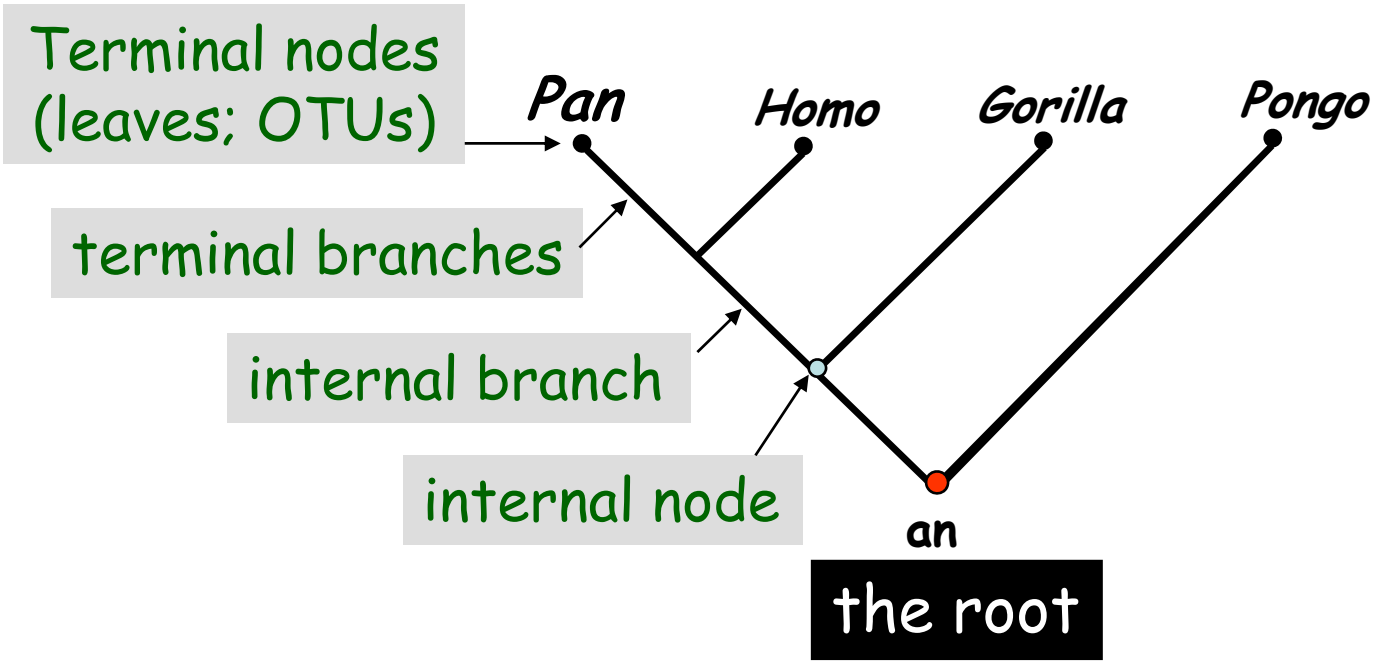
Introduction



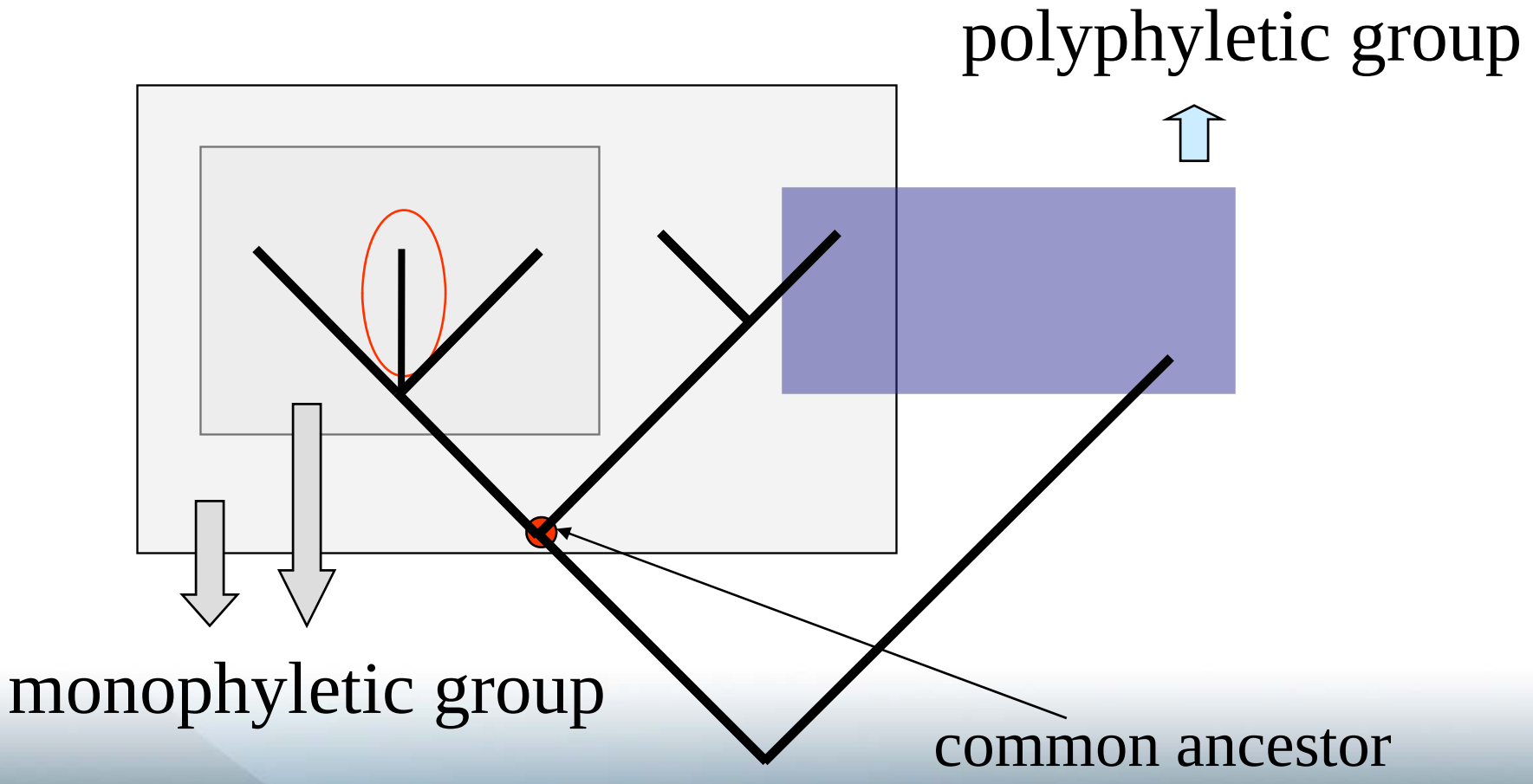
Newick format

= ((d,a),(b,(c,e))) or ((b,(e,c),(a,d))

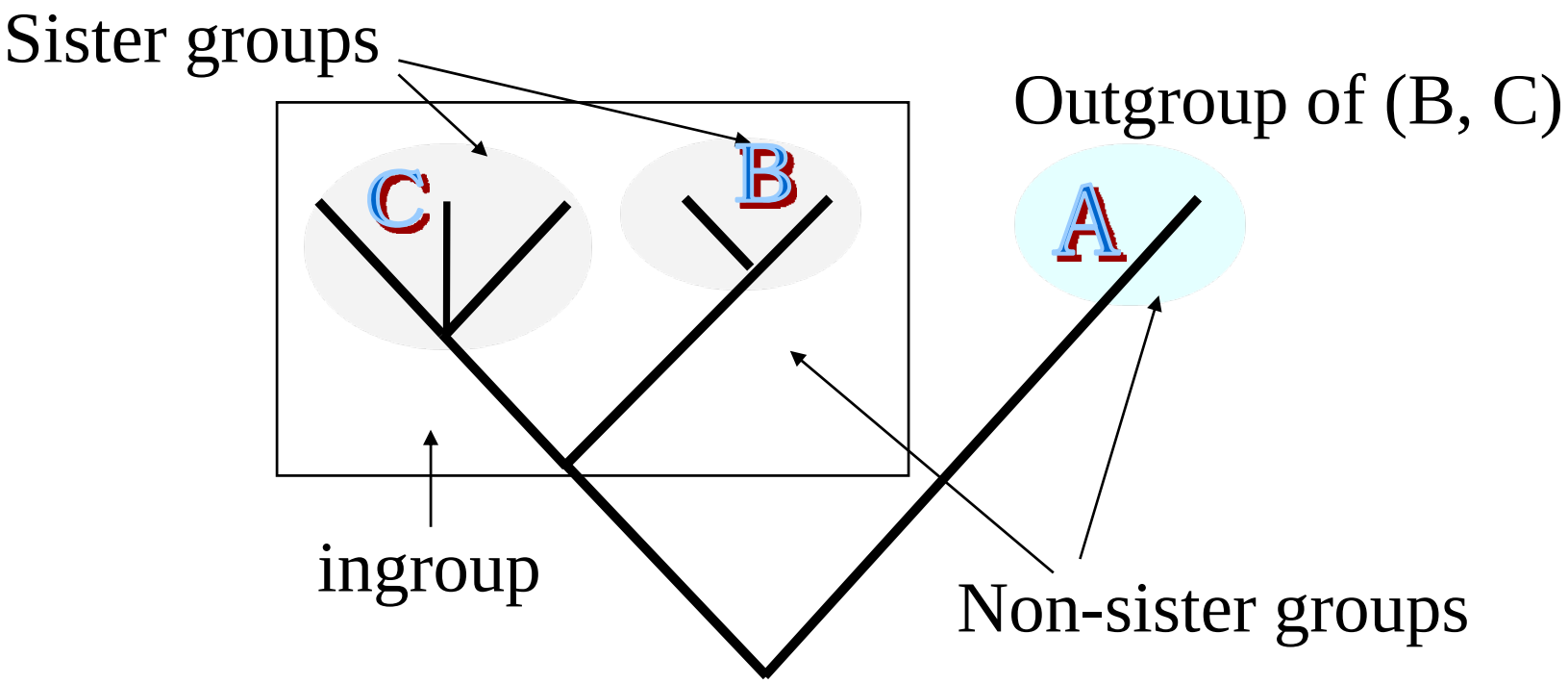
Introduction



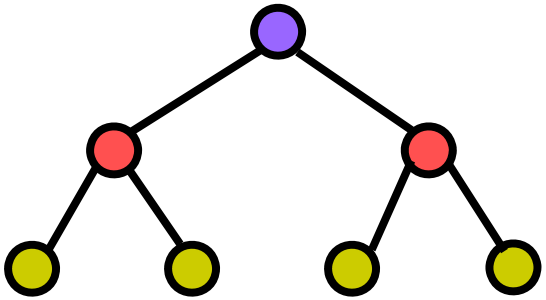
Monophyly & Polyphyly

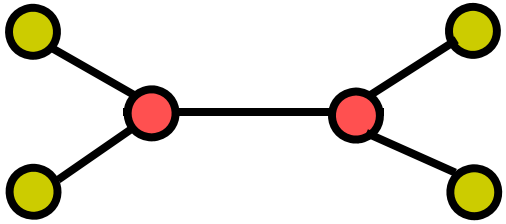


Sister group & Outgroup



Rooted & Unrooted Trees





类群数为m

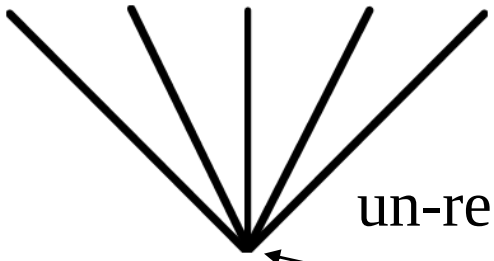
有根二歧树:

$$[(2m-3)!]/[2^{m-2}(m-2)!] \quad (m>=2)$$

无根树：

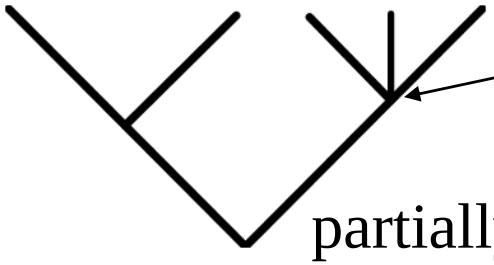
用**m-1**代替上式中的**m**。

Polytomy & Dichotomy



un-resolved relationships

polytomy

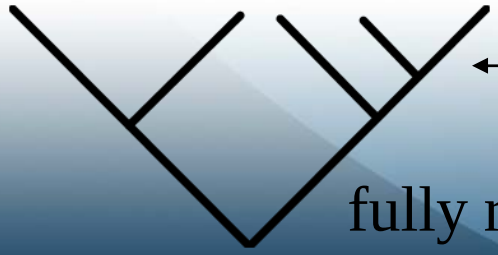


partially resolved relationships

Polytomy: an internal node with more than 2 immediate descendants

"hard" polytomy: simultaneous divergence

"soft" polytomy: unsure order of divergence



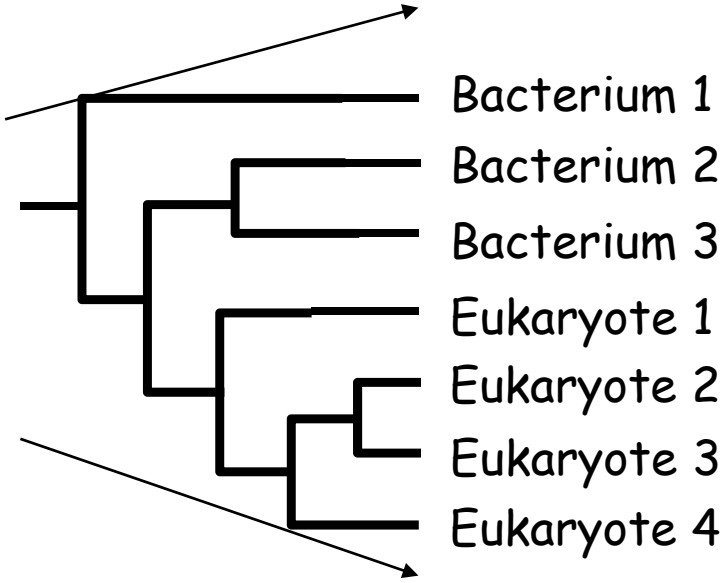
dichotomy

fully resolved relationships

Binary bifurcating tree

Cladograms & Phylograms

Relative Time



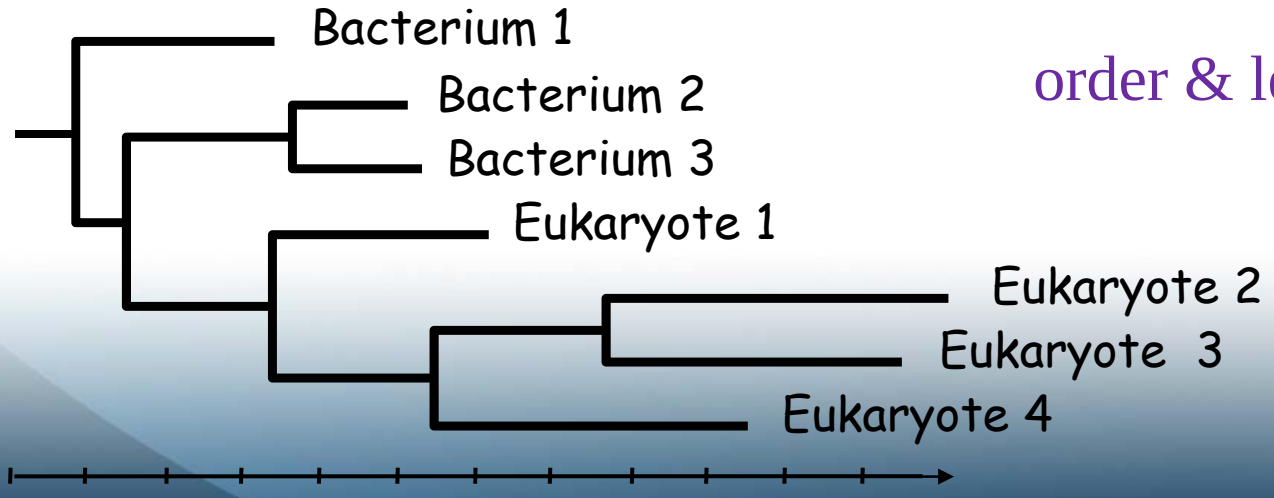
Cladograms:

branching order

Phylograms:

order & lengths

Absolute Time or Divergence



Contents

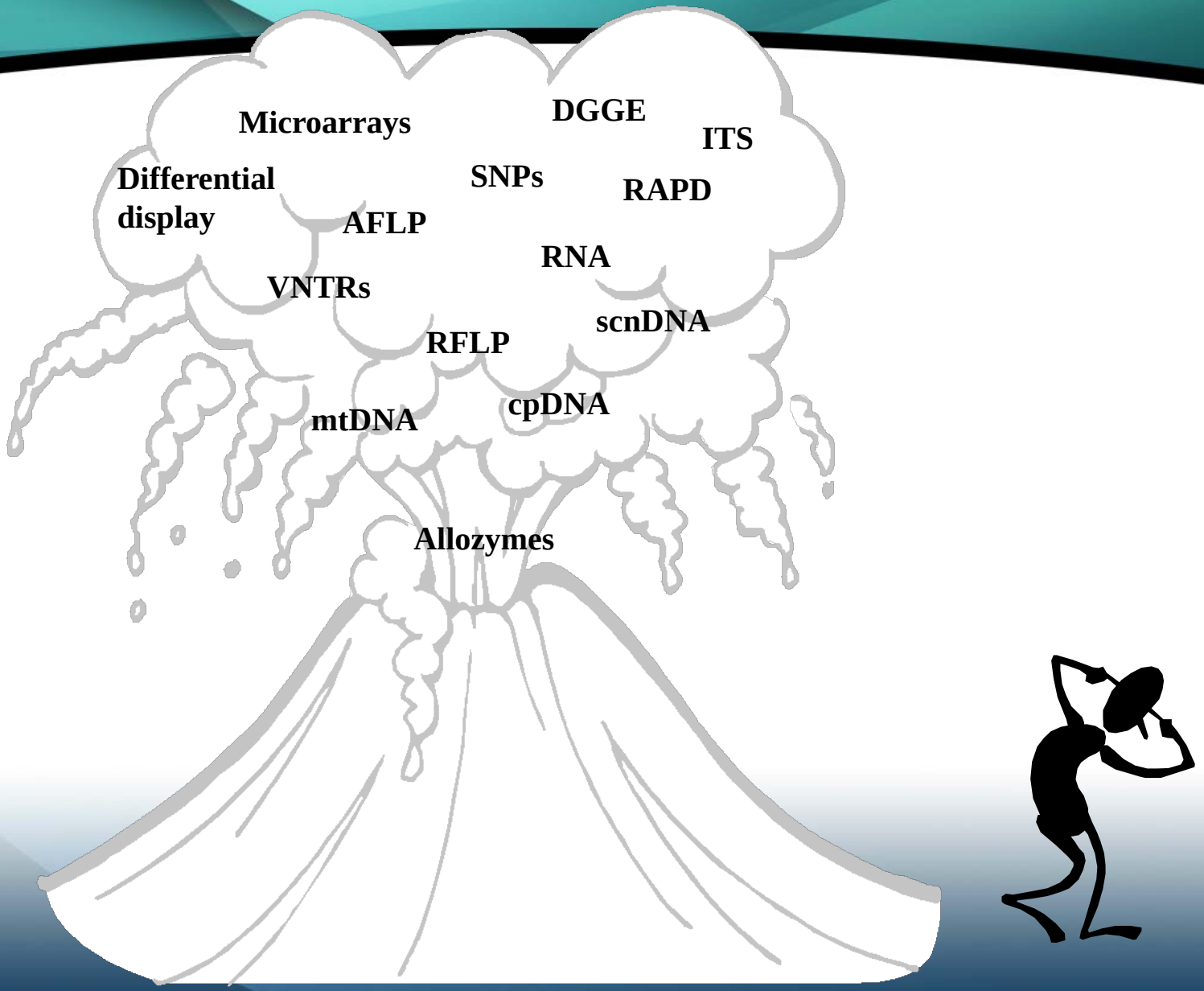
Tree Construction

1. Molecular Markers
2. Sequences Alignment
3. Models of Evolution
4. Tree building
5. Tree evaluation

Choosing Molecular Markers

- **Nucleotide sequence data**
- **Protein sequence data**

Explosion in molecular markers



Choosing Molecular Markers

- **Genome Information**
- **Cytoplasm Information**

  a. chloroplast DNA

  b. mitochondria DNA
- **Nuclear genes Information**

  single/low copy nuclear genes
- **SNP (Single nucleotide polymorphism)**

Sequences Alignment

1、Computer alignment：

Clustal W/X
T-Coffee
bioedit

2、Manuel Adjustment

Choosing Models of Evolution

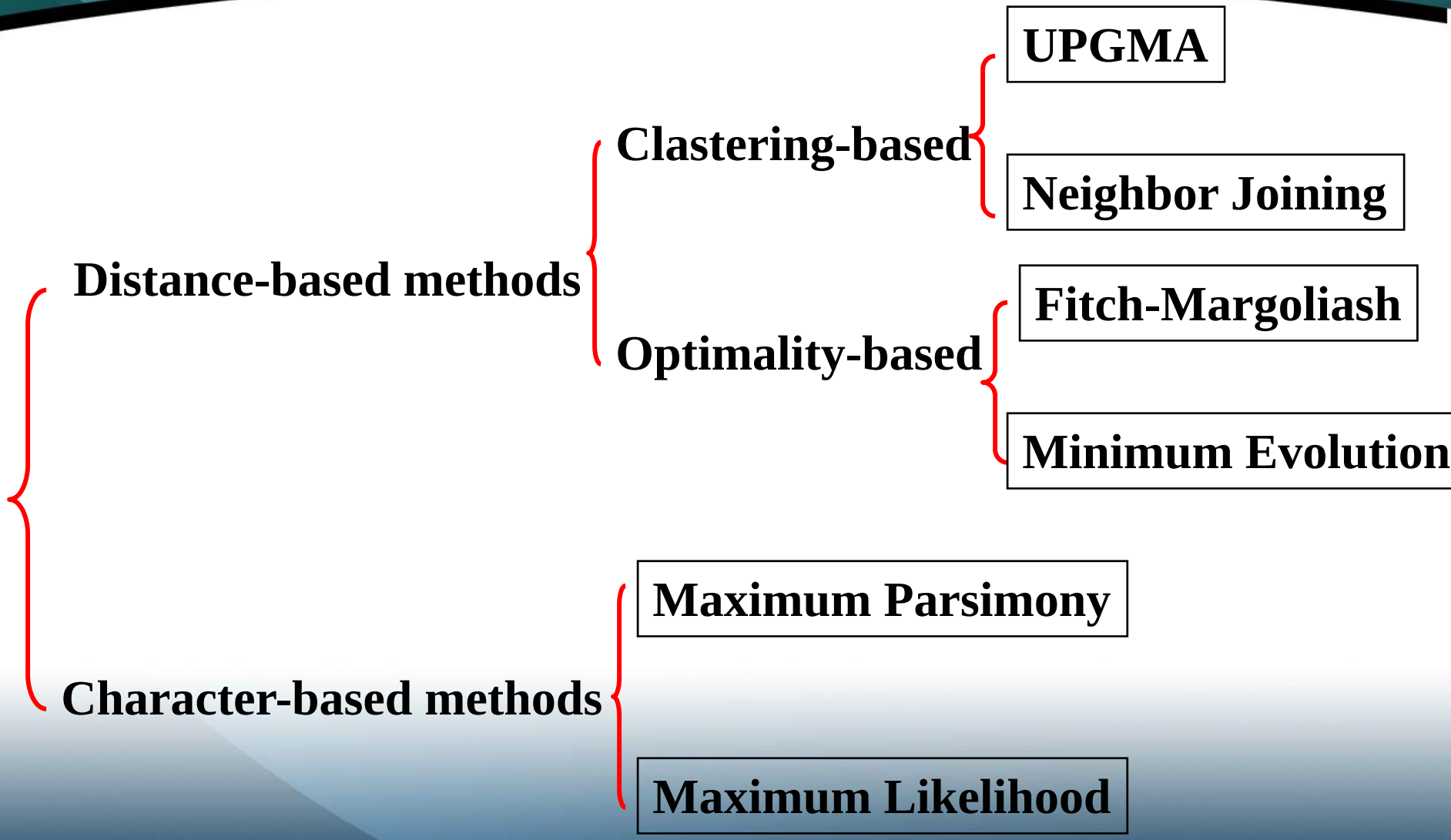
Nucleotide sequence：

Jukes-Cantor Kimura

Protein sequence：

PAM JTT

Tree Building



Tree Building

Assumptions:

Distance-based methods:

1、All sequences involved are homologous.

2、Tree branches are additive.

Character-based methods:

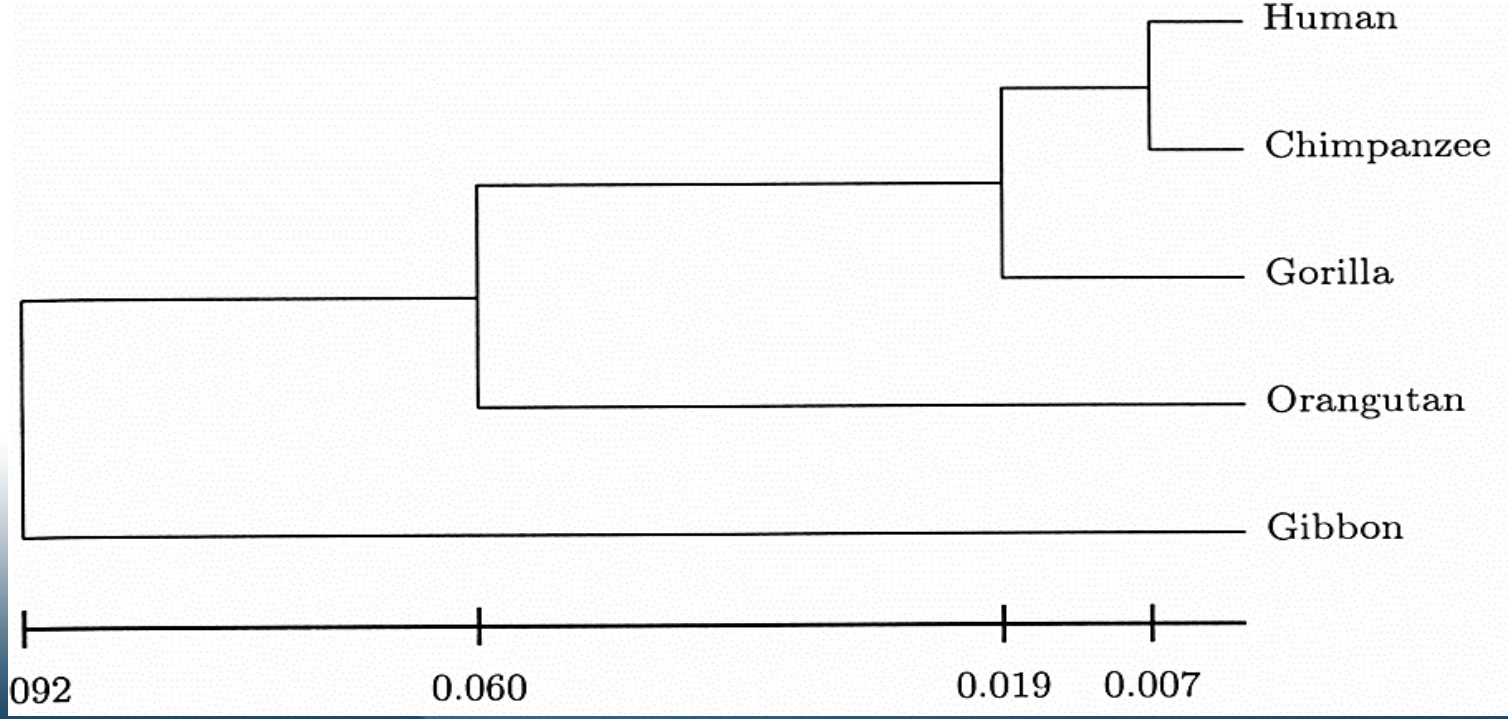
1、Characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved.

2、Each character evolves independently and is treated as an individual evolution unit.

UPGMA

Unweighted Pair Group Method with Arithmetic mean

Closest pair→ Cluster → Average distances



UPGMA

Advantages:

1、Simple and fast method

2、Extensively used in clustering analysis of DNA microarray data.

Disadvantages:

1、Assumes a molecular clock implies a root.

2、 Be susceptible to **Long Branch Attraction.**

Neighbor Joining (NJ)

Neighbor joining (NJ) :

the phylogenetic tree is constructed from a star-like tree by grouping OTUs with shortest distance of branch length together

The best use of an NJ tree is:

As a **starting point** for a model-based analysis such as Maximum likelihood or Bayesian.

Neighbor Joining (NJ)

Advantages:

1、Relatively rapid, it is suitable for analyzing a large dataset.

2、Calculate the branch length.

Disadvantages:

1、Construct only one possible tree.

2、Yield a biased tree under some condition.

3、Compress sequences information.

Optimality-based methods

Optimality-based methods:

Comparing possible tree topologies and select a tree that best fits the actual evolutionary distance matrix.

Fitch-Margoliash(FM) :

Select a best tree based on minimum deviation between the distance calculated in the all branches in the tree and distance in the original dataset.

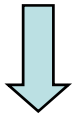
Minimum evolution(ME):

Uses a different optimality criterion that finds a tree among all possible trees with a minimum overall branch length.

Maximum Parsimony

Parsimony is the **principle** that an evolutionary explanation requiring the **fewest steps** should be preferred.

the simpler, the better



tree length becomes our optimality criterion

The best tree is the shortest which requires the least number of substitutions

Maximum Parsimony

Advantage:

1、A simple method and not depend on an explicit model of evolution

2、Gives both trees and associated hypotheses of character evolution

Disadvantage:

- With more data the it becomes increased that parsimony will give the wrong tree due to long branch attraction.

Maximum Likelihood

- Maximum likelihood use statistical tool to evaluate a hypothesis about evolutionary history.

- It constructs all possible trees of evolutionary history from an observed data set.

Maximum Likelihood

Advantages:

1、All the sequence information is used

2、Evaluate all possible trees

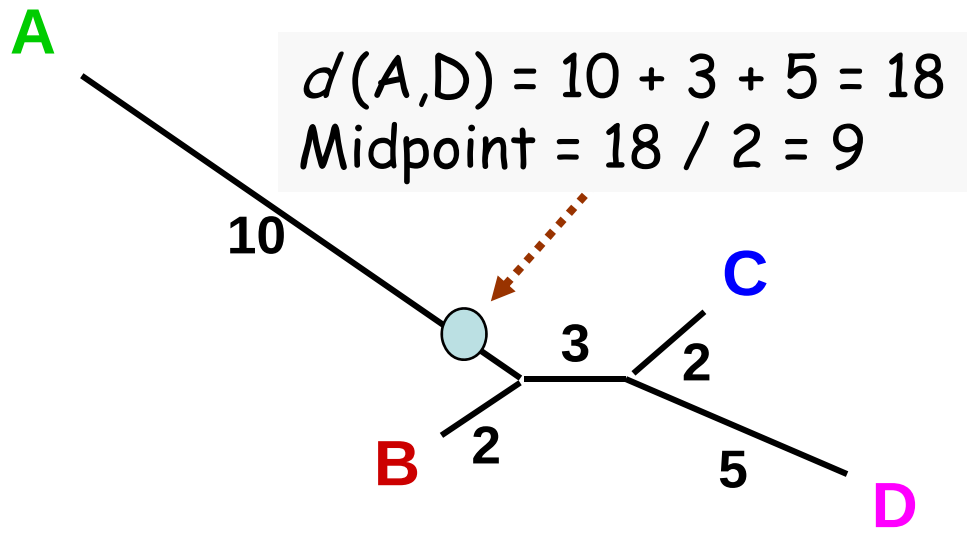
3、Sampling errors have least effect on the method.

Disadvantages:

1、Extremely slow.

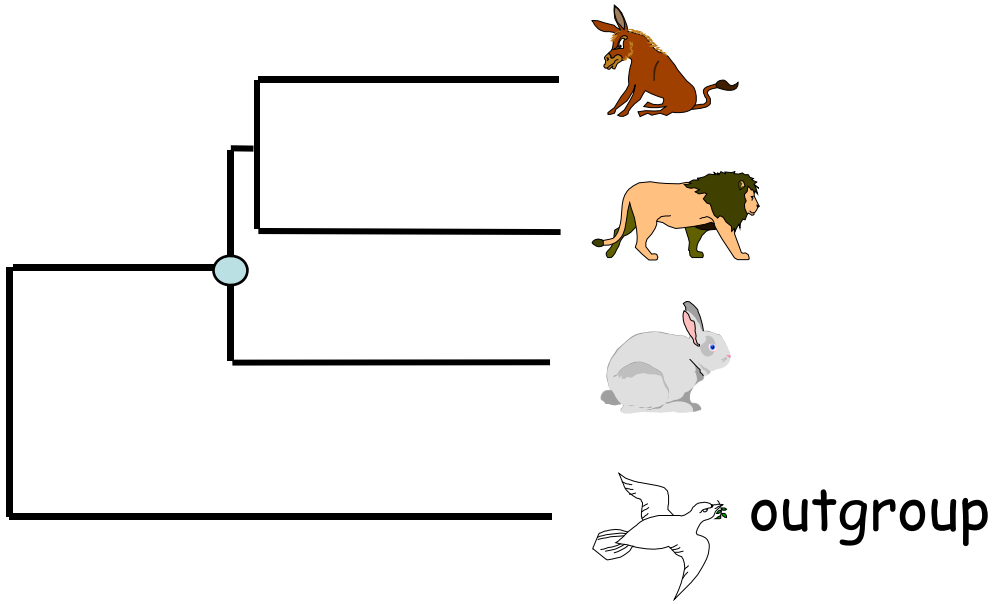
2、Impractical for analyzing large data set.

Rooting the Tree



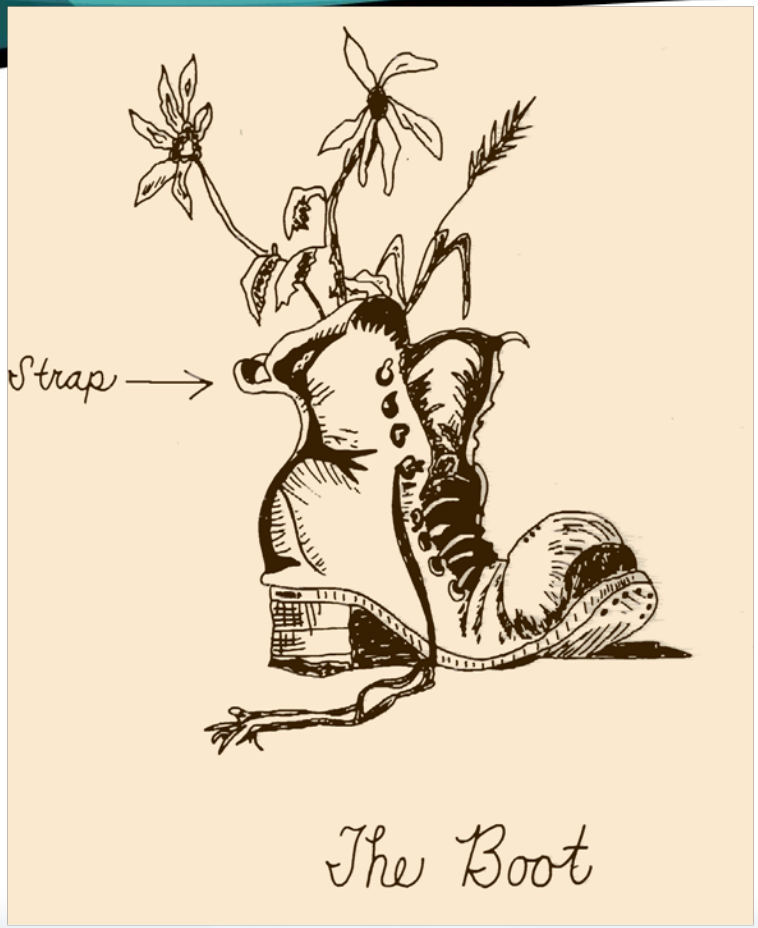
1) **Midpoint rooting:** Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This is used mostly for the distance-based tree building methods.

Rooting the Tree



2) Outgroup rooting: Use “outgroup” taxa that are known to fall outside of the “ingroup” (group under study).

Tree evaluation



A statistical technique that uses computer intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter

Bootstrap is a pseudo-repeats data collecting method to estimated reliability of the tree.

How confident am I that my tree is orrect?

- ☺ **High bootstrap values don't mean that your tree is the true tree!**

- ☺ **Alignment and evolutionary assumptions are key!**

Contents

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	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

Some useful softwares

(1)FORCON——数据格式转换软件
(2)ClustalX——序列比对及系统树构建软件
(3)PAUP——构建系统发育树的基本软件
(4)Treeview——查看以及编辑系统树的软件
(5)Phylip
(6)MEGA
(7)PAML
(8)Mrbayes

演示

Phylip

Phylip有多种不同平台的版本（包括windows，Macintosh，DOS，Linux，Unix和OpenVMX）。

Phylip包含了35个独立的程序，这些独立的程序都实现特定的功能，这些程序基本上包括了系统发生分析的所有方面。

Phylip是目前最广泛使用的系统发生分析程序，主要包括以下几个程序组：分子序列组，距离矩阵组，基因频率组，离散字符组，进化树绘制组。

Phylip

分子序列组:

1. 蛋白质序列: protpars, proml, promlk, protdist;
2. 核酸序列: dnapenny, dnapars, dnamove等。

距离矩阵组: Fitch, kitsch, neighbor

基因频率组: Gendist, contml

离散字符组: Pars, mix, move, penny, dollop等

进化树绘制组: drawtree, drawgram

其他: restdist, restml, seqboot, contrast, treedist, consense, retree。

Phylip软件包的应用

1、选择适当的程序

DNA数据-就在核酸序列分析类中选择程序;

离散数据-如突变位点数据，就在离散字符组里面选择程序。

2、选择适当的分析方法

DNA数据，可以选择简约(DNAPARS),似然法(DNAML)距离法等。

3、进行分析

选择好程序后，执行，读入分析数据，选择适当的参数，进行分析，结果自动保存为outfile，outtree。

实际应用

- 出发数据－已经排列好的氨基酸序列。
- 重构算法－距离法(protdist.exe)
  最大简约法(protpars.exe)
  最大似然法(proml.exe)
- 统计分析－拔靴法(bootstrap)

1、双击执行protdist.exe,根据提示输入分析的文件名。

2、设定各个参数，执行程序，获得距离矩阵数据输出文件outfile。

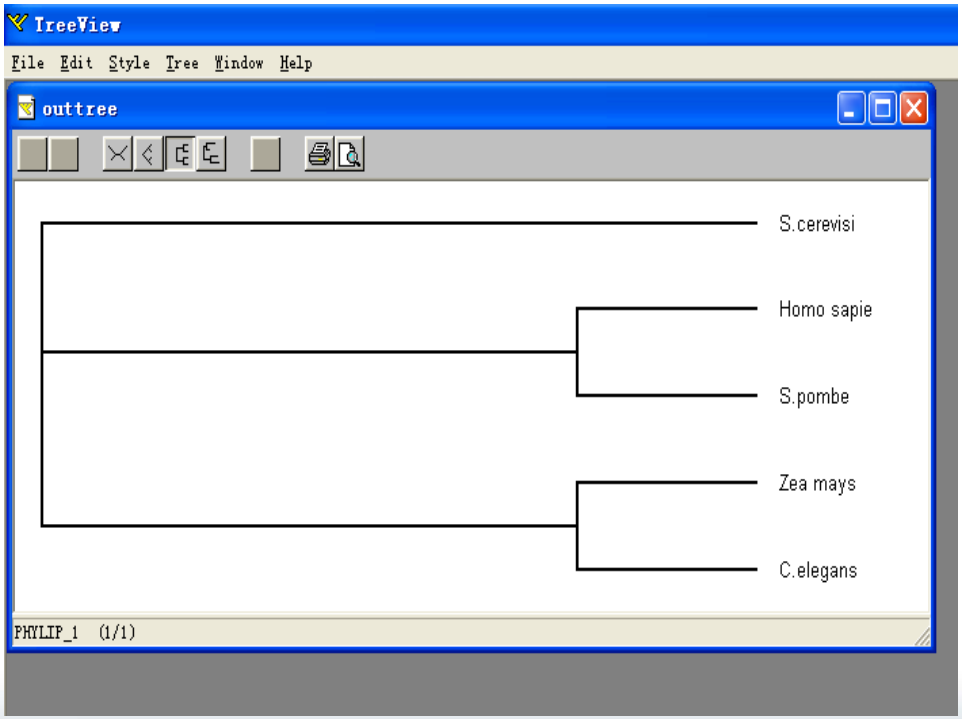
3、选择通过距离矩阵推测进化树的算法(fitch.exe，kotsch.exe，neighbor.exe)。

4、将刚获得的输出文件改名为infile，执行选择的推测算法(neighbor.exe)。设置好参数后执行程序，获得outfile和outtree两个结果输出。

outfile - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
   5 Populations

Neighbor-Joining/UPGMA method version 3.6a3


 Neighbor-joining method

 Negative branch lengths allowed


    +--------------------------Homo_sapie
  +-2
  ! +--------------------------------------S.pombe
  !
  !       +--------------------Zea_mays
  3-------1
  !       +--------------------------------C.elegans
  !
  +--------------S.cerevisi


remember: this is an unrooted tree!

Between        And            Length
-------        ---            ------
   3             2           0.03451
   2          Homo_sapie     1.33969
   2          S.pombe        1.92211
   3             1           0.39696
   1          Zea_mays       1.03167
   1          C.elegans      1.64333
```



outfile

outtree

Bootstrap分析

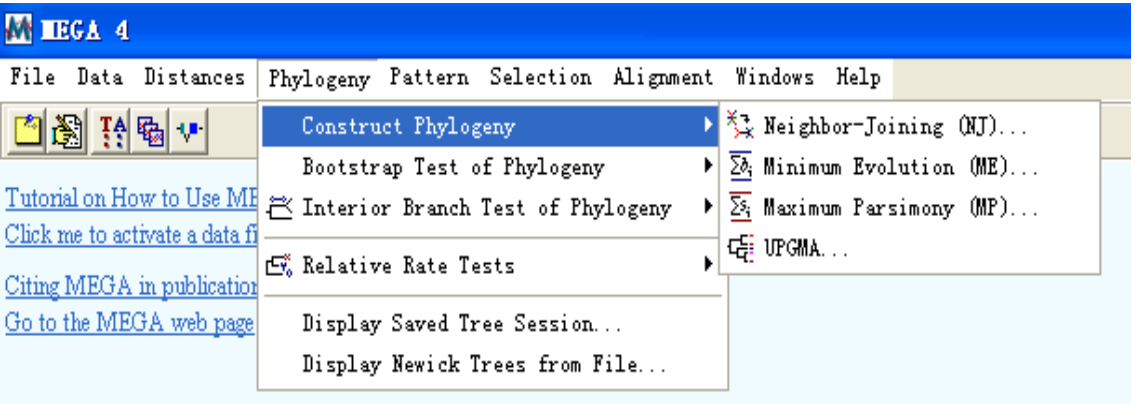
两个用于执行bootstrap分析的程序:

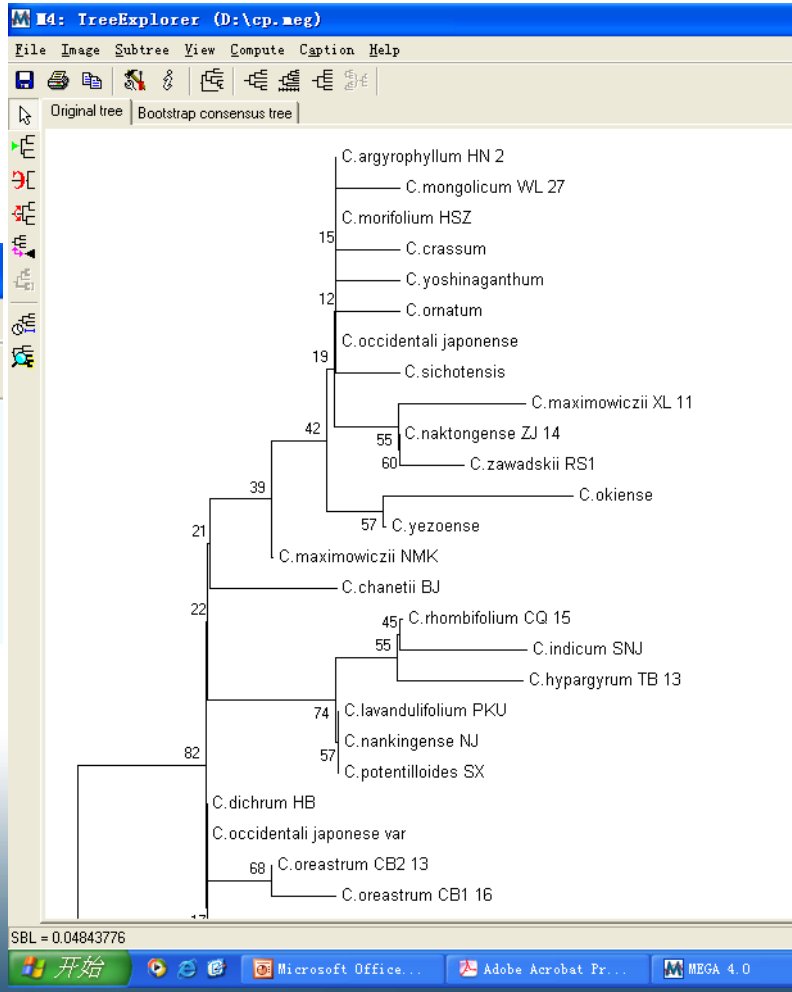
seqboot.exe和consence.exe。

分析过程:

1. Seqboot产生大量的数据组。
2. 应用选择的算法对产生的数据组进行分析。
3. 由consence获得最优树。

MEGA


MEGA 4
File Data Distances Phylogeny Pattern Selection Alignment Windows Help
Construct Phylogeny
Bootstrap Test of Phylogeny
Interior Branch Test of Phylogeny
Relative Rate Tests
Display Saved Tree Session...
Display Newick Trees from File...
Neighbor-Joining (NJ)...
Minimum Evolution (ME)...
Maximum Parsimony (MP)...
UPGMA...
Tutorial on How to Use ME
Click me to activate a data fi
Citing MEGA in publicatio
Go to the MEGA web page


M4: TreeExplorer (D:\cp.meg)
File Image Subtree View Compute Caption Help
Original tree
Bootstrap consensus tree
C.argyrophyllum HN 2
C.mongolicum WL 27
C.morifolium HSZ
C.crassum
C.yoshinaganthum
C.ornatum
C.occidentali japonense
C.sichotensis
C.maximowiczii XL 11
C.naktongense ZJ 14
C.zawadskii RS1
C.okiense
C.yezoense
C.maximowiczii NMK
C.chanetii BJ
C.rhombifolium CQ 15
C.indicum SNJ
C.hypargyrum TB 13
C.lavandulifolium PKU
C.nankingense NJ
C.potentilloides SX
C.dichrum HB
C.occidentali japonese var
C.oreastrum CB2 13
C.oreastrum CB1 16
15
12
19
42
55
60
39
57
21
22
45
55
74
57
82
68
SBL = 0.04843776
开始
Microsoft Office...
Adobe Acrobat Pr...
MEGA 4.0

Phylogenetic Analysis by Maximum Likelihood (PAML)

Ziheng Yang

Table of contents

Introduction

PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood. It is maintained and distributed for academic use free of charge by Ziheng Yang. ANSI C source codes are distributed for UNIX/Linux/MAC OS X, and executables are provided for MS Windows.

This document is about downloading and compiling PAML and getting started. See the manual (pamlDOC.pdf) for more information about running programs in the package.

A summary of the types of analyses performed by different programs in the package is given below.

- **baseml**: ML analysis of nucleotide sequences: estimation of tree topology, branch lengths, and substitution parameters under a variety of nucleotide substitution models (JC69, K80, F81, F84, HKY85, TN93, REV); constant or gamma rates for sites; molecular clock (rate constancy among lineages) or no clock, among-gene and within-gene variation of substitution rates; models for combined analyses of multiple sequence data sets; calculation of substitution rates at sites; reconstruction of ancestral nucleotides.
- **basemlg**: ML analysis of nucleotide sequences under the model of gamma rates among sites. The (continuous) gamma model

Modeltest 3.7

Number of substitution types (nst)

1:

JC	JC+I	JC+G	JC+I+G
F81	F81+I	F81+G	F81+I+G

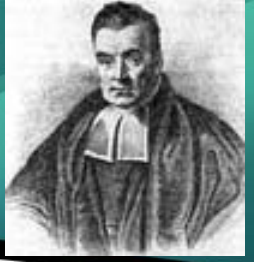
2:

K80	K80+I	K80+G	K80+I+G
HKY	HKY+I	HKY+G	HKY+I+G

6:

TrNef	TrNef+I	TrNef+G	TrNef+I+G
TrN	TrN+I	TrN+G	TrN+I+G
K3P	K3P+I	K3P+G	K3P+I+G
K3Puf	K3Puf+I	K3Puf+G	K3Puf+I+G
TIMef	TIMef+I	TIMef+G	TIMef+I+G
TIM	TIM+I	TIM+G	TIM+I+G
TVMef	TVMef+I	TVMef+G	TVMef+I+G
TVM	TVM+I	TVM+G	TVM+I+G
SYM	SYM+I	SYM+G	SYM+I+G
GTR	GTR+I	GTR+G	GTR+I+G

MrBayes



Prior distribution of tree topologies and uses (MCMC) methods to search tree space and infer the posterior distribution of topologies.

Allows for rate variation among sites and a variety of models of sequence evolution (relativeley free from long branch attraction).

Acknowledge

- **Prof. Luo**

- **Prof. Rao**

Thank You !