Molecular Phylogeney and Phylogeney Tree Construction

分子系统发育学及系统发育树构建

孙田舒

指导老师： 罗静初 教授

特别感谢： 康菊清 博士

北京大学生命科学学院

What is molecular phylogeney

- Phylogeney is the study of the **evolutionary history** of living organisms using tree like diagrams to represent pedigrees of these organisms.

- Molecular phylogeney: The study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules.

Fossil Records Vs. Molecular Approach

- **Fossil records: hard evidence, often biased**

Limitations:

Available for certain species

Limited by abundance, habitat est.

Fragmentary and ambiguous

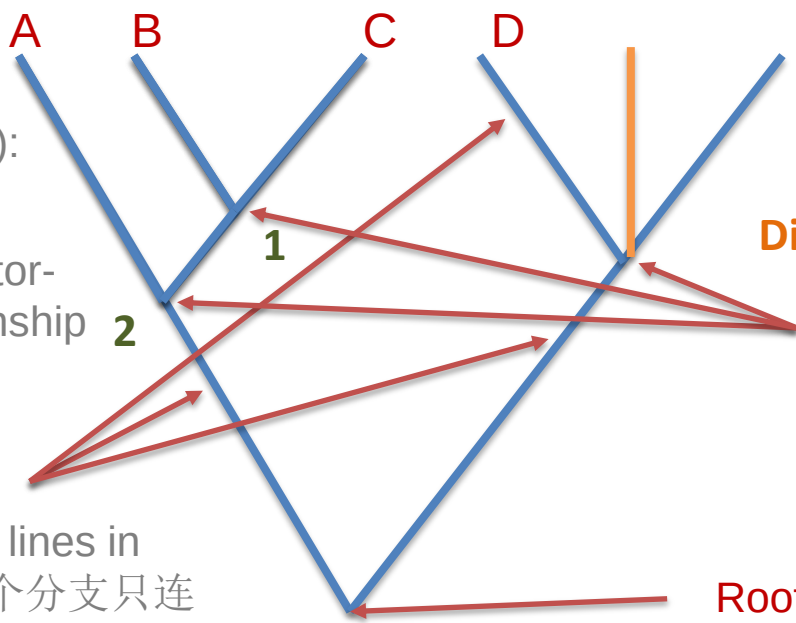
- **Molecular approach: DNA or Protein sequences**

Assumptions:

Molecular sequences are homologous

Each position in a sequence evolved independently

Terminology



Taxon(taxa), OTUs
分类/可操作分类单元,
terminal nodes

Lineage 世系(家系):

the branch path depicting an ancestor-descendant relationship on a tree

Dichotomy or Polytomy

Internal nodes
内部节点
An inferred ancestor of extant taxa.
推断得到的现存分类的祖先

Branches
分支: All the lines in the tree. 一个分支只连接相邻的两个节点。

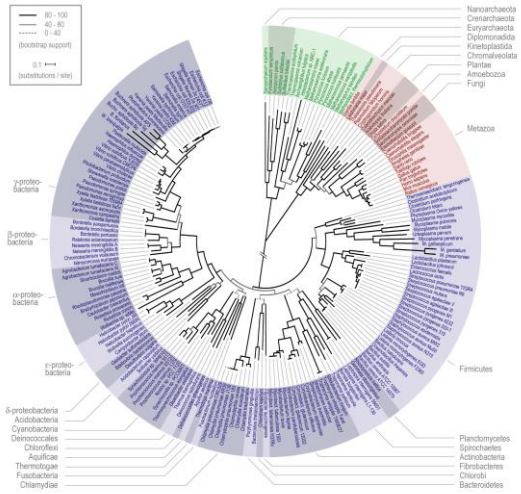
Root node，root 根节点
The common ancestor of all members of the tree.

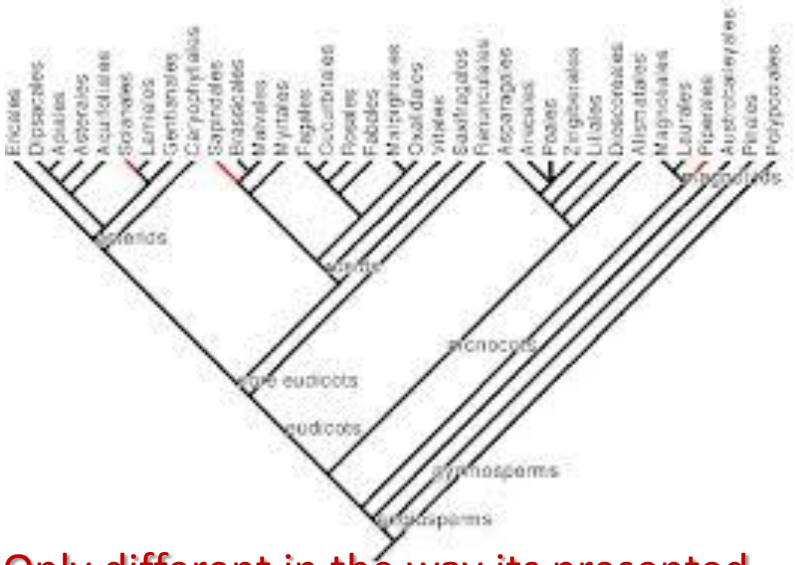
Any way, node is a taxonomic unit, it locates in the bifurcating branch point.

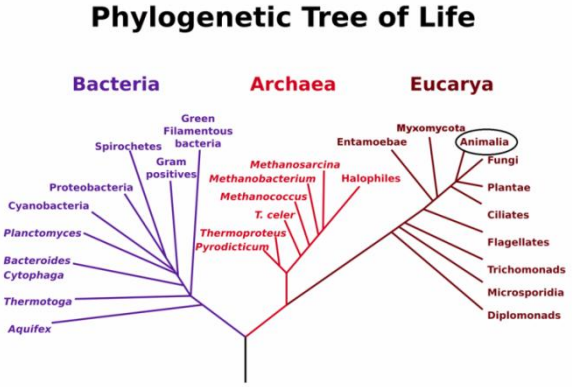
Clades/monophyletics 进化枝/单系类群

A group of taxon consists of a single common ancestor and all its descendants.

Types of phylogenetic trees







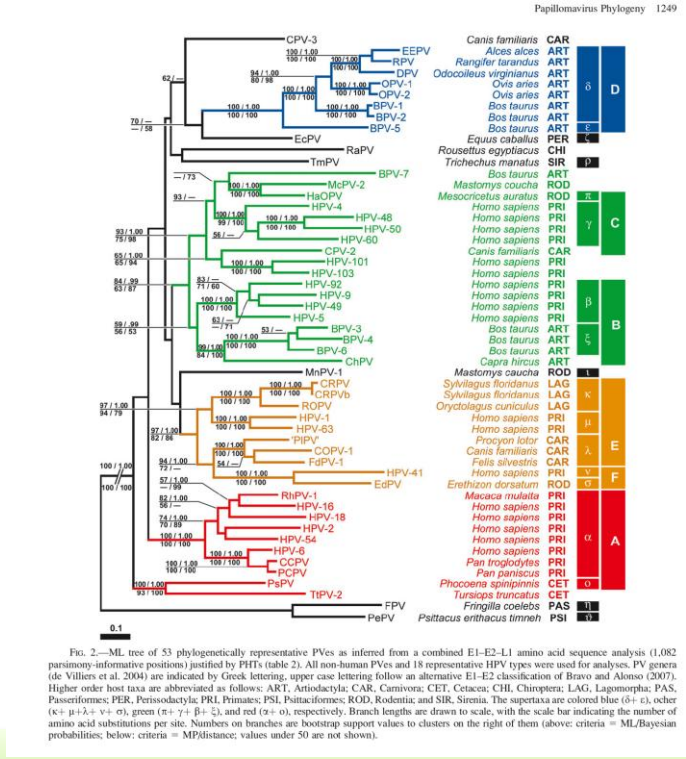
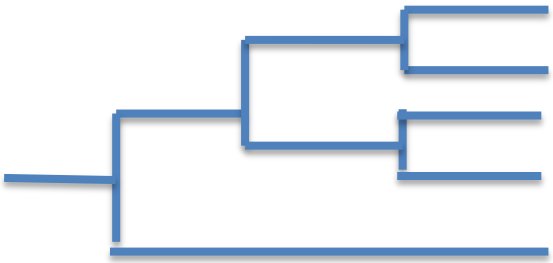


Fig. 2.—ML tree of 53 phylogenetically representative PVes as inferred from a combined E1–E2–L1 amino acid sequence analysis (1,082 parsimony-informative positions) justified by PHTs (table 2). All non-human PVes and 18 representative HPV types were used for analyses, PV genera (de Villiers et al. 2004) are indicated by Greek lettering, upper case lettering follow an alternative E1–E2 classification of Bravo and Alonso (2007). Higher order host taxa are abbreviated as follows: ART, Artiodactyla; CAR, Carnivora; CET, Cetacea; CHI, Chiroptera; LAG, Lagomorpha; PAS, Passeriformes; PER, Perissodactyla; PRI, Primates; PSI, Psittaciformes; ROD, Rodentia; and SIR, Sirenia. The supertaxa are colored blue (δ+ ε), ocher (κ+ μ+λ+ ν+ σ), green (π+ γ+ β+ ξ), and red (α+ ο), respectively. Branch lengths are drawn to scale, with the scale bar indicating the number of amino acid substitutions per site. Numbers on branches are bootstrap support values to clusters on the right of them (above: criteria = ML/Bayesian probabilities; below: criteria = MP/distance; values under 50 are not shown).

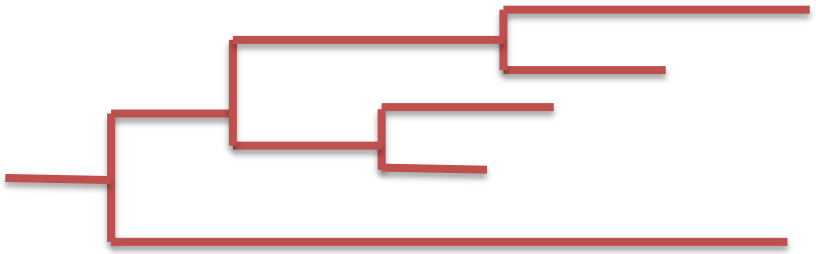
Only different in the way its presented

Unscaled tree VS. Scaled tree

Cladogram 进化分支图



Phylogram 系统发育图

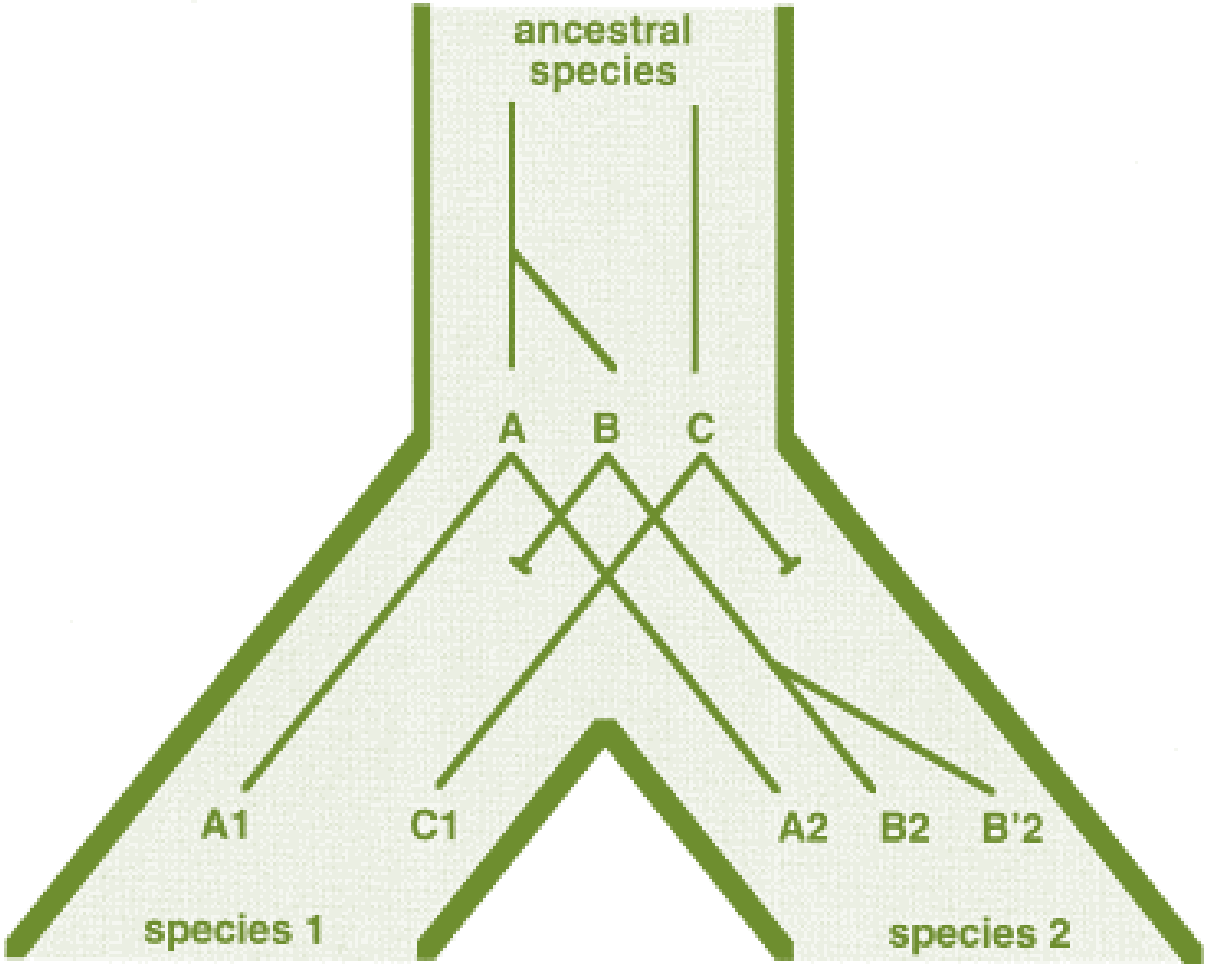


Unscaled trees - all branches in the tree are the same length. (only topology)

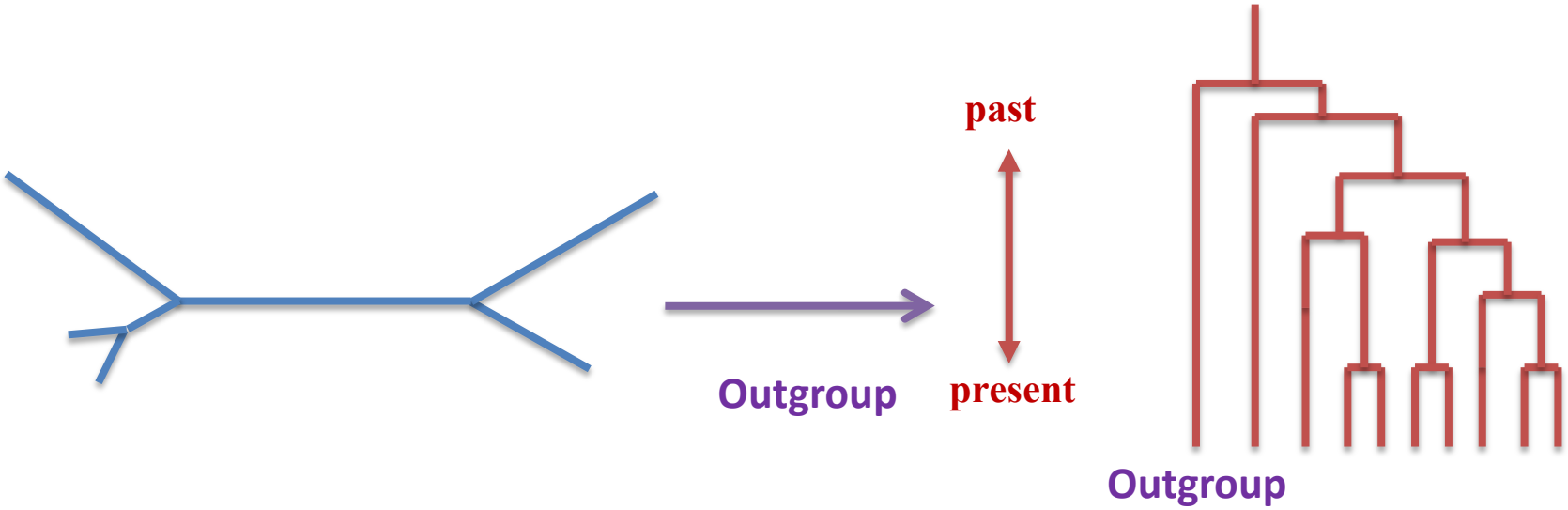
Scaled trees - branches will be different lengths based on the number of evolutionary changes or distance. (branch length & topology)

Difference: **phylogram** is scaled, but **cladogram** not.

Gene tree VS. Species tree



Unrooted tree Vs. rooted tree



Unrooted tree

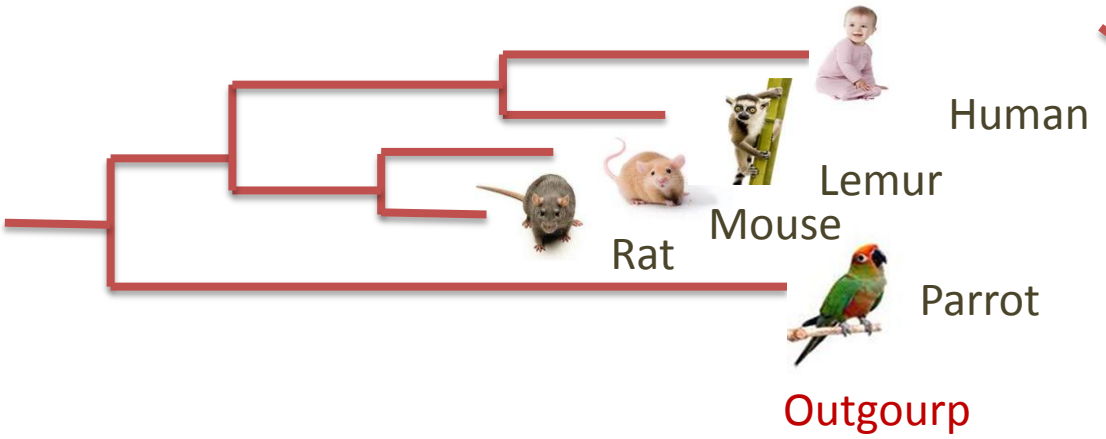
not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships.

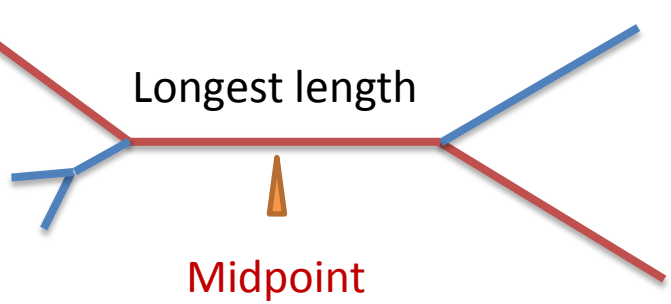
Rooted tree

All sequences have a common ancestor or root note which a unique evolutionary path leads to all other nodes.

Rooting Approach

- Outgroup, is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time.

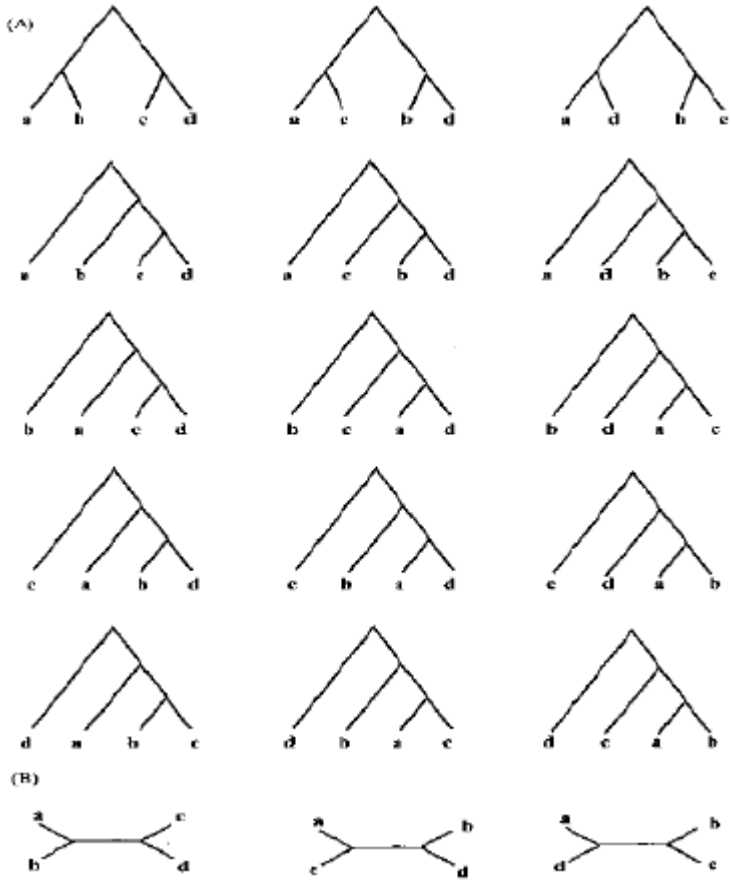




- Midpoint, the midpoint of the two most divergent groups judged by overall branch lengths is assigned as the root.

Finding a true tree is difficult

Species	Unrooted	Rooted
Number	$N_U=(2n-5)!/^{2n-3}(n-3)!$	$N_R=(2n-3)!/^{2n-2}(n-2)!$
4	3	15
5	15	105
6	105	945
10	2027025	34459425
20	2.22×10^{20}	8.20×10^{21}
50	2.84×10^{74}	2.75×10^{78}



Tree construction

① • Choosing molecular marker

② • Performing multiple sequence alignment

③ • Determining a tree building method

④ • Assessing tree reliability

Protein or Nucleotide Sequence

The decision depends on the properties of the sequences and the purposes of the study. Generally, for studying very closed organisms, nucleotide sequences can be used. While if the phylogenetic relationships to be delineated are at the deepest level using protein sequences makes more sense.

- **Protein sequences' characters:**

①More conserved: 61 codons→20AAs
②No different evolutionary rates
③No preferential codon usage
④More sensitive alignment
⑤Gap doesn’t cause frameshift errors

- **Nucleotide sequences' advantages:**

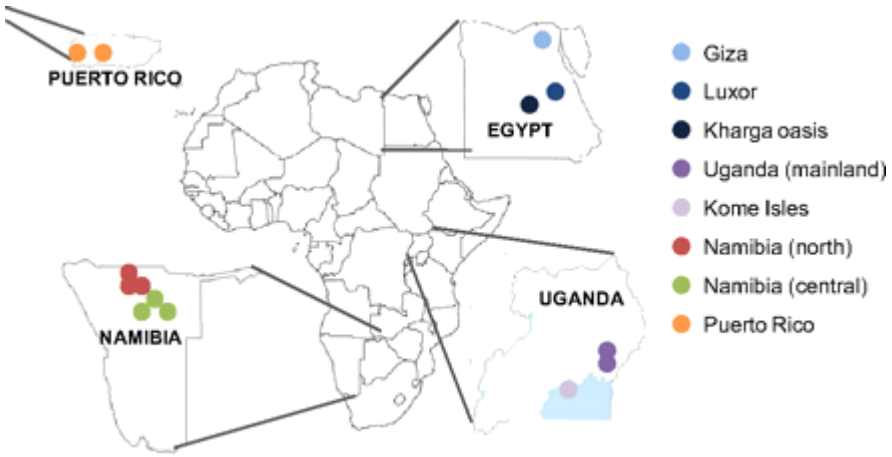
①Rapid evolutionary rates can be informative for closely related sequences.
②depict synonymous and nonsynonymous substitutions, revealing evidence of positive or negative selection.

Origin of Domestic Dogs

- mitochondrial DNA (mtDNA)
- 654 worldwide domestic dogs
- A larger genetic variation in East Asia than in other regions and the pattern of phylogeographic variation suggest an **East Asian** origin for the domestic dog, ~15,000 years ago.

Savolainen et al, 2002, Science

Controversy and Resampling

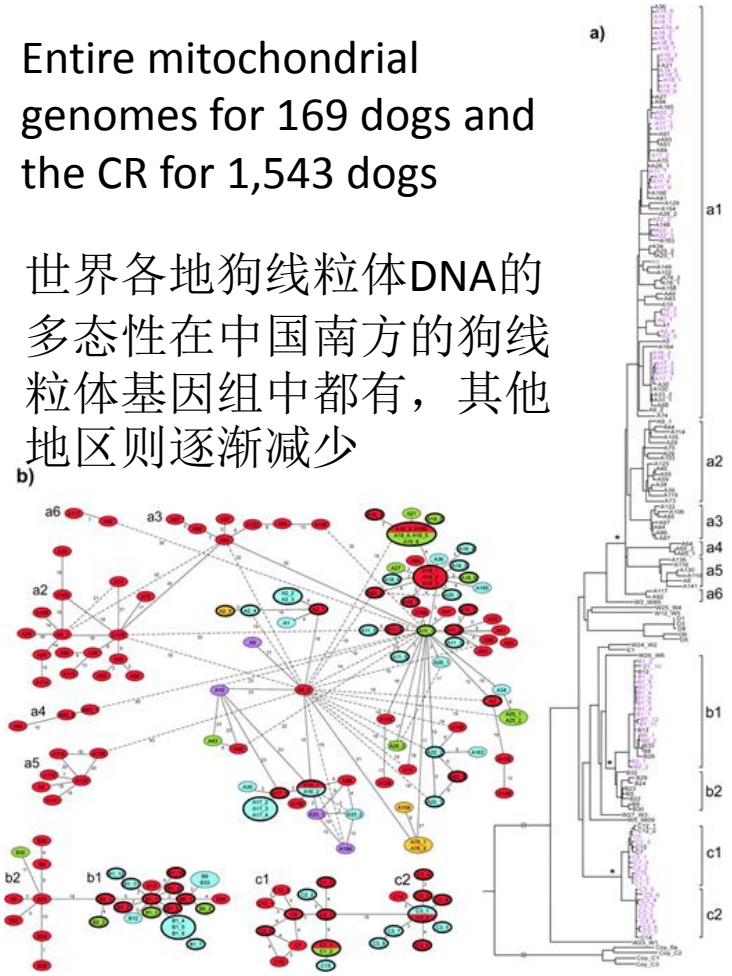


Surprisingly, we find similar mtDNA haplotype diversity in African and East Asian village dogs, potentially calling into question the hypothesis of an East Asian origin for dog domestication.

Boyko et al, 2009, PNAS

Entire mitochondrial genomes for 169 dogs and the CR for 1,543 dogs

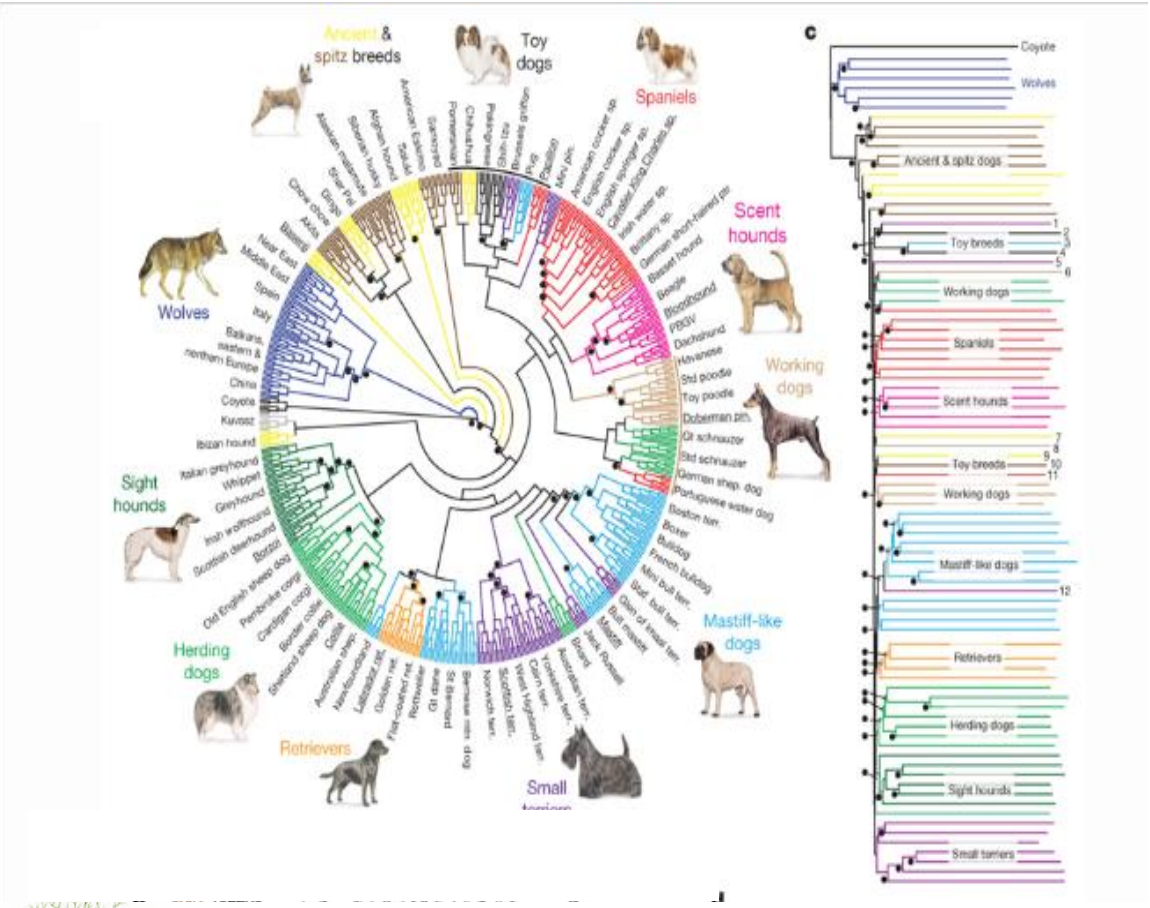
世界各地狗线粒体DNA的多态性在中国南方的狗线粒体基因组中都有，其他地区则逐渐减少



Pang et al, 2009, Molecular Biology and Evolution

狗起源于约一万多年前中国南方的几百只大灰狼

Genome-wide SNP and haplotype analyses



Extensive genome-wide survey of more than 48,000 single nucleotide polymorphisms in dogs and their wild progenitor, the grey wolf.

Middle Eastern wolves were a critical source of genome diversity, although interbreeding with local wolf populations clearly occurred elsewhere in the early history of specific lineages.

Bridgett M. vonHoldt et al, 2010, Nature

Alignment

Alignment may be the most critical step because it establishes positional correspondence in evolution.

- **What can you do?**

①Manual editing: correcting mismatching of key cofactor residues and residues of similar physicochemical properties

②Full alignment or parts of it (domain only)

③Remove ambiguously aligned regions(subjective process)

④Automatic approach: Rascal, NorMD and Gblocks

⑤Statistical models to correct homoplasy

⑥Using a γ correction factor to correct site-dependent rate variation

- **Choosing substitution models**

①Nucleotide: Juke-Cantor Model and Kimura Model

②Protein: PAM or JTT amino acid substitution matrix

Phylogenetic Tree Construction Method

NJ

邻接法Neighbor-joining

- 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

MP

最大简约法(Maximum Parsimony)

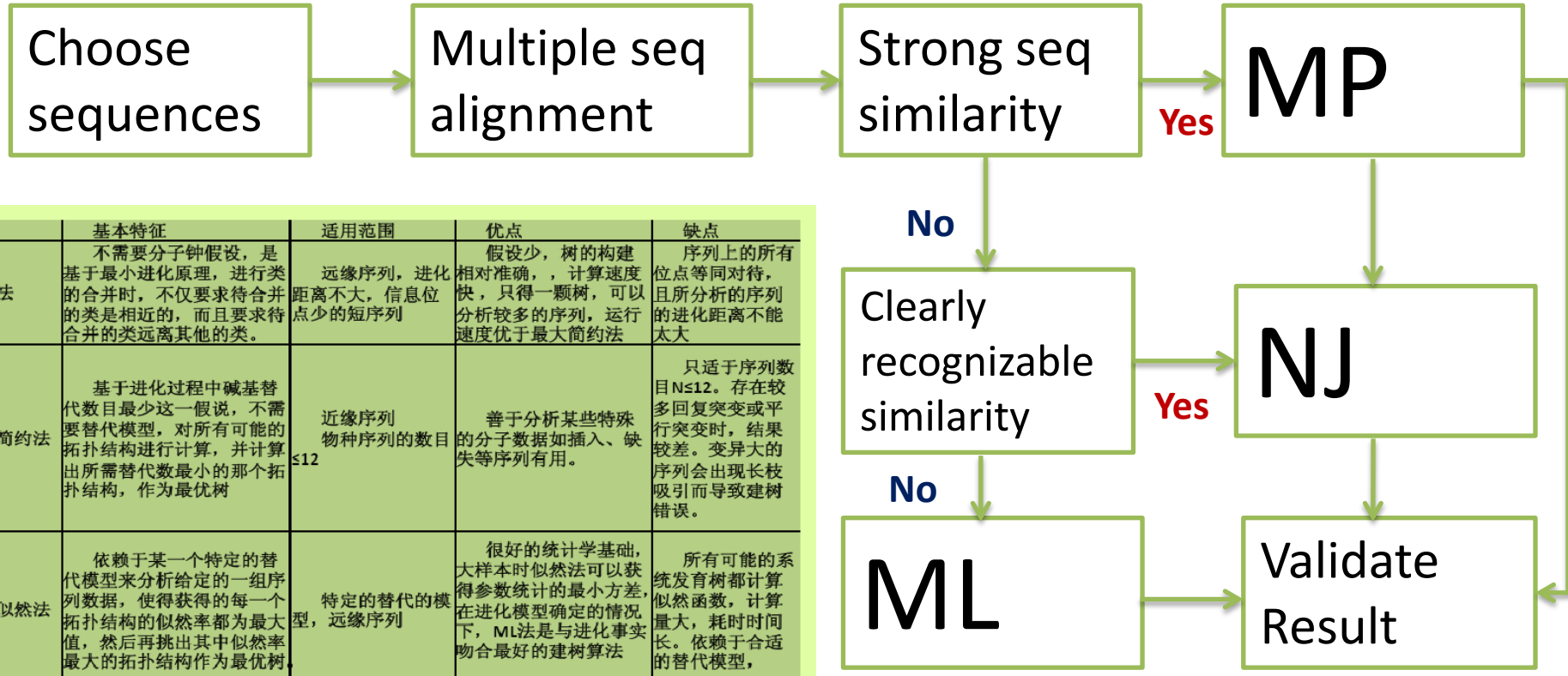
- 最大简约法根据序列的多重比对结果，对所有可能正确的拓扑结构进行计算并挑选出所需替代数最少的拓扑结构作为最优树，即能够利用最少的步骤去解释多重比对中的碱基差异。

ML

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。

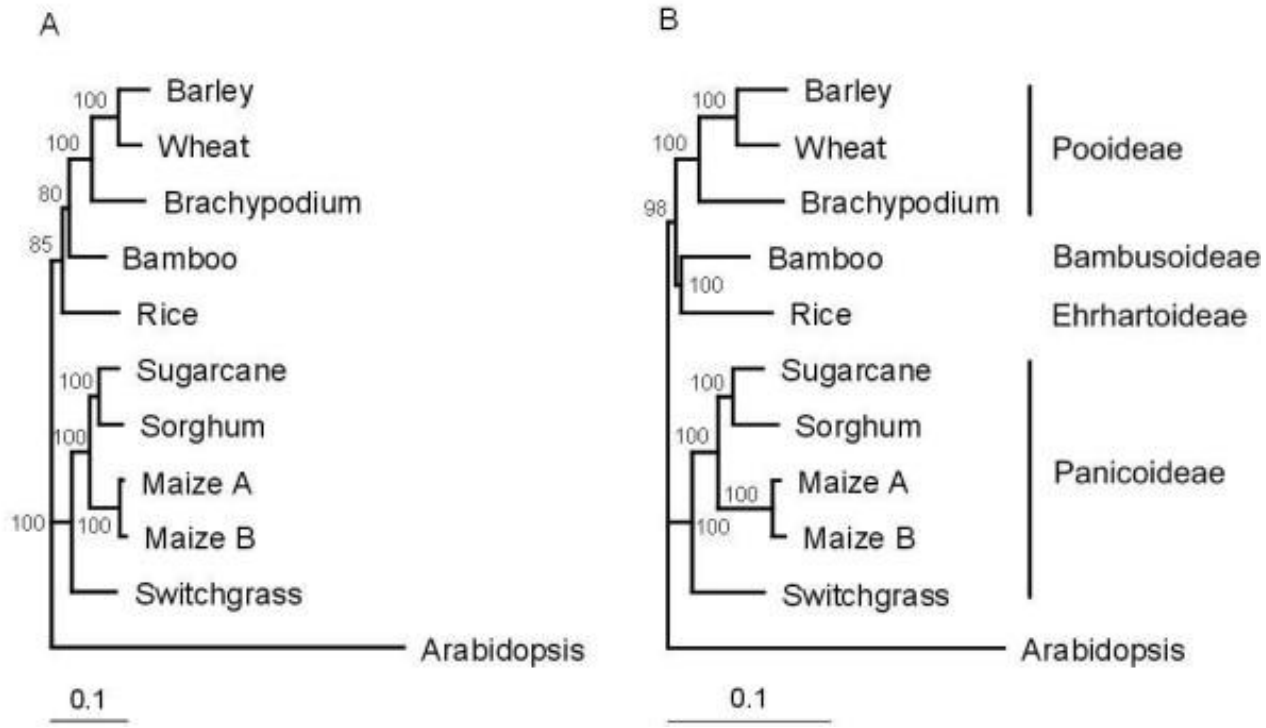
General method



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，

Combine Different Methods for Consensus

Genome-wide characterization of the biggest grass, bamboo, based on 10,608 putative full-length cDNA sequences



Phylogeny of grasses inferred from concatenated alignment of 43 putative orthologous cDNA sequences. (A) Tree inferred from maximal likelihood method. Bayes inference yielded the same topology. (B) Tree inferred from neighbor joining method. Branch length **...**

Tree Evaluation

Boot strapping
自举法

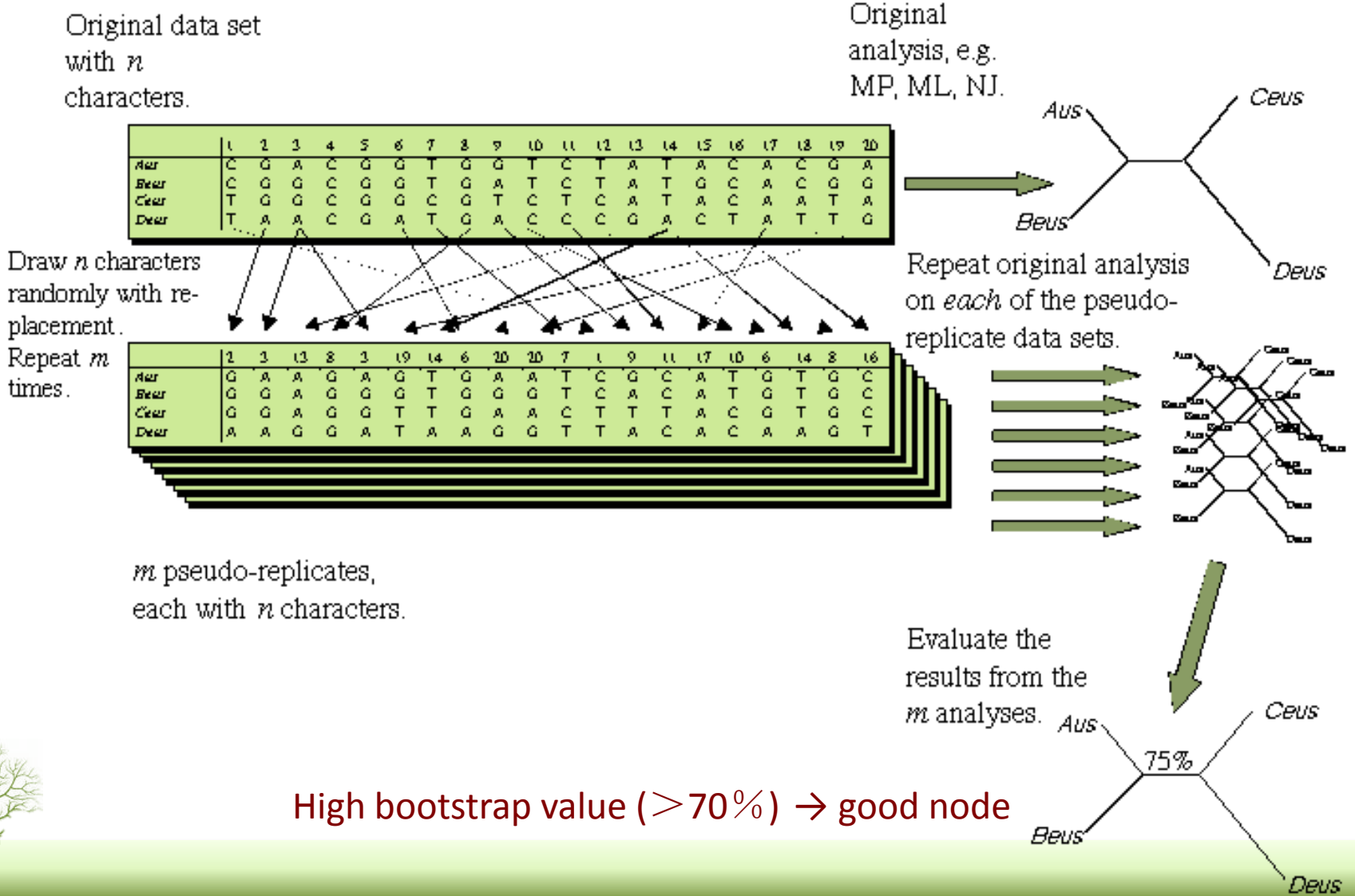
- 是对所比较序列上的替换位点作多次随机取样，根据每次取样的数据可以得到新的树形图，相同的组合出现在某一个节点上的次数占总取样次数的百分比就是该节点的bootstrap值。

Jackknifing
折刀法

- Half of the sites in a dataset are randomly deleted, each new dataset is subjected to phylogenetic tree construction using the same method as the original.

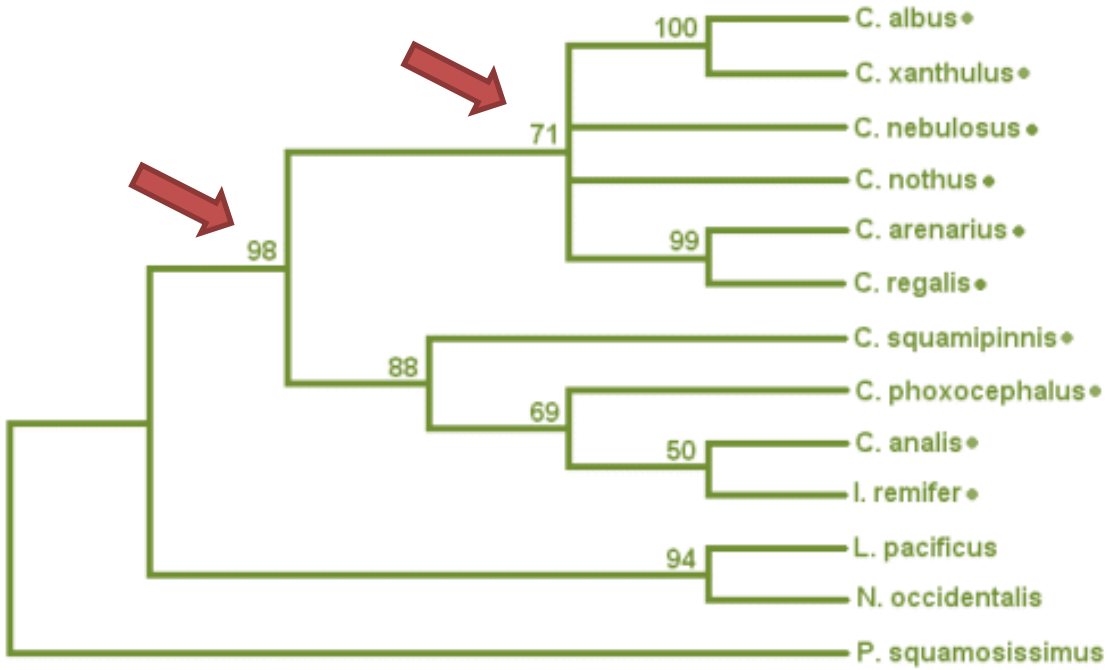
Bootstrapping

To verify the solidity of each node



High bootstrap value ($>70\%$) → good node

Bootstrap Value



Phylogenetic Tree Construction Programs

MEGA	• Molecular Evolutionary Genetics Analysis
PHYLIP	• PHYLogeny Inference Package
PAUP	• Phylogenetic Analysis Using Parsimony
PAML	• Phylogenetic Analysis by Maximum Likelihood

http://evolution.genetics.washington.edu/phylip/software.html

392 phylogeny packages and 54 free web servers

Who needs molecular phylogeny tree

- Functional genomic
- Stress/pathway related
- Species

Sugar transporters for intercellular exchange and nutrition of pathogens

Phylogenetic tree for SWEETs from different species.

Phylogenetic tree of the SWEET superfamily (PFAM PFO3083). Distances were calculated from a multiple sequence alignment (ClustalW) using the neighbor-joining method. The tree displays bootstrap values (percentage of 1000). SWEET genes fall into four clades. All sequences were obtained from NCBI or the Aramemnon database.

The family members were color-coded indicating the species:

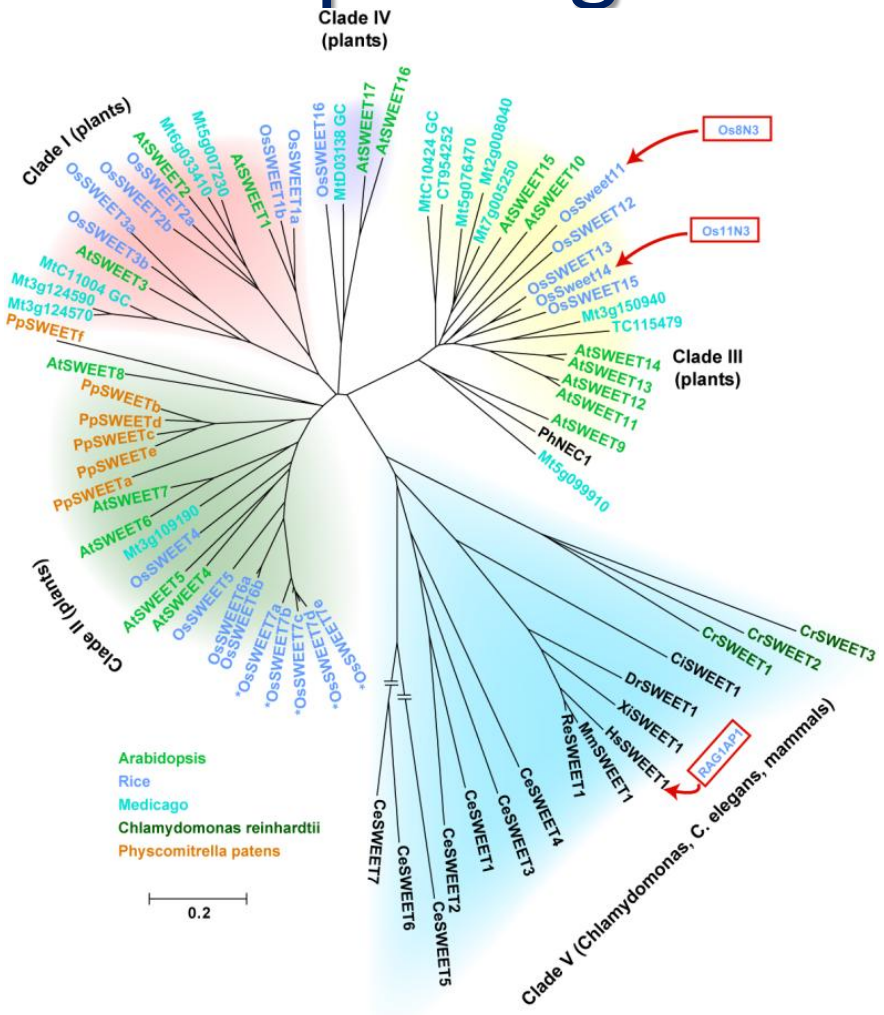
At, Arabidopsis thaliana (light green);

Os, Oryza sativa (blue);

Mt, Medicago trunculata (cyan);

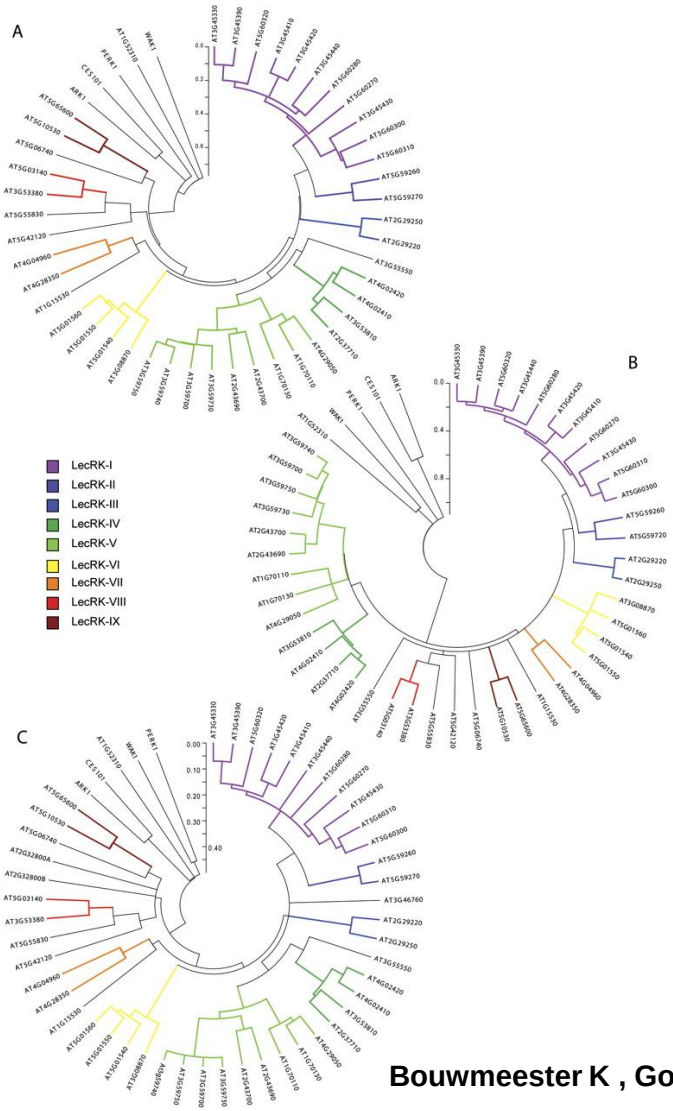
Chlamydomonas reinhardtii (darkgreen),

Physcomitrella patens (orange).



Li-Qing Chen, et al. (2010)

Phylogenetic analysis and classification of Arabidopsis LecRK proteins.



Phylogenetic analysis and classification of Arabidopsis LecRK proteins. Phylograms of
(A) 43 full-length LecRK amino acid sequences,
(B) 43 lectin domains and,
(C) 46 kinase domains, including the kinase domain of LecRK-S.2 and the two kinase domains of LecRK-S.3. Each LecRK clade is a depicted by a different colour.

Bouwmeester K , Govers F J. Exp. Bot. 2009;60:4383-4396

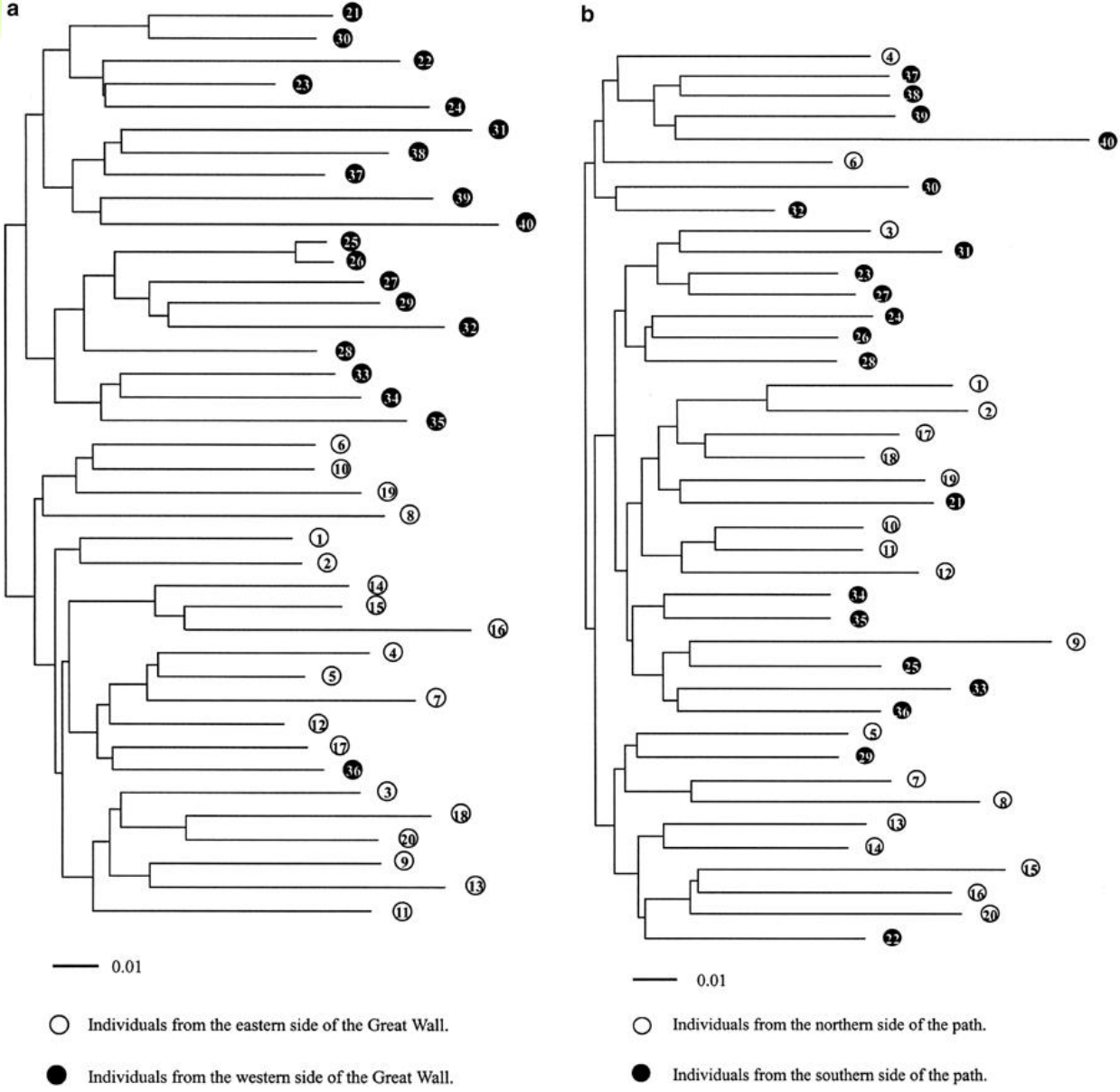
JOURNAL OF Experimental Botany

Application to the real world

- Satisfying One's Curiosity
- Solving Crimes

The Great Wall of China: a physical barrier to gene flow?

Therefore, it is reasonable to deduce that the Juyong-guan Great Wall has served as a physical barrier to gene flow between subpopulations separated for more than 600 years.



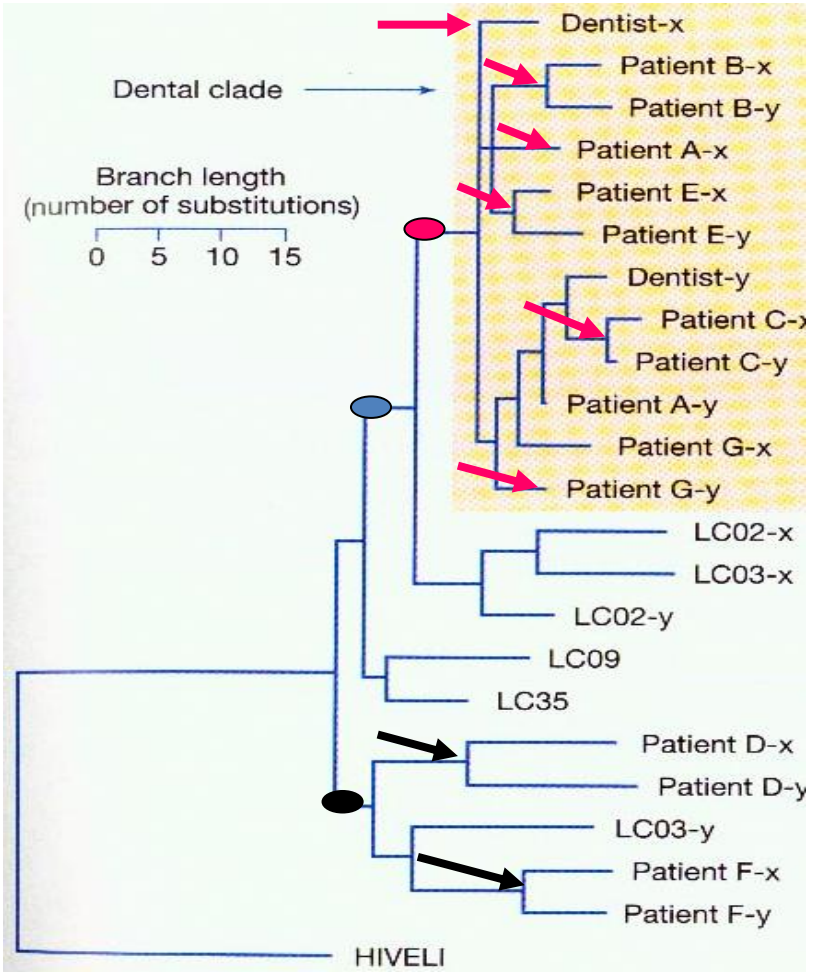
(a) Neighbour-joining dendrogram of 40 P. armeniaca individuals from the Great Wall site. (b) Neighbour-joining dendrogram of 40 P. armeniaca individuals from the control site.

20世纪90年代美国的一件诉讼案

Florida的一位牙医**80**年代末得了艾滋病，他的一位病人**90**年代初被诊断为**HIV**阳性

他的病人均去检查，又发现了几个**HIV**阳性，便起诉牙医

美国**CDC**用牙医及其病人体内不同时间分离的**HIV**基因序列建树



结论：

- 有两位病人从其他来源感染了**HIV**
- 有**5**位病人从牙医那感染了**HIV**（同时考虑了其他因素）

该事件促使医院加强保护措施

The most important step: picking up sequences

- Be aware:
- 1.All the sequences are extracted from reliable source and correct
- 2.All the sequences are **orthologous**
- 3.Choose the right region
- 4.Random evolution
- 5.Each position evolve independently

Thank You!

Sun Jianshu