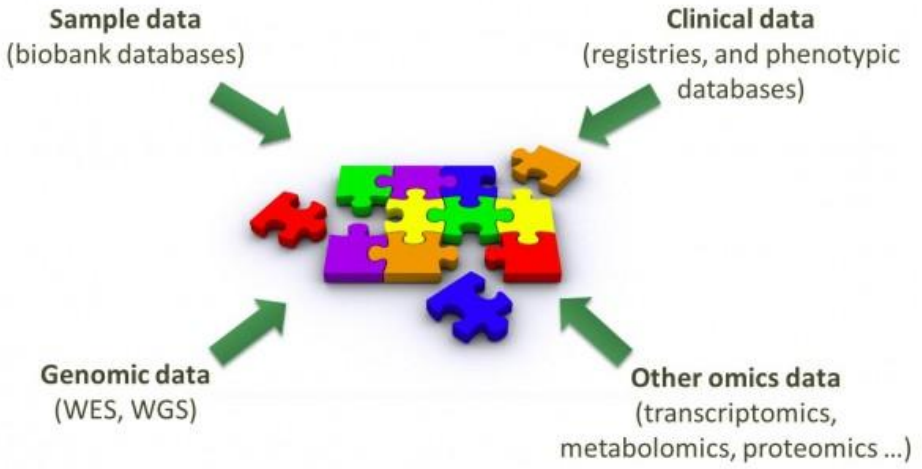
生物信息学：
组学时代的生物信息数据挖掘和理解

2020年秋



有关信息

- 授课教师：宁康，张礼斌，陈鹏
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927

- 课程网页
  - http://www.microbioinformatics.org/Bioinformatics.html
  - QQ群：

课程安排

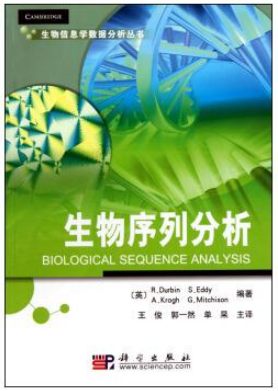
（生物信息中的算法设计与概率统计模型）

- 生物背景和课程简介
- 生物信息学和生物数据挖掘
  - 生物数据的格式及其意义
    - 序列数据
    - 树状数据
    - 网络数据
    - 表达数据等
  - 生物数据库及其用法
  - 生物信息基本算法
    - 双序列联配
    - 多序列联配
    - 基因组组装算法
    - 基因预测和功能注释
    - 系统发育树构建
    - 蛋白质结构预测
    - 生物调控网络解析
  - 组学数据分析方法
    - 基因组变异分析
    - 基因表达和比较分析
    - 非编码RNA分析
    - 蛋白组分析
    - 宏基因组分析
  - 系统生物学与交叉科学
- 面向生物大数据挖掘的深度学习

研究对象：
生物序列，
进化树，
生物网络，
基因表达
…

方法：
生物计算与生物信息

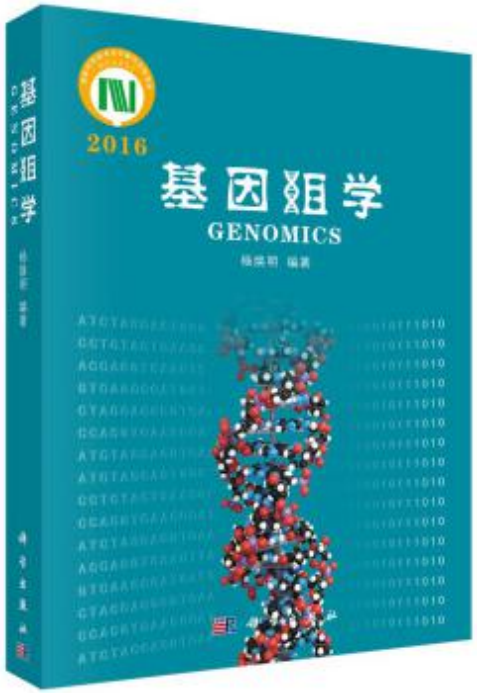
教材及参考书目

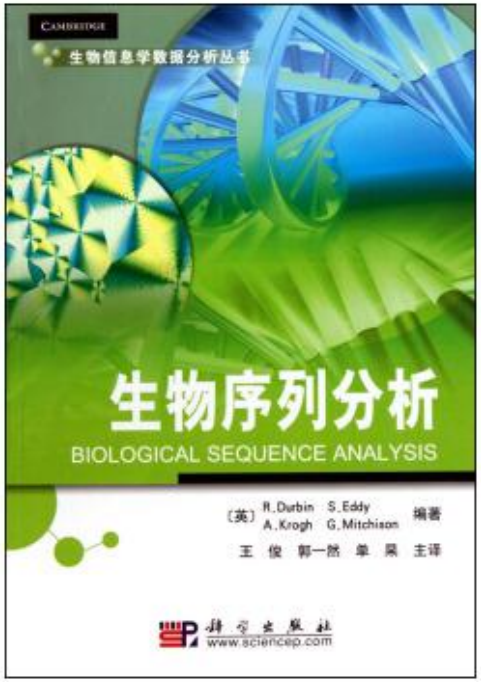


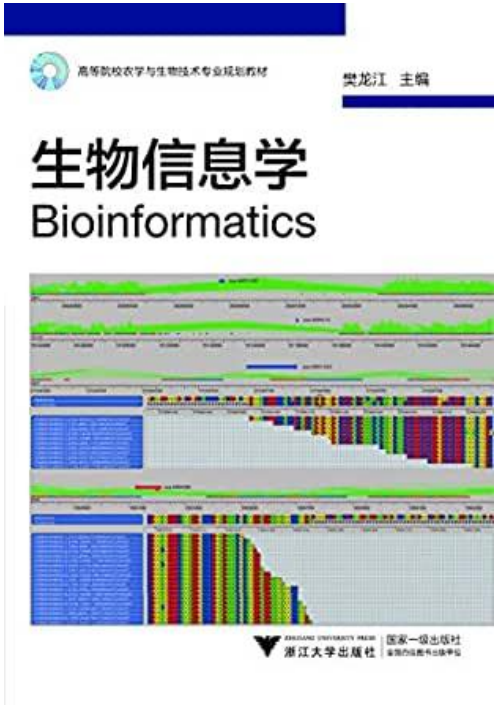
- **教学参考书：**
- 《生物序列分析》（第1版）.科学出版社. 2010年8月出版. R. Durbin等编著，王俊等主译.

- **课外文献阅读：**
- 《生物信息学》（第1版）. 浙江大学出版社. 2017年3月出版. 樊龙江主编.
- 《基因组学》（第1版）. 科学出版社. 2016年10月出版. 杨焕明主编.

References


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单渠 主译
www.sciencep.com


樊龙江 主编
生物信息学
Bioinformatics
浙江大学出版社

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

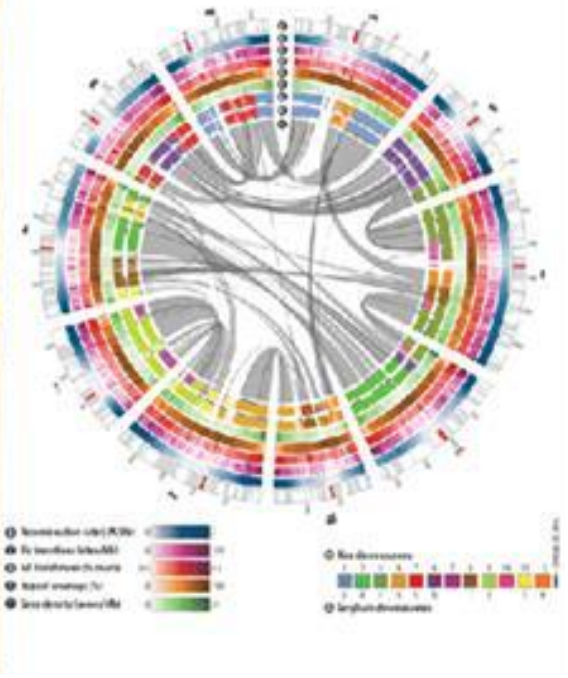
——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

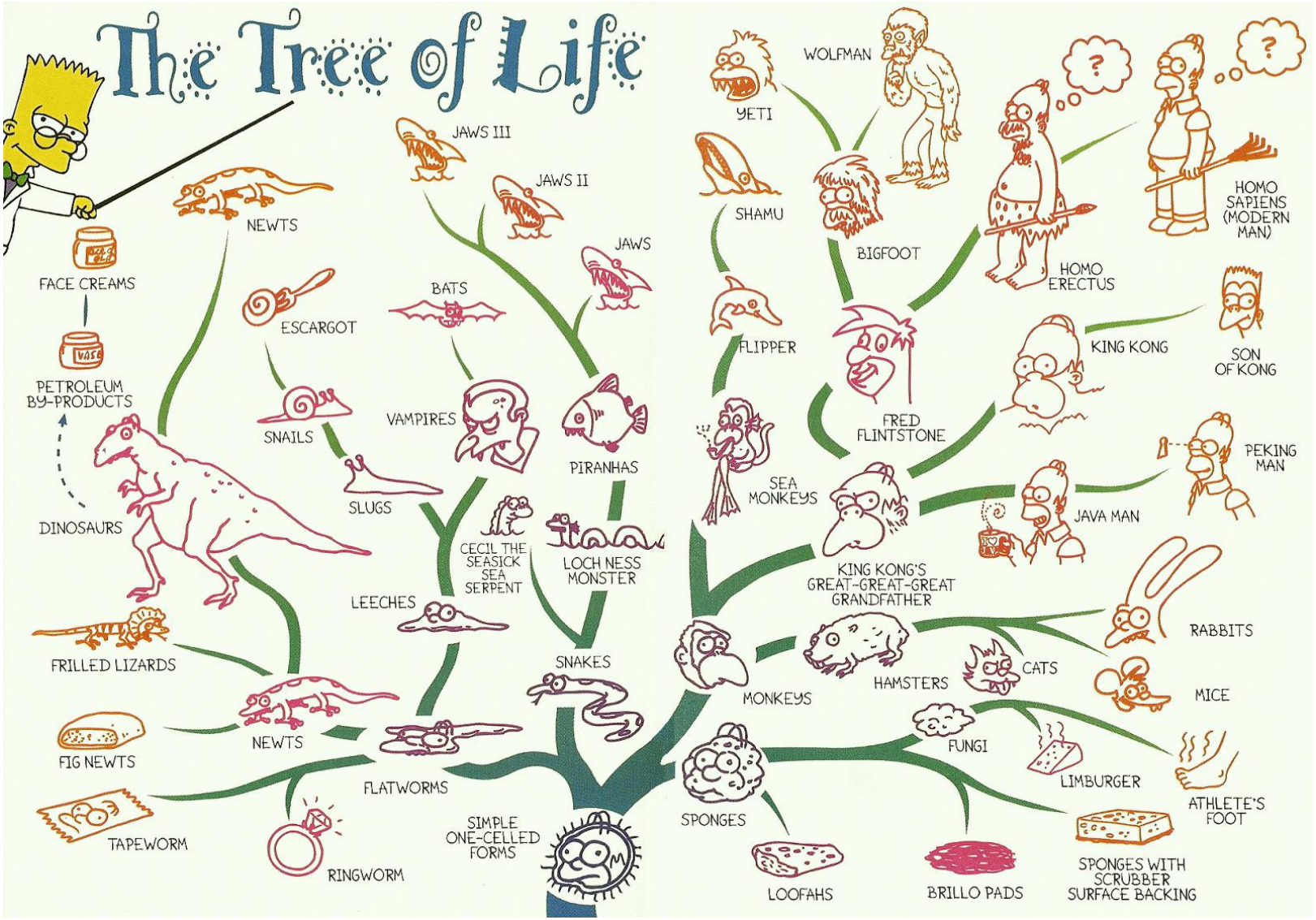
——Deep Learning in the Life Sciences: MIT

Phylogenetic Tree

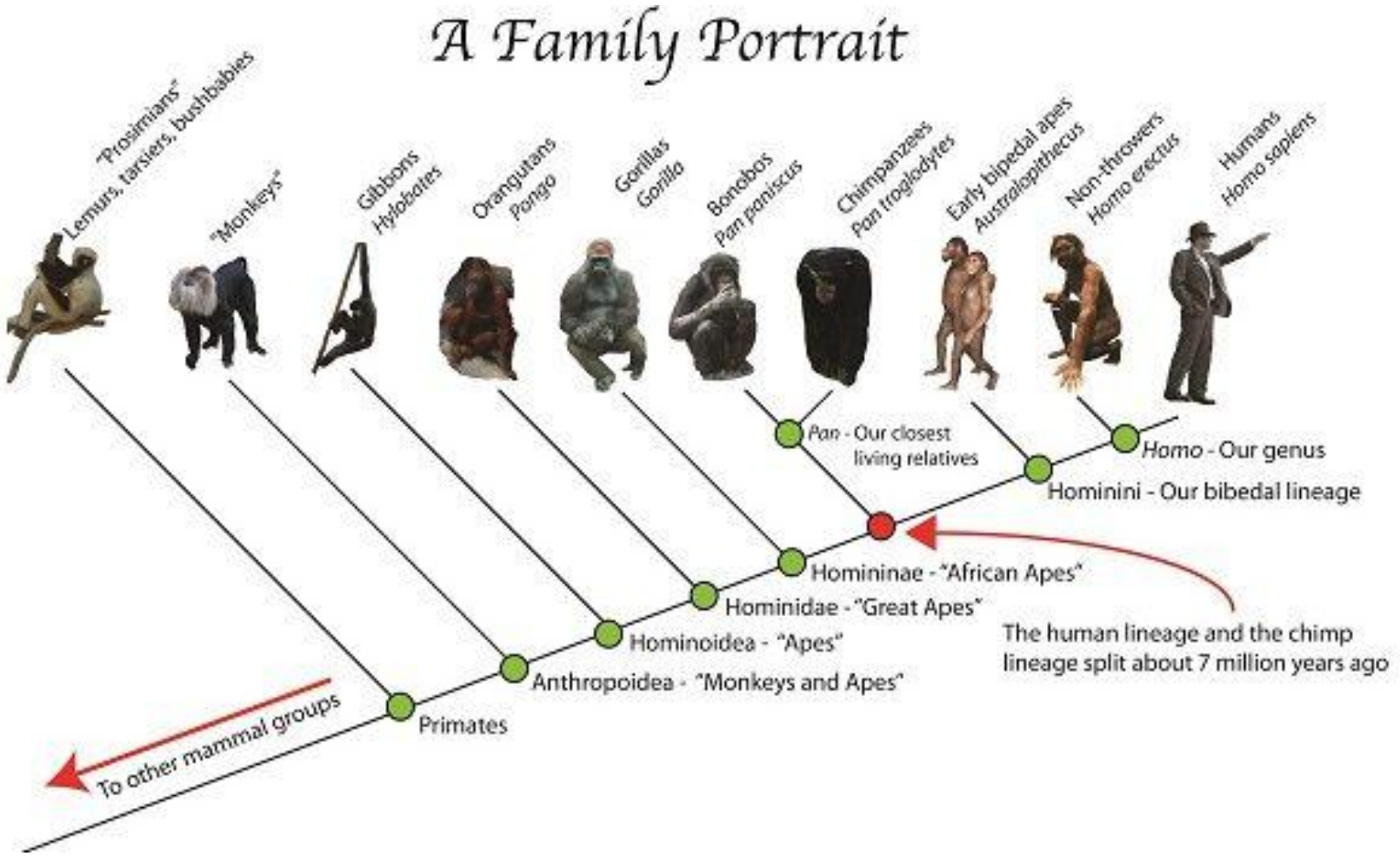
Everyone what to know how this tree look like...



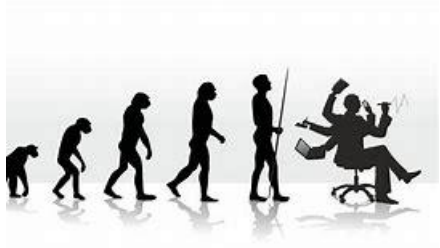
Phylogenetic Tree

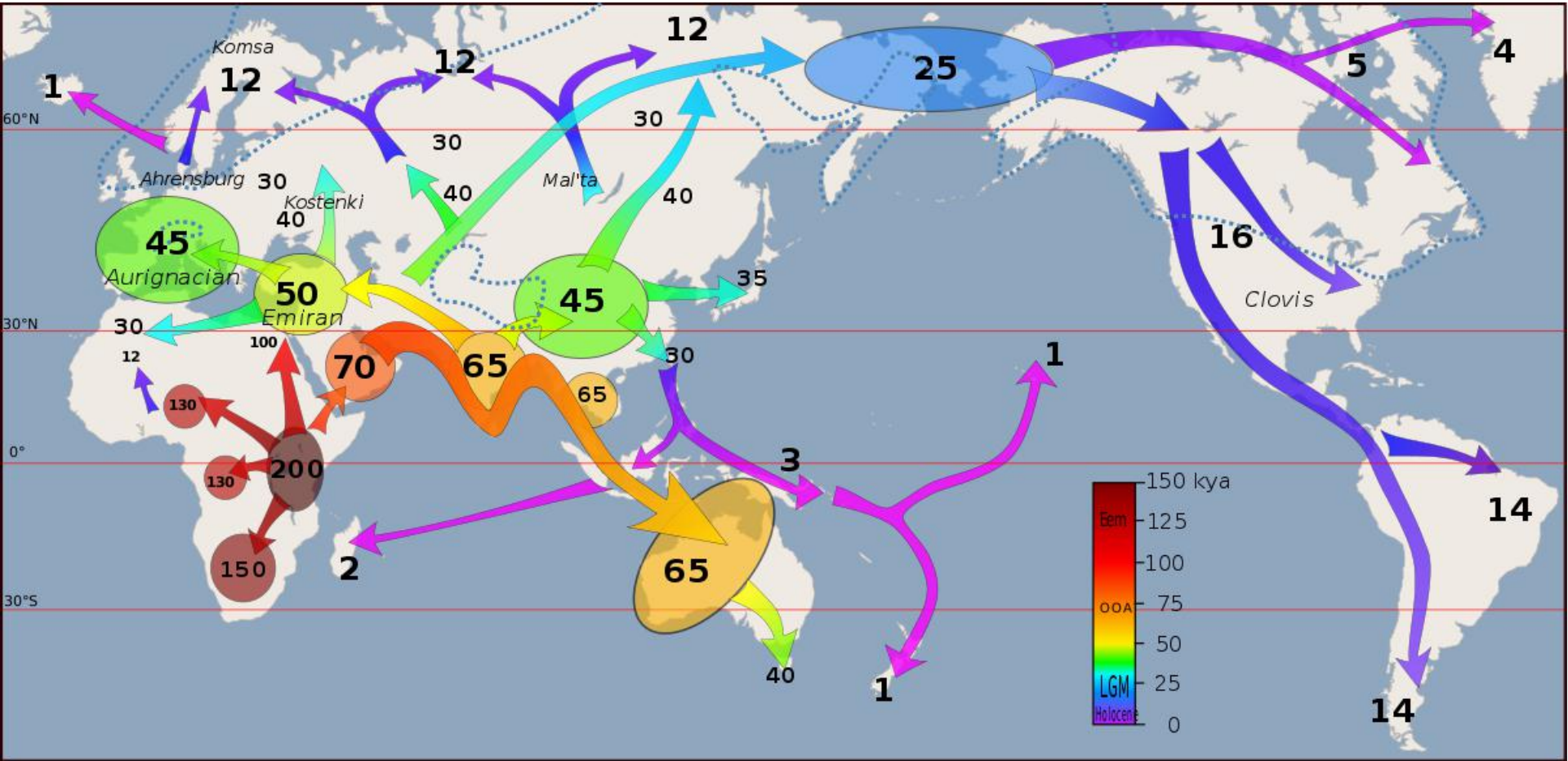


Evolution everywhere



Evolution everywhere

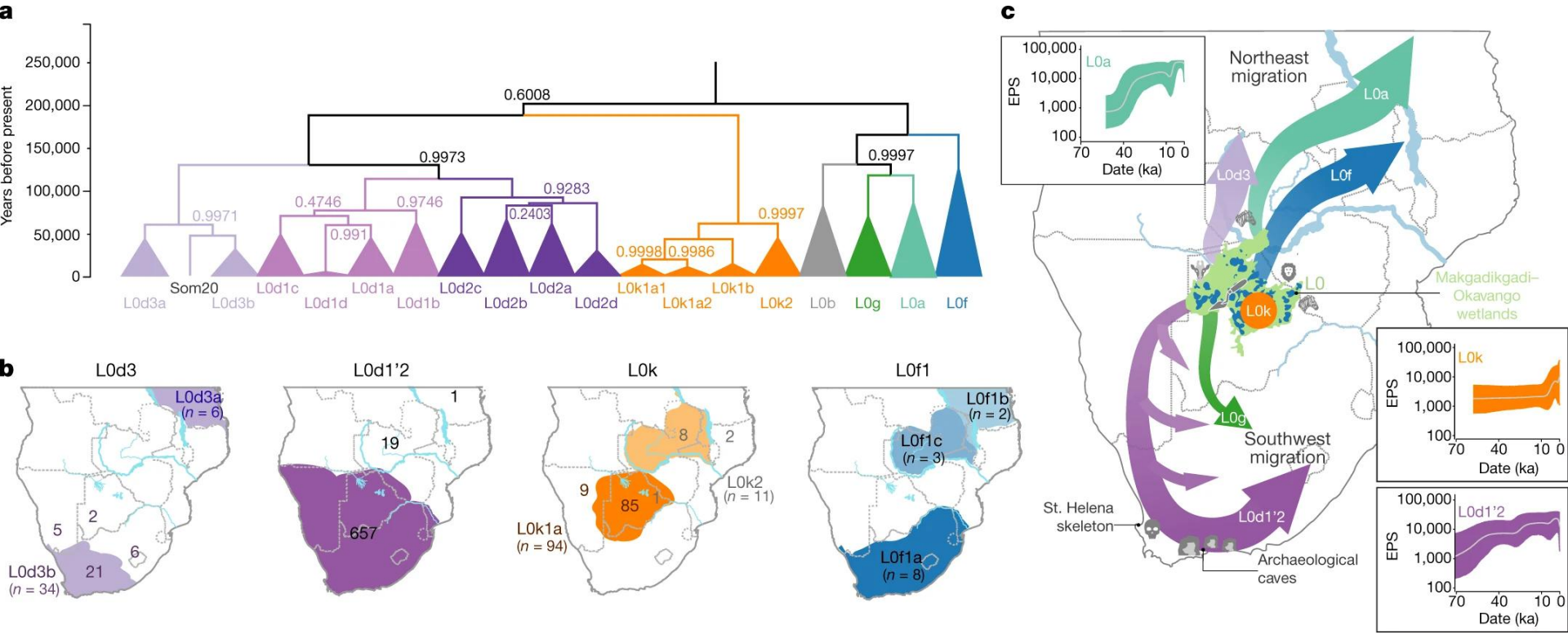




Evolution everywhere

Evolution everywhere

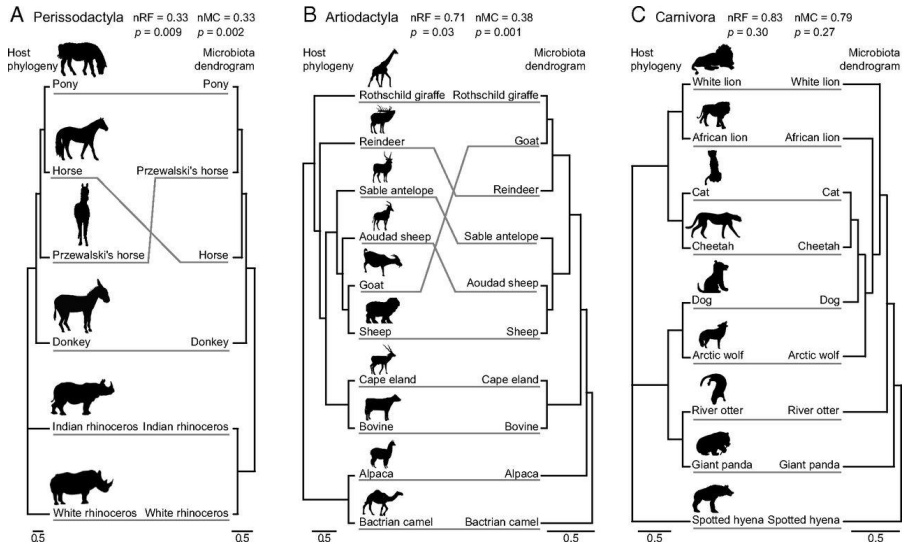
Out of “homeland”?



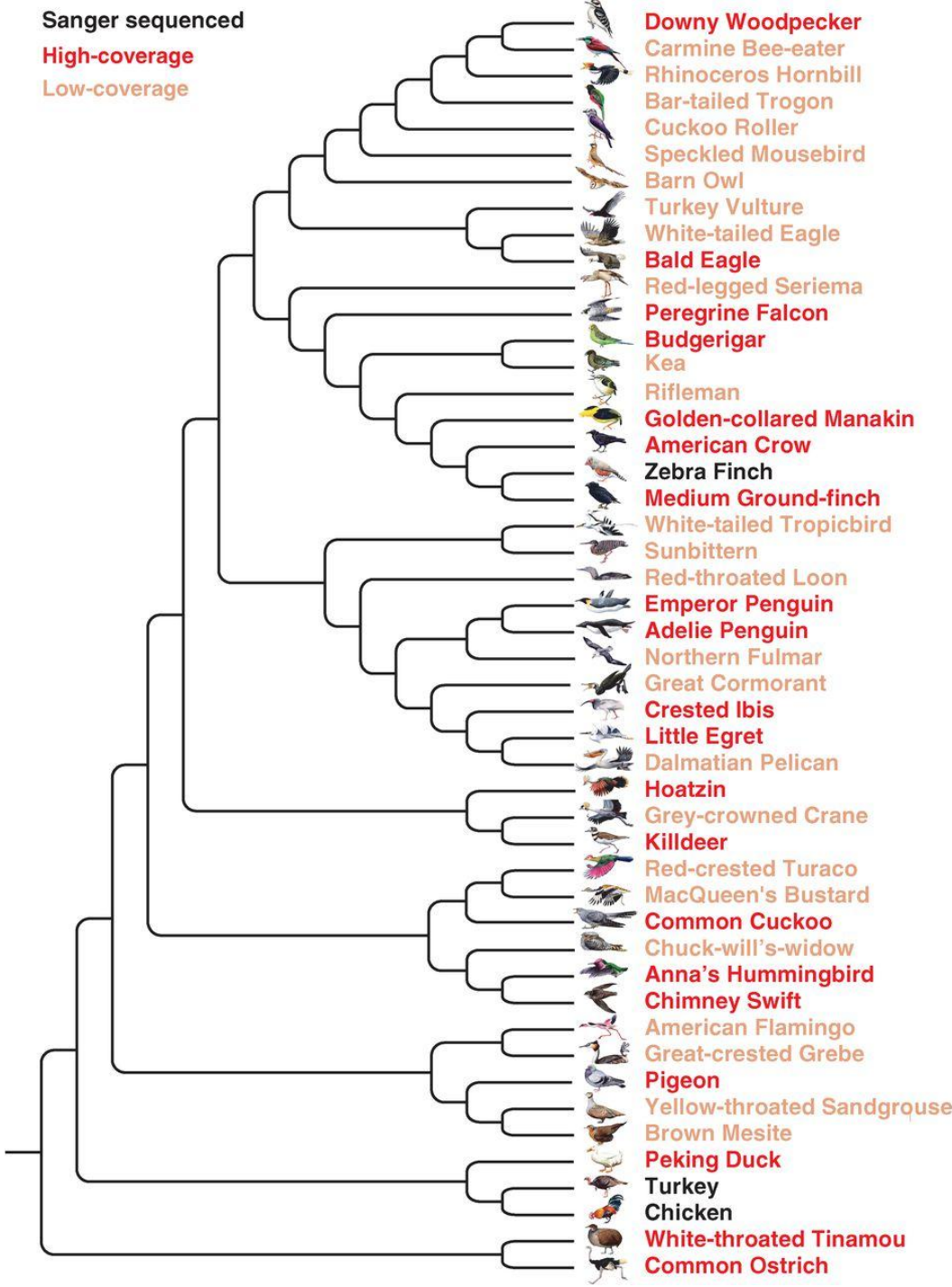
Eva K. F. Chan et al., *Nature*, 2019

Evolution everywhere

Microbiome evolution also exists



Phylogenetic Tree

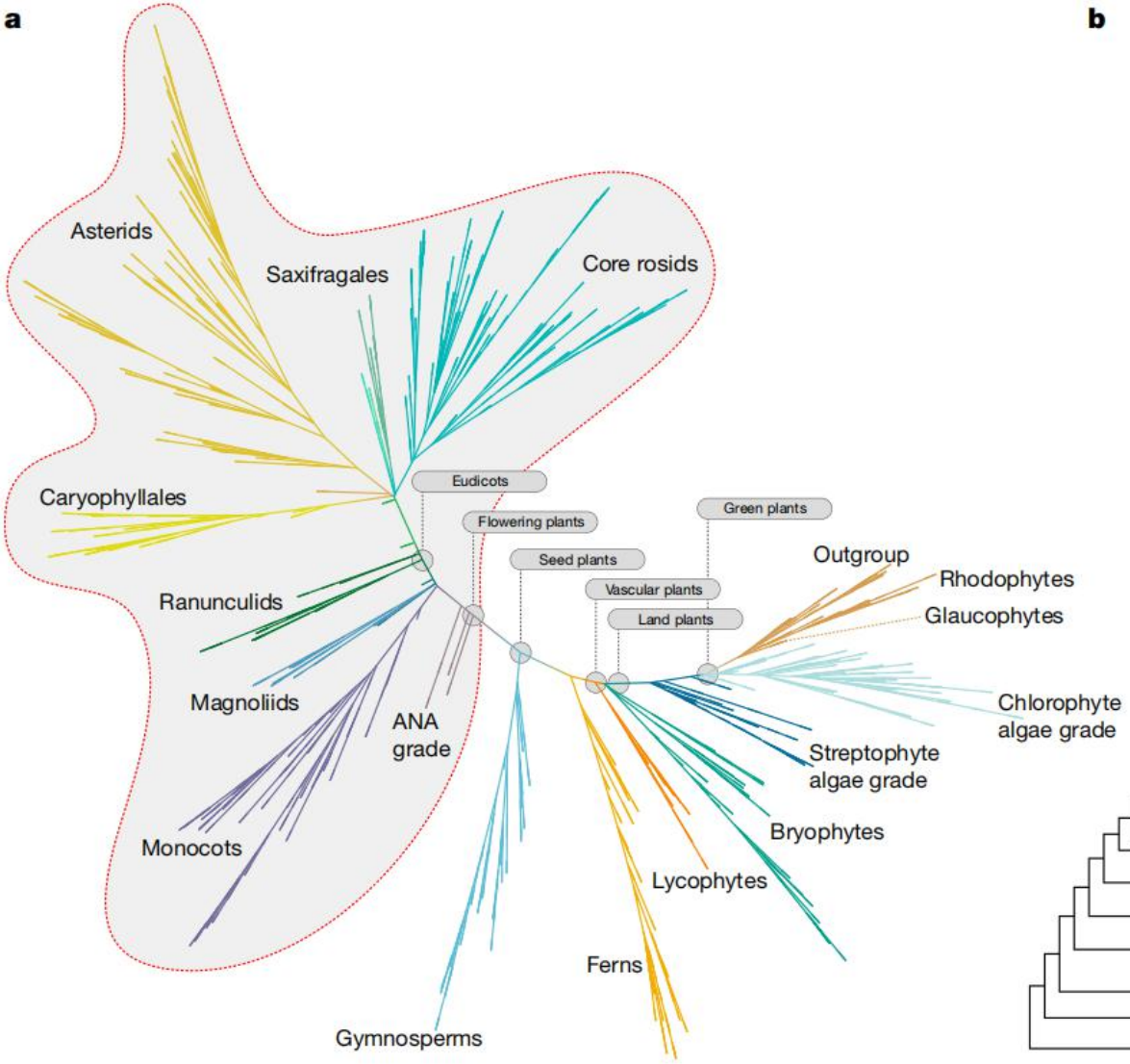


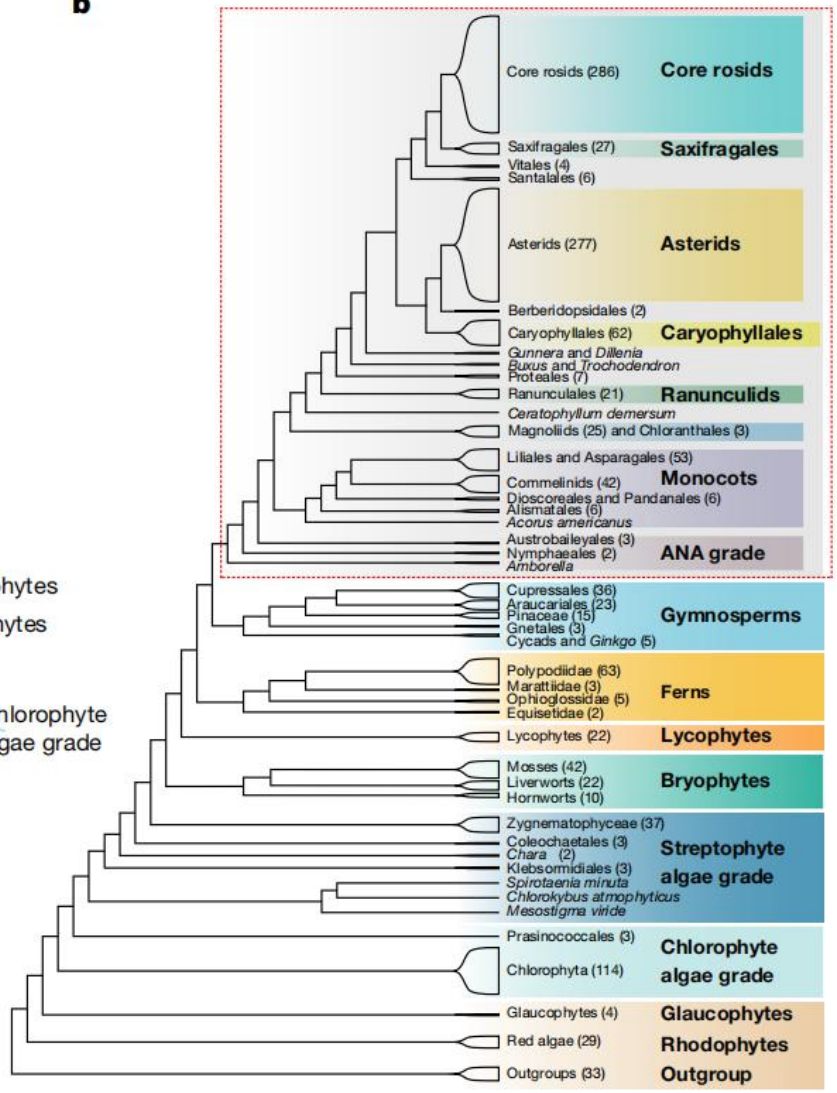
Reference:
https://b10k.genomics.cn/

Phylogenetic Tree

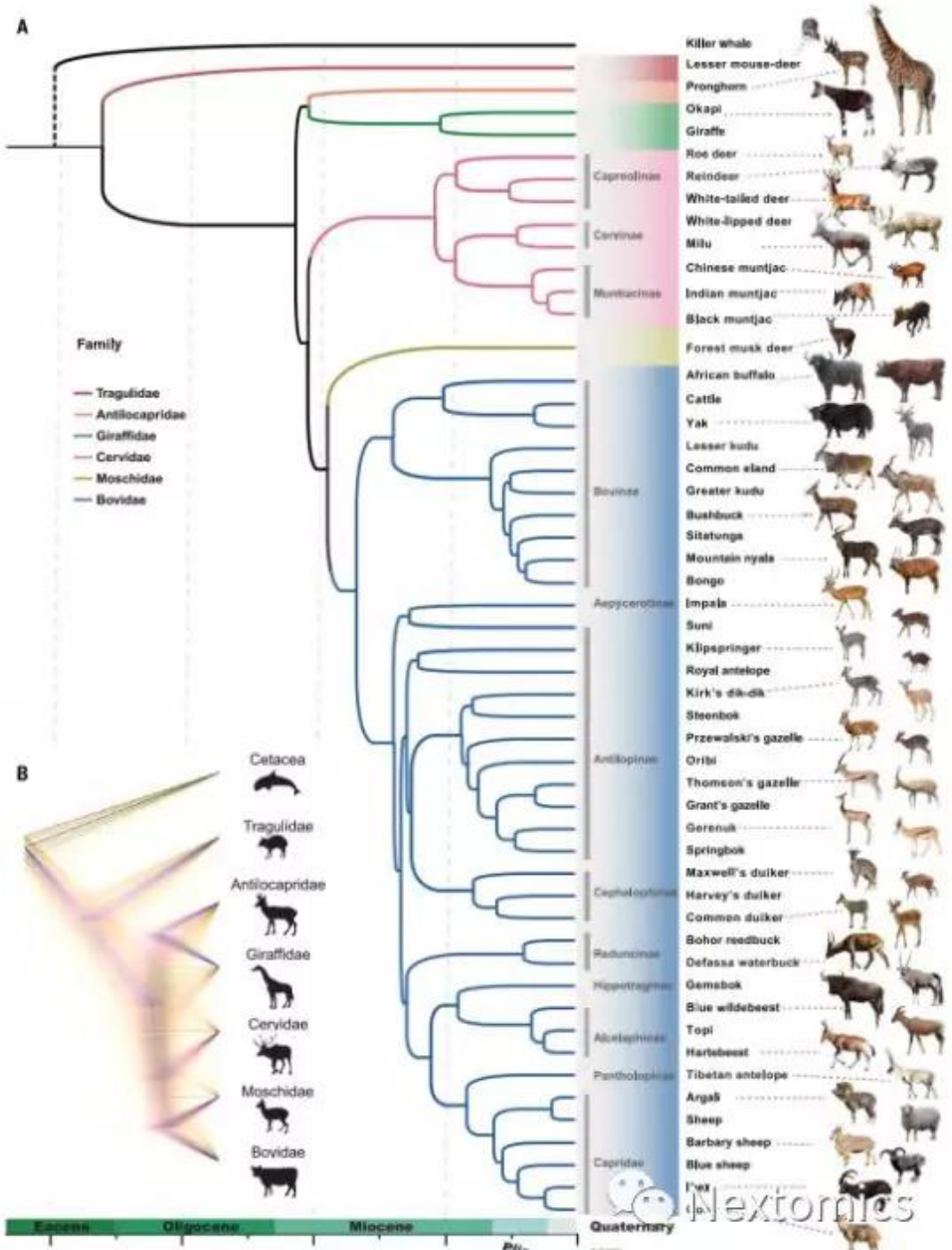
Reference:

One thousand plant transcriptomes and the phylogenomics of green plants, *Nature*, 2019

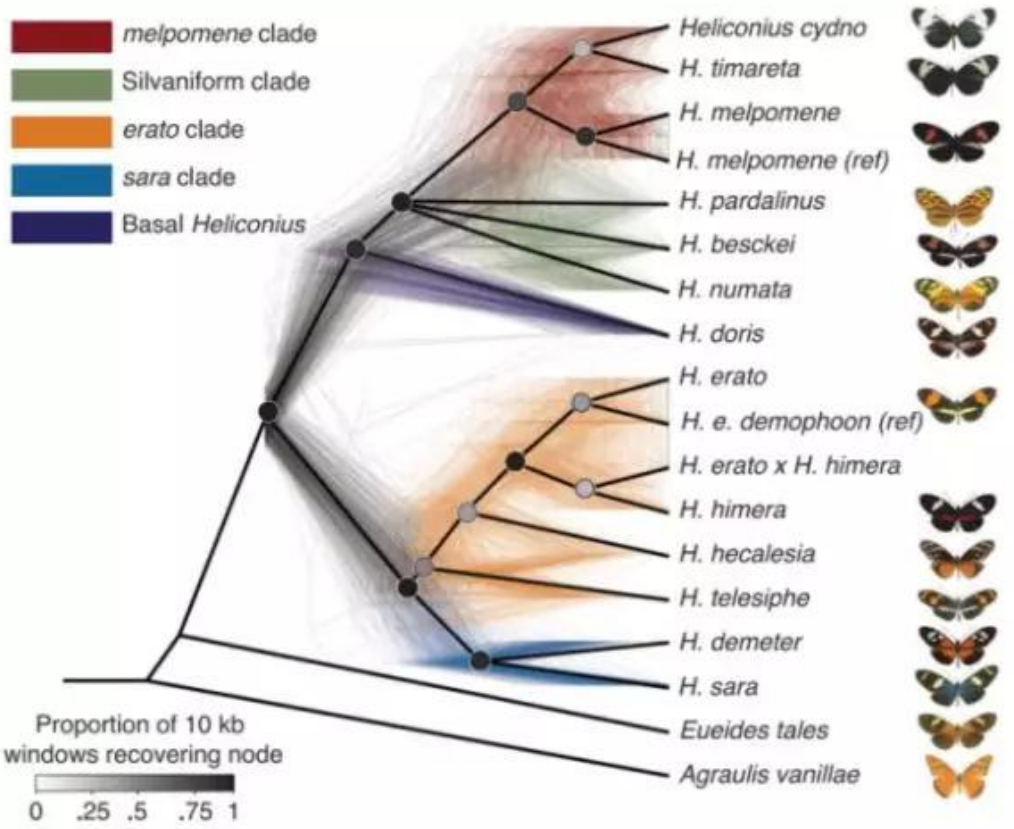




Phylogenetic Tree



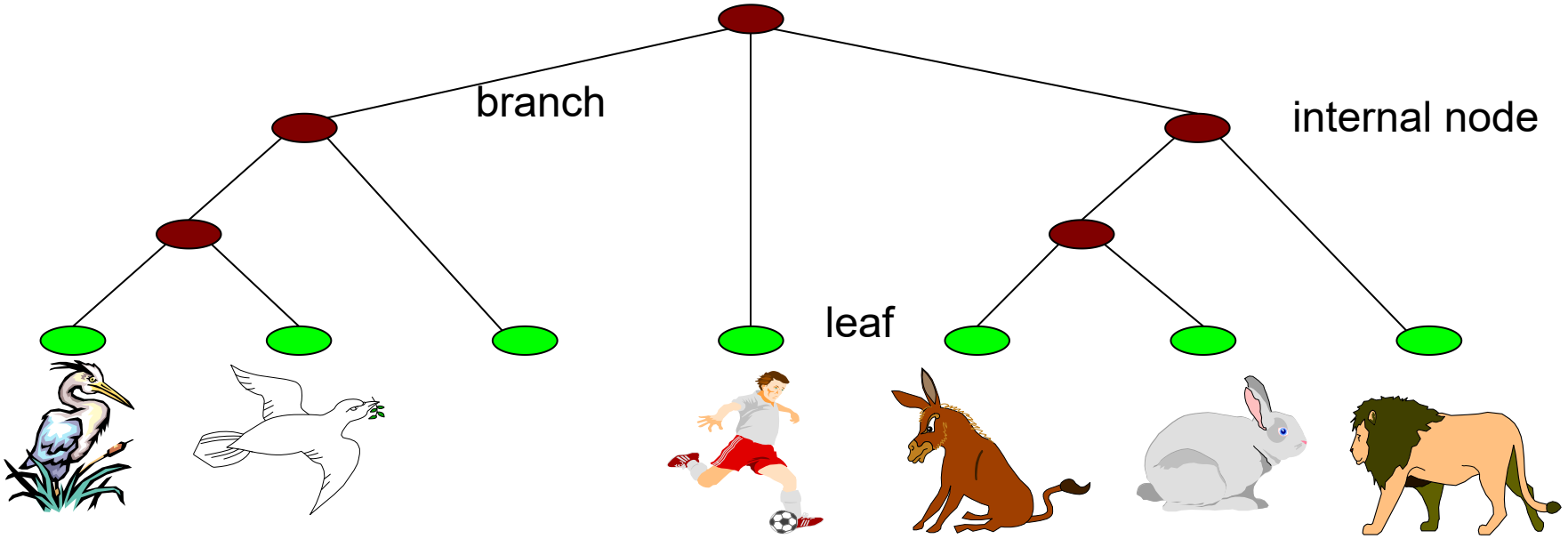
Phylogenetic Tree



Nextomics

Phylogenetic Tree

Phylogenetic Tree



- Topology: bifurcating
  - Leaves - *1…N*
  - Internal nodes *N+1…2N-2*
- Branch length

Reference:
https://itol.embl.de/

构建进化树算法

- 基于距离的构建方法：
  - UPGMA（Unweighted pair group method with arithmetic mean，平均连接聚类法）
  - ME（Minimum Evolution，最小进化法）
  - NJ（Neighbor-Joining，邻接法）

- 基于特征的构建方法：
  - 最大简约法（MP法）
  - 最大似然法（ML法）
  - 进化简约法（EP法）
  - 相容性方法

基于距离的构建方法

- UPGMA

①以已求得的距离系数,所有比较的分类单元的成对距离构成一个t×t方阵,即建立一个距离矩阵M。

②对于一个给定的距离矩阵,寻求最小距离值Dpq。

③定义类群p和q之间的分支深度Lpq=Dpq/2。

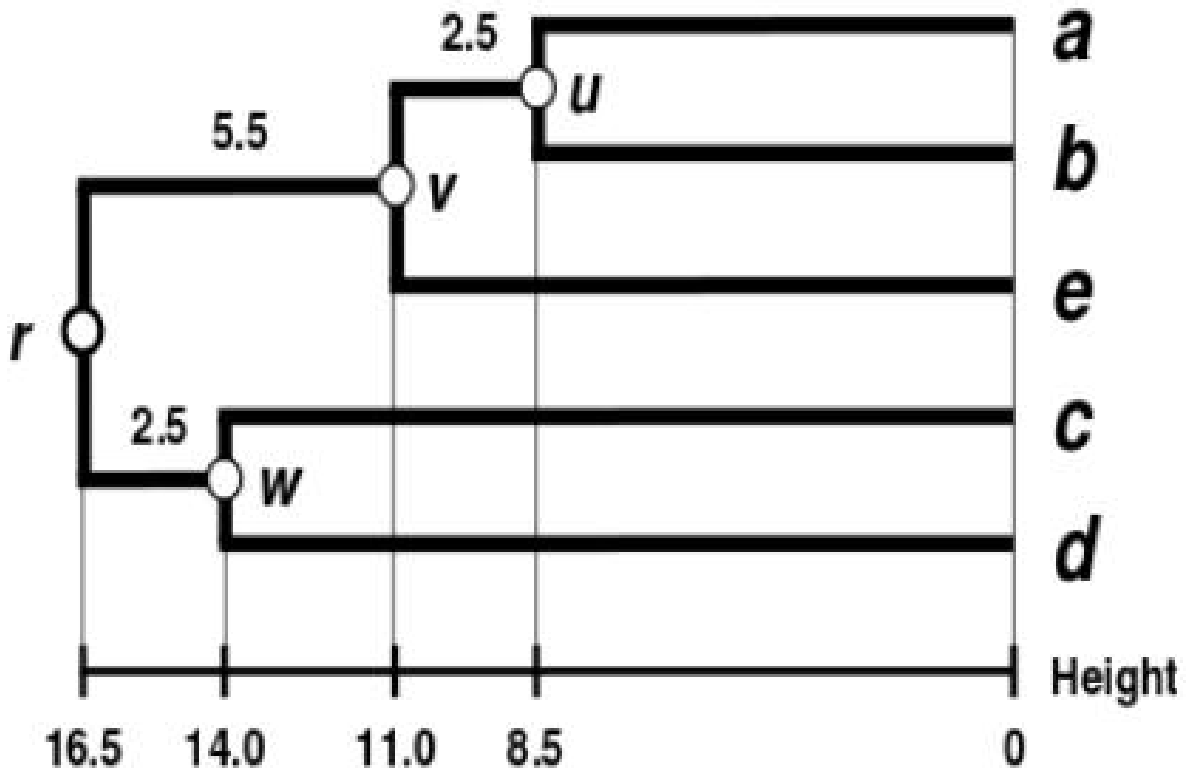
④若p和q是最后一个类群,侧聚类过程完成,否侧合并p和q成一个新类群r。

⑤定义并计算新类群r到其他各类群i(i≠p和q)的距离Dir=(Dpi+Dqi)/2。

⑥回到第一步,在矩阵中消除p和q,加入新类群r,矩阵减少一阶,重复进行直至达到最后归群。

基于距离的构建方法

- UPGMA



基于距离的构建方法

- Neighbor-Joining

① 对于给定距离矩阵中的每一端结i,用下式计算与其它分类单元之间的净趋异量(Ri) (t:矩阵中的分类单元数)

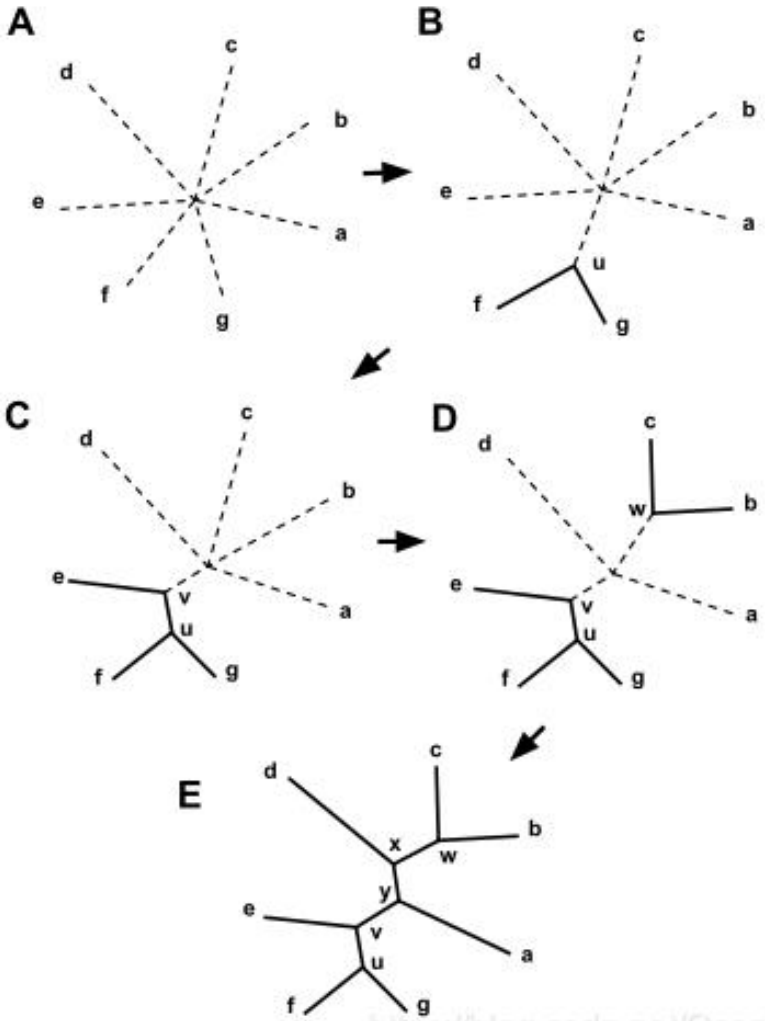
② 建立一个速率校正距离矩阵M,其元素由下式确定:

③ 定义一个新节点u,u的三个分支分别与节点i,j和树的其余部分相连,并且Dij为矩阵中距离最小者,u到节点i和j的分支长度定义为

④ 定义u到树的其它节点k(k≠i和j外的所有节点)的距离:

⑤ 从距离矩阵中删除i和j的距离,矩阵减少一阶。

⑥ 如果矩阵仍然多于两个的节点,重复第①-⑤步,否测除最外两个节点的分支长度来确定外,树上其余节点都确定,最后是剩余的2个的分支长度Sy=Dij



构建进化树算法

- 基于距离的构建方法：
  - UPGMA（Unweighted pair group method with arithmetic mean，平均连接聚类法）
  - ME（Minimum Evolution，最小进化法）
  - NJ（Neighbor-Joining，邻接法）

- 基于特征的构建方法：
  - 最大简约法（MP法）
  - 最大似然法（ML法）
  - 进化简约法（EP法）
  - 相容性方法

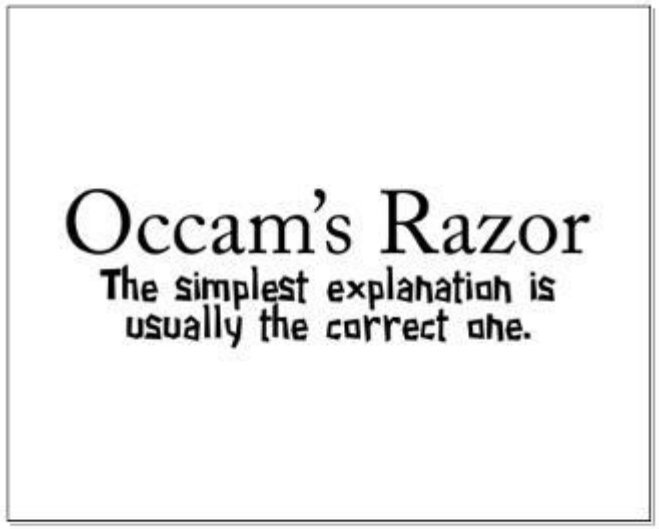
最大简约法（Maximum Parsimony）

最大简约法的理论基础是奥卡姆（Ockham）哲学原则，这个原则认为：解释一个过程的最好理论是所需假设数目最少的那一个。

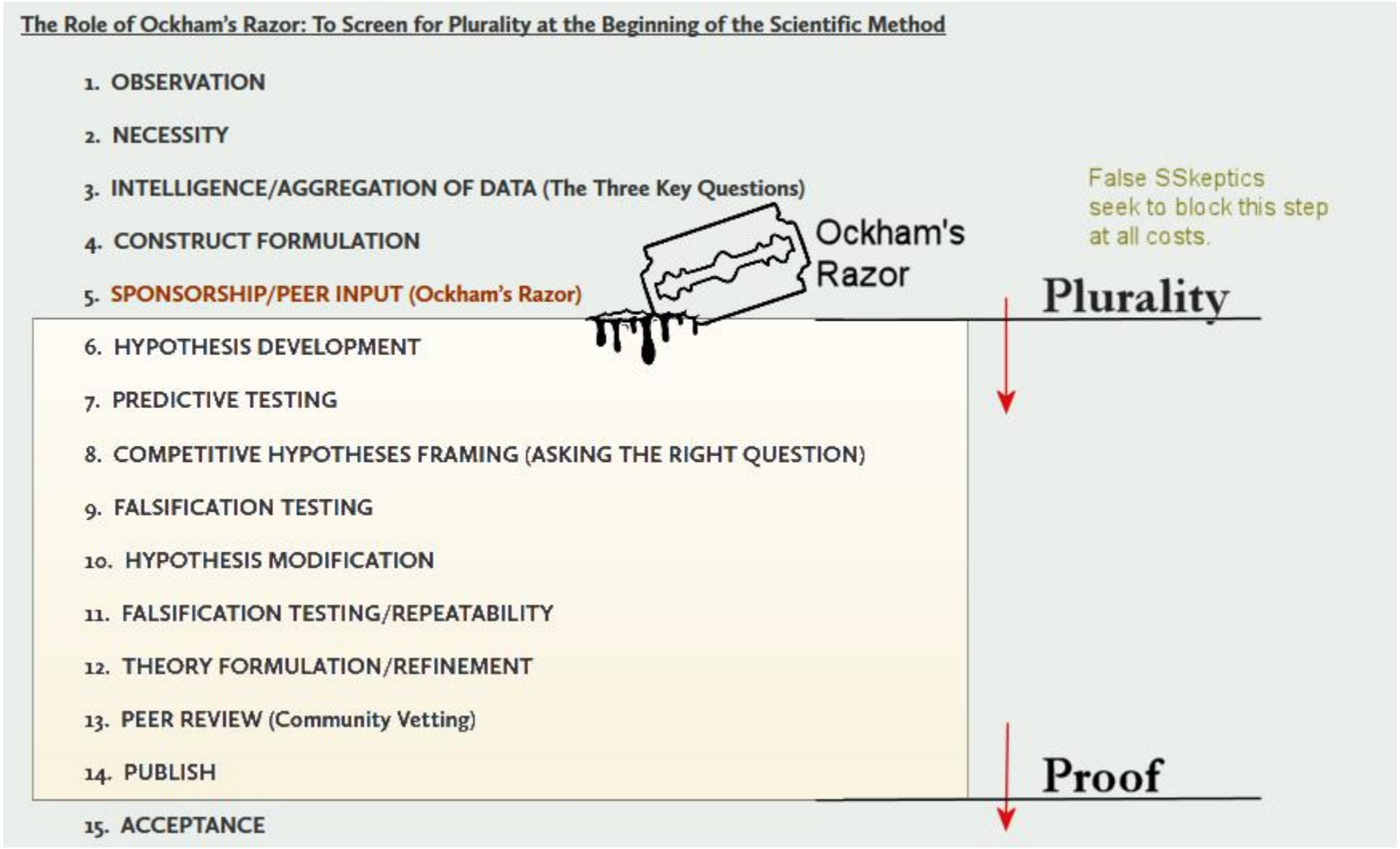
方法：

计算所有可能的拓扑结构；

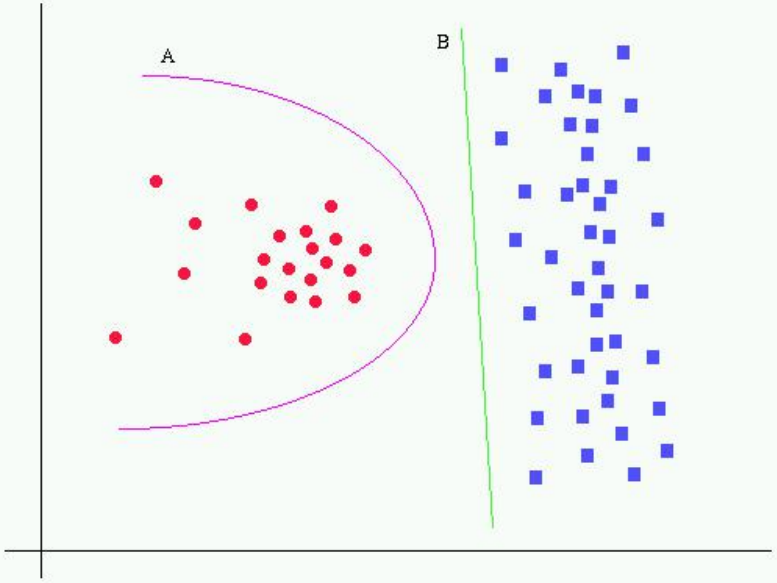
计算出所需替代数最小的那个拓扑结构，作为最优树。

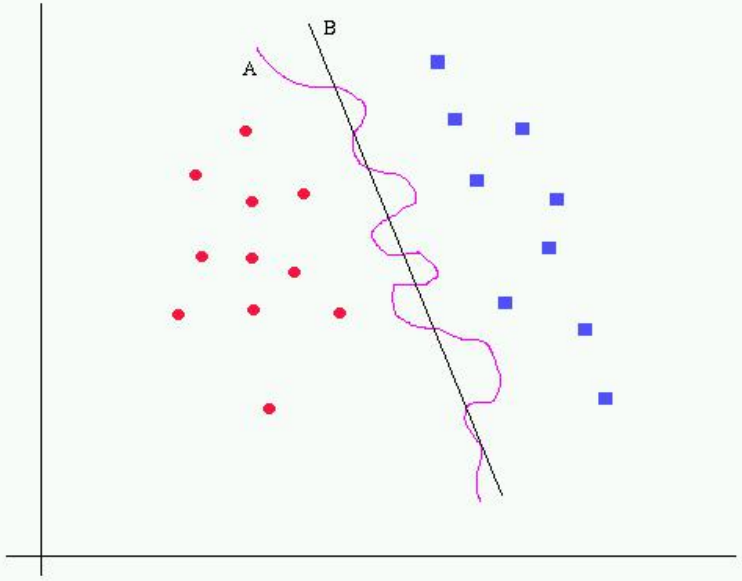


occam's razor method



occam's razor method





最大似然法（Maximum Likelyhood）

ML法对所有可能的系统发育树都计算似然函数，似然函数值最大的那棵树即为最可能的系统发育树。

利用最大似然法来推断一组序列的系统发生树，需首先确定序列进化的模型，如Jukes—Cantor模型、Kimura二参数模型及一般二参数模型等。在进化模型选择合理的情况下，ML法是与进化事实吻合最好的建树算法。其缺点是计算强度非常大，极为耗时。

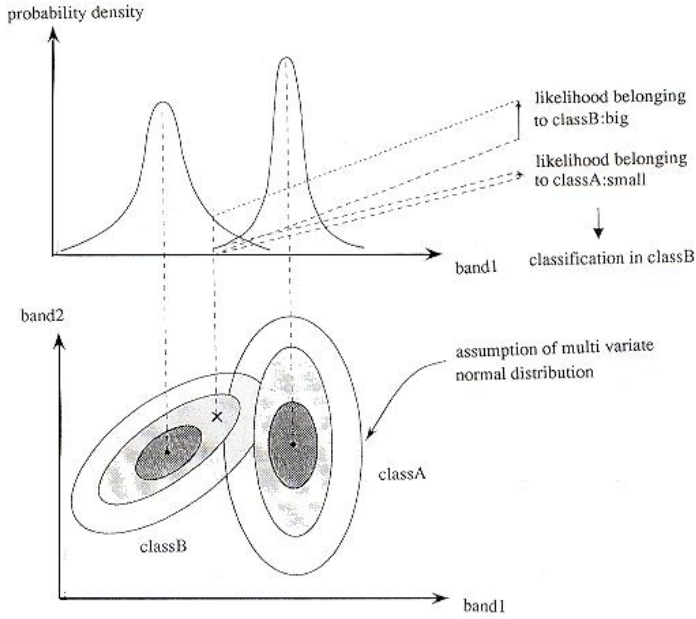
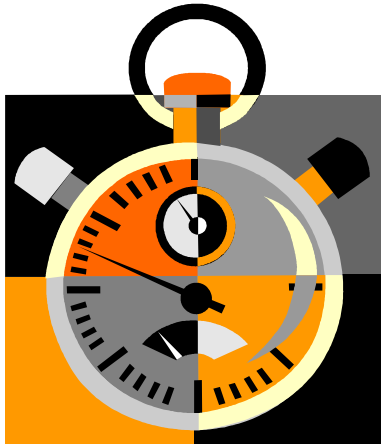


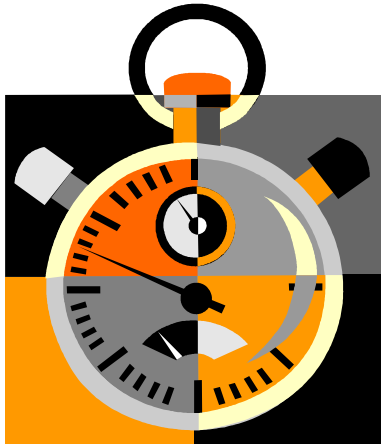
Figure 11.7.1 Concept of Maximum Likelihood Method

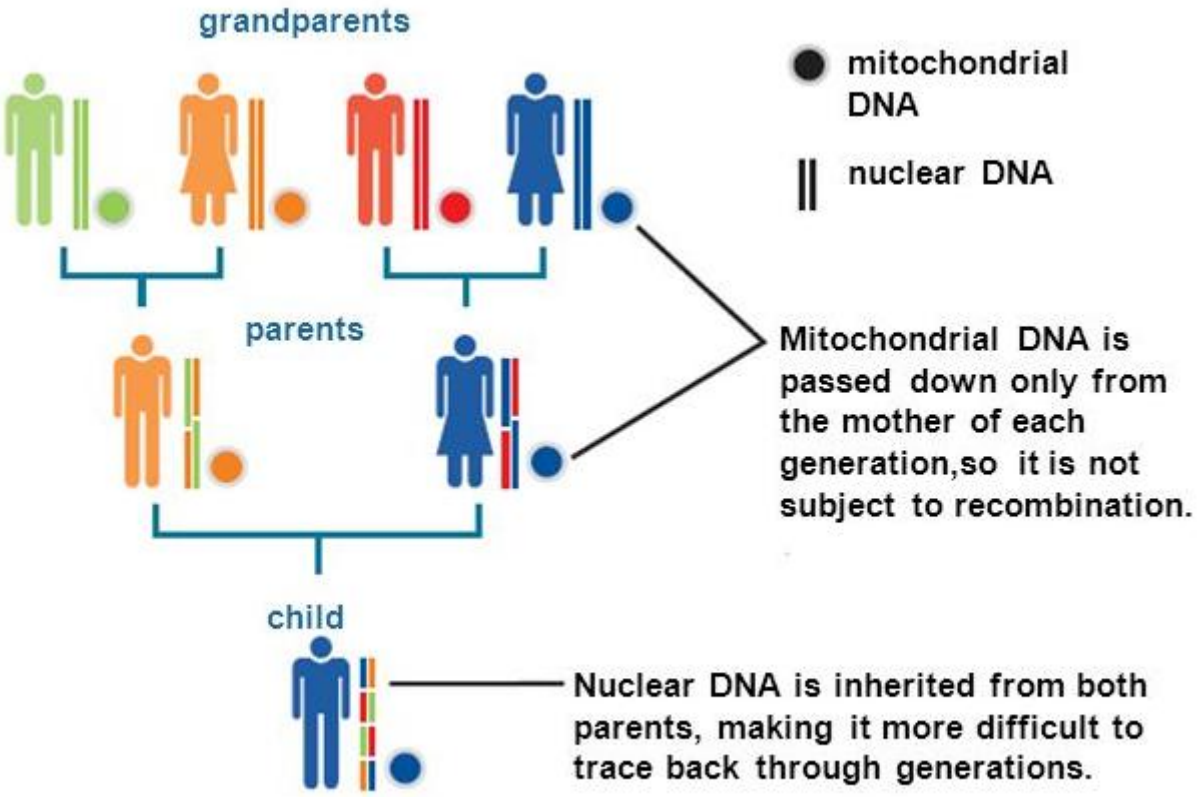
Molecular Clock Hypothesis



- Amount of genetic difference between sequences is a function of time since separation.
- Rate of molecular change is constant (enough) to predict times of divergence

Molecular Clock Hypothesis





Likelihood of a Tree

- Given:
  - n aligned sequences M= $X_1,\ldots,X_n$
  - A tree T, leaves labeled with $X_1,\ldots,X_n$
- Reconstruction t*:
  - Labeling of internal nodes
  - Branch lengths

Goal: Find optimal reconstruction t* : One maximizing the likelihood P(M|T, t*)

Probabilistic Methods

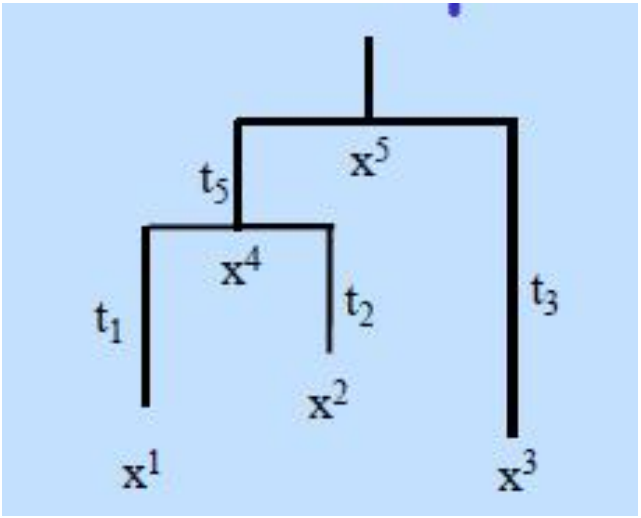
- The phylogenetic tree represents a generative probabilistic model (like HMMs) for the observed sequences.
- Background probabilities: q(a)
- Mutation probabilities: P(a|b, t)
- Models for evolutionary mutations
  - Jukes Cantor
  - Kimura 2-parameter model
- Such models are used to derive the probabilities

Probabilistic Model

- Assumptions:
  - Each character is independent
  - The branching is a Markov process: The probability that a node x has a specific label is only a function of the parent node y and the branch length t between them
  - The probabilities $P(x|y,t)$ are known

Example

- Given the tree



$$P(x_1, x_2, x_3, x_4, x_5|T, t^*)$$
$$=P(x_1|x_4, t_1)P(x_2|x_4, t_2)P(x_3|x_5, t_3)P(x_4|x_5, t_5)$$

Molecular Evolution

Q: How can we model evolution on nucleotide level? (ignore gaps, focus on substitutions)

A: Consider what happens at a specific position for small time interval Δt

- $P(t)$ = vector of probabilities of {A,C,G,T} at time t
- μ_{AC} = rate of transition from A to C per unit time
- $\mu_A = \mu_{AC} + \mu_{AG} + \mu_{AT}$ rate of transition out of A
- $p_A(t+\Delta t) = p_A(t) - p_A(t)\,\mu_A\,\Delta t + p_C(t)\,\mu_{CA}\,\Delta t + \ldots$

Molecular Evolution

In matrix/vector notation, we get

$$P(t + \Delta t) = P(t) + QP(t)\Delta t$$

where Q is the substitution rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$

Molecular Evolution

- This is a differential equation:

$$P'(t) = Q\,P(t)$$

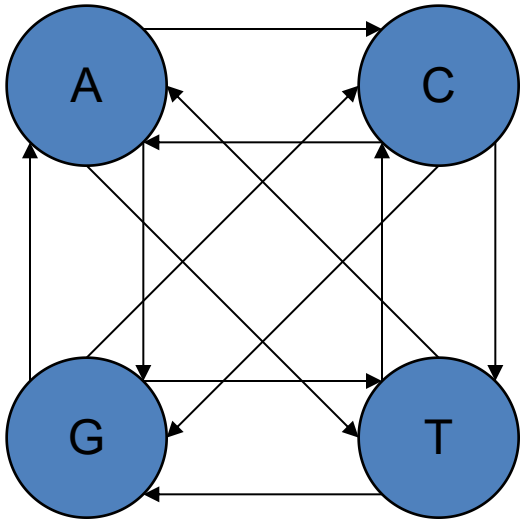
- A substitution rate matrix Q implies a probability distribution over {A,C,G,T} at each position, including stationary (equilibrium) frequencies π_A, π_C, π_G, π_T
- Each Q is an evolutionary model (some work better than others)

Mutation Probabilities

P(t) satisfy the following two property:

- **Lack of memory**:
  - $P_{a\to c}(t+t') = \sum_b P_{a\to b}(t) P_{b\to c}(t')$
- **Reversibility:**
  - Exist stationary probabilities $\{P_a\}$ s.t.

$$P_a P_{a\to b}(t) = P_b P_{b\to a}(t)$$



PAM矩阵

- Point accepted mutation (Dayhoff et al 1978)
- Given an tree of protein family, the frequence matrix A_{ab} counting the occurrence of an “a” in the ancestral sequence was replaced by a “b” in the descendant.
- Estimate the conditional probability p(b|a)

$$P(b|a) = B_{a,b} = \frac{A_{ab}}{\sum_c A_{ac}}$$

PAM矩阵

- Scaling B

$$C_{ab} = \sigma B_{ab}, C_{aa} = \sigma B_{aa} + (1 - \sigma)$$

- Such that the expected number of substitution is 1%, i.e.

$$\sum_{ab} q_a q_b C_{ab} = 0.01$$

- Then the PAM(1) matrix is given by

$$S(1) = (C_{ab})$$

Calculating the Likelihood for Ungapped Alignments

$$P\left(x^1, x^2 \mid T, t_1, t_2\right) = \prod_{u=1}^{N} P\left(x_u^1, x_u^2 \mid T, t_1, t_2\right)$$

$$P\left(x_u^1, x_u^2 \mid T, t_1, t_2\right) = \sum_a q_a P\left(x_u^1 \mid a, t_1\right) P\left(x_u^2 \mid a, t_2\right)$$

Assuming Jukes-Cantor model & $q_C = q_G = q_A = q_T = \frac{1}{4}$:

$$P\left(C, C \mid T, t_1, t_2\right) = q_C r_{t_1} r_{t_2} + q_G s_{t_1} s_{t_2} + q_A s_{t_1} s_{t_2} + q_T s_{t_1} s_{t_2} = \frac{1}{4}\left(r_{t_1} r_{t_2} + 3 s_{t_1} s_{t_2}\right)$$

$$P(C, G \mid T, t) = P(G, C \mid T, t) = \frac{1}{4}\left(r_{t_1} s_{t_2} + s_{t_1} r_{t_2} + 2 s_{t_1} s_{t_2}\right)$$

$$\Rightarrow P\left(x^1, x^2 \mid T, t_1, t_2\right) = 16^{-(n1+n2)} \left(1 + 3e^{-4\alpha(t_1+t_2)}\right)^{n1} \left(1 - e^{-4\alpha(t_1+t_2)}\right)^{n2}$$

where $n1$=matches,n2=mismatches

Calculating the Likelihood for Ungapped Alignments

- n sequences of length N, site u=1...N
- Given a rooted tree contains *2n - 1* nodes*, 1... n* being the leaf nodes, n+1 ... 2n-1 non-leaf, tree lengths t1, …, t_{2n-1}.
- Let a(i) denote the ancestor of node a^i

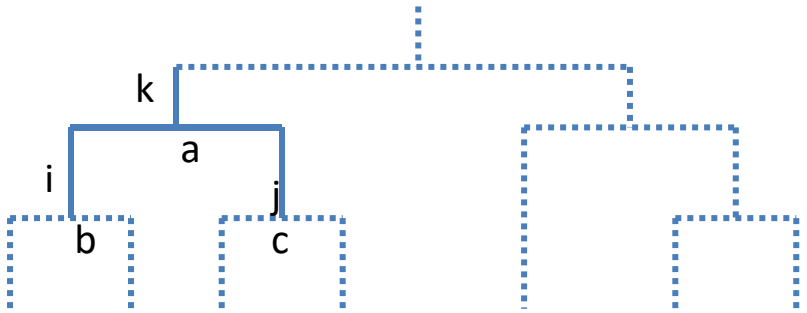
$$P(x^1, \cdots, x^n | T, t) = \prod_{u=1}^{N} P(x_u^1, \cdots, x_u^n | T, t)$$

$$P(x_u^1, \cdots, x_u^n | T, t) = \sum_{a^{n+1}, \cdots, a^{2n-1}} q_{a^{2n-1}} \prod_{i=n+1}^{2n-2} P(a^i | a^{\alpha(i)}, t_i)$$

$$\times \prod_{i=1}^{n} P(x_u^i | a^{\alpha(i)}, t_i)$$

Felsenstein's Recursive Algorithm

- Let $P(L_k|a)$ denote the probability of all the leafs below node k given that the residue at k is a.
- Then we compute $P(L_k|a)$ from the probabilities $P(L_i | b)$ and $P(L_j |c)$ for all b and c, where i and j are the daughter nodes of k.



Felsenstein's Recursive Algorithm

- Initialization: set k=2n-1
- Recursion: Compute $P(L_k \mid a)$ for all a as follows:
  - If k is leaf node: $P(L_k|a)=1$ only if $a = x_u^k$.
  - If k is not a leaf node:
    - Compute $P(L_i|a)$, $P(L_j|a)$ for all a at the daughter nodes i,j , and set $P(L_k|a) = \sum_{bc} P(b|a,t_i)P(L_i|b)P(c|a,t_j)P(L_j|c)$
- Temination: Likelihood at site u,

$$P(x_u|T,t) = \sum_a P(L_{2n-1}|a)q_a$$

Reversibility & Independence of Root Position

- The score of the optimal tree is independent of the root position if and only if:
  - the substitution matrix is **multiplicative**
  - the substitution matrix is **reversible**
- A substitititution matrix is reversible if for all a,b and t:

$$P(b|a,t)q_a = P(a|b,t)q_b$$

Maximum Likelihood (ML)

- Score each tree by
  - Assumption of independent positions “m”
- Branch lengths *t* can be optimized
  - Gradient Ascent
  - EM
- We look for the highest scoring tree
  - Exhaustive
  - Sampling methods (Metropolis)

Computational Problem

- Such procedures are computationally expensive!
- Computation of optimal parameters, per candidate, requires non-trivial optimization step.
- Spend non-negligible computation on a candidate, even if it is a low scoring one.
- In practice, such learning procedures can only consider small sets of candidate structures

构建进化树算法

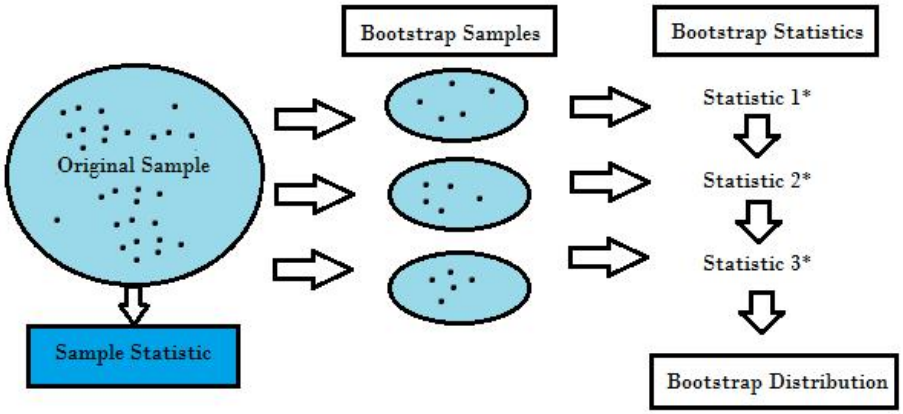
- 如果序列的相似性较高，各种方法都会得到不错的结果，模型间的差别也不大.
- 若有合适的分子进化模型可供选择，用最大似然法构树获得的结果较好.
- 对于近缘物种序列，通常情况下使用最大简约法.
- 而对于远缘物种序列，一般使用邻接法或最大似然法.
- 对于相似度很低的序列，邻接法往往出现长枝吸引(branch attraction)现象，有时严重干扰进化树的构建.

邻接法和最大似然法是需要选择模型的:

蛋白质序列的构树模型一般选择Poisson correction(泊松修正).

核酸序列的构树模型一般选择Kimura 2-parameter (Kimura-2参数).

构建进化树算法



在重建进化树过程中，均需选择bootstrap进行树的检验：

- 一般bootstrap的值>70，则认为重建的进化树较为可靠。
- 如果bootstrap的值太低，则有可能进化树的拓扑结构有错误，进化树是不可靠的。
- 一般推荐用两种以上不同的方法构建进化树，如果所得到的进化树类似，且bootstrap值总体较高，则得到的结果较为可靠。
- 通常情况下，只要选择了合适的方法和模型，构出的树均是有意义的，研究者可根据自己研究的需要选择最佳的树进行分析。

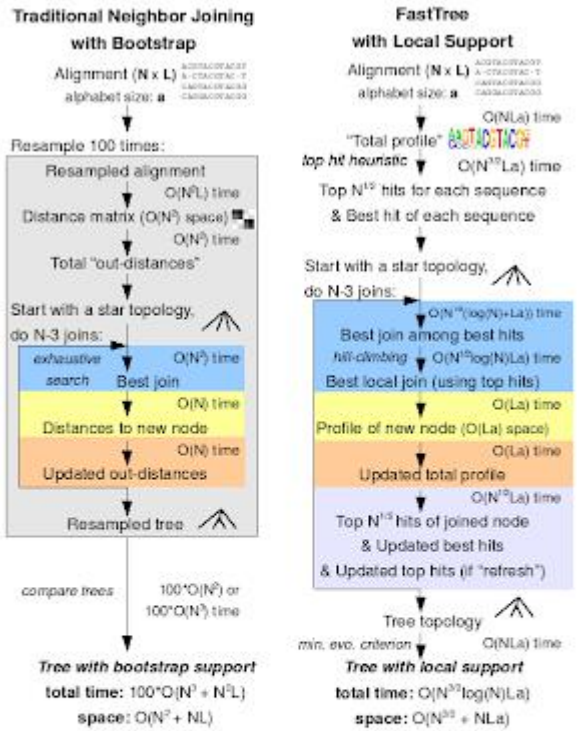
构建分子进化树相关的软件

软件	网址	说明
ClustalX	http://bips.u-strasbg.fr/fr/Documentation/ClustalX/	图形化的多序列比对工具
ClustalW	**http://www.cf.ac.uk/biosi/resear ... loads/clustalw.html**	**命令行格式的多序列比对工具**
GeneDoc	http://www.psc.edu/biomed/genedoc/	多序列比对结果的美化工具
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	序列分析的综合工具
MEGA	**http://www.megasoftware.net/**	**图形化、集成进化分析工具，不包括ML**
PAUP	http://paup.csit.fsu.edu/	商业软件，集成的进化分析工具
PHYLIP	http://evolution.genetics.washington.edu/phylip.html	免费的、集成的进化分析工具
PHYML	**http://atgc.lirmm.fr/phyml/**	**最快的ML建树工具**
PAML	**http://abacus.gene.ucl.ac.uk/software/paml.html**	**ML建树工具**
Tree-puzzle	http://www.tree-puzzle.de/	较快的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
TreeView	**http://taxonomy.zoology.gla.ac.uk/rod/treeview.html**	**进化树显示工具**

MEGA：建树平台

	Likelihood	Distance	Parsimony	Bayesian	Visual Explorer	Caption Expert
Phylogeny	✓	✓	✓		✓	✓
Bootstrap	✓	✓	✓		✓	✓
Distance/Diversity	✓	✓			✓	✓
Model Selection	✓					
Substitution Pattern	✓				✓ XL	✓
Rate Variation	✓				✓ XL	✓
Ancestral Sequence			✓	✓	✓	✓
Clock Test	✓	✓			✓ XL	✓
Time Tree	✓	✓			✓	✓
Selection Test	✓	✓			✓	✓
Disease Mutation	✓	✓			✓ XL	✓

FastTree：建树平台



Neighbor Joining Criterion: find the join that minimizes

$$d^*(A,B) = d(A,B) - r(A) - r(B)$$

Distances to joined nodes:

Neighbor joining: $d(AB,C) = (d(A,C)+d(B,C))/2 - d(A,B)/2$

FastTree: $= P((A+B)/2,C) - u(C) - u(AB)$

Average out-distances:

Neighbor joining: $r(A) = \sum d(A,X)/(n-2)$

FastTree: $= \frac{n*P(A, \sum X/n) - P(A,A) - \sum (u(A)+u(X))}{n-2}$

Speed

Computing maximum-likelihood trees

			FastTree 2.0.0			RAxML	PhyML 3
Alignment	# Distinct Sequences	#Positions	Settings	Hours	Memory (GB)	Hours	Hours
Efflux permeases (COG2814)	8,362	394 a.a.	JTT+CAT	0.25	0.35	>1,200	>1,200
ABC transporters (PF00005)	39,092	214 a.a.	JTT+CAT	1.0	0.96	--	--
16S ribosomal RNAs, distinct families	15,011	1,287 nt.	GTR+CAT	0.66	0.56	99	>360
16S ribosomal RNAs, distinct families	15,011	1,287 nt.	JC+CAT	0.49	0.36	--	--
16S ribosomal RNAs	237,882	1,287 nt.	JC+CAT, -fastest	21.8	5.8	--	--

All of the timings are on a single CPU. The FastTree times include the SH-like local support values. For huge alignments, FastTree 2.1 with -fastest is about twice as fast as 2.0, and the multi-alignment I ran RAxML 6 with the fast hill-climbing option (not RAxML 7), and I ran PhyML was run with the fastest settings (no variation in rates across sites and no SPR moves).

In theory, FastTree takes $O(N\ L\ a + N^{1.5})$ space and $O(N^{1.5}\ \log(N)\ L\ a)$ time, where N is the number of unique sequences, L is the width of the alignment, and a is the size of the alphabet. W time complexity are dominated by initializing the top-hits lists and maintaining them during the neighbor-joining phase. The minimum-evolution NNIs and SPRs take $O(N\ L\ a)$ time per roun $O(N\ \log(N)\ L\ a^2)$ time total, and $O(N\ L\ a)$ space. Similarly, the local supports take $O(N\ L\ a^2)$ time and $O(N\ L\ a\ )$ space. In practice, the maximum likelihood NNIs are usually the slowest step.

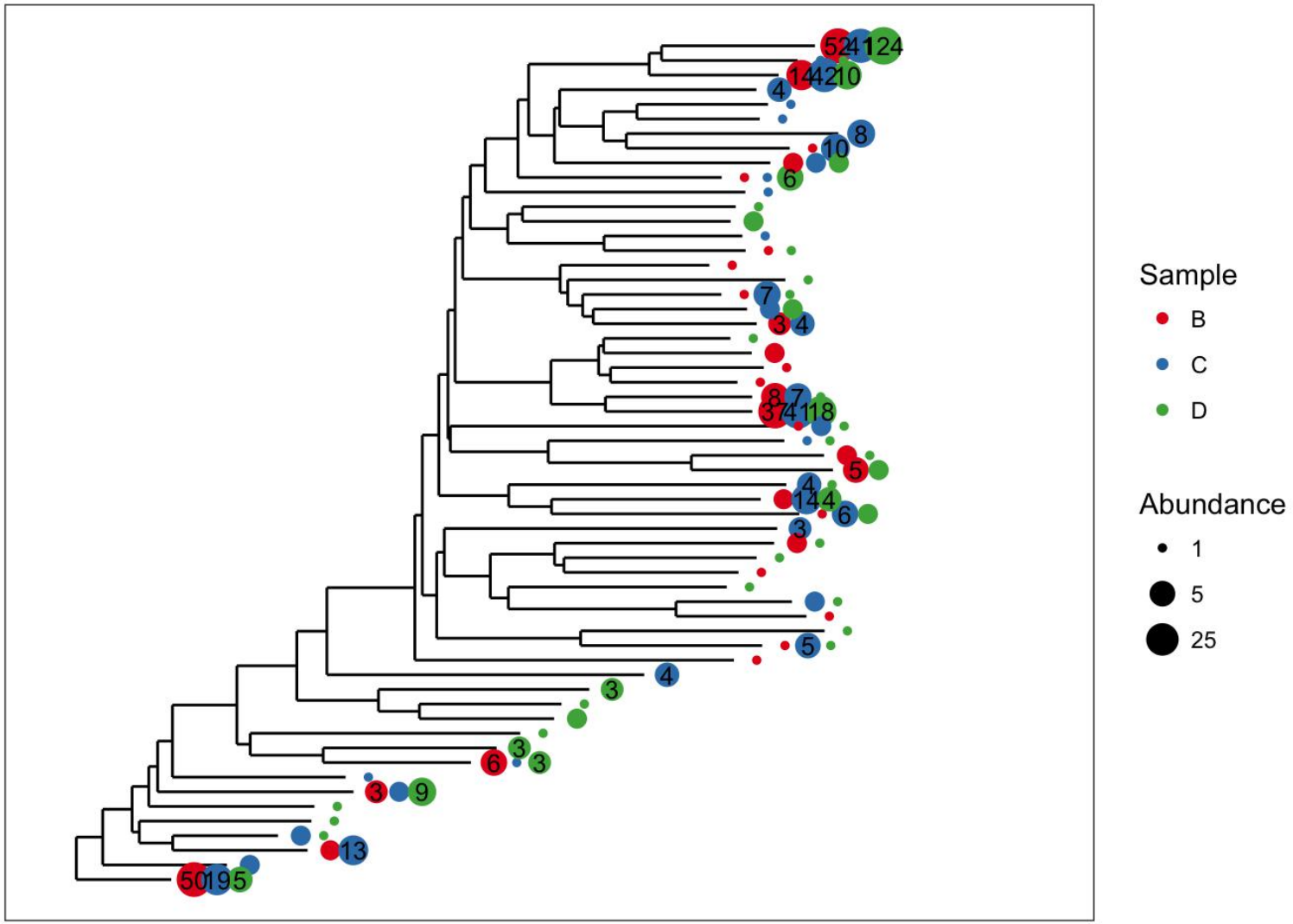
Accuracy

FastTree is slightly more accurate than PhyML3 with NNI moves because it has a better starting tree (thanks to the minimum-evolution SPR moves). FastTree is much more accurate than mi likelihood methods that do a more intensive search of topology space, such as PhyML with SPR moves or RAxML. However, for large alignments, the more accurate methods are orders-of- poor support.

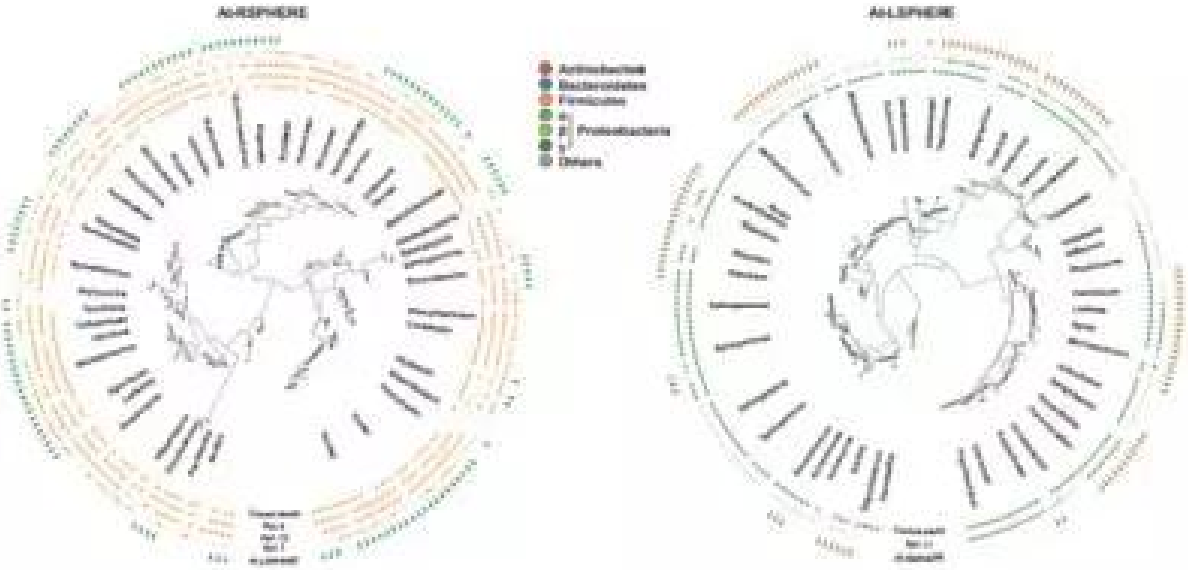
Topological accuracy for simulated alignments with varying numbers of sequences

#Sequences	250	1,250	5,000	78,132
Type	a.a.	a.a	a.a.	nt
RAxML 7 (JTTCAT + SPRs)	90.5%	88.4%	88.4%	--
PhyML 3.0 (Γ_4 + SPRs)	89.9%	--	--	--
FastTree 2.0.0 (JTT+CAT or JC+CAT)	86.9%	83.7%	84.3%	92.1%
PhyML 3.0 (Γ_4, no SPR)	86.0%	--	--	--
PhyML 3.0 (no gamma, no SPR)	81.7%	80.1%	--	--
FastME 1.1 (log-corrected distances)	79.6%	77.7%	75.3%	--
BIONJ (max-lik. distances)	77.7%	73.7%	73.1%	--
Parsimony (RAxML)	76.8%	76.5%	69.4%	--
BIONJ (log-corrected distances)	76.6%	73.0%	72.3%	--
Neighbor-joining (log-corrected distances)	76.0%	72.6%	71.6%	66.1%
Clearcut 1.0.8 (log-corrected distances)	75.5%	72.3%	71.5%	58.1%

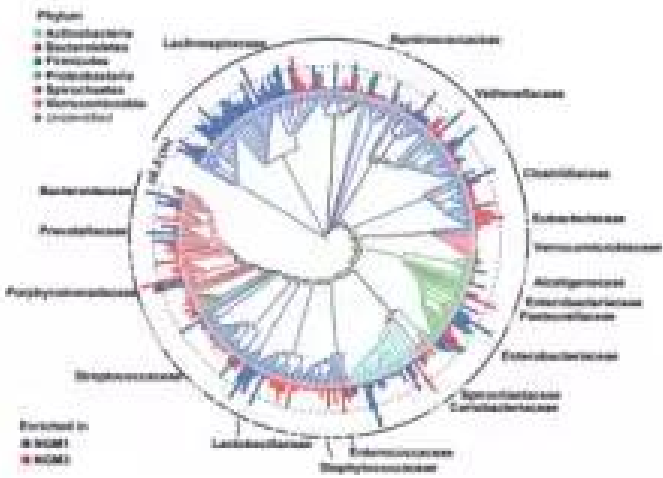
ggplot：物种进化关系分析



iTOL：物种进化关系分析



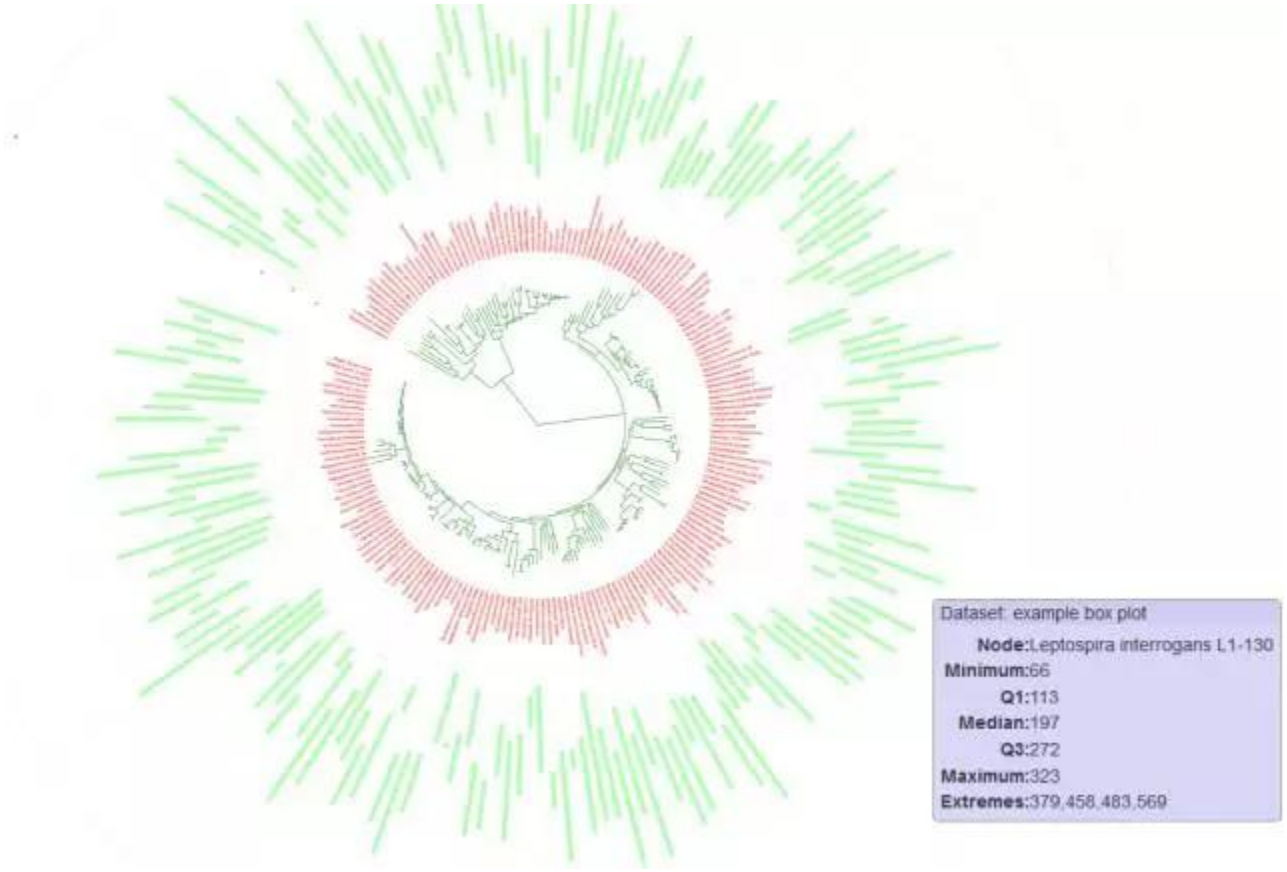
Nature 528, 364–369 (2015)



Nature Medicine 22, 1187–1191 (2016)

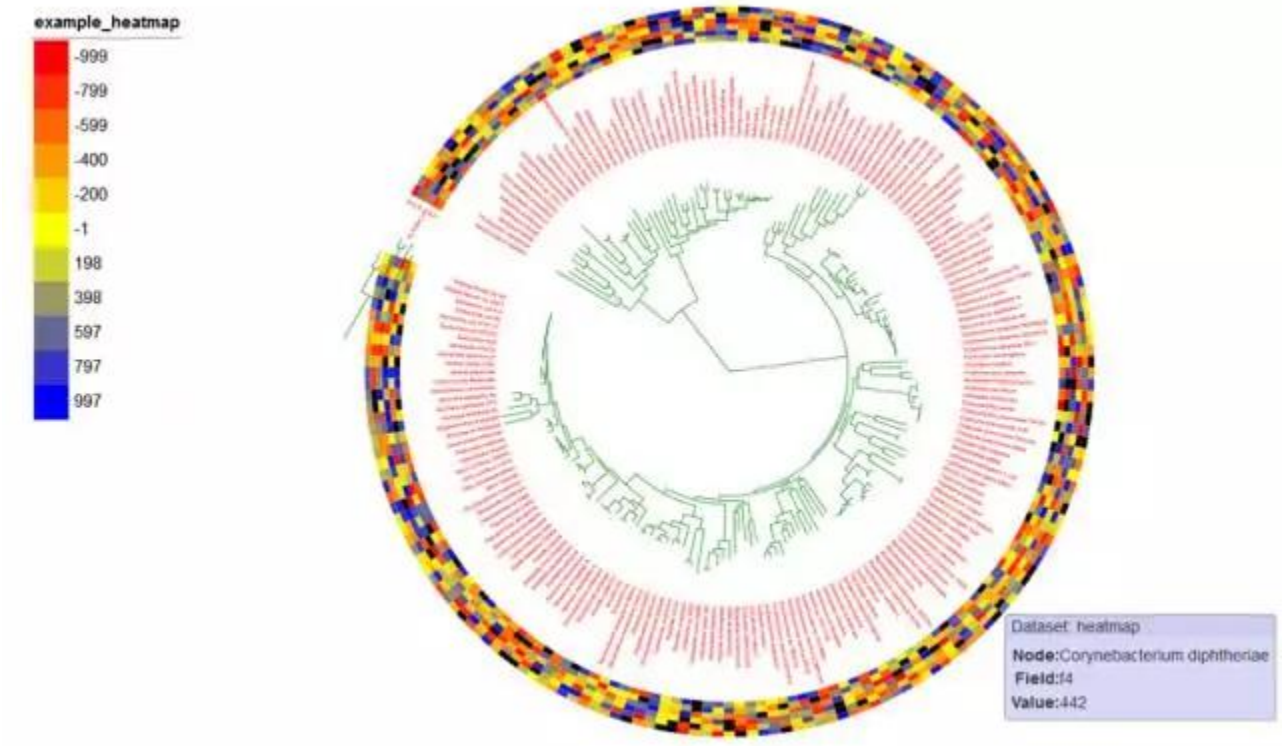
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



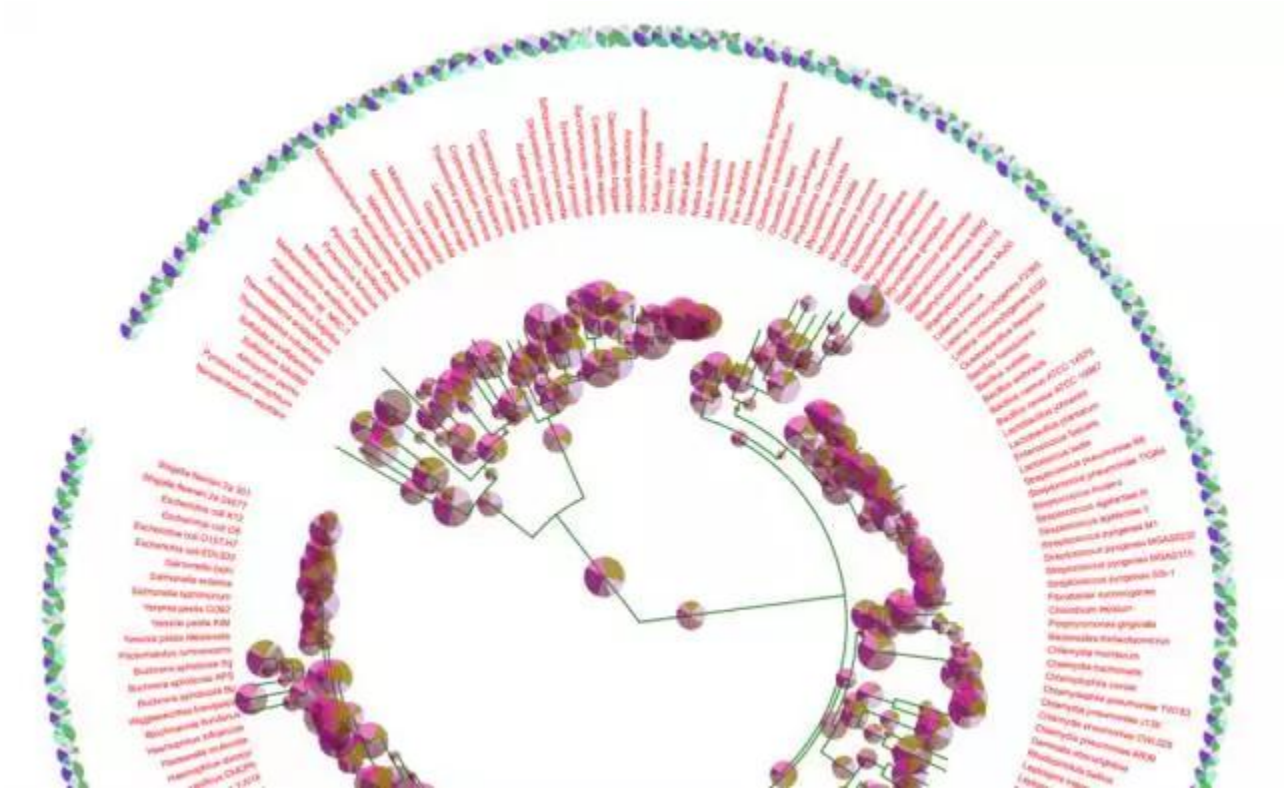
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



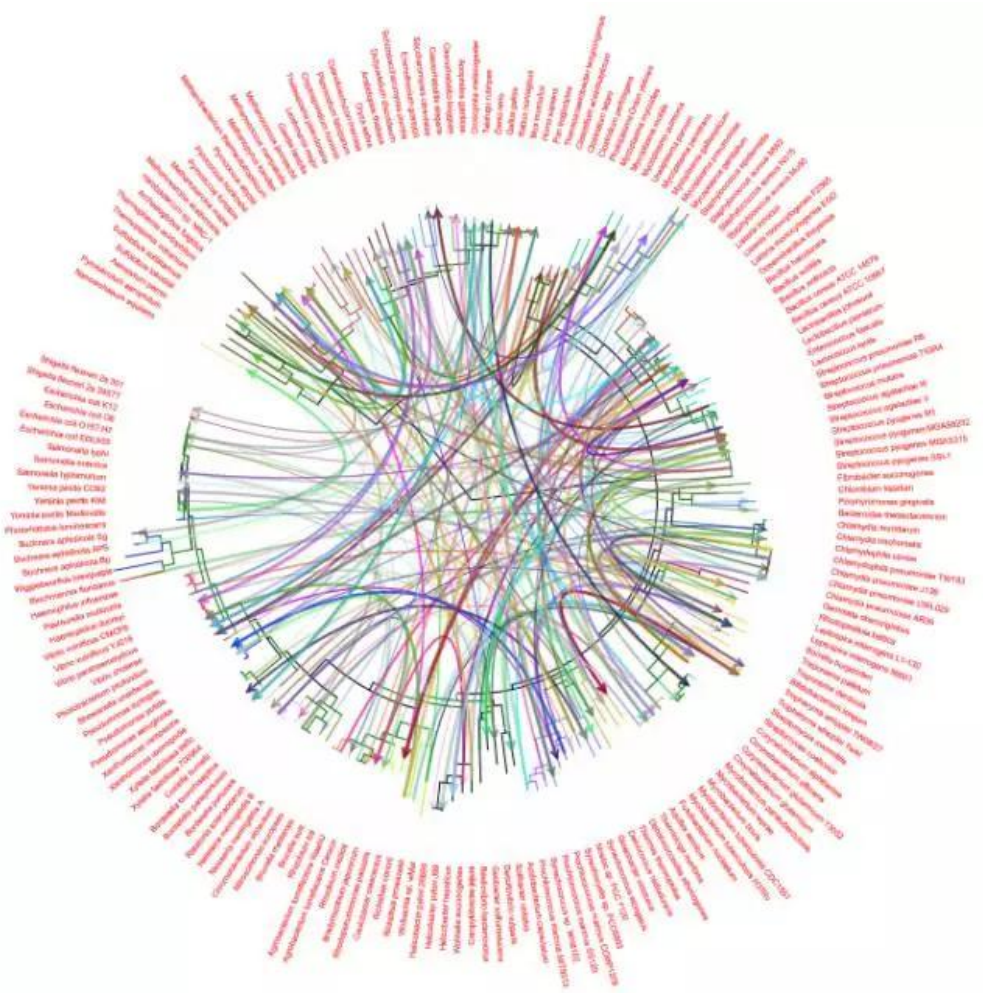
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



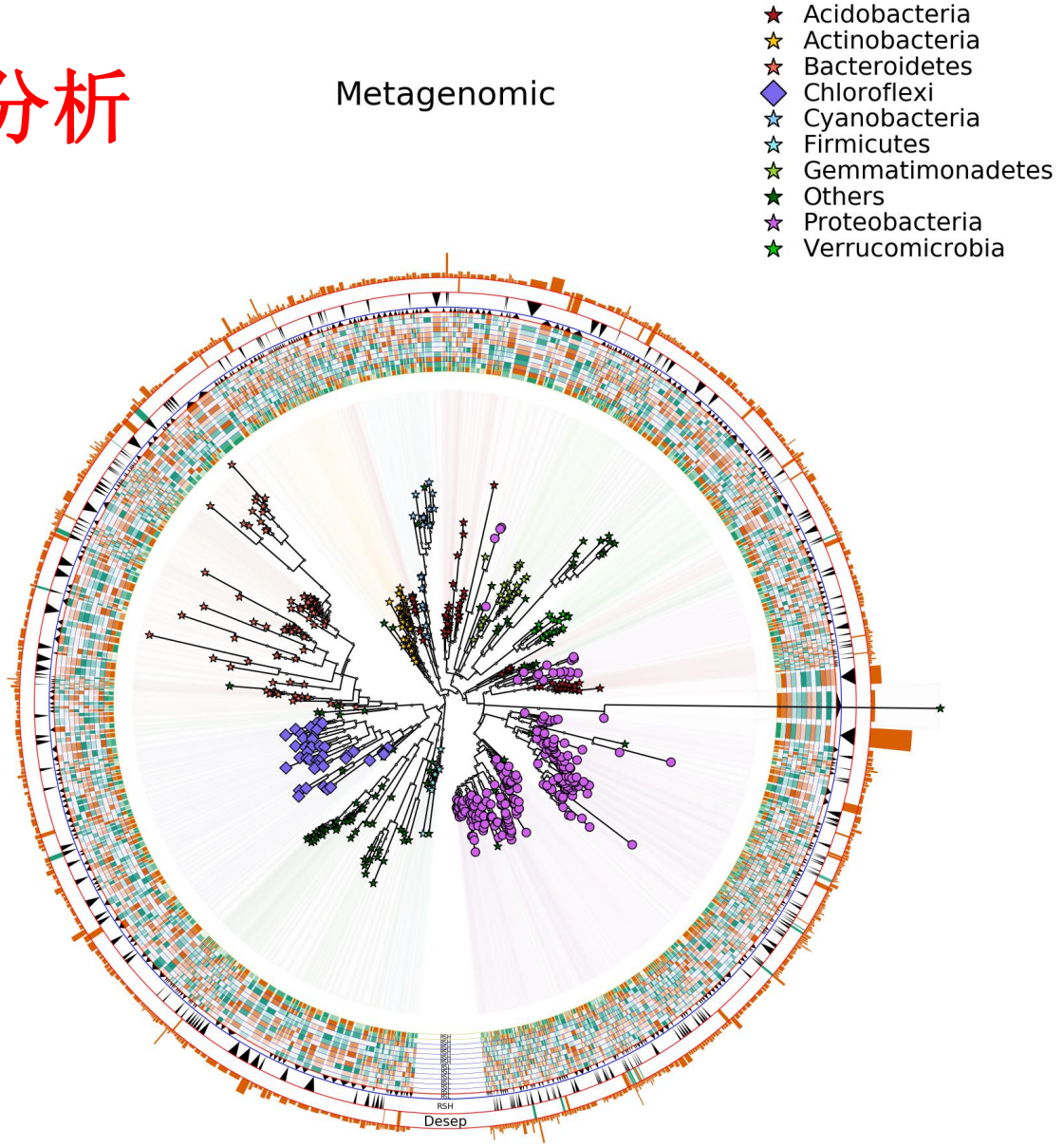
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



iTOL: https://itol.embl.de/

GraPhlAn：分类树分析



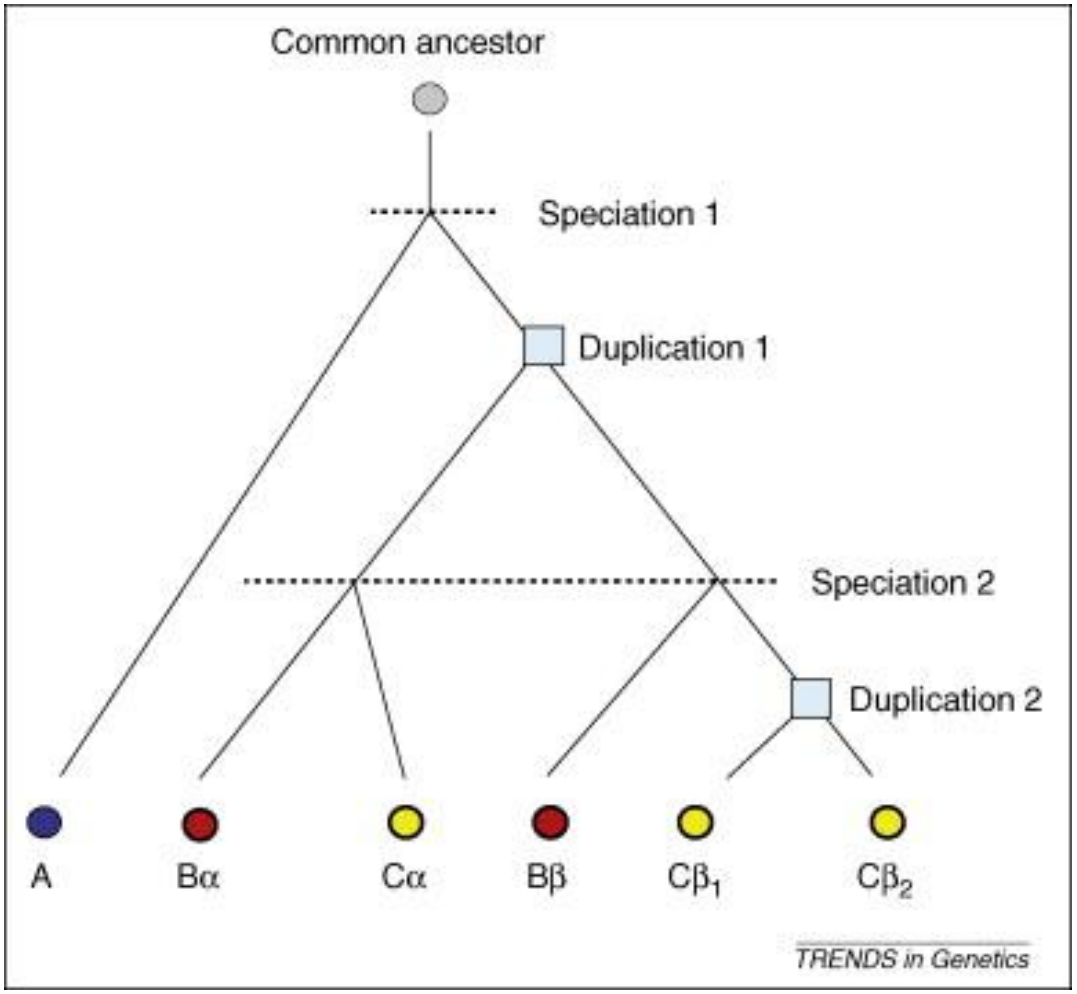
构建分子进化树相关的软件

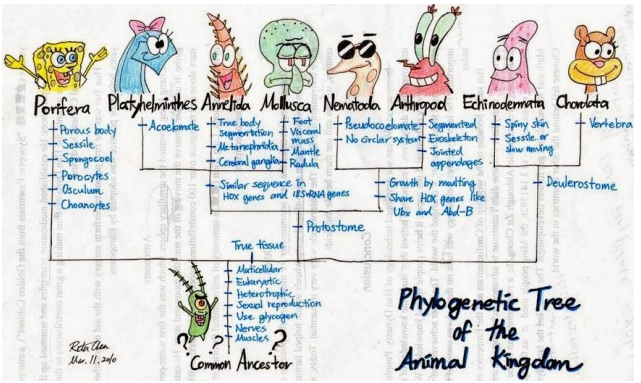
- 就进化树而言，iTOL功能最为全面。iTOL无限制添加的数据集，外环可以制作各种图形包括箱线图等，可以使用多种符号填充外环。但Graphlan就不能这么随便了，只能使用两个符号填充外环。再多的环属性也就是设置环数量和颜色，透明度了。

- ggtree最容易上手，但是就一张圈图来说，它不能添加除了热图以外的其他图形，但是在非圈图的模式下，可以对多种数据进行合并，方法将更为简单，操作也容易一些。

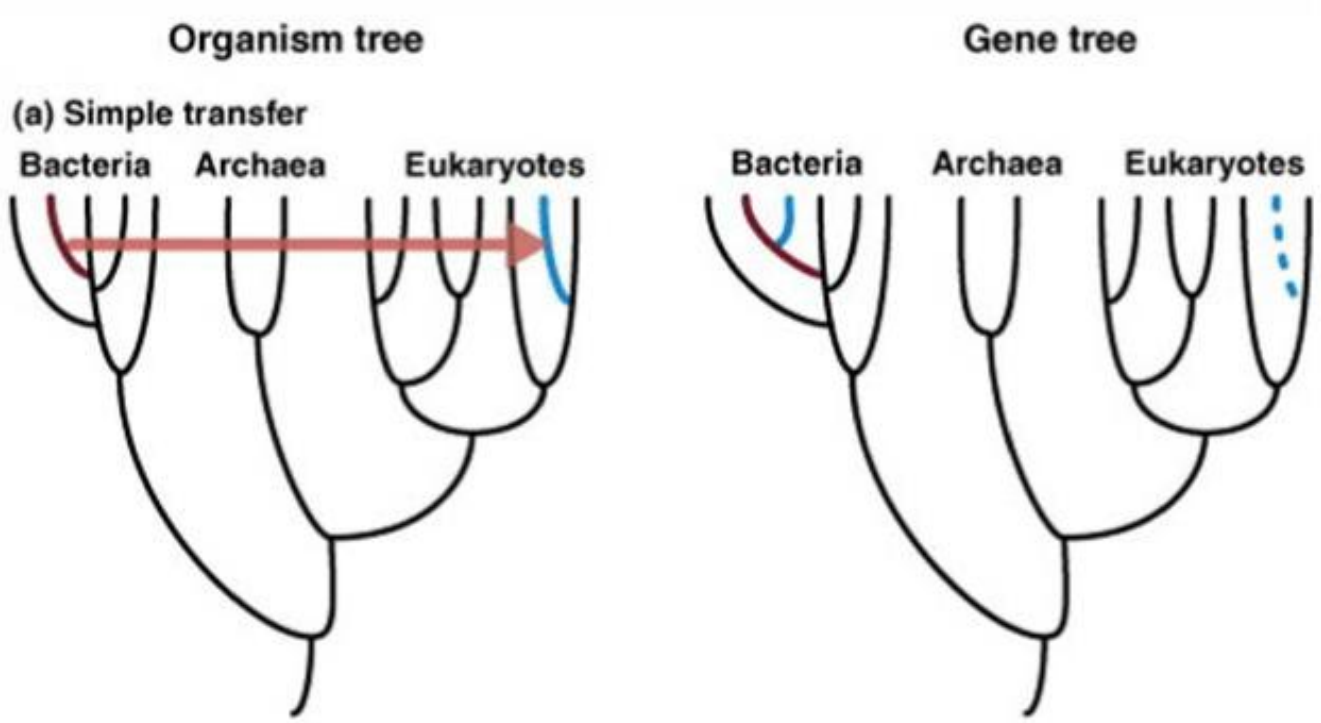
- Graphlan可以制作分类树，是它不同于R包ggtree和iTOL的地方。

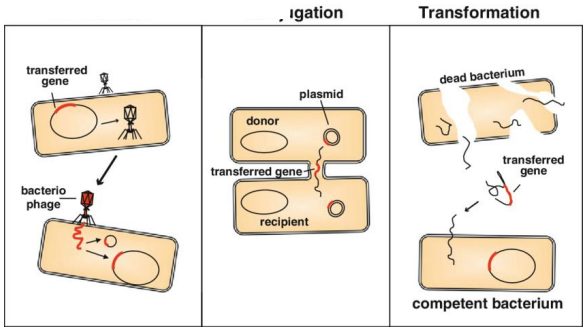
Gene: ortholog and paralog





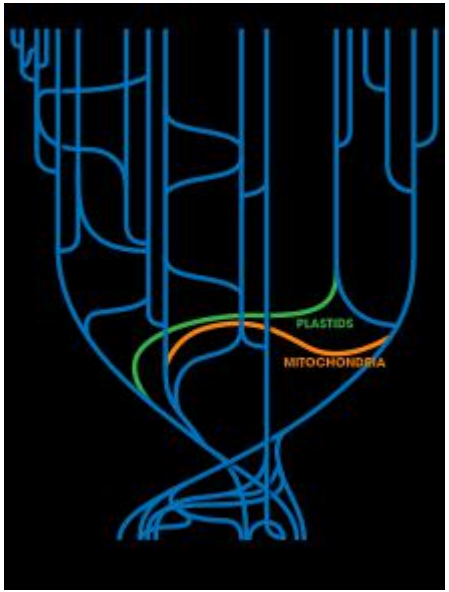
Gene: HGT

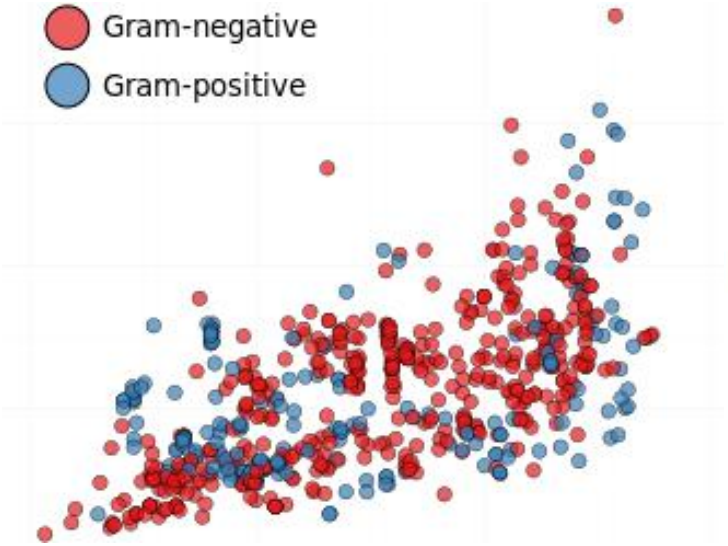

Organism tree
Gene tree
(a) Simple transfer
Bacteria
Archaea
Eukaryotes
Bacteria
Archaea
Eukaryotes


igation
Transformation
transferred gene
bacterio phage
plasmid
donor
transferred gene
recipient
dead bacterium
transferred gene
competent bacterium

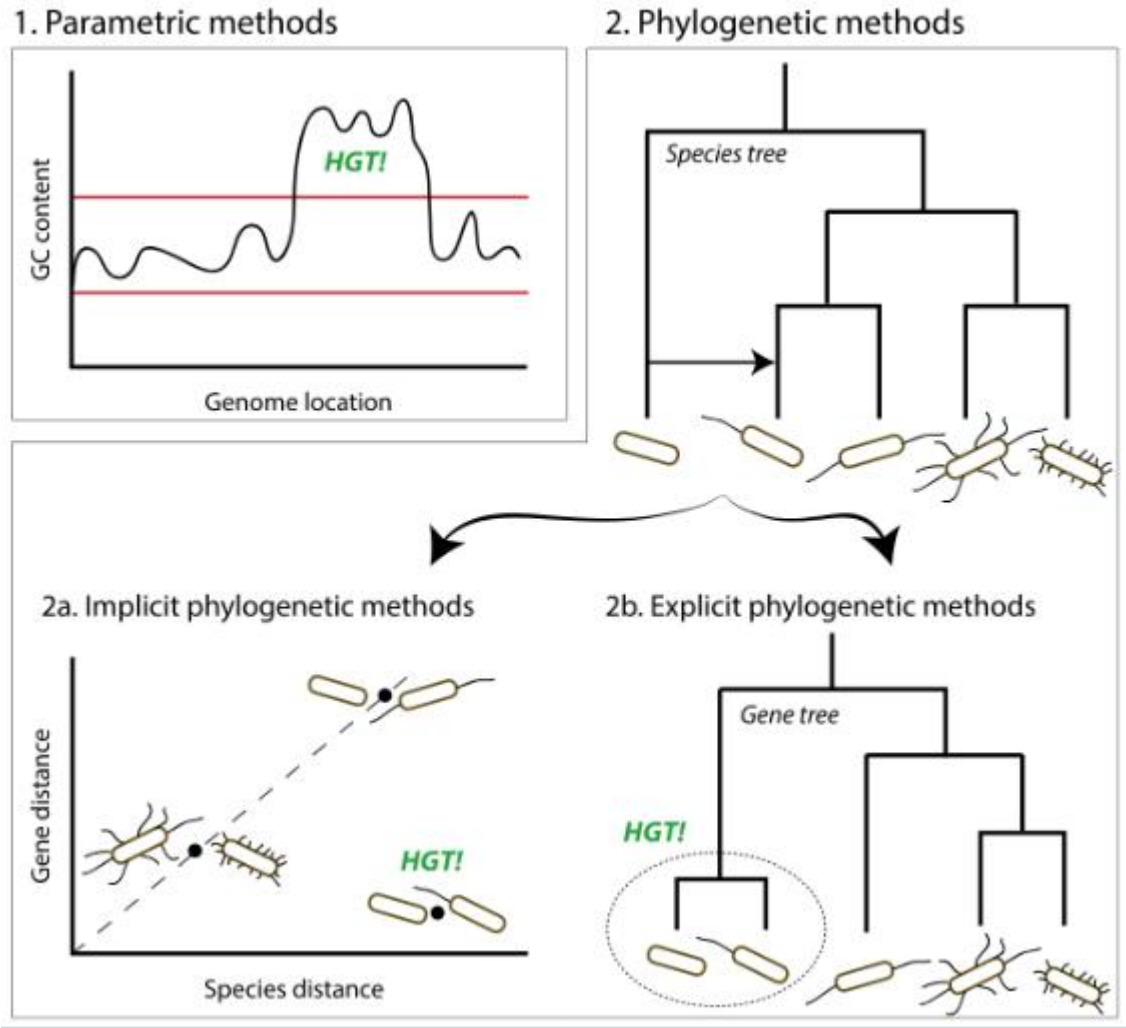
Gene: HGT

Think about it:
how does HGT affect the molecular clock calculation?

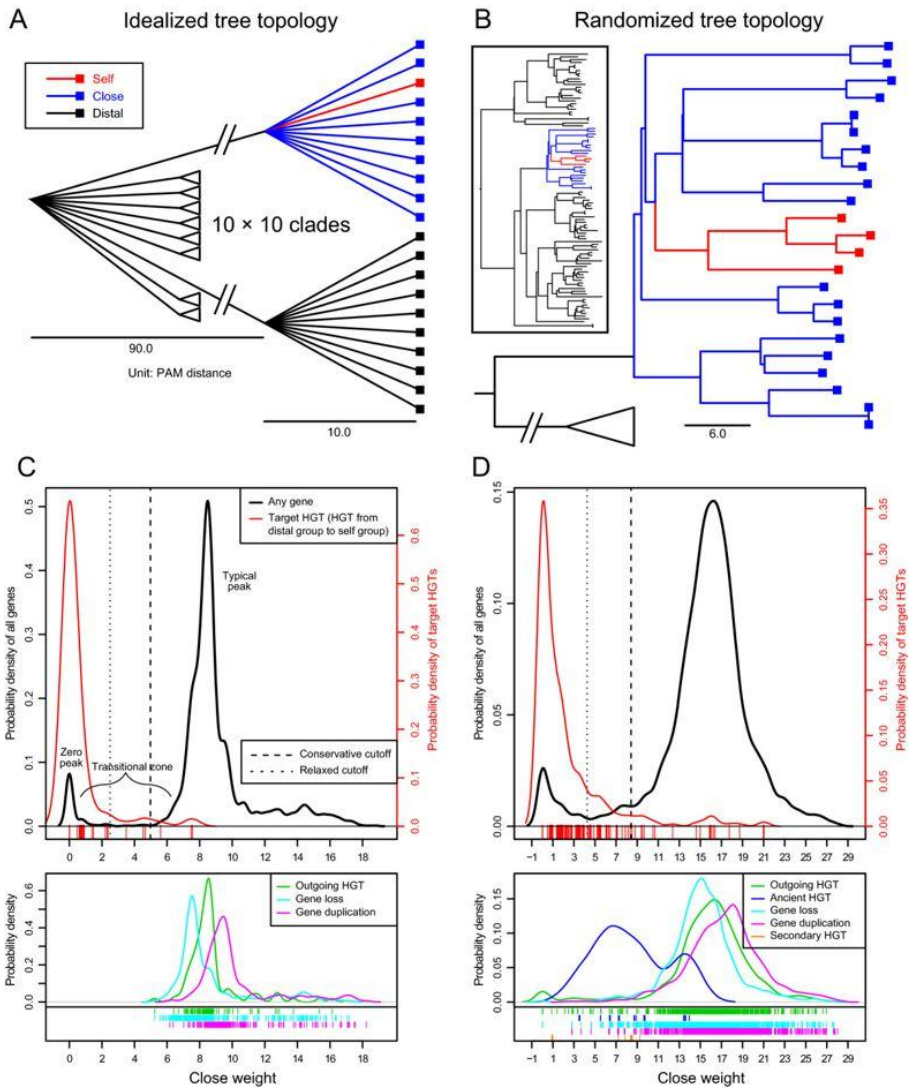




HGT identification methods (phylogeny analysis)



HGT identification methods (phylogeny analysis)



参考文献

- R. Durbin, S. Eddy, A. Krogh and G. Mitchison. Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids. 1998, Cambridge University Press.

一、总括

系统发生学(phylogenetics)

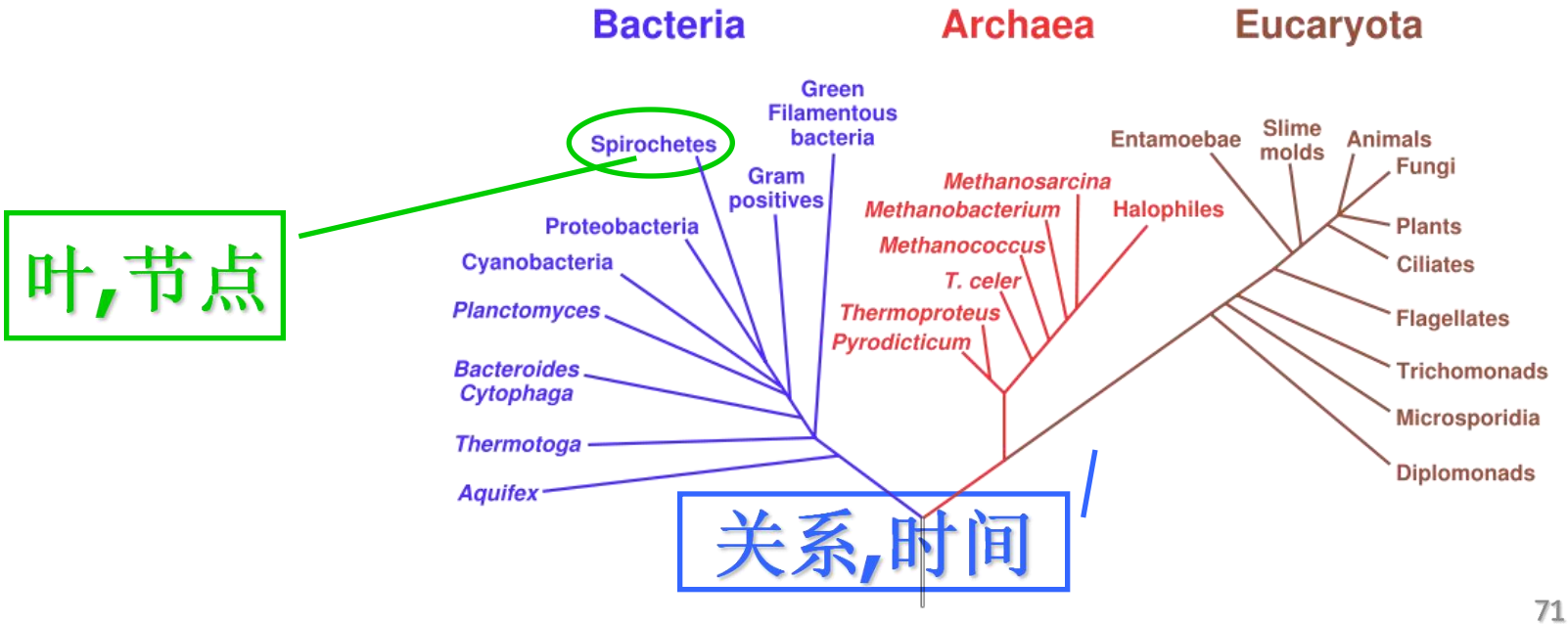
- 亦称系统学,种系发生学,种系发生系统学(phylogenetic systematics)
- 在希腊文中
  - *phylon* = tribe, race(种系)
  - *genesis* = birth
- *研究生物群体(如:物种,种群)之间的进化关系*

相关概念

- phylogenetic taxonomy(系统发生分类学)
  - 是系统学的一个分支
  - *根据进化相关度对生物群体分类*
- phylogeny (=phylogenesis系统发生)
  - *生物群体的产生和进化*
- 分子系统学(molecular phylogenetics)
  - 将核酸,氨基酸序列作为进化特征

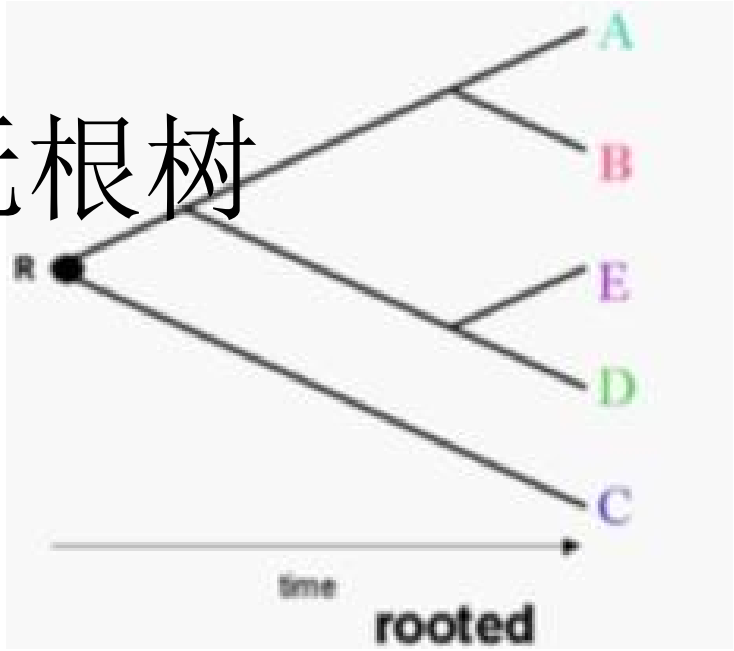
系统发生树(phylogenetic tree)

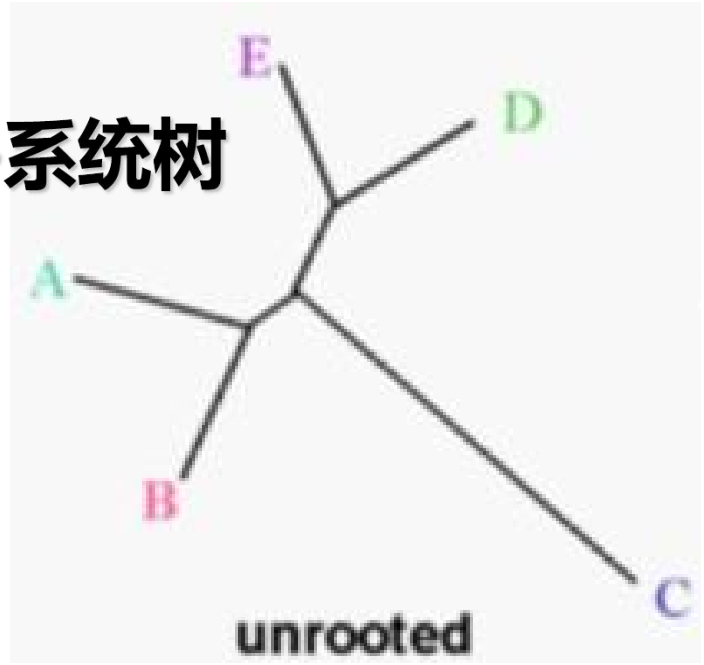
- 也叫系统树,进化树(evolutionary tree),生命树(tree of life)
- *对物种之间的进化关系的一种描述,这些物种被认为有共同祖先*



有根树和无根树

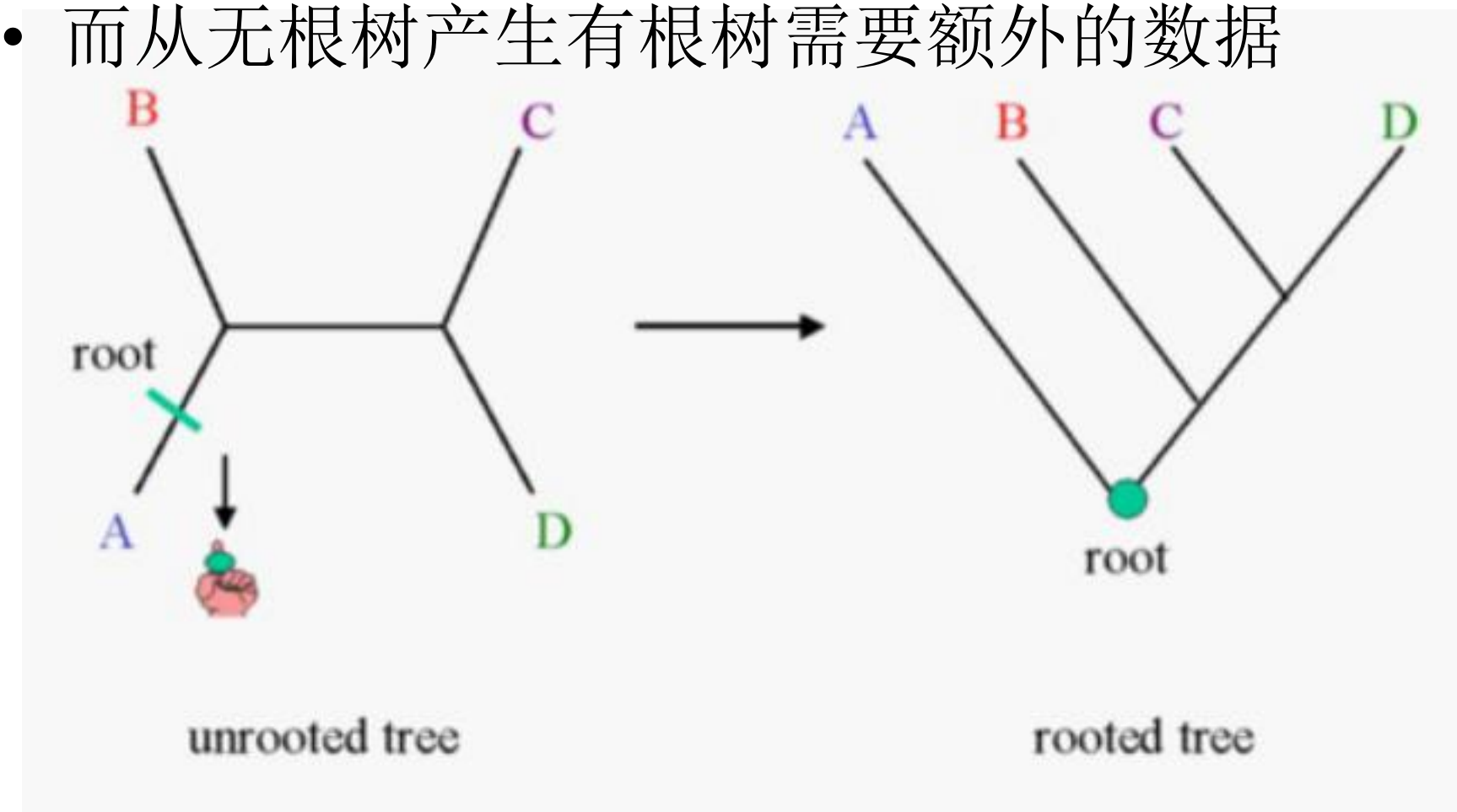
- 有根树(rooted tree)
  - 有共同祖先
- 无根树(unrooted tree)
- 树空间(tree space)
  - 从已知序列可以产生许多系统树
  - 来自几何





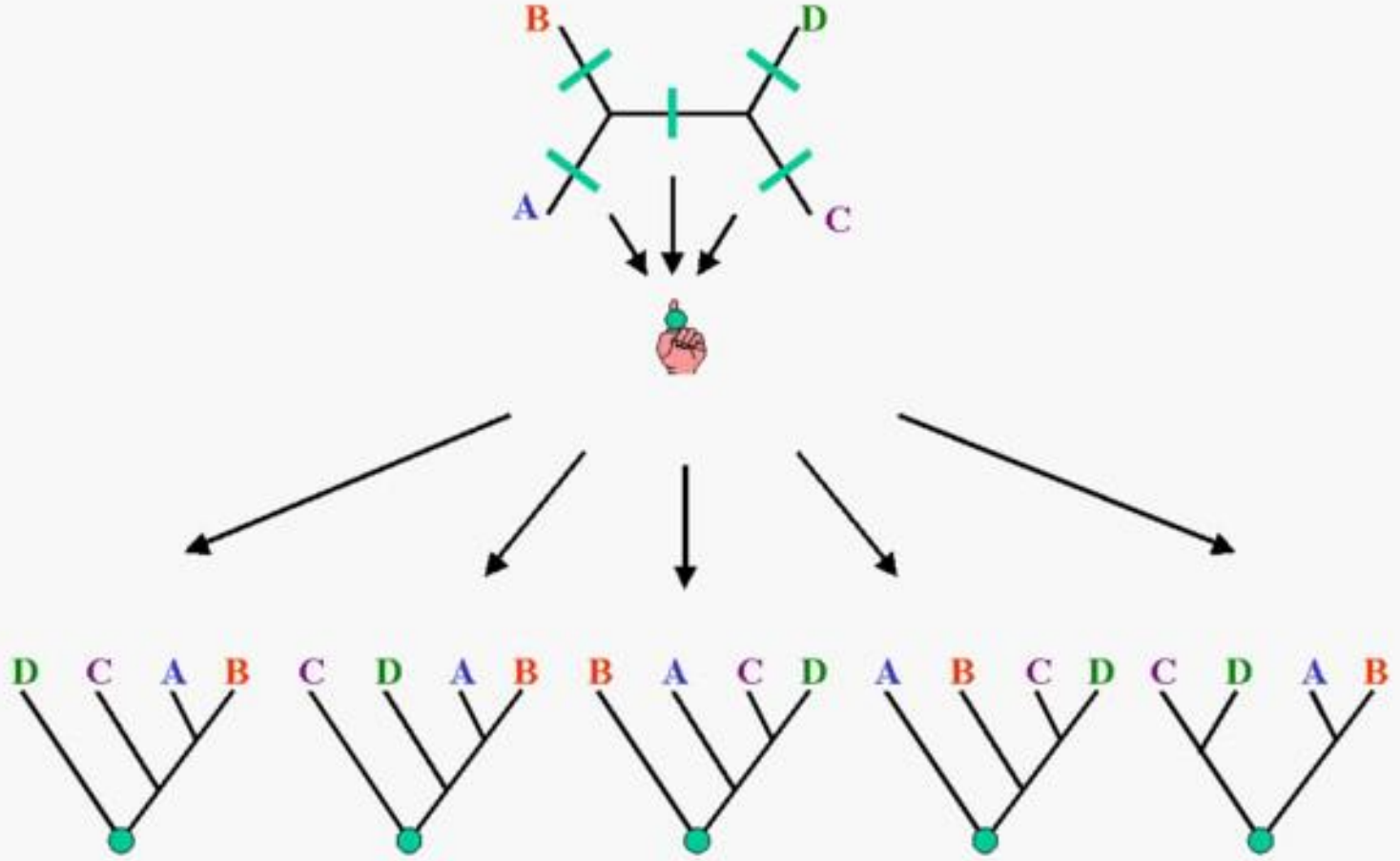
无根树和有根树的关系(1)

- 从一棵有根树总可以产生一棵无根树
- 而从无根树产生有根树需要额外的数据



无根树和有根树的关系(2)

- 一棵无根树可以产生多棵有根树



Willi Hennig (1913-1976)

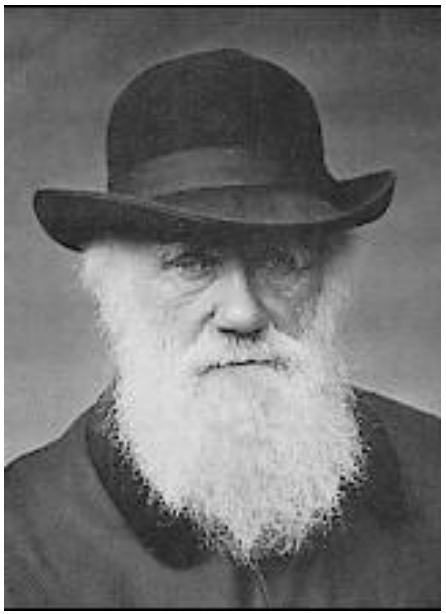
- 德国生物学家,被认为是系统发生学和分类学(cladistics; 也叫cladogram)的奠基人
- 据已知资料来看,他的观点并不是最早被阐述
- 属达尔文学派;类似的观点另一学派的Lamarck和Rosa也有阐述
- 可以认为是系统发生学的集大成者

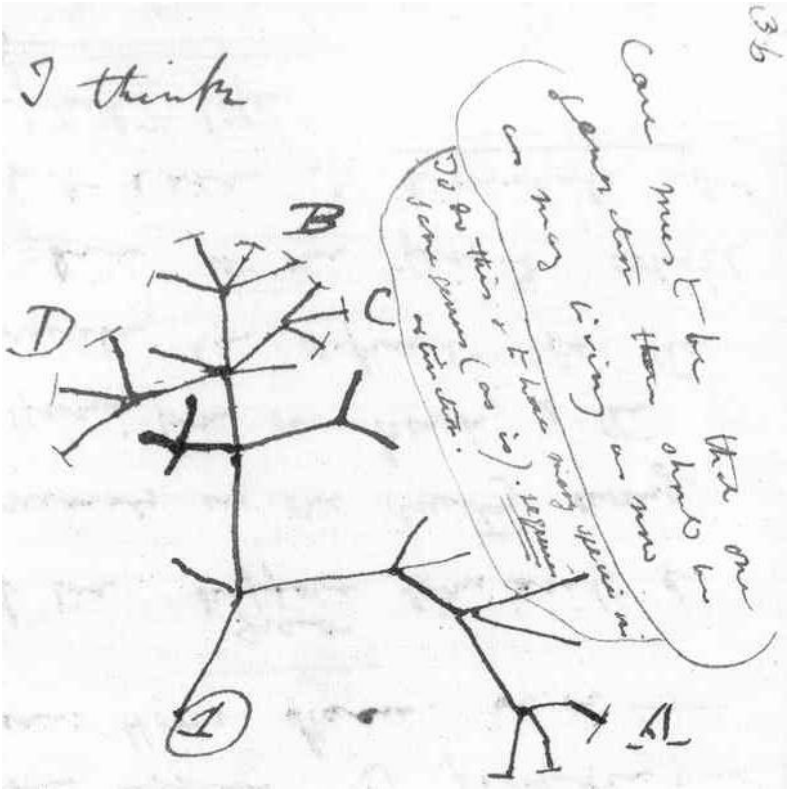
历史上的系统树

- **海克尔(Ernst Haeckel)首次制成了当时所有已知生物的系统树**
  - **(1834-1919)著名的德国生物学家,哲学家,医生,教授,艺术家**
  - **创建了重演论(recapitulation theory)**
  - **命名了许多生物学术语(如:门,系统发生,生态学,原生生物)和几千物种**
  - **出版了著名的*Kunstformen der Natur* (*Artforms of Nature*)**

可能是最早的系统树

- C. Darwin,1837





特征选取的变迁

- 经典系统发生学
  - 主要是比较大的物理或表型特征
  - 如生物体的大小,颜色,牙齿个数,行为特征
  - 缺点: 不易量化(连续), 难以选取合适特征
- 现代系统发生学
  - 分子水平: 核酸或氨基酸序列
  - 优点: 易量化(离散),易获取,适于自动化,更本质
  - 例子: (现代人起源) 通过对线粒体DNA的研究,认为所有现代人都是一个非洲女性的后代(“夏娃”)

系统发生学研究方法

- 目的
  - 在树空间中寻找正确的系统树
- 分析步骤
  1. 多(重)序列比对(multiple sequence alignment, MSA)
  2. 构建系统树
  3. 评价结果

三种构建系统树中使用的搜索算法

- 穷尽法
  - 搜索整个空间(所有可能的树),然后根据评价标准选择一棵最优的树
- 分支约束方法
  - 根据一定的约束条件将搜索空间限制在一定范围内
- 启发式或经验性方法(heuristic)
  - 根据目前的搜索情况指导下一步的搜索方向
  - 根据先验知识或一定的指导性规则压缩搜索空间

两类数据: 距离和离散特征

- 距离
  - *描述序列之间的差别* (遗传距离)
  - 一般用距离矩阵(distance matrix)表示
  - 距离往往由序列比对产生(如错配的比例)
- 离散特征
  - 二态特征 (如: DNA序列上的某个位点是否剪切位点)
  - 多态特征 (如: 某一位点可能的碱基有A,T,G,C)

两大类构建系统树的算法

1. 基于距离的构建方法 (distance-matrix methods)
   - 邻近归并法 (或称邻接法,neighbor-joining)
     - 非加权组平均法 (UPGMA)
   - Fitch-Margoliash法
   - 最小进化方法
2. 基于离散特征的构建方法
   - 最大简约法 (MP)
   - 进化简约法 (EP)
   - 最大似然法 (ML)
   - 相容性方法

注意: 系统树的限制

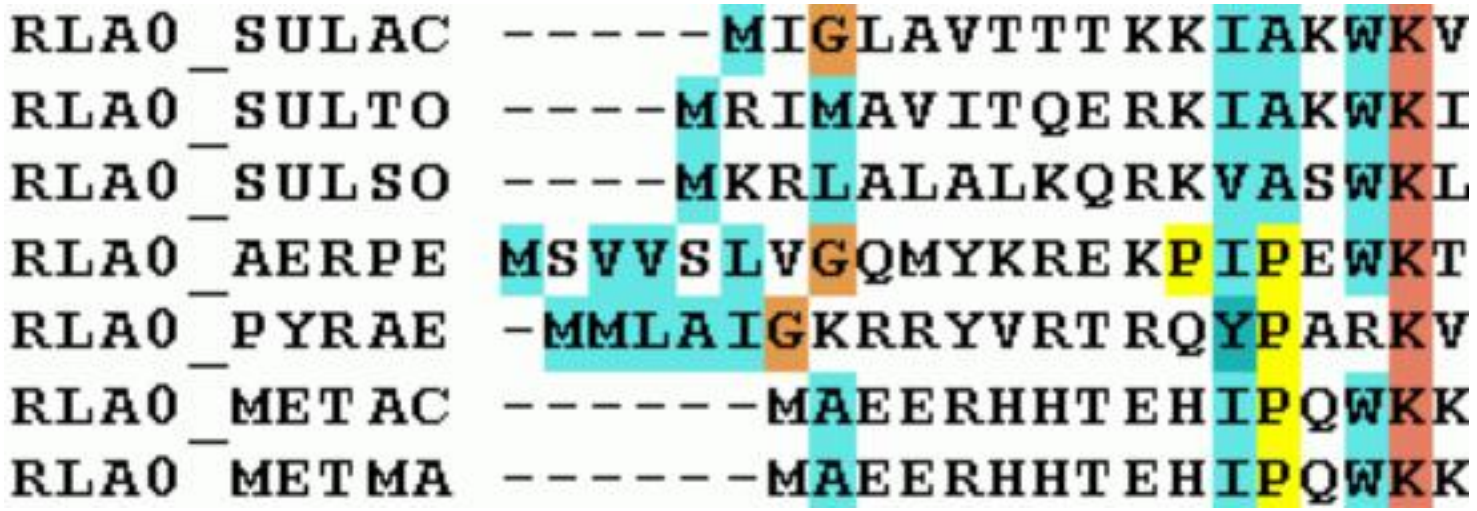
- 有人认为生物的系统关系不一定是树状的
- 系统树不一定代表进化历史

1. 有很多干扰分析的因素
   - 噪音(noisy)
   - 水平基因转移(horizontal gene transfer;网状)
   - 杂交,重组等 (网状)
2. 用不同基因或蛋白产生的树往往不同
3. 已经灭绝的物种只能作为叶节点

二、多序列比对

例子

- 多物种核糖体Rplp0蛋白比对



ClustalW生成(颜色表示氨基酸保守性)

多序列比对方法

- 动态规划(dynamic programming)
  - 慢,耗内存
  - 改进: 使用“sum of pairs”目标函数
- 渐进法(progressive method;或称分级法 hierarchical,建树法tree)
- 迭代法(iterative method)
- 基序法(motif finding;或称轮廓分析法profile analysis)
- 来自计算科学的算法
  - HMM, GA, SA
- 星形比对,树形比对

动态规划法

- 是两两比对所用动态规划方法的直接扩展
- 步骤
  1. 用两两比对的方法比对所有的序列对
  2. 建立n维矩阵(n为序列个数)
  3. 产生多序列比对
- 优点
  - 理论上适用于任意多个序列
  - 保证能得到较好结果
- 缺点
  - 耗费大量时间,内存
  - 实际上很少用于多于3个序列的比对

逐对加和法(sum of pairs, SP)

- 步骤
  1. 进行所有两两比对,并给每个比对打分
  2. 将所有的得分相加
  3. 找到最优多序列比对,使得总得分(目标函数objective function)最高
- 例子
  - 对于这个蛋白多序列(3个)比对,求总分
  - 已知得分(K,R)=3,间隔罚分为-12

(-12)+(-12)+3=-21

K
–
R

Clustal

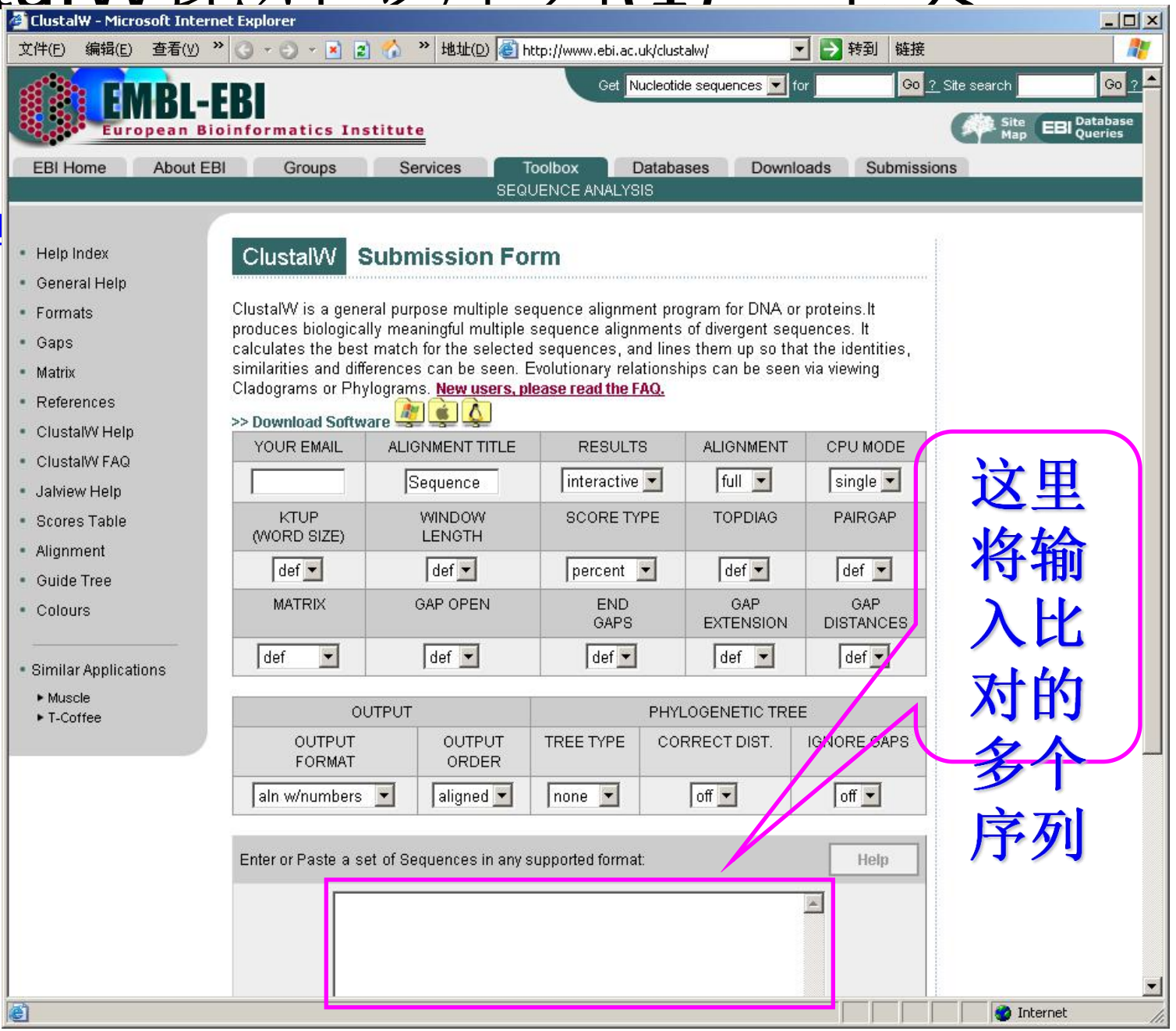
- 可能是使用最广的多序列比对软件
- 算法
  1. 用Needleman-Wunsch全局算法做所有两两比对
  2. 得到距离矩阵,从而产生引导树 (guide tree; 利用UPGMA,见后; 得到dnd文件)
  3. 渐进式比对 (先处理距离最近的2个序列,再加次最近的...; 得到aln文件)
- 两个主要形式
  - ClustalW (命令行)
  - ClustalX (图形用户界面GUI)
- 适用于Windows,Mac OS,Unix/Linux

Clustal的输入输出文件格式

- **输入**
  - **FASTA**
  - **Clustal**
  - **NBRF/PIR**
  - **GCC/MSF**
  - **GDE**
  - **EMBL/Swissprot**
  - **GCG9 RSF**
- **输出**
  - **PHYLIP**
  - **Clustal**
  - **NBRF/PIR**
  - **GCG/MSF**
  - **GDE**
  - **NEXUS**

ClustalW比对多序列(1) – 主页

http://www.ebi.ac.uk/clustalw/



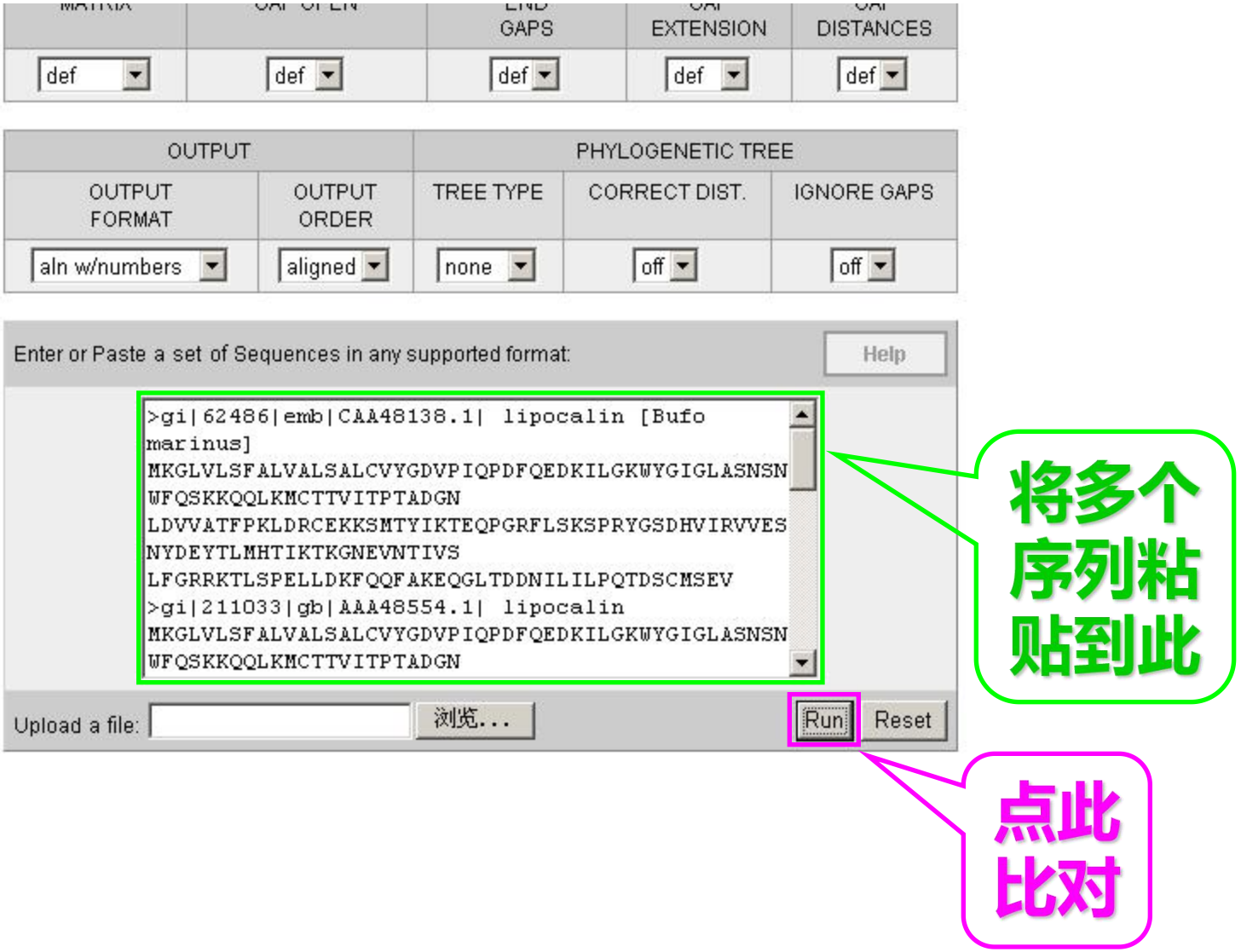
ClustalW (2) – 获取FASTA格式的序列

选择格式

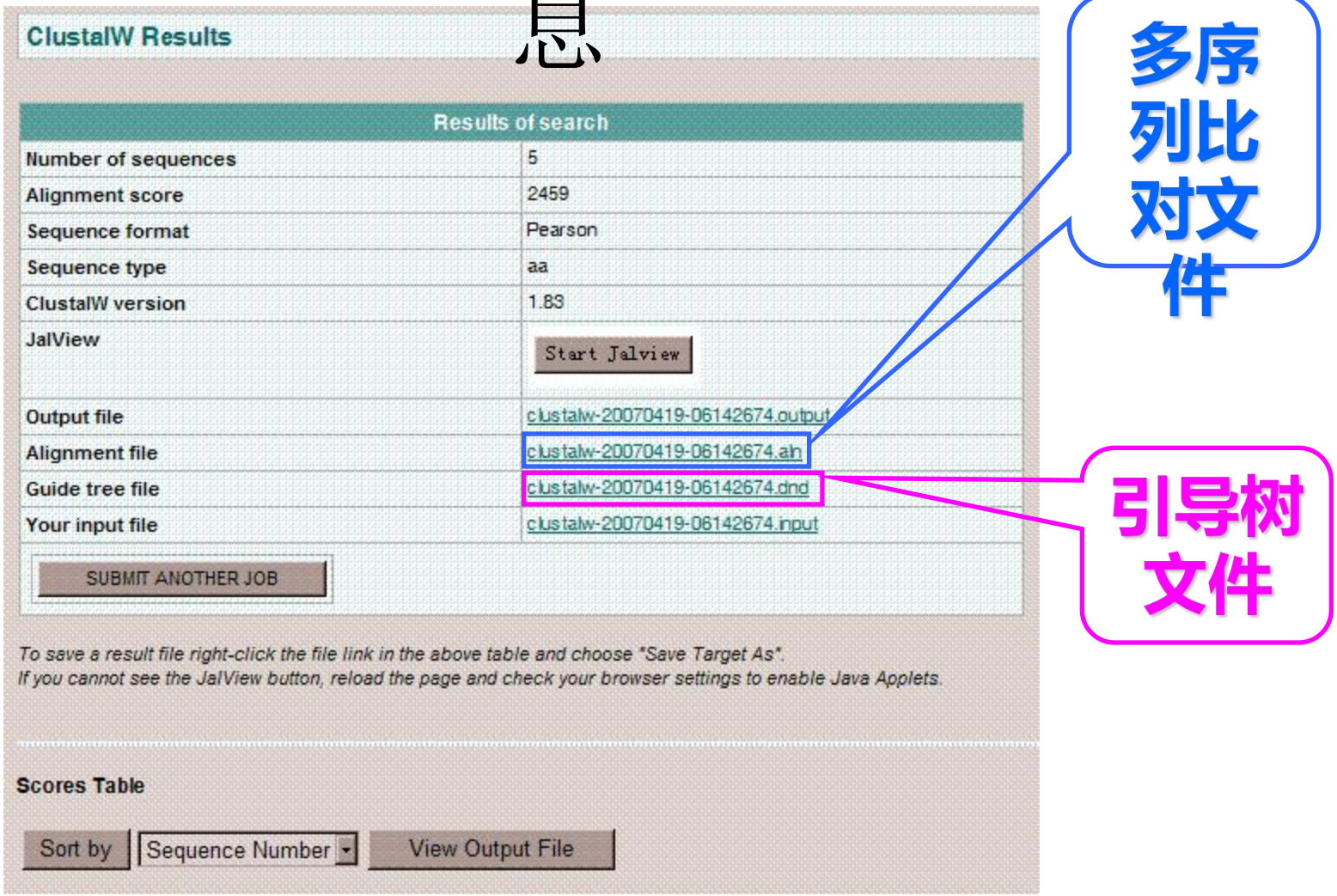
或将这里改为Text,更易拷贝

拷贝这些部分

ClustalW (3) – 将多个序列输入



ClustalW (4) – 比对结果(1) 基本信息



ClustalW (5) – 比对结果(2) 比对图

Alignment

Hide Colors | View Alignment File

```
CLUSTAL W (1.83) multiple sequence alignment

gi|62486|emb|CAA48138.1|          MKGLVLSFALVALSALCVYG--------------DVPIQPDFQEDKILGK 36
gi|211033|gb|AAA48554.1|          MKGLVLSFALVALSALCVYG--------------DVPIQPDFQEDKILGK 36
gi|1478202|emb|CAA57283.1|        MALSVMCLGLALLGVLQSQAQDSTQNLIPAPSLLTVPLQPDFRSDQFRGR 50
gi|28933463|ref|NP_084235.1|      -MVLLLVLGLVLSLATAQFN-------------LHTAVRRDYNLARISGT 36
gi|30794262|ref|NP_828875.1|      -MKLEMALSIALALAVVSWT-----------QEFFPKEAQTLNWSKFSGF 38
                                       : :.:.   .                          .  :: *

gi|62486|emb|CAA48138.1|          WYGIGLASNSNWFQSKKQQLKMCTTVITPTADGNLDVVATFPKLDR--CE 84
gi|211033|gb|AAA48554.1|          WYGIGLASNSNWFQSKKQQLKMCTTVITPTADGNLDVVATFPKLDR--CE 84
gi|1478202|emb|CAA57283.1|        WYVVGLAGNAVQKKTEG-SFTMYSTIYELQENNSYNVTSILVRDQDQGCR 99
gi|28933463|ref|NP_084235.1|      WYLDSIASDNMTRIEENGDLRLFIRNIKLLNNGSLQFDFHFMLQGE--CV 84
gi|30794262|ref|NP_828875.1|      WYIIAIATDTQGFLPARDKRKLGASVVKVHKTGQLRVVIAFSRPRG--CQ 86
                                  **  .:* :         .  :          ..  .   :       *

gi|62486|emb|CAA48138.1|          KKSMTYIKTEQPGRF-LSKSPRYG--SDHVIRVVESNYDEYTLMHTIKTK 131
gi|211033|gb|AAA48554.1|          KKSMTYIKTEQPGRF-LSKSPRYG--SDHVIRVVESNYDEYTLMHTIKTK 131
gi|1478202|emb|CAA57283.1|        YWIRTFVPSSRAGQFTLGNMHRYPQVQSYNVQVATTDYNQFAMVFFRKTS 149
gi|28933463|ref|NP_084235.1|      AVTMVCEKTKNNGEFSVAYEGKNK------VLLLETDYSMYIIFYMQNIK 128
gi|30794262|ref|NP_828875.1|      SREVTLKKDRKRPVFRNTLKGVKG------FHVLSTDY-TYGLVYLRLGR 129
                                      .     .   *               . :  ::*  : :..

gi|62486|emb|CAA48138.1|          GNEVNTIVSLFGRRKTLSPELLDKFQQFAKEQGLTDDNILILPQTDSCMS 181
gi|211033|gb|AAA48554.1|          GNEVNTIVSLFGRRKTLSPELLDKFQQFAKEQGLTDDNILILPQTDSCMS 181
gi|1478202|emb|CAA57283.1|        ENKQYFKITLYGRTKELSPELKERFTRFAKSLGLKDDNIIFSVPTDQCID 199
gi|28933463|ref|NP_084235.1|      NGTKTQVLALYGRSILLDKTHQREFENICNLYGLDSQNIIDMTKKDFCFL 178
gi|30794262|ref|NP_828875.1|      GGSNYKSLLLFNRQNISSFLSLREFLDTCHILQLT-KQATILPKDDSCAH 178
                                   .     : *:.*    .     .*   .:   *  .:       * *

gi|62486|emb|CAA48138.1|          EV-- 183
gi|211033|gb|AAA48554.1|          EV-- 183
gi|1478202|emb|CAA57283.1|        N--- 200
gi|28933463|ref|NP_084235.1|      ----
gi|30794262|ref|NP_828875.1|      TILP 182
```

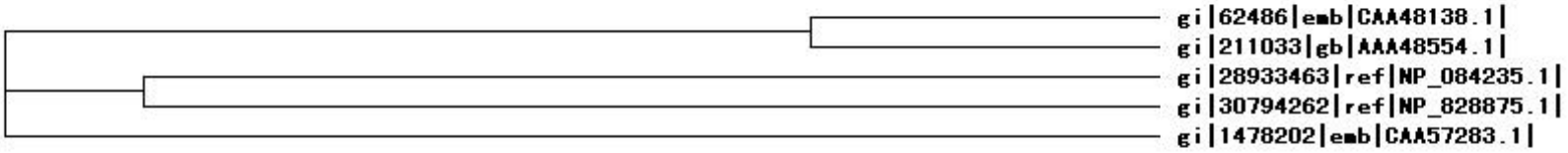
ClustalW (6) – 比对结果(3)引导树

Guide Tree

Show as Phylogram Tree | Show Distances | View DND File

```
(
(
gi|62486|emb|CAA48138.1|:0.00000,
gi|211033|gb|AAA48554.1|:0.00000)
:0.33984,
(
gi|28933463|ref|NP_084235.1|:0.37336,
gi|30794262|ref|NP_828875.1|:0.41877)
:0.05858,
gi|1478202|emb|CAA57283.1|:0.41426);
```

Cladogram



Show as Phylogram Tree | Show Distances | View DND File

Right-click on the above tree to see display options.
Problems printing? Read how to print a Phylogram or Cladogram.

MSA数据库

- Pfam (profile HMM library)
- SMART
- CDD (HMM; NCBI DART; =Pfam+SMART)
- BLOCKS (HMM)
- PRINTS
- PROSITE
- PopSet
- DOMO (Gapped MSA)
- PRODOM (PSI-BLAST)
- MetaFAM
- INTERPRO
- iProClass

MSA软件(维基的列表)

NAME	Description	Sequence Type	Alignment Type	Link	Author	Year
MSA	Dynamic programming	Both	Local or Global	download	D.J. Lipman *et al.*	1989 (modified 1995)
MultAlin	Dynamic programming/clustering	Both	Local or Global	server	F. Corpet	1988
PSAlign	Alignment preserving non-heuristic	Both	Local or Global	download	S.H. Sze, Y. Lu, Q. Yang.	2006
ClustalW	Progressive alignment	Both	Local or Global	EBI PBIL EMBNet GenomeNet	Thompson *et al.*	1994
Kalign	Progressive alignment	Both	Global	server	T. Lassmann	2005
T-Coffee	More sensitive progressive alignment	Both	Local or Global	server	C. Notredame *et al.*	2000
AMAP	Sequence annealing	Both	Global	server	A. Schwartz and L. Pachter	2006
MAVID	Progressive alignment	Both	Global	server	N. Bray and L. Pachter	2004
Multi-LAGAN	Progressive dynamic programming alignment	Both	Global	server	M. Brudno *et al.*	2003
MUSCLE	Progressive/iterative alignment	Both	Local or Global	server	R. Edgar	2004
MAFFT	Progressive/iterative alignment	Both	Local or Global	GenomeNet MAFFT	K. Katoh *et al.*	2005
Geneious	Progressive/Iterative alignment; ClustalW plugin	Both	Local or Global	download	A.J. Drummond et al.	2005 / 2006
CHAOS/DIALIGN	Iterative alignment	Both	Local (preferred)	server	M. Brudno and B. Morgenstern	2003
PRRN/PRRP	Iterative alignment (especially refinement)	Protein	Local or Global	PRRP PRRN	Y. Totoki (based on O. Gotoh)	1991 and later
POA	Partial order/hidden Markov model	Protein	Local or Global	download	C. Lee	2002
SAM	Hidden Markov model	Protein	Local or Global	server	A. Krogh et al.	1994 (most recent 2002)
ProbCons	Probabilistic/consistency	Protein	Local or Global	server	C. Do *et al.*	2005
SAGA	Sequence alignment by genetic algorithm	Protein	Local or Global	download	C. Notredame et al.	1996 (new version 1998)
Ed'Nimbus	Seeded filtration	Nucleotides	Local	server	P. Peterlongo *et al.*	2006
RevTrans	Combines DNA and Protein alignment, by back translating the protein alignment to DNA.	DNA/Protein (special)	Local or Global	server	Wernersson and Pedersen	2003 (newest version 2005)

其他MSA软件

- Opal (Bioinformatics 23(13);2007/7/1;免费)
  - aligning alignments
- Murlet (Bioinformatics 23(13);2007/7/1;开源)
  - for RNA
- SQUINT (Bioinformatics 23(12);2007/6/1)
- Probalign (Bioinformatics 22(22))
- PileUp (全局渐进)
- PIMA (局部渐进)
- BaliBase (比较MSA算法)

- **AMAS**
- **CINEMA**
- **HMMT**
- **Match-Box**
- **Musca**

MSA算法比较

- 全局(global)算法往往优于局部(local)算法
- 迭代(iterative)算法(如PRRP, SAGA)往往优于渐进式(progressive)算法(如Clustal)

- (a recent review) Recent Evolutions of Multiple Sequence Alignment Algorithms. Cédric Notredame. PLoS Computational Biology. 3(8). 2007

三、构建系统树

非加权分组平均法

- UPGMA (Unweighted Pair Group Method with Arithmetic mean)
- 算法(基于距离)
  1. 使每个物种自成一类
  2. 执行下列循环
  - 寻找最小距离的两个类,建立一个新的聚类
  - 连接这两个类形成新节点
  - 在距离矩阵中删除这两个类相应的行和列,为新类加入新的行和列(非加权)
  3. 重复循环,直到仅剩一个类
- 思想跟连锁聚类方法、渐进法类似
  - 是一种改进了的邻近归并法

最大简约法(1)

- **Maximum Parsimony (MP)**
- **思想：最好的树应该用最少的进化上的变化来解释数据**
  - **基于离散特征的方法**
  - **枝长来自该枝进化上变化的数目**
  - **有时会存在多棵最大简约树**

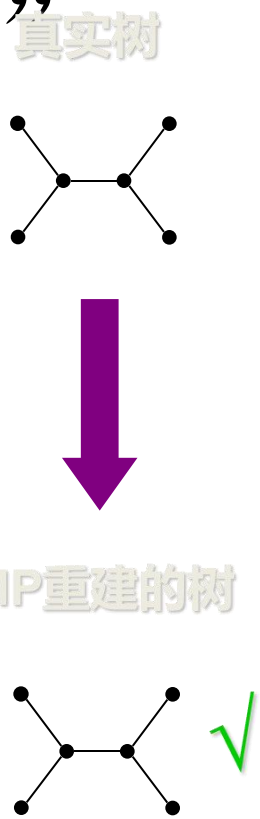
最大简约法(2)

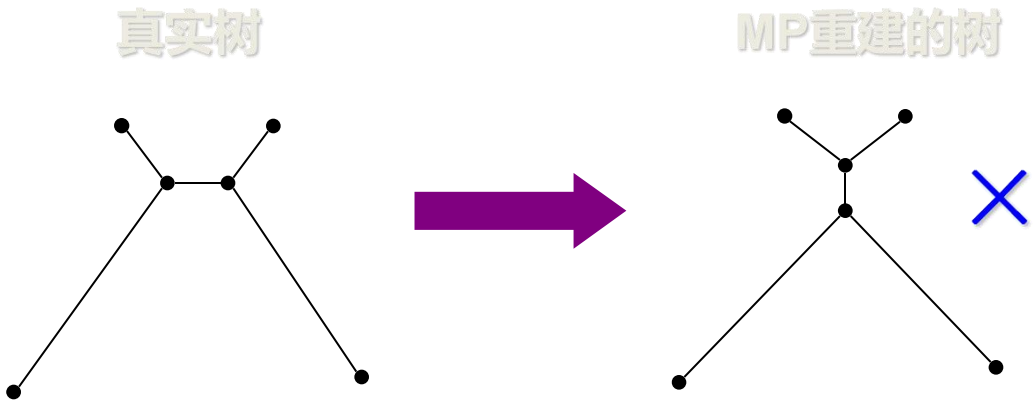
- 计算量太大 → 考虑部分位点
- *信息位点 (informative sites)*
  - 若在某个位点上至少有两个等位基因，而每个等位基因至少存在于两条序列，该位点称为信息位点

	位点								
序列	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

最大简约法(3) – “长枝吸引”

- Long Branch Attraction (LBA)
- 若两个物种的变异率较大，导致：
  1. 长枝
  2. 可能存在共同变异
  - 结果：若这些变异多于那些能区别它们共同祖先的变异，MP将产生错误的树





评价结果

- 问题
  - 整棵树和它的组成部分(分支)的置信度是多少?
  - 这样得到正确的树的可能性比随机选出一棵是正确的树的可能性大多少?
- 方法
  - 自举检验 (bootstrap)
  - 参数检验

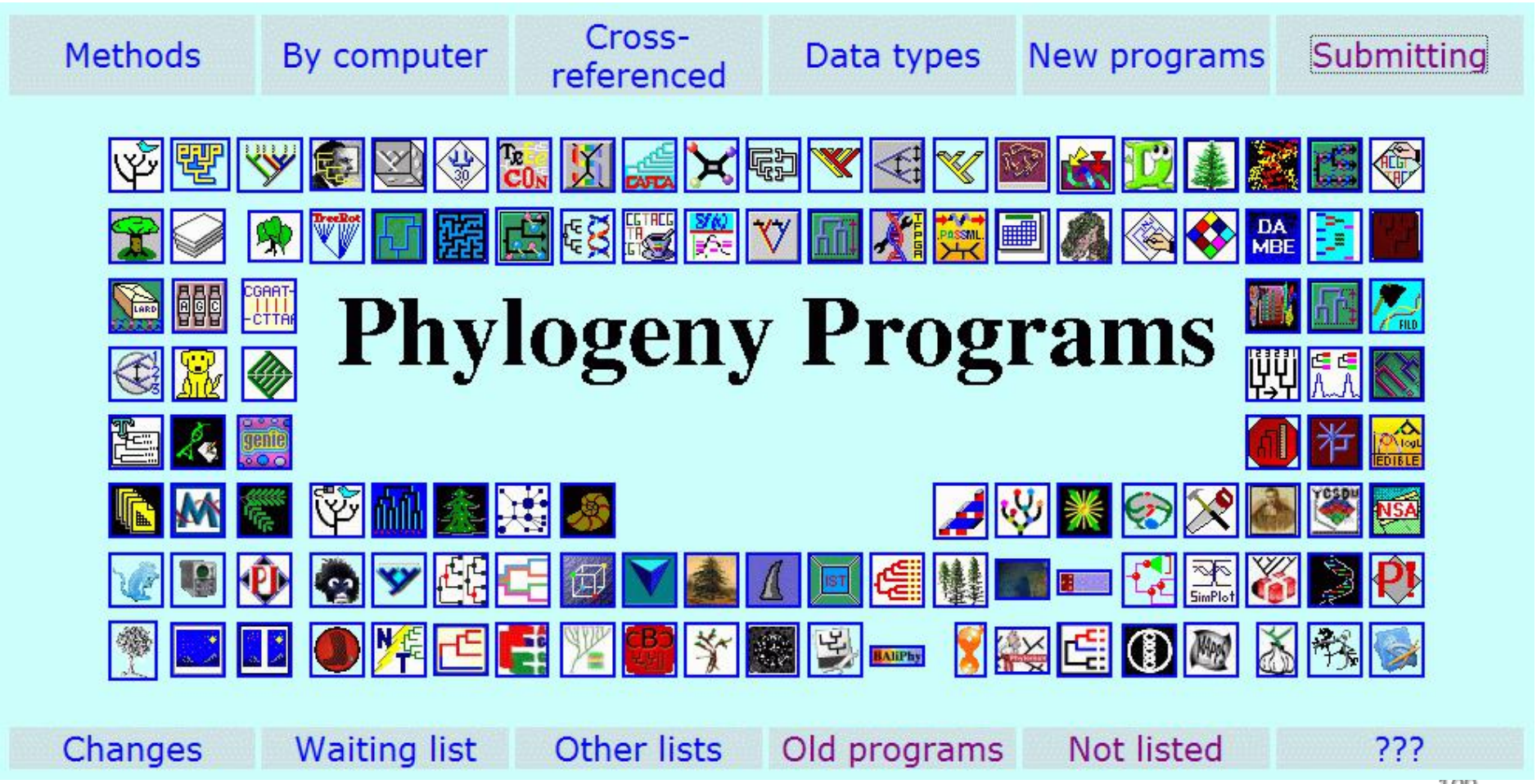
全基因组的系统发生分析

- 基于多棵系统发生树的方法
- 基于基因内容的方法
- 基于蛋白质折叠结构的方法
- 基于基因次序的方法
- 基于连接的直向同源蛋白的方法
- 基于代谢途径(pathway)的方法

四、系统发生软件

Joe Felsenstein's list of Phylogeny Programs (最全的列表)

- http://evolution.gs.washington.edu/phylip/software.html



JF's list (简介;包含309种软件)

- 三种分类依据
  - 软件所使用的方法
  - 软件使用的系统
  - 软件所分析的数据
- 其他列表
  - 最近加入的软件
  - 最近更新的软件
  - 以前列出但已经不再发行的软件
  - 等待加入的软件
  - 不被列出的软件
  - 其他系统发生软件的列表

JF's list (所有软件按方法分类)

1. General-purpose packages 一般目的 11
2. Parsimony programs 简约法 37
3. Distance matrix methods 距离矩阵 65
4. Computation of distances 计算距离 58
5. Maximum likelihood and Bayesian methods 最大似然、贝叶斯 77
6. Quartets methods 四重奏 11
7. Artificial-intelligence and genetic algorithms methods 人工智能、遗传算法 4
8. Invariants (or Evolutionary Parsimony) methods 不变量/进化简约 4
9. Interactive tree manipulation 24
10. Looking for hybridization or recombination events 19
11. Bootstrapping and other measures of support 63
12. Compatibility analysis 9
13. Consensus trees, subtrees, supertrees, distances between trees 22
14. Tree-based alignment 20
15. Gene duplication and genomic analysis 6
16. Biogeographic analysis and host-parasite comparison 8
17. Comparative method analysis 26
18. Simulation of trees or data 21
19. Examination of shapes of trees 13
20. Clocks, dating and stratigraphy 32
21. Model Selection 12
22. Description or prediction of data from trees 9
23. Tree plotting/drawing 38
24. Sequence management/job submission 20
25. Teaching about phylogenies 4

(方法后数字为该分类的软件个数)

JF's list (一般目的软件)

- PHYLIP
- PAUP*
- MEGA
- Phylo_win
- ARB
- DAMBE
- PAL
- Bionumerics
- Mesquite
- PaupUp
- BIRCH

JF's list (软件按数据分类)

- Microsatellite data
  1. RSTCALC
  2. POPTREE
  3. Microsat
  4. Populations
  5. MSA
  6. YCDMA
  7. Network
  8. IM

JF's list (按数据分类)

- RAPDs, RFLPs, or AFLPs
  1. tfpga
  2. RAPDistance
  3. Fingerprinting II Informatix Software
  4. GelCompar II
  5. Bionumerics
  6. Winboot
  7. REAP
  8. RESTSITE
  9. MVSP
  10. DENDRON
  11. Phyltools
  12. Network
  13. BIRCH

JF's list (按数据分类)

- Continuous quantitative characters
  1. PHYLIP
  2. Mesquite
  3. ANCML
  4. COMPARE
  5. CMAP
  6. PDAP
  7. ACAP
  8. Phylogenetic Independence
  9. APE
  10. CAIC
  11. TreeScan
  12. PHYLOGR
  13. IDC
  14. CoMET
  15. OUCH
  16. Brownie
  17. BayesTraits
  18. TNT
  19. PHYSIG

JF's list (按数据分类)

- Gene frequencies (aside from microsatellite loci)
  1. PHYLIP
  2. DAMBE
  3. DISPAN
  4. GDA
  5. POPGENE
  6. YCDMA
  7. FSTAT
  8. Arlequin
  9. DnaSP
  10. APE
  11. DIVAGE
  12. GeneStrut
  13. POPTREE
  14. Genepop
  15. SPAGeDi

免费开源软件

- http://digitaltaxonomy.infobio.net/ (Digital Taxonomy)
- 还包括systematics(分类学/系统学)、morphometrics(形态测定学)方面的软件

维基(Wiki)的列表

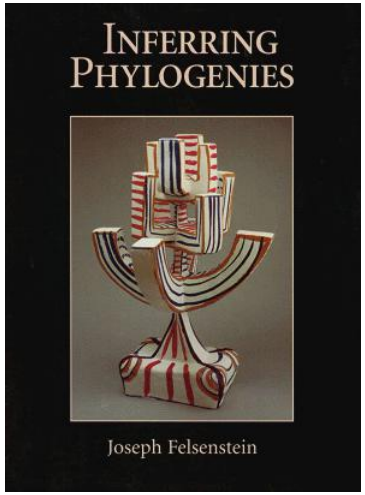
NAME	Description	Methods	Link	Author
PHYLIP	Phylogenetic inference package	Maximum parsimony, distance matrix, maximum likelihood	download	J. Felsenstein
PAUP	Phylogenetic analysis using parsimony	Maximum parsimony, distance matrix, maximum likelihood	purchase	D. Swofford
PAML	Phylogenetic analysis by maximum likelihood	Maximum likelihood	download	Z. Yang
ClustalW	Progressive multiple sequence alignment	Distance matrix/nearest neighbor	EBI PBIL EMBNet GenomeNet	Thompson et al.
QuickTree	Tree construction optimized for efficiency	Neighbor-joining	server	K. Howe, A. Bateman, R. Durbin
MOLPHY	Molecular phylogenetics (protein or nucleotide)	Maximum likelihood	server	J. Adachi and M. Hasegawa
TreeGen	Tree construction given precomputed distance data	Distance matrix	server	ETH Zurich
fastDNAml	Optimized maximum likelihood (nucleotides only)	Maximum likelihood	download	G.J. Olsen
TREE-PUZZLE	Maximum likelihood and statistical analysis	Maximum likelihood	download	H.A. Schmidt, K. Strimmer, A. von Haeseler
TreeAlign	Efficient hybrid method	Distance matrix and approximate parsimony	server	J. Hein
PhyloQuart	Quartet implementation (uses sequences or distances)	Quartet method	download	V. Berry
MrBayes	Posterior probability estimation	Bayesian inference	download	J. Huelsenbeck et al.

杂项

- PhyloCode (不是软件;是种系发生命名法的一些规则)
- TOPD/FMTS (Bioinformatics 23(12); 2007-6-1)

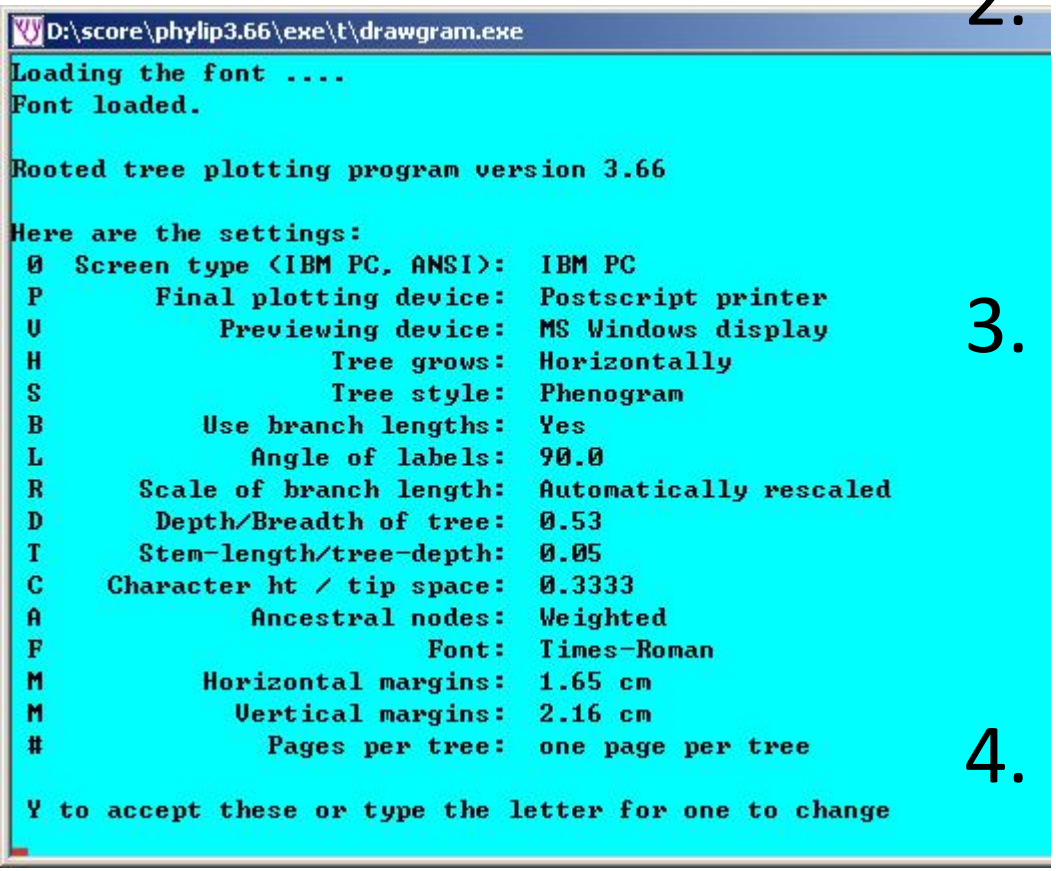
PHYLIP

- 种系发生软件包
- Joseph Felsenstein,华盛顿大学
- 下载(Windows版)
  - http://evolution.genetics.washington.edu/phylip/getme.html



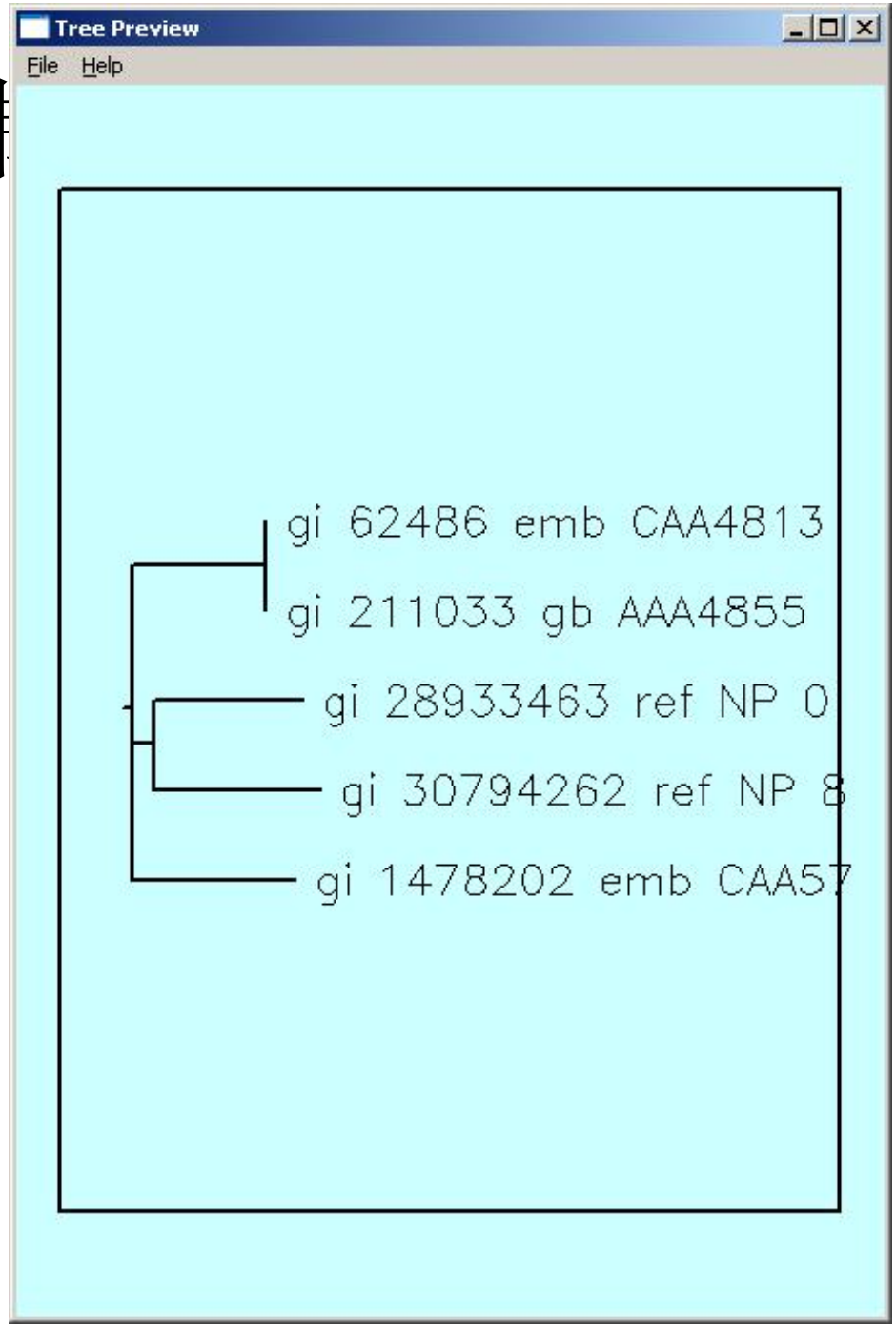
用PHYLIP绘制树(1)

1. 安装目录里有一个exe目录
2. 将该目录中某一个font文件改名为fontfile
3. 将ClustalW中保存的引导树文件移入该目录并改名为intree
4. 运行程序drawgram.exe



绘制

5. 输入 “y”,回车


Tree Preview
File Help
gi 62486 emb CAA4813
gi 211033 gb AAA4855
gi 28933463 ref NP 0
gi 30794262 ref NP 8
gi 1478202 emb CAA57