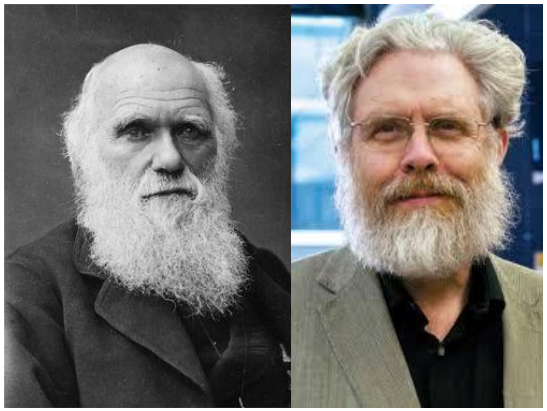
生物统计学：
生物信息中的概率统计模型

2019年秋



有关信息

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

2019生物统计学

扫一扫二维码，加入该群。

- 课程网页
  - http://www.microbioinformatics.org/teach/#
  - QQ群：882140516

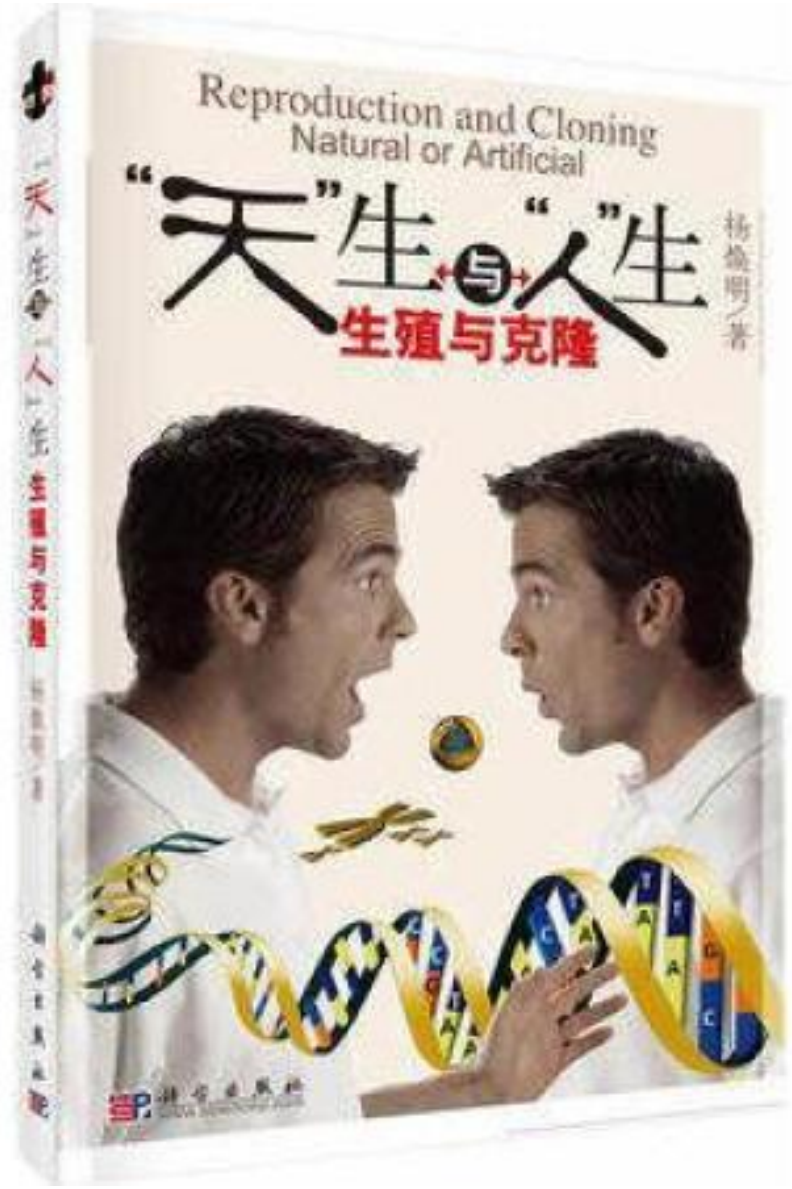
课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III)
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 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
...

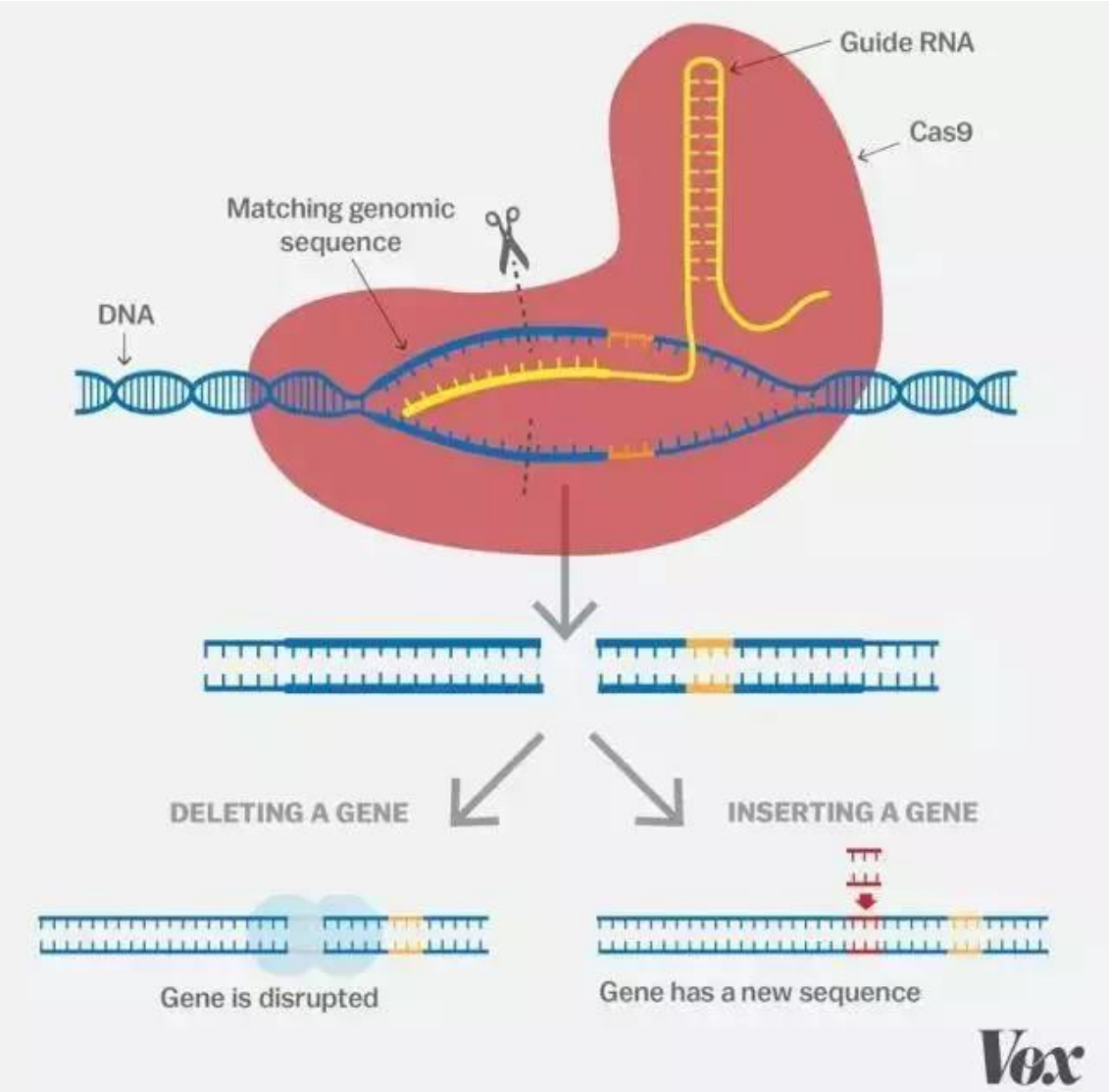
方法:
生物计算与生物统计

读物推荐

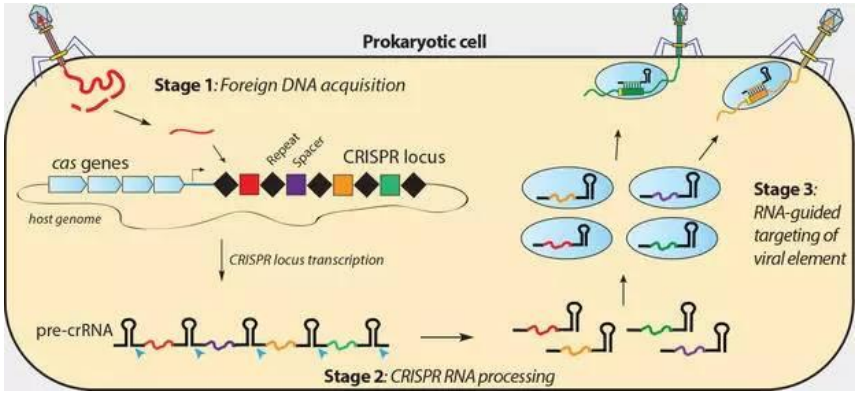

Reproduction and Cloning
Natural or Artificial
“天”生与“人”生
生殖与克隆
杨焕明 著

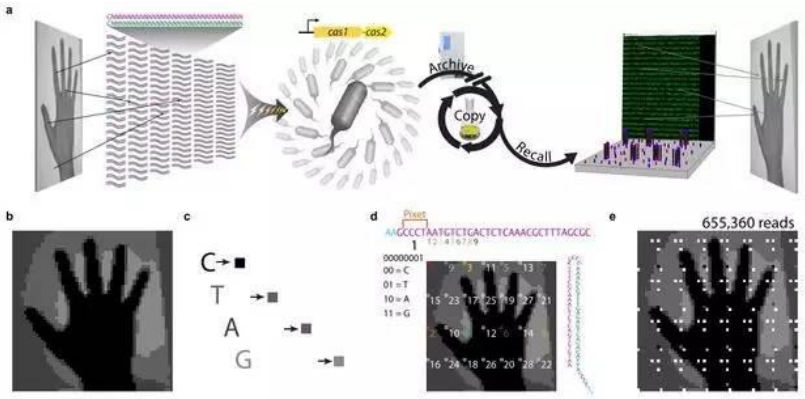

How Synthetic Biology Will Reinvent Nature and Ourselves
Regenesis
[美] 乔治·丘奇（George Church） [美] 艾德·里吉西（Ed Regis） 著
再创世纪
合成生物学将如何重新
创造自然和我们人类
周东 译

CRISPR和基因编辑技术



Understand it, create it!



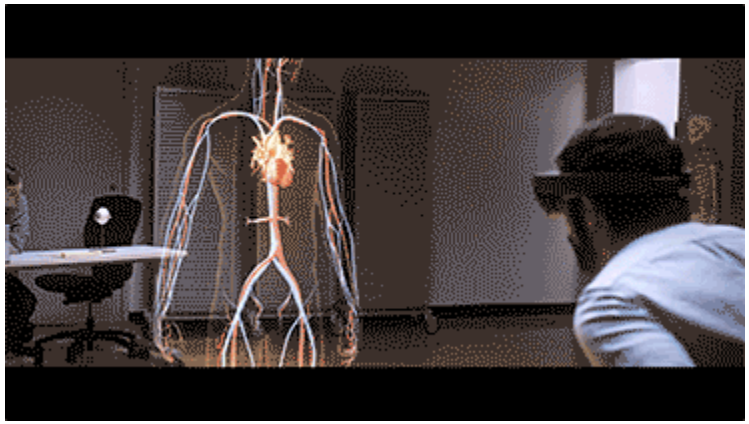


Original Image

原始图像

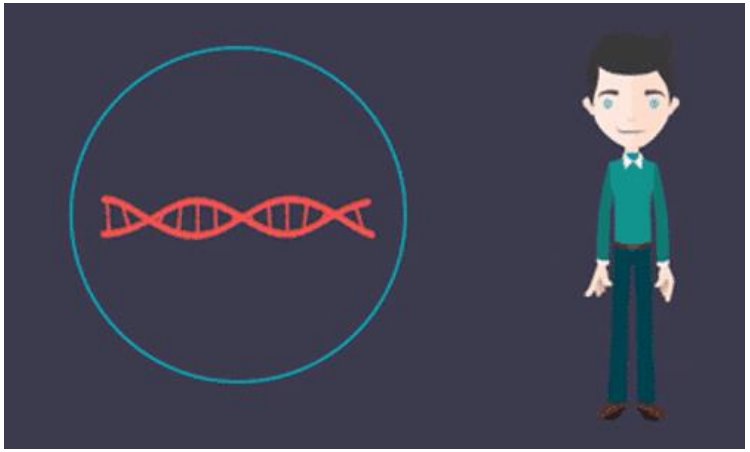
Image Reconstructed From Bacteria

从细菌DNA还原的图像



See it!

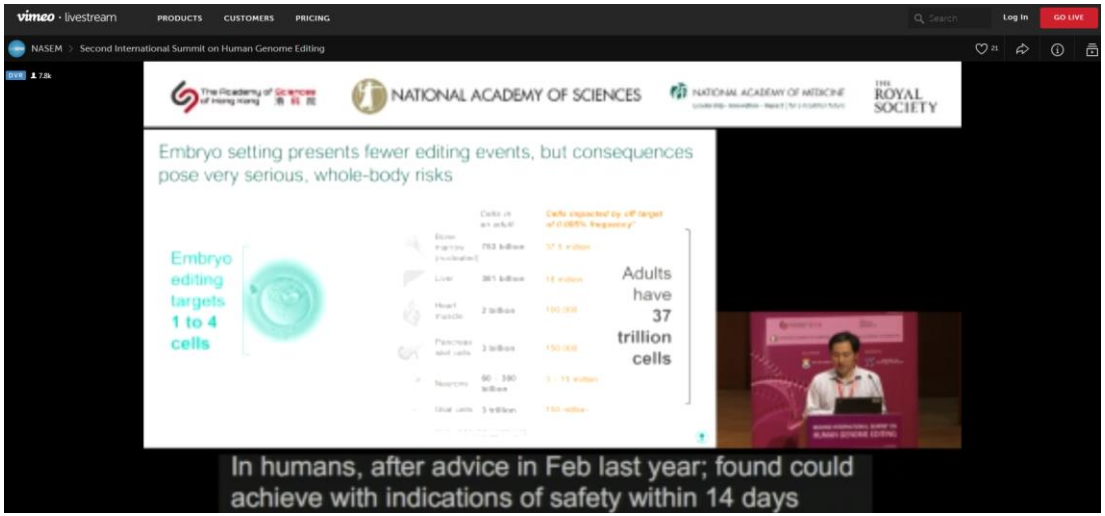
Understand it!

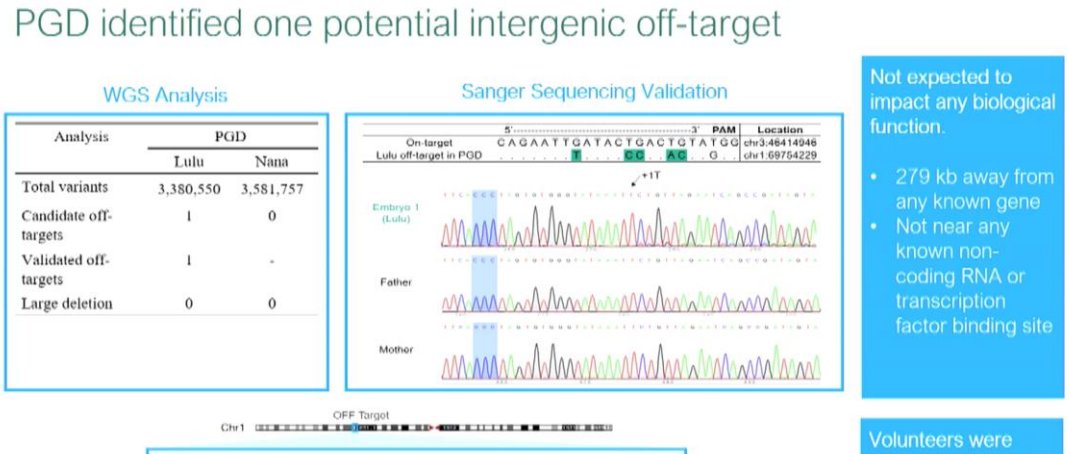


Create it!

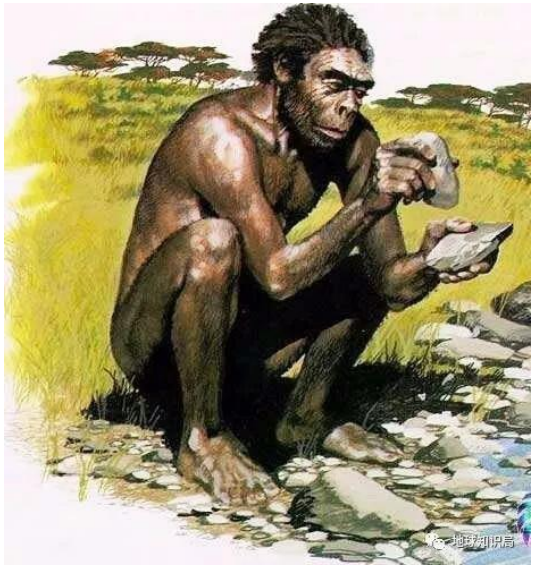
CRISPR和基因编辑技术

要清楚什么可以做，什么不可以做！



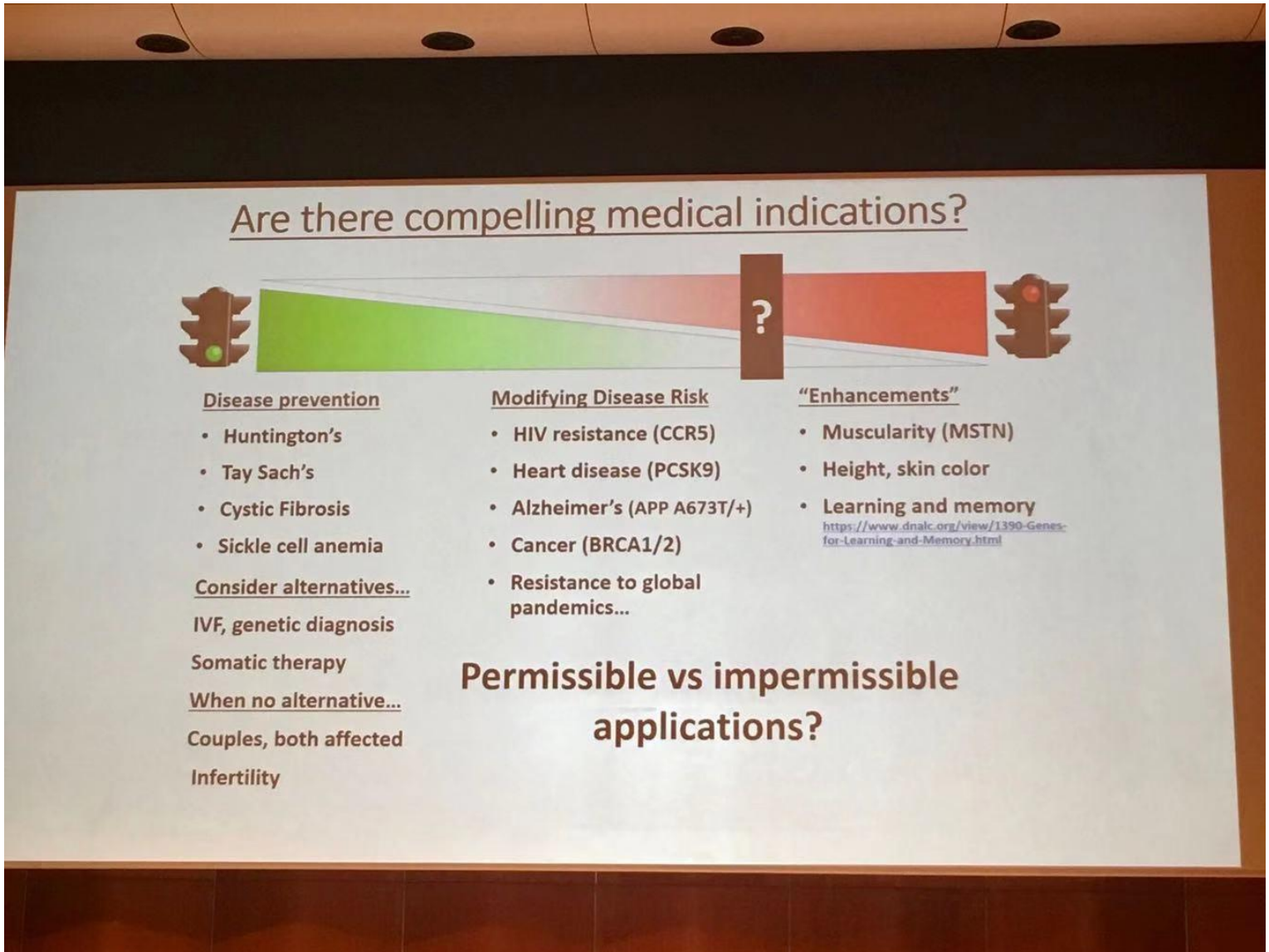


CRISPR和基因编辑技术



OR

CRISPR和基因编辑技术

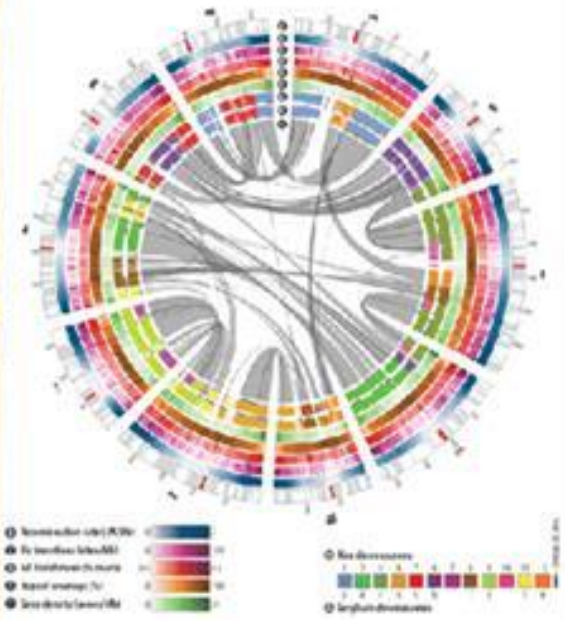


第4章：进化树构建的概率方法

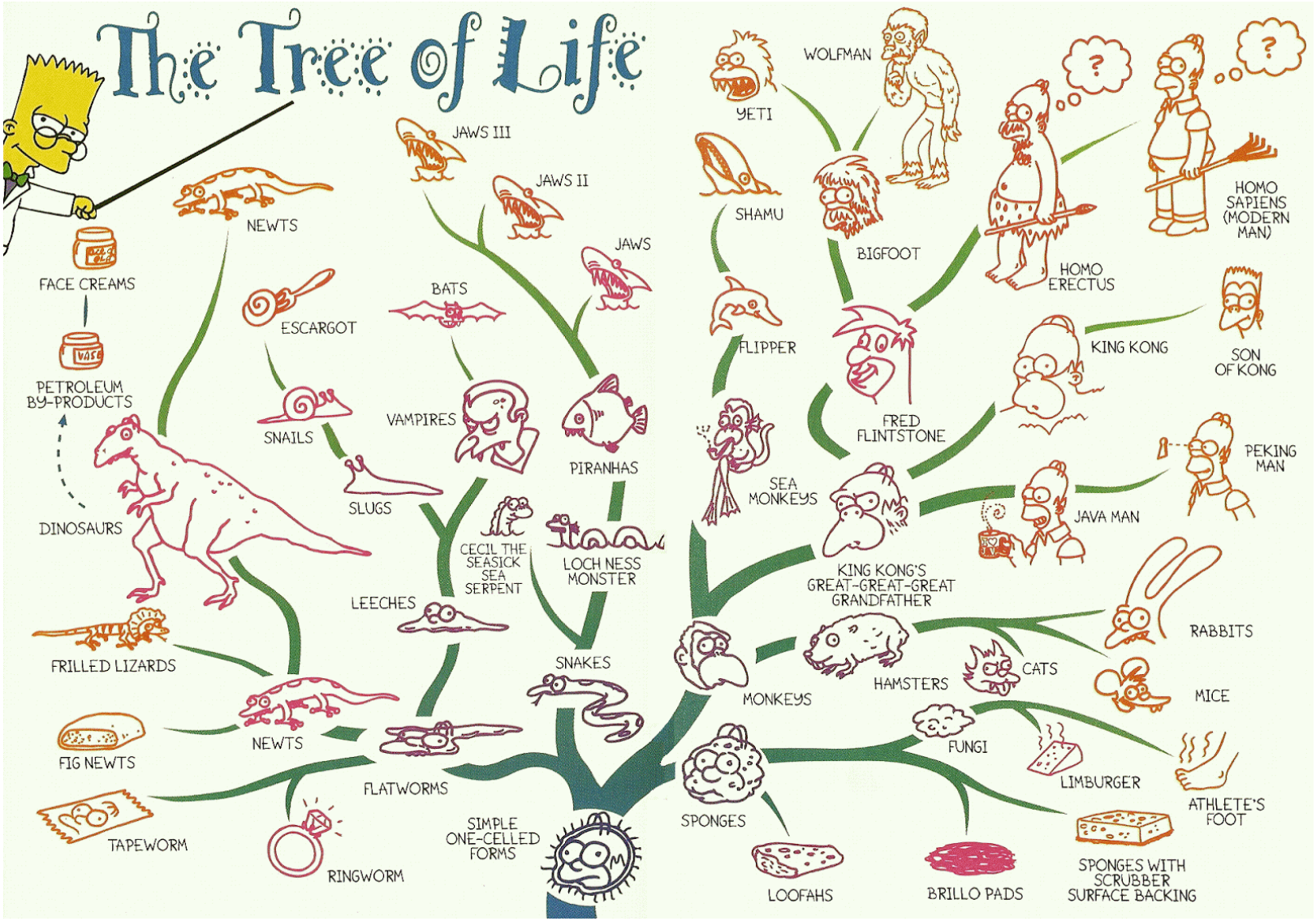
- 问题介绍
- 进化树构建方法的概率方法

Phylogenetic Tree

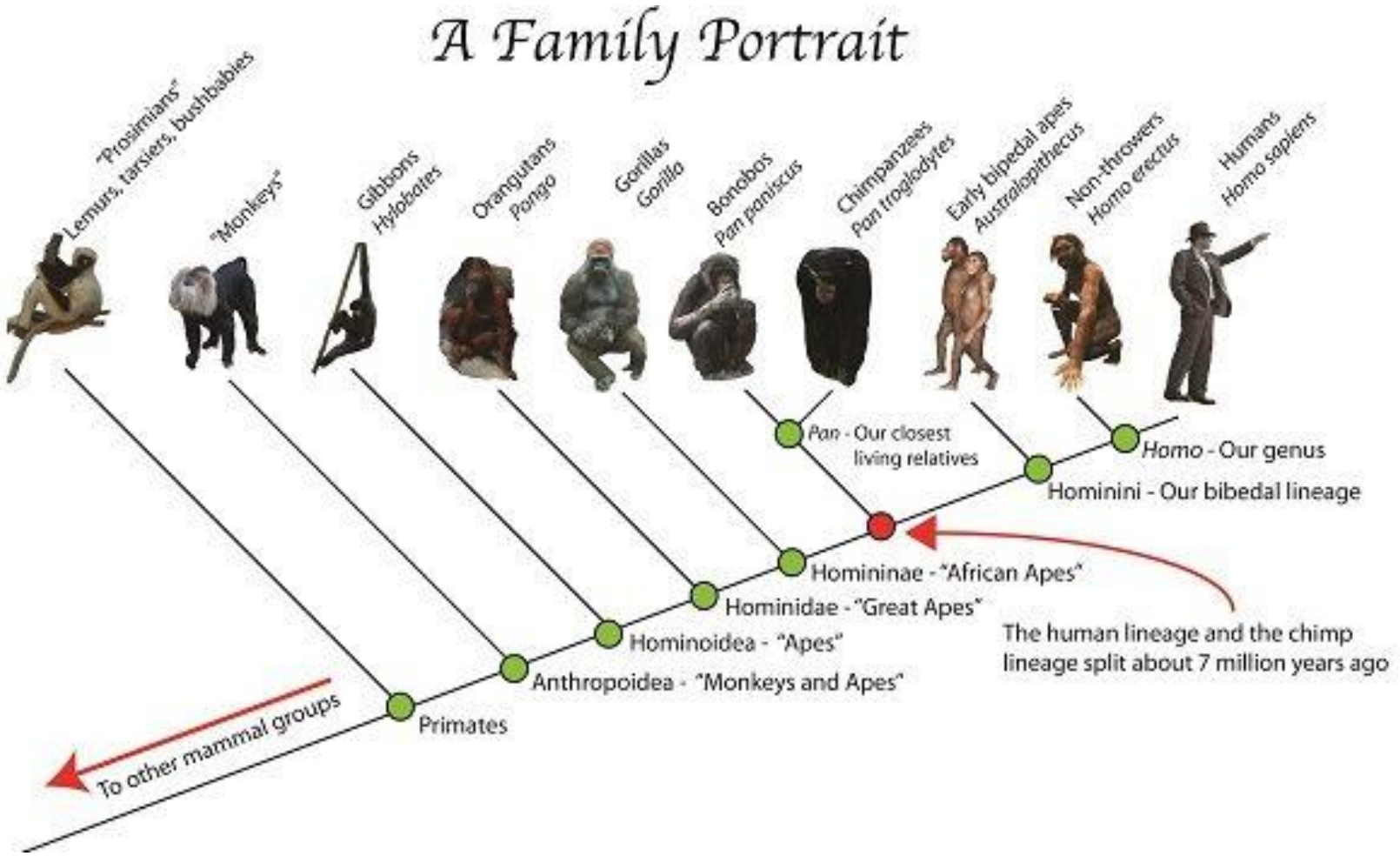
Everyone what to know how this tree look like...



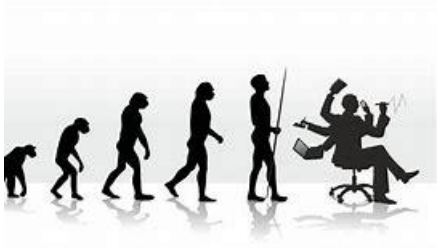
Phylogenetic Tree

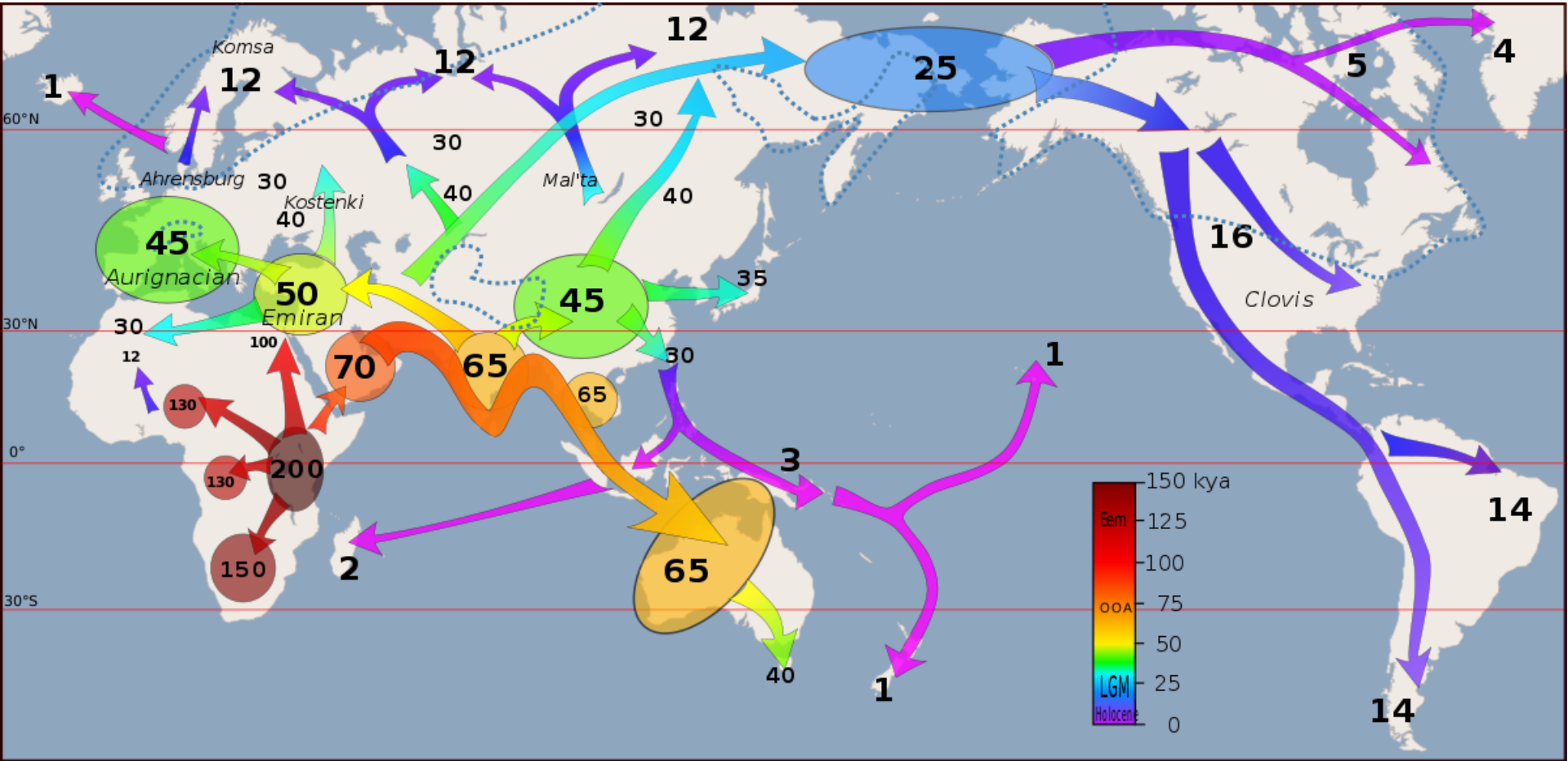


Evolution everywhere



Evolution everywhere

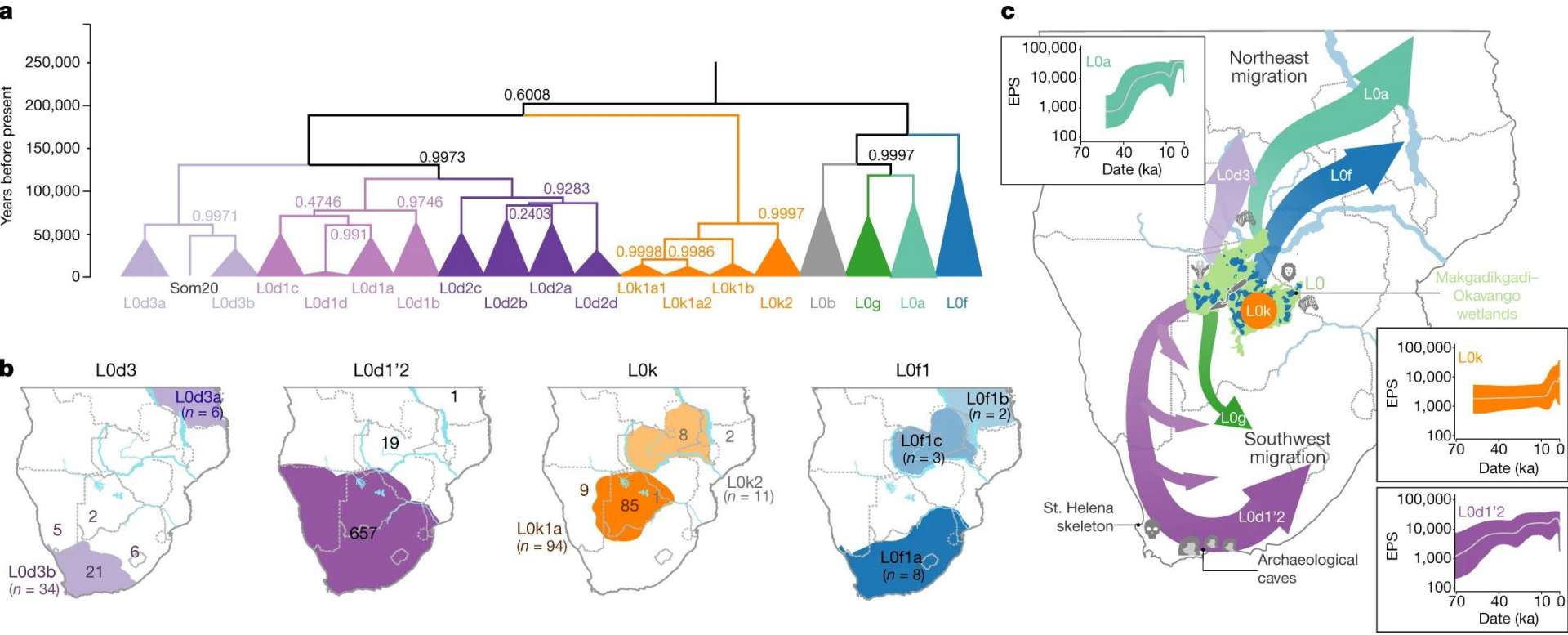




Evolution everywhere

Evolution everywhere

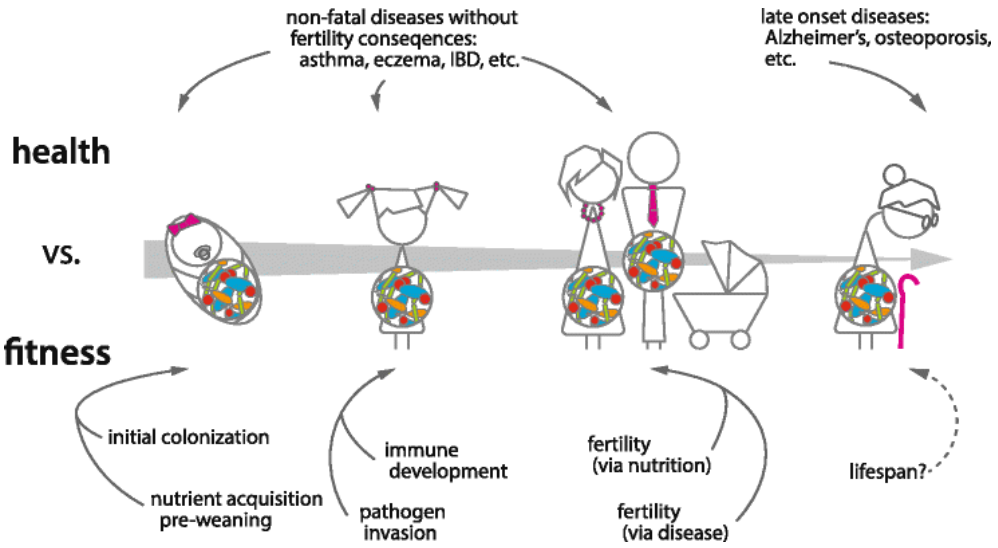
Out of “homeland”?

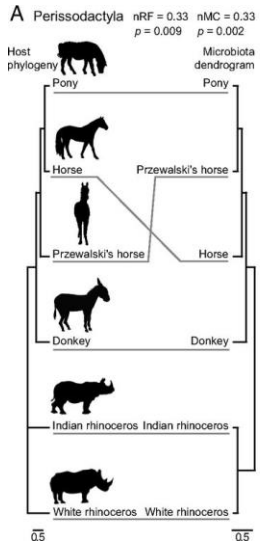


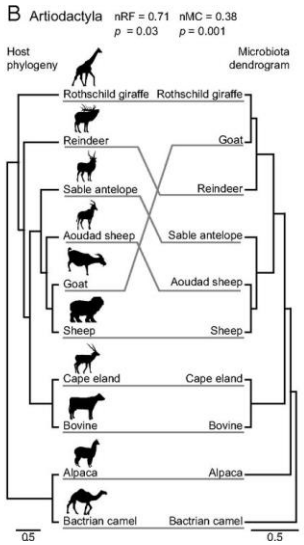
Eva K. F. Chan et al., *Nature*, 2019

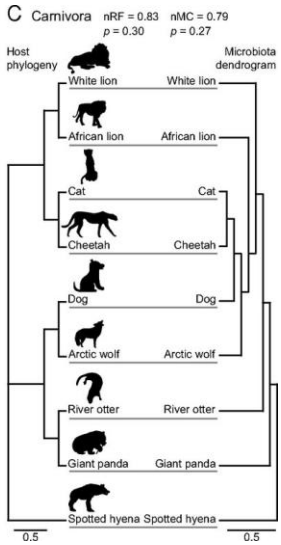
Evolution everywhere

Microbiome evolution also exists

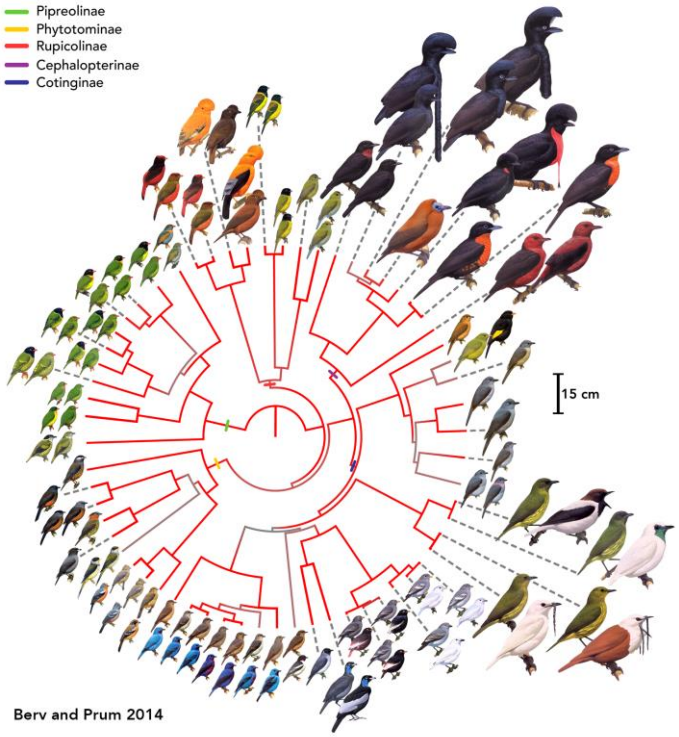


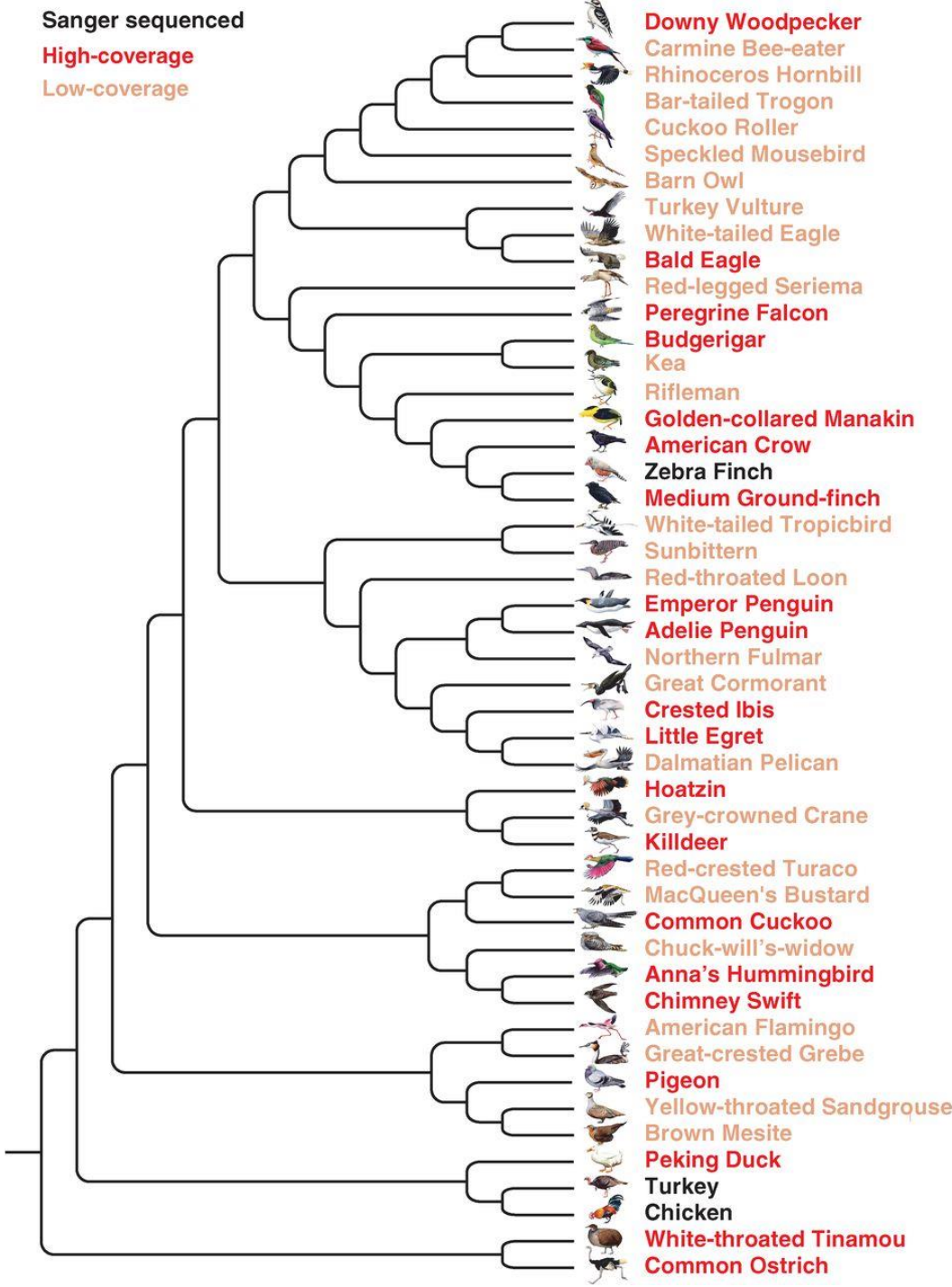






Phylogenetic Tree



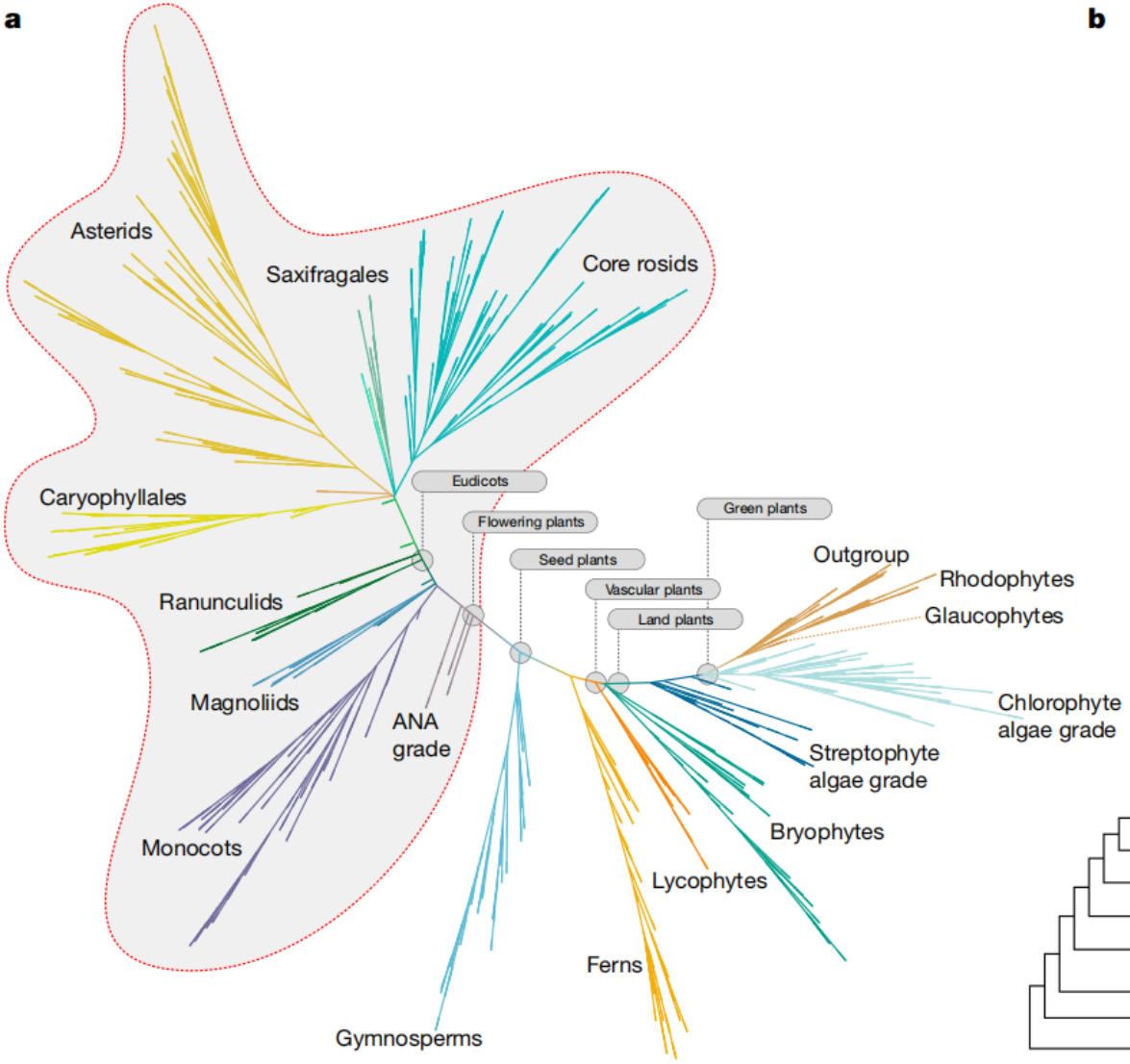


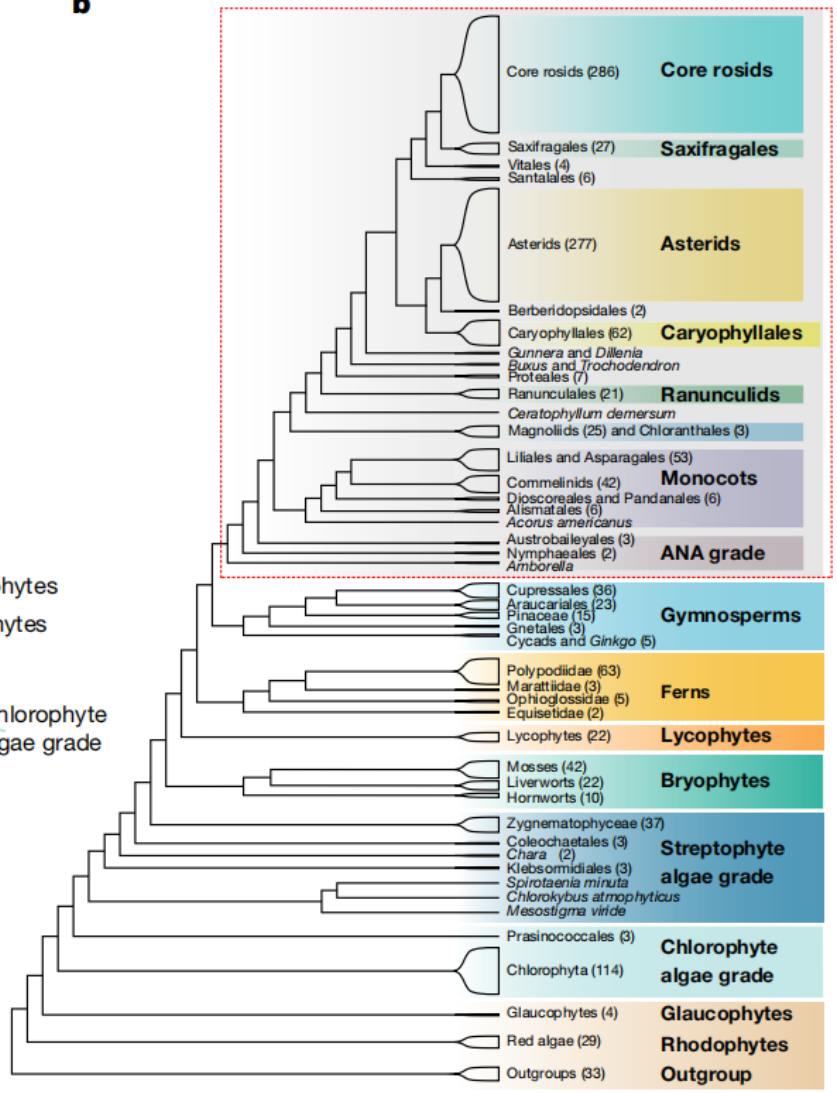
Reference:
https://b10k.genomics.cn/

Phylogenetic Tree

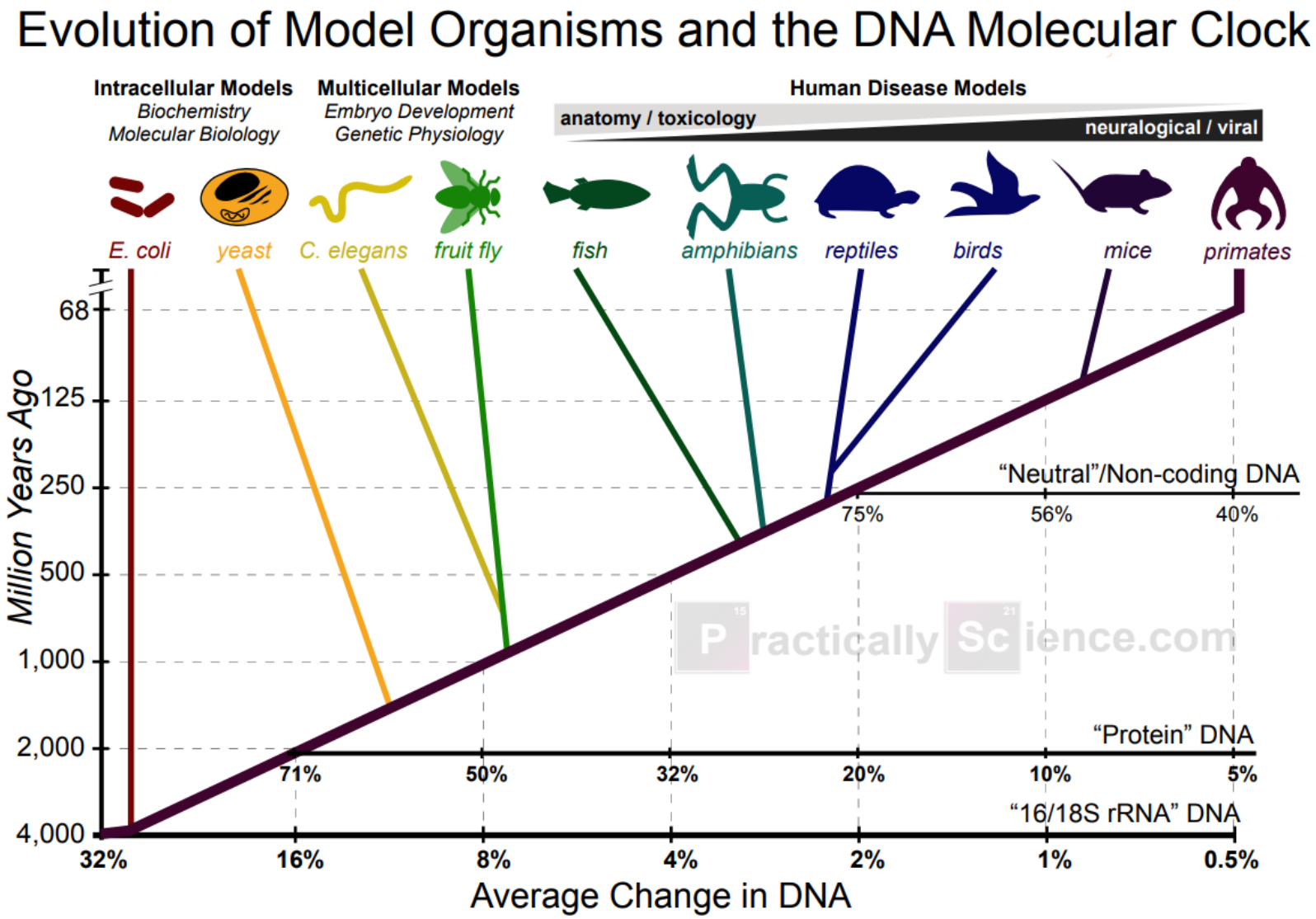
Reference:

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

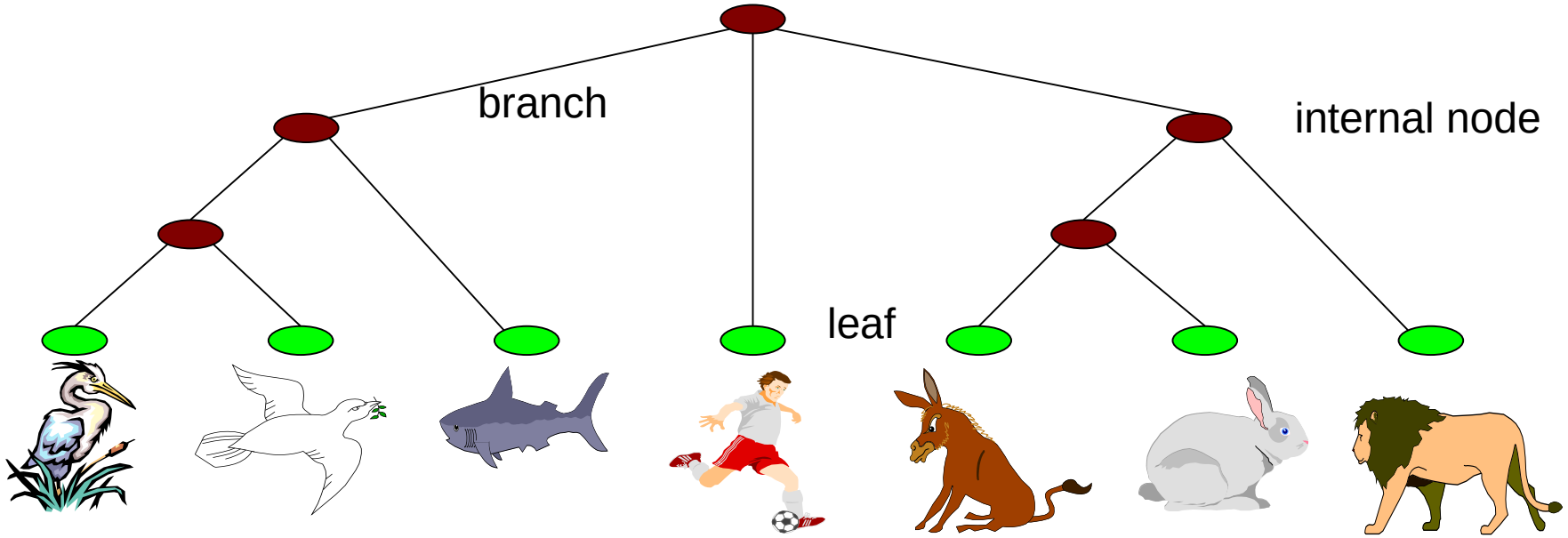




Phylogenetic Tree



Phylogenetic Tree



- 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。

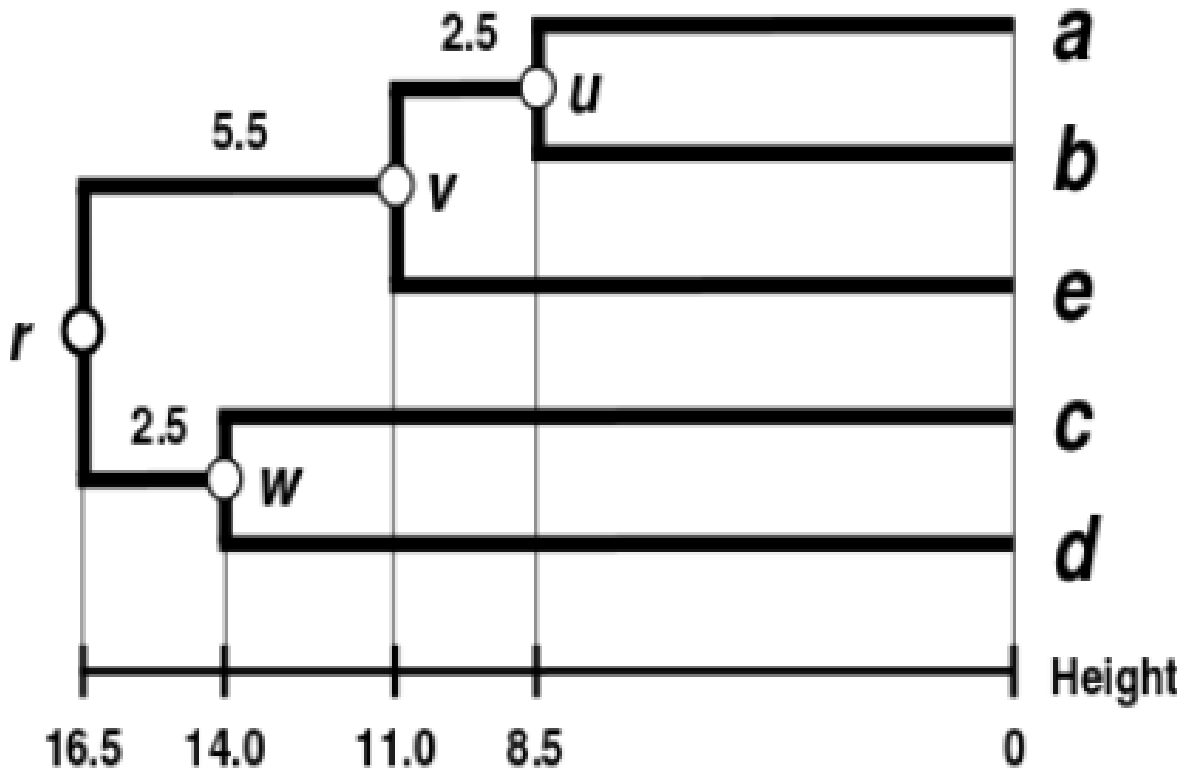
④若p和q是最后一个类群,侧聚类过程完成,否侧合并p和q成一个新类群r。

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)

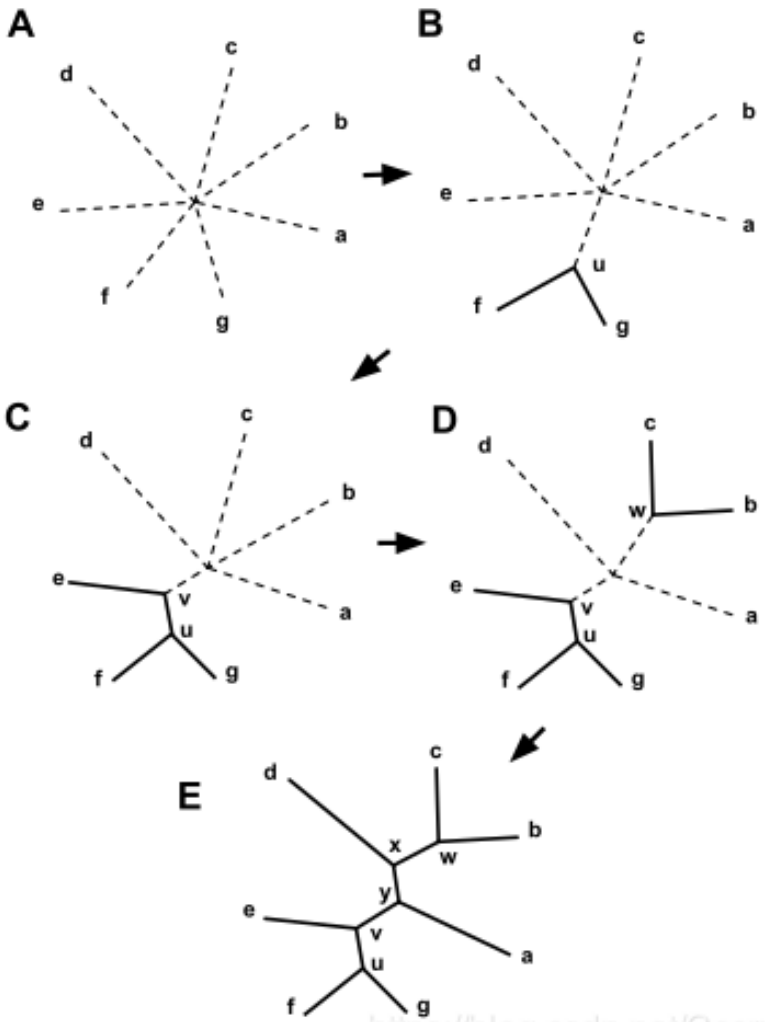
② 建立一个速率校正距离矩阵M,其元素由下式确定:

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

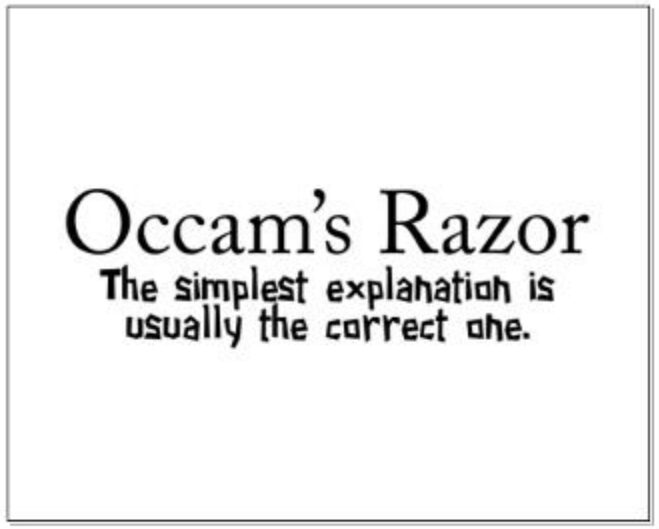
最大简约法（Maximum Parsimony）

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

The Role of Ockham's Razor: To Screen for Plurality at the Beginning of the Scientific Method

1. OBSERVATION
2. NECESSITY
3. INTELLIGENCE/AGGREGATION OF DATA (The Three Key Questions)
4. CONSTRUCT FORMULATION
5. SPONSORSHIP/PEER INPUT (Ockham's Razor)

Ockham's Razor

False SSkeptics seek to block this step at all costs.

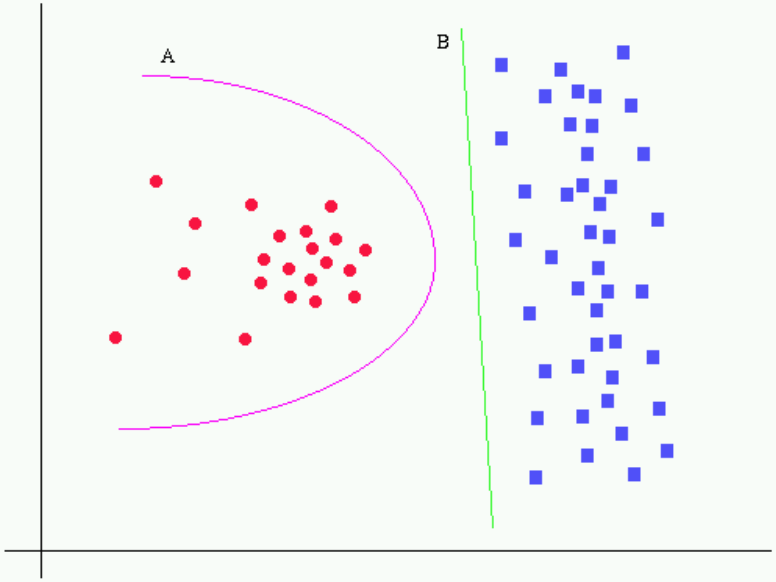
Plurality

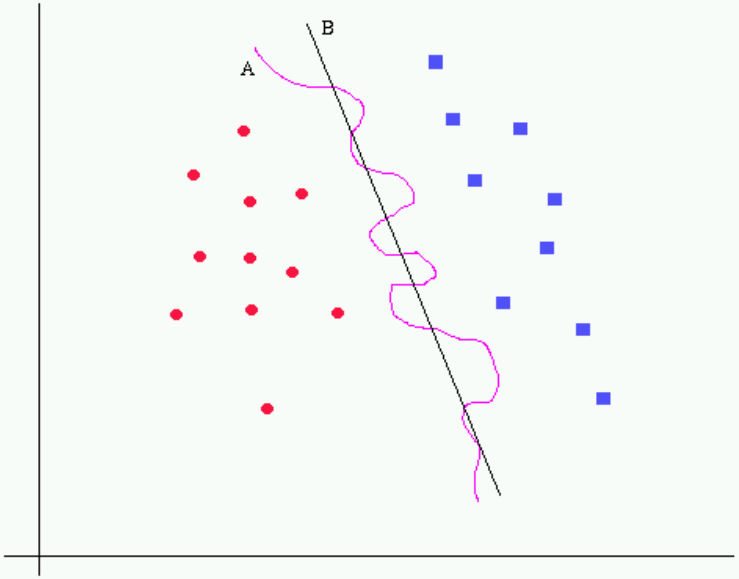
6. HYPOTHESIS DEVELOPMENT
7. PREDICTIVE TESTING
8. COMPETITIVE HYPOTHESES FRAMING (ASKING THE RIGHT QUESTION)
9. FALSIFICATION TESTING
10. HYPOTHESIS MODIFICATION
11. FALSIFICATION TESTING/REPEATABILITY
12. THEORY FORMULATION/REFINEMENT
13. PEER REVIEW (Community Vetting)
14. PUBLISH

Proof

15. ACCEPTANCE

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。

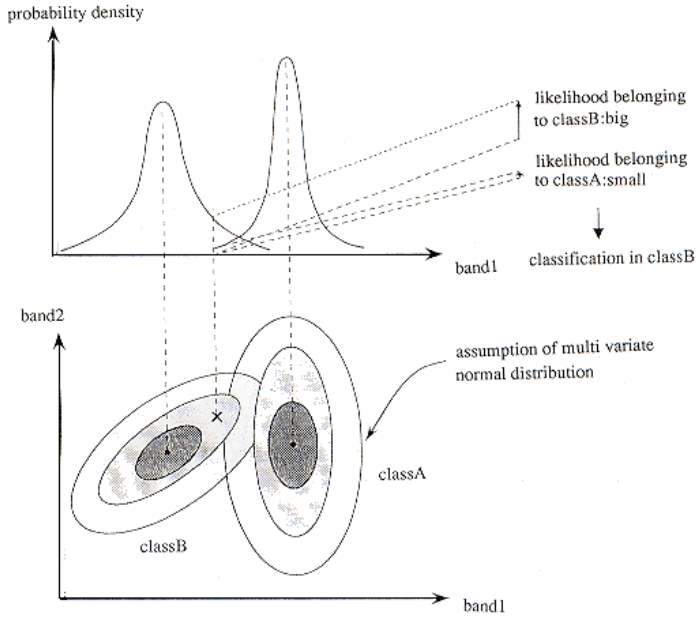
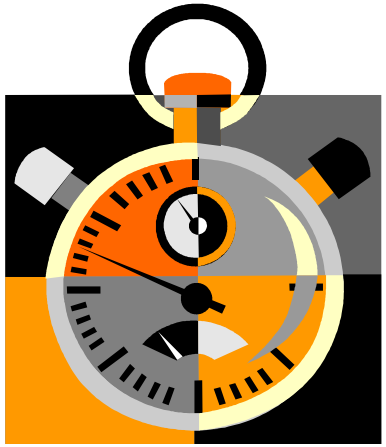


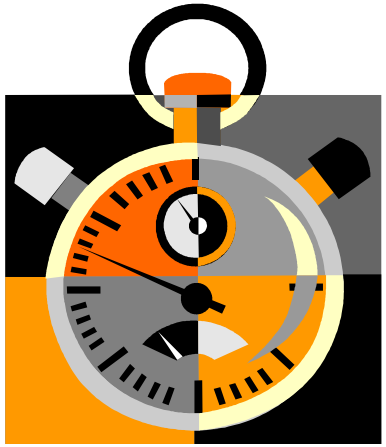
Figure 11.7.1 Concept of Maximum Likelihood Method

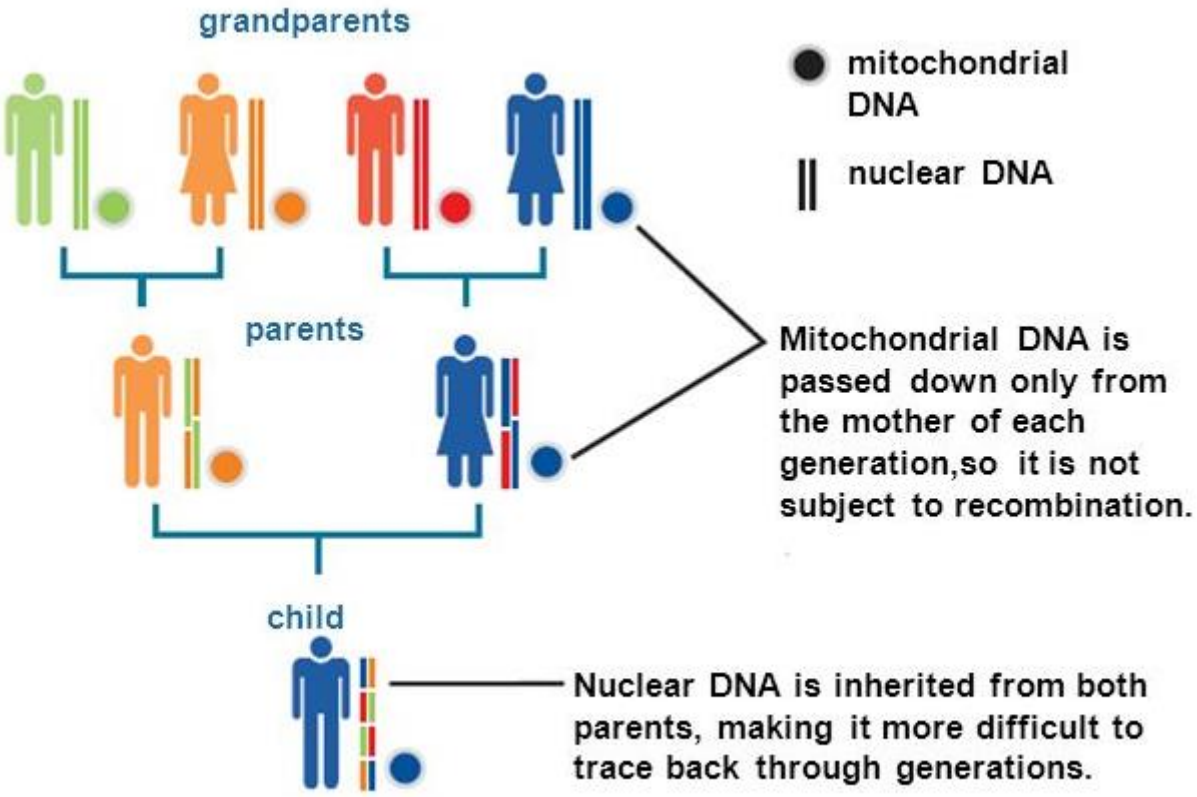
Molecular Clock Hypothesis



- 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

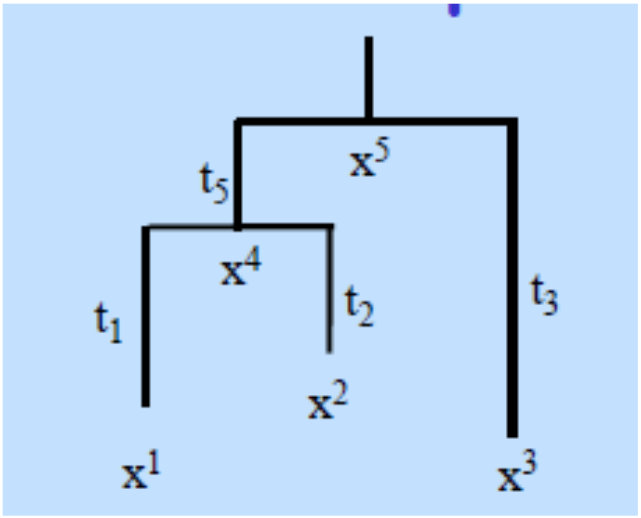
- The phylogenetic tree represents a generative probabilistic model (like HMMs) for the observed sequences.
- Background probabilities: q(a)
- Mutation probabilities: P(a|b, t)
- Models for evolutionary mutations
  - Jukes Cantor
  - Kimura 2-parameter model
- Such models are used to derive the probabilities

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:

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

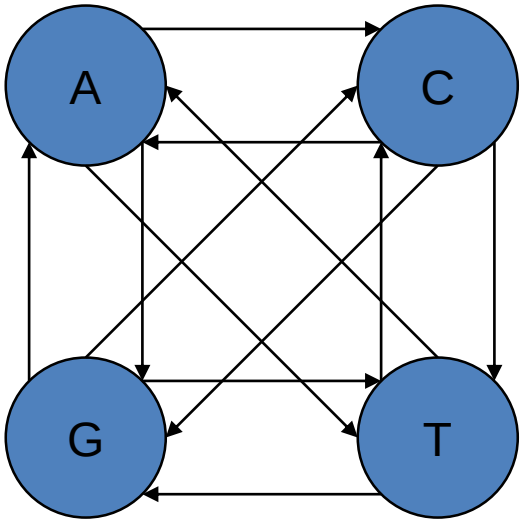
- A substitution rate matrix Q implies a probability distribution over {A,C,G,T} at each position, including stationary (equilibrium) frequencies π_A, π_C, π_G, π_T
- Each Q is an evolutionary model (some work better than others)

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

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

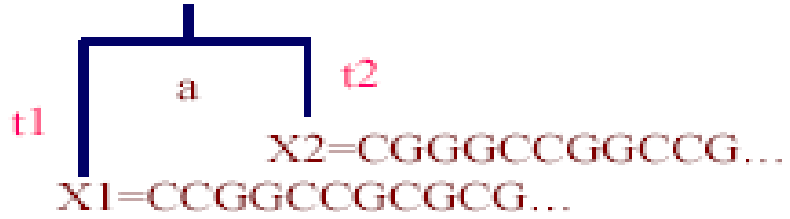
- Such that the expected number of substitution is 1%, i.e.

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

- 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 & $q_C = q_G = q_A = q_T = \frac{1}{4}$:

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

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

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

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

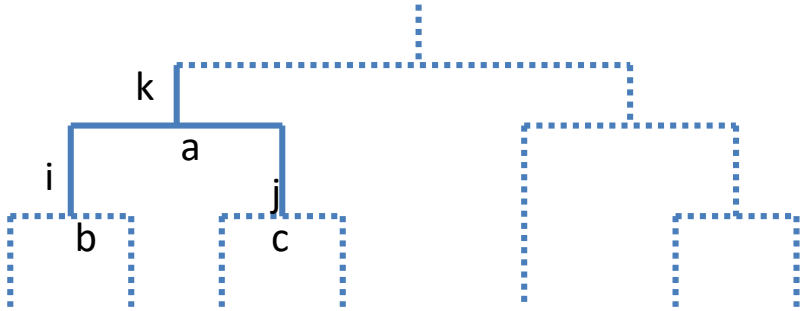
$$P(x^1, \cdots, x^n | T, t) = \prod_{u=1}^{N} P(x_u^1, \cdots, x_u^n | T, t)$$

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

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

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

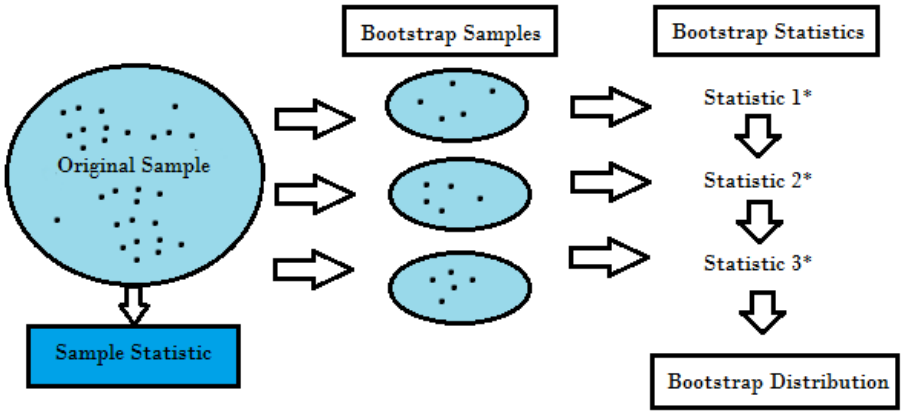
- 如果序列的相似性较高，各种方法都会得到不错的结果，模型间的差别也不大.
- 若有合适的分子进化模型可供选择，用最大似然法构树获得的结果较好.
- 对于近缘物种序列，通常情况下使用最大简约法.
- 而对于远缘物种序列，一般使用邻接法或最大似然法.
- 对于相似度很低的序列，邻接法往往出现长枝吸引(branch attraction)现象，有时严重干扰进化树的构建.

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



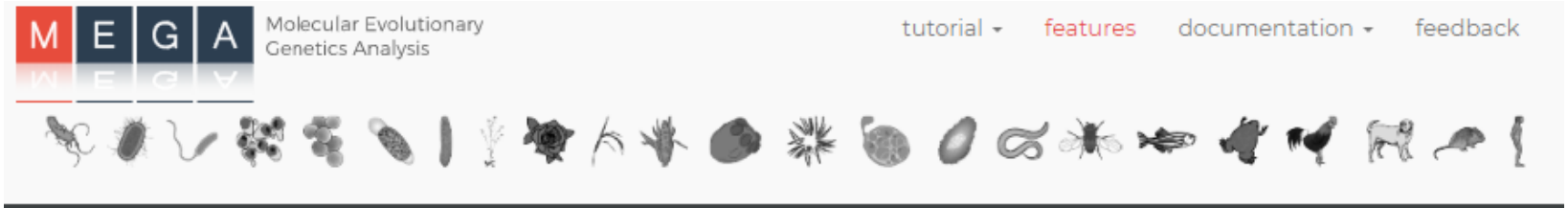
在重建进化树过程中，均需选择bootstrap进行树的检验：

- 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

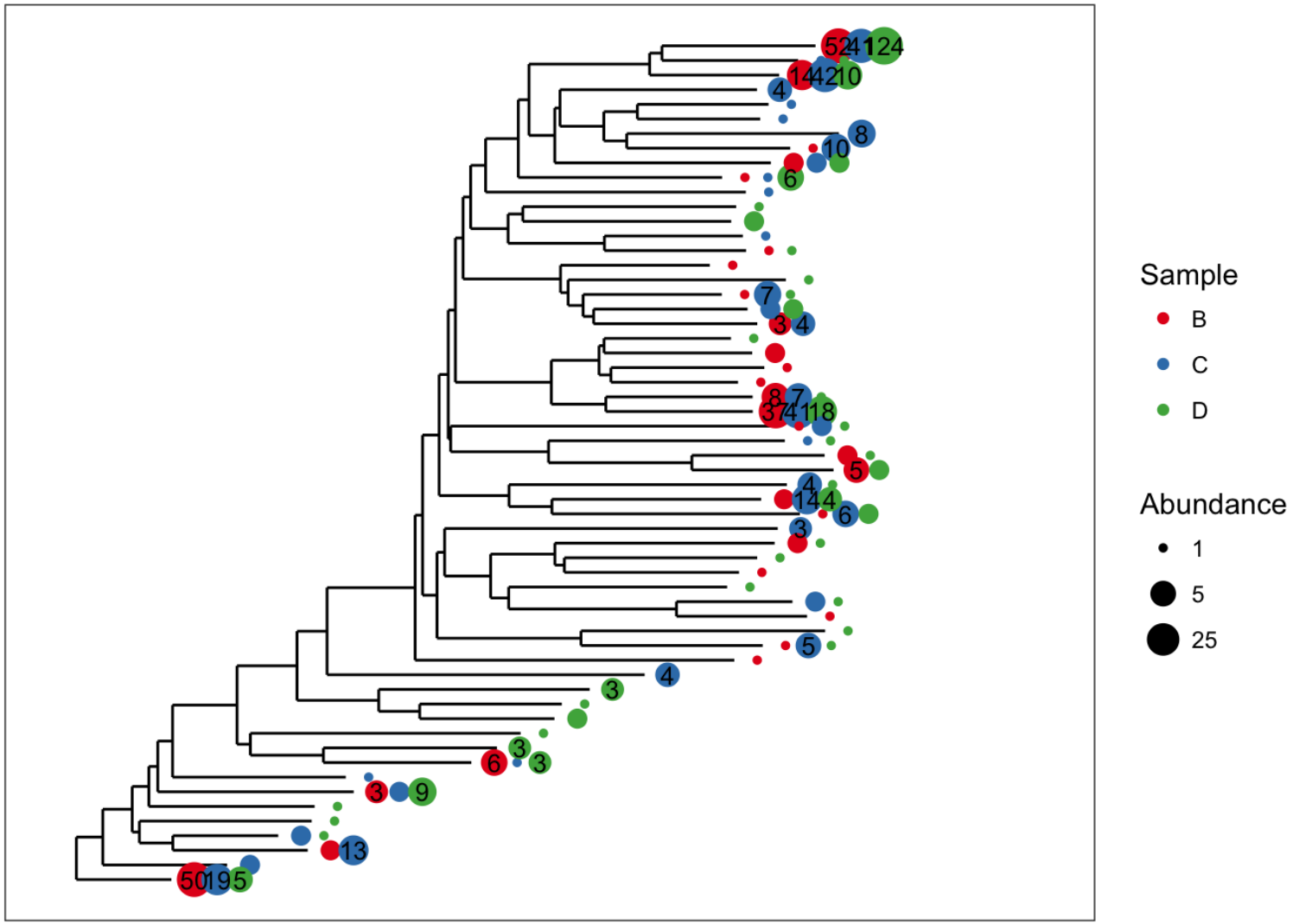
软件	网址	说明
ClustalX	http://bips.u-strasbg.fr/fr/Documentation/ClustalX/	图形化的多序列比对工具
ClustalW	**http://www.cf.ac.uk/biosi/resear ... loads/clustalw.html**	**命令行格式的多序列比对工具**
GeneDoc	http://www.psc.edu/biomed/genedoc/	多序列比对结果的美化工具
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	序列分析的综合工具
MEGA	**http://www.megasoftware.net/**	**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

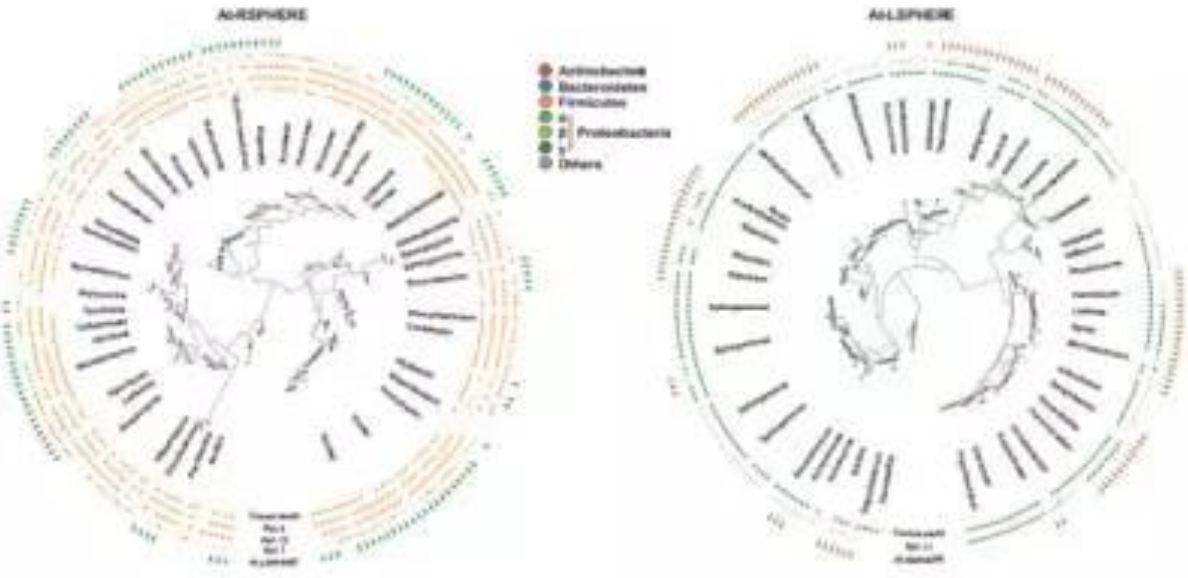


	Likelihood	Distance	Parsimony	Bayesian	Visual Explorer	Caption Expert
Phylogeny	✓	✓	✓		✓	✓
Bootstrap	✓	✓	✓		✓	✓
Distance/Diversity	✓	✓			✓	✓
Model Selection	✓					
Substitution Pattern	✓				✓ XL	✓
Rate Variation	✓				✓ XL	✓
Ancestral Sequence			✓	✓	✓	✓
Clock Test	✓	✓			✓ XL	✓
Time Tree	✓	✓			✓	✓
Selection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

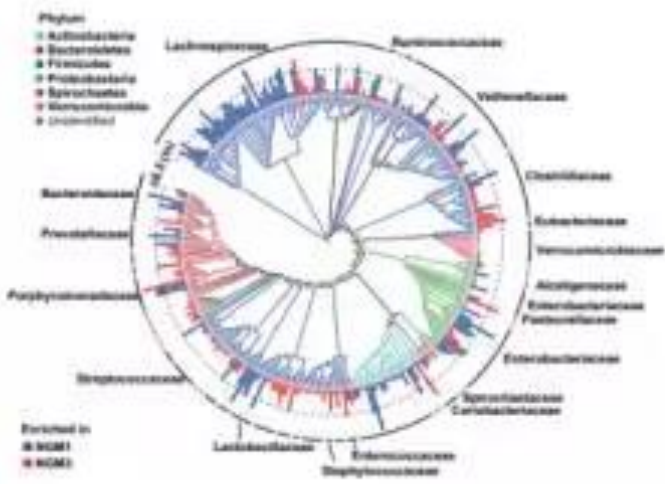
ggplot：物种进化关系分析



iTOL：物种进化关系分析



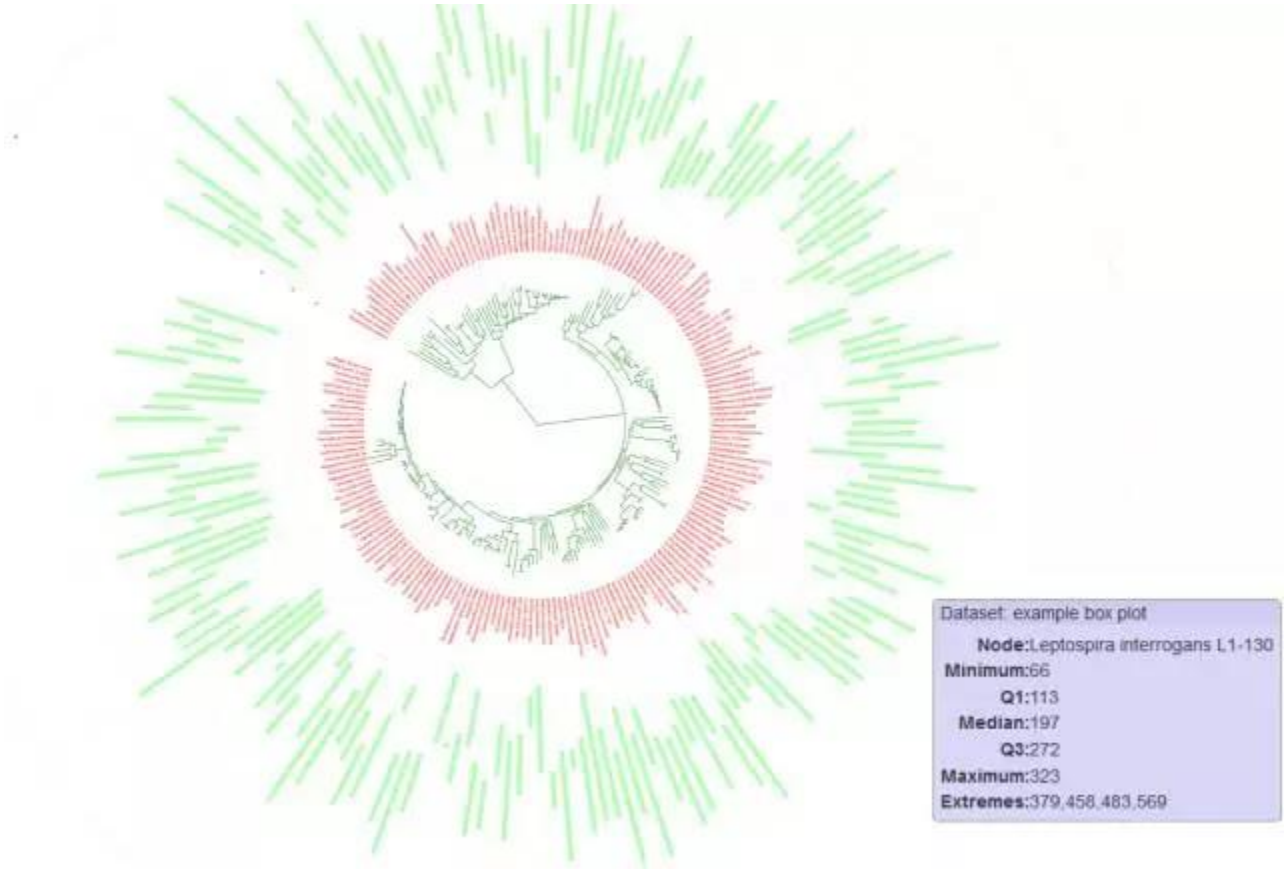
Nature 528, 364–369 (2015)



Nature Medicine 22, 1187–1191 (2016)

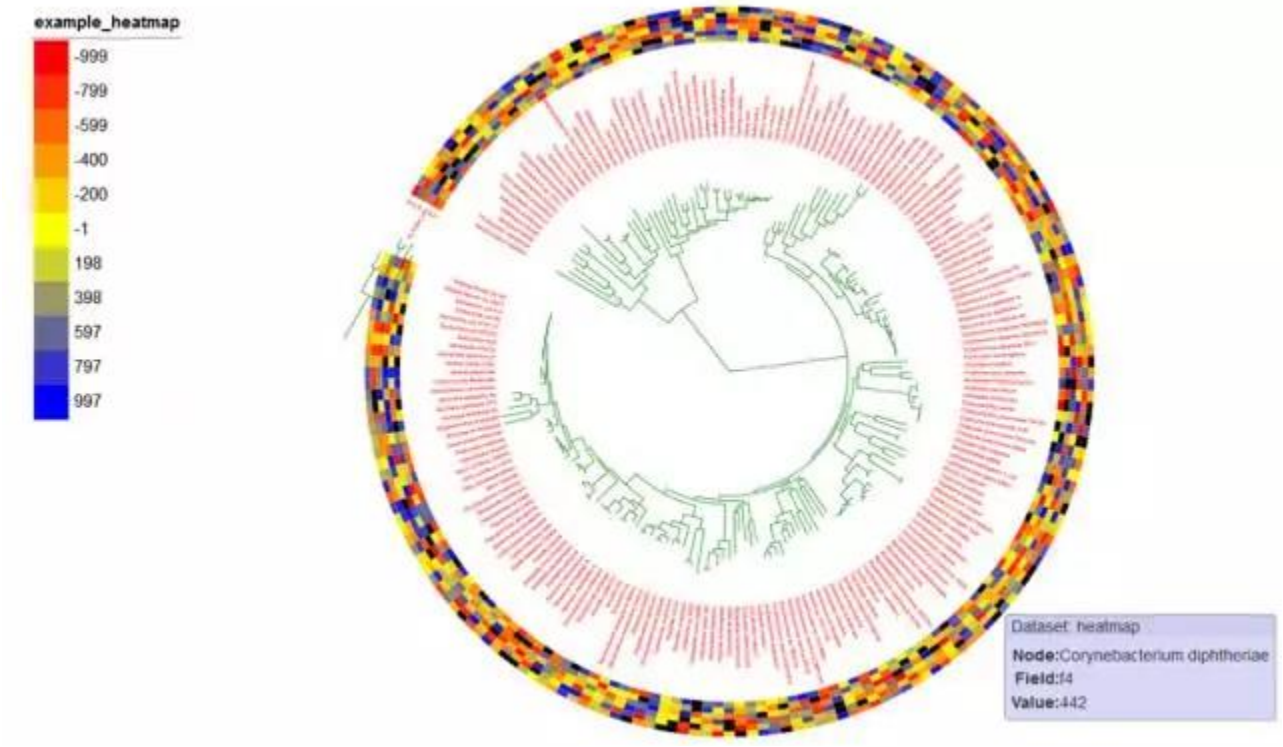
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



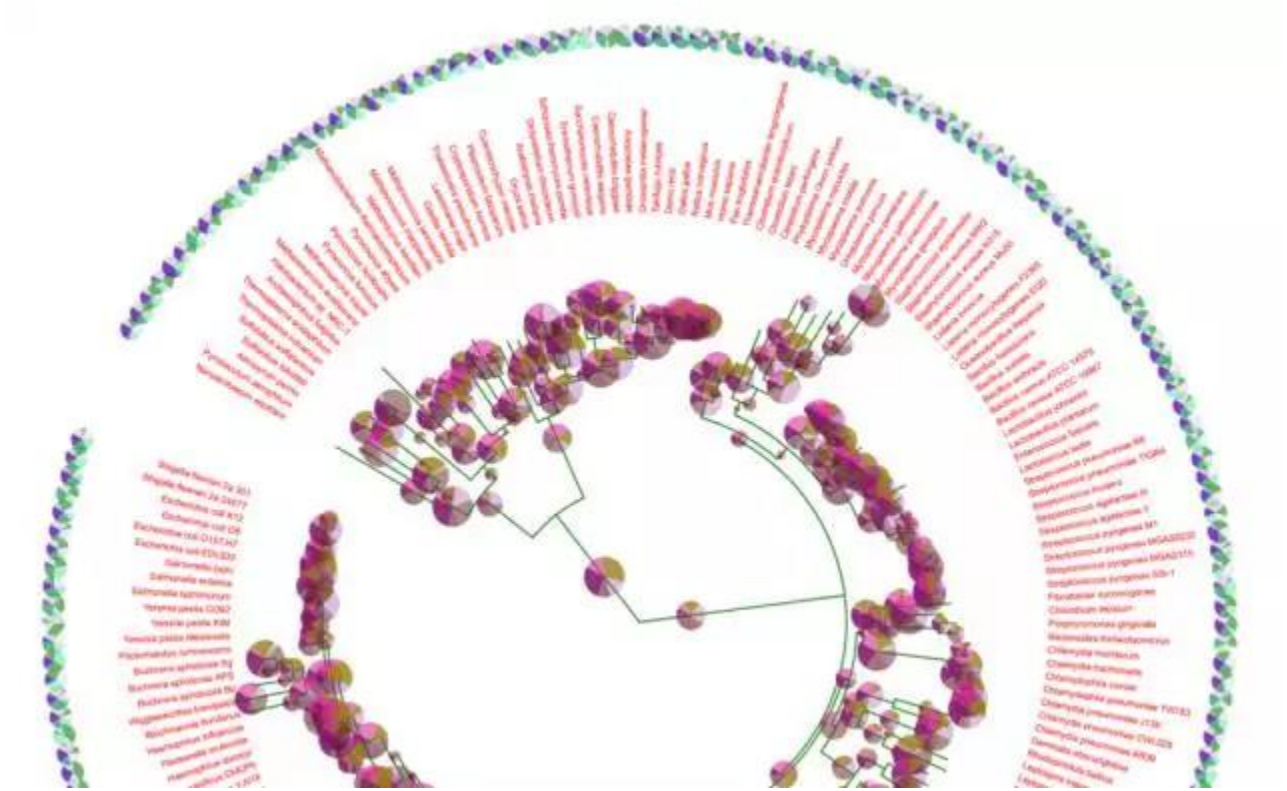
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



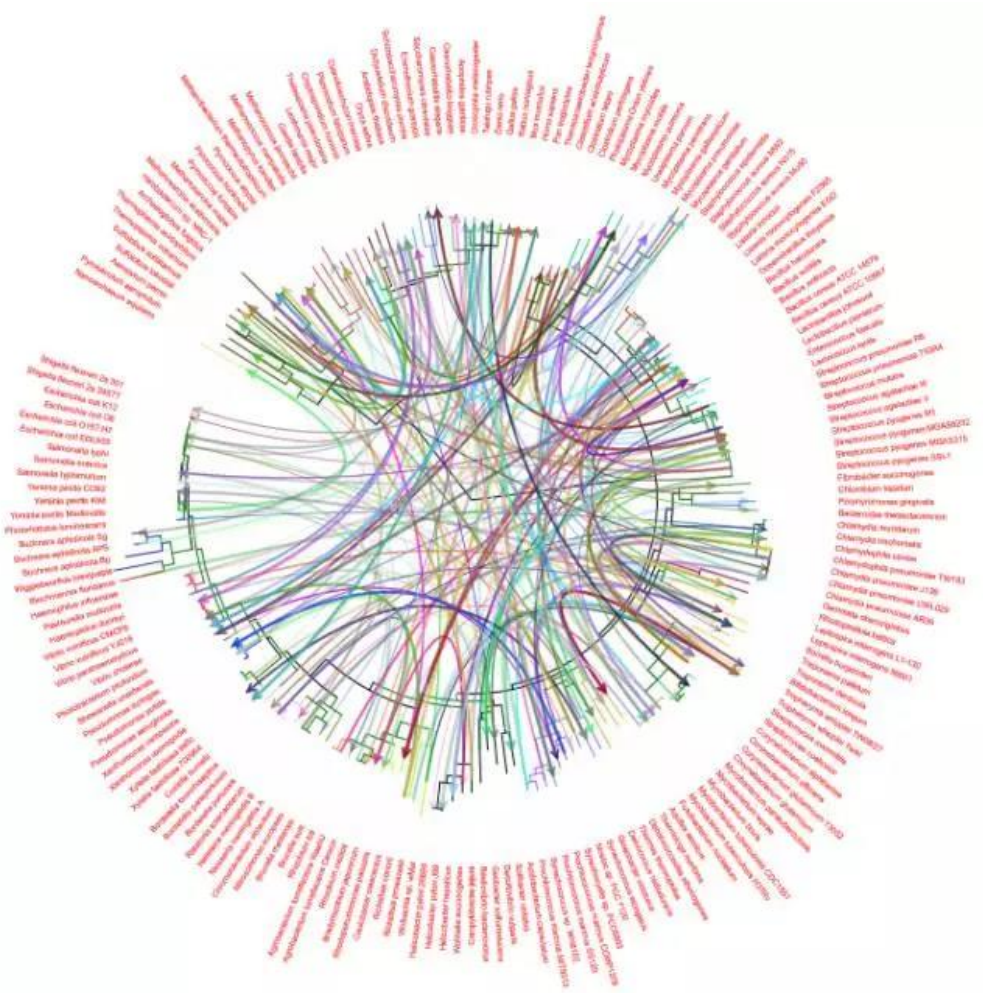
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



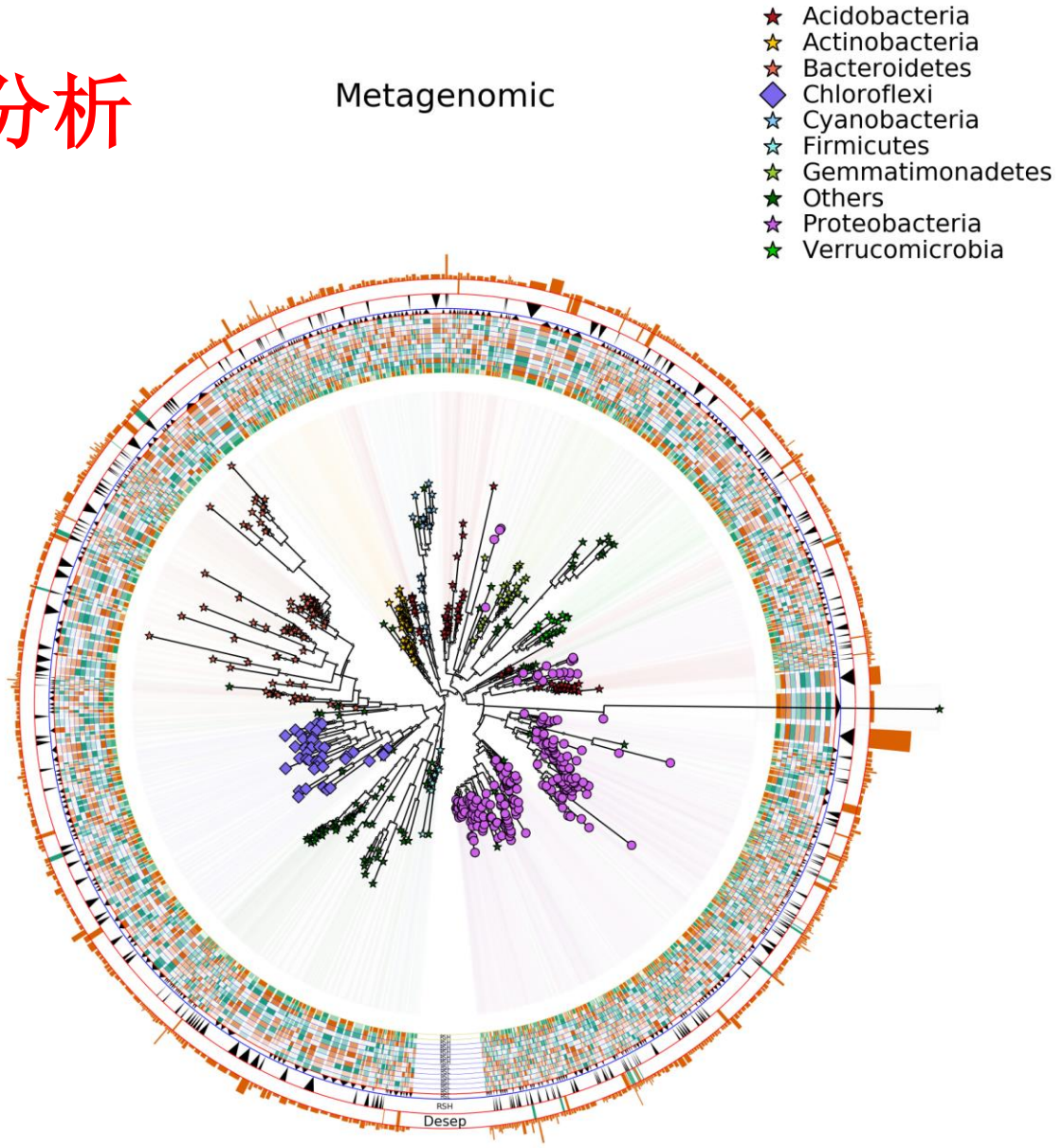
iTOL: https://itol.embl.de/

iTOL：物种进化关系分析



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

GraPhlAn：分类树分析



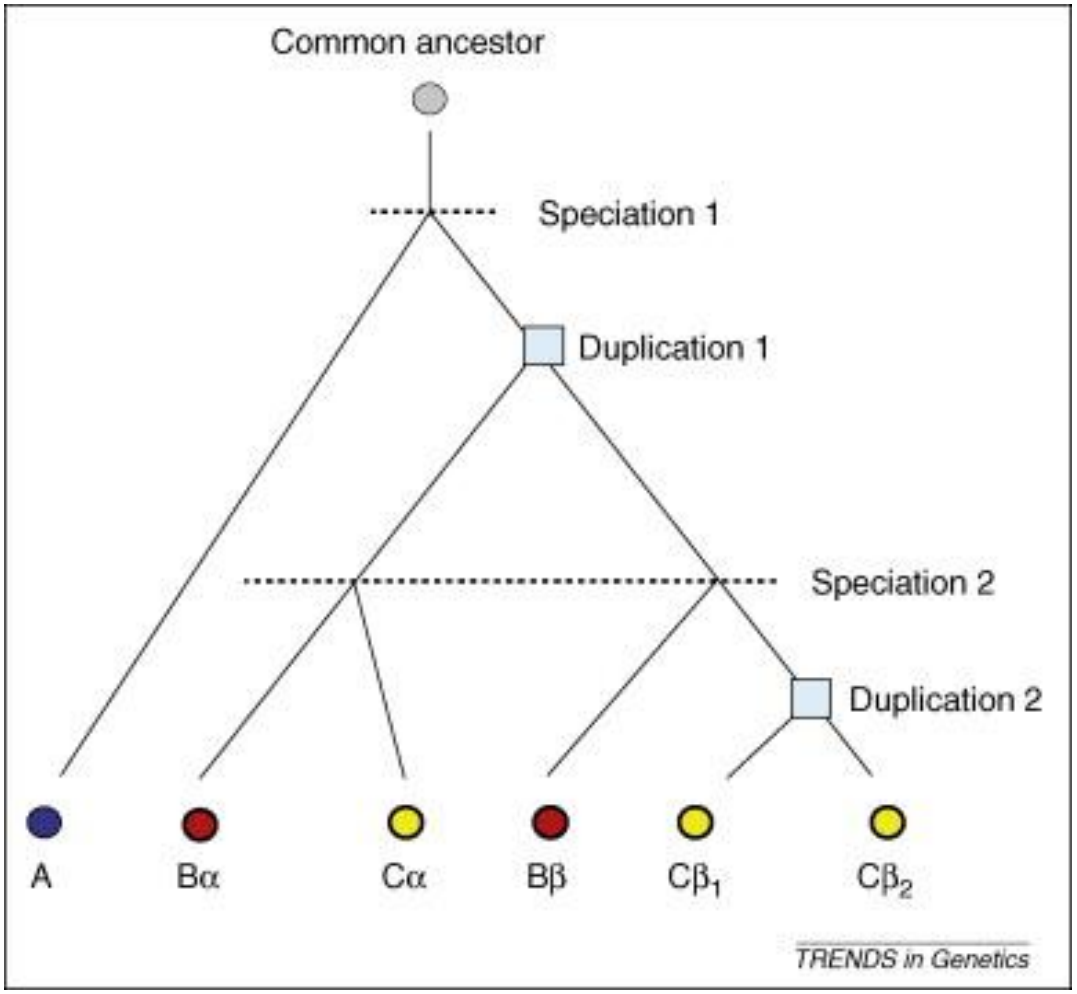
构建分子进化树相关的软件

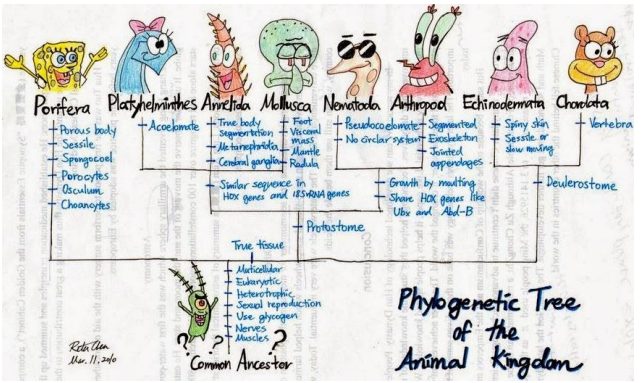
- 就进化树而言，iTOL功能最为全面。iTOL无限制添加的数据集，外环可以制作各种图形包括箱线图等，可以使用多种符号填充外环。但Graphlan就不能这么随便了，只能使用两个符号填充外环。再多的环属性也就是设置环数量和颜色，透明度了。

- 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。

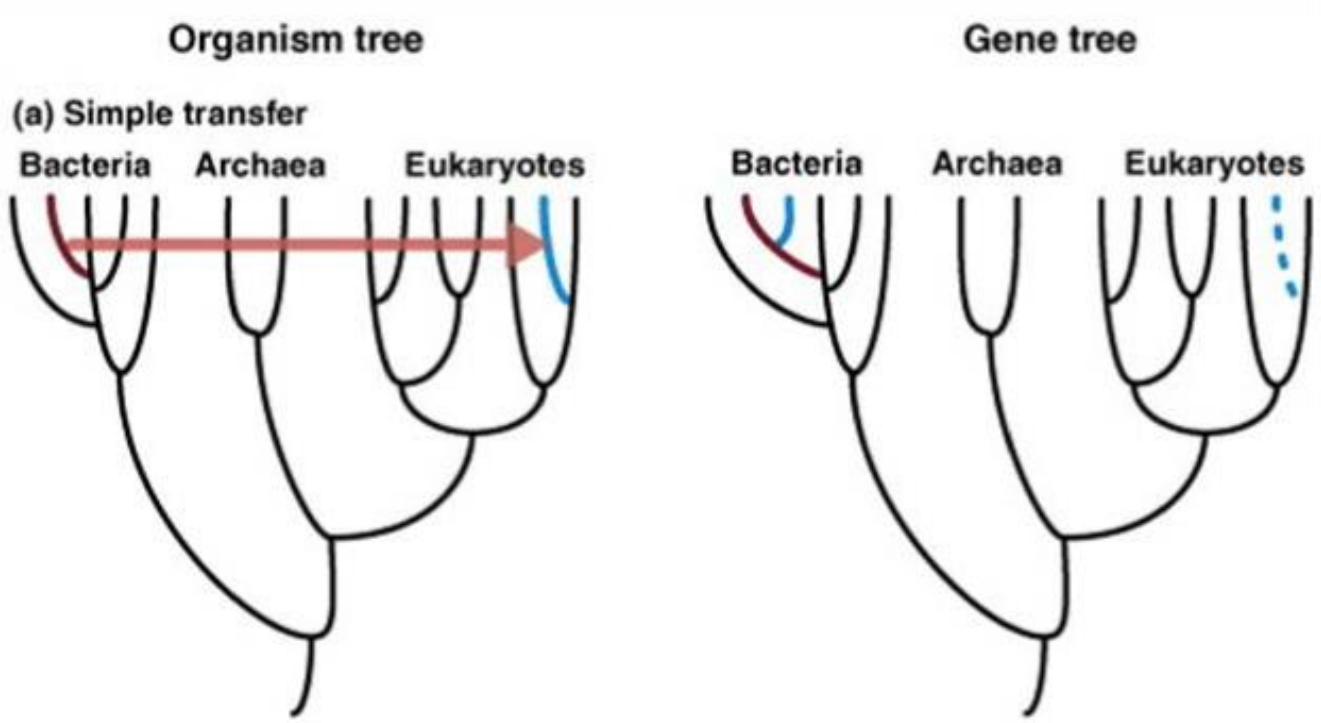
- Graphlan可以制作分类树，是它不同于R包ggtree和iTOL的地方。

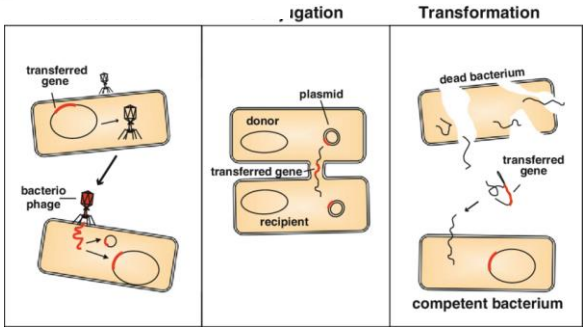
Gene: ortholog and paralog





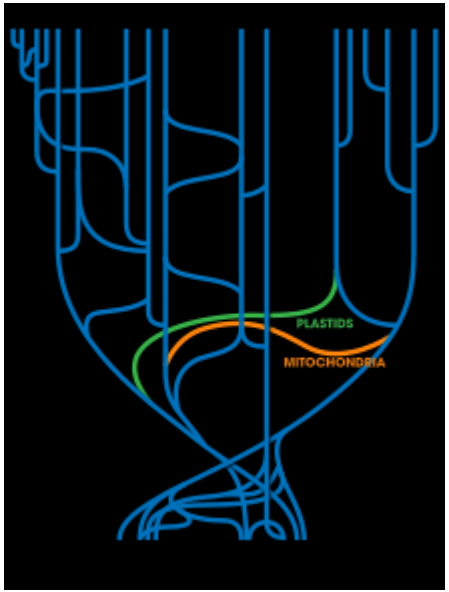
Gene: HGT

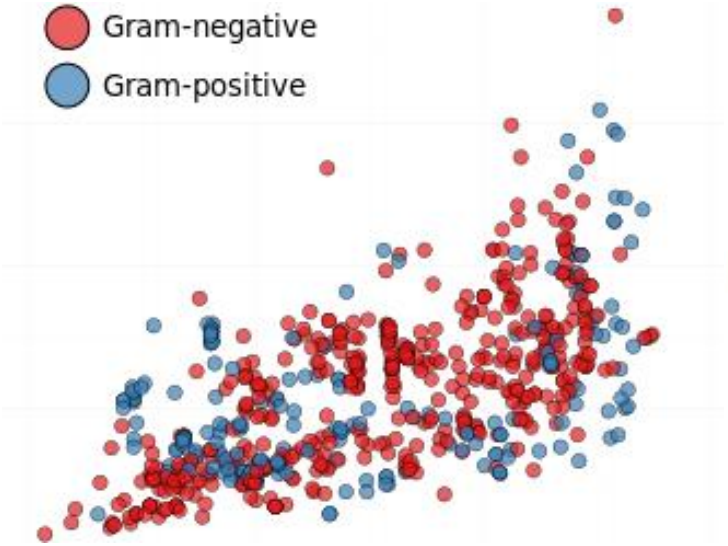

Organism tree
Gene tree
(a) Simple transfer
Bacteria
Archaea
Eukaryotes
Bacteria
Archaea
Eukaryotes


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

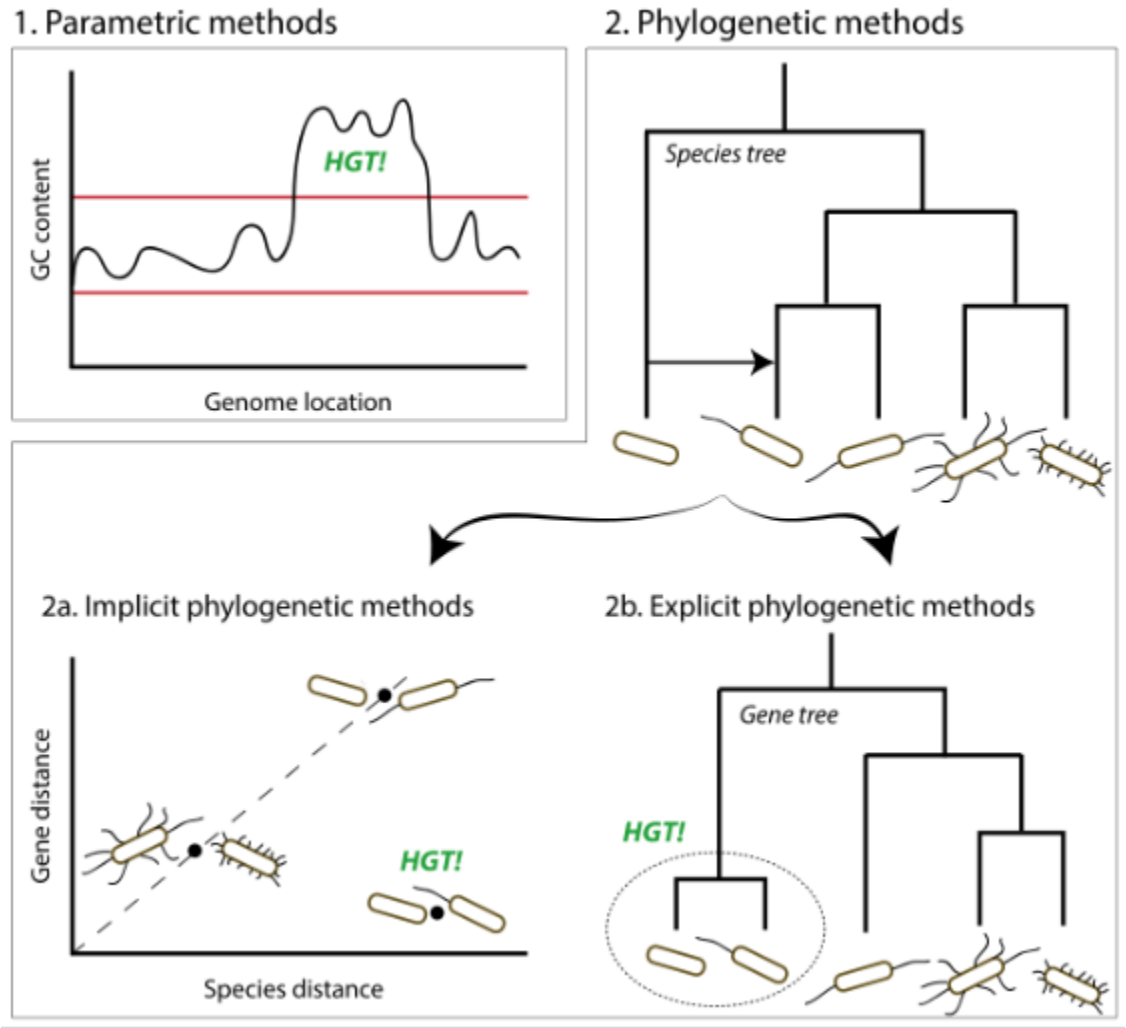
Gene: HGT

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

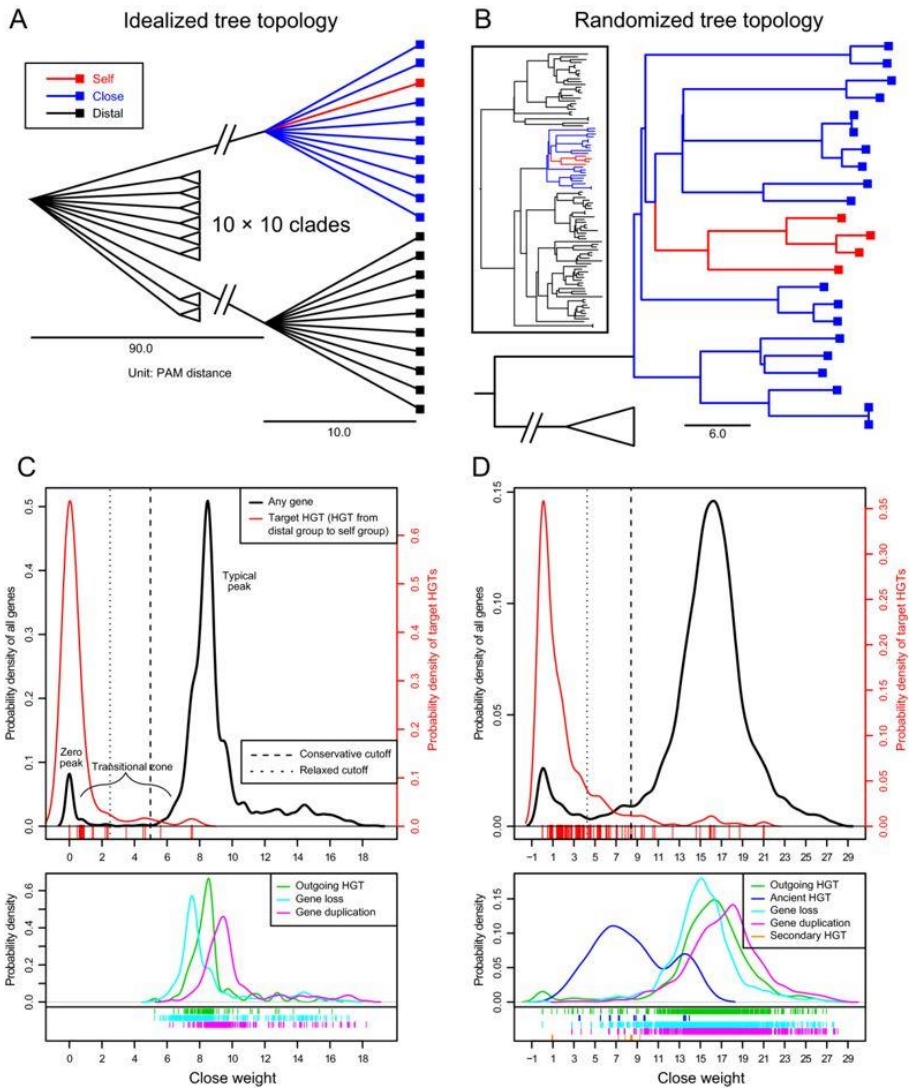




HGT identification methods (phylogeny analysis)



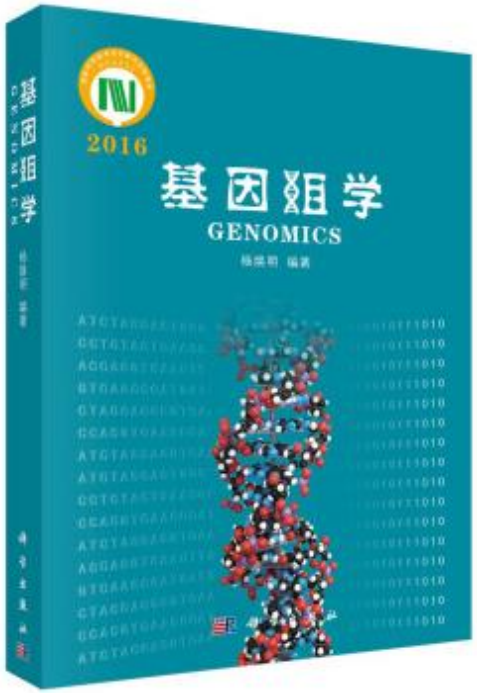
HGT identification methods (phylogeny analysis)

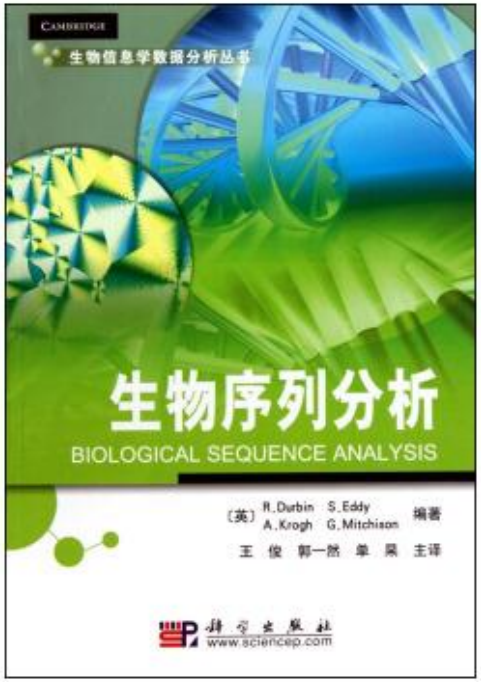


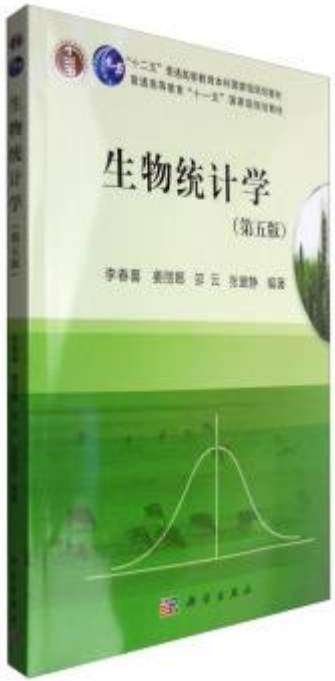
参考文献

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

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 单昊 主译
www.sciencep.com


生物统计学
(第五版)