


Introduction to Computational Biology

《计算生物学导论》

专题二

分子进化与基因组演化

Peking University

Introduction to Computational Biology

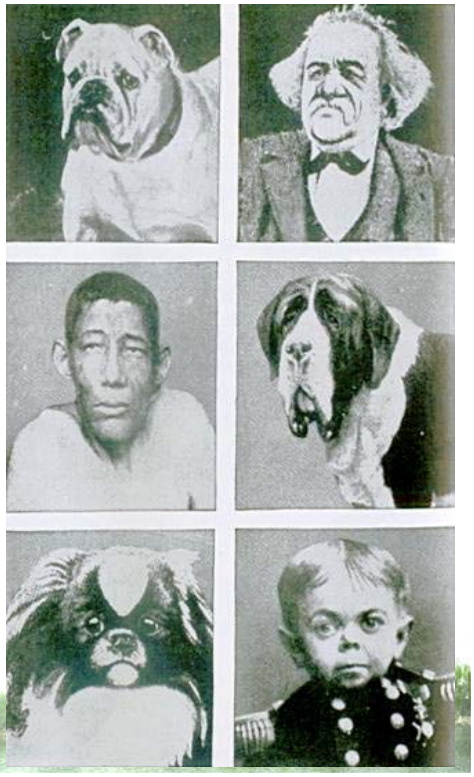
§1 DNA与氨基酸的序列比对

序列比对（sequence alignment）：目前最常用的鉴定和发现同源蛋白的方法。如果两个基因或它们的蛋白序列能够较好地对准，则意味着这两个基因很可能同源。

两序列比对(Pairwise alignment): The process of lining up two sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

```
Query: 181 catcaactacaactccaaagacacccttacaccca
           |||||||| |||| |||||| ||||| | ||||||
Sbjct: 189  catcaactgcaaccccaaagccacccct-caccca

Query: ctaggatatcaacaaacctacccac 240
       ||||||||||||||||||||||||
Sbjct: ctaggatatcaacaaacctacccac 247
```



两序列比对（pairwise alignment）：

- **Protein is more informative** (20 vs. 4 characters); many amino acids share related biophysical properties
- **Codons are degenerate**: changes in the third position often do not alter the amino acid that is specified
- Protein sequences offer **a longer "look-back" time**
- DNA sequences can be translated into protein, and then used in pairwise alignments

Many times, DNA alignments are appropriate

- to confirm the identity of a cDNA
- to study noncoding regions of DNA
- to study DNA polymorphisms

两序列比对广泛应用于基因组研究：

- to decide if two proteins (or genes) are related structurally or functionally
- to identify domains or motifs that are shared between proteins
- the basis of BLAST searching
- used in the analysis of genomes

Peking University
Introduction to Computational Biology

几个重要概念

Homology（同源性）

Similarity attributed to descent from a common ancestor.

Identity（一致性）

The extent to which two (nucleotide or amino acid) sequences are invariant.

```
RBP:         26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVA 59
                   + K++ + ++      GTW++ MA  +      L +      A
glycodelin:  23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKA 55
```

Conservation（保守性）

Changes at a specific position of an amino acid or DNA sequence that preserve the physico-chemical properties of the original residue.

Similarity（相似性）

The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

Pairwise alignment of retinol-binding protein (RBP) and b-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |  |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .       | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Identity (bar)

Very similar (two dots)

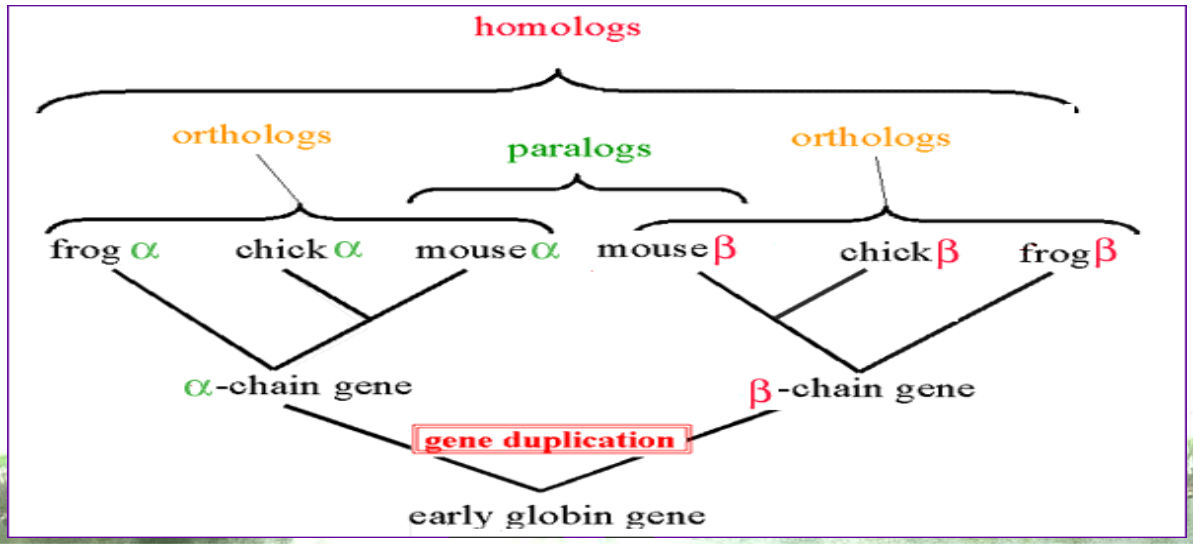
Somewhat similar (one dot)

Peking University
Introduction to Computational Biology

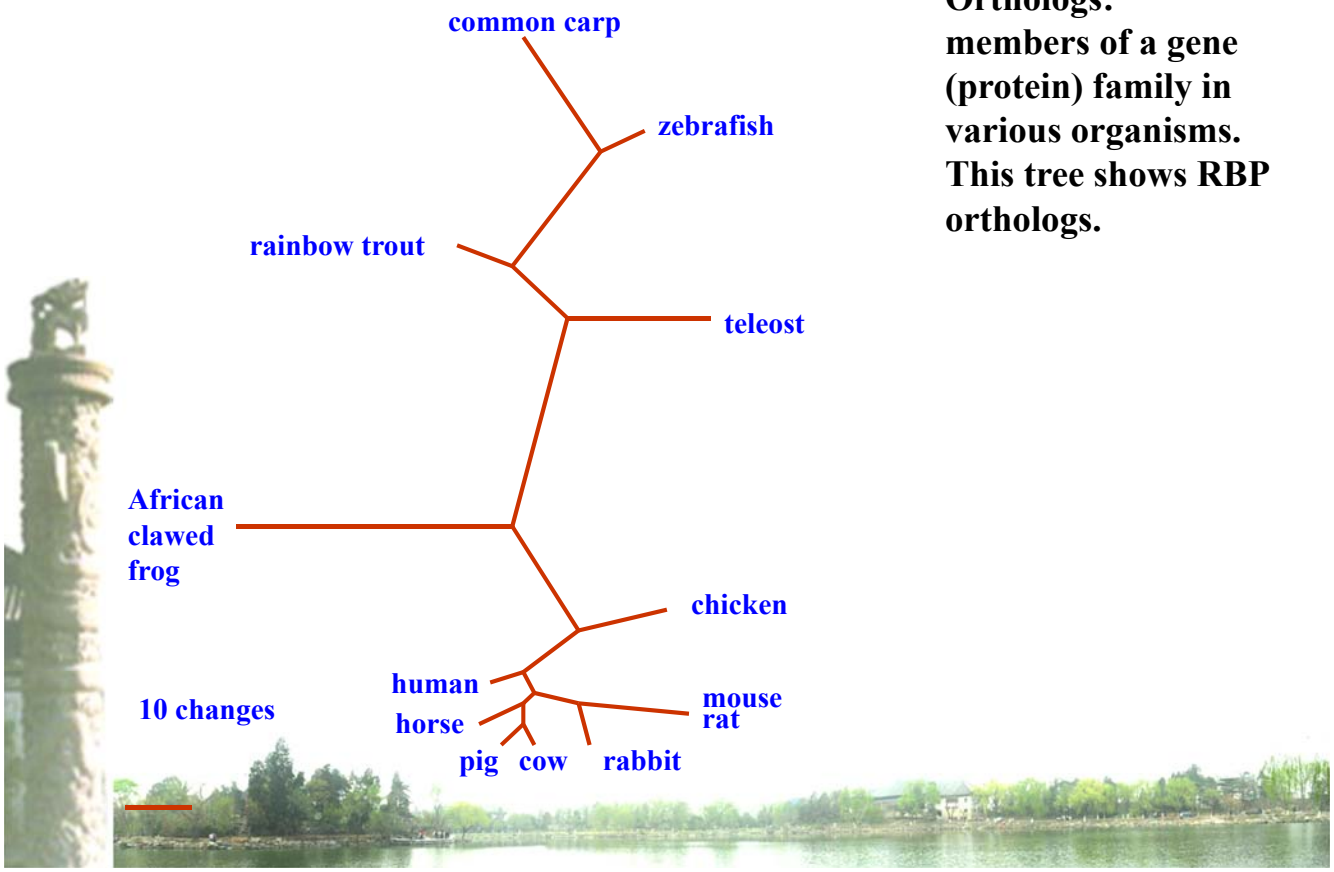
两个重要概念

Orthologs（直向同源）： Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Paralogs（横向同源）： Homologous sequences within a single species that arose by gene duplication; usually are not responsible for the same function.

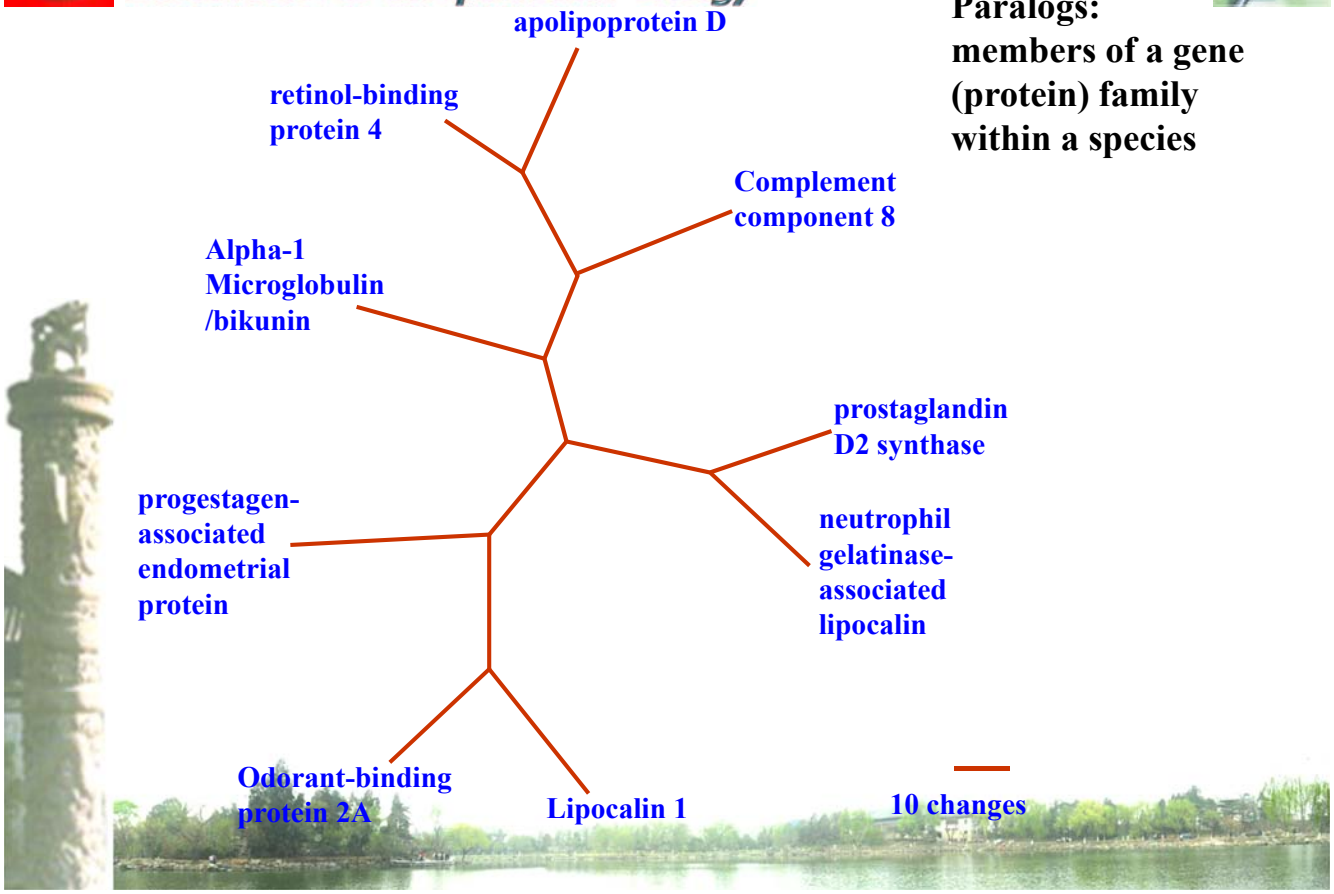


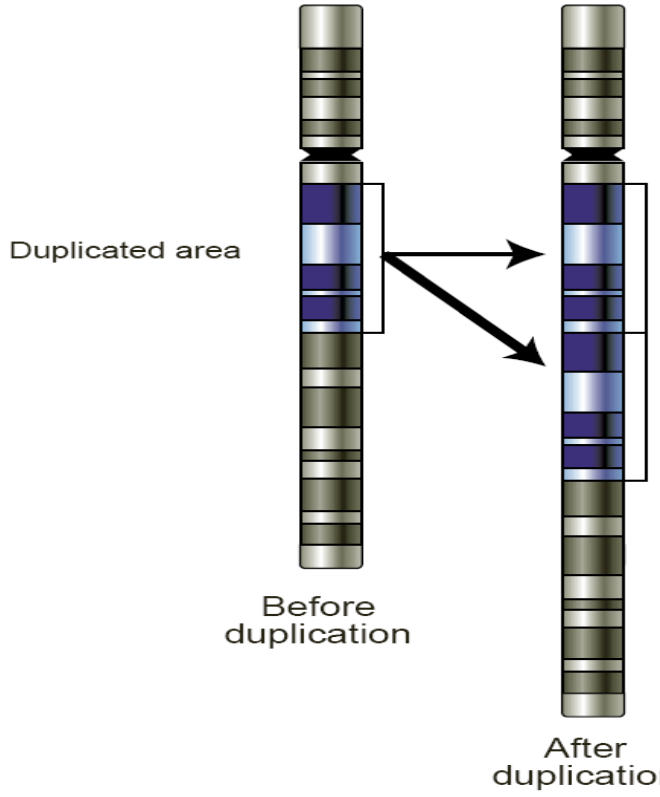
Orthologs: members of a gene (protein) family in various organisms. This tree shows RBP orthologs.



Peking University
Introduction to Computational Biology

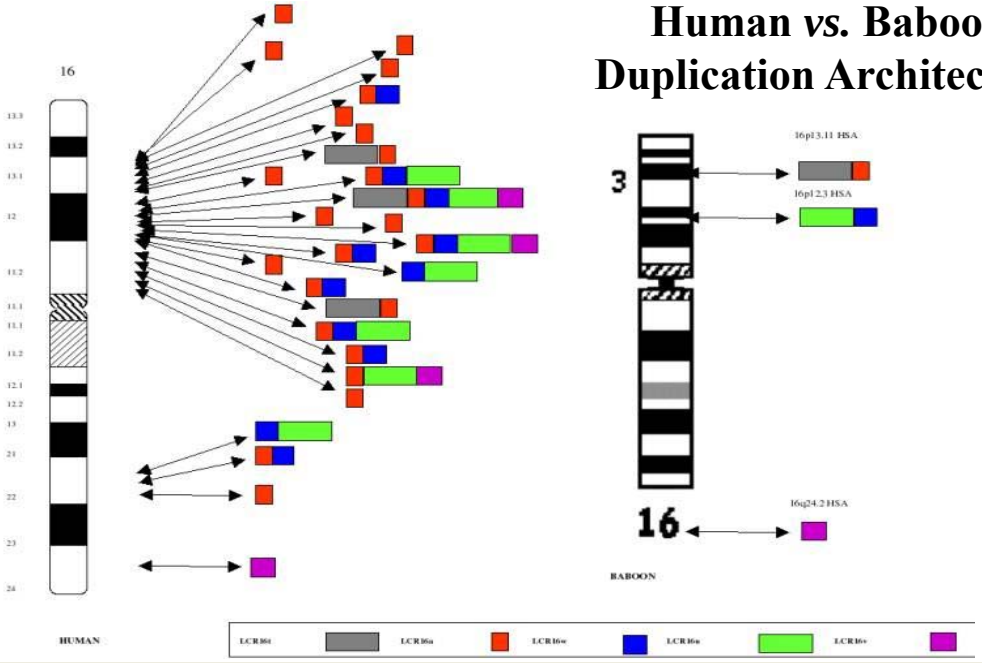
Paralogs: members of a gene (protein) family within a species





A region of a chromosome before and after a duplication event

Human *vs.* Baboon Duplication Architecture

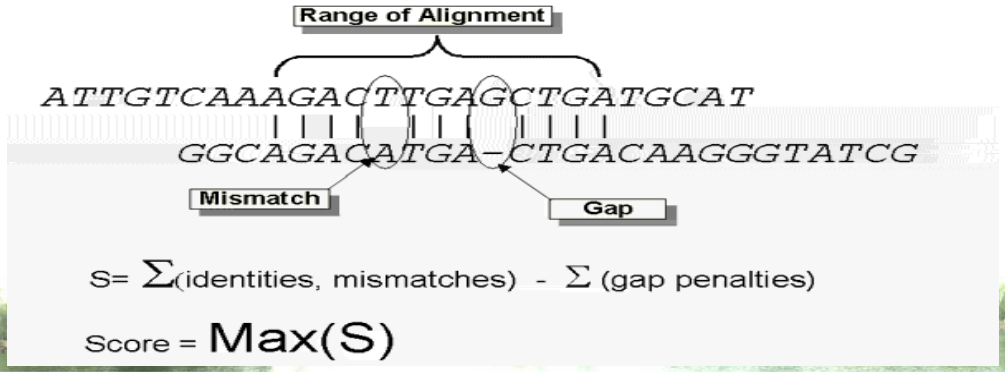


The organization of five LCR16 (low copy repeats on chromosome 16) segmental duplications is compared between human and baboon. In humans, the duplications range in size from 19-75 kb, are 97-99.5% identical and are distributed in different permutations to 27 different map positions shown along the ideogram. In baboons (and other Old World monkeys), the corresponding segments are not duplicated and map to a single locus. The data suggest a dramatic expansion of segmental duplications during hominoid evolution on this chromosome. Note: The LCR16a duplication (red) contains a novel gene family (morpheus) that shows positive selection only in humans and the great-apes. (eichlerlab.gs.washington.edu/primate.html)

General approach to pairwise alignment

- Choose two sequences
- Select an algorithm that generates a score
- Allow gaps (insertions, deletions)
- Score reflects degree of similarity
- Alignments can be global or local
- Estimate probability that the alignment occurred by chance

DNA序列比对的打分



Peking University
Introduction to Computational Biology

氨基酸序列比对的打分：PAM矩阵

Accepted Point Mutations (PAMs)

Pioneer in Bioinformatics
Dr. Margaret O. Dayhoff
(1925-1983)

Substituent residue
(Percentage of total residue sites at which the substituent occurs)

Sequence (original amino acid)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R									50			58				25				
N	33			47					33			33				33	33			
D	44		22				47	34	22			28				25				
C	(66)																			
Q				56			30		40			70								
E	50			44				38				41			24					
G	51			33			30					27				36				
H				26							26	30				22	22			
I	39										58									46
L	21									23		23		28						30
K	23	21		28			31	23			21					21				
M	22									22	89			22						45
F									22		61									
P	50			43			57	43				21								
S	49			24			24	36				24					40			
T	32						28	24				24				52				
W	(40)										(40)			(60)						
Y									(33)					(50)						
V	36									21	43	21								

PAM matrices: Point-accepted mutations

- PAM matrices are based on global alignments of closely related proteins.

- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.

- Other PAM matrices are extrapolated from PAM1. For PAM250, 250 changes have occurred for two proteins over a length of 100 amino acids.

- All the PAM data come from closely related proteins (>85% amino acid identity).

Peking University
Introduction to Computational Biology

PAM250 log odds scoring matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

BLOSUM矩阵

- BLOSUM matrices are based on local alignments.
- BLOSUM stands for blocks substitution matrix.
- BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	⟷	*More divergent*

More conserved
Rat versus mouse RBP

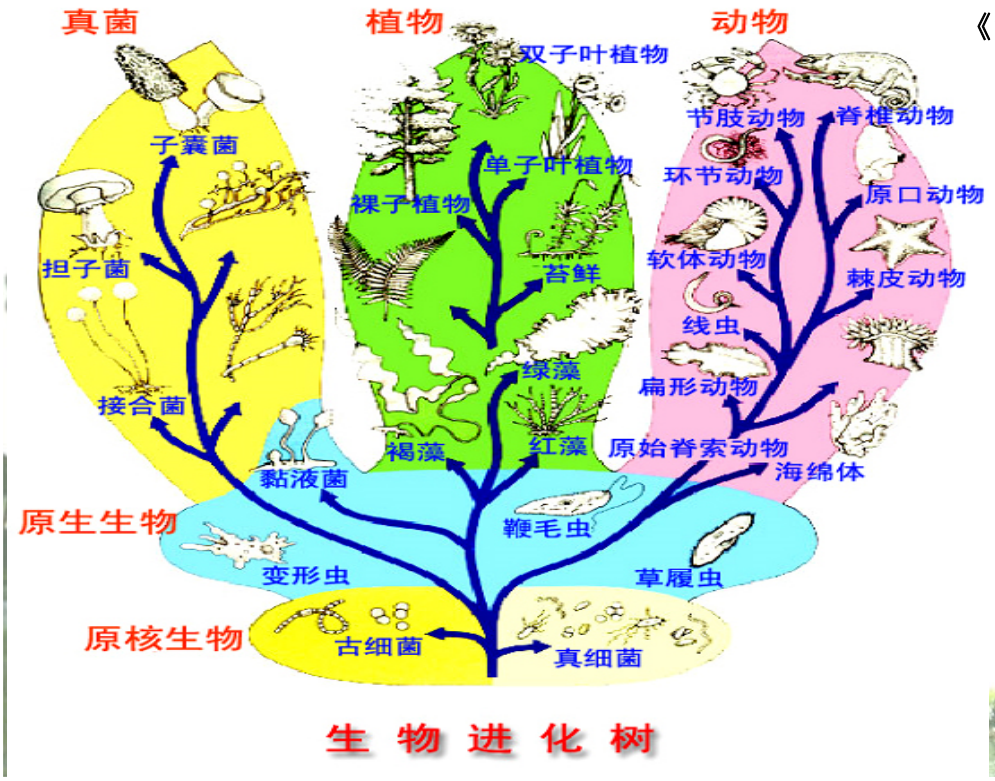
Less conserved
Rat versus bacterial lipocalin

Peking University
Introduction to Computational Biology

Blosum62 scoring matrix

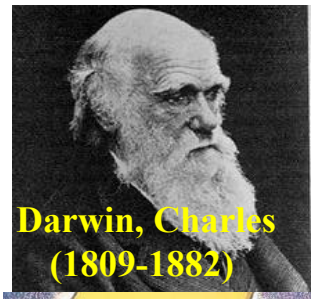
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

§2 生物进化的分子基础

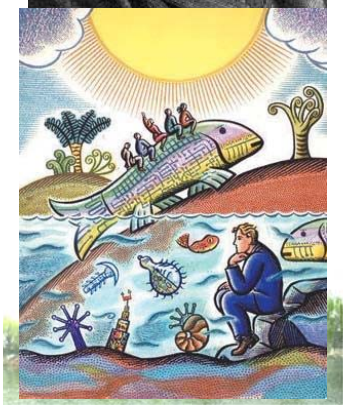


生 物 进 化 树

《*The Origin of Species*》
（1859）



Darwin, Charles
(1809-1882)

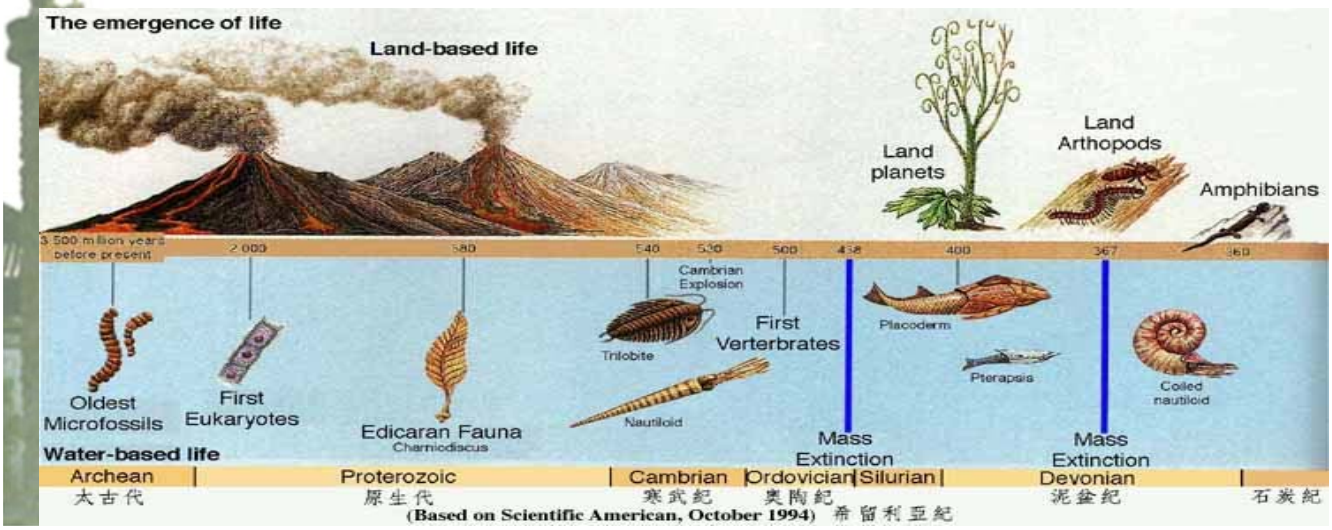


Peking University
Introduction to Computational Biology

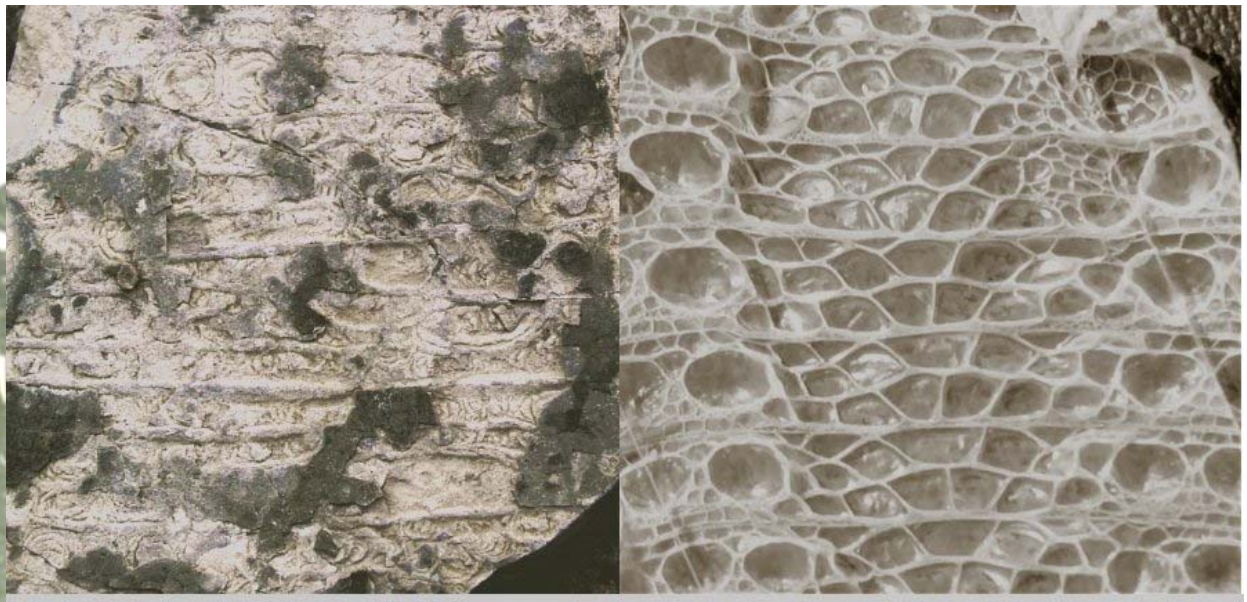
经典的进化研究方法

- 化石证据
  (Fossil)
- 比较形态学证据
  (Comparative morphology)
- 比较生理学证据
  (Comparative physiology)

→ 系统学(Systematics)
分类学(Taxonomy)



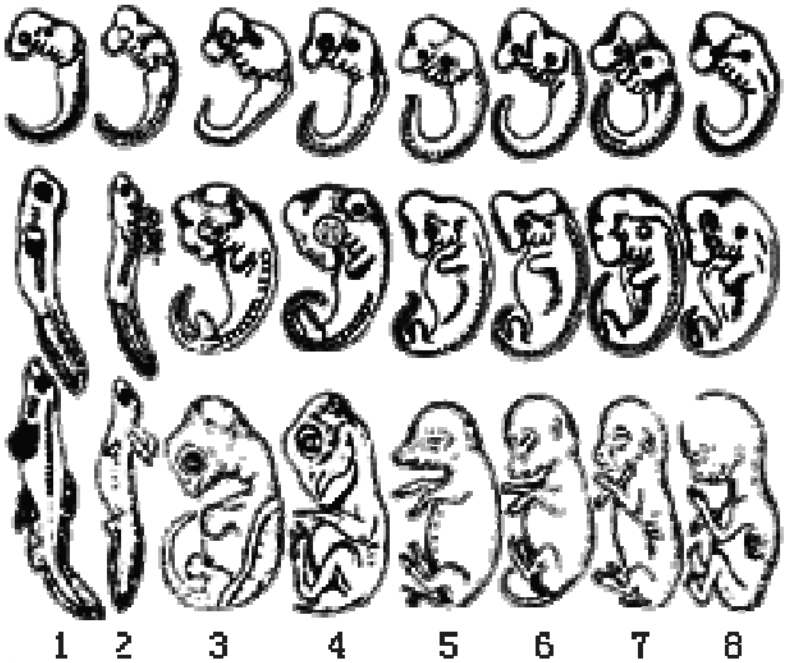
化石比较证据
(Fossil comparison)



Comparison of markings found at the Portland dinosaur footprint site by Stuart Tabner and shed skin of a modern reptile. The fossil impression is on the left while on the right, is shed skin of a living gecko lizard, much enlarged. *Ian West (c) 2003.*

Peking University
Introduction to Computational Biology

比较形态学证据 (Comparative morphology)



进化学的分子途径

- 普适性
  由4种核酸组成 ➔ 分子水平的进化表现为：DNA序列的演化、氨基酸序列演化、蛋白质结构的演化
- 可比较性
  比较不同物种的有关DNA序列 ➔ 建立DNA序列的演化模型、氨基酸序列的演化模型（数学模型）
  蛋白质结构的演化模型
  （形态、性状的演化模型？）
- 基因组编码信息的丰富
  与形态、性状包含的信息相比，基因组序列包含更多、更复杂的信息结构

Peking University
Introduction to Computational Biology

What can we do for molecular evolution?

序列比较：源于同一祖先DNA/氨基酸序列的两条DNA/氨基酸序列，考察二者的差异。

序列差异：进化过程中分子突变的痕迹

分子进化：以累计在DNA/氨基酸分子上的历史信息为基础，研究分子水平的生物进化过程和机制。

分子系统发育学
Molecular Phylogenetics
分子系统学
Molecular Systematics

分子系统学为生物分类问题提供了许多崭新的见解。

生物进化的分子机制

1、核苷酸替代、插入/缺失、重组
2、基因转换

基因突变 → 遗传漂变 / 自然选择 → 固定在生物个体以及物种内

产生新的形态、性状 → 传递给后代

分子系统学是研究进化机制的一个重要工具。

Peking University
Introduction to Computational Biology

DNA序列的突变

DNA分子的改变 → 性状改变

- 核苷酸替代 substitution
- 核苷酸缺失 deletion
- 核苷酸插入 insertion
- 核苷酸倒位 invertion

替代

Thr Tyr Leu Leu
ACC TAT TTG CTG

ACC TCT TTG CTG
Thr Ser Leu Leu

插入

Thr Tyr Leu Leu
ACC TAT TTG CTG

ACC TAC TTT GCT G
Thr Tyr Phe Ala

缺失

Thr Tyr Leu Leu
ACC TAT TTG CTG

ACC TAT TGC TG-
Thr Tyr Cys -

倒位

Thr Tyr Leu Leu
ACC TAT TTG CTG

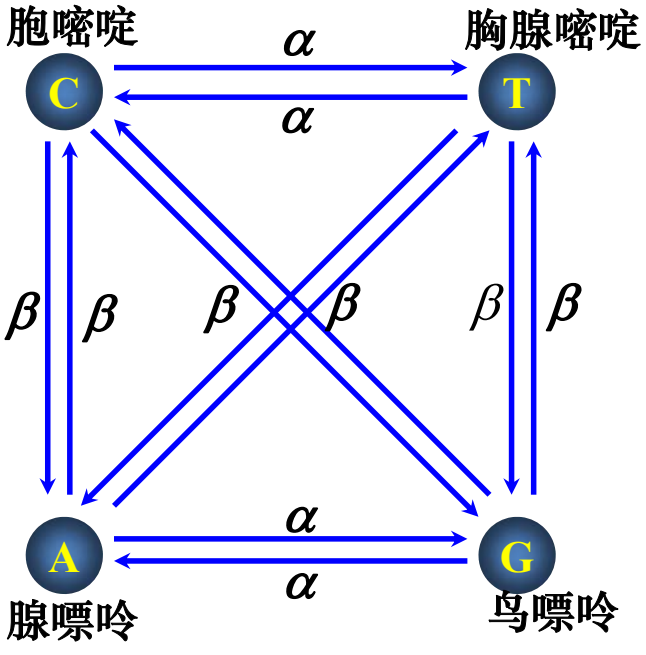
ACC TTT ATG CTG
Thr Phe Met Leu

Peking University
Introduction to Computational Biology

核苷酸替代的几种分类

转换 α
(transition)
嘌呤 ➡ 嘌呤
嘧啶 ➡ 嘧啶

颠换 β
(transvertion)
嘌呤 ➡ 嘧啶
嘧啶 ➡ 嘌呤



在大多数DNA片段中，转换出现的概率高于颠换出现的概率。

DNA序列突变对氨基酸序列的影响

- 同义（沉默）替代（synonymous / silent substitution）

仍然为同义密码子的核苷酸替代

如： TAT ➡ TAC

Tyr ➡ Tyr

- 非同义替代（nonsynonymous substitution）

导致产生非同义密码子的核苷酸替代

如： TAT ➡ AAT

Tyr ➡ Asn

- 无义突变（nonsense mutation）

导致产生终止密码子的核苷酸突变

如： TAT ➡ TAA

Tyr ➡ STP

思考题：假设所有密码子以同一概率出现，上述三种突变的比例分别是多少？（25%，71%，4%）

Peking University
Introduction to Computational Biology

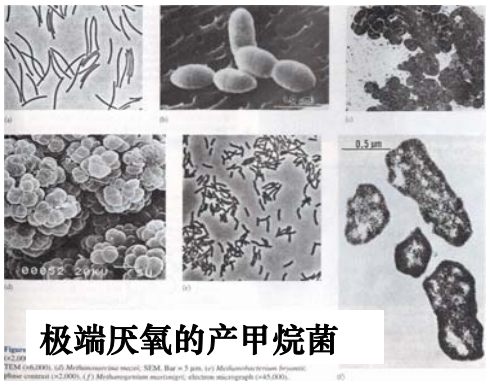
密码子使用频率(codon usage)

遗 传 密 码 表

第一位 (5'端)	第二位 U	第二位 C	第二位 A	第二位 G	第三位 (3'端)
U	UUU phe	UCU ser	UAU tyr	UGU cys	U
	UUC phe	UCC ser	UAC tyr	UGC cys	C
	UUA leu	UCA ser	UAA 终止	UGA 终止	A
	UUG leu	UCG ser	UAG 终止	UGG trp	G
C	CUU leu	CCU pro	CAU his	CGU arg	U
	CUC leu	CCC pro	CAC his	CGC arg	C
	CUA leu	CCA pro	CAA gln	CGA arg	A
	CUG leu	CCG pro	CAG gln	CGG arg	G
A	AUU ile	ACU thr	AAU asn	AGU ser	U
	AUC ile	ACC thr	AAC asn	AGC ser	C
	AUA ile	ACA thr	AAA lys	AGA arg	A
	AUG 起始	ACG thr	AAG lys	AGG arg	G
G	GUU val	GCU ala	GAU asp	GGU gly	U
	GUC val	GCC ala	GAC asp	GGC gly	C
	GUA val	GCA ala	GAA glu	GGA gly	A
	GUG val	GCG ala	GAG glu	GGG gly	G

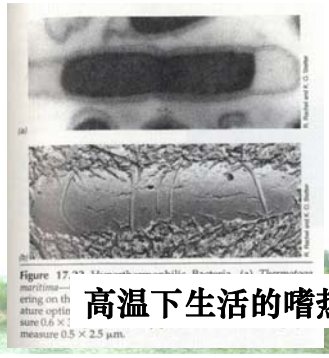
§3 系统发育树（Phylogenetic tree）

1964年美国生物学家在黄石公园的温泉源头发现了微生物，第二年夏天又发现了在60摄氏度的水中生活的水藻，还有在82摄氏度的水温下生存的微生物。



极端厌氧的产甲烷菌

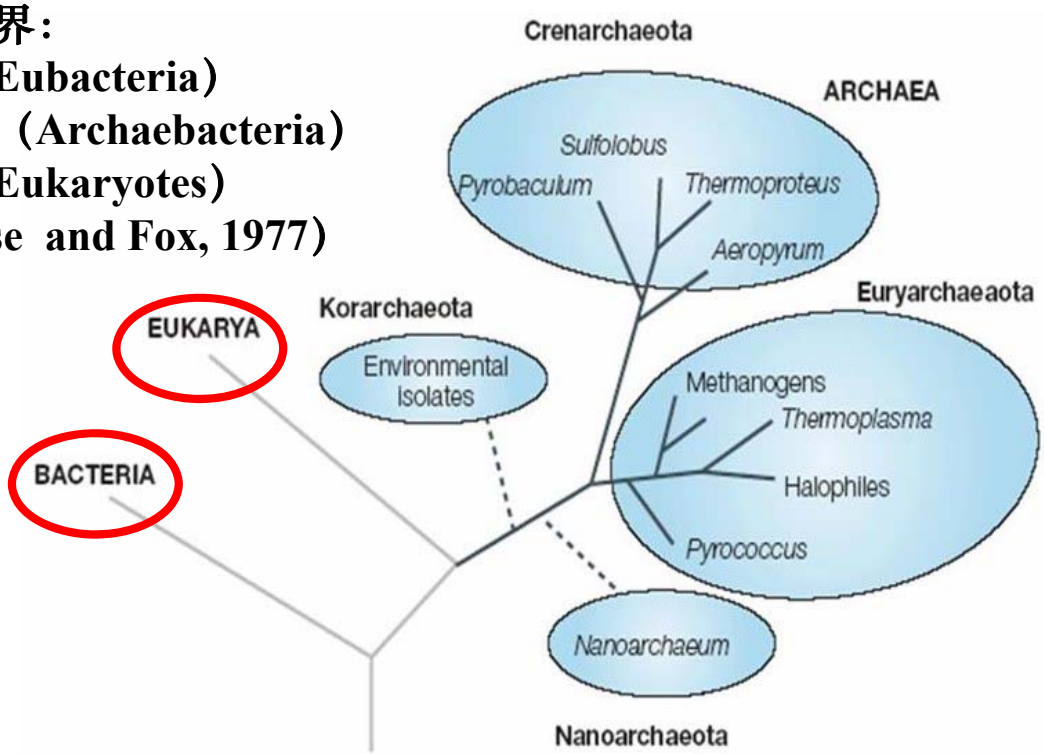
美国黄石公园内有许多温泉，水温从20°C到100°C，其中生活着一些喜欢热的微生物，用显微镜观察，这些微生物呈杆状。



高温下生活的嗜热菌

Peking University
Introduction to Computational Biology

生命三界：
细菌（Eubacteria）
古细菌（Archaebacteria）
真核（Eukaryotes）
（Woese and Fox, 1977）



基于16S/18S核糖体RNA序列比对得到的古细菌系统发育树
（Ettema等，2005）

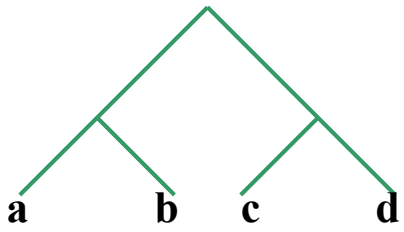
系统发育树的种类
——有根树、无根树

理论上，一个DNA序列在物种形成或基因复制时，分裂成两个子序列，因此系统发育树一般是二歧的。

一般考虑二歧的树结构：二歧树

拓扑结构：

有根树：反映时间顺序

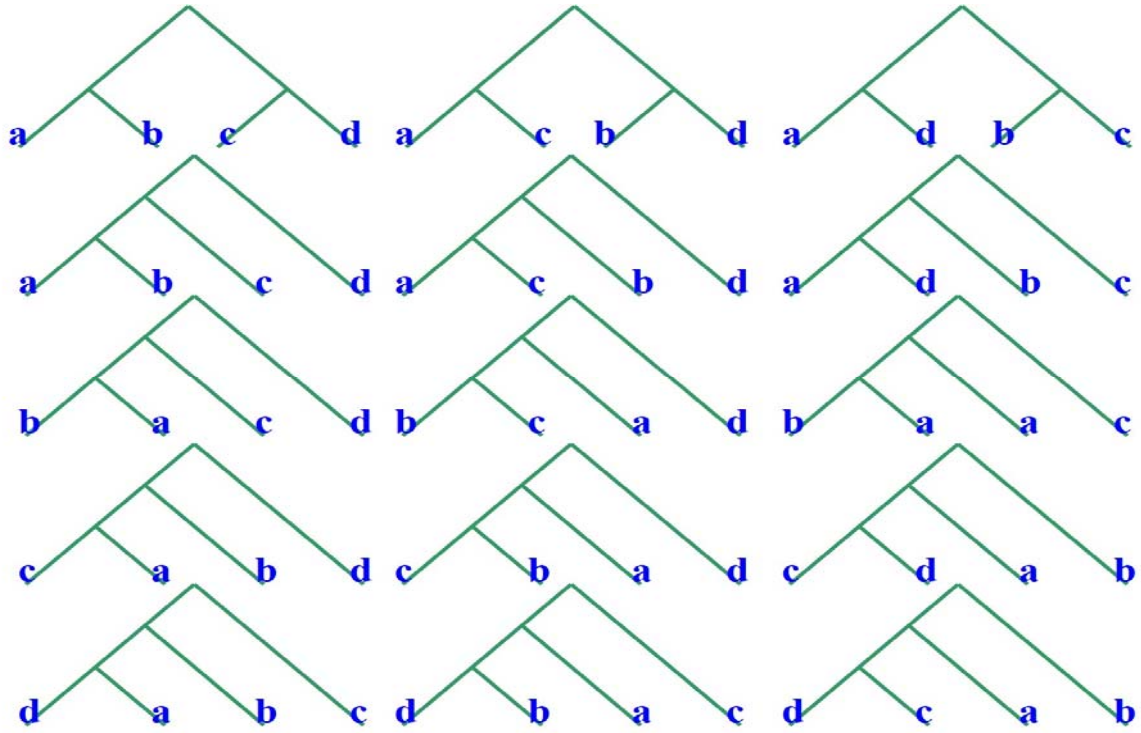


节点：
内部节点
外部节点

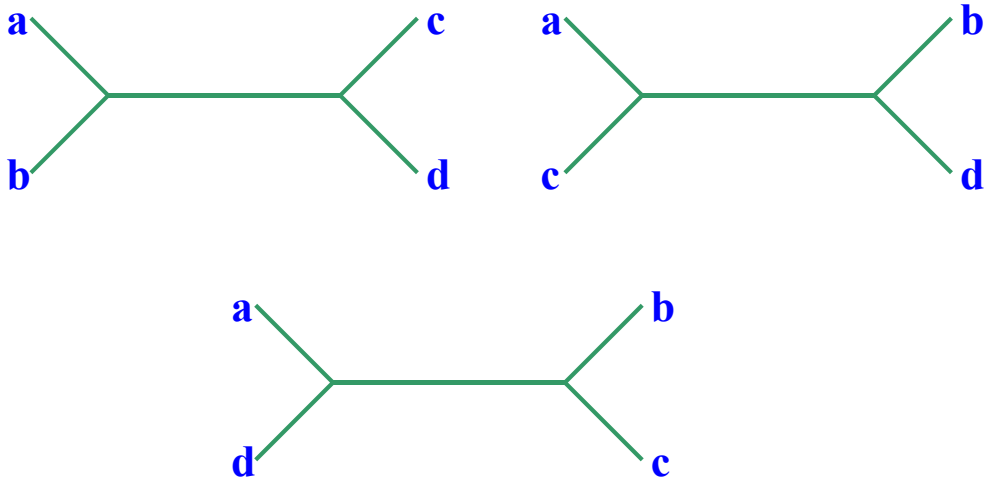
无根树：反映距离

分支：
内部分支
外部分支

Peking University
Introduction to Computational Biology



考虑4个分类群时，共有15种可能的有根树



考虑4个分类群时，共有3种可能的无根树

Peking University
Introduction to Computational Biology

考察类群数为 m ($m \geq 3$)的系统树，其可能的拓扑结构数目为：

有根树 $\dfrac{(2m-3)!}{2^{m-2}\cdot(m-2)!}$ m=10: 34,459,425种

无根树 $\dfrac{(2m-5)!}{2^{m-3}\cdot(m-3)!}$ m=10: 2,027,025种

当 m较大时，选出真实树的拓扑结构十分困难。

分支数目：

有根树 $2m-2$　　无根树 $2m-3$

内部分支数目：

有根树 $m-2$　　无根树 $m-3$

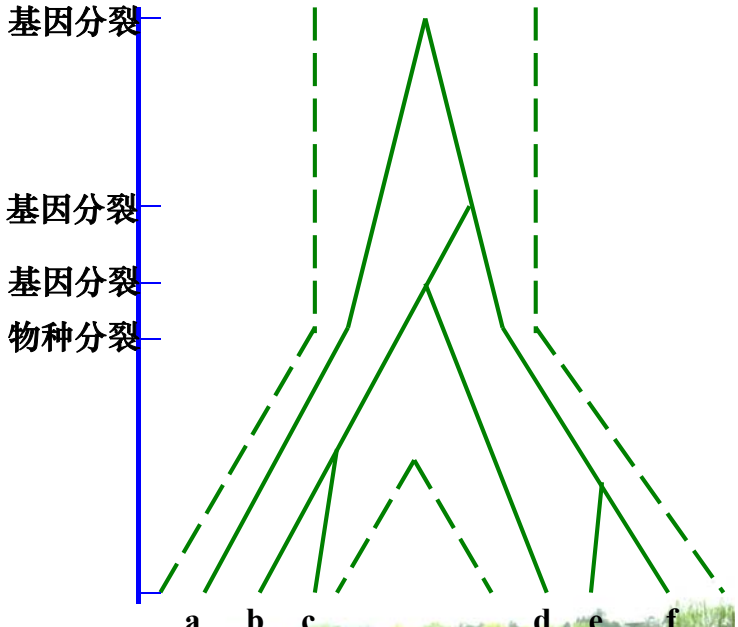
内部节点数目：

有根树 $m-1$　　无根树 $m-2$

Peking University
Introduction to Computational Biology

系统发育树的种类
——基因树、物种树

- 物种树：
  代表一个物种或群体进化历史的系统发育树
  两个物种分歧的时间：两个物种发生生殖隔离的时间
- 基因树：
  由来自各个物种的一个基因构建的系统发育树（不完全等同于物种树），表示基因分离的时间。



系统发育树的种类
——期望树、现实树和重建树

理论上：

假设所研究的序列无限长，从中随机抽样进行统计分析。

实际情况：

所研究的序列是短序列，统计得到的替代数目存在大量随机误差。

期望树：
一个用无限长的序列或每一分支的期望替代数构建的树

现实树：
建立在实际替代数基础上的树

构树方法

重建树

Peking University
Introduction to Computational Biology

系统发育树的构建

构建系统发育树的数据

1、特征数据(character data)：
提供了基因、个体、群体或物种的信息

2、距离数据(distance data)或相似性数据(similarity data)：
涉及的则是成对基因、个体、群体或物种的信息。
➔ 距离矩阵

距离数据可以由特征数据计算得到。
反之?

构造系统发育树的主要方法

•距离法　根据每对物种之间的距离直接计算得到。所生成的树的质量取决于距离尺度的质量
•简约法　通过寻求物种间最小的变更数来完成的
•似然法　通过标准的统计推断建立系统发育的概率模型
•其它方法：神经网络方法、Hadamard结合法……

构建系统发育树的主要过程

1、拓扑结构的判别（从大量的拓扑结构中搜寻、判别）
2、一个既定拓扑结构的分支长度的估计

最优原则

Peking University
Introduction to Computational Biology

系统进化树的构建方法：距离法

1、首先要获得所有分类群之间的进化距离。
2、系统进化树的构建是基于进化距离之间的关系。

如何获得所有分类群之间的进化距离

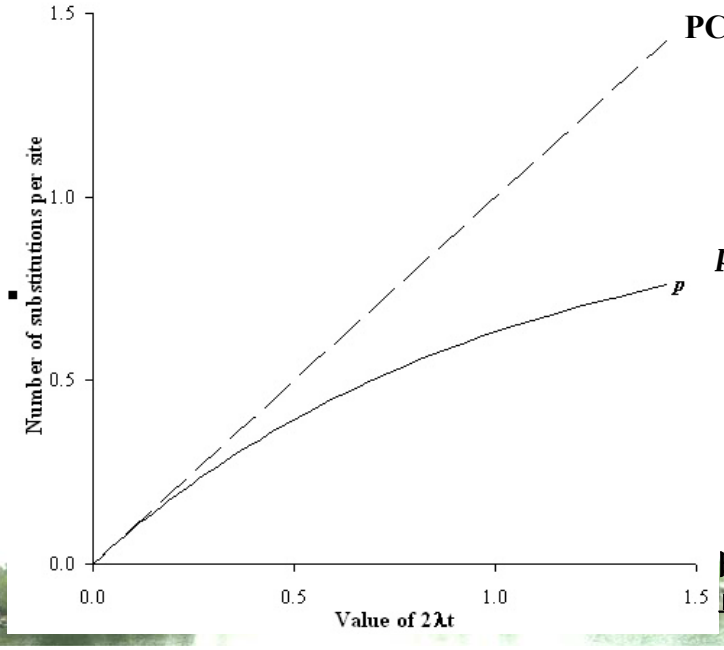
1、选定分类群共同的特征序列——氨基酸序列、核苷酸序列
例：人、马、牛、袋鼠、蝾螈、鲤鱼的血红蛋白α链的氨基酸序 列（140aa）；
人、猕猴、黑猩猩的线粒体DNA中细胞色素b基因的核苷酸序列（1,125bp）；

2、比较两两序列之间的差异 *p*
（序列比对算法）

3、根据不同的概率统计模型，由两条序列的差异*p*值构建它们的进化距离

氨基酸序列：PC（Poisson校正）距离、Γ距离

核苷酸序列：Jukes-Cantor模型、Kimura模型、HKY模型等……



时间
单位：千万年

Peking University
Introduction to Computational Biology

系统进化树的构建方法：最大简约法
(Maximum Parsimony Method)

MP算法基本思想

(Fitch, 1971; Hartigan, 1973)

考虑m个核苷酸（或氨基酸）序列（$m \geq 4$），假定4种核苷酸（或20种氨基酸）可突变为与自身不同的任何一种。

1）对于任一给定的拓扑结构，可以推断每个位点的祖先状态；

2）对于该拓扑结构，可以计算出用来解释整个进化过程所需的核苷酸（或氨基酸）的最小替代数目；

3）对所有可能正确的拓扑结构计算它们的最小替代数目，选择其中最小的作为最优拓扑结构。

Okkham's Razor/Accam's Razor

Entities should *not* be multiplied unnecessarily
如无必要，勿增实体!

Pluralitas non est ponenda sine necessitate.
Frustra fit per plura quod potest fieri per pauciora.
Entia non sunt multiplicanda praeter necessitatem

•万事万物应该尽量简单，而不是更简单。（爱因斯坦）

哲学家、圣方济各会修士

奥卡姆的威廉
(1284-1347)

Peking University
Introduction to Computational Biology

MP法适用的问题

（1）位点不存在回复突变、平行突变；

（2）被分析的序列较长，核苷酸或氨基酸数目很大；

（3）序列的相似度较高；

（4）核苷酸或氨基酸替代速率较稳定。

详细内容请参考《分子进化与系统发育》（高等教育出版社）

系统进化树的构建方法：最大似然法
(Maximum Likelihood Method)

ML算法基本思想
(Felsenstein, 1981; Kishino, 1990)

以一个特定的替代模型分析一组给定的核苷酸（或氨基酸）序列数据，使获得的每一个拓扑结构的似然率均为最大，挑选其中最大似然率最大的拓扑结构，选为最终系统树。

ML法考察的既可以是拓扑结构，也可以是既定拓扑结构的分支长度。

ML法采用了标准的统计方法，以建立进化的概率模型。

计算量非常大。

Peking University
Introduction to Computational Biology

常用分子进化与系统进化分析的软件

软件名称	网址	说明
PHYLIP	http://evolution.gs.washington.edu/phylip.html	It includes programs to carry out parsimony, distance matrix methods, maximum likelihood, and other methods on a variety of types of data, including DNA and RNA sequences, protein sequences, restriction sites, 0/1 discrete characters data, gene frequencies, continuous characters and distance matrices.
PAUP	http://paup.csit.fsu.edu/	It includes parsimony, distance matrix, invariants, and maximum likelihood methods and many indices and statistical tests.
Tree of Life	http://phylogeny.arizona.edu/tree/program/program.html	Arizona大学开发的软件
MEGA	http://www.megasoftware.net	美国宾州州立大学Masatoshi Nei开发 (It carries out parsimony, distance matrix and likelihood methods for molecular data.)

软件名称	网址	说明
MOLPHY	http://www.ism.ac.jp/software/ismlib/softother.e.html#molphy	日本国立统计数理研究所开发。(Carrying out maximum likelihood inference of phylogenies for either nucleotide sequences or protein sequences.)
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	英国伦敦学院ZH YANG开发。(A package of programs for the ML analysis of nucleotide or protein sequences.)
PUZZLE	ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle	应用Quarter puzzling方法（一种最大简约法）构建系统进化树
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	A program for displaying trees on Apple Macs and Windows PCs. It can draw rooted and unrooted trees, display bootstrap values, and supports the native font and graphics file formats of both Macs and PCs.
phylogeny	http://www.ebi.ac.uk/biocat/phylogeny.html	EBI的系统进化树分析软件

Peking University
Introduction to Computational Biology

§4 基因组的进化

Mobile Elements: Drivers of GENOME EVOLUTION

Mobile elements （events）包括：

转座子（transposon），水平基因转移（Horizontal gene transfer），基因复制（Gene duplication），缺失（Loss），……

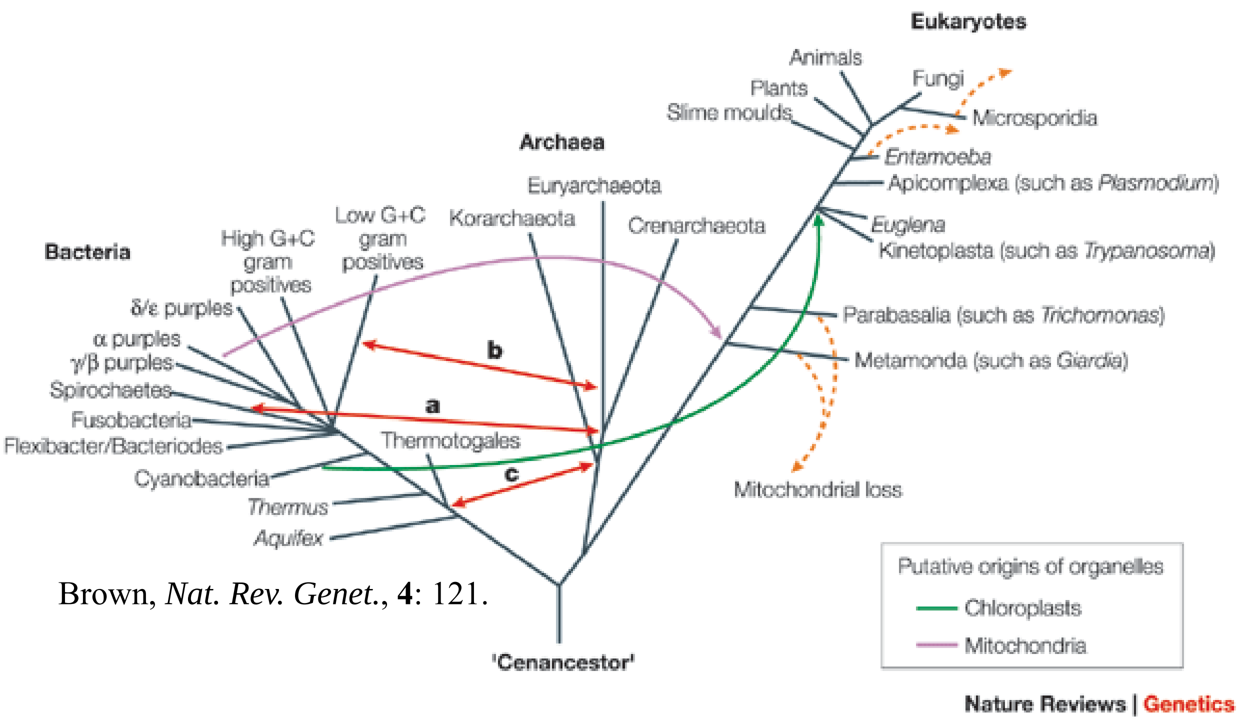
包括两方面含义：

- Mobile materials：遗传物质在基因组内部、基因组之间的交换、迁移和扩增
- Mutations：交换和迁移的遗传物质通过分子水平的进化从而产生功能变化

水平基因转移

<u>Horizontal gene transfer (HGT)</u> or <u>Lateral gene transfer (LGT)</u>: any process in which an organism transfers genetic material to another cell that is not its offspring.

- 首次报告HGT的发现 (Ochiai et al., 1959)；Syvanen (1984) 进一步确认HGT的存在，并指出**HGT "has biological significance, and is a process that shaped evolutionary history from the very beginning of life on earth."**
- HGT is common among bacteria, even very distantly-related ones.
- HGT has also occurred within eukaryotes, from their chloroplast and mitochondrial genome to their nuclear genome.



Brown, *Nat. Rev. Genet.*, **4**: 121.

In addition to putative HGT events that are associated with the origin of mitochondria and chloroplasts (green and purple arrows), HGT between other groups has also been observed, such as between spirochaetes and Archaea (arrow a), between low G+C GRAM-POSITIVE BACTERIA and Archaea (arrow b), and between thermophilic Bacteria and Archaea (arrow c).

HGT对基因组进化的作用

1. 对受体的影响

Positive:

- 通过HGT获得的基因给受体带来新的生物合成和降解能力
- 新基因编码的蛋白质所增加的某些动力学、生物化学和生物物理学特性，也使受体能更好地适应环境，提高对环境的抵抗力

Negative:

- 受体在最大可能获得新基因带来的好处的同时，也不得不对基因组的表达进行相应的调整
- 原核生物基因组大小比较稳定，在接受外来基因时要舍弃更多的序列
- 接受外来序列后，必须进行内源基因重组

Peking University
Introduction to Computational Biology

2. 对原核基因组操纵子结构的影响

HGT被认为是形成操纵子结构的一个重要途径

3. 对物种形成的影响

- HGT是基因组进化、生物多样化的重要途径
- HGT给物种进化史研究带来挑战和革命

基因复制、基因缺失及其它进化事件

基因复制（Gene Duplication）

Gene duplication (or **chromosomal duplication**) is any duplication of a region of DNA that contains a gene; it may occur as an error in homologous recombination, a retro-transposition event, or duplication of an entire chromosome.

●**Gene duplication is believed to play a major role in evolution**

●**基因复制的类型：**

－整个基因组的重复（The entire yeast genome underwent duplication about 100 million years ago. Plants are the most prolific genome duplicators.）

－一个染色体或染色体片段的重复

－一个基因或基因簇的重复

Peking University
Introduction to Computational Biology

基因缺失（Gene Loss）

基因缺失——改变进化的动力

指的是基因突变导致的基因功能的丧失

Am. J. Hum. Genet. 64:18–23, 1999

MOLECULAR EVOLUTION '99

When Less Is More: Gene Loss as an Engine of Evolutionary Change

Maynard V. Olson

Departments of Medicine (Division of Medical Genetics) and Genetics, University of Washington, Seattle

Evolutionary change results from differences in the reproductive success of individuals with different genotypes. The downside of this process is easy to grasp: selection constantly purges deleterious mutations from the gene pool. However, we know remarkably little about evolution's upside—that is, about the types of mutations that commonly lead to increased fitness. To understand the biology of natural populations—including, most notably, that of the human—we need testable ideas about the types of mutations that evolution is likely to have favored in the recent past. Here I explore one such idea, the proposal that loss of gene function may represent a common evolutionary response of populations undergoing a shift in environment and, consequently, a change in the pattern of selective pressures. I propose the testable view that gene loss is a major motif of molecular evolution.

Well-known human examples of conditionally advantageous mutations—those that improve fitness in particular environments—include a number of biallelic and multiallelic systems in which heterozygotes enjoy a conditional advantage. In these cases, alleles that are clearly maladaptive when present in homozygous form are nevertheless maintained at high frequency in some populations. For example, enteric disease and iron-deficient diets, respectively, have been proposed as selective pressures that may confer a heterozygote advantage on mutations that, when homozygous, cause cystic fibrosis and hemochromatosis (Gabriel et al. 1994; Crawford et al. 1995). Similarly, hemoglobinopathies are common in

其它进化事件

协同进化（Co-evolution）

The mutual evolutionary influence between two species. Each party in a co-evolutionary relationship exerts selective pressures on the other, thereby affecting each others' evolution.

趋同进化（Convergent evolution）

The process whereby organisms not closely related (not monophyletic), independently evolve similar traits as a result of having to adapt to similar environments or ecological niches

Peking University
Introduction to Computational Biology

Less is more:

（Olsen, 1999）

Loss of gene function may represent a common evolutionary response of populations undergoing a shift in environment and, consequently, a change in the pattern of selective pressures. Adaptive loss of function may occur frequently and may spread rapidly through small populations.

例子：

趋化因子受体基因的缺失可以增加人体细胞抵抗AIDS和疟原虫感染的能力，因为该基因有助于病原体侵入靶细胞。编码表面蛋白酶的*ompT*基因的缺失可以降低志贺氏菌的感染概率。