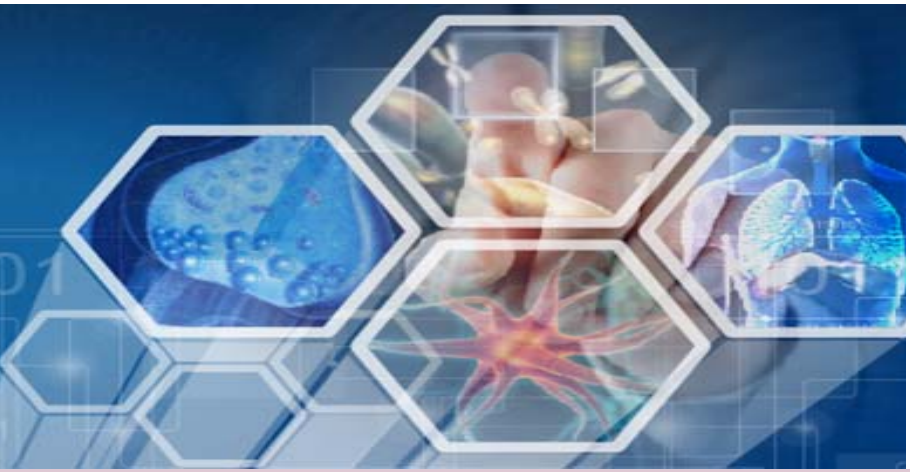
生物信息学



分子进化分析

第一节　引言

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

Tree of Life

- 重建所有生物的进化历史并以系统树的形式加以描述

生物进化理论

- 达尔文进化论：
  - 进化：变异的遗传
  - 自然选择：解释为何演变发生的机制

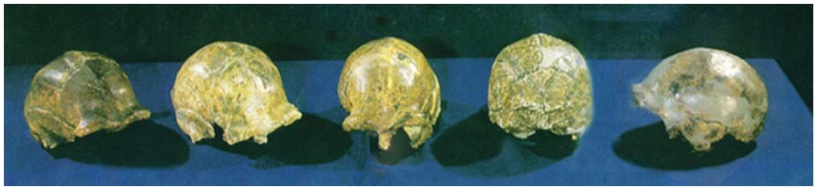
种群中个体变异的遗传学基础：孟德尔遗传

  - 孟德尔豌豆实验：杂交的表现特征是基因表达的结果，而不是基因杂交遗传

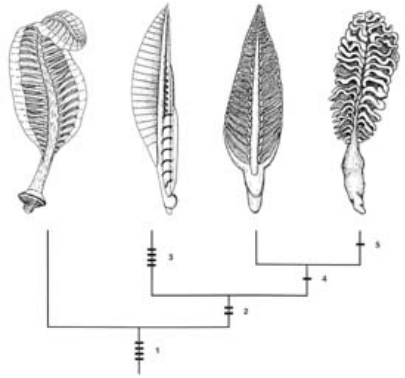
- 中性进化论：并非所有种群中保留下来的突变都由自然选择所形成；大多数突变是中性或接近中性，不妨碍种群的生存与繁衍。

研究生物进化历史的途径

- **1.** 最确凿证据是：生物化石！—— 零散、不完整



- **2.**比较形态学、比较解剖学和生理学等：确定大致的进化框架 —— 细节存很多的争议



分子进化

- **1964**年，**Linus Pauling**提出分子进化理论；
- 从物种的一些分子特性出发，从而了解物种之间的生物系统发生的关系。
- 发生在分子层面的进化过程：**DNA, RNA**和蛋白质分子
- 基本假设：核苷酸和氨基酸序列中含有生物进化历史的全部信息。

- 分子进化开始于20世纪60年代，近20年来由于分子遗传学资料的迅速积累，成为计算生物学和生物信息学等新兴学科的重要组成部分。
- 尤其人类基因组测序后，推动了分子进化的进一步发展，序列保守性，基因表达和网络的进化等研究内容不断的出现在最新的研究中，充实了生物信息学的研究范围。

分子进化的模式

- **DNA突变的模式：替代，插入，缺失，倒位；**
- **核苷酸替代：转换 (Transition) & 颠换 (Transversion)**
- **基因复制：多基因家族的产生以及伪基因的产生**
  - A. 单个基因复制 - 重组或者逆转录
  - B. 染色体片断复制
  - C. 基因组复制

DNA突变的模式

1. Substitution. 替代

Thr Tyr Leu Leu
ACC TAT TTG CTG
↓
ACC TCT TTG CTG
Thr Tyr Leu Leu

3. Insertion. 插入

Thr Tyr Leu Leu
ACC TAT TTG CTG
↓
ACC TAC TTT GCT G—
Thr Tyr Phe Ala

2. Deletion. 缺失

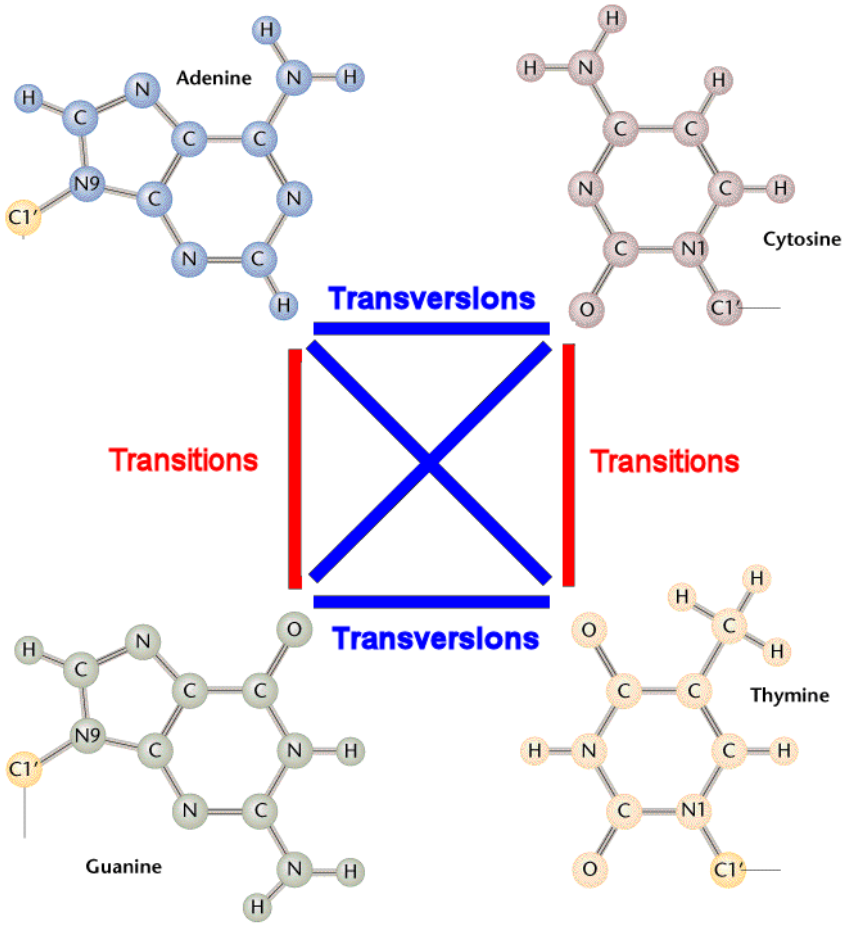
Thr Tyr Leu Leu
ACC TAT TTG CTG
↓
ACC TAT TGC TG-
Thr Tyr Cys

4. Inversion. 倒位

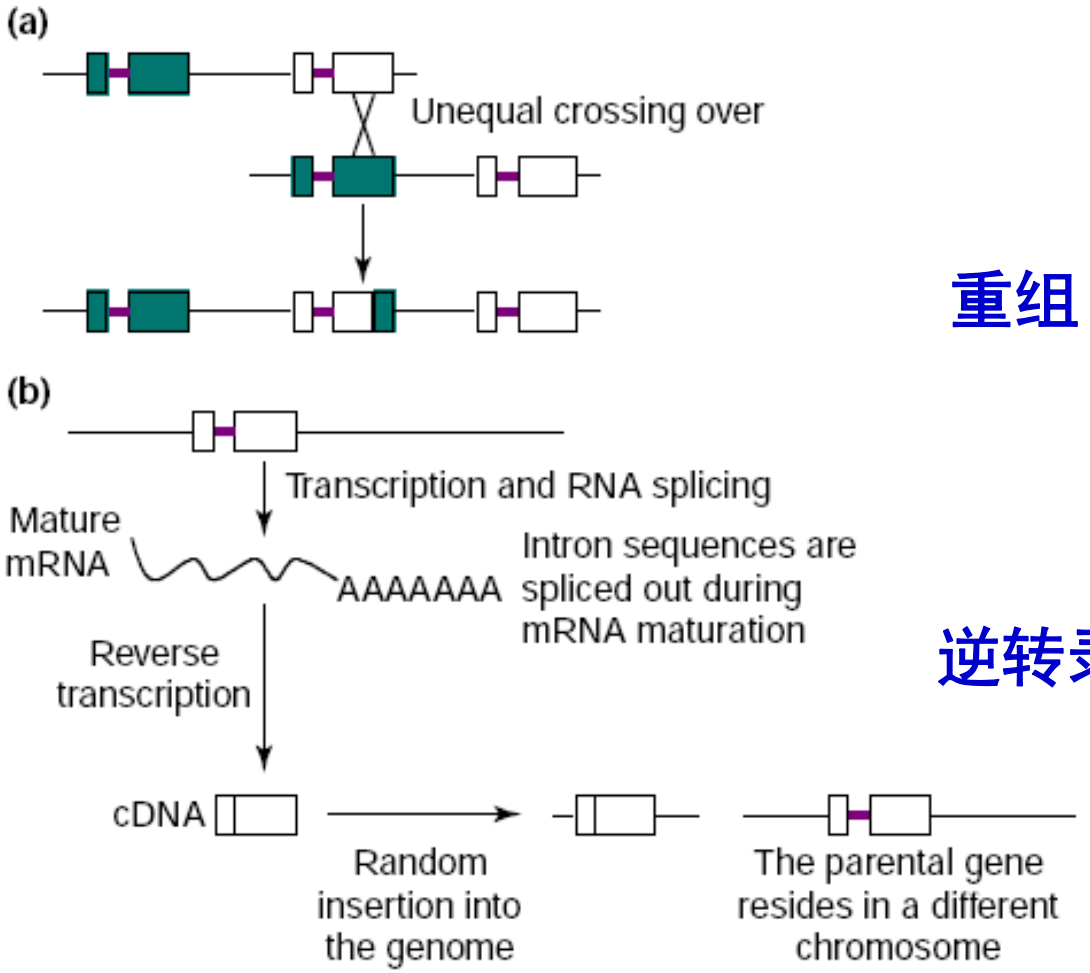
Thr Tyr Leu Leu
ACC TAT TTG CTG
↓
ACC TTT ATG CTG
Thr Phe Met Leu

核苷酸替代：转换 & 颠换

- 转换：嘌呤被嘌呤替代，或者嘧啶被嘧啶替代
- 颠换：嘌呤被嘧啶替代，或者嘧啶被嘌呤替代



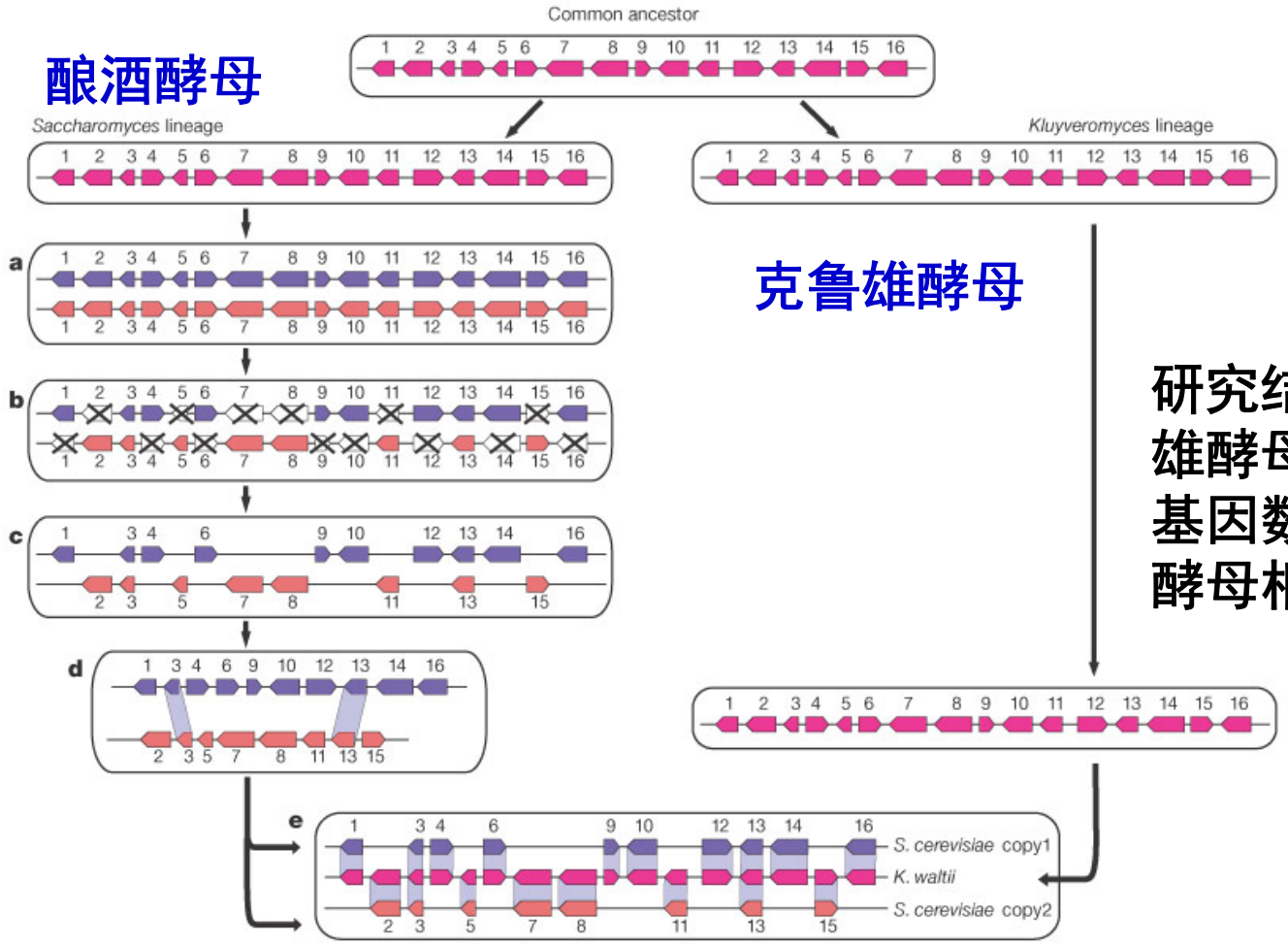
基因复制：单个基因复制



重组

逆转录

基因复制：基因组复制



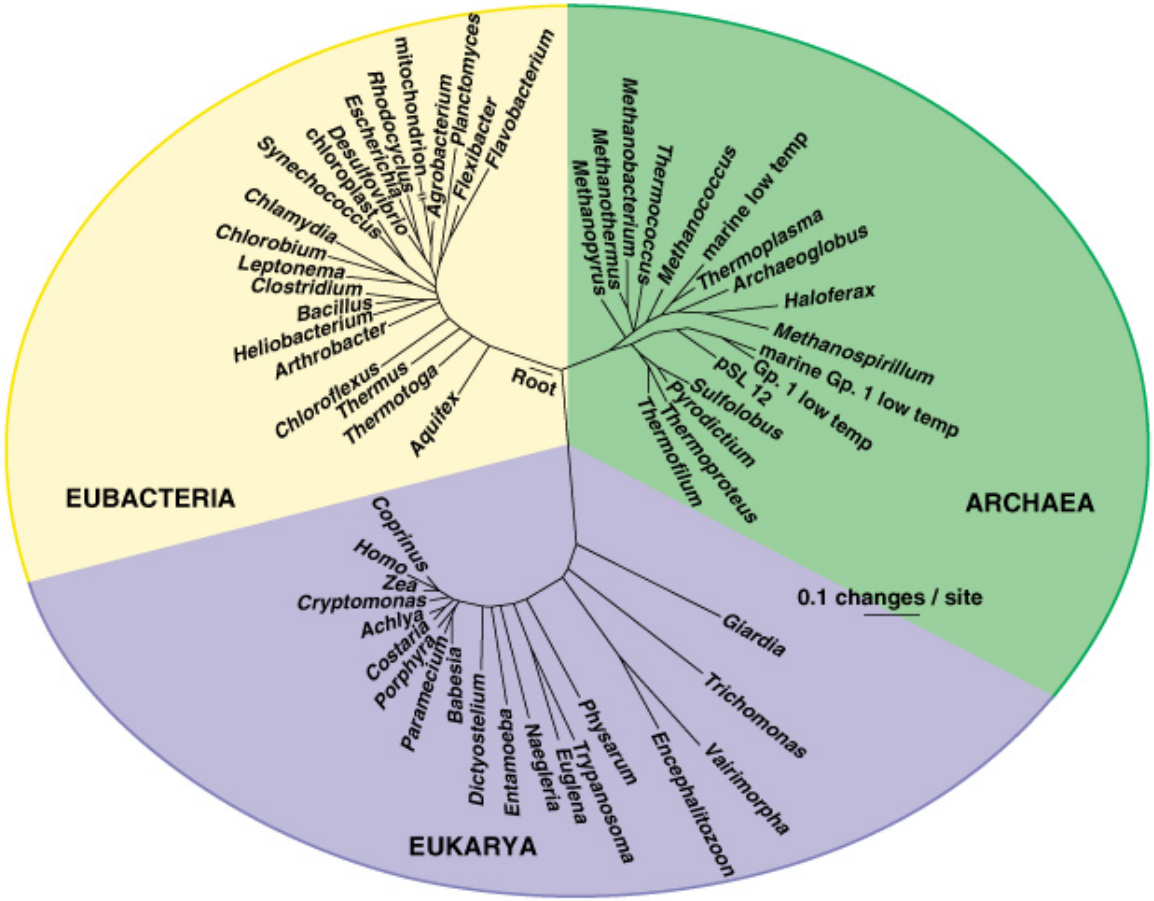
研究结果：克鲁雄酵母中的同源基因数量与酿酒酵母相比为1：2

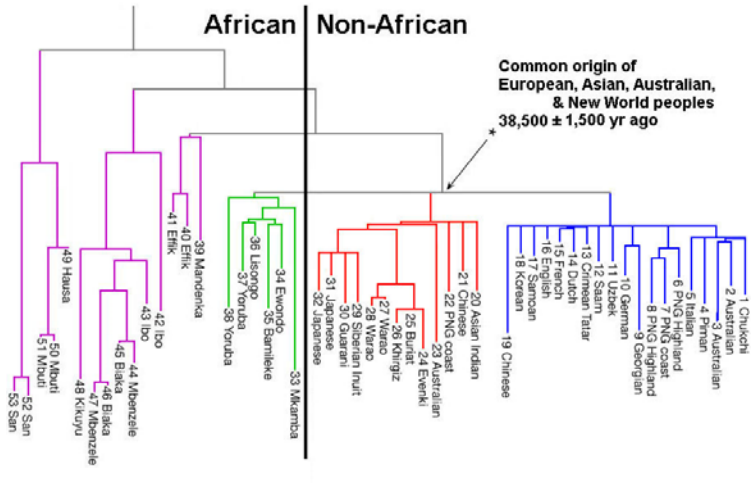
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

分子进化研究的目的

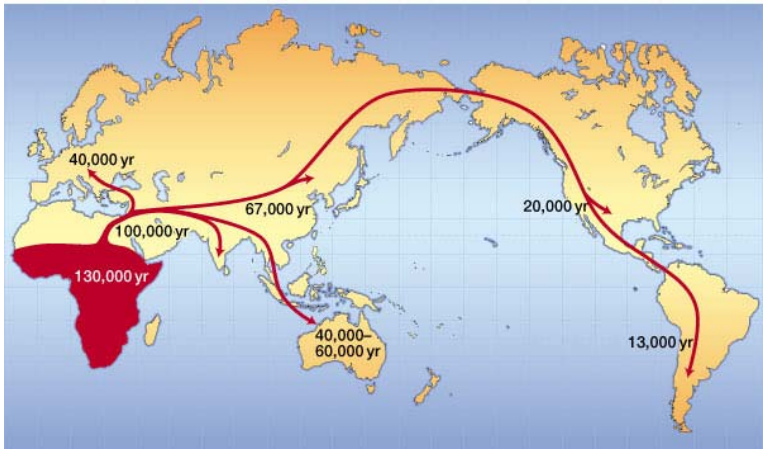
- 物种分类及关系：从物种的一些分子特性出发，构建系统发育树，进而了解物种之间的生物系统发生的关系 —— **tree of life**
- 大分子功能与结构的分析：同一家族的大分子，具有相似的三级结构及生化功能，通过序列同源性分析，构建系统发育树，进行相关分析；功能预测
- 进化速率分析：例如，**HIV**的高突变性；哪些位点易发生突变？

Tree of Life: 16S rRNA





53个人的线粒体基因组(16,587bp)



人类迁移的路线

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

第二节
系统发生分析与重建

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

一、核苷酸置换模型及氨基酸置换模型

（一）DNA序列进化分析

- **DNA序列的进化演变比蛋白质序列的演变更复杂，因为有多种多样的DNA区域，如蛋白质编码区、非编码区、外显子、内含子、侧翼区、重复DNA序列和插入序列等。因此，弄清所研究的DNA类型和功能是十分重要的。即便我们单独考虑蛋白质编码区，密码子第一、二和三位的核苷酸替代式样也不尽相同。何况，某些区比其他区更易受到自然选择的影响，使得DNA的不同区域呈现不同的进化模式。**

基因的编码区和非编码区

- 基因的DNA由编码区（Coding region）和非编码区（Non-coding region）构成；
- 编码区可以转录信使RNA，进而调控蛋白质的合成；
- 非编码区不能转录成信使RNA，但是它可以调控遗传信息的表达；
- 原核基因：编码区全部编码蛋白质；

  真核基因：编码区分为外显子和内含子,只有外显子能编码蛋白质；

分子进化选择压力

- 进化选择压力：
  - A. 编码区：阳性选择 1%；阴性选择19%；中性进化80%；
  - B. 非编码区：~100%的中性进化
- 中性进化：
  - 同义突变， 约占核苷酸置换总数的四分之一；
  - 非编码区DNA序列的突变对蛋白质的合成很少有影响。

1.两个序列间的核苷酸差异

- 对于一种同源的核酸分子来说，它在亲缘关系越近的生物之间差异就越小，相反差异 就越大，即两同源分子分歧的时间与它们之间的序列差异成正比。
- 同一条祖先序列传衍的两条后裔序列，它们的核苷酸差异随时间而增加。一个简便的描述序列分歧大小的测度是两条后裔序列中不同核苷酸位点的比例。

$$\hat{p} = n_d / n$$

- 以下，我们称此估计为**p**距离。

（二）氨基酸序列进化分析

1.氨基酸差异和不同氨基酸的比例

- 蛋白质或肽链的进化演变研究开始于两个或多个氨基酸序列的比较。图**4-1**显示了人、马、牛、袋鼠、蝾螈和鲤鱼的血红蛋白α链的氨基酸序列。图中，不同的氨基酸分别用不同的单字母代表。

图4-1　六种脊椎动物血红蛋白α链的氨基酸序列

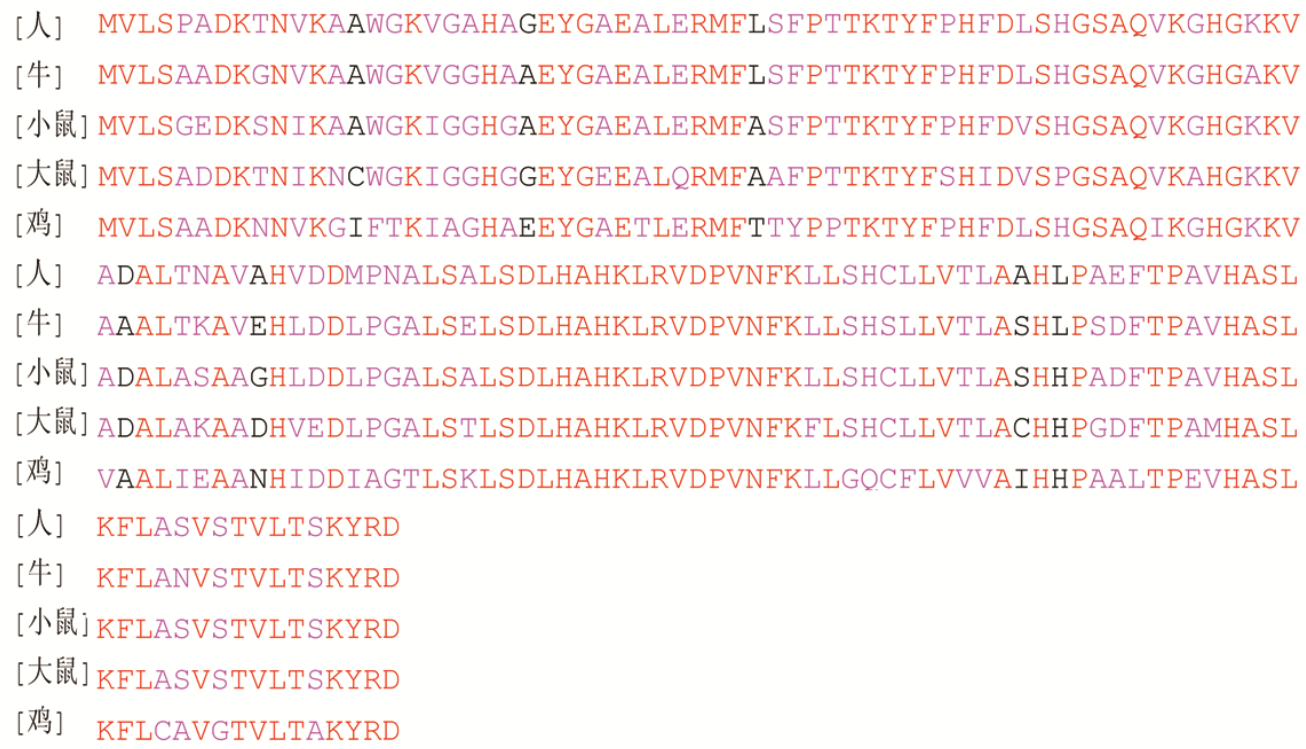


表4-4以及不同氨基酸的比例（下对角线）

不同脊椎动物血红蛋白α链中不同氨基酸的数目（上对角线）

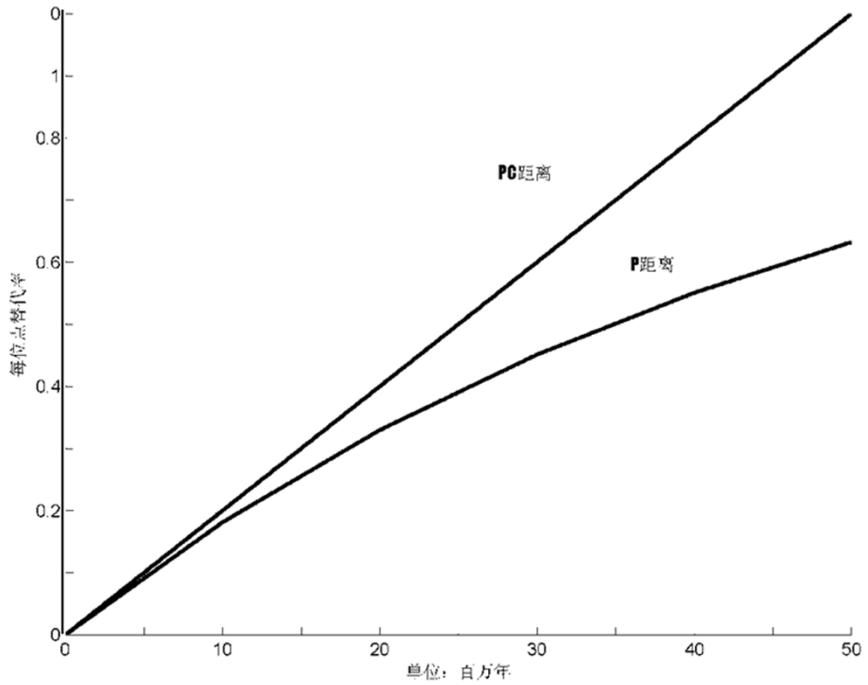
	人	马	牛	袋鼠	蝾螈	鲤鱼
人		17	17	26	61	68
马	0.121		17	29	66	67
牛	0.121	0.121		25	63	65
袋鼠	0.186	0.207	0.179		66	71
蝾螈	0.436	0.471	0.450	0.471		74
鲤鱼	0.486	0.479	0.464	0.507	0.529	

注：计算排除了缺失和插入，使用的氨基酸总数为**140**。

- 在图中所给出的例子中，删除所有间隔后可比较的总氨基酸位点数为**140**。因此，在此例中。值出现在表中对角线上部，可以很容易地计算出，列于对角线下部。

- 当所比较的物种亲缘关系很远时（如人和鲤鱼），值较大，而当亲缘关系较近的物种比较时（如人和马），值较小。这说明随着两个物种的分歧时间增大，氨基酸的替代数也将增大，但并不严格与分歧时间成比例。

图4-2 *p*距离和泊松校正(PC)距离随分歧时间变化的关系



2. 泊松校正（PC） 距离

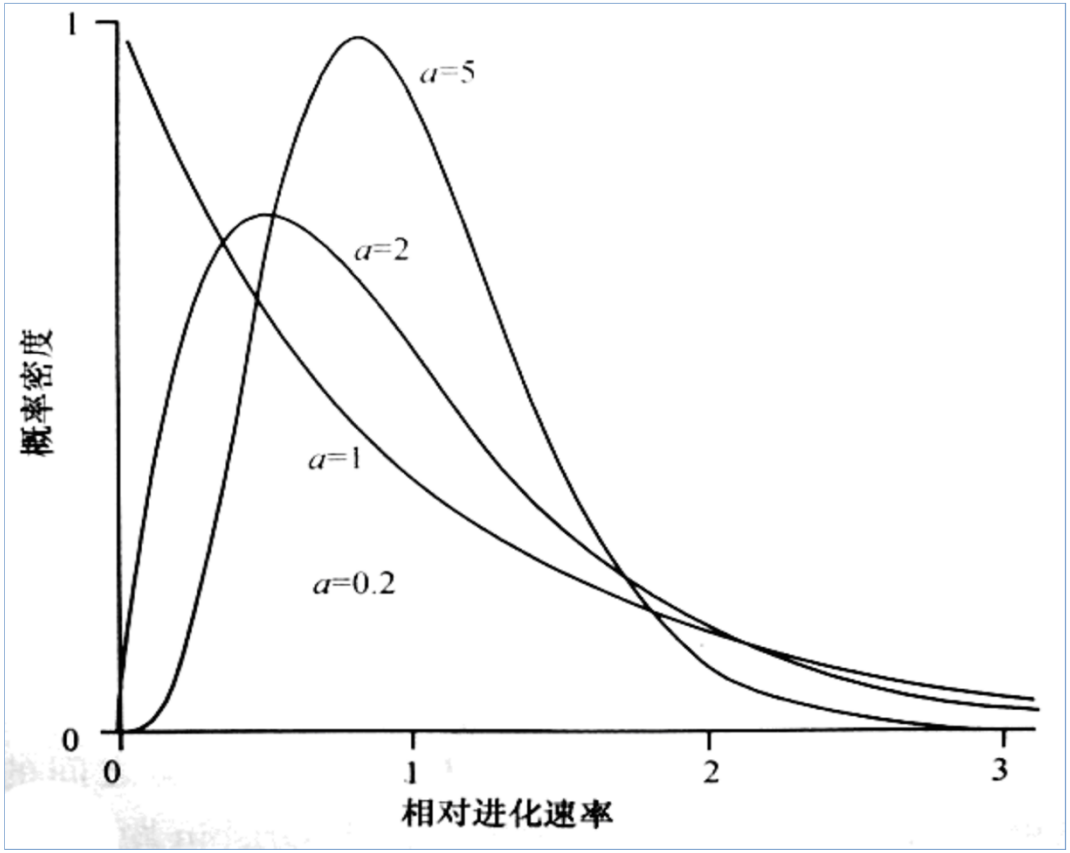
- ***p***与***t***的变化呈现非线性关系的原因之一是当多个氨基酸替代出现在同一位点时，n_d偏离实际氨基酸的替代数将会逐渐增加。更精确估计替代数的方法之一是运用泊松分布的概念。令***r***为一个特定位点每年的氨基酸替换率，并且为简便起见假设所有位点的***r***都相同，在时间**t**年后，每个位点氨基酸替代的平均数是***rt***。在一个给定位点氨基酸替代数***k***（***k*=0, 1, 2, 3, …**）的发生频率遵循泊松分布，即，

$$P(k;t) = e^{-rt}(rt)^k / k!$$

- 若已知每个位点的氨基酸替代率按分布的话，每个位点氨基酸替代的观察值将按负二项式分布。因此，**Uzzell**和**Corbin**研究建议，不同位点的替代率都按分布估计，即

$$f(r)=[b^{a}/\Gamma(a)]e^{-br}r^{a-1}$$

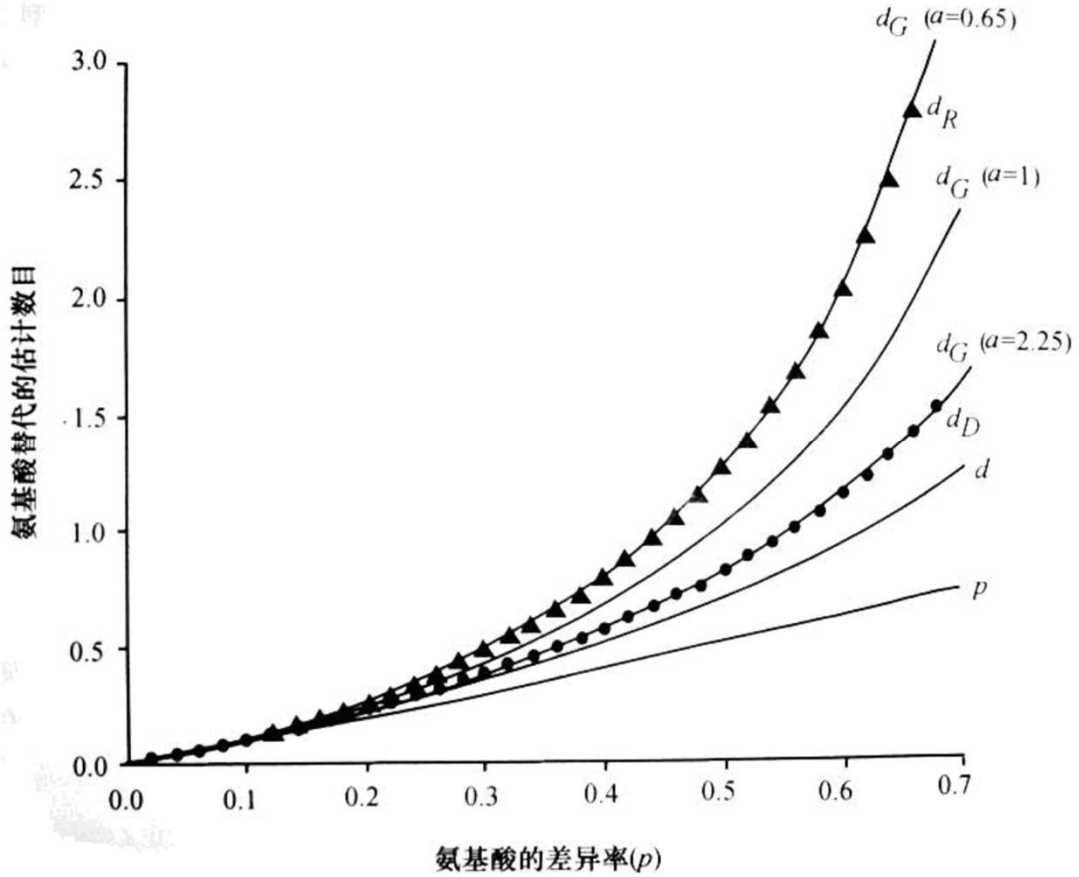
- ***f*** **(*r*)**的分布形状由***a***决定，***a***常称为形状参数或参数，而***b***则称为尺度因子。分布是非常柔性的，有多种多样形状，由形状参数***a***决定。


1
0
概率密度
a=5
a=2
a=1
a=0.2
0
1
2
3
相对进化速率

- 当***r***遵循分布时，就有可能估计出平均每个位点的氨基酸替代数。为此，让我们考虑在时间***t***时两个序列间某一位点上的氨基酸相同的概率，按公式（**4.4**）计算。然后，对所有位点的**q**求均值，为

$$\overline{q} = \int_0^{\infty} qf(r)\mathrm{d}r = \left[\frac{a}{a + 2\overline{r}t}\right]^a$$

$$d_G = a[(1-p)^{-1/a} - 1]$$


d_G (a=0.65)
d_R
d_G (a=1)
d_G (a=2.25)
d_D
d
p
氨基酸替代的估计数目
氨基酸的差异率(p)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

[例] 血红蛋白链的进化距离和氨基酸替代率的估计

- 表4-5表示出了6种脊椎动物血红蛋白链成对比较的有差异氨基酸的数目的比例（$\hat{p}$）。我们用这些值来估计PC距离（d_G）和 距离（Γ）。

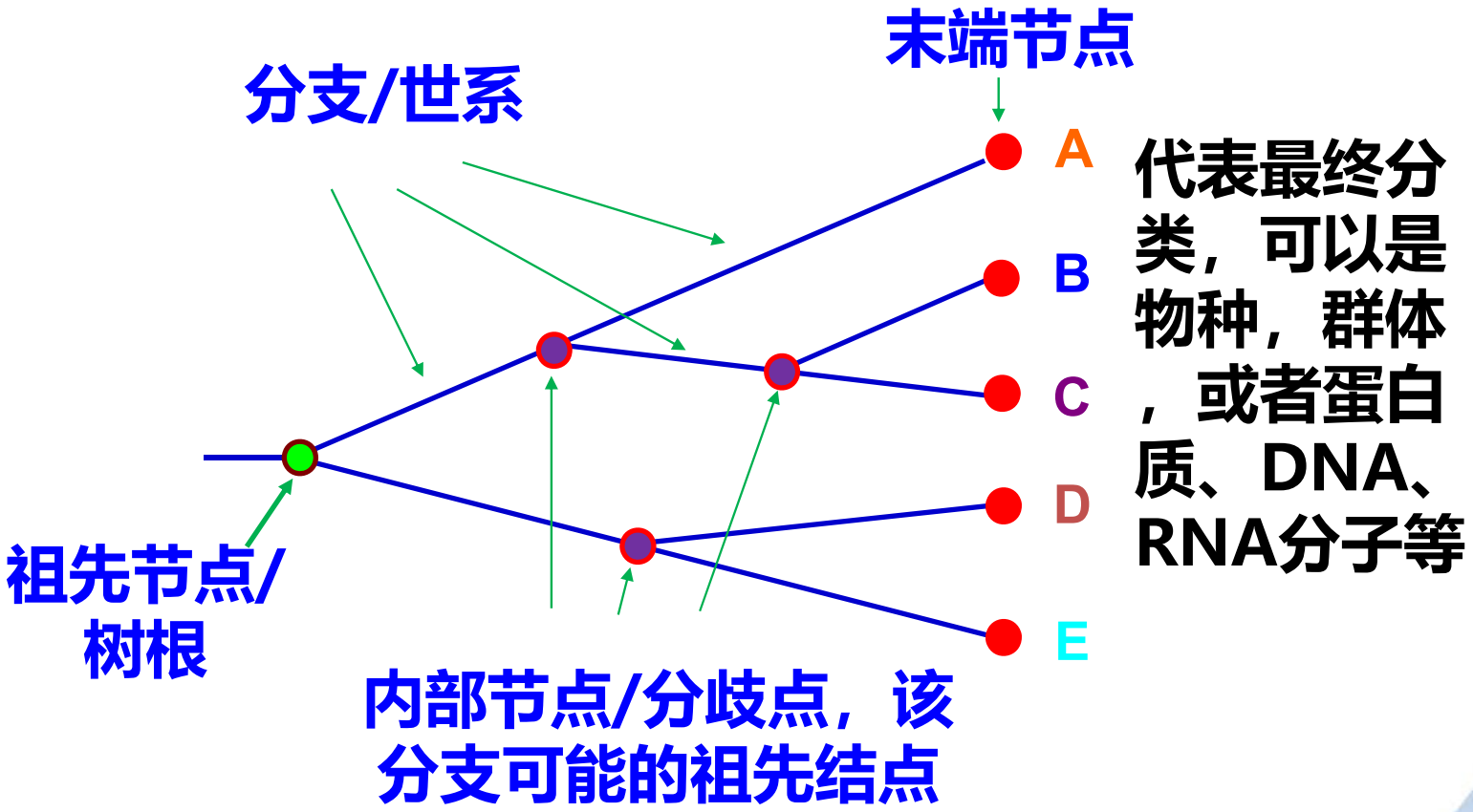
表4-5 解析法估算的PC距离的标准误（下对角阵）及自展法估算的PC距离的标准误（上对角阵）

	人	马	牛	袋鼠	蝾螈	鲤鱼
人		0.129	0.129	0.205	0.572	0.665
马	0.134		0.129	0.232	0.638	0.651
牛	0.134	0.134		0.197	0.598	0.624
袋鼠	0.216	0.246	0.207		0.638	0.708
蝾螈	0.662	0.751	0.697	0.751		0.752
鲤鱼	0.789	0.770	0.733	0.849	0.913	

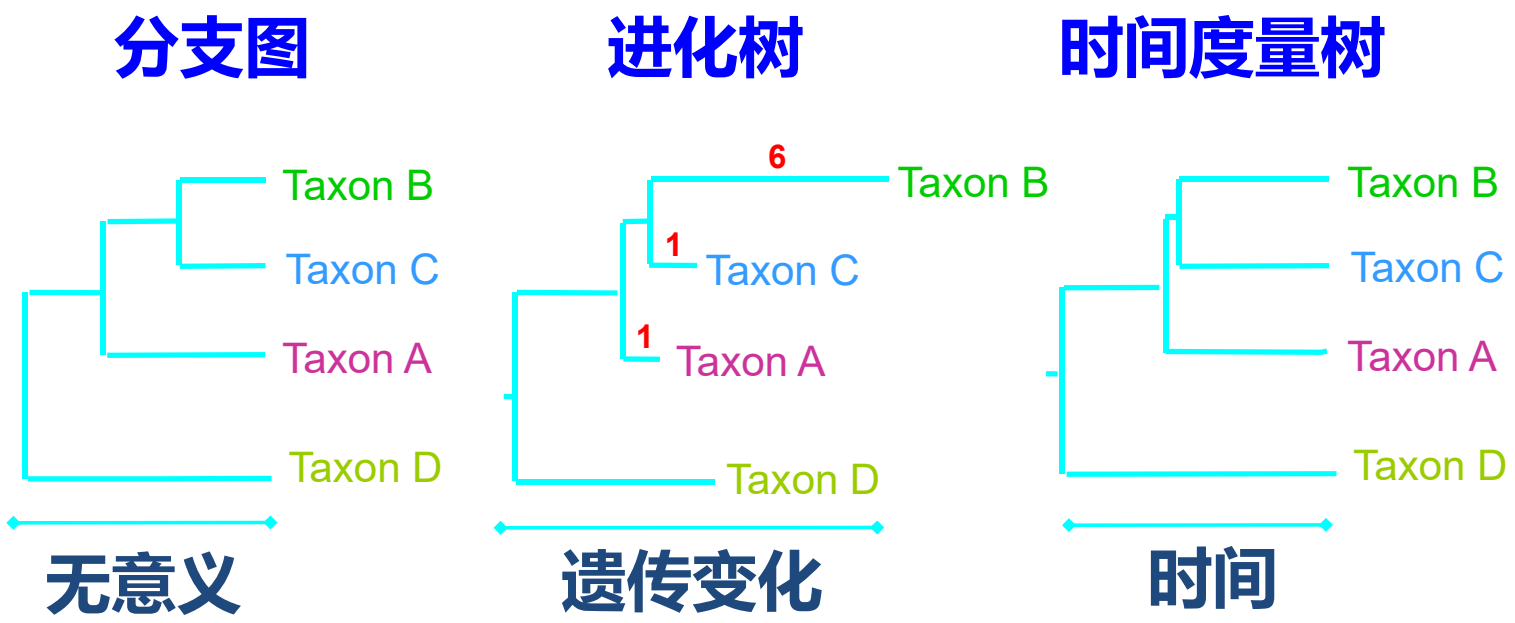
系统发生树的基本概念及搜索方法

- 在研究从病毒到人类的各种生物的进化历史中，DNA或蛋白质序列的系统发育分析已经成为一个重要的工具。
- 由于不同的基因或DNA片段的进化速率存在较大的差异，我们可以通过这些基因或DNA片段来估计几乎所有水平上的有机体间的进化关系（例如，界、门、科、属、种以及种内群体）。

系统发育树: 术语

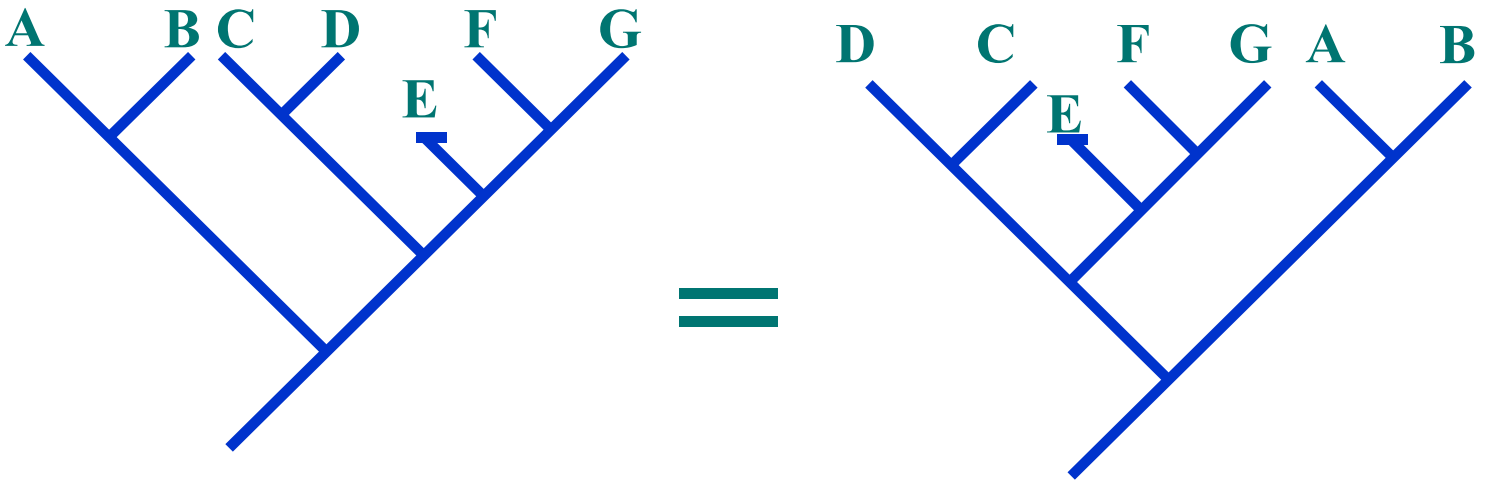

末端节点
分支/世系
A
B
C
D
E
代表最终分类，可以是物种，群体，或者蛋白质、DNA、RNA分子等
祖先节点/树根
内部节点/分歧点，该分支可能的祖先结点

系统发育树：三种类型



以上三种类型的系统发育树表示相同的分支状况，相同的进化关系

树只代表分支的拓扑结构

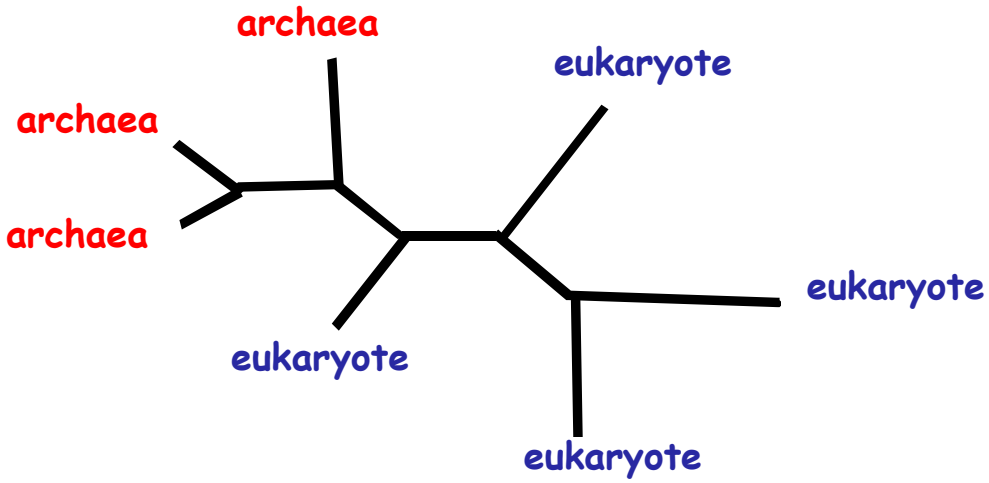


（一）系统发育树的种类

1.有根树和无根树

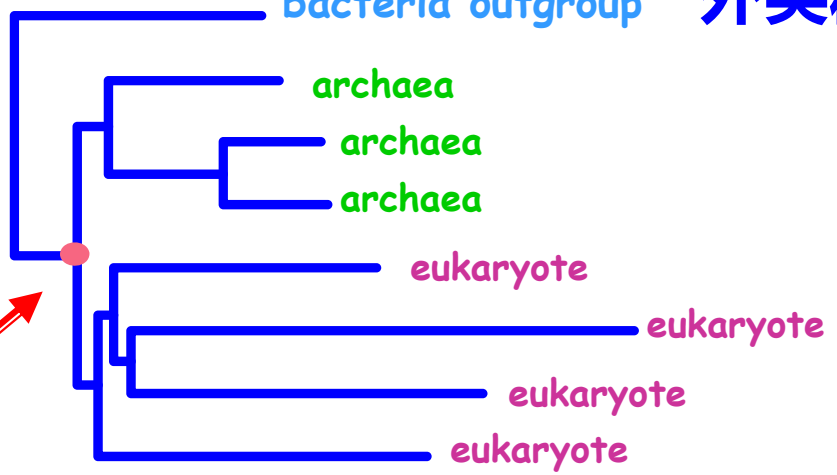
- 基因或生物体的系统发育关系常常用有根或无根的树形结构来表示，即有根树和无根树。

无根树，有根树，外类群



无根树

通过外类群
来确定树根

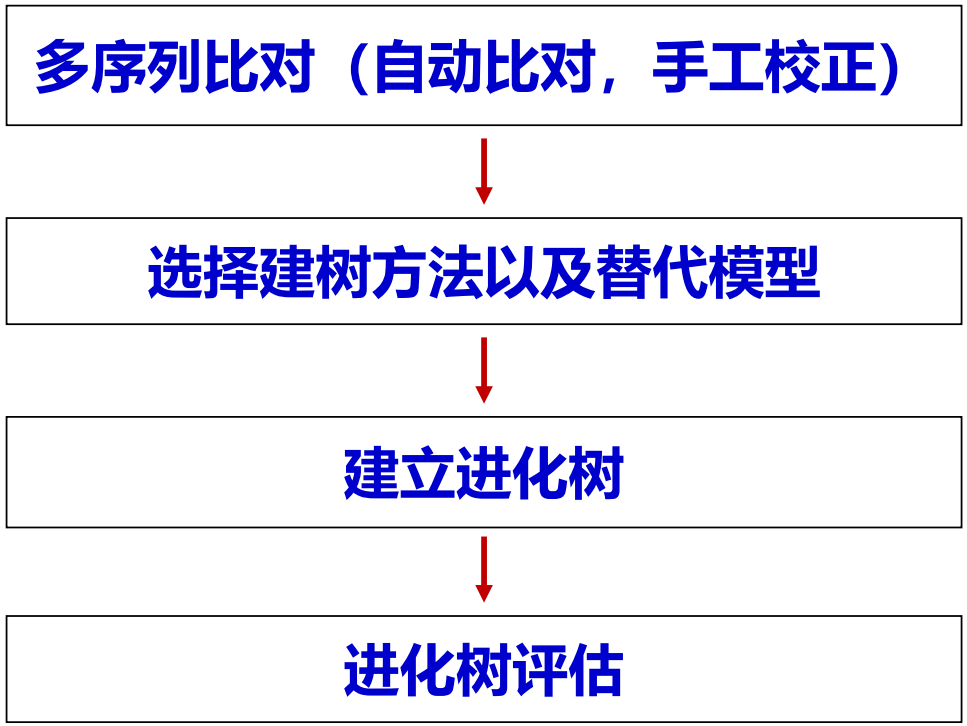


外类群

有根树

根

系统发育树重建分析步骤



（二）基于距离法构建系统发生树

- 通常使用的方法分为3大类：
- **1. 最大简约法 (maximum parsimony, MP)**
  - 适用序列有很高相似性时
- **2. 距离法 (distance)**
  - 适用序列有较高相似性时
- **3. 最大似然法 (maximum likelihood, ML)**
  - 可用于任何相关序列集合
- 计算速度：
  - 距离法 >最大简约法 >最大似然法

2. 距离法

- 又称距离矩阵法，首先通过各个物种之间的比较，根据一定的假设（进化距离模型）推导得出分类群之间的进化距离，构建一个进化距离矩阵。再依据进化距离，分别依次将序列合并聚类，构建进化树。

简单的距离矩阵

A. Sequences

sequence A	ACGCGTTGGGCGATGGCAAC
sequence B	ACGCGTTGGGCGACGGTAAT
sequence C	ACGCATTGAATGATGATAAT
sequence D	ACACATTGAGTGATAATAAT

B. Distances between sequences, the number of steps required to change one sequence into the other.

n_{AB}	3
n_{AC}	7
n_{AD}	8
n_{BC}	6
n_{BD}	7
n_{CD}	3

C. Distance table

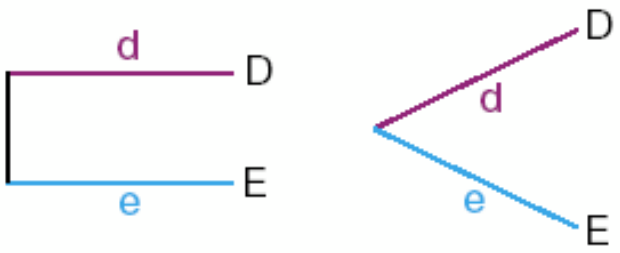
	A	B	C	D
A	–	3	7	8
B	–	–	6	7
C	–	–	–	3
D	–	–	–	–

通过距离矩阵建树的方法

- 由进化距离构建进化树的方法有很多，常见有：
  - (1) Fitch-Margoliash Method (FM法): 对短支长非常有效
  - (2) Neighbor-Joining Method (NJ法/邻接法):求最短支长，最通用的距离方法
  - (3) Neighbors Relaton Method(邻居关系法)
  - (4) Unweighted Pair Group Method with Arithmetic Mean (UPGMA，非加权组平均法)

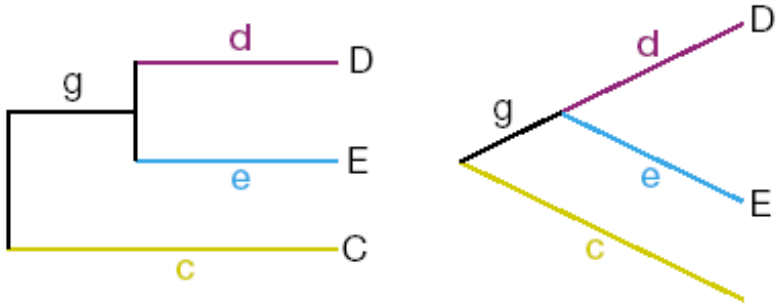
UPGMA法

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—



d=e=10/2=5

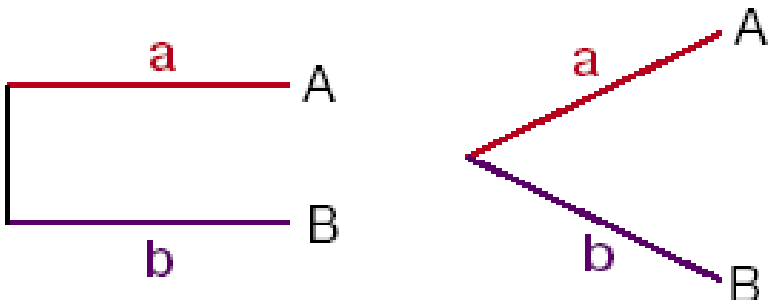
	A	B	C	(DE)
A	—	22	39	40
B	—	—	41	42
C	—	—	—	19
(DE)	—	—	—	—



c=19/2=9.5

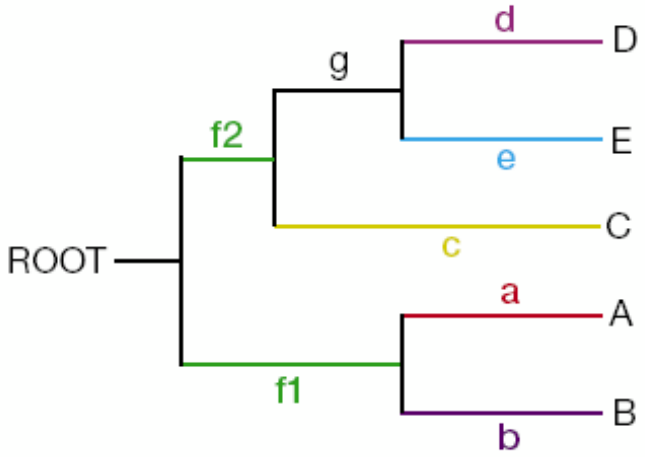
g=c-d=9.5-5=4.5

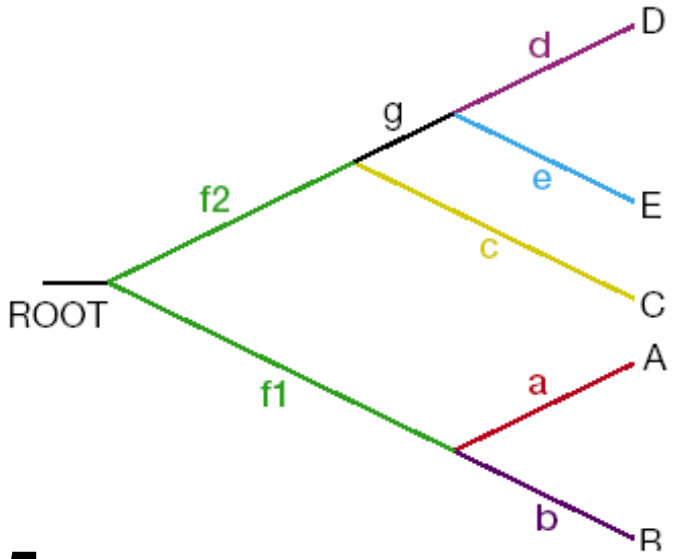
	A	B	(CDE)
A	-	22	39.5
B	-	-	41.5
(CDE)	-	-	-



a=b=22/2=11

	(AB)	(CDE)
(AB)	-	40.5
(CDE)	-	-





f1+a=f2+c=40.5/2=20.25

f1=9.25 , f2=11.75

最小二乘法

- 最小二乘法（**LS**）将成对距离矩阵作为给定数据，通过匹配那些尽可能近的距离来估计一棵树上的枝长。

- 设物种i和j之间的距离为d_{ij}，树上物种i到j间通路的枝长和为d_{ij}。**LS**方法对所有独立的i和j对求距离差的平方$(d_{ij}-\hat{d}_{ij})^2$的最小值，使得这棵树与距离之间的拟合尽可能地近。

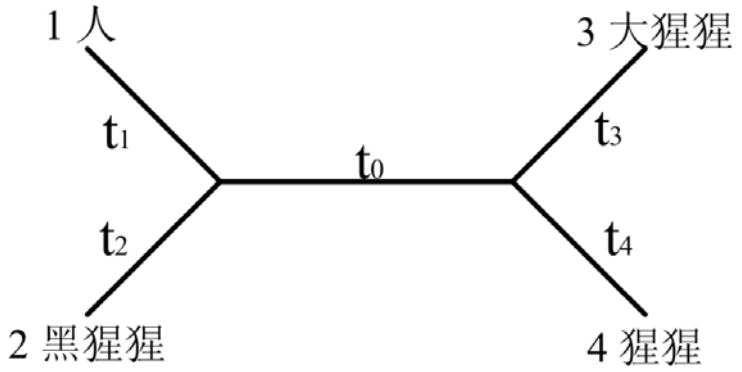


表4-7 线粒体 DNA 序列的成对距离

	1.人	2.黑猩猩	3.大猩猩	4.猩猩
1.人				
2.黑猩猩	0.0965			
3.大猩猩	0.1140	0.1180		
4.猩猩	0.1849	0.2009	0.1947	

$$S = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

$$= (d_{12} - \hat{d}_{12})^2 + (d_{13} - \hat{d}_{13})^2 + (d_{14} - \hat{d}_{14})^2 + (d_{23} - \hat{d}_{23})^2$$

$$+ (d_{24} - \hat{d}_{24})^2 + (d_{34} - \hat{d}_{34})^2$$

K80 模型（Kimura，1980）下的最小二乘法

树	t_0	t_1	t_2	t_3	t_4	S_j
$\tau:((H,C),G,O)$	0.008840	0.043266	0.05328	0.058908	0.135795	0.000035
$\tau:((H,C),C,O)$	0.000000	0.046212	0.05623	0.061854	0.138742	0.000140
$\tau:((H,C),C,O)$	同上					
$\tau:(H,G,C,O)$	同上					

三、分子钟假说

（一）概述

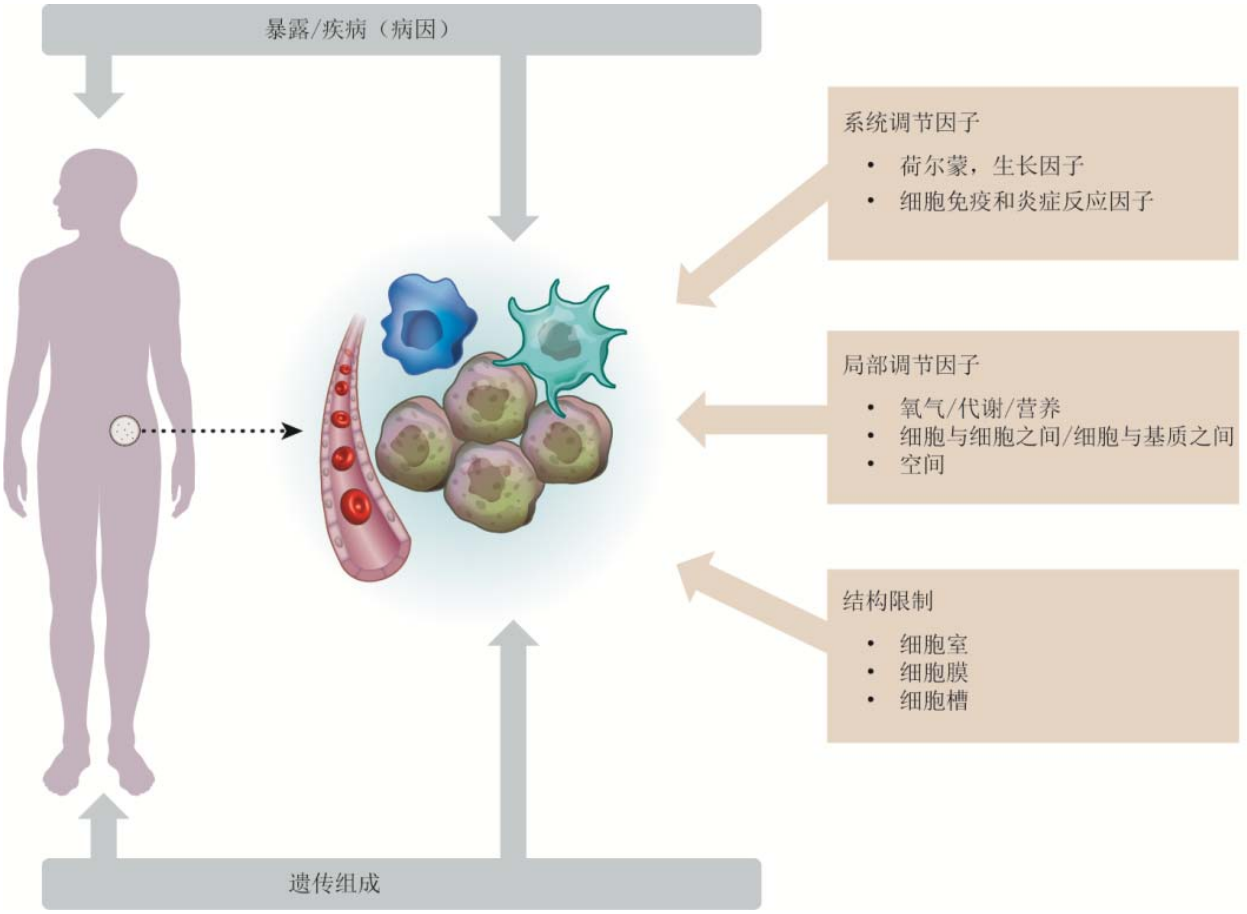
- 分子钟（**molecular clock**）假说认为DNA或蛋白质序列的进化速率随时间或进化谱系保持恒定。
- 化石数据是被用来校定分子钟的，即将序列间的距离转换成绝对地质时间和置换率。

（二）相对速率检验

- 最简单的分子钟假设检验是采用第三个物种C（外类群）来检验两个物种A和B是否以相同的速率进化。这一检验称为相对速率检验（relative-rate test），其实几乎所有的分子钟检验比较的都是相对速率而不是绝对速率。
- 确定灵长类分歧时间。

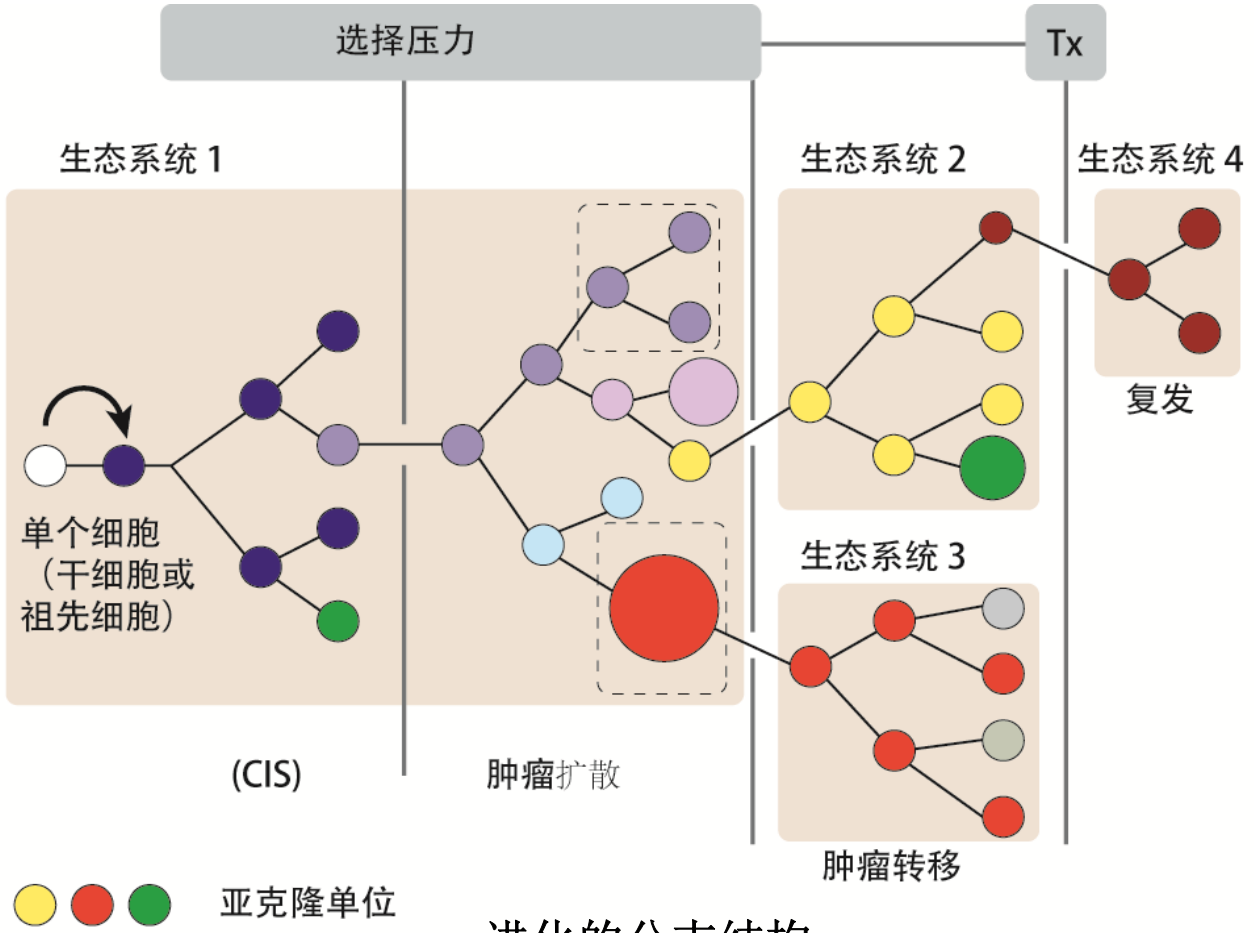
四、肿瘤细胞微进化

- 癌症组织的生态环境提供了适应性进化所需要的条件。组织内的微环境是具有多种组分的复杂的、动态的环境，这些可以影响癌症克隆的进化。例如，转化生长因子-β是癌症环境里的调控分子，其他的如炎症细胞组分也是癌症细胞生态环境的调控子。



组织的生态环境的复杂性

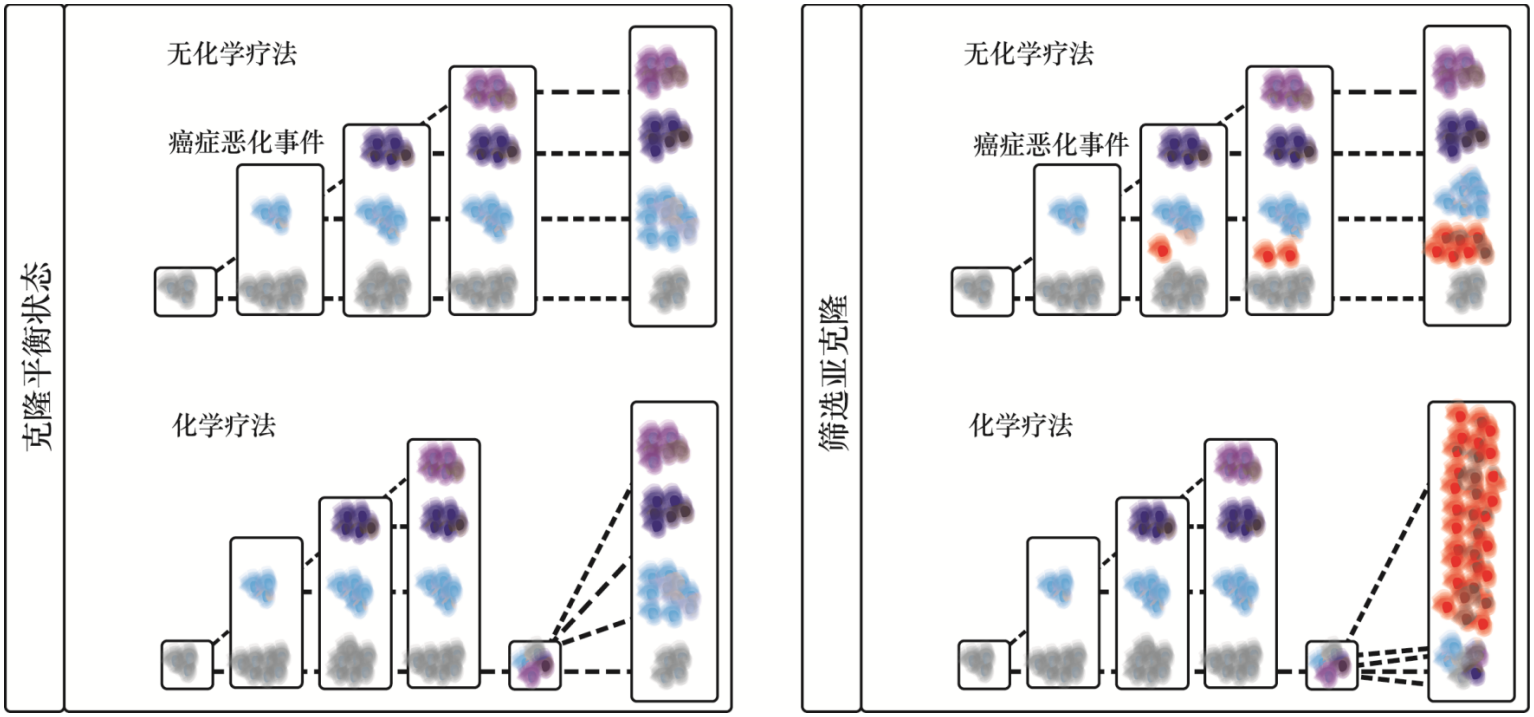
- 克隆进化的经典模型认为伴随着一系列的连续突变，一些亚克隆在群体中会占据优势或选择性清除。疾病进展的病理学证据（腺瘤，癌和转移）支持这一模型。从单细胞测序分析的数据表明，进化的轨迹是复杂的和分支的，就像诺埃尔提出的与达尔文的进化形态相似的物种发育树。



进化的分支结构

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

➢ 应用实例：慢性淋巴细胞白血病突变进化研究



肿瘤细胞的克隆进化

- 近年来，由于序列数据的快速积累，分子进化领域也经历了爆炸性的增长，计算机软件和硬件的能力逐年提高，精细的统计方法也是逐渐攀升。基因组的大规模数据也需要更强的统计方法去分析和解释，这无论是在概念上还是计算上都非常具有挑战性。本章中既有经典的分子进化统计方法，也涉及到了最近前沿的进展。与此同时，在生物信息学发展的带动下，分子进化与生物信息的结合领域也迅速出现。

MOOC内容

5月9日学习

- 第四章 分子进化