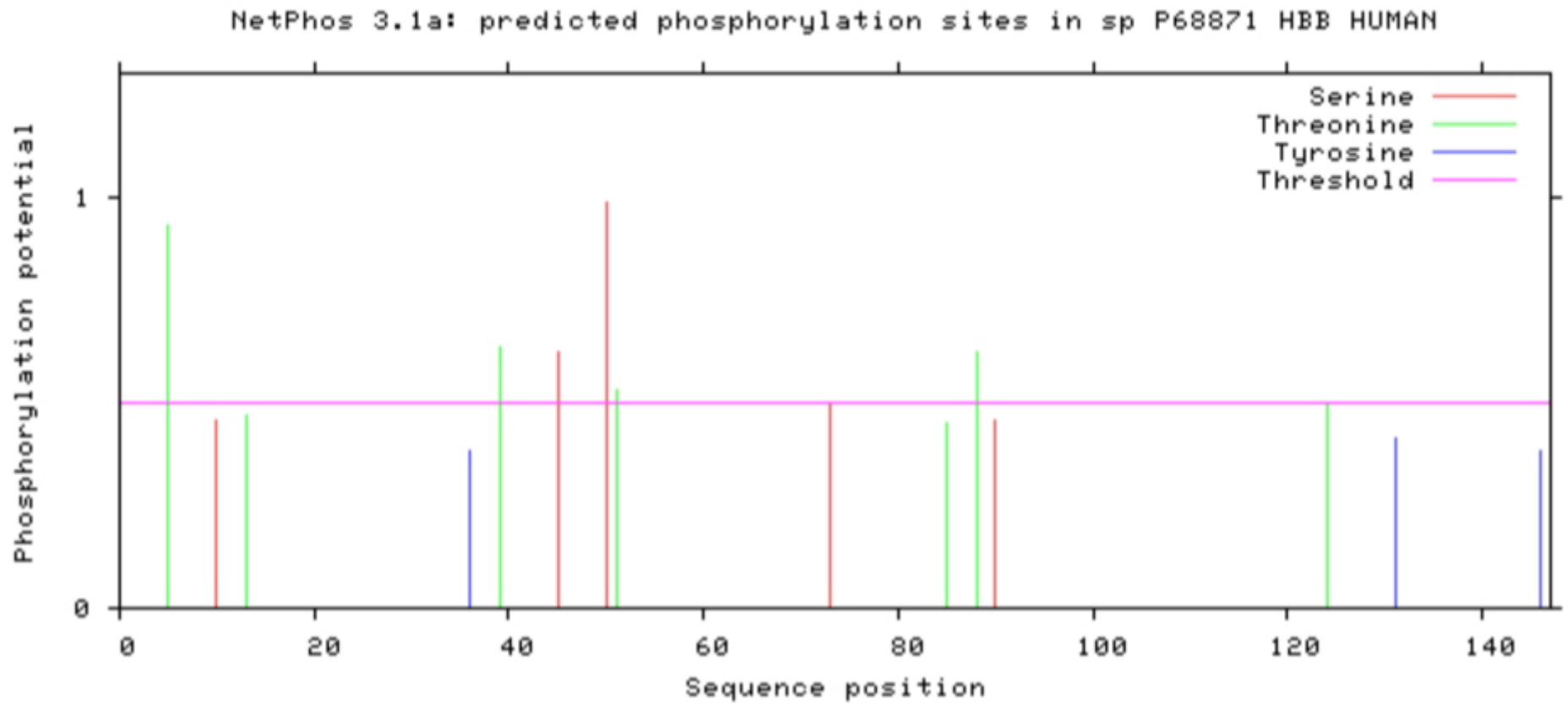
磷酸化位点，9个YES，6个线上

```
#
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.620    unsp        YES
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.571    CKI         YES
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.462    GSK3         .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.451    CaM-II       .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.377    CKII         .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.364    RSK          .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.359    PKA          .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.353    DNAPK        .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.351    cdc2         .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.330    ATM          .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.320    PKG          .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.278    p38MAPK      .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.204    cdk5         .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.169    PKC          .
# sp_P68871_HBB_HUMAN     45 S    RFFESFGDL    0.104    PKB          .
#
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.987    unsp        YES
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.463    GSK3         .
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.454    CKII         .
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.413    CaM-II       .
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.403    cdc2         .
# sp_P68871_HBB_HUMAN     50 S    FGDLSTPDA    0.384    DNAPK        .
```

磷酸化位点，9个YES，6个线上

```
# sp_P68871_HBB_HUMAN     146 Y    LAHKYH---    0.058    unsp          .
#
     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS    #      50
     TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    #     100
     PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH       #     150
61   ....T.....................................T.....S....S    #      50
61   T.....................................T.............    #     100
61   ..............................................
```



生物信息学

Bioinformatics

研究内容：对象（生物信息）

- **目前主要包括：**

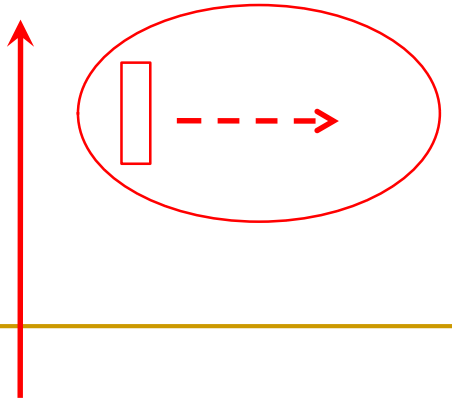
  macromolecular sequences;

  macromolecular structures;

  expression profiles; (EST; RNA-seq; microarrays; 2D-PAGE)

  biochemical network; (Interactions and reactions)

  evolution history.



生物信息学分析

章节	源数据	结果知识	种类
四、序列分析 *	DNA序列	基因等特征序列	Seq.
	蛋白质序列	特征域、特性	
	EST	表达基因（mRNA）	Expr.
五、系统发生分析	DNA/RNA/蛋白质序列	进化历史	Evol.
六、基因组分析	基因组序列	基因位置、功能、物种进化历史	Seq. Evol.
（转录组分析）	Microarray	表达基因（mRNA）	Expr.
	RNA-seq		
七、蛋白质组分析	2D-Page	表达基因（蛋白质）	Expr.
	Y2-hybrid ...	蛋白质相互作用…	Net.
八、结构分析	蛋白质序列	蛋白质结构	Struct.
	RNA序列	RNA结构	

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

什么是分子系统发生分析?*

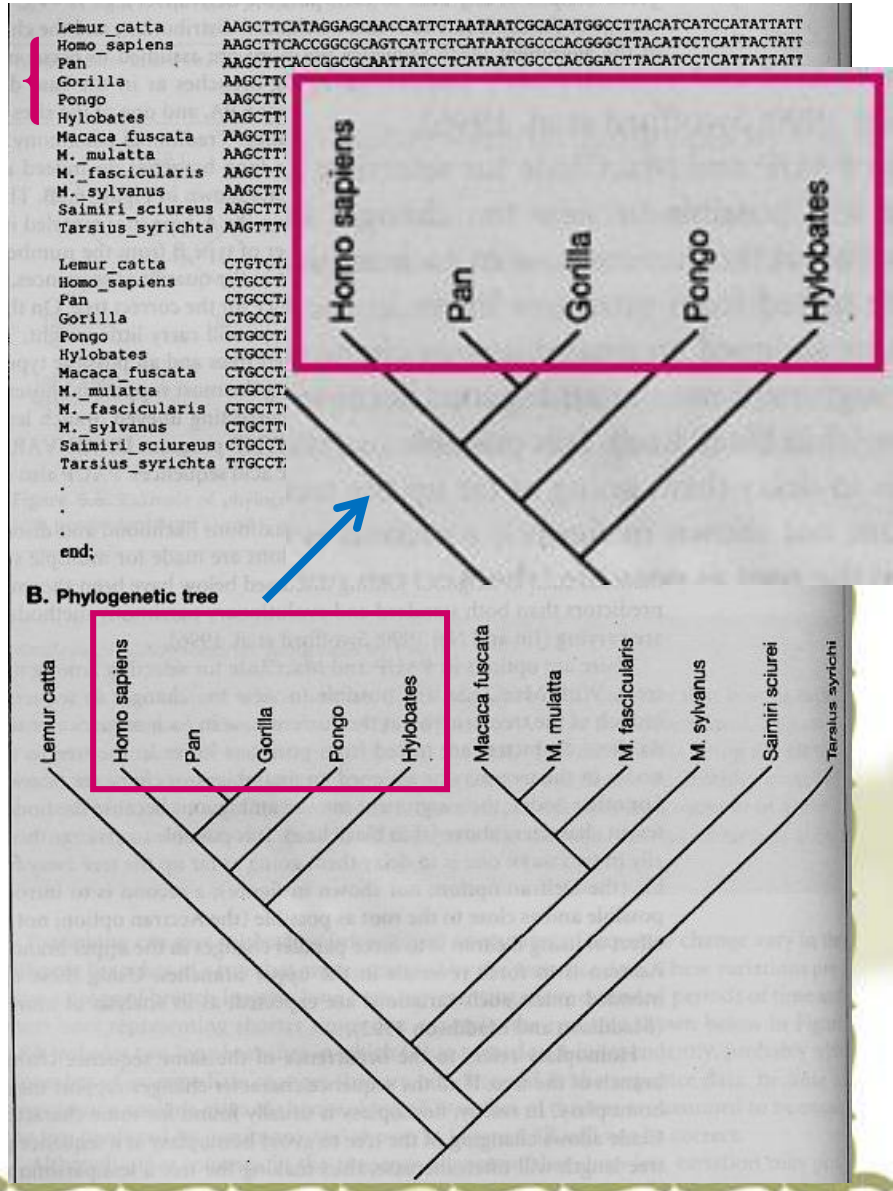
个体发生 （Ontogeny） ------- 发育

- 系统发生 (Phylogeny) —— 进化。
- 系统发生分析 —— 进化分析。
- 分子系统发生分析 —— 分子进化分析。

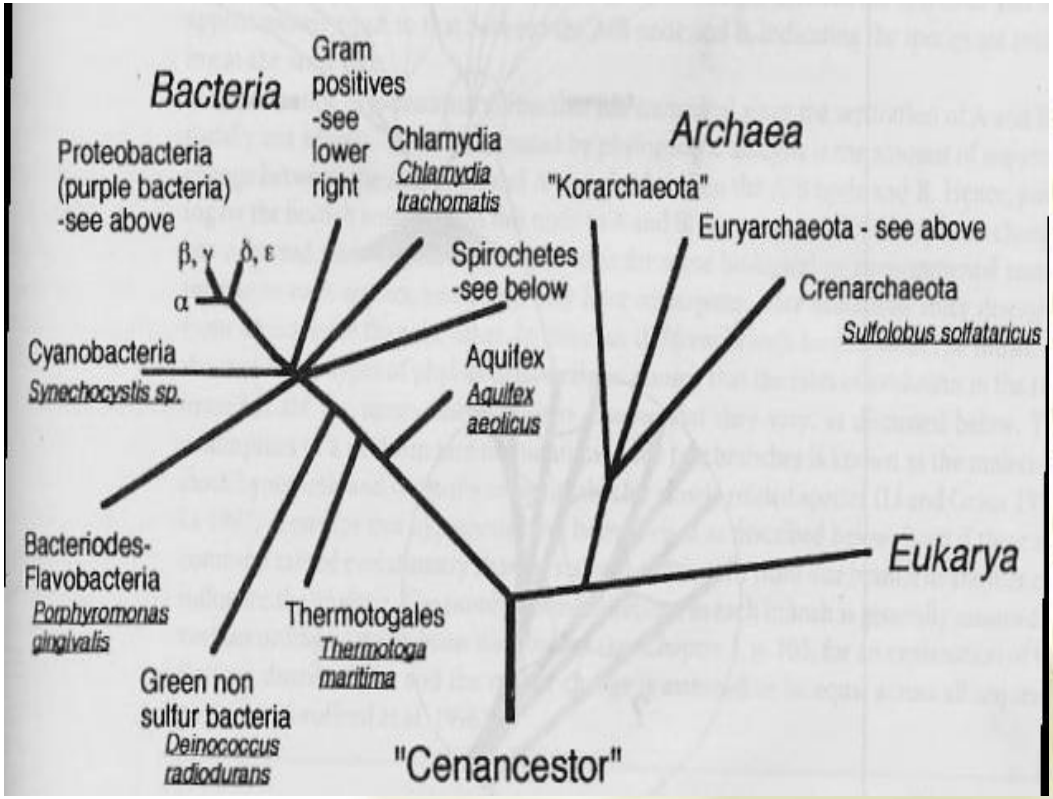
通过生物大分子序列比对，构建进化树，研究生物在分子水平的进化式样、方向、速率。

What is phylogenetic analysis?**

- A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.
- Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them.
- The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.



Tree of Life —— 生命之树



网上资源——生命之树（1）

Tree of Life Web Project - Windows Internet Explorer

http://tolweb.org/tree/

home browse help features learning contribute about

Search

TREE OF LIFE web project

Explore the Tree of Life

Browse the Site

Root of the Tree
Popular Pages
Sample Pages
Recent Additions
Random Page
Treehouses
Images, Movies,...

Search

News

Darwin 200: the celebration

Learn about ...

Ciidae

(minute tree-fungus beetles)

image info

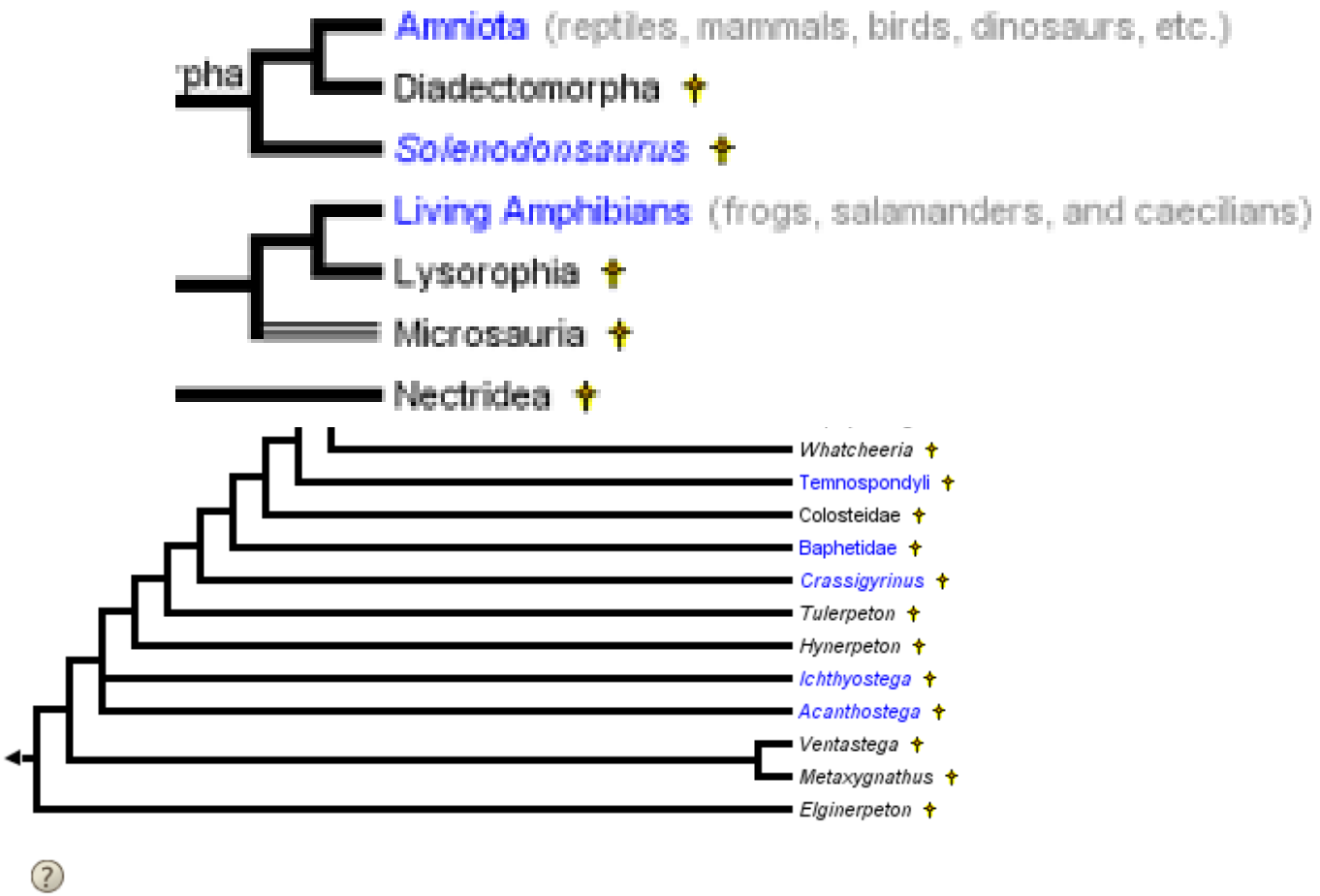
Ciidae is a family of minute beetles with nearly 650 described species in 42 genera, and hundreds of new forms known from museum and private collections...

read more

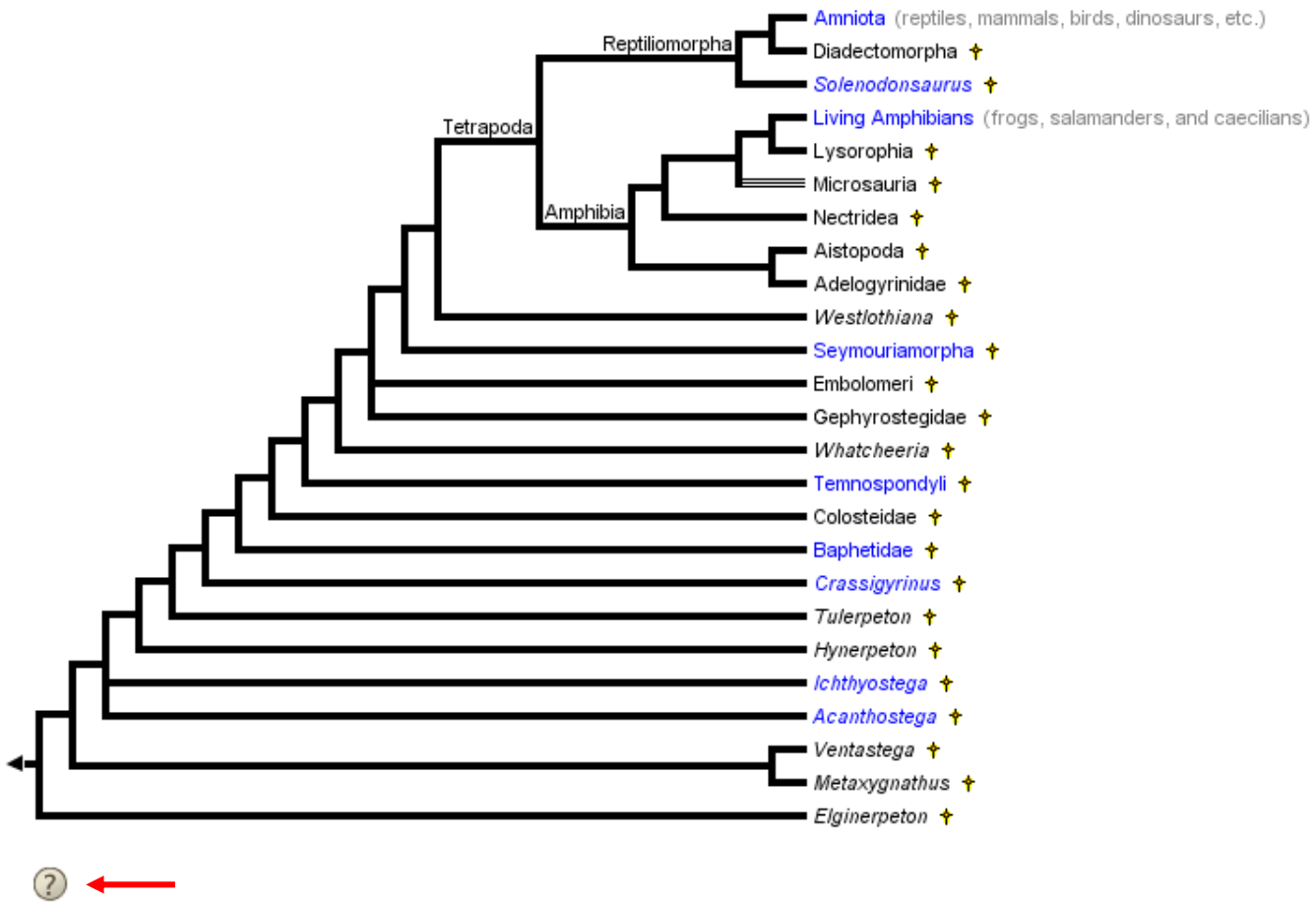
Internet 100%

开始 生物信息学2011 3 Microsoft P... Tree of Life W... My Pictures - ... 15:16

网上资源——生命之树（2）


Amniota (reptiles, mammals, birds, dinosaurs, etc.)
pha
Diadectomorpha
Solenodonsaurus
Living Amphibians (frogs, salamanders, and caecilians)
Lysorophia
Microsauria
Nectridea
Whatcheeria
Temnospondyli
Colosteidae
Baphetidae
Crassigyrinus
Tulerpeton
Hynerpeton
Ichthyostega
Acanthostega
Ventastega
Metaxygnathus
Elginerpeton

网上资源——生命之树（2）


Amniota (reptiles, mammals, birds, dinosaurs, etc.)
Reptiliomorpha
Diadectomorpha
Solenodonsaurus
Tetrapoda
Living Amphibians (frogs, salamanders, and caecilians)
Lysorophia
Microsauria
Amphibia
Nectridea
Aistopoda
Adelogyrinidae
Westlothiana
Seymouriamorpha
Embolomeri
Gephyrostegidae
Whatcheeria
Temnospondyli
Colosteidae
Baphetidae
Crassigyrinus
Tulerpeton
Hynerpeton
Ichthyostega
Acanthostega
Ventastega
Metaxygnathus
Elginerpeton

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树

- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率

- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

同源性与相似性*

- 同源性（Homology）：反映的是进化过程中的“亲缘关系”——有无及远近 。
- 相似性（similarity）：…… 进化分析中，其针对的是“性状” (traits) —— 可以是宏观的形态或结构，也可以是微观的大分子序列或结构等。

相似性高一般意味着同源性高 ——

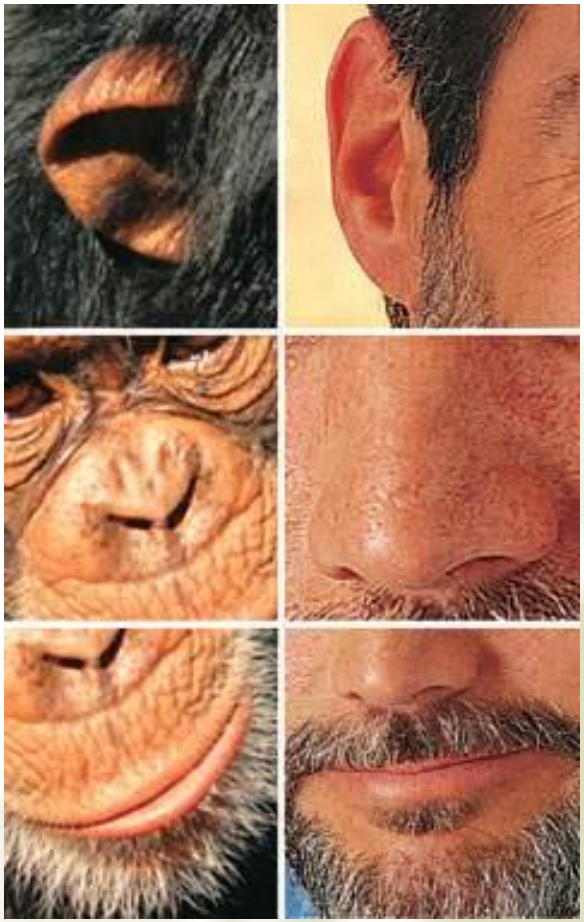
“进化总体来说是趋异的，而且是连续的”。

—— 用相似性来推断（infer）同源性……

用相似性推断同源性**

- 核酸或蛋白质分子的序列之间相似性可以更准确地定量表示，并且实际上更“本质”地标征同源性。

1. ATCCCGGTACGA……..
2. ACCCCGGTATGA……..
3. ATCGCGACACGA……...



网上资源——生命之树（2）


Amniota (reptiles, mammals, birds, dinosaurs, etc.)
pha
Diadectomorpha
Solenodonsaurus
Living Amphibians (frogs, salamanders, and caecilians)
Archosauromorpha (crocodiles, birds, and their extinct relatives)
Sauria
Lepidosauromorpha (lizards, snakes, Sphenodon, and their extinct relatives)
?
Sauropterygia
?
Ichthyosauria
Tulerpeton
Hynerpeton
Ichthyostega
Acanthostega
Ventastega
Metaxygnathus
Elginerpeton

分子进化中的同源类型

- 直系同源（orthology）：由“物种分化”而产生，有功能一致性，序列相似性一般比较高。
  比如，人的肌红蛋白和小鼠的肌红蛋白。
- 并系同源（paralogy）：由基因“多重化（duplicating）＋功能分化” 而产生。
  比如，肌红蛋白和血红蛋白中的α、β亚基。
- 其它的同源性概念：异同源（xenology），多异同源（paraxenology），部分同源（plerology）等。

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）
  类群　（*单系类群）
  系统发生树
- 基本原理

  置换模型　序列分歧度（*置换和相异性）
  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

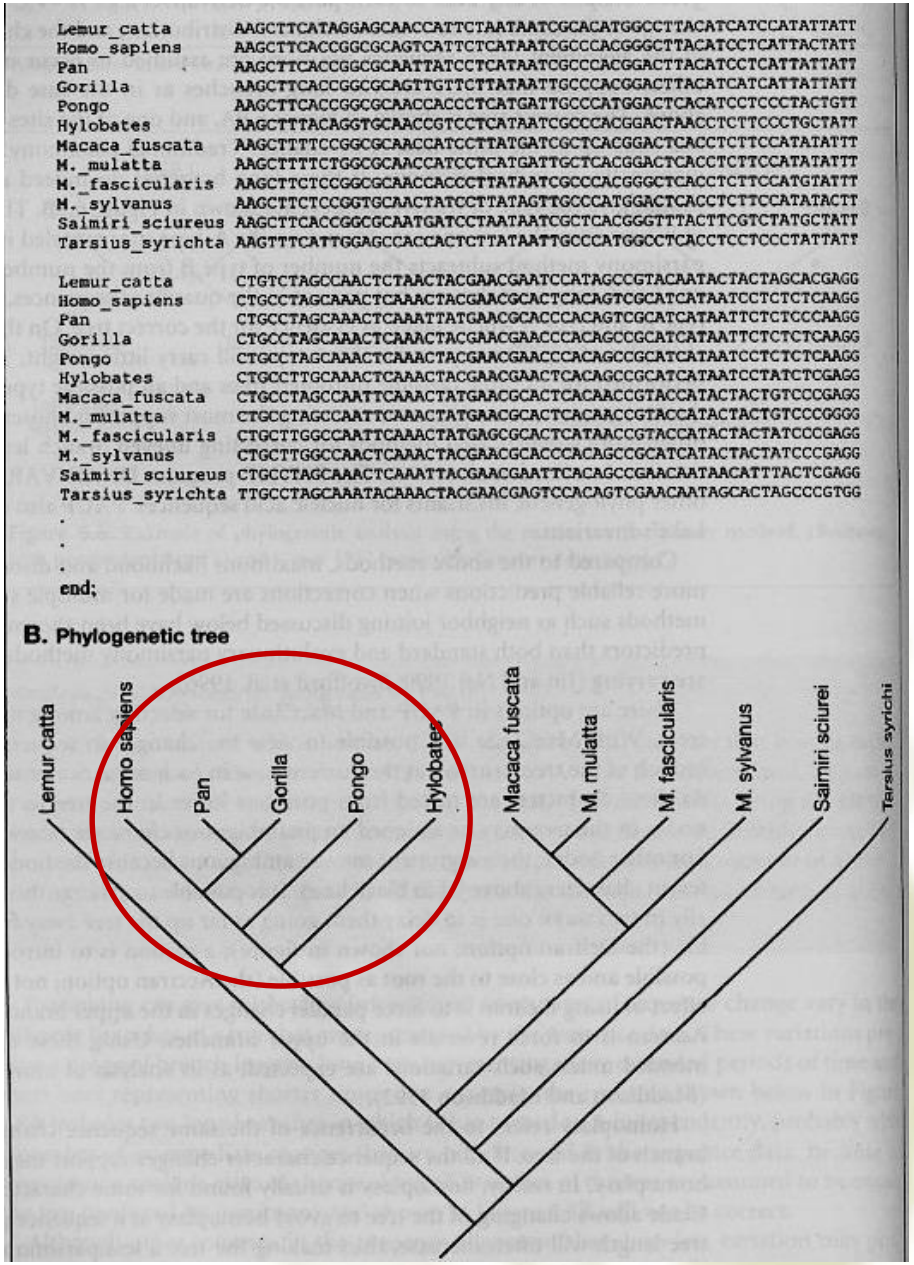
分类单元
Taxon

(or Operational Taxonomic Unit, OTU)

节点（Node）

类群 (Group)

Taxon及node的集合



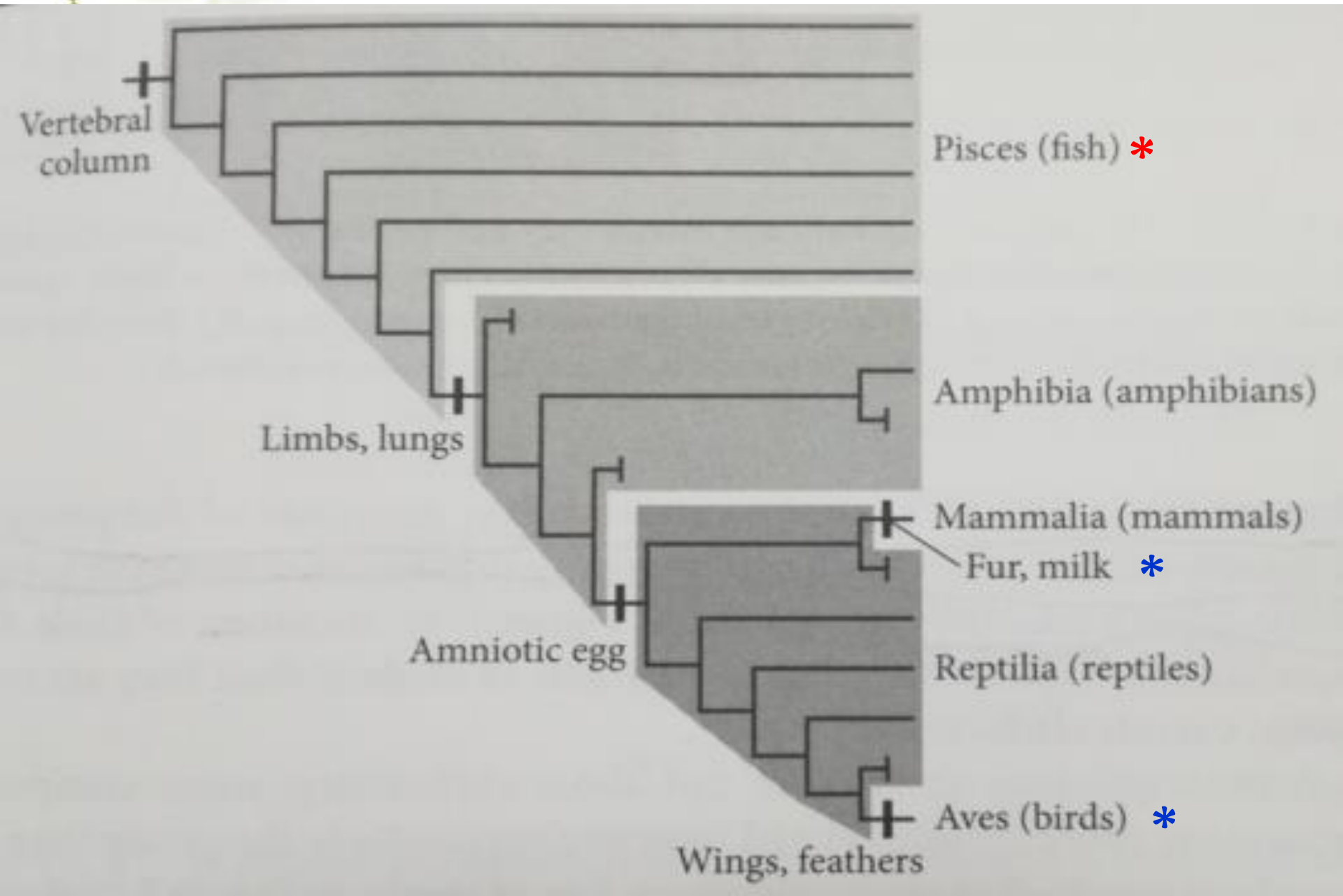
```
Lemur_catta          AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATT
Homo_sapiens         AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATT
Pan                  AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATT
Gorilla              AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATT
Pongo                AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTT
Hylobates            AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATT
Macaca_fuscata       AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTT
M._mulatta           AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTT
M._fascicularis      AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTT
M._sylvanus          AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTT
Saimiri_sciureus     AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATT
Tarsius_syrichta     AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATT

Lemur_catta          CTGTCTAGCCAACTCTAACTACGAACGAATCCATAGCCGTACAATACTACTAGCACGAGG
Homo_sapiens         CTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGG
Pan                  CTGCCTAGCAAACTCAAATTATGAACGCACCCACAGTCGCATCATAATTCTCTCCCAAGG
Gorilla              CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG
Pongo                CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATCCTCTCTCAAGG
Hylobates            CTGCCTTGCAAACTCAAACTACGAACGAACTCACAGCCGCATCATAATCCTATCTCGAGG
Macaca_fuscata       CTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTCCCGAGG
M._mulatta           CTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTCCCGGGG
M._fascicularis      CTGCTTGGCCAATTCAAACTATGAGCGCACTCATAACCGTACCATACTACTATCCCGAGG
M._sylvanus          CTGCTTGGCCAACTCAAACTACGAACGCACCCACAGCCGCATCATACTACTATCCCGAGG
Saimiri_sciureus     CTGCCTAGCAAACTCAAATTACGAACGAATTCACAGCCGAACAATAACATTTACTCGAGG
Tarsius_syrichta     TTGCCTAGCAAATACAAACTACGAACGAGTCCACAGTCGAACAATAGCACTAGCCCGTGG
.
.
.
end;
```

B. Phylogenetic tree

类群的种类(1)

- 祖先类群（ancestral group）--- 原始类群；
- 子裔类群（descendant group) --- 后代类群；
- 单系类群（monophyletic group）--- 一个祖先类群及其所有子裔类群的集合，或称为“进化枝”（clade）；
- 并系类群（paraphyletic group）：一个祖先类群及其部分子裔类群的集合；
- 复系类群（polyphyletic group）：涉及多个祖先类群及其子裔类群。

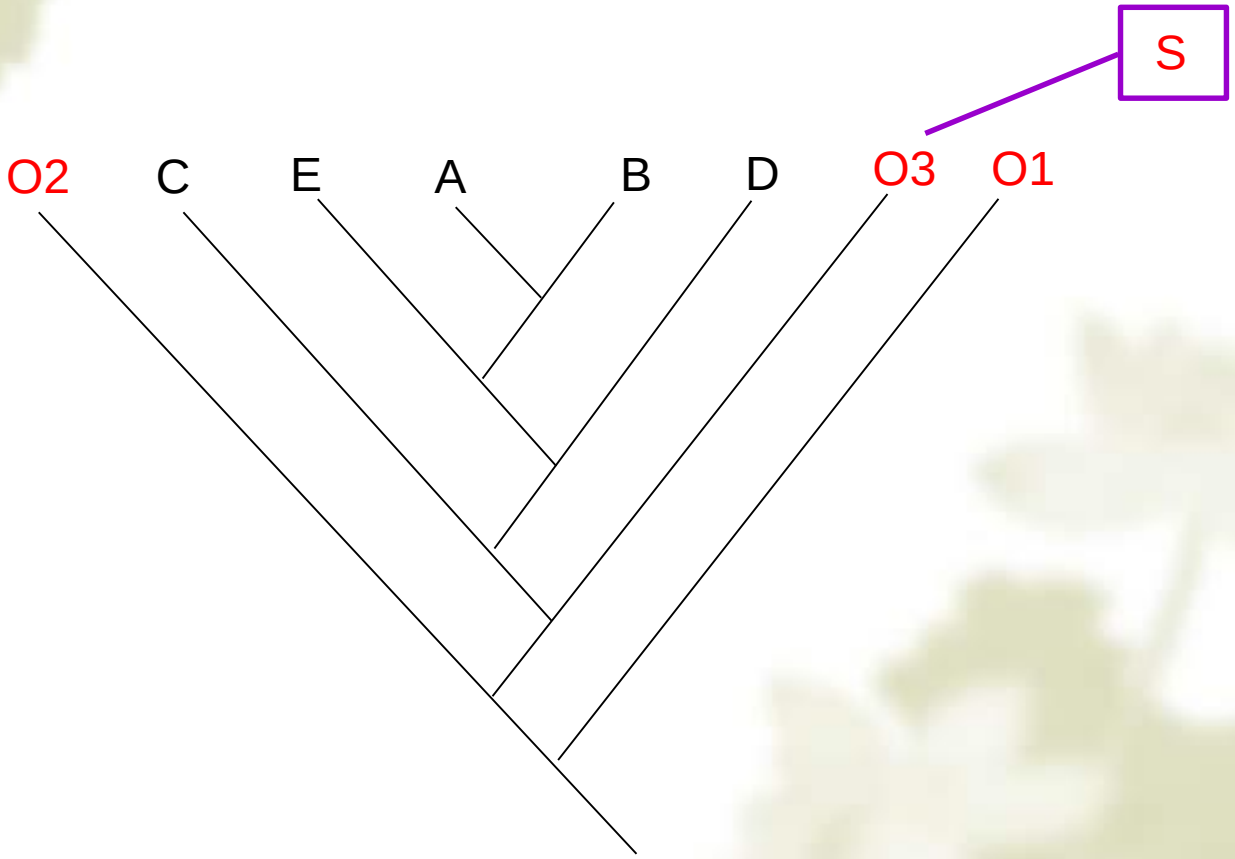
单系类群、并系类群、复系类群



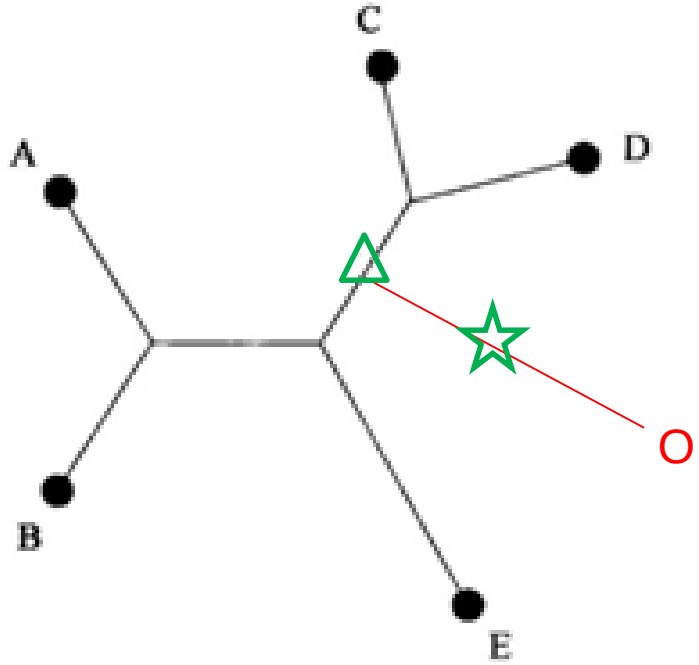
类群的种类(2)

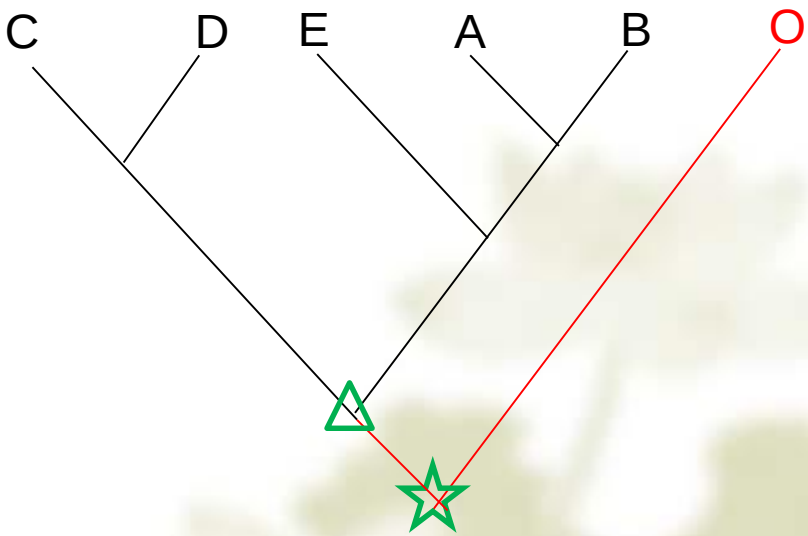
- 内类群（**ingroup**）：一项研究所涉及的某一特定类群可称为内类群；
- 外类群（**outgroup**）：不包含在内类群中（又与之有一定关系）的类群可称为外类群；
- 姐妹群（**sister group**）：与某一类群在谱系关系上最为密切的类群称为姐妹群。

内类群、外类群和姐妹群



利用外类群确定树根*





第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

系统发育树（Phylogenetic tree）

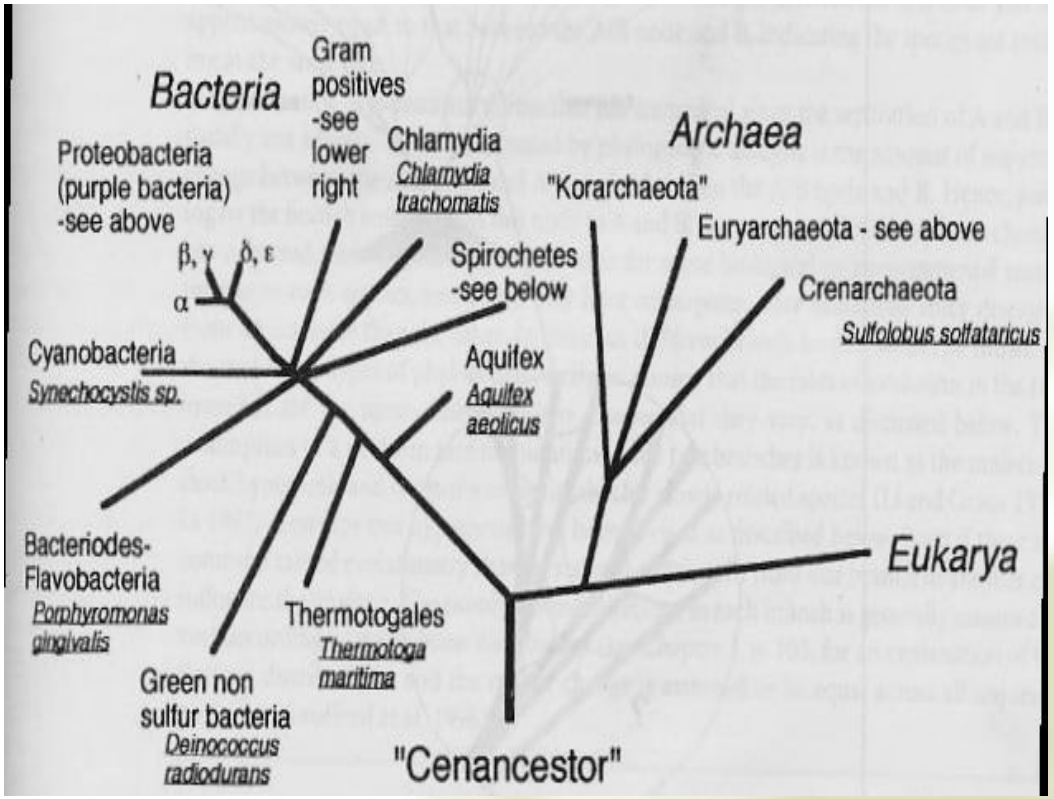
- **简称系统树(或进化树)：**

  表达类群间系统发育关系（进化关系）的一种树状图。

——Cladogram

二歧分枝和多歧分枝

- 多歧分枝多为无法确定分化先后次序而形成，并非真正同时分化。



有根树和无根树

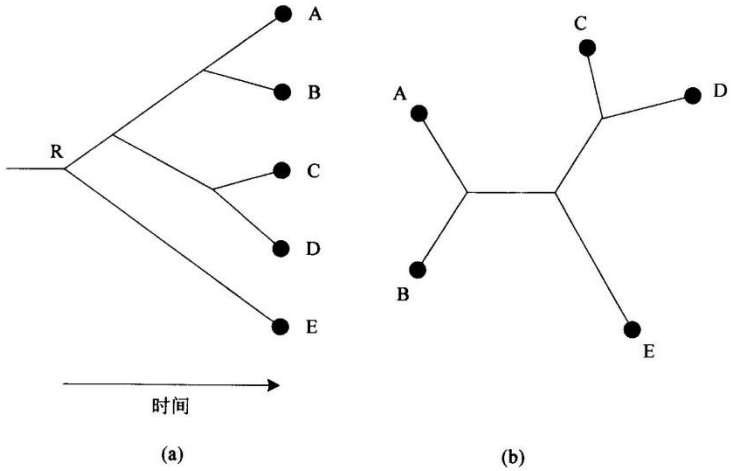


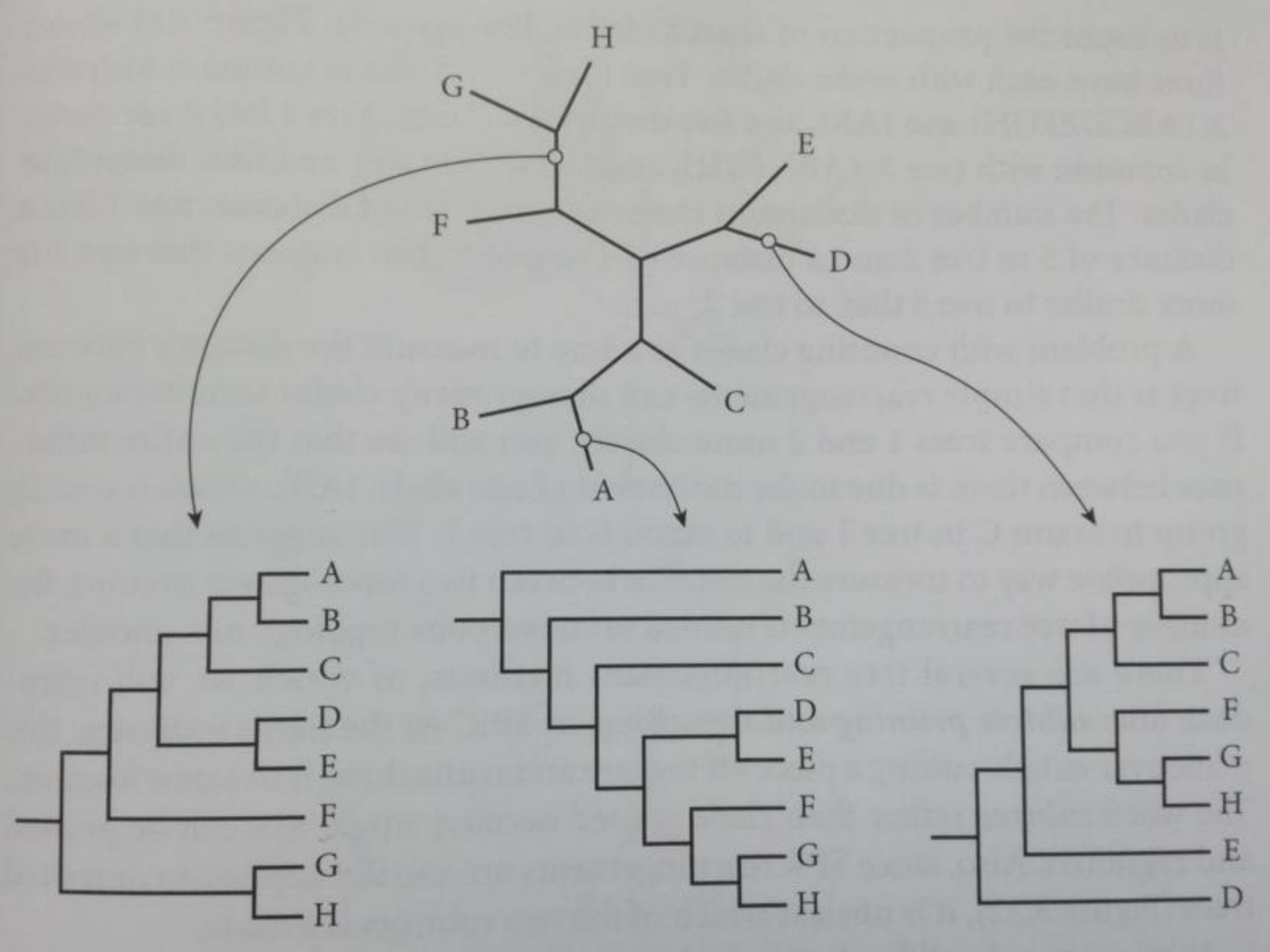
图 5－3　有根树(a)和无根树(b)

- **有根树（rooted tree）**
- **无根树（unrooted tree）**

当分类单位（OTU）数目n>2时， 对于二歧分支的情况：

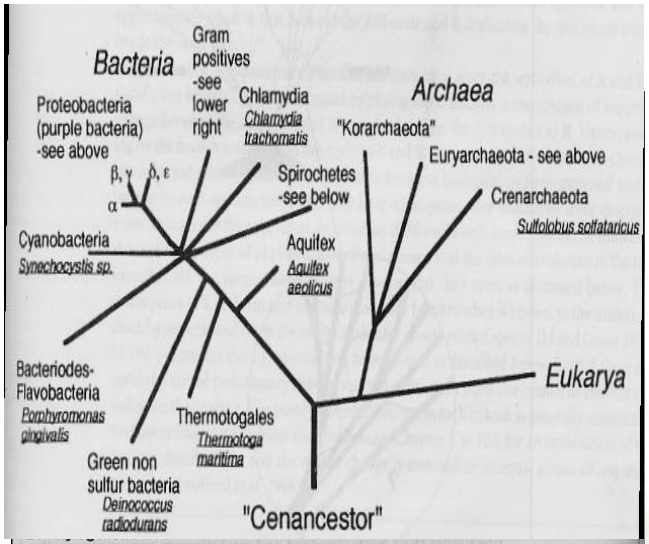
全部有根树的数目为 $(2n-3)! / [2^{n-2}(n-2)!]$ ；

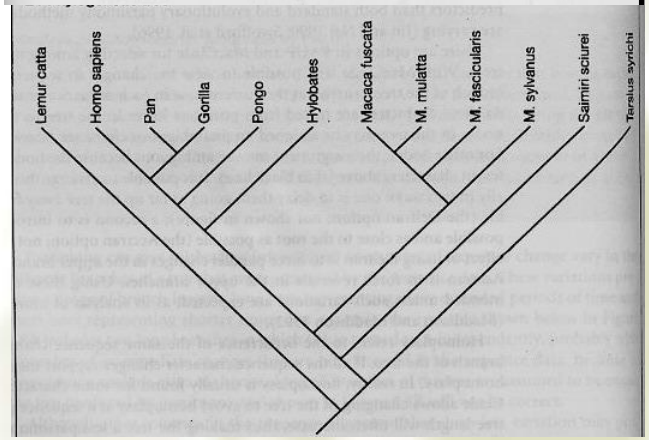
全部无根树的数目为 $(2n-5)! / [2^{n-3}(n-3)!]$ 。


H
G
E
F
D
C
B
A
A
B
C
D
E
F
G
H
A
B
C
D
E
F
G
H
A
B
C
F
G
H
E
D

标度（Scale）

- 标度树枝（**scaled branch**）系统树：树枝长度代表性状变异的数量。
- 非标度树枝（**unscaled branch**）系统树：树枝长度并不表示性状变异的数量。但其中节点的位置通常与分化时间相对应。





基因树和物种树

- 基因树（gene tree）：根据同源基因所构建的进化树。
- 物种树（species tree）：反映物种进化路径的进化树。

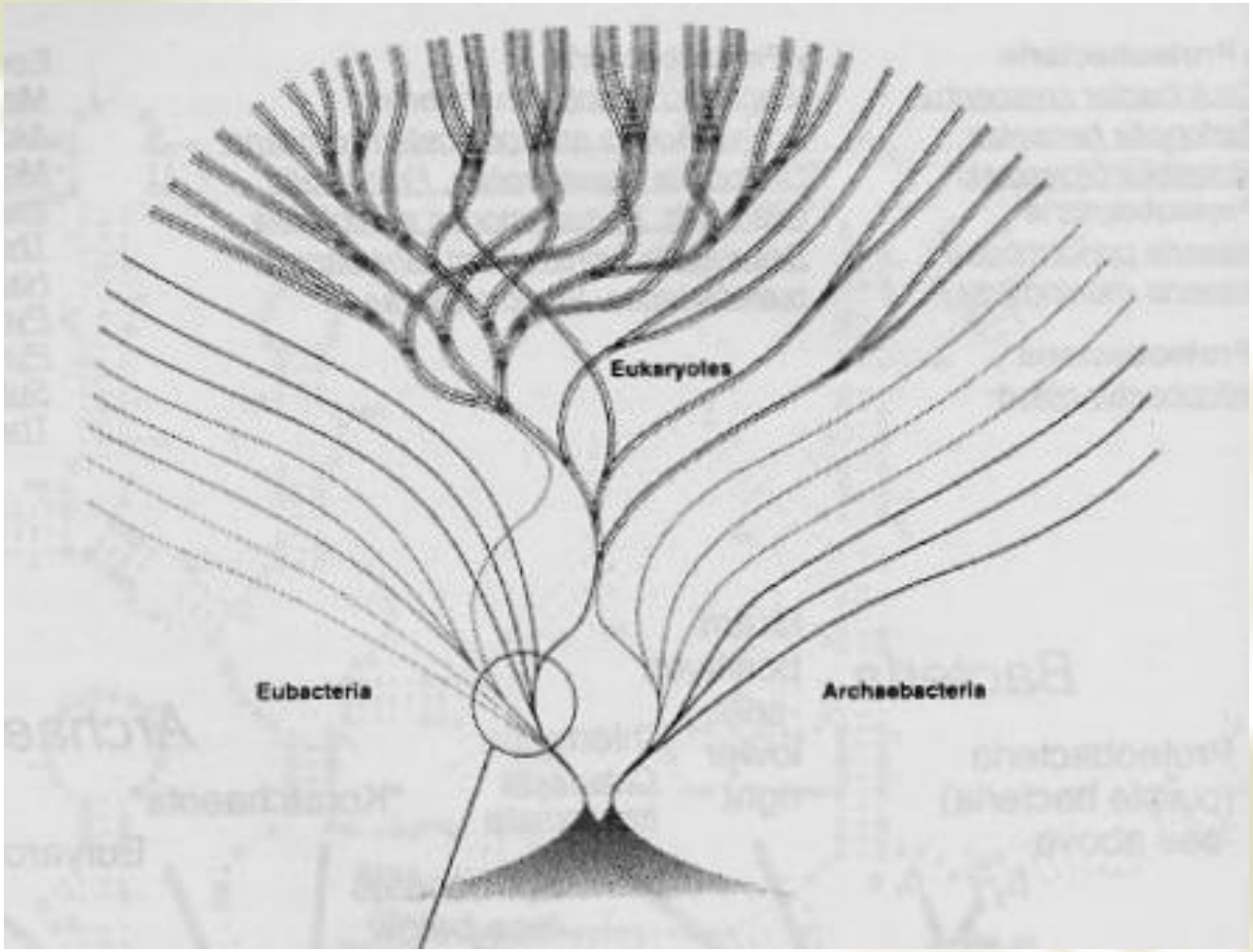
分子进化中的同源类型

- 直系同源（orthology）：由“物种分化”而产生，有功能一致性，序列相似性一般比较高。
  比如，人的肌红蛋白和小鼠的肌红蛋白。
- 并系同源（paralogy）：由基因“多重化（duplicating）＋功能分化” 而产生。
  比如，肌红蛋白和血红蛋白中的α、β亚基。
- 其它的同源性概念：异同源（xenology），多异同源（paraxenology），部分同源（plerology）等。

基因树和物种树

- 基因树（gene tree）：根据同源基因所构建的进化树。
- 物种树（species tree）：反映物种进化路径的进化树。
- 如何将多个（直系同源）基因树综合成一个物种树，是分子系统学的一个重要问题。
- （直系同源）基因树可能彼此不同，不一定是构树过程的问题，可能是进化过程本身造成……

网状树



第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

核苷酸置换模型（1）

- 进化的一个基本过程就是核苷酸随时间而变化（置换）。
- 核苷酸置换模型可以用矩阵来表示。

核苷酸置换模型（2）*

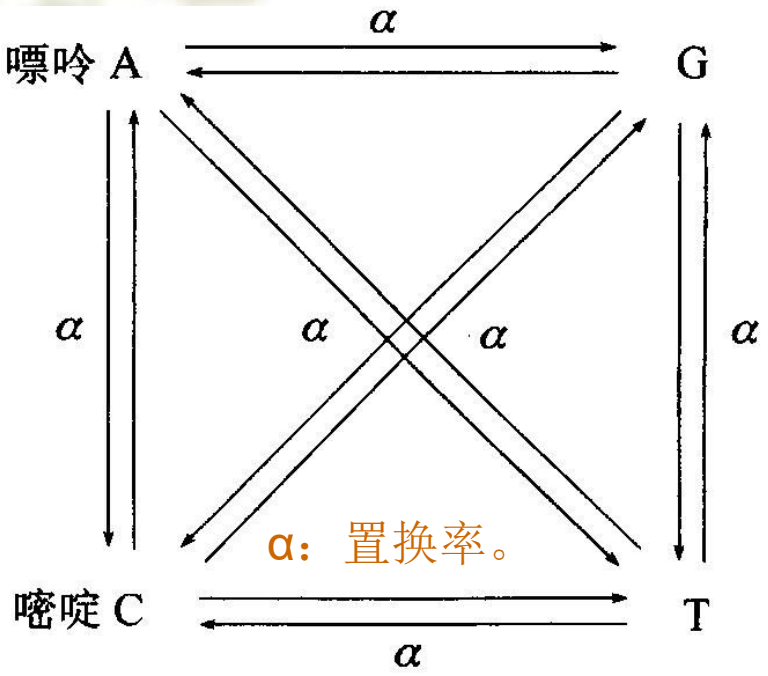


图 5－6　Jukes-Cantor 单参数模型

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}
\end{array}
$$

设n=0时为A，即$P_A(0) = 1$;

n=1时仍为A的概率：$P_A(1) = 1-3\alpha$;

$P_A(2) = (1-3\alpha)P_A(1) + \alpha[1-P_A(1)]$

$= P_A(1) - 4\alpha P_A(1) + \alpha$;

核苷酸置换模型（3）

$P_A(2) = P_A(1) - 4\alpha P_A(1) + \alpha;$

$P_A(3) = P_A(2) - 4\alpha P_A(2) + \alpha;$

……

$P_A(n+1) = P_A(n) - 4\alpha P_A(n) + \alpha;$

$\Delta P_A(n) = - 4\alpha P_A(n) + \alpha;$

$\Delta P_A(n) / 1 \sim \Delta P_A(t)/\Delta t$ ；

$P_A(n) \sim P_A(t)$ ；

求变化率，连续化 ——

$\Delta P_A(t)/\Delta t = - 4\alpha P_A(t) + \alpha;$

$\Delta t \rightarrow 0$ 时，

$dP_A(t)/dt = - 4\alpha P_A(t) + \alpha;$

积分得：

$$P_A(t) = 1/4 + 3/4\, e^{-4\alpha t},$$

相关地，可以推算出：

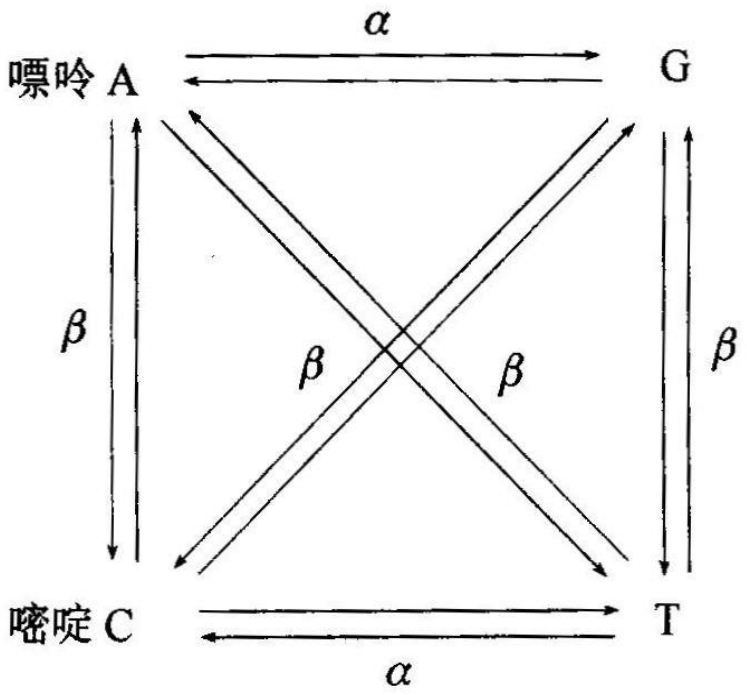
$$P_T(t) = 1/4 - 1/4\, e^{-4\alpha t},$$

$$P_C(t) = 1/4 - 1/4\, e^{-4\alpha t},$$

$$P_G(t) = 1/4 - 1/4\, e^{-4\alpha t},$$

取代函数

核苷酸置换模型（4）



$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{bmatrix}
-\alpha-2\beta & \beta & \alpha & \beta \\
\beta & -\alpha-2\beta & \beta & \alpha \\
\alpha & \beta & -\alpha-2\beta & \beta \\
\beta & \alpha & \beta & -\alpha-2\beta
\end{bmatrix}
\end{array}
$$

Kimura两参数模型：

转换（transition）率为α；

颠换（transversion）率为β。

核苷酸置换模型与氨基酸置换模型

- **Jukes－Cantor单参数模型和Kimura两参数模型是核苷酸置换模型中最为常用的，其它的还有K3ST、F81、GTR、HKY85、SYM、TrN等更为复杂的模型。**
- 氨基酸置换模型涉及密码子基础（即考虑变异），以及特性关系（如极性）（即考虑自然选择）。常见的氨基酸置换模型有：**Dayhoff（PAM）、BLOSUM、Jones-Taylor-Thornton、mtREV模型等。**

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

序列分歧度（相异性指数）

- 设两个序列比对长度为L，残基差异值为N，则差异率P＝N/L;
- 序列分歧度（sequence divergence）K ——

DNA序列分歧度:

单参模型，$K=-\frac{3}{4}\ln(1-4P/3)$

蛋白质序列分歧度:

假设氨基酸之间的置换率相同，$K=-\ln(1-P)$ 。

置换和相异性*

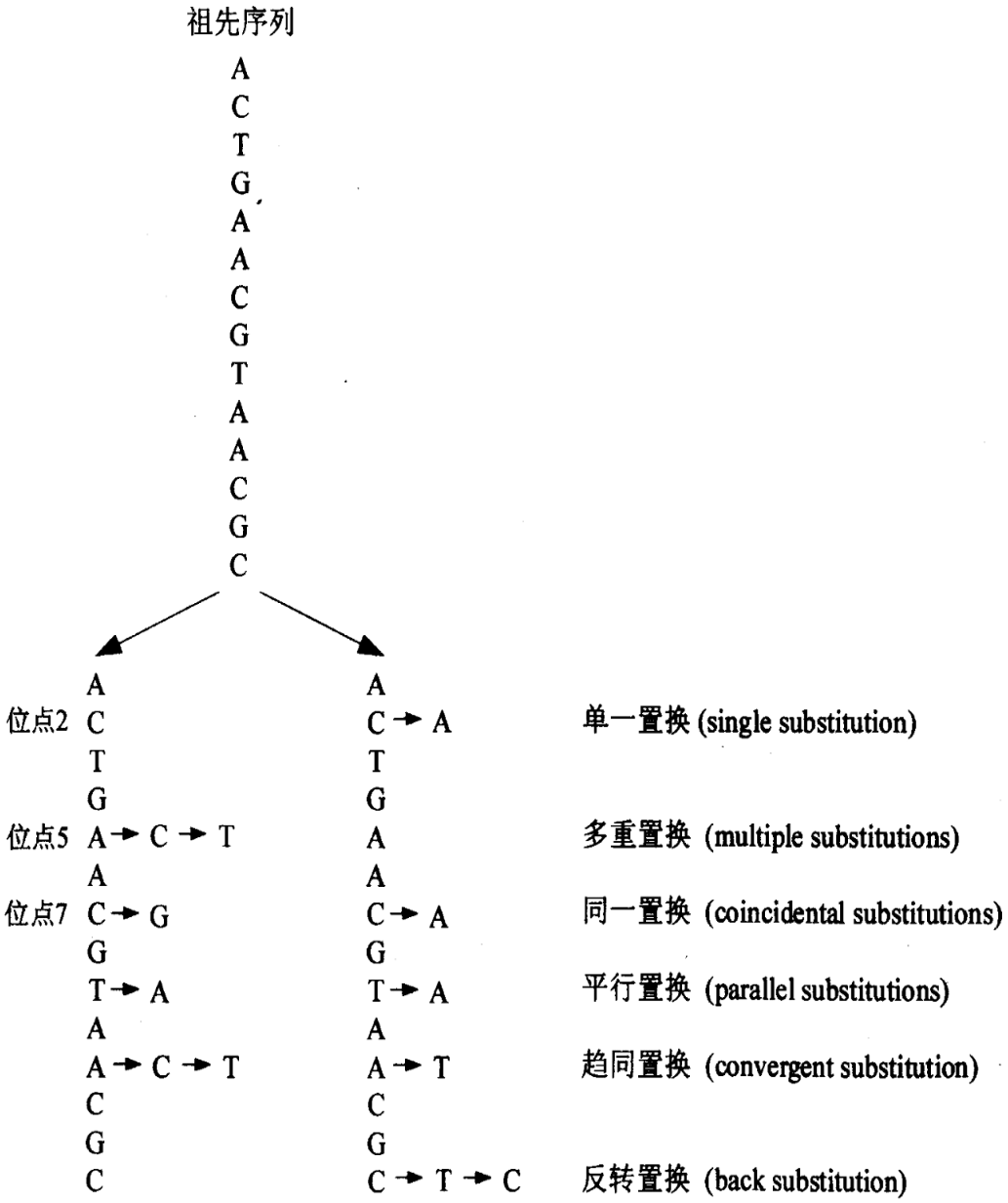


图 5－8　同源序列间的核苷酸置换

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

进化速率问题*

- **变异**本身的速率有差异。

  RNA病毒：易错复制酶，如HIV、流感病毒、新冠病毒。

  线粒体DNA：易错复制酶，高氧环境。（进化很快）

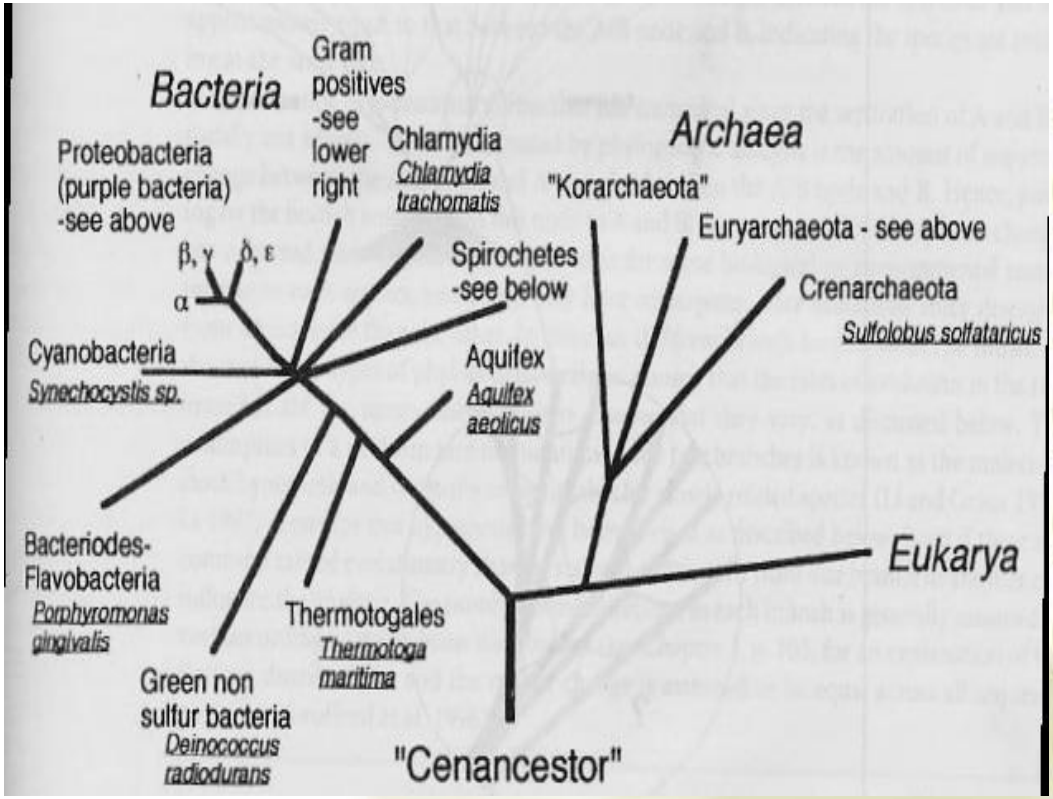
- **自然选择**影响进化速率（保守性）。

  组蛋白和rRNA等重要基因进化很慢；非编码区进化很快。

—— 进化分析要求序列相似性不能太低，或太高。

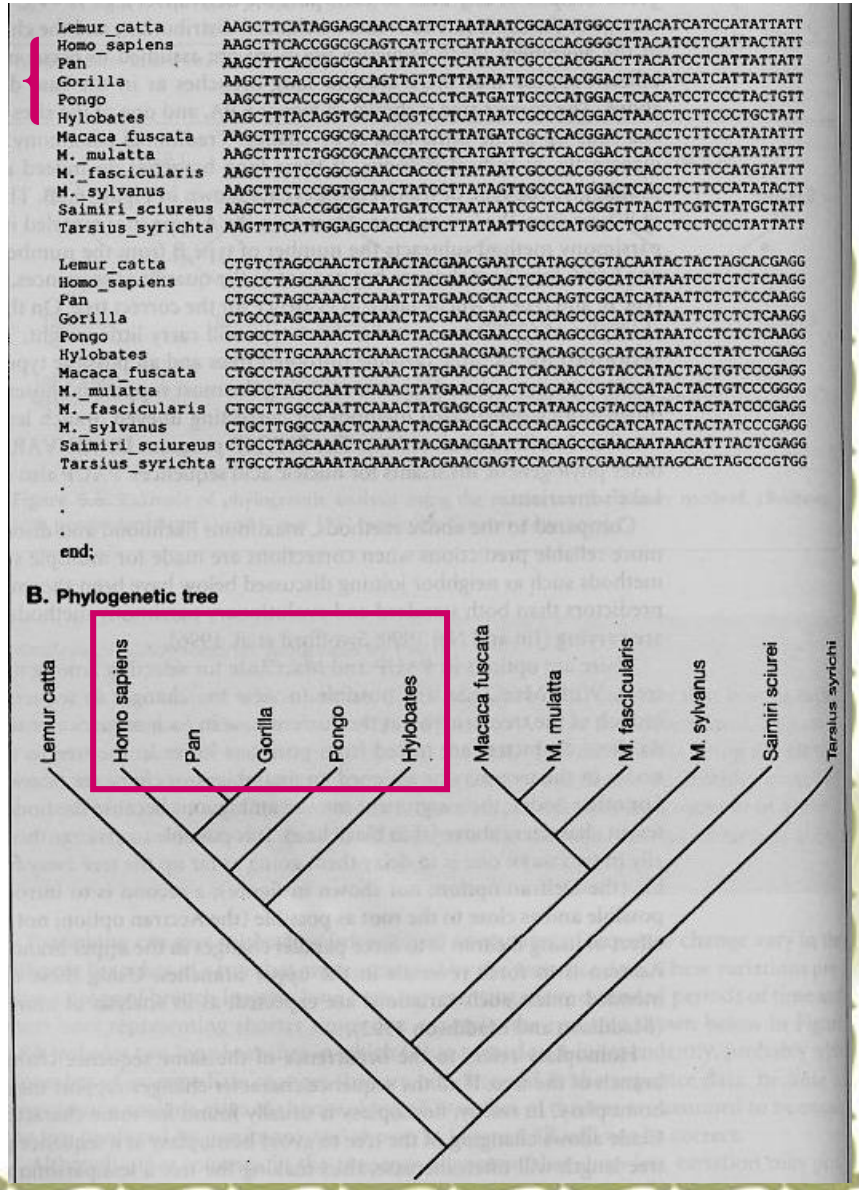
问题：研究亲缘关系远的物种进化关系应该用进化快的序列还是进化慢的序列?

Tree of Life —— 生命之树



What is phylogenetic analysis?**

- A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.
- Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them.
- The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.



第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树

- 基本原理

  置换模型
  序列分歧度　}（*置换和相异性）
  进化速率

- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

分子进化树的几种主要构建方法

- 距离矩阵法
  （distance matrix method)
- 简约法
  （parsimony method)
- 最大似然法
  （maximum likelihood method）

距离矩阵法（distance matrix method)

- 首先，简单地计算两个序列的**差异数量**。
- 这个数量被看作进化的距离，以这些距离构造矩阵，即距离矩阵。
- 然后，运行一个聚类算法，即从最相似（也就是之间距离最短）的序列开始，通过距离矩阵计算出实际的进化树；或者通过将总的树枝长度最小化而优化出进化树。

距离矩阵法的种类

著名的距离矩阵法有：

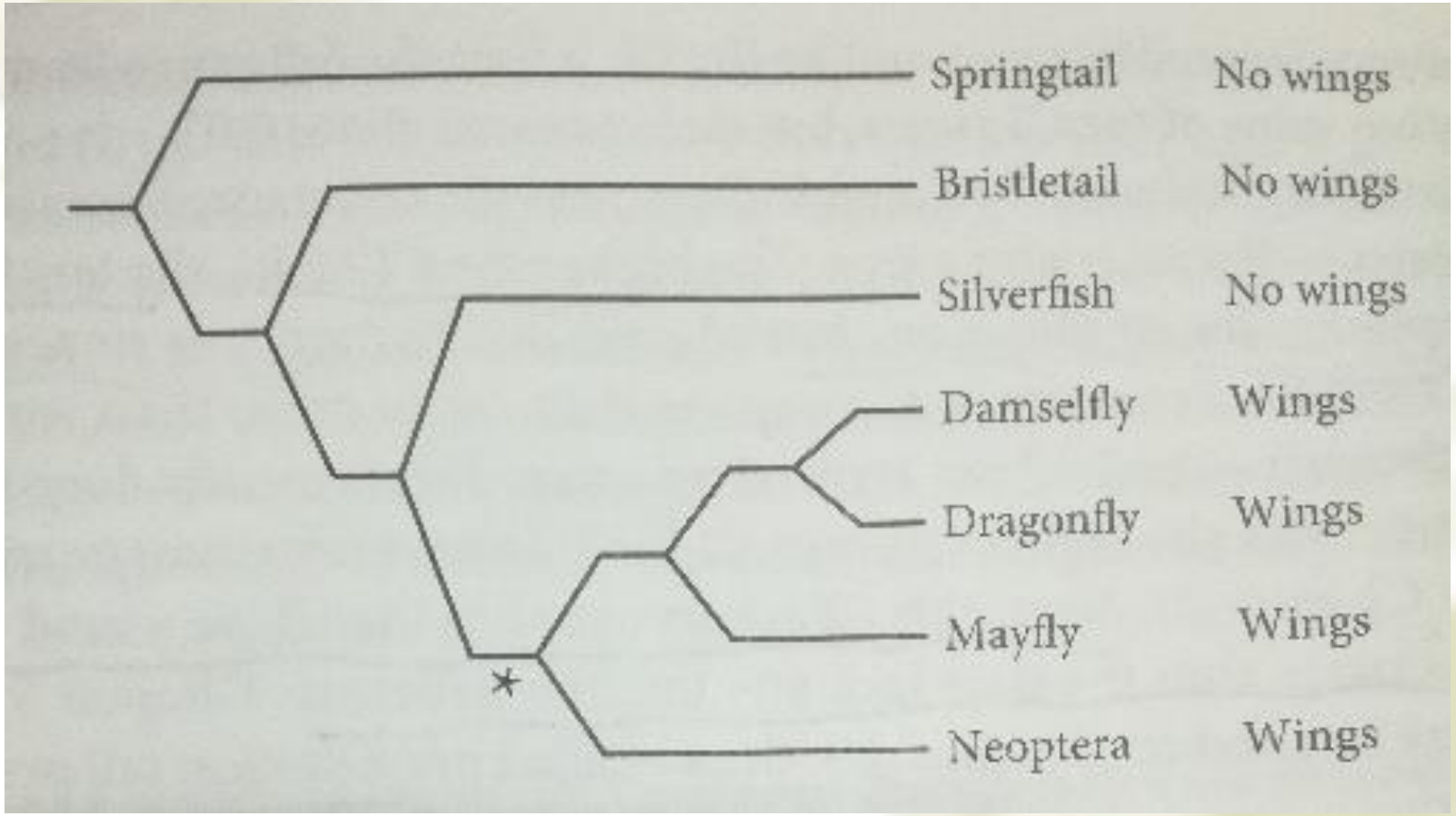
- UPGMA—— Unweighted Pair Group Method with Arithmetic Mean （算术平均的不加权对群法）
- NJ —— neighbor-joining （邻接法）
- FM —— Fitch-Margoliash
- ME —— Minimum Evolution （最小进化法）

简约法（parsimony method)

- 最大简约法：要求用最少的改变来解释所研究的分类群之间的差异。
- 进化简约法（加权最大简约法）：

  考虑实际改变（替换）的代价不同，如转换和颠换的区分。

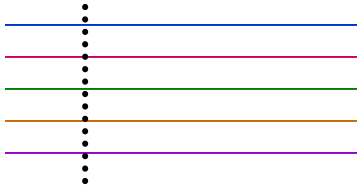
简约法的原理



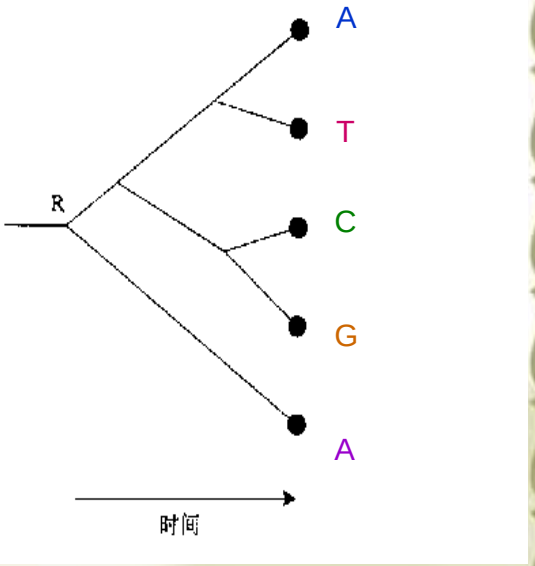
用过程中最少的改变解释所观测到的结果差异

最大似然法（maximum likelihood method）*

评估所选定的进化树能够产生实际观察到的数据的可能性：（基于置换）



➡ 针对一个位点的进化，

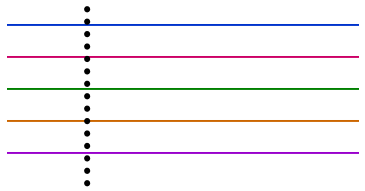


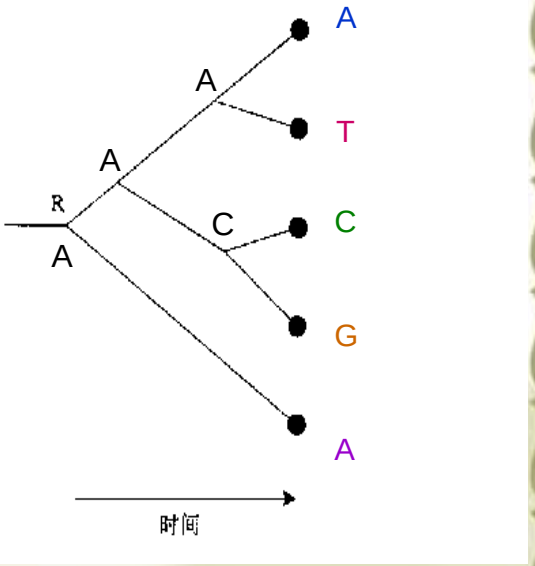
换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

最大似然法（maximum likelihood method）

评估所选定的进化树能够产生实际观察到的数据的可能性：（**基于置换**）

针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据**取代函数**计算每一段进化的可能性，将所有段的这种可能性相？，得到**此进化树**以**此组合**为进化途径产生**此位点**数据的可能性（似然值）；





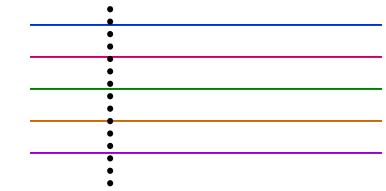
换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

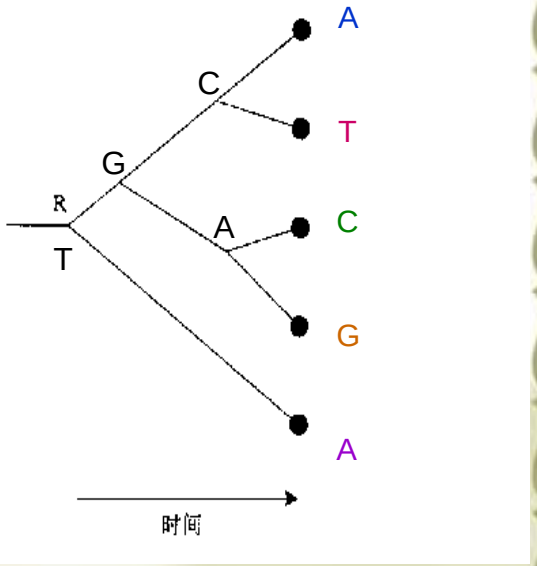
最大似然法（maximum likelihood method）

评估所选定的进化树能够产生实际观察到的数据的可能性：（**基于置换**）

针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据**取代函数**计算每一段进化的可能性，将所有段的这种可能性相？，得到**此进化树**以**此组合**为进化途径产生**此位点**数据的可能性；

➡ 换组合，再算。





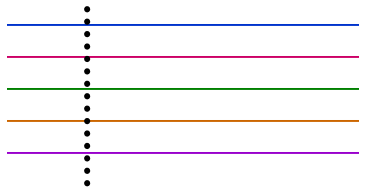
换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

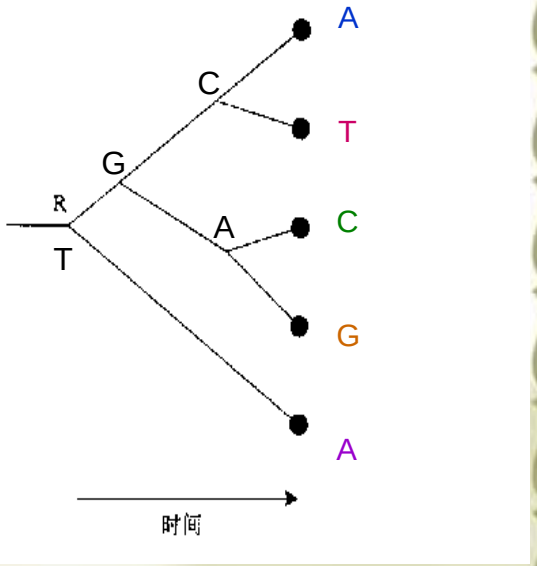
最大似然法（maximum likelihood method）

评估所选定的进化树能够产生实际观察到的数据的可能性：（**基于置换**）

针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据**取代函数**计算每一段进化的可能性，将所有段的这种可能性相？，得到**此进化树**以**此组合**为进化途径产生**此位点**数据的可能性；

换组合，再算。将所有组合的这种可能性相？，





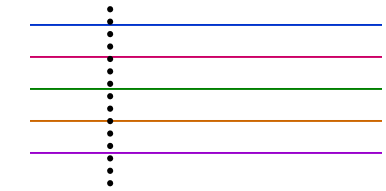
换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

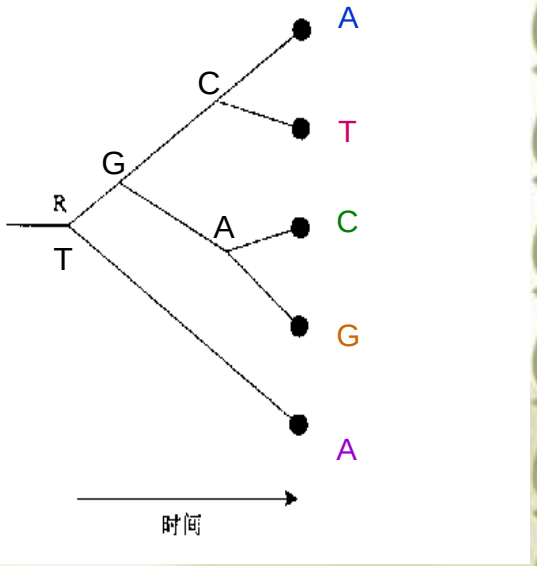
最大似然法（maximum likelihood method）

评估所选定的进化树能够产生实际观察到的数据的可能性：（**基于置换**）

针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据**取代函数**计算每一段进化的可能性，将所有段的这种可能性相？，得到**此进化树**以**此组合**为进化途径产生**此位点**数据的可能性；

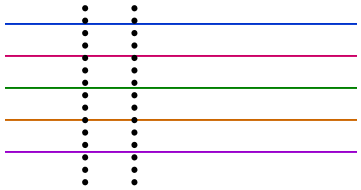
换组合，再算。将所有组合的这种可能性相？，得到**此进化树**产生**此位点**数据的可能性；





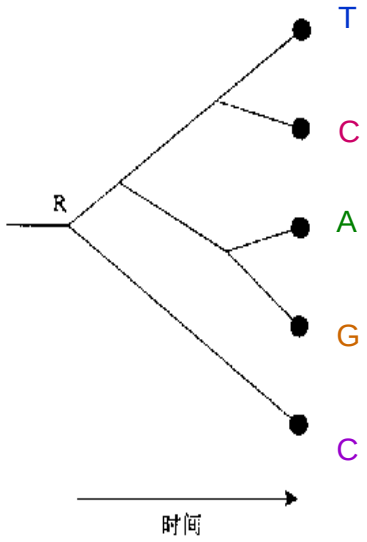
换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

最大似然法（maximum likelihood method）



评估所选定的进化树能够产生实际观察到的数据的可能性：（基于置换）

针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据**取代函数**计算每一段进化的可能性，将所有段的这种可能性相？，得到**此进化树**以**此组合**为进化途径产生**此位点**数据的可能性；



换组合，再算。将所有组合的这种可能性相？，得到**此进化树**产生**此位点**数据的可能性；

换位点，再算。将所有位点的这种可能性相？，得到**此进化树**产生整个序列组数据的可能性。

换进化树，再评估……相比较。产生此数据可能性最大的进化树被认为是最可能的实际进化树。

贝叶斯推断法

- 似然值其实是某树产生某数据的可能性；
- 而我们真正感兴趣的是：已知数据的情况下各个树的可能性---应该借助贝叶斯公式。
- P(AB) = P(A)*P(B|A) = P(B)*P(A|B)
- ➔ P(B|A) = P(A|B)*P(B)/P(A)
- $$\mathrm{P}(Tree|Data) = \frac{\mathrm{P}(Data|Tree)*\mathrm{P}(Tree)}{\mathrm{P}(Data)}$$

构建系统树的各种方法之比较

1.假设---相关敏感性

- UPGMA：进化速率一致--- 敏感；序列长 --- 敏感。
- 邻接法：距离系数准确 --- 敏感；序列长--- 敏感。
- 最大简约法：进化速率一致 --- 敏感；序列很相似 --- 敏感。
- 最大似然法：进化速率一致 --- 不敏感；某种置换模型 --- 不敏感。 （Robust）

构建系统树的各种方法之比较

2.估计一致性（Consistency）

——数据（样本证据）增加时，是否能倾向正确的树。

- 距离矩阵法：进化速率恒定时一致；进化速率变化时不一致或难一致。
- 简约法：不一致。
- 最大似然法：比较一致；其中特别是，只要置换模型合理、正确，就能很好地一致。

构建系统树的各种方法之比较

3.符合程度评价

- “创立”进化过程，再用各种构树方法推测进化过程：

  A.实际进化：预先得到了实际的进化树（如实验室控制进化），再来检验分子构树的各种方法。

  B.计算机模拟进化。

- 结果：（一般来说）

  A.进化速率恒定时：简约法≤距离矩阵法；

  最大似然法依赖于进化模型。

  B.进化速率可变时：简约法≤距离矩阵法＜最大似然法。

构建系统树的各种方法之比较

4.计算时间

距离矩阵法 < 简约法 < 最大似然法

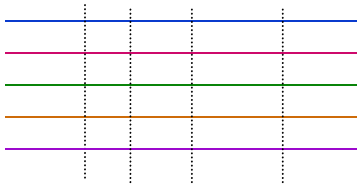
一般构树方法选用策略

- 序列间有极高相似性：简约法。
- 序列间有较明显的相似性：距离矩阵法。
- 序列间没有较明显的相似性：最大似然法。

分子系统树的可靠性

- 一般没有绝对有效的方法来确定一个构出的结果树是否代表真正的进化历史。
- 如果不同方法构出的树一样，则较可靠，即很可能真正反映进化历史。
- 另外，可以通过统计学方法辅助增进其可靠性。

自展法（Bootstrap method）



- 自展法利用对原始数据随机抽样产生的自展数据获得大量“虚拟”树，然后求一致树。

序列组——自展——构树（大量虚拟树）——一致树

Bootstrapping

TABLE 9.2 Generating a bootstrap replicate from a set of observed data

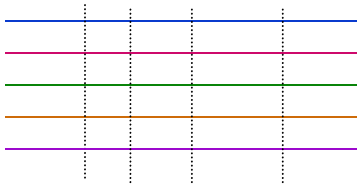
Original data set

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
A	T	T	T	C	C	T	T	T	C	A	G	G	T	A	T	T	A	T	G	A	G	A	T	A	C	G	T	A	C	T	G	A	A	A	A	A	G	T	C	C
B	T	T	T	C	C	T	T	T	T	A	G	G	T	T	T	G	A	T	G	A	G	A	T	A	C	A	T	T	A	C	G	A	A	A	G	A	G	T	C	A
C	T	T	T	G	C	T	T	C	T	C	G	G	T	A	C	T	A	C	A	A	T	A	T	A	T	A	T	A	C	C	A	G	A	A	A	A	G	T	C	A
D	T	T	T	G	C	T	T	C	C	G	A	C	T	A	C	A	A	A	G	G	C	A	T	A	C	G	T	A	G	C	T	G	A	A	A	A	G	G	C	G
E	C	T	T	G	C	C	T	A	C	T	G	T	T	G	C	A	A	T	A	A	T	A	T	A	C	G	A	A	G	C	T	A	A	A	A	A	G	T	C	G
F	T	T	C	G	T	C	C	C	C	G	G	C	T	A	C	A	A	T	G	G	T	A	T	A	T	G	T	A	C	T	C	G	A	A	A	A	G	A	T	G
G	G	T	T	G	T	T	T	C	C	G	G	C	T	A	C	A	G	T	G	A	T	A	T	A	C	G	T	A	C	C	C	G	A	G	A	A	C	T	T	G
H	T	T	T	A	T	T	T	C	C	G	G	C	T	A	C	A	G	T	G	A	T	A	T	A	C	G	T	G	C	C	C	G	A	G	A	A	G	T	T	G

Bootstrap data set

	02	39	35	22	36	31	40	05	16	23	15	35	35	40	03	06	24	33	06	07	14	20	35	01	36	09	13	22	11	25	26	33	03	09	16	20	08	18	17	32
A	T	C	A	A	A	G	C	C	T	T	T	A	A	C	T	T	A	A	T	T	A	A	A	T	A	C	T	A	G	C	G	A	T	C	T	A	T	T	A	A
B	T	C	G	A	A	G	A	C	G	T	T	G	G	A	T	T	A	A	T	T	T	A	G	T	A	T	T	A	G	C	A	A	T	T	G	A	T	T	A	A
C	T	C	A	A	A	A	A	C	T	T	C	A	A	A	T	T	A	A	T	T	A	A	A	T	A	T	T	A	G	T	A	A	T	T	T	A	C	C	A	G
D	T	C	A	A	A	T	G	C	A	T	C	A	A	G	T	T	A	A	T	T	A	G	A	T	A	C	T	A	A	C	G	A	T	C	A	G	C	A	A	G
E	T	C	A	A	A	T	G	C	A	T	C	A	A	G	T	C	A	A	C	T	G	A	A	C	A	C	T	A	G	C	G	A	T	C	A	A	A	T	A	A
F	T	T	A	A	A	C	G	T	A	T	C	A	A	G	C	C	A	A	C	C	A	G	A	T	A	C	T	A	G	T	G	A	C	C	A	G	C	T	A	G
G	T	T	A	A	A	C	G	T	A	T	C	A	A	G	T	T	A	A	T	T	A	A	A	G	A	C	T	A	G	C	G	A	T	C	A	A	C	T	G	G
H	T	T	A	A	A	C	G	T	A	T	C	A	A	G	T	T	A	A	T	T	A	A	A	T	A	C	T	A	G	C	G	A	T	C	A	A	C	T	G	G

自展法（Bootstrap method）



- 自展法利用对原始数据随机抽样产生的自展数据获得大量“虚拟”树，然后求一致树。

序列组——自展——构树（大量虚拟树）——一致树

Bootstrapping

TABLE 9.2 Generating a bootstrap replicate from a set of observed data

Original data set

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
A	T	T	T	C	C	T	T	T	C	A	G	G	T	A	T	T	A	T	G	A	G	A	T	A	C	G	T	A	C	T	G	A	A	A	A	A	G	T	C	C
B	T	T	T	C	C	T	T	T	T	A	G	G	T	T	T	G	A	T	G	A	G	A	T	A	C	A	T	T	A	C	G	A	A	A	G	A	G	T	C	A
C	T	T	T	G	C	T	T	C	T	C	G	G	T	A	C	T	A	C	A	A	T	A	T	A	T	A	T	A	C	C	A	G	A	A	A	A	G	T	C	A
D	T	T	T	G	C	T	T	C	C	G	A	C	T	A	C	A	A	A	G	G	C	A	T	A	C	G	T	A	G	C	T	G	A	A	A	A	G	G	C	G
E	C	T	T	G	C	C	T	A	C	T	G	T	T	G	C	A	A	T	A	A	T	A	T	A	C	G	A	A	G	C	T	A	A	A	A	A	G	T	C	G
F	T	T	C	G	T	C	C	C	C	G	G	C	T	A	C	A	A	T	G	G	T	A	T	A	T	G	T	A	C	T	C	G	A	A	A	A	G	A	T	G
G	G	T	T	G	T	T	T	C	C	G	G	C	T	A	C	A	G	T	G	A	T	A	T	A	C	G	T	A	C	C	C	G	A	G	A	A	C	T	T	G
H	T	T	T	A	T	T	T	C	C	G	G	C	T	A	C	A	G	T	G	A	T	A	T	A	C	G	T	G	C	C	C	G	A	G	A	A	G	T	T	G

Bootstrap data set

	02	39	35	22	36	31	40	05	16	23	15	35	35	40	03	06	24	33	06	07	14	20	35	01	36	09	13	22	11	25	26	33	03	09	16	20	08	18	17	32
A	T	C	A	A	A	G	C	C	T	T	T	A	A	C	T	T	A	A	T	T	A	A	A	T	A	C	T	A	G	C	G	A	T	C	T	A	T	T	A	A
B	T	C	G	A	A	G	A	C	G	T	T	G	G	A	T	T	A	A	T	T	T	A	G	T	A	T	T	A	G	C	A	A	T	T	G	A	T	T	A	A
C	T	C	A	A	A	A	A	C	T	T	C	A	A	A	T	T	A	A	T	T	A	A	A	T	A	T	T	A	G	T	A	A	T	T	T	A	C	C	A	G
D	T	C	A	A	A	T	G	C	A	T	C	A	A	G	T	T	A	A	T	T	A	G	A	T	A	C	T	A	A	C	G	A	T	C	A	G	C	A	A	G
E	T	C	A	A	A	T	G	C	A	T	C	A	A	G	T	C	A	A	C	T	G	A	A	C	A	C	T	A	G	C	G	A	T	C	A	A	A	T	A	A
F	T	T	A	A	A	C	G	T	A	T	C	A	A	G	C	C	A	A	C	C	A	G	A	T	A	C	T	A	G	T	G	A	C	C	A	G	C	T	A	G
G	T	T	A	A	A	C	G	T	A	T	C	A	A	G	T	T	A	A	T	T	A	A	A	G	A	C	T	A	G	C	G	A	T	C	A	A	C	T	G	G
H	T	T	A	A	A	C	G	T	A	T	C	A	A	G	T	T	A	A	T	T	A	A	A	T	A	C	T	A	G	C	G	A	T	C	A	A	C	T	G	G

第五章 分子系统发生分析

- 什么是分子系统发生分析？
- 基本概念

  同源性（*直系同源）

  类群　（*单系类群）

  系统发生树
- 基本原理

  置换模型

  序列分歧度　}（*置换和相异性）

  进化速率
- 分子系统树的构建方法（*最大似然法）
- 分子系统发生分析软件

分子系统发生分析软件包

- PHYLIP：经典软件，支持多种操作系统，功能强大（包含几十个应用程序），运行速度快；但用户界面简单（命令行），特别是在使用多个不同程序时手工操作较多。
- PAUP (Phylogenetic Analysis Using Parsimony)
- MEGA (Molecular Evolutionary Genetics Analysis)
- PHYML、PAML、BEAST、MrBAYES……