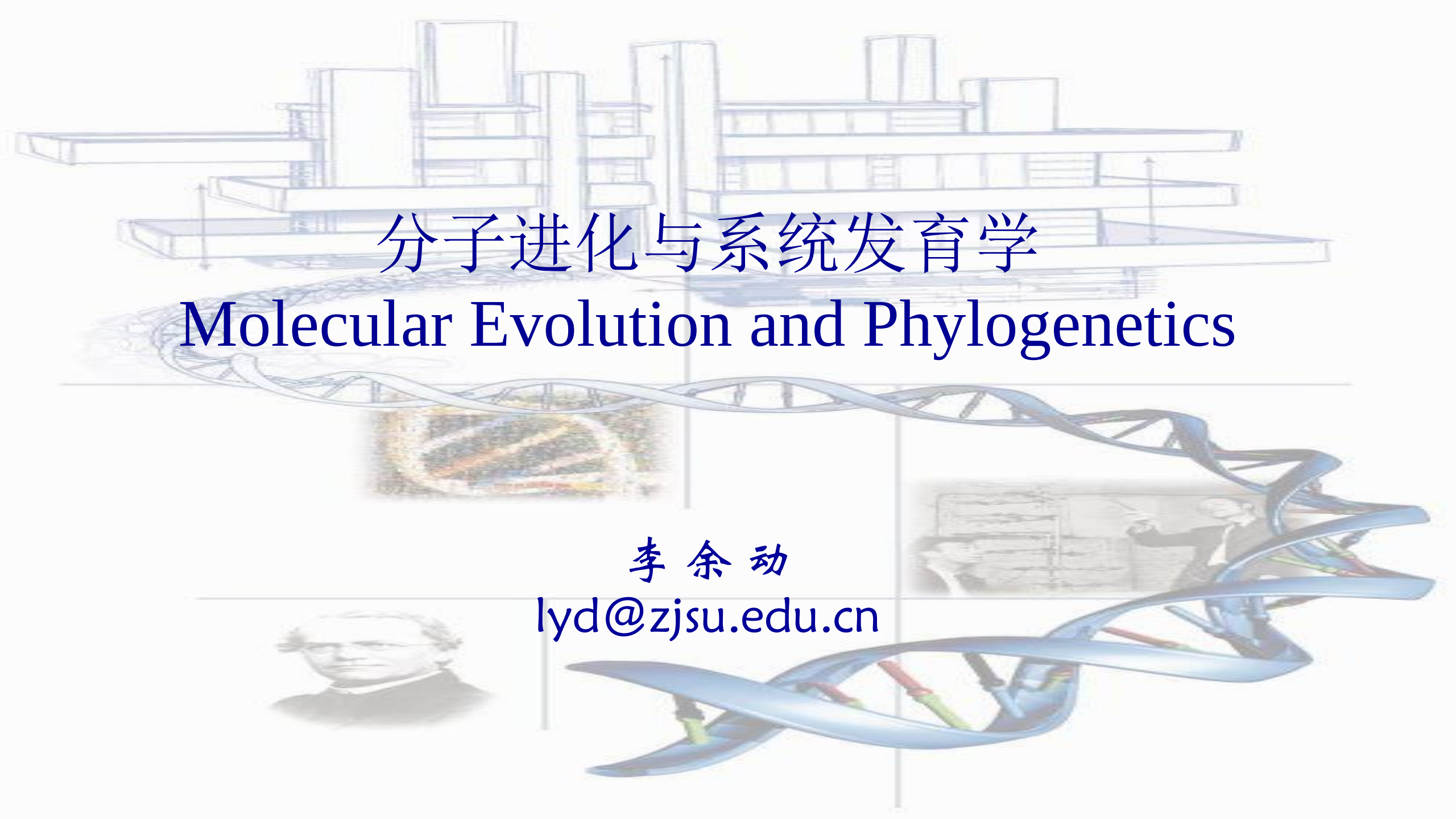
分子进化与系统发育学
Molecular Evolution and Phylogenetics



李 余 动
lyd@zjsu.edu.cn

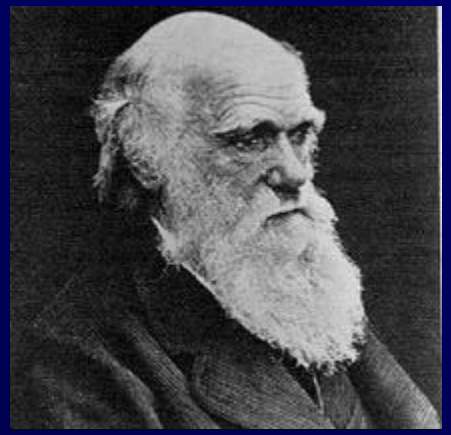
Outline

- 分子进化的基本概念
- 进化树的一些名词
- 系统发育树的构建方法
- MEGA构建简单进化树
  - 理解自举检验方法(bootstrapping)

达尔文与自然选择学说

*"Natural selection is daily, hourly, scrutinising the **slightest variations**, rejecting those that are bad, preserving and adding up all those that are good."*

- The Origin of Species



Darwin, Charles (1809-1882)

物竞天择
适者生存

现代进化研究方法－分子进化

1. 1964年，Linus Pauling提出分子进化理论:
   - 从物种的一些分子(DNA, RNA和蛋白质)特征出发，从而了解物种之间的生物系统发生的关系。
2. 发生在分子层面的进化过程：
   - DNA & RNA: 4种碱基；
   - 蛋白质分子：20种氨基酸

- Darwin's comparison of morphological features of the Galapagos finches led him to postulate the theory of natural selection.
- When you compare the sequences of genes and proteins, you are performing the same type of analysis, just at another level.

分子进化

分子进化

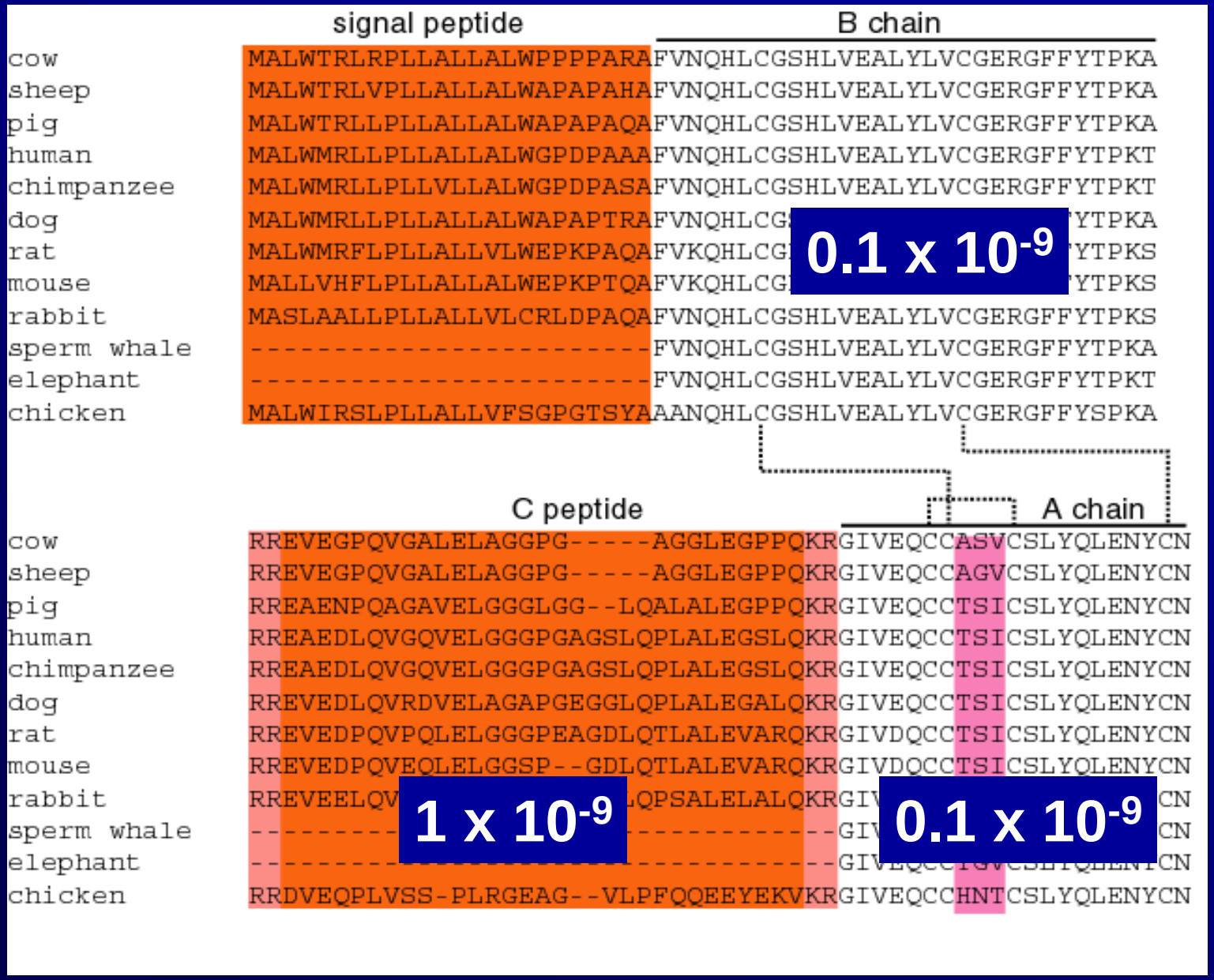
生物大分子进化

研究基因和蛋白质的进化速率及其变异模式

分子系统发育学

研究进化树构建方法及推断基因或生物体的进化历史

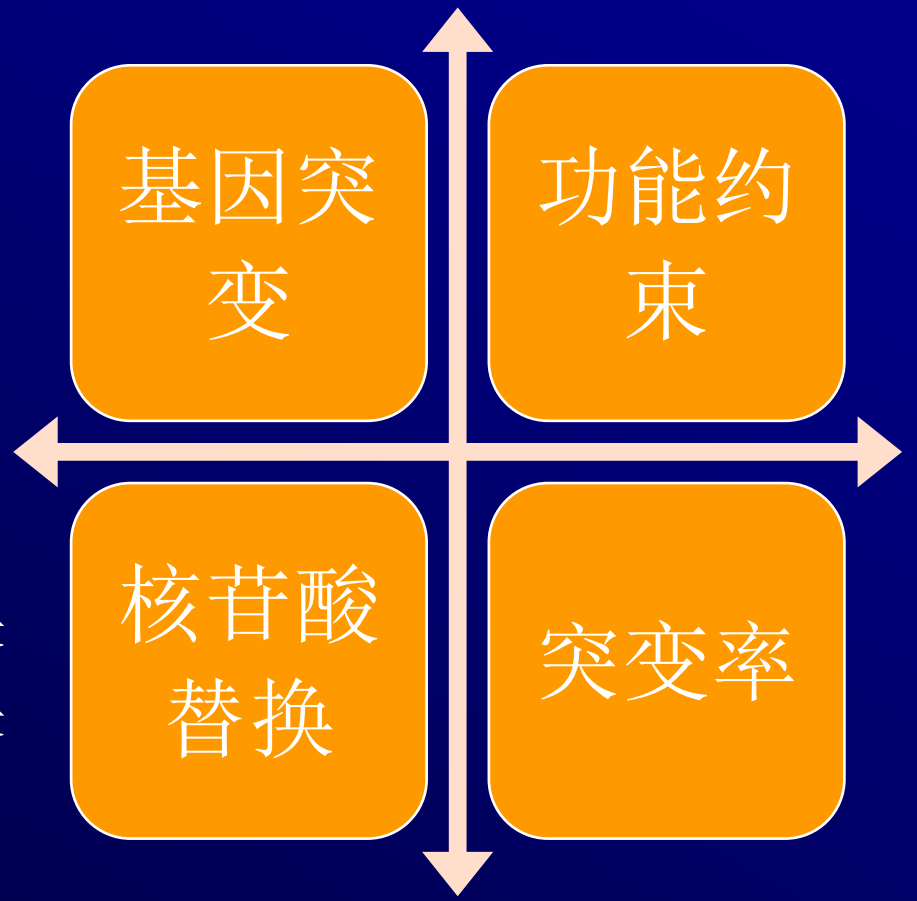
研究生物进化过程中核酸和蛋白质等生物分子的演化规律



Number of nucleotide substitutions/site/year

生物大分进化

替换
插入
缺失
倒位



转换
颠换

基因突变造成蛋白质催化性能或结构特征的变化，受到自然选择约束

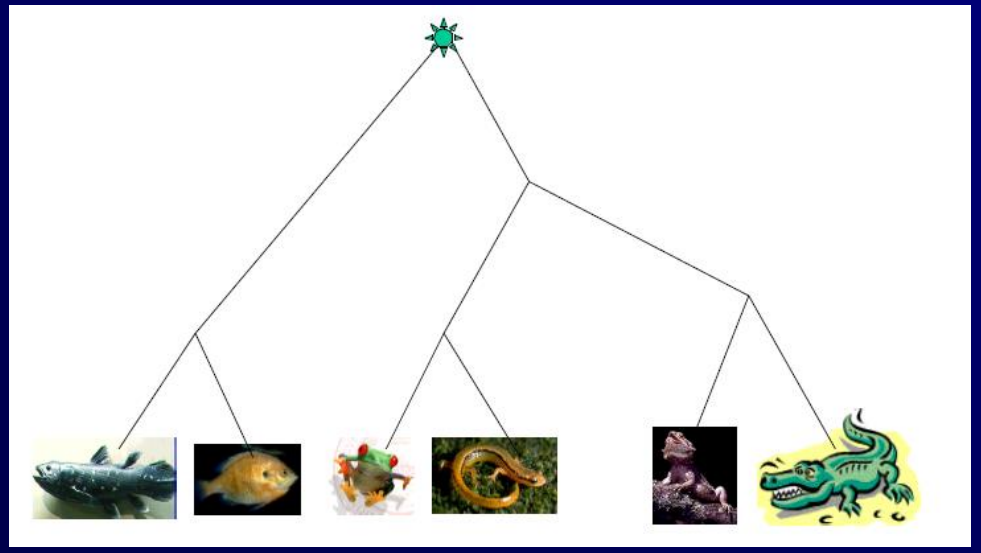
r=K/(2t)
K-遗传距离
t-同源序列从共同祖先分歧的时间
r-替换率

系统发育(phylogeny)

phylogeny is the inference of evolutionary relationships.

系统发育树：对一组实际对象（如基因，物种等）的世系关系的描述。

pattern and timing of evolutionary branching events (“evolutionary tree”)



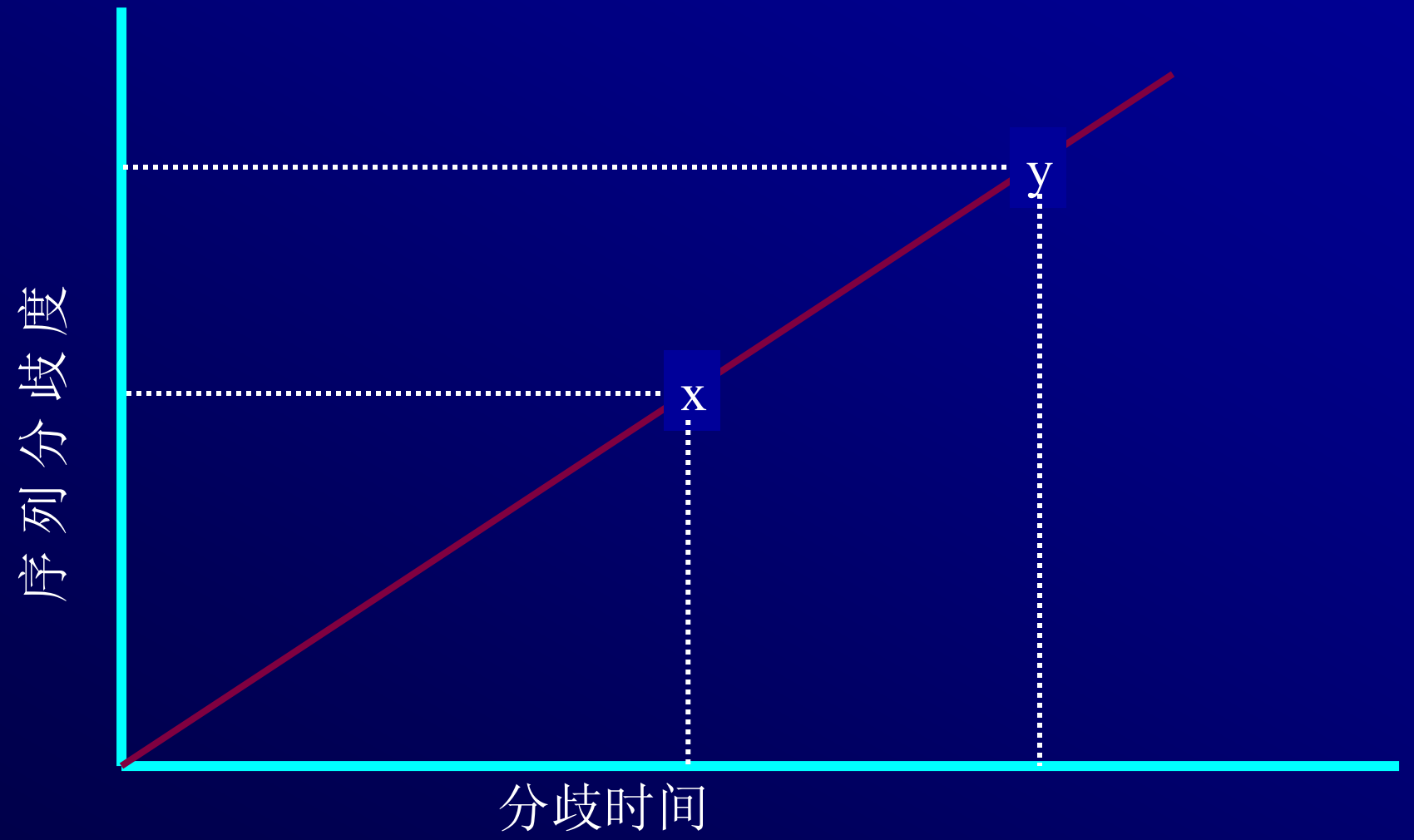
生物大分子用于研究系统发育的基本假设:

(1)生物分子包含了物种的进化历史信息；

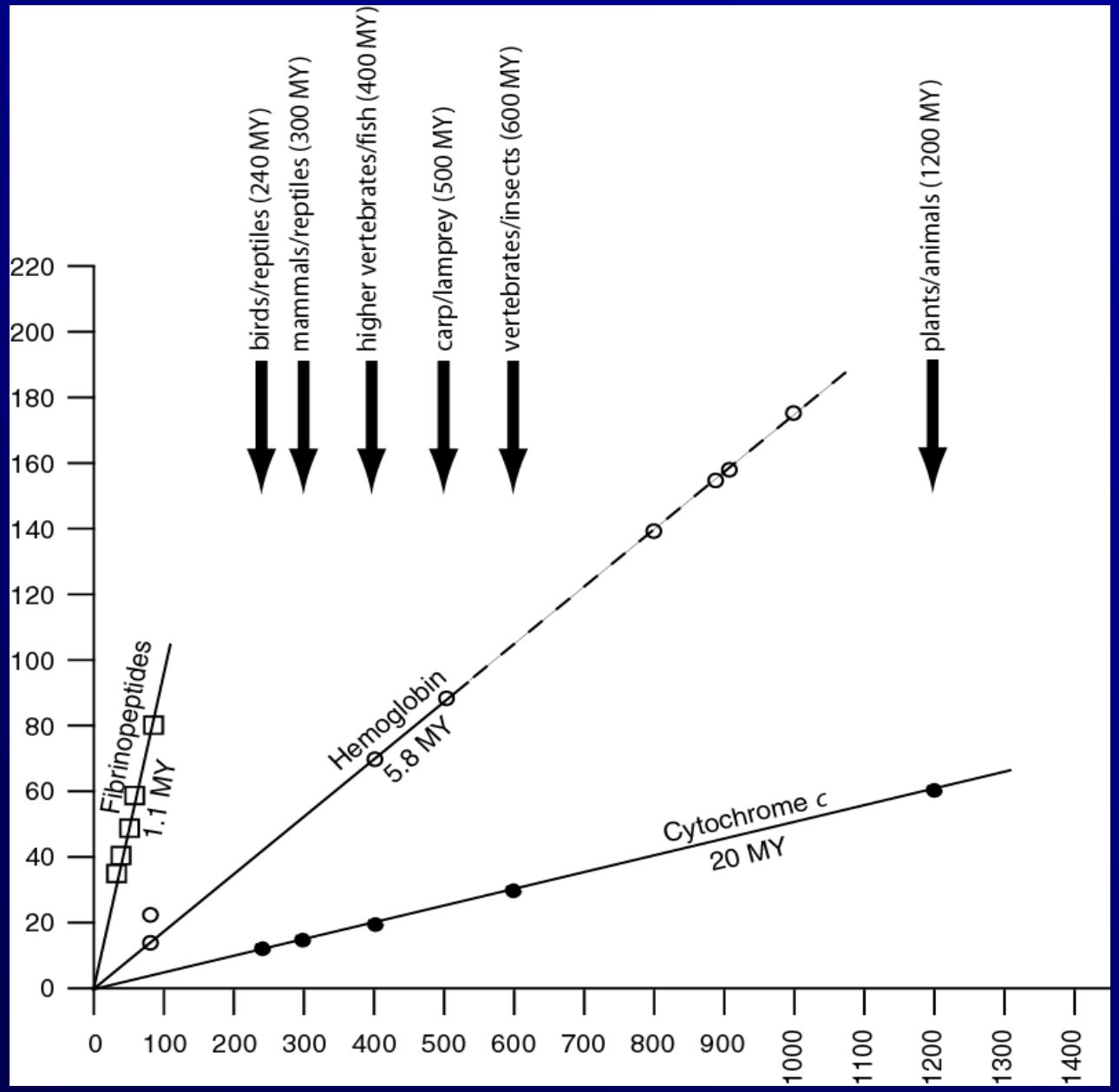
(2)分子钟理论。

分子钟(molecular clock)理论

在各种不同的发育谱系及足够大的进化时间尺度中，许多序列的进化速率几乎是恒定不变的。



corrected amino acid changes per 100 residues (*m*)



Millions of years since divergence

Dickerson (1971)

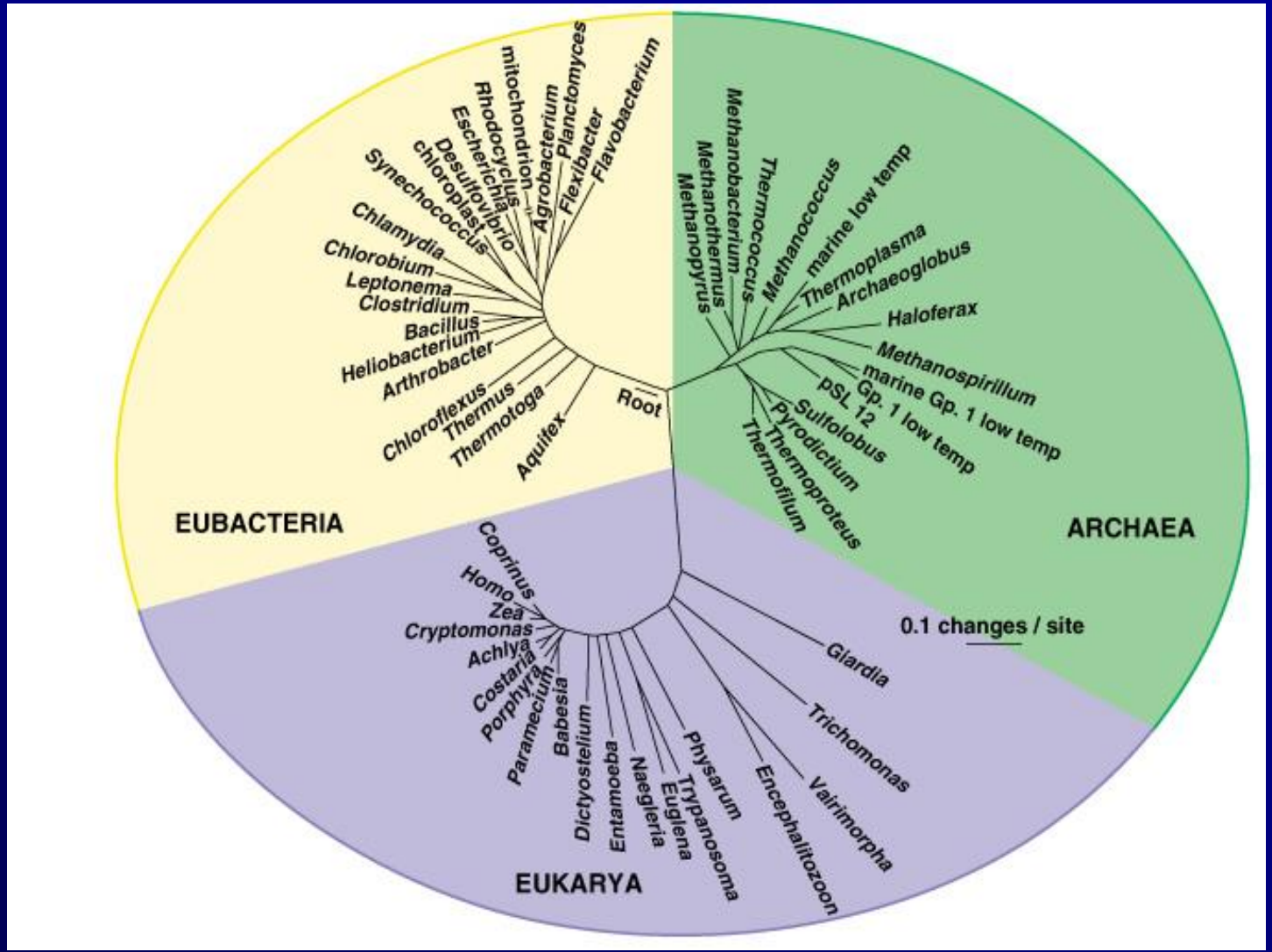
Tree of Life: 16/18S rRNA

生命三界：

- **细菌（Eubacteria）**
- **古细菌（Archaebacteria）**
- **真核生物（Eukaryotes）**

（Woese and Fox, 1977）

真核生物与细菌或古细菌哪个近？



中性进化学说

1968年，日本遗传学家木村资生(Motoo Kimura)提出了分子进化中性学说，向达尔文的自然选择学说提出了挑战。

“The theory states that most evolutionary changes at the molecular level are caused by random genetic drift of selectively neutral nucleotide substitutions.”

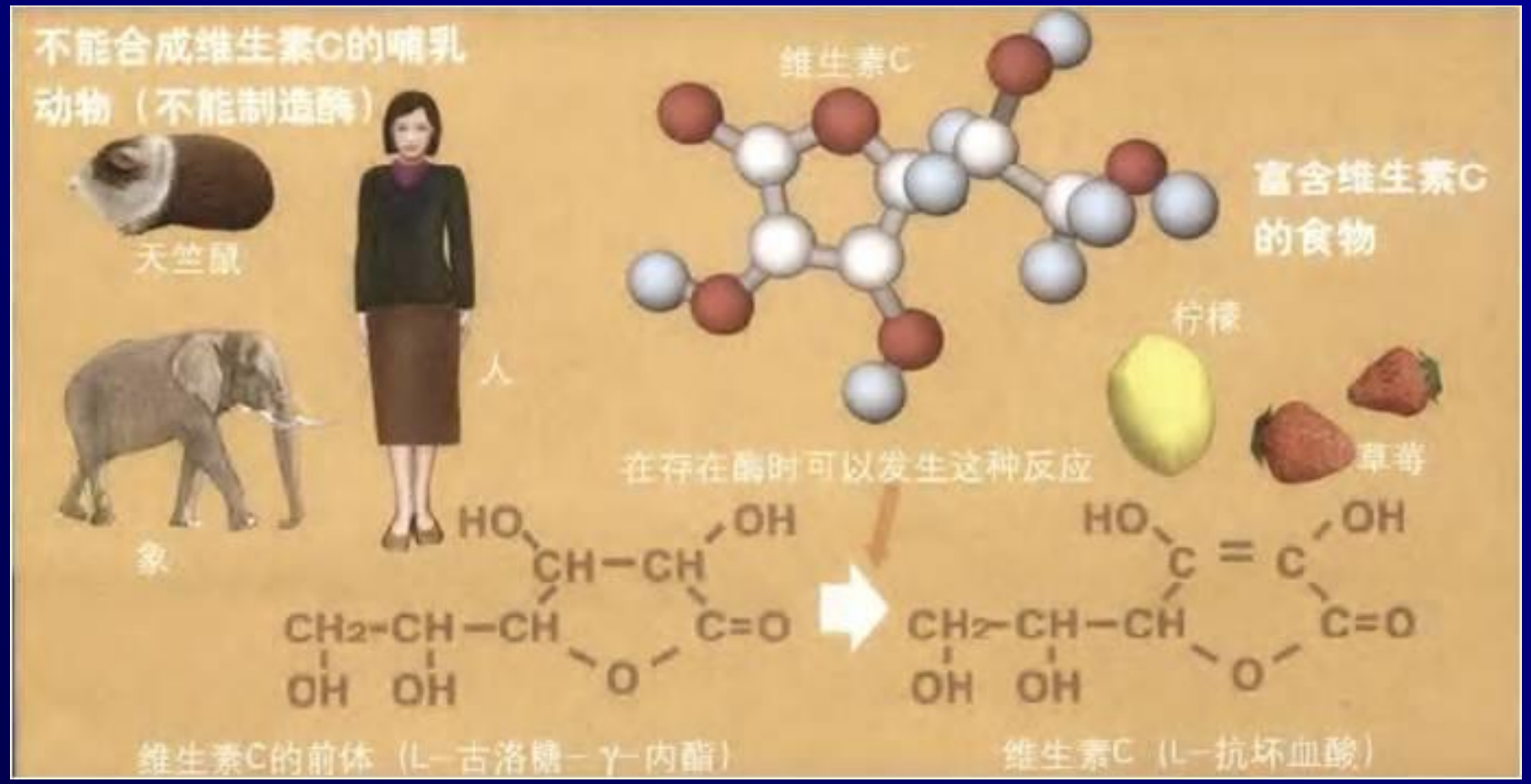
注：在小的种群中，基因频率可因**偶然**的机会，而不是由于选择而发生变化，这种现象称遗传漂变 (genetic drift)。

中性学说要点:

①认为生物进化的主导因素不是自然选择，而是不好不坏的中性选择；新种的形成主要不是由微小的长期有利变异积累而成，而是由那些无适应性的、无好坏利害之分的中性突变积累而成。

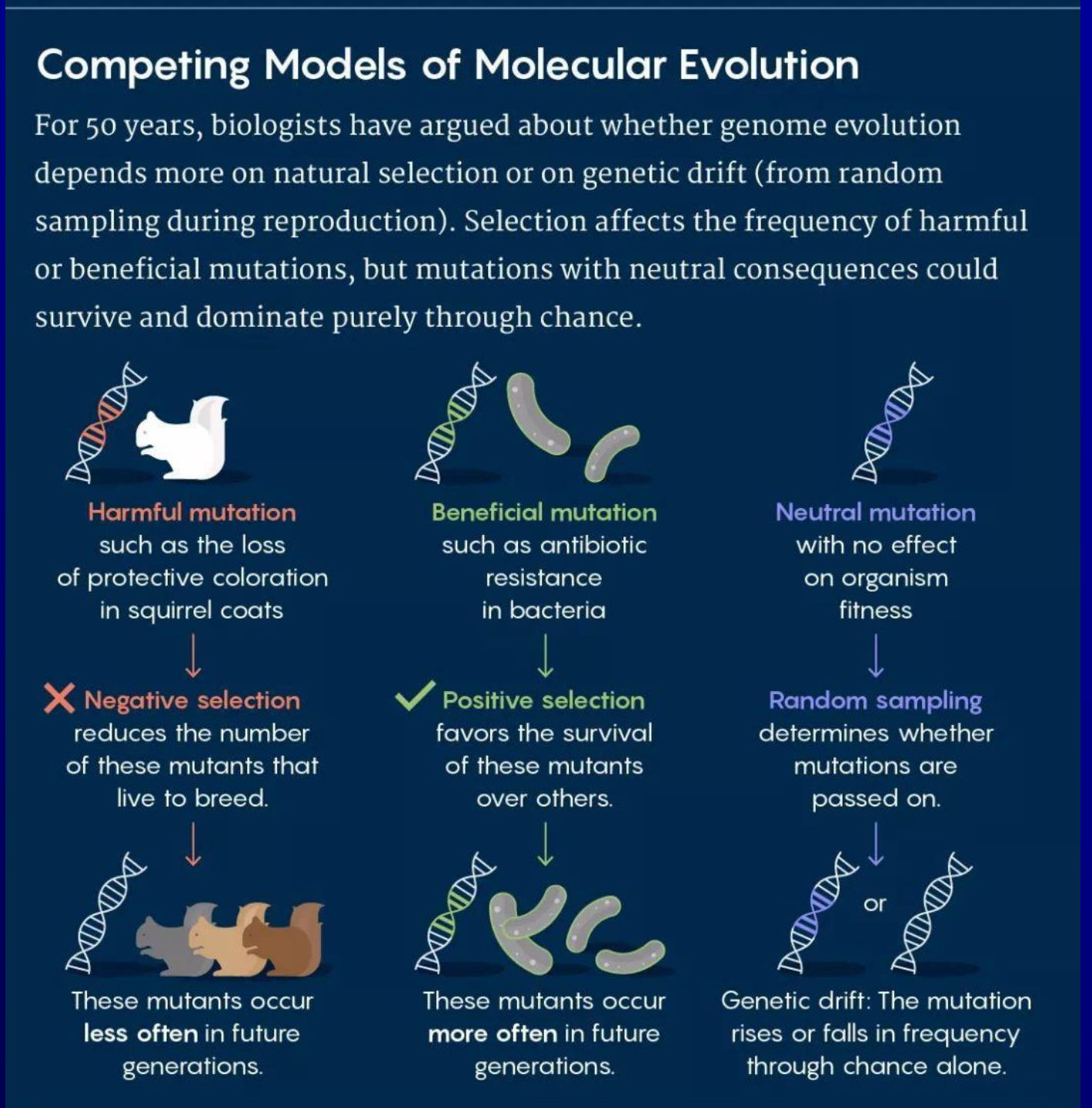
②中性突变通过遗传漂变而被固定下来或消失，由突变提供的进化原材料是偶然的，进化的途径和方向也在很大程度上由几率决定。

在人类祖先的生活环境，生成酶基因是中性基因。那种基因偶然出现了不能合成维生素C的突变，通过遗传漂变，最后在种群中扩散，稳定下来。



分子进化模型

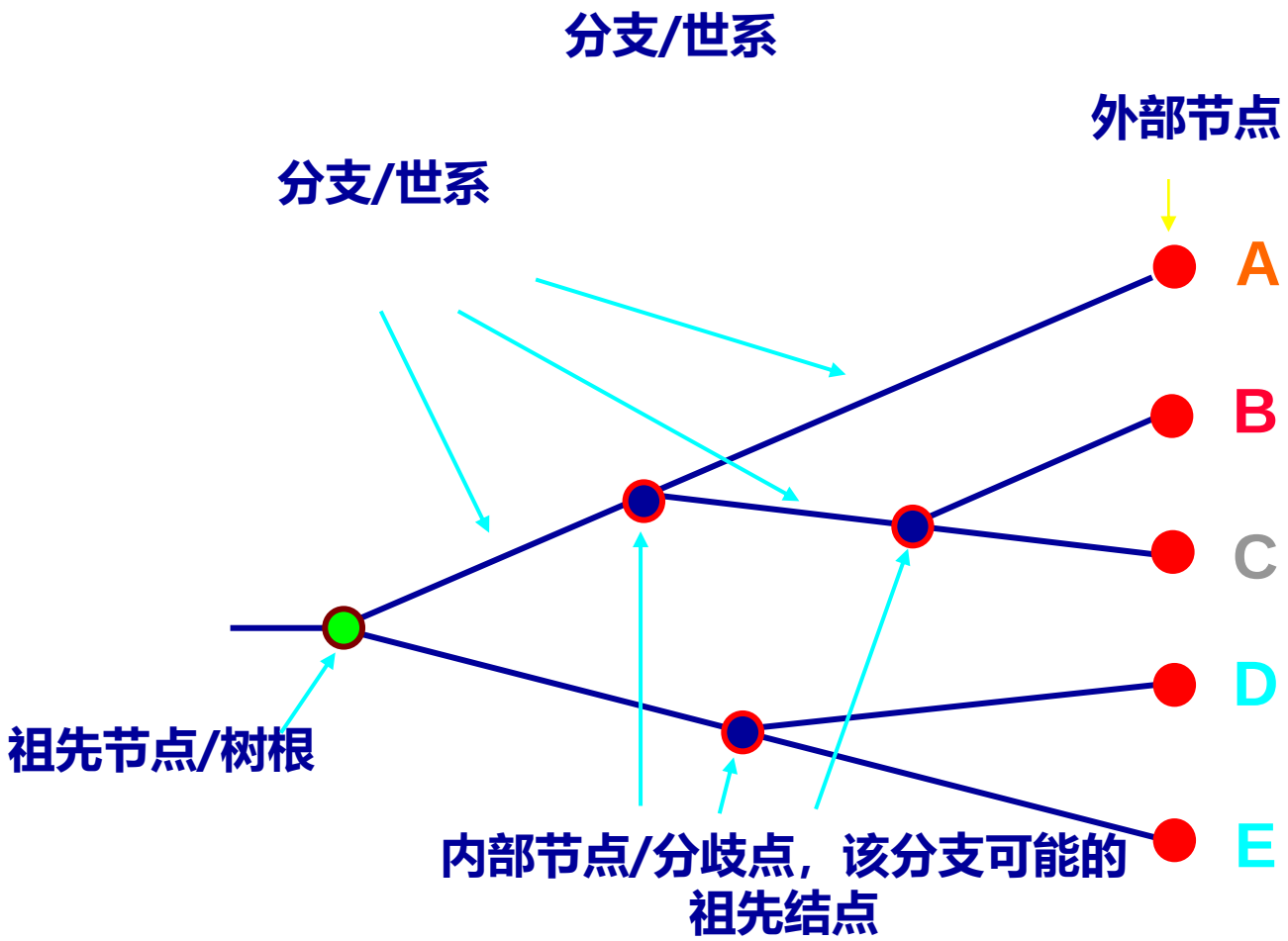
- 负选择(Negative selection)，也称为净化选择（purifying selection）
- 正选择（Positive selection）
- 遗传漂变(Genetic drift)



Outline

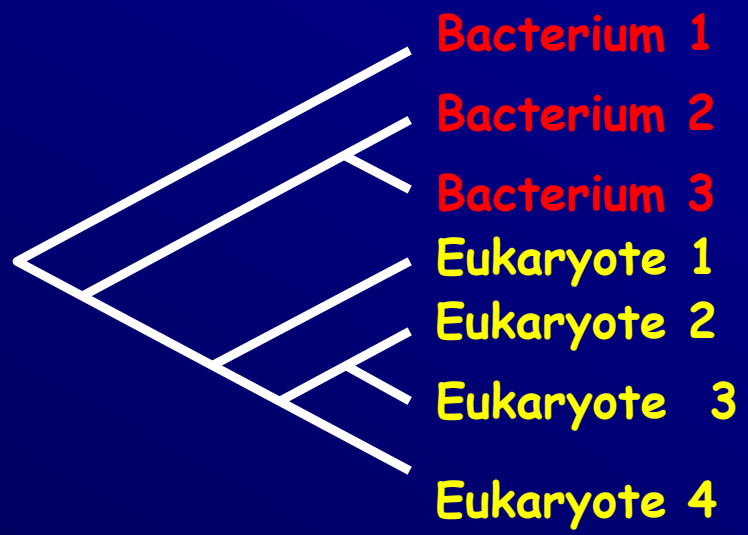
- 分子进化的基本概念
- 进化树的一些名词
- 系统发育树的构建方法
- 构建简单进化树
  - 理解自举检验方法(bootstrapping)

系统发育树：Phylogenetic tree



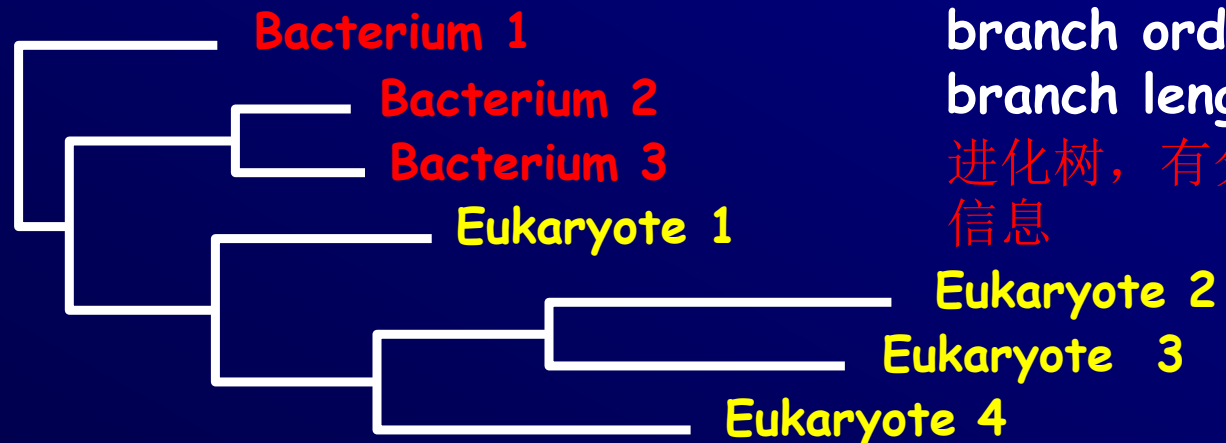
代表最终分类，可以是物种，群体，或者蛋白质、DNA、RNA分子等

进化分支图和进化树



Cladograms show branching order - branch lengths are meaningless

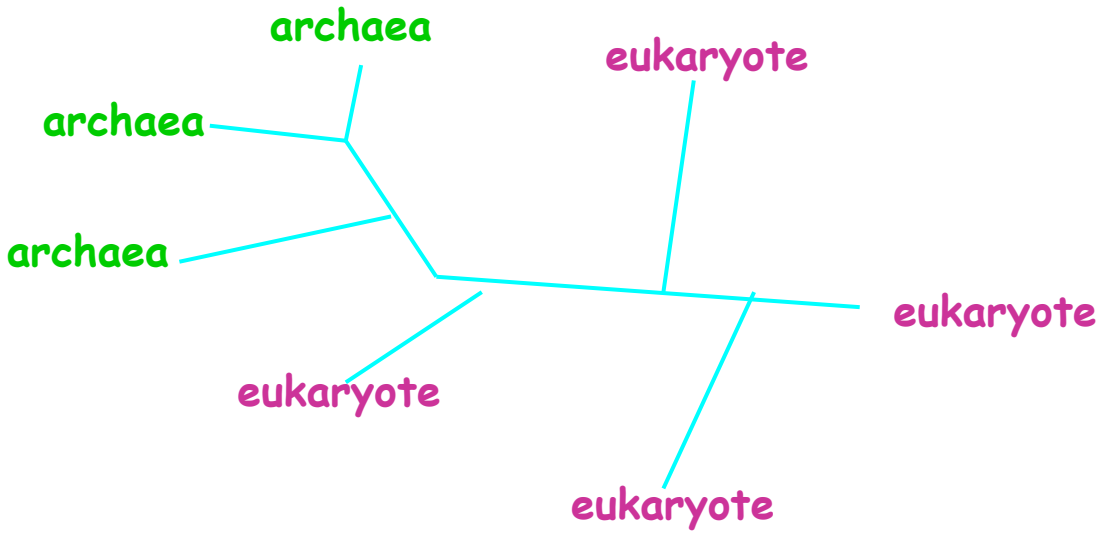
进化分支图，只用分支信息，无支长信息。



Phylograms show branch order and branch lengths

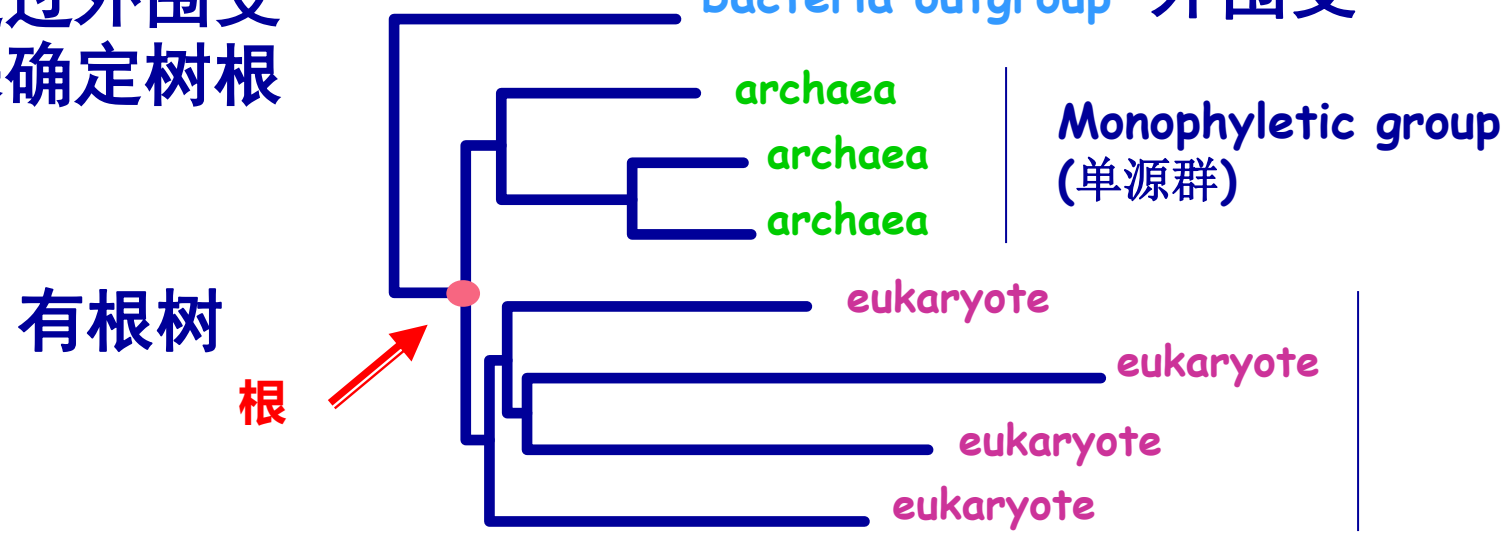
进化树，有分支和支长信息

无根树，有根树，外围支



无根树

通过外围支
来确定树根



外围支

Monophyletic group
(单源群)

有根树

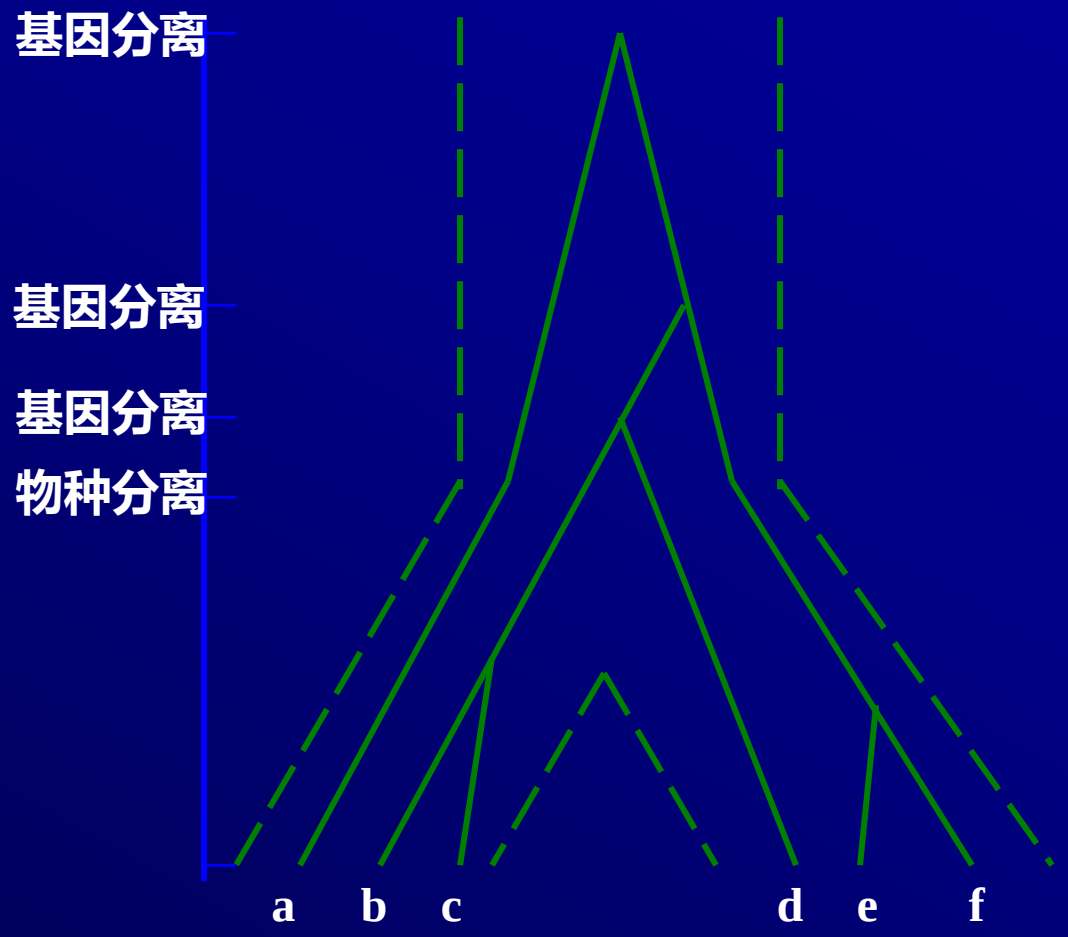
基因树与物种树

基因树

基于单个同源基因差异构建的系统发育树

物种树

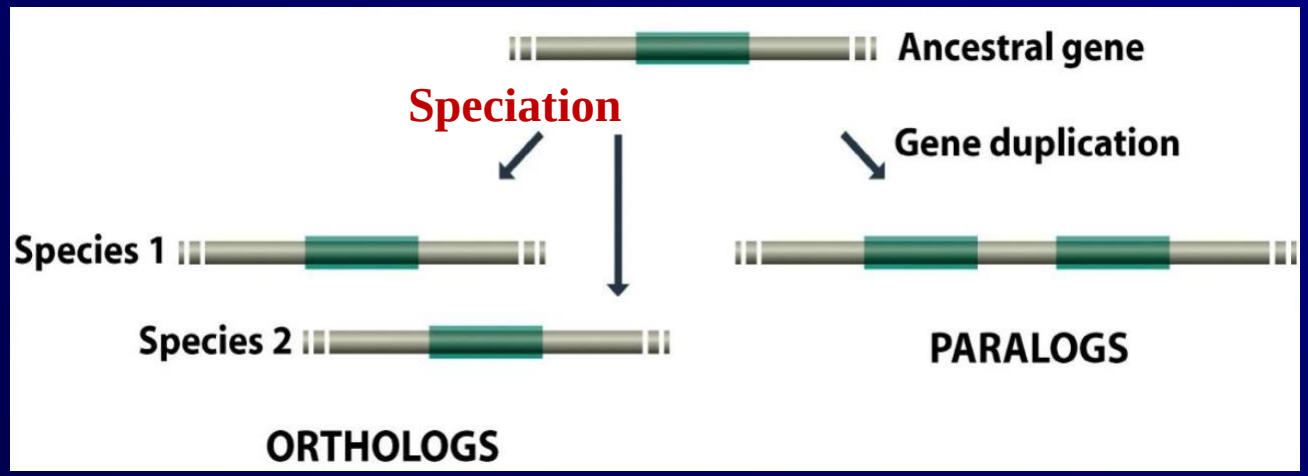
代表一个物种或种群进化历史的系统发育树



同源基因(Homolog)分类

- 直系同源(orthologs): 同源的基因是由于共同的祖先基因进化而产生的.
- 旁系同源(paralogs): 同源的基因是由于基因复制产生的.

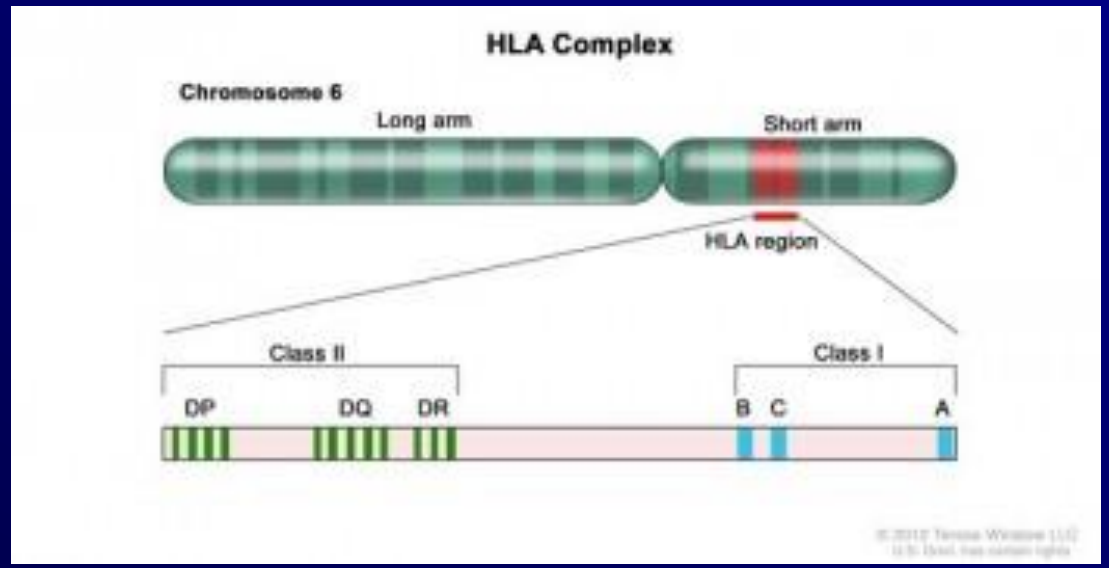
(以上定义源自Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113)



直系同源基因往往具有相似的功能，而旁系同源基因往往会发生功能变化。

HLA基因的多态性

- 人类白细胞抗原(human leukocyte antigen, HLA)是人类的主要组织相容性复合体（MHC）的表达产物，该系统是所知人体最复杂的多态系统。
- HLA基因的多态性起源先于物种分歧，如只用HLA基因构建物种树，有些人类个体可能会与大猩猩分类在一起，而不是和其他人类个体分类在一起。



Outline

- 进化的基本概念
- 进化树的一些名词
- 系统发育树的构建方法
- MEGA构建简单进化树
  - 理解自举检验方法(bootstrapping)

How to construct a tree: A simple example

A. Sequences

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

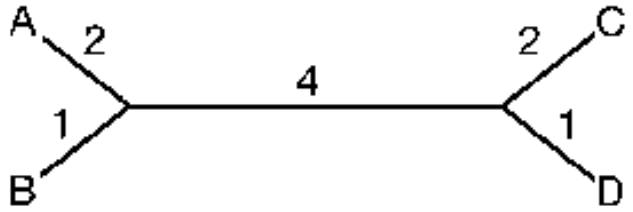
B. Distances between sequences, the number of steps required to change one sequence into the other.

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

C. Distance table

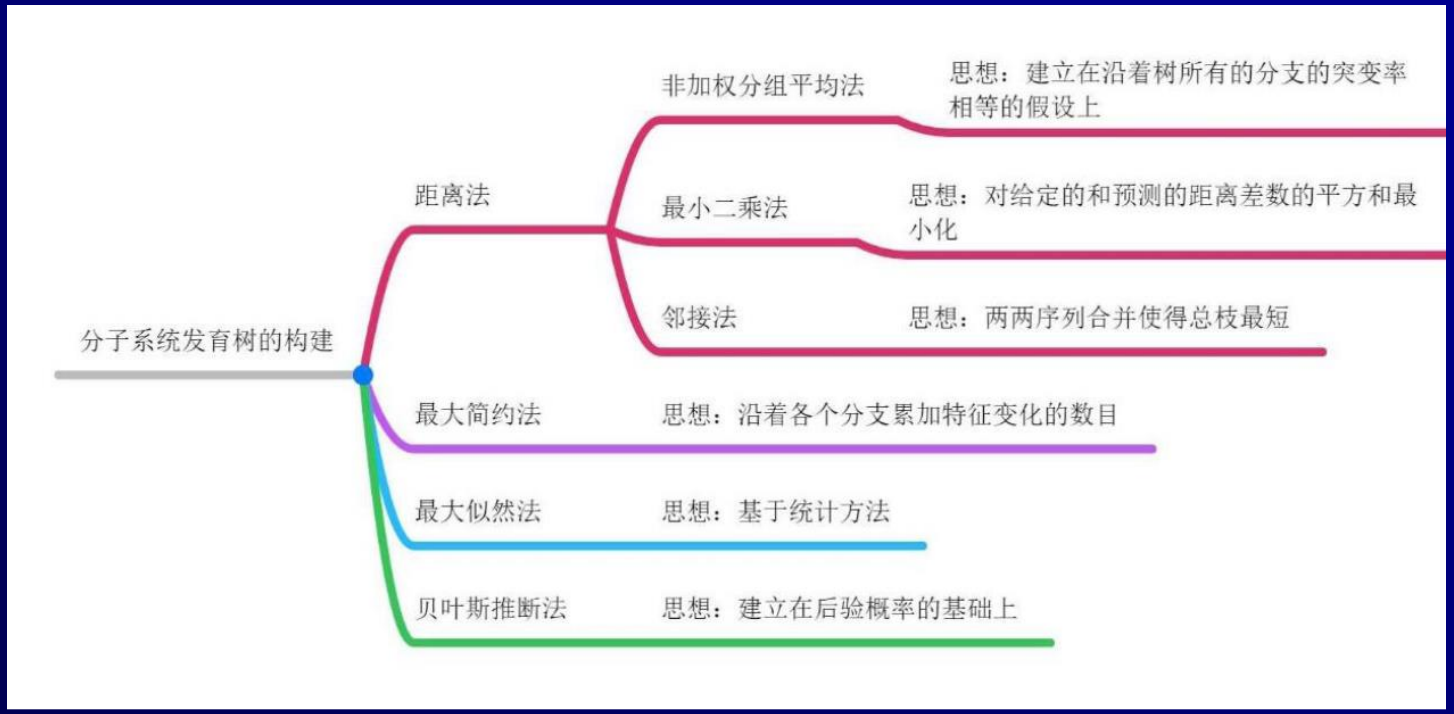
	A	B	C	D
A	–	3	7	8
B	–	–	6	7
C	–	–	–	3
D	–	–	–	–

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.

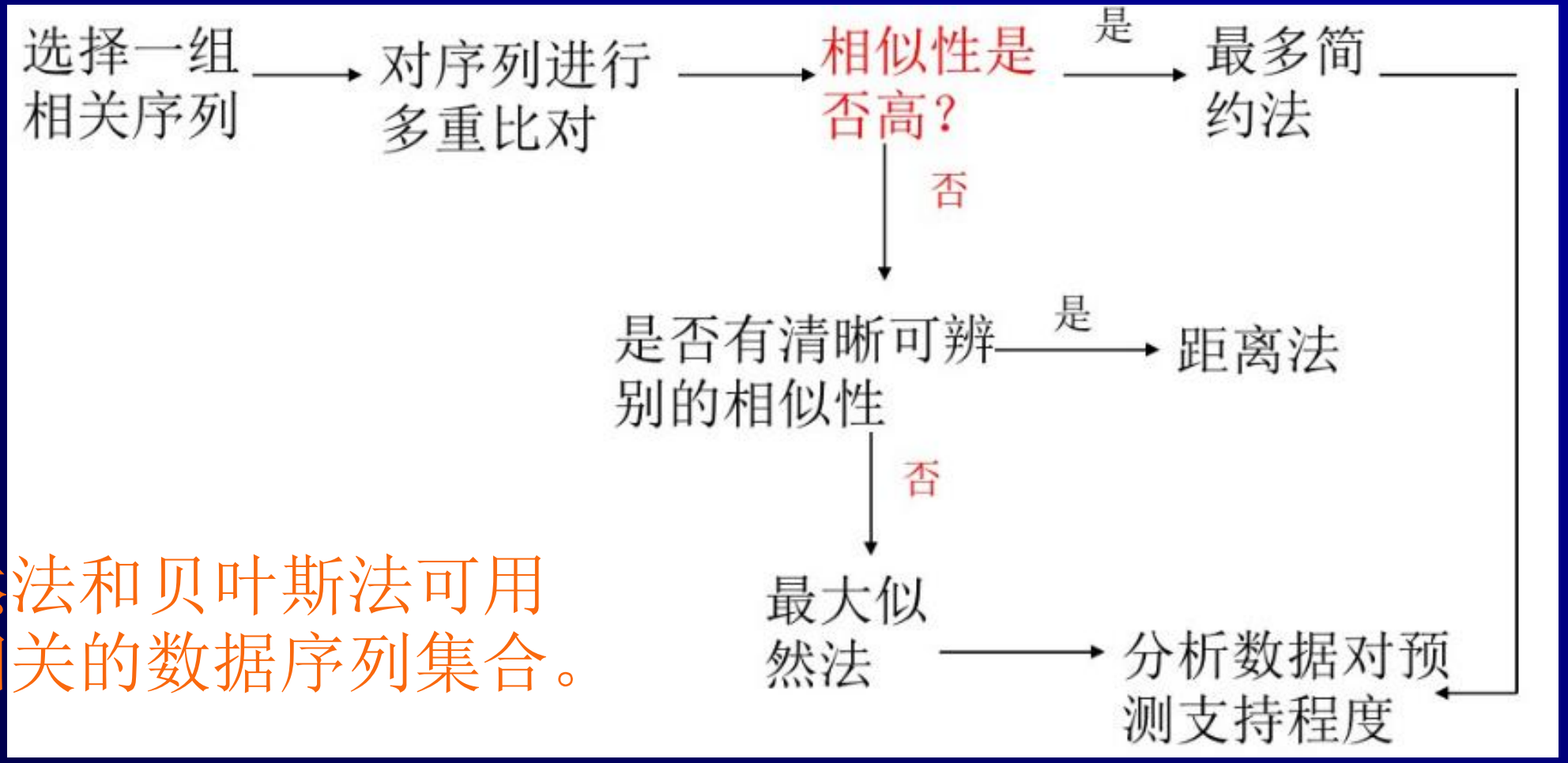


系统发育树重建的基本方法

- 最大简约法(maximum parsimony, MP
- 距离法中的邻接法(neighbor joining, NJ)
- 最大似然法(maximum likelihood, ML)
- 贝叶斯推论法(Bayesian inference, BI)



进化树构建方法选择

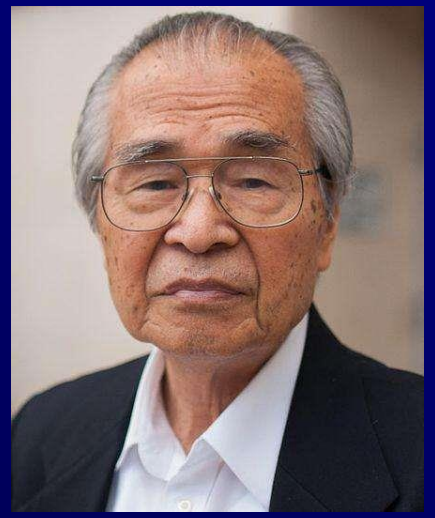


最大似然法和贝叶斯法可用于任何相关的数据序列集合。

邻接法

最小进化（ME）思想：在所有可能的拓扑结构中，选择分支长度和S最小作为最优树。（全局优化思想）(Edwards & Cavalli-Sforza,1963)

Saitou & Nei (1987)：在每一阶段应用最小进化原理，是ME方法的简化。



Masatoshi Nei（根井正利, 1931-2023）

邻接法构建系统进化树

构建距离矩阵

	B	C	D	E
A	11	12	17	24
B		9	16	24
C			16	24
D				24

$$S_X = \sum_{i=1}^{N} d_{xi}$$

S_A = 11+12+17+24=64

同理可得S_B=60 S_C=61

S_D=73 S_E=96

δij=dij-（Si + Sj）/（N-2）

S_{AB} = 11-（64+60）/3 = -30.3

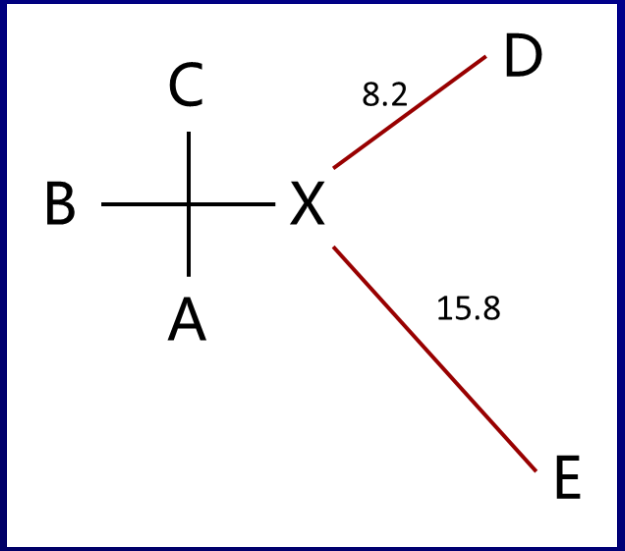
同理可得S_{AC} S_{AD} S_{AE}

得到矩阵

	B	C	D	E
A	-30.3	-29.7	-28.7	-29.3
B		-29.7	-28.3	-28
C			-28.7	-28.3
D				-32.3

最小

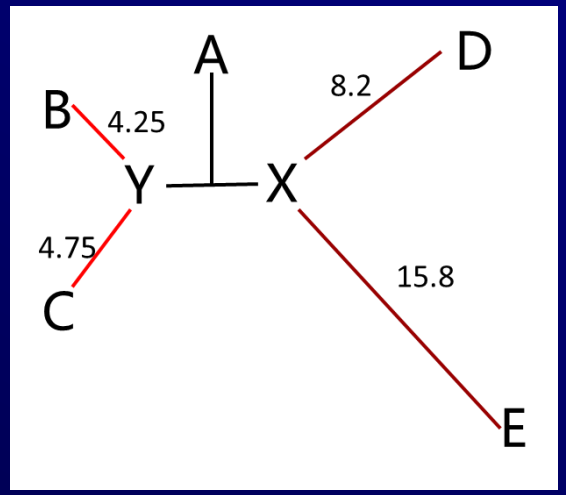
$$d_{DX} = \frac{d_{DE} + \frac{S_D - S_E}{N-2}}{2} = \frac{24 + \frac{73-96}{3}}{2} = 8.2$$



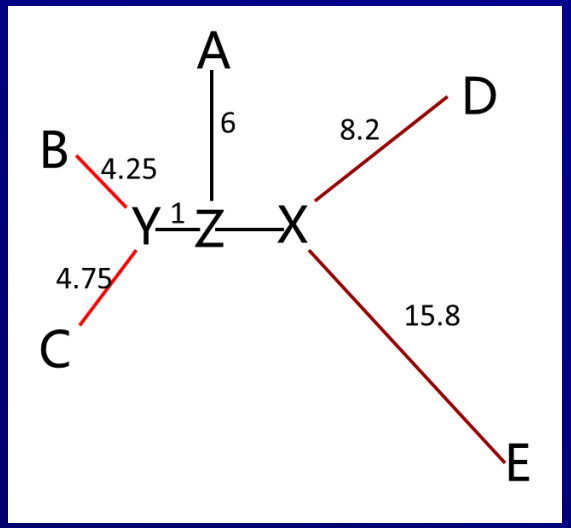
建立一个用X替代了D和E的矩阵

	B	C	X
A	11	12	8.5
B		9	8
C			8

新的节点Y



新的节点Z



Outline

- 理解进化的基本概念
- 理解进化树的一些名词
- 系统发育树的构建方法
- MEGA构建简单进化树
  - 理解自举检验方法(bootstrapping)

系统发育树重建分析步骤

序列准备

多序列比对（自动比对，手工校正）

建立取代模型（建树方法）

建立进化树

进化树评估

序列数据的准备

- 进化树的质量依赖于序列数据的质量：
  - Garbage in ⇔ garbage out

- Most phylogenetic methods work on Proteins and DNA sequences
  - If your DNA sequences are coding and have more than 70% identity . . .
    - Compute the **tree** on the **DNA** multiple-sequence alignment
  - If your DNA sequences are coding and have less than 70% identity . . .
    - Compute the **tree** on the **protein** multiple-sequence alignment

1.基因序列获取

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☐	insulin preproprotein [Homo sapiens]	223	223	100%	5e-77	100.00%	NP_000198.1
☐	insulin, isoform 2 precursor [Homo sapiens]	112	112	56%	6e-32	100.00%	NP_001035835.1
☐	insulin-like growth factor II isoform 1 preproprotein [Homo sapiens]	49.7	49.7	93%	6e-08	33.98%	NP_000603.1
☐	insulin-like growth factor II isoform 2 [Homo sapiens]	50.1	50.1	93%	8e-08	33.98%	NP_001121070.1
☐	insulin-like growth factor I isoform 2 precursor [Homo sapiens]	47.0	47.0	74%	4e-07	34.15%	NP_001104754.1

BLAST检索，挑选至少五个不同物种的胰岛素蛋白质序列

选择第一个人的胰岛素序列

insulin [Homo sapiens]

GenBank: AAA59172.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       AAA59172                  110
DEFINITION  insulin [Homo sapiens].
ACCESSION   AAA59172
VERSION     AAA59172.1
```

FASTA序列文件

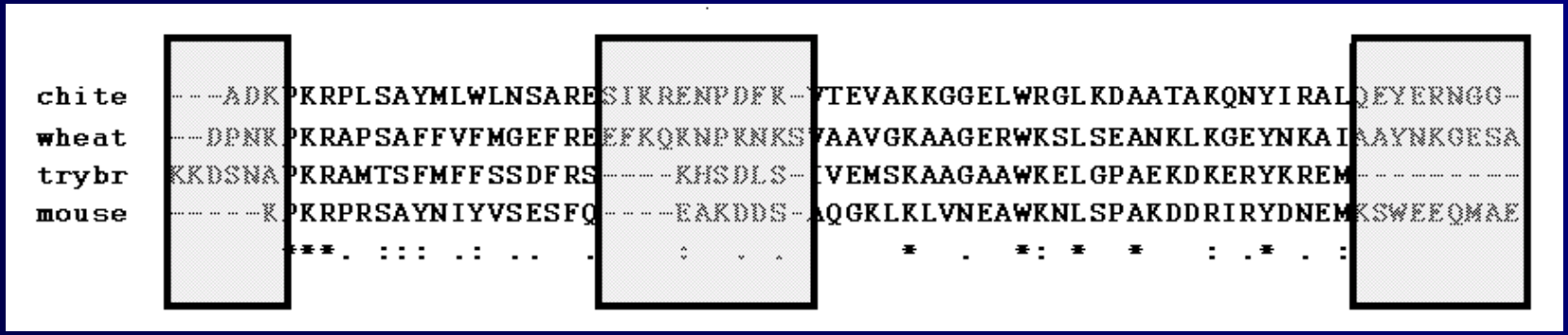
homologene.txt - Notepad

File Edit Format View Help

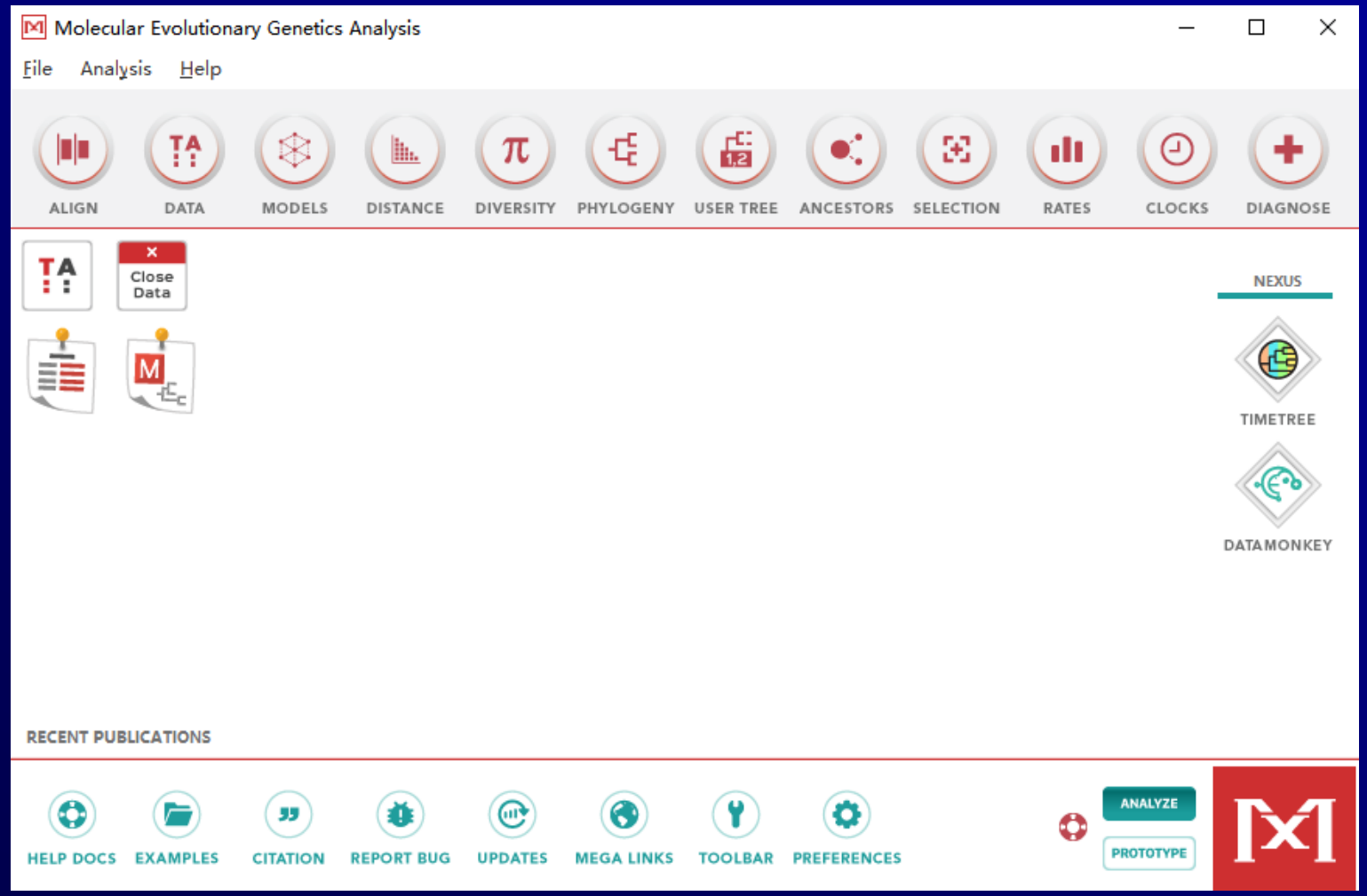
```
>gi|7661534|ref|NP_054862.1| CD274 molecule [Homo sapiens]
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHG
EEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNAPYNKINQR
ILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVTSTLRINTTTNEIFYCT
FRRLDPEENHTAELVIPELPLAHPPNERTHLVILGAILLCLGVALTFIFRLRKGRMMDVKKCGIQDTNSK
KQSDTHLEET
>gi|114623690|ref|XP_001140705.1| PREDICTED: CD274 antigen isoform 2 [Pan troglodytes]
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHG
EEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNAPYNKINQR
ILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVTSTLRINTTTNEIFYCT
FRRLDPEENHTAELVIPELPLAHPPNERTHLVILGAILLCLGVALTFIFCLRKGRMMDVKKCGIQDTNSK
KQSDTHLEET
>gi|73946918|ref|XP_541302.2| PREDICTED: similar to CD274 antigen [Canis familiaris]
MRMFSVFTFMAYCHLLKAFTITVSKDLYVVEYGGNVTMECKFPVEKQLNLFALIVYWEMEDKKIIQFVNG
KEDLKVQHSSYSQRAQLLKDQLFLGKAALQITDVRLQDAGVYCCLIGYGGADYKRITLKVHAPYRNISQR
ISVDPVTSEHELMCQAEGYPEAEVIWTSSDHRVLSGKTTITNSNREEKLFNVTSTLNINATANEIFYCTF
QRSGPEENNTAELVIPEVSHHSSRSLAGNFL
>gi|119900350|ref|XP_613366.3| PREDICTED: similar to programmed death ligand 1 [Bos taurus]
MECQAFTITVSKDLYVVEYGSNVTLECRFPVDKQLNLLVLVVYWEMEDKKIIQFVNGKEDPNVQHSSYHG
RAQLLKDQLFLGKAALQITDVKLQDAGVYCCLISYGGADYKRITLKVNAPYRKIYHTISVDPVTSEHELT
CQAEGYPEADVIWTSSDHQVLSGKTSITSSKREEKLFNVTSTLRINTTADKIFYCTFRRLGHEENNTAEL
VIPEPYLDPAKKRNHLVTLGALFLCLSVTLAVIFCLKRDVRMMDVEKCDTRDMNSKQQNDQRYAVGQGAA
DDGELKKPKLRKQKLRKKRRTKEEGIKVPWKETLVLPNGGRLINTCEKEKGHF
>gi|11230798|ref|NP_068693.1| CD274 antigen [Mus musculus]
MRIFAGIIFTACCHLLRAFTITAPKDLYVVEYGSNVTMECRFPVERELDLLALVVYWEKEDEQVIQFVAG
EEDLKPQHSNFRGRASLPKDQLLKGNAALQITDVKLQDAGVYCCIISYGGADYKRITLKVNAPYRKINQR
ISVDPATSEHELICQAEGYPEAEVIWTNSDHQPVSGKRSVTTSRTEGMLLNVTSSLRVNATANDVFYCTF
WRSQPGQNHTAELIIPELPATHPPQNRTHWVLLGSILLFLIVVSTVLLFLRKQVRMLDVEKCGVEDTSSK
NRNDTQFEET
>gi|109460012|ref|XP_574652.2| PREDICTED: similar to CD274 antigen [Rattus norvegicus]
MRIFAVLIVTACSHVLAAFTITAPKDLYVVEYGSNVTMECRFPVEQKLDLLALVVYWEKEDKEVIQFVEG
EEDLKPQHSSFRGRAFLPKDQLLKGNAVLQITDVKLQDAGVYCCMISYGGADYKRITLKVNAPYRKINQR
ISMDPATSEHELMCQAEGYPEAEVIWTNSDHQSLSGETTVTTSQTEEKLLNVTSVLRVNATANDVFHCTF
WRVHSGENHTAELIIPELPVPRLPHNRTHWVLLGSVLLFLIVGFTVFFCLRKQVRMLDVEKCGFEDRNSK
NRNVRGDVSSVEPSEPRGGIGSLWSVKERARGTWVQGLKNGTGEEKRTKKVLEEEPGTKDISTGDTAKQV
THQSSRAIS
>gi|118103980|ref|XP_424811.2| PREDICTED: similar to B7-H1 [Gallus gallus]
MMEKLLLLHIFLFCWRSLNALFTVEAPKSLYTAELGSNVTMECVFPVNGKLKFRDLSVIWEKKDEVRKDV
YILLKGKEDSGSQHSDFQGRIKLLKENLDFGQSLLQISNVKLRDAGLYHCLIEYGGADYKTINLKVQAPY
RTITQEVVSTGDKEWKLTCQSEGYPKAEVMWQNGECQDLTDKANTSYETGSDQLYRVTSTLTVKNRTCEN
FRCIFWNKEIQENTSANLYILDSADDVLWTESRRFVWPVLIVSALVGSVPITVCIRKARASKDCRTRMAK
SSIHITKLSKDKGAHDCRGPSFEDAELKYIQIETT
```

Building the Right MSA

- 构建进化树的很重要一个步骤是准备高质量的多序列比对数据（Multiple Sequence Alignment, **MSA**）
- Your MSA should have as few gaps as possible.
  - Adjust gap creation and extension penalties as needed to optimize the alignment

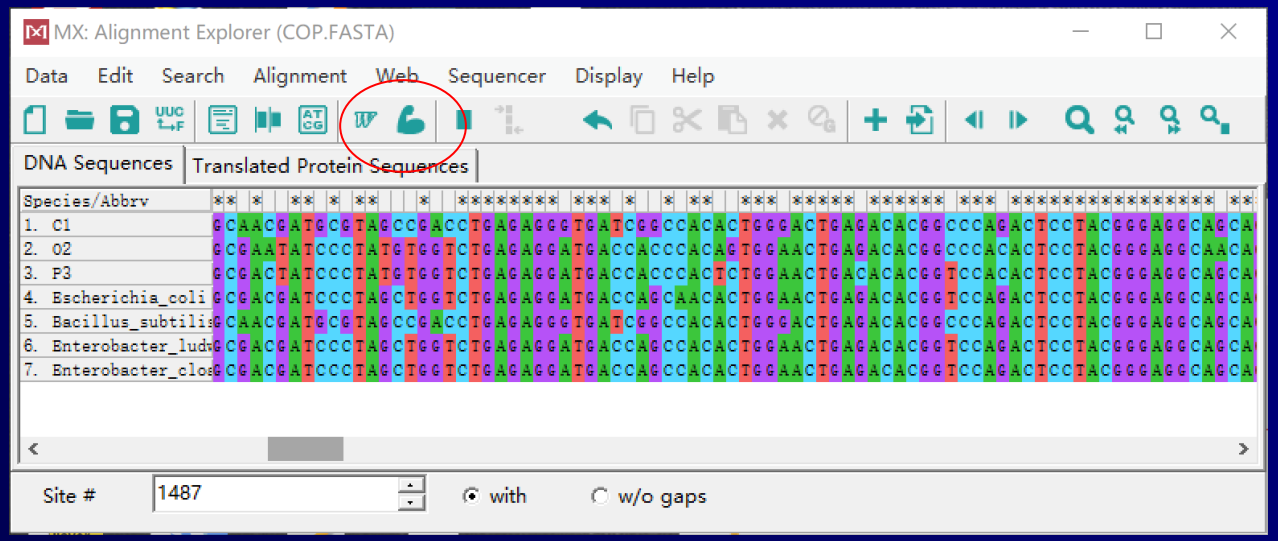


MEGA软件构建进化树



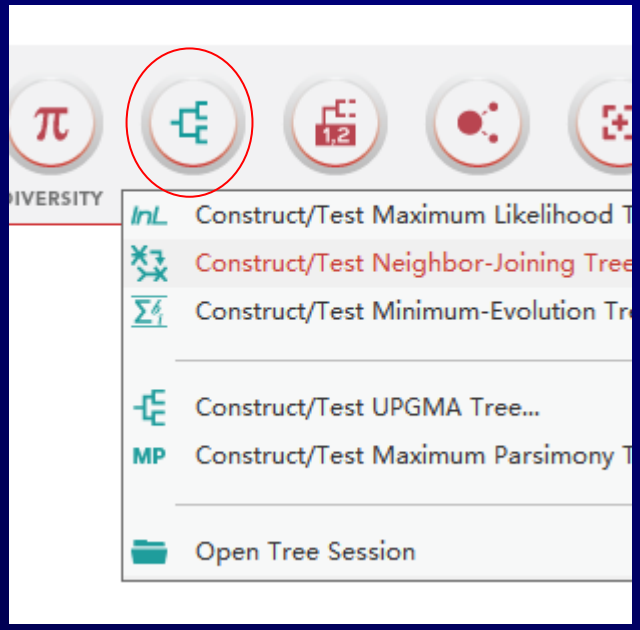
MEGA-X

2.多序列比对


MX: Alignment Explorer (COP.FASTA)
Data Edit Search Alignment Web Sequencer Display Help
DNA Sequences
Translated Protein Sequences
Species/Abbrv
1. C1
2. O2
3. P3
4. Escherichia_coli
5. Bacillus_subtilis
6. Enterobacter_lud
7. Enterobacter_clo
Site #
1487
with
w/o gaps


MX: Alignment Explorer (insulin_seq.meg)
Data Edit Search Alignment Web Sequencer Display Help
Protein Sequences
Species/Abbrv
1. AAA59172.1_insulin_Homo_sapiens
2. NP_000198.1_insulin_preproprotein
3. NP_001035835.1_insulin_isoform_2_
4. NP_000603.1_insulin-like_growth_f
5. NP_001121070.1_insulin-like_growt
6. NP_001104754.1_insulin-like_growt

3.构建进化树



MX: Analysis Preferences

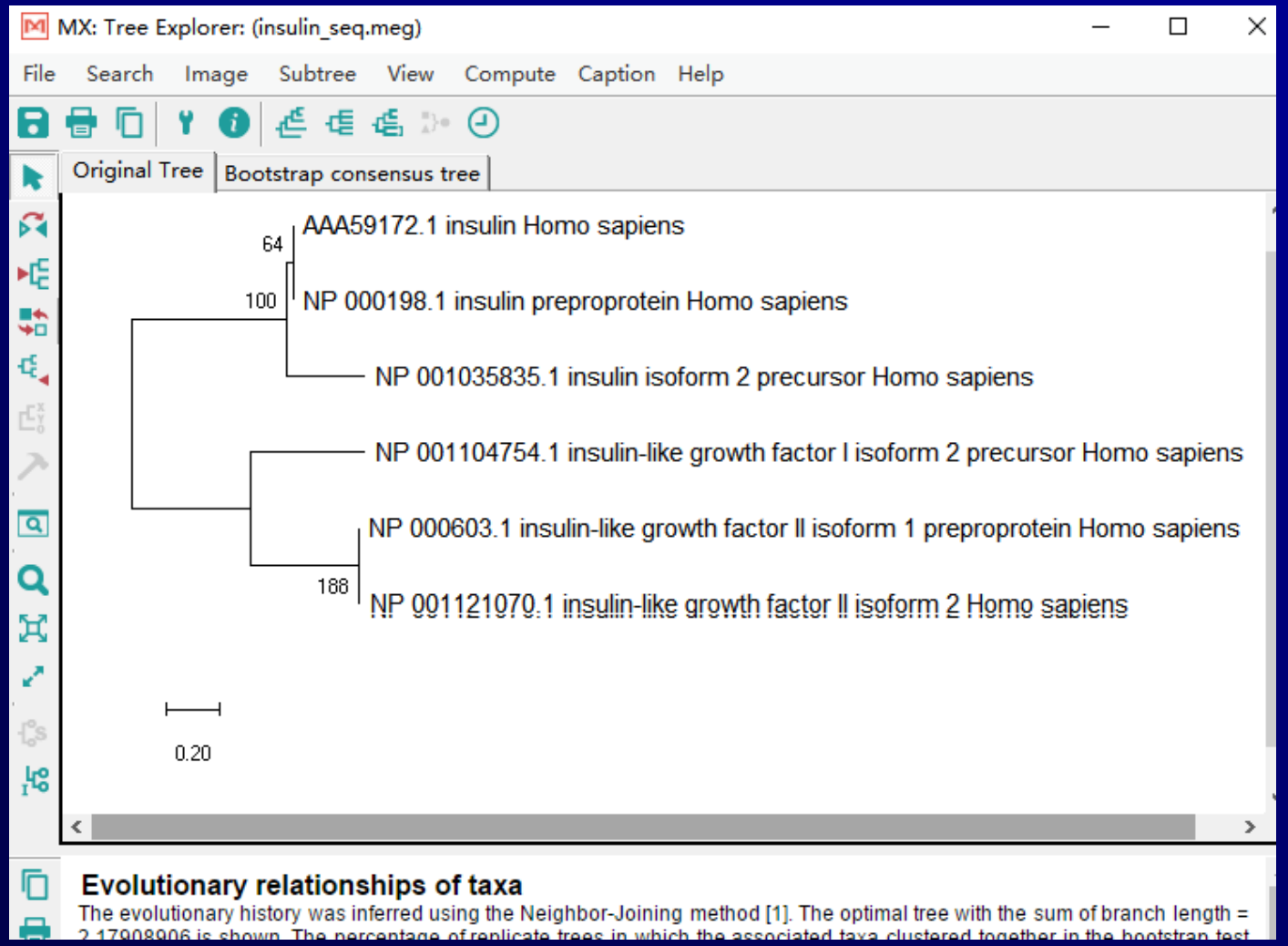
Phylogeny Reconstruction

Option		Setting
ANALYSIS		
Scope	→	*All Selected Taxa*
Statistical Method	→	*Neighbor-joining*
PHYLOGENY TEST		
Test of Phylogeny	→	*Bootstrap method*
No. of Bootstrap Replications	→	*1000*
SUBSTITUTION MODEL		
Substitutions Type	→	*Nucleotide*
Model/Method	→	*Maximum Composite Likelihood*
Substitutions to Include	→	*d: Transitions + Transversions*
RATES AND PATTERNS		
Rates among Sites	→	*Uniform Rates*
Gamma Parameter	→	Not Applicable
Pattern among Lineages	→	*Same (Homogeneous)*
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	*Pairwise deletion*
Site Coverage Cutoff (%)	→	Not Applicable
SYSTEM RESOURCE USAGE		
Number of Threads	→	*3*

Help Cancel OK

MEGA可以在构树过程中自动进行自展检验。

进化树



三要素：
节点
支长
置信度

4.进化树的可靠性评估

Bootstrapping（自展法）

- Bootstrapping可以检验每个节点的可信度。
- Bootstrapping基本步骤:

1. 从排列的多序列中随机有放回的抽取某一列，构成相同长度的新的排列序列
2. 重复上面的过程，得到多组新的序列
3. 对这些新的序列进行建树，再观察这些树与原始树是否有差异，以此评价建树的可靠性

原始排列

Alpha	AACAAC
Beta	AACCCC
Gamma	ACCAAC
Delta	CCACCA
Epsilon	CCAAAC

Bootstrap1

Alpha	ACAAAC
Beta	ACCCCC
Gamma	ACAAAC
Delta	CACCCA
Epsilon	CAAAAC

Bootstrap2

Alpha	AAAACC
Beta	AACCCC
Gamma	CCAACC
Delta	CCCCAA
Epsilon	CCAACC

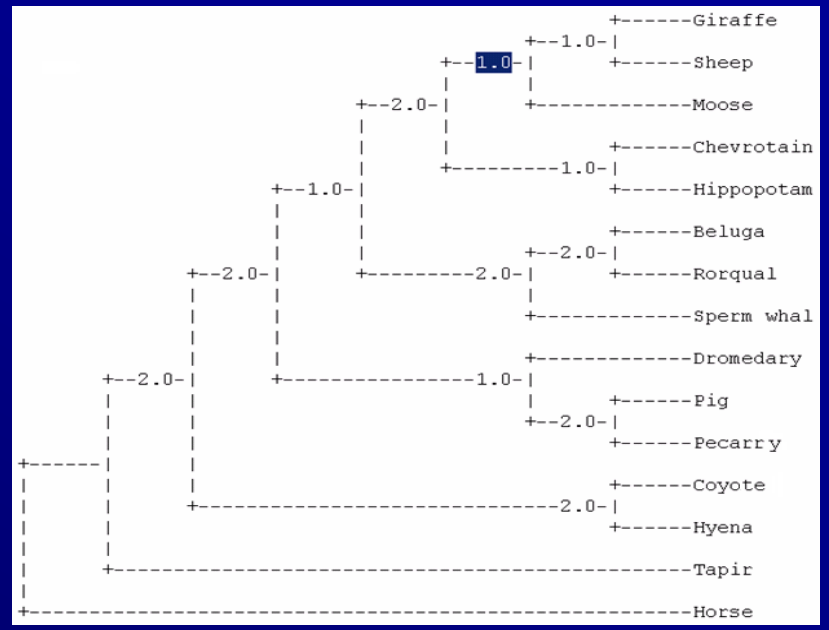
Bootstrap3

Alpha	ACAAAC
Beta	ACCCCC
Gamma	CCAAAC
Delta	CACCCA
Epsilon	CAAAAC

Bootstrapping步骤

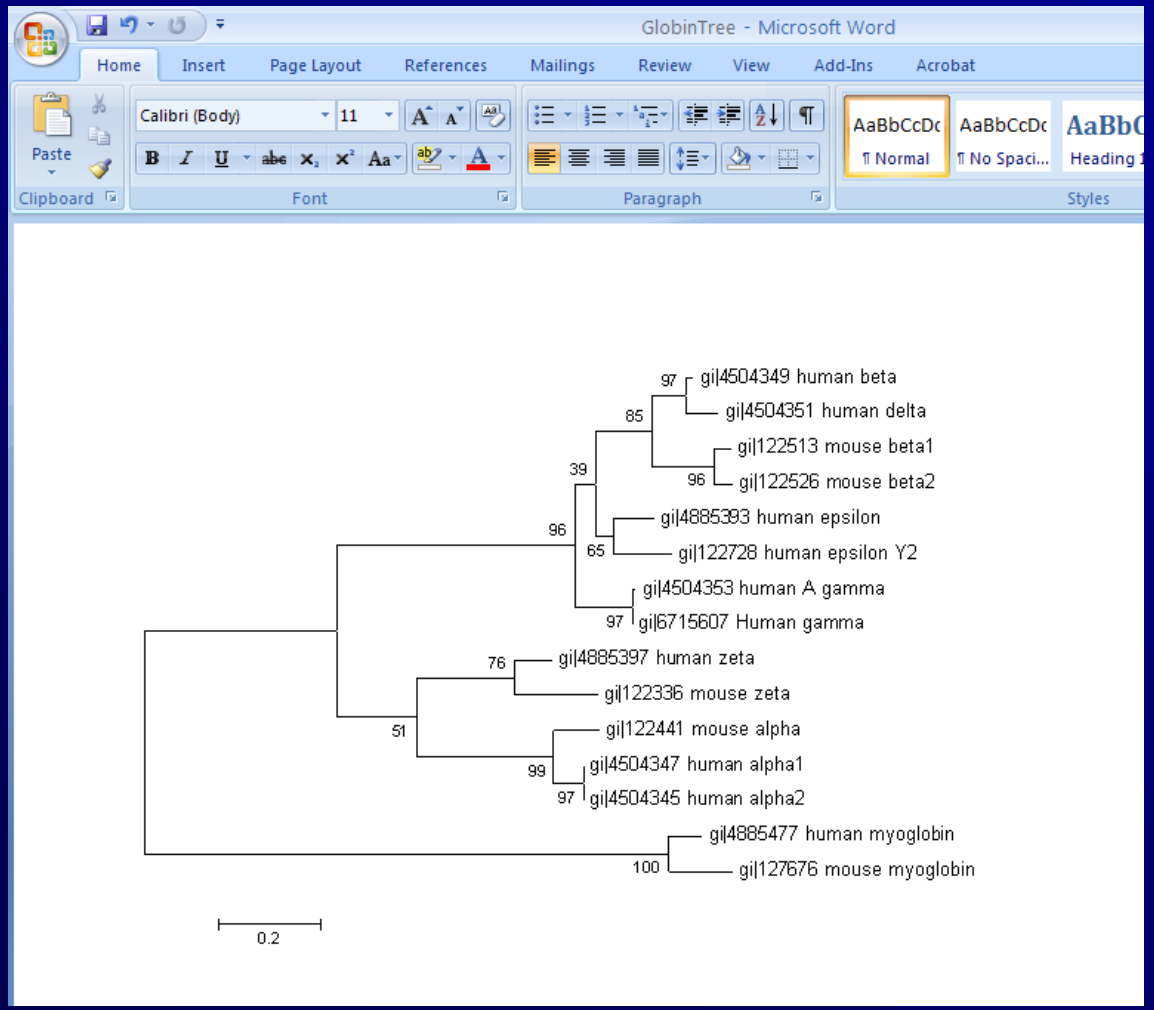
1. 选择序列MSA的一部分序列
2. 以这一部分序列重新进行建树
3. 重复N次步骤1与2 (N为100或1000)
4. 计算以上N个树得到一致树（consensus tree）
5. 计算N个树中的每个节点与一致树中相应节点相同的数量，即为自展值(bootstrap value)或支持度(百分数%)。

- **每个节点的自展值的大小为0与N之间，越大越好.**
- 实际应用中，至少要有多于100次循环。



A Bootstrapped Tree
(此树由2次自展循环产生)

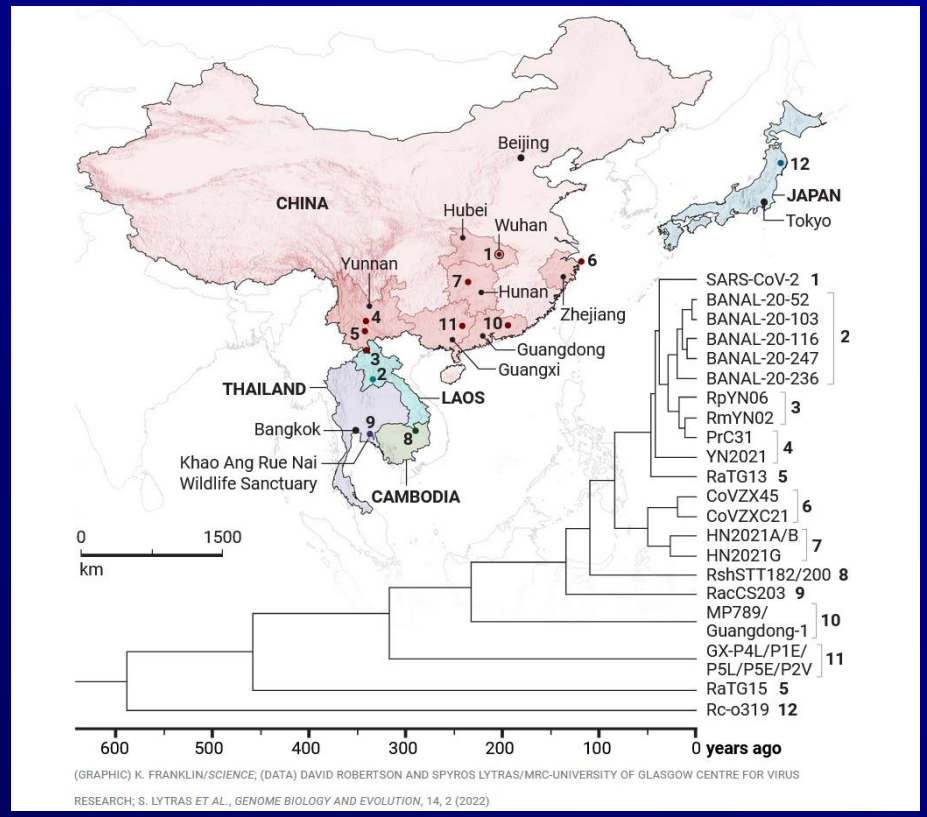
Copy and Paste to Word



作业

- 请用MEGA构建新冠病毒各种病毒株，如原始病毒(Wu-Han-1)、南非变异株(B.1.351)，及蝙蝠(bat)、穿山甲(Pangolin)等冠状病毒的S蛋白序列进化树。根据进化树结果，讨论不同冠状病毒之间的进化关系？

- 提示：
  - ✓ NCBI选nucleotide数据库，并搜索“SARS-COV-2+病毒株名称”；搜索结果页面为genbank格式，按“ CTRL+F ”搜索“Spike ”定位到S蛋白的注释（CDS特征区域），下载氨基酸序列；
  - ✓ MEGA参数设置，模型选p-distance，Gaps选partial deletion, Cutoff: 50%。



https://www.science.org/content/article/hunt-new-viruses-bat-trapping-scientist-hopes-prevent-future-pandemics