

分子进化与系统发育分析（2）-工具篇

教学要求：

了解基本概念，会使用MEGA软件构建进化树。

本节介绍一些术语以及MEGA的使用

同源性分析→功能相似性

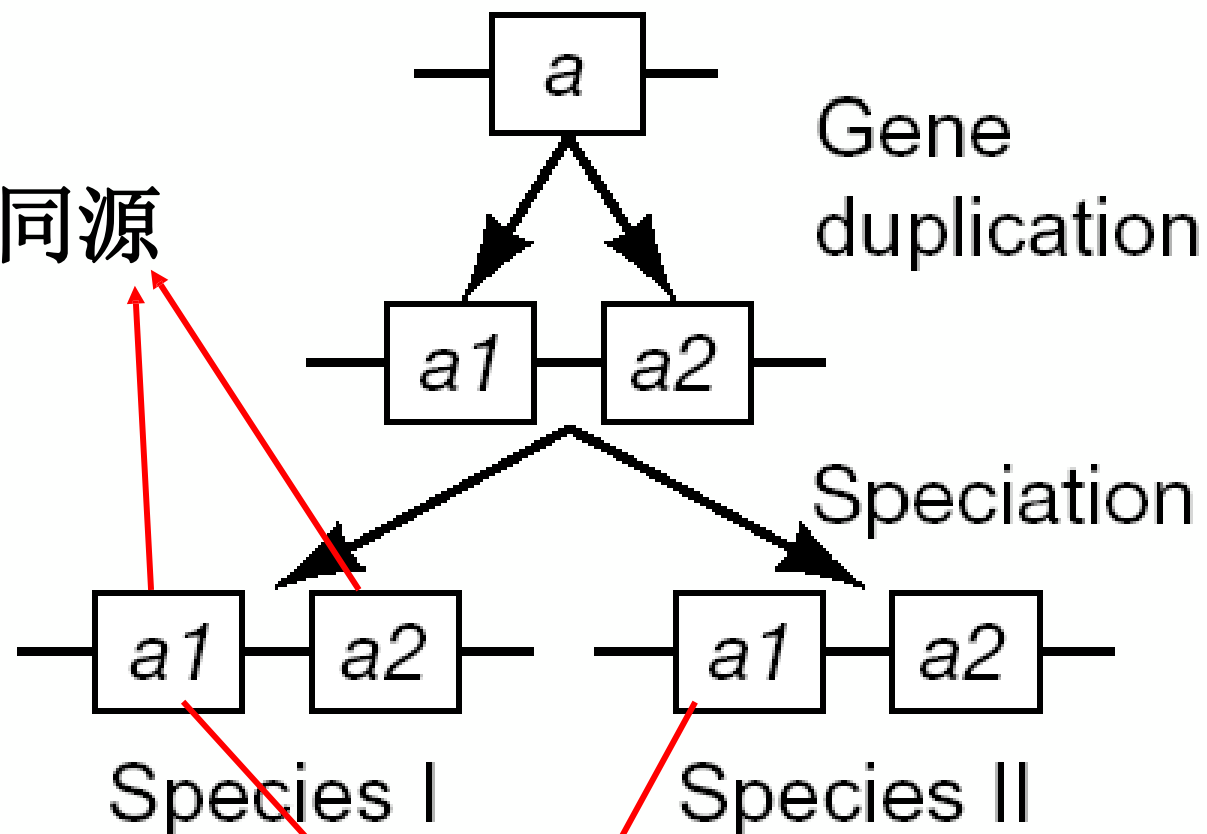


- Ortholog（直系同源序列）：两个基因通过物种形成的事件而产生，或源于不同物种的最近共同祖先的两个基因，或者两个物种中的同一基因，一般具有相同的功能
- Paralog（旁系同源序列）：两个基因在同一物种中，通过至少一次基因复制的事件而产生

直系同源序列 vs. 旁系同源序列



paralogs 旁系同源

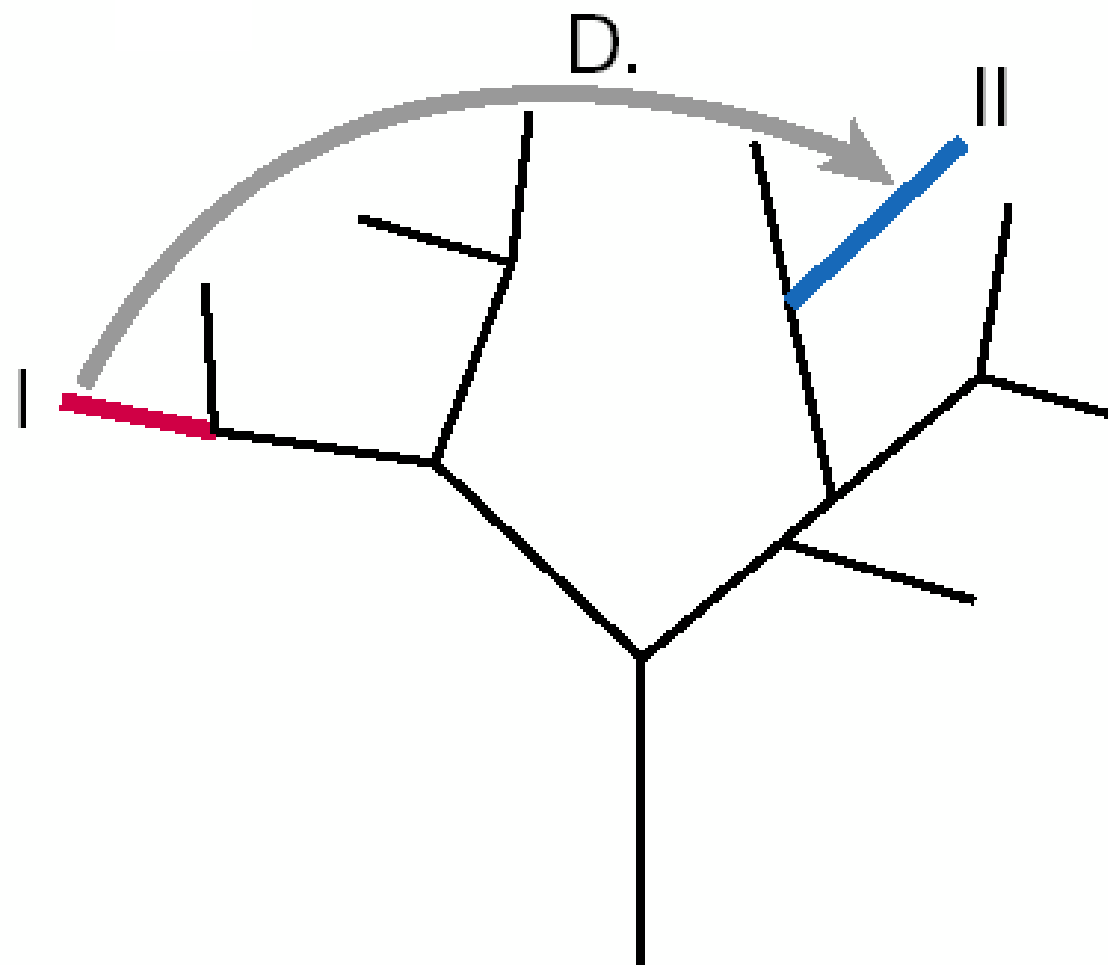


orthologs 直系同源

异同源序列



- ❑ **Xenolog**（异同源序列）：
由某一个水平基因转移
事件而得到的同源序列

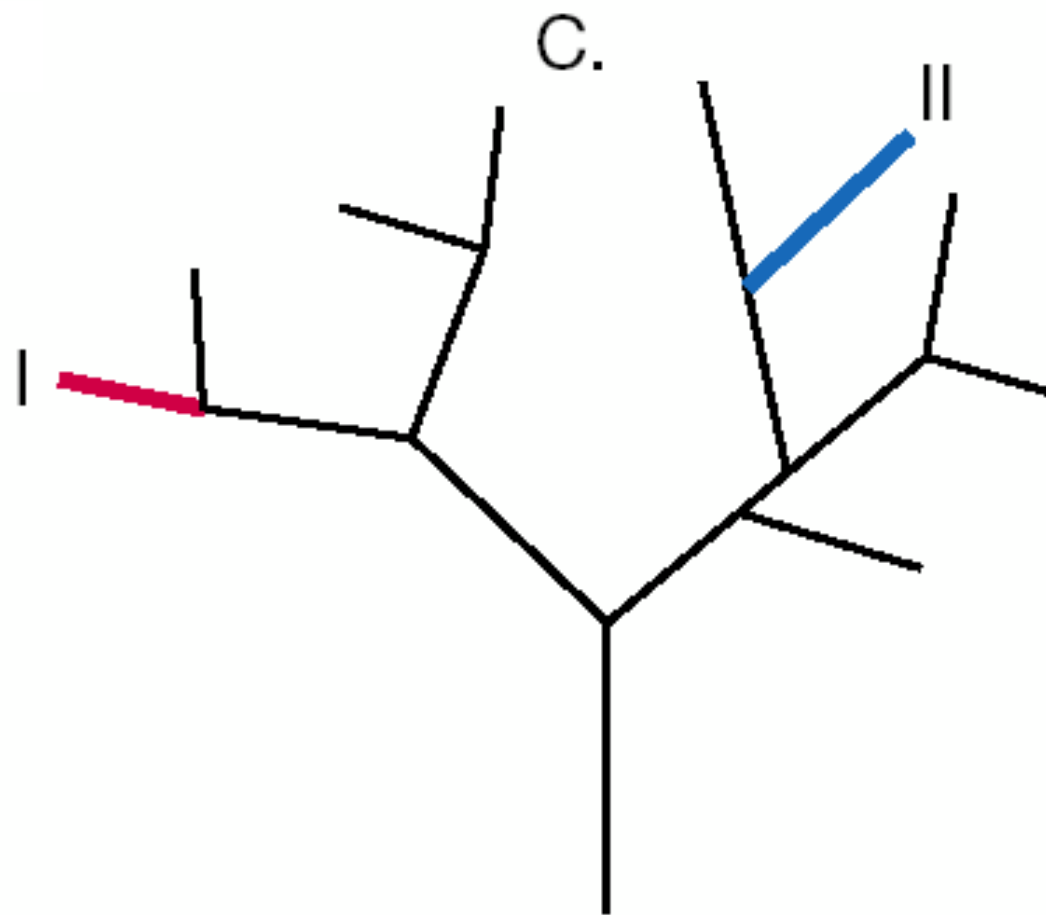


基因的趋同进化



通过不同的进化途径获得保守/相似的功能

- **Convergent evolution**（趋同进化）：
通过不同的 进化途径获得相似的功能，或者功能替代序列



直系、旁系、异同源之间区别

	<u>ortholog</u>	<u>paralog</u>	<u>xenolog</u>
source	from different species	in the same species	from different species
mechanism	Speciation event	Gene duplication	Horizontal gene transfer
function	same	Similar or complementary	similar

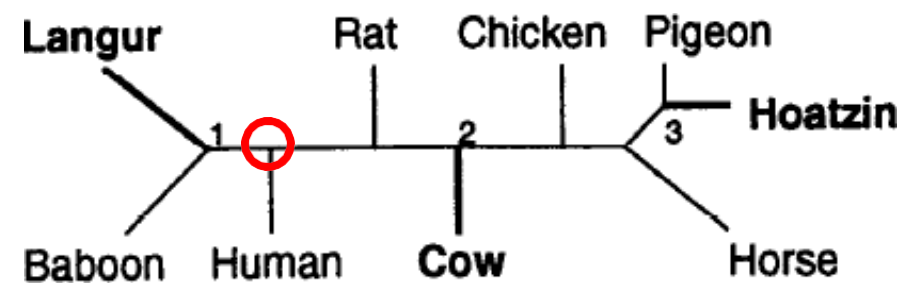
趋同进化: Langur



食叶猴



RNASE: 纤维素分解、消化



密码子偏好及相应分析



- ❑ 密码子（codon）：在随机或者无自然选择 的情况下，各个密码子出现频率将大致相等
- ❑ 密码子偏好：各个物种中，编码同一氨基酸 的不同同义密码子的频率非常不一致
- ❑ 可能的原因：密码子对应的同功tRNA丰度 的不同 – 反密码子（Anticodon）

大肠杆菌RNA聚合酶



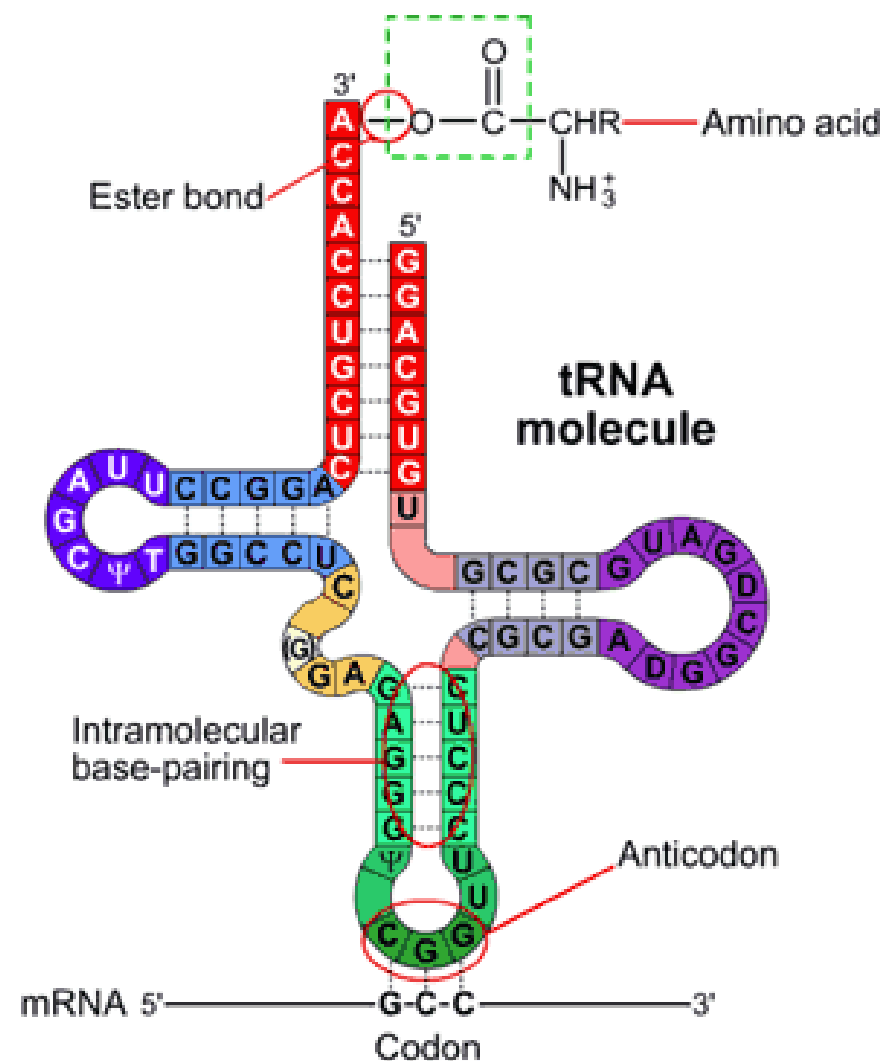
Phe UUU UUC Leu UUA UUG	15 (0.51) 44 (1.49) 2 (0.07) 8 (0.27)	Ser UCU UCC UCA UCG	32 (1.86) 38 (2.21) 2 (0.12) 5 (0.29)	Tyr UAU UAC Ter UAA Ter UAG	18 (0.64) 38 (1.36) 	Cys UGU UGC Trp UGA UGG	5 (1.00) 5 (1.00) 8 (1.00)
Leu CUU CUC CUA CUG	11 (0.36) 18 (0.60) 1 (0.03) 141 (4.67)	Pro CCU CCC CCA CCG	9 (0.48) 0 (0.00) 11 (0.59) 55 (2.93)	His CAU CAC Gln CAA CAG	5 (0.36) 23 (1.64) 15 (0.34) 73 (1.66)	Arg CGU CGC CGA CGG	89 (3.93) 46 (2.03) 1 (0.04) 0 (0.00)
Ile AUU AUC AUA Met AUG	29 (0.69) 98 (2.31) 0 (0.00) 60 (1.00)	Thr ACU ACC ACA ACG	12 (0.78) 63 (2.57) 3 (0.12) 13 (0.53)	Asn AAU AAC Gln AAA AAG	4 (0.11) 66 (1.89) 77 (1.35) 37 (0.65)	Ser AGU AGC Arg AGA AGG	3 (0.12) 23 (1.34) 0 (0.00) 0 (0.00)
Val GUU GUC GUA GUG	55 (1.53) 21 (0.58) 34 (0.94) 34 (0.94)	Ala GCU GCC GCA GCG	30 (0.94) 19 (0.59) 30 (0.94) 49 (1.53)	Asp GAU GAC Glu GAA GAG	60 (0.83) 66 (1.17) 147 (1.52) 46 (0.48)	Gly GGU GGC GGA GGG	78 (2.40) 47 (1.45) 0 (0.04) 5 (0.15)

- 密码子偏好非常明显；例如
- 同为编码Phe的同义密码子UUU和UUC，二者出现的次数显著不等，UUU (15次)，UUC (44次)；
- 再如：编码Arg的四个密码子CGU，CGC，CGA，CGG，出现次数分别为：89，46，1，0。
- 提示：对应CGG的同功tRNA可能不存在！

tRNA & Anticodon



- ❑ 每一个密码子，对应一个 tRNA
- ❑ tRNA通过Anticodon来识别codon,联系 mRNA和氨基酸序列的合成
- ❑ 密码子的使用偏好：由密码子对应的tRNA的进化及丰度来决定



碱基出现的频率



- ❑ 假如：每个核苷酸位点上的替代是随机发生的， 则 A,T,C,G出现的频率应该大致相等
- ❑ 实际情况：DNA受到自然选择的压力，各个位点的碱基出现频率并不相等
- ❑ 需要解决的问题：
 - ✿ 每个位点上受到什么样的选择压力？
 - ✿ 各个位点的碱基频率反映了什么样的规律？
- ❑ 表征/统计的方法：计算G+C的含量，并进行比较

分子进化的理论



- 阳性选择，适应性进化，达尔文进化：
 - ✿ DNA分子显著出现非同义替代，改变编码蛋白质的氨基酸组成，并产生新的功能
- 阴性选择，净化选择：
 - ✿ DNA分子的同义替代显著，较少改变蛋白质的氨基酸组成，其原来的功能高度保守
- 中性进化（木村资生，Motoo Kimura）：
 - ✿ 同义替代与非同义替代比例相当，突变不好不坏，不改变或轻微改变蛋白质的功能

同义替代 vs. 非同义替代



GCG**GTT**TGGGAG

GCG**GTCT**GCGAC

64个密码子，编码20个氨基酸

GTT
GTC
GTA
GTG

脯氨酸P

四倍简并

CGT
CGC

组氨酸H

二倍简并

同义替代

TGG → 色氨酸W

TGC → 半胱氨酸C

非同义替代

编码区 vs. 非编码区



- ❑ 编码区：DNA上编码功能性的基因的部分
- ❑ 非编码区：或称基因组序列，绝大部分无功能
- ❑ 选择压力：
 - ✿ 编码区：阳性选择 1%；中性进化80%；阴性进化 19%
 - ✿ 非编码区：~100%的中性进化

编码区：密码子



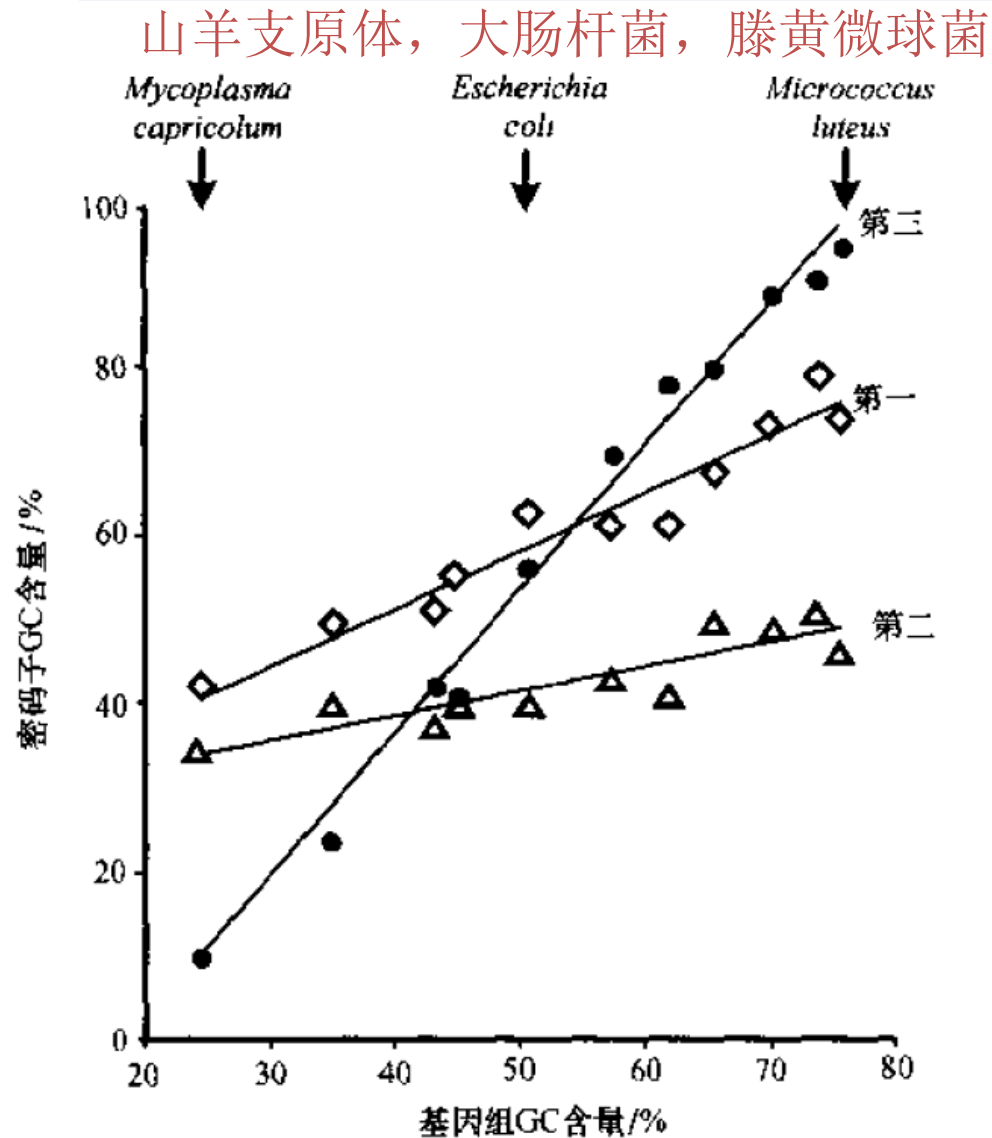
- ❑ 对于同义的密码子，第一位少部分可以允许不同，例如，编码Ser的六个密码子：TCT, TCC, TCA, TCG, AGT, AGC
- ❑ 第二位必须相同
- ❑ 第三位绝大多数可以不同 -> 近似随机
- ❑ 因此：
 - ✿ 第一位：阴性进化占大部分，中性进化占小部分
 - ✿ 第二位：阴性进化
 - ✿ 第三位：阴性进化占小部分，中性进化占大部分

编码区 & 密码子: 推论



- 密码子第三位的碱基出现概率接近基因组序列的碱基频率
- 第二位的碱基出现频率与基因组序列的基础频率相差最大

基因组与GC含量的关系



细菌基因组的 **GC 含量**：
25%~75%

一般认为任何两种微生物在 GC 含量上的差别超过了 10 %，这两种微生物就肯定不是同一个种。
因此可利用 G+C mol %来鉴别各种微生物种属间的亲缘关系及其远近程度。

密码子偏好的应用及计算



- ❑ 基本假设：在**高表达的基因**中，密码子的选择，更倾向于使用**“优化”的同义密码子**
- ❑ 推论1：给定一个物种的一些高表达的基因，我们可以估算优化的同义密码子的分布
- ❑ 推论2：接着，我们可以对给定的一个未知基因的序列进行密码子分布的分析，预测该基因的表达量！
- ❑ 推论3：对于一个表达量很低的基因，我们是否能够通过将少量的密码子改变成优化密码子，从而显著提高基因的表达量？

相对密码子使用频率RSCU



- 相对密码子使用频率（relative synonymous codon usage, RSCU）
- 定义：观测到的某一同一密码子的使用次数，除以“期望”的该密码子出现次数

编码第i个氨基酸
的第j个密码子的
RSCU值

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}$$

编码第i个氨基酸
的第j个密码子的
观测值

如果密码子使用频率没有偏好，则该密码子的RSCU值等于1。当某一密码子的RSCU值大于1，则表明其的使用频率相对较高。

编码第i氨基酸的同
义密码子的数目

密码子相对适用度：the relative adaptation



□ 编码第*i*个氨基酸的第*j*个同义密码子的“相对适应度”：

$$w_{ij} = \frac{RSCU_{i,j}}{RSCU_{i,max}} = \frac{X_{i,j}}{X_{i,max}}$$

□ 即该同义密码子的观察值，除以编码该氨基酸的同义密码子的最大值

密码子适应指数CAI: codon Adaptation Index



$$CAI = \left(\prod_{k=1}^L w_k \right)^{\frac{1}{L}} = \left(\prod_{k=1}^L \frac{RSCU_{i,j}}{RSCU_{i,max}} \right)^{\frac{1}{L}} = \left(\prod_{k=1}^L \frac{X_{i,j}}{X_{i,max}} \right)^{\frac{1}{L}}$$

大肠杆菌 & 酵母



		E.coli		Yeast				E.coli		Yeast	
		RSCU	w	RSCU	w			RSCU	w	RSCU	w
Phe	UUU	0.456	0.296	0.203	0.113	Ser	UCU	2.571	1.000	3.359	1.000
	UUC	1.544	1.000	1.797	1.000		UCC	1.912	0.744	2.327	0.693
Leu	UUA	0.106	0.020	0.601	0.117		UCA	0.198	0.077	0.122	0.036
	UUG	0.106	0.020	5.141	1.000		UCG	0.044	0.017	0.017	0.005
Leu	CUU	0.225	0.042	0.029	0.006	Pro	CCU	0.231	0.070	0.179	0.047
	CUC	0.198	0.037	0.014	0.003		CCC	0.038	0.012	0.036	0.009
	CUA	0.040	0.007	0.200	0.039		CCA	0.442	0.135	3.776	1.000
	CUG	5.326	1.000	0.014	0.003		CCG	3.288	1.000	0.009	0.002
Ile	AUU	0.466	0.185	1.352	0.823	Thr	ACU	1.804	0.965	1.899	0.921
	AUC	2.525	1.000	1.643	1.000		ACC	1.870	1.000	2.063	1.000
	AUA	0.008	0.003	0.005	0.003		ACA	0.141	0.076	0.025	0.012
Met	AUG	1.000	1.000	1.000	1.000		ACG	0.185	0.099	0.013	0.006
Val	GUU	2.244	1.000	2.161	1.000	Ala	GCU	1.877	1.000	3.005	1.000
	GUC	0.148	0.066	1.796	0.831		GCC	0.228	0.122	0.948	0.316
	GUA	1.111	0.495	0.004	0.002		GCA	1.099	0.586	0.044	0.015
	GUG	0.496	0.221	0.039	0.018		GCG	0.796	0.424	0.004	0.001

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}$$

$$w_{ij} = \frac{RSCU_{i,j}}{RSCU_{i,max}} = \frac{X_{i,j}}{X_{i,max}}$$



菌的rpsU

		E.coli		Yeast				E.coli		Yeast	
		RSCU	w	RSCU	w			RSCU	w	RSCU	w
Phe	UUU	0.456	0.296	0.203	0.113	Ser	UCU	2.571	1.000	3.359	1.000
	UUC	1.544	1.000	1.797	1.000		UCC	1.912	0.744	2.327	0.693
Leu	UUA	0.106	0.020	0.601	0.117		UCA	0.198	0.077	0.122	0.036
	UUG	0.106	0.020	5.141	1.000		UCG	0.044	0.017	0.017	0.005
Leu	CUU	0.225	0.042	0.029	0.006	Pro	CCU	0.231	0.070	0.179	0.047
	CUC	0.198	0.037	0.014	0.003		CCC	0.038	0.012	0.036	0.009
	CUA	0.040	0.007	0.200	0.039		CCA	0.442	0.135	3.776	1.000
	CUG	5.326	1.000	0.014	0.003		CCG	3.288	1.000	0.009	0.002
Ile	AUU	0.466	0.185	1.352	0.823	Thr	ACU	1.804	0.965	1.899	0.921
	AUC	2.525	1.000	1.643	1.000		ACC	1.870	1.000	2.063	1.000
	AUA	0.008	0.003	0.005	0.003		ACA	0.141	0.076	0.025	0.012
Met	AUG	1.000	1.000	1.000	1.000		ACG	0.185	0.099	0.013	0.006
Val	GUU	2.244	1.000	2.161	1.000	Ala	GCU	1.877	1.000	3.005	1.000
	GUC	0.148	0.066	1.796	0.831		GCC	0.228	0.122	0.948	0.316
	GUA	1.111	0.495	0.004	0.002		GCA	1.099	0.586	0.044	0.015
	GUG	0.496	0.221	0.039	0.018		GCG	0.796	0.424	0.004	0.001

□ rpsU包含70个codon,部分序列如下:

.CCG.GTA.ATT.AAA.GTA.

$$\text{CAI}_{\text{obs}} = (3.288 \times 1.111 \times 0.466 \times 1.596 \times 1.111 \times \dots)^{1/70}$$
$$\text{CAI}_{\text{max}} = (3.288 \times 2.244 \times 2.525 \times 1.596 \times 2.244 \times \dots)^{1/70}$$

大肠杆菌和酵母：部分基因的CAI



<u>E.coli</u>		<u>yeast</u>	
gene	CAI	gene	CAI
17 RPs	0.467-0.813	16 RPs	0.529-0.915
<u>rpsU</u>	0.726	histones	0.532-0.733
<u>rpoD</u>	0.582		
<u>dnaG</u>	0.271	2u plasmid	0.099-0.106
<u>lacI</u>	0.296	<u>GAL 4</u>	0.116
<u>trpR</u>	0.267	<u>PPR 1</u>	0.114
<u>lpp</u>	0.849 ^a	<u>GPD 1</u>	0.929 ^a
<u>hsdS</u>	0.218 ^b	<u>mat A2</u>	0.098 ^b

RPs - ribosomal protein genes.

^a highest CAI value among data set.

^b lowest CAI value among data set.

异源基因：在其他物种中的CAI



Heterologous gene	Host	
	<u>E.coli</u>	Yeast
Human alpha interferon	0.218	0.099
Human insulin	0.307	0.043
Human growth hormone	0.287	0.082
Human factor VIII	0.205	0.114
Human factor IX	0.263	0.176
Bovine chymosin	0.326	0.086

氨基酸序列的进化演变



- ❑ 分子进化的分析：基于氨基酸序列的分析早于DNA序列
- ❑ 优势：氨基酸序列更为保守，对年代跨度大的进化分析有帮助；数学模型较DNA远为简单
- ❑ p距离：p-distance
- ❑ 泊松校正，d距离

p-distance



- 另两条蛋白质序列之间的氨基酸差异数为 n_d ，
所有序列的氨基酸数目相同为 n ，则

$$\text{P距离} \longrightarrow \hat{p} = \frac{n_d}{n}$$

所有的插入/缺失都要删除！

不同物种的血红蛋白 α 链中不同氨基酸的数目及比例。长度：140aa

	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	

PC: 泊松校正



- 序列差异的百分比 (p) 与分歧时间t的关系: t较短的时候, 回复突变较少, 两者大致成线性关系; 当t较大时, 回复突变增多, 二者成非线性关系
- 令 γ 为某一位点每年的氨基酸替代率, 并假设所有位点的 γ 都相同: 基本假设
- 在时间t年之后, 每个位点替代的平均数为: γt ; 给定一个位点, 氨基酸替代数 k ($k=0,1,2,3,\dots$) 的可能性遵循泊松分布, 即

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

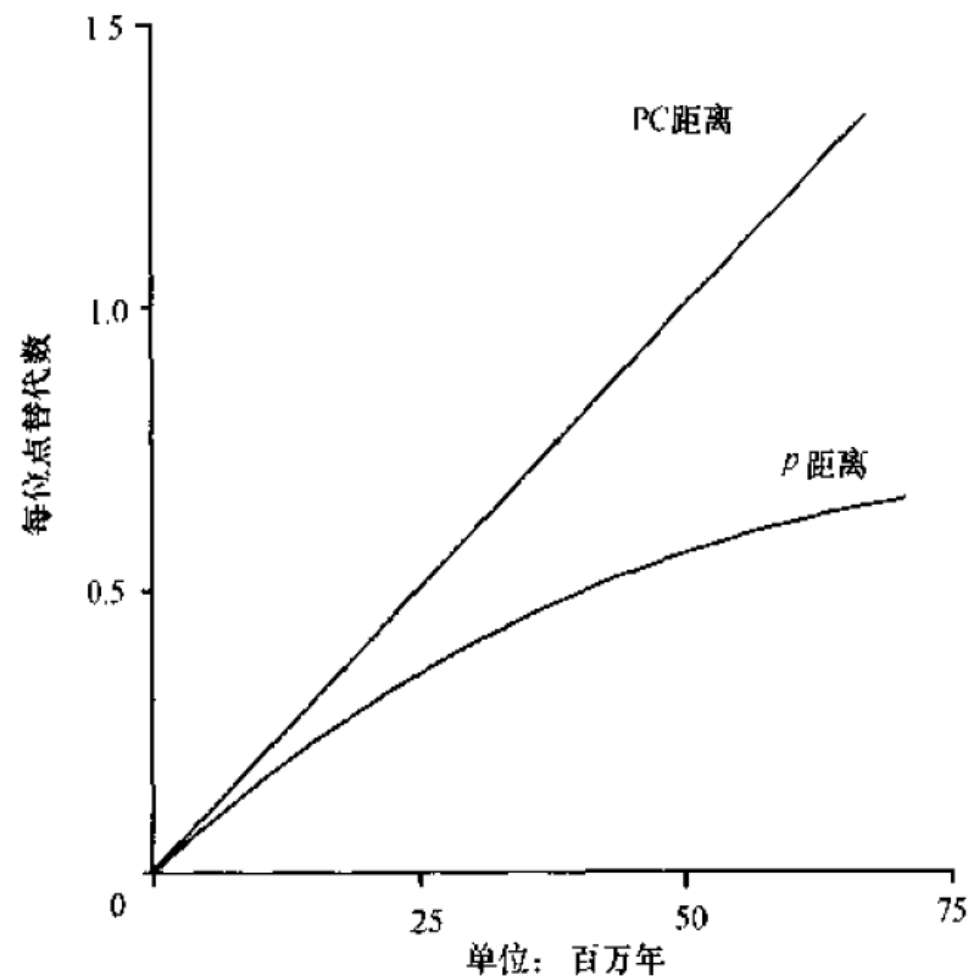
- 因此, 某一位点氨基酸不变的概率为 $P(0;t) = e^{-rt}$

PC: 泊松校正 (2)



- ❑ 祖先序列未知：不知道当前的序列从何演化而来
- ❑ 解决方案：对两条分化 t 年的序列，一条序列无替代的概率为 e^{-rt} ，两条序列则为： $q = (e^{-rt})^2 = e^{-2rt}$
- ❑ $q=1-p$
- ❑ 泊松校正距离 $d=2rt$
- ❑ 因此， $q = (e^{-rt})^2 = e^{-2rt} \Leftrightarrow$
 $1-p = e^{-d} \Leftrightarrow -d = \ln(1-p)$
 $\Leftrightarrow d = -\ln(1-p)$
- ❑ $d=-\ln(1-p)$ ，即泊松距离

p-距离 vs. 泊松距离



DNA序列的进化演变



- ❑ 基因组上存在着多种多样的DNA区域，例如蛋白质编码区，非编码区，内含子，侧翼区，重复片段以及插入序列等
- ❑ 本章考虑蛋白质与RNA的编码区的DNA序列的进化演变模型
- ❑ 进化模型：Jukes-Cantor法与Kimura两参数法

两条DNA序列的差异



- 对于两条长度为 n 的DNA序列，不同的碱基对为 n_d ，则两条序列的差异性可表示为：

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

- 核苷酸的改变：转换 P 、颠换 Q ,则 $p=P+Q$
- 当 p 较小时，如果核苷酸替代是随机发生的， $Q=2P$ ；通常转换比颠换出现频率高；
- 转换/颠换比：

$$\hat{R} = \frac{\hat{P}}{\hat{Q}}$$

核苷酸替代数的估计



	A	T	C	G		A	T	C	G
	(A) Jukes-Cantor 模型					(E) HKY 模型			
A	—	α	α	α		—	βg_T	βg_C	αg_G
T	α	—	α	α		βg_A	—	αg_C	βg_G
C	α	α	—	α		βg_A	αg_T	—	βg_G
G	α	α	α	—		αg_A	βg_T	βg_C	—
	(B) Kimura 模型					(F) Tamura-Nei 模型			
A	—	β	β	α		—	βg_T	βg_C	$\alpha_1 g_G$
T	β	—	α	β		βg_A	—	$\alpha_2 g_C$	βg_G
C	β	α	—	β		βg_A	$\alpha_2 g_T$	—	βg_G
G	α	β	β	—		$\alpha_1 g_A$	βg_T	βg_C	—
	(C) Equal-input 模型					(G) General reversible 模型			
A	—	αg_T	αg_C	αg_G		—	αg_T	βg_C	$c g_G$
T	αg_A	—	αg_C	αg_G		αg_A	—	$d g_C$	$e g_G$
C	αg_A	αg_T	—	αg_G		βg_A	$d g_T$	—	$f g_G$
G	αg_A	αg_T	αg_C	—		$c g_A$	$e g_T$	$f g_C$	—
	(D) Tamura 模型					(H) 无限制模型			
A	—	$\beta \theta_2$	$\beta \theta_1$	$\alpha \theta_1$		—	a_{12}	a_{13}	a_{14}
T	$\beta \theta_2$	—	$\alpha \theta_1$	$\beta \theta_1$		a_{21}	—	a_{23}	a_{24}
C	$\beta \theta_2$	$\alpha \theta_2$	—	$\beta \theta_1$		a_{31}	a_{32}	—	a_{34}
G	$\alpha \theta_2$	$\beta \theta_2$	$\beta \theta_1$	—		a_{41}	a_{42}	a_{43}	—

Jukes-Cantor法



- 假定任一位点的核苷酸替代的频率相等，且每一位点的核苷酸每年以 α 的概率演变为其他三种核苷酸的一种
- 因此，一个核苷酸演变为其他三种核苷酸之一的概率为 $\gamma = 3\alpha$
- 假设，在 t 年前分化出两条核酸序列 X 和 Y ， q_t 表示 X 和 Y 值之间相同核苷酸的比例值， $p_t = 1 - q_t$ ，表示 X 和 Y 之间不同的核苷酸的比例值

Jukes-Cantor法 (2)



- 对于X和Y之间相同(q_t)的核苷酸的一个位点，在时间 $t+1$ 时(过了一年)，以 $(1-\gamma)^2$ 的概率保持不变；当 γ 较小时， γ^2 可以忽略，则 $q_{t+1}=1-2\gamma$
- 对于X和Y之间不同($1-q_t$)的位点，假设在时间 t 时，X序列上的位点 i ，Y序列上为 j ：如果X的 i 变成 j ，而Y上的 j 不变，则二者将相同；事件发生的概率为 $\alpha(1-\gamma)=\gamma(1-\gamma)/3$ ；反之的概率是相等的。因此事件的总概率为： $2\gamma(1-\gamma)/3$ ， γ^2 忽略，则近似为： $2\gamma/3$

Jukes-Cantor法 (3)



□ 因此，差分方程为：

□ 令 $\frac{d_q}{d_t} = q_{t+1} - q_t$ ，则

$$\frac{d_q}{d_t} = \frac{2}{3}\gamma - \frac{8}{3}\gamma q$$

当初始条件 $t = 0$ 且 $q = 1$ 时，

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{8}{3}\gamma t})$$

$$q_{t+1} = (1 - 2\gamma)q_t + (2/3)\gamma(1 - q_t)$$

$$\Leftrightarrow q_{t+1} = q_t - 2\gamma q_t + \frac{2}{3}\gamma - \frac{2}{3}\gamma q_t$$

$$\Leftrightarrow q_{t+1} = q_t + \frac{2}{3}\gamma - \frac{8}{3}\gamma q_t$$

$$\Leftrightarrow q_{t+1} - q_t = \frac{2}{3}\gamma - \frac{8}{3}\gamma q_t$$

□ 两条序列每一位点的替代期望值 $d = 2\gamma t$ ，代入

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{8}{3}\gamma t}), \text{ 且 } d = 2\gamma t, \text{ 则}$$

$$1 - p = 1 - \frac{3}{4}(1 - e^{-\frac{4}{3}d}) \Leftrightarrow \frac{4}{3}p = 1 - e^{-\frac{4}{3}d}$$

$$\Leftrightarrow e^{-\frac{4}{3}d} = 1 - \frac{4}{3}p \Leftrightarrow d = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

Kimura两参数法



- 对于实际数据，转换替代速率通常高于颠换速率；因此，每年每个位点转换替代率为 α ，颠换替代率 2β

- 可计算P,Q值为：
$$P = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t})$$

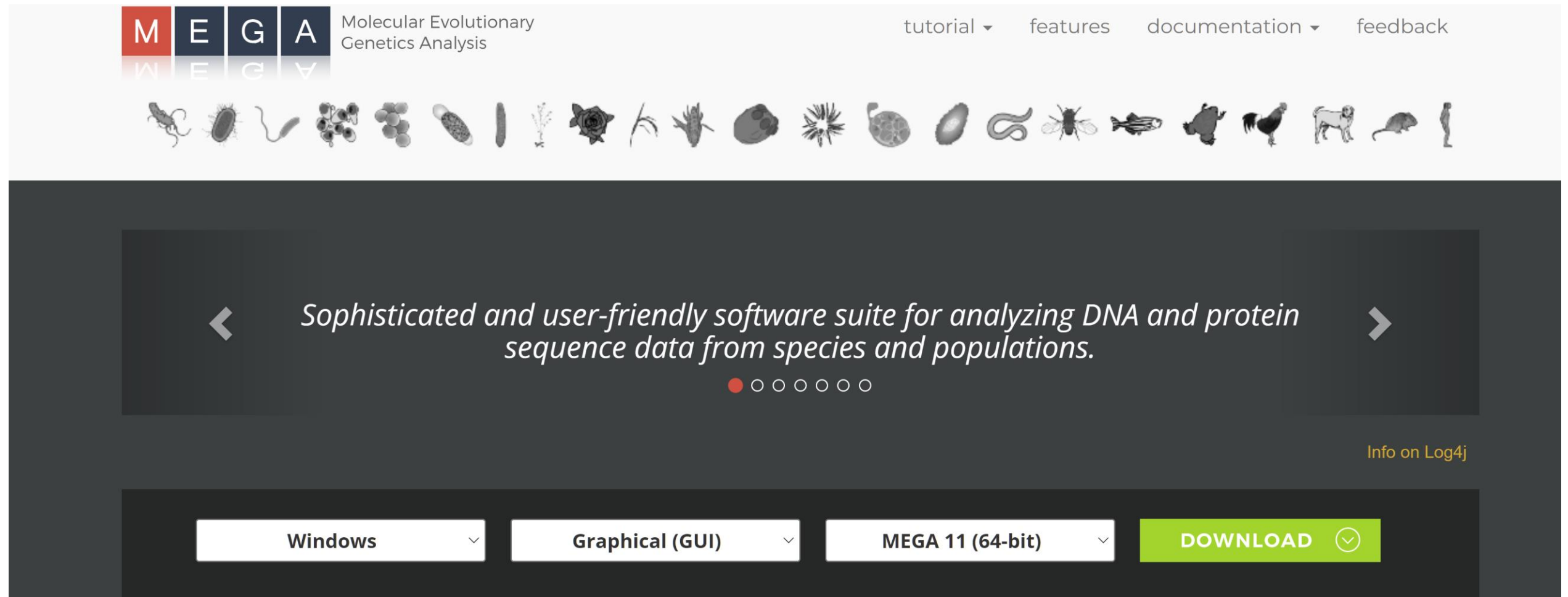
$$Q = \frac{1}{2}(1 - e^{-8\beta t})$$

$$d = 2\gamma t = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$$

- P和Q可以从两条比对的序列中观测并计算得到

MEGA: Molecular Evolutionary Genetics Analysis

网址: <https://megasoftware.net>



The screenshot shows the homepage of the MEGA (Molecular Evolutionary Genetics Analysis) software. At the top, the MEGA logo is displayed with the text "Molecular Evolutionary Genetics Analysis" to its right. Navigation links for "tutorial", "features", "documentation", and "feedback" are located in the top right corner. Below the navigation bar is a horizontal row of 20 small icons representing various organisms, including bacteria, fungi, plants, and animals. The main content area features a dark grey background with a central text box containing the description: "Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations." Below this text is a row of seven circles, with the first one filled in red. To the right of the text box is a link labeled "Info on Log4j". At the bottom, there is a row of four buttons: "Windows", "Graphical (GUI)", "MEGA 11 (64-bit)", and a green "DOWNLOAD" button with a checkmark icon.

M E G A Molecular Evolutionary Genetics Analysis

tutorial ▾ features documentation ▾ feedback

Icons representing various organisms: bacteria, fungi, plants, and animals.

◀ *Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations.* ▶

● ○ ○ ○ ○ ○ ○ ○

[Info on Log4j](#)

Windows ▾ Graphical (GUI) ▾ MEGA 11 (64-bit) ▾ **DOWNLOAD** ✓

M11: Alignment Explorer (TIR.fasta)

DataEditSearchAlignmentWebSequencerDisplayHelp

Protein Sequences

Species

1. sp QCLHFDLPWYLRMLGQCTQTWHF

2. sp QGDLWYCFHLCALPWRGRQS

3. sp QHHLFYWDVWFIYVCLAKVKGY

4. sp QHLYFWDVWYIYHFCAKIKGY

5. sp QYLDLPWYLRMVCQWTQTRRAI

6. sp OTKFRGFCFICYKTAQRLVFKDF

7. sp OKFYFHLMLLAGCIKYGREN

8. sp OEGWRISFYWNVSVHRVLGFKEL

9. sp OHRFHGLWYMKMMAWLQAKRKPR

10. sp QSYLDLPWYLRMVCQWTQTRRRAR

Align by ClustalW

Align by ClustalW (Codons)

Align by MUSCLE

Align by MUSCLE (Codons)

Mark/Unmark Site

Align Marked Sites

Unmark All Sites

Delete Gap-Only Sites

Auto-Fill Gaps

Alignment Explorer

比对前

Align by ClustalW

Align by ClustalW (Codons)

Align by MUSCLE

Align by MUSCLE (Codons)

Mark/Unmark Site

Align Marked Sites

Unmark All Sites

Delete Gap-Only Sites

Auto-Fill Gaps

Alignment Explorer

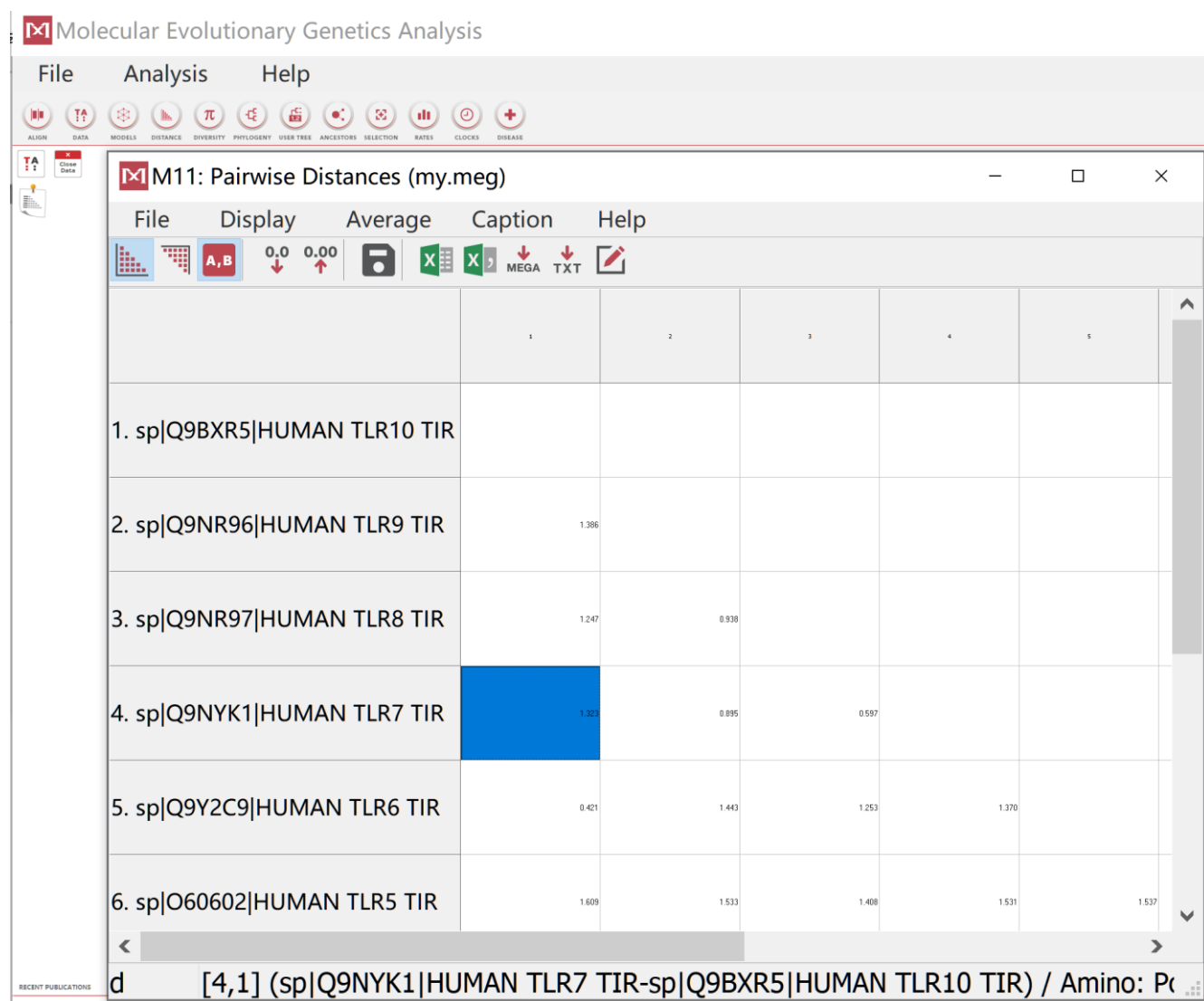
比对后

Protein Sequences

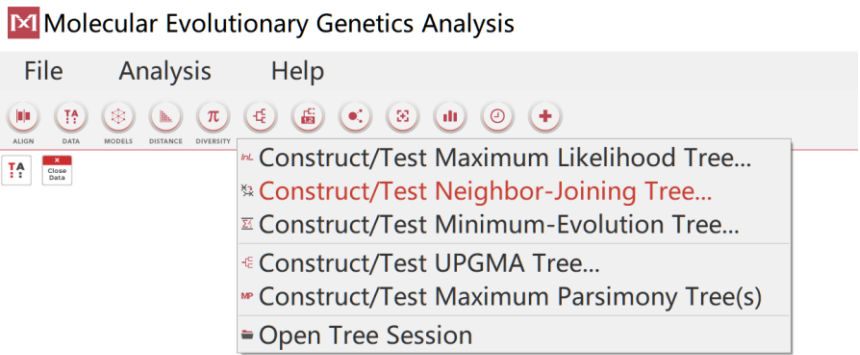
Species: 1. sp, 2. sp, 3. sp, 4. sp, 5. sp, 6. sp, 7. sp, 8. sp, 9. sp, 10. sp

Protein Sequences: QCLHFDLPWYLRMLGQCTQTWHRVRKTTQEQLKRVRFHAFISYSEHD---SLWVKNELIPNLE---KEDGSILICLYESYFDPGKSISENIVSF-IEKSYKSI FVLSPNFVQNEWCH-YEFYFAHNLFHENS DHIILILLEPIPFYCIPTRYHKLKALLEKKA; -----GWDLWYCFHLCLAWLPWRGRQSGRDEALPYDAFVVFDTQSAVADWVYNELRGQLEEC-RGRWALRLCLEERDWLPGKTLFENLWAS-VYGSRKTLFVLAHTDRVSGLLR-ASFLLAQQRLLIEDRKDVVVLVILSPDGR---RSRYVRLRQRLCRQS; ---HHLFYWDVWFIYNVCLAKV--KGYRSLSTSQTF-YDAYISYDTKDASVTDWVINELRYHLEES-RDKNVL-LCLEERDWDPLAIIIDNLMS-INQSKKTVFVLTKKYAKSWNFK-TAFYLALQRLMDENMDVIIIFILLEPVLQ---HSQYLRLRQRICKSS; ---HLYFWDVWYIYHFC AKI--KGYQRLISPDCC-YDAFIVYDTKDPAVTEWVLAELVAKLEDP-REKHFN-LCLEERDWLPGPVLENLSQS-IQLSKKTVFVMTDKYAKTENFK-IAFYLSHQRLMDEKVDVIIIFLEKPFQ---KSKFLQRLKRLCGSS; ---YLDLPWYLRMVCQWTQTRRRARNIPLEELQRNLQFHAFISYSEHD---SAWVKSELVPYLE---KED--IQICLHERNFVPGKSIVENIINC-IEKSYKSI FVLSPNFVQSEWCH-YELYFAHNLFHEGSNLILILLEPIPQNSIPNKYHKLKALMTQRT; ---TKFRGFCFCICYKTAQRLVFKDHPQGTEPDMYKYDAYLCFSSKD---FTWVQNALLKHLDTQYSDQNRNLCFEERDFVPGENRIANIQDA-IWNSRKIVCLVSRHFLRDGWCL-EAFSYAQGRCLSDLNSALIMVVVGSLSQYQ-LMKHQSIRGFVQKQQ; ---KFYFHLMLLAGCIKYGR-----GENIYDAFVIYSSQD---EDWVRNELVKNLE---EGVPPFQLCLHYRDFIPGVAIAANIIEGFFHKSRKIVVVVSHQFIQSRWCI-FEYELIAQTWQFLSSRAGIIFIVLQKVEK-TLLRQQVELYRLLSRNT; ---EGWRISFYWNVSVHRVLG-FKEIDRQTEQFEYAAIYI HAYKD---KDWVWEHFS-SME---KEDQSLKFCLERDFEAGVFLEAIVNS-IKRSRKIIFVITHLLKDP LCKRFKVHHAVQQAIEQNLD SIILVFLEEIPDYKLNHALCLRRGMFKSHC; ---HRFHGLWYMKMMWALQAKRKPRKAP---SRNICYDAFVSYSERD---AYWVENLMVQELE---NFNPPFKLCLHKRDFIPGKWIIDNIIDS-IEKSHKTVFVLS ENFVKSEWCK-YELDFSHFRLFDENNDAAILILLEPIEKKAIPQRFCKLRKIMNTKT; t---SYLDLPWYLRMVCQWTQTRRRARNIPLEELQRNLQFHAFISYSGHD---SFWVKNELLPNLE---KEG--MQICLHERNFVPGKSIVENIITC-IEKSYKSI FVLSPNFVQSEWCH-YELYFAHNLFHEGSNLILILLEPIPQNSIPSSYHKLKSLMARRT

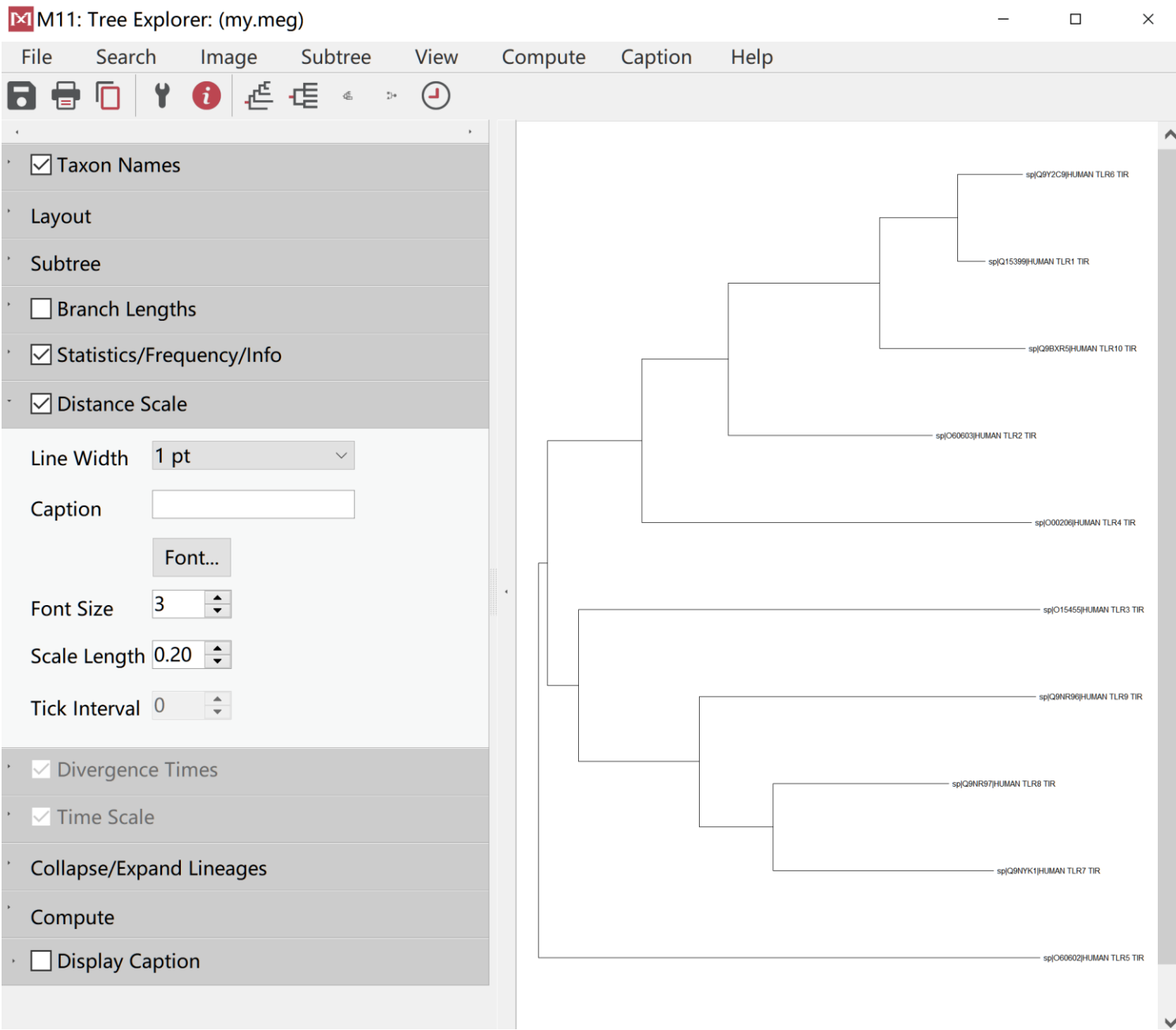
选择 **C**查看保守列。



计算距离矩阵



构建进化树



ME Molecular Evolutionary Genetics Analysis

FileAnalysisHelp

ALIGN

DATA

TA

Close Data

M

Models

Distance

Diversity

Phylogeny

User Tree

Ancestors

Selection

Rates

Clocks

Diagnose Mutation(s)

ELECTION

RATES

CLOCKS

DISEASE

Construct/Test Maximum Likelihood Tree...

Construct/Test Neighbor-Joining Tree...

Construct/Test Minimum-Evolution Tree...

Construct/Test UPGMA Tree...

Construct/Test Maximum Parsimony Tree(s)

Open Tree Session

M11: 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	→ 100
SUBSTITUTION MODEL	
Substitutions Type	→ Amino acid
Model/Method	→ Poisson model
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	→ 7

Help

Cancel

OK

Original Tree

Bootstrap consensus Tree

