生物物理学报　第二十三卷　第三期　二〇〇七年六月
ACTA BIOPHYSICA SINICA　Vol.23 No.3　Jun. 2007

密码对的使用与基因组进化

王芳平，　李　宏
（内蒙古大学理工学院物理系，呼和浩特 010021）

摘要：以 5 种真核、20 种细菌、10 种古菌生物的基因组为样本，分析了编码序列中密码对和基因间序列中三联体对的相对模式数随频数的分布，验证了这种分布符合 $\Gamma(\alpha,\beta)$分布。发现分布形状参数 α 值与生物基因组进化存在明显的相关性；编码序列与基因间序列的进化方式截然不同。随着进化，编码序列的分布形状逐渐向随机分布靠近（α 值逐渐增大）。而对基因间序列，古菌与真核生物的分布形状接近，与细菌的分布相差明显。

关键词：基因组进化；密码对；三联体对；$\Gamma(\alpha,\beta)$分布
中图分类号：Q61

0　引　　言

在密码中蕴藏了“生命机器”的工作原理，包含了生命形成和进化的信息。生命既能保持物种的繁衍，又能表现个体差异，这归功于其机体内遗传密码的作用和基因表达的差异。多年的研究表明，同义密码子使用是非随机的[1~3]，其偏好使用反映了两个主要的进化力量：mRNA 翻译效率的选择[4~7]与作用于 DNA 序列的突变漂移[8,9]。同样，密码对的使用也是非随机的[10~13]。最初 Gutman 和 Hatfield[10]基于大肠杆菌 237 个基因的研究发现：即使在消除密码子与氨基酸对的偏好性之后，密码对的使用还是非随机的。近年来，伴随着多种生物全基因组测序的完成，Ross[14]、Boycheva[15]和 Moura[16]等人从全基因组的角度出发，进一步证实了密码对的使用确实是非随机的。这种非随机性受很多因素的影响：包括密码子的偏好性、二肽的偏好性、二核苷酸的偏好性、密码子的前后文关系、序列的 GC 含量以及转录期间翻译调节信号的潜在驱动力等等[17~20]。如：原核生物偏好使用稀有密码子的组合，真核生物避免这种组合[15,16]；大肠杆菌基因组中起始密码对使用与翻译起始效率有关[21]等。最近 Ross 等[14]运用统计学与信息聚类及多变量分析，对 16 种生物密码对进行了较全面的研究，发现 tRNA 的结构是促进密码对最优搭配的主要因素。

目前关于密码对的非随机使用的研究内容，大多都局限在密码对的使用与翻译效率的关系，及其同源 tRNA 的特性上，但由于涉及参量很多（约为 64^2=4096 个），有关密码对使用的可利用信息匮乏。影响因素的多样性及生物学背景的复杂性，使所得结论仅在问题的表面，或个别几个点上，对造成密码对非随机使用的根源，还不是很清楚，特别是密码对的非随机使用与生物基因组进化关系的研究很少。因此，这个问题是科学研究者共同面临的一个极具挑战性的课题。本文试图从基因组进化的角度对不同生物基因组密码对的使用作全面的统计分析，寻求较为普适的、限制构建编码序列自由度的一种原因，揭示密码对非随机使用的进化约束。从密码对（或三联体对）的相对模式数随频数的分布出发，分析已测序的 10 种古菌、20 种细菌和 5 种真核生物基因组中密码对的组合模式数随频数的分布，得到这种分布所遵循的理论模型，构成编码序列与基因间序列组成结构与生物进化的关系。这是系统研究这一问题的理论尝试，期望对研究密码对的非随机使用有所启发。

1　数据资料和分析方法

1.1　数据资料

从　http://www.cbi.pku.edu.cn/pub/bio-mirror/genebank/genomes/ 下载了 10 个古菌、20 个真细菌和 5 个真核生物的最新基因组（不包括内含子）、DNA 全序列及注释资料。已经测序的细菌全基因

收稿日期：2006-10-27
基金项目：国家自然科学基金项目（30660044）和高等学校博士点基金项目（20050126003）
通讯作者：李宏，电话：(0471)6678889，
E-mail：lihong499@hotmail.com

组（430 个）很多，为提高统计分布的可靠性，减少统计涨落，我们选取基因组的依据是：测序时间较早，ORF 中已确定为编码序列的比例相对较高。只选择了 ORF 数目大于 2000 的细菌基因组。对果蝇由于其他染色体数据的不确定，我们只分析了其染色体 2、3、4 和 X。

考虑到 DNA 双链的互补性对统计结果的影响，这里只分析单链上（文中选择主链）编码序列和对应的基因间序列。除去了非 ATG、GTG 和 TTG 起始的基因，选取的物种名称和短名称以及基因数目、统计密码对的数目、各个物种所对应 DNA 单链上基因间序列的三联体对数目如表 1。

Table 1 The numbers of codon pairs and triplet pairs, the numbers of ORFs and intergenic sequences, and the 35 organisms' names and their abbreviations

Name	Abbr.	Domain	Gene No.	Codon pairs	Intergenic No.	Triplet pairs
Halobacterium_sp	Hsp	A	1008	287228	867	110437
Pyrococcus_abyssi	Paby	A	929	259977	800	109608
Pyrococcus_furiosus	Pfur	A	1054	287673	807	142880
Sulfolobus_solfataricus	Ssol	A	1491	423850	1148	305052
*Thermococcus_kodakaraensis_KOD*1	Tkod	A	1128	319922	888	134423
Archaeoglobus_fulgidus	Aful	A	1178	330763	850	164724
Thermosynechococcus_elongatus	Telo	A	1277	388263	1126	183358
Thermotoga_maritima	Tmar	A	1011	318605	631	110203
Methanobacterium_thermoautotrophicum	Mthe	A	1093	256360	867	101240
*Sulfolobus_acidocaldarius_DSM*_639	Saci	A	939	306941	769	212688
*Acidobacteria_bacterium_Ellin*345	Abac	B	2350	824165	2022	334167
*Anabaena_variabilis_ATCC*_29413	Avar	B	2668	926976	2535	416665
Arthrobacter_aurescens_TC1	Aaur	B	1968	661409	1650	346712
Bacillus_subtilis	Bsub	B	1941	595440	1680	268282
Bacillus_thuringiensis_Al_Hakam	Bthu	B	2320	701783	2026	412159
Bacteroides_thetaiotaomicron_VPI–5482	Bthe	B	2366	935946	2182	313039
*Escherichia_coli_K*12	Ecol	B	2070	649189	1756	269509
Gloeobacter_violaceus	Gvio	B	2189	674425	1938	342060
Jannaschia_CCS1	Jccs	B	2218	688899	1544	264010
*Mycobacterium_tuberculosis_CDC*1551	Mtub	B	2080	648475	1669	329744
*Nitrobacter_hamburgensis_X*14	Nham	B	1884	582774	2658	348975
Oceanobacillus_iheyensis	Oihe	B	1736	492986	1532	240549
*Pseudomonas_entomophila_L*48	Pent	B	2511	840891	2162	371158
Pyrobaculum_aerophilum	Paer	B	1361	350638	1069	183361
*Rhizobium_etli_CFN*_42	Retl	B	2069	647452	1833	268978
*Rhodopseudomonas_palustris_BisB*18	Rpal	B	2393	773048	2052	376368
*Shewanella_frigidimarina_NCIMB*_400	Sfri	B	2079	674741	1869	313709
*Shigella_sonnei_Ss*046	Sson	B	2076	629815	1789	447497
*Trichodesmium_erythraeum_IMS*101	Tery	B	2212	760875	2156	1033523
*Xanthomonas_oryzae_KACC*10331	Xory	B	2189	690992	1675	504847
Saccharomyces erevisiae	Sere	E	2944	1459439	2927	1239721
*Debaryomyces_hansenii_CBS*767	Dhan	E	3300	1532383	3160	1152782
Caenorhabditis elegans	Cele	E	11575	5102516	10143	16574978
Arabidopsis thaliana	Atha	E	15310	6215657	20459	24832614
Drosophila_melanogaster	Dmel	E	9841	5518997	7211	19651710

Note: E means *eukaryote*，B means bacteria and A means *archaea*

1.2 分析方法

1.2.1 密码对的频数

密码对的统计方法：从每个基因（已知的编码序列和理论预测的编码序列）起始密码子 A_1 开始，依次为第二个密码子 A_2 和第三个密码子 A_3 等，则密码对的组合是按照 A_1A_2、A_2A_3、…、$A_{n-1}A_n$ 的规则得到的（见图 1），其中 A_n 是紧邻终止密码子的有义密码子。全部有义密码对的组合模式数是 61^2=3721 个。我们没有考虑 A_nA_{stop} 密码对的组合模式（183 个模式）。统计了所有编码序列的各种密码对出现的频数。

在此引入“密码对的相对模式数”的概念：在密码对频数区组（见下文）内，实际出现的密码对组合模式数与总组合模式数之比。

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	…	…	A_n	A_{STOP}
AUG	CAC	CAA	GCG	UCA	CGG	CUA	UGG	…	…	CCG	UAG

Fig.1 Grouping of codons into codon pairs. Each codon from a coding sequence is assigned ‘A’. The numeric identifier（n）shows the codon position

1.2.2 三联体对的频数

基因间三联体对的统计方法是：从与编码序列对应的单链上取出基因与基因之间的序列，剔除随机重复序列、端粒序列及其互补链属于编码区的序列之后，与密码对的统计方法类似，从所得基因间序列的第一个三联体 B_1 开始，依次为三联体 B_2、B_3、…、B_N，则三联体对是按照 B_1B_2、B_2B_3、…、$B_{N-1}B_N$ 规则得到的，共有 64^2=4096 个三联体对的组合模式。按照这种方式得到所有基因间序列三联体对出现的频数。

引入“三联体对的相对模式数”的概念：在三联体对频数区组（见下文）内实际出现的三联体对组合模式数与总组合模式数之比。

需要说明的是基因间序列三联体对的取法，由于非编码序列不像编码序列那样有三联体密码子读框，我们对非编码序列的三种可能读框方式分别进行了统计，发现得到的三联体对相对模式数随其频数的分布没有实质差别，所以文中只选择从序列的第一个碱基开始的读框方式进行分析。

1.2.3 相对模式数随频数的分布

以密码对或三联体对的相对模式数为纵坐标，以频数为横坐标，得到不同物种基因组中密码对或三联体对的相对模式数随其频数的分布。

对于基因组较小的古菌和细菌，按密码对或三联体对出现频数的大小，以频数 20 为一个区组，频数在 0~19 记为 20，20~39 记为 40，依此类推。对基因组较大的真核生物取频数 50 为频数区组间隔。分别统计出每个生物在不同区组内密码对或三联体对出现的相对模式数，记为数组 $\{x,y\}$，x 表示其频数，y 表示在此区间内的相对模式数。 发现 35 种生物基因组中密码对或三联体对相对模式数随频数的分布规律一致，而且如果不考虑数组中一些偏离分布规律的异常数据点，则 x 与 y 可能满足一种分布。为了寻找 x 与 y 的关系，已采用各种分布对以上分布情况进行了拟合，最终发现分布曲线形状与 $\Gamma(\alpha,\beta)$分布最为相似，所以，选择用 $\Gamma(\alpha,\beta)$分布函数对这种分布进行拟合。

1.2.4 $\Gamma(\alpha,\beta)$分布函数与拟合方法

对于 $\Gamma(\alpha,\beta)$分布函数，

$$D(x)=x^{\alpha-1}\beta^{-\alpha}e^{-x/\beta}/\Gamma(\alpha) \quad (1)$$

式中 α、β 为参数，此分布的平均值为 $<x>=\alpha\beta$，n 阶矩 $\Delta^{(n)}=[<(x-<x>)^n>]^{1/n}$，其中 2 阶矩就是标准差，其形式还可以写为 $\Delta^{(2)}=\sqrt{\alpha}\,\beta$。按照 $\Gamma(\alpha,\beta)$分布模型，参数 α 是反映分布形状的一个重要参数，它与分布的偏态系数和峰态系数紧密相关。$\Gamma(\alpha,\beta)$分布的偏态系数（r_1）和峰态系数（r_2）分别为 $r_1=2\alpha^{-1/2}$，$r_2=6/\alpha$。理论上，一个分布的偏态系数（r_1）和峰态系数（r_2）都趋于零时，则它的性质趋于正态分布[22]，一般认为，当 $\alpha\geqslant10$ 时，$\Gamma(\alpha,\beta)$分布可近似认为是正态分布。参数 β 反映了样本的规模，与样本规模正相关。如果被拟合的分布与 $\Gamma(\alpha,\beta)$分布完全一致，则两者的平均值和 n 阶矩相等。

$\Gamma(\alpha,\beta)$分布中，x 所在的区组内，密码对模式数目为：

$$y_k(x_k)=N\cdot D(x_k)\cdot\Delta x_k \quad (2)$$

$$N=\sum y_k(x_k) \quad (3)$$

N 为总的密码对模式数 3721，Δx_k 为区组间隔，下

标 k 为区组编号。

二维{x,y}数组中统计平均值为：

$$\langle x\rangle=\frac{\sum x_k \bullet y_k}{\sum y_k} \quad (4)$$

n 阶矩为

$$\Delta^{(n)}=\left(\frac{\sum (x_k-\langle x\rangle)^n}{\sum y_k}\right)^{\frac{1}{n}} \quad (5)$$

x_k 的求和是对整个区组范围。

要获得最佳的拟合结果，需通过调节 $\Gamma(\alpha,\beta)$分布的特征参数 α 和 β 值，使得实际分布与 $\Gamma(\alpha,\beta)$分布的平均值和 n 阶矩值最接近。考虑到实际分布的离散性特点，我们的拟合办法是：分别计算 35 种生物密码对的相对模式数随频数分布的二阶矩 $\Delta_1^{(2)}$ 和长度平均值 $\overline{x_1}$，$\Gamma(\alpha,\beta)$分布函数的二阶矩 $\Delta_2^{(2)}$和长度平均值 $\overline{x_2}$，在 $0.8\leqslant\overline{x_1}/\overline{x_2}\leqslant1.2$ 的条件下，在二阶矩之差 $|\Delta_1^{(2)}-\Delta_2^{(2)}|$ 最小的邻域微调，使两者的三阶矩之差 $|\Delta_1^{(3)}-\Delta_2^{(3)}|$ 极小，来确定参数 α 和 β。

2　结　　果

2.1　密码对和三联体对相对模式数目随其实际频数的分布

按照上述方法，对 35 种生物基因组中编码序列的密码对、基因间序列三联体对的相对模式数随其实际频数的分布，用 $\Gamma(\alpha,\beta)$分布进行拟合，得到各基因组的分布参数 α 和 β 值，结果见图 2。图 3 作为一个例子，给出了大肠杆菌编码序列和基因间序列的实际分布和拟合曲线。表 2 列出了古菌、细菌和真核生物分布参数 α 和 β 的平均值，作为对比，我们还给出了随机序列中三联体对的相对模式数目随频数分布做的 $\Gamma(\alpha,\beta)$分布拟合结果，随机序列的长度与大肠杆菌基因组等长。

对所分析的 35 个基因组而言，密码对和三联体对的相对模式数目随其实际频数的分布与 $\Gamma(\alpha,\beta)$分布基本符合，实际分布和 $\Gamma(\alpha,\beta)$分布二阶矩在 $0.999\leqslant\Delta_1^{(2)}/\Delta_2^{(2)}\leqslant1.001$ 之间，密码对三阶矩的差别在 0.5%～16%以内，三联体对三阶矩的差别在 5.3%～23%以内。拟合结果说明实际分布可以用 $\Gamma(\alpha,\beta)$分布很好地描述（图 3）。

随机序列的 α 值接近 10，而实际分布的 α 值远小于 10，图 2 中，无论是编码序列还是基因间序列，α 值均小于 4，说明这种分布明显是非随机分布。

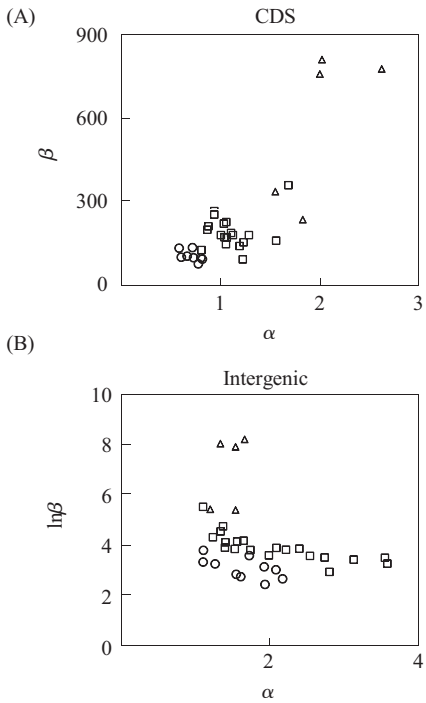

Fig.2　(A) Correlation between α and β of the fitting results of the distribution of the real frequencies of codon pairs in CDS in the 35 genomes. (B) Correlation between α and lnβ of the fitting results for the distribution of the real frequencies of intergenic triplet pairs in the 35 genomes. ○: *Archaea*; □: Bacteria; △: *Eukarya*

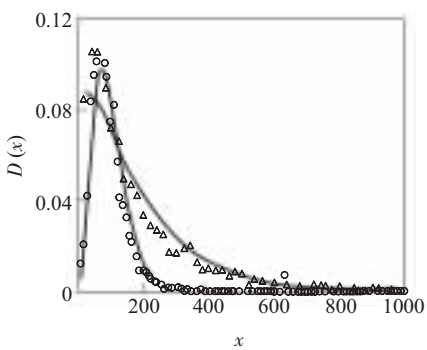

Fig.3　The fitting curves of $\Gamma(\alpha,\beta)$ distribution for codon pairs and triplet pairs with the variation of their real frequencies in *E.coli* genome. The x-axis denotes the frequencies of codon pairs and triplet pairs, the y-axis denotes the fraction from all possible types of codon pairs and triplet pairs. △: CDS; ○: Intergenic

对编码序列，α 的平均值按照古菌、细菌、真核生物的顺序逐渐增大。古菌的 α 平均值最小且小于 1，细菌的 α 值比古菌高 1.54 倍，真核生物最大，比古菌高 2.8 倍，比细菌高 1.8 倍。三类生

Table 2 The average $\bar{\alpha}$ and $\bar{\beta}$ of $\Gamma(\alpha,\beta)$ distributions for the real frequencies of codon pairs and intergenic triplet pairs in 35 genomes and two parameters α and β of $\Gamma(\alpha,\beta)$ distribution for a random sequence triplet pairs

Domain	$\bar{\alpha}_1$	$\bar{\beta}_1$	$\bar{\alpha}_2$	$\bar{\beta}_2$
Archaea	0.72±0.08	106±15	1.63±0.28	24±10
Bacteria	1.11±0.22	188±57	2.08±0.34	60±50
Eukarya	2.01±0.35	581±271	1.51±0.19	1963±1623
Random	9.51	41		

The subscript 1 represents codon pairs and the subscript 2 denotes triplet pairs

物拟合结果的分布参数有明显的区别（图 2 和表 2）。以它们的 α 和 β 平均值为参数做出相应的 $\Gamma(\alpha,\beta)$分布曲线（见图 4）。可以直观地看出，密码对的相对模式数随其实际频数的分布有明显的差别，随着生物的进化，分布逐渐向随机分布靠近，也就是说密码对的使用逐渐朝向随机方向进化，或者说紧邻密码子之间的关联在逐渐减弱。古菌的 α 平均值小于 1，其分布类似于一个指数下降的分

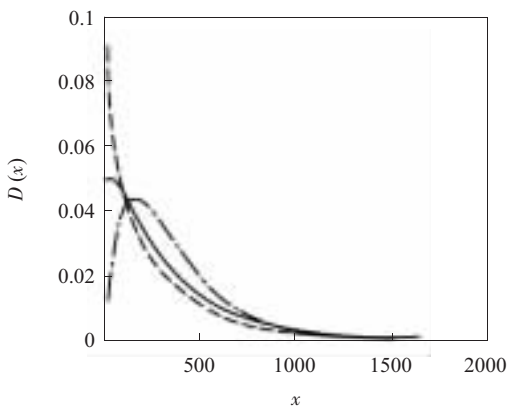

Fig.4 The $\Gamma(\alpha,\beta)$ distribution curves for the distribution of frequency of occurrence of codon pairs. ---: *Archaea*; ——: Bacteria; —·—: *Eukarya*

布，也就是说，绝大多数密码对模式出现的频数都很少，密码对的非随机使用非常明显。大多数细菌和所有真核生物的 α 值大于 1，其分布有一个峰值，特点是出现频数很少的密码对模式和出现频数很大的密码对模式都很少。从进化的角度来看，古菌密码子之间的搭配受到很强的选择压力，由于处于特定的环境，在选择压力下，密码子之间的非随机搭配随进化变化不大，因此其编码序列仍旧保持了古生物的特征，细菌和真核生物则不同，随着环境压力的变化，密码子之间的关联变得越来越弱。同是原核生物，古菌和细菌的密码对使用存在明显的差别，古菌与真核生物密码对的使用差别更大。这个结果与三界理论一致[23]。

对基因间序列，α 平均值的变化规律与编码序列明显不一样。古菌和真核生物的 α 平均值接近，它们明显小于细菌的 α 平均值，即分布偏离随机序列更远。这说明相对于细菌而言，古菌和真核生物基因间序列三联体对的搭配更有序或碱基之间的关联更强。同样是原核生物，古菌和细菌的基因间序列走的不同进化道路，古菌和真核生物一致，而细菌则与它们明显不同。这结论似支持两界理论。

β 值在理论上反映了样本的规模，从表 2 可以看出，它的确反映了基因组的规模，无论是编码序列还是基因间序列，基因组越大，β 的平均值越大（由于所统计真核生物样本少且基因组之间规模差别很大，导致 β 值的标准差较大）。这点也间接说明实际分布是遵从 $\Gamma(\alpha,\beta)$分布的。

生命的复杂性之一在于它的特异性。密码对的拟合结果显示沼泽红假单胞菌（Rpal）和蓝细菌（Avar）这两种细菌很特殊。沼泽红假单胞菌的 α 和 β 值分别为：0.82 和 123，其 α 值在古菌的平均值范围之内，接近古菌 Tkod（α=0.82）、Ssol（α=0.81）和 Telo（α=0.81），β 值是古菌 β 平均值的 1.16 倍。蓝细菌的 α 和 β 值分别为：1.65 和 355，其 α 值和 β 值在真核生物的平均值范围内。$\Gamma(\alpha,\beta)$分布的参数 α 反映了分布的形状，以上两种细菌的 α 值分别靠近古菌和真核生物（见图 2），反映了这两种生物在进化上的特殊性。也就是说，沼泽红假单胞菌在密码对使用上具有古菌的特征，而蓝细菌具有真核生物的特征。蓝细菌在细胞结构上既有与细菌相似之处又有与真核藻类相似而与细菌不同的特征[24]，在我们的结论中反映出其在密码对的使用上也靠近真核。这或许证明密码对的非随机性除了表现某类生物的共性之外，同时也是物种特异的选择压力的结果[25,26]。正如果蝇在密码对使用与密码子的关系上体现出与原核生物相似之处[14]。

2.2　密码对和三联体对相对模式数随其相对频数的分布

在理论上，Γ(α,β)分布的参数 α 反映了分布的形状，参数 β 反映了样本的规模。如果实际分布确实是 Γ(α,β)分布的话，这一特性必须满足。上节的结果仅仅说明实际分布可以用 Γ(α,β)分布很好地描述，并未验证实际分布就是 Γ(α,β)分布。为了进一步验证密码对和三联体对相对模式数随其频数的分布就是 Γ(α,β)分布，将各个生物基因组的规模折合成一样的大小，即将各个模式生物中统计的密码对和三联体对的实际频数折合成等长基因组序列内出现的相对频数，折合的序列长度均为 100 万个密码对（约为闪烁古生球菌 Aful 基因组的大小）。如果实际分布确实是 Γ(α,β)分布，则在折合成相等基因组规模下，拟合得到的 α 值应该与实际基因组规模下得到的 α 值一样。这样还可以直接比较编码序列和基因间序列的分布变化，了解不同类生物密码对和三联体对使用的进化差别。

基因组里全部编码序列中某个密码对模式出现的相对频数，记为 N。基因组中全部基因间序列中某个三联体对模式出现的相对频数，记为 N'。其定义分别为：

$$N=\frac{10^6}{N_{\mathrm{TOT}}}\times N_{\mathrm{OBS}} \tag{6}$$

$$N'=\frac{10^6}{N'_{\mathrm{TOT}}}\times N_0 \tag{7}$$

N_{OBS} 是基因组里所有编码序列中某个密码对模式出现的频数，N_{TOT} 是对应生物的所有编码序列的密码对总数，N_0 是全部基因间序列中某个三联体对模式出现的频数，N'_{TOT} 是全部基因间序列三联体对的总数。

这样可利用与前面相同的统计方法，对 35 种生物基因组中所有编码序列的密码对和基因间序列三联体对的相对模式数目，随其相对频数的分布用 Γ(α,β)分布进行拟合，结果见图 5。为了和实际频数的分布比较，这里给出大肠杆菌的编码序列密码对（三角形数据点）与基因间序列三联体对（圆圈数据点）相对频数的拟合曲线（见图 6）。表 3 给出拟合后三类生物 α 与 β 的平均值。

消除了统计序列规模的影响后，与结果 **2.1** 比较，分布参数 α 基本没有改变，β 值的改变与基因组规模改变成正相关。在基因组规模相等的条件下，因为分布的平均值 <x>=αβ 不变而造成 β 随 α 变化。这一结论也可从表 2 和表 3 中平均值以及图

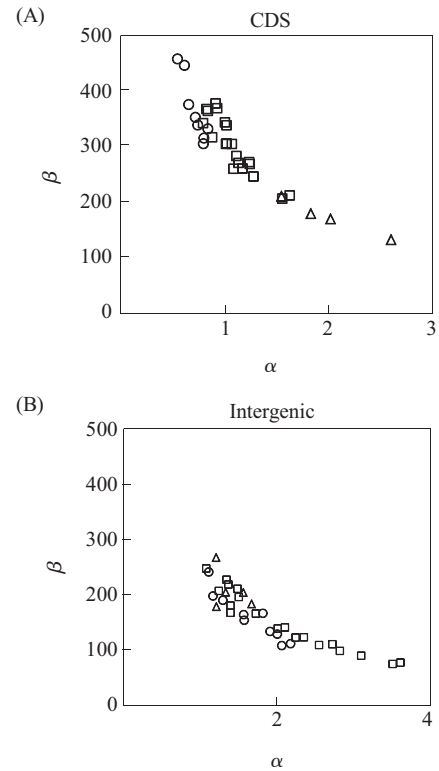

Fig.5　(A) Correlation between α and β of the fitting results of the distribution of the relative frequencies of codon pairs in CDS in the 35 genomes. (B) Correlation between α and β of the fitting results for the distribution of the relative frequencies of intergenic triplet pairs in the 35 genomes. ○: *Archaea*; □: Bacteria; △: *Eukarya*

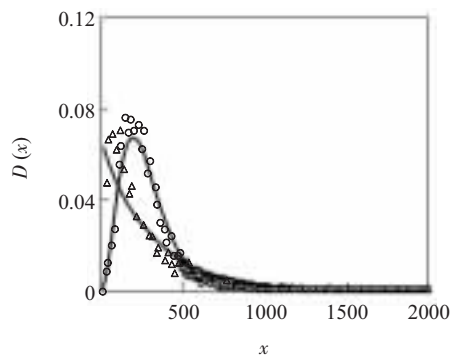

Fig.6　The fitting curves of Γ(α,β) distribution for codon pairs and triplet pairs with the variation of their relative frequencies in *E.coli* genome. The *x*-axis denotes the relative frequencies of codon pairs and triplet pairs, the *y*-axis denotes the fraction from all possible types of codon pairs and triplet pairs. △: CDS; ○: Intergenic

3 和图 6 的比较看出。这说明拟合参数 α 与基因组的规模无关，只反映分布的形状。也就是说实际分布确实遵循 Γ(α,β)分布。

Table 3 The average $\bar{\alpha}$ and $\bar{\beta}$ of $\Gamma(\alpha,\beta)$ distributions for the relative frequencies of codon pairs and intergenic triplet pairs in 35 genomes

Domain	$\overline{\alpha_1}$	$\overline{\beta_1}$	$\overline{\alpha_2}$	$\overline{\beta_2}$
Archaea	0.72±0.08	360±50	1.65±0.28	158±41
Bacteria	1.08±0.22	297±34	2.06±0.34	143±52
Eukarya	2.01±0.35	170±50	1.50±0.19	205±37

The subscript 1 represents codon pairs and the subscript 2 denotes triplet pairs

将三联体对分布和密码对分布放在一起比较：尽管三联体对的模式总数（含 3 个终止密码子）比密码对的模式总数多 375 个，但差别只有 9.16%，对其分布参数 α 比较不会有原则性的改变。比较表 3 可以看出基因间序列的分布与编码序列的分布的差别。在同一类生物内部，基因间序列和编码序列的 α 值明显不一样，古菌和细菌基因间序列的 α 值明显大于其对应编码序列的 α 值，分别为 2.26 和 1.87 倍，即相对而言基因间序列的分布比编码序列的分布更随机一些，原核生物编码序列紧邻密码子的搭配比基因间序列三联体对之间的搭配更有序，也可以说原核生物编码序列的组织比基因间序列的组织更加有序。而真核生物基因间序列的 α 值小于其编码序列的 α 值，即相对而言，真核生物编码序列紧邻密码子的搭配比基因间序列三联体对之间的搭配更随机一些。这一特点也许是真核生物与原核生物的本质区别。

一些研究者通过计算全基因组序列 k-mer（k 碱基联体）的模式反映基因组进化的关系[27]。根据本文的结论，编码序列与基因间序列的进化方式不同，若不区分编码序列和非编码序列而整体考虑基因组 DNA 序列，得到的进化结果是不可靠的。

3 讨 论

密码对的非随机使用已经成为大家公认的事实。哪些因素影响密码对的非随机性，所反映出问题的实质是什么，是摆在人们面前的一个较复杂的问题。从宏观上看，有些密码对的过表达或许是蛋白质折叠的需要。从特定密码对的核苷酸相互作用的微观角度来看，这些因素之间的规律还远没有搞清。可以肯定的是，各个物种所特有的因素影响着密码子的搭配[28]。应该看到，从密码对（或三联体对）的相对模式数随频数的分布出发，得到构成编码序列与基因间序列组成结构与生物进化较为明确的关系，是系统研究这一问题的一个理论突破口。

近年来，$\Gamma(\alpha,\beta)$分布被广泛用于系统发生学的研究[29-31]，密码对（或三联体对）的相对模式数随频数的分布是 $\Gamma(\alpha,\beta)$分布，这里面还包含着什么深刻的内涵以及为什么它遵守 $\Gamma(\alpha,\beta)$分布而不是别的什么分布，是值得进一步思考的。

密码对（或三联体对）的相对模式数随频数的分布与生物进化的关系，在一定程度上反映了编码序列和基因间序列的结构与生物进化的关系。根据本文的结论，从进化角度看，细菌和真核生物这两类序列结构的进化方式明显不同，细菌的编码序列进化相对保守（$\Gamma(\alpha,\beta)$分布偏离随机序列更远），而真核生物基因间序列的进化相对保守。讨论生物进化最有争议的是古菌，古菌是 90 年代分类学上新定义的一类微生物，具有区别于细菌、真菌的独特生理生化特性。其基因组序列遵循不同的进化方式，构成了古菌独特的进化方式。它在某些方面像细菌，在某些方面像真核生物，如在翻译起始机制上[32]。这些独特之处使人们认识到，在基因组进化历程中，古细菌与细菌和真核微生物之间以及各个物种之间，显然皆发生过大量的基因交换，对单细胞进化而言，基因除垂直传递外，横向的或称侧向的基因转移也十分繁多。从我们的分析结果看，古菌编码序列的特征离真核最远，而基因间序列则更接近真核生物。因此不区分编码序列和非编码序列而笼统分析基因组全序列来研究生物进化，其结果会淹没组成基因组序列不同结构的内在信息。

总之，有关密码对使用的这种分布与生物进化的相关性，使我们确信了密码对使用的非随机性与生物进化的关系，进一步的工作重点应该是在消除同义密码子使用的差别和序列 GC 含量的差别后，构造一组特征参量，探讨每个密码对的使用与基因组进化的关系，寻找与进化敏感的密码子搭配及其所反映的生物学意义，相信随着研究的深入，将有更多的特征信息和内在规律被发现和利用。

参考文献：

[1] Andersson SGE, Kurland CG. Codon preferences in free living microorganisms. *Microbiol Rev*, 1990,54:198~210

[2] Kurland CG. Major codon preference: theme and variation. *Biochem Soc Trans*, 1993,21:841~845

[3] Sharp PM, Matassi G. Codon usage and genome evolution. *Curr Opin Genet Dev*, 1994,4:851~860

[4] Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 1985,2:12~34

[5] Berg OG, Silva PJ. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res*, 1977,25:397~404

[6] Fedorov A, Saxonov S, Gilbert W. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res*, 2002,30:1192~1197

[7] Mcvean GAT, Hurst GDD. Evolutionary lability of context-dependent codon bias in bacteria. *J Mol Evol*, 2000,50: 264~275

[8] Jukes TH, Bhushan V. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *J Mol Evol*, 1986,24:864~875

[9] Sueoka N. Directional mutation pressure, selective constraints, and genetic equilibria. *J Mol Evol*, 1992,34:95~114

[10] Gutman GA, Hatfield GW. Non-random utilization of codon pairs in *Escherichia coil*. *Proc Natl Acad Sci USA*, 1989, 86(10):3699~3703

[11] Cheng L, Goldman E. Absence of effect of varying Thr-Leu codon pairs on protein synthesis in a T7 system. *Biochemistry*, 2001,40:6102~6106

[12] Irwin B, Heck JD, Hatfield GW. Codon pair utilization biases influence translational elongation step times. *J Biol Chem*, 1995,270:22801~22806

[13] Yarus M, Folley LS. Sense codons are found in specific contexts. *J Mol Biol*, 1985,182:529~540

[14] Ross Buchan J, Aucott LS, Stansfield I. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res*, 2006,34(3):1015~1027

[15] Boycheva S, Chkodrov G, Ivanov L. Codon pairs in the genome of *Escherichia coli*. *J Bioinformatics*, 2003,19 (8): 987~998

[16] Moura G, Pinheiro M, Silva R, Miranda I, Afreixo V, Dias G, Freitas A，Oliveira JL，Santos MA. Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biol*, 2005,6:R28

[17] Folley LS，Yarus M. Codon contexts from weakly expressed genes reduce expression *in vivo*. *J Mol Biol*, 1989,209: 359~378

[18] Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G. Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res*, 984,12:6663~6671

[19] Smith D, Yarus M. tRNA-tRNA interactions within cellular ribosomes. *Proc Natl Acad Sci USA*, 1989,86:4397~4401

[20] Curran JF, Poole ES, Tate WP, Gross BL. Selection of aminoacy1-tRNAs at sense codons: the size of the tRNA variable loop determines whether the immediate 3' nucleotide to the codon has a context effect. *Nucleic Acids Res*, 1995, 23:4104~4108

[21] Stenström CM, Jin HN, Major LL, Tate WP, Isaksson LA. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene*, 2001,263:273~284

[22] 方开泰, 许建伦. 统计分布. 北京: 科学出版社, 1987. 331~332

[23] Skophammer RG, Herbold CW, Rivera MC, Servin JA, Lake JA. Evidence that the root of the tree of life is not within the archaea. *Molecular Biology and Evolution*, 2006,23 (9): 1648~1651

[24] 北京大学生命科学学院编写组. 生命科学导论. 北京: 高等教育出版社，2000. 151~177

[25] Purvis IJ, Bettany AJ, Santiago TC, Coggins JR, Duncan K, Eason R, Brown AJ. The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. *J Mol Biol*, 1987,193:413~417

[26] Rothman JE. Polypeptide chain binding proteins: catalysts of protein folding and related processes in cells. *Cell*, 1989,59: 591~601

[27] Li W, Fang W, Ling L, Wang JH, Xuan ZY, Chen RS. Phylogeny based on whole genome as inferred from complete information set analysis. *J Biol Phys*, 2002,28(4):439~447

[28] Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF. Crystal structure of the ribosome at 5.5 Å resolution. *Science*, 2001,292:883~896

[29] Hsieh LC, Luo LF, Ji FM. Minimal model for genome evolution and growth. *Physical Review Letters*, 2003,90:1~4

[30] 王树林, 王 戟, 陈火旺, 张鼎兴. *k*- 长 DNA 子序列频数分布研究. *生物物理学报*, 2006,22(3):178~195

[31] 冯立芹, 李宏. 基因组中开阅读框架长度的分布模型与基因组进化. *生物物理学报*, 2004,20(5):375~381

[32] Saito R, Tomita M. Computer analyses of complete genomes suggest that some archaebacteria employ both eukaryotic and eubacterial mechanisms in translation initiation. *Gene*, 1999, 238:79~83

CODON PAIRS USAGE AND GENOME EVOLUTION

WANG Fang-ping, LI Hong

(*Department of Physics, College of Sciences and Technology, Inner Mongolia University, Hohhot 010021, China*)

Abstract: The distributions of the relative mode numbers of codon pairs in protein coding sequences and triplet pairs in intergenic sequences are analysed in ten *archaea*, twenty bacteria and five *eukaryote* genomes. It is proposed that these distributions are in accord with $\Gamma(\alpha,\beta)$ distribution. The shape parameter α of $\Gamma(\alpha,\beta)$ distribution has distinctive correlation with the genome evolution. The modes of evolution for protein coding sequences and intergenic sequences are significantly different. In the course of evolution, the shape of $\Gamma(\alpha,\beta)$ distribution for protein coding sequences moves toward the random distribution (The α value increased gradually). For intergenic sequences, however, the shape of $\Gamma(\alpha,\beta)$ distribution of *archaea* is similar to *eukaryote*, and distinctly different from bacteria.

Key Words: Genome evolution; Codon pairs; Triplet pairs; $\Gamma(\alpha,\beta)$ distribution

---

This work was supported by grants from The National Natural Science Foundation of China (30660044) and The Special Foundation of Education Ministry of China for Doctorial Degree Authorization Unit (20050126003)

Received: Oct 27, 2006

Corresponding author: LI Hong, Tel: +86(471)6678889, E-mail: lihong499@hotmail.com