

20个物种同义密码子偏性的比较分析^{*}

刘庆坡¹, 冯英¹, 董辉²

(1 浙江大学 农业与生物技术学院 农学系, 浙江 杭州 310029;

2 新疆克拉玛依市独山子区 晟通热力有限公司, 新疆 独山子 833600)

[摘要] 对6种双子叶植物、4种单子叶植物、7种多细胞真核动物、2种单细胞真核生物及1种原核生物(大肠杆菌)的编码序列(Coding DNA Sequence, CDS)进行了比较研究。结果发现, 各供试物种密码子的用法存在明显的位置依赖, 编码区域不同区段对密码子选择使用的程度差异很大, 表明基因的同义密码子偏性与其编码区不同区域的碱基环境密切相关; 基因组差异是造成密码子使用偏性的首要因素。对“翻译起始区”和“翻译终止区”的特征分析发现, 大部分供试物种编码序列中的“翻译起始区”碱基偏置强烈, “翻译终止区”相对较弱, 暗示“翻译起始区”的密码子使用对提高蛋白翻译的效率和精确性更为重要。不同物种中G+C含量对其同义密码子偏性的贡献率差异很大, 但并非是影响同义密码子使用偏性的唯一因素。可以认为, 基因编码区域密码子偏性的研究能为搜寻物种进化的分子机理提供线索。

[关键词] 密码子使用偏性; 位置依赖; G+C含量

[中图分类号] Q943.2

[文献标识码] A

[文章编号] 1671-9387(2004)07-0067-05

编码同一氨基酸的密码子称为同义密码子(Synonymous Codons)。在蛋白质编码过程中, 某一物种或某一基因通常倾向于使用一种或几种特定的同义密码子, 这种现象称为同义密码子的使用偏性(Synonymous Codon Usage Bias)^[1]。

从原核生物到真核生物, 其基因组中同义密码子使用偏性的现象广泛存在, 这一现象的产生与诸多因素有关, 如基因的表达水平^[2,3]、翻译起始效应^[4,5]、基因的碱基组分^[6]、某些二核苷酸的出现频率^[7,8]、G+C含量^[9,10]、基因的长度^[11]、tRNA的丰富度^[12,13]、蛋白质的结构^[14]以及密码子-反密码子间结合能的大小^[15]等。在不同物种或细胞器^[16]基因中, 其偏性程度因作用的因素数量以及各因素在同义密码子偏性产生过程中所占决定系数的大小而不同。

基因编码区的同义密码子偏性与其所在基因组环境的碱基组成有关^[17~19]。比如, 人类基因组的平均G+C含量约为50%, 并且其66%的基因位于富含和极端富含G+C的区域^[9]。而线虫基因组的G+C含量仅为36%, 其明显偏爱A或U结尾的密码子^[20], 而人类基因则更喜欢使用以C或G, 尤其C结尾的同义密码子^[9]。

研究^[21,22]表明, 基因密码子的使用与基因编码

蛋白的结构和功能, 以及基因表达的生理功能有密切联系。mRNA中稀有密码子的使用与蛋白质结构域的连接区和规则二级结构单元的连接区有关^[23]。顾万君等^[14]采用聚类分析方法研究了蛋白编码基因的密码子使用偏性, 发现密码子的使用与蛋白质的三级结构有很大的相关性, 从而证明DNA的一维信息中蕴含着蛋白质的三级结构信息。

本研究同时对多种植物、动物和微生物(原核生物)的包含完整编码序列(Coding DNA Sequence, CDS)的全长cDNA序列进行了比较, 重点分析了基因编码区不同区段密码子使用的特点及与G+C含量的相关性, 以期对蛋白质三级结构的预测及搜寻物种进化的分子机理提供线索。

1 材料和方法

1.1 材料

供试物种有6种双子叶植物(拟南芥、烟草、欧洲油菜、番茄、菜豆、豌豆)、4种单子叶植物(普通小麦、玉米、水稻、大麦)、7种多细胞真核动物(大鼠、家鼠、人、猪、斑马鱼、果蝇、线虫)、2种单细胞真核生物(酿酒酵母、裂殖酵母)和大肠杆菌共20个物种。从GenBank数据库(<ftp://ftp.ncbi.nih.gov/>)

^{*} [收稿日期] 2003-06-09

[作者简介] 刘庆坡(1976-), 男, 河北曲阳人, 在读博士, 主要从事基因组学研究。Email: liuqp@genomics.org.cn

genbank/, release 135) 中下载这 20 个物种所有包含完整编码序列(CDS)的全长 cDNA 序列。编写 Perl 程序, 对所需样本按标准格式初步解读, 即过滤掉那些不以 ATG 作起始密码子和不以 TAA, TAG, TGA 为终止密码子, 且编码序列中含有除 A, T, C, G 以外字符的 CDS 序列。

1.2 方法

编写 Perl 程序完成全部序列处理和统计分析, 利用 SPSS 10.0 多元分析软件进行相关性显著性测验。

1.2.1 同义密码子偏性的衡量指标 采用密码子偏好参数(Codon Preference Parameter, CPP)^[24]来度量同义密码子的使用偏好。CPP 值为 0~18, 越接近 18 表示密码子被非随机使用的程度越高。其对基因编码区域总的碱基组成不敏感, 适于比较基因间或物种间密码子使用偏性的大小。

1.2.2 编码区同义密码子偏性使用的特征分析 编写 Perl 程序, 将每个供试物种中所获得的 cDNA 序列均分成 6 个部分, 分别计算相应区段的平均 CPP 值, 分析各个区段同义密码子的使用特点。真核生物的信号肽一般长 22~66 个氨基酸(<http://www.cbs.dtu.dk/services/SignalP/sp-lengths.html>)。故将紧邻 ATG 下游的 66 个碱基组成的序列(+4 至+69 位点)定义为“翻译起始区”; 终止密码子上游 66 bp 长的序列段定义为“翻译终止区”; 考察各供试物种同义密码子偏性在这两个区域的使用情况。

2 结果与分析

2.1 编码区 G+C 和 GC3 含量与 CPP 的关系

供试 cDNA 序列编码区域的 G+C 和 GC3 含量与 CPP 相关性测验的结果(表 1)表明, 禾本科植物(水稻、普通小麦、大麦和玉米)的 G+C 和 GC3 含量与 CPP 均呈高度正相关(r 值分别为 0.590~

0.801 和 0.670~0.805), 表明在这些物种中, G+C 含量在其同义密码子使用偏性的产生过程中起着决定性的作用, 而且同义密码子偏性强烈的基因往往偏爱使用 C 或 G 结尾的密码子。双子叶植物 G+C 和 GC3 含量与 CPP 的相关性与单子叶植物不同, 其中欧洲油菜中两者的相关系数不显著, 即 G+C 含量对该物种同义密码子使用偏性的影响甚微; 除拟南芥和菜豆为显著或极显著正相关外, 豌豆、番茄和烟草 G+C、GC3 与 CPP 的相关系数均为负值; 豌豆和番茄中 G+C 及 GC3 与 CPP 的相关系数均为极显著负相关($r = -0.136^{**}$, -0.128^{**} 和 -0.185^{**} , -0.200^{**}), 表明其基因偏爱使用 A+T 含量高, 尤其 A 或 T 结尾的密码子; 烟草的 G+C 与 CPP 的相关系数为极显著负相关($r = -0.159^{**}$), 而 GC3 与 CPP 的相关性不大($r = -0.012$), 表明 G+C 含量虽然影响其同义密码子的使用偏性, 但作用较小, 而且对密码子“摇摆”位点的选择性不大。多细胞真核动物中, 除线虫外其他物种的相关系数均为极显著正值, 但明显低于单子叶植物, 其中人最高(0.459, 0.444), 大鼠最小(0.271, 0.278)。线虫中 G+C 和 GC3 含量与 CPP 的相关系数均呈高度负相关(分别为 -0.571^{**} 和 -0.532^{**}), 表明 A+T 含量对其同义密码子的使用偏性占有较大的决定系数。单细胞真核生物酿酒酵母和裂殖酵母的相关系数相对较低(0.080~0.177, 0.097~0.232)。拟南芥、菜豆、人、猪、斑马鱼、家鼠、大鼠、果蝇、酿酒酵母以及裂殖酵母等 10 个物种中, G+C 和 GC3 与 CPP 均达到极显著或显著正相关(r 值小于或约等于 0.5), 但彼此间变异较大。大肠杆菌中 G+C 和 GC3 与 CPP 的相关系数均为负向极显著($r = -0.221^{**}$ 和 -0.165^{**}), 但数值较小, 表明其对同义密码子偏性的影响可能不及其他因素。

表 1 供试物种 G+C 及 GC3 含量与密码子偏性指标 CPP 的相关性测验

Table 1 Correlation coefficients between CPP and G+C, GC3 in twenty species

物种 Species	r_{G+C}	r_{GC3}	物种 Species	r_{G+C}	r_{GC3}
水稻 <i>Oryza sativa</i>	0.801**	0.805**	人 <i>Homo sapiens</i>	0.459**	0.444**
普通小麦 <i>Triticum aestivum</i>	0.790**	0.791**	猪 <i>Sus scrofa</i>	0.391**	0.459**
大麦 <i>Hordeum vulgare</i>	0.590**	0.670**	家鼠 <i>Mus musculus</i>	0.428**	0.427**
玉米 <i>Zea mays</i>	0.697**	0.719**	大鼠 <i>Rattus norvegicus</i>	0.271**	0.278**
拟南芥 <i>Arabidopsis thaliana</i>	0.249**	0.159**	斑马鱼 <i>Danio rerio</i>	0.292**	0.275**
欧洲油菜 <i>Brassica napus</i>	0.106	-0.007	果蝇 <i>Drosophila melanogaster</i>	0.276**	0.333**
菜豆 <i>Glycine max</i>	0.146**	0.104*	线虫 <i>Caenorhabditis elegans</i>	-0.571**	-0.532**
豌豆 <i>Pisum sativum</i>	-0.136**	-0.185**	酿酒酵母 <i>Saccharomyces cerevisiae</i>	0.080**	0.097**
番茄 <i>Lycopersicon esculentum</i>	-0.128**	-0.200**	裂殖酵母 <i>Schizosaccharomyces pombe</i>	0.177**	0.232**
烟草 <i>Nicotiana tabacum</i>	-0.159**	-0.012	大肠杆菌 <i>Escherichia coli</i>	-0.221**	-0.165**

综上所述, 绝大多数物种编码区域总的碱基组成显著(正或负向)影响其同义密码子的使用偏性, 但作用的方向及影响程度不同。20 个供试物种中, 只有欧洲油菜的两个相关系数不显著, 这不是偶然现象, 因为同义密码子偏性的产生是个十分复杂的事件, 是多种因素综合作用的产物。

2.2 编码区不同区段同义密码子偏性使用的特点

在“翻译起始区”和“翻译终止区”间的同义密码子偏性上, 不同物种有明显差异(表 2)。单子叶植物“翻译起始区”的同义密码子偏性均大于“翻译终止区”; 暗示“翻译起始区”密码子的使用对提高蛋白翻译的效率和精确性更为重要。而双子叶植物“翻译起

始区”的偏性与“翻译终止区”差异较小, 前者是 2.8~4.8, 后者为 2.6~4.7。值得注意的是, 菜豆、豌豆和烟草起始区密码子的偏性变异反而低于终止区。除斑马鱼、果蝇和线虫外, 多细胞真核动物中的“翻译起始区”同义密码子偏性均大于“翻译终止区”; 这一点与单子叶植物有相似之处, 只是程度稍小。起始区和终止区 CPP 值分别为 4.3~4.9 和 2.6~4.6。斑马鱼、果蝇、线虫以及单细胞真核生物酿酒酵母和裂殖酵母起始区和终止区 CPP 值的变化刚好与其相反, 暗示在这些物种中, “翻译终止区”的密码子用法或许比较重要。大肠杆菌两个区域的 CPP 值分别为 3.77 和 3.82, 密码子偏性差异不明显。

表 2 20 个物种“翻译起始区”和“翻译终止区”密码子偏性的比较

Table 2 Comparison of codon usage between the translational initiation region and translational termination region in twenty species

物种 Species	翻译起始区 Transla- tional initiation region	翻译终止区 Transla- tional tem ination region	物种 Species	翻译起始区 Transla- tional initiation region	翻译终止区 Transla- tional tem ination region
水稻 <i>Oryza sativa</i>	8.210	5.578	人 <i>Homo sapiens</i>	4.328	2.641
普通小麦 <i>Triticum aestivum</i>	7.867	6.363	猪 <i>Sus scrofa</i>	4.918	4.640
大麦 <i>Hordeum vulgare</i>	8.167	6.585	家鼠 <i>Mus musculus</i>	4.527	3.113
玉米 <i>Zea mays</i>	6.932	5.570	大鼠 <i>Rattus norvegicus</i>	4.461	3.877
拟南芥 <i>Arabidopsis thaliana</i>	2.974	2.615	斑马鱼 <i>Danio rerio</i>	2.667	4.072
欧洲油菜 <i>Brassica napus</i>	3.573	3.257	果蝇 <i>Drosophila melanogaster</i>	4.116	4.851
菜豆 <i>Glycine max</i>	2.858	3.353	线虫 <i>Caenorhabditis elegans</i>	3.627	4.275
豌豆 <i>Pisum sativum</i>	3.774	4.639	酿酒酵母 <i>Saccharomyces cerevisiae</i>	3.557	4.115
番茄 <i>Lycopersicon esculentum</i>	4.832	4.694	裂殖酵母 <i>Schizosaccharomyces pombe</i>	5.211	5.412
烟草 <i>Nicotiana tabacum</i>	3.046	3.965	大肠杆菌 <i>Escherichia coli</i>	3.771	3.818

从表 3 可以看出, 所有供试物种的同义密码子偏性程度均呈现双峰型, 即第 1 和第 4 区段 CPP 值相对高于其他区段。其中 4 个单子叶植物(水稻、普通小麦、大麦和玉米)编码区域第 1 区段的 CPP 值均超过 9(9.4~11.5), 并显著高于其他区段(2.8~7.1), 这可能与上述 4 个物种具有信号肽或与存在 G+C 的含量梯度有关。而 6 个双子叶植物(拟南芥、烟草、欧洲油菜、番茄、菜豆及豌豆)中, 6 个区段

同义密码子偏性的程度则无明显差异, 其 CPP 值为 3.8~8.0。大鼠、家鼠、人、猪、斑马鱼、果蝇、线虫、酿酒酵母、裂殖酵母等物种中, 第 1 和第 4 区段的密码子偏性程度比较接近, 而且均略高于其他区段, 暗示这些物种编码序列的不同区域对同义密码子的选择可能不太强烈。不过, 原核生物大肠杆菌 1、4 区段的 CPP 值(分别为 9.2 和 9.7)明显高于其余 4 个区段(3.5~4.4)。

表 3 20 个物种编码区 6 个不同区段密码子偏性的比较

Table 3 Comparison of codon usage among six different sections in coding region in twenty species

物种 Species	区段 Section						物种 Species	区段 Section					
	1	2	3	4	5	6		1	2	3	4	5	6
水稻 <i>Oryza sativa</i>	11.463	4.173	3.299	7.090	3.733	4.230	人 <i>Homo sapiens</i>	6.854	4.614	4.593	6.496	4.746	4.903
普通小麦 <i>Triticum aestivum</i>	11.498	4.012	3.614	6.843	3.858	4.129	猪 <i>Sus scrofa</i>	6.312	3.972	4.404	6.235	4.047	4.796
大麦 <i>Hordeum vulgare</i>	11.417	3.966	2.852	6.901	3.509	4.303	家鼠 <i>Mus musculus</i>	6.553	4.721	4.827	6.159	4.933	5.082
玉米 <i>Zea mays</i>	9.390	3.764	3.349	6.532	3.499	3.503	大鼠 <i>Rattus norvegicus</i>	8.000	4.263	3.914	7.549	4.180	4.323
拟南芥 <i>Arabidopsis thaliana</i>	7.344	5.060	5.405	6.784	5.402	5.643	斑马鱼 <i>Danio rerio</i>	5.480	4.565	4.641	5.932	4.803	4.990
欧洲油菜 <i>Brassica napus</i>	6.170	3.838	4.894	5.911	5.066	4.943	果蝇 <i>Drosophila melanogaster</i>	8.000	4.263	3.914	7.549	4.180	4.323
菜豆 <i>Glycine max</i>	6.533	4.563	5.068	7.581	5.922	5.793	线虫 <i>Caenorhabditis elegans</i>	7.480	6.388	6.402	7.719	6.223	6.017
豌豆 <i>Pisum sativum</i>	8.064	5.815	6.372	7.623	6.028	6.331	酿酒酵母 <i>Saccharomyces cerevisiae</i>	8.036	6.316	6.368	7.970	6.322	6.761
番茄 <i>Lycopersicon esculentum</i>	7.599	5.663	5.822	7.401	6.162	6.366	裂殖酵母 <i>Schizosaccharomyces pombe</i>	8.099	5.865	5.879	7.911	5.745	5.833
烟草 <i>Nicotiana tabacum</i>	6.994	5.380	5.483	7.131	5.933	6.249	大肠杆菌 <i>Escherichia coli</i>	9.180	3.459	3.505	9.676	4.430	4.048

上述结果表明,不同物种对同义密码子的偏好不仅取决于基因的总碱基组成,编码区域不同区段的碱基环境也至关重要,即同义密码子用法存在明显的位置效应。可以推测,核苷酸序列不同区域对同义密码子的选择可能是蛋白质折叠所必需的。

3 讨 论

本研究通过对 20 个物种同义密码子使用的系统比较发现,进化上亲缘关系较近的物种在密码子的使用上比较接近,而动物、植物和微生物间由于分歧的年代久远,其对密码子的选择有很大不同,基因组间的差异可能是造成密码子使用偏性的首要因素。

Hooper 等^[25]指出,密码子用法是沿基因序列变化的,但他并没有明确指出密码子在特定位置的变化情况。本研究发现,所有供试物种在其编码区域不同区段间,密码子的用法均存在较大差异,说明同义密码子的偏性与基因编码区各区域的碱基组成密切相关,而这种特殊的基因组环境也必将直接影响到其蛋白折叠的三维空间结构^[14]。有研究^[26]表明,在翻译起始事件中,真核生物 AUG 的 5 端邻近序

列(-20 至+6)常作为核糖体亚基识别翻译起始点的重要信号。在大肠杆菌^[5]、酵母^[27]等物种中,紧靠起始密码子 3 端的密码子,即+2 位密码子与蛋白翻译起始有关。故有望通过基因工程修饰编码区 5 端的 DNA 序列,以提高特定蛋白的表达水平。

通常认为基因的碱基组成影响着密码子的使用,但在莱因哈德衣藻(*C. reinhardtii*)^[28]和棘皮蚧(*Echinococcus* spp.)^[29]基因组中,密码子的使用并未受到基因组 G+C 含量的影响。本研究中,G+C 含量是影响单子叶植物水稻、普通小麦、大麦、玉米以及线虫等物种同义密码子偏性的主要因子。部分双子叶植物(拟南芥和菜豆)、人、猪、斑马鱼、家鼠、大鼠、果蝇及酿酒酵母、裂殖酵母等物种中,两者的相关系数较小但呈极显著正相关;豌豆、番茄、烟草和大肠杆菌中,G+C 含量对其同义密码子偏性起负向促进作用,不过决定系数较小。表明在这些物种中,同义密码子偏性是 G+C 含量与其他诸多因素综合作用的结果。欧洲油菜中两者的相关性不显著,表明该物种基因的同义密码子偏性主要受到其他因素的影响。

[参考文献]

- [1] 赵翔,李至,陆身枫,等. 酵母 *Yarrowia lipolytica* 的密码子用法分析[J]. 复旦大学学报(自然科学版), 1999, (10): 510- 516
- [2] Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *A. rabidopsis*[J]. Proc Natl Acad Sci U S A, 1999, 96(8): 4482- 4487.
- [3] Carlini D B, Chen Y, Stephan W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *A dh* and *A dhr*[J]. Genetics, 2001, 159(2): 623- 633
- [4] Sajau H, Washio T, Saito R, et al. Correlation between sequence conservation of the 5' untranslated region and codon usage bias in *Mus musculus* genes[J]. Gene, 2001, 276(1- 2): 101- 105
- [5] Stenstrom CM, Jin HN, Major LL, et al. Codon bias at the 3' side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*[J]. Gene, 2001, 263(1- 2): 273- 284
- [6] Fedorov A, Saxonov S, Gilbert W. Regularities of context-dependent codon bias in eukaryotic genes[J]. Nucleic Acids Res, 2002, 30(5): 1192- 1197.
- [7] Ohno S. Universal rule for coding sequence construction: TA/GC deficiency-TG/CT excess[J]. Proc Natl Acad Sci U S A, 1988, 85(24): 9630- 9634
- [8] Karlin S, Campbell AM, Mazek J. Comparative DNA analysis across diverse genomes[J]. Annu Rev Genetics, 1998, 32: 185- 225
- [9] 石秀凡, 黄京飞, 柳树群, 等. 人类基因同义密码子偏好的特征以及与基因 GC 含量的关系[J]. 生物化学与生物物理进展, 2002, 29(3): 411- 414
- [10] Rodriguez-Trelles F, Tarrío R, Ayala F J. Switch in codon bias and increased rates of Amino Acid substitution in the *Drosophila saltans* species group[J]. Genetics, 1999, 153(1): 339- 350
- [11] Moriyama EN, Powell J R. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli* [J]. Nucleic Acids Res, 1998, 26(13): 3188- 3193
- [12] Moriyama EN, Powell J R. Codon usage bias and tRNA abundance in *Drosophila*[J]. J Mol Evol, 1997, 45(5): 514- 523
- [13] Moszer I, Rocha E P C, Danchin A. Codon usage and lateral gene transfer in *Bacillus subtilis*[J]. Current Opinion in Microbiology, 1999, 2(5): 524- 528
- [14] 顾万君, 马建民, 周董, 等. 不同结构的蛋白编码基因的密码子偏性研究[J]. 生物物理学报, 2002, 18(1): 81- 86

- [15] 石秀凡, 黄京飞, 梁宠荣, 等. 人类基因中同义密码子的偏好与密码子-反密码子间的结合强度密切相关吗? [J]. 科学通报, 2000, 45(23): 2520- 2525
- [16] Morton B R. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages[J]. J Mol Evol, 1998, 46(4): 449- 459
- [17] Bernardi G. The human genome: organization and evolution history[J]. Annu Rev Genet, 1995, 29: 445- 476
- [18] Thanaraj T A, Argos P. Ribosome mediated translational pause and protein domain organization[J]. Protein Sci, 1996, 5(8): 1594- 612
- [19] Yu J, Hu S N, Wang J, et al A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*) [J]. Science, 2002, 296(5): 79- 92
- [20] Stenico M, Lloyd A T, Sharp P M. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases[J]. Nucleic Acids Res, 1994, 22(13): 2437- 2446
- [21] Kennedy S P, Ng W V, Salzberg S L, et al Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence[J]. Genome Res, 2001, 11(10): 1641- 1650
- [22] Helene C, Frederique L, Michel C, et al Codon usage and gene function are related in sequences of *Arabidopsis thaliana* [J]. Gene, 1998, 209: 1- 38
- [23] Richard J, Epstein R J, Lin K, et al A functional significance for codon third bases[J]. Gene, 2000, 245(2): 291- 298
- [24] Warrick H M, Spudich J A. Codon preference in *Dicystostelium discoideum* [J]. Nucleic Acids Res, 1988, 16(14): 6617- 6635
- [25] Hooper S D, Berg O G. Gradients in nucleotide and codon usage along *Escherichia coli* genes[J]. Nucleic Acids Res, 2000, 28(18): 3517- 3523
- [26] Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation[J]. J Biol Chem, 1991, 266(30): 19867- 19870
- [27] 吕军, 李宏, 马克健. 酵母编码区及非编码区的统计分析[J]. 内蒙古大学学报(自然科学版), 2001, 32(3): 147- 151
- [28] Naya H, Romero H, Carels N, et al Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii* [J]. FEBS Letters, 2001, 501(2- 3): 127- 130
- [29] Fernandez V, Zavala A, Musto H. Evidence for translational selection in codon usage in *Echinococcus* spp. [J]. Parasitology, 2001, 123(2): 203- 209

Comparative studies on synonymous codon usage bias in twenty species

L IU Qing-po¹, FENG Ying¹, DONG Hu²

(1 Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang 310029, China;

2 Shengtongreli Corporation Limited, Dushanzi District, Kelamayi City, Dushanzi, Xinjiang 833600, China)

Abstract: A comparison of synonymous codon usage bias was made in twenty species from micro to higher plants and animals which involves six Dicotyledoneae, four Monocotyledones, seven multicellular animals, two unicellular eukaryotes and one prokaryote (*Escherichia coli*). Results showed that the codon usage was significantly position-dependent. Every species selectively used synonymous codons at different sections in coding region, suggesting that codon usage bias was closely related to CDS base composition environment. In a word, codon usage bias was primarily determined by difference among investigated species. Most species in this study showed significantly stronger variation in codon usage in translational initiation region than that in translational termination region, which suggested that codon usage in translational initiation region was more important for increasing protein-translation efficiency or accuracy. The contribution of G+C content to codon usage bias was extremely changeable in different species, probably according to its base composition, which suggested that G+C content in coding region was only one of important factors in shaping codon usage bias. Studies on codon usage bias in the coding region of genes could give some clues for searching the molecular mechanism how the species evolves.

Key words: codon usage bias; position-dependence; G+C content