毕赤酵母的密码子用法分析

赵 翔 霍克克 李育阳

(复旦大学遗传学研究所遗传工程国家重点实验室 上海 200433)

摘 要 通过分析 *Pichia pastoris* 的 28 个蛋白编码基因的同义密码子使用情况并计算该酵母的密码子用法,首次确定出 *P. pastoris* 的 19 个高表达优越密码子。这些结果经与已知的 *Saccharomyces cerevisiae* 及 *Kluyveromyces lactis* 的密码子用法基本相似,但在氨基酸谷氨酸的密码子选择上截然相反,提示这可能属于 *P. pastoris* 所偏爱的密码子用法。

关键词 毕赤酵母,密码子用法,优越密码子

中图分类号 Q755,Q939.5 **文献标识码** A **文章编号** 1000-3061(2000)03-0308-04

酵母表达系统是基因工程研究中最常用的表达系统之一。近年来,除酿酒酵母外,还发展了 *Pichia pastoris*, *Kluyveromyces lactis*, *Yarrowia lipolytica* 等非常规酵母表达系统,其中尤以 *Pichia pastoris* 用得最广[1]。

酵母菌对外源基因的表达也和外源基因密码子的选用有关。了解表达系统宿主在密码子使用上的偏爱性对从翻译水平分析外源基因表达的规律有重要意义,也为改造外源基因或改造宿主细胞提供依据[2]。至今尚未见有人对 *Pichia pastoris* 的密码子用法作过分析。本文报道的就是我们对此所作的分析结果。

1 材料和方法

Pichia pastoris 基因的资料:来自 Genbank;密码子有效数(Effective number of codons) *Nc* 的计算,相对同义密码子用法(Relative synonymous codon usage) *RSCU* 的计算和高表达优越密码子的确定方法参见文献[3];*Saccharomyces cerevisiae*, *Kluyver-omyces lactis* 密码子的用法:分别见文献[4,5]。

2 结果和讨论

2.1 *P. pastoris* 基因的选择

由 Genbank DNA 顺序信息库检索得到 *P. pastoris* 所有的全长基因序列共 44 个,经筛选得蛋白表达基因序列 28 个作为样本总体(见表 1),筛选按文献[5]所述原则进行。

2.2 目的基因密码子频率的统计

使用 GCG 程序' codonfrequency '命令在 SGI 工作站上对每个目的基因进行密码子频率分析而得。

2.3 样本总体 *RSCU* 的计算和比较

统计酵母 *P. pastoris* 样本总体中各个密码子的观察数,按照方法 2 计算每个密码子的 *RSCU* 值(见表 2),这个值代表了该种酵母在密码子使用的整体情况。表 2 同时给出 *S. cerevisiae* 和 *K. lactis* 的密码子 *RSCU* 值。通过比较可以发现,*P. pastoris* 的密码子情况和 *S. cerevisiae* 及 *K. lactis* 比较一致。同时我们还计算了解脂耶氏酵母(*Y. lipolytica*)目前已知的 48 个蛋白编码基因的密码子用法,发现它与 *P. pastoris* 有着较大的差异。进一步比较发现,对 Leu,Pro,Glu,Asp,Gln,Ser,Ala 等氨基酸的密码子,*P. pastoris* 中以 UUA,CCA,GAA,GAU,CAA,UCA,GCA 为多数,而在 *Y. lipolytica* 中分别以 CUG,CCC,GAG,GAC,CAG,UCC,GCC 为多数。这一现象与两者的(G+C)%值是一致的,已知在 *Y. lipolytica* 中(G+C)%大于 50%,而在 *P. pastoris* 中(G+C)%只有 40%~42%,相应的在密码子选择上表现出差异时,在 *P. pastoris* 中占优势的密码子较之在 *Y. lipolytica* 中占优势的含有更多的嘌呤。对 *S. cerevisiae*, *K. lactis* 的检索也证实了这种看法,这两个酵母的(G+C)%分别为 40% 和 41%。所以从整个表 2 来看 *RSCU* 的物种间变化,大致反映出物种彼此间(G+C)%的差异。

收稿日期:1999-05-17,修回日期:2000-02-21。
基金项目:国家高技术研究发展计划项目资助(863-102-11-02-02)。

表 1　*P. pastoris* 蛋白编码基因

Table 1　The coding sequences of *P. pastoris*'s

Gene	Description	ACC#	L	Nc	Ref.
PPU62648	Glyceraldehyde-3-phosphate dehydrogenase (GAP) gene	U62648	334	28.4	*Gene*, **186**(1): 37
PPU73376	6-phosphofructokinase alpha subunit (PFK1) gene	U73376	991	33.1	*J Cell Sci*, **110**: 1935
PPU96968	Alcohol oxidase (AOX2) gene	U96968	664	34.2	*Yeast*, **5**(3): 167
PPINO1	Inositol 1-phosphate synthase (INO1) gene	AF078915	526	40.3	*Direct Submission*
PPU28658	Inducible acid phosphatase (PHO1) gene	U28658	469	41.7	*Gene*, **163**(1): 19
PPTRP1877#2	Inorganic pyrophosphatase gene	AJ001000	286	43.1	*Yeast*, **14**: 861
PPPYC1	Pyruvate carboxylase gene	Y11106	1190	43.3	*Yeast*, **14**(7): 647
PPFLD1	Formaldehyde dehydrogenase (FLD1) gene	AF066054	380	44.5	*Gene*, **216**: 93
PPHIS4G	HIS4 gene	X56180	845	47.8	*Direct Submission*
PPU14126	Phosphoribosyl-ATP pyrophosphohydrolase	U14126	843	48.0	*Curr Genet*, **26**: 443
PPPAS5GN	PAS5 gene	Z22556	1166	48.8	*J Cell Sci*, **123**: 535
PPDAK	Dihydroxyacetone kinase (DAK) gene	AF019198	609	48.9	*Yeast*, **14**(8): 759
PPU49510	DNA polymerase gamma gene	U49510	1013	50.2	*N A Res*, **24**: 1481
PPPRC1GEN	PRC1 gene	X87987	524	50.9	*Yeast*, **12**(1): 31
YSRPER3G	PER3 gene	L40485	714	51.0	*J. B. C.* **270**: 10940
PPPER6P	PER6P gene	X96945	462	52.0	*Mol Cell Biol*, **16**: 2527
PPPAS1G	PAS1 gene	Z36987	1158	52.6	*J. C. B.* **127**: 1259
PPPAS8A	PAS8A gene	Z19592	577	53.3	*J. C. B.* **121**: 761
PPU59222	PTS1 receptor gene	U59222	577	53.5	*J. C. B.* **135**: 85
PPKEX1	Carboxypeptidase kex1 precursor (KEX1) gene	AF095574	624	53.6	*Direct Submission*
PPRPEX7	Receptor for PTS2-containing proteins	AF021797	377	54.2	*J. C. B.* **140**(4): 807
PPU12511	Ubiquitin-conjugating enzyme Pas4 (PpPAS4) gene	U12511	205	54.5	*J. B. C.* **269**: 21835
PPU70067	PpPex13p(PpPEX13) gene	U70067	381	54.7	*Direct Submission*
PPU70066	PpPex10p(PpPEX10) gene	U70066	420	55.5	*Mol Cell Biol*, **15**: 6406
PPU58140	Pas10p(PAS10) gene for a zinc-binding protein	U58140	410	56.1	*EMBO*, **15**: 3275
PPPAS2P	PAS2P gene	Z72390	456	57.2	*Direct Submission*
PPU69170#1	Imidazole glycerolphosphate dehydratase (HIS3) gene	U69170	225	57.5	*Direct Submission*
PPIRP1877#1	N-(5′-phosphoribosyl)anthranilate isomerase gene	AJ001000	238	60.0	*Yeast*, **14**: 861

L: Number of codons in the genes　ACC#: Access number

表 2　*P. pastoris* 蛋白编码顺序的密码子用法

Table 2　The codon usage of *P. pastoris*'s coding sequences

		N	RSCU	S. c.	K. l.			N	RSCU	S. c.	K. l.			N	RSCU	S. c.	K. l.			N	RSCU	S. c.	K. l.
Phe	TTT	333	1.04	1.08	0.69	Ser	TCT	368	1.87	1.83	2.24	Tyr	TAT	245	0.89	1.02	0.83	Cys	TGT	150	1.3	1.34	1.59
	TTC	308	0.96	0.92	1.31		TTC	265	1.34	1.09	1.15		TAC	322	1.11	0.98	1.17		TGC	84	0.7	0.66	0.41
Leu	TTA	227	0.94	1.6	1.16		TCA	216	1.06	1.16	1.02	ter	TAA	12	1.36	1.56	2.18	ter	TGA	4	0.55	0.84	0.55
	TTG	517	2.1	2.08	2.9		TCG	88	0.45	0.51	0.49	ter	TAG	12	1.09	0.6	0.27	Trp	TGG	167	/	/	/
Leu	CTT	248	0.95	0.66	0.59	Pro	CCT	255	1.42	1.18	1.07	His	CAT	153	0.98	1.2	1.12	Arg	CGT	115	1.05	0.99	0.97
	CTC	112	0.43	0.28	0.27		CCC	102	0.55	0.54	0.24		CAC	158	1.02	0.8	0.88		CGC	33	0.29	0.29	0.14
	CTA	168	0.64	0.79	0.82		CCA	316	1.71	1.88	2.4	Gln	CAA	404	1.27	1.46	1.43		CGA	48	0.46	0.3	0.2
	CTG	252	0.94	0.59	0.27		CCG	58	0.32	0.39	0.29		CAG	248	0.73	0.54	0.57		CGG	28	0.29	0.17	0.19
Ile	ATT	459	1.51	1.47	1.29	Thr	ACT	361	1.67	1.5	1.7	Asn	AAT	386	0.94	1.11	0.87	Ser	AGT	168	0.85	0.84	0.68
	ATC	300	0.94	0.89	1.36		ACC	240	1.07	0.97	1.21		AAC	414	1.06	0.89	1.13		AGC	94	0.43	0.57	0.42
	ATA	168	0.56	0.63	0.35		ACA	215	0.89	1.06	0.76	Lys	AAA	426	0.91	1.05	0.71	Arg	ACA	326	2.96	3.2	3.77
Met	ATG	344	/	/	/		ACG	80	0.37	0.47	0.33		AAG	548	1.09	0.95	1.29		AGG	110	0.95	1.05	0.72
Val	GTT	460	1.81	1.73	1.81	Ala	GCT	519	1.93	1.73	2.05	Asp	GAT	522	1.1	1.25	1.27	Gly	GGT	499	1.88	2.35	2.95
	GTC	244	0.95	0.96	1.18		GCC	284	0.98	0.97	0.87		GAC	446	0.9	0.75	0.73		GGC	136	0.52	0.65	0.3
	GTA	144	0.54	0.66	0.41		GCA	269	0.93	0.95	0.82	Glu	GAA	596	1.17	1.46	1.58		GGA	334	1.3	0.64	0.52
	GTG	182	0.7	0.65	0.61		GCG	46	0.17	0.35	0.26		GAG	431	0.83	0.54	0.42		GGG	81	0.3	0.37	0.23

K. l and S. c are the codon usage of *Kluyveromyces lactis* and *Saccharomyces cerevisiae* respectively

2.4 高、低表达基因样本组的抽取和高表达优越密码子的确定

虽然通过分析样本总体的密码子分布已经可以帮助我们对 *P. pastoris* 的密码子偏爱有所了解，但在大多数情况下，我们更需要知道有关高表达基因的密码子偏爱情况，这就需要进行高、低表达基因组的抽样和高表达优越密码子的确定工作。

在获得了 *P. pastoris* 样本总体后，按照方法 1 计算每个基因的 *Nc* 值并进行排列（见表 1）。从这一排列的最两端抽取约 10%（3 个基因）的样本分别组成高、低表达样本组。在高表达样本组（低 *Nc* 值）中包括 *PPU*62648，*PPU*96968，*PPU*73376 三个基因，它们都属于能量代谢相关基因，这些基因的功能直接关系到细胞存活，所以一经转录（是否转录与其启动子的调控有关）就要求能迅速和大量地合成，以应付生存需要。这样在密码子的选用上受到选择

表 3 *P. pastoris* 中高/低表达样本组的密码子用法

Table 3 The codon usage of high/low expression sample group in *P. pastoris*

		High		Low				High		Low				High		Low				High		Low	
		N	RSCU	N	RSCU			N	RSCU	N	RSCU			N	RSCU	N	RSCU			N	RSCU	N	RSCU
Phe	TTT	10	0.30	23	1.21	Ser	TCT	50	2.65	16	1.45	Tyr	TAT	2	0.06	8	1.00	Cys	TGT	18	1.50	10	1.67
	TTC	57	1.70	15	0.79		TCC	52	2.76	11	1.00		TAC	70	1.94	8	1.00		TGC	6	0.50	2	0.33
Leu	TTA	5	0.26	20	1.18		TCA	2	0.11	11	1.00	ter	TAA	2	2.00	1	1.00	ter	TGA	0	0.00	1	1.00
	TTG	74	3.79	25	1.47		TCG	2	0.11	6	0.55	ter	TAG	1	1.00	1	1.00	Trp	TGG	21	/	8	/
Leu	CTT	18	0.92	15	0.88	Pro	CCT	30	1.21	10	1.43	His	CAT	8	0.31	15	1.30	Arg	CGT	13	1.04	7	1.02
	CTC	4	0.21	12	0.71		CCC	2	0.08	6	0.86		CAC	43	1.69	8	0.70		CGC	0	0.00	3	0.44
	CTA	2	0.10	11	0.65		CCA	67	2.71	10	1.43	Gln	CAA	26	1.53	27	1.29		CGA	0	0.00	7	1.02
	CTG	14	0.72	19	1.12		CCG	0	0.00	2	0.29		CAG	8	0.47	15	0.71		CGG	0	0.00	5	0.73
Ile	ATT	43	1.45	26	1.07	Thr	ACT	50	1.90	12	1.07	Asn	AAT	9	0.23	28	0.89	Ser	AGT	5	0.27	16	1.45
	ATC	46	1.55	24	0.99		ACC	47	1.79	13	1.16		AAC	68	1.77	35	1.11		AGC	2	0.11	6	0.55
	ATA	0	0.00	23	0.95		ACA	4	0.15	13	1.16	Lys	AAA	16	0.31	33	1.20	Arg	AGA	61	4.88	13	1.90
Met	ATG	37	/	15	/		ACG	4	0.15	7	0.62		AAG	86	1.69	22	0.80		AGG	1	0.08	6	0.88
Val	GTT	58	2.13	19	1.46	Ala	GCT	72	248	22	1.54	Asp	GAT	24	0.46	35	1.19	Gly	GGT	114	3.08	20	1.36
	GTC	45	1.65	9	0.69		GCC	36	1.24	13	0.91		GAC	81	1.54	24	0.81		GGC	7	0.19	9	0.61
	GTA	3	0.11	13	1.00		GCA	8	0.28	18	1.26	Glu	GAA	41	0.84	33	1.06		GGA	27	0.73	25	1.69
	GTG	3	0.11	11	0.85		GCG	0	0.00	4	0.28		GAG	57	1.16	29	0.94		GGG	0	0.00	5	0.34

N：Number of the codon

表 4 *P. pastoris*，*S. cerevisiae* 及 *K. lactis* 中的高表达优越密码子列表

Table 4 The optimal condon list in *P. pastoris*，*S. cerevisiae* and *K. lactis*

		P.p.	S.c.	K.l.			P.p.	S.c.	K.l.			P.p.	S.c.	K.l.			P.p.	S.c.	K.l.
Phe	TTT				Ser	TCT	*	*	*	Tyr	TAT				Cys	TGT		*	
Phe	TTC	*	*	*	Ser	TCC	*	*	*	Tyr	TAC	*	*	*	Cys	TGC			
Leu	TTA				Ser	TCA				ter	TAA				ter	TGA			
Leu	TTG	*	*	*	Ser	TCG				ter	TAG				Trp	TGG			
Leu	CTT				Pro	CCT				His	CAT				Arg	CGT			
Leu	CTC				Pro	CCC				His	CAC	*	*	*	Arg	CGC			
Leu	CTA				Pro	CCA	*	*	*	Gln	CAA		*	*	Arg	CGA			
Leu	CTG				Pro	CCG				Gln	CAG				Arg	CGG			
Ile	ATT	*	*		Thr	ACT	*	*		Asn	AAT				Ser	AGT			
Ile	ATC	*	*	*	Thr	ACC	*	*	*	Asn	AAC		*	*	Ser	AGC			
Ile	ATA				Thr	ACA				Lys	AAA				Arg	AGA	*	*	*
Met	ATG				Thr	ACG				Lys	AAG	*	*	*	Arg	AGG			
Val	GTT	*	*		Ala	GCT	*	*	*	Asp	GAT				Gly	GGT	*	*	*
Val	GTC	*	*	*	Ala	GCC				Asp	GAC	*	*	*	Gly	GGC			
Val	GTA				Ala	GCA				Glu	GAA		*	*	Gly	GGA			
Val	GTA				Ala	GCG				Glu	GAG	*			Gly	GGG			

* Optimal codons

压力要避免引入稀有密码子以防在翻译过程中产生瓶颈效应。在低表达样本组(高 *Nc* 值)中包括 *PPPAS2P*，*PPU69170*＃1，*PPTRP1877*＃1 三个基因，这些基因属于膜结合蛋白、氨基酸代谢等范畴。已知这些范畴涉及的蛋白表达量较低，不会有因为大量翻译而出现瓶颈效应的现象，所以在密码子选择上，缺乏选择压力而保留了较多种类的稀有密码子。对高、低表达样本组样各自计算 RSCU 值(见表 3)，然后对两组数据进行 t-检验，取 P 值≤0.05，共有 19 个密码子被确认为 *P. pastoris* 的高表达优越密码子。表 4 给出了这 19 个密码子和已知的 *S. cerevisiae* 及 *K. lactis* 的高表达优越密码子，通过比较可以发现 *P. pastoris* 在高表达优越密码子的选择上同样是接近于 *S. cerevisiae* 及 *K. lactis*。基本上 *P. pastoris* 可以使用目前已知的酿酒酵母高表达密码子表进行基因改造和统计，但个别氨基酸仍有例外，表现在对氨基酸谷氨酸(Glu)同义密码子 GAG 和 GAA 的选择上：*P. pastoris* 以 GAG 为高表达优越密码子，而 *S. cerevisiae* 和 *K. lactis* 以 GAA 为高表达优越密码子。表 5 进一步给出了这两个同义密码子在不同酵母高表达基因中的使用情况，从中可以很清楚地看出两个密码子在不同酵母高表达基因中的分布很不相同。提示我们这可能是 *P. pastoris* 特有的密码子偏爱倾向。

表 5　*P. pastoris*、*S. cerevisiae* 及 *K. lactis* 中密码子 GAG 和 GAA 的 *RSCU* 值

Table 5　The *RSCU* value of GAA and GAG in three types of yeasts

Codon	*P. pastoris*		*S. cerevisiae*		*K. lactis*	
	T	H	T	H	T	H
GAG	0.83	1.16	0.54	0.04	0.42	0.00
GAA	1.17	0.84	1.46	1.96	1.58	2.00

T：All genes used for analysis；H：Highly expressed genes

3　结束语

在本工作中，我们对 *P. pastoris* 中目前已知的基因进行了筛选和分析。在有关的分析工作中产生了 *Nc*，*RSCU* 等指数，并初步确定了该酵母的密码子用法和高表达优越密码子。通过将有关分析结果和另几种 *S. cerevisiae*、*Y. lipolytica* 和 *K. lactis* 已知的密码子使用情况比较，发现 *P. pastoris* 的密码子用法不同于 *Y. lipolytica* 但接近于 *S. cerevisiae*。同时在个别氨基酸密码子的偏爱上，*P. pastoris* 和 *S. cerevisiae* 也不尽相同。我们的统计结果不仅可以成为在利用 *P. pastoris* 进行外源基因表达方面的参考指标之一，同时也为研究 *P. pastoris* 在系统进化中的位置提供了依据。需要指出的是，外源基因在酵母中的表达受到多种因素的共同作用，而来自密码子用法的影响只是其中之一。

参　考　文　献

[1] Michael A, Carol A S, Jeffery J C. *Yeast*, 1992, **8**: 423～488

[2] Hockema A, Kastellin R A, Vasser M *et al*. *Mol Cell Biol*, 1987, **7**: 2914～2924

[3] 赵　翔，李　至，陆身枫等. 复旦学报(自然科学版)，1999，**38**(5)：510～516

[4] Paul M S, Elizabeth C. *Yeast*, 1991, **7**: 657～678

[5] Andrew T L, Paul M S. *Yeast*, 1993, **9**: 1219～1228

Synonymous Codon Usage in *Pichia pastoris*

ZHAO Xiang　HUO Ke-Ke　LI Yu-Yang

(*State Key Laboratory of Genetic Engineering*, *Institute of Genetics*, *Fudan University*, *Shanghai*　200433)

Abstract　According to the synonymous codons used in 28 open reading frames from *Pichia pastoris*, the codon usage in this species was calculated and 19 codons have been inferred to be its optimal codons. The results show that pattern of the codon usage in *P. pastoris* is similar to that in *S. cerevisiae* and in *K. lactis* except for the synonymous codon of glutamic acid, which may be the special bias of *P. pastoris*.

Key words　*Pichia pastoris*, codon usage, optimal codons