P53基因三周期性与密码子偏好性的相关性*

王其强 谈承杰 朱平†

(江南大学理学院, 无锡 214122)

(2013 年 9 月 15 日收到; 2013 年 10 月 25 日收到修改稿)

为了进一步研究分析P53抑癌基因的性质, 对P53基因的蛋白质编码区以及对应的mRNA 的三周期性进行比较分析. 通过同义密码子相对使用度和拟同义密码子相对使用度方法对其分别进行计算, 分析了蛋白质编码区密码子的偏好性以及mRNA 密码子的偏好性. 结果表明: P53蛋白质编码区具有很强的三周期性, 而对应的mRNA 不具有三周期性; P53蛋白质编码区的密码子偏好G 或C结尾的密码子程度强于对应的mRNA. 说明P53 基因密码子的偏好性与三周期性紧密相关, 密码子的偏好程度影响着三周期性. 进一步从生物角度诠释了P53 蛋白质编码区和对应mRNA 三周期性的异同. 结合P53基因这一特性, 有助于提高其基因识别的正确率, 对P53 基因的深入研究具有重要的意义.

关键词: 三周期性, 偏好性, P53基因, 同义密码子相对使用度

PACS: 87.10.–e, 87.14.gk, 87.15.–v **DOI:** 10.7498/aps.63.048701

1 引 言

P53基因是一种抑癌基因, 是迄今为止发现与人类肿瘤相关性最高的基因之一, 一半以上的人类肿瘤由P53基因发生突变产生. 因此, 其已成为当前生物信息研究中最受关注的基因. 在DNA损伤时, P53蛋白就会增加, 阻止DNA复制, 为DNA修复争取时间, 若修复失败, 则诱导细胞凋亡. 如果P53 基因发生突变, 则细胞增生失去控制, 发生癌变[1]. 因此, P53基因性质的研究中, 蛋白质编码区及mRNA密码子偏好性显得尤为重要. 基于P53基因的重要性, 国内外很多学者对其开展了研究. Xia和Jia[2] 构建了由DNA 损伤信号引发的P53振动网络的数学模型, 研究P53调控细胞中DNA损失的机制. Yan和Zhu[3] 构建了代数结构, 建立了基因的扩展密码子集C_{343}, 对P53基因的突变进行研究. 张丽娟等[4] 研究了在基因治癌上有特殊意义的P53-Mdm2负反馈回路.

对于蛋白质编码区的三周期性以及外显子的三周期性已有很多研究[5]. 蛋白质编码区具有三周期性, 是生物长期进化形成的[6]. 张静和石秀凡[7] 研究中发现三周期性是个全局性质, 序列中局部具有三周期性, 那么该序列就具有三周期性. 文献[8] 对P53外显子三周期性的研究中, 发现其外显子具有三周期性. 对于P53 基因密码子偏好性的研究, 也已有很多成果. 石秀凡等[9] 利用基于氨基酸编码下的同义密码子相对使用度 (relative synonymous codon usage, RSCU) 来分析密码子的偏好性. 朱平等[10] 定义基于拟氨基酸下的同义密码子相对使用度 (quasi relative synonymous codon usage, QRSCU) 分析了P53基因密码子的拟偏好性. 这为本文研究奠定了基础.

在我们之前的研究中, 主要对P53家族外显子的三周期性进行了讨论, 根据外显子的长度, 分析了外显子三周期性的强弱[8], 有关蛋白质编码区的三周期性以及mRNA的三周期性都没有进行研究, 对三周期性与密码子的偏好性也没有进行深入探讨. 蛋白质的编码区和对应的mRNA都是由外显子拼接而成的, 因此可依据局部性来判断全局性. P53基因的蛋白质编码区和对应的mRNA

* 国家自然科学基金 (批准号: 11271163) 和中央高校基本科研业务费 (批准号: JUSRP51317B) 资助的课题.

† 通讯作者. E-mail: zhuping@jiangnan.edu.cn

是否都具有三周期性, 蛋白质编码区部分的密码子偏好性和mRNA的密码子偏好性又如何, 这些问题都值得深入探讨. 为了研究这些问题, 本文在Z-curve映射[11]下, 利用信号处理和分析的方法, 通过信噪比和频率两个指标相结合的方法[8], 讨论了P53蛋白质编码区和对应mRNA的三周期性; 利用Codon W软件计算P53蛋白质编码区和对应mRNA的RSCU值和QRSCU值; 依据基本的划分方法[12], 讨论分析了其密码子偏好性的异同; 分析了三周期性与密码子偏好性的关系, 这有助于提高P53 家族基因识别的正确率; P53密码子偏好性的讨论, 一方面有利于P53基因的后续研究, 另一方面进一步从生物学角度解释了P53蛋白质编码区与对应mRNA三周期性的异同.

2 材料与方法

2.1 材料来源

从GenBank (www.NCBI.com)数据库中选取14条人类(Human beings) P53基因mRNA以及相应的蛋白质编码区(CDS)进行研究. 这些mRNA所对应的序列号为:

NM_001031685.2, NM_000546.4,

NM_001126114.2, NM_147184.3,

NM_022112.2, NM_004881.4,

NM_001195194.1, NM_001258324,

NM_001258320.1, NM_001126116.1,

NM_001126112.1, NM_001126113.1,

NM_001126115.1, NM_001126117.1.

为了后续说明方便, 用1—14数字对其进行编号.

2.2 方 法

2.2.1 数值映射

随着生物信息学的发展, 对于序列中碱基进行数值化的映射有很多, 包括Voss映射、实数映射、Z-curve 映射等[11], 及Yan和Zhu[3]建立的映射$\varphi: GF(7^3) \to C_{343}$. 而通常情况下认为Z-curve映射比Voss映射更具有生物学意义[11], 为此本文采用Z-curve映射.

Z-curve映射是在Voss映射的基础上定义的. 依据Voss映射有DNA序列S的4个指示序列$\{u_b[n]\}$, $b \in I = \{A, C, G, T\}$, $n = 0, 1, 2, \cdots, N-1$, 作其累积序列: $b_n(n = 0, 1, \cdots, N-1)$ 为 $b_n = \sum_{i=0}^{n-1} u_i[i]$. 那么, 定义如下三个序列$x[n], y[n], z[n]$为

$$\begin{cases} x[n] = 2(A_n + G_n) - n, \\ y[n] = 2(A_n + C_n) - n, \\ z[n] = 2(A_n + T_n) - n, \end{cases}$$

且令$x[-1] = 0$, $y[-1] = 0$和$z[-1] = 0$, 定义$\Delta x_n[n] = x[n] - x[n-1]$, $\Delta y_n[n] = y[n] - y[n-1]$和$\Delta z_n[n] = z[n] - z[n-1]$. 那么Z-curve映射定义如下:

$$\begin{pmatrix} \Delta x[n] \\ \Delta y[n] \\ \Delta z[n] \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix}.$$

例如对于DNA序列$S(n) =$ ACGTTAG, 则对应的Z-curve映射为:

$$\begin{cases} \{\Delta x[n]\} = \{1, -1, 1, -1, -1, 1, 1\}, \\ \{\Delta y[n]\} = \{1, 1, -1, -1, -1, 1, -1\}, \\ \{\Delta z[n]\} = \{1, -1, -1, 1, 1, 1, -1\}. \end{cases}$$

2.2.2 功率谱、信噪比及三周期性判断方法

利用离散傅里叶变换(DFT)可以将上述三个序列离散化, 对于序列$\Delta x_n[n]$可得到如下复数序列[13]:

$$\Delta X[k] = \sum_{n=0}^{N-1} \Delta x[n]\, \mathrm{e}^{-\mathrm{i}\frac{2\pi nk}{N}}, \quad (k = 0, 1, \cdots, N-1).$$

对于$\Delta y_n[n]$和$\Delta z_n[n]$同样做DFT可以得到复数序列$\Delta Y_n[n]$和$\Delta Z_n[n]$.

根据上面三个复数序列, 定义序列的功率谱

$$P[k] = |\Delta X[k]|^2 + |\Delta Y[k]|^2 + |\Delta Z[k]|^2;$$

信噪比定义如下

$$R = \frac{P\left[\frac{N}{3}\right]}{\bar{E}} = \frac{\left|\Delta X\left[\frac{N}{3}\right]\right|^2 + \left|\Delta Y\left[\frac{N}{3}\right]\right|^2 + \left|\Delta Z\left[\frac{N}{3}\right]\right|^{2\prime}}{\bar{E}},$$

$\bar{E}$ 指该序列功率谱的平均功率谱

$$\bar{E}=\frac{\sum_{k=0}^{N-1}P[k]}{N}.$$

为了判断基因序列是否具有三周期性, 一方面, 重新定义信噪比如下:

$$\bar{R}=\frac{P_{\max}[k]}{\bar{E}},$$

其中 $P_{\max}[k]$ 指的是序列的最大功率谱. 另一方面, 为了方便计算, 我们将频率转化为出现最大功率谱处的碱基位置与 $\frac{N}{3},\frac{2N}{3}$ 的差值来计算, 进行如下构造:

$$d_1=\left|m_1-\frac{N}{3}\right|,d_2=\left|m_2-\frac{2N}{3}\right|,\bar{d}=\frac{d_1+d_2}{2},$$

其中 m_1 和 m_2 为出现最大峰值的位置, 若序列长度不是 3 的倍数, 则四舍五入后进行计算. 在判断一个序列是否具有三周期性时, 同时考虑 $\bar{R}$ 和 $\bar{d}$ 两个指标, 当 $\bar{R}\geqslant 4$ 且 $\bar{d}\leqslant 6$ 时, 认为该序列具有三周期性.

判断是否具有三周期性的算法:

步骤 1 利用数值映射将序列数值化;

步骤 2 根据 DFT, 将序列离散化, 并根据定义计算序列的功率谱和信噪比;

步骤 3 收缩功率谱的最大值, 并根据最大值重新计算信噪比, 并输出取得最大功率谱的序列位置;

步骤 4 计算输出位置与序列 $\frac{N}{3},\frac{2N}{3}$ 处的差值, 计算平均差值 $\bar{d}$;

步骤 5 根据 $\bar{R}\geqslant 4$ 且 $\bar{d}\geqslant 6$ 这一条件, 判断有无三周期性.

2.2.3 RSCU 与 QRSCU 方法使用

石秀凡等 [9] 利用基于氨基酸编码 RSCU, 分析了密码子的偏好性, RSCU 具体计算公式如下:

$$\mathrm{RSCU}_{ij}=\frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}},$$

其中 RSCU_{ij} 指的是编码第 i 个氨基酸第 j 个密码子的相对密码子使用值; x_{ij} 指编码第 i 个氨基酸第 j 个密码子的出现数目; n_i 指编码第 i 个氨基酸同义密码子的数量 (取值 1—6).

朱平等 [10] 提出了 QRSCU, 具体计算公式定义如下:

$$\mathrm{QRSCU}_{ij}=\frac{y_{ij}}{\frac{1}{4}\sum_{j=1}^{4}y_{ij}},$$

其中 QRSCU_{ij} 指的是编码第 i 个氨基酸第 j 个密码子的相对密码子使用值; y_{ij} 指编码第 i 个氨基酸第 j 个密码子的出现数目. 该式中拟氨基酸是一个重要的定义, 其定义如下. 令

$$\begin{aligned}
&\mathrm{WC}\equiv\{\mathrm{tgg,tgt,tga,tgc}\},\\
&\mathrm{RS}\equiv\{\mathrm{agg,agt,aga,agc}\},\\
&\mathrm{LF}\equiv\{\mathrm{ttg,ttt,tta,ttc}\},\\
&\mathrm{MI}\equiv\{\mathrm{atg,att,ata,atc}\},\\
&\mathrm{ED}\equiv\{\mathrm{gag,gat,gaa,gac}\},\\
&-\mathrm{Y}\equiv\{\mathrm{tag,tat,taa,tac}\},\\
&\mathrm{KN}\equiv\{\mathrm{aag,aat,aaa,aac}\},\\
&\mathrm{QH}\equiv\{\mathrm{cag,cat,caa,cac}\},\\
&\mathrm{S'}\equiv\{\mathrm{tcg,tct,tca,tcc}\},\\
&\mathrm{R'}\equiv\{\mathrm{cgg,cgt,cga,cgc}\},\\
&\mathrm{L'}\equiv\{\mathrm{ctg,ctt,cta,ctc}\}.
\end{aligned}$$

令 ZU = {G, V, ED, A, WC, LF, -Y, S', RS, MI, KN, T, R', L', QH, P }, 则 ZU 称为拟氨基酸集, 这里, −Y 为新终止子, 但是 (tga ∈) WC 不是终止子.

3 结果与分析

3.1 P53 CDS 与对应 mRNA 的三周期性

通过 Matlab7.0, 编程实现了 P53 CDS 及 mRNA 功率谱、信噪比 (R') 以及 $\bar{d}$ 的计算, 并生成了频谱图. 并根据算法判断有无三周期性. 其具体的数据分别如表 1 和表 2 所示.

为了进一步呈现两者周期性的强弱, 考虑长度的影响, 以长度为权重, 定义下列公式:

$$\begin{aligned}
\bar{R}'&=\sum_{i=1}^{14}\frac{N_i}{N_m}R'_i,\\
D&=\sum_{i=1}^{14}\frac{N_i}{N}\bar{d},\\
\eta&=\frac{m}{14},
\end{aligned}$$

其中 N_i 为每条 CDS 或 mRNA 的长度, N_m 为其的总长度, m 为具有三周期性的 CDS 或 mRNA 条数, 按前面给定判断, 当 $\bar{R}\geqslant 4$ 且 $\bar{d}\leqslant 6$ 时认为具有三周期性. 所得结果如表 3 所示, 其中具有较强三周期性的 CDS 频谱及对应 mRNA 的频谱分别如图 1 和图 2 所示.

通过表3及图1和图2我们可以得到如下结果:

1) 所选取的CDS中, 仅有4条不具有明显的三周期性, 其他均具有三周期性, 而且$\bar{R}$很大且$\bar{d}$很小, 三周期性十分强;

表1 P53 CDS的相关数据

序号	P	R'	m_1	$\frac{N}{3}$	m_2	$\frac{2N}{3}$	N	$\bar{d}$
1	69898.00	27.49	1136	1135	2271	2270	3405	1
2	4527.00	5.15	395	394	789	788	1182	1
3	3186.11	4.18	342	342	686	684	1026	1
4	6746.00	9.05	334	333	667	666	999	1
5	919.10	3.30	19	125	358	250	375	107
6	6746.00	9.05	334	333	667	666	999	1
7	1021.85	3.78	144	121	221	242	363	22
8	6396.00	15.14	191	190	381	380	570	1
9	6396.00	15.14	191	190	381	380	570	1
10	2124.00	4.52	211	210	421	420	630	1
11	4527.00	5.15	395	394	789	788	1182	1
12	3300.46	4.26	347	347	696	694	1041	1
13	3368.00	5.76	263	262	525	524	786	1
14	2072.84	4.31	4	215	643	430	645	212

表2 相应P53 mRNA的相关数据

序号	P	R'	m_1	$\frac{N}{3}$	m_2	$\frac{2N}{3}$	N	$\bar{d}$
1	57129.69	16.36	1558	1557	3114	3113	4670	1
2	15349.36	7.94	2	862	2586	1724	2586	861
3	17513.30	8.60	9	908	2717	1816	2724	900
4	7623.86	6.22	549	548	1096	1095	1643	1
5	4849.53	5.10	8	426	1272	852	1278	419
6	13561.99	8.94	2	681	2042	1361	2042	680
7	1622.00	3.62	96	200	506	400	600	105
8	22114.11	8.43	6	1189	3563	2378	3567	1184
9	23330.43	8.39	15	15	3748	2506	3759	621
10	22382.97	12.44	8	801	2398	1603	2404	794
11	15357.06	7.96	2	861	2583	1722	2583	860
12	15066.06	7.62	2	882	2646	1764	2646	881
13	12688.74	7.47	8	757	2265	1514	2271	750
14	19400.61	11.12	8	777	2325	1554	2331	770

表3 $\bar{R}'$, D及η的计算结果

	$\bar{R}'$	D	η/%
CDS	11.70	14.20	78.57
mRNA	9.48	660.57	14.29

2) 所对应选取的mRNA中, 仅有两条具有较明显的三周期性, 其他均无明显三周期性, 而且虽然$\bar{R}$较大, 但$\bar{d}$也很大, 从$\bar{d}$值看出出现频谱最大值的位置靠近两端, 没有三周期的典型特性.

文献[7]指出, 只要一个长序列中的一部分具有三周期性, 那么整个序列也具有三周期性. 文献[8]对P53外显子的三周期性研究中, 发现其外显子具有三周期性. 而本文研究中发现P53的CDS具有很强的三周期性, 而同样由外显子拼接成的mRNA却不具有三周期性. 这与文献[7]中研究的基因所表现出的性质不同. 因而, 从P53基因的CDS或mRNA的局部三周期性判断其全局三周期性不具普遍性意义. 三周期性是体现编码蛋白质序列的特性, 而mRNA并不是所有片段都编码蛋白质, mRNA三周期性不强体现了这一本质特征.

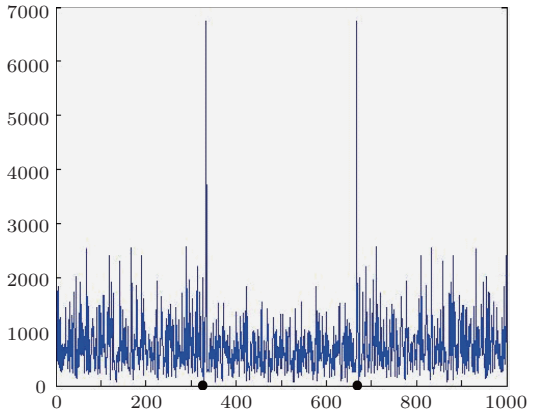

图1 第六条P53蛋白质编码区频谱 横坐标为碱基位置, 纵坐标为P值, 黑色标记处为$N/3$, $2N/3$, 最大峰值越大且出现位置越靠近黑色圆标记处, 三周期越明显; R值较大且最大峰值靠近$N/3$, $2N/3$, 其是具有三周期性的CDS

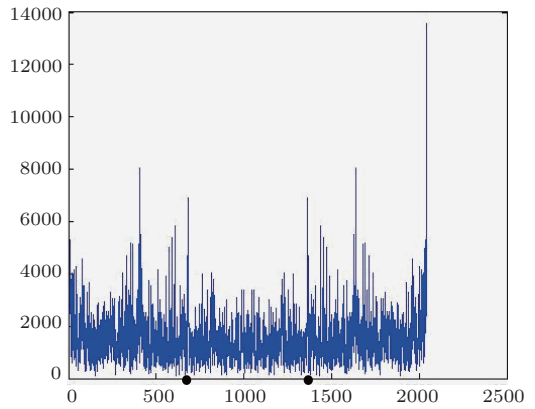

图2 第六条P53 mRNA频谱 横坐标为碱基位置, 纵坐标为P值, 黑色圆标记处为$N/3$, $2N/3$, 最大峰值越大且出现位置越靠近黑色圆标记处, 三周期越明显; R值较大, 但最大峰值远离$N/3$, $2N/3$, 其是不具有三周期性的mRNA

表 4 14 条 CDS 和 mRNA 的 RSCU 值

氨基酸	同义密码子	CDS		mRNA		氨基酸	同义密码子	CDS		mRNA	
		数量	RSCU	数量	RSCU			数量	RSCU	数量	RSCU
Phe							TAC	68	1.13	70	1.13
	TTT	55	0.89	253	1.28	Ter					
	TTC	68	1.11	143	0.72		TAA	4	0.86	113	0.79
Leu							TGA	8	1.71	242	1.70
	TTA	22	0.30	87	0.45		TAG	2	0.43	72	0.51
	TTG	63	0.86	174	0.89	His					
	CTT	65	0.88	231	1.18		CAT	48	0.88	153	0.79
	CTC	79	1.07	240	1.23		CAC	61	1.12	234	1.21
	CTA	38	0.52	93	0.48	Gln					
	CTG	175	2.38	348	1.78		CAA	62	0.53	141	0.68
Ile							CAG	173	1.47	271	1.32
	ATT	35	0.70	132	1.24	Asn					
	ATC	100	2.00	98	0.92		AAT	82	0.89	114	1.07
	ATA	15	0.30	89	0.84		AAC	102	1.11	99	0.93
Met						Lys					
	ATG	126	1.00	127	1.00		AAA	83	0.71	203	1.05
Val							AAG	151	1.29	184	0.95
	GTT	71	1.03	138	1.01	Asp					
	GTC	63	0.91	141	1.03		GAT	106	1.03	106	0.90
	GTA	13	0.19	77	0.57		GAC	99	0.97	129	1.10
	GTG	130	1.88	189	1.39	Glu					
Ser							GAA	125	0.81	182	0.92
	TCT	76	1.21	269	1.26		GAG	183	1.19	212	1.08
	TCC	94	1.49	288	1.34	Gys					
	TCA	60	0.95	242	1.13		TGT	38	0.80	223	0.89
	TCG	12	0.19	58	0.27		TGC	57	1.20	276	1.11
Pro						Trp					
	CCT	136	1.30	301	1.04		TGG	49	1.00	356	1.00
	CCC	119	1.14	405	1.40	Arg					
	CCA	108	1.04	344	1.19		CGT	36	0.79	60	0.29
	CCG	54	0.52	111	0.38		CGC	65	1.42	84	0.41
Thr							CGA	40	0.88	45	0.22
	ACT	45	0.79	176	1.06		CGG	35	0.77	78	0.38
	ACC	96	1.68	195	1.17		AGA	43	0.94	273	1.34
	ACA	61	1.07	245	1.47		AGG	55	1.20	277	1.36
	ACG	27	0.47	61	0.37	Ser					
Ala							AGT	52	0.83	174	0.75
	GCT	102	1.35	243	1.14		AGC	84	1.33	289	1.25
	GCC	119	1.57	293	1.38	Gly					
	GCA	59	0.78	209	0.98		GGT	57	0.71	149	0.68
	GCG	23	0.30	106	0.50		GGC	92	1.15	211	0.96
Tyr							GGA	90	1.13	218	0.99
	TAT	52	0.87	54	0.87		GGG	80	1.00	301	1.37

表 5 14 条 CDS 和 mRNA 的 QRSCU 值

拟氨基酸	同义密码子	CDS		mRNA		拟氨基酸	同义密码子	CDS		mRNA	
		数量	QRSCU	数量	QRSCU			数量	QRSCU	数量	QRSCU
G						ED					
	GGG	80	1.00	301	1.37		GAG	183	1.43	212	1.35
	GGT	57	0.71	149	0.68		GAT	106	0.83	106	0.67
	GGA	90	1.13	218	0.99		GAA	125	0.97	182	1.16
	GGC	92	1.15	211	0.96		GAC	99	0.77	129	0.82
WC						−Y					
	TGG	49	1.29	356	1.30		TAG	2	0.06	72	0.93
	TGT	38	1.00	223	0.81		TAT	52	1.65	54	0.70
	TGA	8	0.21	242	0.88		TAA	4	0.13	113	1.46
	TGC	57	1.50	276	1.01		TAC	68	2.16	70	0.91
RS						KN					
	AGG	55	0.94	277	1.09		AAG	151	1.44	184	1.23
	AGT	52	0.89	174	0.69		AAT	82	0.78	114	0.76
	AAT	82	0.78	114	0.76		AAA	83	0.79	203	1.35
	AGA	43	0.74	273	1.08		AAC	102	0.98	99	0.66
						QH					
	AGC	84	1.44	289	1.14		CAG	173	2.01	271	1.36
R							CAT	48	0.56	153	0.77
	CGG	35	0.80	78	1.17		CAA	62	0.72	141	0.71
	CGT	36	0.82	60	0.90		CAC	61	0.71	234	1.17
	CGA	40	0.91	45	0.67	A					
	CGC	65	1.48	84	1.26		GCG	23	0.30	106	0.50
V							GCT	102	1.35	243	1.14
	GTG	130	1.88	189	1.39		GCA	59	0.78	209	0.98
	GTT	71	1.03	138	1.01		GCC	119	1.57	293	1.38
	GTA	13	0.19	77	0.57	S					
	GTC	63	0.91	141	1.03		TCG	12	0.20	58	0.27
LF							TCT	76	1.26	269	1.26
	TTG	63	1.21	174	1.06		TCA	60	0.99	242	1.13
	TTT	55	1.06	253	1.54		TCC	94	1.55	288	1.34
	TTA	22	0.42	87	0.53	T					
	TTC	68	1.31	143	0.87		ACG	27	0.47	61	0.36
MI							ACT	45	0.79	176	1.04
	ATG	126	1.83	127	1.14		ACA	61	1.07	245	1.45
	ATT	35	0.51	132	1.18		ACC	96	1.68	195	1.15
	ATA	15	0.22	89	0.80	P					
	ATC	100	1.45	98	0.88		CCG	54	0.52	111	0.38
L							CCT	136	1.30	301	1.04
	CTG	175	1.96	348	1.53		CCA	108	1.04	344	1.19
	CTT	65	0.73	231	1.01		CCC	119	1.14	405	1.40
	CTA	38	0.43	93	0.41						
	CTC	79	0.89	240	1.05						

3.2 基于RSCU和QRSCU对三周期性分析

利用Codon W计算出14条CDS和RNA的RSCU和QRSCU值. 为了分析两者的偏好性与三周期性的关系, 类似文献[10, 12]的划分, 分别对两种编码下的偏好性强度值进行了的划分. 基于氨基酸编码下CDS和RNA的偏好性最高值为2.38, 最低值为0.19, 将这之间的数值划分为4个等级: 无偏好性RSCU $\leqslant$ 1.07; 低度偏好性1.07 $<$ RSCU $<$ 1.15; 中度偏好性1.15 $\leqslant$ RSCU $\leqslant$ 1.34; 高度偏好性1.34 $<$ RSCU. 基于拟氨基酸编码下CDS和RNA的偏好性最高值为2.16, 最低值为0.06, 将这之间的数值划分为4个等级: 无偏好性QRSCU $\leqslant$ 1.07; 低度偏好性1.07 $<$ QRSCU $<$ 1.30; 中度偏好性1.30 $\leqslant$ QRSCU $\leqslant$ 1.40; 高度偏好性QRSCU $>$ 1.40. 14条CDS和RNA的RSCU和QRSCU值见表4和表5, 偏好性分布情况见表6和表7.

基于氨基酸编码和拟氨基酸的高度、中度和低度偏好的密码子类型分别如表6和表7所示.

表6 基于氨基酸编码下CDS和mRNA的密码子偏好性

		低度偏好	中度偏好	高度偏好
CDS	密码子	CCC TAC CAC AAC GGA	TCT CCT AAG GAG TGC AGC AGG GGC	CTG ATC GTG TCC ACC GCT GCC CAG TGA TGC
	总数	5	8	10
	C或G结尾	4	6	8
mRNA	密码子	TCA GCT AAT CAC GAG TGC	TTT CTT CTC ATT TCT TCC ACC CAC AGC AGA	CTG GTG CCC ACA GCC TGA CAG GGG AGG
	总数	6	10	9
	C或G结尾	3	5	7

表7 基于拟氨基酸编码下CDS和mRNA的密码子偏好性

		低度偏好	中度偏好	高度偏好
CDS	密码子	GGA GGC TTG TGG TCT CCC	TTC GCT CCT	TGC AGC CGC GTG ATG ATC CTG GAG TAT TAC AAG CAG GCC TCC ACC
	总数	6	3	15
	C或G结尾	4	1	14
mRNA	密码子	AGG AGA AGC CGG CGC ATG ATT GAA AAG CAC GCT TCT TCA ACC CCA	GGG TGG GTG GAG AAA CAG GCC TCC CCC	TTT CTG TAA ACA
	总数	15	9	4
	C或G结尾	8	8	1

从表6和表7可以得到:

1) 基于氨基酸编码下, CDS不论是低度偏好、中度偏好还是高度偏好, 以C或G结尾的密码子的个数和所占比例都比mRNA中的高. 因而, CDS中偏好使用C或G结尾的密码子, 并且偏好程度比mRNA强;

2) 基于拟氨基酸编码下, 偏好程度异同主要表现在高度偏好上, CDS 高度偏好中 15 个偏好类型有 14 个以 C 或 G 结尾, 仅编码拟氨基酸-Y 的一个同义密码子偏好 T 结尾, 而 mRNA 中仅有 1 个高度偏好 C 或 G 结尾的密码子, 因而, CDS 中的密码子拟偏好 C 或 G 结尾的密码子, 偏好程度强于 mRNA.

从氨基酸编码和拟氨基酸编码角度都说明了 P53 蛋白质编码区偏好使用以 C 或 G 结尾的密码子, 且偏好程度远强于对应的 mRNA. 文献 [6, 7] 指出, 产生三周期性的原因是由于密码子使用偏向和蛋白质对某些氨基酸的使用偏向, 这两者都与同义密码子的使用偏向有关. 上述研究中发现 P53 的 CDS 偏好使用 G 或 C 结尾的密码子, 而 mRNA 偏好性不明显. P53 中 CDS 具有三周期性, 对应的 mRNA 不具有三周期性, 这说明 P53 基因密码子的偏好程度与其三周期性紧密相关. 密码子偏好性影响编码的蛋白质, 三周期性是生物长期进化的结果. 这从生物学上解释了 CDS 具有三周期性, 而 mRNA 不具有三周期性.

4 结 论

P53 抑癌基因与肿瘤息息相关, 已经成为癌症治疗研究领域的焦点, P53 基因性质研究的重要性越来越突出. 本文运用 Z-curve 映射将序列数值化, 利用离散傅里叶变换, 采用信号处理和分析的方法对 P53 基因的蛋白质编码区和对应的 mRNA 的三周期性做了比较分析, 得到蛋白质编码区具有很强的三周期性, 而 mRNA 不具有三周期性的结果. 蛋白质编码区与 mRNA 都是由外显子拼接而成的, 外显子具有三周期性, 所以局部三周期性并不能反映全局三周期性. 三周期性是编码蛋白质片段所反映出的特性. 基于氨基酸编码和拟氨基酸编码分析密码子的偏好性, 都得出蛋白质编码序列偏好 G 或 C 结尾的密码子的程度强于对应的 mRNA. 而蛋白质编码区和对应 mRNA 的三周期性又存在异同, 说明 P53 密码子偏好程度与三周期性紧密相关, 密码子的偏好性影响着序列的三周期性. 这也进一步从生物学角度诠释了 P53 蛋白质编码区具有三周期性, 而对应的 mRNA 不具有三周期性的特性. 同时, 计算结果还说明了基于拟氨基酸编码下比基于氨基酸编码下计算更能明显地展现密码子家族中对同义密码子的一致偏好性, 数据显示, 在 QRSCU 分类下理化性质是充分显现的, 同义密码子的偏好与密码子-反密码子间的结合更加紧密. 结合 P53 基因密码子的偏好性, 有利于提高 P53 家族基因识别的正确率, 对 P53 基因性质的进一步研究具有重要的意义.

参考文献

[1] Nantajit D, Fan M, Duru N, Wen YF, Li J J 2010 *Plos. One* **5** e12341

[2] Xia J F, Jia Y 2010 *Chin. Phys. B* **19** 040506

[3] Yan Y Y, Zhu P 2011 *Chin. Phys. B* **20** 018701

[4] Zhang L J, Yan S W, Zhuo Y Z 2007 *Acta Phys. Sin.* **56** 2442 (in Chinese)[张丽娟, 晏世伟, 卓益忠 2007 物理学报 **56** 2442]

[5] Hota M K, Srivastava V K 2010 *Int. J. Computat. Biology Drug Design* **3** 259

[6] Tian Y X, Chen C, Zou X Y, Qiu J D, Cai P X, Mo J Y 2005 *Acta Chim. Sin.* **63** 1215 (in Chinese)[田元新, 陈超, 邹小勇, 邱建丁, 蔡沛祥, 莫金垣 2005 化学学报 **63** 1215]

[7] Zhang J, Shi X F 2002 *Prog. Biochem. Biophys.* **29** 267 (in Chinese) [张静, 石秀凡 2002 生物化学与生物物理进展 **29** 267]

[8] Wang Q Q, Tan C J, Yan H B, Zhu P 2013 *Acta Biophys. Sin.* **29** 296 (in Chinese)[王其强, 谈承杰, 晏寒冰, 朱平 2013 生物物理学报 **29** 296]

[9] Shi X F, Huang J F, Liang C R, Liu S Q, Xie J, Liu C Q 2000 *Chin. Sci. Bull.* **45** 2520 (in Chinese)[石秀凡, 黄京飞, 梁宠荣, 柳树群, 谢君, 刘次全 2000 科学通报 **45** 2520]

[10] Zhu P, Gao L, Xu Z Y 2009 *Acta Phys. Sin.* **58** 714 (in Chinese) [朱平, 高雷, 徐振源 2009 物理学报 **58** 714]

[11] Sharma S D, Shakya K, Sharma S N 2011 *International Conference on Computer, Communication and Electrical Technology-ICCCET 2011* March, 18–19 p71

[12] Zhao J J, Qi B, Ding L J, Tang X Q 2010 *J. Food Sci. Biotechnol.* **29** 755 (in Chinese)[赵静静, 齐斌, 丁利娟, 唐旭清 2010 食品与生物技术学报 **29** 755]

[13] Berryman M J, Allison A 2005 *Fluct. Noise Lett.* **5** 13

Relation between the 3-base periodicity of P53 gene and codon usage bias*

Wang Qi-Qiang Tan Cheng-Jie Zhu Ping†

(*School of Science, Jiangnan University, Wuxi 214122, China*)

(Received 15 September 2013; revised manuscript received 25 October 2013)

Abstract

To further study the properties of P53 suppressor gene, the 3-base periodicity of P53 coding sequence (CDS) and the corresponding mRNA are analyzed. And the codon biases of P53 CDS and mRNA are discussed through analyzing their relative synonymous codon usage and quasi relative synonymous codon usage values. The results show that the CDS of P53 exhibits 3-base periodicity, whereas the corresponding mRNA of P53 does not, and that the P53 CDS has a stronger bias towards C and G ending codons than the mRNA. This suggests that the 3-base periodicity is closely related to the codon usage bias of P53 gene, and the degree of codon bias has an effect on the 3-base periodicity, which further explains the difference in 3-base periodicity between P53 CDS and mRNA from the point of view of biology. This characteristic of P53 gene may be useful in increasing the correct rate of gene recognition and the extensive investigation of P53 gene.

Keywords: 3-base of periodicity, codon usage bias, P53 gene, relative synonymous codon usage

PACS: 87.10.–e, 87.14.gk, 87.15.–v **DOI:** 10.7498/aps.63.048701

* Project supported by National Natural Science Foundation of China (Grant No. 11271163) and Fundamental Research Fund for the Central Universities of China (Grant No. JUSRP51317B).

† Corresponding author. E-mail: zhuping@jiangnan.edu.cn