不具有 3-碱基周期性的编码序列初探*

张 静[1)**] 石秀凡[2)]

([1)]云南大学应用统计中心，昆明 650091；[2)]中国科学院昆明动物研究所，昆明 650223)

摘要 对 120 个较短编码序列 (< 1 200 bp) 的 Fourier 频谱进行分析表明，3-碱基周期性在短编码序列中并不是绝对存在的. 统计分析提示，编码序列有无 3-碱基周期性与序列的碱基组成和分布、所编码蛋白质氨基酸的选用和顺序以及同义密码子的使用都有一定的关系. 一般地，非周期-3 序列中 A+ U 含量高于 G+ C 含量，周期-3 序列的情况则相反；非周期-3 序列中碱基在密码子三个位点上的分布比周期-3 序列中的分布均匀；非周期-3 序列密码子和氨基酸的使用偏向没有周期-3 序列的大. 在利用 Fourier 分析方法预测 DNA 序列中的基因和外显子时，应充分考虑到这些现象.

关键词 编码序列，Fourier 分析，3-碱基周期性，非 3-碱基周期性

学科分类号 Q617

在基因中，编码蛋白质的序列具有 3-碱基周期性 (周期-3 性质)，这是许多研究工作的结果[1~3]. 一般认为，产生此性质的原因是两个方面：a. 密码子的使用偏向. 分析表明 $(RNY)_n$ (R 表示 A 或 G，Y 表示 C 或 U，N 表示任意碱基) 是较多出现的密码形式，而且密码子第一位对 R 的偏向主要是使用了 G；$(GCU)_n$ 型密码被认为能够在翻译过程中保持正确的阅读框架[4, 5]. Lee 等[2]作了透彻的分析后指出，一个碱基序列是否具有周期-3 性质的充分必要条件，是碱基在三个位点上的分布是否均匀，而 $(GCU)_n$ 型序列正好是非均匀分布的. b. 自然界中存在的蛋白质对某些氨基酸的使用偏向[6]. 当然，氨基酸的使用偏向亦可导致密码子的使用偏向，因此关于密码子的使用偏向，主要是考虑同义密码子的使用偏向. 关于这两方面还有一个共同点就是认为密码子和氨基酸的偏向使用都与早期的进化有关，因此，3-碱基周期性亦就应该是伴随基因出现与进化的一个本质特征.

随着基因数据库的不断丰富，对大量序列的进一步研究证实，3-碱基周期性只是蛋白质编码序列特有的，非编码序列 (如内含子等) 都不具有此性质[7, 8]. 于是，有不少研究者利用这一性质来寻找 DNA 序列中可能的蛋白质编码序列以及基因. Tiwari 等[9]对几个不同的物种进行识别试验，其灵敏度随着物种的不同而有差异，其中，对人基因中外显子识别的灵敏度达 0.9，酵母中为 0.86，对一个综合序列 ALLSEQ 检测的灵敏度为 0.66. 对基因识别的灵敏度更高一些. 虽然识别的灵敏度不低，但是我们还是需要对错判或漏判外显子的原因进行一番考察. 首先想到的自然是识别的出发点，即编码序列或外显子的周期-3 性质. 注意到以往研究外显子 (或编码序列) 短程关联性的方法一般是将多个外显子连接起来进行分析，求出所连序列的自相关函数，然后进行 Fourier 变换得到其频谱，发现 3 碱基处出现高峰，由此断定外显子序列具有周期-3 性质[10]. 可见这样得到的周期-3 是一个“全局”性质，即“长外显子”所具有的性质，事实上，用这样的分析方法，只要这个长序列中有一部分序列具有周期-3 性质，整个序列的频谱也会显示出周期-3 信号. 因此，用这样一个“全局”特征去判断一个“局部”序列 (单个外显子) 是否具有整体序列的性质，有可能会得出相反的结论.

问题变为每个外显子序列 (或短编码序列) 是否都具有 3-碱基周期性？对此，我们分析了大量较短的编码序列，发现其中有一些并不具有周期-3 性质. 进一步的分析还发现，无周期-3 的编码序列的碱基组成及分布、密码子和氨基酸使用情况与具有周期-3 的序列有些差异，这是以前不曾注意到的现象.

1 样本和方法

1.1 本文的目的主要是研究不具有3碱基周期性

* 国家教育部科研重点基金资助项目 (00246).

** 通讯联系人.

Tel: 0871-5036207, E-mail: jzhang213@263.net

收稿日期：2001-07-10，接受日期：2001-08-27

的编码序列，前已提及一个长序列中若有一部分序列具有某个周期性，则整个序列的频谱中多少都会显示出该周期峰. 所以选取样本时就尽量避免这种情况，即选取一些小蛋白质的编码序列. 所选蛋白质的氨基酸数一般小于 400，亦即每个编码序列的碱基数不超过 1 200. 我们选取了包括真核生物和原核生物的编码序列共计 120 个.

1.2 画出每个编码序列的频谱图并分析其周期性. 具体方法是：设碱基序列为 $(a_1, a_2, a_3 \cdots, a_N)$，其中 $a_i \in$ (A, G, C, U)，$i = 1, 2, \cdots, N$. 定义序列的自相关函数为:

$$p(n) = \sum_{i=1}^{N} p_i(n)$$

其中

$$p_i(n) = \begin{cases} 1, \text{当 } a_i = a_{i+n} \\ 0, \text{其他, } i + n > N \end{cases}$$

然后对 $p(n)$ 作 Fourier 变换:

$$s(f) = \frac{1}{N} \sum_{n=1}^{N} p(n) e^{-j2\pi nf}$$

$$(f = k/N,\ k = 1, 2, \cdots, N/2)$$

由此可得序列自相关频谱的强度 $P(f) = |s(f)|$，从而作出频谱图，如图 1 所示.

$$P = \frac{2}{N} \sum_{k=1}^{N/2} P(k/N)$$

当信噪比 $R = P\ (1/3)\ /P > 4$ 时，就认为序列具有周期-3 性质.

本文的计算和图形显示是用 MATLAB5.3 软件完成的.

2 结　果

2.1 分析120个编码序列的频谱图发现，绝大多数序列表现出较强的周期-3 性质 (图 1a)，然而有 25 个序列不具有周期-3 性质 (图 1b、图 1c、图 1d)，其中一些序列虽然没有周期-3 信号，但是有其他周期信号.

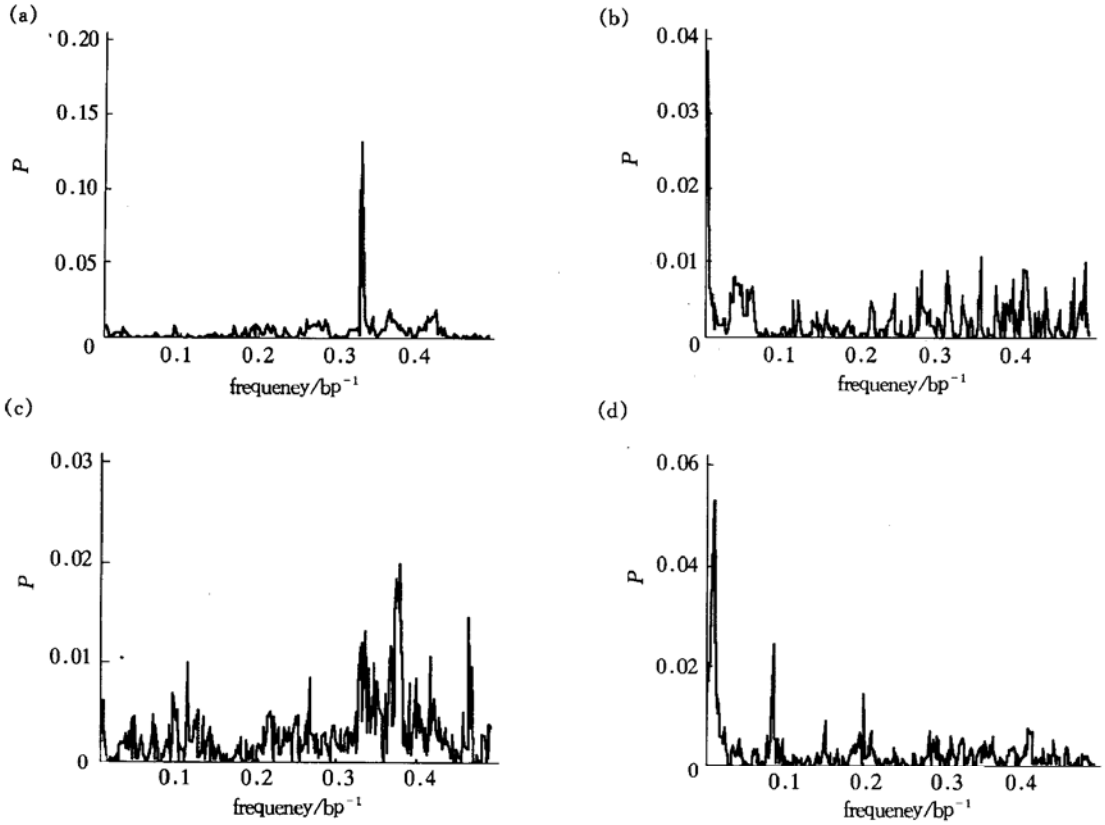

Fig. 1 The Fourier spectra for the coding sequences of (a) 2hhb (identifier of protein in PDB, similarly hereinafter), (b) 1mdy, (c) 2stv and (d) 3cla

A strong signal occurs at 1/3 bp in (a). No peak is shown at 1/3 bp in (b), (c) and (d).

2.2 为了探究编码序列有无周期-3 的原因，我们分别对具有较强周期-3 信号和无周期-3 信号 ($R < 4$) 序列的碱基组成和分布、密码子及氨基酸的使用情况进行了统计分析. 有较强周期-3 信号的编码序列中，A+U 的含量较低，平均为 44%，G+C 含量平均为 56%；而在无周期-3 信号的编码序

列中，A+ U 含量平均为 52%，G+ C 为 48%．各种碱基在两类序列中密码子三个位点上分布的平均标准差列于表 1，从表 1 中可见分布的离散程度．

Table 1 Average standard deviations of base distributions in three codon positions for the period-3 and the non-period-3 coding sequences respectively

	Non-period-3 coding sequence				Period-3 coding sequence			
Base	A	U	G	C	A	U	G	C
Average standard deviation	7	7	9	4	17	11	18	19

2.3 两种序列中氨基酸的平均出现频率 (= 某氨基酸的出现数/各种氨基酸出现的总数) 和密码子的相对使用值 (= 某密码子的使用数/61 个密码子的平均使用数) 列于表 2．从表 2 中可以看出，氨基酸在两种序列中的出现频率虽然相差不是太大，但是也有一些偏向，特别是 Val、Ala 和 Gly 的差异较大，在周期-3 序列中出现的频率高于非周期-3 的频率至少是 1.9%，这 3 个氨基酸的密码子第一位碱基均为 G．就各类序列而言，有周期-3 的序列中氨基酸使用的偏向要比无周期-3 序列中的大一些．有周期-3 的序列中，氨基酸的出现频率最高为 9.0%，最低的只有 1.2%；而在无周期-3 的序列中，最高为 7.6%，最低的为 1.7%．

Table 2 The usages of codons and amino acids occurring in the period-3 and in the non-period-3 coding sequences respectively

Amino acid	Codon	Amino acid	Codon
Phe (3.3/3.9)[1)]	UUU (0.8/1.3)[2)]	Ala (9.0/6.6)	GCU (1.2/1.2)
	UUC (1.2/1.1)		GCC (2.3/1.4)
Leu (8.3/9.3)	UUA (0.2/0.9)		GCA (0.8/0.8)
	UUG (0.4/1.0)		GCG (1.3/0.6)
	CUU (0.5/0.8)	His (1.8/2.3)	CAU (0.5/0.7)
	CUC (1.0/0.7)		CAC (0.6/0.7)
	CUA (0.2/0.6)		
	CUG (2.7/1.7)	Gln (3.3/4.4)	CAA (0.6/1.2)
Ile (4.6/4.7)	AUU (0.9/1.2)		CAG (1.5/1.5)
	AUC (1.6/1.0)	Asn (4.5/5.6)	AAU (1.0/1.8)
	AUA (0.3/0.7)		AAC (1.7/1.6)
Met (1.9/2.2)	AUG (1.2/1.4)	Lys (5.2/4.9)	AAA (1.6/1.6)
Val (7.5/5.6)	GUU (0.8/1.0)		AAG (1.6/1.4)
	GUC (1.3/0.6)	Asp (6.0/4.8)	GAU (1.4/1.8)
	GUA (0.5/0.6)		GAC (2.3/1.1)
	GUG (2.0/1.5)	Glu (6.0/6.7)	GAA (1.7/2.0)
Ser (7.1/7.6)	UCU (0.9/0.8)		GAG (1.9/2.1)
	UCC (1.1/0.9)	Cys (2.7/2.9)	UGU (0.6/1.0)
	UCA (0.3/0.8)		UGC (1.0/0.8)
	UCG (0.4/0.3)	Arg (5.1/6.2)	CGU (0.5/0.5)
	AGU (0.4/0.9)		CGC (1.1/0.7)
	AGC (1.3/1.1)		CGA (0.2/0.3)
Pro (5.2/5.4)	CCU (0.6/1.1)		CGG (0.5/0.6)
	CCC (1.1/0.8)		AGA (0.4/1.2)
	CCA (0.5/1.0)		AGG (0.4/0.6)
	CCG (0.9/0.5)	Gly (9.0/6.8)	GGU (1.3/0.9)
Thr (5.5/5.5)	ACU (0.8/0.9)		GGC (2.6/1.2)
	ACC (1.5/1.0)		GGA (0.8/1.2)
	ACA (0.5/1.1)		GGG (0.9/0.9)
	ACG (0.6/0.4)	Trp (1.2/1.7)	UGG (0.8/1.0)
Tyr (2.9/3.1)	UAU (0.6/1.0)		
	UAC (1.1/0.9)		

[1)] The left (right) number of "/" denotes the occurrence frequency (%) of amino acids in the period-3 (non-period-3) sequences.
[2)] The left (right) number of "/" denotes the relative codon usage (RCU) in the period-3 (non-period-3) sequences.

从表 2 中还可以看出，某些同义密码子的使用在两类序列中有差异. 例如 Ile 在两类序列中的使用频率很接近，分别为 4.6% 和 4.7%，但是密码子 AUC 在具有周期-3 序列中的相对使用值为 1.6，而在非周期-3 序列中的相对使用值为 1.0；密码子 AUA 在具有周期-3 序列中的相对使用值为 0.3，而在非周期-3 序列中的相对使用值为 0.7. Leu 密码子在两类序列中的使用偏向也较大. 一般地，周期-3 序列中偏向使用含 G、C 多的密码子，非周期-3 序列的情况则相反，偏向使用含 A、U 的密码子. 这种偏向刚好与两种序列不同的碱基成分相对应，反映了序列的碱基成分对密码子使用度的影响.

3 讨　论

以上结果显示了有无周期-3 编码序列碱基组分的一般情况，即具有周期-3 的序列中 G+C 含量 (56%) 高于 A+U 含量 (44%)，而在非周期-3 的序列中情况相反，A+U (52%) 高于 G+C (48%). 当然，这只是一个统计平均数，并不是 A+U 含量高的编码序列一定没有周期-3 性质，事实上，我们分析的样本中就有少数几个序列的 A+U 含量很高 (> 60%)，但却有较强的周期-3 信号. G+C 含量高的序列中也有无周期-3 性质的，只是数量较少. 例如 1mdy 的编码序列中 G+C 含量高达 73%，但它却没有周期-3 性质 (图 1b). 在无周期-3 的序列中，虽然 A+U 含量比 G+C 含量高得不是太大，但是如果考虑到 61 个密码子中 A+U 含量比 G+C 的少 (三个终止密码子 UAA、UAG 和 UGA 主要由 A、U 组成)，那么可以说在非周期-3 的编码序列中，使用较多的密码子应为 A+U 含量较高的；而具有周期-3 的序列中，使用较多的密码子则为 G+C 含量较高的，如 CUG、GUG、GCC 和 GGC 等. 由于密码子的碱基组成涉及到两个方面，即氨基酸的选择和同义密码子的使用，所以两类序列中不同的碱基组分与各自序列对氨基酸的选择及同义密码子的使用最终是相互联系的，亦即两类序列对氨基酸和密码子的使用各有偏向，表 2 的结果便体现了这一点.

Lee 等[2]所作的理论分析表明，一个碱基序列的谱线在 1/3bp 处出现峰的充分必要条件是碱基在密码子三个位点上的分布不均匀，这亦说明，如果碱基在三个位点上的分布均匀，则谱线在 1/3 bp 处不会出现峰. 表 1 的结果显示，与周期-3 序列相比，非周期-3 序列中碱基在三个位点上分布的标准差要小得多，这说明非周期-3 序列中碱基在三个位点上的分布比周期-3 序列中的分布均匀得多，按 Lee 等的结果，其谱线在 1/3bp 处自然就不会有明显的峰值.

表 2 的结果显示同义密码子的不同选用与周期-3 性质似乎有一些联系，为了证实这个结果，我们对几个序列进行了碱基变换分析. 例如 1ifc 编码序列的 Fourier 谱在 1/3 bp 处的信号极弱 ($R < 4$)，而在 1/5bp 处有较强的信号 (图 2a)，将部分密码子的第三位碱基进行同义变换 (即保持氨基酸及其序列不变)，变换后的序列 Fourier 谱中，1/3 bp 处出现了很强的信号，而 1/5 bp 处却没有谱峰 (图 2b). 这说明同义密码子的使用会影响序列的周期性. 这个结论与 Tiwari 等[9]有所不同，他们对几个序列做同义变换后认为，周期-3 性质也许与同义密码子的使用无关. Zhurkin[6]认为，编码序列的周期性与蛋白质二级结构有关，我们的研究也显示，蛋白质的二级结构与密码子前后碱基的使用有关[11, 12]，这似乎也说明周期性与同义密码子使用有一定的关联性.

此外，除了氨基酸的成分，氨基酸的排列序列对周期性也有影响. 例如 1hdr 的编码序列有较强的周期-3 信号 (图 3a)，我们保持密码子不变，而将密码子的顺序 (亦即氨基酸序列) 进行随机调整后得到一个新碱基序列，新序列的图谱在近 1/5 bp 处显示出较强的信号，而在 1/3 bp 处的信号变得很弱 (图 3b). 我们注意到，1hdr 的编码序列中有几个 $(GCU)_n$ 片段，研究认为，这种类型的片段从结合能上看有利于 rRNA 位点与 mRNA 的相互作用，所以在翻译过程中能保持正确的阅读框架[4, 5]. 这样的序列片段显然是周期-3 的生成因素[2]. 因此，编码序列中普遍存在的 3-碱基周期性也许与核糖体的翻译机制和基因的进化有关；而非周期-3 编码序列的存在似乎提示了翻译机制和基因的复杂性.

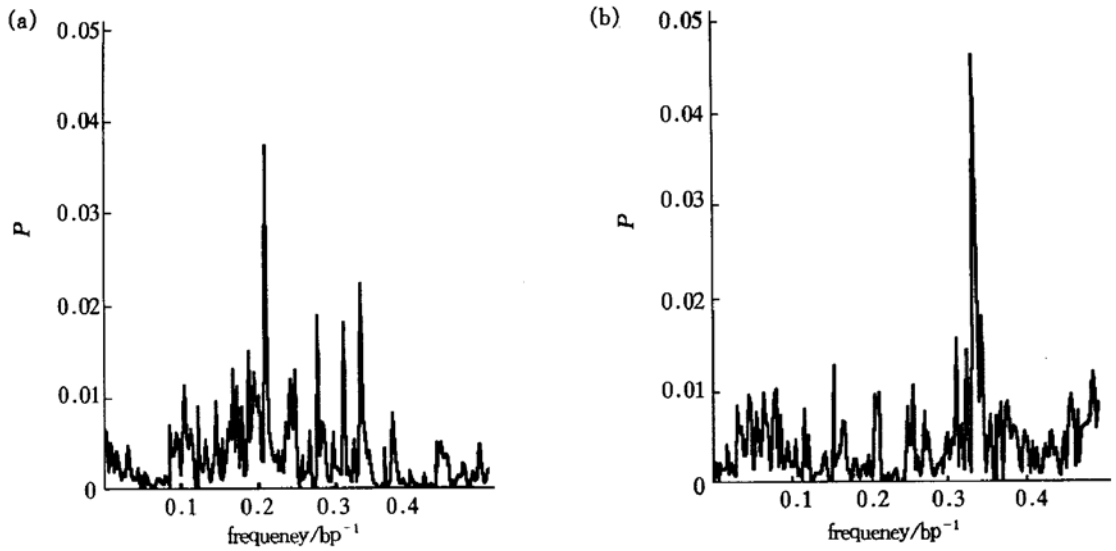

Fig. 2　The Fourier spectra for (a) the coding sequence of 1ifc and (b) the sequence in which the third bases of some codons of 1ifc are changed synonymously

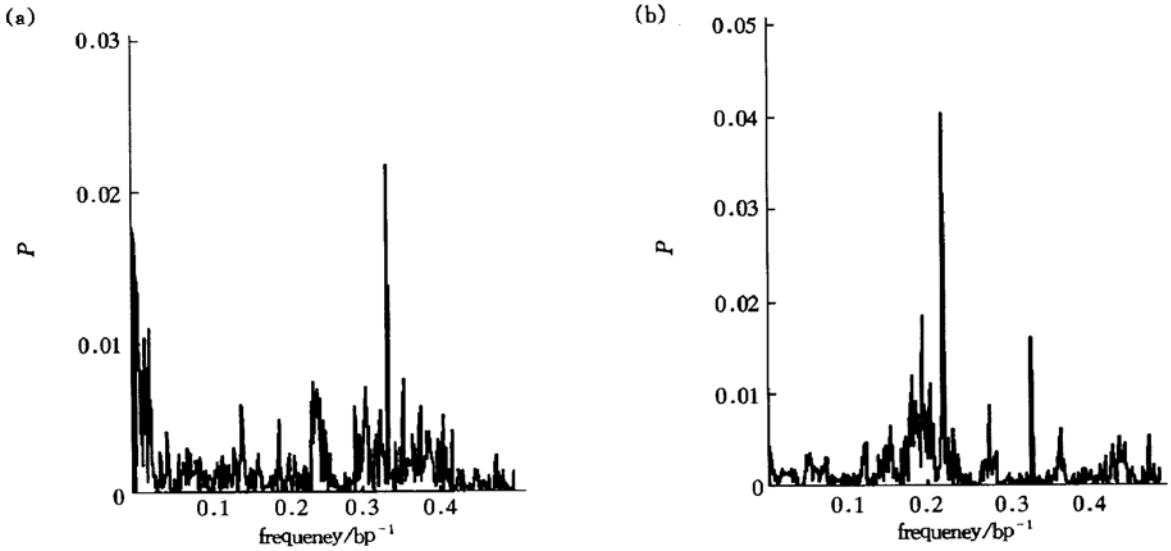

Fig. 3　The Fourier spectra of (a) the coding sequence of 1hdr and (b) the sequence in which the codon order of 1hdr is shuffled

综上所述，编码序列是否具有周期-3 性质可能是由三个因素决定的：a. 碱基的成分和分布；b. 氨基酸的使用和排列顺序；c. 同义密码子的使用度.

还需要说明的是，长编码序列一般都有周期-3 信号，但是周期-3 信号的强弱与序列的长度没有太大的关系. 在我们所研究的 120 个样本中，除了一个序列外，无周期-3 的序列都少于 1 000 bp. 但是在很短的序列 (< 500 bp) 中也有一些具有很强的周期-3 性质. 当然，无周期-3 的编码序列多为短序列. 因此我们提醒研究者，在利用 Fourier 方法预测外显子序列时，要注意这个现象，以尽量避免遗漏，本文的结果可为调整预测方法提供一些参考. 由于非周期-3 编码序列的数量毕竟较少，样本的收集有一定困难，本文的研究中还只收集到 25 个样本，但从分析结果看非周期-3 和周期-3 序列中碱基的成分和分布、氨基酸的使用和排列顺序以及同义密码子的使用等因素的差异是存在的，特别是碱基在密码子三个位点上的分布与理论结果的吻合更说明了这一点. 当然本文的研究还只是一个初步的探索，更深入的研究还在进行中.

参 考 文 献

1　Voss R F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett, 1992, **68** (25): 3805~3808

2　Lee W J, Luo L F. Periodicity of base correlation in nucleotide

sequence. Phys Rev E, 1997, **56** (1): 848~ 851
3 Trifonov E N, Sussman J L. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA, 1980, **77** (7): 3816~ 3820
4 Trifonov E N. Translation framing code and frame-monitoring mechanism as sugested by the analysis of mRNA and 16S rRNA nucleotide sequenes. J Mol Biol, 1987, **194** (4): 643~ 652
5 Trifonov E N. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A, 1998, **249** (1~ 4): 511~ 516
6 Zhurkin V B. Periodicity in DNA-primary structure is defined by secondary of the coded protein. Nucleic Acids Res, 1981, **9** (8): 1963~ 1971
7 Peng C K. Long-range correlations in nucleotide sequences. Nature, 1992, **356** (6365): 168~ 170
8 Buldyrev S V, Goldberger A L, Havlin S. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E, 1995, **51** (5): 5084~ 5094
9 Tiwari S, Ramachandran S, Bhattacharya A, et al. Prediction of probaale genes by Fourier analysis of genomic sequences. Comput Appl Biosci, 1997 (3), 263~ 270
10 Li W, Marr T G, Kaneko K. Understanding long-range correlations in DNA sequences. Physica D, 1994, **75**: 392~ 416
11 张 静, 顾宝洪, 石秀凡, 等. 人基因中密码子前后碱基使用与蛋白质结构. 生物物理学报, 2000, **16** (4): 769~ 774
Zhang J, Gu B H, Shi X F, *et al*. Acta Biophysica Sinica, 2000, **16** (4): 769~ 774
12 张 静, 顾宝洪, 石秀凡, 等. 大肠杆菌基因中密码子前后碱基的使用与蛋白质结构. 生物物理学报, 2001, **17** (1): 174~ 180
Zhang J, Gu B H, Shi X F, *et al*. Acta Biophysica Sinica, 2001, **17** (1): 174~ 180

Tentative Study of the Coding Sequences Without 3-base Periodicity*

ZHANG Jing**
(The Center of Applied Statistics, Yunnan University, Kunming 650091, China)
SHI Xiu-Fan
(Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming 650223, China)

Abstract Fourier spectra of 120 short coding sequences (< 1 200 bp) show that not all coding sequences are characterized by 3-base periodicity. Statistical analysis suggests that whether a coding sequence has 3-base periodicity may be related to the composition and distribution of bases, the usage and the order of the amino acids of the encoded protein as well as the synonymous codon usage. Generally, the content of A+ U is higher than that of G+ C in non-period-3 sequences, inversely in period-3 sequences. In the three codon positions, the base distribution in the non-periodic-3 sequences is more uniform than in the periodic-3 sequences. The usage biases of the amino acids and the codons in non-period-3 sequences are weaker than that in period-3 sequences. All of these phenomena should be considered sufficiently in predicting the genes and exons of DNA sequences by Fourier analysis method.

Key words coding sequence, Fourier analysis, 3-base periodicity, non-3-base periodicity

* This work is supported by a grant from Ministry of Education of China for Key Program in Science Researches (00246).
** Corresponding author. Tel: 86-871-5036207, E-mail: jzhang213@263.net
Received: July 16, 2001 Accepted: August 27, 2001