Chinese Chemical Letters Vol. 16, No. 7, pp 939-942, 2005
http://www.imm.ac.cn/journal/ccl.html

Fourier Power Spectrum Analysis of Exons for the Period-3 Behavior

Yuan Xin TIAN[1], Chao CHEN[1], Xiao Yong ZOU[1,3]*, Jian Ding QIU[2], Pei Xiang CAI[1], Jin Yuan MO[1]

[1]School of Chemistry and Chemical Engineering, Sun Yet-Sun University, Guangzhou 510275
[2] Department of Chemical Engineering, Pingxiang College, Pingxiang 337055
[3] State Key Laboratory of Chem/Biosensing and Chemometrics, Hunan University, Changsha 410082

Abstract: The period-3 behaviors of 105 exons from 20 genes in human were studied by Fourier power spectrum. The results indicated that not all exons show the period-3 behavior. The exons were adjusted in order to make them accord with the order of the protein translated, and we found that the period-3 character is relation to the length of exons and the bases distribution in the three codon position. Furthermore, as long as the exons with period-3 behavior accord with the order of protein translated, they would exhibit the synonymous codons usage preference, and the codons with g/c at the third position are used in higher frequency. The results are significant to the gene prediction and the research on the introns.

Keywords: Period-3 behavior, Fourier power spectrum, exons.

The number of available complete genomes is increasing at a fast pace and as a result it imposes a great challenge to the scientists. The most important task is extracting more and more information from DNA sequences, which is valuable to the biology, medicine and pharmacy. The analysis of biology information is a new hotspot of analytical chemistry[1, 2]. Gene is a DNA strings arranged by adenosine (A)，guanine (G)，cytosine (C)，and thymine(T). The genetic information would be transferred correctly through 64 genetic codons and instruct protein synthesis, then carry out the whole life. In advanced organisms, protein coding regions are typically separated into several isolated subregions called exons, the regions between two successive exons are called introns, and they are eliminated before protein coding through a process called splicing. It is well known that the protein coding regions have exhibited period-3 behavior, which is the base of gene prediction[3, 4]. In the past, coding protein regions would be studied after being connected[5]. Zhang[6] considered as long as a subregion exhibits period-3 behavior, the whole sequences would exhibit the same character. It would draw a wrong conclusion if the 'whole' character was used instead of the 'partial'. In this paper, we focus on the period-3 behavior of separate exons.

The period-3 behavior of coding protein region refer to the maximum of Fourier power spectrum (FPS) at the position of 1/3 frequency. Fourier transform (FT) is used

* E-mail: ceszxy@zsu.edu.cn

to process discrete numerical sequence, while DNA sequence is a symbolic sequence. So the DNA sequences should be mapped to numerical sequences firstly. A given symbolic sequence is converted into four binary signal using projection operators X_α, α=A, T, G, C which replace the symbol sequence by a digital signal containing 1 in those position where the base is α and 0 elsewhere[7], for example, X_A= 1011010000---for the sequence AGAATACTGC----. Then each of the four signals is FT analyzed. The power spectrum of the whole sequence is the sum of the four signals power spectrums.

The 105 exons of 20 genes in human are chosen randomly from GenBank. These genes lie in different chromosome with different length. The direction of translation is all from 5′ to 3′. Each gene codes only one protein. The reason of period-3 behavior is due to the presence of 3-nucleotide code structure in protein coding region. Therefore we first consider whether the exons are integer multiple of 3. The results indicated that most of exons are not the integer multiple of 3, and not all the exons have period-3 behavior. Then the exons were adjusted to make them satisfy the following criterion: (1) accord with the order of protein translation, (2) to be the integer multiple of 3. For example, sequence 2 in **Figure 1** depicts that the four exons of gene TNF lying in No.6 chromosome. The bases with shadow and sash were deleted and then all the exons were satisfied with the two criterions above. The amide acid sequences coded by the adjusted exons are parts of protein sequence (sequence 1 in **Figure 1**). If not, the amide acid sequences are not the parts of protein sequence (sequence 3 in **Figure 1**).

Figure 1 The comparison of translated exons

Sequence 1[*1] M S------E E F P-------A V R S S ----- V V A N P ------A L *
Sequence 2[*2] atg agc---gaa gag | ttc ccc-----gca gtc a | ga tca tct-----gtt gta g | ca aac cct---- gcc ctg tga
Sequence 3[*3] M S--------E E | F P----------A V | D H ------- L * | Q T L----- PC
exon1(186bp) exon2(46bp) exon3(48bp) exon4(422bp)
(*1: the amide acid of protein; *2: the coding-protein sequence in gene; *3: the amide acid sequences translated from exons)

105 exons were adjusted as mentioned above, and FPS of every exons (including the original and after adjusted) were calculated. **Table 1** displayed the frequency corresponding to maximum of FPS. The results demonstrated that most of exons are not the integer multiple of 3, they have no period-3 behavior before adjusted. The exons, which is not the integer multiple of 3 but accord with the order of protein translated, exhibit the period-3 behavior after adjusted (such as the No1. and No3. exon of gene RETN in **Table 1**). So the integer multiple of 3 has influenced on the period-3 behavior, and the genetic code is the inherent factor. We also noticed that so far as any exon in a gene has the period-3 behavior (both original and after adjusted), the whole coding-protein sequence would exhibit the period-3 behavior. These factors should be considered in gene prediction.

It was reported that the main reason of the period-3 behavior is the lopsided distribution of bases in the three codon position and the G/C content at the third position is higher than other positions [8]. We also studied the distribution of bases in the exons, which are the integer multiple of 3 and exhibit the period-3 behavior. **Figure 2** displays

Table 1 The period-3 behavior statistics of exons

Gene name	No.	Length（bp）	Freq (O)*[1]	Freq (C)*[2]
SERPINE1	1	271	0.0148	0.3333
	2	234	0.3333	0.3333
	3	195	0.3333	0.3057
	4	199	0.3317	0.3333
	5	101	0.0990	0.1010
	6	87	0.3448	0.3452
	7	84	0.3333	0.3333
	8	38	0.1597	0.1667
RETN	1	118	0.4746	0.3333
	2	78	0.1282	0.1333
	3	131	0.4427	0.3333

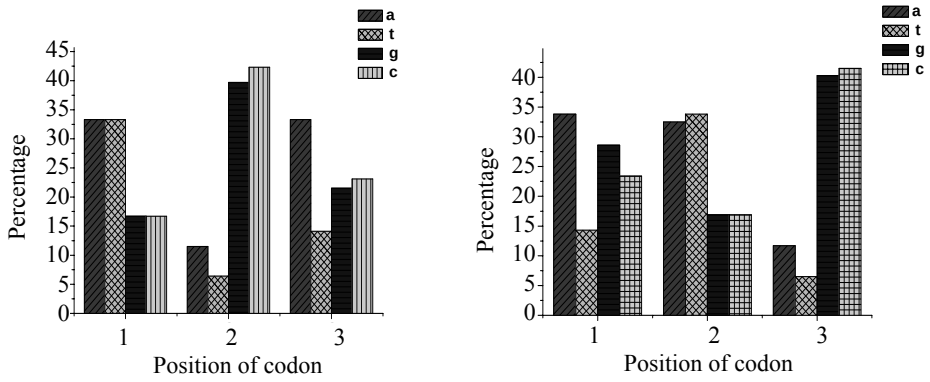
*[1] Freq(O)refer to the frequency corresponding to the maximum of FPS of the origin exons;
*[2] Freq(C) refer to the frequency corresponding to the maximum of FPS of the adjusted exons. The shadow on the number show both the original exon and the adjusted exon appear the period-3 behavior.

the bases distribution of the No 2. exon from gene SERPINE1. The exon exhibited period-3 behavior before and after being adjusted. It is observed that the base distribution is imbalance while GC content is not higher in the third position of codon before adjusted, but the exon also appears the period-3 behavior, shown in **Figure 2(A)**. After the exon is adjusted in order to accord with the order of protein translated, GC content is higher at the third position of the codon, shown in **Figure 2(B)**. This is an evidence that codon with g/c at the third position was used preferentially when protein come into being in the nature.

In order to make a thorough study on the influence of the synonymous codon usage preference, the codon usage of coding protein sequences from 20 genes was investigated. For the whole coding-protein sequences, the statistic results revealed that the codons with g/c at the third position are used more frequently in the sequences, which exhibit the period-3 behavior. Otherwise, the codons are used evenly. The synonymous codons usage of the separated exons has been studied, and we found that the sequences according with the translation order appear the codons usage preference, just the same as the whole coding protein sequences. It comes to a conclusion that the codons usage preference is another character of the coding-protein sequences again.

Consequently, the period-3 behavior of exons is related to the lopsided distribution of bases. So far as the bases distribution is imbalance, the exons probably show period-3 behavior. The length of exons, whether it is the integer multiple of 3 or not, influenced on the behavior. And the genetic code is the inherent factor. The codons usage preference is the unique character of the coding-protein sequences. Exons appear the codons usage preference if it accord with the protein translation order for every exon. It is promising to improve the gene prediction accuracy, if the period-3 behavior and the codon usage preference are considered.

Figure 2 The lopsided distribution of bases in the three position of codon before adjust (A) and after adjust (B)



(A) before adjust (B) after adjust

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants 20475068), the Natural Science Foundation of Guangdong Province (Contact No. 031577), and the Opening Foundation of State Key Laboratory of Chem/Biosensing and Chemometrics of Hunan University (2003).

References

1. J. D. Qiu, R. P. Liang, X. Y. Zou, J. Y. Mo. *Chin. Chem. Lett.*, **2004**, *15*(6), 711.
2. X. Y. Chen, L. J. Bao, J. Y. Mo, P. X. Cai. *Chin. Chem. Lett.*, **2003**, *14*(5), 503.
3. P.P. Vaidyanathan, B. J. Yoon. *Journal of the Franklin Institute*, **2004**, *341*(1-2), 111.
4. G. Aggarwal, R. Ramaswamy. *J. Biosci.*, **2002**, *27*(1), 7.
5. W. Li, T. G. Marr, K. Kaneko, *Physica D*, **1994**, *75*, 392.
6. J. Zhang, X. F Shi, Prog. *Biochem. Biophys.*, **2002**, *29*(2), 267.
7. R. F. Voss, *Phys. Rev. Lett.*, **1992**, *68*(25), 3805.
8. W. J. Lee, L. F. Luo, *Phys. Rev. E,* **1997**, *56*, 848.

Received 16 Augest, 2004