Journal of Southeast University (English Edition) Vol. 37, No. 2, pp. 177 - 183 June 2021 ISSN 1003—7985

Evolutionary selection on synonymous codons in RNA G-quadruplex structural region

Xu Yuming Qi Ting Gu Wanjun Lu Zuhong

(School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China)
(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract: To investigate how synonymous codons have been adapted to the formation of ribonucleic acid (RNA) G-quadruplex (rG4) structure, a computational searching algorithm G4Hunter was applied to detect rG4 structures in protein-coding sequences of mRNAs in five eukaryotic species. The native sequences forming rG4s were then compared with randomized sequences to evaluate selection on synonymous codons. Factors that may influence the formation of rG4 were also investigated, and the selection pressures of rG4 in different gene regions were compared to explore its potential roles in gene regulation. The results show universal selective pressure acts on synonymous codons in rG4 regions to facilitate rG4 formation in five eukaryotic organisms. While G-rich codon combinations are preferred in the rG4 structural region, C-rich codon combinations are selectively unfavorable for rG4 formation. Gene's codon usage bias, nucleotide composition, and evolutionary rate can account for the selective variations on synonymous codons among rG4 structures within a species. Moreover, rG4 structures in the translational initiation region showed significantly higher selective pressures than those in the translational elongation region.

Key words: ribonucleic acid (RNA) structure; G-quadruplex; synonymous codons; evolution; selection
DOI: 10.3969/j.issn.1003-7985.2021.02.007

mRNAs of protein-coding genes in eukaryotic organisms carry the genetic information to encode amino acid sequences and contain multiple regulatory and structural signals[1]. The pivot provision enabling mRNA to hold these regulatory functions is the redundancy of the genetic code that allows for many "silent" mutations at synonymous codon sites[2]. Synonymous nucleotide mutations do not change the amino acid sequences of the encoded proteins; however, they can confer dramatic differences to the structure and functions of mRNA[1-3]. Much evidence has shown that synonymous codons are selected for optimized RNA stability[4-5], proper nucleosome positioning[6], efficient mRNA splicing[7-8], correct microRNA targeting[9], efficient translation initiation[10-12] and elongation[13], and proper protein co-translational folding[14-15].

The RNA G-quadruplex (rG4) is a non-canonical super-secondary structure around the G-rich sites in RNA sequences, which consists of the stacking of G-quartets formed by the G-G Hoogsteen hydrogen bonding[16]. Several experimental studies have shown that rG4 is ubiquitous in untranslated regions and protein-coding sequences (CDSs) of eukaryotes[17]. RG4 structures have been shown to perform many diverse and vital functions in a wide range of biological processes, such as pre-mRNA splicing[18], alternative polyadenylation[19], mRNA localization[20-21], microRNA targeting[22], and translational regulation[23-28]. Due to the presence of the 2α-hydroxyl property, the rG4 structure is more stable than DNA G-quadruplex[29]. In a recent study, Mirihana Arachchilage et al[30] have demonstrated that the most stable G4s appeared to be significantly under-represented within the CDS using specific synonymous codon combinations. However, several key problems regarding synonymous codon usage, rG4 formation, and evolution remain unaddressed.

In this study, the evolutionary choices of synonymous codons were evaluated around putative rG4 structures (pG4) at the whole transcriptome scale in multiple eukaryotic species. The searching algorithm G4Hunter was applied to detect rG4 structures in the protein-coding sequences of mRNAs in five eukaryotic species. Then, the native sequences forming rG4s were compared with randomized sequences to evaluate the selection of synonymous codons. Factors that may influence the formation of rG4 were also investigated, and the selection pressures of rG4 in different gene regions were compared to explore its potential roles in gene regulation. These analyses may help describe the evolutionary selections acting on synonymous codons in protein-coding regions.

1 Materials and Methods

1.1 Data

The nucleotide sequences and exonic structures of all

Received 2020-09-20, **Revised** 2021-04-26.

Biographies: Xu Yuming (1988—), male, Ph. D. candidate; Lu Zuhong (corresponding author), male, doctor, professor, zhlu@seu.edu.cn.

Foundation items: The National Key Research and Development Program of China (No. 2018YFC1314900, 2018YFC1314902), the National Natural Science Foundation of China (No. 61571109), the Fundamental Research Funds for the Central Universities (No. 2242017K3DN04).

Citation: Xu Yuming, Qi Ting, Gu Wanjun, et al. Evolutionary selection on synonymous codons in RNA G-quadruplex structural region[J]. Journal of Southeast University (English Edition), 2021, 37(2): 177 - 183. DOI: 10.3969/j.issn.1003-7985.2021.02.007.

protein-coding genes were downloaded in five eukaryotic species, including *H. sapiens* (GRCh38.p13), *M. musculus* (GRCg6a), *G. gallus* (GRCm38.p6), *D. rerio* (GRCz11), and *D. melanogaster* (BDGP6.28), using Ensembl BioMarts (release 97) [31]. Only protein-coding genes with a coding sequence of more than 150 nucleotides were included. Furthermore, miRNA target sites in the protein-coding regions of human and mouse genomes were downloaded from the miRDB database[32-33].

To explore the factors that affect the selection for rG4 structure formation, gene codon usage bias, nucleotide composition, and its evolutionary rate were considered. The effective number of codons (ENC) was used to measure codon usage bias, and ENC values for each gene were calculated as suggested by Wright[34]. A lower ENC value indicates stronger codon bias[34]. For each gene, we also calculated nucleotide compositions, including G and C contents. Moreover, the D_n and D_s values of all human and mouse orthologous genes were also retrieved from Ensembl BioMarts[35]. As a popular indicator of selection acting on protein-coding sequences, D_n/D_s quantifies the mode and strength of selection by comparing synonymous substitution rates (D_s) with nonsynonymous substitution rates (D_n)[36]. Generally, D_n/D_s close to one indicates neutrality; values greater than one are interpreted as positive selection (selection promoting change), and values less than one usually indicate purifying selection (selection suppressing protein change). Here, the ratio of D_n and D_s values for each gene are calculated and the value of D_n/D_s is used as the measurement of the evolutionary rate of genes in mice and humans.

1.2 rG4 structure in the protein-coding region

To locate rG4 structures in the protein-coding region, the G4Hunter algorithm[37] was exploited to systematically search for potential rG4 forming sites in the protein-coding sequences in all five species. G4Hunter considers G-richness and G-skewness of a given sequence and presents a quadruplex formation propensity score, G4Hscore, as output. G4HunterApps was run with a window size at 25 nt and a cutoff at 1.2[38] to identify all potential rG4 structures (pG4) in all protein-coding sequences. G4Hunter algorithm was chosen with these two parameters, since a comprehensive evaluation of computational methods for rG4 prediction has suggested that G4Hunter has the best performance in predicting G4 structures[39]. 65 562, 42 153, 38 311, 25 194, and 14 184 rG4 structures were identified in protein-coding sequences for *H. sapiens, M. musculus, G. gallus, D. rerio,* and *D. melanogaster*, respectively. To validate the observations of predicted rG4 structures, experimentally detected rG4 sites were also downloaded from the supplemental materials of Guo et al[40]. The experimental rG4 data were achieved by high throughput RT-stop techniques in human *HEK*293*T* cells and mouse *mESC* cells.

1.3 mRNA randomization

If the choice of synonymous codons influences the formation of rG4 structures in coding sequences, the G4Hscore of mRNA sequences in the naturally occurring pG4 region should be statistically different from that of randomized sequences. Thus, synonymous codons in the coding sequence were randomly shuffled, keeping the same amino acid sequence, GC composition, and codon usage bias. For each CDS sequence, the shuffling process was repeated 1 000 times to obtain a set of randomized artificial sequences. The G4Hscore of each 30 nt window in the native CDS sequence and each permutated sequence were calculated using the G4Hunter algorithm[37]. The difference of G4 forming potential between the native sequence and the randomized sequences was determined by calculating the *Z*-score of the G4Hscore (Z_{G4S}) for each sliding window using the following formula

$$Z_{G4S} = \frac{S_N - \overline{S_R}}{\sqrt{\frac{(S_{Ri} - \overline{S_R})^2}{n-1}}} \tag{1}$$

where S_N is the G4Hscore for the native sequence in the window; S_{Ri} is the G4Hscore of the corresponding window of the *i*-th randomized sequence; $\overline{S_R}$ is the mean of S_{Ri} overall randomized sequences; n is the total number of randomized sequences.

Similarly, the difference between the G (or C) compositions of the native sequence and randomized sequences was evaluated. The *Z*-score of the G content (Z_G) and that of the C content (Z_C) for each sliding window can be calculated as follows:

$$Z_G = \frac{G_N - \overline{G_P}}{\sqrt{\frac{(G_{Pi} - \overline{G_P})^2}{n-1}}} \tag{2}$$

$$Z_C = \frac{C_N - \overline{C_P}}{\sqrt{\frac{(C_{Pi} - \overline{C_P})^2}{n-1}}} \tag{3}$$

where G_N and C_N are the G content and the C content of the native sequence in the window; G_{Pi} and C_{Pi} are the G content and the C content of the corresponding window of the *i*-th randomized sequence, respectively; $\overline{G_P}$ and $\overline{C_P}$ are the means of G_{Pi} and C_{Pi} overall randomized sequences, respectively.

2 Results

2.1 Selection of rG4 structures on synonymous codons

The rG4 forming propensity score, G4Hscore[37], was calculated along the mRNA sequences with a sliding win-

dow scheme. For each pG4 structure in the transcriptome, a 30 nt window was moved both upstream and downstream from the start position of the pG4 structure with a 30 nt step, and the G4Hscore of the sequence in each of the 13 windows was calculated. To estimate the background distribution of the formation propensity of rG4 structures, the mRNA sequences were randomized by shuffling synonymous codons 1 000 times, and the G4Hscore in the corresponding sliding windows was calculated. The G4Hscore of the real mRNA sequence in a sliding window was compared with that of 1 000 corresponding sliding windows in the shuffled sequences. Z_{G4S} was calculated to assess the deviation of rG4 formation in the observed mRNA sequence from a random expectation. A positive Z_{G4S} value means synonymous codons are selected to facilitate the formation of rG4 structures, while a negative Z_{G4S} value indicates a selective pressure that prevents the formation of rG4 structures.

The sliding window analysis was performed in five eukaryotic species, including *H. sapiens*, *M. musculus*, *G. gallus*, *D. rerio*, and *D. melanogaster*. A significantly positive Z_{G4S} value was observed in the window of pG4 structures in all five species (see Fig. 1). When sliding windows move to the upstream or downstream of the pG4 structures, Z_{G4S} values drop quickly in the flank region of pG4 structures and oscillate around zero for sliding windows away from the pG4 structures (see Fig. 1). When in vitro experimentally identified rG4 structures in human *HEK*293*T* cells and mouse *mESC* cells were analyzed by the same procedure, a similar pattern of Z_{G4S} was found changing along the sliding windows (see Fig. 2).

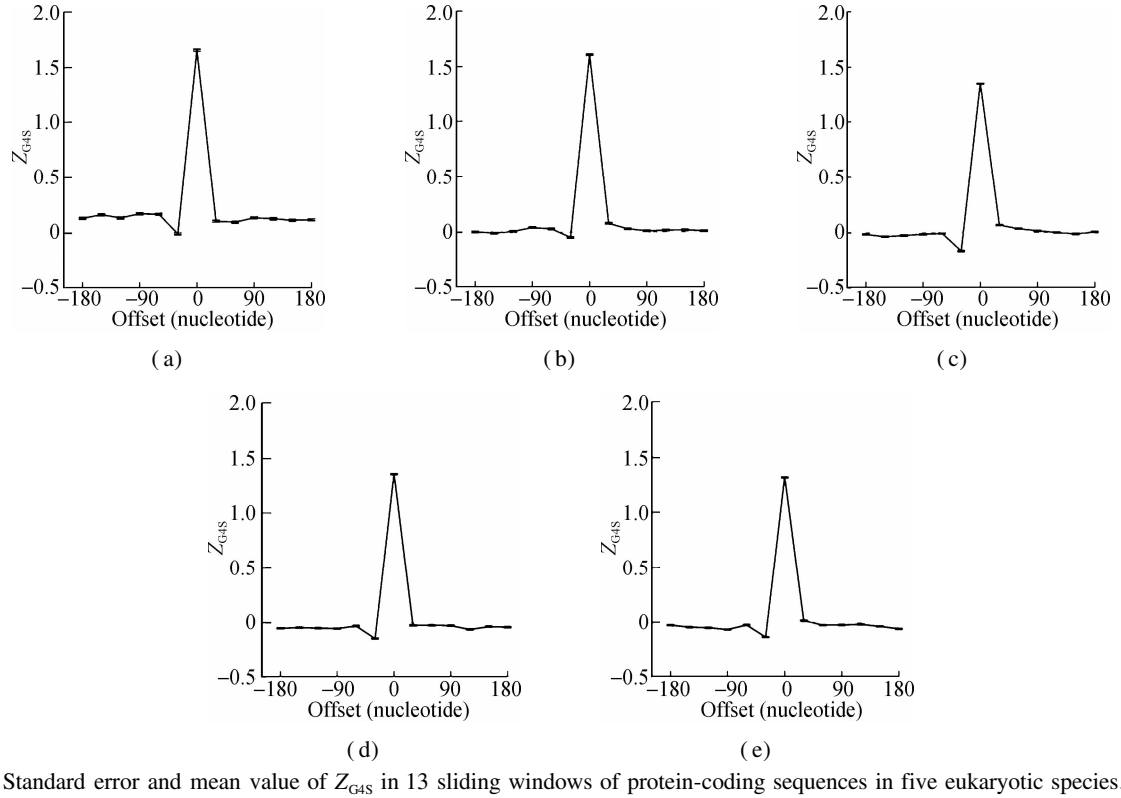

Fig. 1 Standard error and mean value of Z_{G4S} in 13 sliding windows of protein-coding sequences in five eukaryotic species. (a) *D. melanogaster*; (b) *D. rerio*; (c) *G. gallus*; (d) *M. musculus*; (e) *H. sapiens*

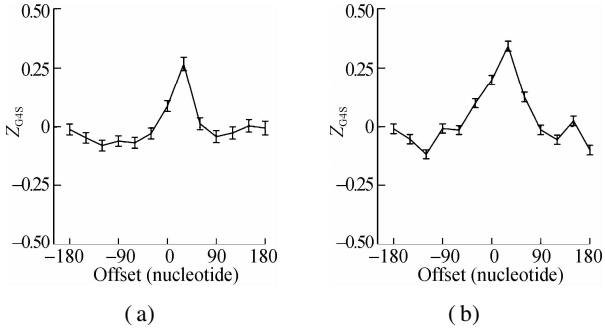

Fig. 2 The standard error and mean of Z_{G4S} of 13 sliding windows around pG4 regions in protein-coding sequences. (a) *HEK*293*T* cells; (b) *mESC* cells

2.2 Selection for rG4 formation on the GC content of codon combinations

To investigate how synonymous codons are selected for rG4 formation, Z_G and Z_C of a 30 nt window were calculated for each pG4 structure. Fig. 3 shows a significant positive correlation between Z_G and Z_{G4S} for pG4 structures in all five species. In comparison, a weaker but significant negative correlation between Z_C and Z_{G4S} of pG4 structures was also observed in all five species (see Fig. 4).

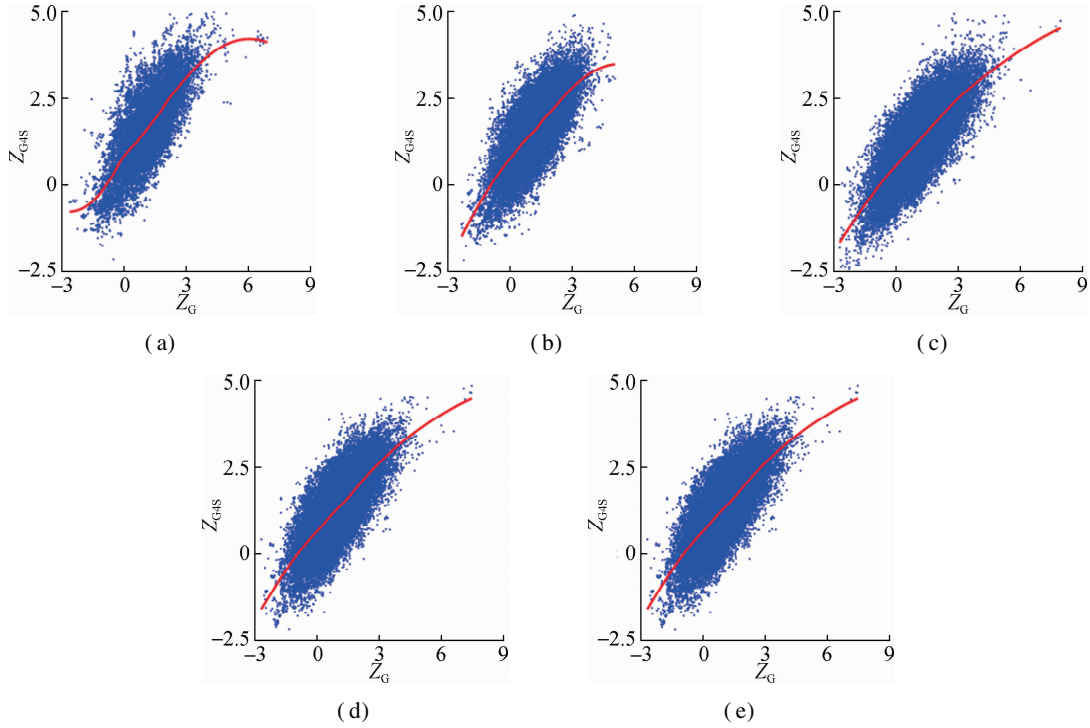

Fig. 3 The correlations between Z_{G4S} and Z_G for each pG4 structure in the protein-coding region. (a) *D. melanogaster*; (b) *D. rerio*; (c) *G. gallus*; (d) *M. musculus*; (e) *H. sapiens*

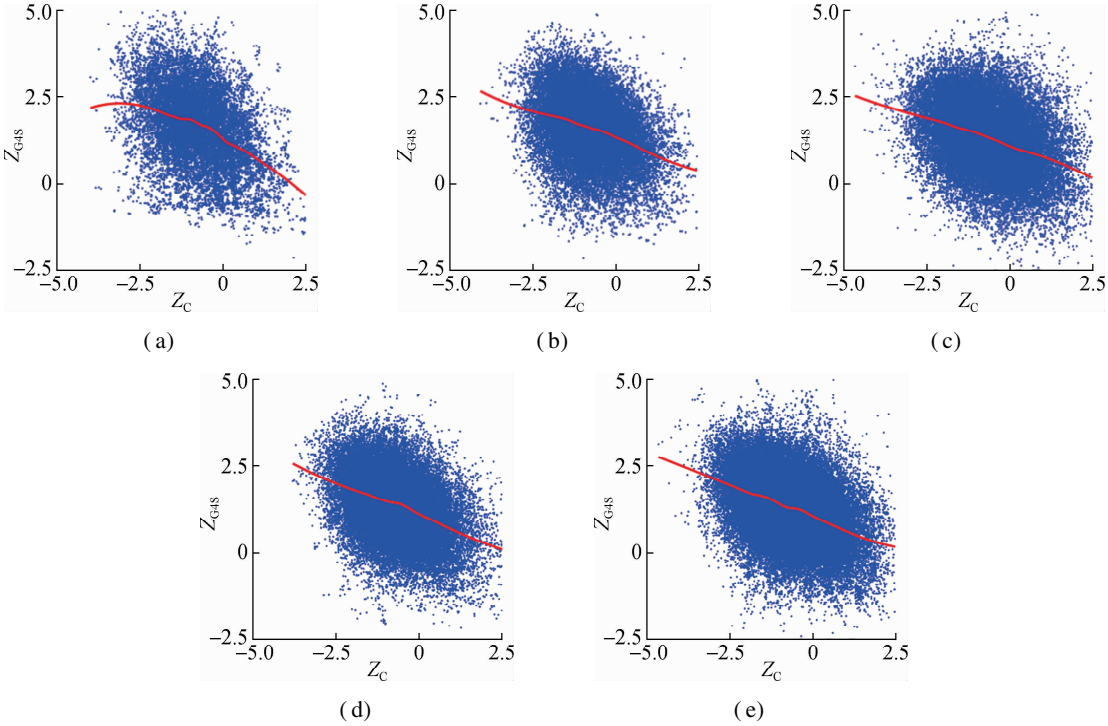

Fig. 4 The correlations between Z_{G4S} and Z_C for each pG4 structure in the protein-coding region. (a) *D. melanogaster*; (b) *D. rerio*; (c) *G. gallus*; (d) *M. musculus*; (e) *H. sapiens*

2.3 Features of the host gene that associate with rG4 selection

Although the mean value of Z_{G4S} is significantly larger than zero for all exonic pG4 structures in all five species (see Fig. 1), there are substantial variations among different pG4 structures within a single organism (see Figs. 3 and 4). To explore the factors that affect the selective pressures on synonymous codons for rG4 formation, several putative gene-level factors were considered, inclu-

ding the codon usage bias, evolutionary rate, and nucleotide compositions of the host gene where the pG4 structure is located. Analyses showed that Z_{G4S} values of pG4 structures in genes with the highest 5% ENC are significantly higher ($p = 1.2 \times 10^{-12}$ in humans and $p = 2.8 \times 10^{-11}$ in mice) than those genes with the lowest 5% ENC values (see Fig. 5(a)). For the top 5% genes with the highest D_n/D_s ratio, Z_{G4S} values of pG4 structures are significantly larger ($p = 4.1 \times 10^{-3}$ in humans and $p = 4.3 \times 10^{-6}$ in mice) than those pG4 structures in the bottom 5% genes with the lowest D_n/D_s ratio (see Fig. 5(b)). When pG4 structures in genes with the highest 5% G content are compared to those with the lowest 5% G content, it showed that pG4 structures located in genes with the highest 5% G content had significantly smaller ($p < 2 \times 10^{-16}$ in humans and $p < 2 \times 10^{-16}$ in mice) Z_{G4S} values (see Fig. 5(c)). Moreover, Z_{G4S} values of pG4 structures in genes with the top 5% C content are also statistically smaller ($p = 8.5 \times 10^{-14}$ in humans and $p = 1.9 \times 10^{-5}$ in mice) than those in genes with the bottom 5% C content.

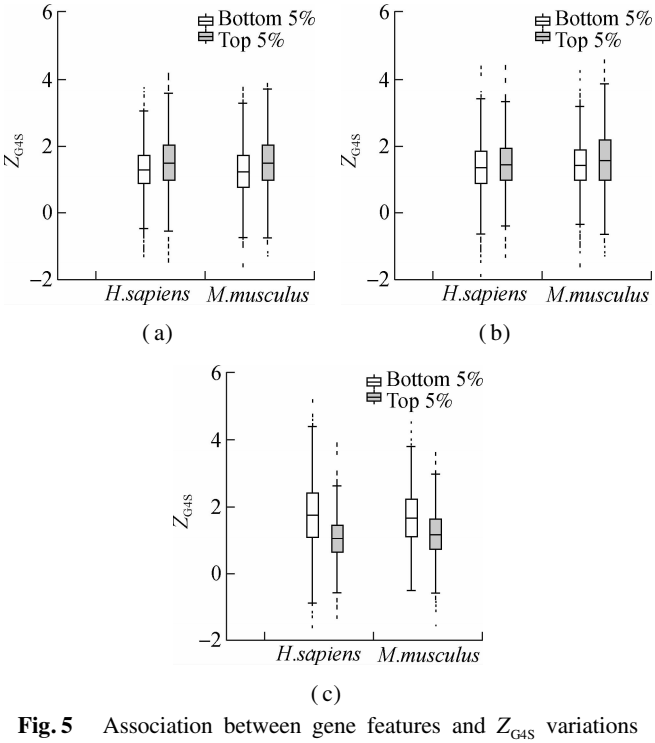

Fig. 5 Association between gene features and Z_{G4S} variations among rG4 structures of human and mice. (a) ENC value; (b) D_n/D_s value; (c) G content

2.4 Different selective pressure for rG4 structures in different gene regions

Other than the features of the host genes, Z_{G4S} differences of rG4 structures in different gene regions were also evaluated. First, pG4 structures were grouped into two categories: those in the translation initiation region (within 70 nt downstream of the start codon) and those out of the translation initiation region. The results showed that pG4 structures in the translation initiation region had significantly higher Z_{G4S} values than those in the translation elongation region (see Fig. 6). Next, Z_{G4S} values of pG4 structures near exonic splicing sites (within 60 nt of splicing sites) were compared with those of pG4s distant from the splicing sites. Although Z_{G4S} values for pG4 structures near the splicing sites tend to be smaller than those distant from the splicing sites (data not shown), the differences are not statistically significant for rG4 structures in both downstream and upstream flank regions of splicing sites. Finally, Z_{G4S} values of pG4 structures in microRNA (miRNA) target sites were compared with Z_{G4S} values of pG4s out of miRNA target region. No significant differences were observed between Z_{G4S} values of pG4 structures in miRNA target region and those out of miRNA target sites.

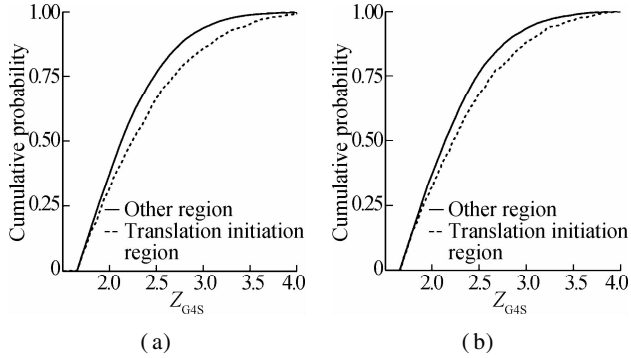

Fig. 6 Cumulative probability curves of Z_{G4S} values of rG4 structures in the translation initiation region and the translation elongation region. (a) *H. sapiens*; (b) *M. musculus*

3 Conclusions

1) Synonymous codons are universally selected for the formation of rG4 structures in the protein-coding sequences of the five eukaryotic organisms under investigation. While G-rich codon combinations are preferred in the rG4 structural region, C-rich codon combinations are selectively unfavorable for rG4 formation.

2) The selective pressures acting on synonymous codons are stronger for pG4 structures in genes with lower G nucleotides, less biased usage of synonymous codons, and a lower evolutionary rate. These differences are consistent for pG4 structures in humans and mice.

3) Synonymous codons in some specific gene regions, such as the translation initiation region, were under different selective pressures for the rG4 formation. However, rG4 structures in the miRNA target region and splicing sites flank region did not show obviously different evolutionary selections on synonymous codons.

References

[1] Shabalina S A, Spiridonov N A, Kashina A. Sounds of

silence: Synonymous nucleotides as a key to biological regulation and complexity[J]. *Nucleic Acids Research*, 2013, **41**(4): 2073−2094. DOI: 10.1093/nar/gks1205.

[2] Plotkin J B, Kudla G. Synonymous but not the same: The causes and consequences of codon bias[J]. *Nature Reviews Genetics*, 2011, **12**(1): 32−42. DOI: 10.1038/nrg2899.

[3] Hunt R C, Simhadri V L, Iandoli M, et al. Exposing synonymous mutations[J]. *Trends in Genetics*, 2014, **30**(7): 308−321. DOI: 10.1016/j.tig.2014.04.006.

[4] Chamary J V, Hurst L D. Biased codon usage near intron-exon junctions: Selection on splicing enhancers, splice-site recognition or something else? [J]. *Trends in Genetics*, 2005, **21**(5): 256−259. DOI: 10.1016/j.tig.2005.03.001.

[5] Stoletzki N. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures[J]. *BMC Evolutionary Biology*, 2008, **8**(1): 1−9. DOI: 10.1186/1471-2148−8-224.

[6] Warnecke T, Batada N N, Hurst L D. The impact of the nucleosome code on protein-coding sequence evolution in yeast[J]. *PLoS Genetics*, 2008, **4**(11): e1000250. DOI: 10.1371/journal.pgen.1000250.

[7] Parmley J L, Chamary J V, Hurst L D. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers[J]. *Molecular Biology and Evolution*, 2006, **23**(2): 301−309. DOI: 10.1093/molbev/msj035.

[8] Warnecke T, Hurst L D. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in drosophila melanogaster[J]. *Molecular Biology and Evolution*, 2007, **24**(12): 2755−2762. DOI: 10.1093/molbev/msm210.

[9] Gu W J, Wang X F, Zhai C Y, et al. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants[J]. *Molecular Biology and Evolution*, 2012, **29**(10): 3037−3044. DOI: 10.1093/molbev/mss109.

[10] Kudla G, Murray A W, Tollervey D, et al. Coding-sequence determinants of gene expression in escherichia coli[J]. *Science*, 2009, **324**(5924): 255−258. DOI: 10.1126/science.1170160.

[11] Gu W J, Zhou T, Wilke C O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes[J]. *PLoS Computational Biology*, 2010, **6**(2): e1000664. DOI: 10.1371/journal.pcbi.1000664.

[12] Tuller T, Waldman Y Y, Kupiec M, et al. Translation efficiency is determined by both codon bias and folding energy[J]. *PNAS*, 2010, **107**(8): 3645−3650. DOI: 10.1073/pnas.0909910107.

[13] Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy[J]. *Molecular Systems Biology*, 2011, **7**: 481. DOI: 10.1038/msb.2011.14.

[14] Thanaraj T A, Argos P. Ribosome-mediated translational pause and protein domain organization[J]. *Protein Science*, 1996, **5**(8): 1594−1612. DOI: 10.1002/pro.5560050814.

[15] Komar A A, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation[J]. *FEBS Letters*, 1999, **462**(3): 387−391. DOI: 10.1016/s0014-5793(99)01566-5.

[16] Kharel P, Balaratnam S, Beals N, et al. The role of RNA G-quadruplexes in human diseases and therapeutic strategies[J]. *Wiley Interdisciplinary Reviews RNA*, 2020, **11**(1): e1568. DOI: 10.1002/wrna.1568.

[17] Kwok C K, Marsico G, Sahakyan A B, et al. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome[J]. *Nature Methods*, 2016, **13**(10): 841−844. DOI: 10.1038/nmeth.3965.

[18] Huang H L, Zhang J, Harvey S E, et al. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF[J]. *Genes & Development*, 2017, **31**(22): 2296−2309. DOI: 10.1101/gad.305862.117.

[19] Beaudoin J D, Perreault J P. Exploring mRNA 3′-UTR G-quadruplexes: Evidence of roles in both alternative polyadenylation and mRNA shortening[J]. *Nucleic Acids Research*, 2013, **41**(11): 5898−5911. DOI: 10.1093/nar/gkt265.

[20] Subramanian M, Rage F, Tabet R, et al. G-quadruplex RNA structure as a signal for neurite mRNA targeting[J]. *EMBO Reports*, 2011, **12**(7): 697−704. DOI: 10.1038/embor.2011.76.

[21] Kanai Y, Dohmae N, Hirokawa N. Kinesin transports RNA: Isolation and characterization of an RNA-transporting granule[J]. *Neuron*, 2004, **43**(4): 513−525. DOI: 10.1016/j.neuron.2004.07.022.

[22] Stefanovic S, Bassell G J, Mihailescu M R. G quadruplex RNA structures in PSD-95 mRNA: Potential regulators of miR-125a seed binding site accessibility[J]. *RNA*, 2015, **21**(1): 48−60. DOI: 10.1261/rna.046722.114.

[23] Beaudoin J D, Perreault J P. 5′-UTR G-quadruplex structures acting as translational repressors[J]. *Nucleic Acids Research*, 2010, **38**(20): 7022−7036. DOI: 10.1093/nar/gkq557.

[24] Bugaut A, Balasubramanian S. 5′-UTR RNA G-quadruplexes: Translation regulation and targeting[J]. *Nucleic Acids Research*, 2012, **40**(11): 4727−4741. DOI: 10.1093/nar/gks068.

[25] Kamura T, Katsuda Y, Kitamura Y, et al. G-quadruplexes in mRNA: A key structure for biological function[J]. *Biochemical and Biophysical Research Communications*, 2020, **526**(1): 261−266. DOI: 10.1016/j.bbrc.2020.02.168.

[26] Kumari S, Bugaut A, Huppert J L, et al. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation[J]. *Nature Chemical Biology*, 2007, **3**(4): 218−221. DOI: 10.1038/nchembio864.

[27] Murat P, Marsico G, Herdy B, et al. RNA G-quadruplexes at upstream open reading frames cause DHX36-and DHX9-dependent translation of human mRNAs[J]. *Genome Biology*, 2018, **19**(1): 229. DOI: 10.1186/s13059-018−1602-2.

[28] Simone R, Fratta P, Neidle S, et al. G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome[J]. *FEBS Letters*, 2015, **589**(14): 1653−1668. DOI: 10.1016/j.febslet.2015.05.003.

[29] Fay M M, Lyons S M, Ivanov P. RNA G-quadruplexes in biology: Principles and molecular mechanisms[J].

Journal of Molecular Biology, 2017, **429**(14): 2127 - 2147. DOI: 10.1016/j.jmb.2017.05.017.

[30] Mirihana Arachchilage G, Hetti Arachchilage M, Venkataraman A, et al. Stable G-quadruplex enabling sequences are selected against by the context-dependent codon bias [J]. *Gene*, 2019, **696**: 149 - 161. DOI: 10.1016/j.gene.2019.02.006.

[31] Kinsella R J, Kähäri A, Haider S, et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space[J]. *Database*, 2011, **2011**(10.1093): database. DOI: 10.1093/database/bar030.

[32] Chen Y H, Wang X W. miRDB: An online database for prediction of functional microRNA targets [J]. *Nucleic Acids Research*, 2020, **48**(D1): D127 - D131. DOI: 10.1093/nar/gkz757.

[33] Liu W J, Wang X W. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data[J]. *Genome Biology*, 2019, **20**(1): 18. DOI: 10.1186/s13059-019 - 1629-z.

[34] Wright F. The 'effective number of codons' used in a gene[J]. *Gene*, 1990, **87**(1): 23 - 29. DOI: 10.1016/0378-1119(90)90491-9.

[36] Nielsen R. Molecular signatures of natural selection[J]. *Annual Review of Genetics*, 2005, **39**: 197 - 218. DOI: 10.1146/annurev.genet.39.073003.112420.

[37] Bedrat A, Lacroix L, Mergny J L. Re-evaluation of G-quadruplex propensity with G4Hunter[J]. *Nucleic Acids Research*, 2016, **44**(4): 1746 - 1759. DOI: 10.1093/nar/gkw006.

[38] Lacroix L. G4HunterApps[J]. *Bioinformatics*, 2019, **35**(13): 2311 - 2312. DOI: 10.1093/bioinformatics/bty951.

[39] Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction[J]. *Nucleic Acids Research*, 2020, **48**(1): 1 - 15. DOI: 10.1093/nar/gkz1097.

[40] Guo J U, Bartel D P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria[J]. *Science*, 2016, **353**(6306): 5371. DOI: 10.1126/science.aaf5371.

RNA G-四联体结构区的同义密码子进化选择

许育铭　齐　婷　顾万君　陆祖宏

(东南大学生物科学与医学工程学院,南京 210096)
(东南大学生物电子学国家重点实验室,南京 210096)

摘要:为探索同义密码子对 RNA G-四联体(rG4)形成的适应效应,利用 G4Hunter 算法搜寻 5 类真核生物 mRNA 蛋白编码区的 rG4 结构.通过对比天然序列和随机序列,评估 rG4 结构对同义密码子的选择偏好,分析影响 rG4 形成的因素,检验基因特定区域 rG4 结构的选择压力,以揭示 rG4 在基因调控中的潜在作用.结果表明,在 5 类真核生物中,rG4 结构区内的同义密码子普遍受到促使 rG4 形成的选择压力.在 rG4 结构区内的蛋白编码序列倾向于使用富含鸟嘌呤的密码子组合,不倾向于选择富含胞嘧啶的密码子组合.密码子使用偏好、碱基成分以及进化速率等因素和 rG4 对同义密码子的选择压力具有相关性.基因翻译起始区的序列受到更强的促使 rG4 结构形成的选择压力.

关键词:RNA 结构;G-四联体;同义密码子;进化;选择

中图分类号:Q311