Quasiperiodic property in Alu repeats

Shihua Zhang and Yi Xiao*

Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

(Received 21 December 2005; revised manuscript received 6 April 2006; published 7 August 2006)

We investigate the possible quasiperiodic property in the sequences of Alu repeats, one of typical noncoding DNA sequences. We calculated the quasiperiods of the right and left monomers of Alu repeats of different families with quasiperiodic matrix algorithm. It is interesting that the right monomers of all families show significant quasiperiod 8 in their sequences while the left monomers show quasiperiods 8 or 5. Our results indicate that there exist common quasiperiods in most Alu repeats. This may be helpful to further explore possible functions of Alu repeats.

DOI: 10.1103/PhysRevE.74.022901 PACS number(s): 87.14.Gg, 87.10.+e

I. INTRODUCTION

Up to now, only 3%–5%, i.e., the coding DNA sequences (genes), of the whole human genome [1–3] are well understood. In other words, there are 95%–97% noncoding sequences (called "Junk" DNA in the past) which are a genuine challenge for us to find out their roles. Among them, there exists a mass of repeated sequences, which can simply be divided into two kinds: consecutive and interspersed repeated sequences. Alu sequences we are discussing in this paper belong to middle repetitive short interspersed nuclei elements, for short SINEs. SINE's appear mainly in primates [4] and account for more than 10% of the entire human genome. There is also about 1 200 000 Alu copies [5], which makes us unavoidably to suppose the potential functions of Alu sequences.

Alu sequences were named after the AluI restriction enzyme site within the consensus Alu sequence [6] (Fig. 1). They are ancestrally derived from the gene specifying 7SL RNA, an abundant cytoplasmic component of the signal recognition particle (SRP [7–11]). About 90% sequence similarity exists between Alu sequences and 7SL RNA. Figure 2 show us a typical Alu sequence, of about 300 nucleotides in length. At the AluI restriction enzyme site, there is a recognizing sequence **agct**$^{(4)}$. Alu sequences have a dimeric structure and are consist of two similar, but distinct monomers linked by an *A*-rich *region*$^{(3)}$. The right Alu monomer contains a 30 bp insert absent from the left monomer. A functional two box (*A* and *B*) RNA polymerase (pol III) promoter is present in the left monomer, but is absent from the right monomer [12]. The box $B^{(2)}$ is necessary to transcription [13] and the box $A^{(1)}$ decides the length and precision of transcription. It is a common character of all the SINEs that they have a *A*-rich *region*$^{(3)}$ and a Poly(*A*) *tail*$^{(5)}$. It is supposed that *A*-rich *region* may affect the distribution of Alu sequences in new chromosome site.

In the genome of the primates, Alu sequences could increase to such a high copy number in the 65 000 000 years [14], so what people care most is whether Alu sequences have certain functions. In the past, Alu sequences have been considered as either "junk," "parasitic," or "selfish" DNA that serve no useful functions [15]. Conversely, recent researches reveal that Alu sequences may have one or several functions: these interspersed Alu sequences show greatly distinct and especially exceptional properties. For instance, the majority of the Alu sequences exist in intron sequences or among the genes. This makes people speculate that Alu sequences may have a certain signal regulating function. At one time, Alu sequences also take on quite high polymorphism [16,17] in the Alu colony. Some polymorphism sites merely exist in a certain kindred. It is supposed [18] that Alu sequences may be related to the regulation of gene transcription and the processing of hnRNA, so they may affect the protein synthesis at a certain extent. All these make scientists suppose that Alu sequences may have important relations with network regulation, namely, Alu sequences may regulate the genes cooperative expressions that are interspersed in genome.

DNA sequences with different functions have different features. For example, exon sequences show a quasiperiod of three nucleic acids. So investigating the quasiperiods of Alu sequences is very helpful to investigate their possible functions [19,20]. In this paper, we investigate the quasiperiodic property of Alu sequences with quasiperiodic matrix algorithm. Through calculating, we found that the two arms of Alu sequences have different quasiperiods. The quasiperiod of the right monomers are the same for all Alu subfamilies while the left monomers also mainly have significant quasiperiod 8 or/and 5.

II. QUASIPERIOD METHOD

There have been many methods [21–27] for investigating the periodicity of DNA sequences, in which we especially mention those developed by Pizzi *et al.* [21]. Pizzi *et al.* proposed an "enhance" algorithm for automatic detection of serial periodicities in a linear sequence. They successfully

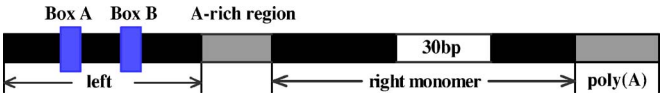

FIG. 1. (Color online) The basic structure of the consensus Alu sequences.

*Corresponding author.

ggcc**gggcgcggtgg**(1)ctcacgcctgtaatcccagcactttgggaggccgaggcgggcggatcacgaggtcagga**g**
atcgagacc(2)atcccggctaaaacggtgaaaccccgtctctact**aaaaatacaaaaaa**(3)ttagccgggcgtagtggc
gggcgcctgtagtccc**agct**(4)acttgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtg
agccgagatcccgccactgcactccagcctgggcgacagagcgagactccgtctc**aaaaaaaaa**(5)

FIG. 2. A typical Alu sequence: (1) Box *A*, (2) Box *B*, (3) *A*-rich region, (4) tAluI restriction enzyme site, and (5) poly(*A*) tail.

revealed the repetitive patterns containing hierarchical periodicities in DNA subtelomeric sequences from *S.cerevisiae* and *P.falciparum*. Their analysis of short-range periodicities also suggested plausible models of pattern evolution from an original array of short repeat units. This is an ideal method for detecting quasi or latent periodicities in DNA sequences, especially those not easily recognizable by other methods such as DFT. To apply this method, the sequences must have sufficient length to make calculated results statistically significant. This is not satisfied by Alu elements of 300 nt. The DFT also cannot recognize the quasiperiods in Alus easily [19]. Therefore, we shall use the quasiperiod method.

The quasiperiod method used here is as follows. For a Alu sequence $S=s_1s_2,\ldots,s_n$, we use a periodic sequence $P=p_1p_2,\ldots,p_n$ to approach it, where P meets that $p_i=p_{i+T}$ $(i=1,2,\ldots,n-T+1)$, i.e., its period is T. For example, sequence $ACACAC$ has a period $T=2$.

The distance between P and S is defined as $H=\Sigma_{i=1}^{n}h(s_i,p_i)$, where $h(s_i,p_i)$ is the closeness of a pair of symbols s_i and p_i which is measured in a Hamming metric [28]

$$h(s_i,p_i)=\begin{cases}1 & s_i\neq p_i\\ 0 & s_i=p_i.\end{cases}$$

Furthermore, we can define average distance between the original and periodic sequences: $D=H/n$, which does not depend on the length of the sequence. Apparently, the average distance is different by using different periodic sequence to approach the original sequence. So we define the quasiperiod of the original sequence as the period of the periodic sequence when their average distance is the smallest. In principle, we can find the quasiperiod of the original sequence by using different periodic sequences to approach it. But in practice this routine is very time consuming. Therefore, in our calculation we shall use a fast matrix algorithm. Figure 2 illustrates the basic idea: assuming the original sequence is $S=ACATAG$ and the period of the sequence used to align with S is $T=2$.

In Fig. 3, $A1$ stands for the number of the base A on the position 1 in the repeat unit, $A2$ stands for the number of the base A on the position 2, and so on. Thus we can construct a matrix: the rows of the matrix denote four kinds of nucleotides, the columns denote the positions in the repeated unit (the size of the repeat units is $T=2$). For example, the matrix element $(A,1)$ denote the number $A1$ that the base A appear on the position 1 in the repeat unit. Then we can get the minimal average distance between the original sequence and the periodic sequence with $T=2$: $D_{2\,\min}=(n-N_2)/n$, where N_2 is the sum of the largest numbers in each column. Using the same method, we can get a series of minimal average distances corresponding to different periodic sequences with

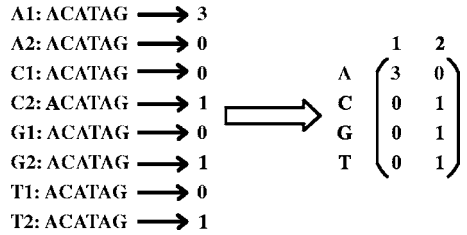

FIG. 3. The sketch of the quasiperiod matrix algorithm.

$T=3,4,\ldots,m:\{D_{2\,\min},D_{3\,\min},\ldots,D_{m\,\min}\}$. The period of the periodic sequence corresponding to the smallest in $\{D_{2\,\min},D_{3\,\min},\ldots,D_{m\,\min}\}$ is just the quasiperiod of the original sequence.

It is noted that the distance between the original and periodic sequences depends also on the period T. The larger the valuee of the period T is and the closer the distance between the original and periodic sequences. Therefore, we introduce a parameter α to eliminate this effect and redefine the average distance between the original and periodic sequences as $D_m=(n-S_m+\alpha T)/n$. To determine the value of α, we calculated the quasiperiod of exon sequences which should have a quasiperiod of 3 bases. We found that this can be achieved when we chose the parameter $\alpha=2.0$.

III. RESULTS AND DISCUSSIONS

A. The quasiperiods of exon and intron sequences

We have collected 203 exon sequences with lengths of 290–300 bp from primates and calculated their quasiperiods. The results show that most of exon sequences have a period of 3 or the multiple of 3 such as 6 and 9 [Fig. 4(a)]. The sequence number with the periods of 3, 6, and 9 accounts for about 85% of the total sequences. Therefore, exon sequences show significant period 3, which accords best with their triple coding property.

On the other hand, we also calculated the quasiperiods of 180 intron sequences with the same method and did not find any distinct signal of quasiperiods in intron sequences [Fig. 4(b)]. Therefore, the introns do not show regular patterns in their sequences. Of course, this may correspond to the intrinsic property of the intron sequences: the intron sequences do not code proteins. These results justify the quasiperiod method.

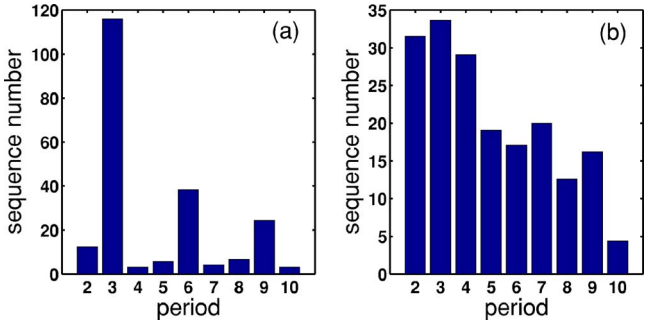

FIG. 4. (Color online) The number of sequences versus period: (a) exon sequences and (b) intron sequences.

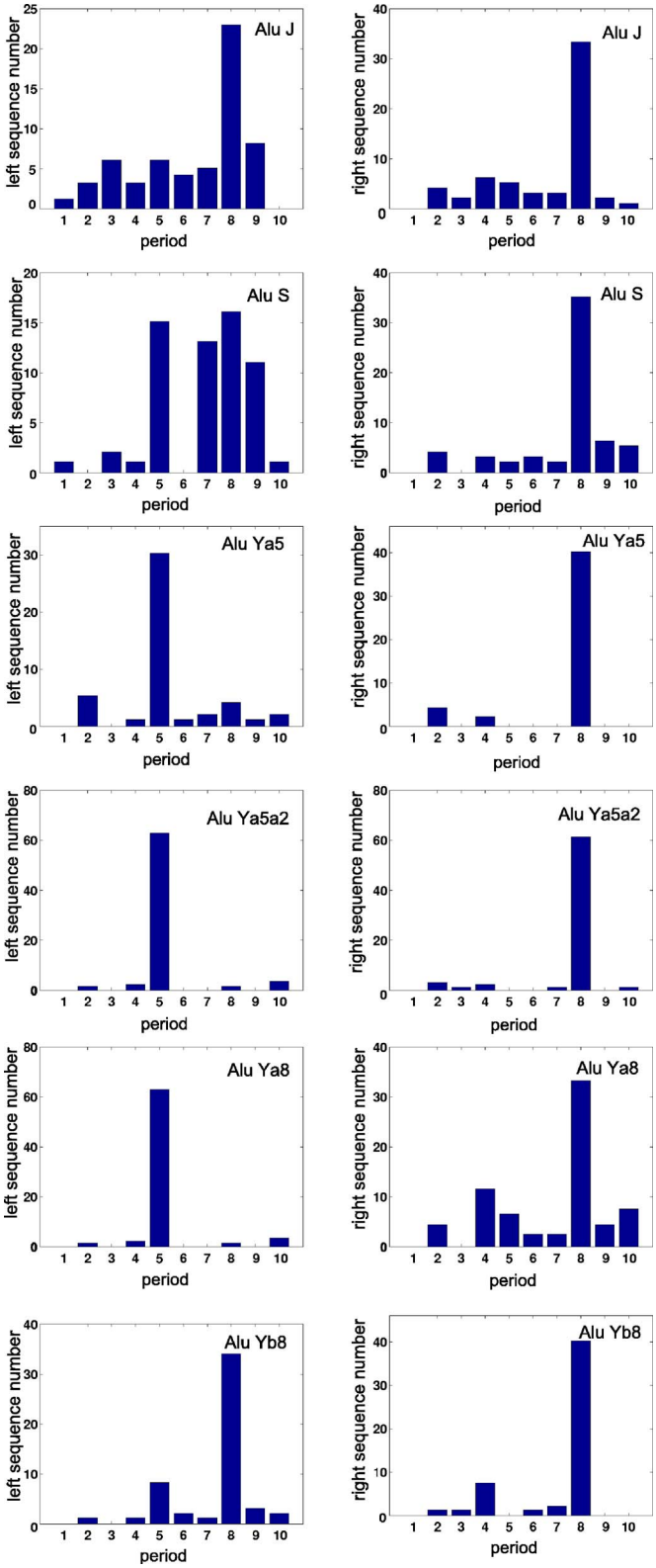

FIG. 5. (Color online) Left and right monomers number versus quasiperiod for old, intermediate, and young Alu*s*.

B. The quasiperiods of the two arms of Alu sequences

Statistical analysis has identified key diagnostic nucleotide positions in Alu sequences that define 12 subfamilies [29]. Alu sequences are generally divided into oldest, intermediate, and youngest. The oldest contains Alu *J*, the intermediate contains Alu *S*, the youngest contains AluYa5, AluYa5a2, AluYa8, and AluYb8. Phylogenic studies indicate that primary Alu sequences may act as the earliest templates for Alu origination. Since there is so many mutations [30] during the evolution of Alu sequences, it is interesting to investigate whether there exist significant quasiperiodic property of Alu sequences and find their possible certain biological functions. For each subfamily, we collected 64 Alu *J*, 64 Alu *S*, 55 AluYa5, 69 AluYa8, 69 AluYa5a2, and 63 AluYb8 from GeneBank.

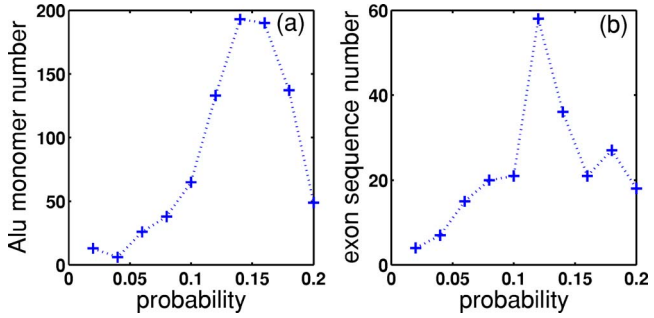

FIG. 6. (Color online) The distribution of the probabilities of occurrence of the quasiperiods of Alu monomers (a) and exon sequences (b) expected from their randomly shuffled sequences.

Alu sequences consist of two monomers. The left monomers are 116–122 bp in length and the right monomers 140–146 bp. We calculated their quasiperiods of the two monomers of each subfamily respectively. In our calculations, we removed the poly(*A*) tails of all Alu sequences.

Figure 5 shows that in all the Alu subfamilies the quasiperiods of most right monomers are 8. For the left monomers, the situation is different. For young Alu*s*, three subfamilies (AluYa5a2, AluYa5, *and* AluYa8) have quasiperiod 5 for most left monomers and only one subfamily (AluYb8) has quasiperiod 8. However, AluYa8 have both quasiperiods 5 and 8. The quasiperiods of the left monomer of Alu *S* also have both quasiperiods 5 and 8. Although the sequence numbers with quasiperiods of 7 and 9 are also comparable, we classify them into quasiperiod 8 for simplicity. Finally, the oldest Alu *J* only has a quasiperiod 8 for the left monomers as the right monomers.

To show the statistical significance of the quasiperiods obtained for Alu*s*, we also calculated the probability of the

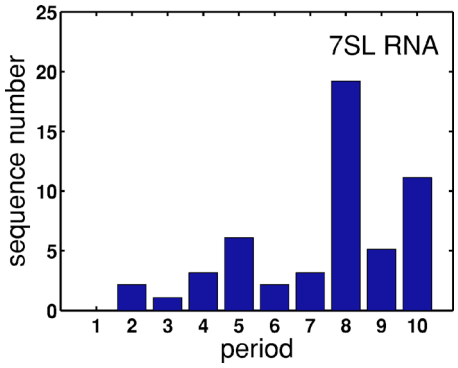

FIG. 7. (Color online) The number against quasiperiods of 7SL RNA sequences

occurrence of the quasiperiod of each Alu monomer expected from randomly shuffled sequences. In our calculation, 1000 randomly shuffled sequences are generated for each Alu repeat and their quasiperiods are calculated. Then we calculate the percentage (i.e., the probability of occurrence) of the number of the shuffled Alus with the same quasiperiod as the original one. Figure 6(a) shows the distribution of the probabilities of occurrence for all the Alu*s* used in this work. It indicates the probabilities of the occurrence of the quasiperiods of most Alu monomers are smaller than 0.14, i.e., the statistical significance of them is higher than 86%. This seems lower than the usual standard, e.g., 95%. To indicate whether this is acceptable, we also calculate the distribution of the probabilities of occurrence of the quasiperiod 3 of each exon in its 1000 randomly shuffled ones. We found the results are similar to that of Alu repeats [Fig. 6(b)]. This indicates that the statistical significance of the quasiperiods of exons and Alu*s* is at the same level. Therefore, we think the statistical significance of the quasiperiods of Alus is meaningful as comparing with that of exons.

The results above are very interesting. First, the quasiperiods of the right monomers have been kept unchanged from the youngest to oldest Alus, although they have undergone many mutations. This implies that the quasiperiod 8 of the right monomers are significant, comparing with significant quasiperiod 3 [Fig. 4(a)] feature in exon sequences (they also have specific function), we believe the right monomers may be related to certain function. For the left monomers, it seems that either they tend to keep the quasiperiod 8 or 5. This also suggests that the left monomers may have some functions.

To understand the quasiperiods 5 and 8 in Alu*s*, we also collected 60 representative 7SL RNA sequences. The calculated quasiperiods of them are shown in Fig. 7. It is clear that most of them have a quasiperiod 8. However, there are also many sequences have quasiperiods 5 and 10. Thus, it is not surprising that Alus have both quasiperiods 5 and 8 because they were much closer to 7SL RNA.

In summary, most Alu sequences have distinct quasiperiods 8 or 5. Our results may be helpful to further explore possible functions of Alu sequences.

ACKNOWLEDGMENTS

This work is supported by the NSFC under Grants No. 30525037 and No. 30470412 and the Foundation of the Ministry of Education of China.

[1] A. M. Roy *et al.*, Genome Res. **10**, 1485 (2000).
[2] A. M. Weiner, P. L. Deininger, and A. Efstratiadis, Annu. Rev. Biochem. **55**, 631 (1986).
[3] N. Okada, Trends Ecol. Evol. **6**, 358 (1991).
[4] Y. Quentin, Nucleic Acids Res. **22**, 2222 (1994).
[5] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, and J. Baldwin, Nature (London) **409**, 860 (2001).
[6] C. M. Houck, F. P. Rinehart, and C. W. Schmid, J. Mol. Biol. **132**, 289 (1979).
[7] D. Y. Chang, K. Hsu, and R. J. Maraia, Nucleic Acids Res. **24**, 4165 (1996).
[8] Y. Chiang and J. K. Vishwanatha, Mol. Cell. Biochem. **155**, 131 (1996).
[9] F. Bovia, N. Wolff, S. Ryser, and K. Strub, Nucleic Acids Res. **25**, 318 (1997).
[10] J. Sarrowa, D. Y. Chang, and R. J. Maraia, Mol. Cell. Biol. **17**, 1144 (1997).
[11] D. Labuda and E. Zietkiewiez, J. Mol. Evol. **39**, 506 (1994).
[12] S. A. Fuhrman, P. L. Deininger, P. Laporte, T. Friedmann, and E. P. Geiduschek, Nucleic Acids Res. **9**, 6439 (1981).
[13] B. Panning and J. R. Smiley, Mol. Cell. Biol. **13**, 3231 (1993).
[14] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[15] W. F. Doolittle and C. Sapienza, Nature (London) **284**, 601 (1980).
[16] L. B. Jorde *et al.*, Am. J. Hum. Genet. **66**, 979 (2000).
[17] W. S. Watkins *et al.*, Am. J. Hum. Genet. **68**, 738 (2001).
[18] A. J. Mighell and A. F. Markham, FEBS Lett. **417**, 1 (1997).
[19] N. Wang, Information analysis of DNA noncoding region sequences and comparative research for full genome, Doctor Dissertation, Institute of Biophysics, Chinese Academy of Sciences, 1999.
[20] Y. Xiao, Y. Z. Huang, M. F. Li, R. Z. Xu, and S. F. Xiao, Phys. Rev. E **68**, 061913 (2003).
[21] E. Pizzi, S. Liuni, and C. Frontali, Nucleic Acids Res. **18**, 3745 (1990).
[22] E. V. Korotkov and D. A. Phoenix, Pac. Symp. Biocomput **222**, 31 (1997).
[23] M. B. Chaley, E. V. Korotkov, and K. G. Skryabin, DNA Res. **6**, 153 (1999).
[24] B. Gary, Nucleic Acids Res. **27**, 573 (1999).
[25] S. Kurtz and C. Schleiermacher, Bioinformatics **15**, 426 (1999).
[26] A. Lefebvre *et al.*, Bioinformatics **19**, 319 (2003).
[27] Y. Wexler, Z. Yakhini *et al.*, J. Comput. Biol. **12**, 7 (2005).
[28] Y. Z. Huang and Y. Xiao, Chin. Phys. Lett. **19**, 434 (2002).
[29] M. A. Batzer *et al.*, J. Mol. Evol. **42**, 3 (1996).
[30] P. L. Deininger and M. A. Batzer, Evol. Biol. **27**, 157 (1993).