Nonlinear analysis of correlations in Alu repeat sequences in DNA

Yi Xiao, Yanzhao Huang, Mingfeng Li, Ruizhen Xu, and Saifeng Xiao
Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
(Received 11 June 2003; published 24 December 2003)

We report on a nonlinear analysis of deterministic structures in Alu repeats, one of the richest repetitive DNA sequences in the human genome. Alu repeats contain the recognition sites for the restriction endonuclease AluI, which is what gives them their name. Using the nonlinear prediction method developed in chaos theory, we find that all Alu repeats have novel deterministic structures and show strong nonlinear correlations that are absent from exon and intron sequences. Furthermore, the deterministic structures of Alus of younger subfamilies show panlike shapes. As young Alus can be seen as mutation free copies from the ''master genes,'' it may be suggested that the deterministic structures of the older subfamilies are results of an evolution from a ''panlike'' structure to a more diffuse correlation pattern due to mutation.

DOI: 10.1103/PhysRevE.68.061913 PACS number(s): 87.14.Gg, 87.10.+e, 05.45.−a

I. INTRODUCTION

Correlation properties of biosequences (DNA and proteins) may indicate their biological functions. A well-known example is the period 3 in coding DNA sequences [1,2]. A linear correlation analysis showed that the power spectrum of the coding sequences shows a strong signal in period 3. This is obviously related to the triplet genetic codes which compose the coding sequences.

DNA consists of coding (exons) and noncoding parts (e.g., the intergenic regions and introns). As the whole DNA sequences of some organisms have been measured, it is clear now that the coding sequences only consist of about 5% of the whole human genome and 95% are noncoding sequences. We know much more about the organization and functions of the coding sequences, but we know little about the noncoding parts and they were considered as ''junk.'' However, more and more investigations show that they are not junk and they should have certain biological functions [3,4], e.g., controlling gene regulation. So we expect that the analysis of correlation properties of noncoding sequences may provide some useful information about their functions. In the present paper, we shall investigate the correlation properties of one of the richest noncoding sequences in the human genome, Alu repeats.

The genome of eukaryotes is known to contain various types of repetitive DNA. Repetitive DNA is any piece of nucleotide sequence which is repeated several to many times in the genome. The function of repetitive DNA is unknown. These repetitive sequences can be classified broadly into two principal components: tandem repeats and interspersed repeats. Alu repeats belong to middle repetitive short interspersed nucleic elements (SINEs), and are specific to primates [5]. Their copy number in a human genome adds up to 500 000–1 000 000, which may account for approximately 10% of the whole human genome. Alu repeats are derived ancestrally from the 7SL RNA gene and mobilize through an RNA polymerase III-derived transcript in a process termed retroposition. Alu repeats share a recognition site for the restriction endonuclease AluI, which is what gave their name.

A typical Alu element is about 300 base pairs long and composed of two halves (about 130 base pair) joined by a middle *A*-rich region along with a 3′ PolyA(Adenine) tail. Alus are selfish DNA but not necessarily junk. Alus contribute to genetic diversity, are prone to gene conversion, and may be used in the modulation of protein expression. Alu sequences accumulate preferentially in gene-rich regions and are not uniformly distributed in the human genome [6]. They may play some role in the regulation of gene expression.

Although linear correlation methods were previously used to study the correlation properties of DNA sequences [7], in the present paper we shall use a nonlinear correlation approach, i.e., the nonlinear prediction technique [8,9]. The nonlinear correlation method defines the correlation property by the determinism of similar segments in a sequence. This method was developed for analyzing the correlation properties of irregular time series and has been previously used successfully to distinguish between chaos and noise in them. It can give specific information of how different regions are characterized and can detect the determinism which is not detected by the standard methods, such as Fourier transformation and power spectrum. It can also give reasonable results for short sequences. Furthermore this method can be extended easily to treat symbolic sequences. So it is reasonable to apply it to analyze DNA sequences which are typical irregular symbolic sequences. The theory of chaos has already been applied to investigate the behaviors of biomolecules and biosequences by many authors [10–16]. In the present paper, we use the nonlinear correlation method to investigate the deterministic structures of Alu repeats. Our results indicate that most Alu repeats have deterministic structures and show significant nonlinear correlations which are absent in exons and introns.

II. METHOD

The nonlinear prediction technique works as follows [8,9,13]. For an arbitrary symbolic series $x_1, x_2, x_3, \ldots, x_N$, one constructs a set of d-dimensional vectors

$$X_1 \equiv (x_1, x_2, \ldots, x_d), \tag{1}$$

$$X_2 = (x_2, x_3, \ldots, x_{d+1}),$$

$$\cdots,$$

$$X_{N-d+1} \equiv (x_{N-d+1}, \ldots, x_N)$$

which correspond to all possible segments of d consecutive symbols. The set contains all possible subsequences of length d in the original symbolic series. Next, for each vector $X_p = (x_p, x_{p+1}, \ldots, x_{p+d-1})$ $(1 \leqslant p \leqslant N-d)$, one searches for its nearest neighbors $X_{H(p)} = (x_{H(p)}, x_{H(p)+1}, \ldots, x_{H(p)+d-1})$ and then compares how close the symbols x_{p+d} and $x_{H(p)+d}$ follow these two vectors. The closeness of a pair of symbols x_i and x_j is measured by the Hamming distance

$$h(x_i, x_j) = \begin{cases} 0, & x_i = x_j, \\ 1, & x_i \neq x_j \end{cases} \tag{2}$$

while the closeness of a pair of vectors X_i and X_j is measured by

$$H(X_i, X_j) = \sum_{k=0}^{d-1} h(x_{i+k}, x_{j+k}). \tag{3}$$

The nearest neighbors $X_{H(p)}$ of a given vector X_p are X_j, such that $H(X_p, X_j)$ is a minimum for $j \neq p$. Once the nearest neighbors $X_{H(p)}$ are determined, we compute the mean local error $\varepsilon_p = \langle h(x_{p+d}, x_{H(p)+d}) \rangle$, where $\langle \cdots \rangle$ denotes the average over all the nearest neighbors of X_p since there are usually more than one nearest neighbor. From this, the overall mean error in the chain is

$$E = \frac{1}{N-d} \sum_{p=1}^{N-d} \frac{\varepsilon_p}{\varepsilon_r}, \tag{4}$$

where we measure ε_p in the expected local error ε_r of a random sequence with a composition identical to that of the Alu sequence. This make it easy to clarify how much the correlation properties of an Alu sequence deviates from random sequences. ε_r is calculated by

$$\varepsilon_r = (1/4) \sum_{i=1}^{4} h(x_{p+d}, \alpha_i) p(\alpha_i),$$

where $\{\alpha_i = A, C, G, T\}$ is the alphabet taken by x_i and $p(\alpha_i)$ is the probability of occurrence for the symbol α_i in the sequence. Consequently, if E for an Alu repeat is close to zero, the Alu sequence is deterministic. If E is close to or larger than 1, the Alu sequence is a random one.

Furthermore, we need to define a value of E on which we consider that a ''significant'' correlation exits. To do this, we calculated the correlation properties of the sequences of randomly shuffled Alu repeats, each of them with a composition identical to that of original Alu repeats but also with a randomly permuted ordering of nucleotides. Figure 1 is the plot of the overall mean error E versus the embedding dimension d for one Alu repeat and its ten randomly shuffled sequences.

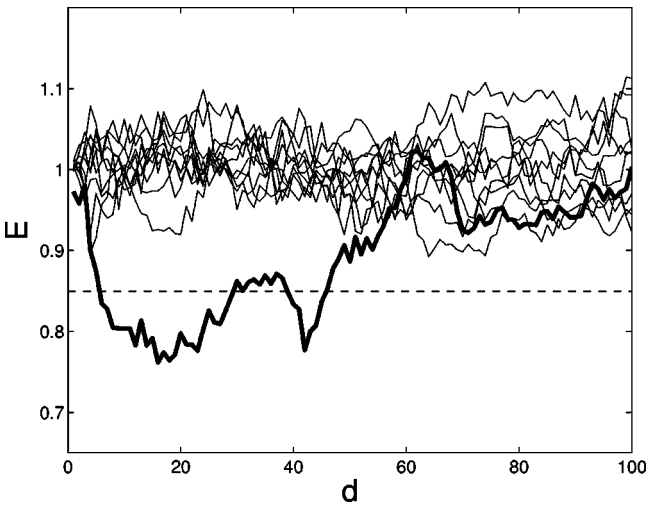

FIG. 1. Error E versus the embedding dimension d for an Alu repeat (thick line) and its 10 randomly shuffled sequences (thin lines) from the HUMHBB DNA sequence. It can be seen that the values of E of all of the randomly shuffled sequences are around 1 and not less than 0.9.

It can be seen clearly that the values of E of all ten randomly shuffled sequences are around 1 and not less than 0.9. Thus we can define that a ''significant'' correlation exits if $E \leqslant 0.85$.

III. RESULTS AND DISCUSSIONS

A. Alu repeats, exons, and introns

We first focus on the human β globin region on chromosome 11 [Genebank code HUMHBB, 73 308 bases]. The reason we choose HUMHBB is that it contains three kinds of typical DNA sequences: exons, introns, and Alu repeats. Therefore, we can give a comparative analysis of their correlation properties.

In HUMHBB, there are six globin genes and eight Alu family repeats. The six genes are the epsilon gene (hbe), G-gamma gene (hbgg), A-gamma gene (hbga), delta gene (hbd), beta gene (hbb), and pseudo-beta-1 gene (hbp). Every gene in HUMHBB contains three exons and two introns. The lengths of the exons are 93, 222, and 129 bases, respectively, and those of the introns are about 120 and 850 bases, respectively.

Figures 2–4 are the plots of the overall mean error E versus the embedding dimension d for exons, introns, and Alu repeats, respectively. In our calculations, the poly(A) tails of the Alu repeats are removed. It can be seen that, in general, the values of E of the exon and intron sequences are close to or larger than 0.85 and show no significant deviation from the random sequences. Some exon and intron sequences show very weak determinism. For only one exon sequence are their values of E smaller than 0.7 for $d > 50$ and show significant correlation. However, most of the Alu repeats have similar deterministic structures and show significant nonlinear correlations for d between 5 and 50. This behavior is very interesting and marks the existence of a common underlying deterministic rule in Alu sequences.

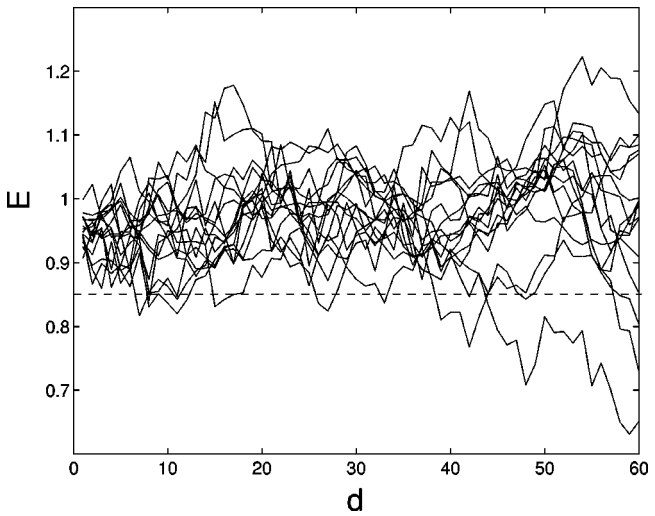

FIG. 2. Error E versus the embedding dimension d for 18 exons from the HUMHBB DNA sequence. It is shown that most exons show no significant deviation from the random sequences.

These results indicate that the correlation properties of Alu repeats are very different from those of exons and introns. Furthermore, although the overall structures of the determinism of the Alu repeats are similar, their details are not the same. Alu repeats are not just exact copies of each other.

B. Right and left halves of Alu repeats

Alu repeats are composed of two halves (about 130 base pair). To see whether the nonlinear correlations shown above are those between the right and left halves, we also studied the nonlinear correlations of them. Figure 5 shows that the right or left half alone does not have the nonlinear deterministic structures of all of the Alu repeats. This indicates that the nonlinear correlations of all the Alu repeats are mainly those between the two halves of the Alu repeats. The Alus are thought to emerge by dimerization of a subfamily free left arm monomer (FLAM) and free right arm monomer (FRAM). FLAM and FRAM arose by a 42 and 11 bp deletion, respectively, from the fossil Alu monomer (FAM). Therefore, although the two arms of the Alus are not identical, they share a common origin and a considerably sequence identity. So it is not surprising that correlations exist. However, it is noted that the significant correlations of all of the Alu repeats occur only for d between 5 and 50. For a dimeric structure with two identical halves, the range of correlation would be identical to the length of one half. A detailed analysis shows that the origin of the reduced correlation lengths from 5 to 50 is complicated because the correlations occur in different ranges of the two halves for different Alu repeats and, therefore, needs further investigation.

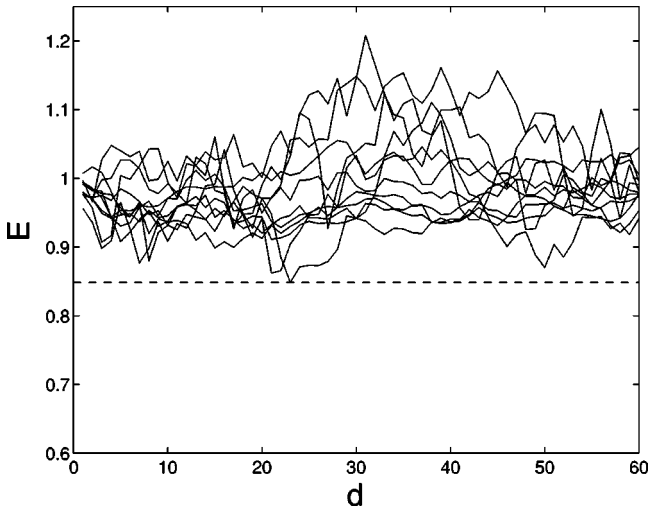

FIG. 3. Error E versus the embedding dimension d for 12 introns from the HUMHBB DNA sequence. It is shown that all introns show no significant deviation from the random sequences.

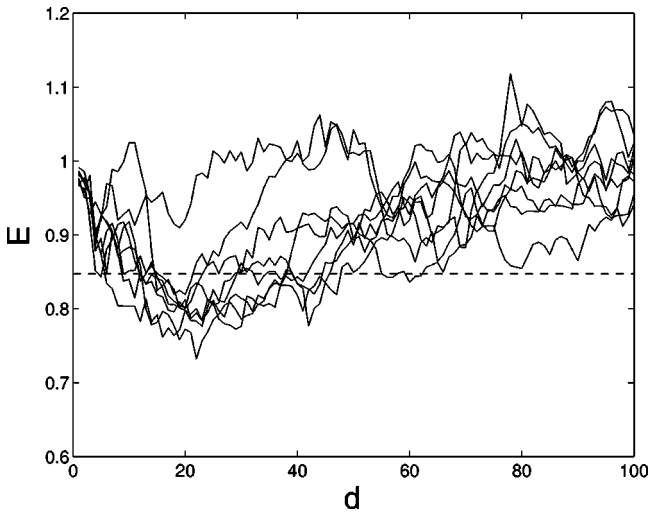

FIG. 4. Error E versus the embedding dimension d for eight Alu repeats from the HUMHBB DNA sequence. It is shown that most alus show significant deviation from the random sequences.

C. Alu repeats in larger dataset

The Alu repeats above are only come from one short segment of human DNA. In order to see whether the determin-

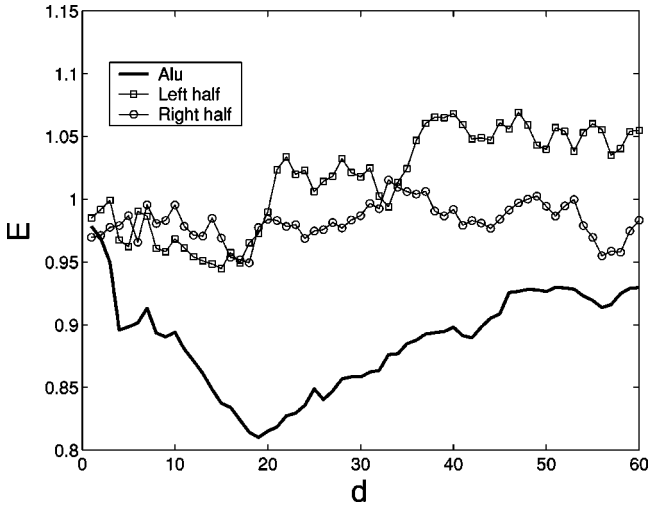

FIG. 5. Average error E versus the embedding dimension d for the left and right halves of the Alu repeats from the HUMHBB DNA sequence.

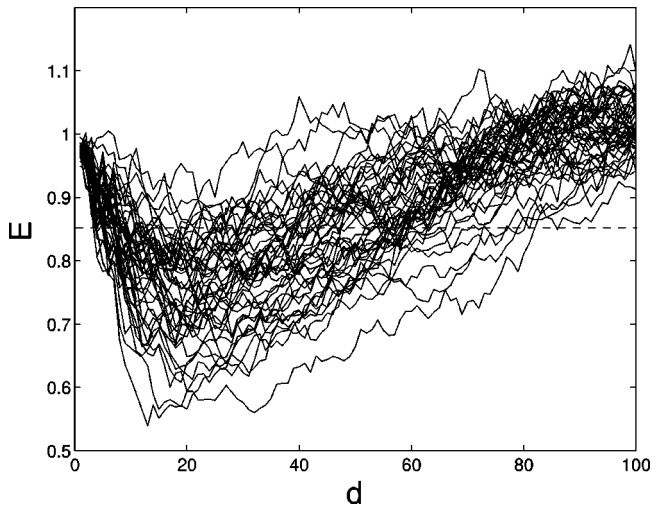

FIG. 6. Error E versus the embedding dimension d for the Alu repeats of the J subfamily. It is shown that most alus show significant correlations but their values of E and the lengths of the correlation segments are diverse.

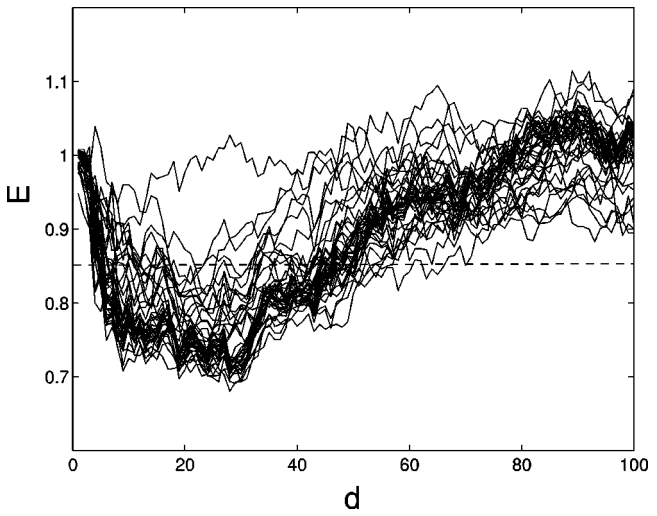

FIG. 8. Error E versus the embedding dimension d for the Alu repeats of the Yb8 subfamily. Most alus of this subfamily show significant correlations but their lengths of the correlation segments become short in comparision with the older subfamilies.

istic structures occur generally, we used a large dataset of human Alu repeats [17]. In our calculation, we randomly selected 50 sequences from each of the five subfamilies: Alu J, Alu S, Alu Yb8, Alu Ya5, and Alu Ya5a2. The major subfamily branches (J, S, and Y) appeared at different evolutionary times, with J being older than S, and S being older than Y. Figures 6–10 show the overall mean error E versus the embedding dimension d for Alu repeats of the five subfamilies, respectively. The results show that the correlations exist in most of the Alu repeats in the five subfamilies. However, the degrees (E) of the correlations and the lengths (d) of the correlation segments are different. For the Alu sequences of the old subfamilies (J and S), the degrees and lengths of the correlations are diverse. The degrees of the correlations are from zero to the case with $E \approx 0.5$. The lengths (d) of the correlation segments also extend over almost the whole range of calculated d's. But for the Alu Y subfamilies, the situation is different. The lengths of the correlation segments are short and less than about 50 bases. The degrees of the correlations for each of the Y subfamilies are not as diverse as the old subfamilies but tend to be similar and most of the sequences show significant correlations between $d \approx 5$ to $d \approx 50$. Furthermore, Figs. 6–10 show that the deterministic structures of the Alu repeats changed gradually, from J to S to Y subfamilies, into panlike shapes of the youngest Ya5a2 subfamily. The flat bottoms of the pans mean that the overall mean errors E and in turn the degrees of the correlations remain almost unchanging for this range of d. This implies that similar segments exist in the two arms and their lengths are about 50 bases. The detailed analysis of the Alus of the

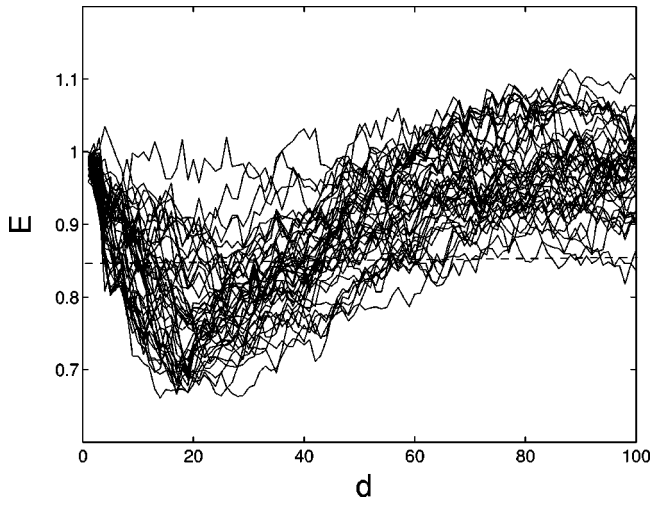

FIG. 7. Error E versus the embedding dimension d for the Alu repeats of the S subfamily. Similar to the J subfamiliy, most alus show significant correlations but their values of E and the lengths of the correlation segments are diverse.

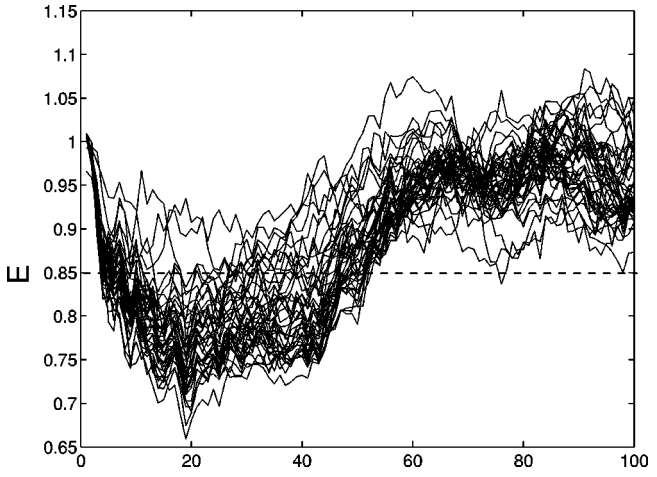

FIG. 9. Error E versus the embedding dimension d for the Alu repeats of the Yb8 subfamily. Most Alus of this subfamily show significant correlations and their lengths of the correlation segments are similar to the subfamily Yb8.

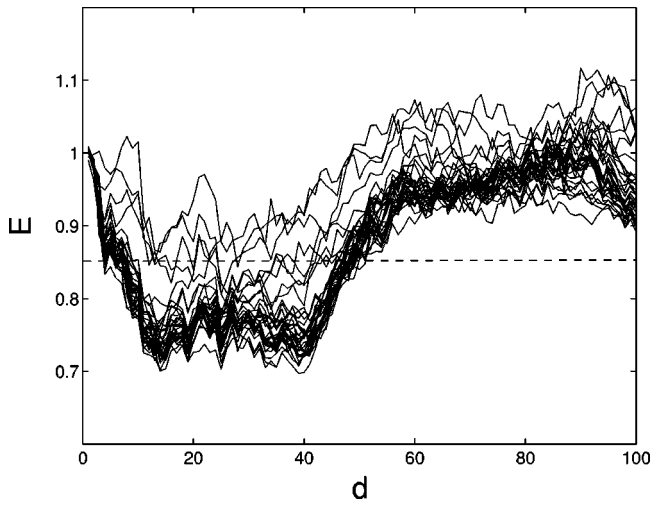

FIG. 10. Error E versus the embedding dimension d for the Alu repeats of the Ya5a2 subfamily. Most Alus of this subfamily show significant correlations and also panlike deterministic structures.

Ya5a2 subfamily shows that the similar segments are mainly the first part of the two arms. Its biological significance needs further investigation.

In conclusion, we calculated the nonlinear correlations of Alu repeats. The results can be summarized as follows.

(1) Most of the Alu repeats show deterministic structures which are clearly absent from those of random sequences and also different from those in exons and introns.

(2) The nonlinear correlations in the Alu repeats are due to their dimeric structures. However, the lengths of the correlation segments are less than those of the two arms.

(3) The deterministic structures of the Alu sequences of the old Alu subfamilies (J and S) are diverse but those of the young subfamilies (Yb8, Ya5, Ya5a2) tend to be similar. Furthermore, from the younger to older subfamilies, the deterministic structures gradually evolved from a panlike structures to a more diffuse correlation pattern. The lengths of the correlation segments of the young subfamilies are shorter and less than about 50 bases. As young Alus can be seen as mutation free copies of the ''master genes'' (of their respective family consensus sequences) it seems very likely that the old AluJ subgroup members showed the same panlike structure when they where ''young'' or mutation free.

It is known that all the members of the Alu repeat superfamily have common tRNA-like secondary structure. It may be possible that the deterministic structures of the Alus are related to this since the base paring of the two arms are needed to form the tRNA-like secondary structures or the base sequences of the two arms are required to have certain similarity or correlation. But this needs further investigations.

ACKNOWLEDGMENTS

The authors thank Professor Runsheng Chen of the Institute of Biophysics of the Chinese Academy of Sinica for calling their attention to the analysis of Alu repeats. This work was supported by the Natural Science Foundation of China under Grant No. 10175023 and 90103031 and by the Foundation of the Ministry of Education of China.

[1] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[2] W. Lee and L. Luo, Phys. Rev. E **56**, 848 (1997).
[3] C. W. Schmid, Prog. Nucleic Acid Res. Mol. Biol. **53**, 283 (1996).
[4] C. W. Schmid, Nucleic Acids Res. **26**, 4541 (1998).
[5] A. M. Weiner *et al.*, Annu. Rev. Biochem. **55**, 631 (1986).
[6] M. A. Batzer and P. L. Deininger, Nat. Rev. Genet. **3**, 370 (2002).
[7] C. K. Peng *et al.*, Nature (London) **356**, 168 (1992); W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992); B. Borstnik *et al.*, Europhys. Lett. **23**, 389 (1993); W. Li and K. Kaneko, *ibid.* **17**, 655 (1992); W. Li, T. Marr, and K. Kaneko, Physica D **75**, 217 (1994); H. E. Stanley *et al.*, Physica A **205**, 214 (1994); S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995); H. Herzel and I. Gross, Physica A **216**, 518 (1995); A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo *et al.*, Physica D **96**, 291 (1996); A. Arneodo *et al.*, Physica A **249**, 439 (1998); Y. Xiao *et al.*, J. Theor. Biol. **175**, 23 (1995); H. Herzel and I. Grosse, Phys. Rev. E **55**, 800 (1997); H. Herzel *et al.*, Physica A **294**, 449 (1998); L. Luo *et al.*, Phys. Rev. E **58**, 861 (1998); L. Luo and L. Tsai, Chin. Phys. Lett. **5**, 421 (1988); E. Coward, J. Math. Biol. **36**, 64 (1997); S. V. Buldyrev *et al.*, Physica A **249**, 430 (1998); Zu-Guo Yu *et al.*, Phys. Rev. E **63**, 011903 (2000); Zu-Guo Yu *et al.*, Chaos, Solitons Fractals **11**, 2215 (2000); Zu-Guo Yu and B. Wang, *ibid.* **12**, 519 (2001).
[8] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997), p. 42.
[9] D. Kaplan and L. Glass, *Understanding Nonlinear Dynamics* (Springer-Verlag, New York, 1995), p. 303.
[10] M. S. El Naschie and T. Kapitaniak, Phys. Lett. A **147**, 275 (1990).
[11] M. S. El Naschie, Chaos, Solitons Fractals **9**, 135 (1998).
[12] L. Cao, Physica A **247**, 473 (1997).
[13] J. Barral *et al.*, Phys. Rev. E **61**, 1812 (2000).
[14] Y. Huang and Y. Xiao, Chaos, Solitons Fractals **17**, 895 (2003).
[15] A. Guiliani *et al.*, Chem. Rev. (Washington, D.C.) **102**, 1471 (2002).
[16] J. P. Zbilut *et al.*, Cell Biochem. Biophys. **36**, 67 (2002).
[17] A. M. Roy-Engel *et al.*, Genome Res. **12**, 1333 (2002).