Computer Analyses of Site-Specific Variabilities in Human Alu Sequences

Michiko Muraki
t93441mm@sfc.keio.ac.jp

Masaru Tomita
mt@sfc.keio.ac.jp

Laboratory for Bioinformatics
Department of Environmental Information, Keio University
5322 Endo, Fujisawa, 252 Japan

1 Introduction

Alu repetitive sequences are Short INterspersed Elements (SINEs) which are about 300bps long. They are said to exist 500,000 to 800,000 copies throughout the human genome. Despite the abundance, their biological function and prolifiration process are poorly understood.

Site-specific variabilities of Alu sequences were investigated by Kapitonov *et al.* [3] and Britten[1] independently in 1994. Kapitonov *et al.* analyzed primary structures of 32 families of dispersed repeats (DR) of eukaryotic genomes: Alu, B1, B2 LINE repeats of mammals. They showed that enhancer-like structures and putative pol III promoters in many of those repetitive sequences are relatively well conserved.

Britten[1], on the other hand, divided Alu sequences into two randomly chosen sets (789 each) and compared all the sequneces in each set with the consensus of recently inserted Alu elements. He, too, argued that the putative "pol III promoters" of Alu sequences are more conserved than other regions.

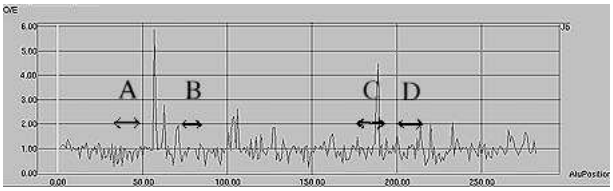
We have conducted comprehensive computer analyses using a larger number of Alu sequences (about 2500) with the following modifications: (1) the Alu sequences were classified into 12 subfamilies according to Batzer *et al.*[4], and each set of sequences has been analyzed respectively; (2) sequence variabilities were computed based on trinucleotides for finer comparison.

2 Materials and Methods

2500 Alu sequences were extracted from the GenBank database and classified into 12 subfamilies using computer programs provided by Jurka[2]. The number of sequences in each subfamily is as follows : Old: Jb(212) and Jo(283); Middle: Sc(161), Sg(178), Sp(188), Sq(164)、 Sx(512), and S(458); Young: Y(278), Ya1(12), Sb1(30) and Yb8(12);

We aligned sequences in each subfamily and computed the Observed/Expected value of mutation rate for each trinucleotide position, where the expected values are site-independent mutation rates, and the observed values are site-specific mutation rates.

3 Results and Discussion



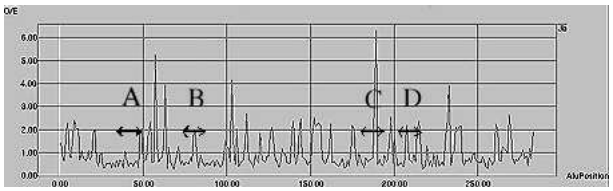


Figure 1 (left) : Site-specific mutation rates (O/E) of Alu Jb subfamily based on trinucleotides
Figure 2 (right): Site-specific mutation rates (O/E) of Alu Jb subfamily based on singlenucleotides

A and C : "core sequences" of pol III promoter"
B : "consensus B box" of "pol III promoter"
D : enhancer-like structure

Figure 1 shows site-specific mutation rates (O/E) of Alu Jb subfamily based on trinucleotides. The O/E values are around 1.0 throughout the sequence, indicating that no particular regions are more conserved than others. Unlike Britten[1] and Kapitonov[3], we are unable to conclude that putative functional sites (such as pol III promoters and enhancer-like structures) in each Alu subfamilies have been conserved in the course of evolution.

We are currently investigating the two outstanding peaks around positions 60 and 185. It may be possible that the consensus sequence of Alu Jb we quoted had errors in those positions.

Figure 2 shows the results of the same experiment as Figure 1 except that it is based on single nucleotides rather than trinucleotides. As we can see in Figure 2, there are a large number of peaks due to local hotspots as CpG, and it is hard to determine whether or not there are site-specific conservations in the sequence. It was thus necessary to compute the sequence variabilities based on trinucleotides.

4 Conclusion

Unlike the previous reports (Britten[1] and Kapitonov[3]), our analyses, based on trinucleotides, could not find any outstanding site-specific conservations among sequences in any of the 12 Alu subfamilies.

Acknowledgement

We thank Dr. Jerzy Jurka (Genetic Information Research Institute, Palo Alto, Ca, U.S.A) and his research group for providing us the software and data on Alu elements. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] R. J. Britten, "Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes" *Proc. Natl Acad. Sci. USA*, 91. 5992-5996, 1994.

[2] J. Jurka, "Origin and Evolution of Alu Repetitive Elements" in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome.* ed. by R.J. Maraia from R.G. Landes Company, 1995.

[3] V. V. Kapitonov, N. A. Kolchanov, "General Characterization of Mobile Genetic Elements and Transposable Sequences in Eukaryotic Genomes" in *Computer Analysis of Genetic Macromolecules.* ed. by N.A.Kolchanov and H.A.Lim from World Scientific, 1994

[4] M. A. Batzer, P. L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C. M. Rubin, C. W. Schmid, E. Zietkiewicz, E. Zuckerkandl, "Standardized Nomenclature for Alu Repeats" *J Mol Evol*, 42:3-6, 1996