Inferring History of Genomic Duplication Using Subclassified Alu Elements*

Yoshimi Toda [1] Masaru Tomita [2]
ytoda@mag.keio.ac.jp mt@sfc.keio.ac.jp
Laboratory for Bioinformatics
[1] Graduate School of Media and Governance
[2] Department of Environmental Information
Keio University
5322 Endo, Fujisawa, 252 Japan

1 Introduction

Alu elements, a family of short interspersed elements (SINEs) which are characteristic to primate genome, can be subclassified into 12 subfamilies (Jo, Jb, Sz, Sx, Sq, Sp, Sg, Sc, Y, Ya1, Ya5, Ya8)[1]. As the evolutionary age of each subfamily is known, Alu elements inserted in genomic sequences can be used as markers to infer the time and order of sequence duplication events.

The human growth hormone locus, HUMGHCSA (GenBank accession number: J03071), was chosen as an example because of three characteristics: five coding sequences have more than 90% similarity [2], the locus is Alu-rich and the same arrangement of Alu elements appear more than once, and some of the Alu elements have common direct repeats as well as poly-a tails. These characteristics suggest that those Alu elements proliferated by genomic duplication rather than transposition.

Using this unique locus as an example, we showed that, using Alu elements as markers, history of duplication can be inferred more precisely and studied in more detail. Our analysis showed consistency in the most part with Chen et al.[2] and gave different insight into the analysis of duplicated genomic sequences as information on Alu subfamilies was incorporated.

2 Methods

The GenBank sequence data of the locus for human growth hormone (GH-1 and GH-2) and chorionic somatomammotropin (CS-1, CS-2, and CS-5) genes (Locus Name: HUMGHCSA; Accession number: J03071) were used for the analysis.

Two sets of software were used to find Alu elements. 40 complete Alu elements (longer than 250 nucleotide length) in the locus were detected and classified into 12 subfamilies using CENSOR[1] additional 4 complete and 4 partial Alu elements were detected by our own program.

A phylogenetic tree of 40 complete Alu elements was drawn to show Alu clusters using ClustalW (Ver. 1.6) and PHYLIP (Ver. 3.54). There were clusters of Alu elements each of which consists of more than one Alu. Each cluster was labeled with its subfamily name and a code unique for each cluster. Within each cluster, similarities were studied in more detail by using direct repeats and poly-A tails as markers.

*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.

[1]We thank Dr. Jerzy Jurka (Genetic Information Research Institute, Palo Alto, CA, U.S.A.) and his research group for providing us with software and data on Alu elements.

3 Results and Discussion

Figure 1 shows 48 labeled Alu elements as well as five coding sequences in the same order as they appear in the locus. Duplication order and time were inferred by comparing directions of Alus, arrangements of labeled Alus, similarity of direct repeats, and distances between Alu elements.

As shown in Figure 1, the same and similar Alu arrangements appear more than once. One possible duplication scenario, which is consistent with Chen et al. [2], based on those information is presented in Figure 1. Region 1 is the original sequence and region 1' is the duplicate of region 1; region 2 takes part of region 1 and whole region 1' as the origin of region 2'; region 3 takes part of region 2 to be duplicated into region 3'.

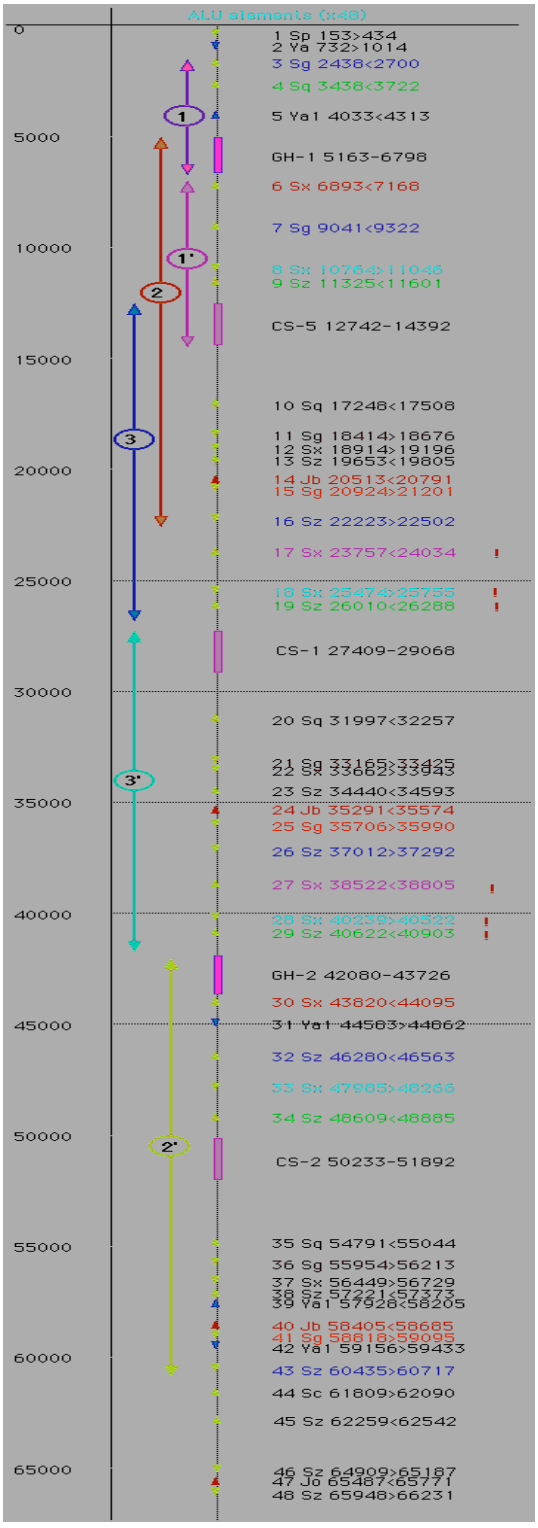

Figure 1. 48 Alu elements and 5 genes in the locus HUMGHCSA, GenBank accession number J03071.

4 Summary

As been described above, Alu subfamily classification, direct repeats, and poly-a tails can be used as markers to refine sequence analysis and infer history of duplication events with high degree of confidence.

References

[1] Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., and Zuckerkandl, E.. "Standardized nomenclature for Alu repeats." *J. of Mol. Evol.*, Vol. 42, pp. 3-6, 1996.

[2] Chen, E.Y., Liao, Yu-Cheng, Smith, D.H., Barrera-Saldana, H.A., Gelinas, R.E., and Seeburg, P.H.. "The human growth hormone locus: Nucleotide sequence, biology, and evolution." *Genomics*, Vol. 4, pp. 479-497, 1989.

[3] Jurka, J.. "Origin and evolution of Alu repetitive elements." in *The impact of short interspersed elements (SINEs) on the host genome*, ed. Maraia, R.J.. Ch. 2. R.G. Landes Company. 1995.

[4] Kapitonov, V. and Jurka, J.. "The age of Alu subfamilies." *J. of Mol. Evol.*, Vol. 42, pp. 59-65, 1996.

[5] Vnencak-Jones, C.L. and John A. Phillips III. "Hot spots for growth hormone gene deletions in homologous regions outside of Alu repeats." *Science*, Vol. 250, pp. 1745-1748, 1990.