Comprehensive Sequence Analyses of 5' Flanking Regions of Primate *Alu* Elements

Yoshimi Toda [1 2] **Rintaro Saito** [1 2] **Masaru Tomita** [1 3]

ytoda@sfc.keio.ac.jp rsaito@sfc.keio.ac.jp mt@sfc.keio.ac.jp

[1] Laboratory for Bioinformatics
[2] Graduate School of Media and Governance
[3] Faculty of Environmental Information, Keio University, 5322 Endo, Fujisawa 252-0816, Japan

Abstract

Retrotransposons have been generally known to integrate randomly into host genomes. Jurka (1997) [3], however, showed consensus sequence patterns at integration sites of certain mammalian retrotransposons, and suggested involvement of sequence specific enzymes that mediate integration.

We have conducted comprehensive sequence analyses of 5' flanking regions of primate *Alu* elements. In contrast to the small but clean data set Jurka (1997) [3] used, (1) larger number of samples were used, (2) wider region of 5' end of *Alu* elements was analyzed, and (3) comparisons were made among different subfamilies for comprehensive analyses in order to identify characteristic sequence pattern(s) preceding 5' end of *Alu* elements. The nucleotide occurrences at each position within 500 bases of 5' end of *Alus* were counted to obtain profiles. Information content at each nucleotide position in the same region was, then, computed. Distinctive difference in the nucleotide composition and information content values that divides the region into two was observed. The region between -20 and 5' end of *Alu* elements is found to be highly adenine-rich and shows significantly higher information content values compared to the rest of the region, implying the existence of certain characteristic sequence pattern in this region. Also, younger subfamilies of *Alu* elements show higher information content values than older subfamilies. This implies that certain characteristic sequence pattern already existed in the region between -20 and 5' end of *Alu* elements at the time of *Alu* integration, and accumulation of mutation in the course of time resulted in the less distinctive sequence pattern in older sequences. Frequencies of all possible triplets (total of 64) were measured in the same region in order to identify characteristic sequence pattern(s). Observation that frequencies of triplets "aaa," "taa" and "tta" in the 5' flanking sequences were high is consistent with Jurka (1997) [3]. Frequencies of some other triplets such as "gaa," "caa," "aac," "ctt," "gtt," "atg," etc. which do not comprise the primary candidates for the nick site in Jurka (1997), also show significantly high frequencies.

1 Introduction

Retrotransposons are a type of repetitive elements commonly found in genomes of various organisms, yet the general mechanism of their integration has not been completely understood (Rogers, 1985 [9]). The general understanding has been that their integration into host genomes occurs at random. However, Jurka (1997) [3] demonstrated that there are possible consensus sequence patterns at integration sites of certain mammalian retrotransposons, which strongly suggests sequence-specific enzymatic involvement that mediates integration. In the study conducted by Jurka, approximately 344 human *Alu* sequences and 56 rodent ID sequences that retain full length with identical flanking repeats at both ends were carefully selected.

We have conducted comprehensive analyses of 5' flanking regions of nearly 30,000 primate *Alu* elements, in order to identify sequence pattern(s) and their location that may be characteristic to the region. The following 3 aspects make our study distinct from Jurka's work: (1) a larger number of samples were used; (2) wider region of 5' upstream of *Alu* elements was considered; and (3) each of 12 subfamilies was individually analyzed.

Alu repetitive sequences constitute an *Alu*-family of short interspersed nucleic elements (SINEs). They are about 300 base pairs (bps) long and are fixed in primate genomes (Weiner et al., 1986 [11]). Their copy number in a human genome is estimated to be several hundred thousand to one million, which accounts for approximately 5 to 10% of the entire human genome (Weiner et al., 1986 [11]; Okada, 1991 [6]; Okada, 1994 [7]). Structurally, an *Alu* sequence consists of 2 similar subunits, connected with an adenine-rich (A-rich) linker, and a 3' poly-A tail. This unit is usually flanked immediately by 5' end and 3' end direct repeats (4-10 bps long on the average) (Rinehart et al., 1980 [8]; Daniels and Deininger, 1985 [2]).

Alu elements are subclassified into three groups, namely, Old, Middle, and Young. They are further classified into 12 subfamilies (Old: Jo, Jb; Middle: Sz, Sx, Sq, Sp, Sg, Sc; Young: Y, Ya1, Ya5, Yb8) in accordance with their putative time of proliferation and based on diagnostic positions in their nucleotide sequences (Batzer et al., 1996 [1]).

In the current study, the nucleotide occurrences at each position in the 5' flanking region of *Alu* elements were counted to obtain profiles of nucleotide composition. We then computed information content value (Schneider et al., 1986 [10]) at each nucleotide position to visualize possible locations of the characteristic nucleotide pattern. In order to specify the sequence pattern(s), frequencies of all possible triplets (total of 64) were also measured.

2 Materials and Methods

2.1 Data

Primate *Alu* sequences longer than 250 bps were extracted from NCBI GenBank file release 103.0, along with their 5' flanking sequences up to 500 bps long, a computer program named CENSOR (Jurka et al. 1996). Altogether, 29,663 *Alu* sequences were extracted and classified into 12 subfamilies. The number of analyzed *Alu* sequences of each subfamily is shown in Table 1.

2.2 Nucleotide profiles

Occurrences of four nucleotides (adenine, thymine, guanine, cytosine) at each nucleotide position up to 500 bps upstream of 5' flanking region were counted in order to obtain nucleotide profiles for each subfamily.

2.3 Information content computation

The following formula was used to compute information content values which were standardized to avoid influence from base composition of analyzed sequence.

$$\text{Information content} \quad = \quad O_i \log_2 O_i / E_i \tag{1}$$

where

$$\begin{aligned} O_i &= \text{Observed frequency of base } i \text{ at specific position} \\ E_i &= \text{Expected frequency of base } i \text{ at specific position} \end{aligned}$$

An observed/expected (O/E) ratio of each nucleotide is defined as the observed frequency divided by the expected frequency, where the expected frequency is computed from the base composition of the whole analyzed sequences. High information content value at a certain position indicates a small variety in base composition, i.e., specific nucleotide(s) appear at the position more often than expected.

2.4 Trinucleotide frequency

Frequency values of 64 possible trinucleotides were measured. The number of trinucleotides observed at a position was divided by the number of sequences analyzed and the resulted frequency was plotted for each position.

3 Results and discussion

3.1 Profile analysis

The nucleotide profiles of sequence composition upstream of 5' end of *Alu* elements show that the region is A-T-rich. Particularly, A is the most prominent nucleotide especially in the region between −20 and −6 of 5' end of *Alu* elements (Table 2). Younger subgroups of *Alu* elements show more distinct A-richness. Table 2(a) shows the nucleotide occurrences of the old subfamilies of *Alu* elements, Table 2(b) shows that of the middle subfamilies, and Table 2(c) shows that of the young subfamilies. With all the subgroups, the rate of A/T exceeds 60% at the positions −20 to −1. The rate of A exceeds 40% in the positions between −16 and −6 with the oldest subgroup, the highest being 56.2%, between −17 and −6 with the middle subgroup, the highest being 65.4%, between −17 and −5 with the youngest subgroup, the highest being 67.1%. A-T-rich characteristics is as expected. It is known that *Alu* elements are inserted into A-T-rich regions that may even include tails of pre-existing *Alu* elements (Rogers, 1985 [9]; Daniels and Deininger, 1985 [2]). The result also suggests that the insertion sites were more A-rich originally at the time of insertion and mutation resulted in slightly smaller occurrences of A at the region which supports Daniels and Deininger (1985) [2].

3.2 Information content

The existence of a characteristic nucleotide pattern was made more apparent from visualization by plotting the information content values onto a graph. In all the analyzed 12 subfamilies, information content is high at the positions between −20 to −15 upstream of the 5' end of *Alu* elements. As shown in Fig. 1, the information content v*Alu*e is the highest with the youngest subfamilies (average information content values of Ya1, Ya5, and Yb8 indicated with "Young" in the graph). This is expected as *Alus* belonging to these subfamilies are evolutionarily young, that is, relatively short period of time has passed since their integration into the host genomes. The subfamilies Jo and Jb, which are evolutionarily the oldest, show lower information content values compared to the younger ones (indicated with "Old" in Fig. 1). It is suspected that not only the sequences of *Alu* themselves, but also flanking regions of old *Alu* elements have accumulated mutations in the course of time and lost their original sequence pattern(s) in the region. However, in contrast to the region between −20 and −10, randomness of nucleotide composition observed in the region further upstream of −20 shows no differences among different subfamilies. This difference of randomness in sequence composition between the flanking region and the region further upstream of it, and among different subfamilies imply that the observed characteristic sequence pattern at the flanking region already existed at the time of each *Alu* integration and random mutation have accumulated since then in the course of time.

The observation that the region, between −20 to −10 preceding the 5' end of *Alu* elements, shows high information content values is consistent with the analysis in Jurka (1997) [3]. Jurka computed χ^2 values of the suspected consensus patterns for each position of 30 bp region preceding 5' end of *Alu* elements and shows that the positions between −19 to −10 preceding the 5' end of *Alu* elements demonstrate significantly high χ^2 values. Consistent with their analysis, our result strongly suggests the existence of a certain characteristic sequence pattern(s) around this region.

3.3 Trinucleotide frequencies

In order to identify DNA sequence pattern(s) involved in the high information content, frequencies of 64 possible triplets were measured. *Alu* sequences and their 5' flanking regions of old, middle and young groups were separately analyzed.

Of all the 64 possible triplet patterns, the most statistically significant are "aaa," "taa," and "tta". Graphical visualization of the results are shown in Fig. 2. Placing these three trinucleotides in order appearing in the analyzed sequences shows "ttaaaa" as the most frequently appearing pattern at the positions between -17 and -13. This is consistent with Jurka (1996 [4–5], 1997 [3]) suggesting that the hexamer "ttaaaa" is the most prominent characteristic sequence pattern.

In our analyses, other triplets such as "aag," "aga," "gaa," "acc," "caa," "aca," "aac," "ctt," "gtt," "cat," "cta," "tct," "atg," etc. also show high standardized frequencies. While their standardized frequencies are lower than that of "aaa," "taa," or "tta," they are still statistically significant. Further investigation is required for those triplets that do not seem to be related to the primary candidates for nick site for *Alu* integration in Jurka (1997).

4 Conclusion

In this study, comprehensive analyses of patterns in large flanking regions of 5' end of *Alu* elements were performed by (1) counting nucleotide occurrences, (2) computing information content, and (3) measuring frequencies of triplet nucleotide patterns. Profiles of nucleotide occurrences show that the flanking region of 5' end of *Alu* elements are rich in adenine which emphasizes the fact that *Alu* elements are inserted into A-T rich regions and particularly into A-rich regions (Daniels and Deininger, 1985 [2]). It is also shown that *Alus* that belong to the younger subgroup are more A-rich than the older ones. A characteristic sequence pattern "ttaaaa" was found through the information content analysis in conjunction with trinucleotide frequency analysis. Change in information content values among evolutionarily different subgroups of *Alus* together with A-richness in younger subgroups of *Alus* further emphasize the possibility that the characteristic sequence pattern existed where Alus would be integrated. Thus as Jurka (1997) [3] suggests, the pattern found to be most frequently appearing in this region can be the primary candidate for the nick site which is involved in *Alu* integration into host genomes. Some patterns which were found to be not consistent with Jurka (1997) [3] need further investigation and the flanking region of 3' end of *Alu* elements also need to be analyzed for more supporting evidence of the enzymatic involvement for *Alu* integration.

Acknowledgement

We would like to thank Dr. Jerzy Jurka, Paul Klonowski, and Jolanta Walichiewicz at the Genetic Information Research Institute, Palo Alto, CA, for data on *Alu* elements. This work is 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] 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. Mol. Evol.*, 42:3–6, 1996.

[2] Daniels, G.R. and Deininger, P.L., Integration site preference of the *Alu* family and similar repetitive DNA sequences, *Nucleic Acids Res.*, 13: 8939–8954, 1985.

Table 1: The number of samples of each subfamily used in our study.

	subfamily	No. of samples
Old	Jo	2853
	Jb	5677
	Sz	4279
	Sq	2728
Middle	Sp	2011
	Sx	4527
	Sg	2462
	Sc	1514
	Y	3132
Young	Ya1	105
	Ya5	169
	Yb8	206
	Total	29663

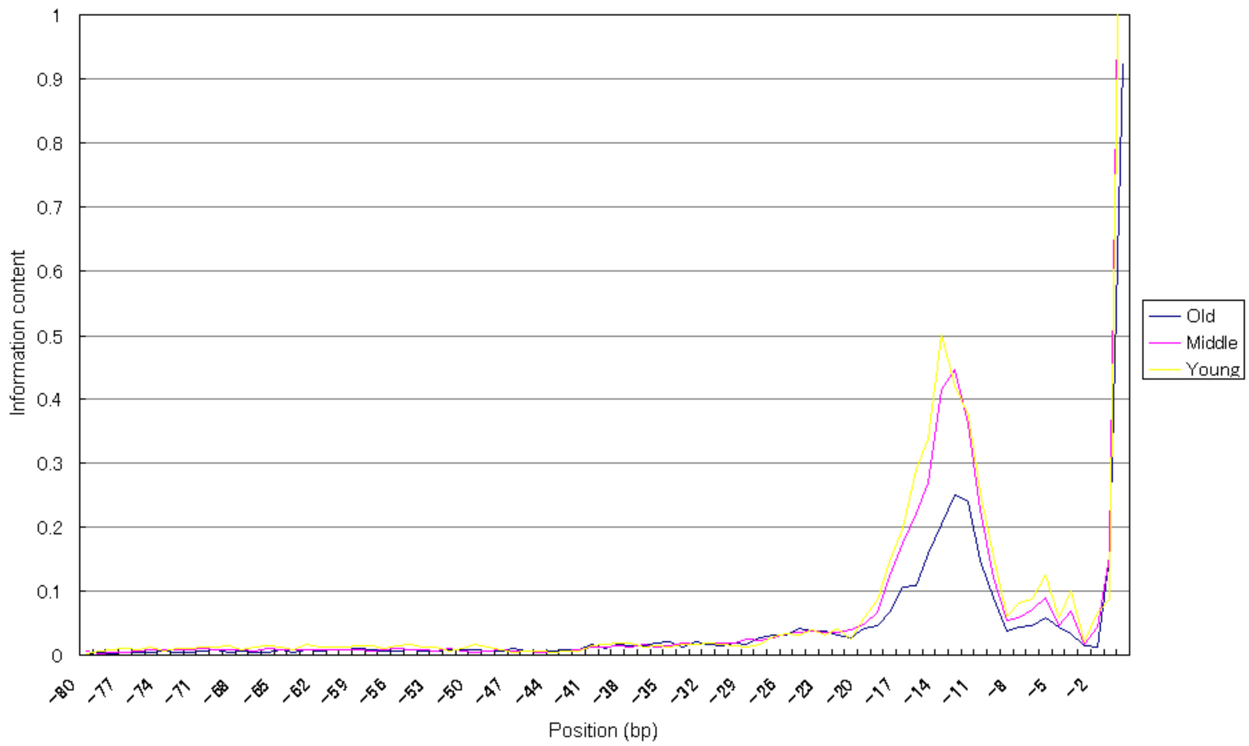


Figure 1: Information content of 5' flanking region of *Alu* elements.

Table 2: Nucleotide occurrences at each nucleic position in the 5' upstream flanking region of *Alu* elements.

Pos	-32	-31	-30	-29	-28	-27	-26	-25	-24	-23	-22	-21
a	33.2	33.2	34	33.3	36.1	35.7	36.5	36.8	34.7	34.2	34.5	34.6
t	28.3	28	27.6	28.5	26.9	28.4	27.4	28.8	30.4	30.9	29.7	29.1
c	18.9	19	18.6	18.5	18.8	17.9	17.9	16.3	16.4	16.5	17	18.2
g	19.5	19.8	19.7	19.6	18.2	18	18.2	18.1	18.4	18.4	18.8	18.2

Pos	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9
a	34.6	35.2	37.5	39.9	41.5	48.5	52.9	56.2	55.2	49.6	44.4	39.1
t	31.2	30.9	31.2	32.6	31.1	26	22.5	20	21.3	21.9	23.6	23.8
c	16.7	17.4	15.8	13.9	14	11.9	10.8	9.8	9.7	12.3	13.4	16.3
g	17.5	16.5	15.6	13.6	13.4	13.6	13.8	13.9	13.8	16.2	18.6	20.8

Pos	-8	-7	-6	-5	-4	-3	-2	-1
a	39.5	40.5	41.8	38.4	33.5	28	24.8	43.8
t	24.8	23	23.4	26.7	31	30.3	30.5	30.2
c	17.5	16.4	15.2	15.1	16	21.7	21.8	7.7
g	18.2	20	19.6	19.8	19.5	20	22.9	18.2

(a) Old subgroup

Pos	-32	-31	-30	-29	-28	-27	-26	-25	-24	-23	-22	-21
a	33.5	33.4	33.6	34.5	34.6	34.9	35.4	35.8	34.8	33.6	32.3	32.9
t	27.4	27.5	27.5	27.4	27.2	27.5	28.1	28.2	29.7	30.2	31.2	31
c	19.3	19.6	19.3	19.4	19.4	18.8	18.5	17.7	17.8	18.4	18.6	19
g	19.9	19.4	19.6	18.7	18.7	18.7	17.9	18.3	17.7	17.8	17.8	17

Pos	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9
a	33.1	33.9	36.6	40.8	48.5	55.1	64.1	65.4	61.1	53.3	47	41.7
t	32	33.2	35.9	35.9	30.6	24.1	18.2	17.2	19.7	23.2	23	21.5
c	18.7	18	15.2	12.5	10	8.9	6.8	5.5	6.8	8.8	12.6	16.6
g	16.2	14.8	12.4	10.7	10.9	11.8	10.9	11.9	12.3	14.7	17.4	20.2

Pos	-8	-7	-6	-5	-4	-3	-2	-1
a	41.6	43	44.6	38.2	34.3	29.1	20.5	43.6
t	23.2	23.1	22.6	26.7	33.6	30	32.5	30
c	16.6	15.5	13.6	15.2	15	20.3	25.1	8
g	18.6	18.4	19.2	19.9	17.1	20.5	21.8	18.3

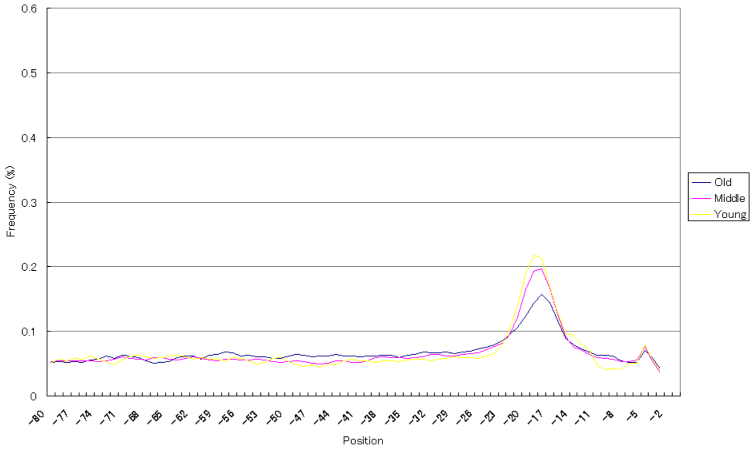
(b) Middle subgroup

Pos	-32	-31	-30	-29	-28	-27	-26	-25	-24	-23	-22	-21
a	33.2	33.9	34.2	33.5	34.1	35.5	35.8	33.7	34.8	31.6	31.5	30.9
t	27	25.5	24.9	25.1	25.4	25.4	26.4	28.7	28.8	30.1	31.3	29.6
c	18.8	20.4	20.2	20.7	20.7	21.3	19.7	17.9	18.6	19.6	19.7	19.4
g	21	20.2	20.7	20.6	19.7	17.8	18.1	19.7	17.8	18.7	17.6	20.2

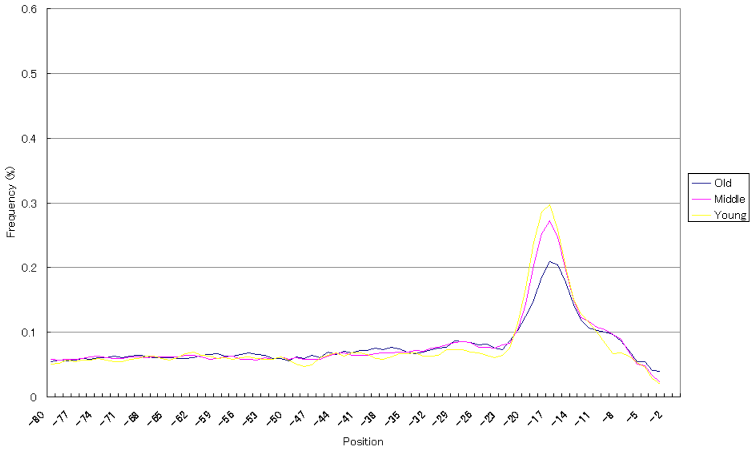
Pos	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9
a	31.4	33.8	36.6	42.3	52.1	59.3	67.1	64.5	61.7	54.9	49.4	42.7
t	33.2	34.2	36.7	35	29.6	21.5	17.9	16.5	19.7	22.3	22.3	18.9
c	18.9	17.3	14.8	12	9	7.5	5.2	6.3	7	8.7	12	18.2
g	16.5	14.6	11.9	10.7	9.3	11.8	9.7	12.8	11.7	14.1	16.3	20.2

Pos	-8	-7	-6	-5	-4	-3	-2	-1
a	44.6	45	47	40	38.2	30.9	20.5	39.8
t	19.1	20.7	23.3	25	32.2	28.6	35.2	29.3
c	18.7	16.3	12.8	14.6	13.3	21.4	24	12.8
g	17.6	17.9	16.9	20.5	16.2	19	20.2	18

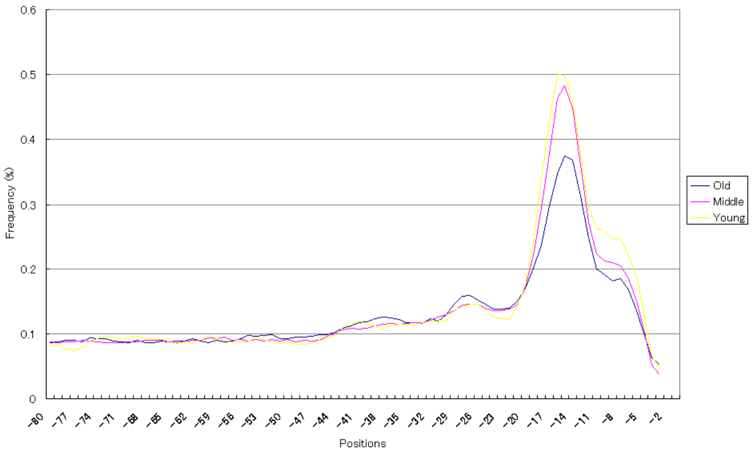
(c) Young subgroup



(a) Frequencies of trinucleotide "tta."



(b) Frequencies of trinucleotide "taa."



(c) Frequencies of trinucleotide "aaa."

Figure 2.

[3] Jurka, J., Sequence Patterns indicate an enzymatic involvement in integration of mammalian retroposons, *Proc. Natl. Acad. Sci. USA*, 94: 1872–1877, 1997.

[4] Jurka, J. and Klonowski, P., Integration of retroposable elements in mammals: Selection of target sites, *J. Mol. Evol.*, 43: 685–689, 1996.

[5] Jurka, J., Klonowski, P., Dagman, V., Pelton, P., CENSOR – a program for identification and elimination of repetitive elements from DNA sequences, *Comput. Chem.*, 20: 119–122, 1996.

[6] Okada, N., SINEs: short interspersed repeated elements of the eukaryotic genome, *TREE*, 6: 358–361, 1991.

[7] Okada, N., Retroposons as time markers of evolution, (Toki no jihyou to shite no retroposon), *Proteins, Nucleic Acids, and Enzymes* (Tanpakushicu, kakusan , kouso), 39:2724–2735, 1994.

[8] Rinehart, F.P., Ritch, T.G., Deininger, P.L. and Schmid, C.W., Renaturation rate studies of a single family of interspersed repeated DNA sequences in human deoxyribonucleic acid, *Biochem.*, 20: 3003–3010, 1980.

[9] Rogers, J.H., The origin of retroposons, *Int. Rev. of Cytol.*, 93: 187–279, 1985.

[10] Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A., Information content of binding sites on nucleotide sequences, *J. Mol. Biol.* 188(3):415–431, 1986.

[11] Weiner, A.M., Deininger, P.L. and Efstratiadis, A., Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information, *Annu. Rev. Biochem.*, 55:631–661, 1986.