**INVARIANT SECONDARY STRUCTURE OF ALU REPEATS PREDETERMINES CLUSTERIZATION OF REGULATORY ELEMENTS IN HUMAN GENOME.**

BLINOV V.M.1+, UVAROV D.L.1, RESENCHUK S.M.1, CHIRIKOVA G.B.1, KISSELEV L.L.2

1Institute of Molecular Biology, State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk region, 633159, Russia;
e.mail: blinov@vector.nsk.su;

2Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 117984, Russia;

+Corresponding author

Keywords: nucleotide sequences, alu repeats, regulatory elements, DNA/RNA secondary structure, compensatory mutations

Alu repeats represent about 5% of the entire human genome. We identified tetraplex GC-rich clusters (GGGC) and (GCCC)in the dimeric Alu repeats of the human genome and demonstrated that these nucleotide sequences form a unique carcass structure for each Alu monomer. Intermediate triplex structures were revealed between GC-reach clusters. The fine structure of repeated and complementary modules of Alu repeats was clarified. Optimal invariant tRNA-like structures were found for all Alu repeats. The carcass model of Alu monomers is repeated at a higher, chromosomal, level. In this case, monomeric, dimeric, and trimeric forms of Alu repeats appear as independent units and are involved in formation of alternative dynamic combinations.

**1. Introduction**

Primate-specific Alu repeats, belonging to the SINE class, constitute about 5% of the entire human genome. They are located predominantly in the R segments of the chromosomes together with the housekeeping genes and tissue-specific genes that replicate at the beginning of the cell cycle
S phase.

The genes containing Alu repeats differ in both the number of Alu repeats and their location. They are usually present as single copies or clusters in the noncoding gene regions, that is, in introns or 5' and 3' flanking regions; however, they can be also found in the translated regions, although rarer.

An Alu repeat contains approximately 300 bp and is consisted of two homologous halves connected with an A-rich linker. The right (R) half contains an additional segment of 31 bp (Fig. 1). Both halves have a common ancestral nucleotide sequence, homologous to 7SL nuclear RNA. The monomeric variant of Alu repeat was discovered in rodents and designated as B1. Despite the differences in their organization, both B1 and Alu repeats retained common features. Both repeats contain an internal two-module promoter (A and B boxes) that is recognized by RNA polymerase III and a poly(A) sequence. It is supposed that the common structural features of B1 and Alu repeats determine their capability of transposition within the genome. Dimeric structure, typical of primates, appeared about 65 million years ago.

Alu repeats are divided into several families, differing in their evolutionary age, basing on DNA divergence and certain "diagnostic" substitutions. The families of Alu repeats differ in their CpG dinucleotide content, the ability to bind RNA polymerase III, and the length of poly(A) regions. Despite drastic differences in the primary structure, all the members of Alu repeat superfamily preserve common tRNA-like secondary structure (Fig. 2). This peculiarity of Alu repeats is of special interest, since the hairpin-forming DNA/RNA structures are functionally important for the 3D structure of chromatin: they are involved in binding HMG proteins, recombination and modification enzymes, in splicing, and regulation of various genetic processes (for example, termination and delays in transcription, selective blocking of mRNA transcription that depends on the secondary structure of the leader sequence. etc.). Hence, the increasing volume of data concerning the role of Alu repeats in regulation of cell differentiation and tissue-specific gene expression are not unexpected.

 **RIGHT ALU=G**

 **G-A**

 **T T**

 **G\*C**

 **C\*G**

 **| C**

 **C\*G**

 **G\*C**

 **A |**

 **G\*C**

 **T\*A**

 **G\*C**

 **A\*T**

 **C\*G**

 **G\*C**

 **LEFT ALU= A T\*A**

 **INSERTION C |**

 **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_G |**

 **G G\*C**

 **C A A\*T**

 **T-G 5' G\*C**

 **T\*A \ C |**

 **5' G\*C T T |**

 **\ A-C T G\*C**

 **G | A A G A**

 **G | G G A G**

 **C\*G\*C============================C\*G\*C**

 **C\*G\*C============================C\*G\*C**

 **| A\*T----------------------------| G-T**

 **G\*C\*G============================G\*C\*G**

 **G\*C\*G============================G\*C\*G**

 **G\*C\*G============================G\*C\*G**

 **C\*G\*C============================C\*G\*C**

 **| A A | A G**

 **| | A | | A**

 **G G\*C----------------------------G G\*C**

 **C T\*A----------------------------T T\*A**

 **| | T | | G**

 **G-T\*A============================G-T\*A**

 **G\*C\*G============================G\*C\*G**

 **T\*A\*T----------------------------T-G\*C**

 **G\*C\*G============================G\*C\*G**

 **G-T\*A----------------------------G-T\*A**

 **| A A | A G**

 **| G A | G A**

 **C\*G\*C============================C\*G\*C**

 **T\*A | G A |**

 **C\*G\*C============================C\*G\*C**

 **A G\*C----------------------------G G\*C**

 **C\*G\*C============================C-A\*T**

 **G\*C\*G============================G\*C\*G**

 **C\*G-T----------------------------C\*G-T**

 **C\*G\*C============================C\*G\*C**

 **T\*A\*T============================T\*A\*T**

 **| G\*C----------------------------| G\*C**

 **G\*C T G T\*A**

 **T C ACAAAAAATACAAAAA-3' T C AAAAAAAAAA-3'**

 **A G A G**

 **A G G G**

 **T\*A==============================T\*A**

 **C\*G==============================C\*G**

 **C\*G==============================C\*G**

 **C\*G==============================C\*G**

 **| | | C**

 **A\*T==============================A\*T**

 **G-T------------------------------G\*C**

 **C T C A**

Figure 1. Triplex structures of the left (А:Т:А) and right (G:C:G) monomers of Alu-J repeat (according to the classification of Jurka [2]) with invariant paired structures (=) and compensatory mutations (-). The right monomer contains an insertion of 31 nt. (\*), complementary pairs; (-) GT pairs.

**5'**

 **\ CCA-3'**

 **G\*C**

 **G-T**

 **C\*G**

 **| A**

 **| T**

 **C\*G**

 **---#G-T#**

 **| #G\*C#**

 **| #G\*C#**

 **| #C\*G# -22.0 kcal/mol**

 **| G\*C**

 **A-box | C\*G**

 **| G\*C**

 **| G G**

 **| T C**

 **| G G-----A box**

 **----G G |**

 **-31.3 kcal/mol | T | -18.6 kcal/mol**

 **| G |**

 **CG #### | G #### TAAAAAC**

 **A ACCCTAATG-TCCGCACTC TGCGGGCC--GAT A**

 **| \*\*\*\*\* \*\*\*\*\* \*\*\* \*\* \*\*\*\* \*\*\* |**

 **C TGGGAGGCCGAGGCGGGAG AC-CCCGTCTCTA T**

 **TT #### G | #### CAAAAAA**

 **A A**

 **T\*A**

 **C\*G**

 **A\*T**

 **C\*G**

 **T\*A**

 **| T**

 **T\*A**

 **G\*C**

 **A A**

 **| A**

 **#G\*C#**

 **#C\*G#**

 **#C\*G#**

 **#C\*G#**

 **A\*T**

 **G\*C -35.5 kcal/mol**

 **G\*C**

 **---A G**

 **| G A---**

 **| T C |**

 **| T C |**

 **| C A |**

 **| G G |**

 **| A |**

 **|\_\_\_\_\_\_\_\_\_\_\_|**

 **B box**

Figure 2. tRNA like structure of the left (L) and a part of the right monomer: core sequences are indicated with ###; two A boxes and one B box, which are the recognition sites of RNA polymerase III, are shown. Energies of interaction for the paired regions are --31.3, --35.5, --18.6, and --22.0 kcal/mol, respectively. B box corresponds to anticodon loop of tRNA; poly() linker between L and R monomers, to  loop.

**2. Materials and Methods**

Secondary structures of the Alu repeats were calculated with the PCFOLD (PC/Gene, Release 6.70) program and a new SSP module of the Alignment Service software [1]. Compensatory mutations in the simulated invariant DNA/RNA secondary structures were mapped using the program SSP-INVAR (S. Resenchuk and V. Blinov, unpublished). Sequences were aligned using modified Alignment Service software [1]. Nucleotide sequences were compared with the sequences from the EMBL Nucleotide Sequence Database and GenBank using FASTA/EMBL SCAN (1993, volume II) and IMAGE (S. Denisov and V. Blinov, unpublished).

**3. Results and Discussion**

To understand the functional role of Alu repeats, it is necessary to consider first the basic principles of chromatin organization and gene structure. Amplified copies of Alu repeats in nontranscribed DNA regions, on the one hand, and exons and introns, on the other, differ in point mutations. Origination of the point mutations in Alu repeats can be explained by selection at the level of DNA/RNA secondary structures due to the so-called compensatory mutations (Fig. 1); origination of the compensatory mutations, in turn, is provided for by elements of invariant RNA/DNA secondary structures of these Alu repeats.

In this work we have demonstrated that such invariant secondary structures occurs actually both in promoter regions and introns of several genes.

Elementary units of the human genome-Alu repeats-have a complex structure and are not only retrotransposons but also rigid carcass elements likely to participate in alternative combinatorial evolution of the primate genome.

The evolution of dimeric (LR) Alu repeats is supposed to begin in the primates. Retroposition of individual (LR) Alu transposons though reverse transcription is likely to be the major expansion pathway of Alu repeats. A great number of recent publications on the Alu clusters in human genome indicate recombination as another major contributor to proliferation/expansion of Alu repeats. Due to these recombination events, in addition to the canonical (LR) Alu repeats, the noncannonical forms of Alu repeats occur in the genome: monomeric L and R; dimeric LL, RR, RL, and LR; trimeric RLL, LRL, LLR, RRL, RLR, LRR, LLL, and RRR; and tetrameric LLRR, RRLL, LRLR, and RLRL. All these forms could evolve further as independent units. In addition, it was demonstrated that the monomers L and R, L and L, and R and R can form heteroduplexes, if the monomers are have an opposite orientation and are located close to one another. Expanded inverted repeats were demonstrated to play the role of silencers and enhancers.

Thus, a discrete location of Alu repeats in the genome reflects not only the traces of Alu expansion but is also likely to be connected with a novel function: alternative interaction between Alu repeats that determines the dynamic unfolding of the information contained in the genome (Alu code). In this case, not only the chromatin state is fixed depending on the type of interaction between Alu repeats, but a definite program of activation/inhibition of information transcription (due to alternative exhaustion) in the human genome is specified. A specified order of the four elements L, R, (L), and (R), each containing 150 bp and representing the direct and inverse copies of the left and right halves of Alu repeats, determines for each of the 23 human chromosomes not only their carcass but determines a higher-order coding system compared to conventional nucleotide representation A-G-C-T---a chromosome code. If we redesignate the elementary units of Alu repeats L, R, (L), and (R) into the corresponding L=**a,**(L)=**t,**R=**g,**and (R)=**c,**then in this new representation**a-g-c-t,**Alu repeats in each chromosome can be shown as**ag-ct-ct-ag-ag-ct-ag-ag-ct-ct-ct-ct...**This coding principle can be applied to any other known or new elements independently of the structure complexity of these elementary units of the genome, one necessary condition provided: each unit must have its complementary counterpart. This new**a-g-c-t**representation takes into account the fractal-like scaling symmetry of the coding levels following the type of Russian matreshka (a wooden doll with smaller ones successively fitting into larger ones) [3].

Linkage, clusterization, the distances between**a, g, c,**and**t,**and their location in either translated or nontranslated regions is a separate problem connected with a higher-order coding. Here we consider only the fine structure of the**a, g, c,**and**t**elements and their universal tRNA-like secondary structure.

Thus, the fine structure of Alu repeats at different levels can be described as follows:

1. Conventional nucleotide code**AGCT**(Scale 1 : 1), where**A**is adenine;**G,**guanine;**C,**cytosine; and**T,**thymine.
2. The**a-g-c-t**code (scale 1 : 150), where**ag**is a dimeric Alu repeat in (+) orientation and**ct**is a dimeric Alu repeat in (-) orientation.

**Acknowledgements**

The work was supported by the Russian National Program on Human Genome.

**References**

1. S.M. Resenchuk and V.M. Blinov, "Alignment Service: creation and processing of alignments of sequences of unlimited length" Comput. Appl. Biosci.,**11,**7 (1995).
2. J. Jurka, E. Zietkiewicz, and D. Labuda, "Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the Mesozoic era" Nucleic Acids Res.,**23,**170 (1995),
3. V.M. Blinov, S.M. Resenchuk, D.L. Uvarov, G.B. Chirikova, S.I. Denisov, and L.L. Kisselev, "Alu elements in human genome: invariant secondary structure of left and right monomers" Molek. biol. (Mosk.)**32,**70 (1998).