World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:2, No:8, 2008

# Bioinformatic Analysis of Retroelement-Associated Sequences in Human and Mouse Promoters

Nadezhda M. Usmanova, and Nikolai V. Tomilin

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

***Abstract*—**Mammalian genomes contain large number of retroelements (SINEs, LINEs and LTRs) which could affect expression of protein coding genes through associated transcription factor binding sites (TFBS). Activity of the retroelement-associated TFBS in many genes is confirmed experimentally but their global functional impact remains unclear. Human SINEs (Alu repeats) and mouse SINEs (B1 and B2 repeats) are known to be clustered in GC-rich gene rich genome segments consistent with the view that they can contribute to regulation of gene expression. We have shown earlier that Alu are involved in formation of cis-regulatory modules (clusters of TFBS) in human promoters, and other authors reported that Alu located near promoter CpG islands have an increased frequency of CpG dinucleotides suggesting that these Alu are undermethylated. Human Alu and mouse B1/B2 elements have an internal bipartite promoter for RNA polymerase III containing conserved sequence motif called B-box which can bind basal transcription complex TFIIIC. It has been recently shown that TFIIIC binding to B-box leads to formation of a boundary which limits spread of repressive chromatin modifications in *S. pombe*. SINE-associated B-boxes may have similar function but conservation of TFIIIC binding sites in SINEs located near mammalian promoters has not been studied earlier. Here we analysed abundance and distribution of retroelements (SINEs, LINEs and LTRs) in annotated sequences of the Database of mammalian transcription start sites (DBTSS). Fractions of SINEs in human and mouse promoters are slightly lower than in all genome but >40% of human and mouse promoters contain Alu or B1/B2 elements within -1000 to +200 bp interval relative to transcription start site (TSS). Most of these SINEs is associated with distal segments of promoters (-1000 to -200 bp relative to TSS) indicating that their insertion at distances >200 bp upstream of TSS is tolerated during evolution. Distribution of SINEs in promoters correlates negatively with the distribution of CpG sequences. Using analysis of abundance of 12-mer motifs from the B1 and Alu consensus sequences in genome and DBTSS it has been confirmed that some subsegments of Alu and B1 elements are poorly conserved which depends in part on the presence of CpG dinucleotides. One of these CpG-containing subsegments in B1 elements overlaps with SINE-associated B-box and it shows better conservation in DBTSS compared to genomic sequences. It has been also studied conservation in DBTSS and genome of the B-box containing segments of old (AluJ, AluS) and young (AluY) Alu repeats and found that CpG sequence of the B-box of old Alu is better conserved in DBTSS than in genome. This indicates that B-box-associated CpGs in promoters are better protected from methylation and mutation than B-box-associated CpGs in genomic SINEs. These results are consistent with the view that potential TFIIIC binding motifs in SINEs associated with human and mouse promoters may be functionally important. These motifs may protect promoters from repressive histone modifications which spread from adjacent sequences. This can potentially explain well known clustering of SINEs in GC-rich gene rich genome compartments and existence of unmethylated CpG islands.

Authors are with Institute of Cytology, Russian Academy of Sciences, Russia (e-mail: nvtom@mail.ru).

***Keywords*—**Retroelement, promoter, CpG island, DNA methylation.

## I. Introduction

IT has been shown earlier that human Alu repeats and mouse B1 repeats are clustered in the heavy GC-rich long DNA segments (isochores), while LINE1 (L1) repeats are mostly associated with the light AT-rich isochores [1]. FISH analysis of human chromosomes with the Alu-specific or L1-specific DNA probes also demonstrated preferential association of Alu with the GC-rich gene-rich R bands, and L1 association with AT-rich gene-poor G-bands [2]. This was confirmed by sequencing the human and mouse genomes. The most abundant and relatively old (>5 MYRA) human Alu repeats and all mouse B1/B2 repeats were found preferentially associated with GC-rich DNA, whereas L1 repeats were mostly found in AT-rich DNA [3, 4]. GC-rich genome segments usually contain high density of genes. Therefore, it was suggested that SINEs concentrate near genes because they may have a role in the control of chromatin structure and gene expression [3]. One suggested mechanism of such control is that SINE transcripts regulate protein synthesis through their interaction with of double-stranded RNA-activated kinase PKR [5] but this hypothesis does not explain why SINEs are clustered near genes. A possibility exists that SINEs control transcription initiation. It has been recently suggested that human Alu can suppress spread of DNA methylation to promoters from adjacent LINE1 elements [6].

The main target for DNA methylation is cytosine in the dinucleotide CpG. This dinucleotide is underrepresented in the most of genomic sequences except for short segments, called CpG islands [7]. CpG islands, associated with 75% of genes [8] were defined as genome segments of more than 200 bp length, having more than 50% GC content, and observed/expected CpG ratio >0.6 [7]. Most of the unmethylated and methylated CpG sequences in mammalian DNA are associated with transposons which become inactive upon methylation [9, 10]. DNA methylation is considered as a mechanism originally arised for the rapid transposon

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

epigenetic silencing [9]. Deamination of methyl-C in methylated CpG is responsible for ~30% of all mutations in germ line and somatic cells [11]. Evolutionary old retroposons of LTR and LINE families have strongly reduced CpG content, which is only 18-19% of expected [10]. Most of SINEs is also methylated. However, the most abundant human SINEs, Alu repeats, contain more than 40% of expected CpG [10]. In part, this reflects relatively late evolutionary origin of these elements [10] but is also caused by their undermethylation at some genome segments. In fact, Alu repeats are weakly methylated in human sperm [12]. 10% of unmethylated DNA in human brain is associated with SINEs [10]. Some Alu repeats overlap with promoter CpG islands [13]. Alu copies adjacent to CpG islands have an increased the observed/expected CpG ratio as compared to randomly located elements [14]. Some genomic Alu repeats also serve as starting points of RNA polymerase III-dependent transcription of microRNA gene clusters [15] indicating that these Alu copies can bind basal transcription complex TFIIIC which is required for transcription by RNA polymerase III.

Binding sites for TFIIIC can limit spread of repressive histone modification in *S. pombe* [16] and promoter-associated TFIIIC binding sites present in many SINEs may have similar function in mammalian cells. Here we studied distribution and conservation of SINE sequences in the mouse and human promoters of the Database of Transcription Start Sites (DBTSS). Our results are consistent with the view that TFIIIC bound to SINEs can protect mammalian promoters.

## II. Methods

### A. Analysis of Promoter Sequences

Human and mouse promoter sequences were downloaded form the Database of Transcription Start Sites (DBTSS), release 5.2.0 (April 4, 2006), based on UCSC hg17 and mm5 and available at (http://dbtss.hgc.jp). Retrotransposons were identified in these sequences using RepeatMasker program (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) with cross_match engine and default sensitivity. Numbers of promoters containing 12-mer sequences conserved in retrotransposons were counted using DBTSS option "Search for TF binding sites" and exact sequence pattern match. In these searches either full target window (-1000 to +200) or specific segments of promoters were studied. 12-mers conserved in retroposons were taken from consensus sequences available at the REPBASE (http://www.girinst.org/repbase/index.html) of the Genetic Information Research Institute, Mountain View, CA. Frequency of retroposon-associated 12-mers was determined using program dna-pattern (strings) available at (http://rsat.ulb.ac.be/rsat) or by direct counting of specific motifs in MS-Word using command "replace".

### B. Analysis of Genomic Sequences

Genomic sequences were obtained from the Ensembl database (http://www.ensembl.org) and analysed for fractions of different retrotransposons - using RepeatMasker (see above). Numbers of specific 12-mers in genomic sequences were counted using program dna-pattern (strings) with zero allowed substitutions available at (http://rsat.ulb.ac.be/rsat) or by direct search of motif (both strands) in MS-Word. Statistical analysis of correlations was done using program STATISTICA 6 (StatSoft Co.).

## III. Results

### A. Abundance of Retroelements in the Upsteam Promoters of Protein-Coding Genes

In order to get quantitative information regarding abundance of different retrotransposons in the upstream promoters of genes, we analyzed the Database of Transcription Start Sites, DBTSS, release 5.2.0 (http://dbtss.hgc.jp). This is based on UCSC hg17 mm5 and contains data for 15,262 human and 14,162 mouse genes. Many genes have alternative promoters and overall number of the human and mouse promoters in DBTSS is 30,964 and 19,023, respectively. Fractions of different reproposons in the human and mouse promoters estimated using RepeatMasker web server are shown in Fig. 1. As expected, fractions of LINEs are strongly underrepresented in promoters which are mainly caused by deficit of L1 sequences. LTRs are also underrepresented with deficit of MaLRs and ERV_class II elements in both human and mouse promoters. ERVL and ERV_class I elements are similarly underrepresented in the mouse DBTSS. A decreased density of some LTRs and L1 sequences in the 5' regions of human genes was noted earlier [17], and it was suggested that L1 and LTR insertions into promoters are negatively selected. However, 7.44% of human promoter sequences are still represented by LINEs and LTRs. This number is close to the overall fraction of Alu repeats (7.96%). Mouse DBTSS have 4.34% of LINEs plus LTRs and 6.35% of B1 plus B2 sequences (Fig. 1).

Besides information about relative length of retroposon sequences Repeat Masker also provides data on the numbers of specific retroelements. In the human DBTSS the number of Alu elements is estimated as 13269, MIR elements – 8264 and L1 elements – 3095. We estimated that ~5% of promoters have more than one Alu element indicating that >12000 (~40%) of all annotated human promoters have at least one Alu sequence. In the mouse DBTSS numbers of B1, B2 and L1 elements are 6160, 5010 and 1733, resp., and fraction of promoters with more than one B1 or B2 elements is <3% indicating that >11000 (~60%) of all annotated mouse promoters have at least one B1 or B2 element.

### B. Analysis of the Retroelement-Asociated Short Sequence Motifs

To identify promoters containing SINEs we analyzed presence in the DBTSS of short sequence motifs known to be strongly conserved in retroposons. For example, motif CACTTTGGGAGG is strongly conserved in all subfamilies of human Alu repeats [18], and motif CTTTAATCCCAG is frequently associated with the mouse B1 repeats (Repbase). To estimate minimal length of these motifs required for unambiguous identification of retroposons we scanned DBTSS with overlapping 7 to 18-mers taken from different segments of the AluSx and B1_Mm consensus sequences of the Repbase, as well as retroposon-unrelated sequence

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

(ATGC)n. It is seen from Fig. 1 that retroposon-unrelated motifs having length >10 nt more detect very few targets in DBTSS but two Alu consensus- and B1-consensus derived blocks having the length >10 nt detect large number of targets in the human and mouse DBTSS, resp.

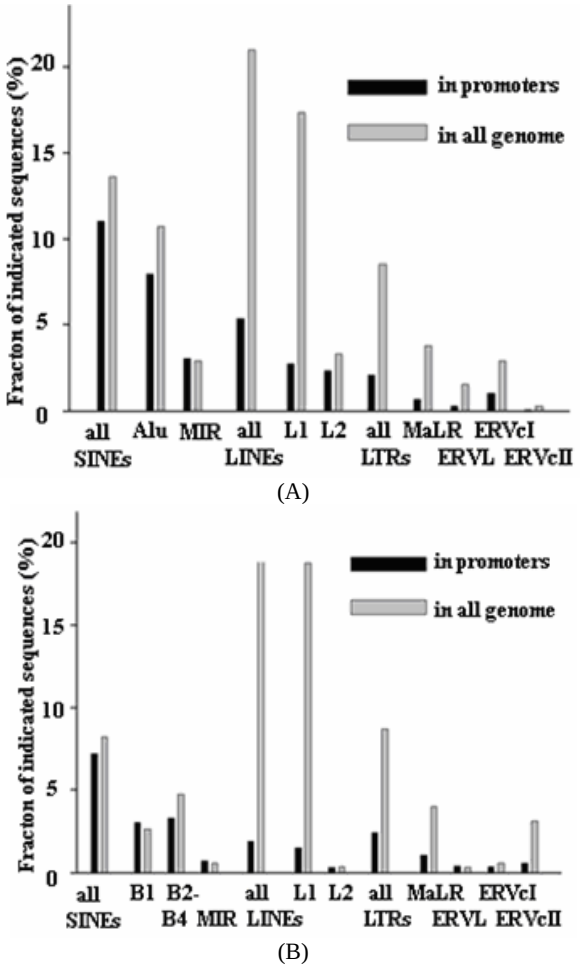



Fig. 1 Retroelement-derived sequences in the human and mouse promoters. Fractions of retroelements in the mouse and human genomes were taken from [4]. Fractions of retroelements in promoters were estimated using RepeatMasker in the sequences downloaded from the DBTSS

The number of targets detected using different B1- or Alu-derived 12-mer blocks varies (Fig. 2) but for a specific block an increase of its length weakly effects the number of detected targets. This indicates that 12-mers associated with Alu- or B1-consensus reliably detect corresponding retroposons in DBTSS but probability of detection of these elements by different 12-mers significantly varies. Apparently, these variations are caused by differential conservation of retroposon subsequences during evolution which is well known for human Alu repeats [19].

To estimate amplitude of this variation we examined DBTSS with different 12-mers from human and mouse SINE consensus sequences using two approaches in one of which numbers of hits (promoters) for a given 12-mer in DBTSS was estimated, and in another approach numbers of the motifs were directly counted in DBTSS sequences (Tables 1 and 2). Both approaches show similar results and indicate that more than 10-fold variations of numbers of different 12-mers from the same retroposon family can be detected in the DBTSS. Slightly higher numbers of motifs detected in the mouse DBTSS using the second approach is explained by the presence of several identical motifs in one promoter (Table I).

TABLE I
ABUNDANCE OF RETROPOSON-ASSOCIATED MOTIFS IN MOUSE DBTSS*

| Motif | Association with indicated retroposon family | Number of hits detected with this motif in DBTSS | Number of motifs found in DBTSS sequences (22.6 Mb) | Number of motifs found in genomic sequence (90.7 Mb) |
|---|---|---|---|---|
| ctaatttccgca | unassociated | 1 | 1 | 2 |
| | | | | |
| ggtgg**cg**ca**cg**c | B1_Mm | 315 | 317 | 484 |
| a**cg**cctttaatc | “ | 628 | 646 | 849 |
| ccagcact**cg**gg | “ | 294 | 296 | 651 |
| ggaggcagaggc | “ | 1236 | 1307 | 3068 |
| ggatttctgagt | “ | 732 | 748 | 1571 |
| ccagcctggtct | “ | 1239 | 1319 | 2964 |
| ctacaaagtgag | “ | 503 | 514 | 1437 |
| aggacagccagg | “ | 1015 | 1062 | 3467 |
| ccctgtct**cg**aa | “ | 518 | 529 | 935 |
| gagtt**cg**aggcc | “ | 742 | 758 | 1394 |
| aggcagaggcag | “ | 1325 | 1478 | 3952 |
| | | | | |
| agatggctcagt | B2_Mm1a | 373 | 379 | 1419 |
| gtgggtaagagc | “ | 116 | 117 | 110 |
| gctcttc**cg**aag | “ | 118 | 120 | 342 |
| ttcaaatcccag | “ | 339 | 345 | 1016 |
| agcaaccacatg | “ | 466 | 484 | 1520 |
| tggtggctcaca | “ | 518 | 532 | 873 |
| caaccatc**cg**ta | “ | 112 | 112 | 236 |
| tctgactccctc | “ | 88 | 90 | 253 |
| gctacagtgtac | “ | 329 | 346 | 963 |
| | | | | |
| atggaaggagtt | L1_MM | 26 | 26 | 1160 |

* Total number of B1 and B2 retroelements in mouse DBTSS sequences was estimated using RepeatMasker: B1 – 6160, B2 – 5010, L1 – 1733. CpG sequences are shown in bold.

Slightly higher numbers of promoters detected in the human DBTSS using the first approach (Table II) is explained by overlap of alternative promoters abundant in human genome when the same motif is counted several times in different promoters. Variations of frequency of different 12-mers of the same retroposon were also found in full genomic sequences (last row in Tables I and II) indicating that variations of retroposon sequences are not specific for promoters but are also typical for retroposons located in introns and intergenic spacers. Together, these results indicate that non-random variation of retroposon subsequences does exist in human and mouse genomes and in DBTSS. In part, lower conservation of some subsequences can be explained by the presence of CpG dinucleotides which can be methylated and lead to frequent CpG to TpG mutations. However, some 12-mers from

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

TABLE II
ABUNDANCE OF RETROPOSON-ASSOCIATED MOTIFS IN HUMAN DBTSS*

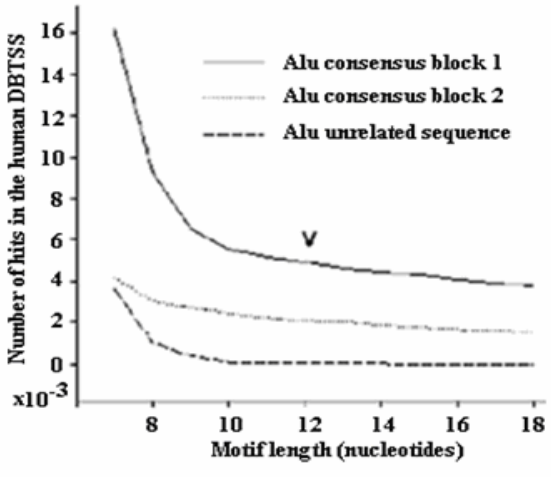
| Motif | Association with indicated retroposon family | Number of hits detected with this motif in DBTSS | Number of motifs found in DBTSS sequences (36.2 Mb) | Number of motifs found in genomic sequence (76.1 Mb) |
|---|---|---|---|---|
| GGAGGGTTTCAC | unassociated | 7 | 6 | 4 |
| | | | | |
| G**CG**C**G**GTGGCTC | AluSx | 1424 | 1298 | 1142 |
| TGGCTCA**CG**CCT | " | 3067 | 2701 | 2783 |
| CACTTTGGGAGG | " | 5564 | 5105 | 8351 |
| C**CG**AGG**CG**GG**CG** | " | 624 | 558 | 470 |
| CTGAGGTCAGGA | " | 1616 | 1346 | 2157 |
| GGTGG**CG**CATGC | " | 316 | 264 | 279 |
| ACT**CG**GGAGGCT | " | 2062 | 1808 | 2509 |
| GGGAGG**CG**GAGC | " | 543 | 456 | 817 |
| ACTCCAGCCTGG | " | 5245 | 4817 | 8035 |
| GAGTT**CG**AGACC | " | 2062 | 1791 | 2048 |
| GAGG**CG**GAGCTT | " | 625 | 525 | 1020 |
| | | | | |
| GAGGATGTGGAG | L1_HS | 65 | 58 | 475 |

* Total number of Alu repeats estimated in the human DBTSS using RepetMasker is 13269, total number of L1 repeats - 3095. All listed 12-mers detect only few hits in the whole *C.elegans* genome (~100 Mb). CpG sequences are shown in bold.

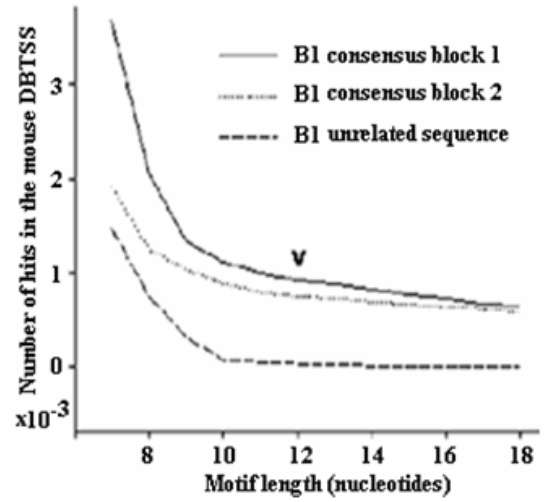
consensus sequences without CpG also show low conservation (Tables I and II). It is important to note that randomly selected 12-mers not associated with SINEs detect only few targets in DBTSS and that direct examination of targets detected using retroposon-associated 12-mers showed that very few non-retroposon targets (<3%) are matched by these motifs. In random sequence any 12-mer is expected for one occurence in $4^{12}$ nt or ~ 16 Mb, and in human DBTSS which is ~36 Mb any 12-mer can occur only two times. We found that a retroposon-unassociated 12-mer is repeated 4 times in the human DBTSS (Table II).

We studied relative distribution of some retroposon-associated conserved motifs along promoters. The number of promoters with these motifs in the vicinity (+/- 200 nt) of TSS is low but is gradually increased in the upstream 900 nt (Fig. 3). 12-mers from the consensus sequences of human (L1HS) and mouse (L1_Mm) LINE1 repeats detect few targets in DBTSS (Fig. 3) which is a consequence of their higher mutational divergence. Apparently, SINE insertions into sequences immediately adjacent to the first exons are negatively selected during evolution. At the same time, insertions of these elements into segments >200 nt upstream of TSS are tolerated, and fixed in many mammalian genes.

We also analysed relative distribution in promoters of dinucleotides CpG and found that the number of promoters with CpG in 100 nt intervals is low in upstream segments and is maximal at +100 nt relative to TSS (Fig. 4AB) which probably reflects high concentration of CpG islands around




(A)




(B)

Fig. 2 Detection of retroposon-derived sequence motifs in the human (A) and mouse (B) DBTSS using short motifs from retroposon consensus sequences. AluSx consensus block 1 was CCCAGCACTTTGGGAGGC started from 5' end 7-mer CCCAGCA to which one nucleotide was sequentially added to 3' end. AluSx consensus block 2 was GAGTTCGAGACCAGCCT started from 7-mer GAGTTCG, and retroposon unrelated sequence was $(ATGC)_4AT$ started from 7-mer ATGCATG. B1_Mm consensus block 1 was GCCTTTAATCCCAGCACT started from 7-mer GCCTTTA, B1 consensus block 2 was GAGTTCGAGGCCAGCCTG started from 7-mer GAGTTCG. (v) marks 12-mers.

TSS. Distribution of CpG sequences mirrors distribution of SINE-associated 12-mers which is apparent from strong negative correlation between Alu- and B1-associated motifs and CpG (Fig. 5AB). It appears, therefore, that in a major fraction of promoter SINEs concentrate upstream of segments having high CpG density. Majority of these segments is apparently associated with CpG islands.

### *C. Conservation of SINE Subsequences in Promoters*

To reveal possible variability of retroposon subsequences in DBTSS and genome in a systematic way we scanned this database and genomic sequences with 12-mers covering full

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

consensus sequence of B1 elements (Repbase B1_Mm) with 2-nucleotide shift. Resulting curves were normalyzed taking obtained values for the conserved B1 12-mer AGGCAGAGGCAG as 1. It is seen from Fig. 6 that conservation of B1 elements is very uneven both in genomic and DBTSS sequences and that there are highly conserved segments (which are seen as peaks) as well as highly variable segments. Conserved subsegments (peaks) do not overlap with CpG sequences (Fig. 6, marked by *) indicating that deamination of meC in CpG contributes significantly to variability of corresponding motifs. However, some subsegments (80-90 nt, 110-120 nt) show low conservation in the absence of CpG possibly reflecting existence of unidentified B1 subfamilies with diagnostic differences in these subsegments. Examination of the two obtained distributions (for genome and DBTSS) using paired t-test showed significant difference between them (T value -6.1352031469, and P value 0.0000000566) which is seen, for example, in 60 – 80 nt interval (Fig. 6). This interval covers well known motif for RNA polymerase III transcription (B-box, GTTCGAGRC) which binds transcription complex TFIIIC and contains CpG sequence [22-24]. Poor conservation of this segment in genome is mostly caused by CpG to TpG mutation after cytosine methylation. We directly estimated fraction of B1 repeats containing CpG and TpG in their B-boxes using searches with diagnostic motifs (CTGAGTT**CG**AGG and CTGAGTTTGAGG, resp.) and found that in DBTSS TpG is present in 25% of B1 repeats and in genome TpG is found in 36% of B1 repeats. This estimation confirms that B-box-associated CpGs are better conserved in DBTSS.

We performed similar analysis of Alu sequences and found that low conservation of some segments may be caused not only by CpG mutations (Table II) but also by variations between known consensus sequences of different Alu subfamilies having different evolutionary age. Therefore, we analysed variations of the B-box containing subsegments of specific Alu subfamilies in genome and DBTSS (Table III). Diagnostic motifs for specific Alu subfamilies (see legend to Table III) were deduced after alignment of consensus sequences of Alu subfamilies taken from Repbase. We found that members of the evolutionary old AluJ subfamily in genome have much more CpG to TpG mutations (71.2%) in B-box than evolutionary young AluY (17.9%) and combined AluS+J subfamilies show intermediate value (52.1%) which is expected result taking into account their evolutionary age. In DBTSS fraction of CpG to TpG mutations was found to be significantly lower compared to genome in AluS+J and AluJ repeats (Table III) and also is slightly lower in AluY. These results indicate that CpG sequences located in the Alu B-box segments of all Alu subfamilies are better conserved in promoters than in genome suggesting their functional role.

TABLE III
CONSERVATION OF SEQUENCES COVERING B-BOX IN DIFFERENT ALU SUBFAMILIES*

| Target sequences from | B-box motif diagnostic for indicated Alu subfamily | % of hits with complete match to consensus motif | % of hits with CpG to TpG mutation |
|---|---|---|---|
| Genome | S + J | 47.9 | 52.1 |
| DBTSS | S + J | 61.2 | 38.8 |
| Genome | Y | 82.0 | 17.9 |
| DBTSS | Y | 83.8 | 16.2 |
| Genome | J | 28.8 | 71.2 |
| DBTSS | J | 43.1 | 56.9 |

*Diagnostic motifs for AluS+AluJ – GAGTT**CG**AGACC, for AluY – AGGAGAT**CG**AGACC, for AluJ – CCAGGAGTT**CG**A. CpG to TpG mutations in target sequences were detected using the same motifs with replacement of CpG (shown in bold) for TpG.

This may be TFIIIC-dependent barrier function for spreading of heterochromatin from adjacent segments suggested for *S.pombe* [16]. As was pointed out above, density of promoter SINEs and, in particular, TFIIIC-binding motifs, negatively correlates with CpG density (Fig. 4) and, therefore, many SINEs are located upstream of CpG islands. This supports the view that they can protect CpG islands from methylation spread from adjacent retroposons L1 and LTR families [6].

## IV. DISCUSSION

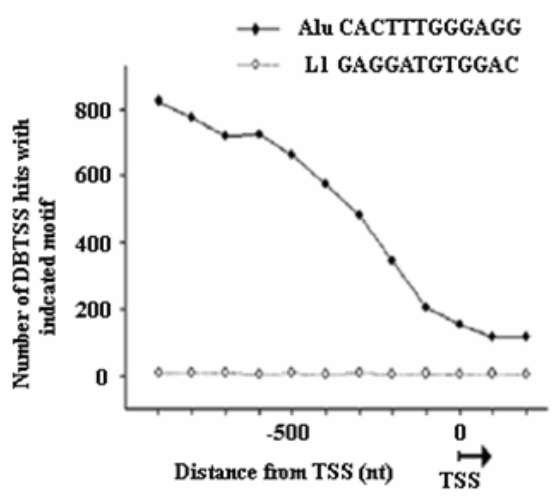
Here we found that major fractions (>40%) of human and mouse promoters contain Alu or B1/B2 sequences. Further, these retrotransposons are unevenly distributed within promoters, and concentrate in the upstream promoter segments (-1000 to -200 relative to TSS). Apparently, presence of SINEs very close to the TSS negatively effects initiation of transcription by RNA polymerase II. Many developmental genes cannot tolerate transposons not only in their proximal 5' promoters but also in the other non-coding segments. In human genome total size of transposon-free regions over 10 kb in length is about 12 Mb, and these regions are strongly enriched with microRNA genes [20]. However, it is unlikely that SINEs are excluded from some regions because they interfere with expression of miRNA genes, since appoximately 20% of miRNA is efficiently transcribed by RNA polymerase III starting from Alu sequences [15].

Alu repeats adjacent to CpG islands show higher conservation of CpG sequences [14] indicating that many of them remain unmethylated. Here we studied sequence divergence in some SINEs (Alu, B1 and B2 repeats) in the DBTSS and in genomic sequences and found that sequence variability is very unevenly distributed along these elements. Uneven sequence divergence along human Alu repeats was detected long time ago and interpreted as indicative of evolutionary selection against mutational changes in Alu subsegments interacting with proteins [19]. Many mammalian transcription factors interact with SINE *in vitro* [21] but SINE subsequences highly conserved in genome and in DBTSS appear to overlap with subsegments free of CpG dinucleotides (Figs. 4 and 5) indicating that SINE methylation also plays

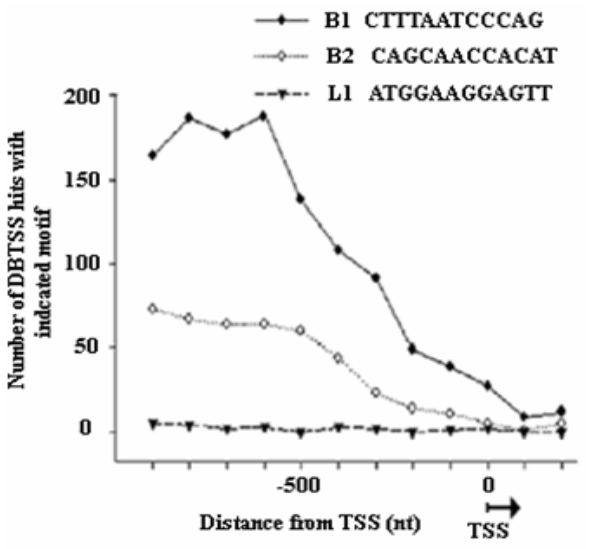
major role in sequence variability. In principle, CpG free SINE subsegments can bind transcription factors even in methylated SINE copies but this seems unlikely because of spread of repressive histone modifications usually correlating with DNA methylation.

We also found that some SINE subsegments are better conserved in promoters than in genomic sequences. One of these subsegments containing CpG is located in the retroposon B-boxes (common motif GTTCGAGRC) which bind transcription complex TFIIIC [22-23]. B-boxes also can bind another nuclear protein which is downregulated in adenovirus-transformed cells [24]. TFIIIC is essential for transcription of the B-box containing genes by RNA polymerase III [25-26]. Methylation of these genes suppresses transcription [27] possibly because of methylation of CpG in their B-box. Therefore, only SINEs in which CpG in B-box is not methylated are likely to be able to bind TFIIIC.

Binding of TFIIIC usually leads to activation of transcription of retroposons by RNA polymerase III. However, recent data indicate that in *S. pombe* TFIIIC binding sites may have another function which is independent of RNA polymerase III [16]. This function is to prevent spreading of repressive histone modifications into neighboring euchromatic regions by recruiting TFIIIC complex, and SINEs are suggested to have similar function in mammalian cells [16]. Our results are consistent with the view that SINE-containing unmethylated binding sites for TFIIIC may serve as functional elements preventing spread of heterochromatin from adjacent non-coding sequences e.g. from heavily methylated LINEs and LTRs [6]. This view is supported by conservation of CpG in Alu sequences adjacent to CpG islands [14], by global clustering of SINEs in GC-rich genome regions containing CpG islands [1-4] and by better conservation of potential TFIIIC binding sites in promoter SINEs. An evidence for a suppressive effect of the Alu on the spread of L1 methylation was also obtained earlier by comparison of the length of the CpG islands and the density of adjacent Alu elements [6], though possible mechanism for this suppression was not suggested. Here, we suggest a novel mechanism of protection of CpG islands by SINEs. Our results are also compatible with another hypothesis in which transcription of promoter-associated retroposons by RNA polymerase III (Pol III) and/or resulting transcripts facilitates assembly of transcription complexes of RNA polymerase II (Pol II). Pol III transcription units can produce transcripts regulating Pol II transcription [28]. Possible regulatory role of nongenic transcription units has been discussed earlier [29]. These units represent about 2/3 of all transcription units in mammalian cells [30-31].




(A)

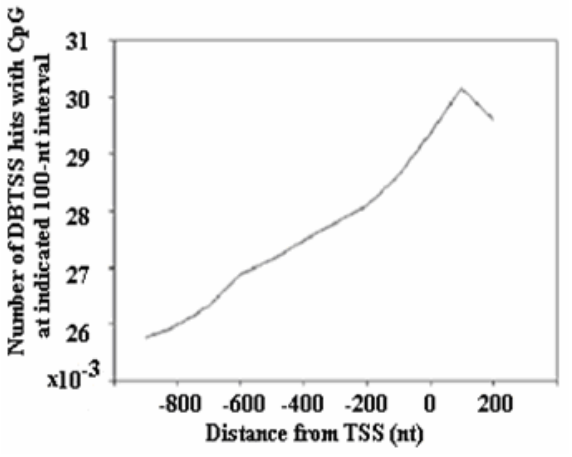



(B)

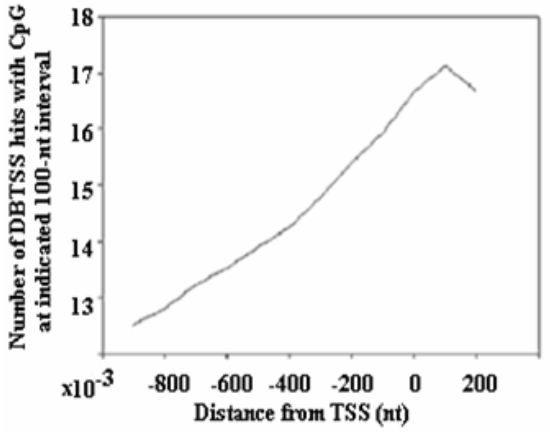
Fig. 3 Uneven distribution of some retroposon-derived motifs within human (A) and mouse (B) promoters of the DBTSS. Transcription Start Sites (TSS) are shown by horizontal arrows. 12-mers are from consensus sequences of AluSx, L1Hs, B1_Mm and L1_MM (Repbase). Numbers of promoters having indicated 12-mer was estimated by scanning DBTSS (option Search for TF binding sites) in 100 nt windows (from -1000 to 200 nt) with this 12-mer motif

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf




(A) H. Sapiens




(B) M. Musculus

Fig. 4 Relative distribution of CpG dinucleotides in the human (A) and mouse (B) promoters. Numbers of promoters having CpG sequence was estimated by scanning DBTSS (option Search for TF binding sites) in 100 nt windows (from -1000 to 200 nt) with motif NCGN

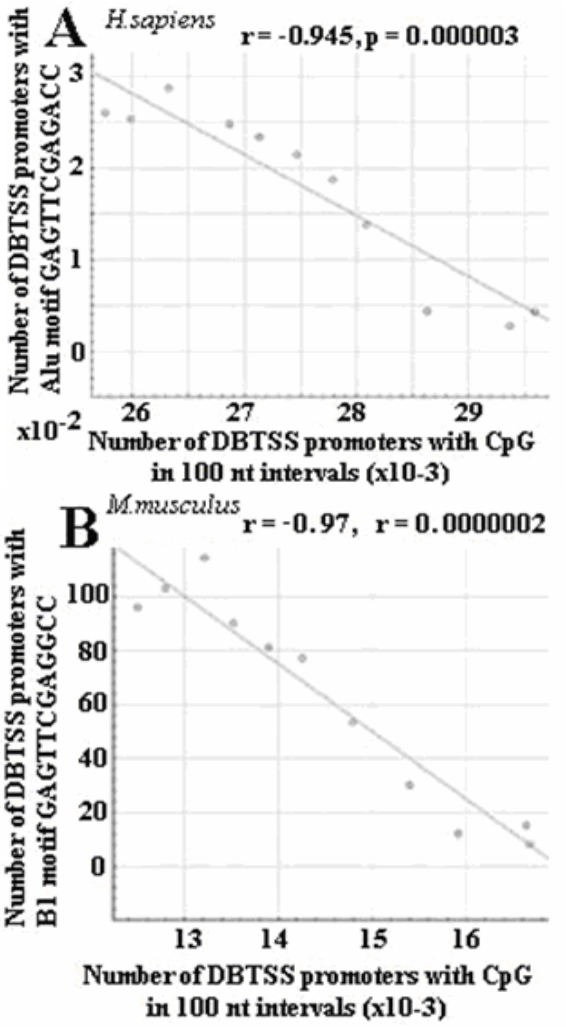



Fig. 5 Correlation between distributions of CpG nucleotides and SINE-associated 12-mer motifs in the human and mouse promoters. Distributions were estimated as described in the legends to Figs. 2 and 3. Correlation was analysed using program STATISTICA

It is well known that SINEs are clustered in GC-rich gene rich genome regions [1-4] which may be explained by their role in the control of chromatin structure and gene expression [3]. Another explanation for SINE clustering in gene-rich regions is that GC-rich SINEs are more stable in isochores, where the surrounding GC-content is similar [32]. However, Alu stability of various age groups is uniform and is irrelevant of GC occupancy [17]. On the other hand, Alu are absent from 245 human transposon-free regions having high (>57%) GC content [20] indicating that GC content of target DNA is not main determinant of the rate Alu fixation during evolution.

We believe that some functional benefits of location of SINEs near genes are more important than GC content. Mammalian retroposons are involved in the control of gene expression [19, 21, 33] and substantial fraction of human regulatory sequences is originated from transposable elements [34]. Alu repeats significantly contributed to regulation of human genes by retinoic acid [35] and by estrogens [36]. Results obtained here are consistent with the view that SINEs can also function to limit spread of repressive histone modifications [16]. Such

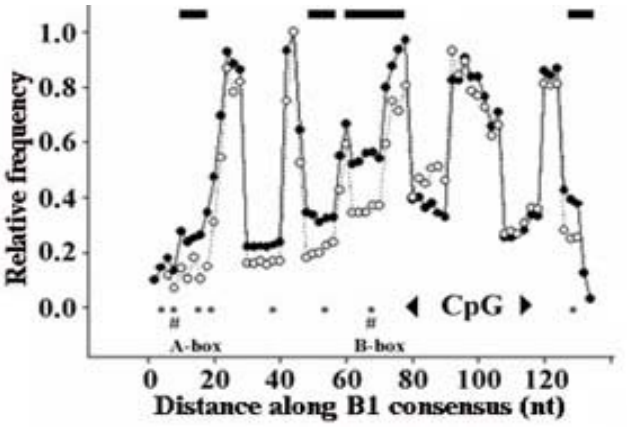



Fig. 6 Conservation of the retroposon B1 subsequences in mouse promoters and in genome (open circles). Mouse DBTSS was scanned with sequential 12-mers of the B1_Mm consensus sequence (Repbase) with 2-nucleotide shift and relative number of promoters with specific 12-mer determined (filled circles). Mouse genomic sequences (total 10 Mb) were randomly selected from chromosome 18 and relative frequency of the same 12-mers was determined in these sequences using program dna-pattern (strings) at http://rsat.ulb.ac.be/rsat (open circles). Horizontal black bars show segments better conserved in promoters, (*) show location of CpG sequences and (#) mark A- and B-boxes of this retroposon.

mechanism is apparent for permanently active housekeeping genes, but tissue-specific genes should be silenced in specific types of cells. This is difficult to achieve by downregulation of TFIIIC, since it is required for continuous synthesis of tRNA and proteins. The binding of TFIIIC to retroposons located in some human promoters can be prevented through competitive tissue-specific binding of the transcription repressor YY1 [37] which binding site is located just downstream from the Alu B-box [13,24]. During mouse development transcription of the growth hormone (GH) gene by RNA polymerase II is activated through the binding of transcription complex TFIIIC to the retroposon B2 located upstream of GH gene [38] but factors stimulating TFIIIC binding to this B2 are not identified. Possible regulatory mechanisms were discussed earlier [39].

## V. Conclusion

Our results indicate that B-box associated CpGs in the human and mouse Alu and B1 retrotransposons, located in promoters of protein-coding genes, are better conserved as compared to that from genomic sequences. This indicates that in promoters they are protected from methylation and, therefore, may have a function. In the future, genes may be identified, which expression depends on Alu- or B1-associated binding site for TFIIIC in their promoters. CpG methylation-dependent binding of TFIIIC to promoter-associated SINEs should be studied further, as well as global distribution of TFIIIC bound to chromatin in the mammalian nucleus.

## List of Abbreviations

TSS, transcription start site; SINEs, short interspersed repeat elements; Alu, major family of human SINEs cleaved by restriction nuclease *Alu*I; B1, B2, ID, mouse families of interspersed repeats; MIRs, mammalian-wide interspersed repeats; LINEs, long interspersed repeat elements; L1, major family of mammalian LINEs; L2, CR1/L3, minor families of mammalian LINEs; LTRs, retroelements containing long terminal repeats; MaLR, mammalian apparent LTR retrotransposon superfamily; ERV_class I, defective endogenous retrovirus similar to type C or γ-retroviruses, ERV_class II, defective endogenous retrovirus similar to type B or β-retrovirus; ERVL, class III of endogenous retroviruses having limited similarity to spuma retroviruses; MYRA, millions of years ago.

## Acknowledgment

Authors thank Dr. Maria G. Samsonova for helpful discussions.

## References

[1] Soriano P, Meunier-Rotival M, Bernardi G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A* 1983, 80:1816-1820.
[2] Korenberg JR, Rykowski MC: Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 1988, 53: 391-400.
[3] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: Initial sequencing and analysis of the human genome. *Nature* 2001, 409: 860–921.
[4] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, 420: 520-562.
[5] Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW: Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol Cell Biol* 1998, 18:58-68.
[6] Kang MI, Rhyu MG, Kim YH, Jung YC, Hong SJ, Cho CS, Kim HS: The length of CpG islands is associated with the distribution of *Alu* and *L1* retroelements. *Genomics* 2006, 87: 580-590.
[7] Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes. *J Mol Biol* 1987, 196: 261–282.
[8] Ioshikhes IP, Zhang MQ: Large-scale human promoter mapping using CpG islands. *Nat Genet* 2000, 26: 61–63.
[9] Yoder JA, Walsch CP, Bestor TH: Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 1997, 13: 335-340.
[10] Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH: Large-scale structure of genomic methylation patterns. *Genome Res* 2006, 16: 157-163.
[11] Kondrashov AS: Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 2003, 21: 12-27.
[12] Rubin CM, VandeVoort CA, Teplitz RL, Schmid CW: *Alu* repeated DNAs are differentially methylated in primate germ cells. *Nucl Acids Res* 1994, 22: 5121-5127.
[13] Oei SL, Babich VS, Kazakov VI, Usmanova NM, Kropotov AV, Tomilin NV: Clusters of regulatory signals for RNA polymerase II transcription associated with *Alu* family repeats and CpG islands in human promoters. *Genomics* 2004, 83: 873-882.
[14] Brohede J, Rand KN: Evolutionary evidence suggests that CpG island-associated *Alu*s are frequently unmethylated in human germline. *Hum Genet* 2006, 119: 457-458.
[15] Borchert GM, Lanier W, Davidson BL: RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* 2006: 13: 1097-1101.

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf

[16] Noma K, Cam HP, Maraia RJ, Grewal SI: A role for TFIIIC transcription factor complex in genome organization. *Cell* 2006, 125: 859-872.
[17] Medstrand P, van de Lagemaat LN, Mager DL: Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 2002, 12: 1483-1495.
[18] Babich V, Aksenov N, Alexeenko V, Oei SL, Buchlow G, Tomilin N: Association of some potential hormone response elements in human genes with the *Alu* family repeats. *Gene* 1999, 239: 341-349.
[19] Britten RJ. Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. *Proc Natl Acad Sci U S A* 1994, 91: 5992-5996.
[20] Simons C, Pheasant M, Makunin IV, Mattick JS: Transposon-free regions in mammalian genomes.*Genome Res* 2006, 16: 164-172.
[21] Tomilin NV: Control of genes by mammalian retroposons. *Int Rev Cytol* 1999, 186: 1-48.
[22] Kochanek S, Renz D, Doerfler W: Probing DNA-protein interactions in vitro with the CpG DNA methyltransferase. *Nucl Acids Res* 1995, 21: 2339-2342.
[23] Chu WM, Wang Z, Roeder RG, Schmid CW: RNA polymerase III transcription repressed by Rb through its interactions with TFIIIB and TFIIIC2. *J Biol Chem* 1997, 272: 14755-14761.
[24] Kropotov AV, Tomilin NV: A human B-box-binding protein downregulated in adenovirus 5-transformed human cells. *FEBS Lett* 1996, 386: 43-46.
[25] Van Dyke MW, Roeder RG: Multiple proteins bind to VA RNA genes of adenovirus type 2. *Mol Cell Biol* 1987, 7: 1021-1031.
[26] Geiduschek EP, Tocchini-Valentini GP: Transcription by RNA polymerase III. *Annu. Rev. Biochem.* 1988, 57: 873-914.
[27] Besser D, Gotz F, Schulze-Forster K, Wagner H, Kroger H, Simon D: DNA methylation inhibits transcription by RNA polymerase III of a tRNA gene, but not of a 5S rRNA gene. *FEBS Lett* 1990, 269: 358-362.
[28] Pagano A, Castelnuovo M, Tortelli F, Ferrari R, Dieci G, Cancedda R: New Small Nuclear RNA Gene-Like Transcriptional Units as Sources of Regulatory Transcripts. *PLoS Genet.* 2007, 3: e1.
[29] Cook PR: Nongenic transcription, gene regulation and action at a distance. *J. Cell Sci.* 2003, 116: 4483-4491.
[30] Mattick JS: Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 2003, 25: 930–939.
[31] Carninci P, Hayashizaki Y: Noncoding RNA transcription beyond annotated genes. *Curr Opin Genet Dev* 2007, 17: 139-144.
[32] Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G: Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 2001, 276: 39-45.
[33] Jordan IK, Rogozin IB, Glazko GV, Koonin EV: Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 2003, 19: 68-72.
[34] Hasler J, Strub K: Alu elements as regulators of gene expression. *Nucleic Acids Res* 2006, 34: 5491-5497.
[35] Laperriere D, Wang TT, White JH, Mader S: Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. *BMC Genomics* 2007, 8: 23.
[36] Norris J, Fan D, Aleman C, Marks JR, Futreal PA, Wiseman RW, Iglehart JD, Deininger PL, McDonnell DP: Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J Biol Chem* 1995, 270: 22777-22782.
[37] Caretti G, Di Padova M, Micales B, Lyons GE, Sartorelli V: The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. *Genes Dev* 2004, 18: 262-272.
[38] Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, Yung Y, Montoliu L, Glass CK, Rosenfeld MG. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 2007, 317: 248-251.
[39] Tomilin NV. Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. *Bioessays* 2008, 30: 338-348.

Open Science Index, Biomedical and Biological Engineering Vol:2, No:8, 2008 publications.waset.org/15711/pdf