Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 4775–4778, July 1988
Evolution

A fundamental division in the *Alu* family of repeated sequences

(evolution/*Alu* subfamilies/secondary structure/CpG dinucleotide)

JERZY JURKA*† AND TEMPLE SMITH‡

*Bionet, 700 East El Camino Real, Mountain View, CA 94040; and ‡Dana–Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, MA 02115

Communicated by Roy J. Britten, March 10, 1988

ABSTRACT **The *Alu* family of repeated sequences from the human genome contains two distinct subfamilies. This division is based on different base preferences in a number of diagnostic sequence positions. One subfamily of the sequences, referred to as the *Alu*-J subfamily, is very similar to 7SL DNA in these positions. The other subfamily, *Alu*-S, can be divided further into well-defined branches of sequences. These findings revise the previous picture of the *Alu* family and expose their complex evolutionary dynamics. They reveal sequence variations of potential importance for the proliferation of *Alu* repeats and relate them to their structural features. In addition, they open the possibility of using different types of *Alu* sequences as natural markers for studying genetic rearrangements in the genome.**

A typical human *Alu* family member is a sequence ≈300 base pairs long and consists of two similar but not identical subunits, *Alu*-left and *Alu*-right, connected by an adenine-rich linker. Both halves of *Alu* elements are related to the 7SL RNA (1). Although *Alu* sequences are the most abundant among middle repetitive elements in the human genome, their biological role remains unclear (2). In this paper, we report on the presence of at least four different types of *Alu* sequences, which probably originated at different times in the history of primates.

METHODS

A set of 125 complete or nearly complete human *Alu* sequences were extracted from the GenBank DNA sequence data base.§ The list of the GenBank loci used, positions of the extracted sequences, and other specifications are given in the legend of Table 1.¶ The statistical analysis described below is based exclusively on pairwise comparisons of each *Alu* sequence with the consensus sequence (see Fig. 1), using the computer algorithm of Smith and Waterman (3). The overall consensus sequence in Fig. 1 was derived from our data and is slightly different from the one recently published (2). The differences are exclusively within CpG doublets, which are known to be variable in *Alu* repeats (4). Taking pairwise comparisons as a starting point, the multiple alignment of the analyzed set of sequences has been done by hand with a specialized sequence editor (5). To detect sequence positions with different base preferences (diagnostic positions), we used "column-correlation" function incorporated in the sequence editor (5). This function was originally designed to perform automatic searches for compensatory mutations.

RESULTS

During a search for compensatory mutations in the multiply aligned set of 125 *Alu* sequences, we noted an unusually high proportion of correlated base occurrences in at least 15 sequence positions. These positions are referred to as diagnostic positions and are listed in column 1 of Table 1 (the position numbers are the same as in Fig. 1). The observed correlations in the diagnostic positions reflect different base occurrences in different *Alu* subfamilies. It is shown below that the most predominant bases in the 15 diagnostic positions belong to only one of the two basic types of *Alu* sequences present in the analyzed set.

To segregate the most predominant type from the remaining *Alu* sequences, we have used computer alignment (3) of each *Alu* element with the *Alu* consensus from Fig. 1. The average overall similarity between the 125 *Alu* sequences and the *Alu* consensus is 83.88% with a SD of 5.63% (gaps counted as single mismatches). These numbers are slightly different if gaps are excluded from the analysis (see Table 3). Given the overall similarity, we assume that the probability of matching between any *Alu* sequence and its consensus in a randomly chosen aligned position equals 0.83. Any *Alu* sequence similar 40% or less to the consensus sequence in the 15 diagnostic positions has been defined as an *Alu*-J element. The probability of only 6 matches or less in 15 randomly chosen aligned positions can be calculated from the binomial distribution and is <0.001. Following the statistical definition, we have found 31 *Alu*-J sequences in the analyzed set of 125 sequences. The remaining 94 sequences are referred to as *Alu*-S sequences. The 3:1 ratio of S/J *Alu* sequences explains why the overall consensus sequences and *Alu*-S consensus sequence overlap. We have found no sequences matching seven or eight diagnostic consensus positions, which suggests that the distinction between J and S sequence types is quite sharp with few or no intermediate forms. As shown in Table 1, in the diagnostic positions the J subfamily maintains consistently different bases from those in the S subfamily. The difference in base preferences between J and S subfamilies is most evident at positions 94, 204, and 275 (Table 1). For example, G-204 is present in 29 of 31 *Alu*-J sequences and in only 1 of 94 *Alu*-S sequences. Similarly, G-94 and C-275 are powerful diagnostic indicators that can be used for preliminary "by eye" identification of *Alu*-J elements.

As illustrated in Table 1 the most frequent bases in the J subfamily are identical with those in 7SL RNA in 14 of 15 diagnostic positions. Furthermore, the differences between J and S *Alu* elements correlate with differences in the adenine-rich linker connecting the left and right halves of the *Alu* dimer (positions 121–133 of the consensus sequence in Fig. 1; data not shown). The triplet TAC in the middle of the linker is present in ≈80% of *Alu*-S compared to only 20% of the *Alu*-J sequences. It is not certain if the homologous TAC triplet was ever present in many *Alu*-J sequences since their

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.
§EMBL/GenBank Genetic Sequence Database (1987) GenBank (IntelliGenetics, Mountain View, CA), Tape Release 46.0.
¶The *Alu* sequences used in this study are available on the Bionet computer in the file <jurka>human-alu.seq. The data can also be obtained by electronic mail from jurka@bionet-20.arpa.

Table 1. Diagnostic base differences between major subfamilies of the *Alu* family

Consensus position	*Alu* subfamily	Frequency of T	C	A	G	(−)	Base in 7SL DNA	Consensus position	*Alu* subfamily	Frequency of T	C	A	G	(−)	Base in 7SL DNA
57 (C)	J	6	3	20	0	2	A	163 (A)	J	1	2	3	25	0	G
	S	36	47	9	2	0			S	2	1	55	36	0	
63 (A)	J	2	0	12	16	1	G	194 (A)	J	1	0	6	22	2	G
	S	1	2	87	4	0			S	1	1	91	1	0	
65 (C)	J	30	1	0	0	0	T	204 (A)	J	0	0	1	29	1	G
	S	20	41	1	1	37			S	0	0	89	1	2	
70 (G)	J	2	21	0	7	1	T	208 (G)	J	0	0	20	11	0	A
	S	0	2	3	89	0			S	2	3	29	59	1	
71 (T)	J	5	23	0	2	1	C	220 (T)	J	6	22	1	1	1	C
	S	90	4	0	0	0			S	80	10	0	1	3	
94 (C)	J	0	0	2	29	0	G	233 (A)	J	22	2	4	0	3	T
	S	4	87	1	1	1			S	0	0	90	3	1	
101 (G)	J	2	0	19	10	0	A	275 (T)	J	1	28	1	0	1	C
	S	0	0	9	85	0			S	85	2	1	2	4	
106 (A)	J	0	0	13	18	0	G								
	S	0	0	93	1	0									

Consensus positions are taken from Fig. 1. (−), Alignment gaps. Loci names and 5′ → 3′ positions of *Alu*-J and *Alu*-S sequences are listed below as they appear in GenBank (release 46.0).§ *Alu* repeats complementary to the consensus sequence are listed in 3′ → 5′ order. Positions preceded by b and c indicate b and c branches of *Alu*-S sequences, respectively, as defined in Table 2 and in the text. *Alu*-J: HUMACHRA7(1580–1295); HUMADAG(4907–5201, 24773–24495, 31460–31747); HUMAPOCII(1982–2235); HUMAPOE4(2562–2849); HUMBLYM1(266–560); HUMERPA(1810–2100); HUMFIXG(24172–24465); HUMFOL5(1577–1847); HUMIFNB3(2663–2405, 13213–13489); HUMIL2R8(1209–1486); HUMLDLR(4193–4485); HUMPOMC2(26–340); HUMPOMC6(26–303); HUMRSAOLD(498–790); HUMRSKPA1(24–291); HUMTBB5(2922–2627, 2949–3239, 5611–5885); HUMTHBNB(3593–3883); HUMTPA(7512–7227, 8862–9165, 10801–10513, 16794–17114, 18878–19167, 20944–21250, 22262–22536, 26941–27228); M13121(1141–856). *Alu*-S: HUMA1ATP(4932–5219); HUMADAG(1672–1369, 2357–2642; c: 5606–5893, 8000–7720, 8484–8193, 13452–13741, 15386–15096, 15806–16094, 17224–16933, 18414–18706, 19900–19613, 22527–22812, 25453–25163, 27269–26979; c: 28032–28320); HUMAGG(b: 1391–1106); HUMALBG(3287–3576, 6046–5759); HUMANFA(c: 1340–1621; b: 1630–1919); HUMAPOAI1(3291–3585, 6421–6709); HUMAPOAII(2571–2860); HUMAPOCII(2254–2542); HUMAPOE4(636–352, 2427–2138, 5049–4773); HUMC1A21(347–60); HUMC1A23(330–45); HUMC1AIN1(992–1285); HUMFIXG(7298–7595; c: 31537–31801, 35947–36248); HUMFOL5(1284–989); HUMGAST2(187–477); HUMGHV(2506–2248); HUMHBA4(2060–1773, 4297–4585, 8548–8836); HUMHBBRT(482–190, 1260–1548); HUMIFNB3(4648–4363, 7265–7545; c: 8975–8688); HUMINS2(69–357); HUMLDLIVS(291–8); HUMLDLR(b: 3715–4011); HUMMHDC3B(b: 3712–3424); HUMMHDRB3(b: 2838–3124; c: 4063–4345); HUMMYCRT(c: 3143–2876); HUMNGFB(c: 5259–5544); HUMPOMC(1392–1102, 7099–6803); HUMPOMC1(333–47); HUMRSA1(c: 508–803); HUMRSA27(1–251); HUMRSA16(b: 168–451); HUMRSAB11(1–269); HUMRSAB13(11–295); HUMRSAB19(1–241); HUMRSAB2(1–288); HUMRSAB6(1–256); HUMRSAB8(1–265); HUMRSAP3(b: 897–1186); HUMRSKA1(21–347); HUMSLJT1(568–280); HUMTBB5(3289–3573, 4115–3849; b: 5241–4953; b: 6799–6516); HUMTBBM40(c: 1828–2113); HUMTHBNB(1165–874, 3418–3110); HUMTPA(5960–5671, 6746–6483, 739–1022, 10066–10355, 12986–12700; b: 17170–17455, 21279–21567, 21940–21651, 25619–25905, 26522–26811, 27879–28149; b: 28803–29090, 32922–33210, 34234–34503); HUMUG2PD(c: 546–260, 1685–1396); M11591(b: 1404–1115); M12036(637–362); M12929(592–302).

linkers vary substantially in both their size and the primary sequence.

Following the same approach, the S subfamily of the *Alu* family has been found to contain other types of *Alu* sequences. Unlike the J/S division, the intra-S division is more difficult to define statistically since the number of simultaneous differences between subsets of *Alu*-S sequences appears to be smaller and there is a number of intermediate sequences virtually absent from the J–S junction. Therefore, we first define the most distinct "b" branch of the S subfamily as containing sequences that match 3 or fewer of the 11 diagnostic positions listed in Table 2 and in Fig. 1. There are 12 such elements in the analyzed set of 94 *Alu*-S sequences. The average overall similarity between each analyzed *Alu*-S sequence and the consensus sequence is 86.59 if every gap is counted as a single mismatch (Table 3). Based on this number, we assume the probability of matching the aligned consensus sequence at a randomly chosen position to be 0.86. As calculated from the binomial distribution, the probability of matching 3 or fewer of the randomly chosen aligned positions is 10^{-4}. The probability of matching exactly 4 and 5 positions equals 1.9×10^{-4} and 1.63×10^{-3}, respectively. We have found 5 sequences matching 4, and 6 matching 5 diagnostic positions in the analyzed set of 94 *Alu*-S sequences. These 11 sequences are arbitrarily defined as a "c" branch of the S subfamily. After segregation of the b and c branches, the remainder of the *Alu*-S subfamily is referred to as an "a-branch." Preliminary analysis of 71 *Alu* elements from this branch revealed the presence of 16 sequences containing simultaneously thymine at position 244 and adenine at position 272, as opposed to C-244 and G-272 in the remaining 55 *Alu* sequences. In addition, 14 of the above 16 sequences contain an extra adenine in position 264. This suggests that the *Alu*-a branch may contain at least two different types of *Alu* sequences and it can tentatively be replaced by "d" and "e" branches containing 16 and 55 sequences, respectively.

As illustrated in Table 2 and Fig. 1, the base preferences are quite similar between *Alu*-b and *Alu*-c sequences up to position 88. Further on, *Alu*-c remain similar to *Alu*-a with the exception of guanine at position 163. Therefore, the c branch can be viewed as an intermediate between the a and b branches of the S subfamily. A unique feature of the c sequences may be the presence of adenine at position 74. Of the 125 *Alu* sequences only 8 contain A-74 of which 7 belong to the *Alu*-c branch defined above. The eighth *Alu* sequence containing adenine at position 74 (HUMPOMC1) has all the *Alu*-c features listed in Fig. 1: deletion at 64 and 65, A-78, T-88, and G-163. Therefore, it can also be considered as an *Alu*-c sequence.

Based on the analysis of phylogenetic trees, other authors (6) have recently identified the *Alu*-b branch as a "subfamily of the *Alu* family." The authors have pointed out differences between the *Alu* consensus and the *Alu*-b sequences at positions listed in Table 2 as well as in Fig. 1 and at other less characteristic positions not included in our analysis.

```
           1            15 16           30 31           45
7SL        -GCCGGGCGCGGTGG CGCGTGCCTGTAGTC CCAGCTACT-CGGGAG
Alu-cons   GGCCGGGCGCGGTGG CTCACGCCTGTAATC CCAGC-ACTTTGGGAG

           46           60 61           75 76           90
7SL        GCTGAGGCTGGAGGA TCGCTTGAGTCCAGG AGTTC....CCAGCC
Alu-J                 a      g t    CC
Alu cons   GCCGAGGCGGGCGGA TCACCTGAGGTCAGG AGTTCGAGACCAGCC
Alu-c                        --      (A)   A          T
Alu-b                        --            A          T

           91          105 106          120 121          135
7SL        TGGGCAACATAGCGA GACCCCGTCTCT
Alu-J        G      a     g
Alu cons   TGGCCAACATGGTGA AACCCCGTCTCTACT AAAAATACAAAAATT
Alu-c
Alu-b          T    C

           136         150 151          165 166          180
7SL        -GCCGGGCGCGGTGG CGCGTGCCTGTAGTC CCAGCTACTCGGGAG
Alu-J                                 g
Alu cons   AGCCGGGCGTGGTGG CGCGCGCCTGTAATC CCAGCTACTCGGGAG
Alu-c                                 G
Alu-b                       G         G

           181         195 196          210 211          225
7SL        GCTGAGGCTGGAGGA TCGCTTGAGTCCAGG AGTTCTGGGCTGTAG
Alu-J                  G          G   a            C
Alu cons   GCTGAGGCAGGAGAA TCGCTTGAACCCGGG AGGCGGAGGTTGCAG
Alu-c
Alu-b                        G  R                   C

          226          240 241          255 256          270
7SL       TGCGCCTGTGA....G CCACTGCACTCCAGC CTGGGCAACATAGCG
Alu-J           T
Alu cons  TGAGCC-GAGATCGCG CCACTGCACTCCAGC CTGGGCGACAGAGCG

           271         285
7SL        AGACCCCGTCTCT
Alu-J          C
Alu cons   AGACTCCGTCTCAAA AAAAA
```

FIG. 1. Consensus sequence for 125 *Alu* sequences and the homologous regions of human 7SL DNA. Major and minor characteristic bases for other types of *Alu* sequences are printed in capital and lowercase letters, respectively, and correlate with the analysis in Table 1. Dots indicate sequence regions absent from 7SL DNA but present in the *Alu* family. The remaining 7SL-specific sequences are not shown. Dashes under positions 64 and 65 indicate bases missing in *Alu*-b and *Alu*-c sequences. Additional characteristic positions not listed in Table 2 are put in parentheses.

The diagnostic position 78 (Table 2) is in the middle of the stretch 77–79, which can pair with base 87–89 containing another diagnostic position 88. Bases 77–79 are within the polymerase III promoter region (bases 74–86 in Fig. 1). Correlation between occurrences of complementary bases at positions 78 and 88 suggests the possibility of a weak secondary interaction in this region. Another potential for secondary interaction, already proposed for 7SL RNA (7, 8), exists between complementary bases 69–75 and 89–95. This region includes 3 of the 15 positions distinguishing between the J and S subfamilies and the complementarity is conserved throughout the *Alu* family. The only A·C mispairing has been found in this region in the *Alu*-c sequences. The role of the above hypothetical structures is not clear, although their location suggests involvement in *Alu* transcription. There is also a possibility of a secondary interaction between bases 244 and 245 and bases 271 and 272 that includes bases at positions diagnostic for putative d and e branches of the *Alu* family discussed above.

Table 3 indicates that the average overall similarity between *Alu*-J and the *Alu* consensus sequence in nondiagnostic positions is lower than the average similarity between *Alu*-S and the *Alu* consensus. This indicates that on average *Alu*-J sequences are more diverse than *Alu*-S sequences. By *t* test, one can find that differences between *Alu*-J/consensus and *Alu*-S/consensus similarities are statistically significant ($P << 0.001$). The conclusion holds true even if the general *Alu* consensus is replaced by the *Alu*-J consensus (data not shown). There is also a significant difference ($P < 0.001$) between analogous numbers for a and b subdivisions of the *Alu* sequences. The differences between *Alu*-b and *Alu*-c sequences are marginally significant ($P < 0.05$), and analogous differences between *Alu*-a and *Alu*-c are insignificant.

As pointed out before (4), CpG doublets undergo rapid mutations in *Alu* sequences. This may result from a deamination of methylated cytosine (for a review, see ref. 9). Average CpG content is lowest in the J subfamily (3.84 ± 2.01) as compared to analogous numbers for *Alu*-a (7.75 ± 2.95), *Alu*-b (16.08 ± 5.01), and *Alu*-c (9.54 ± 3.75) branches of the S subfamily. Significance levels for the differences in the CpG content are virtually identical to those for the similarity differences discussed in the preceding paragraph.

DISCUSSION

Given the similarity between *Alu*-J and 7SL RNA sequences in the diagnostic positions, the large intra-subfamily diversity and the low CpG content, we find the J sequences to be good

Table 2. Base preferences in the S subfamily branches

Consensus position	Branches of *Alu*	Frequency of T	C	A	G	(−)
65 (C)	a	20	41	1	1	8
	c	0	0	0	0	11
	b	0	0	0	0	12
66 (T)	a	62	3	3	0	3
	c	4	0	0	0	7
	b	4	0	0	0	7
78 (T)	a	67	0	4	0	0
	c	1	0	9	1	0
	b	0	0	12	0	0
88 (G)	a	2	1	2	65	1
	c	9	0	1	1	0
	b	11	0	1	0	0
95 (C)	a	2	68	1	0	0
	c	2	9	0	0	0
	b	12	0	0	0	0
100 (T)	a	66	1	2	1	1
	c	7	4	0	0	0
	b	1	10	1	0	1
153 (C)	a	10	35	1	24	1
	c	5	1	0	5	0
	b	0	1	0	11	0
163 (A)	a	1	0	53	17	0
	c	1	1	1	8	0
	b	0	0	1	11	0
197 (C)	a	15	50	2	4	0
	c	3	6	1	1	0
	b	0	0	0	12	0
200 (T)	a	65	3	2	1	0
	c	10	0	1	0	0
	b	1	0	4	7	0
219 (G)	a	0	1	2	64	0
	c	0	1	0	10	0
	b	0	11	0	0	0

(−), Alignment gaps.

candidates for the early *Alu* elements derived from the 7SL RNA (1). The base differences in the diagnostic positions and the linker regions may be important for understanding how this transformation occurred and are good targets for experimental analysis. On the other hand, the least diverse *Alu*-b sequences can be viewed as a relatively young branch of the *Alu* family. There are three published examples of *Alu* sequences that are believed to be inserted relatively recently on the evolutionary time scale: in the α-satellite DNA of African green monkey (10), in the gorilla β-globin gene cluster (11), and at the Mlvi-2 locus of human cell lymphoma (12). All these *Alu* sequences belong to the b branch defined above.

While this paper was in review, other authors (13) reported on a subdivision of the *Alu* family into three different subfamilies corresponding to our J subfamily and two branches (a and b) of the S subfamily. These two branches, as well as the branch c, are virtually equally different from the J subfamily of *Alu* sequences and similar to each other in the

Table 3. Average overall similarities with the *Alu* consensus sequence

Alu type	Gaps as mismatches Mean	SD	Gaps excluded Mean	SD	Total
All	83.88	5.63	86.39	4.38	125
J	79.20	4.35	82.83	2.27	31
S	86.59	3.39	88.75	1.98	94
a	86.52	3.26	88.59	1.79	71
c + b	89.20	4.14	91.15	3.08	23
c	87.04	4.25	89.45	2.87	11
b	91.22	2.92	92.71	2.45	12

Sequence alignments have been made by using the computer algorithm (2). The diagnostic positions have been excluded from similarity calculations.

diagnostic positions from Table 1. Therefore, we consider them as members of the S subfamily. The authors draw their conclusions from analysis of pairwise difference distribution among *Alu* sequences involving both the diagnostic differences discussed in this paper and a mutational noise. Our analysis is based on multiple sequence comparisons, which permits more rigorous distinction between diagnostic and background differences. With this level of resolution we are able to classify each *Alu* sequence individually. This, and the analysis of the CpG content discussed in the accompanying paper (14), opens a way to date the invasion of individual genes by different types of *Alu* sequences and of genetic rearrangements associated with this process.

We thank Donald Faulkner for professional computer assistance and Roy Britten, Douglas Brutlag, Terry Friedemann, David Kristofferson, and Randall Smith for critical and useful comments on the manuscript.

1. Ullu, E. & Tschudi, C. (1984) *Nature (London)* **312**, 171–172.
2. Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. & Matsubara, K. (1987) *Gene* **53**, 1–10.
3. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **145**, 195–197.
4. Bains, W. (1986) *J. Mol. Evol.* **23**, 189–199.
5. Faulkner, D. V. & Jurka, J. (1988) *Trends Biochem. Sci.*, in press.
6. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. & Deininger, P. (1987) *Mol. Biol. Evol.* **4**, 19–29.
7. Gundelfinger, E. D., Di Carlo, M., Zopf, D. & Melli, M. (1984) *EMBO J.* **3**, 2325–2332.
8. Zwieb, K. (1985) *Nucleic Acids Res.* **13**, 6105–6124.
9. Bird, A. P. (1987) *Trends Genet.* **3**, 342–347.
10. Grimaldi, G. & Singer, M. F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1497–1500.
11. Trabuchet, G., Chebloune, Y., Savatier, P., Laucher, J., Faure, C., Verdier, G. & Nigon, V. M. (1987) *J. Mol. Evol.* **25**, 288–291.
12. Economou-Pachnis, A. & Tsichlis, P. N. (1985) *Nucleic Acids Res.* **13**, 8379–8387.
13. Willard, C., Nguyen, H. T. & Schmid, C. W. (1987) *J. Mol. Evol.* **26**, 180–186.
14. Britten, R. J., Baron, W. F., Stout, D. & Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770–4774.