Nucleic Acids Research, Vol. 20, No. 3 ***487–493***

Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes

Yves Quentin
Theoretical Biology and Biophysics Group, T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received October 24, 1991; Revised and Accepted January 9, 1992

ABSTRACT

In the primate genome, a typical Alu element corresponds to a dimeric structure composed of two different but related monomeric sequences arranged in tandem. However, the analysis of primate sequences found in GenBank reveals the presence of free left and free right Alu elements. Here, we report the statistical study of those monomeric elements. We found that only a small fraction of them results from a deletion of a dimeric Alu sequence. The majority derives from the amplification of monomeric progenitor sequences and constitutes two families of monomeric elements: a family of free left Alu monomers that is composed of two subfamilies and a small family of free right Alu monomers. Both families predated the dimeric Alu elements, and a phylogenetic analysis strongly suggests that the first progenitor of the dimeric Alu family arose through the fusion of a free left monomer with a free right monomer.

INTRODUCTION

The Alu family is the dominant family of dispersed, repeated elements in the genomes of all primates studied to date. More than 500,000 members of that family are distributed through the human genome (1–4).

A typical Alu element corresponds to a head-to-tail tandem dimer approximately 300 bp long, the right monomer containing an additional 31 bp segment relative to the left monomer (2). It has been proposed that both monomers arose from the 7SL RNA through a deletion of the central 7SL-specific sequence (5). Alu elements are thought to have been spread throughout the primate genomes via a process, called retroposition (6), involving an RNA intermediate. The transcription is controlled by the RNA polymerase III promoter located in the left half of the element (7,8). This RNA is used as a template by the reverse transcriptase to generate a cDNA that is subsequently integrated at a new locus in the genome. The integration produces short direct repeats that flank the element.

Since the Alu dimeric elements seem to be specific to the primate genomes, they have amplified in the past 65 million years (9). The various studies previously done on the Alu DNA sequences have shown that this family can be subdivided into different subfamilies of related elements, each subfamily being characterized by a specific set of mutations called diagnostic positions (10–15). The study of the diagnostic positions and the divergence observed between the members of the different subfamilies suggest that they have appeared at different evolutionary times through mutations of a member sequence belonging to a previously existing subfamily. The amplification of a given subfamily seems to be limited in time. Supporting this hypothesis, the few human Alu sequences that are known to be polymorphic insertions in the human genome belong to the youngest subfamily found only in that genome (16–18).

If the large number of Alu elements present in the primate genomes attests to their efficiency to amplify, more recent studies have shown that only a small fraction of the Alu elements is able to duplicate (10–15). Those elements have been called source genes, and the elements unable to transpose have been referred to as pseudogenes (12). Those definitions assume a function for the Alu elements that has not yet been defined. Therefore, in the present study we will use a definition that refers only to the mobility of the Alu elements. However, this definition does not exclude a possible function for the Alu sequences but seems more appropriate given the present controversy about the functionality of the Alu sequences (12,15,19). The Alu sequences able to duplicate will be called progenitor elements (9), and the elements unable to transpose will be called sterile elements.

The most numerous Alu elements are dimeric, but the presence of a family composed exclusively of free left Alu monomers (FLAM) has also been reported (14,20,21). This family has been shown to predate the appearance of the first Alu dimeric element and could be representative of an ancestral Alu monomeric family (14,20,21). The increasing number of Alu sequences in the databases allowed us to undertake the DNA sequence analysis of both free left and free right Alu monomeric (FRAM) elements.

In the present study we characterized two new families of repetitive elements in the primate genomes that preceded the emergence of the dimeric Alu elements. The first family is composed of free left Alu monomers (FLAMs); it has been subdivided into two subfamilies. The second family is formed by only one group of free right Alu monomers (FRAMs). Sequence comparisons suggest that both families share a common ancestor issued from a 7SL RNA gene and that just after the

emergence of the primate lineage, a member of the FLAM family merged with a member of the FRAM family to form the first Alu dimeric element.

MATERIALS AND METHODS

The sequences were extracted from GenBank (Release 67) (22) by applying the program FIRE (Fast Identification of Repetitive Elements; Quentin, unpublished). This program allows a fairly exhaustive indentification of the Alu elements, even if those elements are partial and/or divergent from the consensus sequences. The histogram of the length of the extracted Alu elements (not shown) presents a major peak corresponding to the length of the dimeric elements, but also presented are two minor peaks attesting to the existence of free left and free right monomers in our sample of sequences. The sequences corresponding to the free left and the free right monomers were selected and aligned against the Alu consensus sequence given in (14) to obtain a multiple alignment. When different copies of the same gene and/or duplicated genes were present, only one occurrence of the Alu elements was conserved for the statistical analysis. A new approach has been used to select the bases involved in the characterization of the subfamilies. Since each subfamily of sequences is defined by a set of diagnostic bases, each diagnostic base of one subfamily will be correlated with all the other diagnostic bases of the same subfamily. Therefore, this implies that only the bases correlated to other bases can convey some information for subfamilies definition.

To measure possible links between bases at different positions in the multiple alignment of sequences, we used the Kullback information distance (KID) (23). For two random variables x and y, this distance measures the statistical dependence between x and y. It tells us how the uncertainty we have on x is decreased by knowing y. Since knowing y cannot make x more uncertain, the distance is positive or null. KID is a symmetric function of x and y, and it is expressed in terms of probabilities and joint probabilities. If $P(x_i)$ is the probability of observing the base x at position i, if $P(y_j)$ is the probability to observe the base y at position j with $j \neq i$, and if $P(x_i,y_j)$ is the joint probability of the base x to appear at position i and the base y to appear at position j, then the Kullback information distance between x and y can be written as:

$$KID(x_i,:y_j) = P(x_i,y_j)\log_2\frac{P(x_i,y_j)}{P(x_i),P(y_j)} + P(\bar{x}_i,y_j)\log_2\frac{P(\bar{x}_i,y_j)}{P(\bar{x}_i),P(y_j)}$$

$$+ P(x_i,\bar{y}_j)\log_2\frac{P(x_i,\bar{y}_j)}{P(x_i),P(\bar{y}_j)} + P(\bar{x}_i,\bar{y}_j)\log_2\frac{P(\bar{x}_i,\bar{y}_j)}{P(\bar{x}_i),P(\bar{y}_j)}$$

where the event $\bar{x}_i$ is the observation of a base different from x or of a deletion at the position i, and the event $\bar{y}_j$ is the observation of a base different from y or of a deletion at the position j.

This distance is equal to zero if the two bases are independent and greater than zero if the two bases are dependent. The upper limit is defined by $\log_2(n)$, where n is the number of possible states taken by the variables under study. Here, we consider if a base is present or absent at a given position. Thus, in a sequence, the variable can only have two different states, and n is equal to 2. Therefore, the upper limit equals $\log_2(2) = 1$. With this approach, the deletions found several times at the same positions in the sequences increase the score obtained with the Kullback information distance. However, when those deletions are found at the 5′ and 3′ ends of the sequences, they are not informatives for the definition of the subfamilies. They simply correspond to partial sequences and have been discarded from our samples.

The Kullback information distance observed between the pairs of bases in the real sequences have been compared with the scores obtained on 10 different sets of resampling sequences. From this comparison we deduced a threshold that has been used to select the pairs of bases having unusual correlations. The resampling files have been obtained with the 'permute' option of the program SEQBOOT of the PHYLIP package (24). This method permutes the bases of each column of the multiple alignment (25,26). The resampling conserves the base composition of the columns but disrupts all possible links that may exist between the positions. Only the pairs of bases having a KID greater than the threshold have been retained for further analysis, except the pairs of bases involved in a CpG dinucleotide. Indeed, the CpG dinucleotide tendency to mutate to TpG or CpA through deamination of the methylated cytosine (27) induces a correlation between the bases C and A on one hand and T and G on the other hand. Nevertheless, due to the rapid rate of mutation of the CpG dinucleotide, in general, those correlations have no significance for the subfamily definition and induce a background noise (12,14).

The remaining pairs of bases showing a score greater than the threshold have been used to define the subfamilies. Each position of the sequences presenting a correlated base has been encoded by a binary variable, which takes the value 1 if the base is observed in the sequence and 0 otherwise. If at the same position, two bases were implicated in a correlation, this position is encoded by two binary variables. The comparison of the sequences has been achieved by two complementary methods: a clustering method (28) and a correspondence analysis [see (29) and (30) for mathematical details, and (14), (30), and (32) for applications of this method to evolutionary analyses]. Both methods are based on the χ^2 metric, and the partition obtained with the cluster analysis can be shown on the graph produced by the correspondence analysis.

RESULTS AND DISCUSSION

Selection of the variables

The selection of the free left and free right monomeric Alu elements from GenBank allows the constitution of a sample of 66 free left monomers and a sample of 39 free right monomers. We computed the Kullback information distances between pairs of bases on those two samples. As described in the method, the values obtained with 10 resampling sets of sequences are used to define a cutoff for each sample. A greater dispersion of the KID values is observed with the resampling sequences of the right monomers than with those on the left. Thus, two different thresholds have been obtained for each set (0.20 for the left monomers and 0.30 for the right monomers), corresponding however to a similar overlap between the scores of the real and the resampling sequences.

Fifty-two pairs of bases present a KID greater than 0.20 in the left monomer sample, and 39 pairs of bases have a value greater than 0.30 in the right monomer sample. Thus, despite the fact that the left monomers are shorter than the right monomers, they contain more correlated bases. Those pairs of correlated bases involved 23 positions in the left monomers and

13 positions in the right monomers. They represent only a small fraction of the total number of positions in the sequences: 19% for the left monomers and 9% for the right monomers. This result shows the importance of using a method that extracts the informative variables from the background noise. Among the selected positions, only 13 (11%) in the left monomers and 6 (4%) in the right monomers are not implicated in a CpG dinucleotide. Twenty-two and 10 bases in the left and right monomers, respectively, are involved in the correlations between those positions. Indeed, at the same position, two different bases can be correlated with other bases at other positions.

Multivariate analysis

To obtain a good description of the different groups of sequences, we discarded from the initial samples of sequences the elements presenting a deletion in at least one of the selected positions. Our final sets are composed of 59 free left monomers and 36 free right monomers. We applied a cluster analysis and a correspondence analysis to those new samples of sequences where the selected correlated bases have been encoded by a binary variable as described in the method. The results are summarized in Figure 1A and 1B. Those figures represent the projection of the sequences and the variables (the selected bases) on the first and second axes of the correspondence analysis. The partition displayed has been obtained independently with the clustering method. On the free left monomers (Figure 1A), the first axis of the correspondence analysis opposes the group Al and Bl defined by the cluster analysis, and the second axis opposes those groups to the third group (Cl). Each group of the partition is associated with a specific set of variables (throughout the text, the numbering refers to the 7SL RNA sequence in order to ease the comparisons between the figures): the Al group with G27, T35, A36, C37, and C39; the Bl group with T70, C276, G283, and A288; and the remaining variables with the last group, except the variable G69 that is shared by the Bl and Cl groups. The Al group corresponds to the M1 subfamily described in Quentin (20) and to the FLA monomers described by Jurka and Zuckerkandl (21), and the Cl group corresponds to the M2 subfamily reported by Quentin (14,20).

The same analysis applied on the free right monomers (Figure 1B) discriminates two groups of sequences (Ar and Br) on the first axis.This discrimination is based on four positions. The variables A58, A68, G77, and A97 are associated with the group Br and the variables G58, G68, T77, and T97 with the group Ar. The second axis opposes the sequences having a T in position 77 with sequences having a G at the same position. However, as pointed out by Jurka and Milosavljevic (15), a split relying upon only one position has no statistical support.

In summary, the results obtained with both methods are in a good agreement and revealed on our samples the presence of three groups of free left monomers and two groups of free right monomers. Those partitions are based on four or more diagnostic positions.

Origin of the free left and free right Alu monomers

Several hypotheses can be proposed for the origin of the free left and right monomers. (i) They may result from the deletion of one half of a dimeric element. For example, the free left monomer can be the by-product of a recombination between the two poly(dA)-rich regions of a dimeric element. This deletion removes precisely the right monomer. (ii) The monomeric elements may also result from a partial duplication of a dimeric element. In this case, the free right monomers correspond to an early ending of the retrotranscription, and the free left monomers result from the retrotranscription of a dimeric element transcript that is initiated in the middle poly(dA)-rich region instead being initiated in the terminal poly(dA)-rich. (iii) Finally, the free monomers may be the product of the amplification of free left

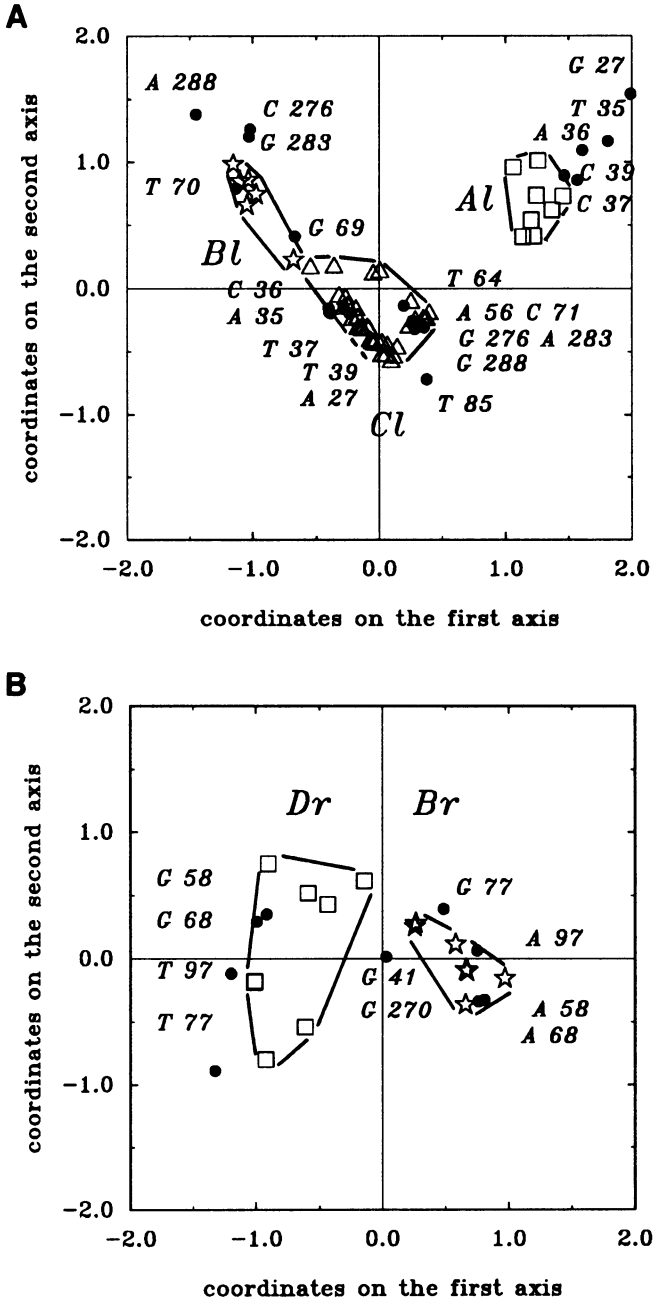

Figure 1. Results of the cluster analysis and of the correspondence analysis on the free left Alu monomers (A) and the free right Alu monomers (B). The plans are the projection of the sequences and the variables on the first two axes of the correspondence analysis. The coordinates of each sequence are indicated by a symbol that refers to the result of the cluster analysis: a square represents the elements of the Al and Ar groups, a star the elements of the Bl and Br groups, a triangle for the elements of the Cl group. Lines are drawn around the groups of sequences found by the cluster analysis. The variables used in the analysis are represented by a closed circle with the description of the base involved and the position. The numbering refers to the 7SL RNA sequence in order to make the comparisons between the figures easier.

and free right monomeric progenitors. These hypotheses are not exhaustive but are sufficient to illustrate the fact that the free monomers can have two different origins: a dimeric or a monomeric origin. To decide between the two possibilities for the origin of the monomeric sequences, we compared each sequence of the different groups with the free left and free right monomers of the progenitor sequences of each subfamily of the dimeric Alu elements previously described. These comparisons have been done on the diagnostic bases given in (15).

Each element of the Bl group and the Br group can be assigned to one of the different subfamilies of dimeric Alu sequences, with the exception of the oldest subfamily (J). Therefore, those two groups of sequences do not form subfamilies of monomeric elements. In our analysis, they formed single groups because we had only a few occurrences corresponding to each dimeric subfamily. Thus, the members of the Bl and Br groups are most likely the result of a deletion or a partial duplication of dimeric elements. If we used only the diagnostic bases, as before, the elements of the Cl, Al, and Ar groups are classified with the J subfamily. However, if we take into account the base changes observed between the 7SL RNA sequence and the consensus sequence of the J subfamily, the elements of the Cl group remain close to the J subfamily when the elements of the Al and Ar groups are clearly located between those two sequences (i.e., they have some bases diagnostic of the J subfamily but also other bases diagnostic of the 7SL RNA sequence). Thus, contrary to the Bl and Br groups, the Al, Cl, and Ar groups are homogeneous and can be characterized by a consensus sequence (Figure 2). There are only two differences between the consensus sequence of the Cl group and the consensus sequence of the left arm of the J subfamily (in positions 62 and 288). This result suggests that the Cl group represents partial dimeric sequences of the J subfamily. However, the following observation raises another explanation. The 9 members of the Bl group are distributed over all the subfamilies of dimeric Alu elements of the S subfamilies (Sa, Sb, and Sc), so they are less than 31 Myr (million years) old (33). The relative age of the J subfamily has been estimated at 55 Myr. Then, if we suppose that the elements of the different subfamilies have the same probability of producing free left monomers, and if we suppose that the 9 Bl elements correspond to partial Sa elements, then we expect 16 (55×9/31) free left monomers corresponding to the J subfamily (this is an overestimate since some elements of the Bl group are less than 31 Myr old; they belong to younger subfamilies than the Sa

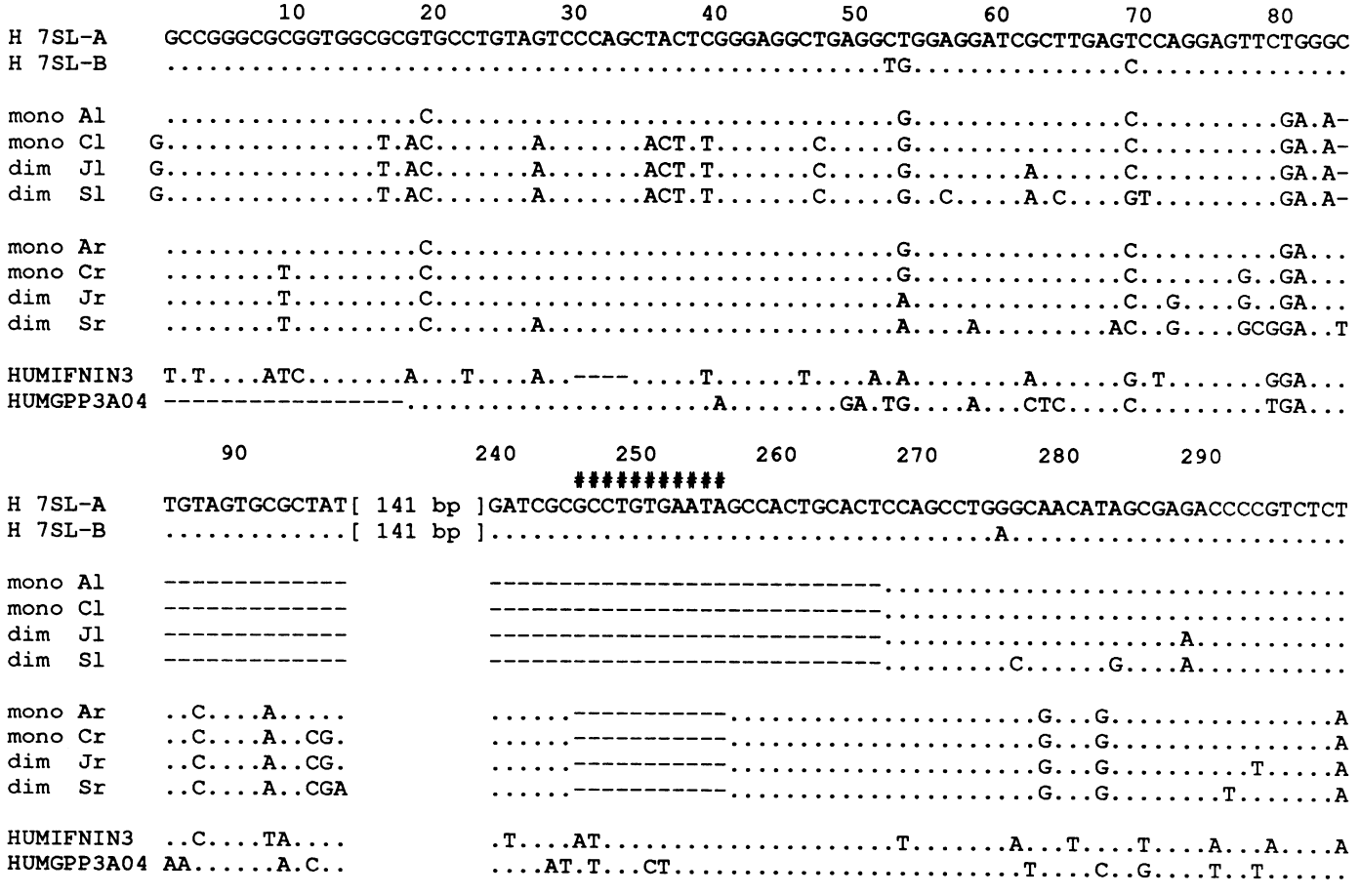

Figure 2. Sequence alignment of the consensus sequences of the Al, Cl, and Ar groups of monomeric elements (mono Al, mono Cl, and mono Ar), with the sequences of the human 7SL RNA (H 7SL-A and H 7SL-B (36)), and the consensus sequences of the S and J subfamilies [dim Jl and dim Sl for the left arms and dim Jr and dim Sr for the right arms (15)]. In the alignment, a dot corresponds to an identity and a dash to a nucleotide deletion. Otherwise, the substituting nucleotide observed in a sequence at a given position is listed. The 7SL-A RNA sequence has been used as reference. The triple substitution TAC to ACT found between 35 and 37 can also be the result of the deletion of a T residue in position 35 and of the insertion of a T residue between positions 37 and 40 where there are three T in a row (21). The sequences HUMIFNIN3 and HUMGPP3A04, discussed in the text, do not have the 11 bp long deletion, marked by #, between positions 245 and 255 found in all the other right Alu monomers.

Al elements

agaactaaaag	[GORAFPA.1]	a_{10}gagaagaagaaccaaaag	10/11
aacaaccctctg a	[HUMGFAP.1]	ataaaaaccaaaagacaaaacaagcctctg	11/12
gcagggag	[HUMKEREP.1]	$ca_{27}(ga)_{10}tt(ga)_{7}$gaagggag	7/8
gctgggtgg	[HUMTKRA.2C]	$a(gaa)_{2}ca_{5}ca_{4}ca_{5}$gaatggtgcgtgggtag	8/10
tcctctcaga	[HUMTPA.10C]	caaaagaatagaatattttcctggcaga	8/10
aactgctcgta	[HUMVWFA31.1]	aacatgtatgta	8/12

Cl elements

aacaaatagg	[HUMAMYAB3.1]	aaaagaacaaagg	8/10
a_{14}caagatggtcttg	[HUMAPOCIA.5]	attta_{14}aaagatggttttg	25/27
aaggagctagg a	[HUMAPOE4.4C]	tgtctctatataaaagagctagg	10/11
$a(ca_{3})_{2}$taagatagtgatgct	[HUMBSF2.1]	acaaaaaaccaaaaccaaa$(ca_{3})_{2}$aatgatagtggtgct	21/24
aatatgctt tga	[HUMEDHB17.4]	gcaaaaaatattaaaaagctt	7/9
agattcttcactt	[HUMHBLOD.3C]	$aca_{5}tta_{8}$ttat[dim]$_{2}a_{18}cta_{4}$agattcttcacct	12/13
caagaagaagaag	[HUMHCF2.4]	attttaaaaaagaaagccaaaagaagaag	12/13
ctaag g	[HUMHMGIB.1]	ctgag	4/5
caaaatgtatcat aatacaa	[HUMHPRTB.12]	aaaaaga$_{8}$taat$_{6}$aca$_{4}$ta$_{5}$caaaatgtatcat	13/13
aaagtgtagaaagtg	[HUMHPRTB.13]	a_{12}gagagaga$_{7}t_{8}$attta$_{4}$aaagtgtagaaagtg	15/15
aatatctcttgatttga	[HUMHPRTB.21]	acaaataataaaaatatctgttgatttga	16/17
gaaagttaaatat g	[HUMHPRTB.24]	a_{4}tatta$_{5}$ta$_{2}$gta$_{3}$ta$_{28}$ta$_{3}$ta$_{5}$gaaggttaagtat	11/13
atatacatcagg	[HUMHPRTB.49]	taatatacatacatatacatcggg	11/12
acaagaaatatttgttg a	[HUMINT02.2]	$a(ta)_{2}t_{5}a_{3}tg(ta)_{2}ga_{7}ga_{3}ga_{5}$acaaataagtgtttgttg	14/18
agaaaatcgat aata	[HUMMHSXA.2C]	aagaaaacaat	9/11
acaacaagatct a	[HUMPADPRP.2]	ttcttt$(ac)_{10}$acacaaaatatct	9/13
atcactctca	[HUMPAIA.20]	(ag)-rich tail (219 bp)atcactctca	10/10
aagaaaggaag tata	[HUMPAPA.2]	accaaaaaaaggaaggaag	10/11
aaaaattt tt	[HUMPRCA.2]	attttta$_{7}$gta$_{8}$gatctaaaaattt	8/8
aaaaaggtaatga gg	[HUMRASR2.2]	taaaaaaaaaaggtgggggaatga	12/18
gagacaggcagcaactcagct	[HUMRASR2.4]	acaaaaatagagacagggagcaactgct	18/21
aaacaa	[HUMSTATH1.1]	cacaaaaacatagaaacaa	6/6
aatctagaacgtg g	[HUMTHB.24]	(a)-rich tail (59 bp)	
	[HUMTHB.25]	$(aaat)_{3}$tga$_{4}$taagatacaatctagaatgtg	12/13
ataaaagaatgttggg	[HUMTKRA.1C]	ataaaaataaaataaaattagctggg	12/16
aaaaagtgtaag g	[HUMTPA.8C]	$aca_{6}taca_{5}tt$[dim]a_{14}aaaaagtataag	11/12
aaatcttaaactcct	[HUMTPA.26]	$aca_{5}tta_{6}$taagaagata$_{5}$ggaaaccttaaactctct	14/16
aaaagcta tgctga	[HUMTS1.8]	acaaaaaaatggaaaagcta	8/8

Ar elements

aataatgtaggact tga	[HUMATP1A2.3]	ttaaaaaaaaatctaggact	12/14
ataaatacaga	[HUMGPP3A04.1]	aaaaaataaatctaga	9/11
aaagttgtgatgcag	[HUMHPRTB.34]	tttaaaaaaagaatgttgtgaggccg	12/15
agacaaagttctggat	[HUMIFNIN3.1]	a_{22}aggcaaagttctagat	14/16
agagaaattggccataggg	[HUMIFNINI.1]	a_{7}taagagattaa$_{7}$aaggaaatttggctataggg	15/17
aagatgtcagactg	[HUMNFR.2]	taaagaaaaaaaaagtcagactg	12/14
ataaactgga	[HUMSISG1.3]	taaagaaataaaaataaatctata	8/11
aaaaaaaagattccca ta	[HUMSNRNP1.1]	taaaaaaatcattcccca	14/18
aaaaaaaaaaattagc	[HUMTHB.28]	aaaaaaaaaaaagctagc	14/16
aactctaccagg	[HUMTS1.6]	aatattta$_{5}$(caa)$_{4}$aactctatcagg	11/12

Figure 3. The sequences flanking the monomeric elements. The sequences of the short direct repeats are underlined. The names correspond to the GenBank entries. The quality of the direct repeats are estimated by the number of matches versus the length of the repeats. The elements HUMHBLOD.3C, HUMTHB.24, HUMTHB.25, and HUMTPA.8C belong to Alu clusters. The notation [dim] means that a dimeric Alu element had become integrated into the dA-rich region of a monomeric element. Note that the dA-rich regions of the elements differ from the consensus sequence (AAAAATACAAAAATTA) found at the end of the left arm of the dimeric Alu elements.

subfamily). 41 such elements have been observed. Thus, this result suggests that either the elements of the J subfamily have a very high probability of becoming free left monomers (compared to the other subfamilies) or that the majority of the elements of the Cl group have arisen from a monomeric progenitor sequence. Another feature supports the second interpretation: 27 elements of the Cl group are flanked by well defined short direct repeats (Figure 3). Such repeats are landmarks for the integration of the elements. We can also notice that the dA-rich regions of the Cl elements differ from the consensus sequence (AAAAATACAAAAATTA) found at the end of the left arm of the dimeric Alu elements (Figure 3). Thus, if few members of the Cl group correspond to partial dimeric Alu elements of the J subfamily, it appears most likely that the other members of this group constitute a subfamily of free left Alu monomers. One element of this group, BC200, presents peculiar features. It has conserved most of its CpG doublets, and it has been found to be expressed in the cytoplasm of primate (cynomolgus monkey) brain and cell lines (34,35). Recently, it has been proposed that it may be an example of a progenitor sequence (15). In the present analysis this sequence has been classified in the Cl group, and no other sequence shares the same base changes observed in the BC200 sequence. Thus, in our study, this sequence does not define a specific subfamily of left monomers. Since we analyzed only a small sample of free left monomers, and mostly from human sequences, this result cannot be a strong argument against the proposed hypothesis.

Phylogenetic relationship between the progenitor sequences

The consensus sequences of the Al, Cl, and Ar groups, the consensus sequences of the two oldest subfamilies of dimeric alu elements [S and J (13)], and the sequences of the human 7SL RNA genes (36) have been used to reconstruct a phylogenetic tree with the maximum likelihood method [DNAML program of the PHYLIP package (24)]. On the topology obtained (Figure 4), the localization of the consensus sequences of the Al, Cl, and Ar groups between the 7SL RNA sequences and the consensus sequence of the first dimeric Alu elements (J), implies that the sequences of those groups have appeared from monomeric progenitors. Thus, the Al and Cl groups form two

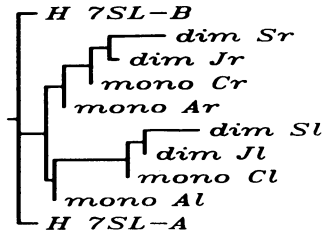

Figure 4. Unrooted tree obtained with the maximum likelihood method (24). The consensus sequences of the Al, Cl, and Ar groups (mono Al, mono Cl, and mono Ar), the consensus sequence of the two oldest subfamilies of dimeric alu elements [dim Jl and dim Sl for the left arms and dim Jr and dim Sr for the right arms (15)], and the sequences of the human 7SL RNA genes [H 7SL-A and H 7SL-B (36)], have been used to reconstruct a phylogenetic tree with the maximum likelihood method [DNAML program of the PHYLIP package (24)]. The length of the branches is proportional to the number of nucleotide substitutions. All branches are significantly positive with a probability $p < 0.01$, except the branch that precedes the divergence between the consensus sequences of the group Cl (mono Cl) and the other left monomers. This branch is significantly positive with a lower confidence ($p < 0.05$).

different subfamilies of a family of free left Alu monomers (FLAMs), and the Ar group constitutes a family of free right Alu monomers (FRAMs). Then, in the primate genomes, the free Alu monomers have two distinct origins. They result either from a dimeric Alu element (the members of the Bl and Br groups and few members of the Cl group) or from the amplification of a monomeric progenitor sequence (the members of the Al and Ar groups and the majority of the elements of the Cl group).

The tree also shows that the consensus sequences of the Al and Ar groups have a common ancestor sequence that derived from the 7SL RNA genes (Figure 4). This ancestor differs by one position (A84) from the consensus sequence of the Al group and by two positions (G279 and G283) from the consensus sequence of the Ar group (Figure 2). Among the elements of the Ar group, two sequences (HUMINFN3 and HUMGPP3A04) do not have the deletion of eleven base pairs found in the other right monomers (Figure 2). This deletion corresponds to the positions 245 and 255 in the 7SL RNA sequence. This feature suggests that those sequences have preceded the elements of the Ar group. In addition, neither sequence has the three base changes (A84, G279 and G283) that have occurred after the divergence between the left and the right monomers. Thus, it is tempting to speculate that HUMINFN3 and HUMGPP3A04 sequences might belong to an ancestral family of elements that has diverged to form the families of left and right monomers. Indeed, both monomers can result from a deletion of this ancestral sequence: a deletion of 42 base pairs for the left monomer and a deletion of 11 base pairs for the right monomer. However, to confirm this hypothesis, more examples of this family are nedeed.

It has been shown that the J subfamily of dimeric Alu elements is nonhomogeneous and can be subdivided into smaller subfamilies (15). Therefore, the progenitor sequence of this family is probably different from its consensus sequence. However, the topology obtained shows that the Alu dimeric elements have appeared from the families of free Alu monomers and confirms the 'fusion model' proposed by Daniels and Deininger (37). The first progenitor of the dimeric Alu element is probably the result of a fusion of a free left Alu monomer (FLAM), coming from the Cl subfamily, with a free right Alu monomer (FRAM), coming from the Ar family. A similar model has been proposed for the formation of the Galago Type II family. This family may result from the insertion of an element belonging to the Type III family in the central dA-rich region of a dimeric Alu element. This model and the fact that the 5′ end dA-rich sequences are a preferential site for the Alu integration (38,39), suggest that either a FRAM inserted in the dA-rich end of a FLAM or a FLAM inserted in the dA-rich sequence that links two FRAMs arranged in tandem fashion. As it has been proposed for the Galago Alu Type II (37,40), the fusion of a FLAM and a FRAM may have occurred several times, but the available data support the hypothesis of only one fusion event.

The monomeric families described here represent a small fraction of the Alu elements in the human genome (a rough estimation from GenBank gives, respectively, 0.7%, 3.2%, and 1.2% for the Al, Cl, and Ar subfamilies). The results obtained suggest that those families started to amplify before the formation of the first dimeric Alu progenitor sequence but that they did not develop new progenitors after this event. However, as the average pairwise similarities among the members of the Al, Cl, and Ar groups (74%, 75% and 71%) are only slightly different from the one observed with the elements of the J subfamily [73% (14,15)], we can assume that the progenitor sequences of the

monomeric families may have been still active when the first progenitor of the first dimeric Alu elements began to spread in the primate genome.The turnover (replacement of the monomeric families by the dimeric family) that we observed can be the result of a competition for the same retroposition machinery. One possibility is the acquisition by the dimeric Alu elements of a more efficient RNA polymerase III promoter [see Daniels and Deininger (40) for such an example in the Galago genome]. In connection with this hypothesis, the success of the Cl subfamily (four times more members than the Al and Ar groups) and of the dimeric Alu elements can be related to the base changes observed between the Al and Cl consensus sequence. They are all clustered in the 5′ part of the sequences (Figure 2), but the same region do not suffer any other mutation in the subsequent dimeric progenitor sequences (21). However, it has been also proposed that fluctuations in the population size may contribute to the turnover of the family of repetitive sequences (14,20). Indeed, in a large population several new progenitor sequences can arise by mutations, but the rate of fixation of the new inserted elements is slow. On the other hand, after an abrupt reduction in the size of the population, only few progenitor sequences will be present; but the rate of fixation is increased. Thus, competition between progenitor sequences and fluctuation in the population size can drive the evolution of the families of small repetitive sequences.

CONCLUSION

We have reported the analysis of free left and free right Alu monomers found in GenBank. The results obtained suggest that only a small fraction of those elements corresponds to partial dimeric Alu sequences. The other ones are the result of the amplification of monomeric progenitor sequences. The available sequences have been subdivided into two subfamilies of free left Alu monomers and one family of free right Alu monomers. Both families share a common ancestor sequence derived from the 7SL RNA genes, and they started to amplify before the formation of the first dimeric Alu progenitor sequence but do not develop new progenitors after this event. We have shown also that the first dimeric progenitor is the result of the fusion of a free left Alu monomer with a free right Alu monomer. This progenitor and the progenitors of the monomeric sequences may have been active at the same time, but at the end the dimeric progenitor superseded the monomeric ones. This replacement of the old monomeric families by the new dimeric family can be the result of a competition between the different progenitor sequences for the same retroposition machinery and/or the consequence of size fluctuations in the primate population.

ACKNOWLEDGEMENTS

We are grateful to C. Burks and G. Fichant for helpful comments and critical readings of the manuscript. We also appreciate careful proofreading by P. Reitemeier. This work was funded by NIH grant GM-37812.

REFERENCES

1. Houck, C. M., Rinehart, F. P. and Schmid, C. W. (1979) *J. Mol. Evol.*, **132**, 289–306.
2. Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. and Schmid, C. W. (1981) *J. Mol. Biol.*, **151**, 17–33.
3. Rinehart, F. P., Ritch, T. G., Deininger, P. L. and Schmid, C. W. (1981) *Biochemistry*, **20**, 3003–3010.
4. Schmid, C. W. and Jelinek, W. R. (1982) *Science*, **216**, 1065–1070.
5. Ullu, E. and Tschudi, C. (1984) *Nature (London)*, **312**, 171–172.
6. Rogers, J. (1983) *Nature (London)*, **301**, 460.
7. Jagadeeswaran, P., Forget, B. G. and Weissman, S. M. (1981) *Cell*, **26**, 141–142.
8. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T. and Gesteland, R. F. (1981) *Cell*, **26**, 11–17.
9. Deininger, P. L. and Daniels, G. R. (1986) *Trends Genet.*, **2**, 76–80.
10. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. and Deininger, P. L. (1987) *Mol. Cel. Biol.*, **4**, 19–29.
11. Willard, C., Nguyen, H. T. and Schmid, C. W. (1987) *J. Mol. Evol.*, **26**, 180–186.
12. Britten, R. J., Baron, W. F., Stout, D. B. and Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4770–4774.
13. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4775–4778.
14. Quentin, Y. (1988) *J. Mol. Evol.*, **27**, 194–202.
15. Jurka, J. and Milosavljevic, A. (1991) *J. Mol. Evol.*, **32**, 105–121.
16. Deininger, P. L. and Slagel, V. K. (1988) *Mol. Biol. Evol.*, **8**, 4566–4569.
17. Matera, A. G., Hellmann, U., Hintz, M. F. and Schmid, C. W. (1990) *Nucleic Acids Res.*, **18**, 6019–6023.
18. Batzer, M. A. and Deininger, P. L. (1991) *Genomics*, **9**, 481–487.
19. Zuckerkandl, E., Latter, G. and Jurka, J. (1989) *J. Mol. Evol.*, **29**, 504–512.
20. Quentin, Y. (1989) Thesis, University of Lyon, France (in french).
21. Jurka, J. and Zuckerkandl, E. (1991) *J. Mol. Evol.*, **33**, 49–56.
22. Burks, C., Cinkosky, M. J., Gilna, P., Hayden, J. E.-D., Abe, Y., Atencio, E. J., Barnhouse, S., Benton, D., Buenafe, C. A., Cumella, K. E., Davison, D. B., Emmert, D. B., Faulkner, M. J., Fickett, J. W., Fischer, W. M., Good, M. Horne, D. A., Houghton, F. K., Kelkar, P. M., Kelley, T. A., Kelly, M., King, M. A., Langan, B. J., Lauer, J. T., Lopez, N., Lynch, C., Lynch, J., Marchi, J. B., Marr, T. G., Martinez, F. A., McLeod, M. J., Medvick, P. A., Mishra, S. K., Moore, J., Munk, C. A., Mondragon, S. M., Nasseri, K. K., Nelson, D., Nelson, W., Nguyen, T., Reiss, G., Rice, J., Ryals, J., Salazar, M. D., Stelts, S. R., Trujillo, B. L., Tomlinson, L. J., Weiner, M. G., Welch, F.J., Wiig, S. E., Yudin, K. and Zins, L. B. (1990) *Meth. Enzymol.*, **183**, 3–22.
23. Kullback, S. (1959) *Statistics and Information Theory*. New York: J. Wiley and Sons.
24. Felsenstein, J. (1989) *Cladistics*, **5**, 164–166.
25. Archie, J. W. (1989) *Systematic Zoology*, **38**, 219–252.
26. Faith, D. P. and Cranston, P. S. (1991) *Cladistics*, **7**, 1–28.
27. Bird, A. P. (1980) *Nucleic Acids Res.*, **8**, 1499–1504.
28. Fages, R. (1978) J. Soc. Franc. Classific, 99.
29. Benzécri, J. P. (1969) In Watanabe, S. (ed.), Methodologies of Pattern Recognition. Academic Press, New York, pp. 35–60.
30. Lebart, L., Morineau, A., Tabard, N. and Warwick, K. M. (1984) *Multivariate Descriptive Statistical Analysis*. New York: J. Wiley and Sons.
31. Mannella, C. A., Frank, J. and Delihas, N. (1987) *J. Mol. Evol.*, **24**, 228–235.
32. Quentin, Y. (1989) *J. Mol. Evol.*, **28**, 299–305.
33. Labuda, D. and Striker, G. (1989) *Nucleic Acids Res.*, **17**, 2477–2491.
34. Watson, J. B. and Sutcliffe, J. G. (1987) *Mol. Cell. Biol.*, **7**, 3324–3327.
35. Brosius J. (1991) *Science*, **251**, 753.
36. Reddy, R. (1988) *Nucleic Acids Res.*, **16**, r71–r85.
37. Daniels, G. R. and Deininger, P. L. (1985) *Nature (London)*, **317**, 819–822.
38. Rogers, J. (1985) *Int. Rev. Cytol.*, **93**, 187–279.
39. Daniels, G. R. and Deininger, P. L. (1985) *Nucleic Acids Res.*, **13**, 8939–8954.
40. Daniels, G. R. and Deininger, P. L. (1991) *Nucleic Acids Res.*, **19**, 1649–1656.