Sequence conservation in Alu evolution

Damian Labuda and George Striker[1]

Génétique Médicale, Centre de Recherche, Hôpital Ste-Justine, Université de Montréal, 3175 Côte Ste-Catherine, Montréal, Québec H3T 1C5, Canada and [1]Max-Planck-Institut für Biophysikalische Chemie, 3400-Göttingen, FRG

Received November 18, 1988, Revised and Accepted March 6, 1989

ABSTRACT

A statistical analysis of a set of genomic human Alu elements is based on a published alignment and a recent classification of these sequences. After separation of the Alu sequences into families, the consensus sequences of these families are determined, using the correct weighting of the unidirectional decay of CG-dinucleotides. For, the tenfold greater mutation rate at CG's requires separate consideration of an independent clock at every stage of analysis. The distributions of the substitutions with respect to the new consensus sequences, taking the CG and the non-CG-nucleotide positions separately, lie far closer to the expected distributions than the total diversity. Computer analysis of the folding of RNAs derived from these sequences indicates that RNA secondary structure is conserved among Alu families, suggesting its importance for Alu proliferation and/or function. The folding pattern, further substantiated by a number of compensatory mutations, includes secondary structure domains which are homologous to those observed in 7SL RNA and a defined region of interaction between the two Alu subunits. These results are consistent with a model in which a small number of conserved Alu master genes give rise via retroposition to the numerous copies of Alu pseudogenes, that then diversify by random substitution. The master genes appeared at different periods during evolution giving rise to different families of Alu sequences.

INTRODUCTION

More than half a million Alu elements (1,2) are interspersed throughout the human genome by retroposition, i.e. by reintegration of reverse-transcribed copies of Alu RNAs (3). Alu elements are composed of two subunits, left and right, both homologous to the 7SL RNA sequence truncated by internal deletions (4). Although no function has been ascribed to this genetic material, its high genomic representation and recent reports of Alu involvement in mutations causing genetic defects (5–8) indicate an important role in primate evolution. Despite variability, all Alu elements can be related to an average consensus sequence, as done by Kariya et al. (9) from an alignment of 50 Alu sequences (for a review see ref.3). Recently, Jurka and Smith (10; cf. also 11–14) were able to classify Alu elements into families, J, Sa, Sc and Sb, based on the presence of correlated nucleotide substitutions in a number of sequence positions. The age of these families is in the above order, as judged both by their mutational diversity and their divergence from the 7SL RNA sequence.

Inspection of the Alu alignment indicates a high proportion of TG and CA dinucleotides in sequence positions corresponding to a CG dinucleotide in the consensus (cf. also 11). This can be explained by a higher transition rate of C to T due to CG methylation at the DNA level (15). A quantitative analysis of this CG-effect on Alu elements seems warranted, for, if this effect is of the order of 10:1 in pseudogenes (16), it should be taken into account in the compilation of their consensus sequences and when analyzing their sequence diversity.

A further question which posed itself was, if the diversity within the families was distributed randomly, and this hypothesis was yet to be quantitatively tested, what can we learn about the mechanism of Alu amplification and their sequence evolution. It was with these questions in mind that we undertook a statistical analysis of a set (9) of genomic Alu sequences and, using these results, a structural study of the secondary folding of Alu RNAs which are presumed intermediates in amplification of these elements by retroposition.

METHODS

The 50 Alu genomic sequences aligned by Kariya et al. (ref. 9, full listing of sequences given there) were used, with some improvement of the alignment to reduce the number of deletions. Programs were written to analyze the differences between the sequences and obtain distributions of these differences, both in terms of differences per nucleotide position and per sequence. These programs were applied both to the entire set of sequences and to the individual families. Gaps are omitted from all the statistics.

The sequences were divided into the families J, Sa, Sb and Sc using 'diagnostic' positions of Jurka & Smith (10). This subdivison proved tractable, except that 3 sequences, essentially more similar to the J family but differing in a number of positions were subclassified in a class J*. Further analysis showed other differences of these sequences to J, so they were omitted from classification.

As there is a far higher tendency for transition mutation at CG-dinucleotides than at other sequence positions (cf. Introduction and Fig. 1), the programs for consensus sequences of the families favored making CG dinucleotides, i.e. if following a C nucleotide a position had a majority of A's, but some G's, a G was chosen, and similarly for a C, mutated to T, to the left of a G. Otherwise the most frequent nucleotide was chosen, resulting in what can be called a weighted consensus sequence.

Given a set of sequences diverging randomly from a common ancestor, so long as that divergence is small it can be treated linearly:

$$m = m(t) = k_r t$$

where m is the linear divergence density found after the elapsed time t, i.e. the number of divergent positions (substitutions) found divided by the total number of possible substitutions. Thus, with a value for the rate k_r, we can estimate the time t since retroposition of the sequence. If we have n sequences with such behavior at a number of positions, and construct a histogram of the number of substitutions, s, at each of these positions, we may expect that histogram, as the sum of independent processes with low probabilities, to follow Poisson statistics:

$$P(s) = e^{-mn} (mn)^s/s!$$

where $P(s)$ is the probability of exactly s substitutions per nucleotide position. We will call sequence positions obeying these statistics 'random positions'.

The high probability of transition mutations at CG dinucleotides and resulting depletion of these dinucleotides renders this process nonlinear. Decay kinetics apply:

$$d = d(t) = 1 - e^{-k_{cg}t}$$

where k_{cg} is the rate of CG decay, and d is the density of CG dinucleotides found to be altered. Another interpretation of the above equation is that d is the poissonian probability of a non-zero event:

$$-\ln (1-d) = k_{cg}t = k$$

in a trial of a process with a probability k of occuring once, k^2 of occuring twice, etc.

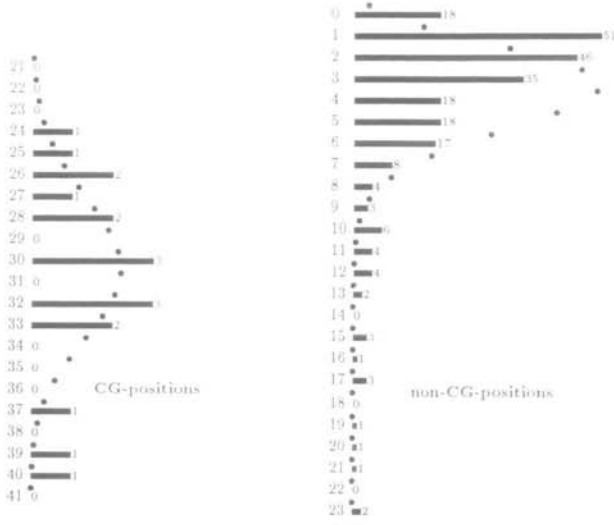

Figure 1. Distribution of the frequency of given number of substitutions per position in the 47 Alu sequences analyzed relative to the general consensus (9). CG-dinucleotides are separated from non-CG positions. The length of the bars corresponds to the number of positions with a given number s (ordinate on the left) of substitutions. Circles correspond to the expected distributions of these frequencies, on the basis of the divergence density found, indicating the probability $P(s)$ of exactly s mutations at a position (see Methods) multiplied by the total number of positions. Note that the scale for CG's is different, there being only 18 dinucleotide positions available.

It is this dinucleotide divergence density k that must be used as a comparison to the linear divergence density m at non-CG positions except that k applies to two nucleotides, and must thus be compared to twice the linear divergence density m. Poisson statistics no longer describe CG-decay due to the high frequency, the binomial distribution must be used instead:

$$P(s) = \binom{n}{s} d^s (1-d)^{n-s}$$

where again $P(s)$ is the probability of finding exactly s dinucleotides of a given CG position altered in n sequences.

Folding of Alu RNAs was investigated using the RNA secondary structure prediction program developed by Zuker & Stiegler (17). A copy of the IBM PC version of this program was kindly provided by Dr M. Zuker (NRC, Ottawa).

RESULTS

The distribution of divergence (per nucleotide position) of all sequences from the general consensus (Fig. 1) divided into CG dinucleotide and non-CG positions shows that the probability of CG-decay is far greater but distributed randomly. The less divergent non-CG positions are distributed around a maximum well below that of the Poisson distribution calculated for their divergence density, while some positions lie far above that distribution. This justifies separation of the sequences into families and determination of weighted consensus sequences for these families (Table I, full computer readable listing of sequences and their family subdivisions is available from the authors).

Table I Comparison of the general consensus of Kariya et al. (9) with 7SL RNA (29) and with the weighted consensus sequences of Alu families, J, Sa, Sc and Sb.

```
    -..C.....C ......G.GU .......G.. .....UAC.C .......U.. 7SL RNA 1-49
  1 GGCTGGGCGT GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCCGA Kariya et al.
    ...C...... ..C..... ......... .........C ......... J
    ...C.....C ......... ......... ......... ......... Sa
    ...--...-. -......... -......... .-........ .......T.. Sc
    ...C.....C ......... ......... ......... ......... Sb

    ..CU..A... ..G.U....U C......... ----..... ...G..... 50-79 267-282
 51 GGTGGGTGGA TCACCTGAGG TCAGGAGTTC AAGACCAGCC TGGCCAACAT
    ..C...A... .TG.T....C C......... G......... ...G.G....
    ..C...C... ......... ..G....... G......... .........
    ......CA.. ....**.... ...A...A.. G......T.. .........C
    ..C...C... ..G.**.... .......A.. G......T.. ....T....C

    A.C..G.... .......--- ---------- ------.... ....C..... 283-299 1-14
101 GGTGAAACCC CGTCTCTACT AAAAATACAA AAATTAGCCG GGCGTGGTGG
    A.C..G.... .........A ......... ......... .........
    ......... ......... ......... ......... .........
    ......... ......... ......... ......... .........
    ......... ......... ......... ......... .........

    ....U..... ..G..... ......... ........U. ...G..... 15-64
151 CGCGCGCCTG TAATCCCAGC TACTCGGGAG GCTGAGGCAG GAGAATCGCT
    ......... ......... ......... ........G. ...G......
    ......... ......... ......... ......... .........
    ..T....... ......... ......... ......... .........
    ..G....... ..G..... ......... ......... ......G..G

    ...GU..A.. ..U.CUG..C ..U....C.. ---.....-. ......... 65-94 101-105
201 TGAACCCGGG AGGTGGAGGT TGCAGTGAGC CGAGATCGCG CCACTGCACT       256-266
    ...G...... ...C.....C ......... ..T....... .........
    ......... ...C..... ...G..... ......... .........
    ......... ...C..... ......... ......... .........
    ......... ...C....C. ......... ......... .........

    ......... .A...U.... ....C..... ..                     267-298
251 CCAGCCTGGG CGACAGAGCG AGACTCCGTC TC
    ..G....... ....G..... ....C..... ..
    ......... ......... ......... ..
    ......... ......... ......... ..
    ..G....... ......... ......... ..
```

Dots indicate agreement with the general consensus. The top line shows two partial copies of 7SL RNA, corresponding to Alu-left and Alu-right subunits, with corresponding numbering on the right. The dashes in that line are above positions in the Alu sequence that do not correspond to 7SL RNA fragments. The dashes in Sc family sequence indicate places where we have no consensus because of a number of deletions; according to ref. 10 these positions do not differ in Sc from other Alu families.

After this division and comparison with the 'local' consensus of the families (Table I) the set of non-CG positions with high divergence is considerably reduced (Fig. 2). Nevertheless, if we compare the observed distribution with that calculated from the divergence density we still see a considerable difference (Fig. 2 *left*). If, however, we calculate that density ignoring all positions with more than 5 nucleotide substitutions, we find that, at least up to four substitutions per nucleotide position the data approaches the

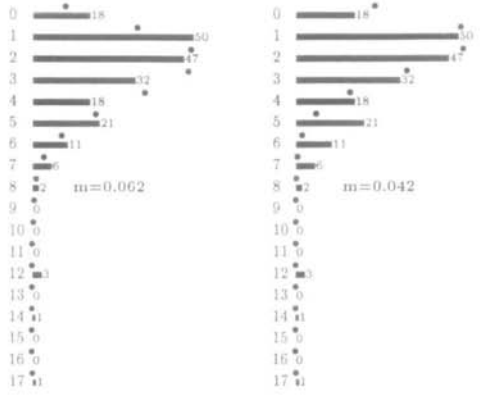

Figure 2. Distribution of the frequency of substitutions (cf. Fig. 1) at non-CG positions but now relative to the local consensus of families. At the *left* the comparison is with the Poisson distribution based on the divergence density at all these positions. The comparison on the *right* is with the Poisson distribution ignoring all positions with more than 5 substitutions (see text). The *left* distribution shows poor agreement with the actual data, while that on the *right* appears to be in agreement for values of s below 5.

Poisson distribution (Fig. 2 *right*). By excluding all positions with more than 5 nucleotide substitutions over the entire set of sequences, we find that the remaining 192 non-CG sequence positions fit their Poisson distributions well in all the families (Fig. 3), and the distribution of substitutions per sequence within each family also appear in agreement with this distribution (data not shown). As more than 98% of this distribution lies in the range considered (less than 10% divergence) we may take these as the set of random non-CG positions, and use the densities of divergence found as the applicable linear densities m characteristic for each Alu family (Table II; cf. Methods). Similarly the distribution of the CG's decayed shows a single albeit far faster process obeying the binomial distribution (Fig. 4), thus indicating random accumulation of substitutions in these positions as well.

A large part of the divergence between the individual Alu sequences is accounted for by the separation of CG dinucleotide positions and division into families (Figs. 1–3). Thus, the divergence of the elements from the consensus of their families consists primarily of a low level of random substitutions at non-CG positions and unidirectional random CG-decay. Most of the divergence that appeared not to be randomly distributed lies in the difference between the family consensus sequences, which thus appear to be 'master' sequences, from which individual Alu repeats originated by retroposition. It will, therefore, be interesting to study the structure of the RNA derived from these master sequences, and see what effect the positional differences between them have on their secondary folding

Figure 3. Distribution of the frequency of substitutions (cf. Fig. 1) at all non-CG positions with a total of five or less subsitutions, by families, compared with the appropriate Poisson distributions. These are the positions referred to as 'random' in the text due to their agreement with this distribution. The scales are the same for all families.

Table II Comparison of two clocks timing Alu diversity.

Family	m	d	k	k_{cg}/k_r
J	0.082	0.82	1 71	20.9
Sa	0 046	0.70	1.21	26.5
Sc	0.036	0.53	0.75	20.9
Sb	0 027	0.38	0.49	17.8

Linear divergence density, m, at 'random' non-CG positions is compared with divergence density, d, at CG-positions which is corrected for CG-decay kinetics giving k. The ratio of k/m corresponds to the ratio of substitution rates, k_{cg}/k_r, characterizing the relative mutation rate in a CG-clock and a general clock. Average ages of the Alu families calculated from these data are given in the text.

(see below). However, as some non-CG positions lie outside the random distribution (Fig. 2) there is still a possibility of further family subdivision. For example, these 'non-random' positions include all identified as variant positions in the Sa family (Table III).

Both the linear divergence density m of the 'random' substitutions and the divergence density $k = -\ln(1-d)$ of CG-decay decrease from the J family to Sa and further to Sc and Sb (cf. Table II). If we divide k by m we get a factor close to 21 for J and Sc and slightly less for the apparently youngest family Sb. The factor for Sa is higher, which, together with its deviation from Poisson statistics might indicate that this family is actually a composite of more than one family with more than one local consensus sequence as suggested above. The factor of 21 corresponds to a 10.5-fold greater mutational rate per nucleotide in CG compared to non-CG random positions. In other words, the 'CG-clock' runs 10.5 times faster than the 'general clock' at non-CG positions.

As seen in Table II, both the general clock and the CG-clock essentially agree in determining the relative ages of the families. If we use $k_r = 1.5*10^{-9}$ per nucleotide position per year as the substitution rate at non-CG positions and the m values from Table II, we obtain 55 Myr (million years) for J family, 31 Myr for Sa, 24 Myr for Sc and 18 Myr for Sb. The same values would be obtained using CG-clock (k's from Table II) except that the average age of the Sa family would then be estimated at 38 Myr. This places the J family at the begining of mammalian radiation and primate divergence (18) and is consistent with the appearance of the Sb family before the divergence of human and gorilla (19). The J sequence also shows more similarities than the S families with

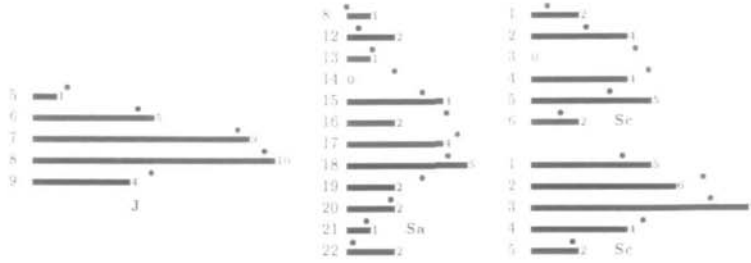

Figure 4. Distribution of the frequency of decay of CG dinucleotides (cf. Fig. 1) from the local consensus, by families, compared with the expected binomial statistics.

Table III Conservation of Alu RNA secondary structure among Alu families.

Base and position	Alu families			
in Sa	J	(Sa)	Sc	Sb
T 13	C unp→C:G152			
T 40	C T:A33→unp			
C 48			T unp→T:A136	
C 53			T C:G112→T:G112	
C 57	A unp			
G 58			A unp	
C 62	T C:G104→T:G104			
A 63	G A:T103→G:C103			G A:T103→G:T103
C 65	T C:G101→T:A101		*	*
T 66			*	*
G 70	C C:G94→G:C94			
T 71	C T:G93→C:G93			
G 73	A G:T91→A:T91		A G:T91→unp	A→G:T91→A:T91
G 74			A G:C90→A:T91	
T 78			A T:A84→A:T88	A T:A84→unp
G 88			T unp→T:A78	T unp→T:A76
G 93		(A) G:T71→A:T71		
C 94	G C:G70→G:C70			
C 95				T C:G69→T:G69
A 96	G unp			
T 100			C unp	C unp
G 101	A G:C65→A:T65			
T 103	C T:A63→C:G65	(A) T:A63→unp		
A 106	G unp			
T 120	A unp			
C 153		(G) C:G11→unp	T C:G11→T:G12	G C:G11→unp
A 163		(G) unp		G unp
A 189	G unp→G:C277			
A 194	G unp			
C 197				G C:G270→unp
T 200				G T:A267→unp
A 204	G unp→G:C261			
G 219				C unp
T 220	C unp			
G 224	A G:T245→A:T245	(A) G:T245→A:T245	A G:T245→A:T245	A G:T245→A:T245
A 233	T unp			
C 244		(T) C:G225→T:G225		
A 253	G unp→G:C214			G unp→G:C214
A 265	G A:T201→G:T201			
G 272		(A) unp		
T 275	C T:A192→C:G191			

List of differences between Sa subfamily and J, Sc, and Sb sequences with the resulting modifications in secondary interactions described on the right. In brackets we include variant nucleotides occuring in Sa at frequencies higher than expected for random mutations. ('unp' denotes an unpaired position and '*' a deletion; '→' denotes the change in base-pairing from the Sa RNA to a subfamily structure; underlined are conserved and compensatory changes including alternative base-pairing; T is used instead of U as the RNA sequences were derived from genomic sequences).

SECONDARY STRUCTURE OF Alu J RNA

ENERGY = -157.5

SECONDARY STRUCTURE OF Alu Sc RNA

ENERGY = -139.8

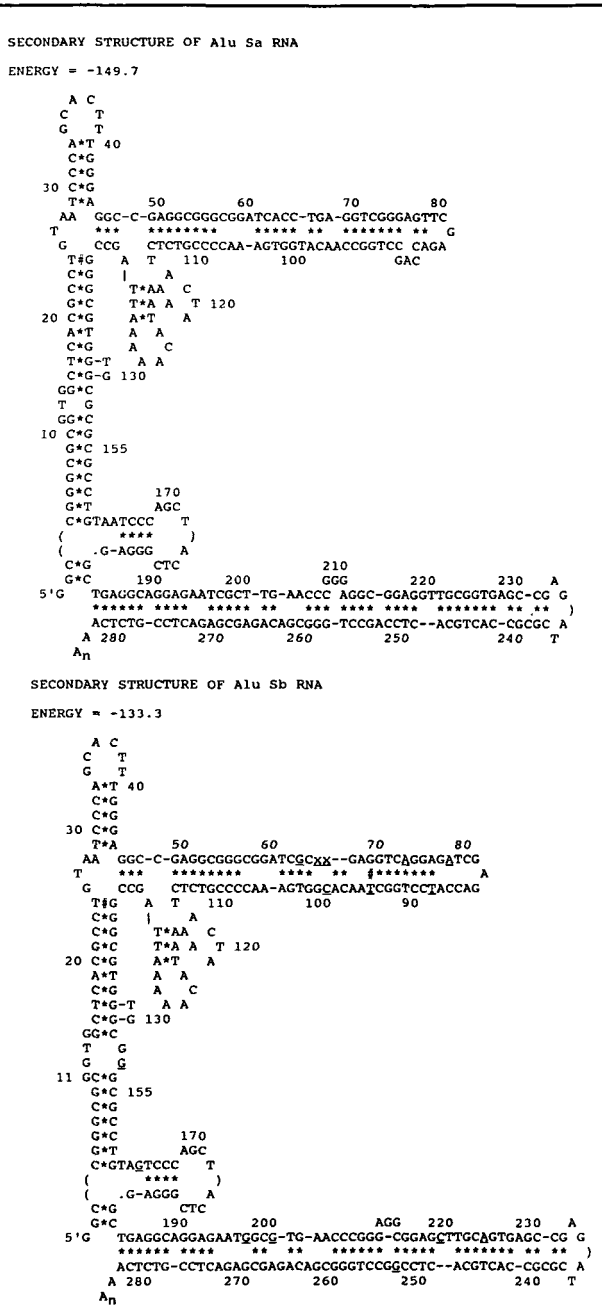

Figure 5. Predicted secondary structures of Alu RNAs of four families. The sequence of Alu Sa RNA is used for comparison with other sequences—underlined positions in J, Sb, and Sc are those which differ from Sa. Free energy is in kcal/mol, 'x' denotes deletion, '*' denotes predicted base-pairing whereas '#' denotes possible base-pairing which was not predicted by the program. T used instead of U emphasizes that these RNA are derived from DNA sequences.

the sequence of 7SL RNA (Table I, ref.10). The estimate of $k_r = 1.5*10^{-9}$ was made from the number of mutations accumulated in non-CG positions in seven orthologous Alu elements in human and chimpanzee since their divergence 5 Myr ago (20) and agrees with the overall rate of DNA mutations in primates estimated previously (21).

Fig. 5 shows the predicted secondary folding of putative Alu RNAs transcribed from Alu J, Sa, Sb, and Sc genes. In all cases the same overall secondary structure is obtained with characteristic secondary structure domains in Alu-left and Alu-right subunits (cf. Introduction and caption to Table I) and with a helical region of secondary interaction between Alu-left and Alu-right, which we will call LRI (left-right interaction). This suggests (i) that sequence conservation among Alu families is connected with the conservation of secondary structure of the transcribed RNAs and (ii) that correlated nucleotide differences among these families do not affect the pattern of RNA folding. Arbitrarily choosing Alu Sa as a standard sequence we have examined the effect of these nucleotide differences on the predicted secondary structures. In Fig. 5 the nucleotide positions in J, Sb and Sc differing from the Sa sequence (cf. Table I) are underlined. Table III summarizes consequences of these changes on base-pairing in individual structures (the effect of variant positions within the Sa family is also examined).

Among 41 positions listed in Table III, twelve positions always occur in unpaired regions. Changes in these positions are consistent with the predicted secondary structures as they do not affect the RNA folding. A tendency is also observed to lose base-pairs going from J through Sa, Sc to Sb, which is also reflected in increasing minima of free energy change, from −157.5 in J to −133.3 kcal/mol in Sb. Most interesting, however, are conservative and compensatory nucleotide changes which preserve the same base pairing pattern among family RNAs providing independent evidence for the predicted secondary interactions. Conservative changes such as A:T G:T or G:T G:C are observed at 9 instances (positions: 53, 62, 71, 73, 93, 95, 224, 244 and 265). Coordinated compensatory changes affect six positions: A63:T103 pair is replaced by a C:G pair in J, C65:G101 by a T:A and C70:G94 by a G:C. Some nucleotide changes result in local rearrangements of secondary interactions without affecting the overall folding pattern; this concerns the LRI region near positions 13 and 153, double stranded regions in Alu-right, near positions 189 and 275, and positions 204 and 253, as well as an extended hairpin in Alu-left involving positions 73 to 91. Deletions in Sb and Sc sequences, at positions 65 and 66 also affect the local base-pairing. For a complete picture of these local rearrangements Fig. 5 must be consulted as modifications in base-pairing also occur in positions adjacent to those listed in Table III.

Since dimeric Alu elements descend from the 7SL RNA sequence (4) (cf. Table I) we examined the 7SL RNA secondary structure predicted using Zuker's approach (Fig. 6 – only 'Alu fragment' of 7SL RNA is shown). This structure is virtually identical to that described earlier from RNase digestion studies and evolutionary comparisons (22, 23). It consists of two hairpins involving positions 3 to 24 and 28 to 43, and of a helical domain formed by complementary regions between positions from 48 to 94 (extending further to position 105) and positions from 256 to 298. This helical domain and the hairpin 28–43 are conserved in both Alu RNA subunits, left and right, whereas the hairpins 3–24 have created a region of secondary interaction between Alu-left and Alu-right (LRI) in the dimeric Alu RNA sequence (Fig. 5). Nine and three nucleotide positions are different in the first 50 nucleotide fragment of 7SL RNA sequence compared to Alu-left and Alu-right, respectively, including a variant purine in position 163 (cf. Table I and Fig. 6). All of these changes

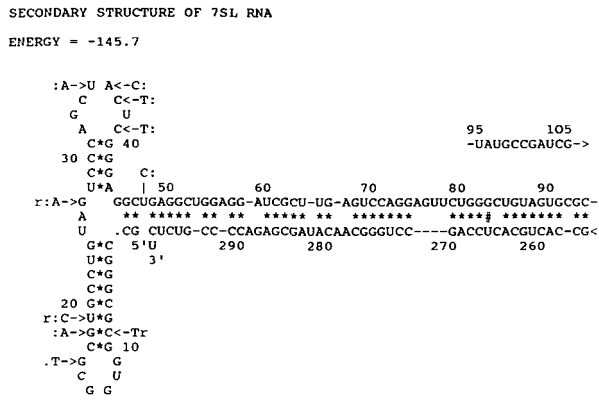

Figure 6. Predicted secondary structure of 7SL RNA. Only fragments homologous to Alu are shown (cf. Table I). Arrows indicate nucleotide positions in the first 50-nucleotide portion of the 7SL RNA sequence which are substituted in Alu-left (marked with ' ') or Alu-right subunit (marked with 'r').

occur in or are adjacent to LRI regions and are conserved among Alu families, except that in the J-left sequence there is a C rather than T in position 13 and position 40 remains C as in 7SL RNA. Their effect on Alu RNA folding has been examined by comparing the predicted Alu RNA structure with that of a hypothetical Alu RNA retaining in both left and right subunits all nucleotides as in 7SL RNA (Fig. 7). In this hypothetical Alu RNA, the LRI based on self-complementarity of hairpins 28–43 is energetically favored over that based on self-complementarity of hairpins 3–24, which are present in left and right subunits. The resulting structure is depicted in Fig. 7*b* (for simplicity only regions involved in LRI are shown). The nucleotide changes from 7SL RNA to Alu sequence are indicated by arrows together with their effect on the observed base-pairing. For comparison, the folding of the same sequence region of the actual Sa sequence is illustrated in Fig. 7*a*, where arrows indicate nucleotides present in 7SL RNA and the consequences of these nucleotide changes on base-pairing are also shown.

In summary, the nucleotide changes from 7SL RNA to Alu sequence in the discussed region lead to a disruption of four base-pairs in the structure shown in Fig. 7*b* and to a creation of two additional base-pairs in the structure in Fig. 7*a*. As a result the latter structure becomes energetically favored. These coordinated nucleotide changes from 7SL RNA to Alu sequence leading to the specific LRI region in Alu RNA provide additional evidence for the predicted folding of Alu RNAs (Fig. 5) indicating as well the importance of Alu RNA secondary structure for Alu function and/or amplification through retroposition.

DISCUSSION

In Fig. 3 we have seen that the differences between the consensus sequences for the Alu families and the actual sequences for a large number of the nucleotide positions not involved in fast CG-dinucleotide decay are Poisson distributed. Similarly, the decay at the positions determined as CG-dinucleotides by weighting these consensus sequences follow the expected binomial distribution (Fig. 4). This is an indication that these consensus sequences are the starting point of sequence diversion, and thus actually correspond to master sequences which were copied into the genome in large numbers by retroposition. This also provides

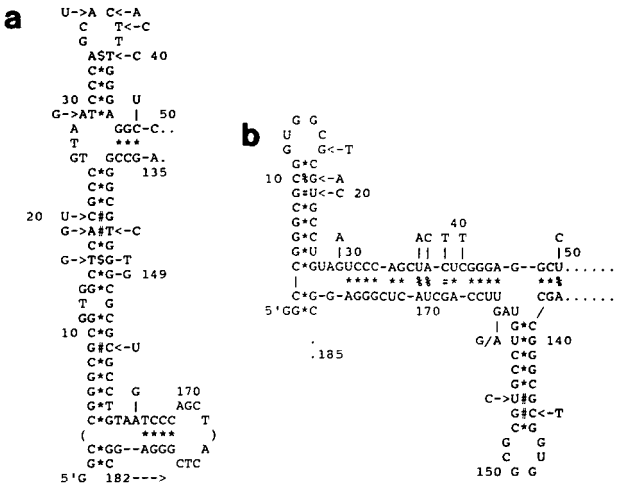

Figure 7. Comparison of the regions of Alu-left and Alu-right interaction (LRI) predicted for Alu Sa RNA (*a*) and for a hypothetical Alu RNA (*b*) constructed from 7SL RNA fragments. Arrows indicate nucleotide positions which were changed in Alu Sa as compared to 7SL RNA sequence. The consequences of these nucleotide changes for the secondary structure in (a) and in (b) are indicated ('$' denotes a gain and '%' a loss of a base-pair, whereas '#' indicates either conservative or compensatory nucleotide charge).

evidence against a 'cascade' model of Alu amplification, in which each newly retroposed Alu could be seen as a potential source of the next daughter Alu retroposon. (In such a situation no random distribution of substitutions from the consensus would be observed.)

Independent evidence for the origin of Alu elements in a small number of conserved sequences comes from the analysis of Alu RNA folding. The weighted consensus sequences of the families (Table I) share most of the sequence both in CG dinucleotide and in non-CG positions. Moreover, they also conserved the same overall folding of the putative transcribed Alu RNAs (Fig. 5) despite numerous differences in their sequences. In fact, the effect of these variant positions on the RNA folding is either neutral or consists of conserved or compensatory substitutions (Table III) which provides an independent confirmation for the predicted secondary structure. While preserving certain features of the 7SL RNA folding which are found repeated in both subunits of their RNA (Figs. 5 and 6) Alu sequences have evolved a secondary interaction betweeen their subunits (Fig. 7). The sequence fragments in Alu-left and Alu-right subunits which participate in this interaction are conserved in all Alu families. All this points to the Alu RNA as a target of selection during Alu evolution and suggests that its secondary structure is an important factor in proliferation of Alu sequences, consistent with their amplification by retroposition. We speak here about proliferation rather than function since heterogeneity of Alu repeats among mammals argues against a specific cellular function for Alu sequences other than that connected with their capacity for retroposition. The particular significance of the human Alu dimeric structure and of different domains of its RNA folding for retroposition remains to be demonstrated.

The weighting we have used to determine family consensus depends upon the realization that the diversification of Alu sequences involves two different clocks. These clocks run simultaneously, albeit with different kinetics (Table II). The general clock is essentially

Figure 8. Frequency by family of all possible dinucleotides found at positions with CG in the local consensus (gaps being ignored). R and Y stand for purine and pyrymidine, respectively. The comparison distribution, indicated by thin bars, is calculated from the assumptions and data of Table II.

linear over short time periods, whereas the fast CG-clock, due to the unidirectional depletion of CG dinucleotides, must be treated exponentially. The occurence of a separate CG-clock can be related to the methylation of CG dinucleotides in Alu integration sites, expected in a heavily methylated human genome. Since this is also the case for other vertebrate genomes (15), this second clock should be applicable to virtually all vertebrate DNA sequences. A ten-fold increase in a CG dinucleotide substitution rate characteristic of Alu sequences (Table II) has also been observed in several pseudogenes (16) and in the human Factor VIII gene (24). This requires an adequate weighting of the CG and the non-CG sequence positions, which in the case of multi-sequence families such as Alu result in the weighted consensus sequences (Table I). Similar weighting should be applied in evolutionary considerations when other, either single or multi-copy, vertebrate sequences are compared.

By separating CG and random non-CG positions we were able to calculate the corresponding divergence densities from the weighted consensus sequences and estimate approximate average ages of the families (cf. Results). Jurka and Smith (10) find a mean agreement, excluding gaps, of 82.8%, 88.6%, 89.5% and 92.7% for J, Sa, Sc and Sb sequences, respectively, which gives overall divergence densities of 0.172, 0.114, 0.105 and 0.073. These numbers are of necessity far higher than our m, as they include CG-decay as well as many positions we consider 'non-random', although they do exclude their diagnostic positions. The m values given in Table II, however, are not just calculated densities, but correspond to Poisson distributions that actually apply in the various families. We saw that the density of 0.062 in Fig. 2 did not fit the data, while using $m=0.042$ for all the data improved the fit. Using the family specific m values for the substitutions in the families improves it further (data not shown). The final step of excluding the 'non-random' positions gives very good agreement between the data and the Poisson distributions (Fig. 3). Similarly, Fig. 4 shows that the CG-dinucleotide positions also decay according to their expected binomial distributions, which additionally provide evidence that these positions are all intact in the master sequences. Nevertheless, the presence of a small portion of 'non-random' positions raises the possibility that further subdivision and thus more master sequences will be found once a significant percentage of the more than half a milion Alu elements have been sequenced. It could be that this will lead to the expected random distribution around their weighted consensus sequences. But it is also possible that there is a small intrinsic heterogeneity in the master genes due to a certain number of branching sequences, or that some diversity is the result of the mechanism of amplification itself.

An analysis of the distribution of depleted CG dinucleotides (Fig. 8) for the various families shows the expected favoring of CA and TG products in Sa, Sb and Sc, explained by the origin of this decay in methylation (15), which no longer favors the C→T or G→A

transition once the partner has already mutated. Nevertheless we find a significantly higher occurence of TA in the J family, statistically compatible with an equal probability of the second mutation, either CA or TG to TA, which cannot be explained by CG methylation alone. The Sa family also has slightly more TA substitutions for CG than would be expected. This effect might be due to a homologous recombination between Alu sequences with corresponding CG dinucleotides mutated in the first (TG) and the second position (CA), respectively. Thus, double T:G or C:A mismatches formed upon exchange of strands in the resulting DNA duplexes, repaired to T:A base-pairs in a considerable percentage of all cases (25) would result in a TA dinucleotide in a sequence position originally occupied by a CG. Such a mechanism occurs rarely until a large percentage of the CG dinucleotides have undergone one substitution, which is the case for the older J family.

In conclusion, Alu elements can be seen as closely related families of pseudogenes amplified through RNA intermediates from a number of conserved and transcriptionally active master genes which appeared at different time periods during primate evolution (cf. also 10–13). As discussed above, conservation of Alu RNA secondary structure suggests specific selection of these sequences for retroposition. According to our analysis the master genes are protected against mutations promoted by CG-methylation. Therefore, one may speculate that *in vivo* only those Alu sequences which reside in non-methylated regions of DNA are independently expressed and can undergo retroposition, thus explaining CG-dinucleotide conservation. Such a preselection for potentially amplifiable Alu sequences at the level of transcription could be additionally controlled by an upstream promoter as is the case of 7SL RNA (26). Alternatively, a selection could occur at the level of RNA structure. Considering the presence of an internal RNA polymerase III promoter in Alu elements which, at least *in vitro*, is sufficient to activate their transcription (27), a variety of Alu retroposons could be transcriptionally active. However, those sequences which become methylated upon reintegration would undergo rapid mutations at CG dinucleotides reducing the chances of the transcribed RNA for retroposition by affecting its secondary structure, whereas those conserving their master sequence could undergo further rounds of retroposition thus promoting survival of the sequence. Experimental characterization of the independently transcribed Alu RNAs, as well as of their transcriptionally active genes should distinguish between these, in principle, non-mutually exclusive possibilities. In this context it is interesting to note that the only primate Alu-like independent RNA transcript sequenced to date, from cynomolgus monkey (28), appears to retain all its CG dinucleotides intact, when compared with the left subunit of the J family consensus sequence.

ACKNOWLEDGEMENTS

The statistical programs were prepared and run on the Vax 8650 of the Gesellschaft für wissenschaftliche Datenverarbeitung, Göttingen (GWDG). The Zuker analyses were run on the IBM/AT PC clone at Ste-Justine Hospital with programs made available by Dr Michael Zuker, which are gratefully acknowledged. We are grateful for the use of BITNET, that has made our cooperation possible. We thank Dr Louise Simard, Daniel Sinnett and Jean-Marc Deragon for discussions. This work was supported by Fondation de l'Hôpital Ste-Justine and by a Scholarship from Fonds de Recherche en Santé du Québec (D.L).

REFERENCES

1. Rinehart, F.P., Ritch, T.G., Deininger, P.L. & Schmid, C.W. (1981) *Biochemistry* **20**, 3003–3010.
2. Hwu, H.R., Roberts, J.W., Davidson, E.H. and Britten, R.J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 3875–3879.

3. Rogers, J.H. (1985) *Int. Rev. Cytol.* **93**, 187–279.
4. Ullu, E. and Tschudi, C. (1984) *Nature* **312**, 171–172.
5. Lehrman, M.A., Russel, D.W., Goldstein, J.L. and Brown, M.S. (1987) *J. Biol. Chem.* **262**, 3354–3361.
6. Myerowitz, R. and Hodikyan, N.D. (1987) *J. Biol. Chem* **262**, 15396–15399.
7. Markert, M.L., Hutton, J.J., Wiginton, D.A., States, J.C and Kaufman, R.E. (1988) *J. Clin. Inv.* **81**, 1323–1327.
8. Rouyer, F , Simmler, M-C., Page, D.C. and Weissenbach, J. (1987) *Cell* **51**, 417–425.
9. Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. and Matsubara, K. (1987) *Gene* **53**, 1–10.
10. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4775–4778.
11. Britten, R.J., Baron, W.F., Stout D.B. and Davidson E.H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770–4774.
12. Quentin, Y. (1988) *J. Mol. Evol.* **27**, 194–202.
13. Willard,C., Nguyen,H.T. and Schmid, C.W. (1987) *J. Mol. Evol.* **26**, 180–186.
14. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. & Deininger, P. (1987) *Mol. Biol. Evol.* **4**, 19–29.
15. Bird, P.A. (1987) *Trends Genet.* **3**, 342–347.
16. Bulmer, M. (1986) *Mol.Biol.Evol.* **3**, 322–329.
17. Zuker, M. and Stiegler, P. (1981) *Nucl. Acids Res.* **9**, 133–148.
18. Gingerich, P.D. (1986) *Mol. Biol. Evol.* **3**, 205–221.
19. Trabuchet, G., Chebloune, Y., Savatier, P., Lachuer,J., Faure, C., Verdier, G. Nigon, V.M (1987) *J. Mol. Evol.* **25**, 288–291.
20. Sawada, I., Willard, C., Shen, C-K.J., Chapman, B , Wilson A.C. & Schmid, C.W. *J. Mol. Evol.* **22**, 316–322 (1985).
21. Britten, R.J (1986) *Science* **231**, 1393–1398
22. Zwieb, C. (1985) *Nucl. Acids Res.* **13**, 6105–6124.
23. Gundelfinger, E.D., DiCarlo, M., Zopf, D. and Melli, M (1984) *EMBO J.* **3**, 2325–2332.
24. Youssoufian, H., Antonorakis, S.E., Bell, W., Griffin, A.M. and Kazazian, Jr. H.H. (1988) *Am. J. Hum. Genet.* **42**, 718–725.
25. Brown, T.C. and Jiricny, J. (1988) *Cell* **54**, 705–711.
26. Ullu, E. and Weiner, A.M. (1985) *Nature* **318**, 371–374.
27. Perez-Stable, C. and Shen, C-K.J. (1986) *Molec. Cell. Biol.* **6**, 2041–2052.
28. Watson, J.B. and Sutcliffe, J.G. (1987) *Molec. Cell. Biol.* **7**, 3324–3327.
29. Ullu, E., Murphy, S. and Melli, M. (1982) *Cell* **29**, 195–202.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.