J Mol Evol (1987) 26:180-186

Journal of
Molecular Evolution

Existence of at Least Three Distinct Alu Subfamilies

Cary Willard, Hiep Thieu Nguyen, and Carl W. Schmid

Department of Chemistry, University of California at Davis, Davis, California 95616, USA

Summary. Computer-assisted sequence analysis of human Alu family members reveals that Alu repeats belong to one of at least three subfamilies. The insertion of human Alu repeats can be represented by three episodic bursts, each of which was founded by a distinct master sequence.

Key words: Alu subfamilies — Human genome — Computer-assisted sequence analysis

Introduction

There are about 500,000 Alu family members dispersed throughout the human genome (Schmid and Jelinek 1982; Schmid and Shen 1985). Individual members of this family share a 282-bp consensus sequence; the average divergence of individual members from this consensus is about 14% (Deininger et al. 1981; Schmid and Shen 1985). This sequence divergence might result from either different founder sequences or the accumulated mutations of individual members subsequent to their genomic insertion.

Phylogenetic comparisons, reviewed in the Discussion, show that some Alus have resided in the human genome for a sufficient time to account for much of their sequence divergence from the consensus (Sawada et al. 1985; Sawada 1986; Sawada and Schmid 1986). Although divergence accounts for some of the sequence heterogeneity of Alu repeats, the Alu family could still be derived from several different founder sequences.

Slagel et al. (1986) found that four Alu repeats form a more closely related subfamily of sequences. The remaining Alu repeats (19 out of 23) in their analysis do not fall into distinct groups. Bains (1986), using a dendrogram analysis on a larger sample of 59 Alu repeats, concluded that there are no recognizable Alu subfamilies, implying that Alu members are derived from a very large pool of precursors.

Additional Alu sequences have been included in this analysis to provide a more extensive test of the disparate conclusions of Bains and Slagel et al. Our approach toward the question of the number of ancestral Alu repeats is statistical. The progeny of a single founder sequence should share a recognizable subfamily consensus sequence. Assuming that mutations occur at random in these members (Discussion), the pairwise divergence of members of a subfamily should generally approximate a binomial distribution of mutations centered about the mean divergence. Using this approach, we find that the Alu family can be resolved into three distinct subfamilies.

Methods

Eighty-nine Alu family member sequences were compiled from the original literature (Willard, Ph.D. thesis, in preparation). Most of these sequences are listed in available reviews (Schmid and Shen 1985; Bains 1986; Slagel et al. 1986). The entire set is available upon request (Willard, thesis). Alignment of these 89 members agrees with the previously derived total consensus sequence at all but five especially variable sites (Fig. 1; Schmid and Shen 1985). Seventeen of these Alu repeats having either deletions of 20 or more nucleotides relative to the consensus or only partially determined sequences were eliminated from the detailed study. These deletions do not occur at the same site nor do these deleted Alu member sequences share other noteworthy features. The remaining 72 Alu family members exhibit 42 ± 12 (SD) differences relative to the 282-nucleotide consensus sequence,

Offprint requests to: C.W. Schmid

```
GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCCGA GGCGGGCGGA TCACCTGAGG TCAGGAGTTC GAGACCAGCC TGGCCAACAT  ALU CONSENSUS
....A..... .........T .......... .......... .......T.. ......A... ....T....C C......... .......... ...G......  DIVERGED
.......... .......... .......... .......... .......... .......... ....--.... .......A.. .......T.. ....T....C  CONSERVED
.......... .......... .......... .......... .......... .......... .......... .......... .......... ..........  MAJOR

.C..A...AT ......A... T.A....... .T......C. ........A. ...A..TA.. .TG.T....C .........T .......... ..AG...G.N  GALAGO

GGTGAAACCC CGTCTCTACT AAAAATACAA AAATTAGCCG GGCGTGGTGG CGCGTGCCTG TAATCCCAGC TACTCGGGAG GCTGAGGCAG GAGAATCGCT  ALU CONSENSUS
A...G..... T........A .......... ........TA ..T....... .AT.C..... ..G....... .......... ........G. ...G...A..  DIVERGED
.......... .........A .......... .......... ..G.C..... ..G....... .......... .......... .......... ......G..G  CONSERVED
.......... .......... .......... ........T. .......... ...A...... .......... .......... .......... ..........  MAJOR
N.CA.G.... TA........ .......G.. ...C....T. ...A...... TAG..A.... ..G....... ....T..... .........A...G......  GALAGO

TGAACCCGGG AGGCGGAGGT TGCAGTGAGC CGAGATCGCG CCACTGCACT CCAGCCTGGG CGACAGAGCG AGACTCCGTC TC         ALU CONSENSUS
...G...A.. ...TT....C .......... .AT......A .......... .......... .A........ ....C.T... ..         DIVERGED
.......... ........C. .......... .......... .......... .......... .......... .......... ..         CONSERVED
.......... .......... .......... ........T. .......... .......... .......... .......... ..         MAJOR
...G...AA. ..TTT..... ...T...... TAT...NNN. ....A..... .T........ T......T. ......T... ..         GALAGO
```

Fig. 1. Comparison of galago and human subfamily Alu consensus sequences to the overall human Alu family consensus (Schmid and Shen 1985; present work). The galago Alu consensus sequence is taken from Daniels et al. (1983); the human Alu subfamily sequences are derived as described in the text.

corresponding to an average value of 14.9% sequence difference (Table 1). Pairwise differences among these 72 Alu family members were computed on a VAX computer. A "compare" program counted pairwise differences and compiled them as a matrix; the "makeplot" program plotted bar graphs using these data (Willard, thesis; Fig. 2). Deletions and insertions are counted as a single difference. The PAUP program (Swafford and Maddison 1986) was employed to construct a maximum parsimony tree. Standard deviations of binomial distributions are calculated as the square root of the product NPq, where N is the number of nucleotides in the Alu consensus (282), P is the average divergence, and q (or 1 − P) is the average similarity.

Results

Identification of Three Possible Alu Subfamilies

The distribution of 2556 pairwise differences were computed for the 72 Alu repeats as described in Methods (Fig. 2A). The mean number of pairwise differences between two Alus is 67 ± 13 (Table 1, Fig. 2a). Assuming all Alu repeats were inserted at about the same time as the product of a single master gene, their pairwise divergence should resemble a simple binomial distribution. The width of the distribution of pairwise differences (SD ±13) is broader than expected for a binomial distribution of mutations (±7.2). The corresponding binomial distribution has no resemblance to the observed distribution (Fig. 2A). According to a chi-square analysis ($\chi^2 \sim 500,000$), there is unit probability that a random sample would resemble the predicted binomial distribution more closely than the observed distribution. As another measure of the difference between these distributions, the width of the pairwise difference distribution (SD ±13) is nearly twice the value expected for the corresponding binomial distribution (±7.2). There are several possible explanations for the increased breadth of the pairwise difference distribution. Results described below show that nonrandom sequence differences existing between members of three distinct sequence subfamilies can account for this increase. A parsimony tree analysis of the Alu family members essentially results in three discernible branches. The question addressed by the following analysis is whether these branches represent distinct subfamilies.

Table 1. Divergence of Alu subfamily members

Group of Alus	Divergence from total consensus	Divergence from subgroup consensus	Pairwise differences
All Alus	42 ± 12	—	67 ± 13
Divergent subfamily	61 ± 6	53 ± 6	81 ± 7
Major subfamily	38 ± 6	38 ± 6	61 ± 8
Conserved subfamily	34 ± 4	22 ± 6	38 ± 7

A significant number of pairwise comparisons show a very small number of sequence differences. For example, 41 differences is two standard deviations below the mean of pairwise differences (Fig. 2A). Eleven closely related Alu members having 45 or fewer pairwise differences appear on a common

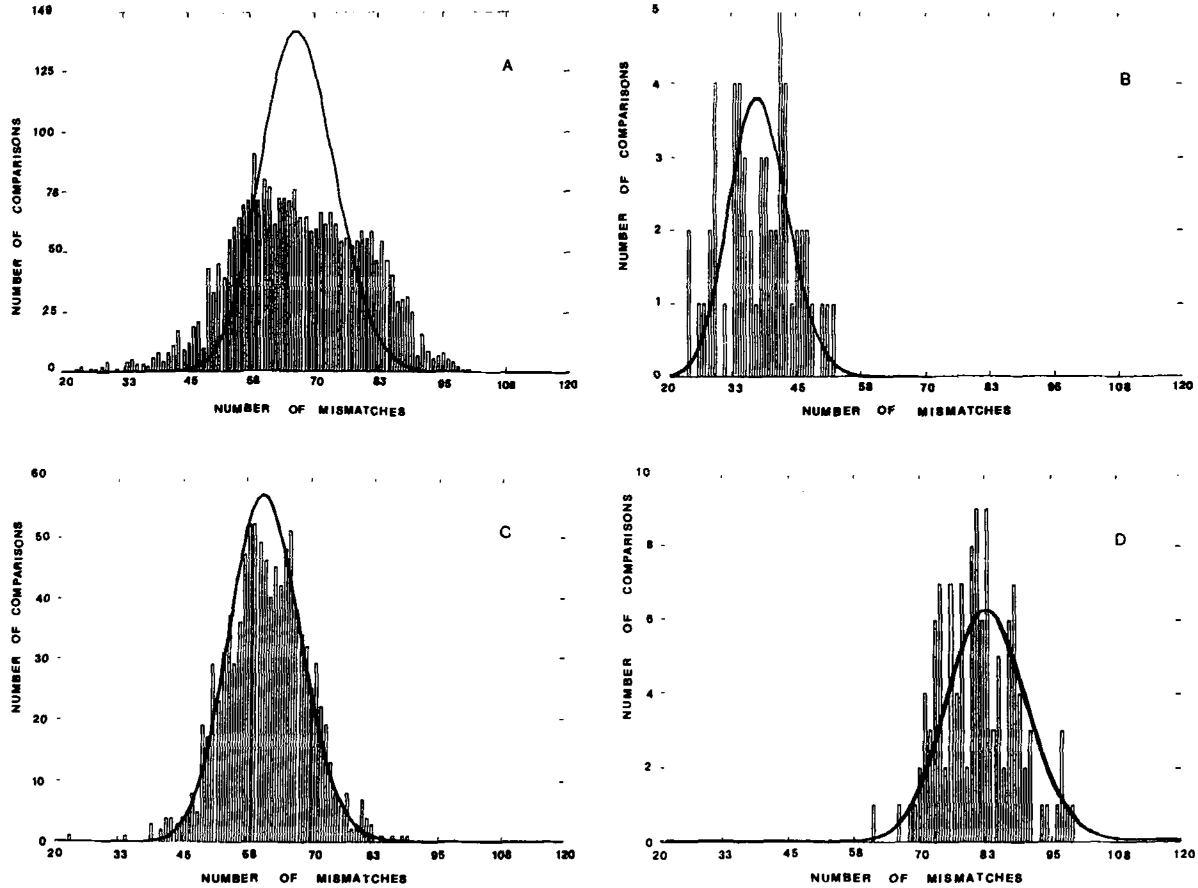

Fig. 2. Pairwise comparisons of Alu family members. **Panel A** has 2556 pairwise comparisons based on 72 Alu sequences. The average divergence is 67.0 ± 13 (SD) differences. **Panel B** has 55 pairwise comparisons based on the 11 Alu repeats assigned to the *conserved subfamily.* The average divergence is 37.8 ± 7 differences. **Panel C** has 990 pairwise comparisons based on the 45 Alu repeats assigned to the *major subfamily.* The average divergence is 61.4 ± 8 differences. **Panel D** has 120 pairwise comparisons based on 16 Alu repeats assigned to the *divergent subfamily.* The average divergence is 81 ± 7 differences. Binomial distributions are superimposed on the pairwise difference distributions. In each case, the binomial distribution is calculated for the number of pairwise comparisons and average number of differences appropriate to that panel with no adjustable parameters.

branch of a parsimony tree (Bains 1986, and Discussion). The 55 pairwise differences among these 11 Alu repeats have an average value of 38 ± 7 pairwise differences (Fig. 2B, Table 1). This value is significantly lower than the 67 ± 13 pairwise differences found for the entire family (Table 1) and, furthermore, the width of this distribution (±7) approximates the value expected for a binomial distribution (±5.7) (Fig. 2B). The chi-square value (2.28) for the fit of these data to a binomial distribution indicates a 30% probability that a random sample would not fit the binomial distribution better than the observed distribution. The chi-square analysis does not reject the hypothesis of a single randomly diverged subfamily. Of course, it is also quite possible that unidentified subgroups comprise this subfamily. We shall in this analysis merely identify the least number of subfamilies required to represent the parent distribution in Fig. 2A.

These 11 closely related Alu repeats define a new subfamily consensus, which has 12 differences compared to the total consensus sequence (Fig. 1). For brevity, we refer to this group of 11 closely related members as the "conserved" subfamily. These 11 conserved subfamily members have an average of 22 ± 6 differences compared to their subfamily consensus sequence, but have 34 ± 4 differences compared to the overall consensus (Table 1). Among other specific differences compared to the overall consensus, the conserved subfamily has two bases deleted at position 65 and an extra A at position 120 (Fig. 1).

Also included within the pairwise distribution of the entire Alu family are members that appear to be especially divergent (Fig. 2A). For example, values of 81 or more pairwise differences are two standard deviations above the mean of the corresponding binomial distribution. Several of these divergent members also appear on a common branch of the parsimony tree and can be used to identify a "divergent" subfamily consisting of 16 members (Bains 1986, and Discussion). There is an average of 81 ±

7 pairwise differences among the 16 members of the divergent subfamily (Fig. 2D, Table 1). Again the width of this distribution (±7) approximates that of a binomial distribution (±7.6) (Fig. 2D). Chi-square ($\chi^2 = 2.6$) indicates a probability of 0.45 that a random sample would not fit the predicted binomial distribution better than the divergent subfamily. As discussed in the preceding example of the conserved subfamily, this probability implies that the members of this group can be regarded as a single divergent subfamily. The 16 members of the divergent subfamily also define a subfamily consensus sequence. Their average divergence from the divergent subfamily consensus (53 ± 6; Table 1) is slightly less than their divergence from the overall consensus (61 ± 6; Table 1). The divergent subfamily consensus sequence has 33 differences compared to the overall consensus (Fig. 1).

The pairwise difference distribution of the remaining 45 Alu family members closely resembles the overall shape of the corresponding binomial distribution, having an average value of 61 differences (Fig. 2C, Table 1). However, the chi-square ($\chi^2 = 34.2$) criterion rejects the goodness of this fit. This relatively large chi-square value results from a small number of sequences that differ markedly from the average. For example, there is a small cluster of comparisons having 80 or more mismatches. These values result from a single Alu family member reported by Miyaki et al. (1983). This particular Alu repeat is unusual in that it is not flanked by short direct repeats, suggesting that it is the result of a recombination between two or more different Alus. Removal of this single Alu from the distribution in Fig. 2C reduces the chi-square value to 20.6. Also contributing substantially to the large chi-square value are several pairs that show relatively little divergence (Fig. 2C). An Alu family member situated 5′ to the $\alpha 2$ globin gene is the source of many of these relatively well matched Alu pairs. This particular Alu family member does not have flanking direct repeats and is known to be a recombination end point in the α globin gene cluster (Hess et al. 1984). Removing this second composite Alu from the pairwise difference distribution reduces chi-square substantially ($\chi^2 = 13.15$). A single Alu pair is observed to have only 21 mismatches. This pair, which is located within the human growth hormone gene cluster, is a sequence duplication (Seeburg 1982), thus accounting for the very small number of mismatches. Removing this single comparison as well as the two previously discussed recombinant Alus reduces chi-square to 5.4, or a 25% probability that a random sample gives no better fit. Further improvements in the chi-square value are of course possible. Removing five additional Alus, which show exceptional similarity or dissimilarity to the major family consensus, improves the chi-square value to over 90% probability that a random sample gives no better fit. However, this improvement is not justified by either the biological identity of these exceptional Alus or the goodness-of-fit criterion, which with the previously noted qualifications accepts the binomial distribution. These findings imply that the members of this subgroup have diverged randomly from either a single consensus founder sequence or, equivalently, several closely related founder sequences. For brevity, this subgroup is called the "major" subfamily.

The consensus sequence of the major subfamily has only three differences at rather variable sites from the overall consensus (Fig. 1). The major subfamily members have an average of 38 ± 6 differences from both their own subfamily consensus and the overall consensus sequences (Table 1).

Significance of the Putative Alu Subfamilies

A heterogeneous distribution often can be divided into component parts. Are these components really subfamilies or merely an artificial mathematical resolution of the original distribution? We can test this issue both statistically and biologically.

The *t*-test for the significance of the difference between the means of the conserved and major subfamilies has a numerical value of 21.4. This value corresponds to a probability of less than 0.0005 of finding the observed separation between the conserved and major subfamilies by chance. The *t*-test criterion indicates even lower probabilities of observing the separations between the major and diverged subfamilies and the diverged and conserved subfamilies by chance. According to these results, there is near certainty that our analysis has resolved nonoverlapping distributions. The question remains whether these subfamily divisions are meaningful.

Assuming that the three subgroups described above are distinct subfamilies, their different consensus sequences should resemble those of their respective founder sequences. The conserved subfamily would include the most recently inserted Alu family members. The primordial Alu sequence should be more closely related to the divergent subfamily consensus and more distantly related to the conserved subfamily consensus. Results from the following sequence comparisons confirm these predictions.

Included among the 11 members of the conserved subfamily identified here are the four Alu family members that Slagel et al. (1986) identified as belonging to a distinct subfamily and an Alu repeat associated with a polymorphism in the Mliv-2 locus (Economou-Pachis and Tsichlis 1985). The Mliv-2

Alu repeat is the only polymorphic member of the family identified to date. It has only 15 sequence differences compared to the conserved subfamily consensus sequence, far less than the average divergence (42 ± 12 differences) of Alu repeats from the total consensus. The inclusion of this polymorphic Alu repeat in the conserved subfamily is direct evidence that this subfamily includes more recently inserted Alu family members.

There are at least two different families of short interspersed repeats in the prosimian primate galago, one of which resembles the human Alu family consensus (Daniels and Deininger 1983; Daniels et al. 1983). Specific sequence differences between the members of the human and equivalent galago Alu families show that the Alu families in these two species are the insertion products of different founder sequences (Daniels et al. 1983). We therefore assume that the human and galago Alu founder sequences diverged from a mutual primordial ancestral Alu sequence. There is a modest but clear trend in homology between the galago and three human subfamily consensus sequences (Fig. 1, Table 2). The conserved subfamily is more distantly related to the galago consensus (24.0%) than is the divergent subfamily (19.5%); the major subfamily is intermediate in its divergence. These differences are summarized on a tree comparing the galago and human Alu subfamilies (Fig. 3).

Although the conserved subfamily consensus shows the most sequence divergence from what should be the ancestral human Alu founder sequence, its individual members show the least divergence from their presumed subfamily founder sequence (Fig. 3). Conversely, the divergent subfamily consensus sequence is apparently closest to what should be the ancestral founder, but its individual members show the greatest sequence divergence relative to their presumed subfamily founder (Fig. 3). The major subfamily and its individual members are intermediate in both regards.

Table 2. Percent divergence of consensus sequences

	Diverged subfamily	Major subfamily	Conserved subfamily
Galago	19.5	21.6	24.0
Conserved subfamily	13.7	5.2	—
Major subfamily	12.1	—	—

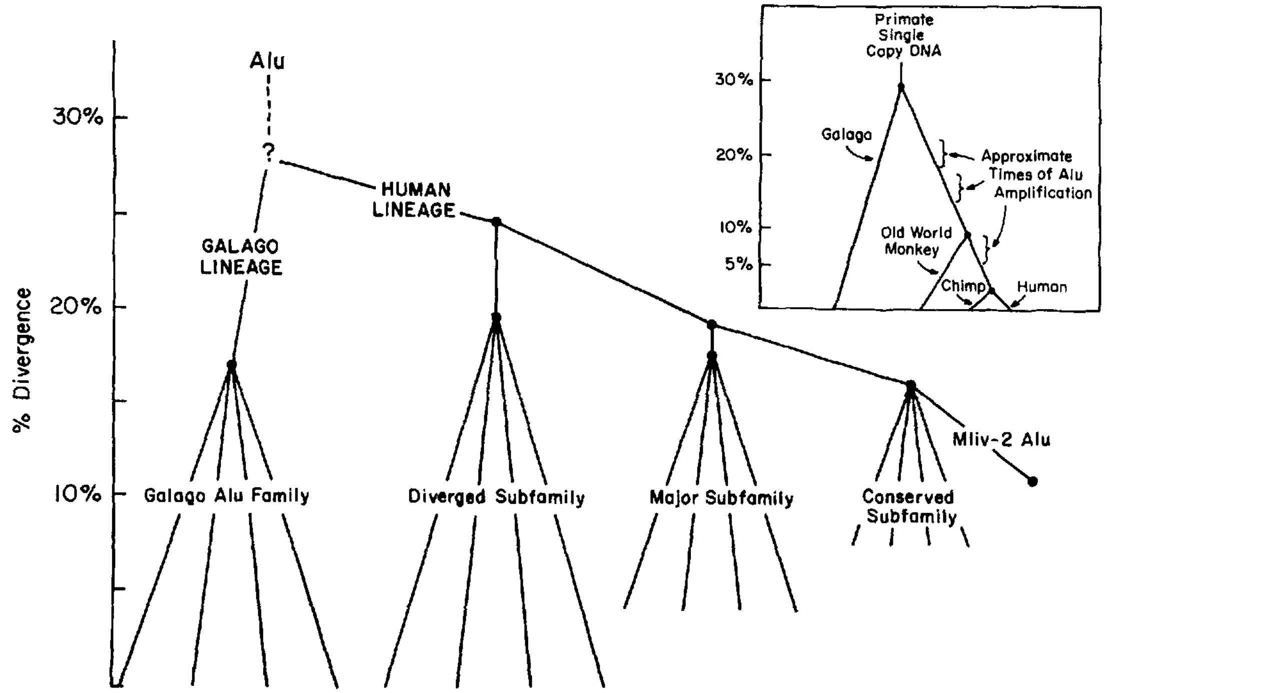

Fig. 3. Schematic representation to scale of the sequence divergence of Alu subfamilies and members. The divergence of the various consensus sequences is taken from the data in Fig. 1 and Table 2. The divergence of individual members from their respective consensus sequences is from the data in Table 1 and in Daniels et al. (1983) for galago. The location of the galago–human Alu branch point is arbitrarily fixed at a midpoint, which assumes equal divergence for galago and the diverged subfamily. For comparison, the insert shows the approximate divergence of single-copy sequences within the primate lineage (Deininger and Schmid 1979; Sawada et al. 1985; and references therein).

The individual members of the conserved subfamily have 8% less sequence divergence from the presumed primate founder sequence than do the individual members of the divergent subfamily (Fig. 3). One possible explanation for this difference is that members of the divergent subfamily have evolved without selection for a longer period of time than have the members of the conserved subfamily. Selection could result from either a required biological function of a master sequence or merely the requirements needed to encode new members of the family. The Alu repeat associated with the polymorphism at the Mliv-2 locus shows relatively less divergence from both its subfamily consensus (5%) and the presumed ancestral human Alu (13%) (Fig. 3). Again, this is consistent with the interpretation that recently mobile members of the family have evolved under selection for a longer time than other family members.

Discussion

Comparing orthologous loci in the alpha globin gene cluster, Sawada and co-workers find that human Alu family repeats were inserted following the divergence of the human and galago lineages, but preceding the divergence of human and chimpanzee, and, at least in some cases, prior to the divergence of the human and Old World monkey lineages (Sawada et al. 1985; Sawada and Schmid 1986). Their conclusions agree with findings from indirect but global sequence comparisons. Randomly selected Alu family members from human and New World monkeys share a common consensus sequence, implying that they originated from a common founder sequence(s) (Daniels et al. 1983). As previously mentioned, the human and equivalent galago Alu families have recognizably distinct consensus sequences and are certainly the product of different founder sequences. The average divergence of the flanking short direct repeats surrounding human Alu family members is also consistent with Alus having an average insertion time following the divergence of the prosimian lineage but preceding the divergence of the human and monkey lineages (Schmid and Shen 1985). Two additional conclusions of Sawada and co-workers are implicit assumptions in the present analysis: once inserted, Alu repeats are not converted or corrected to some other master sequence, but rather diverge at the rate expected for nonselected sequences (Sawada et al. 1985; Sawada 1986; Sawada and Schmid 1986).

Compared to the divergence of single-copy DNA, the pairwise divergence of the members of the three Alu subfamilies is qualitatively consistent with the insertion of most human Alu repeats at a time following the divergence of the galago lineage but preceding the divergence of the Old World monkey lineage (Fig. 3). Although each subfamily is represented as resulting from a single episodic burst, the detailed history of the formation of any of the three subfamilies could be more complex than this simplified representation. As an example, each subfamily could result from an accumulation of numerous founder sequences. However, the pairwise divergence of each subfamily is adequately represented by a binomial distribution (Fig. 2B–D). For this reason, we are unable to detect any additional subgroupings within the three subfamilies and are unable to resolve any additional details concerning their formation.

The present results confirm and extend Slagel et al.'s (1986) report of a closely related Alu subfamily. These conclusions contrast with Bains's (1986) finding that all Alu repeats are about equally divergent from the center of a dendrogram. The issue is whether all Alu repeats are competent to code for new family members, as implied by Bains's conclusion, or whether a select sequence(s) codes for new members, as implied by Slagel et al.'s conclusion. We believe the existence of at least three Alu subfamilies is demonstrated by the previously described results: (1) the nearly binomial distribution of pairwise differences in each subfamily compared to that of the total family; (2) the inclusion of a polymorphic Alu repeat within the conserved subfamily; and (3) the relative divergence of the three human subfamilies from that of galago. Bains's dendrogram analysis correctly segregates branches containing the conserved and diverged subfamilies from each other as well as from most branches of the major subfamily. [Specifically, using Bains's notation, sequence numbers 11, 32, 41, 42, 46, 50, 54, and 55 (designating members of the conserved subfamily) are concentrated on a single branch, whereas sequence numbers 26, 37, 43, 44, and 51 (designating members of the diverged subfamily) are located primarily on another single branch.] Unfortunately, this segregation is difficult to discern by the dendrogram analysis, which is overwhelmed by the random sequence divergence of the major subfamily and further obscured by the difficulty in identifying the central point of the dendrogram. The excellent agreement of the pairwise difference distribution of the major subfamily with a binomial distribution (Fig. 2C) confirms Bains's finding that the majority of Alu repeats does not belong to discernible subfamilies. For these reasons, it is not surprising that the three subfamilies described above were not detected by the dendrogram analysis.

It is widely believed that Alu family members are dispersed by way of an RNA intermediate (Schmid and Shen 1985; Weiner et al. 1986). However, Alu family members, although present as read-through transcripts in heterogeneous nuclear RNA, do not generally promote transcription in vivo (Weiner 1980; Schmid and Jelinek 1982; Paulson and Schmid 1986). The finding of distinct Alu subfamilies suggests that only certain Alu family members have been competent to code for new family members. The present results also suggest that the competent Alu repeats have been subject to selection at the level of their sequences.

Acknowledgments. This research was supported by USPHS grant GM 21346. We thank D. Swafford for providing the PAUP program.

References

Bains W (1986) The multiple origins of human ALU sequence. J Mol Evol 23:189–199

Daniels GR, Deininger PL (1983) A second major class of Alu family repeated DNA sequences in a primate genome. Nucleic Acids Res 11:7595–7610

Daniels GR, Fox M, Lowensteiner D, Schmid CW, Deininger PL (1983) Species-specific homogeneity of the primate Alu family of repeated DNA sequences. Nucleic Acids Res 11: 7579–7593

Deininger PL, Schmid CW (1979) A study of the evolution of repeated DNA sequences in primates and the existence of a new class of repetitive sequences in primates. J Mol Biol 127: 437–460

Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW (1981) Base sequence studies of 300 nucleotide renatured repeated human DNA clones. J Mol Biol 151:17–33

Economou-Pachis A, Tsichlis PN (1985) Insertion of an Alu SINE in the human homologue of the Mlvi-2 locus. Nucleic Acids Res 13:8379–8387

Hess JF, Schmid CW, Shen C-KJ (1984) A gradient of sequence divergence in the human adult alpha-globin duplication units. Science 226:67–70

Miyaki T, Migita K, Sakaki Y (1983) Some KpnI family members are associated with the Alu family in the human genome. Nucleic Acids Res 11:6837–6846

Paulson KE, Schmid CW (1986) Transcriptional inactivity of Alu repeats in HeLa cells. Nucleic Acids Res 14:6145–6158

Sawada I (1986) Evolution of the primate alpha globin gene cluster and its interspersed Alu family repeats. Thesis, University of California, Davis

Sawada I, Schmid CW (1986) Primate evolution of the alpha-globin gene cluster and its Alu-like repeats. J Mol Biol 192: 693–709

Sawada I, Willard C, Shen C-KJ, Chapman B, Wilson AC, Schmid CW (1985) Evolution of Alu family repeats since the divergence of human and chimpanzee. J Mol Evol 22:316–322

Schmid CW, Jelinek WR (1982) The Alu family of dispersed repetitive sequences. Science 216:1065–1070

Schmid CW, Shen C-KJ (1985) The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, pp 323–358

Seeburg PH (1982) The human growth hormone gene family: nucleotide sequences show recent divergence and predict a new polypeptide hormone. DNA 1:239–249

Slagel V, Flemington E, Traina-Dorge V, Bradshaw H, Deininger P (1986) Clustering and subfamily relationships of the Alu family in the human genome. Mol Biol Evol 4:19–29

Swafford DL, Maddison WP (1986) Math Biosci (submitted)

Weiner AM (1980) An abundant cytoplasmic 7S RNA is complementary to the dominant interspersed middle repetitive DNA sequence family in the human genome. Cell 22:207–218

Weiner AM, Deininger PL, Efstradiatis A (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Received December 22, 1986/Revised and accepted March 25, 1987