Nucleic Acids Research, Vol. 20, No. 13 3397–3401

Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements

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

Received March 26, 1992; Revised and Accepted May 28, 1992

ABSTRACT

The Alu dimeric elements are a common feature of the primate genomes, where they constitute a family of related sequences (1). The identification of a free left Alu monomer (FLAM) family plus a free right Alu monomer (FRAM) family suggests that the dimeric structure results from the fusion of a FLAM sequence with a FRAM sequence (2). Here, we describe a very old Alu-like monomeric family, referred to as FAM for fossil Alu monomer. This family arose from a 7SL RNA sequence and gave birth to the FLAM and FRAM families. From the results obtained, the evolution of the Alu family can be subdivided into two phases. The first phase, which involves only monomeric elements, is characterized by deep remodelling of the progenitor sequences and ends with the appearance of the first Alu dimeric element through the fusion of a FLAM and a FRAM element. The second phase, still in progress, starts with the first Alu dimeric element. This phase is characterized by the stabilization of the progenitor sequences.

INTRODUCTION

The most extensively studied family of retroposons is the primate Alu family. A typical member of the Alu family is composed of two related sequences (monomers) tandemly arranged (3). An adenine rich (dA-rich) region is found at the end of each monomer, and the entire element is about 300 bp long. In genomic sequences, it is flanked by short direct repeats that correspond to the duplication of the insertion site. Each monomer is partially homologous with the 7SL RNA (4,5), which is a component of the signal recognition particle implicated in protein secretion (6).

Some Alu elements are transcribed by the RNA polymerase III, and it has been proposed that the RNA obtained can be used as a template by a reverse transcriptase to generate a DNA copy that is subsequently integrated at a new locus in the genome (7–9). This process, called retroposition (9,10), would be responsible for the amplification of more than 500,000 Alu elements per human haploid genome (1). However, only a very small set of sequences called source genes (11), master genes (12), or progenitor sequences (13) are used as template. The replacement of the progenitor elements during evolutionary time produced different subfamilies of Alu elements, which can be characterized by a specific set of mutations (11, 14–19). The identification of a human-specific subfamily indicates that Alu progenitor sequences were still active after the emergence of the human lineage (12, 18, 20–24).

Recently, we described a family of free left Alu monomers (FLAMs, or FLAs in Ref. 25), composed of two subfamilies (Al and Cl) and a small family of free right Alu monomers (FRAMs, or FRAs in Ref. 25) (2). The phylogenetic analysis of these new families suggested that the first progenitor of the Alu dimeric family arose through the fusion of a free Alu monomer (FLAM-Cl) with a free right Alu monomer (FRAM). The older subfamily of FLAM (Al) and the family of FRAM have a common ancestor sequence that derived from the 7SL RNA genes (2). A monomeric family has been also described in the prosimian, *Galago crassicaudatus* (26). However, elements of this family are not related to the Alu sequences. They are assumed to have derived from a methionine transfer RNA gene (26).

In our previous study, we found two free Alu monomers (HUMINF3 and HUMGPP3A04) that were very close to the ancestor of the FLAM and FRAM (2), but they do not have the 11 bp deletion characteristic of the right monomers. Those features suggest that both sequences belong to an ancestral family of Alu-like elements that descended from a 7SL RNA sequence but predated the FLAM and FRAM families (2). In the present study we confirm this hypothesis by the identification of seven other sequences sharing the same features.

MATERIALS AND METHODS

We screened the updated versions of GenBank (27) and EMBL (28) for other members of the ancestral family (see the introduction). We used the facility provided by the GenBank server through the networks. We submitted the following query to the FASTA program (29): 5′ 'CTATGGATCGCGCC TGTGAATAGCCACTGCAC' 3′. This sequence corresponds to the 11 bp deletion (underlined) with its flanking regions in the right Alu monomers (Figure 1).

The phylogenetic tree has been reconstructed using the maximum likelihood method [DNAML program of the PHYLIP package (30)].

RESULTS AND DISCUSSION

Once the 7SL RNA genes and pseudogenes were discarded, the screening of the updated versions of GenBank (27) and EMBL (28) revealed the presence of seven new candidates in the primate subdivision, characterized by the absence of the 11 bp deletion between positions 247 and 259 (Figure 1; throughout the text and the figures, the numbering refers to the 7SL RNA sequence).

The sequence alignments of the nine elements have been edited, comparing them to the 7SL RNA sequences (31) and the progenitor sequence of the FLAM and FRAM families (2) (Figure 1). The sequences HUMALPI and HUMGPP3A04 present a deletion of 77 bp and 17 bp, respectively, in the 5′ end of the Alu element; in databases the sequence M27852 and HUMFURIN only start with the second half of the Alu element. The other seven sequences correspond to full length monomers that are similar to the right monomer of a typical Alu element. All sequences have a dA-rich region in the 3′ end, and six of them are flanked by short direct repeats (Figure 1). These features suggest that they correspond to the insertion of monomeric elements. The elements HUMRSAIFN and HUMIFNIN3 present a high homology (87.7%), and they are flanked by similar sequences (Figure 1). Thus, they probably correspond to a single Alu insertion followed by a gene duplication. In this paper, we consider only one of them (HUMIFNIN3).

In addition to the absence of the 11 bp deletion, the selected elements do not have the diagnostic bases of the FLAM (A in position 83) and FRAM (G in position 278 and G in position 282) families (Figure 1). Those elements can be characterized by a progenitor sequence that corresponds to the consensus sequence, although sites that have high proportions of CpA, CpG, and TpG dinucleotides in the sequences are considered as CpG in the progenitor sequence. Several positions, signaled by a question mark in Figure 1, are ambiguous in the progenitor sequence. All of them are involved in CpG dinucleotides, either in the 7SL RNA or in the other progenitor sequences.

The 7SL RNA genes, and the FAM, FLAM, and FRAM progenitor sequences, have been used to reconstruct a phylogenetic tree of the early stages of the evolution of the Alu-like families (Figure 2). Only the sites (1 to 83 and 267 to 299)

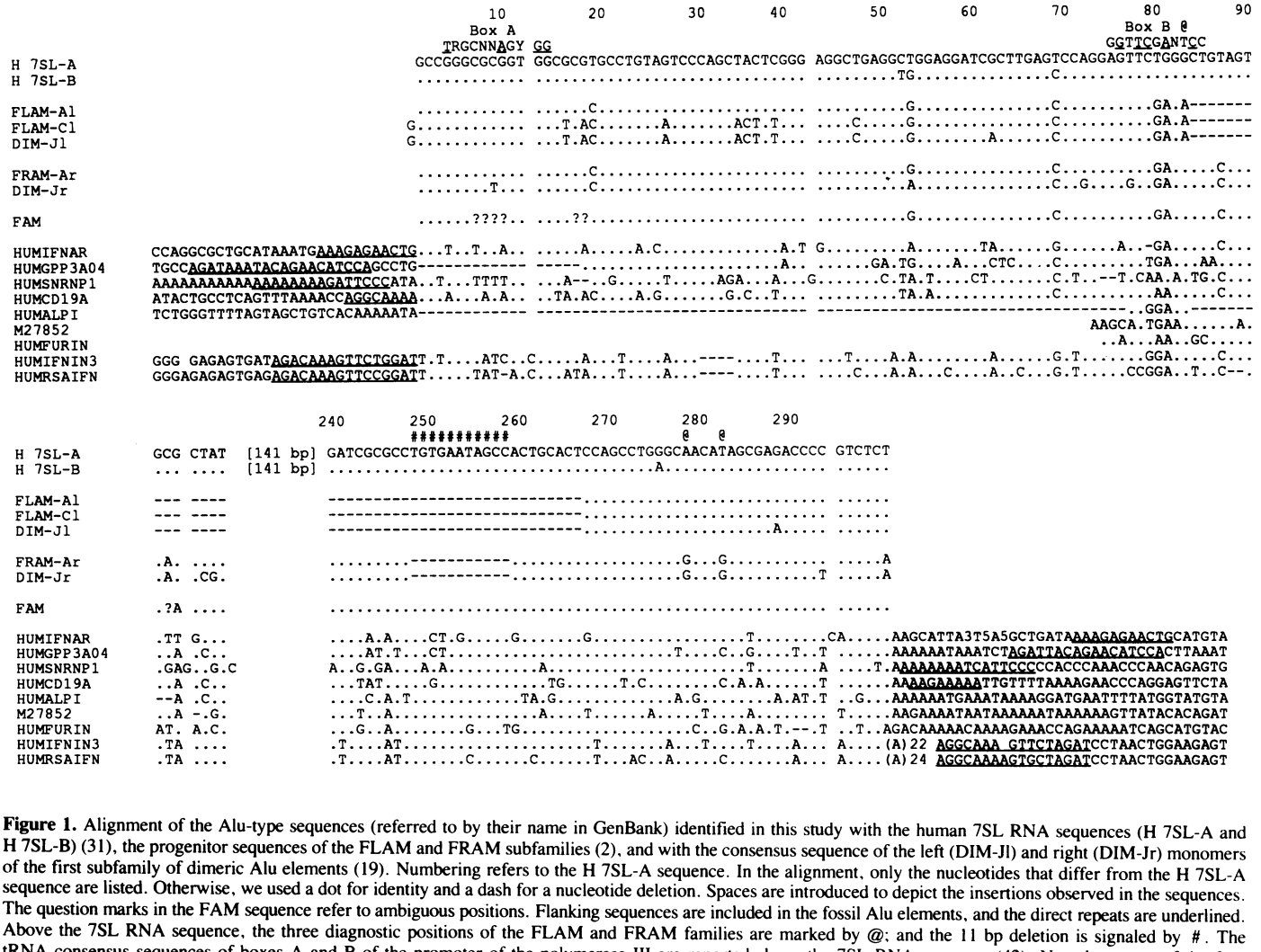

Figure 1. Alignment of the Alu-type sequences (referred to by their name in GenBank) identified in this study with the human 7SL RNA sequences (H 7SL-A and H 7SL-B) (31), the progenitor sequences of the FLAM and FRAM subfamilies (2), and with the consensus sequence of the left (DIM-Jl) and right (DIM-Jr) monomers of the first subfamily of dimeric Alu elements (19). Numbering refers to the H 7SL-A sequence. In the alignment, only the nucleotides that differ from the H 7SL-A sequence are listed. Otherwise, we used a dot for identity and a dash for a nucleotide deletion. Spaces are introduced to depict the insertions observed in the sequences. The question marks in the FAM sequence refer to ambiguous positions. Flanking sequences are included in the fossil Alu elements, and the direct repeats are underlined. Above the 7SL RNA sequence, the three diagnostic positions of the FLAM and FRAM families are marked by @; and the 11 bp deletion is signaled by #. The tRNA consensus sequences of boxes A and B of the promoter of the polymerase III are reported above the 7SL RNA sequence (42). N replaces any of the four bases, and the underlined letters correspond to the invariant bases.

that were not involved in the large deletions were taken into account. On the topology obtained, the progenitor sequence of the new family is localized between the sequences of the 7SL RNA genes and the progenitor sequences of the FLAM and FRAM families (all branches are significantly positive at the $p < 0.05$ level). Therefore, this tree strongly suggests that the Alu monomers selected could represent the fossils of an old Alu-like family that predated the divergence between the left and the right Alu monomers. We shall refer to these elements as fossil Alu monomers (FAMs).

On the other hand, the FAM, FLAM, and FRAM hypothetical progenitor sequences also differ by large deletions, which were not taken into account in the phylogenetic tree (Figure 1). If we take the 7SL RNA sequence as reference, the FAM and FRAM have the same 141 bp deletion (between positions 97 and 239), but the FRAM has an additional deletion of 11 bp (between positions 247, 259). Thus, the first FRAM would have been produced either by a single 11 bp deletion from the FAM sequence, or by two deletions (141 bp and 11bp) from the 7SL RNA. However, the second hypothesis implies that the same deletion at the same location occurred independendly in the FAM and FRAM sequences. It is quite unlikely that it happened by chance. Thus, this observation strongly supports the idea that the FAM preceded the FRAM. The FLAM has a deletion of 183 bp (between positions 83 and 267) that include the 141bp and 11bp deletions. Therefore, the FLAM can result from a single deletion from any of the other sequences: a 183 bp deletion from the 7SL RNA, a 42 bp deletion from the FAM, and a 31 bp deletion from the FRAM. Nevertheless, the 3 substitutions observed between the FLAM and the FRAM progenitor sequences do not support the last hypothesis. The average pairwise similarities among the members of the FAM family (68%) is smaller than the one observed with the elements of the FLAM (74%) and FRAM (71%) families (2). Therefore, all together, the phylogenetic tree presented in Figure 2, the 141 bp deletion, and the average pairwise similarities strongly support that the FAM family predated the FLAM and FRAM families.

The three large deletions are localized in the internal part of the 7SL RNA. In the probable secondary structure of the 7SL RNA, this region corresponds to the end of a stable helix (helix 5 in Ref. 32) that held together the Alu parts of the 7SL RNA. It has been proposed that such deletion may have occurred at the RNA level through nuclease attacks (5, 33). The secondary structure of the FAM hypothetical progenitor sequence preserves the helix 5 (data not shown) and brings closer the end points of the 42 bp deletion. Thus, this deletion may also have occurred at the RNA level. The 11 bp deletion does not correspond to a specific feature of the secondary structure of the FAM progenitor sequence (data not shown), but it is flanked by short perfect repeats of three base pairs (GCC), and it can be the result of a homologous recombination between those repeats at the DNA level.

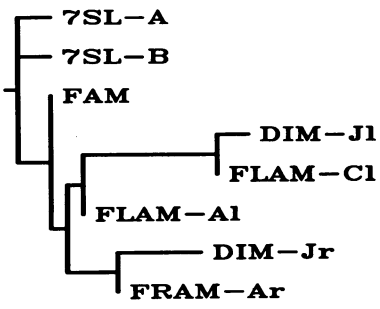

Figure 2. Phylogenetic tree of the origin of the Alu family. Positions 1 to 83 and 267 to 299 in the alignment presented in Figure 1 were used to carry out a tree reconstruction with the maximum likelihood method (30). The lengths of the horizontal lines correspond to the relative divergences from a common ancestor. All branches are significantly positive at the $p < 0.001$ level, except the two smaller ones that are significantly positive with a lower confidence ($p < 0.05$).

From those observations, the evolution of the Alu family can be subdivided into two phases. The first phase, which involves only monomeric elements, is characterized by deep remodelling of the sequences and ends with the appearance of the first Alu dimeric element through the fusion of a FLAM and a FRAM element. The second phase starts with the first Alu dimeric element and is characterized by the stabilization of the progenitor sequences. Indeed, up to the present time the progenitor sequences of the Alu dimeric elements have evolved only through base substitutions and small insertions/deletions (11, 14–19). It has been proposed that large alterations might have been required in the first step of the evolution to abolish the competition between the parent gene and the amplified elements (26). However, the central deletions in the FAMs, FLAMs, and FRAMs conserved the first half of the secondary structure of the 7SL RNA, which has been also maintained by the successive Alu dimeric progenitor sequences (34, 35). This domain, in association with two proteins, confers the elongation-arresting activity to the 7SL RNP particle (36). Thus, from the beginning, the Alu progenitor sequences could have retained the capacity to interact with cellular components, suggesting that they are functionally important for the host genome (11, 19). On the other hand, this RNA secondary structure could have some affinity for reverse transcriptases or other components of the retroposition machinery (37, 38), and its conservation in the monomeric and Alu dimeric sequences could be related to their mobility. Indeed, this structure is first found in the 7SL RNA sequences that are prone to retroposition (39), and it is also retained by the progenitor sequences of the B1 family in the rodent genomes (34, 40). Nevertheless, both hypotheses (secondary structure involved in a cellular function or in the reverse transcription) are not mutually exclusive.

In addition to a stable RNA secondary structure, other structural features may have been decisive for the emergence of the first family of Alu-like elements. For example, the acquisition of a dA-rich sequence near the 3′ end of the element can allow the self-priming of the polymerase III transcripts during reverse transcription (7,8). Another feature is the acquisition of an efficient promoter for the RNA polymerase III. The general organization of the polymerase III promoter of the Alu elements is similar, although not identical, to the split promoter of the tRNA genes (41). The first component is located between positions 3 and 36 in Figure 1 and contains box A of the tRNA gene promoters; it increases the transcriptional efficiency. The second component is located between positions 69 and 81 and overlaps with box B of the tRNA promoter. This element alone is sufficient and necessary for an accurate initiation of the transcription (41). The consensus sequence for box B of the tRNA gene promoter is given as GGTTCGANTCC, where N replaces any of the four bases and the underlined letters correspond to the invariant bases (42). In the 7SL RNA sequence, this signal is weak (A.....TG.G.T, where a dot means the same base as in box B). The FAM progenitor sequence suffered only two substitutions in regard to its ancestral 7SL RNA sequences (G80

and A81), but both of them increased the similarity to box B consensus sequence (A..G.T). Thus, throughout those mutations, the first Alu monomers could have evolved a new promoter and escaped the transcriptional regulation of the 7SL RNA genes. That could have been a crucial step in the evolution of the FAM sequences.

Since the RNA polymerase III promoter is part of the transcript, each new element would be capable of transcription and retroposition. However, the presence of the internal promoter is not sufficient to ensure an efficient amplification of the element. There is evidence that the flanking sequences may play a role in the RNA polymerase III initiation (43). Moreover, to be conserved by evolution, the transcription and retroposition should occur in the germline. Therefore, the success of progenitor sequences should also depend on their chromosomal location.

The results presented here suggest that the first stages of the evolution of the Alu family are characterized by successive replacements of Alu-like progenitor sequences. The FAM family arose from a 7SL RNA sequence; this family gave birth to the FRAM and FLAM families, and then the fusion of a FLAM sequence with a FRAM sequence produced the first Alu dimeric element. This is a replacement since the progenitor sequences of a family would no longer be active after the emergence of a new family (2). A turnover can be easily explained if the successive progenitors sequentially derived from one to the other (11, 24). However, here we have a situation where at least two different progenitor sequences were active at the same time: the FLAM and FRAM progenitor sequences (see Ref. 19 for other examples). Therefore, the replacement observed can also be the result of a competition between different progenitor sequences for the same retroposition machinery. One possibility is the improvement of the RNA polymerase III promoter (see Ref. 44 for such an example in the Galago genome). For instance, the success of the FLAM family, at least 3 times larger than the FRAM family (2), can be related to the large deletion that replaces a T by a C at the end of box B (this modification came with a substitution in position 83 that did not modify the similarity with the consensus signal of box B). Some other features, such as the stem-loop extensively discussed by Jurka and Zuckerkandl (25), might be also related to the success of progenitor sequences.

The FAM sequences have been found only in the primate section of GenBank and EMBL. However, we cannot exclude the fact that they predated the common ancestor of rodents and primates since the rodents B1 elements are close to the left Alu monomer (1). The FAMs are very rare in the human genome (8 sequences in the 13,658,513 bases of the primate section of GenBank release 71). Their absence in the other sections of the databases does not imply that they have no equivalent in other mammalian genomes. The accumulation of sequences in databases and the direct analysis of complete genomes, with a probe specific to the FAM sequences, should help to clarify the origin of the Alu-type elements in the mammalian genomes.

CONCLUSION

We have reported the analysis of Alu-like monomers that represent the fossils of an old family: the FAM family. Three independent observations (a phylogenetic tree, the analysis of large deletions, and average pairwise similarities) strongly suggest that (i) the FAM family arose from a 7SL RNA sequences through a 141 bp deletion, (ii) the first FLAM resulted from a 42 bp deletion of a FAM sequence, and (iii) the first FRAM was created by an 11 bp deletion in a FAM sequence. Therefore, the first step in the evolution of the Alu family involves only monomeric sequences and is characterized by deep remodelling of the sequences. However, those deletions preserved a structural domain of the 7SL RNA, suggesting that an RNA secondary structure is in some way involved in the success of the Alu sequences. The FAM sequences suffered few substitutions from the 7SL RNA sequence, but two of them are located in box B of the polymerase III promoter and enhance its similarity to the consensus sequence for box B of the tRNA genes. Thus, through those mutations, the first Alu monomers could have evolved a new promoter and escaped the transcriptional regulation of the 7SL RNA genes. The first phase of the evolution of the Alu family ends with the fusion of a FLAM sequence and a FRAM sequence; that produced the first Alu dimeric element. This element opened a second phase, characterized by dimeric progenitor sequences. If the dimeric elements are typical of primate genomes, we cannot exclude the fact that the FAM sequences predated the emergence of the primate lineage.

ACKNOWLEDGEMENTS

I thank G.Fichant, D.Torney, and C.Burks for critical readings of the manuscript, and two anonymous reviewers whose comments helped to clarify the text. I also appreciate careful proofreading by P.Reitemeier. This work was funded by NIH grant GM-37812 and DOE grant F118.

REFERENCES

1. Schmid, C. W. and Jelinek, W. R. (1982) *Science* **216**, 1065–1070
2. Quentin, Y. (1992) *Nucleic Acids Res.* **20**, 487–493
3. Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. and Schmid, C. W. (1981) *J. Mol. Biol.* **151**, 17–33
4. Weiner, A. M. (1980) *Cell* **22**, 209–218
5. Ullu, E. and Tschudi, C. (1984) *Nature (London)* **312**, 171–172
6. Walter, P. and Blobel, G. (1982) *Nature (London)* **299**, 691- 698
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. Weiner, A. M., Deininger, P. L. and Efstratiadis, A. (1986) *Ann. Rev. Biochem.* **55**, 631–661
10. Rogers, J. (1983) *Nature (London)* **301**, 460
11. Britten, R. J., Baron, W. F., Stout, D. B. and Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770–4774
12. Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Desselle, T. D., Hoppens, C. L. and Deininger, P. L. (1990) *Nucleic Acids Res.* **18**, 6793–6798
13. Deininger, P. L. and Daniels, G. R. (1986) *Trends Genet.* **2**, 76–80
14. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. and Deininger, P. L. (1987) *Mol. Biol. Evol.* **4**, 19–29
15. Willard, C., Nguyen, H. T. and C. W. Schmid, (1987) *J. Mol. Evol.* **26**, 180–186
16. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4775–4778
17. Quentin, Y. (1988) *J. Mol. Evol.* **27**, 194–202
18. Deininger, P. L. and Slagel, V. K. (1988) *Mol. Cel. Biol.* **8**, 4566–4569
19. Jurka, J. and Milosavljevic, A. (1991) *J. Mol. Evol.* **32**, 105–121
20. Matera, A. G., Hellmann, U., Hintz, M. F. and Schmid, C. W. (1990) *Nucleic Acids Res.* **18**, 6019–6023
21. Matera, G. A., Hellmann, U. and Schmid, C. W. (1990) *Mol. Cell. Biol.* **10**, 5424–5432
22. Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J. and Deininger, P. L. (1991) *Nucleic Acids Res.* **13**, 3619–3623
23. Batzer, M. A. and Deininger, P. L. (1991) *Genomics* **9**, 481–487
24. Shen, M. R., Batzer, M. A. and Deininger, P. L. (1991) *J. Mol. Evol.* **33**, 311–320
25. Jurka, J. and Zuckerkandl, E. (1991) *J. Mol. Evol.* **33**, 49–56
26. Daniels, G. R. and Deininger, P. L. (1985) *Nature (London)* **317**, 819–822

27. 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.
28. Hamm, G. H. and Cameron, G. N. (1986) *Nucleic Acids Res.* **14**, 5–9
29. Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448
30. Felsenstein, J. (1989) *Cladistics* **5**, 164–166
31. Reddy, R. (1988) *Nucleic Acids Res.* **16**, r71-r85
32. Larsen, N. and Zwieb, C. (1991) *Nucleic Acids Res.* **19**, 209–215
33. Gundelfinger, E. D., Krause, E., Melli, M. and Dobberstein, B. (1983) *Nucleic Acids Res.* **11**, 7363–7374
34. Quentin, Y. (1989) Thesis, University of Lyon, France (in french)
35. Sinnett, D., Richer, C., Deragon, J-M. and Labuda, D. (1991) *J. Biol. Chem.* **266**, 8675–8678
36. Siegel, V. and Walter, P. (1986) *Nature (London)* **320**, 81–84
37. Jurka, J. (1989) *J. Mol. Evol.* **29**, 496–503
38. Okada, N. (1990) *J. Mol. Evol.* **31**, 500–510
39. Ullu, E. and Weiner, A. M. (1984) *EMBO J.* **3**, 3303–3310
40. Labuda, D., Sinnett, D., Richer, C., Deragon, J-M. and Striker, G. (1991) *J. Mol. Evol.* **32**, 405–414
41. Perez-Stable, C., Ayres, T. M., and Shen, C. -K. J. (1984) *Proc. Natl. Acad. Sci. USA* 81, 5291–5295
42. Geiduschek, E. P. and Tocchini-Valentini, G. P. (1988) *Ann. Rev. Biochem.* **57**, 873–914
43. Ullu, E. and Weiner, A. M. (1985) *Nature (London)* **318**, 371–374
44. Daniels, G. R. and Deininger, P. L. (1991) *Nucleic Acids Res.* **19**, 1649–1656