170–175 Nucleic Acids Research, 1995, Vol. 23, No. 1

Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era

Jerzy Jurka, Ewa Zietkiewicz[1] and Damian Labuda[1,*]

Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306, USA and [1]Centre de Recherche, Hôpital Ste-Justine, Département de Pédiatrie, Université de Montréal, Montréal, Québec H3T1C5, Canada

Received August 24, 1994; Revised and Accepted November 25, 1994

ABSTRACT

Short interspersed elements (SINEs) are ubiquitous in mammalian genomes. Remarkable variety of these repeats among placental orders indicates that most of them amplified in each lineage independently, following mammalian radiation. Here, we present an ancient family of repeats, whose sequence divergence and common occurrence among placental mammals, marsupials and monotremes indicate their amplification during the Mesozoic era. They are called MIRs for abundant Mammalian-wide Interspersed Repeats. With approximately 120,000 copies still detectable in the human genome (0.2–0.3% DNA), MIRs represent a 'fossilized' record of a major genetic event preceding the radiation of placental orders.

INTRODUCTION

Rapid anatomical and behavioral evolution of mammals has long been associated with emergence of new patterns of gene regulation, promoted by an increase in genomic variability [1]. Studies documenting direct implication of interspersed repetitive elements in DNA rearrangements [2–9] suggest that promiscuous proliferation of such repeats could have been a major force in the evolution of mammals. Indeed, dispersal of distinct, lineage-specific families of SINEs accompanied the evolution of mammalian orders [10–12]. In contrast, repetitive elements involved in the radiation of mammals should have amplified earlier and be therefore common to different lineages. We found that a 70-nucleotide repeat, which was first reported in the human prothrombin gene [13] and then described in a variety of genomic loci [14–16], had the expected characteristics. We have identified 455 such elements in the GenBank and, because of their ubiquitous distribution in mammals, named them MIRs, for Mammalian-wide Interspersed Repeats.

These genomic repeats were used to derive the putative MIR sequence that spread in mammalian lineages during the Mesozoic era. Abundant genomic fossil record was earlier used to reconstruct active sequences of Alu subfamilies retroposing in the past during primate evolution (for the most recent review see [17]). This approach characterizes a discipline that can be called paleogenomics; ancestral sequences that do not exist in an active form any more are here inferred from bits and pieces of their mutated copies still found in the genome. This differs from the usual phylogenetic reconstruction, where ancestral states are deduced by comparing homologous sequences that are active in contemporary genomes. Recently presented example of a paleogenomic reconstruction, including proof of a function of the deduced ancient sequence, illustrates an even greater potential of this approach [18].

MATERIALS AND METHODS

GenBank search and derivation of the consensus sequence

Preliminary consensus sequence was compared with all the mammalian sequences from GenBank release 69.0 using Dasher2 program (D.V. Faulkner, 1986, unpublished). The search was followed by a pairwise alignment of all identified MIR elements with the consensus sequence using LOCAL program [19]. All sequences matching the MIR consensus with an arbitrarily chosen LOCAL score of 20.0 or more were selected. Using this criterion 455 MIR sequence segments were extracted from all the available mammalian sequences. The pairwise alignments have been integrated into a multiply aligned set of MIR sequences using BLOWUP program (A. Milosavljevic and J. Jurka, 1989, unpublished). The new consensus sequence was created and the pairwise alignment was repeated several times until the consensus remained unchanged. The detailed analyses of the multiply aligned set have been done using the sequence editor MASE [20].

Inter-MIR PCR

DNA samples were either isolated from peripheral blood or tissue samples as described [21] or were generously provided by different individuals. The PCR was carried out in 10 mM Tris–HCl, pH 9.0, 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin, 0.01% Triton X-100, 2% formamide, 200 μM dNTPs (each), 1 μM primer (5′-end labeled), with 25 ng of genomic DNA per reaction (20 μl). After an initial denaturation of 7 min at 94°C and the addition of 1 U Taq DNA polymerase at 4°C, 27 cycles of amplification (each consisting of 30 s at 94°C, 45 s at 56°C and 120 s at 72°C) were followed by a final extension at 72°C.

* To whom correspondence should be addressed

Sequences of MIR-specific primers (*mir17*, *omir17*, *mil17*, *omil17* and 3′*mir19*) are given in Figure 1; that of the randomized sequence primer *nicmir17* (synthesized based on the same nucleotide composition as *mir17*) was 5′-TCTGAGATGG-CATTCGA-3′. PCR products were analyzed by electrophoresis in 6% non-denaturating polyacrylamide gel (acrylamide to N-N′-methylene-bisacrylamide ratio of 29:1 in 90 mM Tris–borate, pH 8.3, 2 mM EDTA).

RESULTS AND DISCUSSION

A general consensus resulting from the alignment of 455 identified MIR elements is presented in Figure 1. It consists of a 67 nucleotide-long 'core' sequence (positions 30 through 96), defined by the most frequent base at each position (upper part of the histogram). This consensus can be extended by 33 nucleotides if the positions where gaps appear to be the most frequent (lower histogram) are taken into account. However, because of the very prevalence of the gaps and of a large contribution of nucleotides from the sites of cutoff during the 'extraction' by the LOCAL program, the proposed sequence at these positions should be considered only tentative. Among MIR elements that contribute to the consensus, 381 represent primate (mostly human) and 55 rodent sequences. The remaining are from other mammalian groups, such as artiodactyls (twelve elements), lagomorphs (three), carnivores (three) and pinnipedia (one). Two marsupial (oppossum) sequences were identified later. Since the GenBank release 69.0 contained approximately 0.3% of the primate genome (i.e. discounting the duplicates [22]), the density of MIR elements can be estimated at about 120,000 copies per haploid genome. This may be compared with an approximatively 5-fold higher genomic abundance of Alu sequences, experimentally determined at half million to a million copies [23–25] and represented by about 1800 copies in the same GenBank release [22]. The estimated genomic density of MIR elements in rodents appears to be at least 5-fold lower than in primates, whereas the sequence information is too fragmentary to consider other groups of mammals. Hence, we examined the distribution of MIR elements in a variety of genomic DNAs utilizing polymerase chain reaction (PCR).

Based on the consensus sequence we have designed five MIR-specific primers that are shown in Figure 1. These primers can be used one at a time to amplify DNA segments between adjacent repeats occuring in an amplifiable distance (inter-MIR PCR, Fig. 2A; see also refs. [26–28]). Primers *mir17*, *omil17* and 3′*mir19* direct the amplification from DNA loci flanked by these repeats in a head to head orientation, while primers *mil17* and *omir17* when MIR elements are in a tail to tail orientation. Typical results of an inter-MIR PCR experiment, using *mil17* primer, are shown in Figure 2B, where a variety of DNA samples have been examined. Multiple bands were revealed, varying in size between 0.2–2.0 kb; their distribution was more similar among related species than between divergent taxa. The number of bands also differed significantly: 30–40 were scored in five rodent species, rabbit (lagomorphs), shrew (insectivores) and tree-shrew (Scandentia), and between 55 and 75 in other mammals such as primates, artiodactyls, perissodactyls and carnivores, whereas about 115 discernible PCR-products were obtained from three marsupial DNA samples. Few bands were also seen in birds and reptiles. These results are diagrammed in Figure 2C which also summarizes those obtained with three other MIR-specific primers (data not shown). Although platypus DNA was not included in the main analysis, its inter-MIR PCRs (as far as bands density is concerned) were comparable with those of marsupials (not shown). PCR carried with a randomized sequence primer *nicmir17* was negative except for two solitary bands seen in orangutan and gibbon DNAs, presumably fortuitously amplified by random priming. The primer *omil17*, which was the most efficient in all mammalian species, also led to DNA amplification in non-mammalian samples: the origin of this relaxed specificity may be related to this primer's G-rich 3′-end. In contrast, a relatively high number of amplified products seen in fishes while using the primer *omir17*, can be explained by a perfect, presumably fortuitous complementarity between this primer 3′-end and repeated sequence occuring in salmon and related fish genomes [29].

PCR analysis (Fig. 2C) confirmed earlier conclusions from GenBank searches. Significant amplification with all four primers occurred only in mammalian DNAs, and rodent DNA was consistently a poorer template in inter-MIR PCR than the primate DNA. This effect can be accounted for by higher substitution rate in rodents than in primates [30], causing thus a faster decay of non selected genomic segments in these species. The apparently lower abundance of MIR elements in insectivores, lagomorphs, scandentia could be explained on the same ground [31]. In contrast, the highest density of bands was documented in marsupials and platypus. Few amplification products observed in birds and reptiles could be either fortuitous or might reflect a very rare occurence of MIR elements in these genomes. Since amplification products were observed using four different primers, the latter seems to be the case; however, given a very low level of the amplification, PCR products probably did not represent inter-MIR segments but resulted from the amplification primed within MIR element at one side and within randomly matching genomic sequence at the other.

The presence of MIR elements in different taxa was further examined by PCR using primer pairs amplifying DNA within the repeat. The use of the *omil17/omir17* primer pair (Fig. 1) resulted in a single product of the expected length in the majority of placental mammals (data not shown). Likewise, DNA fragment of about 60 bp, predicted while using the primer pair 5′*mir19/mir17*, was observed in all mammalian species. We also observed an amplification in three avian species. This would be consistent with a rare occurence of a similar sequence element in birds, as suggested by fingerprinting data (Fig. 2).

Fig. 3A presents a pairwise comparison of 9 orthologous loci between human and non-rodent mammals aligned with the MIR consensus (six shared loci between human and rodents were reported earlier [14–16]). The average pairwise nucleotide divergence (K, see Fig. 3B) between the MIR elements in humans and other mammals (including both MIR and flanking segments) is 0.30 or 30%. If we assume that the eutherian radiation occured 65 million years ago, then these elements have been diverging from their common ancestral sequence at about 0.23% per million years [i.e. (30%/2)/65 million years]. This rate of nucleotide substitution falls within the range estimated for primates and rodents [32]. All genomic repeats appear to have accumulated mutations randomly with respect to the consensus, except for the shared substitutions among orthologous repeats confirming their common history prior to mammalian radiation. The average pairwise divergence between the MIR consensus and either the human elements or the other, non-human MIRs is 0.29 or 0.33,

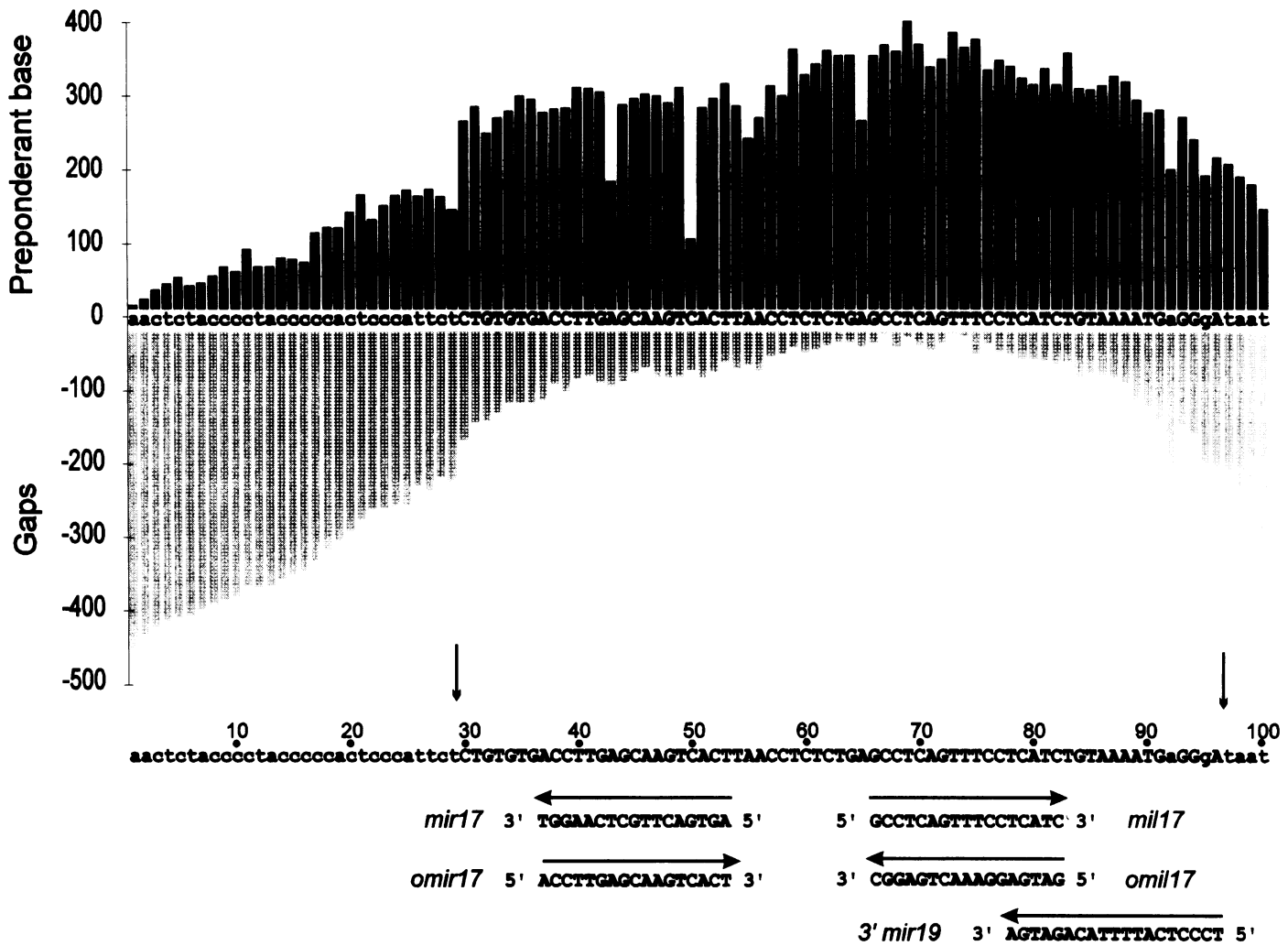


Figure 1. Alignment of 455 MIR elements and derivation of the consensus sequence. Upper histogram shows the observed number of the most frequent base at each sequence position, while the lower one the number of gaps, i.e. 'empty' positions reflecting frequent truncation of the genomic MIR elements at both termini. The consensus and the sequences of oligonucleotides used as MIR-specific PCR-primers are indicated below. The most abundant nucleotides at each position are indicated by upper case letters and define a core consensus sequence, delimited by arrows. Lower case letters at both ends of the core sequence (positions 1–29 and 97–100) represent preponderant nucleotides in the non-core fragments, where truncation based on the arbitrarily chosen LOCAL [19] cutoff value was introduced in a majority of sequences (see Materials and Methods). Lower case letters inside the core region (a and g at positions 92 and 95, respectively) show nucleotides which were the second most abundant to alignment gaps at these positions.

respectively (Fig. 3B). If we assume that the consensus MIR is a reasonable approximation of the original MIR sequence and that the elements in modern genomes have been diverging from it at about 0.23% per year, then we estimate that the original MIR element existed about 130 million years ago [i.e. 29% (or 33%)/(0.23%/million years)]. In other words, the average time elapsed between MIR amplification and the present can be estimated at about 130 Myr, ie. twice that separating contemporary orders of mammals. This value can be possibly extended back in time, since the detection of MIRs in both GenBank searches and PCR-experiments may preferentially target the youngest and/or less mutated sequences.

Our data raise the possibility of MIRs originating from sequences common to mammals and birds. The first MIR elements could have appeared at the beginning of the Mesozoic era, contemporarily to early mammals [33]. Their subsequent amplification presumably occured independently in placental and non-placental lineages, consistent with a higher abundance of MIR elements in marsupials (Fig. 2) and monotremes, and with the results shown in Figure 3. There is no evidence of recent MIR amplification, at least in placental orders such as primates and/or rodents. It is likely that MIR amplification ceased in ancestors of placental mammals giving place to dispersal of other short sequence families that are order-specific (such as primate Alu or rodent B1 and B2).

Most of mammalian SINEs [10–12] dispersed by retroposition, i.e. using as intermediates their RNAs transcribed by polymerase III. In the presented MIR consensus we did not find any Pol III recognition signals nor other sequence features typical for retroposable elements, such as direct repeats and A-rich tails.

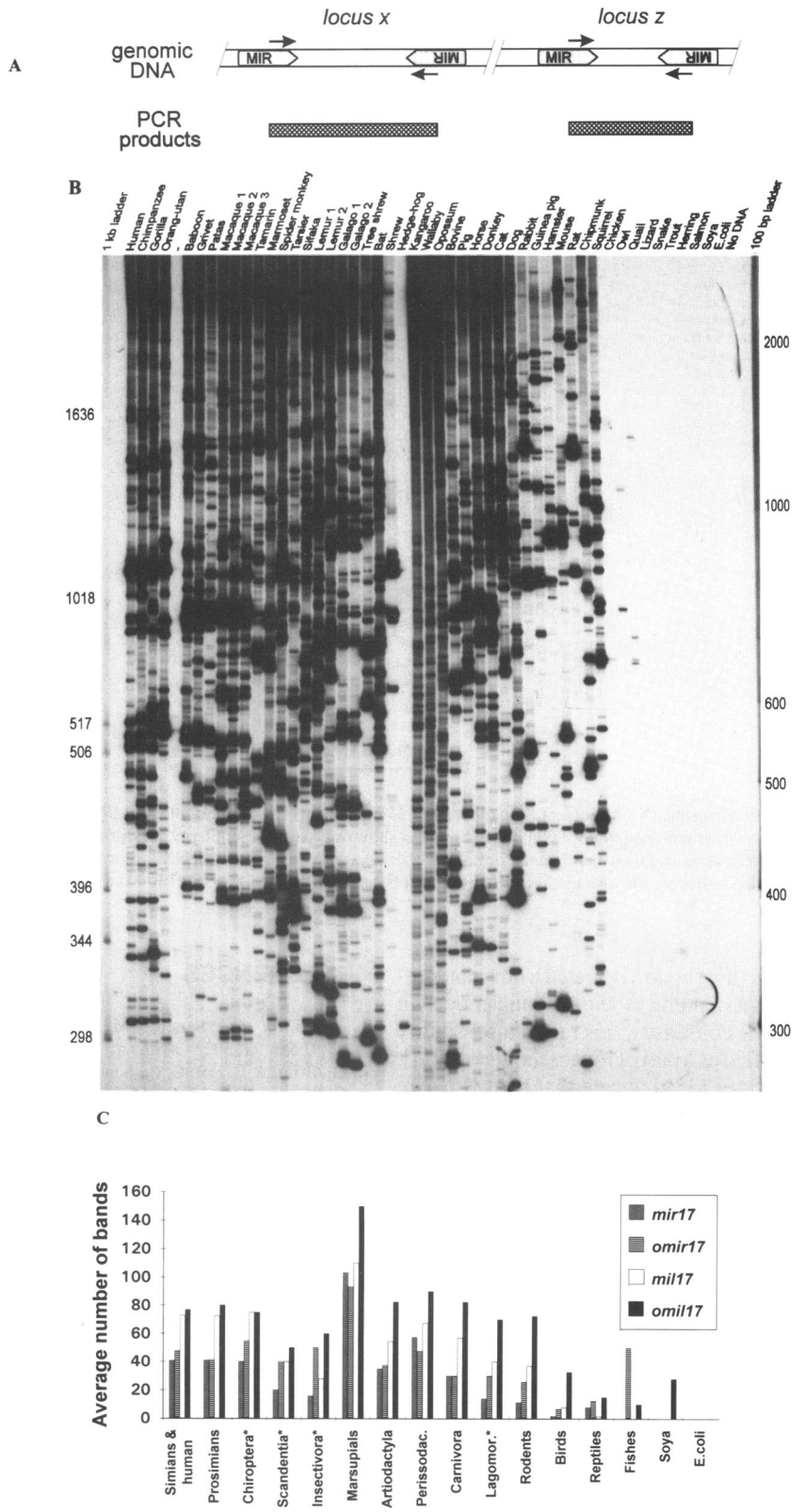
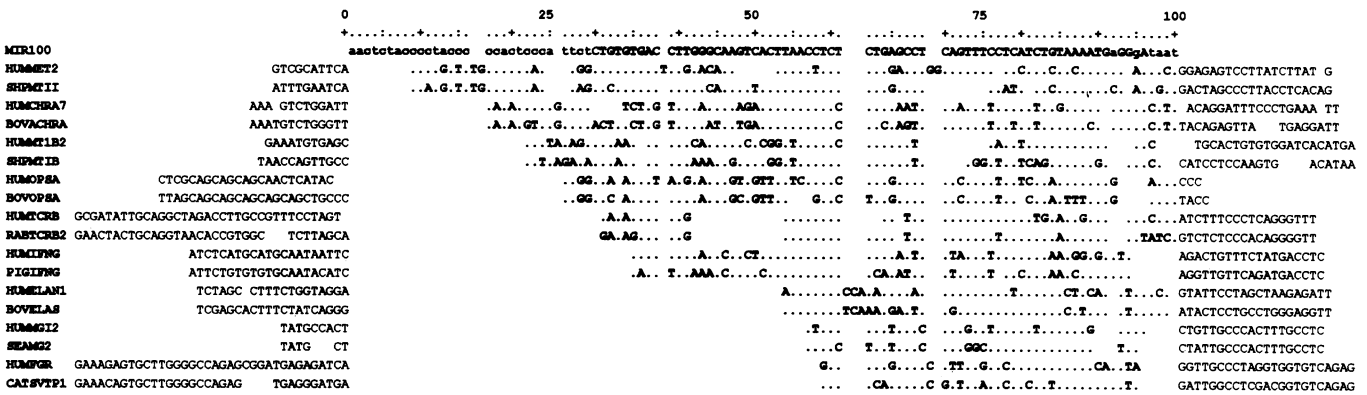


Figure 2. PRC-amplification of genomic segments flanked by MIR-elements. **(A)** Scheme of the inter-MIR PCR. **(B)** The autoradiogram of PCR products obtained using the primer *mil*17. The origin of DNA samples is indicated at the top; M- molecular weight marker (100 bp ladder Pharmacia, 1 kb ladder, BRL). **(C)** The estimated number of distinct PCR products in 0.2–2 kb size range obtained with MIR-specific primers, *mir*17, *omir*17, *omil*17 and *mil*17 (shown in Fig. 2B). Related species are grouped together whereas mammalian orders marked with an asterisk are only represented by a single species.

A



B

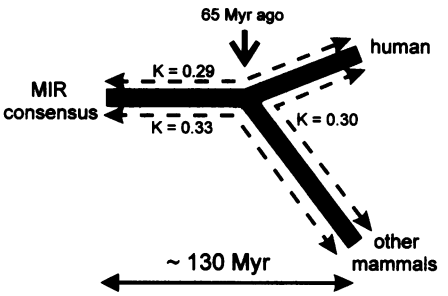


Figure 3. Proposed timing of MIR elements. **(A)** Alignment of the 9 pairs of MIR-elements found in orthologous positions in human and non-primate mammalian DNA. Identity with the MIR consensus is indicated by dots; flanking regions that show only pairwise similarity are shown to the left and right. **(B)** Distance tree between the MIR consensus, human MIR elements and their mammalian orthologs. Pairwise distances were corrected according to [36]. Assuming molecular clock and 65 Myr ago as time period of eutherian radiation, the average age of MIR elements can be estimated at about 130 Myr.

However, statistical analysis (not shown) detected a non-randomness in the flanking regions extending at least 30 nucleotides in both directions from the MIR consensus; this raised a possibility that MIRs could have amplified as a part of longer sequences [34]. Recent data by Smit and Riggs [35] indicate that the MIR core sequence corresponds to a segment of a longer retroposable SINE element. Analysis of MIR elements from distant genomes, and especially those from non-eutherian species, will certainly shed light on the origin of these elements that have marked mammalian evolution during the Mesozoic.

ACKNOWLEDGEMENTS

We are grateful to Drs Clément Lanthier, Robert Patenaude, Daniel Sinnett, Pascale Benoist, Morris Goodman, Jerry Slightom, Michael Stanhope, Annie Depeiges, Antoni Rafalski, Jennifer M. A. Graves and Barbara Cofman for help and collaboration in collecting tissue or DNA samples from a variety of species and to Arian F. A. Smit and Arthur D. Riggs for communicating their data prior to publication. We would like to thank Raffaela Ballarano and Micheline Patenaude for typing the manuscript. This study was supported by The Cancer Research Society, Inc., (D. L.) and the U.S. Department of Energy, Human Genome Program, Grant No. DE-FG03-91ER61152. (J. J.).

REFERENCES

1 Wilson, A.C., Sarich, V.M., Maxson, L.R. (1974) *Proc. Natl. Acad. Sci. U.S.A.* **71**, 3028–3030.
2 Brini, A.T., Lee, G.M., Kinet, J-P. (1993) *J. Biol. Chem.* **268** 1355–1361.
3 Brosius, J. and Gould, S.J. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10706–10710.
4 Garret, M., McHendry-Rinde, B., Spickofsky, N., Margolskee, R. (1992) *Gene* **111**, 215–222.
5 Kim, J.H., Yu, C.Y., Bailey, A., Hardison, R., Shen, C.K.J. (1989) *Nucleic Acids Res.* **17**, 5687–5700.
6 Lehrmann, M.A., Russell, D.W., Goldstein, J.L., Brown, M.S. (1987) *J. Biol. Chem.* **7**, 3354–3361.
7 Onda, M., Kudo, S., Rearden, A., Mattei, M.G., Fukuda, M. (1993) *Proc. Natl Acad. Sci. U.S.A.* **90**, 7220–7224.
8 Rouyer, F., Simmler, M.C., Page, D.C., Weissenbach, J. (1987) *Cell* **51**, 417–425.
9 Vidal, F., Mougneau, F., Gaichenhaus N., Vaigot P., Darmon M., Cuzin F. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 208–212.
10 Singer, M.F. (1982) *Int. Rev.Cytol.* **76**, 67–112.
11 Weiner, A.M., Deininger, P.L., Efstratiadis, A. (1986) *Ann. Rev.Biochem.* **55**, 631–661.
12 Mobile DNA, Berg, D.E. and Howe, M.M., Eds. (American Society for Microbiology, Washington DC, 1989).
13 Degen, S.J., Davie, E.W. (1987) *Biochemistry* **26**, 6165–6167.
14 Donehower, L.A., Slagle, B.L., Darlington, G., Butel, J.S. (1989) *Nucleic Acids Res.* **17**, 699–710.
15 Armour, J.A.L., Wong, Z., Wilson, V., Royle, N.J., Jeffrey, A.J. (1989) *Nucleic Acids Res.* **17**, 4925–4935.
16 Jurka, J. (1990) *Nucleic Acids Res.* **18**, 137–141.
17 Schmid, C. and Maraia, R. (1992) *Curr. Opin. Genet. Develop.* **2**, 874–882.

18 Adey, N.B., Tollefsbol, T.O., Sparks, A.B., Edgell, M.H., Hutchinson III, C.A. (1994), *Proc. Natl. Acad. Sci. U.S.A.* **91**, 1569–1573
19 Smith,T.F., Waterman, M.S., (1981) *J. Mol. Biol.* **145**, 195–197.
20 Faulkner, D.V. and Jurka, J. (1988) *Trends in Biochem Sci.* **13**, 321–322.
21 Sinnett D., Lavergne L., Melancon SB., Dallaire L., Potier M., Labuda D. (1988) *Hum Genet.* **81**: 4–8.
22 Jurka, J., Kaplan D.J., Duncan C.H., Walichiewicz J., Milosavljevic A., Murali G., Solus J.F. (1993) *Nucleic Acids Res.* **21**, 1273–1279.
23 Hwu, H.R., Roberts, J.W., Davidson, E.H., Britten, R.J. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3875–3879.
24 Rinehart, F.P., Ritch, T.G., Deininger, O.L., Schmid, C.W. (1981) *Biochemistry* **20**, 3003–3010.
25 Moyzis, R.K., Torney D.C., Meyne J., Buckingham JM., Wu J-R., Burks C., Sirotkin KM., Goad WB. (1989) *Genomics* **4**, 273–289.
26 Sinnett, D., Deragon, J-M., Simard, L.R., Labuda, D. (1990) *Genomics* **7**, 331–334.
27 Zietkiewicz, E., Labuda, M., Sinnett, D., Glorieux, F.H., Labuda D. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8448–8451.
28 Zietkiewicz, E., Rafalski, A., Labuda, D. (1994) Genomics **20**, 176–183.
29 Kido, Y., Aono M., Yamaki T., Matsumoto, K-I., Murata, S., Saneyoshi, M., Okada N. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 2326–2330.
30 Li, W-H. and Tanimura, M. (1987) *Nature* **326**, 93–96.
31 Martin, A.P. and Palumbi, S.R. (1993) *Proc. Natl. Acad. Sci U.S.A.* **90**, 4087–4091.
32 Li, W.,H. & Tanimura, M. (1987) *Nature* **326**, 93–99
33 Novacek, M.J. (1992) *Nature* **356**, 121–125.
34 Smit, A.F.A. (1993) *Nucleic Acids Res.* **21**, 1863–1872.
35 Smit, A.F.A., Riggs, A.D. (1995), *Nucleic Acids Res.*, **23**, 98–102.
36 Jukes,T.H. and Cantor, C.R. in *Mammalian Protein Metabolism*, H.N. Munro, Ed., (Academic Press New York 1969), pp.21–123.