98–102 Nucleic Acids Research, 1995, Vol. 23, No. 1

MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation

Arian F. A. Smit* and Arthur D. Riggs

Department of Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010, USA

Received September 26, 1994; Revised and Accepted October 25, 1994

ABSTRACT

Short Interspersed Nucleotide Elements (SINEs) are highly abundant in mammalian genomes. The term SINE has come to be restricted to short retroposons with internal RNA polymerase III promoter sites in a region derived from a structural RNA (usually a tRNA). Here we describe a novel, 260 bp tRNA-derived SINE, some fragments of which have been noted before to be repetitive in mammalian DNA. Unlike previously reported SINEs, which are restricted to closely related species, copies of this element can be found in all mammalian genomes, including marsupials. It is therefore called MIR for mammalian-wide interspersed repeat. Their high divergence and their presence at orthologous sites in different mammals indicate that MIRs, at least in part, amplified before the mammalian radiation. Next to Alu, MIRs are the most common interspersed repeat in primates with an estimated 300 000 copies still discernible, which account for 1 to 2% of our DNA. Interestingly, a small, central region of MIR appears to be much better conserved in the genomic copies than the rest of the sequence.

INTRODUCTION

SINEs are interspersed repetitive nucleotide elements of 100–300 bp which are found in most vertebrates as well as in invertebrates (1,2). Typical SINEs share a number of common structural features: they contain an internal RNA polymerase III promoter, are flanked by variable length insertion site duplications and usually end in an A- or T-rich tail or a short simple sequence repeat. The internal promoter is in a region apparently derived from a structural RNA and may be essential for the formation of a SINE (3). Most SINEs seem to be fusion products of a tRNA-derived gene and an unrelated sequence (2,4–6).

A wide variety of uncharacterized short interspersed repeats has been catalogued in human DNA (7). In an effort to determine the origin and nature of these sequences we derived full consensus sequences from the genomic copies and found that many of these sequences are fragments of LTR transposons, LINE1 3′ ends and putative DNA transposons (8,9, unpublished results).

One of these repetitive sequences is ubiquitous in all placental mammals (10) and therefore has been named MIR for Mammalian-wide Interspersed Repeat (7). The repetitive character of MIR was first noted in 1987 (11) and has been rediscovered several times (e.g. 7,12,13). Korotkov (14) believed that this same fragment resembled the mirror image (in purine and pyrimidine sequence) of the rodent B1 repeat, and named it MB1 (Mirror of B1). Elements at orthologous sites in different mammalian species form evidence that the distribution of MIRs took place before the mammalian radiation (10). In the accompanying paper, Jurka *et al.* (15) show that MIRs are even highly repetitive in marsupial and monotreme genomes.

All authors (10,14,15) describe MIRs as about 70 bp long elements without the typical features of 'generic' SINEs, as outlined above. It has, therefore, been suggested that MIRs represent a separate class of repetitive elements (10). Below we will show that this 70 bp MIR element and two other previously reported uncharacterized interspersed repetitive sequences are fragments of a 260 bp, classic tRNA-derived SINE. A second, ancient and abundant short interspersed repeat with limited sequence similarity to MIR, here named MIR2, is presented as well.

MATERIALS AND METHODS

We screened GenBank, Release 82.0, with the published 70 bp MIR consensus sequence (10), using the IFind program (16) in the IntelliGenetics sequence analysis package. IFind was used with default parameters: word-length 4 (search) or 2 (alignment), gap-penalty 4, window-size 40, density = less. After removal of more recently inserted elements, like Alu, regions containing this sequence were compared in pairs and the consensus sequence was extended in either direction as far as entries showed continued similarity to each other. Using the extended consensus sequence, additional MIR copies were found and added to the alignment, leading to improvement of the consensus. The MIR consensus sequence was optimized by expansion of the data set by successive searches with improved consensus sequences until addition of new copies had no further effect on the consensus.

Searches with the blastn program in the NCBI e-mail server (17) were performed, with default parameters, to obtain information on the conservation of fragments of MIR. For calculation of the number of matches, the cutoff score was set so that only one match is expected to occur at random ($E = 1$, $P \leq 0.99$). Redundant entries were disregarded, but multiple matches in one sequence entry were counted.

*To whom correspondence should be addressed at: Department of Molecular Biotechnology, University of Washington, FJ-20, Seattle, WA 98195, USA

RESULTS AND DISCUSSION

MIRs are generic SINEs

By aligning over 80 sequences containing MIR similarities, we could construct a MIR consensus of 260 bp (Fig. 1). The consensus may be considered an approximation of the original, transpositionally active element. It has consensus RNA polymerase A and B boxes and an A/T-rich 3′ end, characteristics of a typical SINE (Fig. 2). The third characteristic of (variable length) direct flanking repeats is likely to have become unrecognizable since the MIR fragments are 25–35% diverged from the consensus. The 5′ 80 bp of MIR containing the A and B boxes are similar to several tRNAs, and even more so to the tRNA-like region of the rodent B2 SINE (4–6). Like many SINEs, MIR thus resembles a fusion product of a tRNA-derived gene and an unrelated sequence. It is hard to identify the exact, ancestral tRNA of MIR (and most other SINEs), since some tRNA sequences are very similar to one another, while the SINE source genes may have diverged considerably from the parental sequence. The 5′ end of the MIR transcript may have retained a tRNA-like

Figure 1. Part of the alignments from which the MIR consensus sequence is derived. A selection of loci (referred to by their GenBank locus names) containing relatively large fragments of MIR is presented in this figure. Ambiguous sites are indicated above the consensus (Y = C/T, R = A/G). Nucleotides that differ from the derived consensus are given as letters. Otherwise, dots denote identical nucleotides, dashes gaps, and numbers insertions of that length. Blank spaces at the beginning and end of entries indicate absence of significant similarity to the consensus.

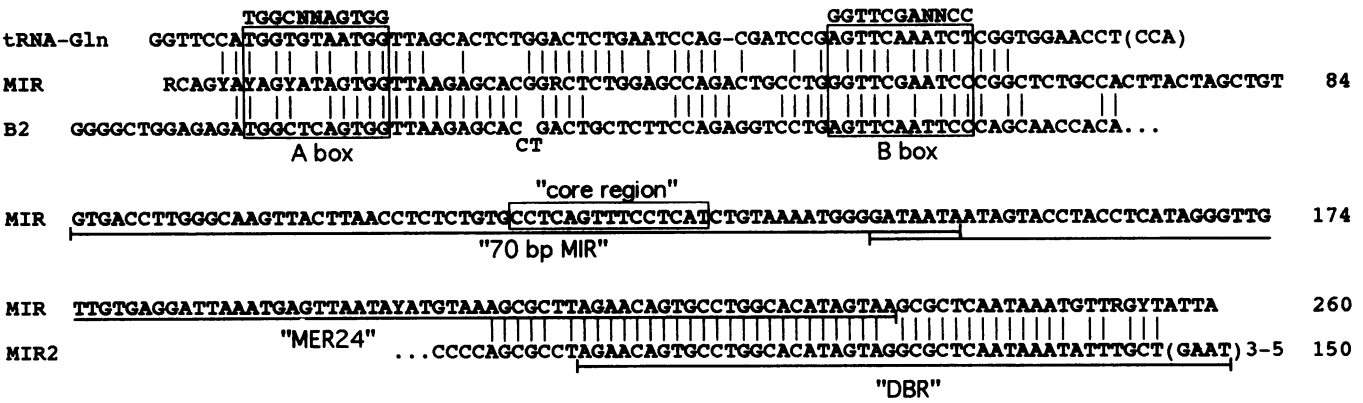

Figure 2. Alignment indicating that MIRs are classic SINEs, consisting of a tRNA-derived region containing a consensus RNA polymerase III promoter (box A and B) fused to an unrelated sequence. The 5′ end of the MIR consensus sequence is compared to the human Gln-tRNA-CUG gene (Q995) (22) and the tRNA-like portion of the mouse B2 consensus sequence as presented in (6). Other tRNAs show considerable similarity to this region as well, but the above sequences were chosen for their high similarity to MIR outside the A and B boxes. The location of the 70 bp MIR consensus and 15 bp core sequence (10), and the prototypes for MER24 and DBR (7) are indicated. They have been published in the opposite orientation as the current consensus.

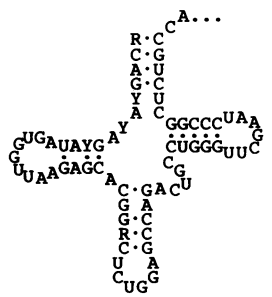

Figure 3. Presentation of the 5′ end of the putative MIR transcript in a tRNA secondary structure. Imperfect stem symmetries are also predicted in other tRNA-derived SINE transcripts (4–6). The only inconsistency with a tRNA structure is the presence of six instead of seven residues in the anticodon loop.

cloverleaf secondary structure (Fig. 3). Although determination of the extent of an interspersed repeat is difficult when flanking repeats are not available, the current consensus sequence probably comprises the whole element, since the 5′ end coincides with that of most tRNA genes, while the 3′ end is an A/T rich tail typical for SINEs.

In addition to the '70 bp MIR' sequence, the consensus sequence includes two more regions which have been described before as separate interspersed repeats: MER24 and DBR (7) (Fig. 2). The latter 'DBR' region occurs with a few diagnostic differences at the end of another interspersed repetitive element of about 150 bp, here named MIR2, which shares no further sequence similarity with MIR (Fig. 4). Like MIR, this element can be found in genomes of distantly related mammals and some elements at orthologous sites (Fig. 4) manifest that its distribution took place, at least in part, before the mammalian radiation as well. The current consensus sequence of MIR2 has an oligo (GAAT) tail, but lacks similarity to a Pol III promoter region. Most MIR2 copies seem truncated at the 5′ end and, although we found no further similarity, it is possible that a full-length MIR2 sequence extends upstream to include such a promoter region.

Most SINEs can be divided into subfamilies of copies that share several diagnostic sites (1). These are thought to represent amplifications from an evolving single or small number of source gene(s). Despite the large number of copies available, we could not group the MIR copies into subfamilies with any confidence, although some members share a few possibly diagnostic mutations from the consensus. However, subfamilies may still exist, since they are increasingly hard to distinguish among more diverged sequences.

The MIR and MIR2 consensus sequences have been deposited in the human repetitive sequence database available via anonymous FTP at ncbi.nlm.nih.gov in the directory repository/repbase.

Number of MIRs in the genome

Next to the primate-specific Alu, MIRs are probably the most abundant interspersed repeats in mammals. The '70 bp MIR' has been estimated to occur several 100 000 times in the human genome (14). Estimates of the total number of MIRs are hindered by their high divergence. Using the blastn program in the NCBI e-mail server (17), 689 human sequence entries were found to contain matches to MIR (October 1994), including redundant entries and cross-matching MIR2 sequences. 132 matches were found among rodent entries (mostly representing B2 repeats) and 61 in other mammals. However, blastn is optimized to find almost identical matches to a query sequence and is likely to overlook most MIR copies. Even when using programs that allow for the introduction of gaps, many MIR copies do not score above background under standard search conditions, and can only be assessed individually. Therefore, we performed a local, more thorough search using the program IFind (16) and identified 72 MIRs (only 18 of which were found by blastn) and 26 MIR2s in the 670 000 bp comprised by 6 large human contigs (GenBank entries HUMHBB, HUMRETBLAS, HUMNEUROF, HUMMMDBC, HUMTCRADCV and HSU07000). The MIRs are not randomly distributed throughout these contigs, but often cluster, a phenomenon also observed for Alu and other SINEs. By extrapolation, one may expect about 300 000 MIRs ($3*10^9 * 72 / 670\,000$) and 100 000 MIR2s to be detectable in the haploid human genome. With an average remaining length of the MIR fragments in these contigs of 170 bp (100 bp for MIR2) we estimate that copies of both elements

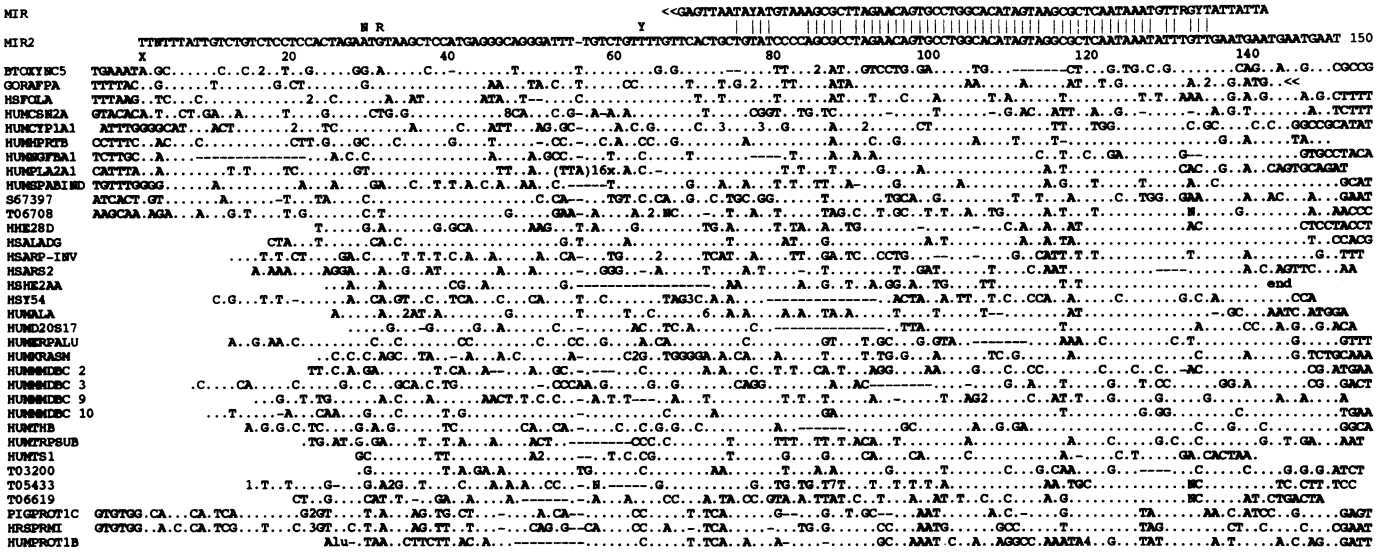

Figure 4. Derivation of the consensus sequence for MIR2. The 3′ end of MIR2 is very similar to that of MIR. See Figure 1 for explanation. The last three sequences are MIR2 elements at orthologous sites in the pig, horse and human protamine 1 gene promoter region.

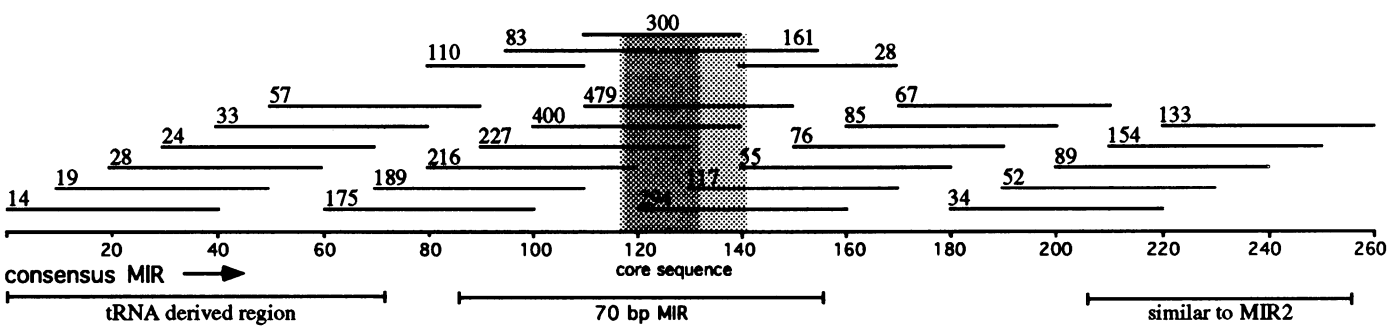

Figure 5. Non-random conservation of the MIR consensus sequence in the genome. 30 bp and 40 bp fragments of the consensus sequence are indicated that have been used as a query in blastn searches of the NCBI databases of October 1994; for each fragment the number of matching sequences ($P < 0.99$) in unique human entries is shown. The location of the 15 bp 'conserved core sequence' reported by Donehower *et al.* (10) is shown as a darkly shaded area; note that all fragments with a high number of matching sequences in the databases include this region, although a core sequence of at least 25 bp (light shading) is more consistent with the presented data. The number of matches to the 3′ end are somewhat higher since they include matches to MIR2.

together constitute about 2% of our DNA. It is likely that an even larger fraction of the genome has its origin in these elements but can not be recognized as such anymore.

A conserved core sequence in MIRs

Using fragments of the 260 bp consensus sequence as queries in database searches, we found many more matching sequences to the central region than to either site of MIR (Fig. 5). This observation may partly explain why previous reports held the length of the MIR element to be about 70 bp, constituting the central region of MIR (Fig. 2). The central region of MIR would be overrepresented if it occurs as a module in one or more other interspersed repeats (like the 'DBR' region in MIR and MIR2). We found no evidence to support this; the analyzed MIR copies usually contain the central region and appear to be truncated at either or both ends (MIR2 copies are usually truncated at the 5′ end), but the similarity to the consensus sequence extends, on average, over 170 bp and ends at random positions. Theoretically, this could be the result of an often incomplete integration process, analogous to that of L1, but effecting both the 5′ and 3′ ends. However, there is no precedent for this in other generic SINEs. A third explanation for the observation would be that the central region has been better conserved than the terminal sequences. The latter have some mutagenic CpG sites, but the difference may be principally in the number of introduced gaps, for which the program blastn is susceptible. A particularly conserved region of about 25 bp (bp 117–141) could account for part of the pattern shown in Figure 5. This region overlaps the 15 bp 'conserved core region' reported by Donehower *et al.* (10). Such conservation is most easily explained if this core region is/was recognized and bound by a protein or has (had) another function in the genome.

Based on their limited species range in mammals (as well as in other organisms) it has been suggested that major SINE amplifications have occurred only in the last 65 million years (after the eutherian radiation), and that SINEs did not exist in quantity prior to this time (18–20). Although a mesozoic origin of the Alu/B1 family of SINEs is likely (21), Alu elements were

certainly not abundant before the mammalian radiation. However, MIRs are generic looking SINEs that amplified in large numbers before this time and probably originated even before the split of eutherians and marsupials; fragments of MIRs are found in intron 2 of the opossum β-hemoglobin β-M gene (OPOHBBB 1210–1332) and in the 3′ untranslated region of the Na/Pi-cotransporter mRNA (OPONAPICO 2218–2325) (Fig. 1). The abundance of MIR copies in the genome is clear evidence that the formation and efficient retrotransposition of tRNA-derived SINEs is not an evolutionary novelty.

ACKNOWLEDGEMENTS

This work was supported by NIH grants GM50575, GM46986 and DOE grant DE-FG03-91ER61137 to A. D. Riggs.

REFERENCES

1 Deininger, P.L. and Batzer, M.A. (1993) *Evol. Biol.* **27**, 157–196.
2 Okada, N. (1991) *Trends Ecol. Evol.* **6**, 358–361.
3 Quentin, Y. (1992) *Nucleic Acids Res.* **20**, 3397–3401.
4 Sakamoto, K. and Okada, N. (1985) *J.Mol.Evol.* **2**, 134–140.
5 Lawrence, C.B., McDonnell, D.P. and Ramsey, W.J. (1985) *Nucleic Acids Res.* **31**, 4239–4252.
6 Daniels, G.R. and Deininger, P.L. (1985) *Nature* **317**, 819–822.
7 Jurka, J., Walichiewicz and Milosavljevic, A. (1992). *J. Mol. Evol.* 35, 286–291
8 Smit, A.F.A. (1993). *Nucl. Acids Res.* **21**, 1863–1872.
9 Smit, A.F.A., Tóth, G., Riggs, A.D. and Jurka, J. (1994) *J. Mol. Biol.*, in press.
10 Donehower, L.A.,, Slagle, B.L., Wilde, M., Darlington, G. and Butel, J.S. (1989) *Nucleic Acids Res.* **17**, 699–710.
11 Degen, S.J. and Davie, E.W. (1987) *Biochemistry* **26**, 6165–6177.
12 Zhou, Y.Z., Slagle, B.L., Donehower, L.A., Van Tuinen, P., Ledbetter, D.H. and Butel, J.S. (1988) *J. Virol.* **62**, 4224–4231.
13 Bancroft, J.D., Schaefer, L.A. and Degen, S.J.F. (1990) *Gene* **95**, 253–260.
14 Korotkov, E.V. (1991) *Mol. Biol. (USSR)* **25**, 250–263.
15 Jurka, J., Zietkiewicz, E. and Labuda, D. (1994) *Nucleic Acids Res.* 00, 000–000.
16 Wilbur, W. J. and Lipman, D. J. (1983). *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
17 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman , D.J. (1990). *J. Mol. Biol.* **215**, 403–410.
18 Weiner, A.M., Deininger, P.L. and Efstradiadis, A. (1986) *Annu.Rev.Biochem.* **55**, 631–661.
19 Deininger, P.L. and Daniels, G.R. (1986) *Trends Genet.* **2**, 76–80.
20 Deininger, P.L. In *Mobile DNA* (eds. Berg, D.E and Howe, M.M.) 619–636 (ASM, Washington, 1989)
21 Quentin, Y. (1994) *Nucleic Acids Res.* **22**, 2222–2227.
22 Sprinzl, M., Hartmann, T., Weber, J., Blank, J. and Zeidler, R. (1989) *Nucleic Acids Res.* **17**, r1–r155.