J Mol Evol (1989) 28:191–199

Journal of
Molecular Evolution

Similarity and Divergence among Rodent Repetitive DNA Sequences

William Bains and Kay Temple-Smith

Department of Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Summary. We have analyzed the sequence of 63 B1 and 71 B2 repetitive elements from published data base sequences. The sequences conform to previously published consensus sequences, but are not identical to them. The B2 sequences show seven regions of high variability between family members, which we show points to the B2 family containing subfamilies; no similar evidence is found for subfamilies of the B1 family. The comparisons show no evidence for the emergence of species-specific variants of B1 or B2 sequences since the separation of murine and hamster lines of descent, nor of their concerted evolution within species in the last 10 million years.

Key words: Repetitive DNA — Rodent — B1 — B2 — Species specificity — Selfish DNA

Introduction

Eukaryotic genomes contain many sequences that are present in more than one copy per haploid chromosome set (Davidson et al. 1973). Such sequences usually fall into families of sequences that are similar but not identical to each other. The most abundant families of dispersed sequences in rodent genomes are the B1 and B2 families (Rogers 1985; Schmid and Shen 1985). The members of the two families conform to a consensus structure, which is usually flanked by short direct repeats not homologous between elements; this suggests that they are integrated reverse transcripts (Rogers 1985). Individual members show variation from this consensus both in their sequence and their overall length.

Such repetitive DNA families have been postulated to spread through mammalian genomes and be maintained within them by a variety of mechanisms including the integration of reverse transcripts of elemental RNA, unequal crossover, and gene conversion within the family (reviewed by Dover 1982). Empirical testing of the results of these mechanisms requires DNA sequences from a number of elements from several closely related species, which have hitherto been unavailable. As B1 and B2 sequences are dispersed throughout rodent genomes, many of them have been sequenced when a gene in which they reside has been analyzed. These sequences provide a large data base of specific instances of family members, which allow more precise characterization of these elements than was possible on their initial description. In this paper we describe the analysis of 63 B1 and 71 B2 elements, and draw conclusions about their family structure and their modes of evolution.

Methods

Sequences were retrieved from the European Molecular Biology Laboratory (EMBL, Heidelberg; Hamm and Cameron 1986) and NIH GenBank (Bolt, Beranek and Newman Inc., for National Institutes of Health, USA; Bilofsky et al. 1986) data bases using LSEARCH and EXTRACT (Soundy, unpublished) running on the Imperial Cancer Research Fund DEC2060. B1 and B2 sequences were located in the data base by their homology to one or more of MUSRSBIA, RATCTRPB, and HAMRSA49D for B1 and MVSRSB2A, RATGH2, and HAMRSA250 for B2. This procedure was designed to reduce any species-specific bias in selecting data from the data base, although in practice nearly all of the resulting B1 and B2 elements were identified by all three of the relevant probes. A B1 or B2 sequence was included in the sequences analyzed here if (1) it showed at least 60% overall similarity to the "trial" sequence in the region in which they overlapped, (2) overlapped the "trial" sequence by at least 30 bp, and (3) did not consist of more than 50% poly-A tail. The sequences analyzed are shown in Table 1. Sequences were ana-

Offprint requests to: W. Bains

Table 1. Sequences used in this study

B1 sequences			B2 sequences		
Number	Name	Species	Number	Name	Species
1	musrsb1a	mouse	1	musrsb2a	mouse
2	muscycp4	mouse	2	musrsab1	mouse
3	musrsb1b	mouse	3	m11284	mouse
4	musrsb1c	mouse	4	musrsam2	mouse
5	musrsb1e	mouse	5	musrsb2ad	mouse
6	musrsaafp	mouse	6	musrsb2b	mouse
7	musmhtlp	mouse	7	musrsb2ab	mouse
8	musrsab1	mouse	8	musrsbac	mouse
9	musrsb1f	mouse	9	musren2g	mouse
10	m11800	mouse	10	musendob2	mouse
11	musmhtlac	mouse	11	m11944	mouse
12	m11741	mouse	12	musmhdd	mouse
13	m12379	mouse	13	musmhld3	mouse
14	musrsb1d	mouse	14	musren2sm	mouse
15	m10246	mouse	15	muserma	mouse
16	m11160	mouse	16	musmhdb	mouse
17	musrplpsc	mouse	17	m12381	mouse
18	m11944	mouse	18	musrplpsa	mouse
19	musmhkk	mouse	19	m11741	mouse
20	musmhab3	mouse	20	musrplpsc	mouse
21	musadfp14z	mouse	21	musmhtlac	mouse
22	musrsrp2	mouse	22	muscyp345	mouse
23	musmhkda	mouse	23	musigkvt2	mouse
24	musrps16	mouse	24	musigkvk	mouse
25	muscyp14x	mouse	25	musc31	mouse
26	musigkjc3	mouse	26	musigkvh1	mouse
27	musigkag5	mouse	27	musigkvj3	mouse
28	mushprt1	mouse	28	musrspr1a	mouse
29	m12561	mouse	29	musmhtlps	mouse
30	musmhdd	mouse	30	musmhcq3	mouse
31	musmdg5	mouse	31	musmhab3	mouse
32	m12976	mouse	32	musablii	mouse
33	musmhtlps	mouse	33	muscyp14x	mouse
34	muserfv42	mouse	34	musmopc	mouse
35	m12379	mouse	35	musgfapd	mouse
36	mus45srna	mouse	36	musins	mouse
37	musgpd1	mouse	37	m11742	mouse
38	musgpd2	mouse	38	muscyp345	mouse
39	muspim	mouse	39	musifna	mouse
40	muserfv41	mouse	40	musifna2m	mouse
41	musgpd3	mouse	41	musil3b	mouse
42	musgpd4	mouse	42	musasp	mouse
43	musmhabz2	mouse	43	musmhtp22	mouse
44	musuvm	mouse	44	musmhtlac1	mouse
45	ratthy	rat	45	musmhtlac2	mouse
46	ratpth	rat	46	muspim	mouse
47	ratctrpb	rat	47	musrpoii1	mouse
48	ratthy1g	rat	48	musrpoii2	mouse
49	ratrsbz1	rat	49	musug6pa	mouse
50	ratcypoxg	rat	50	ratgh2	rat
51	ratmt1pa	rat	51	ratgh1	rat
52	ratcyp45c	rat	52	ratmt1pa	rat
53	ratmt12c	rat	53	ratmhc2	rat
54	ratelaiii	rat	54	ratctrpb	rat
55	ratrhl1	rat	55	m12894	rat
56	ratmyl2g	rat	56	ratmt12c	rat
57	ratelai3	rat	57	ratmt1pb	rat
58	hamrsa49d	hamster	58	ratptry24	rat
59	hamrsa63	hamster	59	ratctrpb	rat
60	hamrsa49b	hamster	60	ratmt1pc	rat
61	hamrsa34	hamster	61	ratwap1	rat
62	hamdes1	hamster	62	ratsv40jn	rat
63	hamhmg	hamster	63	ratalac	rat

Table 1. Continued

B1 sequences			B2 sequences		
Number	Name	Species	Number	Name	Species
			64	ratrgb23	rat
			65	ratmt1pb	rat
			66	ratrs1b11	rat
			67	ratrsrg31	rat
			68	hamrsa250	hamster
			69	hamrsa49c	hamster
			70	hamprp2	hamster
			71	hamprp2-2	hamster

The GenBank/EMBL data base names of the sequences from which B1 and B2 sequences were extracted are given here. Where only some of the sequences present in a genomic sequence entry are used, those most similar to the "search" sequences are the ones used in this study

lyzed by MULTAN (Bains 1986a) as previously described (Bains 1986b). Sequence comparisons between elements included the poly-dA tail, which is believed to be part of the integrated reverse transcript, but not the poly-dA nose, which is believed to be a target sequence feature.

Results

B2 Sequences

The consensus sequence for 71 B2 sequences is shown in Fig. 1A. Shown is the consensus sequence (central line) and the degree to which the data sequences adhere to it ["Adherence" (Bains 1986a)].

The sequence can be divided into three regions on the basis of the Adherence. The 5′ end shows a "poly-dA nose" similar to that shown 5′ of the human *Alu* family consensus (Bains 1986b), which is of low Adherence. The 3′ end shows an A-rich "poly-dA tail" common to all retroposons (Rogers 1985). In between, the element itself has a high Adherence of between 80% and 100%, with occasional bases showing more variation.

The poly-dA nose has been suggested to be the result of an integration site preference for the human *Alu* element (Daniels and Deininger 1985; Bains 1986b), and hence to lie outside the *Alu* element proper. Supporting a similar interpretation of this data, the start of the B2 element as defined by the position of the direct repeats flanking each element is approximately at the base numbered 1 in Fig. 1A (Rogers 1985, Fig. 6). The poly-dA tail lies inside the B2 element as defined by direct repeats.

In those regions of the consensus that are not A-rich, several bases show Adherence substantially less than 80%. These are all isolated bases or dinucleotides: no long runs of sequence different from the consensus are shared between more than two sequences. Nine of these variable bases are members of CG, CA, TG, or TA dinucleotides. These are expected to mutate rapidly due to both the rapid mutation of CG to TG and CA through deamination of methylated cytosines and because of the inherently greater relative mutation rates of pyrimidine-purine dinucleotides over other dinucleotides (Bains and Bains 1987). Thus, these variations could represent rapid mutation at these sites. However, five bases, labeled "a" to "e" in Fig. 1A, are not members of these rapidly mutating dinucleotides but nevertheless show greater than usual variability among family members. Two other variations are also seen very frequently: the base 34 (labeled "f") and bases 23 to 34 (region labeled "g") are frequently deleted. It is notable that the labeled bases also show considerable variation in the analysis of Rogers (1985), and two of the four positions where the consensus in Fig. 1A differs from his are at bases "c" (A instead of G here) and "d" (A instead of T). These common variations could represent other mutation hotspots, in which case they would be expected to occur randomly in different family members: the presence of a variant base in one position in a sequence would not bias the probability of occurrence of a variant base at another position. Alternatively, such a pattern would be expected if the B2 family consisted of two or more subfamilies, distinguished by base differences at these sites whose members were present in similar numbers in our data base. Below we consider the evidence for this second hypothesis.

B2 Subfamily Structure

Figure 2 shows a dendrogram of the B2 sequences analyzed here. The dendrogram was calculated by MULTAN according to the Farris (1972) algorithm. There is no clear subfamily structure in the main section of the diagram: the relatively small size of the elements and the number analyzed make it likely that some pairs of elements will be more similar than others by chance (Tajima 1983). In particular, the topology of links joining the upper central re-

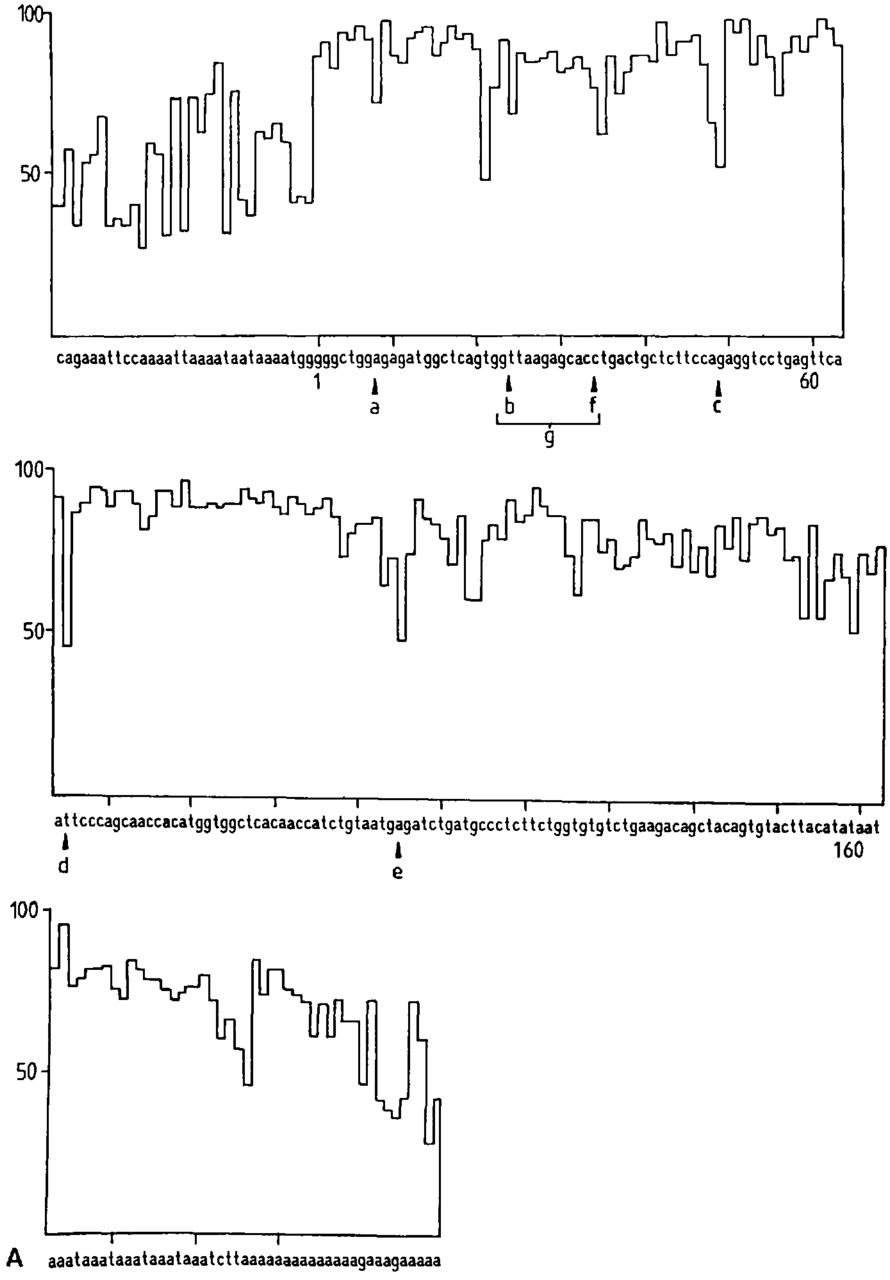

Fig. 1. Consensus sequences for B1 and B2 elements. **Central line** Consensus sequence. **Upper line** Plot of Adherence of the data sequences to this consensus. **A** B2 sequences. Seven points of high variability that are not part of CG dinucleotides, labeled a–g, are referred to in the text and listed in Table 2. **B** B1 sequences.

gions, the left quadrant, and the mid-right region is determined by several very short branch lengths in the center of the diagram, and only a few base changes altering these branch lengths could alter this topology (Farris 1972; Bains 1986a). This in itself suggests that the topology is not significant, and that most of the B2 sequences may be considered to be rooted on the tree to a common point near the consensus. The exception is a potential family of elements defined by their unusual divergence from the consensus and their much smaller intragroup divergence (elements 23, 24, 26, 37, 39, 40, 41, 58), seen in the lower right of the figure. The links between sequences 13 or 16 and the node joining sequence 37 to the tree would have to be deleted to abolish this subfamily. Thus, this is a significant clustering detected by dendrogram analysis. However, these elements cannot be responsible for the presence of variable positions in the consensus, as they are too few substantially to affect the Adherence statistics of the whole database (8/71 sequences).

Figure 3A shows a dendrogram of the B2 sequences in which only the seven variable positions labeled "a" to "g" in Fig. 1A have been used to calculate sequence difference. This, then, is a diagram that should reveal any subfamily relationships between the B2 elements that generate the variable positions. Many of the B2 elements are identical with respect to their sequence at loci "a" to "g." In addition, the dendrogram falls into two sections, with three sequences (31, 48, and 63) falling between them, with substantial branch lengths separating the

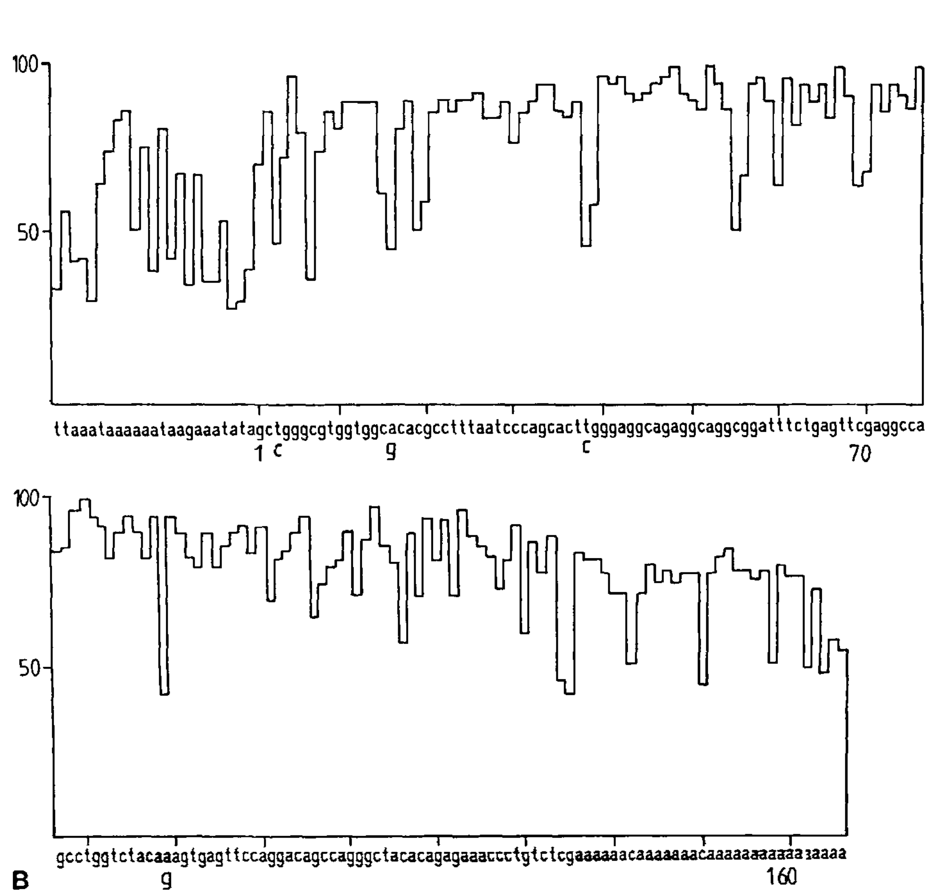

Fig. 1. Continued

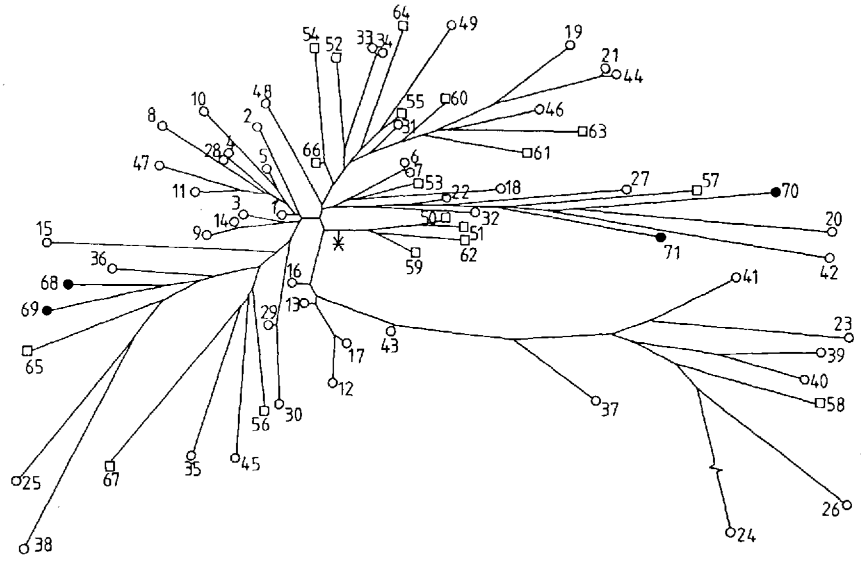

Fig. 2. Unrooted Wagner network (dendrogram) of B2 sequences. The total length of the lines between any two sequences is proportional to the difference between them. The spacial arrangement is arbitrary. ○ = mouse sequences, □ = rat sequences, ● = hamster sequences. The line leading to sequence 24, interrupted by a Z, has been reduced to half its correct length for convenience.

sections (compare Fig. 2). The sections are separated by dotted lines in Fig. 3A. (Not all sequences used in Fig. 2 are used in Fig. 3A, as some are too partial to generate significant comparisons over the small number of points considered.) To show that these groupings could represent real subfamilies, they have been flagged on the dendrogram in Fig. 2. If the variable regions were mutational hot spots, then, regardless of their relationship to each other, they would be expected to be distributed randomly among the elements in Fig. 2, as 5 base substitutions out of 190 bases is insufficient difference to substantially alter a sequence's overall difference from another sequence (deletions are not included in the difference calculation generating Figs. 2, 3B, and 4). The resulting flagged dendrogram is shown in Fig. 3B: vertical arrowheads flag sequences from the left of Fig. 3A, horizontal arrowheads those from the right. The elements identified on the basis of their position in Fig. 3A are highly nonrandomly distributed among

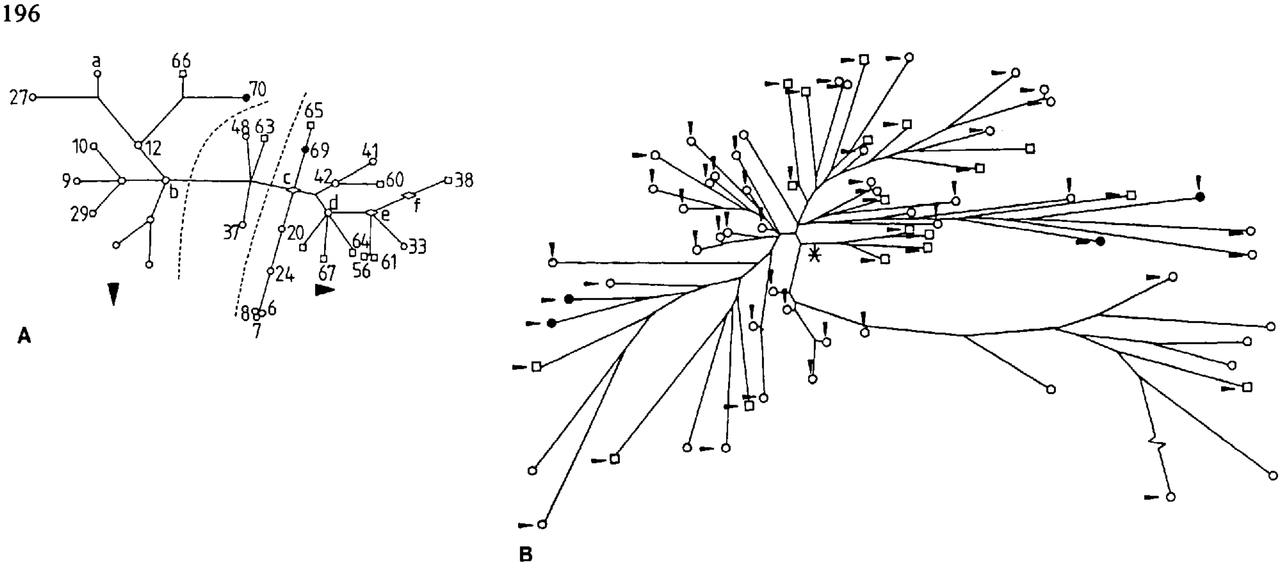

Fig. 3. Detection of subgroups of B1 sequences. **A** Dendrogram of sequences showing the difference between the seven loci listed in Table 2. The detection in location "g" is regarded as a single change for the construction of this diagram. *Abbreviations:* A = sequences 13, 14, 17, 18, 43; B = sequences 1, 4, 5, 15, 47; C = sequences 36, 51, 52, 54, 58; D = sequences 21, 45, 46, 49; E = sequences 19, 30, 50, 53, 62, 68; F = sequences 37, 57, 71. Symbols as for Fig. 2, plus ◇ = sequences from more than one species at this node. The dendrogram is separated into two major regions by the dashed lines. **B** Position of subgroup sequences from 3A on the dendrogram from Fig. 2. Vertical arrowheads: sequences from left of 1A. Horizontal arrowheads: sequences from right of 3A. Other symbols as for Fig. 2.

Table 2. Variable sites in B2 sequences

Code	Location	Consensus	Alternative	No. consensus	No. alternative
a	8	A	T	40	12
b	24	T	G	41	15
c	49	G	A	34	25
d	65	T	A	29	26
e	105	A	G	31	27
f	34	C	Δ	21	25
g	23–34	GTTAAGAGCACC	Δ	42	6

Column 1: letter code for location in Fig. 1A. Column 2: location of site. Column 3: base(s) in consensus. Column 4: common alternative base. Column 5: number of sequences showing the consensus base. Column 6: number of sequences showing the alternative base. Note that figures in columns 4 and 5 do not sum to 71 as not all sequences are full length. Δ = sequence deleted in alternative version. Note that for site "f" the alternative is actually more numerous than the consensus base. This arises because MULTAN places a base in the consensus unless more than 75% of the data sequences have a gap at that position (Bains 1986a)

the regions of the Fig. 2 dendrogram. (Their scattering between the "quadrants" of the figure is probably not significant because of the tight clustering of branch points in the center of Fig. 2 mentioned above.)

Positions a, b, c, d, and f are identified by Rogers (1985) as being those that are different between his subfamilies I and II: I has the same sequence at each of these five positions as the consensus in Fig. 1. However, subfamily consensus I is not identical with our consensus, and neither does subfamily II possess the alternative bases at all positions a–g listed in Table 2. Neither Rogers's subfamilies nor the alternatives identified here are the same as the subfamilies of Deininger and Daniels (1986). Thus, the method of defining a subfamily appears to affect which sequences are incorporated in the subfamily, and hence its consensus. We might note that the deletion of segment "g" is found in only one of Rogers's data set sequences, a subfamily I sequence.

Distribution of Variability

Table 2 shows the alternative bases that are found at each of the loci "a" to "g," and their relative frequency in the data sequences. The chance that the most likely combination of variants—that of the consensus—should actually occur in one sequence is 4.6%, and so 2.8 of the 62 B2 sequences complete enough to be shown in Fig. 3A should have this combination. In fact, six clusters of more than three sequences occur in Fig. 3A, labeled "A" to "E" and containing, respectively, five, five, five, four, and six sequences: Poisson probabilities of four-, five-, and six-sequence clusters are 0.169, 0.095, and 0.044, respectively. Thus, Fig. 3A shows far more clusters of identical sequences than would be expected from

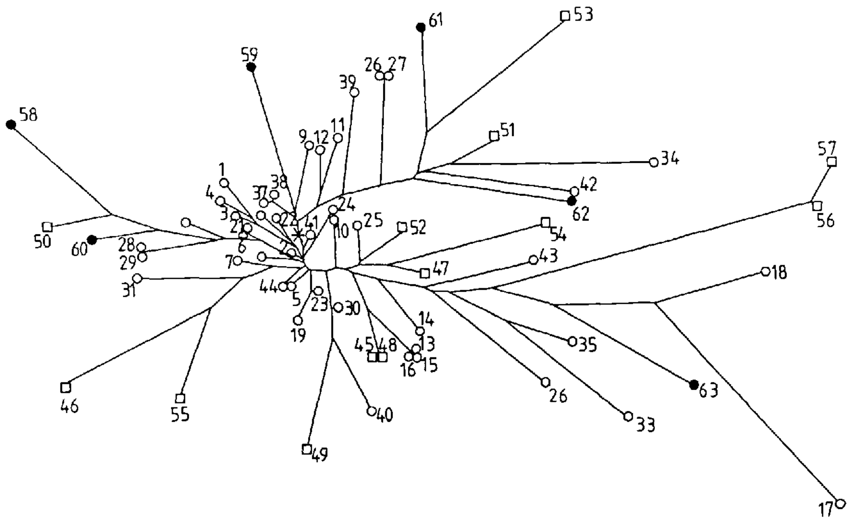

Fig. 4. Dendrogram of B1 sequences. Symbols as for Fig. 2.

random distribution. These lines of evidence suggest that the major portion of the B2 family is divided into at least two subfamilies.

Other Variations in the Consensus

The consensus shown in Fig. 1A also differs from that of Rogers (1985) (subfamily I) at position 155 (ACA here, ATA in Rogers) and at the 3′ end, where Rogers finds three repeats of the TAAA motif where we find five. This latter difference may be due to expansion of this short tandem repeat by unequal crossover, as occurs in "minisatellite" DNA in mammalian genomes (Jeffreys et al. 1985).

B1 Sequences

Figure 1B shows the consensus for the 63 B1 elements analyzed. This consensus also shows a poly-dA nose that probably lies outside the integron itself (Kalb et al. 1983) and hence that probably represents a target site specificity. The consensus shown in Fig. 1B differs in four positions from that of Kalb et al. (1983): these are indicated below the consensus sequence. In only one site (base 89, G in Kalb et al., A here) do these variations fail to fall in CG, TG, or CA pairs. Thus, unlike the case with the B2 elements, the Adherence of the B1 elements to the consensus gives no clear evidence that the data used here contains subfamilies. The dendrogram of the sequences is shown in Fig. 4; this also shows that no subfamilies can be detected in these data. The possibility that minor subfamilies that are similar to the consensus derived here exist remains untested by these approaches.

Species Distribution of Sequences

A notable feature of Figs. 2 and 4 is that the B1 and B2 sequences from different species are not limited to specific parts of the diagram. Indeed, the sequences from the three rodents analyzed here appeared to be scattered throughout the diagram, with only some clustering of murine sequences observable near the centers of both Fig. 2 and Fig. 4. To test whether cospecific sequences were more similar to each other than to other sequences, we compared them in all species–group combinations (Table 3). From this it is obvious that the "average" rat B1 sequence is actually more closely related to the "average" mouse B1 sequence than to other rat sequences, although the differences are small in all cases. Clearly the mouse B1 sequences are more similar to the consensus than the hamster sequences, which are themselves more similar than the rat ones. For the B2 sequences, the rat sequences are most similar to the consensus, with mouse being intermediate between them and hamster. Artifacts of sequence selection could account for these differences: the rat sequences originate purely from fortuitous sequencing, the mouse ones are a combination of fortuitous sequencing and selective sequencing of B1 and B2 elements, and the hamster sequences are largely derived from work aimed at sequencing repetitive DNA. The effects seen are small, and so it is also questionable whether they are statistically significant even if they correctly represent all genomic B1 and B2 sequences. However, it is also possible that different regions of the genome contain repetitive DNAs originating at different times, and that different species contain families having differing divergence patterns.

Discussion

We have presented an analysis of 63 B1 and 71 B2 sequences collected from published data base sequences. Several points arise from the properties of these sequences.

Table 3. Mean differences between B1 and B2 elements from different species

	Mean fractional difference (SD)		
	Mouse	Rat	Hamster
		B1	
Consensus	0.2435 (0.0958)	0.3092 (0.1116)	0.2880 (0.0928)
Mouse	0.3323 (0.1132)	0.3941 (0.1171)	0.3800 (0.1175)
Rat		0.4509 (0.1325)	0.4210 (0.1478)
Hamster			0.4171 (0.0781)
		B2	
Consensus	0.2638 (0.1158)	0.2251 (0.0917)	0.3928 (0.1027)
Mouse	0.3742 (0.1437)	0.3677 (0.1289)	0.4609 (0.1080)
Rat		0.3189 (0.1166)	0.4547 (0.1098)
Hamster			0.5081 (0.1225)

Shown are the mean fractional difference figures (standard deviation in parentheses) between all the elements of rat, mouse, and hamster, and between the elements and the overall consensus

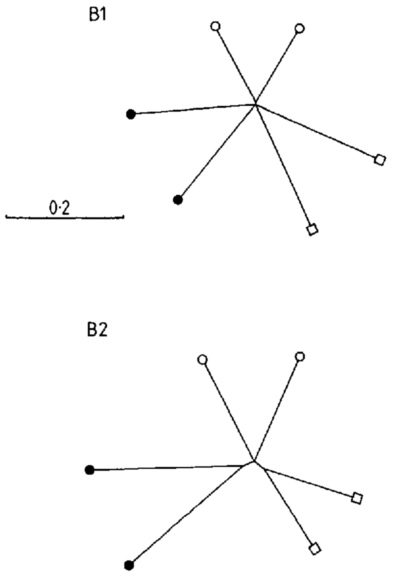

Fig. 5. Dendrogram of hypothetical "typical" sequences from mouse, rat, and hamster. Symbols as for Fig. 2. The distances between terminal symbols for each species are proportional to the mean difference between B1 and B2 sequences in those species. The terminal symbols do not represent actual sequences.

Subfamily Structure

The B2 elements show strong evidence for falling into at least three subfamilies: the two detected by the high variability shown by the variable positions "a" to "g," and the third, highly diverged group of eight elements appearing in the lower right sector of Fig. 2. That the B2 family falls into subfamilies is not a novel observation (Deininger and Daniels 1986), but we present this evidence at some length to show that this method of Adherence analysis may be used to demonstrate the existence of such subfamily structure even when the subfamilies are very similar, and hence indistinguishable on the basis of dendrogram analysis. A similar analysis of B1 elements (Fig. 4) and of human *Alu* sequences (Bains 1986b, Fig. 1) shows only one such variable position that is not a member of a CG dinucleotide in each case, and, although the B1 and *Alu* sequences could be classified according to the base at this point, this is insufficient grounds for suggesting that the two groups resulting are "subfamilies." This suggests that these two families of sequences are not subdivided into families whose genomic frequencies are similar, although they may be divided into subfamilies whose frequency ratio is greater than the 80% : 20% limit at which variation would cease to be obvious in Fig. 1B.

Figure 3 also suggests that the B2 subfamilies are present in different ratios in different species. While this may be true, we do not believe that these data support such a view, because of the potential for artifacts of selection of sequences. Two selection steps—selection of DNA segments for molecular cloning and sequencing and selection of data base sequences for analysis—could affect species representation of different subfamilies, and while we have tried to reduce the latter to a minimum the former certainly occurs as described above. However, species-specific or family-specific sequence biases may exist: this possibility is examined below.

Species Specificity of Sequences

It is commonly stated that, in such families of repetitive DNAs, elements within a species are more closely related to other elements in the species than to elements of the same family from different species (Dover 1982). This is a consequence of the homogenization of sequences through transposition and gene conversion, a process termed "concerted evolution." These data allow us to estimate the amount of such homogenization from the number of pairs of extremely similar sequences (which are candidates for the products of recent homogenization events) and from the difference between sequences in different species. Figures 2 and 4 show few identical or nearly identical pairs, suggesting a small amount of homogenization: quantitative measures of homogenization rates could be derived from such data using theoretical models derived by Ohta and Dover (1983, 1984). The degree to which elements in one species are distinct from those in another is examined in Table 3; while it is clear that the highly repetitive families of man and rodents are distinct, we find no evidence for a similarly clear distinction between highly repetitive sequences in rodents.

Mouse B1 sequences in this data set are more closely related to other mouse B1 sequences than to rat sequences solely because they are more closely related to the consensus, and hence rat B1 sequences are also more closely related to mouse sequences than to rat sequences.

Sequences from one species are slightly more closely related to each other than to sequences from other species. Seven of 13 rat B1 sequences and 8 of 18 rat B2 sequences are more closely related to a rat sequence than to a mouse sequence, a slightly higher ratio than would be expected from the ratios of rat : mouse sequences among those analyzed (rat : mouse = 13:44 for B1, 18:49 for B2). A dendrogram of pairs of "typical" rat and hamster sequences (i.e., sequences that give the mutual species divergence figures shown in Table 3) shows that the individual species' B1 and B2 elements are slightly more closely related to each other than to other species' elements (Fig. 5). However, there is no sharp distinction between mouse and rat sequences, or even between hamster and nonhamster sequences: it is not possible to identify an element's species of origin from its sequence. Figure 3 shows that the B2 subfamilies are not confined to one species, and hence are not species-specific variants. One subfamily is more numerous in mouse than in rat, but is not exclusive to it, nor is the only B2 family in mouse. Thus, we must conclude that species-specific concerted evolution of these sequences has not occurred in the last 10 million years. If these elements are evolving "in concert" (Dover 1982), the concert is remarkable for its dissonance.

Acknowledgments. We thank the Imperial Cancer Research Fund for the generous availability of their computers and software. W.B. is the holder of a grant from the Muscular Dystrophy Group of the UK.

References

Bains W (1986a) MULTAN: a program to align multiple DNA sequences. Nucleic Acids Res 14:159–177

Bains W (1986b) The multiple origins of human Alu sequences. J Mol Evol 23:189–199

Bains W, Bains J (1987) Rate of base substitution in mammalian nuclear DNA is dependent on local sequence context. Mutat Res 179:65–74

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Wayne PR, Swindell CD, Tung C-S (1986) The Genbank genetic sequence database. Nucleic Acids Res 14:1–4

Daniels GR, Deininger PL (1985) Integration site preferences of the Alu family and similar repetitive sequences. Nucleic Acids Res 13:8939–8954

Davidson EH, Graham DE, Neufeld BR, Chamberlin ME, Amenson CS, Hough BR, Britten RJ (1973) Arrangement and characterisation of repetitive sequence elements in animal DNAs. Cold Spring Harbor Symp Quant Biol 38:295–301

Deininger PL, Daniels GR (1986) The recent evolution of mammalian repetitive DNA elements. Trends Genet March: 76–80

Dover GA (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Farris (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

Hamm GH, Cameron GN (1986) The EMBL data library. Nucleic Acids Res 14:5–9

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73

Kalb VF, Glasser S, King D, Lingrel JB (1983) A cluster of repetitive elements within a 700 base pair region in the mouse genome. Nucleic Acids Res 11:2177–2184

Ohta T, Dover GA (1983) Population genetics of multigene families that are dispersed into two or more chromosomes. Proc Natl Acad Sci USA 80:4079–4083

Ohta T, Dover GA (1984) The cohesive population genetics of molecular drive. Genetics 108:501–521

Rogers J (1985) Retroposons. Int Rev Cytol 93:187–279

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, Oxford, pp 323–358

Tajima F (1983) Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460

Received December 22, 1987/Revised February 28, 1988