Microsatellites in Different Eukaryotic Genomes: Survey and Analysis

Gábor Tóth,[1,3] Zoltán Gáspári,[1] and Jerzy Jurka[2]

[1]Department of Genetics, Eötvös Loránd University, Budapest, Hungary; [2]Genetic Information Research Institute, Sunnyvale, California 94089 USA

We examined the abundance of microsatellites with repeated unit lengths of 1–6 base pairs in several eukaryotic taxonomic groups: primates, rodents, other mammals, nonmammalian vertebrates, arthropods, *Caenorhabditis elegans*, plants, yeast, and other fungi. Distribution of simple sequence repeats was compared between exons, introns, and intergenic regions. Tri- and hexanucleotide repeats prevail in protein-coding exons of all taxa, whereas the dependence of repeat abundance on the length of the repeated unit shows a very different pattern as well as taxon-specific variation in intergenic regions and introns. Although it is known that coding and noncoding regions differ significantly in their microsatellite distribution, in addition we could demonstrate characteristic differences between intergenic regions and introns. We observed striking relative abundance of $(CCG)_n$•$(CGG)_n$ trinucleotide repeats in intergenic regions of all vertebrates, in contrast to the almost complete lack of this motif from introns. Taxon-specific variation could also be detected in the frequency distributions of simple sequence motifs. Our results suggest that strand-slippage theories alone are insufficient to explain microsatellite distribution in the genome as a whole. Other possible factors contributing to the observed divergence are discussed.

Microsatellites or simple sequence repeats (SSRs) are tandemly repeated tracts of DNA composed of 1–6 base pair (bp) long units. They are ubiquitous in prokaryotes and eukaryotes, present even in the smallest bacterial genomes (Field and Wills 1996; Hancock 1996a). A subset of SSRs, namely trinucleotide repeats, are of great interest because of the role they play in many human neurodegenerative disorders (fragile X syndrome, Huntington's disease, myotonic dystrophy, spinal-bulbar muscular atrophy, spinocerebellar ataxia, etc.; for reviews, see Warren and Nelson 1993; Bates and Lehrach 1994; Reddy and Housman 1997) and in some human cancers, e.g. hereditary nonpolyposis colorectal carcinoma (Wooster et al. 1994; Arzimanoglou et al. 1998). The alteration responsible for these genetic diseases is the expansion of triplet repeats, where the rate of mutation depends on the number of tandem units within the repeat. Hence the term 'dynamic mutation' was coined by Richards and Sutherland (1992).

Microsatellites can be found anywhere in the genome, both in protein-coding and noncoding regions. Because of their high mutability, microsatellites are thought to play a significant role in genome evolution by creating and maintaining quantitative genetic variation (Tautz et al. 1986; Kashi et al. 1997). In promoter regions, the length of SSRs may influence transcriptional activity (Kashi et al. 1997). Length of polyglutamine or polyproline tracts encoded by SSRs may affect protein–protein interactions involving transcription factors (Gerber et al. 1994; Perutz et al. 1994).

It has been shown that SSRs in exons are less abundant than in noncoding regions (Hancock 1995), and that different taxa exhibit different preferences for SSR types (Beckmann and Weber 1992; Lagercrantz et al. 1993; Tautz and Schlötterer 1994). Moreover, the overall microsatellite content in the genome correlates with the genome size of the organisms (Hancock 1996b).

SSRs are inherently unstable. Two models have been proposed to explain microsatellite generation and instability: DNA polymerase slippage and unequal recombination. The first model involves transient dissociation of the replicating DNA strands, followed by misaligned reassociation (Richards and Sutherland 1994). The slipped structure may be stabilized by hairpin, triplex, or quadruplex arrangement of DNA strands (for review, see Pearson and Sinden 1998; Sinden 1999). Thus, it is expected that those repeats that are able to form such alternative DNA conformations would be generated more frequently than others. The possible structures of triplet repeats involved in human diseases have been studied extensively. The repeats that show a considerable potential to form alternative structures include $(CTG)_n$•$(CAG)_n$, $(CCG)_n$•$(CGG)_n$, $(GAA)_n$•$(TTC)_n$, $(AGG)_n$•$(CCT)_n$, and $(TGG)_n$•$(CCA)_n$ (Gacy et al. 1995; Bidichandani et al. 1998; Usdin 1998). However, some sequences with theoretically high hairpin-forming potential [e.g. $(CCG)_n$] show the slowest in vitro slippage rate (Schlötterer and Tautz 1992). Moreover, the rate of alterations is likely to be

[3]Corresponding author.
E-MAIL tothg@ludens.elte.hu; FAX (+36-1) 266-2694.

controlled at multiple steps in vivo. An active role of the DNA mismatch repair system to stabilize simple sequence repeats has been revealed in *Escherichia coli*, yeast, and humans (for review, see Sia et al. 1997). Although a number of experimental results argue in favor of the above model, homologous recombination may also result in genetic instability of certain SSRs (Jakupciak and Wells 1999).

We can expect that the fixation of de novo-generated SSRs is determined by the interplay of several factors, of which the repeat type, the genomic position of the SSR, and the genetic-biochemical background of the cell are the most important. In our study we addressed the questions of whether the abundance of various microsatellite types is similar or not in different taxonomic groups and how SSR frequencies differ in exons, introns, and intergenic regions. We intended to give a detailed picture analyzing all possible (501) SSR motifs to complement the results of a previous study on primate DNA sequence data (Jurka and Pethiyagoda 1995), and place them into comparative evolutionary perspective.

RESULTS

We examined the distribution of perfect SSRs over 12-bp long, so if not explicitly stated otherwise, our results described here apply to microsatellites meeting this criterion. To assess expandability of the repeats, we also analyzed perfect repeats longer than 24 bp (see Methods) and compared the results to those obtained using the shorter cutoff length. Data presented below always refer to duplex DNA, even if we show only the sequence of the repeated motif on one strand for simplicity, i.e. notations like AC and $(AC)_n \bullet (GT)_n$ are equivalent.

The nonoverlapping groups of DNA sequences used in this study will be referred to as taxonomic groups or taxa. These groups represent either individual species (*Caenorhabditis elegans* and *Saccharomyces cerevisiae*), or groups of related species such as Primates, Rodentia, and Mammalia. Thus our taxa are defined rather arbitrarily based primarily on sequence availability (see Methods). We carried out the analyses on sequences classified into three genomic regions (intergenic regions, introns, and exons), and on a superset referred to as all sequences. The latter contained all sequence entries that passed the filtering criteria described in Methods, even if they could not be assigned to genomic regions.

To estimate database bias caused by the use of GenBank, we also included the full sequence of the human chromosome 22 in our study. The results obtained for chromosome 22 are in good agreement with those for all primate sequences, confirming the validity of our approach. The 30% increase in total microsatellite content in the full chromosomal sequence (see the last column of Table 1) is mostly due to greater abundance of (A + T)-rich repeats, especially poly(A/T) tracts (Tables 2 and 6).

To assess the contribution of repeated unit length to microsatellite abundance, we calculated the total lengths of all mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats per megabase pair (Mbp) of DNA sequence (Table 1). In exons, trinucleotide repeats are invariably the most abundant in all taxa, with hexanucleotide repeats being the second most common. Intergenic regions and introns, however, contain more hexanucleotide repeats than exons do, Embryophyta and *S. cerevisiae* introns being the only exceptions to this rule.

In primates, mononucleotide repeats are the most copious. In introns and intergenic regions they are more than twice as frequent as di- and tetranucleotide repeats. The latter are of similar abundance, and interestingly, much more frequent than trinucleotide repeats. In rodents, repeats with dinucleotide units are about three times more frequent than those with mononucleotides. Dinucleotide repeats are dominant in introns and intergenic regions of many other taxa, except for Primates, Embryophyta, *S. cerevisiae*, and Fungi. In rodent introns and intergenic regions, the rarity of triplet repeats is also quite pronounced in comparison to di- and tetranucleotide repeats.

The relative abundance of tetranucleotide over trinucleotide repeats in introns and intergenic regions is characteristic of all vertebrate taxa but not of any other taxonomic group studied. In all mammalian taxa, even pentanucleotide repeats are more frequent in introns and intergenic regions than triplet repeats. In invertebrates and fungi, tetranucleotide repeats constitute the less frequent class of microsatellite in introns and intergenic regions, whereas in vascular plants they are comparably rare as hexanucleotide repeats.

When comparing various taxonomic groups, it is evident that rodents adopt much more microsatellites than any other group we examined. *C. elegans*, however, contains the least SSRs per one Mbp of DNA, less than *S. cerevisiae* and other fungi.

A more detailed picture could be drawn when we analyzed the distribution of SSRs by the sequence of the repeated motif. Results obtained for mono-, di-, and trinucleotide repeats are shown in Tables 2–5. The most frequent tetra-, penta-, and hexanucleotide repeats are listed in Tables 6–9. More data are available online in our SSRDB database at http://genetics.elte.hu/ssr.

Mononucleotide Repeats

In general, poly(A/T) tracts are more abundant in each taxon than poly(C/G) sequences (Tables 2–5). This difference is the least characteristic in *C. elegans* and most

Table 1. Total Lengths[a] of Simple Sequence Repeats by Repeated Unit Length

Taxonomic group	Genomic region	Length of repeated motif (bp) 1	2	3	4	5	6	Total
Primates								
	all	3429	1643	477	1368	898	341	8156
	intergenic regions	3880	1709	517	1464	991	385	8946
	introns	4137	1506	424	1428	988	392	8875
	exons	49	10	1126	29	57	244	1515
Human chromosome 22								
	all	5141	1511	604	1906	1097	419	10678
Rodentia								
	all	1839	5461	1196	2942	1417	1034	13889
	intergenic regions	2192	5928	1230	2823	1577	740	14490
	introns	2182	5837	1123	3009	1399	922	14472
	exons	62	70	1557	63	116	620	2488
Mammalia								
	all	1397	2312	532	915	774	693	6623
	intergenic regions	1954	4666	531	1529	1115	1155	10950
	introns	1967	2202	395	792	685	637	6678
	exons	69	88	876	19	18	356	1426
Vertebrata								
	all	1418	2449	1069	1279	709	220	7144
	intergenic regions	2193	3363	1127	1766	1201	320	9970
	introns	1476	3193	861	1502	585	142	7759
	exons	49	0	823	0	26	75	973
Arthropoda								
	all	985	1403	956	439	732	875	5390
	intergenic regions	1462	2259	1128	621	1110	1090	7670
	introns	950	1627	728	461	735	917	5418
	exons	12	34	1566	0	21	591	2224
C. elegans								
	all	428	556	337	144	225	449	2139
	intergenic regions	573	822	414	198	310	574	2891
	introns	512	549	228	169	283	556	2297
	exons	43	54	308	18	38	116	577
Embryophyta								
	all	1245	1067	880	184	491	272	4139
	intergenic regions	2012	1715	869	303	781	334	6014
	introns	1380	1322	576	260	547	207	4292
	exons	18	50	1119	2	29	303	1521
S. cerevisiae								
	all	1075	580	646	93	204	406	3004
	intergenic regions	3140	1875	512	273	494	532	6826
	introns	3012	1437	516	162	509	288	5924
	exons	36	19	706	7	52	330	1150
Fungi								
	all	905	272	485	194	395	426	2677
	intergenic regions	2080	555	550	421	925	548	5079
	introns	2075	1013	951	458	659	661	5817
	exons	9	4	381	2	35	219	650

[a]Base pairs (bp) per megabase of DNA.

pronounced in primates. The total length of mononucleotide repeats, taking together both patterns, is also greatest in primates (Table 1). Nonmammalian vertebrates show the second highest ratio of poly(C/G)

Table 2. Total Lengths of Mono-, Di-, and Trinucleotide Repeats in All Sequences[a]

Repeated unit	Taxonomic group									
	primates	human chr 22[b]	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
A	3418	5126	1634	1291	1051	875	227	1221	1069	872
C	11	15	205	106	367	110	201	24	6	33
AC	1033	981	3468	1333	1496	825	163	85	67	81
AG	293	261	1619	854	334	235	222	333	19	58
AT	314	266	354	109	616	340	170	648	494	133
CG	3	3	20	16	3	3	1	1	–	–
AAC	110	128	149	49	66	189	19	139	125	108
AAG	42	55	217	19	36	17	105	332	119	59
AAT	161	211	122	30	350	100	70	108	176	119
ACC	25	56	136	17	54	58	28	42	6	36
ACG	0	1	1	15	3	26	12	11	28	12
ACT	6	2	8	4	21	16	11	15	17	15
AGC	32	32	222	201	166	387	23	38	80	55
AGG	42	60	226	97	182	58	15	52	13	31
ATC	26	35	34	4	93	86	48	119	82	32
CCG	33	24	81	96	98	19	6	24	–	18

[a]Base pairs (bp) per megabase of DNA.
[b]Human chromosome 22 sequence.

to poly(A/T). Besides *C. elegans*, they constitute the only group where poly(C/G) repeats appear in exons in a proportion comparable to poly(A/T) (Table 5). Intergenic regions show an interesting preference for poly(C/G) over poly(A/T) in *C. elegans* (Table 3). Introns contain more poly(A/T) than poly(C/G) repeats in each taxon (Table 4).

Dinucleotide Repeats

Dinucleotide repeats are most abundant in rodents and the least frequent in fungi (Table 1). Characteristic differences between taxa can only be observed for intergenic regions and introns (Tables 3 and 4) because of the rarity of dinucleotide repeats in exons (Table 5). Curiously, we have found one 16-bp long CG repeat in

Table 3. Total Lengths of Mono-, Di-, and Trinucleotide Repeats in Intergenic Regions[a]

Repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
A	3864	1956	1868	1612	1333	252	1975	3121	2024
C	16	236	86	581	129	321	37	19	56
AC	1049	3733	2654	2070	1159	249	115	266	140
AG	327	1799	1766	294	437	353	521	57	136
AT	329	363	246	999	663	219	1077	1552	279
CG	4	33	–	–	–	1	2	–	–
AAC	115	156	28	83	437	16	106	23	104
AAG	40	280	31	50	36	143	377	35	84
AAT	188	168	77	497	196	98	198	321	162
ACC	23	136	–	45	47	23	21	–	60
ACG	–	–	–	–	32	14	10	9	6
ACT	4	3	13	8	10	16	14	36	30
AGC	25	154	236	84	255	18	14	39	19
AGG	45	204	56	148	32	14	26	–	34
ATC	19	41	–	112	75	65	88	49	43
CCG	58	88	90	100	8	7	15	–	8

[a]Base pairs (bp) per megabase of DNA.

Table 4. Total Lengths of Mono-, Di-, and Trinucleotide Repeats in Introns[a]

Repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
A	4125	1893	1743	1033	851	335	1341	2994	1868
C	12	289	224	443	99	177	39	18	207
AC	1012	3782	1348	2155	749	151	168	26	511
AG	268	1741	698	412	469	173	444	26	140
AT	221	300	124	626	409	224	710	1385	362
CG	5	14	32	–	–	1	–	–	–
AAC	114	156	194	105	78	11	121	–	145
AAG	33	266	20	11	33	70	176	20	34
AAT	171	165	24	343	199	67	170	434	537
ACC	14	118	–	34	51	9	11	–	14
ACG	–	–	–	–	18	6	–	39	–
ACT	10	9	–	–	56	8	16	23	–
AGC	16	114	117	43	195	16	13	–	103
AGG	39	253	40	168	11	12	7	–	45
ATC	26	42	–	157	78	26	57	–	53
CCG	1	–	–	–	9	3	5	–	20

[a]Base pairs (bp) per megabase of DNA.

the protein-coding region of beta one adrenergic receptor gene from *Canis familiaris*. Otherwise, CG repeats are very rare.

In all vertebrates and arthropods, AC is the most frequent dinucleotide repeat motif (Tables 2–4). *C. elegans* prefers AG in intergenic regions, AT in introns. In embryophytes, yeast, and fungi, AT repeats are the most frequent in general, except for introns in fungi where AC is more abundant (Table 4).

Trinucleotide Repeats

Trinucleotide repeats can be found in each genomic region with a significant frequency (Tables 2–5). However, the frequency distribution by repeat type shows

Table 5. Total Lengths of Mono-, Di-, and Trinucleotide Repeats in Exons[a]

Repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
A	49	62	69	30	12	23	17	36	9
C	–	–	–	19	–	20	1	–	–
AC	4	29	–	–	21	16	4	5	2
AG	6	24	69	–	6	29	39	–	–
AT	–	17	–	–	7	9	7	14	2
CG	–	–	19	–	–	–	–	–	–
AAC	8	–	–	–	220	33	253	156	107
AAG	57	29	22	43	–	81	317	147	55
AAT	–	–	–	–	10	19	4	123	34
ACC	58	184	22	46	142	65	89	12	26
ACG	5	–	65	15	21	10	20	27	12
ACT	5	–	–	–	–	6	13	5	5
AGC	381	889	376	337	954	31	108	112	67
AGG	192	143	219	210	83	14	106	21	27
ATC	26	22	18	15	58	41	138	103	22
CCG	394	290	154	157	78	8	71	–	26

[a]Base pairs (bp) per megabase of DNA.

Table 6. The Most Frequent Tetra-, Penta-, and Hexanucleotide Repeats in All Sequences[a]

Length of repeated unit	Taxonomic group									
	primates	human chr 22[b]	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
4	AAAT (378) AAAG (225) AAAC (216)	AAAT (537) AAAG (270) ATCC (263)	AGAT (620) AAAG (397) AAAC (370) AAGG (346)	AAAG (208) AAGG (206) AAAT (197)	AGAT (372) ACAG (167) AAAT (125)	ACAT (81) AAAT (73) AAAC (34) ACTG (28) AGCC (28)	AAAT (59) ACCT (20)	AAAT (51) AAAG (31) AAAC (20)	AAAT (38) ACAT (17)	AAAT (56) AAAG (26) AAAC (16)
5	AAAAC (285) AAAAT (195)	AAAAC (339) AAAAT (257)	AAAAC (432) AGCTC (133) AAAAG (120) AAAAT (91)	AACTG (147) AGCTC (107) AAAAT (96) AAAAC (68)	AAAAT (93) AAAAC (71) CCCCG (57) AAGGG (47) AGAGG (45) AGCGG (31) AAAAG (26)	AAAAT (48) AAAAC (36) AATAT (32) AAATT (27) AATCG (24) AACTG (23) AAACC (22) AAATG (22) AAACG (19) ACTCC (19) AACAG (17) AATCC (16) ATCCC (15) ATCCG (15) AAAGT (14) AAAAG (13) AAATC (13)	AAAAT (58) AAATT (42) AAAAC (19)	AAAAT (133) AAAAC (60) AAAAG (50) AAATT (25)	AAAAG (53) AGATG (32) AAAAC (18)	AAAAT (90) AAAAG (40) AAAAC (24) AAATT (23) AATAT (10) ACTAT (10) AAACC (8)
6	AAAAAC (99) AAAAAT (66) AAAAAG (38)	AAAAAC (123) AAAAAT (86) AAAAAG (52)	ACAGGC (171) AGAGGC (146) AAAAAC (96) ACAGAG (71) AGAGGG (59)	AGAGCG (151) ACACGC (95) AAAAAC (53) ACAGCC (50)	AACCCT (30) AAAAAG (20) AATCCC (20) AATAGT (13) AGCTCC (12) etc.[c]	ACAGAT (52) AACAGC (32) AGCAGG (19) ACATCC (18) AACTGC (16) AATGGG (15) AAATAT (14) AATCCC (14) AGCTCC (14) AAAAAT (12) etc.[c]	AAGCCT (252)	AAAAAT (34) AAAAAC (19) AAAAAG (14) etc.[c]	ACACCC (50) AACAGC (49) AAAAAG (22) AAAAAC (21) AAGATG (17) AAGAGG (13) AAAAAT (12) etc.[c]	AAAAAT (65) AACCCT (62) AAAAAG (16) AAAGAG (13) AACCAG (13) etc.[c]

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here (see also Table 1). The total length (bp) of repeats per megabase of DNA is in parenthesis. Repeats with identical total lengths are sorted alphabetically.
[b]Human chromosome 22 sequence.
[c]Hexanucleotide repeats for which total length per megabase of DNA is <12 bp are not shown. Complete lists are available at http://genetics.elte.hu/ssr.

Table 7. The Most Frequent Tetra-, Penta-, and Hexanucleotide Repeats in Intergenic Regions[a]

Length of repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
4	AAAT (444) AAAC (243) AAAG (233)	AAAC (487) AAAG (410) AAAT (380) AAGG (380)	AAGG (469) AAAG (335)	AGAT (665) ACAG (224)	ACAT (169) AAAT (82) AAAC (75)	AAAT (85) ACCT (20)	AAAT (101) AAAG (44) AAAC (30)	AAAT (118) ACAT (46)	AAAT (114) AAAG (66) AAAC (28) ATCC (21)
5	AAAAC (310) AAAAT (223)	AAAAC (484) AAAAG (125) AAGAG (120) AAAAT (104)	AACTG (257) AGCTC (171) AGGGG (111) AAAAC (77)	AAAAT (156) AAAAC (133) AAGGG (133) AGAGG (120) CCCCG (88)	AAAAT (84) AAAGT (76) AAAAC (73) AAATT (36) ACACC (35) AAATG (32) AATCC (30) ATCCG (30) AAACG (28) AAAAG (26) AATAT (26) AAACC (25) AACTG (25) ACAGC (22) ACGAT (22)	AAAAT (68) AAATT (54) AAAAC (33)	AAAAT (236) AAAAG (82) AAAAC (80)	AAAAG (116) AGATG (78) AATAT (60)	AAAAT (202) AAAAG (91) AAATT (62) AAAAC (52) ACTAT (31) AATAT (27)
6	AAAAAC (117) AAAAAT (78)	AGAGGC (130) AAAAAC (95) AGAGGG (74) ACAGAG (69) AAAAAG (34)	AGAGCG (467) ACACGC (293)	AACCCT (56) AAAAAG (37) AGCTCC (34) AAAAAC (19) AAAAAT (19)	ACAGAT (43) AGCAGG (40) AATACT (38) AATTAC (28) ACATCC (28) AGCTCC (27) AAACAG (23) AAAATT (20) AAAAAC (17) AATAGT (17) ACTCGC (17) etc.[b]	AAGCCT (265) AGGCAT (41)	AAAAAT (50) AAAAAC (32) AAAAAG (23) etc.[b]	ACACCC (119) AAAAAC (57) AAAAAT (57) AAAAAG (33)	AAAAAT (134) AAAAAC (23) AAAAAG (16) AAATAG (16) etc.[b]

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here (see also Table 1). The total length (bp) of repeats per megabase of DNA is in parenthesis. Repeats with identical total lengths are sorted alphabetically.
[b]Hexanucleotide repeats for which total length per megabase of DNA is <16 bp are not shown. Complete lists are available at http://genetics.elte.hu/ssr.

Table 8. The Most Frequent Tetra-, Penta-, and Hexanucleotide Repeats in Introns[a]

Length of repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
4	AAAT (448) AAAC (233) AAAG (214)	AAAG (510) AAGG (407) AAAC (387) AAAT (343)	AAGG (183) ATCC (162) AAAT (129)	AGAT (418) ACAG (188) AGGG (179)	AAAT (101) ACAT (99) AAAC (43)	AAAT (64) ACCT (33)	AAAT (46) AAAG (44) AAAC (36) AATT (30)	AAAT (78) AATC (42)	AAAT (119) ACCC (86) AATT (67)
5	AAAAC (347) AAAAT (173)	AAAAC (476) AAAAG (151) AGGGG (130)	AACTG (135) AAAAC (121) AAAAT (101)	AAAAT (87) AAAAC (64) AAATC (45) AAAAG (27) ACAGG (27) AATTC (23) AGCCT (23)	AAAAT (84) AAATT (72) AACCG (39) AATAT (36) ACACT (33) AAGAG (30) AAAGC (30) AAACC (27) ATCCG (27)	AAAAT (87) AAATT (60)	AAAAT (111) AAAAC (84) AAAAG (51) AAATC (28)	AAAAG (183) AAAAC (78)	AAAAT (196) AGGGG (70) AAAAG (56) AAGAG (42)
6	AAAAAC (98) AAAAAT (68) AAAAAG (42)	AGAGGC (140) AAAAAC (123) ACAGAG (80) ACCCCC (65) ACATAT (50) AAGGAG (48)	AAAAAC (137) ACAGCC (113) ACCCCC (73)	AATAGT (41) AACCCT (27) AACGGG (18)	ACTGAT (105) ACAGAT (54) AATACT (47) AAAAAC (33) AATACC (29) AAGATC (25) AATCCC (25) ACTAGG (22) AAAAAT (22) AAATAT (22) AAATTT (22) AACTCC (22) AAGTGG (22) AATGCG (22) AATTAC (22)	AAGCCT (353)	AAAAAT (29) AAAAAC (17) AAAAAG (14) AATCAC (11) AAACAC (8) AATGAT (8) AAAACC (6) AAAACT (6) AAAATT (6)	AAAAAC (94) AAAAAG (55)	AAAAAT (241) AAAAAG (67) AAAATT (39) ACAGGG (39)

[a]Only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here (see also Table 1). The total length (bp) of repeats per megabase of DNA is in parenthesis. Repeats with identical total lengths are sorted alphabetically.

Table 9. The Most Frequent Tetra-, Penta-, and Hexanucleotide Repeats in Exons[a]

Length of repeated unit	Taxonomic group								
	primates	rodentia	mammalia	vertebrata	arthropoda	*C. elegans*	embryophyta	*S. cerevisiae*	fungi
4	AAAT (11) AATC (8) AAAC (5) AATG (5)	AAGG (28) AAAC (14) CCCG (9) AACC (6) AGGC (6)	AAAC (19)	—	—	ATCC (5) AAAG (2) AAAT (2) AATG (2) ACAT (2) etc.[b]	AAAT (1) ATCG (1)	ACAT (3) AAAG (2) AATG (2)	AGGG (2)
5	AAAAC (11) AAAGG (6) AAAAG (5) AAGAG (5) AATGG (5)	AAGAG (41) AGAGG (19)	CCCGG (18)	ACGCC (13)	AATCC (6) AAAAC (5)	AAAAG (6) AAAAT (6) AAATC (3) AAATG (2) AAGAG (2)	ACCCG (5) AAGAG (5) AAAAC (3) AAAAG (3)	AAAAG (21) AAAAC (6)	AAAAG (10) AGCTC (3) AAAAT (2) AAAGG (2) AAAGT (2)
6	CCCCGG (51) AGCTCC (17) AGGGCG (15) ACGCCC (13) AAGAGG (11) ACCCGC (11) AGGCCC (11)	ACAGGC (346)	AAGGCC (108) AGCCGC (43) AGAGGC (29)	ACTGCT (30) ACCCTC (15)	AACAGC (78) AACTGG (34) ACGCCC (23) AGCGGC (23) AAGCCC (21) AAGGAC (17) AAGGAG (17) ACCTGG (17) AGCAGG (17) AGCTCC (17) AAGATG (13) AACACC (11) AATGGC (11)	ACCAGG (11) ACCTCC (8) AGCTCC (8) AAGCCT (7) AACAGC (5) AACTAC (4) AAGATG (4) AAAAAG (3) AAATGG (3) AAATTC (3) AAGAGG (3)	AAGGAG (17) AGCCTG (13) ACCATC (11) AACAGC (9) AAGATG (9) ACTGAG (9) AAAGGC (8) ACGGCG (8) AACACC (7) AAGAGG (7) ACCGCC (7) ACCTCC (7) ACGCCG (7) ACCGGC (6) AGGCGG (6) AAAACC (5) ACAGTG (5) AGCTCC (5) CCGGCG (5) AAAGCC (4)	AACAGC (61) AAGATG (19) ACGATG (18) AAGAGG (15) AAAAAG (12) AAAATC (12) AAGACG (12) AAGCTG (9) AGCCTG (8)	AACCAG (26) AACAGC (20) AAGAGG (14) AAGATG (12) ACCAGT (8) AACCGC (6) ACCTGG (6) ACTGCC (6) AGGGCG (6) AAAATG (5) AAAAAG (4)

[a]For penta- and hexanucleotide repeats, only the repeat motifs that together comprise 50% of all repeats of the particular unit length are shown here (see also Table 1). The total length (bp) of repeats per megabase of DNA is in parenthesis. Repeats with identical total lengths are sorted alphabetically.
[b]Repeats for which total length per megabase of DNA is <2 bp are not shown. Complete lists are available at http://genetics.elte.hu/ssr.

major differences in various genomic regions and among taxa. In all vertebrates, (G+C)-rich repeats dominate in exons, whereas they are less pronounced in other regions. AAC and AAG are the most frequent repeat types in Embryophyta exons and interesting relative abundance of (A+T)-rich repeats can also be observed in the exons of yeast and other fungi.

Generally there is an underrepresentation of ACG and ACT repeats in most taxa. The lack of ACG repeats is worth noting, because the triplet repeat with the same base composition (AGC=CAG) is found much more frequently in all regions. There is also a noticeable excess of AGC repeats in exons compared to introns and intergenic regions. In primates and rodents, CCG constitutes the second most frequent repeat type in exons. CCG repeats are almost totally absent from introns. ACC repeats are relatively infrequent in intergenic regions and introns, with the exception of rodents, where their occurrence exceeds that of ATC repeats.

Apart from these general trends, a relatively unique pattern of distribution can be observed for each taxon. While intergenic CCG repeats are quite significant in all vertebrates, they are underrepresented in other taxa. In sharp contrast with this, there is a lack of CCG repeats in vertebrate introns (Tables 3 and 4). Rodents have a relatively balanced distribution of most triplet repeat types in intergenic regions and introns showing generally higher frequencies than most other taxa. AAT repeats are the most abundant in the introns of primates, vertebrates, arthropods, yeast and other fungi, whereas they come out third after AGG and AAG in rodents. Interestingly, in mammalian introns, AAC turns out to be the most frequent triplet repeat.

Tetranucleotide Repeats

Exons contain almost no tetranucleotide repeats (Tables 1 and 9). Therefore, data can only be evaluated for introns and intergenic regions. The abundance of tetranucleotide repeats in vertebrate introns and intergenic regions exceeds that of trinucleotide repeats. Repeat frequency by type shows a general dependence on the base composition of the repeat unit. Repeats with <50% of G+C are generally more abundant (Tables 6–8). There are, however, a few notable exceptions, e.g. AAGG, which constitutes the second most frequent tetranucleotide repeat in mammals, and the fourth one in primates and rodents. Repeats of the type AAAB, where B denotes any base other than A, are very abundant in primates and rodents. AAAG and AAAT are also highly represented in other mammals.

Pentanucleotide Repeats

In all mammalian taxa, pentanucleotide repeats are at least as abundant as triplet repeats both in introns and intergenic regions (Table 1). They are underrepresented in exons of all taxa, whereas their frequency is comparable to that of trinucleotide repeats in introns and intergenic regions of nonmammalian genomes. In nonvertebrate taxa, they are invariably more frequent than tetranucleotide repeats. Within the whole genome, among the most common types we can always find (A+T)-rich ones, such as AAAAC in primates, rodents or AAAAT in vertebrates, arthropods, *C. elegans*, vascular plants, and fungi as dominant tract (Tables 6–9). The exclusive dominance of AAAAB type repeats is clear for primates and a bit less striking for rodents, and occurs in vascular plants and fungi. An interesting finding is that the CpG-containing CCCCG repeat is present in the top 50% of pentanucleotide repeats found in vertebrate intergenic regions.

Hexanucleotide Repeats

Hexanucleotide repeats constitute the second most frequent type after trinucleotide repeats in exons (Table 1). In introns and intergenic regions of nonvertebrate taxa, they are generally more abundant than tetranucleotide repeats, and in *C. elegans* their density also exceeds that of pentanucleotide repeats.

The repeat motifs present in exons show a great variation and are relatively (G+C)-rich (Table 9). A dominance of (A+T)-rich repeats can be observed in primate, plant, yeast, and fungal introns and intergenic regions (Tables 7 and 8). A few telomere-like repeat motifs are also found, like AACCCT in vertebrates and fungi, or AATCCC in vertebrates and arthropods. Interestingly, AACCCT repeats are present in vertebrate introns and intergenic regions. The presence of the (G+C)-rich ACCCCC motif in the top 50% of simple sequence repeats in introns of rodents and mammals is also noteworthy. Two CpG-containing repeats (AGAGCG and ACACGC) are relatively abundant in mammalian intergenic regions.

Rare Repeats

We could not find in our database subsets any of the following 27 sequence motifs in repeats longer than 12 bp: the pentanucleotide ACGCT, the hexanucleotides AAACGT, AAAGCG, AACGAG, AACGCG, AACGCT, AACGTT, AAGAGT, AAGCGC, ACACCG, ACACTG, ACCGAG, ACGACT, ACGATC, ACGCCT, ACGCGT, ACGCTC, ACGGCT, ACTAGC, AGATCT, AGCGCT, AGCTCG, ATATCG, ATCGCG, ATGCGC, CCCGGG, and CCGCGG. It should be noted here that 23 of them contain the dinucleotide CpG and four of them contain two CpG motifs. Ten of them are palindromes. Of the four hexanucleotides that do not contain the CpG dinucleotide (AAGAGT, ACACTG, ACTAGC, AGATCT), the first three include the trinucleotide duplex (ACT)•(AGT), and three contain a stop codon in at least one frame. Considering the cumulated size (>380 Mbp, see Table 10) of the sequences we analyzed, the

Table 10. Cumulated Lengths of Sequences Analyzed

Taxonomic group	All (Mbp)[a]	Intergenic regions (Mbp)[a]	Introns (Mbp)[a]	Exons (Mbp)[a]
Primates	160.08	38.29	17.78	3.17
Human chromosome 22	33.48	—	—	—
Rodentia	21.26	6.82	3.51	2.59
Mammalia	3.61	1.17	0.74	0.84
Vertebrata	5.47	1.92	1.32	1.19
Arthropoda	28.76	3.62	1.66	3.17
C. elegans	81.55	32.97	25.38	19.08
Embryophyta	48.17	15.34	6.44	10.76
S. cerevisiae	15.18	3.28	0.77	7.77
Fungi	17.78	5.79	1.07	9.28

[a](Mbp) megabase pair.

total absence of a repeat type may well indicate either a sequence unpreferred for the mechanism generating repeats or strong selective pressure against repeated occurrence of the particular sequence. The very low frequency of ACT trinucleotide repeats in all sequences is also striking (Table 2). It cannot be explained by the presence of a stop codon on one strand since genomic regions other than exons are also affected.

Repeats Longer than 24 bp

The above results apply to repeats longer than 12 bp. To be able to estimate the instability of the various repeat motifs, we also analyzed repeats longer than 24 bp and defined the expandability of a repeat motif as the total length of repeats longer than 24 bp divided by the total length of repeats longer than 12 bp. The overall distribution of these longer repeats follows comparable trends as presented above for all repeats considered (data not shown; for details see the SSRDB database at http://genetics.elte.hu/ssr). The contribution of SSRs with different unit lengths is generally similar to that observed for repeats longer than 12 bp, albeit with modified ratios. Mononucleotide repeats are, however, replaced by dinucleotide repeats as the dominant repeat type in primate, plant and yeast intergenic regions and introns. Although the abundance of the repeats longer than 24 bp is much lower and some motifs are missing, the relative frequencies of various motifs are mostly conserved. An interesting exception is the AAC repeat in the exons of embryophytes, being much more abundant using the greater length threshold than AAG, which is the most frequent repeat at the shorter threshold (101bp/Mbp vs. 18bp/Mbp compared with 253bp/Mbp vs. 317bp/Mbp for AAC vs. AAG).

The contribution of repeats longer than 24 bp to the observed SSR distribution is well represented by the expandability values, which not surprisingly, turn out to be repeat- and taxon-dependent. In all sequences, rodents show the highest and arthropods the lowest values (data not shown). The expandability of AC, AG, and AT repeats is almost uniformly high, although a preference for long $(AC)_n \bullet (GT)_n$ repeats is observed in primates. However, consistent with their general underrepresentation, no CG repeats longer than 24 bp were found. In rodent intergenic regions and introns, AC, AG, and AT dinucleotide repeats show very high expandability values (55%–80%), and most of these repeats are longer than 24 bp in rodent exons (79%–100%), even though dinucleotide repeats are generally rare in exons. In the case of trinucleotide repeats, repeat abundance and expandability rarely correlate: e.g., in primate intergenic regions, the second most abundant AAC displays the lowest expandability (10%), whereas 45% of the total length of the moderately frequent AAG originates from tracts longer than 24 bp. Trinucleotide repeats in exons exhibit uniformly low expandability: AGC is the only trinucleotide motif for which repeats longer than 24 bp can be found in all taxa. However, the expandability values for AGC in exons vary between 3% (arthropods) and 57% (rodents).

DISCUSSION

We examined the distribution of microsatellites composed of motifs 1–6 bp long in primates, other mammals, other vertebrates, arthropods, *C. elegans*, embryophytes, *S. cerevisiae* and other fungi. To obtain a detailed picture, we analyzed the frequencies of perfect SSRs longer than 12 bp in exons, introns, and intergenic regions for all of these taxa. Our results show that the abundance of certain repeat types varies with the genomic region and distribution is also characteristic of the taxonomic group examined.

It should be noted here that due to biased sequence availability in the databases, our results apply mainly to those regions of the genomes that contain protein-coding genes. Even in the case of 'all' sequences, where we did not select for genes (see Methods), the contribution of gene-rich sequences is considerable, as can be judged from the relatively high ratio of exon sequences compared to the total (Table 10). In an attempt to analyze regions less represented in GenBank, we included the human chromosome 22 sequence. Data obtained for this chromosome agree well with those obtained for all primate sequences, although an increase in (A+T)-rich microsatellites could be observed. We suggest that the poly(A/T) tails of densely scattered retroposed sequences, like Alu, LINE-1, and processed pseudogenes are responsible for this higher proportion of (A+T)-rich repeats. Chromosome 22 sequence, however, includes only the euchromatic portion, namely the relatively gene-rich long arm, 22q (Dunham et al. 1999). Thus, any interpreta-

tion of the results should bear in mind that telomeric regions or genomic regions with very low gene density are not covered in the present analysis. Repeat abundance and distribution in such regions may differ from those presented here.

Nonetheless, analysis of the datasets resulted in several noteworthy findings. First, it is very interesting to compare repeat occurrence in introns and intergenic regions. Whereas the constraints shaping protein-coding DNA sequences obviously differ from those that affect these two regions of the genome, comparison of the latter could reveal some less trivial differences. In all vertebrates, the microsatellite distribution in introns and intergenic regions is quite similar but the abundance of CCG triplets differs: Introns do not contain this type of repeat whereas it is relatively abundant in intergenic regions. Because CCG is one of the most abundant repeats in vertebrate exons, a potential bias caused by error in distinguishing exons and intergenic regions cannot be ignored (see Methods). However, we have taken sufficient and appropriate measures to avoid such errors, and we argue that the observed difference is not due to incorrect assignment of exon sequences to intergenic regions. A short calculation carried out on primate data supports this argument: Assuming that microsatellite distribution in intergenic sequences is identical to that of introns, and the increased length of CCG repeats observed in the intergenic regions can be attributed only to exonic sequences, the expected total length of AGC repeats (the dominant trinucleotide repeat of all vertebrate exons) would be almost three times greater in intergenic regions than the observed value.

The absence of CCG and ACG repeats from introns of all vertebrates could be explained by the presence of the highly mutable CpG dinucleotide within the motif. The elevated level of CCG repetition could be found in intergenic regions of all vertebrates but not in the other taxonomic groups examined. This result suggests that intergenic sequences containing regulatory DNA elements are unmethylated sufficiently in all vertebrates to prevent 5-methyl-cytosine-directed spontaneous mutations that would efficiently disrupt repeated stretches of the CCG triplet, as it is observed for intronic sequences. An alternative explanation would be that a specific mechanism exists to maintain the observed level of CCG repeats in intergenic regions of all vertebrates. The role of cytosine methylation in histone deacetylation, chromatin remodeling, and gene silencing (Razin 1998) and the presence of CpG islands (Bird 1986) may account for this phenomenon. Coffee et al. (1999) demonstrated histone deacetylation as a consequence of CGG (=CCG) repeat expansion at the 5′ end of *FMR1* in fragile X-syndrome cells. Although the association with acetylated histones depends on the methylation state of DNA, we suggest that the length of the repetitive tract may be an important factor determining the level of methylation, not only in the CGG microsatellite but also in the proximal CpG island of *FMR1*. Boyes and Bird (1992) demonstrated that transcriptional repression by DNA methylation depends on CpG density. Thus, $(CCG)_n \bullet (CGG)_n$ repeats may play an active role in vertebrates by allowing regulatory switches via the processes of DNA methylation/demethylation and, consequently, histone acetylation/deacetylation. The low level of CCG repeats in intergenic regions of species that do not methylate their DNA (*C. elegans*, *Drosophila* and yeast) suggests that, even in the absence of methyl-directed CpG suppression, CCG repeats are not favored outside the protein-coding regions. This supports the idea that either the maintenance of CCG repeats in intergenic regions of vertebrates or their suppression in most nonvertebrate sequences is an active process.

Another interesting problem is the absence of CCG from introns. In addition to the above mentioned effect of the CpG dinucleotide, CCG repeats may also be selected against because of the requirements of the splicing machinery. Repeated elements containing the motif GGG located at the 5′ end of human introns proved to be involved in splice site selection (Sirand-Pugnet et al. 1995). Long CCG sequences could compete with this region in recruiting splicing machinery components resulting in inadequate splicing. Furthermore, CCG repeats, which exhibit considerable hairpin- and quadruplex-forming potential, may influence the secondary structure of the pre-mRNA molecule. If we consider the observations showing that intron self-complementarity (Howe and Ares 1997) and mRNA secondary structure (stem loops, Coleman and Roesser 1998; hairpins, Goguel et al. 1993) modulate the efficiency and accuracy of splicing, we can assume that the presence of repeated CCG tracts would interfere with the formation of mature mRNA.

Differences between introns and intergenic regions can also be observed in nonvertebrate taxa. Intergenic regions of arthropods and vascular plants show excess of AAC and AAG repeats, respectively, when compared to introns of the same taxon. In fungi, AAT is the most frequent trinucleotide repeat in both intergenic regions and introns, but its abundance is much higher in the latter. Other biases (e.g., C, AG, and AAG in *C. elegans*; AC in yeast and other fungi) also suggest that the selective forces acting on intergenic regions and introns differ from each other in a taxon-specific manner.

It is also worth noting that tetranucleotide repeats represent a higher proportion of all vertebrate genomes than triplet repeats (Table 1), in spite of the fact that exons seem to tolerate only trinucleotide and hexanucleotide repeats effectively. The observed dependence of repeat abundance on repeated unit length is

very much biased from the expected trend of gradual decrease. SSRs with even unit length seem to be favored strongly in rodent introns and intergenic regions, and, to a lesser extent, in other vertebrates. In sharp contrast to this, penta- and hexanucleotide repeats are almost invariably more frequent than tetranucleotide repeats in all nonvertebrate taxa. This varying dependence on repeat unit length suggests fundamental differences between vertebrates and other taxa in the mechanisms of generation and fixation of simple repetitive DNA.

Although our analysis cannot measure microsatellite polymorphism per se, the maximum, average, and variance of SSR lengths may give good indication of the expected instability (data available online). As a rough estimate for this expandability, we compared the abundance of SSRs longer than 24 bp to that of repeats longer than 12 bp. AC, AG, and AT dinucleotide repeats show a striking dominance among long SSRs in introns and intergenic regions of all taxa, except for fungi. This suggests that dinucleotide repeats other than CG are the most expandable types in higher eukaryotes, a statement well supported by the numerous dinucleotide microsatellite markers used in mapping studies.

Our study confirmed the previous results indicating that the microsatellite patterns of coding and noncoding regions in eukaryotes show divergence that can be explained on the basis of differential selection (Hancock 1995). However, where Hancock (1995) — using a different approach — found high correlation between introns and intergenic regions in *Homo sapiens*, *C. elegans* and *S. cerevisiae*, we observed characteristic differences between the two regions in all taxa examined. The notion of differential selection can also be invoked to explain these differences. Moreover, our results clearly demonstrate that the preferred SSR types in exons and other genomic regions are taxon-dependent. Each repeat type that was shown to be flexible in forming various nonconventional intra- or interstrand structures (Pearson and Sinden 1998; Sinden 1999) can be found in relatively high frequencies in one or more, but never in all, taxa. This observation may indicate differences in repair enzyme specificities or other divergent factors acting at the level of selection.

Our results show, in accordance with many other studies, that strand-slippage theories alone cannot explain microsatellite distribution in the genome as a whole. The inherent potential of a sequence to form alternative DNA conformations can be important for the generation of SSRs, but cannot account for the differences observed among taxa. Enzymes and other proteins involved in various aspects of DNA-processing (i.e., replication and repair) and chromatin remodeling may be responsible for the taxon-specificity of microsatellite abundance. It should be emphasized that not only does the repetitiveness of the genomes differ (Hancock 1996b), but also the preferred microsatellite types are quite different. This may indicate that SSRs play an important role in genome evolution whereas the processes responsible for SSR generation and fixation must also have undergone alteration during evolution.

METHODS

DNA Sequences

Sequences were obtained from GenBank releases 107 (for primates), 109 (for rodents, mammals, and vertebrates) and 110 (for all other taxa) (ftp://ncbi.nlm.nih.gov/genbank). The taxonomic groups examined were the following: primates, rodents, other mammals (excluding primates and rodents), other vertebrates (excluding mammals), arthropods, *C. elegans*, embryophytes, *S. cerevisiae*, and other fungi. The human chromosome 22 sequence superlink was obtained from the Sanger Center web site (http://www.sanger.ac.uk/HGP/Chr22). Only genomic (chromosomal) sequences were included in our study. To decrease the effect of database bias as much as possible, we eliminated all GenBank entries defined as either tandem repeats, microsatellites, minisatellites, SSRs, telomeric or centromeric sequences. All mRNA, cDNA, and structural RNA sequences were excluded from the analysis. Standard UNIX tools (e.g., grep, awk) and Perl scripts were used to carry out the necessary filtering steps. From the remaining sequences, we selected those ⩾250-bp long (1000 bp in the case of primate sequences). The redundancy of sequences present in the database was minimized using the program CLEANUP (Grillo et al. 1996). We eliminated sequences that were ⩾95% similar to and overlapped by ⩾60% with another, longer sequence. The sizes of the database subsets used for the analysis, also broken down to intergenic regions, introns, and exons, are listed in Table 10. The taxonomic groups are rather arbitrarily defined, primarily based on sequence availability. The species contributing to >5% of sequences in the appropriate database subset are listed in Table 11.

Although full chromosomal sequences are available for *S. cerevisiae* and *C. elegans*, the unconfirmed nature of the majority of sequence annotations prevented their meaningful use in our study. The potential risk of incorrectly classifying DNA fragments into exons, introns, and intergenic regions cannot be neglected even for sequences derived from the traditional GenBank database sections. Although the extent of such bias did not seem to be large, we tried to minimize it by excluding from the analysis all such entries that contained no CDS line and by a rather conservative handling of alternative splicing (either biologically relevant or due to uncertain predictions or database errors). We eliminated from our analysis all DNA fragments where exon–intron junctions of a protein-coding gene was specified in two or more different, contradictory ways. We also ignored putative intergenic regions before and after such genes. Despite our precautions, there still may be a few exon or intron sequences specified incorrectly as intergenic regions. We think, however, that the resultant bias should not affect our conclusions.

Because most of our results were obtained from sequences containing protein-coding genes, we were also interested in whether or not this caused a bias in the SSR distribution. To test this, we also carried out the analysis on the full

Table 11. Contribution of Various Species to the Taxonomic Groups Studied

Taxonomic group	Number of species	Major species[a]	Sequence lengths (%)
Primates	64	*Homo sapiens*	99.43
Rodentia	81	*Mus musculus*	73.71
		Rattus norvegicus	18.25
Mammalia	203	*Bos taurus*	27.26
		Sus scrofa	20.72
		Oryctolagus cuniculus	19.09
		Ovis aries	10.59
		Canis familiaris	6.62
Vertebrata	353	*Gallus gallus*	32.20
		Fugu rubripes	17.76
		Xenopus laevis	12.15
Arthropoda	586	*Drosophila melanogaster*	84.27
		Drosophila sp.[b]	7.93
Embryophyta	1313	*Arabidopsis thaliana*	79.18
Fungi	1164	*Schizosaccharomyces pombe*	48.41

[a]Only the species representing >5% of the cumulated sequence length are listed.
[b]All other *Drosophila* species included in the analysis (159 species).

sequence of the human chromosome 22. The sequence was used as a whole, i.e., no attempt was made to assign portions of the chromosome 22 sequence to exon, intron, or intergenic regions.

SSR Analysis

From the database subsets obtained for each taxa, we extracted all perfect tandem repeats with a maximum unit size of six that contained at least two consecutive units, as described by Jurka and Pethiyagoda (1995). The SSRs were then grouped according to their localization in the genome (i.e., within exons, introns, or intergenic regions) using Perl scripts. This classification was based on the information provided in the CDS feature table lines of the GenBank entries. Intergenic regions were defined as being the part of DNA from the end of the last exon of one gene to the beginning of the first exon of the following gene (similar to Hancock 1995). Fragments derived from entries containing no CDS line were not classified to regions but were retained in all sequences.

Further data analysis (classification of SSRs by unit patterns and computing the values listed in the tables) was carried out as described by Jurka and Pethiyagoda (1995). In the present analysis, repeats with unit patterns being circular permutations and/or reverse complements of each other were grouped together as one type. The total number of such nonoverlapping types is 501 for 1–6-bp long motifs (for details see Jurka and Pethiyagoda 1995).

We mainly examined the distribution of perfect repeats >12-bp long. Because microsatellites are often disrupted by single base substitutions, the contribution of various repetitive motifs to the overall repetitivity of the genome could be better estimated using this relatively short cutoff length. However, to assess expandability of the repeats, we also identified repeats longer than 24 bp. For a particular motif, expandability is defined as the total length of repeats longer than 24 bp divided by the total length of repeats longer than 12 bp.

To allow direct comparisons regardless of the cumulated size of genomic regions in the database subsets, normalized total lengths of the microsatellites were calculated for 1 Mbp of the appropriate genomic sequence type.

ACKNOWLEDGMENTS

This work was supported by grant OTKA T19278 from the Hungarian National Scientific Research Fund. We thank Ágnes Major for helpful discussion and Paul Klonowski for the computer program of tandem repeat extraction. We also thank the anonymous referees for their useful comments and suggestions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Arzimanoglou, I.I., Gilbert, F., and Barber, H.R.K. 1998. Microsatellite instability in human solid tumors. *Cancer* **82:** 1808–1820.

Bates, G. and Lehrach, H. 1994. Trinucleotide repeat expansions and human genetic disease. *BioEssays* **16:** 277–284.

Beckmann, J.S. and Weber, J.L. 1992. Survey of human and rat microsatellites. *Genomics* **12:** 627–631.

Bidichandani, S.I., Ashizawa, T., and Patel, P.I. 1998. The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. *Am. J. Hum. Genet.* **62:** 111–121.

Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209–213.

Boyes, J. and Bird, A. 1992. Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.* **11:** 327–333.

Coffee, B., Zhang, F. Warren, S.T., and Reines, D. 1999. Acetylated histones are associated with *FMR1* in normal but not fragile X-syndrome cells. *Nat. Genet.* **22:** 98–101.

Coleman, T.P. and Roesser, J.R. 1998. RNA secondary structure: an important *cis*-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry* **37:** 15941–15950.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Field, D. and Wills, C. 1996. Long, polymorphic microsatellites in simple organisms. *Proc. R. Soc. Lond.* **263:** 209–215.

Gacy, A.M., Goellner, G., Juranić, N., Macura, S., and MacMurray, C.T. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **81:** 533–540.

Gerber, H.-P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S., and Schaffner, W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263:** 808–811.

Goguel, V., Wang, Y. and Rosbash, M. 1993. Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. *Mol. Cell. Biol.* **13:** 6841–6848.

Grillo, G., Attimonelli, M., Liuni, S., and Pesole, G. 1996. CLEANUP: A fast computer program for removing redundancies from nucleotide sequence databases. *Comput. Appl. Biosci.* **12:** 1–8.

Hancock, J.M. 1995. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* **41:** 1038–1047.

———. 1996a. Simple sequences in a 'minimal' genome. *Nat. Genet.* **14:** 14–15.

———. 1996b. Simple sequences and the expanding genome. *BioEssays* **18:** 421–425.
Howe, K.J. and Ares, M., Jr. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc. Natl. Acad. Sci.* USA **94:** 12467–12472.
Jakupciak, J.P. and Wells, R.D. 1999. Genetic instabilities in (CTG•CAG) repeats occur by recombination. *J. Biol. Chem.* **274:** 23468–23479.
Jurka, J. and Pethiyagoda, C. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* **40:** 120–126.
Kashi, Y., King, D., and Soller, M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* **13:** 74–78.
Lagercrantz, U., Ellegren, H., and Andersson, L. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucl. Acids Res.* **21:** 1111–1115.
Pearson, C.E. and Sinden, R.R. 1998. Trinucleotide repeat DNA structures: Dynamic mutations from dynamic DNA. *Curr. Opin. Struct. Biol.* **8:** 321–330.
Perutz, M.F., Johnson, T., Suzuki, M., and Finch, J.T. 1994. Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci.* **91:** 5355–5358.
Razin, A. 1998. CpG methylation, chromatin structure and gene silencing—a three-way connection. *EMBO J.* **17:** 4905–4908.
Reddy, P.S. and Housman, D.E. 1997. The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.* **9:** 364–372.
Richards, R.I. and Sutherland, G.R. 1992. Dynamic mutations: A new class of mutations causing human disease. *Cell* **70:** 709–712.
———. 1994. Simple repeat DNA is not replicated simply. *Nat. Genet.* **6:** 114–116.
Schlötterer, C. and Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucl. Acids Res.* **20:** 211–215.
Sia, E.A., Jinks-Robertson, S., and Petes, T.D. 1997. Genetic control of microsatellite instability. *Mutation Research* **383:** 61–70.
Sinden, R.R. 1999. Trinucleotide repeats: Biological implications of the DNA structures associated with disease-causing triplet repeats. *Am. J. Hum. Genet.* **64:** 346–353.
Sirand-Pugnet, P., Durosay, P., Brody, E., and Marie, J. 1995. An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken β-tropomyosin pre-mRNA. *Nucl. Acids Res.* **23:** 3501–3507.
Tautz, D. and Schlötterer, C. 1994. Simple sequences. *Curr. Opin. Genet. Dev.* **4:** 832–837.
Tautz, D., Trick, M., and Dover, G. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322:** 652–656.
Usdin, K. 1998. NGG-triplet repeats form similar intrastrand structures: Implications for the triplet expansion diseases. *Nucl. Acids Res.* **26:** 4078–4085.
Warren, S.T. and Nelson, D.L. 1993. Trinucleotide repeat expansions in neurological disease. *Curr. Opin. Neurobiol.* **3:** 757–759.
Wooster, R., Cleton-Jansen, A.-M., Collins, N., Mangion, J., Cornelis, R.S., Cooper, C.S., Gusterson, B.A., Ponder, B.A.J., von Deimling, A., Wiestler, O.D. et al. 1994. Instability of short tandem repeats (microsatellites) in human cancer. *Nat. Genet.* **6:** 152–156.

Received January 5, 2000; accepted in revised form May 4, 2000.