J. Biosci., Vol. 20, Number 5, December 1995, pp 613-627. © Printed in India.

Analysis of CAG/CTG triplet repeats in the human genome: Implication in transcription factor gene regulation

RASHNA BHANDARI and SAMIR K BRAHMACHARI*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

MS received 15 September 1995; revised 14 November 1995

Abstract. Instability and polymorphism at several CAG/CTG trinucleotide repeat loci have been associated with human genetic disorders. In an attempt to identify novel sites that may be possible loci for expansion of CAG/CTG repeats, we searched all human sequences in the EMBL nucleotide sequence database for $(CAG)_5$ and $(CTG)_5$ repeats. We have identified 121 human DNA sequences of known and unknown functions that contain stretches of five or more CAG or CTG repeats. Many repeat stretches were interrupted by variant triplets, a significant number of which differ from the repeat triplet only by a single base, suggesting that these evolved from the parent triplet by point mutations. A large number of human transcription factor genes were found to contain CAG repeats within their coding sequences. Analysis of the EMBL transcription factors database showed that many transcription factor genes of other eukaryotes, including genes involved in *Drosophila* embryo development, possess these repeats. Interestingly, CAG repeats are absent from prokaryotic transcription factors. Different sequence entries for the human TATA box binding protein showed a polymorphism in the length of the CAG repeat in this gene, suggesting that loci other than those already known to be associated with genetic diseases may be possible sites for repeat instability related disorders. On the basis of our findings in this database analysis, we propose a role for CAG repeats as cis-acting regulatory elements involved in fine-tuning gene expression.

Keywords. Triplet repeats; sequence analysis; transcription factor genes; TBP.

1. Introduction

Trinucleotide repeat regions in the genome have recently been shown to be unstable and polymorphic. The expansion of trinucleotide repeats in humans has been associated with several human genetic diseases (Sutherland and Richards 1995). In the Fragile X syndrome, there is expansion of a CGG repeat in the 5'UTR of the FMR–1 gene (Kremer *et al* 1991). Myotonic dystrophy (DM) is associated with the expansion of a CTG repeat in the 5'UTR of the myotonin kinase gene (Mahadevan *et al* 1992; Fu *et al* 1992). Six neurodegenerative disorders—Huntington's disease (HD; Huntington's Disease Collaborative Research Group 1995), spinal and bulbar muscular atrophy (SBMA; La Spada *et al* 1991), spinocerebellar ataxia type 1 (SCA-1; Orr *et al* 1995), dentatorubral pallidoluysian atrophy (DRPLA; Koide *et al* 1994), Haw river syndrome (Burke *et al* 1994), and Machado-Joseph disease (MJD; Kawaguchi *et al* 1994)—are caused by an amplified CAG repeat

Corresponding author (Fax: 91-080-3341683; Email: skb@mbu.iisc.ernet.in.).

within the coding sequence of the gene, resulting in an increase in the length of a polyglutamine tract in the corresponding protein. Despite a variation in the location of the repeat within the coding region, all these inherited diseases share the characteristics of increased severity with increasing repeat length. The age of onset of disease decreases in subsequent generations, thus exhibiting genetic anticipation (Richards and Sutherland 1992).

With the discovery that trinucleotide repeat expansions are involved in genetic diseases, attempts have been made to identify other repeat regions that may be possible sites for similar expansions. The screening of cDNA libraries has revealed that CAG and CGG repeats are found in close proximity to several gene loci (Riggins *el al* 1992; Li *et al* 1993). By computer aided sequence analysis of approximately 10 Mb human DNA sequence in GenBank, Han *et al* (1994) have studied the frequency distribution of all the 64 possible trimers. They found that the disease associated trimers, GGC and CAG are over-represented in the genome. They also identified 51 human genes that contain $(GGC)_{\geq 4}$ repeats. A search carried out by Stallings (1994) on approximately 16 Mb GenBank human DNA sequences for all 10 possible microsatellie motifs with a perfect string of eight or more trinucleotides, showed that CAG is the most abundant triplet repeat unit in the human genome.

There are, to date, no composite data available on the location of CAG repeats in the human genome, except for the loci known to be associated with genetic diseases· It is possible that there are other regions in the genome containing long stretches of CAG/CTG repeats that show length polymorphism. The identification and analysis of genes associated with such regions may provide a clue to understanding the functional significance of these repetitive DNA sequences in the genome and perhaps also of the genetic diseases associated with them. In an attempt to identify other sites in the human genome that may be possible loci for expansion of CAG/CTG repeats, we carried out a search on the EMBL nucleotide sequence database. All human sequences in the EMBL database (Release 41, Dec. 1994) were screened for the presence of a stretch of at least 5 tandem repeats of CAG or CTG· We observed that a large number of human transcription factor genes contain $(CAG)_{\geq 5}$ repeats in their coding regions. Analysis of the transcription factors of other organisms listed in the EMBL transcription factors database showed that this repeat is present in many eukaryotic transcription factor genes, but is absent from the genes encoding prokaryotic transcription factors. Based on our analysis, we have examined the potential role of CAG/CTG repeats as *cis*-acting regulatory elements that quantitatively modulate gene expression.

2. Materials and methods

In order to identify human sequences containing CAG or CTG repeats, the FASTA program was used to search all primate sequences in the EMBL nucleotide sequence database (Release 41, Dec. 1994) for the strings. $(CAG)_5$ and $(CTG)_5$. Since the FASTA program searches for a string only once in a sequence, the sequences containing a $(CAG)_5$ or $(CTG)_5$ repeat were put through the GCG Find program (Devereux *et al* 1984) to search for $(CAG)_5$ strings elsewhere in the same sequence.

The annotation of each sequence containing a $(CAG)_{\geq 5}$ or $(CTG)_{\geq 5}$ repeat was then examined to determine whether the repeat occurred in the 5' untranslated

region (5'UTR), coding DNA sequence (CDS), 3'UTR, or in an intron. If a repeat was present in the coding sequence, the amino acid coded for by the triplet repeat was determined by a simple calculation using the start position of the coding sequence given in the feature table of the EMBL sequence entry. The location of the repeat region could not be determined for unannotated sequences and sequences for which the entire CDS was not defined.

The sequence in the vicinity of each stretch of five repeats was then examined. In several sequences we found clusters of CAG/CTG triplets around the tandem $(CTG/CTG)_{\geq 5}$ repeat. CAG/CTG triplets in this region, that were separated by not more than 12 nucleotides, were included in the repeat stretch shown in table 1. The ends of the repeat were defined by the presence of at least two tandem CAG/CTG triplets that were separated from the nearest CAG/CTG by not more than 12 nucleotides. In most cases the CAG/CTG triplets were interrupted by multiples of three nucleotides. The sequences of variant triplets (T) that interrupted the perfect repeats were noted and the frequency of occurrence of each interrupting triplet was determined.

A large number of human transcription factor genes were found to contain $(CAG)_{\geq 5}$ stretches. To examine the transcription factors of other organisms, the EMBL transcription factors database was searched for (CAG)5 stretches using the GCG Find program. All non human transcription factors containing at least five tandem CAG/CTG repeats were examined for repeat and non-repeat triplets in that region, as above.

3. Results

3.1 *Distribution of CAG repeats in human DNA sequences*

We have screened approximately 54 Mb DNA from 52549 primate sequence entries in the EMBL nucleotide sequence database and have identified 121 unique sequences in the human genome that contain $(CAG/CTG)_{\geq 5}$ repeats. The observed number of repeats is significantly higher than expected [$P << 0{\cdot}01$, based on Poisson distribution, with mean number of $(CAG/CTG)_5$ repeats per sequence = 10^{-6}]. Table 1 lists the human sequences that contain five or more tandem repeats of CAG/CTG. The subscript denotes the length of CAG(R) and CTG(R') repeats found in these sequences. For those sequence entries in which the CDS is defined, we have determined the location of the repeat with respect to the coding sequence. The repeat CAG codes for G1n, Ser, or Ala in the three reading frames. Similarly, the repeat CTG codes for Leu, Cys, or Ala. In 22 of the 57 sequences in which the reading frame of CAG repeats present within the coding region could be determined, the repeat CAG codes for a stretch of Gin residues. In ten cases, the reading frame is AGC, coding for Ser, and in five, GCA, encoding Ala. The CTG repeat codes for Leu in all cases.

From among the 48 $(CAG)_{\geq 5}$ repeats in human sequences for which the location with respect to the transcribed region could be determined, it was found that none were located within an intron. However three $(CTG)_{\geq 5}$ repeats (i.e., CAG on the opposite strand) out of 44 were found within introns. In other words, CAG may be excluded from introns of human genes in a strand specific manner, with no

Table 1. Human sequence entries in the EMBL database that contain (CAG/CTG) $_{\geq 5}$ repeats.

Entry name	Description	Repeat	Location	Amino acid[a]
HS02013*	Sterol regulatory element binding protein	$R_2TR_2TR_6T_2RTR_3$	CDS	Ser
HS04840	Onconeural ventral antigen-1 (Nova-1)	R_5	CDS	Ala
HS07857	18 kDa Alu RNA binding protein	$R_8T_3R_2TRT_3R_2T_2R_2TR_2$	CDS	Ala
HS9515	T-lymphocyte repeat containing mRNA	R_5	—	—
HSAF9X*	AF-9	$R_8TR_{10}TR_{10}TR_3TR_6$	CDS	Ser
HSAFPEBP*	α-fetoprotein enhancer binding protein	$R_7T_2RTR_2$	CDS	Gln
		$R_5T_3R_4$	CDS	Gln
HSALMG1	Calmodulin gene exon 1 (CALM 1)	R_7	—	—
HSANDREC*[b,c]	Androgen receptor	$R_{20}N_{18}R_6$	CDS	Gln
HSASH1A*	Achaete scute homologous protein (ASH 1)	R_{14}	CDS	Gln
HSAUTANT	Auto antigen	$R_5T_2R_6TR_2$	CDS	Ser
HSBNPA	Brain natriuretic peptide	R_7TR_4	5′UTR	—
HSCDP*	CCAAT displacement protein	R_5	—	—
HSCEI5A	CEI5, 3′flank	R_6	3′UTR	—
HSCOUPII*	Chick ovalbumin upstream promoter transcription factor II	R_5	5′UTR	—
HSCYT2A	Cytokeratin 2	R_7	CDS	Ser
HSDB11*	DB1 (IL-3 promoter binding zinc finger protein)	R_7	CDS	Gln
HSDNAPLYC	Repeat region DNA (wg1d10)	R_8	—	—
HSDRPLA	DRPLA mRNA	R_{10}	CDS	Gln
HSEGR1*[c]	Early growth response protein 1	$R_6N_{33}R_6$	CDS	Ser
HSEGR2A*	Early growth response protein 2	R_5	CDS	Ala
HSENKB4	Preproenkephalin B	R_2TR_7	3′UTR	—
HSEPKER	Type I epidermal keratin	R_6	CDS	Ser
HSERF2	ERF-2	R_7	CDS	Gln
HSERK1	ERK1 (protein serine threonine kinase)	R_6	5′UTR	—
HSFRI	Chromosome 13q12 DNA fragment	R_5	—	—
HSHBRM*	hbrm (homologue of *Drosophila* brm)	$R_3TR_{13}TR_2T_4RTRT_2R_2$	CDS	Gln
HSHCCA	Splicing factor (CC1.3)	R_7	5′UTR	—
HSHSNFA*	hSNF2a (estrogen receptor)	$R_3TR_{13}TR_2TR_3$	—	—
HSI9R	Interleukin 9 receptor	R_2TR_8	CDS	Ser
HSIFNRTF*	IFN responsive transcription factor	R_6	CDS	Ser
HSJUNA*	c-jun proto-oncogene	R_5	CDS	Gln
HSKER16A1	Keratin type 16 gene, exon 1	R_5	—	—
HSKERP1	Keratin psuedogene, exon 1	R_5	—	—
HSMEF2*	Myocyte-specific enhancer factor 2	R_{11}	CDS	Gln
HSMMDA	Human DNA from cosmid MMDA from chromosome 19q13	R_5	—	—
HSMRNAC	Human brain cDNA (clone CTG-A4)	$R_4TR_9TRTR_6T_2R_2$	—	—
HSMRNAD1	Human brain cDNA (clone CTG-B10)	R_5TR_9	—	—

Table 1. (*Contd*)

Entry name	Description	Repeat	Location	Amino acid[a]
HSMYC1*	c-myc oncogene	R_5	CDS	Gln
HSNATPEP	Natriuretic peptide precursor	R_7TR_4	5′UTR	—
HSNFYA*	CAAT box DNA binding protein subunit A	R_5	CDS	Gln
HSNOCTPOU*	N-Oct3, N-Oct5a and N-Oct5b	$R_3TR_3TRTR_3TR_2TR_6$	CDS	Gln
HSNOTIC	Human chromosome 13q12-specific mRNA	R_5	—	—
HSNUCLEOB*	Nucleobindin	R_6	CDS	Gln
HSOFR1A	ORF of a gene expressed in human hypothalamus	R_{14}	CDS	Ser
HSORFG	Human mRNA for randomly sequenced product	$R_2TR_5TRN_8RN_{12}R_2$	CDS	Ala
HSPDEE	Phosphodiesterase	R_5	5′UTR	—
HSPIM1A	pim-1 proto-oncogene	R_8	5′UTR	—
HSREPRA	Simple repeat region (wg1a9)	$R_3T_3R_{12}$	—	—
HSRNAF	Brain cDNA (clone CTG-B37)	R_{15}	—	—
HSRNAG	Brain cDNA (clone CTG-B43A)	$R_3TR_5TR_2TR_5TR_{10}$	—	—
HSRSRFC49*	SRF-related proteins RSRFC4 and RSRFC9	R_9	CDS	Gln
HSRWG1A10	Simple DNA sequence region (clone wg1a10)	R_5	—	—
HSRWG1A4	Simple DNA sequence region (clone wg1a4)	R_{10}	—	—
HSRWG1A5	Simple DNA sequence region (clone wg1a5)	R_9	—	—
HSRWG1F9	Simple DNA sequence region (clone wg1f9)	R_9	—	—
HSSATB1A*	MAR/SAR DNA binding protein	$R_2N_2RTRTR_3TRTR_7$	CDS	Gln
HSSATM01	DNA sequence 3′-flanking mini-satellite pMS1	R_6	—	—
HSSCA1	Ataxin (SCA-1)	$R_{12}TRTR_{15}N_5R_2$	CDS	Gln
HSSRP14A	Signal recognition particle subunit 14	$R_8T_3R_2TR_2T_2R_2T_2R_2TR_2$	CDS	Ala
HSTFEB*	Transcription factor TFEB	R_6TR_3	CDS	Gln
HSTFIID*†	TATA binding protein (TFIID) (HeLa cells)	$R_3T_3R_9TRTR_{18}TRNRN_3$ $RN_3RN_5R_2N_9R_2$	CDS	Gln
HSTFIIDAA*†	Transcription factor TFIID (human hypothalamus)	$R_3T_3R_{11}TRTR_{16}TRNR$ $N_3RN_3RN_5R_2N_9R_2$	CDS	Gln
HSTFIIDX*†	Transcription factor TFIID (Namalwa cell line)	$R_3T_3R_9TRTR_{14}TRNRN_3$ $RN_3RN_5R_2N_9R_2$	CDS	Gln
HSTGFBC	Transforming growth factor-β	$R_2T_2R_6$	CDS	Gln
HSTRINUC	Huntington's disease gene (HD)	R_{21}	CDS	Gln
HSTUPLE1*	TUP1-like enhancer of split gene 1	R_5	CDS	Gln
HSWG1F4RP	Repeat region DNA (clone wg1f4)	$R_2N_4R_{10}$	—	—

Table 1. (*Contd*)

Entry name	Description	Repeat	Location	Amino acid[a]
S62539	Insulin receptor substrate-1	R_8	5′UTR,	—
		R_7	CDS	Ser
		R_6	CDS	Gln
A08001	Repeat region	R'_6	—	—
A18521	Mullerian inhibiting substance	R'_5	CDS	Leu
HS03398	Receptor 4-1BB ligand	R'_6	CDS	Leu
HS04806	FLT3/FLK2 receptor tyrosine kinase ligand	R'_5	CDS	Leu
HS09510	Glycyl tRNA synthetase	R'_5	5′UTR	—
HS10886	Density enhanced phosphatase-1	R'_6	CDS	Leu
HS8955	β-spectrin (HSPTB1)	R'_5	Intron	—
HSA3NARSP	Nicotine acetylcholine receptor α-3 subunit	$R'_6TR'_2$	CDS	Leu
HSAK1	Cytosolic adenylate kinase (AK1)	R'_7	Intron	—
HSALPP	Placental alkaline phosphatase	R'_7	CDS	Leu
HSAMYB15	Amyloid-β protein (APP)	R'_6	Intron	—
HSAP2*	Transcription factor AP2	$R'_3TR'_5$	3′UTR	—
HSAPB01	Apolipoprotein B-100	$R'_2TR'_2TR'T_2R'_6$	CDS	Leu
HSAREB6*	Transcription factor AREB6	R'_7	3′UTR	—
HSARSBX	Aryl sulphatase B	R'_5	CDS	Leu
HSBPGM1	2,3-bisphosphoglycerate mutase	R'_8	5′UTR	—
HSC1INH2	C1 inhibitor	$R'_2T_2R'_5$	CDS	Leu
HSCA1V	Collagen α-1 (V) chain	$R'_2T_2R'_6$	CDS	Leu
HSCN2	Skin collagenase	R'_5	CDS	Leu
HSCP2101	Mutant CYP 21 B (steroid 21 mono-oxygenase)	$R'_5TR'_2T_4R'_2$	CDS	Leu
HSD9S118	DNA segment containing GT repeat (D9S118)	R'_8	—	—
HSDIPEPA	Dipeptidyl aminopeptidase like protein	R'_5	5′UTR	—
HSEFL2	EHK-1 receptor tyrosine kinase ligand (EFL-2)	R'_6	CDS	Leu
HSENA78	ENA-78 (neutrophil activating peptide)	R'_5	CDS	Leu
HSEPOHYDD	Epoxide hydrolase	R'_6	3′flank	—
HSERK1	ERK1 (protein serine threonine kinase)	R'_5	3′UTR	—
HSFLA1A	Leucocyte associated molecule-1 α-subunit	R'_5	CDS	Leu
HSGAS	Growth-arrest-specific protein	R'_6	CDS	Leu
HSGFAP	Glial fibrillary acidic protein	R'_6	5′UTR	—
HSGLUT5	Glucose transport like 5 (GLUT5)	$R'_5N_{11}R'N_9R'_2$	CDS	Leu
HSHGFAP	Hepatocyte growth factor activator precursor	R'_6	CDS	Leu
HSHNF4	Hepatocyte nuclear factor 4	R'_5	CDS	Leu
HSIIP	γ-interferon-inducible protein (IP3)	$R'_2TR'T_2R'_5$	CDS	Leu

Table 1. (*Contd*)

Entry name	Description	Repeat	Location	Amino acid[a]
HSK1AAC	ORF	R'_5	CDS	Leu
HSLHHCGR	Leutenising hormone-choriogonadotropin receptor	$R'_2TR'_5$	CDS	Leu
HSLPAP	Lymphocyte phosphatase associated protein	$R'_5T_2R'_2$	CDS	Leu
HSMDAPKB†	Myotonic dystrophy associated protein kinase	R'_{11}	3′UTR	—
HSMYOPK†	Myotonin protein kinase associated triplet repeat region	R'_5	3′UTR	—
HSNNAR	α-3 neuronal nicotinic acetyl-choline receptor subunit	$R'_6TR'_2$	CDS	Leu
HSNUCPHOF	Processed pseudogene of nucleophosmin/B23	R'_7	—	—
HSORFM	Randomly sequenced product	$R'_6T_3R'_2$	—	—
HSPA1V	pro-α-1 (V) collagen	$R'_2T_2R'_6$	CDS	Leu
HSPPR	Protective protein	R'_8	CDS	Leu
HSPSPBA	Pulmonary surfactant associated protein PSP-B	$R'_2T_2R'_5N_8R'N_{11}R'_2$	CDS	Leu
HSREPRC	Simple repeat region (wg1d6)	R'_6	—	—
HSRNAE	Human brain cDNA (clone CTG-B33)	R'_{13}	—	—
HSRWG1E10	Simple DNA sequence region (clone wg1e10)	R'_8	—	—
HSRWG2A11	Simple DNA sequence region clone (wg2a11)	R'_7	—	—
HSRWG2B8	Simple DNA sequence region clone (wg2b8)	$R'_2T_3R'_6N_4R'_2$	—	—
HSRWG2D12	Simple DNA sequence region clone (wg2d12)	$R'_2T_2R'T_4R'TR'T_2R'_2$ $TR'T_2R'_8T_4R'_4$	—	—
HSSIRPOCK	Simple repeat polymorphism	R'_7	—	—
HSSUBTDA	Subtelomeric DNA sequence	$R'_6TR'TR'_2$	—	—
HSTGFBG1	Transforming growth factor-β precursor	$R'_5T_2R'_2$	CDS	Leu
HSUVOECAD	Uvomorulin (E-cadherin)	R'_5	CDS	Leu
HSVLDLR01	Very low density lipoprotein receptor	R'_5	5′flank	—
S57777	N-acetyl galactosamine-4-sulphatase	R'_5	CDS	Leu
S57793	Lutenising hormone receptor	$R'_2TR'_5$	CDS	?

R, CAG triplet; R', CTG triplet; T, any triplet; N, single nucleotide·

[a]The amino acid coded for by the $(CAG/CTG)_{\geq 5}$ repeat is given· The reading frame of neighbouring CAG/CTG triplets may not be the same·

[b]Different sequence entries for androgen receptor have number of repeats ranging from 15 to 20.

[c]Two repeats found within the same sequence are close enough to be located within the same nucleosome·

*Human transcription factor genes·

†Different entries for the same sequence, which differ in the number of triplet repeats·

?, Author given protein sequence is in conflict with the conceptual translation·

CAG repeats being seen in the sense strand. A similar result was obtained by Stallings (1994) from a search on the GenBank DNA database, wherein it was found that none of the 15 $(CAG)_{\geq 8}$ repeats found in mouse, human or rat sequences was located within intronic sequences. It has been suggested that the exclusion of CAG from introns is possibly because of its similarity to the 3' consensus splice site, CAGG (Stallings 1994).

In 43 sequences we found a stretch of interrupted CAG or CTG repeats. Long stretches of interrupted repeats were seen in the Alu RNA binding protein, the human homologue of *Drosophila* brm, the signal recognition particle subunit 14, the transcription factor N-Oct, the human TATA-binding protein, and in apolipoprotein B-100. A striking feature was that interruptions by multiples of three nucleotides were significantly higher than expected ($\chi 2$ =113·9 for interruptions of CAG and $\chi 2$ = 46·4 for CTG; $P << 0{\cdot}001$). In most cases, the interruptions were by triplets, such that the reading frame of successive perfect repeats was maintained. Non-triplet interruptions were seen only in eight sequences. The number of intervening nucleotides (N) and triplets (T) within each repeat sequence is shown in table 1.

Table 2 lists the interrupting triplets (T) that were in frame with the $(CAG/CTG)_{\geq 5}$ repeats, and the number of times each interrupting triplet occurred. By simple statistical analysis it can be seen that if the interrupting triplets had arisen randomly, then only 9/63 (i.e., 14%) of them would differ from CAG or CTG by one base. However, we see that 88% of the interrupting triplets differ from the flanking CAG

Table 2. Frequency distribution of triplets interrupting (CAG/CTG) repeats.

Parent triplet	Type of mutation	Interrupting triplet	Base change	Occurrence	Frequency (%)
CAG	Point mutation (transition)	TAG	C → T	8	9·1
		CGG	A → G	10	11·4
		CAA	G → A	41	46·6
	Point mutation (transversion)	AAG	C → A	2	2·3
		GAG	C → G	3	3·4
		CCG	A → C	5	5·7
		CTG	A → T	2	2·3
		CAC	G → C	3	3·4
		CAT	G → T	3	3·4
	Two or three base changes			11	12·4
CTG	Point mutation (transition)	TTG	C → T	12	21
		CCG	T → C	5	8·8
		CTA	G → A	4	7
	Point mutation (transversion)	ATG	C → A	1	1·8
		GTG	C → G	1	1·8
		CAG	T → A	1	1·8
		CGG	T → G	—	—
		CTC	G → C	5	8·8
		CTT	G → T	—	—
	Two or three base changes			28	49

triplets by one base change, and in the case of CTG repeats, 51 % of the variant triplets differ by a single base. Since these values are significantly higher than expected, the null hypothesis of "random origin" can be rejected. (Expected frequencies are computed using binomial distribution, $X^2 = 384{\cdot}2$ for CAG interruptions and $X^2 = 62{\cdot}8$ for CTG interrupting triplets; $P << 0{\cdot}001$). The indication is that the interrupting triplets are very likely to have arisen by point mutations in an ancestral CAG or CTG repeat sequence. We also observe that single base transitions, i.e., A↔G or C↔T, are more frequent than transversions. Transitions were approximately three-fold more common than transversions in the case of CAG and approximately two-fold for CTG. It has been seen that transitions are two times more likely than transversional mutations (Kimura 1983). The high frequency of CAG to CAA mutations may be due to the synonymous nature of this substitution.

3.2 *Several transcription factor genes contain CAG repeats*

A large number of human transcription factors were found to contain CAG repeats in their coding sequences. To look for transcription factors in other organisms that contain CAG repeats, sequences listed in the EMBL transcription factors database were searched for strings of $(CAG)_5$. Out of 563 transcription factor gene entries from 86 different species analysed, 48 transcription factors from 12 species were found to contain $(CAG)_{\geq 5}$ repeats (tables 1 and 3). A number of *Drosophila* genes involved in embryo development, such as *dorsal, knirps, deformed* and *hunchback,* possess long CAG stretches in their coding regions. In most cases, CAG repeats are present in the sense strand of transcription factors. Only in human transcription factors AP2 and AREB6 are CTG stretches found in the coding strand. Most transcription factors monitored contain CAG repeats in the reading frame that codes for glutamine, but some also contain serine and alanine stretches coded for by this repeat. An interesting observation was that although CAG repeats were present in the transcription factors of lower and higher eukaryotes, none of the prokaryotic transcription factor genes contained CAG repeats.

3.3 *Length polymorphism in human TATA binding protein*

Human TATA binding protein has a stretch of CAG repeats similar in length to loci such DRPLA, MJD, and SCA-1 that show dynamic mutations leading to repeat expansion. It is therefore quite possible that the repeat stretch at the TFIID locus may similarly undergo dynamic mutation. Figure 1 shows the alignment of the repeat region of the MJD la and SCA-1 loci with the different sequence entries for human TBP that are derived from different sources, viz., HeLa cells, Namalwa cells, and human hypothalamus. It can be seen that hTBP contains three CAG repeats that are interrupted by the variant triplet CAA, which also codes for glutamine. Similar variations on the CAG motif are also seen in the MJD la locus. The 5' end of the CAG repeat in this locus contains two variant sequences, CAA and AGG, at three positions. The positions at which the variant CAA triplets occur is conserved in all the sequence entries of hTBP, except at two positions in the TBP cloned from human hypothalamus.

It is interesting to note that there is a polymorphism in the length of the repeats in TBP derived from different sources. The number of repeats in the 3' CAG

Table 3. Non-human transcription factor genes containing $(CAG/CTG)_{\geq 5}$ repeats.

Entry name	Description	Species	Repeat	Location	Amino acid
BTTRFSQA	Transcription factor	*Bos taurus* (cattle)	R_5	—	—
DM74E	74E	*D. melanogaster*	$R_5TR_2TRT_2R_2$	CDS	Gln
			$R_2N_{10}R_5$	CDS	Ser
DMDFD	Deformed (Dfd)	*D. melanogaster*	$R_2T_2R_5TR_4TR_2TR_2$	CDS	Gln
DMDORSAL	Dorsal	*D. melanogaster*	$R_2T_3R_4TR_4TRTRT$		
			$RTR_2T_3R_9$	CDS	Gln
			R_4TR_7	CDS	Gln
DMHBG	Hunchback	*D. melanogaster*	R_5TR_2	CDS	Gln
DMKNIRPS	Knirps	*D. melanogaster*	R_7		
DMTATABF	TFIID (TATA binding factor)	*D. melanogaster*	R_5TR_2	CDS	Gln
DVHB	Hunchback (hb)	*D. virilis*	R_2TR_7	CDS	Gln
FCFTT	Feline proviral (FTT) v-myc	*Felis catus* (cat)	R_5	CDS	Gln
GDCMD1	CMD1 (mouse myoD1 homologue)	*G. domesticus* (chicken)	R_6	CDS	Ser
MAHCRMYC	Fused hcr and myc genes	*Marmota monax* (woodchuck)	R_5	CDS	Gln
MMCJUN	c-jun oncogene	*Mus musculus* (mouse)	R_5	CDS	Gln
MMGRLA	Glucocorticoid receptor form A	*Mus musculus*	R_8	CDS	Gln
PTCMYC	c-myc proto-oncogene	*Pan troglodytes* (chimpanzee)	R_5	CDS	Gln
REFCVMYC	v-myc gene	*Feline leukemia* virus	R_5	CDS	Gln
RNANDREC	Androgen receptor	*Rattus norvegicus* (rat)	$R_4TRTR_{11}TR_3$	CDS	Gln
RNCBFB	CCAAT binding transcription factor-B subunit	*R. norvegicus*	R_5	CDS	Gln
RNNGFIA	Nerve growth factor-induced (NGFI) gene	*R. norvegicus*	R_9	CDS	Ser
RNRJG9	RJG-9 gene for c-jun	*R. norvegicus*	R_5	CDS	Gln
SCADR6	ADR6	*S. cerevisiae*	R_6	CDS	Gln
SCHAP2	HAP2	*S. cerevisiae*	R_7	CDS	Gln
SCINO4X	INO4	*S. cerevisiae*	R_5	CDS	Ser

stretch is 18 in TBP from HeLa cells and 14 in TBP from Namalwa cells (figure 1). The polymorphism could reflect an instability at this locus in the general population. Analysis of TBP from 48 unrelated individuals has shown that the PCR products encompassing the CAG repeat region of this gene are polymorphic in length (Polymeropoulos *et al* 1991). Perhaps a narrow range of length variation is allowed in the TBP gene, but expansion beyond a certain number could be deleterious.

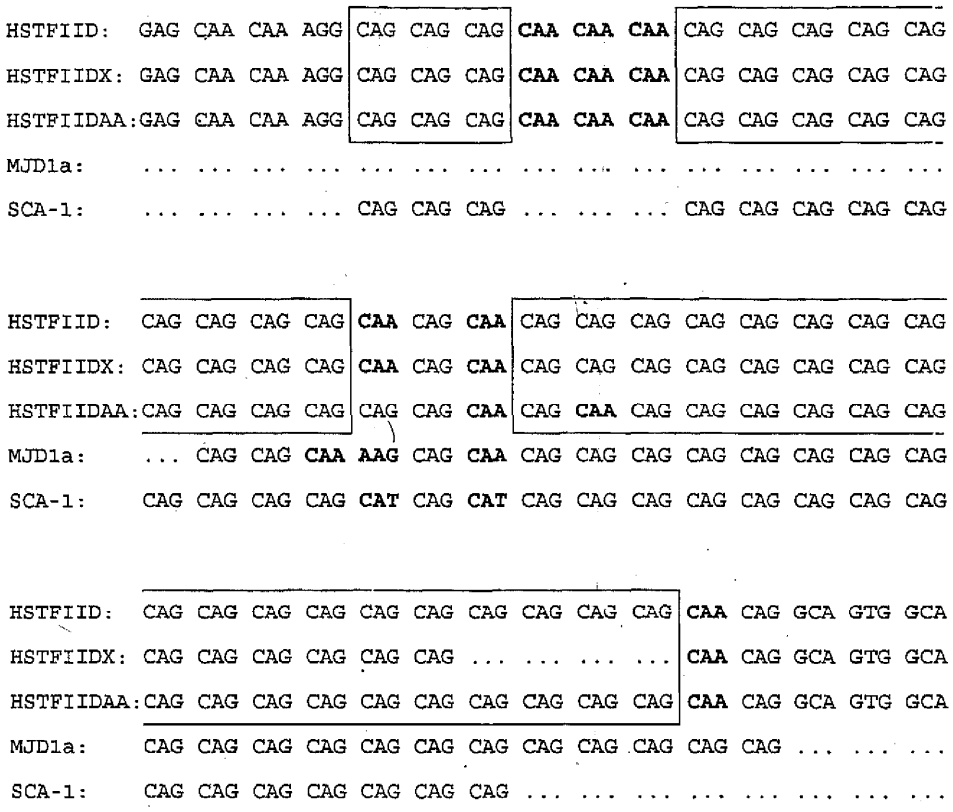

Figure 1. The CAG repeat region of hTBP derived from HeLa cells (HSTFIID), Namalwa cells (HSTFIIDX), and human hypothalamus (HSTFIIDAA), is compared with the CAG repeat of the MJDla and SCA-1 loci. Gaps have been introduced to align the sequences. The three CAG repeat stretches in hTBP are boxed.

4. Discussion

We have identified 121 human DNA sequences that contain a stretch of five or more CAG or CTG repeats. This list encompasses a variety of different proteins from different tissues, with diverse functions. Apart from the five loci known to be associated with triplet repeat expansion related diseases, i.e., androgen receptor, DRPLA, SCA-1, HD and myotonin kinase, we found 15 loci that have 10 or more tandem CAG/CTG repeats. Of these, 7 are random DNA sequences identified by screening of cDNA libraries (Li *et al* 1995;-Armour *et al* 1994; Phillips *et al* 1995). Other than these, several genes of known function, namely, AF-9, achaete-scute homologous protein, human estrogen receptor, myocyte- specific enhancer factor, human homologue of *Drosophila* brm, and human TATA-box binding protein, contain stretches of $(CAG/CTG)_{\geq 10}$. There are also some sequences that contain long stretches of interrupted CAG repeats, which are usually in frame; that is, the repeats are interrupted, in most cases, by triplets.

Earlier studies have suggested that interruption of a repeat stretch serves to stabilize the repeat region. It has been shown that for the SCA-1 locus, 98% of normal alleles show interruption of the CAG repeat by CAT triplets, while disease alleles consist of pure CAG repeats (Chung *et al* 1993). Also, in the case of the FRAXA locus, the CGG repeat is interrupted, at least in normal alleles, by one or two AGG triplets. The DNA sequence in the region containing AGG has been shown to be relatively invariant. Most of the variation at the FRAXA locus occurs at the 3' end, where the longest tracts of uninterrupted CGG are found (Kunst and Warren 1994). In the case of MJD, the CAG repeats contain two variations on the CAG motif at three positions, towards the 5' end of the repeat. CAG expansions in patients are seen to occur on the 3' side of the variant sequences (Kawaguchi *et al* 1994). From figure 1 it can be seen that human TBP has three CAG repeat stretches that are bordered by CAA triplets. It is possible that the presence of CAA triplets prevents expansion of the $(CAG)_3$ repeat located at the 5' end and the $(CAG)_9$ stretch in the centre of the repeat region. However, a polymorphism is evident in the long stretch of CAG triplets at the 3' end. The absence of intervening CAA triplets within this stretch may allow variation in repeat length by expansion or contraction during DNA replication. From these observations one can speculate that intervening triplets may serve to stabilize repeat region DNA and preclude expansion of the microsatellite.

Our observation is in agreement with the strand slippage model for expansion of repetitive DNA sequences (Streisinger *el al* 1966). It has been suggested that the loss and gain of repeat units occurs because of errors resulting from strand slippage during DNA replication, which remain uncorrected as a result of defective post replication heteroduplex repair (Strand *et al* 1993). In this picture, if a stretch of CAG repeats is interrupted by other triplets, slippage and looping out of this region would not be allowed, since this would result in extensive mismatch in the loop region and instability of the loop (Kang *et al* 1995). Therefore, expansion would be allowed only at long uninterrupted CAG repeats. It is also observed that the intervening triplets differ from the parent triplet repeat usually by a single base change. A simple explanation for this would be that these interrupting triplets arose during the course of evolution by mutation of the ancestral CAG/CTG triplet, perhaps during incorrect repair following strand slippage, and that these mutants were selected for because they stabilized the length of the repeat region.

Many genes encoding transcription factors have been found to contain long CAG/CTG repeats. It is known that the transcriptional activation domains of many transcription factors contain polyglutamine tracts, and that the length of these tracts may modulate transcription factor activity (Gerber *et al* 1994). It is postulated that these amino acid tracts serve as interfaces for intermolecular protein-protein interactions during transcription activation (Hoffmann *et al* 1990; Stott *et al* 1995). However, our finding that the CAG repeat motif occurs in all three reading frames in transcription factors, implies that stretches of polyserine, and polyalanine are also encoded by this repeat. Therefore in addition to the involvement of CAG encoded polyamino acid tracts in protein-protein interactions, the CAG repeat motif may also play a regulatory role at the DNA level.

It has been hypothesised that tandem repeats may modulate gene expression by variation in their copy numbers (Trifonov 1989; Tripathi and Brahmachari 1991). We have observed that CAG repeats are found only in the transcription factors of

eukaryotes, whose DNA is organized into nucleosomes, and are not found in prokaryotic transcription factors. This suggests that CAG repeats may play a role in histone-DNA interactions. Wang *et al* (1994) have shown that long CTG repeats may act as strong nucleosome positioning signals. It has been shown that expansion of a CTG repeat alters the DNAase I hypersensitivity of the repeat region, perhaps due to a change in nucleosome positioning (Otten and Tapscott 1995). These findings suggest that CAG repeats may serve as nucleosome phasing signals, thereby determining promoter accessibility, and thus regulating transcription of the gene. It is possible that length polymorphism of the repeat is able to affect gene expression by changes in nucleosome affinity. Certain genes such as transcription factors require finely regulated levels of expression. Down regulation of transcription may be achieved by increased binding to histones due to the presence of CAG repeats in these genes, and this may be reversed by histone phosphorylation.

Neurodegenerative disorders that are caused by expansion of CAG repeats within the coding region involve a loss of function, which in some cases has been correlated with decreased mRNA levels (Fu *et al* 1993). Experimental work carried out in our laboratory has shown that the presence of CTG repeats within a reporter gene results in a decrease in the level of gene expression in *Escherichia coli* and yeast, along with a decrease in mRNA level (Brahmachari *et al* 1995). Duplex DNA fragments containing the CTG repeat were seen to exhibit enhanced electrophoretic mobility with increase in repeat length, suggesting that adoption of a compact structure could be responsible for decreased gene expression *in vivo*. Spectroscopic and gel electrophoretic studies carried out in our laboratory on single stranded CAG and CTG repeat sequences show that these oligonucleotides adopt a compact structure. Mitas *et al* (1995) have shown that a single stranded oligonucleotide containing 15 CTG repeats forms a hairpin structure. In addition, a number of RNA binding proteins, some of which recognize hairpin structures, have been identified in eukaryotic systems (Chen and Frankel 1994). Richards *et al* (1993) have shown the existence of a protein that specifically binds CGG repeats. It is quite possible that there may be a similar CAG repeat binding protein Formation of secondary structure at the DNA and/or mRNA level, and interaction with specific binding proteins may regulate the efficiency of transcription and/or translation of genes containing CAG/CTG repeats.

In the light of our findings in this database search, and experimental evidence, we hypothesise that apart from the involvement of CAG encoded poly amino acid tracts in protein function, the CAG repeat motifs also play a critical role as *cis*-acting regulatory elements at the DNA and the mRNA level. Increased histone binding of CAG containing regions may serve to autoregulate the expression of genes in which they are present. We find that long CAG/CTG repeats occur at many loci, such as the estrogen receptor, myocyte specific enhancer factor, AF-9, TATA binding protein, and in many sequences of unknown function. In addition, polymorphism is observed at the TBP locus. This suggests that in addition to the loci known to be associated with disease, CAG repeats may show polymorphism at many other loci. Slight variations in repeat length at these regions may lead to quantitative differences in the level of expression of these genes, which in turn may be manifested as subtle or overt differences in phenotype. It is possible that the loci at which we have now shown the existence of triplet repeats may be associated with repeat length expansion related diseases that are as yet unidentified.

Acknowledgements

The authors thank Ms. Shyamla at Distributive Information Centre (IISc) and Harish Rao for their assistance in database analysis. We thank Dr N V Joshi, Centre for Ecological Sciences (IISc), for his invaluable help in statistical analysis of the data. SKB wishes to thank Department of Biotechnology, New Delhi, for financial support. We acknowledge the help rendered by T R Raghunand and S Unniraman in data tabulation and compilation. Our thanks are due to S Raghavan and U S Shaligram for their helpful suggestions and critical comments.

References

Armour J A L, Neumann R, Gobert S and Jeffreys A J 1994 Isolation of human simple repeat loci by hybridisation selection; *Hum. Mol. Genet.* **3** 599-605

Brahmachari S K, Meera G, Sarkar P S. Balagurumoorthy P, Tripathi J, Raghavan S, Shaligram U S and Pataskar S S 1995 Functional significance of simple repetitive sequences in the genome; *Electrophoresis* **16** 1705-1714

Burke J R. Kingfield M S, Lewis K E. Roses A D. Lee J E. Hulette C, Pericak-Vance M A and Vance J M 1994 The Haw River Syndrome: Dentatorubropallidoluysian atrophy (DRPLA) in an African-American family; *Nature Genet.* **7** 521-524

Chen L and Frankel A D 1994 An RNA-binding peptide from bovine immunodeficiency virus Tat protein recognises an unusual RNA structure; *Biochemistry* **33** 2708-2715

Chung M Y, Ranum L P, Duvick L A, Servadio A, Zoghbi H Y and Orr H T 1993 Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I; *Nature Genet.* **5** 254-258

Devereux J, Haeberli P and Smithies O 1984 A comprehensive set of sequence analysis algorithms for the VAX; *Nuclei Acids Res.* **12** 387-395

Fu Y H, Friedman D L, Richards S, Pearlman J A, Gibbs R A, Pizzuti A, Ashizawa T, Perryman M B, Scarlato G, Fenwick R G and Caskey C T 1993 Decreased expression of myotonin-protein kinsase messenger RNA and protein in adult form of myotonic dystrophy; *Science* **260** 235-238

Fu Y H, Pizzuti A, Fenwick R G Jr, King J, Rajnarayan S, Dunne P W, Dubell J, Nasser G A, Ashizawa T, Jong P D, Wieringa B, Korneluk R, Perryman M B, Epstein H F and Caskey C T 1992 An unstable triplet repeat in a gene related to myotonic muscular dystrophy; *Science* **255** 1256–1258

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

Han J, Hsu C, Zhu Z, Longshore J W and Finley W H 1994 Over-representation of the disease associated (CAG) and (CGG) repeats in the human genome; *Nucleic Acids Res.* **22** 1735-1740

Hoffmann A, Sinn E. Yamamoto T, Wang J, Roy A, Horikoshi M and Roeder R G 1990 Highly conserved core domain and unique N terminus with presumptive regulatory motifs in human TATA factor (TF1ID); *Nature (London)* **346** 387-390

Huntington's Disease Collaborative Research Group 1993 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes; *Cell* **72** 971-983

Kang S, Jaworski A, Ohshima K and Wells R D 1995 Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli; Nature Genet.* **10** 213-218

Kawaguchi Y, Okamoto T, Taniwaki M. Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I, Kimura J, Narumiya S and Kakizuka A 1994 CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1; *Nature Genet.* **8** 221-228

Kimura M 1983 *The neutral theory of molecular evolution* (Cambridge: Cambridge Univ. Press)

Koide R, Ikeuchi T, Onodera O, Tanaka A H, Igarashi S, Endo K, Takahashi H, Kondo R, Ashikawa A, Hayashi T, Saito M, Tomoda A, Miike T, Naito H, Ikuta F and Tsuji S 1994 Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA); *Nature Genet.* **6** 9-12

Kremer E J, Pritchard M, Lynch M, Yu S, Holman K, Baker E. Warren S T, Schlessinger D, Sutherland G R and Richards R I 1991 Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)$_n$; *Science* **252** 1711-1714

Kunst C B and Warren S T 1994 Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles; *Cell* **77** 853-861

La Spada A R, Wilson E M, Lubahn D B, Harding A E and Fishbeck K H 1991 Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy; *Nature (London)* **352** 77-79

Li S H, McInnes M G, Margolis R L, Antonarakis S E and Ross C A 1993 Novel triplet containing genes in human brain: Cloning, expression, and length polymorphism; *Genomics* **16** 572-579

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemeiya C, Jansen G, Neville C, Narang M. Barcelo J, O'Hoy K, Leblund S, Earle-Macdonald J, deJong P J, Wieringa B and Komeluk R G 1992 Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene; *Science* **255** 1253–1255

Mitas M, Yu A, Dill J, Kamp T J, Chambers E J and Haworth I S 1995 Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: (CTG)$_{15}$; *Nucleic Acids Res.* **23** 1050-1059

Orr H T, Chung M, Banfi S, Kwiatkowski T J, Servadio A, Beaudet A L, McCall A E, Duvick L A, Ranum L and Zoghbi H Y 1993 Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1; *Nature Genet.* **4** 221-226

Otten A D and Tapscott S J 1995 Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure; *Proc. Natl. Acad. Sci. USA* **92** 5465-5469

Phillips K L, Gartrell D M, Roses A D and Lee J E 1993 A triplet repeat polymorphism in a gene expressed in human hypothalamus; *Hum. Mol. Genet.* **2** 1332

Polymeropoulos M H, Rath D S, Xiao H and Merril CR 1991 Trinucleotide repeat polymorphism at the human transcription factor IID gene; *Nucleic Acids Res.* **19** 4307

Richards R I, Holman K, Yu S and Sutherland G R 1993 Fragile X syndrome unstable element. p(CCG)$_n$, and other simple tandem repeat sequences are binding sites for specific nuclear proteins; *Hum. Mol. Genet.* **2** 1429-1435

Richards R I and Sutherland G R 1992 Dynamic mutations: A new class of mutations causing human disease; *Cell* **70** 709-712

Riggins G J, Lokey L K, Chastain J L. Leiner H A, Sherman S L. Wilkinson K D and Warren S T 1992 Human genes containing polymorphic trinucleotide repeats; *Nature Genet.* **2** 186-191

Stallings R L 1994 Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implications for human genetic diseases; *Genomics* **21** 116-121

Streisinger G. Okada Y. Emnch J, Newton J, Tsugita A, Terzaghi F and Inouye M 1966 Frameshift mutations in the genetic code; *Cold Spring Harb. Symp. Quant. Biol.* **31** 77-84

Strand M, Prolla T A, Liskay R M and Petes T D 1993 Destabilisation of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair; *Nature (London)* **365** 274-276

Stott K, Blackburn J M, Butler P J G and Perutz M 1995 Incorporation of glutamine repeats makes protein oligomerize: Implications for neurodegenerative diseases; *Proc. Natl. Acad. Sci. USA* **92** 6509-6513

Sutherland G R and Richards R I 1995 Simple tandem DNA repeats and human genetic disease; *Proc. Natl. Acad. Sci. USA* **92** 3636-3641

Trifonov E N 1989 The multiple codes of nucleotide sequences; *Bull. Math. Biol.* **51** 417-432

Tripathi J and Brahmachari S K 1991 Distribution of simple repetitive (TG/CA)$_n$ and (CT/AG)$_n$ sequences in human and rodent genomes; *J. Biomol. Struc. Dyn.* **9** 387-397

Wang Y H, Amirhaeri S, Kang S, Wells R D and Griffith J D 1994 Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene; *Science* **265** 669-671

Corresponding editor: VIDYANAND NANJUNDIAH