Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 327 (2005) 276–286
BBRC
www.elsevier.com/locate/ybbrc

A novel rat genomic simple repeat DNA with RNA-homology shows triplex (H-DNA)-like structure and tissue-specific RNA expression

Indranil Dey, Pramod C. Rath*

Molecular Biology Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi-110067, India

Received 30 November 2004

Abstract

Mammalian genome contains a wide variety of repetitive DNA sequences of relatively unknown function. We report a novel 227 bp simple repeat DNA (3.3 DNA) with a d{$(GA)_7A(AG)_7$} dinucleotide mirror repeat from the rat (*Rattus norvegicus*) genome. 3.3 DNA showed 75–85% homology with several eukaryotic mRNAs due to $(GA/CU)_n$ dinucleotide repeats by nBlast search and a dispersed distribution in the rat genome by Southern blot hybridization with [^{32}P]3.3 DNA. The d{$(GA)_7A(AG)_7$} mirror repeat formed a triplex (H-DNA)-like structure in vitro. Two large RNAs of 9.1 and 7.5 kb were detected by [^{32}P]3.3 DNA in rat brain by Northern blot hybridization indicating expression of such simple sequence repeats at RNA level in vivo. Further, several cDNAs were isolated from a rat cDNA library by [^{32}P]3.3 DNA probe. Three such cDNAs showed tissue-specific RNA expression in rat. pRT 4.1 cDNA showed strong expression of a 2.39 kb RNA in brain and spleen, pRT 5.5 cDNA showed strong expression of a 2.8 kb RNA in brain and a 3.9 kb RNA in lungs, and pRT 11.4 cDNA showed weak expression of a 2.4 kb RNA in lungs. Thus, genomic simple sequence repeats containing $d(GA/CT)_n$ dinucleotides are transcriptionally expressed and regulated in rat tissues. Such $d(GA/CT)_n$ dinucleotide repeats may form structural elements (e.g., triplex) which may be sites for functional regulation of genomic coding sequences as well as RNAs. This may be a general function of such transcriptionally active simple sequence repeats widely dispersed in mammalian genome.

Keywords: Simple sequence repeats; $(GA/CT)_n$ dinucleotides; Triplex (H-DNA); RNA-homology; RNA expression; cDNAs

Approximately, 95% of mammalian genome is non-coding in nature and composed of mainly repetitive DNA from intragenic, extragenic, and gene-poor regions. About 30% of mammalian- and 50% of human genomes consist of repetitive DNA of either unknown or poorly known biological function [1,2]. However, repetitive DNA has been proposed to influence structure, function, and evolution of chromosomes in many organisms. Centromeric and telomeric repeats, retrotransposon-like elements, and *Alu*-transcripts provide examples for function of repeat sequences. Pathological expansion of trinucleotide repeats leading to hereditary neurodegenerative diseases in humans has highlighted the repeat sequences in mammalian genome [3,4]. Simple repeat sequences such as 'microsatellites' are of considerable interest due to their use as markers for genotyping, genome mapping, and species diversity. Certain simple sequence repeats are linked to regulation of gene expression. Eukaryotic genomes have a differential distribution pattern for simple sequence repeats [5] with a bias for long stretches of contiguous oligopurine tracts that may possibly influence genome function through flexibility of structure [6] leading to local variations in three-dimensional structure of DNA and interaction with specific proteins [7]. For example, poly(d(A · T)) in promoters stimulated transcription through its intrinsic DNA structure [8], $d(TG)_n \cdot d(CA)_n$ repeats formed Z-DNA conformation in vitro [9], $d(CG)_5 \cdot d(GC)_5$

* Corresponding author. Fax: +91 11 26717586.
E-mail address: pcrath@mail.jnu.ac.in (P.C. Rath).

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.12.015

sequence in the mouse c-Ki-ras gene promoter adopted *Z*-conformation under neutral pH in vitro [10], and $d(CT)_n \cdot d(GA)_n$ repeats in the *Drosophila hsp26* promoter [11] and $d(G)_{18} \cdot d(C)_{18}$ repeats in the chicken β-globin promoter [12] formed triplex (H-DNA) structure in vitro. Simple sequence repeats are also dynamic in terms of their mutation [13].

The present study reports a novel rat genomic simple repeat DNA (GenBank Accession No. X97459) containing $d\{(GA)_n \cdot d(CT)_n\}$ dinucleotides, such sequences have potential to form triplex (H-DNA)-like structure and are expressed into RNAs in a tissue-specific manner which may be of functional consequence.

Materials and methods

Animals, biochemicals, and reagents. Adult female rats (*Rattus norvegicus*); analytical grade biochemicals and molecular biology reagents (Sigma Chemical); restriction enzymes and other DNA modifying enzymes (New England Biolab); and PCR, in vitro transcription, and DNA sequencing reagents (Promega) were used for the study. Oligonucleotides (Genmed) were M13PUC primer (5′-CGCCAGGG TTTTCCCAGTCACGAC-3′) for sequencing of 3.3 DNA, Rb primer (5′-TGAGCGCGCGTAATACGACTCACTATAGGCAG-3′) and M13 reverse primer (5′-AACAGCTATGACCATG-3′) for inserting a G residue after T_7 promoter in p7SKM3.3 plasmid by PCR to facilitate in vitro transcription of 3.3 DNA by T_7 RNA polymerase. Radiolabels [α-^{32}P]dATP and [α-^{32}P]UTP, specific activity: 3000 Ci/mmol were purchased from BRIT, India.

P7SKM3.3 plasmid. Digestion of purified nuclei (chromatin-DNA) by DNase I (deoxyribonuclease I) [14] from rat skeletal muscle was carried out, the DNA fragments were purified and resolved into 10 nucleotide nucleosomal ladder by denaturing polyacrylamide gel electrophoresis [15]. A DNA bank was prepared from a gel-purified 'relatively DNase I-resistant' chromatin-DNA band of approximately 220 bp. A 227 bp DNA fragment (arbitrarily named as 3.3 DNA, GenBank Accession No. X97459) was isolated by colony hybridization as a recombinant DNA cloned at *Nsi*I site in pBlot7* vector (*one *Nsi*I site was introduced in between TATA box of the T_7 promoter and the multiple cloning site in pBluescript KS^+ plasmid, a gift from the Institute for Molecular Biology I, University of Zurich, Switzerland). The resulting recombinant plasmid was named as p7SKM3.3. 3.3 DNA was sequenced by M13PUC primer to determine the 227 bp DNA sequence. NBlast analysis was carried out at NCBI web site to find out the homology of 3.3 DNA with DNA/RNA sequences in the database.

Genomic Southern blot analysis. Genomic DNA was isolated from purified nuclei (chromatin-DNA) of rat liver, digested by *Bam*HI, *Eco*RI, *Hin*dIII, and *Pst*I and used for Southern blot-hybridization with random primed [^{32}P]3.3 DNA probe (10^7–10^8 CPM/μg DNA) [15,16]. Hybridization and washing conditions were maintained at low stringency (50 °C).

Structure of 3.3 DNA. Structure of 3.3 DNA in the plasmid state was determined as follows: (a) In the first experiment, in vitro elongation of DNA was carried out by *Escherichia coli* DNA polymerase I (Klenow fragment) and T_7 primer at 25 °C by using [α-^{32}P]dATP. In vitro PCR-amplification of 3.3 DNA was carried out at 72 °C by T_7-, T_3 primers and *Taq* DNA polymerase. In vitro synthesis of 3.3 RNA was carried out by T_7 RNA polymerase from T_7 promoter at 30 °C by using 3.3 DNA template and [α-^{32}P]UTP. The DNA and RNA products were analyzed by agarose gel electrophoresis. (b) In the second experiment, 2.5 μg of p7SKM3.3 plasmid DNA in 10 μl H_2O + 10 μl of 100 mM sodium acetate (NaOAc, pH 5.2) was mixed with 2 μl DEPC (diethyl pyrocarbonate) and incubated at 37 °C for 30 min. DNA was precipitated by ethanol and dissolved in 30 μl TE (10 mM Tris–HCl + 1 mM EDTA, pH 7.5). Modified 3.3 DNA was released by *Pvu*II, the DNA fragment was purified through agarose gel, cleaved by 100 μl of 1 M piperidine at 90 °C for 30 min, dried and washed with 200 μl H_2O. Finally, the DNA was dissolved in 20 μl of 10 mM Tris–HCl, pH 7.5, and transcribed by T_3 RNA polymerase using [α-^{32}P]UTP in vitro. Different lengths of [^{32}P]3.3 RNA produced from the cleaved-3.3 DNA template resolved by denaturing polyacrylamide gel electrophoresis. Standard dideoxy DNA sequencing reactions of the DNA template were carried out for comparison. Transcription termination sites were identified from the sequence and were used to map the DEPC + piperidine cleavage sites in 3.3 DNA in plasmid state from which the structure of 3.3 DNA was derived. (c) In the third experiment, computational analysis of 3.3 DNA sequence was carried out by dot plot online (http://antheprot-pbil.ibcp.fr/dot_matrix_plot.html), M-fold and Zucker plot (based on free energy calculation by D. Stewart and M. Zucker, 2002, Washington University, http://www.bioinfo.rpi.edu/ or http://www.bioinfo.rpi.edu/applications/mfold/old/dna/).

RNA isolation and Northern blot analysis. Total cellular RNA was isolated [17] from fresh rat tissues by extraction in lysis buffer (6 M urea, 3 M LiCl, 50 mM NaOAc, 0.1% SDS, and 200 μg/ml heparin) precipitated on ice and centrifuged at 4 °C. The pellet was washed in 8 M urea + 4 M LiCl and dissolved in buffer (200 mM NaOAc, pH 5.0, 0.2% SDS, and 1 mM EDTA), extracted by phenol:chloroform (1:1), and precipitated by ethanol. The RNA pellet was dissolved in 100 μl of H_2O. Northern blot analysis [15,16] of total cellular RNA from various rat tissues was carried out by resolving RNA in 1% formaldehyde–agarose gel, capillary blotting of RNA to nylon membrane, and hybridization of RNA-filter with random-primed [^{32}P]3.3 DNA probe or [^{32}P]cDNA probe (specific activity = 10^7–10^8 CPM/μg DNA) in formamide-hybridization buffer. Autoradiography of ^{32}P-labelled-agarose gels, ^{32}P-labelled-sequencing gels, and ^{32}P-labelled-hybridization blots was carried out by exposure to X-ray films.

Isolation of cDNAs by 3.3 DNA probe and RNA expression. A rat testis cDNA library in λgt11 vector (Clonetech) was screened three rounds by [^{32}P]3.3 DNA probe and several positive cDNA plaques were isolated. cDNA inserts in individual plaques were checked by PCR using λgt11-specific primers flanking the cloning junction. cDNA clones were validated by Southern blot hybridization with [^{32}P]3.3 DNA probe. Positive plaques were isolated from the plates and individually grown in liquid culture. cDNAs were isolated from purified genomic DNA of λgt11-cDNA clones by *Eco*RI-digestion and subcloned into *Eco*RI site of pBluescript vector. Finally, a total of 28 cDNA clones were obtained. In the present work, three rat cDNAs (pRT 4.1, pRT 5.5, and pRT 11.4) were checked for RNA expression. Random-primed [^{32}P]cDNA probes were hybridized with total cellular RNA isolated from various rat tissues by Northern blot analysis.

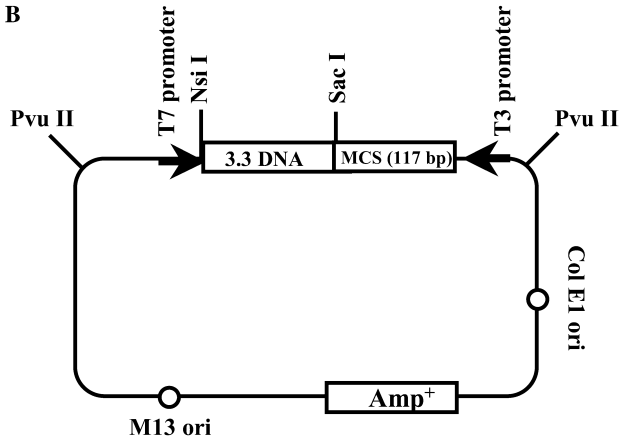
Results

DNA sequence and its homology

Fig. 1A shows 227 bp sequence of 3.3 DNA. It is purine-rich and has a mirror repeat of $(GA)_7A(AG)_7$ and GACA, $(CA)_2$, $(GAAA)_{10}$, $(AGAGAA)_4$, and $(AGAGGAAA)_2$ simple repeats, polypurine $(GA)_n$ in one strand, and polypyrimidine $(CT)_n$ in the complementary strand. Fig. 1B shows 3.3 DNA flanked by T_7 and T_3 promoters in p7SKM3.3 plasmid. In Fig. 2, rat genomic DNA digested by *Bam*HI, *Eco*RI, *Hin*dIII, and *Pst*I shows repetitive DNA bands, and heterogeneous DNA

A
1 22 26 37 43
*GAAA*GAG*GAAA*TAACATA*CACAGACAGAGAAA*CA*GAAAGACACA*
51 65 79
*GAGAGA***GAGAGAGAGAGA***GAAA***GAGAGAGAGAGAG**CGCTGCAGAGATGCT
109 140
*GAAA*CAAAAACAGAT*AAAG*AACA*GAAA*GAG*GAAA*GCACATACACAGTGA*GACA*
150 173 190 200
*GAGAAA*GA*GAAAGACACACACACACAC*TAATTTGGAAGGGTAAT*GAGA*TGGGCA
220 227
AGGGTATC*CAGAGAGA*CCCCTAGGAT

3.3 DNA sequence



p7SKM3.3 plasmid

Fig. 1. 3.3 simple repeat DNA (X97459). (A) 3.3 DNA sequence and (B) map of p7SKM3.3 recombinant plasmid. d$(GA)_7$A$(AG)_7$ mirror repeat is underlined and other repeats are in italics. Nucleotide numbers are referred in the results for Fig. 4C. Amp, ampicillin resistance; MCS, multiple cloning site.

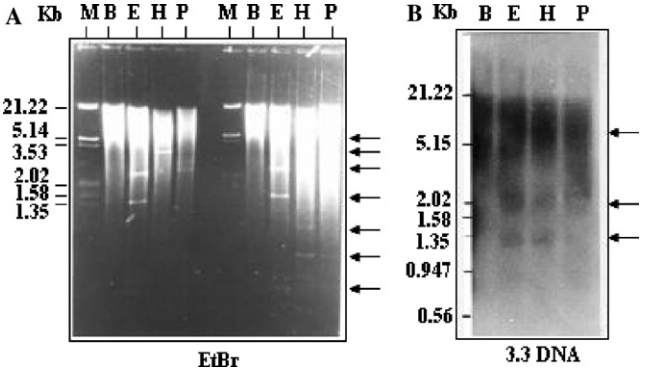


Fig. 2. Southern blot analysis of rat genomic DNA digested by *Bam*HI (B), *Eco*RI (E), *Hin*dIII (H), and *Pst*I (P), and hybridized with [^{32}P]3.3 DNA probe. (A) Ethidium bromide (EtBr) stained agarose gel and (B) hybridization with 3.3 DNA probe. Arrow indicates repeat DNA elements. bp, base pair.

fragments after agarose gel electrophoresis (a). After Southern blot hybridization under low stringency condition (50 °C), [^{32}P]3.3 DNA strongly detected DNA fragments of heterogeneous size producing a smear and weakly detected *Eco*RI, *Hin*dIII, and *Pst*I DNA fragments (b), indicating that sequences like 3.3 DNA are dispersed throughout the rat genome. Nucleotide sequence homology of 3.3 DNA by nBlast search showed many related genomic simple repeat DNA sequences and at least 31 different mRNA/protein/cDNA/exon sequences from various eukaryotic genomes primarily due to homology with $(GA/CU)_n$ dinucleotide repeats (Table 1). These 31 mRNAs from the rat, mouse, human, and *Drosophila* showed 75–89% homology with $(GA)_n$ or $(CU)_n$ dinucleotide repeats, suggesting that DNA strands containing either $(CT)_n$ or $(GA)_n$ dinucleotide repeats were used as template for transcription of RNA from the corresponding genes. Such repeats occurred either in amino acid coding regions (exons) or in 5′/3′ untranslated region (5′/3′ UTR) of mature mRNAs. Thus, $(GA/CT)_n$ repeat has widespread occurrence in the coding compartment of eukaryotic genome. Due to their polypurine/polypyrimidine nature such sequences may possess intrinsic structural property.

Structure of 3.3 DNA in vitro

Oligopurine sequences are known to form alternative non-B structures. We, therefore, checked the possible structure of 3.3 DNA. When *Sac*I linearized

Table 1
Examples of homology of various mRNAs/cDNAs containing simple $(GA)_n$ or $(CU/T)_n$ dinucleotide repeats from different species with the 3.3 DNA sequence

No.	Feature	% homology	Sequence
1	Rat thymocyte mRNA cell surface protein	83	CT
2	*Drosophila melanogaster* garnet mRNA	78	GA
3	*c-maf* (proto-oncogene) mRNA	86	GA
4	*Catharanthus roseus* mRNA for CYP P450	81	GA
5	Mouse Evi-1-zinc finger protein gene, exon 1	86	CT
6	*Mus musculus* mRNA for retinoid X receptor-γ.	88	GA
7	Mouse *Hox-2.5* mRNA	88	CT
8	*Rattus* sp. zinc finger protein RIZ mRNA	83	GA
9	Brn-3c class V pou transcription factor	87	CT
10	Human fetal brain cDNA 5′ end	87	GA
11	Rat catalase gene exon 1	80	CT
12	Rat neuropeptide gamma gene exons 1 and 2	86	GA
13	Mouse interleukin-3 receptor mRNA	87	CT
14	Human HMG I-C mRNA	82	CT
15	Rat brain mRNA for sodium channel protein II	85	CT
16	Rat genes for Leu, Glu, Asp, and Gly specific tRNA	76	CT
17	*Carassius auratus ras* gene exon 1	75	CT
18	Rat *fibrinogen gamma chain* gene exons 1–4	87	GA
19	*Rattus norvegicus* lung surfactant protein-c gene	88	CT
20	Rat β-*casein* gene, exon-1	84	GA
21	Rat *pyruvate kinase* gene, exons 3–9	82	CT
22	Rat *elastase-II* gene, exons 1–2	81	CT
23	Rat mRNA.	89	GA
24	Human complete GS2 mRNA	77	CT
25	Rat β-1/3 retinoic acid 5′region, exons 1–4	79	CT
26	γ-Aminobutyric acid receptor, α-2 subunit (human, fetal brain mRNA)	80	CT
27	*R. norvegicus* clone ndf22 neu differentiation factor mRNA	82	CT
28	*R. norvegicus* mRNA for 6-phospho-2 kinase exon 1 (a–d)	77	CT
29	*R. norvegicus* E-selectin (*ELAM-1*) mRNA	77	GA
30	Rat seven transmembrane helix receptor mRNA	86	GA
31	Rat calcitonin receptor like (*CRLR*) mRNA	86	GA

p7SKM3.3 plasmid DNA was used as template for DNA synthesis by *E. coli* DNA polymerase I (Klenow fragment), T_7 primer, and [α-^{32}P]dATP at 25 °C in vitro, 3.3 DNA was extended from 5′ end but it prematurely terminated at two major sites producing [^{32}P]3.3 DNA fragments of approximately 23 and 56 nt (Fig. 3A, lane 2). By sequence analysis, these two ''Klenow-fall-off'' sites were identified to be **G** in ACACA**G**ACA GA at 23 nt position and **A** in GAGAG**A**GAGAA sequence at 56 nt. The **G** and **A** corresponded to two possible termination sites. A control DNA (4.3 DNA) template of approximately same size but without any $(GA/CT)_n$ repeat generated full length DNA under identical conditions (lane 1). Fig. 3B shows that full length 3.3 DNA (along with 436 bp multiple cloning site of pBluescript vector) was synthesized by *Taq* DNA polymerase at 72 °C using Rb and M13 reverse primers by PCR, indicating that the previously observed inhibition of DNA chain elongation was abrogated at higher temperature. Full length 3.3 RNA (239 nt) was synthesized by T_7 RNA polymerase from T_7 promoter at 30 °C in vitro (Fig. 3C). DNA template for transcription was generated from *Sac*I linearized p7SKM3.3 plasmid by PCR using Rb and M13 reverse primers. This indicated that unlike *E. coli* DNA polymerase I (Klenow fragment), T_7 RNA polymerase could overcome the possible structural barrier in 3.3 DNA.

Figs. 4A and B show analysis of 3.3 DNA structure by DEPC-modification of P7SKM3.3 plasmid DNA and piperidine-cleavage of the gel-purified *Pvu*II fragment containing 3.3 DNA followed by its transcription by T_3 RNA polymerase from T_3 promoter (117 nt upstream of the 227th nt of 3.3 DNA). The 448 bp *Pvu*II fragment from the untreated vector (pBluescript KS+) (lane 1), and the 675 bp *Pvu*II fragment from the DEPC-treated p7SKM3.3 plasmid (lane 2) produced full length [^{32}P]RNA for the vector and 3.3 DNA, respectively. An additional larger transcript was generated which corresponded to initiation from upstream of *lac* promoter. However, when DEPC-modified and piperidine-cleaved *Pvu*II fragment from p7SKM3.3 plasmid was used, [^{32}P]3.3 transcripts terminated at several sites (as indicated for lane 3). Sequencing reactions from the T_3 primer were run in parallel to determine the size of these transcripts. The major termination sites corresponded between 26–43, 51–109, 140–190, and 220–237 nt of 3.3 DNA (lane 3). The 237 nt transcript coincided to 10 nt downstream from the 3′ end of 3.3

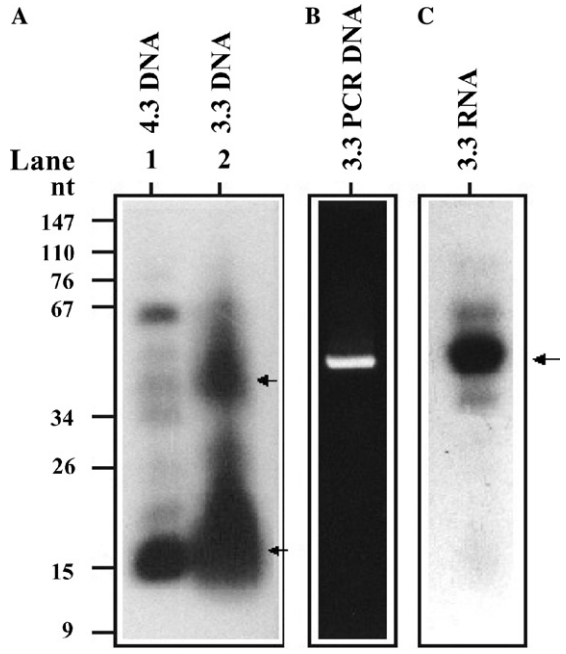


Fig. 3. Structure of 3.3 DNA. (A) Termination of 3.3 DNA chain elongation by *E. coli* DNA polymerase I (Klenow fragment) at 25 °C. Arrows indicate major termination sites at 23 and 56 nt of 3.3 DNA. (B) Full length (346 bp) 3.3 DNA amplified by *Taq* DNA polymerase at 72 °C by PCR. (C) Full length (227 nt) [^{32}P]3.3 RNA synthesized in vitro by T_7 RNA polymerase at 30 °C. nt, nucleotide, bp, base pairs.

DNA (i.e., in the vector DNA). DEPC-modification alone produced full length transcript from 3.3 DNA with very few terminations (lane 2), indicating that piperidine cleavage was necessary for the terminations. The transcriptional termination sites (lane 3) corresponded to G/A residues modified by DEPC due to exposure of the single stranded region of 3.3 DNA sequence in the plasmid state followed by piperidine cleavage. As shown in Fig. 4C in the d{$(GA)_7A(AG)_7$} mirror repeat of 3.3 DNA, the second stretch of d$(AG)_7$, i.e., (66–79 nt) was most likely single stranded. The stretch of $(TC)_7$ at 66–79 nt of the complementary strand possibly folded back on the first stretch of $(GA)_7$ at 51–64 nt to form a triplex (H-DNA) structure involving **C** · **G** · C and **T** · **A** · T triads, where **C** · **G** and **T** · **A** were held together by Watson–Crick base pairing, and **G** · C and **A** · T were held together by Hoogsteen base pairing leaving the A/T (64 nt) as a point of folding back. Although 31 (A) and 43 (C) nt positions as well as 25 (C) and 49 (G) nt positions are not identical bases, a second triplex (H-DNA) structure could also be formed by 22–36 nt (AGA**C**AGAGA**A**ACAGA) of 3.3 DNA when compared with 38–52 nt (AGACA**C**AGAGA**G**A GA) of the same strand taking the 37 nt (A) as the center of the mirror repeat. This would mean that at least two types of intra-molecular triplexes are possible in 3.3 DNA sequence suggesting how the 23 and 56 nt. DNA fragments fell off the Klenow enzyme as shown in Fig. 3A (lane 2) and confirmed by transcription termination experiment (Fig. 4B). The triplex (H-DNA) structure of 3.3 DNA might be favored by negative supercoiling of the plasmid.

Computational analysis

Computational analysis showed that 3.3 DNA could potentially form a self-folding stem–loop structure (Fig. 5) as predicted by dot plot (A), RNA fold (B), and Zucker plot (C) programs. Homologous sequences within 3.3 DNA can help self-folding. A diagonal string of dots corresponds to a similar stretch in the DNA while isolated dots represent random matches and are insignificant (A). 3.3 DNA has the potential to form self-folding structure between 42–58 and 64–74 nt. making a stem–loop structure as predicted by M-fold program (B). A model for a stem–loop structure was predicted for 3.3 DNA sequence by Zucker plot (C). Large loops are shown with small stems representing intra-strand hydrogen bonding. Two loops could be formed and some small bulges were also possible. These loops and bulges in case of RNA derived from 3.3 DNA may provide sites for interaction with proteins.

RNA expression for 3.3 DNA

Expression of RNA homologous to 3.3 DNA from rat genomic DNA was checked by Northern blot analysis (Fig. 6). Total cellular RNA from rat brain, liver, and skeletal muscle was resolved in denaturing agarose gel (A) and hybridized with [^{32}P]3.3 DNA probe (B). Interestingly, two large RNAs of 7.5 and 9.1 kb were detected (b) specifically in brain (lane 1) but not in liver (lane 2) and skeletal muscle (lane 3). Signals from 5, 10 pg of 3.3 DNA (b) or expression of α-actin mRNA by [^{32}P-α]actin PCR–DNA probe in all three tissues (C) served as hybridization controls. Self-hybridizations of 1, 5, and 10 pg of 3.3 DNA (D) and 5, 10, and 50 pg of α-actin DNA (E) by dot blot hybridization are also shown. This showed that genomic sequences like 3.3 DNA are transcribed into RNA in rat brain under in vivo condition.

Isolation of cDNAs by 3.3 DNA and RNA expression for the cDNAs

In order to assess biological significance of transcripts containing sequences homologous to 3.3 DNA in rat tissues, we screened a rat testis λgt11 cDNA library by [^{32}P]3.3 DNA probe and isolated several cDNA clones. Three cDNAs (pRT 5.5, pRT 4.1, and pRT 11.4) were checked for RNA expression in rat tissues using [^{32}P]cDNA probes. Fig. 7 shows RNA expression by Northern blot analysis for pRT5.5 (A), pRT4.1 (B),

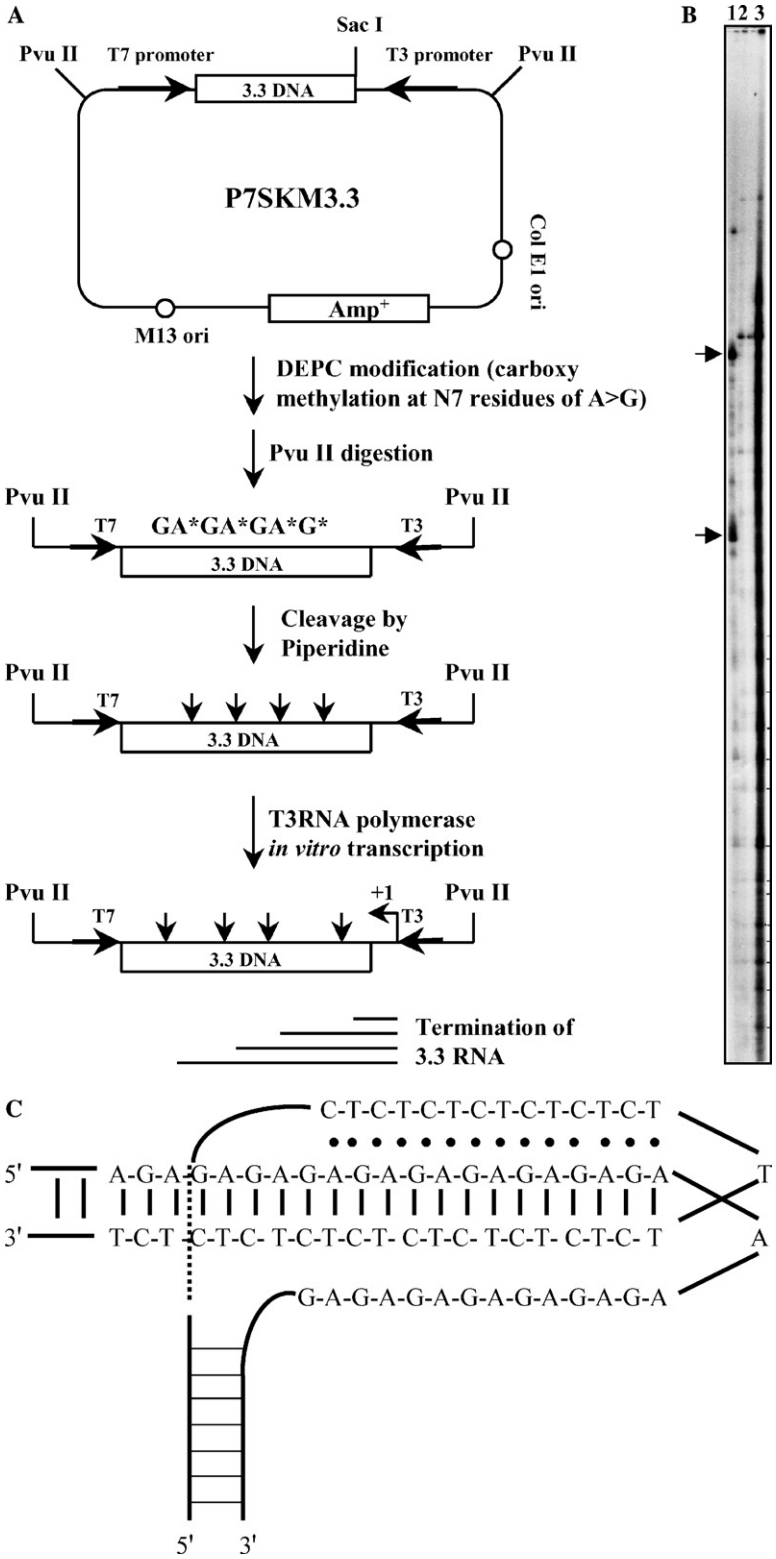


Fig. 4. Determination of 3.3 DNA structure. (A) Schematic representation of the assay. * Carboxymethylation at G/A by diethyl pyrocarbonate (DEPC), ↓ piperidine cleavage, +1 is transcription start site. (B) Termination of T_3 RNA polymerase-mediated transcription of 3.3 DNA (*Pvu*II fragment) after DEPC + piperidine cleavage. Transcription started 117 nt upstream of 3′ end of 3.3 DNA. *Pvu*II DNA templates are from: lane 1, unmodified vector (pBluescript KS+); lane 2, DEPC-modified; and lane 3, DEPC-modified, piperidine-cleaved p7SKM3.3 plasmid. Major full length transcripts are marked by arrow for lanes 1, 2 and transcription termination sites (calculated from DNA sequencing reactions) are marked for lane 3. (C) A probable intra-molecular triplex (H-DNA) structure by GA/CT mirror repeat showing interaction of $(GA)_7$ in duplex with $(CT)_7$ strand leaving $(AG)_7$ single stranded. | indicates Watson–Crick base pairing and ● indicates Hoogsteen base pairing. The $C \cdot G \cdot C$, and $T \cdot A \cdot T$ triads represent triplex (H-DNA) structure.

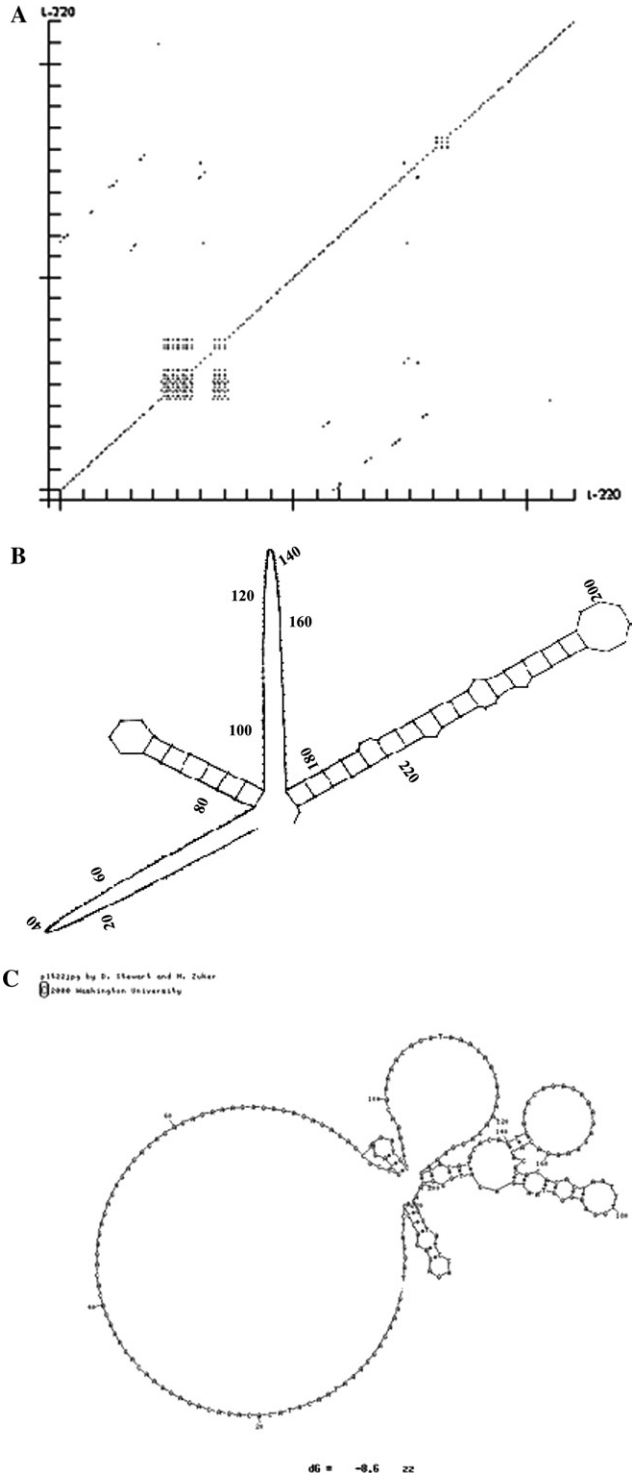


Fig. 5. Predicted structure of 3.3 DNA and 3.3 RNA. (A) Dot plot showing self-folding structure of 3.3 DNA. (B) Stem–loop structure of 3.3 RNA as determined by RNA fold (M-fold) program. (C) Zucker plot of 3.3 DNA.

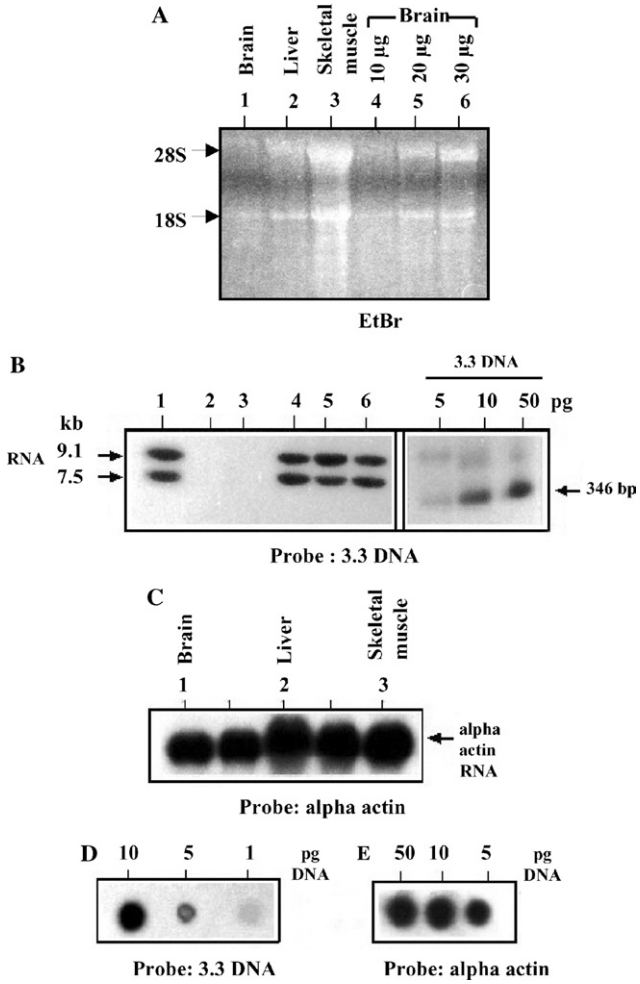


Fig. 6. Tissue-specific RNA expression in rat brain detected by [^{32}P]3.3 DNA probe. (A) Ethidium bromide (EtBr) stained 1% agarose gel showing 30 µg total cellular RNA from rat brain (lane 1), liver (lane 2), and skeletal muscle (lane 3), 28S and 18S rRNAs are marked. (B) Northern blot analysis of total cellular RNA from rat tissues was carried out with [^{32}P]3.3 DNA probe. Two RNAs (9.1 and 7.5 kb) were detected in brain (lane 1) but not in liver (lane 2) and skeletal muscle (lane 3). Lanes 4, 5, and 6 show the 9.1 and 7.5 kb RNAs from 10, 20, and 30 µg RNA from brain. 5, 10, and 50 pg PCR-amplified 3.3 DNA was used as positive control for hybridization. (C) ^{32}P-labelled-chicken α-actin DNA probe was used to detect α-actin mRNA as internal reference. (D) 1, 5, and 10 pg of 3.3 DNA and (E) 5, 10, and 50 pg chicken α-actin DNA were used for dot blot self-hybridization controls. kb, kilo bases.

and pRT 11.4 (C) cDNAs along with RNA gel for loading control (D), and hybridization with α-actin DNA as positive control (E). Total cellular RNA from rat brain, kidney, liver, lungs, skeletal muscle, and spleen was used. Fig. 7A shows that pRT 5.5 cDNA probe detected strong expression of a 2818 nt RNA in brain (lane 2) but a 3981 nt RNA in lungs (lane 5) while rest of the tissues were negative. Self-hybridization of pRT 5.5 cDNA is shown in lane 1. Similarly, Fig. 7B shows that pRT 4.1 cDNA detected strong expression of a 2398 nt RNA both in brain (lane 8) and spleen (lane 13) but weak expression in kidney (lane 9) and liver (lane 10), while rest of the tissues were negative. Interestingly, both pRT 4.1 (lane 14) and pRT 11.4 (lane 15) showed

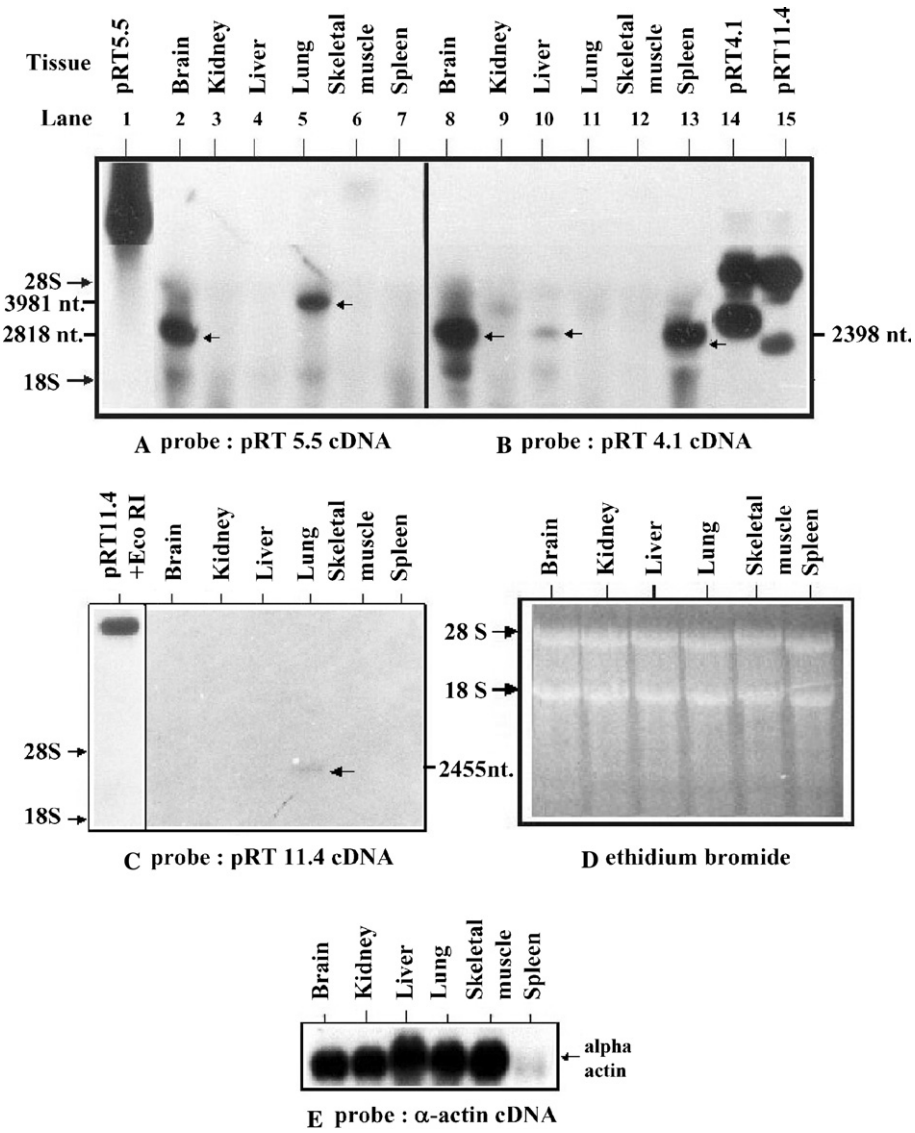


Fig. 7. Tissue-specific RNA expression in rat brain, lungs, and spleen for three rat cDNAs isolated by 3.3 DNA probe. Northern blot analysis of total cellular RNA from rat brain, kidney, liver, lungs, skeletal muscle, and spleen was carried out with [^{32}P]cDNA probes: (A) pRT 5.5; (B) pRT 4.1; and (C) pRT 11.4, respectively. Self-hybridization controls of the cDNAs are shown for pRT 5.5 (lane 1) in (A), pRT 4.1 (lane 14), and pRT 11.4 (lane 15) with pRT 4.1 in (B) and pRT 11.4 (lane 1) in (C). (D) One percent of formaldehyde–agarose gel stained by ethidium bromide (EtBr) showing 30 μg total cellular RNA from various rat tissues as RNA loading control. (E) ^{32}P-labelled-chicken α-actin DNA probe was used to detect α-actin mRNA as internal reference. 28S and 18S rRNAs are marked in (A–D). nt, nucleotides.

hybridization with pRT 4.1 cDNA probe, possibly due to repeat sequences common to both. In Fig. 7C, pRT 11.4 cDNA detected weak expression of a 2455 nt RNA only in lungs while rest of the tissues were negative. Self-hybridization of the cDNA is shown in lane 1. 28S and 18S ribosomal RNAs (A–D) are indicated as loading controls (D). Tissues negative for cDNA probes showed positive expression of α-actin mRNA (E). These results showed tissue-specific RNA expression for three cDNAs (pRT 4.1, pRT 5.5, and pRT 11.4) isolated by 3.3 DNA probe in the rat. Such RNA expression may be regulated by transcriptional control of the corresponding genomic regions.

Discussion

Simple repeat sequences in mammalian genome

Mammalian genome has abundant repetitive DNA. Repetitive non-coding sequences provide a 'genomic environment' in which genes are organized, gene expression is regulated, and gene sequences are genetically, and epigenetically modified. Simple repeat sequences are of diverse types and they are dispersed in mammalian genome. Polymorphic simple repeats mark genomic positions independent of its coding potential. Due to complementary sequences at different locations of chro-

mosomes, simple repeats may promote recombination, and DNA rearrangement during evolution. Simple repeats have potential to form non-B-DNA structures which may provide recognition sites for nucleic acids, proteins, and drugs. Such structures may also regulate movement of DNA-, RNA polymerases, and other protein complexes in chromosomes. Transcriptionally active simple repeat sequences are of special importance because the repeated RNA motif can be more versatile in structure than its DNA counterpart. Such structures may not only regulate the RNA itself but also regulate the function of other RNAs. Simple repeats in RNAs corresponding to protein coding regions may be of special interest due to the repeated amino acid motif(s) it codes in the proteins. This will influence both structure and function of the proteins. Simple repeats in coding regions may expand by "slippage" during DNA replication leading to pathological mutations in the genes. In such abnormal situation, it will not only affect the function of the protein but also deplete the amino acid pool thereby causing secondary metabolic problems. Such altered proteins may elicit pathological response in cells leading to genetic diseases. Simple repeats in promoters may influence binding of transcription factors and gene expression. Organization of nucleosomes on DNA may also be different in simple repeat sequences leading to variations in chromatin structure and function. In mammalian genome, $(TG/CA)_n$ repeats are widespread, conserved in many loci, associated with nucleosomes, and possibly play a role in chromatin organization. They have potential for Z-DNA conformation and may absorb superhelical stress during transcription [18]. Human telomeric repeats, $(TTAGGG)_n$, adopt a quadruplex (G4-DNA) structure by folding of double stranded GC-rich sequence onto itself forming Hoogsteen base pairing [19]. This unusual structural motif may be involved in chromosome pairing and stability [20]. Intragenic amplification of $(CTG)_n/(CAG)_n$ trinucleotide repeats is associated with several genetic disorders, and can downregulate gene expression in vivo [21].

In the present study, we have isolated a 227 bp novel genomic simple repeat DNA (3.3 DNA) from a relatively "DNase I-resistant" chromatin-DNA fragment of rat (*R. norvegicus*) skeletal muscle (Fig. 1). 3.3 DNA-like simple repeat sequences are dispersed throughout the rat genome (Fig. 2). Besides microsatellites present in genes, 3.3 DNA showed 75–89% homology with $(GA/CU)_n$ containing mRNAs/cDNAs from the database (Table 1). Such $(GA/CU)_n$ sequences in mRNAs may occur in 5′ untranslated region (5′UTR), 3′ untranslated region (3′UTR), and in amino acid coding region, and may form secondary structures. The $(GA)_n$ sequence may code for glutamic acid (GAG) or arginine (AGA), while the $(CU)_n$ sequence may code for leucine (CUC) or serine (UCU). Secondary structures formed by $(GA)_n$ and $(CU)_n$ sequences at 5′UTR or 3′UTR of mRNAs may provide sites for regulation of mRNAs, e.g., control of half-life at post-transcriptional level and control of mRNA translation. In both cases, such secondary structures may be sites for RNA:RNA or RNA:protein interactions.

Simple repeats as units for structural organization

Our results showed that 3.3 DNA could form a triplex (H-DNA)-like structure due to $(GA/CT)_n$ repeats in plasmid state in vitro (Fig. 4). This possibly inhibited DNA chain elongation by *E. coli* DNA polymerase I (Klenow fragment) at 25 °C but not by *Taq* DNA polymerase at 72 °C or strong polymerases like T_3 or T_7 RNA polymerases at 30 °C (Fig. 3). Computational analysis of 3.3 DNA showed self-folding stem–loop structures characteristic of RNA structure and function (Fig. 5). There are many reports which argue in favor of biological significance of simple repeat sequences. $(GA/CT)_n$ tracts occur at regular intervals along vertebrate chromosomes indicating their role in structural organization of chromosomes [22]. Conserved $(CT)_n$ microsatellite loci at orthologous positions in closely related mammals indicate evolutionary significance of such sequences [23]. Although simple repeat sequences usually occur as dispersed polymorphic microsatellites, they may also occur as discrete stretches in genes as potential sites for interaction with specific proteins, e.g., nuclear proteins binding with $(GAAA)_n$ [24] and $(GA)_n$ single strand [25]. Further, Raf-1 kinase targeted GA-binding protein in HIV-1 [26]. $(GA)_n$ oligonucleotides formed stable triple helices under physiological conditions [27] and triplex (H-DNA) conformation was formed by $A \cdot A \cdot T$, $T \cdot G \cdot C$, $A \cdot G \cdot C$, and $T \cdot A \cdot T$ base triads [28], as well as a naturally occurring $(dT–dC)_{18}$ repeat [29]. Simple repeats hybridizing with $d(CAC)_n/d(GTG)_n$ oligonucleotides were expressed in human lymphocyte mRNAs [30]. Heterogeneity in distribution of $(GACA)_n$ simple repeats in primate and mouse karyotypes [31], random polypyrimidine tracts in rodent and primate genomes [28] and uniform $(CT)_n$ tract throughout the chromosome arm of ten different vertebrate species [32] argue in favor of evolutionary significance of such genomic sequences. Triplex (H-DNA)-like structure formed by 3.3 DNA in vitro may represent its potential to form such structures at chromosomal loci. However, functional significance of $(GA)_n$ or $(CU)_n$ dinucleotides in specific eukaryotic mRNAs needs further investigation.

Simple sequence repeats and RNA expression

Transcriptionally active simple repeat DNA is suitable for studying the function of such sequences. Our results (Fig. 6) showed that microsatellite sequences with 3.3 DNA-like dinucleotide repeats are not only dis-

persed in rat genome but also transcribed into RNA in rat brain under in vivo conditions. Their specific expression at RNA level in rat brain needs further investigation. Simple repeats, by their structure, can influence the function of RNA through promoter as well as 5′ and 3′ untranslated (UTR) regions. Some examples are as follows. $(GA)_n$ and $(CT)_n$ repeats in heat shock elements regulate transcriptional activation of *Drosophila hsp26* gene by binding with the transcriptional activator, GAGA-factor [33]. GAGA-factor also binds to $(GA)_n$ sequences in the 5′ upstream region of many constitutive genes. Such sequences can form triple-stranded DNA as well as other non-B-DNA conformation [20], influence transcriptional machinery, and regulate gene expression in vivo [21]. Triplex forming oligonucleotides [34] and peptide nucleic acids (PNA) [35] can inhibit gene expression at transcriptional level [36] and inhibit DNA conformational change preventing the interaction of transcription factor [37]. Intra-molecular triplex (H-DNA) can recognize certain regions in a linear DNA where DNA–protein interaction may take place [38]. Interestingly, initiation of DNA replication by DNA polymerases (T_7 and T_4 DNA polymerases, *E. coli* DNA polymerase Klenow fragment) from primers form a triple helix structure [18]. RNA structures are responsible for regulatory processes such as hairpins or pseudoknots involved in ribosomal frame-shifting and TAR-RNA element for HIV-1 Tat protein. Binding of selective ligands to regulatory RNA motifs with high affinity may serve as tools for dissecting molecular mechanisms or as prototype agents to control gene dysfunction and pathogen multiplication. For example, efficient inhibition of translation, splicing, or reverse transcription in cell-free systems, in cultured cells, or in vivo by oligomers complementary to RNA has been reported [39]. $(GA)_n/(CT)_n$ dinucleotide sequence has also been involved in chromosomal integration of functional gene array [40]. It is possible that $(GA/CT)_n$ sites in genes may be prone to pathological mutations similar to trinucleotide repeats.

Due to strong homology with mRNAs, we used [^{32}P]3.3 DNA probe to detect RNA expression in rat tissues in vivo and to isolate cDNAs from a rat testis cDNA library. We detected rat brain-specific RNA expression, and isolated 28 cDNAs from rat testis and tissue-specific RNA expression for three such cDNAs (pRT 4.1, pRt 5.5, and pRT 11.4) are reported here. Interestingly, 3.3 DNA and the three cDNA probes detected large size RNAs from rat tissues and expression patterns of the cDNAs were distinct from each other qualitatively and quantitatively. Detection of large size RNAs by the cDNAs may indicate either their protein coding potential or their regulatory role through mechanisms such as RNA-interference or RNA-mediated gene silencing [41,42] and pseudogene-RNA regulated gene expression [43] and others. Tissue-specific RNA expressions in the rat brain detected by 3.3 DNA, and in rat brain, lungs, and spleen detected by the cDNAs suggest biological significance of expression of RNA from genomic simple repeat sequences. However, further investigation is necessary to correlate the in vitro property of such sequences with in vivo function. These cDNAs may be useful candidates for studying role of simple repeat sequences in the coding compartment of mammalian genome.

Accession number of the 227 bp 3.3 simple repeat DNA in p7SKM3.3 plasmid is X97459.

Acknowledgments

The kind gift of the rat cDNA library from Dr. Y.K. Jaiswal, School of Studies in Biochemistry, Jiwaji University, Gwalior, India; financial support from the Department of Science and Technology (D.S.T.), the Council of Scientific and Industrial Research (C.S.I.R), the University Grants Commission (U.G.C), the "University of Potential for Excellence" (UPOE) grant of the U.G.C., Govt. of India, to P.C.R. as well as the "Centre of Advanced Study" (C.A.S.) grant of the U.G.C., the F.I.S.T. grant of the D.S.T. to the School of Life Sciences are gratefully acknowledged. ID was supported by J.R.F. and S.R.F. fellowships from the U.G.C.

References

[1] E.S. Lander et al., International human genome sequencing consortium initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[2] J. Venter et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[3] R.R. Sinden, Biological implications of the DNA structures associated with disease-causing triplet repeats, Am. J. Hum. Genet. 64 (1999) 346–353.
[4] P. Ferrigno, P.A. Silver, Polyglutamine expansions:proteolysis, chaperones, and the dangers of promiscuity, Neuron 26 (2000) 9–12.
[5] M.V. Katti, P.K. Ranjekar, V.S. Gupta, Differential distribution of simple sequence repeats in eukaryotic genome sequences, Mol. Biol. Evol. 18 (2001) 1161–1167.
[6] M.J. Behe, An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes, Nucleic Acids Res. 25 (1995) 689–695.
[7] H.L. Paulson, Protein fate in neurodegenerative proteinopathies:polyglutamine diseases join the (mis)fold, Am. J. Hum. Genet. 64 (1999) 339–345.
[8] V. Iyer, K. Struhl, Poly(dA:dT) a ubiquitous promoter element that stimulates transcription via it's intrinsic DNA structure, EMBO J. 14 (1995) 2570–2579.
[9] J.L. Weber, Informativeness of human $(dC–dA)_n \cdot (dG–dT)_n$ polymorphism, Genomics 7 (1990) 524–530.
[10] D.G. Pestov, A. Dayn, S.E. Yu, D.L. George, S.M. Mirkin, H-DNA and Z-DNA in the mouse c-Ki-ras promoter, Nucleic Acids Res. 19 (1991) 6527–6532.

[11] H. Granok, B.A. Leibovitch, C.D. Shaffer, S.C. Elgin, ga–ga factor over GAGA, Curr. Biol. 5 (1995) 238–241.
[12] D. Gross, W.T. Garrard, The ubiquitous potential Z-forming sequence of eukaryotes, $(dT–dG)_n \cdot (dC–dA)_n$, is not detectable in the genomes of eubacteria, archaebacteria, or mitochondria, Mol. Cell. Biol. 6 (1986) 3010–3013.
[13] C. Macaubas, L. Jin, J. Hallmayer, A. Kimura, E. Mignot, The complex mutation pattern of a microsatellite, Genome Res. 6 (1997) 635–641.
[14] M.M. Chaturvedi, M.S. Kanungo, Analysis of chromatin of the brain of young and old rats by micrococcal nuclease and DNase I, Biochem. Int. 6 (1985) 357–363.
[15] I. Dey, A study of chromatin structure and transcription using LINE DNA and simple repeat DNA probes, Ph.D. Thesis, Jawaharlal Nehru University, New Delhi, (2000).
[16] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, second ed., Cold Spring Harbor Laboratory Press, New York, 1989.
[17] C. Auffray, F. Rougeon, Purification of mouse immunoglobulin heavy chain messenger RNAs from total myeloma tumor RNA, Eur. J. Biochem. 107 (1980) 303–314.
[18] C. Rocher, R. Dalibar, T. Letellier, G. Precigoux, P. Lestienne, Initiation of DNA replication by DNA polymerases from primers forming a triple helix, Nucleic Acids Res. 29 (2001) 3320–3326.
[19] D. Sen, W. Gilbert, Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis, Nature 334 (1988) 364–366.
[20] E. Jimenez-Garcia, A. Vaquero, M.L. Espinás, R. Soliva, M. Orozco, J. Bernués, F. Azorín, The GAGA factor of *Drosophila* binds triple-stranded DNA, J. Biol. Chem. 273 (1998) 24640–24648.
[21] S.S. Ririe, R.V. Guntaka, An RNA oligonucleotide corresponding to the polypyrimidine region of the rat alpha 1(I) procollagen promoter forms a stable triplex and inhibits transcription, Biochem. Biophys. Res. Commun. 249 (1998) 218–221.
[22] A.K.C. Wong, H.A. Yee, J.H. Van de Sande, J.B. Rattner, Distribution of CT rich tract is conserved in vertebrate chromosomes, Chromosoma 99 (1990) 344–351.
[23] R.L. Stallings, Conservation and evolution of $(CT)_n/(GA)\,n$ microsatellite sequences at orthologous positions in diverse mammalian genomes, Genomics 25 (1995) 107–113.
[24] J.T. Epplen, A. Kyas, W. Mäueler, Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins, FEBS Lett. 389 (1996) 92–95.
[25] A. Aharoni, N. Baran, H. Manor, Characterization of a multisubunit human protein which selectively binds single stranded $d(GA)_n$ and $d(GT)_n$ sequence repeats in DNA, Nucleic Acids Res. 21 (1993) 5221–5228.
[26] F. Flory, A. Hoffmeyer, U. Smola, U.R. Rapp, J.T. Bruder, Raf-1 kinase targets GA-binding protein in transcriptional regulation of the human immunodeficiency virus type 1 promoter, J. Virol. 76 (1996) 2260–2268.
[27] P.B. Arimondo, T. Garestier, C. Helene, J.S. Sun, Detection of competing DNA structures by thermal gradient gel electrophoresis: from self-association to triple helix formation by (GA)-containing oligonucleotides, Nucleic Acids Res. 29 (2001) E15.
[28] J. Klysik, An intra-molecular triplex structure from non-mirror repeated sequence containing both $Py \cdot Pu \cdot Py$ and $Pu \cdot Pu \cdot Py$ triads, J. Mol. Biol. 245 (1995) 499–507.
[29] H. Htun, J.E. Dahlberg, Single strands, triple strands, and kinks in H-DNA, Science 241 (1988) 1791–1796.
[30] C. Epplen, J.T. Epplen, Expression of $(cac)_n/(gtg)_n$ simple repetitive sequences in mRNA of human lymphocyte, Hum. Genet. 93 (1994) 35–41.
[31] I. Nanda, C. Deubelbeiss, M. Guttenbach, J.T. Epplen, M. Schmid, Heterogeneities in the distribution of $(GACA)_n$ simple repeats in the karyotypes of primates and mouse, Hum. Genet. 85 (1990) 187–197.
[32] G.D. Burkholder, L.J.P. Latimer, J.S. Lee, Immunofluorescence staining of mammalian nuclei and chromosomes with a monoclonal antibody to triplex DNA, Chromosoma 97 (1988) 185–195.
[33] Q. Lu, L.L. Wallrath, H. Granok, S.C. Elgin, $(CT)_n \cdot (GT)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the *Drosophila* hsp26 gene, Mol. Cell. Biol. 13 (1993) 2802–2814.
[34] M.P. Knauert, P.M. Glazer, Triplex forming oligonucleotides: sequence-specific tools for gene targeting, Hum. Mol. Genet. 10 (2001) 2243–2251.
[35] S. Diviacco, V. Rapozzi, L. Xodo, C. Helene, F. Quadrifoglio, C. Giovannangeli, Site-directed inhibition of DNA replication by triple helix formation, FASEB J. 14 (2001) 2660–2668.
[36] B.P. Casey, P.M. Glazer, Gene targeting via triple-helix formation, Prog. Nucleic Acids Res. Mol. Biol. 67 (2001) 163–192.
[37] A. Majumdar, A. Khorlin, N. Dyatkina, F.L. Lin, J. Powell, J. Liu, Z. Fei, Y. Khripine, K.A. Watanabe, J. George, P.M. Glazer, M.M. Seidman, Targeted gene knockout mediated by triple helix forming oligonucleotides, Nat. Genet. 20 (1998) 212–214.
[38] C. Giovannangeli, C. Helene, Triplex-forming molecules for modulation of DNA information processing, Curr. Opin. Mol. Ther. 3 (2000) 288–296.
[39] J.J. Toulme, C. Di Primo, S. Moreau, Modulation of RNA function by oligonucleotides recognizing RNA structure, Prog. Nucleic Acids Res. Mol. Biol. 69 (2001) 1–46.
[40] A.D. Bailey, T. Pavelitz, A.M. Weiner, The microsatellite sequence $(CT)_n \times (GA)_n$ promotes stable chromosomal integration of large tandem arrays of functional human U2 small nuclear RNA genes, Mol. Cell. Biol. 18 (1998) 2262–2271.
[41] O.A. Kent, A.M. MacMillan, RNAi:running interference for the cell, Org. Biomol. Chem. 2 (2004) 1957–1961.
[42] V. Schramke, R. Allshire, Those interfering little RNAs! silencing and eliminating chromatin, Curr. Opin. Genet. Dev. 14 (2004) 174–180.
[43] S. Hirotsune, N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Yakahashi, K.-I. Yagami, A. Wynshaw-Boris, A. Yoshiki, An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene, Nature 423 (2003) 91–96.