Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$

Michael Mitas*, Adong Yu, Jeffrey Dill, Timothy J. Kamp[1], Eric J. Chambers[2] and Ian S. Haworth[2]

Department of Biochemistry and Molecular Biology, Oklahoma State University, 246 Noble Research Center, Stillwater, OK 74078, USA, [1]Department of Medicine, Division of Cardiology, Johns Hopkins University, 844 Ross Building, Baltimore, MD 21205, USA and [2]Department of Pharmaceutical Sciences, University of Southern California, 1985 Zonal Avenue, Los Angeles, CA 90033, USA

Received September 22, 1994; Revised and Accepted January 30, 1995

ABSTRACT

Although triplet repeat DNA sequences are scattered throughout the human genome, their biological function remains obscure. To aid in correlating potential structures of these nucleic acids with their function, we propose their classification based on the presence or absence of a palindromic dinucleotide within the triplet, the G+C content, and the presence or absence of a homopolymer. Five classes of double-stranded (ds) triplet repeats are distinguished. Class I repeats, which are defined by the presence of a GC or CG palindrome, have the lowest base stacking energies, exhibit the lowest rates of slippage synthesis [Schlötterer and Tautz (1992) *Nucleic Acids Res.*, 20, 211] and are uniquely associated with triplet repeat expansion diseases. The six single-stranded (ss) triplet repeats within Class I also have the potential to form hairpin structures, as determined by energy minimization. To explore the possibility of hairpin formation by ss Class I triplet repeats, studies were performed with a ss oligonucleotide containing 15 prototypic CTG repeats [ss $(CTG)_{15}$]. Electrophoretic, P1 nuclease and $KMnO_4$ oxidation data demonstrate that ss $(CTG)_{15}$ forms a hairpin containing base paired and/or stacked thymines in the stem. Potential functions of hairpins containing Class I triplet repeats are discussed with respect to protein translation and mRNA splicing. Further, potential roles of hairpin structures in triplet repeat expansion events are discussed.

INTRODUCTION

The term 'minisatellite' DNA was coined by Jeffreys *et al.* in 1985 to describe the occurrence of polymorphic tandem repeats of DNA that were shorter than classic satellite DNA (1). 'Microsatellite' DNA soon referred to a subset of minisatellite DNA composed of repeating units of very short oligonucleotides (~1–5 bp). Mini- or microsatellites are thought to accumulate by DNA slippage and mispairing during replication (2–5) or by extension of single-strand ends (1,6). The frequency of events leading to repeat length alteration can be as high as 10^{-4}–10^{-2}, values much higher than that of classical mutations (7). Due primarily to their location within or near genes, micro- or minisatellite DNAs were implicated in a range of functions, including gene regulation (8–10), gene recombination (11), signals for DNA binding proteins (12,13) and signals for gene conversion (1,14). However, since no clear function was ascribed to these regions of DNA, it was assumed by some that polymorphic simple sequence DNA merely reflected internal genomic mechanisms that had the tendency to dynamically produce and delete these sequences (15). Indeed, because microsatellite DNA was so polymorphic in length, its only apparent function was in DNA fingerprinting (16–18).

Recently, seven human genetic diseases [herein referred to as triplet repeat expansion diseases (TREDs)] associated with expansions of microsatellite triplet repeat sequences have been identified. TREDs are characterized by the coincidence of disease manifestation with expansion of a G+C-rich trinucleotide repeat contained within a specific gene. Continued expansion of the repeat is observed in offspring of affected individuals, resulting in increased severity of the disease and/or an earlier age of onset, phenomena clinically referred to as 'anticipation'. Two fragile X syndromes (19,20), Kennedy's disease (21,22), Huntington's disease (23), myotonic dystrophy (24,25), spinocerebellar ataxia type I (26) and hereditary dentatorubral-pallidoluysian atrophy (27) are members of this class of disease. Typically, an individual afflicted with a TRED contains greater than 50 trinucleotide repeats within a specific gene, while unaffected individuals contain between five and 30 repeats. Expansions of CGG triplet repeats are also associated with hypermethylation of the two genes containing these expansions and fragility of the X chromosome at their loci (28–30). At least two of the expanded triplet repeats are within non-coding regions of their respective genes, suggesting that these sequences may serve some type of biological function. The function of trinucleotide repeats and mechanisms leading to their expansion are among the most

* To whom correspondence should be addressed

perplexing topics in modern biology, as evidenced by the number of recent review articles and letters that address TREDs (31–40).

To aid in correlating potential structures of triplet repeat nucleic acids with their function, we first describe a sequence-based classification system for them. Class I repeats, which are defined by the presence of a GC or CG palindrome, have the lowest base stacking energies, exhibit the lowest rates of slippage synthesis (52) and are uniquely associated with TREDs. All six complementary single strands of Class I triplet repeats potentially form stable hairpin structures, as determined by energy minimization. The hairpin stability was greatest for $(CTG)_n$ and least for $(CAG)_n$. The experimental results presented below demonstrate that a ss oligonucleotide containing 15 prototypic CTG repeats exhibited rapid electrophoretic mobility. The middle of the triplet repeat region of ss $(CTG)_{15}$ was cleaved by $KMnO_4$/piperidine or P1 nuclease, suggesting that its rapid mobility was due to a hairpin structure.

MATERIALS AND METHODS

Oligonucleotides

All oligonucleotides were synthesized on an Applied Biosystems 381A oligonucleotide synthesizer (Foster City, CA) with the trityl group on and purified with oligonucleotide purification cartridges (Cruachem, Glasgow, UK). Sequences of oligonucleotides were:

$(ATC)_{15}$, $GATCC(ATC)_{15}TAGA$;
$(GAT)_{15}$, $AGCTTCTA(GAT)_{15}G$;
VR4, (upper strand) GATCCGGCTGACCGAGGGTTAAGCGCGCTGCGGCCGCGGGCCGCCCGGTTAACCCCCTCGGA;
VR4, (lower strand) AGCTTCCGAGGGGGTTAACCGGGCGGCCCGCGGCCGCAGCGCGCTTAACCCTCGGTCAGCCG;
linker, (upper strand) GATCCCCGCGGTGATCAACGCGTCTGCAGAGATCTTCTAGAACTAGTGATATCGTCGACA;
linker, (lower strand) AGCTTGTCGACGATATCACTAGTTCTAGAAGATCTCTGCAGACGCGTTGATCACCGCGGG;
$(CTG)_{15}$, $GATCC(CTG)_{15}GGTACCA$;
$(CAG)_{15}$, $AGCTTGGTACC(CAG)_{15}G$;
markers, $CGATA(CTG)_nACGTA$, where $n = 1, 3, 5$ or 7.

The sequence of VR4 was derived from the 5′-upstream region of the rat brain a_1 subunit Ca^{2+} channel gene (41) and contains two direct repeats of near perfect 8 bp palindromic sequences separated by 31 G+C-rich nucleotides. There are 16 bp and 44 hydrogen bonds in each of the hairpin structures of VR4-U and VR4-L.

Calculations of base stacking energies

Base stacking energies of the various triplet repeat sequences were determined by summing the three individual dinucleotide base stacking energies, obtained from published data (42).

Energy minimization calculations

The starting structures for the hairpins were constructed using the QUANTA 3.2.3 package (Molecular Simulations, Burlington, MA). Molecules of the sequences $5'$-$CGATA(XYZ)_nACGTA$ (where XYZ = CTG, CAG, CCG, CGG, GTC, GAC and $n = 5$, 7 or 15) were manipulated to assume a hairpin structure. The duplex part of the hairpin had two Watson–Crick G·C base pairs and a single mismatched base pair for each pair of interacting repeats. In each of the mismatches the nucleotides were retained in an *anti* conformation. The 5′- and 3′-ends of the sequence were maintained in a conformation lacking base pairs. The central trinucleotide of each sequence formed a three-membered loop. Several conformations for this loop were constructed and each subjected to 2000 cycles of minimization, with the remainder of the structure kept rigid. The lowest energy loop conformation was used in a subsequent 4000 step minimization of the whole structure. All the molecular mechanics calculations were performed on a Silicon Graphics Indigo workstation using the AMBER 4.0 force field (43–46). Standard all-atom force field parameters and STO-3G m.e.p.-derived charges were assigned to the DNA. In each of these minimizations, a distance-dependent dielectric of the form $e = 4r_{ij}$ was used to reproduce the effect of solvent on electrostatic interactions and a 12 Å, residue-based cut-off was applied to the non-bonded interactions. To provide a comparison with the hairpin energies, structures of the same sequences (initially in the conformation of a duplex single strand) lacking base pairs were minimized for 4000 steps in a similar fashion.

Plasmid DNA preparation

Five micrograms of each synthetic oligonucleotide was phosphorylated with ATP and T4 polynucleotide kinase (Boehringer Mannheim, Indianapolis, IN) in buffer containing 1 mM DTT, 1 mM ATP, 50 mM Tris–HCl (pH 7.6), 10 mM $MgCl_2$, 0.1 mM spermidine, 0.1 mM EDTA and 20 U T4 polynucleotide kinase (New England Biolabs, Beverly, MA) in a final volume of 100 µl. Complementary pairs were combined, heated to 100°C and allowed to cool to room temperature over a period of at least 4 h. Annealed oligonucleotide pairs were extracted with one volume of 25:24:1 phenol:chloroform:isoamyl alcohol (PCI), ethanol precipitated and resuspended in 40 µl H_2O. Fifteen micrograms of pBLCAT2 vector (47) was digested to completion with *Bam*HI and *Hin*dIII and treated with calf intestinal phosphatase (CIP) (Boehringer Mannheim). Reaction mixtures were extracted with PCI, ethanol precipitated and resuspended in 200 µl H_2O. Ligation reactions were performed in a 10 µl volume for at least 4 h at room temperature and contained 1 U T4 ligase (Boehringer Mannheim), 1 µl CIP-treated pBLCAT2 and 1 µl of either a 1:1, 1:10, 1:100, 1:1000 or 1:10 000 dilution of annealed oligonucleotides. Ligation mixtures were transformed into XL1-Blue competent cells (Stratagene, La Jolla, CA) and plated onto agar containing 50 µg/ml ampicillin. Plasmid DNAs containing the desired inserts were purified through CsCl gradient centrifugation for sequence analysis and probe purification. Plasmid DNAs were named according to the sequence of the cloned pyrimidine-rich strand. For example, the plasmid containing 15 CTG repeats was named pCTG15.

Probe preparation

Labeling of ss fragments for electrophoretic mobility analysis. For labeling of an oligonucleotide strand containing a pyrimidine-rich triplet repeat, 15 µg pCTG15, pATC15, plinker or pVR4 plasmid DNAs were digested with 50 U *Hin*dIII (Boehringer Mannheim) for 1 h at 37°C in a volume of 70 µl. Recessed ends were labeled at the 3′-terminus by adding 5 µl [α-^{32}P]dCTP, 5 µl [α-^{32}P]dATP (each 3000 Ci/mmol; ICN, Irvine, CA), 2.5 µl 5 mM dTTP and dGTP and 25 U Klenow enzyme (New England Biolabs). Reactions were incubated for 1 h at room temperature. Plasmid DNAs were extracted with PCI and precipitated with ethanol. Resuspended DNAs were digested with 50 U *Bam*HI in a volume of 70 µl and applied directly to a Nuctrap column (Stratagene) for further removal of unincorporated [^{32}P]dNTPs. Labeling of oligonucleotides containing purine-rich (or VR4-L and linker-L) triplet strands was performed in an identical manner except that the order of restriction enzymes was reversed. Probes were diluted to 2 × 10^4 d.p.m./µl with H_2O. For electrophoretic analysis, DNAs (4 × 10^4 d.p.m., ~1 fmol) were diluted to 10 µl in buffer containing 8% glycerol, 10 mM HEPES, pH 8.5, 50 mM KCl, 1 mM EDTA. One microliter of loading dye (50% glycerol, 0.4% bromophenol blue) was added to the DNA samples prior to gel electrophoresis. During electrophoresis, 15°C tap water was circulated through a Hoeffer SE 600 series unit (San Francisco, CA).

5′ End-labeling for P1 nuclease digestion, $KMnO_4$ oxidation and dimethyl sulfate studies. For labeling of an oligonucleotide strand containing a pyrimidine-rich triplet repeat, 15 µg of pCTG15 or pATC15 plasmid DNAs were digested with *Bam*HI and dephosphorylated with calf intestinal phosphatase. DNAs were 5′ end-labeled with ^{32}P by incubation in buffer containing 1 mM DTT, 1 µl 7000 Ci/mmol [γ-^{32}P]ATP (ICN, Irvine, CA), 50 mM Tris–HCl (pH 7.6), 10 mM $MgCl_2$, 0.1 mM spermidine, 0.1 mM EDTA and 20 U T4 polynucleotide kinase (New England Biolabs). The labeled pyrimidine-rich strand (annealed to its unlabeled purine-rich complement) was liberated from the plasmid by digestion with *Hin*dIII. Unincorporated DNA was removed by size exclusion chromatography (Nuctrap column, Stratagene). Labeled DNAs were subjected to electrophoresis in a 2% agarose gel. Oligonucleotides containing triplet repeats were excised from gels and purified from vector DNA with glass beads (Mermaid Kit; Bio101, La Jolla, CA).

Labeling of synthetic oligonucleotides used for markers of nucleotide length. Five micrograms of column-purified synthetic oligonucleotide containing the sequence CGATA(CTG)$_n$. ACGTA was 5′ end-labeled with [γ-^{32}P]ATP by incubation in buffer containing 1 mM DTT, 1 µl 7000 Ci/mmol [γ-^{32}P]ATP (ICN), 50 mM Tris–HCl (pH 7.6), 10 mM $MgCl_2$, 0.1 mM spermidine, 0.1 mM EDTA and 20 U T4 polynucleotide kinase (New England Biolabs). Unincorporated [γ-^{32}P]ATP was removed by size exclusion chromatography (Nuctrap column, Stratagene).

$KMnO_4$ oxidation

Unlabeled synthetic oligonucleotide (1.4 pmol) of the same sequence as the labeled strand was added to 4 × 10^3 d.p.m. (0.7 fmol) of 5′ end-labeled DNA, placed in a boiling water bath for 5 min and then cooled at room temperature for 5 min. $KMnO_4$ oxidation was performed essentially according to the method of McCarthy and Rich (48). Briefly, DNAs were incubated at room temperature in 50 mM sodium cacodylate, pH 7.0, 2 mM EDTA and various concentrations of $KMnO_4$ for 4 min. Final volume of the reaction was 50 µl. Reactions were stopped by addition of 150 µl ice-cold solution containing 98% v/v ethanol, 1% v/v β-mercaptoethanol and 0.02 µg/µl tRNA. DNA was precipitated at –70°C for 1 h after addition of 5 µl 3 M sodium acetate. Pelleted DNA was washed with 750 µl 70% ethanol and dried. DNA was resuspended in 40 µl 1 M piperidine, heated at 92°C for 30 min and dried (×3). Dried samples were dissolved in 10 µl H_2O and 17 µl formamide loading buffer (80% formamide, 10 mM NaOH, 1 mM EDTA, 0.1% xylene cyanol and 0.1% bromophenol blue) was added. Samples were placed in a boiling water bath for 5 min, chilled on ice for 5 min and loaded onto a 12% polyacrylamide gel containing 8 M urea. During electrophoresis, 55°C tap water was circulated through a Hoeffer SE 600 series unit. Electrophoresis was performed at 25 mA/gel in 45 mM Tris–borate, 1 mM EDTA. Gel plates were 14 cm (length) × 16 cm (width) × 1.5 mm (thickness). Electrophoresis was stopped when the bromophenol blue marker migrated 10 cm. Dried gels were placed between two intensifying screens (Dupont) and exposed to Fuji RX film for 3 h–overnight at –80°C.

P1 nuclease digestion

Unlabeled synthetic oligonucleotide (1.4 pmol) of the same sequence as the labeled strand was added to 4 × 10^3 d.p.m. (0.7 fmol) of 5′ end-labeled DNA, placed in a boiling water bath for 5 min and then cooled at room temperature for 5 min. P1 nuclease digestions were performed essentially according to the method of Wohlrab (49). Briefly, DNA was incubated for 3 min at 37°C in 10 µl buffer containing 0.2 M NaCl, 50 mM sodium acetate (pH 7.4), 1 mM $ZnSO_4$, 5% glycerol and various amounts of P1 nuclease (Sigma, St Louis, MO). Reactions were stopped by the addition of 1 µl 0.5 M EDTA. Seventeen microliters of formamide loading buffer were added. DNAs were purified and subjected to polyacrylamide gel electrophoresis as described in the $KMnO_4$ oxidation studies.

Dimethyl sulfate reaction

The CTG-containing strand was liberated from pCTG15 and labeled as described above for $KMnO_4$ oxidation studies. Reactions were performed with 21 mM DMS essentially according to the method of Maxam and Gilbert (50).

RESULTS

A classification of triplet repeats

To aid in correlating potential structures of triplet repeat nucleic acids with their function, we first describe a sequence-based classification system for ds triplet repeats (Table 1). Class V triplet repeats [(AAA/TTT)$_n$, (CCC/GGG)$_n$] contain the homopolymers. Classes I–IV contain non-homopolymeric repeats. Class I triplets [(CTG/CAG)$_n$, (CCG/CGG)$_n$, (GTC/GAC)$_n$] contain a GC or CG palindromic dinucleotide and are G+C-rich. Class II triplets [(CAC/GTG)$_n$, (CTC/GAG)$_n$] do not contain a GC or CG palindromic dinucleotide and are G+C-rich. Class III repeats [(ATC/GAT)$_n$, (TAC/GTA)$_n$, (ATA/TAT)$_n$] contain an AT or TA palindromic dinucleotide and are A+T rich. Class IV triplets [(AGA/TCT)$_n$, (ACA/TGT)$_n$] do not contain a palindromic dinucleotide and are A+T-rich.

Table 1 lists features of ds triplet repeat sequences: base pairing and base stacking energies, slippage synthesis (51) rates (measured by Schlötterer and Tautz; 52), estimated frequency in the Entrez database and the number of associated TREDs. Slippage synthesis of repetitive DNA sequences is a measure of the ability of DNA polymerase to synthesize ds fragments that are longer than template DNA. Class I triplet repeats have the lowest base stacking energies and the lowest slippage synthesis rates, indicating that this class of triplet repeats has unique properties. TAT triplet repeats exhibited the highest rate of slippage synthesis (3.6 bp/min), while CGG triplet repeats exhibited the lowest (no detectable rate).

Table 1. Classification of double-stranded triplet repeat DNA sequences

Class	Sequence	G+C or A+T rich?	Dinucleotide palindrome	kcal/mol double-stranded triplet			Slippage syn. rate bp/min	Database frequency	No. of associated TREDs
				Base-pairing	Base-stacking	Total			
I	$(CCG/CGG)_n$	G+C	GC and CG	-50.37	-32.58	-82.95	<0.5	67	2
	$(GTC/GAC)_n$	G+C	CG only	-40.58	-30.01	-70.59	0.5	5	0
	$(CTG/CAG)_n$	G+C	GC only	-40.58	-28.15	-68.73	0.5	147	5
II	$(CAC/GTG)_n$	G+C	none	-40.58	-25.55	-66.13	0.9	46	0
	$(CTC/GAG)_n$	G+C	none	-40.58	-24.85	-65.43	1.8	75	0
III	$(ATC/GAT)_n$	A+T	AT only	-30.79	-22.95	-53.74	1.0	40	0
	$(ACT/AGT)_n$	A+T	TA only	-30.79	-21.11	-51.90	3.2	14	0
	$(TAT/ATA)_n$	A+T	AT and TA	-21.00	-15.76	-36.76	3.6	213	0
IV	$(TGT/ACA)_n$	A+T	none	-30.79	-22.45	-53.24	1.1	96	0
	$(TCT/AGA)_n$	A+T	none	-30.79	-21.96	-52.75	2.3	75	0
V	$(CCC/GGG)_n$	G+C	none	-50.37	-24.78	-75.15	nd[a]	33	0
	$(TTT/AAA)_n$	A+T	none	-21.00	-16.11	-37.11	nd	nd	0

The 12 double-stranded triplet repeat DNA sequences were classified according to presence of absence of homopolymeric sequences, presence of absence of a palindromic dinucleotide and G+C content. Five classes of triplet repeats are distinguished. Various properties of the double-stranded triplet repeats are listed. Base-stacking energies were obtained by adding the three individual base-stacking components (42) of a triplet repeat sequence. Slippage synthesis rates were determined by Schlotterer and Tautz (52). Database frequency was estimated by searching the Entrez database (National Center for Biotechnology Information, National Library of Medicine, NIH, release 10.0) for the number of perfect matches to eight reiterations of the triplet repeat sequences listed. Searches were performed using the MacVector program on a Macintosh IIsi computer. Excluded from the search were bacterial and synthetic sequences, obvious multiple entries of the same gene, sequences identified by the use of probes containing triplet repeats and matches to $(AAA)_8$ (due to polyadenlylation of mRNA).

[a] nd, not determined.

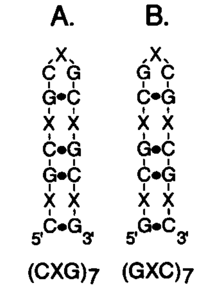

Figure 1. Potential hairpin structures of ss oligonucleotides containing seven Class I triplet repeats.

Five TREDs were associated with expansion of ds CAG repeats, while two TREDs were associated with expansion of ds CGG repeats. However, no TRED was associated with expansion of a ds Class I GAC repeat. The Entrez database search revealed only five perfect matches to eight reiterations of GAC or GTC, none of which were to human DNA. These results suggest that the relative absence of this triplet repeat from DNA may perhaps not only account for its lack of association with a TRED, but also suggest that a selection against a ds GAC triplet repeat exists.

Hairpin formation of Class I triplet repeats predicted by energy minimization

The classification described above also revealed another unique feature of Class I triplet repeats; since the GC or CG dinucleotide is palindromic, ss Class I triplet repeats could form imperfect but stable hairpin structures with 'mismatched' third bases (Fig. 1). Hairpins might also form with Class III repeats, since they contain a palindromic TA or AT dinucleotide. However, T·A base pairs are not as stable as G·C base pairs (-7.00 versus -16.79 kcal) and relatively more AT-containing triplet repeats should be required before a hairpin structure is formed.

To provide some indication of the potential stability of hairpin structures formed from Class I triplet repeats, we performed energy minimizations of hairpin conformations of a series of 5′-CGATA$(XYZ)_n$ACGTA oligonucleotides (where XYZ is a given triplet). These calculations were performed for all six ss Class I repeats. An approximate measure of the stability of the hairpin over the same sequence lacking base pairs was determined from the energies of the minimized structures and is given in Table 2. Further analysis of the energetics of these structures allowed us to further probe the behavior of the different sequences. These data are given in Table 3.

Table 2. Computed stabilization (ΔE) of hairpin over sequence lacking base pairs of 5′-CGATA$(XYZ)_n$ACGTA oligonucleotides and increased stability/trinucleotide repeat (E/rpt) on going from an $n = 5$ to $n = 7$ hairpin and an $n = 7$ to $n = 15$ hairpin

Sequence $(XYZ)_n$	ΔE ($n = 5$)	ΔE ($n = 7$)	ΔE ($n = 15$)	E/rpt (5→7)	E/rpt (7→15)
CTG	-38.3	-64.7	-174.7	-13.2	-13.7
CCG	-35.2	-60.6	-157.1	-12.7	-12.1
GTC	-33.9	-58.9	-150.6	-12.5	-11.5
CGG	-27.3	-50.4	-130.7	-11.6	-10.0
GAC	-29.5	-52.0	-131.2	-11.2	-9.9
CAG	-28.8	-49.8	-127.6	-10.5	-9.7

Table 3. Components of the stabilization energy of hairpin over sequence lacking internal base pairs of 5′-CGATA(XYZ)$_{15}$ACGTA oligonucleotides

Sequence (XYZ)$_{15}$[a]	E_{int}[b] (tri–tri)	ΔE_{int}[c] [tri–tri(+1)]	E_{int}[d] [tri–tri(+1c)]	E_{int}[e] [tri–tri(−1c)]
CTG	−208.4	−8.1	−6.2	12.6
CCG	−197.9	−7.2	−8.1	14.5
CGG	−195.2	9.8	−8.1	4.0
CAG	−184.5	3.6	−7.8	11.6
GAC	−165.3	10.1	−3.4	−23.6
GTC	−157.7	11.5	−1.5	−25.6

[a] A trinucleotide repeat is defined as a 5′-pXpYpZ.

[b] Sum of the interactions between trinucleotides 1 and 15, 2 and 14, 3 and 13, 4 and 12, 5 and 11, 6 and 10, 7 and 9.

[c] Difference in energy of the interactions between trinucleotide repeats 1 and 2, 2 and 3, 3 and 4,...,14 and 15 in the hairpin and in the sequence lacking base pairs.

[d] Sum of the interactions between trinucleotides 1 and 14, 2 and 13, 3 and 12, 4 and 11, 5 and 10, 6 and 9.

[e] Sum of the interactions between trinucleotides 2 and 15, 3 and 14, 4 and 13, 5 and 12, 6 and 11, 7 and 10.

Of the pyrimidine-rich Class I sequences, hairpin structures formed from CTG repeats were the most stable, whilst hairpins formed from GTC repeats were the least stable. Of the purine-rich Class I sequences, hairpin structures formed from CGG repeats were the most stable, while hairpin structures formed from CAG repeats were predicted to be the least stable. As few as five repeats of each Class I triplet were predicted to be capable of forming a hairpin structure.

Examination of the structural and energetic basis for the above provided a number of interesting insights. We were particularly concerned with understanding the difference between the CTG and GTC sequences which have identical base content. In Table 3 several components of the hairpin energies are given for each of the six Class I repeats 5′-CGATA(XYZ)$_{15}$ACGTA. This analysis suggests the main differences between different sequences lie in the respective stacking energies. The E_{int}(tri–tri) term gives the total interaction between ‘directly base paired’ trinucleotides (for example, triplet 1 with triplet 15, 2 with 14, etc.) in the hairpin structure. These are direct stabilization energies, since they obviously do not occur in the sequences lacking base pairs. This term strongly favors the (CTG)$_{15}$ hairpin over the (GTC)$_{15}$ hairpin. Upon breaking this down further (data not shown), we found that the total energies of the Watson–Crick base pairing and the T–T mismatches was similar in the two sequences. The large difference between (CTG)$_{15}$ and (GTC)$_{15}$ arises from stacking, particularly between bases on ‘opposite’ strands of the hairpin.

Further differences in the energetics of the different sequences which arise from stacking effects are shown by computing the interactions between ‘non-directly base paired trinucleotides’ (Table 3). ΔE_{int}[tri–tri(+1)] (an energy difference, since this term is not zero in the sequence lacking base pairs) describes the total stabilization energy upon hairpin formation due to interactions between successive trinucleotide repeats (i.e. between 1 and 2, 2 and 3, etc., excluding the loop). This term favors hairpin formation in (CTG)$_{15}$ and in (CCG)$_{15}$, but opposes it in the other sequences. The E_{int}[tri–tri(+1c)] term gives the interaction between a given trinucleotide and the complementary trinucleotide of the +1 repeat (for example, 1 with 14, 2 with 13, etc., and favors hairpin formation in all six sequences. The E_{int}[tri–tri(−1c)] term similarly gives the interaction between a given trinucleotide and the complementary trinucleotide of the −1 repeat (for example, 2 with 15, 3 with 14, etc.). This term is dramatically different in the GXC repeat sequences compared with the CXG sequences, strongly favoring hairpin formation in the former, but opposing hairpin formation in the latter.

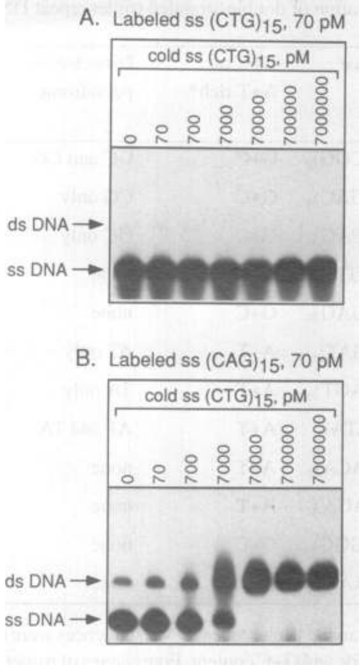

Figure 2. Structure of ss (CTG)$_{15}$ is concentration-independent. A ds oligonucleotide containing (CTG)$_{15}$ was excised from plasmid pCTG15 as described in Materials and Methods. Strands end-labeled with polynucleotide kinase were (**A**) ss (CTG)$_{15}$ and (**B**) ss (CAG)$_{15}$. Prior to gel electrophoresis, unlabeled synthetic oligonucleotide containing (CTG)$_{15}$ at the indicated concentration was added to the indicated labeled strand, placed in a boiling water bath for 10 min and then cooled at room temperature for 15 min. DNA samples were applied to a native 8% polyacrylamide gel.

The electrophoretic mobility of an oligonucleotide containing ss (CTG)$_{15}$ is concentration-independent

To analyze DNA containing a Class I triplet repeat, a ds oligonucleotide containing 15 prototypic CTG triplet repeats was cloned into a plasmid as described in Materials and Methods. Oligonucleotides liberated from the plasmid were utilized for studies, since they are unequivocally full-length. Non-repetitive sequences were also included in the termini of the oligonucleotides to help prevent ‘slippage’ of the hairpin structure and to provide restriction sites for release of the oligonucleotide from the plasmids.

The sequence of the labeled ss oligonucleotide containing (CTG)$_{15}$ [ss (CTG)$_{15}$] was G1-A2-T3-C4-C5-(CTG)$_{15}$-G51-G52-T53-A54-C55-C56-A57-A58-G59-C60-T61. The nucleotides within the presumed loop region were C27-T28-G29. The predicted hairpin structure of ss (CTG)$_{15}$ contained 47 hydrogen bonds. To determine whether ss (CTG)$_{15}$ exhibited properties of an intramolecular hairpin or some type of intermolecular structure, various studies were performed. First, the molecular

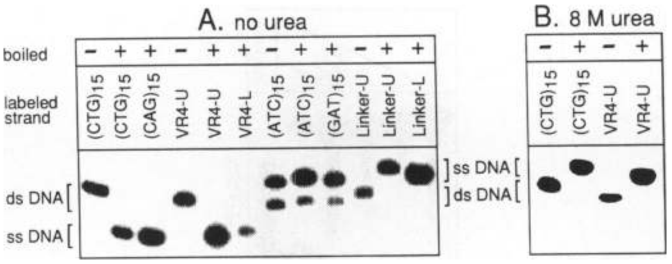

Figure 3. Electrophoretic analyses of various DNAs. Oligonucleotides containing $(CTG)_{15}$, VR4, $(ATC)_{15}$ and linker were prepared from pCTG15, pVR4, pATC15 and plinker respectively, as described in Materials and Methods. Double-stranded oligonucleotides were labeled on one strand only. The sequence of the labeled strand is indicated in the figure. Where indicated, the DNA samples were boiled for 10 min and cooled at room temperature for 10 min prior to gel loading. Labeled vector DNA, which remained at or near the well origin, is not shown. Electrophoresis was performed at 15°C. (**A**) Eight per cent polyacrylamide gel, pH 8.5. (**B**) Eight per cent polyacrylamide gel containing 8 M urea, pH 8.5. Synthetic unlabeled ss DNAs containing ss $(CTG)_{15}$ or VR4-U were added (2×10^3-fold molar excess, final DNA concentration = 140 nM) to reactions containing ss $(CTG)_{15}$ or VR4-U respectively, to prevent re-annealing of complementary strands.

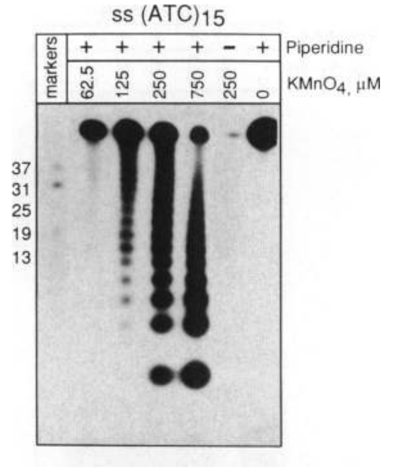

Figure 4. $KMnO_4$ uniformly oxidizes all thymines of ss $(ATC)_{15}$. A ss oligonucleotide containing $(ATC)_{15}$ was prepared as described in Materials and Methods and labeled at the 5′ end with polynucleotide kinase. Oxidation reactions were performed at room temperature as described in Materials and Methods at the above indicated concentration of $KMnO_4$. Reactions were performed with and without piperidine as indicated. The marker lane contained synthetic 5′ end-labeled ss oligonucleotides with the sequence 5′-CGATA$(CTG)_n$ACGTA-3′, where n = 1, 3, 5 or 7, corresponding to lengths of 13, 19, 25, 31 and 37 nt respectively. The gel contained 12% polyacrylamide and 8 M urea. In the lane containing 250 µM $KMnO_4$ the second signal from the bottom corresponds to the thymine in the first ATC repeat.

composition of the structure(s) formed with ss $(CTG)_{15}$ was investigated by performing electrophoretic studies with labeled ss $(CTG)_{15}$ mixed with various amounts of unlabeled ss synthetic oligonucleotide containing ss $(CTG)_{15}$ (Fig. 2A). If ss $(CTG)_{15}$ formed a stable intramolecular hairpin structure, increasing the concentration of unlabeled synthetic ss $(CTG)_{15}$ should have no effect on the amount of hairpin formed.

In the absence of added unlabeled ss synthetic oligonucleotide, a predominant species of DNA with a relatively fast electrophoretic mobility was observed, corresponding to ss DNA (Fig. 2A). Addition of a 10^5-fold molar excess (final DNA concentration 7 µM) of unlabeled ss synthetic oligonucleotide of the same sequence as ss $(CTG)_{15}$ did not result in formation of slow migrating complexes (as anticipated), indicating that ss $(CTG)_{15}$ formed a stable unimolecular structure.

A minor, slower migrating species of DNA (barely detectable in the autoradiograph) was also detected (Fig. 2A). Addition of a 10^3-fold molar excess of complementary unlabeled ss synthetic oligonucleotide resulted in complete conversion of the fast migrating species to the slow migrating species (data not shown), indicating that the slow migrating species was the ds form of $(CTG)_{15}$.

To demonstrate that the unlabeled ss synthetic oligonucleotide containing $(CTG)_{15}$ was not degraded and contained CTG repetitive sequences, a control experiment was performed with labeled ss $(CAG)_{15}$ (Fig. 2B). Addition of increasing amounts of unlabeled ss synthetic oligonucleotide containing $(CTG)_{15}$ to the labeled ss $(CAG)_{15}$ probe resulted in complete conversion of the fast migrating ss form to the slow migrating ds form.

Single-stranded oligonucleotides that lack predicted hairpins migrate slower than their ds forms

The results described above suggested that the ss form of $(CTG)_{15}$ migrated faster than its ds form. With respect to ds $(CTG)_{15}$, the relatively rapid mobility of ss $(CTG)_{15}$ may have been due to reduced mass or increased charge density. Alternatively, the relatively rapid mobility could have been due to secondary structure. To provide electrophoretic evidence for the latter possibility, it was necessary to demonstrate that: (i) under native conditions, a ss DNA containing a hairpin also migrated faster than its ds form; (ii) under native conditions, a ss DNA containing no predicted hairpin did not migrate faster than its ds form; (iii) under partial denaturing conditions, the electrophoretic mobility of ss $(CTG)_{15}$ was reduced relative to ds $(CTG)_{15}$.

To test these hypotheses, the electrophoretic mobilities of various DNA sequences were analyzed and compared to $(CTG)_{15}$ and $(CAG)_{15}$. The test DNAs included two complementary sequences (referred to as VR4-U and VR4-L) that formed a hairpin with equivalent numbers of nucleotides (n = 66) and equivalent numbers of hydrogen bonds (n = 44), two complementary sequences with equivalent numbers of nucleotides (n = 64) that contained no predicted hairpin (referred to as linker-U and linker-L) and two oligonucleotides that contained 15 Class III triplet repeats [ss $(ATC)_{15}$ or ss $(GAT)_{15}$]. Single-stranded $(ATC)_{15}$ and ss $(GAT)_{15}$ contain AT dinucleotides that might form base pairs in a hairpin. Since the stacking energy in the d(AT/AT) dinucleotide duplex is 1.7 times greater than the stacking energy in the d(TA/TA) dinucleotide duplex (42), oligonucleotides containing ATC or GAT repeats are probably more likely to form hairpin structures compared with oligonucleotides containing TAC or GTA repeats.

The relative electrophoretic mobilities of the DNAs are shown in Figure 3. Under native conditions (Fig. 3A), the electrophoretic mobilities of ss $(CTG)_{15}$, ss $(CAG)_{15}$, ss VR4-U and VR4-L were greater than their ds forms. In contrast, the electrophoretic mobilities of ss linker-U, ss linker-L, ss $(ATC)_{15}$ and ss $(GAT)_{15}$ were less than their ds forms. These results indicate that ss $(CTG)_{15}$, ss $(CAG)_{15}$, ss VR4-U and VR4-L form secondary structures that are more compact compared with those of ss linker-U, ss linker-L, ss $(ATC)_{15}$ and ss $(GAT)_{15}$.

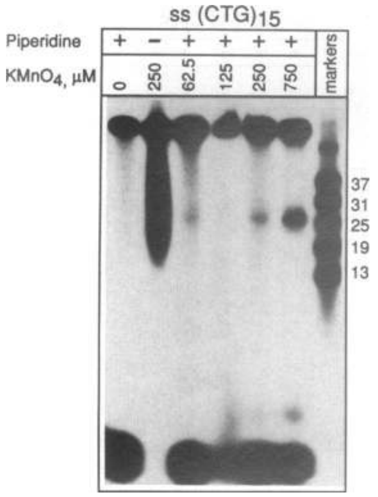

Figure 5. $KMnO_4$ oxidizes a single thymine in the middle of the triplet repeat region of ss $(CTG)_{15}$. A ss oligonucleotides containing $(CTG)_{15}$ was prepared as described in Materials and Methods and labeled at the 5′ end with polynucleotide kinase. Oxidation reactions were performed at room temperature as described in Materials and Methods at the indicated concentration of $KMnO_4$. Reactions were performed with and without piperidine as indicated. The gel contained 12% polyacrylamide and 8 M urea. The marker lane contained synthetic 5′ end-labeled ss oligonucleotides with the sequence 5′-CGATA$(CTG)_n$ACGTA-3′, where n = 1, 3, 5 or 7, corresponding to lengths of 13, 19, 25, 31 and 37 nt respectively.

To demonstrate that the electrophoretic mobility of ss $(CTG)_{15}$ was reduced relative to ds $(CTG)_{15}$ under partial denaturing conditions, electrophoresis was performed at 15°C in a polyacrylamide gel containing 8 M urea, conditions which do not denature ds $(ATC)_{15}$ or ds $(CTG)_{15}$ (M. Mitas, unpublished results). At 15°C in 8 M urea, the electrophoretic mobilities of ss $(CTG)_{15}$ and ss VR4-U were less than their respective ds forms (Fig. 3B). These results indicate that 8 M urea partially or completely denatured the secondary structures of ss $(CTG)_{15}$ and ss VR4-U, a result consistent with the hypothesis that each contained a hairpin. The results of the electrophoretic studies suggest that ss $(CTG)_{15}$, ss $(CAG)_{15}$, ss VR4-U and VR4-L contain secondary structure, while ss linker-U, ss linker-L, ss $(ATC)_{15}$ and ss $(GAT)_{15}$ do not contain secondary structure.

$KMnO_4$ oxidizes a single thymine in the triplet repeat region of ss $(CTG)_{15}$

$KMnO_4$ preferentially oxidizes unpaired or unstacked thymines, resulting in strand cleavage upon subsequent treatment with piperidine (53,54). $KMnO_4$ oxidation experiments were performed first with ss $(ATC)_{15}$. Treatment of ss $(ATC)_{15}$ with $KMnO_4$ at 23°C resulted in uniform oxidation of all thymines (Fig. 4), indicating that these residues do not participate in base pairing or base stacking interactions. All thymines of ss $(GAT)_{15}$ were oxidized in a similar manner (data not shown). These results indicate that ss $(ATC)_{15}$ and ss $(GAT)_{15}$ do not form hairpin structures.

Treatment of ss $(CTG)_{15}$ with $KMnO_4$/piperidine (Fig. 5) resulted in a single cleavage product within the triplet repeat region at T28 (predicted loop region C27-T28-G29). A second product of short nucleotide length was also observed near the bottom of the gel, corresponding to T3, a thymine that formed part of the *Bam*HI restriction recognition site. These results indicate that ss $(CTG)_{15}$ formed a structure that contained a single loop located in the middle of the triplet repeat region. The results further indicated that the thymines within the presumed stem region were involved in base pairing and/or base stacking interactions.

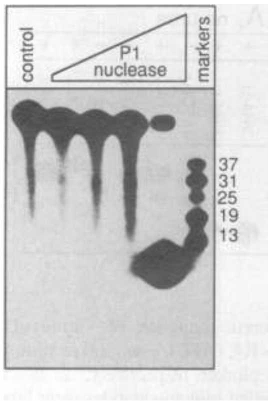

Figure 6. P1 nuclease digestion of ss $(ATC)_{15}$. P1 nuclease digestions were performed at 37°C with ss $(ATC)_{15}$ as described in Materials and Methods. The amounts of P1 nuclease used to digest ss $(ATC)_{15}$ were (from left to right) 0, 2.6, 7.5, 25 and 75 × 10^{-4} U. Marker lane contained synthetic 5′ end-labeled ss oligonucleotides with the sequence 5′-CGATA$(CTG)_n$ACGTA-3′, where n = 1, 3, 5 or 7, corresponding to lengths of 13, 19, 25, 31 and 37 nt respectively. The gel contained 12% polyacrylamide and 8 M urea.

P1 nuclease cleaves at the predicted loop region of single-stranded $(CTG)_{15}$

Nuclease sensitivity studies were performed with P1, a single-strand-specific endonuclease that exhibits no apparent sequence specificity (49). Experiments were first performed at 37°C with ss $(ATC)_{15}$ (Fig. 6). The results with ss $(ATC)_{15}$ show that P1 nuclease did not significantly cleave the triplet repeat region (nt 6–51).

Treatment of ss $(CTG)_{15}$ with increasing amounts of P1 nuclease (Fig. 7) produced a fragment of size similar to that generated by treatment of ss $(CTG)_{15}$ with $KMnO_4$/piperidine. Triplets 1–7 and 9–15 were not digested, indicating that the nucleotides in the presumed stem region participated in base pairing interactions. The results of P1 digestion are in agreement with $KMnO_4$ studies and are consistent with a hairpin structure of ss $(CTG)_{15}$ in which the thymines in the stem are base paired and/or stacked.

Energy minimization of ss $(CTG)_{15}$ containing hydrogen bonded thymines

To evaluate the energetics of potential T·T base pairs within ss $(CTG)_{15}$, energy minimization was performed. In the first step, the distances of the hydrogen bonds between thymines in ss $(CTG)_{15}$ were fixed at 2.0 Å. The hydrogen bonds were formed between H3 and O4 of the thymines on the first strand (i.e. trinucleotide repeats 1–7) and with O2 and H3 of the thymines on the second strand (i.e. trinucleotide repeats 9–15) respectively. In the second step, the entire structure was relaxed with no constraints and subjected to 2000 cycles of minimization. The final result was a structure in which the hydrogen bonds were

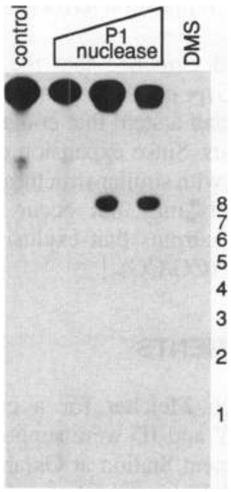

Figure 7. P1 nuclease digestion of ss $(CTG)_{15}$. P1 nuclease digestions and DMS reactions were performed at 37°C with ss $(CTG)_{15}$ as described in Materials and Methods. Numbers indicate approximate position of the respective CTG triplet repeats. The amounts of P1 nuclease used to digest ss $(CTG)_{15}$ were (from left to right) 3.8, 12 and 34 × 10^{-3} U respectively. Electrophoretic conditions were as described in Materials and Methods for $KMnO_4$ oxidation studies except that the gel contained 20% polyacrylamide. The signal in the control lane that migrates between full-length ss DNA and triplet eight corresponds to a minor contaminant of ds DNA. The ^{32}P label in the ds DNA is at the 5′-terminus of a 4 nt overhang and susceptible to P1 digestion.

retained. Surprisingly, the structure was almost identical in energy to the previous ss $(CTG)_{15}$ conformation; the difference in energies between the 'hydrogen bonded' conformation and the one described earlier without T·T hydrogen bonds was only 3 kcal/mol (in favor of the 'hydrogen bonded' conformation). Therefore, the formation of the hydrogen bonds appears to compromise other parts of the duplex.

DISCUSSION

To aid in correlating potential structures of triplet repeat nucleic acids with their function, we describe a sequence-based classification system for ds triplet repeats (Table 1). Class I repeats, which are defined by the presence of a GC or CG palindrome, have the lowest base stacking energies, exhibit the lowest rates of slippage synthesis (52) and are uniquely associated with TREDs. The six single-stranded (ss) triplet repeats within Class I also have the potential to form hairpin structures, as determined by energy minimization (Table 2 and 3). These results provide evidence for the validity of the classification system and suggest that Class I triplet repeats may perform unique biological functions.

Class I triplet repeats may form stable hairpin structures

Due to the unique structural features of Class I triplet repeats and due to their unique association with TREDs, various experiments were performed with a ss oligonucleotide containing 15 prototypic CTG repeats. Electrophoretic analysis revealed that ss $(CTG)_{15}$ migrated rapidly in native polyacrylamide gels (Figs 2 and 3), a result consistent with a hairpin structure and inconsistent with a random coil structure (55,56). Similar results were obtained with ss oligonucleotides containing predicted hairpin structures: ss $(CAG)_{15}$, ss VR4-U and ss VR4-L (Fig. 3). In contrast to the results obtained with ss $(CTG)_{15}$, two oligonucleotides of equivalent length lacking predicted hairpins (linker-U and linker-L), as well as ss $(ATC)_{15}$ and ss $(GAT)_{15}$, migrated slower than their ds forms in native polyacrylamide gels (Fig. 3). These results suggest that the oligonucleotides containing ss linker-U, ss linker-L, ss $(ATC)_{15}$ and ss $(GAT)_{15}$ do not form hairpins.

$KMnO_4$ oxidation and P1 nuclease digestion studies of ss $(CTG)_{15}$ revealed cleavage at triplet number eight, suggesting that the structure causing its rapid electrophoretic mobility was a hairpin. Similar $KMnO_4$ oxidation results to those described above have also been observed at 50°C, indicating that the hairpin of ss $(CTG)_{15}$ is thermally stable (A. Yu, J. Dill and M. Mitas, unpublished results). Energy minimization data further suggested that the hairpin structure may be stabilized by hydrogen bonds formed between mismatched thymines. In support of this conclusion, hydrogen bonds formed between the T–T mismatches of the homoduplex d(CTG/CTG)$_3$ have been observed in ^1H NMR studies (X. Gao, University of Houston, personal communication).

We suspect that since ss $(CAG)_{15}$ also migrated with a relatively rapid electrophoretic mobility, ss oligonucleotides containing other Class I triplet repeats will also form hairpin structures, although perhaps not as stable as those formed from CTG triplet repeats.

Single-stranded d$(ATC)_{15}$: a sequence that forms a structure lacking stacked bases?

In contrast to the results obtained with ss $(CTG)_{15}$, all thymines of ss $(ATC)_{15}$ and ss $(GAT)_{15}$ were uniformly oxidized by $KMnO_4$ (Fig. 4 and data not shown), indicating that these sequences did not form hairpins. These results suggest that structures of ss triplet repeat DNAs from Class III are different from those containing GC or CG palindromic dinucleotides. P1 nuclease degraded the ATC triplet region in ss $(ATC)_{15}$ very poorly, perhaps suggesting that ss $(ATC)_{15}$ contained considerable secondary structure. This conclusion apparently contradicts the results of electrophoretic mobility and $KMnO_4$ oxidation studies, which suggested ss $(ATC)_{15}$ contained little or no secondary structure. A possible explanation of these apparently conflicting data is that ss $(ATC)_{15}$ formed a nuclease-resistant structure lacking stacked bases. The properties of the ss $(ATC)_{15}$ structure may parallel those of d$(A^+G)_{10}$, which forms an acid-induced intramolecular α-helical-like structure stabilized not by stacked bases or hydrogen bonded base pairs, but instead by ionic bonds between positively charged adenine residues and distal negatively charged phosphates (57,58). If ss $(ATC)_{15}$ and ss $(GAT)_{15}$ contain similar α-helical-like properties, their structures should be characterized by marked circular dichroism and little or no UV hypochromicity.

Alternatively, it is possible that at higher salt concentrations (e.g. ≥200 mM Na^+), ss $(ATC)_{15}$ forms a hairpin containing a nuclease-resistant loop. In the studies described above, P1 nuclease digestions were performed at 200 mM Na^+, whereas $KMnO_4$ studies were performed in 50 mM Na^+. If $KMnO_4$

reactions were performed at higher Na^+ concentrations, the thymines may not have been oxidized.

Biological implications of hairpin structures formed from Class I triplet repeats

We suspect that hairpin structures will also form with ss $(CUG)_{15}$ and other Class I RNA triplet repeats. These putative RNA hairpins can potentially regulate protein translation efficiency and mRNA splicing. Regulation of protein translation by hairpin structures has been documented in procaryotic and eucaryotic systems (59–64). In addition, a number of RNA binding proteins have been identified in eucaryotic systems, some of which recognize hairpin structures (65–67). Therefore, hairpins formed from Class I triplet repeats have the potential to regulate protein translation. Further, since the lengths of triplet repeats often vary between individuals, hairpin structures formed from them may differ in size and thermal stability. The stability of these hairpins may lead to subtle variations in levels of translated protein. Also, mRNA hairpins can affect splicing; an mRNA hairpin containing as few as 6 bp in the stem can sequester the 5′ splice site and inhibit the early steps of spliceosome assembly (68). mRNA hairpins can also increase splicing efficiency, as well as affect splice site selection (69).

Hairpin structures formed from Class I triplet repeats may play an important role in triplet repeat expansion events

Seven TREDs have been described, all of which are due to expansions of Class I triplet repeats. If TREDs were evenly distributed among the triplet repeats and the six ss Class I triplet repeats comprised 6/24 of triplet repeats present in the human genome, the probability of random association of only Class I repeats with TREDs is $(1/4)^7 = 0.00006$. Thus it seems likely that TREDs do not involve all classes of repeats. However, this analysis is oversimplified, in that three of the seven genes (HD, SCA-1 and DPA) associated with TREDs were isolated exclusively by their ability to hybridize to DNA probes containing repeats of CAG or CTG. Therefore, the estimate of P is between 0.004 and 0.00006. This analysis, and the results presented in this paper, perhaps suggest that hairpin structures formed from Class I triplet repeats play an important role in triplet repeat expansion events.

A mechanism of triplet repeat expansion was recently proposed by Sinden and Wells that involved blockage of DNA replication by an unknown structure, slippage of the replicated leading strand and stabilization of the slipped strand through formation of hairpin structures (70). Presumably, hairpin formation of the slipped strand is necessary to prevent nucleolytic degradation of the ss structure. In support of the later aspect of this mechanism, we have presented evidence that at least one of the triplet repeat sequences associated with TREDs forms a hairpin. However, concerning other aspects of the mechanism proposed by Sinden and Wells, we suspect that 'blockage' to DNA replication does not occur during leading strand synthesis, but instead occurs during lagging strand synthesis. Unlike leading strand synthesis, lagging strand synthesis requires exposure of large regions of ss DNA, which, in the case of Class I triplet repeats, can form hairpin structures that may slow or block DNA replication. It is known that hairpin structures at the lagging strand frequently contribute to several types of replication errors, such as deletions and mutations (71–74).

We have provided evidence that the unimolecular hairpin structure of ss $(CTG)_{15}$ contains a loop that is accessible to a nucleolytic enzyme and a stem that contains base paired and/or base stacked thymines. Since expansion of nucleotide elements may require hairpins with similar structural features, it is possible that expansion events may not occur with repetitive DNA elements that form hairpins that exclusively contain Watson–Crick base pairs [i.e. $(CG/CG)_n$].

ACKNOWLEDGEMENTS

We thank Dr Ulrich Melcher for a critical review of this manuscript. MM, AY and JD were supported by the Oklahoma Agricultural Experiment Station at Oklahoma State University. TJK was supported by a post-doctoral fellowship from the American Heart Association, Maryland Affiliate.

REFERENCES

1 Jeffreys,A.J., Wilson,V. and Thein,S.L. (1985) *Science*, **314**, 67–73.

2 Streisinger,G., Okada,Y., Emrich,J., Newton,J., Tsugita,A., Terzachi,E. and Inouye,M. (1966) *Cold Spring Harbor Symp. Quant. Biol.*, **31**, 77–84.

3 Drake,J.W., Glickman,B.W. and Ripley,L.S. (1983) *Am. Scient.*, **71**, 621–630.

4 Wells,R.D., Ohtsuka,E. and Khorana,H.G. (1965) *J. Mol. Biol.* **14**, 221–240.

5 Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritx,R.A., DeRiel,J.K., Forget,B.G., Weissman,S.M., Slightom,J.L., Blechi,S.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot,N.J. (1980) *Cell*, **21**, 653–668.

6 Tautz,D and Renz,M. (1984) *Nucleic Acids Res.*, **12**, 4127–4138.

7 Dallas,J.F. (1992) *Mammalian Genet.*, **3**, 452–456.

8 Wang,H.H.J., Quigley,G.J., Kolpak,F.J., Van Boom,J.L., Van der Marel,G. and Rich,A. (1979) *Nature (London)*, **282**, 680–686.

9 Weintraub,H. and Groudine,M. (1976) *Science*, **193**, 348–356.

10 Hentschel,C.C. (1982) *Nature (London)*, **295**, 714–716.

11 Yamazaki,H., Nomoto,S., Mishima,Y. and Kominami,R. (1992) *J. Biol. Chem.*, **267**, 12311–12316.

12 Collick,A. and Jeffreys,A. (1990) *Nucleic Acids Res.*, **18**, 625–629.

13 Richards,R.I., Holman,M.G., Yu,S. and Sutherland,G.R. (1993) *Hum. Mol. Genet.*, **2**, 1429–1435.

14 Shen,S.H., Slighton,J.L. and Smithies,O. (1981) *Cell*, **26**, 191–203.

15 Tautz,D., Trick,M. and Dover,G.A. (1986) *Nature*, **322**, 652–656.

16 Tautz,D. (1989) *Nucleic Acids Res.*, **17**, 6463–6471.

17 Turner,B.J., Elder,J.F.,Jr, Laughlin,T.F. and Davis,W.P. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 5653–5657.

18 Schafer,R., Zischler,J. and Epplen,J.T. (1988) *Nucleic Acids Res.*, **16**, 5196.

19 Fu,Y.-H., Kuhl,D.P.A., Pizzuti,A., Pierreti,M., Sutcliffe,S.S., Richards,S., Verkerk,A.J.M.H., Holden,J.J.A., Fenwick,R.G.,Jr, Warren,S.T., Oostra,B.A., Nelson,D.L. and Caskey,C.T. (1991) *Cell*, **67**, 1047–1068.

20 Ververk,A.J.J.H., Pieretti,M., Sutcliffe,J.S., Fu,Y.-H, Kuhl,D.P.A., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F., Eussen,B.E., van Ommen,G.-J.B., Blonden,L.A.J., Riggins,G.J., Chastain,J.L., Kunst,C.B., Galjaard,H., Caskey,C.T., Nelson,D.L., Oostra,B.A. and Warren,S.T. (1991) *Cell*, **65**, 905–914.

21 La Spada,A.R., Wilson,E.M., Lubahn,D.B., Harding,A.E. and Fischbeck,K.H. (1991) *Nature (London)*, **352**, 77–79.

22 Edwards,A., Hammond,H.A., Jin,L., Caskey,C.T. and Chakraborty,R. (1992) *Genomics*, **12**, 241–253.

23 The Huntingon's Disease Collaborative Research Group (1993) *Cell*, **72**, 971–983.

24 Brook,J.D., McCurrash,A.E., Harley,H.G., Buckler,A.J., Church,D., Aburatani,H, Hunter,K., Stanton,V.P., Thirion,J.-P., Hudson,T., Sohn,R., Zemelman,B., Snell,R.G., Rundle,S.A., Crow,S., Davies,J., Shelbourne,P., Buxton,J., Jones,C., Juvonen,V., Johnson,K., Harper,P.S., Shaw,D.J. and Housman,D.E. (1992) *Cell*, **68**, 799–808.

25 Mahadevan,M., Tsilfidis,C., Sabourin,L., Shutler,G., Amemiya,C., Jansen,G., Neville,C., Narang,M., Barcelo,J., O'Hoy,K., Leblond,S., Earle-Macdonald,J., de Jong,P.J., Wieringa,B. and Korneluk,R.G. (1992) *Science*, **255**, 1253–1255.

26 Orr,H.T. Chung,M.Y., Banfi,S., Kwiatkowski,T.J., Servakio,A., Beaudet,A.L., McCall,A.E., Duvick,L.A., Ranum,L.P. and Zoghbi,H.Y. (1993) *Nature Genet.*, **4**, 221–226.

27 Koide.R., Ikeuchi,T., Onodera,O., Tanaka,H., Igarashi,S., Endo,K., Takahashi,H., Kondo,K., Ishikawa,A., Hatashi,T., Saito,S., Tomoda,A., Miike,T., Naito,H. and Ikuta,F. (1994) *Nature Genet.*, **6**, 9–13.

28 Bell,M.V., Hirst,M.C., Nakahori,Y., Mackinnon,R.N., Roche,A., Flint,T.J., Jacobs,P.A., Tommerup,N., Tranebjaerg,L., Froster-Iskenius,U., Kerr,B., Turner,G., Lindenbaum,R.H., Winter,R., Pembrey,M., Thideau,S. and Davies,K.E. (1991) *Cell*, **64**, 861–866.

29 Sutclffe,J.S., Nelson,D.L., Zhang,F., Pieretti,M., Caskey,C.T., Saxe,D. and Warren,S.T. (1992) *Hum. Mol. Genet.*, **1**, 397–400.

30 Pieretti,M., Zhang,F., Fu,Y.-H., Warren,S.T., Oostra,B.A., Caskey,C.T. and Nelson,D.L. (1991) *Cell*, **66**, 817–822.

31 Caskey,C.T., Pizzuti,A., Fu,Y.-H., Fenwick,R.G.,Jr and Nelson,D.L. (1992) *Science*, **256**, 784–789.

32 Tarleton,J.C. and Saul, R.A. (1993) *J. Pediat.*, **122**, 169–185.

33 Ezzell,C. (1993) *J. Natl. Inst. Hlth Res.*, **5**, 54–58.

34 Nelson,D.L and Warren,S.T. (1993) *Nature Genet.*, **4**, 107–108.

35 Morrison,P.J. (1993) *Lancet*, **342**, 385–386.

36 Martin,J.B. (1993) *Science*, **262**, 674–676.

37 Miwa,S. (1994) *Nature Genet.*, **6**, 3–4.

38 Richards,R.I. and Sutherland,G.R. (1992) *Cell*, **70**, 709–712.

39 Wrogemann,K., Biancalana,V., Devys,D., Imbert,G., Trottier,Y. and Mandel,J.L. (1993) *Exs.*, **67**, 141–152.

40 Sutherland,G.R. and Richards,R.I. (1994) *Am. Scient.*, **82**, 157–163.

41 Kamp,T.J., Mitas,M., Fields,K.L., Asoh,S., Chin,H., Marban,E. and Nirenberg,M. (1995) *Cell. Mol. Neurobiol.* in press.

42 Saenger,W. (ed.) (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, NY, p. 139.

43 Pearlman,D.A., Case,D.A., Caldwell,J.C., Seibel,G.L., Singh,U.C., Weiner,P. and Kollman,P.A. (1991) University of California, San Francisco.

44 Weiner,P. and Kollman,P.A. (1981) *J. Comput. Chem.*, **2**, 287–303.

45 Weiner,P., Kollman,P.A., Nguyen,D.T. and Case,D.A. (1986) *J. Comput. Chem.*, **7**, 230–252.

46 Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,G., Alagona,G., Profeta,S. and Weiner,P. (1984) *J. Am. Chem. Soc.*, **106**, 765–784.

47 Luckow,B. and Schutz,G. (1987) *Nucleic Acids Res.*, **15**, 5490.

48 McCarthy,J.G. and Rich,A. (1991) *Nucleic Acids Res.*, **19**, 3421–3429.

49 Wohlrab,F. (1992) *Methods Enzymol.*, **212B**, 294–301.

50 Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.

51 Korberg,A., Bertsch,L.L., Jackson,J.F. and Khorana,H.G. (1964) *Proc. Natl. Acad. Sci. USA*, **51**, 315–323.

52 Schlötterer,C and Tautz,D. (1992) *Nucleic Acids Res.*, **20**, 211–215.

53 Hayatsu,H. and Ukita,T. (1967) *Biochem. Biophys. Res. Commun.*, **29**, 556–561.

54 Rubin,C.M. and Schmid,C.W. (1980) *Nucleic Acids Res.*, **8**, 4613–4619.

55 Henderson,E., Harkin,C.C., Walk,S.K., Tinoco,I.,Jr and Blackburn,E.H. (1987) *Cell*, **51**, 899–908.

56 Balagurumoorthy,P., Brahmachari,S.K., Mohanty,D.Bansal,M. and Sasisekharan,V. (1992) *Nucleic Acids Res.*, **20**, 4061–4067.

57 Dolinnaya,N.G. and Fresco,J.R. (1992) *Proc. Natl. Acad. Sci., USA*, **89**, 9242–9246.

58 Dolinnaya,N.G., Braswell,E.H., Fossella,J.A., Klump,H. and Fresco,J.R. (1993) *Biochemistry*, **32**, 10263–10270.

59 Peterson,C. (1992) *Mol. Microbiol.* **6**, 277–282.

60 Dix,D.J., Lin,P.N., McKenzie,A.R., Walden,W.E. and Theil,E.C. (1993) *J. Mol. Biol.*, **231**, 230–240.

61 Theil,E.C. (1993) *Biofactors*, **4**, 87–93.

62 Yoon,H., Miller,S.P., Pabich,E.K. and Donahue,T.F. (1992) *Genes Dev.*, **5**, 2463–2477.

63 Liebhaber,S.A., Cash,F. and Eshleman,S.S. (1992) *J. Mol. Biol.*, **226**, 609–621.

64 Vega Laso,M.R., Zhu,D., Sagliocco,F., Brown,A.J., Tuite,M.F. and McCarthy,J.E. (1993) *J. Biol. Chem.*, **268**, 6453–6462.

65 van Gelder,C.W., Gunderson,S.I., Jansen,E.J., Boelens,W.C., Polycarpou-Schwarz,M., Mattaj,I.W. and van Venrooi,J.W.J. (1993) *EMBO J.*, **12**, 5191–5200.

66 Chen,L. and Frankel,A.D. (1994) *Biochemistry*, **33**, 2708–2715.

67 Pause,A., Methot,N., Svitkin,Y., Merrick,W.C. and Sonenberg,N. (1994) *EMBO J.*, **13**, 1205–1215.

68 Goguel,V., Wang,Y. and Rosbash,M. (1993) *Mol. Cell. Biol.*, **13**, 6841–6848.

69 Goguel,V. and Rosbash,M. (1993) *Cell*, **72**, 893–901.

70 Sinden,R.R. and Wells,R.D. (1992) *Curr. Opin. Biotechnol.*, **3**, 602–622.

71 Sinden,R.R., Zheng,G., Brankamp,R.G. and Allen,K.N. (1991) *Genetics*, **129**, 991–1005.

72 Tmh,T.Q. and Sinden,R.R. (1991) *Nature (London)*, **352**, 544–547.

73 Trinh,T.Q. and Sinden,R.R. (1993) *Genetics*, **134**, 409–422.

74 Peirce,J.C., Kong,D. and Masker,W. (1991) *Nucleic Acids Res.*, **19**, 3901–3905.