Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA
Christopher E Pearson*† and Richard R Sinden*

Models for the disease-associated expansion of (CTG)$_n$°(CAG)$_n$, (CGG)$_n$°(CCG)$_n$, and (GAA)$_n$°(TTC)$_n$ trinucleotide repeats involve alternative DNA structures formed during DNA replication, repair and recombination. These repeat sequences are inherently flexible and can form a variety of hairpins, intramolecular triplexes, quadruplexes, and slipped-strand structures that may be important intermediates and result in their genetic instability.

Introduction
The expansion of triplet repeats in genomic DNA is associated with human disease (reviewed in [1]). Nine different disease loci have unstable (CAG)$_n$°(CTG)$_n$ repeats, five chromosomal fragile sites contain unstable (CGG)$_n$°(CCG)$_n$ repeats, and another disease is associated with an unstable (GAA)$_n$°(TTC)$_n$ repeat (Table 1). The expansion mutation is strongly dependent upon the repeat tract length, such that the expanded alleles are more likely to undergo further expansion [2]. In addition, the purity (or homogeneity) of the repeat tract can affect its genetic stability (Table 1). Models for expansion invoke the participation of alternative DNA structures in aberrant DNA replication, repair and recombination [1]. Here we review the unusual helical properties of the repeats and the variety of alternative DNA structures formed by them (Table 2) — characteristics that may be biologically relevant in their genetic instability.

Unusual helical characteristics associated with (CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats
Many unusual properties have been identified for DNA containing (CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats. Kohwi et al. [3] initially presented evidence for an unusual (CTG)$°$(CAG) helix, demonstrating the hyper-reactivity of even a single (CTG)$°$(CAG) trinucleotide in DNA to chloroacetaldehyde modification. This hyper-reactivity, which is dependent upon Zn$^{2+}$ or Co$^{2+}$ ions and DNA supercoiling, suggests a DNA structure in which cytosines possess unpaired character. Subsequently, unusual electrophoretic mobility, hyper-flexibility, and unusual properties of nucleosome assembly have been identified.

(CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats exhibit unusual fast mobility on polyacrylamide gels
Chastain et al. [4] reported that duplex (CTG)$_n$°(CAG)$_n$-containing DNA from the myotonic dystrophy (DM) locus and (CGG)$_n$°(CCG)$_n$-containing DNA from the fragile X mental retardation (FRAXA) locus migrated faster than expected in polyacrylamide gels (up to 20% faster for (CTG)$_{156}$°(CAG)$_{156}$). This unusual mobility is dependent upon repeat length, gel temperature, and the percentage of acrylamide. The physical basis for this unusually rapid electrophoretic mobility, which has rarely been observed previously (see, for discussion, [4,5]), is not yet understood. Based on a reptation model for electrophoresis [6], Chastain et al. [4] suggested that this increase in mobility may have resulted from a 20% increase in the apparent persistent length of the DNA from $P_{app} = 43–51$ nm, which is possibly indicative of a straighter and perhaps less flexible helix than typical B-DNA. More recent data suggest, however, that the rapid mobility may be a result of the presence of a long, flexible helical region (see below) [5].

(CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats are inherently flexible
Two recent studies have shown that (CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats are inherently flexible. (CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$ repeats exhibit unusually high molar cyclization factors ($\lambda_m$), suggesting that they are more flexible or curved molecules than normal mixed sequence B-DNA [5,7]. By analyzing the cyclization kinetics for a series of molecules with increasing numbers of repeats, Bacolla et al. [7] reported bending moduli of $1.13 \times 10^{-19}$ and $1.27 \times 10^{-19}$ erg cm for (CTG)$_n$°(CAG)$_n$ and (CGG)$_n$°(CCG)$_n$, respectively [7]. These values are 40% less than that for random B-DNA, and suggest persistence length values about 60% that for mixed sequence B-DNA. Values for the torsional moduli and helix repeat ($h$) were essentially unchanged from that for mixed sequence B-DNA. When measured using a topological band shift method, $h$ varied dramatically as a function of the repeat length — from 9.6 to 10.3 for (CTG)$_n$°(CAG)$_n$ and from 10.0 to 11.1 for (CGG)$_n$°(CCG)$_n$ [7]. This deviation in $h$ with respect to repeat length could be explained if the different lengths of repeats had a variable twist, which is inconsistent
Table 1

<table>
<thead>
<tr>
<th>Disorder/disease locus</th>
<th>Repeat</th>
<th>Normal</th>
<th>Premutation</th>
<th>Full expansion</th>
<th>Interruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragile X/A/FRAXA</td>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>6–52</td>
<td>59–230</td>
<td>230–2000</td>
<td>AGG (threshold ≥ 34 pure repeats)</td>
</tr>
<tr>
<td>Fragile XE/FRAXE</td>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>4–39</td>
<td>*(31–61)</td>
<td>200–900</td>
<td>None</td>
</tr>
<tr>
<td>Fragile XF/FRAXF</td>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>7–40</td>
<td>*</td>
<td>306–1008</td>
<td>(GCCGTC)$_3$–$4$ complex</td>
</tr>
<tr>
<td>FRA16A</td>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>16–49</td>
<td>*</td>
<td>1000–1900</td>
<td>None</td>
</tr>
<tr>
<td>Kennedy's disease/SBMA</td>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>11</td>
<td>80</td>
<td>100–1000</td>
<td>None</td>
</tr>
<tr>
<td>Myotonic dystrophy/DM</td>
<td>(CTG)$_n$*(CAG)$_n$</td>
<td>14–32</td>
<td>*</td>
<td>40–55</td>
<td>None</td>
</tr>
<tr>
<td>Huntingdon's disease/HD</td>
<td>(CAG)$_n$*(CTG)$_n$</td>
<td>5–37</td>
<td>50–80</td>
<td>80–1000, 2000–3000</td>
<td>None</td>
</tr>
<tr>
<td>Spino-cerebellar ataxia 1/SCA1</td>
<td>(CAG)$_n$*(CTG)$_n$</td>
<td>10–34</td>
<td>36–39</td>
<td>40–121</td>
<td>None</td>
</tr>
<tr>
<td>Machado Joseph disease/MJD</td>
<td>(CAG)$_n$*(CTG)$_n$</td>
<td>6–39</td>
<td>None</td>
<td>40–81</td>
<td>CAT (threshold ≥ 40 pure repeats)</td>
</tr>
<tr>
<td>Spino-cerebellar ataxia 2/SCA2</td>
<td>(CAG)$_n$*(CTG)$_n$</td>
<td>14–31</td>
<td>None</td>
<td>34–59</td>
<td>CAA</td>
</tr>
<tr>
<td>Spino-cerebellar ataxia 3/SCA3</td>
<td>(CAG)$_n$*(CTG)$_n$</td>
<td>13–44</td>
<td>*</td>
<td>60–84</td>
<td>None</td>
</tr>
<tr>
<td>Haw river syndrome/HRS (also DRPLA)</td>
<td>(GAA)$_n$*(TTC)$_n$</td>
<td>6–29</td>
<td>*(&gt;34–40)</td>
<td>200–900</td>
<td>(GAGGAA) (threshold ≥ 34–40 pure repeats)</td>
</tr>
</tbody>
</table>

*A possible mutagenic intermediate length. Not all diseases are associated with premutation and protomutation clinical or repeat length status.

Table 2

<table>
<thead>
<tr>
<th>Repeat</th>
<th>Alternative structure</th>
<th>Properties of duplex DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CGG)$_n$*(CCG)$_n$</td>
<td>S-DNA</td>
<td>Fast electrophoretic gel mobility, Hyperflexible, resists assembly into nucleosomes</td>
</tr>
<tr>
<td>(CGG)$_n$</td>
<td>SI-DNA</td>
<td></td>
</tr>
<tr>
<td>(CGG)$_n$*(CAG)$_n$</td>
<td>Intrastrand hairpins, interferstrand duplexes, quadruplex DNA</td>
<td></td>
</tr>
<tr>
<td>(CTG)$_n$*(CAG)$_n$</td>
<td>S-DNA</td>
<td></td>
</tr>
<tr>
<td>(CTG)$_n$*(CAG)$_n$</td>
<td>SI-DNA</td>
<td></td>
</tr>
<tr>
<td>(CTG)$_n$*(CAG)$_n$</td>
<td>Intrastrand hairpins, interferstrand duplexes</td>
<td></td>
</tr>
<tr>
<td>(CAG)$_n$</td>
<td>Intrastrand hairpins, interferstrand duplexes</td>
<td></td>
</tr>
<tr>
<td>(GAA)$_n$*(TTC)$_n$</td>
<td>Intramolecular triplex DNA</td>
<td></td>
</tr>
</tbody>
</table>

Differential assembly of (CTG)$_n$*(CAG)$_n$ and (CGG)$_n$*(CCG)$_n$ repeats into nucleosomes

(CTG)$_n$*(CAG)$_n$ and (CGG)$_n$*(CCG)$_n$ repeats assemble into nucleosomes with much higher efficiency than any other DNA sequences.

with the $j_\alpha$ calculations, or if the repeat had a quite different writhe from that of mixed sequence DNA. Bacolla et al. [7**] suggested that (CTG)$_n$*(CAG)$_n$ and (CGG)$_n$*(CCG)$_n$ repeats are both more flexible and more highly writhed than random B-DNA.

Chastain and Sinden showed that short (CTG)$_n$*(CAG)$_n$ tracts are inherently flexible by examining the effect of triplet repeats upon the electrophoretic mobility of curved and noncurved DNA polymers [5**]. Triplet repeats interspersed between regions of straight DNA showed a slightly reduced electrophoretic mobility, excluding the possibility that (CTG)$_n$*(CAG)$_n$ repeat tracts possess static curvature. When (CTG)$_n$*(CAG)$_n$ was interspersed between A-tract curves, the reduced electrophoretic mobility due to DNA curvature was dramatically abrogated, consistent with interspersion of a torsionally flexible and/or bendable region that would destroy the helical phasing required for A-tract curvature. The effect of (CTG)$_n$*(CAG)$_n$ was almost identical to that afforded by two consecutive T* T mismatches [(TT)$(T'T')$]. This study shows that even the very short (CTG)$_n$*(CAG)$_n$ tract imparts a very large degree of flexibility to a DNA helix [5**].
yet identified [8-10], whereas (CGG)\textsubscript{n}(CCG)\textsubscript{n} repeats assemble into nucleosomes with the least efficiency of any known sequence [11,12]. This difference is surprising given the general observation that flexible DNAs and DNAs that contain inherent curved regions typically assemble preferentially into nucleosomes [13]. Presumably the inherent flexibility of the (CTG)\textsubscript{n}(CAG)\textsubscript{n} repeats allows the sequence to easily adopt the conformation required in order to stably bend around nucleosomes. (CGG)\textsubscript{n}(CCG)\textsubscript{n} repeats are inherently flexible, although they may not easily adopt the stable radius of curvature required for the assembly of DNA into nucleosomes. Length-dependent differences have been reported for the effects of methylation at CpG dinucleotides within (CGG)\textsubscript{n} repeats: methylated short repeats (n = 13) and long repeats (n = 74-76) assemble into nucleosomes more and less easily than nonmethylated repeats, respectively [12,14]. Both triplet repeats are flexible [5*,7**], although they almost certainly have very different biological properties in terms of the chromatin organization upon expansion in the disease state, and also depend upon the state of methylation.

**Inter- and intramolecular association of triplet repeats: mismatched duplexes and hairpins**

Single-stranded oligonucleotides of (CGG)\textsubscript{n} or (CGG)\textsubscript{n} repeats can form intermolecular duplexes and intramolecular hairpin duplexes (Figure 1). On native polyacrylamide gels, the migration of intermolecular triplet hairpins is characteristically faster than that of their complementary base-paired duplexes [15-23], whereas the migration of the intermolecular duplexes is slower than that of their corresponding Watson–Crick duplexes [20,23,24].

Two alternative schemes of inter or intrastrand (hairpin) mispairing exist for (CGG)\textsubscript{n} repeats: one in which the G of the GpC step participates in a (G).(G) mismatch.

**Figure 1**

<table>
<thead>
<tr>
<th>(a)</th>
<th>Watson–Crick</th>
<th>(b)</th>
<th>Interstrand</th>
<th>(c)</th>
<th>Hairpins</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>(\text{CGG CGG CGG CGG CGG CGG - } )</td>
<td>(i)</td>
<td>(\text{CGG CGG CGG CGG CGG CGG - } )</td>
<td>(i)</td>
<td>(\text{CGG CGG CGG CGG CGG CGG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GCC GCC GCC GCC GCC GCC} - )</td>
<td>(\text{GCC GCC GCC GCC GCC GCC} - )</td>
<td>(\text{GCC GCC GCC GCC GCC GCC} - )</td>
<td>(\text{GCC GCC GCC GCC GCC GCC} - )</td>
<td></td>
</tr>
<tr>
<td>(ii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(ii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(ii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GAC GAC GAC GAC GAC - } )</td>
<td>(\text{GAC GAC GAC GAC GAC - } )</td>
<td>(\text{GAC GAC GAC GAC GAC - } )</td>
<td>(\text{GAC GAC GAC GAC GAC - } )</td>
<td></td>
</tr>
<tr>
<td>(iii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(iii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(iii)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td></td>
</tr>
<tr>
<td>(iv)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(iv)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(iv)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td></td>
</tr>
<tr>
<td>(v)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(v)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(v)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td></td>
</tr>
<tr>
<td>(vi)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(vi)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
<td>(vi)</td>
<td>(\text{CTG CTG CTG CTG CTG - } )</td>
</tr>
<tr>
<td></td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td>(\text{GTC GTC GTC GTC GTC - } )</td>
<td></td>
</tr>
</tbody>
</table>

**Current Opinion in Structural Biology**

Structures assumed by the individual single strands of triplet repeats. (a) Canonical Watson–Crick (CGG)\textsubscript{n}(CCG)\textsubscript{n} and (CTG)\textsubscript{n}(CAG)\textsubscript{n} duplexes between complementary strands containing triplet repeats (i and ii, respectively). (b) Intermolecular duplexes formed between individual (CGG)\textsubscript{n}, (CCG)\textsubscript{n}, and (CTG)\textsubscript{n} strands (i, iii and iv, respectively). Intermolecular duplexes have not been reported for (CAG)\textsubscript{n} strands. Single strands of (CCG)\textsubscript{n} can form an interstrand duplex, called an e-motif, in which the Cs of the GpC dinucleotide step are extrahelical. (CTG)\textsubscript{n} and (CGG)\textsubscript{n} duplexes contain T.T and GoG mispairs, respectively. (c) A variety of intramolecular duplexes (hairpins) have been reported for (CGG)\textsubscript{n}, (CCG)\textsubscript{n}, and (CAG)\textsubscript{n} (vi) repeats. (CGG)\textsubscript{n} and (CCG)\textsubscript{n} repeats adopt two different folding patterns in which the GpG or CpC mismatch is flanked by either (CG)\textsubscript{n}(CG)\textsubscript{n} or (GC)\textsubscript{n}(GC)\textsubscript{n} (i or ii and iii or iv). The preferred structures are indicated in bold. The larger font denotes the mispaired or extrahelical bases.
The loops of the CGG hairpins are sensitive to diethyl pyrocarbonate modification [22].

(CCG)₁₀ hairpins can contain mismatched Cs that involve the C of the CpG and GpC steps (Figure 1cii and iv). The former appears to be the preferred conformation [20], but this may be influenced by the length [21]. (CCG)₁₀ with n = 2–5, forms a staggered interstrand duplex stabilized by Watson–Crick (CG)°(CG) base-pairs, in which the Cs of the GpC dinucleotide are extrahelical and located diagonally across the minor groove (called an e-motif) (Figure 1bii) [25]. The extrahelical Cs appear to be in equilibrium with the protonated stacked form. The longer (CCG)₁₅ tract appears to form an intramolecular hairpin [26] with characteristics very similar to an e-motif in which the C residues are unpaired and extrahelical. (CCG)₁₀ hairpins are reportedly more stable than (CGG)₁₀ hairpins [20].

Single stranded (CTG)₁₀ or (CTG)₃₀ oligonucleotides can form inter and intramolecular duplexes (Figure 1b,c) [23]. (CAG)₁₀ and (CAG)₃₀, however, only form intramolecular hairpin structures. The hairpin loop typically contains three or four bases [27]. The tip of the hairpin loops of both (CTG)₁₀ and (CAG)₁₀ hairpins are sensitive to P1 nuclease [17,18]. Only at high temperatures (60–70°C) were the (T)°(T) mismatches in the (CTG)₁₀ hairpins [28,29] measured the length independence of the Tₘ values for (CTG)₁₀ or (CAG)₁₀ hairpins; the differences were 0°C and 1°C, respectively, for repeats of n = 10 and 30 [23]. Similarly, only a minor difference of 7°C existed between the Tₘ values for (CTG)₁₀•(CAG)₁₀ and (CAG)₁₀•(CAG)₃₀ duplexes. These findings contrast with the predicted direct relationship between repeat length and Tₘ [21,30]. The actual length independence of Tₘ suggests that rather than forming a single large hairpin, the longer tract may form multiple short hairpin structures [23].

Hairpin base-base mismaps

In (CGG)ₙ hairpins with n = 11–15, the (G)°(G) mismatches appear to be hydrogen bonded [22,24], whereas with n = 3, the mismatches appear to be unpaired and flexible in various dynamic conformational isomers [21]. Additionally, the mismatches alternate G(syn)°G(anti) throughout the GGG hairpin [16,20,24]. Although no C–C interactions should exist in the e-motif, hydrogen bonds have been detected in (CCG)ₙ hairpins with n = 5–7. This is consistent with the presence of a (G)°(G) mismatch [21,24]. Longer (CCG)ₙ molecules, with n > 15, appear to form an e-motif-like intramolecular hairpin [26], in which the cytosine residues are unpaired and extrahelical. In (CTG)ₙ hairpins, the (T)°(T) mismatches are stacked into the helix and appear to be stabilized by two hydrogen bonds [21,27,28], similar to a wobble (G)°(T) mispair [29]. In (CAG)ₙ hairpins, the (A)°(A) mismatches exhibit conformational instability, showing little, if any, base–base interaction [21,28]. These results provide an explanation for the reduced biophysical stability of CAG hairpins compared with CTG hairpins [17,23,28,30].

In either the (CCG)ₙ or (CCG)₁₀ intermolecular duplexes or intramolecular hairpins, the cytosine residues of the CpG step are either (C)°(C) mismatches, or are destabilized by either the adjacent mispaired (G)°(G) residues or the adjacent extrahelical cytosine residues (e-motif) (Figure 1). Interestingly, the cytosine residues of the CpG step are methylated in expanded FRAXA repeats. Presumably, the destabilization facilitates the flipping of the cytosine away from the helix and into the active site of the CpG methylase.

Thermodynamics of mispaired duplexes

Analysis of the thermodynamic parameters (Tₘ,ΔH, and ΔGₛₘ) of single-stranded repeats revealed that the formation of intrastrand structures (hairpins) from Watson–Crick duplexes is favored for the disease-associated (CTG)₁₀•(CAG)₁₀ and (CAG)₁₀•(CAG)₃₀ repeats relative to the nondisease-associated (GTC)₁₀•(GAC)₁₀ repeats [21].

Petruska et al. [23] measured the length independence of the Tₘ values for (CTG)₁₀ or (CAG)₁₀ hairpins; the differences were 0°C and 1°C, respectively, for repeats of n = 10 and 30 [23]. Similarly, only a minor difference of 7°C existed between the Tₘ values for (CTG)₁₀•(CAG)₁₀ and (CAG)₁₀•(CAG)₃₀ duplexes. These findings contrast with the predicted direct relationship between repeat length and Tₘ [17,30]. The actual length independence of Tₘ suggests that rather than forming a single large hairpin, the longer tract may form multiple short hairpin structures [23].

Triplex DNA structures in (GAA)ₙ•(TTC)ₙ

Homopurine·homopyrimidine tracts that contain mirror repeat symmetry can form four different intramolecular triplex DNA structural isomers (see, for review, [31]). This sequence requirement is met by the (GAA)ₙ•(TTC)ₙ tract, which forms intramolecular triplex DNA (Figure 2a) [32]. Within this sequence, the 5’ half of the TTC strand folds into the major groove of the other half of the Pu•Py tract, thus forming Hoogsteen hydrogen bonds and creating C°•G•C and 7°•A°•T base triads (the italicized base represents the third strand). The formation of this DNA structure is favored by a low pH (as a result of the required protonation of the third strand cytosine) and DNA supercoiling. These structures can exist in cells [31]. An intramolecular triplex may also form if an appropriate single-stranded triplet-containing strand is available (Figure 2b).
The triplex DNA structures formed by \((GAA)_n(TTC)_n\) repeats. (GAA)_n*(TTC)_n repeats can form (a) intramolecular triplex DNA structures in which the 3' half of the TTC tract folds into the major groove of the 5' half of the repeat tract [31]. (b) Intermolecular triplex structures may also form provided a single pyrimidine-rich groove of the 5' half of the repeat tract is available. A filled bullet (*) indicates Watson-Crick hydrogen bonds. An empty bullet (○) indicates Hoogsteen hydrogen bonds.

**Formation of quadruplex DNA structures within triplet repeats**

Several reports have shown that, under some conditions, quadruplex DNA structures (Figure 3) can form within single-stranded CGG tracts. Fry and Loeb [33] showed that short oligonucleotides composed of CGG residues (but not CCG) formed complexes in the presence of K+, Na+, or Li+ ions that migrated as four times their molecular weight on polyacrylamide gels. Usdin and Woodford [34] have interpreted specific pauses in *in vitro* replication as an indication of quadruplex formation. The pause required the presence of K+ ions and the CGG template strand, conditions required for quadruplex formation. Chemical modification studies, including protection of the N7 of guanine [33,34], have supported the contention that guanines are involved in Hoogsteen hydrogen bonding, a characteristic of quadruplex DNA (Figure 3c). Kettani *et al.* [35] detected the quadruplex with two G*C*G*C tetrads flanked by two G*G*G*C tetrads (G quartets) shown in Figure 3d. The Kettani structure, which was expected from studies of the folding of an intramolecular hairpin (Figure 3f), is different from that proposed by Usdin and Woodford [34] (Figure 3e). The design of the Kettani sequence probably precluded the formation of any other quadruplex, and it is not known whether this would be a preferred folding scheme within longer repeats. Many triplet repeat tracts clearly form hairpins, and these triplet repeats may fold into quadruplex or other higher order structures. A rigorous demonstration of any such structures, however, is not yet available.

**Slipped-stranded DNA structures**

There are two types of slipped-stranded DNA structures — 'homoduplex' Slipped structures (S-DNA) [36], which are formed between two complementary strands with the same number of repeats paired in an out of register fashion, and 'heteroduplex' Slipped Intermediates (SI-DNA) [37*], in which the strands have different numbers of repeats (Figure 4). SI-DNA may arise through replication slippage, resulting in repeat expansions or deletions (Figure 4a), whereas strand slippage in the absence of replication will result in S-DNA (Figure 4b).

**S-DNA**

Although long proposed (see, for reviews, [1*,31]), the first biochemical evidence for the existence of slipped-stranded DNA structures (S-DNA) was obtained using DNA containing (CTG)_n*(CAG)_m and (CGG)_n*(CCG)_m repeats [36]. S-DNA formed following the denaturation and renaturation (reduplexing) of DNA. S-DNAs migrate in polyacrylamide gels as a broad distribution of distinct products ranging up to twice their actual size. This reduced gel mobility is consistent with that expected for DNA containing bends introduced from the three- and four-way junctions generated by the looped-out strands [38]. The broad distribution was as expected from the multiple possibilities for strand slippage involving different lengths of loop outs at multiple locations throughout the repeat tract (see Figure 4b). In fact, the complexity of the S-DNA populations, as well as the percentage of S-DNA formed, increased with increasing length of the repeat tract [36,37*,39*,40*]. The effect of length on the propensity of S-DNA formation correlates with the effect of length on genetic stability. Electron microscopic analyses confirm that S-DNA secondary structure occurs within the repeat tract, that the S-DNA is subtended by linear mixed-sequence duplex DNA, and that the S-DNAs are a heterogeneous population of products [39**]. S-DNA does not require superhelical tension for formation or stability; the structures are remarkably thermostable, and display minimal interconversion between isomers and the linear duplex form under physiological conditions [36,37*,39**]. If the looped-out regions of the slipped-stranded DNA were unpaired, then branch migrations could occur, resulting in the reformation of linear duplex DNA. The remarkable stability of (CTG)_n*(CAG)_m and (CGG)_n*(CCG)_m S-DNAs may result from the formation of intrastrand hairpins, as branch migration would require breaking base pairs at the slipped-out junction and within the slipped-out hairpin.
The quadruplex structures formed by triplet repeats. Guanine-rich DNA sequences can form a variety of quadruplex DNA structures including those shown in (a–c) (see, for a review, [31]). Short tracts of thymine (not shown for clarity) typically intersperse guanine tracts. (d) The quadruplex structure detected by Kettani et al. [35]. (e) The quadruplex structure suggested by Usdin and Woodford [34]. This could form within a single (CGG)n tract. The duplex/hairpin precursor shown, involving G-G Hoogsteen hydrogen bonds, has not been observed and quadruplex formation probably occurs without formation of this intermediate. The precursor shown with the Kettani-type quadruplex (f) is the preferred conformation of a (CGG)n intrastrand hairpin. The quadruplex shown in (e) may be more probable than the Kettani-type quadruplex due to a higher proportion of G-quartets.

The effect of sequence interruptions upon S-DNA formation

Sequence interruptions within several trinucleotide repeat tracts confer an increased genetic stability upon the repeat tract.
The inhibitory structural phenomenon specific to interrupted repeats in one-stranded DNAs (S-DNAs) can form following denaturation and renaturation of a triplet repeat tract in an out of register fashion (i-iv). Following continued replication, the DNA contains an excess of SI-DNAs that can be electrophoretically separated, suggesting that they are structurally distinct [37]. Similarly, two sister (CGG)**(CCG) SI-DNAs could also be electrophoretically separated (CE Pearson, unpublished data). The structural differences between SI-DNAs containing either looped-out CTG or CAG repeats may reflect the different single-stranded character of the looped-out strands [36,37]. In S-DNA, the CAG strand is preferentially susceptible to mung bean nuclease compared to the CTG strand [36], indicating the greater single-stranded character of the CAG strand (consistent with the lower thermal stability of the CAG hairpin relative to the CTG hairpin). Structural differences between sister SI-DNAs have also been detected by the binding of the human mismatch repair protein hMSH2 [37].

**Conclusions**

DNA is not a passive participant in its procreation. Through subtle variations in helix shape, DNA controls its intimate association with regulatory proteins. Through variations in helix trajectory, stiffness, and flexibility, the DNA can specify both its pattern of conjugal assembly with nucleosomes, or its organization into specific three-dimensional structures that regulate gene expression or facilitate genetic recombination. Within the past seven years, a novel type of repeat expansion mutation leading to human disease, called dynamic mutation [2], has emerged. In repeat expansion models, the repeat tract is considered to be a single-stranded DNA (S-DNA) that is capable of forming secondary structures such as a hairpin through sequence interruptions. Sequence interruptions of either (CAG)**(CTG) or (CGG)**(CCG) repeats suggest a possible molecular explanation for the biology associated with the protective mechanism of interruptions; interruptions may inhibit the formation of slipped mutagenic intermediates during replication and, in this fashion, inhibit genetic length changes. Thus, the effects both of the length and purity of the repeat tract on the propensity of S-DNA formation correlates with their effects on genetic stability and disease.

**Slipped intermediate DNA structures (SI-DNA)**

SI-DNAs have been formed by reannealing DNAs containing two different lengths of repeats [37]. SI-DNAs migrate more slowly in polyacrylamide gels than their respective linear duplexes or S-DNAs that are formed from either length of repeat. Two sister SI-DNAs composed of (CTG)\textsubscript{n}*(CAG)\textsubscript{n} and (CTG)\textsubscript{n}*(CAG)\textsubscript{n} (where n ≠ m) could be electrophoretically separated, suggesting that they are structurally distinct [37]. Similarly, two sister (CGG)**(CCG) SI-DNAs could also be electrophoretically separated (CE Pearson, unpublished data). The structural differences between SI-DNAs containing either looped-out CTG or CAG repeats may reflect the different single-stranded character of the looped-out strands [36,37]. In S-DNA, the CAG strand is preferentially susceptible to mung bean nuclease compared to the CTG strand [36], indicating the greater single-stranded character of the CAG strand (consistent with the lower thermal stability of the CAG hairpin relative to the CTG hairpin). Structural differences between sister SI-DNAs have also been detected by the binding of the human mismatch repair protein hMSH2 [37].
Models for the protective mechanism of repeat sequence interruptions. The open circles represent (CAG)°(CTG) and (CGG)°(CCG) repeats, and the filled circles represent the (CAT)°(ATG) and (AGG)°(CCT) interruptions. (a) The reduced percentage of S-DNA formed by an interrupted repeat tract relative to a pure (uninterrupted) tract may result from the reduced stability of the inter and intrastrand slipped DNAs containing these interruptions. These sequence interruptions will destabilize both interstrand and intrastrand duplexes. (b) Sequence interruptions may limit the position, and thereby the length, of looped-out regions. Loop-outs in sequence interrupted repeat tracts may occur preferentially between the sequence interruptions or form with the sequence interruptions at the tip of a hairpin loop (not shown). (c) Pure repeat tracts can have loop-out strands of variable lengths throughout the repeat tract.
unusual helical properties, such as very flexible DNA, and alternative DNA secondary structures, including a variety of intrastrand hairpins containing mispairs, slipped-strand alternative DNA secondary structures, including a variety and loci-specific expansion leading to human disease although any direct relationship between DNA structure (GAA)n'(TTC)n repeats represent dynamic DNA associations. Clearly, (CTG)n'(CAG)n, (CGG)n'(CCG)n, and (GAA)n'(TTC)n repeats represent dynamic DNA associated with dynamic mutations. Much has been revealed, although any direct relationship between DNA structure and loci-specific expansion leading to human disease remains, at present, an enigma.

Note added in proof
While this manuscript was in preparation, there were two publications indicating the formation of triplex DNA structures in (GAA)n'(TTC)n repeats [50,51] and one demonstrating the in vivo formation of hairpins from long inverted repeats having (CGG)n'(CCG)n repeats at their centre [52].

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:
• of special interest
** of outstanding interest

This article contains a thorough review of triplet-repeat diseases, the DNA sequences involved, and expansion models. It is part of a comprehensive book describing the current state of this field.


5. Chastain PD, Sinden RR: CTG repeats associated with human genetic disease are inherently flexible. J Mol Biol 1998, 275:405-411. This paper demonstrates that (CTG)n'(CAG)n repeats are unusually flexible. Internspection of (CTG)n'(CAG)n between A-tract curves abrogates overall DNA curvature to the same extent as interspersion of two consecutive T-T mismatches. This is a remarkable degree of flexibility for a very short repeat tract.


This paper presents a careful analysis of the characteristics of a d(CG)n oligonucleotide hairpin, which forms the 'e-motif'. The methods used in this paper are representative of those used in many similar analyses of this and other triplet repeats published during 1995 and 1996.


This paper presents the purification, characterization, and mechanism of binding to the human mismatch repair protein (hMSH2) to SI-DNA and S-DNA. SI-DNA (slipped intermediate DNA) contains heteroduplex triplet repeat regions in which the lengths of the repeat tracts in the complementary strands are different. Sister heteroduplexes migrate to different positions on a polyacrylamide gel. hMSH2 preferentially binds to SI-DNA containing looped out CAG residues compared with CTG strands.


A paper that continues the description of the properties and characteristics of the S-DNA structures formed in (CTG)n(CAG) repeats. Following denaturation and renaturation, a complex, heterogeneous distribution of S-DNA isomers was formed. These structures were characterized biochemically and visualized using electron microscopy. The majority of S-DNASs from (CTG)n(CAG) repeats contain short slipped-out regions that are visible only as kinks. Other structures, containing the loops and three- and four-way junctions expected of slipped-stranded structures, are visible, especially in (CTG)155(CAG)155 S-DNASs.


This paper demonstrates the fact that sequence interruptions within triplet repeat tracts reduce both the percentage and the complexity of the S-DNA formed following denaturation and renaturation. A model is proposed for this effect that correlates with the effect of repeat interruptions upon the stability of repeats in humans.


This paper reveals stalling of the replication fork within the repeat tract in Escherichia coli. Pausing is dependent upon repeat length and orientation with respect to the origin of plasmid replication. This paper is representative of many showing an in vivo biological effect of triplet repeats, and the authors suggest that the formation of an unusual DNA structure is responsible. To date, however, these papers have not presented biochemical data consistent with the formation of any alternative triplet repeat DNA structure in vivo. This is likely to be an important area for future research.


