Computational and Structural Biotechnology Journal 19 (2021) 2819–2832

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Molecular conformations and dynamics of nucleotide repeats associated with neurodegenerative diseases: double helices and CAG hairpin loops

Feng Pan [a], Yuan Zhang [a], Pengning Xu [b], Viet Hoang Man [b], Christopher Roland [b], Keith Weninger [b], Celeste Sagui [b,*]

[a] *Department of Statistics, Florida State University, Tallahassee, FL 32306, USA*
[b] *Department of Physics, North Carolina State University, Raleigh, NC 27695, USA*

ARTICLE INFO

Article history:
Received 3 February 2021
Received in revised form 13 April 2021
Accepted 16 April 2021
Available online 26 April 2021

Keywords:
Trinucleotide hexanucleotide repeats
CAG
GAC
CTG
CUG
CCG
GGC
GGGGCC
GGCCCC
Hairpin loops
smFRET
Molecular dynamics

ABSTRACT

Pathogenic DNA secondary structures have been identified as a common and causative factor for expansion in trinucleotide, hexanucleotide, and other simple sequence repeats. These expansions underlie about fifty neurological and neuromuscular disorders known as “anticipation diseases”. Cell toxicity and death have been linked to the pathogenic conformations and functional changes of the RNA transcripts, of DNA itself and, when trinucleotides are present in exons, of the translated proteins. We review some of our results for the conformations and dynamics of pathogenic structures for both RNA and DNA, which include mismatched homoduplexes formed by trinucleotide repeats CAG and GAC; CCG and CGG; CTG(CUG) and GTC(GUC); the dynamics of DNA CAG hairpins; mismatched homoduplexes formed by hexanucleotide repeats (GGGGCC) and (GGCCCC); and G-quadruplexes formed by (GGGGCC) and (GGGCCT). We also discuss the dynamics of strand slippage in DNA hairpins formed by CAG repeats as observed with single-molecule Fluorescence Resonance Energy Transfer. This review focuses on the rich behavior exhibited by the mismatches associated with these simple sequence repeat noncanonical structures.

Contents

* Corresponding author.
E-mail address: sagui@ncsu.edu (C. Sagui).

https://doi.org/10.1016/j.csbj.2021.04.037
2001-0370/

1. Introduction

Simple sequence repeats (SSRs) typically consist of units of 1 to 6 nucleotides that are repeated up to 30 times, or more [1]. They represent about 3% of the entire human genome sequence [2]. Trinucleotide repeats (TRs) represent one of the most common type of SSRs in the exome of all eukaryotic genomes [3]. TRs may be selectively neutral sequences, or may play an important functional role. Many TRs exhibit "dynamic mutations" that do not follow Mendelian inheritance, which asserts that mutations in a single gene are stably transmitted between generations [4]. This can lead to genetic diseases where, with successive generations, the age of disease onset decreases and the disease severity increases, and the probability that this type of mutation results in disease also increases [5]. These mutations are caused by the intergenerational expansion of TRs. In addition, the repeats increase their length in somatic cells during the lifespan of the affected individual. After a certain threshold in the repeat number of the TR, the probability of further TR expansion and severity of the diseases increases with the number of repeats. In particular, the dynamic mutations in human genes associated with TRs cause severe neurodegenerative and neuromuscular disorders known as Trinucleotide (or Triplet) Repeat Expansion Diseases (TREDs) that lead to cell toxicity and death [6–8].

To date, approximately fifty DNA expandable SSR diseases have been identified and their number is expected to grow [9,10]. See Fig. 1 for a schematic illustrating some of the most common SSRs (note that they are mostly TRs). The SSR expansions are believed to be caused by some sort of slippage during DNA replication, repair, recombination or transcription. Cell toxicity and death have been linked to the pathogenic conformation and functional changes of the RNA transcripts, of DNA itself and, when TRs are present in exons, of the translated proteins [10,11], mainly in the group of polyglutamine (polyQ) diseases. Abnormal nuclear foci can result when the expanded RNA transcripts with pathogenic-related secondary structures sequester regulatory proteins. What makes these pathological mechanisms even more complicated is that antisense transcripts of the expansions – which result from the bidirectional transcription of the DNA TRs [12,13] can also form nuclear RNA foci that contribute to toxicity, and that both sense and antisense expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional repeat-associated non-ATG (RAN) translation [14]. Table 1 lists some of the most common SSRs and the associated diseases along with the normal and pathological range of repeats, as well as a short-hand notation for the most common molecular mechanism behind the disease. Much is still unknown about these mechanisms, thus our notation is only indicative of some common experimental findings. For instance, in the polyQ diseases listed on the table, polyQ stretches are well known to trigger various abnormal cellular processes that lead to neurodegeneration. However, mutant transcripts formed by expanded CAG repeats are also toxic and contribute to cellular pathology, but the exact nature and relevance of RNA toxicity in polyQ diseases are only starting to be studied [15].

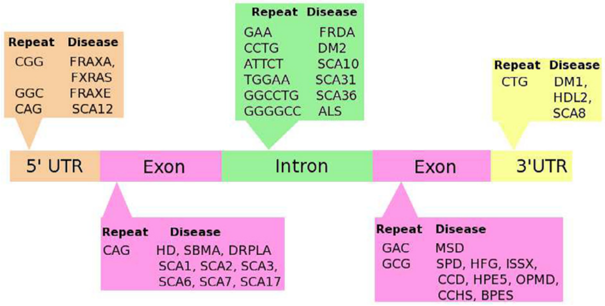

Fig. 1. Schematic illustrating the occurence of some of the most common SSRs (note that most are TRs), and abbreviations of the most common diseases that they lead to.

Although the mechanisms underlying TREDs are believed to be extremely complex, it turns out that some simple trends are remarkably robust. In particular, there is a correlation between the repeat number beyond the repeat threshold and the probability of further expansion and increased pathology. Another important breakthrough has been the recognition that the critical step in all models of repeat instability is the transient formation of pathogenic non-B DNA stable secondary structures in the expandable repeats [16]. In fact, expandable repeats are known to display pathogenic structural characteristics such as hairpin structures, Z-DNA, triple helices, G-quadruplexes and various slipped-stranded duplexes. Clearly, it is important to understand the structural and dynamical characteristics of these pathogenic secondary structures that trigger the cascade of molecular mechanisms ultimately resulting in disease.

In this article, we review our work in the area of pathogenic secondary structures of SSR nucleotides associated with TREDs. We discuss both classical Molecular Dynamics (MD) simulations and experimental results based on single-molecule Fluorescence Resonance Energy Transfer (smFRET) techniques [17]. Primarily, we have been concerned with the structure and dynamics of DNA and RNA double helix and loop structures associated with the most common TRs and hexanucleotide repeats. These include for DNA/RNA CAG, GAC, CCG, GGC, and CTG (CUG for RNA) and GTC (GUC) TRs, and GGGGCC and CCCCGG hexanucleotide repeats. For the latter, we discuss not only the double helices but also the G-quadruplex structures that may be formed.

2. Results

Hairpins represent perhaps the most common pathogenic secondary structure associated with TREDs [18–28]. Because relatively little is known about these structures at the molecular level, our work has focused on the structural and dynamical characteristics of both homoduplexes (representing the stem of a long hairpin) and hairpins. We have used large scale classical MD simulations using the AMBER simulation package [29] with state-of-the art force fields to explore the DNA and RNA duplexes associated with selected TRs which are characteristic of hairpin stems [30–32], and in the case of CAG TRs, the loops as well [17]. The loops were investigated both experimentally by means of smFRET techniques [17] augmented with MD simulations. In addition, we have also studied the case of GGGGCC/GGCCCC hexanucleotide repeats, characterized by both duplex and, in the case of the G-rich repeat, G-quadruplex structures [33,34]. Since structural transitions in DNA/RNA typically take place over time scales often precluded by straightforward MD simulations, we used special methods such as the Adaptively Biased Molecular Dynamic (ABMD) method with suitably chosen collective variables [35] to explore duplex and loop conformations. In addition, Steered Molecular Dynamics (SMD) [36] was used to explore transition mechanisms.

2.1. Helical homoduplexes and hairpins formed by CAG and GAC TRs

Of all the different known SSRs, it is the CAG TRs that are associated with the largest number of neurodegenerative diseases. Spinocerebellar ataxia type 12 (SCA12) is the result of CAG repeats in the 5'-UTR part of the PPP2R2V gene. Associated with the exon

Table 1
Table summarizing most common SSRs and associated diseases. Here NRR indicates the normal range of repeats and PRR the pathological range of repeats. In terms of the mechanisms, polyQ indicates the polyglutamine diseases briefly mentioned in main text; RNA multiple indicates either RNA loss or gain of function (sometimes both). Other mechanisms include abnormal methylation, impaired transcription leading to defective proteins, etc.

SSR	Associated diseases	NRR	PRR	Mechanism
CAG	Huntington's Disease (HD)	6–35	36–250	polyQ
	Spinal and bulbar atrophy (SBMA)	4–34	35–72	polyQ
	Dentatorubal-pallidolysian atrophy (DRPLA)	6–35	49–88	polyQ
	Spinocellular ataxia 1 (SCA1)	6–35	35–72	polyQ
	SCA2	14–32	33–77	polyQ
	SCA3	12–40	55–86	polyQ
	SCA6	4–18	21–30	polyQ
	SCA7	7–17	38–120	polyQ
	SCA12	7–41	43–51	polyQ
	SCA17	25–42	47–63	polyQ
GAC	Epiphyseal dysplasia	5	6	Impaired transcription
	Pseudoachondroplasis	5	4 or 7	Impaired transcription
CGG	Fragile X mental retardation (FRAXA)	6–60	230+	Abnormal methylation
	Fragile X tremor ataxia syndrome (FXTAS)	6–53	55–200	Increased expression
CCG	X-linked mental retardation (FRAXE)	6–39	200+	Abnormal methylation
CTG	Myotonic dystrophy type 1 (DM1)	5–37	50+	RNA based
	SCA8	16–34	74+	Unknown
	HD L2	7–28	66–78	polyQ
CCTG	Myotonic dystrophy type 2 (DM2)	10–26	75+	RNA based
GGGGCC	Amyotropic lateral sclerosis (ALS)	20	70+	RNA multiple
	Frontotemporal dementia (FTD)	20	70+	RNA multiple
GGGCCT	SCA36 (SCA36)	5–14	800+	RNA multiple
GAA	Friedreich's ataxia (FRDA)	7–34	100+	Impaired transcription
ATTCT	SCA10 (SCA10)	10–20	500+	Unknown
TGGAA	SCA31	0	560+	RNA multiple

part of various other genes with CAG repeats are nine late-onset progressive neurodegenerative disorders such as Huntington's disease (HD), spinal and bulbar atropy (SBMA), dentatorubral-pallidolysian atrophy (DRPLA), as well as several other spinocerebellar ataxias (SCAs), some of which are summarized in Table 1. These are also generically termed polyglutamine (polyQ) diseases [37], since the CAG expansions in these genes lead to polyglutamine expansions despite the fact that – depending on the reading frame – the codons CAG, AGC and GCA code for glutamine, serine and alanine, respectively. Polyglutamine diseases are associated with expansions greater than a specific repeat length [37], which is also a characteristic of other TREDs. For example, in HD the normal CAG repeat number is between 10 to 34 repeats, while repeats in the 36 to 250 range are pathologically high leading to disease expression. While different TREDs have different pathologies, they all share a similar feature: the formation of polyglutamine aggregates [38] where the fully formed fibrils are held together by cross-β conformations, which eventually result in neuronal death [39,40,33].

Subsequent to the understanding that CAG repeats are associated with neurological diseases, it was also discovered that GAC TRs are involved with a very different set of diseases from the TREDs. These specific diseases are the result of relatively small changes in the repeat number. The human gene for cartilage oligomeric matrix protein, for example, is characterized by a $(GAC)_5$ repeat. Epiphyseal dysplasia is caused by the expansion by one repeat; expansion by two repeats (or deletion by one repeat) results in pseudoachondroplasis [41]. The specific structure of the different duplexes depend on the pH of the solution and the ionic strength [42]. While the CAG trinucleotide leads to expansion, the GAC trinucleotide does not (except for at most two extra repeats). This perhaps can be attributed to the fact that CAG tracks are over-represented in the human genome, while GAC tracks are not. Indeed, a study in 2010 [43] showed that tracks equal or longer than six repeats occur 1055 times in the human genome with 300 tracks in the exons, while GAC tracts appear 16 times in the entire human genome, with only three tracts located in the exons. If the track length is increased to ten or more repeats, there are 136 CAG tracks in the human genome with 33 tracks in the protein coding regions, which represents a 43-fold enrichment over a random expectation value [44].

2.1.1. Helical homoduplexes formed by CAG and GAC TRs

Given the importance of CAG TRs (where the Watson–Crick base pairs between the mismatches exhibit GpC steps) and GAC TRs (with CpG steps between the Watson–Crick base pairs), we have investigated their helical homoduplex structures with a focus on the A$\cdot$A mismatches [30]. Our main results are as follows. The global minimum conformation of the duplexes is characterized by the A$\cdot$A mismatches stacked inside the core of the helix with nucleotide torsion angles in an *anti-anti* conformation for RNA (this corresponds to torsion angles of $\sim 180 - 200^{\circ}$) and *(high-anti)-(high-anti)* for DNA (torsion angles of $\sim 230 - 260^{\circ}$). In terms of free energy, the next minimum corresponds to *anti-syn* conformations, followed by *syn-syn* conformations which pay the highest price in free energy. These conformations are illustrated in Fig. 2 and the results are consistent with experimental X-ray studies on CAG-RNA homoduplexes, in which the mismatches are in an *anti-anti* conformation and/or anti-syn depending on the mismatch flanking sequences [45]. The differences between the RNA and DNA anti conformations is explained by the presence of the additional hydroxyl group characteristic of RNA sugar ring. This hydroxyl group interacts with the RNA backbone, pulling the sugar ring at one end and causing a twist in the other, thereby leading to an overall reduction of the torsion angles [30].

In terms of dynamics, we find that DNA helices near the global minimum are very dynamic, characterized by large fluctuations [30]. RNA helices also fluctuate, but to a considerably lesser degree. The most relevant fluctuations of the DNA helix correspond to a coupling between the bending and unwinding modes of the helix. RNA helices close to the global free energy minimum are very stable. They exhibit a wider major groove and a substantial decrease of the inclination angle with respect to the canonical A-RNA form. We have also studied transitions from *anti-syn* → *anti-anti* and *syn-syn* → *anti-syn*, and different mechanisms have been identified for both the major and minor grooves. These transitions

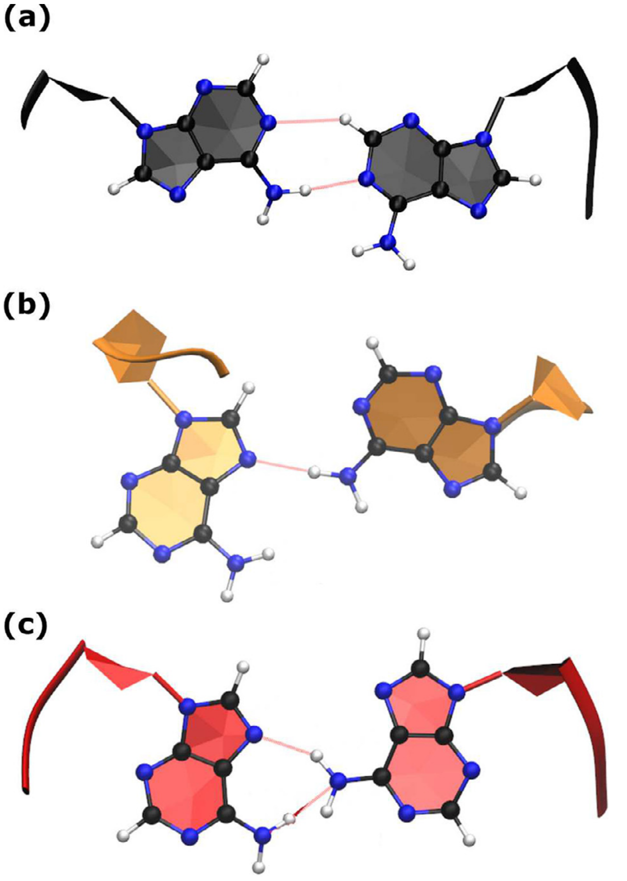

Fig. 2. Configurations for the A·A mismatches for CAG TRs: (a) anti-anti; (b) anti-syn; (c) syn-syn. Associated hydrogen bonds are indicated [30].

involve local distortions of the duplexes around the mismatches. For the *anti-syn* → *anti-anti* transition, the mechanism in both major and minor grooves occurs through base flipping. The *syn-syn* → *anti-syn* transition, on the other hand, involves base flipping in the minor groove and a combination of base stacking and rotation in the major groove. CAG-DNA and GAC-DNA homoduplexes in their *anti-anti* conformations experience some degree of unwinding, with unwinding in CAG-DNA occurring at the mismatches and in GAC-DNA at the CpG steps. No evidence was found for the formation of local left-handed structures as associated with Z-DNA. However, the duplex structure does strongly depend on the pH of the solution and the ionic strength. CD and UV absorption spectroscopy experiments do reveal the presence of Z-DNA in GAC repeats (but not in CAG repeats) under conditions of low alkaline pH, and high concentrations of NaCl and other various divalent ions [42].

2.1.2. Conformations and dynamics of DNA CAG loops: smFRET and MD studies

In order to elucidate the structure and dynamics of DNA CAG loops, we recently carried out a combined experimental and computational investigation to directly probe the conformational ensemble and dynamic slipping of a CAG TR hairpin [17]. We used smFRET techniques to directly observe the slipping dynamics in $(CAG)_n$ hairpins by an integer number of CAG units (turns out to be predominantly two units). For the experiments, we designed a two-stranded system involving an anchor strand and a hairpin strand with the donor (Cy3) and acceptor (Atto647N) fluorophores placed at consistent positions for all hairpin structures considered (see Fig. 3a. When the hairpin closes, the donor and acceptor get into close proximity and therefore a high FRET signal is expected; likewise, when the hairpin is open, a low signal is obtained.

We measured smFRET signals from DNA containing 14 and 15 CAG repeat units, designated as $(CAG)_{14}$ and $(CAG)_{15}$ respectively. These two structures are indicative of hairpin systems with an even and odd parity. smFRET results are shown in Fig. 3b. For $(CAG)_{14}$, the time-trace results show that there are transitions between three different FRET efficiencies of 0.01 (barely visible in Fig. 3b). The 0.01 state was rarely visited and, by calibrating against an A31 loop (data not shown) [17], turns out to be associated with a completely open state. The 0.65 state turns out to be the most stable, with the hairpin populating that state the majority of time.

Likewise, Fig. 3b also shows smFRET results for $(CAG)_{15}$ which show transitions between four states with efficiencies 0.01, 0.25, 0.46 and 0.73. Again, the 0.01 state is associated with an open hairpin structure. The 0.73 and 0.46 states have similar populations, indicating that they have similar stabilities. A more detailed analysis indicates [17] that for the 0.73 state, the donor and acceptor are closer compared to 0.65 state of $(CAG)_{14}$, which happens when the hairpin slips by one CAG unit towards the donor on the anchor strand (*i.e.*, a −1 slip). By the same token, the 0.46 state is associated with the hairpin slipping by one CAG unit away from the donor (*i.e.*, a + 1 slip). This back and forth slippage allows for $(CAG)_{15}$ to form a AGCA tetraloop with the stem assembling into CAG/GAC aligned pairings. Note that aligning the CAG/GAC at the end of the $(CAG)_{15}$ hairpin (*i.e.*, a 0 slip) results in a triloop consisting of a single CAG unit. The smFRET and simulation results show that these kinds of loops are considerably less stable than the tetraloops. Proceeding in a similar fashion, we found the 0.31 state of $(CAG)_{14}$ could be associated with a slippage of two CAG units, and the 0.25 state of $(CAG)^{15}$ with a slippage of three CAG units (*i.e.*, slips of +2 and +3, respectively). Similar results were obtained for smFRET experiments on $(CAG)_n$ structures, with different integers n. Thus, there are systematic differences between the behavior of loops with even and odd number of CAG units. Even number repeats (such as $(CAG)_{14}$) accommodate an AGCA tetraloop with either a fully paired stem or a stem slipped by two CAG units. For a hairpin with an odd number of CAG units, a paired-end stem is associated with a CAG triloop, which spontaneously slips back and forth to form a AGCA tetraloop with a hanging CAG trinucleotide in the stem. Thus, there is a difference in stability of slipped CAG states in hairpin systems with an even and odd number of repeats which indicates a balance between the stem and tri-and tetraloops energies.

Both the smFRET populated states and the MD simulations indicate greater stability for 5'-AGCA-3' tetraloops, compared with the alternative 5'-CAG-3' triloops. Fig. 4 illustrates MD results showing a triloop as it transitions to a tetraloop and a stable tetraloop configuration. The slipping kinetics depends on the repeat parity of $(CAG)_n$ (*n* even or odd). As already noted, to accommodate the tetraloop, even (odd)-numbered repeats have an even (odd) number of hanging bases in the hairpin stem. In particular, a paired-end tetraloop (no hanging TR) is very stable in $(CAG)_{n=even}$, but such situation cannot occur in $(CAG)_{n=odd}$, where the hairpin is "frustrated" and slips back and forth between states with one TR hanging at the 5' or 3' end.

The difference in stability between the loops is explained as follows. In the 5'-CAG-3' triloop the three nucleotides are in anti conformation, the C base flips out and the weak sheared C·G pair is held by a single hydrogen bond. The triloop is then "locked" by a weak AG/CA step (where the A bases are mismatched). In contrast, 5'-AGCA-3' tetraloops are stabilized relative to triloops by favorable stacking energy within the loop, less bending deformation of the backbone; and locking by a GC/GC step (see Fig. 4b. Considering the χ torsion angle, the 5'-AGCA-3' tetraloop shows two

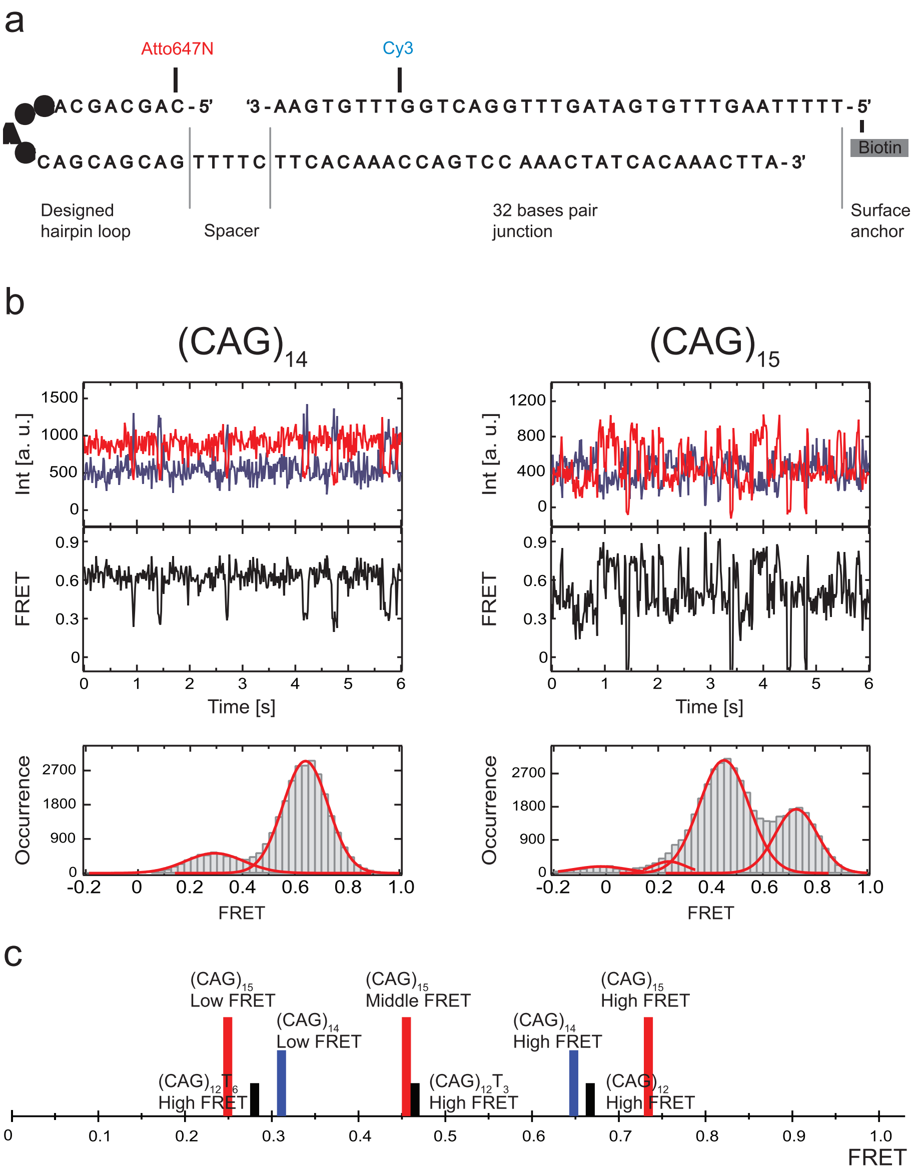

Fig. 3. Schematic DNA design and smFRET analysis result of $(CAG)_{14}$ and $(CAG)_{15}$ at 10 mM. (a) The hairpin loop of interest is immobilized to slide by a partial complementary DNA anchor strand. The spacer helps reduce the interaction between the hairpin and the junction duplex. (b) Representative smFRET time traces of $(CAG)_{14}$ and $(CAG)_{15}$ (top panel). The bottom histograms show all the timepoints of different states from multiple picked traces. Each histogram is fitted into a gaussian function (black line). (c) FRET $(CAG)_{15}$ states (red) and $(CAG)_{14}$ states (blue) [17]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

preferred conformations, where the AGCA nucleotides are either in anti-anti-anti-syn, or anti-anti-anti-anti conformations. These conformations, in turn, display subpopulations of single and double base stacks within the loop [17].

Even-numbered sequences can form a tetraloop while forming a paired-end stem without any overhangs, which should minimize the free energy of the entire hairpin structure. Odd-numbered sequences, by contrast, can only accommodate an AGCA tetraloop if one strand is displaced from the other by at least one TR, thereby forming an even-numbered hairpin with one hanging base. The latter, of course, takes extra energy but not enough to deter the formation of a tetraloop. Ultimately, this leads to very different dynamics for the hairpins: hairpins with an even number of repeats spend most of their time in the paired-end state, with occasional slips by 2 TR units, while odd-numbered hairpins slip back and forth in one direction or another in their bid to form a tetraloop. The simulations suggest that the transition a $(CAG)_{odd}$ from a triloop to a tetraloop is triggered by a disruption of the A·A mismatch closest to the loop, with the A base on the 3' strand switching towards to minor groove allowing for the formation of a temporary GACG tetraloop. This slipping process may be a crucial element for the expansion of CAG TRs. Indeed, it is interesting to note that chemical and enzymatic probing of RNA CAG, CCG, CGG, and CUG repeats form hairpins that slip dynamically with several possible 3' overhangs [46]. One possible scenario would occur when the initial complementary strands in a (CAG)·(CTG) duplex separate, for instance under negative supercoiling, giving rise to opposite CAG and CTG hairpins, as has been suggested previously [17,23]. Strand slipping by trinucleotide units would thus allow these hairpins to travel apart in a soliton-like wave. Finally, if a single-stranded cleavage happens in the CAG strand facing the CTG hairpin, then the CTG hairpin may relax and stretch leaving a gap in the CAG strand. The subsequent filling of this gap by different proteins in the cell machinery would then result in the TR expansion of the CAG strand. This mechanism is illustrated in Fig. 5.

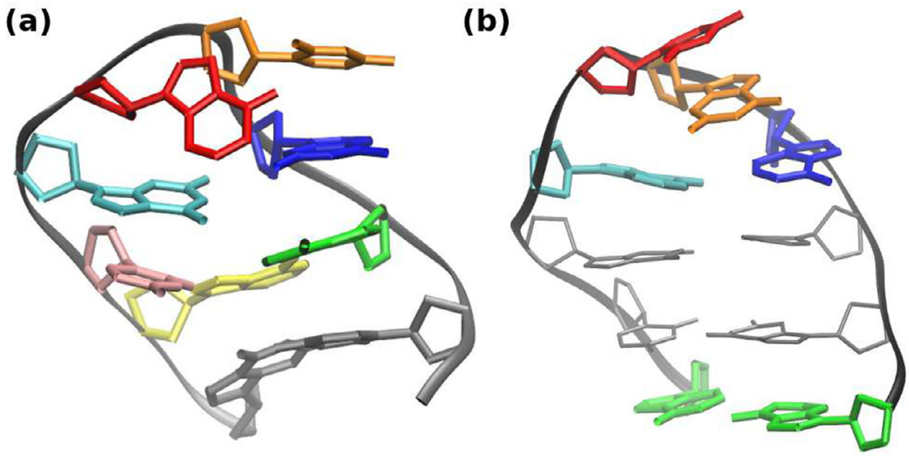

Fig. 4. Snapshots of two CAG hairpin loops: (a) $(CAG)_{15}$ triloop in100 mM excess salt (note that only selected residues are shown) with A20 (green), G21 (blue), C22 (orange), A23 (red), G24 (cyan), C25 (pink) and A26 (yellow). The figure shows the CAG triloop (C22-A23-G24) as it deforms into the tetraloop configuration. (b) One of the stable conformations of the 5'-ACGA-3' tetraloop in a $(CAG)_{14}$ hairpin, with bases shown in color: A20 (blue), G21 (orange), C22 (red), A23 (cyan), A17 and A26 (green). The tetraloop is stabilized by and A20-G21-C22 triple stack with A20, G21 and C22 all in the anti conformations, and A23 in syn conformation. The first mismatch in the stem, A17-A26 is in an anti-anti conformation [17]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

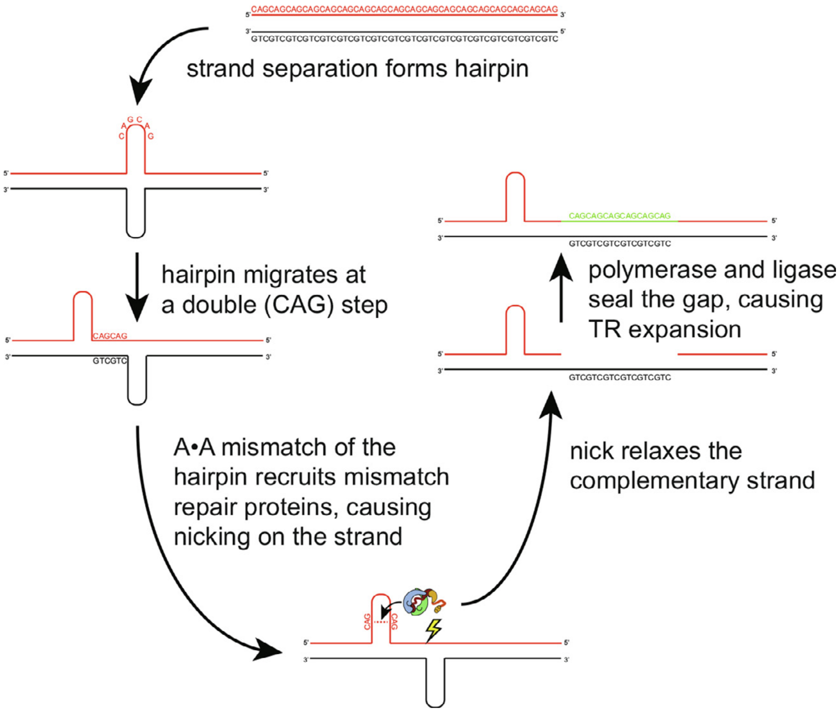
5' CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG 3'
3' GTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTC 5'
strand separation forms hairpin
A G C A
C G
5'
5'
3'
3'
hairpin migrates at a double (CAG) step
CAGCAG
GTCGTC
A•A mismatch of the hairpin recruits mismatch repair proteins, causing nicking on the strand
CAG
CAG
nick relaxes the complementary strand
GTCGTCGTCGTCGTCGTC
polymerase and ligase seal the gap, causing TR expansion
CAGCAGCAGCAGCAGCAG
GTCGTCGTCGTCGTCGTC

Fig. 5. Schematic of proposed model or CAG TR expansion mechanism.

2.2. Helical homoduplexes formed by CCG and CGG TRs

CGG and CCG TRs are overexpressed in the human genome exons. TRs of CGG are encountered in the 5'-untranslated region (5'-UTR) of the fragile X mental retardation gene (FMR1) [47]; TRs of CCG are similarly found in the 5'-UTR and the translated parts of more than one gene. The normal repeat length of CGG TRs is in the 5 to 54 range. The higher range increases the probability to disease expression in descendants [48,49]. A longer repeat number (55 to 200) CGGs is associated with fragile X tremor ataxia syndrome (FXTAS) in males [50] and premature ovarian failure in females [51]. Repeats greater than 200 CGGs cause the inherited fragile X mental retardation syndrome [52]. CCG TRs are associated to three TREDS, with the longest expansion being associated with the FRM2 gene which results in chromosome X-linked mental retardation (FRAXE) [53]. These repeats also appear to play a role in HD, and type 1 myotonic dystrophy [54].

We have investigated the conformation and dynamics of the CGG and CCG TR homoduplexes both for DNA and RNA [31]. As is the case with other TRs, the structural characteristics of the duplexes are largely determined by the characteristics of the C·C and G·G mismatches. Here, it is important to consider the nature of the Watson–Crick pairs that surround the mismatches. Sequences of the form 5'-(CGG)-3' and 5'-(CCG)-3' are characterized by GpC steps between the Watson–Crick base pairs, while sequences of the form 5'-(GGC)-3' and 5'-(GCC)-3' exhibit CpG steps between the Watson–Crick base pairs. Hence, when both RNA and DNA are considered, this results in eight different nonequivalent helical duplexes.

Using the ABMD free energy methods [35], the global minima associated with C·C mismatches in the four C-rich homoduplexes correspond to anti-anti conformations, with *ap* anti in RNA duplexes and *ac* anti in DNA duplexes. As with CAG and GAC TRs, the anti-syn mismatch conformation is about 5 (7.5) kcal/mol higher than the anti-anti conformation for RNA (DNA), and syn-syn conformations have an even higher free energy. By contrast, the G-rich duplexes favor the anti-syn conformation, followed by the anti-anti conformation. This is shown in Fig. 6, which illustrates the (χ_5, χ_{14}) free energy landscapes for single G·G mismatches (here χ represents the dihedral angle associated with the mismatched G's [31]). The exception here is the RNA-CGG structure for which the anti-syn and anti-anti conformations appear to have the same value within the limits of the calculations. This inability of the free energy calculations to resolve the differences between the two minima is most likely due to the strong triple G-base stacking not present in the other GGC structures. Theoretically, the results have been tested using three different AMBER force fields (BSC0 [55], BSC1 [56], and OL15 [57]), which all give similar results. Experimentally, there is crystallographic data for two of the eight duplexes, mainly RNA CGG and CCG sequences (with GpC steps between the Watson–Crick base pairs), and these are in agreement with our free energy calculation results [58–60].

We have also run extensive MD simulations of the RNA/DNA duplexes with the mismatches placed in different conformations, thereby being able to investigate the transitions to the ground state conformations. Generally, intrahelical C·C mismatches transition to their global anti-anti minimum faster than G·G mismatches that get trapped in metastable states. In particular, the G bases are subject to a stacking interaction, which tends to slow down the transition to the anti-syn conformation. Interestingly, C mismatches in DNA-GCC homoduplexes may be extruded to form the so-called “e-motif” [32], discussed more extensively below. The mismatched duplexes were also observed to form characteristic sequence-dependent patterns such that the twist is more regular in intrahelical C-mismatched sequences, while the largest twist variations were observed in the G-mismatched sequences [30].

2.3. Helical homoduplexes formed by DNA CTG (RNA CUG) and GTC (GUC) TRs

Myotonic dystrophy belongs to a group of inherited neuromuscular disorders called muscular dystrophies, that typically begin in adulthood [61,62]. The disease is caused by either CTG TRs (myotonic dystrophy type 1) or CCTG tetranucleotide repeats (myotonic dystrophy type 2). The CTG TRs are located in the 3’-UTR

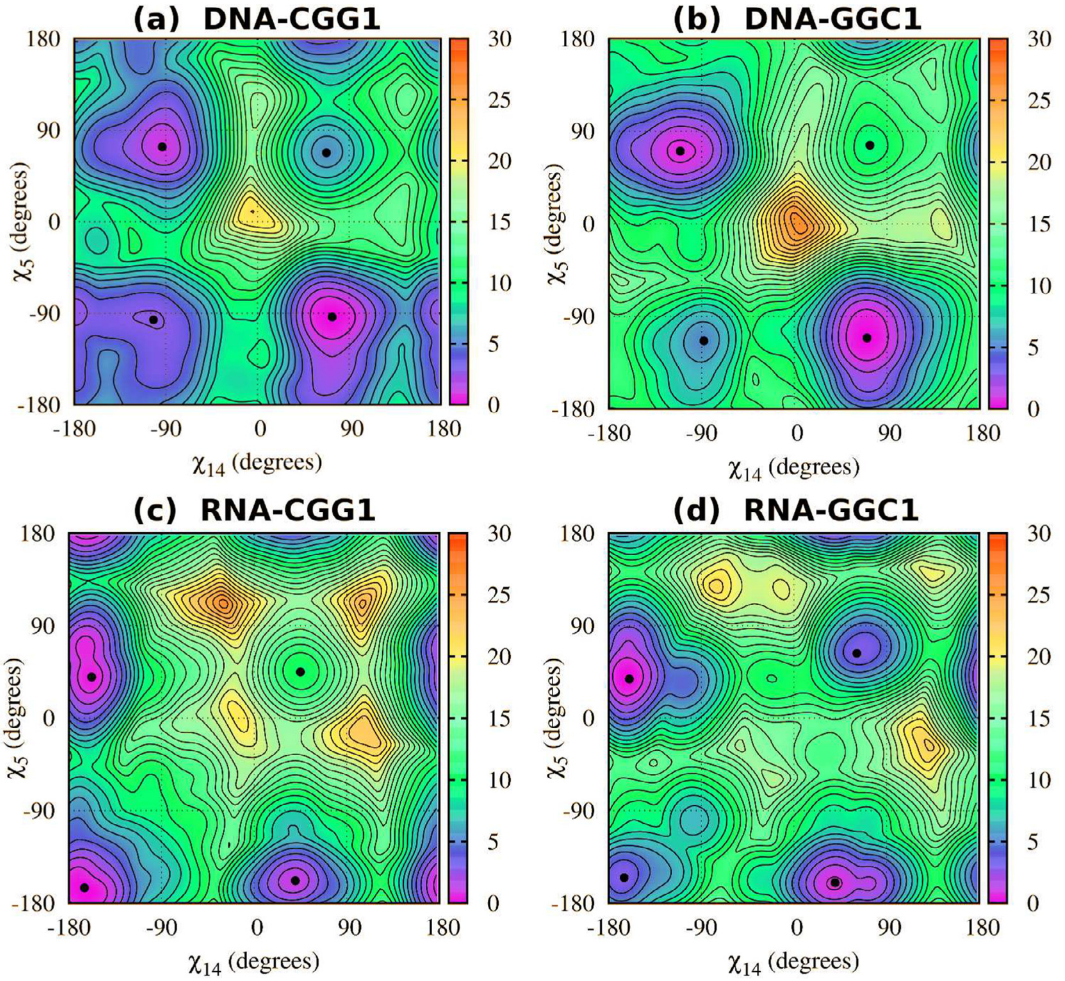

Fig. 6. The (χ_5, χ_{14}) free energy landscapes for a single G·G mismatches (unit kcal/mol) for: (a) DNA-CGG1; (b) DNA-GGC1; (c) RNA-CGG1; (d) RNA-GGC1. Here the unity in the label indicates that the simulated structure contains a single mismatch; the collective variables χ are the dihedral angles associated with the mismatch nucleotides. Roughly speaking, for DNA the anti-syn minima are located at $(-96^\circ, 73^\circ)$ (and its mirror image) for CGG1 and $(-113^\circ, 70^\circ)$ (and mirror image) for GGC1. For RNA, the anti-syn minima are located at $(-160^\circ, 40^\circ)$ (and mirror images) for both CGG1 and GGC1[31].

of the dystrophia myotonic protein kinase gene while CCTG tetranucleotide repeats are found in the zinc finger 9 (ZNF9) gene [63]. The RNA from the transcribed genes contains CUG or CCUG repeats which are known to fold into RNA hairpins. Hence, in this particular case, the disease is associated with toxic mRNA gain-of-function [64]. Experimentally, CUG RNA structures have been investigated with X-ray diffraction [65–67]. The results indicate that RNA CUG TRs form a double-helix homoduplex in A-RNA form, where the conformation of the U-U mismatches is quite dynamic forming what has been termed a "stretched U-U wobble" form [65] with hydrogen bond numbers ranging from 2 to 0. Mismatches with two and one hydrogen bond appear to be most frequent, estimated to be about 40% [68]. As with other TRs, we have investigated the helical homoduplexes formed by DNA CTG (RNA CUG) and GTC (GUC) TRs. In agreement with another computational study and the experimental work [68], we find that RNA (DNA) U·U (T·T) mismatches are primarily located inside the helical core in an anti-anti conformation in RNA and in (high-anti)-(high-anti) conformation in DNA. We also characterized the dynamics of these helical duplexes and their mismatches; we are currently finishing the characterization of the structural differences between the CTG (CUG) repeats (with GpC Watson–Crick basepair steps) and the non-equivalent GTC (GUC) repeats (with CpG Watson–Crick basepair steps). Electrophoresis experiments indicate that RNA CUG repeats form "slippery" hairpins [69], which dynamically slip as noted with the CAG hairpins [17]. As may be expected, hairpins with longer stems tend to be more stable. Since this study was based on biochemical methods, it does not unambiguously determine the actual loop structure, although the data is consistent with both 5'-GCUGC-3'and 5'-UGCU-3' loops, with varying number of stem overhangs.

2.4. Helical homoduplexes formed by GGGGCC and GGCCCC hexanucleotide repeats

Amyotropic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are two neurodegenerative diseases that share similar neurological and genetic pathways. FTD, which is due to the degeneration of the frontal and anterior temporal lobes, is a common cause of early-onset dementia; ALS, on the other hand, is associated with progressive weakening of the muscles and paralysis of the motor neurons in the spinal cord and brain. It turns out that a (GGGGCC) hexanucleotide repeat (HR) expansion in the first intron of the C9ORF72 gene is the major cause behind both ALS and FRD [70,71]. While the normal, unaffected population is characterized by fewer than 20 HRs, patients with FTD and ALS have large expansions greater than 70 repeats and often in the 250–1600 range.

As already noted, nucleotide repeat disorders can cause toxicity through different but not exclusive mechanisms. While the expansions originate in the DNA itself, these expansions can alter the local chromatin structure, and change the RNA transcription and protein translation of the gene. For FTD/ALS, the transcribed introns with these anomalously enlarged expansions give rise to the neuropathology both through a loss of function as mRNA levels in the C9ORF72 gene are decreased, as well as through a gain of function as transcripts with the (GGGGCC) HRs accumulate in the nuclear foci of the frontal cortex and spinal chord resulting in the sequestration of RNA-binding proteins [12,72]. Complicating the disease pathology is evidence that the antisense (GGCCCC) HR expansion transcripts that result from a bidirectional transcription of the DNA HR also form nuclear RNA foci. The translated repeats may cause toxicity in the formed protein and its interaction partners. It is also known that even though the HRs are to be found in a non-coding region of the C9ORF72 gene, these expansions can trigger protein translation even in the absence of the ATG start coding. This leads to the formation of unconventional repeat-associated non-ATG (RAN) translations [12,72,14]. Such RAN translations of the (CCCCGG) expansions may give Gly-Pro, Gly-Ala and Gly-Arg polydipeptide expansions. Likewise, RNA translations of the antisense (GGCCCC) expansion may give Pro-Ala, Pro-Gly and Pro-Arg expansions. These generically coined "C9RAN" dipeptide proteins have been found in the brain of C9FTD/ALS model mice [73]; and both sense and antisense C9RAN proteins in all three reading frames have been found in the central nervous system of C9FTD/ALS patients and culture cells [74,13]. In fact, poly(Gly-Pro) has been proposed as a disease biomarker [13,75].

Based on enzymatic and chemical probing of the r(GGGGCC) expansion, the generally accepted scenerio is that the repeat expansion adopts a hairpin conformation with the G·G mismatches in equilibrium with a quadruplex structure [75,76]. This equilibrium is temperature dependent with $T = 37^\circ$ favoring hairpin structures, and higher temperatures favoring quadruplex formation, The structural transition is also controlled by ion type with the larger K^+ ions favoring G-quadruplexes and the smaller Na^+ favoring hairpins [75].

We now discuss the results of a simulation study of the conformations and dynamics of all possible DNA and RNA homoduplexes that can be formed from the GGGGCC sense and the GGCCCC antisense HRs [33], leaving the analysis of the associated quadruplex structures for the next section [34]. Generating all the possible duplexes via a shifting of the reading frames results in three different homoduplexes for either G-rich or C-rich sequences, both for DNA and RNA, which gives a total of twelve different homoduplexes as shown in Fig. 7. The structures differ in the pattern of "steps" (which also includes the mismatches) and are therefore not the same. Each conformation when repeated have the same number of G·G or C·C mismatches and the same number of Watson–Crick base pairs. What is different though is that "double G" (DG) and "double C" (DC) duplexes have neighboring double mismatches separated by four Watson–Crick base pairs, while the "single G" (SG) and "single C" (SC) homo–helices are characterized by single mismatches separated by two Watson–Crick base pairs. We have carried out large scale MD simulations of each of these structures in order to probe the behavior of the local mismatches, the ion distributions and bindings, and relative stability [24].

G-rich double helices share common features. The inner G-G mismatches stay inside the helix in G_{syn}-G_{anti} conformations and form two hydrogen bonds between the Watson–Crick edge of G_{anti} and the Hoogsteen edge of G_{syn}. Also, G_{syn} in RNA is associated with a base-phosphate hydrogen bond; whilst inner G·G mismathces lead to a local unwinding of the helix. The neutralizing Na^+ ions are typically located in the major groove and help stabilize the double mismatches through the formation of ion bridges that join two G's in a mismatch with bases in neighboring Watson–Crick base pairs, or the four G's composing the double mismatch. G-rich helices are more stable than C-rich ones due to a better stacking and due to the hydrogen bond formation associated with the anti-syn conformation of the G·G mismatches.

While the C-rich double helices are characterized by a variety of conformations, one common feature is that the inner mismatched C bases are all in the anti configuration. The most unstable C-rich RNA and DNA helices consist of single mismatches separated by two Watson–Crick base pairs (SC-3 structure shown in Fig. 7). For DNA, the mismatched Cs tend to flip out of the helical core. This is in contrast to RNA, where the C–C mismatches remain inside the helix inclined to either the major or minor groove. The DNA DC-1 helix accommodates mismatches via the formation of e-motifs (Fig. 8), where mismatched bases flip towards the minor groove and point in the 5' direction of their respective strands. The e-motif was first described in NMR experiments in a solution conformation of a DNA CCG TR [77]. Once formed, the e-motif appears to be particularly stable [33]. While the DNA DC-2 duplex is stable

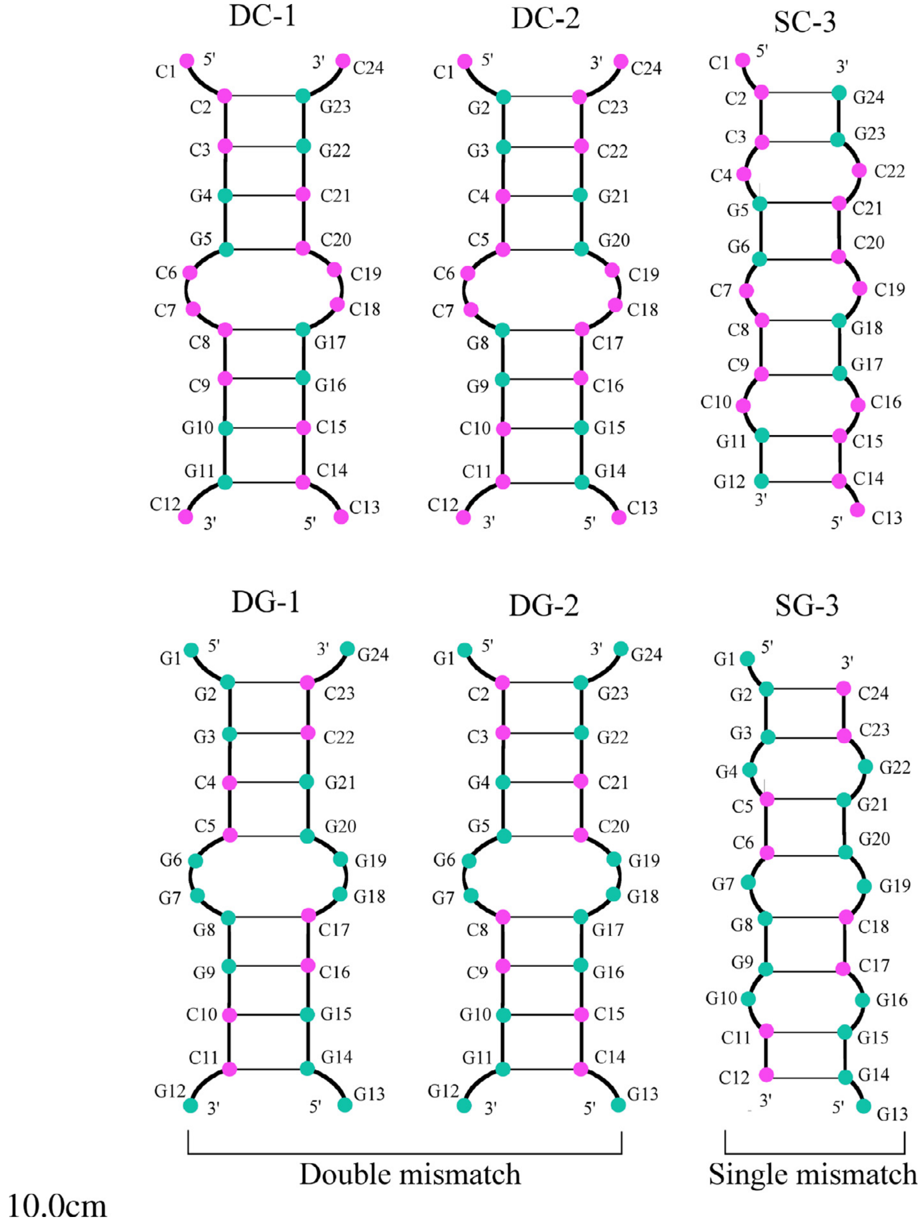

Fig. 7. Schemes for the six GGGGCC and CCCCGG hexanucleotide repeat nonequivalent homoduplexes. The G's are marked in green, and C's in purple. The helical homoduplexes were built and simulated for both RNA and DNA[33].

(with bases in the mismatches alternating between the major and minor grooves), the e-motif structure associated with DNA DC-1 gives a superior stacking arrangement thereby providing additional stabilty and leading to a more compact double helix. Flipped C bases at the i^{th} position form G(N3)-C(N4) hydrogen bonds with the $i-2^{th}$ G base, so that DC-1 is the preferred sequence for the e-motif. By contrast the DNA SC-3 structure is unstable; the structure either unfolds or evolves into the more stable DC-1 structure. For the DNA antisense CCCCGG HR duplexes, the stability of the structures is therefore ranked as DC-1 > DC-2 > SC-3.

With RNA, the DC-1 structure is characterized by two stable conformations. In one conformation, one inner mismatch pair forms a N3-N4 hydrogen bond between the C bases of the mismatch, while the other inner mismatch pair points towards the minor groove. This leaves the major groove relatively unoccupied and results in the bending of the helix towards the major groove. For the second structure, two bases belonging to two consecutive mismatches stack on top of each other, whilst their partners bend towards the major groove which causes the helix to straighten. The RNA DC-2 structure was found to be stable in this latter conformation. The RNA SC-3 structure is the least stable of all the RNA conformations. However, in contrast to its DNA counterpart, the duplex remains stable with mismatches exhibiting a synchronized oscillation in motion between the major and minor grooves. Thus, if one base of the mismatched pair turns towards the minor groove, the other turns towards the major groove and vice versa. This eventually leads to a slow reorganization of the hydrogen bonds between the mismatched bases.

With regards to the neutralizing Na^+ ions, in the C-rich DNA duplexes, these occupy the minor groove close to the C·C mismatches. In the RNA DC-2 structure, the ions favor the major groove – also around the mismatches, In RNA DC-1 and SC-3, the neutralizing ions in the major groove tend to be associated with

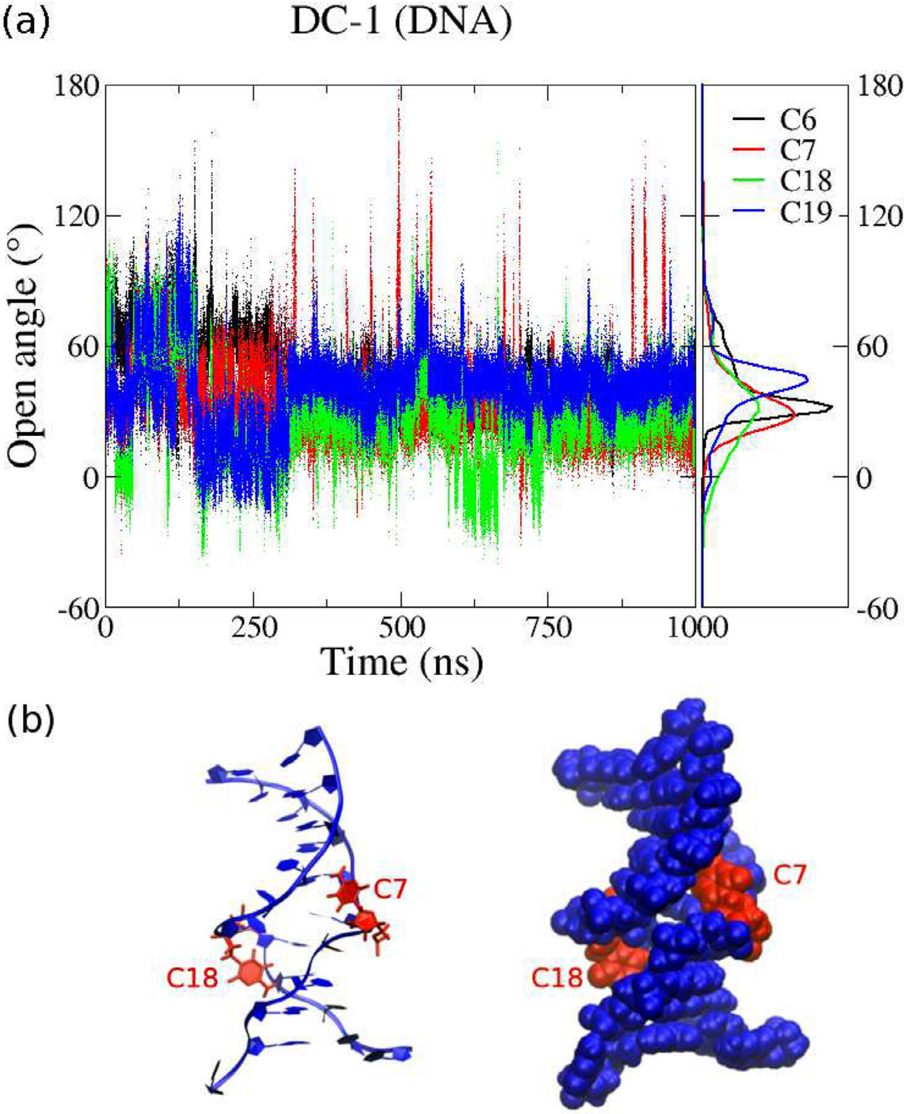

Fig. 8. For C-rich CCCCGG hexanucleotide repeats: (a) open-angle of DNA DC-1 as a function of time (left) along with its distribution (right); (b) the e-motif configuration [33].

the Watson–Crick pairs. Again, there is a non-negligible ion presence in the minor groove near the mismatches. Direct ion binding to the C·C mismatches is a contributing factor to the stability of these duplex structures.

2.5. DNA and RNA quadruplexes formed by GGGGCC and GGGCCT hexanucleotide repeats

As noted in the previous section, a GGGGCC HR expansion in the first intron of the C9ORF72 gene has been shown to be a major cause behind both ALS and FTD. Pathogenic structures associated with these HRs are both hairpins and G-quadruplexes, which we now discuss. There is, however, another closely related HR (GGGCCT), found in intron 1 of the NOP56 gene located on chromosome 20 that has been associated with spinocerebellar ataxia 36 (SCA36) [78], which also leads to quadruplex structures.

Alleles of the NOP56 gene that are normal have 5 to 14 HRs, while faulty genes have repeats in the 800 to 2000 range and lead to a heterogenous group of neurodegenerative disorders – characterized by a loss of balance and limb ataxia – associated with autosomal-dominant spinocerebellar ataxias. Similarly, SCA36 is associated with late-onset motor neuron involvement with symptoms like ALS and a sensorineural loss of hearing. In contrast to ALS and other SSR diseases, the severity of SCA36 does not seem to vary with the repeat length of the HRs [78].

The GGGGCC HR expansion belongs to the class of G-rich sequences that form stable G-quadruplex structures that are associated with a right-handed helicity that comes from the hydrophobic stacking of two or a larger number of G-quartets. A G-quartet (or tetrad) is composed of a planar array of four guanines held together by a cyclic array of hydrogen bonds from the Watson–Crick and Hoogsteen faces. These quartets then stack to form a quadruplex stem of varying length. The stem is usually stabilized by monovalent cations located within the central channel of the stem in direct contact with the carbonyl groups of the guanines. Many times, there are also loops associated with specific G-quadruplexes, which form as the strands fold to form the structure. Human telomeric DNA based on d(TTAGGG) repeats are associated with G-quadruplex formation, where the structure functions to maintain telomere length. Other SSRs such as CGG TRs found in the 5'-UTR of the FRAXA gene FMR1 and the regulatory region of the insulin gene [79,80] are also implicated with G-quadruplex formation. DNA d(GGGGCC) oligomer of varying repeat length have also been shown to form inter- and intra-molecular

G-quadruplexes. NMR and circular dichroism (CD) spectroscopy experiments show that these may form in either a parallel or antiparallel orientation [81].

As already noted, the stability of the G-quadruplexes is linked to the ion size, with large ions favoring quadruplex structures and smaller ions favoring the formation of hairpins. The decreased stability associated with decreasing ion size, *i.e.* $K^+ > Na^+ > Li^+$ appears to be a characteristic of other kinds of G-quadruplexes as well.

Here, we report on the results of an MD simulation study of G-quadruplexes associated with GGGGCC and GGGCCT HRs [34]. Some of the initial structures for the study are shown in Fig. 9. Motivated by NMR and circular dichroism spectroscopy, the resulting G-quadruplex structures may be classified as either parallel quadruplex (PQ) or antiparallel quadruplex (AQ). The PQ model has all four strands parallel to each other and their guanines in an anti conformation, so that all parallel structures are very similar. Antiparallel strands, on the other hand, have considerably more variation in their structure leading to significant polymorphism. Given the large number of possible structures, we have opted to focus on two models, AQ-1 and AQ-2. In the AQ-2 model in Fig. 9, each strand has two adjacent, opposite-direction nearest neighbor strands such that the diagonal strands are parallel (it is a fully antiparallel model). The hybrid antiparallel model AQ-1 has one neighbor strand running in the same direction and one

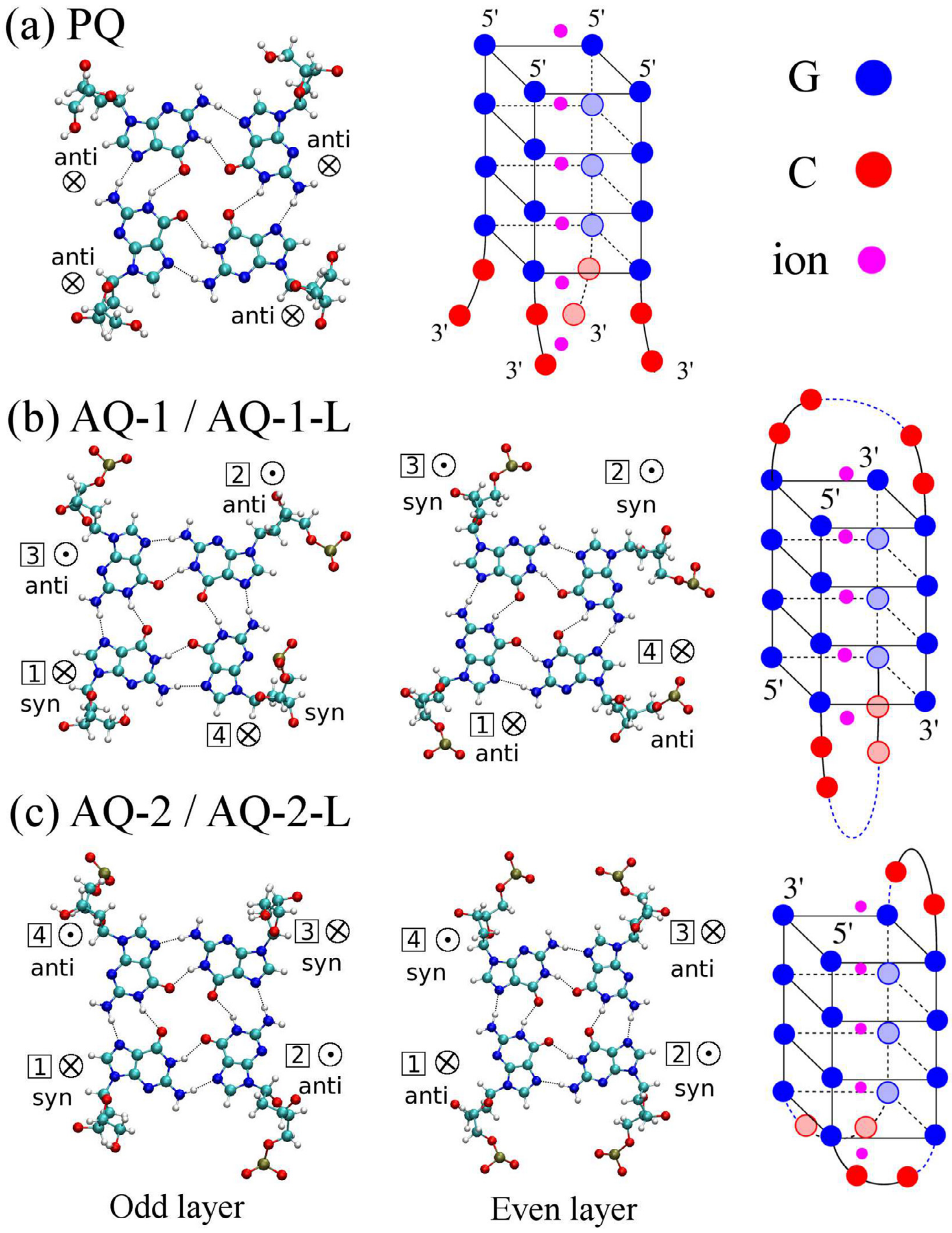

Fig. 9. Quadruplex geometries. On the left, the geometry of the guanine residues within a quartet (where relevant, odd layers in first column and even layers in second column). On the right, the full quadruplex configurations. (a) PQ; (b) AG-1 and AG-q-L; (c) AQ-2 and AQ-2-L. Here, G's are blue, C's in red and ions in magenta [34].

running in the opposite direction. The AQ-1-L and AQ-2-L models have loops, as shown in Fig. 8. Within the AQ-1 model, adjacent G-quartets are characterized by a syn-syn-anti-anti and anti-anti-syn-syn adjacent quartet topology, while AQ-2 structures are characterized by anti-syn-anti-syn and syn-anti-syn-anti adjacent quartet topology. In all types of G-quadruplexes, strands that have the same (opposite) direction have the same (opposite) glycosidic conformation. Adjacent guanines along the same strand have alternating syn and anti glycosidic orientations.

All in all, we tested a total of 22 different G-quadruplex models over a 1 μsec time scale for each model, some with different ions [34]. The results indicate that all DNA – either parallel or antiparallel, either with or without loops – are stable. For RNA, on the other hand, only the PQ and AQ-1-L structures were found to be stable. The latter was found to be stabilized by the presence of the two diagonal loops. The stability and unfolding of the unstable RNA structures was investigated by tracking the quadruplex twist and quadruplex buckle displacements along with the backbone and glycosidic torsion angles. Generally speaking, twist values remain constant for DNA and RNA parallel models, but quickly decay for RNA antiparallel models signaling the structural unwinding.

It is interesting to note that for both DNA and RNA, the parallel G-quadruplex stabilize the adjacent C bases into a C-quartet thereby effectively forming a mixed quadruplex of at least 5 layers. This C-quartet is stabilized by the stacking interactions with the guanine bases in the preceding G-quartet and by C(N4)-C(O2) hydrogen bonds and an ion between the G-tetrad and the C-tetrad. In absence of stacking with a G-quartet, the second, free-floating layer formed by C bases is not stable. However, it could probably become stable for longer sequences where two C-quartets could become "sandwiched" between four-layered G-tetrads. We also studied the stability of quadruplexes as a function of K^+ and Na^+ ions [25]. Both types of ions favor a stable twist distribution with Na^+ giving smaller twist angles. However, the buckle distributions of Na^+ is larger as compared to K^+, indicating that the G-quadruplexes are more stable in the presence of the latter ions in agreement with experiments.

2.6. E-motif formed by extrahelical cytosine bases in C-rich DNA homoduplexes

To understand the mechanisms underlying sequence expansion, gene hypermethylation, and folate-induced chromosomal fragile sites, it is crucial to elucidate the secondary structure adopted by the C-rich sequences d(CCG)$_n$ of various repeat length n. Sequences of this kind attracted considerable interest twenty years ago, when it was observed that homoduplexes d(CCG)·d(CCG) exhibited an unusual DNA secondary structure termed "e-motif" [77], already mentioned and shown in Fig. 8. This structural motif was observed in a solution NMR DNA antiparallel duplex, with each strand consisted of two 5'-(CCG)$_2$-3' repeats (PDB ID1NOQ). In this duplex, a slipping of the strands resulted in two unpaired 5'-C terminals, and a central C·C mismatch surrounded by Watson–Crick pairs in the center. This mismatch led to the formation of an "e-motif" in which the mismatched C bases symmetrically flip out in the minor groove, pointing their base moieties to the 5' direction on each of the strands. There is evidence that CCG repeats form more complicated structures. For instance, in the PDB ID4PZQ structure [82], two dT(CCG)$_3$A strands associate to form a tetraplex structure with an i-motif [83,84] core containing four C:C+ pairs flanked by two G:G homopurine base pairs as a structural motif. The tetraplex core is attached to a short parallel-stranded duplex. Each hairpin itself contains a central CCG loop in which the nucleotides are flipped out and stabilized by stacking interactions. This superficially resembles an e-motif, but it is not, as these extruded bases come from the same strand due to the formation of the loop as opposed to different strands in a helical duplex as is characteristic of e-motifs.

In the course of our investigations of the DNA d(CCCCGG) HRs, we noted the formation of a stable e-motif in the DC-1 structures [33]. Motivated by this observation, we extended our studies to other C-rich sequences in order to determine which sequences give rise to the e-motif, both as an isolated extrusion of a mismatch or as an extended e-motif formed by consecutive extrahelical C·C mismatches, and to characterize their conformations and dynamics [32]. Specifically, we examined the TRs with two non-equivalent reading frames (GCC)$_n$ and (CCG)$_n$, and the three non-equivalent reading frames associated with HRs: (CCCGGC)$_n$, (CGGCCC)$_n$ and (CCCCGG)$_n$. The salient results are as follows [32].

In the e-motif, the C bases of the i^{th} residue in a mismatch are symmetrically flipped out of the minor groove pointing towards the i-2 residue, *i.e.*, in the 5' direction on each strand. A single e-motif is partially stabilized by the formation of hydrogen bonds between the extruded C base of the mismatch and i-2 base along the same strand, which is a C base in the case of (GCC) or a G base in the case of (CCCGGC) in DC-1. Creation of the e-motif is favored by the formation of pseudo GpC steps between non-adjacent base pairs when the bases in the C·C mismatches are extruded. Consequently, the e-motif is stable in paired-end homoduplexes of (GCC) and (CCCGGC) SSRs, but not in the other reading frames [22]. As a consequence of the coupling between the particular step arrangement and the rotation paths followed by the extruded base, the extruded mismatched C bases in an e-motif are always found in the minor groove. Finally, the extended e-motif (in which all the C bases in the mismatches are extruded) is stabilized by highly cooperative interactions. In addition to the favorable stacking afforded by pseudo GpC steps, and the hydrogen bonds between the mismatched bases and other nucleotides, the extended e-motif is further stabilized by the stacking of the extruded C bases themselves. The net result is a very stable anomalous secondary structure.

We have also probed RNA (GCC) duplexes for e-motif formation [31] and, although the C bases occasionally flip into either groove, an e-motif is never formed. We believe that this is due to the additional hydroxyl group on RNA which can form hydrogen bonds with neighboring sugar and backbone atoms thereby hindering the extrusion of the C bases. In addition, the A-form of RNA precludes good stacking for either pseudo GpC and CpC steps.

3. Summary

In the search for a molecular understanding of the anticipation diseases generated by the expansion of SSRs, an important breakthrough has been the recognition that the critical step in all models of repeat instability is the transient formation of pathogenic non-B DNA stable secondary structures in the expandable repeats [16]. The transcripts of the expanded tracks also present pathogenic-related secondary structures which result in functional changes ultimately leading to cell toxicity and death. Experimental determination of these pathogenic structures and their dynamics at the atomic level is scarce, especially for DNA. Here, we reviewed some of our results for the conformations and dynamics of pathogenic structures for both RNA and DNA, which include the four non-equivalent homoduplexes for CAG and GAC TRs; the eight different homoduplexes generated by GGC and GCC TRs; CTG (CUG) and GTC (GUC) homoduplexes; dynamics of CAG DNA hairpins of varying lengths (in conjunction with smFRET experiments); the twelve different homoduplexes that can be formed from GGGGCC and GGCCCC HRs; and a number of parallel and antiparallel quadruplexes formed by GGGGCC and GGGCCT HRs. Knowledge of these various noncanonical nucleic acid structural motifs at the atomic level may ultimately prove to be very important for the development of mechanistic models of SSR neurodegenerative diseases.

Author contributions

As a review of our joint work, authors C. Roland, C. Sagui and K. Weninger primarily wrote this article. Authors F. Pan, V. Man, P. Xu and Y. Zhang performed the simulations, experiments and analysis discussed in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported by the National Institute of Health [NIH-R01GM118508]; the National Science Foundation (NSF) [SI2-SEE-1534941]; and the Extreme Science and Engineering Discovery Environment (XSEDE) [TG-MCB160064]. We thank the NC State HPC Center for computational support.

References

[1] Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 2004;5:435–45.
[2] Subramanian S, Madgula V, George R, Mishra R, Pandit M, Kumar C, Singh L. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics 2003;19:1455.
[3] Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 2000;10:967–81.
[4] Caburet S, Cocquet J, Vaiman D, Veitia R. Coding repeats and evolutionary agility. BioEssays 2005;27(6):581–7.
[5] Mirkin SM. DNA structures, repeat expansions and human hereditary disorders. Curr Opin Struct Biol 2006;16:351–8.
[6] Wells RD, Warren S. Genetic instabilities and neurological diseases. San Diego, CA: Academic Press; 1998.
[7] Orr H, Zoghbi H. Trinucleotide repeat disorders. Annu Rev Neurosci 2007;30:575.
[8] Mirkin S. Expandable DNA repeats and human disease. Nature 2007;447:932.
[9] Pearson C, Edamura K, Cleary J. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 2005;6:729–42.
[10] Ranum LPW, Cooper TA. RNA-mediated neuromuscular disorders. Ann Rev Neurosci 2006;6:259–77.
[11] Li L-B, Bonini NM. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. Trends Neurosci 2010;33:292–8.
[12] Gendron TF et al. Antisense transcripts of the expanded C9ORF72 hexanucleotide repeat form nuclear RNA foci and undergo repeat-associated non-ATG translation in c9FTD/ALS. Acta Neuropathol 2013;126:829–44.
[13] Mori K, Arzberger T, Grasser FA, Gijselinck I, May S, Rentzsch K, Weng S-M, Schludi MH, van der Zee J, Cruts M, Van Broeckhoven C, Kremmer E, Kretzschmar HA, Haass C, Edbauer D. Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. Acta Neuropathol 2013;126:881–93.
[14] Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, Margolis J, Peterson M, Markowski TW, Ingram MA, et al. Non-ATG-initiated translation directed by microsatellite expansions. Proc Natl Acad Sci USA 2011;108:260–5.
[15] Galka-Marciniak P, Urbanek MO, Krzyzosiak WJ. Triplet repeats in transcripts: structural insights into RNA toxicity. Biol Chem 2012;393:1299–315.
[16] McMurray C. DNA secondary structure: a common and causative factor for expansion in human disease. Proc Natl Acad Sci USA 1999;96:1823–5.
[17] Xu P, Pan F, Roland C, Sagui C, Weninger K. Dynamics of strand slippage in DNA hairpins formed by CAG repeats: role of sequence parity and trinucleotide interrupts. Nucl Acids Res 2020;48:2232.
[18] Schmidt MH, Pearson CE. Disease-associated repeat instability and mismatch repair. DNA Repair 2016;38:117–26.
[19] Mitas M, Yu A, Dill J, Haworth I. The trinucleotide repeat sequence D(CGG) (15) forms a heat-stable hairpin containing G(syn).G(anti) base-pairs. Biochemistry 1995;34:12803–11.
[20] Gacy A, Goellner G, Juranic N, Macura S, McMurray C. Trinucleotide repeats that expand in human-disease form hairpin structures in-vitro. Cell 1995;81:533–40.
[21] Darlow JM, Leach DR. The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in Escherichia coli suggest hairpin folding preferences in vivo. Genetics 1995;141:825–32.
[22] Petruska J, Hartenstine MJ, Goodman MF. Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. J Biol Chem 1998;273:5204–10.
[23] Hartenstine MJ, Goodman MF, Petruska J. Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. J Biol Chem 2000;275:18382–90.
[24] Figueroa AA, Cattie D, Delaney S. Structure of even/odd trinucleotide repeat sequences modulates persistence of non-B conformations and conversion to duplex. Biochemistry 2011;50:4441–50.
[25] Huang J, Delaney S. Unique length-dependent biophysical properties of repetitive DNA. J Phys Chem B 2016;120:4195–203.
[26] Cleary JD, Nichol K, Wang YH, Pearson CE. Evidence of cis-acting factors in replication-mediated trinucleotide repeat instability in primate cells. Nat Genet 2002;1:37–46.
[27] Napierala M, Bacolla A, Wells RD. Increased negative superhelical density in vivo enhances the genetic instability of triplet repeat sequences. J Biol Chem 2005;280:37366–76.
[28] Hou C, Chan NL, Gu L, Li GM. Incision-dependent and error-free repair of (CAG)(n)/(CTG)(n) hairpins in human cell extracts. Nat Struct Mol Biol 2009;16:869–75.
[29] Case D et al. AMBER 20. San Francisco: University of California; 2020.
[30] Pan F, Man VH, Roland C, Sagui C. Structure and dynamics of DNA and RNA double helices of CAG and GAC trinucleotide repeats. Biophys J 2017;113:19–36.
[31] Pan F, Zhang Y, Man VH, Roland C, Sagui C. E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats. Nucleic Acids Res 2018;46:942–55.
[32] Pan F, Man V, Roland C, Sagui C. Structure and dynamics of DNA and RNA double helices obtained from the CCG and GGC trinucleotide repeats. J Phys Chem B 2018;122:4491.
[33] Zhang Y, Roland C, Sagui C. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. ACS Chem Neurosci 2017;8:578–91.
[34] Zhang Y, Roland C, Sagui C. Structural and dynamical characterization of DNA and RNA quadruplexes obtained from the GGGGCC and GGGCCT hexanucleotide repeats associated with C9FTD/ALS and SCA36 diseases. ACS Chem Neuro 2018;9:1104.
[35] Babin V, Roland C, Sagui C. Adaptively biased molecular dynamics for free energy calculations. J Chem Phys 2008;128:134101.
[36] Izrailev S, Stepaniants S, Isralewitz B, Kosztin D, Lu H, Molnar F, Wriggers W, Schulten K. Steered molecular dynamics. Computational molecular dynamics: challenges, methods, ideas. Springer-Verlag: Berlin, Germany; 1998. p. 39–65..
[37] Zoghbi HY, Orr HT. Glutamine repeats and neurodegeneration. Annu Rev Neurosci 2000;23:217–47.
[38] Davies SW, Turmaine M, Cozens BA, DiFiglia M, Sharp AH, Ross CA, Scherzinger E, Wanker EE, Mangiarini L, Bates GP. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. Cell 1997;90:537–48.
[39] Sikorski P, Atkins E. New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils. Biomacromolecules 2005;6:425–32.
[40] Man VH, Roland C, Sagui C. Structural determinants of polyglutamine protofibrils and crystallites. ACS Chem Neurosci 2015;6:632–45.
[41] Délot E, King LM, Briggs MD, Wilcox WR, Cohn DH. Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. Hum Mol Genet 1999;8:123–8.
[42] Vorlickova M, Kejnovska I, Tumova M, Kypr J. Conformational properties of DNA fragments containing GAC trinucleotide repeats associated with skeletal displasias. Eur Biophys J 2001:30;197–85.
[43] Kozlowski P, de Mezer M, Krzyzosiak WJ. Trinucleotide repeats in human genome and exome. Nucleic Acids Res 2010;38:4027–39.
[44] Schaefer MH, Wanker EE, Andrade-Navarro MA. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res 2012;40:4273–87.
[45] Yildirim I, Park H, Disney MD, Schatz GCA. Dynamic structural model of expanded RNA CAG repeats: a refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. JACS 2013;135:3528–38.
[46] Sobczak K, de Mezer M, Michlewski G, Krol J, Krzyzosiak W. RNA structure of trinucleotide repeats associated with human neurological diseases. Nucleic Acids Res 2003;31:5469–82.
[47] Fu Y-H, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkert AJ, Holden JJ, Jr RGF, Warren ST, Oostra BA, Nelson DL, Caskey C. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 1991:67;1047–1058.
[48] Zhong N, Ju W, Pietrofesa J, Wang D, Dobkin C, Brown WT. Fragile X gray zone alleles: AGG patterns, expansion risks, and associated haplotypes. Am J Med Genet 1996;64:261–5.
[49] Dombrowski C, Lévesque S, Morel ML, Rouillard P, Morgan K, Rousseau F. Premutation and intermediate-size FMR1 alleles in 10572 males from the general population: loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. Hum Mol Genet 2002;11:371–8.
[50] Hagerman R, Leehey M, Heinrichs W, Tassone F, Wilson R, Hills J, Grigsby J, Gage B, Hagerman P. Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. Neurology 2001;57:127–30.

[51] Sherman SL. Premature ovarian failure among fragile X premutation carriers: parent-of-origin effect?. Am J Hum Genet 2000;67:11–3.

[52] Glass I. X linked mental retardation. J Med Genet 1991;28:361–71.

[53] Gu Y, Shen Y, Gibbs RA, Nelson DL. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. Nat Genet 1996;13:109–13.

[54] Braida C, Stefanatos RK, Adam B, Mahajan N, Smeets HJ, Niel F, Goizet C, Arveiler B, Koenig M, Lagier-Tourenne C, Mandel J-L, Faber CG, de Die-Smulders CE, Spaans F, Monckton DG. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. Hum Mol Genet 2010;19:1399–412.

[55] Ivani I, Dans PD, Noy A, Pérez A, Faustino I, Hopsital A, Walther J, Andrio P, Goñi R, Balaceanu A, et al. Parmbsc1: a refined force field for DNA simulations. Nat Methods 2016;13:55–8.

[56] Pérez A, March'an I, Svozil D, Sponer J, Cheatham III TE, Laughton CA, Orozco M. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. Biophys J 2007;92:3817–29.

[57] Zgarbová M, Otyepka M, Šponer J, Mládek A, Banáš P, Cheatham III TE, Jurečka P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. J Chem Theory Comput 2011:7;2886–2902.

[58] Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. Nucleic Acids Res 2011;39:7308–15.

[59] Kumar A, Fang P, Park H, Guo M, Nettles KW, Disney MD. A crystal structure of a model of the repeating r(CGG) transcript found in fragile X syndrome. Chembiochemistry 2011:12;2140–2142.

[60] Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Crystallographic characterization of CCG repeats. Nucleic Acids Res 2012;40:8155–62.

[61] Brook J et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 1992;68:799.

[62] 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, Wieringa B, Korneluk R. Myotonic dystrophy mutation: an unstable CTG repeat in the untranslated region of the gene. Science 1992;255:1253.

[63] Davis B, McCurrach M, Taneja K, Singer R, Housman D. Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retension of transcripts. Proc Natl Acad Sci USA 1997;94:7388.

[64] Philips A, Timchenko L, Cooper T. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. Science 1998;280:737–41.

[65] Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Structural insights into CUG repeats containing the 'stretched U-U wobble': implications for myotonic dystrophy. Nucleic Acids Res 2009;37:4149–56.

[66] Coonrod LA, Kohman JR, Berlund JA. Utilizing the GAAA tetra loop/receptor to facilitate crystal packing and the determination of the structure of a CUG RNA helix. Biochemistry 2012;42:8330.

[67] Tamjar J, Katorcha E, Popov A, Malinina L. Structural dynamics of double-helical RNAs composed of CUG/CUG and CUG/CGG repeats. J Biomol Struct Dyn 2012;30:505.

[68] Yildirim I, Chakraborty D, Disney M, Wales D, Schatz G. Computational investigation of RNA CUG repeats responsible for myotonic dystrophy 1. J Chem Theor Comput 2015;11:4943.

[69] Napierala M, Krzyzosiak WJ. CUG repeats present in myotonin kinase RNA form metastable slippery hairpins. J Biol Chem 1997;272:31079.

[70] DeJesus-Hernandez M, Machkenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 2011;72:245–56.

[71] Renton AE et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 2011;72:257–68.

[72] Mori K, Weng S-M, Arzberger T, May S, Rentzsch K, Kremmer E, Schmid B, Kretzschmar HA, Cruts M, Van Broeckhoven C, Haass C, Edbauer D. The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science 2013;339:1335–8.

[73] Chew J et al. C9ORF72 repeat expansions in mice cause TDP-43 pathology, neuronal loss, and behavioral deficits. Science 2015;348:1151–4.

[74] Ash PEA, Bieniek KF, Gendron TF, Caulfield T, Lin W-L, DeJesus-Hernandez M, van Blitterswijk MM, Jansen-West K, Paul III JW, Rademakers R, Boylan KB, Dickson DW, Petrucelli L. Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ ALS. Neuron 2013;77:639–46.

[75] Su Z et al. Discovery of a biomarker and lead small molecules to target r(GGGGCC)-associated defects in c9FTD/ALS. Neuron 2014;83:1043–50.

[76] Haeusler AR, Donnelly CJ, Periz G, Simko EA, Shaw PG, Kim M-S, Maragakis NJ, Troncoso JC, Pandey A, Sattler R, Rothstein JD, Wang J. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature 2014;507:195–200.

[77] Gao X, Huang X, Smith G, Zheng M, Liu H. New antiparallel duplex motif of DNA CCG repeats that is stabilized by extrahelical basis symmetrically located in the minor-groove. J Am Chem Soc 1995;117:8883–4.

[78] Kabayashi H, Abe K, Matsuura T, Ikeda Y, Hitomi T, Akechi Y, Habu T, Liu W, Okuda H, Koizumi A. Expansion of intronic GGCCTG hexnucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvment. Am J Human Genet 2011;89:121.

[79] Maizels N. G4-associated human diseases. EMBO Rep 2015;16:910.

[80] Wu Y, Bosch RM. G-quadruplex nucleic acids and human disease. FEBS J 2010;277:3470.

[81] Sket P, Pohleven J, Kovanda A, Stalekar M, Zupunski V, Zalar M, Plavec J, Rogelj B. Characterization of DNA G-quadruplex species forming from C9ORF72 G_4C_2-expanded repeats associated with amyotrophic lateral sclerosis and frontotemporal lovar degeneration. Neurobiol Aging 2015;36:1091–6.

[82] Chen YW, Jhan CR, Neidle S, Hou MH. Structural basis for the identification of an i-motif tetraplex core with a parallel-duplex junction as a structural motif in CCG triplet repeats. Angew Chem Int Ed Engl 2014;53:10682.

[83] Assi H, Garavis M, Gonzalez C, Damha M. i-Motif DNA: structural features and significance to cell biology. Nucleic Acids Res 2018;46:8038.

[84] Zeraati M, Langley D, Schofield P, Moye A, Rouet R, Hughes W, Bryan T, Dinger M, Christ D. I-motif DNA structures are formed in the nuclei of human cells. Nat Chem 2018;10:631.