Published in final edited form as:

ACS Chem Neurosci. 2017 March 15; 8(3): 578–591. doi:10.1021/acchemneuro.6b00348.

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

Yuan Zhang^{†,‡}, Christopher Roland^{†,‡}, and Celeste Sagui^{†,‡,*}

[†]Department of Physics, North Carolina State University, Raleigh, North Carolina 27695-8202, United States

[‡]Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, North Carolina 27695-8202, United States

Abstract

A (GGGGCC) hexanucleotide repeat (HR) expansion in the C9ORF72 gene, and its associated antisense (CCCCGG) expansion, are considered the major cause behind frontotemporal dementia and amyotrophic lateral sclerosis. We have performed molecular dynamics simulations to characterize the conformation and dynamics of the 12 duplexes that result from the three different reading frames in sense and antisense HRs for both DNA and RNA. These duplexes display atypical structures relevant not only for a molecular level understanding of these diseases but also for enlarging the repertoire of nucleic-acid structural motifs. G-rich helices share common features. The inner G-G mismatches stay inside the helix in G_{syn} - G_{anti} conformations and form two hydrogen bonds (HBs) between the Watson–Crick edge of G_{anti} and the Hoogsteen edge of G_{syn} . In addition, G_{syn} in RNA forms a base-phosphate HB. Inner G-G mismatches cause local unwinding of the helix. G-rich double helices are more stable than C-rich helices due to better stacking and HBs of G-G mismatches. C-rich helix conformations vary wildly. C mismatches flip out of the helix in DNA but not in RNA. Least (most) stable C-rich RNA and DNA helices have single (double) mismatches separated by two (four) Watson–Crick basepairs. The most stable DNA structure displays an “e-motif” where mismatched bases flip toward the minor groove and point in the 5′ direction. There are two RNA conformations, where the orientation and HB pattern of the mismatches is coupled to bending of the helix.

*Corresponding Author: sagui@ncsu.edu. Phone: 919-515-3111.

ORCID

Celeste Sagui: 0000-0002-2750-5186

Author Contributions

Y.Z. ran simulations and analyzed data; all authors helped formulate problems and wrote paper.

Notes

The authors declare no competing financial interest.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acchemneur-o.6b00348. Further details on simulations and analysis, open angle definition and additional data, additional PCA data, ion occupation in duplex models, hydrogen bond population for DNA and RNA for GG and CC mismatches, structural predictions of hairpins based on m-fold analysis, and radial distribution functions associated with ions (PDF)

Graphical Abstract

Keywords

Nucleotide repeat disorder; hexanucleotide repeat; C-C and G-G mismatches; C9FTD; ALS; e-motif

INTRODUCTION

Frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) are two neurodegenerative diseases with similar genetic and neurological pathways. FTD is the most common cause of early onset dementia due to degeneration of the frontal and anterior temporal lobes, while ALS is characterized by progressive muscle weakness and paralysis due to loss of motor neurons in the brain and the spinal cord. These diseases can occur simultaneously, and are believed to be part of the same spectrum.¹ A (GGGGCC) hexanucleotide repeat (HR) expansion in the first intron of the C9ORF72 gene has been shown to be the major cause behind both FTD (here specifically indicated as C9FTD) and ALS.^{2,3} While the unaffected population carries fewer than 20 repeats (generally no more than a couple), large expansions greater than 70 repeats and usually encompassing 250–1600 repeats have been found in C9FTD and ALS patients.

Nucleotide repeat disorders can cause toxicity through different but nonexclusive mechanisms. First, the expansions originate in the DNA itself, and these expansions can alter the local chromatin structure, affecting RNA transcription and protein translation in the gene. Second, transcribed RNA can cause gain and/or loss of function. The transcribed introns containing these large expansions seem to contribute to neuropathology both through *loss of function*, as mRNA levels of C9ORF72 are decreased in C9FTD/ALS patients;^{2,3} and through *gain of function*, as RNA transcripts containing the (GGGGCC) HRs are accumulated in nuclear foci in the frontal cortex and spinal cord, leading to the sequestration of RNA-binding proteins.² Contributing to the complexity of the pathological mechanisms, there is also evidence that antisense transcripts of the expansion, i.e., (CCCCGG) expanded repeats resulting from the bidirectional transcription of the DNA HR expansions, also form nuclear RNA foci.^{4,5} Third, translated repeats can also cause toxicity in the corresponding protein and its interaction partners. Even though the hexanucleotide expansions reside in a noncoding region of the C9ORF72 gene, it has been shown that these expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional *repeat-associated non-ATG* (RAN) translation.^{5–8} RAN translation of the (GGGGCC) expansion can lead to Gly-Ala, Gly-Pro, and Gly-Arg poly dipeptide expansions, while RAN

translation of the antisense (CCCCGG) expansion can lead to Pro-Gly, Pro-Ala, and Pro-Arg polypeptide expansions. Generically, these are known as “C9RAN” proteins, and have been detected in C9FTD/ALS patients.

For simple-sequence nucleotide repeat diseases, an important breakthrough has been the recognition that stable atypical DNA secondary structure in the expanded repeats is “a common and causative factor for expansion in human disease”.⁹ Chemical and enzymatic probing of a r(GGGGCC) repeat expansion points to a general scenario where the repeat expansion adopts a hairpin structure with G-G mismatches in equilibrium with a quadruplex structure.¹⁰ Recently, circular dichroism (CD), optical melting and 1D ¹H NMR spectroscopy, combined with chemical and enzymatic analysis reveal that r(GGGGCC) expansions fold into a G-quadruplex in equilibrium with hairpin structures.¹¹ The equilibrium is temperature dependent, with $T = 37^\circ$ favoring hairpins and higher annealing temperatures favoring quadruplexes. The equilibrium is also controlled by the type of ion (with K^+ ions favoring G-quadruplexes and Na^+ ions favoring hairpins), and ionic strength.¹¹

In this paper, we present an extensive and detailed analysis in silico of the structure and dynamics of all possible DNA and RNA duplexes that can be formed from the GGGGCC sense and CCCCCG antisense HRs. First, we carefully generate the possible duplexes that can be obtained by shifting the reading frame in the HRs. This results therefore in three duplexes for each G-rich or C-rich DNA or RNA helical duplex, for a total of 12 duplexes. Then we carry out 1 μ s classical molecular dynamics simulations to investigate the global conformational space of dodecamers and associated dynamics; characterize the local conformation of the mismatches and the neutralizing ion distribution and binding; and, when possible, compare the relative stability of the double helices.

RESULTS

Initial Structures

We start with ideal B-DNA and ideal A-RNA. In order to achieve a more comprehensive structural characterization at the atomic level, we specifically consider all the possible linear duplexes built from sense GGGGCC sequences with G-G mismatches and antisense CCCCCG sequences with C-C mismatches in all possible reading frames. A scheme for two repeats is shown in Figure 1. In order to identify the independent structures, it is necessary to consider the pattern of steps. To illustrate this, consider a duplex formed by two strands of two HRs each with sequence 5'-GGGGCC-GGG-GCC-3'. Now define “steps” as steps with Watson–Crick bonding, A = GG/CC = CC/GG, D = CG/CG (CpG step), E = GC/GC (GpC step), or as steps that include mismatches, $m_1 = GG/GC = GC/GG$, $m_2 = GG/CG = CG/GG$, and M = GG/GG. The step or stacking pattern in this sequence becomes $Am_1Mm_1ADAm_1Mm_1A$, and one can see that it repeats periodically. Now consider the sequence 5'-GCCGGG-GCC-GGG-3'. The step pattern in this case becomes $m_1ADAm_1Mm_1ADAm_1$. For long sequences, this is exactly the same pattern as before (same reading frame), and the only differences are simply due to end effects. Thus, both sequences give rise to the same type of helix, shown as duplex “DG-2” (DG stands for double G mismatch) in the scheme in Figure 1. Now consider the sequence 5'-GGGCCG-

GGGCCG-3' as shown in duplex "DG-1" in Figure 1. The step pattern is m₂AEAm₂Mm₂AEAm₂, which is different from the previous two. All in all, neglecting end effects, there are three independent duplexes for each sequence, as shown in Figure 1. The third type, SG-3 (SC-3) stands for single G-G (C-C) mismatch (of course, the number of mismatches is the same in all cases).

In addition to these duplexes, we created two more duplexes for testing. After 1 μ s, we found that the duplex DNA SC-3 became unstable in both sets of simulations: in set A, the DNA helix slowly unfolded while in set B, the second strand shifted one base, transforming it into DNA DC-1. We hypothesized that this behavior reflected a lesser stability for DNA SC-3, and decided to test this in three ways: (i) rerunning SC-3 with different random seed; (ii) adding a small restrain (1 kcal/(mol Å)) between bases C2 and G24 as well as G12 and C14 (we call the new restrained SC-3, DNA R-SC-3); (iii) adding a C at the 3' end of first strand and the 5' end of second strand to form a single mismatching model without hanging bases (duplex SC-3+C). Finally, all the duplex structures were built by using the NAB and LEAP programs in AMBER v.14, and Na⁺ ions were added for neutralization. A summary of all these models is presented in Table 1.

One final consideration for defining the initial structures is the initial glycosidic angle χ of the mismatched base pairs. Cytosine tends to always favor the *anti* state, even in C-C mismatches¹² (unless it is a terminal mismatch, where it can rotate more freely). Guanine, on the other hand, can rotate around the glycosidic bond adopting the *syn* conformation. G_{syn}-C_{anti} Watson-Crick base pairs are observed in both Z-DNA and Z-RNA duplexes.¹³⁻¹⁷ More importantly, G_{syn}-G_{anti} mismatches have been observed in the crystal structures of CGG RNA repeats.^{18,19} There is, however, a considerable energy barrier for this rotation,¹⁷ and we have chosen to insert the right symmetry in the initial structures. Thus, we have set all the initial C-C mismatches in the anti-anti conformation. For DNA and RNA G-rich duplexes, we have chosen G_{syn}-G_{anti} mismatches alternating so as to preserve the symmetry of the strands (and therefore, the symmetry of the steps as described above). This is possible in duplex types DG-1 and DG-2, that exhibit neighboring G-G mismatches. In model SG-3, an anti/syn distribution that preserves the symmetry of the strands results in alternating G_{anti}-G_{anti} and G_{syn}-G_{syn} mismatches, which are energetically less favorable. We have therefore chosen for the initial structures alternating G_{syn}-G_{anti} mismatches, even though the initial symmetry of the strands is broken.

Lastly, for the analysis below, we define a stable duplex as one that keeps its double helical structure and its Watson-Crick hydrogen bonds in the G-C pair during the course of the 1 μ s simulations.

G- and C-Rich Duplexes

Figure 2 shows the distribution of conformations over the second half of the simulation projected onto the first three principal components obtained from the PCA,²⁰ which correspond to the largest positional deviation and most of the positional fluctuations of the DNA and RNA duplexes. The G-rich structures in both DNA and RNA (DG-1, DG-2, and SG-3) in Figure 2 show Gaussian distributions corresponding to independent harmonic motions that have converged. For the DNA models, the first principal component

corresponds to bending around the central hinge and the second, corresponds to bending in DG-1, but to groove breathing in DG-2 and SG-3. For RNA, the first component corresponds to bending, while the second component also corresponds to bending in DG-1 and DG-2, but to groove breathing in SG-3.

In contrast to the G-rich duplexes, non-Gaussian distributions are observed for C-rich duplexes. In particular, for DNA, we find that DC-1 is Gaussian-distributed, with the first two modes corresponding to groove breathing motions. However, DC-2 is not: the first component corresponding to bending exhibits considerable spread, and the second component corresponding to groove breathing has a dominant and a secondary peak. To track the origin of the non-Gaussian distribution, we plotted the time behavior of the first PC. Figure S2 shows that DNA DC-2 exhibits an end effect at later times where the 3' end of one of the strands (C24 here) bends over and interacts with the C6 base on the other strand, which has flipped out. This end effect also considerably distorts the main chain and is reflected in the non-Gaussian distribution of the second principal component. The three different versions of SC-3, SC-3, shown in Figure 2, and R-SC-3 and SC-3+C, shown in Figure S3 all display a bimodal distribution for the first principal component, but are Gaussian-distributed for the next two components. The time behavior of DNA SC-3 (Figure S4) shows two main conformations: In both cases the helical structure is unwound at the top, but the later one (820–1000 ns) is also unfolding at the bottom. Due to the restraint at the terminal bases, DNA R-SC-3 strives to maintain a relative stable structure during the 1 μ s simulation, facing different stages of unwinding at both ends (Figures S3 and S5). Similar behavior is observed in DNA SC-3+C. In Figure S6, we show a state (500–880 ns) corresponding to a half unwound structure and a later state (880–1000 ns) where the duplex has healed its helical structure. All in all, the data for the three versions of DNA SC-3 seem to indicate oscillations between unfolding and partially refolding events.

The C-rich RNA duplexes in Figure 2 show the following behavior: SC-3 is Gaussian-distributed while DC-1 and DC-2 are not. Tracking the time behavior (Figure S7), we see that there are two conformations in DC-1: In conformation A, described in more detail below, the C6–C19 mismatches stay inside the helix and occasionally form a hydrogen bond, while the C7 and C18 bases point toward the minor groove, facing each other. In this case, the major groove is less occupied and the duplex is bent toward the major groove. In conformation B, C7 and C19 stay inside the helix, stacking on top of each other, while the complementary mismatches C18 and C6 form an N3–N4 cross hydrogen bond and turn to the major groove, unbending the duplex (see also Figure S10). Both cases seem to correspond to equilibrium distributions, as the C6 and C18 bases slowly turn toward the minor and major groove, slightly bending or unbending the duplex. Finally, RNA DC-2 stays only in conformation B, with a small amplitude oscillation motion (Figure S8) of the mismatched bases, which results in the perfectly bimodal distribution in Figure 2.

These general results can be further quantified by analyzing the behavior of backbone sugar torsion δ and the glycosyl torsion χ , as well as the “open angle”, as defined in the Supporting Information (SI). Figure 3 shows the χ and δ torsion angle distributions of duplex models in DNA form (left) and RNA form (right) corresponding to internal G-G mismatches (top three panels) and internal C-C mismatches (bottom three panels; end

mismatched bases are excluded). The salient features of these graphs are that the distributions associated with RNA are relatively narrow, while there is considerably dispersion with those associated with DNA. For the G-G mismatches, the two distributions associated with G_{syn} and G_{anti} are very distinct. The bases in the C-C mismatches are all in some form of anti conformation. The time evolution graphics of the χ torsion angle for the internal mismatches for the DC-1 and DC-2 forms of DNA and RNA are shown in Figure S9. For the DNA DC-2 model, the distributions are centered around -106° and -130° , corresponding to anti-anti configurations. For the DNA DC-1 model, the χ angles of C6–C19 are Gaussian distributed at -102° and -80° , which correspond to anti and high anti configurations. The pair C7–C18, on the other hand, has a bimodal distribution with values around -112° and -68° for C7; and -120° and -70° for C18. This reflects the presence of the “e-motif”, described below, where the C7 and C18 bases flip out into the minor groove. The distributions for the RNA forms are centered around -160° (corresponding to anti periplanar conformations).

The open angle (see definition in the SI) distributions of DNA and RNA are shown in Figure 4. Both C and G bases that participate in Watson–Crick hydrogen bonds have a Gaussian distribution around 60° . Distributions for G-G mismatches are centered at $30\text{--}35^\circ$ and 75° , indicative of flipping toward minor and major grooves, respectively. Distributions for C-C mismatches are spread-out and generally multimodal, reflecting the rotating activity of these bases. The time evolution of the open angle for the mismatches and the associated conformations are described below.

Now that we have an overall view of the main structural features of the duplexes and the time behavior of the first principal component and glycosyl angle χ , both shown in the SI, we can concentrate on more particular structural features. First, we consider the single C (SC-3) mismatch models. For DNA, as discussed previously, the various conformations associated with the non-Gaussian peaks in the first principal component distribution (shown in Figures 2 and S3) suggest that the various versions of SC-3 are not as stable as those for DC-1 and DC-2. Figure 5 shows this in more detail. In the “set A” simulations (corresponding to the data presented here, see the Methods), we see partial unfolding of the structure: (i) the base C1 at the 5' end of one strand bends over and interacts with the base moiety of C4; (ii) the mismatched C10 base in the same strand flips into the major groove; (iii) Watson–Crick pairs G11–C15 and G12–C14 have been broken; (iv) ending bases C13 and C14 shift toward the 3'–5' direction, and bend and get trapped in the minor groove. In the “set B” simulations (corresponding to a slightly less extensive equilibration process), one of the strands shifts, transforming the duplex into a partial DC-1 structure with reconstitution of four successive Watson–Crick hydrogen bonds, as shown in Figure 5. For both sets of simulations, SC-3 $\neq$ C shows clear signs of unwinding. Finally the “restricted” simulation R-SC-3 seems to be relatively stable for set A but mutates into DC-1 for set B. Figure S3 compares the structures of R-SC-3 and SC-3 $\neq$ C for set A. Interestingly, the RNA SC-3 duplex—in spite of exhibiting a perfect Gaussian distribution for the first principal component, with a very narrow distribution in the χ – δ map also undergoes hydrogen bond shifting, as explained below.

Next, we consider the double C mismatches, DC-1 and DC-2, which are stable up to 1 μ s for both DNA and RNA (except for end effect fluctuations). Let us first consider DNA. Figure 6 shows the time evolution of the open angle for the internal mismatches in DC-1 and DC-2 respectively, along with associated conformations. The internal mismatches correspond to the pairs C6–C19 and C7–C18. In the DC-1 structure in Figure 6a,b, the mismatched bases C7 and C18 start to turn toward the minor groove early in the simulation: around 160 ns, C7 flips out and toward the 5' direction in the strand, interacting with G5; and C18 flips out and toward the 5' direction of its own strand at about 300 ns, interacting with G16. Both G5–C18 and C18–G16 form G(N3)–C(N4) hydrogen bonds. This gives rise to the rarely seen “e-motif”, where the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the $i-2$ residue (i.e., toward the 5' direction in each strand), as shown in Figure 6a,b. Once formed, the “e-motif” is stable from 300 ns to the end of the simulation at 1 μ s. The other pair, C6–C19, initially shows some flipping activity (C19 turning toward the minor groove and C6 turning toward the major groove), but after 300 ns they both settle inside the helical structure. Now we turn to DNA DC-2 in Figure 6c,d. In this case, the C7–C18 mismatch mainly stays inside the helix and the conformation of the C6–C19 mismatch evolves slowly. During the first 560 ns, C6–C19 and C7–C18 either form an internal pair or slightly turn to the major and minor grooves as shown in Figure 6d. After 560 ns, the bases exhibit different behavior. During 560–820 ns, C6 turns toward the minor groove and C19 flips out into the major groove. During the last 180 ns, C6 flips completely out into the minor groove, and upward toward the 5' direction of its strand, while C19 stays in the major groove. During 560–1000 ns, C18 still stays inside the helix, while C7 turns toward the major groove forming an O2–N4 hydrogen bond with C18 (Figure 11). Finally, a comparison between the two DNA DC duplexes shows that the e-motif structure is better stacked, more compact than the structure in DNA DC-2. This can be observed, for instance, in Figure 6 and can be quantified by measuring the distance d_i between neighboring layers for each base pair (i.e., for base pair i , we compute the distances between base pairs $i-1$ and $i+1$). This distance is mainly the same for layers 3, 4, 5, 9, and 10, but is shorter around the mismatches for layers 6, 7, and 8 for DC-1, which is a measure of the better packing of the e-motif. (The average distances for DC-1 are $d_6 = 5.15$ Å, $d_7 = 3.13$ Å, and $d_8 = 4.90$ Å; and for DC-2 are $d_6 = 6.52$ Å, $d_7 = 7.11$ Å, and $d_8 = 7.81$ Å.)

Here, we make some observations with respect to the flipping of the bases. An analysis of all the trajectories for C-rich DNA shows that when a C base completes its flipping into the minor groove, it folds back and interacts with bases toward the 5' direction on its own strand. This directionality has also been observed in careful MD studies of base flipping with conformational flooding (Figure 7a in ref 21). On the other hand, when the C base flips toward the major groove, the bases have more space and therefore are more flexible, interacting less with the rest of the helix. We have plotted the backbone dihedral angles as a function of the open angle and found that most angles do not give a clear indication of the transition but instead reflect increased fluctuations and transitions. Indeed, same values of the open angle can correspond to completely different values of the backbone torsion angles. However, by plotting the backbone torsion angles as a function of time, we were able to pick up two distinctive trends for flipping into the major versus the minor grooves. Figure 7 shows the torsion angles α and γ for the C7–C18 mismatch in DNA DC-1. While most

backbone angles do not reflect the transition, α decays at the transition points, and then it assumes the characteristic anticorrelation with γ , where $\alpha + \gamma$ stays constant. Alternatively, if the sum $\alpha + \gamma$ is depicted, it decreases approximately by 100° , when the corresponding base, C7 at around 160 ns and C18 at around 300 ns, flips out. The torsions ϵ and ζ also show small jump, synchronized transitions at these times, with constant difference $\epsilon - \zeta \approx -90^\circ$, corresponding to a BI backbone conformation for this particular mismatch. For the other mismatch, C6–C19, α and γ stay constant, but the C6 angles ϵ , defined as the C4'–C3'–O3'–P($i+1$) torsion, and ζ , defined as the C3'–O3'–P($i+1$)–O5'($i+1$) torsion, show an increase at 160 ns from $\epsilon = 188^\circ$ to $\epsilon = 290^\circ$, and a decrease from $\zeta = 277^\circ$ to $\zeta = 81^\circ$, clearly due to the coupling of these angles to the configuration of C7. For C6, this represents a transition from $\epsilon - \zeta \approx -90$ (BI) to $\epsilon - \zeta \approx 210$. The flipping out into the major groove, as occurs with C19 after 560 ns in DNA DC-2, is signaled by an increase in ϵ and a decrease in ζ : ϵ increases from $\epsilon = 188^\circ$ to $\epsilon = 283^\circ$ and ζ decreases from $\zeta = 277^\circ$ to $\zeta = 81^\circ$ ($\epsilon - \zeta \approx -90$ to $\epsilon - \zeta \approx 200$). C18 also turns toward the major groove but it never flips out: its ϵ angle increases from $\epsilon = 193^\circ$ to $\epsilon = 283^\circ$, but ζ stays the same.

Now, we consider the RNA C-rich duplexes. Typical conformations are depicted in Figures 8 and 9. There is a clear coupling between mismatches that shuttle between major and minor grooves and bending of RNA. Figure S10 shows the open angle of C18 superposed to the radius of gyration (R_g) of the DC-1 RNA duplex as a function of time (bending is hard to characterize for such a short duplex, so we show radius of gyration instead). We observe two main mismatch conformations in RNA DC-1 (labeled A and B in Figure 8). During the 0–180, 310–600, and 880–1000 ns time intervals, DC-1 is under configuration A, where C6–C19 mismatches stay inside the helix and occasionally form a N3–N4 hydrogen bond (Figure 11), while the C7 and C18 bases point toward the minor groove, facing each other with the open angle of C7 and C18 dropping to 0° . Since both C7 and C18 are in the minor groove in conformation A, the major groove is less occupied and bases $i-1$ and $i+1$ can interact pulling the backbones closer so that the duplex exhibits a smaller R_g , or is bent, as shown in Figure S10. In conformation B, both C7 and C19 stay inside the helix stacked together, but C6 and C18 turn to the major groove forming an O2–N4 hydrogen bond (Figure 11) in their corresponding C6–C19 and C7–C18 mismatches. In this conformation, the occupation of the major groove by C6 and C18 pushes the strands away, unbending the duplex, which adopts a slightly larger R_g value. In the RNA DC-2 duplex, only conformation B is observed, as shown in Figure 8, with C6 and C18 in the major groove. Figure S10 shows that the radius of gyration stays at a constant, “unbent” value of approximately 14 Å (there is an abnormal R_g value from 375 to 500 ns corresponding to one end bending toward the main chain; the end effect disappears after 500 ns).

For the RNA SC-3 duplex shown in Figure 9, the C4–C22 and C7–C19 pair bases favor a synchronized oscillation motion between minor and major grooves, where if one base of the mismatched pair turns toward the minor groove, the other turns toward the major one. The pair C10–C16 is under the same oscillation motion before 450 ns, but after that, the Watson–Crick hydrogen bonds in base pair G11–C15 and G12–C14 are broken, and new Watson–Crick hydrogen bonds are formed in the base pairs G11–C16 and G12–C15, while C10 remains trapped inside the helix.

By contrast, the G-rich duplexes are all stable, both for DNA and RNA, with internal mismatches always inside the helix. In particular, the G_{syn} - G_{anti} conformations for the internal G-G mismatches are stable throughout the simulations (of course, end mismatches are subject to fluctuations). The alternating syn/anti conformation causes G_{syn} to point toward the minor groove, with an open angle value centered around 35° , while G_{anti} points toward the major groove around an open angle 75° , while still remaining inside the helix. Figure 10 shows the twist angle based on the $C1'-C1'$ vectors (see the SI for a definition of this twist). Following the definition of “steps” given in the Initial Structures subsection, the steps 2,..., 10 correspond to AEA₂Mm₂AEA for DG-1, ADA₁Mm₁ADA for DG-2, and m₁m₁Am₂m₂Am₁m₁A for SG-3. Converged structures result in a twist that reflects the inversion symmetry of the sequence. Notice that the middle mismatches result in local unwinding of the helix. The G-G mismatch stays inside the helix and is stabilized by two hydrogen bonds, N7–N2 and O6–N1, present in both DNA and RNA; and a third hydrogen bond, N2–OP2, only present in RNA (Figure 11).

Figure 11 shows the most common hydrogen bond patterns for the mismatches. Generally in both DNA and RNA there are two hydrogen bonds between the Watson–Crick edge of G_{anti} and the Hoogsteen edge of G_{syn} : O6–N1H and N7–N2H. In addition, RNA mismatches are further stabilized by a hydrogen bond between N2H in G_{syn} and its phosphate oxygen atom OP2 (see Table S1 for hydrogen bond populations). The population of the hydrogen bonds in the C-C mismatches, on the other hand, varies wildly (see Table S2). As Figure 11c shows, there are few hydrogen bonds in DNA DC-1, reflecting the tendency of the bases to flip out. The presence of hydrogen bonds increases for DNA DC-2, especially for O2–N4 in C7–C18. Hydrogen bonds in RNA DC-1 also seem to form a considerable amount of the time, and RNA DC-2 is remarkable in the sense that there are no N3–N4 hydrogen bonds, but the O2–N4 bonds are populated all the time. The hydrogen bonds populations are given in Tables S1 for G-G mismatches and Table S2 for C-C mismatches.

Now we turn to ion occupation whose definition is given in the SI (assignment to major or minor groove is slightly tricky around the mismatches). Figures S11 and S12 show the neutralizing Na^+ ion occupation as a function of base pairs in the major and minor grooves, and Figure 12 shows some typical patterns for ions involved in direct binding to the mismatched bases. We know from studies of ion distributions in the sequence $(CG)_6$ ²² that neutralizing Na^+ ions sit preferentially around the G bases in the major grooves, both in B-DNA and A-RNA forms. This trend is clearly seen in the G-rich duplexes in Figures S11 and S12 where the amount of ions in the minor groove is completely negligible and most of them sit in the major groove, especially around the G-G mismatches. Notice that for both DNA and RNA, the neutralizing ion distributions are nicely converged reflecting the symmetry of the molecules. A Na^+ ion can stabilize a G-G mismatch by forming an ion bridge connecting the G(O6) atoms in the four G bases in a double G-G mismatch (Figure 12a); or the two bases in a single G-G mismatch and the two bases in the adjacent G-C base pair (Figure 12b). These pockets are very similar to the Na^+ ion trapped in a $(CG)_6$ duplex, where the ion is held by two G(O6) and two C(N4), in both B-DNA and A-RNA (see Figure 14a and b in ref 22).

For the C-rich motifs the story is slightly more complicated. While a single Na^+ ion is needed to bridge four G bases in a double mismatch, two Na^+ ions are required for a double C-C mismatch, as shown in Figure 12c. This situation occurs in C-C mismatches that tend to form an N3–N4 hydrogen bond. Figure 12d shows an ion bridge between two bases that are involved in O2–N4 hydrogen bonds. Figure 12e shows an ion trapped between a Watson–Crick bonded base pair and the immediately following C-C mismatch. For RNA, ions still prefer to sit in the major groove. While in RNA DC-2 the ions are still more localized in the central mismatches in the major groove, in both RNA DC-1 and RNA SC-3 the neutralizing ion distribution in the major groove favors the Watson–Crick pairs and there is a non-negligible distribution of ions in the minor groove around the C-C mismatches. RNA DC-2 exhibits only conformation B, where one of the mismatches turns toward the minor groove, and the other stays inside the helix, both bases linked by an O2–N4 hydrogen bond. The internal pairing of the C bases reduces the distance between the two strands of the duplex resulting in a narrower minor groove that admits no ions (Figure 12d,e). In this configuration, the ion trapped at the major groove around the C mismatches is also held in place by the adjacent G-C base pair, as shown in Figure 12e. In RNA DC-1, where both conformations A and B exist, the ion can be trapped at the minor groove when the bases form a N3–N4 hydrogen bond, as shown in Figure 12c. However, if the two bases form an O2–N4 hydrogen bond, the ion favors the major groove (Figure 12d), as in the RNA DC-2 model. By contrast, in C-rich DNA the ion population in the minor groove considerably exceeds the population in the major groove, and it is distributed around the C-C mismatches (the DC-1 population is not completely symmetric due to the presence of the e-motif). A distribution where the minor groove Na^+ ion population is larger than that in the major groove is curiously reminiscent of the ion distribution around $(\text{CG})_6$ Z-DNA.²² In addition, Figures S13 and S14 present radial distribution functions for the various C-C mismatches for C(O2) atoms in the minor groove and C(N4) atoms in the major groove. Here we have concentrated on the main differences between C-C and G-G tracts, but it is obvious that ions are attracted to other electronegative groups in the sugar rings and backbones. Figure 12 showcases two such cases: panel (f) shows two Na^+ ions directly bound to OP1 atoms in the minor groove of DC-1 DNA, while panel (g) shows a bridge ion linking the ends of the helix and thus aiding in the bending of a DG-2 RNA duplex. Situations like this are also observed, for instance, in an A-RNA $(\text{CG})_6$ helix.²²

Finally, for the sake of reference, we include in Table S3 results for the structure prediction of 10 hexanucleotide repeats and their corresponding free energy as calculated by the Mfold web server.²³ Figure S15 shows the predicted minimum free energy structures. The results are only for reference, as Mfold does not consider duplexes but single strands and therefore the resulting structures are hairpins, and the particular geometry of the mismatches is not considered.

DISCUSSION

The fact that the (GGGGCC) hexanucleotide repeat (HR) expansion in the first intron of the C9ORF72 gene is the major cause behind both C9FTD and ALS diseases was only recently established.^{2,3} This HR has thus joined the family of simple-sequence nucleotide repeat (SSR) diseases, consisting of 1–6 (and now 12) nucleotides that are repeated up to 30 times

(and more for pathological cases).²⁴ SSRs exhibit “dynamic mutations” that do not follow Mendelian inheritance. In fact, the intergenerational expansion of SSRs gives rise to the “genetic anticipation” diseases,^{25–27} where the age of onset decreases and the severity of the disease increases with successive generations. In addition, the “penetrance”, i.e., the probability that this type of mutation results in disease, also increases with successive generations.

An important breakthrough in the field of SSR diseases has been the recognition that stable atypical DNA secondary structure in the expanded repeats is “a common and causative factor for expansion in human disease”.⁹ There is, therefore, a considerable urge to determine these atypical nucleotide structures as they drive the interaction with other molecules resulting in further expansion of the SSRs. With respect to the (GGGGCC) HRs, the experimental data is rather scarce. Chemical and enzymatic probing of a r(GGGGCC) repeat expansion points to a general scenario where the repeat expansion adopts a hairpin structure with G-G mismatches in equilibrium with a quadruplex structure.¹⁰ Recently, CD, optical melting and 1D ¹H NMR spectroscopy, combined with chemical and enzymatic analysis reveal that r(GGGGCC) expansions fold into a G-quadruplex in equilibrium with hairpin structures.¹¹ The equilibrium is temperature dependent, with $T = 37^\circ$ favoring hairpins with long stems and higher annealing temperatures favoring quadruplexes. Currently there is no experimental data with atomic resolution for any of these hairpins or duplexes. We have therefore carried out a structural and dynamical study based on MD simulations of the 12 dodecamers that result from the three different reading frames in sense (GGGGCC) and antisense (CCCCGG) HRs for both DNA and RNA.

The three possible G-rich and C-rich double helices are shown in Figure 1. As described in the Methods, for either the sense or antisense sequences, the double helices (when repeated periodically to avoid differences due to end effects) differ in the pattern of “steps” (that also include mismatches) and, therefore, are not equivalent. All structures, when repeated periodically, initially have the same number of G-G or C-C mismatches, and the same number of Watson–Crick bonded G-C base pairs. One obvious difference is that the “double G” DG and “double C” DC helices have adjacent, double mismatches, separated by four Watson–Crick G-C base pairs, while the “single G” SG and “single C” SC helices have single mismatches, separated by two Watson–Crick G-C base pairs.

Before discussing specific results, it is worthwhile to discuss two general trends. First, for a given sequence, mismatched RNA duplexes are more stable than mismatched DNA duplexes, as supported by the fact that the few crystal structures of G-G and C-C (and also A-A and U-U) mismatched double helices are only available for RNA structures. For example: (i) crystal structures with G-G mismatches (GCGGCGGC duplex,¹⁸ PDB id 3R1C; GC(8BrG)GCGGCGGC duplex,¹⁸ PDB id 3R1D; copies of the 5'-CGG/3'-GGC motif,¹⁹ PDB id 3JS2); (ii) crystal structures with C-C mismatches (GCCGCCGC duplex,¹² PDB id 4E59); (iii) crystal structures with A-A mismatches (GGC-AGCAGCC duplex,²⁸ PDB ids 3NJ6, 3NJ7). Second, the G-G mismatches are more stable than the C-C mismatches.^{12,18,29} This is also not surprising, as the G-G mismatches favor both good hydrogen bonds and good stacking within the helix, but the C-C have weak hydrogen bonding and considerably less favorable stacking. This is what makes G-quadruplexes

relatively abundant in the PDB, while continuous C-quadruplexes are nonexistent. These general trends are also supported by detailed free energy maps that we are at present computing in the context of trinucleotide GGC and CCG repeats, and by the general trends in the least Gibbs free energy structures as calculated with the Mfold server and presented in Table S3.

G-Rich Double Helices

There are several features that the six DNA and RNA helices with G-G mismatches share:

(1) The G-rich double helices are stable—This seems to be the case, at least up to 1 μ s, for the set of simulations presented here (set A) and for a second set of 1 μ s simulations that followed a shorter equilibration cycle (set B). There are occasional end effects, where the terminal mismatched bases flip back, but these do not affect the structure of the inner helix. Indeed, for the six DNA and RNA cases, the PCA reveals smooth Gaussian distributions for the first three principal components in each case, that correspond to bending or to groove breathing modes.

(2) The preferred conformation for the internal G-G mismatches is G_{syn} - G_{anti} —This is clearly illustrated in the map for the backbone sugar torsion δ and the glycosyl torsion χ , shown in Figure 3, where distinct distributions are associated with G_{syn} and G_{anti} . The syn and anti conformations correlate with the value of the open angle associated with the corresponding base. The open angle distributions of DNA and RNA are shown Figure 4. Both C and G bases that participate in Watson–Crick hydrogen bonds have a Gaussian distribution around 60°. In the mismatches, G_{syn} points toward the minor groove, with an open angle value centered around 35°, while G_{anti} points toward the major groove with an open angle around 75°.

(3) The inner G-G mismatches stay inside the helix core and form hydrogen bonds—Figure 11 shows the most common hydrogen bond patterns for the mismatches. For both DNA and RNA, there are two hydrogen bonds between the Watson–Crick edge of G_{anti} and the Hoogsteen edge of G_{syn} : O6–N1H and N7–N2H. In addition, RNA (but no DNA) mismatches are further stabilized by a hydrogen bond between N2H in G_{syn} and its phosphate oxygen atom OP2. These hydrogen bonds have been observed experimentally for (CGG) trinucleotide repeat RNA duplexes¹⁸ (the rather high bond populations are listed on Table S1).

(4) The inner G-G mismatches cause local unwinding of the helix—The twist angle based on the C1'–C1' vectors in Figure 10 reflects the inversion symmetry of the sequences. A local unwinding of the helix is reflected in a considerable but localized decay in twist in the inner mismatches.

(5) Ions sit in the major groove, especially around the G-G mismatches, and help stabilize G-G double mismatches via ion bridges—This trend occurs for both DNA and RNA, while ion occupation in the minor groove is completely negligible. In addition, ion bridges such as those shown in Figure 12 help stabilize the mismatches; in this

case, by connecting the G(O6) atoms of the four mismatched G bases, or the G(O6) atoms of a G-G pair with the G(O6) and C(N4) atoms of an adjacent Watson–Crick base pair.

Finally, for either DNA or RNA, the current simulations are not enough to predict which model (DG-1, DG-2, or SG-3) is more stable.

C-Rich Double Helices

Unlike their G-rich counterpart, the C-rich helices show a wide variety of conformations, with the one common characteristic that the inner mismatched C bases are all in some form of anti conformation. We thus need to consider the duplexes separately, and first consider DNA.

(1) C mismatched base flipping—Unlike Watson–Crick hydrogen-bonded base pairs where base flipping takes of the order of milliseconds or longer, and enhanced sampling techniques are needed in simulation studies,²¹ base flipping in C-C mismatches occur in the microsecond time scale and can be observed in regular molecular dynamics simulations. While RNA C-C mismatches remain inside the helix, inclined toward the major or minor groove, DNA mismatched Cs tend to flip out of the helix core. An analysis of all the trajectories for C-rich DNA shows that when a C base completes its flipping into the minor groove, it folds back and interacts with bases toward the 5' direction on its own strand. This directionality has also been observed in careful MD studies of base flipping with conformational flooding.²¹ When the C base flips toward the major groove, the bases have more space, are more flexible and interact less with the rest of the helix. Linear combinations of the backbone torsion angles signal the flipping transition: a decay of $\alpha + \gamma$ of about 100° indicates a flipping toward the minor groove, and a jump of $\epsilon - \zeta$ of about 290° signals a flipping into the major groove.

(2) DNA DC-1 is stable and accommodates mismatches via the formation of e-motifs—PCA shows nice Gaussian distribution for the three principal components, with the first two modes corresponding to groove breathing, as the shortening of the helix due to the e-motif hinders bending for the dodecamer. Figure 6d tracks the evolution of the open angle for the internal mismatches C6–C19 and C7–C18. The mismatched bases C7 and C18 start to turn toward the minor groove early in the simulation: C7 flips out and toward the 5' direction in the strand, interacting with G5; and C18 flips out and toward the 5' direction of its own strand, interacting with G16. Eventually, both G5–C18 and C18–G16 form G(N3)–C(N4) hydrogen bonds. This gives rise to the rarely seen “e-motif”, first described in a solution conformation of a DNA CCG triplet repeat sequence in NMR experiments,³⁰ where the C bases (*i* residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the *i* – 2 residue (i.e., toward the 5' direction in each strand). Once formed around 300 ns, the “e-motif” is stable to the end of the simulation at 1 μ s. After some initial flipping activity (C19 turning toward the minor groove and C6 toward the major groove), C6–C19-both settle inside the helical structure at around 300 ns, with occasional hydrogen bonding (Table S2).

(3) DNA DC-2 is stable with the bases of a mismatch alternating between the minor and major grooves—Figure 6c,d shows that the C7–C18 mismatch mainly stays inside the helix and the conformation of the C6–C19 mismatch evolves slowly. During 500–820 ns, C6 turns toward the minor groove and C19 flips out into the major groove. During the last 180 ns, C6 flips completely out into the minor groove, and upward toward the 5' direction of its strand, while C19 stays in the major groove.

(4) DNA SC-3 is unstable—This was confirmed through six 1 μ s runs, with different conditions on the terminal bases to make sure that this is not due to the presence of a single hanging base at the end. The duplexes in these runs either unfold or convert into DC-1. Thus, it seems that C-rich DNA needs at least four consecutive Watson–Crick base pairs in order to preserve stability (notice that Mfold also favors DC-1, and when SC-3 is enforced, it still switches to DC-1; see Table S3).

(5) The DC-1 e-motif structure is better stacked than the DNA DC-2 structure—A comparison of the DC-1 and DC-2 duplexes after 500 ns shows that the e-motif structure is better stacked, more compact than the structure in DNA DC-2. This can be observed in Figure 6 and can be quantified, for instance, by measuring the distance between neighboring layers for each base pair (i.e., for base pair i , we compute the distances between base pairs $i - 1$ and $i + 1$). This distance is the same for both duplexes except near the mismatches, where it is shorter for layers 6, 7, and 8 for DC-1. Thus, the distance between basepairs 5 and 9 is 10.05 Å for DC-1 and 14.33 Å for DC-2. This is a measure of the better stacking of the e-motif.

(6) For the DNA antisense (CCCCGG) HR duplexes, the stability is ranked as DC-1 > DC-2 > SC-3 (unstable)—We have shown that the e-motif is better packed than the DC-2 resulting structure, and probably it is the optimum structure for the C-rich hexanucleotide repeats. We know that the DC-1 sequence can adopt this motif, but we have not shown that this motif could not be present (if the simulations were long enough) in the other two models, DC-2 and SC-3. Fortunately, in this case it is not necessary to run any longer to realize that the e-motif has to be less stable, or unstable, in the other two structures. This is due to the nature of the steps surrounding the mismatches. Indeed, going back to Figure 1, one can see that the inner mismatched bases in DC-1 can flip toward the 5' strand and encounter a G base with which to form a G(N3)–C(N4) hydrogen bond. On the other hand, in DC-2 there are no G bases in the 5' direction to help stabilize the flipping mismatches. Finally, in SC-3, a flipping C in the 5' direction does encounter a G base but contributes to the destabilization of the helix, as it compromises one of the two hydrogen bonds that separate the mismatches.

(7) For the DNA antisense (CCCCGG) HR duplexes, the ions occupy the minor groove and are centered around the mismatches—A Na⁺ ion distribution around the minor groove has also been observed in (CG)₆ Z-DNA.²²

Now we consider the RNA C-rich duplexes.

(8) RNA DC-1 displays two stable conformations: A, with one mismatch pair pointing toward the minor groove, and the other mismatch forming a N3–N4 hydrogen bond; and B, with two mismatching bases forming an O2–N4 hydrogen bond. The occupation state of the major groove by mismatched bases is coupled to the helix bending—In conformation A, shown in Figures 8 and S10, the C6–C19 mismatches stay inside the helix and occasionally form an N3–N4 hydrogen bond (Figure 11), while the C7 and C18 bases point toward the minor groove, facing each other. In this case, the major groove is less occupied and bases $i - 1$ and $i + 1$ interact pulling the backbones and bending the duplex. In conformation B, C7 and C19 stay inside the helix, stacking on top of each other, while the complementary mismatches C18 and C6 turn to the major groove. The mismatches C6–C19 and C7–C18 interact through O2–N4 hydrogen bonds, and occasionally C6 and C18 form a N3–N4 cross hydrogen bond. In this conformation, the occupation of the major groove by C6 and C18 pushes the strands away, unbending the duplex.

(9) RNA DC-2 is stable in conformation B—Figures 8 and S10 show that only conformation B is observed in the RNA DC-2 duplex, with O2–N4 hydrogen bond forming between C6 and C19 as well as C7 and C18.

(10) RNA SC-3 is the least stable RNA conformation, exhibiting some reorganization of hydrogen bonds—For the RNA SC-3 duplex shown in Figure 9, the C4–C22, C7–C19, and C10–C16 pair bases favor a synchronized oscillation motion between minor and major grooves, where if one base of the mismatched pair turns toward the minor groove, the other turns toward the major one. After 450 ns, the Watson–Crick hydrogen bonds in base pair G11–C15 and G12–C14 are broken, and new Watson–Crick hydrogen bonds are formed in the base pairs G11–C16 and G12–C15, while C10 remains trapped inside the helix. Obviously, a reorganization of hydrogen bonds breaks the symmetry that defines SC-3. Interestingly, in the crystallization of RNA C-(CCG)₂-C, it was observed that the C-(CCG)₂-C oligomer formed a “slippery duplex” with overhanging nucleotides and terminal C bases expelled out of the helix core such that the C-C mismatch was surrounded on both sides by four Watson–Crick C-G base pairs.¹² This supports the idea that C-C mismatches in RNA, like those in their DNA counterpart, need to be separated by at least four successive Watson–Crick base pairs in order to preserve helix stability.

(11) In RNA DC-2, neutralizing Na⁺ ions favor the major groove around the C-C mismatches, while in RNA DC-1 and SC-3 ions in the major groove favor the Watson–Crick base pairs, and there is a non-negligible ion occupation in the minor groove around the C-C mismatches. Ion direct binding contributes to the stability of the duplexes—While a single Na⁺ ion is needed to bridge four G bases in a double mismatch, two Na⁺ ions are required for a double C-C mismatch (Figure 12c). This situation occurs in C-C mismatches that tend to form an N3–N4 hydrogen bond. Ion bridges also occur between two bases involved in O2–N4 hydrogen bonds (Figure 12d) and between mismatched bases and Watson–Crick base pairs (Figure 12e).

CONCLUSION

We have carried out the first comprehensive study based on MD simulations of the 12 dodecamers that result from the three different reading frames in sense (GGGGCC) and antisense (CCCCGG) HRs for both DNA and RNA. Our observations can be summarized as follows. 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} : O6–N1H and N7–N2H. In addition, G_{syn} in RNA forms a base-phosphate OP2–N2H hydrogen bond. Inner G-G mismatches cause local unwinding of the helix. Na^+ ions sit in the major groove and help stabilize G-G double mismatches via ion bridges that join either four G bases belonging to a double mismatch, or two Gs in a mismatch with the bases in the adjacent Watson–Crick G-C base pair. G-rich double helices are more stable than C-rich helices due to better stacking and to the hydrogen bonds of the G_{syn} - G_{anti} mismatches.

Unlike their G-rich counterpart, the C-rich helices show a wide variety of conformations, with the one common characteristic that the inner mismatched C bases are all in some form of anti conformation. Least (most) stable C-rich RNA and DNA helices have single (double) mismatches separated by two (four) Watson–Crick basepairs. For the C-rich DNA duplexes, mismatched Cs tend to flip out of the helix core, while RNA C-C mismatches remain inside the helix. A decay of the DNA backbone torsion angles $\alpha + \gamma$ of about 100° indicates a flipping of a C base toward the minor groove, and a jump of $\epsilon - \zeta$ of about 290° signals a flipping into the major groove.

For DNA, the duplex stability is ranked as DC-1 > DC-2 > SC-3 (unstable). The DNA DC-1 helix accommodates mismatches via the formation of e-motifs, where mismatched bases flip toward the minor groove and point in the $5'$ direction of their respective strands. The DNA DC-2 duplex is stable with the bases of a mismatch alternating between the minor and major grooves, but the e-motif in DNA DC-1 provides far better stacking and stability to the double helix. The fact that the flipped C bases at position i in an e-motif form G(N3)–C(N4) hydrogen bonds with the $i - 2$ G base makes DC-1 the preferred sequence for the e-motif. There are two stable conformations in the C-rich RNA duplexes: conformation A, with one mismatch pair pointing toward the minor groove and the other mismatch pair forming a N3–N4 hydrogen bond; and conformation B, with two mismatching bases forming O2–N4 hydrogen bonds. The occupation state of the major groove by mismatched bases is coupled to the helix bending. RNA DC-1 displays both conformation A and B. In conformation A, one mismatch remains inside the helix while the other points toward the minor groove, thus leaving the major groove relatively unoccupied and causing the helix to bend toward the major groove. In conformation B, two bases belonging to different mismatches stack on top of each other, while their partners turn to the major groove, causing unbending of the helix. On the other hand, RNA DC-2 only exhibits conformation B. In C-rich DNA helices, neutralizing Na^+ ions occupy the minor groove, centering around the C-C mismatches. In RNA DC-2, these ions favor the major groove around the C-C mismatches, while in RNA DC-1 and SC-3, neutralizing ions in the major groove prefer the Watson–Crick base pairs, and there is a non-negligible ion occupation in the minor groove around the C-C

mismatches. Ion direct binding to C-C mismatches contributes to the stability of the duplexes.

It is important to emphasize that currently there is no experimental data with atomic resolution for any of these duplexes or for hairpins. Thus, our study is timely and we hope that this characterization of these HR structures will ultimately help in understanding the origins and mechanisms behind nucleotide repeat disorder diseases.

METHODS

The simulations were carried out with the AMBER 14 simulation package³¹ with the ff12SB force field³² with parameters ff99BSC0³³ for DNA and 99BSC0+ χ_{OL3} ^{34–36} for RNA. We note that the AMBER force fields are considered the best among all the force fields presently available for the simulation of nucleic acids. All the duplex structures were built by using the NAB and LEAP programs in AMBER v.14. Na⁺ ions were added for neutralization by using Joung/Cheatham ion parameters³⁷ for TIP3P water (radius = 1.369 Å, mass = 22.99 *u*, and ϵ = 0.0874393 kcal/mol). The TIP3P water model³⁸ was used for the explicit solvent simulations with periodic boundary conditions in truncated octahedron water boxes (larger than (61 Å)³) with more than 5000 waters each. Electrostatics were handled by the PME method,³⁹ with a direct space cutoff of 9 Å, and with an average mesh size of approximately 1 Å for the lattice calculations. We used Langevin dynamics⁴⁰ with a coupling parameter γ = 2.0 ps⁻¹. The NPT simulations were carried out mainly at 300 K and 1 atm with the Berendsen barostat⁴¹ with an isothermal compressibility of β = 44.6 × 10⁻⁶ bar⁻¹ and pressure relaxation time τ_p = 1.0 ps. The length of the simulation was 1 μ s for each model, and the coordinate information was collected every 5 ps.

The equilibration process took place in five steps. First, we carried out a steepest descent simulation followed by conjugate gradient minimization keeping the nucleic acid and ion atoms fixed at their initial positions. Second, the system was slowly heated up from 0 to 300 K in five steps, for a total of 1 ns, under constant volume condition and restraints (100 kcal/(mol Å)) on the nucleic acid and ion atoms. Third, another 1 ns NVT run with restraints on the nucleic acid and ion atoms was carried out at 300 K. Fourth, 1 ns NPT runs with slowly decreasing restraints from 100 to 0 kcal/(mol Å) in five steps were carried out for a total of 1 ns. In this step, the system reached a stable density, around 1.0 g/mL. Fifth, 1 ns NPT Langevin dynamics with constant pressure at 300 K without restraints was carried out as a final check. These final structures were used for the 1 μ s simulations. This set of simulations was labeled “set A”. We also have another set of 1 μ s simulations that were carried out with a less extensive equilibration process. These simulations (labeled “set B”, not shown here except for a few cases), give very similar results to the ones presented here, with the main differences due to end effects in DNA, and different behavior in DNA SC-3 (shown in the Results), which we believe is due to the lesser stability of this arrangement.

Principal component analysis (PCA) analysis was carried by using the PCA suite software.⁴² The hydrogen bonds population is calculated by using CPPTRAJ⁴³ in AMBER v.14. A hydrogen bond is defined as being between an acceptor heavy atom A, a donor hydrogen atom H, and a donor heavy atom D. The alignment angle θ is defined as the angle AH–D

with a cutoff 135°, and the hydrogen bond length as the distance between atoms A and D with a cutoff 3.5 Å. Definitions or comments for the PCA, helix open angle, helix twist angle, and ion occupancy are given in the SI.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work has been supported by the National Institute of Health (NIH) via R01GM118508; and the National Science Foundation (NSF) via SI2-SEE-1534941.

We also thank the NC State HPC and especially the Extreme Science and Engineering Discovery Environment (XSEDE) TG-MCB160064 for computational support.

References

1. Lillo P, Hodges JR. Frontotemporal dementia and motor neurone disease: Overlapping clinic-pathological disorders. *J Clin Neurosci*. 2009; 16:1131–1135. [PubMed: 19556136]
2. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, Kouri N, Wojtas A, Sengdy P, Hsiung GYR, Karydas A, Seeley WW, Josephs KA, Coppola G, Geschwind DH, Wszolek ZK, Feldman H, Knopman DS, Petersen RC, Miller BL, Dickson DW, Boylan KB, Graff-Radford NR, Rademakers R. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-Linked FTD and ALS. *Neuron*. 2011; 72:245–256. [PubMed: 21944778]
3. Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L, Kalimo H, Paetau A, Abramzon Y, Remes AM, Kaganovich A, Scholz SW, Duckworth J, Ding J, Harmer DW, Hernandez DG, Johnson JO, Mok K, Rytten M, Trabzuni D, Guerreiro RJ, et al. A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron*. 2011; 72:257–268. [PubMed: 21944779]
4. Gendron TF, Bieniek KF, Zhang YJ, Jansen-West K, Ash PEA, Caulfield T, Daugherty L, Dunmore JH, Castanedes-Casey M, Chew J, Cosio DM, van Blitterswijk M, Lee WC, Rademakers R, Boylan KB, Dickson DW, Petrucelli L. 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; 126:829–844. [PubMed: 24129584]
5. Mori K, Arzberger T, Grasser FA, Gijssels I, May S, Rentzsch K, Weng SM, Schludi MH, van der Zee J, Cruts M, Van Broeckhoven C, Kremmer E, Kretschmar HA, Haass C, Edbauer D. Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. *Acta Neuropathol*. 2013; 126:881–893. [PubMed: 24132570]
6. Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, Margolis J, Peterson M, Markowski TW, Ingram MAC, Nan Z, Forster C, Low WC, Schoser B, Somia NV, Clark HB, Schmechel S, Bitterman PB, Gourdon G, Swanson MS, Moseley M, Ranum LPW. Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci U S A*. 2011; 108:260–265. [PubMed: 21173221]
7. Ash PEA, Bieniek KF, Gendron TF, Caulfield T, Lin WL, DeJesus-Hernandez M, van Blitterswijk MM, Jansen-West K, Paul JW III, Rademakers R, Boylan KB, Dickson DW, Petrucelli L. Unconventional translation of C9ORF72 GGGGCC expansion generates Insoluble polypeptides specific to c9FTD/ALS. *Neuron*. 2013; 77:639–646. [PubMed: 23415312]
8. Mori K, Weng SM, Arzberger T, May S, Rentzsch K, Kremmer E, Schmid B, Kretschmar HA, Cruts M, Van Broeckhoven C, Haass C, Edbauer D. The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTL/ALS. *Science*. 2013; 339:1335–1338. [PubMed: 23393093]

9. McMurray C. DNA secondary structure: A common and causative factor for expansion in human disease. *Proc Natl Acad Sci U S A*. 1999; 96:1823–1825. [PubMed: 10051552]
10. Haeusler AR, Donnelly CJ, Periz G, Simko EA, Shaw PG, Kim MS, 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. [PubMed: 24598541]
11. Su Z, Zhang Y, Gendron TF, Bauer PO, Chew J, Yang WY, Fostvedt E, Jansen-West K, Belzil BB, Desaro P, Johnston A, Overstreet K, Oh SY, Todd PK, Berry JD, Cudkowicz ME, Boeve BF, Dickson D, Floeter MK, Traynor BJ, Morelli C, Ratti A, Silani V, Rademakers R, Brown RH, Rothstein JD, Boylan KB, Petrucelli L, Disney M. Discovery of a biomarker and lead small molecules to target r(GGGGCC)-associated defects in c9FTD/ALS. *Neuron*. 2014; 83:1043–1050. [PubMed: 25132468]
12. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Crystallographic characterization of CCG repeats. *Nucleic Acids Res*. 2012; 40:8155–8162. [PubMed: 22718980]
13. Wang AH, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*. 1979; 282:680–686. [PubMed: 514347]
14. Hall K, Cruz P, Tinoco I, Jovin T, van de Sande J. Z-RNA - a left-handed RNA double helix. *Nature*. 1984; 311:584. [PubMed: 6482970]
15. Popenda M, Milecki J, Adamiak RW. High salt solution structure of a left-handed RNA double helix. *Nucleic Acids Res*. 2004; 32:4044. [PubMed: 15292450]
16. Placido D, Brown B, Lowenhaupt K, Rich A, Athanasiadis A. A left-handed RNA double helix bound by the Z-alpha domain of the RNA-editing enzyme ADAR1. *Structure*. 2007; 15:395. [PubMed: 17437712]
17. Moradi M, Babin V, Roland C, Sagui C. Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Res*. 2013; 41:33–43. [PubMed: 23104380]
18. 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; 39:7308–7315. [PubMed: 21596781]
19. 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 - Syndrome. *ChemBioChem*. 2011; 12:2140–2142. [PubMed: 21766409]
20. Amadei A, Linssen AB, Berendsen HJC. Essential Dynamics of Proteins. *Proteins: Struct, Funct, Genet*. 1993; 17:412–425. [PubMed: 8108382]
21. Bouvier B, Grubmuller H. A molecular dynamics study of slow base flipping in DNA using conformational flooding. *Biophys J*. 2007; 93:770–786. [PubMed: 17496048]
22. Pan F, Roland C, Sagui C. Ion distributions around left- and right-handed DNA and RNA duplexes: a comparative study. *Nucleic Acids Res*. 2014; 42:13981–96. [PubMed: 25428372]
23. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003; 31:3406–3415. [PubMed: 12824337]
24. Ellegren H. Microsatellites: Simple sequences with complex evolution. *Nat Rev Genet*. 2004; 5:435–445. [PubMed: 15153996]
25. Oberle I, Rousseau F, Heitz D, Devys D, Zengerling S, Mandel J. Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am J Hum Genet*. 1991; 49:76. [PubMed: 1829582]
26. Campuzano V, Montermini L, Molto M, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Canizares J, Koutnikova H, Bidichandani S, Gellera C, Brice A, Trouillas P, DeMichele G, Filla A, DeFrutos R, Palau F, Patel P, DiDonato S, Mandel J, Coccozza S, Koenig M, Pandolfo M. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*. 1996; 271:1423–1427. [PubMed: 8596916]
27. Mirkin SM. DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol*. 2006; 16:351–358. [PubMed: 16713248]
28. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res*. 2010; 38:8370–8376. [PubMed: 20702420]

29. Tikhomirova A, Beletskaya IV, Chalikian TV. Stability of DNA duplexes containing GG, CC, AA, and TT mismatches. *Biochemistry*. 2006; 45:10563–10571. [PubMed: 16939208]
30. 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; 117:8883–8884.
31. Case, DA., et al. AMBER 14. University of California; San Francisco: 2014.
32. Case, DA., et al. AMBER 12. University of California; San Francisco: 2012.
33. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys J*. 2007; 92:3817–3829. [PubMed: 17351000]
34. Banas P, Hollas D, Zgarbova M, Jurecka P, Orozco M, Cheatham TE, Sponer J, Otyepka M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J Chem Theory Comput*. 2010; 6:3836–3849.
35. Zgarbova M, Otyepka M, Sponer J, Mladek A, Banas P, Cheatham TE, Jurecka 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. [PubMed: 21921995]
36. Besseova I, Banas P, Kuhrova P, Kosinova P, Otyepka M, Sponer J. Simulations of A-RNA Duplexes. The Effect of Sequence, Solute Force Field, Water Model, and Salt Concentration. *J Phys Chem B*. 2012; 116:9899–9916. [PubMed: 22809319]
37. Joung IS, Cheatham TE. Determination of Alkali and Halide Monovalent Ion Parameters for use in Explicitly Solvated Biomolecular Simulations. *J Phys Chem B*. 2008; 112:9020–9041. [PubMed: 18593145]
38. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983; 79:926–935.
39. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *J Chem Phys*. 1995; 103:8577–8593.
40. Allen, MP., Tildesley, DJ. *Computer Simulation of Liquids*. Clarendon Press; Oxford: 1987.
41. Berendsen HJC, Postma JPM, van Gunsteren WF, Di Nola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys*. 1984; 81:3684–3690.
42. Meyer T, Ferrer-Costa C, Pérez A, Rueda M, Bidon-Chanal A, Luque FJ, Laughton CA, Orozco M. Essential Dynamics: A Tool for Efficient Trajectory Compression and Management. *J Chem Theory Comput*. 2006; 2:251–258. [PubMed: 26626512]
43. Roe DR, Cheatham TE III. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput*. 2013; 9:3084–3095. [PubMed: 26583988]

Figure 1.
Structures of duplex configurations. G's are in blue and C's in red.

Figure 2.

PCA distributions of G-G (first column) and C-C (fourth column) mismatched duplexes over the 500 ns to 1 μ s interval projected onto the first three principal components for DNA models (a) and RNA models (b). And the corresponding visualization of the atomic motions involved in the first (blue) and second (red) principal components.

Figure 3. Backbone sugar torsion δ and the glycosyl torsion χ angle distributions for DNA duplexes (left) and RNA duplexes (right) corresponding to G mismatches (top three rows) and C mismatches (bottom three rows).

Figure 4.

Open angle distribution of G mismatching models (left two columns) and C mismatching models (right two columns) in DNA (odd columns) and RNA (even columns) forms. Black and red lines show the distribution of G and C bases, respectively, that participate in Watson–Crick bonding. The green line shows the distribution for the internal mismatched bases (G or C).

Figure 5.

Simulations results for DNA SC-3. (a) DNA SC-3 evolution in the simulation set A from 0 ns (after equilibration) to 1 μ s (Final simulation time). First, one strand shifts by one residue in the 3'-5' direction, and then the free ending bases bond to the Watson-Crick paired bases, causing the breaking of the hydrogen bonds. (b) In the simulation set B, one strand shifts by one residue in the 3'-5' direction, transforming the duplex into a partial DC-1 structure. The FLipping out of a C basis allows the retention of the SC-3 configuration in the second half of the duplex.

Figure 6.

(a) Open angle of DNA DC-1 as a function of time (left) and corresponding distribution (right). (b) The e-motif configuration. (c) Open angle of DNA DC-2 as a function of time (left) and corresponding distribution (right). (d) Evolution of the C6–C19 mismatch (side views) and top view of the C6–C19 evolution.

Figure 7.

(a) Torsion angle α (black) and γ (red) of C7 and C18 as a function of time for DC-1 in DNA form. Sum of torsion angles ($\alpha + \gamma$) for C7 (blue) and C18 (orange) as a function of time for DC-1 in DNA form. (b) Torsion angle ϵ (black) and ζ (red) of C18 and C19 as a function of time for DC-2 in DNA form. Difference of torsion angles ($\epsilon - \zeta$) for C18 (blue) and C19 (orange) as a function of time for DC-2 in DNA form.

Figure 8. Open angle of RNA DC-1 (a) and DC-2 (b) as a function of time (left) and corresponding distribution (right). (c) Configurations A and B corresponding to the average structures of DC-1 at 400–500 and 650–700 ns, and typical hydrogen bonds associated with them.

Figure 9.

(a) Open angle of RNA SC-3 as a function of time (left) and corresponding distribution (right). (b) Typical structures of C10–C16 pair base.

Figure 10.

Average over last 100 ns of twist angles of DNA duplexes (top) and RNA duplexes (bottom) as a function of step numbers for G-rich sequences.

Figure 11.

Typical hydrogen bond of G bases (a) and C bases (b) in mismatched bases. Due to the syn/anti configuration, a G-G pair forms two hydrogen bonds in DNA form, N7–N2 and O6–N1, and forms three hydrogen bonds in RNA form, N7–N2, O6–N1, and OP2–N2. The C-C pair forms either N3–N4 or O2–N4 hydrogen bonds. (c) Presence of hydrogen bonds in only N3–N4 (red), in only O2–N4 (blue), or formed simultaneously (green), as a function of time for DC-1 (top 6) and DC-2 (bottom 6) in DNA (left), and RNA (right).

Figure 12.

Atomistic details of direct binding sites of Na⁺ for different mismatches. (a) G-G mismatch in DG models, (b) G-C base pair adjacent to a G-G mismatch, (c) C-C mismatch where the bases stay inside the helix and form N3–N4 H-bonds, (d) C-C mismatches where the bases form O2–N4 H-bonds, and (e) G-C base pair adjacent to a C-C mismatch can also contribute to trap the ion. In addition, the figure shows two examples of ions binding to the helix backbone. (f) Ion bridging four OP1 atoms in DNA DC-1; (g) ion linking the ends of a DG-2 RNA helix.

Table 1

Characterization of the Initial Conformations

structure	model name	sequence	arrangement	form
duplex	DG-1	(GGGCCG) ₂	antiparallel	DNA, RNA
	DG-2	(GCCGGG) ₂	antiparallel	DNA, RNA
	SG-3	H-(GGGGCC) ₂	antiparallel	DNA, RNA
	DC-1	(CCCGGC) ₂	antiparallel	DNA, RNA
	DC-2	(CGGCCC) ₂	antiparallel	DNA, RNA
	SC-3	H-(CCCCGG) ₂	antiparallel	DNA, RNA
duplex (special)	R-SC-3	H-(CCCCGG) ₂	antiparallel	DNA
	SC-3 $\neq$ C	(CCCCGG) ₂ -C	antiparallel	DNA