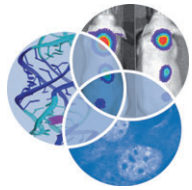
Overview



Structures of RNA repeats associated with neurological diseases

Leszek Błaszczyk, Wojciech Rypniewski and Agnieszka Kiliszek*

All RNA molecules possess a 'propensity' to fold into complex secondary and tertiary structures. Although they are composed of only four types of nucleotides, they show an enormous structural richness which reflects their diverse functions in the cell. However, in some cases the folding of RNA can have deleterious consequences. Aberrantly expanded, repeated RNA sequences can exhibit gain-of-function abnormalities and become pathogenic, giving rise to many incurable neurological diseases. Most RNA repeats form long hairpin structures whose stem consists of noncanonical base pairs interspersed among Watson–Crick pairs. The expanded hairpins have an ability to sequester important proteins and form insoluble nuclear foci. The RNA pathology, common to many repeat disorders, has drawn attention to the structures of the RNA repeats. In this review, we summarize secondary structure probing and crystallographic studies of disease-related RNA repeat sequences. We discuss the unique structural features which can contribute to the pathogenic properties of the repeated runs. In addition, we present the newest reports concerning structural data linked to therapeutic approaches.

How to cite this article:
WIREs RNA 2017, 8:e1412. doi: 10.1002/wrna.1412

MICROSATELLITES: TINY REPEATS WITH LARGE-SCALE EFFECTS

Microsatellites are tandem repeated tracts of 1–6 nucleotides, known also as 'Short Tandem Repeats' (STR). They are a specific type of repetitive DNA which constitutes of a major part of genomes (ca 50% in humans).[1] Microsatellite DNA is ubiquitous in Procaryota and Eucaryota and is found in coding and noncoding regions of their genomes.[2,3] A characteristic feature of microsatellites is their instability which leads to length polymorphism of the repeated tracts. It is thought that they are one of the sources of genetic variation that drives genome evolution.[4,5] Although the function of microsatellites is not well understood, they possess a 'dark side' which was revealed in recent years. They are causative agents in the development of incurable repeat expansion disorders, such as Huntington's disease (HD), fragile X-associated tremor ataxia syndrome (FXTAS), myotonic dystrophies (MD), spinocerebellar ataxias (SCA), and many others.[6–9] A majority of microsatellites associated with diseases are trinucleotide CNG repeats (N is one of the four natural nucleotides) but also tetra-, penta-, and hexanucleotide repeats belong to this group.[6,7] Although each disease shows characteristic symptoms they all start when an abnormal amplification of repeated units exceeds a specific threshold.[10,11] The number of expanded repeats correlates with disease symptoms: the more repeats, the earlier the disease appears and the greater its severity. Pathologies of repeat associated disorders are diverse and depend on the localization of repeated runs in the gene. If the expansion occurs in an open reading frame the repeats are translated into toxic proteins containing homopolymeric tracts of amino acids (glutamine or alanine). This is observed in

*Correspondence to: kiliszek@ibch.poznan.pl

Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

Conflict of interest: The authors have declared no conflicts of interest for this article.

polyQ diseases such as HD and several SCAs.[12–14] This pathologic protein gain-of-function can affect specific metabolic pathways as well as induce toxic protein aggregates which lead to cell dysfunction. When expanded tracts of microsatellites are present in the noncoding parts of genes (5′ or 3′ untranslated regions or introns), the RNA is toxic. In transcripts, repeated runs fold into defined secondary structures (mostly long hairpins) gaining new functions. They sequester important cellular factors and form insoluble foci (e.g., MD type 1 and 2, HD-like 2).[7,15,16] If sequestration concerns splicing factors such as MBNL1 (Muscleblind-like protein 1) aberrantly spliced transcripts are produced.[6,7,17] Recently, other coexisting pathways were discovered: repeat-associated non-AUG (RAN) translation and bidirectional transcription which added complexity to the existing protein and RNA gain-of-function mechanisms. RAN translation is a noncanonical translation initiation which occurs without the need of the AUG codon.[18,19] As a result, homopolymeric peptides are translated that can exhibit toxicity, similar to polyQ diseases. It is suggested that one of the triggers of RAN translation is the structure formed by expanded repeats.[19] Bidirectional transcription of repeated regions results in antisense RNA which can hybridize to sense RNA and form double-stranded structures. They can be processed into small interfering RNA (siRNA), activating a silencing mechanism.[20,21]

Although many pathogenic pathways can occur simultaneously, still the RNA and protein-mediated pathogenesis are believed to be the major mechanisms. The RNA pathology is common to many repeat associated disorders which has drawn attention to the study of their structures. Earliest work was focused on probing the secondary structure, but in recent years also crystallography and NMR have been employed. The structural features of expanded RNA repeats are unique among RNA structures found within the cell. Most of the repeats form hairpins whose major part is a stem formed by many repeated blocks consisting of Watson–Crick pairs punctuated by noncanonical base pairs (Figure 1(a)). These systems seem simple, but are able in fact to sequester a variety of proteins. Moreover, the three-dimensional structures of RNA repeats display unique structural features that add to our general structural knowledge concerning RNA molecules. Thus, the structures of repeats reflect and extend the structural richness of the RNA world.

SECONDARY STRUCTURE OF RNA REPEATS

Determination of secondary structures formed by RNA repeats is an important step in understanding the roles of toxic RNAs in the development of neurodegenerative disorders. The folding and structural stability of repeated runs has been investigated with a range of probing techniques including classic (nucleases, Pb^{2+} ions) and modern (SHAPE) approaches as well as biophysical methods [ultra-violet (UV) melting, circular dichroism (CD) spectroscopy]. Although the secondary

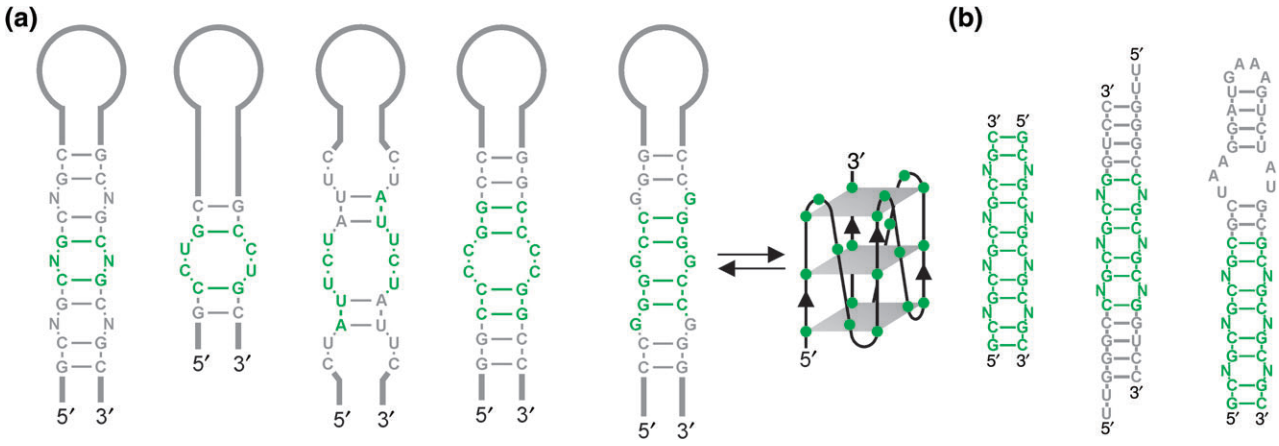


FIGURE 1 | (a) Base pair arrangements in RNA repeats (green) used in structural studies. Hairpin stems consist of noncanonical base pairs flanked by canonical Watson–Crick pairs. In the case of CNG repeats N denotes one of the four natural nucleotides (A, C, U, or G). The GGGGCC repeats show an alternative quadruplex structure. (b) Constructs used in crystallographic studies. In most cases, the crystallized RNA were oligomers composed of pure repeats (left). To facilitate crystallization, flanking sequences (gray) have been added in some cases (middle) or RNA repeats were crystallized as a part of GAAA tetraloop/receptor (right, gray). Secondary structures were generated using RNA structure and VARNA software.[22,23]

structure of RNA repeats has been studied extensively, most of the results describe RNA molecules having a number of repeats within their normal range and surrounded by relatively short flanking sequences.

Structure of CNG Repeats

Structural analyses of pure CNG runs composed of up to 20 triplets revealed that all of them form relatively stable hairpin structures. The stems of such hairpins consist of G-C and C-G Watson–Crick base pairs separated by noncanonical N-N pairs[24,25] (Figure 1(a)). Thermodynamic studies have shown that CNG hairpins have comparable stabilities, with CGG being the most stable, followed by CAG, CUG, and CCG repeats.[25,26] It turned out that RNA composed of pure CAG, CUG, and CCG, but not CGG, repeats are prone to form slippery structures resulting in alternative alignments of the hairpins which, was evident from nuclease mapping and electrophoresis under nondenaturing conditions.[24] This microheterogeneity can be overcome by adding artificial G-C clamps at the hairpin base. This modification helped to establish that the only difference in the hairpin structure formed by the investigated CNGs was the hairpin loop size (from 3 nt for CUG or CGG repeats to 7 nt for CAG or CCG repeats).[24] Moreover, the loop size varied according to whether the number of CNG repeats is odd or even. This suggested that the structure and stability of CNG repeats could be affected not only by their sequence but also by natural flanking sequences. In this view, structural studies performed on RNA oligomers composed of pure CNG repeats were insufficient to address completely their potential role in pathogenesis.

In vitro secondary structure probing of CNG repeats situated in the context of natural flanking regions was performed for the majority of disease-related transcripts (Table 1). The results identified not only factors influencing the structure and stability of CNG repeats but also characterized for the first time the folding of RNA having pathological lengths.[27–30] In many cases, the CNG hairpin stability is influenced by direct pairing of 5′ and 3′ flanking sequences. This interaction leads also to the structural isolation of the CNG hairpin from its surroundings. For example, hairpins formed by a normal number of CAG repeats in CACNA1A and ATXN1 transcripts implicated in SCA6 and SCA1, respectively, are stabilized by G-C-rich flanking sequences which increase the length of the hairpin stem (Figure 2(a) and (b)).[31,32] Importantly, the same interaction is formed in the ATXN1 transcript containing a pathogenic number of CAG repeats.[32] Flanking sequences can also interact with CNG runs, which was observed for CGG repeats in FMR1 transcript related to FXTAS.[33] In this case, the base of an CGG hairpin is stabilized *via* an interaction of the 5′ part of the repeated region and the downstream natural sequence (Figure 2(c)).

An interesting type of interaction between repeated runs and flanking sequences occurs in AR (androgen receptor) and HTT (huntingtin) transcripts (implicated in spinal and bulbar muscular atrophy and HD, respectively) which beside potentially pathogenic CAG repeats contain three adjacent CUG (AR) or seven CCG (HTT) runs (Figure 2(d) and (e)).[34] In the HTT mRNA, the expandable region is separated from downstream CCG repeats by 12 nucleotides of a natural sequence. Secondary structure probing of HTT transcripts containing the normal or mutated number of CAG repeats revealed direct interaction between CAG and CCG repeats within the base of the hairpin stem (Figure 2(e)). The CAG/CGG interaction forms a double-stranded region with noncanonical A-C pairs. This is followed by a region of interaction between CAG repeats and the 12 nt of natural sequence, and the apical part of the hairpin composed of only CAG repeats.[34] In AR mRNA CUG repeats, which directly precede the expandable region, base pair with the last three CAG repeats forming an ideal double-stranded region at the hairpin base, also in the mutated transcript (Figure 2(d)).[34]

In contrast to the aforementioned examples, in some transcripts flanking sequences revealed no contribution to the folding of CNG hairpin structure and their stability. Chemical and enzymatic probing of CUG tracts in DMPK mRNA (related to myotonic dystrophy type 1), and CAG tracts in ATXN2 and ATN1 transcripts (related to SCA2 and dentatorubral-pallidoluysian atrophy, respectively) revealed that sequences adjacent to CNG repeats are folded into separate autonomous structures or remain single-stranded (Figure 3(a)–(c)).[31,35,36] This is probably because those regions are rather AU-rich which may not interact with the repeated, GC-rich regions. As a result, slippery hairpins are formed even when transcripts contain an expanded number of repeats.[31,35,36] In one case (ATXN3 mRNA related to SCA3), although base pairing between CAG repeats and 5′ flanking sequence was observed it did not increase the stability of the hairpin since slippery structures formed even when 65 CAG repeats were present in the RNA molecule (Figure 3(d)).[31]

Normally, a number of genes associated with neurodegenerative disorders contain specific

TABLE 1 | RNA Repeats Used for Secondary Structure Probing Experiments

Repeat type	Gene	Localization	Number of repeats used in secondary structure probing experiments	Effect of flanking sequences	Effect of interruptions	Disease	Refs.
CAG	*CACNA1A*	Coding region	4–14	Stabilization (pairing)	—	Spinocerebellar ataxia 6 (SCA6)	31
CAG	*ATXN1*	Coding region	53–54 (pure) 27–34 (CAU interrupted)	Stabilization (pairing)	CAU interruptions induce presence of large apical loop, internal loops, or branched structures	Spinocerebellar ataxia 1 (SCA1)	32
CAG	*AR*	Coding region	9–46	Stabilization (pairing). CUG repeats present in 5′ flanking sequence	—	Spinal and bulbar muscular atrophy (SBMA)	34
CAG	*HTT*	Coding region	17–70	Stabilization (pairing). CCG repeats present in 3′ flanking sequence	—	Huntington's disease (HD)	34,37
CAG	*ATXN2*	Coding region	36–37 (pure) 14–29 (CAA interrupted)	Slippery structures (no interaction)	CAA interruptions induce presence of branched structures	Spinocerebellar ataxia 2 (SCA2)	36
CAG	*ATN1*	Coding region	6–15	Slippery structures (no interaction)	—	Dentatorubral-pallidoluysian atrophy (DRPLA)	31
CAG	*ATXN3*	Coding region	15–65	Slippery structures (no interaction)	—	Spinocerebellar ataxia 3 (SCA3)	31
CGG	*FMR1*	5′UTR	19–97 (pure) 23–47 (AGG interrupted)	Slippery structures (no interaction)	AGG interruptions induce presence of large apical loop or branched structures	Fragile X-associated tremor ataxia syndrome (FXTAS)	28,33
CUG	*DMPK*	3′UTR	11–140	Slippery structures (no interaction)	—	Myotonic dystrophy type 1 (DM1)	30,35
CCUG	*ZNF9*	Intron	14–17	Slippery structures (no interaction)	—	Myotonic dystrophy type 2 (DM2)	24
AUUCU	*ATXN10*	Intron	9–17	Unknown	—	Spinocerebellar ataxia 10 (SCA10)	38
GGGGCC	*C9orf72*	Intron	4–8	Unknown	—	Amyotrophic lateral sclerosis/ frontotemporal dementia (ALS/FTD)	39,40

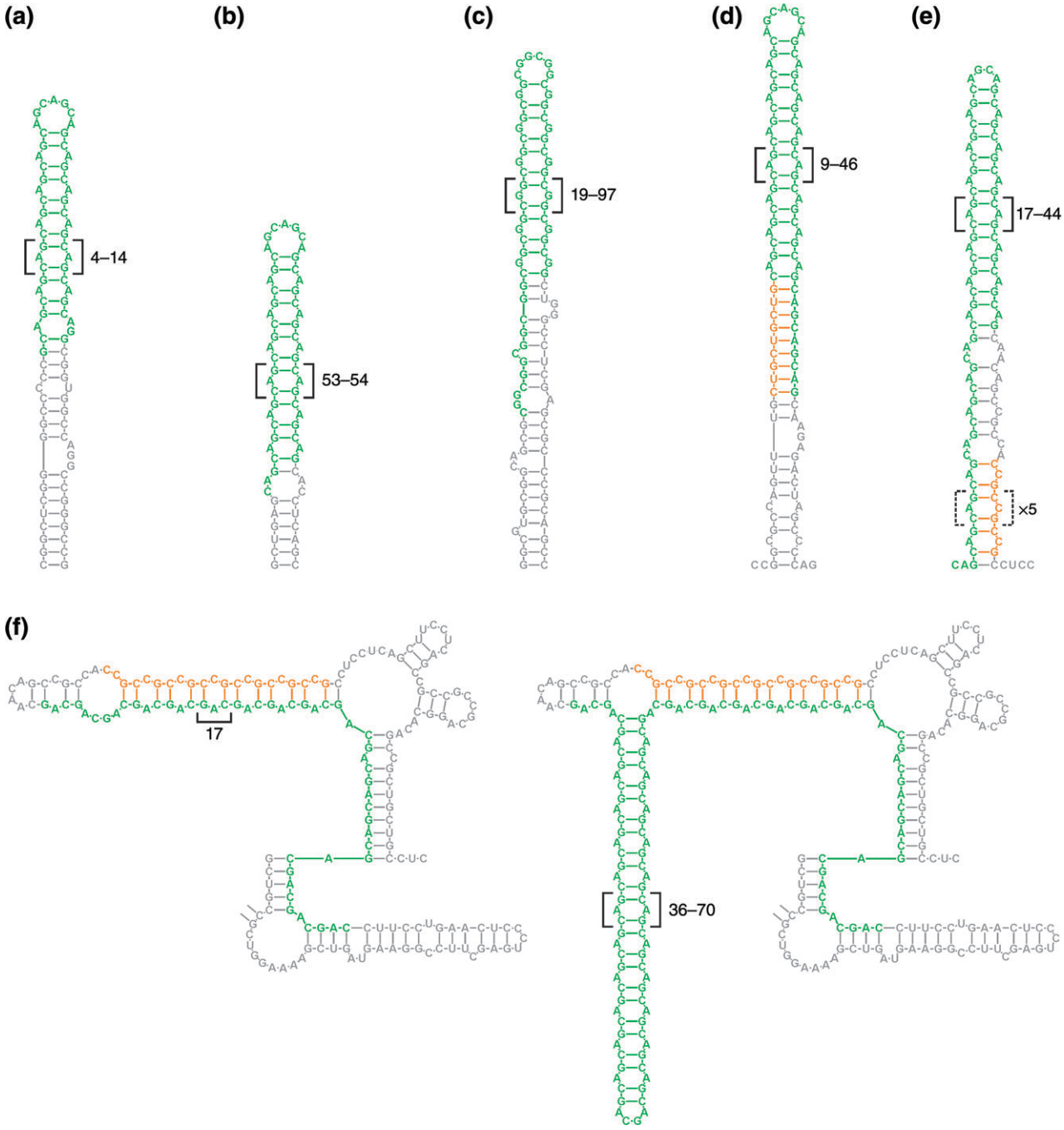


FIGURE 2 | Stabilization effect of flanking sequences on hairpin structures formed by CNG repeats. (a) CAG repeats in CACNA1A mRNA,[31] (b) CAG repeats in ATXN1 mRNA,[32] (c) CGG repeats in FMR1 mRNA,[33] (d) CAG repeats in AR mRNA,[34] (e and f) CAG repeats in HTT mRNA. Three secondary structures are shown: (e) Reprinted with permission from Ref 34. Copyright 2011 Oxford University Press (f) Reprinted with permission from Ref 37. Copyright 2013 Oxford University Press. CNG repeats are depicted in green while flanking sequences in gray. Adjacent CUG repeats in AR mRNA and CCG repeats in HTT mRNA are orange. Brackets denote the number of CNG repeats used in secondary structure probing experiments. Secondary structures were generated using RNAstructure and VARNA software.[22,23]

interruptions (single nucleotide substitutions) located in CNG repeats regions. Importantly, mutated alleles are deprived of those interruptions forming long homogenous tracts. Structural studies of the role of the interruptions were conducted for ATXN1 (CAG repeats with CAU interruptions), ATXN2 (CAG repeats with CAA interruptions), and FMR1 (CGG repeats with AGG interruptions) transcripts.[32,33,36]

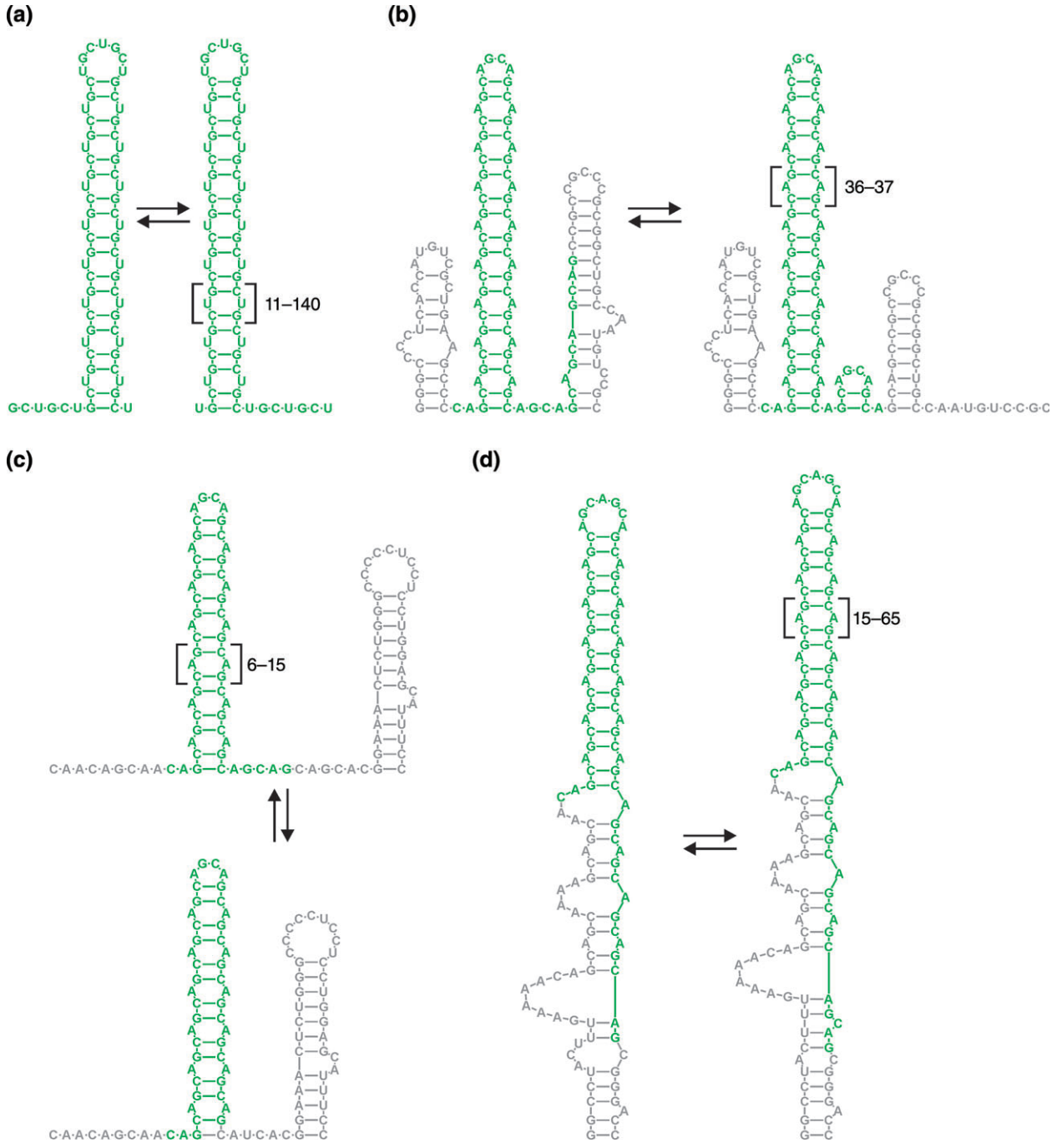


FIGURE 3 | In several CNG-containing transcripts, flanking sequences revealed no contribution to folding and stability of the hairpin structure, and the structures show strand slippage. (a) CUG repeats in DMPK mRNA,[35] (b) CAG repeats in ATXN2 mRNA,[36] (c) CAG repeats in ATN1 mRNA,[31] (d) CAG repeats in ATXN3 mRNA.[31] In each panel, alternative structural arrangement of CNG hairpin is shown. CNG repeats are depicted in green while flanking sequences in gray. Brackets denote the number of CNG repeats used in secondary structure probing experiments. Secondary structures were generated using RNAstructure and VARNA software.[22,23]

It turned out that the presence of interruptions in normal alleles had a profound effect on the structure of repeated runs. The presence of even a single interruption leads to perturbations of CNG hairpin folding. Depending on the location and number of interruptions they can introduce an enlargement of the apical loop, incorporation of internal loops along the hairpin stem and formation of branched (Y-shaped) structures composed of shorter hairpins (Figure 4(a)-(c)). It was

suggested that the destabilization effect of interruptions may prevent the formation of long single hairpins in some premutation carriers and thus attenuate disease symptoms.[33]

Folding of the full length mRNAs *in vivo* can be modulated by environmental conditions and proteins. The information gathered from secondary structure probing of CNG repeats suggests that at least in some transcripts flanking sequences can influence the structure and stability of CNG hairpins. However, the secondary structures of the majority of CNG repeats were determined only in *in vitro* conditions and for transcripts bearing relatively short flanking sequences. Their folding in full length transcripts remains unknown. The only exception is the structure of 15 CAG repeats present in the full length ATXN3

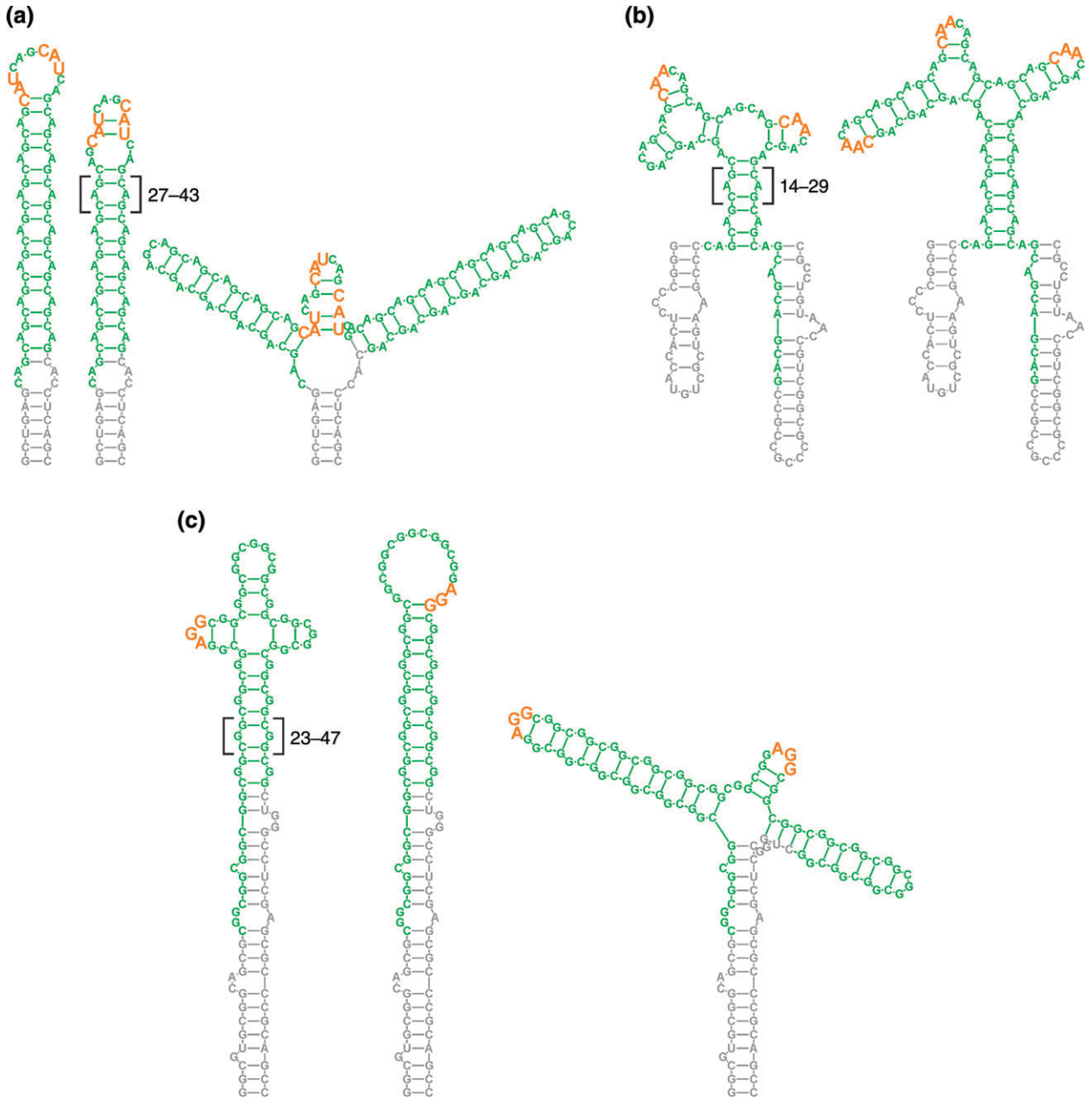


FIGURE 4 | Effect of interruptions on the structure of CNG repeats. (a) CAG repeats in ATXN1 mRNA,[32] (b) CAG repeats in ATXN2 mRNA,[36] (c) CGG repeats in FMR1 mRNA.[33] The CNG repeats are depicted in green, flanking sequences in gray. Orange letters represent the interruptions. Brackets denote the number of CNG repeats used in secondary structure probing experiments. Secondary structures were generated using RNA structure and VARNA software.[22,23]

mRNA. The study confirmed the presence of a hairpin structure similar to shorter RNA transcripts.[31] In another study, a nuclease mapping was performed of CAG repeats (with CAU interruptions) in ATXN1 mRNA in whole-cell extract from human fibroblasts.[32] It turned out that its secondary structure closely resembled the structure obtained under *in vitro* conditions, which suggested that similar structures of CAG and perhaps other CNG repeats can be formed *in vivo*. On the other hand, a recent study from the Weeks laboratory determined that flanking regions had a greater impact on CNG hairpin structure than previously anticipated.[37] Using the SHAPE approach they investigated the secondary structure of normal and mutated CAG repeats from HTT mRNA in the context of long natural sequences (entire exon 1 and c.a. 100 nucleotides of 3′ flanking sequence, including CCG polyproline repeats). In transcripts containing the normal number of repeats (17 or 23) the CAG hairpin was absent or very short (Figure 2(f)). This was due to extensive pairing of CAG repeats with the 3′ flanking region. The CAG hairpin became evident only when the mutant number of CAG repeats was introduced into the transcript (Figure 2(f)). This suggest a potential role of allele-specific drugs with an ability to discriminate between healthy and mutated transcripts. Moreover, it emphasized a need for structural studies of CNG and other repeats in the context of their full length mRNAs.

Secondary Structure of Other RNA Repeats Associated with Diseases

Besides the extensively studied trinucleotide repeats, longer microsatellites are also implicated in neurodegenerative diseases. However, their structural studies are currently limited to a small number of examples. Tetranucleotide CCUG repeats are located in intron 1 of ZNF9 pre-mRNA and their expansion (usually above 40 repeats) is implicated in myotonic dystrophy type 2.[17] The secondary structure of CCUG repeats has been analyzed only for transcripts containing less than 20 repeats and without any flanking sequences.[24] The CCUG repeats fold into hairpins very similar to those of CNG repeats (Figure 1(a)). The difference is the presence of two noncanonical C-U pairs along the hairpin stem and a larger (6 or 10 nucleotides) apical loop. This is manifested by lower predicted thermodynamic stability of the CCUG hairpin than of the CUG repeats.[24] Similar to CAG, CUG, and CCG repeats, pure CCUG repeats form slippery hairpins. The pentameric AUUCU repeats are located in intron 9 of ATXN10 pre-mRNA and are associated with spinocerebellar ataxia type 10 (SCA10).[41] The secondary structure of AUUCU repeats (up to 17 repeats) was investigated using S1 and V1 nucleases as well as CD spectroscopy and NMR.[38] The pattern of S1 nuclease cleavages and NMR analysis indicates that at low temperature (20°C) the AUUCU repeats are folded into a hairpin structure (Figure 1(a)). The stem of the hairpin is composed of a symmetric 3-nucleotide internal loop 5′-UCU-3′/3′-UCU-5′ separated by two A-U pairs. However, the AUUCU hairpin seems to be unstable since the pattern of nuclease cleavages and CD spectra revealed that AUUCU repeats are unstructured at temperatures above 37°C. Hexanucleotide repeats GGGGCC are present in intron 1 of C9orf72 pre-mRNA and normal alleles contain less than 25 repeats (usually 2). Large expansion (up to 1600 repeats) of the GGGGCC repeats has been linked to ALS (amyotrophic lateral sclerosis) and FTD (frontotemporal dementia) *via* the formation of nuclear foci and induction of RAN translation.[42] An initial study showed that GGGGCC runs formed highly stable quadruplexes.[43] Next, the secondary structure of GGGGCC repeats was investigated and revealed that they formed a mixture of hairpins and quadruplexes. Based on DMS and T1 nuclease mapping it was proposed that the repeated unit in hairpin stem formed two G-C and two C-G base pairs separated by noncanonical G-G pairs[39,40] (Figure 1(a)). The apical loop of GGGGCC hairpin contains seven nucleotides. The formation of the hairpin was promoted by low annealing temperature (heating at 37°C followed by slow cooling and equilibration at room temperature) or lack of KCl. However, in the presence of KCl and higher annealing temperature a quadruplex is a preferable structure.

WHAT WE HAVE LEARNED FROM CRYSTALLOGRAPHIC AND NMR STUDIES?

In the last decade a number of crystal structures of RNA associated with repeat expansion disorders has been solved (Table 2). In most cases the sequences of the crystallized RNA contained several repeats. They formed self-complementary duplexes representing the hairpin stem (Figure 1(b)). In some cases the crystallized oligomer had additional flanking sequences to facilitate crystallization or was attached to the hairpin of GAAA tetraloop/receptor (Figure 1(b)). All the obtained crystallographic structures of repeats formed helices showing A-RNA character. Interestingly, the CCG, CCUG, and CCCCGG repeats formed slippery duplexes with dangling nucleotides at one end of each strand. According to the secondary structure probing the helices formed by CNG

TABLE 2 | Crystal Structures of RNA Repeats

PDB	Type of repeat	Number of repeats	Additional sequence	Ending of the helix	Type of noncanonical base pair according to Leontis–Westhof nomenclature	N-N pairing interaction	C1′-C1′ distance for N-N pairs (Å)	C1′-C1′ distance for Watson–Crick pairs (Å)	Twist (°)	Rise (Å)	Resolution (Å)	*R*/R_{free} (%)	Refs
1ZEV	CUG	6	No	Blunt ended	Unclear due to ambiguous map	?	?	?	~34	~3	1.58	21.8/27.9	[44]
3GLP	CUG	2	No	Blunt ended	U•U cWW	N3-H3···O4	10.5	10.6	33.4	2.7	1.23	14.7/18.4	[45]
3GM7	CUG	6	No	Blunt ended	U•U cWW	N3-H3···O4	10.3	10.5	33.6	2.6	1.58	21.9/26.2	[45]
3SYW	CUG	3	5′ and 3′ flanking sequences	5′ dangling ends	U•U cWW (wobble)	N3-H3···O4 and N3-H3···O2	8.8	10.6	30.5	2.6	1.57	16.7/18.7	[46]
					U•U cWW	None	9.9						
3SZX	CUG	3	5′ and 3′ flanking sequences	5′ dangling ends	U•U cWW	N3-H3···O4	10.5	10.5	32.8	2.6	2.20	22.2/27.1	[46]
					U•U cWW	None	10.3						
4E48	CUG	6	No	Blunt ended	U•U cWW	N3-H3···O4	10.5	10.7	32.8	2.6	2.50	20.3/27.9	[47]
					U•U cWW	None	9.7						
4FNJ	CUG	2	GAAA tetraloop/ receptor	Blunt ended	U•U cWW	N3-H3···O4	10.6	10.4	36.3	2.3	1.95	20.8/26.6	[48]
					U•U cWW (wobble)	N3-H3···O4 and N3-H3···O2	8.6						
3NJ6	CAG	2	No	Blunt ended	A•A cWW (wobble)	C2-H2···N1	11.1	10.7	28.9	2.8	0.95	10.6/N/A	[49]
3NJ7	CAG	2	No	Blunt ended	A•A cWW (wobble)	C2-H2···N1	11.0	10.7	28.0	2.9	1.90	21.2/24.8	[49]
4J50	CAG	3	5′ and 3′ flanking sequences	5′ dangling ends	Unclear due to static or dynamic disorder	?	?	10.7	28.9	2.6	1.65	16.9/18	[50]

(*continued overleaf*)

TABLE 2 | Continued

PDB	Type of repeat	Number of repeats	Additional sequence	Ending of the helix	Type of noncanonical base pair according to Leontis–Westhof nomenclature	N-N pairing interaction	C1′-C1′ distance for N-N pairs (Å)	C1′-C1′ distance for Watson–Crick pairs (Å)	Twist (°)	Rise (Å)	Resolution (Å)	R/R_{free} (%)	Refs
4YN6	CAG	3	5′ and 3′ flanking sequences	5′ dangling ends	Unclear due to static or dynamic disorder	?	?	10.7	30.1	2.6	2.3	21.7/26.1	[51]
3R1C	CGG	2	No	Blunt ended	G•G cWH	N1H···O6 and *exo*-N2H···N7	11.3	10.7	30.6	3.1	2.05	21.6/25.7	[52]
3R1D	CGG	3	No	Blunt ended	G•G cWH	and intra-molecular *exo*-N2H···O2	11.3	10.8	30.0	2.9	1.45	23.2/27	[52]
3R1E	CGG	2	No	Blunt ended	G•G cWH		11.4	10.7	32.2	2.8	0.97	13.7/N/A	[52]
3SJ2	CGG	3	5′ and 3′ flanking sequences	5′ dangling ends	G•G cWH		11.3	10.8	31.4	2.9	1.36	15.4/18.5	[53]
4E58	CCG	2	No	3′ dangling ends	C•C cWW	*exo*-N4H···N3	10.8	10.6	31.2	3.0	1.95	25.8/30.1	[54]
					C•C cWW	None	10.7						
4E59	CCG	2	No	5′ dangling ends	C•C cWW	*exo*-N4H···N3	10.7	10.6	33.9	2.6	1.54	25.5/30.3	[54]
5EW4	CCCCGG	3	No	5′ dangling ends	C•C cWW	Conformational variability of C-C pairs	10.0	10.6	31.7	2.6	1.47	21.5/23.9	[55]
5EW7	CCCCGG	3	No	5′ dangling ends	C•C cWW	Conformational variability of C-C pairs	10.0	10.6	31.5	2.6	1.75	23.5/25.7	[55]
4 K27	CCUG	3	GAAA tetraloop/ receptor	Blunt ended	C•U cWW	Possible protonation or tautomerisation and H-bonding as in 4XW0, 4XW1 below	8.5	10.6	32.1	2.6	2.35	19.4/24	[56]
					C•U cWW	None	11.7						

TABLE 2 | Continued

PDB	Type of repeat	Number of repeats	Additional sequence	Ending of the helix	Type of noncanonical base pair according to Leontis–Westhof nomenclature	N-N pairing interaction	C1′-C1′ distance for N-N pairs (Å)	C1′-C1′ distance for Watson–Crick pairs (Å)	Twist (°)	Rise (Å)	Resolution (Å)	*R*/R_{free} (%)	Refs
4XW0	CCUG	2	No	3′ dangling ends	C•U cWW	Three H bonds (protonation or tautomerisation): between *exo*-N4 and O4, N3 and N3, O2, and O2 atoms	8.4	10.7	31.4	2.2	1.81	22.4/29.8	[57]
4XW1	CCUG	2	No	3′ dangling ends	C•U cWW	Three H bonds (protonation or tautomerisation): between *exo*-N4 and O4, N3 and N3, O2, and O2 atoms	8.4	10.6	32.1	2.3	2.3	19.1/21.3	[57]
5BTM	AUUCU	2	GAAA tetraloop/ receptor	Unwound ends	U•U cWW (wobble)	N3-H3···O4 and N3-H3···O2	8.7	10.4	NA	NA	2.78	17.6/22.4	[58]
					C•C cWW	N3-H3···O2	7.8		NA	NA			
4PCJ	CUG	2	GAAA tetraloop/ receptor	Blunt ended	U•Ψ cWW	N3-H3···O4	11.0	10.5	36.6	2.2	1.9	19.6/26.8	[59]
5EME	rCAG/pCTG (RNA/PNA)	2	No	Blunt ended	None (only Watson–Crick base pairs)	—	—	10.6	25.5	2.4	1.15	13.9/17.7	[60]
5EMF	rCUG/pCAG (RNA/PNA)	2	No	Blunt ended	None (only Watson–Crick base pairs)	—	—	10.6	25.4	2.4	1.14	12.2/15.3	[60]
5EMG	pCTG (PNA)	2	No	Blunt ended	T•T cWW (wobble)	N3-H3···O4 and N3-H3···O2	9.0	10.7	19.7	3.9	1.06	16.3/19.3	[60]

cWW, *cis* Watson–Crick/Watson–Crick pair.
Helical parameters were calculated using 3DNA software. Reprinted with permission from Ref 61. Copyright 2009 Wiley VCH.

repeats have C-G and G-C pairs interrupted by single noncanonical base pairs. The CCUG repeats have two C-U and U-C noncanonical base pairs. In the case of CCCCGG repeats (the antisense sequence of GGGGCC repeats in C9orf72 pre-mRNA) the duplex has two C-C pairs flanked by two C-G and two G-C pairs (Figure 1(a)).

Noncanonical Base Pairs of Repeated Runs

The most interesting part of the RNA runs are the noncanonical N-N pairs: the factor differentiating the features of RNA repeats structures. Except for RNA containing AUUCU repeats, all N-N pairs are located in a specific structural context. They are surrounded by stable Watson–Crick C-G and G-C pairs which dominate the structure. The noncanonical pairs consist of the same type of nucleotide residues, namely either two interacting pyrimidines or two larger purines.

Noncanonical U-U Pairs

The CUG repeats are the most studied type of repeated sequences. In the PDB repository there are seven different crystal structures (see Table 2).[44–48] Three of them contain pure CUG repeats while the others have additional sequences to facilitate crystallization.

Most of the U-U pairs form one hydrogen bond between the N3 atom of one uridine residue and the O4 carbonyl atom of the second U, which is inclined toward the minor groove (Figure 5(b)) (PDB code: 3GLP, 3GM7, 3SZX, 4E48, 4FNJ).[45–48] The degree of inclination is indicated by the λ angle determined between the line connecting the C1′-C1′ atoms and the N-glycosidic bond (Figure 5(a)). For Watson–Crick base pairs, the λ angle is around 55°, while for the inclined uridine it is only 31°. The distance between the C1′ atoms of the paired uridines is 10.5 Å, similar to that of canonical base pairs (10.6 Å). Owing to the specific conformation of the inclined uridine this type of noncanonical pair was named stretched U-U wobble. According to Leontis–Westhof nomenclature this would be defined as *cis* Watson–Crick/Watson–Crick pair (U-U cWW).[62,63] Such pairing is unique among the other known RNA structures. Only tRNA-Gln and one pair in 16S rRNA shows the stretched U-U wobble while in other RNA structures the uridine residues form two hydrogen bonds and are only 8.6 Å apart (C1′-C1′ distance).[45]

In the structures of CUG repeats, two other conformations of U-U pairs are observed. In one, the uridine residues are aligned *vis-à-vis* and do not form any H-bonds (five examples) (Figure 5(c)). This conformation was named symmetric H-nonbonded U-U cWW pair (PDB code: 3SYW, 3SZX, 4E48).[46,47] It is suggested that this conformation is an intermediate state between the two alternative conformations of the stretched U-U pair. Perhaps the inclination of uridine is not fixed and the U residue can swap between inclined and noninclined states. The second alternative conformation is the common U-U cWW pair with two hydrogen bonds (three examples) (Figure 5(d)) (PDB code: 3SYW, 4FNJ). Although the U-U pairs can adopt different conformations, the stretched U-U wobble is predominantly observed within the CUG repeats structures (20 of 28 observed unique pairs).[45–48] This is also confirmed by NMR studies followed by molecular dynamics calculations where 76.5% species exhibited one H-bond.[46]

Noncanonical A-A Pairs

The CAG repeats are represented by four crystallographic models.[49–51] Two of them are structures of pure repeats while the other two contain flanking sequences.

In the two native models (PDB code: 3NJ6, 3NJ7), where one is an atomic resolution structure (0.95 Å), both adenosine residues are in *anti* conformation and form one weak H-bond between C-H···N atoms (A-A cWW pair) (Figure 6(a)).[49] This type of interaction is weak because a carbon atom is a poorer proton donor than oxygen or nitrogen atoms which form most inter- and intramolecular H-bonding interactions. The C1′-C1′ distance between the adenine residues is 11 Å. Both adenosines are inclined toward the major groove. The λ values are 64° and 87°.

In two other independent studies, authors crystallized the same oligomer containing three CAG repeats with additional flanking sequences on both sides of the duplex (Figure 6) (PDB code: 4 J50, 4YN6).[50,51] In the first study the authors observed that the two closing A-A pairs located at the ends of the duplex show a *syn–anti* conformation. According to Leontis–Westhof nomenclature this would be defined as *cis* Watson–Crick/Hoogsteen pair (A-A cWH). The A-A pairs form one H-bond between the *exo*-amino group of A(*syn*) and N1 atom of A(*anti*) (Figure 6(b) and (f)).[50] Moreover, the N1 atom of A (*syn*) probably interacts with the 2′OH group of a disordered uridine residue of the flanking sequence that is tucked in the major groove. The second group interpreted nearly the same electron density map in a different way. They concluded that one of the closing A-A pairs showed similar *syn–anti* conformation while the second one had *anti-anti* conformation with no H-bonds (A-A cWW pair) (Figure 6(c) and (g)).[51] In both models a noncanonical pair located in the

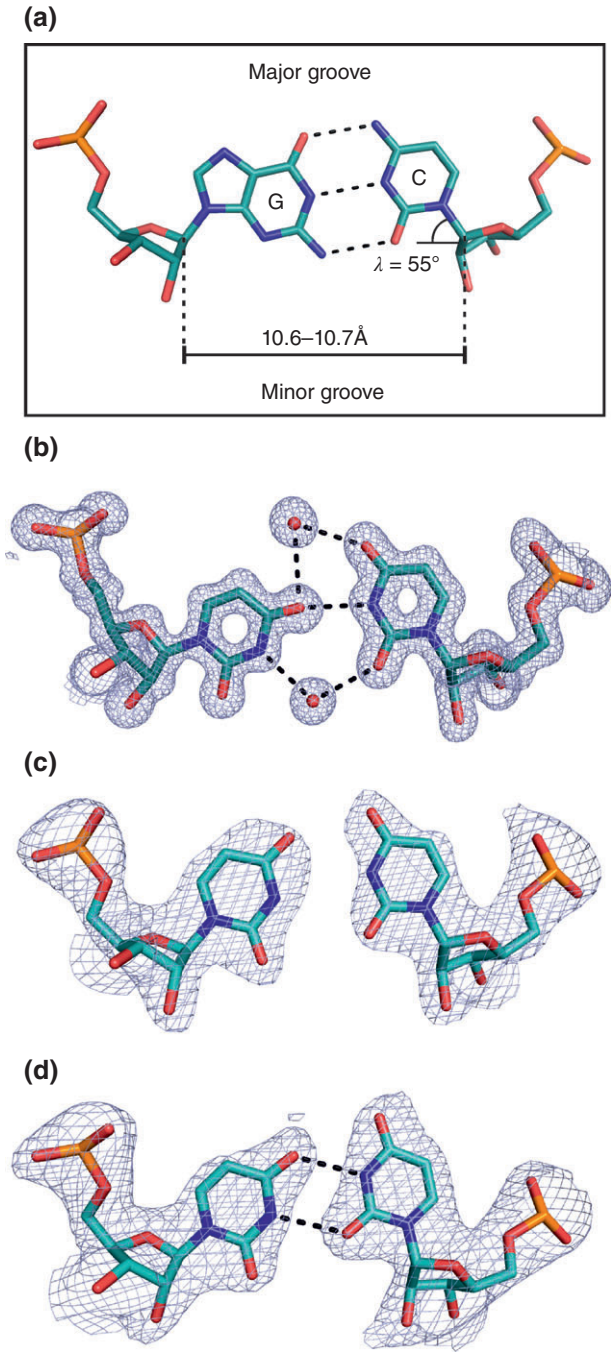


FIGURE 5 | The Watson–Crick G-C (a) and noncanonical U-U pairs (b–d). The G-C pair interacts by three hydrogen bonds (dashed lines). The distance between the C1′-C1′ atoms (black line) and the λ angle are indicated. In the CUG repeats most of the U-U pairs form the stretched U-U *cis* Watson–Crick/Watson–Crick pair (cWW) wobble with one hydrogen bond (b). U-U cWW pairs with zero (c) or two H-bonds (d) have also been observed. The $2F_o$-F_c electron density map is light blue, contoured at 1σ level. Red spheres are water molecules. The presented base pairs were derived from the following PDB entries: (b) code 3GLP[45] (resolution 1.23 Å, R/R_{free} = 14.7/18.4%), (c) code 4E48[47] (resolution 2.5 Å, R/R_{free} = 20.3/27.9%), (d) code 4FNJ[48] (resolution 1.95 Å, R/R_{free} = 20.8/26.6%).

middle of duplex shows *anti-anti* conformation (A-A cWW) (Figure 6(d) and (e)). In the first study the authors claim that the central A-A pair forms one H-bond while in the second study the adenosines do not interact. The reason for the differences between these models can rise from poor and ambiguous electron density maps of most adenosine residues, compared to the atomic resolution structure (Figure 6). The calculated electron density maps of both structures indicate that A-A pairs show static (two or more equally plausible conformations) or dynamic (thermally induced motion of residues) disorder. Thus, the conformation of the A-A pair within these two models remains unclear. The question if the A-A pair can have different conformations within CAG repeats remains open.

Noncanonical G-G Pairs

In the case of CGG repeats there are four known crystal structures. One is a native structure containing two repeats (PDB code: 3R1C).[52] Two contain one modified guanosine residue with substituted bromine atom at position 8 in the base ring (PDB code: 3R1D, 3R1E).[52] The last is a structure (PDB code: 3SJ2) having three CGG repeats and flanking sequences similar to the structure of CUG repeats (PDB code: 3SYW, 3SZX).[53]

All crystal structures present the same conformation of G-G pairs (Figure 7). One guanosine residue is in the *syn* conformation while the second remains *anti*. The pair forms two hydrogen bonds between the Watson–Crick edge of G(*anti)* and the Hoogsteen edge of G(*syn*): O6···N1H and N7···N2H (G-G cWH pair). In addition, an intramolecular bond is observed between the *exo*-amino group and the phosphate group of the G(*syn*) residue. The average distance between the C1′ atoms of the guanosines is 11.3 Å. This type of G-G pair is widely observed in many RNA structures obtained by crystallography or NMR and it seems to be preferred in helical regions of RNA. The characteristic feature of the G-G pair is the alternation from *syn-anti* to *anti-syn* orientation. This static disorder (two possible base pairing geometries) is observed in two crystal structures and also under conditions present during NMR measurements.[52,64]

Noncanonical C-C Pairs

Among all noncanonical base pairs of the CNG repeats, the C-C pairs show the most variability in conformation. In the two crystallographic models of CCG repeats three different C-C cWW pairs are observed characterized by unique pairing (PDB code: 4E58, 4E59).[54] One of the C-C pairs does not form H-bonds (Figure 8(a)). In another C-C pair, the

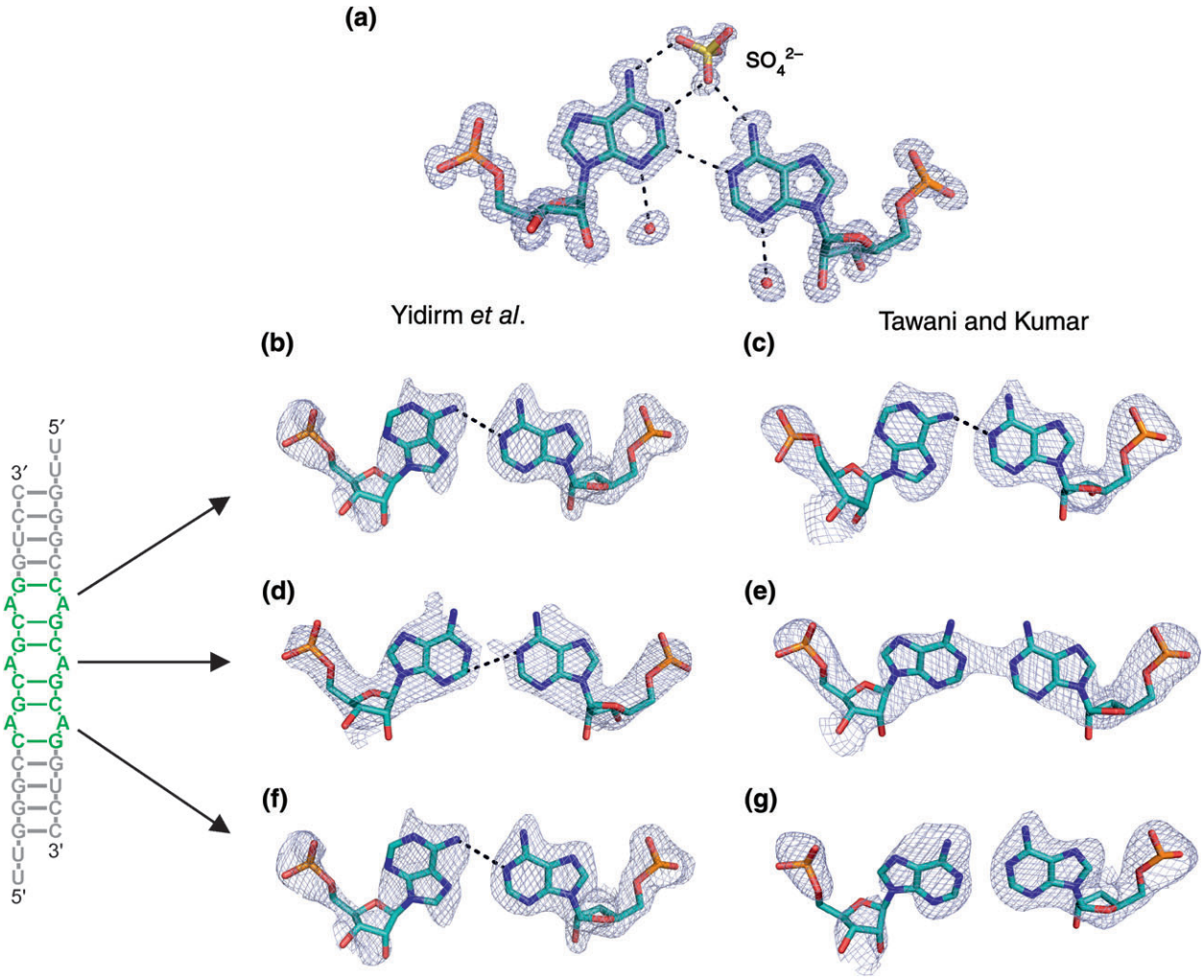


FIGURE 6 | The A-A *cis* Watson–Crick/Watson–Crick pair (cWW) wobble pair at atomic resolution (a) and in lower resolution structures analyzed by Yildrim et al. (b, d, f)[50] and Tawani and Kumar (c, e, g).[51] On the left is a secondary structure of the crystallized duplex containing CAG repeats (green) and flanking sequences (gray). The $2F_o$-F_c electron density map (blue) is contoured at 1σ level. Red spheres are water molecules. The presented base pairs were derived from the following PDB entries: (a) code 3NJ6[49] (resolution 0.95 Å, R/R_{free} = 10.6%/NA), (b, d, f) code 4J50[50] (resolution 1.65 Å, R/R_{free} = 16.9/18.0%), (c, e, g) code 4YN6[51] (resolution 2.3 Å, R/R_{free} = 21.7/26.1%).

cytosine residues probably form one weak interaction between the *exo*-amino group and the N3 atom (3.6 Å) (Figure 8(b)). In the third case, the C-C conformation is similar to the stretched U-U wobble pair. One of the cytosines is inclined toward the minor groove (λ = 31°) and one H-bond is formed between N4H···N3 atoms (Figure 8(c)). In all cases the *exo*-amino groups of the paired cytosines are relatively close but they avoid clashing by twisting the bases relative to each other.[54] Despite the different conformational arrangement of the C-C pairs the C1′-C1′ distances are about 10.8 Å.

Other examples of C-C cWW pairs are found in two similar models (r.m.s.d. 0.28 Å) of RNA containing hexanucleotide CCCCGG repeats of the C9orf72 antisense RNA (PDB code: 5EW4, 5EW7).[55] In each duplex six independent C-C pairs are observed. The conformation of the C-C pairs and the hydrogen bond orientation is described as 'consistent and reproducible' but detailed analysis of the deposited structures indicated some conformational variability of the noncanonical pairs. Each C-C pair forms at least one hydrogen bond but the number of H-bonds and the functional groups involved in the interactions differ (Figure 8(d)). Also the λ angle of both cytosine residues varies as well as the distance between the C1′ atoms of the paired C.

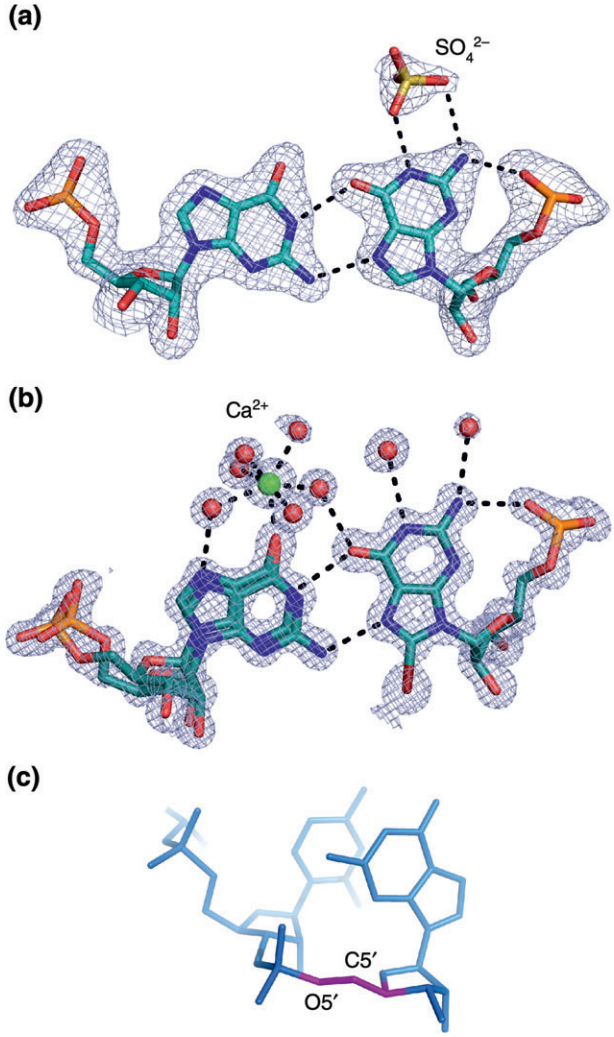


FIGURE 7 | G-G cWH pairs interacting with a sulfate anion (a) or with a Ca^{2+} cation (b) bound in the major groove. In G(*syn*) the O5′-C5′ bond is flipped (purple) (c). The $2F_o$-F_c electron density map (blue) is contoured at 1σ level. Red spheres are water molecules. The presented base pairs were derived from the following PDB entries: (a) code 3R1C[52] (resolution 2.05 Å, R/R_{free} = 23.2/27.0%), (b) code 3R1D[52] (resolution 0.97 Å, R/R_{free} = 13.7%/NA).

Noncanonical C-U Pairs

Noncanonical C-U pairs are present in the structures of CCUG repeats.[56,57] One of the available models contains three CCUG repeats adjacent to GAAA tetraloop/receptor motifs (PDB code: 4 K27).[56] In the structure, six independent C-U cWW pairs are observed. Two interact by one hydrogen bond between N4H (C) and O4(U) (Figure 9(a)). Both residues are inclined toward the minor groove and the C1′-C1′ distance between them is 11.6 Å. In addition, in the minor groove a bridging water molecule interacts with Watson–Crick edges of C and U residues. In the other four noncanonical base pairs, the C and U residues are much closer than the Watson–Crick base pairs (Figure 9(b)). The average C1′-C1′ distance between them is only 8.6 Å. These C-U pairs form two hydrogen bonds: one between the *exo*-amino group of C and O4 carbonyl atom of U, and the second between the N3-imino atom of C and the amino N3H group of U. The carbonyl oxygen atoms of cytosine and uracil are in close proximity, at an average distance of 3.2 Å. This is explained as a result of H-bond formed between the N3 atoms that overcame the repulsive interaction between the carbonyl O atoms.[56]

In another study two structures containing pure CCUG repeats were obtained (4XW0, 4XW1).[57] Three independent C-U cWW pairs show similar conformations having the Watson–Crick edges aligned *vis-à-vis*. The nucleotide residues interact by three hydrogen bonds (Figure 9(c)). One is formed between the N4 *exo*-amino group and the O4 carbonyl atom, and the second between the N3-imino group and the N3-amino group. The third H-bond is observed between two carbonyl O2 atoms (distance between the atoms is 2.9–3.0 Å). This suggests that one of the carbonyl groups becomes a proton donor due to protonation or tautomerization. Thus, the C-U forms a C(enol+)-U or C(imino)-U(enol) or C(imino+)-U(enol) pair (Figure 9(d)). It is possible that the same effect occurs in the structure of CCUG repeats obtained by Childs-Disney et al.[56] If this is the case, most of the observed C-U pairs would have three H-bonds.

The Structure of AUUCU Repeats

In the PDB repository there is a crystal structure of AUUCU repeats embedded within the GAAA tetraloop/receptor motifs (PDB code: 5BTM).[58] In the crystal asymmetric unit, two independent molecules are observed. The AUUCU repeats form a duplex that consists of two canonical A-U and U-A pairs followed by noncanonical U-U, C-C, and U-U pairs (Figure 1(a)). It is likely that the uridines interact by two hydrogen bonds, forming the most common U-U cWW pair (Figure 10 (a)). The cytosine residues are in close proximity (the C1′-C1′ distance is 7.9 Å) and form one hydrogen bond (C-C cWW pair) (Figure 10(b)). As indicated by the authors, the ends of the hairpins, which comprise the AUUCU repeats, show dynamic disorder. In consequence, the electron density map is poor, and only one and a half of a repeat (AUUCUAUU sequence) could be modeled unambiguously.

Accommodation of N-N Pairs Within The A-RNA Helix

It is assumed that the observed RNA structures represent free energy minima. Thus, in the structure of the

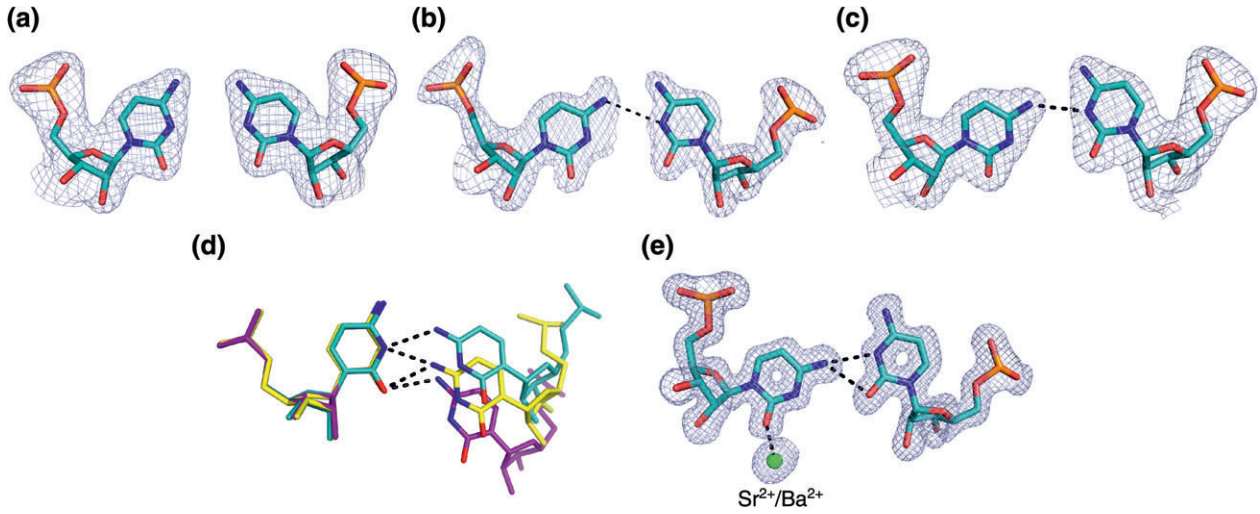


FIGURE 8 | Noncanonical C-C *cis* Watson–Crick/Watson–Crick pair (cWW) pairs observed in CCG (a-c) and CCCCGG repeats (d and e). The C-C pairs in the CCCCGG repeats show a large variety of conformation, the number of hydrogen bonds and values of the λ angles and C1′-C1′ distance (d). One of the noncanonical C-C pairs forms a bifurcated H-bond (e). The $2F_o$-F_c electron density map (blue) is contoured at 1σ level. The presented base pairs were derived from the following PDB entries: (a, c) code 4E58[54] (resolution 1.95 Å, R/R_{free} = 25.8/30.1%), (b) code 4E59[54] (resolution 1.54 Å, R/R_{free} = 25.5/30.3%), (e) code 5EW4[55] (resolution 1.47 Å, R/R_{free} = 21.5/23.9%).

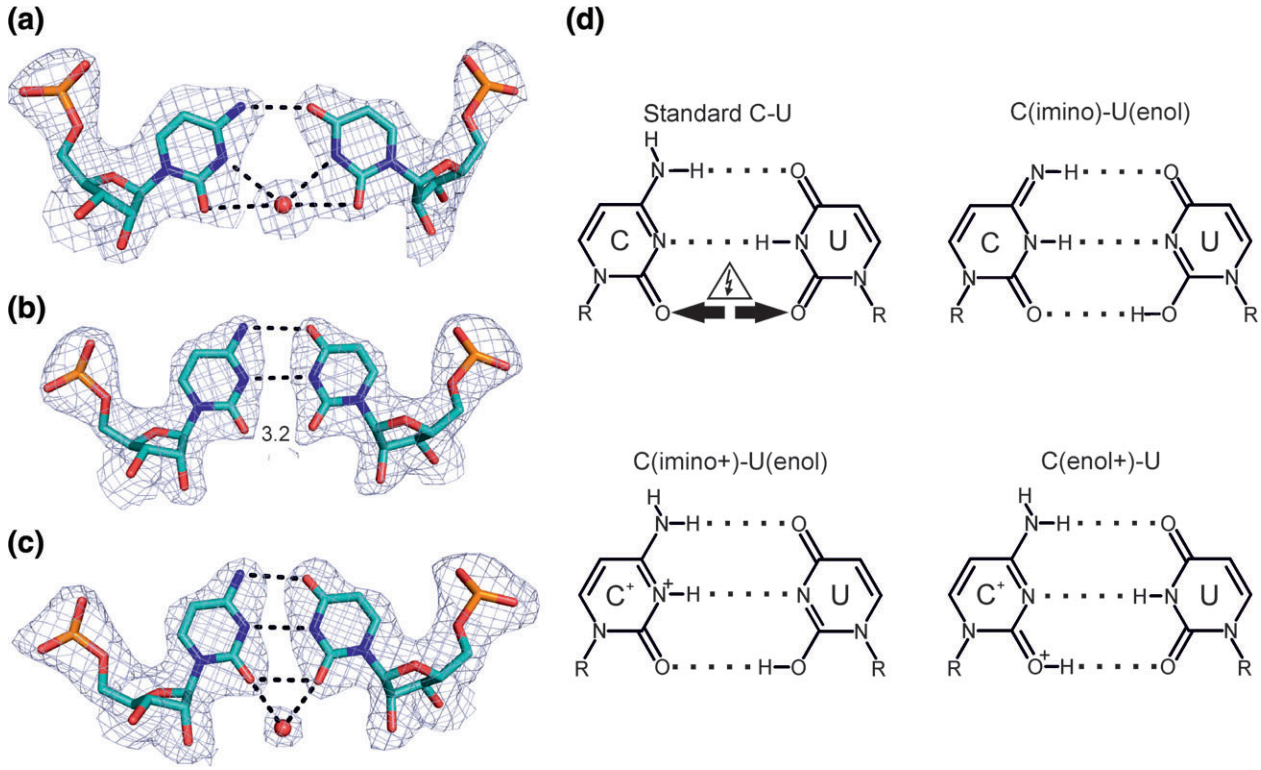


FIGURE 9 | C-U *cis* Watson–Crick/Watson–Crick pair (cWW) pairs observed in CCUG repeats analyzed by Childs-Disney et al.[56] (a and b) and Rypniewski et al. (c).[57] Standard and tautomeric or protonated forms within the C–U pairs are shown on panel d. Water molecules (red spheres) are located in the minor groove. Distance in Å between the O2 atoms is indicated in panel b. The $2F_o$–F_c electron density map (blue) is contoured at 1σ level. Red spheres are water molecules. The presented base pairs were derived from the following PDB entries: (a, b) code 4K27[56] (resolution 2.35 Å, R/R_{free} = 19.4/24.0%), (c) code 4XW1[57] (resolution 2.3 Å, R/R_{free} = 19.1/21.3%).

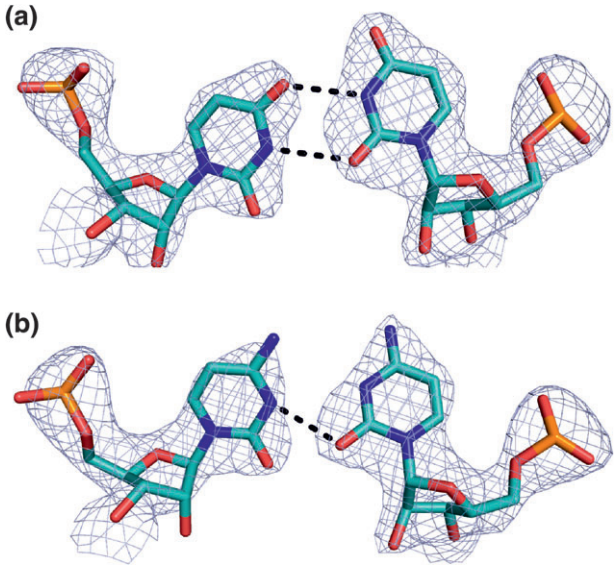


FIGURE 10 | Noncanonical U-U cWW (a) and C-C cWW (b) base pairs in the structure of AUUCU repeats. H-bonds are indicated as they were interpreted by the authors.[58] The $2F_o-F_c$ electron density map (blue) is contoured at 1σ level. The presented base pairs were derived from PDB entry 5BTM[58] (resolution 2.78 Å, R/R_{free} = 17.6/22.4%).

repeats, the mode of pairing of the noncanonical base pairs show a balance between optimizing the H-bonding interactions and responding to constraints imposed by the A-RNA form maintained by the canonical C-G and G-C pairs. Thus, each of the N-N pairs shows a specific mode of accommodation within the RNA helix. The double-helical structure of the RNA restricts the possible conformations of the noncanonical pairs, as reflected in the distance between the C1′-C1′ atoms of paired nucleotide residues (approximately 10.5-10.8 Å). Most of the noncanonical base pairs approximately maintain this distance as they fit within the double-stranded region between Watson–Crick base pairs (Table 2). They do not form bulges or cause a significant deformation of the sugar-phosphate backbone.

The G-G cWH and A-A cWW pairs consist of two bulky purine rings. Although the available space for them in the A-RNA is relatively small both the noncanonical pairs have the distance between the C1′-C1′ atoms only slightly larger than for the G-C and C-G pairs. Nevertheless, the A-A and G-G pairs show a different manner of adjustment into the A-RNA structure. In the G-G pair one of the guanosines assumes the *syn* conformation and positions itself above the ribose ring, thus making space for the second G which remains in the *anti* conformation (Figure 7).[52,53] In addition, the O5′-C5′ dihedral angle of G(*syn*) is rotated to allow for a local 'straightening' of the sugar-phosphate backbone (Figure 7(c)). This conformation seems to be necessary for forming the internucleotide bond between the *exo*-amino group of G(*syn*) and the phosphate O atom. In the A-A pair of the best resolved structure both residues present the *anti* conformation and form cWW pair (Figure 6(a)).[49] They are accommodated by a shift toward the major groove. One A is more inclined, which enables formation of a H-bond. In other structures of CAG repeats the conformation of A-A pairs is difficult to determine due to ambiguous electron density maps.

The uracil ring is less bulky than the purine base and if two uridine residues were aligned *vis-à-vis* at 10.4 Å apart, they would not form any H-bonds (Figure 5(c)). However, the unique conformation of the stretched U-U cWW enables H-bond formation because one of the uridine residues is inclined toward the minor groove (Figure 5(b)). This reduces the distance between the H-bonding functional groups of the uracil rings. In the case of the U-U cWW pair with two H-bonds the residues are closer, approximately 8.8 Å apart, which causes narrowing of the helix. Interestingly, all the observed U-U pairs with two H-bonds are located at the ends of duplexes. Perhaps the local distortion of the width of the helix can only be accommodated at this position.

Accommodation of C-C cWW and C-U cWW pairs within the A-helix is different than for the other noncanonical base pairs. Instead of the expected base pairing we observe duplexes with dangling nucleotides. The strand slippage causes a reduction of the number of noncanonical C-C and C-U base pairs while the number of Watson–Crick base pairs is maximized. This suggests that structures with dangling nucleotides are thermodynamically favored over structures with additional C-C or C-U pairs. If noncanonical C-C pairs are formed, they do not possess one predominant conformation (like U-U or G-G pairs) (Figure 8). Nevertheless, they mostly maintain the C1′-C1′ distance typical of A-RNA (Table 2). In the case of C-U pairs the available data suggest that a majority of the noncanonical pairs form (due to protonation or tautomerization) three H-bonds, and with the C1′-C1′ distance shorter than for the Watson–Crick base pairs[56,57] (Figure 9(c)).

In the structures of CUG repeats, despite the presence of two paired pyrimidines, strand slippage is not observed.[44,45,47] The main difference between C and U is the substitution of the *exo*-amino group of cytosine, which is bulkier than the carbonyl group of uridine. The thermodynamic data show that CCG

and CCUG repeats are less stable than the CUG repeats.[25,57,65] This could explain why the presence of C-C or C-U pairs is associated with strand slippage.

Impact of Noncanonical Base Pairs on The Global Structure of A-RNA

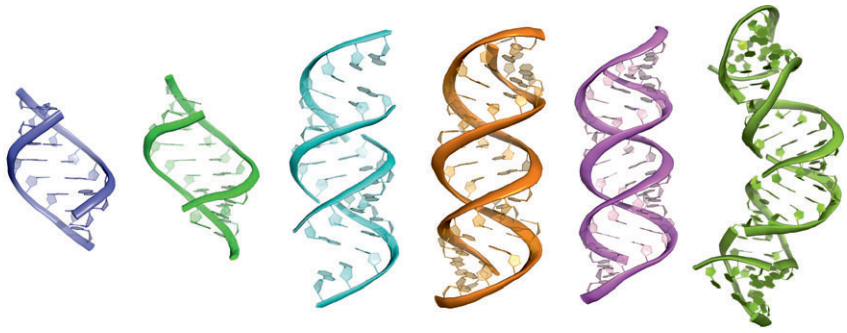
Although the C-G and G-C pairs dominate in most of the structures containing RNA repeats, the noncanonical base pairs can also affect the shape of the RNA helix, distorting it from A-form. This is evident for CGG and CAG repeats, both of which have bulky N-N pairs (Figure 11). Each duplex contains residues showing flipped O5′-C5′ bond. The flipping can be defined in terms of the α and γ dihedral angles (Figure 7(c)). α is almost half a turn from the typical value of −68° while γ deviates ca 120° from the standard value of 58°.[66] This unusual backbone conformation is correlated with a reduction of the helical twist. For the CGG repeats this effect is local and is compensated in other parts of the structure. Thus, the overall helical twist (31°) is only slightly lower than for A-RNA (32.7°). In the case of CAG repeats the effect is global. The helix unwinds and the major groove extends and opens up (>20 Å) (Figure 11). The average helical twist for CAG repeats is only 28.4°.

The C-C cWW pairs affect the structure mostly by inducing a strand slippage. The values of helical parameters are similar to A-RNA form, which is also observed for the CUG repeats. The helical twist is 33° for both CCG and CUG and 31.6° for CCCCGG repeats. The roll, buckle, or propeller values vary to some extent but the changes are local and limited to regions with noncanonical base pairs. The C-U cWW pairs also cause strand slippage. In addition, they are responsible for narrowing and bending of the helix compared to a canonical A-RNA (Figure 11). In consequence, the alignment of C-shaped molecules into a pseudo-infinite helix amounts to supercoiling. The bending is accompanied by shrinking of the rise parameter (2.3 versus 2.8 Å in A-RNA) and closing of the major groove. On balance the duplex opens up at the ends. The C-U pairs seem to have the biggest effect on the RNA structure (PDB code: 4XW0, 4XW1), perhaps because, as the only noncanonical base pairs, they are in equal number with the Watson–Crick base pairs.

Interactions With Solvent Molecules

Interactions with the solvent molecules can provide information about the properties of the RNA structure. In addition, it can serve as a guide for designing molecules to target the RNA.

In RNA repeats the way of pairing of most noncanonical base pairs does not saturate their H-bonding capacity. Functional groups located at the Watson–Crick edges of the base rings show the highest potential for interaction. In the canonical base pairs these groups are involved in H-bonding. In the stretched U-U cWW pair the inclined uridine exposes the Watson–Crick edge toward the minor groove.



	CGG	CAG	CCG	CUG	CCCCGG	CCUG
Twist (°)	31.0	28.4	32.6	33.2	31.6	31.8
Rise (Å)	2.9	2.8	2.4	2.6	2.6	2.3
Major groove width (Å)	18	24	15	13	13	Variable

FIGURE 11 | All the crystallized RNA repeats fold into the A-RNA form. The presence of G-G cWH and A-A *cis* Watson–Crick/Watson–Crick pair (cWW) pairs is accompanied by unwinding of the helix and widening of the major groove. The helix of CCUG repeats is bent and twisted, which amounts to supercoiling. The presented helices were derived from the following PDB entries: 3R1C,[52] 3NJ6,[49] 4E59,[54] 4E48,[47] 5EW4,[55] 4XW1.[57]

Moreover, both the carbonyl groups of the second U are also exposed. The U-U pair fulfills its binding potential by interacting with two water molecules (Figure 5(b)). One is located in the minor groove and is H-bonded to the N3 amino group of the inclined U and to the O2 carbonyl atom of the other U. The second water molecule is found in the major groove. It interacts with the carbonyl O4 atoms of each uridine residues. The C-C cWW pairs can interact with a water molecule bound in the minor groove. Similarly to the 'U-U water' it interacts with the N3 and O2 atoms of the inclined C and the O2 carbonyl atom of the second C. In addition, in the structures of CCCCGG repeats one of the inclined cytosine residues interacts with the Sr^{2+} or Ba^{2+} ions located in the minor groove (Figure 8(e)). The O2 carbonyl atom forms an inner complex with the cation. In the major groove of C-C pairs no characteristic solvent molecules are observed. In the case of the C-U cWW pairs with three H-bonds, a water molecule is present in the minor groove (Figure 9(c)). It is H-bonded to both the O2 carbonyl atoms. When a C-U pair forms one H-bond, the water molecule is wedged between C and U and interacts with the N1 and O2 atoms of each residue (Figure 9(a)).

In the G-G cWH pair one guanosine residue is in the *syn* conformation, exposing its Watson–Crick edge toward the major groove. The binding capacity of G(*syn*) is fulfilled by the interaction with the sulfate anion which seems to fit well in terms of the geometric and chemical properties (Figure 7(a)). Oxygen atoms of the sulfate ion are good proton acceptors and the distance between them corresponds to the distance between the two amino groups of a guanosine residue. The G-G pair can also interact with a Ca^{2+} ion in the major groove (Figure 7(b)). Cation interacts with the O6 carbonyl group of the G(*anti*) residue. When a suitable ligand is not present in the medium it is replaced by interactions with water molecules.

Adenosine residues of the A-A cWW pairs interact with the sulfate ion located in the major groove (Figure 6(a)). The anion is wedged between A residues and forms one hydrogen bond with the *exo*-amino group of the less inclined A and two hydrogen bonds with the N1-imino and *exo*-amino group of the second adenosine. Similar to G-G pairs, when a sulfate ion is not present, a water molecule is bound. In the minor groove of the A-A pair two characteristic water molecules are observed (Figure 6(a)). Each of them interacts with the N3 atom of one A. Similar hydration can be observed for canonical base pairs but in the case of A-A pairs these interactions seem to be more specific.

Biological Aspects of RNA Repeats

The length polymorphism of microsatellite sequences is an important source of genetic variability. However in certain cases, when a repeated region expands abnormally, it folds into aberrant RNA structures, a prerequisite for pathogenesis. Although structural studies have been conducted on short RNA oligomers, the majority of structures of RNA repeats resemble the stem of a long hairpin composed of expanded runs. In the crystal lattice, neighboring duplexes interact by stacking (blunt-ended duplexes) or form Watson–Crick base pairs between dangling residues. As a result, unwinding of the ends of the duplexes is not observed and RNA molecules stack end-to-end forming pseudo-infinitive helices. The best example is the native crystal structure of CGG repeats.[52] In the asymmetric unit there are 18 independent duplexes which can be arrange into a 32-repeat long helix giving a snapshot how three-dimensional structure of mutated RNA runs can look like *in vivo*. An exception from the stacking rule are only AUUCU repeats. In the crystal structure the ends of the molecule are not visible due to dynamic disorder which suggest lower stability of the repeats.[58] This is supported by the biochemical data showing that the AUUCU runs form a hairpin structure only at 20°C.[38]

Some of the crystallized CNG repeats have additional base pairs included to facilitate crystallization, stabilizing the ends of the duplexes and inducing stacking interactions between molecules in the crystal. However, one could ask how these structures are relevant to the biological structures in the living cell. A comparison between the known RNA models with or without flanking sequences shows that the A-helical conformation is maintained but interactions within the noncanonical pairs can be affected.[45–51] Sometimes, adding flanking sequences is advisable because natural 'clamps' occur also in native RNA (see section Secondary structure of RNA repeats). Crystal lattice contacts do not seem to have much effect on the structure; same RNA oligomers crystallized in different crystal forms are similar. Crystallization conditions are certainly different from *in vivo* but one should note that crystals of biological molecules usually have a solvent content of around 50%. The nucleic acids in the crystal are highly solvated and most of their surface is surrounded by ordered solvent molecules. In fact, crystallography is by far the best method to observe the detailed interactions with the solvent (water, ions, small ligands), which can be related to conditions in the cell.

Secondary structure probing and crystallographic studies suggest that most of the RNA repeats form hairpins. In the case of CGG and GGGGCC repeats

tetraplex structures have been proposed.[43,67–70] However, crystallographic data clearly show that CGG repeats form duplexes.[52,53] Even an introduction of 8-bromoguanosine into the CGG repeats, which is suggested to promote tetraplex folding, resulted in crystals of the duplex.[52] These observations indicate that hairpin is the primary structure for CGG repeats. The association of CGG duplexes into tetraplex cannot be ruled out but this could require specific conditions. In the case of GGGGCC repeats, the three-dimensional structure is not known and the question concerning structural arrangement remains open.

So far the reasons why structures of RNA repeats have the ability to sequester so many proteins are elusive. Regardless of particular scenarios the structural features of hairpins composed of repeated tracts can be important in RNA-driven pathogenesis. First, unusually long A-RNA like stem, common to most of the expanded RNA repeats, can serve as a platform for excessive RNA/protein interactions. Second, the repetitive, noncanonical base pairs of the long hairpin stem have unsaturated binding capacity that can attract the protein ligands. The N-N pairs correspond also to unique character of each repeat. For example, one of the proteins sequestered by RNA repeats is MBNL1 which was shown to bind CUG, CAG, and CCG but not CGG repeats.[71–73] In all three types of repeats the noncanonical base pairs usually have only one H-bond which can be crucial for MBNL1 recognition. This hypothesis is supported by a recent observation that MBNL1 binds single-stranded RNA.[74,75] In that case, unwinding of CUG, CAG, and CCG helices would be easier than for CGG repeats. Although this needs to be verified experimentally, it points to a future research direction: structural studies of RNA repeats/protein complexes which will be important in understanding the pathogenic properties of RNA runs.

TOWARDS A THERAPY AGAINST PATHOGENIC RNA REPEATS

Diseases associated with expanding RNA repeats are neurodegenerative, progressive, and incurable. The current treatment aims only to minimize the secondary characteristics associated with the disorder. Thus, many ongoing studies are devoted to development of effective therapies. For targeting expanded RNA molecules two main approaches were employed. In the first the goal is to degrade RNA molecules using antisense oligomers or site-specific RNA endonucleases.[76–83] In the second approach, pathogenic properties of RNA are blocked by interactions with small molecules or antisense oligonucleotides.[76,84–90]

Although a number of biochemical and chemical studied has been carried out on a therapy of repeat associated disorders most of them are not supplemented by structural data. One of the few exceptions is the structure of CUG repeats having uridine residues replaced by pseudouridine (Ψ) (PDB code: 4PCJ).[59] It was shown that the introduction of Ψ stabilizes the double-stranded or helical structure formed by CUG repeats. In consequence, the pseudouridine-enriched CUG repeats increase the thermal stability which results in a reduced affinity for MBNL1 protein, sequestered by CUG runs in myotonic dystrophy type 1. In the crystal structure, the noncanonical Ψ-U cWW pair, similar to U-U cWW pair, have a stretched wobble conformation with one H-bond (Figure 12(a)). The C1′-C1′ distance is larger (11.0 Å) than for the native U-U pair (10.5 Å). In the major groove a water molecule is present that interacts with the O6 carbonyl group of U and with the N1 amine group of Ψ. Molecular dynamics simulation revealed that Ψ-Ψ pair requires more energy to open than a native pair.[59] Moreover, the bridging water molecule found in the major groove adds two hydrogen bonds to the Ψ-U pair which probably stabilizes its conformation and the overall RNA structure.

The next example are crystal structures of CUG and CAG repeats in complex with their antisense PNA oligomers as well as the structure of PNA-PNA duplex having CTG repeats (PDB code: 5EME, 5EMF, 5EMG).[60] PNA (peptide nucleic acid) is a homolog of a nucleic acid having the sugar-phosphate backbone replaced by a peptide backbone.[91] The study was aimed to see how PNA antisense oligomers recognize RNA repeats and to understand PNA's outstanding sequence selectivity. The two obtained structures of RNA–PNA complexes turned out to be isomorphic. Despite different sequences, they form identical helices with fully complementary Watson–Crick base pairs. The helices have the A-from, but some helical parameters show deviations from the canonical values. For example, the helical twist (26°) and rise (2.4 Å) are low. The structure of PNA-PNA duplex contains C-G and G-C pairs and two non-Watson–Crick T-T pairs. The T-T cWW pairs have two hydrogen bonds between N3 and O4 and O2 and N3 atoms (Figure 12(b)). One of T residues is inclined toward the minor groove more than the other. The distance between T-T is 9.0 Å (10.7 Å for neighboring Watson–Crick base pairs). The PNA–PNA duplex has the P-PNA form with a low helical twist (19.7°) and very high helical rise (3.9 Å) indicating different conformational preferences of PNA from RNA. The data show that the RNA–PNA helix is an intermediate of

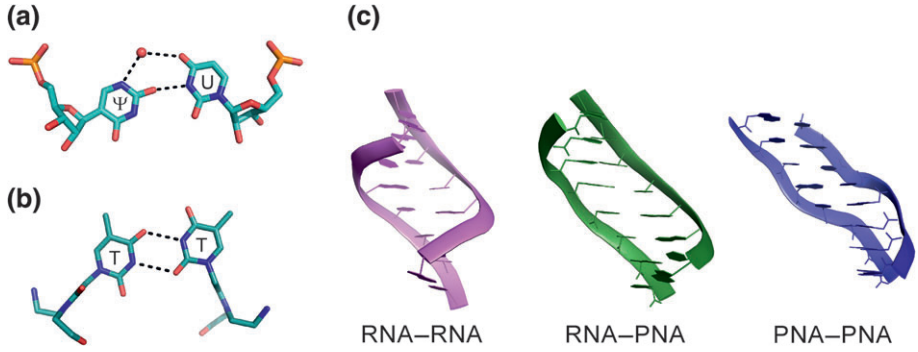


FIGURE 12 | The noncanonical Ψ-U pair inserted into CUG repeats (a) and T-T pair found in a PNA-PNA duplex (b). Comparison of the overall structure of a mixed RNA–PNA duplex against RNA–RNA and PNA-PNA (c). Red sphere is a water molecule. The presented base pairs and helices were derived from the following PDB entries: (a) code 4PCJ,[59] (b) code 5EMG,[60] (c) code 3GLP,[45] 5EME,[60] 5EMG.

the A-RNA and P-PNA structures (Figure 12(c)). Both the RNA and PNA molecules adapt to one another to form a fully-complementary duplex. Formation of mismatches in RNA–PNA helix would require additional conformational changes which seem to be difficult to overcome.

CONCLUSION

Herein we presented all the known structures linked to RNA repeat diseases. Our intention was to perform a comprehensive analysis of the available models. The obtained data provide information about the unique and characteristic features of RNA repeats. Their structural profile have been generated and can serve as a guide in a drug design process or as an inspiration for further biochemical and chemical studies.[26] Structural data can be also considered as a source of basic knowledge about RNA structures. We believe that future research devoted to structural studies of RNA repeats complexed with ligands and proteins will accelerate the discovery of new therapies. We would like also to encourage all scientists to explore the crystallographic and NMR models by themselves. Useful validation metrics can be found in the PDB database, namely real-space refinement Z-score (RSRZ). A related measure is provided in the Coot software in the module 'density fit analysis'.[92] They indicate a fit of the atomic coordinates to the electron density map. Inspection of the atomic B-factors provides similar information.

ACKNOWLEDGMENTS

We thank Marta Kowalik and Jacek Kocur for valuable comments and constant support. This work was supported by the National Science Centre (Poland) [UMO-2011/01/B/NZ1/04429]; Ministry of Science and Higher Education (Poland) [0450/IP1/2013/72,01/KNOW2/2014].

REFERENCES

1. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2012, 13:36–46.
2. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Res* 2000, 10:62–71.
3. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000, 10:967–981.
4. Kashi Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* 1997, 13:74–78.
5. Tautz D, Trick M, Dover GA. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 1986, 322:652–656.
6. Echeverria GV, Cooper TA. RNA-binding proteins in microsatellite expansion disorders: mediators of RNA toxicity. *Brain Res* 2012, 1462:100–111.

7. Mohan A, Goodwin M, Swanson MS. RNA-protein interactions in unstable microsatellite diseases. *Brain Res* 2014, 1584:3–14.
8. Nelson DL, Orr HT, Warren ST. The unstable repeats—three evolving faces of neurological disease. *Neuron* 2013, 77:825–843.
9. Walsh MJ, Cooper-Knock J, Dodd JE, Stopford MJ, Mihaylov SR, Kirby J, Shaw PJ, Hautbergue GM. Invited review: decoding the pathophysiological mechanisms that underlie RNA dysregulation in neurodegenerative disorders: a review of the current state of the art. *Neuropathol Appl Neurobiol* 2015, 41:109–134.
10. Lee DY, McMurray CT. Trinucleotide expansion in disease: why is there a length threshold? *Curr Opin Genet Dev* 2014, 26:131–140.
11. Mirkin SM. Expandable DNA repeats and human disease. *Nature* 2007, 447:932–940.
12. Almeida B, Fernandes S, Abreu IA, Macedo-Ribeiro S. Trinucleotide repeats: a structural perspective. *Front Neurol* 2013, 4:76.
13. Fan HC, Ho LI, Chi CS, Chen SJ, Peng GS, Chan TM, Lin SZ, Harn HJ. Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant* 2014, 23:441–458.
14. Katsuno M, Watanabe H, Yamamoto M, Sobue G. Potential therapeutic targets in polyglutamine-mediated diseases. *Expert Rev Neurother* 2014, 14:1215–1228.
15. Pettersson OJ, Aagaard L, Jensen TG, Damgaard CK. Molecular mechanisms in DM1—a focus on foci. *Nucleic Acids Res* 2015, 43:2433–2441.
16. Wojciechowska M, Krzyzosiak WJ. Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Hum Mol Genet* 2011, 20:3811–3821.
17. Dickson AM, Wilusz CJ. Repeat expansion diseases: when a good RNA turns bad. *Wiley Interdiscip Rev RNA* 2010, 1:173–192.
18. Green KM, Linsalata AE, Todd PK. RAN translation—What makes it run? *Brain Res* 2016, 1647:30–42.
19. Wojciechowska M, Olejniczak M, Galka-Marciniak P, Jazurek M, Krzyzosiak WJ. RAN translation and frameshifting as translational challenges at simple repeats of human neurodegenerative disorders. *Nucleic Acids Res* 2014, 42:11849–11864.
20. Groh M, Silva LM, Gromak N. Mechanisms of transcriptional dysregulation in repeat expansion disorders. *Biochem Soc Trans* 2014, 42:1123–1128.
21. La Spada AR, Taylor JP. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat Rev Genet* 2010, 11:247–258.
22. Darty K, Denise A, Ponty Y. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 2009, 25:1974–1975.
23. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 2010, 11:129.
24. Sobczak K, de Mezer M, Michlewski G, Krol J, Krzyzosiak WJ. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res* 2003, 31:5469–5482.
25. Sobczak K, Michlewski G, de Mezer M, Kierzek E, Krol J, Olejniczak M, Kierzek R, Krzyzosiak WJ. Structural diversity of triplet repeat RNAs. *J Biol Chem* 2010, 285:12755–12764.
26. Kiliszek A, Rypniewski W. Structural studies of CNG repeats. *Nucleic Acids Res* 2014, 42:8189–8199.
27. Galka-Marciniak P, Urbanek MO, Krzyzosiak WJ. Triplet repeats in transcripts: structural insights into RNA toxicity. *Biol Chem* 2012, 393:1299–1315.
28. Handa V, Saha T, Usdin K. The fragile X syndrome repeats form RNA hairpins that do not activate the interferon-inducible protein kinase, PKR, but are cut by Dicer. *Nucleic Acids Res* 2003, 31:6243–6248.
29. Krzyzosiak WJ, Sobczak K, Wojciechowska M, Fiszer A, Mykowska A, Kozlowski P. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res* 2012, 40:11–26.
30. Tian B, White RJ, Xia T, Welle S, Turner DH, Mathews MB, Thornton CA. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* 2000, 6:79–87.
31. Michlewski G, Krzyzosiak WJ. Molecular architecture of CAG repeats in human disease related transcripts. *J Mol Biol* 2004, 340:665–679.
32. Sobczak K, Krzyzosiak WJ. Imperfect CAG repeats form diverse structures in SCA1 transcripts. *J Biol Chem* 2004, 279:41563–41572.
33. Napierala M, Michalowski D, de Mezer M, Krzyzosiak WJ. Facile FMR1 mRNA structure regulation by interruptions in CGG repeats. *Nucleic Acids Res* 2005, 33:451–463.
34. de Mezer M, Wojciechowska M, Napierala M, Sobczak K, Krzyzosiak WJ. Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. *Nucleic Acids Res* 2011, 39:3852–3863.
35. Napierala M, Krzyzosiak WJ. CUG repeats present in myotonin kinase RNA form metastable "slippery" hairpins. *J Biol Chem* 1997, 272:31079–31085.
36. Sobczak K, Krzyzosiak WJ. CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J Biol Chem* 2005, 280:3898–3910.
37. Busan S, Weeks KM. Role of context in RNA structure: flanking sequences reconfigure CAG motif folding in huntingtin exon 1 transcripts. *Biochemistry* 2013, 52:8219–8225.

38. Handa V, Yeh HJ, McPhie P, Usdin K. The AUUCU repeats responsible for spinocerebellar ataxia type 10 form unusual RNA hairpins. *J Biol Chem* 2005, 280:29340–29345.

39. Haeusler AR, Donnelly CJ, Periz G, Simko EA, Shaw PG, Kim MS, Maragakis NJ, Troncoso JC, Pandey A, Sattler R, et al. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* 2014, 507:195–200.

40. Su Z, Zhang Y, Gendron TF, Bauer PO, Chew J, Yang WY, Fostvedt E, Jansen-West K, Belzil VV, Desaro P, et al. Discovery of a biomarker and lead small molecules to target r(GGGGCC)-associated defects in c9FTD/ALS. *Neuron* 2014, 83:1043–1050.

41. Teive HA, Munhoz RP, Arruda WO, Raskin S, Werneck LC, Ashizawa T. Spinocerebellar ataxia type 10—a review. *Parkinsonism Relat Disord* 2011, 17:655–661.

42. Gendron TF, Belzil VV, Zhang YJ, Petrucelli L. Mechanisms of toxicity in C9FTLD/ALS. *Acta Neuropathol* 2014, 127:359–376.

43. Fratta P, Mizielinska S, Nicoll AJ, Zloh M, Fisher EM, Parkinson G, Isaacs AM. C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. *Sci Rep* 2012, 2:1016.

44. Mooers BH, Logue JS, Berglund JA. The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc Natl Acad Sci USA* 2005, 102:16626–16631.

45. 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, 37:4149–4156.

46. Kumar A, Park H, Fang P, Parkesh R, Guo M, Nettles KW, Disney MD. Myotonic dystrophy type 1 RNA crystal structures reveal heterogeneous 1 x 1 nucleotide UU internal loop conformations. *Biochemistry* 2011, 50:9928–9935.

47. 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, 30:505–523.

48. Coonrod LA, Lohman JR, Berglund JA. Utilizing the GAAA tetraloop/receptor to facilitate crystal packing and determination of the structure of a CUG RNA helix. *Biochemistry* 2012, 51:8330–8337.

49. 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.

50. Yildirim I, Park H, Disney MD, Schatz GC. A dynamic structural model of expanded RNA CAG repeats: a refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J Am Chem Soc* 2013, 135:3528–3538.

51. Tawani A, Kumar A. Structural insights reveal the dynamics of the repeating r(CAG) transcript found in Huntington's disease (HD) and spinocerebellar ataxias (SCAs). *PLoS One* 2015, 10:e0131788.

52. 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.

53. 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. *Chembiochem* 2011, 12:2140–2142.

54. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Crystallographic characterization of CCG repeats. *Nucleic Acids Res* 2012, 40:8155–8162.

55. Dodd DW, Tomchick DR, Corey DR, Gagnon KT. Pathogenic C9ORF72 antisense repeat RNA forms a double helix with tandem C:C mismatches. *Biochemistry* 2016, 55:1283–1286.

56. Childs-Disney JL, Yildirim I, Park H, Lohman JR, Guan L, Tran T, Sarkar P, Schatz GC, Disney MD. Structure of the myotonic dystrophy type 2 RNA and designed small molecules that reduce toxicity. *ACS Chem Biol* 2014, 9:538–550.

57. Rypniewski W, Banaszak K, Kulinski T, Kiliszek A. Watson-Crick-like pairs in CCUG repeats: evidence for tautomeric shifts or protonation. *RNA* 2016, 22:22–31.

58. Park H, Gonzalez AL, Yildirim I, Tran T, Lohman JR, Fang P, Guo M, Disney MD. Crystallographic and computational analyses of AUUCU repeating RNA that causes spinocerebellar ataxia type 10 (SCA10). *Biochemistry* 2015, 54:3851–3859.

59. deLorimier E, Coonrod LA, Copperman J, Taber A, Reister EE, Sharma K, Todd PK, Guenza MG, Berglund JA. Modifications to toxic CUG RNAs induce structural stability, rescue mis-splicing in a myotonic dystrophy cell model and reduce toxicity in a myotonic dystrophy zebrafish model. *Nucleic Acids Res* 2014, 42:12768–12778.

60. Kiliszek A, Banaszak K, Dauter Z, Rypniewski W. The first crystal structures of RNA-PNA duplexes and a PNA-PNA duplex containing mismatches-toward antisense therapy against TREDs. *Nucleic Acids Res* 2016, 44:1937–1943.

61. Zheng G, Lu XJ, Olson WK. Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res* 2009, 37:W240–W246.

62. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA* 2001, 7:499–512.

63. Sweeney BA, Roy P, Leontis NB. An introduction to recurrent nucleotide interactions in RNA. *Wiley Interdiscip Rev RNA* 2015, 6:17–45.
64. Zumwalt M, Ludwig A, Hagerman PJ, Dieckmann T. Secondary structure and dynamics of the r(CGG) repeat in the mRNA of the fragile X mental retardation 1 (FMR1) gene. *RNA Biol* 2007, 4:93–100.
65. Broda M, Kierzek E, Gdaniec Z, Kulinski T, Kierzek R. Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry* 2005, 44:10873–10882.
66. Saenger W. *Principles of nucleic acid structure*. New York: Springer-Verlag; 1984.
67. Gudanis D, Popenda L, Szpotkowski K, Kierzek R, Gdaniec Z. Structural characterization of a dimer of RNA duplexes composed of 8-bromoguanosine modified CGG trinucleotide repeats: a novel architecture of RNA quadruplexes. *Nucleic Acids Res* 2016, 44:2409–2416.
68. Khateb S, Weisman-Shomer P, Hershco I, Loeb LA, Fry M. Destabilization of tetraplex structures of the fragile X repeat sequence (CGG)n is mediated by homolog-conserved domains in three members of the hnRNP family. *Nucleic Acids Res* 2004, 32:4145–4154.
69. Khateb S, Weisman-Shomer P, Hershco-Shani I, Ludwig AL, Fry M. The tetraplex (CGG)n destabilizing proteins hnRNP A2 and CBF-A enhance the in vivo translation of fragile X premutation mRNA. *Nucleic Acids Res* 2007, 35:5775–5788.
70. Ofer N, Weisman-Shomer P, Shklover J, Fry M. The quadruplex r(CGG)n destabilizing cationic porphyrin TMPyP4 cooperates with hnRNPs to increase the translation efficiency of fragile X premutation mRNA. *Nucleic Acids Res* 2009, 37:2712–2722.
71. Ho TH, Savkur RS, Poulos MG, Mancini MA, Swanson MS, Cooper TA. Colocalization of muscleblind with RNA foci is separable from mis-regulation of alternative splicing in myotonic dystrophy. *J Cell Sci* 2005, 118:2923–2933.
72. Kino Y, Mori D, Oma Y, Takeshita Y, Sasagawa N, Ishiura S. Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. *Hum Mol Genet* 2004, 13:495–507.
73. Yuan Y, Compton SA, Sobczak K, Stenberg MG, Thornton CA, Griffith JD, Swanson MS. Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res* 2007, 35:5474–5486.
74. Fu Y, Ramisetty SR, Hussain N, Baranger AM. MBNL1-RNA recognition: contributions of MBNL1 sequence and RNA conformation. *Chembiochem* 2012, 13:112–119.
75. Teplova M, Patel DJ. Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat Struct Mol Biol* 2008, 15:1343–1351.
76. Fiszer A, Krzyzosiak WJ. Oligonucleotide-based strategies to combat polyglutamine diseases. *Nucleic Acids Res* 2014, 42:6787–6810.
77. Francois V, Klein AF, Beley C, Jollet A, Lemercier C, Garcia L, Furling D. Selective silencing of mutated mRNAs in DM1 by using modified hU7-snRNAs. *Nat Struct Mol Biol* 2011, 18:85–87.
78. Gonzalez-Barriga A, Mulders SA, van de Giessen J, Hooijer JD, Bijl S, van Kessel ID, van Beers J, van Deutekom JC, Fransen JA, Wieringa B, et al. Design and analysis of effects of triplet repeat oligonucleotides in cell models for myotonic dystrophy. *Mol Ther Nucleic Acids* 2013, 2:e81.
79. Lee JE, Bennett CF, Cooper TA. RNase H-mediated degradation of toxic RNA in myotonic dystrophy type 1. *Proc Natl Acad Sci USA* 2012, 109:4221–4226.
80. Mulders SA, van den Broek WJ, Wheeler TM, Croes HJ, van Kuik-Romeijn P, de Kimpe SJ, Furling D, Platenburg GJ, Gourdon G, Thornton CA, et al. Triplet-repeat oligonucleotide-mediated reversal of RNA toxicity in myotonic dystrophy. *Proc Natl Acad Sci USA* 2009, 106:13915–13920.
81. Sobczak K, Wheeler TM, Wang WL, Thornton CA. RNA interference targeting CUG repeats in a mouse model of myotonic dystrophy. *Mol Ther* 2013, 21:380–387.
82. Wheeler TM, Leger AJ, Pandey SK, MacLeod AR, Nakamori M, Cheng SH, Wentworth BM, Bennett CF, Thornton CA. Targeting nuclear RNA for in vivo correction of myotonic dystrophy. *Nature* 2012, 488:111–115.
83. Zhang W, Wang Y, Dong S, Choudhury R, Jin Y, Wang Z. Treatment of type 1 myotonic dystrophy by engineering site-specific RNA endonucleases that target (CUG)(n) repeats. *Mol Ther* 2014, 22:312–320.
84. Arambula JF, Ramisetty SR, Baranger AM, Zimmerman SC. A simple ligand that selectively targets CUG trinucleotide repeats and inhibits MBNL protein binding. *Proc Natl Acad Sci USA* 2009, 106:16068–16073.
85. Childs-Disney JL, Parkesh R, Nakamori M, Thornton CA, Disney MD. Rational design of bioactive, modularly assembled aminoglycosides targeting the RNA that causes myotonic dystrophy type 1. *ACS Chem Biol* 2012, 7:1984–1993.
86. Gareiss PC, Sobczak K, McNaughton BR, Palde PB, Thornton CA, Miller BL. Dynamic combinatorial selection of molecules capable of inhibiting the (CUG) repeat RNA-MBNL1 interaction in vitro: discovery of lead compounds targeting myotonic dystrophy (DM1). *J Am Chem Soc* 2008, 130:16254–16261.
87. Konieczny P, Stepniak-Konieczna E, Sobczak K. MBNL proteins and their target RNAs, interaction and splicing regulation. *Nucleic Acids Res* 2014, 42:10873–10887.

88. Warf MB, Nakamori M, Matthys CM, Thornton CA, Berglund JA. Pentamidine reverses the splicing defects associated with myotonic dystrophy. *Proc Natl Acad Sci USA* 2009, 106:18551–18556.
89. Wheeler TM, Sobczak K, Lueck JD, Osborne RJ, Lin X, Dirksen RT, Thornton CA. Reversal of RNA dominance by displacement of protein sequestered on triplet repeat RNA. *Science* 2009, 325:336–339.
90. Wojtkowiak-Szlachcic A, Taylor K, Stepniak-Konieczna E, Sznajder LJ, Mykowska A, Sroka J, Thornton CA, Sobczak K. Short antisense-locked nucleic acids (all-LNAs) correct alternative splicing abnormalities in myotonic dystrophy. *Nucleic Acids Res* 2015, 43:3318–3331.
91. Gambari R. Peptide nucleic acids: a review on recent patents and technology transfer. *Expert Opin Ther Pat* 2014, 24:267–294.
92. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 2010, 66:486–501.

