Published online 11 May 2009

Nucleic Acids Research, 2009, Vol. 37, No. 12 ***4149–4156***
doi:10.1093/nar/gkp350

Structural insights into CUG repeats containing the 'stretched U–U wobble': implications for myotonic dystrophy

Agnieszka Kiliszek, Ryszard Kierzek, Wlodzimierz J. Krzyzosiak and Wojciech Rypniewski*

Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

Received January 15, 2009; Revised April 22, 2009; Accepted April 23, 2009

ABSTRACT

Tracks containing CUG repeats are abundant in human gene transcripts. Their biological role includes modulation of pre-mRNA splicing, mRNA transport and regulation of translation. Expanded forms of CUG runs are associated with pathogenesis of several neurodegenerative diseases, including myotonic dystrophy type 1. We have analysed two crystal structures of RNA duplexes containing the CUG repeats: G(CUG)$_2$C and (CUG)$_6$. The first of the structures, analysed at 1.23 Å resolution, is of an oligomer designed by us. The second model was obtained after 'detwinning' the 1.58 Å X-ray data previously deposited in the PDB. The RNA duplexes are in the A-form in which all the C–G pairs form Watson–Crick interactions while all the uridine pairs can be described as U•U *cis* wobble having only one hydrogen bond between the bases. The residue, which accepts the H-bond, is inclined towards the minor groove. This previously unreported base pairing can be described as 'stretched U–U wobble'. The regular hydrogen-bonding pattern of interactions with the solvent, the electrostatic charge distribution and surface features indicate the ligand binding potential of the CUG tracks.

INTRODUCTION

The CUG repeats are among the most abundant trinucleotide repeats in human transcripts, and their over-representation in coding regions implies a functional significance of these sequences. In mature mRNAs, the CUG repeat tracts occur most frequently in their protein-coding parts followed by 5′ and 3′ untranslated regions (1). The documented biological functions of CUG repeats in transcripts include modulation of efficiency and accuracy of pre-mRNA splicing (2), mRNA transport (3) and regulation of translation (4,5).

The CUG repeats are better known for the multiple system dysfunctions they cause in the mutated form that occurs in myotonic dystrophy type 1 (DM1) patients (6). The mutation leading to DM1 is the expansion of a CTG repeat, located in the 3′UTR of *dystrophia myotonica* protein kinase (*DMPK*) gene from normal 5–37 repeats to mutated 50–3000 repeats (7). A key feature of the expanded CUG repeats is misregulation of alternative splicing of numerous developmentally regulated transcripts (8). The misregulation is caused by altered interactions of the implicated transcripts with two types of antagonistic splicing regulators: the CUG repeat binding protein (CUG-BP) (9) and the muscleblind like (MBNL) protein (10). The expanded CUG repeats cause a decrease in the cellular level of free MBNL in DM1 cells by its sequestration to nuclear foci (11,12) and at the same time an increase in the CUG-BP level by a yet unknown mechanism (13).

Structural studies of the CUG repeats have begun with the demonstration that short repeat tracts remain single-stranded in the DMPK transcript, whereas longer repeats form hairpins whose stability increases with length (14). The single-stranded CUG repeats are known to bind CUG-BP (9), while the double-stranded stem of the CUG repeat hairpin interacts with MBNL in a length-dependent manner (10). Further biochemical studies provided more information on the sequence specificity of CUG-BP (15,16) and MBNL (17) as they bind to CUG repeats and focused on defining natural targets of MBNL (18,19). It has been indicated that MBNL recognises GC-rich hairpins containing pyrimidine mismatches. In a recently published X-ray structure of zinc-finger domains of the MBNL proteins in complex with single-stranded runs of r(CGCUGU) (20), it has been shown that the protein interacts mainly with the GC elements of the sequence. This structure is relevant to the regulation of

*To whom correspondence should be addressed. Tel: +48 61 852 8503; Fax: +48 61 852 0532; Email: wojtekr@ibch.poznan.pl

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

alternative splicing and perhaps also throws light on the way MBNL recognizes the double-stranded CUG repeats, as they also contain GC steps.

There is however no model yet of the protein interacting with double-stranded CUG runs. Electron microscopic examination revealed the formation of dsRNA by long CUG repeats and confirmed that MBNL bound to the double-stranded stem of CUG-repeat hairpins, while CUG-BP bound to single-stranded repeats (21). Recently, the same method was used to disclose more details of the interaction between the CUG-repeat hairpin and MBNL (19). The melting profiles of CUG-repeat transcripts were analysed and found consistent with a single type of secondary structure (22), and accurate thermodynamic parameters were determined for the U–U mismatches within the duplexes formed by CUG repeats (23). NMR studies also showed that the CUG-repeat fragments adopt a double-stranded form (24). In 2005 the first crystal structure of synthetic RNA, composed of six CUG repeats, was determined with 1.58 Å resolution (25). The structure was originally described as statically disordered and the resulting model consisted of two superimposed duplexes. The double helices contained U–U pairs flanked by G-C pairs, as expected. The duplexes in the crystal lattice stacked end-to-end, forming long pseudo-continuous helices resembling stem structures of long CUG-repeat hairpins. The overall structure was similar to the A-form RNA, as expected, but the disambiguation of the electron density was difficult. It was determined that the distances between the C1 atoms of the paired uridines were ~10 Å but the U–U pairs appeared to lack hydrogen bonds.

In this study we present two crystal structures of RNA containing CUG repeats: a high resolution model of $G(CUG)_2C$ duplex designed by us, and an unambiguous model of $(CUG)_6$ duplex obtained after detwinning the X-ray data previously deposited in the PDB by Mooers *et al*. (25). To our knowledge these two oligomers are to date the only TRED-related RNA molecules whose structures have been analysed empirically in atomic detail. The CUG repeats form regular, well defined structural motifs, whose characteristic hydrogen-bonding pattern, interactions with the solvent, the electrostatic charge distribution and surface features, define their properties and indicate the ligand binding potential of the CUG tracks.

MATERIALS AND METHODS

Synthesis, purification and crystallization of CUG oligoribonucleotides

Oligoribonucleotides were synthesized on an Applied Biosystems DNA/RNA synthesizer using cyanoethyl phosphoramidite chemistry. Commercially available C, G and U phosphoramidites with 2′-O-tetrbutyldimethylsilyl were used for synthesis of RNA (Glen Research, Azco, Proligo). The details of deprotection and purification of oligoribonucleotides were described previously (26). $r(GCUGCUGC)_2$ was dissolved in 5 mM $MgCl_2$ in water to the final RNA concentration of 1 mM and annealed for 5 min at 65°C, then cooled overnight to room temperature. Crystals were grown by the hanging drop/vapour diffusion method at 19°C. Initially, drops contained 2 µl of RNA and 2 µl of reservoir solution (50 mM sodium acetate, pH 5.5, 100 mM $MgCl_2$, 1.5 M Li_2SO_4). Crystals appeared within 2–3 days.

X-ray data collection, structure solution and refinement

X-ray diffraction data were collected at 100 K to the resolution of 1.23 Å from $r(GCUGCUGC)_2$ crystal cryoprotected with 25% glycerol (v/v), on the EMBL X13 beam line at the DESY synchrotron in Hamburg. The data were integrated and scaled using the program suite DENZO/SCALEPACK (27). The space group was assigned as C2, although β was 90°. The X-ray data are summarized in Supplementary Table 1. The structure was solved by molecular replacement using PHASER (28) and refined using Refmac5 (29) from the CCP4 program suite (30). Five percent of reflections were set aside and used for R-free calculation. The last few cycles of the refinement were carried out with SHELXL (31), during which the occupancy factors were refined for the sulphate ions, glycerol and those parts of the RNA model with alternative conformations. The program Coot was used for visualization of electron density maps $2F_o–F_c$ and $F_o–F_c$ and manual rebuilding of the atomic model (32). Solvent water molecules were added by ARP/wARP working in the default solvent building mode (33). Towards the end of the refinement anisotropic temperature factors were refined for all atoms. At the end of the refinement additional few refinement cycles were performed using all data, i.e. including the reflections used for calculating R-free. The final model is summarized in Table 1.

The second RNA model, r(CUGCUGCUGCUGCUG CUG)$_2$, was obtained by detwinning the X-ray structure factors deposited in the PDB (code 1zev) by Mooers *et al*. (25) who originally described the structure as disordered. The structure factor amplitudes were examined for merohedral twinning and the corresponding Patterson function was inspected for evidence of pseudo-translation, using program PHENIX (34). The initial twin fraction was calculated with the aid of the Yeates & Fam UCLA twinning server (35) and subsequently refined together with the atomic model in SHELXL (31).

The helical parameters were calculated using 3DNA (36). Sequence-independent measures were used, based on vectors connecting the C1′ atoms of the paired residues, to avoid computational artefacts arising from non-canonical base-pairing. Program PDB2PQR was used to assign partial charges and radii to atoms of the models, according to the AMBER force field (37). Subsequently, the surface electrostatic potential for the RNA models was calculated with APBS (38). All pictures were drawn in PyMOL v0.99rc6 (39). The coordinates of both crystallographic models have been deposited with the Protein Data Bank (PDB). The accession codes are 3glp for the monoclinic structure and 3gm7 for the rhombohedral structure.

RESULTS

The $[G(CUG)_2C]_2$ model

In the monoclinic structure, the asymmetric unit contains five RNA $G(CUG)_2C$ strands forming two complete RNA duplexes (strands A + B and C + D), while the third duplex is formed by strand E and its symmetry equivalent, related by the 2-fold crystallographic axis (Supplementary Figure 1A). The duplexes stack end-to-end, forming semi-infinite columns parallel to the *a*–*c* lattice plane and inclined at ~45° to the axes *a* and *c*. The model also contains ordered water molecules, two sulphate ions and one glycerol molecule (Table 1).

The $[(CUG)_6]_2$ model

The analysis of the structure factors deposited (pdb code 1zev) by Mooers *et al.* (25) indicated twinning, as described in Supplementary Notes. Refinement of atomic model against the 'perfectly' twinned data using SHELXL (31) resulted in electron density that was largely unambiguous (Figure 1 and Supplementary Figure 2). The asymmetric unit contains one RNA duplex of $(CUG)_6$, strands G and H (Supplementary Figure 1B), and 53 ordered water molecules. The crystal lattice consists of RNA duplexes running parallel to the crystallographic 3-fold axes and stacking end-to-end.

The RNA duplex conformation and base-pairing

In both crystal structures the RNA duplexes are in the A-form. Most of the sugar residues are in the 3′-*endo* conformation, except for seven which have the 2′-*exo* pucker. Sequence-independent helical parameters have been calculated using the C1′ atoms of the base-paired residues. Displacement, angle (inclination between the inter-atomic C1′-C1′ vector and the helix axis) and rise do not indicate any significant effects that can be attributed to the non-canonical base pairing. The average values are 6.7 Å, 13.4°, 2.7 Å, respectively (Supplementary Table 2). Helical twist shows irregularity within the duplex A + B in the monoclinic structure (standard deviation = 8.1°). The values are elevated for both C–G/U–U steps (above 40°) compared to the other steps within this duplex (about 30°). The other duplexes do not show such variability (s.d. = 3.6° for duplex C + D, 2.6° for E + E*—asterisk denotes a symmetry-related molecule) and 3.2° for $[(CUG)_6]_2$. Nevertheless, the average values for the helical twist are very similar for each duplex: 32–34°, which is typical of A-form. The different $[G(CUG)_2C]_2$ duplexes can be superposed with root-mean-square deviation (r.m.s.d.) of atomic coordinates between 0.9 and 1.4 Å. They can also be fitted onto matching segments of the $[(CUG)_6]_2$ model with r.m.s.d. between 1.0 and 1.7 Å.

Table 1. Summary of the models and refinement statistics

	$(GCUGCUGC)_2$	$[(CUG)_6]_2$ by Mooers *et al.* (25)	$[(CUG)_6]_2$ after detwinning
Overall mean B-factor (Å^2)	22.8	28	33.8
Number of reflections: work/test	30062/1602	9965/753	9925/1084
R-value (%)	14.8	21.8	21.9
R-free (%)	18.4	27.9	26.2
RNA atoms	934	1500 (half occupied)	750
Water molecules	194	81 (half occupied)	53
Ligand molecules	2 sulphate, 1 glycerol	–	–
R.m.s.d. in bonds/target (Å)	0.018/0.21	0.011/0.21	0.01/0.02
R.m.s.d. in angles/target	2.75/3.0°	2.06/3.0°	0.028/0.04 Å

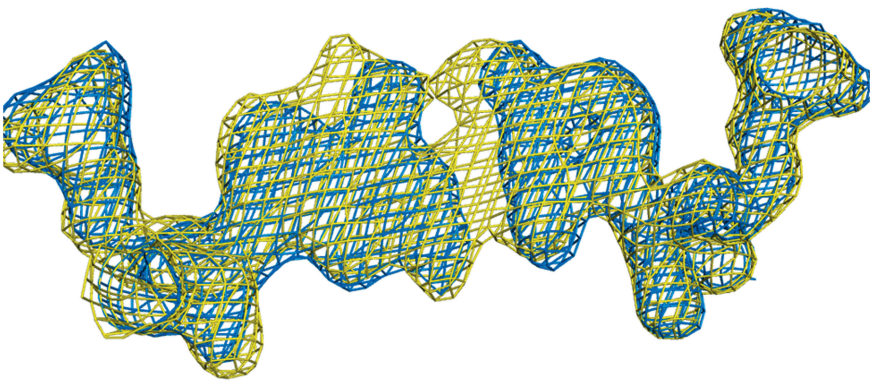

Figure 1. Comparison of the 2Fo–Fc electron density maps calculated using the model deposited by Mooers *et al.* (25) (yellow) and after data detwinning (blue).

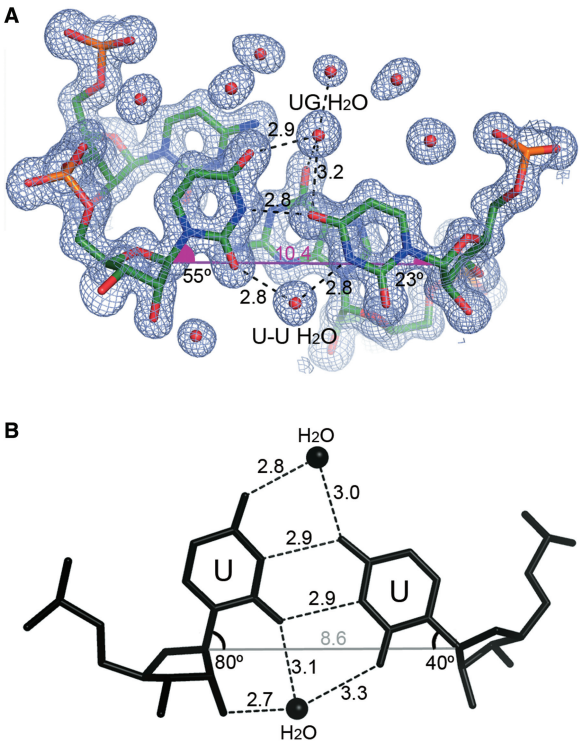

Figure 2. A representative 'stretched U–U pair' with a single H-bond N3-O4, as observed in the monoclinic structure (**A**). All the pairs in both analysed crystal forms show the same conformation. One of the uridines is inclined towards the minor groove, and the λ angle, between the glycosidic bond and the line connecting C1′ atoms (green line), is 30° or less, as opposed to the typical value of 55°. The distance C1′–C1′ for the 'stretched U–U pair' is about 10.4 Å, similar to the average value for an A-helix. The corresponding distance for standard U–U pair (**B**), calculated from all 582 available U(*anti*)-U(*anti*) pairs in the SwS server, is 8.6 Å, and the uridines interact via two H-bonds. Each type of U–U pair is solvated by two water molecules, one in each groove. The interactions of the water in the minor groove are very different between the two types of U–U pairs. The environment of the water in the major groove also changes due to the inclination of one U.

All the observed C–G base pairs form Watson–Crick interactions, while all the U–U pairs interact *via* only one hydrogen bond between the carbonyl O4 atom of one base and the N3 amino group of the second U. The residue accepting the H-bond is inclined towards the minor groove, as indicated by angle λ (between the glycosidic bond and the line joining the base-paired C1′ atoms) (Figure 2A). The value for the inclined bases is small, 30°, compared to the average value for nucleotides of 55°. The inter-strand distance measured between the C1′ atoms of the paired uridines remained typical for A-RNA—about 10.4 Å (the average for the analysed duplexes is 10.5 Å, with standard deviation of 0.2 Å). The base pair opening for all U–U pairs is −7.5°, irrespective of which U is inclined (Supplementary Table 2). The above features are preserved in all the observed U–U pairs. According to the nomenclature introduced by Leontis and Westhof (40) the pairing of uridines could be described as 'U•U *cis* (wobble) W+C+/W+C+', with the additional clarification that there is only one hydrogen bond between the bases. This base pairing can be described as 'stretched U–U wobble'.

Overall, each CUG repeat assumes one of two distinct conformations depending on whether the uridine is inclined towards the minor groove (low λ) or not. In the A+B duplex, both uridines on strand A are inclined, thus the two strands are structurally different. Similarly, in the C+D duplex both uridines of strand D are inclined. The duplex E+E* is crystallographically symmetric and has the second U inclined. In the rhombohedral structure the first and the third U of strand G are inclined (and the remaining four U of strand H).

RNA hydration and ligand interactions

Ordered water molecules are associated with the U–U pairs, forming a characteristic pattern in both grooves (Figure 2A). In the minor groove one water molecule H-bonds with the N3 amino group of the inclined uridine (low λ) and with O2 of the other U. This pattern is observed for all six U–U pairs in the monoclinic structure and for four of the six U–U pairs in the detwinned rhombohedral structure. In the major groove, a water molecule is bound to the O4 carbonyl of the non-inclined U and to the O6 carbonyl of the nearest guanosine on the opposite strand. These interactions are observed in all cases in the monoclinic structure and in three U–U pairs in the rhombohedral structure.

C–G hydration also exhibits regularity (Supplementary Figure 2). Most guanosines in the high resolution structure are observed to interact with four ordered water molecules. Two of them are in the major groove: one H-bonded to the N7 group and the other to the O6 carbonyl atom. The two water molecules in the minor groove interact with the *exo*-amino and the imine groups. The cytosines are typically associated with two water molecules, one in each groove. In three cases the C *exo*-amino group in the major groove interacts with a sulphate anion or a glycerol molecule instead of water. One of the sulphate ions is located between the A+B duplex and its symmetry-equivalent duplex, in the space between two sugar moieties. Two of its oxygen atoms interact each with a different O2′ atom: from 3U of chain B, and 2C, chain A*. Another sulphate oxygen is H-bonded to the N2 *exo*-amino group of 1G A*.

Two ligands bind in the major groove in an ordered manner: a glycerol molecule is bound to duplex A+B and the second sulphate ion interacts with C+D. Each ligand forms two hydrogen bonds: with the amino group of 2C (chain A for glycerol or D for sulphate) and with the nearby 'U–G water' of the major groove, associated with 3U–6U pair. Each ligand is half-occupied and associated with a local disorder in the RNA strands (sulphate with chain C and glycerol with B) and interacts with one of two distinct conformers. The two strands are in contact in the crystal lattice and their conformations are co-related. In consequence, either the sulphate can bind to A+B duplex or glycerol to C+D (Supplementary Figure 3). In addition, the third OH group of glycerol interacts with the *exo*-amino group of 5C in chain B.

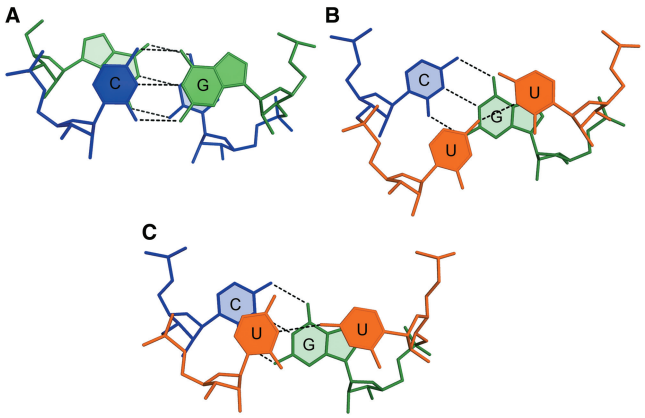

Figure 3. Stacking interactions for GC/GC step (**A**) and two kinds of CU/UG steps (**B** and **C**) depending on the conformation of the U–U pair.

Stacking interaction

Three kinds of intramolecular stacking interactions can be distinguished in the analysed structures: two for the CU/UG step, depending on the conformation of the U–U pair, and one for the GC/GC step (Figure 3). The latter, characterized by Watson–Crick pairing and typical for A-form also shows extensive stacking overlaps (Figure 3A). The steps involving the non-canonical pairing have more limited stacking interactions. In all observed cases, uridines stack against the five-membered ring of the neighbouring guanosines, but stacking of U against C depends on the conformation of U. If the U is inclined towards the minor groove, there is no interaction with the neighbouring C, only limited stacking with the six-membered ring of G from the opposite strand (Figure 3B). If the uridine is not inclined, it interacts weakly with both C and the opposite G (Figure 3C).

Surface of electrostatic potential

The surface of potential shows a similar charge distribution for all structures (Figure 4). The major groove is predominantly electronegative with patches of positive potential due to amino groups of cytosines. These are the binding sites of glycerol and sulphate. The potential of the minor groove is complex and forms a pattern of alternating bands of positive and negative potential along the direction of the helix axis. The negative bands are formed by the electropositive atoms of stacking C, G and U, and the positive bands by the carbonyl oxygen atoms of U and C residues. The carbonyl groups of the inclined uridines protrude out of the minor groove and form bulges with high negative potential.

DISCUSSION

The two presented models reveal characteristic features of RNA duplexes containing CUG repeats. The shorter oligomers show the high resolution detail, while the longer molecule, analysed at lower resolution, contains more

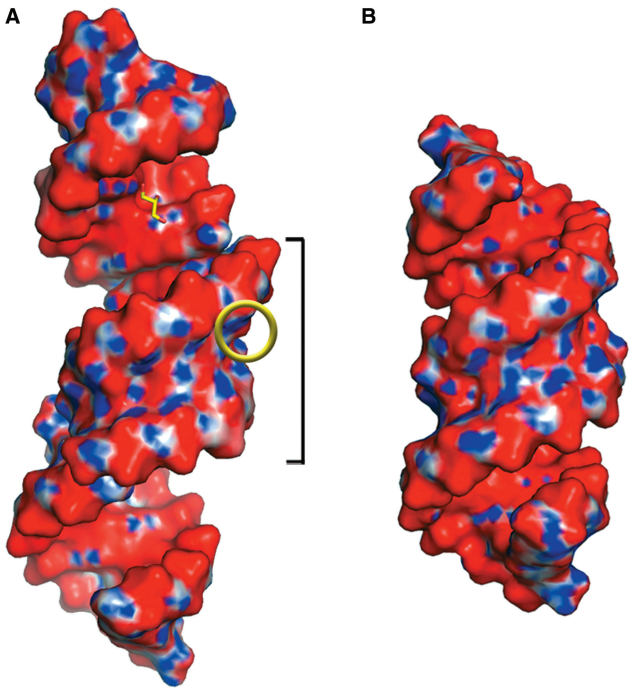

Figure 4. The electrostatic potential surface for (**A**) the monoclinic structure, showing the three consecutive duplexes in the asymmetric unit (the middle duplex is indicated by a brace) and (**B**) the detwinned rhombohedral structure. Red is negative, blue is positive. A glycerol molecule (sticks) is shown interacting with electropositive patches in the major groove. A bulge in the minor groove formed by the O2 carbonyl group of one inclined uridine is indicated by a ring.

repeats and therefore corresponds more closely to the biological trinucleotide runs.

Detwinning of the X-ray diffraction data (pdb id 1zev) enabled us to interpret the structure factor amplitudes in terms of a single unambiguous model, instead of two overlapping models presented before (Figure 1). Details of solvation, previously unobserved, have now appeared. Interpreting perfectly twinned data (twin fraction 0.5) is difficult and laden with uncertainty, because the structure factors cannot be proportioned algebraically. Nevertheless, it is possible to refine an atomic model against such data. The final model we obtained is in good agreement with electron density maps, is stereochemically valid, shows reasonable H-bonding interactions and is consistent with the related high-resolution structure in terms of helical parameters and details of base pairing and hydration. The consistency of the structures of different lengths, obtained under different crystallization conditions and localized in different packing environments, indicates that the observed features represent a major stable form characteristic of the sequence rather than external factors. The lack of clear 'end-effects' at the duplex termini can be explained by the close packing of molecules that form pseudo-infinite helices.

Given that the crystal structure of the CUG repeat appears to be independent of the length of the oligomer in which it is embedded, it is hard to explain why longer CUG tracks are less sensitive to lead-induced cleavage

(14). There are two possibilities. The structures analysed crystallographically cover a relatively narrow range of two to six repeats, whereas the lead digestion experiments included up to 49 repeats. It is possible that structural differences become apparent only when short sequences are compared with much longer ones. Alternatively, it is possible that the sensitivity to digestion of CUG tracks depends on the hairpin loop that was present in the molecules studied by Napierala and Krzyzosiak (14) and absent in the X-ray study.

Despite the recurrence of the U–U pairs, all four helices in the two crystal structures retain the A-form, as evidenced by the predominance of the C3′-*endo* conformation and regular inter-strand distance of 10.5 Å. The U–U pairs are accommodated in the duplex without a significant effect on the strand separation, with one U strongly inclined towards the line connecting the opposite C1′ atoms (λ 30°) and with a single H-bond between the uridines (Figure 2A). A comparison with U–U pairs deposited in PDB reveals significantly shorter C1′–C1′ distance: 8.6 Å on average, with standard deviation 0.27 Å, based on 582 U–U pairs (Figure 2B) extracted by the SwS web server (41). The common U–U pairs have relatively large λ angles (40–80°) and there are two H-bonds (O4–N3 and N3–O2). The unusually wide separation between the uridines in the $(CUG)_n$ duplexes and the single H-bond between them can be explained by the stabilizing effect of the sturdy canonical C–G pairs interleaved with the U–U pairs.

The single H-bond of the 'stretched U–U pair' does not exhaust the bonding potential of the paired uridines and additional bonds are formed with water molecules: the 'U–U water' in the minor groove and the 'UG bridging water' in the major groove (Figure 2A). The two solvent molecules form a characteristic structural pattern around the U–U pairs and deserve to be considered a stable part of the structure. At the same time, they point to the specific H-bonding capacity of the CUG repeat. The solvation pattern in the minor groove strongly depends on the interactions between the uridines. In the 'stretched U–U pair' the N3 atom of the inclined U (low λ) is H-bonded to the water molecule (the 'U–U water' in Figure 2A), whereas in the typical U–U pair the nitrogen interacts with the second U and the water interacts with O2 (Figure 2B). Thus in the case of the 'stretched pair', the U–U water in the minor groove has to be a donor and an acceptor, while typically it is a donor of two bonds. In the major groove the O4 carbonyl oxygen of the inclined U is less accessible to the solvent than in the typical U–U pair. The water makes a clear H-bond with the non-inclined U but the H-bond with the other O2 appears weaker (3.2 Å). The second favoured acceptor seems to be the guanosyl O6 atom of the neighbouring C–G pair. In the typical U–U pair both carbonyl O4 atoms are easily accessible to solvent and accept two H-bonds from a single water molecule (Figure 2B). The solvent structure around biological molecules reveals their potential for interactions with ligands and can be a useful guide in designing pharmacophores. In the monoclinic crystal, the ligands (glycerol or sulphate) bound in the major groove interact with the 'UG water' (between 3U and 4G) rather than displace it. The water molecule, together with NH2 from 2C, provides a specific environment for accepting an H-bond from the glycerol (in duplex A + B) or sulphate (C + D). The common feature in both ligands is a hydroxyl group, which, having both capacities, accepts an H-bond from the NH2 group and donates one to the 'UG water' (Supplementary Figure 3). One could also consider the possibility that the ordered water molecules are replaced by ligands. The UG water in the major groove donates two hydrogen bonds to the two O4 carbonyl oxygen atoms, which means that any other group binding specifically in its position should possess similar H-bonding capacity, e.g. an amino group. The U–U water donates one H-bond and accepts one. Such H-bonding capacity is shared by hydroxyl or imine groups.

The wobble U–U interaction and the way the pair stacks with other base pairs have consequences for the accessible surface of the grooves and the surface electrostatic potential. The inclined base forms a clear indentation in the major groove while it bulges out in the minor groove. The electrostatics potential depends on which of the carbonyl oxygen atoms forms a H-bond and is therefore obscured (Figure 4).

There is evidence that the U–U pairs within the CUG repeats are central to the recognition by proteins that control the appropriate splicing of mRNA. Replacing the U–U by Watson–Crick pair almost completely abolishes MBNL binding (18). The analysis of the crystal structure indicates that the key to the specific properties of the CUG repeat is the 'stretched wobble' of the U–U pairs. The consequence of this conformation is an environment that can be clearly mapped in terms of electrostatics, surface features of the minor and the major grooves, and specific H-bonding potential—and it probably determines the possible interactions with proteins and smaller ligands.

An interesting feature of the U–U interaction is its structural asymmetry. Either one of the bases can be inclined towards the minor groove, which breaks the symmetry of the chemically symmetric CUG duplex. It is unclear what determines which uridines must be inclined, but the structures we have analysed show both possibilities realised along the sequence. Of the three short duplexes, one is symmetric (one U inclines on each strand) and two are asymmetric. In the $[(CUG)_6]_2$ structure the U–U alternate in a seemingly random manner. The consequence for longer RNA chains is that despite the simple, palindromic nature of duplexes made up of CUG repeats, the two alternative modes of U–U wobble vastly expand their repertoire in terms of available three-dimensional structures. If each CUG repeat can take either conformation, the number of possible conformations of longer duplexes grows rapidly as the number of repeats (N) expands [$2^N/2$ for odd N; $(2^N + 2^{N/2})/2$ for even N]. This has interesting implications for the structure of the RNA and its interactions with ligands. For a repetitive structure that simply increased in length, the affinity for ligand binding would be expected to grow proportionally; whereas a flexible structure has a much broader and rapidly growing range of possibilities to interact as its size expands. Only a modest increase in binding affinity for MBNL has been

observed for expanded CUG runs (18). On the other hand, pathogenesis-related nuclear foci are formed by the association of mRNA transcripts containing expanded CUG runs together with MBNL and related proteins (11,12). The condensation of cell components is a cooperative process and is difficult to explain in terms of the usual properties of the constituent parts, which do not normally exhibit tendencies to aggregate. The emergent structural richness of expanding CUG repeats may be the key to explaining the formation of nuclear foci.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Z. Dauter for enlightening discussion, Dr Andrew R. Jones and Ms Erika Lenz for help in preparing the manuscript and Mr Heinz-Dieter Genz for an excellent logistics support.

FUNDING

Polish Ministry of Science and Higher Education (N-N301-0171634, PBZ-MNiI-2/1/2005 and PBZ-KBN-124/P05/2004); European Community Research Infrastructure Action under the FP6 'Structuring the European Research Area Programme' (contract number RII3/CT/2004/5060008). Funding for open access charge: Polish Ministry of Science and Higher Education.

Conflict of interest statement. None declared.

REFERENCES

1. Jasinska,A., Michlewski,G., de Mezer,M., Sobczak,K., Kozlowski,P., Napierala,M. and Krzyzosiak,W.J. (2003) Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Res.*, **31**, 5463–5468.
2. Philips,A.V., Timchenko,L.T. and Cooper,T.A. (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science*, **280**, 737–741.
3. Taneja,K.L., McCurrach,M., Schalling,M., Housman,D. and Singer,R.H. (1995) Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. *J. Cell Biol.*, **128**, 995–1002.
4. Raca,G., Siyanova,E.Y., McMurray,C.T. and Mirkin,S.M. (2000) Expansion of the (CTG)(n) repeat in the 5′-UTR of a reporter gene impedes translation. *Nucleic Acids Res.*, **28**, 3943–3949.
5. Sasagawa,N., Saitoh,N., Shimokawa,M., Sorimachi,H., Maruyama,K., Arahata,K., Isiura,S. and Suzuki,K. (1996) Effect of artificial (CTG) repeat expansion on the expression of myotonin protein kinase (MtPK) in COS-1 cells. *Biochim. Biophys. Acta*, **1315**, 112–116.
6. Groenen,P. and Wieringa,B. (1998) Expanding complexity in myotonic dystrophy. *Bioessays*, **20**, 901–912.
7. Brook,J.D., McCurrach,M.E., Harley,H.G., Buckler,A.J., Church,D., Aburatani,H., Hunter,K., Stanton,V.P., Thirion,J.P., Hudson,T. *et al.* (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, **68**, 799–808.
8. Ranum,L.P. and Cooper,T.A. (2006) RNA-mediated neuromuscular disorders. *Ann. Rev. Neurosci.*, **29**, 259–277.
9. Timchenko,L.T., Timchenko,N.A., Caskey,C.T. and Roberts,R. (1996) Novel proteins with binding specificity for DNA CTG repeats and RNA CUG repeats: implications for myotonic dystrophy. *Hum. Mol. Genet.*, **5**, 115–121.
10. Miller,J.W., Urbinati,C.R., Teng-Umnuay,P., Stenberg,M.G., Byrne,B.J., Thornton,C.A. and Swanson,M.S. (2000) Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J.*, **19**, 4439–4448.
11. Fardaei,M., Rogers,M.T., Thorpe,H.M., Larkin,K., Hamshere,M.G., Harper,P.S. and Brook,J.D. (2002) Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. *Hum. Mol. Genet.*, **11**, 805–814.
12. Mankodi,A., Urbinati,C.R., Yuan,Q.P., Moxley,R.T., Sansone,V., Krym,M., Henderson,D., Schalling,M., Swanson,M.S. and Thornton,C.A. (2001) Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. *Hum. Mol. Genet.*, **10**, 2165–2170.
13. Timchenko,N.A., Wang,G.L. and Timchenko,L.T. (2005) RNA CUG-binding protein 1 increases translation of 20-kDa isoform of CCAAT/enhancer-binding protein beta by interacting with the alpha and beta subunits of eukaryotic initiation translation factor 2. *J. Biol. Chem.*, **280**, 20549–20557.
14. Napierala,M. and Krzyzosiak,W.J. (1997) CUG repeats present in myotonin kinase RNA form metastable "slippery" hairpins. *J. Biol. Chem.*, **272**, 31079–31085.
15. Mori,D., Sasagawa,N., Kino,Y. and Ishiura,S. (2008) Quantitative analysis of CUG-BP1 binding to RNA repeats. *J. Biochem.*, **143**, 377–383.
16. Takahashi,N., Sasagawa,N., Suzuki,K. and Ishiura,S. (2000) The CUG-binding protein binds specifically to UG dinucleotide repeats in a yeast three-hybrid system. *Biochem. Biophys. Res. Commun.*, **277**, 518–523.
17. Kino,Y., Mori,D., Oma,Y., Takeshita,Y., Sasagawa,N. and Ishiura,S. (2004) Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. *Hum. Mol. Genet.*, **13**, 495–507.
18. Warf,M.B. and Berglund,J.A. (2007) MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. *RNA*, **13**, 2238–2251.
19. Yuan,Y., Compton,S.A., Sobczak,K., Stenberg,M.G., Thornton,C.A., Griffith,J.D. and Swanson,M.S. (2007) Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res.*, **35**, 5474–5486.
20. Teplova,M. and Patel,D.J. (2008) Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat. Struct. Mol. Biol.*, **15**, 1343–1351.
21. Michalowski,S., Miller,J.W., Urbinati,C.R., Paliouras,M., Swanson,M.S. and Griffith,J. (1999) Visualization of double-stranded RNAs from the myotonic dystrophy protein kinase gene and interactions with CUG-binding protein. *Nucleic Acids Res.*, **27**, 3534–3542.
22. Tian,B., White,R.J., Xia,T., Welle,S., Turner,D.H., Mathews,M.B. and Thornton,C.A. (2000) Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA*, **6**, 79–87.
23. Broda,M., Kierzek,E., Gdaniec,Z., Kulinski,T. and Kierzek,R. (2005) Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry*, **44**, 10873–10882.
24. Leppert,J., Urbinati,C.R., Hafner,S., Ohlenschlager,O., Swanson,M.S., Gorlach,M. and Ramachandran,R. (2004) Identification of NH ... N hydrogen bonds by magic angle spinning solid state NMR in a double-stranded RNA associated with myotonic dystrophy. *Nucleic Acids Res.*, **32**, 1177–1183.
25. Mooers,B.H., Logue,J.S. and Berglund,J.A. (2005) The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc. Natl Acad. Sci. USA*, **102**, 16626–16631.
26. Xia,T., SantaLucia,J. Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.
27. Otwinowski,Z and Minor,W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.*, **276**, 307–325.

28. Storoni,L.C., McCoy,A.J. and Read,R.J. (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 432–438.
29. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D. Biol. Crystallogr.*, **53**, 240–255.
30. Collaborative Computational Project 4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **50**, 760–763.
31. Sheldrick,G.M. and Schneider,T.R. (1997) SHELXL: high-resolution refinement. *Methods Enzymol.*, **277**, 319–343.
32. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 2126–2132.
33. Lamzin,V.S. and Wilson,K.S. (1993) Automated refinement of protein models. *Acta Crystallogr. D. Biol. Crystallogr.*, **49**, 129–147.
34. Adams,P.D., Grosse-Kunstleve,R.W., Hung,L.W., Ioerger,T.R., McCoy,A.J., Moriarty,N.W., Read,R.J., Sacchettini,J.C., Sauter,N.K. and Terwilliger,T.C. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 1948–1954.
35. Yeates,T.O. (1997) Detecting and overcoming crystal twinning. *Methods Enzymol.*, **276**, 344–358.
36. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.J., Neidle,S., Shakked,Z. *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
37. Dolinsky,T.J., Nielsen,J.E., McCammon,J.A. and Baker,N.A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
38. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.
39. DeLano,WL (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, Palo Alto, CA, USA.
40. Leontis,N.B. and Westhof,E. (1998) Conserved geometrical base-pairing patterns in RNA. *Q. Rev. Biophys.*, **31**, 399–455.
41. Auffinger,P. and Hashem,Y. (2007) SwS: a solvation web service for nucleic acids. *Bioinformatics.*, **23**, 1035–1037.