2602–2613 Nucleic Acids Research, 2023, Vol. 51, No. 6 Published online 2 March 2023
https://doi.org/10.1093/nar/gkad117

Structural polymorphism driven by a register shift in a CGAG-rich region found in the promoter of the neurodevelopmental regulator *AUTS2* gene

Aleš Novotný[1,2], Janez Plavec[1,2,3,*] and Vojč Kocman[1,3,*]

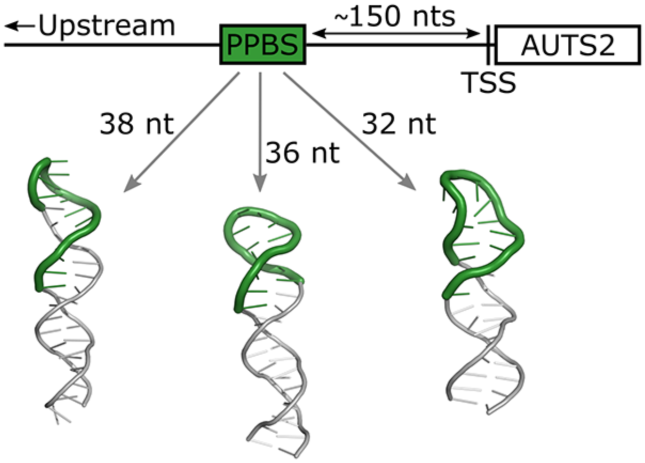
[1]Slovenian NMR Centre, National Institute of Chemistry, Ljubljana SI-1000, Slovenia, [2]Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana SI-1000, Slovenia and [3]EN-FIST Centre of Excellence, Ljubljana SI-1000, Slovenia

Received December 01, 2022; Revised February 03, 2023; Editorial Decision February 06, 2023; Accepted February 09, 2023

ABSTRACT

The *AUTS2* gene has been shown to influence brain development by controlling the number of neurons, promoting the growth of axons and dendrites and regulating neuronal migration. The expression of two isoforms of AUTS2 protein is precisely regulated and misregulation of their expression has been correlated with neurodevelopmental delay and autism spectrum disorder. A CGAG-rich region, which includes a putative protein binding site (PPBS), d(AGCGAAAGCACGAA), was found in the promoter region of *AUTS2* gene. We show that oligonucleotides from this region adopt thermally stable non-canonical hairpin structures stabilized by G:C and sheared G:A base pairs arranged in a repeating structural motif we termed CGAG block. These motifs are formed consecutively, in a way that exploits a shift in register throughout the whole CGAG repeat to maximize the number of consecutive G:C and G:A base pairs. The differences in CGAG repeat shifting affect the structure of the loop region, where PPBS residues are predominantly located, specifically the loop length, types of base pairs and the pattern of base-base stacking. Finally, we propose a previously unexplored mechanism, by which different folds in the CGAG-rich region could cause a switch in expression between the full-length and C-terminal isoforms of AUTS2.

GRAPHICAL ABSTRACT



INTRODUCTION

It is well established that DNA is not necessarily involved just in storage of genetic information encoded in the sequence of nucleotides and carried from one generation to the other in the form of double helix. A variety of non-canonical structures were identified with structural characteristics strongly dependent on the nucleotide sequence and external conditions such as pH, type of ions and their concentration (1–5). Formation of such structures was shown to affect vital cellular processes such as DNA replication, gene transcription, and elongation of telomeres indicating a potential mechanism for natural or ligand-induced regulation (6–9). Further supporting their role in biological processes, different studies employing structure-specific antibodies and immunofluorescence showed that the formation of non-canonical DNA structures is dependent on the stage in the cell cycle (10,11). Sequences predisposed to form structures other than B-DNA were also linked to increased mutagenesis and genetic instability (12–14).

[*]To whom correspondence should be addressed. Tel: +386 1 4760 416; Fax: +386 1 4760 300; Email: vojc.kocman@ki.si
Correspondence may also be addressed to Janez Plavec. Tel: +386 1 4760 353; Email: janez.plavec@ki.si

Non-canonical structures differ from canonical B-DNA double helix in various structural elements, which makes them potential targets for protein recognition and disease therapy (15–19).

More than 50, mostly neurological, diseases were associated with the expansion of short (3–12 nt) nucleotide repeats scattered across both coding and non-coding regions of genes (20,21). A significant number of nucleotide repeats were shown to adopt various types of non-canonical structures *in vitro* including hairpins, cruciform, i-motifs and G-quadruplexes and there is a constantly growing body of data that support the formation of these structures *in vivo* (22–25). Detailed knowledge of the relationship between the varying number of repeats and the corresponding structure of the DNA is crucial for understanding the pathology at the molecular level.

Misregulations in the Autism Susceptibility gene 2 (*AUTS2*) have been linked with a genetic condition commonly referred to as AUTS2 syndrome, which is characterized by a wide range of phenotypes including intellectual disability, developmental delay, microcephaly, feeding difficulties, attention deficit hyperactivity disorder and autism spectrum disorder (26–30). Recent studies of biological functions of AUTS2 protein identified its role as the crucial transcription regulator during brain development (31,32). AUTS2 is also proposed to be involved in RNA metabolism by associating with RNA-binding protein complexes and in regulation of cytoskeletal dynamics. The *AUTS2* gene spans 1.2 million nts and comprises 19 exons as well as transcription start sites (TSS) located upstream of exon 1 (TSS-1) and exon 9 (TSS-2). The transcription starting at TSS-1 and TSS-2 results in a full-length 1259 amino-acid (aa) isoform and a C-terminal 711-aa isoform, respectively (26). A study employing mouse embryonic stem cells showed that the full-length isoform is expressed in undifferentiated cells, which is replaced by the C-terminal isoform during differentiation of neurons. In developing human brain, the expression of C-terminal isoform is dominant, whereas in the adult human brain the expression of both isoforms is decreased to low levels. The overexpression of the full-length isoform was shown to delay the neuronal differentiation in mouse embryonic cells (33). In developing mouse brain, the full-length isoform was detected in all embryonic stages as well as after birth, but the C-terminal isoform was expressed only transiently during the embryonic stages and was barely detected postnatally (34). The molecular mechanism responsible for switching between the expression of the full-length and C-terminal isoforms remains unknown.

In this study, we set out to structurally characterize a CGAG-rich region of the *AUTS2* promoter, which could play a role in switching between the expression of the full-length and C-terminal isoforms of the AUTS2 protein. In the genome, the 60 nt long CGAG-rich region (here termed A60) is found approximately 150 nts upstream of the TSS-1. A60 is comprised of a 14 nt long putative protein binding site (PPBS), 5'-d(AGCGAAAGCACGAA)-3', which is flanked by two and six CGAG repeats on its 5'- and 3'-sides, respectively. As A60 may play a role in initiating and regulating transcription at TSS-1, we set out to obtain insights into its structure. Based on reports on similar repeats (35–39), we formulated a hypothesis that A60 adopts a hairpin structure stabilized by canonical G:C and non-canonical sheared G:A base pairs with PPBS positioned in the loop region. However, due to structural flexibility and polymorphism of A60, we were unable to interpret the experimental data in terms of a single, well-defined 3D structure. Therefore, we prepared a set of truncated oligonucleotides comprising different number of CGAG repeats surrounding the PPBS that enabled us to uncover a complex interplay of their structural and kinetic features that define the conformations adopted by the number of CGAG repeats embedding the PPBS. Utilizing a divide and conquer approach, we obtained important insights into structural and dynamic features of CGAG repeats that are involved in making PPBS available for interaction with transcription regulating proteins. Finally, we deduced certain structural features of A60, which could help in understanding the recognition of PPBS and the mechanism of the expression switch.

MATERIALS AND METHODS

Preparation of DNA oligonucleotides

Natural isotope abundance oligonucleotides as well as residue-specific low-enrichment (10%) ^{13}C- and ^{15}N-labelled DNA oligonucleotides were synthesized on K&A Laborgeraete GbR DNA/RNA Synthesizer H-8 using standard phosphoramidite chemistry in DMT-off or DMT-on mode. The oligonucleotides synthesized in DMT-on mode were purified using Glen-Pak DNA cartridges according to the instructions of the supplier. Prior to desalting, the solutions of oligonucleotides were heated to 90°C for 5 min and left to cool at room temperature. The desalting was performed by multiple rounds of centrifugal concentration using Amicon Ultra-15 Centrifuge filters with 3 kDa cutoff and deionized Milli-Q water. Potassium phosphate buffer solution (KPi) was used in the last two rounds of desalting. The concentration of the DNA in the final solution was determined by UV absorbance at 260 nm at 90°C measured on Varian CARY-100 BIO UV-VIS spectrometer.

Nuclear magnetic resonance (NMR) spectroscopy

All NMR data were collected on Bruker Avance Neo 800 MHz and 600 MHz NMR spectrometers equipped with triple resonance HCN and quadruple resonance HCNP cryoprobes. Oligonucleotide strand concentrations ranged from 0.3 to 1.2 mM in 90%/10% H_2O/D_2O or 100% D_2O buffered by 20 mM KPi at pH 7.2. The signal of water was suppressed using the excitation sculpting pulse scheme. The ^{13}C-edited heteronuclear single quantum coherence spectra (^{1}H–^{13}C HSQC) were measured on isotopically labelled samples. 2D nuclear Overhauser effect spectroscopy (NOESY) spectra were acquired with 150 or 200 ms mixing times to obtain estimates of interproton distances. Double quantum filtered ^{1}H–^{1}H (DQF-COSY) and constant time ^{1}H–^{31}P (^{1}H-^{31}P COSY) correlation spectroscopy experiments were conducted to obtain the information regarding sugar puckers and B^I/B^{II} conformations of the phosphate backbone, respectively. The NMR spectra were analyzed using TopSpin (Bruker), MestreNova (Mestrelab Research) and NMRFAM-Sparky softwares (40).

Circular dichroism (CD) spectroscopy

CD wavelength scans were carried out on an Applied Photophysics Chirascan spectrophotometer. The experiments were acquired at 0.3–0.4 mM DNA strand concentration in 0.1 mm quartz cuvettes at 25°C with a 1 nm spectral bandwidth, 60 nm·min^{-1} scan rate, 220–330 nm spectral range, and 1 nm steps. Scans were repeated 3 times and averaged. The resulting curve was smoothed using a Savitzky-Golay 20-point quadratic function. All manipulation was done in Origin 2018.

UV melting experiments

UV melting experiments were carried out on Agilent Cary 3500 UV-Vis spectrophotometer with the Cary Win UV Thermal program in 1.0 cm path length quartz cuvettes. The samples were prepared from the same stock solution as NMR samples and diluted to 5 μM DNA concentration by 20 mM KPi buffer of pH 7.2. The samples were heated and cooled in a temperature range from 10 to 90°C at a rate of 0.5°C·min^{-1}. The absorbance was monitored at 260 nm in 0.5°C steps. The heating and cooling were repeated twice and only the second round was used for the analysis. The temperatures of mid-transition ($T_{1/2}$) were determined from a zero-crossing of the second derivative of heating/cooling curves, which were smoothened using a Savitzky-Golay 20-point quadratic function. The error in determination of $T_{1/2}$ is estimated to be ±1°C based on the repeated measurements. The data processing was done in Origin 2018.

Molecular modelling

All molecular dynamics (MD) simulations were performed in Amber 2020 software suite using CUDA version of pmemd module and OL15 forcefield (41–43). The initial rod-like structures were generated from the nucleotide sequence in the *tleap* module. The structural ensembles were obtained using a simulated annealing (SA) protocol with NMR-derived restraints, *vide infra*. The SA was carried out in implicit solvent (igb = 2) with Langevin temperature control (ntt = 3), an integration step of 1 fs and SHAKE restraining bonds involving hydrogen (ntc = 2) (44–46). During the force evaluation step, the bond interactions involving hydrogen atoms were excluded (ntf = 2). The default value of 9999 Å was used for the non-bonded cutoff. The SA was run for 2 ns in three stages – heating to 1500 K for 50 ps, equilibration of temperature at 1500 K for 250 ps and cooling to 0 K for 1700 ps. In total, the SA was repeated 100 times yielding an ensemble of 100 structures. The ensemble was refined by exclusion of structures having maximum distance violation larger than 0.3 Å and maximum torsion angle violation larger than 10°. Ten structures satisfying the above-described criteria were selected based on the lowest sum of potential and restraint energy (lowest energy) yielding the refined ensemble of structures. The lowest energy structure of each refined ensemble was chosen as a representative model.

Interproton distances were calculated from the volumes of NOE cross-peaks using an average volume of well-resolved cytosine H5-H6 NOE cross-peaks and H5-H6 distance (2.45 Å) as a reference. To reduce the spectral overlap and simplify the integration of cross-peaks, the NOESY spectra were acquired on samples of oligonucleotides in 100% D_2O with mixing time of 150 ms at 298 K (A36 and A38) and 318 K (A32_mod). The calculated interproton distances were divided into three categories (strong: 1.8–3.6 Å, medium: 2.6–5.0 Å and weak: 3.5–6.5 Å), which were used as distance restraints during SA protocol. Additional distance restraints were obtained from NOESY spectra recorded in 90%/10% H_2O/D_2O at 278 K (A38) and 273 K (A36 and A32_mod). These NOEs were included as restraints with broader zero-penalty plateau (1.8–6.5 Å). The hydrogen bonds were restrained based on the NMR data obtained at low temperatures. The χ-torsion angles were restrained in 25°–95° range for purines in *syn*-conformation, in 200°–280°range for purines in *anti*-conformation and in 170°–310° for pyrimidines in *anti*-conformation. The sugar puckers were restrained using pseudorotation phase angle in 144°–180° range for South-type (*C2'-endo*) and in 0°–36° range for North-type (*C3'-endo*) conformation, which was recalculated to a set of five torsion angles using standard AMBER tools. Chirality restraints, employed in a form of improper torsion angles derived from the initial structures using standard AMBER tools, were used to avoid changes in configuration of atoms at high temperature stages of SA. The calculations of the final ensemble of structures employed restraints with the following force constants: 20 kcal·mol^{-1}·Å^{-2} for NOE-derived distances and H-bonds, 50 kcal·mol^{-1}·rad^{-2} for sugar puckers and χ-torsion angles and 100 kcal·mol^{-1}·rad^{-2} for chirality restraints.

RESULTS

Oligonucleotides A36 and A38 adopt hairpins dominated by G:C and G:A base pairs exhibiting characteristic NMR features

Examination of the downfield region of the 1D ^1H NMR spectra of A60 revealed two well-separated clusters of signals with chemical shifts around δ ∼12.7 and ∼10.1 ppm, which are characteristic for imino-protons of guanine residues involved in canonical G:C and non-canonical base pairs, respectively (Figure 1A). Further analysis with the use of 2D ^1H–^1H NOESY and ^1H–^{13}C HSQC spectra indicated that A60 adopts either multiple conformations or comprise dynamic segments, which make structural determination challenging (Supplementary Figure S1). To obtain insights into the structural features of A60, we prepared six truncated variants, which differ in the number of CGAG repeats flanking the PPBS at the 5'- and 3'-ends (Table 1). All variants fold into secondary structures, which is evident from the presence of imino-proton signals indicating base pairing (Figure 1A). CD spectra of all variants show negative bands at ∼245 nm and positive bands at ∼275 nm with shoulders at ∼290 nm (Supplementary Figure S2), which are similar to d(GAGC)$_5$ oligonucleotide (47). All truncated variants as well as A60 show very similar imino-proton chemical shifts and superimposable CD spectra. These data suggest that the truncated variants and A60 comprise a similar structural motif. All oligonucleotides adopt thermally stable structures with T_m ≥50°C (Table 1, Supplementary Figure S3). For further structural analysis, we selected three variants A38, A36 and A32 exhibiting both sharp signals

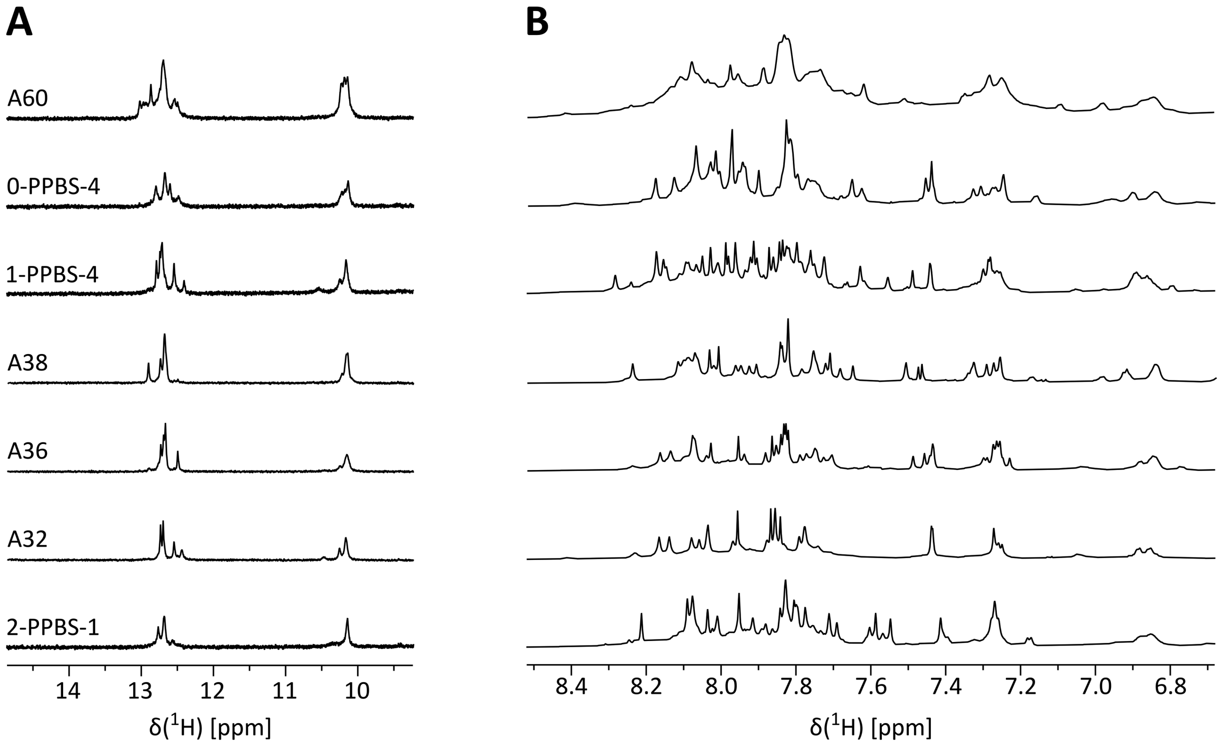


Figure 1. (**A**) Imino and (**B**) aromatic regions of 1D ^{1}H NMR spectra of A60 and its truncated variants. Both regions originate from the same NMR experiment. The vertical scale of the imino regions in panel A was increased 4×. The spectra were recorded at 0.3–0.6 mM strand concentration in 90%/10% H_2O/D_2O buffered by 20 mM KPi at pH 7.2 and 25°C on a 600 MHz NMR spectrometer.

Table 1. A list of oligonucleotides derived from A60

2'-Deoxyribonucleotide[a]	Abbreviation	Length (nt)	T_m (°C)
(GC)$_4$ACT(CGAG)$_2$CGAG-CGAA-AGCA-CGAA-(CGAG)$_6$C	A60	60	77
CGAG-CGAA-AGCA-CGAA-(CGAG)$_4$	0-PPBS-4	32	62
CGAG-CGAG-CGAA-AGCA-CGAA-(CGAG)$_4$	1-PPBS-4	36	64
(CGAG)$_2$CGAG-CGAA-AGCA-CGAA-(CGAG)$_3$CG	A38	38	69
(CGAG)$_2$CGAG-CGAA-AGCA-CGAA-(CGAG)$_3$	A36	36	64
(CGAG)$_2$CGAG-CGAA-AGCA-CGAA-(CGAG)$_2$	A32	32	60
(CGAG)$_2$CGAG-CGAA-AGCA-CGAA-CGAG	2-PPBS-1	28	50
GAGA-CGAG-CGAG-CGAA-AGCA-CGAA-CGAG-TCTC	A32_mod	32	55

[a] Underlined residues denote the sequence of putative protein binding site (PPBS).

and minimal signal overlap in the aromatic region of ^{1}H NMR spectra (Figure 1B). Compared to A60, the variants A38, A36 and A32 are truncated by 11 residues at the 5'-end and by 11, 13 and 17 residues at the 3'-end, respectively (Table 1).

By examining the aromatic-H2'/H2" region of 2D ^{1}H–^{1}H NOESY spectra of A36 and A38, we identified a repetitive pattern of NOE contacts linking residues within the CGAG repeat (here termed CGAG walk pattern). Specifically, the residues C(n), G(n + 1), A(n + 2) and G(n + 3) are linked by the standard sequential (H2'/H2")$_n$-H8$_n$ and (H2'/H2")$_n$-H8$_{n+1}$ NOE contacts. The CGAG walk pattern is repeated for every CGAG repeat in both A36 and A38 (Figure 2A and B). It is noteworthy that the clustering of ^{1}H signals is a result of very similar shielding environment of the corresponding residues in each CGAG repeat. Unique features distinguishing the CGAG walk pattern include the following: (i) downfield signal of G(n + 1)H2' at δ ~2.8 ppm and upfield signal of A(n + 2)H2' proton at δ ~1.0 ppm, (ii) low intensity of sequential G(n + 1)H2'/H2"-A(n + 2)H8 NOE contacts and (iii) overlapping signals of G(n + 3)H2'/H2" (Figure 2 and Supplementary Figure S4). All four residues of CGAG repeat adopt *anti*-conformations across the glycosidic bond as indicated by medium intensities of the intra-residual H1'-H8 NOE and the presence of sequential (H1'/H2'/H2")$_n$-(H8/H6)$_{n+1}$ NOE cross-peaks. The DQF-COSY spectra revealed that residues C(n), G(n + 1) and A(n + 2) adopt South-type sugar puckers. The sugar pucker of G(n + 3) residue could not be determined due to spectral overlap (Supplementary Figures S5 and S6). C(n)P and G(n + 3)P resonate within δ −1.2–0.0 ppm range of ^{31}P, which is typical for B^I conformation of sugar-phosphate backbone (defined by difference ε–ζ ~ −90°) observed in canonical B-DNA (Supplementary Figure S7). Slightly upfield ^{31}P signals at δ −1.3 ppm were assigned to G(n + 1)P connecting C(n) and G(n + 1). The

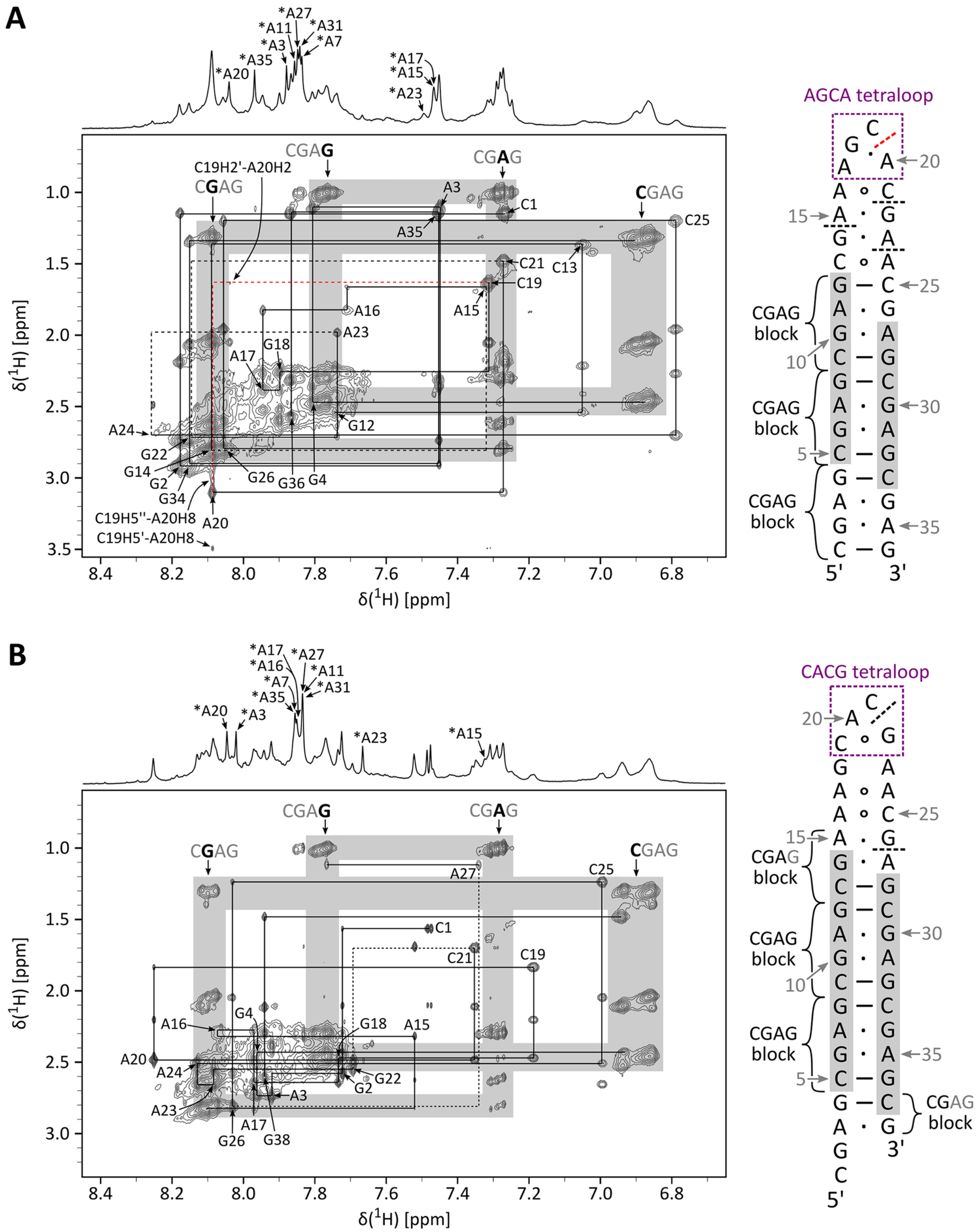


Figure 2. Aromatic-H2'/H2" region of 2D ^{1}H–^{1}H NOESY spectra of A36 (**A**) and A38 (**B**) oligonucleotides with their corresponding secondary structures. The sequential walk is represented by a solid black line. The steps in which the sequential walk is very weak or missing are marked by dashed black lines and the interruption associated with a *syn*-conformation of the following residue is marked by a red dashed line (applies also to the schematic of structure). We note that the sequential NOE contacts in G26–A27 step in A38 oligonucleotide are weak and hidden below the lowest contour shown. The assignments are indicated next to the intra-residual cross-peak. For clarity, the assignments of ^{1}H signals within the CGAG walk pattern is omitted. The grey areas in NOESY spectra denote the CGAG walk pattern. The ^{1}H signals of residues marked grey in the structure on the right side resonate within the CGAG walk pattern. The assignment of adenine H2 protons are indicated with asterisk (*) in the 1D ^{1}H spectrum shown above the NOESY spectrum. The dashes and dots in the structure denote experimentally verified canonical and non-canonical base-pairs, respectively. Circles denote base pairs suggested by molecular modelling. The NOESY mixing time was set to 150 ms. The spectra were recorded at 25°C, 1.2 mM strand concentration in 100% D_2O buffered by 20 mM KPi at pH 7.2.

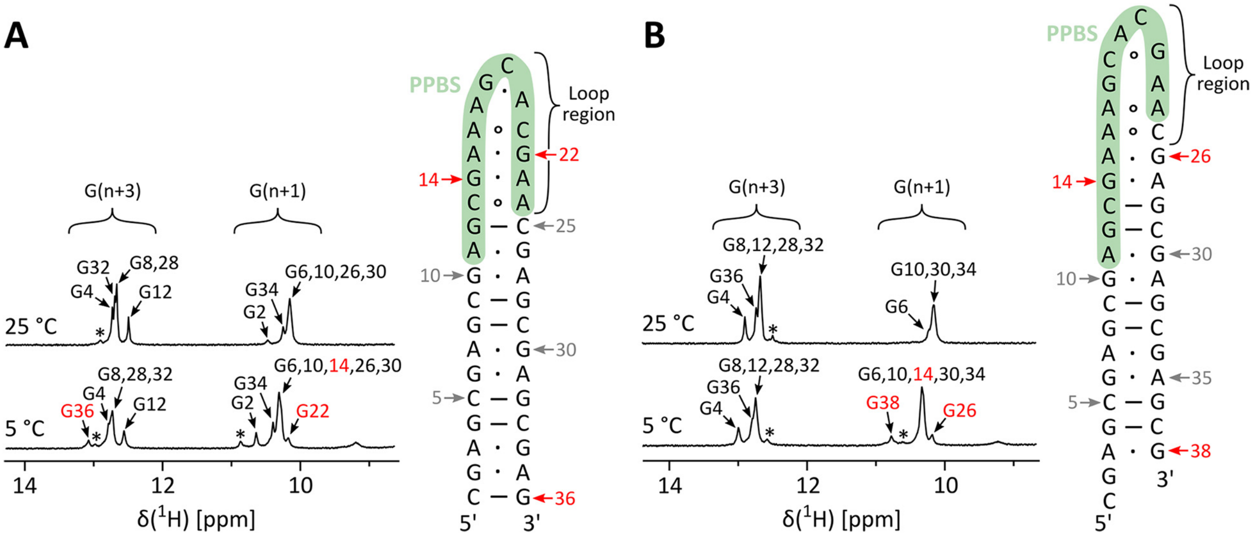


Figure 3. Imino-proton regions of 1D ^{1}H NMR spectra of A36 (**A**) and A38 (**B**) with their corresponding secondary structures. The assignment of imino-proton signals is indicated. The red labels denote residues involved in base pairs only at 5°C. The asterisks (*) mark unassigned imino-proton signals. The PPBS is highlighted in green. The spectra were recorded at 1.2 mM and 0.6 mM strand concentration for A36 and A38, respectively, in 90%/10% H_2O/D_2O buffered by 20 mM KPi at pH 7.2.

downfield ^{31}P signals in δ 0.5–1.5 ppm range were assigned to A(n + 2)P, which connects residues G(n + 1) and A(n + 2) involved in sheared G:A base pairs. The deshielding of ^{31}P atoms is connected to the increased population of B^{II} conformation ($\varepsilon-\zeta \sim +90°$) in comparison to B^{I} (48,49). Additional structural information was obtained with the use of NOESY spectra recorded at low temperature (Supplementary Figures S8 and S9). The imino-proton signals in δ 12–13 ppm range were assigned to guanines G(n + 3) utilizing NOE contacts between cytosine amino-protons and guanine imino-proton within a G:C base pair (Figure 3, Supplementary Figures S10 and S11). The imino-protons signals in δ 10–11 ppm range were assigned to guanines G(n + 1) through multiple cross-strand NOE contacts to sugar and aromatic protons of guanine in the neighboring G:A base pair (Supplementary Figures S10 and S11). Similarly, we observed several cross-strand NOE contacts of A(n + 2)H2 proton to sugar and aromatic protons of adenine in the neighboring G:A base pair (Supplementary Figure S12). These NOE contacts arise from cross-strand G-over-G and A-over-A stacking between two sheared G:A base pairs. Observed spectral features concur with previous reports on structures comprising symmetrical pyrimidine–GA–purine tetranucleotides oriented in tandem (50–58). Altogether, our data indicate a tandem orientation of CGAG repeats forming a canonical C(n):G(m + 3) base pair next to sheared G(n + 1):A(m + 2) and A(n + 2):G(m + 1) base pairs followed by a canonical G(n + 3):C(m) base pair as illustrated in the schematic below:

5′- C(n) G(n + 1) A(n + 2) G(n + 3) -3′
3′- G(m + 3) A(m + 2) G(m + 1) C(m) -5′

We hereby refer to the structural motif above as CGAG block. The NMR data indicate that structure of A36 is stabilized by three CGAG blocks (Figure 2A). The structure of A38 comprises two CGAG blocks flanked on each side by one incomplete CGAG block, which refers to C:G followed by at least one sheared G:A base pair (Figure 2B). We use the term 'loop region' to describe residues, which connect the CGAG blocks (Figure 3).

The hairpins adopted by A36 and A38 feature long purine-rich loop regions

The ^{1}H NMR signals of residues in the loop regions of A36 and A38 were assigned with the aid of ^{1}H–^{13}C HSQC spectra recorded on residue-specifically ^{13}C- and ^{15}N-labelled oligonucleotides (Supplementary Figure S13, for assignment strategy see Supplementary Materials). The loop region of A36 hairpin comprises 12 residues (C13-A24), whereas it comprises 10 residues (A16-C25) in A38 hairpin. In A36, 12 residues of PPBS are located in the loop region d(AGCGAAAGCACGAA), whereas only 9 residues of PPBS d(AGCGAAAGCACGAA) are located in the loop region of A38 (Figure 3A and B). The residues of PPBS that are not part of the loop regions are involved in the stem.

In A36, the residues C13–C19 and C21–A24 adopt *anti*-conformations across their glycosidic bonds. High intensity of C19H1'/H2'-A20H2 and A20H1'-H8 NOE contacts indicates *syn*-conformation of residue A20, which was corroborated by the downfield signal of A20C8 at δ 140.5 ppm characteristic of a purine residue in *syn*-conformation (Supplementary Figures S12 and S13) (59–63). Two imino-proton signals assigned to G14 and G22 (Figure 3A) reflect formation of two sheared G14:A23 and G22:A15 base pairs in the loop region. Signals of C19H5' and C19H5" resonate upfield at δ 3.5 and 3.0 ppm, respectively, which suggests that the sugar moiety of C19 is positioned above a neighboring base (Figure 2A and Supplementary Figure S8). DQF-COSY spectrum revealed that residue A15 adopts North-type sugar pucker, whereas residues G18, C19, A20 and G22 adopt South-type sugar pucker (Supplementary Figure S5).

In A38, all residues adopt *anti*-conformations. The weak sequential NOE connectivity in C21–G22 step (Figure 2A) likely results from a poor stacking caused by the turn of

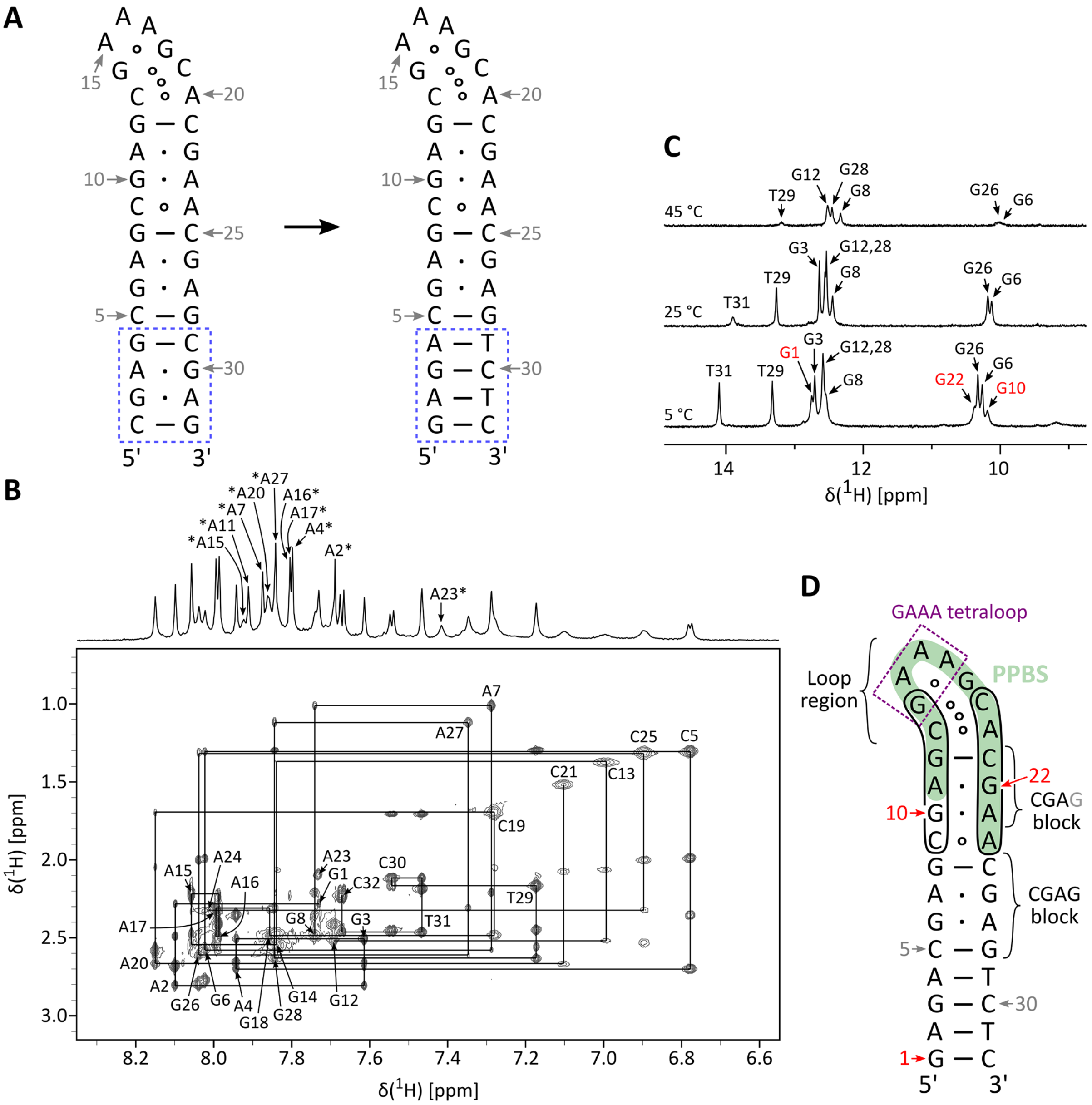


Figure 4. ^{1}H NMR spectra and secondary structure of A32_mod. (A) Schematics of secondary structures of A32 (left) and A32_mod (right). (B) H2'/H2"-aromatic region of 2D NOESY of A32_mod spectrum recorded at 45°C with a mixing time of 150 ms. The sequential walk is represented by a solid black line. The sequential walk was not traced in segments C9–A11 and G22–A23. We note that the sequential NOE contacts in C13–G14 step are weak and hidden below the lowest contour shown. The assignment of the aromatic protons is indicated next to the intra-residual cross-peaks. The assignment of adenine H2 protons is indicated with asterisk (*) in the 1D ^{1}H spectrum shown above the NOESY spectrum. (C) The imino-proton regions of 1D ^{1}H spectra of A32_mod, the assignment of imino-proton signals is indicated. The red labels denote residues involved in base pairs only at 5°C. (D) The secondary structure of A32_mod. The dashes and dots denote experimentally verified canonical and non-canonical base-pairs, respectively. Circles denote base pairs suggested by molecular modeling. The ^{1}H signals of residues in the encircled segments exhibit broadening. The PPBS is highlighted in green. The spectra were recorded at 0.6 mM strand concentration in (B) 100% D_2O and (C) 90%/10% H_2O/D_2O buffered by 20 mM KPi at pH 7.2.

the sugar-phosphate backbone in the loop. An unusual upfield chemical shift is observed for C21H4', which resonates at δ 3.5 ppm and indicates proximity of neighboring base to sugar moiety of C21 (Supplementary Figure S6). Only residues C19 and A23 from A16–C25 segment were determined to adopt South-type sugar pucker (Supplementary Figure S6). The sugar pucker could not be determined for the remaining residues of the loop region due to signal overlap and local dynamics.

Hairpin of A32_mod involves a dynamic loop region

Due to chemical exchange, the truncated variant A32 was not suitable for determination of high-resolution structure and only partial assignment of ^{1}H signals was achieved (Supplementary Figures S14–S16). Therefore, we prepared a modified oligonucleotide A32_mod by substituting the first and the last CGAG repeat in A32 with GAGA and TCTC segments, respectively (Table 1 and Figure 4A). A32_mod exhibits good quality of NMR spectra at 45°C

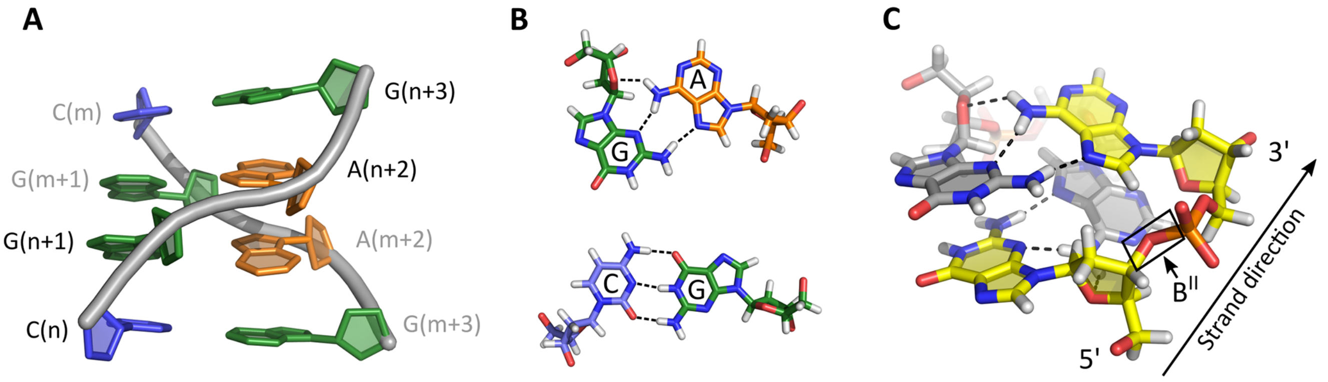


Figure 5. Structure of the conserved CGAG block constituting the stems of A36, A38 and A32_mod hairpins. (**A**) Cartoon representation of CGAG block adopted by two CGAG repeats. dC is in blue, dG in green and dA in orange. Note the cross-strand stacking between the two sheared G:A base pairs. (**B**) Non-canonical sheared G:A and canonical G:C base pairs constituting the CGAG block. (**C**) Detailed view on two G:A base pairs formed in tandem. The strands are colored yellow and grey. Note the cross-strand G-over-G and A-over-A stacking. The B^{II} conformation of the sugar-phosphate backbone is marked by rectangle. For clarity, the B^{II} conformation is highlighted only on the strand in the foreground.

without any signs of chemical exchange (Supplementary Figure S16). Negligible differences in ^{1}H chemical shifts between A32 and A32_mod suggest that the structure of the loop region is unaffected by the modification (Supplementary Figure S17). Therefore, we carried out the structure characterization using A32_mod. The H2'/H2''-aromatic region of NOESY spectrum of A32_mod recorded at 45°C shows an uninterrupted sequential walk in segments G1–G8, G12–C21 and A24–C32 (Figure 4B). The sequential walk in C5–G8 and C25–G28 segments resembles the CGAG walk pattern indicating formation of one CGAG block, which is corroborated by the downfield signals of A7P and A27P (Supplementary Figure S18). A broadening of ^{1}H signals is observed for residues C9–G14 and C19–A24, which was reduced at 0°C (Supplementary Figure S19). However, the ^{1}H signals of G14 and C19 remain significantly broadened even at 0°C indicating an absence of their well-defined position within the structure. The residues G12–C21 constitute the loop region of A32_mod hairpin. All residues adopt *anti*-conformations and the majority adopts South-type sugar pucker (Supplementary Figure S20).

The imino-region of ^{1}H NMR spectra recorded at 5°C include four imino-proton signals at δ ~10.3 ppm, five signals at δ ~12.6 ppm and two signals at δ ~13.4 and ~14.1 ppm, which indicate formation of four sheared G:A, five G:C and two A:T base pairs (Figure 4C). The imino-imino NOE contacts confirm formation of two G:C and two A:T base pairs followed by CGAG block (Figure 4D and Supplementary Figure S21). The remaining signal at δ ~12.6 ppm was assigned to G12H1, which confirmed the formation of G12:C21 base pair. The imino-proton signals at δ ~10.3 ppm present only at 5°C were assigned to G10 and G22 involved in G10:A23 and G22:A11 base pairs. The base pairs G12:C21, G10:A23 and G22:A11 form an incomplete CGAG block in the stem of A32_mod. The disappearance of G10 and G22 imino-proton signals at 25°C indicate destabilization of G10:A23 and G22:A11 base pairs, which explains the increased dynamics of C9–G14/C19–A24 segments observed at elevated temperatures (*vide supra*).

Formation of CGAG blocks and stable loops determine the arrangement of PPBS

Using MD simulations and experimentally derived restraints, we calculated structural ensembles of A36, A38 and A32_mod (Supplementary Figure S22, Supplementary Tables S1-S3). The stems of the three hairpins are very similar and predominantly composed of CGAG blocks stabilized by canonical G:C and sheared G:A base pairs (Figure 5A and B). In addition to CGAG blocks, the stem of A32_mod comprises one C:A, two G:C and two A:T base pairs. The stem of A38 also features an unpaired overhang formed by residues C1 and G2. The G:A base pairs are stabilized by two base-base H-bonds and one base-sugar H-bond formed between amino-proton H61 of adenine and O4' of guanine (Figure 5B) (35). Two sheared G:A base pairs formed in tandem are well-known for their cross-strand stacking (Figure 5C). Such arrangement of nucleobases results in B^{II} conformation of the sugar-phosphate backbone. The structures of A36, A38 and A32_mod demonstrate that the formation of multiple consecutive CGAG blocks is sterically allowed and energetically favourable.

The loop region of A36 is comprised of 12 residues C13–A24, which exhibit efficient base-stacking (Figure 6A). The bases are stacked sequentially in segments C13–G14, A15–C19, A20–G22 and A23–A24. The interruptions in G14–A15 and G22–A23 steps are result of cross-strand G22-over-G14 and A15-over-A23 stacking. In total, two C:A and three G:A base pairs are formed within the loop region. The sheared G14:A23 and G22:A15 base pairs are flanked by C13:A24 and C21:A16 base pairs. Even though both C:A base pairs are stabilized by a single H-bond, the base pair geometries are different (Figure 6D). Our results are in agreement with reports that C:A base pairs can adopt two interconverting geometries (64–66). The loop region is closed by residues A17–G18–C19–A20 forming an AGCA tetraloop, which involves an unpaired A17 residue, G18:A20 base pair with A20 adopting a *syn*-conformation and C19 facilitating the antiparallel orientation of strands. The G18:A20 base pair does not adopt a sheared geometry. Instead, it is stabilized by three H-bonds G18H22·A20N1, A20H61·G18N3, A20H62·G18O4' and features a signif-

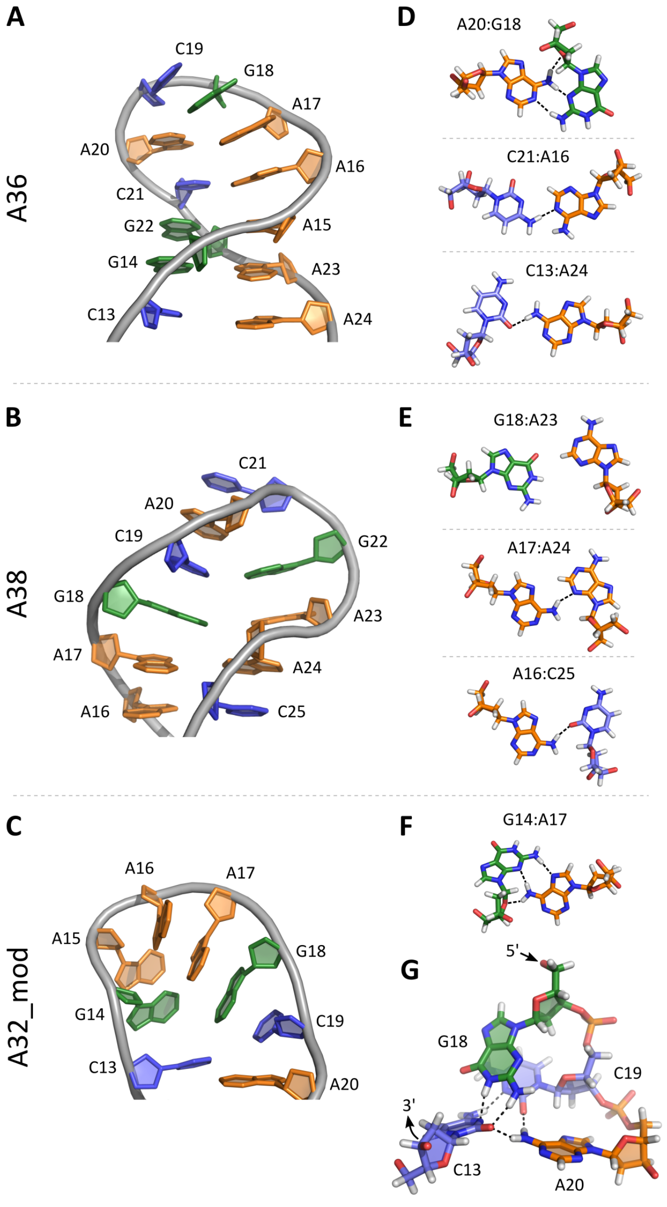


Figure 6. Loop regions in the structures of A36, A38 and A32_mod. (**A**–**C**) Cartoon representations of the loop regions with (**D**–**F**) detailed views of the non-canonical base pairs and (**G**) the structure adopted by residues C13 and G18–C19–A20 (C/GCA section) in A32_mod. dC is in blue, dG in green and dA in orange. The H-bonds are denoted by dashed black lines.

icant propeller twist (–22°), buckle (22°) and opening (–45°). The sugar moiety of C19 is positioned above the base of A20, which explains the upfield signals of C19H5' and C19H5" (Supplementary Figure S8). The AGCA tetraloop of A36 adopts a structure that somewhat resembles a structure of GNA triloop (N = any nucleotide), a thermodynamically stable DNA loop characterized by closing sheared G:A base pair and upfield H4' signal of residue N at δ ∼1.9 ppm (67–69).

The loop region of A38 hairpin encompasses 10 residues A16–C25 (Figure 6B). The bases are well-stacked just as in case of A36, however, only sequential stacking is observed in A38. The sequential stacking is interrupted only in C21–G22 step, which facilitates the reversal of strand directionality. It is important to note, that there is a considerable degree of variability in the base pair geometries of A16:C25, A17:A24 and G18:A23 within the refined structural ensemble of A38. It is likely due to high adenine content of the loop region (50%) and formation of semi-stable base pairs. In the lowest energy structure, the base pairs A16:C25 and A17:A24 are stabilized by H-bonds A16H62·C25O2 and A17H62·A24N3, respectively (Figure 6E). The bases G18 and A23 are not stabilized by any H-bonds. The closing C19:G22 base pair adopts a canonical geometry with three H-bonds and considerable buckle (34°) probably imposed by the turn of the sugar-phosphate backbone. The sugar moiety of residue C21 is positioned above the base of G22, which is experimentally supported by the upfield signal of C21H4' at δ 3.5 ppm (Supplementary Figure S6). The loop region of A36 is closed by residues C19–A20–C21–G22 forming a CACG tetraloop, which belongs to CNNG family of tetraloops exhibiting extraordinary thermal stability (70). The structure of CACG tetraloop in A38 is strikingly similar to the one reported previously for a hairpin formed by d(CGCACGCG) (71). Additionally, in d(CGCACGCG), C5H4' resonates upfield at δ 3.7 ppm, which is similar to chemical shift of C21H4' in A38.

The hairpin adopted by A32_mod involves the shortest loop region of the three hairpins comprising only eight residues C13–A20 (Figure 6C). G18–C19 bulge prevents efficient base-stacking between several residues in A32_mod. The axis of the loop region is tilted relative to the axis of the stem. The loop region is capped by G14–A15–A16–A17 tetraloop comprising a sheared G14:A17 base pair (Figure 6F). The GAAA tetraloop belongs to thermodynamically stable GNNA family of DNA tetraloops (70). The residues C13 and G18–C19–A20 (termed C/GCA section) adopt a peculiar arrangement that is stabilized by five H-bonds C13H41·C19N3, G18H1·C13N3, G18H21·C13O2, A20H61·C13O2 and A20H62·C19O2 in the lowest energy structure (Figure 6G). A slightly different arrangement of C/GCA section is found in the remaining models of the refined structural ensemble of A32_mod, which may explain the broadening of NMR signals observed for residues of C/GCA section (Supplementary Figure S23).

DISCUSSION

The two isoforms of the AUTS2 protein are expressed at different stages of brain development. However, the molecular mechanism for the expression switch is unknown. By performing a genome search, we identified a CGAG-rich region, which we termed A60, in the promoter of *AUTS2* gene capable of adopting a variety of non-canonical secondary structures that may be involved in switching between expression of the two isoforms of AUTS2. Utilizing Ensembl showed that A60 comprises putative protein binding site, PPBS, for the transcription regulatory proteins ERF, MAX and FLI1 (72). However, due to structural flexibility and polymorphism, the structural characterization of A60 turned out to be rather challenging. We therefore focused on three shorter oligonucleotides A38, A36 and A32_mod to deduce the main driving forces shaping the structure of A60. Utilizing extensive NMR analysis and

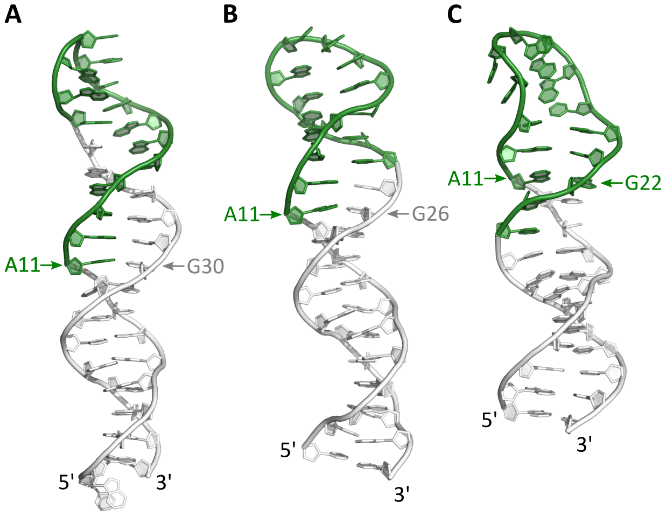


Figure 7. Different arrangements of PPBS in the high-resolution structures of A38 (**A**), A36 (**B**) and A32_mod (**C**). The PPBS (A11–A24) is colored green, the remaining residues are shown in grey. The base pairs involving residue A11 are shown to emphasize the register-shift between the three structures. Compared to the structure of A38, the structures of A36 and A32_mod are register-shifted by four and eight residues, respectively.

MD, we obtained high-resolution structures of A38, A36 and A32_mod. All three oligonucleotides adopt hairpins composed predominantly of non-canonical base pairs with various relative positions of PPBS in the loop regions (Figure 7). Importantly, the thermal stability of the structures adopted by A38, A36, A32_mod and A60 is well above the physiological temperature, which supports the possibility of their formation *in vivo*. The main contributor to the formation and stability of the hairpins are the consecutive G:C and G:A base pairs arranged into CGAG blocks. Specifically, we observed that oligonucleotides with several CGAG tracts have a propensity to arrange in a way to maximize the number of consecutive G:C and G:A base pairs. If we compare the structure of A38 with structures of A36 and A32_mod, we find that the strands are register-shifted by 4 and 8 residues, respectively, which results in different arrangements of PPBS (Figure 7). The strands shift in register by four (or multiples of four) residues as any other register-shift prevents the formation of CGAG blocks. Consequently, the structures of A38, A36 and A32_mod have similar stem regions dominated by CGAG blocks, but the register-shift causes the loop regions to be comprised of different residues, which leads to variability in their structure. Since the loop regions of all three structures have high purine content (75% for A36 and A32_mod, 70% for A38) and feature an efficient base-base stacking, it is clear that the π–π stacking is an important stabilizing force shaping the structure of the loop regions. Additionally, the loop regions tend to arrange in a way to enable formation of thermodynamically stable tri- and tetraloops. In A38, the CACG tetraloop, which belongs to CNNG family, is formed at the tip of the loop region. In A36, the AGCA tetraloop may be considered as a GNA triloop flanked at its 5'-end by the unpaired A17 residue. In A32_mod, the peculiar arrangement of residues C13/G18–C19–A20 enables formation of GAAA tetraloop of GNNA family. Importantly, based on the NMR spectra of A60 and the structural information gathered from the A38, A36 and A32_mod hairpins, we propose that A60 adopts multiple hairpin structures that differ in register-shift stabilized by G:C and sheared G:A base pairs arranged in CGAG blocks. The hairpins of A60 most likely contain dynamic regions such as bulges and internal loops.

Observing the impact of the register-shift on the structure and the propensity to form thermodynamically stable tri- and tetraloops enables us to propose different mechanisms how A60 could be implicated in regulation of AUTS2 expression. During processes of DNA replication and transcription, A60, or parts of it, can be unfolded from the double-helical structure. Dependent on the number of CGAG repeats that are unfolded, the register-shift mechanism could lead to formation of other hairpins with PPBS located mostly in the stem region. Positioning PPBS in the loop region of the hairpin could sequester the PPBS from the transcription machinery. On the other hand, the loop regions could act as binding sites for yet to be identified biomolecules. A clear way of changing the structure of the PPBS from mostly loop associated to being predominantly present in a stem has the potential to be relevant for exploring the implications on the regulation of the *AUTS2* gene. In the case that the entire A60 region is in a single strand state, multiple hairpin structures might be formed, which could lead to structural interactions between them and also raise possibilities of their relevance for replication and transcription. To answer these important questions more experiments including *in vivo* studies will have to be carried out and we believe that this work will stimulate further research into the influence of A60 on the regulation of *AUTS2* gene.

DATA AVAILABILITY

Atomic coordinates of A38, A36 and A32_mod NMR structures are deposited in the Protein Data Bank under accession numbers 8BM7, 8BM6 and 8BM4, respectively. The lists of chemical shifts of A38, A36 and A32_mod are deposited in Biological Magnetic Resonance Data Bank under accession numbers 34770, 34769, 34768, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge the CERIC-ERIC Consortium for access to experimental facilities.

FUNDING

Slovenian Research Agency (ARRS) [P1-0242, Z1-3192]; Young Researcher grant [to A.N.]. Funding for open access charge: Slovenian Research Agency [P1-0242].
Conflict of interest statement. None declared.

REFERENCES

1. Neidle,S. and Balasubramanian,S. (2006) In: *Quadruplex Nucleic Acids The Royal Society of Chemistry*. London, UK.

2. Kaushik,M., Kaushik,S., Roy,K., Singh,A., Mahendru,S., Kumar,M., Chaudhary,S., Ahmed,S. and Kukreti,S. (2016) A bouquet of DNA structures: emerging diversity. *Biochem. Biophys. Rep.*, **5**, 388–395.
3. Galer,P., Wang,B., Šket,P. and Plavec,J. (2016) Reversible pH switch of two-quartet G-quadruplexes formed by human telomere. *Angew. Chem. Int. Ed.*, **55**, 1993–1997.
4. Jana,J. and Weisz,K. (2021) Thermodynamic stability of G-quadruplexes: impact of sequence and environment. *Chembiochem*, **22**, 2848–2856.
5. Matsumoto,S., Tateishi-Karimata,H., Takahashi,S., Ohyama,T. and Sugimoto,N. (2020) Effect of molecular crowding on the stability of RNA G-quadruplexes with various numbers of quartets and lengths of loops. *Biochemistry*, **59**, 2640–2649.
6. Kim,N. (2019) The Interplay between G-quadruplex and transcription. *Curr. Med. Chem.*, **26**, 2898–2917.
7. Spiegel,J., Adhikari,S. and Balasubramanian,S. (2020) The structure and function of DNA G-quadruplexes. *Trends Chem*, **2**, 123–136.
8. Simone,R., Fratta,P., Neidle,S., Parkinson,G.N. and Isaacs,A.M. (2015) G-quadruplexes: emerging roles in neurodegenerative diseases and the non-coding transcriptome. *FEBS Lett*, **589**, 1653–1668.
9. Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res*, **43**, 8627–8637.
10. Zeraati,M., Langley,D.B., Schofield,P., Moye,A.L., Rouet,R., Hughes,W.E., Bryan,T.M., Dinger,M.E. and Christ,D. (2018) I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.*, **10**, 631–637.
11. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
12. Georgakopoulos-Soares,I., Morganella,S., Jain,N., Hemberg,M. and Nik-Zainal,S. (2018) Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res*, **28**, 1264–1271.
13. Guiblet,W.M., Cremona,M.A., Harris,R.S., Chen,D., Eckert,K.A., Chiaromonte,F., Huang,Y.-F. and Makova,K.D. (2021) Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res*, **49**, 1497–1516.
14. Khristich,A.N. and Mirkin,S.M. (2020) On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.*, **295**, 4134–4170.
15. Balasubramanian,S. and Neidle,S. (2009) G-quadruplex nucleic acids as therapeutic targets. *Curr. Opin. Chem. Biol.*, **13**, 345–353.
16. Brown,S.L. and Kendrick,S. (2021) The i-motif as a molecular target: more than a complementary DNA secondary structure. *Pharmaceuticals*, **14**, 96.
17. Tateishi-Karimata,H. and Sugimoto,N. (2020) Chemical biology of non-canonical structures of nucleic acids for therapeutic applications. *Chem. Commun.*, **56**, 2379–2390.
18. Neidle,S. (2017) Quadruplex nucleic acids as targets for anticancer therapeutics. *Nat. Rev. Chem.*, **1**, 0041.
19. Neidle,S. (2020) In: *Quadruplex Nucleic Acids As Targets For Medicinal Chemistry*. 1st edn., Academic Press, Cambridge, USA.
20. Paulson,H. (2018) Repeat expansion diseases. *Handb. Clin. Neurol.*, **147**, 105–123.
21. Depienne,C. and Mandel,J.-L. (2021) 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.*, **108**, 764–785.
22. Krafcikova,M., Dzatko,S., Caron,C., Granzhan,A., Fiala,R., Loja,T., Teulade-Fichou,M.-P., Fessl,T., Hänsel-Hertsch,R., Mergny,J.-L. *et al.* (2019) Monitoring DNA–ligand interactions in living human cells using NMR spectroscopy. *J. Am. Chem. Soc.*, **141**, 13281–13285.
23. Krafčík,D., Ištvánková,E., Džatko,Š., Víšková,P., Foldynová-Trantírková,S. and Trantírek,L. (2021) Towards profiling of the G-quadruplex targeting drugs in the living human cells using NMR spectroscopy. *Int. J. Mol. Sci.*, **22**, 6042.
24. Cheng,M., Qiu,D., Tamon,L., Ištvánková,E., Víšková,P., Amrane,S., Guédin,A., Chen,J., Lacroix,L., Ju,H. *et al.* (2021) Thermal and pH stabilities of i-DNA: confronting in vitro experiments with models and in-cell NMR data. *Angew. Chem. Int. Ed.*, **60**, 10286–10294.
25. Poggi,L. and Richard,G.-F. (2021) Alternative DNA structures in vivo: molecular evidence and remaining questions. *Microbiol. Mol. Biol. Rev. MMBR*, **85**, e00110-20.
26. Beunders,G., Voorhoeve,E., Golzio,C., Pardo,L.M., Rosenfeld,J.A., Talkowski,M.E., Simonic,I., Lionel,A.C., Vergult,S., Pyatt,R.E. *et al.* (2013) Exonic deletions in AUTS2 cause a syndromic form of intellectual disability and suggest a critical role for the C terminus. *Am. J. Hum. Genet.*, **92**, 210–220.
27. Amarillo,I.E., Li,W.L., Li,X., Vilain,E. and Kantarci,S. (2014) De novo single exon deletion of AUTS2 in a patient with speech and language disorder: a review of disrupted AUTS2 and further evidence for its role in neurodevelopmental disorders. *Am. J. Med. Genet. A.*, **164**, 958–965.
28. Liu,Y., Zhao,D., Dong,R., Yang,X., Zhang,Y., Tammimies,K., Uddin,M., Scherer,S.W. and Gai,Z. (2015) De novo exon 1 deletion of AUTS2 gene in a patient with autism spectrum disorder and developmental delay: a case report and a brief literature review. *Am. J. Med. Genet. A.*, **167**, 1381–1385.
29. Beunders,G., Kamp,J.v., Vasudevan,P., Morton,J., Smets,K., Kleefstra,T., Munnik,S.A.d., Schuurs-Hoeijmakers,J., Ceulemans,B., Zollino,M. *et al.* (2016) A detailed clinical analysis of 13 patients with AUTS2 syndrome further delineates the phenotypic spectrum and underscores the behavioural phenotype. *J. Med. Genet.*, **53**, 523–532.
30. Sanchez-Jimeno,C., Blanco-Kelly,F., López-Grondona,F., Losada-Del Pozo,R., Moreno,B., Rodrigo-Moreno,M., Martinez-Cayuelas,E., Riveiro-Alvarez,R., Fenollar-Cortés,M., Ayuso,C. *et al.* (2021) Attention deficit hyperactivity and autism spectrum disorders as the core symptoms of AUTS2 syndrome: description of five new patients and update of the frequency of manifestations and genotype-phenotype correlation. *Genes*, **12**, 1360.
31. Pang,W., Yi,X., Li,L., Liu,L., Xiang,W. and Xiao,L. (2021) Untangle the multi-facet functions of Auts2 as an entry point to understand neurodevelopmental disorders. *Front. Psychiatry*, **12**, 580433.
32. Biel,A., Castanza,A.S., Rutherford,R., Fair,S.R., Chifamba,L., Wester,J.C., Hester,M.E. and Hevner,R.F. (2022) AUTS2 syndrome: molecular mechanisms and model systems. *Front. Mol. Neurosci.*, **15**, 858582.
33. Monderer-Rothkoff,G., Tal,N., Risman,M., Shani,O., Nissim-Rafinia,M., Malki-Feldman,L., Medvedeva,V., Groszer,M., Meshorer,E. and Shifman,S. (2021) AUTS2 isoforms control neuronal differentiation. *Mol. Psychiatry*, **26**, 666–681.
34. Hori,K., Nagai,T., Shan,W., Sakamoto,A., Taya,S., Hashimoto,R., Hayashi,T., Abe,M., Yamazaki,M., Nakao,K. *et al.* (2014) Cytoskeletal regulation by AUTS2 in neuronal migration and neuritogenesis. *Cell Rep*, **9**, 2166–2179.
35. Gao,Y.-G., Robinson,H., Sanishvili,R., Joachimiak,A. and Wang,A.H.-J. (1999) Structure and recognition of sheared tandem G·A base pairs associated with human centromere DNA sequence at atomic resolution. *Biochemistry*, **38**, 16452–16460.
36. Chen,H., Yang,P., Yuan,C. and Pu,X. (2005) Study on the binding of base-mismatched oligonucleotide d(GCGAGC)2 by cobalt(III) complexes. *Eur. J. Inorg. Chem.*, **2005**, 3141–3148.
37. Chou,S.H., Zhu,L. and Reid,B.R. (1997) Sheared purine x purine pairing in biology. *J. Mol. Biol.*, **267**, 1055–1067.
38. Robinson,H., van Boom,J.H. and Wang,A.H.-J. (1994) 5'-CGA motif induces other sequences to form homo base-paired parallel-stranded DNA duplex: the structure of (G-A)n derived from four DNA oligomers containing (G-A)3 sequence. *J. Am. Chem. Soc.*, **116**, 1565–1566.
39. Novotný,A., Novotný,J., Kejnovská,I., Vorlíčková,M., Fiala,R. and Marek,R. (2021) Revealing structural peculiarities of homopurine GA repetition stuck by i-motif clip. *Nucleic Acids Res*, **49**, 11425–11437.
40. Lee,W., Tonelli,M. and Markley,J.L. (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics*, **31**, 1325–1327.
41. Case,D.A., Belfon,K., Ben-Shalom,I.Y., Brozell,S.R., Cerutti,D.S., Cheatham,T.E. III, Cruzeiro,V.W.D., Darden,T.A., Duke,R.E., Giambasu,G. *et al.* (2020) *Amber 2020*.
42. Zgarbová,M., Šponer,J., Otyepka,M., Cheatham,T.E., Galindo-Murillo,R. and Jurečka,P. (2015) Refinement of the sugar–phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput.*, **11**, 5723–5736.
43. Galindo-Murillo,R., Robertson,J.C., Zgarbová,M., Šponer,J., Otyepka,M., Jurečka,P. and Cheatham,T.E. (2016) Assessing the

current state of amber force field Modifications for DNA. *J. Chem. Theory Comput.*, **12**, 4114–4127.

44. Onufriev,A., Bashford,D. and Case,D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, **55**, 383–394.
45. Loncharich,R.J., Brooks,B.R. and Pastor,R.W. (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. *Biopolymers*, **32**, 523–535.
46. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J.C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
47. Kypr,J., Kejnovská,I. and Vorlíčková,M. (2007) Conformations of DNA strands containing GAGT, GACA, or GAGC tetranucleotide repeats. *Biopolymers*, **87**, 218–224.
48. Heddi,B., Foloppe,N., Bouchemal,N., Hantz,E. and Hartmann,B. (2006) Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.*, **128**, 9170–9177.
49. Cheng,J.-W., Chou,S.-H. and Reid,B.R. (1992) Base pairing geometry in GA mismatches depends entirely on the neighboring sequence. *J. Mol. Biol.*, **228**, 1037–1041.
50. Chou,S.-H., Cheng,J.-W. and Reid,B.R. (1992) Solution structure of (d(ATGAGCGAATA))2: adjacent G: a mismatches stabilized by cross-strand base-stacking and BII phosphate groups. *J. Mol. Biol.*, **228**, 138–155.
51. Greene,K.L., Jones,R.L., Li,Y., Robinson,H., Wang,A.H.-J., Zon,G. and Wilson,W.D. (1994) Solution structure of a GA mismatch DNA sequence, d(CCATGAATGG)2, determined by 2D NMR and structural refinement methods. *Biochemistry*, **33**, 1053–1062.
52. Li,Y., Zon,G. and Wilson,W.D. (1991) NMR and molecular modeling evidence for a G.A mismatch base pair in a purine-rich DNA duplex. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 26–30.
53. Li,Y. and Agrawal,S. (1995) Oligonucleotides containing G.A pairs: effect of flanking sequences on structure and stability. *Biochemistry*, **34**, 10056–10062.
54. Chou,S., Chin,K. and Wang,A.H.-J. (2003) Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res*, **31**, 2461–2474.
55. Chou,S.H., Cheng,J.W., Fedoroff,O. and Reid,B.R. (1994) DNA sequence GCGAATGAGC containing the human centromere core sequence GAAT forms a self-complementary duplex with sheared G.A pairs in solution. *J. Mol. Biol.*, **241**, 467–479.
56. Lane,A., Martin,S.R., Ebel,S. and Brown,T. (1992) Solution conformation of a deoxynucleotide containing tandem G.A mismatched base pairs and 3’-overhanging ends in d(GTGAACTT)2. *Biochemistry*, **31**, 12087–12095.
57. Lam,S.L. and Chi,L.M. (2010) Use of chemical shifts for structural studies of nucleic acids. *Prog. Nucl. Magn. Reson. Spectrosc.*, **56**, 289–310.
58. Katahira,M., Sato,H., Mishima,K., Uesugi,S. and Fujii,S. (1993) NMR studies of G:a mismatches in oligodeoxyribonucleotide duplexes modelled after ribozymes. *Nucleic Acids Res*, **21**, 5418–5424.
59. Greene,K.L., Wang,Y. and Live,D. (1995) Influence of the glycosidic torsion angle on 13C and 15N shifts in guanosine nucleotides: investigations of G-tetrad models with alternating syn and anti bases. *J. Biomol. NMR*, **5**, 333–338.
60. Fonville,J.M., Swart,M., Vokáčová,Z., Sychrovský,V., Šponer,J.E., Šponer,J., Hilbers,C.W., Bickelhaupt,F.M. and Wijmenga,S.S. (2012) Chemical shifts in nucleic acids studied by density functional theory calculations and comparison with experiment. *Chem. Eur. J.*, **18**, 12372–12387.
61. Bielskutė,S., Plavec,J. and Podbevšek,P. (2021) Oxidative lesions modulate G-quadruplex stability and structure in the human BCL2 promoter. *Nucleic Acids Res*, **49**, 2346–2356.
62. Lenarčič Živković,M., Rozman,J. and Plavec,J. (2018) Adenine-driven structural switch from a two- to three-quartet DNA G-quadruplex. *Angew. Chem. Int. Ed.*, **57**, 15395–15399.
63. Marušič,M. and Plavec,J. (2019) Towards understanding of polymorphism of the G-rich region of human papillomavirus Type 52. *Molecules*, **24**, 1294.
64. Boulard,Y., Cognet,J.A.H., Gabarro-Arpa,J., Le Bret,M., Sowers,L.C. and Fazakerley,G.V. (1992) The pH dependent configurations of the C.A mispair in DNA. *Nucleic Acids Res*, **20**, 1933–1941.
65. Boulard,Y., Cognet,J.A.H., Gabarro-Arpa,J., Le Bret,M., Carbonnaux,C. and Fazakerley,G.V. (1995) Solution structure of an oncogenic DNA duplex, the K-ras gene and the sequence containing a central C·A or A·G mismatch as a function of pH: nuclear magnetic resonance and molecular dynamics studies. *J. Mol. Biol.*, **246**, 194–208.
66. Granzhan,A., Kotera,N. and Teulade-Fichou,M.-P. (2014) Finding needles in a basestack: recognition of mismatched base pairs in DNA by small molecules. *Chem. Soc. Rev.*, **43**, 3630–3665.
67. Hirao,I., Kawai,G., Yoshizawa,S., Nishimura,Y., Ishido,Y., Watanabe,K. and Miura,K. (1994) Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nucleases and heat. *Nucleic Acids Res*, **22**, 576–582.
68. Chou,S.-H., Zhu,L. and Reid,B.R. (1996) On the relative ability of centromeric GNA triplets to form hairpins versus self-paired duplexes. *J. Mol. Biol.*, **259**, 445–457.
69. Pavc,D., Wang,B., Spindler,L., Drevenšek-Olenik,I., Plavec,J. and Šket,P. (2020) GC ends control topology of DNA G-quadruplexes and their cation-dependent assembly. *Nucleic Acids Res*, **48**, 2749–2761.
70. Nakano,M., Moody,E.M., Liang,J. and Bevilacqua,P.C. (2002) Selection for thermodynamically stable DNA tetraloops using temperature gradient gel electrophoresis reveals four motifs: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc). *Biochemistry*, **41**, 14281–14292.
71. Ippel,H.H., van den Elst,H., van der Marel,G.A., van Boom,J.H. and Altona,C. (1998) Structural similarities and differences between H1- and H2-family DNA minihairpin loops: NMR studies of octameric minihairpins. *Biopolymers*, **46**, 375–393.
72. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R. *et al.* (2022) Ensembl 2022, release 108. *Nucleic Acids Res*, **50**, D988–D995.