1604–1617 *Nucleic Acids Research, 2005, Vol. 33, No. 5*
doi:10.1093/nar/gki307

Structural roles of CTG repeats in slippage expansion during DNA replication

Lai Man Chi and Sik Lok Lam*

Department of Chemistry, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received November 25, 2004; Revised December 29, 2004; Accepted February 27, 2005

ABSTRACT

CTG triplet repeat sequences have been found to form slipped-strand structures leading to self-expansion during DNA replication. The lengthening of these repeats causes the onset of neurodegenerative diseases, such as myotonic dystrophy. In this study, electrophoretic and NMR spectroscopic studies have been carried out to investigate the length and the structural roles of CTG repeats in affecting the hairpin formation propensity. Direct NMR evidence has been successfully obtained the first time to support the presence of three types of hairpin structures in sequences containing 1–10 CTG repeats. The first type contains no intra-loop hydrogen bond and occurs when the number of repeats is less than four. The second type has a 4 nt TGCT-loop and occurs in sequences with even number of repeats. The third type contains a 3 nt CTG-loop and occurs in sequences with odd number of repeats. Although stabilizing interactions have been identified between CTG repeats in both the second and third types of hairpins, the structural differences observed account for the higher hairpin formation propensity in sequences containing even number of CTG repeats. The results of this study confirm the hairpin loop structures and explain how slippage occurs during DNA replication.

INTRODUCTION

Genetic instabilities of triplet repeat sequences including $(CTG)_n{\cdot}(CAG)_n$, $(CGG)_n{\cdot}(CCG)_n$ and $(GAA)_n{\cdot}(TTC)_n$ have been found to associate with at least 14 genetic neurodegenerative diseases (1,2). The onset of Huntington's disease, myotonic dystrophy and spinocerebellar ataxia type 1 is related to the expansion of CAG and CTG repeats in specific gene regions (1). The unusual biology associated with these repeats is probably owing to the inherent flexibility (3,4) and genetic instability originating from unusual DNA structures, such as single loop, double loop, hairpins, flap intermediates and slipped-strand structures (4–7). Surface probing by anti-DNA antibodies has also revealed the presence of cruciform and Z-DNA structures in sequences containing slipped CAG and CTG repeats (8). Although biophysical and biochemical studies have shown that CAG and CTG repeats form stable hairpin structures with mismatched base pairs in the stem region during DNA replication (9,10), little is known about the underlying chemical forces that govern these surprisingly stable structures.

The presence of mismatches in unusual DNA structures has been hypothesized to be the origin of genetic instabilities that lead to DNA mutations (11–16). Therefore, structural information about these unusual structures and mismatches is useful for understanding the triplet repeat expansion process. During DNA replication, repair and recombination, slippage in single strands of CAG and CTG repeats can occur, forming slipped-strand structures with A·A and T·T mismatches, respectively (17–19). In CAG repeat sequences, both extrahelical (10,20) and intrahelical (21) A·A mismatches have been reported, whereas intrahelical T·T mismatches have been shown to remain stacked in CTG repeat sequences (20,22–24). Not much information, however, has been obtained to elucidate how the repeat length affects the structural roles of CTG repeats during replication.

In order to understand how structures affect the slippage mechanism, the present study aims at investigating the structural details of CTG repeat sequences with different repeat lengths using NMR spectroscopy. Electrophoretic study has also been performed to analyze the hairpin formation propensity of these CTG repeat sequences. Owing to the repetitive nature of the sequences, it has been difficult to analyze triplet repeats using NMR spectroscopy due to the severe overlaps of resonance signals. Moreover, individual strands of CTG repeats tend to form homoduplex at NMR sample concentration. This reduces the population of the hairpin conformer and thus hampers spectroscopic studies of the structural roles of triplet repeats in the hairpin loops. To overcome the above-mentioned difficulties, the present work focuses on NMR

*To whom correspondence should be addressed. Tel: +852 2609 8126; Fax: +852 2603 5057; Email: lams@cuhk.edu.hk

investigation of the less crowded imino and methyl proton regions of CTG repeats. Since guanine and thymine imino protons are involved in Watson–Crick G–C and A–T base pairs, respectively (Figure 1A), the appearance of these imino signals provides evidence for the presence of Watson–Crick base pairs. The presence of interactions in T·T mismatches is also evidenced by the more upfield thymine imino signals. In order to enhance the population of hairpins, CTG repeat sequences have been designed to contain a complementary stem region to restrict the flexibility at both ends and adenine nucleotides have been inserted to the stem–loop junction (25). Fast cooling procedures and modifications by lengthening the stem region have also been carried out and demonstrated to be useful to promote the hairpin population.

The present work analyzed DNA sequences with up to 10 repeats because triplet repeat expansion with higher repeat number may be resulted from the formation of more folded structures of shorter repeats with similar stability (21). This is supported by previous ultraviolet (UV) melting studies that showed the folds formed by 30 CAG or CTG repeats have no higher melting temperatures than those formed by 10 repeats (20). Based on the successful resonance assignments of the imino and methyl signals, three types of CTG repeat hairpin structures have been identified. This study provides direct structural evidence for investigating the effect of repeat length on the stability and the propensity of CTG repeat hairpins. It also gives a clearer picture on how slippage occurs in steps of two triplets during expansion (18,19).

MATERIALS AND METHODS

Sample design

$(CTG)_n$ hairpins have been designed to contain seven complementary Watson–Crick base pairs in the stem region. Figure 1B shows the positions of the stem nucleotides that are sequentially labeled from 1 to 7 on the 5′-stem and 8 to 14 on the 3′-stem. The increasing number of CTG repeats allows the

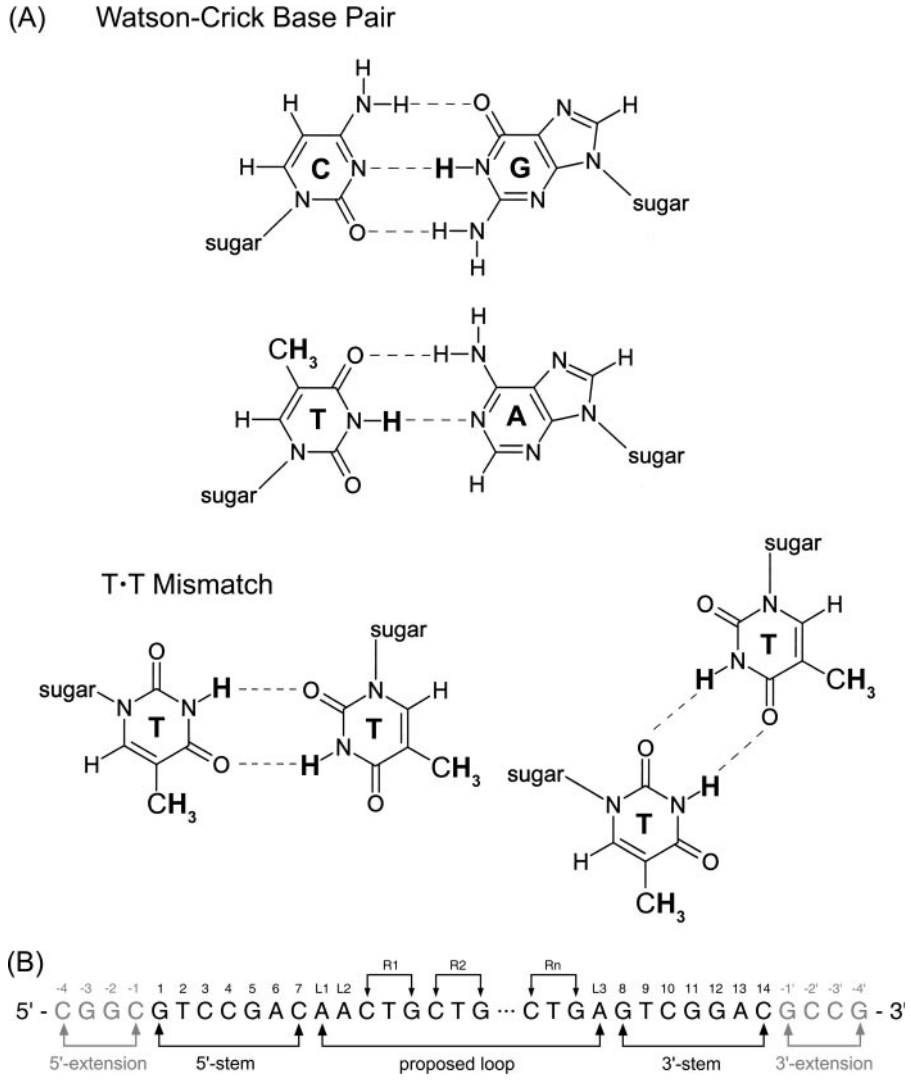

Figure 1. (**A**) Imino and methyl protons (in bold) in Watson–Crick base pair and T·T mismatch. Owing to the asymmetric arrangement of the T·T mismatches, two T·T pairing modes are present. Rapid transition between the two pairing modes is possible. (**B**) Numbering scheme of $(CTG)_n$. Nucleotides in gray represent the extended stem regions that promote hairpin formation.

understanding of how repeat length affects DNA structures. 'R*n*' is used to label the *n*th CTG repeat in the sequence. Two adenine nucleotides have been inserted to the 5′ end of the CTG repeats and one adenine nucleotide has been inserted to the 3′ end in order to promote the formation of hairpins (25). This step is important because the present work aims at studying how the CTG repeats behave in the hairpin loops. Their positions are labeled with L1, L2 and L3 as shown in Figure 1B. To assist the resonance assignments of $(CTG)_n$, extended sequences of $(CTG)_3$, $(CTG)_4$ and $(CTG)_5$ have been synthesized. Two G–C base pairs have been added to the hairpin stem of $(CTG)_3$, whereas four G–C base pairs have been added to the hairpin stems of $(CTG)_4$ and $(CTG)_5$ in order to increase the hairpin population. The numbers −1, −2, −3 and −4 are used to label the positions of nucleotides in the extended 5′ terminal, whereas −1′, −2′, −3′ and −4′ are used for the extended 3′ terminal. In order to confirm the resonance assignments of $(CTG)_4$ and $(CTG)_5$, the thymine nucleotide in the first CTG repeat of extended $(CTG)_4$ and the thymine nucleotide in the second CTG repeat of $(CTG)_5$ were substituted by a cytosine nucleotide. These sequences are called extended TR1C-$(CTG)_4$ and extended TR2C-$(CTG)_5$ with the substituted nucleotides TR1C and TR2C, respectively.

Sample preparation

All CTG repeat samples were synthesized using solid-phase phosphoramidite chemistry in an Applied Biosystems Model 392 DNA synthesizer. The samples were purified using denaturing PAGE, followed by electroelution and diethylaminoethyl Sephacel anion exchange column chromatography. Centricon-3 concentrators were finally used to remove the high salt contents of the samples. Sample quantities were determined by UV absorbance at 260 nm.

Electrophoretic study

Non-denaturing gels containing 20% polyacrylamide were prepared to investigate the hairpin and duplex populations of $(CTG)_n$ sequences. In order to determine how the thermodynamic stabilities of different structures affect the hairpin formation propensity, the DNA samples were first heated to 95°C for 5 min to remove all structural heterogeneity and then allowed to cool slowly and stand at room temperature for at least 4 days. This slow cool procedure promotes the formation of the thermodynamically more stable conformer and thus allows an accurate estimation of the hairpin formation propensity based on the relative thermodynamic stabilities of the different conformers. As higher DNA concentrations favor the formation of duplex, 10 μM and 1 mM samples were prepared in an NMR buffer containing 150 mM sodium chloride, 10 mM sodium phosphate at pH 7 and 0.1 mM 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS). The results from these two concentrations were useful for determining the concentration effect on the hairpin formation propensity. DNA samples in gels were stained with stains-all solution for detection. The hairpin and duplex populations were analyzed using Genetool Gel Analysis Software from Syngene.

NMR study

NMR samples were prepared by dissolving 1 mM purified DNA samples into 500 μl NMR buffer and put into 5 mm Wilmad PP528 NMR tubes. Before NMR experiments, DNA samples were first heated to 95°C for 5 min and then immediately placed in an ice-water bath for another 5 minutes in order to promote the hairpin population. This fast cool procedure allows the detection and investigation of the structural features from the hairpin conformers. DSS was used to serve as an internal chemical shift reference standard and its most upfield signal was set to 0 p.p.m. All experiments were performed using a Bruker ARX-500 spectrometer operating at 500.13 MHz. The spectrometer is equipped with a BGU2 gradient unit and a BGPA 10 gradient power amplifier. All NMR data were acquired at 25°C unless stated otherwise. They were processed using Bruker XWIN-NMR software.

For studying the exchangeable imino protons, the samples were dissolved in 90% H_2O/10% D_2O NMR buffer. 1D proton NMR experiments were performed using the water suppression by gradient-tailored excitation (WATERGATE) pulse sequence (26,27). 2D WATERGATE-NOESY experiments were performed at 300 ms mixing time with the time-proportional phase incrementation (TPPI) method (28). Spectral width was set to 21 p.p.m. with the carrier frequency placed at the water signal. The delay for the binomial water suppression was set to 67 μs. A pair of 1 ms gradient pulse at 16 G/cm was used in the WATERGATE sequence. For experiments involving only non-exchangeable protons, the solvent was exchanged to 99.96% D_2O and a 2 s presaturation pulse was applied before the acquisition pulse to suppress the residual HDO signal. 2D NOESY experiments were also performed at 300 ms mixing time with the TPPI method. Spectral width was set to 11 p.p.m. For both types of NOESY experiments, the recycling delays were set to 2 s and 512 FIDs were collected. Each FID consists of 4096 complex data points and at least 32 scans were accumulated. The acquired data matrix was finally zero-filled to give a 4k × 4k data set with cosine window function applied to both dimensions.

Translational diffusion coefficients were determined at least three times for each sample using the bipolar pulse pairs longitudinal encode–decode (BPP-LED) pulse sequence (29) and the following equation (30):

$$-\ln(I/I_0) = D_t\gamma_H^2\delta^2G_z^2(\Delta - \delta/3)$$

where D_t is the translational diffusion coefficient, I and I_0 are the measured and maximum peak intensities, respectively, γ_H is the gyromagnetic ratio of a proton, δ is the gradient length, Δ is the time between the gradients and G_z is the gradient strength. A plot of $-\ln(I/I_0)$ versus $\gamma_H^2\delta^2G_z^2(\Delta - \delta/3)$ gives a slope equal to D_t. Full gradient strength was calibrated using a D_t of 1.902×10^{-5} cm^2/s for a 99.9% D_2O sample (31).

RESULTS

Electrophoretic and NMR studies on 10 CTG repeat hairpins, namely $(CTG)_n$, where n = 1–10, have been performed. Figure 1B shows the exact sequence and the numbering scheme for each nucleotide. Based on the NMR characteristics of the imino and methyl protons, three types of hairpin structures have been identified in CTG repeat sequences (Figure 2). They are Type I: hairpin with no intra-loop hydrogen bond; Type II: hairpin with a 4 nt TGCT-loop; and Type III: hairpin with a 3 nt CTG-loop.

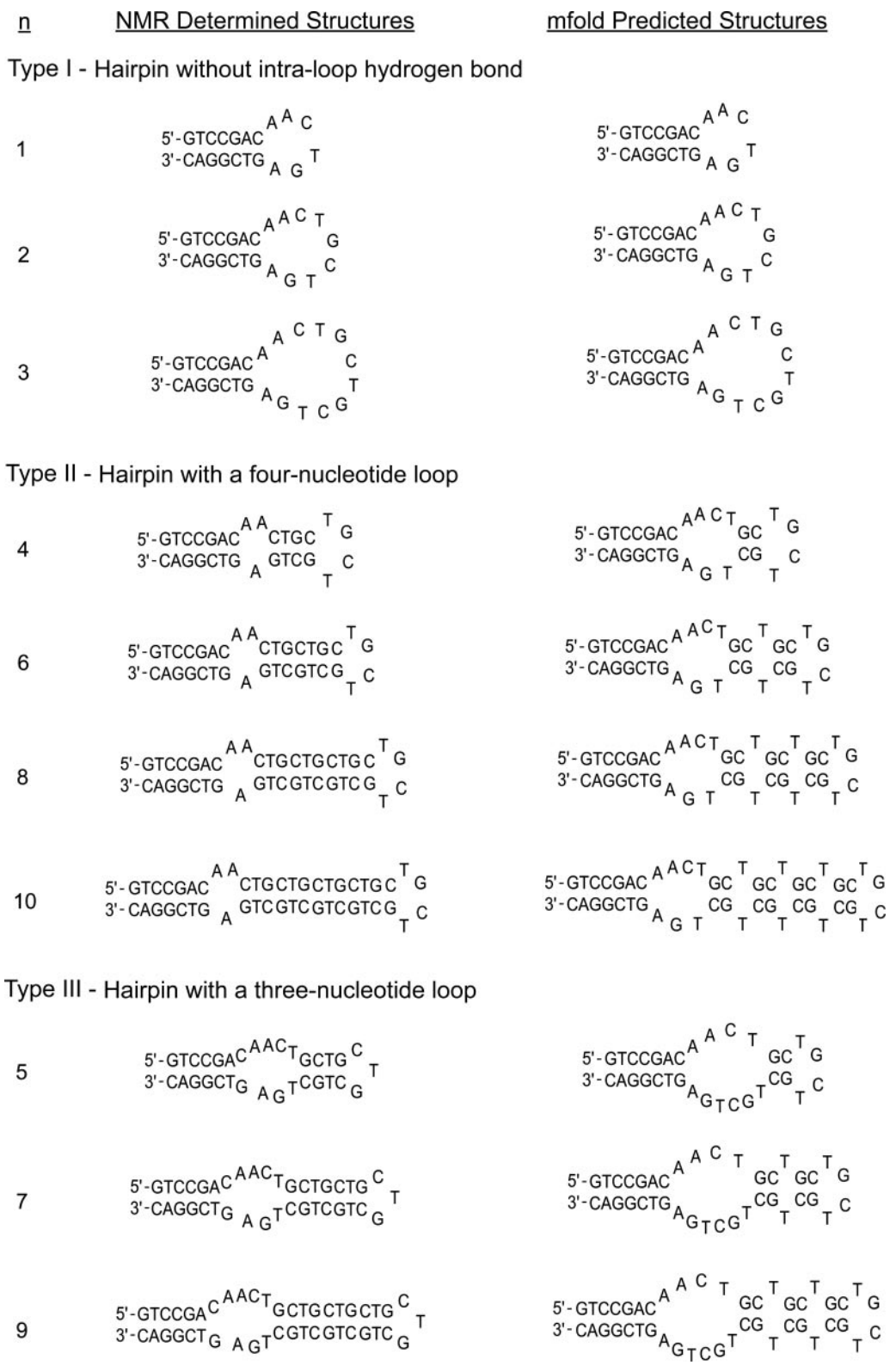

Figure 2. NMR determined and mfold predicted structures of $(CTG)_n$.

Electrophoretic results indicate the hairpin formation propensities of $(CTG)_n$. Figure 3A and B shows the non-denaturing gel pictures of $(CTG)_n$ at 10 μM and 1 mM, respectively. In general, two bands were observed for all CTG sequences except $(CTG)_1$ and $(CTG)_2$. Owing to the higher electrophoretic mobility of the hairpin conformer, the lower bands were assigned to the hairpin conformers, whereas the upper bands were assigned to the duplex conformers.

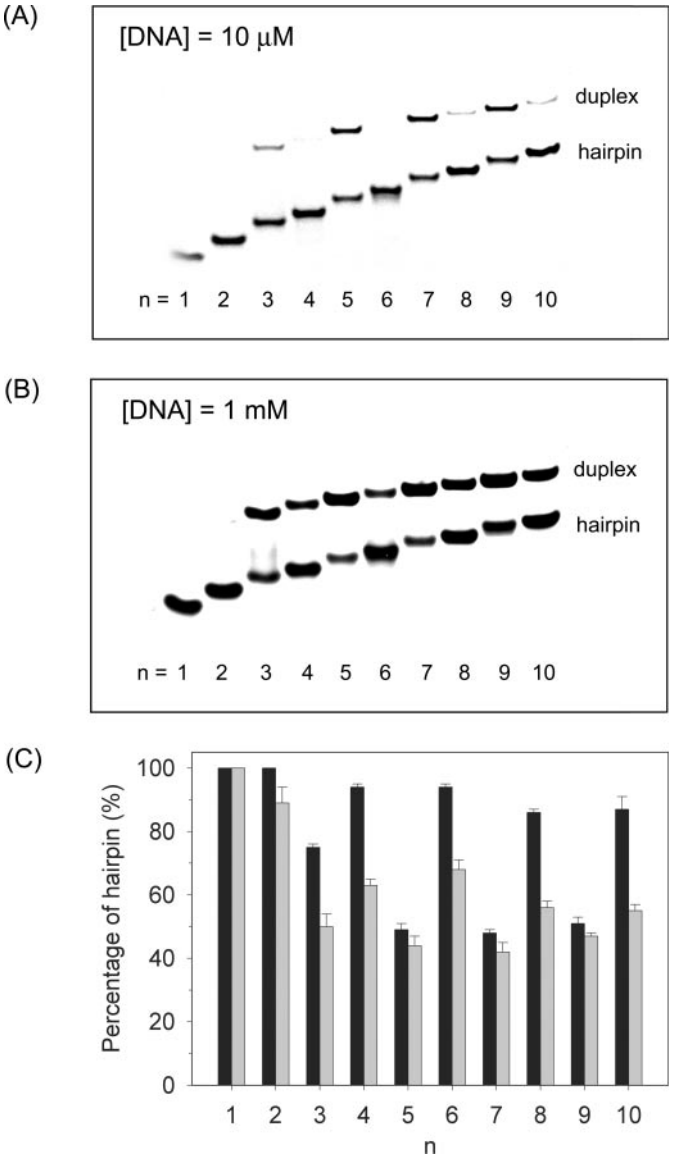

Figure 3. Non-denaturing gel of $(CTG)_n$ at (**A**) 10 μM and (**B**) 1 mM. (**C**) Hairpin populations of $(CTG)_n$ determined at 10 μM (black) and 1 mM (gray). The populations were calculated from the average of four trials and the uncertainties were determined from the standard deviations of the average values.

The hairpin populations as determined from the band intensities are shown in Figure 3C. At 10 μM sample concentration, both $(CTG)_1$ and $(CTG)_2$ adopt only the hairpin conformation. A decrease in the hairpin population was observed when the repeat length was increased to three. The hairpin population was predominant when increased to four. All sequences containing even number of CTG repeats appear to have a higher hairpin population than the duplex. For $(CTG)_4$, $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$, their hairpin populations were higher than those in $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$. At 1 mM sample concentration, there remain significant populations of hairpins, allowing NMR investigation of the hairpin structures. The structural details of these hairpins have been obtained to account for the different hairpin formation propensities observed between even and odd number of repeats. Based on the NOESY spectra, sequential assignments of the aromatic H6/H8, sugar H1′ and imino protons have been completed for $(CTG)_1$ and $(CTG)_4$ (Supplementary Material S1 and S2). As most of the other sequences showed serious spectral overlap in their NOESY spectra, very limited assignment information has been obtained. Therefore, subsequent assignments of the imino and methyl protons and structural analysis of these sequences were carried out with the help of the assignment results of $(CTG)_1$ and $(CTG)_4$.

Type I: hairpin without intra-loop hydrogen bond

$(CTG)_1$, $(CTG)_2$ and $(CTG)_3$ were found to adopt hairpin structure of this type. For $(CTG)_1$, the proton NMR spectrum shows only a single set of proton peaks, indicating this sequence adopts a single conformation. $(CTG)_1$ contains 20 nt and its D_t was found to be 1.22×10^{-6} cm^2/s (s = ±0.01) from the intensities of T2 and T9 methyl peaks. This value is similar to the extrapolated zero concentration D_t (1.23×10^{-6} cm^2/s, s = ±0.02) of a dodecamer duplex that contains 24 nt (32), supporting $(CTG)_1$ is monomolecular. Based on the resonance assignments of the aromatic peaks, the imino signals were also assigned and they were all from the expected hairpin stem region (Figure 4A). The absence of imino signals from TR1 and GR1 suggests that $(CTG)_1$ contains no intra-loop hydrogen bond.

Figure 4B shows the methyl assignments of $(CTG)_1$, $(CTG)_2$ and $(CTG)_3$. For $(CTG)_2$, apart from T2 and T9 (1.33 and 1.35 p.p.m.) in the expected hairpin stem region and TR1 and TR2 (1.77 and 1.74 p.p.m.) in the expected loop region, two weak methyl signals corresponding to a small amount of duplex TR1·TR2 appeared between 1.5 and 1.6 p.p.m. (Figure 4B). The methyl assignments were confirmed by a D_t of 0.83×10^{-6} cm^2/s (s = ±0.04) for the duplex signals and a D_t of 1.26×10^{-6} cm^2/s (s = ±0.02) for the hairpin signals. As the imino region of $(CTG)_2$ shows no imino signals from the loop region (Figure 4A), no intra-loop hydrogen bond is suggested to be present in $(CTG)_2$ hairpin.

For $(CTG)_3$, the resonance assignments were based on the assignment results of $(CTG)_1$, $(CTG)_2$ and extended $(CTG)_3$. In the methyl region (Figure 4B), TR1, TR2 and TR3 of the hairpin loop are located more downfield than those of the duplex. In the imino region (Figure 4A), T9 (13.51 p.p.m.) and G8 (12.46 p.p.m.) of the duplex were found to be slightly different from those of the hairpin (13.54 and 12.40 p.p.m., respectively). These differences are probably due to a small structural variation at the stem–loop junctions of the hairpin and duplex. Although two T·T mismatch imino peaks were observed at 10.76 and 10.52 p.p.m., these signals did not imply the presence of intra-loop hydrogen bonds in the hairpin structure of $(CTG)_3$ because both of them were originated from the minor duplex conformer. To verify this, the sample was treated with the slow cool procedure that allows some kinetically trapped $(CTG)_3$ hairpins to convert to the duplex conformer. Figure 5 shows the imino and methyl regions of $(CTG)_3$ after the slow cool treatment. The G8 and T9 imino protons and the TR1, TR2 and TR3 methyl protons of the hairpin conformer were found to decrease but not the T·T mismatch imino protons, indicating these T·T signals belong to the duplex conformer.

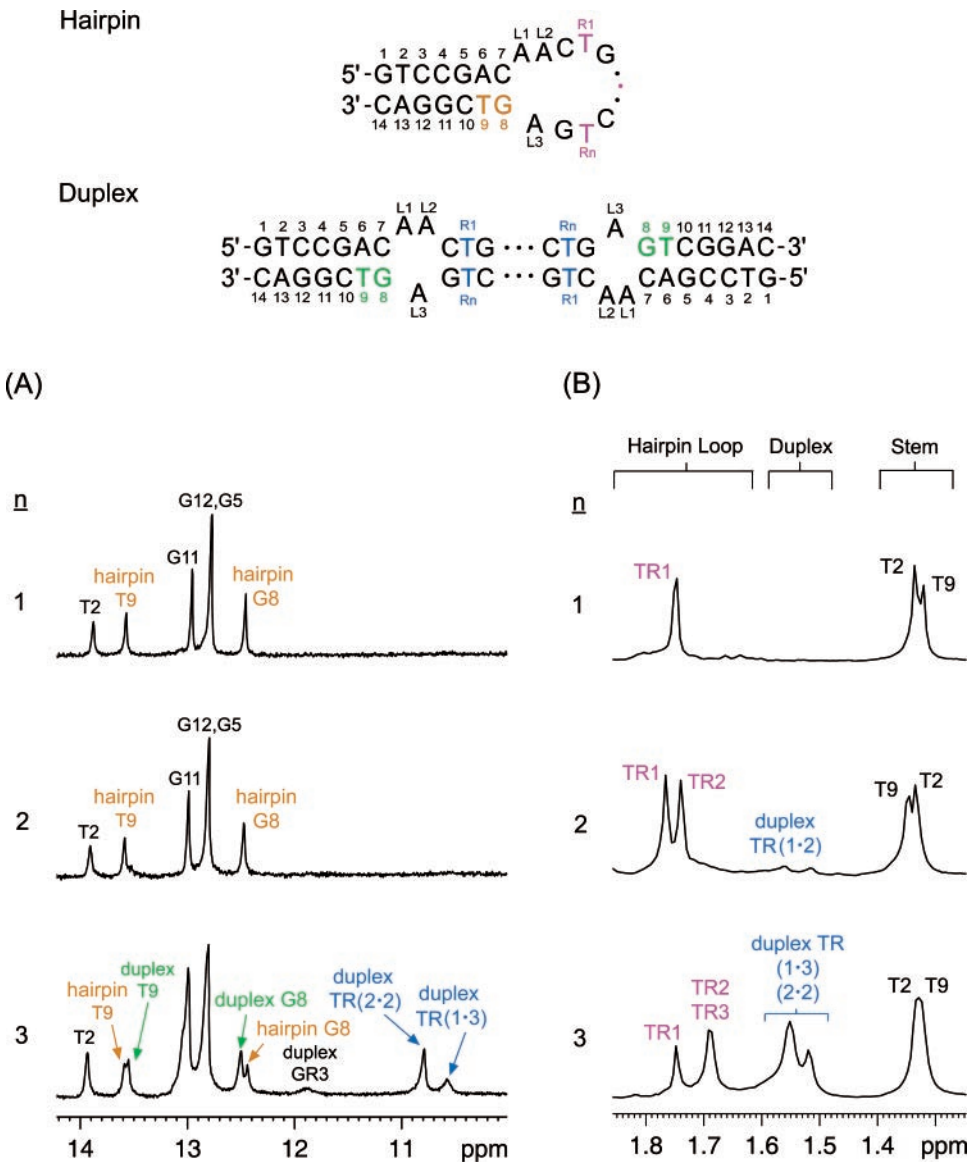

Figure 4. (**A**) Imino and (**B**) methyl regions of $(CTG)_n$ with n = 1–3. The label TR(m·n) refers to the assignment of TRm·TRn.

To further confirm the absence of intra-loop hydrogen bond in the $(CTG)_3$ hairpin, the proton NMR spectrum of an extended $(CTG)_3$ sequence with enhanced hairpin population was acquired. This extended sequence contains two additional G–C base pairs in the hairpin stem of $(CTG)_3$. A simultaneous decrease in the peak intensities of the T·T mismatch imino, duplex methyl and duplex G8, T9 and GR3 imino protons was observed (Figure 5A), confirming no intra-loop hydrogen bond. With the consideration of symmetry, the larger imino peak at 10.76 p.p.m. was assigned to duplex TR2·TR2 and the one at 10.52 p.p.m. was assigned to duplex TR1·TR3. These T·T mismatch chemical shift values are similar to those found in several duplexes (33). The possibility that these signals correspond to flipped-out solvent exposed thymines is unlikely because their imino signals are usually not observed at room temperature owing to their faster exchange rates with water (34,35).

Type II: hairpin with a 4 nt loop

The hairpins of $(CTG)_4$, $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$ were found to contain a 4 nt TGCT-loop closed by a 5′-C 3′-G base pair, in which the C is on the 5′ end and the G is on the 3′ end of the loop. CTG repeats in the proposed loop region were found to interact with each other, forming a T·T-containing stem region that separates the 1 × 2 internal loop and the TGCT-loop (Figure 2). This structure was supported by the appearance of (i) additional Watson–Crick guanine imino signals and (ii) T·T mismatch imino signals from CTG repeats of the hairpin conformer. Figure 6 shows the imino region of $(CTG)_4$ and the resonance assignments were completed (Supplementary Material S2). Apart from the imino signals from the hairpin stem region, additional guanine signals GR1, GR3 and GR4 were observed, indicating the presence of Watson–Crick base pairs GR1–CR4, GR3–CR2 and GR4–CR1, respectively.

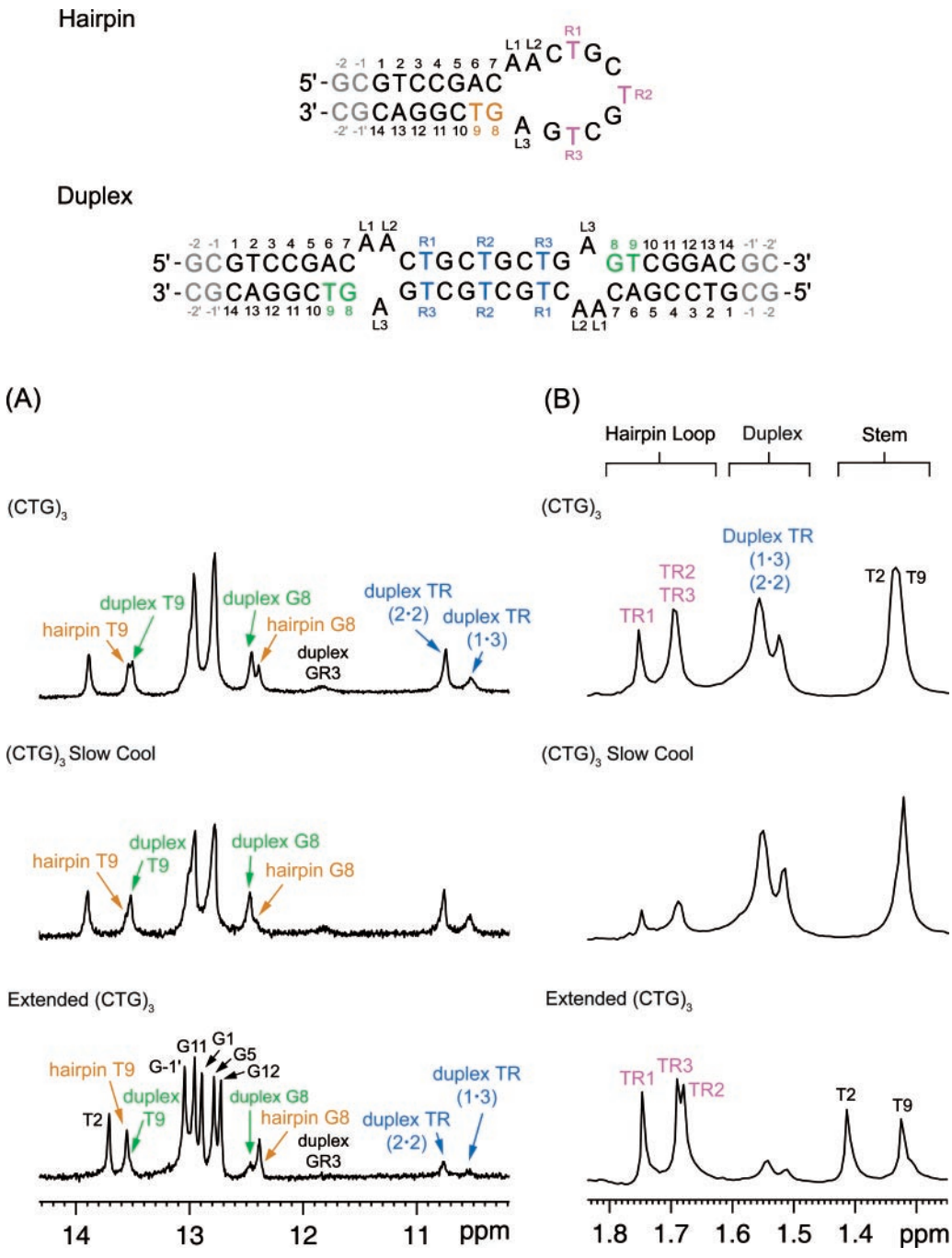

Figure 5. (**A**) Imino and (**B**) methyl regions of $(CTG)_3$, $(CTG)_3$ after slow cool treatment and extended $(CTG)_3$. By comparing the G8 and T9 imino intensities and the methyl intensities of the hairpin and duplex conformers, the duplex population was found to increase in the slow cool state but decrease in the extended $(CTG)_3$ sequence.

To confirm these hairpin assignments, the spectrum of $(CTG)_4$ was re-acquired after the slow cool treatment to promote the duplex population. The imino signals from GR1 and GR3 were found to decrease (Figure 6), indicating these signals belong to the hairpin conformer. The intensity of GR4 was found to remain unchanged, indicating this peak belongs to both the hairpin and duplex conformers. As the structures near the 1×2 internal loop regions between both conformers are expected to be similar, it is reasonable that both conformers have the same GR4 chemical shift.

Owing to the structural similarity in the duplex conformers of $(CTG)_3$ and $(CTG)_4$, the two $(CTG)_4$ imino peaks at 10.53 and 10.76 p.p.m. were assigned to TR1·TR4 and TR2·TR3, respectively (Figure 6). The remaining peak at 10.60 p.p.m. was assigned to TR1·TR4 of the hairpin conformer. To verify these, NMR studies on an extended $(CTG)_4$ sample was performed. The extension of the stem region with four additional G–C base pairs causes an increase in the hairpin population and a decrease in the duplex population. As both the imino peaks at 10.53 and 10.76 p.p.m. were found to decrease

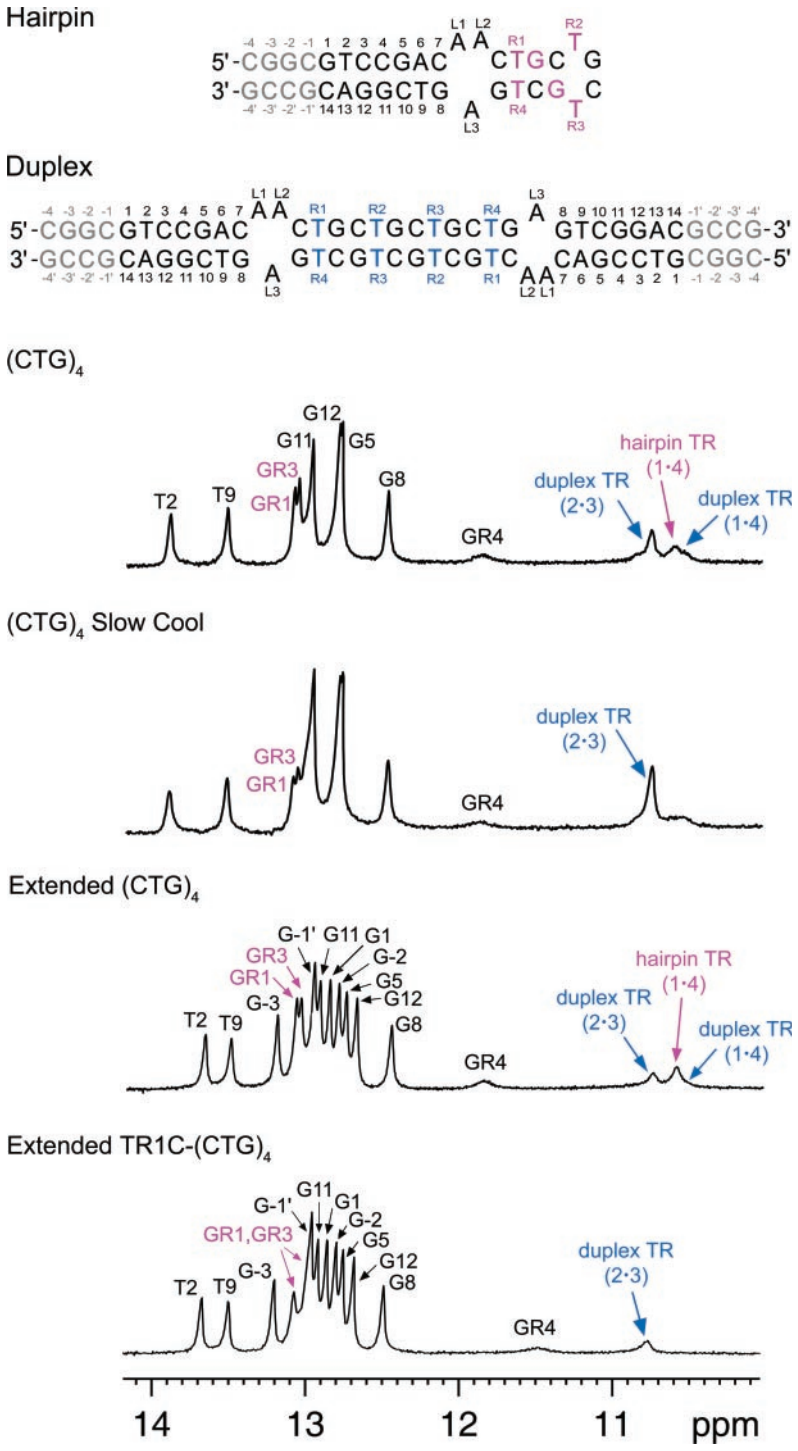

Figure 6. Imino regions of $(CTG)_4$, $(CTG)_4$ after slow cool treatment, extended $(CTG)_4$ and extended TR1C-$(CTG)_4$. The extended $(CTG)_4$ sample contains four additional G–C base pairs at the stem terminals. The extension of the stem region promotes the hairpin population and thus a decrease in the duplex population. The imino peaks at 10.53 and 10.76 p.p.m. were found to decrease accordingly, confirming that these two peaks belong to the duplex conformer. In the extended TR1C-$(CTG)_4$ sample, TR1 was substituted by a cytosine nucleotide TR1C, the imino region shows no signal at 10.53 and 10.60 p.p.m. As no TR1C·TR4 pairing was expected in both the hairpin and duplex conformers, the absence of signal at these chemical shifts confirms the assignments of TR1·TR4 in the $(CTG)_4$ duplex and hairpin, respectively.

(Figure 6), these two peaks were confirmed to belong to the duplex conformer. The hairpin and duplex TR1·TR4 imino assignments were also confirmed by the disappearance of the imino signals at 10.60 and 10.53 p.p.m. in the extended TR1C-$(CTG)_4$ spectrum (Figure 6). In this sample, TR1 was substituted with a cytosine nucleotide and thus no TR1·TR4 pairing was expected in both the hairpin and duplex conformers.

For $(CTG)_6$ hairpin, two pairs of CTG repeats, namely R1–R6 and R2–R5, were found to interact with each other. The 4 nt loop was closed by CR3–GR4 base pair. This structure was again supported by the presence of corresponding signals in the imino spectrum of $(CTG)_6$. By comparing the imino region of $(CTG)_6$ with that of $(CTG)_4$ in Figure 7, the most upfield T·T mismatch signal was assigned to TR1·TR6 of both the hairpin and duplex conformers owing to the similarity of their local structures. The peak at 10.76 p.p.m. was assigned to duplex TR2·TR5 and TR3·TR4, whereas the sharp peak at 10.84 p.p.m. was assigned to hairpin TR2·TR5.

Similar hairpin structures and imino signals were also observed for $(CTG)_8$ and $(CTG)_{10}$. Their imino assignments shown in Figure 7A were made according to the assignment

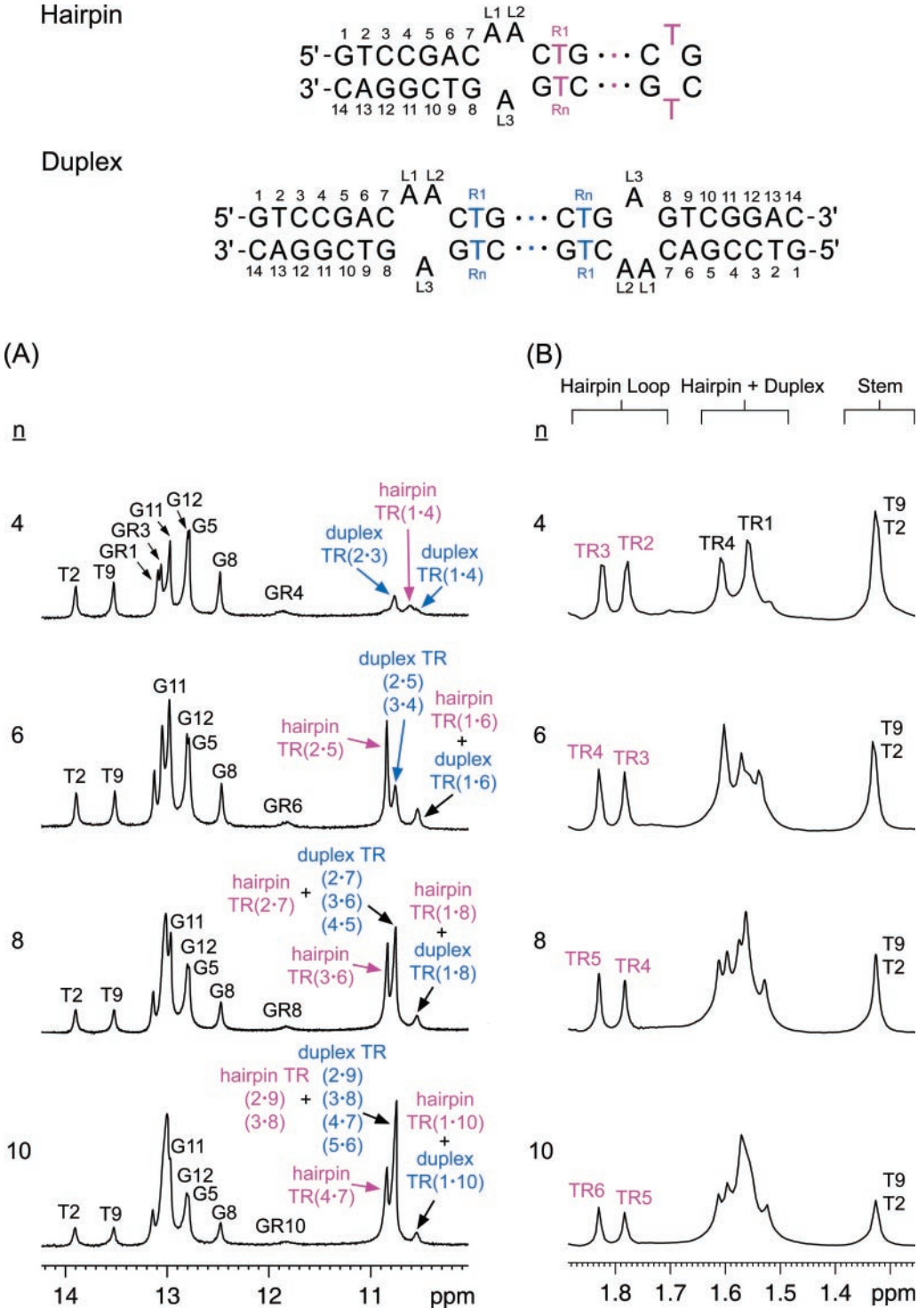

Figure 7. (**A**) Imino and (**B**) methyl regions of $(CTG)_n$ with n = 4, 6, 8 and 10.

results of $(CTG)_6$. For $(CTG)_8$, the most upfield imino signal was assigned to TR1·TR8 of both the hairpin and duplex conformers. The imino signal at 10.76 p.p.m. was found to belong to hairpin TR2·TR7 and duplex TR2·TR7, TR3·TR6 and TR4·TR5. The imino protons of hairpin TR3·TR6 appear at 10.84 p.p.m. For $(CTG)_{10}$, owing to the structural similarity when compared with $(CTG)_6$ and $(CTG)_8$, the most upfield imino signal was assigned to TR1·TR10 of both the hairpin and duplex conformers. The imino protons of hairpin TR2·TR9 and TR3·TR8 and the duplex TR2·TR9, TR3·TR8, TR4·TR7 and TR5·TR6 were found to locate at 10.76 p.p.m., whereas the sharp peak at 10.84 p.p.m. was assigned to hairpin TR4·TR7.

Apart from the imino region, the methyl region also exhibits the characteristics of the TGCT-loop with a 5′-C 3′-G closing base pair. Figure 7B shows the methyl regions of $(CTG)_4$, $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$. Based on the sequential assignment of $(CTG)_4$, the two most downfield methyl resonances at 1.77 and 1.83 p.p.m. were assigned to TR2 and TR3 of $(CTG)_4$ hairpin, respectively. These TGCT-loop signals also appear in the spectrum of $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$. Besides, other thymine methyl signals of the hairpin and the duplex conformers were found to appear between 1.49 and 1.70 p.p.m. The most upfield methyl signal corresponds to the stem T2 and T9 of both conformers.

Type III: hairpin with a 3 nt loop

The hairpins of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$ possess a 3 nt CTG-loop closed by a 5′-G 3′-C base pair (Figure 2). The C7–G8 base pair in the hairpin stem region was weakened when compared with the above two types of hairpins. Apart from the first and last repeats and the repeat involved in the internal loop, the thymine nucleotides in all other repeats were found to interact with each other, forming a T·T-containing stem region that separates the internal loop and the CTG-loop. This hairpin structure was supported by (i) the presence of T·T mismatch imino signals between the two CTG repeats that locate on both sides of the CTG-loop, (ii) the increase in the intensities of these T·T mismatch signals at longer repeat lengths and (iii) the absence of imino signals from hairpin TR1·TR*n*, CR1–GR*n* and C7–G8.

For $(CTG)_5$ hairpin, the third CTG repeat was involved in the loop and the most downfield T·T imino peak at 10.84 p.p.m. was assigned to hairpin TR2·TR4 (Figure 8A). This is based on the resonance assignments of $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$ hairpins. All these T·T mismatches are involved in the CTG repeats that close the loops and thus the chemical environment around these thymine nucleotides is expected to be similar. Other T·T mismatch signals were assigned to the duplex and verified by the spectrum of extended $(CTG)_5$. Four G–C base pairs have been added to the stem region of $(CTG)_5$. Although the sequential assignment of this extended sequence is practically not feasible owing to the severe overlap of cross peaks and the interfering cross peaks from the duplex conformer in the NOESY spectrum, most of the hairpin stem imino assignments were carried out by referencing the assignment results of extended $(CTG)_4$. Because of the increase in the hairpin population of extended $(CTG)_5$ as evidenced by the hairpin methyl signals, the increase in the peak intensity of the most downfield T·T imino signal confirms that this peak belongs to the hairpin. On the other hand, the intensities of the peaks at 10.53 and 10.76 p.p.m. were found to decrease, indicating that they are both from the duplex conformer. With the consideration of symmetry and structural similarity of the CTG repeats in the duplex, the imino peak at 10.53 p.p.m. was assigned to duplex TR1·TR5, whereas the peak at 10.76 p.p.m. was assigned to duplex TR2·TR4 and TR3·TR3.

Apart from the T·T mismatch imino signals, the imino signals of G8 and GR5 were also found to decrease in extended $(CTG)_5$ (Figure 8A), indicating that these signals belong to the duplex. As a result, no evidence has been obtained to support the presence of C7–G8 and CR1–GR5 base pairs in the hairpin. The T9 imino intensity remains similar in the spectra of $(CTG)_5$ and extended $(CTG)_5$, indicating the presence of A6–T9 base pair in the hairpin stem. Further confirmations of the above assignments were obtained from the observed changes in the corresponding imino peak intensities of extended $(CTG)_5$ after the slow cool treatment. Upon a decrease in the hairpin population as evidenced by the hairpin methyl signals (Figure 8B), the imino signals of hairpin TR2·TR4 were found to decrease, whereas those of duplex TR2·TR4, TR3·TR3, GR5 and G8 were found to increase (Figure 8A). By substituting a cytosine nucleotide with TR2, no TR2·TR4 mismatch formation is expected in extended TR2C-$(CTG)_5$. The disappearance of the imino peak at 10.84 p.p.m. and the decrease in the peak intensity at 10.76 p.p.m. indicate that these peaks belong to both the hairpin and duplex TR2·TR4, respectively. The methyl region of this substituted hairpin also assists the assignment of the most downfield signal to TR2 in the $(CTG)_5$ and extended $(CTG)_5$ spectra (Figure 8B).

The imino and methyl regions of $(CTG)_7$ and $(CTG)_9$ are very similar to those of $(CTG)_5$, except a higher peak intensity was observed for the most downfield T·T mismatch imino at 10.84 p.p.m. (Figure 9). Therefore, their hairpin structures are expected to be similar to that of $(CTG)_5$. The imino peak at 10.84 p.p.m. was assigned to hairpin TR3·TR5 in the $(CTG)_7$ spectrum and hairpin TR4·TR6 in the $(CTG)_9$ spectrum. The increase in peak intensity at longer repeat lengths suggests that more CTG repeats interact with each other and thus stabilizes the T·T mismatches near the CTG-loop. As no TR1·TR5 imino signal was observed for $(CTG)_5$ hairpin, no TR1·TR7 and TR1·TR9 mismatches were also expected in the hairpin structures of $(CTG)_7$ and $(CTG)_9$, respectively. Therefore, only two pairs of CTG repeats, namely R2–R6 and R3–R5, were found to interact with each other in $(CTG)_7$. The chemical shifts of the imino signals of hairpin TR2·TR6 and duplex TR2·TR6, TR3·TR5 and TR4·TR4 were found to be the same at 10.76 p.p.m. owing to the similarity of their local structures in both conformers. The imino peak at 10.53 p.p.m. was assigned to duplex TR1·TR7. For $(CTG)_9$, three pairs of CTG repeats, namely R2–R8, R3–R7 and R4–R6, were found to interact with each other. Hairpin TR2·TR8 and TR3·TR7 and duplex TR2·TR8, TR3·TR7, TR4·TR6 and TR5·TR5 have the same imino chemical shift at 10.76 p.p.m. The imino peak at 10.53 p.p.m. was assigned to duplex TR1·TR9.

Figure 9B shows the methyl regions of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$. Owing to the lower hairpin populations compared with sequences containing even number of CTG repeats, the signals from the hairpin conformers are relatively weak.

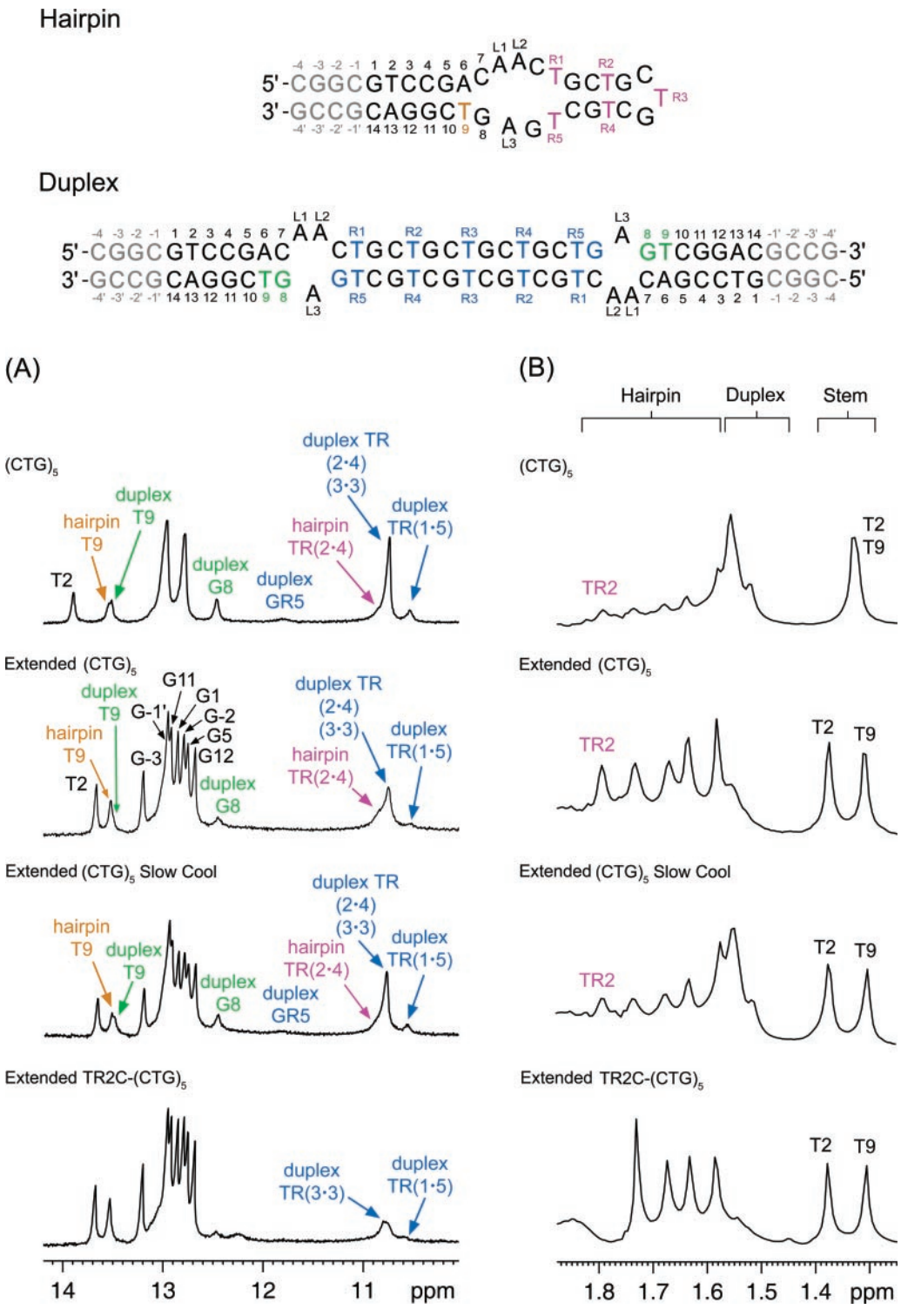

Figure 8. (**A**) Imino and (**B**) methyl regions of $(CTG)_5$, extended $(CTG)_5$, extended $(CTG)_5$ after slow cool treatment and extended TR2C-$(CTG)_5$.

Unlike the hairpins with the TGCT-loop (Figure 7B), no characteristic peak has been identified in the methyl regions for these hairpins with the CTG-loop. The methyl region of extended $(CTG)_5$ hairpin in Figure 8B also differs significantly from those with the TGCT-loop (Figure 7B), indicating that the hairpin loops in these CTG repeat sequences are different.

DISCUSSION

Three types of CTG repeat hairpins have been identified in this study. For the first type in which no intra-loop hydrogen bond was observed, the increase in the number of CTG repeats causes an increase in the loop size. $(CTG)_1$ contains a 6 nt loop,

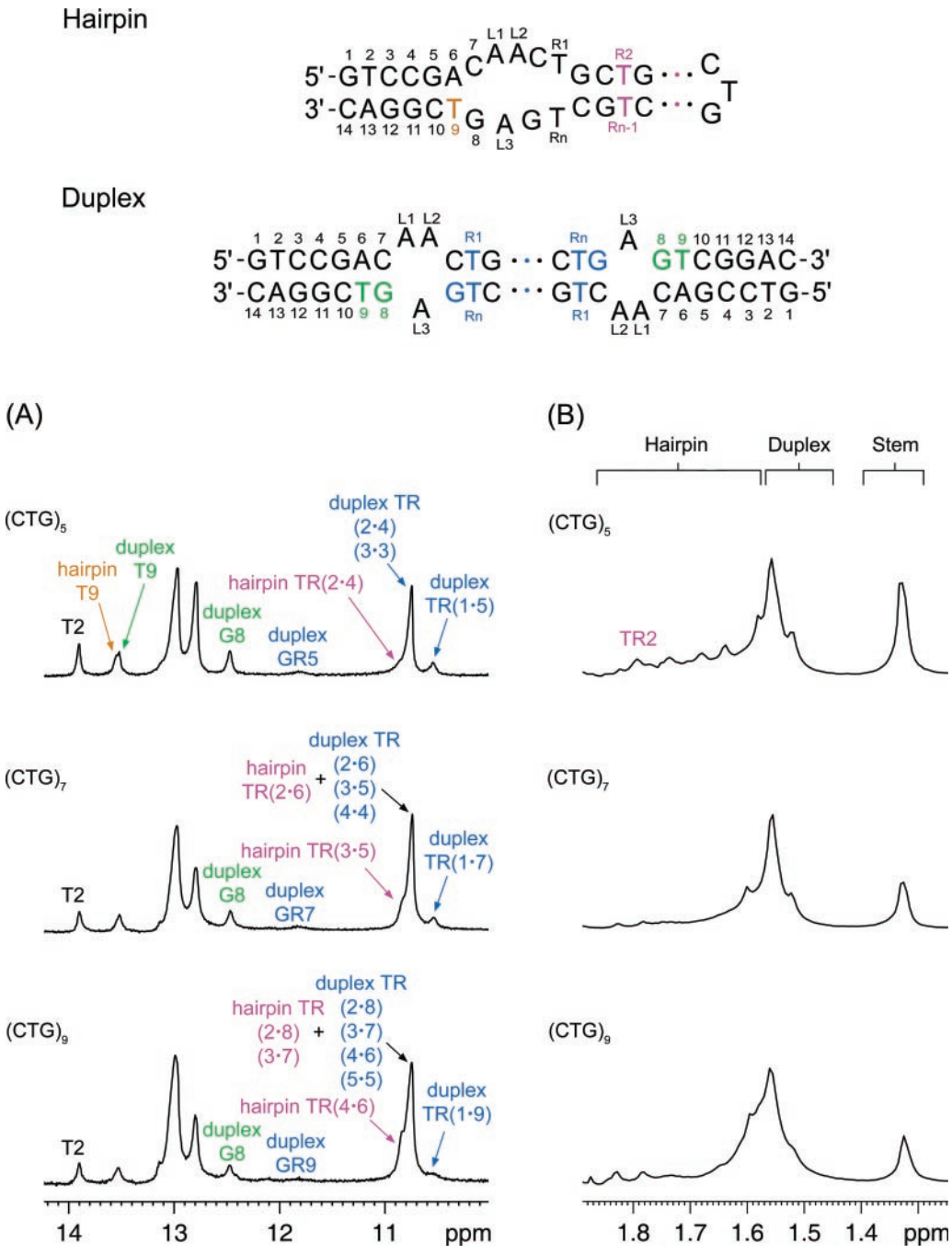

Figure 9. (**A**) Imino and (**B**) methyl regions of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$.

whereas $(CTG)_2$ and $(CTG)_3$ contain a 9 and a 12 nt loop, respectively. The formation of this type of hairpin is probably owing to the presence of the three flanking adenines. However, these adenines reduce the duplex population and allow us to elucidate the structural roles of the CTG repeats in sequences containing 1–3 repeats. The absence of intra-loop hydrogen bond in this type of hairpin suggests that during DNA replication, there is probably no stabilizing interaction in the triplet repeat region when the sequence contains less than four CTG repeats and thus the formation of slipped-strand structures is not likely.

Figure 2 shows the results obtained using the mfold web server (36), which predicts the secondary structures of single strand DNA based on the nearest-neighbor model free energies (37). The predicted results agree well with the experimental structures. From the electrophoretic results, increasing the repeat length to three led to a drop in the hairpin population. This indicates that the hairpin formation propensity decreases at longer repeat lengths in this type of hairpins. However, this propensity indeed depends not only on the thermodynamic stability of the hairpin but also on the stability of the duplex. Therefore, an increase in the hairpin stability

does not necessary cause an increase in the hairpin formation propensity.

When the repeat length was increased to four, instead of a further decrease, the hairpin population was found to increase. This is probably related to the formation of the TGCT-loop and the presence of stabilizing interactions between the CTG repeats. The 4 nt TGCT-loop structure is consistent with the previous findings on the loop structure of the sequence containing six CTG repeats (38). Structure predictions using mfold also support this 4 nt loop structure (Figure 2). The only difference is that no base pair is expected between CR1 and GR4, i.e. the 2 nt right next to the 1×2 internal loop.

The TGCT-loop has also been suggested to behave like a 2 nt loop in which the two thymines form hydrogen bonds and stack with the loop-closing base pair (9). In Figure 6, a small shoulder peak was observed at 10.85 p.p.m. in the $(CTG)_4$ spectrum. This is probably owing to the two thymines that weakly interact with each other in the TGCT-loop. The presence of these interactions is supported by the imino spectra at 15 and 5°C (Supplementary Material S3), in which enhancement of the hairpin TR2·TR3 signals was observed owing to the improved stability at lower temperatures.

The stability of the hairpins with the CTG-loop also depends on the repeat length because there are more CTG repeats interacting with each other at longer repeat lengths. For $(CTG)_5$ hairpin, although the guanine imino signals from GR2 and GR4 have not been identified probably due to the fact that they remain unresolved with the guanine peaks from the stem, CR2–GR4 and GR2–CR4 base pairs are likely to exist because these base pairs can improve the stability of TR2·TR4 mismatch through base pair stacking. The appearance of the TR2·TR4 signals implies the presence of these two flanking G–C base pairs. This situation has been observed in the hairpin structures of $(CTG)_4$ and extended $(CTG)_4$. The appearance of the hairpin TR1·TR4 imino signals was accompanied by the hairpin GR1 and GR4 imino signals. GR1–CR5 base pair may also be present in $(CTG)_5$ hairpin owing to the stacking interactions with CR2–GR4 base pair. As no TR1·TR5 imino signal has been observed, GR1–CR5 base pair is probably not very stable.

For the hairpin structures of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$, the TGCT-loop is expected to form according to mfold predictions (Figure 2). However, the present NMR results indicate the formation of the CTG-loop instead of the TGCT-loop. Although thermodynamic findings have shown that DNA hairpin loops of 4–5 nt are more stable than other smaller or larger loops (39–41), the formation of the CTG-loop instead of the TGCT-loop in $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$ allows more nucleotides in the CTG repeats to interact with each other and thereby increasing the overall stabilities of the hairpins. This loop structure is also consistent with previous NMR findings on the loop structure of a DNA sequence containing five CTG repeats (38).

The hairpin populations of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$ were found to be lower than any of those hairpins with the TGCT-loop. When compared with the TGCT-loop, the CTG-loop was found to destabilize the hairpin conformation. In Figure 9, the absence of the hairpin G8 signal indicates that C7–G8 base pair is not stable. The hairpin peaks were slightly broadened probably owing to more dynamical motions in the CTG repeat region. This destabilizing effect exemplifies the importance of the loop structure in affecting the hairpin formation propensity. Apart from the geometric constraints between the 4 and 3 nt loops, the TGCT-loops of $(CTG)_4$, $(CTG)_6$, $(CTG)_8$ and $(CTG)_{10}$ are closed by a 5′-C 3′-G base pair instead of a 5′-G 3′-C base pair for the CTG-loops of $(CTG)_5$, $(CTG)_7$ and $(CTG)_9$. This 5′-pyrimidine 3′-purine arrangement has been found to be particularly favorable for loop closing (42,43) and thereby leading to an increase in the stabilities of hairpins with the TGCT-loop.

For both types of hairpins with the TGCT-loop and the CTG-loop, more CTG repeats were found to interact with each other at longer repeat lengths. Previous findings have shown that hairpin stability increases linearly with repeat length (23). However, the hairpin populations of the same type were found to be similar, indicating that the hairpin formation propensity has no dependence on the repeat length. This is probably owing to the fact that increasing the repeat length not only improves the stability of the hairpin conformer but also the duplex conformer. As more CTG repeats are involved in interacting with each other in the duplex conformer, the enthalpic contribution to the thermodynamic stability is even more pronounced compared with the hairpin conformer. However, this enthalpic contribution is compensated by the entropic effect that the duplex formation requires two individual DNA strands. As a result, no significant change has been observed in the hairpin formation propensity for sequences with different repeat lengths.

During DNA replication, triplet repeat expansion has been proposed to occur via the formation of slipped-strand structures (17–19). Polymerase-catalyzed extensions of self-priming hairpin folds have shown that CAG and CTG repeats predominantly form 4 nt loops that slip in steps of two triplets during expansions (18,19). The results of this study demonstrate that DNA sequences containing even number of CTG repeats have a higher propensity to form hairpins than those containing odd number of repeats. The difference in the propensities between these types of structures is owing to the presence of the TGCT-loop closed by a 5′-C 3′-G base pair instead of the CTG-loop closed by a 5′-G 3′-C base pair. In addition, there are more stabilizing interactions between the CTG repeats in the hairpins with the TGCT-loop than the CTG-loop. All these contribute significantly to the higher hairpin formation propensity in sequences containing even number of CTG repeats, thus lead to the slippage process occurs in step of two triplets.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Dr Victor Hsu at Oregon State University for his comments on the manuscript. The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. CUHK4255/01P). Funding to pay the Open Access publication charges for this article was provided by the Research Grants Council of the Hong Kong Special Administrative Region, China.

Conflict of interest statement. None declared.

REFERENCES

1. Wells,R.D. and Warren,S.T. (eds) (1998). *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA.
2. Cummings,C.J. and Zoghbi,H.Y. (2000) Trinucleotide repeats: mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.*, **1**, 281–328.
3. Chastain,P.D.,II, Eichler,E.E., Kang,S., Nelson,D.L., Levene,S.D. and Sinden,R.R. (1995) Anomalous rapid electrophoretic mobility of DNA containing triplet repeats associated with human disease genes. *Biochemistry*, **34**, 16125–16131.
4. Chastain,P.D. and Sinden,R.R. (1998) CTG repeats associated with human genetic disease are inherently flexible. *J. Mol. Biol.*, **275**, 405–411.
5. Pearson,C.E., Wang,Y.H., Griffith,J.D. and Sinden,R.R. (1998) Structural analysis of slipped-strand DNA (S-DNA) formed in $(CTG)_n$·$(CAG)_n$ repeats from the myotonic dystrophy locus. *Nucleic Acids Res.*, **26**, 816–823.
6. Sinden,R.R., Potaman,V.N., Oussatcheva,E.A., Pearson,C.E., Lyubchenko,Y.L. and Shlyakhtenko,L.S. (2002) Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J. Biosci.*, **27**, 53–65.
7. Veeraraghavan,J., Rossi,M.L. and Bambara,R.A. (2003) Analysis of DNA replication intermediates suggests mechanisms of repeat sequence expansion. *J. Biol. Chem.*, **278**, 42854–42866.
8. Tam,M., Erin Montgomery,S., Kekis,M., David Stollar,B., Price,G.B. and Pearson,C.E. (2003) Slipped (CTG).(CAG) repeats of the myotonic dystrophy locus: surface probing with anti-DNA antibodies. *J. Mol. Biol.*, **332**, 585–600.
9. Darlow,J.M. and Leach,D.R. (1995) The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *Escherichia coli* suggest hairpin folding preferences *in vivo*. *Genetics*, **141**, 825–832.
10. Mitas,M. (1997) Trinucleotide repeats associated with human disease. *Nucleic Acids Res.*, **25**, 2245–2254.
11. Ashley,C.T.,Jr and Warren,S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.
12. Gellibolian,R., Bacolla,A. and Wells,R.D. (1997) Triplet repeat instability and DNA topology: an expansion model based on statistical mechanics. *J. Biol. Chem.*, **272**, 16793–16797.
13. Pearson,C.E. and Sinden,R.R. (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr. Opin. Struct. Biol.*, **8**, 321–330.
14. Vincent,J.B., Paterson,A.D., Strong,E., Petronis,A. and Kennedy,J.L. (2000) The unstable trinucleotide repeat story of major psychosis. *Am. J. Med. Genet.*, **97**, 77–97.
15. Mosemiller,A.K., Dalton,J.C., Day,J.W. and Ranum,L.P. (2003) Molecular genetics of spinocerebellar ataxia type 8 (SCA8). *Cytogenet. Genome Res.*, **100**, 175–183.
16. Lenzmeier,B.A. and Freudenreich,C.H. (2003) Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. *Cytogenet. Genome Res.*, **100**, 7–24.
17. Ohshima,K. and Wells,R.D. (1997) Hairpin formation during DNA synthesis primer realignment *in vitro* in triplet repeat sequences from human hereditary disease genes. *J. Biol. Chem.*, **272**, 16798–16806.
18. Petruska,J., Hartenstine,M.J. and Goodman,M.F. (1998) Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J. Biol. Chem.*, **273**, 5204–5210.
19. Hartenstine,M.J., Goodman,M.F. and Petruska,J. (2000) Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *J. Biol. Chem.*, **275**, 18382–18390.
20. Petruska,J., Arnheim,N. and Goodman,M.F. (1996) Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res.*, **24**, 1992–1998.
21. Mariappan,S.V., Silks,L.A.,III, Chen,X., Springer,P.A., Wu,R., Moyzis,R.K., Bradbury,E.M., Garcia,A.E. and Gupta,G. (1998) Solution structures of the Huntington's disease DNA triplets, $(CAG)_n$. *J. Biomol. Struct. Dyn.*, **15**, 723–744.
22. Mitas,M., Yu,A., Dill,J., Kamp,T.J., Chambers,E.J. and Haworth,I.S. (1995) Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$. *Nucleic Acids Res.*, **23**, 1050–1059.
23. Gacy,A.M., Goellner,G., Juranic,N., Macura,S. and McMurray,C.T. (1995) Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell*, **81**, 533–540.
24. Smith,G.K., Jie,J., Fox,G.E. and Gao,X. (1995) DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes. *Nucleic Acids Res.*, **23**, 4303–4311.
25. Barbault,F., Huynh-Dinh,T., Paoletti,J. and Lanceloti,G. (2002) A new peculiar DNA structure: NMR solution structure of a DNA kissing complex. *J. Biomol. Struct. Dyn.*, **19**, 649–658.
26. Piotto,M., Saudek,V. and Sklenar,V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR*, **2**, 661–665.
27. Sklenar,V., Piotto,M., Leppik,R. and Saudek,V. (1993) Gradient-tailored water suppression for ^{1}H-^{15}N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson. Ser. A*, **102**, 241–245.
28. Marion,A. and Wüthrich,K. (1983) Application of phase sensitive two-dimensional correlated spectroscopy (COSY) for measurement of 1H–1H spin–spin coupling constants in proteins. *Biochem. Biophys. Res. Commun.*, **113**, 967–974.
29. Wu,D., Chen,A. and Johnson,C.S.,Jr (1995) An improved diffusion-ordered spectrsocopy experiment incorporating bipolar-gradient pulses. *J. Magn. Reson. A*, **115**, 260–264.
30. Stejskal,E.O. and Tanner,J.E. (1965) Spin diffusion measurements: spins echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.*, **42**, 288–292.
31. Longsworth,L.G. (1960) The mutual diffusion of light and heavy water. *J. Phys. Chem.*, **64**, 1914–1917.
32. Lapham,J., Rife,J.P., Moore,P.B. and Crothers,D.M. (1997) Measurement of diffusion constants for nucleic acids by NMR. *J. Biomol. NMR*, **10**, 255–262.
33. Peyret,N., Seneviratne,P.A., Allawi,H.T. and SantaLucia,J.,Jr (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry*, **38**, 3468–3477.
34. Mooren,M.M., Pulleyblank,D.E., Wijmenga,S.S., van de Ven,F.J. and Hilbers,C.W. (1994) The solution structure of the hairpin formed by d(TCTCTC-TTT-GAGAGA). *Biochemistry*, **33**, 7315–7325.
35. van Dongen,M.J., Mooren,M.M., Willems,E.F., van der Marel,G.A., van Boom,J.H., Wijmenga,S.S. and Hilbers,C.W. (1997) Structural features of the DNA hairpin d(ATCCTA-GTTA-TAGGAT): formation of a G–A base pair in the loop. *Nucleic Acids Res.*, **25**, 1537–1547.
36. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
37. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
38. Mariappan,S.V., Garcoa,A.E. and Gupta,G. (1996) Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucleic Acids Res.*, **24**, 775–783.
39. Blommers,M.J., Walters,J.A., Haasnoot,C.A., Aelen,J.M., van der Marel,G.A., van Boom,J.H. and Hilbers,C.W. (1989) Effects of base sequence on the loop folding in DNA hairpins. *Biochemistry*, **28**, 7491–7498.
40. Haasnoot,C.A., de Bruin,S.H., Berendsen,R.G., Janssen,H.G., Binnendijk,T.J., Hilbers,C.W., van der Marel,G.A. and van Boom,J.H. (1983) Structure, kinetics and thermodynamics of DNA hairpin fragments in solution. *J. Biomol. Struct. Dyn.*, **1**, 115–129.
41. Haasnoot,C.A., Hilbers,C.W., van der Marel,G.A., van Boom,J.H., Singh,U.C., Pattabiraman,N. and Kollman,P.A. (1986) On loop folding in nucleic acid hairpin-type structures. *J. Biomol. Struct. Dyn.*, **3**, 843–857.
42. Davison,A. and Leach,D.R. (1994) Two-base DNA hairpin-loop structures *in vivo*. *Nucleic Acids Res.*, **22**, 4361–4363.
43. Hilbers,C.W., Heus,H.A., van Dongen,M.J.P. and Wijmenga,S.S. (1994) The hairpin elements of nucleic acid structure: DNA and RNA folding. *Nucleic Acids Mol. Biol.*, **8**, 56–104.