Unusual Structures of the Tandem Repetitive DNA Sequences Located at Human Centromeres†

Paolo Catasti,‡ Goutam Gupta,§ Angel E. Garcia,§ Robert Ratliff,‖ Lin Hong,‡ Peter Yau,‡ Robert K. Moyzis,‖ and E. Morton Bradbury*,⊥

Theoretical Biology and Biophysics Group, Division T-10, M/S K710, Life Sciences Division, M/S M881, P.O. Box 1663, and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, and Department of Biological Chemistry, School of Medicine, University of California at Davis, Davis, California 95616

Received April 19, 1993; Revised Manuscript Received December 13, 1993®

ABSTRACT: The presence of the highly conserved repetitive DNA sequence $d(AATGG)_n$·$d(CCATT)_n$ in human centromeres argues for a special role for this sequence in recognition, most probably through the formation of an unusual structure during mitosis. Quantitative one- and two-dimensional nuclear magnetic resonance (1D/2D NMR) spectroscopic studies reveal that the Watson–Crick duplex $d(AATGG)_n$·$d(CCATT)_n$ adopts the usual B-DNA conformation as illustrated by taking $d(AATGG)_3$·$d(CCATT)_3$ as an example, whereas the $d(CCATT)_n$ strand is essentially a random coil. In contrast, the $d(AATGG)_n$ strand adopts an unusual stem–loop motif for repeat lengths n = 2, 3, 4, and 6. In addition to normal Watson–Crick A·T pairs, the stem–loop structures are stabilized by mismatched A·G and G·G pairs in the stem and G-G-A stacking in the loop. Stem–loop structures of $d(AATGG)_n$ are independently verified by gel electrophoresis and nuclease digestion studies and were also previously shown to be as stable as the corresponding Watson–Crick duplex $d(AATGG)_n$·$d(CCATT)_n$ [Grady et al. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89,* 1695–1699]. Therefore, the sequence $d(AATGG)_n$ can, indeed, nucleate a stem–loop structure at little free energy cost, and if, during mitosis, it is located on the chromosome surface, it can provide specific recognition sites for kinetochore function.

Genomes of all higher organisms contain subsets of DNA sequences in multiple copies. Whereas many classes of these repetitive DNA sequences are interspersed throughout the genome (Moyzis et al., 1989), other classes are localized, e.g., in extremely long tracts at the telomeric or centromeric regions of the chromosome (Zakian et al., 1989). Tracts of $d(TTAGGG)_n$ [where n refers to the number of repeats] sequences are found at the termini of all human chromosomes and represent the functional human telomere (Moyzis et al., 1988). The ability of telomeric DNA sequences to adopt unusual DNA structures in which the G-rich tracts form a tetraplex stem–loop structure (Kang et al., 1992) is thought to relate to their biological function (Zakian et al., 1989). Other unusual DNA structures of possible biological function include a triple helix through the formation of a stem–loop motif in the purine-rich regions of the eukaryotic gene promoters (Durland et al., 1991) and Z-DNA, which may be involved in transcription and supercoiling (Rich et al., 1984).

The highly conserved repetitive DNA sequences $d(AATGG)_n$ present in human centromeres were shown to have unusually high thermal stabilities, particularly for a sequence lacking perfect Watson–Crick duplex complementarity (Grady et al., 1992). UV melting studies on the synthetic analogs of centromeric DNA repeats suggested that the high stability of this sequence results from the utilization of A·G and possible G·G base pairing (Grady et al., 1992) in an unusual DNA structure. One- and two-dimensional (1D/2D) NMR spectroscopy has been used to determine the three-dimensional structure of $d(AATGG)_n$ and the Watson–Crick duplex $d(AATGG)_n$·$d(CCATT)_n$. The duplex $d(AATGG)_n$·$d(CCATT)_n$ was shown to adopt the usual B-DNA double helix with Watson–Crick G·C and A·T pairs. However, whereas $d(CCATT)_n$ was found to exist in the random-coil state (Grady et al., 1992), the complementary $d(AATGG)_n$ strands displayed stable ordered structures for various repeat lengths n = 2, 3, 4, and 6. Stabilities of $d(AATGG)_n$ structures varied with the repeat length (n) and the ionic conditions. Previously, we have reported structural studies using 1D/2D NMR spectroscopy (Gupta et al., 1987; Garcia et al., 1990) for other DNA sequences and shown that a monomeric hairpin is favored at low DNA concentrations and low salt concentrations, whereas at higher DNA concentrations and higher salt concentrations the duplex is typically favored. A stem–loop structure is stabilized by specific base-pairing schemes in the stem and specific base stacking in the loop. The accurate determination of the stem–loop structure of repetitive DNA sequences requires the characterization of the base-pairing scheme in the stem region, the base stacking arrangement in the loop region, and, most importantly, the conformations of the individual nucleotides. All these structural features of a stem–loop motif can be determined in solution by one- and two-dimensional nuclear magnetic resonance (1D/2D NMR) experiments. Here we report the determination of the solution structures of the highly conserved human centromeric DNA repeats $d(AATGG)_n$ of lengths n = 2, 3, 4, and 6.

In the text below, the single-stranded loop-folded structures (the hairpins or the stem–loop motifs) are denoted as $d(AATGG)_n$, whereas the end-stacked hairpin dimers (also called the stem–loop motifs) are denoted as $[d(AATGG)_n]_2$.

† Support for this work from the U.S. Department of Energy to the Human Genome Center and the Life Sciences and Theoretical Divisions of LANL and to E.M.B. (Grant DE-FG03-88ER60673) is greatly appreciated.

* Corresponding author (telephone, 505-667-2690; FAX, 505-665-3024).

‡ Department of Biological Chemistry, University of California at Davis.

§ Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

‖ Center for Human Genome Studies, Los Alamos National Laboratory.

⊥ Life Sciences Division, Los Alamos National Laboratory.

® Abstract published in *Advance ACS Abstracts,* February 15, 1994.

0006-2960/94/0433-3819$04.50/0

MATERIALS AND METHODS

NMR Experiments. NMR spectra were recorded on a GE-Omega 500 spectrometer. 1D NMR experiments in H_2O were conducted using the 11-echo pulse sequence of Sklenar and Bax (1987). The acquisition parameters for phase-sensitive 2D NOESY/COSY experiments were as follows: sweep width = 5000 Hz, complex data points in t_2 = 2048, complex FIDs in t_1 = 256, number of transients = 32, and relaxation delay = 1.5 s. The mixing times, τ_m, for NOESY experiments were 100 and 250 ms, respectively. The data in t_1 was zero-filled to 1024 before Fourier transformation of the 2048 × 1024 data matrix. The data were not symmetrized.

Sequential Assignment. First, the sequential assignment of the spin system H8/H6, H1′, H2′, H2″ was obtained from the NOESY cross sections H8/H6 vs H2′, H2″ at various mixing times. Second, the spin system H1′, H2′, H2″, H3′, H4′ was sequentially assigned by monitoring the intranucleotide interactions (NOE or *J*-coupling) involving H1′–H2′, H1′–H2″, H2′–H3′, H2″–H3′, H3′–H4′ in the NOESY/COSY cross sections.

Structural Analyses. The following steps were adopted to interpret the 1D/2D NMR data. First, the nature of H-bonding in the structure was characterized by monitoring the temperature dependence and the solvent-exchange properties of the exchangeable imino signals and by performing 1D NOE experiments. Second, a set of interproton distances (i.e., average values and associated dispersions) was extracted for various pairwise interactions by performing full-matrix NOESY simulation and associated *R*-factor tests by comparing the corresponding calculated and observed NOESY intensities (Gupta et al., 1988). The sugar puckers of different residues were estimated by monitoring the corresponding *J*-coupling parameters of the H1′–H2′, H1′–H2″, H2′–H3′, H2″–H3′, etc., interactions in the corresponding phase-sensitive COSY cross sections. Third, these interproton distances were used as structural constraints for constant high-temperature (400 K) 200-ps molecular dynamics (MD) simulations after 3 ps of temperature equilibration. The starting configuration for MD simulation is an energy-minimized structure that satisfies the NOE distance constraints and the observed base-pairing scheme. Fourth, 200 snapshots (one after every 1 ps) were extracted from the MD trajectory, and constrained energy minimization on each snapshot was used to map local minima on the sampled energy surface; this is the *temperature quenching step*. Fifth, 200 energy-minimized structures were assigned to different disjoint clusters such that conformationally similar hairpins belong to the same cluster while conformationally distinct stem–loop structures belong to different clusters (Gupta et al., 1993). Finally, full-matrix NOESY simulation and the associated *R*-factor tests were performed on the representative structures of different clusters to check the agreement with the NOESY data (Gupta et al., 1988). Steps 3–5 are collectively referred to as "high-temperature MD simulation followed by rapid temperature quenching, HTMD/RTQ" (Stillinger & Weber, 1983). During HTMD/RTQ all NOE-derived distance constraints were imposed by using appropriate constraint energy functions. Therefore, all the final 200 energy-minimized structures are in agreement with the NMR data. In order to distinguish local and global rearrangements of atoms or groups among different structures, we defined a hierarchy of structures by progressively dividing structures among different clusters (Gupta et al., 1993). The mean-square distance between all pairs of structures is used as a discriminating parameter for this purpose.

MD and energy minimization were performed using the all-atom force field of Weiner et al. (1986) in AMBER 3.0. All calculations were done *in vacuo* with a constant dielectric coefficient of 78.4 (Garcia & Soumpasis, 1989; Garcia et al., 1990) and without any nonbonding cutoff. High-temperature (400 K) simulations were performed with a set of strong H-bonding constraints (k = 100 kcal mol^{-1} Å^{-2} for the A·T, A·G, and G·G pairs in the structure). Strong constraints are also imposed for interproton distances ≤2.5 Å. The A·G base-pairing constraints in the G-G-A loop segments were not imposed in the calculations.

Gel Electrophoresis and Nuclease Digestion. Electrophoretic patterns of $d(AATGG)_{2,3,4,6}$ were monitored in a nondenaturing gel: 12% polyacrylamide and 0.5 × TBE buffer. Oligonucleotides were labeled by ^{32}P. Samples were heated to 80 °C and then gradually cooled down to 4 °C. Gels were run in a cold room with an ambient temperature of 4 °C. To keep the gel cool, the gel plates were kept in direct contact with the cold circulating buffer. About 12 μL of the sample containing 0.1 mg of DNA was loaded in each lane. No mobility difference was found when the DNA concentration was changed.

The mung bean nuclease (a probe for single-stranded regions in DNA) was used to map the single-stranded regions expected in the stem–loop structures. Oligonucleotides were labeled by ^{32}P before digestion. Reaction conditions were 30 mM sodium acetate, 50 mM NaCl, and 15 μM $ZnCl_2$, pH 5.0. Reactions containing about 20 ng of oligonucleotides with 11 units of mung bean nuclease were run at 0 °C for 2.5, 5.0, and 10.0 min. The reaction was stopped at different times by addition of 50 mM EDTA. Denaturation DNA gels were used: 15% polyacrylamide (20:1) and 7 M urea in Tris buffer.

RESULTS

The structural details are based upon extensive 1D/2D NMR data that include (i) identification of A·T, A·G, and G·G pairs by 1D/2D NOE experiments involving H-bonding (exchangeable) protons (Gupta et al., 1988), (ii) determination of sugar puckers by phase-sensitive COSY and determination of interproton distances by full-matrix NOESY simulations and the associated *R*-factor calculations with respect to the 2D NOESY experiments at two mixing times (Gupta et al., 1988), and finally (iii) derivation of quantitative structures in agreement with the NMR data by HTMD/RTQ (Gupta et al., 1993).

The Hairpin Monomer of $d(AATGG)_2$ and the Stem–Loop Motif of $[d(AATGG)_2]_2$. The 1D NMR spectra of $d(AATGG)_2$ at different solution conditions are shown in Figure 1A. They show the imino, amino, and base proton signals. For low $d(AATGG)_2$ concentration (0.4 mM in strand) and low salt (25 mM NaCl, pH 7), spectrum 1 shows the imino proton signals characteristic of a monomeric hairpin structure. Note that (i) the two imino signals within 13.5–13.2 ppm correspond to two A·T pairs in the stem as confirmed by the irradiation of the imino signals (Figure 1A, spectrum 3), (ii) the broad signal at 11 ppm is due to the imino signals of G's in the loop, (iii) the imino signal from the terminal A·G pair is not clearly visible due to solvent exposure and dynamically rapid open ↔ close exchange, and (iv) the terminal unpaired G shows no imino signal. NOESY experiments (mixing time, τ_m = 250 ms) for this hairpin revealed that all the nucleotides in this hairpin belonged to the C2′-*endo*, *anti* conformation (data not shown) and weaker NOEs were found for the proton pairs in the loop compared to the proton pairs in the stem, as previously reported for other hairpin structures (Gupta et al., 1987; Blommers et al., 1989; Williamson et al., 1989). At the

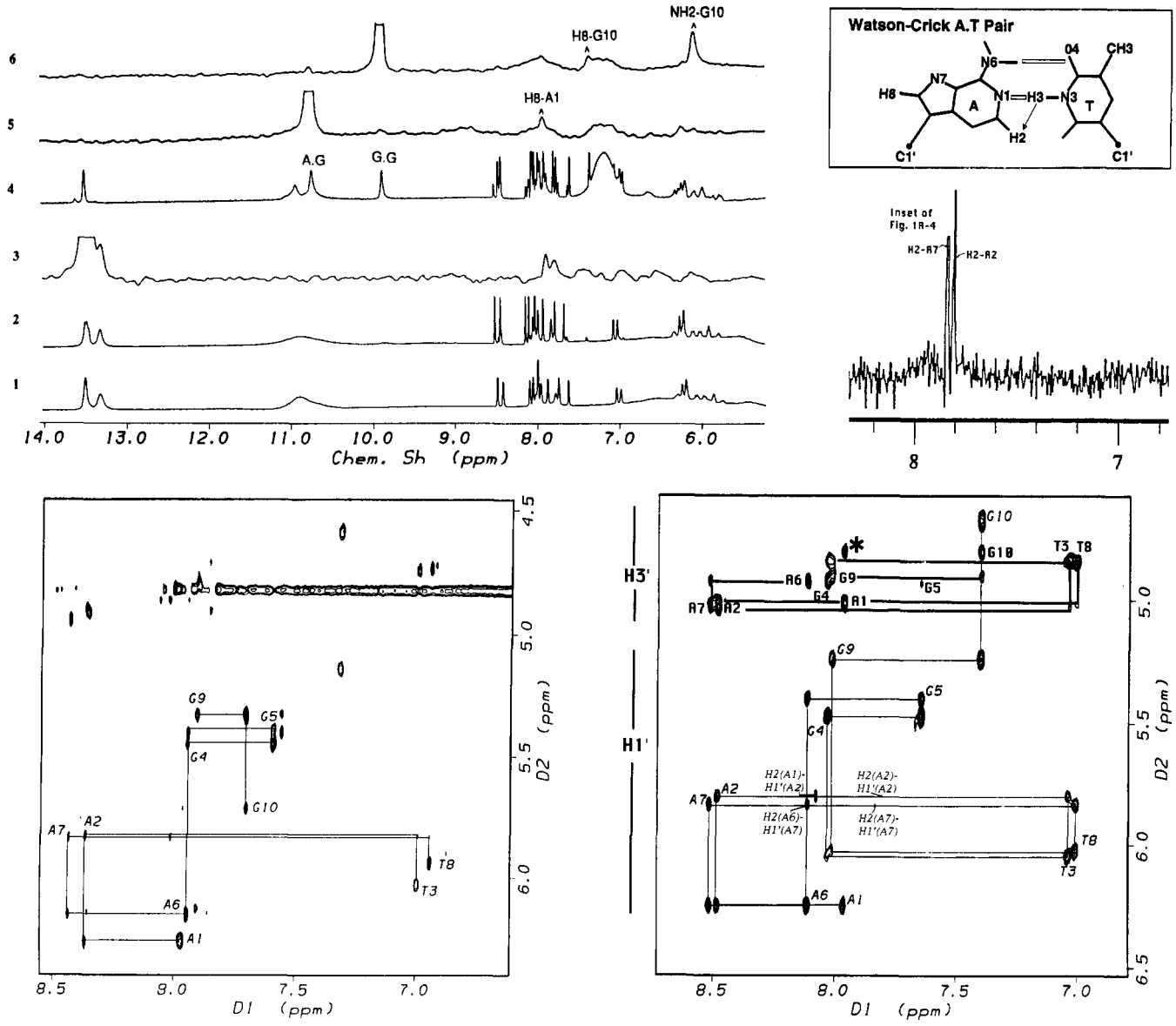

FIGURE 1: (A, top left) 500-MHz 1D proton NMR spectra of d(AATGG)$_2$ showing the imino, amino, and base protons under different conditions of DNA and salt concentrations. The spectra were recorded in a H_2O–D_2O (9:1) mixture using the 1–1 echo pulse sequence of Sklenar and Bax (1987). Spectrum 1 (0.4 mM in strand, 25 mM NaCl, pH 7, temperature 10 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs and a broad signal at 11 ppm due to the imino signals of G's in the loop. Spectrum 2 (2.4 mM in strand, 25 mM NaCl, pH 7, temperature 16 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs and a broad signal at 11 ppm due to the imino signals of G's in the loop. The similarity of spectra 1 and 2 indicates that (AATGG)$_2$ can form a monomeric hairpin over a wide range of DNA concentrations. Spectrum 3 shows a 1D NOE at H2's of A upon irradiation of the low-field imino signal due to one of the A·T pairs (presaturation time = 150 ms). Spectrum 4 (2.4 mM in strand, 1 M NaCl, pH 7, temperature 3 °C) shows the presence of two imino signals (13.5–13.2 ppm) due to two A·T pairs, a broad signal at 11 ppm due to the imino signals of G's in the loop, a signal at 10.8 ppm due to an A·G pair, and a signal at 9.9 ppm due to a G·G pair. Also shown as an inset (top right) are two H2's of A2 and A7 as the sites of NOE when the imino signals of the Watson–Crick A·T pairs are irradiated for 100 ms. The Watson–Crick A·T pair is also shown in the inset along with the expected NOE pathway (shown by an arrow) when N3-H is irradiated. Spectrum 5 shows a 1D NOE at H8 of A1 upon irradiation of the imino signal at 10.8 ppm (presaturation time = 400 ms); such an NOE pattern is consistent with the A1·G9 pair as shown in Figure 2B with an NOE pathway N1-H(G) → N2-H(G) → H8(A) (NOEs are denoted as arrows). Spectrum 6 shows 1D NOEs at the NH_2 region and H8 of G10 upon irradiation of the imino signal at 10.8 ppm (presaturation time = 400 ms); such an NOE pattern is consistent with the G10·G10 pair as shown in Figure 2C, where N1-H(G) is close to N2-H(G) of the same base and H8(G) of the pairing partner (NOEs are denoted as arrows). Therefore, the imino signal and 1D NOE pattern in spectra 4–6 suggest the presence of a stem–loop motif (as shown in panel C) under conditions of high DNA concentration and high salt concentration. The H atom and the acceptor atom involved in H-bonding for the A·T, A·G, and G·G base pairs are joined by an open bar. (B, bottom left) 2D NOESY (τ_m = 250 ms) spectrum of d(AATGG)$_2$ in D_2O for the H1′ vs H8/H6 cross section (2.4 mM in DNA strand, 25 mM NaCl, pH 7, temperature 16 °C). The internucleotide connectivity pattern is indicative of a monomeric hairpin structure (Gupta et al., 1987). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms reveal that all nucleotides adopt a C2′-*endo*, *anti* conformation. Cross-peaks not on the H1′–H8/H6 connectivity route are due to the minor population of a stem–loop motif. (C, bottom right) 2D NOESY (τ_m = 250 ms) spectrum of [d(AATGG)$_2$]$_2$ in D_2O for the H1′,H3′ vs H8/H6 cross section (2.4 mM in DNA strand, 1 M NaCl, pH 7, temperature 3 °C). The intra- and internucleotide NOEs reveal that nine nucleotides (A1 through G9) exist predominantly in *C2′-endo, anti* conformations while one of two G10's shows an *anti* to *syn* conversion to facilitate the G10·G10 pair. Internucleotide NOEs involving H1′(A2/A7)–H2(A1/A6) (weak NOE) and H1′(G9)–H8(G10) (strong NOE) are indicative of special stacking patterns at the T-G and G-G steps, as discussed later in Figure 4. Note the high-field shift of H8,H1′(G10). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms allow us to extract 100 independent interproton distances as independent constraints for structure derivation. Intranucleotide H3′–H8/H6 NOEs are shown. Internucleotide N3′(i–1)–H8/H6(i) NOEs are also observed, but the connectivity pattern is not shown to preserve the clarity of the diagram. H3′ and H1′ chemical shift regions are nonoverlapping except for G10. Note the presence of the intermolecular NOESY cross-peak (marked *) between A1 and G10.

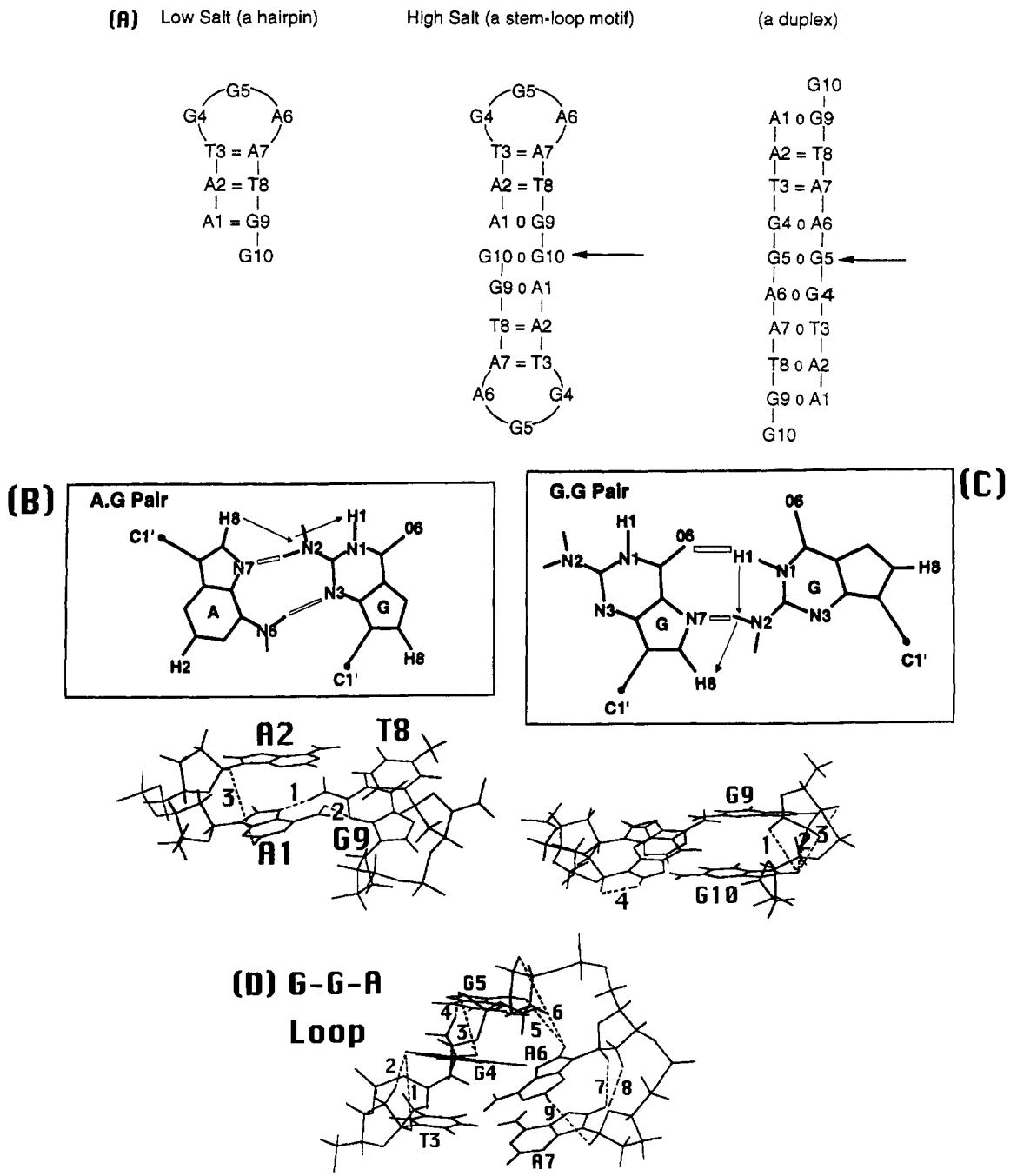

FIGURE 2: Schematic representations of (A) a monomeric hairpin, a stem–loop motif (in which two hairpins anneal with an internal 2-fold symmetry), and a duplex for d(AATGG)$_2$. (B) A·G pairing scheme and the crucial H-bonding and interproton contacts at the A1pA2·T8pG9 step of the representative model of [d(AATGG)$_2$]$_2$ (Figure 7): distance 1 (=2.02 Å), N7(A1)–HN2(G9); the corresponding H-bond angle N7(A1)–HN2(G9)–N2(G9) = 177°; distance 2 (=2.22 Å), N3(G9)–HN6(A1); the corresponding H-bond angle N3(G9)–HN6(A1)–N6(A1) = 132°; distance 3 (=4.11 Å), H2(A1)–H1′(A2). The H-bond lengths and angles show acceptable H-bonding geometry. (C) G·G pairing scheme consistent with the imino proton profile and the NOE pattern at the G9pG10·G10-A1 step of the representative model of [d(AATGG)$_2$]$_2$ (Figure 7). Both G10's can be in *syn* or *anti* conformation (but not simultaneously) in order to preserve the G(*syn*)·G(*anti*) pairing scheme. The following interproton contacts are predicted in the G9(*anti*)pG10(*anti*) fragment from our model of Figure 7: distance 1 (=2.11 Å), H2″(G9)–H8(G10); distance 2 (=2.97 Å), H1′(G9)–H8(G10); distance 3 (=4.46 Å), H3′(G9)–H8(G10). In the G9(*anti*)pG10(*syn*) fragment the only prominent distance is 4 (=2.2 Å), H1′(G10)–H8(G10). Because of the *syn*/*anti* flip-flop of G10 the corresponding observed NOE should reflect the averages of the four distances. (D) Expected NOE pattern in the G-G-A loop of the representative model of [d(AATGG)$_2$]$_2$ (Figure 7): distance 1 (=4.44 Å), H1′(T3)–H8(G4); distance 2 (=2.28 Å), H2″(T3)–H8(G4); distance 3 (=3.73 Å), H1′(G4)–H8(G5); distance 4 (=2.42 Å), H2″(4)–H8(G5); distance 5 (=3.35 Å), H1′(G5)–H8(A6); distance 6 (=5.52 Å), H2″(G5)–H8(A6); distance 7 (=3.88 Å), H1′(A6)–H8(A7); distance 8 (=2.54 Å), H2″(A6)–H8(A7); distance 9 (=3.96 Å), H2(A6)–H1′(A7). Note that G5 is the most flexible part of the G-G-A loop. However, a sampling of different stacking conformations of G5 results in the effective (averaged) unbroken NOE pathway.

higher d(AATGG)$_2$ concentration (2.4 mM in strand), the same hairpin conformation was retained, as evidenced by imino proton spectrum 2 of Figure 1A. Under the conditions of high DNA concentration and low salt concentration, NOESY experiments at τ_m = 250 and 100 ms also confirmed a hairpin structure. Figure 1B shows the H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) typical of a monomeric hairpin (Gupta et al., 1987; Blommers et al., 1989; Williamson et al., 1989). The hairpin motif that is fully consistent with the NMR data is shown in Figure 2A. When the salt concentration is increased to 1 M NaCl (pH 7), two hairpins of d(AATGG)$_2$ form an end-stacked dimer (i.e., [d(AATGG)$_2$]$_2$) as shown

in Figure 2A. The NMR evidence of such a structure is discussed below.

(*i*) *The Presence of Watson–Crick A·T Pairs.* In our model the number of Watson–Crick A·T pairs equals the number of T's present in the sequence. For example, in the stem–loop structure of $[d(AATGG)_2]_2$ (Figure 2A), only two A·T pairs are expected, and this is exactly what we observe by 1D NOE experiment. In this experiment, the imino proton of T is irradiated (inset of Figure 1A, spectrum 4), and a strong NOE is observed at H2 of A even at 100 ms of presaturation time—a characteristic feature of a Watson–Crick A·T pair. NOEs are observed only at two H2's of A, suggesting that there are two Watson–Crick A·T pairs. And out of the four H2's belonging to four A's in $[d(AATGG)_2]_2$, H2's belonging to H-bonded A·T pairs show high-field shifts as expected. By thermal melting and chemical substitution studies, it was also shown that A·T pairs are crucial to the stability of the centromeric structures (Grady et al., 1992).

(*ii*) *The Nature of A·G Pairs.* Once the imino signals above 12 ppm are accounted for by the two A·T pairs, the signals below 11 ppm remain to be identified. As discussed below, they belong to G's in the A·G and G·G pairs and to G's in the loop. The sharp signal at 10.8 ppm belongs to the A·G pairs. It is clear that the A·G pairing is not through the imino proton of G because that should give the imino G signals above 12 ppm. This rules out the possibility of A(*syn*)·G(*anti*) and A(*anti*)·G(*anti*) pairing as observed in the single crystals (Kennard, 1988; Prive et al., 1988). The A(*syn*)·G(*anti*) pairing is also inconsistent with the NMR data because no A was observed in a *syn* conformation. The A(*anti*)·G(*anti*) pairing is also ruled out because the irradiation of the exchangeable signals below 12 ppm did not show any strong NOE at H2 of A (Kan et al., 1983). This leaves two other types of A·G pairings that involve amino protons of G instead of the imino protons (Li et al., 1991). One such pairing that is consistent with our NOE data is shown in Figure 2B. In this pairing, as observed in Figure 1A, spectrum 5, the irradiation of the imino proton of G at 10.8 ppm should show a secondary NOE at H8 of A via NH2 of G. Another additional feature of such an A·G pairing (as shown in Figure 2B) is the NOE between the H2 of A of the A·G pair and the H1′ of the neighboring 3′ A·T pair. Observation of such an NOE is shown in Figure 1C. Even though such an H-bonding, as shown below, has propeller-twisted A·G pairs, it has acceptable geometry and is free of any short sugar–base contacts. In this pairing scheme the imino protons of G do not participate in H-bonding; however, because of A-G-G stacking, the imino proton of the central G involved in the A·G pairing is excluded from solvent and hence not exchanged. Li et al. (1991) also demonstrated such an A·G pairing in a DNA duplex where the imino protons of G were not readily exchanged and located within 10–11 ppm.

(*iii*) *The Nature of G·G Pairing in the Stem–Loop Motif.* If the G·G pairing involved two imino protons of G's, then these two protons are expected to be located at two distinct chemical shifts and strong NOEs are expected between them. For the stem–loop structure of $[(AATGG)_2]_2$ (Figure 2A), the irradiations of the signals at 10.8 ppm did not produce any NOE at 9.9 ppm or vice versa (Figure 1A, spectra 5 and 6) as observed by Cognet et al. (1991) for a DNA duplex with G·G pairs. However, the G·G pairing scheme shown in Figure 2C is consistent with the NOE pattern shown in Figure 1A, spectrum 6, in which irradiation of the G imino signal at 9.9 ppm results in a primary NOE at NH2 of G and secondary NOE at H8 of G. In this pairing scheme G10 is in the *syn* conformation. In view of the fact that a single G10 signal is observed in panels A and C of Figure 1, it appears that there is a rapid *syn*/*anti* flip-flop without exposing the imino proton to the solvent. The observed intranucleotide H1′(G10)–H8(G10) NOE, though strong, is only the average of *syn*/*anti* conformations. The strong internucleotide H1′(G9)–H8(G10) contact is also consistent with the pairing shown in Figure 2C. The chemical shifts of the imino protons of G·G pairs as high-field-shifted as 9.9 ppm are not uncommon in the literature (Cognet et al., 1992; Smith & Feigon, 1992; Gupta et al., 1993). The G·G part happens to be the most flexible region of the stem–loop structures of $[d(AATGG)_2]_2$. Thermal melting and base substitution studies also confirm this observation (Grady et al., 1992); i.e., substitutions of the G·G pairs by any other mismatches have little effect on the melting temperature.

(*iv*) *The NOE Pattern of the G-G-A Loop.* Even under conditions of 1 M NaCl the imino signals of G's in the loop were still present, and these signals, as expected for a hairpin, did not show any specific NOE. Therefore, under the conditions of high DNA concentration and high salt concentration, two hairpins most probably anneal to form a stem–loop motif as shown in Figure 2A. The H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) of the stem–loop motif is shown in Figure 1C. Comparison of panels B and C of Figure 1 shows the following high salt induced changes: (i) a high-field shift of H8,H1′(G10) and strong H1′(G9)–H8(G10) NOEs (with a distance of 2.5–3.0 Å) providing information regarding stacking at the G-G step in the stem, (ii) *anti* to *syn* conversion of one of the G10's (as evidenced by the presence of a strong intranucleotide H1′–H8 NOE) consistent with the G·G pairing shown in Figure 2C, and (iii) emergence of weak but observable internucleotide H1′(A2/A7)–H2(A1/A6) NOEs (with distances of 3.8–4.2 Å) in agreement with the A·G pairing scheme shown in Figure 2B and also providing information regarding stacking at the T-G steps. Comparison of the base pairing in the duplex with the stem–loop motif (Figure 2A) shows a relatively small difference. Whereas the duplex has four internal A·T pairs, two internal A·G pairs, one internal G·G pair, two terminal A·G pairs, and two unpaired G's, the stem–loop motif has the same number of internal A·T, A·G, and G·G pairs. The expected difference in the two structure lies in the fact that two terminal A·G pairs and two unpaired G's in the duplex are replaced in the stem–loop motif by two G-G-A loops. However, as will be discussed later, the NOE constraints and energetic considerations can lead to an A·G pairing involving the first and the third base in the **G-G-A** loop (paired bases are shown in bold). Therefore, a stem–loop motif and a duplex can have the same number of base pairs and the same type of base stacking for all internal base pairs (Figure 2A). In this case, the energy difference in stacking between G in the loop of the stem–loop motif and the terminal G in the duplex essentially determines the difference in stability between these two structures. In the case of a stem–loop structure with three nucleotides (G-G-A) in the loop, the relative stabilities of a duplex and the corresponding end-stacked dimer (from now on referred to as the stem–loop structure) are determined by the difference in energy between *base stacking and base pairing in the duplex and base stacking and loop entropy in the stem–loop structure.*

In summary, the NMR data described above show that the centromeric repeat, $d(AATGG)_2$, at a low salt concentration forms a monomeric hairpin while at a high salt concentration two such hairpins anneal to form a stem–loop structure of $[d(AATGG)_2]_2$ that is stabilized by A·T, A·G, and G·G pairs (Figure 2A). The NOE contacts in the G-G-A loop as expected from our stem–loop structure of $[(AATGG)_2]_2$ are shown in

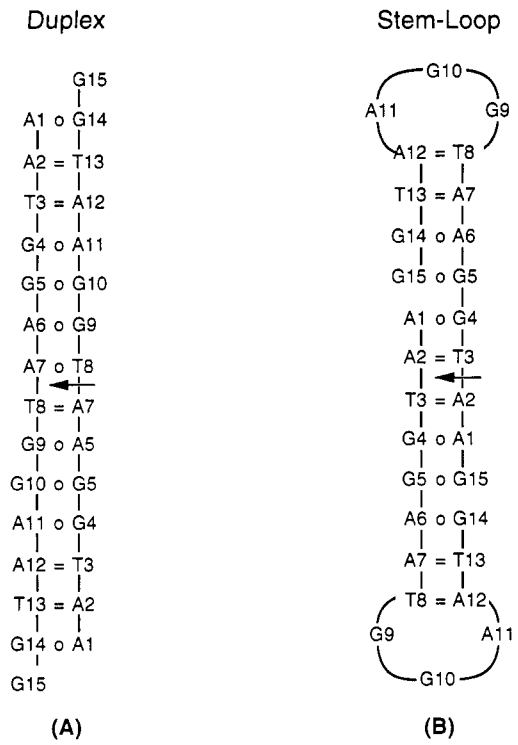

FIGURE 3: Schematic representations of (A) a duplex and (B) a stem–loop motif for $[d(AATGG)_3]_2$.

Figure 2D. Note the presence of continuous H1′(*i*)–H8(*i*+1) NOE contacts in the loop region and also the presence of the H2(A6)–H1′(A7) NOE. The NOE data shown in Figure 1C are in complete agreement with the structural model of Figure 2D. Although the NOE connectivities are often broken in the loop segment of a hairpin, such a criterion is never an absolute requirement of a loop structure. However, H2″(G*i*)–H8-(A*i*+1) NOEs in the loop segment are either weak or absent.

The Stem–Loop Motif of $[d(AATGG)_3]_2$. Results from the preceding NMR studies on $[d(AATGG)_2]_2$ show that the stem–loop motif utilizes all T's in A·T pairs. For $[d(AATGG)_3]_2$, utilization of all three T's in A·T pairing is achieved either in (A) a duplex or in (B) a stem–loop motif as shown in Figure 3. Spectra 1, 2, and 3 in Figure 4A show the imino signals of $d(AATGG)_3$ under different conditions: (i) for 1.8 mM $[d(AATGG)_3]_2$ concentration and low salt concentration (spectrum 1), (ii) for 1.8 mm $[d(AATGG)_3]_2$ and high salt concentration (spectrum 2), and (iii) in the presence of an equimolar complementary Watson–Crick strand, $[d(CCATT)_3]_2$ (spectrum 3). Spectra 1 and 2 show imino signals corresponding to three A·T pairs, two A·G pairs, one G·G pair, and G's in the loop, whereas spectrum 3 shows imino signals typical of Watson–Crick G·C and A·T pairs which were characterized by performing 2D NOESY experiments in (9:1) H_2O–D_2O (data not shown). The nature of the imino proton pattern in spectrum 3 and NOESY experiments in D_2O at various mixing times, τ_m = 150 and 50 ms (data not shown), unequivocally show that $d(AATGG)_3$·$d(CCATT)_3$ forms the usual Watson–Crick B-form duplex. Similar experiments on $d(AATGG)_2$·$d(CCATT)_2$ also indicate the presence of the usual Watson–Crick B-DNA conformation (data not shown). However, spectra 1 and 2 indicate that $d(AATGG)_3$ forms a stem–loop motif of $[d(AATGG)_3]_2$ under a wide range of salt concentrations. Figure 4B shows the H1′ vs H8/H6 NOESY cross section (τ_m = 250 ms) under conditions of high DNA concentration and low salt concentration. Note that NOE signatures of a stem–loop motif, as observed for $[d(AATGG)_2]_2$ in Figure 1C, are also preserved here. These are (i) a high-field shift of H8,H1′(G5/G15) and strong H1′(G4/G14)–H8(G5/G15) NOEs (with distances of 2.5–2.8 Å) providing information regarding stacking at the G-G steps in the stem, (ii) *anti* to *syn* conversion of one of the G5/G15's (as evidenced by the presence of a strong intranucleotide H1′–H8 NOE) consistent with the G·G pairing shown in Figure 2C, and (iii) emergence of weak but observable internucleotide H1′(A2/A7/A11)–H2(A1/A6/A11) NOEs (with distances of 3.8–4.2 Å) in agreement with the A·G pairing scheme shown in Figure 2B and also providing information regarding stacking at the T-G steps. Analysis of NOESY data (τ_m = 250 and 100 ms) of $d(AATGG)_3$ under conditions of high salt also revealed the same stem–loop motif of $[d(AATGG)_3]_2$. However, as will be discussed, in addition to the three-nucleotide G-G-A loop, the NOE constraints and energetic considerations also indicate the possibility of an A·G pairing involving the first and the third base in the **G-G-A** loop (paired bases are shown in bold).

In summary, structural analyses of $[d(AATGG)_{2,3}]_2$ show that the stem–loop structure in each case (Figure 2A and Figure 3B) is stabilized by utilizing all T's in A·T pairs and formation of A·G and G·G pairs (which show enhanced stability at high salt). The melting profiles of $[d(AATGG)_{2,3}]_2$ show that the imino signals of the A·G and G·G pairs disappear at a much lower temperature than the imino signals of the A·T pairs. Therefore, it appears that the base pairing and base stacking at the A-T steps nucleate these structures.

The Stem–Loop Motifs of $[d(AATGG)_{4,6}]$*—an Extension of* $[d(AATGG)_{2,3}]_2$. Detailed analyses of 1D/2D NMR data of $d(AATGG)_{4,6}$ (data not shown) were unambiguous and revealed the presence of the stem–loop motifs shown in panels A and B of Figure 5. The stem–loop motif of $d(AATGG)_4$ (Figure 5A) is similar to the stem–loop motif of $[d(AATGG)_2]_2$ (under conditions of high salt, Figure 2A) except for an additional phosphodiester linkage. Also, the stem–loop motif of $d(AATGG)_6$ (Figure 5B) is similar to the stem–loop motif of $[d(AATGG)_3]_2$ (Figure 3B) except for an additional phosphodiester linkage.

The 1D NMR spectrum of $d(AATGG)_4$ showed the presence of imino signals belonging to four A·T pairs, two A·G pairs, one G·G pair, and the G's in the loop. Similarly, the stem–loop motif of $d(AATGG)_6$ (Figure 5B) is almost the same as the stem–loop motif of $[d(AATGG)_3]_2$ (Figure 3B) except for a phosphodiester covalent linkage. The 1D NMR spectrum of $d(AATGG)_6$ confirms the presence of the imino signals of six A·T pairs, four A·G pairs, two G·G pairs, and the G's in the two G-G-A loops (1D NMR spectra are included in Figure 8 of the supplementary material).

Comparisons of NOESY (τ_m = 250 and 100 ms) cross sections of $d(AATGG)_4$ and those of $[d(AATGG)_2]_2$ at high salt reveal a striking similarity: H8, H1′ of the nucleotides in the A-A, A-T, T-G, T-G, G-G steps in the stem region and H8, H1′ of the nucleotides in the loop G-G-A segment show similar chemical shifts and NOE patterns for the two repeat lengths. Further comparisons of NOESY (τ_m = 250 and 100 ms) cross sections of $d(AATGG)_6$ with those of $[d(AATGG)_3]_2$ show similarities in the chemical shift and NOE pattern (2D NMR spectra are included in Figures 9 and 10 of the supplementary material). Figure 5C shows the average stem–loop structures of $d(AATGG)_{4,6}$. The atoms are color coded (i.e., C = green, N = blue, O = red, P = yellow); the H-atoms are omitted to enhance the clarify of the diagram. In each case the average structure, taken over 200 local minima, is in agreement with the NMR data. In Figure 5C, the stem–loop structure of $d(AATGG)_4$ is shown on the left and the

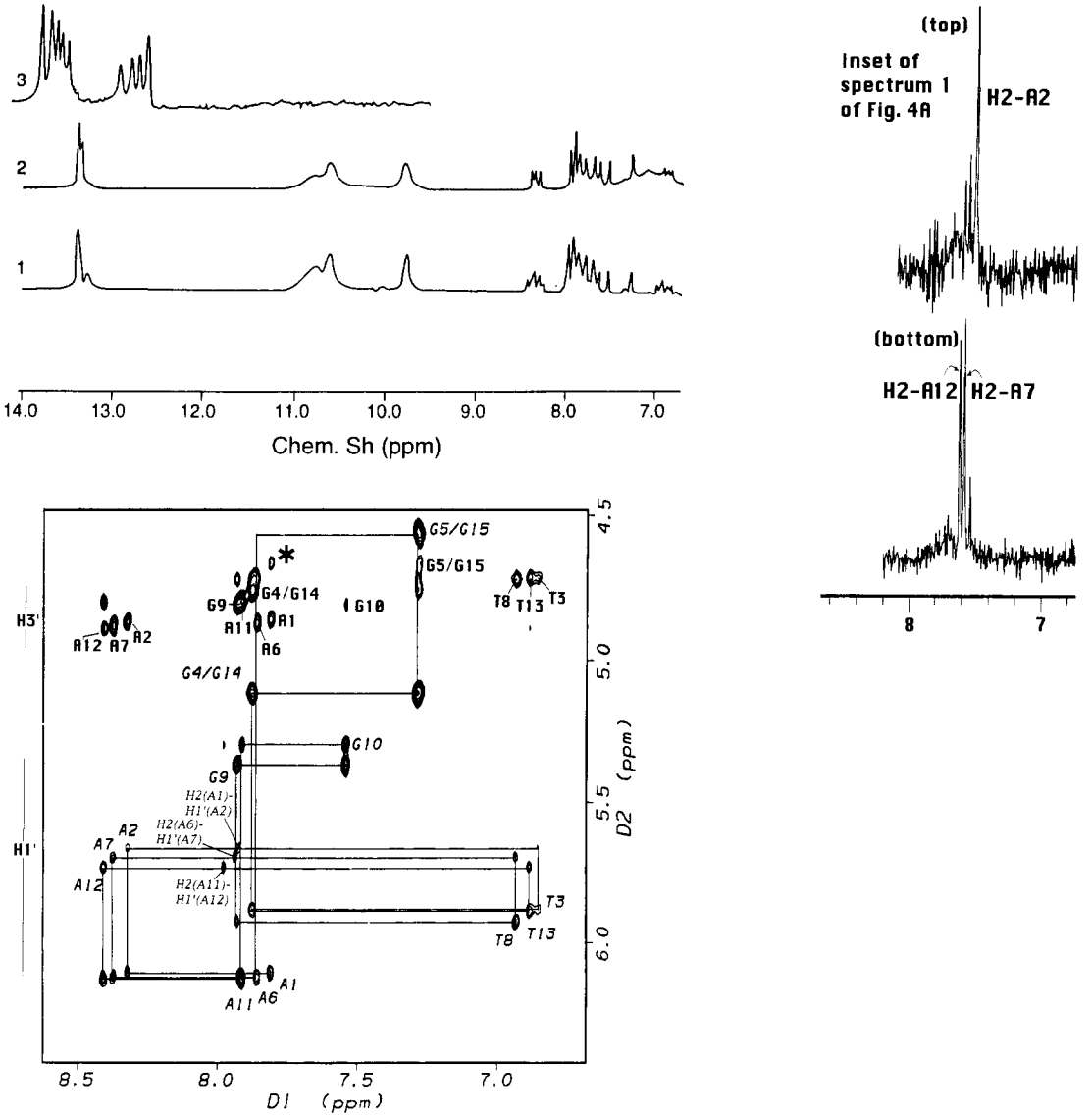

FIGURE 4: (A, top left) 1D NMR spectra of $d(AATGG)_3$ in a H_2O–D_2O (9:1) mixture under three different conditions: (spectrum 1) high DNA concentration (1.8 mM in strand) and low salt concentration (25 mM NaCl, pH 7) at 3 °C, (spectrum 2) high DNA concentration and high salt concentration (1 M NaCl, pH 7) at 3 °C, and (spectrum 3) in the presence of the complementary Watson–Crick strand $(CCATT)_3$ (1:1 molar ratio of two strands, 1 mM in duplex, 100 mM NaCl, pH 7, temperature 20 °C). 1D NOE data are shown in the inset (top right). When the smaller imino signal of A2·T8 in spectrum 1 is irradiated for 100 ms, a strong NOE is observed at H2 of A2 (the top NOE spectrum), and when the combined imino signals of T3·A7 and T13·A12 are irradiated, strong NOEs are observed at H2's of A7 and A12 (the bottom NOE spectrum). (B, bottom) 2D NOESY (τ_m = 250 ms) spectrum of $[d(AATGG)_3]_2$ in D_2O for the H1′,H3′ vs H8/H6 cross section (1.8 mM in DNA strand, 25 mM NaCl, pH 7, temperature 3 °C). The intra- and internucleotide NOEs reveal that 13 nucleotides (A1 through G4, A6 through G9, A11 through G14, and G10) exist predominantly in C2′-*endo*, *anti* conformations while either G5 or G15 shows an *anti* to *syn* conversion to facilitate the G5·G15 pairs. G10 resides in the loop segment. Internucleotide NOEs involving H1′(A2/A7/A12)–H2-(A1/A6/A11) (weak NOEs) and H1′(G4/G14)–H8(G5/G16) (strong NOEs) are indicative of special stacking patterns at the A-G and G-G steps, as discussed later. Note the high-field shift of H8,H1′(G5/G15). Full-matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms allow us to extract 150 independent interproton distances as structural constraints for molecular model building of a stem–loop motif. Intranucleotide H3′–H8/H6 NOEs are shown. Internucleotide H3′(i–1)–H8/H6(i) NOEs are also observed, but the connectivity pattern is not shown to preserve the clarity of the diagram. Note the presence of the intermolecular NOE (marked by *) between A1 and G15.

stem–loop structure of $d(AATGG)_6$ is shown on the right. Also shown are the approximate loop-folding axes drawn through the two central G's on the two loops. This clearly illustrates how the polynucleotide chain folds back onto itself to form a base-paired stem and two single-stranded loops. Stereoviews and different parts of the stem–loop structure are described in greater detail in Figure 7.

Additional Support for the Stem–Loop Structure. A stem–loop motif of $d(AATGG)_n$ differs from the corresponding non-Watson–Crick duplex (Figures 2, 3, and 5) in the following respects: (i) the overall size, e.g., as shown in Figure 5, the stem–loop motifs of $d(AATGG)_{4,6}$ are about half the length of the corresponding non-Watson–Crick duplex $d(AATGG)_{4,6}$·$d(AATGG)_{4,6}$, and (ii) the present of internal single-stranded loops in the stem–loop motif that are absent from the corresponding duplex. Therefore, the electrophoretic mobility of the stem–loop structure in a nondenaturing gel should be different from that of the duplex. Further, the internal single-stranded loops of the stem–loop structure should be susceptible to single-strand-specific *mung bean nuclease*, whereas the corresponding duplex should not be susceptible to single-strand scission.

Figure 6A shows the electrophoretic mobilities of $d(AATGG)_n$ in a nondenaturing gel. Note that $d(AATGG)_{4,6}$ migrate faster than the Watson–Crick duplexes $d(AATGG)_{4,6}$·$d(CCATT)_{4,6}$. This is consistent with the monomeric stem–loop

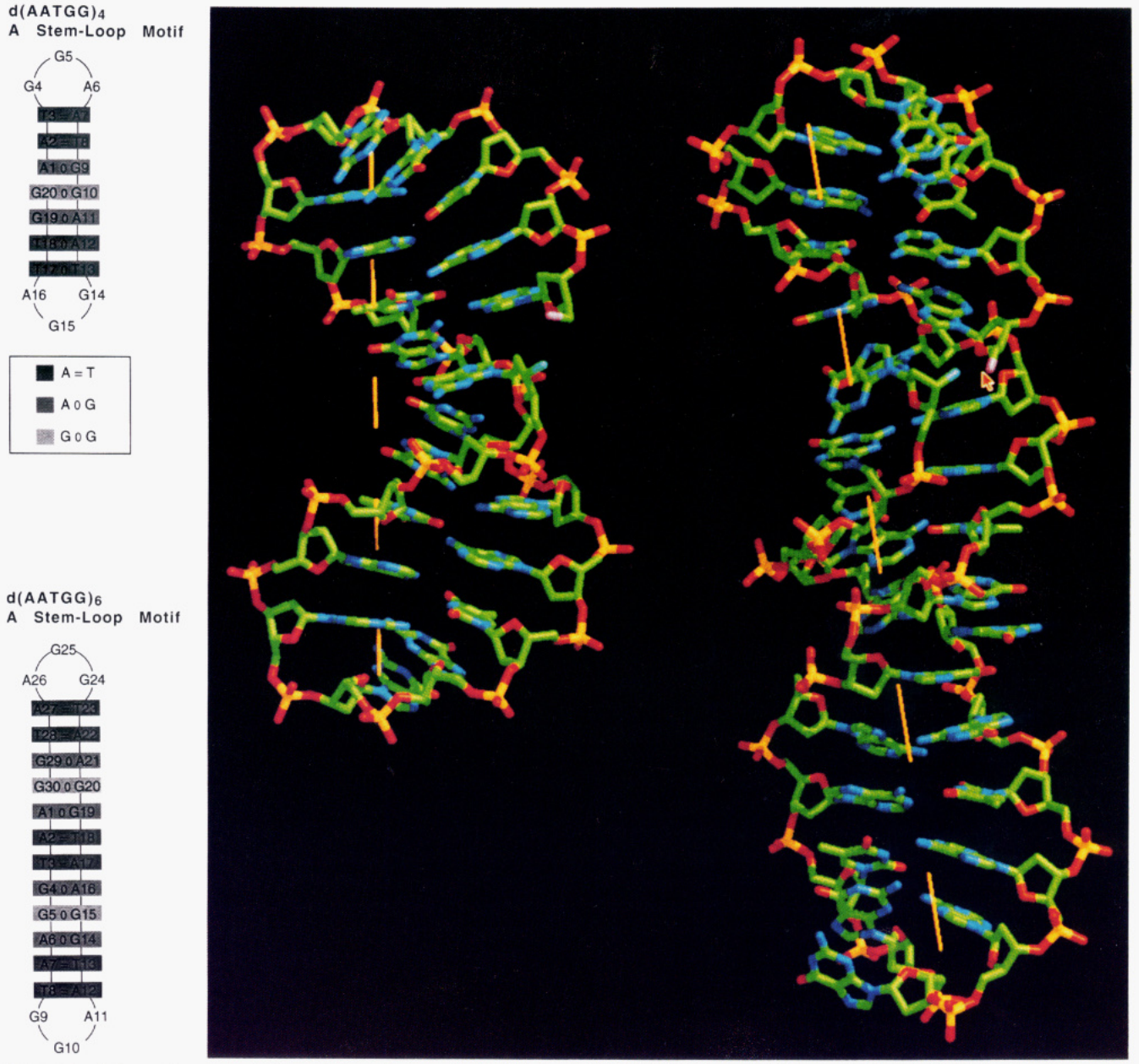

FIGURE 5: Schematic representations of stem–loop motifs for (A, top left) d(AATGG)$_4$ and (B, bottom left) d(AATGG)$_6$. (C, right) The skeletal model of the stem–loop structure of d(AATGG)$_4$ on the left and that of d(AATGG)$_6$ on the right. Only the non-hydrogen atoms are shown; C = green, N = blue, P = yellow, and O = red. The 5′-end of the structures is colored magenta while the 3′-end is colored cyan. In the stem–loop structure of d(AATGG)$_6$ the arrow is placed close to the 5′-end. The approximate axis of folding is also indicated by a dashed line. The structures represent the average of 200 sampled local minima obtained after simulated annealing subject to the NOE constraints.

structures of d(AATGG)$_{4,6}$ and not with the non-Watson–Crick duplexes d(AATGG)$_{4,6}$·d(AATGG)$_{4,6}$ because the latter is expected to migrate in a manner similar to that of the Watson–Crick duplex of the same size. Note that d(AATGG)$_6$ has the same gel mobility as the marker d(CCATT)$_4$, which is of shorter length and is a random coil under experimental conditions. Also note that d(AATGG)$_6$ migrates even faster than the Watson–Crick duplex d(AATGG)$_4$·d(CCATT)$_4$; this is consistent with the fact that the former is shorter in length than the latter. Similarly, d(AATGG)$_4$ migrates a little faster than the marker d(CCATT)$_3$ and much faster than the Watson–Crick duplex d(AATGG)$_4$·d(CCATT)$_4$. These observes support the presence of the stem–loop motifs for d(AATGG)$_{4,6}$. The gel electrophoresis patterns for d(GGATT)$_{4,6}$ are included in lanes 6 and 7 to show (without performing detailed NMR studies) that they also adopt similar stem–loop structures.

Figure 6B shows the digestion pattern of the stem–loop structures and different markers as produced by mung bean nuclease in a denaturing gel for different times of digestion. As shown in Figure 5B, two internal single-stranded loops are present in the stem–loop structure of d(AATGG)$_6$: one within nucleotides 8–12 and the other within nucleotides 23–27. Therefore, a single nick at any one of the loops is likely to produce DNA fragments of length greater than 20 but less than 25, whereas double nicks at both the loops are likely to produce DNA fragments of length greater than 10 but less than 15. Digestion of d(AATGG)$_6$ for 2.5 and 5 min produces such digestion products (lanes 2 and 3, Figure 6B) as expected for a stem–loop motif of d(AATGG)$_6$ (Figure 5B). Similarly, the single or the double nicks of d(AATGG)$_4$ (Figure 5A) are likely to produce DNA fragments of lengths greater than 10 but less than 15. Here also the nuclease treatment of d(AATGG)$_4$ for 2.5 and 5 min produces such digestion products (lanes 4 and 5, Figure 6B) as expected for a stem–loop motif of d(AATGG)$_4$ (Figure 5B).

The Structure of the Stem–Loop Motifs. Analyses of NOESY data (τ_m = 250 and 100 ms) of [d(AATGG)$_{2,3}$]$_2$ and

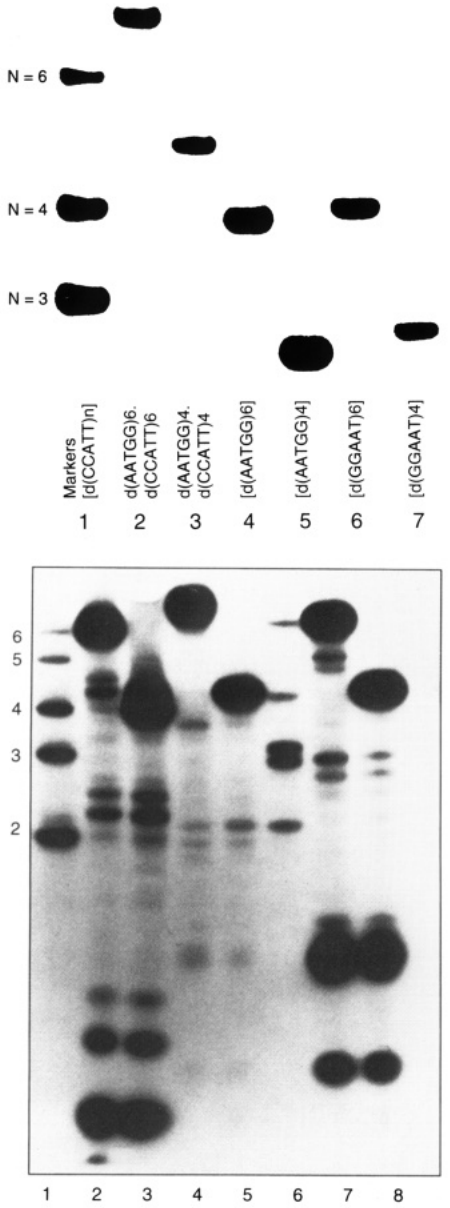

FIGURE 6: (A, top) Electrophoretic pattern of centromeric DNA repeats in nondenaturing gel. About 12 μL of the sample containing 0.1 mg of DNA was loaded in each lane. No mobility difference was found when the DNA concentration was changed. Lanes: 1, mixture of markers; 2, Watson–Crick duplex $d(AATGG)_6$·$d(CCATT)_6$; 3, Watson–Crick duplex $d(AATGG)_4$·$d(CCATT)_4$; 4, $d(AATGG)_6$; 5, $d(AATGG)_4$; 6, $d(GGAAT)_6$; 7, $d(GGAAT)_4$. (B, bottom) Digestion profiles of various centromeric repeats by mung bean nuclease (a probe for single-stranded regions in DNA). Lanes: 1, untreated $d(AATGG)_{2,3,4,5,6}$ used as markers; 2, digestion of $d(AATGG)_6$ for 2.5 min; 3, digestion of $d(AATGG)_6$ for 5 min; 4, digestion of $d(AATGG)_4$ for 2.5 min; 5, digestion of $d(AATGG)_4$ for 5 min; 6, untreated $d(GGAAT)_{2,3,4,5,6}$ and a 14-mer DNA used as markers; 7, digestion of $d(GGAAT)_6$ for 2.5 min; 8, digestion of $d(AATGG)_6$ for 5 min. Note that the duration of digestion does not alter the nature of cleavage.

$d(AATGG)_{4,6}$ with the aid of full-matrix NOESY simulations (Keepers et al., 1984; Gupta et al., 1988) resulted in a set of average interproton distances for various repeat lengths. The number of independent interproton distances was as follows: ~100 for $d(AATGG)_2$, ~150 for $d(AATGG)_3$, ~200 for $d(AATGG)_4$, and ~300 for $d(AATGG)_6$ (for illustration, chemical shift values and the NOESY data for the stem–loop structure of $[d(AATGG)_3]_2$ are included in Tables I and II of the supplementary material). Here we discuss the structures of $[d(AATGG)_2]_2$ and $d(AATGG)_4$ corresponding to the schematic models shown in Figures 2A and 5A, respectively. Using the interproton distances as structural constraints, we performed MD and energy minimization calculations (Weiner et al., 1986) in order to determine three-dimensional structures that satisfy the NMR data. A starting model of the stem–loop motif was constructed with a right-handed helical stem connecting two G-G-A loops. A left-handed helix that satisfied the observed NOEs could not be constructed. All structural parameters were taken close to B-DNA for the helical stem, except for the G·G base pair region where one of the G's in the pair adopted a *syn* conformation. For the central G·G base pair, two possibilities [i.e., G(*syn*)·G(*anti*) and G(*anti*)·G(*syn*)] were considered in our calculations. Constrained MD simulations were carried at 400 K, for 200 ps, with the purpose of sampling local and large-scale variations around a starting model (Garcia, 1992; Gupta et al., 1993). Snapshots of configurations along the MD trajectory were taken every 1 ps and subjected to constrained energy minimization (temperature quenching).

For $d(AATGG)_4$, an analysis of the rms distances among all pairs of quenched configurations followed by a hierarchical tree analysis (Gupta et al., 1993) revealed that two main families of stem–loop configurations were sampled. The main differences between both families of structure reside in the G-G-A loop region. The first family exhibits T3-G4-G5 stacking at the 5′-end and A6-A7 stacking at the 3′-end, with a two-hydrogen-bonded G4·A6 base pair similar to the base pairing found in the stem (Figure 5A). Thus, in this family two loops have the same conformation with only one unpaired base, and the G4·A6 base pair forms a part of the stem. The second family exhibits T3-G4 stacking at the 5′-end and G5-A6-A7 stacking at the 3′-end. This stacking is characteristic of DNA/RNA hairpin sequences with a Watson–Crick base-paired stem (Gupta et al., 1987; Cheong et al., 1990). This loop does not contain a G4·A6 base pair and therefore consists of three bases. In this family, only one loop adopts this conformation, whereas the other loop contains only one base for the first family of structures as described above. It is expected that each loop will independently exchange between the single- and three-base form, thus leading to four families of configurations: two symmetrical stem–loop motifs with two identical G-G-A loops (with or without a G·A base pair) and two nonsymmetrical stem–loop motifs with two different G-G-A loops (one with a G·A base pair and the other without a G·A base pair). The average energy of all minimized structures is 89.2 ± 2.3 kcal/mol, with 86.3 and 98.1 kcal/mol being the minimum and maximum energies, respectively. The average mean square distance among all structures is 1.48 Å^2. The energy differences among all structures (which account for all the conformational variants) are within 12.2 kcal/mol.

It may be noted that the energy difference between the loops with one and three unpaired bases is rather small. This becomes apparent by analyzing the relevant interaction energies in the loop segment of the stem–loop structures of $d(AATGG)_4$. For example, the interaction energy between G and A nucleosides in the loop with one unpaired base is approximately −5.0 kcal/mol of G·A, whereas the interaction energy between G and A nucleosides in the loop with three unpaired bases is approximately −3.0 kcal/mol. This difference is compensated by stacking interactions between bases in the loop and those in the stem such that the final stabilization energy difference is only 0.3 kcal/mol in favor of the conformer with loops of one unpaired base. It may also be noted that the energies in our HTMD/RTQ are only enthalpic contributions to the free energy. It is reasonable to expect that the loops with three unpaired bases will gain additional stability from the loop entropy by virtue of being inherently more

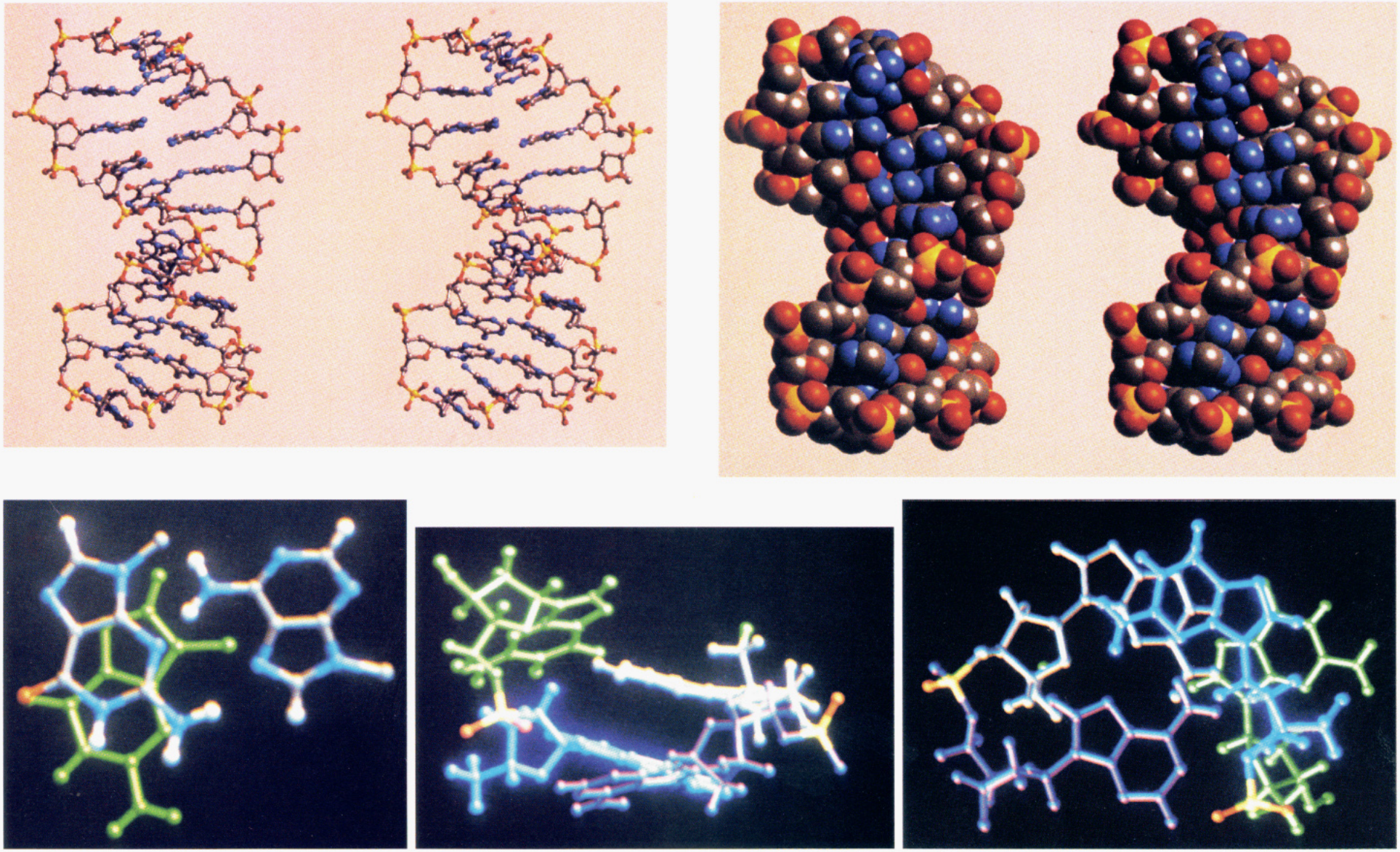

FIGURE 7: Interproton distances extracted from the NOESY data used to obtain energy-minimized models of stem–loop motifs for d(AATGG)$_4$. Only the non-hydrogen atoms are shown; C = gray, N = blue, P = yellow, and O = red. Note that in the model the stem regions contain A·T, A·G, and G·G pairs while the G-G-A constitutes the loop segments. (A, top left) Ball-and-stick model of the symmetric d(AATGG)$_4$ shown along the stem axis. (B, top right) CPK model of the symmetric d(AATGG)$_4$ shown along the stem axis. (C, bottom left) Horizontal view of the G-G-A stacking in this model. The unpaired G is shown in green. Note the presence of an A·G base pair involving the 5′G (left) and the 3′A (right). (D, bottom middle and right) Two views of stacking at the central G10–A11 step. A kink at this step leads to a localized bend of 31° for the stem. Note that the continuity of intrastrand stacking is broken at this base step. Because of unusual base pairing and stacking at the G-G-A segment of the stem, the helix describing the stem region is substantially unwound.

flexible. The activation barrier between two conformers (i.e., one with loops of three unpaired bases and the other with loops of a single unpaired base) is also small because such a transition can be locally achieved simply by moving or rotating away the 5′G and the 3′A in the loop. In view of the fact that the different loop configurations are almost equally stable and only a small barrier separates them, the 5′G and the 3′A are expected to show conformational equilibrium between the paired and unpaired states. The fact that the correspnoding G-NH signal is broad and sensitive to temperature change indicates a fast exchange of this proton within the NMR time scale, and hence this proton (and the corresponding A·G pair) is not locked only in the paired state. An RNA hairpin with an A·G pair between the 5′G and the 3′A has been reported (Heus & Pardi, 1991); in this hairpin two unpaired bases are present in the loop. Hairpins with loops containing a single nucleotide, though uncommon, are also observed in the single crystal structure of the human telomeric DNA d(GGGTTAGGG). In this structure, the 5′T and the 3′A (marked in bold) are involved in a Hoogsteen pair stacked on top of the G·G paired stem, thus leaving a single T in the loop (Alex Rich, personal communication, MIT).

Panels A and B of Figure 7 show the ball-and-stick and CPK models of the symmetrical stem–loop structure of $d(AATGG)_4$. The atoms are color coded (i.e., C = gray, N = blue, O = red, P = yellow). In this model both G-G-A loops have G·A base pairs between the 5′G and the 3′A in the loop (marked in bold). The base stacking arrangement in the loop is shown in Figure 7C where the unpaired G in the loop is colored uniformly in the green. The 5′G (on the right) and the 3′A (on the left) are colored by atoms (i.e., C = gray, N = blue, O = red, H = white).

The stem region of $d(AATGG)_4$ is a right-handed double helix unwound at the G·G pair in the stem; the unwinding also extends one base up and down the G·G pair in the stem. The stem pseudo-2-fold symmetry is broken by a G·G base pair, where one G is in a *syn* and the other is in an *anti* conformation. The stem regions separated by the G·G base pair are quite similar. A fitting of these two short helices to straight helices revealed that the two helix axes are kinked by 31°, resulting in a localized 31° bend at the central G·G base pair (Soumpasis et al., 1991). This feature is evident in Figure 7D, which shows the vertical and horizontal views of the central G10-A11 step. Note that the continuity of intrastrand stacking is broken at this base step.

As previously stated, the stem–loop motif of $[d(AATGG)_2]_2$ (Figure 2B) is similar to the $d(AATGG)_4$ model (Figure 5A) except for the lack of a phosphodiester linkage between G10 and A11. We therefore included the ball-and-stick (Figure 7E) and CPK (Figure 7F) stereoviews of $[d(AATGG)_2]_2$ in the supplementary material. As shown in Figure 7E,F, two identical hairpins are annealed in the stem–loop structure of $[d(AATGG)_2]_2$. In this model an A·G pair is present in the G-G-A loop with a stacking arrangement similar to that shown in Figure 7C.

Examination of the models in Figure 7 shows that Watson–Crick A·T pairs and stacking at A-T steps, A·G pairing and stacking at T-G steps, and G-G-A loop folding determine the stability of the stem–loop motifs of $d(AATGG)_{4,6}$ produced by two loop folding of the single strand. This is in agreement with the systematic thermal stability studies (Grady et al., 1992) that showed that any change in sequence that disrupted any one of these interactions reduced the stability of the structure.

DISCUSSION

A highly conserved repeated DNA sequence, $d(AATGG)_n$, has been identified in human centromeric regions, and for the reasons already discussed (Grady et al., 1992), it may represent a functional component of the human centromere. The high degree of sequence conservation exhibited by this sequence among diverse species (Grady et al., 1992), as great as or greater than that exhibited for animal telomeric DNA sequences (Meyne et al., 1989), suggests ongoing selective pressures to maintain this sequence. It is tempting to speculate that, like telomeric sequences (Zakian et al., 1989; Kang et al., 1992), an unusual DNA structure rather than an invarient nucleotide sequence may be the consensus recognition element for this sequence. As a non-self-complementary sequence, repeats of d(AATGG) exhibit unusual thermal stability, implying an intra- or interchain conformation with extensive base pairing (Grady et al., 1992).

Here we show by multi-dimensional NMR spectroscopy and nuclease digestion studies that repeats of d(AATGG) form a stable stem–loop structure. Formation of such a structure is expected to cost little free energy because the Watson–Crick complementary sequence d(CCATT) could easily fold back and form a triple helix with an enhanced stability at acidic pH to facilitate C^+·G·C triplets (Durland et al., 1991). Structural changes in the chromosome during the processes of chromosome condensation, particularly changes in DNA supercoiling, could induce a transition from a B-DNA duplex to the stem–loop or triple helix conformation, providing DNA binding sites for the proteins associated with the kinetochore (Pluta et al., 1990). The possibility of such a structural intermediate can be tested in two ways: (i) by using NMR spectroscopy to detect and determine the structure of a triple helix in isolated $d(AATGG)_n$·$d(CCATT)_n$ fragments at different pHs and (ii) by determining the ability of a $d(AATGG)_n$·$d(CCATT)_n$ insert in a plasmid DNA to induce negative superhelicity as a function of pH. A third approach is to search for proteins that recognize the stem–loop structures of the $d(AATGG)_n$ sequence under ionic conditions that stabilize them.

ACKNOWLEDGMENT

We thank Drs. Neville Kallenbach, Luis Marky, P. Reitemeier, C.-S. Tung, and C. Burks for their comments on the manuscript. The majority of the NMR work was done at the NMR Facility at the University of California, Davis, using the GE 500-MHz spectrometer (funded by NSF Grant DIR-88-04739 and USPHS Grant RR04795). Some of the NMR work was also done at Iowa State University, Ames, IA. The help and hospitality of Dr. A. Kintanar at the NMR facility of Iowa State University are greatly appreciated. We thank the Advanced Computing Laboratory for providing their facilities.

SUPPLEMENTARY MATERIAL AVAILABLE

Four figures showing the stem–loop structure of $[d(AATGG)_2]_2$ and 1D NMR spectra and 2D NOESY spectra of $d(AATGG)_4$ and $d(AATGG)_6$ and three tables giving chemical shift values of the stem–loop structures of $[d(AATGG)_2]_2$ and $[d(AATGG)_3]_2$ and NOESY data for $[d(AATGG)_3]_2$ (15 pages). Ordering information is given on any current masthead page.

REFERENCES

Blommers, M. J. J., Walters, J. A. L. I., Hassnoot, C. A. G., Aelen, J. M. A., van der Marel, G. A., van Boom, J. H., &

Hilbers, C. W. (1989) Effects of base sequence on the loop folding in DNA hairpins, *Biochemistry 28*, 7491–7498.

Cheong, C., Varini, G., & Tinoco, I., Jr. (1990) Solution structure of an unusually stable RNA hairpin, 5′GGAC(UUCG)GUCC, *Nature 346*, 680–682.

Cognet, J. A. H., Gabarro-Arpa, J., Bret, M. L., van der Marel, G. A., van Boom, J. H., & Fazakerley, G. V. (1991) Solution conformation of an oligonucleotide containing G·G mismatches determined nuclear magnetic resonance and molecular mechanics, *Nucleic Acids Res. 19*, 6771–6779.

Durland, R. H., Kessler, D. J., Gunnel, S., Duvic, M., Pettitt, A., & Hogan, M. E. (1991) Binding of triple helix forming oligonucleotides to sites in gene promoters, *Biochemistry 30*, 9246–9255.

Garcia, A. E. (1992) Large-amplitude nonlinear motions in proteins, *Phys. Rev. Lett. 68*, 2696–2699.

Garcia, A. E., & Soumpasis, D. M. (1989) Harmonic vibrations and thermodynamic stability of a DNA oligomer in monovalent salt solution, *Proc. Natl. Acad. Sci. U.S.A. 86*, 3160–3164.

Garcia, A. E., Gupta, G., Soumpasis, D. M., & Tung, C. S. (1990) Energetics of the hairpin to mismatched duplex transition of d(GCCGCAGC) on NaCl solution, *J. Biomol. Struct. Dyn. 8*, 173–186.

Grady, D. I., Ratliff, R. L., Robinson, D. L., McCanlies, E. C., Meyne, J., & Moyzis, R. K. (1992) Highly conserved repetitive DNA sequences are present at human centromeres, *Proc. Natl. Acad. Sci. U.S.A. 89*, 1695–1699.

Gupta, G., Sarma, M. H., Sarma, R. H., Bald, R., Engelke, U., Oei, S. L., Gessner, R., & Erdmann, V. A. (1987) DNA hairpin structures in solution: 500-MHz two-dimensional ^{1}H NMR studies on d(CGCCGCAGC) and d(CGCCGTAGC), *Biochemistry 26*, 7715–7723.

Gupta, G., Sarma, M. H., & Sarma, R. H. (1988) On the question of DNA bending: two-dimensional NMR studies on d(GTTTAAAAC)$_2$ in solution, *Biochemistry 26*, 7909–7919.

Gupta, G., Garcia, A. E., & Hiriyanna, K. T. (1993) Sampling of the conformations of the d(CGCTGCGGC) hairpin in solution by two-dimensional nuclear magnetic resonance and theoretical methods, *Biochemistry 32*, 948–960.

Heus, H. A., & Pardi, A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops, *Nature 253*, 191–194.

Kan, L.-S., Chandrasegaran, S., Pulford, S. M., & Miller, P. S. (1983) Detection of a guanine·adenine base pair in a decadeoxyribonucleotide by proton magnetic resonance spectroscopy, *Proc. Natl. Acad. Sci. U.S.A. 80*, 4263–4265.

Kang, C. H., Zhang, X., Ratliff, R., Moyzis, R., & Rich, A. (1992) Crystal structure of four-stranded *Oxytricha* telomeric DNA, *Nature 356*, 126.

Keepers, J. W., & James, T. L. (1984) A theoretical study of distance determinations from NMR. Two-dimensional nuclear overhauser effect spectra, *J. Magn. Reson. 57*, 404–426.

Kennard, O. (1988) Structural studies of base pair mismatches and their relevance to theories of mismatch formation and repair, in *Structure and Expression: DNA and its Drug Complexes* (Olson, W. K., Sarma, R. H., Sarma, M., & Sundaralingam, M. S., Eds.) Vol. 2, pp 1–26, Adenine Press, Schenectady, NY.

Li, Y., Zon, G., & 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.

Meyne, J., Ratliff, R. L., & Moyzis, R. K. (1989) Conservation of the human telomere sequence (TTAGGG)$_n$ among vertebrates, *Proc. Natl. Acad. Sci. U.S.A. 86*, 7049–7053.

Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L., & Wu, J.-R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)$_n$, present at the telomeres of human chromosomes, *Proc. Natl. Acad. Sci. U.S.A. 85*, 6622–6626.

Moyzis, R. K., Torney, D. C., Meyne, J., Buckingham, J. M., Wu, J.-R., Burks, C., Sirotkin, K. M., & Goad, W. B. (1989) The Distribution of interspersed repetitive DNA sequences in the human genome, *Genomics 4*, 273–289.

Pluta, A. F., Cooke, C. A., & Earnshaw, W. C. (1990) Structure of the human centomere at metaphase, *Trends Biochem. Sci. 15*, 181–185.

Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M., & Dickerson, R. E. (1988) Structural studies of base pair mismatches and their relevance to theories of mismatch formation and repair, in *Structure and Expression: DNA and its Drug Complexes* (Olson, W. K., Sarma, R. H., Sarma, M., & Sundaralingam, M. S., Eds.) Vol. 2, pp 27–48, Adenine Press, Schenectady, NY.

Rich, A., Nordhem, A., & Wang, A. H.-J. (1984) The Chemistry and Biology of Left-Handed Z-DNA, *Annu. Rev. Biochem. 53*, 791–846.

Sklenar, V., & Bax, A. (1987) Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra, *J. Magn. Reson. 74*, 469–479.

Smith, F. W., & Feigon, J. (1992) Quadruplex structure of *Oxytricha* telomeric DNA oligonucleotides, *Nature 356*, 164–168.

Soumpasis, D. M., Tung, C.-S., & Garcia, A. E. (1991) Rigorous description of DNA structures. II. On the computation of best axes, planes, and helices from atomic coordinates, *J. Biomol. Struct. Dyn. 8*, 867–888.

Stillinger, F. H., & Weber, T. A. (1983) Dynamics of structural transitions in liquids, *Phys. Rev. A 28*, 2408–2416.

Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986) An all atom force for simulations of proteins and nucleic acids, *J. Comput. Chem. 7*, 230–245.

Williamson, J. R., & Boxer, S. G. (1989) Multinuclear NMR studies of DNA hairpins. 1. Structure and dynamics of d(CGCGTTGTTCGCG), *Biochemistry 28*, 2831–2836.

Zakian, V. A. (1989) Structure and function of telomeres, *Annu. Rev. Genet. 23*, 579–604.