LA-UR-93-3073

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

TITLE: STEM-LOOP STRUCTURES OF THE REPETITIVE DNA SEQUENCES
LOCATED AT HUMAN CENTROMERES

AUTHOR(S): GOUTAM GUPTA
ANGEL E. GARCIA
PAOLO CATASTI
LIN HONG
PETER YAU
E.MORTON BRADBURY
ROBERT RATLIFF
ROBERT K. MOYZIS

SUBMITTED TO:

PROCEEDINGS OF BIOMOLECULAR STEREODYNAMICS

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Stem-Loop Structures of the Repetitive DNA Sequences Located at Human Centromeres

Goutam Gupta+ and Angel E. Garcia+
+Theoretical Biology and Biophysics Group
Division T-10, M/S K710
Los Alamos National Laboratory
Los Alamos, NM 87545

Paolo Catasti*, Lin Hong*, Peter Yau*, and E. Morton Bradbury^*
^Life Sciences Division, M/S K881,
P. O. Box 1663,
Los Alamos National Laboratory
Los Alamos, NM 87545
*Department of Biological Chemistry,
School of Medicine,
University of California at Davis,
Davis, CA 95616

Robert Ratliff# and Robert K. Moyzis#
#Center for Human Genome Studies,
Los Alamos National Laboratory,
Los Alamos, NM 87545

The Corresponding Author: GOUTAM GUPTA,
T-10, MS K710
Theoretical Biology and Biophysics,
Los Alamos National Laboratory,
NM 87545
Tel (505)-665-6463
Fax (505)-665-3493
E-Mail gxg@temin.lanl.gov

This work is supported by the U. S. Department of Energy

Abstract

The presence of the highly conserved repetitive DNA sequences in the human centromeres argues for a special role of these sequences in their biological functions - most likely achieved by the formation of unusual structures. This prompted us to carry out quantitative one- and two-dimensional nuclear magnetic resonance (1D/2D NMR) spectroscopy to determine the structural properties of the human centromeric repeats, d(AATGG)$_n$.d(CCATT)$_n$.

The studies on centromeric DNAs reveal that the complementary sequence, d(AATGG)$_n$.d(CCATT)$_n$, adopts the usual Watson-Crick B-DNA duplex and the pyrimidine-rich d(CCATT)$_n$ strand is essentially a random coil. However, the *purine-rich* d(AATGG)$_n$ strand is shown to adopt unusual stem-loop structures 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 mismatch 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. Thermal melting studies show that the DNA repeats, d(AATGG)$_n$, are as stable as the correponding Watson-Crick duplex d(AATGG)$_n$.d(CCATT)$_n$. Therefore, the sequence d(AATGG)$_n$ can, indeed, nucleate a stem-loop structure at little free-energy cost and if, during mitosis, they are located on the chromosome surface they can provide specific recognition sites for kinetochore function.

The stem-loop structures of d(AATGG)$_{4,6}$ are derived subject to the gross structural constraints obtained from the gel electrophoresis data and H-bonding and inter-proton distance constraints obtained from the NMR data.

A methodology combining high temperature molecular dynamics and rapid temperature quenching (HTMD/RTQ), subject to the experimental constraints, is used to sample the local energy minima of a given stem-loop structure. In view of the fact that the stem-loop structures contain non-Watson-Crick elements (*e.g.*, the mismatch A•G and G•G pairs in the stem and the single-stranded G-G-A loop), the analyses of the energy-minimized structures prove quite informative regarding the role of different non-Watson-Crick elements. The analyses reveal the intrinsic flexibility of the G-G-A loop, the spatial relation of the A•G and G•G mismatches with respect to their Watson-Crick A•T neighbor, and the relative strengths of various nearest-neighbor interactions in the overall stability of the structure. Finally, the most noteworthy aspect of our studies is the observation that the repetitive DNA sequences can self-associate through chain reversals at two sites of the single strand such that the two ends of the single strand are brought close together. In the case of the centromere, while the presence of the Watson-Crick A•T pair and the mismatch A•G pair in the stem is crucial for stability, the G•G pair can be replaced by any other mismatch without altering the stability. Interestingly, natural mutations in the centromeric repeat occur more frequently only at one of the Gs expected to form the G•G pair in the stem-loop structure.

Introduction

In the past decade, we have witnessed the experimental demonstrations of novel DNA structures for various functional elements of the geonomic DNA. These unusual DNA structures show drastic departures from the classical right-handed A- or B-DNA double helix [1,2]. These structures fall into four catagories: (i) DNA duplexes that retain the same Watson-Crick pairing but are different in terms of the handedness- *e.g.*, the left-handed Z-DNA [3,4], and the unwound B-Z junctions [5,6]; (ii) triple helices that have a third chain in the major-groove of the Watson-Crick duplex forming Hoogsteen pairing with the bases in one of the Watson-Crick paired duplex strands [7,8]; (iii) tetraplex structures with ordered layers of circularly non-Watson-Crick H-bonded G-quartets [9-12]; and (iv) hairpin structures with stems containing Watson-Crick pairs [13-15], mismatches [16,17], or ordered layers of G-quartets [18-21]. Although a lot of research work is reported in literature describing the structure and thermodynamic stabilities of these non-Watson-Crick or unusual DNA structures, it is only quite recently that the biological relevance of these structures are beginning to be understood.

We at LANL have also been actively engaged in understanding the structure and stability of unusual DNA motifs [22,23]. Recently the discovery of the highly conserved repetitive DNA sequences at the human telomere [24] and centromere [25] have added a new dimension to our research on unusual DNA structures. Figure 1 shows different functional loci on the human chromosome. As indicated in Figure 1, although the DNA sequences at the centromere or telomere do not code for any proteins, they are highly conserved and constitute important functional loci during mitosis (for centromere) and during protection against the end-damage and for the end-replication (for telomere) [26]. The highly conserved nature of the centromere and telomere repeats argues for special three-dimensional structures of these DNA motifs that are finally responsible for their functions. Because of the repetitive nature of the sequence at the telomere and centromere, these DNA repeats are capable of forming self-associated structure; and when they do, they invariably contain elements of non-Watson-Crick DNA structures. The most important questions about these repetitive DNA sequences are then: can the sequence self associate? if so, what are the structure and stability of the self-associated form? and how do the structure and stability depend upon the repeat length and the solution conditions? The self-associated hairpin and tetraplex forms of telomeric DNA repeats [9-12, 18-21] are reported in the literature. In this article, we describe the self-associated stem-loop structures of the purine-rich strand of the human centromere DNA, $d(AATGG)_n.d(CCATT)_n$, for repeat lengths n=2-6. The structural parameters are derived from the NMR, gel electrophoresis, and the nuclease digestion data. Molecular modeling studies are performed in conjunction with these experimental data by using a high temperature molecular dynamics simulation (HTMD) and rapid temperature quenching (RTQ) [22].

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 due to Sklenar and Bax [27]. 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, 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 X 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 intra-nucleotide 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 inter-proton distances (i.e., average values and associated dispersions) was extracted for various pair-wise interactions by performing Full-Matrix NOESY Simulation and associated R-Factor tests by comparing the corresponding calculated and the observed NOESY intensities (methodology is described in Ref. 28). The sugar puckers of different residues were estimated by monitoring the corresponding J-coupling parameters of H1'---H2', H1'---H2", H2'---H3', H2"---H3', etc., interactions in the corresponding phase-sensitive COSY cross-sections. *Third*, these inter-proton distances were used as structural constraints for constant high temperature (400K) molecular dynamics (MD) simulations after 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*, snapshots were extracted at regular intervals 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*, all 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 [methodology described in Ref. 22]. *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 [28]. Steps *three to five* are collectively referred to as "high temperature MD simulation followed by rapid temperature quenching, HTMD/RTQ" [22]. During HTMD/RTQ all NOE-derived distance constraints were made to satisfy by using appropriate constraint energy functions. Therefore, all the final 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 [22]. 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. [29] in AMBER 3.0/4.0. All calculations were done *in vacuum* with a constant dielectric coefficient of 78.4 [30,31] and without any non-bonding cut-off. High temperature (400K) 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).

Gel Electrophoresis and Nuclease Digestion

Electrophoretic studies in a non-denaturing gel and single-strand specific nuclease digestion studies were performed to independently verify the stem-loop structures of the centromeric repeats [32].

Electrophoretic patterns of $d(AATGG)_{2,3,4,6}$ were monitored in a non-denaturing gel; 12% polyacrylamide, 0.5 X TBE buffer. Oligonucleotides were labelled 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 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 were loaded in each lane. No mobility difference was found by changing the DNA concentration.

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 labelled by ^{32}P before digestion. Reaction conditions: 30 mM sodium acetate, 50 mM NaCl, 15 μM $ZnCl_2$, pH 5.0. Reactions of 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 minutes. The reaction was stopped at different times by adding 50 mM EDTA. Denaturation DNA gels were used: 15% polyacrylamide (20:1), 7 M urea in Tris buffer.

Results

The technical details of the experimental studies on the human centromeric DNA are already published elsewhere [32]. As stated in the abstract, the Watson-Crick duplex, $d(AATGG)_n.d(CCATT)_n$, adopts the usual B-DNA while the $d(CCATT)_n$ strand is essentially a random coil. However, the $d(AATGG)_n$ strand, is shown to adopt an unusual stem-loop motif for repeat lengths, n=2,3,4, and 6 [32]. Here we only discuss the essential NMR evidences that lead to the derivation of the final stem-loop structures of $d(AATGG)_{4,6}$. The NMR evidences include the following: (i) the signatures of 1D NOE due to the exchangeable imino (NH) and amino (NH2) protons that lead to identification of the Watson-Crick A•T pairs and the mismatch A•G and G•G pairs; (ii) the intra-nucleotide NOEs involving the base and sugar protons and the J-coupling pattern of the sugar protons revealing *(C2'-endo,anti)* conformations of all the nucleitides in the structure excepting one of the Gs in the G•G pair (that adopts C2'-endo, syn conformation; and (iii) inter-nucleotide NOE pattern indicating the presence of stem-loop structures.

(i) The identification of the base pairing pattern

Figure 2 shows the exchangeable NH and NH2 and base proton H8/H6/H2 signals for (A) $d(AATGG)_4$ and (B) $d(AATGG)_6$, respectively. The NH signals above correspond to the Watson-Crick A•T pairs because as shown in Figure 3, they show strong NOE at the in-plane H2 of A of the same A•T pair. It turns out that all the Ts in the centromeric repeats are involved in Watson-Crick A•T pairs, *e.g.*, there are four A•T pairs in $d(AATGG)_4$ and there are six A•T pairs in $d(AATGG)_6$. The participation of one set of As in the Watson-Crick A•T pairs leaves the other set available for A•G pairs. The NH signal of G (or G-NH) near 10.7 ppm in Figures 2A and 2B correpond to the A•G pairs [32]. This signal containing two or more NH protons shows NOE at the in-plane H8s of the A•G pairs as shown in Figure 3. It may be pointed out that the G-NH signal in the A•G pair appears at a field of 10.7 ppm. This chemical shift value is much higher than those reported for A*(anti)*•G*(anti)* and A*(syn)*•G*(anti)* pairs in which the G-NH proton participates in H-bonding. The low field shifted G-NH signal in the two latter cases is the consequence of the ring current deshielding induced by the proximity of another aromatic ring to the G-NH signal. Therefore, the high field shifted G-NH signal gives a qualititative indication that the G-NH does not take part in H-bonding of the A•G pair [33]. In addition, the absence of a strong NOE from G-NH to the in-plane A-H2 of the A•G pair rules out the possibility of the A(anti)•G(anti) pair involving G-NH in the H-bonding. Similarly, the absence of syn conformation for any A in the sequence discards the possibility of A(syn)•G(anti) with G-NH participating in the H-bonding. These observations suggest that G-NH2 protons are involved in the H-bonding of the A•G pair. The nature of NOE favors the A•G pairing scheme shown in Figure 3. Note that, in the A•G pair, the A-H2 of the pair occupies the edge; whereas, in the Watson-Crick A•T pair, the A-H2 of the pair resides at the core. Thus, as shown later, the set of A-H2s belonging to the A•G pairs are deshielded and occur at a lower field; whereas, A-H2s belonging to the A•T pairs are deshielded and occur at a higher field. The G-NH signal at 9.8 ppm shows in-plane NOE of the type G-NH->G-NH2->G-H8 as shown in Figure 3. This is consistent with the G(anti)•G(syn) pairs. The broad G-NH signals near 10.9 ppm in Figure 2A and 2B show no specific NOEs and are extremely sensitive to temperature changes. These properties are characteristic of the NH signals belonging to a single- stranded loop. There are two ways in which one or two Gs can reside in a single-stranded loop of a structure in which A•T, A•G, and G•G pairs can still be accommodated. They are: (i) a hairpin with a G-G-A loop at the center or (ii) a stem-loop structure with two G-G-A loops connected by a base paired stem. For the reasons discussed below stem-loop structures of d(AATGG)$_{4,6}$ are consistent with NMR, gel electrophoresis, and nuclease digestion data.

(ii) Experimental evidence of the stem-loop structures of d(AATGG)$_{4,6}$

Figure 4 displays the schematic representation of the stem-loop structures of (A) d(AATGG)$_4$ and (B) d(AATGG)$_6$, respectively. Note that different base pairs are shown in different colors and the presence of two (G-G-A) loops that connect the ends of the stem. As expected for the nucleotides in the duplex stem [28], a continuous NOE connectivity was observed for H2"(i-1)...H8/H6(i) and H1'(i-1)...H8/H6(i) interactions. However, H2"(G_{i-1})...H8(A_i) connectivity was either very weak or absent for the two Gs for d(AATGG)$_4$ and d(AATGG)$_6$. Such an NOE pattern is expected for the central G in the G-G-A loop; and the presence of two such Gs indicate the presence of two G-G-A loops in the d(AATGG)$_4$ and d(AATGG)$_6$ structures. Further, the proof that d(AATGG)$_4$ and d(AATGG)$_6$ adopt stem-loop structures came from the NMR data on [d(AATGG)$_2$]$_2$ and [d(AATGG)$_3$]$_2$ [31]. Both [d(AATGG)$_2$]$_2$ and [d(AATGG)$_3$]$_2$ form stem-loop structures in which two identical hairpins are end-stacked. Therefore, the stem-loop structure of [d(AATGG)$_2$]$_2$ is similar to that of d(AATGG)$_4$ except for a missing covalent link (Figure 4A). Therefore, as shown in Figure 5A, the NOE pattern of [d(AATGG)$_2$]$_2$ is very similar to that of d(AATGG)$_4$ except for the fact that in the latter case there are twice as many residues because the the covalent link between G10 and A11 in d(AATGG)$_4$ removes the two-fold symmetry present in [d(AATGG)$_2$]$_2$. Similarly, as shown in Figure 5B, the NOE pattern of [d(AATGG)$_3$]$_2$ is very similar to that of d(AATGG)$_6$ except for the fact that in the latter case there are twice as many residues because the covalent link between G15 and A16 in d(AATGG)$_6$ removes the two-fold symmetry present in [d(AATGG)$_3$]$_2$. The presence of two chemically non-equivalent G-G-A loops in d(AATGG)$_6$ and d(AATGG)$_6$ are evident in the NOESY diagrams of Figures 5A and 5B, respectively.

The formation of the stem-loop structures of d(AATGG)$_6$ and d(AATGG)$_6$ [Figures 4A and 4B] are also verified by analyzing the electrophoretic pattern of these DNA oligomers in a non-denaturing gel and by examining the *mung bean nuclease* (a single-strand specific enzyme) digestion pattern in a denaturing gel.

Figure 6A shows the electrophoretic mobilities of [d(AATGG)$_n$] in a non-danutaring gel. Note that d(AATGG)$_{4,6}$ migrate faster than the Watson-Crick duplexes d(GGAAT)$_{4,6}$.d(ATTCC)$_{4,6}$. This is consistent with the monomeric stem-loop structures of d(AATGG)$_{4,6}$ (Figures 4A and 4B] and not with a non-Watson-Crick duplex d(AATGG)$_{4,6}$.d(AATGG)$_{4,6}$ because the latter is expected to migrate in a similar manner as the Wason-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(AATGG)$_4$; 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(AATGG)$_4$. These observations support the presence of the stem-loop motifs for d(AATGG)$_{4,6}$.

Figure 6B shows the digestion pattern of the stem-loop structures and different markers as produced by *mung bean nuclease* in denaturing gels for different times of digestion. As shown in Figure 4B, 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 minutes produces such digestion products (Lanes 2 and 3 Figure 6B) as expected for a stem-loop motif of $d(AATGG)_6$ (Figure 4B). Similarly the single or the double nicks of $d(AATGG)_4$ (Figure 4A) 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 minutes produces such digestion products (Lanes 4 and 5 Figure 6B) as expected for a stem-loop motif of $d(AATGG)_4$ (Figure 4A).

The Structure of the Stem-Loop Motifs

Analyses of NOESY data (τ_m = 250 and 100 ms) of $d(AATGG)_{4,6}$ with the aid of Full-Matrix NOESY simulations [28,32] resulted in a set of average inter-proton distances for various repeat lengths. The number of independent inter-proton distances was as follows: ~200 for $d(AATGG)_4$, and ~300 for $d(AATGG)_6$. Using the inter-proton distances as structural constraints, we performed MD and energy minimization calculations [22] 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 Gs in the pair adopted a *syn* conformation. For the central G•G base pair, the two possibilities [i.e., G(*syn*)•G(*anti*) and G(*anti*)•G(*syn*)] were considered in our calculations.

$d(AATGG)_4$

For $d(AATGG)_4$, an analysis of the MSDs among all pairs of quenched configurations followed by a hierarchical tree analysis [32] revealed that two main families of stem-loop configurations were sampled. The main differences between both families of structures reside in the G-G-A loop region. The first family exhibits T3-G4-G5 stacking at the 5'-end, 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 (Figures 3 and 4A). Thus, in this family both 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 hairpin sequences with Watson-Crick base paired stem [13-15, 22]. 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 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 non-symmetrical 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/Mole with 86.3 and 98.1 KCal/Mole being the minimum and maximum energies, respectively. The average MSD among all structures is 1.48 Å^2. The energy differences among all structures (which accounts for all the conformational variants) are within 12.2 KCal/Mole. 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/Mole of G•A; whereas, the interaction energy between G and A nucleosides in the loop with three unpaired bases is approximately -3.0 KCal/Mole. This difference is ultimately 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/Mole in favor of the conformer with loops of one unpaired base. It may also be noted that the energies in our HTMD/RTQ calculations are only estimates of enthalpic contributions to the free energy. It is reasonable to expect that the loops with three unpaired bases will gain extra stability from the loop-entropy by virtue of being inherently more flexible. The activation barrier between two conformers, (i.e., one with loops of three unpaired bases and the other with loops of 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 corresponding 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. 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 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 (personal communication, Alex Rich, MIT).

Figure 7A shows the skeleton model of the symmetrical stem-loop structure of d(AATGG)$_4$. The atoms are color coded (i.e., C=green, 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 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 two-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 degrees resulting in a localized 31-degree bend at the central G•G base pair [34].

d(AATGG)$_6$

The stem-loop structure of d(AATGG)$_6$ is inherently more complex than the structure of d(AATGG)$_4$. We, therefore, discuss in detail the modeling studies subject to the NMR, gel electrophoresis, and nuclease digestion data.

2D NMR, gel electrophoresis, and nuclease digestion experiments suggest a stem-loop structure for d(AATGG)$_6$ as shown in Figure 4B. The structure contains Watson-Crick A-T base pairs, G•A and G^{syn}•G^{anti} mismatches, and two G-G-A loops. The numbering scheme of the bases relevant for the present discussion is shown in Figure 8 (Inset). This secondary structure allows the formation of a nicked dumbbell with one nick at positions A1, A11, A16, or A26. Of these, pairs of structures nicked at A1 or A16 and A11 or A26 are topologically equivalent. Here we choose to model a structure shown in Figure 8 (Inset). Another degeneracy that should be considered is the positioning of the $G^{syn} \cdot G^{anti}$ base pairs at positions G10, G15, G25, and G30. There are four possible arrangements of the G•G base pairs:

1) $G30^{syn} \cdot G10^{anti}$ and $G25^{syn} \cdot G15^{anti}$;
2) $G10^{syn} \cdot G30^{anti}$ and $G25^{syn} \cdot G15^{anti}$;
3) $G30^{syn} \cdot G10^{anti}$ and $G15^{syn} \cdot G25^{anti}$;
4) $G10^{syn} \cdot G30^{anti}$ and $G15^{syn} \cdot G25^{anti}$.

If we neglect the effect of the nicked sequence, arrangements 1 and 3 and arrangements 2 and 4 give identical structures. As indicated in Figure 8 (Inset), we choose arrangement 1, which contains two different stem-loops, 5'- G^{syn} ...loop ... G^{anti}-3' and 5' -G^{anti} ...loop ... G^{syn} - 3'.

Initial structures were generated by starting from ideal right handed double helices for the stem region of the molecule and the G-G-A loops obtained for the previously studied centromeres of repeat lengths 2 and 4, respectively. Modifications in the sugar backbone and base pairings were made in order to satisfy the $G^{syn} \cdot G^{anti}$ and the G•A base pairings described above. NOE constraints were imposed by using an energy function that is flat between the lower and upper bounds for the NOE distances and harmonic for distances smaller than the lower bounds or larger than the upper bounds. Similar constraints were used for constraining hydrogen bonding distances in the stem region (i.e., excluding the GGA loops). Harmonic force constants of 10.0 Kcal/mol-A^2 were used for NOE constraints, and 35.0 Kcal/mol-A^2 were used for hydrogen bonding constraints.

The initial structures were energy minimized (rms derivative ~0.01 KCal/Mole/Å) without any NOE constraints. NOE constraints were included in the energy functions and were retained for all subsequent calculations. The NOE-constrained energy-minimized model was used as a starting structure for a 400K molecular dynamics simulation. The temperature of the system was slowly increased to 400K during a 10 ps MD simulation. Afterwards, a 100 ps simulation was carried out. Configurations were saved for every 2 ps along the 100 ps trajectory. Each of the 50 configurations collected were subjected to a 2500-step conjugate gradient energy minimization. This sequence of high temperature MD followed by energy minimization (quenching) allowed the mapping of local minima along the trajectory [22]. The resulting 50 local minima were further studied in terms of energetics and mean square distances (MSD) among local minima. Given that each structure was strictly different from every other structure, a clustering algorithm based on the MSD was used to obtain a set of structures that best represented clusters of configurations.

The resulting structures vary in energy from 186 to 209 Kcal/mol, with average energy of 195 ± 5 Kcal/mol. The largest MSD between any two structures is 5.7 Å^2, the smallest 1.1 Å^2 with average of 1.9 Å^2. Figure 7B shows the average stem-loop structure of d(AATGG)$_6$ that is consistent with the experimental data. The O5'-atom of the 5'-A1 is colored magenta while the O3'-atom of the 3'-G30 is colored cyan. The folding pattern of the sugar-phosphate chain is readily traced from this view. The yellow dashed line shows the approximate chain-axis of folding. Figure 8 shows the two structures in the cluster that differ the most in terms of MSD. The bottom molecule is bent relative to the top structure. This bend is due to the flexible hinge at one of the G•G base pairs and include the hairpin section of the molecule formed by bases 15 to 25. Bends also occur at the other G•G base pair. The numbering scheme of the bases in the stem-loop structure and the possible source of bend or kink are schematically shown in Figure 8 (Inset). The loop bases G5 and G20 exhibit the largest fluctuations with $<x^2>$ values of 4.0 and 5.5 $Å^2$, respectively. The sequences T3-G4-G5-A6 and T18 -G19-G20-A21 show $<x^2>$ values larger than the average MSD while all the stem region show $<x^2>$ values below the MSD for all structures, indicating a rigid and well determined stem region. In the stem-loop structure of $d(AATGG)_4$, the presence of a single G-G step at the center of the stem caused a kink. However, in the stem-loop structure of $d(AATGG)_6$ two G-G steps in the central stem are located on two diagonally opposite faces of the stem [see Figure 8 (Inset)]. The net effect is the cancellation of the two kinks and straightening of the central stem region.

In view of the fact that different non-Watson-Crick structural elements are assembled in the stem-loop structure, it is necessary to examine the relative strengths of all nearest-neighbor interactions. Figure 9 shows three different contributions of nearest-neighbor interactions and also the total energy: (A) pairing energy of the bases on opposite faces of the molecule, (B) 5'-3' intra-strand stacking, (C) diagonal (or inter-strand) stacking, and (D) total interaction energy. These interaction energies are computed for the lowest energy structure in the cluster. The phosphate groups are omitted for calculating the interaction energies for nucleoside pairs. Figure 9A shows that the lowest base pairing energy is that of the G•G base pairs, followed by the Watson-Crick A•T base pairs. The most stabilizing base pairing energy corresponds to G•A base pairs, with the largest G•A pairing energies occurring for bases in the first and third positions of the loops. Given the strong stabilizing energies of these two bases in the loop, we can consider the loop to consist of only one base (no H-bonding constraints were imposed between the G and A bases in the loop). G5 and G20 show zero pairing energy since they are not paired. Figure 9B shows the 5'-3' stacking energies between neighboring bases in the sequence. The bar shown over the i-th base corresponds to an interaction between the i-th and the (i+1)-th nucleosides. Among the most relevant features, we should emphasize that the GpG stacking between the 5' nucleoside in the loop and the second base in the loop is the strongest interaction (-12.6 Kcal/Mole). In addition, there is a partial stacking between the second and the third nucleosides in the loop (-3.3 Kcal/Mole). Stem region ApA, ApT, and TpG stacking interactions average to -10.1 Kcal/Mole, with ApA being the strongest; whereas, the stem GpA and GpG interactions average to -3.3 Kcal/mol. The smallest stacking interaction occurs between G30^{syn}-A1 (1.1 Kcal/Mole) and G15^{anti}-A16 (-0.7 Kcal/Mole). The loss in interaction energy at these two steps is compensated by a diagonal stacking of A1 on G20 (-7.9 Kcal/Mole) and that of G15 on G24 (-9.6 Kcal/Mole). Note that in both cases the G^{anti} nucleoside stacks with another purine, but in the first case G20 stacks on the nucleoside at the 5' end (A1) of the pairing base (G30), while in the second case G15 stacks on the nucleoside at the 3'-end (G24) of the pairing base (G25). Strong inter-strand stacking interactions between purine bases in DNA duplexes have been previously described [33]. Diagonal stacking interaction energies (i.e., inter-strand interaction in the case of double helices) are shown in Figure 9C. Note that the highly stabilizing diagonal interactions purine-purine stacking energies for A1, G10, G15, and G24. Note that these interactions are more than twice as strong as the A2-A7, A12-A27, and A17-A22 among purines involved in regular Watson-Crick base pairings. The diagonal interaction energies of G14 are also large. This interaction results from larger than average interactions between G14 and G25 (-5.7 Kcal/Mole) and normal stacking interactions between G14 and A27 (-3.6 Kcal/Mole). The sum of all the inter-nucleosides interactions is shown in Figure 9D. The mean total interaction per nucleoside is -24.3 ± 4.5 Kcal/mol, with values ranging from -14.0 Kcal/mol for G20 to -30.6 Kcal/mol for G4. This analysis suggests that the most stabilizing interactions in this sequence result from the G•A pairing and the GpA stacking of the three loop nucleosides and in the G•A base pairings and ApA stackings in the stem.

Concluding Remarks

It is now being recognized that repetitive genomic DNA sequences have a special biological role primarily dictated by their three-dimensional structures. The human centromeric DNA repeats have a special role in the mitotic phase. Although these sequences do not code for any protein

(Figure 1), they remain highly conserved. The stem-loop structures of $d(AATGG)_4$ and $d(AATGG)_6$ (Figures 4 and 7) described in this article, clearly explain how the repetitive nature of the sequence preserves the single-strand fold for different repeat lengths. The extent of conservation of the three-dimensional structure and the associated stability could be tested by introducing naturally occurring single- or double-site mutations in the skelton of the stem-loop structures of $d(AATGG)_4$ and $d(AATGG)_6$. Experiments in this direction are in progress and will be reported elsewhere. Another important question regarding the stem-loop structures of $d(AATGG)_n$ is: how such a structure can be initiated from the parent Watson-Crick duplex, $d(AATGG)_n.d(CCATT)_n$? We are at present carrying out experiments to stabilize the stem-loop structure of $d(AATGG)_n$ and simultaneously stabilize the pyrimidine-rich strand, $d(CCATT)_n$.

Acknowledgment

Support for this work from the United States Department of Energy to the Human Genome Center, the Life Sciences and Theoretical Divisions of LANL, and to E. Morton Bradbury [grant number DE-FG03-88ER60673] is greatly appreciated. We thank Drs. Neville Kallenbach, Luis Marky, P. Reitemeier, and C. -S. Tung for their comments and input. The NMR work was done at the NMR Facility at U.C. Davis using the GE 500 MHz spectrometer (funded by NSF DIR-88-04739 and PHS RR04795) and at the Iowa State University, Ames, Iowa. The help and hospitality of Drs. D. Scott and A. Kintanar at the NMR facility of the Iowa State University is greatly appreciated.

References

1. Langridge, R., Wilson, H. R., Hooper, C. W., Wilkins, and Hamilton, L. D. (1960), *J. Mol. Biol.*, **2**, 19-37.

2. Fuller, W., Wilkins, M. H. F., Wilson, H. R., Hooper, C. W., and Hamilton, L. D. (1965), *J. Mol. Biol.*, **12**, 60-81.

3. Rich, A., Nordhem, A., and Wang, A. H. -J., (1984), *Ann. Rev. Biochem.*, **53**, 791-846.

4. Drew, H. R., Takano, T., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980), *Nature*, **286**, 567-573.

5. Gupta., G., Bansal, M., and Saisisekharan, V. (1980), *Biochem. Biophys. Res. Commun.*, **95**, 1258-1267.

6. Garcia, A. E. and Gupta, G. (1988), *Biophysical Jl., in the Book of Abstract of the Biophysical Soc.*, 1988.

7. Rajagopal, P. and Feigon, J. (1989), *Biochemistry*, **28**, 7859-7870.

8. Durland, R. H., Kessler, D. J., Gunnel, S., Duvic, M., Pettitt, A., and Hogan M. E. (1991), *Biochemistry*, **30**, 9246-55.

9. Kang, C., Zhang, X., Ratliff, R., Moyzis, R., and Rich, A. (1992), *Nature*, **356**, 126-131.

10. Smith, F. W. and Feigon, J. (1992), *Nature,* **356**, 164-168.

11. Lu, M. Guo, Q. and Kallenbach, N. R. (1992), *Biochemistry* , **31**, 2455-2459.

12. Gupta, G., Garcia, A. E., Guo, Q., Lu, M., and Kallenbach, N. R. (1993), *Biochemistry*, **32**, 7098-7103.

13. 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., and Hilbers, C. W. (1989), *Biochemistry*, **28**, 7491-7498.

14. Gupta, G., Sarma, M. H., Sarma, R. H., Bald, R., Engelke, U., Oei, S. L., Gessner, R., and Erdmann, V. A. (1987), *Biochemistry*, **26**, 7715-7723.

15. Williamson, J. R. and Boxer, S. G., (1989),*Biochemistry*, **28**, 2831-2836.

16. van de Ven, F. J. M. and Hilbers, C. W. (1988), *Eur. J. Biochem.*, **173**, 1-38.

17. Raghunathan, G., Jernigan, R. L., Miles, H. T., and Sasisekharan, V. (1991), *Biochemistry*, **30**, 782-790.

18. Henderson, E. R., Hardin, C. C., Walk, S. K., Tinoco, I., and Blackburn, E. H. (1987), *Cell* , **51**, 899-908.

19. Sen, D. and Gilbert, W. (1988), *Nature*, **334**, 364-366.

20. Sundquist, W. I. and Klug, A. (1989), *Nature*, **342**, 825-829.

21. Williamson, J. R., Raghuraman, M. K., and Cech, T. R. (1990), *Cell*, **59**, 871-880.

22. Gupta, G., Garcia, A. E., and Hiriyanna, K. T., (1993), *Biochemistry*, **32**, 948-960.

23. Garcia, A. E., Gupta, G., Sarma, M. H., and Sarma, R. H. (1988), *J. Biomol. Str. &. Dyn.*, **6**, 525-540.

24. Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L., and Wu., J. -R. (1988), *Proc. Natl. Acad. Sci, USA*, **85**, 6622-6626.

25. Grady, D. I., Ratliff, R. L., Robinson, D. L., McCanlies, E. C., Meyne, J., and Moyzis, R. K. (1992), *Proc. Natl. Acad. Sci., USA*, **89**, 1695-1699.

26. Moyzis, R. K. (1991), *Scientific American*, **265**, 48-55.

27. Sklenar, V. and Bax, A. (1987), *J. Magn. Reson.*, **74**, 469-479.

28. Gupta, G., Sarma, M. H., and Sarma, R. H. (1988),*Biochemistry*, **26**, 7909-7919.

29. Weiner, S. J., Kollman, P. A., Nguyen, D. T., and Case, D. A. (1986),*J. Comp. Chem.*, **7**, 230-245.

30. Garcia, A. E. and Soumpasis, D. M. (1989),*Proc. Nat. Acad. Sci., USA,* **86**, 3160-3164.

31. Garcia, A. E., Gupta, G., Soumpasis, D. M., and Tung, C. S. (1990),*J. Biomol. Str. & Dyn.*, **8**, 173-186.

32. Catasti, P., Gupta, G., Garcia, A. E., Ratliff, R., Hong, L.,Yau, P., Moyzis, R. K., and Bradbury, E. M. (1993), *Biochemistry*, in press.

33. Li, Y., Zon, G., and Wilson, W. D. (1991), *Proc. Natl. Acad. Sci., USA*, **88**, 26-30.

34. Soumpasis, D. M., Tung, C., -S., and Garcia, A. E. (1991), *J. Biomol. Str. Dyn.,* **8**, 867-888.

35. Kennard, O. (1988), *in Structure & Expression : DNA and its Drug Complexes*, vol. 2, pp. 1-26.

36. Prive, G. G., Heinemann, U., Chandrasegaran, Kan, L. S., Kopka, M., and Dickerson, R. E. (1988), *in Structure & Expression : DNA and its Drug Complexes*, vol. 2, pp. 27-48.

Figure Legends

Figure 1 Schematic representation of the critical parts of the human chromosome. The cylindrical parts of the chromosome represent stretches of DNA that do not code for any protein. These stretches of DNA are located at various functional loci, *e.g.*, telomere, matrix attachment sites, and centromere. In this article, structural studies are described for the human centromeric DNA repeats that are responsible for proper segregratiion of duplicated chromosomes during cell division. This diagram is taken from Ref. 26.

Figure 2 1D NMR spectra of $d(AATGG)_4$ and $d(AATGG)_6$ in $H_2O:D_2O$ (9:1) mixture. (A) $d(AATGG)_4$ - high DNA (1.8 mM in DNA strand) and low salt concentration (25 mM NaCl; pH 7) and at 3°C, (B) $(AATGG)_6$ - high DNA (1.8 mM in DNA strand) and low salt concentration (25 mM NaCl; pH 7) and at 3°C.

Figure 3 Identification of different base pairs. *(i) The Presence of Watson-Crick A•T Pairs.* The number of Watson-Crick A•T pairs equals the number of Ts present in the centromeric DNA sequence. For example, in $d(AATGG)_4$, four A•T pairs are expected, and this is exactly what we observe by 1D NOE experiment. In this experiment, the imino protons of T (> 13 ppm in Figure 2) are irradiated, and strong NOEs are observed at H2s of A even at 100 ms of presaturation time -- a characteristic feature of a Watson-Crick A•T pair. NOEs are observed only at four H2s of A suggesting that there are four Watson-Crick A•T pairs. And out of the eight H2s belonging to eight As in $d(AATGG)_4$, H2s belonging to four H-bonded A•T

pairs show high-field shifts as expected. Similarly, for d(AATGG)$_6$, six Watson-Crick A•T pairs are identified. Temperature dependence of the NH signals in Figure 2 show that the T-NH signals from the Watson-Crick A•T pairs dissappear at the highest temperature. This suggests that the stacked Watson-Crick A•T pairs form the core of the centromere structure. By thermal melting and chemical substitution studies, it was also shown that A•T pairs are crucial to the stability of the centromere structures [23]. *(ii) The Nature of A•G Pairs.* Once the imino signals above 12 ppm are accounted for by the the A•T pairs, the signals below11 ppm remain to be identified. They belong to Gs in the A•G and the G•G pairs and the Gs 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 [35,36]. 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 [33]. This leaves two other types of A•G pairings that involve amino protons of G instead of the imino protons [32,33]. One such pairing that is consistent with our NOE data is shown here. In this pairing, the irradiation of the imino proton of G shows a secondary NOE at H8 of A via NH2 of G. Another additional feature of such an A•G pairing is the NOE between 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 5. Even though such an H-bonding 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. [33] 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.* If the G•G pairing involved two imino protons of Gs then these two protons are expected to be located at distinct chemical shifts and strong NOEs are expected between them [12]. However, the irradiations of the signals at 10.8 ppm did not produce any NOE at 9.9 ppm or vice-versa. However, the irradiation of the G-NH signal at 9.9 ppm results in a primary NOE at NH2 of G and secondary NOE at H8 of G. This is consistent with the G(anti)•G(*syn*) pairing as shown here. The NMR data shows that both Gs in the G•G pair undergo rapid *syn/anti* flip-flop without exposing the imino proton to the solvent. The observed intra-nucleotide H1'(G)---H8(G) NOE, though strong, is only the average of *syn/anti* conformations. The strong inter-nucleotide H1'(Gi-1)-H8(Gi) contact is also consistent with the G•G pairing shown here. 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 [12]. The G•G part happens to be the most flexible region of the structures of d(AATGG)$_{4,6}$. Thermal melting and base substitution studies also confirm this observation [32], *i. e.*, substitutions of the G•G pairs by any other mismatches have little effect on the melting temperature.

Figure 4 Schematic representations of the stem-loop structures for **(A)** d(AATGG)$_4$ and **(B)** d(AATGG)$_6$. The A•T, A•G, and G•G base pairs are shown in different colors.

Figure 5A *(Left)* 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 inter-nucleotide NOEs reveal that nine nucleotides (A1 through G9) exist predominantly in (*C2'-endo, anti*) conformations while one of two G10s shows an *anti* to *syn* conversion to facilitate the G10•G10 pair. Inter-nucleotide 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 inter-proton distances as independent constraints for structure derivation. Intra-nucleotide H3'---H8/H6 NOEs are shown. Inter-nucleotide H3'(i-1)---H8/H6(i) NOEs are also observed but the connectivity pattern is not shown in order to preserve the clarity of the diagram. H3' and H1' chemical shift regions are non-overlapping except for G10. Note the presence of the inter-molecular NOESY cross-peak (marked *) between A1 and G10.*(Right)* 2D NOESY (τ_m = 250 ms) spectrum of d(AATGG)$_4$ in D_2O for the H1' vs. H8/H6 cross-section (1.8 mM in DNA strand, 25 mM NaCl, pH 7, temperature 3 °C). Note that the corresponding NOESY cross-section for the stem-loop motif of [d(AATGG)$_2$]$_2$ shows a close similarity with this cross-section. (G4-G5-A6) and (G14-G15-A16) form the two loop segments while the rest of the nucleotides form the stem with A•T, A•G, and A•G pairs.

Figure 5B *(Left)* 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 inter-nucleotide 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. Inter-nucleotide NOEs involving H1'(A2/A7/A12)--H2(A1/A6/A11) (weak NOEs) and H1'(G4/G14)--H8(G5/G15) (strong NOEs) are indicative of special stacking patterns at the A-G and G-G steps. Note the high field shift of H8,H1'(G5/15). Full-Matrix NOESY simulations with respect to the observed data at τ_m = 250 and 100 ms allow us to extract 150 independent inter-proton distances as structural constraints for molecular model building of a stem-loop motif. Intra-nucleotide H3'---H8/H6 NOEs are shown. Inter-nucleotide 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 inter-molecular NOE (marked by *) between A1 and G15. *(Right)* 2D NOESY (τ_m = 250 ms) spectrum of d(AATGG)$_6$ in D_2O for the H1' vs. H8/H6 cross-section (1.8 mM in DNA strand, 25 mM NaCl, pH 7, temperature 3 °C). Note that the corresponding NOESY cross-section for the stem-loop motif of [d(AATGG)$_3$]$_2$ shows a close similarity with this cross-section. In this structure, (G9-G10-A11) and (G24-G25-A26) form the two loop segments while the rest of the nucleotides form the stem with A•T, A•G, and G•G pairs.

Figure 6A Electrophoretic pattern of centromeric DNA repeats in non-denaturing gel. About 12 µl of the sample containing 0.1 mg of DNA were loaded in each lane. No mobility difference was found by changing the DNA concentration. Lane 1: Mixture of markers; Lane 2: Watson-Crick duplex

d$(GGAAT)_6$.d$(ATTCC)_6$; Lane 3: Watson-Crick duplex d$(GGAAT)_4$.d$(ATTCC)_4$; Lane 4: d$(AATGG)_6$; Lane 5: d$(AATGG)_4$, Lane 6: d$(GGAAT)_6$; and Lane 7: d$(GGAAT)_4$.

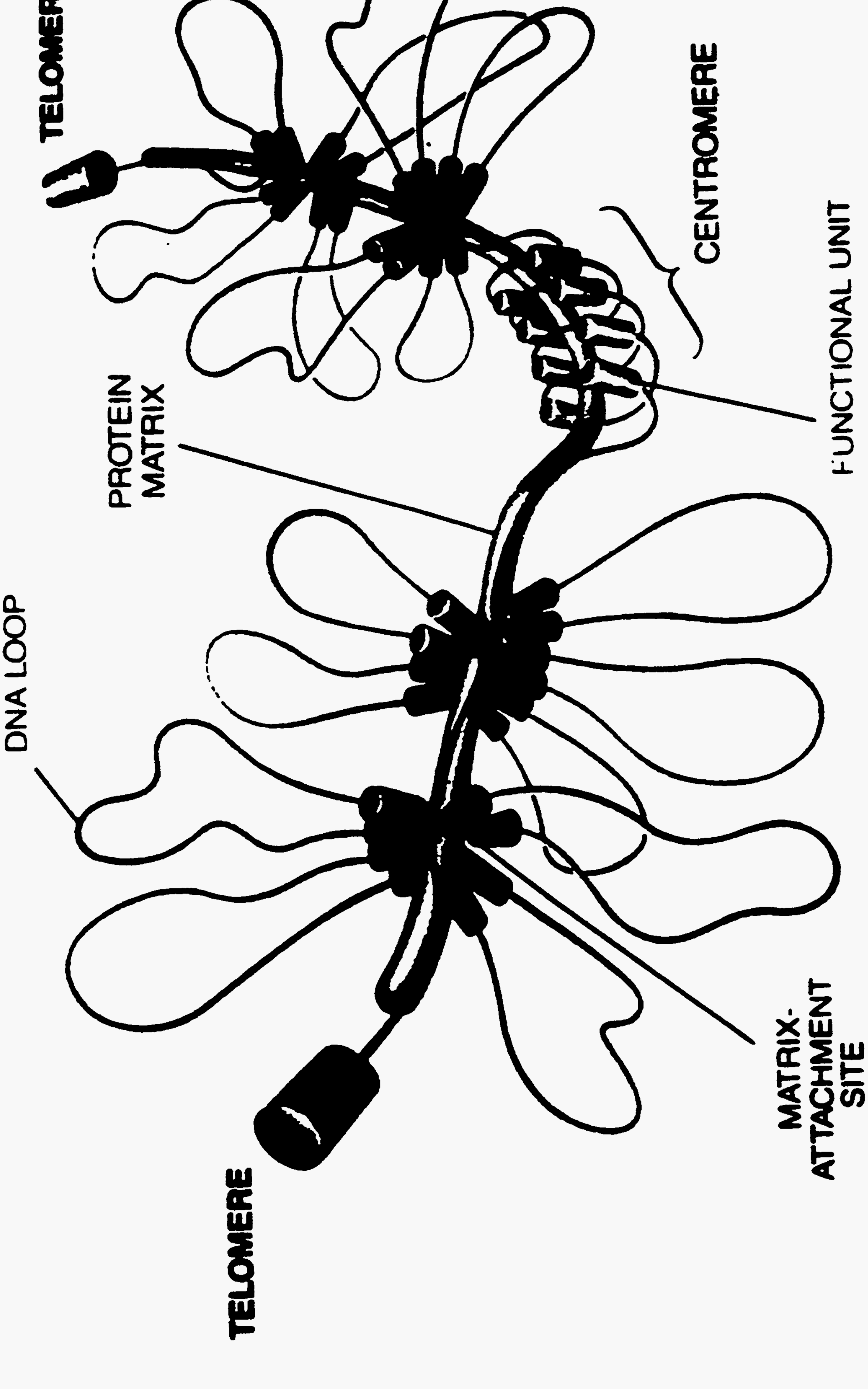
Figure 6B Digestion profiles of various centromeric repeats by mung bean nuclease (a probe for single-stranded regions in DNA). Lane 1: untreated d$(AATGG)_{2,3,4,5,6}$ used as markers; Lane 2: digestion of d$(AATGG)_6$ for 2.5 minutes; Lane 3: digestion of d$(AATGG)_6$ for 5 minutes; Lane 4: digestion of d$(AATGG)_4$ for 2.5 minutes; Lane 5: digestion of d$(AATGG)_4$ for 5 minutes; Lane 6: untreated d$(GGAAT)_{2,3,4,5,6}$ and a 14-mer DNA used as markers; Lane 7: digestion of d$(GGAAT)_6$ for 2.5 minutes; and Lane 8: digestion of d$(AATGG)_6$ for 5 minutes. Note that the duration of digestion does not alter the nature of cleavage.

Figure 7 The skeletal model of the stem-loop structure of (A) d$(AATGG)_4$ on the left and that of (B) d$(AATGG)_6$ on the right. Only the non-hydrogen atoms are shown; C=green, N=blue, P=yellow, O=red. The 5'-end of the structures are 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 all sampled local minima obtained after HTMD/RTQ calculations subject to the NOE constraints.

Figure 8 The presence of segmental flexibility in the stem-loop structure of d$(AATGG)_6$. (Inset) The numbering scheme of the bases in the structure. Three independent segments of the structure are boxed. Also marked are the diagonal or inter-strand interactions at the junctions. The

flexibility at the two junctions lead to two types of stem-loop structures. Stereo-views of them are shown. (A) The straight structure where the top and bottom loop segments are aligned with the middle segment and (B) the bent or kinked structure where the top and the bottom loop segment are bent with respect to the middle segment. In A & B, the backbone atoms are shown in orange-red in a skeleton representation while the base atoms are shown as green *van der Waal* spheres.

Figure 9 Nearest-neighbor interaction energies (in KCal/Mole) for all the nucleosides in the stem-loop structure of d(AATGG)$_6$. Note that the positions of A1 & A16 and A11 & A26 are interchangeable in terms of the NMR data and the final structure. The energy values given here correspond to a representative structure that has the lowest energy in the cluster. The nearest-neighbor interactions are computed after taking out the phosphate groups from the lowest-energy structure. Therefore, interaction energy between two neighbors refer to the energy of interaction between the corresponding nucleosides. (A) Pairing energy between two nucleosides facing each other, i.e., G4 & A6, T3 & A7, G30 & G10, etc.,. Note that the nucleosides G5 & G20 at the tip of the two loops do not have any pairing energy. (B) Intra-strand 5'-3' stacking energy; interaction energy between G30 & A1 is also included as stacking energy. The 5'-3' stacking energy for A1 is between A1 & A2 and similarly for other nucleotides. (C) Inter-strand (or diagonal) stacking energies between bases located in one plane up and down in the opposing halves of the molecule. For example, for A1 these interactions will include the pairs (A1, T8) and (A1,G10). (D) The total interaction energy, comprising all the terms in (A)-(C), for each nucleoside.


TELOMERE
DNA LOOP
PROTEIN MATRIX
MATRIX-ATTACHMENT SITE
FUNCTIONAL UNIT
CENTROMERE
CENTROMERE
TELOMER

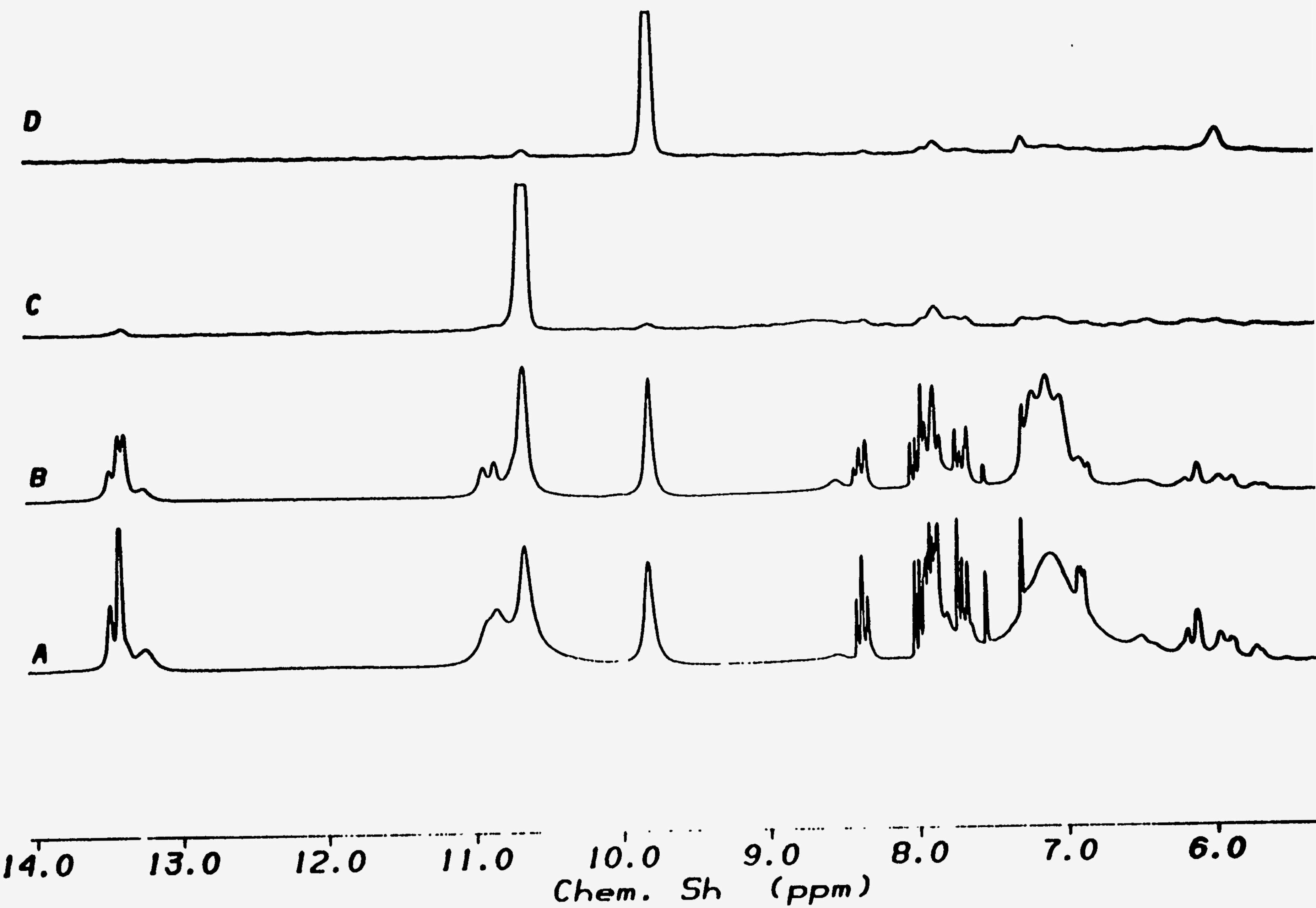
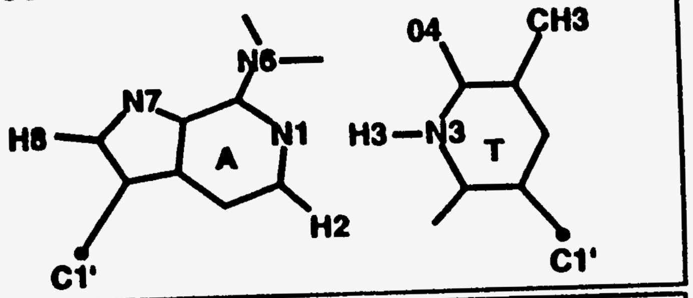


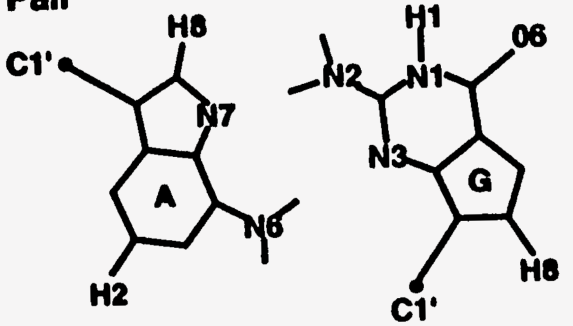
Fig. 8

IS-1/93-1317 6/93

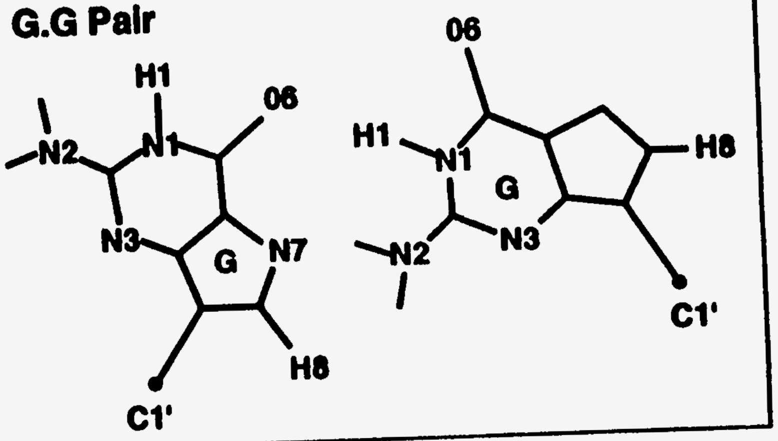
Watson-Crick A.T Pair



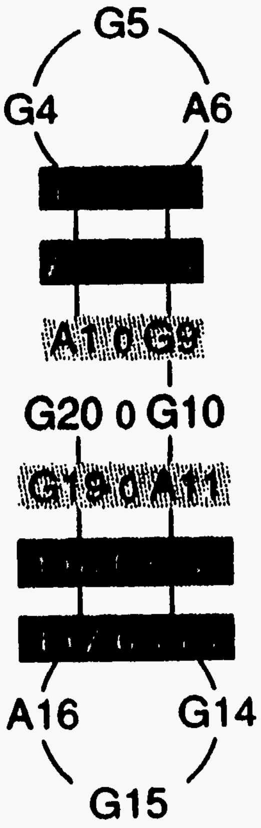
A.G Pair



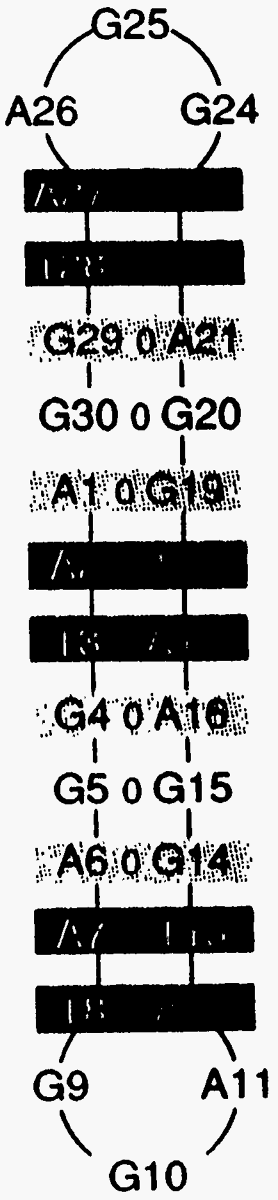
G.G Pair



$(AATGG)_4$
A Stem-Loop Motif



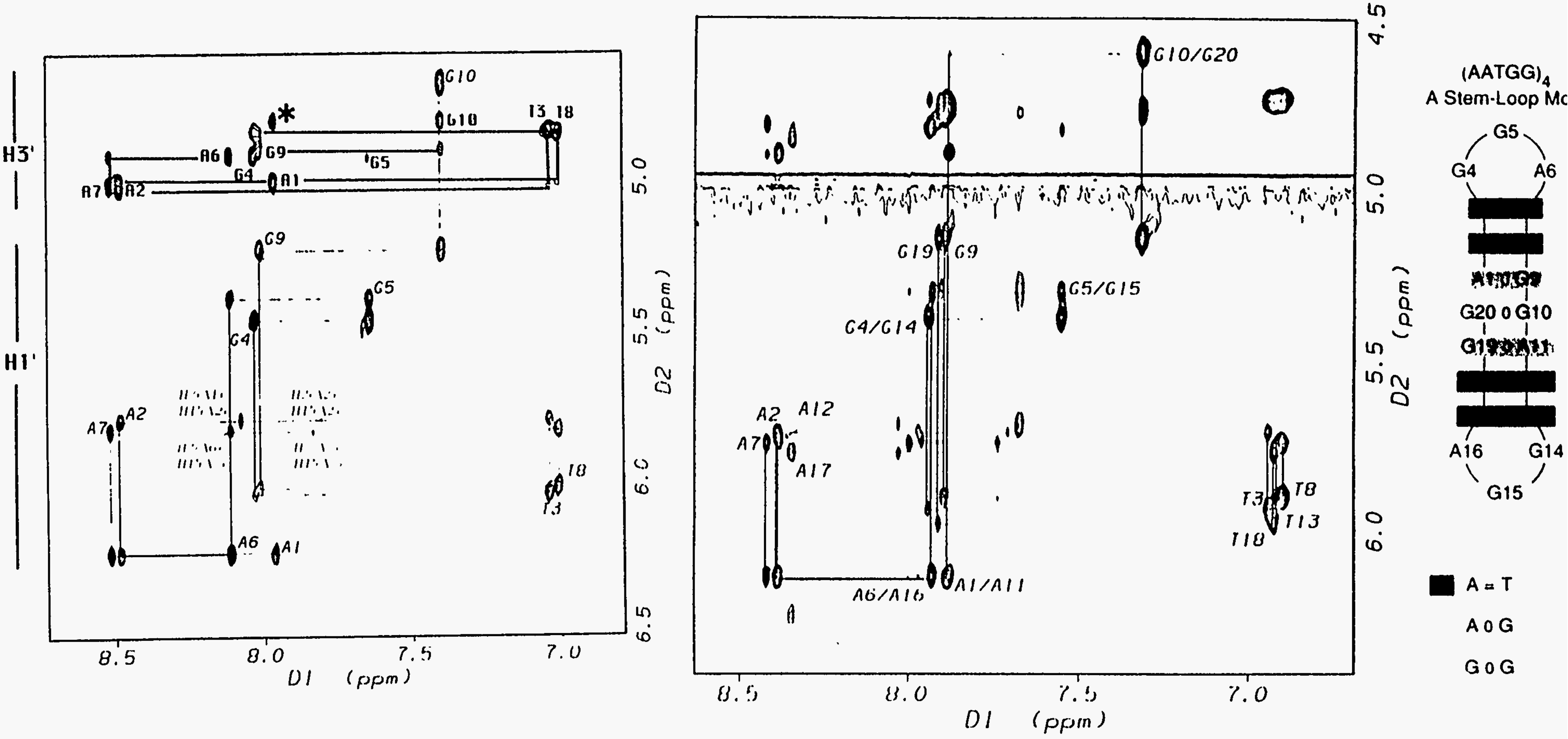
$(AATGG)_6$
A Stem-Loop Motif

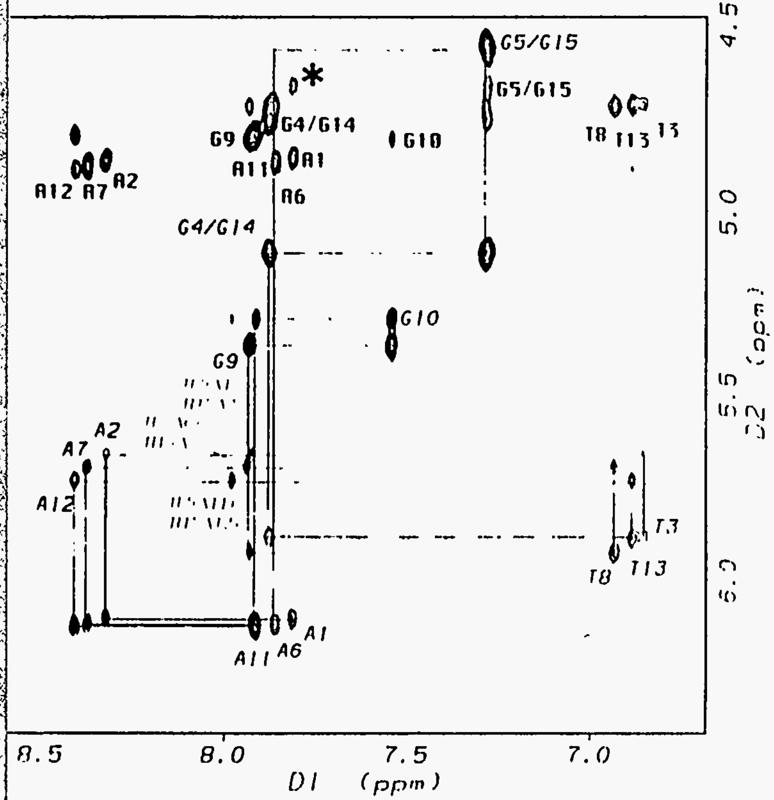


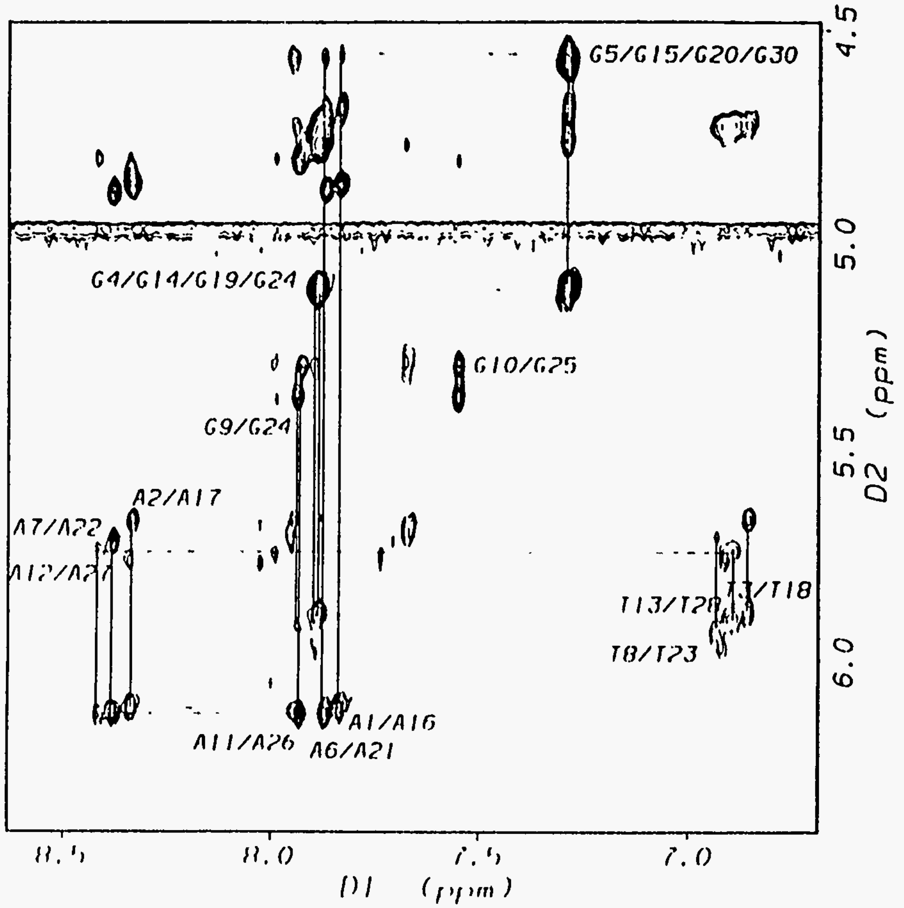
■ A = T

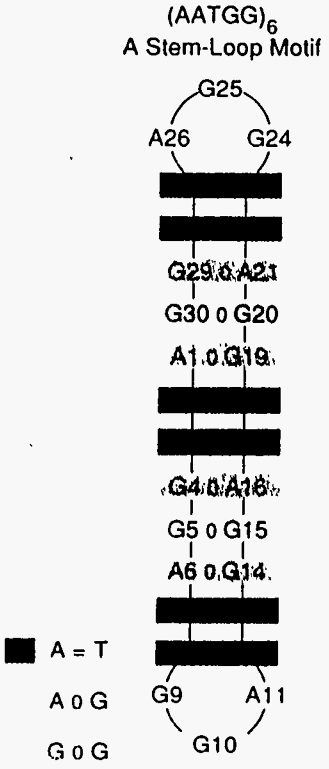
▨ A 0 G

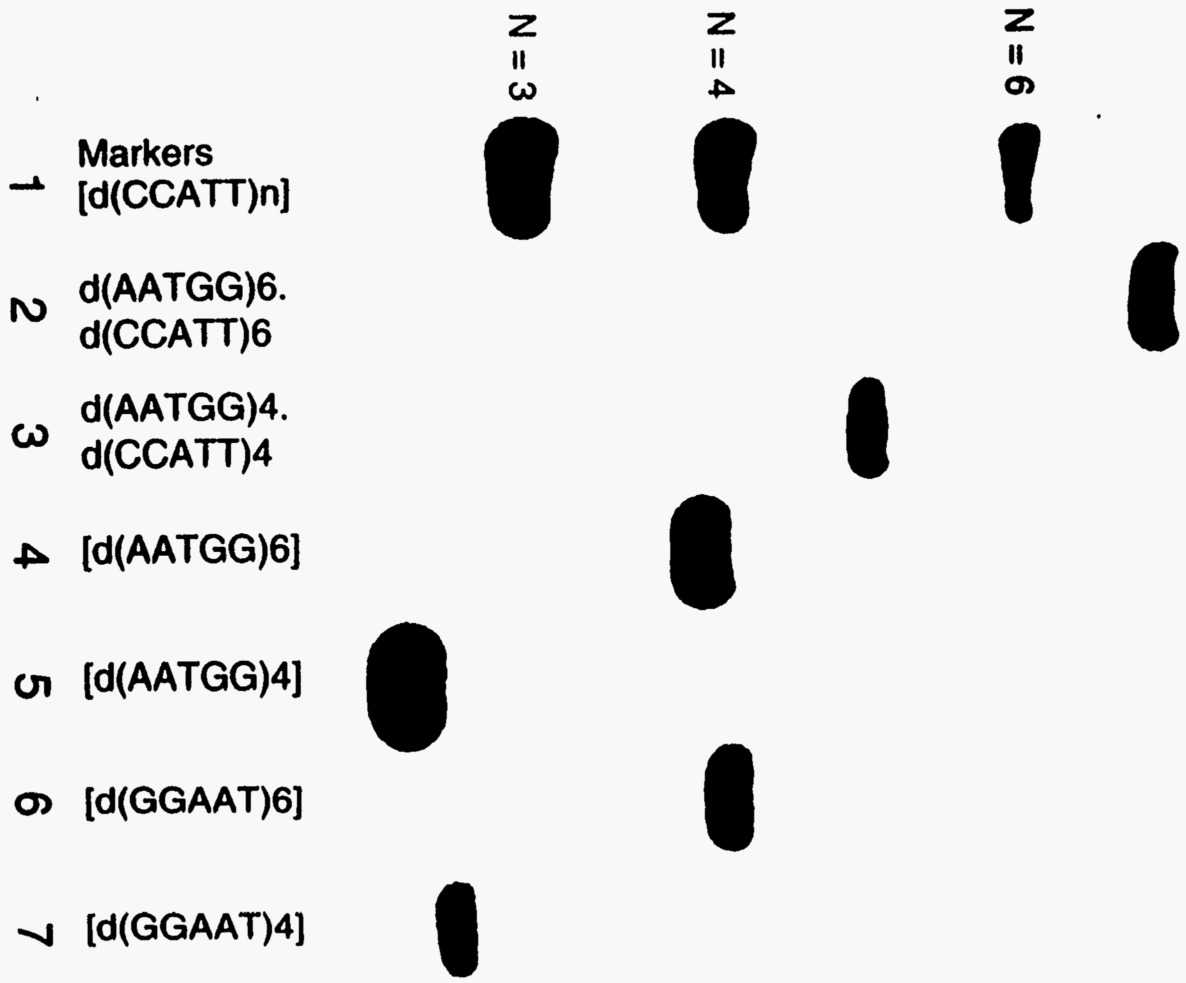
G 0 G

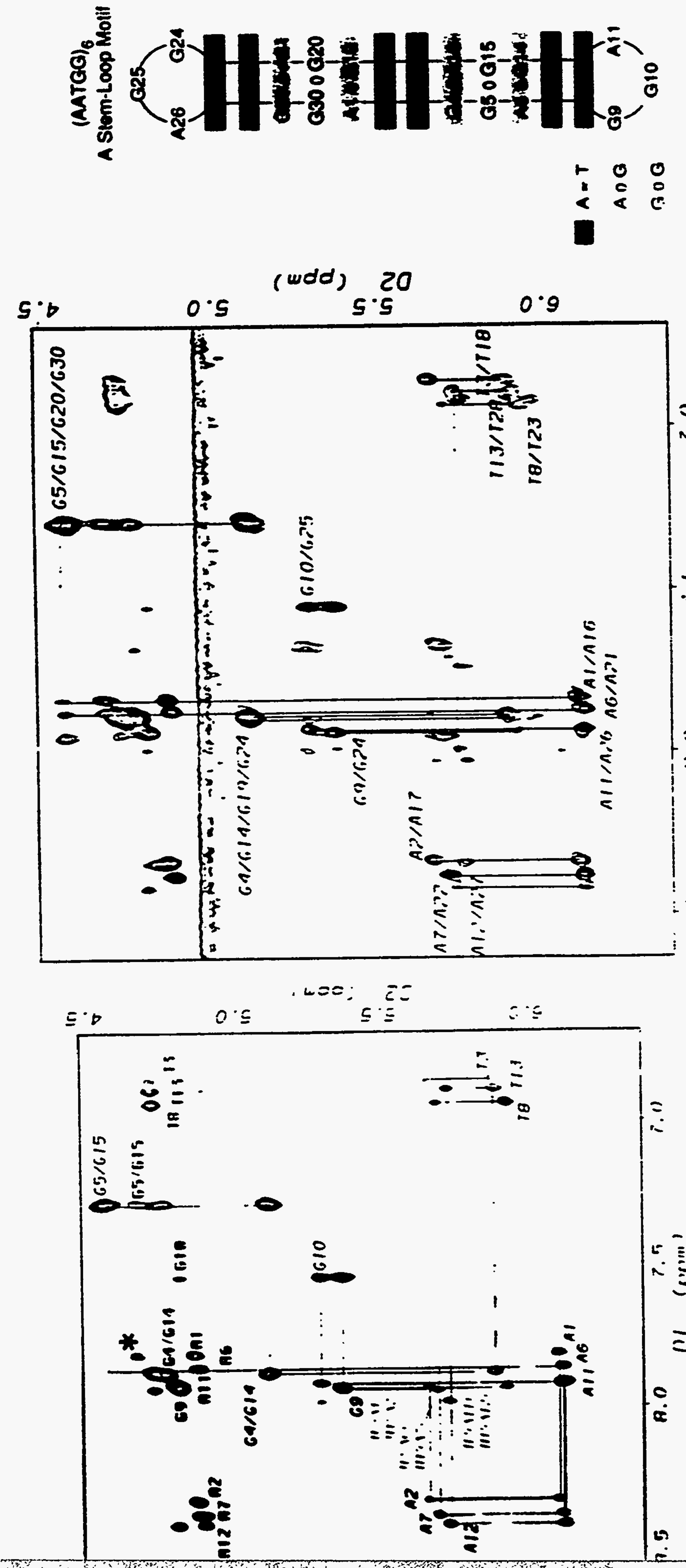

H3'
H1'
G10
G10
T3 18
*
G9
G5
A6
G4
A1
A7
A2
G9
G5
G4
A7
A2
18
T3
A6
A1
8.5
8.0
7.5
7.0
D1 (ppm)
5.0
5.5
6.0
6.5
D2 (ppm)
G10/G20
G19
G9
G5/G15
G4/G14
A2
A12
A7
A17
T8
T3
T13
T18
A6/A16
A1/A11
8.5
8.0
7.5
7.0
D1 (ppm)
4.5
5.0
5.5
6.0
D2 (ppm)
(AATGG)$_4$
A Stem-Loop Mc
G5
G4
A6
G20 o G10
A16
G14
G15
A = T
A o G
G o G

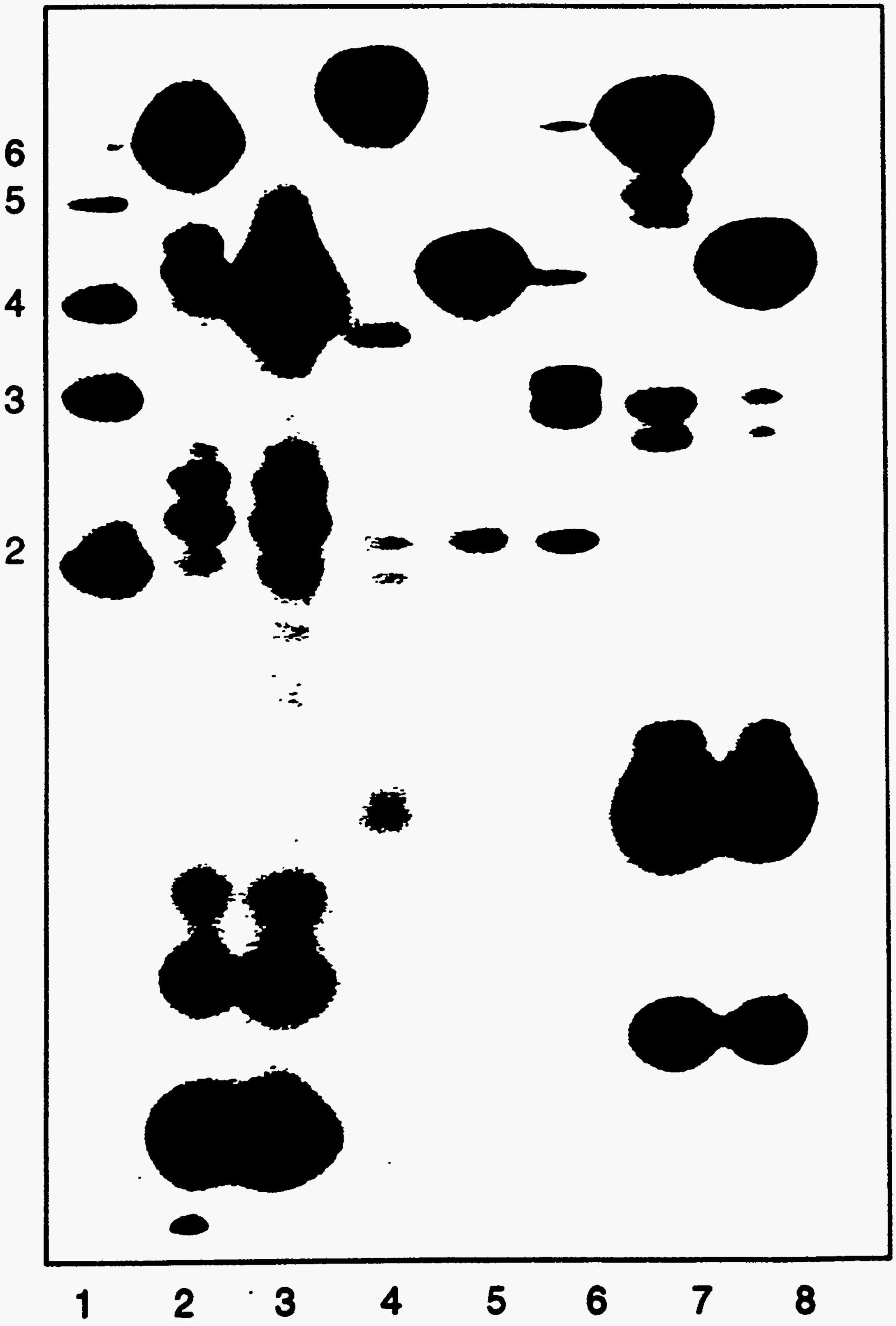

G5/G15
G5/G15
*
G4/G14
G9
A11
A1
A6
A12
A7
A2
G10
T8
T13
T3
G4/G14
G10
G9
A2
A7
A12
T3
T8
T13
A1
A6
A11
4.5
5.0
5.5
6.0
D2 (ppm)
8.5
8.0
7.5
7.0
D1 (ppm)

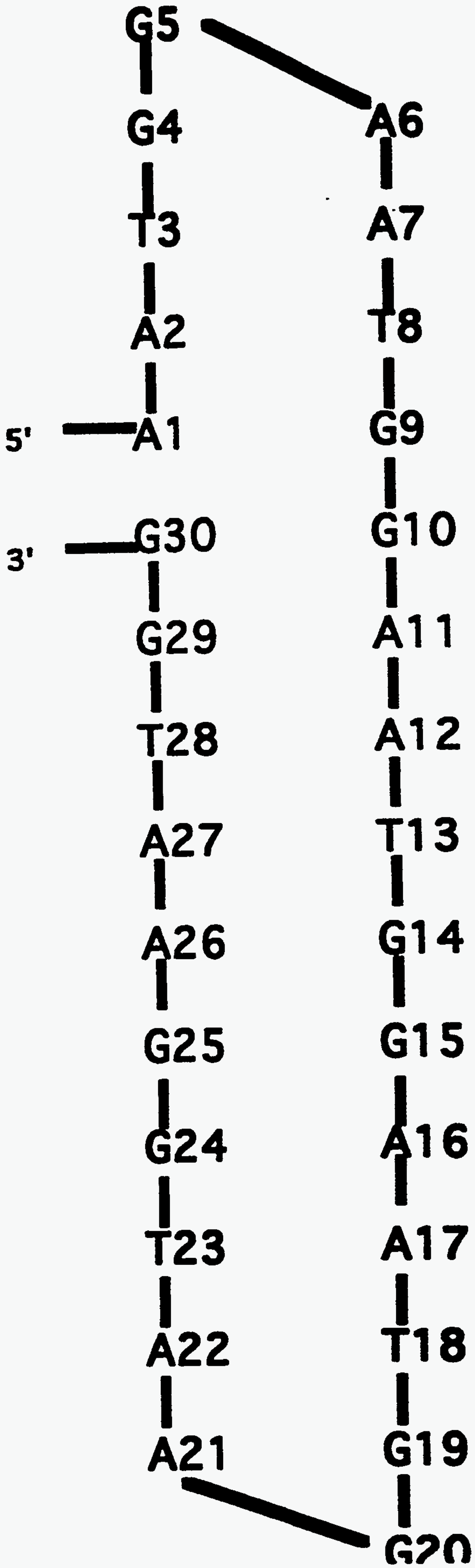

G5/G15/G20/G30
G4/G14/G19/G24
G10/G25
G9/G24
A2/A17
A7/A22
A12/A27
T13/T28
T3/T18
T8/T23
A1/A16
A11/A26
A6/A21
4.5
5.0
5.5
6.0
D2 (ppm)
8.5
8.0
7.5
7.0
D1 (ppm)

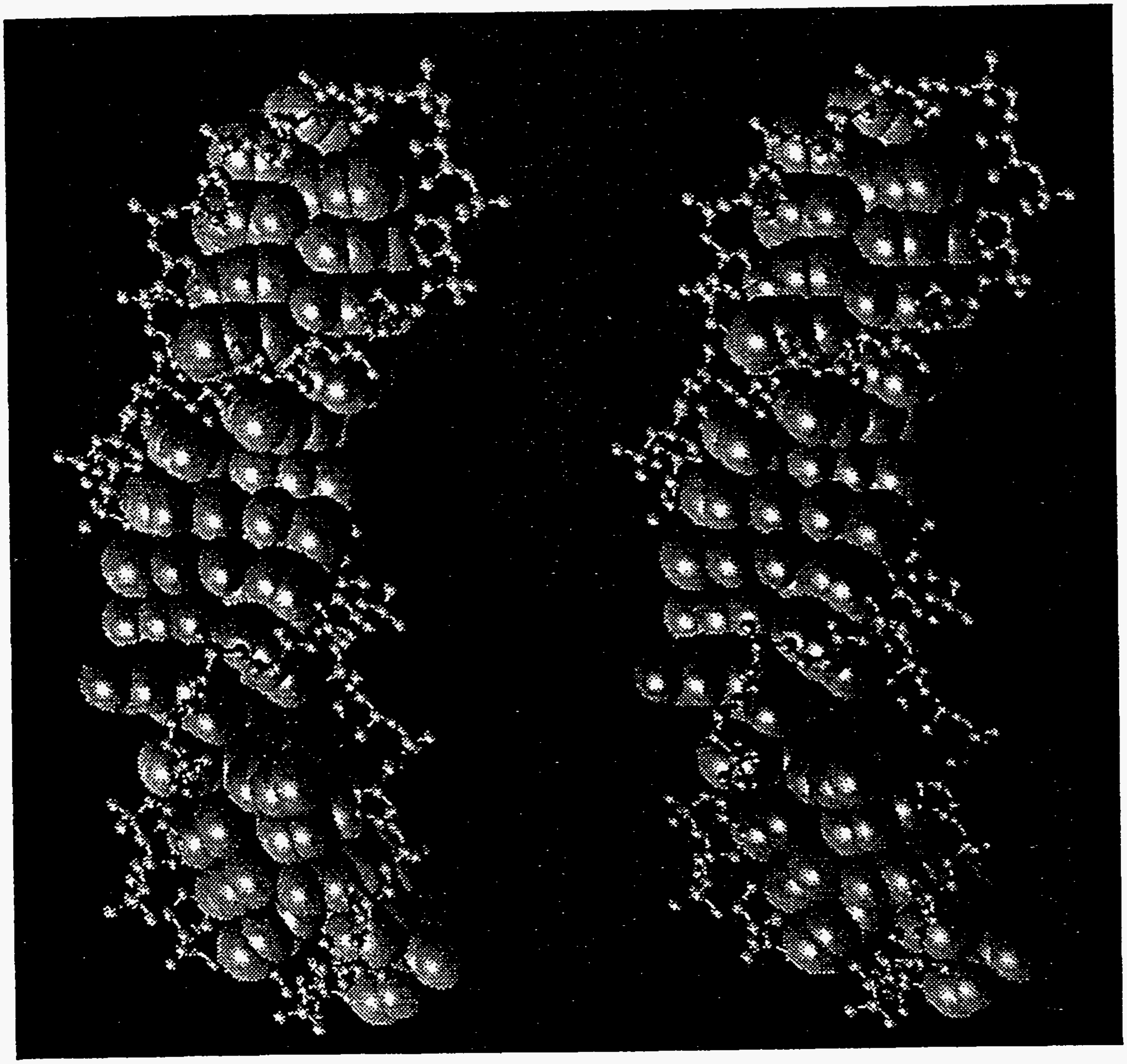

(AATGG)$_6$
A Stem-Loop Motif
G25
A26
G24
G29 0 A21
G30 0 G20
A1 0 G19
G4 0 A16
G5 0 G15
A6 0 G14
G9
A11
G10
A = T
A 0 G
G 0 G

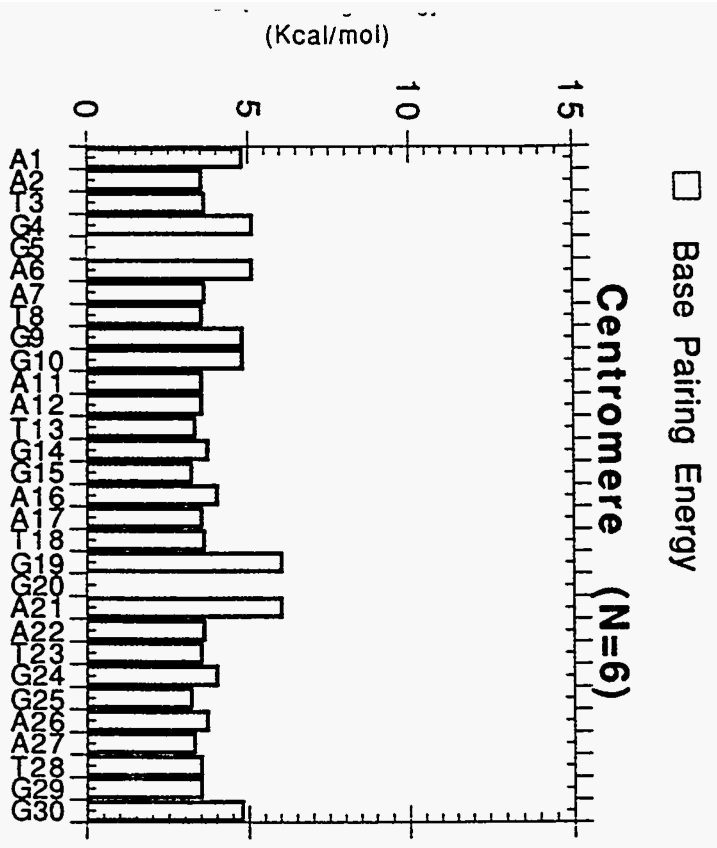

N = 3
N = 4
N = 6
1 Markers [d(CCATT)n]
2 d(AATGG)6. d(CCATT)6
3 d(AATGG)4. d(CCATT)4
4 [d(AATGG)6]
5 [d(AATGG)4]
6 [d(GGAAT)6]
7 [d(GGAAT)4]


(AATGG)$_6$
A Stem-Loop Motif
G25
A26
G24
G30 ○ G20
G5 ○ G15
G9
G10
A11
A = T
A ○ G
G ○ G
D2 (ppm)
4.5
5.0
5.5
6.0
D1 (ppm)
7.0
7.5
8.0
8.5
G5/G15/G20/G30
T13/T28
T18
T8/T23
G10/G25
A1/A16
A6/A21
A11/A26
G9/G24
G4/G14/G19/G24
A2/A17
A7/A22
A12/A27
G5/G15
T8
T13
T3
G10
G4/G14
G9
A11
A1
A6
A2
A7
A12
*

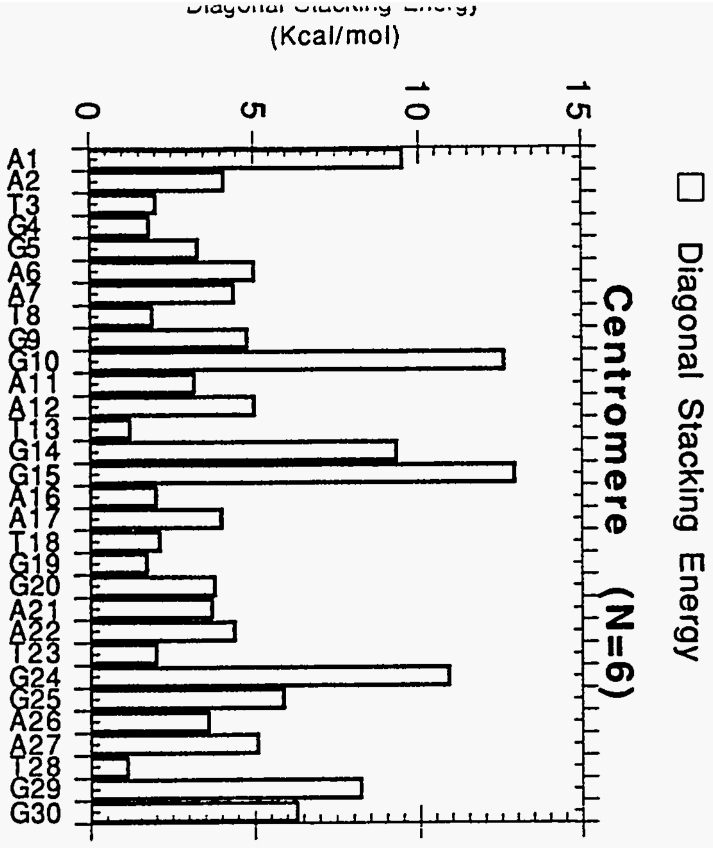

6
5
4
3
2
1
2
3
4
5
6
7
8


G5
G4
T3
A2
5'
A1
3'
G30
G29
T28
A27
A26
G25
G24
T23
A22
A21
A6
A7
T8
G9
G10
A11
A12
T13
G14
G15
A16
A17
T18
G19
G20

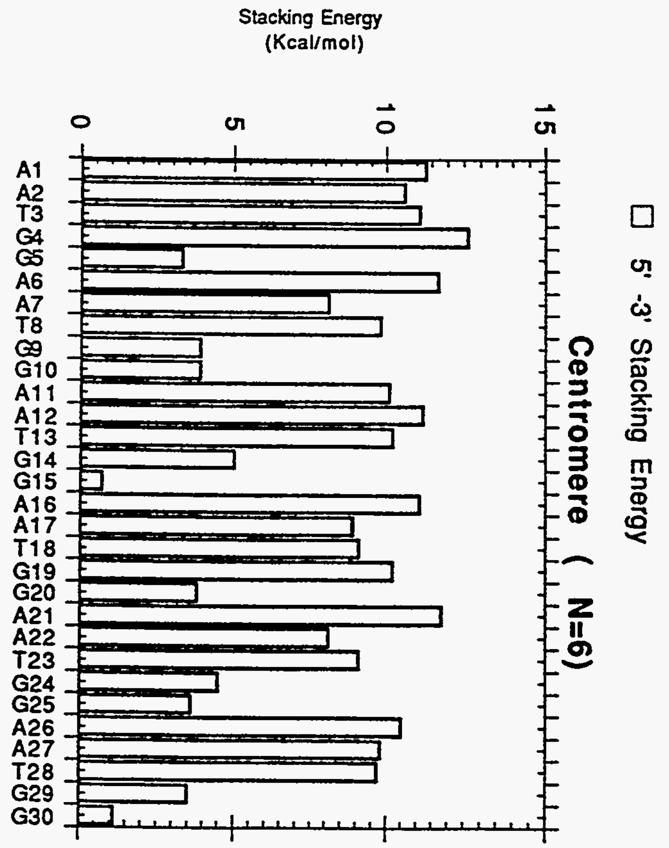




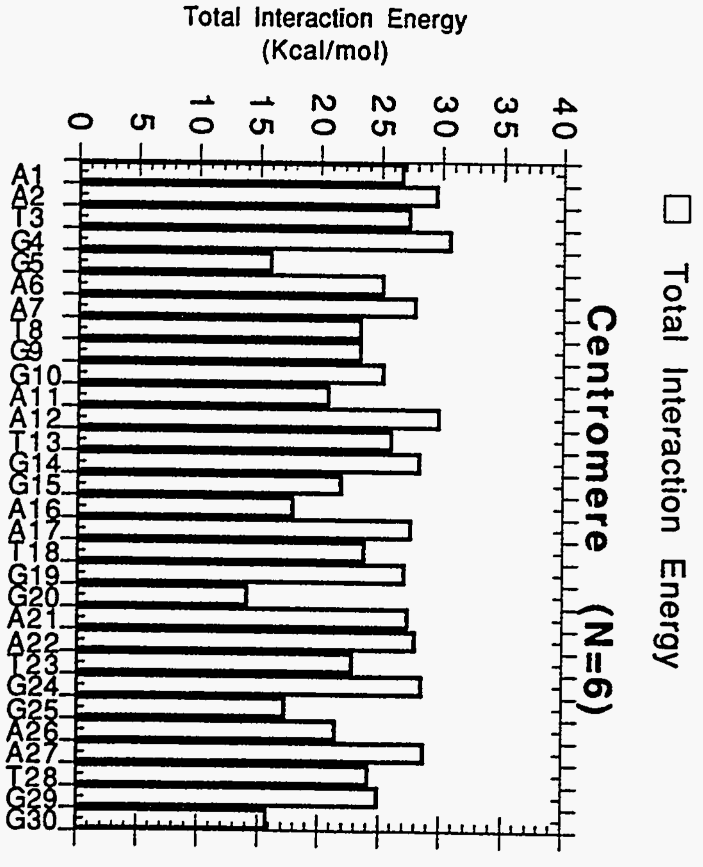
(Kcal/mol)
0
5
10
15
□ Base Pairing Energy
Centromere (N=6)
A1
A2
T3
G4
G5
A6
A7
T8
G9
G10
A11
A12
T13
G14
G15
A16
A17
T18
G19
G20
A21
A22
T23
G24
G25
A26
A27
T28
G29
G30



Diagonal Stacking Energy
(Kcal/mol)
0
5
10
15
□ Diagonal Stacking Energy
Centromere (N=6)
A1
A2
T3
G4
G5
A6
A7
T8
G9
G10
A11
A12
T13
G14
G15
A16
A17
T18
G19
G20
A21
A22
T23
G24
G25
A26
A27
T28
G29
G30



Stacking Energy
(Kcal/mol)
0
5
10
15
□ 5' -3' Stacking Energy
Centromere (N=6)
A1
A2
T3
G4
G5
A6
A7
T8
G9
G10
A11
A12
T13
G14
G15
A16
A17
T18
G19
G20
A21
A22
T23
G24
G25
A26
A27
T28
G29
G30



Total Interaction Energy
(Kcal/mol)
0
5
10
15
20
25
30
35
40
□ Total Interaction Energy
Centromere (N=6)
A1
A2
T3
G4
G5
A6
A7
T8
G9
G10
A11
A12
T13
G14
G15
A16
A17
T18
G19
G20
A21
A22
T23
G24
G25
A26
A27
T28
G29
G30