Sampling of the Conformations of the d(CGCTGCGGC) Hairpin in Solution by Two-Dimensional Nuclear Magnetic Resonance and Theoretical Methods†

Goutam Gupta* and A. E. García

Theoretical Division, Theoretical Biology and Biophysics Group (T-10), Los Alamos National Laboratory, Los Alamos, New Mexico 87545

K. T. Hiriyanna

Zoology Department, Iowa State University of Science and Technology, Ames, Iowa 50011

Received October 24, 1991; Revised Manuscript Received July 16, 1992

ABSTRACT: Most NMR studies of DNA oligomers have focused on rigid structures that show a strong preference for one or a small set of ground-state conformations. There is an increasing interest in extending NMR methods to investigate DNA systems in which this preference does not exist. A DNA hairpin is one such system where a large number of low-energy structures coexist in solution. In this article we show how 1D/2D NMR data of the d(C1-G2-C3-T4-G5-C6-G7-G8-C9) hairpin are used to map the conformational space of this molecule. First, we characterize the gross morphology of the hairpin by monitoring the exchangeable imino signals in the molecule. Second, we extract a set of inter-proton distances (i.e., the average values and the associated dispersions) for various pairwise interactions by performing full-matrix NOESY simulation with respect to the observed NOESY data for mixing times of 250 and 100 ms. Third, we use these distances as structural constraints to perform a 300-ps molecular dynamics simulation at 500 K. Fourth, we extract 600 snapshots (one after every 0.5 ps) from the MD trajectory and perform constrained energy minimization to map local minima on the sampled energy surface (we call this the rapid temperature quenching step). Fifth, we assign 600 structures to 14 disjoint clusters such that conformationally similar hairpins belong to the same cluster while conformationally distinct hairpins belong to different clusters. Finally, we interpret the NOESY data in terms of conformationally distinct structures by recalculating NOESY contributions taken from representative structures of different clusters. Our analyses clearly demonstrate that the NMR data correspond to an ensemble of distinct structures, i.e., a set of energetically stable but conformationally distinct structures that satisfies the constraints of loop folding in the d(C1-G2-C3-T4-G5-C6-G7-G8-C9) hairpin. Two types of loop folding consistent with NMR data are obtained: (i) a hairpin with two G·C pairs in the stem and four residues in the loop and (ii) a hairpin with two G·C pairs and a reverse wobble G·T pair in the stem plus two residues in the loop.

Short DNA oligomers with a self-complementary sequence can adopt, as a function of the solution properties, a great variety of conformations. A double-helical conformation is formed at high DNA concentration, moderately low salt concentration (0.15 M), and low temperature. At higher temperatures (or lower salt and DNA concentrations) a monomeric hairpin structure is formed, which can be taken as an intermediate state between a double helix and a coil structure (Marky et al., 1983). By changing the solution conditions or by replacing Watson–Crick pairs with mismatches in the self-complementary double helix, it is possible to drive the equilibrium at low temperature from a duplex to a hairpin conformation (Gupta et al., 1987; Howard et al., 1991; Raghunathan et al., 1991; Williamson & Boxer, 1989a,b). The equilibria between mismatched duplexes and monomeric hairpins may play an important role in the recognition of tertiary structure alterations caused by mismatches. In addition to the recognition of mismatches as local defects on the DNA duplex, the mismatch repair mechanism may recognize global changes in structure—such as the formation of hairpins. This mechanism may act as an additional recognition of mismatches in otherwise palindromic sequences near recognition or control sites in DNA. Formation of cruciforms (i.e., two hairpins) is known to release superhelical stress in a circular DNA (Lilley, 1985). Monomeric hairpins form integral structural components in single-stranded virus replication (Bearns & Bohernzky, 1987; Chen et al., 1989). They also constitute primase binding sites in single-stranded circular fd and G4 phage DNA (Bouche et al., 1978) and termination signals for transcription (Briat et al., 1987).

Hairpins contain two structurally and dynamically distinct domains: a base-paired stem and a single-stranded loop connecting the two halves of the stem. The stems of hairpins show the same response to changes in solution conditions as double-helical DNA oligomers; e.g., hairpin stems can form right-handed B (Hare & Reid, 1986) or left-handed Z helices (Antosiewicz et al., 1988; Xodo et al., 1986) as a response to solution salt conditions. The thermodynamic stability of the stem region does not vary significantly from that of the corresponding duplex structure (Rentzeperis et al., 1991). It was commonly believed that the stem of a hairpin should be at least 3 base pairs long, but recent studies have shown that stable hairpin structures are possible with 2 G·C base pairs in the stem (Gupta et al., 1987). In contrast to the double-helical stem, the loop region shows a wide range of folding patterns that depend on sequence. The relative stability between a duplex and hairpin structure depends on the counterbalance among various interactions, e.g., the number of hydrogen bonds broken during the duplex-to-hairpin transition, the stacking of the bases in the loop (if any), and the entropic stabilization of the loop region due to large

† Work supported by the U.S. Department of Energy.

0006-2960/93/0432-948$04.00/0

dynamical fluctuations of atoms around their equilibrium positions.

Current understanding concerning the stability of the loop region of hairpins is limited. It is known that the stacking of the bases in the loop region of hairpins depends on the length (Benight et al., 1989) and the sequence of the loop region (Senior et al., 1988). Calorimetric and NMR studies of DNA hairpin structures with identical stem sequences showed that the hairpins are most stable when there are four bases in the loop region (Benight et al., 1989; Marky et al., 1983; van de Ven & Hilbers, 1988): Homosequences in the loop showed the following stabilities: T4 > C4 > A4 > G4 (Senior et al., 1988). NMR studies indicate that the 3′-end base of the loop region stacks very well with the 3′-side of the base-paired stem. However, details of the loop folding will depend on the nature of base–base and base–backbone interactions in the loop segment. For example, the first base in the 5′-end of the loop can form a base pair with the last base in the 3′-end of the loop (van de Ven & Hilbers, 1988). In our previous studies (García et al., 1988), on the d(C1-G2-C3-C4-G5-C6-A7-G8-C9) hairpin, we observed an H-bond between the first base (i.e., N4 of C4) and the fourth base (i.e., N7 of A7) of the loop.

In this work we present 1D/2D NMR and theoretical studies on the oligomer d(C1-G2-C3-T4-G5-C6-G7-G8-C9) in the hairpin conformation. By comparison of this sequence to previously studied sequences (Gupta et al., 1987), the replacements of C4 with T4 and A7 with G7 in this hairpin sequence leads to the possibility of forming a two-H-bonded T4–G7 base pair, which is easily detected in the NMR spectra. We expect to find either a hairpin with 3 base pairs in the stem and 2 bases in the loop or a hairpin with 2 base pairs in the stem and 4 bases in the loop. While the first conformation is stabilized by a longer stem region (3 base pairs) and may be destabilized by a shorter loop region, the second conformation has a less stable stem region but a longer and presumably more stable loop region. Therefore, the counterbalance of stabilities of the loop and the stem region will determine the predominant conformation.

Generally, the number of distance constraints obtained from NMR data is far less than the number of variables in the system. Therefore, NMR data cannot uniquely define the structure of the system; i.e., more than one structure can satisfy the same NMR data. As a result, the interpretation of NMR data should strictly involve the characterization of an ensemble of structures that satisfies the NMR data. The description of the structures consistent with NMR data using such an ensemble is particularly necessary for flexible systems like hairpins, for which several conformationally distinct but energetically stable structures exist in solution and interconvert within the NMR time scale.

In contrast to most NMR studies of more rigid duplex structures, we show that a wide range of structures can fulfill the NMR constraints. We show this by generating a series of 600 structures corresponding to local potential energy minima that are consistent with the NMR data. These structures are computed using a molecular dynamics simulated annealing (MDSA) followed by a rapid temperature quenching at various points along the dynamics trajectory (Stillinger & Weber, 1984). In order to distinguish local and global rearrangements of atoms in those structures, we define a hierarchy of structures by dividing the structures progressively among clusters. These clusters are constructed by using the *mean square* distance between all pairs of structures. A *single linkage* (Lebart et al., 1984) distance is used to define the distances between clusters once a cluster contains more than one structure. The resulting hierarchy is ultrametric (Rammal et al., 1986). This hierarchy has 600 disjoint clusters at one end, where all structures are considered globally distinct (unless the distance is exactly zero), and only one cluster at the other end, where every change in structure is considered a local variation around the average structure. The level along the hierarchy at which one *cuts* the tree is arbitrary. By cutting closer to the initial points we get a larger set of clusters, while cutting closer to the end we get fewer clusters. For the oligomer studied here we chose to study 14 disjoint clusters. Representative structures of each cluster show various patterns of loop bases stacking, as well as two extremely distinct loop foldings. These results are consistent with our data, which show the coexistence of the two hairpins: one minor population with a T4·G7 wobble base pair and a major population without this base pair. The hairpin structures with T4·G7 base pairs showed a reverse wobble G·T pairing where T4 adopts a glycosyl torsion in the low anti domain (χ = 180° – 200°) consistent with the 2D NOESY and ROESY data for the minor conformer.

METHODOLOGY

Oligomer Synthesis and Purification. The DNA oligomer was synthesized using the Applied Biosym synthesizer and purified on an HPLC column. The product was then ethanol-precipitated and dried several times. The dried oligomer was then used for making appropriate solutions for NMR experiments.

NMR Experiments. All the NMR experiments were carried out on a 500-MHz Varian (UNITY) instrument at the NMR Facility, Iowa State University, Ames, IA. Data were processed on a SPARC workstation using Varian software. Chemical shift values are given with respect to DSS as an internal reference. 1D spectra in H_2O–D_2O (9:1) were recorded using $1\bar{1}$ spin-echo pulse sequence of Sklenar and Bax (1987). NOESY spectra in D_2O (mixing times, τ_m = 100 and 250 ms) were collected using the method of States et al. (1982), and the HDO signal was presaturated during 1.5-s recycle delay (RD). ROESY spectra (τ_m = 250 ms) were collected using the pulse sequence of Bothner-by et al. (1984) and Bax and Davis (1985) with a spin-lock power of 4 kHz during the mixing period, and the carrier frequency was set at the HDO signal; the HDO signal was presaturated during the 1.5-s recycle delay.

Derivation of the Structure. The nature of base pairing in the hairpin is deduced from the imino proton profiles of the oligomer at different temperatures. A set of average inter-proton distances for pairwise interactions is obtained by performing full-matrix NOESY simulation with respect to the NOESY data at τ_m = 100 and 250 ms, in conjunction with stereochemical modeling, using linked-atom least-squares refinement [as described in Gupta et al. (1989)]. One energy-minimized structure, satisfying the inter-proton distance constraints, is derived. This structure is used as the starting configuration in a molecular dynamics simulated annealing (MDSA) calculation. In our calculations, inter-proton distance constraints (consistent with 2D NMR data) and Watson–Crick base-pairing constraints for the stem region are included as a harmonic potential, with a large spring constant. We choose a spring constant strength of 100 kcal/mol for Watson–Crick hydrogen bonding in the stem base pairs, 10.0 kcal/mol for NOE constraints with distances less than 3.5 Å, and 1.0 kcal/mol for NOE constraints with distances larger than 3.5 Å. No inter-proton distances or base pairings are forced in the loop region. This may yield structures with glycosidic angles in disagreement with the data (e.g.,

H1′, H2′ to H6 or H8 cross-peaks). In such cases, the incorrectly oriented glycosidic angles corresponding to a loop base are rotated by 180°. Then the new conformation is further energy-minimized. All calculations are done in vacuo, including all nonbonding pairs of interactions, and with a dielectric constant of 78.5 (García et al., 1990). All other energy terms are calculated using the *all-atom* force field of Weiner et al. (1986). The nine nucleotide systems contain 183 atoms. The initial system was heated and equilibrated to 500 K in a 10-ps constant-temperature molecular dynamics (MD) simulation. The resulting conformation at the end of this heatig period is then used as input for a 300-ps constant-energy MD simulation. We divide the configurational energy surface into different potential energy basins by taking conformations along the molecular dynamics trajectory (snapshots) and moving downhill from that point along the direction of the steepest descent of the potential energy (i.e., energy minimization). The result of this energy minimization is to subtract any vibrational motions from the displacements away from the local minimum. Minimization will relax the system to the bottom of the energy minimum sampled at the time of the snapshot. This is equivalent to an instantaneous temperature quenching (Stillinger & Weber, 1984). The whole sequence of energy minima describes transitions from one minimum to another. Strictly speaking, two structures are different if the distance between them, in the root mean square sense (McLachlan, 1979), is not zero. However, the bottom of a minimum well is reached asymptotically, and it will be practically impossible to find two structures to have identical conformations obtained at different times along the trajectory.

In order to distinguish local and global rearrangements of atoms among structures, we define a hierarchy of structures by progressively dividing the structures among clusters. The mean square distance between all pairs of structures is used for this purpose. At each subsequent level in the clustering, the two closest neighbors are clustered together into a new set in the hierarchy (Lebart et al., 1984). The distance between any two sets is taken as the smallest distance between any two structures belonging to the two different sets. This defines a single linkage distance (Lebart et al., 1984) where $d(i,j) \leq \min(d(i,k),d(j,k))$.

The resulting hierarchy for this oligomer will be described later in the text.

RESULTS

Characterization of Optimum Solution Conditions for Hairpin Stability. The gross morphology of a hairpin conformation is derived from the spectra of the imino protons. Parts B–E of Figure 1 show the spectra of the imino (NH) protons of d(CGCTGCGGC) in low DNA concentration (1.0 mM in strand) and in low salt (5 mM in NaCl, pH 7) at various temperatures. At 5 °C, we observe that five distinct NH signals appear in the spectrum (Figure 1B). We put them in three groups: (i) two N1-H signals (at 13.10 and 13.00 ppm) from two G·C pairs in the stem (Gupta et al., 1987), (ii) two NH signals from a G·T pair—one [N3-H of T] at 11.90 ppm and the other [N1-H of G] at 10.75 ppm, and (iii) one N1-H of G from the loop of the hairpin.

These spectra can be interpreted in terms of a duplex bulged out at the center or of a mismatched duplex and hairpin equilibrium. If the latter situations were true, the NH spectrum should be dependent on the DNA concentration. However, the spectrum in Figure 1A shows that when the DNA concentration is lowered to 0.5 mM in strand, a similar NH spectrum is observed. Therefore, under the solution

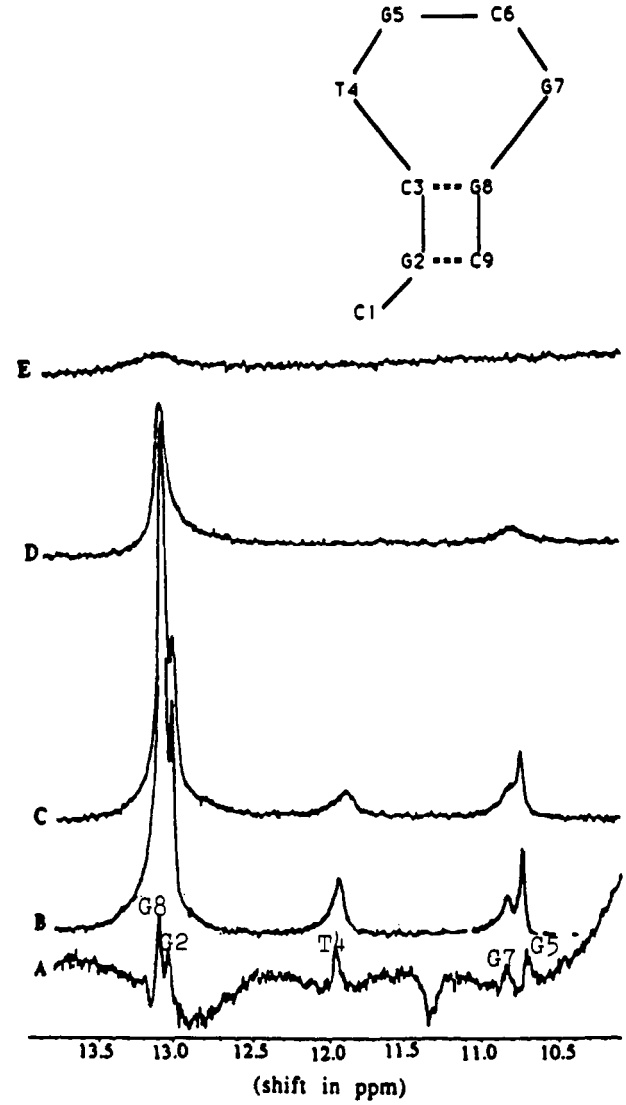

FIGURE 1: 500-MHz proton NMR spectra of the d(CGCTGCGGC) hairpin in 90% H_2O + 10% D_2O (10 mM phosphate buffer with 4 mM EDTA) recorded under various conditions using the pulse sequence of Sklenar and Bax (1987). The locations of different imino protons are marked. (A) DNA concentration, 0.5 mM in strand; salt concentration, 25 mM in NaCl; pH 7; temperature, 5 °C; number of scans, NS = 200; relaxation delay, RD = 1.2 s. (B) DNA concentration, 1.0 mM in strand; salt concentration, 25 mM in NaCl; pH 7; temperature, 5 °C; NS = 1000; RD = 1.2 s. (C–E) Conditions same as in (B) but at temperatures of 15, 30, and 40 °C.

conditions of Figure 1B, a monomeric hairpin is formed for d(CGCTGCGGC).

We observe three groups of NH signals that show different temperature-dependent solvent-exchange properties. Signals belonging to G·T pairs are the first to disappear at $T > 15$ °C. The signal from the G5 in the loop disappears above 30 °C, while the two NH signals from G·C pairs in the stem are the last ones to disappear at $T > 40$ °C. The nature of temperature dependence of the NH signals (especially those from the stem and G5) is typical of a hairpin structure (Gupta et al., 1987); i.e., NH protons of the bases in the (rigid) stem are less sensitive to temperature than those in the (flexible) loop. One interesting observation here is the dynamic equilibrium in which G·T pairs between T4 and G7 are formed and broken in this hairpin.

The NH spectra of d(CGCTGCGGC) are also shown for various temperatures (Figure 2A–G) under conditions of high salt (i.e., 1 M in NaCl, pH 7) while other solution conditions are identical to those in Figure 1B. Note that the signals

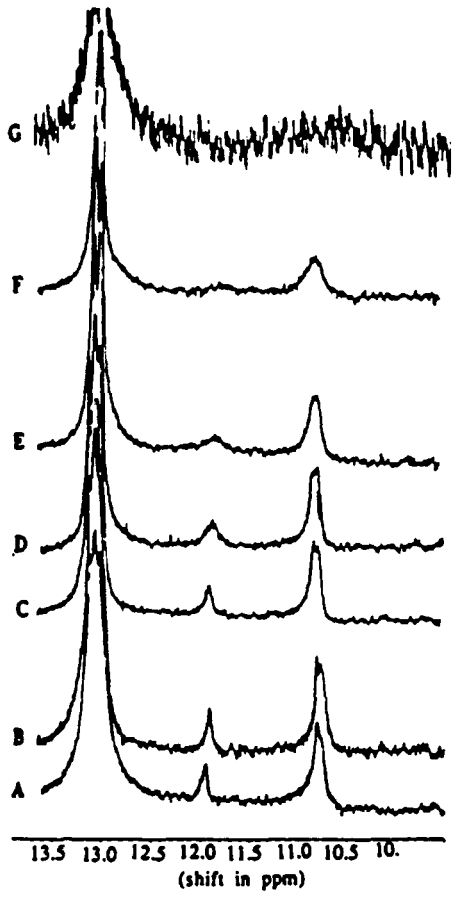

FIGURE 2: 500-MHz proton NMR spectra of the d(CGCTGCGGC) hairpin in 90% H_2O + 10% D_2O (10 mM phosphate buffer with 4 mM EDTA) recorded under various conditions using the pulse sequence of Sklenar et al. (1987): NS = 500; RD = 1.2 s. (A) DNA concentration 1.0 mM in strand, salt concentration 1.0 mM in NaCl, pH 7, temperature 5 °C. (B–G) Conditions same as (A) but at temperatures of 10, 15, 20, 25, 30, and 40 °C.

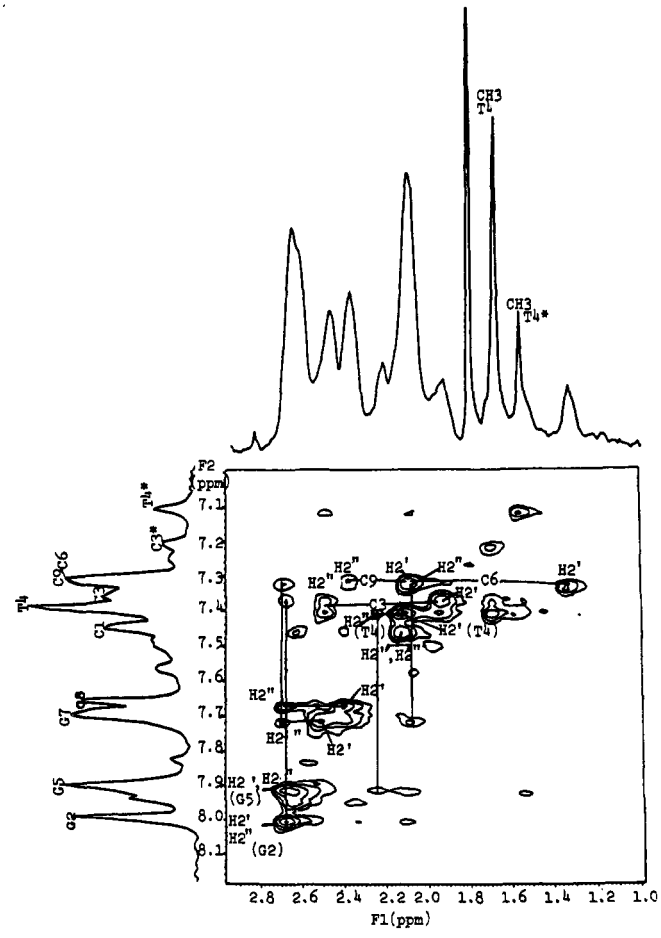

FIGURE 3: 500-MHz proton phase-sensitive NOESY (τ_m = 100 ms) cross section showing H8/H6–H2′ and H2″ connectivities in the d(CGCTGCTTC) hairpin: DNA concentration, 1.0 mM in strand; salt concentration, 25 mM in NaCl; pH 7; temperature 5 °C. [The 3.0–1.0 ppm range defines the region of H2′, H″, and CH3(T) protons, while the 8.4–7.0 ppm range defines the region of H8/H6 protons]. Signals in the base region that are not marked belong to a small population of coil present because these signals sharpen with increasing temperature. HDO is presaturated during recycle delay in all NOESY experiments. Note that, in addition to the dominant hairpin population, there is a minor population present that is marked by C3* and T4* (i.e., at least H6 of C3* and T4* separate out). There is a distinct difference in the NOE patterns of H6 of T4 (the major population) and H6 of T4* (minor populations): a strong NOE for H6–H2′(T4) and a weak NOE for H6–H2′(T4*). Cross-peaks parallel to the H2′, H2″ axis correspond to intra-nucleotide NOE connectivity, while the line joining the H2″ of two different residues defines the intra-nucleotide H2″(i)–H8/H8(i + 1) connectivities.

characteristic of a hairpin are still present. However, there is a noticeable increase of the area under the NH signals due to G·C pairs. This implies that a certain population of a duplex is stabilized under the solution conditions of Figure 4. An exclusive presence of a self-complementary mismatched duplex should be marked by the presence of three NH signals from three G·C pairs and one pair of NH signals from one G·T pair (note the equivalence of the NH signals around the center of symmetry in the duplex). The exclusive presence of the duplex also implies total absence of the NH signal of G5 near 10.7 ppm; i.e., the area under the signal at 10.7 ppm should be approximately the same as that under the NH signal at 11.9 ppm (as expected for a G·T pair). In an attempt to drive the equilibrium toward the mismatched duplex, we monitored the NH spectrum at higher DNA concentration (2.5 mM in strand) and high salt (1 M in NaCl)—a sizable population of the hairpin conformation was still present (data not shown). This is consistent with our previous work on the AC-mer (Sarma et al., 1987). The AC-mer, even at high salt (0.5 M in NaCl, pH 7) and high DNA concentrations (8 mM in strand) showed a finite population of a hairpin unless the temperature was dropped to 5 °C and pH to 4.5. Therefore, we conclude that a still higher DNA concentration is needed for the exclusive presence of a mismatched duplex for d(CGCTGCGGC). At the same time, from these studies (Figures 1 and 2) it is clear that a hairpin conformation of d(CGCTGCGGC) is present under a wide range of solution conditions; i.e., salt, DNA concentration, and temperature. Based on these studies we have chosen the following solution conditions for structural analysis of a hairpin for d(CGCTGCGGC): DNA concentration, 1.0 mM in strand, salt concentration, 25 mM in NaCl; pH 7; temperature 5–10 °C. Under these conditions, a monomeric hairpin is predominantly present for this oligomer.

Sequential Assignment of the Spin System H8/H6,H5/CH3,H1′,H2′,H2″,H3′, etc. NOESY experiments were performed at τ_m = 100 and 250 ms. Analyses of the NOESY data indicated that all residues exhibited an average C2′-endo, anticonformation. For such a hairpin conformation, a characteristic NOESY cross-connectivity is observed, i.e., H1″(i−1)–H8/H6(i)–H2′(i)–H2″(i)–H8/H6 and so on. One can make sequential assignment of H8/H6, H2′, H2″ protons from such a NOESY pattern (Gupta et al., 1987). Figure 3 shows a NOESY (τ_m = 100 ms) for the H8/H6 vs H2′,H2″ cross section; the intra- and inter-nucleotide connectivities are discussed in Figure 3. For a C2′-endo sugar, intra-nucleotide NOEs are expected for H1′–H2″ (d = 2.4 Å) and H2′ – H3′ (d = 2.7 Å) interactions. Figure 6 shows the

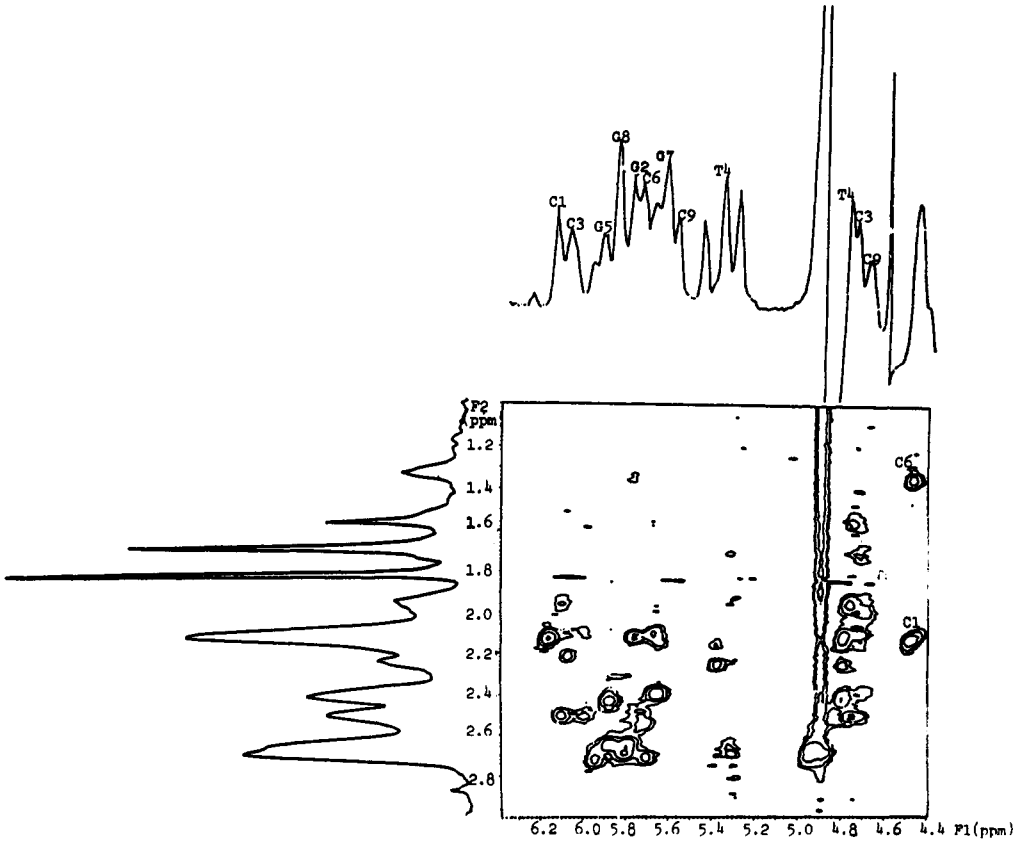

FIGURE 4: 500–MHz proton phase-sensitive NOESY (τ_m = 100 ms) cross section showing H1′, H3′–H2′, and H2″ cross-peaks in the d(CGCTGCTGGC) hairpin: DNA concentration, 1.0 mM in strand; salt concentration, 25 mM in NaCl; pH 7; temperature, 5 °C. The 6.5–5.2 ppm range defines the region of H1′ and H5(C), while the 5.1–4.3 ppm range defines the region of H3′.

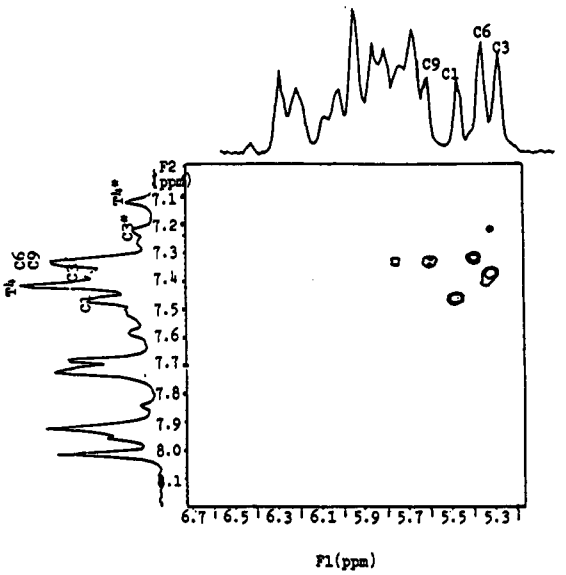

FIGURE 5: 500-MHz proton phase-sensitive NOESY (τ_m = 100 ms) cross section showing H6–H5(C) cross-peaks in the d(CGCTGCTGGC) hairpin: DNA concentration, 1.0 mM in strand; salt concentration, 25 mM in NaCl; pH 7; temperature, 5 °C. Also, note the presence of the inter-nucleotide H1′(G2)–H6(C3) cross-peak.

NOESY (τ_m = 100 ms) cross section for H1′, H3′ vs H2′, H2″. Combination of the data in Figures 3 and 4 allows the sequential assignment of H8/H6, CH3, H1′, H2′, H2″, and H3′. Figure 5 shows the H6 vs H5 cross-connectivities in C, which allows sequential assignment of H5 of C. Table I lists the chemical shift values of various protons of d(CGCTGCGGC) in the hairpin conformation.

Note that in Figure 5, no NOE cross-peaks (τ_m = 100 ms) are observed for H1′–H8/H6 interactions. Thus the H1′-(i)–H8/H6(i + 1) cross-connectivity pattern cannot be used to obtain the sequential assignment as done routinely for a B-DNA-like structure. Such a NOE pattern (i.e., absence of H1′–H8/H6 cross-peaks at τ_m = 100 ms) is a typical NOE fingerprint (and not an experimental limitation) of DNA hairpins with two G·C pairs in the stem and four bases in the loop (Gupta et al., 1987). However, this does not pose a major problem because inter-nucleotide H2″(i)–H8/H6(i + 1) cross-connectivity can be used for sequential assignment; i.e., first sequential assignment of the spin system H2′, H2″, H8/H6 is obtained from the H2′, H2″ vs H8/H6 cross section (Figure 3), and then the spin system H1′, H3′ is sequentially assigned in the H1′, H3′ vs H2″, H2″ cross section by monitoring intra-nucleotide H1′–H2′, H1′–H2″, H2′–H3′, and H2″–H3′ NOESY (τ_m = 100 ms) cross-peaks (Figure 4). In addition to the NOESY data, ROESY data (τ_m = 250 ms), which show *J*-coupled spin systems, are also used to reconfirm the assignment of H1′, H2′, H2″, H3′, and H6, H5. NOESY (τ_m = 100 and 250 ms) slices through H2′ are also used to assign H3′ because H2′–H3′ shows a strong cross-peak.

Even though intra- and inter-nucleotide cross-peaks are missing at τ_m = 100 ms (Figure 5), when τ_m is raised to 250 ms, the following intra- and inter-nucleotide weak cross-peaks show up in the H1′, H5 vs H8/H6 cross section (data not shown): H1′(G2)–H8(G2), H1′(G5)–H8(G5), H1′(G5)–H6-(C6), H1′(C1)–H6(C1), H1′(C6)–H6(C6), H1′(G7)–H8-(G7), H1′(G7)–H8(G8), H1′(G8)–H1′(G8), H1′(C9)–H6-(C9), H5(C6)–H8(G5), H5(C3)–G2(C8) [due to the partial overlap of H1′(C9) and H6(C9), the H5(C9)–H8(G8) cross-peak could not be unequivocally discerned]. Analysis of cross-peaks in the H1′, H5 vs H8/H6 cross section at τ_m = 250 ms also leads to the reconfirmation of the sequential assign given in Figures 3–6 and Table I.

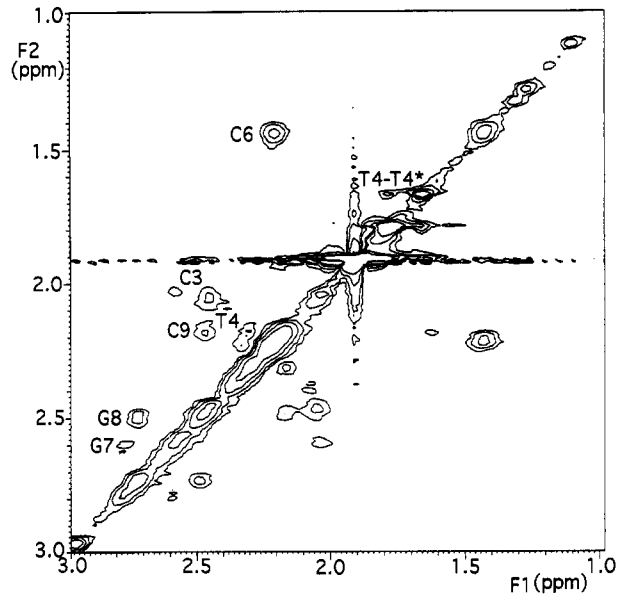

FIGURE 6: ROESY cross section (τ_m = 250 ms) showing H2′–H2″ and CH_3–CH_3 interactions. Solution conditions are same as Figures 3–5 except the temperature was raised to 10 °C to facilitate the transition involving major ⇌ minor conformation. The carrier offset was at the HDO peak. A spin-lock power of 4 kHz was used during the mixing phase. The HDO signal was presaturated during the 1.5-s recycle delay. Cross-peaks (without asterisks) involving the protons of the major form show negative intensities with respect to the diagonal. Cross-peaks involving the protons of different species, e.g., CH_3 of T4 and T4*, show positive intensities with respect to the diagonal. (The residues belonging to the minor species are marked by asterisks).

Table I: Chemical Shift Values (in ppm with Respect to DSS as an Internal Standard) of the Protons in the d(C1-G2-C3-T4-G5-C6-G7-G8-C9) Hairpin[a]

residue	H8/H8	CH3/H5	H1′	H2′	H2″	H3″
C1	7.52	5.47	6.18	2.14	2.14	4.71
G2	8.01		5.80	2.68	2.68	4.96
C3	7.36	5.32	6.11	1.96	2.51	4.77
C3*	(7.22)	5.32				
T4	7.40	1.70	5.40	2.12	2.24	4.80
T4*	(7.12)	(1.60)				
G5	7.93		5.95	2.66	2.66	4.80
C6	7.30	5.40	5.76	1.33	2.06	4.49
G7	7.73		5.71	2.52	2.68	4.89
G8	7.67		5.87	2.38	2.68	4.92
C9	7.31	5.60	5.68	2.08	2.37	4.80

[a] H4′, H5′, H5″ protons also assigned but not shown here because of their extensive overlap. Signals of C3* and T4* belonging to the minor conformer are shown within the parentheses.

It may be noted that even for d(ATCCTA*TTTT*TAGGAT) hairpin, which has a long stem of 6 base pairs, the intra-/interresidue NOEs involving H1′ and H8/H6 are either weak or totally absent even at τ_m = 300 ms (Blommers et al., 1989). Therefore, a lack of NOEs involving H1′ and H8/H6 for the nucleotides in the loop segment provides stuctural characteristics rather than a paucity of data due to experimental limitations. Williamson and Boxer (1989a,b) also reported similar observations for a different hairpin.

Analysis of the NOESY and ROESY Data. NOESY experiments were conducted for the d(CGCTGCGGC) hairpin for mixing times τ_m = 100 and 250 ms. The NOESY data at τ_m = 100 ms essentially reflected primary NOE, while primary as well as higher order NOEs were present in the NOESY spectrum at τ_m = 250 ms. The prominent NOE pattern involving H2″(i − 1)–H8/H6(i)–H2(i) (Figure 3) suggested C2′-endo, anti conformation of the constituent nucleotides. This is further confirmed by examining the H8/H6–H1′ cross section (Figure 5), where weak or no NOE is observed for H8/H6–H1′ interactions. The NOE pattern shown in Figures 3 and 5 is consistent with the C2′-endo, anti conformation of the nucleotides in which intra-nucleotide distance H8/H6(i)–H2′(i) ~ 2.2 Å and intra-nucleotide distance H8/H6(i)–H1′(i) ~ 3.8 Å. It may be noted that the NOESY pattern shown in Figure 5 unequivocally rules out the possibility of a syn conformation for any of the residues in the hairpin because for such a conformation a strong intra-nucleotide NOE for H8/H6(i)–H1′(i) (corresponding distance, ~2.2 Å) is expected.

Figure 3 also shows the presence of signals belonging to C3* and T4* of a minor conformer (residues in the minor conformation are marked *). The signals belonging to the remaining residues of the minor conformer are not resolved in this cross section. The NOESY cross-peaks of T4* indicate a different conformation for this residue in the minor conformer; note that the cross-peak H6(T4*)–H2′(T4*) is much weaker than the cross-peak H6(T4*)–CH_3(T4*) (fixed distance of 2.8 Å). This implies that the glycosyl torsion of T4* in the minor conformation is 190° ± 20° (low anti conformation) so that the distance H6(T4*)–H2′(T4*) is 4.2 ± 0.4 Å. However, the NOESY data corresponding to H6-(C3*)–H2′(C3*) interaction suggest a typical C2′-endo, anti confirmation for C3* in the minor conformation.

We conducted ROESY experiments (Bothner-by et al., 1984; Bax & Davis, 1985) for the d(CGCTGCGGC) hairpin at τ_m = 250 ms to verify the presence of two conformers. In the ROESY spectra, peaks due to *J*-coupling (as in COSY) and net magnetization transfer (as in NOESY) have peak heights of signs opposite to the diagonals, while peaks originating due to conformation exchange (for example, a major ⇌ minor conformer equilibrium, as in our case) have heights of the same sign as the diagonals. Figure 6 shows the diagonal H2′, H2″ cross sections for the d(CGCTGCGGC) hairpin. As indicated in Figure 6, the cross-peaks due to intraspecies interactions are negative with respect to the diagonal peaks, while cross-peaks due to interspecies equilibrium are positive with respect to the diagonal. The positive cross-peak CH_3(T4)–CH_3(T4*) suggests confirmation equilibrium between two conformations. The stereochemical features of these two conformations of the d(CGCTGCGGC) are discussed later in the text.

Extraction of a Set of Average Inter-Proton Distances: Use of Full-Matrix NOESY Simulations. The structural variables in the system are the dihedral angles for all the nucleotides, i.e., α(O3′–P–O5′–C5′), β(P–O5′–C5′–C4′), γ-(O5′–C5′–C4′–C3′), δ(C5′–C4′–C3′–O3′), ε(C4′–C3′–O3′–P), ζ(C3′–O3′–P–O5′), and χ(O1′–C1′–N1, N9–C2, or C4). All the bond lengths and bond angles are kept at standard values (Saenger, 1983) except the bond angles in the sugars, where small changes are made (when necessary) for ring closure. Two structural constraints are used for arriving at the starting structure of a hairpin: (i) H-bonding pattern in the stem as evident from the NH, NH2 profiles of the DNA (Figure 1) and (ii) C2′-endo, anti conformation for all the constituent nucleotides obtained from the primary NOE pattern (Figures 3 and 5). Constraint i translates into a set of distances due to H-bonding. For example, the distance d_a(N1-H–N3) in a G·C pair should correspond to an H-bond length of d_{hb} ~ 2.0 Å; in other words, a constraint of $G_j = d_a - d_{hb} = 0$ should be satisfied. Constraint ii implies that a subset of the dihedral angles should belong within a specified domain. For C2′-endo anti conformation, this means δ = 115° − 160°, χ = 230° − 270°.

Hairpin structures that satisfy constraints i and ii are obtained using a linked-atom least-squares refinement in the dihedral angle space (Gupta et al., 1989).

Several initial guesses are used for this least-squares refinement procedure, which lead to a set of refined structures that satisfy constraints i and ii. Nonbonded contracts are calculated for each refined structure, and only the structures that have allowed stereochemical contacts are chosen for the full-matrix NOESY simulations.

Stereochemically allowed hairpin structures of d(C1-G2-C3-T4-G5-C6-G7-G8-C9) in the chosen set show the following conformational features for the major conformers: (i) all nucleotides belong to C2′-endo, anti conformations as in B-DNA. (ii) β, γ, and ϵ are within the range of values found in B-DNA; i.e., $\beta = 180° \pm 30°$, $\gamma = 60° \pm 20°$, and $\epsilon = 200° \pm 30°$. (iii) All P–O torsions belong to g^-, g^- conformations as in B-DNA (Fratini et al., 1982), except that the set of P–O torsions connecting the loop and the stem on the 5′-side of the hairpin exhibited a g^-, t conformation.

All the stereochemically allowed models of the hairpin are then tested for their agreement with the NOESY data by performing full-matrix NOESY simulations and associated *R*-factor tests with respect to the observed data at $\tau_m = 100$ and 250 ms [methodology described in detail in Gupta et al. (1989) and Sarma et al. (1990)]. NOESY slices for $\tau_m = 100$ and 250 ms through H8/H6, H1′, H2′, H2″, H3′, CH3(T4) and H5(C) of various residues are used to obtain experimental NOESY intensities for pairwise interactions. Slices through H8/H6 show NOESY cross-peaks at sugar protons H1′, H2′, H2″, H3′, etc., and base protons CH3 and H5; while slices through sugar protons H1′, H2′, H2″, and H3′ show NOESY cross-peaks at other sugar protons in the same ring and intra- and inter-nucleotide sugar–base connectivities. Therefore, combination of NOESY data involving slices through H1′, H2′, H2″, and H3′ of a sugar and slices through H8/H6 of the base connected to the same sugar or to a neighboring one in sequence results in distance estimates of intra- and inter-nucleotide distances between base and sugar protons by using full-matrix NOESY simulation. For example, NOESY slices through sugar proton H1′ and base proton H8/H6 connected to the same sugar, are used to obtain estimates of the same intra-nucleotide H1′–H8/H6 distance; NOESY slices through sugar protons H2′ and H2″ and base protons H8/H6 connected to the same sugar are used to obtain the same intra-nucleotide H2′–H8/H6 (H2′–H2″ being fixed at 1.8 Å, and sp^3-linked H2″–H8/H6 distance is dependent on the corresponding H2′–H8/H6 distance). In summary, NOESY cross-peak intensity data are larger in size than the number of pairwise inter-proton distances derived from them because two or more cross-peak intensities result in the same inter-proton distance measurement.

After performing NOESY simulations, we arrive at a set of stereochemically allowed structures that satisfies the NOESY data for the major conformer. The ranges of inter-proton distances shown by the models in this set of various observed NOESY peaks, are as follows: H2′(C1)–H6(C1) = 2.4–2.8 Å, H2′(G2)–H8(G2) = 2.3–2.7 Å, H8(G2)–H5(C3) = 3.6–4.0 Å, H2′(C3)–H6(C3) = 2.3–2.7 Å, H2″(G2)–H6(C6) = 2.9–3.3 Å, H2′(T4)–H6(T4) = 2.5–3.2 Å, H2″(C3)–H6(T4) = 3.5–3.9 Å, H6(C3)–CH_3(T4) = 3.8–4.2 Å, H2′(G5)–H8(G5) = 2.3–2.7 Å, H2″(G5)–H6(C6) = 2.9–3.3 Å, H2′(C6)–H6(C6) = 2.3–2.7 Å, H8(G5)–H5(C6) = 3.8–4.1 Å, H2′(G7)–H8(G7) = 2.3–2.7 Å, H2″(G6)–H8(G7) = 2.6–3.2 Å, H2′(G8)–H8(G8) = 2.3–2.7 Å, H2″(G7)–H8(G8) = 2.5–3.0 Å, H2′(C9)–H6(C9) = 2.3–2.7 Å, H2″(G8)–H6(C9) = 2.4–2.8 Å, H1′(G8/G2)–H6(C9/C3) = 3.3–3.6 Å, and H8(G8)–H5(C5) = 3.5–4.0 Å. In addition, we have two other types of intra-nucleotide distances, H1′–H8/H6 = 3.7–4.1 Å and H3′–H8/H6 = 3.8–4.3 Å (typical for a C2′-endo, anti conformation). In all, 38 inter-proton distances are extracted as independent structural constraints from the NOESY data.

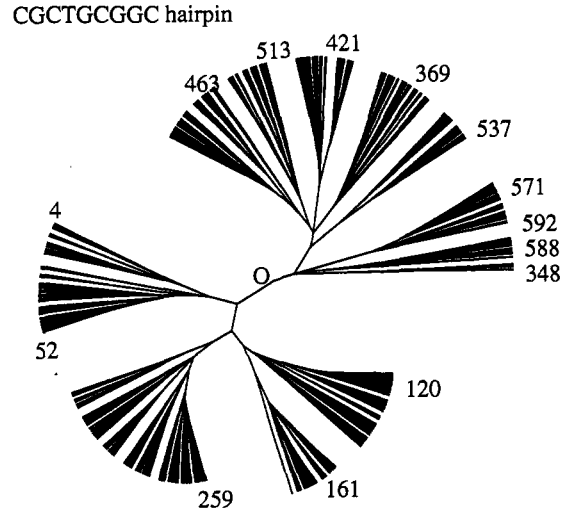

FIGURE 7: Radical tree representation of structures in different clusters.

For the minor conformer, the ranges of distances for various pairwise interactions remain similar to those in the major conformer except the distance H2′(T4*)–H6(T4*) = 3.9–4.3 Å, indicative of a low anti conformation of T4* ($\chi = 190° \pm 20°$) that probably facilitates the formation of a reverse wobble T4·G7 pair in the hairpin.

The estimates of the inter-proton distances derived from the full-matrix NOESY simulations are more reliable than those obtained from a two-spin model because experimental data hardly ever correspond to such an ideal situation. However, it may be noted that the NOESY simulations are carried out for a single correlation time for all pairwise interaction. In practice, it need not necessarily be so, especially for a hairpin where different nucleotides can show different mobilities. We addressed this situation in the following manner. We have interpreted the NOESY data in terms of an ensemble of structures (not a single one) such that different ranges of distances are obtained for different pairwise interactions. Thus, even when we keep the same correlation time for all pairwise interactions, the different proton pairs show different distance ranges, and therefore, differential mobility of different pairwise interactions are automatically incorporated in our analysis.

Mapping of the Hairpin Conformations. The single-stranded loop segment (T4-G5-C6-G7) of the d(C1-G2-C3-T4-G5-C6-G7-G8-C9) hairpin folds in such a way that G8 and C3 form a Watson–Crick pair and the G8·C3 pair stack on the top of the C9·G2 in the B-DNA geometry. It is obvious that even under the constraint of a B-DNA-like GpC stem, the loop segment can adopt a variety of conformations. We are interested in identifying distinct conformations that show different patterns of loop folding. We identify these conformationally distinct hairpin structures in the following manner: First, we isolate 600 structures from the 300-ps constrained MD trajectory (i.e., one structure after every 0.5 ps) and perform constrained energy minimization for 2500 steps or until the root mean square value of the energy first derivative is below 0.1 kcal/mol·Å. In this way we are able to obtain 600 local minima on the sampled potential energy surface of the hairpin. Potential energies of the 600 local

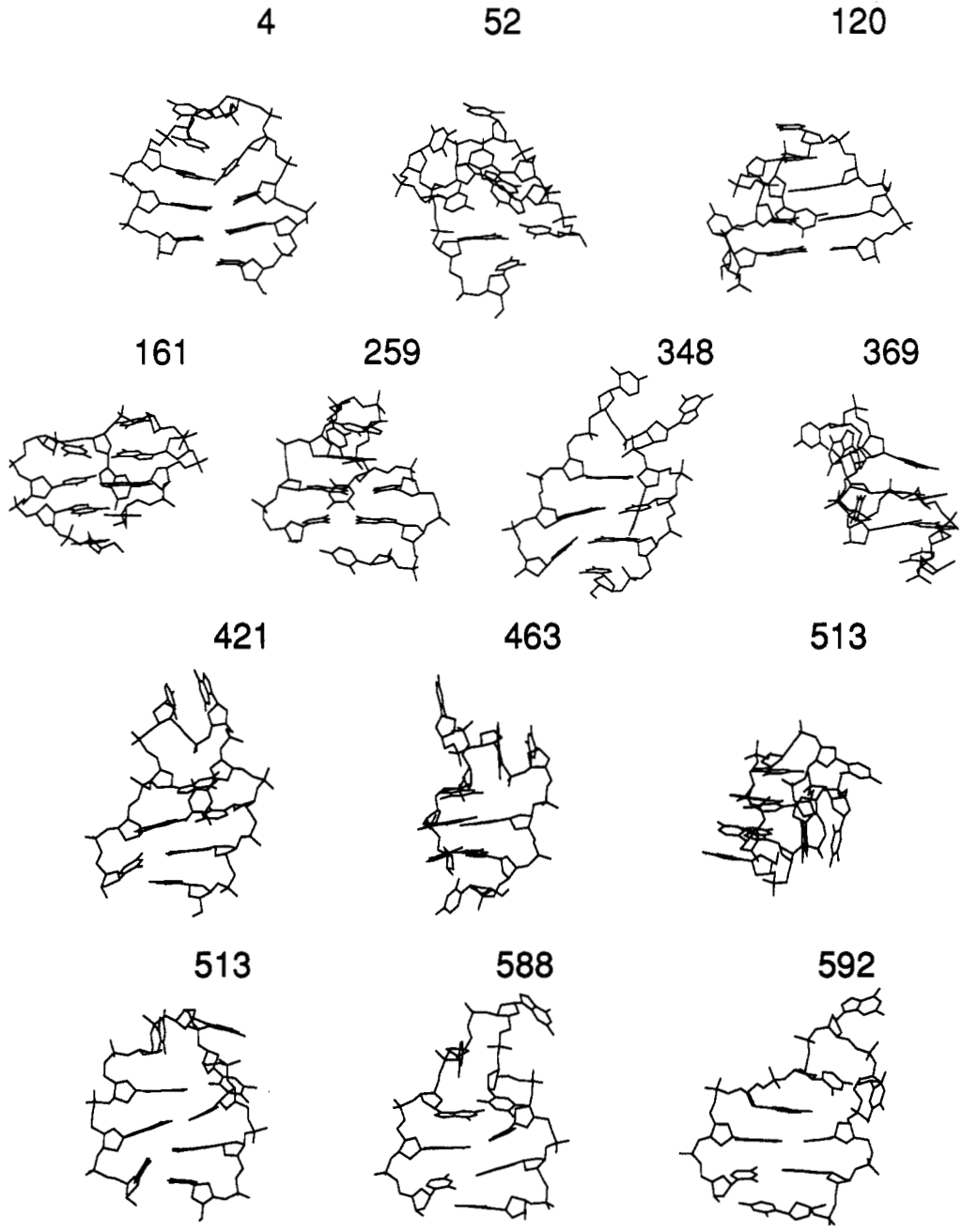

FIGURE 8: Plots of the conformationally distinct hairpin structures for the oligomer d(CGCTGCGGC) labeled in the tree in Figure 7. The numbering scheme of the structures is as follows: The structure number 1 refers to the minimized structure starting from the configuration at t = 0.5 ps on the MD trajectory, the structure number 2 refers to the minimized structure starting from the configuration at t = 1.0 ps on the MD trajectory, and similarly for the rest. The presence of distinctly different hairpin structures indicates the extent of the inherent flexibility of the loop region. (a) Seven structures representing clusters 4, 52, 120, 161, 259, 348, and 369; (b) six structures representing clusters 421, 463, 513, 588, and 592.

minima lie around 45 ± 20 kcal/mol of the 9-mer. Second, we construct a 600 × 600 matrix, where each element in the matrix gives an estimate of the structural difference between i and j. The maximum value of $(d_{ij}^{\text{rms}})_{\text{max}}$ is ~ 9.0 Å. Third, we construct a hierarchy in the manner described before. By choosing a level on the hierarchy that goes at least two levels of nodes below the largest cluster, we obtained 14 conformationally distinct clusters of hairpin structures that retain the same B-DNA geometry for the GpC stem but show different patterns of loop foldings. Below, we describe the approach we followed in partitioning a classification.

Figure 7 shows two graphical representations of the classification, an axial tree representation, Figure 7, and a hierarchical tree representation (figure provided as supplementary material). The axial representation is the best representation to see features such as branching and the number of structures in a cluster. The hierarchical representation is the best representation for determining the partitioning of the tree.

Figure 7 shows an axial representation of the classification constructed in the manner described before. Each point at which two lines meet is called a node. Notice that there is only one path from one node to another. A node represents a cluster, and the members of the cluster are all the structures that belong to the two nodes that joined to form the new one. In this plot, the minimum mean square distance between elements of the clusters represented by different nodes is shown as the length of the path that joins the two nodes. Notice that the center of this tree consists of one point, labeled O, representing structures, and the periphery has 600 clusters (i.e., 600 clusters with only one structure in each one). The numbers around this tree indicate the structure that best

represents the cluster (i.e., is closer to the average structure), with the exception of 4 and 161, which are the lowest energy structures for the cluster. The numbering of a structure refers to the point of time on the 300-ps MD trajectory at which the structure is extracted for energy minimization. For example, structure 1 is the local minimum obtained from the 0.5-ps MD snapshot, structure 2 refers to the local minimum obtained from the 1.0-ps MD snapshot, and so on. Notice that each cluster contains a different number of structures. The cluster represented by structure 259 has the largest number of structures, while the cluster represented by 348 has the smallest number of structures.

In the hierarchical representation, the mean square distance among clusters is represented along the vertical direction in the plot (the corresponding diagram included as supplementary material). Horizontal lines passing through the nodes are drawn to separate lines from different branches, but do not have any meaning regarding the distance among clusters. The partitioning of this hierarchy is done by drawing a horizontal line across the tree. Each intersection of this line with branches of the tree represents clusters containing structures that are as far as d_{cut} among themselves, and farther than d_{max} from structures belonging to other clusters. The partitioning cut has to be made above the lower levels of the tree where there are many clusters and below the upper levels of the tree where various groups of structures are clustered together. Starting from the top level of the tree, labeled O, there are two branches: one on the right-hand side (right) and another on the left-hand side (left) of the tree. The left branch shows a large number of nodes as the vertical level of partitioning is lowered, while the right branch shows only a few nodes. In order to obtain enough details about the clusters branching from the right branch of the tree, we partition the tree at a level that is at least two levels of nodes below the largest cluster.

The structures used as representatives of each cluster are shown in Figure 8a,b. The structural differences among structures in the right branch (structures 4, 52, 120, 161, and 259) and the left branch, described above, are mainly in the hairpin loop folding. All structures in the right branch show as a common feature a G7-G8-C9 stacking on the 3′-side and G2-C3-T4 stacking on the 5′-side and the conformations of the stem. The other structures show the G7-G8-G9 stacking, but T4 does not stack with the stem although T4-G5-C6 exhibit some degree of stacking among themselves. The structures belonging to the left branch of the tree show large exposure of the bases to what would be occupied by the solvent. The structures belonging to the right branch are expected to show favorable interaction with the solvent. Two members of this branch (4 and 161) will be described in detail later. The stem ($\langle^{G8\cdot C3}_{C9\cdot G2}\rangle$) interactions are constrained in our calculations both for MD simulation and energy minimization).

The average root mean square distance for structures within a cluster is 2.2 Å. The distance from a structure closest to the average structure for a cluster to another such structure in other clusters varies from 3.5 to 7.2 Å. The average root mean square distance for all 600 structures is 3.47 Å, with a maximum distance between a pair of structures of 9.0 Å. The structures chosen as the closest structure to the average structure for a cluster at this level in the hierarchy are, on the average, 1.5 Å away from the average structure for the set. This is, the average structure at higher levels on the hierarchy, where the sets represent global differences in structure changes of the hairpins, is not representative of a member of the group itself. This behavior is typical of systems that sample multiple potential minima basins. The structure closest to the average structure of the whole 600-structure set is structure 167 (83.5 ps), which is very close to structure 161 in Figure 7. This structure shows a loop folding with 2 bases in the loop and 3 base pairs in the stem.

Table II: Mean Square Displacements (Å^2) of Nucleotides

		$\langle x^2 \rangle$			
nucleo seq	phosphate	total	backbone[a]	phosphate[b]	base[c]
C1		18.7	13.9		23.8
	p1			10.7	
G2		4.5	4.6		2.3
	p2			6.1	
C3		5.5	6.9		
	p3			7.7	
T4		14.0	7.8		21.9
	p4			13.2	
G5		28.6	14.6		45.8
	p5			14.1	
C6		18.8	9.8		28.8
	p6			9.1	
G7		8.5	6.9		9.7
	p7			9.1	
G8		4.8	5.2		3.0
	p8			6.1	
C9		6.5	8.7		4.3
av[d]		12.1	8.7	9.5	15.2

[a] Backbone atoms are C1′, H1′, C2′, H2′, H2″, C3′, H3′, C4′, H4′, C5′, H5′, H5″, and O1′. [b] Phosphate atoms are O5′, P, OA, OB, and O3′. [c] Given that different bases have a different number of atoms, the average over all bases is not equal to the average values for the individual bases. [d] Average over the complete sequence.

Table II shows the average mean square displacements for all nine nucleotides in the hairpin. The average mean square displacements for the nucleotides range from largest to smallest in the order

$$G5 \gg C1 \geq C6 > T4 \gg G7 > C9 > C3 \geq G8 \geq G2$$

The most flexible base, G5, belongs to the loop region. This nucleotide samples many structures. The characteristic conformation is when G5 forms part of a stacked array (3′-stacked array) G5, C6, G7, G8, C9, shown in Figure 8, structure 4. Another characteristic conformation is when G5 forms favorable van der Waals contacts with the stem bases on what would have been the major groove of a longer B-DNA double-helical stem. In this conformation, G5 slides over the stem surface, thus adopting many compact structures. These conformations are shown in Figure 8, structure 161. The unpaired C1 base at the 5′-end of the molecule is the second most flexible base. This base easily alternates between stacked (with G2) and unstacked conformations. It is interesting to note that G2 is the least mobile base in the hairpin. It is commonly believed that unpaired bases at the 5′-end add stability (entropic) to oligonucleotides due to its larger flexibility. Here we observe a large flexibility of C1, as expected, but we also observed a reduced flexibility of G2.

The most flexible region of the whole molecule is the region consisting of P-G5-P-C6. The most rigid part of the molecule is the stem region.

The involvement of T4, G5, and C6 in loop folding is different in different structures. The class of structures represented by 1 shows an array of stacking involving (G5-C6-G7-G8-C9) on the 3′-side—a feature typical of hairpins with four residues in the loop (García et al., 1988). The class of structures represented by 161 shows an interesting pattern of loop folding in which T4 and G7 are involved in a reverse wobble pair and G5 goes out of stack. Features of this structure are discussed later in the text.

In view of the fact that no constraint was imposed for the residues in the loop segment, we are able to identify a wide

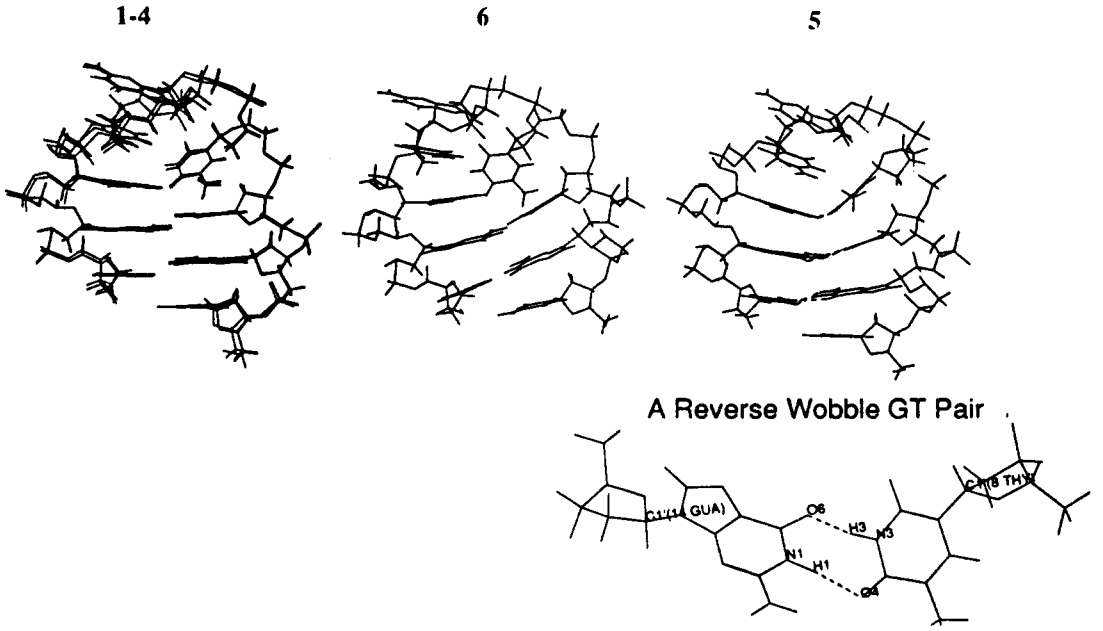

FIGURE 9: A set of representative hairpin structures in agreement with the NMR data of the oligomer d(CGCTGCGGC) under conditions of low salt and low DNA concentration. Structures 1–4 and 6 have four bases in the loop of the hairpin and two G·C pairs in the stems, while structure 5 has the T4·G7 pair in the beginning of the loop. The formation of a reverse wobble T4·G7 pair leads to a low anti conformation for χ of T4 (i.e., $\chi \sim 180^{\circ}$). However, note that all the structures 1–6 show favorable stacking on the 3′-side (i.e., for G5-C6-G7-G8-C9). Structure 5 has the following additional H-bonds involving G7 and T4: N1-H(G7)–O4(T4) = 2.0 Å; O6(G7)–H-N3(T4) = 1.95 Å.

variety of loop folding patterns. [The efficacy of this method to sample the allowed dihedral angle space is exhibited by the fact that the most of the dihedral angles sample all three staggered (i.e., g^+, t, g^-) conformations.] It may be pointed out that even though our method unequivocally demonstrates the inherent conformational flexibility in the loop segment of the hairpin, the structures from Figure 8 should be appropriately screened before they are tested for their agreement with the NMR data. All calculations reported here are done in vacuo, and consequently, as expected, some loop folding patterns in Figure 8 show one or more bases in the loop completely exposed to the solvent or phosphate groups partially buried in the core of the molecule—situations highly improbable under conditions of NMR experiments. Therefore, we chose structures that showed minimum exposure of the bases and maximum exposure of the phosphate groups to the solvent. For example, structures 1 and 161 are in the set of our chosen structures for NMR analyses. In order to obtain structures consistent with the NMR data, we went back to the MD snapshot (i.e., t = 0.5 ps) corresponding to structure 1, imposed the distance constraints for all nine residues in the hairpin derived from the NOESY data (τ_m = 100 and 250 ms), and performed a new energy minimization. In this way we arrived at a low-energy hairpin structure consistent with the NMR data. Figure 9 shows six representative hairpin structures; models 1–4 and 6 have four residues, T4, G5, C6, and G7, in the single-stranded loop segment (which constitutes the major form) while model 5 has a reverse wobble G·T pair between T4 and G7 with low anti χ for T4 (the minor conformer as indicated in Figure 3).

However, note that the hairpins in Figure 9 with or without T4·G7 reverse wobble pair show an array of stacking involving G5-C6-G7-G8-C9 and that what distinguishes models 1–4, 6, and 5 (with a T4·G7 pair) is a subtle difference in the nucleotide geomerty of the T4 residue.

Figure 10 shows the theoretical and observed NOESY slices for various base protons; theoretical slices are averaged over structures 1–4 and 5. Note the agreement between the theoretical and experimental spectra for the major conformer. We have actually obtained a library of several conformationally distinct structures that agree with the NMR data of the major conformer. In Figure 9, we merely include a select few to illustrate our methodology.

In addition to structure 5 in Figure 9, we also identify a set of structures clustered around model 161 in Figures 7 and 8. As shown in Figure 11, these 10 structures clustered around model 161 show the T4·G7 reverse wobble pair as in structure 5 in Figure 9 (i.e., T4·G7 pair formation is facilitated by low anti χ for T4). However G5, in all 10 models in Figure 11, does not show any stacking with the neighboring bases as does G5 in model 5 of Figure 9. G5 in all 10 models of Figure 11 dips into the major groove of the stem and forms two H-bonds involving N7(G5)–HN4(C3) and OB(C6)–HN2(G5). Comparison of the structures in Figures 9 and 11 indicates that, while only subtle changes are needed to go from a hairpin with four residues in the loop (as in structures 1–4 and 6 of Figure 9) to a hairpin with a T4·G7 pair (as in structure 5 of Figure 9), drastic conformational changes are needed to go from structures 1–4 and 6 of Figure 9 to the 10 structures of Figure 11. Therefore, it is conceivable that the structures of Figure 11 constitute the minor conformation in slow exchange with the major conformers (i.e., structures like 1–4 and 6 of Figure 9). Note that the structures in Figure 11 show a sharp turn at C6 similar to that suggested by Orbons et al. (1987) for the d(CGC*G*TGCG) hairpin, where G and T at the center form the single-stranded loop. Full-matrix NOESY simulations of structures corresponding to clusters 5 and 161 do satisfy the NOESY data corresponding to the minor conformation. Structural characteristics of the major (a hairpin with 4 bases in the loop) and the minor (a hairpin with 2 bases in the loop) are obtained from the NMR data (Figures 1–6). A slow major $\leftrightarrow$ minor equilibrium is also evident from the NMR data (Figure 6). However, we have not accounted for this equilibrium in our full-matrix NOESY simulations because our major interest is to characterize experimentally observable different hairpin structures for the same DNA sequence.

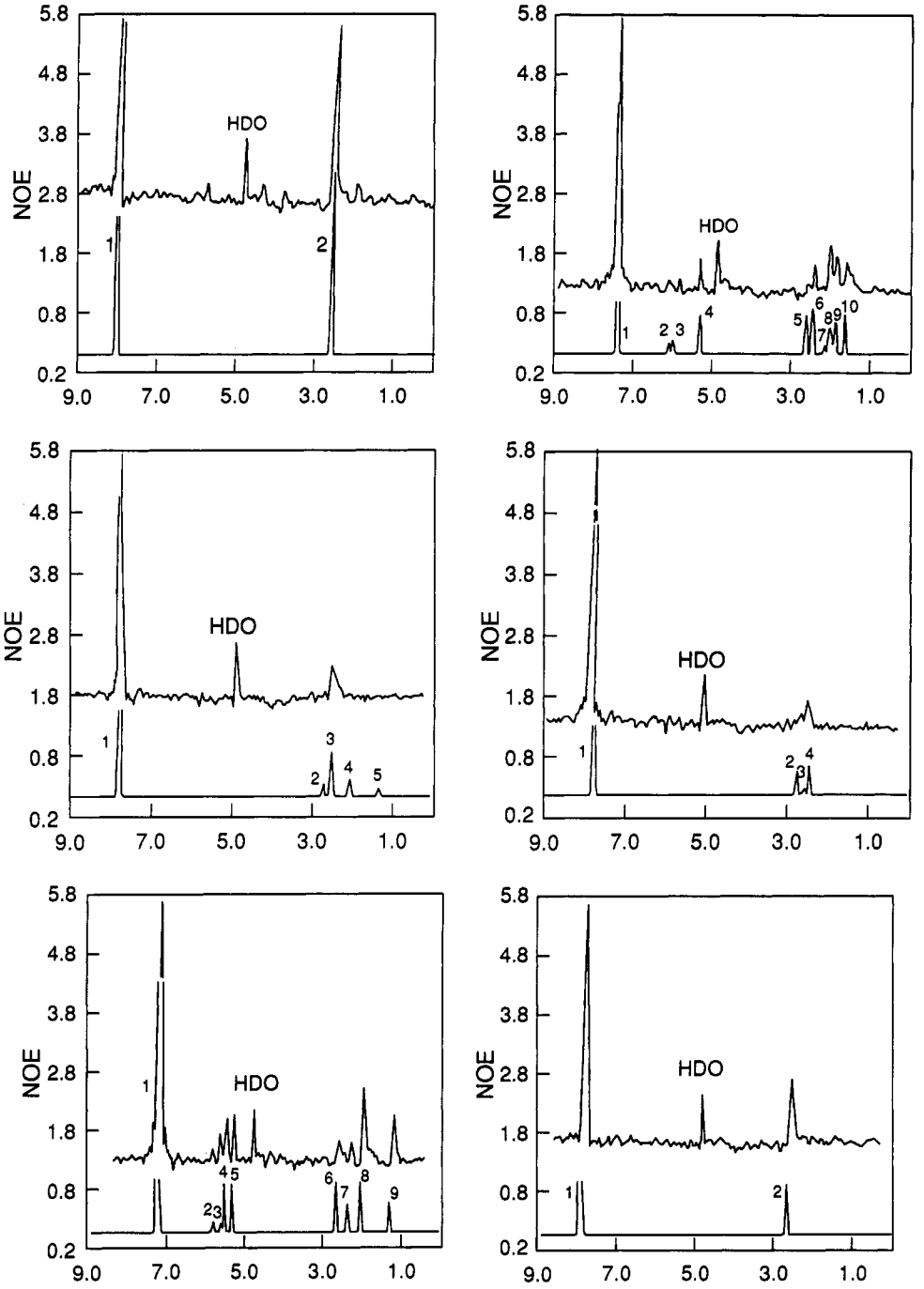

FIGURE 10: Theoretical and experimental NOESY slices for the base protons of the d(CGCTGCGGC) hairpin. Theoretical (average) NOESY slices are due to structures 1–4 and 6. The methodology of constructing theoretical NOESY slices is described previously in Sarma et al. (1990). (A, top left) Slices through H8 of G2: 1, H8 of G2; 2, H2′, H″ of G2. (B, top right) Slices through H6 of C3 and T4: 1, H6 of C3 and T6; 2, H1′ of C3; 3, H1′ of G2; 4, H5 of C3, 5, H2′ and H2″ of G2; 6, H2″ of C3; 7, H2″ of T4; 8, H2′ of T4; 9, H2′ of C3′; 10, CH_3 of T4. (C, middle left) Slices through H8 of G7: 1, H8 of G7; 2, H2″ of G7; 3, H2′ of G7; 4, H2″ of C6; 5, H2′(C6). (D, middle right) Slices through H8 or G8: 1, H8 of G8; 2, H2″ of G7 and G8; 3, H2′(G7); 4, H2′ of G8. (E, bottom left) Slices through H6 of C6 and C9: 1, H6 of C6 and C9; 2, H1′ of G8; 3, H1′ of G5 and C6; 4, H5 of C6; 5, H5 of C6; 6, H2′, H2″ of G5 and H2″ of G8; 7, H2′ of G8; 8, H2′ of C9 and H2″ of C6; 9, H2′ of C6. (F, bottom right) Slices through H8 of G5: 1, H8 of G5; 2, H2′ and H2″ of G5.

DISCUSSION

In this article, using a quantitative approach, we demonstrate the presence of multiple conformations of the d(C1-G2-C3-T4-G5-C6-G7-G8-C9) hairpin consistent with 2D NMR data. We have used molecular dynamics simulated annealing, followed by rapid temperature quenching, as well as cluster analysis, to study the multiple potential energy basins consistent with 2D NMR data. The MDSA and the temperature quenching techniques were used to extract a set of 600 structures. However, not all these structures show large changes in conformations, and therefore, it is not trivial to put every structure into different folding classifications. By using the root mean square distance between structures as a measurement of the difference between structures and the single linkage distance between clusters, we were able to construct an indexed hierarchy in which structures with distinct folding features are separated from others.

Our theoretical results are consistent with experimental results that indicate that the hairpin adopts two forms of hairpin under conditions of low salt and low DNA concentration: The major form has four residues in the single-stranded loop while the minor form has two residues in the single-stranded loop and a reverse wobble T4·G7 pair in the beginning of the loop. The major population shows stacking on the 3′-side

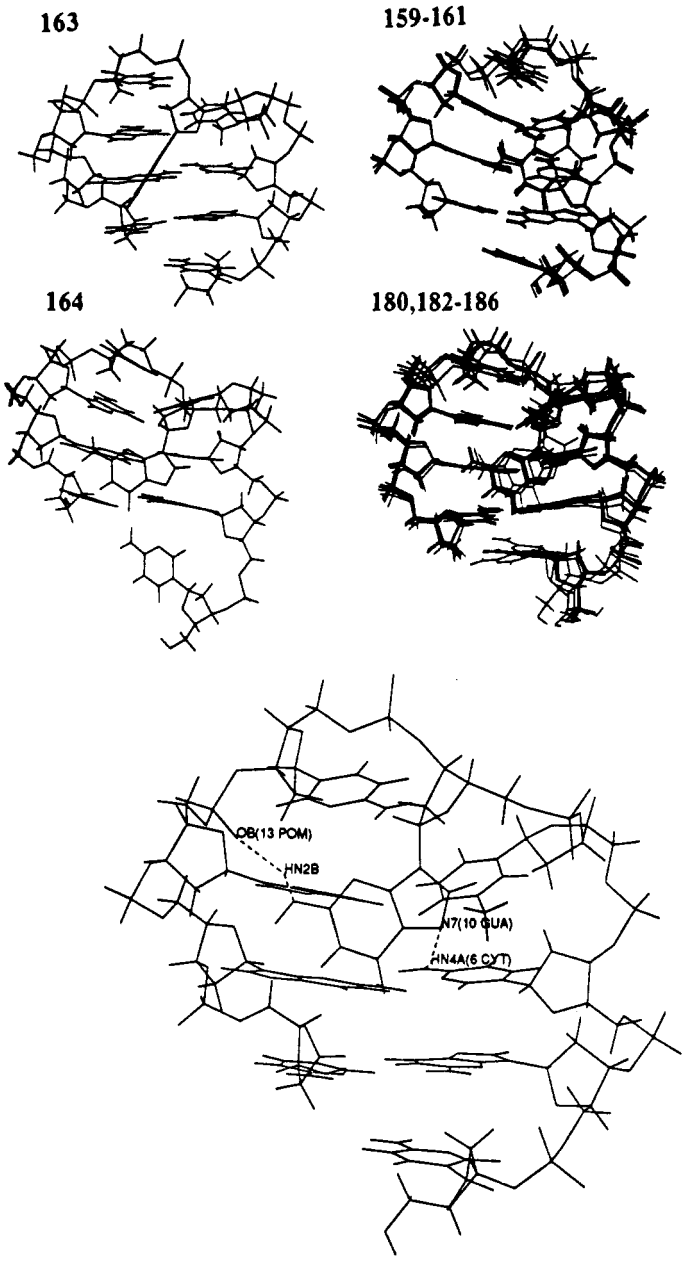

FIGURE 11: Another set of stable hairpin structures with the T4·G7 reverse wobble pair but with a different base-stacking arrangement in the loop (compare with structure 5 in Figure 7). In this set of structures, although G5 is destacked, it can make favorable electrostatic contacts with the G8·C3 pair in the stem and a phosphate oxygen in the loop (also shown). Note that the 10 structures are put in four different clusters. Following the numbering scheme in Figure 7, the members of the four clusters are indicated in the diagram. H-bond lengths and angles involving G5: OA(C6/G7)–H-N2(G5) = 2.12 Å with H-bond angle, N–H–OB = 145°, and H-N4(C3)–N7(G5) = 1.98 Å with H-bond angle N4–H–N7 = 150°.

(i.e., for the segment G5-C6-G7-G8-C9) while T4 is partially unstacked. Subtle conformational changes in structures 1–4 and 6 of Figure 9 lead to the formation of a T4·G7 reverse wobble pair in the minor conformer (structure 5 in Figure 9) in which we observe a low anti conformation of χ for T4 ($\chi \sim 180°$) in agreement with the NOE data of Figure 3 [see the NOEs due to H6(T4*)]. It may be pointed out that the T4·G7 pair is not an extension of the stem but only a conformational state of T4 and G7 in the loop in which they are paired. It was interesting to note that such a hairpin with the T4·G7 reverse wobble pair is also obtained in another way in which G5 (Figure 11) is destacked but assumes a compact shape while making favorable H-bonds with a phosphate in the loop segment and G8·C3 pair in the stem. In this class of structures (Figure 11), T4·G7 forms an extension of the stem region, and the single-stranded loop region has only two bases (Orbons et al., 1987).

In addition we have obtained information regarding the tendency of bases at different positions in the loop to show large (or small) atomic fluctuations about multiple equilibrium positions. These results are in agreement with previous results by Orborns et al. (1987), which indicated that the last base in the 3′-base in the loop stacks with the next base in the stem region. We cannot make any further generalizations about this stacking since it may be sequence-dependent.

The approach followed here to extract various structures consistent with 2D NMR data is not exhaustive in the sense that there may be many other structures that were not sampled during the MDSA computation. However, there are some indications that most of the configurational space in the neighborhood of the starting structure was sampled. For example, there are cases where structures separated by large time intervals belong to the same cluster. This is indicative of the recurrence of a given structural state. However, we did

not extend the simulation past 300 ps to see if the trajectories are bounded. At this point we should mention that the calculations done here have consumed hundreds of CPU hours on a Convex C220 computer, and extending this calculation for a longer periods of times does not seem appropriate. Although we have applied the MDSA, temperature quenching, and cluster analysis to determine a set of structures for a flexible DNA hairpin, this framework is equally applicable to studying the structural variability of double-helical DNA and globular proteins (García, 1992).

ACKNOWLEDGMENT

All computations were performed at the Los Alamos Advanced Computer Laboratory (ACL) facilities. The NMR work was done at the NMR facility of the Iowa State University, Ames, IA, under the biotechnology program. We are especially thankful to Drs. David Scott and Augustine Kintanar for their help and hospitality. We also thank the Nucleic Acid Facility of the Iowa State University for the supply of the purified oligomer. We thank Dr. G. Fichant and Dr. Y. Quentin for their many discussions concerning cluster analysis and the generation of trees.

SUPPLEMENTARY MATERIAL AVAILABLE

A truncated hierarchical tree representation of structures of different clusters (2 pages). Ordering information is given on any current masthead pages.

REFERENCES

Antosiewicz, J., German, M. W., Van der Sande, J. H., & Porshchke, D. (1988) *Biopolymers 27*, 1319.
Bax, A., & Davis, D. G. (1985) *J. Magn. Reson. 63*, 207.
Benight, A. S., Wang, Y., Amaratuna, M., Chattopadhyaya, R., Henderson, J., Hanlow, S., & Ikuta, S. (1989) *Biopolymers 28*, 3323.
Berns, K. I., & Bohernzky, R. A. (1987) *Adv. Virus Res. 32*, 243.
Blommers, M. J. J., Walters, J. A. L. I., Haasnoot, C. A. G., Aelen, J. M. A., van der Marel, G. A., van Boom, J. H., & Hilbers, C. W. (1989) *Biochemistry 28*, 7491.
Bothner-by, A. A., Stephens, R. L., Lee, J., Warren, C. D., & Jeanloz, R. W. (1984) *J. Am. Chem. Soc. 106*, 811.
Bouche, J. P., Rowen, L., & Kornberg, A. (1978) *J. Biol. Chem. 253*, 765.
Briat, J. F., Bollag, G., Kearney, C. A., Molineaux, I., & Chamberlin, M. J. (1987) *J. Mol. Biol. 198*, 43.
Chen, K. C., Tyson, J. J., Lederman, M., Stout, E. R., & Bates, R. C. (1989) *J. Mol. Biol. 208*, 283.
Cheong, C., Varani, G., & Tinoco, I., Jr. (1990) *Nature 346*, 680.
Fratini, A. V., Kopka, M. L., Drew, H. R., & Dickerson, R. E. (1982) *J. Biol. Chem. 257*, 14686.
García, A. E. (1992) *Phys. Rev. Lett. 68*, 2696.
García, A. E., Gupta, G., Sarma, M. H., & Sarma, R. H. (1988) *J. Biomol. Struct. Dyn. 6*, 525.
García, A. E., Gupta, G., Soumpasis, D. M., & Tung, C. S. (1990) *J. Biomol. Struct. Dyn. 8*, 173.
Gupta, G., Sarma, M. H., Sarma, R. H., Bald, R., Engelke, U., Oei, S. L., Gessner, R., & Erdmann, V. A. (1987) *Biochemistry 26*, 7715.
Gupta, G., Umemoto, K., Sarma, M. H., & Sarma, R. H. (1989) *Int. J. Quantum Chem., Quantum Biol. Symp. 16*, 17.
Hare, D., & Reid, B. R. (1986) *Biochemistry 25*, 5341.
Howard, B., Chen, C.-Q., Ross, P. D., & Miles, H. T. (1991) *Biochemistry 30*, 779.
Lebart, L., Morineau, A., & Warwick, K. M. (1984) *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*, John Wiley & Sons, New York.
Lilley, D. M. J. (1985) *Nucleic Acids Res. 13*, 1443.
Marky, L. A., Blumefeld, K. S., Kozlowski, S., & Breslauer, K. J. (1983) *Biopolymers 22*, 1247.
McLachlan, J. (1979) *J. Mol. Biol. 128*, 49.
Orbons, L. P. M., van der Marel, G. A., van Boom, J. H., & Altona, C. A. (1987) *J. Biomol. Struct. Dyn. 4*, 939.
Raghunathan, G., Jernigan, R. L., Miles, H. T., & Sasisekharan, V. (1991) *Biochemistry 30*, 782.
Rammal, R., Toulouse, G., & Virasoro, M. A. (1986) *Rev. Mod. Phys. 58*, 765.
Rentzeperis, D., Kharakoz, D. P., & Marky, L. A. (1991) *Biochemistry 30*, 6276.
Saenger, W. (1984) *Principles of Nucleic Acid Structures*, Springer-Verlag, New York.
Sarma, M. H., Gupta, G., Sarma, R. H., Bald, R., Engelke, U., Oei, S. L., Gessner, R., & Erdmann, V. A. (1987) *Biochemistry 26*, 7707.
Sarma, M. H., Gupta, G., García, A. E., Umemoto, K., & Sarma, R. H. (1990) *Biochemistry 29*, 4723.
Senior, M. A., Jones, R. A., & Breslauer, K. J. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 6242.
Sklenar, V., & Bax, A. (1987) *J. Magn. Reson. 74*, 469.
States, D. J., Haberkorn, R. H., & Ruben, D. J. (1982) *J. Magn. Reson. 48*, 286.
Stillinger, F. H., & Webber, T. A. (1984) *Science 225*, 983.
van de Ven, F. J. M., & Hilbers, C. W. (1988) *Biochemistry 30*, 3280.
Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986) *J. Comput. Chem. 7*, 230.
Williamson, J. R., & Boxer, S. G. (1989a) *Biochemistry 28*, 2819.
Williamson, J. R., & Boxer, S. G. (1989b) *Biochemistry 28*, 2831.
Xodo, L. E., Manzini, G., Quadrifoglio, F., van der Mavel, G. A., & van Boom, J. H. (1986) *Nucleic Acids Res. 14*, 5389.