Occurrence, solution structure and stability of DNA hairpins stabilized by a GA/CG helix unit

Peter Sandusky⁺, E. Wrenn Wooten, Alexander V. Kurochkin, Thomas Kavanaugh¹, Wlodek Mandecki¹ and Erik R. P. Zuiderweg^{*}

Biophysics Research Division, The University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109-1055, USA and ¹Molecular Diagnostics, Abbott Laboratories, Abbott Park, IL 60064, USA

Received June 12, 1995; Revised and Accepted October 14, 1995

ABSTRACT

The occurrence and NMR solution structure of a class of biloop hairpins containing the sequence 5'-CGXYAG are presented. These hairpins, which are variations on a sequence found in the reverse transcript of the human T-cell leukemia virus 2 (HLV2), show elevated melting points and high chemical stability toward denaturation by urea. Hairpins with the 5'-CGXYAG configuration have melting points 18–20° higher than hairpins with 5'-CAXYGG or 5'-GGXYAC configurations. The identities of the looping bases, X and Y above, play a negligible role in determining the stability of this DNA hairpin stability. This is very different from G-A based loops in RNA, where the third base must be a purine for high stability [the GNRA loops; V.P. Antao, S.Y. Lai and I. Tinoco, Jr (1991) *Nucleic Acids Res.*, 19, 5901–5905]. We show that these properties are associated with a four base helix unit that contains both a sheared GA base pair and a Watson–Crick CG base pair upon which it is stacked. As an understanding of the significance of AG base pairs has become increasingly important in the structural biology of nucleic acids, we compute an 0.7–0.9 Å precision ensemble of NMR solution structures using iterative relaxation matrix methods. Calculations performed on NMR-derived structures indicate that neither base–base electrostatic interactions, nor base–solvent dispersive interactions, are significant factors in determining the observed differences in hairpin stability. Thus the stability of the 5'-CGXYAG configuration would appear to derive from favorable base–base London/van der Waals interactions.

INTRODUCTION

Consideration of the occurrence and significance of non-Watson–Crick base pairs has become increasingly important to the study of nucleic-acid structural biology. AG mismatches specifically occur with great frequency. They have been identified as a component in one of the two common loop motifs found ubiquitously in ribosomal RNA hairpins (1,2). They are also found in DNA hairpins (3) and are essential components of the 5S

ribosomal internal 'E-loop' helix (4), the *T.thermophilus* tRNA_{ser} (5), the hammerhead ribozyme (6) and possibly the hairpin ribozyme (7). The wide occurrence of AG mismatches in RNA structures has led to the recent suggestion that they often play a role in the docking of single stranded sequences into the major groove of A-form helices (7) and several structural and thermodynamic studies of the GNRA tetra loop, where the second position (N) can be any base and the third position is a purine (R) have appeared (2,8–10). AG mismatches are also found in B-form duplex DNA (8–16), where they often increase helix stability (20). Recently Hirao and coworkers have described an unusually stable class of DNA hairpins containing the sequence 5'-CGAAG (3,21). An NMR-derived solution structure of this species showed a tight monolooop GAA turn with an AG mismatch and single bridging, base-stacked adenosine. It was postulated that the unusual stability of this species derived from a very regular B-form conformation of the helix stem. Recent experiments with λ phage infection of *Escherichia coli* have established the existence of tight DNA hairpin loops *in vivo* (22).

Here we report on the occurrence, stability and solution structure of a class of DNA hairpins derived from the sequence, 5'-C₃G₄X₅Y₆A₇G₈, where X and Y may be any base. These 'bilooop' species share many properties with the GAA 'monolooop' hairpin, including: the structure of the AG base pair, elevated melting points and conformational stability in the presence of 8 M urea. However, in contrast to what has been postulated for the monolooop species, we demonstrate that the origin of stability in these bilooop species derives largely from a very specific G4A7/C3G8 base stacking interaction. Comparison of the NMR-derived solution structure of the stable hairpin, TGC GXYA GCA, with a model of the less stable TGG GXYA CCA hairpin, indicates that more favorable London/van der Waals interactions are associated with the G4A7/C3G8 base stacking configuration.

MATERIALS AND METHODS

DNA synthesis and sample preparation

DNA oligomers were synthesized in μ mol quantities by the automated cyanoethyl phosphoramidite method. The sequences of oligomers synthesized are listed in Table 1. Sequences I, II and IV were synthesized at Abbott Laboratories. All other oligomers discussed in this report were synthesized at the University of

^{*} To whom correspondence should be addressed

⁺Present address: Chemistry Department, Clark University, Worcester, MA 01610, USA

Michigan Biomedical Core Facility. The water soluble fractions of the crude DNA oligomers were purified by two cycles of ethanol precipitation. The final ethanol pellets were resuspended into 0.5 ml of pH 6.5 20 mM phosphate buffer, which also contained 100 mM NaCl. When NMR spectra were to be taken in D₂O, the samples were lyophilized to dryness 4-fold and resuspended in 99.96% D₂O on each cycle. Unless otherwise indicated the single strand DNA oligomer concentration in the NMR samples was between 1 and 2 mM. When ³¹P spectra of a sample were to be taken in the presence of 8 M urea, 150 µl of the sample was diluted to 500 µl with pH 6.5 buffer containing 20 mM phosphate, 100 mM NaCl and 12 M urea.

Table 1. Oligonucleotide sequences and hairpin melting points

Sequence	T _m (°C)	Stability in 8 M urea
I <u>TGC</u> <u>GGCA</u> <u>GCA</u> ACAGC	75	Stable
II <u>TGC</u> <u>GGCA</u> <u>GCA</u> ACAAC	76	Stable
III <u>TGC</u> <u>AGCG</u> <u>GCA</u> AC AAC	58	
IV <u>TGC</u> <u>AGAA</u> <u>GCA</u> AC AGC	56	Denatured
V <u>TGC</u> <u>GGCA</u> <u>GCA</u>	76	Stable
VI <u>TGC</u> <u>GGAA</u> <u>GCA</u>	75	Stable
VII <u>TGC</u> <u>GAAA</u> <u>GCA</u>	75	Stable
VIII <u>TGC</u> <u>AGAA</u> <u>GCA</u>	57	Denatured
IX <u>TGC</u> <u>CGCA</u> <u>GCA</u>	63	Denatured
X <u>TGC</u> <u>GGAG</u> <u>GCA</u>	64	Denatured
XI <u>TGC</u> <u>AGAG</u> <u>GCA</u>	56	
XII <u>TGC</u> <u>AGCG</u> <u>GCA</u>	58	
XIII <u>TGG</u> <u>GGCA</u> <u>CCA</u>	57	
XIV <u>TGC</u> <u>G A A</u> <u>GCA</u>	74	

Helix stem bases are underlined and positions 4 and 7 are in bold case.

NMR spectroscopy

NMR experiments were performed on Bruker AMX 500 (500 MHz for ¹H, 202 MHz for ³¹P) and Bruker AMX 600 (600 MHz for ¹H) spectrometers. Unless otherwise indicated all NMR experiments were performed at 16°C. Assignments were based on standard presaturation COSY performed on samples in D₂O buffer, presaturation NOESY performed on samples in D₂O buffer, jump-return NOESY performed on samples in H₂O buffer and ¹H-³¹P hetero-COSY. Standard pulse sequences were employed with a 2.5 s relaxation time and NOESY mixing times of 400 ms.

For determination of NOE build-up curves in D₂O, solvent suppression by presaturation was omitted, the mixing times were randomly varied by ±15–20% to suppress zero quantum artifacts (23) and the relaxation delay was lengthened to 7 s (24). In this manner data were taken at 25, 65, 150 and 300 ms average mixing times. NMR data were processed on a Silicon Graphics INDY R4000 workstation using the FELIX software package (Hare Inc., Seattle, WA). ¹H spectra were referenced to H₂O at 4.9 p.p.m. (16°C).

Melting profiles

Melting profiles of pre-annealed and refolded DNA oligomers were constructed from 1-D proton NMR spectra taken at 5°C increments from 15 to 95°C. Samples were equilibrated for 5 min at each temperature. The reference chemical shift for H₂O was corrected by -0.1 p.p.m. for each 10°C increase in temperature. Chemical shifts for cytosine position 5 and thymine methyl protons were plotted versus temperature. These p.p.m. plots were then converted into plots of fraction linear strand (α) versus temperature by following the method described by Marky and Breslauer (25).

Melting enthalpies were estimated from the Van't Hoff equation:

$$\Delta H_m = 4R(T_m)^2(\partial\alpha/\partial T)_m$$

where R is the gas constant (kcal/mol-K), T_m is the hairpin melting point in degrees K and (∂α/∂T)_m is the slope of the fraction linear strand versus temperature plot evaluated at the melting point (25).

Structure generation and refinement

Structure generation and refinement were performed on a Silicon Graphics Indigo R3000 workstation and a 4D360TX computer using the Insight II, Discover and NMRchitect software packages licensed from Biosym Technologies (San Diego, CA). In general, distance geometry calculations were used to generate ensembles of 10 base hairpin structures using the DGII program within the Biosym NMRchitect module. Selected hairpin structures were then further refined using the iterative relaxation matrix analysis/restrained molecular dynamics (IRMA/RMD) method of Boelens and colleagues (26,27), as implemented in the Biosym NMRchitect module.

Distance geometry calculations (DGII). Approximate interproton distances were determined from initial NOE volume build-up rates using the isolated two-spin approximation (ITSA). The cytosine H5–H6 distance of 2.4 Å was used for calibration. These approximate distances were loosely weighted with ±0.5 Å uncertainties. In general, the NOE derived distances were supplemented with various combinations of base-paired hydrogen bonds, as described in detail below and were then used as distance restraints in the DGII calculations. Chiral restraints at the C1', C3' and C4' deoxyribose positions were routinely employed.

A linear 10 base B-form single stranded sequence, TGC GGCA GCA, was used as the source of the topology file for the DGII calculations. In order to detect internal inconsistencies, each given distance restraint set was subjected to sequential tetragonal smoothing. If the given restraint set passed this test, then subsequent calculations were pursued using triangular smoothing to approximate the Euclidian limits. The distances were then embedded in four dimensions and the resulting structures were optimized by the simplified DGII simulated annealing procedure with a 1500 cal initial energy and a 0.4 fs step size.

IRMA/RMD calculations. Cleanly integratable NOE peak build-up data were used as an input to further refine the DGII generated structures following the IRMA/RMD method. The motional model correlation time was based on simulations of cytosine H5–H6 NOE build-up curves as described below. The restrained molecular dynamics (RMD) leg of each RMA/RMD iteration was performed using the AMBER force field as implemented in

Figure 1. Model for the species II asymmetric hairpin. This basic structure applies to all 15 base species discussed in this report.

the Biosym Discover program 3.0. The calculations were performed without solvent molecules using a distance dependent dielectric function of the form $4\pi\epsilon(r)$ (r in Å). The phosphate charges were reduced to $-0.32e$ to approximate the effect of counterions (28). Both the RMA calculated and fixed distance restraints were weighted with 40 kcal/mol-Å^2 upper and lower bound force constants and a maximum force of 1000 kcal/m-Å . Each RMD leg consisted of an initial steepest descent/conjugate gradient minimization, followed by dynamics at 300 K with a 1 fs step size, followed by a final steepest descent/conjugate gradient minimization. 'R' values used were calculated as:

$$R = \frac{\left[\sum_{\tau_m} \sum_{NOE} \left\{ \tau_m \left(A_{ex}^{1/6} - A_{calc}^{1/6} \right) \right\}^2 \right]^{1/2}}{\sum_{\tau_m} \sum_{NOE} \tau_m A_{ex}^{1/6}}$$

with A_{ex} and A_{ca} the experimental and back-calculated NOE volumes respectively and τ_m the NOE mixing times.

Finite difference Poisson–Boltzmann calculations of electrostatic potentials

In order to estimate the magnitude of the electrostatic contribution to the base stacking energies for various nucleotide base configurations, finite difference Poisson–Boltzmann calculations were performed using the Biosym *DelPhi* module. In general the computational procedures described by Friedman and Honig were followed (29). Calculations were performed with an internal dielectric constant of 2 and external dielectric constant of 78.5, a 1.4 Å solvent probe radius and a 2.0 Å ion exclusion radius. AMBER partial charges and atom radii were used. Focusing grids were employed in order to achieve a resolution of $3.5 \text{ grid points per Å}$ in all directions. Calculations were performed at a variety of ionic strengths ranging from 0 to 0.22 M . When the electrostatic interaction between base pair WX and base pair YZ in the WX/YZ base stacked configuration was to be determined, an initial *DelPhi* calculation was performed in which all atoms in the WXYZ bases were assigned AMBER partial charges, q_i . A second *DelPhi* calculation was then performed in which the charges on the base atoms of the Y and Z nucleotides were set to zero and the resulting changes in the electrostatic potentials at

each W and X base atom i , $\Delta\phi_i$, were determined. The electrostatic component of the base stacking energy between the WX base pair and the YZ base pair was then calculated as: $E_{es} = \sum_i q_i \Delta\phi_i$, where the sum is taken over all atoms i in the W and X bases.

RESULTS AND DISCUSSION

Occurrence of urea stable DNA hairpins

In the course of a study at Abbott Laboratories on the human T-cell leukemia virus 2 (HLV2), it was observed that a 15 base fragment of the virus' reverse transcript (sequence I in Table 1) could not be used as a DNA sequencing primer and had the electrophoretic mobility of a nine base oligonucleotide on denaturing acrylamide gels with 8 M urea. Certain variants on the HLV2 sequence also electrophoresed short (e.g. sequence II), whereas other variants electrophoresed normally as 15 base oligonucleotides (e.g. sequence IV). All of these 15 base sequences electrophoresed as 15 base fragments in 90% formamide.

A selection of these 15 base oligonucleotides (sequences I, II and III) were examined by standard presaturation COSY, presaturation NOESY and jump-return NOESY experiments (30). In general the presaturation NOESY spectra, taken in 99.96% D_2O , showed sequential connectivities indicative of B-form base stacking running in two segments from position 15 to position 7 and from position 4 to position 1. Proton 1D jump-return spectra taken in 10% D_2O showed three hydrogen bonded imino-proton peaks between 12 and 14 p.p.m. . The two sharper, lower field, imino peaks had NOESY connectivities, via two sets of amino-protons, with the cytosine ring protons of C_3 and C_9 . Similar sequential NOESY connectivities and hydrogen-bonded imino-protons were observed in both urea stable and urea sensitive sequences. In general the proton NMR spectra of these oligonucleotides were largely unchanged by 10- and 100-fold dilutions, indicating that the secondary structures of these species are unimolecular. Thus, both urea stable and urea sensitive species form asymmetric hairpin loops with a helix stem composed of $T_1A_{10}/G_2C_9/C_3G_8$ and a five base dangle composed of positions 11–15 (Fig. 1).

Melting profiles and urea stability

In order to determine which sequence positions contributed to the unusual urea stability of the asymmetric hairpins, melting profiles were constructed for four 15 base oligonucleotide and nine 10 base oligonucleotide sequence variations. The results of these experiments are summarized in Table 1 and Figure 2. For a selected group of sequences, I, II, III, IV, V, X, XII and XIII, the concentration dependence of the melting profiles were also studied. In all these cases the melting profiles were invariant with concentration, as should be the case for the melting of unimolecular hairpins (25).

Examination of the melting properties of the 15 base sequences indicated that the short-running urea stable sequences had melting points of $73\text{--}75^\circ\text{C}$, $18\text{--}20^\circ\text{C}$ higher than those of the normal-running species. Further, comparison between the melting profiles of the 15 base and 10 base species indicates that the presence or composition of the 3' end dangle contributed nothing significant to the hairpin stability.

Figure 2. Melting profiles for 10 base variants of the species II hairpin sequence. Open symbols were used for the species with G₄A₇ base pairing. Note that species V, VI and VII all have a G₄A₇/C₃G₈ configuration, while species XIII has a G₄A₇/G₃C₈ configuration. Species XI has a A₄G₇/C₃G₈ configuration.

The unusual stability of the high melting-point hairpins derives from the sequence C₃G₄X₅Y₆A₇G₈, where X and Y may be any base. It is notable that the melting profiles of species with this sequence are essentially superimposable (species I, II, V, VI and VII). The substitution of bases at positions 5 and 6 has no effect on the melting profile, whereas substitutions at either G₄ or A₇ result in a decrease of 11–20°C in the melting point (species III, IV, VIII, IX, X, XI, XII). This indicates that the high melting species are in fact 'biloops' closed by an A₇ G₄ base pair. Switching the positions of A₇ and G₄, to give a G₇A₄ configuration, results in an 18–19°C decrease in the hairpin melting point (species III, XI and XII).

The absence of an effect on the melting profile when positions 5 and 6 were varied is interesting. This indicates that there are no stabilizing loop interactions involving the Y base in the GX₅Y₆A₇ loop similar to the cytosine to phosphate hydrogen bond that significantly stabilizes the UNCG RNA loop (32). More importantly, we do not see a melting-temperature effect on changing the Y position in our loops from pyrimidine (species I, II, V) to purine (species VI and VII). This is at variance with the observations made by Antao *et al.* for the RNA GNRA-class tetra loops, where purine at the Y position significantly enhances stability (8,9). In the RNA GNRA loops, it is found that the N7 of the R-base is involved in a hydrogen bond to the 2' OH of the G-base (10). This explains the difference between Tinoco's results on RNA GNRA loops and our results on DNA GX₅Y₆A₇ loops: without a 2' OH in the DNA species the hydrogen-bonding potential of the Y-base becomes irrelevant. We also note that the melting point of the monoloop species, XIV, is essentially the same as that of the analogous biloop species, V, VI and VII. This result agrees with the report by Hirao and coworkers who observed the same melting points for the GC GAA GC monoloop and GC GAAA GC biloop species (21). However, we observe that the slope of the α versus T plot is somewhat steeper for the monoloop species, XIV, than for the analogous biloop species V, VI and VII. This indicates that there are some additional stabilizing interactions in the monoloop species not found in the biloop species. Consider-

ing the superimposability of the melting curves for species V, VI and VII, it is clear that no significant base interaction occurs between the G₄A₇ base pair and the looping bases at positions 5 and 6. Thus it is reasonable to conclude that the stacking of the looping A with the base paired G, which Hirao observed in the monoloop species GC GAA GC (3), results in a slightly higher melting enthalpy relative to that of the biloop species.

Switching C₃ and G₈ results in a decrease of 18°C in the hairpin melting point as well (compare species V and XIII). This corresponds to a decrease in the melting enthalpy of 3 kcal/mol. Altering the G₂C₉/C₃G₈ interface to G₂C₉/G₃C₈ should result in the loss of only 0.1 kcal/mol in the base stacking enthalpy between these stem base pairs, assuming a B-form helix (31). It thus follows that the biloop stability must derive largely from the specific base interactions in the G₄A₇/C₃G₈ stem/loop unit. Such interactions may also explain the unusual stability of the CGATAG hairpin reported by Antao *et al.* (8), which as they note, does not conform to the GNRA consensus thought to be the necessary and sufficient basis for the stability of the G.A family of RNA tetraloops (8–10). We have not found reports from other workers on the effects of stem C.G to G.C interchange on the stability of the G.A family of RNA and DNA tetraloops, but note that the CUUCGG tetra loop is considerably more stable than the corresponding GUUCGC loop (9). This indicates that stem-loop basepair interaction is also important for the stability of the RNA UNCG loops (32).

In order to correlate the melting profile results with retention of secondary structure in 8 M urea, the ³¹P spectral signatures of selected species were examined in the presence and absence of urea. In all cases examined the secondary structures signatures ³¹P spectra of the high melting, G₄A₇/C₃G₈ species were stable in 8 M urea (sequences I, II, V, VI, VII), whereas the secondary structure ³¹P spectra of low melting, non-G₄A₇/C₃G₈ species were sensitive to urea (sequences IV, VIII, IX, X, XI) (see results in Table 1).

We conclude that the unusual stability of the 5'-C₃G₄X₅Y₆A₇G₈ loop derives from (i) the presence of the G₄A₇ mismatch basepair exclusively in an 5'-GX₅Y₆A₇ polarity and (ii) the interaction of this GA basepair with the stem pair C₃G₈ exclusively in the 5'-CGX₅Y₆A₇G₈ polarity. The sequence is strongly reminiscent of the also unusually stable GNRA loop-sequences abundantly found in RNA (7,35). The DNA variety of these loops differs from the RNA species in the important aspect that the third base is irrelevant for stability. This difference can directly be traced back to the presence/absence of a 2'OH group on the ribose moieties. The importance of loop-stem basepair interaction as we find in the DNA species has not been systematically investigated for RNA, but some isolated examples of such effects for those molecules have also appeared in the literature. We will show below that the stabilizing loop-stem basepair interaction in the DNA CGX₅Y₆A₇G₈ loop can be attributed to London-type stacking interactions.

Even though the formation of the stable GA/CG biloop derives from four instead of only two (3) specifically placed bases, its occurrence must be quite high in genomic DNA. Whether this has any relevance with respect to *in vivo* transcription is at present unknown, although hairpin formation effects have been observed in studies of viral replication (22). From a practical point of view it is clear that these sequences would be poor PCR primers and should be avoided for those purposes.

Figure 3. Standard DQF-COSY spectra for species V (the 10 base analogue of the species I and species II hairpins). The connectivity pattern indicates predominately 2' endo sugar conformations except at A10.

The structure of the G4A7/C3G8 hairpin stabilizing helix unit

A structure determination of the sequence II hairpin and its 10 base analogue, sequence V, was undertaken in order to understand the structural basis of the G4A7/C3G8 helix unit stability.

NMR data analysis. Assignments of the species II hairpin and species V hairpin [^1H] and ^{31}P spectra were undertaken using standard procedures. The assignments are available as supplementary material. Assignments of 2' and 2'' protons were based on the relative intensities of the $1' \times 2'$ and $1' \times 2''$ NOESY cross peaks. The assignments of the 10 base hairpin species (V) for positions from T₁ to C₉ mapped easily onto the assignments of its 15 base analogue (II), with only a slight parabolic ring shift effect centered at positions 5 and 6. (This shift pattern was maximal at positions 1 and 9, with shifts of roughly 0.10 and 0.15 p.p.m.).

The pattern of COSY connectivities with strong $J_{1'2'}$, $J_{1'2''}$ and $J_{2'3'}$ peaks and weak (absent) $J_{2'3''}$ and $J_{3'4'}$ peaks; and the magnitudes of $J_{1'2'}$ (7.5–11.5 Hz) and $J_{1'2''}$ (4.0–6.5 Hz), all indicate that the deoxyribose ring conformations are predominately 2' endo from position T1–C9 (Fig. 3). There is, however, evidence of minor 3' endo contributions to some of the ring conformations, for instance weak but detectable 3' × 4' COSY cross peaks, at positions 1, 3, 5, 6, 9 and 10.

Generation of hairpin structures by distance geometry: a question of the G₄A₇ mismatch hydrogen-bonding motif. NOESY build-up data were taken for the sequence II species at average mixing times of 25, 65, 150 and 300 ms. The cytosine H5–H6 cross peak build-up curves were linear up to 150 ms. A variation in correlation times was evident in the cytosine H5–H6 cross-peak data, with shorter correlation times at the loop and dangle positions, C₆, C₁₂ and C₁₅, than at the helix positions, C₃ and C₉.

Initial slope data were converted into 102 interproton distances for base positions T₁–A₁₀, using the C₃ and C₉ H5–H6 cross peak slopes for calibration.

The G4 to C3 B-form sequential NOESY connectivities define the G4 base conformation as *anti*. However three different G4A7 base pair hydrogen bonding motifs consistent with an *anti* G4 conformation are possible (10–19). These are shown in Figure 4 [geometries adapted from Saenger (33)]. Initial distance geometry calculations for the species II hairpin based on the NOESY derived distances and T₁G₂C₃ to A₁₀C₉G₈ Watson–Crick hydrogen bonding resulted in ensembles of B-form helix hairpins. These initial calculations however did not define the G4A7 base pair hydrogen bonding motif.

In order to determine which of the three possible G4A7 base pair motifs were compatible with the set of NOESY derived distances, three separate sets of DGII calculations were performed. In each set of calculations the basic distance restraint set was supplemented with hydrogen bond restraints taken from one of the three G4A7 mismatch motifs. Two of these three restraint sets, those supplemented with G4A7 hydrogen bonding motifs #2 and #3, readily gave high percentages of reasonable structures with good planar G4A7 base pairing and close G4A7/C3G8 base stacking. The 1# G4A7 mismatch motif failed to give any reasonable structures. We note that the G4A7 mismatch motif #1 has a A₇ N₉–G₄ N₉ distance of ~13 Å, whereas the canonical B-form GC and AT base pairs have N1–N9 distances of ~7 Å. Thus the #1 G4A7 mismatch motif appears to be simply too wide to fit inside the hairpin loop. Consequently inclusion of the #1 motif G4A7 hydrogen bonds in the restraint set resulted in DGII generated structures with radically domed G4A7 base pairing (A₇–G₄ angles of 90° or less). Attempts to add additional restraints in order to promote a planar type #1 base pairing

Figure 4. The three AG base-pairing hydrogen bond motifs found in DNA which are consistent with an anti G conformation [structures adapted from Saenger (33)]. See Katahira *et al.* (18) for a recent discussion of the literature on the occurrence of these structures

resulted in restraint sets which were inconsistent with the triangle inequality condition.

A distinction can readily be made between the ensemble of DGII structures based on the G4A7 #2 motif and that based on the #3 motif. Although both sets of structures show good near planar base pairing and good G4A7/C3G8 base stacking, the A7 H1'-C1'-N9-C8 dihedral angles in the #2 motif ensemble of DGII structures are rotated by roughly 180° relative to the usual purine B-form conformation. Significantly this places the A7 H8 base proton equidistant from the A7 H1', H2' and H2'' protons, at roughly 2.9 Å from each. However, the initial NOESY build-up rate for the A7 H8-A7 H2' cross-peak is >2-fold that of the A7 H8-A7 H2'' cross-peak and 8-fold that of the H8-H1' cross-peak (Fig. 5). Thus the A7 base to deoxyribose dihedral angle defined by the #2 G4A7 mismatch motif is inconsistent with the NOESY data and structures involving the #2 motif were eliminated from further consideration on this basis.

In contrast, the structures generated using the #3 G4A7 base pairing motif all have A7 base to deoxyribose dihedral angles close to that found for purines in canonical B-form helices and this places the A7 H8 roughly 2 Å from H2', 3 Å from H2'' and 3.8 Å from H1'. This is a geometry completely consistent with the NOESY build-up data. Back calculated NOESY volumes generated using typical motif #2 and #3 structures are compared for A7 cross peaks in Table 2. It is clear that only the #3 G4A7 base pairing motif is consistent with the NOESY data. This point was further verified by relaxation matrix calculations (see below). The #3 motif, which is sometimes called the 'sheared' AG base pairing, is also the base pairing motif found in the GNRA ribosome hairpins (2,8-10), the hammerhead ribozyme (6), the *T.thermophilus* tRNA_{ser} (5), some duplex B-form DNA sequences (16-19) and the DNA 'monolooop' hairpin (3).

Out of an ensemble of 20 structures generated using a restraint set supplemented with the #3 motif A7 to G4 hydrogen bond

Figure 5. NOESY build-up spectra of species II taken at 600 MHz in 99.96 % D₂O. Numbers given refer to average mixing times. Features associated with A7 conformation are indicative of an anti conformation.

Figure 6. Ensemble of 12 DGII generated/IRMA/RMA refined structures ($R = 0.03$ – 0.06) for the species V stem helix. For purposes of clarity the looping positions 5 and 6, which were not structurally defined, are not shown.

restraints, three structures were discarded because of poor helix structure at the $A_{10}T_1$ base pair. A further five structures were discarded because of doming of the G_4A_7 base pair, which prevented good G_4A_7/C_3G_8 base stacking. The remaining 12 structures, all of which showed well defined helices, near planar G_4A_7 base pairing and good G_4A_7/C_3G_8 base stacking, were further refined using the IRMA/RMA method.

Table 2. NOESY Crosspeak intensities for position A_7

Atom 1	Atom 2	65 ms			150 ms		
		Motif #2	Motif #3	Exp	Motif #2	Motif #3	Exp
H8	H2''	0.159	0.126	0.150	0.325	0.297	0.291
H8	H2'	0.154	0.337	0.328	0.297	0.597	0.536
H8	H1'	0.147	0.030	0.060	0.278	0.070	0.100

Back calculated intensities based on typical structures calculated using motif #2 and motif #3 A_7G_4 hydrogen bonding restraints: $\tau_c = 1.5$ ns.

IRMA/RMD calculations. Build-up curves for the cytosine H5–H6 NOESY cross-peaks were simulated in order to determine the motional model correlation time (τ_c). It was evident that, due to variation in the secondary structure, the model effective τ_c s varied at different points in the asymmetric loop structure. While the build-up curves associated with the helix cytosines, C_3 and C_9 , modeled with a τ_c of 1.5 ns, the build-up curves for the loop and dangle cytosines, C_6 , C_{12} and C_{15} , were clearly governed by shorter τ_c s. Because of this variation in dynamics, the IRMA refinement of the DGII generated structures was limited to the core of the helix, including the AG mismatch pair. The NOESY volume inputs consequently consisted of data from positions G_2 , C_3 , G_4 , A_7 , G_8 and C_9 , plus inter base connectivities between G_2 and T_1 , A_7 and G_5 and A_{10} and C_9 . NOESY volume data for mixing times 65, 150 and 300 ms from 42 cleanly integratable peaks were supplemented with Watson–Crick base pair hydrogen bond restraints between $T_1G_2C_3$ and $A_{10}C_9G_8$, but no hydrogen-bonding restraints were used for the A.G mismatch pair.

In general the calculations converged structurally in three IRMA/RMD cycles, so that the RMSD values of the bases C_3 , G_4 ,

A_7 and G_8 were 0.7–0.9 Å for superposition of the second and third structures. The final 'R' values from these calculations were in the range of 0.03–0.06, indicating that the distances calculated from the simulated NOESY volumes would agree with distances calculated from experimental NOESY volumes to within a few percent. An ensemble of 12 DGII generated/IRMA/RMD refined structures for the hairpin of species V is shown in Figure 6. We emphasize that each structure in this ensemble predicts the NMR spectrum equally well. The G_2C_9/C_3G_8 helix unit of any of these structures can be superimposed on a canonical B-form helix with an RMSD of 0.7–1 Å. The G_4A_7/C_3G_8 base stacking conformation found in these structures is shown in Figure 7.

Structural heterogeneity of tetraloops lacking G_4A_7 mismatch base pairing. The COSY spectra of species containing the G_4A_7/C_3G_8 helix unit, for example sequences II, V and VI, show well defined, intense $1' \times 2'$, $1' \times 2''$ and $2' \times 3'$ cross peaks for every deoxyribose present (Fig. 3). In contrast the COSY spectra of species lacking the G_4A_7 mismatch base pairing (sequences III, VIII and XII) show well defined, intense deoxyribose cross peaks only for the helix positions T_1 , G_2 , C_3 , G_8 and C_9 . This indicates that the tetra-loop positions 4, 5, 6 and 7 do not assume a single well defined conformation in those species lacking the G_4A_7 mismatch.

The structure of the G_4A_7/C_3G_8 helix unit and the basis of its stability. According to literature, the stability of RNA GNRA loops is governed by (i) the sheared-A.G mismatch base pairing, (ii) a specific hydrogen bond between the amino protons of the A-base and the phosphate backbone of the G residue, (iii) a specific hydrogen bond between the purine N7 and the 2'OH of the G nucleotide and (iv) the 'stacking' of the N, R and A bases (10). Which of these interactions might be of importance for the enhanced stability of the CGXYAG DNA loop? The hydrogen bonding interactions (i) and (ii) can certainly also contribute in DNA if they can in RNA, but the specific G-Y hydrogen bonding interaction (iii) cannot exist. Base stacking within the loop is also not expected to be a major contribution to the stability for the DNA loop: we observe a shorter correlation time for the base on the Y position suggesting that it is pointing towards solution. On the other hand, we have observed that changing CGXYAG to

Figure 7. Base stacking of the G4A7 base pair (yellow) on the C3G8 base pair (red) in the species V hairpin. Adapted from the ensemble shown in Figure 6.

GGXYAC does lower the melting temperature by 19°C in our DNA tetra loop, strongly implicating the importance of the interactions between the G.A mismatch pair and the stem Watson–Crick pair. We present in the following estimations of the differences of electrostatic interaction, solvent interface exposure (water–nucleotide London/Van Der Waals and ‘hydrophobic’ interaction) and base–base London/Van Der Waals between the bases of the experimental CGXYAG structure and of a hypothetical structure for the GGXYAC sequence. These estimations suggest that base–base London/Van Der Waals is the most important term in the interaction between the G.A mismatch and stem Watson–Crick base pairs.

The melting enthalpy associated with the G4A7/C3G8 configuration is 3–4 kcal/mol higher than that of the G4A7/G3C8 configuration. In order to consider the structural implications of the G4A7/G3C8 base stacking configuration a set of models for the TGG GGCA CCA sequence (species XIII) were constructed. Starting with the ensemble of NMR-derived structures for species V, several structures were chosen and G3C8 base pairs substituted for the C3G8 base pairs. The species XIII models were then refined by

Figure 8. Models for the hypothetical species XIII G4A7/G3C8 structure. G4A7 base pairs in yellow and G3C8 base pairs in red. Note poor G3–G4 base stacking.

restrained molecular dynamics using the same RMD schedule employed in the IRMA/RMD refinement of the species V structures. The restraints used consisted of hydrogen bond distance restraints between T₁G₂G₃G₄ and A₁₀C₉C₈A₇ and dihedral angle restraints used to maintain the sugar puckers in the 2′ endo configuration. To a large degree the conformation of the G4A7/G3C8 base stacking was unaffected by the RMD refinement. Typical G4A7/G3C8 model structures are shown in Figure 8.

It is important to note, when considering the significance of the species XIII models, that the actual species XIII structure almost certainly does not incorporate a sheared #3 motif G4A7 base pair. The H₂O exchangeable features of the various G4A7/C3G8 species all include a peak near 10 p.p.m. (10.9 p.p.m. in species V). Such peaks, usually assigned to the mismatch G imino proton, are signatures for a sheared #3 AG base pair (3,16–19). No such peak is observed in the spectrum of species XIII. In the case of duplex DNA, which of the four possible AG mismatch motifs will actually occur is largely determined by the identity of the adjacent base pairs (13,15,16,20). Clearly a similar phenomenon occurs in the hairpins. The actual structure of the species XIII hairpin may

incorporate an G4A7 base pair with a base-pairing motif other than #3, or the species XIII structure may not have a well defined G4A7 base pairing at all. Thus, in generating structures for species XIII which incorporate a sheared G4A7 base pair, we are modeling a structure which probably does not occur in order to consider why such a hypothetical species XIII structure is unstable relative to the closely related and very stable, species V structure.

Empirical values for base stacking enthalpies and entropies in duplex B-form DNA have been tabulated by Breslauer and coworkers (31). The base stacking enthalpies are strongly sequence dependent and vary from 5.6 kcal/mol for GC/AT to 11.9 kcal/mol for CG/GC. Friedman and Honig have published a computational analysis which identifies three factors which could contribute to sequence specific variations in base stacking enthalpies: variations in the electrostatic interactions between the stacked bases, variations in the London/van der Waals (Lennard-Jones potential) interactions between the stacked bases and variations in the base-solvent London interactions derived from variations in the base-solvent interface (29). It should be noted that the change in heat capacity (C_p) associated with the melting of DNA is small (34). This indicates that the structuring of water molecules around unstacked portions of the base faces, an effect which would be analogous to the hydrophobic effect which dominates the thermodynamics of soluble protein folding, is a negligible factor in the case of nucleic acids (35).

In order to assess the magnitude of the electrostatic contributions to the base stacking interactions in the G4A7/C3G8 and G4A7/G3C8 helix units, *Delphi* calculations were performed on several species V and species XIII structures. The range of values for electrostatic base stacking interactions was -0.1-0.5 kcal/mol for the G4A7/C3G8 structures and 0.2-0.5 kcal/mol for the hypothetical G4A7/G3C8 model structures. Thus the electrostatic interactions in the G4A7/C3G8 configuration might be either slightly stabilizing or slightly destabilizing relative to those in the hypothetical G4A7/G3C8 configuration. Either way, the magnitude of the electrostatic interactions indicates that they are too small to be the determining factor stabilizing the species V configuration relative to the species XIII configuration.

Careful comparison of the Connolly surfaces for the species V and species XIII structures indicates that the base-solvent interface is 10-12 Å² larger in the G4A7/G3C8 structures. Accepting the value of 59 kcal/mol-Å² for base-solvent dispersive interactions given by Friedman and Honig (29), the change in the base-solvent surface would then correspond to a 0.6 kcal/mol stabilization of the hypothetical species XIII configuration relative to the species V configuration, in contrast with the observations.

Having ruled out significant contributions from base-base electrostatic and base-solvent London interactions, we postulate significantly more favorable base-base London/van der Waals interactions (Lennard-Jones potential) in the G4A7/C3G8 configuration relative to the G4A7/G3C8 configuration. Simple inspection of the two structures shows a tighter base stacking in the G4A7/C3G8 configuration (Figs 7 and 8). The G₃ and G₄ bases of the hypothetical G4A7/G3C8 structure appear particularly poorly base stacked. In order to account for the differences in the species V and species XIII melting curves, a difference of at least 3 kcal/mol in the base-base dispersive and steric interactions must be postulated. Considering that the range of values for Lennard-Jones interactions calculated by Friedman and Honig is 2.5 kcal/mol for single base to base steps in single stranded B-form helices (29), a value of 4 or 5 kcal/mol as a

difference in the Lennard-Jones interactions between two double helix structures does not seem unreasonable.

ACKNOWLEDGEMENTS

We thank Mr Jose Aparicio for pointing out to us the unusual properties of oligonucleotide I and Dr Hong Wang for very useful discussions concerning structure calculation methods. This work was partially supported by NIH grant G1252406. The National Science Foundation, The National Institutes of Health and Parke-Davis Co. are acknowledged for support of the NMR instrumentation.

REFERENCES

- 1 Woese, C.R., Winkler, S. and Gutell, R.R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 8467-8471.
- 2 Heus, H. and Pardi, A. (1991) *Science*, **253**, 191-194.
- 3 Yoshizawa, S., Ueda, T., Ishido, V., Miura, K., Watanabe, K. and Hirao, I. (1994) *Nucleic Acids Res.*, **22**, 2217-2221.
- 4 Wimberly, B., Varani, G. and Tinoco, I. (1993) *Biochemistry*, **32**, 1078-1087.
- 5 Biou, V., Yaremchuk, A., Tukalo, M. and Cusack, S. (1994) *Science*, **263**, 1404-1410.
- 6 Pley, H.W., Flaherty, K.M. and McKay, D.B. (1994) *Nature*, **372**, 68-74.
- 7 Wimberly, B. (1994) *Nature Struct. Biol.*, **1**, 820-827.
- 8 Antao, V.P., Lai, S.Y. and Tinoco, I. (1991) *Nucleic Acids Res.*, **19**, 5901-5905.
- 9 Antao, V.P. and Tinoco, I. (1992) *Nucleic Acids Res.*, **20**, 819-824.
- 10 Jucker, F.M. and Pardi, A. (1995) *RNA*, **1**, 219-222.
- 11 Brown, T., Hunter, W.N., Kneale, G. and Kennard, O. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 2402-2406.
- 12 Prive, V.I., Heienmann, U., Chandrasegaren, S., Kan, L.S., Kopta, M.L. and Dickerson, R.E. (1987) *Science*, **238**, 498-504.
- 13 Gao, X. and Patel, D.J. (1988) *J. Am. Chem. Soc.*, **110**, 5178-5182.
- 14 Brown, T., Leonard, G.A., Booth, E.D. and Chambers, J. (1989) *J. Mol. Biol.*, **207**, 455-457.
- 15 Carbonnaux, C., van der Martel, G.A., van Boom, J.H., Guschlbauer, W. and Fazakerley, G.V. (1991) *Biochemistry*, **30**, 5449-5458.
- 16 Cheng, J.W., Chou, S.H. and Reid, B.R. (1992) *J. Mol. Biol.*, **228**, 1037-1041.
- 17 Katahira, M., Sato, H., Mishima, K., Uesugi, S. and Fujii, S. (1993) *Nucleic Acids Res.*, **21**, 5418-5424.
- 18 Katahira, M., Kanagawa, M., Sato, H., Uesugi, S. and Fujii, S., Kohno, T. and Maeda, T. (1994) *Nucleic Acids Res.*, **22**, 2752-2759.
- 19 Green, K.L., Jones, R.L., Li, Y., Robinson, H., Wang, A.H.J., Zon, G. and Wilson, W.D. (1994) *Biochemistry*, **33**, 1053-1062.
- 20 Ebel, S., Lane, A.N. and Brown, T. (1992) *Biochemistry*, **31**, 12 083-12 086.
- 21 Hirao, I., Kawai, G., Yoshizawa, S., Nishimura, Y., Ishido, Y., Watanabe, K. and Miura, K. (1994) *Nucleic Acids Res.*, **22**, 576-582.
- 22 Davison, A. and Leach, R.F. (1994) *Nucleic Acids Res.*, **22**, 4361-4363.
- 23 Macura, S., Huang, Y., Suter, D. and Ernst, R.R. (1981) *J. Magn. Reson.*, **43**, 259.
- 24 Wang, H., Zuiderweg, E.R.P. and Glick, G.D. (1995) *J. Am. Chem. Soc.*, **117**, 2981-2991.
- 25 Marky, L.A. and Breslauer, K. (1987) *Biopolymers*, **26**, 1601-1620.
- 26 Boelens, R., Koning, T.M.G., van der Martel, G.A., van Boom, J.H. and Kaptein, R. (1989) *J. Magn. Reson.*, **82**, 290-308.
- 27 Boelens, R., Koning, T.M.G. and Kaptein, R. (1988) *J. Mol. Struct.*, **173**, 299-311.
- 28 Tidor, B., Irikur, K.K., Brooks, B.R. and Karplus, M. (1983) *J. Biomol. Struct. Dyn.*, **1**, 231-252.
- 29 Friedman, R.A. and Honig, B. (1992) *Biopolymers*, **32**, 145-159.
- 30 Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.
- 31 Breslauer, K., Frank, R., Blocker, H. and Marky, L.A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3746-3750.
- 32 Cheong, C., Varani, G. and Tinoco, I. (1990) *Nature*, **346**, 680-682.
- 33 Saenger, W. (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag, New York.
- 34 Vesnaver, G. and Breslauer, K. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 3569-3573.
- 35 Searle, W. and Williams, D.H. (1993) *Nucleic Acids Res.*, **21**, 2051-2056.