Selection for Thermodynamically Stable DNA Tetraloops Using Temperature Gradient Gel Electrophoresis Reveals Four Motifs: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc)†

Mariko Nakano,‡ Ellen M. Moody, Jing Liang,§ and Philip C. Bevilacqua*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received July 18, 2002; Revised Manuscript Received September 28, 2002

ABSTRACT: Hairpins play important roles in the function of DNA, forming cruciforms and affecting processes such as replication and recombination. Temperature gradient gel electrophoresis (TGGE) and in vitro selection have been used to isolate thermodynamically stable DNA hairpins from a six-nucleotide random library. The TGGE−selection process was optimized such that known stable DNA tetraloops were recovered, and the selection appears to be exhaustive. In the selection, four families of exceptionally stable DNA loops were identified: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc). (Lowercase denotes the closing base pair; N = A, C, G, or T; and B = C, G, or T.) It appears that the known stable d(cGNAg) triloop motif can be embedded into a tetraloop, with the extra nucleotide inserted into either the middle of the loop, d(cGNNAg), or at the 3′-end of the loop, d(cGNABg). For d(cGNNAg) and d(cGNABg), a CG closing base pair was strongly preferred over a GC, with $\Delta\Delta G^\circ_{37} \approx 2$ kcal/mol. Members of the two families, d(cCNNGg) and d(gCNNGc), are similar in stability. The loop sequences and closing base pairs identified for exceptionally stable DNA tetraloops show many similarities to those known for exceptionally stable RNA tetraloops. These data provide an expanded set of thermodynamic rules for the formation of tetraloops in DNA.

The hairpin is the most common secondary structural element in RNA. Hairpins play roles in initiating RNA folding and forming tertiary structure and protein binding sites (*1*, *2*). Hairpins are also important for the function of DNA. In ssDNA, hairpins are involved in regulating replication and transcription (*1*, *3*), and the loop composition can affect binding of antibiotics and antitumor agents (*4*). In dsDNA, hairpins can form transiently as cruciforms (*5*) or double-hairpin elements (DHE) (*6*). Hairpins formed from dsDNA have been implicated in numerous biological processes including telomere replication, deletion mutations, and V(D)J recombination (*7*−*9*). Structural and functional knowledge about thermodynamically stable DNA hairpin loops may therefore lead to a deeper understanding of biological processes.

Recent advances have been made by SantaLucia and coworkers on the prediction of DNA secondary structure from sequence (*10*). DNA structure prediction is valuable in primer design for PCR[1] and DNA microarrays and for DNA therapeutics and molecular beacons. Thermodynamic parameters are available for Watson−Crick base pairs and select secondary structure defects of DNA, including single internal mismatches and dangling ends (*10*−*17*). However, there is a paucity of thermodynamic data for other DNA motifs, including hairpin loops.

The structure and stability of hairpins have been most extensively studied in RNA, where they are the most common. Phylogenetic studies indicated that certain loop-closing base pair combinations are unusually common in RNA, including r(cUNCGg), r(GNRA), and r(gCUUGc) (*18*). Thermodynamic studies showed that these loops are exceptionally stable and that the closing base pair makes an important contribution to the stability of r(cUNCGg) sequences (*19*). The structures of numerous RNA hairpin loops have been solved by NMR and X-ray methods and have revealed a structural basis for stability in terms of hydrogen bonding, stacking, and shape complementarity (*20*−*22*). A nearest-neighbor model for RNA hairpins has been developed on the basis of extensive experimentation by Serra, Turner,

† Supported by National Science Foundation CAREER Grant MCB-9984129 and a Camille Dreyfus Teacher-Scholar Award and Sloan Fellowship to P.C.B.

* To whom correspondence should be addressed. Phone: (814) 863-3812. Fax: (814) 863-8403. E-mail: pcb@chem.psu.edu.

‡ Current address: Department of Chemistry, Faculty of Science and Engineering, Konan University, 8-9-1 Okamoto Higashinada-ku, Kobe 658-8501, Japan.

§ Current address: Department of Biophysics and Biophysical Chemistry, Johns Hopkins University, Baltimore, MD 21218.

[1] Abbreviations: B = C, G, or T; BSA, bovine serum albumin; CD, circular dichroism spectroscopy; C_T, total strand concentration; EDTA, ethylenediaminetetraacetic acid; I, inosine; minihairpins, shorter DNA oligonucleotides used for biophysical study with the general sequence 5′-d(ggaX$L_1L_2L_3L_4$X′tcc), where X and X′ are complementary nucleotides forming the closing base pair; L_n, loop position *n*; N = A, C, G, or T; NTP, nucleoside triphosphate; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; $P_{10}E_{0.1}$, 10 mM sodium phosphate and 0.1 mM Na_2EDTA (pH 7.0); R = A or G; SSB, single-stranded DNA binding protein; $TBEK_{50}$, 100 mM Tris, 83 mM boric acid, 1 mM Na_2EDTA, and 50 mM KCl (pH 8.8); TE, 10 mM Tris and 1 mM Na_2EDTA (pH 7.5); TGGE, temperature gradient gel electrophoresis; T_M, melting temperature; Tris, tris(hydroxymethyl)-aminomethane; UV, ultraviolet spectroscopy.

10.1021/bi026479k CCC: $22.00
Published on Web 11/05/2002

and co-workers that successfully predicts the stability of many RNA hairpin loops (*23*, *24*).

DNA hairpins have received less attention than RNA hairpins. It has been known for some time that the stability of DNA hairpins is affected by the sequence of the loop and closing base pair. Early studies showed that the stability of hairpins with homogeneous tetraloop sequences is dependent on the loop sequence and closing base pair (*25*, *26*). A study of 28 sequences containing homogeneous and heterogeneous tetraloop sequences also demonstrated a dependence of hairpin stability on sequence (*27*). Studies by Antao and Tinoco showed that certain DNA stem−loops are exceptionally stable; for example, d(gCTTGc) is significantly more stable than d(gTTCGc), with a $\Delta\Delta G^\circ_{37}$ of 2.5 kcal/mol, and d(cGAAAg) is considerably more stable than d(cAAAAg), with a $\Delta\Delta G^\circ_{37}$ of 1.6 kcal/mol (*19*). Also, they demonstrated that the DNA motif, d(cTTCGg), unlike its RNA counterpart, is not exceptionally stable. Hirao and co-workers showed that d(cGNNAg) comprises a stable tetraloop family (*28*); this loop contains a sheared GA pair (*29*) that is qualitatively similar to the r(GNRA) tetraloop in RNA (*21*). In addition, the CG closing base pair makes an important contribution to stability of the d(cGNNAg) motif, with $\Delta T_M \approx 20$ °C relative to a GC closing base pair (*29*). DNA triloops with the motif d(cGNAg) have a sheared GA base pair and are especially stable (*30*, *31*). Related structures with sheared GA base pairs form in the d(GCA) triloop with a TA closing base pair, which may be important in triplet-repeat diseases and human centromeres (*32*, *33*), and in the d(GTTA) tetraloop with an AT closing base pair (*34*).

Despite these advances, many outstanding questions remain for DNA loops: What are the families of exceptionally stable DNA tetraloops? Do closing base pairs play important roles in such hairpins? What are the similarities and differences between stable RNA and DNA tetraloops? Answers to these questions may lead to a better appreciation of the biological importance of DNA hairpins and to a deeper understanding of the thermodynamic and structural principles for nucleic acid hairpin stability.

We have been developing a combinatorial approach to studying the thermodynamic stability of structural motifs in nucleic acids. This involves a combination of temperature gradient gel electrophoresis (TGGE) and in vitro selection. TGGE has been shown to be capable of resolving RNA and DNA molecules with small differences in free energy (*35*−*38*). In particular, model stem loops with a $\Delta\Delta G^\circ_{37}$ of only 0.3−0.5 kcal/mol, including AU to UA transversions, were shown to be resolvable by TGGE (*38*). TGGE−selections were used to identify RNA triloop hairpins with enhanced stability and instability (*39*). Also, an RNA tetraloop library of 4096 sequences was recently studied with TGGE−selection, and a stable tetraloop motif was identified (*40*).

Here we introduce the application of TGGE−selections to DNA secondary structure motifs. TGGE−selection of a DNA tetraloop combinatorial library with 4096 members was carried out. We optimized the PCR component of the selection process to ensure that known examples of stable sequences were present in the final pool. Four families of exceptionally stable hairpins were isolated and characterized by base substitution, UV melting, and CD spectroscopy.

MATERIALS AND METHODS

Preparation of DNA. Solid-phase synthesized DNA was deblocked and desalted by the manufacturer (IDT). The library has six random nucleotides (N6) to allow both the tetraloop and the closing base pair to be optimized. The DNA sequences of the libraries are as follows: library 1, 5′-GGGAGAGGATTTAATTTANNNNNTAAATTGGATCCGCAACG; library 2, 5′-GACCCAGCGCGAGGATTTAATTTANNNNNTAAATTGGATCCGCAGCCAAGGCC. Library 2 was found to be optimal for the selections and was used in the majority of the experiments. Positions of randomization were prepared by mixing the four phosphoramidites for a single coupling reaction in a 1:1:1:1 ratio. The coupling reactivities were reported by the manufacturer (IDT) at 10:12:13:15 for A:C:G:T, respectively; since the library was relatively small, many copies of each sequence were present. All DNA hairpins used for selections were purified by 10% denaturing PAGE, crush and soak recovery, and ethanol precipitation as previously described (*38*). Only the top half of the band was cut from the gel to help to remove $n - 1$ sequences. The DNA oligonucleotides were stored in TE (=10 mM Tris, 1 mM Na_2EDTA, pH 7.5) at −20 °C. Prior to TGGE, the initial DNA tetraloop library was 5′-end labeled using polynucleotide kinase and [γ-^{32}P]-ATP. Radiolabeled DNAs were desalted and separated from unincorporated ATP by G-25 size-exclusion Sephadex Quick Spin columns (Boehringer Mannheim), and the DNA was quantitated by scintillation counting (Beckman LS 7500) (*38*).

For UV melting and CD experiments, minihairpins were designed that have short stems and the sequence 5′-d(ggaX$L_1L_2L_3L_4$X′tcc), where "X" and "X′" indicate the closing base pair and "L" indicates a loop nucleotide. For representative sequences, the purity was confirmed by observation of a single band of end-labeled DNA on denaturing PAGE (in 8.0 M urea at 50 °C); to prevent secondary structure, a small amount of DNA was glyoxylated and directly electrophoresed (*41*).

Temperature Gradient Gel Electrophoresis. The tetraloop library was electrophoresed by perpendicular TGGE in 10% polyacrylamide (29:1 acrylamide:bisacrylamide) with 2 M urea to favor melting near the center of the ≈15−60 °C range. The temperature gradient was measured by insertion of an Omega HH21 microprocessor thermometer directly into the gel and was linear with distance across the perpendicular gel. The buffer in the gel and reservoirs was 1× $TBEK_{50}$ (=100 mM Tris, 83 mM boric acid, 1 mM Na_2EDTA, 50 mM KCl, pH 8.8, at 20 °C), and gels were preelectrophoresed for 30 min at 350 V with both water baths running to equilibrate the temperature in the gel. To minimize dimerization, the appropriate mixture of DNA strands was renatured in TE by heating for 3 min at 90 °C, followed by cooling on the bench for at least 10 min. For perpendicular TGGE, 160 μL of the sample and dye were loaded into a 13 cm lane. For parallel TGGE, 20 μL of sample without dye was loaded into 10 mm lanes; additional lanes with dye only (=1× $TBEK_{50}$, 10% glycerol, 0.5% xylene cyanol, and 0.125% bromophenol blue) were loaded for parallel TGGE. Gels were run at 350 V for 2−2.5 h depending on the temperature gradient. The gradient was linear with position and did not fluctuate over the course of an electrophoresis

run (*39*). The reservoir buffers were emptied, mixed, and refilled every hour to prevent large pH gradients from developing. Analytical gels were dried on Whatman 3M paper and analyzed using a PhosphorImager (Molecular Dynamics).

Selection Procedure. For selection, preparative parallel TGGE gels were exposed to film at 4 °C without drying, and bands were visualized by autoradiography. The desired DNA species was excised with a razor blade and recovered by crush and soak. Selected DNAs from library 1 were annealed with a 5′-^{32}P-labeled top strand primer, 5′-p*GGGAGAGGATTTAATTT, and a 5′-biotinylated bottom strand primer, 5′-biotin-CGTTGCGGATCCAATTT. Selected DNAs from library 2 were annealed with a 5′-^{32}P-labeled top strand primer, 5′-p*GACCCAGCGCG-AGGATTTAATTTA, and a 5′-biotinylated bottom strand primer, 5′-biotin-GGCCTTGGCTGCGGATCCAATTTA. PCR buffer [20 mM Tris-HCl, pH 8.8, 2 mM $MgSO_4$, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton X-100, 0.1 mg/mL BSA, 1 mM dNTP] was mixed with primers and template, and additional $MgCl_2$ was added for a final concentration of 2.5 mM Mg^{2+}. The PCR reaction (library 2) was initiated with *Pfu* turbo DNA polymerase (Stratagene) and run at 97 °C for 1 min; 15× (97 °C, 45 s; 62 °C, 45 s; 72 °C, 90 s); 72 °C, 10 min; 4 °C, 10−20 min. The PCR product was affinity purified using streptavidin paramagnetic particles (Promega), and the top strand was released with 0.15 M NaOH (*42*). For the final selected DNAs, PCR reactions used the top strand primer, 5′-GCGCGAATTC-GACCCAGCGCGAGGATTTAATTTA-3′, which introduced an *Eco*RI cloning site. PCR products were digested with *Eco*RI and *Bam*HI and cloned into pUC19, with transformation and *Escherichia coli* DH-5α growth at 30 °C to minimize mutations. Sequences of isolated clones were determined by the dideoxy method. To avoid multiple bands in four lanes, 1.1 μg of single-stranded DNA binding protein (SSB) was added to the labeling reaction tube (*43*), and a reverse sequencing primer was used. Before addition of the stop solution, 1 μL of proteinase K (0.1 μg/mL) was added to each tube, and the resultant solution was incubated at 65 °C for 20 min to degrade the SSB. Only sequences without mutations in the fixed sequence stem were tabulated.

UV Melting Experiments. The DNAs used in UV melting studies were minihairpins (see above). UV absorbance melting profiles were obtained in $P_{10}E_{0.1}$ [=10 mM sodium phosphate, 0.1 mM Na_2EDTA (pH 7.0)] at 280 nm, using a Gilford Response II spectrophotometer with a heating rate of ≈1 °C/min. Melts were performed in 5 or 10 mm path length cuvettes, at at least three different strand concentrations (C_T) ranging from 1 to 50 μM. Melts were found to be independent of concentration, consistent with the hairpin conformation; if duplexes were forming, calculations using a nearest-neighbor model predicted a T_M change of ≈8 °C over a 20-fold difference in C_T (*10*−*12*, *15*, *16*), which is outside the error limits of the experiments (±1 °C). At the start of each experiment, the minihairpin was renatured in melting buffer by heating to 90 °C. Data for forward and reverse melts were similar, consistent with reversibility of the melting transition. Concentrations were calculated using absorbance values at 90 °C and extinction coefficients from a nearest-neighbor analysis (*44*, *45*). Thermodynamic parameters were obtained by fitting to a two-state model with sloping baselines using a nonlinear least-squares program (*46*), which provided good fits to the data. Briefly, this program fits the data to six variables, the slopes and intercepts of the two baselines and the $\Delta H°$ and $\Delta S°$ of the transition. It is assumed that the transition is two state and that the enthalpy is temperature-independent. The T_M and $\Delta G°_{37}$ are calculated from the standard thermodynamic relationships for hairpins.

Circular Dichroism. Circular dichroism data were obtained with an AVIV 62DS spectrometer at 25 °C in $P_{10}E_{0.1}$ (pH 7.0) and ≈10 μM C_T, conditions under which the major species was the folded hairpin for all sequences. Data are the average of three scans from 220 to 320 nm in 1 nm increments using an integration time of 2 s. A buffer blank was subtracted from all spectra. CD spectra were converted to molar residue ellipticity ($\Delta\epsilon$ in units of M^{-1} cm^{-1}) according to the equation:

$$\Delta\epsilon = \theta/[Cnb \times 33] \quad (1)$$

where θ is the measured ellipiticity in degrees, C is the DNA concentration in molar, n is the number of nucleotides, and b is the cuvette path length in centimeters (0.5) (*47*, *48*).

RESULTS

Design and Optimization of the DNA Tetraloop Library. The library was designed with six randomized positions to allow the sequence of the tetraloop and closing base pair to be investigated (Figure 1A,B). In certain cases, selected tetraloops may have Watson−Crick base pairing between positions 1 and 4 of the loop (see below). Nevertheless, we consider such loops tetraloops rather than diloops since all Watson−Crick base pairs were not found to substitute at these positions. This nomenclature has been adopted elsewhere for four-membered DNA hairpin loops (*49*).

The fixed sequence portion of the library was designed to incorporate several features (Figure 1A,B) (*38*). Primer binding sites are present at the 5′- and 3′-ends of the hairpin; a staggered stem sequence is used to avoid slipped pairings; and GT wobble pairs are used to make the downstream *Bam*HI cloning site unique. The hairpin library was also designed with a 12 bp stem to favor a large structural change upon unfolding.

Initially, in vitro selection was completed using DNA library 1 (Figure 1A). Conditions for the selection included PCR with an annealing temperature of only 50 °C to accommodate the short primer binding sites, primers that extended only to the penultimate base pair at the top of the stem, and cloning at 37 °C. However, sequencing of 47 clones revealed none of the known stable d(cGNNAg) tetraloop sequences (*28*). In addition, the closing base pair frequencies (in brackets) were GC [29], non-Watson−Crick [16], AT [1], and CG [1], with no preference toward the known stable CG closing base pair (*29*). Apparently, the selection of library 1 was biased in some way against stable tetraloops.

To investigate this problem, we performed a PCR reaction on an equal mixture of library 1 and a hairpin with the same stem but the stable d(cGAAAg) motif. Upon PCR, and cloning and sequencing, no d(cGAAAg) sequences were recovered. In addition, deletions were observed to occur when cloning at 37 °C. Subsequently, the library was redesigned as library 2 (Figure 1B), which has extended

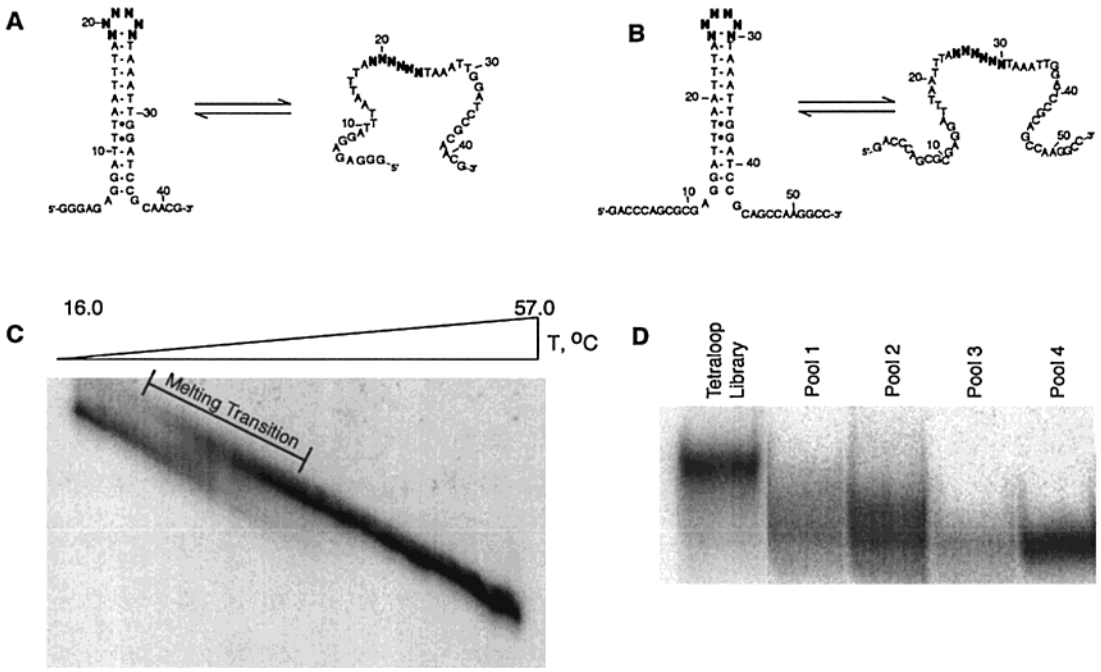

FIGURE 1: Melting and selection of the tetraloop library. Equilibrium between folded and unfolded forms of DNA hairpins for (A) library 1 with shorter primer binding sites and (B) library 2 with longer primer binding sites. Watson−Crick base pairs are denoted with a dash and wobble pairs with a dot. The letter N denotes an approximately equimolar mixture of A, C, G, and T. Each tetraloop library is a mixture of 4096 different sequences. (C) Perpendicular TGGE of ^{32}P-labeled library 2 (pool 0). The gel contained 2 M urea, and the temperature in the gel ranged from 16 to 57 °C (measured in the gel). The melting transition was between ≈21 and 37 °C and is indicated by the label "Melting Transition". (D) Parallel TGGE of library 2 after zero to four rounds of selection. The lane labeled Tetraloop Library is pool 0 containing 4096 sequences. The bottom portion just below the major region of the smear was excised in each case and amplified by PCR. Mobility increased with increasing rounds of selection, suggesting an enrichment for stable sequences. The gel contained 2 M urea, and the circulating water baths were set to 35 and 42 °C for the top and bottom of the gel, respectively.

flanking regions to allow usage of longer PCR primers and higher annealing temperatures. Also, primers were extended to the final fixed base pair in the top of the stem to aid disruption of stable tetraloops. An optimal PCR annealing temperature of 62 °C was determined by gradient thermocycling. As before, an equal mixture of library 2 and a hairpin with the same stem but the stable d(cGAAAg) loop motif was subjected to PCR and was cloned at 30 °C. Of the 12 clones sequenced, 6 were random and 6 were d(cGAAAg) without mutated stems. On the basis of this outcome, library 2 was used in the subsequent TGGE−selection experiments.

Selection of Stable DNA Tetraloops. Library 2 was subjected to gel electrophoresis with temperature gradients either perpendicular or parallel to the electric field (Figure 1C,D). Moderate concentrations of urea (2 M) were added to all gels to tune the T_M of the DNA library into an experimentally convenient range (*38*). Perpendicular TGGE of the initial library (pool 0) on a 16−57 °C gradient resulted in a "smeary" melting transition with the major region between ≈21 and 37 °C (Figure 1C). The smearing of the transition is consistent with a mixed population of sequences with different stabilities. Upper baselines were nearly coincident for all hairpins as were lower baselines (Figure 1C), consistent with similar folded and unfolded structures (Figure 1B). These observations suggest that differences in electrophoretic mobility on parallel TGGE gels should be associated primarily with differences in thermodynamic stability.

Parallel TGGE experiments were run between the temperatures of the major melting transition in perpendicular TGGE (Figure 1C,D). Parallel TGGE gels were designed to enrich for stable sequences, which have faster electrophoretic mobilities (*38*). Parallel TGGE of pool 0 (labeled "Tetraloop Library") revealed one major band with a smear below it (Figure 1D), which suggested that the initial library contains a small population of exceptionally stable sequences. The lower portion of the smear was excised, and the DNA was amplified by PCR using a 5′-^{32}P-labeled top strand primer and a 5′-biotin-labeled bottom strand primer. The top strand was separated from the bottom strand as described in the Materials and Methods section. Enrichment of stable sequences was confirmed by faster mobility for pool 1 (Figure 1D). In a similar fashion, pools 2, 3, and 4 were obtained from the lower portion of the band from the previous round. Pool 4 was chosen for sequencing and characterization. Electrophoresis of all pools under completely denaturing conditions revealed identical mobilities (data not shown), consistent with faster mobility on the parallel TGGE gel being due to differences in stability rather than PCR deletion artifacts.

Identification of Stable DNA Hairpins. We sequenced 114 clones from pool 4 of library 2 (Table 1). Among these, 63 have a CG closing base pair, and 51 have a GC closing base pair. The loops can be divided into four different motifs with the following sequences and frequencies: d(cGNNAg) [25/114], d(cGNABg) [18/114], d(cCNNGg) [11/114], d(gCNNGc) [49/114], and other loops [11/114] (Table 1).

For the d(cGNNAg) motif, 13 of the 16 possible sequences were found, and the closing base pair was almost exclusively a CG [25/26] (Table 1). This is in agreement with published loop and closing base pair preferences of the d(cGNNAg) motif (*28*, *29*). Recovery of the d(cGNNAg) motif along with

Table 1: Distribution of Loop and Closing Base Pair Sequences in Stable Hairpins from Selection[a]

motif[b]	loop sequence	occurrences[c]	melt data[d]	CD data[d]
loops with CG closing base pairs				
cGNNAg (25)	GCGA	3	y	y
	GCTA	3	y	y
	GTAA	3		
	GTGA	3		
	GAAA	2	y	y
	GCAA	2	y	y
	GCCA	2		
	GTCA	2		
	GACA	1		
	GGAA	1		
	GGCA	1	y	y
	GGGA	1		
	GTTA	1		
cGNABg (18)	GCAT	5	y	y
	GCAC	3	y	y
	GGAC	3	y	y
	GAAT	2		
	GCAG	1	y	y
	GGAG	1	y	y[e]
	GGAT	1	y	y
	GTAC	1	y	y
	GTAT	1		
cCNNGg (11)	CGCG	3		
	CGTG	3	y	y
	CGAG	2	y	y
	CAGG	1		
	CATG	1		
	CTTG	1	y	y[e]
others (9)	GACC	2	y	
	GACT	2	y	
	AACA	1		
	ACCT	1		
	GCTT	1	y	y
	GGCC	1		
	CGAAGT[f]	1	y	y[e]
loops with GC closing base pairs				
gCNNGc (49)	CATG	7	y	y[e]
	CGCG	7	y	y
	CGTG	7	y	y
	CTGG	6		
	CGAG	4	y	y
	CTAG	4		
	CTCG	4		
	CAAG	2		
	CCCG	2		
	CCTG	2	y	y[e]
	CACG	1		
	CAGG	1	y	y[e]
	CGGG	1		
	CTTG	1		
others (2)	GTCA	1	y	y[e]
	TGAG	1		

[a] Summary of stable loop sequences isolated from in vitro selection experiments. The total number of clones sequenced was 114. All sequences are DNA. [b] Sequences are organized into consensus motifs, and occurrences for a motif are provided in parentheses. [c] Listed by occurrences and then alphabetically. [d] Sequences for which melting and CD data were collected. [e] Data not shown. [f] This sequence may form a stable GAA triloop with a CG closing base pair and a bulged T.

numerous other sequences suggests that unknown tetraloops present in the final pool may also be stable.

The second motif that was found, d(cGNABg), appears to be an expansion of the known stable triloop, d(cGNAg) (*30*, *31*). For this motif, 9 of the 12 possible sequences were found, and the closing base pair was exclusively a CG (Table 1). Similarity of loop and closing base pair preferences to the d(cGNNAg) and d(cGNAg) motifs suggests that the d(cGNABg) motif may be related in structure. We are not aware of any prior reports on this motif.

The other two motifs identified were d(cCNNGg) and d(gCNNGc). For the d(cCNNGg) motif, 6 of the 16 possible sequences were found, although only 11 of the 114 clones sequenced fit this motif (Table 1). For the d(gCNNGc) motif, 14 of the 16 possible sequences were found (Table 1). We are aware of only one study involving these two motifs, namely showing that d(gCTTGc) is exceptionally stable (*19*). This sequence was originally studied because it is similar to the r(gCUUGc) tetraloop sequence that is common in rRNA (*18*) and is exceptionally stable (*19*). Although similar in sequence and stability, the d(cCNNGg) and d(gCNNGc) motifs may adopt different folds on the basis of CD and mutagenesis studies (see below).

Last, 6 sequences (8 clones) were found with CG closing base pairs that did not conform to one of these motifs, and 2 sequences (2 clones) were found with GC closing base pairs that did not conform to the d(gCNNGc) motif. For these sequences, 7/8 have a purine at position 1, with only one having the possibility of a sheared GA between position 1 and positions 3 or 4. One occurrence of the sequence, d(CGAAGT), was found. This sequence might form a GAA triloop closed by a CG base pair with a bulged T 3′ of the CG base pair. This would be consistent with the observation that the d(cGAAg) triloop is the most stable DNA loop sequence investigated (see below).

Thermodynamic Characterization of d(cGNNAg) and d(cGNABg) Minihairpins. To determine the stability of the selected tetraloop−closing base pair combinations, UV melting experiments were conducted on minihairpins. Minihairpins are stem−loops containing the selected loop−closing base pair combination but with stems of only four base pairs (*39*). These smaller hairpins were used to facilitate two-state behavior and to suppress alternative folds and kinetically trapped dimers. Table 2 summarizes results of melts on representative sequences from the selection and related variants. In all cases, melts were conducted between ≈1 and 50 μM C_T, and parameters were independent of concentration as expected for hairpins (see Materials and Methods).

The most stable sequence studied was the d(cGCAg) triloop (not selected), which has a T_M of 76.1 °C and a $\Delta G°_{37}$ of −3.7 kcal/mol (Table 2). This sequence is a member of the stable d(cGNAg) motif that has been previously characterized (*30*, *31*). Sequences from the d(cGNNAg) and d(cGNABg) motifs were also exceptionally stable and only slightly less stable than d(cGCAg) (see Figure 2 for representative melting profiles and reversibility of melting). For the d(cGNNAg) motif, 5 sequences were melted, resulting in T_M values from 72.7 to 68.2 °C and $\Delta G°_{37}$ values from −3.3 to −2.6 kcal/mol (Table 2). For the d(cGNABg) motif, 11 sequences were melted, resulting in T_M values from 67.9 to 64.4 °C and $\Delta G°_{37}$ values from −3.1 to −2.0 kcal/mol (Table 2).

The closing base pair for the selected d(GNNA) and d(GNAB) loop sequences was almost exclusively a CG (43/44 occurrences) (Table 1). The thermodynamic preference for the CG closing base pair was investigated, and a strong preference for a CG was found (Figure 2B, Table 3). For

Table 2: Thermodynamic Parameters for Minihairpin Formation Measured by UV Melting[a]

sequence[b]	$\Delta H°$ (kcal mol^{-1})	$\Delta S°$ (cal mol^{-1} K^{-1})	$\Delta G°_{37}$ (kcal mol^{-1})	$\Delta G°_{55}$[c] (kcal mol^{-1})	T_M (°C)	occurrences
$cL_1L_2L_3L_4g$[d]						
GCA	−33.0 ± 1.3	−94.4 ± 3.7	−3.69 ± 0.12	−1.99 ± 0.06	76.1 ± 0.7	0
GCTA	−32.3 ± 1.0	−93.6 ± 3.1	−3.26 ± 0.11	−1.58 ± 0.05	71.8 ± 0.2	3
GGCA	−30.8 ± 1.6	−89.2 ± 4.7	−3.19 ± 0.17	−1.58 ± 0.09	72.7 ± 0.5	1
GCAA	−30.3 ± 1.4	−87.6 ± 4.0	−3.12 ± 0.12	−1.61 ± 1.14	72.5 ± 0.3	2
GTAC	−34.8 ± 1.1	−102.1 ± 3.2	−3.09 ± 0.07	−1.25 ± 0.04	67.3 ± 0.5	1
GIAC	−33.5 ± 1.5	−98.2 ± 4.6	−3.03 ± 0.09	−1.26 ± 0.07	67.9 ± 1.1	0
GCAT	−33.6 ± 1.2	−98.5 ± 3.5	−2.99 ± 0.10	−1.21 ± 0.03	67.3 ± 0.5	5
GCAC	−33.0 ± 1.0	−97.0 ± 2.1	−2.96 ± 0.10	−1.21 ± 0.06	67.5 ± 0.3	3
GCGA	−29.7 ± 1.0	−86.6 ± 3.0	−2.83 ± 0.06	−1.27 ± 0.03	69.6 ± 0.5	3
CTTG	−34.0 ± 0.5	−100.6 ± 1.3	−2.81 ± 0.12	−1.00 ± 0.10	64.9 ± 0.9	1
GGAC	−31.7 ± 0.9	−93.2 ± 2.8	−2.76 ± 0.03	−1.08 ± 0.05	66.6 ± 0.8	3
GGAT	−32.2 ± 2.1	−95.1 ± 6.3	−2.73 ± 0.15	−1.02 ± 0.06	65.8 ± 0.7	1
CGTG	−32.0 ± 1.2	−95.0 ± 3.7	−2.56 ± 0.08	−0.85 ± 0.02	63.9 ± 0.3	3
GAAA	−27.9 ± 0.1	−81.8 ± 0.3	−2.55 ± 0.03	−1.05 ± 0.05	68.2 ± 0.3	2
GCAI	−30.7 ± 0.8	−90.8 ± 2.5	−2.55 ± 0.07	−0.91 ± 0.05	65.0 ± 0.6	0
GCAG	−28.0 ± 1.1	−82.7 ± 3.5	−2.32 ± 0.08	−0.83 ± 0.06	65.1 ± 1.4	1
GCI	−29.3 ± 1.2	−87.2 ± 3.8	−2.29 ± 0.09	−0.72 ± 0.09	63.3 ± 1.2	0
GACC	−32.9 ± 1.0	−98.9 ± 2.9	−2.26 ± 0.05	−0.48 ± 0.02	59.9 ± 0.3	2
CGAG	−30.9 ± 1.6	−92.7 ± 4.8	−2.19 ± 0.10	−0.52 ± 0.06	60.6 ± 0.7	2
GACT	−32.7 ± 1.0	−98.5 ± 3.1	−2.18 ± 0.06	−0.40 ± 0.06	59.2 ± 0.4	2
GAAG	−26.3 ± 0.6	−78.0 ± 1.8	−2.14 ± 0.02	−0.64 ± 0.17	64.4 ± 0.5	0
CGCG	−27.6 ± 1.7	−82.4 ± 5.0	−2.08 ± 0.17	−0.62 ± 0.10	62.2 ± 0.7	3
GGAG	−24.7 ± 2.1	−73.2 ± 6.5	−2.04 ± 0.06	−0.72 ± 0.06	65.0 ± 1.6	1
CGAAGT[d]	−22.6 ± 2.0	−66.3 ± 6.6	−2.04 ± 0.02	−0.85 ± 0.14	68.1 ± 3.5	1
ICA	−29.1 ± 1.5	−87.5 ± 4.7	−1.99 ± 0.07	−0.41 ± 0.04	59.7 ± 0.7	0
GGTC	−30.7 ± 0.6	−92.6 ± 2.0	−1.95 ± 0.07	−0.29 ± 0.07	58.1 ± 0.8	0
GCTC	−30.3 ± 0.9	−91.7 ± 2.9	−1.90 ± 0.06	−0.25 ± 0.07	57.7 ± 0.8	0
ICAC	−27.2 ± 0.7	−82.8 ± 2.2	−1.51 ± 0.05	−0.02 ± 0.06	55.2 ± 0.8	0
ICAT	−27.3 ± 0.9	−83.3 ± 2.9	−1.50 ± 0.03	0.00 ± 0.06	55.1 ± 0.7	0
IGAC	−29.8 ± 1.6	−91.2 ± 4.9	−1.54 ± 0.06	0.11 ± 0.07	53.8 ± 0.8	0
GCTT	−26.2 ± 0.8	−80.0 ± 2.7	−1.42 ± 0.04	0.02 ± 0.08	54.7 ± 1.0	1
GCIG	−25.5 ± 1.4	−77.9 ± 4.2	−1.35 ± 0.10	0.05 ± 0.10	54.3 ± 1.3	0
$gL_1L_2L_3L_4c$						
CGTG	−33.5 ± 0.4	−99.4 ± 1.2	−2.69 ± 0.03	−0.90 ± 0.02	64.1 ± 0.3	7
CGCG	−31.6 ± 1.2	−94.0 ± 3.5	−2.46 ± 0.11	−0.77 ± 0.06	63.2 ± 0.5	7
CCTG	−32.6 ± 1.3	−97.4 ± 4.1	−2.37 ± 0.08	−0.62 ± 0.06	61.4 ± 0.8	2
CGAG	−31.3 ± 0.5	−93.8 ± 1.5	−2.22 ± 0.02	−0.53 ± 0.03	60.7 ± 0.4	4
CATG	−31.2 ± 0.7	−93.8 ± 2.1	−2.10 ± 0.07	−0.41 ± 0.06	59.4 ± 0.7	7
CAGG	−28.9 ± 0.4	−88.5 ± 1.3	−1.49 ± 0.02	0.10 ± 0.02	53.9 ± 0.2	1
GTCA	−26.6 ± 1.9	−81.1 ± 5.9	−1.40 ± 0.06	0.06 ± 0.09	54.4 ± 1.0	1
GCTA	−27.7 ± 1.5	−85.2 ± 4.6	−1.27 ± 0.04	0.26 ± 0.07	52.0 ± 0.7	0
GCAT	−26.0 ± 1.6	−80.2 ± 5.2	−1.11 ± 0.02	0.33 ± 0.08	50.9 ± 0.8	0
GCAC	−26.5 ± 1.0	−82.0 ± 2.9	−1.08 ± 0.12	0.39 ± 0.11	50.2 ± 1.4	0
GGTC	−23.9 ± 0.8	−75.1 ± 2.7	−0.59 ± 0.06	0.76 ± 0.08	44.8 ± 0.9	0
GAAA	−19.9 ± 1.9	−62.5 ± 5.9	−0.56 ± 0.08	0.57 ± 0.03	45.9 ± 0.6	0
$aL_1L_2L_3L_4t$						
CGTG	−32.0 ± 1.4	−98.4 ± 4.4	−1.53 ± 0.03	0.24 ± 0.09	52.6 ± 0.8	0

[a] Solutions were $P_{10}E_{0.1}$ (pH 7.0). Melting is of minihairpins, 5′-d(ggaX$L_1L_2L_3L_4$X′tcc). Parameters are the average of at least three independently prepared samples between 1 and 50 μM C_T, and errors are standard deviations. The T_M was constant between 1 and 50 μM C_T. All sequences are DNA. [b] Listed by $\Delta G°_{37}$ and then alphabetically. [c] $\Delta G°_{55}$ values are provided since they are closer to the T_M for the sequences melted and may therefore have less error associated with them. Similar trends were observed with $\Delta G°_{37}$ values, and these were chosen for the text. [d] See footnote *f* of Table 1.

the d(cGNNAg) motif, changing the closing base pair from a CG to a GC for d(GAAA) and d(GCTA) loops resulted in ΔT_Ms of −22.3 and −19.9 °C, respectively, and $\Delta\Delta G°_{37}$s of 2.0 kcal/mol (Table 3). These values are in agreement with ΔT_Ms from a previous study on a d(GGCA) loop sequence (*29*). For the d(cGNABg) motif, changing the closing base pair from a CG to a GC for d(GCAC) and d(GCAT) loops resulted in ΔT_Ms of −17.3 and −16.3 °C, respectively, and $\Delta\Delta G°_{37}$s of 1.9 kcal/mol (Table 3). The expected $\Delta\Delta G°_{37}$ for a CG to GC switch next to an AT base pair is only 0.16 kcal/mol based on unified nearest-neighbor parameters (*10*, *11*). [These parameters are at 1 M NaCl, while the thermodynamic parameters reported here were measured at ≈10 mM Na^+ to minimize duplex formation; however, preliminary melting experiments for d(GNNA) and d(GNAB) loop sequences at 1 M NaCl revealed similar $\Delta\Delta G°_{37}$ and ΔT_M values for CG to GC or AT closing base pair changes (T. M. Denkenberger and P. C. Bevilacqua, unpublished data).] Thus, for both d(cGNNAg) and d(cGNABg), the CG closing base pair makes a large contribution to overall stability that cannot be explained by the standard nearest-neighbor model. Similarly, large thermodynamic contributions of CG closing base pairs have been observed for certain RNA tetraloops and triloops (*19*, *39*).

To investigate the d(cGNABg) motif in more detail, we made changes at each position in the loop and determined the effect on hairpin stability (Figure 3). Changing position 1 from G to I (inosine), which substitutes the amino group

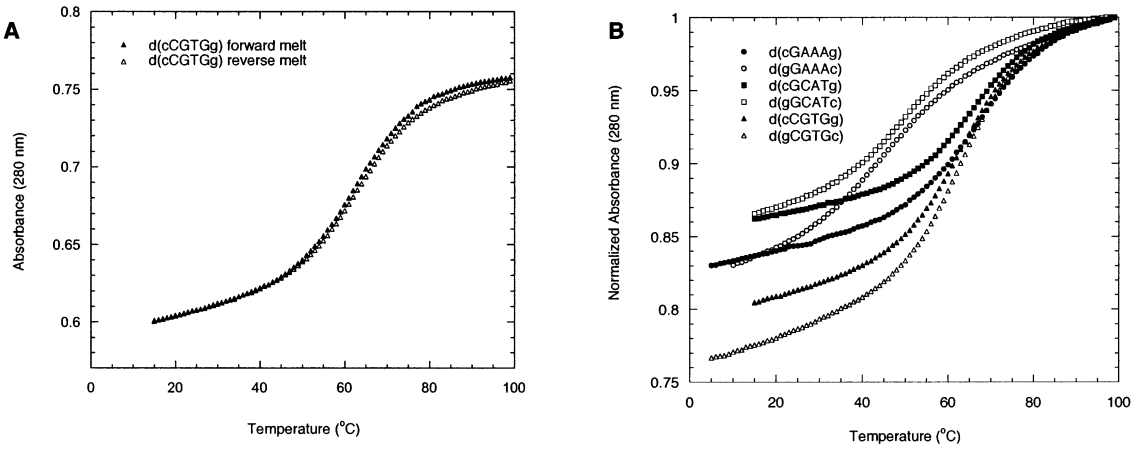

FIGURE 2: Representative UV melting curves for minihairpins from each of the four stable families and controls. (A) Forward (▲) and reverse (△) melts for d(cCGTGg). Melts were carried out as described in the Materials and Methods section. Data are nearly superimposable, consistent with reversibility of melting. (B) Loop sequences are as follows: d(cGAAAg) (●), d(gGAAAc) (○), d(cGCATg) (■), d(gGCATc) (□), d(cCGTGg) (▲), and d(gCGTGc) (△). Sequences were from pool 4, except for d(gGAAAc) and d(gGCATc). Data were collected in $P_{10}E_{0.1}$ (pH 7.0), and $C_T \approx 10$ mM. Thermodynamic parameters are given in Table 2.

Table 3: Thermodynamic Comparisons for Closing Base Pair Changes for All Tetraloop Motifs

sequence[a]	$\Delta\Delta H°$ (kcal mol^{-1})	$\Delta\Delta S°$ (cal mol^{-1} K^{-1})	$\Delta\Delta G°_{37}$ (kcal mol^{-1})	$\Delta\Delta G°_{55}$[b] (kcal mol^{-1})	ΔT_M (°C)
GNNA					
gGAAA**c**-**c**GAAA**g**	8.0 ± 1.9	19.3 ± 5.9	1.99 ± 0.09	1.64 ± 0.04	−22.3 ± 0.7
gGCTA**c**-**c**GCTA**g**	4.6 ± 1.8	8.5 ± 5.5	1.99 ± 0.91	1.84 ± 0.09	−19.9 ± 0.7
GNAB					
gGCAC**c**-**c**GCAC**g**	6.5 ± 1.4	15.0 ± 4.2	1.88 ± 0.16	1.60 ± 0.12	−17.3 ± 1.4
gGCAT**c**-**c**GCAT**g**	7.6 ± 2.0	18.4 ± 6.3	1.88 ± 0.10	1.55 ± 0.10	−16.3 ± 1.0
CNNG					
aCGTG**t**-**c**CGTG**g**	−0.0 ± 1.8	−3.4 ± 5.7	1.03 ± 0.08	1.09 ± 0.09	−11.4 ± 0.9
gCGAG**c**-**c**CGAG**g**	−0.4 ± 1.6	−1.1 ± 5.0	−0.04 ± 0.10	−0.01 ± 0.07	0.1 ± 0.8
gCGTG**c**-**c**CGTG**g**	−1.5 ± 1.3	−4.4 ± 3.8	−0.13 ± 0.08	−0.05 ± 0.03	0.1 ± 0.4
gCGCG**c**-**c**CGCG**g**	−4.0 ± 2.1	−11.6 ± 6.1	−0.38 ± 0.02	−0.15 ± 0.12	0.9 ± 0.9

[a] Bold font indicates bases being compared. Pairs listed in order of most penalizing closing base pair conversion. All sequences are DNA. Errors were propagated by standard methods (*65*). [b] $\Delta G°_{55}$ values are provided since they are closer to the T_M for the sequences melted and may therefore have less error associated with them. Similar trends were observed with $\Delta G°_{37}$ values, and these were chosen for the text.

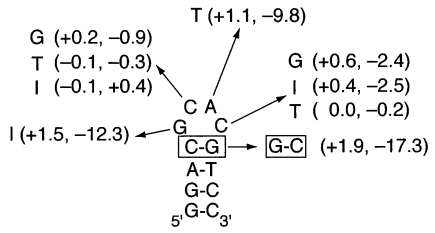

FIGURE 3: Summary of thermodynamic effects for substitutions in the d(cGNABg) motif. Base substitutions made at a particular position are shown. Effects of substitutions on $\Delta\Delta G°_{37}$ (kcal/mol) and ΔT_M (°C) relative to d(cGCACg) are provided in parentheses ($\Delta\Delta G°_{37}$, ΔT_M). Original data are provided in Tables 2−4.

of G with a hydrogen, substantially destabilized the hairpin; for d(GCAT) and d(GCAC) loops, G to I substitutions resulted in ΔT_Ms of −12.3 °C and $\Delta\Delta G°_{37}$s of 1.5 kcal/mol (Table 4). Likewise, changes at position 3 of the loop substantially destabilized the hairpin; replacement of the A with T or I resulted in ΔT_Ms ranging from −12.8 to −9.8 °C and $\Delta\Delta G°_{37}$s ranging from 1.6 to 1.0 kcal/mol (Table 4). In contrast, changes at positions 2 and 4 of the loop resulted in much smaller ΔT_M and $\Delta\Delta G°_{37}$ effects (Table 4). Overall, the data for substitutions on the d(cGNABg) motif are consistent with the importance of positions 1 and 3 and the CG closing base pair and with the unimportance of positions 2 and 4, as expected.

The destabilizing effects of the I substitutions at positions 1 and 3 of the loop are consistent with expected effects for a sheared GA base pair between positions 1 and 3. Indeed, the $\Delta\Delta G°_{37}$ values for inosine substitutions at loop positions 1 and 3 of the d(cGNABg) motif are similar to those for positions 1 and 3 of the d(cGNAg) motif (Table 2). This would be consistent with d(cGNABg) representing expansion of structure of the d(cGNAg) triloop which contains a sheared GA base pair (*30*).

Thermodynamic Characterization of d(cCNNGg) and d(gCNNGc) Minihairpins. The other two motifs found in the selection are d(cCNNGg) and d(gCNNGc). For the d(cCNNGg) motif, 4 sequences were melted, resulting in T_M values from 64.9 to 60.6 °C and $\Delta G°_{37}$ values from −2.8 to −2.1 kcal/mol (Table 2). For the d(gCNNGc) motif, 6 sequences were melted, resulting in T_M values from 64.1 to 53.9 °C and $\Delta G°_{37}$ values from −2.7 to −1.5 kcal/mol (Table 2); the lower value on the d(gCNNGc) motif is from the one occurrence of a d(gCAGGc) sequence, and the next lowest stability sequence in this motif has thermodynamic parameters that are similar to the lower ones for the other motifs. Thus, the d(cCNNGg) and d(gCNNGc) motifs are similar

Table 4: Thermodynamic Comparisons for Base Changes in cGNABg Loops[a]

loop position	sequence[b]	$\Delta\Delta H°$ (kcal mol^{-1})	$\Delta\Delta S°$ (cal mol^{-1} K^{-1})	$\Delta\Delta G°_{37}$ (kcal mol^{-1})	$\Delta\Delta G°_{55}$[c] (kcal mol^{-1})	ΔT_M (°C)
1	**I**CAT-**G**CAT	6.2 ± 1.5	15.2 ± 4.5	1.49 ± 0.10	1.21 ± 0.08	−12.3 ± 0.8
	ICAC-**G**CAC	5.8 ± 1.2	14.2 ± 3.7	1.45 ± 0.11	1.19 ± 0.09	−12.3 ± 0.8
2	G**G**AT-G**C**AT	1.3 ± 2.4	3.5 ± 7.2	0.26 ± 0.18	0.19 ± 0.08	−1.6 ± 0.9
	G**G**AC-G**C**AC	1.4 ± 1.4	3.8 ± 4.1	0.20 ± 0.11	0.13 ± 0.07	−0.9 ± 0.9
	G**I**AC-G**C**AC	−0.5 ± 1.8	−1.2 ± 5.5	−0.07 ± 0.14	0.05 ± 0.09	0.4 ± 1.2
	G**T**AC-G**C**AC	−1.7 ± 1.5	−5.2 ± 4.4	−0.13 ± 0.13	−0.04 ± 0.07	−0.3 ± 0.6
3	GC**T**T-GC**A**T	7.3 ± 1.4	18.6 ± 4.4	1.57 ± 0.10	1.24 ± 0.09	−12.8 ± 1.0
	GC**T**C-GC**A**C	2.7 ± 1.4	5.3 ± 4.1	1.06 ± 0.12	0.97 ± 0.09	−9.8 ± 0.8
	GC**I**G-GC**A**G	2.5 ± 1.7	4.8 ± 5.5	0.97 ± 0.13	0.89 ± 0.14	−10.8 ± 1.9
4	GCA**G**-GCA**C**	5.1 ± 1.5	14.3 ± 4.6	0.64 ± 0.13	0.38 ± 0.11	−2.4 ± 1.4
	GCA**I**-GCA**C**	2.3 ± 1.3	6.2 ± 3.9	0.42 ± 0.13	0.30 ± 0.08	−2.5 ± 0.6
	GCA**T**-GCA**C**	−0.5 ± 1.6	−1.6 ± 4.6	−0.03 ± 0.14	0.00 ± 0.08	−0.2 ± 0.6

[a] All sequences shown are for DNA and have a CG closing base pair. Errors were propagated by standard methods (*65*). [b] Bold font indicates bases being compared. Pairs listed in order of most penalizing base changes. [c] $\Delta G°_{55}$ values are provided since they are closer to the T_M for the sequences melted and may therefore have less error associated with them. Similar trends were observed with $\Delta G°_{37}$ values, and these were chosen for the text.

in stability and only slightly less stable than members of the d(cGNABg) motif.

The closing base pair preferences for the selected d(CNNG) loop sequences were investigated by UV melting (Figure 2B, Table 3). As seen in Figure 2B, CG and GC closing base pair sequences give similar melting profiles for d(CNNG) loops but not for d(GNNA) and d(GNAB) loops. Thermodynamic comparisons of CG and GC closing base pair changes are summarized in Table 3; for d(CGAG), d(CGTG), and d(CGCG) loops, the ΔT_Ms and $\Delta\Delta G°_{37}$s were close to zero and ranged from 0.9 to 0.1 °C and −0.4 to −0.04 kcal/mol, respectively. In addition, we changed the closing base pair for the d(CGTG) loop to an AT base pair; compared to a CG closing base pair, this gave a ΔT_M of −11.4 °C and a $\Delta\Delta G°_{37}$ of 1.0 kcal/mol (similar values are obtained when comparing the AT to the GC closing base pair). The expected $\Delta\Delta G°_{37}$ for a CG to AT switch next to an AT base pair in the stem is 0.44 kcal/mol based on unified nearest-neighbor parameters (*10*, *11*). Thus, for d(cCNNGg) and d(gCNNGc), the closing base pair makes a somewhat larger contribution to overall stability than expected from a simple nearest-neighbor model, although the non-nearest-neighbor effect is not as large as for the d(cGNNAg) and d(cGNABg) motifs.

Similarity of sequences and thermodynamics for the d(cCNNGg) and d(gCNNGc) motifs might suggest that these two motifs have the same structure. However, several melting and CD results suggest that there may be differences. The thermodynamic effects of swapping positions 1 and 4 of the loop were examined for the sequences d(cCGTGg) and d(gCGTGc). For d(cCGTGg) this swap resulted in a modest destabilization, with a ΔT_M of only −5.8 °C and a $\Delta\Delta G°_{37}$ of 0.61 kcal/mol; in contrast, for d(gCGTGc) this swap resulted in a large destabilization with a ΔT_M of −19.3 °C and a $\Delta\Delta G°_{37}$ of 2.1 kcal/mol (Table 2). These differences suggest that the d(cCNNGg) and d(gCNNGc) motifs may not be identical.

Hilbers and co-workers showed that purine−pyrimidine Watson−Crick complementarity between positions 1 and 4 of a DNA tetraloop leads to a less stable hairpin than pyrimidine−purine Watson−Crick complementarity (*49*, *50*). The low frequency of purine−pyrimidine Watson−Crick complementarity of positions 1 and 4 found in our selections

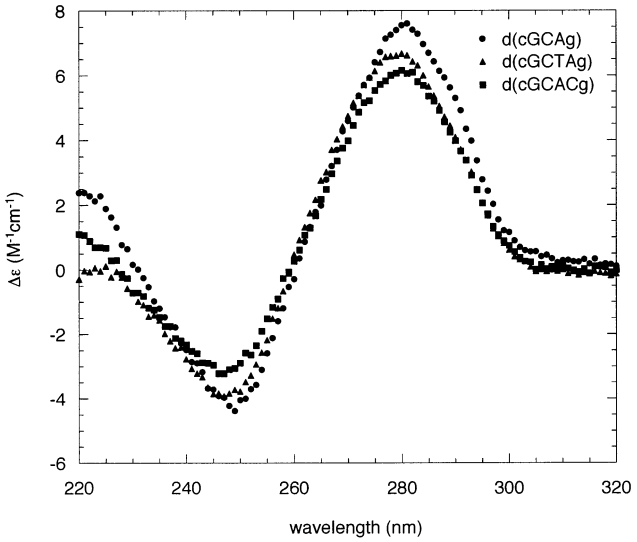

FIGURE 4: Representative CD spectra from d(cGNAg)-related sequences. Loop sequences are as follows: d(cGCAg) (●), d(cGCTAg) (▲), and d(cGCACg) (■). Spectra were collected in $P_{10}E_{0.1}$ (pH 7.0) at 25 °C, and $C_T \approx 10$ μM.

is generally consistent with this model. Nevertheless, when such changes were made, they were less penalizing for CG closing base pairs than for GC closing base pairs. Studies from Hilbers and co-workers focused on GC and AT closing base pairs (*49*); thus, having a CG closing base pair may improve the ability of the loop to tolerate purine−pyrimidine Watson−Crick complementarity between positions 1 and 4 of the loop.

Characterization of Minihairpins by CD Spectroscopy. Circular dichroism is effective at reporting certain changes in nucleic acid conformation. For example, A-form and B-form dsDNAs have distinctive spectroscopic features, including differing behavior at 260 and 280 nm (*40*, *51*). Circular dichroism was used to compare the d(cGNAg), d(cGNNAg), and d(cGNABg) families. The CD spectra of sequences from these three motifs are nearly identical, with a maximum near 280 nm and a minimum near 248 nm (Figure 4, Supporting Information). Similarity among these spectra is consistent with similar folds for these three motifs. Similarity of d(cGNAg) and d(cGNNAg) is expected on the basis of both containing sheared GA base pairs (*29*, *30*) and is also expected for d(cGNABg) on the basis of the

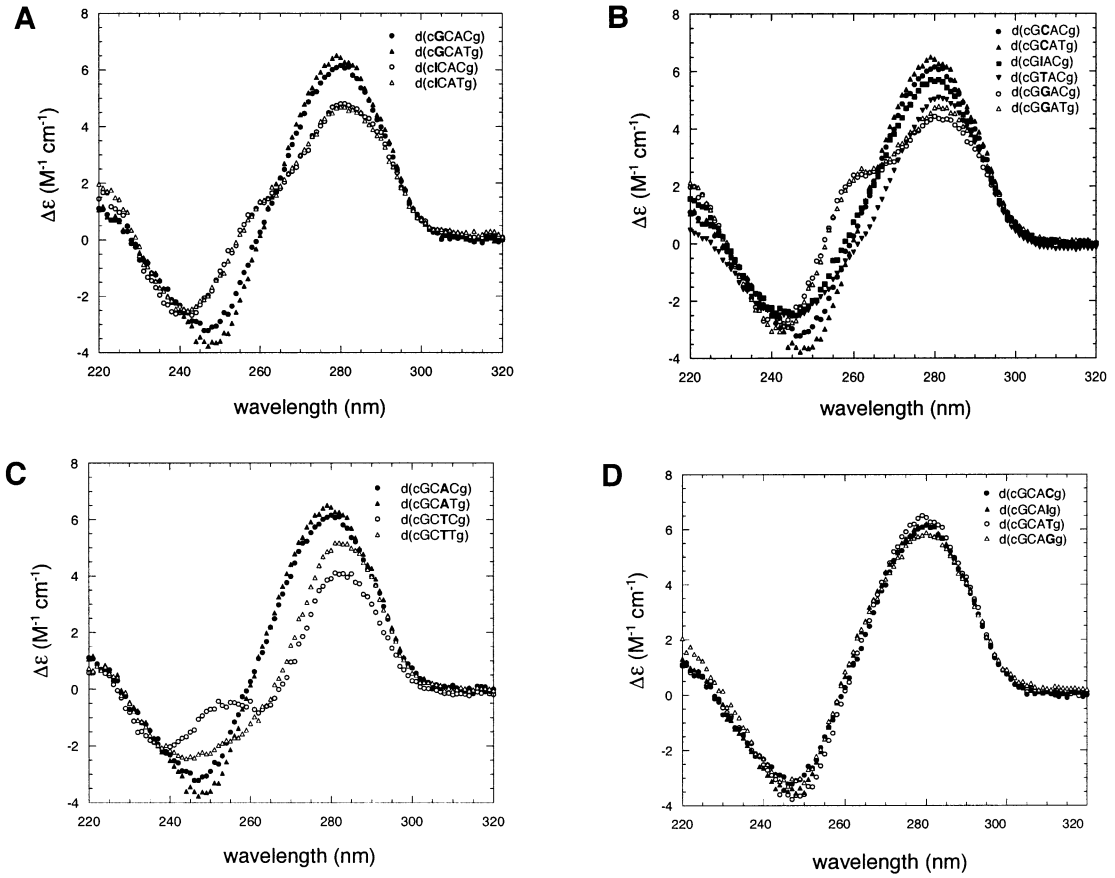

FIGURE 5: Position dependence of CD spectra for the d(cGNABg) motif. (A) Effect of substitutions at position 1 of the loop. Loop sequences are as follows: d(cGCACg) (●), d(cGCATg) (▲), d(cICACg) (○), and d(cICATg) (△). (B) Effect of substitutions at position 2 of the loop. Loop sequences are as follows: d(cGCACg) (●), d(cGCATg) (▲), d(cGIACg) (■), d(cGTACg) (▼), d(cGGACg) (○), and d(cGGATg) (△). (C) Effect of substitutions at position 3 of the loop. Loop sequences are as follows: d(cGCACg) (●), d(cGCATg) (▲), d(cGCTCg) (○), and d(cGCTTg) (△). (D) Effect of substitutions at position 4 of the loop. Loop sequences are as follows: d(cGCACg) (●), d(cGCAIg) (▲), d(cGCATg) (○), and d(cGCAGg) (△). Spectra were collected in $P_{10}E_{0.1}$ (pH 7.0) at 25 °C, and $C_T \approx 10$ μM.

thermodynamic effects of substitutions (Figure 3, Tables 2−4). In addition, these spectra are similar to that for B-form DNA (*51*), suggesting that these loops do not severely distort the stem.

To investigate the d(cGNABg) motif in greater detail, we examined the CD spectra for substitutions at positions 1−4 of the loop (Figure 5). As expected from thermodynamic studies and selection statistics (see above), changes at positions 1 and 3 of the loop resulted in changes in the CD spectra (Figure 5A,C). In both cases, the ellipticity changes at the maximum and minimum are less intense, with subtle changes elsewhere in the spectrum. Changes at position 4 of the loop did not affect the CD spectrum, as expected from thermodynamic studies and selection statistics (Figure 5D). Also, most of the changes at loop position 2 did not affect the CD spectrum, with the exception of a G substitution (Figure 5B). This result was unexpected and gave rise to a less intense peak at 280 nm, a shoulder at ≈255 nm, and a minimum at ≈240 nm. Subtraction of the d(cGGACg) and d(cGCACg) spectra resulted in a trace that was qualitatively similar to that reported for the dinucleotide, d(GpGp) (*52*) (Figure 6). This result suggests that the two Gs in d(cGGACg) may stack strongly. This would be consistent with the structure of d(cGAAg), in which the G and A at positions 1 and 2 of the triloop stack (*30*).

The CD spectra of representative d(cCNNGg) and d(gCNNGc) sequences were also compared (Figure 7). The CD spectra for d(CGAG), d(CGTG), and d(CGCG) loops were similar when the closing base pair was CG; likewise, the CD spectra of the loop sequences were similar when the

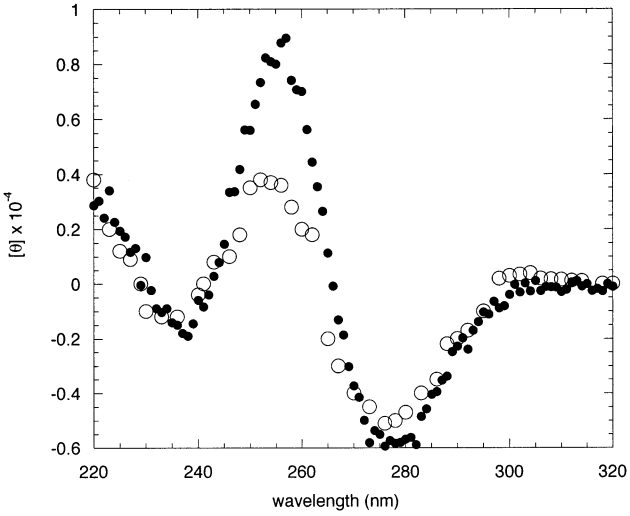

FIGURE 6: Difference CD spectrum between the d(cGGACg) and d(cGCACg) (●) compared to d(GpGp) (○). The difference spectrum was obtained by subtraction of the buffer-corrected spectra for d(cGGACg) and d(cGCACg) (Figure 5B), followed by division by 3.03 to convert from $\Delta\epsilon$ to $[\theta]$ (*47*). The spectrum for d(GpGp) was from the literature (*52*) and was obtained under conditions similar to those used here.

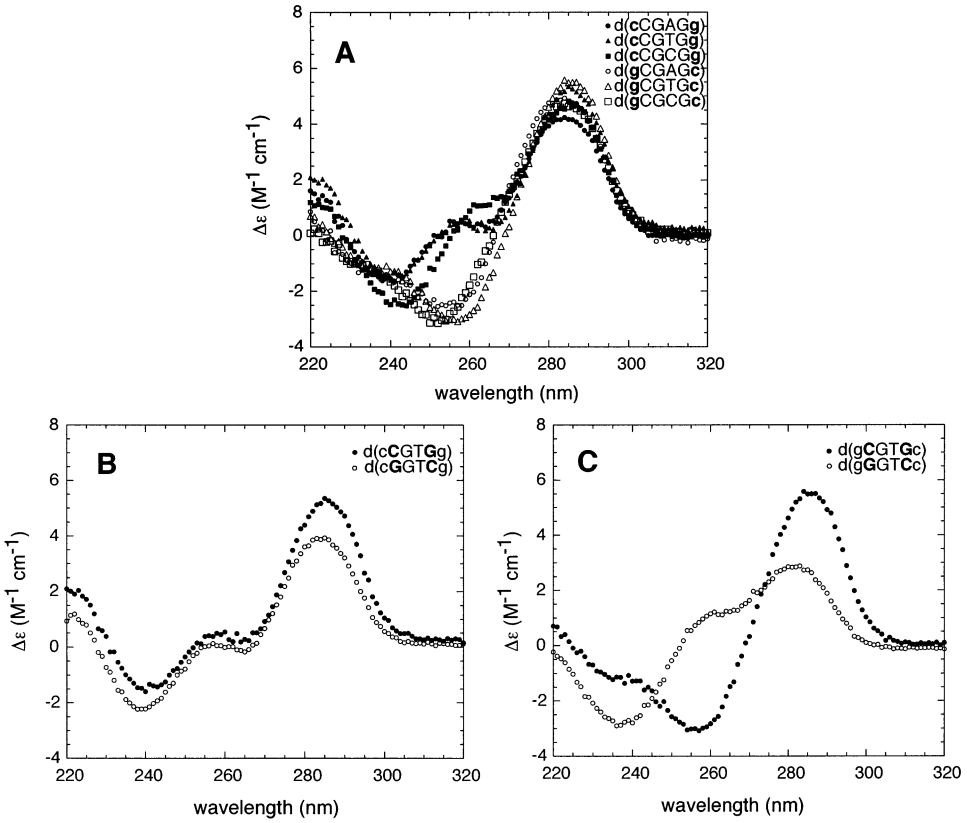

FIGURE 7: Comparison of CD spectra for the d(cCNNGg) and d(gCNNGc) motifs. (A) Effect of closing base pair changes. Loop sequences are as follows: d(CGAG) (circle), d(CGTG) (triangle), and d(CGCG) (square); closed and open symbols represent CG and GC closing base pairs, respectively. (B) Effect of swapping positions 1 and 4 of a d(cCNNGg) tetraloop. Loop sequences are as follows: d(cCGTGg) (●) and d(cGGTCg) (○). (C) Effect of swapping positions 1 and 4 of a gCNNGc tetraloop. Loop sequences are as follows: d(gCGTGc) (●) and d(gGGTCc) (○). Spectra were collected in $P_{10}E_{0.1}$ (pH 7.0) at 25 °C, and $C_T \approx 10$ μM.

closing base pair was GC (Figure 7A). Interestingly, these two sets of spectra differed from each other. While they both show a maximum at ≈285 nm, the spectra with a CG closing base pair have a shoulder near 255 nm and a minimum at ≈240 nm; the spectra with a GC closing base pair, on the other hand, have a minimum at ≈255 nm without a shoulder (Figure 7A). These differences occurred despite identical base composition. This suggests that the structures of these loops may change depending on the closing base pair, although we cannot judge the nature or extent of these differences at this time.

To investigate these differences in more detail, we compared the CD spectra for swapping positions 1 and 4 of the loop for d(cCGTGg) and d(gCGTGc). As mentioned above, this swap resulted in a more severe thermodynamic destabilization for d(gCGTGc) than d(cCGTGg). The CD spectra for d(cCGTGg) and d(cGGTCg) have very similar shapes (Figure 7B). This suggests that the position 1 and 4 swap does not change the structure of the loop significantly, consistent with the small destabilization observed for this swap. In contrast, the CD spectra for d(gCGTGc) and d(gGGTCc) have different shapes, with the maximum at 285 nm diminished and moved to ≈280 nm, the minimum shifted from 255 to ≈238 nm, and a shoulder appearing at 255 nm, much like the spectrum for the d(cCNNGg) sequences (Figure 7C). This result indicates a change in the structure of the loop, consistent with the significant destabilization observed for this swap.

DISCUSSION

The TGGE−selection method was originally developed as a way to isolate RNA sequences that have unusual thermodynamic stability for a given motif. The method has been successfully applied to RNA model systems and to triloop and tetraloop libraries (*38*−*40*). In vitro selection can also be applied to DNA. For example, DNA aptamers to a variety of ligands have been isolated (*42*, *53*), and catalytic DNA molecules have been evolved (*54*). We carried out in vitro selections on DNA libraries using TGGE and isolated stable DNA tetraloop sequences. These sequences should be useful for prediction of DNA structure from sequence and for understanding the importance of DNA stem−loops in nature.

One of the desired features of TGGE−selections is that the selection be exhaustive. We optimized the library and PCR conditions in order to recover known stable tetraloops including d(gCTTGc) and d(cGNNAg) sequences (*19*, *28*) and showed that selected unknown tetraloops are also thermodynamically stable when embedded in small hairpins. In addition, we are not aware of other stable DNA tetraloop sequences that were not found in the selection. Thus, it appears that the selection is exhaustive, or nearly so. This is desirable since one can now search for DNA tetraloop−closing base pair combinations in nature and know whether such sequences are thermodynamically stable. These sequences should help toward developing hypotheses on

biological functions of stable DNA stem−loops. Also, these sequences should help toward an understanding of why certain sequences are not found in nature; for example, thermodynamic stability may be undesirable for certain functions.

Of the tetraloop motifs found, three appear to be unknown or only partially studied. The d(cGNABg) motif appears to be closely related to the well-known d(cGNNAg) and d(cGNAg) motifs, both of which have been structurally and thermodynamically characterized (*19*, *29*−*31*). The d(cGNAg), d(cGNNAg), and d(cGNABg) motifs have a strong thermodynamic preference for a CG closing base pair (Tables 2−4) and the potential for a sheared GA base pair in the loop (Table 4). In addition, the CD spectra are similar for sequences from these motifs (Figure 4, Supporting Information). Both the d(cGNNAg) and d(cGNABg) motifs appear to be related to the d(cGNAg) motif through expansion of the triloop with an extra nucleotide. Similar behavior has been reported for RNA loops; for example, the r(cGNRAg) tetraloop motif can be expanded to a penta- or hexaloop motif (*55*).

The d(gCNNGc) and d(cCNNGg) motifs appear to be unrecognized in the literature, although a few individual sequences from the d(gCNNGc) motif have been characterized. The sequence, d(gCTTGc), is related to the RNA loop, r(gCUUGc), which is common in rRNA (*18*), and was found to be exceptionally stable (*19*). Most of the possible d(gCNNGc) sequences (14/16) were found in the present study, and these sequences are also exceptionally stable. In addition, the structures of variants related to d(gCNNGc) have been studied (see below).

The similarities and differences between stable RNA and DNA tetraloops are of interest. Tetraloops with sheared GA pairs occur in both cases, loops can be expanded with extra nucleotides, and CG closing base pairs can make unusual contributions to the stability of certain loops. Also, both RNA and DNA have a limited number of stable loops with GC closing base pairs; in these cases, C and G are required at positions 1 and 4 of the loop, respectively. The structures of these loops in RNA and DNA may be similar as well.

The structure of the r(gCUUGc) revealed Watson−Crick base pairing of the closing base pair and of the first and last bases in the loop, with the U at loop position 2 hydrogen bonding with these two base pairs in the minor groove (*56*). NMR studies on a hairpin with the loop, d(g^{a}CTAGc) (^{a}C = arabinofuranosylcytosine), also showed Watson−Crick base pairing of the closing base pair and of the first and last bases in the loop, with the loop T in the minor groove (*57*). Furthermore, NMR and model building suggested that the loop, d(aCTTGt), forms a similar structure (*49*), as does the loop of d(m^{5}CGm^{5}CGTGm^{5}CG) (m^{5} = 5-methylcytosine) (*58*). Therefore, it is possible that d(gCNNGc) sequences, in general, fold into a structure that is similar to that found for r(gCUUGc), although the details of these structures and loop position 2 interactions are presently unknown. Recently, we found examples of both r(cCNNGg) and r(gCNNGc) motifs in a selection on RNA tetraloops (*40*). As with the DNA tetraloops (Figure 7), the CD spectra were very different for RNA loops with the same CNNG loop but with CG versus GC closing base pairs (P. C. Bevilacqua and D. J. Proctor, unpublished data).

Differences also exist between RNA and DNA hairpin loops. The stable r(cUNCGg) tetraloops are not stable as DNA sequences (*19*). In addition, in RNA, r(GNRA) is a stable motif, and in DNA, d(GNNA) is a stable motif; apparently, this is because the N7 of the purine at position 3 of the r(GNRA) makes a hydrogen bond to the 2′OH of the G at position 1 of the loop (*21*, *60*). It is hoped that comparison of the properties of stable RNA and DNA hairpin loops will help to drive understanding of the structural and thermodynamic principles for hairpin folding.

DNA hairpins are found numerous times in nature and have a number of important biological functions (*1*). Others have pointed out that the d(cGNAg) motif occurs in regions of DNA implicated in transcriptional regulation in bacteriophage N4 double-stranded DNA (*61*) and that the d(cGNNAg) motif occurs in the replication origin of phage G4 single-stranded DNA (*62*). We have also found instances in which the d(cGNABg) motif occurs in nature. For example, the sequence repeat, d(GTAC)$_{n}$, has been shown to favor cruciform extrusion in plasmids (*63*). This sequence has the potential of forming a d(cGTACg) tetraloop motif in both strands of the cruciform. The d(cGTACg) sequence was found in our selection and was the most stable d(cGNABg) sequence melted (Tables 1 and 2). Although d(cGNAg) sequences are slightly more stable than d(cGNABg) sequences (Table 2), the latter motif can be palindromic and thus has the potential to drive the formation of stable loops in *both* strands of the DNA, which may be an important determinant for cruciform formation.

An additional occurrence of a d(cGNABg) motif was found. Habig and Loeb found that a small DNA hairpin negatively regulates priming during reverse transcription in avian hepatitis B virus (HBV) (*64*). In duck HBV, the small hairpin is capped with the stable triloop motif, d(cGAAg), while in heron HBV, the small hairpin is capped with the stable tetraloop motif, d(cGAATg), found twice in our selections (Table 1). Habig and Loeb found that disrupting the CG closing base pair with a CC or GG mismatch led to enhanced priming near this hairpin but that inverting this closing base pair to a GC only partially restored blocked priming (*64*). (The GC closing base pair was 2.2-fold poorer than the CG in negatively regulating reverse transcription.) The inability of the GC closing base pair to regulate reverse transcription to the full extent of the CG closing base pair is consistent with the poorer thermodynamic stability of GC closing base pairs for the d(GNAB) motif (Table 3). It should prove interesting to see if stable DNA tetraloops play other important roles in nature.

ACKNOWLEDGMENT

We thank Professor Doug Turner and members of the Bevilacqua laboratory for comments on the manuscript.

SUPPORTING INFORMATION AVAILABLE

One figure showing CD spectra for five d(cGNNAg) minihairpins. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Varani, G. (1995) *Annu. Rev. Biophys. Biophys. Chem. 24*, 379−404.

2. Tinoco, I., Jr., and Bustamante, C. (1999) *J. Mol. Biol. 293*, 271−281.
3. Glucksmann-Kuis, M. A., Dai, X., Markiewicz, P., and Rothman-Denes, L. B. (1996) *Cell 84*, 147−154.
4. Wadkins, R. M., Tung, C. S., Vallone, P. M., and Benight, A. S. (2000) *Arch. Biochem. Biophys. 384*, 199−203.
5. Lilley, D. M. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 6468−6472.
6. Paquin, B., Laforest, M. J., and Lang, B. F. (2000) *Mol. Biol. Evol. 17*, 1760−1768.
7. Choi, K. H., and Choi, B. S. (1994) *Biochim. Biophys. Acta 1217*, 341−344.
8. Bzymek, M., and Lovett, S. T. (2001) *Genetics 158*, 527−540.
9. Gellert, M. (2002) *Annu. Rev. Biochem. 71*, 101−132.
10. SantaLucia, J., Jr. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 1460−1465.
11. Allawi, H. T., and SantaLucia, J., Jr. (1997) *Biochemistry 36*, 10581−10594.
12. Allawi, H. T., and SantaLucia, J., Jr. (1998) *Biochemistry 37*, 2170−2179.
13. Allawi, H. T., and SantaLucia, J., Jr. (1998) *Nucleic Acids Res. 26*, 4925−4934.
14. Allawi, H. T., and SantaLucia, J., Jr. (1998) *Biochemistry 37*, 9435−9444.
15. Allawi, H. T., and SantaLucia, J., Jr. (1998) *Nucleic Acids Res. 26*, 2694−2701.
16. Peyret, N., Seneviratne, P. A., Allawi, H. T., and SantaLucia, J., Jr. (1999) *Biochemistry 38*, 3468−3477.
17. Bommarito, S., Peyret, N., and SantaLucia, J., Jr. (2000) *Nucleic Acids Res. 28*, 1929−1934.
18. Woese, C. R., Winker, S., and Gutell, R. R. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 8467−8471.
19. Antao, V. P., Lai, S. Y., and Tinoco, I., Jr. (1991) *Nucleic Acids Res. 19*, 5901−5905.
20. Cheong, C., Varani, G., and Tinoco, I., Jr. (1990) *Nature 346*, 680−682.
21. Heus, H. A., and Pardi, A. (1991) *Science 253*, 191−194.
22. Allain, F. H., and Varani, G. (1995) *J. Mol. Biol. 250*, 333−353.
23. Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) *J. Mol. Biol. 288*, 911−940.
24. Dale, T., Smith, R., and Serra, M. J. (2000) *RNA 6*, 608−615.
25. Senior, M. M., Jones, R. A., and Breslauer, K. J. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 6242−6246.
26. Paner, T. M., Amaratunga, M., Doktycz, M. J., and Benight, A. S. (1990) *Biopolymers 29*, 1715−1734.
27. Vallone, P. M., Paner, T. M., Hilario, J., Lane, M. J., Faldasz, B. D., and Benight, A. S. (1999) *Biopolymers 50*, 425−442.
28. Hirao, I., Nishimura, Y., Tagawa, Y., Watanabe, K., and Miura, K. (1992) *Nucleic Acids Res. 20*, 3891−3896.
29. Sandusky, P., Wooten, E. W., Kurochkin, A. V., Kavanaugh, T., Mandecki, W., and Zuiderweg, E. R. (1995) *Nucleic Acids Res. 23*, 4717−4725.
30. Hirao, I., Kawai, G., Yoshizawa, S., Nishimura, Y., Ishido, Y., Watanabe, K., and Miura, K. (1994) *Nucleic Acids Res. 22*, 576−582.
31. Yoshizawa, S., Kawai, G., Watanabe, K., Miura, K., and Hirao, I. (1997) *Biochemistry 36*, 4761−4767.
32. Zhu, L., Chou, S. H., Xu, J., and Reid, B. R. (1995) *Nat. Struct. Biol. 2*, 1012−1017.
33. Zhu, L., Chou, S. H., and Reid, B. R. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 12159−12164.
34. van Dongen, M. J., Mooren, M. M., Willems, E. F., van der Marel, G. A., van Boom, J. H., Wijmenga, S. S., and Hilbers, C. W. (1997) *Nucleic Acids Res. 25*, 1537−1547.
35. Wartell, R. M., Hosseini, S. H., and Moran, C. P., Jr. (1990) *Nucleic Acids Res. 18*, 2699−2705.
36. Ke, S. H., and Wartell, R. M. (1993) *Nucleic Acids Res. 21*, 5137−5143.
37. Zhu, J., and Wartell, R. M. (1997) *Biochemistry 36*, 15326−15335.
38. Bevilacqua, J. M., and Bevilacqua, P. C. (1998) *Biochemistry 37*, 15877−15884.
39. Shu, Z., and Bevilacqua, P. C. (1999) *Biochemistry 38*, 15369−15379.
40. Proctor, D. J., Schaak, J. E., Bevilacqua, J. M., Falzone, C. J., and Bevilacqua, P. C. (2002) *Biochemistry 41*, 12062−12075.
41. Latham, J. A., Zaug, A. J., and Cech, T. R. (1990) *Methods Enzymol. 181*, 558−569.
42. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H., and Toole, J. J. (1992) *Nature 355*, 564−566.
43. Chou, Q. (1992) *Nucleic Acids Res. 20*, 4371.
44. Borer, P. N. (1975) in *Handbook of Biochemistry and Molecular Biology: Nucleic Acids* (Fasman, G. D., Ed.) p 597, CRC Press, Cleveland, OH.
45. Richards, E. G. (1975) in *Handbook of Biochemistry and Molecular Biology: Nucleic Acids* (Fasman, G. D., Ed.) p 597, CRC Press, Cleveland, OH.
46. McDowell, J. A., and Turner, D. H. (1996) *Biochemistry 35*, 14077−14089.
47. Cantor, C. R., and Schimmel, P. R. (1980) in *Biophysical Chemistry, Part II: Techniques for the Study of Biological Structure and Function*, W. H. Freeman, San Francisco, CA.
48. Kierzek, R., Li, Y., Turner, D. H., and Bevilacqua, P. C. (1993) *J. Am. Chem. Soc. 115*, 4985−4992.
49. Blommers, M. J., Walters, J. A., Haasnoot, C. A., Aelen, J. M., van der Marel, G. A., van Boom, J. H., and Hilbers, C. W. (1989) *Biochemistry 28*, 7491−7498.
50. Haasnoot, C. A., Blommers, M. J., and Hilbers, C. W. (1987) *Springer Ser. Biophys. 1*, 212−216.
51. Tunis-Schneider, M. J., and Maestre, M. F. (1970) *J. Mol. Biol. 52*, 521−541.
52. Cantor, C. R., and Warshaw, M. M. (1970) *Biopolymers 9*, 1059−1077.
53. Ellington, A. D., and Szostak, J. W. (1992) *Nature 355*, 850−852.
54. Breaker, R. R., and Joyce, G. F. (1994) *Chem. Biol. 1*, 223−229.
55. Abramovitz, D. L., and Pyle, A. M. (1997) *J. Mol. Biol. 266*, 493−506.
56. Jucker, F. M., and Pardi, A. (1995) *Biochemistry 34*, 14416−14427.
57. Pieters, J. M., de Vroom, E., van der Marel, G. A., van Boom, J. H., Koning, T. M., Kaptein, R., and Altona, C. (1990) *Biochemistry 29*, 788−799.
58. Orbons, L. P., van Beuzekom, A. A., and Altona, C. (1987) *J. Biomol. Struct. Dyn. 4*, 965−987.
59. Hirao, I., Kawai, G., Kobayashi, K., Nishimura, Y., Miura, K., Watanabe, K., and Ishido, Y. (1992) *Nucleic Acids Symp. Ser. 27*, 127−128.
60. Jucker, F. M., Heus, H. A., Yip, P. F., Moors, E. H., and Pardi, A. (1996) *J. Mol. Biol. 264*, 968−980.
61. Dai, X., Kloster, M., and Rothman-Denes, L. B. (1998) *J. Mol. Biol. 283*, 43−58.
62. Hirao, I., Ishida, M., Watanabe, K., and Miura, K. (1990) *Biochim. Biophys. Acta 1087*, 199−204.
63. Naylor, L. H., Yee, H. A., and van de Sande, J. H. (1988) *J. Biomol. Struct. Dyn. 5*, 895−912.
64. Habig, J. W., and Loeb, D. D. (2002) *J. Virol. 76*, 980−989.
65. Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.

BI026479K