Folding of a DNA Hairpin Loop Structure in Explicit Solvent Using Replica-Exchange Molecular Dynamics Simulations

Srinivasaraghavan Kannan and Martin Zacharias
School of Engineering and Science, Jacobs University Bremen, D-28759 Bremen, Germany

ABSTRACT Hairpin loop structures are common motifs in folded nucleic acids. The 5′-GCGCAGC sequence in DNA forms a characteristic and stable trinucleotide hairpin loop flanked by a two basepair stem helix. To better understand the structure formation of this hairpin loop motif in atomic detail, we employed replica-exchange molecular dynamics (RexMD) simulations starting from a single-stranded DNA conformation. In two independent 36 ns RexMD simulations, conformations in very close agreement with the experimental hairpin structure were sampled as dominant conformations (lowest free energy state) during the final phase of the RexMDs (~35% at the lowest temperature replica). Simultaneous compaction and accumulation of folded structures were observed. Comparison of the GCA trinucleotides from early stages of the simulations with the folded topology indicated a variety of central loop conformations, but arrangements close to experiment that are sampled before the fully folded structure also appeared. Most of these intermediates included a stacking of the C_2 and G_3 bases, which was further stabilized by hydrogen bonding to the A_5 base and a strongly bound water molecule bridging the C_2 and A_5 in the DNA minor groove. The simulations suggest a folding mechanism where these intermediates can rapidly proceed toward the fully folded hairpin and emphasize the importance of loop and stem nucleotide interactions for hairpin folding. In one simulation, a loop motif with G_3 in *syn* conformation (dihedral flip at *N*-glycosidic bond) accumulated, resulting in a misfolded hairpin. Such conformations may correspond to long-lived trapped states that have been postulated to account for the folding kinetics of nucleic acid hairpins that are slower than expected for a semiflexible polymer of the same size.

INTRODUCTION

Hairpin loop structures in nucleic acids consist of a base-paired stem structure and a loop sequence with unpaired or non-Watson-Crick-paired nucleotides. These common structural motifs can be of functional importance as ligand recognition elements or folding initiation sites. A number of trinucleotide sequences at the center of palindromic sequences in DNA can form compact and stable hairpin loops (1–11). Formation of stable DNA hairpin structures can influence supercoiling of DNA and DNA replication and transcription (6,7,12–14). It has been proposed that hairpin formation of triplet repeat sequences during DNA replication could play a role in the expansion of such repeats associated with several genetic diseases (15–20).

Hairpin loops with a central GNA trinucleotide motif (G, guanine; N, any nucleotide; A, adenine) have been found to form particularly stable structures (1,8–11,20–22). For example, for the sequence 5′-GCGCAGC, a melting transition for disruption of the hairpin structure of 67°C has been reported (8). The thermodynamic stability of the GCA trinucleotide loop, the influence of loop expansion, and the influence of closing and flanking sequences have been characterized extensively (1,3,8–11). In addition, structural studies using NMR spectroscopy have revealed a characteristic compact folding topology for the GNA-loop (1,3) with a B-DNA form stem, a sheared G:A loop closing basepair, and the central loop base stacking on top of the G:A basepair pointing toward the major groove. Several studies on base modifications allowed us to elucidate the contribution of individual hydrogen bonds and other nonbonded contacts to folding stability (9–11). However, the molecular mechanism of DNA hairpin structure formation and characterization of possible stable intermediate states has so far not been possible experimentally.

Due to the small size and characteristic fold, DNA trinucleotide motifs are well suited for theoretical and computational studies on loop structure and dynamics. DNA trinucleotide hairpin loops have been investigated in multistart energy minimization (23) and conformational scanning search (24) approaches employing a generalized Born-type implicit solvent model to characterize possible stable conformational substates. In principle, molecular dynamics (MD) simulations are well suited to follow the structure formation process of structural motifs in nucleic acids. However, the accessible timescale and sampling efficiency strongly limits the usefulness of standard MD simulations to study nucleic acid structure formation processes. Formation of hairpin loops in DNA has been found to occur on the order of microseconds (depending on DNA length and sequence) beyond current maximum MD simulation timescales (25–29). Interestingly, the kinetics of nucleic acid hairpin folding can display non-Arrhenius temperature dependence following multiple transition rates (25–29).

Submitted March 2, 2007, and accepted for publication June 22, 2007.

Address reprint requests to Martin Zacharias, Jacobs University Bremen, School of Engineering and Science, Campus Ring 1, D-28759 Bremen, Germany. Tel.: 49-421-200-3541; Fax: 49-421-200-3249; E-mail: m.zacharias@jacobs-university.de.

Editor: Tamar Schlick.

0006-3495/07/11/3218/11 $2.00

doi: 10.1529/biophysj.107.108019

This might be due to formation of transiently trapped misfolded states that follow different transition kinetics toward the folded state (26,29). So far, multiple MD simulations starting from thousands of different start structures have been used to observe folding transitions of RNA tetraloop structures with the central GCAA sequence that forms a characteristic RNA structural motif (30–32). In a very small fraction of the total number of simulations (19 out of 10,000), folding transitions to near-native structures were observed (32). Such simulation studies are very useful to characterize the rapid transition from a few starting conformations to the folded form and to estimate the folding rate (and mean folding time). However, without prior knowledge of the native folded structure, it is not possible to select those simulation events that lead to native structure formation. With only a very small fraction of simulations resulting in near-native structures, it is also not possible to identify this state as the most favorable conformational state (with lowest free energy).

To overcome the sampling limitations of standard MD methods, we employed the replica-exchange MD simulation methodology (RexMD) (33–35) in explicit solvent to study structure formation of the 5′-GCGCAGC motif in DNA. During RexMD simulations, several replicas of a system are simulated at different temperatures in parallel, allowing for exchanges between replicas at frequent intervals (33–35). This technique allows significantly improved sampling of conformational space and has already been used for folding simulations and structure prediction of peptides and small proteins (35–38) and the analysis of dinucleotide stacking in DNA (39–41) but so far much less to study the dynamics of DNA oligonucleotides.

Two independent RexMD simulations were started from single-stranded nucleic acid conformations using different starting conditions and using 16 replicas ranging in temperature from 315 K to 425 K. Both simulations lead to conformations in very close agreement with the experimental hairpin loop structure as the final dominate state with highest population at the replica run with the lowest temperature. Cluster analysis of structures sampled at early and later stages during the simulations allowed us to characterize stable intermediate states accessible during the structure formation process. The simulations indicate that the characteristic loop motif with a sheared guanine:adenine (G:A) basepair and not fully formed stem basepairs can occur at an early stage of the simulations followed by a rapid subsequent formation of the stem basepairs. In one of the two RexMD simulations an alternative loop motif with the loop guanine base in a *syn* conformation (which corresponds to an altered dihedral state around the *N*-glycosidic sugar-base bond compared to the more common anticonformation) was formed and accumulated to some degree as a stable alternative loop structure. This misfolded structure may correspond to a transiently trapped state that has to undergo partial or complete unfolding to form the ‘‘correctly’’ folded structure and may correspond to a fraction of slowly folding hairpins.

The article is organized as follows. We first compare sampled DNA conformations during continuous MD and RexMD and analyze the accumulation of near-native folded DNA hairpins during independent RexMD simulations. In the following paragraphs the accumulation of intermediates and misfolded sampled conformations is analyzed to determine which intermediates contribute productively to the folding process. Finally, the accumulation of near-native structures over time and at different temperatures has been investigated. The simulation results demonstrate that advanced sampling methods based on current force fields and including explicit solvent and ions allowed the folding of stable DNA hairpin loop structures, in close agreement with experiment and as the dominant conformational state (of lowest free energy). The relatively modest computational demand may allow us to systematically study the sequence dependence of hairpin folding and the characterization of stable intermediate structure.

MATERIALS AND METHODS

RexMD simulations were started from an extended single-stranded DNA structure of the sequence 5′-GCGCAGC. The start structure was generated using the *Nucgen* program of the Amber8 (Assisted Model Building with Energy Restraints, (42)) program package with a B-DNA-type geometry followed by energy minimization. Initial positions of 6 K^+ counterions were placed using the xleap module of the Amber8 package. The structure was solvated in an octahedral box with 1127 TIP3P water molecules (43) leaving at least 10 Å between solute atoms and the borders of the box. This corresponds to an ion concentration of ~200 mM.

Initial energy minimization (2500 steps) of the solvated systems was performed with the *sander* module of the Amber8 package and using the parm99 force field (44). After minimization the system was gradually heated from 50 to 300 K with positional restraints (force constant: 50 kcal mol^{-1} $Å^{-2}$) on DNA over a period of 0.25 ns, allowing water molecules and ions to move freely. A 9 Å cutoff for the short-range nonbonded interactions was used in combination with the particle mesh Ewald option (45), using a grid spacing of ~0.9 Å to account for long-range electrostatic interactions. The Settle algorithm (46) was used to constrain bond vibrations involving hydrogen atoms, and a time step of 1 fs was used during RexMD simulations (2 fs for standard MD). During an additional 0.25 ns the positional restraints were gradually reduced to allow final unrestrained MD simulation of all atoms over a subsequent equilibration time of 2 ns. This procedure was repeated for the same starting structure using different randomly assigned initial atom velocities.

The replica-exchange simulations were conduced under constant volume using 16 replicas. An exponentially increasing temperature series along the replicas was used which gives approximately uniform acceptance ratios for exchanges between neighboring replicas (37) with the following simulation temperatures (in Kelvin): 315.0, 317.0, 320.6, 324.8, 329.6, 335.0, 341.0, 347.6, 354.8, 362.6, 371.0, 380.0, 389.6, 399.8, 410.6, and 422.0. These simulation temperatures resulted in exchange probabilities between neighboring replicas of ~20% (attempted exchanges every 750 steps). Both RexMD simulations A and B were continued for 36 ns. For comparison, two standard 75 ns MD simulations starting from the same start structure but different initial atomic velocities were run at 330 K (same starting conformation as for RexMD simulations).

An experimental high-resolution structure of the GCA trinucleotide loop is only available in the context of two flanking T:A basepairs (PDB entry: 1ZHU) (3). A reference structure for comparison with the current simulation results (with the sequence 5′-GCGCAGC) was constructed by isosterical replacement of the T:A basepairs (in the first structure of the 1ZHU entry) by G:C stem basepairs using the program Jumna (47). The structure was energy minimized (1000 steps) to remove any residual sterical clashes which resulted in only very small changes from the experimental loop structure (Rmsd < 0.4 Å).

Cluster analysis was based on the pairwise Cartesian Rmsd (only heavy atoms) between conformations with an Rmsd cutoff of 2 Å and using the kclust program in the MMTSB-tools (48). The visual molecular dynamics program (49) was used for visualization of trajectories and preparation of figures.

RESULTS AND DISCUSSION

Conformational flexibility of single-stranded DNA during continuous MD simulations

Both continuous and replica-exchange (Rex)MD simulations were started from single-stranded 5′-GCGCAGC DNA molecules in a stacked B-type conformation with different initial velocity assignments. This type of start structure was chosen since there is experimental evidence that especially purine-rich single-stranded DNAs adopt stacked structures in solution as dominant conformational states (50–53). The 5′-GCGCAGC sequence adopts a very stable GCA trinucleotide hairpin loop structure flanked by two G:C Watson-Crick basepairs in solution that has been investigated using NMR spectroscopy (1,3,8–11). However, an experimental high-resolution structure of the GCA trinucleotide loop is only available in the context of two flanking T:A basepairs (PDB entry: 1ZHU) (3). A reference structure for comparison with the current simulation results (with the sequence 5′-GCGCAGC) was constructed by isosterical replacement of the two T:A basepairs by corresponding G:C basepairs using the program Jumna (47) followed by a short energy minimization (see Materials and Methods).

The dynamics and stability of the single-stranded start conformation was first investigated during two independent 75 ns standard MD simulations at 330 K started with different initial atom velocities. An elevated simulation temperature slightly below the expected hairpin melting temperature (~340 K) was chosen because it should accelerate conformational transitions including those to the native structure compared to simulations at room temperature. The generated DNA structures showed considerable fluctuations with significant deviations from the start conformation (Fig. 1). Structural transitions included several unstacking events along the single-stranded DNA, in particular at the termini of the nucleic acid molecule (not shown). However, no folding transitions to a structure close to the experimental hairpin loop conformation were observed. The root mean-square deviation (Rmsd) from the reference hairpin structure (heavy atoms) remained around 5–8 Å in both simulations over the entire simulation time.

Hairpin structure formation during replica-exchange MD simulation

During the RexMD simulations, the initial Rmsd from the experimental hairpin structure was ~7 Å and started to decrease at around 5–7 ns in the lowest temperature replica run (Fig. 2). At a simulation time of ~9 ns and 12 ns during simulations A and B, respectively, conformations with an Rmsd of ~2 Å from experiment were sampled. After ~15–20 ns simulation time conformations as close as 1.2–1.6 Å (heavy atoms) with respect to the reference hairpin conformation were sampled as the dominant conformational states (Fig. 2). These structures show the same characteristic arrangement of loop and stem bases and the same hydrogen (H-)bonding pattern as the experimental structure of the GCA loop motif (Fig. 3). The Rmsd probability distributions at the various stages of the simulations (Fig. 2) indicate that in the final stage of both 36 ns RexMD simulations conformations within an Rmsd of 2 Å from the reference structure accounted for 35% (simulation A) and 40% (simulation B) of sampled conformations, respectively. Comparison with the earlier stages of the simulation showed that in both simulations the fraction of native-like conformations increased over time with a dramatic difference between early and middle part of the simulation and only a modest change during the final stage of both simulations (Fig. 2).

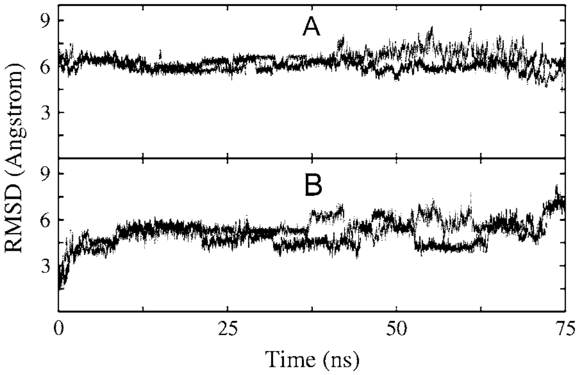

FIGURE 1 Heavy atom Rmsd of sampled DNA conformations (5′-GCG-CAGC) from (*A*) folded hairpin structure and (*B*) single-stranded start structure versus simulation time. Results are shown for two independent 75 ns simulations starting from the same single-stranded DNA with different initial atomic velocity assignments (*dotted* and *continuous lines*, respectively).

Interestingly, in simulation A cluster analysis of the final part of the trajectory (lowest temperature replica) indicated a significantly populated cluster of conformations relatively close to the experimental trinucleotide hairpin structure (Rmsd ~2.5–3 Å, ~15% of sampled conformations) but with the G_3 nucleotide in the *syn* conformation (Fig. 3 *C*) instead of the regular anticonformation at the *N*-glycosidic bond (bond between sugar and base). Such *syn* conformations are frequently found in the case of purin bases in folded RNA structures (e.g., UNCG hairpins, (54,55)). However, for the present loop structure the *syn*-G_3 conformation allows for stacking interactions with neighboring bases but prevents formation of stable H-bonds with the A_5 as seen in the sheared basepair arrangement of the native loop conformation (Fig. 3 *A*). *Syn*-G_3 conformations were also observed in simulation B, however, mainly during the first part of the simulation (at least in the lowest temperature replica) lacking the basepaired stem

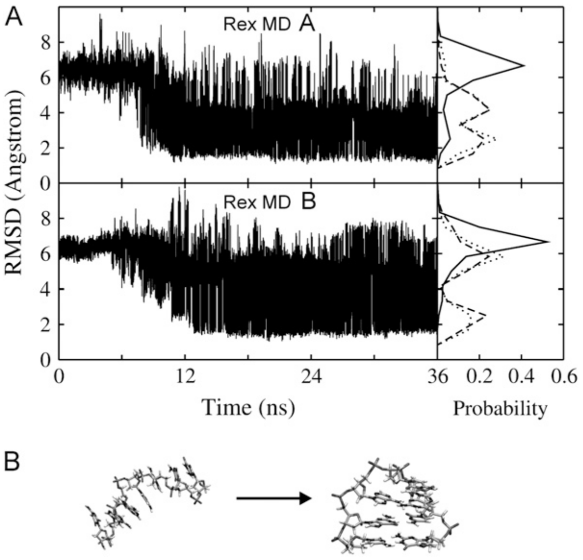

FIGURE 2 (*A*) Rmsd (heavy atoms) of the 5′-d(GCGCAGC) conformations (from lowest temperature run of each RexMD simulation) with respect to the folded hairpin reference structure versus simulation time. The panel on the right of each Rmsd plot corresponds to the Rmsd probability distribution during the first (*continuous line*), second (*dashed line*), and last (*dotted line*) 12 ns of each simulation. (*B*) Single-stranded start structure and fully folded hairpin loop structure (sampled as the dominant state of both simulations after ~20 ns).

and no significant accumulation of completely folded hairpin loops (with a *syn*-G_3). It indicates that a ‘‘misfolded’’ trinucleotide loop with a *syn*-G_3 once it has formed a complete basepaired stem structure corresponds to a long-lived trapped structure that can only refold to the native-loop structure after complete unfolding of the stem region. Hence, it is separated from the native structure by a large energy barrier that even in a RexMD simulation requires significantly longer simulation times (than the present 36 ns) to completely disappear in the final conformational ensemble.

This result suggests the possibility that such *syn* conformations of nucleobases may also form during other structure formation processes of nucleic acids (e.g., double-strand formation) and may in general result in long-lived trapped misfolded structures. It is also integrated with the observation that hairpin formation is overall slower than expected from estimated end-to-end contact formation of a semiflexible polymer and may be characterized by multiple rates due to the formation of long-lived trapped states (26,29).

Accumulation of intermediates and misfolded structures

A variety of nucleic acid conformational states were sampled during the RexMD simulations. Cluster analysis was performed for conformations formed during the first, second, and third intervals (each 12 ns) of both simulations (a cluster represents structures within an Rmsd of 2 Å from the cluster center). During the first 12 ns the dominant cluster was in both simulations formed by conformations close to the stacked singled-stranded state (not shown). Other significantly populated clusters included single-stranded conformations with kinks (unstacking) at various positions along the DNA and structures that started to form compact states near the 5′- or the 3′-ends of the DNA chain (representative structures are shown in the first row of Fig. 4). Characteristic for most of the sampled states are stretches of stacked bases ranging from 2 to 4 consecutive nucleotides. Even during this first phase (12 ns) of the simulations the near-native structures formed a significantly populated cluster (structures illustrated in Fig. 3 *B*).

The last 24 ns (phase II and III) in both simulations were already dominated by conformations close to the native folded hairpin structure (forming the highest populated cluster). However, several alternative compact states were also sampled that included kink turns at various positions along the DNA molecule. A subset of conformations close to the average structures (cluster centers) of clusters populated with at least 1% of all recorded conformations is shown in Fig. 4. Several of these partially folded structures contained structural elements that are similar to elements in the native folded structures (e.g., a topological arrangement of the central trinucleotide loop similar to the arrangement in the native structure; see next paragraph). However, several other conformational clusters indicate stacking and basepair arrangements that strongly deviate from the native structure (lower two rows of Fig. 4) and are presumably (indicated by the low population) of higher free energy than structures close to the native state.

Due to the exchanges with neighboring replicas in the RexMD simulation, the conformations at one temperature do not represent continuous trajectories. However, it is possible to look at the pattern and accumulation of conformations that occur before any native-like folded hairpin structure first appears. Structures with a low Rmsd with respect to the trinucleotide hairpin loop motif alone (only the central three nucleotides) appeared at an earlier stage of both RexMD simulations than structures with the native-like stem structure

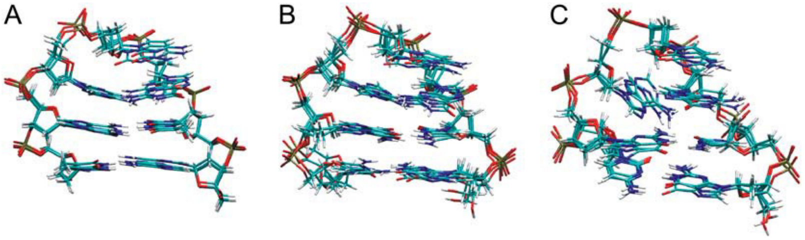

FIGURE 3 Comparison of an ensemble of NMR structures of the (*A*) GCA trinucleotide loop (four structures of PDB entry: 1ZHU; sequence: 5′-dATGCAAT) and four randomly selected structures obtained during the final stage of the (*B*) RexMD simulation A with a heavy atom Rmsd of <2 Å from the folded reference hairpin structure. (*C*) Superposition of ‘‘misfolded’’ DNA hairpin structures with the loop guanine (G_3) in a *syn* conformation and the loop adenine (A_5) partially stacked in the DNA minor groove.

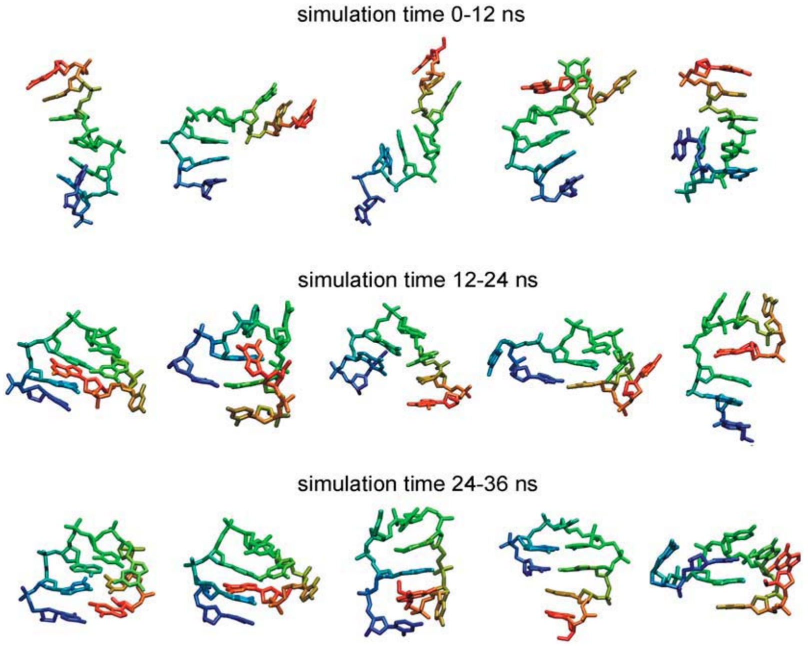

FIGURE 4 Representative structures (*stick representation*) of conformational clusters obtained during three different phases of the RexMD simulations. Each structure corresponds to a conformation closest to the average structure of a cluster (cluster centroid) with a cluster population around or above 1% of all recorded structures during the corresponding time interval. Cluster analysis was performed with an Rmsd cutoff of 2 Å and using the kclust program of the MMTSB package (48). The color in the stick representation goes gradually from red (5′-DNA end) to blue (3′-DNA end) to get an impression of the chain orientation. For clarity, hydrogen atoms have been omitted.

(Fig. 5). However, the delay time between trinucleotide loop formation and first occurrence of conformations with correctly formed loop and stem was only ~1 ns in the case of the simulation A. It amounted to ~4 ns in the second RexMD simulation (Fig. 5). The accumulation of intermediate native-like trinucleotide loop structures with varying conformations of the stem nucleotides (Fig. 5 *B*) is consistent with negative free energy estimates of −0.4 to −0.3 kcal/mol for GCA loop formation alone (after subtraction of the stem contribution) by Yoshizawa et al. (8).

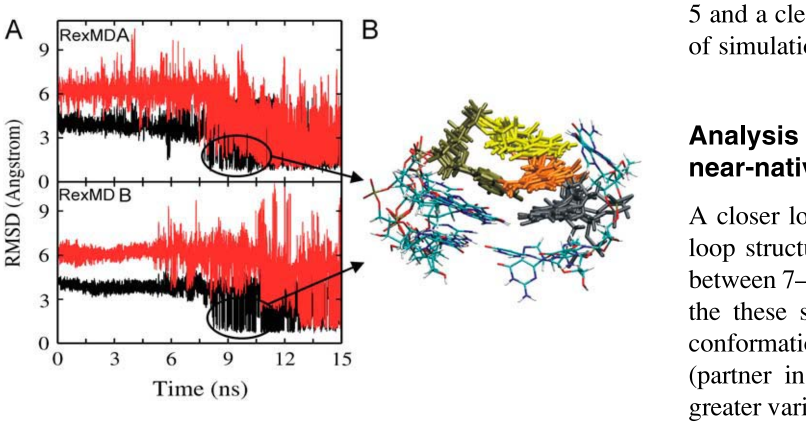

FIGURE 5 (*A*) Rmsd of sampled conformations (during lowest temperature run) with respect to the native trinucleotide loop structure (only of the three central nucleotides, in *black*) and with respect to the stem structure of the folded hairpin conformation (considering only the two stem basepairs, in *red*). (*B*) Superposition of five conformations obtained during the 7–10 ns simulation time interval with near-native trinucleotide loop structure but not correctly formed stem structure. Loop nucleotides C_2 (*gray*) to A_5 (*green*) are shown as bond sticks and using color coding according to residue number.

Note that the estimated loop formation free energy of most sequences is positive. For example, even the well-known UNCG loop in RNA (54,55) has a positive free energy of formation (~1 kcal/mol after subtraction of the stem contribution; 56,57). A two-dimensional (2D) plot of the trinucleotide loop Rmsd from the native loop structure versus Rmsd of the stem with respect to the native structure indicates that at no stage of both simulations was a native-like stem structure observed without formation of a near-native loop structure (Fig. 6). The plot indicates for RexMD simulation A an almost simultaneous loop and stem formation consistent with the short delay between loop and stem formation seen in Fig. 5 and a clearer separation of both folding events in the case of simulation B.

Analysis of intermediate structure with near-native loop structure

A closer look at sampled conformations with a near-native loop structure (but still incorrect stem) in the time interval between 7–10 ns of both simulations indicates that in most of the these structures the C_2 residue is in a stable stacked conformation with respect to the G_3 base. The opposing G_6 (partner in the fully folded hairpin loop) adopts a much greater variety of conformations (illustrated in Fig. 5 *B*). The reduced mobility of C_2 (compared to, for example, the G_6) is likely due to favorable stacking interactions with G_3 but also due to the A_5 nucleotide. In conformations near the native loop topology, the A_5 base frequently contacts the G_3 (correct H-bonding partner in the native loop structure) but also, frequently, the C_2 base (located below the G_3 in a stacked arrangements) and in some conformations both bases.

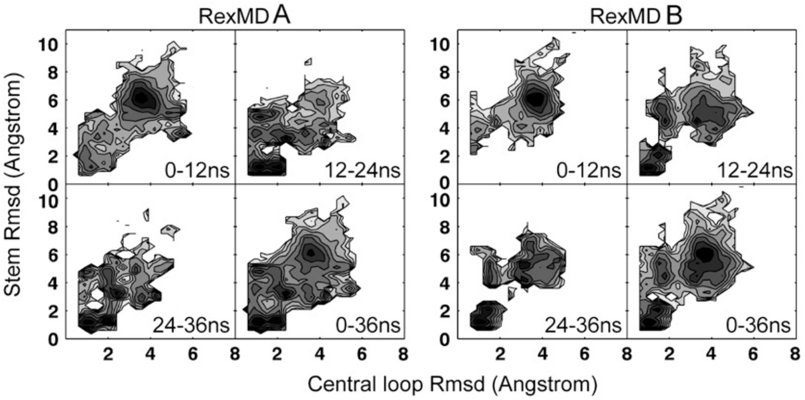

FIGURE 6 Deviation of the central three nucleotides (*x* axis) and four stem nucleotides (*y* axis) from the folded reference DNA hairpin structure during four different time intervals of the RexMD simulations. Dark/light regions in the 2D plots indicate a high/low probability, respectively, for a given pair of central loop and stem Rmsds.

Interestingly, the analysis of the distribution of solvent molecules revealed one site in the minor groove of the DNA where a frequently bound water molecule bridges the C_2 and A_5 bases (forming simultaneous H-bonds with the O_2 of the C_2 base and N_1 of the A_5 base; Fig. 7). This water molecule was found in >90% of all recorded structures with a near-native loop structure (but not necessarily fully formed stem). The high occupancy of the bridging water molecule indicates that solvent may have a specific role in stabilizing the topologically ''correct'' hairpin loop motif. Three of such topologically almost correctly folded trinucleotides loop motifs are shown in Fig. 8. Apparently, during the folding process the stacking of C_2, G_3, (and probably also C_4) and the bridging water molecule in the minor groove are important to provide a stable template for the A_5 to search for the ''correct'' H-bonding partner during loop formation. Conversely, the C_2-G_3 stacking is stabilized by H-bond formation of the A_5 with C_2 or both C_2 and G_3. The importance of the C_2-G_3 stacking, as indicated in the simulation here, is supported by the experimental observation that the stability and folding of GNC trinucleotide loops is especially sensitive to the destabilization of C_2-G_3 interactions (9). The introduction of a three-carbon linker between C_2 and G_3 that mimics the insertion of one nucleoside (without a base) increases the distance between the bases and disturbs the C_2-G_3 interactions and has a strongly destabilizing effect on loop formation (by ~1.6 kcal mol^{-1}) (9). Insertion of the same linker at other positions in the loop has only a minor effect on loop formation (9).

In Fig. 8 near-native loop motifs that were observed shortly before the appearance of the first near-native folded hairpin loops (including the stem) are compared with alternative ''misfolded'' loop structures that cannot directly proceed toward the correctly folded structure. An exception is the already mentioned loop motif with a *syn*-G_3 conformation that is also sterically compatible with a progression toward a fully folded hairpin structure (Fig. 8) and provides at least favorable stacking interactions of the loop bases (but not the native H-bonds as seen in the sheared G:A basepair). In the 2D plot of the trinucleotide loop versus stem Rmsd (Fig. 6 *A*), this conformational state in the case of simulation A also shows up as a second peak close to the peak that corresponds to the native-like state with a slightly larger Rmsd of the loop segment from experiment compared to the native-like structure. Comparison of different time intervals of the simulation indicates that the *syn* conformation of the loop adenine results in a relatively stable ''trapped'' and nonnative hairpin loop structure.

Since on the timescale of the RexMD simulations the population did not significantly change within the last ~20 ns, this nonnative hairpin loop structure may have a low free

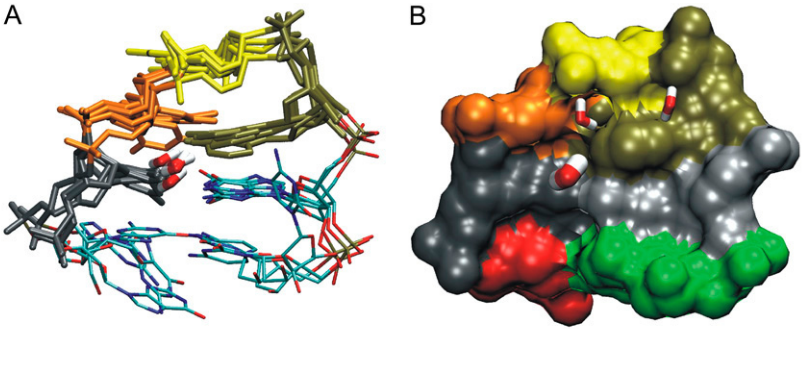

FIGURE 7 Specific water binding to the hairpin loop motif in the DNA minor groove. (*A*) Superposition of four sampled structures with the near-native trinucleotide loop structure and a water molecule bridging the O_2 atom of C_2 (*gray*) and the N_1 atom of the A_5 (*green*) nucleobase. A water molecule was found at this position in more than 90% of the recorded conformations where the loop had correctly formed. The view is into the minor groove and using the same color coding as in Fig. 5. (*B*) Accessible surface area representation of one simulation snapshot (color coding of residue numbers) with a bound water molecule bridging C_2 (*gray*) and A_5 (*bold bond stick model*). Two minor water binding sites (*thin bond stick water model*) bridging phosphate groups and the A_5 base (occupancy ~40% in recorded conformations with a native-like trinucleotide loop structure) are also indicated.

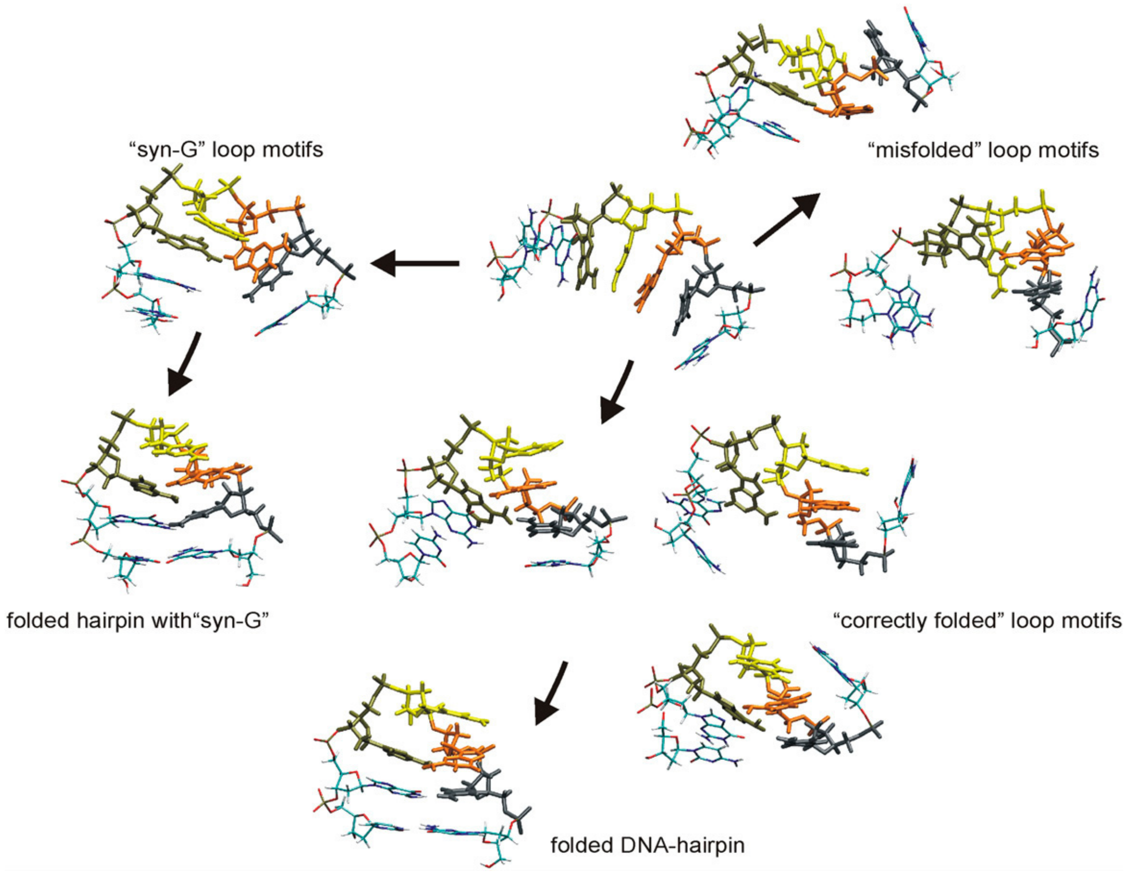

FIGURE 8 Folding intermediates of the DNA-trinucleotide hairpin loop. Each of the snapshots from various stages of the RexMD simulations contains a frequently found structural motif of the central nucleotides (color coded and using *bold sticks*). ''Correctly folded'' loop motifs correspond to a similar helical arrangement of the central loop nucleotides as the native hairpin structure. These intermediates are likely to rapidly progress toward the fully folded conformation. The *syn*-G_3 loop motif is sterically also compatible with a fully folded hairpin, but it retains the misfolded helical arrangement of the central loop nucleotides. ''Misfolded loop'' motifs strongly deviate from the native trinucleotide loop structure (only a few examples are shown) and are unlikely to progress rapidly toward a fully folded hairpin structure.

energy similar to the native state. This likely was an artifact of the simulation force field since in the experimental structure of the GCA trinucleotide loop such a *syn*-G_3 conformation is not observed. However, it is also possible that the ''refolding'' to a conformation with an anti-G_3 conformation requires the complete unfolding of the hairpin loop since for sterical reasons the compact hairpin loop structure does not allow the transition to an anticonformation in the compact folded form. The RexMD simulation in principle allows for such transitions due to the replica exchanges. Indeed, at the higher temperature replicas, single-stranded DNA conformations are significantly populated throughout the whole simulation (Fig. 9). However, in a RexMD simulation stable trapped conformations once formed do not disappear but can only evolve toward native-like structures by ''traveling'' along the temperature coordinate to overcome energetic barriers. Due to the thermodynamic stability of the alternative hairpin loop structure, complete unfolding toward a single-stranded structure that allows for *syn* antitransitions even during the RexMD is a rare event and may require much longer simulation timescales to reach a fully equilibrated probability distribution of sampled conformations.

Temperature dependence of hairpin loop stability

The population of native-like structures during the simulations varies between different stages of the simulations. However, in both simulations the accumulated fraction of near-native DNA hairpin conformations (within 2 Å of the reference structure) at the lowest temperature replica approaches ~35% (Fig. 9). In a fully equilibrated simulation the population at the lowest temperature is expected to be much higher because it is significantly below the hairpin melting temperature. The fraction depends on the Rmsd cutoff to distinguish between folded and unfolded structures (~45% if one chooses an Rmsd cutoff of 2.5 Å). This suggests that the hairpin folding free energy at the lowest temperature replica (42°C) is close to zero.

The experimental folding free energy from calorimetric studies for the same sequence is, however, $\Delta G_{\text{fold}} = -2.7$

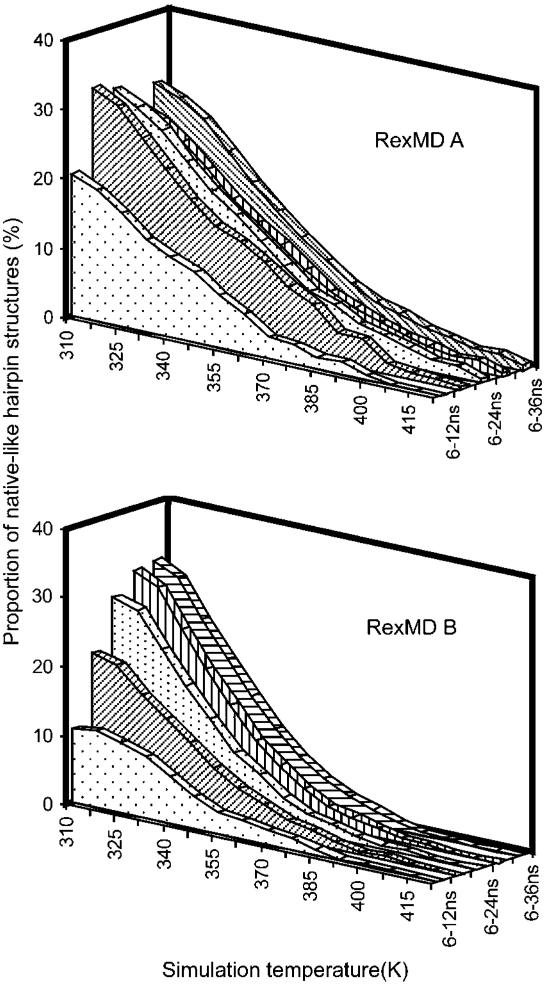

FIGURE 9 Contribution of native-like hairpin loop structures (within an Rmsd of <2.0 Å of the folded reference structure) at various stages of the RexMD simulations (indicated by different plot textures). Contributions are given as percentage of the total ensemble at each replica simulation temperature.

kcal mol^{-1} (in 1 M NaCl at 37°C; with little changes in the melting behavior at 0.1 M and 1 M NaCl, (8)). The RexMD simulations on the timescale here clearly underestimate the fraction of native-like loop conformations at the lowest temperature replica. In principle, it is possible to use the fraction of near-native hairpin structures from all simulation temperatures (all replicas) to extract thermodynamic quantities. However, in addition to the possibility of insufficient convergence, one needs to also keep in mind that inaccuracies of the force field and water model (designed for room temperature simulations) are likely to have a significant impact at the higher simulation temperatures.

Nevertheless, the overall shape of the population curve looks similar for the different time intervals, and it is possible to extract the temperature at which the level of near-native conformations has dropped to half of the lowest temperature level (melting temperature). This results in a rough estimate of the melting temperature of ~340–350 K (67°C–77°C) quite close to the experimental melting temperature of 67°C (8). A van 't Hoff analysis of the change in near-native population versus temperature results in a $\Delta H_{\rm fold} \approx -10$ kcal mol^{-1}. For monomolecular processes such as hairpin formation and assuming a two-state unfolding-folding transition and no temperature dependence of $\Delta H_{\rm fold}$, one can estimate $\Delta G_{\rm fold}(T) = \Delta H_{\rm fold}\,(1 - T/T_{\rm m}) \approx -0.9$ kcal mol^{-1}at 37°C. The magnitude of the calculated $\Delta H_{\rm fold}$ is ~3 times smaller than the experimental $\Delta H_{\rm fold}$ (−30.4 kcal mol^{-1}).

The discrepancy is due to an ''underestimation'' of the population at near-native structures at the low temperature replicas and/or an overestimation of the population of near-native structures at the higher temperature replicas. Insufficient conformational sampling but also force field artifacts, especially at the higher simulation temperatures as discussed above, are likely reasons for the discrepancy. It should be emphasized that the simulations here demonstrate that the force field approach is sufficiently accurate to generate near-native DNA hairpin structures as most populated conformation at the lowest simulation temperature. However, accurate description of the temperature dependence of the conformer stability may require further force field improvement. It also indicates that care should be taken if one combines ensembles generated at the various temperatures of a RexMD simulation to extract thermodynamic quantities due to possible force field artifacts.

CONCLUSIONS

Hairpin loop structures are an important structural motif in nucleic acids and have been shown to play important roles in many biological processes. Understanding the structure formation process of nucleic acid hairpin structures at atomic detail is of major importance to fully understand the function of hairpins and the folding of larger nucleic acids that contain hairpin motifs.

We used replica exchange MD simulations in explicit solvent to study the structure formation of the stable GCA trinucleotide DNA hairpin with a characteristic loop structure and flanked by two stem basepairs.

The RexMD simulations employed a completely flexible single-stranded DNA without adding any restraints to bias the simulations toward a folded hairpin structure. This goes beyond a previous systematic conformational search study on the same system employing an implicit solvent model (23). In this study only the central loop structure was flexible, assuming a basepaired stem structure. During two independent RexMD simulations, folding of a single-stranded start structure to conformations close to an experimental hairpin structure as the dominant state was observed. In both simulations the population of near-native structures reached ~35% at the lowest temperature replica after ~20 ns (Fig. 9) with only small changes at later stages of the simulations. However, the population of alternative (misfolded) loop structures (e.g., with a *syn*-G_3 conformation) differed between both

RexMD simulations even at the final stages of the simulations. This result indicates that an appropriate sampling of alternative conformations and the possible refolding of trapped intermediate structures toward a correctly folded structure requires longer simulation times.

The analysis of intermediates at or shortly before the occurrence of fully folded hairpin structures indicated the formation of near-native trinucleotide loop conformations (without fully formed stems) and a variety of alternative intermediate structures. Folding to the native hairpin structure appeared to occur almost simultaneously or quickly after the formation of the near-native trinucleotide loop. This agrees qualitatively with results on the structure formation of an RNA tetraloop (central GCAA sequence) by Sorin et al. (32) using massively parallel independent MD simulations. In a small fraction of simulations the authors observed hairpin folding. Both a sequential folding mechanism (first loop and subsequent formation of stem basepairs) as well as compaction and simultaneous loop formation were observed (32). However, in contrast to the folding mechanism proposed by Sorin et al. (32) for an RNA tetraloop, in these simulations no hydrophobic collapse of the loop structure before loop formation was observed. The stable ''folding nucleus'' was formed by the central DNA trinucleotide loop element. This could be due to the fact that formation of the trinucleotide loop itself (without the stem) might be thermodynamically slightly favored, as proposed by Yoshizawa et al. (8).

In most of the sampled conformations with a near-native trinucleotide loop arrangement, the C_2 nucleotide adopted a stacked conformation with respect to the first loop nucleotide (the G_3 nucleotide of the GCA loop). This arrangement provides a hydrogen-bonding interface for the A_5 nucleotide of the loop to stabilize different loop fine structures but an overall helical arrangement or topology of the three loop nucleotides in close agreement with the native loop structure. This form can then rapidly proceed toward the fully folded hairpin loop structure. It appears to be further stabilized by a specifically bound water molecule at a cavity in the minor groove of the DNA that bridges the O_2 atom of the C_2 base and the N_1 of the A_5 base. Water molecules were also found to play a structural role during formation of RNA tetraloop structures by stabilizing partially formed stem basepairs (32). During folding of the DNA triloop the water molecule that bridges C_2 and A_5 stabilizes a specific stacking arrangement of the bases that form the native loop structure.

The proposed folding mechanism is supported by the experimental observation that the insertion of a three-carbon spacer in between the C_2 and G_3 nucleotide (destabilization of C_2-G_3 interactions) has a strongly destabilizing effect on loop formation (9). It is also consistent with time-resolved fluorescence spectroscopy of single-stranded DNA, which indicates that interactions of loop nucleotides and stem nucleotides can have a strong influence on the kinetics of hairpin formation (29). It is important to note that these RexMD simulations allow characterizing populations of near-native hairpin conformations and accumulation of intermediate structures. It is also possible to extract the order of appearance of such intermediate structures. However, the folding kinetics that is the exact transition times and transition rates between the various sampled structures cannot be determined. Characterization of folding kinetics might be possible in future studies using very long continuous MD simulations.

Hairpin formation in nucleic acids has been found to occur on a longer timescale than expected from the expected end-to-end contact formation rates of a semiflexible polymer (25–28). This has been attributed to the possible formation of trapped long-lived intermediate states that slow down structure formation (27,28) and may also lead to deviations from single-exponential kinetics of hairpin formation (29). Consistent with this experimental finding, the simulations show many ''misfolded'' intermediates that are unlikely to rapidly undergo direct transitions to the native loop structure. In addition, accumulation of an alternative loop structure containing a *syn*-G_3 conformation and an otherwise similar loop structure with respect to the native structure was observed. This loop structure also allowed formation of a fully folded structure with the G_3 trapped in the *syn* conformation. Indeed, in one of the RexMD simulations a significant fraction of the sampled structures even at the final stage of the simulation contained a *syn*-G_3. A slow decrease of the population over simulation time indicates that the loop structure with a *syn*-G_3 may correspond to a stable (long-lived) trapped conformation that requires unfolding and refolding to proceed toward the native hairpin loop structure.

The misfolding of nuleobases (especially of purines) at the *N*-glycosidic bond to form a *syn* conformation and the trapping of stable misfolded structures as seen in the simulations here might be of relevance for the folding of other nucleic acid structural motifs. The simulations indicate that it is possible to systematically study structure formation processes of small nucleic acid structural motifs using MD simulations in explicit solvent and advanced sampling methods. It can form the basis for systematic studies on characterizing the sequence dependence of hairpin folding in nucleic acids and on characterizing possible stable intermediate structures.

This study was performed using the computational resources of the Computational Laboratory for Analysis, Modeling and Visualization at Jacobs University Bremen and supercomputer resources of the Environmental Molecular Science Laboratories at the Pacific Northwest National Laboratories; grant gc9593. The work was supported by a grant (I/80485) from the VolkswagenStiftung to M.Z.

REFERENCES

1. Hirao, I., G. Kawai, S. Yoshizawa, Y. Nishimura, Y. Ishido, K. Watanabe, and K. Miura. 1994. Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nuclease and heat. *Nucleic Acids Res.* 22:576–582.
2. Yu, A., J. Dill, and M. Mitas. 1995. The purine-rich trinucleotide repeat sequences d(CAG)15 and d(GAC)15 form hairpins. *Nucleic Acids Res.* 23:4055–4057.

3. Zhu, L., S. H. Chou, and B. R. Reid. 1996. Structure of a single cytidine hairpin loop formed by the DNA triplet GCA. *Nat. Struct. Biol.* 2:1012–1017.

4. Chou, S. H., L. Zhu, Z. Gao, J. W. Cheng, and B. R. Reid. 1996. Hairpin loops consisting of single adenine residues closed by sheared A:A or G:G pairs formed by DNA triplets AAA and GAG: solution structures of the d(GTACAAAGTAC) hairpin. *J. Mol. Biol.* 264:981–1001.

5. Chou, S. H., Y. Y. Tseng, and S. W. Wang. 1999. Stable sheared A:C pair in DNA hairpins. *J. Mol. Biol.* 287:301–313.

6. Chou, S. H., Y. Y. Tseng, and B. Y. Chu. 1999. Stable formation of a pyrimidine-rich loop hairpin in a cruciform promoter. *J. Mol. Biol.* 292:309–320.

7. Aslani, A. A., O. Mauffret, F. Sourgen, S. Neplaz, G. Maroun, E. Lescot, G. Tevanian, and S. Fermandjian. 1996. The hairpin structure of a topoisomerase II site DNA strand analysed by combined NMR and energy minimization methods. *J. Mol. Biol.* 263:776–788.

8. Yoshizawa, S., G. Kawai, K. Watanabe, K. Miura, and I. Hirao. 1997. GNA trinucleotide loop sequences producing extraordinarily stable DNA minihairpins. *Biochemistry.* 36:4761–4767.

9. Moody, E. M., and P. C. Bevilacqua. 2003. Thermodynamic coupling of the loop and stem in unusually stable DNA hairpins closed by CG base pairs. *J. Am. Chem. Soc.* 125:2032–2033.

10. Moody, E. M., and P. C. Bevilacqua. 2003. Folding of a stable DNA motif involves a highly cooperative network of interactions. *J. Am. Chem. Soc.* 125:16285–16293.

11. Moody, E. M., and P. C. Bevilacqua. 2004. Structural and energetic consequences of expanding a highly cooperative stable DNA hairpin loop. *J. Am. Chem. Soc.* 126:9570–9577.

12. Glucksmann-Kuis, M. A., C. Malone, P. Markiewicz, and L. B. Rothman-Denes. 1992. Specific sequences and a hairpin structure in the template strand are required for N4 virion RNA polymerase promoter recognition. *Cell.* 70:491–500.

13. Glucksmann-Kuis, M. A., X. Dai, P. Markiewicz, and L. B. Rothman-Denes. 1996. E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell.* 84:147–154.

14. Gellert, M. 2002. V(d)j recombination: rag proteins, repair factors, and regulation. *Annu. Rev. Biochem.* 71:101–132.

15. Gacy, A. M., G. Geollner, N. Juranic, S. Macura, and C. T. McMurray. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell.* 81:533–540.

16. Chen, X., S. V. Santhana-Mariappan, P. Catasti, R. Ratliff, R. K. Moyzis, A. Laayoun, S. S. Smith, E. M. Bradbury, and G. Gupta. 1995. Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Natl. Acad. Sci. USA.* 92:5199–5203.

17. Mitas, M., A. Yu, J. Dill, and I. S. Haworth. 1995. The trinucleotide repeat d(CGG)15 forms a heat-stable hairpin containing G*syn*:Ganti base pairs. *Biochem.* 34:12803–12811.

18. Gellibolian, R., A. Bacolla, and R. D. Wells. 1997. Triplet repeat instability and DNA topology: an expansion model based on statistical mechanics. *J. Biol. Chem.* 272:16793–16797.

19. Völker, J., N. Makube, G. E. Plum, H. H. Klump, and K. J. Breslauer. 2002. Conformational energetics of stable and metastable states formed by DNA triplet repeat oligonucleotides: implications for triplet expansion diseases. *Proc. Natl. Acad. Sci. USA.* 99:14700–14705.

20. Pavia, A. M., and R. D. Sheardy. 2004. Influence of sequence context and length on the structure and stability of triplet repeat DNA oligomers. *Biochem.* 43:14218–14227.

21. Chou, S. H., K. H. Chin, and A. H. Wang. 2003. Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res.* 31:2461–2474.

22. Nakano, M., E. M. Moody, J. Liang, and P. C. Bevilacqua. 2002. Selection for thermodynamically stable DNA tetraloops using temperature gradient gel electrophoresis reveals four motifs: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc). *Biochemistry.* 41:14281–14292.

23. Zacharias, M. 2001. Conformational analysis of DNA-trinucleotide-hairpin-loop structures using a continuum solvent model. *Biophys. J.* 80:2350–2363.

24. Villescas, G., and M. Zacharias. 2004. Efficient search on energy minima for structure prediction of nucleic acid motifs. *J. Biomol. Struct. Dyn.* 22:355–364.

25. Ansari, A., S. V. Kuznetsov, and Y. Shen. 2001. Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc. Natl. Acad. Sci. USA.* 98:7771–7776.

26. Ansari, A., and S. V. Kuznetsov. 2005. Is hairpin formation in single-stranded polynucleotide diffusion-controlled? *J. Phys. Chem.* 109:12982–12989.

27. Wallace, M. I., L. Ying, S. Balasubramanian, and D. Klenerman. 2001. Non-Arrhenius kinetics for the loop closure of a DNA hairpin. *Proc. Natl. Acad. Sci. USA.* 98:5584–5589.

28. Wang, X., and W. M. Nau. 2004. Kinetics of end-to-end collision in short single-stranded nucleic acids. *J. Am. Chem. Soc.* 126:808–813.

29. Kim, J., S. Doose, H. Neuweiler, and M. Sauer. 2006. The initial step of DNA hairpin folding: a kinetic analysis using fluorescence correlation spectroscopy. *Nucleic Acids Res.* 34:2516–2527.

30. Sorin, E. J., Y. M. Rhee, B. J. Nakatani, and V. S. Pande. 2003. Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. *Biophys. J.* 85:790–803.

31. Sorin, E. J., B. J. Nakatani, Y. M. Rhee, G. Jayachandran, V. Vishal, and V. S. Pande. 2004. Does native state topology determine the RNA folding mechanism? *J. Mol. Biol.* 337:789–797.

32. Sorin, E. J., Y. M. Rhee, and V. S. Pande. 2005. Does water play a structural role in the folding of small nucleic acids? *Biophys. J.* 88: 2516–2524.

33. Swendsen, R. H., and J. S. Wang. 1986. Replica Monte Carlo simulations of spin glasses. *Phys. Rev. Lett.* 57:2607–2609.

34. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

35. Sanbonmatsu, K. Y., and A. E. Garcia. 2002. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins.* 46:225–236.

36. Zhou, R., and B. J. Berne. 2002. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. USA.* 99:12777–12782.

37. Zhou, R. 2004. Exploring the protein folding free energy landscape: coupling replica exchange method with P3ME/RESPA algorithm. *J. Mol. Graph. Model.* 22:451–463.

38. Yoshida, K., T. Yamaguchi, and Y. Okamoto. 2005. Replica-exchange molecular dynamics simulation of small peptide in water and in ethanol. *Chem. Phys. Lett.* 41:2280–2284.

39. Murata, K., Y. Sugita, and Y. Okamoto. 2004. Free energy calculations for DNA base stacking by replica-exchange umbrella sampling. *Chem. Phys. Lett.* 385:1–7.

40. Murata, K., Y. Sugita, and Y. Okamoto. 2005. Molecular dynamics simulations of DNA dimmers based on replica-exchange umbrella sampling I: test of sampling efficiency. *J. Theoret. Comput. Chem.* 4:411–432.

41. Murata, K., Y. Sugita, and Y. Okamoto. 2005. Molecular dynamics simulations of DNA dimers based on replica-exchange umbrella sampling II: free energy analysis. *J. Theoret. Comput. Chem.* 4:433–448.

42. Case, D. A., D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. 2003. Amber 8. University of California, San Francisco.

43. Jorgensen, W., J. Chandrasekhar, J. Madura, R. Impey, and M. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

44. Duan, Y., C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. 2003. A point-charge force field for molecular mechanics

simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* 24:1999–2012.

45. Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. Pedersen. 1995. A smooth particle mesh Ewald potential. *J. Chem. Phys.* 103:8577–8593.

46. Miyamoto, S., and P. A. Kollman. 1992. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13:952–962.

47. Lavery, R., K. Zakrzewska, and H. Sklenar. 1995. JUMNA (junction minimization of nucleic acids). *Comput. Phys. Com.* 91:135–158.

48. Feig, M., J. Karanicolas, and C. L. Brooks. 2004. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* 22:377–395.

49. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

50. Luzzati, V., A. Mathis, F. Mason, and J. Witz. 1964. Structure transitions observed in DNA and polyA in solution as a function of temperature and pH. *J. Mol. Biol.* 10:28–41.

51. van Holde, K. E., J. Brahms, and A. M. Michelson. 1965. Base interactions of nucleotide polymers in aqueous solution. *J. Mol. Biol.* 12:726–739.

52. Mills, J. B., E. Vacano, and P. J. Hagerman. 1999. Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence length of poly(dT) and poly(dA). *J. Mol. Biol.* 285:245–257.

53. Isakson, J., S. Acharya, J. Barman, P. Cheruka, and J. Chattopadhyaya. 2004. Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochem.* 43:15996–16010.

54. Cheong, C., G. Varani, and I. Tinoco. 1990. Solution structure of an unusually stable RNA hairpin, 5GGAC(UUCG)GUCC. *Nature.* 346: 680–682.

55. Ennifar, E., A. Nikulin, S. Tishchenko, A. Serganov, N. Nevskaya, M. Garber, B. Ehresmann, C. Ehresmann, S. Nikonov, and P. Dumas. 2000. The crystal structure of UUCG tetraloop. *J. Mol. Biol.* 304: 35–42.

56. Antao, V. P., S. Y. Lai, and I. Tinoco. 1991. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* 19: 5901–5905.

57. Antao, V. P., and I. Tinoco. 1992. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* 20:819–824.