A Computational Approach to Modeling Nucleic Acid Hairpin Structures

Chang-Shung Tung

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

ABSTRACT Hairpin is a structural motif frequently observed in both RNA and DNA molecules. This motif is involved specifically in various biological functions (e.g., gene expression and regulation). To understand how these hairpin motifs perform their functions, it is important to study their structures. Compared to protein structural motifs, structures of nucleic acid hairpins are less known. Based on a set of reduced coordinates for describing nucleic acid structures and a sampling algorithm that equilibrates structures using Metropolis Monte Carlo simulation, we developed a method to model nucleic acid hairpin structures. This method was used to predict the structure of a DNA hairpin with a single-guanosine loop. The lowest energy structure from the ensemble of 200 sampled structures has a RMSD of <1.5 Å, from the structure determined using NMR. Additional constraints for the loop bases were introduced for modeling an RNA hairpin with two nucleotides in the loop. The modeled structure of this RNA hairpin has extensive base stacking and an extra hydrogen bond (between the CYT in the loop and a phosphate oxygen), as observed in the NMR structure.

INTRODUCTION

A hairpin motif, made of a double-helical stem and a single-stranded loop, is one of the simplest yet most important structural elements in nucleic acid molecules. For example, it serves as a major building block for the folded three-dimensional (3-D) RNA structure. The very stable hairpins were proposed to be nucleation sites (Shen et al., 1995) for RNA folding. The torsional stress accumulated in a negatively supercoiled circular DNA molecule can be relaxed by extruding a palindromic sequence from a duplex conformation to an alternative hairpin structure (Boulard et al., 1991).

Hairpins are involved in specific interactions with a variety of proteins to regulate gene expression. The hairpin structure is required in the HIV TAR RNA for transactivation by tat protein (Feng and Holland, 1988). Both the R17 and Qβ coat protein bind to RNA hairpin loop structures and repress the translation of the phage replicase gene (Romaniuk et al., 1987; Witherell and Uhlenbeck, 1989). Other molecules such as IRE (iron-responsive element) binding protein (Leibold et al., 1990; Barton et al., 1990), sarcin and racin (Endo et al., 1988), ribosomal proteins (Freedman et al., 1987; Philippe et al., 1990; Climie and Friesen, 1987), and a number of phage coat proteins (Witherell et al., 1991) all recognize and bind to an RNA hairpin motif. A cruciform often occurs in regions of the DNA associated with biological controls (Boulard et al., 1991; Zhou and Vogel, 1993). It was proposed that DNA hairpins play an important role in the control of transcription, translation, and other biological functions (Blatt et al., 1993; Raghunathan et al., 1991).

It is important to study nucleic acid hairpin structures to understand how these structures perform their associated functions. Compared to protein structural motifs, atomic structures of nucleic acid hairpins are much less known. For many years, hairpins from tRNA molecules (structures solved by x-ray crystallography (Jack et al., 1976; Sussman et al., 1978)) were the only ones with known three-dimensional (3-D) structures. In recent years, a large effort has been devoted to the study of nucleic acid hairpin structures using NMR combined with molecular modeling techniques (e.g., constrained molecular dynamics, distance geometry). Structures of hairpins with various loop sizes were determined (e.g., Cheong et al., 1990; Orita et al., 1993; Davis et al., 1993; Puglisi et al., 1992; Gupta et al., 1994; Cain et al., 1995). Both RNA and DNA hairpin structures were solved by using x-ray crystallography (Valegard et al., 1994; Chattopadhyaya et al., 1988).

Several attempts were made to study nucleic acid hairpin structures using different theoretical approaches. Erie et al. (1993b) studied hairpins with T_n and A_n (n = 3, 4, 5) loops based on a library of feasible dinucleotide structures (Erie et al., 1993a). Conformational flexibilities were observed in these hairpins. The constructed hairpin loop models exhibit hydrophobic and hydrophilic surfaces. Perpendicular aromatic interactions of bases in the loop are observed in many of the modeled hairpin structures. Based on the observation that single-stranded RNA assumes an A-RNA-like conformation, Kajava and Ruterjans (1993) proposed a mechanism to construct a hairpin turn by changing only one backbone torsional angle, α (Seeman et al., 1976). Applying a very complex modeling approach that includes the use of MOGLI on an Evans and Sutherland system, Quanta on a Silicon Graphics, CHARMM, and an adapted basis Newton-Raphson method, Raghunathan et al. (1991) demonstrated that a hairpin loop containing only two nucleotides (IA) can be readily formed. These different theoretical methods had produced some very useful insights. Here we describe an alternative method that is developed for general

Received for publication 8 July 1996 and in final form 1 November 1996.

Address reprint requests to Dr. Chang-Shung Tung, Los Alamos National Laboratory, Mail Stop K-710, Group T-10, Los Alamos, NM 87545. Tel.: 505-665-2597; Fax: 505-665-3493; E-mail: cst@t10.lanl.gov.

0006-3495/97/02/876/10 $2.00

and efficient searches of hairpin structures in the conformational space.

To a first approximation, the stem region of a nucleic acid hairpin can be adequately modeled by using a canonical duplex conformation (i.e., A-form for RNA and B-form for DNA). With this approximation, the modeling of nucleic acid hairpin structures is reduced to the modeling of the structures of the single-stranded loop that connected the two strands of the duplex. Using a set of reduced coordinates developed in our laboratory, we have constructed an algorithm that is capable of generating structures of single-stranded loops with a pair of fixed ends (Tung et al., 1996). This algorithm allows efficient structural sampling of the loop in the conformational space. We combine this loop-generating algorithm and a modified Metropolis Monte Carlo algorithm to form a structural simulation package for studying nucleic acid hairpin structures. This method is general and can be used to model structures of both DNA and RNA hairpins with different lengths and sequences.

METHODS AND RESULTS

A set of reduced coordinates for nucleic acid structures

During the past decade, we have been involved in the development of a set of reduced coordinates for describing and modeling nucleic acid structures. Because of the double-bond-like nature, each base in the molecule is planar and can be treated as a rigid body. Choosing the principal axes of inertia (Goldstein, 1959) as the internal coordinate system for both bases and base pairs, we have developed a set of reduced coordinates that uses six orthogonal parameters for describing the structure of every base in the molecule (Soumpasis and Tung, 1988). This set of reduced coordinates was modified and expanded such that the new set is capable of describing noncanonical structures such as multiple-stranded molecules (e.g., triplexes, quadruplexes), parallel-stranded molecules, and duplexes that have mismatched base pairs (Tung et al., 1994). This new set of parameters conforms with a requirement (Dickerson et al., 1989) agreed upon in the Cambridge 1988 workshop that the values of helical parameters (twist, roll, tilt, etc.) should be the same (up to sign change) calculated from the two opposite directions of an antiparallel structure.

There are six backbone torsional angles (Seeman et al., 1976) involved in the sugar-phosphate connection between two neighboring bases. These torsional angles plus the flexibility in the sugar ring (Cremer and Pople, 1975) make the nucleic acid backbone structure inherently very flexible. To properly model the flexible backbone structures, a second set of reduced coordinates was developed. In our representation (Tung, 1993), three parameters (glycosidic angle, out-of-plane distortion, pseudo-rotational angle) are used for defining sugar structures, and one torsional angle (a rotation of the group with respect to the $O3'_n$-$C5'_{n+1}$ pseudobond) and two bond angles ($O3'_n-P_{n+1}-O5'_{n+1}$, $P_{n+1}-O5'_{n+1}-C5'_{n+1}$) completely describe the phosphate structure.

Modeling of the loop region

For a nucleic acid hairpin, the loop is connected to the 3′ and 5′ ends of the two strands in a double-helical conformation (as shown in Fig. 1). Under physiological conditions, the structure of the stem can be approximated by a fixed conformation (A-form for RNA and B-form for DNA). Therefore, the two ends of the loop in a hairpin are known. Various nucleic acid chain-closure algorithms were proposed by Zhurkin et al. (1978) and Miller (1979) and further developed by Olson's group (Srinivasan and Olson, 1987; Erie et al., 1993). With proper end constraints introduced, hairpin loop structures can be generated. In practice, a large body of generated structures that do not satisfy the end constraints has to be discarded. Here we introduce a direct and efficient method that allows the generation of loop structures with a pair of fixed ends. Fig. 2 shows a loop consisting of a single nucleoside and two flanking phosphate groups. As in the virtual bond approach introduced by Olson and Flory (1972), in our representation every phosphate group (Tung, 1993) is replaced by a vector (see $\overrightarrow{AB}$ and $\overrightarrow{CD}$ in the figure), and every sugar (Altona and Sundaralingam, 1972) is also replaced by a vector (see $\overrightarrow{BC}$ in the figure). For a given pair of end points (A, D),

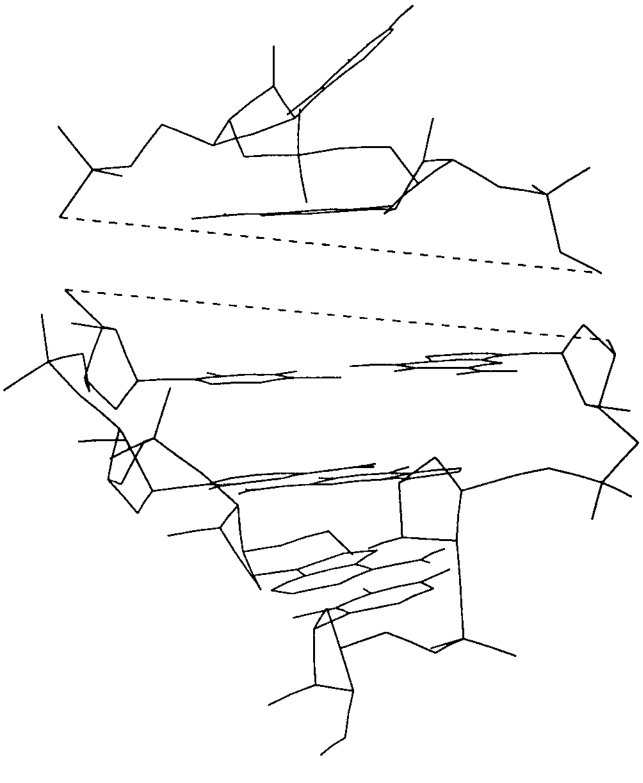

FIGURE 1 The plot of a hairpin with two bases in the loop and three base pairs in the stem. The dashed lines connect the O3′ and the C5′ atoms where the loop region joins the stem region.

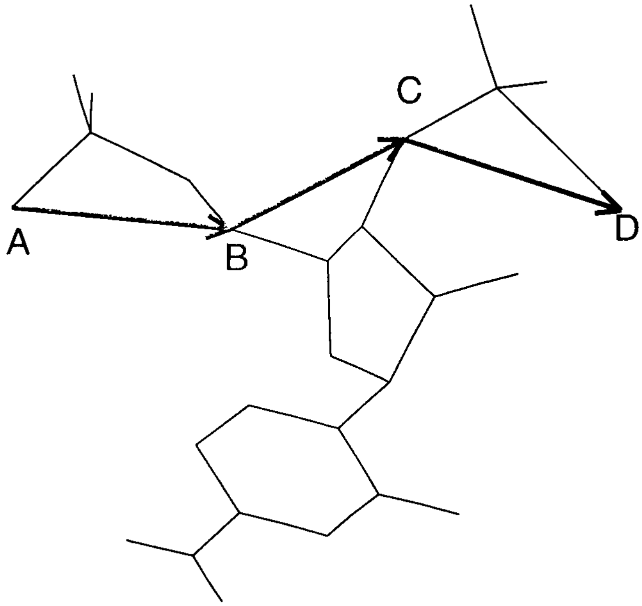

FIGURE 2 The plot of a loop consisting of a single nucleoside and two flanking phosphate groups. In this representation, each of the sugar and phosphate groups is replaced by a vector.

there exists a torsional angle (β_{ABCD}) for the three vectors that has the exact end-to-end distance (l_{AD}). This torsional angle (β_{ABCD}) can be written as a function of three lengths (l_{AB}, l_{BC}, l_{CD}), two angles (θ_{ABC}, θ_{BCD}), and the end-to-end distance (l_{AD}) as shown in the following equation (Tung, 1993):

$$\cos\beta = \frac{l_1^2\sin^2\theta_1 + l_3^2\sin^2\theta_2 + (l_1\cos\theta_1 + l_3\cos\theta_2 - l_2)^2 - d}{2l_1l_3\sin\theta_1\sin\theta_2}, \tag{1}$$

where l_1 is l_{AB}, l_2 is l_{BC}, l_3 is l_{CD}, θ_1 is θ_{ABC}, θ_2 is θ_{BCD}, and d is l_{AD}. There exists a degeneracy in the solution of Eq. 1. The parameter that defines the sign of the torsional angle is required in addition to the three bond lengths and the two bond angles to completely describe the structure of $\vec{\mathrm{A}}\mathrm{BCD}$ with a fixed l_{AD}. Each of the vectors in this representation also consists of three bonds (O3′−P, P-O5′, O5′−C5′ for the phosphate group and C5′−C4′, C4′−C3′, C3′−O3′ for the sugar). With a fixed conformation for the connecting vectors $\vec{\mathrm{A}}\mathrm{BCD}$, a correct conformation of the phosphate group or the sugar can be found to replace each of the vectors with one rotational degree of freedom. These rotational degrees of freedom can be defined as one torsional angle [O3′(i − 1) − C5′(i) − O3′(i) − C3′(i)] for the ith sugar and one torsional angle [C5′(i) − O3′(i) − C5′(i + 1) − P(i)] for the ith phosphate groups. The base can be attached to the sugar with the specified glycosidic angle. Therefore, a set of parameters that consists of three lengths (l_{AB}, l_{BC}, l_{CD}), two bond angles (θ_{ABC}, θ_{BCD}), five torsional angles (one each for the sugar, the base, the two phosphates, and the orientation of the ABCD group), and three sign parameters completely describes the structure of a single-nucleoside loop with a pair of fixed ends. A recursive application of this approach was used for developing software to generate loop sizes up to five nucleotides (as used for our modeling of a pseudo-knot structure in *Escherichia coli* 16S rRNA). Expanding this algorithm to generate structures of larger loops with more than five nucleotides is an easy and straightforward task.

This algorithm allows the easy generation of different conformations for a single-stranded molecule with a pair of fixed ends. Fig. 3 shows different conformations of a single-nucleoside loop, all of which share a pair of identical ends. Each of the conformations corresponds to a set of 13 parameters, with the end-to-end distance as the input parameter. The end-to-end distance (l_{AD} in Fig. 2) in this case is 8.7 Å. This length compares favorably with the averaged length of single-nucleoside loops (8.9 Å) calculated from crystal structures of six different RNA molecules (6tna, 1tn2, 1tra, 4tra, 5tra, 1rmn) from the Protein Data Bank (PDB) (Bernstein et al., 1977), as shown in Table 1.

Modeling the structure of a DNA hairpin with a single base loop

We chose to model structures of the hairpin d(AT-GGA-AT) as our first study because 1) it is a structural motif for the human centromeric repeating sequence d(AATGG)$_n$ (Moyzis et al., 1988) and 2) the structure was solved using NMR and constrained molecular dynamics (Gupta et al., 1994). In this hairpin, the GUA-ADE forms a reverse wobble base pair, leaving only one nucleotide (GUA) in the loop region. Another compact hairpin d(GC-GAA-GC) with a single-nucleoside (ADE) loop was also observed by Hirao and colleagues (Hirao et al., 1994). One of the important advantages of studying a compact hairpin with a single-nucleoside loop is the relatively small conformational space associated with the molecule. An exhaustive sampling of the conformational space associated with the loop structure is feasible.

There is a maximum distance that a single base loop can bridge. A GUA-ADE wobble base in a canonical B-conformation has an O3′-C5′ distance of 11.4 Å (see the *dashed line* at the left image of Fig. 4). This distance is too large for even a completely stretched single base loop to make the connection. Based on a systematic search, allowing base

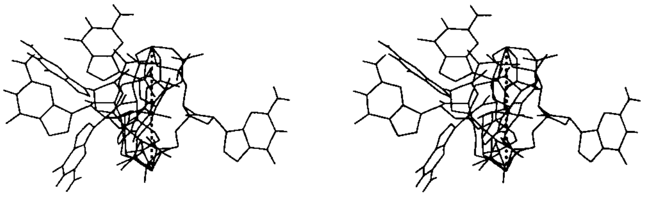

FIGURE 3 Different conformations of a single-nucleoside loop that all share a pair of identical ends (shown as the *dotted line*).

TABLE 1 End-to-end distances* of various loop sizes (in Å)

No. of bases	Averaged length	SD
1	8.86	0.89
2	13.58	1.73
3	17.26	2.56
4	19.90	3.40
5	21.65	4.34
6	22.82	5.27
7	23.69	6.10
8	24.45	6.87

*Calculated using six structures (1rmn, 1tn2, 1tra, 4trn, 5tra, 5tna) from PDB.

pair buckle (κ), propeller twist (ω) angle, and the two glycosidic angles (χ_1, χ_2) to vary, we have found that a particular GUA-ADE conformation (k: 48°, w: 20°, χ_1: $-30°$, χ_2: $-20°$) has reduced this O3′-C5′ distance to 8.6 Å. This shorter distance is favorable for the single-nucleoside loop to make the connection. For the backbone to make the connection, structural adjustments from the canonica B-conformation were made for the base pair dimer d(TG)•d(AA). The resultant structure of the d(ATG)•d(AAT) double-helical stem is shown as the right image of Fig. 4. Table 2 lists the structural parameters associated with this duplex.

We essentially used the AMBER force field (Weiner et al., 1984, 1986) plus a solvent correction term, as follows, to calculate the conformational energy of the hairpin (three base pairs in the stem plus the single base loop):

$$E_{\text{total}} = \sum_{\text{angle}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedral}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] + E_{\text{PMF}}. \quad (2)$$

Because all bond lengths were kept at their equilibrium values when the set of reduced coordinates for constructing structures was used, the bond energy term vanished. To increase the speed of computation, those nonbonded pairs with both atoms in the stem region of the molecule are excluded from the list in calculating van der Waals and electrostatic energies. E_{PMF} represents the correction of the total energy due to the solvent effects and is calculated from the hyper-netted chain approximation, as shown by Hummer and Soumpasis (Soumpasis, 1984; Hummer and Soumpasis, 1993). Implicitly including the solvent effects properly takes care of the overrepresented charge-charge interactions between the phosphate groups.

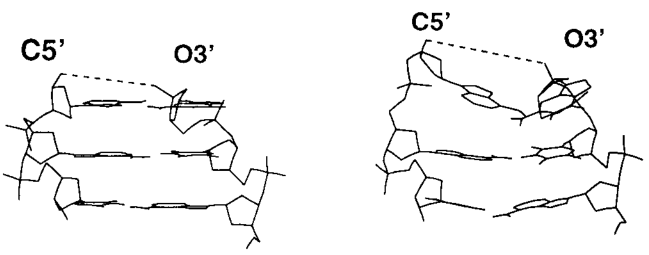

FIGURE 4 Structure of the double helix with a GUA-ADE mismatched base pair. The O3′-C5′ distance (indicated by the *dashed line*) is 11.4 Å for the duplex in a canonical B-conformation (shown as the *left image*). This distance has reduced to 8.6 Å for the same duplex in a modified B-conformation. The shorter O3′-C5′ distance is more favorable for the single-nucleoside loop to make the connection.

TABLE 2 Structural parameters* for d(ATG) · d(AAT)

Parameters	A1-T6	T2-A5	G3-A4
Tilt	—	−0.9	3.6
Twist	—	34.2	−43.6
Roll	—	4.4	3.7
Shift	—	0.0	−1.4
Slide	—	0.0	1.3
Rise	—	3.3	3.5
Buckle	15.2	19.1	48.2
Opening	0.0	0.0	5.4
Propeller twist	18.9	15.0	20.0
Shear	0.0	0.0	−6.2
Stretch	0.0	0.0	5.7
Stagger	0.0	0.0	−0.4

*Following the Cambridge convention as described by Dickerson et al. (1989).

Based on the modeled structure of the double-helical stem, random structures of the single-stranded region were constructed using our developed loop-generating algorithm. The random structures were generated with bond angles of the loop parameters (as described in the previous section) that varied between 60° and 170°, torsional angles of the loop parameters that varied between 0° and 360°, and sign parameters of either +1 or −1. Those structures with a total energy smaller than a cutoff value (typically 10,000 kcal/mol) were selected for energy refinement by a short run (1000 cycles) of Metropolis Monte Carlo Simulation (Metropolis et al., 1953). The purpose of the preselection of initial structures for subjection to energy refinement is to prevent the trapping of conformations in high energy minima. Every energy-refined structure is treated as an accepted structure of the hairpin. It took approximately 6 h of UNIX time on a Silicon Graphics Indy workstation to generate 200 accepted structures for this particular hairpin. During the simulation, approximately one-tenth of the computational time was spent on the generation of random structures for the preselection procedure, and the rest of the time was spent on the equilibration of the selected initial random structures.

Using the root-mean-square deviation (RMSD) as a measure for uniqueness between different structures of the hairpin, we are able to classify the 200 accepted structures into families of distinct structures. The number of unique structures (N_{unique}) is defined to be the maximum number of structures in the ensemble that have pairwise RMSDs all greater than a cutoff value (d_c). Each unique structure represents a distinct family of structures. Each of the remaining accepted structures is grouped into the family with the unique structure that gives the smallest RMSD. Therefore, the pairwise RMSDs between structures within a family are all less than the cutoff value. Fig. 5 shows the plot of N_{unique} against the number of structures simulated with different cutoff values. It is obvious that a larger number of unique structures corresponds to a smaller d_c value (a fine sampling), whereas a smaller number of unique structures corresponds to a larger d_c value (a coarse sampling). For example, 25 unique structures were obtained with a d_c of 3 Å, whereas 4 unique structures were obtained with a cutoff of 7 Å, based on the set of 200 accepted structures. Fig. 5 also indicates that in general, a larger number of structures had to be simulated for a smaller d_c before the corresponding N_{unique} curve reaches its plateau. This figure provides an estimate for lengths of simulation required for different d_cs before a good sampling of the conformational space has been reached.

If one chooses an intermediate graininess of sampling with a d_c equal to 5 Å, nine unique structures (as shown in Fig. 6) can be deduced from the 200 accepted structures. Of these nine structures, four (*top row left, middle row middle, bottom row left and right*) have the loop base (GUA) unstacked with bases in the stem. Three of the structures (*top row middle, middle row left, and bottom row middle*) have the loop base stacked from the minor groove, and the remaining two structures have the loop base stacked from the major groove.

We have ranked the 200 accepted hairpin structures according to their conformational energies (E_{total}). The energetic differences between these accepted structures are relatively small. For example, the 50 lowest-ranked hairpin structures have energies differing by less than 15 kcal/mol. After a closer investigation, we have found that four of the five hairpins that rank as the lowest in energy of the whole ensemble belong to one family of distinct structure (corresponding to the right molecule on the top row in Fig. 6). The structure that corresponds to the lowest energy (number 44 in the ensemble) is our predicted structure for the hairpin. When compared to the two centromeric hairpin structures solved by using NMR (Gupta et al., 1994), it has RMSD values of 1.48 Å and 1.31 Å, respectively. Fig. 7 shows our predicted structure of the hairpin plotted along with the two structures (two loop structures of the double hairpin) solved by using NMR. Both our predicted structure and the NMR-determined structures have the loop base GUA stacked with the GUA base in the stem from the major groove. According to Shen et al. (1995), a RMSD of 1.6 Å between two conformations of a hairpin was considered structurally similar. We have calculated the structural parameters of the predicted and the two NMR-determined structures of the DNA hairpin. These parameters are listed in Table 3. All of the backbone parameters for the predicted structure are consistent with those for the NMR-determined structures. The slightly larger values in tilt, roll, slide, and shift indicate that the GUA-GUA step does not stack as well in the predicted structure as those in the NMR-determined structures. The differences in comparing nucleic acid structures using the RMSD and the structural parameters were discussed in our previous work (Tung and Soumpasis, 1996).

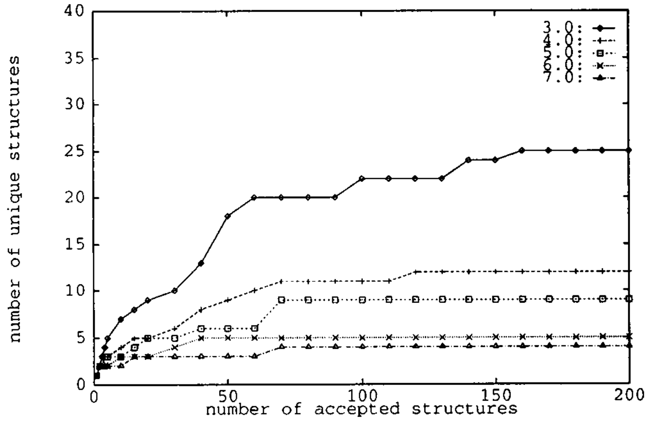

FIGURE 5 The plot of the number of unique structures versus the length of the simulation for the hairpin d(ATGGAAT) with different cutoff values.

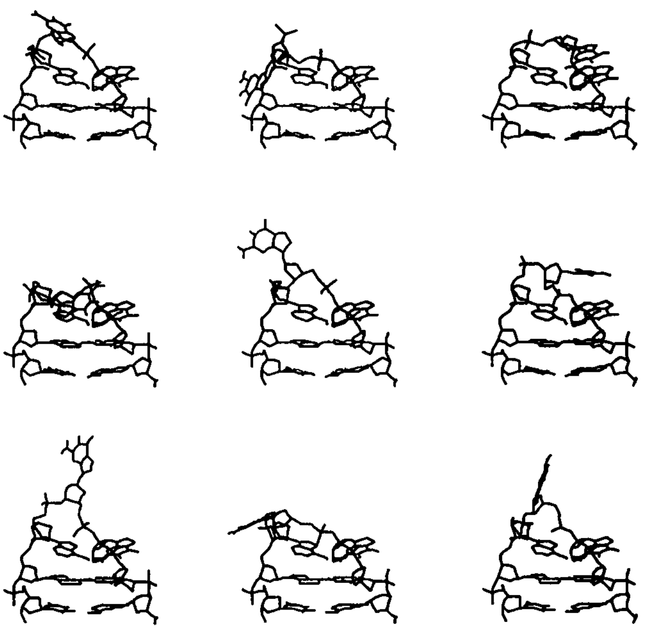

FIGURE 6 Nine unique structures of the hairpin d(ATGGAAT) deduced from the 200 accepted structures with all pairwise RMSDs greater than 5 Å.

Using only the information of the base-pairing and structural constraints based on connectivities but without the use of measured (NMR) distance constraints, we were able to predict the hairpin structure that is very close to the two derived experimentally.

Modeling the structure of an RNA hairpin with a two-base loop

Because of the success in predicting the structure of one of the most compact hairpins with a single-nucleoside loop, we have extended our study to include an RNA hairpin r(AC-

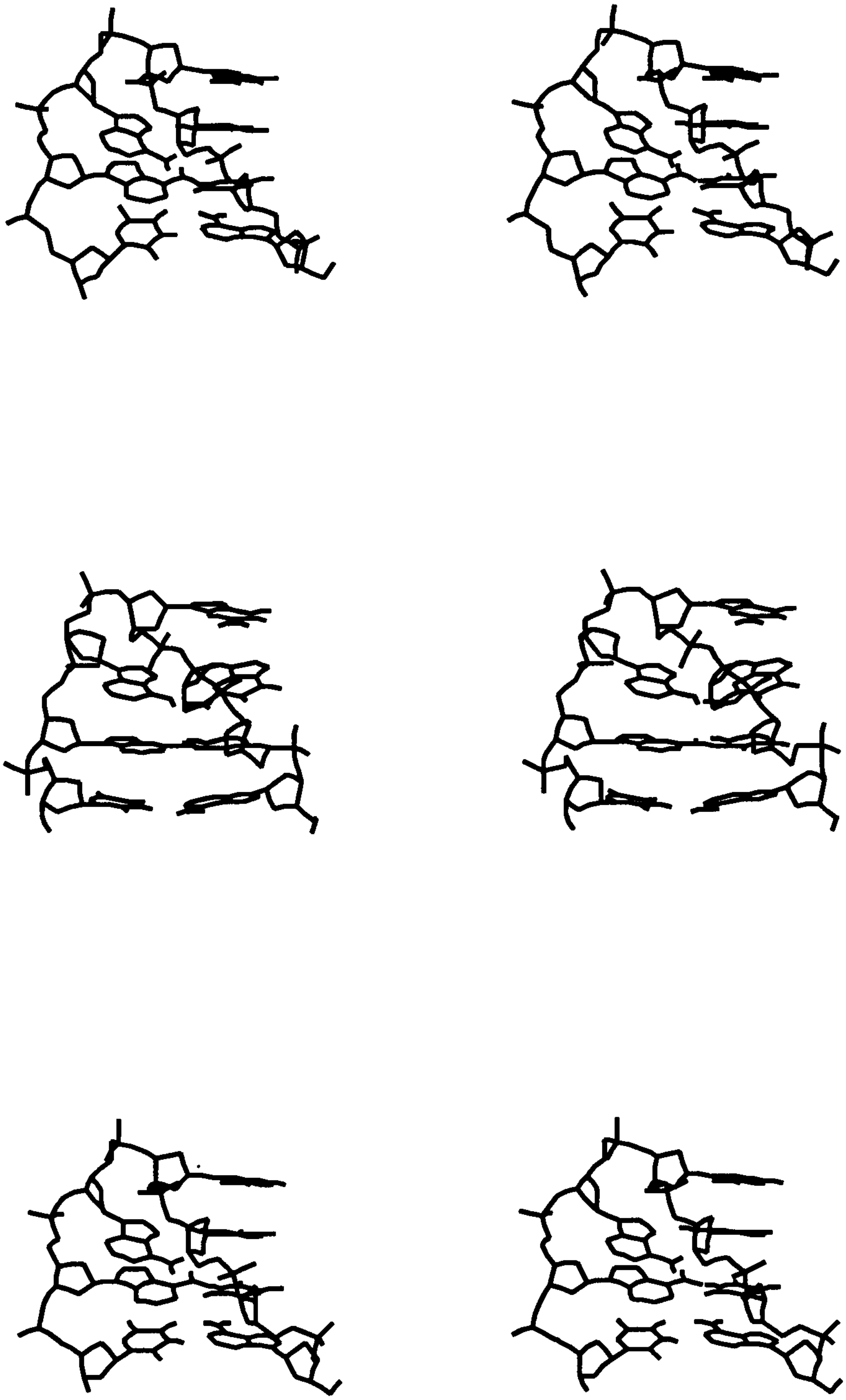

FIGURE 7 The stereo plot of the predicted structure (*middle image*) of the hairpin d(ATG-GAAT) together with the two loop structures (*top and bottom images*) of the double hairpin solved using NMR.

UUCG-GU). This hairpin has one of the very common tetraloop sequences, UUCG (Tuerk et al., 1988). The structure of this hairpin was solved by Tinoco and co-workers (Cheong et al., 1990) using NMR and a combination of distance geometry and constrained energy minimization. Similar to the d(AT-GGA-AT) hairpin, there exists a mismatched base pair (U-G) that closes the loop, leaving only two nucleotides in the single-stranded region. Extensive stacking for the loop bases and an additional hydrogen bond between one of the loop bases and a phosphate oxygen make the hairpin very stable. The ultrastable hairpins with a UUCG tetraloop were proposed to be nucleation sites for RNA folding (Cheong et al., 1990).

Because a two-nucleoside loop can bridge a much larger distance (averaged length of 13.58 Å as shown in Table 1) than a single-nucleoside loop, geometrically the URA-GUA mismatch can be modeled as a planar reverse wobble base pair in a duplex that assumes a canonical A-conformation

TABLE 3 Structural parameters for the GUA-4 in the loop of the DNA hairpin d(ATGGAAT)

	Parameters	Predicted	NMR-1	NMR-2
Base*	Tilt	−13.5	3.6	3.6
	Twist	51.7	33.4	37.5
	Roll	9.7	3.3	2.4
	Shift	−2.0	0.8	0.8
	Slide	2.5	−1.1	−1.1
	Rise	3.1	3.2	3.2
Backbone#	α	−67.5	−69.8	−69.8
	β	124.7	172.9	173.6
	γ	47.7	58.8	61.1
	δ	146.0	122.1	128.3
	ϵ	−80.5	−92.0	−92.6
	ζ	108.8	127.9	135.7
	χ	−86.8	−124.6	−123.8
	W	163.9	127.9	135.7

*The base parameters are calculated for the GUA-GUA step in the DNA hairpin based on the method of Tung et al. (1994).

#The backbone torsional angles are defined according to the method of Seeman et al. (1976). *W* is the pseudorotational angle (Cremer and Pople, 1975) for the sugar attached to the guanine in the loop.

without any modification. The modeled stem structure with an URA-GUA mismatch is shown in Fig. 8. The O3′-C5′ distance (shown as the *dotted line* in Fig. 8) between the ends of the two strands is 14.8 Å.

There is a larger number of degrees of freedom involved in a two-nucleoside loop than a single-nucleoside loop. We have run a simulation to generate 500 accepted structures for the hairpin. A plot similar to Fig. 4 had shown that with a cutoff value of 3 Å the curve had not reached a plateau at the end of the simulation. A much longer simulation has to be performed before a good sampling of the loop structure in the accessible conformational space can be reached. A closer examination of the 500 accepted structures showed that none of the structures has all of the features (i.e., an additional hydrogen bond between CYT-5 and a phosphate oxygen, stacking between URA-4 and the sugar attached to CYT-5, and stacking between CYT-5 and URA-3) observed in the NMR derived structure. One particular structure (number 363) in the ensemble has the amino group of CYT-5 close to a phosphate oxygen, as shown in Fig. 9. But the stacking between CYT-5 and URA-3 is poor in this structure. URA-4 is also far away from the sugar attached to CYT-5.

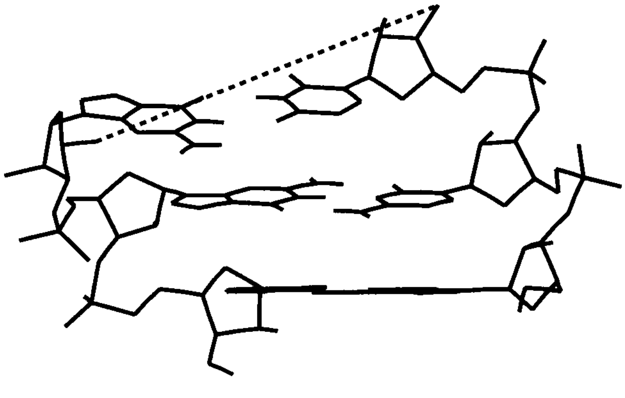

FIGURE 8 The stem structure of the hairpin r(ACUUCGGU) with an URA-GUA mismatch. The two ends for the loop connnection are highlighted by the connecting dotted line.

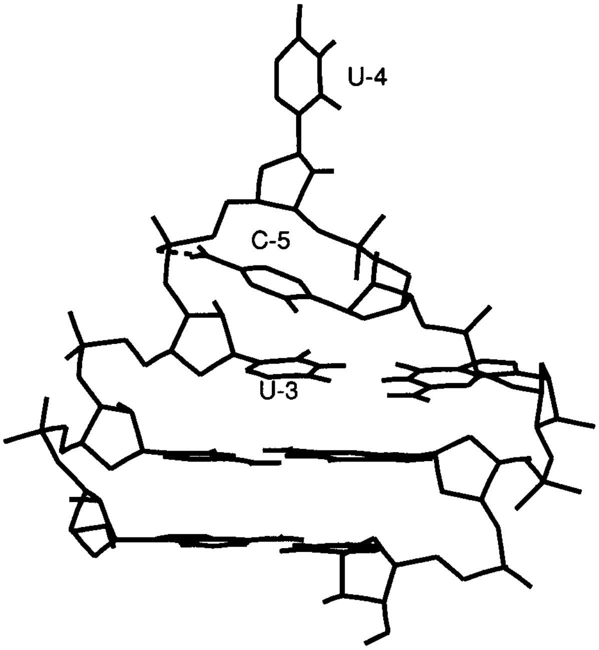

FIGURE 9 One hairpin structure (number 363) in the ensemble of 500 accepted structures has the amino group of CYT-5 close to a phosphate group, as indicated by the dotted line. This particular structure has URA-4 completely unstacked and a poor stacking between CYT-5 and URA-3.

To further test our method, instead of extending our simulation immediately, we took an alternative approach. Introducing constraints that represent the stacking of the loop bases (the distance between the centers of CYT-5 and URA-3 and the distance between the centers of URA-4 and the sugar attached to CYT-5 were both constrained to 3.5 Å) and the hydrogen bond (between CYT-5 and the phosphate

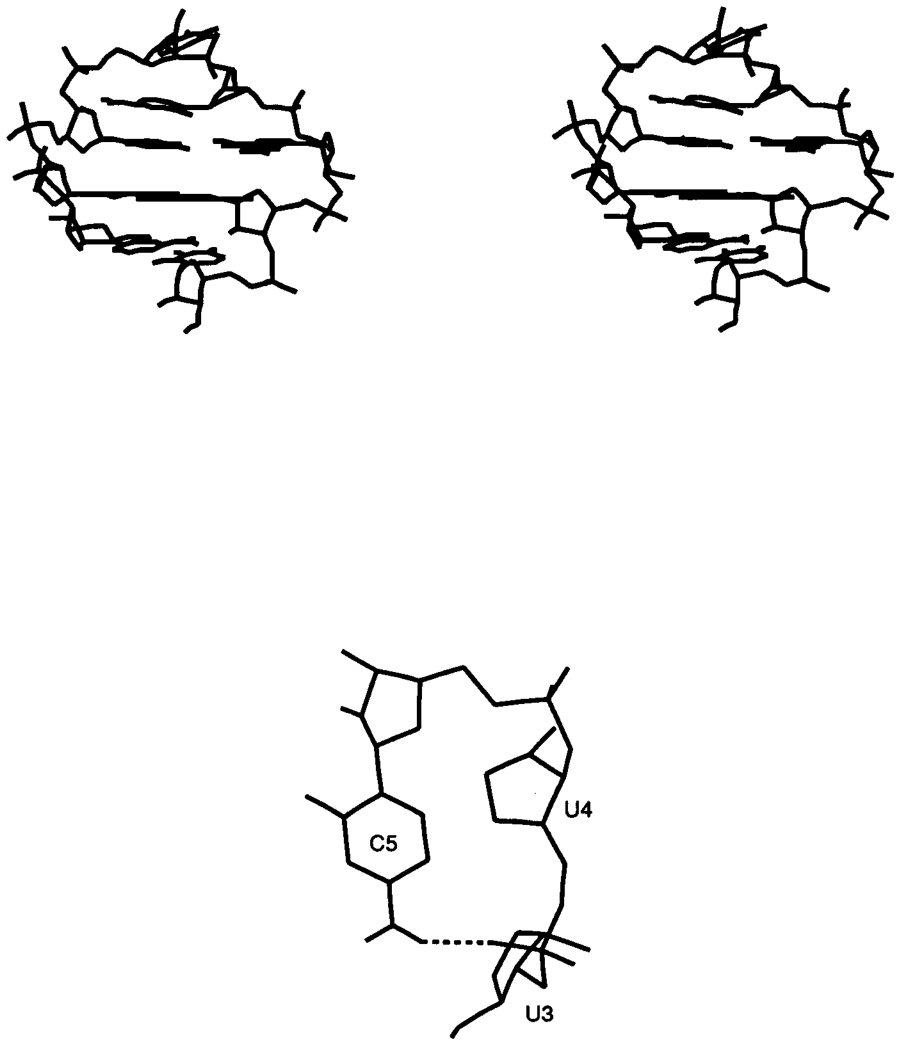

FIGURE 10 The predicted structure of the hairpin r(ACUUCGGU). The stereo image at the top shows the extensive stacking of the loop bases, and the image at the bottom shows the hydrogen bond between CYT-5 and the phosphate group that connects URA-3 and URA-4.

connecting URA-3 and URA-4), we were able to construct a predicted structure for the hairpin based on generating 200 accepted structures. The predicted structure (the one with the lowest energy) is shown in Fig. 10. Because of the fact that the structure of this hairpin solved using NMR was not deposited in PDB, we were not able to make a direct comparison in terms of RMSD between our predicted structure and the structure determined experimentally. Our predicted structure has the characteristics described in the paper published by Cheong et al. (1990), including an additional hydrogen bond between a cytosine and a phosphate oxygen, and extensive base stacking for bases in the loop. After we deduced our predicted structure for the RNA hairpin, the simulation was also extended until a reasonably good sampling was achieved when a small cutoff value (3 Å) was used. The simulation took approximately 9 days on an SGI Indy workstation to generate 7000 accepted structures. A plot (Fig. 11) of the number of unique structures against the number of accepted structures shows that the curve leveled off after 4500 accepted structures were generated. Of the accepted structures in the ensemble, we were able to find one (number 4265) that resembles the predicted structure with a RMSD of 1.8 Å. Fig. 12 shows the superposition of structure 4265 together with the predicted structure.

At this stage of development, our sampling algorithm seems to scale exponentially with loop size. To study hair-

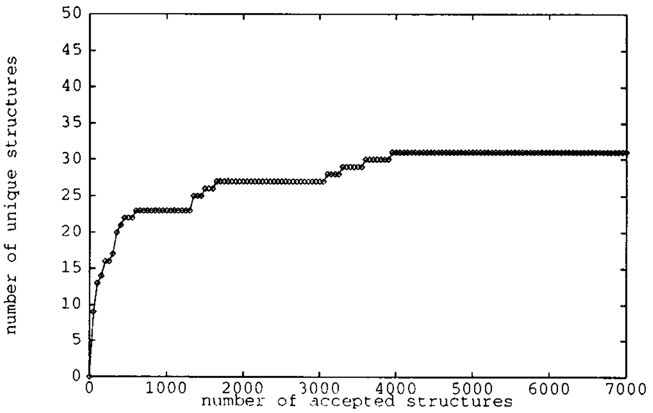

FIGURE 11 The plot of the number of unique structures versus the length of the simulation for the hairpin r(ACUUCGGU) with a cutoff of 3 Å.

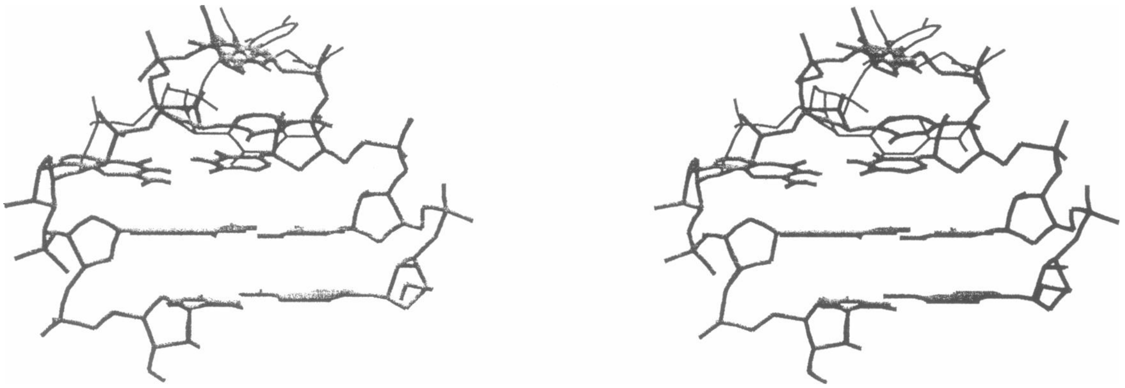

FIGURE 12 The stereo plot of a structure (*thin lines*) from the long simulation that resembles the predicted structure (*thick lines*) of the hairpin r(ACUUCGGU). The RMSD between the two structures is 1.8 Å.

pin structures with larger loop sizes, it is a great advantage if some structural constraints are available.

CONCLUSION

Based on a set of reduced coordinates and a modified Monte Carlo algorithm, we have developed a computational approach for modeling structures of both DNA and RNA hairpins. This approach is general and suitable for studying hairpin structures with different loop lengths and loop sequences.

Depending on the graininess of the sampling for the study, different numbers of distinct structures are generated. The history of the simulation provides a guideline with regard to the length of simulation required for a given fineness of sampling. A shorter simulation is needed if a coarse sampling is intended. As a first step, this approach allows the narrowing down of the study to only a small set of distinct structures. Without any additional imposed constraint, our predicted structure for d(AT-GGA-AT) has an RMSD of <1.5 Å from that determined using NMR. There is a larger number of degrees of freedom associated with a hairpin that has two nucleotides in the loop. A simulation that sampled 500 accepted structures was unable to produce a satisfactory structure for the hairpin r(AC-UUCG-GU). By introducing three additional distance constraints, we were able to generate a modeled structure of this RNA hairpin that resembles the NMR structure. Without the additional distance constraints, the procedure must be extended extensively before a reasonably good sampling of the conformational space is reached.

The coordinates of the NMR structure for the DNA hairpin were kindly provided to us by Drs. A. Garcia and G. Gupta.

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

REFERENCES

Altona, C., and M. Sundaralingam. 1972. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.* 94:8205–8212.

Barton, H. A., R. S. Eisenstein, A. Bomford, and H. N. Munro. 1990. Determinants of the interaction between the iron-responsive element-binding protein and its binding site in rat L-ferritin mRNA. *J. Biol. Chem.* 265:7000–7008.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Blatt, N. B., S. E. Osborne, R. J. Cain, and G. D. Glick. 1993. Conformational studies of hairp in sequences from the ColE1 cruciform. *Biochimie.* 75:433–441.

Boulard, Y., J. Gabarro-Arpa, J. A. Cognet, M. Le Bret, A. Guy, R. Teoule, W. Guschlbauer, and G. V. Fazakerley. 1991. The solution structure of a DNA hairpin containing a loop of three thymidines determined by nuclear magnetic resonance and molecular mechanics. *Nucleic Acids Res.* 19:5159–5167.

Cain, R. J., E. R. Zuiderweg, and G. D. Glick. 1995. Solution structure of a DNA hairpin and its disulfide cross-linked analog. *Nucleic Acids Res.* 23:2153–2160.

Chattopadhyaya, R., S. Ikuta, K. Grzeskowiak, and R. E. Dickerson. 1988. X-ray structure of a DNA hairpin molecule. *Nature.* 334:175–179.

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

Climie, S. C., and J. D. Friesen. 1987. Feedback regulation of the rplJL-rpoBC ribosomal protein operon of *Escherichia coli* requires a region of mRNA secondary structure. *J. Mol. Biol.* 198:371–381.

Cremer, D., and J. A. Pople. 1975. A general definition of ring puckering coordinates. *J. Am. Chem. Soc.* 97:1354–1358.

Davis, P. W., W. Thurmes, and I. Tinoco, Jr. 1993. Structure of a small RNA hairpin. *Nucleic Acids Res.* 21:537–545.

Dickerson, R. E., M. Bansal, C. R. Calladine, S. Diekmann, W. N. Hunter, O. Kennard, R. Lavery, H. C. Nelson, W. K. Olson, W. Saenger, Z. Shakked, H. Sklenar, D. M. Soumpasis, C. S. Tung, E. von Kitzing, A. H. C. Wang, and V. B. Zhurkin. 1989. Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.* 8:1–4.

Endo, Y., Y. L. Chan, A. Lin, K. Tsurugi, and I. G. Wool. 1988. The cytotoxins alpha-sarcin and ricin retain their specificity when tested on a synthetic oligoribonucleotide (35-mer) that mimics a region of 28 S ribosomal ribonucleic acid. *J. Biol. Chem.* 263:7917–7920.

Erie, D. A., K. J. Breslauer, and W. K. Olson. 1993a. A Monte Carlo method for generating structures of short single-stranded DNA sequences. *Biopolymers.* 33:75–105.

Erie, D. A., A. K. Suri, K. J. Breslauer, R. A. Jones, and W. K. Olson. 1993b. Theoretical predictions of DNA hairpin loop conformations: correlations with thermodynamic and spectroscopic data. *Biochemistry.* 32:436–454.

Feng, S., and E. C. Holland. 1988. HIV-1 tat trans-activation requires the loop sequence within tar. *Nature.* 334:165–167.

Freedman, L. P., J. M. Zengel, R. H. Archer, and L. Lindahl. 1987. Autogenous control of the S10 ribosomal protein operon of *Escherichia* *coli*: genetic dissection of transcriptional and posttranscriptional regulation. *Proc. Natl. Acad. Sci. USA.* 84:6516–6520.

Goldstein, H. 1959. Classical Mechanics. Addison-Wesley, Reading, MA.

Gupta, G., A. E. Garcia, P. Catasti, R. Ratliff, E. M. Bradbury, R. K. Moyzis. 1994. Stem-loop structures of the repetitive DNA sequences located at human centromeres. *In* Proceedings of the Eighth Conversation. R. H. Sarma and M. H. Sarma, editors. Adenine Press, New York. 137–154.

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 nucleases and heat. *Nucleic Acids Res.* 22:576–582.

Hummer, G., and D. M. Soumpasis. 1993. Correlations and free energies in restricted primitive model descriptions of electrolytes. *J. Chem. Phys.* 98:581–591.

Jack, A., J. E. Ladner, and A. Klug. 1976. Crustallographic refinement of yeast phenylalanine transfer RNA at 2.5 angstrom resolution. *J. Mol. Biol.* 108:619–649.

Kajava, A., and H. Ruterjans. 1993. Molecular modelling of the 3-D structure of RNA tetraloops with different nucleotide sequences. *Nucleic Acids Res.* 21:4556–4562.

Leibold, E. A., A. Laudano, and Y. Yu. 1990. Structural requirements of iron-responsive elements for binding of the protein involved in both transferrin receptor and ferritin mRNA post-transcriptional regulation. *Nucleic Acids Res.* 18:1819–1824.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–1092.

Miller, K. J. 1979. Interactions of molecules with nucleic acids. I. An algorithm to generate nucleic acid structures with an application to the B-DNA structure and a counterclockwise helix. *Biopolymers.* 18: 959–980.

Moyzis, R. K., J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. Meyne, R. L. Ratliff, and J. R. Wu. 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci. USA.* 85: 6622–6626.

Olson, W. K., and P. J. Flory. 1972. Spatial configuration of polynucleotide chains. I. Steric interactions in polyribonucleotides: a virtual bond model. *Biopolymers.* 11:1–23.

Orita, M., F. Nishikawa, T. Shimayama, K. Taira, Y. Endo, and S. Nishikawa. 1993. High-resolution NMR study of a synthetic oligoribonucleotide with a tetranucleotide GAGA loop that is a substrate for the cytotoxic protein, ricin. *Nucleic Acids Res.* 21:5670–5678.

Philippe, C., C. Portier, M. Mougel, M. Grunberg-Manago, J. P. Ebel, B. Ehresmann, and C. Ehresmann. 1990. Target site of *Escherichia coli* ribosomal protein S15 on its messenger RNA. Conformation and interaction with the protein. *J. Mol. Biol.* 211:415–426.

Puglisi, J. D., R. Tan, B. J. Calnan, A. D. Frankel, and J. R. Williamson. 1992. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science.* 257:76–80.

Raghunathan, G., R. L. Jernigan, H. T. Miles, and V. Sasisekharan. 1991. Conformational feasibility of a hairpin with two purines in the loop. 5′-d-GGTACIAGTACC-3′. *Biochemistry.* 30:782–788.

Romaniuk, P. J., P. Lowary, H. N. Wu, G. Stormo, and O. C. Uhlenbeck. 1987. RNA binding site of R17 coat protein. *Biochemistry.* 26: 1563–1568.

Seeman, N. C., J. M. Rosenberg, F. L. Suddath, J. J. Park Kim, and A. Rich. 1976. A simplified alphabetical nomenclature for dihedral angles in the polynucleotide backgone. *J. Mol. Biol.* 104:142–143.

Shen, L. X., Z. Cai, and I. Tinoco, Jr. 1995. RNA structure at high resolution. *FASEB J.* 9:1023–1033.

Soumpasis, D. M. 1984. Statistical mechanics of the B-Z transition of DNA: contribution of diffuse ionic interactions. *Proc. Natl. Acad. Sci. USA.* 81:5116–5120.

Soumpasis, D. M., and C. S. Tung. 1988. A rigorous basepair oriented description of DNA structures. *J. Biomol. Struct. Dyn.* 6:397–420.

Srinivasan, A. R., and W. K. Olson. 1987. Nucleic acid model building: the multiple backbone solutions associated with a given base morphology. *J. Biomol. Struct. Dyn.* 4:895–938.

Sussman, J. L., S. R. Holbrook, R. W. Warrant, and S.-H. Kim. 1978. Crystal structure of yeast phenylalanine tRNA. I. Crystallographic refinement. *J. Mol. Biol.* 123:607–630.

Tuerk, C., P. Gauss, C. Thermes, D. R. Groebe, M. Gayle, N. Guild, G. Stormo, Y. d'Aubenton-Carafa, O. C. Uhlenbeck, and I. Tinoco, Jr. 1988. CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci. USA.* 85:1364–1368.

Tung, C. S. 1993. Computation of Biomolecular Structures: Achievements, Problems, and Perspectives. D. M. Soumpasis and T. M. Jovin, editors. Springer Verlag, New York. 87–97.

Tung, C. S., T. I. Oprea, G. Hummer, and A. E. Garcia. 1996. Three-dimensional model of a selective theophylline-binding RNA molecule. *J. Mol. Recognit.* (in press).

Tung, C. S., and D. M. Soumpasis. 1996. Structural prediction of A- and B-DNA duplexes based on coordinates of the phosphorus atoms. *Biophys. J.* 70:917–923.

Tung, C. S., D. M. Soumpasis, and G. Hummer. 1994. An extension of the rigorous base-unit oriented description of nucleic acid structures. *J. Biomol. Struct. Dyn.* 11:1327–1344.

Valegard, K., J. B. Murray, P. G. Stockley, N. J. Stonehouse, and L. Liljas. 1994. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature.* 371:623–626.

Weiner, S. J., P. A. Kollman, D. A. Case, U. Chandra Singh, C. Ghio, G. Algona, S. Profeta, Jr., and P. Weiner. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106:765–784.

Weiner, S. J., P. A. Kollman, D. T. Nguyen, and D. A. Case. 1986. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* 7:230–252.

Witherell, G. W., J. M. Gott, and O. C. Uhlenbeck. 1991. Specific interaction between RNA phage coat proteins and RNA. *Prog. Nucleic Acids Res. Mol. Biol.* 40:185–220.

Witherell, G. W., and O. C. Uhlenbeck. 1989. Specific RNA binding by Q beta coat protein. *Biochemistry.* 28:71–76.

Zhou, N., and H. J. Vogel. 1993. Two-dimensional NMR and restrained molecular dynamics studies of the hairpin d(T8C4A8): detection of an extraloop cytosine. *Biochemistry.* 32:637–645.

Zhurkin, V. B., Y. P. Lysov, and V. I. Ivanov. 1978. Different families of double-stranded conformations of DNA as revealed by computer calculations. *Biopolymers.* 17:377–412.