JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

RNA Conformational Sampling: II. Arbitrary Length Multinucleotide Loop Closure

C. H. Mak,*[,†] Wen-Yeuan Chung,[‡] and Nikolay D. Markovskiy[†]

[†]Department of Chemistry, University of Southern California, Los Angeles, California 90089-0482, United States

[‡]Department of Mechanical Engineering, Chinese Cultural University, Taipei, Taiwan, Republic of China

ABSTRACT: In this paper, we describe how the inverse kinematic solution to the loop closure problem may be generalized to reclose a RNA segment of arbitrary length containing any number of nucleotides without disturbing the atomic positions of the rest of the molecule. This generalization is made possible by representing the boundary conditions of the closure in terms of a set of virtual coordinates called RETO, allowing the inverse kinematics to be reduced from the original six-variable/six-constraint problem to a four-variable/four-constraint problem. Based on this generalized closure solution, a new Monte Carlo algorithm has been formulated and implemented in a fully atomistic RNA simulation capable of moving loops of arbitrary lengths using torsion angle updates exclusively. Combined with other conventional Monte Carlo moves, this new algorithm is able to sample large-scale RNA chain conformations much more efficiently. The utility of this new class of Monte Carlo moves in generating large-loop conformational rearrangements is demonstrated in the simulated unfolding of the full-length hammerhead ribozyme with a bound substrate.

1. INTRODUCTION

The loop closure problem was first considered by Go and Scheraga in 1970 for polypeptide chains.[1] In biopolymers like proteins and nucleic acids, their bond lengths and bond angles are largely fixed. The flexibility of these molecules is therefore primarily derived from bond torsions. But for very long chains, even minute motions in a single torsion angle affect the coordinates of many atoms simultaneously, which may lead to excessive steric overlaps among the atoms being moved. For torsion angle moves in the conformational sampling of linear polymers to be practical, multiple torsion angles must be moved at the same time so that the atomic positions of the majority of the chain can remain relatively unperturbed. The loop closure problem seeks solutions for the possible sets of torsion angles inside a chain segment that have to be moved simultaneously in order to maintain the atomic coordinates of the rest of the chain fixed. Various formulations and extensions of the loop closure problem have been reported since Go and Scheraga's work.[2−6] When applied to simulations of polymer conformational sampling, these loop closure solutions may be incorporated into a class of Monte Carlo (MC) schemes called concerted rotation or "conrot" moves.[7−15] These concerted rotation algorithms have been successfully used in conformational sampling to generate trial moves for short segments along the backbone of linear polymers, proteins, and nucleic acids. A related set of methods called "rebridging MC"[16,17] have also been used previously to reclose peptide chains as well as small ring structures.

In the original formulation of Go and Scheraga, they considered the case of a protein where the rotatable (ϕ,ψ) angles are separated by fixed peptide bonds in the trans conformation. They concluded that at least six torsion angles belonging to a tripeptide sequence must be moved simultaneously in order for the atomic coordinates of the rest of the chain to remain fixed. Dodd et al.[4] and Deem and Bader[9] used this to develop enhanced MC schemes to update the conformation of short segments in linear polymer chains and proteins. Later, Dinner[10] extended this formulation to allow arbitrary values for the fixed intervening torsion angles between the rotatable ones. Extensions of this formulation have also been applied to nucleic acids.[10,15] Different formulations of the same loop closure problems were also reported recently for proteins by Coutsias et al.[6,18] and for single nucleotides and ribose by Mak.[19]

In this paper, we consider the closure of an arbitrarily long segment on the backbone of a linear polymer. We generalize the closure problem so that it may be applied to reclose loops of any length. This generalized formulation applies to structures from as small as a single nucleotide in a RNA to loops of any length with fixed intervening torsion angles within three connecting segments. This generalized formulation relies on the reduction of the original six-variable/six-constraint problem to a four-variable/four-constraint problem, and a simple solution is derived using geometrics of rigid bodies well-known in the field of the kinematics of mechanisms. In various limits, our formulation is related to the conrot algorithm[7−15] as well as the rebridging MC method.[16,17] The generalized formulation allows for an efficient numerical or analytical solution. We show how this generalized closure solution may be used to devise several MC moves that are able to sample large-scale loop conformational changes in RNAs. We demonstrate the utility of these new MC moves in an all-atom simulation, studying the possible unfolding pathways of a ribozyme containing 63 nucleotides.

2. SINGLE-NUCLEOTIDE CLOSURE AND THE RETO COORDINATES

We begin with a brief review of the single-nucleotide RNA closure problem. This has been considered by one of us in a

Received: December 2, 2010
Published: March 11, 2011

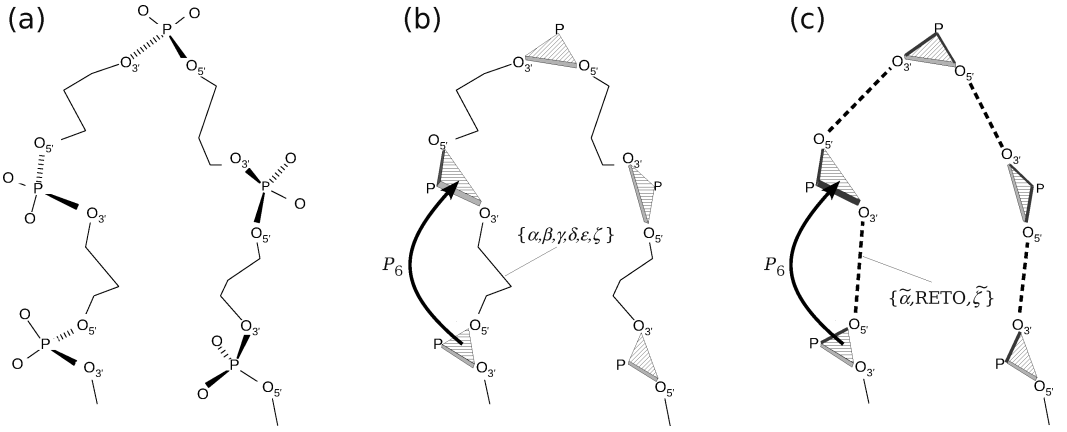

Figure 1. (a) Drawing shows a portion of the nucleic acid backbone. Starting from the 5′ end, given the coordinates of the first phosphate group, the coordinates of all subsequent phosphate groups are uniquely determined if all intervening torsion angles are known. (b) Since bond lengths and bond angles are fixed, the coordinates of all five atoms in each phosphate group are uniquely determined if only the position of P and the orientation of the triangle $O_{3'}PO_{5'}$ are known. From one phosphate triangle, the position and orientation of the next phosphate triangle is known by either specifying the six intervening torsion angles, $\{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta\}$, or by P_6. (c) Starting from one phosphate triangle, the position and orientation of the next phosphate triangle is uniquely determined if the six variables {, RETO, $\tilde{\zeta}$} are specified. These are represented by dashed lines.

previous paper,[19] which we shall refer to as paper (I). Figure 1a shows several nucleotides along the backbone of a nucleic acid. For clarity, the side chains and sugar rings have been removed for the drawing. In the simplest variant of the loop closure problem, we assume all bond lengths and bond angles are fixed and consider bond torsions as the only motions in the molecule. (This assumption may easily be relaxed if desired.) Starting from the 5′ end, if the position and orientation of the first phosphate group is known, then the positions and orientations of all subsequent phosphate groups can be uniquely determined if all intervening torsion angles are specified. As such, the coordinates of every atom along the backbone of the chain may be easily reconstructed based on a knowledge of the coordinates of the first phosphate group and all the backbone torsion angles. In the case of RNAs, the forward kinematic problem is to solve for the positions of all the phosphate groups given the backbone torsion angles. Clearly, the forward kinematic problem of transforming from torsion angles to real-space coordinates is rather trivial.

The inverse kinematic problem of transforming from real-space coordinates back to torsion angles is much more involved. We can phrase the problem this way: If the coordinates of each phosphate group along the chain are specified, is it possible to solve for the intervening torsion angles given that the bond lengths and bond angles are rigid?

To begin, we recognize first that while every phosphate group has five atoms, the coordinates of all of them are uniquely determined if only the position of P and the orientation of the $O_{3'}PO_{5'}$ triangle is known. For convenience, we will call this the "phosphate triangle". The coordinates of the rest of the atoms in this phosphate group are then fixed by the rigid bond lengths and bond angles. The phosphate triangles are shown in Figure 1b, where each phosphate group has been replaced by an oriented phosphate triangle. Using this reduction, the inverse kinematic problem may be restated alternatively as: If the position and orientation of every phosphate triangle is specified, is it possible to solve for all the torsion angles along the chain?

To be a mathematically well-posed problem, the number of degrees of freedom must be greater than or equal to the number of constraints. Six constraints (three translations and three Euler angles) are involved in specifying the position and orientation of each phosphate triangle relatively to the previous one. We will call these six variables P_6. For the inverse kinematic problem to be solvable, there must be a minimum of six torsion angles in each nucleotide. Interestingly, there are exactly six backbone torsion angles inside each nucleotide. These six torsion angles, shown in Figure 2, are conventionally labeled α to ζ. Consequently, the smallest reclosible loop in a RNA is a single nucleotide, and the inverse kinematic problem is to seek the transformation $P_6 \rightarrow \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta\}$. This is known as the 6R problem in robotics.[20] (If there were fewer than six free torsion angles in a nucleotide, the inverse kinematic problem can no longer be phrased in terms of the phosphate triangles, but the problem can still be solved for any segments with six torsion angles.)

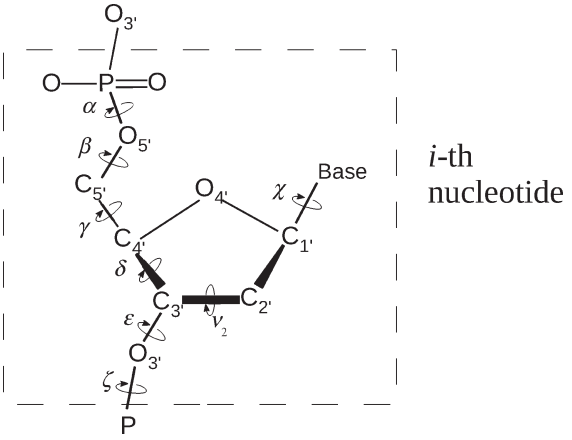

Figure 2. The six torsion angles along the backbone of a nucleotide are conventionally labeled α to ζ. For simplicity, $O_{2'}$ has been omitted from the drawing.

To further simplify the loop closure problem, paper (I) shows that the position and orientation of one phosphate triangle relative to the previous one may be given in terms of a new set of internal coordinates. These internal coordinates are defined in Figure 3, where r is the distance from O'_5 to the next O'_3, η is the

dx.doi.org/10.1021/ct100681j |*J. Chem. Theory Comput.* 2011, 7, 1198–1207

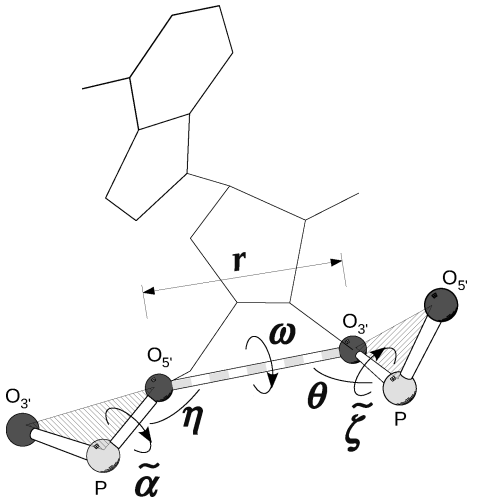

Figure 3. Internal coordinates used in the RETO representation to describe the relative position and orientation of two consecutive phosphate triangles along the backbone of a RNA. (Portions of this and other figures in this paper were generated using molscript).[21]

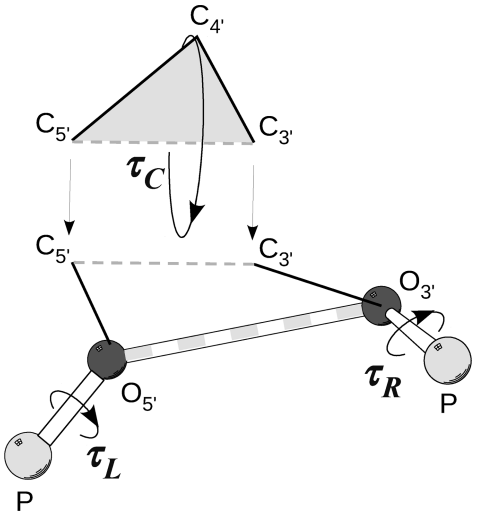

Figure 4. Drawing illustrating the single-nucleotide closure solution. RETO is fixed by the positions of P, $O_{5'}$, $O_{3'}$, and the next P. The chain is divided into three rigid elements: $O_{5'}$ with its bonds to P and $C_{5'}$ form the left element, $C_{4'}$ with its bonds to $C_{5'}$ and $C_{3'}$ the center element, and $O_{3'}$ and its bonds to $C_{3'}$ and P the right element. The closure is obtained by the rotations τ_L and τ_R to match the native $C_{5'}-C_{3'}$ distance, the reattachment of the center element, and the rotation τ_C around the $C_{5'}-C_{3'}$ axis to produce the correct $O_{5'}-C_{5'}-C_{4'}$ bond angle.

$P-O'_5-O'_3$ virtual bond angle, θ is the $O'_5-O'_3-P$ virtual bond angle, and ω is the $P-O'_5-O'_3-P$ virtual torsion angle. These four variables, $\{r, \eta, \theta, \omega\}$, are given the special name "RETO coordinate" because they play a special role in the generalized loop closure solution. The four atoms, P, $O_{5'}$, $O_{3'}$ and the next P, are said to form a structure called a "RETO element". The RETO coordinates are simply the minimal set of variables needed to uniquely specify the internal geometry of any RETO element. We also define to be the $O'_3-P-O'_5-O'_3$ virtual torsion angle on the 5′ end outside the RETO element, and $\tilde{\zeta}$ to be the $O'_5-O'_3-P-O'_5$ virtual torsion angle on the 3′ end. Together, this set of six variables, {, RETO, $\tilde{\zeta}$}, specifies the position and orientation of one phosphate triangle relative to the previous one. This is shown in Figure 1c, where the phosphate groups are depicted as connected RETO elements, and their relative positions and orientations are now specified by the RETO variables and the torsion angles and $\tilde{\zeta}$. This is shown in greater detail in Figure 3, which illustrates that starting from one phosphate triangle, specifying {, RETO, $\tilde{\zeta}$} will uniquely determine the position and orientation of the next phosphate triangle. In terms of the new RETO coordinates, the inverse kinematic problem is now to seek the transformation {, RETO, $\tilde{\zeta}$} → {α, β, γ, δ, ε, ζ}, which corresponds to the mapping from Figure 1c back to 1b.

Paper (I) shows that with the introduction of the RETO coordinates, the original six-constraint/six-variable problem may be reduced to a simpler four-constraint/four-variable problem. This is because the variable maps directly onto α, while $\tilde{\zeta}$ maps onto ζ. In mathematical terms, the Jacobian matrix (J) of the transformation {, RETO, $\tilde{\zeta}$} → {α, β, γ, δ, ε, ζ}:

$$\mathbf{J} = \frac{\partial(\tilde{\alpha}, \mathrm{RETO}, \tilde{\zeta})}{\partial(\alpha, \beta, \gamma, \delta, \varepsilon, \zeta)} \qquad (1)$$

turns out to be block diagonal, with forming a 1 × 1 block with α, RETO forming a 4 × 4 block with {β, γ, δ, ε}, and $\tilde{\zeta}$ forming a 1 × 1 block with ζ. Furthermore, it is easy to show that the 1 × 1 block between and α is unity for all values of , and the same is true for the 1 × 1 block between $\tilde{\zeta}$ and ζ.

A detailed solution for the inverse kinematic transformation RETO → {β, γ, δ, ε} has been given in paper (I). Briefly, the solution is illustrated in Figure 4 and it proceeds as follows: Given the RETO coordinates, the relative positions of P, $O_{5'}$, $O_{3'}$, and the next P are fixed. With the known $P-O_{5'}$ and $O_{5'}-C_{5'}$ bond lengths and the $P-O_{5'}-C_{5'}$ bond angle, we can consider the three atoms P, $O_{5'}$, and $C_{5'}$ as a rigid element (the left element). Similarly, with the known $C_{3'}-O_{3'}$ and $O_{3'}-P$ bond lengths and the $C_{3'}-O_{3'}-P$ bond angle, we can also consider the three atoms $C_{3'}$, $O_{3'}$, and P as another rigid element (the right element). Finally, the atom $C_{4'}$ and its two adjacent bonds to $C_{5'}$ and $C_{3'}$ may be considered as yet another rigid element (the center element), which is shown in Figure 4 as the gray triangle. To construct the closure solution, we imagine first detaching the center element from the loop. The left element is then free to rotate about the $P-O_{5'}$ bond through some rotation angle τ_L. Similarly, the right element can also rotate about the $P-O_{3'}$ bond through angle τ_R. In order for the center element to be reattached properly, the distance between $C_{5'}$ and $C_{3'}$, shown as a gray dashed line between the left and right elements, must match the $C_{5'}-C_{3'}$ distance on the lower edge of the gray triangle of the center element. If we consider τ_L as the input, then the output angle τ_R that produces the correct $C_{5'}-C_{3'}$ distance will form a discrete set of points for every input τ_L. For each one of these τ_R values, we can then reattach the center element. But in order to also produce the correct $O_{5'}-C_{5'}-C_{4'}$ bond angle, the center element must also be rotated about the $C_{5'}-C_{3'}$ axis by the proper angle(s) τ_C. After this, we can measure the $C_{4'}-C_{3'}-O_{3'}$ angle. If this matches the correct $C_{4'}-C_{3'}-O_{3'}$ angle, then the closure is solved. The solution of the closure problem is therefore obtained by expressing the output $C_{4'}-C_{3'}-O_{3'}$ angle as a function of the input τ_L. This is followed by a root search to determine the value(s) of τ_L which produce(s) the correct angle matching the native $C_{4'}-C_{3'}-O_{3'}$ angle. After the solution of this four-constraint/four-variable problem is obtained, the last two torsion angles and $\tilde{\zeta}$ are easily determined since they are simply equal to their counterparts α and ζ, plus some offsets.

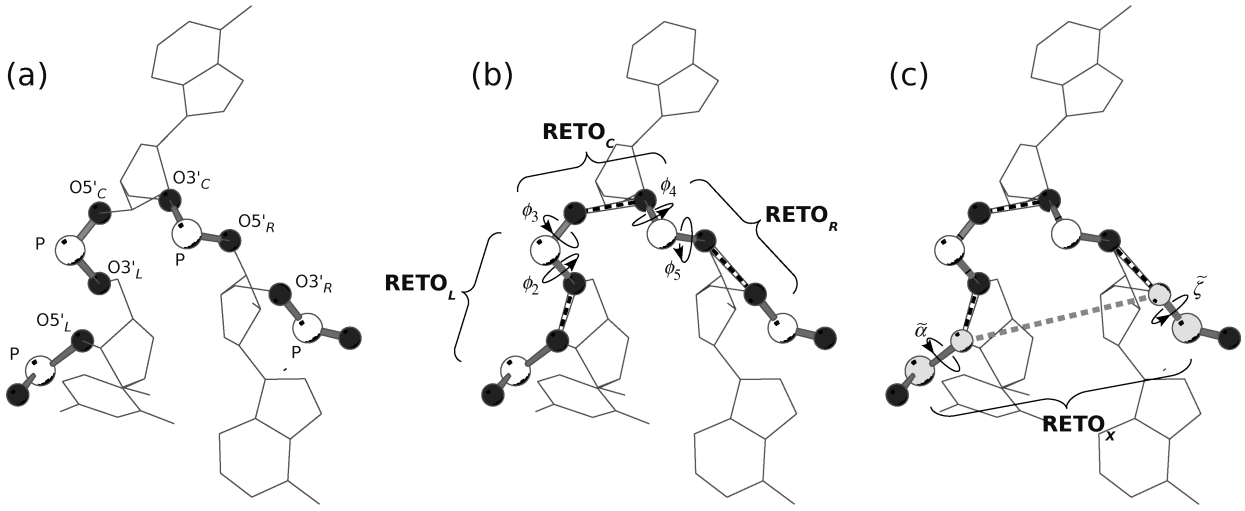

Figure 5. The three-nucleotide loop closure problem. (a) Drawing showing phosphate triangles in three consecutive nucleotides. (b) The RETO representation of the left, center and right elements, each one containing a one-nucleotide segment. Their relative orientations are defined by four intervening virtual torsion angles: ϕ_2, ϕ_3, ϕ_4, and ϕ_5. (c) The boundary conditions are specified by $RETO_X$, which defines the geometry of the four boundary atoms shown in gray for this three-nucleotide sequence. The orientations of the rest of chain to the left (the 5′ end) and the right (the 3′ end) are specified by two additional virtual torsion angles and $\tilde{\zeta}$.

In addition to the backbone torsion angles, the ribose also has internal torsions. These are coupled to the mainchain torsion angle δ. When the nucleotide backbone is reclosed, δ is modified, which requires the ribose to be reclosed simultaneously. This problem of reclosing the ribose is analogous to the conformation of prolines in a polypeptide, which has been considered by Ho et al.[18] who shows that the standard loop closure solution may be applied with slight modification to reclose any five-membered ring structure. As such, one can simply use the same closure algorithm to reclose the ribose, and full details have been given in paper (I).

3. MULTINUCLEOTIDE CLOSURE

The inverse kinematic solution for reclosing a RNA backbone sequence containing more than a single nucleotide is a straightforward extension of the single-nucleotide closure solution. The extension is facilitated by the RETO coordinates described in the last section. To illustrate how this works, we will start with the closure of a three-nucleotide loop, a schematic of which is shown in Figure 5a. The P atoms are colored white, and the O atoms are colored black. For clarity, only the two phosphate oxygens on the backbone are shown. The rest of the chain is displayed as sticks. The 5′ end of the chain is on the left. We identify the O atoms on the leftmost nucleotide by the subscript L, those on the center nucleotide by C, and the ones on the rightmost nucleotide by *I*.

Figure 5b shows how this three-nucleotide sequence may be divided into three connected RETO elements. $RETO_L$ on the left consists of the P, $O_{5'}$, and $O_{3'}$ atoms from the first nucleotide plus the P atom from the second nucleotide. $RETO_C$ in the center consists of P, $O_{5'}$, and $O_{3'}$ from the second nucleotide plus P from the third and similarly with $RETO_R$ on the right. Two consecutive RETO elements are connected through the P atom they share. In Figure 5b, we have indicated the virtual bond between $O_{5'}$ and $O_{3'}$ in each RETO element using a thick dashed line. As described in Section 2, the geometry of each RETO element is specified by its four RETO variables $\{r, \eta, \theta, \omega\}$.

In addition to defining the RETO variables of the three internal elements, we can also define a set of RETO variables specifying the relative positions of the four terminal atoms, P and $O_{5'}$ on the 5′ end and $O_{3'}$ and P on the 3′ end. We called these the $RETO_X$ variables. The four atoms that define $RETO_X$ are colored gray in Figure 5c. Similar to the single-nucleotide case in Figure 3, two additional virtual torsion angles and $\tilde{\zeta}$ are needed to determine how $RETO_X$ is oriented relative to the rest of the chain outside.

If the internal structure of each RETO element is fixed, the complete conformation of this three-nucleotide sequence can be uniquely reconstructed if we specify the RETO coordinates of each element as well as the two virtual torsion angles between each pair of connected RETO elements. These torsion angles are depicted in Figure 5b. The two virtual torsion angles connecting $RETO_L$ and $RETO_C$ are ϕ_2, the $O_{5'}-O_{3'}-P-O_{5'}$ angle, and ϕ_3, the $O_{3'}-P-O_{5'}-O_{3'}$ angle. Similarly, the two virtual torsion angles connecting $RETO_C$ and $RETO_R$ are ϕ_4, the $O_{5'}-O_{3'}-P-O_{5'}$ angle, and ϕ_5, the $O_{3'}-P-O_{5'}-O_{3'}$ angle. Together with $RETO_L$, $RETO_C$, and $RETO_R$, specifying these four virtual torsion angles will uniquely define the complete conformation of the chain, producing the final positions of $O_{3'}$ and P on the 3′ end relative to the positions of P and $O_{5'}$ on the 5′ end.

In terms of the variables defined above, the inverse kinematic problem of closing a three-nucleotide loop is to seek the transformation $RETO_X \rightarrow \{\phi_2, \phi_3, \phi_4, \phi_5\}$, given known $RETO_L$, $RETO_C$, and $RETO_R$ and fixed bond lengths and bond angles. Once this problem is solved, the two virtual torsion angles and $\tilde{\zeta}$ in Figure 5c are then trivially given by the native torsion angle α on the 5′ end and ζ on the 3′ end plus some offsets, just as in the single-nucleotide closure problem.

The solution of the problem $RETO_X \rightarrow \{\phi_2, \phi_3, \phi_4, \phi_5\}$ follows closely the single-nucleotide solution described in the last section. The only difference is that the left, center, and right rigid elements are now the $RETO_L$, $RETO_C$ an $RETO_R$ in Figure 5, instead of the three-atom segments shown in Figure 4. Other than this, the rest of the solution is identical to the single-nucleotide case, and its details will not be repeated again.

At this point, it should also be clear that the solution described above is applicable to not only a three-nucleotide loop, but it can also be applied to close loops of any size. To close larger loops, the only modification needed is to redefine the $RETO_L$, $RETO_C$, or $RETO_R$ elements so that each may encompass more than one nucleotide. For the three-nucleotide closure above, each of the $RETO_L$, $RETO_C$, and $RETO_R$ elements contain only one nucleotide, and we may call this a $1 + 1 + 1$ closure. To close larger loops, such a nine-nucleotide loop for instance, we may take each of $RETO_L$, $RETO_C$, and $RETO_R$ to be a rigid three-nucleotide sequence and perform a $3 + 3 + 3$ closure. But the same nine-nucleotide loop may also be closed using a $2 + 5 + 2$ or a $2 + 6 + 1$ closure or any other combination of sizes. Furthermore, if desired, the closure may even be applied in a cascading or a recursive manner. For example, we may carry out a $1 + 1 + 1$ closure for each of $RETO_L$, $RETO_C$, and $RETO_R$ and then reclose the composite nine-nucleotide loop by a second $3 + 3 + 3$ closure. Clearly, by using different combinations of single-nucleotide and $l_L + l_C + l_R$ multinucleotide closures, we may reclose loops of any arbitrary length.

4. MONTE CARLO ALGORITHMS USING MULTINUCLEOTIDE CLOSURE

We can incorporate the multinucleotide closure solution above to construct a set of novel MC moves that are capable of updating the conformation of an arbitrarily long RNA sequence. We will consider three variants of this MC method, which differ in complexity. In certain limits, they are related to the conrot algorithm[7−15] in various ways, but a driver angle is not used in our method. We will describe each MC algorithm by referring to the drawings in Figure 5, but it is important to remember that any of the $RETO_L$, $RETO_C$, and $RETO_R$ elements may contain more than one nucleotide.

4.1. Monte Carlo Variant MC1. This is the simplest of all three MC variants. In this MC move, we will keep the $RETO_X$ geometry as well all three internal RETO elements (L, C, and R) rigid and let the closure select a new solution to reclose the loop into a different conformation. Since we are not altering $RETO_X$, the rest of the chain to the left on the 5′ end of the drawing in Figure 5 as well that to the right on the 3′ end are fixed during this MC move. Because the internal structures of $RETO_L$, $RETO_C$, and $RETO_R$ are also frozen, the reclosing of the loop may easily be done by randomly picking one closure solution from among the possible sets of $\{\phi_2, \phi_3, \phi_4, \phi_5\}$ that solve the closure problem with $RETO_X$ as the boundary condition. Because the constraints are in the $RETO_X$ instead of in torsion angle space, a Jacobian is involved. The Jacobian that is needed is

$$J = \left| \frac{\partial(\phi_2, \phi_3, \phi_4, \phi_5)}{\partial(r_X, \eta_X, \theta_X, \omega_X)} \right| \tag{2}$$

for the transformation $\{\phi_2, \phi_3, \phi_4, \phi_5\} \rightarrow RETO_X$. This is more easily computed from its reciprocal $|\partial(r_X, \eta_X, \theta_X, \omega_X)/\partial(\phi_2, \phi_3, \phi_4, \phi_5)|$ using numerical differentiation. This MC move therefore consists of three simple steps:

(1) Randomly select a residue as the starting point of the $RETO_L$ element. Assign the next l_L nucleotides to the $RETO_L$ element, l_C nucleotides to $RETO_C$, and then l_R nucleotides to $RETO_R$.
(2) With $RETO_X$, $RETO_L$, $RETO_C$, and $RETO_R$ fixed, use a multinucleotide closure to generate all loop solutions. Randomly select one of the solutions to reclose the loop.
(3) Accept or reject the new loop conformation based on its energy compared to the energy of the old state using the Metropolis[22] rule:

$$P = \min\left[1, \frac{J' \exp(-E'/kT)}{J \exp(-E/kT)}\right] \tag{3}$$

where E is the energy of the old conformation, E' is the energy of the new, J' and J are the Jacobians of the new and old conformations, respectively, k is Boltzmann's constant, and T is the temperature.

Clearly, steps 1 and 2 generate the trial state with a symmetric transition probability. Together with the Metropolis acceptance criterion in step 3, this MC strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L, l_C, and l_R may either be fixed or chosen randomly.

4.2. Monte Carlo Variant MC2. In the second MC variant, we allow the internal RETO elements to be flexible while keeping $RETO_X$ fixed. But instead of moving the variables $\{r, \eta, \theta, \omega\}$ in each of $RETO_L$, $RETO_C$, and $RETO_R$ directly, we simply apply random displacements to each of their internal torsion angles to arrive at a new RETO geometry for each element. For instance, when applying this MC variant to a four-nucleotide sequence using a $1 + 1 + 2$ closure, we would randomly displace the β, γ, δ, and ε torsion angles inside the $RETO_L$ and $RETO_C$ elements first. This will generate new $RETO_L$ and $RETO_C$ geometries. Then for the two-nucleotide-long $RETO_R$, we will displace all torsion angles between β of the first residue in $RETO_R$ and ε in the last residue of $RETO_R$. This would result in new geometries for all three internal RETO elements, and we would then reclose the loop for the fixed $RETO_X$ on the outside into a new loop conformation. Since the density of conformational states is uniform in torsion angle space, generating displacements for $RETO_L$, $RETO_C$, and $RETO_R$ this way will not require an additional Jacobian other than the one already used in MC1. This MC move therefore consists of the following steps:

(1) Randomly select a residue as the starting point of the $RETO_L$ element. Assign the next l_L nucleotides to the $RETO_L$ element, l_C nucleotides to $RETO_C$, and then l_R nucleotides to $RETO_R$.
(2) Using the current geometries of $RETO_L$, $RETO_C$, and $RETO_R$, obtain the number of loop closure solutions N with boundary condition fixed by $RETO_X$.
(3) Apply random displacements to all internal torsion angles in each of the $RETO_L$, $RETO_C$, and $RETO_R$ elements.
(4) With the new geometries for the $RETO_L$, $RETO_C$, and $RETO_R$ elements, use a multinucleotide closure to generate all new loop solutions for the fixed $RETO_X$. Let the number of new solutions be N'.
(5) Randomly select from one of the new solutions to reclose the loop.
(6) Accept the new loop conformation using the probability:

$$P = \min\left[1, \frac{N'J' \exp(-E'/kT)}{NJ \exp(-E/kT)}\right] \tag{4}$$

It can be shown easily that this MC scheme strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L, l_C, and l_R may either be fixed or chosen randomly.

4.3. Monte Carlo Variant MC3. In this third and final MC variant, we keep the three internal RETO elements (L, C, and R) frozen and allow only $RETO_X$ to move. To move $RETO_X$, we have no alternative but to displace $\{r_X, \eta_X, \theta_X, \omega_X\}$ directly. So we would reclose the loop using the original $RETO_L$, $RETO_C$, and $RETO_R$ elements onto the new $RETO_X$ constraint. This MC move therefore consists of the following steps:

(1) Randomly select a residue as the starting point of the $RETO_L$ element. Assign the next l_L nucleotides to the $RETO_L$ element, l_C nucleotides to $RETO_C$, and then l_R nucleotides to $RETO_R$.
(2) Using the current geometries of $RETO_L$, $RETO_C$, and $RETO_R$, obtain the number of loop closure solutions N with boundary condition fixed by $RETO_X$.
(3) Generate a trial $RETO_X$ by applying random displacements to $\{r_X, \eta_X, \theta_X, \omega_X\}$.
(4) Displace the rest of the molecule on the 5′ end external to $RETO_X$ and on the 3′ end so that their relative positions and orientations are consistent with the trial $RETO_X$. Then apply random displacements to , which connects $RETO_X$ to the rest of the molecule on the 5′ end, and to $\tilde{\zeta}$, which connects $RETO_X$ to the rest of the molecule on the 3′ end.
(5) Use a multinucleotide closure to generate all new loop solutions for the new $RETO_X$ using the old $RETO_L$, $RETO_C$, and $RETO_R$ geometries. Let the number of new solutions be N'.
(6) Randomly select from one of the new solutions to reclose the loop.
(7) Accept the new loop conformation using the probability:

$$P = \min\left[1, \frac{N'J' \exp(-E'/kT)}{NJ \exp(-E/kT)}\right] \qquad (5)$$

It can be shown easily that this MC scheme strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L, l_C and l_R may either be fixed or chosen randomly.

5. DISCUSSION

While it is straightforward to construct MC moves based on the single- or multinucleotide closure solutions given above, the practical usefulness of these MC moves in an atomistic simulation is not guaranteed. In fact, our experience with loop-closure MC simulations shows that the effectiveness of MC moves based on loop closure is often severely limited by steric problems. If we start from an already folded conformational state in a long-chain biomolecule, then steric collisions will certainly prevent most of the new loop solutions from being accepted. Even though loop closure has the potential to generate large-scale conformational changes, the reality is that any trial loop conformation in a sterically congested region of the molecule will produce so many new steric overlaps that almost all trial conformations are rejected most all of the time.

As a test, we have used MC variant MC1 on single-nucleotide loops for a number of RNA structures in the PDB database. We discovered that the acceptance rate of any single-nucleotide loop conformations other than the native one is almost always zero. This result should not be too surprising. Any new loop conformation other than the native one almost always cause too many steric collisions, and if the loop is to be reclosed, then it will almost always close back into the native conformation. Therefore, MC1 would be a very ineffective choice for moving single-nucleotide loop or very short loops. For these, MC2 or MC3 would be a better choice. In our MC simulations, we would use MC3 exclusively for single-nucleotide closures. However, single-nucleotide moves will only produce small-scale local motions that are very similar to the thermal motions typically seen in a molecular dynamics (MD) simulation. If an atomistic MC simulation of RNAs is based on single-nucleotide MC3 moves alone, then it will not be a very effective algorithm for studying large-scale loop motions. Multinucleotide MC moves must be added.

Using loop-closure MC moves to close larger loops introduces other problems. When we tried using MC2 or MC3 to reclose loops with nine or more nucleotides starting from the native structure of several RNAs in the PDB database, we discovered that not only was excessive steric collisions a frequent problem but also the loss of favorable base-pairing and base-stacking interactions when the structures of the RETO elements are altered leading to large energy costs. Therefore, while the MC moves MC2 and MC3 allow the RNA molecules to acquire larger conformational changes than MC1, the energy cost associated with this increased flexibility is mostly unfavorable. In fact, the larger the amplitude of the random displacements in $RETO_L$, $RETO_C$, $RETO_R$, or $RETO_X$ we applied, the lower the acceptance rate became. So for reclosing loops larger than single-nucleotide ones, we relied exclusively on MC1, which is the simplest variant to implement numerically.

Since the majority of the new loop conformations are sterically impossible when long loops are reclosed, loop-closure MC simulations are intrinsically highly inefficient. If it was not for the fact that these MC moves possess the unique capability of generating large-scale motions that molecular dynamics or regular MC simulations cannot see, they would have been useless in a practical sense. In order to increase their efficiency and turn loop-closure MC into a practical simulation tool, we have considered various ways to try to speed up the computation. Since the calculation of the total energy is one of the most expensive operations in the algorithm, we tried to accelerate the closure moves by detecting excessive steric overlaps early and screening out sterically impossible closure solutions before their full energies were calculated. To do this, we used a minimum steric radius of 1.5 Å for each atom. If any pair of atoms in the new loop solution come closer than twice the minimum steric radius, then the configuration was rejected. This simple change dramatically reduced the overhead required for calculating the full energy for conformations which were sterically too costly. With this we were able to screen a much larger number of new loop conformations with a smaller amount of computational effort, making loop-closure MC a reasonably practical simulation method for finding large-scale loop motions.

Furthermore, when large loops are reclosed, the choice of the size of the L, C, and R RETO elements is arbitrary. Therefore, we are free to try to find combinations of l_L, l_C, and l_R that would optimize the performance of the loop closure MC. If MC1 is used, the three elements $RETO_L$, $RETO_C$, and $RETO_R$ are fixed; therefore, the internal coordinates of all the atoms inside each element are fixed relative to each other. Since the closure solution rearranges these three elements, keeping each one as a rigid structure, any base pairing or base stacking interactions internal to any element will be preserved, while the base pairing or base stacking interactions between two different RETO elements may

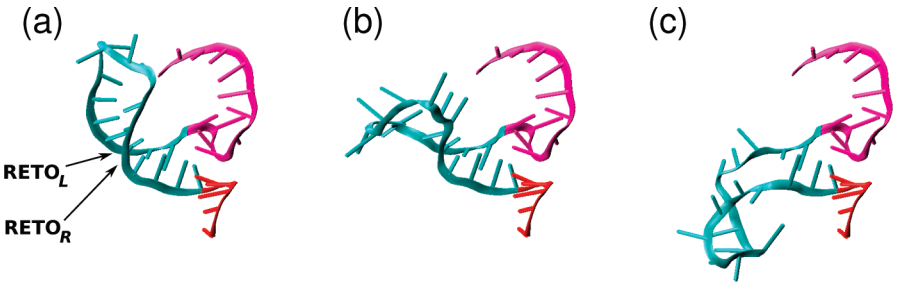

Figure 6. Sample conformations illustrating the effect of a 3+ R + 3 MC1 move on a nucleotide segment: (a) The RETO$_L$ and RETO$_R$ elements randomly selected by this MC move. (b) New conformation generated by reclosing the loop with a rigid RETO$_C$ in between. (c) Another new conformation generated using a similar loop-closure MC move.

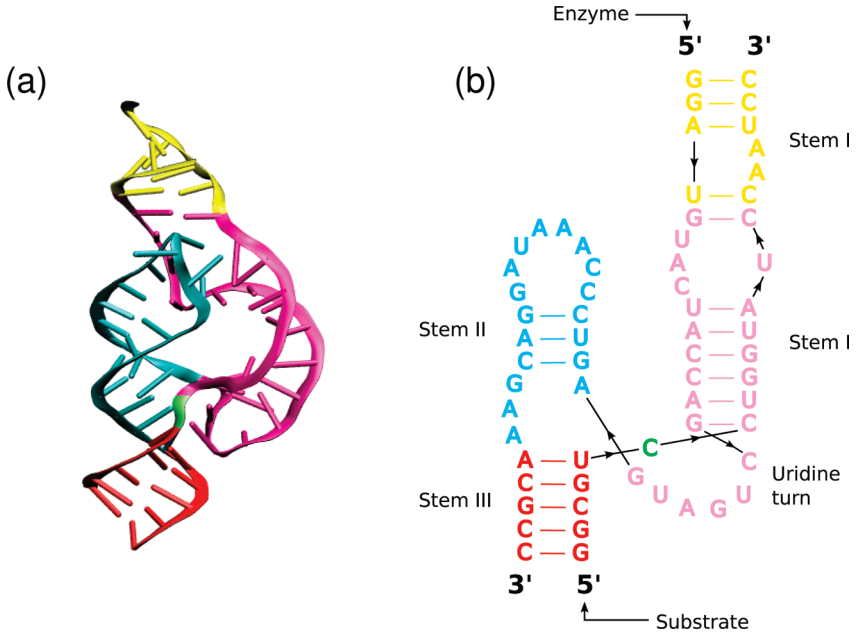

Figure 7. (a) Three-dimensional tertiary structure of the hammerhead ribozyme with a bound substrate (PDB code: 2GOZ). The substrate is the shorter chain, with its 5′ end on the helix colored red (lower left of the figure), going through the nucleolytic site colored green, passing through the stem region colored purple, and finally ending in another stem region colored yellow (top of the figure). (b) Secondary structure corresponding to the three-dimensional structure in (a) using the color coding adopted from Martick and Scott.[23] (Portions of this and other figures in this paper were generated using VMD).[24]

be disrupted when the loop is reclosed into a new conformation. Therefore, in order not to disturb the most favorable interactions, the L, C, and R RETO elements could be chosen so that there are no strong base pairing or base stacking interactions between different RETO elements. The details of how this idea may be implemented in a rigorous MC algorithm will be addressed in a forthcoming paper.

Since the main purpose of this paper is to demonstrate how loop closure solutions may be used to design a set of MC moves to study large-scale loop motions in RNAs, we will only present results from a MC simulation where l_L and l_R are preselected and fixed, while l_C is chosen randomly. In this simulation, we chose RETO$_L$ and RETO$_R$ to be three nucleotide long each. From our experience, the segment lengths $l_L = l_R = 3$ are long enough to provide the loop to be reclosed with adequate flexibility. In this way, the RETO$_L$ and RETO$_R$ elements are essentially acting as hinges that could allow the RETO$_C$ element to unfold. Because closure requires at least one nucleotide on the left and right outside of RETO$_X$ and because RETO$_L$ and RETO$_R$ take up three nucleotide each, while RETO$_C$ must have a minimum of one nucleotide, the length l_C can be chosen randomly according to $l_C = \text{int}(R \times [L - 9]) + 1$, where R is a uniformly distributed random number between 0 and 1, and L is the sequence length of the RNA. We will call this a 3 + R + 3 MC1 move because it employs MC variant 1 with fixed lengths $l_L = l_R = 3$ for RETO$_L$ and RETO$_R$ and a random length l_C for RETO$_C$.

Figure 6 illustrates the unfolding of a stem loop in a small RNA fragment by a 3 + R + 3 MC1 move. In every MC pass of a 3 + R + 3 MC1 move, the first residue of RETO$_L$ as well as the length l_C were selected randomly according to the criteria given above. Corresponding to these choices, RETO$_C$ and RETO$_R$ elements were then determined. A loop closure MC move was then carried out using MC variant 1 described in Section 4.1. Figure 6b gives an example of a new loop conformation derived from a 3 + R + 3 loop closure MC1 move, showing that the three-nucleotide RETO$_L$ and RETO$_R$ elements provide the loop

with sufficient conformational flexibility for it to unfold. Figure 6c shows another unfolded conformation of the same loop derived from the same loop-closure MC move.

6. EXAMPLE: UNFOLDING OF THE HAMMERHEAD RIBOZYME

In this section, we present MC simulation results to demonstrate how loop closure MC moves may be used to study large-scale loop motions in RNAs. The molecule we have chosen for the test is the full-length *Schistosoma* hammerhead ribozyme with a bound substrate, whose structure was determined recently by Martick and Scott[23] (PDB code: 2GOZ). The three-dimensional folded structure of the molecule is shown in Figure 7a with its secondary structure given in Figure 7b, and the color coding has been adopted from Martick and Scott's paper. Together, the ribozyme and substrate contain 63 nucleotides in two chains, A and B, with approximately 2000 atoms in total. The substrate is the shorter chain B, which in Figure 7a is positioned in the front. In the folded structure, the bulge in stem I interacts with the loop in stem II via a number of tertiary contacts.

The results described below were obtained from a 298 K simulation using $3 + R + 3$ MC1 moves plus single-nucleotide MC3 moves with an all-atom model for the hammerhead as well as the substrate. Potential energies were evaluated using the Amber ff98 force field[25,26] and generalized Born/surface area (GB/SA)[27,28] to model the solvent, with no explicit counterions. Without counterions, the native folded state may no longer be the most stable conformation in the simulation because counterions are known to be essential for RNA folding, and the native state may therefore unfold during the simulation.

To demonstrate that the loop closure MC we have proposed is useful for simulating large-scale loop motions in RNAs, we used it to investigate the possible unfolding motions of the hammerhead ribozyme. We carried out 256 independent simulations each starting with the same native structure. Each run consisted of 1000 MC passes, with one pass defined as the molecule having attempted one MC3 single-nucleotide move for each residue and $N/(3 + l_C + 3)$ 3+ $l_C + 3$ MC1 moves for the entire molecule, where l_C was chosen randomly for each pass according to the procedure described in the last section, with the total sequence length of the molecule L = 63. At the end of each run, we examined the degree of unfolding of the final structure by computing its root-mean-square deviation (rmsd) from the native folded state using all atoms in the molecule.

The rmsd data from the simulations are summarized in Figure 8, which shows a histogram of all the results. While a majority of the final structures were very similar to the native folded state (with rmsd <3 Å), there were also a significant number of final structures having rather different conformations from the native state. The final structures obtained from the MC simulations fall into several overlapping but identifiable clusters.

First, structures in cluster A with rmsd between 0 to 3 Å are those conformations that are similar to the native state. Structures in this cluster account for almost 70% of the final population. An example of a structure in this cluster is shown in Figure 9A.

The second cluster with the most distinctive structures is label E in Figure 8. These have rmsd between approximately 21 and 28 Å from the native state. A typical structure from this cluster is shown in Figure 9E, with the native conformation in translucent orange superposed on it. Structures in this cluster are

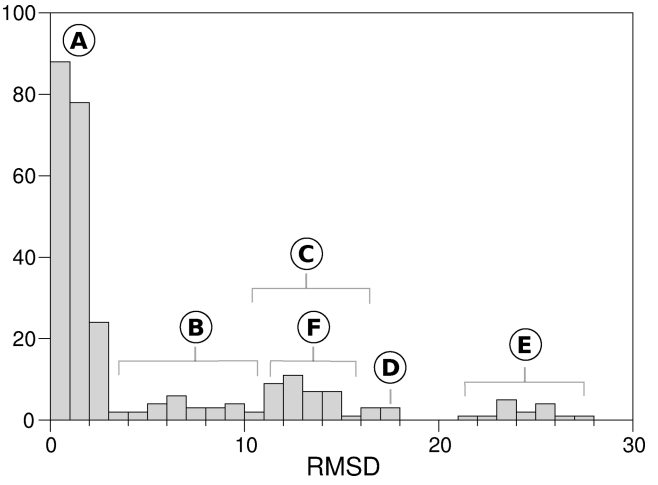

Figure 8. Histogram showing rmsd of final structures derived from 256 independent MC simulations of the hammerhead ribozyme starting from the native conformation. The simulations were carried out using only single-nucleotide MC3 moves plus $3 + R + 3$ MC1 loop closure moves, with rigid bond lengths and bond angles. The final structures fall into six identifiable clusters, labeled A−F.

characterized by the complete unfolding of stem II in chain A, with one hinge ($RETO_L$) in the uridine turn (purple, chain A) and the other ($RETO_R$) in stem III (red, chain A). This class of structures may be important conformational intermediates which act as gateway states to the correct positioning of the enzyme for it to carry out the nucleolytic reaction it is to catalyze.

Structures in cluster B are similar to those in cluster E, but they are characterized by the partial instead of complete unfolding of the stem loop structure in stem II. An example is shown in Figure 9B, with the native structure in translucent orange superposed. For structures in this cluster, the unfolding of stem II is achieved with both hinges ($RETO_L$ and $RETO_R$) on stem II (cyan, chain A). The catalytic site in these structures are largely intact.

The rest of the clusters in Figure 8, namely C, D, and F, were derived from the native state via a very different type of loop motions. While the multinucleotide loop closure solution described in Section 4 has been developed to reclose a chain with a contiguous backbone, an interesting feature is that the same algorithm may actually be used to reclose a segment that spans two or more disjoint chains. For example, if the beginning and end of the $RETO_C$ element happen to be chosen such that $RETO_C$ straddles two different chains, reclosing the loop using MC1 will intrinsically preserve the relative coordinates of both segments in the interior of $RETO_C$, even though they belong to disjoint chains. In this way, if the molecule consists of two or more chains, rigid segments containing residues from multiple chains may be moved at the same time using MC1. Figure 9C and D shows two examples of structures derived from MC moves of this type. The structures belonging to clusters C and D in Figure 8 primarily come from the unfolding of stem I, which is formed from the hybridization of chains A and B. The example shown in Figure 9C came from the bending of a short section of stem I, using one hinge on chain A (purple) and another hinge on chain B (yellow). The one in Figure 9D involves the unfolding of stem I outside the uridine turn, with one hinge on chain A (purple) and the other on chain B (purple). Structures falling under cluster F in Figure 8 are also derived from a similar

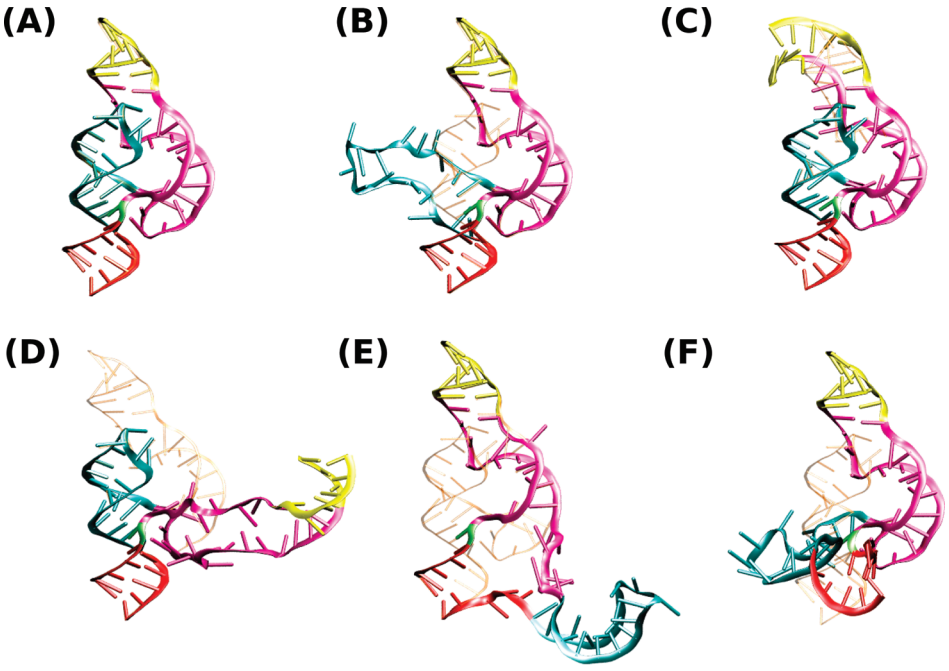

Figure 9. Sample structures from each cluster under the histogram in Figure 8. Structures are color coded using the same scheme as Figure 7. Each structure has been aligned for maximal overlap with the native conformation shown in translucent orange.

interchain loop closure move, but they involve hinges from the two different chains very close to the nucleolytic site, which is colored green, with one hinge on the uridine turn (purple, chain A) and the other bing on stem III (red, chain B). This kind of loop motions leads to a complete unraveling of the catalytic site.

Since the histogram in Figure 8 was derived from an equilibrium sampling of the conformations of the molecule, the logarithm of the population should be proportional to the negative of the free energy divided by kT. Therefore, Figure 8 could be interpreted as a picture of the probability of the conformation as a function of rmsd away from the native state. However, we will refrain from taking this interpretation, since the potential energy function we have used is at best incomplete without a rigorous treatment of the counterions, and by using only two loop-closure MC move ($3 + R + 3$ MC1 and single-nucleotide MC3), the conformations sampled in the simulations may not have been completely equilibrated. Nonetheless, the results in Figure 8 show a close correspondence with the experimental results obtained by Liley et al.[29−31] using fluorescence resonance energy transfer measurements, which show that in the absence of added salt the hammerhead unfolds into an extended three-way junction. This is consistent with the distribution in Figure 8 as well as some of the extended structures in Figure 9 identified by the simulations. This example has clearly demonstrated the feasibility of loop closure MC for investigating large-scale loop motions in RNA simulations.

7. CONCLUSION

In this paper, we have described the formulation of a loop closure problem that is applicable to multinucleotide loops of arbitrary lengths in RNAs. By representing the boundary constraints in the original loop closure problem using a new set of variables called the RETO coordinates, the original six-variable/six-constraint closure problem can be reduced to a simpler four-variable/four-constraint problem. This generalization permits a simple solution of the multinucleotide loop closure problem. In various limits, this formulation is related to the conrot algorithm and the rebridging MC method. Using this generalized solution, we have developed new MC algorithms that are able to reclose loops of any arbitrary lengths to study large-scale loop motions in an all-atom RNA simulation. We have demonstrated the feasibility of the proposed method on the hammerhead ribozyme with a bound substrate and have shown that the simulation produced a large diversity of loop reconfigurations which were otherwise difficult to obtain from a conventional molecular dynamics or MC simulation.

AUTHOR INFORMATION

Corresponding Author

*To whom correspondence should be addressed E-mail: cmak@usc.edu.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under CHE-0713981.

REFERENCES

(1) Go, N.; Scherage, H. Ring closure and local conformational deformations of chain molecules. *Macromolecules* **1970**, *3*, 178.

(2) Wakana, H.; Wako, H.; Saito, N. Monte-Carlo Study on Local and Small-Amplitude Conformational Fluctuation in Hen Egg-White Lysozyme. *Int. J. Pept. Prot. Res.* **1984**, *23*, 315.

(3) Knapp, E. W. Long-Time Dynamics of a Polymer with Rigid Body Monomer Units Relating to a Protein Model - Comparison with the Rouse Model. *J. Comput. Chem.* **1992**, *13*, 793.

(4) Dodd, L. R.; Boone, T. D.; Theodorou, D. N. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.* **1993**, *78*, 961.

(5) Knapp, E. W.; Irgensdefregger, A. Off-Lattice Monte-Carlo Method with Constraints - Long-Time Dynamics of a Protein Model without Nonbonded Interactions. *J. Comput. Chem.* **1993**, *14*, 19.

(6) Coutsias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A. A kinematic view of loop closure. *J. Comput. Chem.* **2004**, *25*, 510.

(7) Bruccoleri, R. E.; Karplus, M. Chain Closure with Bond Angle Variations. *Macromolecules* **1985**, *18*, 2767.

(8) Sartori, F.; Melchers, B.; Bottcher, H.; Knapp, E. W. An energy function for dynamics simulations of polypeptides in torsion angle space. *J. Chem. Phys.* **1998**, *108*, 8264.

(9) Deem, M. W.; Bader, J. S. A configurational bias Monte Carlo method for linear and cyclic peptides. *Mol. Phys.* **1996**, *87*, 1245.

(10) Dinner, A. R. Local deformations of polymers with nonplanar rigid main-chain internal coordinates. *J. Comput. Chem.* **2000**, *21*, 1132.

(11) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

(12) Favrin, G.; Irback, A.; Sjunnesson, F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* **2001**, *114*, 8154.

(13) Ulmschneider, J. P.; Jorgensen, W. L. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J. Chem. Phys.* **2003**, *118*, 4261.

(14) Ulmschneider, J. P.; Jorgensen, W. L. Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation. *J. Am. Chem. Soc.* **2004**, *126*, 1849.

(15) Ulmschneider, J. O.; Jorgensen, W. L. Monte Carlo backbone sampling for nucleic acids using concerted rotations including variable bond angles. *J. Phys. Chem. B* **2004**, *108*, 16883.

(16) Wu, M. G.; Deem, M. W. Efficient Monte Carlo methods for cyclic peptides. *Mol. Phys.* **1999**, *97*, 559.

(17) Wu, M. G.; Deem, M. W. Analytical rebridging Monte Carlo: Application to cis/trans isomerization in proline-containing, cyclic peptides. *J. Chem. Phys.* **1999**, *111*, 6625.

(18) Ho, B. K.; Coutsias, E. A.; Seok, C.; Dill, K. A. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* **2005**, *14*, 1011.

(19) Mak, C. H. RNA conformational sampling: 1. Single-nucleotide loop closure. *J. Comput. Chem.* **2008**, *29*, 926.

(20) Lee, H.-Y.; Liang, C.-G. Displacement analysis of the general spatial 7-link 7R mechanism. *Mech. Mach. Theor.* **1988**, *23*, 219.

(21) Kraulis, P. J. MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures. *J. Appl. Crystallogr.* **1991**, *24*, 946.

(22) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of the state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087.

(23) Martick, M.; Scott, W. G. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **2006**, *126*, 309.

(24) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33.

(25) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(26) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. A Modified Version of the Cornell et al. Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.

(27) Still, W. C.; Tempczyk, A.; Hawlay, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(28) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate radii. *J. Phys. Chem. A* **1997**, *101*, 3005.

(29) Bassi, G. S.; Murchie, A. I.; Walter, F.; Clegg, R. M.; Lilley, D. M. Ion-induced folding of the hammerhead ribozyme: a fluorescence resonance energy transfer study. *EMBO J.* **1997**, *16*, 7481.

(30) Hammann, C.; Lilley, D. M. Folding and activity of the hammerhead ribozyme. *ChemBioChem* **2002**, *3*, 690.

(31) Penedo, J. C.; Wilson, T. J.; Jayasena, S. D.; Khvorova, A.; Lilley, D. M. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA* **2004**, *10*, 880.