Efficient global biopolymer sampling with end-transfer configurational bias Monte Carlo

Gaurav Arya and Tamar Schlick[a)]
Department of Chemistry and Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012

(Received 1 November 2006; accepted 5 December 2006; published online 29 January 2007)

We develop an "end-transfer configurational bias Monte Carlo" method for efficient thermodynamic sampling of complex biopolymers and assess its performance on a mesoscale model of chromatin (oligonucleosome) at different salt conditions compared to other Monte Carlo moves. Our method extends traditional configurational bias by deleting a repeating motif (monomer) from one end of the biopolymer and regrowing it at the opposite end using the standard Rosenbluth scheme. The method's sampling efficiency compared to local moves, pivot rotations, and standard configurational bias is assessed by parameters relating to translational, rotational, and internal degrees of freedom of the oligonucleosome. Our results show that the end-transfer method is superior in sampling every degree of freedom of the oligonucleosomes over other methods at high salt concentrations (weak electrostatics) but worse than the pivot rotations in terms of sampling internal and rotational sampling at low-to-moderate salt concentrations (strong electrostatics). Under all conditions investigated, however, the end-transfer method is several orders of magnitude more efficient than the standard configurational bias approach. This is because the characteristic sampling time of the innermost oligonucleosome motif scales quadratically with the length of the oligonucleosomes for the end-transfer method while it scales exponentially for the traditional configurational-bias method. Thus, the method we propose can significantly improve performance for global biomolecular applications, especially in condensed systems with weak nonbonded interactions and may be combined with local enhancements to improve local sampling. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2428305]

I. INTRODUCTION

Biopolymers such as RNA, DNA, proteins, polysaccharides, and chromatin pose a considerable challenge to thermodynamic sampling. The sheer number of degrees of freedom, and hence, favorable configuration regions involved has made molecular dynamics, Brownian/Langevin dynamics, and Monte Carlo simulations all viable yet limited approaches. Strong nonbonded intramolecular interactions ($>k_BT$) like hydrogen bonding in proteins and polysaccharides and strong electrostatics in single-stranded RNA/DNA and chromatin confer a highly corrugated free energy landscape that further exacerbates their sampling. Hence, prohibitively long simulations are required to sample the entire energy landscape of the biopolymer due to their tendency to get trapped in deep local energy minima separated by large energy barriers. In the case of chromatin, the nucleoprotein complex in which protein spools called nucleosomes are linked to one another by wound DNA, structural complexity resulting from charged histone tails protruding from each nucleosome core which lead to numerous possible folding arrangements make configurational sampling of the thermally accessible configurational space challenging.

Brownian dynamics (BD) and Langevin dynamics (LD) methods, which replace the solvent molecules of molecular dynamics by a more computationally affordable stochastic heat bath, offer a direct route to obtaining the configurational and dynamical properties of biopolymers.[1,2] The simulations are, however, limited by the magnitude of the integration time step and generally lead to regional sampling. Furthermore, the computation of the hydrodynamic diffusion tensor in BD can be very demanding, limiting the overall sampling. The BD/LD methods are, hence, more suited to simulating biopolymers with weak intramolecular nonbonded interactions, especially where hydrodynamic interactions are important.

Monte Carlo (MC) methods, on the other hand, offer more flexibility with the type and size of jumps attempted in the configurational phase space, though they are clearly recognized as having limitations as the number of variables increases. Several MC methods, many tailored for polymer simulations, have been developed to sample the complex energy landscape of biopolymers. The simplest involves local perturbations to monomer positions or bond/torsion angles with a straightforward Metropolis acceptance criteria

$$P_{\mathrm{acc}} = \min[1, \exp(-\Delta U/k_BT)], \tag{1}$$

where ΔU is the energy difference between the proposed and original configurations. Examples of such local moves include translations, rotations, crank-shaft rotations, and pivot rotations (see Ref. 3 for a comprehensive review). Such local moves are typically combined with larger perturbations to

[a)] Author to whom correspondence should be addressed. Electronic mail: schlick@nyu.edu sample space more globally, such as regrowing the entire or large portions of the biopolymer. The configurational-bias MC (CBMC) method[4] lies at the heart of these global MC moves.

It is well-known that the acceptance probability of regrowing a self-avoiding lattice chain via a MC procedure where each new segment is placed randomly at one of the neighboring sites of the previously inserted segment (whether occupied or not) decreases exponentially with the length of the polymer due to an increased tendency for segment/segment overlaps.[5] The original CBMC scheme[6,7] was developed to overcome this limitation by inserting new segments at one of the unoccupied neighbor positions either with equal probabilities $[P(j)=1/n$, where n is the number of trial sites and j is the selected site$]$ or with probabilities proportional to the trial position's Boltzmann weights $[P(j)=\exp(-\beta U_j)/\Sigma_{k=1}^{n}\exp(-\beta U_k)$, where U_j is the selected position's external energy$]$. This bias toward successful chain insertion is then corrected in the acceptance criterion through the computation of the Rosenbluth factor W, which equals the product of the sum of the Boltzmann weights of trial positions for each segmental insertion: $W=\Pi_{i=1}^{N}\Sigma_{k=1}^{n}\exp(-\beta U_k)$, where N is the number chain segments. The new acceptance criteria is then given by

$$P_{\text{acc}}=\min\left[1, \frac{W_{\text{new}}}{W_{\text{old}}}\right], \tag{2}$$

where W_{new} and W_{old} are the Rosenbluth factors of generating the new polymer configuration and retracing the old configuration using the Rosenbluth scheme.

The CBMC scheme was generalized to continuously deformable polymers[8,9] with weak and strong intramolecular interactions (bond stretching, bending, and torsion) by modifying the trial segment generation procedure—to sample trial segments from a Boltzmann distribution corresponding to the bonded interaction energies. Other variants of the CBMC method include the end-bridging algorithm,[10,11] which regrows interior portions of the polymer while keeping end segments fixed; the recoil growth algorithm,[12] which prevents polymers from reaching "dead-alleys" by looking several monomers ahead before attempting a move; and the pruned-enriched Rosenbluth algorithm, which selectively enriches polymer conformations with high Rosenbluth weights.[13]

We propose a simple modification of the traditional CBMC scheme, which we call the end-transfer CBMC method, that leads to dramatic improvement in its sampling efficiency. Rather than regrowing deleted portion of a polymer from the same "cut" end, as in the traditional CBMC, we regrow the deleted portions of the polymer at the opposite end in the same spirit as the reptation algorithm.[3]

After describing this CBMC modification in detail and showing that it conserves microscopic reversibility, we apply it to sampling oligonucleosomes modeled using our recent coarse-grained flexible-tail model[14] and compare its performance to other moves (local, pivot rotations, and traditional CBMC). We characterize sampling efficiency using parameters relevant to the different degrees of freedom of the biopolymer. We find that, under all conditions tested here (medium to high salt), our method improves sampling by several orders of magnitude compared to the traditional CBMC scheme. Our approach also leads to better sampling than other moves under high salt conditions, but falls short of the pivot rotation method at low to medium salt conditions. The difficulty at low salt can be explained by reduced electrostatic salt-screening effects which lead to large electrostatic energies/barriers that dramatically reduce the method's acceptance ratios. Thus, though the sampling of complex but moderately charged/neutral biopolymers may be substantially improved through the end-transfer CBMC scheme, adequate sampling of highly charged systems such as chromatin remains a challenge, especially when salt-screening effects are small.

II. METHOD DEVELOPMENT

A. Concept

The main concept of the end-transfer CBMC scheme (deleting portions of the biopolymer from one end and regrowing them at the opposite end) is sketched in Fig. 1. Without loss of generality, it suffices to focus on the simulation of single biopolymers where only intramolecular interactions are considered; extensions to biopolymers interacting with the surroundings is straightforward. We denote the polymer's two ends by "head" and "tail," arbitrarily. Each repeating unit of the biopolymer (depicted as circles in Fig. 1) could represent a nucleosome plus the associated linker DNA in the case of chromatin, four actin monomeric subunits in the case of an F-actin filament, etc. In general, the repeating unit may be asymmetric (e.g., in chromatin) so the direction of regrowth depends upon which of the two end motifs is selected. The size of the cut portion must be an integer multiple of the size of the repeating motif to preserve the polymer's integrity in the case of asymmetric biopolymers. To preserve microscopic reversibility, both "head-to-tail" ($H\rightarrow T$) and "tail-to-head" ($T\rightarrow H$) must be employed and applied with equal probability of 1/2.

If the above "cut and regrow" scheme is performed randomly (without any biases in the transition matrix) and in a stepwise manner beginning at the polymer termini, the acceptance probability is given by the standard Metropolis criterion

$$P_{\text{acc}}=\min\left[1, \exp\left(-\frac{U_{\text{grow}}-U_{\text{cut}}}{k_BT}\right)\right]. \tag{3}$$

Here, U_{grow} and U_{cut}, respectively, represent the interaction energies of the regrown motif at its new and original locations/configurations with the rest of the polymer. For example, in the first end-transfer move in Fig. 1, U_{cut} represents the energy of interaction between motif 7 at the tail end with motifs 1–6 plus its internal energy, and U_{grow} represents the interaction energy of the regrown motif 7 at the head end with motifs 1–6 plus its internal energy. Unfortunately, such a random regrowth of polymer ends usually results in very low acceptance rates.

As in CBMC, the acceptance probability of the end-transfer moves may be considerably improved by employing the Rosenbluth scheme for regrowing biopolymer motifs.

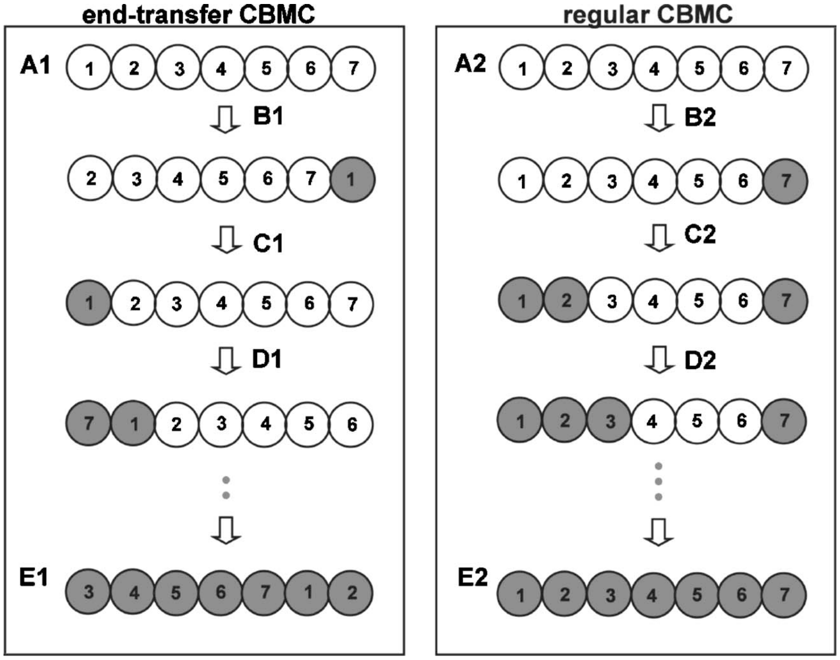

FIG. 1. Comparison of end-transfer CBMC (left box) and regular CBMC (right box) schemes for a biopolymer comprising of seven repeating motifs depicted as circles. Empty and shaded circles represent unsampled and sampled (regrown) motifs, respectively. Biopolymers A1 and A2 represent the starting conformations, and E1 and E2 represent final well-sampled conformations. The following sequence of moves are illustrated for the end-transfer CBMC scheme: transfer/regrowth of segments 1 from head to tail (B1), transfer/regrowth of segment 1 from tail to head (C1), and transfer/regrowth of segment 7 from tail to head (D1). The following sequence of moves are illustrated for the standard CBMC scheme: regrowth of segment 7 (B2), regrowth of segments 1 and 2 (C2), and regrowth of segments 1–3 (D2).

Consider a biopolymer as N repeats of each building block, itself of n_R segments. The total $n_R N$ segments interact within the biopolymer through a bonded force field, U_{bond}, comprising of stretching, bending and torsion terms, and a nonbonded force field, U_{nonb}, comprising long-ranged interactions such as van derWaals and Coulomb.

Consider first the implementation of the $T \rightarrow H$ end-transfer move. A pseudocode for this move is provided in the Appendix. According to the Rosenbluth scheme, the n_R segments of the motif are inserted one after the other at the head end in the following order: $i=n_R, \ldots, 1$. First, k trial positions of segment n_R are generated by sampling from the Boltzmann distribution corresponding to the bonded interaction energy between segments n_R and n_R+1, i.e., the probability of choosing a particular position j for segment i is given by

$$P^{j}_{\text{bond}} = A^{i}_{\text{norm}} \exp(-U^{i,j}_{\text{bond}}/k_B T), \tag{4}$$

where A^{i}_{norm} includes the Jacobian and the distribution's normalization constant.[4] The nonbonded (external) Boltzmann factor of each trial position j is computed and summed to yield the Rosenbluth weight for the insertion of segment i,[15] as given by:

$$w^{i}_{H} = \sum_{j=1}^{k} \exp(-U^{i,j}_{\text{nonb}}/k_B T). \tag{5}$$

Next, one of the trial position s is selected with a probability

$$P^{s}_{\text{nonb}} = \exp(-\beta U^{i,s}_{\text{nonb}}) / \sum_{j=1}^{k} \exp(-\beta U^{i,j}_{\text{nonb}}). \tag{6}$$

The remaining segments $n_R - 1, \ldots, 1$ are also inserted using the above procedure and their Rosenbluth weights are recorded to yield the cumulative Rosenbluth weight

$$W_H = \prod_{i=1}^{n_R} w^{i}_{H}. \tag{7}$$

The cumulative Rosenbluth weight of the original motif configuration at the tail end, W_T, is computed by retracing the old configuration from segments $n_R N - n_R + 1$ onwards. For each segment, $k-1$ trial positions of segment i are generated; the original segment position becomes the kth trial position. The Rosenbluth weight for each segment i is then computed as

$$w^{i}_{T} = \exp(-U^{i,o}_{\text{nonb}}/k_B T) + \sum_{j=1}^{k-1} \exp(-U^{i,j}_{\text{nonb}}/k_B T), \tag{8}$$

where $U^{i,o}_{\text{nonb}}$ represents the external energy of segment i in the old conformation o. The cumulative Rosenbluth for the motif in the old conformation/position is given by

$$W_T = \prod_{i=n_R N - n_R + 1}^{n_R N} w^{i}_{T}. \tag{9}$$

Finally, the regrown H motif is chosen with a probability of

$$P^{T\rightarrow H}_{\text{acc}} = \min\left[1, \frac{W_H}{W_T}\right]. \tag{10}$$

Note that the acceptance probability expression of Eq. (8) is typical of all configurational bias methods. The acceptance probability for the reverse move, namely the $H \rightarrow T$ move, is given by

$$P_{\text{acc}}^{H\rightarrow T} = \min\left[1, \frac{W_T}{W_H}\right]. \tag{11}$$

The transfer and regrowth of polymer segments from one end to the other is what differentiates the end-transfer CBMC scheme from traditional CBMC schemes. Standard CBMC schemes involve cutting the polymer at a random position along its entire length and then regrowing the shorter half of the polymer via a Rosenbluth scheme (right-hand side of Fig. 1). The size of the regrown portion therefore varies uniformly between 1 and $N/2$ segments, where N is the total number of segments in the polymer. Assuming that the acceptance probability of regrowing a single motif is given by a, where $a<1$, the acceptance probability of regrowing portions of polymer containing n motifs decreases exponentially (power law) with n as given by a^n. This means that polymer sampling by the traditional CBMC method is rate limited by the sampling of the innermost regions of the polymer which require the largest cut sizes with the smallest acceptance probabilities. The centermost segment of the polymer is therefore sampled approximately with an acceptance probability of $a^{N/2}$. If we define sampling time τ_1 as the average number of MC steps taken before the regrowth of a single motif is accepted, then $\tau_1 \simeq a^{-1}$, and the average MC steps required to sample the entire length of the polymer, τ_N, is given by

$$\tau_N \simeq \tau_1^{N/2}. \tag{12}$$

Consequently, the sampling time of polymers increases in a power-law fashion with the length of the polymer, making the standard GCMC method require prohibitively long simulations for sampling long polymers chains.

In the end-transfer move, instead, regrowing such large polymer segments is not required to sample the innermost segments; the random transfer of polymer motifs from one end to the other routinely exposes inner regions of the biopolymer for sampling. Thus, the size of the regrown portions remains short enough (equal to one motif here) to allow regrowth with high acceptance probabilities. The polymer chain is exhaustively sampled when each motif has been transferred from one end to the other at least once. It can be shown that for a polymer chain composed of N repeating motifs, the average MC steps required to sample the entire polymer is given by

$$\tau_N \simeq \tau_1 \frac{N(N+1)}{2}, \tag{13}$$

where τ_1 is defined above [see derivation of Eq. (13) in the Appendix]. Note that the average time needed to sample a polymer chain now scales quadratically with the chain length, as opposed to a power-law in the traditional CBMC.

The end-transfer method resembles the reptation algorithm commonly employed in polymer simulations. In a reptation move, a move simply involves deleting a segment from one end and randomly placing it at one of the neighboring lattice positions of the terminal segment at the other end; a configurational bias method is not required, and the acceptance criteria is simply given by Eq. (3). In complex biopolymers, the repeating unit generally consists of many segments (n_R); the configuration bias MC approach is preferable to implementing Eq. (3) because of its efficiency in generating an energetically favorable configuration.

The end-transfer scheme is not limited to transfer/regrowth of single repeating motifs. Larger portions comprised of more than one motif may be regrown, though the acceptance probability of regrowing larger portion grows exponentially with size. Thus, even though the entire polymer will be sampled with a fewer number of accepted moves, the number of MC steps taken to accept a single end-transfer move will increase drastically with the size of the regrown portion. Specifically, if x motifs are transferred/regrown at each end-transfer step, the sampling time τ_N will roughly vary as

$$\tau_N \simeq \tau_1^x \frac{(N/x)[(N/x)+1]}{2}. \tag{14}$$

If each motif is difficult to sample, i.e., τ_1 is large, the end-transfer moves actually become less efficient as the size of the transferred portions is increased.

B. Proof of microscopic reversibility

Microscopic reversibility for the acceptance criteria in Eqs. (10) and (11) can be proven by considering the detailed balance between two states T and H. The two states are connected to one another through an end-transfer move $H \rightarrow T$ and its reverse $T \rightarrow H$ as

$$\rho(T)\alpha(T\rightarrow H)\text{acc}(T\rightarrow H) = \rho(H)\alpha(H\rightarrow T)\text{acc}(H\rightarrow T), \tag{15}$$

where $\rho(T)$ and $\rho(H)$ are the thermodynamic probabilities of occurrence of states T and H, respectively; $\alpha(T\rightarrow H)$ and $\alpha(H\rightarrow T)$ represent the probabilities of generating the trial configuration H when already in state T, and vice versa, respectively; and $\text{acc}(T\rightarrow H)$ and $\text{acc}(H\rightarrow T)$ represent the probability of accepting the generated trial moves. We now assume that $\rho(T)$ and $\rho(H)$ are proportional to their Boltzmann factors, i.e.,

$$\rho(T) = \prod_{i=1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{T,i,o}/k_B T)\exp(-U_{\text{nonb}}^{T,i,o}/k_B T), \tag{16}$$

$$\rho(H) = \prod_{i=1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{H,i,s}/k_B T)\exp(-U_{\text{nonb}}^{H,i,s}/k_B T). \tag{17}$$

In the above expressions, the overall Boltzmann weights have been separated into the Boltzmann weights of individual segments i, and by bonded and nonbonded terms. Recall that the prefactor A_{norm}^i in Eqs. (16) and (17) contains terms related to the normalization and Jacobian of the bonded interaction of segment i (without double counting). Note that the configuration/position of motifs in the range 1 to $n_R N - n_R + 1$ in the initial (T) configuration remains unchanged during the $T \to H$ transition, i.e.,

$$U_{\text{bond}}^{H,i+n_R,s} = U_{\text{bonb,}}^{T,i,o} \quad \text{when} \quad i = \cdots, n_R N - n_R + 1,$$
$$U_{\text{nond}}^{H,i+n_R,s} = U_{\text{nonb,}}^{T,i,o} \quad \text{when} \quad i = \cdots, n_R N - n_R + 1. \tag{18}$$

The probability of attempting a $T \to H$ transition is then given by

$$\alpha(T \to H) = \frac{1}{2} \prod_{i=1}^{n_R} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{H,i,s}/k_B T) \times \frac{\exp(-U_{\text{nonb}}^{H,i,s}/k_B T)}{\Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{H,i,j}/k_B T)}, \tag{19}$$

where the prefactor 1/2 accounts for the probability of choosing one of the two end motifs for transfer/regrowth; the second term accounts for sampling from the probability distribution corresponding to the bonded force field, and the third term represents the Boltzmann-factor biased probability of picking configuration s from the k trial positions. Similarly, the probability of attempting the reverse move ($T \to H$) is given by

$$\alpha(H \to T) = \frac{1}{2} \prod_{i=n_R N - n_R + 1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{T,i,o}/k_B T) \times \frac{\exp(-U_{\text{nonb}}^{T,i,o}/k_B T)}{\Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{T,i,j}/k_B T)}. \tag{20}$$

Substituting Eqs. (16) and (20) into Eq. (15), we obtain

$$\frac{\text{acc}(T \to H)}{\text{acc}(H \to T)} = \left[\frac{\Pi_{i=1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{H,i,s}/k_B T) \exp(-U_{\text{nonb}}^{H,i,s}/k_B T)}{\Pi_{i=1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{T,i,o}/k_B T) \exp(-U_{\text{nonb}}^{T,i,o}/k_B T)} \right] \times \left[\frac{\Pi_{i=n_R N - n_R + 1}^{n_R N} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{T,i,o}/k_B T) \exp(-U_{\text{nonb}}^{T,i,o}/k_B T) / \Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{T,i,j}/k_B T)}{\Pi_{i=1}^{n_R} A_{\text{norm}}^i \exp(-U_{\text{bond}}^{H,i,s}/k_B T) \exp(-U_{\text{nonb}}^{H,i,s}/k_B T) / \Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{H,i,j}/k_B T)} \right]. \tag{21}$$

Upon canceling like terms and substituting Eqs. (5), (7)–(9), and (18), the above expression simplifies to

$$\frac{\text{acc}(T \to H)}{\text{acc}(T \to H)} = \frac{\Pi_{i=1}^{n_R} \Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{H,i,j}/k_B T)}{\Pi_{i=n_R N - n_R + 1}^{n_R N} \Sigma_{j=1}^k \exp(-U_{\text{nonb}}^{T,i,j}/k_B T)} = \frac{W_H}{W_T}. \tag{22}$$

Finally, using the Metropolis criterion, we obtain

$$\text{acc}(T \to H) = \begin{cases} W_H/W_T, & \text{if} \quad W_H < W_T \\ 1, & \text{if} \quad W_H \geq W_T \end{cases}, \tag{23}$$

which is equivalent to our acceptance criterion in Eq. (10). The acceptance probability of the reverse move [acc($H \to T$)] presented in Eq. (11) can be proven similarly (not shown).

III. APPLICATION TO CHROMATIN

We test the efficiency of the proposed end-transfer CBMC scheme in sampling short chromatin segments. Chromatin, the system that motivated our algorithm, is a suitable model system for two reasons. First, chromatin is composed of negatively charged double-stranded DNA and mostly positively charged histone proteins. Thus, intramolecular nonbonded interactions are strong, making adequate sampling a challenge. Second, the complicated architecture of chromatin, even in the coarse-grained formulation that we use,[14,16] presents hierarchical interactions that are difficult to balance. We model short segments of chromatin (oligonucleosomes) using the recently developed mesoscopic "flexible-tail" model, sketched in Fig. 2. The flexible tail model is briefly described next; Ref. 14 contains full details.

A. Flexible tail model of an oligonucleosome

An oligonucleosome consists of a chain of repeating units, where each repeating motif consists of a nucleosome core, linker DNA, and histone tails. The nucleosome core is treated as a rigid body whose surface is uniformly spanned by 300 discrete charges; the charges are optimized to reproduce the electric field around the atomistic nucleosome core.[17] Each charge is also assigned an effective excluded volume using a Lennard–Jones potential to prevent overlap of nucleosome core with the other chromatin components. Each linker DNA is represented as a chain of $n_{lb}=6$ charged beads (for 60 base pairs) with appropriate excluded volume, stretching, bending and twisting terms in its force field.[18–20] The charges have been optimized using the procedure of Stigter[21] to reproduce the far-field electrostatic potential of double-stranded DNA. The oligonucleosomal arrays are formed as the collection of these nucleosomes, connected by linker DNAs. The linker DNAs enclose an angle of 90° about the center of the nucleosome core, and are separated by a distance of 3.6 Å normal to the plane of the nucleosome core (see Fig. 2), following the crystal structure.

Each nucleosome core also serves as the origin for $n_h = 10$ histone tails: two copies each of the N-termini of H2A, H2B, H3, and H4 histones, and C-terminus of each H2A

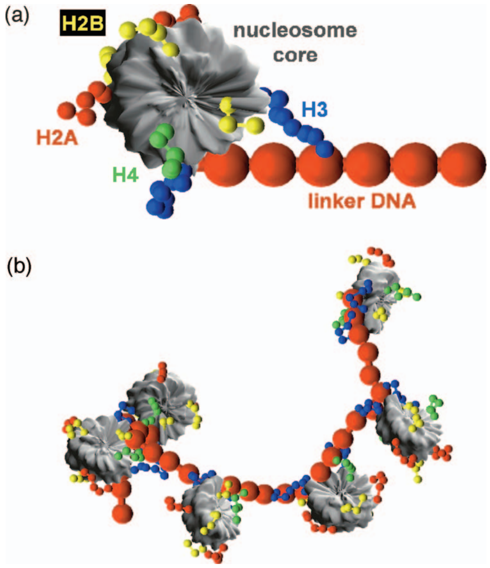

FIG. 2. (Color) Coarse-grained oligonucleosome. (a) Repeating motif of an oligonucleosome consisting of a rigid nucleosome core, six linker DNA beads and ten histone tails (two hidden from view), and (b) a typical six-unit oligonucleosome in a moderately unfolded state. The nucleosome itself is composed of 300 uniformly distributed spherical charges.

C-terminal domain. Each histone tail is treated as a chain of coarse-grained beads with a force field comprising of stretching, bending, charge/charge interaction and excluded volume terms.[14] Briefly, the $n_{hb}=50$ histone tail beads per nucleosome core correspond to 4 for H2A N-termini ($n_{b1}=n_{b2}=4$), 3 for H2A C-termini ($n_{b3}=n_{b4}=3$), 5 for H2B ($n_{b5}=n_{b6}=5$), 8 for H3 ($n_{b7}=n_{b8}=8$), and 5 for H4 ($n_{b9}=n_{b10}=5$). The stretching and bending potentials for interbead lengths and bond angles (defined by three consecutive beads) are represented by harmonic potentials with parameters that reproduce configurational properties of the atomistic histone tails. A Lennard–Jones potential provides excluded volume to each protein bead. Appropriate charges are also assigned to each histone tail bead to mimic its electrostatics. Each histone chain is attached to the nucleosome core using a stiff spring. The nucleosome core, linker DNA, and histone tail charges interact electrostatically with each other through an effective salt-dependent Debye–Hückel potential.

Using N basic building blocks, the total number of major (nucleosome core+linker DNA) components in the entire oligonucleosome is $(n_{lb}+1)N$, and the total number of histone tail beads is $n_{hb}N$. Hence, an N-unit oligonucleosome contains $(n_{lb}+n_{hb}+1)N$ interacting "particles;" a 12-unit oligonucleosome has 684 particles. The array head corresponds to the first unit and the array tail to the last (N) unit (see Fig. 2).

B. Implementation of end-transfer CBMC

To implement the end-transfer scheme to oligonucleosomes, we consider the $T \rightarrow H$ move. The first step involves computing the Rosenbuth weight W_T of the existing tail motif in the oligonucleosome. This is done by retracing the position and orientation of the nucleosome core, the six linker DNA beads, and the histone tails in that order, and computing the segmental Rosenbluth weights at each step.

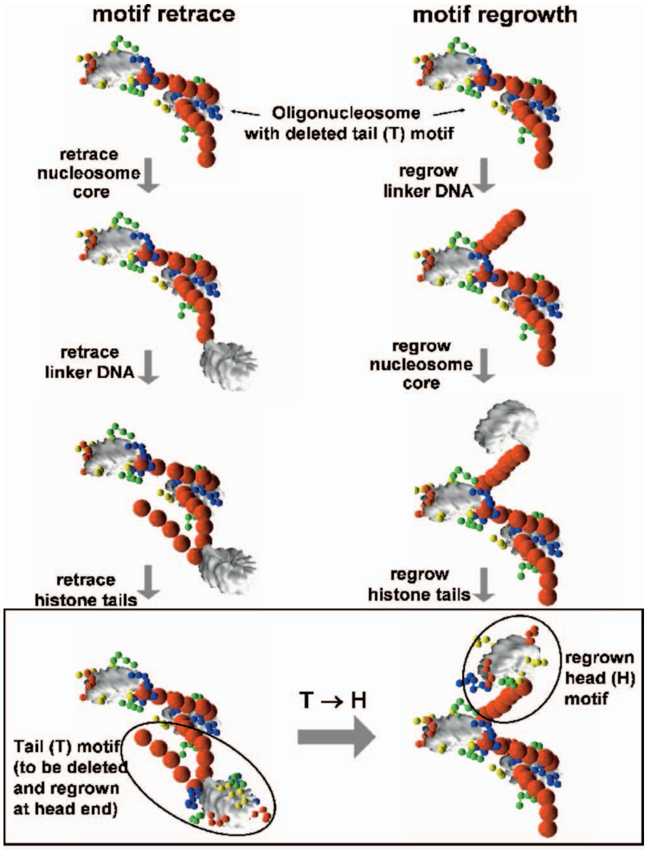

FIG. 3. (Color) Implementation of an $T \rightarrow H$ end-transfer CBMC move in a trinucleosome illustrating the retrace and regrowth steps. Boxed region encloses the original and proposed conformation of the trinucleosome.

Figure 3 illustrates this retracing procedure for a trinucleosome. The overall Rosenbluth weight of the retraced end motif is given by

$$W_T = (w_{C,T})\left(\prod_{i=1}^{n_{lb}} w^{i}_{L,T}\right)\left(\prod_{i=1}^{n_h}\prod_{j=1}^{n_{bi}} w^{i,j}_{T,T}\right), \tag{24}$$

where the first term represents the Rosenbluth weight of inserting the last nucleosome core at its original position; the second term accounts for the Rosenbluth weight for the insertion of the last six linker beads; and the last term accounts for the insertion of n_{hb} beads on the last nucleosome core. The histone tail bead positions are retraced starting from the bead attached to the nucleosome core.

The next step involves regrowth of a complete motif at the head end of the oligonucleosome. Again, the sequence of insertions is as follows. The six linker DNA beads are inserted first, beginning with the linker bead attached to the head nucleosome. Next, the nucleosome core is regrown from the last-inserted linker DNA bead. Finally, the histone tails are regrown from the inserted nucleosome core, each tail regrowth beginning with the bead attached to the nucleosome core. The overall Rosenbluth weight of the regrown motif is then given by

$$W_H = (w_{C,H})\left(\prod_{i=1}^{n_{lb}} w^i_{L,H}\right)\left(\prod_{i=1}^{n_h}\prod_{j=1}^{n_{bi}} w^{i,j}_{H,T}\right), \tag{25}$$

where the first term again is the Rosenbluth weight of inserting the first nucleosome core, the second term is the Rosenbluth weight for inserting the six linker beads, and the last term is the Rosenbluth weight for the insertion of n_{bi} beads of n_{ht} histone tails at nucleosome core 1. The overall ($T \rightarrow H$) move is then accepted with the probability given by Eq. (10). The reverse move ($H \rightarrow T$) is implemented similarly with the sequence of retracing and regrowth reversed.

C. Simulation details

We use the end-transfer method in combination with three local Monte Carlo moves to enhance efficiency: translation and rotation of nucleosome cores and linker DNA beads and histone tail regrowths.

Translation is performed by choosing a randomly oriented axis passing through a randomly selected linker DNA bead/nucleosome core, and shifting that component (and its associated histone tails if the chosen component is a nucleosome core) along the axis by a distance sampled from a uniform distribution in the range (0, 0.6 nm). Rotation involves rotating the selected nucleosome core/linker DNA bead about one of its axis by an angle uniformly sampled from the range (0, 36°). Note that while the nucleosome core may be rotated about either of its three axes, the linker DNA bead may only be rotated about its interbead axis (see Ref. 14). The translation and rotation moves are accepted/rejected based on the standard Metropolis criterion.

Tail regrowth involves selecting a histone tail randomly and regrowing it in a different conformation using the standard CBMC algorithm by generating N_t=4 trial positions.

The end-transfer moves are performed by transferring portions of one repeating motif using N_t=10 trials for linker DNA and nucleosome core regrowth and N_t=4 trials for tail regrowths. The four moves—translation, rotation, tail regrowth, and end-transfer—are performed with frequencies 0.1:0.1:0.6:0.2; the tails are sampled more frequently to account for their larger numbers (recall that there are 50 tails beads per nucleosomes). We call simulations with the above set of moves "L+E" (for local+end transfer).

For comparison, we perform three additional simulations with different sets of Monte Carlo moves. The first ("L") involves only local moves (translational, rotational, and tail regrowth) in frequencies 0.2:0.2:0.6, respectively. The second ("L+P") employs pivot rotations in addition to the three local moves; pivoting involves randomly choosing one linker DNA bead or nucleosome core, selecting a random axis passing through the chosen component, and then rotating the shorter part of the oligonucleosome about this axis by an angle chosen from a uniform distribution within (0, 20°). The relative frequencies of attempting the translation/rotation/tail-regrowth/pivot moves is fixed at 0.1:0.1:0.6:0.2, respectively. The third type of simulations ("L+C") employs the standard CBMC algorithm to regrow large portions of the oligonucleosome in addition to the three local moves. The size of the regrowth portion is chosen uniformly between 1 and $N/2$ (if N even) or $N/2+1$ (if N odd) motifs, and the oligonucleosome end to be regrown is selected with a probability of 1/2. The mechanics of the regrowth procedure in the CBMC moves are similar to that of the end-transfer moves, with the exception that the size of the regrown portions in the CBMC method is not restricted to single motifs. The relative attempt frequencies of the translation/rotation/tail regrowth/pivot/oligonucleosome regrowths are given by 0.1:0.1:0.6:0.2, respectively.

All simulations are performed in the canonical ensemble at T=293 K and at different monovalent salt concentrations to assess the impact of the strength of electrostatic interactions on sampling efficiency. We choose two salt concentrations: 0.2 M (medium salt; close to physiological salt concentration), where oligonucleosomes are moderately folded; and 0.5 M (high salt), where electrostatic interactions are mostly screened and oligonucleosomes exhibit slightly more extended configurations as compared to those at 0.2 M.

Oligonucleosome sizes range from N=3 to N=12 nucleosomes. The length of the simulations varies between 10 and 50 million MC moves. The initial configuration of the oligonucleosome corresponds to the square solenoid nucleosomal pattern of Ref. 14. We found that the initial state does not influence the results. All results reflect averages from a set of four runs started from a different random number generator seed.

D. Sampling efficiency measures

We quantify the degree of sampling using four parameters that characterize four separate degrees of freedom of the oligonucleosomes that should be properly sampled for thermodynamic convergence.

To examine how well an oligonucleosome samples the surrounding volume, i.e., its translational degree of freedom, we compute the mean square deviation of the center of mass position of the oligonucleosomes versus the number of MC steps t separating the two oligonucleosomes, as given by

$$\xi(t) = \langle|\mathbf{r}_{\mathrm{CM}}(t+t_0) - \mathbf{r}_{\mathrm{CM}}(t_0)|^2\rangle, \tag{26}$$

where $\mathbf{r}_{\mathrm{CM}}(t_0)$ and $\mathbf{r}_{\mathrm{CM}}(t+t_0)$ are the center of mass coordinates of the oligonucleosome computed from its nucleosome core positions at times t_0 and $t+t_0$, respectively, and $\langle\cdot\rangle$ is an ensemble average over different origins (t_0) and simulation runs with different random number generator seeds. The rate of translational sampling, R, may be quantified by taking the slope of $\xi(t)$ versus MC steps t as $t \rightarrow \infty$, as is normally done for computing diffusion coefficients using the Einstein relation

$$R \sim \lim_{t\rightarrow\infty} \frac{\xi(t)}{t}. \tag{27}$$

To quantify rotational sampling of oligonucleosomes, we compute an autocorrelation of the end-to-end unit vector of the oligonucleosome, as given by

$$\rho(t) = \langle \mathbf{e}(t_0 + t) \cdot \mathbf{e}(t_0) \rangle, \tag{28}$$

where $\mathbf{e}(t_0)$ is the end-to-end unit vector of the oligonucleosome after t_0 steps (pointing from nucleosome 1 to nucleosome N), as given by

$$\mathbf{e}(t_0) = \frac{\mathbf{r}_N(t_0) - \mathbf{r}_1(t_0)}{|\mathbf{r}_N(t_0) - \mathbf{r}_1(t_0))}, \tag{29}$$

where $\mathbf{r}_N(t_0)$ and $\mathbf{r}_1(t_0)$ are the positional coordinates of the two end nucleosomes. The autocorrelation can be fitted to an exponential curve $\rho(t) = \exp(-t/\tau_\rho)$ to obtain the rotational relaxation time τ_ρ in terms of number of MC steps.

To characterize internal sampling of oligonucleosomes, we compute "translation-rotation corrected" mean square deviation in the positions of nucleosome cores of oligonucleosomes separated by t MC steps during the course of the simulation, as given by

$$\lambda(t) = \left\langle \min\left(\sum_{i=1}^{N} |\mathbf{r}_i^{\dagger}(t + t_0) - \mathbf{r}_i(t_0)|^2 \right) \right\rangle, \tag{30}$$

where "min" implies that the two oligonucleosome have been superimposed onto each other to minimize their relative deviation from each other before computing the mean square deviation; $\mathbf{r}_i(t_0)$ is the coordinate vector of nucleosome i of the oligonucleosome after t MC steps and $\mathbf{r}_i^{\dagger}(t+t_0)$ is the coordinate vector of nucleosome i after $t+t_0$ steps after the superposition. The superposition involves determining the rotation matrix and translation vector that gives the best fit superimposition of the two sets of molecules based on the program PDBSUP created by Rupp and Parkin.[22] Such a superposition therefore captures only the internal structural arrangements within the oligonucleosome without corruption from purely translational and rotational modes of the oligonucleosome. The characteristic timescale associated with the sampling of the internal structure of the oligonucleosome (relaxation time τ_λ) can then be obtained by fitting the $\lambda(t)$ versus t plot to a stretched exponential function of the form

$$\lambda(t) = \alpha\{1 - \exp[-(t/\tau_\lambda)]\}, \tag{31}$$

where α and β are best fit parameters.

Note that $\lambda(t)$ does not differentiate between the sampling of the inner portions of the biopolymer from the rest. Because the innermost portions of the biopolymer are sampled infrequently in traditional CBMC simulations, we define $\chi(t)$ to characterize the rate of sampling of the innermost motif (i.e., the mth nucleosome core, where $m = \text{int}\{c_n/2\}+1$, and the ensuing n_{lb} linker DNA beads) by computing deviations in its internal structure with MC steps using the same superposition formalism as above

$$\chi(t) = \left\langle \min\left[\sum_{i=1}^{n_{lb}+1} |\mathbf{r}_i^{\dagger}(t + t_0) - \mathbf{r}_i(t_0)| \right] \right\rangle, \tag{32}$$

where $\mathbf{r}_i(t_0)$ is the position of the ith nucleosome core/linker DNA bead at time t_0, and $\mathbf{r}_i^{\dagger}(t+t_0)$ is its position at time $t+t_0$ after superposition. As above, we derive the characteristic sampling time τ_χ by fitting the $\chi(t)$ to Eq. (31).

IV. RESULTS AND DISCUSSION

We compare sampling efficiency in terms of the measures above for simulations employing the end-transfer method ($L+E$) versus standard CBMC method ($L+C$), pivot rotations ($L+P$), and local moves only (L) at the two salt conditions (0.2 and 0.5 M) (Fig. 4).

Figure 4(a) analyzes translational sampling for the different sets of simulations. Mean square deviations in the oligonucleosome center of masses $\xi(t)$ increase linearly with t as is typical for a random walk (a representative ξ-t plot is shown). The rate of sampling is computed from the slopes of $\xi(t)$ versus t and plotted on a logarithmic scale. As expected, sampling becomes more difficult as the length of the polymers increases and/or the salt concentration is lowered.

Significantly, the efficiency of the end-transfer method ($L+E$) compared to the other methods in sampling translational degrees of freedom of the oligonucleosomes is excellent: the $L+E$ simulations enhance sampling by a factor of 10 (medium salt) and 100 (high salt) over other methods. This advantage stems from the reptationlike nature of the move where the biopolymer advances/recedes an entire repeating motif at a time. Indeed, reptation moves commonly employed in lattice polymer simulations are well known for providing efficient sampling.[3]

The $L+P$ approach provides the second best sampling efficiency, followed by $L+C$, and L simulations. Surprisingly, the conventional CBMC method ($L+C$) provides poor translational sampling, especially for oligonucleosomes at low salt, due to its extremely low acceptance rates. The conservative local moves in L simulations provide the slowest translational sampling as expected.

The rotational sampling efficiency in Fig. 4(b) shows that the autocorrelation of the end-to-end separation vector $\rho(t)$ decays to zero with a characteristic rotational relaxation time, τ_ρ (see representative plot). A small τ_ρ is characteristic of fast rotational sampling of the oligonucleosomes and vice versa. Such autocorrelations are typically fitted to an exponential, but we found that stretched exponentials with exponents in the range 0.4–0.8 fit the data better, possibly due to coupling of multiple rotational modes in the autocorrelation. The end-transfer moves provide the best rotational sampling at high salt and second (to pivot moves) at medium salt. As for translational sampling, local moves and standard CBMC regrowth of oligonucleosomes are worst. In fact, the rotational relaxation time corresponding to local moves for 12-unit oligonucleosomes exceeds the simulation length, which explains the missing data.

Figure 4(c) analyzes internal degree of freedom sampling via mean square deviations $\lambda(t)$ for a trinucleosome at high salt separated by a sampling time t (best fit superposition). We note that $\lambda(t)$ increases monotonically until it plateaus at a characteristic internal relaxation time, τ_λ. The computed τ_λ values increase dramatically with the length of the oligonucleosomes. At high salt, the end-transfer moves provides the best internal configurational sampling; at medium salt, pivoting improves sampling by an order of magnitude than end-transfer. The CBMC move provides the next best

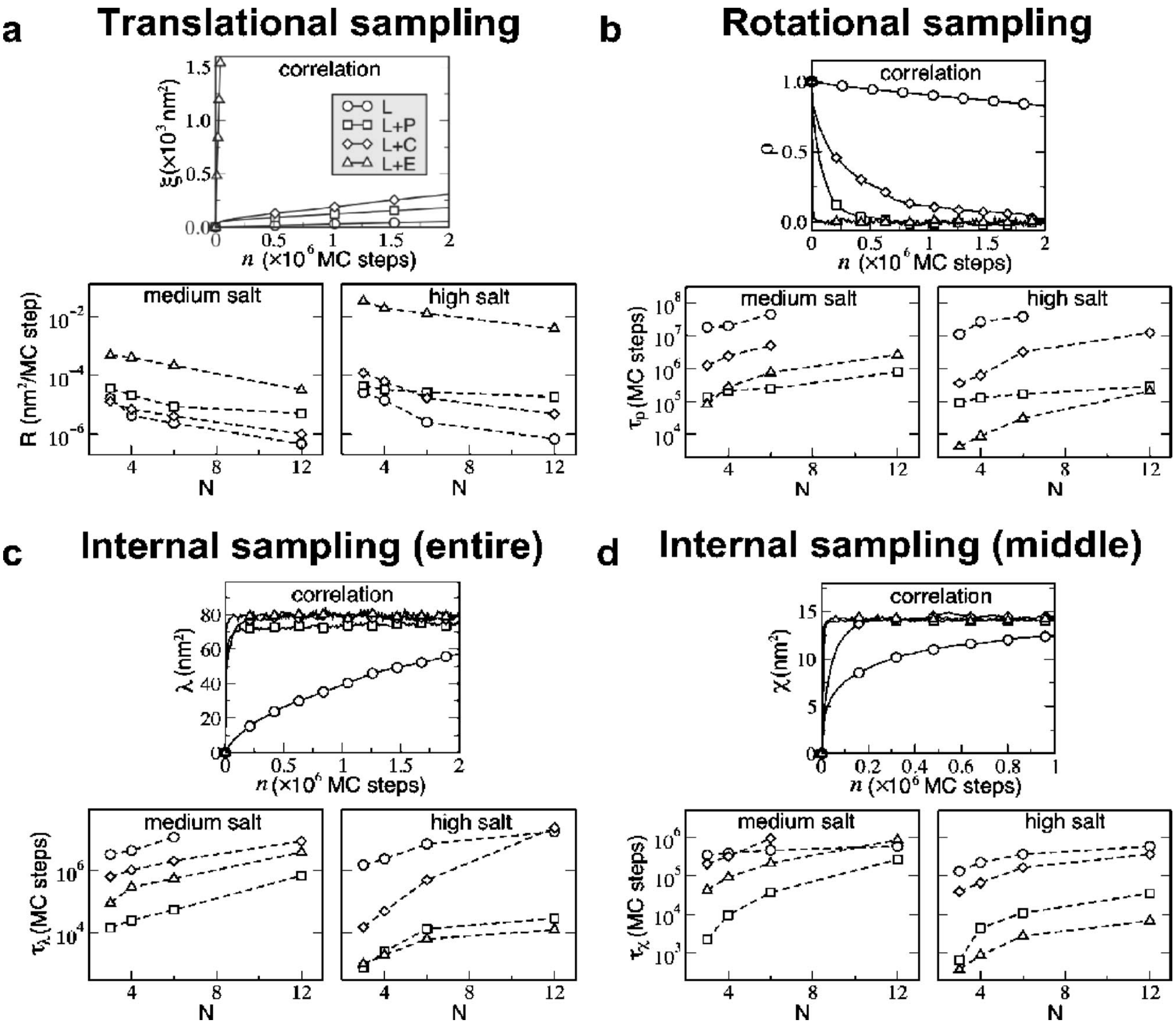

FIG. 4. Sampling efficiency of various MC method in terms of four sampling measures: (a) translational, (b) rotational, (c) internal (entire oligonucleosome), and (d) internal (middle motif). Each box contains a representative correlation function (for a trinucleosome at high salt) used for computing the relevant sampling measure (top), and the relevant measures for different-sized oligonucleosomes at medium (left) and low salt (right); the dashed lines are guides to the eye. Results from simulations employing local (*l*), pivot ($L+P$), standard CBMC ($L+C$), and end-transfer CBMC ($L+E$) moves are represented by black, red, blue, and green symbols/lines, respectively.

internal sampling and the local moves provide the worst sampling for both neutral and charged systems.

Figure 4(d) compares results in terms of sampling only the innermost motif of the oligonucleosome. The evolution of the mean square deviation of the center unit also follows a stretched exponential behavior. The corresponding relaxation times τ_χ show that the end-transfer moves again provide the best internal sampling at high salt, while pivoting appears the best for oligonucleosomes at medium salt. Significantly, the standard CBMC method performs very poorly—almost like local moves. The rarity of accepting regrowth of innermost motifs constitutes the most severe drawback of the traditional CBMC method.

Thus, the end-transfer CBMC approach provides excellent sampling at high salt but its performance degrades below that of pivot moves at medium salt. To understand the origin of this salt-dependent efficiency, we compute in Fig. 5 the mean acceptance probabilities of the pivot, standard CBMC and end-transfer CBMC moves with respect to the chain length at medium and high salt. At high salt, the acceptance probability of the end-transfer move is roughly two orders of magnitude smaller that that of pivot moves. As the salt concentration is lowered, this disparity between the two acceptance probabilities widens. In fact, at medium salt, this gap is already larger than three orders of magnitude. At even lower salt concentrations (0.01 M), the acceptance of an end-transfer move becomes smaller than one in a million. This drastic reduction in the acceptance probability of the end-transfer move compared to pivot moves explains why the end-transfer method is less efficient than the pivot moves at medium salt. The same reasoning also explains why the standard CBMC—with even lower acceptance probabilities than the end-transfer CBMC for long chains—yields poor sampling.

The low probability of acceptance of the end-transfer moves is expected given the number of histone tails than

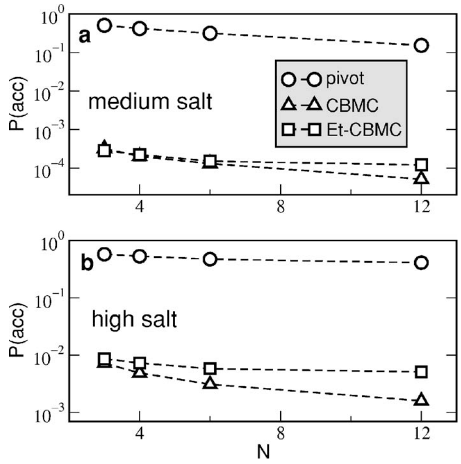

FIG. 5. Acceptance probabilities of the pivot, standard CBMC, and end-transfer CBMC moves for different oligonucleosome sizes at (a) moderate and (b) high salt. Dashed lines are guides to the eye.

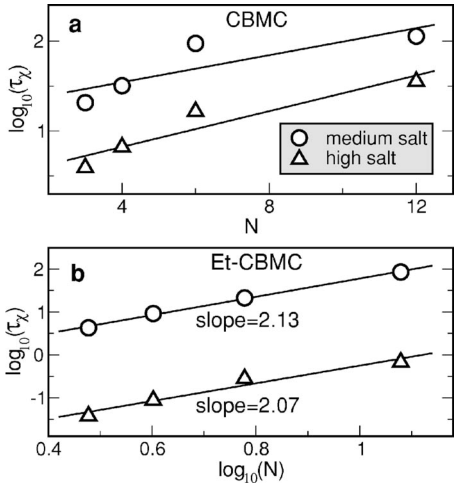

FIG. 6. Characteristic sampling time of the oligonucleosome's innermost motif with (a) regular CBMC, and (b) end-transfer CBMC moves at different oligonucleosome lengths indicating a power-law and quadratic dependence of sampling time with chain length.

need to be regrown at one step within the highly corrugated electrostatic landscape surrounding each nucleosome core and linker DNA. The fact that a single histone tail regrowth is accepted on an average with a probability of about 0.25 clearly points to the difficulty in inserting all ten tails at a time. Surprisingly, despite the low acceptance probabilities of the end-transfer moves, the method still yields remarkably good sampling efficiency.

It is also instructive to examine the scaling of the structural relaxation times of the central motif (τ_χ) obtained from the end-transfer and regular CBMC methods with chain length N and compare them with analytical predictions based on scaling arguments [Eqs. (12) and (13)]. Figure 6(a) plots the standard CBMC relaxation times for both the neutral and fully charged tail oligonucleosomes versus the chain length in a log-linear plot; those corresponding to the end-transfer CBMC method are plotted ina log-log plot in Fig. 6(b). The linear relationship between $\log(\tau_\chi)$ and N observed in standard CBMC confirms that the sampling time of the centermost motif of the oligonucleosome scales in a power-law fashion with the length of the oligonucleosome. The linear relationship with a slope of roughly 2 in the bottom figure confirms the quadratic dependence of sampling time with chain length. Hence, for large N, the end-transfer CBMC approach quickly becomes superior to the regular CBMC in sampling the interior portions of the biopolymer.

The added computational cost in implementing the end-transfer moves compared to pivot moves stems from generating multiple trial positions and energy computations for each segment of the regrown and retraced motif. The CPU factor is about 3–4. We have not attempted to optimize the end-transfer moves, but it may be possible to reduce the CPU cost through careful optimization of parameters involved (e.g., the number of trials N_t) and through reusage of computed Rosenbluth weights until an end-transfer move is rejected. Also, we have used the end-transfer CBMC method only alongside local moves. Efficiency can be further improved, especially for biomolecules with strong intramolecular nonbonded interactions, by combining this approach with global pivot rotations. Such a combination would provide superior translational sampling as well as good rotational and intramolecular sampling.

In sum, the developed end-transfer CBMC method provides excellent sampling of structurally complex biopolymers when nonbonded interactions of an electrostatic origin are relatively weak. The method, however, loses its superiority over pivot moves when electrostatics effects begin to dominate, as in low salt conditions, though it still provides good sampling of the translational degrees of freedom. Still, the method consistently provides orders of magnitude better sampling than the traditional CBMC method. This exceptional translational sampling and good rotational/internal sampling makes it highly suitable for sampling condensed polymers and biopolymers with weak intermolecular and nonbonded intramolecular interactions to study their collective properties (as opposed to single-molecule properties) like phase behavior, thermodynamics and structure.

V. CONCLUSION

Our extension of the well-known configurational bias Monte Carlo methodology for better sampling of phase space in complex biopolymers, motivated by reptation moves, involves randomly transferring a repeating motif in a complex biopolymer from one end to the other and regrowing it using the efficient Rosenbluth scheme. We assessed efficiency on mesoscale simulations of oligonucleosomes compared to traditional CBMC and pivot rotations for four degrees of freedom of the biopolymer (translation, rotation, and intramolecular sampling of the entire chain/middle portion). Our method yields very efficient sampling when charge-charge attractions within the oligonucleosomes are not too strong. When nonbonded charge-charge attractions become strong, however, sampling suffers due to its low acceptance probabilities. Still, the end-transfer method provides several orders of magnitude better sampling than the traditional CBMC method. This is particularly due to the quadratic scaling between the simulation time required to sample the innermost regions of the biopolymer and the chain length, as compared to the drastic exponential scaling observed for the traditional CBMC method. Our extension to the standard CBMC method may thus be further coupled to other techniques like parallel tempering or replica exchange[23–25] (where configurations are switched between ensembles at different temperatures) or stochastic tunneling[26] (where large energy barriers are smoothed out to allow barrier crossings) to better sample the vast rugged energy landscape of biopolymers.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support by NSF Award No. MCB-0316771 and NIH Grant No. R01 GM55164 to T. Schlick. Acknowledgment is also made to the donors of the American Chemical Society Petroleum Research Fund for support (or partial support) of this research (Award No. PRF39225-AC4 to T. Schlick). They also thank Dr. Chia Chang for useful discussions that led to the proof for the scaling of sampling time with chain length in the end-transfer scheme.

APPENDIX: PSEUDOCODE FOR IMPLEMENTING *T*→*H* END-TRANSFER CBMC MOVE

```
! N: total number of repeating units in the biopolymer
! n_R: number of segments in each repeating unit
! n_T: number of trial positions generated for each segment
! k_B: Boltzmann constant
! T: temperature
! R(n_R*N): current coordinates of biopolymer segments
! W_old: Rosenbluth weight of old configuration
! W_new: Rosenbluth weight of new configuration
! RETRACE OLD CONFIGURATION AT TAIL END
W_old=1
Do i=n_R*N-n_R+1, n_R*N
    w=0
    Do t=1,n_T-1
        call genpos (r(t)) ! generate trial position of segment
        ! based on internal constraints (e.g., bond)
        call energy (r(t),U(t)) ! calculate external energy of trial
        w=w+exp(-U(t)/k_B/T)
    Enddo
    call energy(r_old(i), U_old)
    w=w+exp(-U_old/k_B/T)
    W_old=W_old*w
Enddo
! REGROW NEW CONFIGURATION AT HEAD END
W_new=1
Do i=n_R-1, 1
    w=0
    Do t=1, n_T
        call genpos(r(t)) ! generate trial position of segment
        ! based on internal constraints (e.g., bond)
        call energy (r(t),U(t)) ! calculate external energy of trial
        w=w+exp(-U(t)/k_B/T)
    Enddo
    W_new=W_new*w
    call rand (q) ! generate a random number between 0 and 1
    p(0)=0
    p(1)=exp(-U(1)/k_B/T)/w
    do t=2, n_T
        p(t)=p(t-1)+exp(U(t)/k_B/T)/w
    Enddo
    do t=1, n_T
        if (q>p(t-1) and q(t)) then
            select =t ! trial position t has been selected
        Endif
    Enddo
    r_new(i)=r(t) ! store new position in a temporary array
Enddo
! ACCEPT OR REJECT NEW CONFIGURATION BASED ON ROSENBLUTH CRITERIA
call rand (q)
if (q<W_new/W_old) then ! accept new configuration
    do i=n_R*N-n_R+1,1 ! shift indices of polymer
        R(i+n_R)=R(i)
    Enddo
    do i=1, n_R-1
        R(i)=r_new(i)
    Enddo
Endif
```

Scaling of the sampling time as a function of polymer length

The end-transfer scheme applied to a biopolymer can be simplified to the following mathematical model. The biopolymer with N repeating motif represents a single file of balls, where each ball represents a single repeating motif (Fig. 7). Initially, all balls are white (unsampled); at each step, analogous to an accepted end-transfer move, a ball is selected randomly from one end and placed at the opposite end; in this process the ball's color is changed from white to black (sampled). Note that if a ball is already black, it remains black irrespective of its transfer. The process is repeated until all the balls become black. The problem of determining the average time to sample the entire polymer then is the equivalent of determining the average number of steps required to turn all the balls black.

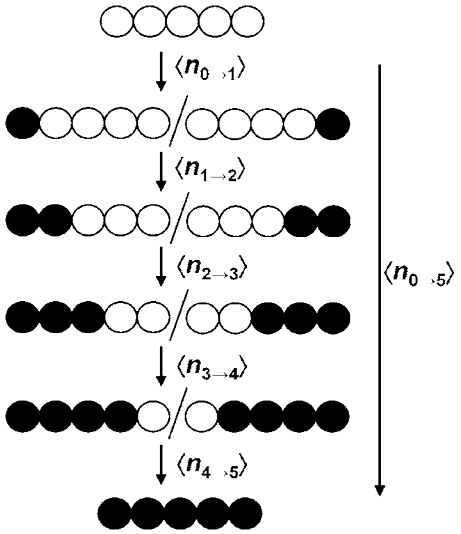

FIG. 7. Breakdown of the end-transfer procedure into intermediate steps. The black balls represent biopolymer motifs that have been transferred at least one from one end to another and the white balls represent motifs that are yet to be transferred.

The following key insight helps solve the problem more readily: for a transition from n to $n+1$ black balls to occur, the black and white balls must be separated out on opposite ends of the file (see Fig. 7). Only such a situation allows the possibility of the transfer (with probability 1/2) of a white ball from one end to the other, and consequently, the color change from white to black. If this is not the case, both ends of the file are capped by a black ball and the n to $n+1$ transition cannot occur. This insight helps us understand why the process of changing zero black balls to N black balls can be broken down into $N-1$ intermediate stages containing $1, \ldots, N-1$ black balls, in sequence, arranged together at one of the ends of the file, as shown in Fig. 7. The total number of steps taken to reach the final configuration with N black balls $\langle n \rangle$ is given by the cumulative sum of average number of steps required to go from one intermediate stage to the next, given by $\langle n_{i-1\to i} \rangle$, where $i=1, \ldots, N-1$.

Clearly, $\langle n_{0\to 1} \rangle = 1$, as it takes a single end-transfer move to turn the first ball black. It can be shown that $\langle n_{1\to 2} \rangle$ is given by the infinite series

$$\langle n_{1\to 2} \rangle = 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{4} + 3 \cdot \tfrac{1}{8} + \cdots, \tag{A1}$$

where each term represents the steps taken to transfer a white ball from one end to other and its associated binomial probability of occurrence. The above sum represents a combined arithmetic/geometric progression whose sum is given by $\langle n_{1\to 2} \rangle = 2$. Similarly, $\langle n_{2\to 3} \rangle = 3$ and, in general, $\langle n_{i-1\to i} \rangle = i$. This implies that the total number of steps needed to change all white balls to black is given by

$$\langle n \rangle = \sum_{i=1}^{N} \langle n_{i-1\to i} \rangle = 1 + 2 + 3 + \cdots + N = \frac{N(N+1)}{2}. \tag{A2}$$

Hence, all motifs of a biopolymer get sampled, on average, in $\tau_N = \tau_1 N(N+1)/2$ MC steps, if each transfer move takes τ_1 MC steps for acceptance.

[1] D. L. Ermak and J. A. McCammon, J. Chem. Phys. **69**, 1352 (1978).

[2] E. Dickinson, Chem. Soc. Rev. **14**, 421 (1985).

[3] K. Binder, in *Computer Simulation of Polymers*, edited by E. A. Colbourn (Longmans Scientific and Technical, London, 1994), p. 91.

[4] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego, 2002).

[5] J. Batoulis and K. Kremer, J. Phys. A **21**, 127 (1988).

[6] J. I. Siepmann, Mol. Phys. **70**, 1145 (1990).

[7] J. I. Siepmann and D. Frenkel, Mol. Phys. **75**, 59 (1992).

[8] J. J. de Pablo, M. Laso, and U. W. Suter, J. Chem. Phys. **96**, 2395 (1992).

[9] D. Frenkel, G. C. A. M. Mooij, and B. Smit, J. Phys.: Condens. Matter **4**, 3053 (1992).

[10] M. Vendruscolo, J. Chem. Phys. **106**, 2970 (1997).

[11] Z. Chen and F. A. Escobedo, J. Chem. Phys. **113**, 82 (2000).

[12] S. Consta, N. B. Wilding, D. Frenkel, and Z. Alexandrowicz, J. Chem. Phys. **110**, 3220 (1999).

[13] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).

[14] G. Arya, Q. Zhang, and T. Schlick, Biophys. J. **91**, 133 (2006).

[15] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).

[16] G. Arya and T. Schlick, Proc. Natl. Acad. Sci. U.S.A. **103**, 16236 (2006).

[17] Q. Zhang, D. A. Beard, and T. Schlick, J. Comput. Chem. **24**, 2063 (2003).

[18] J. Sun, Q. Zhang, and T. Schlick, Proc. Natl. Acad. Sci. U.S.A. **102**, 8180 (2005).

[19] D. A. Beard and T. Schlick, Structure (London) **9**, 105 (2001).

[20] S. A. Allison, R. Austin, and M. Hogan, J. Chem. Phys. **90**, 3843 (1989).

[21] D. Stigter, Biopolymers **16**, 1435 (1977).

[22] B. Rupp and S. Parkin, *PDBSUP–A FORTRANP Program That Determines the Rotation Matrix and Translation Vector for Best Fit Superposition of Two PDB Files by Solving the Quaternion Eigenvalue Problem* (Lawrence Livermore National Laboratory, Livermore, CA, 1996).

[23] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[24] Q. Yan and J. J. de Pablo, J. Chem. Phys. **111**, 9509 (1999).

[25] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

[26] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003).