Monte Carlo Simulation Algorithm for B-DNA

Steven C. Howell,[a] Xiangyun Qiu,[b] and Joseph E. Curtis*[a]

Understanding the structure–function relationship of biomolecules containing DNA has motivated experiments aimed at determining molecular structure using methods such as small-angle X-ray and neutron scattering (SAXS and SANS). SAXS and SANS are useful for determining macromolecular shape in solution, a process which benefits by using atomistic models that reproduce the scattering data. The variety of algorithms available for creating and modifying model DNA structures lack the ability to rapidly modify all-atom models to generate structure ensembles. This article describes a Monte Carlo algorithm for simulating DNA, not with the goal of predicting an equilibrium structure, but rather to generate an ensemble of plausible structures which can be filtered using experimental results to identify a sub-ensemble of conformations that reproduce the solution scattering of DNA macromolecules. The algorithm generates an ensemble of atomic structures through an iterative cycle in which B-DNA is represented using a wormlike bead–rod model, new configurations are generated by sampling bend and twist moves, then atomic detail is recovered by back mapping from the final coarse-grained configuration. Using this algorithm on commodity computing hardware, one can rapidly generate an ensemble of atomic level models, each model representing a physically realistic configuration that could be further studied using molecular dynamics. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24474

Introduction

The conformation and dynamics of DNA directly impact biological processes at a fundamental level, but there are few methods to rapidly generate model structures that agree with experimental measurements of DNA, thus limiting the understanding its structure–function relationship. Within eukaryotic cells, DNA is packaged as chromatin, a DNA–protein complex which serves to compact, organize, and protect DNA while also modulating genetic expression. The nucleosome core particle (NCP) is the molecular building block of chromatin and is composed of 147 base pair (bp) of DNA wrapped 1.67 times around a symmetric octamer of histone proteins.[1] In a nucleosome array, each NCP is linked by 10–90 bp of DNA as "beads-on-a-string".[2] The structure and packaging of an array of nucleosomes to form chromatin is critically important as it determines the accessibility of the genetic code, and therefore affects DNA-directed processes, including transcription, replication, recombination, and repair.[3,4]

Despite decades of effort to reveal chromatin structure–function relationship, the structure of chromatin and its connection to gene regulation remains an active area of research.[4] Efforts to determine the structure of chromatin using X-ray crystallography have led to atomic models for systems as simple as a bare DNA double helix,[5,6] and as complicated as a NCP[7–9] and short arrays of nucleosomes.[10,11] Cryo-electron microscopy experiments have begun to expand these results to additional nucleosome array constructs.[12] While much has been learned by studying NCPs and nucleosome arrays in these biologically artificial environments, understanding the structure–function relationship of chromatin *in vivo* requires experiments in solution. To date, efforts to identify model structures from solution studies of nucleosomes and nucleosome arrays have been limited to qualitative comparisons,[13] or structure models generated from dummy spheres,[14] manual modifications of atomic structures,[15] or by replacing nucleosome-bound DNA with linear DNA fragments.[16] Robust structure models of these and other nucleosome complexes require efficient simulation tools to complement the experimental methods used; in this context, a robust structure model refers to an atomic representation of a physically realistic molecular structure.

Small-angle X-ray and neutron scattering (SAXS and SANS) are well suited for studying a wide variety of DNA structures in biologically relevant conditions. In small-angle scattering (SAS) experiments, a beam of collimated radiation (photons or neutrons) interacts with inhomogeneities in the sample. The elastic scattering from these interactions encodes the temporal and spatial average of the pair-distance distribution of atoms within the sample. This ensemble average provides structural information related to characteristic internal distances, as well as the overall size and shape of the scattering particles, for example, molecular weight, radius of gyration.

More detailed structural information can be extracted from SAS data by matching theoretical scattering profiles from model structures to experimental scattering data. For many molecules, or molecular subunits, atomic coordinates are available

[a] *S. C. Howell, J. E. Curtis*
Neutron Condensed Matter Science Group, NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8562
E-mail: joseph.curtis@nist.gov
[b] *X. Qiu*
Department of Physics, The George Washington University, Washington, District of Columbia, 20052
Contract grant sponsor: EPSRC; Contract grant number: EP/K039121/1; Contract grant sponsor: National Science Foundation; Contract grant number: CHE-1265821; Contract grant sponsor: National Science Foundation; Contract grant number: MCB-1616337

from X-ray crystallography measurements. Such atomic models serve as a starting point for evaluating the differences between the static and solution structures.[17]

To determine solution structures that have scattering profiles consistent with experiment, ensembles of robust macromolecular structures need to be rapidly generated and then their theoretical scattering profiles must be compared to experimental data. For systems containing regions of flexible amino acid or single-stranded nucleic acid, given an all-atom structure model of the system, the program SASSIE can be used to rapidly sample the conformation of the flexible regions to generate ensembles of structures.[18] No such tool exists for rapidly generating configuration ensembles of B-DNA, the native form of double-stranded DNA.

The various techniques currently used for modeling DNA include user-directed modifications of all-atom models,[16,15,19] *ab initio* modeling using dummy spheres,[20,21] rigid body modeling,[22,23] coarse-grained (CG) simulations,[24–37] and all-atom molecular dynamics (MD).[38–44] Each of these modeling techniques has different advantages and disadvantages, but only CG and all-atom simulations incorporate physics of the DNA molecule. While all-atom MD is the most exhaustive, the large number of atoms in nucleosomes and nucleosome arrays, over 25,000 atoms per nucleosome, make this method intractable using commodity hardware. MD simulations on high-performance computing resources may not sample enough configurations to find structures that match experimental results in a reasonable amount of time. To overcome the limitation caused by molecular complexity, CG simulations reduce the degrees of freedom by replacing many atoms with a single CG bead. Many of these models maintain the ability to simulate DNA melting and hybridization while replacing each nucleotide with as few as 6–7,[27,32,36] or even 3 CG beads.[25,28–30,33,34,37] Of these, only those with 6–7 CG beads per nucleotide provide the ability to map back from a CG representation to an all-atom model after a simulation.[27,32,36] Though well suited for many different purposes, none of these techniques offer the capability to rapidly modify model structures thereby generating an ensemble of structural configurations with atomic detail.

In this report, we present a Monte Carlo (MC) algorithm for B-DNA, not designed for *a priori* structure prediction but for rapidly generating ensembles of all-atom macromolecular structures to compare to experimental data. For nucleosomes and nucleosome arrays, bending and twisting of B-DNA are the dominant mechanisms for modifications of the macromolecular structure. To simplify the simulations while allowing for B-DNA bend and twist moves, we represent each DNA bp using a single CG bead. This level of simplicity allows for a straightforward implementation of MC sampling, a mechanism which typically explores configuration space more rapidly than MD simulations but for which there are few examples. To recover atomic detail after performing CG MC sampling, our algorithm uses the final orientation of each CG bead to reinsert the atoms of each bp. The resulting model must then be energy minimized to eliminate bond strains between the simulated base pairs.

To validate this algorithm, we compare resulting structural metrics to both experimental and theoretical properties of B-DNA. As CG moves create discontinuities in the resulting structure, we demonstrate that a short energy minimization relaxes these discontinuities thereby producing robust atomic models. We illustrate ensembles of structure models for a single nucleosome and a short array of nucleosomes, filtered against mock experimental SAXS profiles. This method of filtering an ensemble using experimental SAS data provides a powerful means to determine the structures of DNA macromolecules in solution.

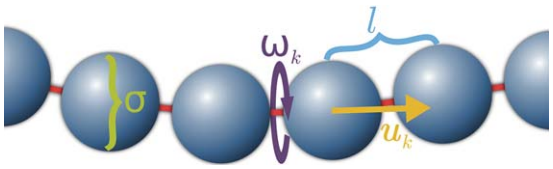

Figure 1. Illustration of the bead–rod model. The CG beads each have width σ and are connected by an inextensible rod of length l. The unit vector $\boldsymbol{u}_k$ describes the orientation of the k^{th} rod. The angle ω_k represents the twist angle between beads k and $k+1$. This model builds on the bead–rod model reported by Wang et al. [31] by allowing for DNA twisting. [Color figure can be viewed at wileyonlinelibrary.com]

Methodology

In designing an algorithm for modeling DNA, our focus was to rapidly generate atomic models that cover a wide range of configurations while employing Metropolis Monte Carlo sampling of DNA. The CG representation we use to achieve this goal is based on the bead–rod model reported by Wang et al.[31] which originates from the classical wormlike chain model.[45] This simple model, illustrated in Figure 1, represents DNA using N beads connected by $N-1$ inextensible rods of length l, for a total contour length, $L=(N-1)l$. The energetics of this model employs a bending penalty between adjacent beads and an excluded volume repulsion between beads.

The bending energy, U_{bend}, is obtained by discretizing the wormlike chain model:

$$U_{\text{bend}} = \frac{k_B T}{2} L_p \int_0^L \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds \quad \rightarrow$$
$$U_{\text{bend}} \approx k_B T \frac{L_p}{l} \sum_{k=1}^{N-2} (1 - \boldsymbol{u}_k \cdot \boldsymbol{u}_{k+1}) \tag{1}$$

where k_B is the Boltzmann constant, T is the temperature, L_p is the experimentally determined salt dependent persistence length of DNA (534 Å in 10 mM Na^+,[46,47]) and $\boldsymbol{u}_k$ is the unit vector along the kth rod. The experimentally determined salt dependent persistence length is a measure of the total persistence length, which is the sum of the inherent and the electrostatic persistence lengths.[48] Consequently, this bending energy term accounts for the non-bonding electrostatic interactions in addition to the bonded interactions of the DNA molecule.

The energy term describing the excluded volume between any two beads, i and j, separated by a distance r_{ij}, is represented using a Weeks–Chandler–Andersen (WCA) form of a purely repulsive Lennard–Jones potential:

$$U_{WCA}=\begin{cases}4k_BT\left[\left(\sigma/r_{ij}\right)^{12}-\left(\sigma/r_{ij}\right)^{6}+1/4\right] & r_{ij}<2^{1/6}\sigma\\ 0 & r_{ij}\geq 2^{1/6}\sigma\end{cases} \quad (2)$$

where the bead width, σ, determines the range of the interaction. We fixed $\sigma=2^{-1/6}l$ so U_{WCA} applies only for $r_{ij}<l$.

Expanding on the bending energetics of the bead–rod model, we added a harmonic twist energy:

$$U_{twist}=\frac{k_BT}{2}\,\kappa\sum_{k=1}^{N-1}\left(\omega_k-\bar{\omega}\right)^2 \quad (3)$$

where $\kappa=0.062\times(1^{\circ})^{-2}$ is the average twist force constant, ω_k is the twist angle between base pairs k and $k+1$, and $\bar{\omega}=+35.4^{\circ}$ is the average twist angle.[49] Performing twist moves based on this harmonic twist energy allows for increased conformational sampling while maintaining the appropriate major and minor groove widths in simulated DNA.

Though this bead–rod model, with its typical level of granularity, was originally developed for long DNA ($L>L_p$), we demonstrate in this manuscript that with the addition of hard-sphere (HS) potentials between CG beads and back-mapped atoms it can also be used to simulate MC moves of short DNA structures ($L<L_p$). Considering a nucleosome array as an example, the conformation of each 10–90 bp of linker DNA significantly change the overall scattering.[13] With such a short region of flexible DNA, using one CG bead to represent each bp maximizes the range of structures generated through MC simulations. One CG bead for each bp translates to an average rod length between beads of $l=3.38$ Å, and a bead width of $\sigma=3.01$ Å. As 3.01 Å is much smaller than the width of B-DNA, the WCA energy term does not fully enclose all the atoms represented by each bead and is therefore not sufficient to prevent overlap between atoms in the final all-atom structures. We avoid such overlap using an additional steric HS potential at both the CG and all-atom levels. At the CG level, we use a HS diameter of 19 Å, corresponding to the atomic width of B-DNA, but only apply this restriction to beads separated by at least 6 beads, or roughly 20 Å in either direction along the DNA chain. After reinserting the transformed all-atom group back into the transformed configuration, we use a HS diameter of 0.8 Å for each non-hydrogen atom.

Sampling bend and twist moves, this algorithm generates an ensemble of all-atom structures using an iterative multi-step process. Figure 2 illustrates this process for a 40 bp linear DNA fragment containing 10 base pairs designated to be the only flexible region. The algorithm randomly selects one of the user designated flexible groups, replaces each flexible DNA bp with a CG bead positioned at the bp reference frame origin,[50] then stores the coordinates of the bp atoms with respect to that reference frame. The algorithm then generates a new structure for the flexible group by selecting a CG bead and sampling a move about one of the three coordinate axes of the DNA bp reference frame. All successive beads and any other post atoms are transformed with respect to the move sampled. This new group structure is accepted according to Maxwell–Boltzmann statistics[51] using the energies in eqs. (1)–(3) and the HS potentials to only accept structures for which no beads nor atoms overlap. After a new group structure is accepted, the algorithm recovers an all-atom representation for the flexible DNA by reinserting the atoms for each CG bead according to its new origin and orientation. A screened Debye–Hückel potential is applied to this complete all-atom structure to account for long range electrostatic interactions. As a last check, the algorithm verifies that the coordinates of the altered group atoms do not overlap with the other atoms in the complete molecule before accepting the new configuration and proceeding to the next iteration. This cycle of selecting a flexible group, coarse-graining the flexible DNA, sampling MC moves, mapping back to all-atoms, and then checking for overlap is repeated for the

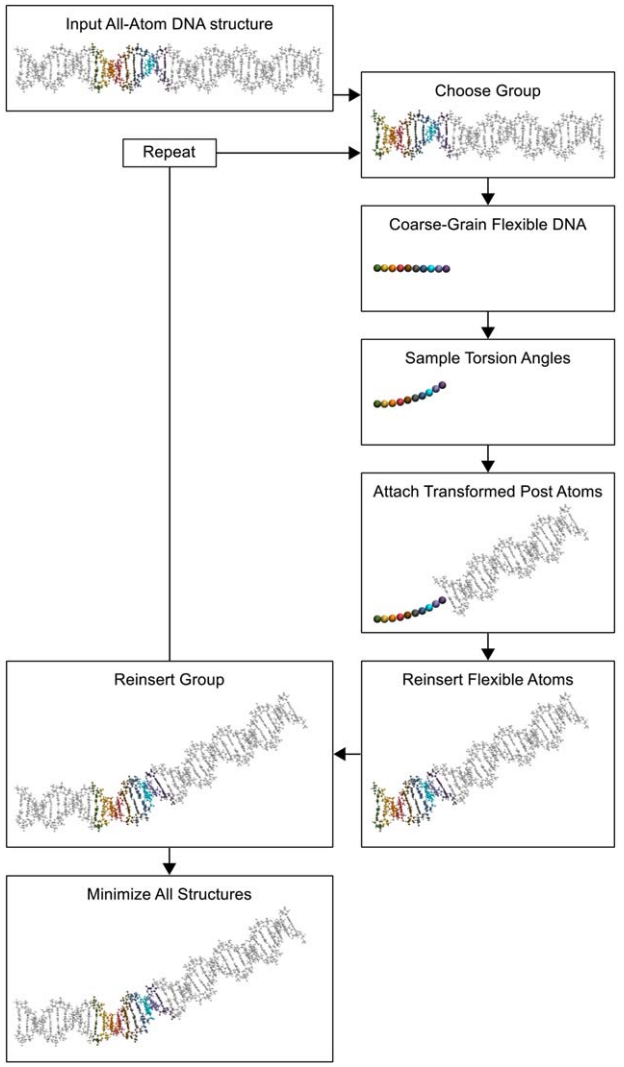

Figure 2. Schematic of the DNA Monte Carlo algorithm. Illustration of MC sampling for a 40 bp linear DNA segment with 10 base pairs of flexible DNA. Each flexible bp is uniquely colored. Note that the bead width used in the WCA energy, $\sigma=3.01$ Å, does not enclose the volume of the atoms represented by each bead. Consequently, the resulting energies do not fully prevent overlap of atoms when recovering an all-atom representation after CG moves. Such overlap is prevented by including a HS potential between beads separated by more than 6 bp. This HS diameter is set at 19 Å, the atomic width of B-DNA. DNA structures rendered using VMD.[65] [Color figure can be viewed at wileyonlinelibrary.com]

designated number of trial steps. After the cycle completes, all structures must be energy minimized to relax atomic bond strains between flexible DNA base pairs.

A key part of this multi-step process is the mechanism of representing a DNA bp using a single bead, then reinserting the atoms from that bp back onto the bead after performing MC sampling. This requires first determining the coordinates of the atoms within each bp with respect to the standard DNA bp reference frame, defined by Olson et al. as follows (illustrated in fig. 1 of Ref. 50). In this reference frame, the *x*-axis is the perpendicular bisector of the C1′...C1′ vector spanning the bp and points in the direction of the pseudo-dyad axis of an ideal Watson–Crick bp. The *y*-axis runs parallel to the C1′...C1′ vector, points in the direction of the sequence strand, and passes through the intersection of the *x*-axis and the vector connecting the pyrimidine Y(C6) and the purine R(C8) atoms. The *z*-axis is defined as the cross product of the *x* and *y* axes, that is, $\boldsymbol{z}=\boldsymbol{x}\times\boldsymbol{y}$, and consequently points along the 5′ to 3′ direction of the sequence strand. Once this reference frame has been determined, a CG bead is placed at the bp center. During the MC sampling process, both the bead position and orientation are updated with each accepted move. The final position and orientation are used to reinsert the atoms for each bp after the MC sampling. Though this produces complete atomic models, an unavoidable consequence of performing CG moves is that atomic bonds between DNA base pairs become compressed or extended. A short energy minimization relaxes these strains. This CG then reverse-CG process sufficiently reduces the complexity of performing MC sampling thereby facilitating the rapid generation of atomic models by varying the DNA structure in an energetically sampled manner.

Wall clock estimates using a standard desktop computer indicate that, without energy minimization, the algorithm would require 24 h to generate 430,000 different structures for a 3800 atom 60 bp linear DNA model, and 700 different structures for a 107,000 atom model of an array of four nucleosomes (tetranucleosome). For these estimates, we designated the middle 58 bp of the linear DNA as flexible and the five 20 bp linkers as flexible in the tetranucleosome model.

This algorithm, with the energy terms presented, is not presently suited to simulate A-DNA or Z-DNA. Implementing the bending energetics described in eq. (1) requires that the bead centers lie approximately on the line passing through the center of the DNA chain. Unlike B-DNA, this is not the case for A-DNA. To account for this spiral would require a different definition for the CG beads. While the bead centers for Z-DNA can be approximated by the DNA center line, there are only limited experimental measurements of the persistence length[52] and the harmonic twist force constant. Consequently, we currently only apply this bead–rod model to simulate B-DNA.

After applying this model to simulate B-DNA, evaluating how well the generated structures represent experimental SAS data requires calculating the theoretical scattering intensity from each structure. In the simplified case of uniform scattering power of all atoms, the scattering intensity vs momentum transfer ($I(Q)$ vs. Q) is calculated using

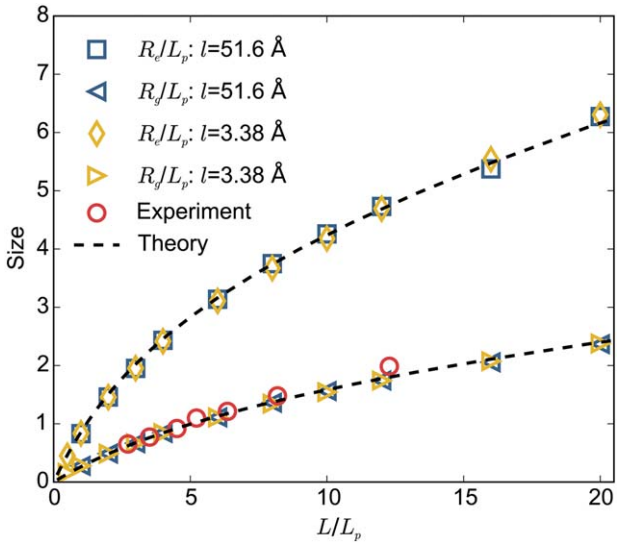

Figure 3. Bulk properties of the wormlike bead–rod model from DNA bend moves. Comparison of the end-to-end distance, R_e, and the radius of gyration, R_g, from simulations (blue and yellow symbols) versus the wormlike chain theory[55] (lines) and experimental R_g measurements[56] (red circles). The R_e and R_g values are normalized by the DNA persistence length, L_p. To validate the implementation of the energy terms in eqs. (1) and (2), these MC simulations only allowed for bend moves. In these simulations, the length between CG beads, l, was set at either 51.6 Å (blue), to reproduce the results reported by Wang et al.,[31] or 3.38 Å (yellow) to correspond to one bp per CG bead. Error values are smaller than the marker size. [Color figure can be viewed at wileyonlinelibrary.com]

$$I(Q)=4\pi\int_0^\infty P(r)\,\frac{\sin(Qr)}{Qr}\,dr \tag{4}$$

where $Q=4\pi\sin(\vartheta)/\lambda$, λ is the incident (photon or neutron) wavelength, 2ϑ is the scattering angle, and $P(r)$ is the pair-distance distribution function describing the probability of atoms being separated by a distance r. For the examples we provide in Figures 8 and 9, we calculated the theoretical scattering intensities using the open-source application FoXS,[53,54] which calculates $I(Q)$ in a two-step process. This application first determines $P(r)$ by explicitly calculating all the interatomic distances and implicitly modeling the first hydration layer of the molecule. It then performs a numerical integration at each point in Q according to eq. (4).

Validation

A primary objective of this algorithm is to generate atomic models of robust biological molecules which could then be selectively studied further. Therefore, it is critically important that the energetics, both bending and twisting, agree with reported experimental results and accepted theory. To validate the energetics of the MC sampling, we compared the structural properties from simulation results to previous studies. Further, we justified the process of energy minimizing the reinserted atomic coordinates by comparing the dihedral angles of resulting structures against those obtained using all-atom MD simulations.

To validate the bending energetics in eqs. (1) and (2), we performed MC simulations that only sampled bend moves (not

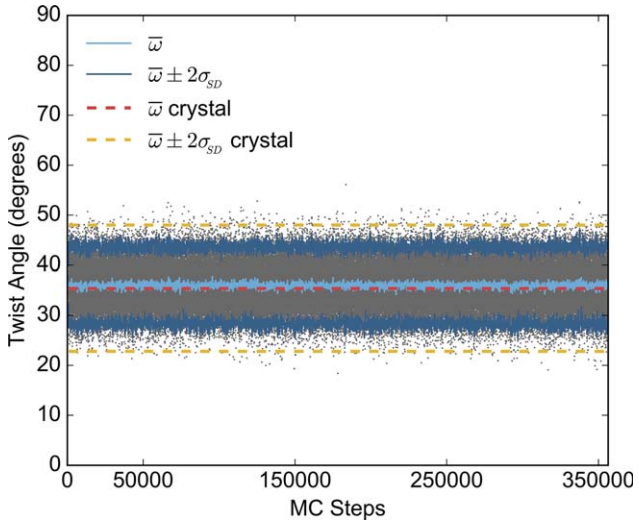

Figure 4. Survey of simulated DNA twist moves. Each gray dot represents the DNA twist angle, ω, between two adjacent simulated DNA bp. At each MC step the light blue shows the average twist angle, $\bar{\omega}$, and the dark blue shows two standard deviations, $2\sigma_{SD}$, from the mean. The dashed red and yellow lines indicate $\bar{\omega}$ and $\bar{\omega} \pm 2\sigma_{SD}$ from a crystal structure survey.[49] The $\bar{\omega}$ in the simulated structures closely matches the crystal structures, $+35.4^{\circ}$, while the distribution about the mean is tighter for simulated structures. [Color figure can be viewed at wileyonlinelibrary.com]

allowing for twist moves) then calculated the root-mean-square end-to-end distance, R_e, and the radius of gyration, R_g. For these simulations, we used two different sets of parameters for the rod length between beads, $l=51.6$ Å, in direct comparison to the implementation of Wang et al.,[31] and $l=3.38$ Å, to reproduce our implementation of 1 bead for each DNA bp. As seen in Figure 3, the results from the simulations using both conditions agree with the accepted wormlike chain theory curves[55] and experimental measurements[56] for DNA ranging in length from 265 Å, to 10,600 Å (0.5–20 $\times L_p$). Each calculated value represents the average of over 300 configurations sampled every 10^5 steps after an initial equilibration of 10^6 steps.

To validate our implementation of the twist energetics in eq. (3), we performed a MC simulation sampling only twist moves then calculated the mean twist angle and standard deviation about that mean. Figure 4 shows the twist angles between 58 flexible DNA base pairs over 350,000 MC steps. Following an initial equilibration of 1000 steps, we calculated the twist angle between flexible base pairs after every 100 MC steps (gray dots). The mean twist angle from our MC results, light blue, has an overall average of $+35.9^{\circ}$, a close match to the expected $+35.4^{\circ}$ used as the equilibrium angle in the harmonic energy, and determined from a crystal structure survey of DNA–protein complexes.[49] The solid dark blue and dashed yellow lines respectively identify two standard deviations from the mean ($2\sigma_{SD}$) for our MC result and the crystal structure survey. The smaller standard deviation seen for MC generated structures indicates that $\kappa = 0.062 \times (1^{\circ})^{-2}$ is a conservative selection, well within the limits for B-DNA.

In addition to validating the DNA bend and twist energetics, we also evaluated the ability of the algorithm to generate

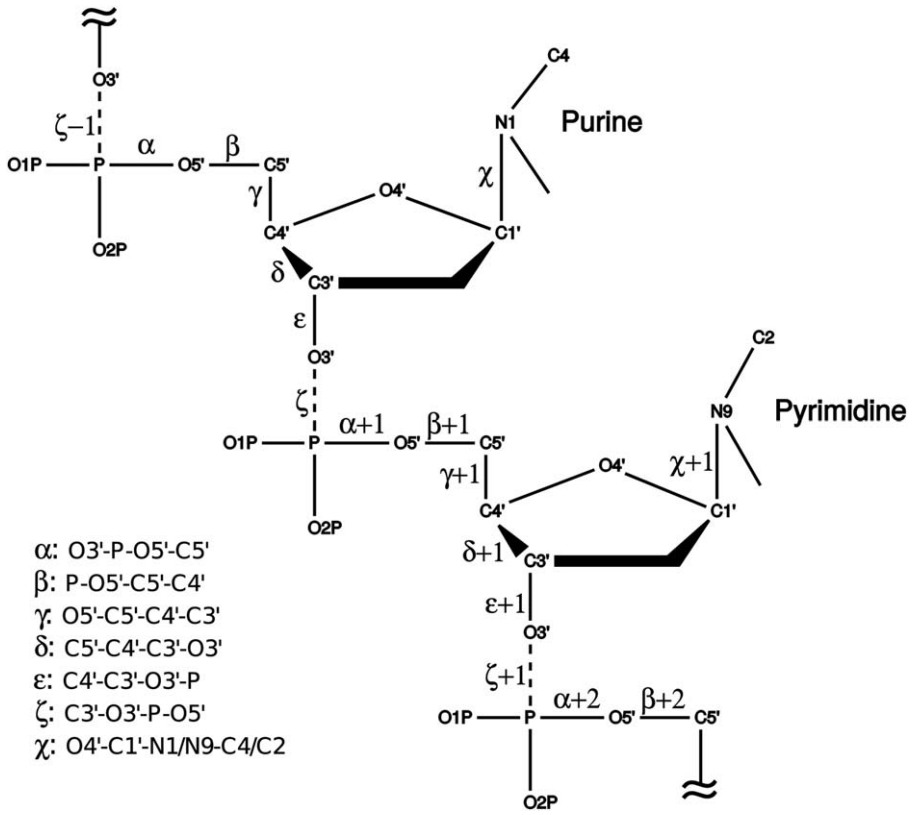

Figure 5. Schematic of DNA illustrating the α, β, γ, δ, ε, η, and χ, dihedral angles. The separation between DNA bases is between the O3′ atom of one base and the P atom of the following base. CG MC moves strain this bond (dashed line) causing a distortion of the ε, ζ, and $\alpha+1$ dihedral angles. Energy minimization relaxes this strain.

Table 1. Summary of DNA dihedral angles. Dihedral angles from a MD simulation of a solvated and ionized B-DNA model, compared to those from 34 B-DNA crystal structures.[59] The tabulated values show the mean plus or minus one standard deviation. Subscripts differentiate BI and BII conformations (and purine versus pyrimidine for χ_I).

Angle	MD		Crystal	
α	300°	±20°	298°	±15°
β_I	165°	±24°	176°	±9°
β_{II}	152°	±26°	146°	±8°
γ	52°	±15°	48°	±11°
δ_I	131°	±21°	128°	±13°
δ_{II}	136°	±13°	144°	±7°
ε_I	185°	±20°	184°	±11°
ε_{II}	247°	±30°	246°	±15°
ζ_I	259°	±22°	265°	±10°
ζ_{II}	182°	±37°	174°	±14°
$\chi_{I_{pur}}$	259°	±16°	258°	±14°
$\chi_{I_{pyr}}$	249°	±22°	241°	±8°
χ_{II}	272°	±16°	271°	±8°

DNA structures which maintain the canonical structure of B-DNA despite performing CG moves. The essential step that relieves the strain caused by CG moves is energy minimization of the final atomic structures. This strain occurs between the O3′ atom on one base and the P atom of the following base. An ideal mechanism for evaluating the degradation caused by this strain is an analysis of the three dihedral angles which contain the O3′–P bond, illustrated in Figure 5: ε, ζ, and $\alpha+1$.

To determine the acceptable ranges for the DNA dihedral angles, α, β, γ, δ, ε, ζ, and χ, we evaluated a 75 ns MD simulation of a solvated and ionized 12 bp B-DNA model (PDB ID: 1I9D[57]). The MD simulation was run using NAMD[58] with the CHARMM36 force field,[38] which is noted for its improved ability to reproduce experimentally observed sampling of the different conformations of B-DNA. Plots of each dihedral angle as a function of time are shown in Supporting Information Figs. 10–16. Table 1 summarizes the resulting average dihedral angles and their standard deviations together with these same values from a survey of 34 B-DNA crystal structures.[59] We note the consistency between the average angles from the MD simulation and the crystal structure survey with the angles from simulation having a larger standard deviation.

Equipped with the expected DNA dihedral angles from MD simulations, we evaluated these same angles for a set of structures generated using our CG MC algorithm with a particular focus on the angles containing the O3′–P bond. For the starting structure, we used a 60 bp linear DNA model generated from a random sequence using the 3D-DART DNA structure modeling server.[60] Using NAMD[58] with the CHARMM36 force fields,[38] we prepared this model for simulations by performing 2000 energy minimization steps followed by 200 MD steps (0.2 ps) then another 2000 energy minimization steps. After this preparation process, all the model dihedral angles were within two standard deviations of the mean values from MD simulations, shown in Table 1. With this prepared starting structure, we iteratively sampled MC moves, using select values for $\delta\vartheta_{max}$, creating a trajectory containing over 1000 times as many accepted steps as the number of flexible DNA bp. For a sample structure generated using 71,000 MC steps, Figure 6 compares select scatter plots of DNA dihedral angles, before and after energy minimization. The shaded regions show the angles within two standard deviations of the mean; yellow and red respectively represent BI and BII backbone conformations. Before energy minimization approximately 20% of the α, ε, and ζ dihedrals are beyond two standard deviations of the mean, but after 2000 energy minimization steps only 3% are beyond the same limit, an acceptable amount for a normal distribution.

To evaluate how the percentage of dihedral angles within the two standard deviation range changes as a function of the number of MC steps, for each $\delta\vartheta_{max}$ trajectory we calculated the dihedral angles from every hundredth structure. Figure 7 plots this percentage as a function of the number of MC steps, comparing angles before (left) and after (right) the 2000 energy minimization steps. Note that for all three of the dihedrals containing the O3′–P bond, the percentage in range, regardless of the $\delta\vartheta_{max}$ used, are relatively indistinguishable after 200 MC steps per bp. Also, after energy minimization, for nearly every structure all seven of the dihedral angles are within the 95% cutoff (dashed line) for a normal distribution. This demonstrates that performing 2000 energy minimization steps sufficiently relaxes the bond strain between stacked bases even

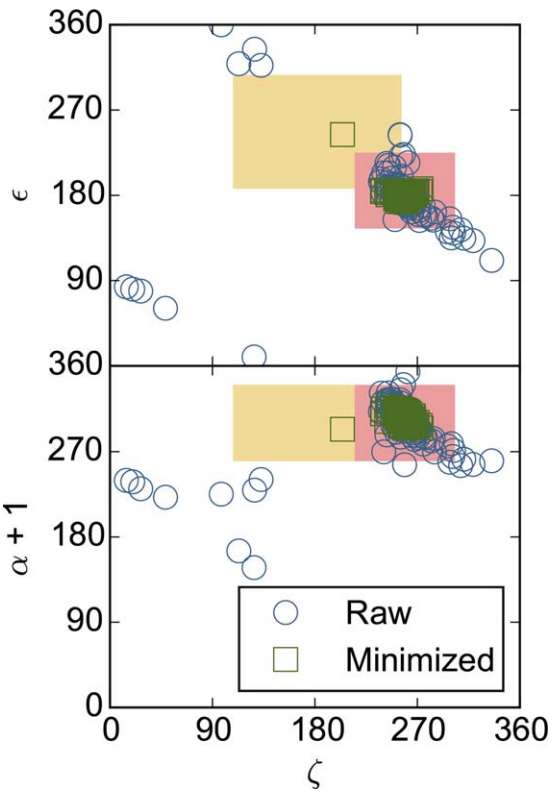

Figure 6. Raw versus energy minimized dihedral angles. Scatter plots of the DNA dihedral angles between adjacent DNA bp. These angles were extracted from a 60 bp DNA model after each of the flexible bases had experienced an average of 1233 MC steps using $\delta\vartheta_{max}=10^{\circ}$. The red and yellow patches respectively show the two standard deviation ranges for BI and BII DNA, based on MD simulation results (Table 1). [Color figure can be viewed at wileyonlinelibrary.com]

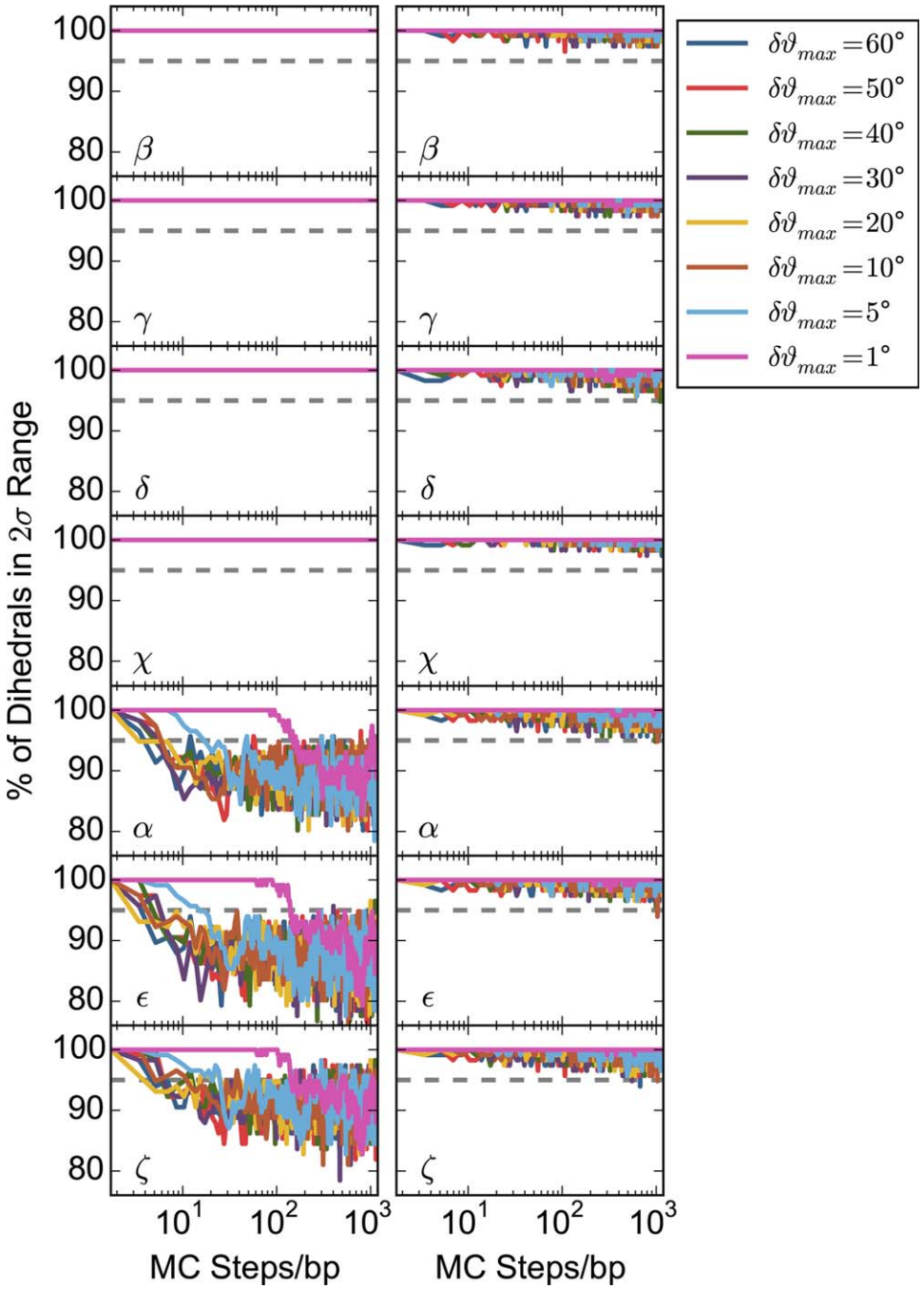

Figure 7. Percentage of DNA dihedral angles, before (left) and after (right) energy minimization, that are within two standard deviations of the mean, based on MD simulation results (Table 1). This percentage is shown as a function of the number of accepted MC steps normalized by the number of flexible base pairs. The dashed lines mark the 95% cutoff for the amount of dihedral angles that should be within two standard deviations of the mean. [Color figure can be viewed at wileyonlinelibrary.com]

after performing 1000 MC steps per flexible bp, with a maximum $\delta\vartheta \leq 60^{\circ}$.

In exploring energy minimization, we recognized that an excessive number of minimization steps lead to degradation of the DNA molecule, as noted in the literature.[61] To investigate this degradation, we iteratively performed 2000 energy minimization steps on the 3D-DART linear DNA without performing any MC simulations. DNA degradation was not apparent from evaluating the dihedral angles as a function of the number of energy minimization iterations, Supporting Information Fig. 17, but was visually apparent after as few as 10 iterations, or 20,000 total energy minimization steps. This degradation results in a collapse of the DNA minor groove. Calculating the root-mean-square deviation (RMSD) provided a better measure for ensuring that the number of energy minimization steps was not excessive. The RMSD was 0.35 after one iteration but increased from 2.3 after 10 iterations to 6.7 after 80 iterations.

Based on these validations, the recommended protocol to produce robust atomic structures, given a robust starting structure, is to perform up to 1000 MC steps per flexible bp

Table 2. Summary of operation and simulation parameters. The "Standard Operation" values are based on the validations discussed and should be followed when applying this algorithm. The "Advanced Inputs" allow users to tune the energetics according to their experimental conditions. The "Internal Parameters" are fixed parameters embedded in our implementation of this model.

	Default
Standard Operation	
Maximum MC angle ($\delta\vartheta_{max}$):	10°
MC steps before minimizing:	$\leq$ 1000 per flexible bp
Energy minimization procedure:	2000 steps
Advanced Inputs	
DNA persistence length (L_p):	534 Å
Twist force constant (κ):	$0.062\times(1^\circ)^{-2}$
Internal Parameters	
Granularity:	1 bp/bead
Rod length (l):	3.38 Å
Bead width (σ):	3.01 Å
Mean twist angle (ω):	35.4°

followed by 2000 energy minimization steps. This means that when using a structure from a previous MC simulation as the starting structure for another MC simulations, one should use the non-minimized structure to avoid excessive minimization, which leads to deformation of the DNA major and minor groove. Additionally, in this situation, one should reduce the total number of MC steps simulated by the number of MC steps already performed, as we have only verified up to 1000 MC steps per flexible bp. After completing the MC simulations, 2000 energy minimization steps should be performed on each structure to relieve the strain caused by CG moves. Subsequently, structures with theoretical scattering profiles that matches an experimental data set can be further studied using MD simulations.

A web server hosting this MC algorithm is freely available at https://sassie-web.chem.utk.edu/sassie2/. Table 2 contains a summary of the simulation parameters used for this server and the following examples. The "Standard Operation" values are those used in the validation of the algorithm and thus are the values users should use when performing B-DNA simulations. The "Advanced Inputs" are settings users can adjust to tune the bend and twist energetics to the experimental conditions they want to simulate. The "Internal Parameters" are values fixed within our implementation of this model. Using these parameters, the MC sampling algorithm provides rapid configuration space coverage reserving local configuration space exploration to more computationally expensive MD simulations.

Examples

We provide two examples to demonstrate the application of this algorithm. The first simulates flexible DNA tails of a NCP, and the second simulates flexible linker DNA between an array of four nucleosomes (tetranucleosome). These examples not only exhibit the creation of an ensemble of structures but also filtering that ensemble against mock experimental data.

We chose these examples because they correspond to two areas of active investigation, DNA breathing and unwrapping from a NCP[15,62], and the configuration of a short array of nucleosomes.[10–13] For the NCP, we designated 30 bp on each end to be flexible (in Figs. 8d and 8e, the flexible DNA is contained inside the white isosurface). For the tetranucleosome, we designated each of the three 20 bp linker DNA fragments to be flexible (in Figs. 9d–9k, the linker DNA is the DNA between nucleosomes) and also roughly 20 bp before the first and after the last nucleosome. For both the NCP and tetranucleosome examples, all proteins and any DNA not designated as flexible were considered rigid bodies throughout the simulations. The atomic coordinates for the starting structures were based on the crystal structures of the 1KX5 NCP[8] and the 1ZBB tetranucleosome.[11] These starting structures were prepared for MC simulations in the same manner as the 60 bp linear DNA model used for the dihedral angle verification, using 2000 energy minimization steps followed by 200 MD steps and then another 2000 energy minimization steps.

To create mock experimental data for both the NCP and tetranucleosome examples, we performed a separate MC simulation on each initial structure and selected a random resulting structure as the goal configuration. We then calculated the theoretical scattering profile from the selected structure using FoXS.[53,54]

The discrepancy between the SAXS profiles from the model data, I_m, and the mock experimental data, I_e, was quantified using the *R*-factor, defined as

$$R=\frac{\sum_{i=1}^{N}|I_e(Q_i)-c\,I_m(Q_i)|}{\sum_{i=1}^{N}|I_e(Q_i)|} \quad (5)$$

summing over N points in Q, and scaling the data sets to match at $Q=0$ using the scale factor $c=I_e(0)/I_m(0)$. We do not use the χ^2 statistic as we do not have an estimate of experimental error.

Figures 8 and 9 respectively show the results from comparing theoretical scattering profiles from 23,044 NCP structures and 86,001 tetranucleosome structures with the mock experimental data. In each figure, panel (a) shows a plot of the *R*-factor versus R_g for each structure with insets illustrating the configurations with the best and worst discrepancy. Panel (b) compares the experimental data with the calculated theoretical scattering from the best-matched, worst-matched, and average of the 500 best matches. Panel (c) shows a convergence analysis for the structure ensemble in both real space and reciprocal space, with an inset comparing the spatial range envelopes from the entire ensemble and from the subensembles of 500 best-matched structures. Panels (d) and (e) show from two angles the spatial range of the best subensemble overlaid on an illustration of the structure with the absolute smallest *R*-factor. This spatial range exemplifies the extent of structures which represent mock experimental data equally well. Panels (f)–(k) further depict the variance in best-matched structures by showing three additional example

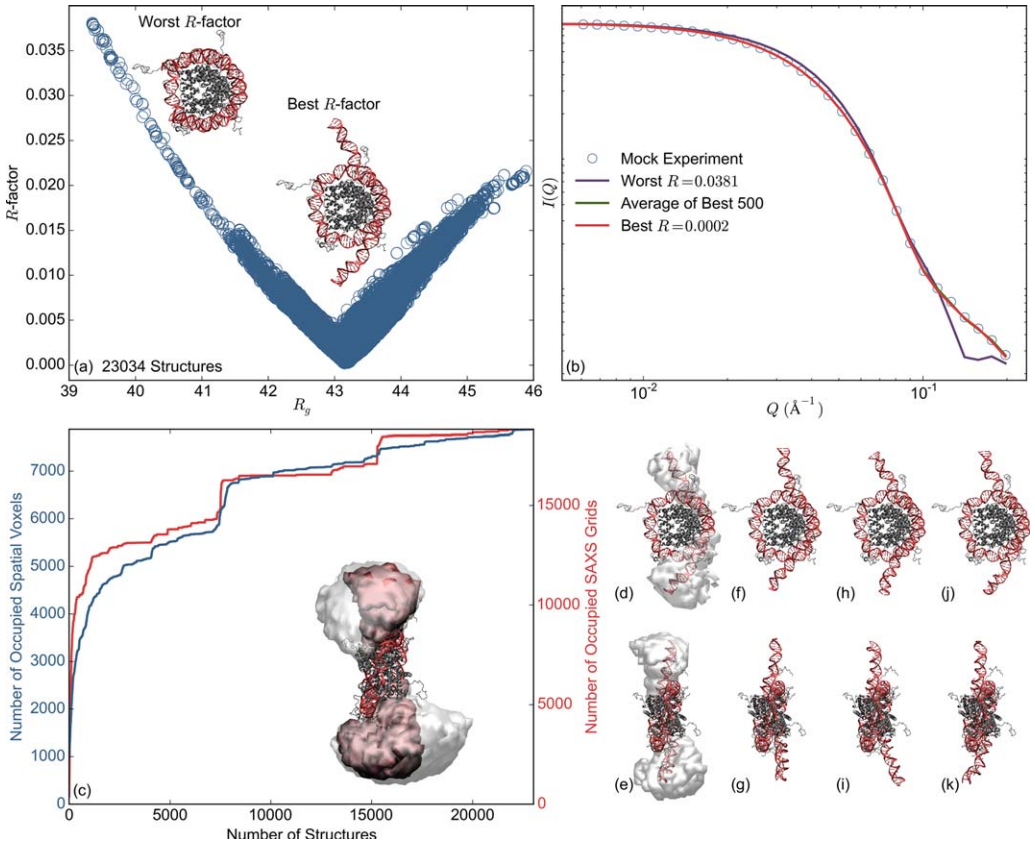

Figure 8. Example NCP simulation. An example of filtering a structure ensemble of NCP models against mock SAXS data. (a) *R*-factor versus R_g for each structure. Labeled insets illustrate the model structures with the best and worst *R*-factor. (b) mock SAXS data compared to the scattering profiles calculated from the worst match (purple), best match (red), and average of the 500 best matches (green). (c) spatial (blue) and SAXS (red) convergence analysis for the structure ensemble. The inset in (c) shows the spatial range envelope of the entire ensemble of structures (white) overlaid on the spatial range envelope for the sub-ensemble of the 500 best-matched structures (red). Panels (d) and (e) show from two angles the spatial range envelope for the best-matched sub-ensemble, overlaid on an illustration of the structure with the smallest *R*-factor. Panels (f)–(k) show three additional example structures from the best-matched sub-ensemble. The view in (d), (f), (h), and (j) is looking at a face of the NCP. The view in (e), (g), (i), and (k) is looking at the NCP dyad. NCP structures rendered using VMD.[65] [Color figure can be viewed at wileyonlinelibrary.com]

structures randomly selected from the sub-ensemble of 500 best structures.

When creating an ensemble of structures, it is important to consider when convergence has been reached, or when have a sufficient number of structures been generated to sufficiently sample the configuration space. To quantify the convergence of an ensemble, we analyze each new structure and its calculated scattering to determine how different it is from the other structures in the ensemble, both in real space, and in the reciprocal space of SAS. Figures 8c and 9c show the results of this analysis for the NCP and tetranucleosome respectively. For the real space analysis, we discretize all space into 5 Å voxels and for each structure count the number of new voxels that are occupied by either an alpha carbon or a phosphate atom. For the reciprocal space analysis, we first consider the calculated scattering curves from all structures in the ensemble to find the overall maximum, $I_{\max}(Q_i)$, and minimum, $I_{\min}(Q_i)$, at each Q_i point. We then discretize the possible $I(Q_i)$ between these extremum into $N_g = 100$ grids. For each new structure we bin the $I(Q_i)$ values using

$$g(Q_i) = \text{Round}\left(N_g \frac{I(Q_i) - I_{\min}(Q_i)}{I_{\max}(Q_i) - I_{\min}(Q_i)} \right) \quad (6)$$

where $g(Q_i)$ specifies the grid for each $I(Q_i)$ value. After each structure we count the total integer number of occupied grids. For both methods, as more structures are generated, the number of occupied voxels and grids will eventually plateau indicating the ensemble has reasonably explored possible spatial configurations and $I(Q_i)$ values. In our examples, both the NCP and tetranucleosome ensembles are reasonably converged. We further explored the spatial convergence of the individual nucleosomes in the tetranucleosome array and found that each nucleosome also demonstrates convergence (see Supporting Information fig. 19).

In Figs. 8 and 9, plots of the spatial range envelopes, shown as insets in panel (c) and also panels (d), and (e), allow one to visualize the extent in space that is covered by the structure ensembles. Envelopes of the best-matched sub-ensembles identify the configuration range of the structures that have scattering profiles that are consistent with the mock

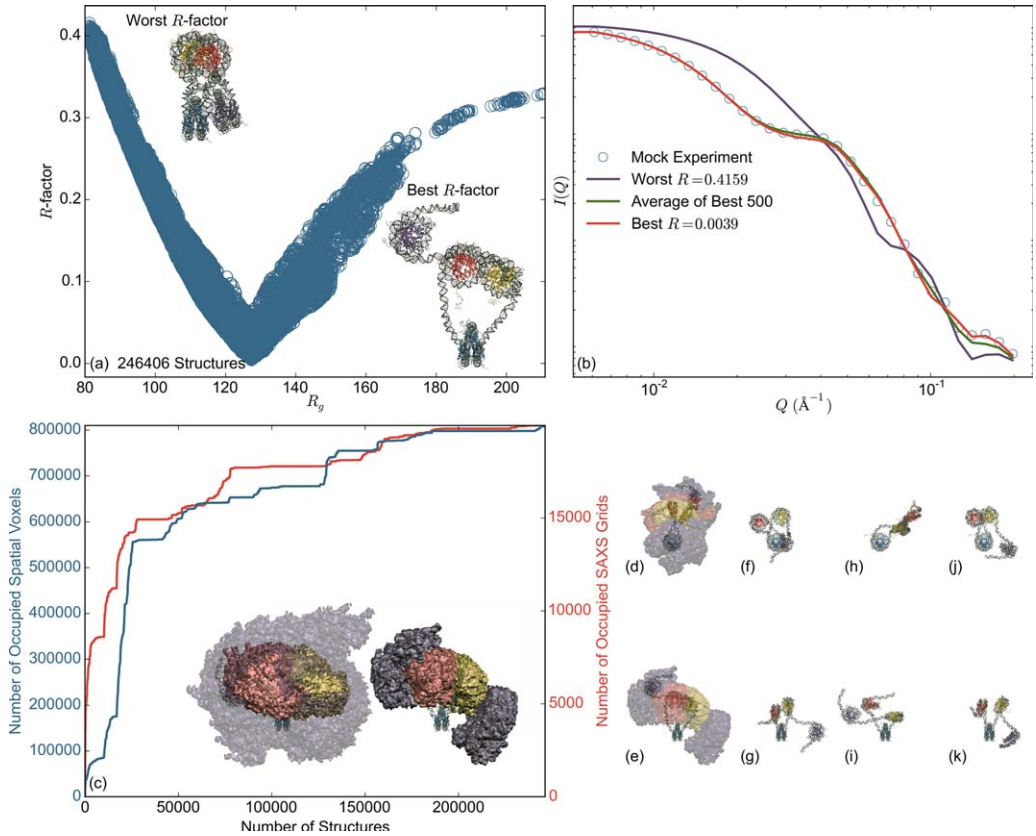

Figure 9. Example tetranucleosome simulation. An example of filtering a structure ensemble of tetranucleosome models against mock experimental SAXS data. (a) *R*-factor versus R_g for each structure. Labeled insets illustrate the model structures with the best and worst *R*-factor. (b) mock SAXS data compared to the scattering profiles calculated from the worst match (purple), best match (red), and average of the 500 best matches (green). (c) spatial (blue) and SAXS (red) convergence analysis for the structure ensemble. The inset in (c) shows the spatial range envelope of the entire ensemble of structures (left) together with the envelope for the sub-ensemble of the 500 best-matched structures (right). As a reference point, each structure in the ensemble was aligned using the second NCP in the array, illustrated with a blue protein core. The different colored regions indicate the range envelopes for the individual NCPs, where red, yellow, and purple correspond to the range of NCP_1, NCP_3, and NCP_4, respectively. Panels (d) and (e) show from two angles the spatial range envelope for the best-matched sub-ensemble, overlaid on an illustration of the structure with the smallest *R*-factor. Panels (f)–(k) show three additional example structures taken from the best-matched sub-ensemble. The illustrated proteins are colored to match the spatial ranges shown in (c). The view in (d), (f), (h), and (j) is looking at a face of NCP_2. The view in (e), (g), (i), and (k) is looking at a side of NCP_2. Nucleosome array structures rendered using VMD.[65] [Color figure can be viewed at wileyonlinelibrary.com]

experimental data. This viewing method can be used to visualize the range of space occupied by different molecular domains. We demonstrate this for the tetranucleosome example, Fig. 9c, using separate colored regions to represent the range of each nucleosome. Supporting Information Fig. 18 further illustrates each of these separate regions for added clarity.

From these examples we obtained single structures and sub-ensembles of structures with SAXS profiles that reproduce the mock experimental data. We note that the number of relevant constraints inherent in a SAS profile is largely unknown, but it is generally accepted that the mathematical problem is largely underdetermined. Such highly underdetermined problems can have an infinite number of solutions.[63] Thus, one should be careful in choosing a single structure, or even a linear combination of structures, as a definitive solution to the problem. We submit that the plots of the spatial range of the best-matched structure ensembles provide a conservative method for representing the solution structures of molecules, as molecules in solution adopt a wide range of configurations.

Conclusions

We have developed a MC algorithm to rapidly generate ensembles of robust molecular structures of B-DNA. This algorithm represents flexible DNA using a CG bead–rod model, samples DNA bend and twist moves, then recovers an all-atom representation based on the final CG configuration. To relax strains of the O3′–P bond between adjacent bases, the final structures must be energy minimized. Our implementation of this algorithm for ensemble modeling of experimental data, rather than for equilibrium structure prediction, complements the underdetermined nature of SAS experiments. We emphasize the significant benefit this algorithm provides for modeling SAS experiments of nucleosomes, nucleosome arrays, and other similar systems.

This algorithm effectively generates structures spanning a wide range of configuration space in a rapid manner. This is demonstrated by example simulations of NCP and tetranucleosome structures. These examples also demonstrate the process

of filtering a structure ensemble against mock experimental data. We recognize the bead–rod model in this algorithm only allows for homogeneous DNA bending and therefore does not explore locally complex configurations, including the potentially significant possibility of DNA kinking and stretching.[64] Such eccentricity can be explored by performing MD simulations on select structures from the MC sub-ensemble best matched to experimental data.

Keywords: structural biology · deoxyribonucleic acid · molecular mechanics · Monte Carlo · computer modeling · small-angle scattering · neutron scattering · X-ray scattering

How to cite this article: S. C. Howell, X. Qiu, J. E. Curtis, *J. Comput. Chem.* **2016**, *37*, 2553–2563. DOI: 10.1002/jcc.24474

Additional Supporting Information may be found in the online version of this article.

[1] T. J. Richmond, C. A. Davey, *Nature* **2003**, *423*, 145.
[2] K. E. van Holde, Chromatin, Springer Series in Molecular Biology; Springer New York, **1989**; pp. 355–408.
[3] E. I. Campos, D. Reinberg, *Ann. Rev. Genet.* **2009**, *43*, 559.
[4] T. Schlick, J. Hayes, S. Grigoryev, *J. Biol. Chem.* **2012**, *287*, 5183.
[5] R. E. Franklin, R. G. Gosling, *Nature* **1953**, *171*, 740.
[6] M. H. F. Wilkins, A. R. Stokes, H. R. Wilson, *Nature* **1953**, *171*, 738.
[7] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, T. J. Richmond, *Nature* **1997**, *389*, 251.
[8] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, T. J. Richmond, *J. Mol. Biol.* **2002**, *319*, 1097.
[9] T. D. Frouws, S. C. Duda, T. J. Richmond, *Proceed. Nat. Acad. Sci.* **2016**, *113*, 1214.
[10] R. K. Suto, R. S. Edayathumangalam, C. L. White, C. Melander, J. M. Gottesfeld, P. B. Dervan, K. Luger, *J Mol Biol* **2003**, *326*, 371.
[11] T. Schalch, S. Duda, D. F. Sargent, T. J. Richmond, *Nature* **2005**, *436*, 138.
[12] F. Song, P. Chen, D. Sun, M. Wang, L. Dong, D. Liang, R. M. Xu, P. Zhu, G. Li, *Science* **2014**, *344*, 376.
[13] M. Hammermann, K. Tóth, C. Rodemer, W. Waldeck, R. P. May, J. Langowski, *Biophys J.* **2000**, *79*, 584.
[14] C. Yang, M. J. van der Woerd, U. M. Muthurajan, J. C. Hansen, K. Luger, *Nucleic Acids Res.* **2011**, *39*, 4122.
[15] K. Andresen, I. Jimenez-Useche, S. C. Howell, C. Yuan, X. Qiu, *PLoS One* **2013**, *8*, e78587.
[16] Y. Chen, J. M. Tokuda, T. Topping, J. L. Sutton, S. P. Meisburger, S. A. Pabit, L. M. Gloss, L. Pollack, *Nucleic Acids Res.* **2014**, *42*, 8767.
[17] S. Krueger, J. H. Shin, S. Raghunandan, J. E. Curtis, Z. Kelman, *Biophys. J.* **2011**, *101*, 2999.
[18] J. E. Curtis, S. Raghunandan, H. Nanda, S. Krueger, *Comput. Phys. Commun.* **2012**, *183*, 382.
[19] D. Norouzi, V. B. Zhurkin, *Biophys. J.* **2015**, *108*, 2591.
[20] D. Franke, D. I. Svergun, *J. Appl. Crystallogr.* **2009**, *42*, 342.
[21] D. I. Svergun, *Biophys. J.* **1999**, *76*, 2879.
[22] M. V. Petoukhov, D. I. Svergun, *Biophys. J.* **2005**, *89*, 1237.
[23] P. V. Konarev, M. V. Petoukhov, D. I. Svergun, *J. Appl. Crystallogr.* **2001**, *34*, 527.
[24] H. L. Tepper, G. A. Voth, *J. Chem. Phys.* **2005**, *122*, 124906.
[25] T. A. Knotts, IV, N. Rathore, D. C. Schwartz, J. J. de Pablo, *J. Chem. Phys.* **2007**, *126*, 084901.
[26] A. Savelyev, G. A. Papoian, *Proceed. Nat. Acad. Sci.* **2010**, *107*, 20340.
[27] P. D. Dans, A. Zeida, M. R. Machado, S. Pantano, *J. Chem. Theory Comput.* **2010**, *6*, 1711.
[28] A. Morriss-Andrews, J. Rottler, S. S. Plotkin, *J. Chem. Phys.* **2010**, *132*, 035105.
[29] M. Maciejczyk, A. Spasic, A. Liwo, H. A. Scheraga, *J. Comput. Chem.* **2010**, *31*, 1644.
[30] M. C. Linak, R. Tourdot, K. D. Dorfman, *J. Chem. Phys.* **2011**, *135*, 205102.
[31] Y. Wang, D. R. Tree, K. D. Dorfman, *Macromolecules* **2011**, *44*, 6594.
[32] T. Cragnolini, P. Derreumaux, S. Pasquali, *J. Phys. Chem. B* **2013**, *117*, 8047.
[33] Y. He, M. Maciejczyk, S. Ołdziej, H. A. Scheraga, A. Liwo, *Phys. Rev. Lett.* **2013**, *110*, 098101.
[34] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, J. J. de Pablo, *J. Chem. Phys.* **2013**, *139*, 144903.
[35] N. Korolev, D. Luo, A. P. Lyubartsev, L. Nordenskiöld, *Polymers* **2014**, *6*, 1655.
[36] J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman, S. J. Marrink, *J. Chem. Theory Comput.* **2015**, *11*, 3932.
[37] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Šulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, et al. *J. Chem. Phys.* **2015**, *142*, 234901.
[38] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, A. D. MacKerell, *J. Chem. Theory Comput.* **2012**, *8*, 348.
[39] N. Alegret, E. Santos, A. Rodríguez-Fortea, F. X. Rius, J. M. Poblet, *Chem Phys. Lett.* **2012**, *525–526*, 120.
[40] C. Sathe, A. Girdhar, J. P. Leburton, K. Schulten, *Nanotechnology* **2014**, *25*, 445105.
[41] A. Savelyev, A. D. MacKerell, *J. Comput. Chem.* **2014**, *35*, 1219.
[42] A. Garai, S. Saurabh, Y. Lansac, P. K. Maiti, *J. Phys. Chem. B* **2015**, *119*, 11146.
[43] M. Alishahi, R. Kamali, O. Abouali, *Eur. Phys. J. E* **2015**, *38*, 1.
[44] S. Bowerman, J. Wereszczynski, *Biophys. J.* **2016**, *110*, 327.
[45] J. Wang, H. Gao, *J. Chem. Phys.* **2005**, *123*, 084906.
[46] S. B. Smith, L. Finzi, C. Bustamante, *Science* **1992**, *258*, 1122.
[47] C. Bustamante, J. F. Marko, E. D. Siggia, S. Smith, *Science* **1994**, *265*, 1599.
[48] G. Maret, G. Weill, *Biopolymers* **1983**, *22*, 2727.
[49] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, V. B. Zhurkin, *Proceed. Nat. Acad. Sci.* **1998**, *95*, 11163.
[50] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X. J. Lu, S. Neidle, Z. Shakked, H. Suzuki, M. Tung, C. S. Westhof, E. Wolberger, C. Berman, M. Helen, *J. Mol. Biol.* **2001**, *313*, 229.
[51] D. Frenkel and B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, No. 1 in Computational Science Series, 2nd ed.; Academic Press: San Diego, **2002**.
[52] T. J. Thomas, V. A. Bloomfield, *Nucleic Acids Res.* **1983**, *11*, 1919.
[53] D. Schneidman-Duhovny, M. Hammel, A. Sali, *Nucleic Acids Res.* **2010**, *38*, W540.
[54] D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, A. Sali, *Biophys. J.* **2013**, *105*, 962.
[55] J. R. C. van der Maarel, Introduction to Biopolymer Physics; World Scientific: Hackensack, NJ, **2007**).
[56] J. E. Godfrey, H. Eisenberg, *Biophys. Chem.* **1976**, *5*, 301.
[57] G. A. Leonard, W. N. Hunter, *J. Mol. Biol.* **1993**, *234*, 198.
[58] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, *J. Comput. Chem.* **2005**, *26*, 1781.
[59] B. Schneider, S. Neidle, H. M. Berman, *Biopolymers* **1997**, *42*, 113.
[60] M. van Dijk, A. M. J. J. Bonvin, *Nucleic Acids Res.* **2009**, *37*, W235.
[61] Y. Liu, E. Haddadian, T. R. Sosnick, K. F. Freed, H. Gong, *Biophys. J.* **2013**, *105*, 1248.
[62] R. Buning, J. van Noort, *Biochimie* **2010**, *92*, 1729.
[63] G. H. Golub, C. F. V. Loan, Matrix Computations, 4th ed; JHU Press: Baltimore, MD, USA, **2012**.
[64] J. P. Peters, L. J. Maher, *Quarter. Rev. Biophys.* **2010**, *43*, 23.
[65] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graphics* **1996**, *14*, 33.

Received: 14 March 2016
Revised: 12 July 2016
Accepted: 23 July 2016
Published online in Wiley Online Library