TMB-iBIOMES: An iBIOMES-Lite Database of Nucleosome Trajectories and Meta-Analysis

Ran Sun,*† Zilong Li,*‡ and Thomas C. Bishop*¶

†College of Engineering and Science, Engineering Physics Program
‡College of Engineering and Science, Computational Analysis and Modeling Program
¶College of Engineering and Science, Departments of Chemistry and Physics

E-mail: rsu007@latech.edu; zli007@latech.edu; bishop@latech.edu
Phone: +1(318)-257-5209

Abstract

Nucleosomes are the fundamental building blocks of chromatin, the biomaterial that houses the genome in all higher organisms. A nucleosome consists of 145-147 base pairs of double stranded DNA wrapped 1.7 times around eight histones. There are approximately 160 atomic resolution structures of the nucleosome available from the protein data bank. Collectively they explore histone mutations, species variations, binding of drugs and ionic effects, but only a few sequences of DNA. Given a four-letter code (A, C, G, T), there are approximately $4^{147}$ or $10^{88}$ possible sequences of DNA that can form a nucleosome. Exhaustive studies are not possible. Here we introduce our TMB-iBIOMES database which serves as both a reference for future comparative, on-demand simulations of nucleosomes and a demonstration of iBIOMES-Lite as a tool for managing a laboratory’s simulation inventory. TMB-iBIOMES contains over 20 microseconds of all atom molecular dynamics simulations for over 500 different realizations of the nucleosome. For every simulation, the original input, output, de-watered trajectories, RMSD, and DNA helical parameter data are provided. Closely related simulations are grouped together, and a meta-analysis of each group is provided. The data can be navigated in a file browser format or downloaded directly with command line tools. Collectively the simulations provide a novel view of nucleosomal DNA. Compared to DNA free in solution, DNA on the nucleosome is not highly deformed or tightly restricted as determined by DNA helical parameter analysis. The overall conformation is restricted to a specific left-handed super helix, but the range of conformations explored by individual base pairs is larger than that observed for DNA free in solution.

Introduction

Chromatin is the biomaterial that contains eukaryotic genomes. The fundamental building block of chromatin is the nucleosome. All genomic mechanisms are therefore either directly or indirectly affected by the structure and dynamics of nucleosomes. There are over 7464 protein-DNA complexes in the protein data bank(https://www.rcsb.org/stats/growth/protein_na_complex). In eukaryotes, the proteins in these complexes either complement or compete with nucleosomes to achieve DNA binding. Of the available protein-DNA complexes, approximately 160 contain nucleosomes.¹ Collectively they represent less than five of the $4^{147}$ sequences of DNA that can be wrapped around a histone octamer. As a method, x-ray crystallographic results are fundamentally limited by symmetry of the DNA
sequence. To explore structure-function relationships governing genomic processes, methods for docking proteins to nucleosomes, workflows for computing thermodynamic ensembles, and tools for analyzing nucleosomes within the context of chromatin must be developed. Given the near infinite variations in the nucleosome based only on sequence, comparative methods are required. This approach differs from exhaustive studies such as the ABC effort to characterize all possible tetra-nucleotide steps by molecular dynamics simulation. We previously demonstrated that, with sufficient resources, comparative molecular dynamics studies of nucleosome can be computed overnight. We employed similar comparative techniques to study DNA kinking and mispositioning of the super strong nucleosome positioning sequence, 601 (in preparation). A recent study of 601 provides six simulations of nucleosomes with different sequences of DNA. Each simulation is over 2 microseconds (µs) long. With current computing resources and techniques, it is more efficient to compute ensembles or comparative simulations of nucleosomes rather than a single trajectory because parallel scaling provides limited returns beyond what can be achieved with one or two nodes. As next generation sequencing and Hi-C like methods enable genome wide association of function at or near base pair resolution, including the positioning of individual nucleosomes, there will be increasing demand to understand sequence specific variations in the structure and dynamics of individual nucleosomes. TMB-iBIOMES thus serves as both a reference for future comparative, on-demand simulations of nucleosomes and as a demonstration of iBIOMES-Lite as a tool for managing a laboratory’s simulation inventory.

The TMB-iBIOMES database at Louisiana Tech contains over 20 µs of all atom molecular dynamics simulations for over 500 different realizations of the nucleosome. The simulations include studies of positioning and mispositioning of nucleosomes for 23 unique sequences of DNA. The data is presented in two formats: as an iBIOMES-Lite database that is convenient for web browsing and obtaining summary data and as a file system tree for direct navigation and access to files. Here we provide a description of this iBIOMES-Lite database and a meta-analysis of the simulations.

**Methods**

**Modeling**

The simulations, experiments in iBIOMES nomenclature, represent seven types of systems as indicated in Table 1. All models except the systems labeled Nuc146 and Nuc147 were based on protein databank entry 1KX5. This structure is the highest resolution and most complete model of the nucleosome. Except for the systems labeled SIN, Nuc146 and Nuc147, the DNA fragment in 1KX5 has been replaced with DNA corresponding to a difference sequence. For some systems the histones are modeled with tails (WT) and for some systems there are no tails (NT). The systems labeled SIN, Nuc146 and Nuc147 included one or more point mutations or modifications to the histone core. Details regarding the choice of DNA sequence, the histone modifications, and the motivation are described below. The sequences utilized are graphically summarized in Figure 1.

**ACGT:** This set of 16 experiments serves as an experimental control. It is designed to represent all possible DNA sequence variants at each of the 146 dimer step locations in an octasome. The initial structure for each experiment was determined by removing the DNA from 1KX5, folding the desired sequence (AA)n, (AC)n, ... (TG)n, (TT)n into the 1KX5 super helix and docking it back onto the histone core by aligning phosphate atoms of the modeled DNA super helix to the phosphate atoms of the 1KX5 super helix. Additional information is provided by Mukherjee. All systems were solvated using the same initial box and simulated for 16 ns.

**Yeast POS and NFR:** This set of threading experiments corresponds to the mostly highly occupied and most well-defined nucleosomes
Table 1: Table for system summaries.

<table>
<thead>
<tr>
<th>System</th>
<th>Model</th>
<th>System Size (atoms)</th>
<th>Number of Unique Systems</th>
<th>Force Field</th>
<th>Water</th>
<th>Solvent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGT NT</td>
<td>1KX5</td>
<td>206,097-206,097</td>
<td>16</td>
<td>ff99bsc0</td>
<td>TIP3P</td>
<td>NaCl 12Å</td>
</tr>
<tr>
<td>Nuc146 FMH/WT</td>
<td>2CV5</td>
<td>266,038-266,083</td>
<td>3</td>
<td>ff14SB</td>
<td>SPC/E</td>
<td>KCl 15Å</td>
</tr>
<tr>
<td>Nuc147 FMH/WT</td>
<td>1KX5/2CV5</td>
<td>255,833-260,216</td>
<td>9</td>
<td>ff14SB</td>
<td>SPC/E</td>
<td>KCl 15Å</td>
</tr>
<tr>
<td>SIN M/WT</td>
<td>1KX5</td>
<td>387,061-387,089</td>
<td>6</td>
<td>ff14SB</td>
<td>TIP3P</td>
<td>NaCl 10Å</td>
</tr>
<tr>
<td>NFR NT/T</td>
<td>1KX5</td>
<td>157,550-158,506</td>
<td>105</td>
<td>ff14SB</td>
<td>TIP3P</td>
<td>NaCl 12Å</td>
</tr>
<tr>
<td>POS NT/T</td>
<td>1KX5</td>
<td>157,550-157,550</td>
<td>336</td>
<td>ff14SB</td>
<td>TIP3P</td>
<td>NaCl 12Å</td>
</tr>
<tr>
<td>NucA WT/T</td>
<td>1KX5</td>
<td>254,394-289,518</td>
<td>21</td>
<td>ff14SB</td>
<td>SPC/E</td>
<td>KCl 15Å</td>
</tr>
<tr>
<td>601 WT/T</td>
<td>1KX5</td>
<td>254,335-297,274</td>
<td>21</td>
<td>ff14SB</td>
<td>SPC/E</td>
<td>KCl 15Å</td>
</tr>
<tr>
<td>601 NT</td>
<td>1KX5</td>
<td>214,211</td>
<td>1</td>
<td>ff14SB</td>
<td>SPC/E</td>
<td>KCl 15Å</td>
</tr>
</tbody>
</table>

for each chromosome of *S. cerevisiae* denoted POS, along with nucleosomes corresponding to nucleosome free regions in chromosomes 2, 4, 5, 8 and 16, denoted NFR. In both cases a sequence corresponding to 147 base pairs was identified based on experimental data. The sequence was extended by 10 nucleotides in both the upstream and downstream directions. For each of the 21 sequences, 21 nucleosomes corresponding to 10 upstream, 10 downstream, and the central position of the selected sequence, Figure 2, were created using the DNA folding and docking scheme developed for ACGT. For these simulations two additional base pair GG and CC were added to each nucleosome to stabilize the DNA ends. These bases are not included in our helix parameter analysis. Each of the 441 unique nucleosomes was sampled for at least 20 ns by distributing the simulations across XSEDE resources.

**SIN**: The experiments labeled SIN-mutants were built manually and represent single-point mutations associated with disruption of the SWI/SNF-independent (SIN) complex, as described by Vijayalakshmi. The mutations correspond to histone H3 E105K, histone H3 R116H, histone H3 T118T, histone H4 R45H, and histone H4 V43I. The systems are labeled accordingly. An additional unmodified structure represents the wild type nucleosome (WT). All nucleosomes were modeled on 1KX5. As indicated in Table 1 an initial sampling of 20 consecutive nanoseconds was obtained for each system, and then 10 independent replicas of each system were utilized to accumulate an additional 150 ns (10 x 15ns) for each system.

**NUC147 and NUC146**: The experiments labeled NUC147 and NUC146 represent a set of simulations based on two different x-ray structures, 1KX5 and 2CV5. These systems are intended to determine if different initial conformations of nucleosomes will converge during standard molecular dynamics simulations. 2CV5 contains 146 base pair of DNA and 1KX5 contains 147. Protein data bank entry 1KX5 almost provides complete structure for the histone tails; however, the sequence does not correspond to wild type. Nonetheless the tails are unstructured. 2CV5 does not have histone tails. 2CV5 contains human histones. 1KX5 contains *xenopus laevis* (frog) histones. Using 1KX5 and 2CV5 as templates nine initial structures were generated. First 1KX5 was modified to be 100% consistent with the sequence of
Figure 1: Sequences used for model building. Graphical representation of the sequences used for each group of simulations. For the NFR and POS studies 147nt subsequences of the sequences indicated were used for model building. For all other studies the entire sequence indicated was used for model building. The ACGT, 147, and SIN studies contained 147 base pairs. The NucA and 601 studies contained 177 base pairs.

*xenopus* by adding missing amino acids to the tails as needed. The modified 1KX5 histone core was aligned with 2CV5’s core then the histone tails and 147 base pair super helix from 1KX5 were transferred onto the histone core of 2CV5. Docking a 147-base pair super helix onto 2CV5 was a deliberate mismatch since 2CV5 contains only 146 base pair. For each of these *xenopus*-based structures amino acid substitutions transforming the frog to human and mouse variants of the nucleosomal histones were introduced. The variants are denoted collectively as the FMH variants of Nuc147. The final set of systems are the FMH variants of 2CV5 with a 146 base pair super helix, denoted as FMH variants of Nuc146. For each of the 9 systems, 500 ns of continuous molecular dynamics have been computed to date.

**MMTV-NucA:** The experiments labeled NucA explore mispositioning of nucleosome A in the mouse mammary tumor virus promoter (MMTV). The MMTV is known to position six nucleosomes, referred to as nucleosomes A to F. Protein databank entry 5F99\(^23\) was released in 2016 and is a 2.63Å resolution structure of nucleosome A of the MMTV. 5F99 is the only x-ray structure available that represents nucleosome positioning for a naturally occurring DNA sequence. This structure was not available when we initiated a mispositioning study of nucleosomes A and B of the MMTV. The NucA systems were modeled by docking the relevant MMTV sequence to 1KX5 as described above. For NucA, 21 systems were created corresponding to 10 upstream, 10 downstream, and ideal positioning of NucA from the MMTV (Figure 2). The 5F99 x-ray structure thus provides a means of validating these simulation methods for the ideally positioned NucA; however, x-ray crystallography cannot provide insights into the alternate positioning of NucA as detected by early nucleosome positioning assays.\(^24\)
Figure 2: Modeling. Left: The four types of nucleosomes employed in the simulations: (i) contains 147 base pairs of DNA and eight histones with the tails truncated; (ii) contains 147 base pairs of DNA and eight histones with full tails; (iii) contains core and linker DNA and eight histones with the tails truncated; (iv) contains core and linker DNA and eight histones with full tails. Right: Graphical representation of the threading of DNA around a histone core to create mispositioned nucleosomes. Positions are numbered relative to ideal positioning at 0, indicated by the red rectangle. Positions -10 through +10 represent mispositioning from 10 base pairs upstream through 10 base pairs downstream of ideal positioning.

601: These experiments explore mispositioning of the super strong nucleosome positioning sequence known at 601. Two x-ray structures containing this sequence exist, namely 3LZ0 and 3LZ1. Both are 2.5Å resolution. Models of the ideally positioned 601 nucleosome and 20 mispositioned nucleosomes (10 upstream and 10 downstream, Figure 2) were built using the 1KX5 mouse model described above with the sequence known as 601-177. The systems were constructed with histone tails (WT) and with no histone tails (NT). Unlike the simulations of mispositioning based on yeast sequences which had only 151 (147+4) base pairs regardless of positioning, all 601 models included the entire length of 177 base pairs of DNA without any additional base pairs added for stability. The ideally positioned 601 contained two equal length linkers of 15 base pairs. Threading from positions -1 to -10 incrementally increased one linker up to 25 base pairs and decreased the other down to 5 base pairs. Threading to positions +1 to +10 did the opposite: incrementally decreased one linker to 5 base pairs and increased the other to 25 base pairs, see Figure 2. For the ideal position 1000 ns long simulations with and without tails are included in the database.

Simulations

Parameters: All of the simulations include explicit solvent (TIP3 or SPC/E) and ions (NaCl or KCl) with an approximate molar concentration of 150mM and initial solvent layer ranging from 10Å to 15Å depending on the study. The systems were all built with Amber’s tleap module using the most up to date modifications of the DNA force field available at the time, Table 2. In all cases NAMD2 was used as the compute engine with periodic boundary conditions, PME based long range electrostatics, a 2 fs time step, rigid constraints for all hydrogen atoms, Berendsen pressure regulation at 1.0 atm and Langevin temperature control at 300 K. Full simulation details are available from the protocol tab in the iBIOMES-Lite view or by downloading the NAMD log or configuration simulation files. Continuous simulation trajectories range from 16 ns to over 1 µs. All trajectories are computed in 1 ns simulation tasks. Several sets of independent parallel replicas are included for the statistical comparison of single versus multiple trajectory approaches. Each replica trajectory was computed as a series of 1 ns simulation tasks independent of all other replicas.
Table 2: Table for simulation parameters.

<table>
<thead>
<tr>
<th>System</th>
<th>Resource</th>
<th>CPUs</th>
<th>Simulations (numbers x ns)</th>
<th>Average time (ns/day)</th>
<th>Total duration (ns)</th>
<th>Volume of data(GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGT NT</td>
<td>LONI HP NAMD 2.6 mpi</td>
<td>64</td>
<td>16 x 16</td>
<td>1.6</td>
<td>256</td>
<td>126</td>
</tr>
<tr>
<td>Nuc146 FMH/WT</td>
<td>QB2 NAMD 2.10 ibverb cuda</td>
<td>180</td>
<td>3 x 499</td>
<td>12.8</td>
<td>1,497</td>
<td>506</td>
</tr>
<tr>
<td>Nuc147 FMH/WT</td>
<td>QB2 NAMD 2.10 ibverb cuda</td>
<td>180</td>
<td>9 x 499</td>
<td>13.5</td>
<td>2,994</td>
<td>1,015</td>
</tr>
<tr>
<td>SIN M/WT</td>
<td>QBI/Kraken NAMD2.9 ibverb/mpi</td>
<td>256</td>
<td>6 x 20 10* x 15</td>
<td>3.8</td>
<td>1,020</td>
<td>357</td>
</tr>
<tr>
<td>NFR NT/T</td>
<td>Lonestar NAMD 2.9 mpich</td>
<td>240</td>
<td>105 x 20</td>
<td>25.9</td>
<td>2,205</td>
<td>689</td>
</tr>
<tr>
<td>POS NT/T</td>
<td>Kraken NAMD2.9/mpi</td>
<td>64</td>
<td>336 x 20</td>
<td>5.9</td>
<td>6,786</td>
<td>3,000</td>
</tr>
<tr>
<td>NucA WT/T</td>
<td>QB2 NAMD 2.10 ibverb cuda</td>
<td>180</td>
<td>21 x 20 10* x 10</td>
<td>13.0</td>
<td>510</td>
<td>192</td>
</tr>
<tr>
<td>601 WT/T</td>
<td>Blue Waters NAMD 2.10 cuda</td>
<td>98</td>
<td>21 x 20 10 x 100 1 x 1000</td>
<td>15.1</td>
<td>4,078</td>
<td>1,500</td>
</tr>
<tr>
<td>601 NT</td>
<td>QB2 NAMD 2.10 ibverb cuda</td>
<td>180</td>
<td>1 x 1000</td>
<td>18.8</td>
<td>1,000</td>
<td>316</td>
</tr>
<tr>
<td>Summary</td>
<td></td>
<td>515</td>
<td></td>
<td>20,346</td>
<td>7,701</td>
<td></td>
</tr>
</tbody>
</table>

**Workflows:** Simulations were run on widely varying architectures including some with and some without GPU acceleration. Benchmarks on each system (not reported) indicate that parallel scaling efficiency dropped off significantly beyond 100-200 processors depending on the computer system. In most cases this was significantly less than the total number of processors available. The number of processors chosen for each computer system represented a balance between parallel scaling efficiency and throughput. Texas Advanced Computing Center’s Lonestar system provided the shortest time to completion of any single task. However, the simulations run on Blue Waters utilized only a small fraction of available resources. Thus, running multiple concurrent simulations on Blue Waters provided significantly higher daily throughput than any other system. The timings reported by Smith suggested by Smith\textsuperscript{7} jobs taking longer than 110% of predicted time to completion can be proactively killed and restarted with a high degree of confidence that it was a hardware glitch rather than intrinsic failure of the modeling. Caution must be exercised in identifying a successful simulation. This is fundamentally more complicated than identifying a successful computation. We use root mean square displacement (RMSD) and DNA helical parameter analysis as indicators of success of our simulations.

**Analysis**

**Post Processing:** All trajectories were de-watered, and the coordinates wrapped to the unit cell before root mean square (RMS) fitting all frames for a given system to the initial conformation. The trajectories in TMB-iBIOMES include periodic box information, ions, histones and DNA. All are available for downloading in DCD format. An Amber formatted parmtop file, nowat.parm, compatible with the de-watered trajectories is included in

6
each directory. Automated tools are recommended for downloading individual trajectories. Each nanosecond trajectory is approximately 300 MB but the entire collection of trajectories is over 6 TB.

**Conformational Analysis**: Two types of analyses were applied to all experiments. The metrics are chosen such that they can be used to compare any two simulations or any group of simulations regardless of which experiment the simulation is associated with. Two methods of comparison are employed: root mean square deviation (RMSD) and DNA helical parameter (HP) analysis.

**RMSD Analysis**: RMSD of the histone core, the DNA core, and the nucleosome core are computed to analyze structural stability of each system. For this purpose, the histone core is defined using residue numbering as in pdbid 1KX5: H3 (T45 to A135), H4 (N25 to G122), H2A (V27 to L96) and H2B (Y37 to K122). This selection is chosen to specifically ignore the mobility of the unstructured histone tails. Inclusion of tails raises the RMSD in a manner that is unpredictable and non-informative. The DNA core is defined as the central 146 or 147 base pairs directly in contact with the histone core. This allows comparison of systems with different lengths of DNA linker. The nucleosome core is defined as consisting of the histone and DNA cores. Again, this allows direct comparison of any simulation data since all simulations in the study are histone octamers. Comparison of FHM variants and SIN mutant is also allowed by these metrics since all mutants and variants correspond to single amino acid substitutions, rather than insertions or deletions.

**HP Analysis**: The DNA helical parameters provide a local description of DNA stacking (inter base pair parameters) and pairing (intra base pair parameters). This metric is a local measure of structure. Values from different simulations can be compared if the data is aligned to represent the same location along the nucleosome super helix. For the analysis presented here, only the 146 or 147 base pair in direct contact with the histone octamer are considered. Thus, helix parameter data from any systems containing 146 or 147 base pair can be compared to any other system containing 146 or 147 base pair, respectively. The NAS-TRUC utility, part of AmberTools package, is used to calculate the DNA helix parameter values. NAS-TRUC provides the same values as 3DNA and is based on El Hassan’s algorithm. Output is formatted in an HDF5 file that contains the inter and intra base pair helix parameter values. In all cases analysis began with an initial structure that allowed NASTRUC to identify proper base pairing. The raw NAS-TRUC outputs and HDF5 formatted data files are available for download.

**Publication**

All data is available in an iBIOMES-Lite and file browser format. The iBIOMES-Lite formatted data is presented in an informative, easily navigable display. The usual iBIOMES-Lite presentation has been modified to include links to the raw data. For each simulation the iBIOMES-Lite interface organizes the data as: Summary, Browse files, Execution info, and Protocol. The Summary tab provides a sample structure displayed and downloadable with JSmol, analysis data, a summary of dynamics, details of the molecular system and details of the computational tasks. The Browse files tab provides a collapsible view of files organized by type. The Execution info tab provides a summary of the computational tasks. The Protocol tab summarizes and displays all simulation variables.

The file browser presentation provides direct access to the simulation data in a similarly organized format. For every simulation the following files are available: initial Amber formatted parameter and topology files (*.parm) and coordinate files (*.crd) with and without solvent (sys.* and nowat.*, respectively), NAMD2 formatted configuration (*.conf) and output files (*.log), de-watered DCD trajectory files (1000
frames at 1 frame/ns in nowat.*.dcd), RMSD trajectory data (rmsd*.png and rsmd*.dat), and inter/intra base pair stacking and step parameters in raw NASTRUCT output format(*.nastruct.dat). The helix parameter data is also available in a self-describing hierarchical data format(hps.hd5). Closely related simulations are grouped together and contain a summary folder with analysis of DNA helical parameter and Cartesian coordinate RMSD.

iBIOMES-Lite provides tools for managing all post processing tasks. The tools include consistency checks to update analysis when new analysis scripts or trajectory data is entered into the database. For the purposes of maintaining and sharing a laboratory’s simulation inventory iBIOMES-Lite has proven sufficient. The complete iBIOMES solution provides additional capabilities for data inventory, sharing and distribution.

Results

All available RMSD and helix parameter data is reduced to three summary graphs, Figure 3, Figure 4, and Figure 5. Histone core and DNA core RMSD are interpreted below as a measure of nucleosome stability during the simulations. The range and mean helix parameter values across all simulations are interpreted as a measure of local conformational freedom or constraint on the nucleosome super helix.

Histone Core RMSD: All simulations except the SIN mutants and 601 NT (not shown) resulted in mean RMSD values for the histone core below 2.5Å. The highest RMSD values observed for the histone core for any system were for 601 NT(1000 ns) the mean RMSD was 3.3Å. This system is discussed in greater detail in the novel observations section.

For the SIN-mutants the range of RMSD values is 2.4 to 3.0Å for the individual replicas, con-
sistently higher than for all other simulations. For H4-V43I the RMSD values range from 2.4 to 2.5Å (in the lowest cluster) and H3-E105K exhibits the highest range of RMSD values 2.9 to 3.0Å (in the highest cluster). The WT SIN-mutant simulation exhibits intermediate RMSD values (2.6 to 2.7Å). The larger deviations observed for these systems may reflect the known destabilizing effect of the SIN mutations or differences in assembly or simulation protocols. The SIN mutants are the only systems assembled outside of our lab and simulation protocol utilized a 20 ns continuous trajectory followed by independent replicas. The fact that the WT SIN-mutant (i.e. unmodified) core RMSD values are higher than all NUC147 RMSD values suggests differences are due to assembly protocols. While caution should be exercised in comparing the SIN mutant simulations to others in the database, we expect that comparing SIN mutants to SIN mutations is valid.

POS and NFR used identical build and simulation protocols. The only differences between the 336 POS simulations and the 105 NFR simulations are the DNA fragments docked to the histone core. The histone RMSD values for NFR are consistently lower than for POS suggesting that variations in DNA material properties can alter the structure of the histone core even on the 20 ns timescale. The magnitude of the effect is related to the level of sequence specific positioning associated with the fragment. The NucA and 601 simulations also utilized identical build and simulation protocols and provide further support for this hypothesis. The NucA sequence is not a super strong positioning sequence; 601 is. The RMSD values observed for the histone core in NucA are lower than for the 601 simulations.

Core DNA RMSD: In all cases the RMSD values for the core DNA are higher than for the histone core. Since DNA wraps around the outside of histones, it can act independently of histone core. If structural transitions of the histone core occur then the DNA must also somehow adapt. In either case, we expect DNA RMSD to be at least as high as that observed for histone core, and this is the general trend.

DNA RMSD is used as a measure of nucleosome super helix stability as a function of DNA sequence. The expectation in the threading simulations was that DNA RMSD would correlate with mispositioning. However, DNA RMSD for the ideally positioned systems was not consistently higher or lower than the mispositioned systems associated with the same sequence (see individual experiment plots). Interestingly, the range of values associated with yeast positioning sequences (POS) is significantly higher than the range of values observed for the nucleosome free regions (NFR). Since these systems utilized identical build and simulate protocols this is a strong suggestion that sequence dependent material properties of DNA affect its behavior even on the 20 ns time of these particular simulations.

Simulation time alone is not a distinguishing factor. Specifically the RMSD for the DNA in the 500 ns long Nuc147 simulations spans the same range as the 20 ns ACGT simulations. The highest DNA RMSD observed for Nuc147 is for 2CV5-Human. In this system the structure of DNA from 1KX5 was docked onto 2CV5. The actual 2CV5 structure contains only 146 base pairs. Given this mismatch the DNA RMSD is expected to be high, but it is not higher than values associated with NFR or POS simulations. POS and NFR used similar assembly and simulation protocols, but contain different core DNA sequences. For the Nuc146 systems (500 ns) the DNA RMSD varies significantly considering that differences in these structures are limited to modifications of the histones (FMH variants) rather than DNA. The range is similar to the range observed in SIN mutants (150 ns) in which single point mutations were introduced into the histone core.

The systems labeled NFR (20 ns), POS (20 ns), NucA (20 ns) and 601 (100 ns) all included mispositioned nucleosomes. NFR and POS used the same build and simulation protocols. NucA and 601 used the same build and simulation protocols. Both the POS and
Figure 4: Range of the DNA Helix Parameters. Left represents the inter base-pair helical parameters. Right represents the intra base-pairs helical parameters. The blue line represents the max values and the green line represents the min values of individual helical parameter for all simulations in database. The three red dotted line (from top to bottom) represent the max, min and average values of free DNA.

601 simulations exhibit a large range of RMSD values. POS includes 336 simulations representing the threading of 16 unique positioning sequences of biologic origin while 601 includes only 21 simulations representing the threading of a single, artificial, super strong positioning sequence. One should expect mispositioning of these sequences to disrupt the DNA super helix and yield a wide range of RMSD values. 601 clearly demonstrates this effect. In further support of the idea that DNA RMSD correlates with positioning ability we note that the range of RMSD values observed for NFR is less than POS and for NucA the range is less than 601. Here, we have only compared simulations that use the same build and simulate protocols.

In case of NucA, inspection of the raw data suggests that an RMSD of 2.5Å can be used to cluster the simulations into two groups. All NucA replicas representing ideal positioning fall under the 2.5Å cut-off. Only replicas representing mispositioned nucleosomes are in the above 2.5Å group.

In case of 601 the lowest RMSD value is obtained for position -1 (2.5Å) rather than the ideal position (3.0Å). Mispositioning of 601 by greater than ±5 base pairs yields a DNA RMSD greater than 3.5Å. Again ideal and near ideal positions are associated with low RMSD values (< 3.3Å in this case), while any RMSD value above 3.5Å is a mispositioned nucleosome. For 601, behavior of linker DNA also plays a role in the high DNA RMSD values, see Novel Observations below.

The ACGT, NFR and POS simulations utilized very different sequences of DNA and similar, but not identical, build and simulation protocols. The range of DNA RMSD values observed in the ACGT simulations is considerably lower than NFR and POS. The ACGT sequences are simple di-nucleotide repeats, suggesting that sequence complexity also plays a role in the stability of DNA on the histone octamer, i.e. the materials properties of DNA effecting nucleosome positioning extend beyond di-nucleotide length scale.

Helix Parameters: We consider the range of helix parameter values observed at any location along the nucleosome super helix regardless of sequence and regardless of the simulation from which the values were obtained, see Figure 4 and Figure 5. Intra and inter base pair helical parameters results are graphed as two sets of six-plots with location reported relative to the nucleosome dyad. The reported range of values represents over 20 µs of nucleosome sim-
ulations and is compared to the range of values observed for DNA free in solution from a similarly large set of simulations.\(^6\) What emerges from this comparison is that rather than being tightly constrained, the pairing and stacking of bases in the nucleosome appears to explore the same range of conformational space as does DNA free in solution. In most instances the range of values observed for nucleosomal DNA is actually larger than for free DNA. The exceptions are the inter base pair helical parameters Slide and Rise and the intra base pair helical parameters Shear, Stretch and Stagger. All exhibit tendencies to be more restricted that free DNA, but only at some locations along the super helix. There are no easily discernible patterns as to where these restrictions occur. The observations are consistent with a stress release mechanism in which transient kinks or other large scale deformations at one or more sites allow other sites to relax or adhere to relatively strict conformational tolerances.\(^39\) For the ACGT simulations we demonstrated that such kinking is dynamic on the nanosecond timescale.\(^8\)

Novel Observations: Something unique occurred during the 601 NT simulation(1000 ns). An inspection of the trajectory shows that within 200 ns the DNA linkers on opposite sides of the nucleosome line up between periodic cells to yield near native inter-nucleosome base pair stacking, Figure 6. These images were obtained from the unmodified trajectories. The underlying cause of the anomalously large RMSD for the histone core observed in this simulation appears to arise from linker-linker interactions between periodic images. The linker-linker interactions are dynamic on a microsecond timescale. They form during the first 200 ns and explore near native stacking conformations until eventually repelling each other from 500 ns to 1000 ns. We also observe linker-linker interactions in the 601 WT (100 ns) at positions +5 and −5. However, rather than interactions across periodic boundaries, the long linker arm loops back for self-interaction with the shorter linker arm within the periodic cell. Both simulations contain 177 base pairs, thus 30 base pairs is sufficient for self-linkage interactions. These linker-linker interactions depend not only on the length of DNA but sequence dependent bending of DNA and, in case of 601 NT, the periodic boundary conditions. The simulations clearly demonstrate that the level of force associated with inter-nucleosome linker-linker interactions is sufficient to disrupt even the histone core structure. The disruption strongly depends on the direction in which these forces are applied and rescaling of the periodic box that may occur. However, the magnitude and temporal properties of the forces arising in our simulations are intrinsic to the nucleosome interactions and thus fundamentally differ from external forces that arise in either pulling simulations\(^41\) or AFM/tweezer experiments.\(^42,43\)
Conclusion

Modeling: The iBIOMES-Lite tools for data sharing and publication are easy to configure and deploy tools for managing large simulation data. For our particular implementation we side-stepped many of the infrastructure requirements (i-RODS and apache tomcat) of the full iBIOMES implementation by providing links within the iBIOMES html directly to the data on disk. The data on disk is organized in a tree structure that is familiar to any practicing molecular modeler and can be conveniently downloaded with various automated tools for further analysis and review. Since the resources available at today’s supercomputing centers far exceed the parallel scaling limits of medium size systems (∼500,000 atoms) tools such as iBIOMES-Lite provide a means of effectively managing analysis and sharing results from comparative studies. Coupling entire model, simulate and analyze workflows with AI or informatics systems provides novel opportunities for the integration of informatics and physics to explore structure-function relationships.

Our computations demonstrate that comparative studies of DNA sequence specific variations in nucleosome structure and dynamics on the nano to micro second time scale can be achieved with widely available computing resources. Even a desktop workstation can compute over 10 ns of nucleosome dynamics in a single day. Atomic details of the variations in the stacking and pairing of bases that affect the binding of proteins and drugs to DNA can be observed on this time scale. From this point of view, comparative molecular modeling and dynamics is an ideal complement to Next Generation Sequencing and nucleosome positioning techniques that can rapidly hone in on a limited region of interest of any genome. With advanced sampling protocols, suitable workflows, and access to medium scale supercomputing resources, comparative modeling can be integrated directly into bioinformatics pipelines to directly probe structure-function relationships in atomic detail. Our TMB-iBIOMES dataset serves as a validation suite for the development of more efficient sampling protocols, and our genome dashboard, G-Dash, provides the missing link to bioinformatics workflows.

Biology: Our simulation of vastly different sequences of DNA docked onto a histone octamer suggests a novel view of nucleosomal DNA. Even within the limitations of our modeling approach, namely inaccuracies due to sampling and force field deficiencies, it is clear that nucleosomal DNA is remarkably similar to DNA free in solution. By this we mean the
required variations in structure associated with super helix formation are well within the range of conformations associated with thermal fluctuations of free DNA. Moreover, comparing the time variation of pairing and stacking of a single base pair or base pair step to the variation of stacking and pairing as a function of location along the nucleosome super helix we find that time variations are larger than variations associated with location. This is true even on the 10-100 ns time scale. Nucleosomal DNA forms a characteristic super helix, but it is not tightly wrapped onto or tightly attached to the histone core.

We have recently updated our tunable metric for DNA kinking and applied it to the threading simulations. Detailed analysis will be reported elsewhere. However it is clear in our studies of nucleosome positioning and mis-positioning that as DNA is thread around the histone core, the conformation and dynamics at the base pair step level falls into two categories. In one category conformation and dynamics are determined by sequence. Thus observations of DNA kinking, tend to shift as the DNA is thread around the octamer core. In the other category location determines the conformation and dynamics. Thus observations of DNA kinking do not simply shift as the DNA is thread around the core.

The histone core is not merely a rigid three dimensional object to which DNA must passively conform. Rather the material properties of DNA, expressed by its sequence, affect the conformation and dynamics of nucleosomal DNA and, to some extent, even the structure and dynamics of the histone octamer. It is for this reason that rules for sequence based nucleosome positioning, beyond those proposed by Widom, have proven so elusive. Since protein-DNA and drug-DNA interactions are often sequence specific, great care must be exercised in selection of a coarse-grained potential for chromatin modeling, if one seeks to understand how protein-DNA and drug-DNA interactions are scaled-up by the nucleosome to affect chromatin structure and dynamics and ultimately achieve a specific biologic outcome.

Acknowledgement This effort was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number 5 P20 GM103424-15 and 3 P20 GM103424-15S1. Partially sup-
port also received from NSF through cooperative agreement OIA-1541079 and the Louisiana Board of Regents.

References


(46) Li, Z.; Sun, R.; Bishop, T. C. G-Dash: A Genome Dashboard Integrating Modeling and Informatics. 2018,