A Parallel Molecular Modeling Framework to Assess DNA Sequence Effects on Nucleosome Stability

By

Alexander S. Garrett

Submitted to the Department of Electrical Engineering and Computer Science
and the Faculty of the Graduate School of the University of Kansas
In partial fulfillment of the requirements for the degree of Master's of Science

____
Dr. Terry Clark: Chairperson

Committee members

____
Dr. Xue-wen Chen

____
Dr. Victor Frost

____
Dr. Krzysztof Kuczera

Date defended: ____________________

The Thesis Committee for Alexander S. Garrett certifies
That this is the approved version of the following thesis:

A Parallel Molecular Modeling Framework to Assess DNA Sequence Effects on Nucleosome Stability

Committee:

____
Chairperson

____

____

____

Date approved: ______________________

Abstract

Central to eukaryotic cellular support such as transcription regulation and chromatin structure are DNA interactions with the nucleosome histone core. The interactions of the histone core with DNA are currently understood partially at best, due in part to the wide range of DNA sequence effects involved in the formation of stable nucleosomes. This paper presents a high-performance computational modeling framework designed to enable greater understanding of DNA sequence effects on nucleosome stability through the automated generation of new DNA-histone configurations. The presented framework includes modules for the mutation of DNA sequences, use of those sequences in molecular simulation, and subsequent energy calculations. Parallel computing enables the configuration of a large number of nucleosome structures and modules are customizable for use in studies such as molecular dynamics. This paper discusses a prototype study using energy minimization to evaluate the high-throughput framework and address the computational feasibility and application to future studies of DNA sequence effects on nucleosome stability.

Contents

List of Figures

Chapter 1

Introduction

All living organisms belong in either one of two superkingdoms: the prokaryotes such as bacteria, and the eukaryotes, comprising all animals, plants, and fungi on the planet. One important feature of the eukaryotic empire is the presence of bundled genetic material in the form of chromosomes. One chromosome is amazingly just a single molecule of double-stranded deoxyribonucleic acid (DNA), organized beautifully with management proteins forming the structure of chromatin. The enormous system of genetics necessary for life is managed by a fundamental chromatin packaging unit of 146 DNA base pairs with a cylindrical protein complex. This particle is known as the nucleosome.

Nucleosomes are the central database administrators for every archived instruction in DNA. The genes for life and elements needed for the regulation, organization, replication, and packaging of those genes are controlled by the delicate role of nucleosomal dynamics. Central to the importance of the nucleosome is its role in the condensation of the molecule of DNA to fit inside the cell nucleus and regulation of transcription events that interpret and process genes. Regulations of transcription events for protein encoding depend on the position the nucleosomes form on the DNA sequence. The interaction of nucleosomal protein and DNA is active and fluid during the life of a cell, where by environmental conditions, DNA sequence effects, and the amino-acid composition of the constituent protein chains all affect the placement and stability of the nucleosome. A deeply interesting factor affecting nucleosome stability is the role that DNA sequence plays in the interaction.

Nucleosome stability is a function of the ability for DNA to bind favorably with the protein complex counterpart. This ability is related to positions of the DNA sequence that allow for bendability in nucleosome formation. Specifics concerning the sequence features necessary to bind nucleosomes steadily emerge from the experimental community; however, a clear energetic understanding of these specifics has not yet fully materialized, nor has a consensus DNA sequence been discovered.

The purpose of the computational framework presented is to assess DNA sequence effects on nucleosome stability by way of high-throughput assessment of computationally engineered DNA sequences using molecular simulations of the nucleosome structure. Design of the framework was done with high-performance and experimental control in mind. The framework shows promise in enabling researchers an automated capability for assessing many DNA sequences to model DNA sequence effects on nucleosome stability not currently understood.

The rest of this paper is as follows. The problem statement of this thesis work is stated in Chapter 2. Background information is Chapter 3 including a discussion of the nucleosome, DNA, and computational molecular modeling. Chapter 4 introduces related work to concepts of this thesis from scientists in the field. Design features are discussed in detail in Chapter 5 relating to the system built to configure new DNA sequences to assess their sequence effects on nucleosome stability. The design discussion includes information on the configuration generator and control system enabling high-throughput studies of many DNA sequences. The configuration generator consists of three main subsystems: DNA mutation, simulation, and energy retrieval. Chapter 6 outlines a proof-of-concept study highlighting the capabilities of the engineered framework. An overview of the experimental issues relating to the prototype study is given with framework specifications that generated the presented results. A discussion of the prototype study and its implications are laid out in Chapter 7 with an emphasis in performance and potential application directions. Comments on the overall framework's extensibility and future work are concluded in Chapter 8.

Chapter 2

Statement of Problem

Nucleosome formation potential is a crucial factor behind chromatin structure, with pivotal roles in gene regulation and replication processes at the centromere. A vast range of DNA sequences are considered important to nucleosome formation. These are studied in experimental work on varying scales and in computational studies involving one to few DNA types. However, it is not possible presently to study routinely suites of DNA in high fidelity models to determine potentially crucial molecular details. Because of this, DNA sequence effects of interest to various molecular biology studies, such as disease states of chromatin, are often deduced from a limited collection of published nucleosome stability data.

Although detailed molecular modeling is conducted on nucleosomes, it is performed on typically one DNA sequence. In principle, large scale studies are realizable, but there is much work to be done in model and infrastructure development. The reported work is an implementation of current modeling technologies in a high-throughput framework to explore the engineering and scientific challenges in this area. These include addressing the computational overheads associated with detailed modeling methods and assessment of a large structure space focused on the problem of sequence effects on DNA curvature on nucleosomes.

This thesis introduces a computational framework to model nucleosome stability involving large suites of DNA sequences. A preliminary database is developed with energy and structural analyses of systematically modified DNA modeled on the nucleosome core. The work develops the methodology and applies it using a test suite of DNA sequences.

Assessments are made about the computational feasibility of the substitution method using other modeling approaches, based on our prototype studies. These projections are geared to assist in the design of further studies.

The framework supports the computation of DNA sequence affinities using molecular modeling modules. In the prototype, energy minimization is used. The system provides for the definition of a suite of DNA molecules, which are automatically generated and modeled. Molecular modeling and analysis is based on the recently solved 1.9 Å Xenopus laevis nucleosome crystal structure. The framework presented addresses computation requirements with parallel computing and by reducing processing requirements on the large number of DNA sequences through incremental nucleotide substitutions. The framework implements a heuristic for ordering the substitutions with stepwise energy relaxation on substituted DNA to minimize perturbations to the base structure.

Analysis of sequence effects on nucleosome stability is performed for a substitution suite consisting of 4096 DNA sequences computationally engineered in a way that is widely thought to impart curvature to the DNA, which is associated with stable nucleosomes. This benchmark suite, which characterizes a widely studied motif found to underlie stable nucleosomes, allows us to assess the prototype framework and compare our simulation results with published data, and other models; for example, we evaluate our sequence suite using a bendability model. These studies provide feedback to develop our modeling framework, test and validate the system, and establish a preliminary database of nucleosome structures for further assessment.

Chapter 3

Background

3.1 Nucleosome

Genetic material in eukaryotes organizes into the nucleoprotein complex chromatin. Chromatin is the central structure used by the cell to package DNA. Genetic processes, both vital such as transcription and replication, and pathological such as cancer and viral infection, depend on the DNA-protein complex comprising chromatin (Davey et al., 2002). Nucleosomes are the interactive building blocks in the structure and dynamics of chromatin; they are the fundamental, repeating unit of protein and its associated DNA.

Nucleosomes are ubiquitous appearing at roughly periodic intervals of 200 DNA base pairs in bulk genomic DNA. The central protein complex is composed of eight histone protein chains: a tetramer of (H3)2(H4)2, and two flanking H2A-H2B dimers. The globular octamer is cylindrical in shape, shown in Figure 3.1 with corresponding DNA.

Figure 3.1 The Nucleosome of Xenopus laevis (Davey et al., 2002). Eight histone chains form a cylindrical protein core wrapped by ~146 DNA base pairs.

Each histone chain is organized into two domains: a central fold that contacts the DNA superhelix, contributing to the compact core of the nucleosome; and amino-terminal tails which extend beyond the core (Kornberg et al., 1999). About 146 DNA base pairs associate directly with the central octamer complex to form the nucleosome, while a linker protein and roughly 60 DNA base pairs connect neighboring nucleosomes to form higher order chromatin structures.

The histone chains comprising the nucleosome core protein are some of the most conserved in nature (Kimball 2006). Arginine side chains in 12 of 14 DNA-histone binding positions are absolutely conserved among the major type histones of all species (Muthurajan et al., 2003). The histone chain H4 in the calf differs from H4 in the pea plant by only two amino acid residues out of the chain’s 102 amino acids. The evolutionary conservation across all eukaryotes conveniently permits the study and understanding of the general biological significance of nucleosomes by looking at a specific one, such as the African clawed frog's nucleosome. The conservation also allows greater emphasis on DNA sequence effects in nucleosome interactions by mitigating the complexity introduced by differences in histones. The reluctance for large histone mutation across species exploits the general and fundamental characteristics of chromatin regulation in cellular replication and transcription.

Modulation of the packaging of DNA is the essential function of chromatin and thus the nucleosome. Condensation and relaxation of the DNA through affiliated proteins serves two main purposes. First, replication of the genetic material undergoes considerable packing of chromatin during prophase of mitosis; proper chromosomal division requires it. For the human diploid chromosome count of 46, the unpackaged DNA length would be over two meters. The ultimate size manageable for replication is just ten micrometers, for all of the hundreds of trillions of reproducing cells in humans (Kimball 2006). The second function of regulating the state of DNA in chromatin structures relates to transcription events, essential to the order and function of life's gentle balance.

Transcriptional regulation via the nucleosome has a functional relationship to sequence binding positions. Sequence factors left unbound in the nucleosome could contain transcription factors for the promotion or enhancement of genes or other elements. A preferentially poor binding sequence positions itself in a more accessible state, simply by eluding nucleosomal protein attraction. Proteins can locate unbound promoting regions and systematically dissociate neighboring nucleosomal DNA for processing. Positions of DNA sequence in the bound state could be recognized by proteins evolved to locate such sequences in the bound state. Nucleosomes can repress transcription by preventing the initiation of RNA polymerases, and thus preventing access to genes hidden in the bound configuration. Acetylation of histone tails allows for the dissociation of nucleosomal DNA for transcription events. Gene regulation depends on the local environmental presence of histone acetyltranferase and deacetyltranferases which usually promote and repress transcription, respectively.

The model nucleosome system presented in this work is PDB code 1KX5, the x-ray crystal structure of the Xenopus laevis nucleosome resolved at 1.9 angstroms (Davey et al., 2002). The resolution of this state-of-the-art structure is specific enough to coordinate hydrogen atoms, ions, and solvent surrounding the DNA and octamer complex. Computational studies are thus permitted with the use of this model system and all-atom force fields in molecular simulation.

3.2 DNA

Deoxyribonucleic Acid (DNA) is the genetic material of life. Contained between its phosphodiester backbones are nucleotide pairs that form the instructions for every gene in an organism. DNA influences nearly every process in cells: its development, function, reproduction, and even death; epigenetic processes coupled with environment appear to take the DNA blueprints in the multitude of life's diverse directions. The polymer structure of DNA allows it to become very long, one of the largest molecules found in nature. The molecule retains its integrity while dynamically interacting with proteins to organize, transcribe, and replicate it. In terms of interacting with the nucleosome core protein, DNA has a number of important characteristics.

Hydrogen bonds and electrostatic forces between the highly negative backbones of DNA and the positively charged amino acids of the histone chains mediate DNA-histone interactions. The effects of specific nucleotides do not directly relate to binding affinity because only the backbone contacts with the protein core. All DNA has a degree of attractive nature to the core protein, some having a relatively higher binding affinity than others. Paradoxically, the nucleosome must form with most all sequences of DNA for chromatin folding and function, but in a way to permit eventual unfolding to expose the DNA.

Central to the principle of nucleosome positioning is energy minima, where by the DNA and protein core maintain a bound state. Positioning of nucleosomes on ~146 DNA base pairs seems to occur preferentially at regions conferring higher binding affinity. DNA and nucleosomes likely evolved together, so that minor sequence features could favorably bind nucleosomes without tremendous effect of the instructions held within the DNA code. Periodicities in DNA sequences are consistent with nucleosome binding principles, particularly in regions known to be repetitive, non-coding, satellite regions of chromosomes.

Sequence specificity relies on one seemingly dominant feature of DNA: anisotropic deformability, in this text called bendability. Numerous derivations in the literature address the primary role of sequence bendability, as discussed in Chapter 4. The most prominent enablers of bendability are specific nucleotide occurrences permitting the left-handed curvature direction of nucleosome binding. These bending features specifically refer to the compression side of DNA on the interior minor groove close to the protein core and the expansion side where the minor groove faces away from the protein. Sequence areas of poor bendability and curvature reflect positions with reduced likelihood for nucleosome formation.

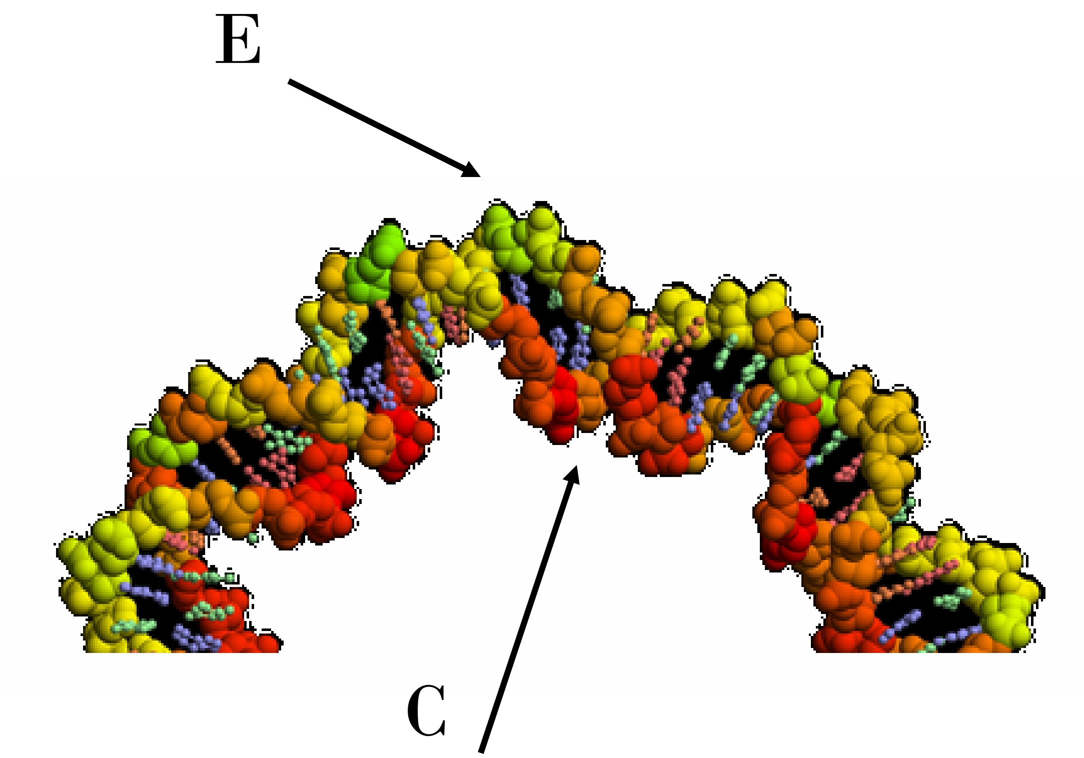

Figure 3.2 DNA Curvature. Bendable steps along every twist of the minor groove (10.3 bp). 'E' labels the expansion site where bendability may favor sequences containing GCC. 'C' shows the compression site which may favor AAA.

Two sequence characteristics explored in this work are particular tri-nucleotide occurrences at strategic locations on the DNA to permit optimal curvature (Figure 3.2). Of these studied are AAA on the compression side and GCC on the expansion side of nucleosomal DNA's minor groove.

A well-curved or bendable sequence in the proper orientation is mechanically more inclined to undergo wrapping a nucleosome core protein, and is more likely to find itself positioned accordingly. DNA curvature and DNA-histone contacts are two principal nucleosome stabilizing factors. Curvature similarities between free and nucleosome bound DNA suggest that DNA bendability can be a major factor in nucleosome stability (Drew & Travers, 1986). The large free energies for bending DNA on the nucleosome in the range of about 75 kcal/mol (Lowary & Widom, 1997) may be reduced by bendable DNA (Shrader & Crothers, 1989, Lowary & Widom, 1997).

3.3 Computational Molecular Modeling

One essential component of the framework presented is the usefulness and importance of computational molecular modeling. This relatively new form of scientific discovery pertains to the computational visualization, analysis, and simulation of molecular systems. Molecular modeling tools give researchers unique control and understanding of the many interesting portions of their studied systems. Of the many fascinating and ever expanding roles computers play within biology, chemistry, and physics, certain key modeling technologies have been kept in mind during the architecture of the presented software system. The most fundamental of these technologies being molecular dynamics and energy minimization engines, molecular topologies and force fields, biopolymer visualization and mutation software, and tools used to enhance the biological coherence of molecular systems removed from their native environments. Parallel and supercomputing technologies extend other modeling technologies into exciting realms.

Among the most advanced and technically profound molecular modeling tools available are those used in molecular simulation. Simulation is this context is the experimentation of molecules using the computer. Molecular dynamics (MD) simulation relates to the technique of evolving the conformation of a set of interacting atoms by integrating Newton's equations of motion. Provided to the molecular dynamics engine are an initial set of coordinates for all atoms in the molecular system, definitions of the way these atoms behave according to their respective chemical properties, and instructions for evolving the system in time. At each time step equations of motion are calculated based on repulsive and attractive forces that each atom has with its neighbors and the overall electrostatic environment. MD simulations have been used extensively in academic and industrial settings to test the conformational dynamics of interesting molecular systems. By simulating the molecular system long enough, statistical principles can be used to infer the overall mechanics of the underlying system, providing valuable insight into thermodynamics and kinetics, for example. Energy minimization studies do not model dynamic motion in the system but, instead, the conformation is optimized to a local minimum through the same principles of force calculation and integration. This method is used, for example, in rapid screening performed for lead drug compounds. In the case of nucleosome studies, both molecular dynamics and energy minimization can be a valuable tool in determining interaction energies in order to compare different sequence effects on nucleosome stability.

Interaction energetics for molecular systems are based on molecular dynamics force fields as the summation of many various energy terms.

$$
\begin{aligned}
E \; = \; & \sum_{\mathrm{bonds}} k_b(r - r_0)^2 + \sum_{\mathrm{angles}} k_\theta(\theta - \theta_0)^2 + \\
& \sum_{\mathrm{proper\ dihedrals}} |k_\phi| - k_\phi cos(n\phi) + \sum_{\mathrm{improper\ dihedrals}} k_\omega(\omega - \omega_0)^2 + \\
& \sum_{pairs, i \neq j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{e r_{ij}} \right] + \\
& \sum \left[\frac{A'}{r_{AD}^{i}} - \frac{B'}{r_{AD}^{i}} \right] cos^m(\theta_{A-H-D}) x cos^n(\theta_{AA-A-H}) + \\
& \sum K_i(r_i - r_{i0})^2 + \sum K_i(\phi_i - \phi_{i0})^2
\end{aligned}
$$

Figure 3.3 CHARMM Force Field (MacKerell, 1995). The value used in the calculation of interacting forces is the summation of many various energy terms.

Force fields are parameterized by experiment and semi-empirical computations. The physical correctness of the simulation is thus directly coupled to the quality of the force field chosen to describe the behavior of atomic interactions. The force fields are used to provide meaningful, chemically relevant information to the simulator for hopeful extrapolation into the corresponding natural environment.

Additional modeling tools permit manipulation of a molecular system in various ways. One tool used in the framework identifies the average direction of the DNA backbone orientation. When a replacement or computational mutation of a DNA nucleotide is requested, the tool switches out the old nucleotide with the new one, maintaining the original directionality of the phosphodiester backbone. Subsequent energy minimizations are used to relax deformations in the structure introduced by this mutation. Other tools permit the solvation and ionization of macromolecules to simulate a biological environment. These tools can calculate electrostatic potential gradients to position counter-ions to neutralize the system's net charge.

The world of molecular modeling grows and matures daily, and as computing power intensifies, the applications for molecular modeling techniques will grow and mature along side. Parallel computing is one such intensification of computing power. An assembly of parallel workstations, or cluster, is capable of undertaking computations simultaneously by decomposing the computation into its constitutive components. In the domain of molecular simulation, the foundational component is the formulation and calculation of each atom's neighboring atoms and their respective forces. Pioneering work done by Clark, et al. (1994) showed that by distributing computation across a computing cluster where the decomposition was done spatially, each processor could calculate a subset of the entire system, and approach near ideal speedups. A large system such as the nucleosome with a large computing cluster benefits in the same regard.

Chapter 4

Related Work

The nucleosome core protein has many available positioning sites on a given stretch of DNA. Often, the core protein will favor some positions for their stability. The stability of a nucleosome is in part correlated to the binding affinity of DNA to the histone octamer core. Nucleosome formation on DNA templates has been shown to be a statistical event with finite probabilities associated with the free energies of binding (Lowary, 1997). From this view, nucleosome formation has a probability associated with its free energy of formation. The causes behind nucleosome spacing and positioning are varied. Numerous interactions of the protein with DNA contributions may contribute to the nucleosome formation potential of a particular sequence. To complicate matters ATP has been found to be required for the assembly of physiologically spaced nucleosome arrays in some in vitro systems. Furthermore, the influence of an excess of histones over DNA in inducing nucleosome formation has been found in vitro (Lowary & Widom, 1997; Solis et al., 2004).

Our studies involve sequence effects on positioning. The effect of DNA sequence on nucleosome stability has been widely studied experimentally, and with statistical models (Scipioni, 2004). In many cases, DNA from nucleosomes selected for stability exhibits periodic nucleotide patterns in phase with DNA helical twist (Drew & Travers, 1985; Shrader & Crothers, 1990), providing evidence that DNA sequence can affect nucleosome stability. The contribution of DNA sequence on the energy of formation appear to vary from having little effect in much of bulk DNA (Lowary & Widom, 1997), to having significant phasing effects based on observed patterns in reconstitution experiments.

Satchwell et al. (1986) were among the first to discover significant patterns in nucleosomal DNA. The researchers employed a technique known as statistical sequencing to investigate sequence similarities in nucleosome forming DNA. By introducing a binding protein to certain nucleotides on a DNA sequence, and introducing a DNA cleaving agent, sequences where protein binding occurred would not cleave as easily as unbound portions. The distribution of uncut sites indicates sequence characteristics. The group discovered nucleotides in phase with the period of approximately 10.2 DNA base pairs, where short runs, two and three, of (A, T) nucleotides are positioned facing the octamer on the minor groove of DNA, and similar runs of (G, C) face outward on the minor groove, away from the core. Fourier analysis of 177 sequences further demonstrated periodic waveforms of GpC and ApA occurrences with a period of about 10.3 base pairs. Maximum frequencies of the dinucleotides were in accordance with sequence discoveries on the respective minor groove locations. The study highlighted dominant sequence features of DNA from stable nucleosome particles.

Reconstitution experiments to discover dominant DNA sequence motifs for nucleosome positioning have been on going for years. The prominent method is salt gradient dialysis where many nucleosome core proteins and DNA fragments are brought together into a highly salty solvent. The ionic strength is incrementally reduced to promote nucleosome formation. A subsequent quantitative analysis shows relative sequence preferences. These methods have populated many of our databases and contributed monumentally to the understanding of DNA-nucleosome interactions. There are, however, a number of limitations (Thastrom et al., 2004). While there is a capability to separate well-forming nucleosome sequences from poor ones, there is not a clear method for ranking DNA sequences with intermediate nucleosome formation capabilities. Researchers have only a partial view of the dominant characteristics of DNA sequence effects and thus there is currently little explanation for the most dominant features of the highest-affinity sequences.

A statistical approach employs the computational alignment of many DNA sequences known to form stable nucleosomes (Ioshikhes, 1996; Bolshoy et al., 1997). A multiple sequence alignment procedure lines up common features using heuristics. In this case, the procedure found patterns related to the frequency and position of base pairs in nucleosome formation with similar results to experiments by Satchwell and colleagues (1986). The repeated occurrence of the dinucleotides AA and TT displayed regular repeats, in phase with a complete rotation of one DNA period. The AA pattern was seen on the octamer facing minor groove, while the complementary TT was discovered six DNA base pairs down sequence in a symmetric assembly. Other nucleotide arrangements were present, but the researchers believed the AA occurrences were the main contributors. Indeed, the studies carried out by Ioshikhes et al. (1996) with myriad of alignment techniques found the same results and positional distributions other than AA and TT were not found to be informative.

Other approaches to understand prominent features involved in DNA-histone interactions highlight some fundamentals of nucleosome formation. Among these methods is work done Ramaswamy et al. (2005) who conducted a normal mode analysis of the dynamics of the nucleosome with histone variants. Interesting results from this work indicated a breathing motion by the nucleosome along its in-plane axis with the dyad whereby the nucleosome expands and compresses in a periodic motion. The breathing causes massive distortion to the nucleosomal DNA relative to the crystal structure. A comprehensive study of histone-DNA interactions, in silico, therefore demands rigorous conformational sampling of the various states corresponding to low frequency motions. Ramaswamy's work also shows that the global dynamics of the nucleosome depend on the specific DNA sequence in the bound state, indicating that sequence effects have a role in the nucleosome's active dynamics.

The requirements for DNA curvature on the nucleosome appear to be non-uniform around the histone octamer. Studies conclude that super-coiled DNA surrounds the pseudo dyad axis with about 10.5 bp per turn, and about 10.0 bp per turn to the sides of the central region (Gale & Smerdon, 1988). Fitzgerald and Anderson (1998) showed DNA sequences with non-curved regions in the center of the strand favor nucleosome positioning better than their curved counterpart does. Highly flexible DNA has been identified near chromosome breakpoint regions that are associated with chromatin disruptions (Goode et al., 1996). Other authors have shown specificity near the dyad to be opposite that of most regions (Satchwell et al., 1986) and the central 15 DNA base pairs completely unbent (Ioshikhes, 1996). This asymmetry and the unclear dyad role in rotational translation are additional motivations for detailed modeling beyond sequence-based curvature models.

Through a survey of the most stable nucleosomal DNA found, runs of AAA, found by Widlund et al. (1997), were six times more likely than the statistical expectation of randomness (reference). Even though phased tracts of adenines and their contribution to curvature have been known for some time (Wu & Crothers, 1989), Widlund suggests that the polar arrangement of adenines may give it a binding advantage to the protein core. The team visualized the presence of highly stable sequences in mouse chromosome via in situ hybridization and found them in the centromeric regions. Uniform phasing of nucleosomes is thought to cause conserved lengths of satellites among various species (Henikoff et al., 2001), with satellite lengths on the order of nucleosome DNA (Hall et al., 2003). Exceptions have been found to uniform spacing in DNA of Drosophila melanogaster correlated with satellite arrays with spacing at ~240 bp intervals instead of the bulk spacing of ~190 bp (Doshi et al., 1994).

In many cases DNA sequence is considered significant to the nucleosome stability affecting the physical outcome. In gene regulation, five-prime ends of tissue-specific genes have been seen to have higher nucleosome formation potential than house-keeping genes (Ganapathi et al., 2005). Rotational positioning has a role in transcription and facilitated binding, such as with the TATA-binding protein. Configurations of nucleosomal DNA in which the TATA box is interior to the nucleosome cannot be accessed for the initiation of transcription without rearrangement (Imbalzano et al., 1994). Many molecular processes involve nucleosomes where positioning and stability factors are not clearly understood.

Chapter 5

Design

5.1 Configuration Generator

The computational framework presented combines three main software technologies into a high-throughput system (Figure 5.1): automated mutation of a nucleosomal DNA sequence, a number of molecular modeling computations, and the retrieval of associated data. Combined with these technologies, a hierarchical control system permits the parallel and incremental modification of nucleosomal DNA.

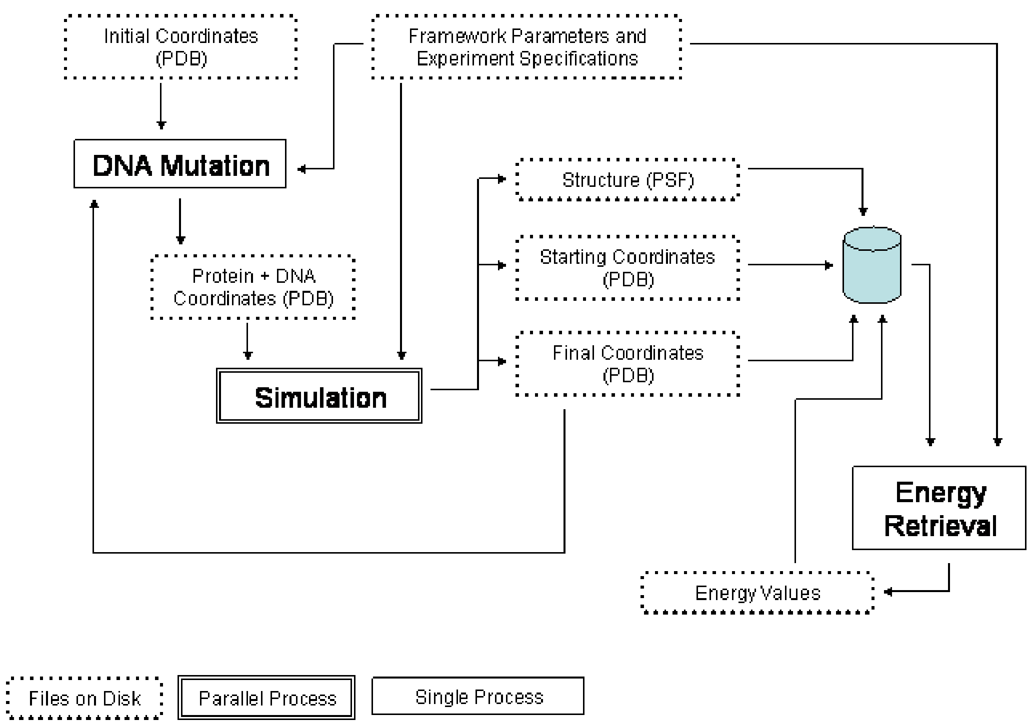

Figure 5.1 Configuration Generator. DNA mutations and subsequent simulations form an automated loop in the framework. All subsystems use a common database to share information.

5.1.1 DNA Mutation

For the purposes of a high-throughput investigation of DNA sequence effects on nucleosome stability, a system for automated mutations was built to integrate seamlessly the production of new sequences into the computational modeling framework. The DNA mutation subsystem composes the creation of proper input to the software making the mutation, the mutation itself along with its corresponding output structure in the form of coordinates, and preparation of that output for use in the simulation subsystem (Figure 5.2).

To achieve a successful mutation of DNA under the overall framework, two important input files are necessary: an input coordinate file on which the mutation will occur, and a script directing detailed execution of the third-party modeling software that performs the mutation. The execution script includes commands specifying coordinate files and mutation calls specific to each appropriate position on the DNA molecule. For each instantiation of a configuration generator, a master input file directs the appropriate mutation. The master input file contains the identification number of the current configuration, the identification number of the coordinate system needed for input into the mutation program, the positions on which the mutation should occur, and the actual residues desired for mutation. The global input file provided to the framework at execution time provides the location of a single file describing every mutation desired by each configuration generator; that is, the controlling process uses the global input file to construct a unique master file for every mutation needed. The system determines base pair matches to the master file (which only directs single stranded mutation on the double stranded DNA molecule). After the determination of the proper input coordinate file, all positions and residues needed for the specific round of mutation, and the output filename, the process that executes the mutation reads from the final input script.

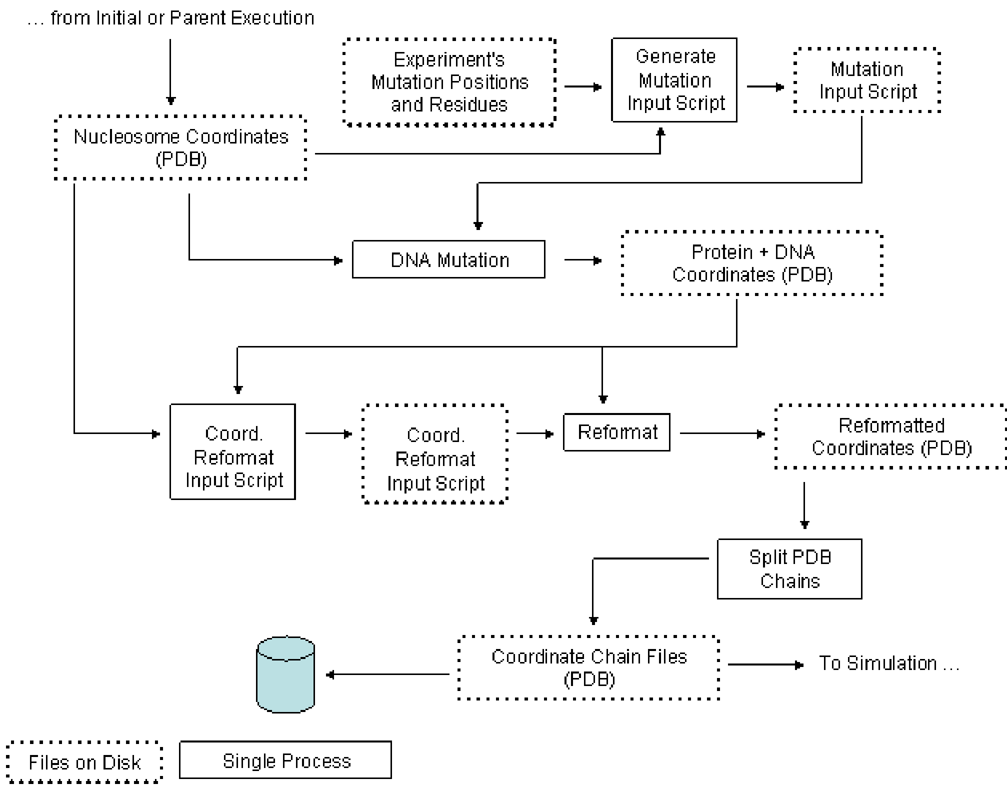

Figure 5.2 DNA Mutation Subsystem. Scripts are automatically generated to instruct a separate mutation program.

After the mutating process completes the replacement of all residues directed to it from the input script, the process writes the new coordinates to disk. These coordinates contain all three-dimensional positions of the protein and mutated DNA. Because there are various representations of coordinate files in the molecular modeling field, a series of reformatting scripts aid the automatic flow of the computational framework. Among these scripts are those needed to rename and organize residues for compatibility in subsequent simulation programs, and those scripts designed to add solvent to the nucleosome system, if desired. Compatibility with simulation packages is often related to the specific molecular topology and force field parameters specified by the researcher; thus, reformatting likely needs customization. Various scripts built for use in this framework are capable of providing compatibility with multiple simulation packages, as well as with the molecular modeling software used for mutation. One additional step needed by the simulation subsystem is the splitting of protein and DNA chains, along with division of the solvent into chains, for structure file creation. The newly split coordinate files are stored in a flat database for subsequent access.

5.1.2 Simulation

The simulation subsystem composes all necessary steps to perform a full-scale molecular simulation, be that an energy minimization or molecular dynamics run. The steps include the construction of a protein structure file, scripts needed to submit a high-performance execution on a super-computing platform, and all directives needed for the actual molecular simulation (Figure 5.3).

Coordinates for the protein and DNA chains composing the molecular system to simulate are products of the mutation subsystem. Initially, a process generates a protein structure file by reading a specified molecular topology and force field parameters from disk and the atomic coordinates contained in PDB files, organized by protein and DNA chains. In addition to building the structure file, the process rebuilds a coordinate file with the atomic coordinates for all atoms from the individual chain files. The structure file and new coordinate file are stored in the database. The new coordinate file contains the starting coordinates for the molecular simulation. According to the usage preferences of the experimenter, the framework produces a submission script to request execution on a parallel cluster. This script requests the allocation of new computing nodes exclusively for the computations needed by the simulation package. The submission script principally contains the number of compute nodes desired and the execution arguments needed by the simulating program. Decisions about how many nodes and the processors-per-node are left to the experimenter and may depend on cluster resources, various user and queue limits, and the time needed to complete a simulation. All cluster input script options are accessible to the experimenter through variables in the configuration generator main program.

The framework builds another script needed directly by the simulation module. This script specifies the numerous details concerning every parameter and option needed for a specific simulation, such as length of time to run, cut-off distances for force calculations, methods for the computation of long-range forces, and others. The simulation input script also contains the location of all necessary files: the structure file, the coordinate file, molecular topology and force field parameter files, and any restart files that may be applicable. Operators also specify output options such as the coordinate capture frequency for a trajectory, and the output path for the final coordinates. The precise requirements of a given experiment will determine the values for each option in the input script, along with the format needed by the simulating program. Variables in the code assist the input script generation as much as possible. However, because simulation packages require custom input scripts, the main configuration generator program will output to file the exact text needed by the experimenter’s preferred package.

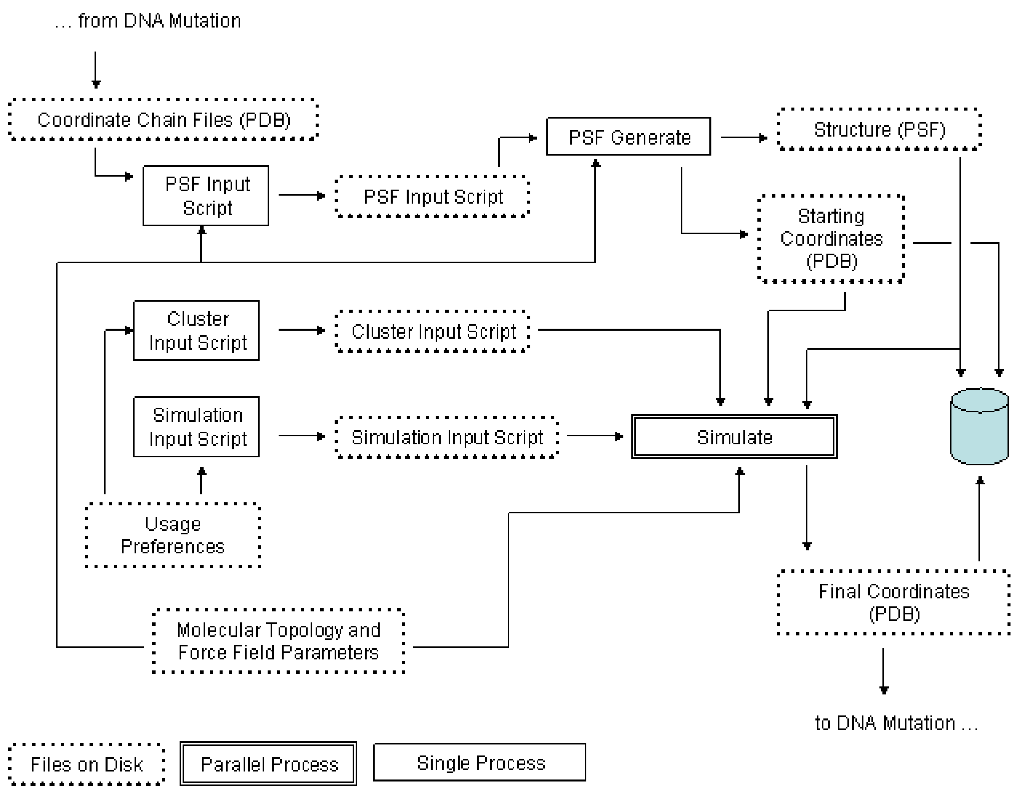

Figure 5.3 Simulation Subsystem. Generated structure and coordinate files are stored in a database for use in analysis.

After simulation scripts are prepared for some single mutation, and the necessary files are available and correctly formatted, a call to the cluster scheduling system by the cluster input script commences the actual molecular simulation procedure. If resources are available, the high-performance simulation will begin while the main worker program waits. The main worker program will also block if the resource request is queued and waiting for compute nodes. If the simulation exits normally, the final coordinate and/or trajectory file is ready and written to the database, making it available for usage in subsequent mutation rounds. The configuration generator can proceed to the final data gathering process.

5.1.3 Energy Retrieval

Every simulation begins with a purpose. One main purpose of energy minimizations is to assess and compare the conformational energy of different molecular configurations. The framework incorporates the need for energetic analysis through an energy retrieval subsystem following every molecular simulation.

The energy retrieval and analysis subsystem uses output from the simulation subsystem to compute and extract various energy values. The framework's database provides many of the input files needed by the energy calculating process, such as the structure and coordinate files (Figure 5.4). The molecular topology and force field parameters are also necessary, but do not have to be the same files used in the simulation. For the energy retrieval of a particular or interesting part of the simulated system, an atom index is accepted. In the case of analyzing stability through the metric of an energy comparison, the energy subsystem automatically builds an index of all atoms making up the nucleosomal DNA. This index is provided to the calculation engine in order examine the self-energy of DNA, while in a minimized conformation with the nucleosome core protein.

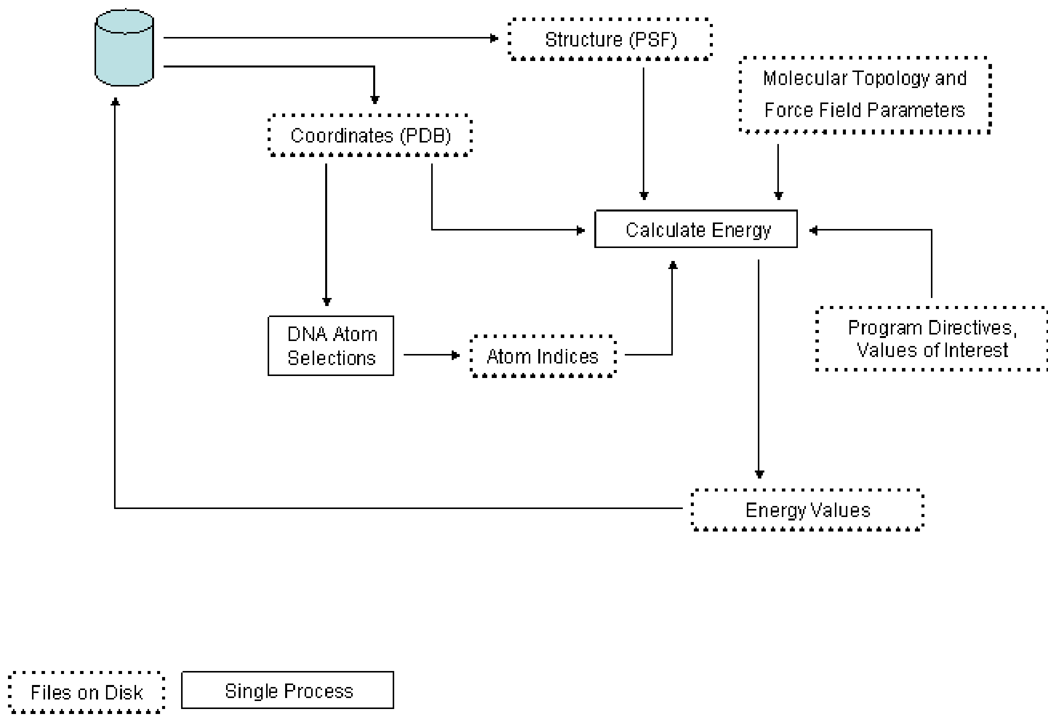

Figure 5.4 Energy Retrieval Subsystem. Atom indices can be generated specific to each coordinate file automatically to facilitate a detailed energy analysis.

Further directives to the calculating process are the types of energy needed, such as total, internal, electrostatic, etc. The energy values specified for each sequence with its index are added to a single file with all other computed configurations in the database, enabling rapid plotting and convenient experimental analysis.

5.2 Control System

A single starting nucleosome structure initiates the direction of the entire framework. The initial structure is the root node of a mutation tree. Each level of the tree corresponds to the number of mutations performed on the root sequence. Each parent node of the tree corresponds to the starting structure for each child node's mutation and subsequent simulation. A child structure is a single mutation from its parent. All nodes on the same level of the tree can undergo configuration generation in parallel, since the nodes on the same level are independent of one another. The dependency-based control structure of the configuration generators mitigates extensive disturbance to the root nucleosome structure. Incremental mutations with energy minimization are explored in this work as a way to ameliorate these perturbations in order to retain structural and dynamic integrity for generating a multitude of configurations when full-scale dynamics of each configuration are not possible.

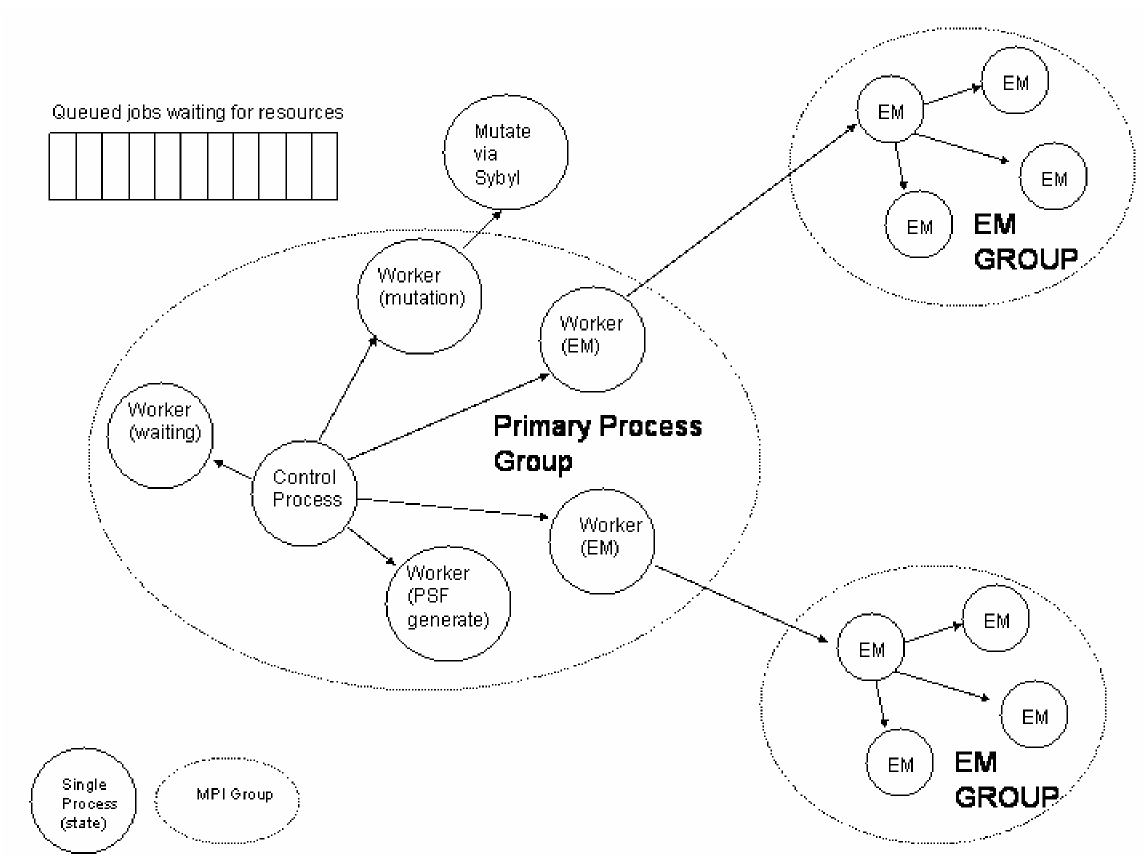

Figure 5.5 Process View. Worker processes can be in various states during the suite of configuration generations.

Parallelism is doubly exploited: configurations are generated in parallel as independent simulations execute in parallel (Figure 5.5). Operationally, this architecture permits configuration generators to be in multiple different states and multiple levels of the mutation tree, such that dependency rules are satisfied. One worker process, an individual configuration generator, may be blocking on the completion of a parallel energy minimization, while another may be generating a protein structure file. A worker may be waiting for queue resources while another is executing the mutation program.

Chapter 6

Proof-of-concept

6.1 Overview

The presented framework, through the use of a high-throughput mutation scheme, molecular simulation, and energy calculations, allows for the detailed comparisons of thousands of sequences. The proof-of-concept study concerns two dominant features for distinguished sequence effects on DNA-histone binding involving tri-nucleotide patterns on the minor groove of DNA. For twelve positions and trinucleotide sequences where this feature is believed to provide flexibility to the DNA wrapping the nucleosome, configurations were generated with the tri-nucleotide motifs for 4096 permutations of this feature on twelve symmetric positions. An analysis of the resulting conformational energy permits the comparison of the substitutions at specific positions toward assessing this feature importance. These sequence effects on curvature are well studied (e.g., Goodsell & Dickerson, 1994), so that this suite of sequences provides a test of our methodology, and data for study sequence effects on the set of structures generated. Study details and hypotheses are as follows:

There are 12 sequence positions on either side of the central position of DNA sequence surrounding a nucleosome core considered important to nucleosome stability.

AAA tri-nucleotides contribute a significant degree to nucleosome stability when found +5, +15, +25, +35, +45, and +55 from the central DNA base pair position.

Correspondingly, the tri-nucleotide contributes significantly at -5, -15, -25, -35, -45, and -55 from the center.

GGC tri-nucleotides contribute a significant degree to nucleosome stability when found +10, +20, +30, +40, +50, and +60 from the central DNA base pair position. Correspondingly, the tri-nucleotide contributes significantly at -10, -20, -30, -40, -50, and -60 from the center.

Mutations are done symmetrically with respect to the DNA sequence center. In other words, a single configuration contains one tri-nucleotide motif at the same offset, both positive and negative from the center position.

Certain positions and combinations of a tri-nucleotide mutation may have a more pronounced affect on stability than others. For example, a single mutation of GGC at a central offset of 40 may contribute more to a lower energy configuration than the same mutation at central offset of 10.

Comparison of DNA sequence effects on nucleosome stability can be interpreted from differences in the total energy calculations of each engineered DNA conformation.

6.2 Starting Structure

An appropriate starting structure is imperative to reduce error associated with the subsequent series of configuration generations. The structure selected for this study was produced by a series of thermalization and equilibration procedures using molecular dynamics. The crystal structure, PDB 1KX5, was downloaded from the Protein Data Bank. The structure was assessed for missing heavy atoms and other anomalies and none were found. The tails of the two histones H3 (PDB chains A and E) were cut 38 amino acids from the N-terminals to reduce configuration variability associated with unpredictable movements of these tails and contact with the DNA. The system was solvated in a 164 angstrom cube, sufficient to absorb a wide-range of motion of the system. One hundred forty sodium ions were added at randomly chosen positions to neutralize the system charge. The solvation and ionization procedures were done using the 'genbox' and 'genion' programs from the GROMACS v3.3 suite (van der Spoel et al., 2005). With water and ions, the total system contained 440,794 atoms. The CHARMM (Brooks et al., 1983) package was used to build a protein structure file in the X-PLOR format for subsequent use in the MD program of choice, NAMD (Phillips et al., 2005).

All thermalization procedures were done in a NPT ensemble in the 164 angstroms periodic cube. A time-step of one femtosecond was used with a cutoff radius of 16 angstroms for short-range interactions and full particle mesh Ewald (PME) for long-range electrostatics. Initially, the system was energy minimized for 1000 steps. The solute was then frozen while the water and ions underwent dynamics at 100K for 120 picoseconds. The solute was then free and the solvent frozen for dynamics at 100K for 120 picoseconds. The two steps were repeated again, only at a higher temperate of 300K for the same amount of time. Lastly, all atoms were unfrozen and free to move at 300K for 220 picoseconds.

After the rigorous thermalization procedure, the final frame of the simulation was selected to begin the study DNA sequence effects on stability. Water and ions were removed from the structure before beginning the framework run. The existing framework has been developed with more complex, explicit solvent systems in mind. Scripts to handle large systems, especially naming of water chains, have been developed and tested independently of this prototype study.

6.3 Framework Specifications

The mutation subsystem utilized the Biopolymer module of the Sybyl 7.2 software from Tripos, Inc. Input scripts were built according to descriptions in Chapter 5. The Sybyl software was executed in serial from an application server networked to the computing cluster. Secure sockets handling was used to execute the input script across the network, capture standard error and output from the Sybyl program, and return proper control to the main worker program. Mutations were done by executing the ‘biopolymer replace’ function from the Sybyl command line. This function preserved the directionality of the DNA backbone during residue replacement.

Simulations and protein structure files were processed with the CHARMM program, version 31b2. Protein structure file generation was done on the worker compute node and produced output in CHARMM and X-PLOR format. In order to assess the prototype framework, an implicit solvent energy minimization in CHARMM was run for each sequence. The Combined CHARMM27 All-Hydrogen Protein and Nucleic Acid topology and parameters (Foloppe et al., 2000) were used in the energy minimizations along with the Analytical Continuum Solvent (ACE) parameters, to model the solvent implicitly (M. Schaefer & M. Karplus, 1996). The low computational overhead of ACE allowed the framework to work in an actual application, and to develop a database of preliminary structures for future analysis. Implicit solvent can be problematic for simulating the highly charged DNA system and may introduce artifacts. The explicit solvation used to generate the base nucleosome configuration was too computationally expensive for the early runs. ACE was suitable for prototyping to relieve unnatural conformations introduced by nucleotide substitution. Energy minimization was run for 400 steps of steepest descent and 100 steps of conjugate gradient. The minimization steps were chosen to minimize undesirable effects caused by the simplified implicit solvent model, but to provide meaningful results for discussion. The minimizations were submitted to the queue system on a compute cluster using four processors.

The energy subsystem used the MDenergy (Kale et al., 1999) program to calculate the internal energy of the resulting structure after energy minimization. An index of only the DNA atoms was generated but not used and the resulting energy values were total energy for the protein and DNA. The atom selection features provided by the index allow specifying molecular subsystems to the energy retrieval module. Each energy value for every sequence was concatenated into a single file listing all energies indexed by sequence number.

6.4 Simulation Results

Energy values were taken from the file produced by the framework and assessed against an existing curvature program called BEND (Goodsell & Dickerson, 1994). The curvature program uses only sequence information to produce a arbitrarily scaled curvature value describing relative curvature potential of a DNA sequence. The BEND program used FASTA sequence files also produced by the framework using different modules not detailed here. The purpose of this comparison is to assess the resulting configuration energies against an existing, experimentally parameterized model. Generally, good curvature of DNA correlates to well-forming nucleosomes, so the assessment can also hold validations of the tri-nucleotide motif observations and hypotheses.

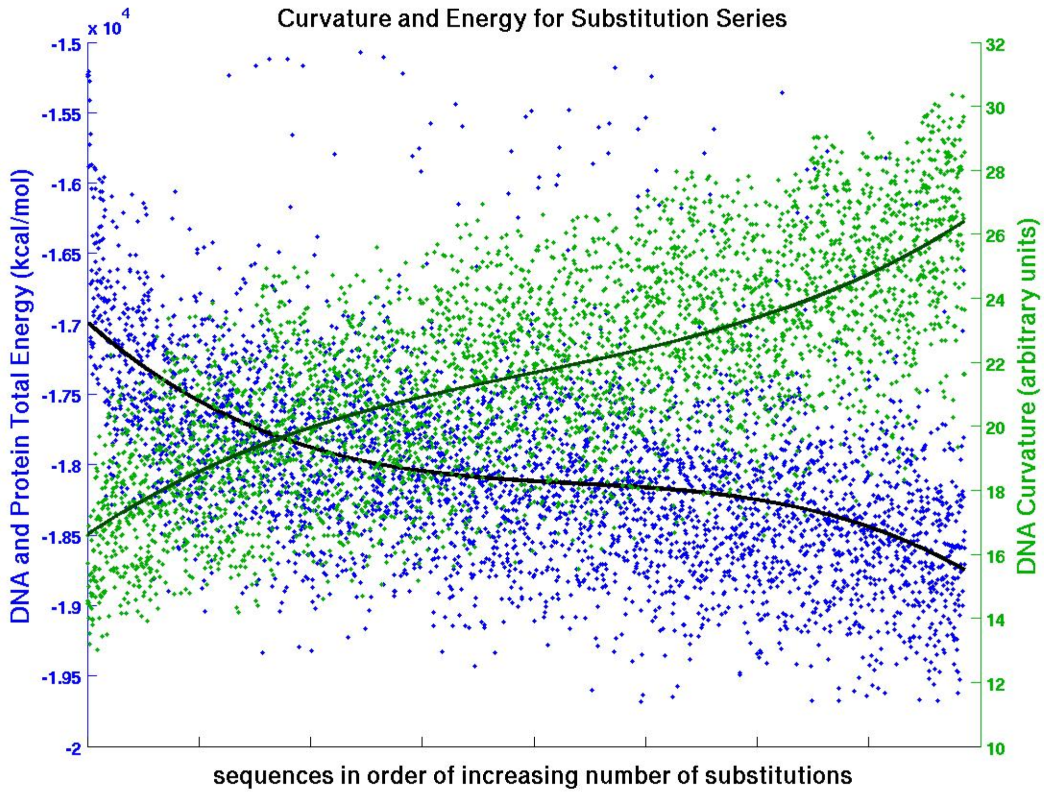

Figure 6.1 Curvature and Energy for Substitution Series. Comparisons between bendability and energy could indicate a correlation between DNA sequence effects and nucleosome stability. The prototype study shows large-scale computational experiments on nucleosomes are possible with the presented framework.

In Figure 6.1, the total energy of the minimized structures is plotted with the curvature computed by BEND. The program BEND uses an arbitrary scale for curvature. The increase in curvature with our substitutions correlates with a decrease in the energy of the substituted structures. The resulting plot may describe the increased affinity for nucleosome forming DNA sequences with favorable mutations at dominant bending positions. Enhanced curvature values match the relatively low energies of highly-mutated sequences. However, the energy trend is at least partially a result of the accumulation of minimization steps on the substituted structures; the number of steps of minimization on a structure increases with the number of substitutions on it. More detailed studies could use a greater number of minimization steps or a change in energy cutoff to reduce the error involved in this study.

Our independent analysis of ACE minimization suggest that excessive application of ACE may alter the nucleosome particle in ways incompatible with an explicitly solvated system while lowering the electrostatic energy; this could be exacerbated by increased minimization steps. Definitive assessment of ACE in this regard requires further study.

The curvature as a predictor of nucleosome stability should also be carefully qualified. DNA curvature models like BEND are developed for free DNA. A variety of approaches reach a common conclusion of more tightly super-coiled DNA around the pseudo dyad axis of about 10.5 bp per turn and about 10.0 bp per turn to either side of the central region (Gale & Smerdon, 1988). This asymmetry in curvature on the nucleosome is one point exemplifying the need for molecular modeling, beyond simple sequence-based curvature models. Furthermore, excessive DNA curvature has been found to impede nucleosome formation (Scipioni et al., 2004).

Chapter 7

Discussion

The framework presented here is geared for molecular modeling of larges classes of DNA sequences to extend the analysis of DNA effects in the complex nucleosome environment. These can include computationally intensive approaches such as explicit solvent models and free energy calculations, which are enabled by the parallel architecture. Numerous possibilities exist for the application of such a framework to address limitations inherent in other investigatory methods.

7.1 Performance

The study presented generated 4096 sequences through mutation and minimization, and performed an energy calculation on them. The parallel architecture of the framework was found effective to model a significant set of nucleosome configurations. The limits to this parallelism are a configuration's dependency on parent configurations, available time, and computing resources.

The prototype model generated and minimized a single configuration with energy calculation in about 305 seconds. Each protein structure file generation took ~135 seconds, the parallel energy minimization needed ~145 seconds, the energy calculation was ~5 seconds, and the mutation with coordinate preparation and extraneous I/O took ~20 seconds. If the prototype study were to be done sequentially, the generation and calculation of all 4096 sequences would take 1249280 CPU seconds, roughly 14.5 days. In the parallel study, 20 processors were originally used to operate the framework: one control processor, and 19 worker processors. Since the time a worker spent in energy minimization was about equal to time processing sequential operations, about 10 workers were utilizing four additional processors at a given time running simulations. Therefore, using 60 total processors, the parallel framework completed the prototype study in 66490 computing seconds, just under18.5 hours. In a parallel run consisting of 19 worker processes, there are 218 rounds of generation necessary to configure all 4096 sequences[1]. Under the same experimental conditions, if 373 total processors were utilized with124 worker processes, only 38 rounds of generation would be necessary, and the total time to configure all sequences would be 3.22 hours. In one month almost a million sequences could undergo configuration and energy calculation about the nucleosome.

Certainly the simulations performed in the prototype were not adequate to conclusively validate bendability hypotheses proposed by biologists in the field. Limitations to the prototype were the lack of solvent to the system, arbitrary minimization times, and conformational disturbance to the nucleosome structure, possibly pushing the system out of an equilibrated state. Protein structure file generation for the 440,794 atom system requires 20 minutes using CHARMM. Molecular dynamics on such a structure with 120 processors requires about 80.5 hours per nanosecond and about one minute per 100 steps of energy minimization. Running full molecular dynamics in NAMD for ten nanoseconds (roughly the minimum time needed to sample the phase space of the nucleosome's dynamics), would take more than a month for the configuration of one sequence. Practicality clearly enters into the

[1] The number of sequences on the same level of the mutation tree from the root sequence goes as the binomial distribution $\binom{d}{m}$, where d is the total number of mutation sites and m is the number of mutations performed. For example, for m=6 mutations on a mutation template with d=12 sites there are 924 possible sequences. The set of possible sequences on one level of the mutation tree is not dependent on the whole of its parent level set, so that branches of the mutation tree can grow as fast as possible. The framework trickles through new configurations while maximizing available parallelism.

experimental design and modeling approximations. If the same solvated structure underwent an energy minimization for 6000 steps instead of dynamics, total time for a single configuration would be around ten hours. With 1160 processors, one could begin analyzing the energies of all 4096 sequences on day 92; with 7565 processors, the time needed to emulate the prototype system would be less than 14 days. Data management is an issue facing the large numbers of coordinate and structure files, along with trajectories of the simulation. The prototype system requires nearly 200 gigabytes of disk space for the suite of 4096 sequences. Larger structure requirements for storage are linear with the number of atoms, as are the trajectory files. Further performance extrapolation hints that a study likely exists to take advantage of the framework in a reasonable way, exploiting the ever increasing availability of resources and computing time with interesting and rewarding results.

7.2 Applications

The free energy of DNA binding to the nucleosome, $\Delta\Delta G$, could in principle be obtained using rigorous free energy simulations. In these, free energies of DNA mutations $\Delta G3$ and $\Delta G4$ would be calculated using the thermodynamic cycle in Figure 7.1 to write $\Delta\Delta G = \Delta G2 - \Delta G1 = \Delta G4 - \Delta G3$. However, the large system size and large number of DNA sequences of interest make free energy simulations unfeasible. The approach developed in this work supports approximations to this modeling. The concept is geared to reduce computational overhead. Generally, this is done by perturbing DNA on structures from fully prepared simulations, followed by abbreviated simulations on the perturbation. The abbreviated simulations may range from short energy minimizations to more computationally demanding molecular dynamics simulations.

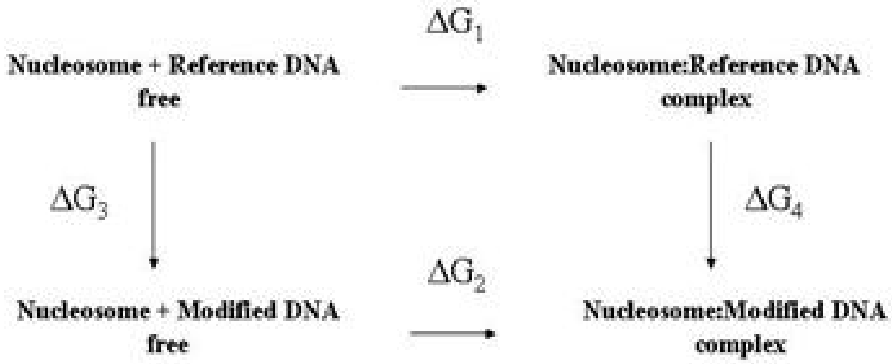

Figure 7.1 Free Energy Perturbation Cycle. A potential study comparing the free energies of a nucleosome complex with that of free DNA and protein could highlight relative binding affinities.

In Figure 7.2, F0 and F1 represent fully simulated systems and their molecular dynamics trajectories. The binary representation 000 indicates DNA without substitutions relative to the base crystal structure; the binary representation of 111 represents substitutions at all sites but fully equilibrated and simulated. In this scheme, substitutions are made on the two full trajectories at the points indicated along it.

The vertical arrows in Figure 7.2 represent short simulations such as energy minimizations (as in this study), or they could be more involved dynamics runs. The key idea is that multiple simulations are performed by perturbing base simulations as is tolerable. In the hypercube at the left, sequences other than 000 and 111 are derived from either 000 or 111; the reported work uses a single root sequence in the substituted series, that of 000. In our prototype study, the simulations are performed in substitution order from the root 000 downward into the tree spanning the hypercube in order to systematically perturb the structure. In the example given in Figure 7.2, substitutions are only one edge from either 000 or 111 so that the substitution dependence is not present. The need for extensive minimization or equilibration could be reduced or practically eliminated.

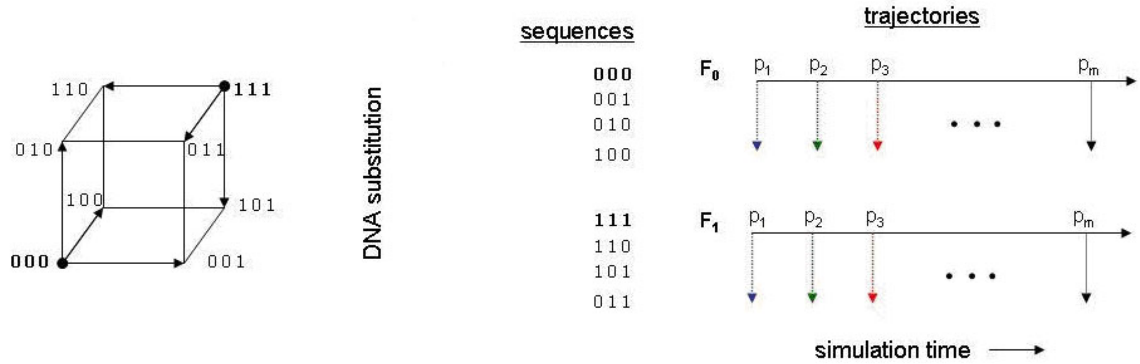

Figure 7.2 Low Perturbation Study. Multiple starting structures can be used in detailed experiments. Notice sequences labeled 000 and 111 form the root node on separate mutation trees. Points along the fully equilibrated trajectories can be used as new starting points to sample the phase space of nucleosome configurations.

Stochastic searching of dominant stabilizing features of nucleosomal DNA could benefit from the presented framework as well. A module extending the use of the energy calculation subsystem could evaluate the sequence's binding affinity and help discover even better sequences, such as in a genetic algorithm. Sequences with relatively low energies would persist, mutate randomly or with calculation, cross-over with sequences on its current level or population, and evolve with other genetic programming techniques. Highly dominant characteristics could rapidly emerge and persist simultaneously with experimentally introduced sequences. The application of a genetic algorithm to an actual simulated genetic system may very well be the first of its kind. Further applications are wide open.

Chapter 8

Conclusion

The study of the nucleosome contains many opportunities for an enhanced understanding of cellular processes. Nucleosome studies can also provide insight into the mechanics of DNA itself. Such motivations are important to unraveling the mysteries of genetics, evolution, and functional genomics. Understanding more about DNA could prove useful in the development of many areas of science: such as the enhancement existing genomic studies involving DNA reagents; formulation of hypotheses concerning genetic disorders and diseases; principles of genetic engineering such as centromere dynamics; and many others. Nucleosome specifics are instrumental to the understanding of transcription processes and chromatin folding--vital processes in life.

Fundamental to the dynamics of the chromatin genetic management strategy are the DNA positions where nucleosomes tend to form. Currently, gaps exist in understanding how and why nucleosomes form in the positions they do. In numerous studies, there appears to be a range of sequence specific effects at work in a stable nucleosome. Detailed dynamical studies to assess the energetics and structural characteristics of a large number of DNA sequences are needed to refine the statistical positioning picture. Three dimensional molecular modeling is a crucial complement to linear sequence studies. The parallel molecular modeling framework presented is an early crucial step in the generation of a publication-class data to enrich the understanding of the nucleosome.

Work to extend and improve the framework would allow even greater discovery and increase relevance of the system. Modularity of the framework encourages the use of separate packages for all subsystems. Certain experiments may require a different mutation scheme or simulation engine. Modules currently built to generate input scripts, for example, could easily be copied and modified in order to establish libraries of subsystems for many diverse applications of the framework. Design features easily scale with the amount of available resources, enabling rigorous and large-scale calculations now and in the future. The parallel framework works to reduce the total amount of idle CPU time but does not eliminate it. Efficiency could improve if worker processes could ready new simulations while waiting for simulations on cluster resources. Opportunities exist to use the valuable technologies of this work and associated collaborations to enhance computational workflow systems. The programs developed by the presented effort, and the parallel-configured modeling tools could bridge automation routines in new environments easily. Such integration would improve the quality of use through provenance and visualization features and extend utilization into deeper biological perceptions.

The underlying computational framework and associated data is only important with an innovative study. This thesis demonstrates that work involving thousands of mutated structures is possible with a reasonable amount of resources. New and imaginative nucleosome studies can place full attention on experimental design and the analysis of interesting data while the parallel framework configures novel structures automatically.

References

Bolshoy A, Shapiro K, Trifonov EN, and Ioshikhes I: Enhancement of the nucleosomal pattern in sequences of lower complexity. Nucleic Acids Research, 25(16):3248-3254, 1997.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, and Karplus M: CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem., 4(2):187-217, 1983.

Clark TW, vin Hanxleden R, McCammon JA, Scott LR: Parallelizing molecular dynamics using spatial decomposition. Proceedings of the Scalable High Performance Computing Conference, 1994.

Davey CA, Sargent DF, Luger K, Maeder AW, and Richmond TJ: Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 Angstrom Resolution. J. Mol. Biol., 319:1097-1115, 2002.

Doshi P, Kaushal S, Benyajati C, and Wu C-I: Molecular analysis of the responder satellite DNA in Drosophila melanogaster: DNA bending, nucleosome structure, and Rsp-binding proteins. Mol. Biol. Evol., 8(5):721-741, 1991.

Drew H and Travers AA: DNA bending and its relation with nucleosome positioning. J. Mol Biol., 186:773-790, 1985.

Fitzgerald DJ and Anderson JN: Unique translational positioning of nucleosomes on synthetic DNAs. Nucleic Acids Research, 26(11):2526-2535, 1998.

Foloppe N and MacKerell AD: All-atom empirical force field for nucleic acids. I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J. Comp. Chem. 21:86-104.

Gale JM and Smerdon MJ: Photofootprint of nucleosome core DNA in intact chromatin having different structural states. J. Mol. Biol., 204:949-958, 1988.

Ganapathi M, Srivastava P, et al.: Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC Bioinformatics, 6:126, 2005.

Goode JS, Kass SU, Hirst MC, and Wolffe AP: Nucleosome assembly on methylated CGG triplet repeats in the Fragile X mental retardation gene 1 promoter. J. Biol. Chem., 271:24325-24328, 1996.

Goodsell DS and Dickerson RE: Bending and curvature calculations in B-DNA. Nucleic Acids Res. 1994, 22:5497-5503.

Hall SE, Kettler G, and Preuss D: Centromere satellites from Arabidopsis populations: maintenance of conserved and variable domains. Genome Research, 13:195-205, 2003.

Henikoff S, Ahmad K, and Malik HS: The centromere paradox: stable inheritance with rapidly evolving DNA. Science, 293:1098-1102, 2001.

Imbalzano AN, Kwon H, Green MR, and Kingston RE: Facilitated binding of TATA-binding protein to nucleosomal DNA. Nature, 370:481-485, 1994.

Ioshikhes I, Bolshoy A, Derenshteyn D, Borodovsky M, and Trifonov EN: Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J. Mol. Biol., 262:129-139, 1996.

Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, and Schulten K: NAMD2: Greater scalability for parallel molecular dynamics. J. Comp. Phys., 151:283-312, 1999. MDEnergy author: Jan Saam. Program available online: http://www.ks.uiuc.edu/Development/MDTools/mdenergy/

Kimball JW: Kimball's Biology Pages. Text available online: http://biology-pages.info, 2006.

Kornberg RD and Lorch Y: Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. Cell, 98:285-294, 1999.

Lowary PT and Widom J: Nucleosome packaging and nucleosome positioning of genomic DNA. Proc. Natl. Acad. Sci., 94:1183-1188, 1997.

Luger K: Dynamic nucleosomes. Chromosome Res., 14:5-16, 2006.

MacKerell AD, Wiorkiewicz-Juczera J, and Karplus M: An all-atom empirical energy function for the simulation of nucleic acids. J. Am. Chem. Soc. 117:11946-11975, 1996.

Muthurajan UM, Park Y-J, Edayathumangalam RS, Suto RK, Chakravarthy S, Dyer PN, and Luger K: Structure and Dynamics of Nucleosomal DNA. Biopolymers, 68:547-556, 2003.

Philips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, and Schulten K: Scalable molecular dynamics with NAMD. J. Comp. Chem., 26:1781-1802, 2005.

Ramaswamy A, Bahar I, and Ioshikhes I: Structural Dynamics of Nucleosome Core Particle: Comparison with Nucleosomes Containing Histone Variants. Proteins: Structural, Function, and Bioinformatics, 58:683-696, 2005.

Satchwell SC, Drew HR, and Travers AA: Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol., 191:659-675, 1986.

Schaefer M and Karplus M: A Comprehensive Analytical Treatment of Continuum Electrostatics. J. Phys. Chem., 100(5): 1578-1599, 1996.

Scipioni A, Pisano S, Anselmi C, Savino M, and De Santis P: Dual role of sequence-dependent DNA curvature in nucleosome stability: the critical test of highly bent *Crithidia fasciculata* DNA tract. Biophys. Chem., 107:7-17, 2004.

Shrader TE and Crothers DM: Artificial nucleosome positioning sequence. PNAS, 86:7418-7422, 1989.

Thastrom A, Bingham LM, and Widom J: Nucleosomal Locations of Dominant DNA Sequence Motifs for Histone-DNA Interactions and Nucleosome Positioning. J. Mol. Biol., 338:695-709, 2004.

Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, and Berendsen HJC: GROMACS: Fast, Flexible, and Free. J. Comp. Chem., 26:1701-1718, 2005.

Widland HR, Cao H, Simonsson S, Magnusson E, Simonnson T, Nielson PE, Kahn JD, Crothers DM, and Kubista M: Indentification and Characterization of Genomic Nucleosome-positioning Sequences. J. Mol. Biol., 267:807-817, 1997.

Widom J: Role of DNA sequence in nucleosome stability and dynamics. Quarterly Reviews in Biophysics, 34(3):269-324, 2001.