ON THE MECHANOCHEMICAL MACHINERY UNDERLYING CHROMATIN REMODELING

By

TAHIR I. YUSUFALY

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

written under the direction of

Wilma K. Olson

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2014

ABSTRACT OF THE DISSERTATION

On the Mechanochemical Machinery Underlying Chromatin Remodeling

By TAHIR I. YUSUFALY

Dissertation Director:

Wilma K. Olson

This dissertation discuss two recent efforts, via a unique combination of structural bioinformatics and density functional theory, to unravel some of the details concerning how molecular machinery within the eukaryotic cell nucleus controls chromatin architecture. The first, a study of the 5-methylation of cytosine in 5′-CG-3′ : 5′-CG-3′ base-pair steps, reveals that the methyl groups roughen the local elastic energy landscape of the DNA. This enhances the probability of the canonical B-DNA structure transitioning into the undertwisted A-like and overtwisted C-like forms seen in nucleosomes, or looped segments of DNA bound to histones. The second part focuses on the formation of salt bridges between arginine residues in histones and phosphate groups on the DNA backbone. The arginine residues are observed to apply a tunable mechanical load to the backbone, enabling precision-controlled activation of DNA deformations.

Acknowledgments

My journey through graduate school was a turbulent one, consisting of the highest of highs and the lowest of lows. No person is an island, and truly, no one can hope to undertake a Ph.D. in isolation while also staying sane. An uncountable number of individuals made this possible, and any attempt to exhaustively acknowledge everyone who deserves to be acknowledged will inevitably fall short. For those whose names do not directly appear below, I hope you will forgive me - know that I could not have done this without you.

The first person that deserves special mention is, of course, my Ph.D. supervisor, Wilma K. Olson. I truly won the lottery in getting the opportunity to work with her. She is a spectacular and brilliant scholar, a patient and wise teacher, and above all, an exceptional human being. Thank you for everything - for believing in me when I didn't believe in myself, for inspiring me to grow both scientifically and personally, and overall for being an example that I will spend the rest of my life striving to live up to.

I am also indebted to several excellent physicists, both at Rutgers and beyond. This dissertation is dedicated in part to the late David Langreth, who taught me solid-state physics in my first year of graduate school, and who spearheaded the development of the crucial van der Waals density functional theory, without which this work would not be possible. David Vanderbilt is one of the most brilliant people I have ever met, demonstrating an invaluable mix of impeccable organization, even-tempered wisdom, and overall five-star professionalism. Karin Rabe has largely influenced my overall attitude and mindset

towards materials theory and design, with her unmatched creativity in using electronic structure methods to do borderline magical things. Premi Chandra has taught me a lot about how to mix microscopic details with phenomenological thermodynamic models that capture essential qualitative insights. Piers Coleman never ceases to inspire me with his enthusiastic childlike passion for discovering new physics in unexpected places. Anirvan Sengupta has helped me out tremendously with invaluable advice and suggestions to ease the transition from from traditional hard condensed matter physics to soft biological matter. Special thanks are afforded to Ned Wingreen and David Botstein of Princeton University, who were generous enough to let me cross-register into their graduate quantitative biology course in the fall of my fourth year. I am grateful to Saurabh Jha and Abdelbaki Brahmia for helping me to grow as a teacher in addition to a researcher. And of course, no list of acknowledgments would be complete without mentioning Ron Ransome, who truly went to bat for us graduate students - I could not have asked for a better person to serve as department chair during these most crucial years.

And finally, I give thanks to all of my family and friends who helped me get to this point. I consider it a sign of how blessed I am to have far too many of you to be able to mention by name. To my four grandparents - Noman Yusufaly (Daada Jaan), Abul Mubashir Siddiqi (Naana Jaan), Momina Siddiqi (Naani Maa) and Khadila Mulla Ibrahimjee (Daadi Maa) - thank you for struggling to give my parents the opportunity to give my sister and I the opportunity to study and succeed in this country. We all stand on your shoulders. To my sister Sara, thank you for believing in me and for pushing me to get back up when I was too tired to continue. Last but not least, thank you Mom and Dad - no amount of words could ever express how much you two mean to me. This dissertation is as much your work as it is mine. I love you.

Dedication

For Daadi Maa, Samina Aunty, David Langreth and all others who perished against cancer. Inna lilahi wa inna ilaihi rajioon.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background and Overview

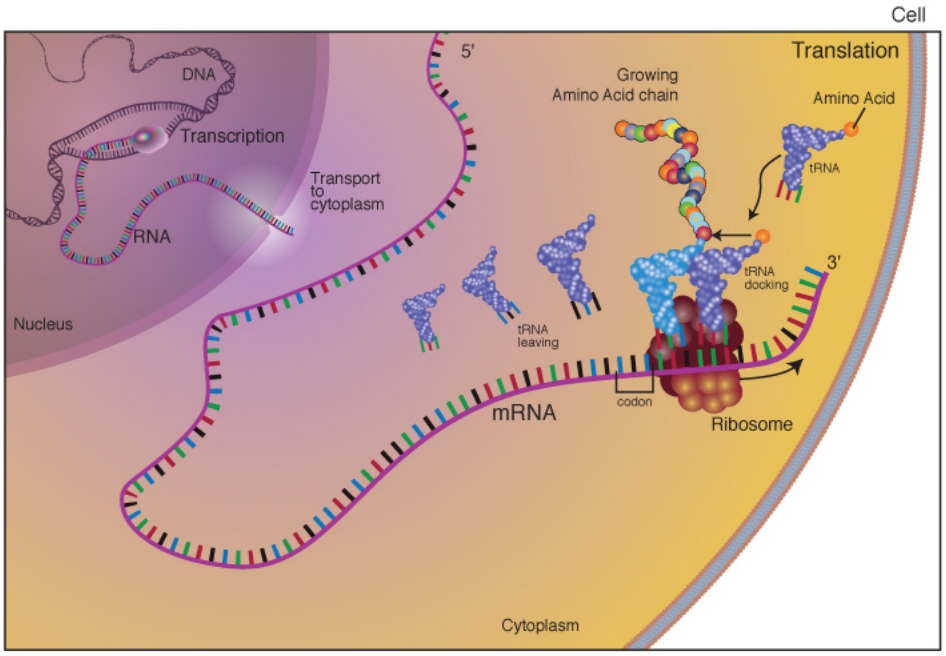

Figure 1.1: This figure illustrates the central tenet of molecular biology for the example of eukaryotic cells, or those containing a membrane-bound nucleus. Within the cellular nucleus, genetic sequences of DNA get transcribed into complementary messenger RNA (mRNA) sequences. These mRNA sequences then get transported into the cytoplasm, where ribosomes translate them into a corresponding polypeptide amino acid sequences with the help of transfer RNA (tRNA). This polypeptide sequence emerges as a protein. Figure courtesy of the National Human Genome Research Institute [37].

The seminal discovery of the double helical structure of deoxyribonucleic acid (DNA) established its role as the molecular carrier of genetic information [101, 24, 103]. The central tenet of molecular biology, as illustrated in Figure 1.1, dictates that specific genetic

sequences of DNA get transcribed into complementary RNA sequences, which then get translated into specific proteins that carry out specialized functions. Darwinian evolution then selects for functional proteins and genetic sequences with optimal evolutionary fitness.

However, genetics is not the sole determinant of the phenotype of an organism. Epigenetics is the study of heritable changes in a phenotype that arise due to changes beyond mutations in the DNA base sequences [97]. An example of this is in cellular differentiation of multicellular organisms, for example, humans. While different types of cells each contain the same core genetic sequence of DNA, the levels of expression for various proteins differ from cell type to cell type, resulting in additional sources of phenotypic variability.

Gene regulation occurs at all levels of expression of biological information. However, one of the most influential points at which it takes place is in the control of transcriptional initiation [100]. Specifically, different genetic sequences are recognized by transcription factors, and therefore transcribed into messenger RNA, at different rates.

1.1 Chromatin Remodeling and Eukaryotic Transcription

At a molecular level, the issue of determining how easily a particular genetic sequence can be transcribed relates to how easily accessible it is to transcriptional machinery. In eukaryotes, the approximately two meters of linear DNA must be packaged into chromosomes that have to fit in a subcellular nucleus that has a diameter of about 10 microns. The basic building block of this chromosomal packaging is a nucleosome, consisting of approximately 146 base pairs of DNA wrapped around a core of structural proteins called histones [56, 17], as illustrated in Figure 1.2. These nucleosomal subunits repeat themselves throughout the genome, connected by segments of DNA referred to as linker DNA. The nucleosomes can

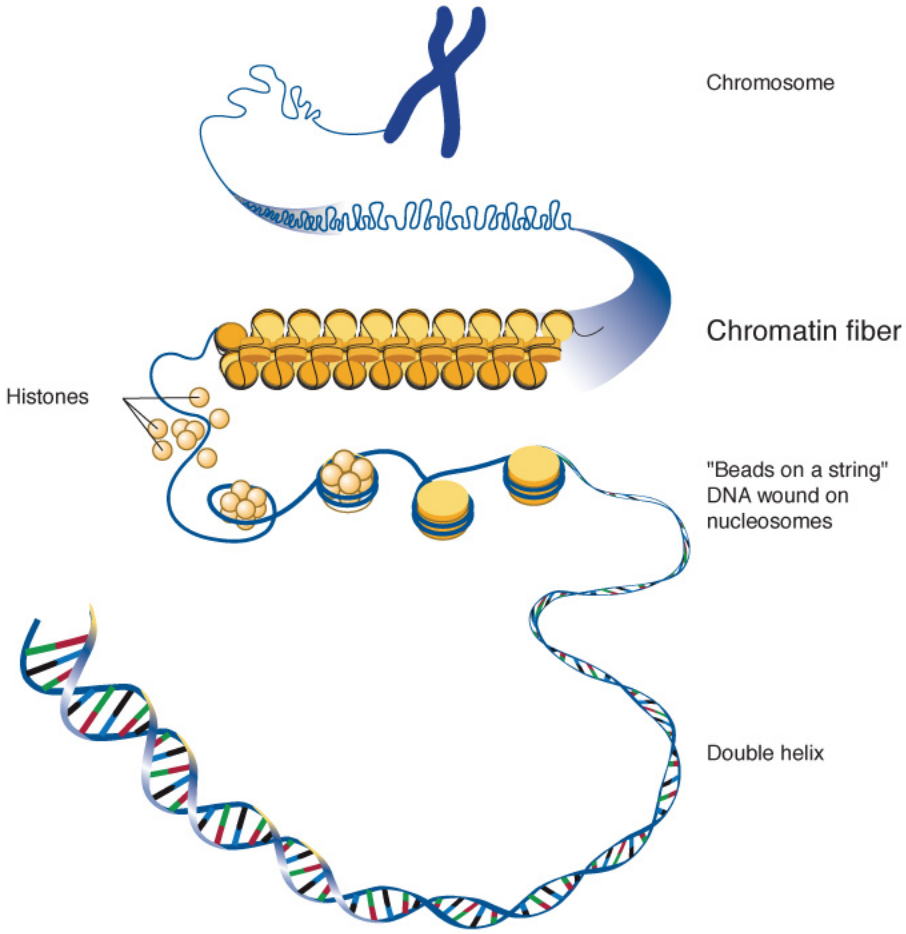

Figure 1.2: In eukaryotic cells, the DNA double helix is packaged into the cellular nucleus by structural proteins called histones. The basic building block of DNA packaging is a nucleosome, which consists of approximately 147 base pairs wrapped around an octamer of histones. These nucleosomes repeat throughout the DNA to form a fibrous structure called chromatin. During mitosis, chromatin condenses into chromosomes. Figure courtesy of the National Human Genome Research Institute [35].

thus be visualized as 'beads' lying on linker DNA 'strings'. A large collection of these beads-on-a-string makes up the bundled assembly of DNA and protein known as chromatin, the fibrous complex that condenses into chromosomes.

Nucleosomal DNA that is wrapped around histones is less easily accessible to binding by regulatory proteins, and is thus less transcriptionally active. Consequently, the density of nucleosomes in a particular region of DNA is a biophysical marker of the level of expression

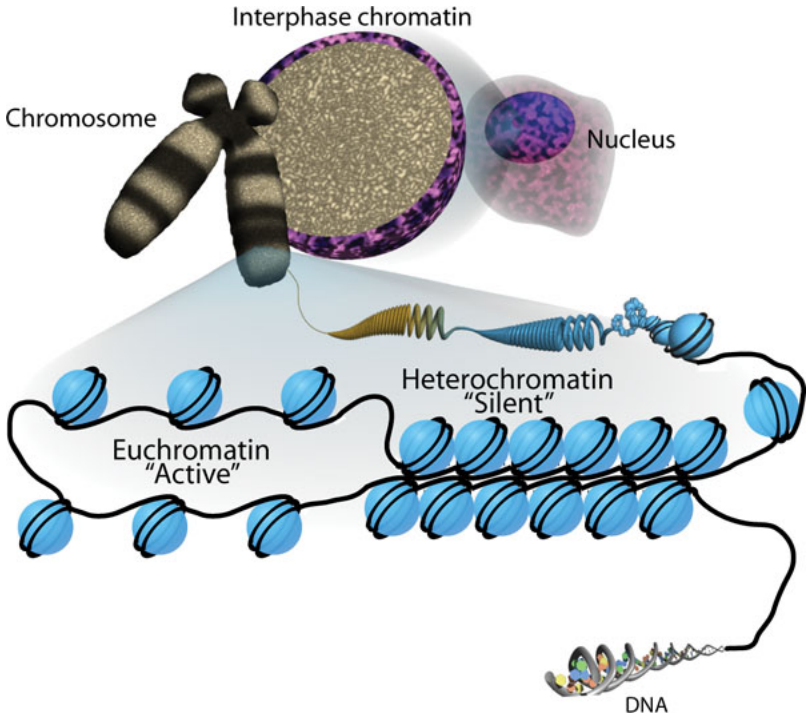

Figure 1.3: To a first approximation, the packaging of the chromatin fiber can be characterized by two distinct forms: 1) Euchromatin, which has a relatively low density of nucleosomes, leading to any genes in the region being more easily accessible to transcription, and therefore 'active', and 2) Heterochromatin, in which the nucleosomes are very tightly packaged, thereby silencing transcriptional expression. Figure courtesy of the Harvard Stem Cell Institute [86].

of genes located within that region [38, 11]. Chromatin is observed to generally transition between two major forms, illustrated in Figure 1.3: 1) Euchromatin, which has relatively low nucleosomal density, and is thus lightly packed and more easily accessible to regulatory machinery, and 2) Heterochromatin, in which the density of nucleosomes is relatively high, and the resulting tightening of the DNA packaging suppresses transcription. The process by which chromatin is dynamically modified to regulate access to transcriptional machinery is known as chromatin remodeling.

This dissertation presents theoretical investigations into two important biomolecular processes in chromatin remodeling: 1) DNA methylation of CG:CG base-pair steps, and 2)

Histone-DNA binding via the formation of arginine-phosphate salt bridges. Both of these are examples of mechanochemical processes, where a chemical modification, be it replacing hydrogen with a methyl group or binding an amino acid to a phosphate receptor, triggers a mechanical response, namely the activation of different kinds of elastic deformations. This situation is analogous to the way that different binding orientations of a ligand on a receptor molecule modify the conformation of the receptor.

In the course of this work, a general procedure for modeling the mechanical response of DNA to various biochemical modifications has been developed. It is through the lens of this procedure that this dissertation shall be structured. The procedure can be viewed as consisting of three basic steps:

1. Identify a local 'reacting region' of DNA that undergoes a specific biochemical modification, as well as the changing chemical environment of the atoms in the reacting region 'before' and 'after' the biochemical modification.

2. Perform principal component analysis on a collection of high-resolution protein-DNA crystal structures containing the reacting region in order to determine likely modes of structural deformation.

3. Use van der Waals density functional theory to calculate the energy landscape of the principal components both in the 'before' and 'after' state of the chemical environment, and extract the mechanical effect of the chemical change.

Details on the latter two steps shall be discussed in Chapter 2, and reports on the research applications and results shall be presented in Chapters 3 and 4, which are preprint versions of two manuscripts created in the course of this work [110, 111]. Chapter 5 completes the dissertation, with a summary and some concluding thoughts and discussions.

However, before computational methods can be applied to generate quantitative insight into a particular biomolecular process, the basic modeling framework of the process must be established, so that it is clear exactly how principal component analysis and density functional theory need to be applied. This is the purpose of the first step in the procedure, which shall be discussed in the reminder of this chapter. It shall be separately illustrated for both cytosine methylation and the formation of arginine-phosphate salt bridges between histones and DNA.

1.2 DNA Methylation of CG:CG Base-Pair Steps

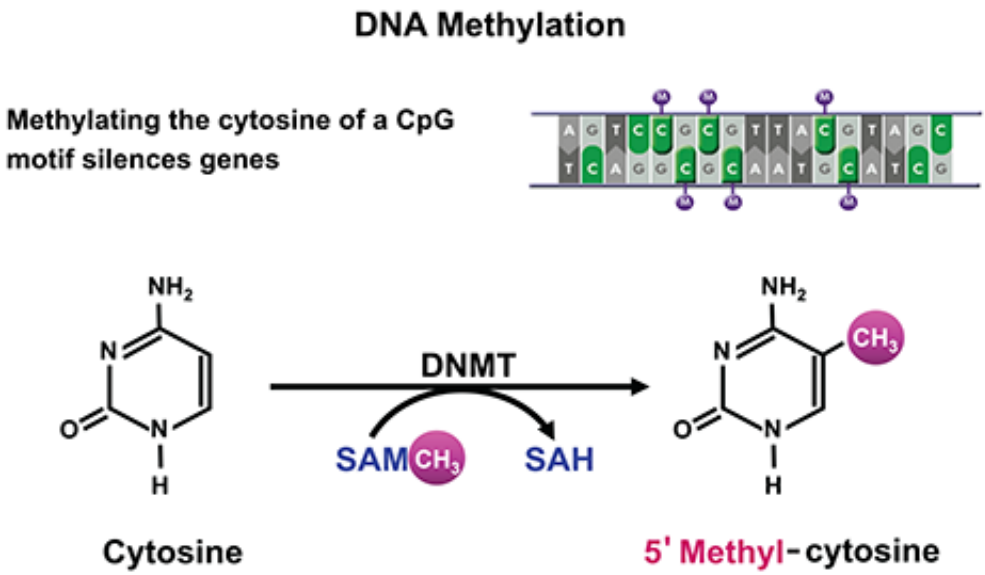

Figure 1.4: One of the most common and important epigenetic modifications is the methylation of cytosine at a CpG dinucleotide sequence motif. The enzyme DNA methyltransferase (DNMT) transfers a methyl group from an S-adenosyl methionine (SAM) donor to the C5′-carbon on the cytosines, replacing the hydrogen that is usually present. Genes in heavily methylated regions are observed to be transcriptionally silenced. Top figure adapted from the National Institute on Alcohol Abuse and Alcoholism [112], and bottom figure courtesy of Klaus Schulten [42].

One very important epigenetic modification is the 5-methylation of cytosine, demonstrated in Figure 1.4. This modification is observed to occur at a particular sequence motif, consisting of a 5′-CG-3′ sequence on one strand, with a complementary 5′-CG-3′ on the opposite strand. Methylation occurs when a DNA methyltransferase enzyme removes the hydrogens from the C5′ carbon on the cytosines, and replaces them with methyl groups that it transfers from an S-adenosyl methionine donor. DNA methylation is observed to correlate with regions of DNA where genes are silenced, or transcription is repressed [15, 7, 23, 16, 109, 106]. Thus, from a biophysical perspective, it is worthwhile to consider what effects the methyl groups have on the local deformability of DNA in their immediate vicinity.

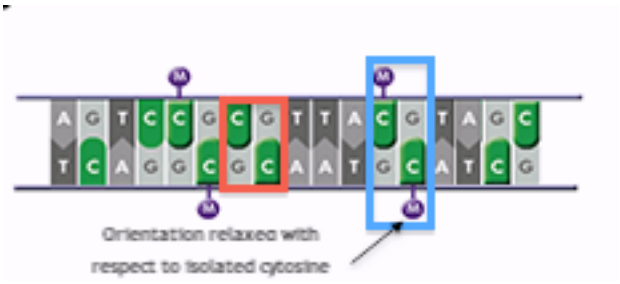

Figure 1.5: For the purposes of studying the methylation of cytosine, the relevant local 'reacting region' of interest is a CG:CG base-pair step, consisting of just the four nucleobases, with the sugar-phosphate backbone stripped off. The reference reacting region consists of CG:CG base-pair steps with no methylation (red box). The effects of the methyl group can be simulated by placing them at the original positions of the C5′-hydrogens, and relaxing the orientation of their covalent bond with respect to an isolated cytosine base. Then, under the approximation that local covalent interactions overwhelm non-covalent interactions, this orientation of the methyl group on the cytosine can then be transferred when performing calculations in the environment of a base-pair step.

The immediate vicinity of interest that we focus on is the CG:CG base-pair step, ignoring the presence of the DNA backbone or any other base-pairs. Locally, this is the part of the DNA that will most strongly feel the effects of methylation, relative to regions that are 'further down' from the methyl groups. Thus, this CG:CG base-pair step may be viewed

as our 'reacting region' of interest.

The changing biochemical 'state' of the CG:CG step is very easily identifiable. Before methylation, the CG:CG base-pair step is just a 'normal' base-pair step, with cytosine and guanine having their typical biochemical structure. After methylation, however, the site where the C5-attached hydrogens were originally located has been replaced with a carbon, and the three hydrogens attached to the carbon form a trigonal pyramidal structure. The orientation of the three hydrogens with respect to the C-C bond can be determined by a simple structural relaxation within cytosine, a feature which is built-in to Open Babel, the software used in this work to add methyl groups [75]. Since bonding interactions are generally much stronger than non-bonding interactions, this favored orientation is unlikely to change appreciably when the cytosine is in a base-pair step.

Thus, in summary, the reacting region is a CG:CG base-pair step, and the beginning and end states correspond to the situations where the C5′ carbons are attached to hydrogens and methyl groups in a structurally relaxed orientation, respectively. The reacting region is illustrated in Figure 1.5.

1.3 Arginine-Phosphate Salt Bridges between Histones and DNA

The nucleosome is stabilized by the formation of histone-DNA binding contacts. These contacts apply a mechanical load at specific sites of the DNA that hold the nucleosomal wrapping in place. There are several different histone-DNA contacts. However, a generally common pattern of binding involves the attraction of positively charged basic amino acids to the negatively charged sugar-phosphate backbone, as illustrated in Figure 1.6. Of the basic amino acids, the one that is observed to most frequently bind to the backbone in nucleosomes is arginine.

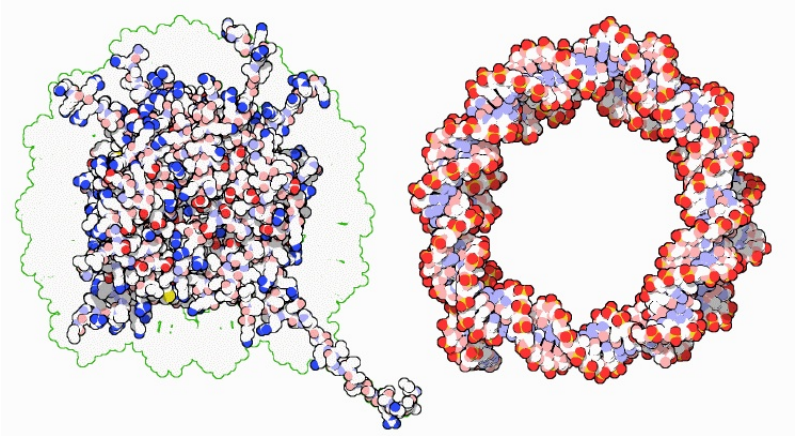

Figure 1.6: The basic building block of chromatin is a nucleosome, which consists of approximately 147 base pairs wrapped around a core set of histone proteins. The histones and DNA form a stable complex due to attractions between the basic, positively charged histone amino acids (colored blue on the left) and the acidic, negatively charged phosphate backbone (colored red on the right). Figures courtesy of the Research Collaboratory for Structural Bioinformatics [28].

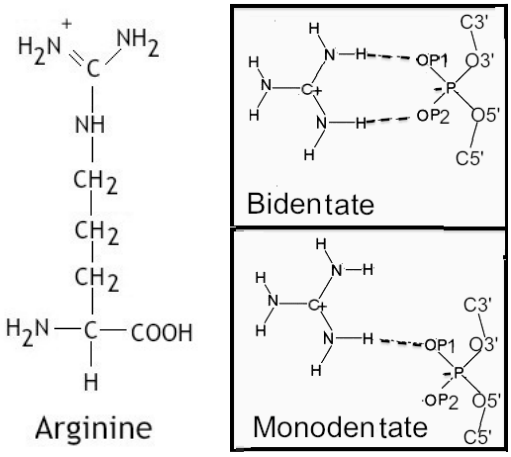

Figure 1.7: The most common mode of direct DNA-histone binding is through the formation of salt bridges between the basic arginine amino acids on the histone and the acidic phosphate group on the DNA backbone. This salt bridge entails a combination of electrostatic attraction and the formation of hydrogen bonds. Depending on whether one or two hydrogen bonds are formed, the salt bridge is classified as monodentate or bidentate, respectively.

In addition to the electrostatic attraction, arginine also forms hydrogen bonds between

its end-group nitrogens and the phosphate-group oxygens. This combination of electrostatics and hydrogen bonding is collectively referred to as a salt bridge [19]. Data analysis of nucleosomal crystal structures (reported in more detail later in this dissertation) further indicates that the arginine-phosphate salt bridge orientations tend to cluster into one of two dominant groups, illustrated in Figure 1.7: 1) Monodentate bridges, in which only one end-group nitrogen forms a hydrogen bond, and 2) Bidentate bridges, in which both do so.

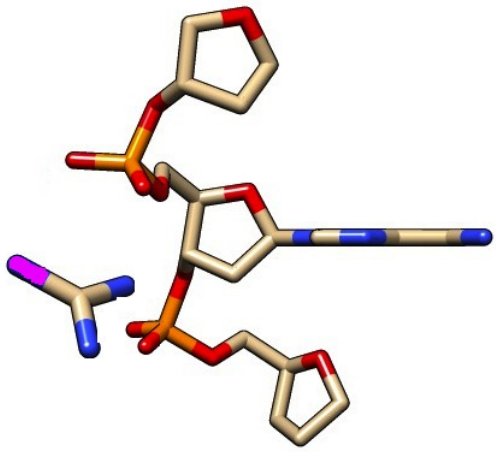

Figure 1.8: The model complex chosen as the 'reacting region' for the formation of histone-DNA salt bridges. The DNA consists of three deoxyribose sugars, with two linking phosphate groups and one central nucleobase included in order to study sequence-specificity. The end-group nitrogens on arginine (colored blue to contrast with the purple colored main-chain nitrogen) form a salt bridge with the phosphate backbone. This modifies the local elasticity and deformability of the DNA backbone. This image was created with Pymol [84].

For the purposes of this work, the main interest is in the effects of the arginine end-group on the deformation of the DNA backbone. Thus, in our choice of a reacting region, we would like to focus primarily on the sugar-phosphate backbone. Any region that is representative of the local flexibility of the region should, at a minimum, account for all interactions between neighboring backbone units. To meet this requirement, we select a combination of three deoxyribose sugars, with two intermediate phosphate groups, and one

central base to account for the sequence-dependence of backbone deformations. This model complex is displayed in Figure 1.8.

Thus, in summary, the reacting region is a combination of three deoxyribose sugars linked by two phosphate chains, with a central nucleobase. There are five different chemical states of interest for this sequence-dependent reacting region: the state in which no arginine is bound to it, and the four different combinations of ways that it can bind, namely, monodentate or bidentate attachment to the phosphate group located on the $5'$ or $3'$ side of the nucleobase.

Chapter 2

Theoretical and Computational Techniques

Having developed a general modeling framework for connecting biochemical epigenetic modifications to DNA mechanics, we now turn our attention to methods for actually inferring and quantitatively characterizing specific mechanochemical signals. In other words, we want to understand how the effective mechanical forces generated by changes to a reacting region's chemical environment act - in particular, we wish to describe precisely how chemical modifications deform the atoms in the reacting region.

This problem is complicated by the fact that even the simplest reacting regions of biological interest tend to have $N \sim 100$ atomic nuclei, and $3N$ corresponding positional degrees of freedom. Clearly, to explicitly track the motion of every single atomic coordinate is a hopelessly inefficient task for all but the most trivial systems.

Furthermore, the *in vivo* biological environment of DNA is a complex, messy place, with solvation, DNA-protein interactions, and a whole array of other biological effects at work. The atoms in any reacting region cannot be treated as if they were in 'vacuum'. The shapes that they adopt in living matter, and the dynamic behavior of those shapes, are the result of billions of years of evolution. The question, therefore, turns to how to figure out the most likely regions in configurational space that the atomic coordinates of DNA may be found *in vivo*. After identifying such a set of likely configurations, analysis of mechanochemical signaling is vastly simplified, as we can then focus our attention on this reduced subspace

that is more likely to be of immediate biological relevance.

In order to do this model reduction, however, we must have a set of measurable quantitative parameters describing the structural features of our two distinct reacting regions: 1) The CG:CG base-pair step, and 2) The three deoxyribose sugars linked by two phosphates, with a central nucleobase attached. We thus take a brief detour into the details of DNA structural biochemistry, where we will learn about some convenient experimentally measurable nucleic acid configurational variables.

2.1 DNA Structural Biochemistry

The basic building blocks of double-stranded DNA are base-paired nucleotides. A single nucleotide consists of a nucleobase connected to a five-membered deoxyribose sugar, with a phosphate group connecting neighboring sugars. The nucleobases on two separate single strands then form hydrogen bonds with each other, resulting in the canonical double-helical duplex. Structurally, it is useful for the purposes of this work to hone in on the particular structural features of the two distinct reacting regions: 1) The CG:CG base-pair step, and 2) The three deoxyribose sugars linked by two phosphates, with a central nucleobase attached.

2.1.1 Base-Pair Steps: The Rigid Body Representation

Nucleobases contain on the order of 10 atoms. However, these atoms are all covalently linked to each other, meaning their bond lengths and angles are, to a very good approximation, fixed [13]. Additionally, the deviations of the aromatic rings of the bases from co-planarity are observed to negligible, apart from methyl groups on thymine and methylated cytosine, which take a fixed orientation with respect to the base plane. The great simplification that this affords is that instead of having to keep track of every single atom, we can view bases

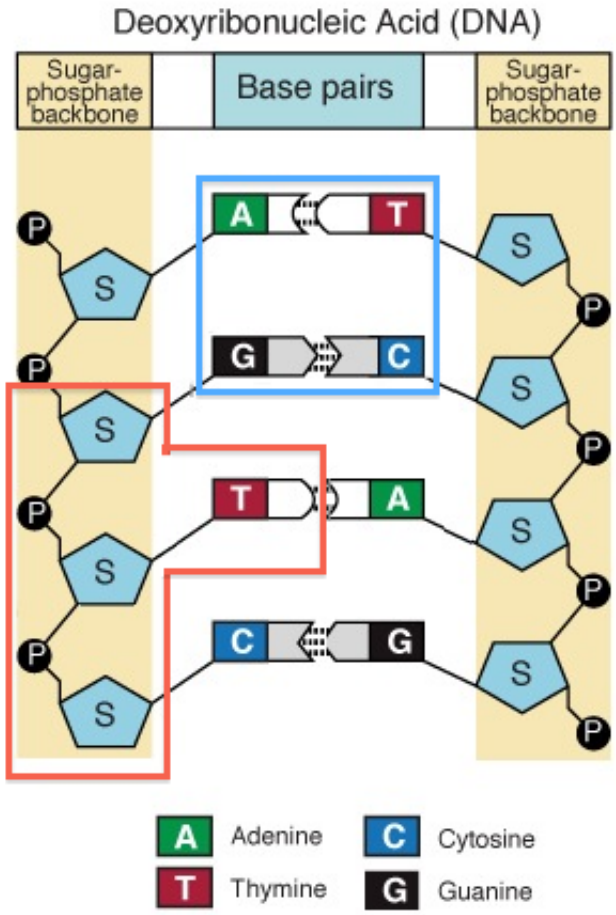

Figure 2.1: Displayed is a schematic diagram of the local chemical architecture of DNA at the level of individual nucleotides. Bases are approximated as rigid planes, and adjacent bases are covalently linked by a deoxyribose sugar ring and phosphate linkages, with the phosphorus atoms denoted by a black dot. For reference, the two reacting region sub-units relevant for this work are enclosed with a blue box (for the base-pair step used in the study of cytosine methylation) and a red box (for the sugar-phosphate backbone linkage used in the study of salt bridges). Details on the specific quantitative parameters describing each of these regions are expanded upon in the remainder of this chapter. Figure adapted from the National Human Genome Research Institute [36].

as rigid two-dimensional planes. The only possible motions of the atomic coordinates are those that correspond to collective translation or rotation of the entire base [71, 69].

Therefore, a base-pair step, such as the CG:CG steps of interest to us, can be viewed as a collection of four rigid planes. The relative position and orientation of any two rigid

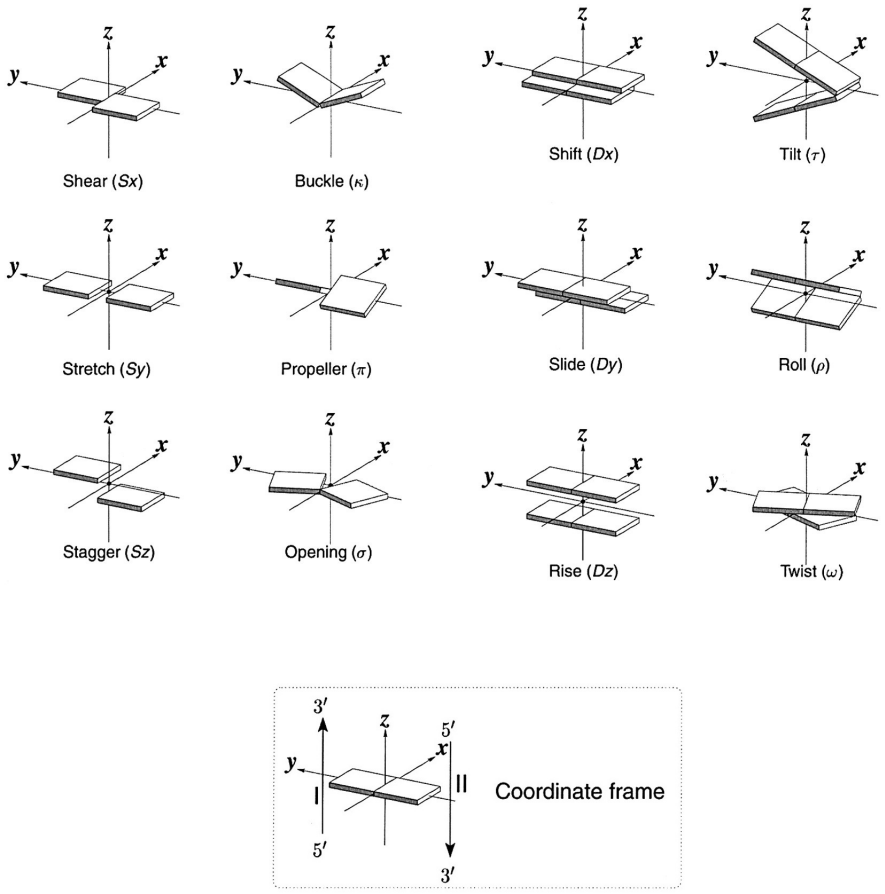

Figure 2.2: DNA nucleobases are very well approximated as rigid planes, and therefore, the configuration of a base-pair step, or four rigid planes, can be described by eighteen rigid body degrees of freedom: six each describing the relative translation and orientation of the base on one strand with respect to its hydrogen-bonded pair (leading to twelve total 'base-pair' parameters) and six describing the relative translation and orientation of stacked base pairs with respect to each other. Displayed above are positive values for each of the parameters. Figures adopted from Web3DNA [113].

bodies is described by six parameters: three corresponding to relative translation of the center of mass, and three corresponding to relative orientation. Therefore, if we fix one of the rigid planes, the relative configuration of the other three planes requires eighteen total rigid-body parameters. Thus, the atomic configuration of a base-pair step can be described by eighteen structural degrees of freedom.

In practice, we choose our rigid-body parameters in the following manner: six are necessary to describe the configuration of each individual base pair. These are known as base-pair parameters, and are labelled shear, stretch, stagger, buckle, propeller and opening. The remaining six describe the relative translation and orientation of the two base pairs with respect to one another. These base-pair step parameters are denoted rise, slide, shift, tilt, roll and twist. For further details concerning the exact specification and computation of atomic coordinates from base-pair and base-pair step parameters, the reader is referred to the literature on 3DNA, a software package developed by the Olson group for analysis and reconstruction of nucleic acid structures [54, 55, 113].

2.1.2 The Sugar-Phosphate Backbone: Dihedral Angles and Sugar Puckers

The description of the configuration of the sugar-phosphate backbone, like that of the base-pair steps, is also simplified by the observation that covalently bonded atoms, tend to have negligible variation in their bond lengths and angles [26]. A notable exception to this is the deoxyribose sugar, which we will discuss momentarily. However, the connecting sugar-phosphate linkages, as well as the covalently attached central nucleobase, do not share this complication. All of their bond lengths and angles can be fixed at experimentally determined values, as built-in to 3DNA, and this is the protocol adopted in this work.

The coordinates of the non-deoxyribose atoms can thus be specified entirely by dihedral angles, or angles between adjacent covalently-bonded planes, as illustrated in Figure 2.3. In other words, the dihedral angle between connected atoms A-B-C-D is given by the angle between the plane formed by A-B-C and that formed by B-C-D. A set of dihedral angles ϵ, ζ, α, β and γ describe the atomic configuration of the phosphate chain connecting two deoxyribose sugars. Additionally, a dihedral angle χ sets the orientation of the central

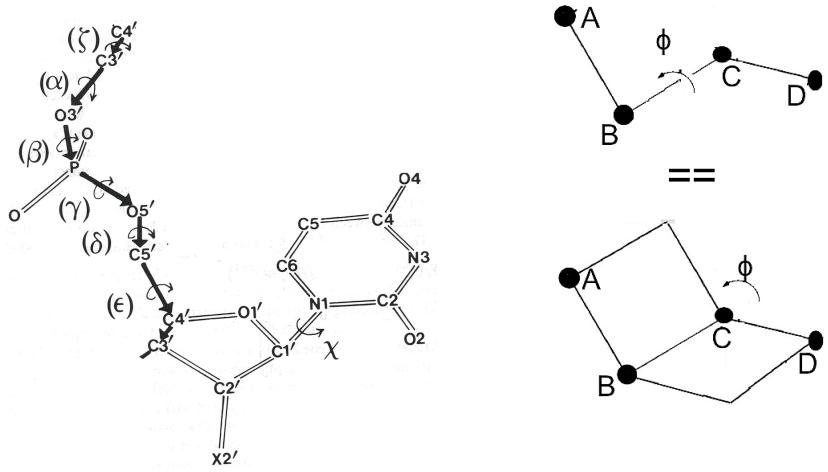

Figure 2.3: The configuration of the covalent chain linking adjacent deoxyribose sugars is specified by a set of five dihedral angles. Dihedral angles are defined for groups of four atoms A-B-C-D, and are equal to the angle that the plane of atoms A-B-C makes with respect to the plane of atoms B-C-D. An additional dihedral angle is used to set the orientation of a base with respect to the deoxyribose sugar. The sign convention is such that perfectly coplanar atoms have zero dihedral angle. Image adapted from Olson [68].

nucleobase with respect to its attached deoxyribose sugar [83].

The only structural degrees of freedom remaining are the coordinates of the five-membered deoxyribose sugar rings themselves. Their description is complicated by the fact that the deoxyribose is a closed structure. Thus, while it is possible for the bond lengths to remain fixed, it is not possible to also keep all bond angles fixed without constraining the sugar to lie in a rigid form, unable to move. Any adjustment of dihedral angles while keeping bond lengths fixed inevitably must change bond angles due to the non-trivial topology of the sugar ring.

Decades of research have identified that the space of observed conformations of the sugar rings, to a very good approximation, lies along a 'pseudorotation' cycle, as shown in Figure 2.4. Movement along this cycle changes the 'puckering' of the ring, or relative co-planarity

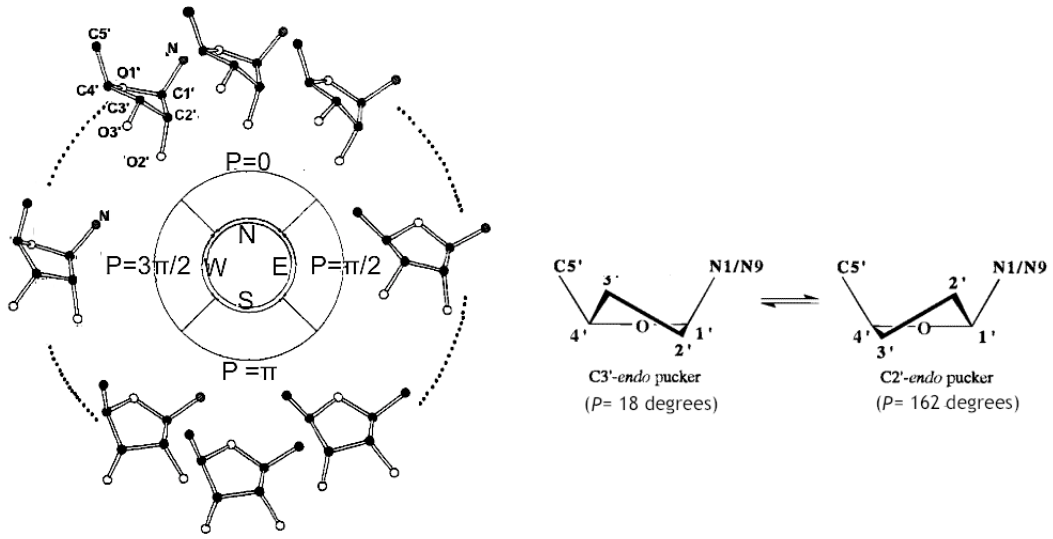

Figure 2.4: The configuration of the deoxyribose sugar ring is in general more complex than that of the rest of the DNA polymer. This is due primarily to the closed ring topology of the furanose ring. Thus, if covalent bond lengths are to remain fixed, the additional geometrical constraints imply that it is not possible to freely change the dihedral angles without also changing some bond angles. In practice, however, the configuration of the sugar ring in nucleic acids is well approximated as lying along a pseudorotation pathway, specified by a single phase angle P. Motion along this pathway changes the sugar puckering, or planarity of the atoms in the deoxyribose ring. Figures adapted from Olson [67].

of different atoms in the ring. In 1972, Altona and Sundaralingam [1] showed how motion along this pathway can be characterized by two numbers: a sugar pucker amplitude τ_m and a pseudorotation phase angle P. Additionally, in nucleic acids, the amplitude τ_m is observed to vary negligibly. Therefore, to a first approximation, the conformation of the deoxyribose sugar is set by the pseudorotation phase angle P. In 1982, Olson [72, 67] developed a knowledge-based procedure for reconstructing the Cartesian coordinates of the deoxyribose sugar for a given phase angle, and this procedure was used in the present work.

2.2 The Nucleic Acid Database and Principal Component Analysis

Having identified a set of experimentally measurable parameters to characterize local DNA architecture, we return to our original goal of trimming the vast conformational space to

a reduced subspace of biologically relevant molecular configurations. To proceed with this reduction of the configurational subspace requires two things: 1) A sufficiently large and diverse dataset of DNA structures, from which we can glean general trends and patterns, and 2) A procedure for statistically analyzing this dataset in order to extract a simplified, reduced set of variables that captures most of the variability in the data. The former is provided by the Nucleic Acid Database, and the latter by principal component analysis.

2.2.1 Nucleic Acid Database

The Nucleic Acid Database, or NDB [5], was initiated at Rutgers University in 1992 by Drs. Helen Berman and Wilma K. Olson of Rutgers University, and David Beveridge of Wesleyan University. It is an up-to-date database of structural information about nucleic acids, including X-ray studies of naked DNA (unbound to any protein) and protein-nucleic acid complexes, as well as NMR structures. An illustrative snapshot is displayed in Figure 2.5. In our work we focus on high-resolution X-ray crystallographic structures of protein-DNA complexes.

The data used for this work consists of a working dataset of 239 protein-DNA crystal complexes of 2.5 Angstroms or better resolution, taken from the NDB. This dataset was generated by Dr. Yun Li in his Ph.D. dissertation [51]. A vitally important task in the generation of the data is to ensure that the structures are non-redundant. Specifically, several protein-DNA crystal structures in the databases are nearly identical sequentially and/or structurally. Having too many data points in one particular protein family or DNA genomic sequence results in sampling bias. This is undesirable since the goal of the present work is to infer a complete landscape of possible structural deformations *in vivo*. Including too many examples of one kind skews the sampling and results in overrepresenting particular

Figure 2.5: Displayed is a snapshot of the Nucleic Acid Database (NDB) website. The NDB provides a repository of experimental nucleic acid structures, of which protein-DNA crystal complexes are selected for analysis. For any particular structure, all necessary structural quantities, such as base-pair and base-pair step parameters, torsion angles, and sugar puckering coordinates are generated from the 3DNA software package and displayed in the database.

regions of configurational space, while potentially missing other important regions. Further details regarding the generation of the working dataset can be found in Dr. Yun Li's Ph.D. dissertation, as well as in the Supplemental Information of the manuscript [110].

2.2.2 Principal Component Analysis

Having identified a relevant collection of variables to describe the structure of our reacting regions, the question turns to how to further reduce this collection to an even smaller set that captures the most essential features. This task is achieved by principal component analysis (PCA) [64, 107, 34, 12].

The general idea behind PCA is to take a high-dimensional dataset and project it into a linear subspace of lower dimensionality which captures as much of the variance in the data as possible. By dimensionality, we mean here the number of parameters needed to describe a single data point. For example, N atoms will have, without further constraint, $3N$ coordinate parameters to describe their total microscopic configuration, and so the

dimensionality will be $3N$.

However, in general, parameters are not all independent. For example if the N atoms were each bonded to another atom in the set, in order to form $N/2$ diatomic molecules, then only $3N-3$ coordinates could move independently, and thus the data would be more efficiently described using a representation in terms of the center of mass coordinates and rotational orientation of each of the $N/2$ diatomic molecules. In other words, a scatter plot of the data points in these parameters would capture most of the statistical fluctuations that a scatter plot in terms of the complete $3N$ atomic coordinates would capture.

To make this more precise, let us take a concrete example of an initially two-dimensional dataset, with two random variables x and y. A single data point is specified by its values of x and y. However, we could just as well describe it in terms of its projection into any arbitrary linear orthonormal basis spanning the space of x and y, for example,

$$x^{'} = \cos(\theta)\ x + \sin(\theta)\ y \tag{2.1}$$

$$y^{'} = \sin(\theta)\ x - \cos(\theta)\ y \tag{2.2}$$

where θ is some angle. The goal of PCA is to select the value of θ, or in other words the choice of parametrizations $x^{'}$ and $y^{'}$, in which as much of the statistical variability of the data is captured by its projection onto a single component.

To make this more explicit, suppose the average value of the data in $(x,\ y)$ space is given by $(\overline{x},\ \overline{y})$ respectively. Then, for the i'th data point $(x_i,\ y_i)$, the squared distance from the average σ_i^2 is given by

$$\sigma_i^2 = (x_i - \overline{x})^2 + (y_i - \overline{y})^2 \tag{2.3}$$

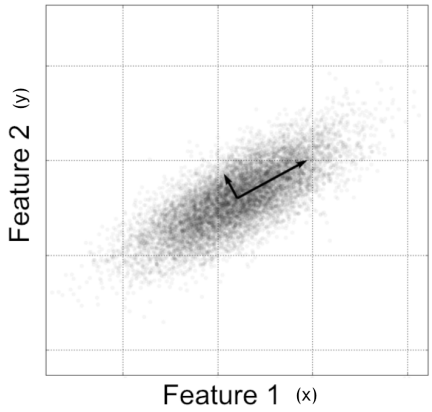

Figure 2.6: In principal component analysis, an original set of variables, with significant statistical variance in multiple directions of parameter space, is traded for a reduced set of principal components from which statistical deviation is minimized. In this example of two random variables x and y, expressing the data in terms of their projection onto the $y = x$ line captures the highest possible fraction of the total variance. Any deviation from this line, or projection onto $y = -x$, is minimized.

Thus, the total variance of an entire dataset of N data points is given by

$$\sigma^2 = \sum_i^N \frac{\sigma_i^2}{N} = \sum_i^N \frac{((x_i - \overline{x})^2 + (y_i - \overline{y})^2)}{N} \tag{2.4}$$

In general, the variance contribution from x will not equal that from y,

$$\sum_i^N (x_i - \overline{x})^2 \neq \sum_i^N (y_i - \overline{y})^2. \tag{2.5}$$

Additionally, if the data are reexpressed in terms of a rotated orthonormal basis $(x^{'}, y^{'})$, the variance distribution between $x^{'}$ and $y^{'}$ will be different from the original variance distribution between x and y. However, for a particular choice of $(x^{'}, y^{'}) = (x_{princ}, y_{princ})$, a maximum possible amount of the total variance will be embedded in the variance of just

one of the components, let us take it to be x_{princ} without loss of generality,

$$\sum_{i}^{N}(x_{princ_i} - \overline{x}_{princ})^2 \geq \sum_{i}^{N}(x_i - \overline{x})^2 \quad \forall\ x_i \tag{2.6}$$

Equivalently, in this optimal representation the remaining orthogonal axis y_{princ} will carry the smallest possible amount of the total variance that it can. This principal axis x_{princ}, therefore, is the best possible approximation for having all of the data points lie entirely on a single line, with fluctuations on axes orthogonal from that line minimized. This argument is displayed graphically in Figure 2.6.

This argument can be straightforwardly generalized to apply to more complex multivariate datasets of arbitrarily large dimension D. The purpose of PCA is to identify a reduced set of $N <$ D principal orthogonal axes that capture the highest possible fraction of the covariance in D-dimensional data space. Equivalently, fluctuations in the remaining N - D dimensions are minimized. Thus, PCA is an invaluable tool for the reduction of large datasets to smaller, more manageable forms.

2.3 Electronic Structure Calculations

Once the dominant configurational landscapes of a reacting region have been identified, the task turns to isolating the effects of specific chemical modifications on the energies of deformation along these landscapes. In order to do this, we require a theoretical technique to extract reliable energies for a given configuration of atomic nuclei. Such a method should, ideally, maintain sufficient chemical accuracy for us to be reasonably confident in any results, yet computationally inexpensive enough for it to be able to be applied to biologically relevant reacting regions of DNA. The method used in this work is density functional theory (DFT).

DFT is an *ab initio* method for quantum-mechanical modeling of real materials. In contrast to quantum chemical methods, which work with the many-electron wavefunction, DFT works with the ground-state electronic charge density, which has significantly fewer degrees of freedom. This is the main reason behind DFT's low computational cost relative to quantum chemical techniques. However, DFT has a major shortcoming: there is no method to systematically improve its accuracy, and thus, figuring out how to usefully apply it to different problems is both an art and a science.

It is this shortcoming that has been the primary reason for the historical relative dearth of applications of DFT to biological systems. Specifically, the 'soft' matter that is ubiquitious in living matter is heavily influenced by non-local London dispersion forces. Traditional density functional theory techniques do not successfully account for these interactions, and the problem has only recently been solved with the advent of van der Waals density functional theory (vdW-DFT), pioneered by the late Dr. David Langreth of Rutgers.

These improved functionals, along with other notable algorithmic advances such as the development of linear-scaling DFT for biological macromolecules in solution by Dr. Darrin York, now of Rutgers [108], have in recent years extended the applicability of DFT to biological matter for the first time. In regards to the latter, it is worthwhile to point out that while this dissertation only focuses on the advances allowed by the novel van der Waals functional, there is also much to be gained in the future by integrating it with other state-of-the-art DFT methodologies.

2.3.1 The Schrodinger Equation and the Born-Oppenheimer Approximation

The starting point for a quantum-mechanical description of materials is the Schrodinger equation [3] for a collection of electrons and nuclei. To make some progress towards simplifying the problem, we employ the Born-Oppenheimer approximation: since nuclei are much more massive than electrons, $m_N >> m_e$, nuclear motion is much slower than electronic motion. Consequently, for the purposes of analyzing the electronic wavefunction, the nuclear coordinates can be well approximated as fixed classical point particles.

With these simplifications, the Schrodinger equation for a collection of $i = 1, 2, 3, ... N$ electrons, for a fixed set of atomic nuclear positions, is

$$(-\sum_{i=1}^{i=N} \frac{\hbar^2}{2m_e}\nabla_i^2 + V_{ee}(\vec{r_i}) + V_{en}(\vec{r_i};\vec{R_N}))\ \psi(\vec{r_i};\vec{R_j}) = E_e(\vec{R_j})\ \psi(\vec{r_i};\vec{R_N}) \tag{2.7}$$

Here, $\vec{r_i}$ refers to the coordinates of the electrons, and $\vec{R_N}$ refers to the fixed positions of the classical atomic nuclei. The first term is the kinetic energy of the electrons, and the potential terms represent Coulombic electromagnetic interactions between all of the electrons and nuclei. In effect, the electrons have generated an additional energy term for the *nuclei*, $E_e(\vec{R_N})$. This can be made more explicit by seeing that the total energy of a given configuration of nuclear coordinates $\vec{R_n}$ is $E_n(\vec{R_N}) = T_n(\vec{R_N}) + V_n(\vec{R_N})$, where T_n is the nuclear kinetic energy and V_n is the total potential energy,

$$V_n(\vec{R_N}) = V_{nn}(\vec{R_N}) + E_e(\vec{R_N}) \tag{2.8}$$

where V_{nn} is the usual Coulombic repulsion between positively charged nuclei, and E_e is an effective energy contribution arising from the electrons and their interactions with each

other and with the nuclei.

2.3.2 Density Functional Theory

While the problem has been slightly simplified, the solution of the electronic Schrodinger equation for a general system of multiple electrons is still largely analytically intractable. Thus, it becomes useful to turn to methods for approximate solution. The field of computational quantum chemistry [99, 3] is built upon developing advanced techniques to solve the Schrodinger equation. There is, in general, a speed-error tradeoff: the faster a calculation takes, the less accurate the solution will be.

However, wavefunction-based techniques are not the only way to go about solving the many-body problem. There is an alternative approach, called density functional theory (DFT), which is based upon physical arguments that recast the many-body problem into a different form. Specifically, DFT opts to work with the total electronic charge density $\rho(\vec{r})$ of N electrons, instead of the complete electronic wavefunction $\chi(\vec{r_i};\vec{R_j})$. For a system of N electrons, this density is given by

$$\rho(\vec{r}) = N \int d^3\vec{r_2} \int d^3\vec{r_3} \; ... \int d^3\vec{r_N} \; |\chi(\vec{r},\vec{r}_2,\vec{r}_3,...\vec{r}_N)|^2 \tag{2.9}$$

where we are implicitly including the dependence of the wavefunction on nuclear coordinates, and also writing out all the variables: $\chi(\vec{r_i};\vec{R_j}) = \chi(\vec{r}_1,\vec{r}_2,....\vec{r}_N)$. The principal computational advantage of working with charge densities is that whereas the N-electron wavefunction requires $3N$ coordinate variables for its description, charge densities only requires 3. This significantly reduces the computational resources, namely time and computer memory, required to perform numerical calculations.

The Hohenberg-Kohn Theorems

The starting points of density functional theory are two theorems called the Hohenberg-Kohn theorems [33, 43]. The first Hohenberg-Kohn theorem states that

First Hohenberg-Kohn Theorem: For a system of N interacting electrons, in the presence of an external potential $v_{ext}(\vec{r})$, the ground state electronic charge density $\rho_0(\vec{r})$ uniquely determines $v_{ext}(\vec{r})$, up to an additive constant.

To prove this theorem, we use the variational principle of quantum mechanics, which states that for any Hamiltonian $\hat{H}$, the energy of the ground state energy $|\psi_0\rangle$ is less than or equal to the energy expectation value of any other possible state, $\langle\psi_0|\hat{H}|\psi_0\rangle \leq \langle\psi|\hat{H}|\psi\rangle$ for all $|\psi\rangle$. To simplify things slightly, let us assume a non-degenerate electronic ground state, which holds for a fairly wide class of systems. This restriction implies that the variational equality can only possibly hold for $|\psi\rangle = |\psi_0\rangle$.

The Hamiltonian of an interacting electron system for an external potential V_{ext} is $\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext}$. Suppose we have two external potentials V^1_{ext} and V^2_{ext} that differ by more than a constant, with corresponding Hamiltonians $\hat{H}_1$ and $\hat{H}_2$ and ground states $|\psi^1_0\rangle$ and $|\psi^2_0\rangle$, respectively. The ground states must necessarily be different, $|\psi^2_0\rangle \neq |\psi^1_0\rangle$, since the potentials differ by more than a constant. Let us label the ground state energies $E_1 = \langle\psi^1_0|\hat{H}_1|\psi^1_0\rangle$ and $E_2 = \langle\psi^2_0|\hat{H}_2|\psi^2_0\rangle$. By the variational principle, and taking advantage of non-degeneracy, we may note that

$$E_1 = \langle\psi^1_0|\hat{H}_1|\psi^1_0\rangle < \langle\psi^2_0|\hat{H}_1|\psi^2_0\rangle \tag{2.10}$$

$$E_2 = \langle\psi^2_0|\hat{H}_2|\psi^2_0\rangle < \langle\psi^1_0|\hat{H}_2|\psi^1_0\rangle \tag{2.11}$$

Now, noting that $\hat{H}_1 = \hat{H}_2 + \hat{V}_1^{ext} - \hat{V}_2^{ext}$, the right-hand sides of the inequality can be rewritten as

$$\langle\psi_0^2|\hat{H}_1|\psi_0^2\rangle = \langle\psi_0^2|\hat{H}_2 + \hat{V}_1^{ext} - \hat{V}_2^{ext}|\psi_0^2\rangle = E_2 + \langle\psi_0^2|\hat{V}_1^{ext} - \hat{V}_2^{ext}|\psi_0^2\rangle \tag{2.12}$$

$$\langle\psi_0^1|\hat{H}_2|\psi_0^1\rangle = \langle\psi_0^1|\hat{H}_1 + \hat{V}_2^{ext} - \hat{V}_1^{ext}|\psi_0^1\rangle = E_1 + \langle\psi_0^1|\hat{V}_2^{ext} - \hat{V}_1^{ext}|\psi_0^1\rangle \tag{2.13}$$

Putting everything together, we realize

$$E_1 < E_2 + \langle\psi_0^2|\hat{V}_1^{ext} - \hat{V}_2^{ext}|\psi_0^2\rangle \tag{2.14}$$

$$E_2 < E_1 + \langle\psi_0^1|\hat{V}_2^{ext} - \hat{V}_1^{ext}|\psi_0^1\rangle \tag{2.15}$$

However, the last terms are integrals of potential differences multiplied by the charge densities

$$\langle\psi_0^2|\hat{V}_1^{ext} - \hat{V}_2^{ext}|\psi_0^2\rangle = \int d^3\vec{r}\,(V_1^{ext}(\vec{r}) - V_2^{ext}(\vec{r}))\;\rho_0^1(\vec{r}) \tag{2.16}$$

$$\langle\psi_0^1|\hat{V}_2^{ext} - \hat{V}_1^{ext}|\psi_0^1\rangle = \int d^3\vec{r}\,(V_2^{ext}(\vec{r}) - V_1^{ext}(\vec{r}))\;\rho_0^2(\vec{r}). \tag{2.17}$$

Therefore, we see that if V_1^{ext} and V_2^{ext} give rise to equivalent charge densities $\rho_0^1 = \rho_0^2$, then adding the two inequalities would yield the contradictory result $E_1 + E_2 < E_1 + E_2$. Thus, the ground state electronic charge density is uniquely determined by the external potential, up to a trivial additive constant. The proof can be readily extended to deal with non-degenerate cases [60].

One of the consequences of this result is the second Hohenberg-Kohn theorem,

Second Hohenberg-Kohn Theorem: There exists a universal functional for the electronic energy E_e in terms of the charge density, $E_e = E_e[\rho(\vec{r})]$, and the exact ground state energy

and charge density are given by minimizing this functional.

This follows from the fact that the Hamiltonian $\hat{H}$ completely specifies observable properties of the system, including the total energy. The Hamiltonian, in turn, can be characterized by the external potential, which by the first theorem, is determined by the charge density. Thus, observable properties like the total energy, as well as each of its individual contributions, can be viewed as functionals of the charge density. Explicitly,

$$E_e = E_e[\rho(\vec{r})] = T[\rho(\vec{r})] + E_{ee}[\rho(\vec{r})] + \int d^3\vec{r}\, V_{ext}(\vec{r})\ \rho(\vec{r}) \tag{2.18}$$

Here, T is the electron kinetic energy, E_{ee} is the energy of electron-electron interaction and V_{ext} is the interaction of the electrons with the external potential, which in our present case corresponds to the Coulombic attraction due to the nuclei. By the variational principle, the ground state energy, which corresponds to E_e evaluated at the ground state charge density $\rho_0(\vec{r})$, is less than or equal to the energy of any other possible charge density $\rho(\vec{r})$, $E^{(0)} = E_e[\rho^0(\vec{r})] \leq E_e[\rho(\vec{r})]$. Therefore, minimizing the functional yields the ground state energy and charge density.

Kohn-Sham Theory and Exchange-Correlation

In principle, solving the many-body Schrodinger equation has been reduced to minimizing the exact Hohenberg-Kohn electronic energy functional $E_e[\rho(\vec{r})]$. However, the catch is that in practice, the exact form of this many-body functional is unknown. Nevertheless, we can make remarkable progress by using an elegant argument devised by Kohn and Sham. The Kohn-Sham procedure is to apply the Hohenberg-Kohn theorem to a set of *non-interacting* electronic wavefunctions $\psi^{KS}(\vec{r})$, called Kohn-Sham orbitals, which give

rise to the *interacting* many-body ground state charge density, $\rho(\vec{r}) = \sum_{i=1}^{N} |\psi_i^{KS}(\vec{r})|^2$. Then, the exact energy functional $E_e[\rho(\vec{r})]$ can be expressed in terms of the kinetic energy of the Kohn-Sham orbitals, T_{KS} and the remainder, which we can conveniently relabel to be the Kohn-Sham potential energy $E_{KS} = E_e - T_{KS}$,

$$E_e[\rho(\vec{r})] = -\frac{\hbar^2}{2m_e} \sum_{i=1}^{N} \int d^3\vec{r}\ \psi_i^{KS*}(\vec{r})\ \nabla_i^2\ \psi_i^{KS}(\vec{r}) + E_{KS}[\rho(\vec{r})]. \tag{2.19}$$

Variational minimization of E_e then yields the Kohn-Sham equations, which are essentially Schrodinger equations for the fictitious Kohn-Sham orbitals $\psi_{KS}(\vec{r})$ in an effective Kohn-Sham potential $V_{KS}(\vec{r}) = \frac{\delta E_{KS}[\rho(\vec{r})]}{\delta \rho(\vec{r})}$

$$(-\frac{\hbar^2}{2m_e}\nabla^2 + V_{KS}(\vec{r}))\ \psi_i^{KS}(\vec{r}) = E_i\ \psi_i^{KS}(\vec{r}). \tag{2.20}$$

We can conveniently decompose the Kohn-Sham potential energy E_{KS}, and thus the Kohn-Sham potential V_{KS}, into three physically distinct pieces: 1) the contribution arising from the 'external' potential due to the nuclei,

$$E_{ext} = \int d^3\vec{r}\ V_{ext}(\vec{r})\ \rho(\vec{r}) \tag{2.21}$$

2) the Hartree contribution arising from electrostatic Coulomb repulsion

$$E_{Hartree} = \frac{1}{8\pi\epsilon_0} \int d^3\vec{r} \int d^3\vec{r}'\ \frac{\rho(\vec{r})\rho(\vec{r'})}{|\vec{r} - \vec{r'}|} \tag{2.22}$$

and 3) the remainder, which we label the exchange-correlation energy E_{xc}, and which we

can take being the consequence of an effective exchange-correlation potential V_{xc}

$$E_{xc} = E_{KS} - E_{ext} - E_{Hartree} = \int d^3\vec{r}\, V_{xc}(\vec{r})\ \rho(\vec{r}) \tag{2.23}$$

The exchange-correlation energy is the major bottleneck of density functional theory. There is no systematic methodology to improving its accuracy, and developing approximations for it is as much an art as a science. Historically, one of the most common classes of approximation are the local-density approximation (LDA) [9, 79, 78], in which the exchange-correlation energy at a point only depends on the density at that point

$$E_{xc}^{LDA}[\rho] = \int d^3\vec{r}\, \rho(\vec{r})\ \epsilon_{xc}(\rho(\vec{r})) \tag{2.24}$$

where ϵ_{xc} is the energy per particle of a homogeneous electron gas with a particular charge density ρ. The next rung in the ladder of exchange-correlation functionals is the generalized gradient approximation (GGA) [4, 47, 77, 32], which allows ϵ_{xc} to depend not only on the local density $\rho(\vec{r})$, but also on gradients of the density $\nabla\rho(\vec{r})$.

$$E_{xc}^{GGA}[\rho] = \int d^3\vec{r}\, \rho(\vec{r})\ \epsilon_{xc}(\rho(\vec{r}), \nabla\rho(\vec{r})) \tag{2.25}$$

These *semi-local* functionals, while an improvement to LDA, fail to capture long-range non-covalent interactions, in particular the London dispersion forces that are ubiquitious in biology. Historically, this limitation has prevented the widespread application of DFT to biological systems. However, in recent decades, several novel methods have been introduced for the inclusion of dispersion forces, extending the realm of DFT to biological matter for the first time. In this work, we focus on the development, by David Langreth and

co-workers, of the *non-local* van der Waals density functional, vdW-DF, and its successor vdW-DF2. Nevertheless, it is worthwhile to note that this is not the only way of dealing with dispersion in DFT. An exhaustive overview of all proposed methodologies is beyond the scope of this work, but a recent review by Kolb and Thonhauser [44] highlights a few of the most promising techniques, including the density functional theory with added empirical dispersion correction (DFT-D) method of Grimme [30] , and the related DFT+vdW technique of Tkatchenko and Scheffler [93].

Van der Waals Density Functional Theory

Complete details regarding the rigorous derivation of the precise functional form can be found in the original literature [18, 91, 14, 82, 48]. Here, we focus on some essential qualitatively important properties of the functional that are of particular relevance to soft and biological matter.

London dispersion forces arise from the correlated interactions between *instantaneous* quantum fluctuations of the charge density, in contrast to the classical electrostatic interactions between *permanent* charge density distributions. From the point of view of density functional theory, London dispersion forces are, at their heart, effects due to *non-local correlation*. It is, therefore, fruitful to separate the exchange-correlation energy into three components: 1) the exchange energy, which is well described by a GGA form $E_x = E_x^{GGA}[\rho]$, 2) the local correlation energy, which is just the LDA correlation $E_c^{local} = E_c^{LDA}[\rho]$, and 3) the remaining non-local correlation $E_c^{nl}[\rho]$

$$E_{xc}[\rho] = E_x^{GGA}[\rho] + E_c^{LDA}[\rho] + E_c^{nl}[\rho] \tag{2.26}$$

In vdW-DFT, this non-local correlation functional is assumed to take the mathematical form

$$E_c^{nl}[\rho(\vec{r})] = \frac{1}{2}\int d^3\vec{r}\int d^3\vec{r}'\,\rho(\vec{r})\phi^{London}(\vec{r},\vec{r}')\rho(\vec{r}') \tag{2.27}$$

where $\phi^{London}(\vec{r},\vec{r}')$ is a kernel, or Green's function, describing the London dispersion interactions between charge densities at points $\vec{r}$ and $\vec{r}'$. The advantage of this representation is that it makes transparent the fundamental action-at-a-distance nature of non-local correlation. In fact, it bears a striking resemblance to the form of classical Coulomb repulsion, a point made especially clear if we rewrite the electrostatic energy in terms of the Coulombic Green's function $\phi^{Coulomb}(\vec{r},\vec{r}') = \frac{1}{4\pi\epsilon_0|\vec{r}-\vec{r}'|}$,

$$E_{Hartree}[\rho(\vec{r})] = \frac{1}{2}\int d^3\vec{r}\int d^3\vec{r}'\,\rho(\vec{r})\phi^{Coulomb}(\vec{r},\vec{r}')\rho(\vec{r}') \tag{2.28}$$

In essence, all non-local influences between distant charge densities $\rho(\vec{r})$ and $\rho(\vec{r}')$ can be completely characterized by the interaction kernel $\phi(\vec{r},\vec{r}')$, which describes exactly how the charge density at one point propagates an influence to a distant point,

$$E_{non-local}[\rho(\vec{r})] = E_{Hartree}[\rho(\vec{r})] + E_c^{nl}[\rho(\vec{r})] = \frac{1}{2}\int d^3\vec{r}\int d^3\vec{r}'\,\rho(\vec{r})\phi(\vec{r},\vec{r}')\rho(\vec{r}'), \tag{2.29}$$

and the interaction kernel $\phi(\vec{r},\vec{r}')$ can be conveniently separated into two pieces, arising from classical electrostatic Coulomb repulsion and quantum-mechanical dispersion

$$\phi(\vec{r},\vec{r}') = \phi^{Coulomb}(\vec{r},\vec{r}') + \phi^{London}(\vec{r},\vec{r}') = \phi^{Coulomb}(|\vec{r}-\vec{r}'|) + \phi^{London}(\vec{r},\vec{r}'). \tag{2.30}$$

In the last equality we have explicitly labeled that the Coulomb interaction kernel depends only on the absolute distance between the two points. This seemingly trivial statement

contains within it a wealth of valuable information. It tells us that the Coulomb interaction is isotropic, depending only on the distance between objects and not their orientation. Furthermore, the Coulomb interaction between two particular points $\vec{r}$ and $\vec{r}^{\,\prime}$ is a fundamentally pairwise effect, depending only on the charge densities at each of the two points in question. In other words, if we keep the charge densities at $\vec{r}$ and $\vec{r}^{\,\prime}$ fixed, then the Coulomb force between these two points is independent of the form of the charge density at all other points.

These properties do not hold in the case of London dispersion forces, which are anisotropic, many-body effects. These complexities manifest themselves in the detailed form of the London kernel for vdW-DFT, which ends up having the variable dependencies

$$\phi^{London}(\vec{r},\vec{r}^{\,\prime}) = \phi(|\vec{r}-\vec{r}^{\,\prime}|f(n(\vec{r}),\nabla n(\vec{r})),|\vec{r}-\vec{r}^{\,\prime}|f(n(\vec{r}^{\,\prime}),\nabla n(\vec{r}^{\,\prime}))) \tag{2.31}$$

where $f(\vec{r})$ has a complex functional dependence on the charge density and charge density gradient at $\vec{r}$. What this property of the kernel implies is that London dispersion forces, in contract to electrostatic forces, not only depend on the absolute distance between two regions of charge, but are also highly dependent on the details of the charge distribution in said regions. Even if we keep the charge densities at $\vec{r}$ and $\vec{r}^{\,\prime}$ fixed, the direct London dispersion interaction between those two points is very sensitive to changes in their chemical environment. This extreme sensitivity, in this author's opinion, is one of the most biologically important features of London dispersion forces, and at the heart of their functionality.

Chapter 3

5-Methylation of Cytosine in CG:CG Base-Pair Steps

3.1 Introduction

Self-consistent Kohn-Sham density functional theory (KS-DFT) [33, 43] has traditionally been one of the most popular tools of choice for ab initio electronic structure calculations of properties of dense matter, due to both its comparable accuracy to quantum chemical methods and relatively small computational cost. However, until recently, the failure of traditional exchange-correlation functionals to account for nonlocal London dispersion forces precluded its application to sparse matter, such as biological molecules. The development of the nonlocal van der Waals density functional vdW-DF [18] and the subsequent, higher accuracy vdW-DF2 [48] has expanded the realm of DFT to a whole new class of systems.

One of the first applications of vdW-DF was the analysis by Cooper et al. [14] of hydrogen bonding between nucleic-acid base pairs and stacking interactions between successive bases in base-pair steps. The authors successfully accounted for sequence-specific trends in base-pair separation and rotation seen in high-resolution crystal structures [29]. Additionally, they demonstrated the role that the methyl group of thymine plays in stabilizing double-stranded DNA over its uracil counterpart in RNA. The initial success of vdW-DF, combined with the ensuing development of the even more accurate vdW-DF2, motivates deeper theoretical study of the structural energetics of biologically relevant nucleobase configurations.

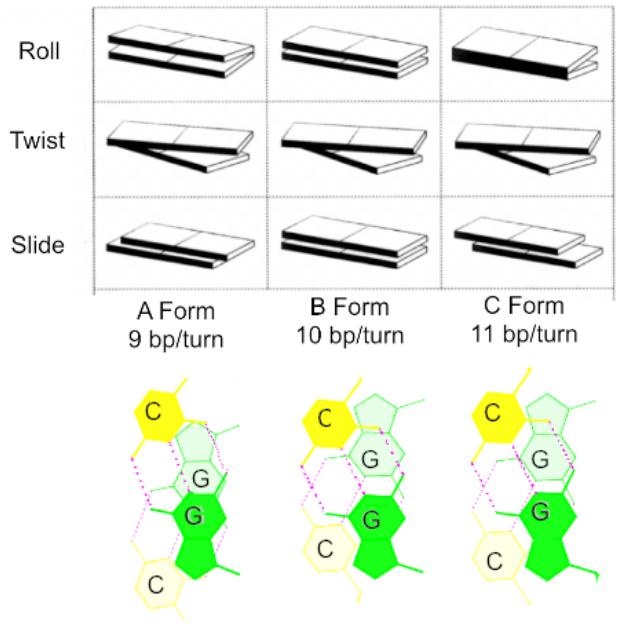

Figure 3.1: Schematic illustration of important base-pair step parameters of canonical A, B and C forms of DNA. Also included are sample stacking diagrams for a CG:CG step, showing the exact spatial displacements of chemical units in fiber models [2]. The lower rigid body in the upper schematics is denoted by the lightly shaded base pair in the lower stacking diagram. The shaded edges on the schematic blocks and the right edges of the stacking diagrams both correspond to the minor-groove edges of base pairs. The pink dashed lines represent hydrogen bonds. Schematics adapted from Reference 6 and stacked CG step images computed with X3DNA [113].

In vivo, DNA predominantly adopts right-handed double-helical structures. Therefore, this class of structures is likely to be the most biologically relevant. In particular, high-resolution DNA crystal structures primarily adopt three right-handed double-helical states, namely A, BI and BII. Repetitions of these local conformations throughout the nucleic acid lead to polymers characterized as A-like, B-like and C-like, respectively [74, 2, 98].

The subtle differences in base-pair step geometries for each of the different helical states are illustrated in Figure 3.1. Most notably, A-like base-pair steps are under-twisted relative to B-like ones, while C-like steps are over-twisted. It has been suggested [94] that the

interconversions between these three configurations are related to sequence-specific mechanisms of nucleosome positioning in chromatin, the bundled assembly of DNA and proteins in eukaryotic cell nuclei. This has consequences for the control of gene silencing.

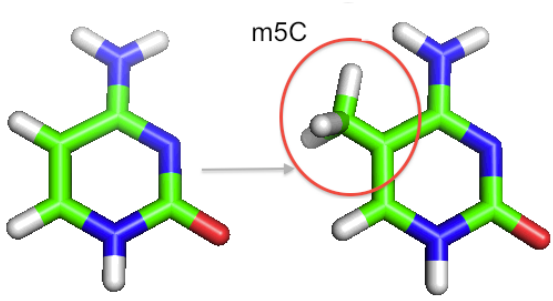

Figure 3.2: 5-methylation of cytosine, an important epigenetic modification. The methyl group that replaces the hydrogen is circled in red. Green represents carbon, white hydrogen, red oxygen, and blue nitrogen. Graphics generated using PyMOL [84].

The stabilizing effect of methylation on thymine found by Cooper et al. inspires the question of whether similar effects occur for other biochemical modifications. Specifically, the methylation of cytosine at the C5 position, as illustrated in Figure 3.2, is an important epigenetic modification. In particular, in CpG dinucleotides, it appears to trigger the protein-assisted compaction of chromatin [7]. This compaction determines whether or not genes can be transcribed into RNA. Elucidating the effects of the 5-methylation of cytosine on the structural energetics of CG:CG steps is therefore a valuable step towards understanding the factors that control eukaryotic gene regulation. Also worth analyzing are possible indirect effects, in particular, the interactions of methylated cytosines of CG:CG steps with immediately adjacent base-pair steps, including AC:GT, TC:GA, CC:GG and GC:GC.

In addition, there has been very little, if any, work connecting first-principles electronic structure calculations of biomolecular structure with bioinformatics analyses. The crystal structures in the Nucleic-Acid Database (NDB) [5] can be utilized to extract the most

commonly observed sequence-specific collective atomic motions. This allows reduction of the complex conformational coordinate space to a simpler subspace that is more likely to be biologically relevant. This information, besides being valuable in its own right, can interface with DFT calculations to determine realistic energy landscapes of stacked base pairs.

Previous quantum chemical studies have focused on base stacking [88, 76], including some recent work on the stacking of Watson-Crick paired bases [89], and the effects of cytosine methylation in the context of reaction kinetics and equilibrium structures [23, 109, 106]. However, these methods have not traditionally been integrated with a bioinformatics approach to look at DNA conformational energetics in the context of the coordinated motions of neighboring base pairs in the double helix. In this study, the analysis of Cooper et al. is extended to the 5-methylation of cytosine, taking into account these considerations.

3.2 Methods

3.2.1 Generation of a Non-Redundant DNA-Protein Dataset

The motions of neighboring base-pair steps were deduced from a non-redundant dataset of 239 protein-DNA crystal complexes of 2.5 Angstroms or better resolution taken from the NDB [51]. The dataset included 101 structures of double-helical DNA bound to enzymes, 121 duplexes determined in the presence of regulatory proteins, 16 complexes with structural proteins, and one DNA bound to a multi-functional protein. The structures were filtered to exclude over-represented complexes in order to obtain a balanced sample of spatial and functional forms. The selection and classification of structures was based on sequential and structural alignment, as well as available protein classification databases, including the SCOP [65] scheme. The working dataset excluded terminal base-pair steps (i.e., residues at chain ends and nicked dimer steps), which may adopt alternate conformations or be

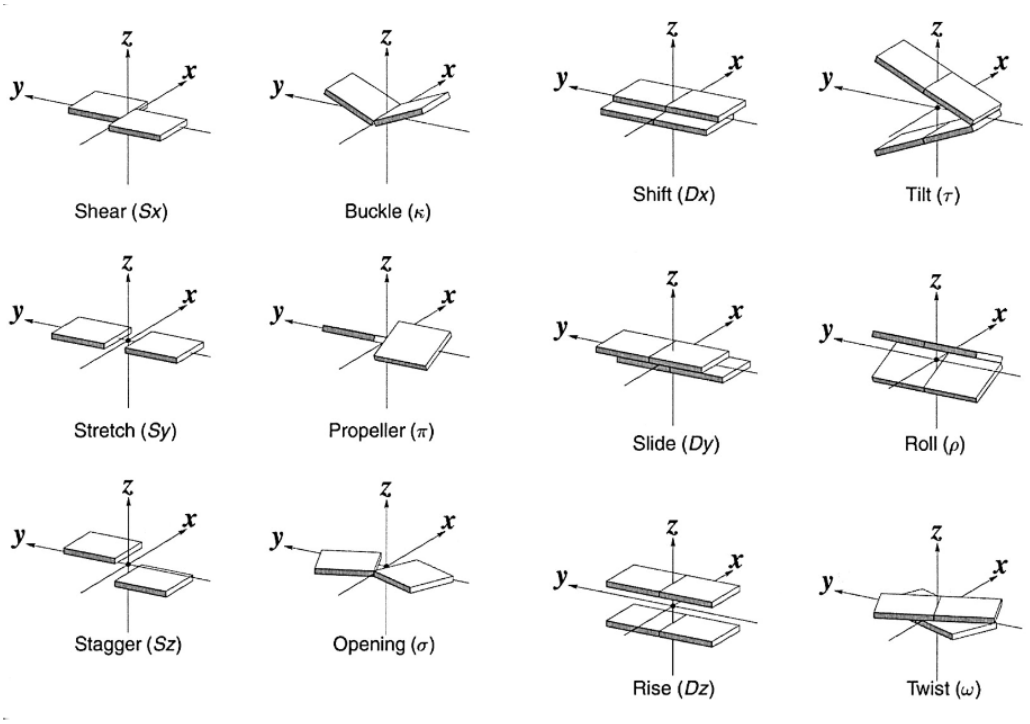

Figure 3.3: The rigid-body configuration of a DNA base-pair step is specified by six local parameters per base pair and six step parameters for successive base-pair steps. There are three translational and three rotational degrees of freedom for each kind of rigid-body motion. Figure adapted with permission from Reference [54].

affected by crystal packing, as well as chemically modified nucleotides, nucleotides involved in non-canonical base pairs, and melted residues, in which complementary base pairs are highly distorted and do not contain the requisite number or types of hydrogen bonds.

3.2.2 Principal Component Analysis

Information on dinucleotide steps was extracted from the NDB files using the 3DNA software package [113]. From this information, an eighteen-parameter data vector characterizing the conformation of a given base-pair step was generated with 3DNA. This included twelve base-pair parameters (six for each base-pair) and six step parameters (Figure 3.3). After converting the data vector to a standardized z-value, principal component analysis

was performed for the CG:CG steps. Using a scree test, the three highest eigenvalues, corresponding to the dominant collective modes of motion, were extracted. This process was repeated for all possible steps that could directly connect to CG:CG, and therefore be affected by methylation.

3.2.3 Density Functional Theory Calculations

In the analysis of individual modes, a sufficient set of points within two standard deviations of the mean along the modal pathways was sampled to determine the shape of the energy landscape of stacked base pairs. Base-pair steps were created with 3DNA and methyl groups were added with Open Babel [75]. Computations were performed using vdW-DF2 as implemented in the Quantum Espresso package [27] via the algorithm of Roman-Perez and Soler [82]. All calculations used standard generalized gradient approximation pseudopotentials [96] ((these pseudopotentials are available free of charge on the Quantum Espresso website, and were generated by Tozzini et. al in [95]), with an energy cutoff of 60 Ry (1 Ry = 313.755 kcal/mol). SCF diagonalizations were performed with the Davidson algorithm, using convergence criteria of 10^{-6} Ry. To reduce spurious interaction between periodic images, the system was placed in a 40 x 30 x 30 cubic Bohr (1 Bohr = 0.529 Angstrom) supercell.

3.3 Results and Discussion

The main focus of this paper regards the effects of methylation on CG:CG steps. The next subsection discusses the qualitative nature of the modal motion of the three dominant principal components, which are termed ‘Opening’, ‘Sliding’ and ’Tearing’ and illustrated in Figures 3.4, 3.5 and 3.6, respectively. This is followed up with a report on the effects of 5-methylcytosine on the stacking energetics of CG:CG steps, including the interplay between

local and global effects. Finally, the indirect effects on neighboring steps are discussed.

3.3.1 Nature of the Principal Components

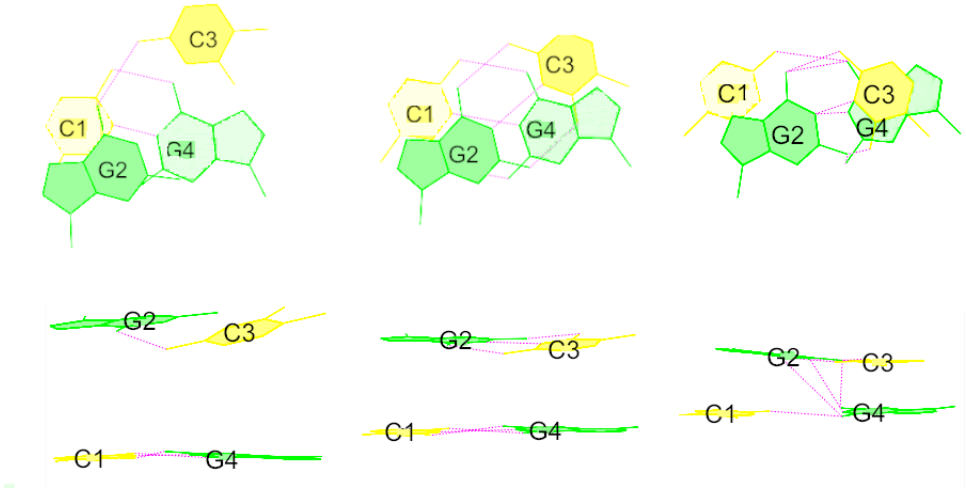

Figure 3.4: The first principal component of CG:CG steps, a tensile 'opening' mode of the crack between DNA strands. From left to right, respectively, are images for steps that are five negative normal mode units from the mean, at the mean, and five positive units away. The upper and lower rows display views from the top-down and looking into the minor groove. The lower base pair is labelled by C1 bonded to G4, while the upper one is labelled by G2 and C3. The pink dashed lines represent hydrogen bonds. Molecular images created with 3DNA [113].

The predominant principal component of CG:CG steps may be interpreted, borrowing a term from the fracture mechanics community [52], as a tensile 'opening' mode between the two DNA strands. It consists of a coherent twisting, sliding and rising of the step, accompanied by out-of-phase stretching and opening of the C:G and G:C base pairs. As the vertical separation between the base pairs decreases and the step is tightened and undertwisted, the lower C:G pair breaks apart and the upper G:C pair is compacted. The most prominent consequence is an enhanced overlap between the C3 and G4 bases.

This breakage of hydrogen bonds is similar to the 'breathing' modes observed in the

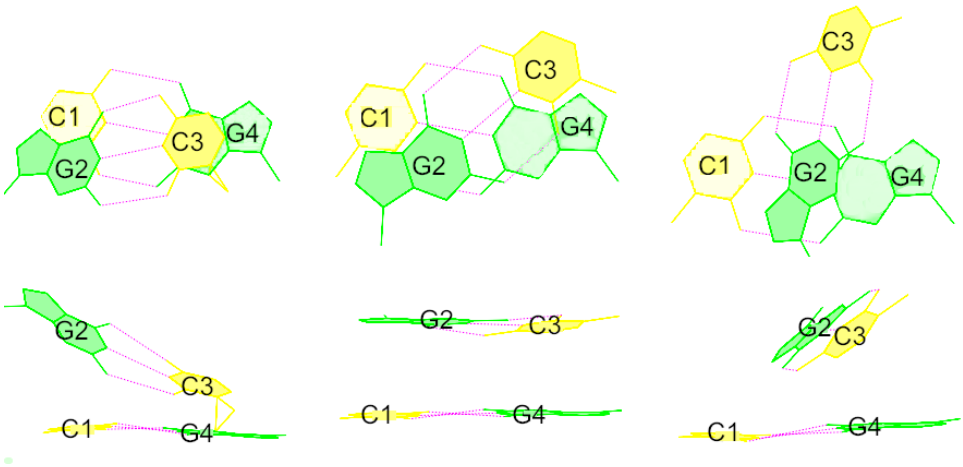

Figure 3.5: The second principal component of CG:CG steps, a shear 'sliding' mode. See the caption of Figure 3.4 for explanation of notations and symbols.

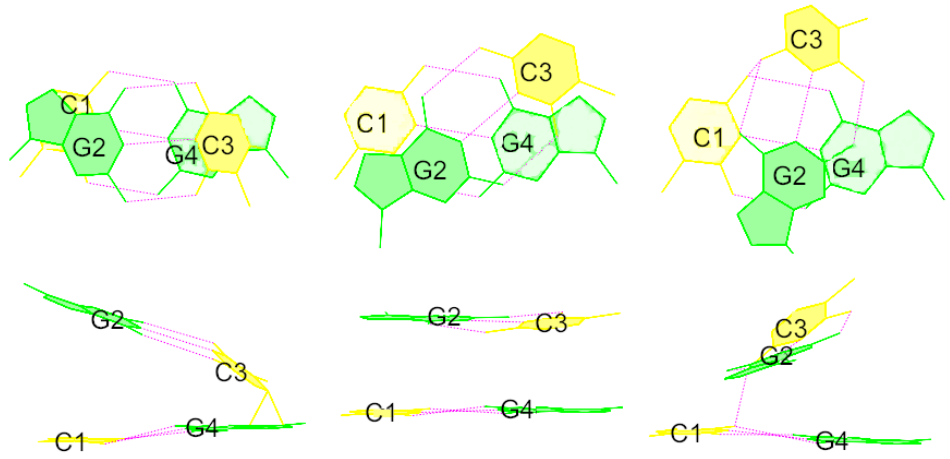

Figure 3.6: The third principal component of CG:CG steps, a shear 'tearing' mode. See the caption of Figure 3.4 for explanation of notations and symbols.

classic work of Mandal, Kallenbach and Englander [58]. The analysis here provides a detailed picture of the possible nature of this breaking, including the symmetric way in which the bond-breaking acts with respect to the equilibrium state.

The two secondary modes share several similarities. Both of them at their extreme ends correspond to the C and G bases along one strand having a very small separation, with

the bases on the opposite strand being consequently more spread out and isolated. Motion along the modes can be interpreted as the two strands alternating between compressing and opening. In both cases, this is primarily controlled by a twisting motion that acts in the opposite direction of tilt and slide, and in tandem with shift. However, the remaining step parameters, namely rise and roll, move in an opposite direction in one of the modes relative to the other, as do dominant local base-pair motions, such as stagger or buckle. Thus, as the modes are traversed, the DNA gradually transitions from the A to B to C-forms.

To continue with the analogy with fracture mechanics, whereas the most dominant principal component represents a tensile 'opening' mode, the two secondary components may be interpreted as an orthogonal pair of shearing modes. Together, they span the fracture plane perpendicular to that of the tensile opening. The conventional terms for this pair [52] are 'tearing' and 'sliding' modes, respectively, and this nomenclature shall be adopted for the remainder of the paper.

While the 'opening' modes correspond to traditional breathing modes, this pair of transverse modes is more along the lines of hydrogen bond bifurcations [66]. Here, the bases on neighboring stacks can become 'mixed', resulting in hydrogen bonding not just between complementary bases in a single base pair, but also between adjacent ones located on the same DNA strand.

3.3.2 Effects of Cytosine 5-Methylation on CG:CG Steps

The stacking energetics and effects of methylation for each of the principal modes are illustrated in Figures 3.7 and 3.8, respectively. For convenience of illustration, the data are plotted against the value of the twist angle along each of the three mode landscapes. Two general trends emerge from the data: 1) The methyl groups globally suppress the

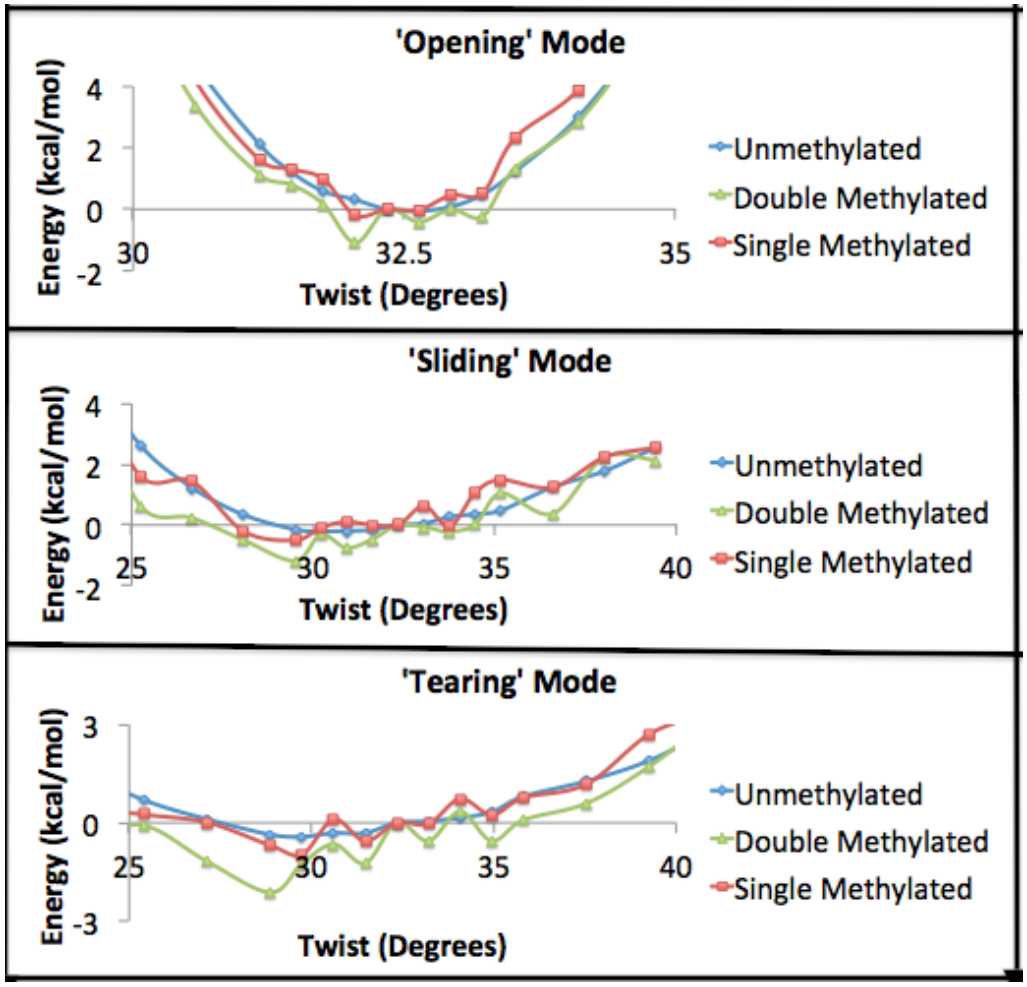

Figure 3.7: Stacking energy landscapes for each of the three principal components, with and without methylation. The horizontal axes are labelled by the variation of twist along each of the modes.

overtwisting of CG:CG steps by pushing the minimum to a lower twist, and 2) This global inhibition is accompanied by local modulations of the potential that soften the landscape by creating an ensemble of intermediate states, including several that are overtwisted.

To demonstrate the physical reason for the global inhibition of twisting, it is informative to look at the 'low-twist' regimes of each of the modes. As illustrated in Figure 3.9, these regimes correspond to a higher degree of stacking overlap between the C3 and G4 bases. As a consequence, the stacking interactions of the C5 carbon in cytosine with the adjacent

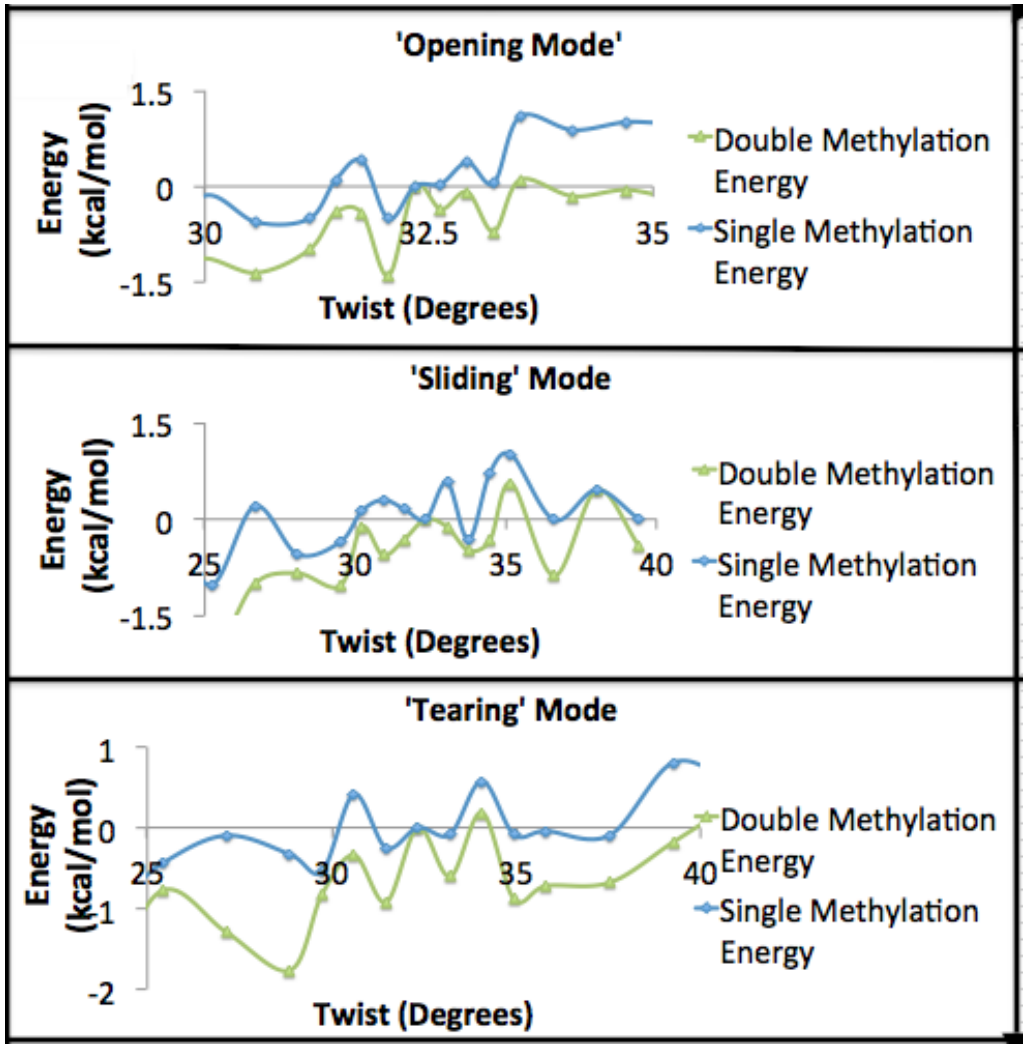

Figure 3.8: The individual contributions of one and two methyl groups to the effective stacking potential of the step, as measured by the energy difference, with respect to the unmethylated state, of calculations with one or both C5 groups methylated, respectively. The horizontal axes are labelled by the variation of twist along each of the modes.

guanine are enhanced. The addition of a methyl group amplifies these stacking forces, serving as an effective 'pinning' field that stabilizes undertwisted configurations.

The origin of the local fluctuations is more subtle. A useful metaphor for illustrating where they come from, at least on a phenomenological level, is to consider the dihedral angle ϕ between four atoms in a covalent bond, as shown in Figure 3.10. As is well known [99], due to the periodicity of the dihedral angle, $\phi = \phi + 2\pi$, the potential energy $V(\phi)$ can

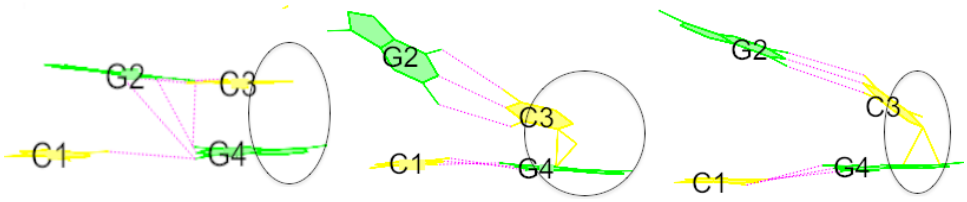

Figure 3.9: From left to right are minor-groove views of the low-twist regimes of the opening, sliding and tearing modes, respectively. As illustrated, in these regimes, the overlap area between C3 and G4 is greater. This leads to an enhancement in the stacking interactions of a methyl group at the C5 position with the adjacent guanine. The pink dashed lines represent hydrogen bonds.

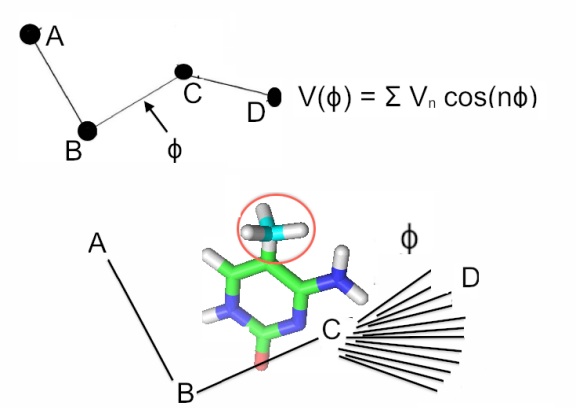

Figure 3.10: As a base-pair step moves along the landscape of its conformational modes, the methyl group may be interpreted as having a set of moving 'non-covalent' dihedral angles with the atoms in the step, generated by the 'angle' that the 5-methylcytosine makes with the remainder of the base-pair step. In analogy to its more commonly discussed analog in covalent bonding, this torsional variation creates an effective torsional 'potential', which, like any periodic potential, can be decomposed into a superposition of harmonics of varying wavelength. In this schematic, A may be interpreted as a methyl group, the B-C line as the cytosine, and D as all other atoms in the step.

be expanded in a harmonic series. Often, one harmonic dominates the torsional energetics, but generally the potential is an incoherent superposition of different frequencies, leading to a complex landscape of metastable vacua.

From the perspective of the CG:CG steps, the fundamental difference upon addition of

the methyl groups is the introduction of several additional variables, which can be interpreted as a set of non-local 'torsional' angles describing the orientation of the methyl group with respect to other atoms in the polymer assembly. These variables, being periodic like dihedral angles, likewise give rise to modulations in the potential energy. The net result of all such contributions is a noisy signal that 'softens' the base-pair step, enabling it to more easily change its shape via fluctuations. This ties in nicely with the expectation that the 5-methylation of cytosine would make the base-pair softer, similar to previous observations by Cooper et al. regarding the impact that the methyl group of thymine has on A:T base pairs compared to A:U [14].

3.3.3 Effects of Methylation on Neighboring Steps

Calculations of the energetics of different base-pair steps adjacent to CG:CG indicate that the interactions of the C5 methyl groups with the neighboring steps have an energetic effect comparable to that of their immediate interactions with the CG:CG step.

Principal component analysis reveals that the classification of dominant collective motions into opening, sliding and tearing modes persists for the different base-pair steps. Furthermore, as illustrated in Figure 3.11, 5-methylcytosine continues to torsionally modulate the stacking energy as the identities of the remaining bases in the step change.

However, there are also some important differences compared to CG:CG. Perhaps most relevant to nucleosome formation is the observation that 5-methylcytosine globally *enhances* overtwisting of several modes in GC:GC steps. This overwinding is a potential mechanism for preserving the double-helical structure of DNA by countering the tendency to melt from CG:CG unwinding.

Additionally, the typical stacking energy of methylation is lower for CG:CG steps than

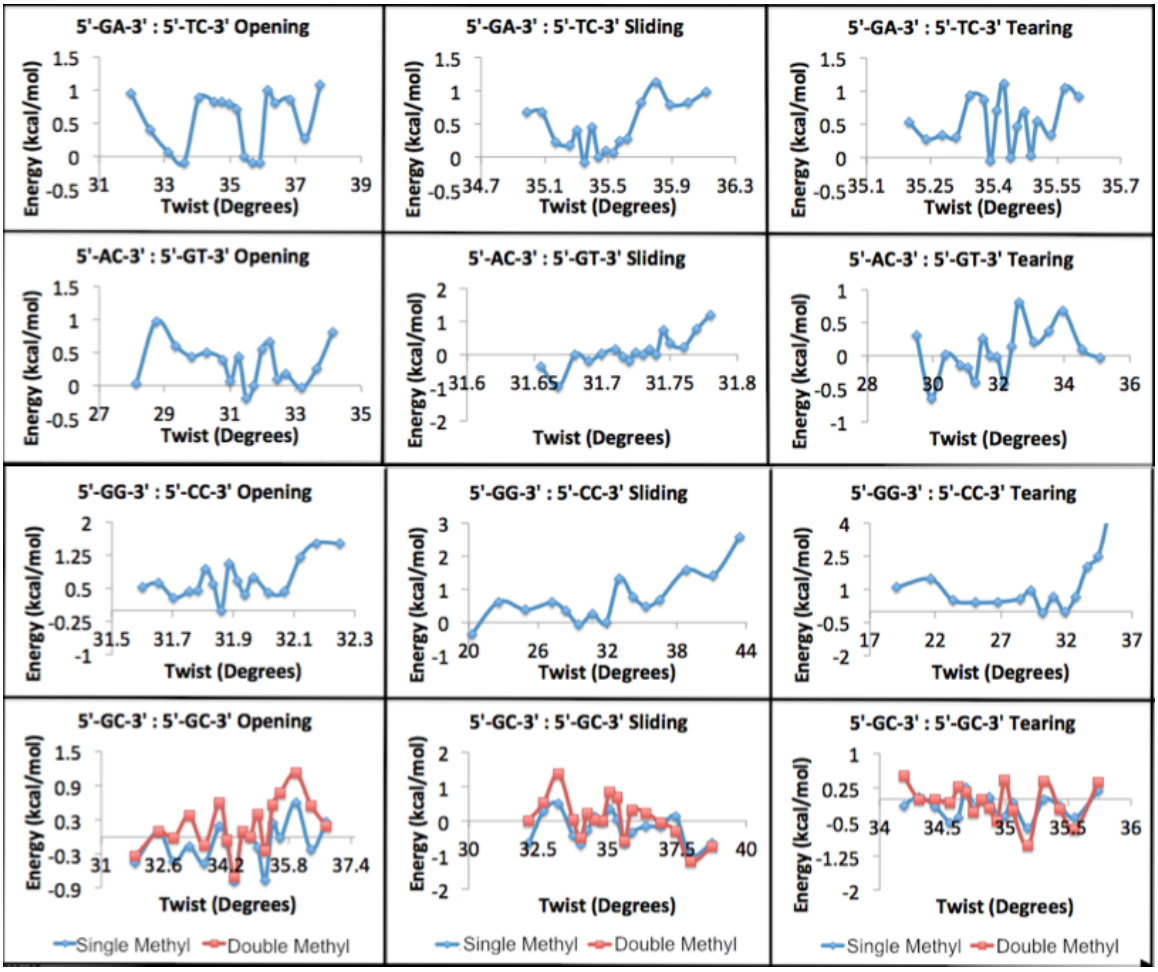

Figure 3.11: Indirect stacking interaction energies of 5-methylcytosine with possible steps neighboring the CG:CG step. For GA:TC, AC:GT, and GG:CC steps, there is only one possible cytosine that can connect to a CG:CG, and thus be potentially methylated. For GC:GC steps, however, it is possible for one or both of its cytosines to be methylated.

it is for its neighbors, as seen by a comparison of Figures 3.8 and 3.11. From a statistical mechanical point of view, methylation enhances the room-temperature Boltzmann partition function of CG:CG steps while decreasing that of its neighbors, corresponding to an increase or decrease in Helmholtz free energy, respectively. Consequently, the methylation of CG:CG steps is more thermodynamically stable than methylation of other possible steps, an argument for why it is more commonly observed.

3.4 Conclusions

In summary, this study has extended the work of Cooper et al. in a systematic study of the effects of C5 methylation on base-stacking energetics. Methylation is seen to have nontrivial effects on on the flexibilities of the opening, sliding and tearing motions of CG:CG steps. Specifically, it globally inhibits overtwisted states while simultaneously generating local potential energy modulations that soften the step. Furthermore, analysis of interactions of the methyl groups with possible neighboring steps indicates that these effects are of comparable importance to those of the methyl group on CG:CG itself.

The mechanisms discussed in this work do not appear to be limited to this specific system. There is consistent evidence that the methyl groups perform a functional role via a combination of long-wavelength and short-wavelength effects, which is suggestive of some more general principles underlying chemical epigenetic modifications and the physical processes responsible for their biological functionality, particularly in a mechanical context.

The results of this work compare favorably with previous experimental data regarding the effects of cytosine methylation on nucleosome positioning. In particular, Davey, Pennings and Allan [15] observed that methylation of nucleosomal DNA prevents the histone octamer from interacting with an otherwise high-affinity chicken β-globin gene positioning sequence. This sequence contains a $(\mathrm{CpG})_3$ motif located 1.5 helical turns from the dyad axis of the nucleosome, with minor-groove edges on the base-pair step that are oriented towards the histone core. When this sequence motif is unmethylated, it is capable of adopting the structural deformations necessary to interact with the histone octamer, and thus enable nucleosome positioning. However, as the current calculations demonstrate, methylation of CG-rich stretches of DNA enhances the formation of the A-DNA polymorph, a helical

form that is more resistant to bending deformations than B-DNA, and which also bends DNA in the opposite sense. Consequently, interactions with the histones are inhibited, and nucleosome formation is suppressed.

Furthermore, a followup study by Davey, et al. [16] indicated that mutations of the $(CG)_3$ sequence motif into either GC:GC or CC:GG base-pair steps affect both the degree of nucleosome formation and the amount of disruption by CG:CG methylation. This ties in with the present finding that the effects of methylation depend on the sequential and structural context of the modified cytosines.

This work, additionally, demonstrates a foundation for future studies of realistic structural biomaterials modeling at the atomistic level via density functional theory. In particular, the consistency between experiments and calculations, in both this work and in the earlier studies of Cooper et al. [14], points to the capability of using first-principles approaches to extract valuable biochemical information on systems in which there is no prior experimental data. Thus, density functional theory calculations can serve as a complement to more traditional single-molecule biophysical experiments.

3.5 Appendix

Displayed below are the mean values of the parameters, and for each principal component, the contribution of each parameter to one positive unit of that component, as measured in standard deviations from the mean. Also displayed is the score for each component, or the fraction of the total variance that it captures.

Not included in this dissertation, but available in the Supplemental Information of the manuscript [110], is comprehensive information on the non-redundant set of protein-DNA crystal complexes used in this study, including references to the original literature.

3.5.1 Results of Principal Component Analyses

Parameter	Mean	Std.	PCA 1 (Opening)	PCA 2 (Sliding)	PCA 3 (Tearing)
Latent Score			0.20	0.16	0.12
$Z_{methylation}$			2.760	2.697	7.088
$\text{Shear}_{C:G}$	0.18	0.38	-0.05	-0.02	0.02
$\text{Stretch}_{C:G}$	-0.09	0.41	**0.18**	0.01	0.03
$\text{Stagger}_{C:G}$	-0.03	0.35	-0.02	-0.08	**0.10**
$\text{Buckle}_{C:G}$	0.00	8.36	0.78	**1.28**	-0.64
$\text{Propeller}_{C:G}$	-6.79	6.48	-0.53	-0.22	-0.12
$\text{Opening}_{C:G}$	-0.68	8.39	**-3.64**	**-1.11**	0.70
$\text{Shear}_{G:C}$	-0.11	0.40	0.02	-0.01	0.08
$\text{Stretch}_{G:C}$	-0.14	0.36	**-0.16**	-0.03	-0.01
$\text{Stagger}_{G:C}$	0.05	0.36	-0.002	**-0.10**	**0.17**
$\text{Buckle}_{G:C}$	2.15	9.14	0.35	**-2.63**	**3.64**
$\text{Propeller}_{G:C}$	-6.21	6.84	0.98	0.49	-0.39
$\text{Opening}_{G:C}$	-0.01	7.75	**3.63**	0.70	0.12
Shift	-0.06	0.75	0.05	**0.11**	0.09
Slide	0.46	0.98	0.02	**-0.41**	**-0.18**
Rise	3.37	0.56	**-0.18**	**0.11**	**-0.15**
Tilt	-0.66	12.59	0.85	**-4.79**	**-5.59**
Roll	6.93	11.93	**-1.11**	**5.66**	-0.49
Twist	32.34	20.46	**-2.94**	**7.10**	**8.67**

Table 3.1: CG:CG base-pair step parameter mean values, standard deviations, and magnitudes for one unit of each principal component. The analysis included $N = 213$ data points. Units of angular parameters are degrees and units of translational parameters are Angstroms. The second row is an estimate of the contribution of the methyl groups to the Boltzmann partition function $Z_{methylation} = e^{-\frac{\Delta F_{methylation}}{k_B T}}$ at room temperature, based on DFT calculations. Base-pair parameters subscripted with A:T or C:G refer to values for the lower A:T or upper C:G pairs, respectively. Particularly dominant parameter motions, namely those greater than 0.1 Angstrom or 1 degree in magnitude, are highlighted in bold.

Parameter	Mean	Std.	PCA 1 (Opening)	PCA 2 (Sliding)	PCA 3 (Tearing)
Latent Score			0.16	0.12	0.11
$Z_{methylation}$			0.708	0.768	0.777
Shear$_{A:T}$	0.13	0.34	0.02	-0.001	0.03
Stretch$_{A:T}$	-0.10	0.18	0.01	0.01	-0.006
Stagger$_{A:T}$	0.01	0.31	0.04	-0.09	-0.05
Buckle$_{A:T}$	3.79	9.48	**3.31**	-0.01	**2.54**
Propeller$_{A:T}$	-8.75	6.55	**-1.30**	**2.27**	0.28
Opening$_{A:T}$	0.77	4.42	-0.56	0.20	**-1.45**
Shear$_{C:G}$	0.15	0.33	-0.02	-0.026	0.04
Stretch$_{C:G}$	-0.09	0.34	**0.11**	**0.13**	**-0.10**
Stagger$_{C:G}$	-0.01	0.30	-0.01	**-0.11**	-0.02
Buckle$_{C:G}$	3.19	8.97	**3.12**	-0.35	**2.66**
Propeller$_{C:G}$	-8.30	6.84	**-1.55**	**1.69**	0.80
Opening$_{C:G}$	0.25	8.52	**-2.61**	**-3.04**	**2.19**
Shift	0.25	0.58	**0.18**	0.04	**0.18**
Slide	-0.59	0.43	**-0.13**	0.07	-0.004
Rise	3.29	0.38	0.08	**-0.12**	**-0.12**
Tilt	0.39	3.45	0.78	0.65	**1.36**
Roll	2.20	4.97	-0.15	**1.89**	-0.96
Twist	31.73	6.38	**2.39**	0.05	**-2.24**

Table 3.2: AC:GT base-pair step parameter mean values, standard deviations, and magnitudes for one unit of each principal component. The analysis included $N = 498$ data points. Units of angular parameters are degrees and units of translational parameters are Angstroms. The second row is an estimate of the contribution of the methyl groups to the Boltzmann partition function $Z_{methylation} = e^{-\frac{\Delta F_{methylation}}{k_B T}}$ at room temperature, based on DFT calculations. Base-pair parameters subscripted with A:T or C:G refer to values for the lower A:T or upper C:G pairs, respectively. Particularly dominant parameter motions, namely those greater than 0.1 Angstrom or 1 degree in magnitude, are highlighted in bold.

Parameter	Mean	Std.	PCA 1 (Sliding)	PCA 2 (Opening)	PCA 3 (Tearing)
Latent Score			0.15	0.12	0.09
$Z_{methylation}$			0.705	0.586	0.460
$\text{Shear}_{G:C}$	0.05	0.38	-0.05	-0.03	-0.05
$\text{Stretch}_{G:C}$	-0.12	0.19	-0.03	-0.04	0.05
$\text{Stagger}_{G:C}$	-0.01	0.32	-0.06	**0.10**	-0.01
$\text{Buckle}_{G:C}$	0.95	8.09	**-1.00**	**2.28**	**-2.15**
$\text{Propeller}_{G:C}$	-10.35	6.35	-0.02	**-1.93**	-0.68
$\text{Opening}_{G:C}$	0.47	5.32	-0.19	0.02	**1.64**
$\text{Shear}_{A:T}$	-0.08	0.36	**0.10**	-0.04	-0.07
$\text{Stretch}_{A:T}$	-0.14	0.20	-0.02	0.004	0.09
$\text{Stagger}_{A:T}$	-0.05	0.32	0.09	**0.12**	0.02
$\text{Buckle}_{A:T}$	0.70	8.60	**2.25**	**3.13**	**-1.61**
$\text{Propeller}_{A:T}$	-9.39	6.70	**-1.08**	-0.82	**-1.83**
$\text{Opening}_{A:T}$	0.21	4.40	**-1.11**	0.44	**2.12**
Shift	-0.16	0.62	**0.20**	**-0.13**	**0.13**
Slide	0.01	0.69	**-0.27**	**0.12**	-0.05
Rise	3.31	0.25	0.08	0.05	0.02
Tilt	-1.34	3.81	**1.81**	-0.09	0.44
Roll	1.97	5.47	-0.28	**-1.35**	**-1.47**
Twist	35.44	5.02	-0.45	**2.31**	0.16

Table 3.3: GA:TC base-pair step parameter mean values, standard deviations, and magnitudes for one unit of each principal component. The analysis included $N = 308$ data points. Units of angular parameters are degrees and units of translational parameters are Angstroms. The second row is an estimate of the contribution of the methyl groups to the Boltzmann partition function $Z_{methylation} = e^{-\frac{\Delta F_{methylation}}{k_B T}}$ at room temperature, based on DFT calculations. Base-pair parameters subscripted with G:C or A:T refer to values for the lower G:C or upper A:T pairs, respectively. Particularly dominant parameter motions, namely those greater than 0.1 Angstrom or 1 degree in magnitude, are highlighted in bold.

Parameter	Mean	Std.	PCA 1 (Sliding)	PCA 2 (Tearing)	PCA 3 (Opening)
Latent Score			0.19	0.15	0.10
$Z_{methylation}$			0.676	1.765	0.768
Shear$_{G:C}$	-0.19	0.38	0.03	0.08	0.02
Stretch$_{G:C}$	-0.17	0.20	0.05	0.01	0.07
Stagger$_{G:C}$	0.00	0.31	-0.08	**0.13**	0.03
Buckle$_{G:C}$	-1.36	8.00	-0.27	**2.58**	**3.95**
Propeller$_{G:C}$	-6.47	7.04	**2.11**	**-2.04**	**1.54**
Opening$_{G:C}$	-1.12	3.61	0.69	-0.56	**1.18**
Shear$_{C:G}$	0.17	0.42	0.002	-0.08	**0.10**
Stretch$_{C:G}$	-0.18	0.19	0.02	0.03	0.01
Stagger$_{C:G}$	-0.03	0.35	**-0.13**	-0.06	0.01
Buckle$_{C:G}$	1.25	9.27	**3.82**	**1.53**	**1.59**
Propeller$_{C:G}$	-6.12	6.62	**1.45**	-0.78	**-2.20**
Opening$_{C:G}$	-1.22	3.88	0.16	0.67	0.54
Shift	-0.01	0.67	0.05	**0.26**	**-0.15**
Slide	-0.27	0.57	-0.02	0.06	-0.05
Rise	3.30	0.41	**-0.14**	0.004	**0.12**
Tilt	0.15	3.51	0.36	**1.69**	-0.29
Roll	1.21	4.94	**1.47**	-0.58	0.83
Twist	34.81	7.02	**-2.64**	-0.62	**1.95**

Table 3.4: GC:GC base-pair step parameter mean values, standard deviations, and magnitudes for one unit of each principal component. The analysis included $N = 264$ data points. Units of angular parameters are degrees and units of translational parameters are Angstroms. The second row is an estimate of the contribution of the methyl groups to the Boltzmann partition function $Z_{methylation} = e^{-\frac{\Delta F_{methylation}}{k_B T}}$ at room temperature, based on DFT calculations. Base-pair parameters subscripted with G:C or C:G refer to values for the lower G:C or upper C:G pairs, respectively. Particularly dominant parameter motions, namely those greater than 0.1 Angstrom or 1 degree in magnitude, are highlighted in bold.

Parameter	Mean	Std.	PCA 1 (Sliding)	PCA 2 (Tearing)	PCA 3 (Opening)
Latent Score			0.22	0.17	0.11
$Z_{methylation}$			0.030	0.148	0.381
$\text{Shear}_{(G:C)_1}$	-0.13	0.36	-0.08	0.07	-0.02
$\text{Stretch}_{(G:C)_1}$	-0.29	0.66	**0.24**	**-0.22**	0.02
$\text{Stagger}_{(G:C)_1}$	0.00	0.33	0.05	-0.03	**0.15**
$\text{Buckle}_{(G:C)_1}$	1.37	8.00	-0.30	0.22	**3.00**
$\text{Propeller}_{(G:C)_1}$	-6.20	7.22	**-1.52**	**1.40**	-0.62
$\text{Opening}_{(G:C)_1}$	2.05	14.47	**-5.32**	**4.65**	-0.99
$\text{Shear}_{(G:C)_2}$	-0.13	0.52	0.02	-0.05	**-0.19**
$\text{Stretch}_{(G:C)_2}$	-0.15	0.19	0.02	-0.03	-0.07
$\text{Stagger}_{(G:C)_2}$	0.01	0.31	0.02	0.001	0.08
$\text{Buckle}_{(G:C)_2}$	0.95	8.11	0.54	-0.07	-0.007
$\text{Propeller}_{(G:C)_2}$	-6.80	6.78	0.21	-0.46	**-1.75**
$\text{Opening}_{(G:C)_2}$	-0.42	4.55	0.34	-0.53	**-1.80**
Shift	-0.12	0.75	0.05	-0.08	**-0.20**
Slide	-0.28	1.06	**-0.36**	**-0.29**	0.07
Rise	3.25	1.16	**0.35**	**-0.45**	**0.11**
Tilt	-1.01	26.83	**9.05**	**10.93**	**-1.65**
Roll	4.91	22.60	**-7.35**	**-9.20**	0.91
Twist	31.86	28.48	**11.57**	**8.54**	-0.26

Table 3.5: GG:CC base-pair step parameter mean values, standard deviations, and magnitudes for one unit of each principal component. The analysis included $N = 467$ data points. Units of angular parameters are degrees and units of translational parameters are Angstroms. The second row is an estimate of the contribution of the methyl groups to the Boltzmann partition function $Z_{methylation} = e^{-\frac{\Delta F_{methylation}}{k_B T}}$ at room temperature, based on DFT calculations. Base-pair parameters subscripted with $(\text{G:C})_1$ or $(\text{G} : \text{C})_2$ refer to values for the lower G:C or upper G:C pairs, respectively. Particularly dominant parameter motions, namely those greater than 0.1 Angstrom or 1 degree in magnitude, are highlighted in bold.

Chapter 4

Histone Arginine - DNA Phosphate Salt-Bridges

4.1 Introduction

The discovery of the double-helical structure of DNA [101] established that genetic information is encoded in the molecular sequence of base-paired nucleotides constituting an organism's genome. This information is transduced into an observable set of characteristics, or phenotype, via the central tenet of molecular biology: gene sequences of DNA are transcribed into complementary RNA sequences, which are subsequently translated into functional proteins. Mutations provide genetic variability, and Darwinian evolution acts on the resulting diversity of phenotypes, selecting for traits that maximize evolutionary fitness.

However, modifications of base sequences are not the only source of phenotypic variability. There exists an additional set of modifications termed the epigenetic code, which modify an organism's hereditary information while leaving the genomic sequence intact [97]. While epigenetic regulation occurs at all levels of gene expression, one of the most prominent mechanisms is at the level of control of transcription. In eukaryotes, this occurs via the dynamic remodeling of the structure of chromatin, the bundled assembly of DNA and proteins in cell nuclei. This remodeling controls the expression of specific genes, by selectively blocking or enabling the binding of transcription factors to particular regions of the genome [11].

Recent theoretical and experimental work [45, 70] has highlighted the role of histones,

the structural proteins that package chromatin, in mediating long-range communication between regulatory elements in the genome. The physical mechanism behind this signaling is the controlled manipulation of DNA elasticity at specific genomic sites. This is accomplished through a complex interplay of direct and water-mediated protein-DNA interactions [56, 17].

Over thirty years ago, Mirzabekov and Rich [62] suggested that histone-DNA interactions control DNA flexibility in chromatin via neutralization of the sugar-phosphate backbone by cationic amino acids. This has inspired experimental investigations into the electrostatic mechanisms of protein-induced DNA bending. These studies have verified that counterion condensation is indeed a major contributing factor to DNA deformability [104].

However, in recent years [59, 81], it has become apparent that this electrostatic neutralization is not the only significant mode of interaction between cationic amino acids and the polyelectrolyte backbone. There exist several additional non-covalent interactions, including hydrogen bonding between amino acids and phosphate groups, cation-π interactions between positively charged amino acids and deoxyribose sugars, and van der Waals forces [83]. These additional forces allow for control of chemical architectures at a higher precision.

A relevant example of the interplay between these different molecular forces is the salt bridge between the side-chain guanidinium cation of arginine and the phosphate group of the DNA backbone, as illustrated in Figure 4.1. This salt bridge is one of the most common mechanisms by which histones bind to DNA [56, 17]. It consists of a combination of electrostatic attraction between the charged molecular entities and hydrogen bonds of the guanidinium nitrogens to the phosphate group oxygens.

The importance of the mechanical manipulation of DNA for the control of gene expression has led to the emergence of single-molecule biophysical [8] experiments that directly probe the molecular machinery operating on DNA at a nanoscale level. However, accurate

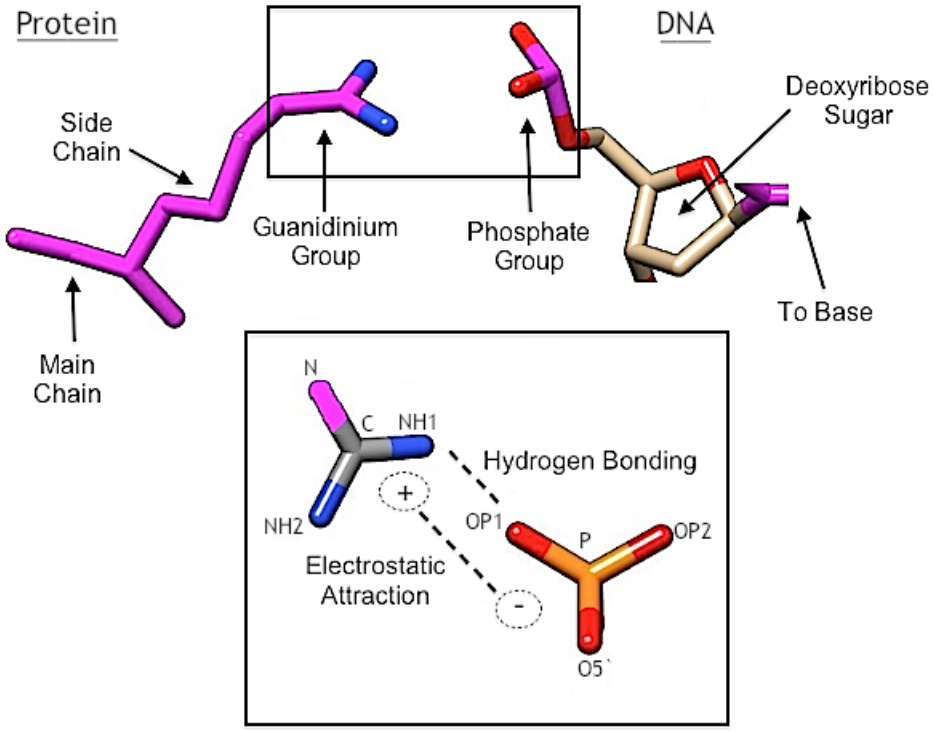

Figure 4.1: In a salt bridge between a histone protein and DNA, the guanidinium side-chain group of the amino acid arginine (top left) binds to the phosphate group of the DNA sugar-phosphate backbone (top right). This is done through a combination of: 1) Electrostatic attraction between the negatively-charged phosphate and positively-charged guanidinium, and 2) Hydrogen bonds between the two end-group nitrogens in guanidinium, labelled NH1 and NH2, and the two side-group oxygens on the phosphate, labelled OP1 and OP2. C and N label the carbon and the non-end group nitrogen on the guanidinium, respectively. $O5'$ is an oxygen connecting to the main chain of the sugar-phosphate backbone. Image created with Pymol [84].

quantum-mechanical modeling and simulation of these systems is relatively less mature. In particular, while first-principles calculations of 'hard' matter have sufficiently advanced to allow the predictive, atomic-level design of new materials before they are synthesized in the laboratory [31], they have not been similarly applied to the 'soft' biomolecular machinery in the cell. Historically, the key reason for this dearth of activity was the inability of traditional Kohn-Sham Density Functional Theory (KS-DFT) [33, 43] to account for the nonlocal London dispersion forces that are ubiquitous in soft matter.

The recent development of van der Waals density functional theory [18, 48] (vdW-DFT) has remedied this situation, expanding the realm of DFT to soft and biological materials. Subsequent applications of vdW-DFT have yielded novel atomistic insight into biologically important mechanochemical processes in DNA. Cooper et. al [14] studied the hydrogen bonding between base pairs and stacking interactions between nearest-neighbor nucleic acid base-pair steps, and illustrated the role of these interactions in determining sequence-specific elasticity. A follow-up study [110] investigated the 5-methylation of cytosines in $5'$-CG-$3'$: $5'$-CG-$3'$ base-pair steps, an epigenetic modification that is thought to trigger the protein-assisted compaction of chromatin [7].

Chemical changes to the nucleobases, however, are only a small piece of the elaborate epigenetic machinery controlling DNA structure. Further advances in the usefulness of density functional theory for molecular biophysics will inevitably require expanding its application to a more diverse group of biomolecular processes. In the context of the regulation of chromatin architecture, the formation of salt bridges between histones and DNA is a timely example of an important process that is ripe for investigation. While there have been previous quantum-mechanical studies of the energetics of the sugar-phosphate backbone [87, 63], including some work on arginine-phosphate interactions [25], such studies have not yet been attempted using the most recent vdW-DFT methods.

Useful application of first-principles calculations to biophysics, however, crucially requires that they not become divorced from the biological context of the problem at hand. In this regard, it is valuable to bridge the traditional gap between the electronic structure theory and structural bioinformatics communities. The latter can help with the judicious selection of biologically relevant molecular configurations to subject to more detailed atomistic modeling. In particular, principal component analysis (PCA) of a statistical ensemble

of experimental crystal structures reduces the intractably large phase space of possible molecular deformations to an 'essential subspace' of slow modes, or low-frequency collective motions most associated with biological functionality [107, 34]. Density functional theory can then provide quantitative information regarding how these functional motions are influenced by specific biochemical perturbations. Electronic structure calculations thus serve as a complement to single-molecule experiments, allowing a microscopic view of the detailed mechanochemical machinery operating within living cells.

With these guidelines in mind, the current work presents a novel investigation into the effects of guanidinium-phosphate salt bridges on the local conformational elasticity of the DNA sugar-phosphate backbone. After an introduction to the basic modeling setup of the problem, the relevant bioinformatics analysis and electronic structure procedures are described. The main results of the work are then presented and discussed. The principal components of fluctuation of the sugar-phosphate backbone are observed to encode sequence information, and DFT calculations illustrate that salt bridges non-covalently interact with the sugar-phosphate backbone in a complex, multi-faceted manner, enabling precision-controlled activation of various backbone deformations. The results have implications for how specific local histone-DNA interactions and positionings can stabilize and control more global, long-range elastic landscapes, an effect that is important for nucleosome positioning.

4.2 Modeling Setup

One of the first tasks in modeling arginine-DNA interactions is the selection of an appropriate 'model complex' that is a realistic representation of the actual salt bridge and is sufficiently simple to allow for detailed statistical analysis and atomistic calculations.

Such a complex should isolate the specific local effects of guanidinium-phosphate hydrogen bonding and electrostatics on DNA deformability.

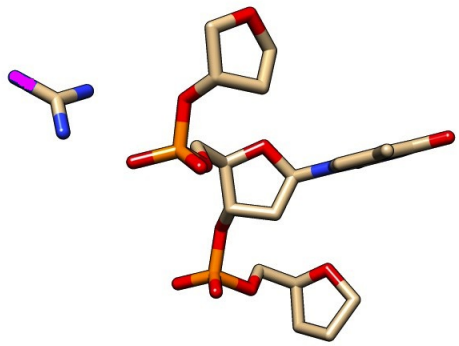

Figure 4.2: The model complex selected for this study consists of a guanidinium cation representative of the end-group of the arginine residues, and a collection of three deoxyribose sugars connected by two intermediate phosphate backbone linkages. Carbon atoms are colored beige, oxygen atoms red, phosphorus atoms orange, and all nitrogens blue except for the non-end group nitrogen of the guanidinium, which is colored purple. Hydrogen atoms are not illustrated for clarity. Image created with Pymol.

The model complex chosen for this study is illustrated in Figure 4.2. It strips off all atoms of the arginine amino acid except for the end-group guanidinium cation, which is the part that binds to the phosphate group. This binding alters the local flexibility of the DNA backbone, which is carried by the covalently-bonded chain of deoxyribose sugars and phosphate groups. Any model compound that is representative of this flexibility should, at a minimum, account for all nearest-neighbor interactions between nucleotide backbone units. One structure that meets these requirements is a combination of three deoxyribose sugars, with two intermediate phosphate groups, as well as one central nucleobase that incorporates the most dominant sources of sequence-dependent motions. Additional non-local interactions beyond neighboring nucleotides, while present, are likely to be less influential to DNA elasticity. They are beyond the scope of this study, and are a subject for future investigation.

4.2.1 Specifying the Configuration of the Model Complex

With the model complex selected, the question turns to determining an appropriate set of variables specifying its atomic coordinates. Such information is necessary both for determining average molecular configurations, and for characterizing the principal modes of fluctuation from this average. There are two parts to the problem: 1) Specifying the configuration of the sugar-phosphate backbone unit, and 2) Specifying the position and orientation of the guanidinium group with respect to the backbone.

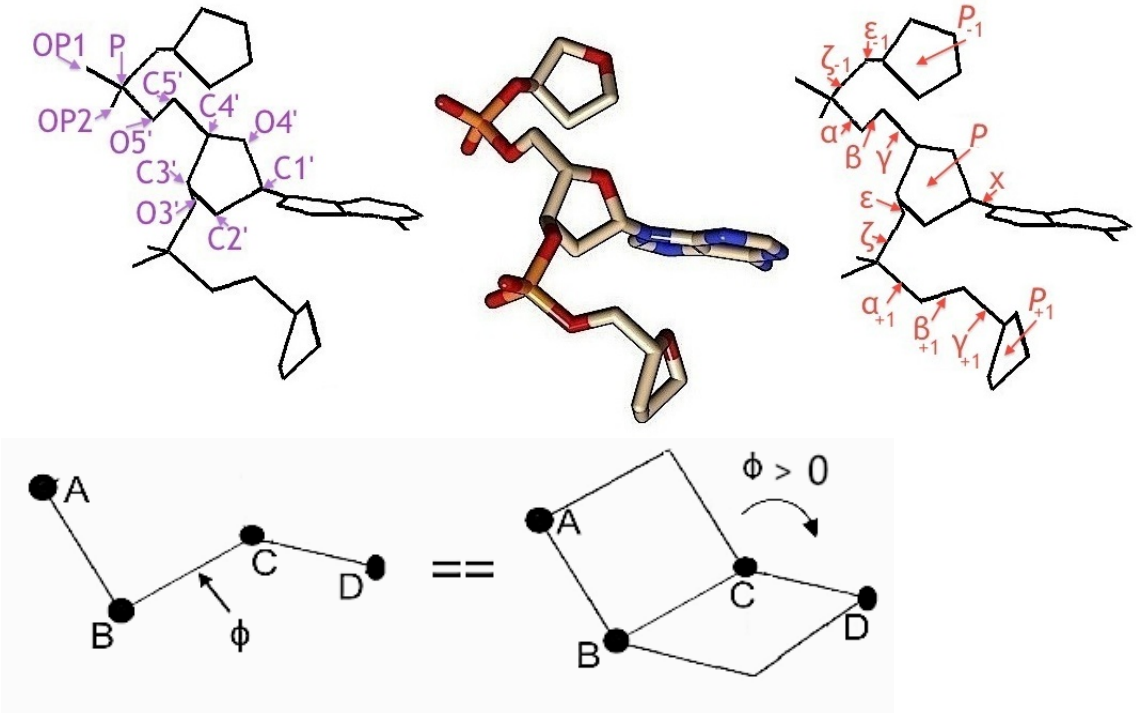

Figure 4.3: The detailed parameters specifying the conformation of the sugar-phosphate backbone in the model complex. (Top) From left to right are a stick image with selected non-hydrogen atoms labeled, an all-atom molecular graphic, and a stick image with the dihedral angles and pseudorotation phase angles labeled. In the all-atom molecular graphic, oxygen is colored red, phosphorus orange, carbon beige, and nitrogen blue, with hydrogens not shown for clarity. (Bottom) Displayed is the chosen positive sign convention for the dihedral angle ϕ between four atoms A-B-C-D, defined to be the angle between the planes formed by A-B-C and by B-C-D, with the angle taken to be zero when the atoms are in a planar, *cis* conformation.

Backbone Conformation

The problem of specifying the backbone coordinates is reduced by the observation that covalent bond lengths in crystal structures are, to a good approximation, fixed at experimentally prescribed values [26]. Furthermore, except for the covalent linkages formed by the deoxyribose sugars, bond angles are also approximately fixed. The conformation of the deoxyribose sugars, meanwhile, is well described by the phase angle of pseudorotation P, which specifies the puckering of the furanose ring [1]. With these simplifications, the backbone conformation is specified by the dihedral angles α, β, γ, ϵ, and ζ describing covalent bond links between adjacent sugars, the glycosydic torsion angle χ connecting the central deoxyribose to the nucleobase, and the pseudorotation phase angles P of the deoxyribose sugars, as illustrated in Figure 4.3.

Salt-Bridge Configuration

The specification of the coordinates of the guanidinium is simplified by the observation that its C-N bond angles and bond lengths vary negligibly from 1.33 Angstroms and 120°, respectively. Thus, the guanidinium cation can be treated as a rigid body with a trigonal planar geometry, and its position and orientation with respect to the backbone reduces to finding three translational and three rotational rigid body parameters, as shown in Figure 4.4. Without loss of generality, the phosphorus atom can be defined to be the origin, with the side-group oxygens OP1 and OP2 positioned symmetrically in the $y-z$ plane. The three translational parameters of the guanidinium can be taken to be the position vector $\vec{r}$ of the central carbon C with respect to the phosphorus. Two angles θ and ϕ then set the orientation of the non-end-group nitrogen N, and an angle ω describes the remaining rotational freedom of the end-group nitrogens NH1 and NH2.

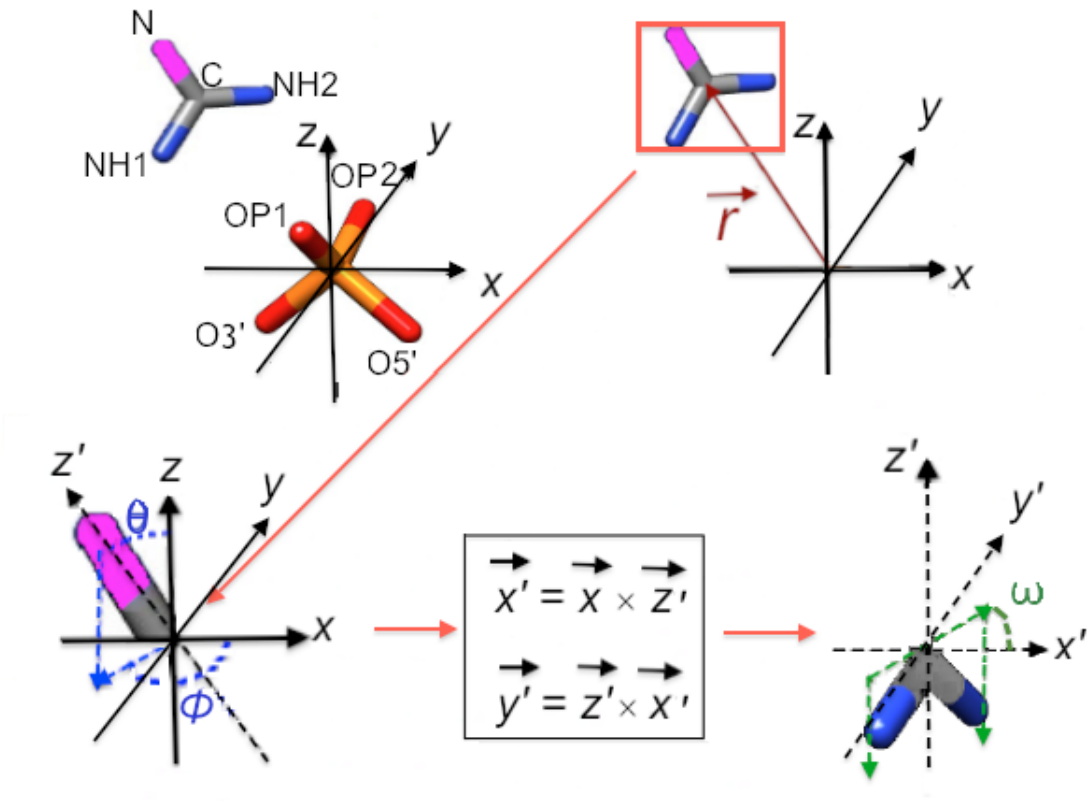

Figure 4.4: Displayed is a schematic of the six variables necessary to represent the configuration of a guanidinium cation with respect to a phosphate group. Coordinates are chosen so that phosphorus lies at the origin, OP1 and OP2 lie in the $y-z$ plane at equal and opposite values of y, and O5$'$ lies in the $x-z$ plane, with positive x and negative z. With this choice of coordinate frame, the translational parameters of the guanidinium are specified by the vector $\vec{r}$ describing the displacement of the guanidinium carbon C from the phosphorus atom P. The position of this carbon is then taken to be the origin of a new set of coordinates $x^{'}$, $y^{'}$, and $z^{'}$. These coordinates are defined such that the non-end-group nitrogen N lies on the positive $z^{'}$-axis, the $x^{'}$-axis is set by the cross product of the x- and $z^{'}$-axes, and the $y^{'}$-axis is set by the cross product of the $z^{'}$- and $x^{'}$-axes. With this set of coordinates, the rotational degrees of freedom are given by the Euler angles θ and ϕ that the $z^{'}$-axis makes with respect to the z-axis, and the angle ω that the NH1-NH2 vector makes with the $x^{'}$-axis. Images created with Pymol.

4.3 Methods

In this section, the procedures for determining the principal components of backbone deformation and primary clusters of guanidinium-phosphate interaction are described. Subsequently, the methodology for electronic structure calculations is expanded upon. Particular focus is given to how these calculations couple to the bioinformatics analyses.

4.3.1 Extracting Functional Motions from Crystal Structures

Principal Component Analysis of the Sugar-Phosphate Backbone

Statistical analysis is performed on a non-redundant dataset of protein-bound DNA obtained from the Nucleic Acid Database [5] and reported in a previous publication [110]. From this dataset, a fourteen-parameter data vector is generated that characterizes the atomic configuration of the model complex illustrated in Figure 4.3. This vector includes a single glycosidic base-sugar torsion angle χ, three sugar pucker phase angles P for each of the three deoxyribose sugars (converted to Cartesian coordinates using an algorithm previously developed by Olson [67]), and ten dihedral angles along the backbone. The total collection of data vectors is then sorted into four groups based on the identity of the central nucleobase, and each group of data is separately standardized and subjected to principal component analysis. Using a scree test, the four highest eigenvalues, corresponding to dominant modes of deformation, are extracted for each group.

Clustering of Guanidinium-Phosphate Salt Bridges

The analysis of DNA-histone interactions is performed on an ensemble of 83 high-resolution crystal structures of nucleosomal DNA. Within this ensemble, arginine-phosphate contacts are observed to be the most common mode of interaction between the histones and the sugar-phosphate backbone. Thus, an initial dataset is created, consisting of 1556 structural examples in which an arginine nitrogen is less than 4.0 Angstroms away from the phosphorus atom.

This dataset is further curated so that it only includes structures in which the minimum distance between an end-group nitrogen (NH1 or NH2) and a side-group oxygen (OP1 or OP2) is at least 1.6 Angstroms less than the minimum distance between the non-end-group

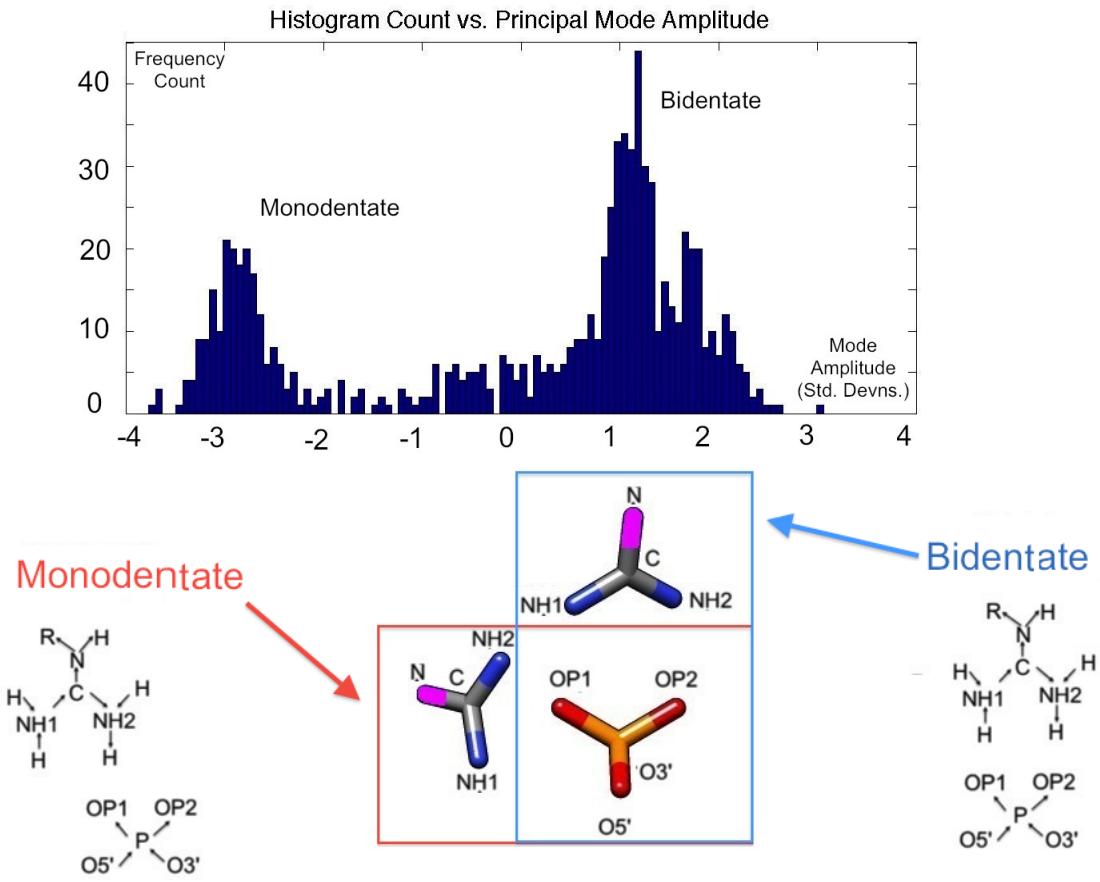

Figure 4.5: Histogram of the frequency count of the amplitude of the dominant principal component, as measured by standard deviations from the average value, with one hundred equally spaced bins from -4 to 4. The component is observed to peak around two central clusters: 1) 'monodentate' bridges, in which only one hydrogen bond is formed between guanidinium and phosphate, and 2) 'bidentate' bridges, in which two hydrogen bonds are formed, as displayed. This justifies, for an initial study, a 'mean-field' approximation in which the configuration of the guanidinium cation can be taken as adopting one of two 'average' values. These average values are determined by K-means clustering. Histogram created in MATLAB [61].

nitrogen (N) and a side-group oxygen. This step is necessary to remove any 'anomalous' structures in which the guanidinium cation may not be hydrogen bonded to the phosphate through the end-group nitrogens. While it is conceivable that arginines may interact with the phosphate in ways different from this, including for example hydrogen bonding of the non-end-group nitrogen to the phosphate, such interactions are beyond the scope of the present analysis, and are a subject for future investigation. As it turns out, the chosen

constraints account for over half of all significant arginine-phosphate interactions, resulting in a working dataset of 790 structural examples of guanidinium-phosphate salt bridges.

From this working dataset, a six-parameter data vector ($\vec{r}$, θ, ϕ, ω) is generated that characterizes the configuration of a salt bridge, as described in Figure 4.4. This collection of data vectors is then standardized and subjected to principal component analysis. A scree test determines that only the first principal component carries a significant fraction of the total variance. Furthermore, a histogram of the frequency distribution of the amplitude of the first principal component, displayed in Figure 4.5, indicates that the data are localized around two strongly peaked regions: 1) A monodentate cluster, in which only OP1 is hydrogen bonded to an end-group nitrogen, and 2) A bidentate cluster, in which both OP1 and OP2 bond to a separate end-group nitrogen. Because of the sharpness of these peaks, it can be assumed, in a 'mean-field' approximation, that the salt bridges only adopt two distinct states corresponding to the centers of each of these clusters. The data are thus sorted by K-means clustering, and the central average of each cluster is taken as one of two possible salt-bridge orientations.

4.3.2 Calculating Energy Landscapes with Density Functional Theory

Having determined both the functional modes of deformation of the backbone, and a representative set of guanidinium-phosphate salt-bridge clusters, the next task is to determine, with vdW-DFT, the elastic energy of deformation of each of the modes in both the absence and presence of different salt bridges. The first step is to sample a series of points along the configurational pathway of each mode, and calculate the energy of each point in the absence of any guanidinium group. From these initial computations, a set of low-energy points along the landscape is determined. Calculations on these low-energy points are then

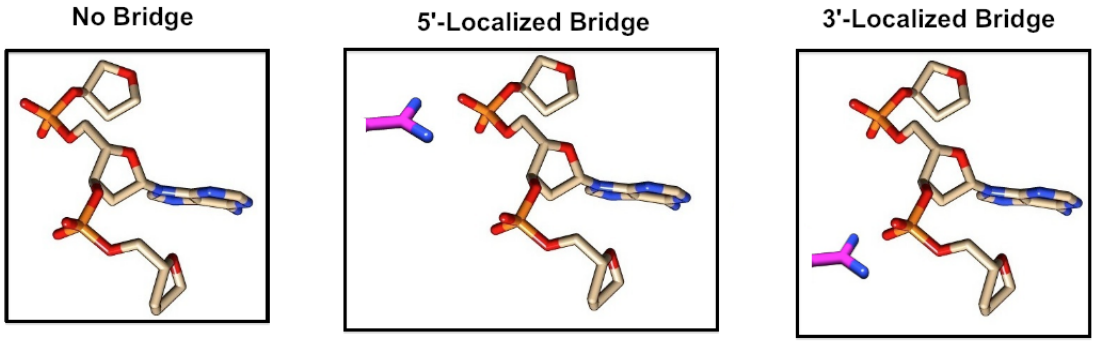

Figure 4.6: The energy landscape of each principal component of the backbone is calculated using density functional theory. The energies are first calculated in the absence of any guanidinium cation (left). Calculations are then repeated with salt bridges localized on the $5'$-phosphate (middle) and $3'$-phosphate (right). This procedure is repeated for both the monodentate and bidentate configurations determined by K-means clustering, leading to a total of four different salt-bridge environments being simulated. The example salt bridge on the left is of a bidentate form, and the example salt bridge on the right is of a monodentate form. Images created with Pymol.

repeated for each of four different types of guanidinium-phosphate salt bridges, namely the set of all combinations of bridges that are localized around the $5'$ or $3'$ phosphate group and which lie in either a monodentate or bidentate orientation.

DFT calculations are performed with the vdW-DF2 functional [48], as implemented in the Quantum Espresso package [27] via the algorithm developed by Roman-Perez and Soler [82]. Standard generalized gradient approximation pseudopotentials [96] are employed (these pseudopotentials are available free of charge on the Quantum Espresso website, and were generated by Tozzini et. al in [95]), with a kinetic energy cutoff of 60 Ry (1 Ry = 313.755 kcal/mol). SCF diagonalizations are performed with a convergence criteria of 10^{-6} Ry. To ensure efficient convergence of the energy in the presence of the net charges of the phosphate ions, a Makov-Payne electrostatic correction term [57] is added. Spurious interaction between artificial periodic images is reduced by placing the system in a cubic supercell of side length 36 Bohr (1 Bohr = 0.529 Angstroms).

4.4 Results and Discussion

The two main results of this work are that: 1) the configurational fluctuations of the sugar-phosphate backbone, as represented by the dominant principal components, display sequence specificity, and 2) the guanidinium cations interact with the sugar-phosphate backbone to tunably 'freeze in' specific backbone deformations. This section begins by discussing the molecular character of the principal components, paying particular attention to signatures of sequence specificity. This is then followed by a presentation of the insights gleaned from DFT calculations, in particular, the observation that the main effect of the guanidinium cations is to apply an approximately linear mechanical stress to the backbone. This stress displays an intricate dependence on many different tunable 'knobs', including the chemical identity of the central nucleobase, the choice of phosphate on which the guanidinium cation is localized, and the number of hydrogen bonds that the guanidinium makes with the phosphate. These effects have direct implications for the robust and adaptive control of nucleosome positioning.

4.4.1 Backbone Motions: Bending Virtual Bonds

Full results regarding the quantitative coefficients and fractions of total variance captured by each of the four highest principal components are presented in the Appendix. Here, the focus shall be on developing an intuition regarding the qualitative character of these principal modes. In order to develop this intuition, it is useful to switch from the all-atom picture of the backbone to a more coarse-grained view, in which the covalent chain connecting the C1$'$ atoms on adjacent sugars is represented as a 'virtual bond' [22, 73], as illustrated in Figure 4.7. This perspective allows the motion of the backbone to be expressed in terms of deformations of a virtual triatomic 'molecule', analogous to the well-studied IR

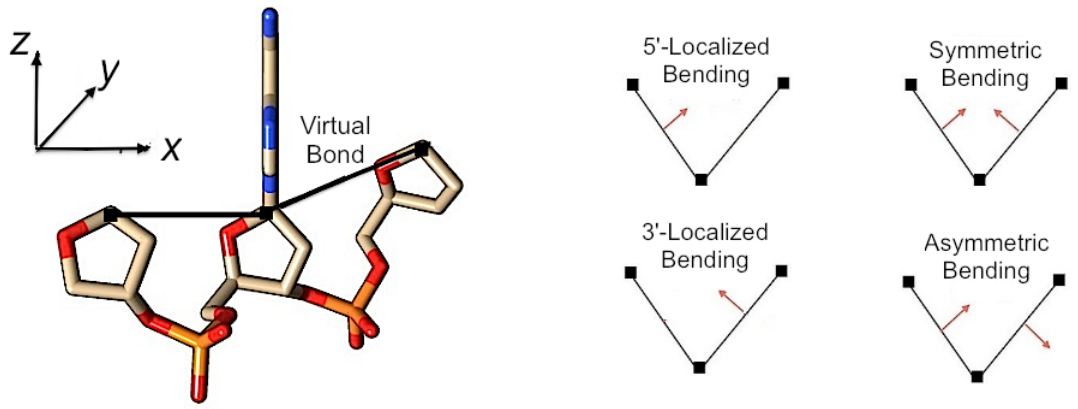

Figure 4.7: (Left) The motions of the sugar-phosphate backbone unit can be simplified by using a reduced description in terms of 'virtual bonds' between the C1′ atoms on each of the deoxyribose sugars. Then, the complicated collection of atoms in the nucleotide is reduced to a simple virtual triatomic 'molecule'. Image created with Pymol. (Right) The deformations of a linear triatomic molecule can be described in terms of the relative motions of each of the two bonds.

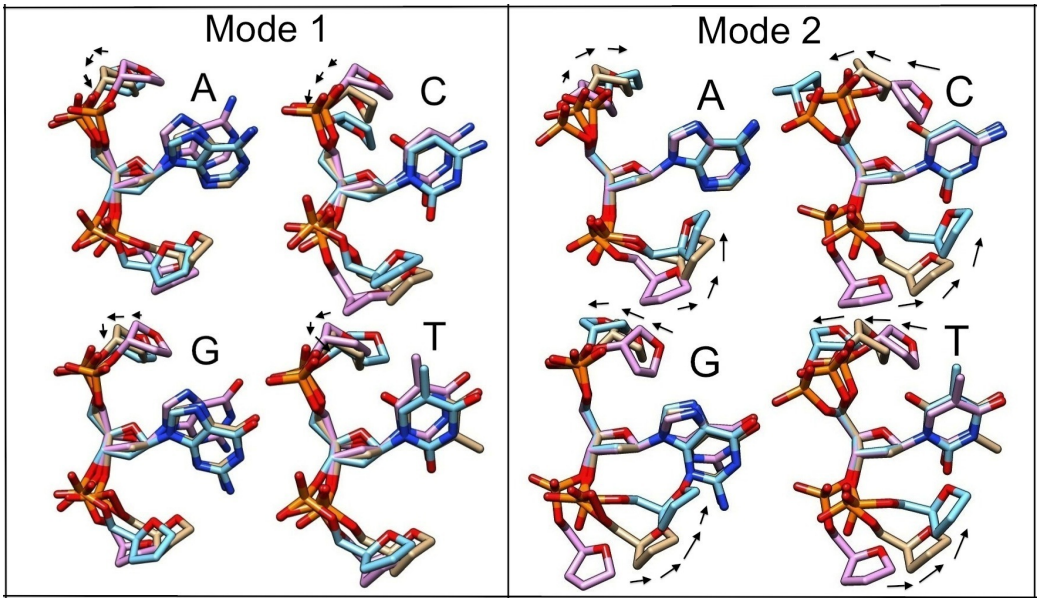

Figure 4.8: Displayed here are the first and second principal modes of deformation, for each of the four different nucleobases. Images are superimposed such that the C1′, C3′ and C4′ atoms of the central deoxyribose sugar are fixed in position. Any bending motions are accompanied by black arrows guiding the direction of motion. The molecular images are color coded such that the beige carbon colored units are associated with the average backbone conformation, and the pink and blue carbon colored units are associated with -1 and 1 standard deviations of deformation away from the average, respectively. Images created with Pymol.

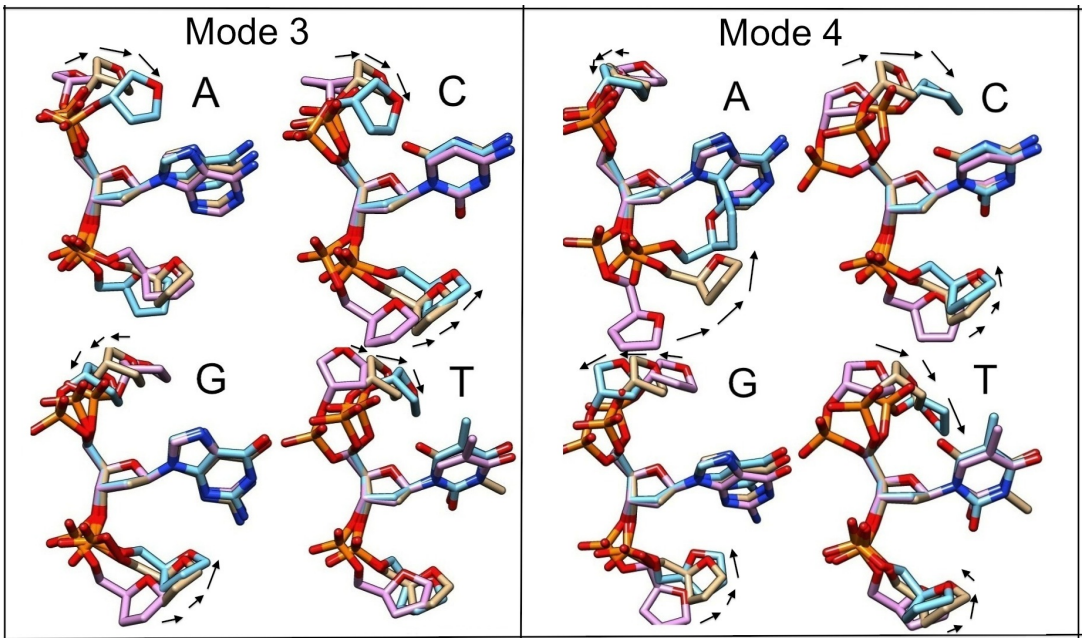

Figure 4.9: Displayed here are the third and fourth principal modes of deformation, for each of the four different nucleobases. For detailed annotation, see the caption of Figure 4.8.

vibrations of more well-known triatomic molecules such as H_2O [3].

For a particular principal component, each virtual bond can be viewed as either increasing, decreasing, or negligibly changing the angle that it makes with respect to the central nucleobase. Then, to a first approximation, a particular combination of bond motions can be characterized as being in one of four classes: 1) $5'$-Localized bending, in which only the $5'$-end sugar appreciably moves; 2) Symmetric bending, in which the two bonds move 'in-phase'; 3) $3'$-Localized bending, in which only the $3'$-end sugar appreciably moves; 4) Asymmetric bending, in which the two bonds move 'out-of-phase'. A schematic of the various bending combinations is displayed in Figure 4.7.

The principal components are observed to display a complex dependence on the chemical identity of the central nucleobase. In spite of this, some general patterns and trends do emerge, as displayed in Figures 4.8 and 4.9. The principal component with the highest amount of the total variance, hereafter labeled the first principal component, displays the

least amount of qualitative sequence dependence, adopting a $5'$-localized bending motion in which the $3'$-end sugar merely rotates in position.

Sequence behavior becomes much more diverse for the second, third and fourth components. The second principal component is observed to take the form of asymmetric bending for cytosine, guanine and thymine, but adopts a symmetric bending for adenine. The third principal component demonstrates adenine and thymine performing $5'$-localized bending, cytosine symmetrically bending, and guanine asymmetrically bending. And finally, the fourth principal component displays a dependence on purine vs. pyrimidine character, being an asymmetric bend for adenine and guanine but a symmetric bend for cytosine and thymine. While these classifications are only qualitative heuristics, they serve to demonstrate the point that the fluctuations of the sugar-phosphate backbone encode sequence information.

4.4.2 Tuning Energy Landscapes via Adjustment of Salt Bridges

The full results of energy vs. mode amplitude for each of the different modes and salt-bridge configurations are relegated to the Appendix. The main text focuses on extracting the energetic effect of the salt bridges, in particular the generation of an approximately linear mechanical stress signal that couples to each of the principal components in a sequence-specific manner.

To extract this signal, plots of energy vs. mode amplitude are generated for each of the modes in both the presence and absence of salt bridges. From these plots, the energy landscapes of salt-bridged modes are decomposed into the sum of a reference landscape with no salt bridge present and a perturbation that reflects the elastic energy contribution arising

from the presence of the guanidinium group. This perturbation contribution is then least-squares fit to a linear function, $\Delta E(\lambda) = \sigma_\lambda \ \lambda$, where λ is the amplitude of the principal component in units of standard deviations from the mean. The resulting coefficient σ_λ is the linear mechanochemical stress along the axis of deformation of the principal component. An illustration of the procedure is given in Figure 4.10 and a full display of the resulting stresses σ_λ is shown in Figure 4.11.

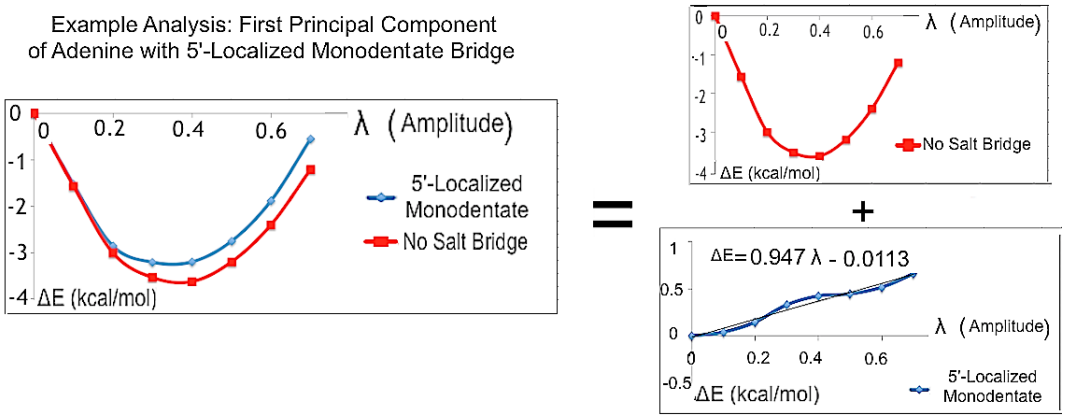

Figure 4.10: Displayed here is an example of the procedure used to extract the mechanochemical stress σ_λ induced on a particular principal component λ by a particular type of salt-bridge configuration. In this case, the illustration is provided by the $5'$-localized monodentate bridge on the first principal component of adenine. The energy landscapes along the mode are computed both with and without the salt bridge, and plots are standardized so that the point of zero mode amplitude is the zero-point reference energy. This allows the energy landscape of the mode in the presence of the salt bridge to be decomposed into the sum of the landscape in the absence of the salt bridge and an approximately linear component representative of the effects of the salt bridge. This component is least-squares fit to a line, and the resulting slope approximates σ_λ. This procedure can then be repeated for each of the other three different salt-bridge configurations, and then further repeated for all the different principal components and nucleobases.

As seen in Figure 4.11, the salt-bridge induced stresses display a complex multi-pronged dependence on base sequence, salt-bridge denticity, and phosphate positioning of the guanidinium group. Even for the first principal component, in which the atomic deformation is a $5'$-localized bending irrespective of base identity, the nature of the salt-bridge induced stresses and their dependence on denticity and positioning differs for adenine as compared

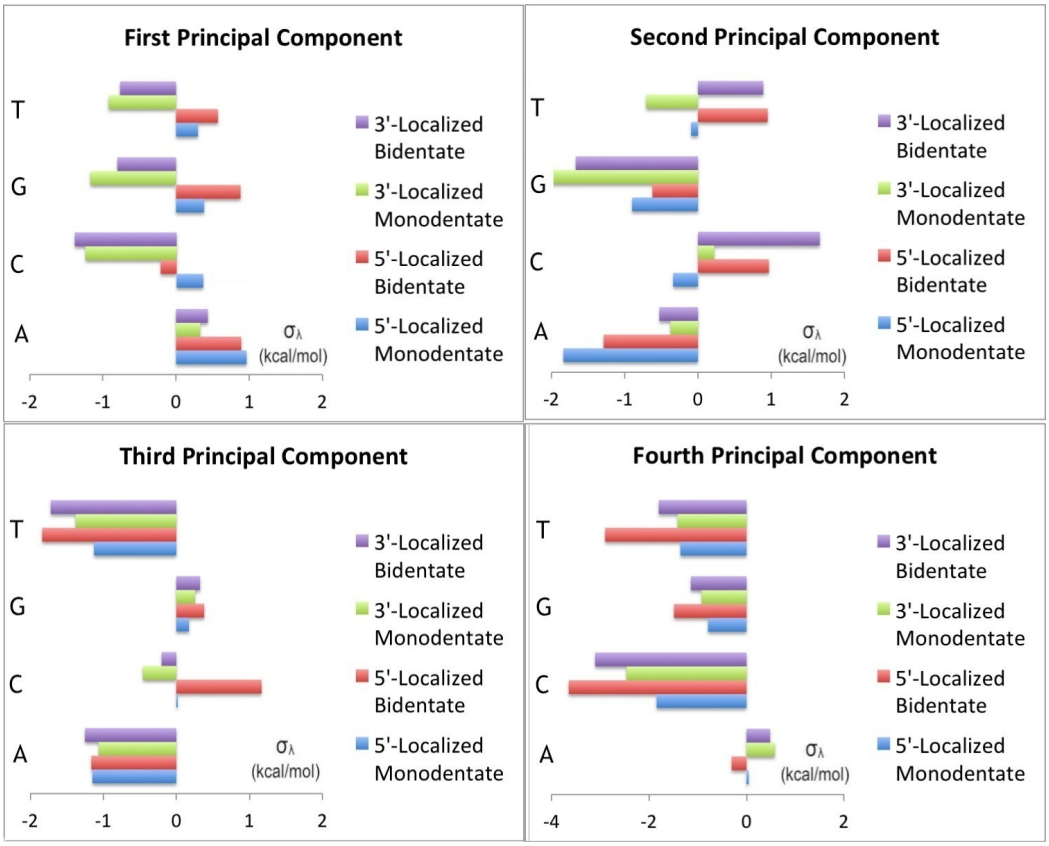

Figure 4.11: Presented here are the results for the different mechanochemical stresses σ_λ for each of the four principal components in the presence of different salt-bridge forms and different central nucleobases. The x-axis displays mechanochemical stresses, with units of kcal/mol resulting from the fact that mechanochemical stresses are defined as changes in energy over change in unitless principal mode amplitude.

to cytosine, guanine, and thymine.

These effects continue to hold true for other groups of similar deformations. For the second principal component, in which cytosine, guanine, and thymine all asymmetrically bend, the coupling of mechanical stress to salt-bridge denticity and positioning is different for the purine guanine compared to the pyrimidines cytosine and thymine. The third principal component, which groups adenine and thymine together as $5'$-localized bends, shows that the mechanical effect of the salt bridge on adenine is slightly weaker than it is on thymine.

4.4.3 Discussion

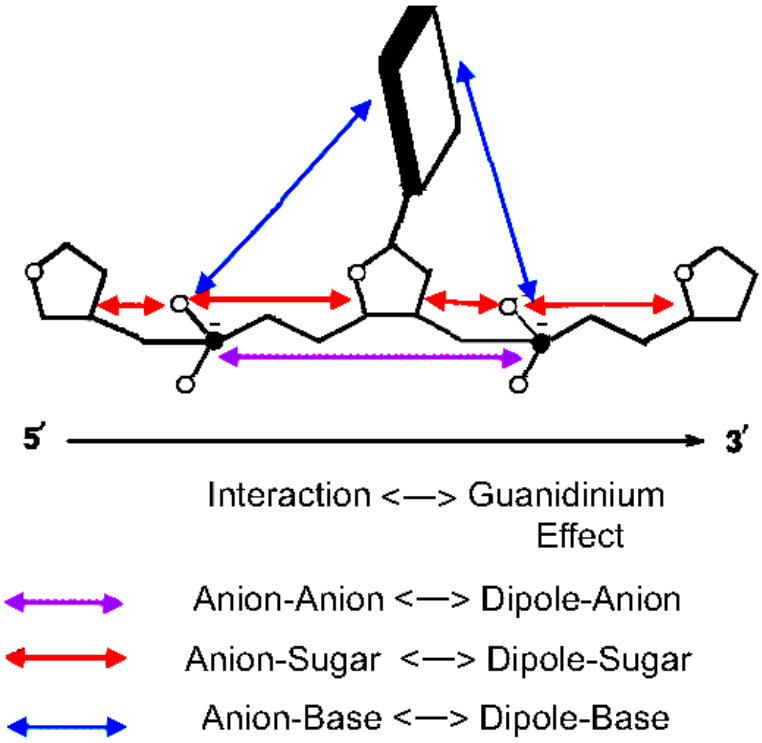

Figure 4.12: The anionic phosphate groups non-covalently interact with each other, the deoxyribose sugars, and the aromatic nucleobases. The combined effect of these nonlocal forces is an 'intermolecular' stress arising near the phosphate group. When a guanidinium cation neutralizes one of the phosphate groups, it also modifies these non-covalent interactions and the resulting intermolecular stress. Modification of denticity and positioning further adjust the character of the non-covalent interactions, allowing for a diverse array of tunable 'knobs' that induce particular kinds of mechanical deformation.

Chemically, the stabilization of various principal components arises from a combination of 'intermolecular', non-covalent effects. The anionic phosphate groups interact electrostatically with each other, and through ion-pi interactions with the aromatic ring sugars and nucleobases. When a guanidinium cation is present, the anionic phosphate group has been reduced to a dipole, and the tuning of denticity tunes the dipole magnitude and orientation. The electrostatic ion-ion interactions of the phosphate groups are reduced to ion-dipole interactions, and the interactions with the aromatic sugars and nucleobases are

likewise modified to have a larger contribution from non-electrostatic London dispersion forces.

The result is an effective mechanical load arising from the complex interplay of these different noncovalent interactions, and it is this mechanical load that causes the linear mechanochemical stress which activates specific combinations of principal component deformations. Altogether, the combination of sequence, salt-bridge positioning, and denticity serves as a collection of tunable 'knobs' that histones can use to locally activate particular combinations of backbone deformations.

Implications for Nucleosome Positioning

From the point of view of nucleosomes, one of the most interesting consequences of the salt bridges is their modulation of the helical periodicity of the DNA backbone. In canonical DNA forms, such as B-DNA or undertwisted A-DNA, the backbone torsion angles display a consistent periodicity commensurate with the spacing between adjacent base-pairs. In other words, the torsion angles α, β, γ, ϵ, and ζ are equal to $\alpha + 1$, $\beta + 1$, $\gamma + 1$, $\epsilon + 1$, and $\zeta + 1$, respectively. However, the principal modes of deformation do not necessarily obey this periodicity, as seen most notably in modes that tend toward $5'$-localized and asymmetric bending type character. As a result, the histones effectively apply an elastic modulating signal to the DNA, arising from the collection of guanidinum-phosphate salt-bridge contacts within the nucleosome. By tuning these local sites of DNA-histone binding, the shape and size of the modulating signal can be controlled. In turn, this size and shape alter the equilibrium positioning of various mechanical deformations, such as the wrapped pathways characteristic of nucleosomes.

A further remarkable feature of biological evolution is the high degree of precision with

which these delicate elastic modulations are controlled. The sensitivity of DNA deformations to multiple different variables endows the chromatin with a tremendous amount of adaptability, which enables it to maintain the homeostatic stabilization of nucleosome positions under a diverse set of possible environmental perturbations. At the same time, however, this substantial sensitivity creates an equally substantial challenge concerning the maintenance of robust control. A greater set of sensitive variables for adaptation also means there is a greater set of variables that need to be tightly regulated to maintain a normal biological stasis.

This significance of evolution in determining nucleosome positioning has been increasingly recognized over the past few decades. It has been suggested that there is a genomic code for nucleosome positioning [85], with evolutionary conservation of high-affinity nucleosome binding sequences. Additionally, it has been further realized that these sites of high-affinity nucleosome binding tend to repeat themselves at well defined 10 base pair periodicity as opposed to being randomly distributed, a phenomenon known as nucleosome phasing [90, 53].

The results of this study show that a possible evolutionary design principle underlying nucleosome phasing is in the selection for variables that sensitively tune DNA backbone deformations in order to control the nucleosomal wrapping. As the present theoretical calculations show, these deformations are sensitive to both the sequence and positioning of histone-DNA contacts. Furthermore, the diverse nature of these contacts displays a much richer phenomenology than simply electrostatic bindings of cationic amino acids and the anionic phosphate backbone, demonstrating the importance of the relatively underappreciated many body van der Waals interactions in controlling chromatin structure at nanoscopic and mesoscopic length scales.

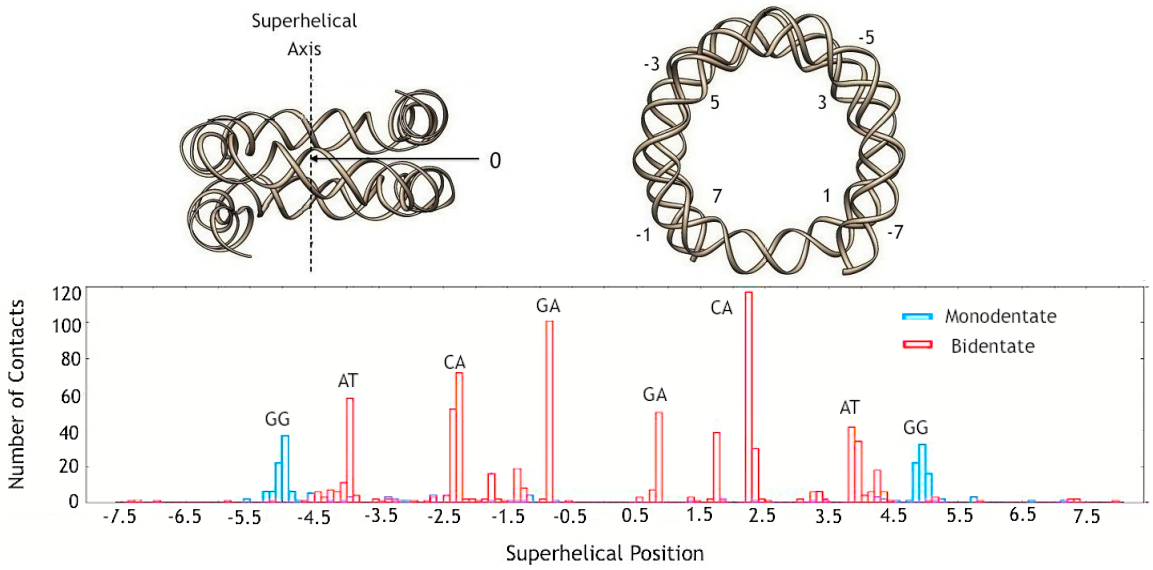

Figure 4.13: (Top) An illustration of the shape of the DNA sugar-phosphate backbone for a high-resolution nucleosomal crystal structure[56], PDB ID 1kx5, displayed from both a side view and a top-down view. The histones and DNA bases have been removed for clarity. The 147 base pairs of nucleosomal DNA can be viewed as consisting of approximately 15 superhelical turns of roughly 10 base pairs each, with position along the nucleosome consequently labelled by these superhelical positions and ranging from -7.5 to 7.5. With this labeling convention, 0 represents the dyad, or the midpoint of the nucleosome that is spatially sandwiched in between the entry and exit points of the nucleosome. (Bottom) A histogram of the frequency of monodentate and bidentate arginine contacts, as a function of superhelical position, for the 83 nucleosomal crystal structures used in this study. The contacts are observed to localize in well defined clusters. Particular clusters, which are observed to favor specific DNA sequences, are further annotated with the corresponding sequence.

Additionally, this work has demonstrated the importance of denticity, a relatively unexplored variable with the potential to be affected by evolution. The number of hydrogen bonds is observed to be of comparable importance to backbone deformation as underlying base sequence and contact positioning. Thus, it is worthwhile to ask if the distribution of such contacts displays similarly non-random behavior characteristic of natural selection. Figure 4.13 displays a histogram of the nucleosome positions of both monodentate and bidentate contacts in the 83 nucleosomal crystal structures analyzed in this work. As the

data show, the positioning of specific types of contacts is far from random, but instead distributed in very well localized clusters, with the clusters being strongly biased to particular sequence motifs.

Monodentate contacts, in particular, are found to be strongly localized at ± 5 superhelical turns with respect to the central nucleosomal dyad. The placement of these contacts is commensurate with regions of the nucleosome that previous researchers [80, 50, 92, 41, 40] have associated with a high affinity for 'invasion' by DNA binding proteins, an effect that is important for active nucleosome remodeling. This makes sense from the point of view of nucleosome stability, since a contact with one hydrogen bond is more easily disrupted than a contact with two. The present work suggests that this requirement of reduced stability must be met while also simultaneously maintaining very specific shape requirements for the nucleosome, resulting in evolutionary selection pressure for a very precise spatial distribution of the necessary van der Waals contacts, and thus of denticity, contact positioning, and DNA sequence. Incidentally, this is also consistent with related work indicating that nucleosome structure and function are highly sensitive to histone sequence, and are in fact disrupted by SIN point mutations of histones [46, 21, 105] which could potentially interfere with the van der Waals contacts.

4.5 Conclusions

In summary, this work has presented a novel integration of structural bioinformatics and van der Waals density functional theory to investigate the effects of a major histone-DNA interaction, the formation of salt bridges between guanidinium arginines and the DNA phosphate group, on the deformations of the DNA sugar-phosphate backbone. Guanidinium-phosphate complexes are observed to occur in both bidentate and monodentate salt-bridge

configurations. The combined interplay between denticity, chemical identity of nucleobases, and positioning of the guanidinium group creates a rich array of different mechanochemical stress input signals. These equip the histones with a versatile toolkit for the precise stabilization and control of nucleosome positioning.

4.6 Appendix

Displayed below are: 1) the raw numerical output results of the principal component analysis, 2) the raw output of the DFT energy calculations, 3) an explanation of the procedure used to convert pseudorotation phase angles of the deoxyribose sugar into Cartesian coordinates, and 4) details on salt-bridge clustering. Not present in this dissertation, but to be included in the Supplemental Information of the published version of the manuscript [111], is a compilation of references to original literature for the 83 nucleosomal crystal structures, including annotated output tables of monodentate and bidentate bridges.

4.6.1 Converting Pseudorotation P to Cartesian Coordinates

In order to generate atomic coordinate input structures for electronic structure calculations, the experimentally determined pseudorotation phase angle P must be translated into a set of Cartesian coordinates. In practice, this is done by specifying, for fixed bond lengths, a set of three-atom bond angles $\{\theta_i\}$ and four-atom dihedral angles $\{\phi_i\}$ sufficient enough to determine the coordinates of the deoxyribose sugar ring and its connection to the backbone main-chain. However, before that can be done, it is first necessary to determine a numerical function for inverting from a given phase angle P to a particular bond angle θ or dihedral angle ϕ, in other words, to determine quantitative forms of $\theta = \theta(P)$ and $\phi = \phi(P)$ for all necessary θ and ϕ angles.

Phase P	0.00	18.00	36.00	54.00	72.00	90.00	108.00	126.00	144.00	162.00
(O4-C4-C3)$_{-1}$	106.17	105.10	103.78	103.51	103.78	104.69	105.32	106.13	106.39	106.39
(C4-C3-O3)$_{-1}$	113.32	113.42	113.14	112.72	112.13	111.67	111.36	111.52	111.97	112.48
(O4-C4-C3-O3)$_{-1}$	207.11	202.09	200.40	202.85	208.83	217.29	227.77	238.19	251.18	261.02
(C4-C3-C2)$_{-1}$	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.23
(O4-C4-C3-C2)$_{-1}$	329.00	324.01	322.19	324.45	330.21	338.47	348.80	359.31	372.52	382.57
(C4-O4-C1)$_{-1}$	110.02	109.67	109.31	107.93	107.50	106.99	107.50	107.93	109.31	109.68
(C3-C4-O4-C1)$_{-1}$	12.04	22.61	31.62	37.79	38.60	36.57	31.32	24.15	11.97	0.57
C5-C4-C3	111.08	111.62	112.26	112.36	112.23	111.80	111.51	111.12	111.00	110.96
C4-C3-O3	113.32	113.42	113.14	112.72	112.13	111.67	111.36	111.52	111.97	112.48
C5-C4-C3-O3	86.26	80.93	79.00	81.41	87.42	95.99	106.67	117.26	130.32	140.23
C5-C4-O4	111.06	111.62	112.22	112.34	112.23	111.86	111.50	111.13	110.99	110.96
C3-C5-C4-O4	242.07	242.70	243.56	243.76	243.57	242.95	242.58	242.06	241.89	241.95
C4-C3-C2	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.23
C5-C4-C3-C2	208.15	202.85	200.78	203.01	208.81	217.17	227.70	238.38	251.66	261.78
C4-O4-C1	110.02	109.67	109.31	107.93	107.50	106.99	107.50	107.93	109.31	109.68
C5-C4-O4-C1	132.91	143.78	153.05	159.24	160.01	157.83	152.42	145.08	132.83	121.36
O4-C1-N	111.44	110.94	110.99	111.13	111.50	111.79	112.22	112.34	112.24	111.70
C4-O4-C1-N	251.20	239.65	227.17	214.92	207.54	202.24	199.98	200.75	206.97	215.17
(C5-C4-C3)$_{+1}$	111.08	111.62	112.26	112.36	112.23	111.80	111.51	111.12	111.00	110.96
(C5-C4-O4)$_{+1}$	111.06	111.62	112.22	112.34	112.23	111.86	111.50	111.13	110.99	110.96
(C3-C5-C4-O4)$_{+1}$	242.07	242.70	243.56	243.76	243.57	242.95	242.58	242.06	241.89	241.95
(C4-C3-C2)$_{+1}$	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.23
(C5-C4-C3-C2)$_{+1}$	208.15	202.85	200.78	203.01	208.81	217.17	227.70	238.38	251.66	261.78
(C4-O4-C1)$_{+1}$	110.02	109.67	109.31	107.93	107.50	106.99	107.50	107.93	109.31	109.68
(C5-C4-O4-C1)$_{+1}$	132.91	143.78	153.05	159.24	160.01	157.83	152.42	145.08	132.83	121.36

Table 4.1: Furanose bond and dihedral angles vs. pseudorotation phase $P < 180$.

Phase P	180.00	198.00	216.00	234.00	252.00	270.00	288.00	306.00	324.00	342.00	360.00
(O4-C4-C3)$_{-1}$	106.17	105.10	103.78	103.51	103.78	104.69	105.32	106.13	106.39	106.44	106.17
(C4-C3-O3)$_{-1}$	113.37	113.41	113.18	112.73	112.14	111.64	111.38	111.52	111.94	112.52	113.32
(O4-C4-C3-O3)$_{-1}$	269.08	274.07	276.01	273.93	268.40	260.37	250.16	239.56	226.16	215.27	207.11
(C4-C3-C2)$_{-1}$	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.20	101.42
(O4-C4-C3-C2)$_{-1}$	391.00	395.99	397.81	395.55	389.79	381.53	371.20	360.69	347.48	336.78	329.00
(C4-O4-C1)$_{-1}$	110.02	109.67	109.31	108.28	107.50	106.99	107.50	107.93	109.31	109.67	110.02
(C3-C4-O4-C1)$_{-1}$	-12.04	-22.61	-31.62	-36.94	-38.60	-36.57	-31.32	-24.15	-11.97	0.51	12.04
C5-C4-C3	111.08	111.62	112.25	112.40	112.24	111.82	111.50	111.13	110.97	110.99	111.08
C4-C3-O3	113.37	113.41	113.18	112.73	112.14	111.64	111.38	111.52	111.94	112.52	113.32
C5-C4-C3-O3	148.19	152.91	154.58	152.39	146.98	139.12	129.09	118.64	105.32	94.39	86.26
C5-C4-O4	111.10	111.61	112.24	112.41	112.24	111.82	111.48	111.18	110.99	110.98	111.06
C3-C5-C4-O4	242.05	242.71	243.55	243.69	243.56	242.97	242.60	242.06	241.92	241.84	242.07
C4-C3-C2	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.20	101.42
C5-C4-C3-C2	270.10	274.84	276.38	274.01	268.36	260.28	250.13	239.77	226.65	215.90	208.15
C4-O4-C1	110.02	109.67	109.31	108.28	107.50	106.99	107.50	107.93	109.31	109.67	110.02
C5-C4-O4-C1	108.84	98.54	89.82	84.59	82.82	84.68	89.77	96.78	108.85	121.39	132.91
O4-C1-N	121.46	110.96	110.96	111.13	111.48	111.86	112.26	112.38	112.24	111.62	111.44
C4-O4-C1-N	226.71	238.63	251.18	261.85	270.22	275.29	277.15	276.24	270.17	261.48	251.20
(C5-C4-C3)$_{+1}$	111.08	111.62	112.25	112.40	112.24	111.82	111.50	111.13	110.97	110.99	111.08
(C5-C4-O4)$_{+1}$	111.10	111.61	112.24	112.41	112.24	111.82	111.48	111.12	110.99	110.98	111.06
(C3-C5-C4-O4)$_{+1}$	242.05	242.71	243.55	243.69	243.56	242.97	242.60	242.06	241.92	241.84	242.07
(C4-C3-C2)$_{+1}$	101.42	101.30	101.81	102.77	104.01	104.98	105.61	105.30	104.40	103.20	101.42
(C5-C4-C3-C2)$_{+1}$	270.10	274.84	276.38	274.01	268.36	260.28	250.13	239.77	226.65	215.90	208.15
(C4-O4-C1)$_{+1}$	110.02	109.67	109.31	108.28	107.50	106.99	107.50	107.93	109.31	109.67	110.02
(C5-C5-O4-C1)$_{+1}$	108.84	98.54	89.82	84.59	82.82	84.68	89.77	96.78	108.85	121.39	132.91

Table 4.2: Furanose bond and dihedral angles vs. pseudorotation phase $P \geq 180$

In practice, this inversion cannot be accomplished by a smooth, analytic function. It instead becomes necessary to approximate it by linear interpolation. In a 1982 manuscript,

J. Am. Chem. Soc. **104** (1): 278-286 (1982), Olson numerically determined a set of coordinate inversions from phase P to specific atomic structures, for a set of twenty values of phase from 0 to 360 degrees, in 18 degree increments. In this work, all relevant bond angles θ and dihedral angles ϕ were measured for the Olson-determined structures at each of the twenty phase values P_i. Then, for any quantity q, the inversion is well-approximated by the interpolation function

$$q(P) = q(P_i) + \frac{P - P_i}{P_{i+1} - P_i}(q(P_{i+1}) - q(P_i)) \qquad \text{for } P_i \leq P \leq P_{i+1} \tag{4.1}$$

Tables V and VI display the necessary table of bond angles and dihedral angles versus P_i phase values. Any quantity q with parentheses $(q)_{-1}$ refers to a variable located on the 5′-end deoxyribose sugar, any with parentheses $(q)_{+1}$ refers to a variable located on the 3′-end deoxyribose sugar, and any that is non-parenthesized refers to a variable located on the central sugar. Quantities q are labelled such that a collection of three atoms indicates a bond angle between those three atoms and a collection of four atoms indicates a dihedral angle between those four atoms.

4.6.2 Raw Energy Data

Below are the raw output data of the energy landscape calculations for each of the principal modes, with energy, in units of Ry, tabulated against principal mode amplitude, in standard deviations from the mean. These data are converted into kcal/mol, and the plots can then be standardized so that the zero of energy is located at the point of zero mode amplitude, allowing a least-squares extraction of an approximately linear mechanochemical stress σ_λ arising from the salt bridge, a procedure illustrated in Figure 10 of the main manuscript.

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
A/1	0	-812.1596	-812.1638	-812.2120	-812.2019	-733.1414
	0.1	-812.1645	-812.1683	-812.2169	-812.2068	-733.1464
	0.2	-812.1687	-812.1730	-812.2213	-812.2110	-733.1510
	0.3	-812.1698	-812.1745	-812.2228	-812.2127	-733.1527
	0.4	-812.1698	-812.1747	-812.2226	-812.2128	-733.1529
	0.5	-812.1684	-812.1732	-812.2210	-812.2113	-733.1516
	0.6	-812.1656	-812.1705	-812.2182	-812.2086	-733.1490
	0.7	-812.1614	-812.1666	-812.2141	-812.2048	-733.1452
A/2	-1.8	-812.1667	-812.1798	-812.2227	-812.2170	-733.1596
	-1.6	-812.1713	-812.1818	-812.2251	-812.2193	-733.1616
	-1.4	-812.1734	-812.1825	-812.2260	-812.2200	-733.1622
	-1.2	-812.1751	-812.1828	-812.2272	-812.2206	-733.1625
	-1	-812.1755	-812.1824	-812.2268	-812.2201	-733.1617
	-0.8	-812.1744	-812.1806	-812.2258	-812.2185	-733.1597
	-0.6	-812.1720	-812.1776	-812.2236	-812.2157	-733.1565
	0	-812.1596	-812.1638	-812.2120	-812.2019	-733.1414
A/3	0	-812.1596	-812.1638	-812.2120	-812.2019	-733.1414
	0.1	-812.1653	-812.1693	-812.2178	-812.2076	-733.1468
	0.2	-812.1697	-812.1736	-812.2221	-812.2117	-733.1507
	0.3	-812.1715	-812.1755	-812.2241	-812.2138	-733.1523
	0.4	-812.1731	-812.1768	-812.2255	-812.2152	-733.1533
	0.5	-812.1720	-812.1758	-812.2243	-812.2141	-733.1517
	0.6	-812.1695	-812.1732	-812.2219	-812.2117	-733.1489
	0.7	-812.1653	-812.1692	-812.2178	-812.2077	-733.1445
	0.8	-812.1590	-812.1631	-812.2116	-812.2017	-733.1381
A/4	0	-812.1596	-812.1638	-812.2120	-812.2019	-733.1414
	0.6	-812.1821	-812.1851	-812.2357	-812.2235	-733.1640
	0.7	-812.1839	-812.1868	-812.2376	-812.2254	-733.1658
	0.8	-812.1848	-812.1874	-812.2385	-812.2260	-733.1667
	0.9	-812.1851	-812.1877	-812.2390	-812.2263	-733.1671
	1	-812.1849	-812.1874	-812.2387	-812.2260	-733.1669
	1.1	-812.1835	-812.1859	-812.2373	-812.2243	-733.1655
	1.2	-812.1802	-812.1823	-812.2339	-812.2208	-733.1621

Table 4.3: Results of energy landscape calculations for Adenine.

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
G/1	-0.3	-844.4221	-844.4191	-844.4748	-844.4564	-765.4077
	-0.15	-844.4312	-844.4288	-844.4832	-844.4659	-765.4169
	0	-844.4365	-844.4353	-844.4886	-844.4721	-765.4227
	0.15	-844.4381	-844.4370	-844.4896	-844.4740	-765.4240
	0.3	-844.4362	-844.4364	-844.4877	-844.4729	-765.4227

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
	0.45	-844.4331	-844.4338	-844.4845	-844.4703	-765.4196
	0.6	-844.4273	-844.4287	-844.4784	-844.4648	-765.4140
G/2	-0.4	-844.4287	-844.4265	-844.4814	-844.4632	-765.4160
	-0.2	-844.4337	-844.4319	-844.4862	-844.4689	-765.4205
	0	-844.4365	-844.4353	-844.4886	-844.4721	-765.4227
	0.2	-844.4382	-844.4372	-844.4899	-844.4739	-765.4235
	0.4	-844.4385	-844.4384	-844.4901	-844.4748	-765.4233
	0.6	-844.4377	-844.4383	-844.4892	-844.4746	-765.4219
	0.8	-844.4355	-844.4372	-844.4870	-844.4731	-765.4193
	1	-844.4319	-844.4351	-844.4835	-844.4700	-765.4154
G/3	0	-844.4365	-844.4353	-844.4886	-844.4721	-765.4227
	0.2	-844.4410	-844.4393	-844.4929	-844.4760	-765.4270
	0.4	-844.4426	-844.4408	-844.4942	-844.4774	-765.4286
	0.6	-844.4425	-844.4409	-844.4942	-844.4774	-765.4288
	0.8	-844.4403	-844.4384	-844.4917	-844.4749	-765.4265
	1	-844.4361	-844.4345	-844.4875	-844.4711	-765.4228
G/4	-0.25	-844.4236	-844.4222	-844.4752	-844.4590	-765.4104
	-0.1	-844.4325	-844.4312	-844.4844	-844.4679	-765.4189
	0	-844.4365	-844.4353	-844.4886	-844.4721	-765.4227
	0.05	-844.4379	-844.4365	-844.4900	-844.4736	-765.4239
	0.2	-844.4402	-844.4387	-844.4926	-844.4758	-765.4257
	0.35	-844.4390	-844.4377	-844.4918	-844.4747	-765.4242
	0.5	-844.4341	-844.4331	-844.4873	-844.4704	-765.4190
	0.65	-844.4266	-844.4256	-844.4802	-844.4630	-765.4111

Table 4.4: Results of energy landscape calculations for Guanine.

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
C/1	-0.45	-792.8073	-792.8125	-792.8575	-792.8500	-713.7860
	-0.3	-792.8184	-792.8241	-792.8688	-792.8618	-713.7971
	-0.15	-792.8259	-792.8327	-792.8768	-792.8704	-713.8050
	0	-792.8290	-792.8365	-792.8801	-792.8742	-713.8082
	0.15	-792.8299	-792.8377	-792.8809	-792.8757	-713.8091
	0.3	-792.8267	-792.8356	-792.8784	-792.8738	-713.8063
	0.45	-792.8199	-792.8296	-792.8719	-792.8676	-713.7997
	0.6	-792.8115	-792.8220	-792.8636	-792.8598	-713.7913
C/2	0	-792.8290	-792.8365	-792.8801	-792.8742	-713.8082

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
	0.15	-792.8344	-792.8414	-792.8850	-792.8787	-713.8137
	0.3	-792.8376	-792.8443	-792.8879	-792.8811	-713.8167
	0.45	-792.8395	-792.8458	-792.8890	-792.8821	-713.8185
	0.6	-792.8397	-792.8456	-792.8888	-792.8815	-713.8186
	0.75	-792.8385	-792.8444	-792.8869	-792.8792	-713.8172
	0.9	-792.8350	-792.8407	-792.8824	-792.8745	-713.8134
	1.05	-792.8298	-792.8355	-792.8765	-792.8684	-713.8079
C/3	-0.45	-792.8203	-792.8272	-792.8729	-792.8656	-713.7997
	-0.3	-792.8239	-792.8308	-792.8758	-792.8691	-713.8032
	-0.15	-792.8272	-792.8345	-792.8789	-792.8725	-713.8065
	0	-792.8290	-792.8365	-792.8801	-792.8742	-713.8082
	0.15	-792.8295	-792.8373	-792.8800	-792.8751	-713.8089
	0.3	-792.8301	-792.8378	-792.8801	-792.8758	-713.8094
	0.45	-792.8283	-792.8365	-792.8775	-792.8740	-713.8076
	0.6	-792.8257	-792.8341	-792.8743	-792.8716	-713.8051
C/4	-1.05	-792.7970	-792.8023	-792.8421	-792.8381	-713.7825
	-0.9	-792.8121	-792.8176	-792.8581	-792.8538	-713.7970
	-0.75	-792.8234	-792.8291	-792.8704	-792.8659	-713.8075
	-0.6	-792.8315	-792.8375	-792.8792	-792.8744	-713.8147
	-0.45	-792.8360	-792.8423	-792.8848	-792.8798	-713.8185
	-0.3	-792.8372	-792.8436	-792.8868	-792.8815	-713.8187
	-0.15	-792.8349	-792.8417	-792.8852	-792.8797	-713.8153
	0	-792.8290	-792.8365	-792.8801	-792.8742	-713.8082

Table 4.5: Results of energy landscape calculations for Cytosine.

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
T/1	-0.05	-817.7471	-817.7531	-817.8021	-817.7901	-738.7336
	0	-817.7499	-817.7560	-817.8049	-817.7930	-738.7364
	0.1	-817.7542	-817.7607	-817.8092	-817.7979	-738.7409
	0.25	-817.7588	-817.7656	-817.8132	-817.8026	-738.7454
	0.4	-817.7583	-817.7660	-817.8131	-817.8032	-738.7456
	0.55	-817.7556	-817.7636	-817.8100	-817.8007	-738.7426
	0.7	-817.7509	-817.7599	-817.8054	-817.7966	-738.7382
	0.85	-817.7438	-817.7532	-817.7980	-817.7897	-738.7310
T/2	0	-817.7499	-817.7560	-817.8049	-817.7930	-738.7364
	0.4	-817.7675	-817.7739	-817.8212	-817.8096	-738.7540
	0.6	-817.7734	-817.7800	-817.8265	-817.8150	-738.7598
	0.8	-817.7776	-817.7845	-817.8300	-817.8185	-738.7639
	1	-817.7793	-817.7866	-817.8311	-817.8195	-738.7657
	1.2	-817.7798	-817.7878	-817.8309	-817.8194	-738.7661

Base/PCA	St Devs	5′-Mono (Ry)	3′-Mono (Ry)	5′-Bi (Ry)	3′-Bi (Ry)	No Arginine (Ry)
	1.4	-817.7781	-817.7867	-817.8286	-817.8171	-738.7643
	1.6	-817.7749	-817.7841	-817.8245	-817.8130	-738.7608
T/3	-2	-817.7511	-817.7556	-817.8015	-817.7904	-738.7447
	-1.8	-817.7651	-817.7694	-817.8158	-817.8047	-738.7580
	-1.6	-817.7754	-817.7794	-817.8266	-817.8151	-738.7678
	-1.4	-817.7831	-817.7874	-817.8347	-817.8231	-738.7748
	-1.2	-817.7879	-817.7925	-817.8401	-817.8285	-738.7791
	-1	-817.7895	-817.7939	-817.8420	-817.8305	-738.7799
	-0.8	-817.7879	-817.7929	-817.8410	-817.8293	-738.7777
	-0.6	-817.7830	-817.7879	-817.8364	-817.8248	-738.7719
	-0.4	-817.7751	-817.7801	-817.8288	-817.8171	-738.7632
	-0.2	-817.7643	-817.7698	-817.8187	-817.8068	-738.7515
	0	-817.7499	-817.7560	-817.8049	-817.7930	-738.7364
T/4	-1.6	-817.7654	-817.7712	-817.8125	-817.8062	-738.7589
	-1.4	-817.7725	-817.7778	-817.8202	-817.8135	-738.7651
	-1.2	-817.7767	-817.7822	-817.8253	-817.8181	-738.7687
	-1	-817.7786	-817.7839	-817.8282	-817.8203	-738.7698
	-0.8	-817.7783	-817.7834	-817.8286	-817.8203	-738.7686
	-0.6	-817.7752	-817.7807	-817.8269	-817.8174	-738.7647
	-0.4	-817.7692	-817.7746	-817.8217	-817.8117	-738.7577
	0	-817.7499	-817.7560	-817.8049	-817.7930	-738.7364

Table 4.6: Results of energy landscape calculations for Thymine.

4.6.3 Results of Principal Component Analysis

Presented here are the raw outputs of the principal component analyses on the model complex, for each of the different central nucleobases. For an annotated key explaining each of the different variables, see Figure 3 of the main manuscript. Displayed are the mean values of the parameters, and for each principal component, the contribution of each parameter to one positive unit of that component, as measured in standard deviations from the mean. Also displayed is the score for each component, or the fraction of the total variance that it captures. These results were generated from 1495 structural examples of adenine, 1183 of guanine, 1175 of cytosine, and 1490 of thymine. For a full list of the protein-DNA crystal complexes used to generate the data, see the Supplementary Information of

the manuscript *J. Phys. Chem. B* **117(51)**, 16436-16442 (2013).

Base	A	G	C	T
P_{-1}	140.75	144.43	147.22	140.29
ϵ_{-1}	198.4	198.42	193.94	189.36
ζ_{-1}	-119.18	-115.15	-107.25	-101.43
α_{-1}	-50.08	-52.51	-44.72	-50.64
β_{-1}	170.47	170.45	173.01	177.65
γ_{-1}	46.38	47.21	42.75	40.4
P	144.73	149.01	139.15	138.5
χ	-107.98	-107.97	-114.48	-112.58
ϵ_{+1}	192.7	193.43	197.82	196.41
ζ_{+1}	-105.48	-114.65	-112.98	-112.61
α_{+1}	-46.83	-48.65	-52.13	-48.81
β_{+1}	173.68	174.57	171.55	171.87
γ_{+1}	41.64	42.56	46.31	45.38
P_{+1}	137.47	144.65	146.07	144.52

Table 4.7: Average conformational parameter values of the sugar-phosphate backbone. Displayed are parameter mean values and standard deviations for each of the four different central nucleobases. Units of angular parameters are degrees.

Base	A	G	C	T
P_{-1}	17.13	15.47	12.42	17.12
ϵ_{-1}	0	0.83	0.86	-1.06
ζ_{-1}	-7.18	-6.34	-5.04	-5.65
α_{-1}	4.9	6.1	8.11	12.46
β_{-1}	0.34	0.76	3.3	2.98
γ_{-1}	-7.67	-11.2	-18.85	-14.45
P	19.55	17.27	18.45	17.94
χ	5.02	8.64	10	6.25
ϵ_{+1}	-0.38	2.12	3.16	0.55
ζ_{+1}	-8.18	-10.52	-13.5	-8.69
α_{+1}	13.9	10.46	9.24	6.72
β_{+1}	4.89	-1.04	-3.28	-0.88
γ_{+1}	-14.96	-13.17	-8.43	-5.46
P_{+1}	17.84	17.04	12.52	13.24
Score	0.22	0.21	0.23	0.25

Table 4.8: Conformational contributions to the first principal component of deformation of the sugar-phosphate backbone. Displayed are magnitudes of one unit of a principal component, as measured in standard deviations from the mean, for each of the four different central nucleobases.

Base	A	G	C	T
P_{-1}	-5.04	-3.29	-4.24	1.87
ϵ_{-1}	-14.18	-5.63	-5.12	-0.43
ζ_{-1}	22	15.5	12.02	1.81
α_{-1}	-3.69	-9.6	-21.06	-13.96
β_{-1}	10.88	1.77	-7.88	-4.35
γ_{-1}	-1.06	2.2	16.91	14.12
P	-1.83	-2.85	-3.5	-0.76
χ	0.9	5.98	1.84	3.07
ϵ_{+1}	3.27	11.79	11.17	11.87
ζ_{+1}	-6.38	-16.72	-16.17	-16.35
α_{+1}	8.72	-0.67	5.17	0.83
β_{+1}	2.76	-9.98	-7.43	-8.77
γ_{+1}	-4.92	8.73	-0.88	1.77
P_{+1}	0.47	-5.31	1.31	-0.05
Score	0.15	0.15	0.15	0.15

Table 4.9: Conformational contributions to the second principal component of deformation of the sugar-phosphate backbone. Displayed are magnitudes of one unit of a principal component, as measured in standard deviations from the mean, for each of the four different central nucleobases.

Base	A	G	C	T
P_{-1}	2.7	1.37	-0.55	-6.69
ϵ_{-1}	-0.85	14.31	13.39	10.74
ζ_{-1}	3	-18.47	-17.12	-6.53
α_{-1}	-3.75	0.12	8.13	14.3
β_{-1}	-1.28	-11.51	-4.99	-1.21
γ_{-1}	-10.3	3.22	8.9	-14.28
P	8.72	-1.24	-4.68	-5.25
χ	7.3	0.86	-0.95	0.66
ϵ_{+1}	3.3	5.61	8.48	3.54
ζ_{+1}	-7.25	-5.32	-5.42	-0.42
α_{+1}	-12.02	-7.16	-6.92	-9.49
β_{+1}	-16.36	-6.55	-8.18	-8.28
γ_{+1}	25.31	11.63	6.95	6.65
P_{+1}	-3.73	-4.66	-7.22	-10.99
Score	0.12	0.13	0.12	0.11

Table 4.10: Conformational contributions to the third principal component of deformation of the sugar-phosphate backbone. Displayed are magnitudes of one unit of a principal component, as measured in standard deviations from the mean, for each of the four different central nucleobases.

4.6.4 Salt-Bridge Clustering

The clustering of arginine-phosphate salt bridges is performed on a set of 83 nucleosomal crystal structures taken from the Nucleic Acid Database. From these structures, 1556 structural examples of arginine-phosphate contacts are generated, and within that set, a reduced group of 790 examples with arginine-phosphate salt-bridge contact are identified, according to the criteria described in the main manuscript. The reduced set of 790 structural

Base	A	G	C	T
P_{-1}	8.17	0.53	-6.11	-5.72
ϵ_{-1}	-6.83	-0.39	-5.82	-5.44
ζ_{-1}	4.01	1.57	7.31	10.7
α_{-1}	6.44	13.37	11.87	0.06
β_{-1}	6	7.26	14.47	13.67
γ_{-1}	-8.3	-16.89	-18.13	-11.56
P	8.91	6.23	-1.18	-0.72
χ	-0.62	5.12	3.76	4.1
ϵ_{+1}	-12.39	-5.9	2.34	-0.95
ζ_{+1}	13.53	4.48	-4.73	-2.41
α_{+1}	-14.34	-21.28	-13.04	-9.74
β_{+1}	-0.8	-7.76	-7.62	-5.34
γ_{+1}	4.79	21.41	17.91	5.45
P_{+1}	4.68	-2.25	-11.57	-7.87
Score	0.12	0.11	0.1	0.1

Table 4.11: Conformational contributions to the fourth principal component of deformation of the sugar-phosphate backbone. Displayed are magnitudes of one unit of a principal component, as measured in standard deviations from the mean, for each of the four different central nucleobases.

examples of arginine-phosphate salt bridges are then divided into two sets by K-means clustering. K-means clustering is an iterative scheme which divides a dataset into two clusters, each with a cluster centroid point, such that the in-cluster sum of total point-point distances is minimized. As this is an iterative approximate scheme, 100 replicates of the trial of performed, and the trial with the lowest overall reported sum is used to divide the dataset into mondentate and bidentate clusters, with one or two hydrogen bonds, respectively. Displayed below is the calculated average cluster values of the six-parameter vector $(\vec{r}, \theta, \phi, \omega) = (x, y, z, \theta, \phi, \omega)$ for monodentate and bidentate forms.

Cluster	x (Angstrom)	y (Angstrom)	z (Angstrom)	θ (°)	ϕ (°)	ω (°)
Monodentate	-2.242	0.829	3.351	47.506	164.839	-4.423
Bidentate	1.447	3.654	-0.121	94.930	62.652	21.717

Table 4.12: Average values of salt-bridge parameters determined by K-means clustering.

Chapter 5

Summary and (Highly Personal) Outlook

This dissertation has presented two theoretical investigations of molecular biophysical processes underlying eukaryotic transcriptional regulation. The results have highlighted the important role of mechanochemistry, or the coupling of chemical change to mechanical deformation, in the functionality of biological matter.

More importantly, however, I hope that this work has highlighted the value that the methods and mindset of soft condensed matter physics can bring to the biological table, and vice versa. By having a deeper understanding of the theoretical design principles that shape the organization of living matter, we can potentially go a long way towards having a more coherent understanding of how evolution shapes biological emergence in general. These insights, in turn, can help condensed matter and materials theorists to more generally understand how atomic structure shapes the emergent effective properties of materials.

I shall momentarily go into more detail concerning a few of my own personal observations of interesting connections between condensed matter physics and biology that I have come to appreciate over the past few years. However, before I do so, I would be remiss not to point out that there is still plenty of work to be done in the development of novel theoretical and computational approaches to quantitatively characterize and predict the properties of complex biological systems. In the course of this work, for example, a significant fraction of the day-to-day labor was spent figuring out how to seamlessly integrate

structural bioinformatics and density functional theory together in a way tailored to the generation of biologically relevant information. There is still much room for improvement in the development of methods and computer algorithms that intelligently and efficiently mix together statistical inference, molecular simulation, and analytical toy modeling. This is especially true, in my opinion, for *ab initio* density functional theory calculations, which are a relative latecomer to the arena of theoretical biophysics. It remains to be seen how they can complement more established, large-scale approaches, such as molecular dynamics and Monte Carlo simulations [20].

5.1 Reflections on Methylation: the Role of Noise in Switching Kinetics

In the first project, an analysis of the 5-methylation of CG:CG base-pair steps, the presence of methyl groups at the C5 positions of cytosines was observed to roughen the smooth elastic energy landscape of canonical B-DNA. This roughening arises because the methyl groups are, to a first approximation, covalently fixed, while also forced to interact with the other atoms in the base-pair step via pi stacking. As these stacking interactions cannot be minimized without breaking the covalent bond, the system is geometrically frustrated, and unable to minimize all interactions simultaneously. This results in a collection of secondary minima generated by an effectively stochastic mechanical stress. This stochasticity, in turn, allows the otherwise rigid B-DNA to fluctuate into the A-like and C-like forms necessary for nucleosome wrapping.

The role of stochasticity in triggering switching between different thermodynamic states is not a new idea. One of its most well-known examples is in the switching of ferromagnetic spins in magnetic recording devices [6]. Specifically, magnetic spins change direction via

the nucleation of domain walls, or regions of space separating up and down spins. Upon nucleation, the domain walls dynamically propagate, coalesce, and become pinned into stable configurations, until the direction of magnetization in the recording material is reversed. This suggests that, in the context of the role of methylation in chromatin remodeling, there is much potential new insight to be gained by looking at the problem from the point of view of first-order phase transition kinetics, a field in which ideas like hysteretic switching and defect nucleation have historically provided a veritable mine of useful information [10].

5.2 Reflections on Salt Bridges: Complementarity and Self Assembly

The second project studied salt bridge contacts between histone arginines and the DNA sugar-phosphate backbone. Specifically, the mechanochemical effect of the histone-DNA binding was calculated as a function of several different parameters, including sequence, contact positioning, and the number of hydrogen bonds formed. What was observed was that in all cases, the effects of the contact on DNA elastic energetics were well-described by a linear mechanochemical stress. Furthermore, all of the different parameters were found to induce sizable mechanical deformations, relative to thermal fluctuations, for at least some of the principal components.

These findings take on a newfound significance when interpreted from the perspective of the relatively young areas of condensed matter science dealing with the molecular self-assembly and design of complex, adaptive materials [49, 102]. A linear mechanochemical stress, in contrast to the stochastic signals seen in methylation, is 'smooth', and characteristic of systems without significant frustration. Put another way, the molecular building blocks 'fit' together like pieces of a puzzle, a phenomenon that has been described in biological communities by terms like lock-and-key complementarity [100].

What is so remarkable about biological materials like histone-DNA contacts, from a materials science perspective, is the sheer number of different locks and keys that significantly influence molecular shape, and the way in which living systems are able to use them in such a precisely controlled manner to do specific organizational tasks. The extreme sensitivity of the materials to so many different parameters has traditionally been associated with chaotic systems with sensitive dependence to initial conditions. Yet evolution has been able to carefully select for parameters and control mechanisms such that the complex sensitivity of the building blocks remains intact, while still behaving in an 'organized' fashion. To use the words of the biologist Stuart Kauffmann, living systems appear to be at the 'edge of chaos' [39]. Clearly, there is a great deal more to be learned about the non-equilibrium statistical mechanics of dynamic soft materials by studying the ultimate example of emergence in living matter.

Bibliography

[1] C. Altona and M. Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation. *J. Am. Chem. Soc.*, 94(23):8205–8212, 1972.

[2] S. Arnott. Polynucleotide secondary structures: a historical perspective, 1999.

[3] Peter W. Atkins and Ronald S. Friedman. *Molecular Quantum Mechanics*. Oxford University Press, 2011.

[4] Axel D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38(6):3098, 1988.

[5] Helen M. Berman, Wilma K. Olson, David L. Beveridge, John Westbrook, Anke Gelbin, Tamas Demeny, Shu-Hsin Hsieh, A.R. Srinivasan, and Bohdan Schneider. The nucleic acid database. a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, 63(3):751–759, 1992.

[6] H Neal Bertram. *Theory of Magnetic Recording*. Cambridge University Press, 1994.

[7] Adrian P. Bird and Alan P. Wolffe. Methylation-induced repression-belts, braces, and chromatin. *Cell*, 99(5):451–454, 1999.

[8] Carlos Bustamante, Zev Bryant, and Steven B. Smith. Ten years of tension: single-molecule dna mechanics. *Nature*, 421(6921):423–427, 2003.

[9] David M. Ceperley and B.J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45(7):566–569, 1980.

[10] Paul M Chaikin and Tom C. Lubensky. *Principles of Condensed Matter Physics*. Cambridge University Press, 2000.

[11] Taiping Chen and Sharon Y.R. Dent. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Gen.*, 15(2):93–106, 2014.

[12] Neil R. Clark and Avi Maayan. Introduction to statistical methods to analyze large data sets: Principal components analysis. *Sci. Signal.*, 4(190):tr3, 2011.

[13] Lester Clowney, Shri C. Jain, A.R. Srinivasan, John Westbrook, Wilma K. Olson, and Helen M. Berman. Geometric parameters in nucleic acids: Nitrogenous bases. *J. Am. Chem. Soc.*, 118(3):509–518, 1996.

[14] Valentino R Cooper, Timo Thonhauser, Aaron Puzder, Elsebeth Schroder, Bengt I. Lundqvist, and David C. Langreth. Stacking interactions and the twist of dna. *J. Am. Chem. Soc.*, 130(4):1304–1308, 2008.

[15] Colin Davey, Sari Pennings, and James Allan. Cpg methylation remodels chromatin structure in vitro. *J. Mol. Biol.*, 267(2):276–288, 1997.

[16] Colin S. Davey, Sari Pennings, Carmel Reilly, Richard R. Meehan, and James Allan. A determining influence for cpg dinucleotides on nucleosome positioning in vitro. *Nuc. Acids Res.*, 32(14):4322–4331, 2004.

[17] Curt A. Davey, David F. Sargent, Karolin Luger, Armin W. Maeder, and Timothy J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 angstrom resolution. *J. Mol. Biol.*, 319(5):1097–1113, 2002.

[18] Max Dion, Henrik Rydberg, Elsebeth Schroder, David C. Langreth, and Bengt I. Lundqvist. Van der waals density functional for general geometries. *Phys. Rev. Lett.*, 92(24):246401, 2004.

[19] Jason E. Donald, Daniel W. Kulp, and William F. DeGrado. Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function, and Bioinformatics*, 79(3):898–915, 2011.

[20] Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Ann. Rev. Biophys.*, 41:429–452, 2012.

[21] A. Flaus, C. Rencurel, H. Ferreira, N. Wiechens, and T. Owen-Hughes. Sin mutations alter inherent nucleosome mobility. *EMBO J.*, 23:343–353, 2004.

[22] Paul Flory, M. Volkenstein, et al. Statistical mechanics of chain molecules, 1969.

[23] Gareth Forde, Leonid Gorb, Oleg Shiskin, Aviane Flood, Curinetha Hubbard, Glake Hill, and Jerzy Leszczynski. Molecular structure and properties of protonated and methylated derivatives of cytosine. *J. Biomol. Struct. Dyn.*, 20(6):819–828, 2003.

[24] Rosalind E. Franklin and Raymond G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, 1953.

[25] D. Frigyes, F. Alber, S. Pongor, and P. Carloni. Arginine–phosphate salt bridges in protein–dna complexes: a car-parrinello study. *Journal of Molecular Structure: THEOCHEM*, 574(1):39–45, 2001.

[26] Anke Gelbin, Bohdan Schneider, Lester Clowney, Shu-Hsin Hsieh, Wilma K. Olson, and Helen M. Berman. Geometric parameters in nucleic acids: Sugar and phosphate constituents. *J. Am. Chem. Soc.*, 118(3):519–529, 1996.

[27] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L. Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Cond. Mat.*, 21(39):395502, 2009.

[28] David S at the RCSB Goodsell. *Molecule of the Month: Nucleosome*, July 2000 (accessed May 18, 2014). http://www.rcsb.org/pdb/101/motm.do?momID=7.

[29] Andrey A. Gorin, Victor B. Zhurkin, and K Wilma. B-dna twisting correlates with base-pair morphology. *J. Mol. Biol.*, 247(1):34–48, 1995.

[30] Stefan Grimme. Semiempirical gga-type density functional constructed with a long-range dispersion correction. *J. Comp. Chem.*, 27(15):1787–1799, 2006.

[31] Jurgen Hafner, Christopher Wolverton, and Gerbrand Ceder. Toward computational materials design: the impact of density functional theory on materials research. *MRS Bulletin*, 31(09):659–668, 2006.

[32] B. Hammer, Lars Bruno Hansen, and Jens Kehlet Norskov. Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals. *Phys. Rev. B*, 59(11):7413, 1999.

[33] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864, 1964.

[34] Jochen S. Hub and Bert L. de Groot. Detection of functional modes in protein dynamics. *PLoS Comp. Biol.*, 5(8):e1000480, 2009.

[35] National Human Genome Research Institute. Chromatin. "http://www.genome.gov/glossary/index.cfm?id=32". Accessed: 2014-05-18.

[36] National Human Genome Research Institute. Nucleotide. "http://www.genome.gov/glossary/index.cfm?id=143". Accessed: 2014-05-18.

[37] National Human Genome Research Institute. Translation. "http://www.genome.gov/glossary/index.cfm?id=200". Accessed: 2014-05-18.

[38] Cizhong Jiang and B. Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Gen.*, 10(3):161–172, 2009.

[39] Stuart A. Kauffman and Sonke Johnsen. Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.*, 149(4):467–505, 1991.

[40] M.L. Kireeva, B. Hancock, G.H. Cremona, W. Walter, V.M. Studitsky, and M. Kashlev. Nature of the nucleosomal barrier to rna polymerase ii. *Mol. Cell*, 18:97–108, 2005.

[41] M.L. Kireeva, W. Walter, V. Tchernajenko, V. Bondarenko, M. Kashlev, and V.M. Studitsky. Nucleosome remodeling induced by rna polymerase ii: loss of the h2a/h2b dimer during transcription. *Mol. Cell*, 9:541–552, 2002.

[42] University of Illinois at Urbana-Champaign Klaus Schulten group. *DNA Methylation and Hydroxymethylation*, 2014 (accessed May 18, 2014). http://www.ks.uiuc.edu/Research/methylation/.

[43] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133, 1965.

[44] Brian Kolb and Timo Thonhauser. Molecular biology at the quantum level: can modern density functional theory forge the path? *Nano LIFE*, 2(02), 2012.

[45] Olga I. Kulaeva, Guohui Zheng, Yury S. Polikanov, Andrew V. Colasanti, Nicolas Clauvelin, Swagatam Mukhopadhyay, Anirvan M. Sengupta, Vasily M. Studitsky, and Wilma K. Olson. Internucleosomal interactions mediated by histone tails allow distant communication in chromatin. *J. Biol. Chem.*, 287(24):20248–20257, 2012.

[46] H. Kurumizaka and A.P. Wolffe. Sin mutations of histone h3: influence on nucleosome core structure and function. *Mol. Cell Biol.*, 17:6953–6969, 1997.

[47] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785, 1988.

[48] Kyuho Lee, Eamonn D. Murray, Lingzhu Kong, Bengt I. Lundqvist, and David C. Langreth. Higher-accuracy van der waals density functional. *Phys. Rev. B*, 82(8):081101, 2010.

[49] Jean-Marie Lehn. Toward self-organization and complex matter. *Science*, 295(5564):2400–2403, 2002.

[50] G. Li and J. Widom. Nucleosomes facilitate their own invasion. *Nature Struct. Molec. Biol.*, 11:763–769, 2004.

[51] Yun Li. *Understanding DNA-protein interactions from the nucleic-acid perspective.* ProQuest, 2006.

[52] Shu Liu, Yuh J. Chao, and Xiankui Zhu. Tensile-shear transition in mixed mode i/iii fracture. *Int. J. Solids Struct.*, 41(22):6147–6172, 2004.

[53] D. Lohr, K. Tatchell, and K.E. Van Holde. On the occurrence of nucleosome phasing in chromatin. *Cell*, 12(3):829–836, 1977.

[54] Xiang-Jun Lu and Wilma K. Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nuc. Acids Res.*, 31(17):5108–5121, 2003.

[55] Xiang-Jun Lu and Wilma K. Olson. 3dna: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protocols*, 3(7):1213–1227, 2008.

[56] Karolin Luger, Armin W. Mader, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature*, 389(6648):251–260, 1997.

[57] G. Makov and M.C. Payne. Periodic boundary conditions in ab initio calculations. *Phys. Rev. B*, 51(7):4014, 1995.

[58] Chhabinath Mandal, Neville R. Kallenbach, and S. Walter Englander. Base-pair opening and closing reactions in the double helix: A stopped-flow hydrogen exchange study in poly (ra) : poly (ru). *J. Mol. Biol.*, 135(2):391–411, 1979.

[59] Amir Marcovitz and Yaakov Levy. Frustration in protein-dna binding influences conformational switching and target search kinetics. *Proc. Nat. Acad. Sci.*, 108(44):17957–17962, 2011.

[60] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.

[61] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[62] Andrei D. Mirzabekov and Alexander Rich. Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal dna and its role in dna bending. *Proc. Nat. Acad. Sci.*, 76(3):1118–1121, 1979.

[63] Arnost Mladek, Miroslav Krepl, Daniel Svozil, Petr Cech, Michal Otyepka, Pavel Banas, Marie Zgarbova, Petr Jurecka, and Jiri Sponer. Benchmark quantum-chemical calculations on a complete set of rotameric families of the dna sugar–phosphate backbone and their comparison with modern density functional theory. *Physical Chemistry Chemical Physics*, 15(19):7295–7310, 2013.

[64] Bruce C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transaction on Automatic Control*, 26(1):17–32, 1981.

[65] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, 1995.

[66] Hillary C.M. Nelson, John T. Finch, Bonaventura F. Luisi, and Aaron Klug. The structure of an oligo (da) : oligo (dt) tract and its biological implications. *Nature*, 330(6145):221–226, 1987.

[67] Wilma K. Olson. How flexible is the furanose ring? 2. an updated potential energy estimate. *J. Am. Chem. Soc.*, 104(1):278–286, 1982.

[68] Wilma K. Olson. Theoretical studies of nucleic acid conformation: Potential energies, chain statistics, and model building, 1982.

[69] Wilma K. Olson, Manju Bansal, Stephen K. Burley, Richard E. Dickerson, Mark Gerstein, Stephen C. Harvey, Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked, et al. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, 313(1):229–237, 2001.

[70] Wilma K. Olson, Nicolas Clauvelin, Andrew V. Colasanti, Gautam Singh, and Guohui Zheng. Insights into gene expression and packaging from computer simulations. *Biophys. Rev.*, 4(3):171–178, 2012.

[71] Wilma K. Olson, Andrey A. Gorin, Xiang-Jun Lu, Lynette M. Hock, and Victor B. Zhurkin. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proc. Nat. Acad. Sci.*, 95(19):11163–11168, 1998.

[72] Wilma K. Olson and Joel L. Sussman. How flexible is the furanose ring? 1. a comparison of experimental and theoretical studies. *J. Am. Chem. Soc.*, 104(1):270–278, 1982.

[73] Wilma King Olson. Configurational statistics of polynucleotide chains. a single virtual bond treatment. *Macromolecules*, 8(3):272–275, 1975.

[74] W.K. Olson, A.R. Srinivasan, A.V. Colasanti, G. Zheng, and D. Swigon. Dna biomechanics, 2009.

[75] Noel M. OBoyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: An open chemical toolbox. *J. Cheminf.*, 3(1):1–14, 2011.

[76] Trent M. Parker, Edward G. Hohenstein, Robert M. Parrish, Nicholas V. Hud, and C. David Sherrill. Quantum-mechanical analysis of the energetic contributions to π stacking in nucleic acids versus rise, twist, and slide. *J. Am. Chem. Soc.*, 135(4):1306–1316, 2013.

[77] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865, 1996.

[78] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, 45(23):13244, 1992.

[79] John P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23(10):5048, 1981.

[80] K.J. Polach and J. Widom. Mechanism of protein access to specific dna sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.*, 254:130–149, 1995.

[81] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[82] Guillermo Roman-Perez and Jose M. Soler. Efficient implementation of a van der waals density functional: Application to double-wall carbon nanotubes. *Phys. Rev. Lett.*, 103(9):096102, 2009.

[83] Wolfram Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, 1984.

[84] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

[85] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature Struct. Molec. Biol.*, 442:772–778, 2006.

[86] Ky Sha and Laurie A Boyer. *The chromatin signature of pluripotent cells*. Harvard Stem Cell Institute, 2009.

[87] Jiri Sponer, Arnost Mladek, Judit E. Sponer, Daniel Svozil, Marie Zgarbova, Pavel Banas, Petr Jurecka, and Michal Otyepka. The dna and rna sugar–phosphate backbone emerges as the key player. an overview of quantum-chemical, structural biology and simulation studies. *Phys. Chem. Chem. Phys.*, 14(44):15257–15277, 2012.

[88] Jiri Sponer, Kevin E. Riley, and Pavel Hobza. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.*, 10(19):2595–2610, 2008.

[89] Jiri Sponer, Judit E. Sponer, Arnost Mladek, Petr Jurecka, Pavel Banas, and Michal Otyepka. Nature and magnitude of aromatic base stacking in dna and rna: Quantum chemistry, molecular mechanics, and experiment. *Biopolymers*, 99(12):978–988, 2013.

[90] K. Struhl and E. Segal. Determinants of nucleosome positioning. *Nature Struct. and Molec. Biol.*, 20:267–273, 2013.

[91] Timo Thonhauser, Valentino R. Cooper, Shen Li, Aaron Puzder, Per Hyldgaard, and David C. Langreth. Van der waals density functional: Self-consistent potential and the nature of the van der waals bond. *Phys. Rev. B*, 76(12):125112, 2007.

[92] H.S. Tims, K. Gurunathan, M. Levitus, and J. Widom. Dynamics of nucleosome invasion by dna binding proteins. *J. Mol. Biol.*, 411:430–448, 2011.

[93] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102(7):073005, 2009.

[94] Michael Y. Tolstorukov, Andrew V. Colasanti, David M. McCandlish, Wilma K. Olson, and Victor B. Zhurkin. A novel roll-and-slide mechanism of dna folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, 371(3):725–738, 2007.

[95] V. Tozzini, A.R. Bizzarri, V. Pellegrini, R. Nifosi, P. Giannozzi, A. Iuliano, S. Cannistraro, and F. Beltram. The low frequency vibrational modes of green fluorescent proteins. *Chemical Physics*, 287:33–42, 2003.

[96] Norman Troullier and Jose Luriaas Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43(3):1993, 1991.

[97] Bryan M. Turner. Defining an epigenetic code. *Nat. Cell Biol.*, 9(1):2–6, 2007.

[98] Lorens van Dam and Malcolm H. Levitt. Bii nucleotides in the b and c forms of natural-sequence polymeric dna: A new model for the c form of dna. *J. Mol. Biol.*, 304(4):541–561, 2000.

[99] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses.* Cambridge University Press, 2003.

[100] James D. Watson. *Molecular Biology of the Gene.* Pearson Education India, 2004.

[101] James D. Watson and Francis H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[102] George M. Whitesides and Bartosz Grzybowski. Self-assembly at all scales. *Science*, 295(5564):2418–2421, 2002.

[103] Maurice H.F. Wilkins, Alex R. Stokes, and Herbert R. Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–740, 1953.

[104] Loren Dean Williams and L. James Maher III. Electrostatic mechanisms of dna deformation. *Annu. Rev. Biophys. Biomol. Struct.*, 29(1):497–521, 2000.

[105] F. Xu, A.V. Colasanti, Y. Li, and W.K. Olson. Long-range effects of histone point mutations on dna remodeling revealed from computational analyses of sin-mutant nucleosome structures. *Nucleic Acids Research*, 38:6872–6882, 2010.

[106] Jin Yang, Lee Lior-Hoffmann, Shenglong Wang, Yingkai Zhang, and Suse Broyde. Dna ctosine methylation: Structural and thermodynamic characterization of the epigenetic marking mechanism. *Biochemistry*, 52(16):2828–2838, 2013.

[107] Lei Yang, Guang Song, Alicia Carriquiry, and Robert L. Jernigan. Close correspondence between the motions from principal component analysis of multiple hiv-1 protease structures and elastic network modes. *Structure*, 16(2):321–330, 2008.

[108] Darrin M. York, Tai-Sung Lee, and Weitao Yang. Quantum mechanical treatment of biological macromolecules in solution using linear-scaling electronic structure methods. *Phys. Rev. Lett.*, 80:5011–5014, 1998.

[109] Ben Youngblood, Fa-Kuen Shieh, Fabian Buller, Tim Bullock, and Norbert O. Reich. S-adenosyl-l-methionine-dependent methyl transfer: observable precatalytic intermediates during dna cytosine methylation. *Biochemistry*, 46(30):8766–8775, 2007.

[110] Tahir I. Yusufaly, Yun Li, and Wilma K. Olson. 5-methylation of cytosine in cg:cg base-pair steps: A physicochemical mechanism for the epigenetic control of dna nanomechanics. *J. Phys. Chem. B*, 117(51):16436–16442, 2013.

[111] Tahir I. Yusufaly, Yun Li, Gautam Singh, and Wilma K. Olson. Arginine-phosphate salt bridges between histones and dna: Intermolecular actuators that control nucleosome architecture. *arXiv:1404.4405*, 2014.

[112] Samir Zakhari. Alcohol metabolism and epigenetics changes. *Alcohol Research: Current Reviews*, 35(1):6, 2013.

[113] Guohui Zheng, Xiang-Jun Lu, and Wilma K. Olson. Web 3dnaa web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nuc. Acids Res.*, 37(Suppl. 2):W240–W246, 2009.