INTERACTION OF ECORI WITH NONCOGNATE DNA SEQUENCES: COMPUTATIONAL INVESTIGATION OF DYNAMICS OF PROTEIN & WATER AND DNA CONFORMATION

VIGNESHWAR RAMAKRISHNAN

NATIONAL UNIVERSITY OF SINGAPORE

2011

INTERACTION OF ECORI WITH NONCOGNATE DNA SEQUENCES: COMPUTATIONAL INVESTIGATION OF DYNAMICS OF PROTEIN & WATER AND DNA CONFORMATION

VIGNESHWAR RAMAKRISHNAN

(B. Tech., PSG College of Technology, India)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMICAL AND BIOMOLECULAR ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2011

ACKNOWLEDGEMENTS

"Curiosity keeps leading us down new paths", Walt Disney said. So did this thesis. What started as an investigation on the effect of macromolecular crowding on biological reactions ended up as a thesis on how proteins recognize their DNA targets with high fidelity. The meandered path, for sure, would have turned into an insurmountable maze if not for the support of several people at different stages and at different scales. The optimistic and encouraging attitude of my parents (Dr. Ramakrishnan and Mrs. Premalatha), despite their long separations (across all the four dimensions) from me, and my very supportive sisters (Mrs. Bhuvaneswari and Dr. Subasree) are indeed the foremost reasons for where I have reached. The warmth and support extended by my cousin Mrs. Deepa and her family throughout my stay in Singapore is incalculable.

"Curiosity killed the cat" goes the popular saying. I would have certainly been a perfect example of this quote if not for my thesis advisor Prof. Raj Rajagopalan. Although he allowed me to cruise on my enthusiastic expeditions, his knack to steer at the right moment was quintessential for me not to have become an iconic example of the above quote. For this, I am greatly indebted to him. I did learn very many things from him and his enthusiasm for Science is very contagious indeed. I am also very much thankful to Prof. Michael Raghunath and Prof. K P Mohanan who helped me shape my perspectives on Science and Education. Particularly, I cherish the debates that I had with Prof. Mohanan on several issues on science education in general. I also thank Prof. Jiang Jianwen who was very supportive particularly during the initial years of my graduate

school when I was transitioning from being an undergraduate student to a graduate student.

A substantial part of the support to meander through the vicissitudes of the graduate school came from my friends at NUS, particularly, Dr. Karthiga Nagarajan, Mr. Vivek Vasudevan, Dr. Satyen Gautam, Mr. Sundaramurthy Jayaraman and my friends elsewhere around the globe, Mr. Gopuraja Dharmalingam, Mr. Thilak Rajasekaran, Dr. Kaushik Raghunathan, Mr. Madhu Balasubramanian, Mr. Santio Ruban and Mr. Vasanthakumar Chandran. My friends in the research team Dr. Karthik Harve, Dr. Søren Enemark, Dr. Abdul Rajjak Shaikh and Mr. Srivatsan Jagannathan were all instrumental in shaping my thesis and providing immense support. Particularly, the tea sessions with Dr. Søren Enemark, Dr. Abdul Rajjak Shaikh and Mr. Srivatsan Jagannathan were fun and refreshing. I also thank our lab officers, Ms. Chow Pek, Ms. Chew Su Mei Novel, Ms. Tay Kaisi Alyssa, Mr. Ang Wee Siong and Ms. Yan Fang who were all very helpful. I also thank Ms. Sivaneswari Raj, Ms. Saroja Ramasamy, Ms. Rita Mary and Ms. Doris How Yoke Leng, our administrative support officers, for their immense help in assisting me with any departmental matters and providing a congenial atmosphere.

There certainly were very many friends who have helped me throughout the graduate school and I might not have listed them all here. To all of them, I express my sincere gratitude.

Thank you everyone!

TABLE OF CONTENTS

SUMMARY

Protein-DNA interactions form the basis for many cellular processes. How a protein rapidly identifies its target (cognate) DNA sequence from among a sea of random (noncognate) sequences is an intriguing area owing to its innate fundamental importance and its role in developing therapeutic gene modulation strategies.

Many DNA-binding proteins, including restriction endonucleases, diffuse linearly along the DNA over short segments in addition to exhibiting 3D diffusion, hopping, intersegmental transfers, etc. The linear diffusion of proteins along the DNA has been suggested as a mechanism by which proteins enhance their ‘searching’ speed. The question then is how proteins discriminate between the cognate and noncognate sequences *as* they slide over the DNA segments. Several factors and/or properties of the binding partners have been proposed to act in concert to bring about the specificity in protein-DNA interactions. Of these, precise positioning of hydrogen bonding donors and acceptors in the protein and DNA interfaces was the one to be proposed first and subsequently confirmed by various studies, primarily x-ray crystallographic structures. The crystal structures of protein-DNA complexes, in addition, also revealed the presence of, in most cases, ‘deformed’ DNA and interfacial waters. These observations collectively led to the idea that specificity is achieved when the protein is able to ‘deform’ the DNA and form the precise hydrogen bonds. Subsequent studies also suggested various roles for water in molecular recognition. However, despite the numerous efforts by various researchers, the question of specificity in protein-DNA interactions still remains incompletely answered and the holy grail of a protein-DNA recognition code unreached. While this is partly because of the inherently complex nature of the problem, it is also because of lack of systematic studies for a *particular* enzyme elucidating its range of structural/dynamical responses and attendant changes as it binds to various noncognate sequences which would provide clues to the various underlying principles in protein-DNA recognition.

The scope of this thesis is to systematically investigate the structural/dynamic responses and the attendant changes when a protein binds to noncognate sequences compared against the cognate sequence. Three factors, namely, intrinsic dynamics of the protein, dynamics and thermodynamics of water in the hydration layer and the sequence-dependent DNA conformational responses for EcoRI, a type II restriction endonuclease, were investigated using molecular dynamics simulations. The choice of EcoRI, one of the first proteins to be co-crystallized with the DNA, stems from the fact that EcoRI minimally restructures upon binding to the DNA. The choice of a minimally restructuring protein allows one to isolate and examine the issues of interest (here, the intrinsic dynamics of the protein, water dynamics and DNA conformation) relatively unfettered and unclouded by the dynamics driving unfolding and folding events. Such cases can serve as a building block for developing an overall picture of protein-DNA interactions.

We first characterized the intrinsic dynamics of the protein and the dynamics and thermodynamics of water in the hydration layer for EcoRI bound to a noncognate sequence (TAATTC) that differs from the cognate sequence (GAATTC) by just a single basepair. The replacement of G with T represents the least perturbation to the protein-DNA complex, that is, a loss of just one hydrogen bond. The TAATTC sequence is also the next-preferred sequence of cleavage for EcoRI. Thus, in essence, we asked how the (a) protein dynamics and (b) water dynamics vary when the protein shows minimal

rearrangement and the perturbation in the substrate is the least. The main results are summarized as follows:

a) Essential dynamics analyses of EcoRI reveal that the overall dynamics of the protein subunits change from a coordinated motion in the cognate complex to a scrambled motion in the noncognate complex. This dynamical difference extends to the protein-DNA interface where EcoRI tries to constrict the DNA in the cognate complex. The motion of the C_α atoms of the residues in the recognition site of the noncognate complex are roughly orthogonal to those in the cognate complex indicating that the motion in the noncognate complex is tangential to the DNA. These differences in the dynamics coupled with structural relaxation of the arms leaves the DNA in the noncognate complex unkinked.

b) The noncognate complex is more hydrated than the cognate complex with 45 more water molecules in the interfacial region. The interfacial and intercalating waters in the noncognate complex exhibit a faster reorientational dynamics, which in turn reduces the water-protein/DNA hydrogen-bond lifetimes in the noncognate complex. The entropy and enthalpy of water in the interfacial and intercalating regions in the two complexes are essentially the same.

Having investigated the changes in the dynamics of the protein and water when EcoRI binds to a minimally mutated DNA sequence, we then asked how the protein (here, EcoRI) environment influences the conformation of DNA sequences that differ by just a single basepair. The results reveal that while the DNA conformational differences are prominent at the basepair *step* level for free DNA chains, the differences become prominent even at the level of basepairs in the protein-bound form. The protein induces

long-range correlations in the DNA conformation in the sequence it is bound to. This long-range correlation and amplification of DNA conformational differences at the basepair-level leads to a ‘structural misfit’ of the DNA in the protein *throughout* the recognition sequence.

The above studies suggest collectively that when EcoRI chances upon its cognate sequence, specific domains in the protein undergo dynamical changes, which, along with the reduction in the dynamics of water in the hydration layer and sequence-dependent DNA conformational changes promote the formation of a stable complex. Even a minimal mutation of the DNA sequence is enough to alter the DNA conformation, the dynamics of the interfacial residues and the dynamics of water sufficient to make the complex unfit for required function.

In summary, this thesis sheds light on the structural/dynamic responses and the attendant changes when a protein binds to minimally mutated noncognate sequences. The cases presented in this work can serve as building blocks for developing an overall picture of protein-DNA interactions.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Q	Total canonical partition function of a system
q	Partition function of individual normal modes
N	Number of atoms
υ	Frequency
$g(\upsilon)$	Density of states distribution
E	Internal Energy
T	Temperature
W_E	Classical weighting function for internal energy
W_A	Classical weighting function for Helmholtz free energy
W_S	Classical weighting function for entropy
S	Entropy
k	Boltzmann constant
A	Helmholtz free energy
V_0	Reference energy
q_{HO}^{Q}	Quantum harmonic partition function
h	Planck's constant
$W_E^Q(\upsilon)$	Quantum weighting function as a function of frequency υ for internal energy
$W_S^Q(\upsilon)$	Quantum weighting function as a function of frequency υ for entropy
$W_A^Q(\upsilon)$	Quantum weighting function as a function of frequency υ for Helmholtz free energy
v_j^k	Velocity of atom j along the k$^{\text{th}}$ component

$F(\upsilon)$	Fourier transformation
$c(t)$	Autocorrelation function of individual molecule
$C(t)$	Sum of the autocorrelation function of all the molecules
$x(t)$	Coordinate of an atom at time t
m_j	Mass of atom j
D	Diffusion coefficient
α	Enskog friction factor
f	Fluidicity factor
ρ	Density
y	Volume fraction
$z(y)$	Compressibility factor
σ_{HS}	Diameter of the hard-sphere particle
ϕ	Rescaled volume fraction
Δ	Normalized diffusivity
$W_E^g(\upsilon)$	Quantum weighting function of internal energy for the gas component
$W_S^g(\upsilon)$	Quantum weighting function of entropy for the gas component
$W_A^g(\upsilon)$	Quantum weighting function of Helmholtz free energy for the gas component
S^{HS}	Entropy of hard sphere
$e^{\mu}(t)$	Unit vector along the direction of the dipole moment vector at time t for water

1 INTRODUCTION

> “*The most beautiful thing we can experience is the mysterious. It is the source of all true art and science. He to whom the emotion is a stranger, who can no longer pause and stand wrapped in awe, is as good as dead; his eyes are closed.*”
>
> – Albert Einstein

Our interest today in a molecular level understanding of protein-DNA interactions, which forms the focus of this thesis, has evolved over thousands of years from man’s curiosity about his inheritance of parental traits. An absorbing interest in the world around him triggered man to seek explanations for his observations in the surroundings. One such observation is the resemblance he saw between a parent and a child, be it in humans, animals or plants. The earliest documented explanation for the inheritance of paternal traits is that of Hippocrates’ (*ca*. 460 BC – *ca*. 370 BC), who proposed the *pangenesis* theory. According to this theory, “inheritance is based on the production of specific particles (“seeds”) by all parts of the body and transmission of these at the time of conception” [1]. If this were the case, then children would only have physical resemblance to their parents. On the contrary, nonphysical features such as voice, gait, etc. were also seen to be inherited by the children. Further, it was noticed that children also inherited the characteristics of their remote ancestors. In addition, if the two parents produced the “seeds” then wouldn’t we expect offsprings with two heads, four arms and so on? These and other arguments were put forth by Aristotle (384-322 BC), who later rejected the pangenesis theory. He asked “Why not admit straight away that the semen (the term was used to refer to the reproductive elements of both sexes which we call as the ova and sperm today) ... is such that out of it blood and flesh can be formed, instead of maintaining that semen is itself both blood and flesh?” [2]. Thus he linked the seemingly disparate fields of genetics and development. Aristotle’s ideas were the conceptual limit on the theory of inheritance for the next ~2000 years after which Charles Darwin (1809-1882 AD) adopted the pangenesis theory and proposed the concept of “gemmules” to explain the huge data he had assembled on his observations of inheritance in animals. Although Darwin’s theory was not successful, it cannot be denied that he laid the foundation for scientific approach in addressing problems. The problem of inheritance was simultaneously studied by Gregor Mendel (1822-1884 AD), who laid the foundation of modern genetics. He associated each trait with a “unit” or “factor” that gets passed on to the descendant and explained the nature of inheritance of these “units” (see [3]). These “units” are now called the genes. Further works by other scientists such as Hugo de Vries, Walter Sutton, William Bateson, Thomas Morgan and several others established an acceptable theory of “transmission” genetics which we know of today. Although an acceptable theory was established by 1930, it was still not known what the chemical nature of a gene was and what precisely it did. The answers to these questions were slowly revealed with the discovery of DNA as the genetic material [4] and the discovery of the double-helical structure of DNA, which were instrumental in explaining the mechanism of DNA replication [5].

Parallel to these investigations on inheritance and the focus on DNA were the investigations on proteins and their composition. The French chemist Antoine Fourcroy (1755 – 1809) identified three distinct varieties of protein from animal sources in 1789, the albumin, fibrin and gelatin [6]. Since then, advances in the analysis of elemental composition of compounds enabled several researchers to investigate proteins for their elemental composition. Particularly, Gerrit Jan Mulder's analyses led him to conclude that the albuminous substances consisted of the radical $C_{40}H_{62}N_{10}O_{12}$ to which varying amounts of sulfur and/or phosphorus were attached [7]. Jöns Jakob Berzelius then suggested the name 'proteine' for this radical [7]. What started with the elemental compositional analysis of proteins slowly evolved and merged into developmental biology (or gene regulation) when Jacobs and Monod propounded the theory of the operon in 1961 [8]. Their theory was based on their observations of induction of the lac gene. The isolation and characterization of the lac repressor, a protein, and the discovery that it actually bound to specific DNA sequences marked the beginning of the investigations on protein-DNA interactions and gene regulation in general. Since then, researchers have made great strides in understanding the molecular basis of gene regulation, and, proteins, no doubt, play a crucial role in gene regulation. Thus the history of protein-DNA interaction stemmed from man's curiosity about what he saw about inheritance of paternal traits. Since then molecular-level understanding of gene regulation and embryo development has taken great strides, including discoveries of other biomolecules that are involved in the process. The focus of this thesis, however, is limited to studying the underlying mechanisms of protein-DNA interactions. In the next section, we give a brief overview of the different classes of DNA-binding proteins before we discuss the current theories on protein-DNA interactions.

1.1 Protein-DNA Interactions

There are several proteins that bind to the DNA inside the cell. Depending on the their functions, they are broadly classified as follows [9] :

1. **Regulatory proteins**: These proteins control the transcription of a particular gene by binding to specific signal sequences such as the 5'-TATA.
2. **DNA cleavage proteins:** These are a class of proteins with varying degrees of specificity to the DNA sequence and cleave the DNA. For example, the DNAseI has little sequence specificity while restriction enzymes such as EcoRI are highly specific in the sequence they cleave.
3. **Repair proteins:** This is an important class of proteins that recognize lesions in the DNA and repair them by excising or joining the breaks in the damaged DNA.
4. **Topology modifying proteins:** These important therapeutic targets wind or unwind DNA prior to replication (e.g. DNA Topoisomerases).
5. **Structural proteins:** Structural proteins are those that maintain the integrity of the folded DNA, e.g. histones in chromatin.
6. **Processing proteins**: These proteins use the DNA as a template for further nucleic acid synthesis. Eg. DNA and RNA polymerases.

As one might see from the above classification, DNA-binding proteins, particularly the regulatory proteins and the DNA cleaving proteins, have a window of DNA sequence preference. Some are extremely specific (e.g., restriction endonucleases such as EcoRI, EcoRV etc) and some bind to a class of DNA sequences (e.g., regulatory proteins binding to a TATA box). These proteins (the regulatory proteins and the DNA cleavage proteins) are instrumental in gene expression, regulation and in self-defense. Given the fact that the long genomic DNA (3.2 Gigabases in a human cell [10]) is packaged inside the cell with multiple hierarchies of DNA folding, the intriguing aspect in such protein-DNA interactions is how these proteins *rapidly* identify their target DNA sequences with such *high fidelity*. The specificity of protein-DNA interactions, coupled with their fast search has been an active area of research for many decades now. The search for a recognition code has been the holy grail of many scientists. In the next few sections, an overview of the efforts towards understanding protein-DNA interactions is described. This is then followed by a discussion of why there is a necessity to study the underlying mechanisms in protein-DNA interactions. The ensuing section then discusses, in the backdrop of all that was discussed, the scope and objectives of this thesis.

1.2 Mechanisms of Protein-DNA Interaction: Status Quo

1.2.1 Facilitated Target Location

After the DNA structure was solved in the 1950's, there was progressive understanding of gene duplication and expression [5, 8, 11]. However, how gene regulation works at molecular level was not clear until 1967. In 1967, Ptashne [12] and Gilbert & Müller-Hill [13] showed that proteins bind directly to specific DNA sequences to regulate (repress in their cases) transcription of the DNA to RNA in contrast to the previous ideas that the repressor protein interacts with the mRNA to prevent translation of the encoded message. In 1970, after three years of the first demonstration that proteins had the ability to bind to specific DNA sequences [12, 13], the first kinetic studies of a sequence-specific association of a protein with DNA were reported by Riggs et al. [14]. The rate constant for the binding reaction was measured to be 7 x 10^9 $M^{-1}s^{-1}$, a value that was noted to be about 100-fold faster than the upper limit estimated for macromolecules of that size by 3D diffusion (by the Smoluchowski equation). Riggs et al. [14] suggested, based on the ionic strength-dependency of the association constant, that the long-range attractive electrostatic forces between the repressor and the DNA accelerated greatly the association reaction than that predicted by the three-dimensional random walk. This surprising observation triggered a series of studies to investigate the possible mechanisms proteins might use to accelerate their search for their target DNA sequence. Seminal works by Peter von Hippel, Otto Berg and others led to several diffusion-based models to explain the rapid association of the protein and the DNA [15-18]. These mechanisms include [19]

(i) **One-dimensional diffusion (sliding)**

In this model, the protein is assumed to exhibit a random walk along the DNA. All throughout the random walk, the protein is in association with the DNA.

(ii) **One-dimensional hopping**

When the protein moves along the DNA by a series of microscopic dissociation and rebinding events, the protein is said to exhibit one-dimensional hopping.

(iii) **Jumping**

In this model, the protein moves over longer distances in the DNA by dissociation at a particular site and rebinding at a different, distal site.

(iv) **Intersegmental transfer**

This model proposes the transfer of proteins between distal sites via a looped intermediate. Eg: lac repressor.

Figure 1 shows a schematic representation of the above-discussed models. The development of these ideas and its proof collectively laid the intellectual ground work for all subsequent studies on facilitated target location studies. Several recent single-molecular studies have now shown the presence of one-dimensional diffusion or the sliding of proteins along the DNA [20-24]. Recently, Gorman et al. showed that eukaryotic proteins hop to overcome obstacles such as other bound proteins [25]. Raghunathan et al. [26] showed that the RecA protein moves 3 nucleotides per step. These observations have, collectively, led to the idea that a combination of 1D and 3D diffusional walks bring about the protein-DNA interactions [27, 28].

Parallel to the investigation of the facilitated-target-search mechanism, efforts were also devoted toward understanding the structural origins of specificity. Studies on the structural aspect of protein-DNA interactions help to make a more thorough picture of protein-DNA interactions and are described in the next section.

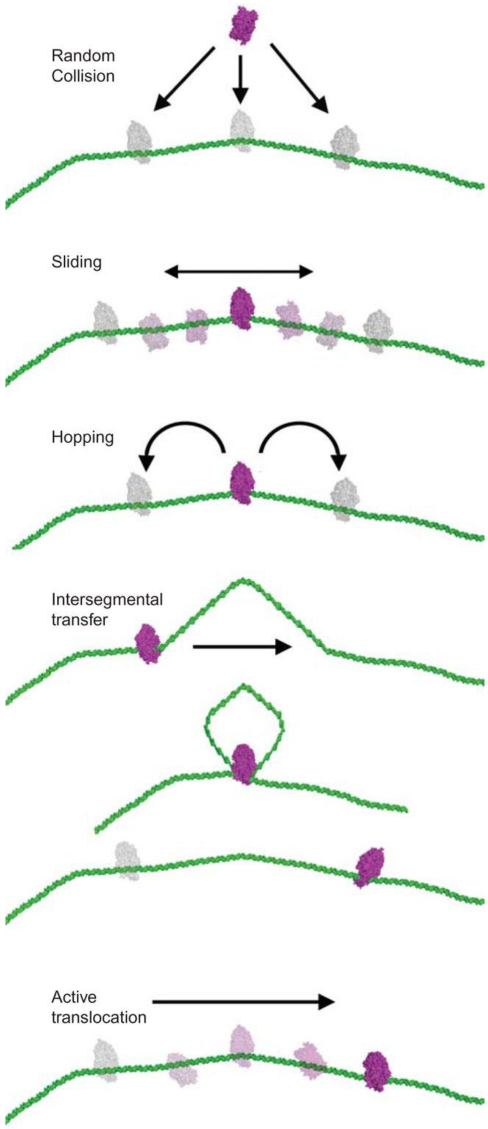


Figure 1-1. Schematic representation of the various diffusion-based models for protein-DNA interactions. (Adopted from Gorman and Greene [19].)

1.2.2 Structural Insights into the Specificity of Protein-DNA Interactions

> “The minimal model implies that only one or very few protein sequences (with regard to hydrogen-bond forming amino-acid) exist which bind one particular DNA sequence. If this is true there must exist rules which describe the binding of protein sequences to DNA sequences” [29].

In 1972, Adler et al. [29] conceived the idea that there must exist rules to the binding of protein sequences to DNA sequences. Four years later, Seeman et al. [30] proposed several hydrogen-bonding interactions that could be a part of this protein-DNA code. They cautiously concluded that

> "Single hydrogen bonding is inadequate for the complete identification of base pairs, but that pairs of hydrogen-bonded interactions may play a role in this process. It is hoped that proposals set forth here will serve to stimulate experiments which may eventually reveal the mechanism for protein-nucleic acid recognition."

As an attestation to their caution, several crystal structures of protein-DNA complexes (lac, EcoRI, EcoRV, Cro repressor), revealed no strict code for DNA recognition. Brian W. Matthews [31] concluded, in 1988,

> "Is there a code whereby certain DNA basepairs are recognized by certain amino acids? … The answer, again is no … The DNA-protein interface is seen to be very complex, with several side-chains sometimes contributing to the recognition of a single base … It is very satisfying now to have in hand the structures of several repressor-operator complexes that vindicate the general principles of DNA-protein recognition that have been developed by many individuals during the past 20 years. But the full appreciation of the complexity and individuality of each complex will be discouraging to anyone hoping to find simple answers to the recognition problem."

Despite the revealing that there cannot be a single recognition code to protein-DNA interaction, the crystal structures were pivotal to revealing at least two of the important aspects in protein-DNA interaction which have gained considerable attention thereafter. These aspects are a) DNA deformability and b) interfacial waters. DNA in most of the protein-DNA complexes was “deformed”. Analysis of several protein-DNA complexes in which the DNA was kinked revealed a DNA sequence-dependent pattern in the deformability of a DNA [32]. Further, the presence of waters at key positions between the protein and the DNA surfaces suggested that water plays an important role in protein-DNA recognition. Thus, it was understood that several factors, in addition to the direct interactions between the protein and the DNA, contribute to the specificity in protein-DNA interaction. In addition, recent works and understanding that biomolecules are dynamic entities and not static entities have led to the proposition that protein intrinsic dynamics plays an important role in determining the mechanisms of its interactions [33, 34].

These observations collectively led to the idea that specificity is achieved when the protein is able to ‘deform’ the DNA and form the precise hydrogen bonds and that the protein dynamics and interfacial waters help to achieve the desired recognition. Questions that remain, however, include how proteins actually deform the DNA as they slide over the DNA? What is the source of the sequence-dependent alteration in the deformability of the DNA as the protein binds? What is the relation between hydration, DNA deformation, and protein binding? What is the relation between the intrinsic dynamics of the protein in binding to DNA and attendant conformational changes? Thus, despite our long strides in understanding several principles of protein-DNA interaction, we are still quite far away from a full picture.

1.3 Why Study the Mechanisms of Protein-DNA Recognition: Therapeutic Importance

As discussed above, protein-DNA interactions represent one of the fundamental biomolecular interactions in the cell and pose intriguing challenges. In addition to the fundamental interest, delineating the mechanisms of protein-DNA interactions holds promises for the *rational* design and development of therapeutic strategies for endogenous gene modulation. Endogenous modulation of gene function is an attractive concept wherein, in contrast to conventional gene therapeutic strategies where the downstream products (mRNA or protein) are targeted, the gene (the DNA sequence) is targeted directly. Thus, it can be very effective because only a fewer copies have to be targeted. Further, this approach does not suffer from problems due to DNA methylation, which leads to loss of function in approaches that integrate gene copies. Central to the gene modulation approaches is the availability of agents that bind to specific DNA sequences. These agents include Triplex Forming Oligonucleotides (TFOs), synthetic polyamides and designer zinc finger proteins. TFO is a synthetic single stranded oligonucleotide which binds to a specific DNA and forms a triple-helical structure (see [35] for a detailed review on these). However, a major limitation to the application of TFOs is that they can only bind to purine-rich target strands [35]. Chemical modifications to TFOs such as modifications to the phospho-diester backbone [36-39] , the ribose [40-43] or the base [44-46] moiety have recently shown a promising potential to overcome the limitation of the affinity to purine-rich targets. In addition to this major limitation, other concerns such as binding affinity and specificity, uptake into cells and in vivo stability [35] necessitate the development of newer and effective DNA-binding agents.

The next class of DNA-binding agents, synthetic polyamides, is a class of agents that has been engineered rationally based on the DNA-binding mechanisms of the natural products netropsin and distamycin. Stretches of these polyamides, containing the aminoacids hydroxypyrrole (Hp), imidazole (Im) and pyrrole (Py), form a hairpin structure that binds via hydrogen bonding to specific basepairs in the minor groove of DNA [35]. Specifically, the polyamide aminoacid pairs Py/Im, Py/Hp, Hp/Py and Im/Py recognize the C-G, A-T, T-A and G-C basepairs respectively [35], thus conferring specificity in binding. The major shortcoming of synthetic polyamides is the shortness of their DNA target sites. Elongation of the aminoacid pairings to recognize a longer DNA target sequence fails because of the over-bending of the polyamide structure relative to the minor groove of the DNA [47]. Several strategies to improvise the use of these class of agents is underway (see [35] for further details).

Zinc finger proteins, or DNA-binding proteins in general, are the other class of DNA-binding agents. This class of agents is promising because of its high target DNA specificity to about 6bp of DNA and its ‘naturalness’. Despite the lack of a “recognition code”, there have been several knowledge-based strategies to engineer the protein to bind to specific DNA sequences [48-50]. Thus we see that there is a need for clear delineation of protein-DNA binding mechanisms either to get inspired for strategies (like that of synthetic polyamides) or to rationally re-engineer protein-DNA interfaces.

1.4 Scope and Objectives of this Thesis

As discussed towards the end of section 1.2, despite our progress in understanding the mechanisms of protein-DNA interactions, we are still far from a complete understanding of how proteins achieve specificity (and such a clear understanding is essential as discussed in section 1.3). While this is partly because of the inherently complex nature of the problem, it is also because of a lack of systematic studies for a *particular* enzyme to elucidate its range of structural/dynamical responses and attendant changes as it binds to various noncognate sequences which would provide clues to the various underlying principles in protein-DNA recognition. The scope of this thesis is thus to systematically investigate the structural/dynamic responses and the attendant changes when a protein binds to various noncognate sequences compared against the cognate sequence. Specifically, three factors, namely, DNA structure, protein dynamics and water dynamics and thermodynamics are investigated for a protein when it is bound to noncognate sequences.

1.5 Choice of a Model

The choice of the DNA-binding protein to investigate the issues of protein-DNA interaction is critical. Restriction enzymes are advantageous and suitable models for the purpose because of their high specificity to short (usually 6 bp) DNA sequences. EcoRI is one such restriction endonuclease which cleaves the DNA at the $(GAATTC)_2$ sequence. It is one of the first proteins to be co-crystallized with its cognate sequence. The availability of crystal structure, extensive kinetic and thermodynamic studies, and several mutational studies make it a suitable candidate for our choice. Furthermore, the minimal restructuring of EcoRI upon binding to its cognate sequence makes it an ideal choice to investigate the issues unfettered and unclouded by the dynamics and attendant protein folding events[1]. Therefore, in this thesis, we focus on the binding of EcoRI to DNA sequences.

1.6 Organization of the Thesis

This thesis is organized into six chapters. Chapter 2 presents an overview of key studies related to EcoRI-DNA interactions including the roles of water and protein dynamics. Chapter 3 investigates the effect of a minimal mutation in the DNA on the intrinsic dynamics of EcoRI, and we show that even such small perturbations in the substrate are enough to alter the dynamics of EcoRI. In Chapter 4, we investigate the dynamic and thermodynamic properties of water around the EcoRI-DNA complex when bound to a cognate and a noncognate DNA sequence and show that the intercalating waters, particularly, show a decreased reorientational dynamics in the cognate sequence. In Chapter 5, we investigate the role of a protein environment on DNA structure and show that the protein (here, EcoRI) alters the DNA conformation in a sequence-dependent manner and that the changes occur at basepair level in addition to basestep levels. Finally, we summarize the key findings in light of the broader picture of protein-DNA recognition and propose some further works based on the insights gained in above-presented investigations in Chapter 6.

[1] The root mean-squared deviation of C_α atoms obtained after fitting the DNA-free crystal structure of EcoRI (pdb id: 1QC9) and the crystal structure of EcoRI with the cognate DNA (pdb id: 1ERI) is 2.06 Å.

2 PROTEIN-DNA RECOGNITION: OVERVIEW & STATUS QUO

> *"I don't know anything, but I do know that everything is interesting if you go into it deeply enough."*
>
> \- Richard Feynman

Restriction endonucleases have been apt models to study the specificity of protein-DNA interactions because of their very high selectivity to short duplex DNA targets. EcoRI is one such restriction enzyme that has been investigated extensively from kinetic, thermodynamic and structural perspectives. EcoRI, in the presence of Mg^{2+} ion, catalyses the cleavage of the phospho-diester bond between guanine and adenine in the palindromic sequence $(GAATTC)_2$. The exceptional selectivity of EcoRI to this DNA site is exemplified by the fact that the difference in the transition-state interaction free energy for sites that differ by just 1 bp is between 6 - 13 kcal/mol and those sites that differ by 2 or more basepairs are not cleaved at all [51]. The high selectivity has been speculated to be the result of various "direct" and "indirect" readout mechanisms that include loss in one or more hydrogen bonds between the protein and DNA, steric clashes that arise out of inappropriate positioning of a functional group in the base and the increased cost in attaining the DNA conformation in the transition complex [51]. "Direct readout" refers to the contacts between the protein and DNA mainly by hydrogen bonds, whereas "indirect readout" refers to other mechanisms (aside from direct protein-DNA contacts) affecting the DNA sequence-dependence of protein-DNA interactions. Considerable effort has been devoted to elucidate the contributions of the direct and indirect readouts towards specificity in EcoRI-DNA interactions and protein-DNA interactions in general [52-55]. Since the direct and indirect readout mechanisms have been extensively reviewed by

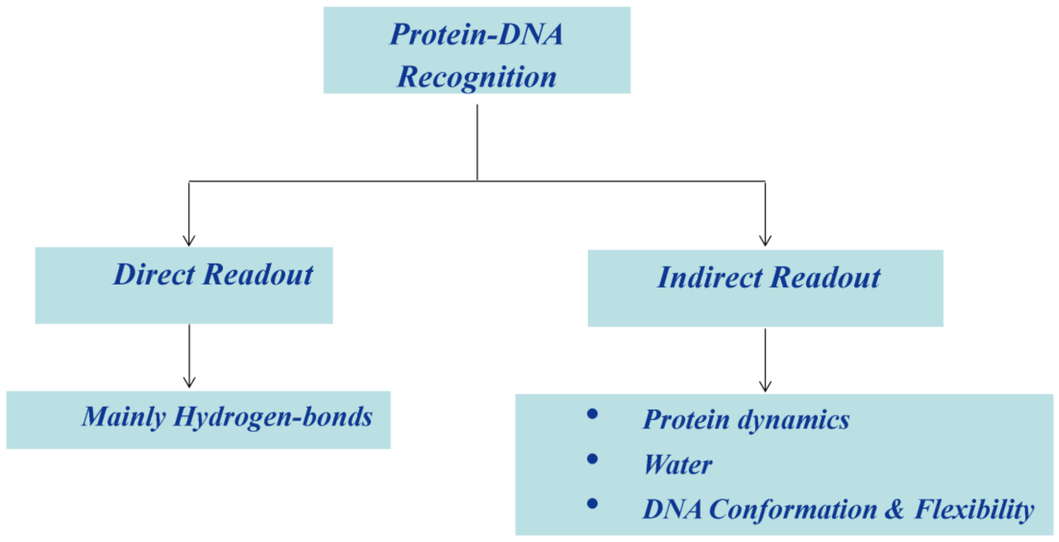


Figure 2-1. An overview of the various protein-DNA recognition mechanisms.

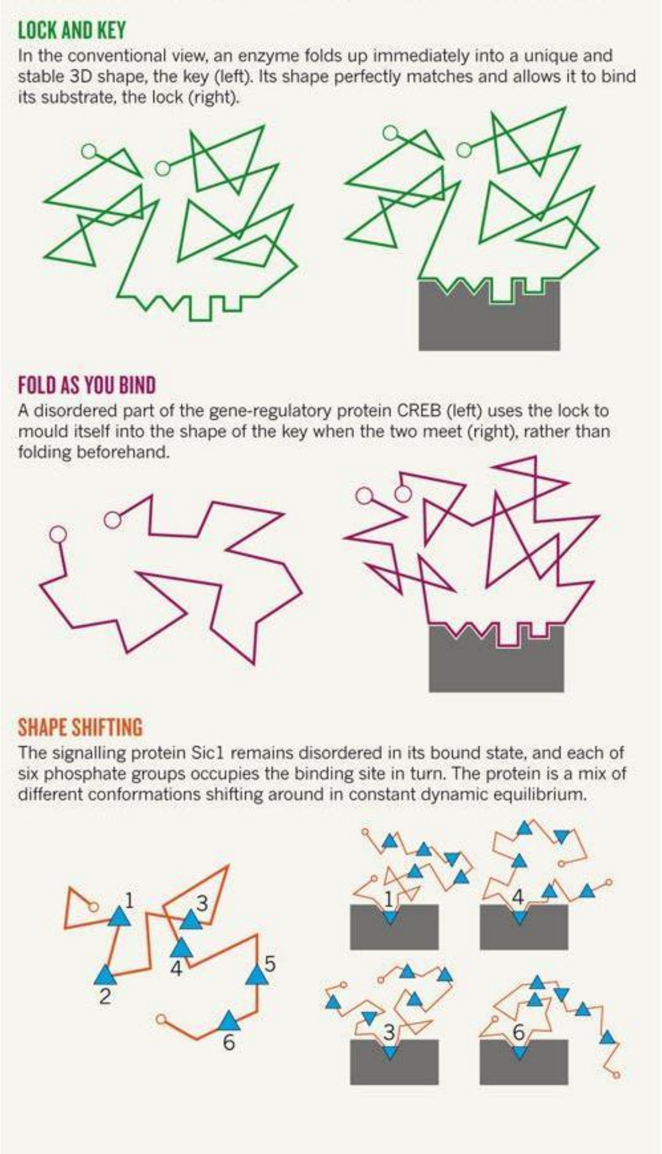
several researchers [33, 56-61] , we restrict the scope of this chapter to discuss only the most essential information. In the next section we discuss the direct readout mechanism in EcoRI before we move on to discuss the indirect readout mechanisms (protein dynamics, role of water and sequence-dependent DNA properties).

2.1 Direct Readout in EcoRI

Structural and mutational studies reveal that EcoRI makes extensive contacts throughout the recognition site. The original recognition model was based on the X-ray crystal structure of EcoRI-DNA complex [62]. According to this model, EcoRI made contacts with the purines, and it was claimed that Arg200 interacted with guanine and that Glu144/Arg145 recognized both the adenines to make a total of twelve hydrogen bonds. However, a subsequent study [63] showed that EcoRI made contacts with the pyrimidines as well. A difference in any of the basepairs in the recognition sequence would, thus, disrupt one or more hydrogen bonds enabling discrimination. Lesser et al. [51] estimated that the introduction of one incorrect basepair into the recognition sequence can cost +6 to +13 kcal/mol in the transition state interaction energy. They further investigated the binding of EcoRI to a set of purine-base analogue sites, each of which was formed by deleting one functional group that forms a hydrogen bond with EcoRI [52] and inferred that, in general, the binding free energy penalty of deletion varies between +1.3 to +1.7 kcal/mol. They also further estimated that the incremental energetic contribution of one protein-base hydrogen bond is about –1.5 kcal/mol. Interestingly, Lesser et al. [52] noted that the deletion of the N6 amino groups in the second adenine of the recognition sequence improved binding by –1.0 kcal/mol and inferred that this favorable effect arises because the penalty of deleting a protein-base hydrogen bond is outweighed by the facilitation of the required DNA distortion. Quantification of the contribution of the contacts enabled Lesser et al. [51] to calculate the total energy of binding as a function of the individual contacts seen in the crystal structure. Interestingly, their study revealed that the total binding energy is not just the sum of energetic contributions from each of the protein-DNA contacts, but that there were additional factors. Further, the crystal structure of the EcoRI-DNA complex showed that the DNA was ‘kinked’ at the central recognition step [64]. From these observations, Lesser et al. [51] concluded that the net protein-DNA binding energy is a result of various other factors that include conformational rearrangements of the protein, DNA, water and ions. In the next sections on indirect readout mechanisms, we first discuss the role of protein dynamics in protein-DNA interactions, role of water and then the sequence-dependent DNA properties.

2.2 Indirect Readout Mechanisms: Protein Dynamics

While significant effort has been invested in investigating the necessary and crucial contacts between the protein and the DNA and the residues involved in binding and catalysis, etc., independent studies have also showed the importance of dynamics of a protein for its function. For example, Eisenmesser et al. [65] showed, using NMR relaxation technique, that the rate of structural rearrangements of specific protein residues of cyclophilin A involved in the catalysis of the substrate is intimately connected to the microscopic rates of substrate turnover. Wang et al. [66] showed that the dynamics of the residues adjacent to the active site of the binase ribonuclease are extremely flexible and facilitate access to the substrate by structural rearrangements of these residues, thus indicating that the dynamics of the protein is crucial in binding events. Recently, Su et al. [67] showed that protein unfolding motions are significantly influenced by structure-



LOCK AND KEY

In the conventional view, an enzyme folds up immediately into a unique and stable 3D shape, the key (left). Its shape perfectly matches and allows it to bind its substrate, the lock (right).

FOLD AS YOU BIND

A disordered part of the gene-regulatory protein CREB (left) uses the lock to mould itself into the shape of the key when the two meet (right), rather than folding beforehand.

SHAPE SHIFTING

The signalling protein Sic1 remains disordered in its bound state, and each of six phosphate groups occupies the binding site in turn. The protein is a mix of different conformations shifting around in constant dynamic equilibrium.

Figure 2-2. An illustration showing the various ideas of protein dynamics in ligand binding (adopted from [68]).

encoded dynamical properties. Martinez et al. [69] showed that aminoacid substitutions in the psychrophilic protease subtilisin S41 lead to a change in the principal fluxional modes allowing the protein to explore a different subset of conformations. In the specific context of protein-DNA interactions, Kalodimos et al. [70] observed from NMR experiments that the conformational substates of the free *lac* DNA Binding Domain

(DBD) redistribute upon binding to the cognate sequence but not when binding to noncognate sequences. They attributed the difference in the redistribution of the conformational substates to a change in the dynamics of the lac DBD upon binding to the cognate DNA sequence. Cave et al. [71] observed that the backbone dynamics of the basic/helix-loop-helix domain of the Pho4 protein from *Saccharomyces cerevisiae* shows large differences upon binding of the protein to the DNA. In addition, they noted that the overall backbone dynamics of the protein remains similar regardless of whether the complexation is with the cognate sequence or the noncognate sequence. However, two of the protein residues do show different backbone flexibility depending on whether the protein is bound to the cognate sequence or the noncognate sequence, suggesting possible role of dynamics in sequence discrimination. Recently, Brown et al. [72] showed using NMR experiments that flexibility of the DNA binding domain of the human papillomavirus E2 protein is essential for the recognition of its target site. Doruker et al. [73], based on the elastic network model of EcoRI, studied the collective dynamics of EcoRI. Uyar et al. [74], based on computational analysis, suggested that the dynamics of the β-strands around the DNA binding region in restriction endonucleases may have a role for target site recognition and cleavage. The dynamics of intrinsically disordered segments of proteins in DNA recognition has also been discussed [33, 75]. Thus we see that protein dynamics plays an important role in protein-DNA interactions. In the next section, we discuss the role of water in protein-DNA interaction.

2.3 Indirect Readout Mechanism: Role of Water

Water plays an important role in biomolecular function [76-79]. Particularly, water at the surface of biomolecules has been shown to play key roles in biological processes such as molecular recognition, biomolecular interactions, etc., [80-84] as noted below. The markedly different dynamic and thermodynamic properties of interfacial waters from bulk water properties [85-87] and the interaction of water with specific groups in the biomolecules have been identified for the influence of water. Kasson et al. [79] show that water between two membranes exhibits decreased mobility compared to the bulk water and that the “structured” water between the two membranes controls the fusion of the membranes. Ahmad et al. [84] show in the case of association of hydrophilic proteins that water in the interfacial gap, in addition to forming an adhesive hydrogen-bond network that stabilizes the intermediates, also generates a preferred directionality for electrostatic interactions that drives the interfaces towards each other. Adkar et al. [77] show that the interaction of water with the polar groups of the enzyme adenylate kinase stabilizes the intermediate state during enzymatic catalysis. In the context of protein-DNA interactions water has been proposed to play a wide variety of roles including being a hydrogen bond donor/acceptor at the protein-DNA interface, filler to maintain packing densities at the interface and buffer to screen unfavorable electrostatic interactions [61, 88]. Specifically, in the *trp* repressor-operator complex [89, 90] and the BamHI-DNA complex [60, 91], water molecules, via a network of hydrogen bonds, allow amino acids which are otherwise out of reach of the bases to make contacts which are required for specific

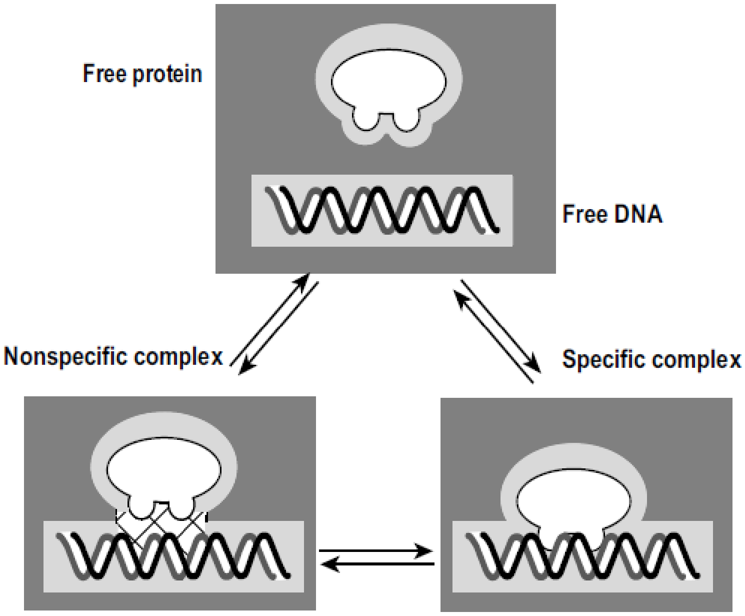


Figure 2-3. An illustration showing the exclusion of water molecules at the interface of the protein and DNA during the formation of the specific complex. (Adopted from [92])

binding [60]. The structures of the specific and nonspecific complexes of glucocorticoid receptor DNA-binding domain (GRDBD) bound to DNA reveal a cluster of seven water molecules at the protein-DNA interface of the nonspecific complex, whereas only a maximum of three or four water molecules were found at the interface of the specific complex [60, 93, 94]. In the case of EcoRI, it has been shown using "osmotic stress" analysis that EcoRI bound to a noncognate sequence sequesters ~110 waters more than when bound to the specific DNA sequence [95]. It was further shown that the dissociation rate of EcoRI-DNA-specific complex is linked to water activity [96]. Thus, the overall consensus is that water plays a major role in protein-DNA binding interaction and/or specificity.

Having discussed the studies on the role of protein dynamics and the role of water, we now turn our focus to the studies on the sequence-dependent DNA properties which are also crucial for protein-DNA interactions.

2.4 Indirect Readout Mechanisms: Sequence-dependent DNA Properties

In the last decade, advances in computational power and techniques enabled researchers to quantify the energy required for the structural adaptation of DNA through molecular simulations. Duan et al. [97] provided an estimate of +63 kcal/mol for the free energy change accompanying configurational changes in the DNA upon EcoRI binding. Subsequently, Jayaram et al. [98] made a detailed analysis evaluating the contributions of selected factors towards the energetics of EcoRI-DNA complexation. They represented the standard free energy of complexation in terms of a thermodynamic cycle of 7 distinct steps decomposed into 24 components. Their results showed the net binding energy of the complex to be a combination of several competing contributions with 10 of 24 terms favoring complexation. In addition to confirming the free energy change for the structural adaptation of DNA as +63.1 kcal/mol, their results showed that the van der Waals interactions and water release favored complexation, while electrostatic interactions were unfavorable. Sen and Nilsson [99], simultaneously, performed a 0.7 ns molecular dynamics simulation of the EcoRI-DNA complex in explicit solvent to investigate the details of interactions that are responsible for the specificity and stability of the EcoRI-DNA complex. They estimated the enthalpic part of the free energy of DNA kink formation to be approximately +31 kcal/mol.

While the above studies elucidated the energy required for deformation, it was also realized that the energy required for deforming the DNA would be different for different DNA sequences. To investigate and quantify the sequence-dependent deformability of DNA, Olson et al. [32] analyzed DNA-protein crystal complexes in the public database and extracted a set of sequence-dependent empirical energy functions from the fluctuations and correlations of structural parameters of DNA in DNA-protein crystal complexes. They found that, in general, the pyrimidine-purine (YR) dimer steps are the most flexible and the purine-pyrimidine (RY) dimer steps are the most rigid. Lankas et al. [53] performed a similar study using molecular dynamics simulations of a *free* DNA and showed that the linear correlations between adjacent basepair steps extend up to 2-3 bp, i.e., the motion of the first basepair is likely to influence the motion of the second basepair and this influence extends up to the third basepair. Subsequently, Fuji et al. [54] studied the influence of the flanking bases on the deformability of a basepair step. Their study revealed that the deformability of the AT steps are least influenced by the flanking sequences while YR steps are greatly influenced by the flanking sequences. A recent study [100] also asserts that the next-nearest-neighbor effects on sequence-dependent DNA features may not be ignored. While many of the above studies focused on basepair step deformability, few studies have also focused on the deformability at the basepair level. For example, Lankas et al. [101] studied the deformability at basepair level and observed that while buckle and propeller parameters are softer than roll, the most flexible basepair step parameter, other parameters such as opening, shear, stretch and stagger are generally comparable to or even stiffer than the basepair step parameters. This indicates that in a free DNA the basepair is more likely to be a rigid plane.

Together, these studies indicate that the deformability of DNA, and hence its deformation cost, is likely to be different for different sequences. This could be one of the factors that determine specificity in protein-DNA interactions.

The literature on protein-DNA interactions clearly shows the importance of various factors in protein-DNA recognition, including direct contacts between the protein and the DNA, the role of protein dynamics, role of water and the DNA sequence-dependent properties. In the next few chapters (Chapters 3, 4 and 5) we present our investigations and conclusions on these indirect readout mechansisms.

3 DNA SEQUENCE-DEPENDENT CHANGES IN INTRINSIC DYNAMICS OF ECORI

"*Our nature consists in motion; complete rest is death.*"

- Blaise Pascal

3.1 Introduction

From the studies described in section 2.2, one infers that the dynamics of the protein is important for its function and that the differences in the dynamics can lead to sequence discrimination in the case of protein-DNA interactions. Even though there is a plethora of studies of protein dynamics when they are present alone [102-113], studies of protein dynamics available in the context of protein-DNA interactions are few in number. Further, the available studies have focused on regulatory proteins that undergo *large* conformational rearrangements upon binding to DNA. In such cases, one expects the protein folding/unfolding dynamics upon binding to play a crucial role in how the protein and the DNA chain interact and accommodate each other. In addition, the noncognate sequences used in these studies differ by at least 6 basepairs from the cognate [70, 114].

In the present chapter, we ask if just a single basepair substitution in the DNA could alter the dynamics of the protein and, if it does, where such changes occur. For this, we choose a minimally restructuring protein (EcoRI). As discussed in section 1.5, the choice of a minimally restructuring protein allows one to isolate and examine the intrinsic dynamics of the protein, relatively unfettered and unclouded by dynamics driving unfolding and folding events. An understanding of the underlying dynamics in such cases serves as a building block for developing an overall picture of the role of dynamics in protein-DNA interactions.

EcoRI, a type II restriction endonuclease, binds to the DNA and catalyzes it at GAATTC. According to Lesser et al. [115], the next preferred sequence in the order of catalysis is TAATTC, followed by AAATTC and CAATTC. Lesser et al. [115] attribute the observed order of catalysis to the changes in the number of hydrogen bonds and appositional interactions with different substitutions. That is, there is a loss of one hydrogen bond when G in the recognition site, which has two hydrogen bonds with the protein, is replaced by T, whereas, the replacement of G with A leads to a loss of one hydrogen bond *along with appositional interactions* in the donor atoms of the protein. Replacement with C, on the other hand, results in the loss of *both* the hydrogen bonds along with appositional interactions in the donor atoms (see Fig. 4 in [115]). Thus, one can see that replacement of G with T represents the least perturbation to the protein-DNA complex, that is, a loss of just one hydrogen bond. In the present work, we ask how the dynamics of the protein would differ in such a case, i.e., when the protein shows minimal structural rearrangements and the perturbation in its substrate is the least.

In what follows, we first describe the methodologies used in this study, including setting up the system for computations, parameters used in molecular dynamics (MD) simulations, and methods of analysis of structure and dynamics. A brief discussion of the temporal variations of root mean squared displacements (RMSD) of all the atoms and the root mean squared fluctuations (RMSF) of individuals residues then follows to assess the approach to equilibrium structures and any differences in residue-level fluctuations in the structures. We then present detailed discussions of the Essential Dynamics (ED) analysis of the whole protein and some specific regions of the protein and the implications of the structural and dynamical differences between the complexes to binding and to recognition. We conclude with some remarks based on our observations.

3.2 Methods

3.2.1 System Setup and MD Simulations

The initial configurations of the protein-DNA complex were obtained from the crystallographic coordinates of 2.5 Å resolution crystal of the EcoRI-DNA complex (PDB entry 1ERI) with the DNA sequence d(CGCGAATTCGCG)$_2$ [116, 117]. Residues of Subunit I of the protein were numbered 1-261 and the residues of Subunit II of the protein were numbered 274-534. The cognate complex contains the EcoRI recognition sequence **G**AATTC, while the noncognate complex corresponds to the DNA with **T**AATTC, both with the flanking sequence mentioned in the above PDB entry. We performed the mutation at the first basepair of the recognition sequence of the DNA using Swiss PDB viewer [118]. The recognition site is divided into two half-sites, with the first half containing the sequence $\frac{\text{GAA}}{\text{CTT}}$ in the cognate complex and $\frac{\text{TAA}}{\text{ATT}}$ in the noncognate complex, respectively, and the second half containing the sequence $\frac{\text{TTC}}{\text{AAG}}$ in the cognate and the noncognate complexes. All simulations were carried out using the molecular dynamics software package GROMACS 4.0.7 [119]. Molecular interactions were represented by the parmbsc0 force field [120] for the DNA and the Amber03 force field for the protein [121], and for water the TIP3P water model [122] was used. The complex was first energy-minimized by the steepest-descent method for 1000 steps and then solvated in a 10x10x10 nm^3 cubic box. After solvation, the system was again energy-minimized using the steepest descent method for 1000 steps. The total charge of the system was –24 units and hence 24 Na$^+$ counter-ions were added to make the system electrically neutral. The ion parameters of Na$^+$ were used based on the results of Joung and Cheatham [123]. We computationally added the Mg^{2+} ion close to the catalytic site of the DNA sequence by replacing one of the water molecules. Energy minimization was done and the system was allowed to be equilibrated for 10 ns to ensure the proper positioning of the magnesium ion. Hexa-coordination of Mg^{2+} ions, as reported by Kurpiewski et al. [124], was also verified. Energy minimization was again performed prior to MD simulations. *Two independent simulations* were performed for each of the cognate and the noncognate complexes. Each simulation was done for 50 ns. Periodic boundary conditions were employed. The van der Waals and short-range electrostatic interactions were estimated within a 10 Å cutoff, whereas the long-range electrostatic interactions were assessed using the Particle Mesh Ewald (PME) method [125]. Bonds involving hydrogen were constrained using the SHAKE algorithm [126]. The total size of the system was about ~100,000 atoms. All the simulations were run in the NPT ensemble. The temperature was kept constant at T = 300K and a pressure of 1 bar.

3.2.2 Analysis of Structural Changes

Structural changes in the protein and DNA were monitored through the root mean-squared deviations of positions of the atoms. In particular, we monitored the root mean-squared displacements of *all atoms* in the two protein chains and the DNA with respect to their positions in the initial, energy-minimized solvated structure and refer to these as RMSD, as commonly done. The RMSDs are examined as a function of time for the cognate and the noncognate complexes. The root mean squared displacements for each protein residue using its constituent atoms are denoted as Root Mean Squared *Fluctuations* (RMSF) and are also examined to see if significant residue-level variations exist between the cognate and the noncognate complexes. The RMSFs are calculated relative to the equilibrium structure, which was taken to be the structure at the end of the equilibration time of the simulation (see the Discussion section below).

3.2.3 Essential Dynamics (ED) Analysis on the Protein

The ED analysis, also known as Principal Component (PC) analysis, separates the *essential or the concerted motions* from the *non-concerted or the local fluctuations*. The *concerted* motions are defined as the motions of a large number of atoms that induce global structural changes in the protein [127]. The ED analysis is a two step process, in which the first step is the fitting of atoms' trajectories to a reference frame so as to filter the translational and rotational motions and to extract only the *concerted motions*. The second step is the construction of the *3N* × *3N* covariance matrix (C) defined as

$$C_{(i,j)} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \qquad [3\text{-}1]$$

The covariance matrix represents the positional deviations of the atoms over the trajectory. The covariance matrix is then diagonalized (see Equation [3-2]) by an orthonormal transformation such that;

$$\mathrm{T}^{\mathrm{T}}\ \mathrm{C}\ \mathrm{T} = \mathrm{diag}\ (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \ldots, \lambda_{\mathrm{N}})\ ;\ \lambda_1 > \ldots > \lambda_{\mathrm{N}} \qquad [3\text{-}2]$$

where N represents the dimensions of the coordinate vector , x_i is the position of an atom "*i*" along a particular axis, $\langle\rangle$ represents the time-average of the property under investigation, and *T* is the transformation matrix. The elements of the i^{th}-column in *T* are the eigenvectors corresponding the eigenvalues, λ_i. The displacements are represented by the eigenvalues, and the direction is represented by the associated eigenvector. The greatest variance of the fluctuations occurs along the first eigenvector, with monotonically decreasing variance occurring along successive eigenvectors.

3.2.4 Porcupine Plots

Porcupine plots illustrate the motion of the residues along a particular principal component and were generated using the Dynatraj software [128]. The porcupine plots generated by the software from the trajectories from the simulations were then visualized and analyzed using VMD [129].

3.2.5 Description of DNA Structure

The structure of the DNA is described using the twelve helicoidal parameters. They are calculated using the software 3DNA [130, 131]. The helicoidal parameters are calculated for the six basepairs of the recognition sequence, i.e. GAATTC for the cognate complex and TAATTC for the noncognate complex. We define the first basepair of the recognition sequence as GC3 in the cognate complex and TA3 in the noncognate complex, the second basepair as AT4, the inner adenine as AT5, the fourth basepair as TA6, the fifth basepair as TA7 and the sixth basepair as CG8.

3.3 Results & Discussion[2]

3.3.1 Choice of Regions of the Protein for Examination

In addition to examining the entire protein for discernible changes in dynamics between the cognate and the noncognate complexes, we selected six specific regions of the protein (i.e., six sets of residues) for closer look. These regions are selected based on information available in the literature on their potential roles in the protein's function as a catalyst as described below and are indicated in Figure 3-1.

(a) **"Catalysis Region I" – Residues Asp75-Lys97 (Region R1)**: Specific residues in the region Asp75-Lys97 are known to coordinate hexavalently with the Mg^{2+} ion, and the whole region is known to be critical for the catalytic action (Kurpiewski et al. [133]).

(b) **"Crosstalk Region" (Region R2)**: It has been noted by Kurpiewski et al. [133] that four residues in the protein are involved in a "cross-talk" between the protein chains and provide mechanical strength to the reaction centers. These residues are Glu128 and Arg129 in each of the subunits. Glu128 of Subunit I is hydrogen-bonded with Arg129 of Subunit II of the protein, and similarly the Glu128 of Subunit II is coupled with Arg129 of Subunit I by a hydrogen bond. Hence, we choose the regions containing five amino acids on either side of Glu128 and Arg129 and look for a possible difference in the dynamics. Hereafter, we define this region as the "crosstalk region". The residues that lie in the crosstalk region are Ala123-Ile134 in each of the subunits of the protein.

[2] The results presented in this chapter have been published as a research article (see [132])

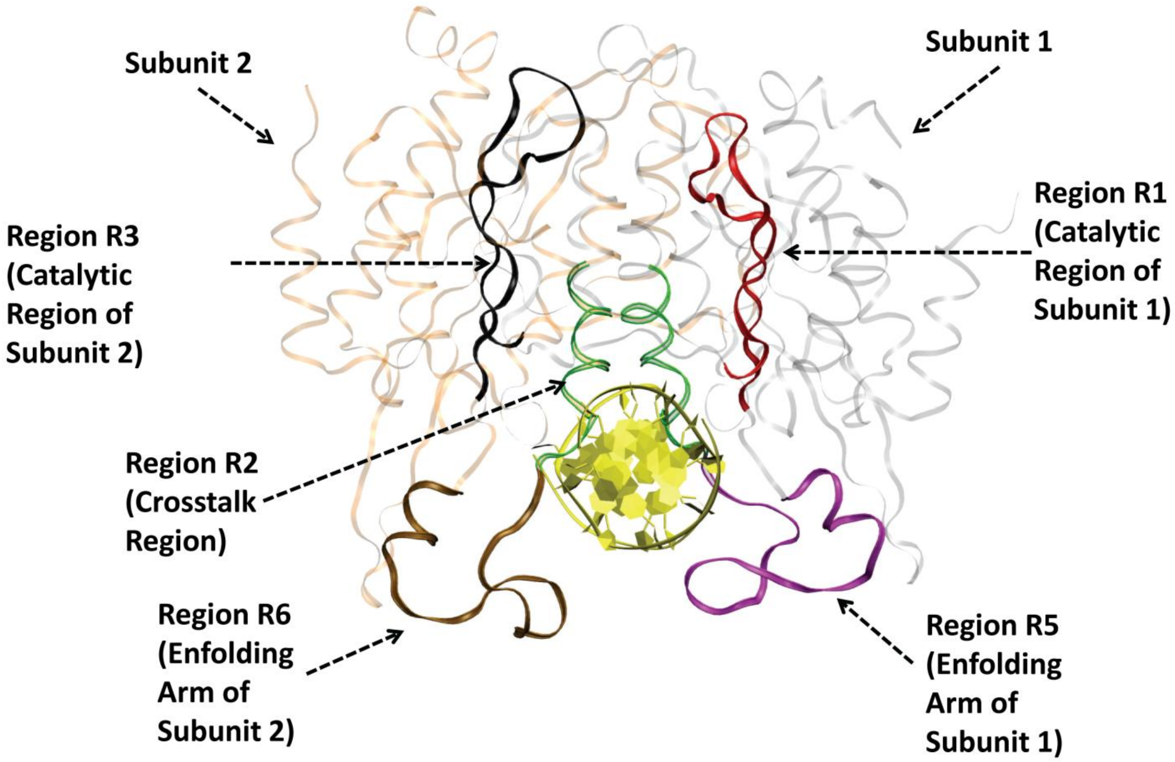

Subunit 2
Subunit 1
Region R1 (Catalytic Region of Subunit 1)
Region R3 (Catalytic Region of Subunit 2)
Region R2 (Crosstalk Region)
Region R6 (Enfolding Arm of Subunit 2)
Region R5 (Enfolding Arm of Subunit 1)

Figure 3-1. A cartoon representation of the EcoRI-DNA complex indicating the various regions chosen for analysis. Region R4 is not shown as it consists of a few unconnected residues. The residues forming Region R4 are shown in Figure 3-6C.

(c) **"Catalysis Region II" -- Region Asp348-Lys370 (Region R3)**: This region, in Subunit II, is the complement of Region R1 in Subunit I, (i.e., Asp75-Lys97, which are involved in the catalysis of the first half-site) and is involved in the catalysis of the second half-site.

(d) **"Protein/DNA Interface Region" – Region within 3.5 Å of Point of Substitution (Region R4)**: It is also instructive to follow the dynamics of regions close to the point of substitution. A distance criterion of 3.5 Å in the equilibrium structure was used for defining residues as "close" to the point of substitution.

The residues in this region are Ile72, Lys73, Asp75, Lys97, Arg129, Lys132, Asn133, and Asn398.

(e) **“Enfolding Arms” (Regions R5 and R6)**: Extending from the globular region of EcoRI are two arms that roughly encircle the DNA. These arms are thought to be essential for DNA binding through non-specific ionic contacts with the DNA phosphate [134]. It has been suggested in the literature that cleavage of the DNA requires the coordinated action of the arm of one subunit and the globular region of the other subunit [134]. We define the residues Asp102-Ala122 that constitute the arm of Subunit I as Region R5 and the residues Asp375- Ala395 that constitute the arm of Subunit II as Region R6.

3.3.2 Examination of Residue Fluctuations Resulting from Substitution

We use the all-atom RMSD of the protein (calculated relative to the starting structure) as a measure of the structural changes in the complexes as the simulation progresses. The RMSD results show that the structural fluctuations of the protein-DNA complexes with respect to the initial structure stabilize beyond about 20 ns (see Figure 3-2) and that the RMSDs remain statistically the same beyond 20 ns. Although, in the strictest sense, this does not mean that the structures have “stabilized” after 20 ns, it does indicate that the structures beyond 20 ns are sufficiently independent of the initial structure. All subsequent analyses were done on the trajectories beyond 20 ns, and the structure at this timeframe (20th ns) was chosen as the representative “equilibrium” structure. Our results on the RMSFs (see Figure 3-3) show that the variations in residue-level fluctuations between the cognate and noncognate complexes are statistically indistinguishable, thereby indicating that the single, minimal mutation introduced in the DNA does not exert a strong enough influence on the fluctuations in the residues.

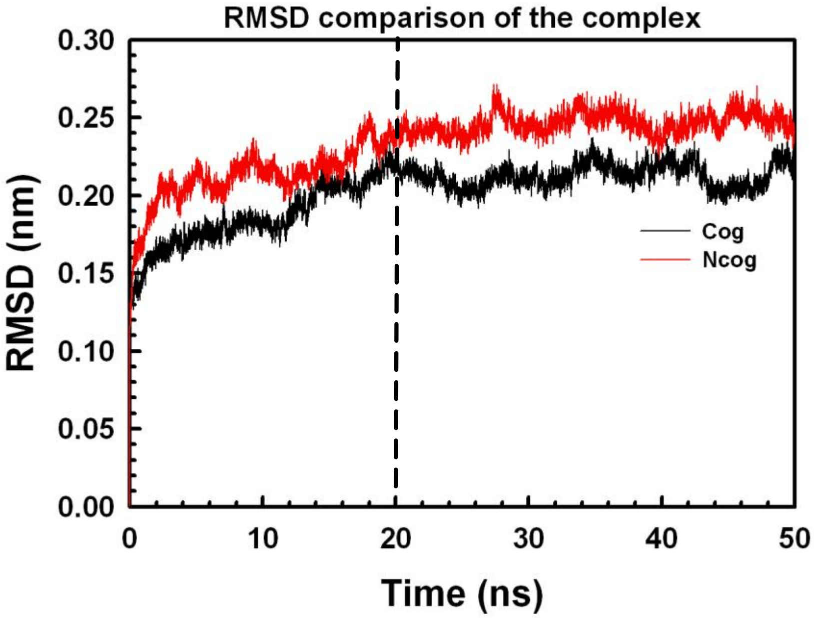


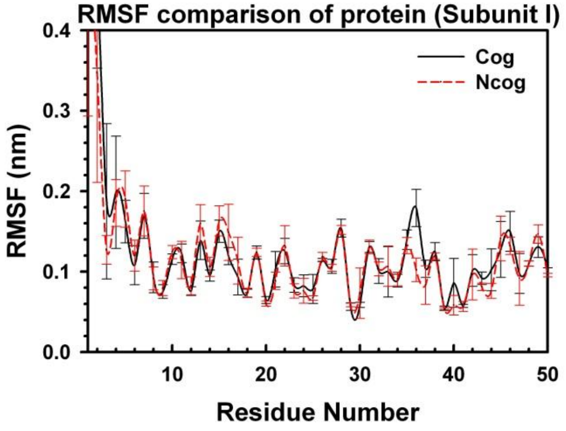
Figure 3-2: Root mean squared deviation of all atoms in the complex shows that the trajectories reach equilibrium at 20ns.

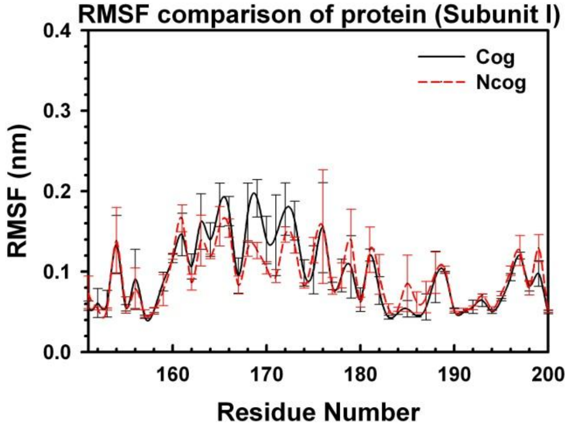
This is perhaps not surprising since EcoRI is known to display minimal restructuring on binding to the DNA, although, as we show later, the mutation does affect the grip of the protein on the DNA. The lack of differences in residue-level fluctuations does not, however, imply that the dynamics of the protein remains unaffected. We shall see in the following section that the *concerted* motions of the protein, at the backbone level, do show interesting differences, including at the protein/DNA interface region (Region R4), for which no noticeable variations are observed at the residue level.

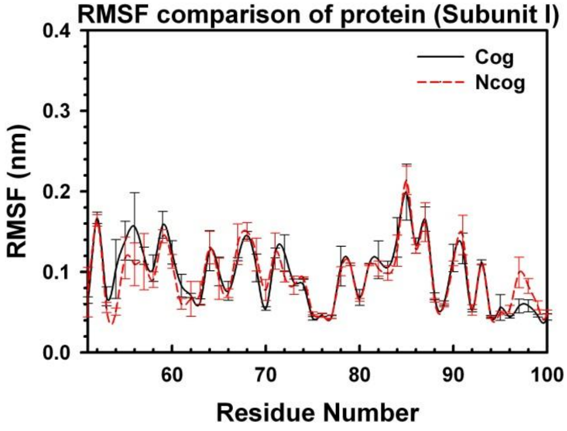
3.3.3 Altered Dynamics of the Protein

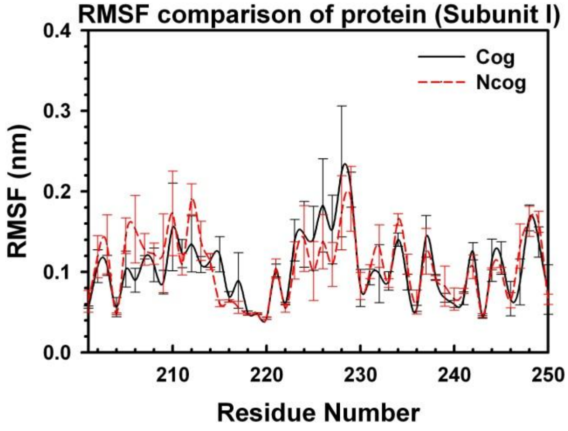
As the RMSF values represent the fluctuations of each residue taken in isolation, we use ED analyses for the regions identified above and for the whole protein to examine concerted, collective motions. The ED analysis (i.e., PC analysis) essentially serves as a filtering tool, so that molecular motions can be better appreciated. The concerted motions are characterized by the eigenvalues and the eigenvectors of the positional covariance matrix *C* in Equation [3-1] (see Methodology Section), with the eigenvalue representing the relative amount of motion along the corresponding eigenvector. The directions of motions of the residues along the principal components are represented by "porcupine plots" where the "porcupine needles" represent the direction of motion of the C_α atoms and the lengths of the needles correspond to the amplitudes of the motion.

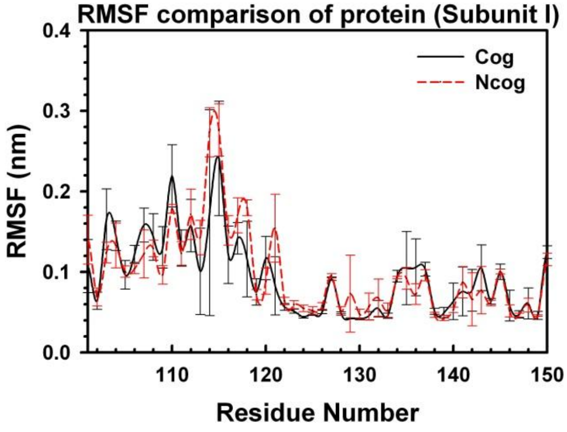
Figure 3-4 presents the porcupine plots for the full protein in the cognate complex and in the noncognate complex, in stereoscopic view, (Figure 3-5 shows the porcupine plots along with the DNA) and Figure 3-6 presents the same for the protein/DNA region (Region 4). Only the plots for the first principal component are shown, for brevity. In the case of the full protein, the first PC contributes about 25% to the motion in the case of the cognate complex, and the first four PCs account for a total of 55% of the motion. In the case of the noncognate complex, the first PC contributes about 20% and the first four collectively contribute about half of the total.

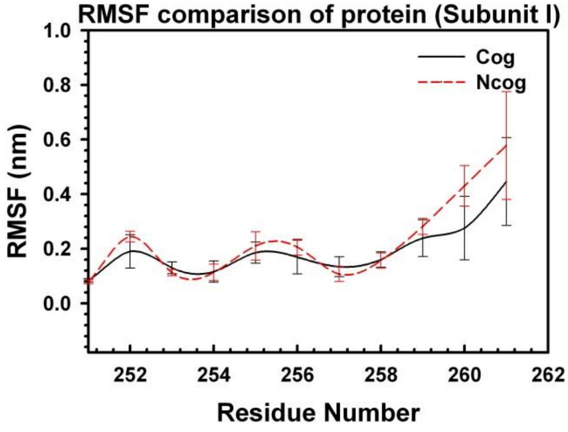

RMSF comparison of protein (Subunit I)


RMSF comparison of protein (Subunit I)


RMSF comparison of protein (Subunit I)


RMSF comparison of protein (Subunit I)


RMSF comparison of protein (Subunit I)


RMSF comparison of protein (Subunit I)

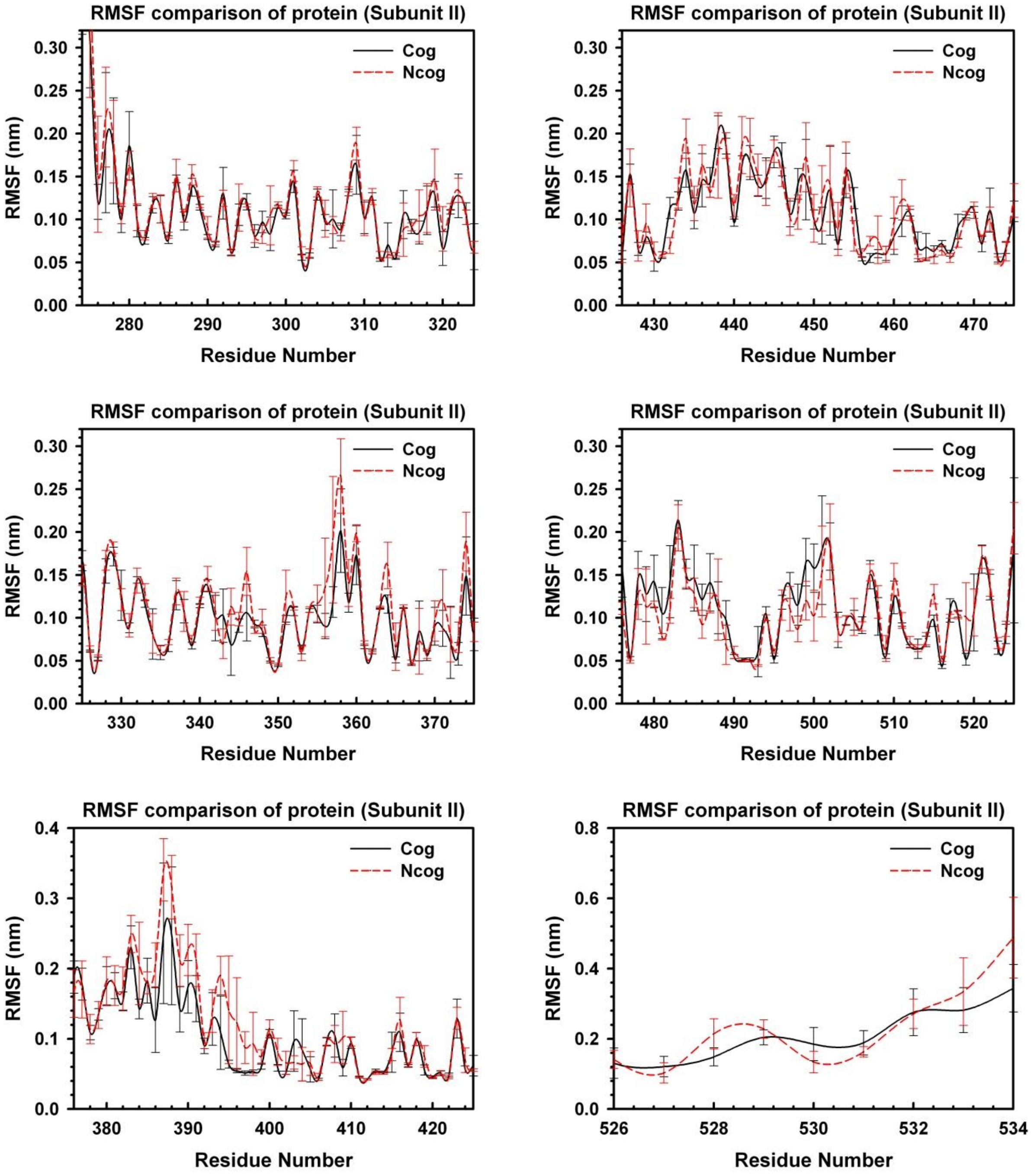


Figure 3-3. Root Mean Squared Fluctuations (RMSF) for each protein residue in the cognate and noncognate complex.

The details are given in Figure 3-7 and Figure 3-8. The general direction and characteristics of the motion in both cases do not change significantly when all the four PCs are combined, and therefore the dynamics that emerges from the first PC sufficiently captures the discussions below. (In the case of Region R4, the first PC contributes 40% to the overall motion, with the first four PCs accounting for about 70% of the motion, in the case of the cognate complex. In the other complex, the first PC contributes 25%, with the first four contributing about two-thirds.) The details on the convergence and sampling of the essential subspace, evaluated from the inner products of the eigenvectors, are given in Table 3-1. Also, shown in Figure 3-9 is a typical plot of the inner product matrix of the cognate and the noncognate complex in the essential subspace which indicates that the direction of the PCs in each of the complex is different.

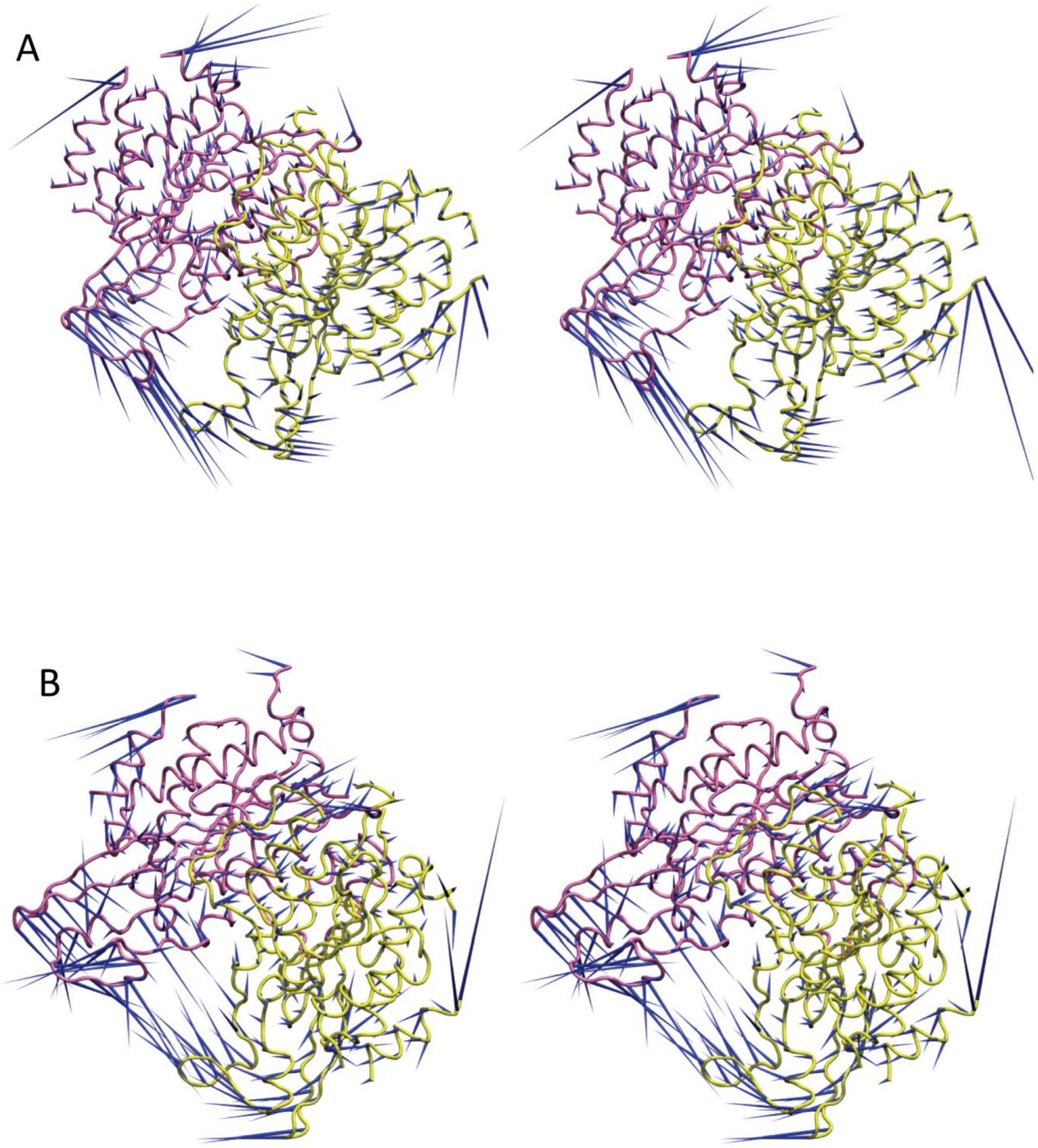


Figure 3-4. Stereo views of the porcupine plots showing the motion of the protein subunits along the first principal component in the cognate complex (A) and in the noncognate complex (B). Subunit 1 is in yellow and Subunit 2 is in mauve.

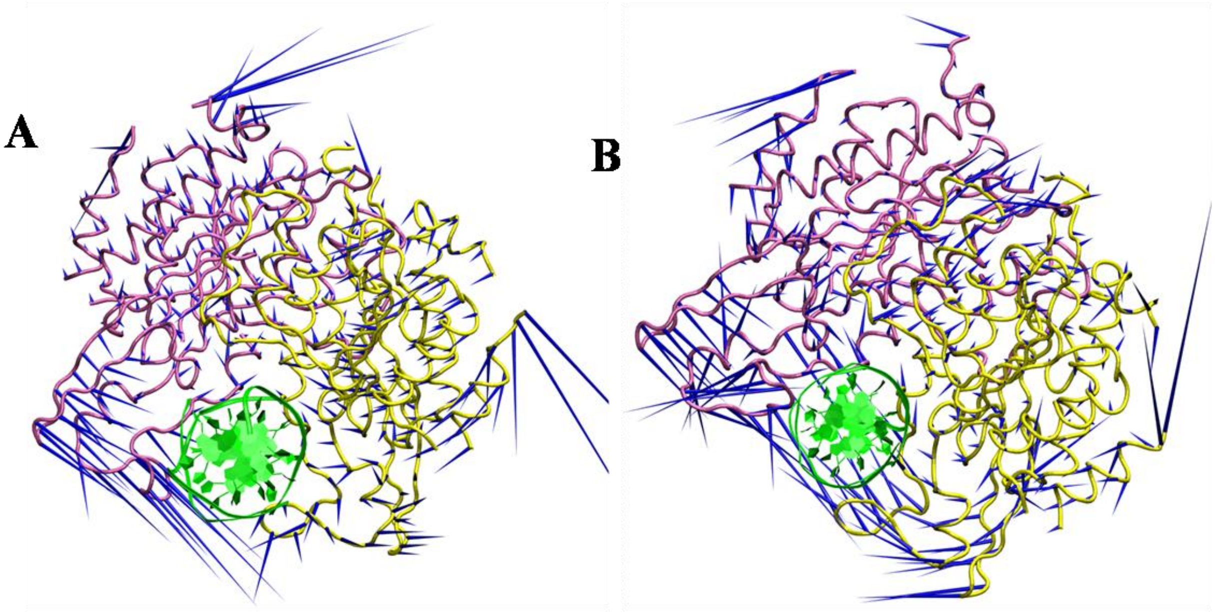
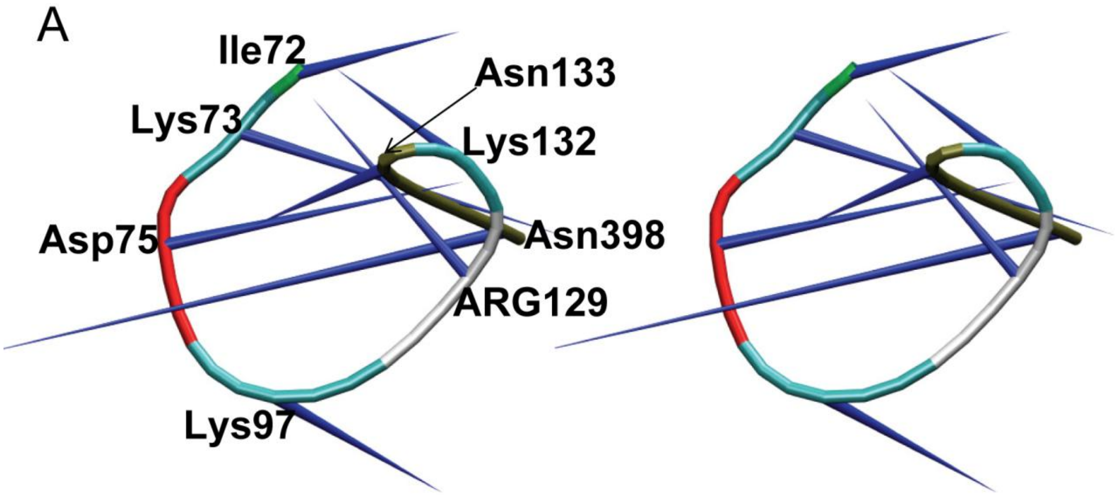


Figure 3-5. Porcupine plots with the DNA showing the motion of the protein subunits along the first principal component in the cognate complex (A) and in the noncognate complex (B). Subunit 1 is in yellow and Subunit 2 is in mauve.



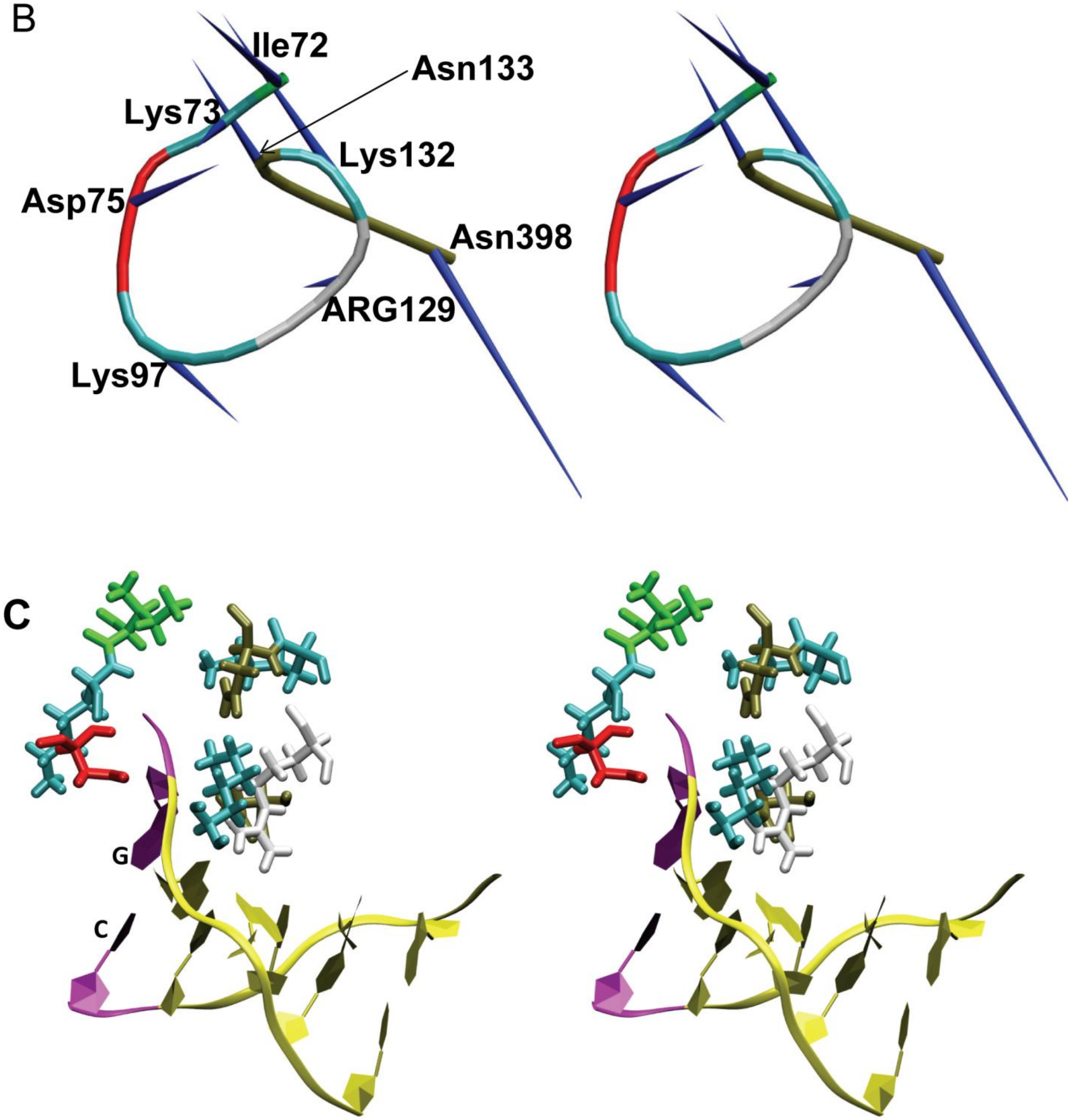


Figure 3-6. Stereo views of the porcupine plots[3] showing the motion of the residues in Region R4 along the first principal component for the cognate complex (A) and for the noncognate complex (B).

[3] The porcupine plots generated by the software show the amino acids as a string (http://s12-ap550.bioch.ox.ac.uk:8078/dynamite_html/collect_xtc_data_v1.5.html). However, the residues straddle the point of substitution (in pink) as shown in (PTO) stereo view in (**C**), and are not actually connected to each other. The residues in (**C**) are color-coded to match the colors used in the porcupine plots. The stereo view taken is from the cognate complex.

Table 3-1. The convergence of the essential subspace was evaluated by splitting the 30ns trajectory into three 10-ns blocks and calculating the eigenvectors in each case. The similarity between the eigenvectors were evaluated by the root mean squared inner product (RMSIP) values given as $RMSIP = \sqrt{\frac{1}{10}\sum_{i=1}^{i=10}\sum_{j=1}^{j=10}(u_i.v_j)^2}$ where u_i and v_j are the eigenvectors from the two different time intervals.

Time interval (ns)	RMSIP (Cognate)	RMSIP (Nocognate)
10 ns vs 10 ns intervals		
20-30 vs 30-40	0.67	0.7
20-30 vs 40-50	0.66	0.64
30-40 vs 40-50	0.70	0.69
10 ns vs 30 ns intervals		
20-30 vs 20-50	0.77	0.76
30-40 vs 20-50	0.80	0.78
40-50 vs 20-50	0.77	0.81
15 ns vs 15/30ns intervals		
20-35 vs 35-50	0.68	0.67
20-35 vs 20-50	0.84	0.82
35-50 vs 20-50	0.82	0.87

As seen from the above table, the eigen subspaces are sampled reasonably well enough as indicated by the relatively lower RMSIP values for 10 ns vs 10 ns intervals compared against the values from 10 ns vs 30 ns intervals. The higher RMSIP value between the eigenvectors estimated at a 10 ns interval and those estimated at 30 ns interval indicate the similarity/convergence of the eigen subspace.

Note: The average RMSIP value calculated for a set of random pairs of orthogonal vectors is 0.083±0.004; that is, the RMSIP values we observe for the protein dynamics are well above those one would expect for two random orthonormal sets of principal components.

It is evident from Figure 3-4 that the dynamics of the whole protein is altered in the noncognate complex as a result of the substitution in the DNA. Figure 3-4A shows that when bound to the recognition sequence the body of the protein on top of the DNA shows a coordinated twisting-type motion in both subunits, perhaps indicative of pre-catalytic posture. Further, the arm and the body of the protein twist in opposite directions in each subunit. On the other hand, even the minimal disturbance in the DNA caused by the substitution appears sufficient to initiate a scrambling of the coordinated action in the body of the protein (although some remnants of the coordination are discernible on close inspection). (See Figure 3-5 for the porcupine plots along with the DNA.) We shall return to this observation and to a discussion of Region R4 (Figure 3-6) later after an examination of the enfolding arms of the protein.

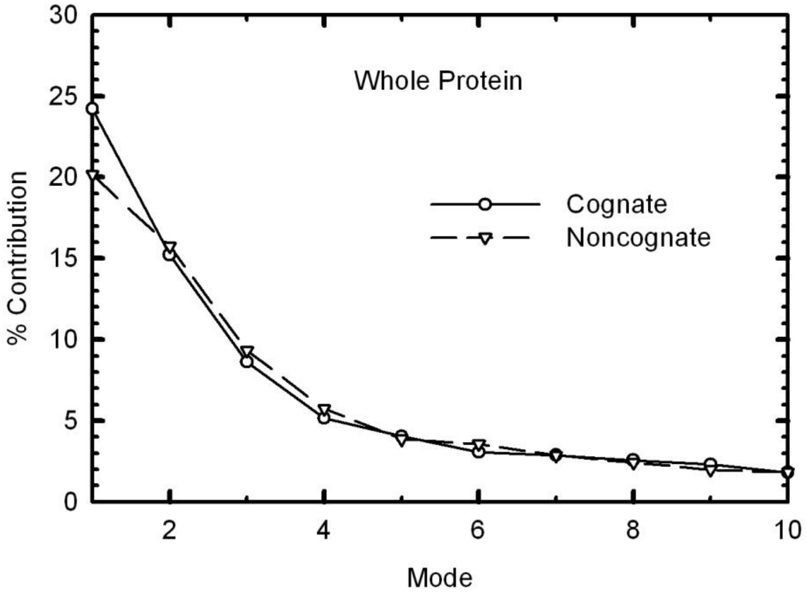


Figure 3-7. Percentage contribution of each mode toward the dynamics of the whole protein.

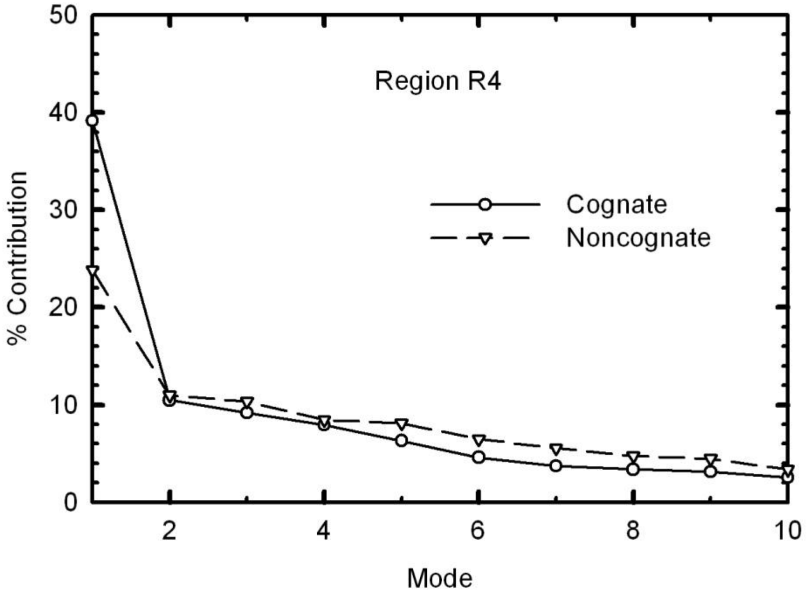


Figure 3-8. Percentage contribution of each mode toward the dynamics of interfacial residues around the point of substitution (Region R4).

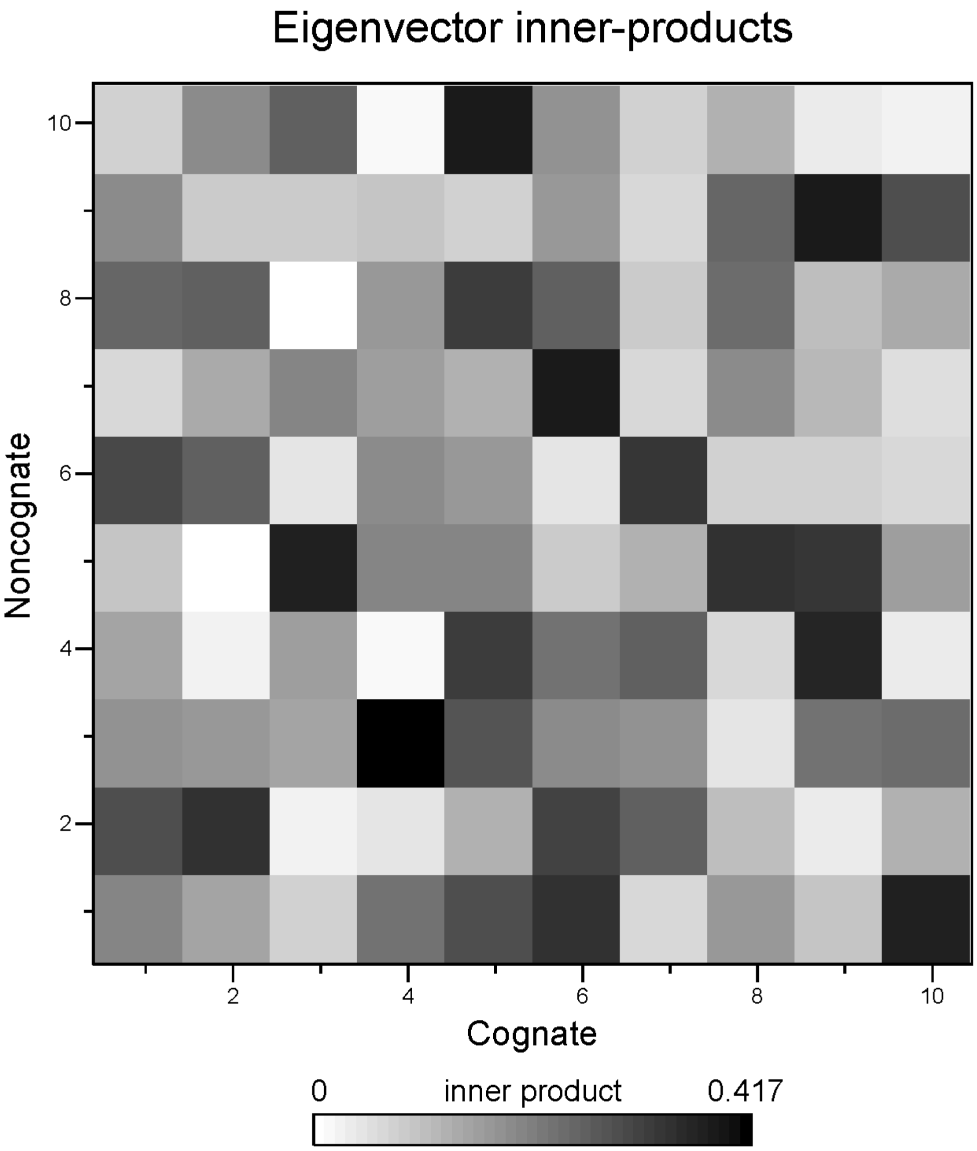


Figure 3-9. A typical plot showing the eigenvector inner products of the cognate and the noncognate complex in the essential subspace defined by the respective first 10 eigenvectors. The maximum value of the inner product is 0.417, thus clearly indicating the dissimilarity in the dynamics of the two complexes.

3.3.4 Structural Relaxation of the Arms in the Noncognate Complex

Although EcoRI restructures itself minimally when binding to a DNA, we observe that the arms of EcoRI show a significant structural relaxation when the protein is bound to a noncognate sequence (see Figure 3-10). An examination of the distances between the arms (defined, for convenience, as the distance between the centers of mass of each of the arms) shows that once the structure has relaxed the distance remains statistically constant with an average distance of 3.93 ± 0.07 nm in the noncognate complex while the corresponding average distance is 3.62 ± 0.02 nm in the cognate complex (see Figure 3-10 for a plot of the inter-arm distance as a function of time), for the timescale of analysis reported here.

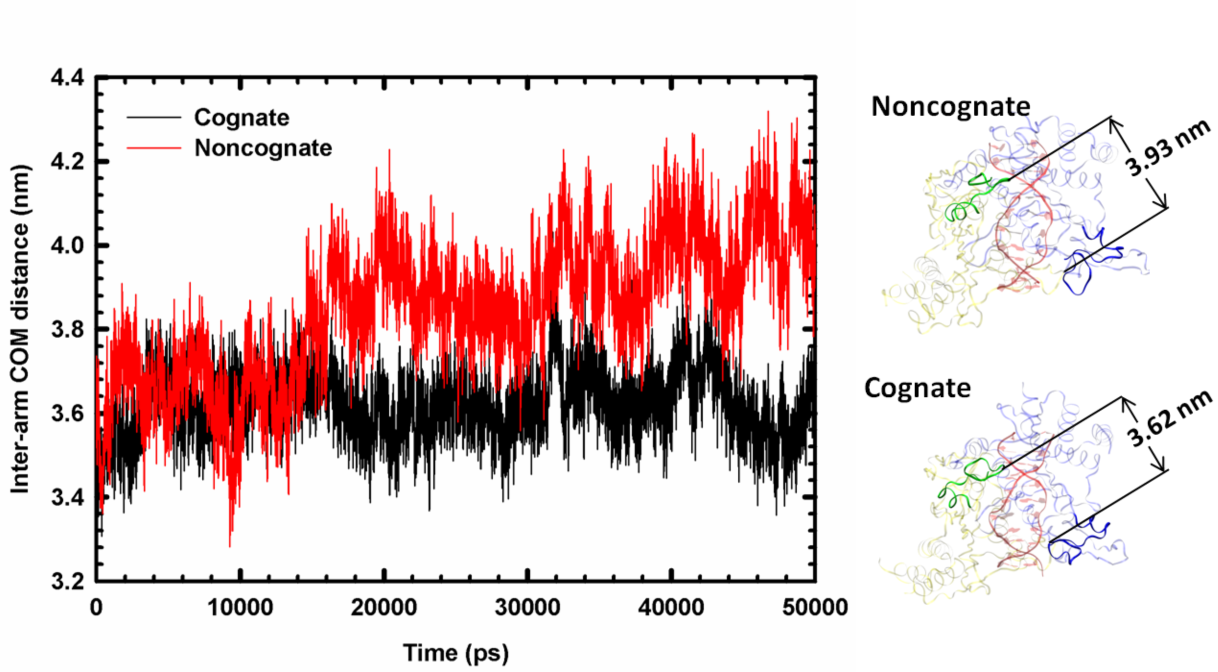


Figure 3-10. The average distances between Arm 1 (green) in Subunit 1 (yellow) and Arm 2 (blue) in Subunit 2 (blue) as a function of time in the cognate complex and in the noncognate complex.

The arms of EcoRI have been proposed, based on N-terminal deletion studies, to be essential for DNA binding and ensuring that the DNA is held in the appropriate configuration [134]. Our results reveal that when there is even a minimal change in the basepair the arms undergo structural and dynamic changes. More specifically, the arms relax and move away from the DNA, indicating that the DNA is no longer tightly bound. Later we show that this structural relaxation of the arms results in a less-kinked DNA.

3.3.5 Altered Dynamics at the Protein/DNA Interface

Although residue-level fluctuations in the various regions identified as functionally important or interesting regions of the protein remain statistically the same, as discussed earlier, variations in the essential dynamics are seen in some of the regions. The details are given in Appendix A, but we shall focus here on the protein/DNA interfacial region, namely, Region R4, consisting of residues Ile72, Lys73, Asp75, Lys97, Arg129, Lys132, Asn133, and Asn398. As mentioned earlier, the porcupine plots for this region for the cognate and the noncognate complexes are presented in Figure 3-6. Figure 3-6A shows that the dominant motion of the residues in this region, which straddle the site of substitution, in the cognate complex constrains and constricts the catalytic region of the DNA, but this coordinated motion is disrupted by the substitution, in the noncognate complex. In fact, an examination of the inner products of the eight residues in this region, with the inner products taken between the first principal vector of a C_α atom in the cognate complex with that of its counterpart in the noncognate complex, shows that the motion in the noncognate complex is roughly orthogonal to the one in the cognate complex (see Figure 3-11), indicating that the motion of the C_α atoms in the noncognate complex are almost tangential to the DNA.

In fact, not only do the interfacial residues in Region R4 show this rough orthogonality of motions between the cognate complex and the noncognate complex, but also all the residues over the entire recognition site show the same behavior (see Figure 3-11).These results show that the loosening of the enfolding arms and the attendant changes in the overall dynamics of the subunits extend to the interfacial region and further contribute to the loosening of the protein's grip on DNA even with the minimal disruption in the recognition sequence.

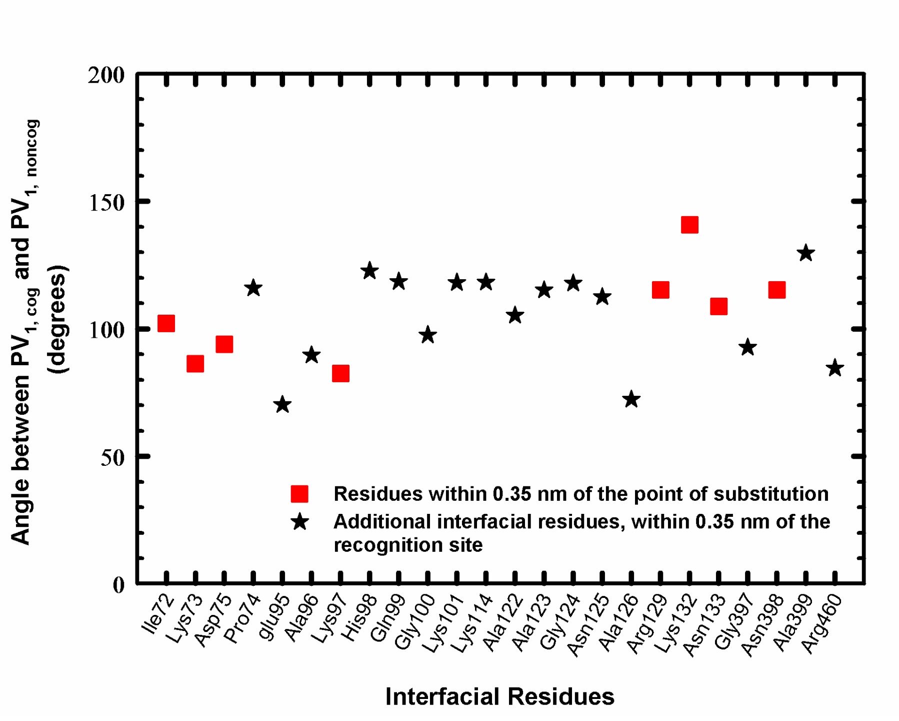
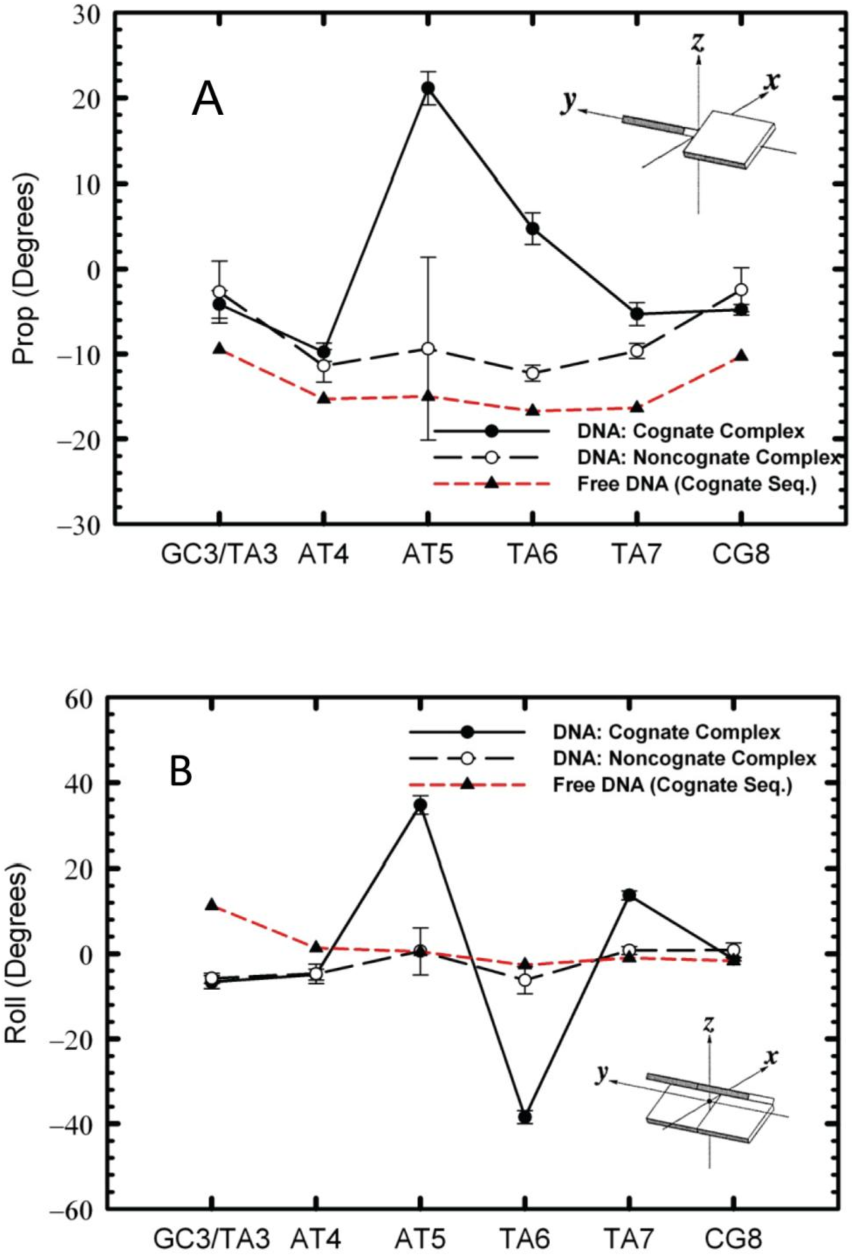


Figure 3-11. Angles between the first principal vector of the interfacial residues in the cognate ($PV_{1,cog}$) and noncognate ($PV_{1,noncog}$) complexes within 0.35 nm of the point of substitution (Region R4) and within 0.35 nm of the full recognition site.

3.3.6 Effect of Changes in Binding on the Structure of the DNA

The above observations on the loosening of the enfolding arms and the interfacial dynamics are further confirmed by the differences in the structures of the DNA between the two complexes. The conformation of a DNA sequence can be effectively described by the basepair parameters (translational: *shear*, *stretch*, *stagger*; rotational: *buckle*, *propeller*, *opening*) and the basepair step parameters (translational: *shift*, *slide*, *rise*; rotational: *tilt*, *roll*, *twist*). We present in Figure 3-12A and Figure 3-12B two of these parameters as examples, namely, propeller and roll, calculated for a free DNA (from the crystal structure of B-DNA dodecamer CGCGAATTCGCG, with the cognate sequence GAATTC; NDB id: BDL084), for the DNA in the cognate complex, and for the mutated DNA in the noncognate complex. (All the other parameters, including the ones shown in Figure 3-13.) In addition, Figure 3-12C and Figure 3-12D show pictorial representations of the DNA in the cognate and the noncognate complexes, respectively. An examination of the basepair parameters in Figure 3-12A and Figure 3-12B for the recognition site, with and without substitution, shows that the parameters in the noncognate complex are significantly different from those in the cognate complex and are, in fact, closer to those of the free DNA. In particular, one notes that AT5 and TA6, the central kinked basepairs in the cognate complex, show noticeable structural relaxations in the noncognate complex. This reduced kinking of the DNA in the noncognate complex and the fact that the overall conformation is closer to that of a free DNA (than to the one in the cognate complex) further confirm that the protein has loosened its grip on the DNA considerably even with minimal disturbance to the recognition sequence.


A
Prop (Degrees)
30
20
10
0
−10
−20
−30
z
y
x
DNA: Cognate Complex
DNA: Noncognate Complex
Free DNA (Cognate Seq.)
GC3/TA3
AT4
AT5
TA6
TA7
CG8
B
Roll (Degrees)
60
40
20
0
−20
−40
−60
DNA: Cognate Complex
DNA: Noncognate Complex
Free DNA (Cognate Seq.)
z
y
x
GC3/TA3
AT4
AT5
TA6
TA7
CG8

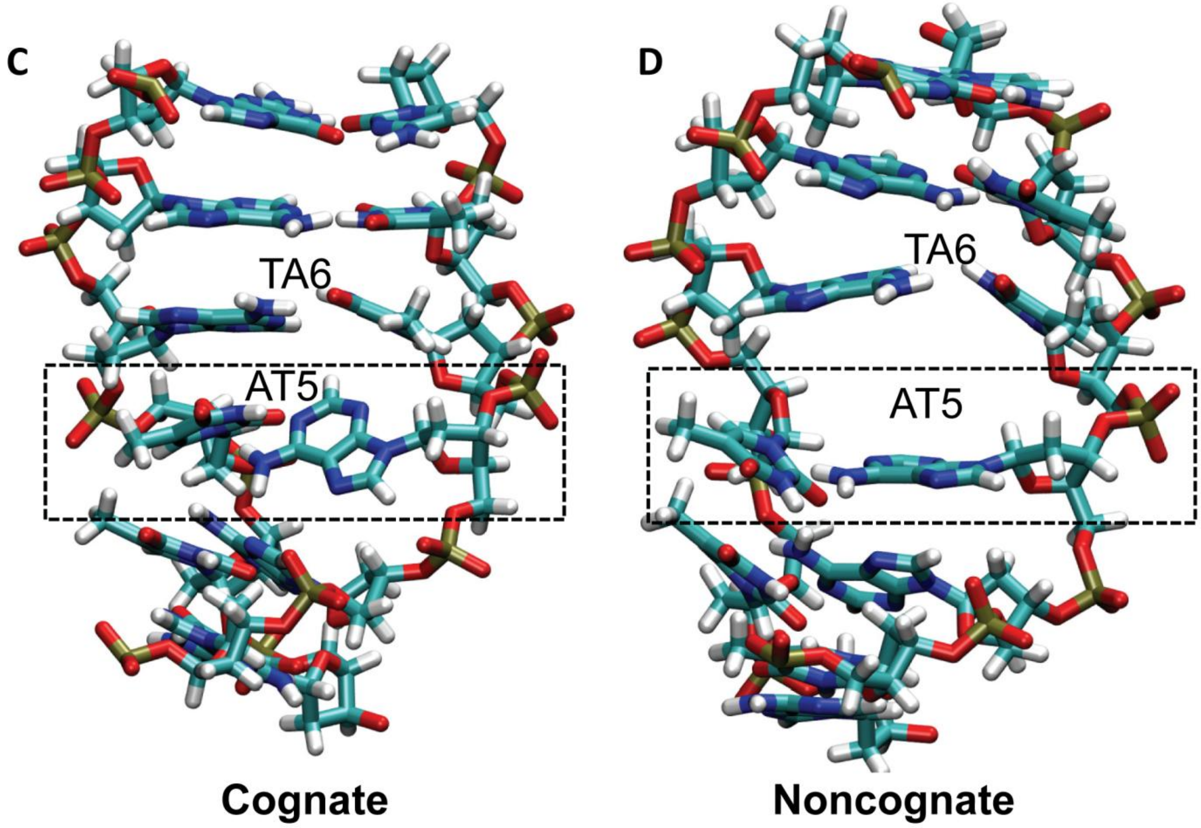
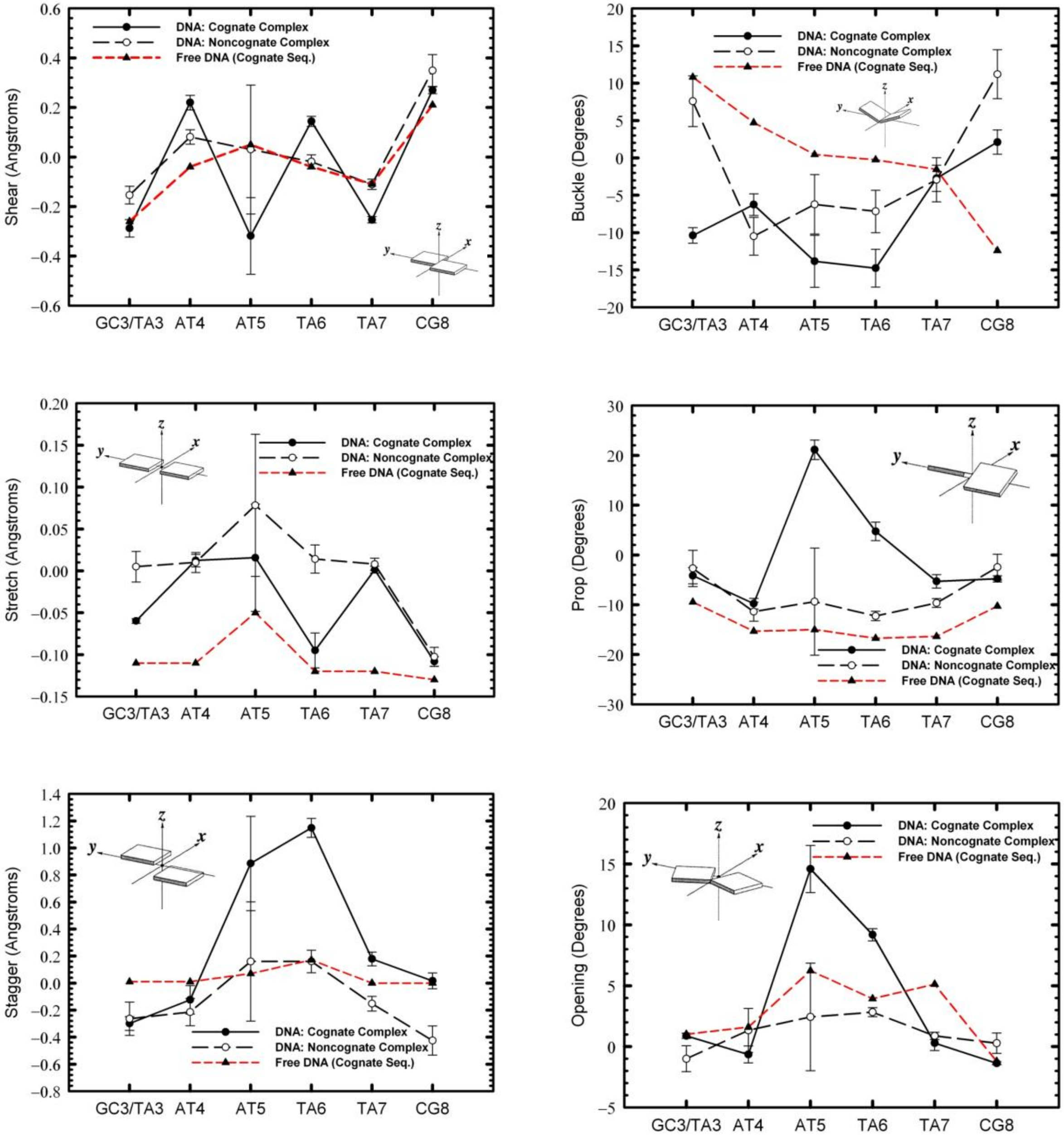


Figure 3-12. Comparison of DNA Structure. The parameters Propeller (A) and Roll (B) in the DNA of the cognate and noncognate complexes relative to those of the crystal structure of a free DNA. Figures (C) and (D) present typical snapshots of the DNA structures, showing the kinking of the central basepair in the DNA in the cognate complex (C) and the reduced kinking in the DNA in the noncognate complex (D).


Shear (Angstroms)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
DNA: Cognate Complex
DNA: Noncognate Complex
Free DNA (Cognate Seq.)
GC3/TA3
AT4
AT5
TA6
TA7
CG8
y
z
x
Buckle (Degrees)
20
15
10
5
0
−5
−10
−15
−20
Stretch (Angstroms)
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
Prop (Degrees)
30
20
10
0
−10
−20
−30
Stagger (Angstroms)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Opening (Degrees)
20
15
10
5
0
−5

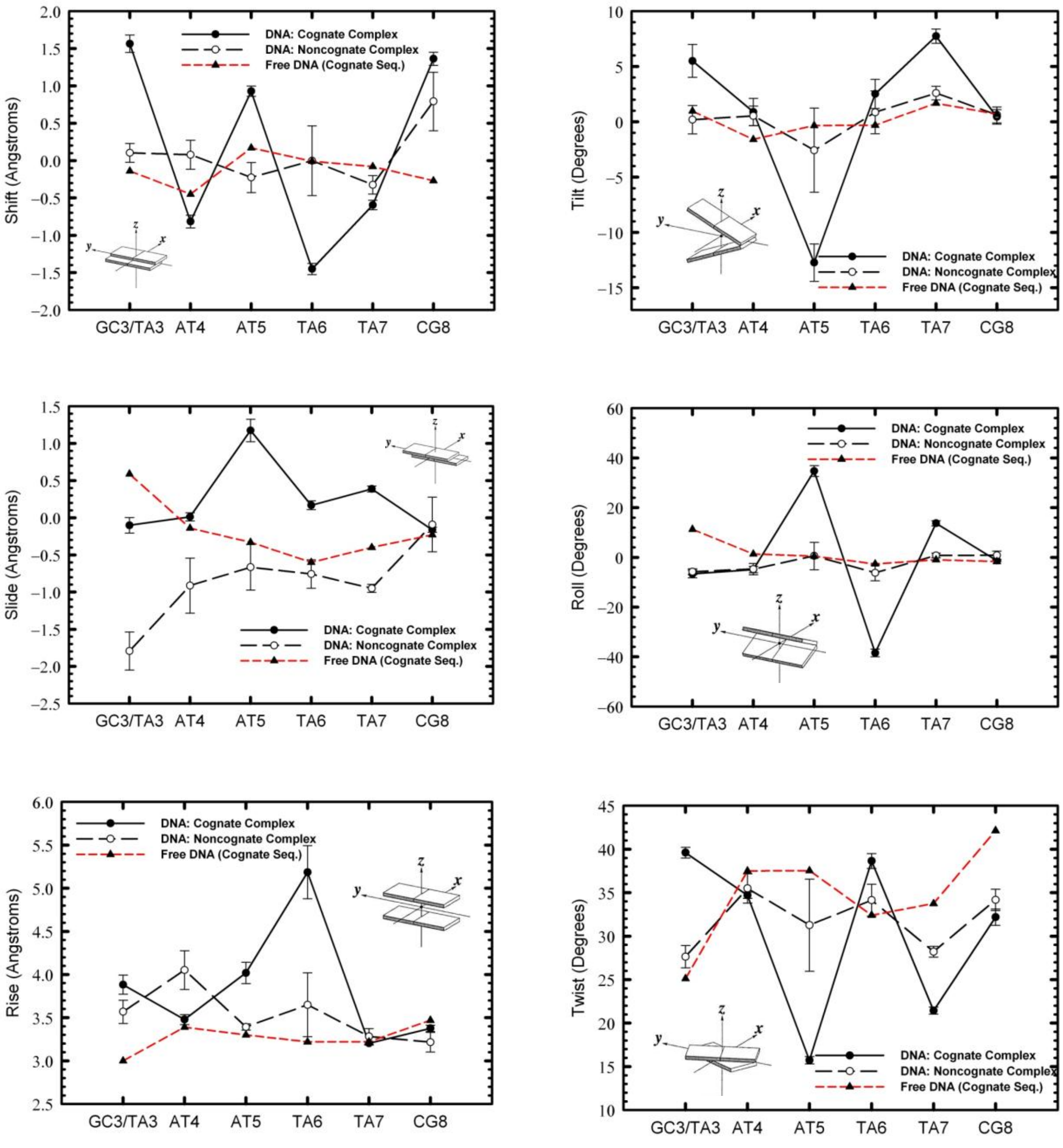


Figure 3-13. The average helecoidal parameters of the DNA in the cognate and noncognate complexes relative to those of the crystal structure of a free DNA.

3.3.7 Implications to Recognition

EcoRI has been the subject of several biochemical and biophysical studies because of interest in delineating the underlying principles of protein-DNA interactions and recognition, and previous studies have identified the residues that are important for DNA binding and/or catalysis. The present study focusing on the dynamics of the protein residues shows that the substitution of the basepairs of the DNA alters the dynamics of the residues in some of the identified regions and that the dynamics of the whole protein shows marked differences when the protein is bound to the noncognate sequence. The results show that even a minimally disrupting, single basepair substitution causes a more “open” protein structure (as revealed by the arms), scrambles a relatively coordinated dynamics in the body of the subunits, makes the DNA less kinked, and loosens the protein’s grip on the DNA. Many of the hydrogen bonds between the protein and the DNA do remain intact upon substituting a single basepair. Nevertheless, it appears that the enzyme, though attached to the DNA, is dynamically and structurally different from that in the cognate complex, and is poised for linear diffusion and further exploration. Alternatively, the results imply that when the protein chances upon the recognition sequence the dynamics of some of the key domains of the protein undergoes changes that serve as a prelude to eventual catalysis.

3.4 Concluding Remarks

Protein-DNA binding is a complex phenomenon brought about by a myriad of factors acting in unison. Experimental evidence has established that a protein generally diffuses linearly along the DNA before it chances upon the cognate sequence. Structural characterization of a protein bound to its cognate and to noncognate DNA sequences have revealed that the protein, in general, shows remarkably different conformation in the two cases. This leads one to suspect that the dynamics of the protein must also be different in the two cases. NMR studies have indeed indicated different dynamics in the protein depending on whether it is bound to the cognate or the noncognate DNA. However, most of these studies have been performed on noncognate sequences that are different at least by 6 basepairs. In this study, we asked if the dynamics of the protein would be sensitive to even the most minimal perturbation in the protein-DNA complex. Our study reveals that even such small perturbations can lead to altered dynamics of the protein. Thus, it is no surprise that regulatory proteins that undergo large structural changes upon DNA binding fail to achieve the required conformation when bound to the noncognate sequence. The difference in the DNA sequence is enough to alter the dynamics in the protein sufficient to make it unfit for the required function. The present study also implies that systematic investigation of the effects of mutations in a protein/DNA complex on protein dynamics could shed light on the machinations behind protein/DNA recognition.

Having investigated the changes in protein dynamics when EcoRI is bound to its noncognate DNA sequence, in the next chapter (chapter 4), we discuss the changes in the dynamics of water in the hydration layer of EcoRI when it is bound to its noncognate DNA sequence.

4 DYNAMICS AND THERMODYNAMICS OF WATER IN ECORI–DNA INTERACTIONS

"In the world there is nothing more submissive and weak than water. Yet for attacking that which is hard and strong nothing can surpass it."

- Lao-Tzu

4.1 Introduction

As discussed in section 2.3, studies on protein-DNA complexes have largely focused on delineating the number of waters at the protein-DNA interface/complex either via structural studies or the osmotic stress method upon protein-DNA and the static roles of water. Protein-water hydrogen bond dynamics has been shown to be important for protein structural relaxation [135]. Recently Grossman et al.[136] showed, in addition to the correlation between kinetics and retarded water dynamics at the active site of a metalloprotease, that the dynamics of water around a specifically bound enzyme-peptide complex is different from a nonspecifically bound complex. Taken together, these studies indicate that protein function and structural relaxation (or dynamics) is tightly coupled with water dynamics. Alternatively, one can also argue that by controlling the interfacial water dynamics, one can alter protein function. Thus, in the context of protein-DNA complex, in addition to quantifying the number of water molecules at the protein-DNA interface, it is also essential to characterize the dynamics (and thermodynamics) of the interfacial waters to establish a functional relation. Recently, Sinha et al.[137] have identified that water molecules between the binding motifs of the protein and the DNA exhibit restricted dynamics due to more frequent reformation of water-water hydrogen bonds. Collectively, all the above studies indicate that the binding of protein to a DNA is associated with changes in dynamics of water molecules at the interface. However, it is not clear if these associated changes in the dynamics of water molecules influence specificity in protein-DNA interactions. Thus the overall focus is to delineate the role of water in the specific binding of protein to the DNA. Specifically, here, we study the dynamics and thermodynamics of water molecules in the different regions around a protein bound to its specific sequence (cognate complex) and a nonspecific sequence (noncognate complex).

We choose the minimally restructuring EcoRI bound to its cognate or a noncognate DNA as our model. The choice of a minimally restructuring protein stems from our interest in developing cases which can eventually serve as building blocks to develop an overall picture of the role of water in protein-DNA interactions and protein-DNA interaction in general [138]. To investigate the binding of EcoRI to a noncognate sequence, we choose the noncognate sequence TAATTC since, as described in Chapter 3, the TAATTC sequence represents the minimal perturbation to the DNA from the cognate sequence (GAATTC). In the present work, we ask how the water dynamics and thermodynamics would differ when the protein shows minimal structural rearrangements and the perturbation in its substrate is the least. We use molecular dynamics simulations and the recently-developed “two-phase thermodynamic (2PT) theory” to estimate the dynamic and thermodynamic properties of water.

In the next section, we outline details on system set-up, MD simulations and other analytic tools used. We also briefly outline the two-phase thermodynamic scheme (commonly called the 2PT theory) for calculating entropy from MD trajectories. This is followed by a discussion of the hydration which is higher for the noncognate complex.

We then show that water around the noncognate complex, particularly in the intercalating region, has a faster dynamics than those in the cognate complex. This is followed by a discussion of the water-protein/DNA and water-water hydrogen-bond and dynamics which indicates relatively long-lived hydrogen bonds in the cognate complex. The thermodynamic properties of water in the defined regions are then discussed. We end with a few concluding remarks on the implications of our results for protein-DNA interactions.

4.2 Methods

4.2.1 System Set-Up and MD Simulations

The system set-up and MD simulation protocols are the same as that described in Chapter 3 except for an additional 100 ps simulation with the trajectory written every 4 fs. For ease of reference, here we briefly outline the methodology. The initial configurations of the protein-DNA complex were taken from the crystallographic coordinates of the EcoRI-DNA complex (PDB entry 1ERI) with the DNA sequence d(CGCGAATTCGCG)$_2$ [116]. The noncognate EcoRI-DNA complex was obtained by mutating the first basepair in the recognition sequence (i.e., G in GAATTC to T) using Swiss PDB viewer [139]. All simulations were carried out using the molecular dynamics software package GROMACS 4.0.7 [119]. Molecular interactions were represented by the parmbsc0 force field [120] for the DNA and the Amber03 [121] force field for the protein; for water the TIP3P water model [122] was used. Each of the system was first energy-minimized by the steepest-descent method for 1000 steps and then solvated in a 10x10x10 nm^3 cubic box. After solvation, the system was again energy-minimized using the steepest descent method for 1000 steps. Counter-ions were added to make the system electrically neutral. The ion parameters of Na$^+$ were used based on the results of Joung and Cheatham [123]. Standard periodic boundary conditions were employed to avoid boundary effects. The van der Waals and short-range electrostatic interactions were estimated within a 10 Å cutoff, whereas the long-range electrostatic interactions were assessed using the Particle Mesh Ewald (PME) method [125]. Bonds involving hydrogen were constrained using the SHAKE algorithm [126]. The total size of the system was about ~100,000 atoms. All the simulations were run in the NPT ensemble. The temperature was kept constant at T = 300K and the pressure at 1 bar. The EcoRI-DNA complexes were simulated for 50 ns. Following this, each of the system was further simulated for 100 ps and the trajectory was written every 4 fs. The 100ps trajectory was used for the further analyses.

4.2.2 Orientational Dynamics of Water

The rotational dynamics of water is investigated by following the reorientational dynamics of its dipole moment vector $\boldsymbol{\mu}$, defined as the vector connecting the oxygen atom of water to the center of the line connecting to the two hydrogen atoms. The angular reorientation of this vector is given by [140]

$$C_l^{\mu} = \frac{\left\langle P_l[\mathbf{e}^{\boldsymbol{\mu}}(t).\mathbf{e}^{\boldsymbol{\mu}}(0)]\right\rangle}{\left\langle P_l[\mathbf{e}^{\boldsymbol{\mu}}(0).\mathbf{e}^{\boldsymbol{\mu}}(0)]\right\rangle} \quad [4\text{-}1]$$

where $\mathbf{e}^{\boldsymbol{\mu}}(t)$ is the unit vector along the dipole moment vector at time t and P_l refers to the l-th Legendre polynomial. The angular brackets denote ensemble averaging. The reorientational dynamics of water in the particular region (interface or the intercalating region) is calculated using those water molecules that reside in that region *throughout* the 100 ps simulation.

4.2.3 Hydrogen-bond Dynamics of Water

The dynamics of water can also be examined using the changes in the hydrogen bonds a water molecule makes with other atoms. Two atoms are considered to form a hydrogen bond if the acceptor-donor distance is < 3.5 Å and the hydrogen-donor-acceptor angle is < 30°. The fluctuations in the hydrogen bond population as a function of time can be characterized by the correlation function $C_{HB}(t)$ as [141, 142]

$$C_{HB}(t) = \frac{\langle h(0)h(t) \rangle}{\langle h(0)h(0) \rangle} \qquad [4\text{-}2]$$

where $h(t)$ is 1 if a hydrogen bond is intact at time *t* and 0 otherwise. The angular brackets denote ensemble averaging. The hydrogen bond lifetime correlation function ($C_{HB}(t)$ in Eq. 4-2) gives the probability that a pair of hydrogen bond that is hydrogen-bonded at time zero is still hydrogen bonded at time *t*, irrespective of whether the bond was present or absent in the intermediate times. Thus the decay of $C_{HB}(t)$, beyond an initial transient period where the decay of $C_{HB}(t)$ is determined by water rotation/libration, is determined by the rearrangement of the hydrogen bond network [135].

4.2.4 2PT Theory for Calculating Thermodynamic properties from MD Trajectories

The 2PT theoretical scheme developed by Lin et al.[143] assumes that the thermodynamic properties of a system (here, water) at liquid-like densities can be obtained as the sum of the corresponding thermodynamic properties at gas-like and solid-like densities. This assumption enables one to account for the fluidicity effects of the liquid-like state. The thermodynamic properties at gas-like and solid-like densities are calculated from the corresponding density of states, which, in turn are calculated from the velocity autocorrelation functions. The readers are referred to Appendix B for a detailed discussion on these. Here, for ease of reference, we present a brief outline of the 2PT approach.

The density of states of a system $g(\upsilon)$ is given as the Fourier transform of the velocity autocorrelation function:

$$g(\upsilon) = \frac{2}{kT}\lim_{\tau\to\infty}\int_{-\tau}^{\tau} C(t)e^{-i2\pi\upsilon t}dt \qquad [4\text{-}3]$$

where $C(t)$ is the mass-weighted translational velocity autocorrelation function or the moment-of-inertia-weighted angular velocity autocorrelation function (see [85, 143, 144]), *k* is the Boltzmann constant, and *T* is the absolute temperature. In the 2PT model, the density of state $g(\upsilon)$ of a system with *3N* degrees of freedom is assumed to be partitioned into a gas-like component $g^{g}(\upsilon)$ and solid-like component $g^{s}(\upsilon)$, i.e.,

$$g(\upsilon) = g^{g}(\upsilon) + g^{s}(\upsilon) \qquad [4\text{-}4]$$

A thermodynamic property P of a system can then be determined by weighting the individual components as follows:

$$P = \int_0^\infty g^s(\upsilon) W_P^{HO}(\upsilon) d\upsilon + \int_0^\infty g^g(\upsilon) W_P^g(\upsilon) d\upsilon \qquad [4\text{-}5]$$

where $W_P^{HO}(\upsilon)$ is the weighting function for the solid phase based on the harmonic oscillator model and $W_P^g(\upsilon)$ is the weighting function corresponding to the choice of gas component. The gas-like component can be taken to be a hard-sphere fluid, for which the density of states can be written as [143]

$$g^g(\upsilon) = g^{HS}(\upsilon) = \frac{g_0}{1 + \left[\frac{\pi g_0 \upsilon}{6 fN}\right]^2} \qquad [4\text{-}6]$$

where g_0 is equal to $g(\upsilon = 0)$, f is the fluidicity factor and N is the number of molecules. The factor f is a measure of the "fluidicity" of the system and indicates the departure of the state of system from the two extremes, namely, the gas-like and solid-like states. Thus, f needs to satisfy two conditions: (i) At high temperatures and in the low-density limit, the system behaves like a gas, here taken to be a hard-sphere gas. Hence, f should be equal to 1. (ii) At the high-density limit, the system becomes a solid, and hence, $f = 0$. Therefore, f determines the apportioning of the chosen property of the liquid phase in terms of the corresponding values for the solid state and the gas state. One can write f as [143]:

$$f = \frac{D(T,\rho)}{D_0^{HS}(T,\rho;\sigma^{HS})} \qquad [4\text{-}7]$$

which satisfies the above two conditions. In the above equation, $D(T,\rho)$ is the self-diffusion coefficient of the molecules and is obtained from the zero-frequency intensity of density of states as:

$$D_{Trans} = \frac{kTg(\upsilon = 0)}{12mN} \qquad [4\text{-}8]$$

for translational diffusion (where *m* is the mass of the water molecule) and

$$D_{Rot} = \frac{kT\sum_{j=1}^{3} g^{j}(\upsilon = 0)}{4N\sum_{j=1}^{3} I_j} \qquad [4\text{-}9]$$

for rotational diffusion (where I_j is the moment of inertia along the *j*th principal axes). The denominator in Eq. [4-7] is the hard-sphere diffusion coefficient in the zero-pressure limit.

Lin et al. [143] developed a universal equation for f which bypasses the need for estimating D_0^{HS} (and hence σ_{HS}), and the equation is given as

$$2\Delta^{-9/2} f^{15/2} - 6\Delta^{-3} f^{5} - \Delta^{-3/2} y^{7/2} + 6\Delta^{-3/2} f^{5/2} + 2f - 2 = 0 \qquad [4\text{-}10]$$

where Δ , the normalized diffusivity, is a function of the material properties and is given as

$$\Delta(T,\rho,m,g_0) = \frac{2g_0}{9N}\left(\frac{\pi kT}{m}\right)^{1/2}\rho^{1/3}\left(\frac{6}{\pi}\right)^{2/3} \quad [4\text{-}11]$$

Thus, once $g_0(=g(\upsilon=0))$ and f are determined, one can determine g^g and $g^s(g^s = g - g^g)$.

Once the individual components of the density of states are determined, one can use Equation [4-5] to obtain the thermodynamic properties. The quantum weighting functions in Equation [4-5] for the solid-like component is given as follows.

$$W_E^Q(\upsilon) = \frac{\beta h\upsilon}{2} + \frac{\beta h\upsilon}{\exp(\beta h\upsilon)-1} \quad [4\text{-}12]$$

$$W_S^Q(\upsilon) = \frac{\beta h\upsilon}{\exp(\beta h\upsilon)-1} - \ln[1-\exp(-\beta h\upsilon)] \quad [4\text{-}13]$$

$$W_A^Q(\upsilon) = \ln\frac{1-\exp(\beta h\upsilon)}{\exp(-\frac{\beta h\upsilon}{2})} \quad [4\text{-}14]$$

The quantum weighting functions for the gas-like component are given as:

$$W_E^g(\upsilon) = W_E^{HS}(\upsilon) = 0.5 \quad [4\text{-}15]$$

$$W_S^g(\upsilon) = W_S^{HS}(\upsilon) = \frac{1}{3}\frac{S^{HS}}{k} \quad [4\text{-}16]$$

$$W_A^g(\upsilon) = W_A^{HS}(\upsilon) = W_E^{HS}(\upsilon) - W_S^{HS}(\upsilon) \quad [4\text{-}17]$$

The energy *E*, entropy *S*, and Helmholtz free energy *A* for a canonical ensemble can then be determined as

$$E = V_0 + \beta^{-1} \int_0^\infty g(\upsilon) W_E^Q(\upsilon) d\upsilon \qquad [4\text{-}18]$$

$$S = k \int_0^\infty g(\upsilon) W_s^Q(\upsilon) d\upsilon \qquad [4\text{-}19]$$

$$A = V_0 + \beta^{-1} \int_0^\infty g(\upsilon) W_A^Q(\upsilon) d\upsilon \qquad [4\text{-}20]$$

4.3 Results & Discussion

4.3.1 Cognate Complex is Less Hydrated

We begin our analysis with an examination of the number of water molecules associated with each of the protein-DNA complexes, as this number is known to be a measure of the "closeness or directness" of the contacts between the protein and the DNA surfaces [95]. We define two kinds of water molecules, namely, intercalating and interfacial, associated with the complex. *Intercalating water molecules* are defined as those water molecules that reside at a distance less than the thickness of the first hydration shell from *both* the protein and the DNA. *Interfacial water molecules* are defined as those molecules that reside in the first hydration shell of *either* the protein or the DNA. The first hydration shell of the protein-DNA complex extends until 4 Å (see Figure 4-1 for a typical plot of the average number of water molecules around the protein-DNA complex as a function of distance from any atom in the complex). Hence, those water molecules that reside within 4 Å of the protein *or* the DNA are called the interfacial water molecules and those water molecules that are present at a distance less than 4 Å from *both* the protein and the DNA are called the intercalating waters.

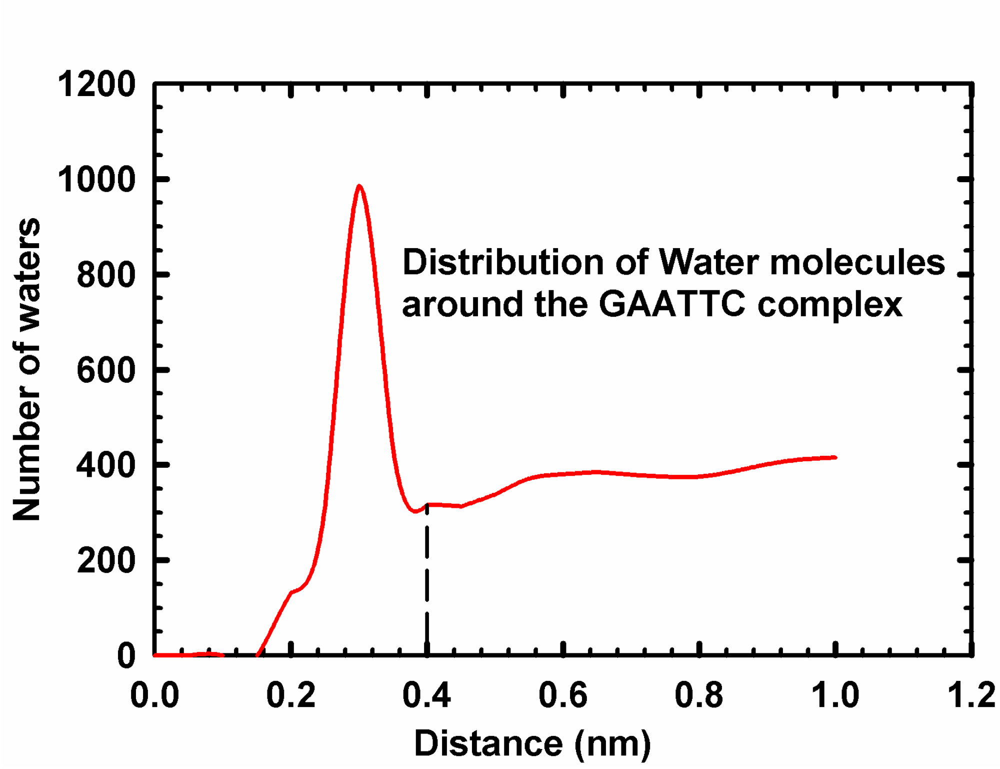


Figure 4-1. Distribution of water molecules around the GAATTC complex indicates that the first hydration shell is about 0.4 nm.

Figure 4-2 shows a pictorial representation of water in the interfacial and intercalating regions. The average numbers of interfacial water molecules in the two protein-DNA complexes differ significantly from each other (see Table 4-1), with ~147 more interfacial waters in the noncognate complex. Moreover, the number of water molecules that reside in the interfacial region *throughout* the 100 ps simulation time is also different, with 45 more interfacial water molecules in the noncognate complex than the cognate complex (Table 4-2). The number of intercalating waters that reside throughout the 100 ps simulation time, however, does not show any difference.

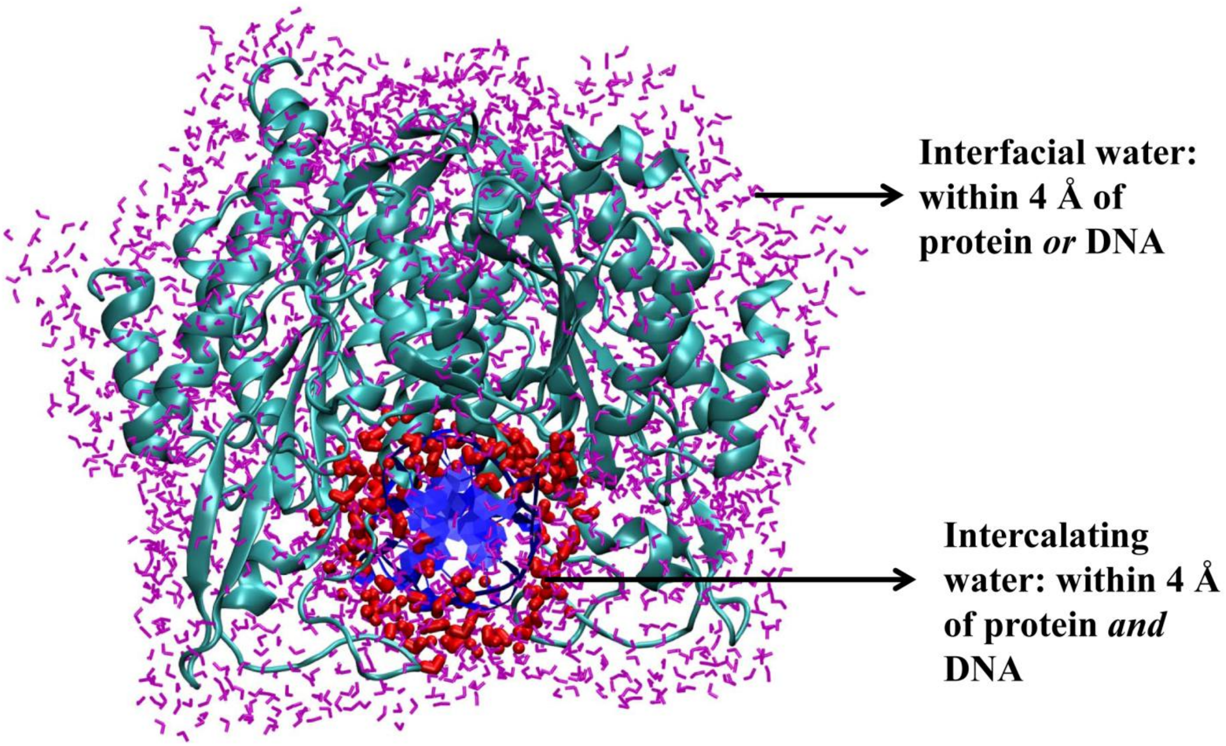

Interfacial water: within 4 Å of protein *or* DNA
Intercalating water: within 4 Å of protein *and* DNA

Figure 4-2. A snapshot of the cognate complex showing the intercalating waters (red) and the interfacial waters (magenta). Protein is shown in cyan and the DNA is shown in blue.

Table 4-1. Average number of waters in the interface and intercalating regions (calculated over 100 ps).

	Intercalating	**Interfacial**
GAATTC complex	141 ± 6	2366 ± 17
TAATTC Complex	157 ± 7	2513 ± 21

Our results are consistent with the experimental observations [145] which show that ~ 70 more waters are associated with the TAATTC noncognate complex at ~0˚C and low osmotic pressures.

Table 4-2. Number of water molecules present throughout the 100 ps simulation in the two regions.

	Intercalating	**Interfacial**
GAATTC complex	33	275
TAATTC Complex	33	320

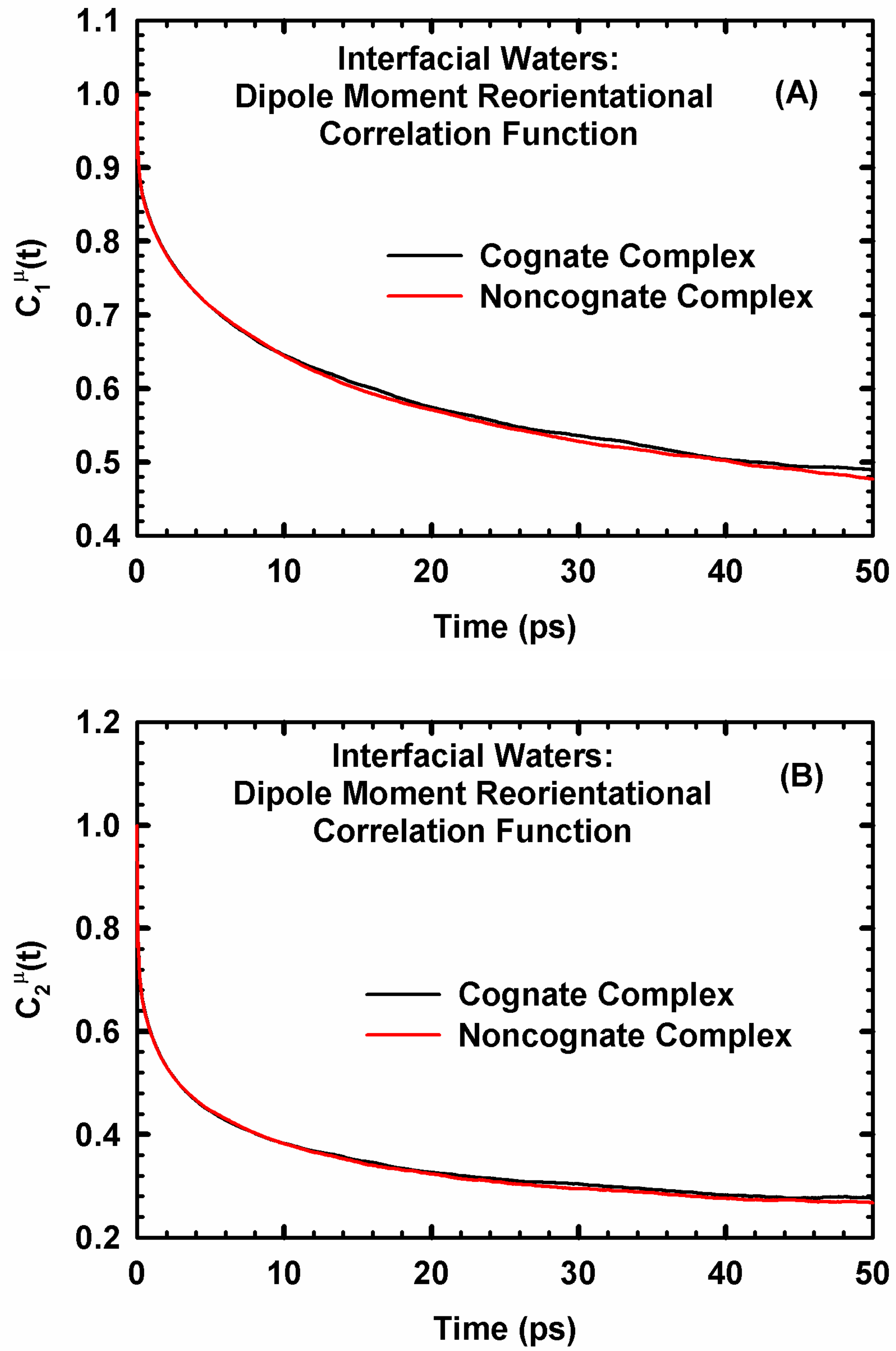
In essence, our results, which indicate that the noncognate complex is much more hydrated than the cognate complex, suggests that protein and the DNA surfaces in the noncognate complex are not as close to each other as in the cognate complex. This "looseness" of the surfaces of the protein and the DNA is also substantiated by the fact that the DNA in the noncognate complex is unkinked [138]. In the next section, we show that the "looseness" of the surfaces also leads to a faster dynamics of the associated water molecules in the noncognate complex.

4.3.2 Intercalating Waters Reorients Faster in the Noncognate Complex

We now turn our attention to the dynamics of the water molecules in the interfacial and the intercalating regions, described here in terms of the rotational and translational motions of the molecules. For this, we focus on those water molecules that reside in the interface or the intercalating region *throughout* the 100 ps simulation time. The rotational dynamics of the molecules is evaluated by their dipole moment reorientational correlation function (Eq. 1). In Figure 4-3 we show the first- and second-rank dipole moment correlation functions of the interfacial and intercalating waters. Since it has been suggested that water reorientation occurs at three characteristic timescales, the fastest corresponding to libration motions, the intermediate timescale corresponding to the restricted motions of the dipole moment vector within a cone of semiangle θ (angular motions) and the slowest corresponding to the overall rotation of the vector without any restriction (tumbling motion) [146, 147], we fit the correlation functions to the triple exponential function

$$C_l^{\mu}(t) = A_0 \exp(-t/\tau_0^l) + A_1 \exp(-t/\tau_1^l) + A_2 \exp(-t/\tau_2^l) + A_3 \qquad [4\text{-}21]$$

where the constant A_3 denotes net polarization [146]. The amplitudes and relaxation times obtained from the data are given in Table 4-3 along with the amplitude-weighted average relaxation time (i.e., $\sum A_i \tau_i$). The amplitudes and relaxation times of the bulk waters are also presented as reference. While a full functional form in Equation [4-21] describes the correlation function of the interfacial waters well, for the intercalating region a statistically better result is obtained for the triple-exponential with A_3 =0. From Figure 4-3 and Table 4-3 one sees that the correlation functions of interfacial waters in the cognate and the noncognate complex are essentially same. When compared to the bulk waters, the second and third relaxation time constants are higher for the interfacial waters. The second and third relaxation time constants correspond to the angular vibration of the dipole vector within a cone of semiangle θ and the overall tumbling of the molecule. The results imply that the interactions of the water with the protein/DNA surface dampen the angular vibrations and tumbling of the interfacial water molecules. In the intercalating region, the waters in the cognate complex have a significantly longer orientational relaxation time than that in the noncognate complex.


Interfacial Waters:
Dipole Moment Reorientational
Correlation Function
(A)
$C_1^{\mu}(t)$
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Cognate Complex
Noncognate Complex
0
10
20
30
40
50
Time (ps)
Interfacial Waters:
Dipole Moment Reorientational
Correlation Function
(B)
$C_2^{\mu}(t)$
1.2
1.0
0.8
0.6
0.4
0.2
Cognate Complex
Noncognate Complex
0
10
20
30
40
50
Time (ps)

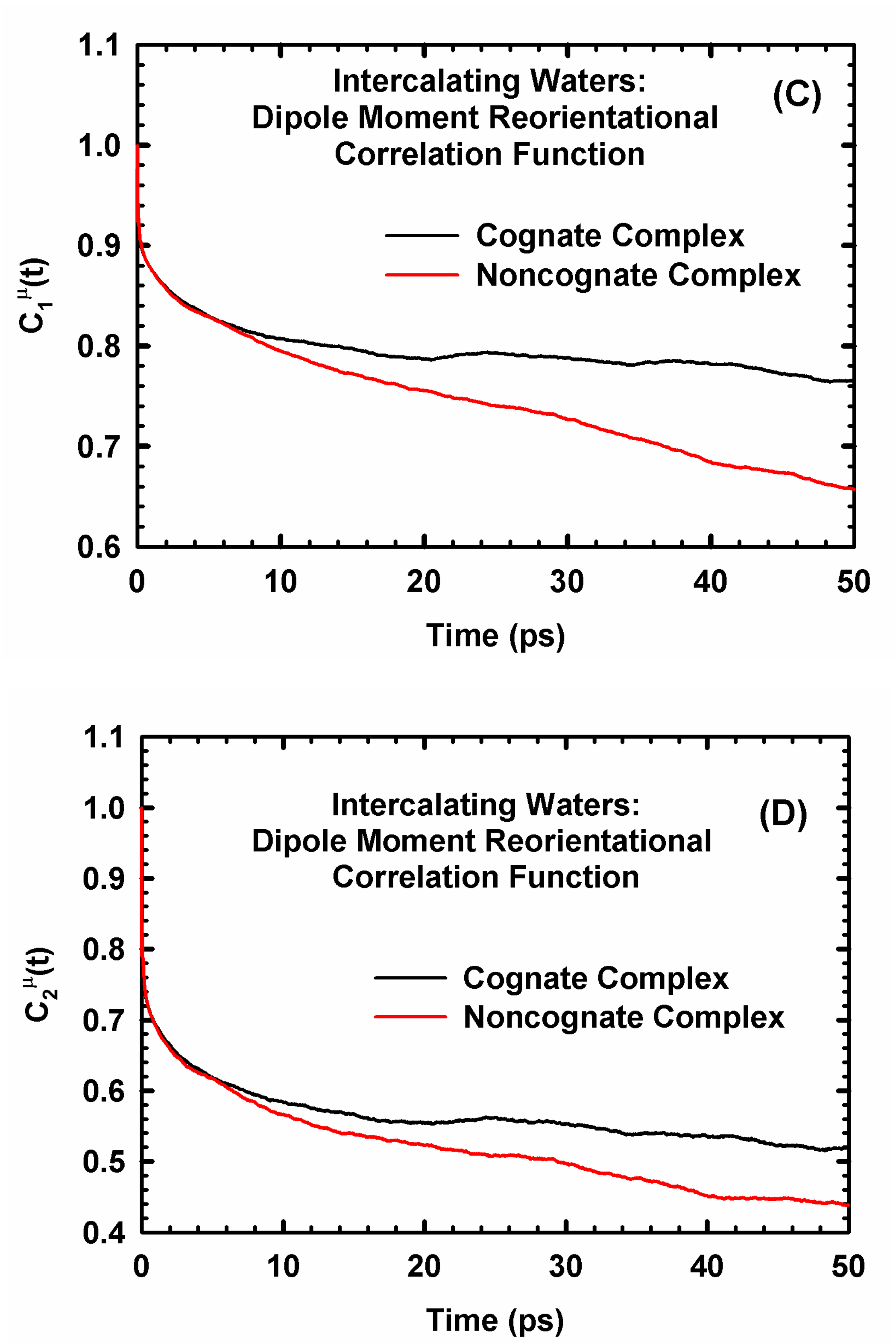


Figure 4-3. First- and second-rank dipole moment reorientation correlation function for interfacial (A and B) and intercalating waters (C and D).

Particularly, the tumbling motion (the slowest of the three relaxation mechanisms) is significantly slower in the cognate complex. The amplitude-weighted average relaxation time of the intercalating waters in the cognate complex for $l = 1$ is ~5 times that of the noncognate complex and for $l = 2$ is ~2.6 times of that in the non cognate complex. In summary, the results indicate that the rotational dynamics of water molecules around the two complexes are different from that of bulk waters and the intercalating waters in the cognate complex have a significantly arrested motion compared to those in the noncognate complex. We have also evaluated the translational dynamics by the root mean-squared displacement of the oxygen atoms in the water molecules as a function of time (see Figure 4-4 and Table 4-4). The results show that both interfacial and intercalating waters show retarded, sublinear diffusion (in contrast to bulk water, which shows linear diffusion). However, there is no significant difference between the cognate and noncognate complexes for both the interfacial and intercalating waters. Since we have used those water molecules that stay in the interfacial or the intercalating region for the entire time period of analysis, understandably, the translational dynamics is not different between the two complexes. In essence, our results indicate that while the translational motions and rotational motions of the water in the interfacial and intercalating regions are relatively ‘arrested’ compared to the bulk waters due to the interactions with the protein/DNA surfaces, the intercalating water molecules in the noncognate complex are freer to rotate than those in the cognate complex. As we show in the next section, the free rotation of the intercalating waters in the noncognate complex alters the hydrogen-bond dynamics of these molecules, thus bringing about a “dynamic” region between the surfaces of the protein and the DNA.

Table 4-3. Amplitudes and relaxation time constants for the reorientational correlation function.

	Order	Cognate			Noncognate		
		Amplitude	**Time (ps)**	**Average (ps)**	**Amplitude**	**Time (ps)**	**Average (ps)**
Interfacial waters	1	0.49 0.12 0.13 0.26	-- 0.08 2.9 24.4	6.7	0.44 0.13 0.14 0.28	-- 0.11 3.9 29.7	8.9
	2	0.30 0.29 0.21 0.20	-- 0.05 1.9 16	3.7	0.28 0.29 0.19 0.24	-- 0.05 1.7 16	4.2
Intercalating Waters	1	0.10 0.09 0.81	0.05 3.8 1088	881.7	0.10 0.07 0.83	0.07 3.2 215	178.1
	2	0.26 0.15 0.59	0.05 3.5 409.0	241.8	0.27 0.14 0.59	0.06 3.7 159.0	94.3
Bulk Waters	1	0.11 0.35 0.54	0.04 1.5 3.4	2.4	--		
	2	0.21 0.27 0.52	0.01 0.3 1.2	0.7			

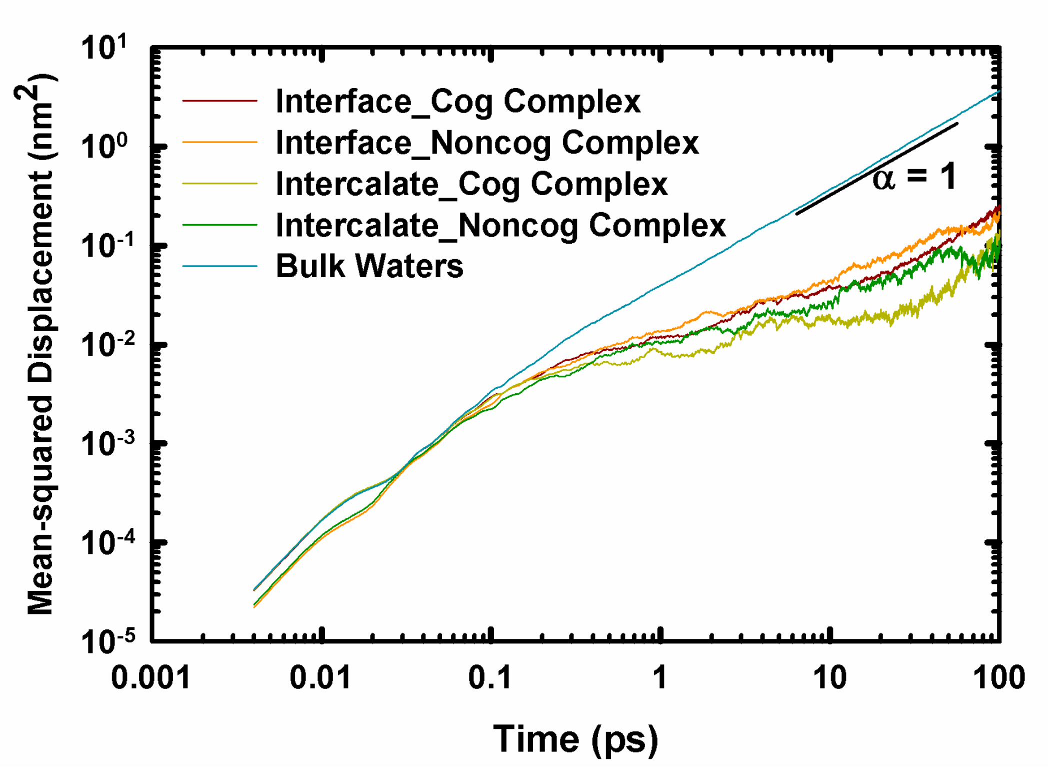


Figure 4-4. Mean-squared displacement of water molecules in the interfacial and intercalating regions.

Table 4-4. Comparison of the exponent α (from mean-squared displacement of water molecules as a function of time) in the interface and the intercalating regions of the cognate and noncognate complexes show the sublinear diffusion in these regions.

		A
Interface	Cognate	0.68 ± 0.01
	Noncognate	0.60 ± 0.01
Intercalate	Cognate	0.55 ± 0.03
	Noncognate	0.48 ± 0.11
Bulk		0.99 ± 0.01

4.3.3 Short-lived Water-Protein/DNA Hydrogen Bonds in the Noncognate Complex

In this section we discuss the dynamics of the hydrogen bonds of interfacial and intercalating waters with the protein or the DNA. As discussed in the Methods section, the hydrogen bond lifetime correlation function ($C_{HB}(t)$ in Equation [4-2]) gives the probability that a hydrogen bond with a given pair of atoms at time zero also exists at time t, irrespective of whether the bond was intact in the intermediate time. In Figure 4-5 we present the lifetime correlation function of the hydrogen bonds of the interfacial and intercalating waters with the protein or the DNA in the cognate and noncognate complexes. It is known (see, for example, Laage et al. [148]) that the dynamics of the hydrogen-bond network of bulk water involves three timescales corresponding to the times taken for (a) the initial breaking of a hydrogen bond, (b) stable rearrangement of the hydrogen-bond network and (c) the diffusion of the hydrogen bonds. Therefore, we use a triple-exponential function

$$C_{HB}(t) = A_0 \exp(-t/\tau_0) + A_1 \exp(-t/\tau_1) + A_2 \exp(-t/\tau_2) \qquad [4\text{-}22]$$

to describe the correlation function $C_{HB}(t)$; the resulting amplitudes and relaxation times are presented in Table 4-5.

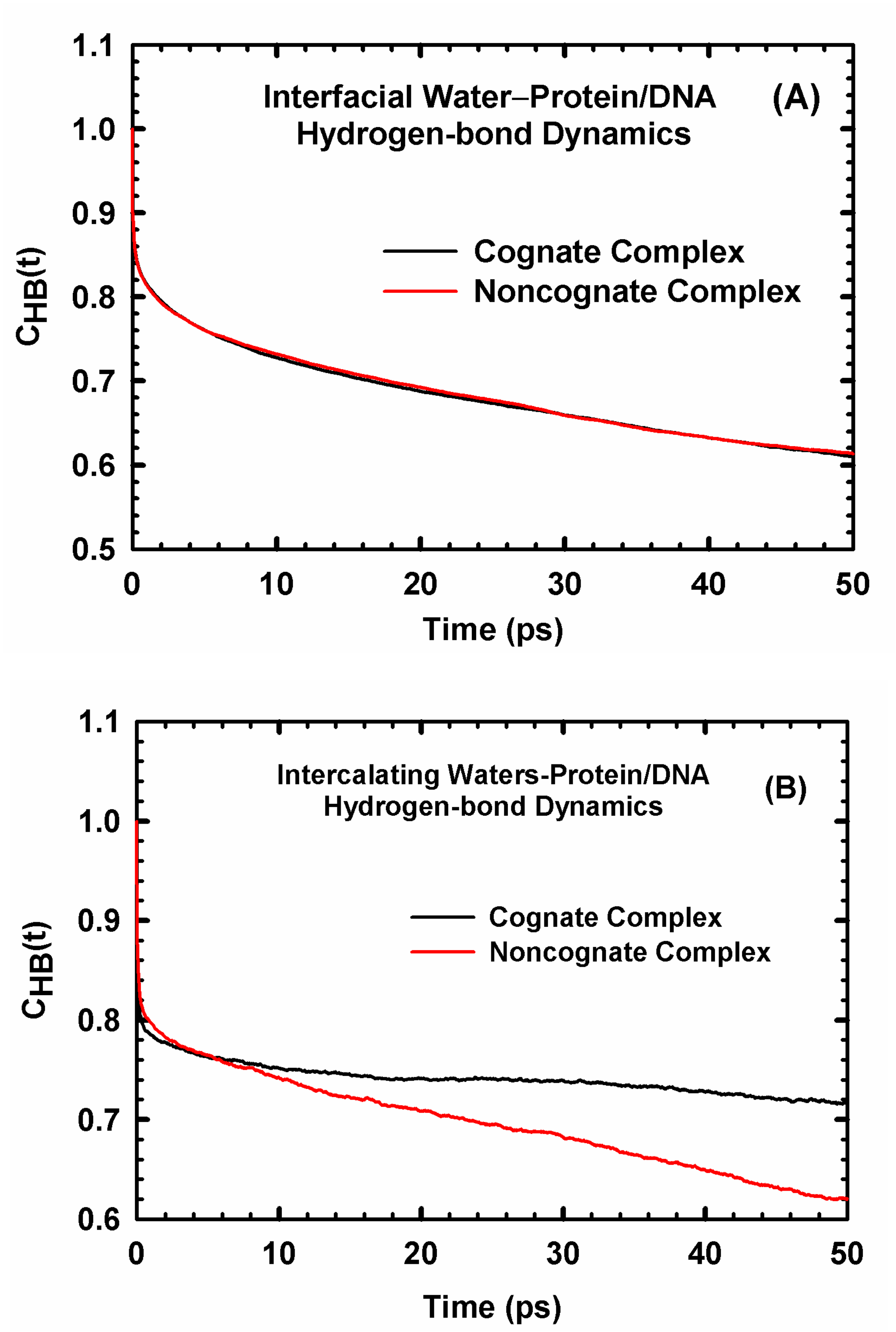


Figure 4-5. Water-protein/DNA hydrogen bond lifetime correlation function of interfacial waters (A) and intercalating (B) waters with the protein or the DNA in the cognate (black) and noncognate (red) complex.

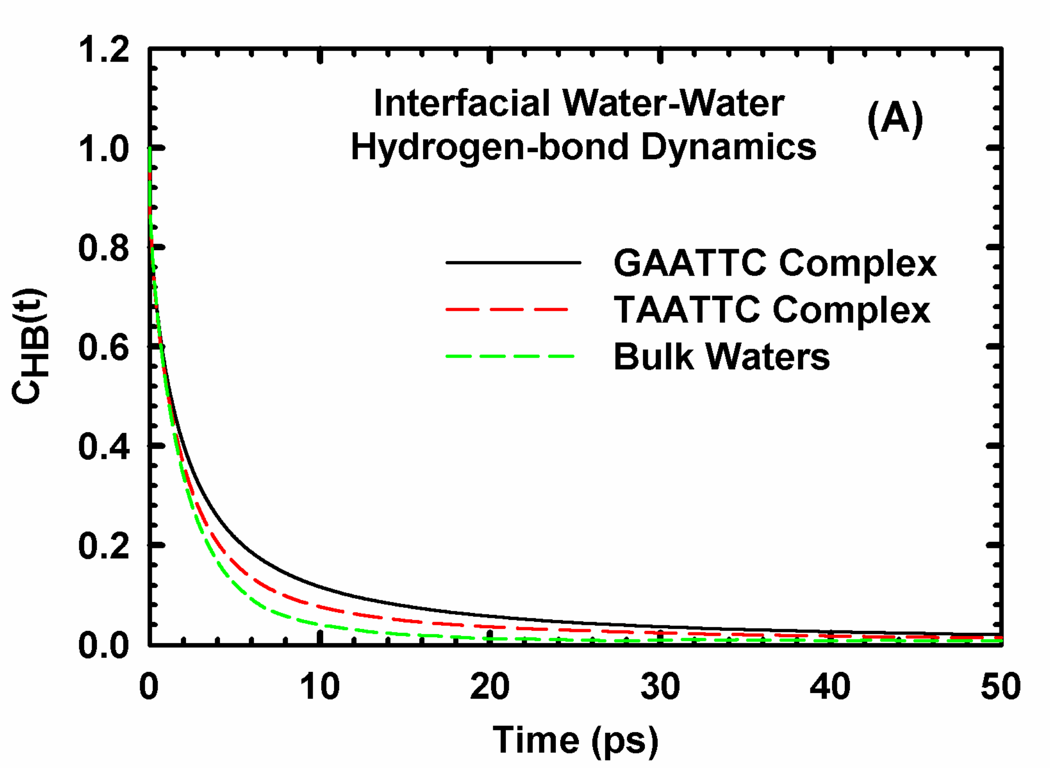
Table 4-5. Amplitudes and relaxation time constants for water-Protein/DNA Hydrogen-bond lifetime correlation function.

Region	**Cognate**			**Noncognate**		
	Amplitude	**Time (ps)**	**Average (ps)**	**Amplitude**	**Time(ps)**	**Average(ps)**
Interfacial	0.16 0.09 0.75	0.07 4.5 242.4	182.2	0.17 0.08 0.75	0.08 4.3 231.5	174.0
Intercalating	0.20 0.04 0.76	0.04 2.9 941.6	715.8	0.18 0.05 0.77	0.05 1.8 208.3	160.5

While, in the interfacial region, the slowest relaxation time, τ_2, in the cognate complex is only ~1.2 times that in the noncognate complex, in the intercalating region, τ_2 for cognate is ~4.5 times higher than that in the noncognate complex, thereby indicating that the hydrogen bonds of the intercalating waters in the cognate complex are significantly long-lived than that in the noncognate complex. This result is consistent with what one would expect from the rotational dynamics results presented in the previous section. The fast orientational relaxation and the short-lived hydrogen bonds of the intercalating water indicates that the water in the intercalating region of the noncognate complex is “dynamic” and is indicative of the transient fluidic nature of the water molecules in the intercalating region in the noncognate complex as against the cognate complex. In the next section we discuss the water-water hydrogen-bond dynamics in the interfacial and intercalating waters. Water-water hydrogen-bond dynamics, particularly in the interfacial region, is important since a dynamic hydrogen-bond network would indicate the “readiness” of the waters to “accommodate” a diffusing protein.

4.3.4 Short-lived Water-Water Hydrogen Bonds in the Noncognate Complex

In Figure 4-6, we show the autocorrelation functions for the interfacial and intercalating water-water hydrogen bonds, respectively. The water-water hydrogen bond in the interface of the noncognate complex decays slightly faster than that of the cognate complex, although no significant difference is seen in the case of intercalating waters. The amplitude-weighted average hydrogen-bond relaxation time is higher for interfacial waters in the GAATTC complex (4.9 ps) than that of the noncognate complex (3.6 ps) (Table 4-6), which means that the water-water hydrogen bonds around the noncognate complex break more quickly than those in the cognate complex.



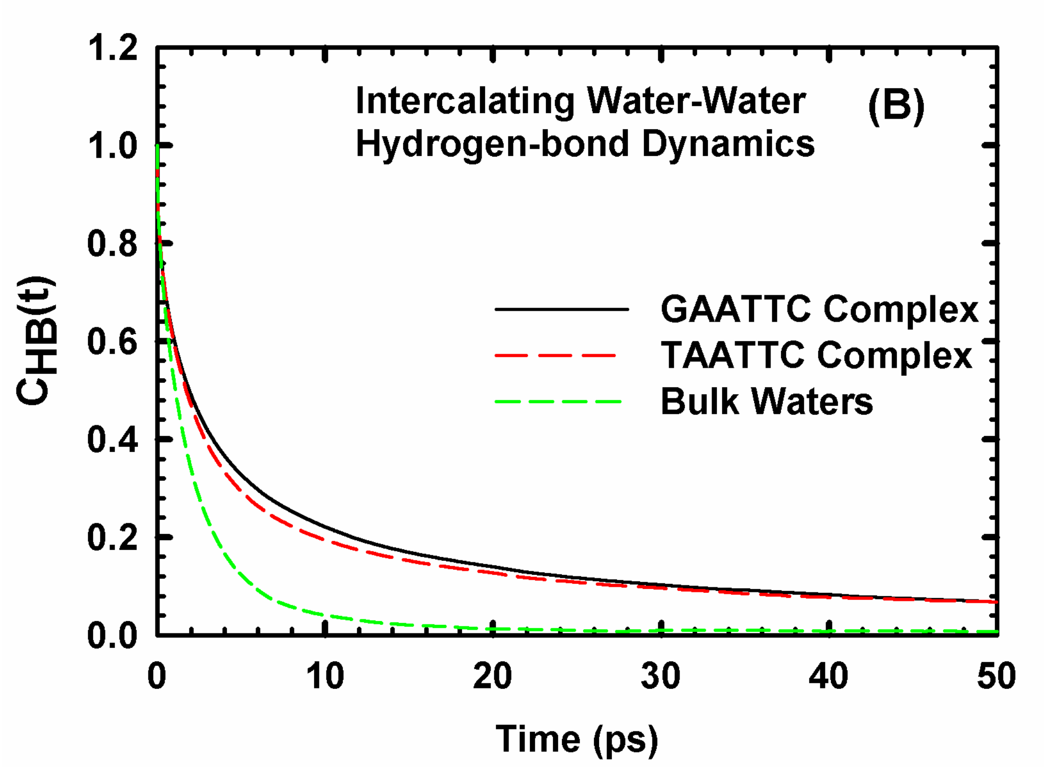


Figure 4-6. Water-water hydrogen bond lifetime correlation function of *interfacial* waters (A) and intercalating waters (B) around the cognate (black) and noncognate (red) complex.

The results are in line with the observations of Grossman et al.,[136] who also observe a difference in the hydrogen bond dynamics of interfacial water between a specifically bound protein-substrate complex and a nonspecifically bound protein-substrate complex. Particularly, Grossman et al. [136] also observe that the lifetime of hydrogen bonds is higher around the specifically bound protein-substrate complex. The water-water hydrogen-bond dynamics indicates the ease with which the network of hydrogen bonds can rearrange. We suggest that the faster breaking of the water-water hydrogen bonds in the interfacial region of the EcoRI-noncognate DNA complex relative to those in the cognate complex indicates the 'readiness' of the water molecules to accommodate the diffusion of the protein.

Table 4-6. Amplitudes and relaxation time constants for water-water hydrogen-bond dynamics.

	Cognate			Noncognate		
	Amplitude	**Time (ps)**	**Average (ps)**	**Amplitude**	**Time (ps)**	**Average (ps)**
Interfacial waters	0.28 0.56 0.16	0.2 2.7 21.1	4.9	0.25 0.65 0.10	0.1 2.3 20.8	3.6
Intercalating Waters	0.3 0.46 0.24	0.2 3.7 37	10.6	0.26 0.52 0.22	0.2 3.1 38.4	10.1
Bulk	0.21 0.74 0.05	0.1 2.3 19	2.6	--		

4.3.5 Thermodynamics of Water in Protein-DNA Binding

From the above results it is clear that the dynamics of the water molecules around the protein-DNA complex differs depending on the DNA sequence. We investigated if this difference in the dynamics also resulted in differences in the entropy and the average interaction energy of the water molecules, which, then would result in a difference in the thermodynamic driving forces for binding. The entropy of the water molecules are calculated using the 2PT scheme. In Figure 4-7 we show the translational and rotational density of states spectrum of the waters in the various regions around the cognate complex. A comparison of the density of states spectra of the waters in the cognate and noncognate complexes is given in Appendix C. The corresponding entropy values in the various regions are given in Table 4-7 and Table 4-8. In essence, the blue shifts of the bands [149, 150] corresponding to the O- - - O - - - O bending mode (at ~50 cm^{-1} in the translational spectrum of Figure 4-7) and the O--O longitudinal oscillations (at 200 cm^{-1} in the translational spectrum of Figure 4-7) indicate that water is severely restricted in its motion as one moves from the bulk water to interfacial and intercalating waters. Such restriction has also been observed for interfacial waters around DNA and bilayers [85, 87, 137, 151].

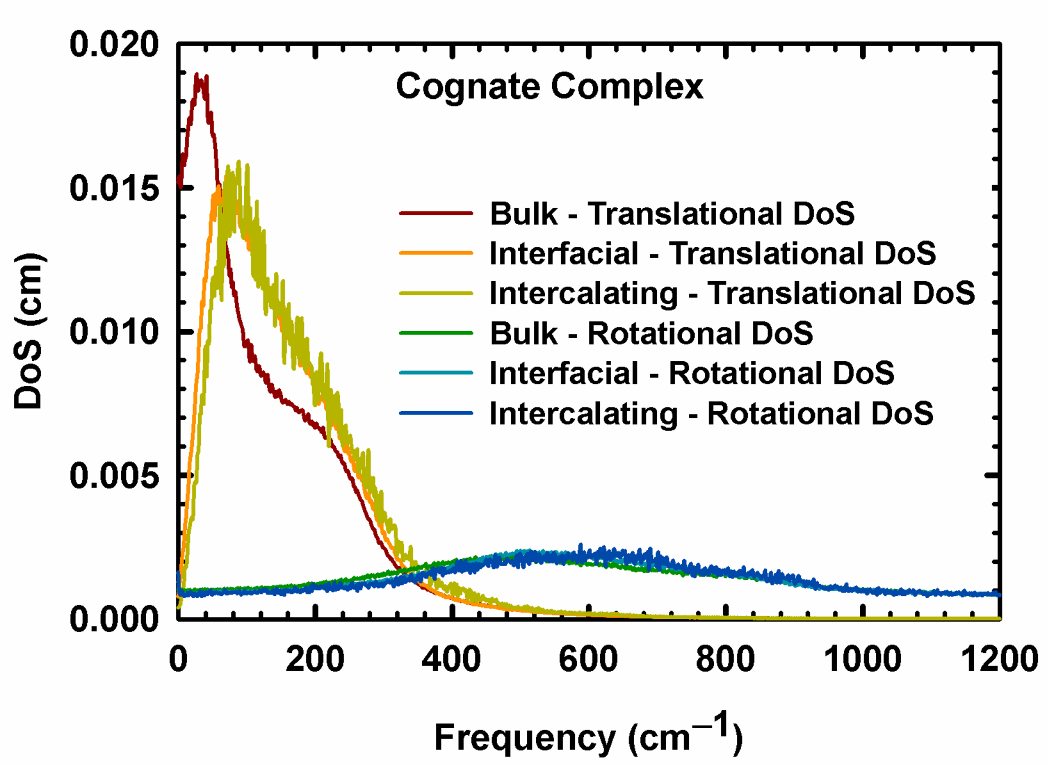


Figure 4-7. Translational and rotational density of states (DoS) spectrum of waters in the various regions around the cognate complex.

Energetically, the intercalating waters are the most stable (Table 4-9). However, there is no significant difference in the positions of the bands (see Appendix C) and the thermodynamic properties (see Table 4-7, Table 4-8 and Table 4-9) between the two complexes, indicating that the entropic and enthalpic driving force for binding is the same for the two complexes.

Table 4-7. Comparison of the translational entropy (J/mol/K) of the intercalating, interfacial and bulk waters in the cognate and noncognate complexes.

	Intercalating Water	Interfacial Water	Bulk Water
GAATTC Complex	36.68±1.42	40.40±0.80	56.62±0.24
TAATTC Complex	36.32±0.84	40.60±0.60	56.73±0.29

Table 4-8. Comparison of the rotational entropy (J/mol/K) of the intercalating, interfacial and bulk waters in the cognate and noncognate complexes.

	Intercalating Water	Interfacial Water	Bulk Water
GAATTC Complex	6.69±0.07	7.06±0.04	7.87±0.06
TAATTC Complex	6.56±0.12	7.13±0.10	7.88±0.03

Table 4-9. Comparison of the average interaction energy (kcal/mol) of the intercalating, interfacial and bulk waters in the cognate and noncognate complexes.

	Intercalating Water	Interfacial Water	Bulk Water
GAATTC Complex	−11.04±0.11	−10.05±0.06	−9.52±0.03
TAATTC Complex	−10.98±0.26	−10.26±0.18	−9.51±0.04

4.4 Concluding Remarks

Protein-DNA binding is brought about by the complex orchestration of several factors. Experimental evidence indicates that the protein exhibits a one-dimensional diffusion along the DNA before it chances upon its cognate sequence. Several studies have pointed out the decrease in the number of waters associated with the protein-DNA complex when the protein binds to its cognate sequence. This suggests that water plays an important role in protein-DNA recognition as the protein moves along the DNA. In this study we investigated the differences in the dynamics of water around EcoRI, a minimally restructuring protein, bound to its cognate sequence and to a minimally mutated noncognate sequence. The results show that even such a minimal mutation in the DNA chain results in higher hydration and faster dynamics of water molecules around the mutated protein-DNA complex. The faster dynamics of water, in turn, results in easily broken hydrogen bonds between the water and the protein/DNA. The results taken together indicate that the regions around the noncognate complex are more poised to allow the protein to diffuse away from the DNA. In Chapter 3, we had shown that such minimal mutations in the DNA can also cause changes in the dynamics of EcoRI sufficient to make it unfit for the required function. The studies together suggest that specific protein-DNA binding is brought about by the reduced dynamics of water around the protein-DNA complex which probably allows the formation of stable contacts between the protein and the DNA along with specific changes in the dynamics of the protein priming it for catalysis of the DNA. In addition to the roles of protein and water dynamics, DNA conformational properties also play a crucial role in protein-DNA interactions. The role of the protein environment on the DNA conformational properties is discussed in the next chapter (chapter 5).

5 PROTEIN-INDUCED SEQUENCE-DEPENDENT DNA CONFORMATIONAL CHANGES

"*DNA neither cares nor knows. DNA just is. And we dance to its music.*"

- Richard Dawkins

5.1 Introduction

The studies described in section 2.4 together suggest that sequence-dependent DNA deformability is likely to play an important role in protein-DNA interaction. However, one must realize that DNA properties also depend on the environment. For example, Williams et al. [152] show that the presence of even a single positive charge in the vicinity of the DNA can alter its flexibility. So, it is not immediately evident how the intrinsic deformability of the DNA would vary in a protein environment. To elucidate possible influence of protein on the DNA deformability, we make a systematic comparison of the *conformation* of DNA sequences that differ by 1 bp in the free and the EcoRI-bound forms. We perform all-atom molecular dynamics simulations of the four different DNA sequences that differ from each other at first base position of the recognition sequence in the free and EcoRI-bound forms. The conformation of the DNA is characterized by the twelve parameters that describe the relative orientations of the basepairs with each other (shear, stretch, stagger, buckle, propeller and opening) and the relative orientations of the basepair steps with each other (shift, slide, rise, tilt, roll and twist). We evaluate the significance of the changes in the conformational alterations in terms of the number of hydrogen bonds between the protein and the DNA. We believe that the hydrogen bonds and the DNA conformation are simple but sufficient parameters to meet our objective. In the next section we discuss the rationale for the choice of sequences in our study. This is followed by a discussion of the methodology that includes the description of how we obtained the substituted DNA sequences and the molecular dynamics simulation protocol. Finally the results are presented and the implication discussed.

5.2 Methods

5.2.1 Choice of Sequences

As stated above, our objective is to investigate possible influence of protein on the DNA deformability. One may achieve this by choosing sequences that vary at any of the three positions in the recognition half site. From Lesser et al. [51] it is known that the penalty of basepair substitution increases with increase in the position in the recognition site, i.e., a substitution at the first position affects the binding the least. The increasing penalty for substitution at the second and third positions in the recognition site might be indicative of even more complex inter-dependence of factors contributing to sequence discrimination. Hence, to keep the task relatively simple, we study the effect of substitution at the first position of the recognition site. Thus the sequences under comparison have the pseudo recognition sites: TAATTC, CAATTC and AAATTC. We shall call these “noncognate sequences” and when associated with the protein “noncognate complexes”.

5.2.2 Basepair Substitution and Molecular Dynamics Simulations

The starting structure for the protein-free DNA was the B-DNA structure available in the protein data bank (pdb id: 1bna) with the sequence 5′ - CGC*GAATTC*GCG - 3′. The initial structures of the noncognate sequences were generated by substituting G in the recognition site with the appropriate bases (A, T or C). The complementary bases were also appropriately substituted (i.e., T, A or G, respectively). The starting structure for the protein-DNA complexes was the crystal structure of EcoRI bound with the DNA sequence 5′ - TCGC*GAATTC*GCG - 3′ available in the protein data bank (pdb id: 1eri). The initial structures of the noncognate complexes were created by substituting G in the recognition site with appropriate bases and the complementary bases were also substituted appropriately. Substitutions were done with 'mutation' option in the Swiss PDB Viewer v4.0.1 [139]. The structures thus created were solvated in SPC water in a 10 x 10 x 10 nm^3 box. This was followed by 500 steps of energy minimization by steepest descent method and by molecular dynamics simulation at 300 K with the molecular interactions represented by the AMBER03 forcefield [121, 153-155] for the protein and ions and the PARMBSC0 force field [120] for DNA. All simulations were done using the Gromacs 4 package [119]. Free DNA was simulated for a total of 15 ns and the protein-DNA complexes were simulated for a total of 50 ns. Long range electrostatics was computed using the particle mesh ewald method. Appropriate numbers of Na+ counterions were added.

5.2.3 Conformational Parameters and Hydrogen Bond

The coordinates of the atoms were written every 10 ps. The six basepair and the six basepair step parameters were calculated for every 10 ps interval using the 3DNA program [130, 131].

5.2.4 Hydrogen-bond Analysis

For the hydrogen-bond analysis, two atoms were considered to form a hydrogen bond if the acceptor-donor distance is <3.5 Å and the hydrogen-donor-acceptor angle is <30°. We define the propensity of two atoms to form a hydrogen bond as the ratio of the number of frames in which they are hydrogen bonded (according to the above criteria) to the total number of frames used in the analysis. Thus, a propensity value closer to 1 indicates that the two atoms under investigation are more likely to be found hydrogen bonded and a value closer to 0 indicates that the two atoms are least likely to be found hydrogen bonded. This helps us to evaluate the significance of a given hydrogen bond.

5.3 Results & Discussion

We begin by discussing the stabilization of the trajectories from the molecular dynamics simulations. This is followed by the discussion of differences in the DNA conformations that arise in the protein-free forms as a result of basepair substitution. We then compare the DNA conformational differences in the protein-bound forms followed by a discussion on the nature of DNA conformational differences between protein-free and protein-bound forms. The influence of the terminal nucleotides and the reproducibility of the calculated conformational parameters are subsequently discussed. Then we present the results from hydrogen bond analyses followed by a discussion of the implications of the results for specificity in protein-DNA interactions.

5.3.1 DNA Conformation

The conformation of a DNA sequence can be effectively described by the basepair parameters (translational: shear, stretch, stagger; rotational: buckle, propeller, opening) and the basepair step parameters (translational: shift, slide, rise; rotational: tilt, roll, twist). Although these parameters describe the relative orientations of the bases and the base steps, any conformational change in the DNA backbone will also be reflected in the orientations of the bases or the base steps. Despite the varying energetic scales of deformation for each of the parameters, these parameters together form a 12-dimensional coordinate set to describe the DNA conformation. Hence none of the parameters can be considered to be more important than the other and a total estimate of the number of parameters that differ significantly between sequences is in itself a quantitative measure of differences in the conformation of two DNA sequences. We first characterize the DNA conformation for each of the protein-free and the protein-bound DNAs. To compare the conformation of each of the noncognate sequence/complex against the cognate sequence/complex, we analyze the number of conformational parameters that differ (that is, beyond one standard deviation) from the cognate sequence/complex in each position in the recognition sequence.

5.3.2 Basepair Substitution Leads to Altered DNA Conformation in the Protein-free State

First, we identify the conformational differences arising in a free DNA upon 1 basepair substitution. In Figures D-1, D-2 and D-3 (Appendix D), we compare the mean values of the conformational parameters of each of the noncognate sequence with the cognate sequence. The error bars denote the standard deviations in the conformational parameter. In Figure 5-1 we show the number of conformational parameters that vary (i.e., the mean values that differ beyond one standard error) for each of the noncognate sequences from the cognate DNA sequence (protein-free DNA) at each base position in the recognition site. The most varying sequence is CAATTC and the least varying is AAATTC sequences. We note that in the first position, the number of parameters that vary is generally higher than compared to the subsequent positions. Understandably, this is because of the substitution at first position. This substitution has led to conformational changes in the adjacent positions as well. One can also note that the differences in the conformation die off with increasing basepair step. There is no difference in the conformation after 2 basepair steps from the point of substitution. This is in accordance with the results of Lavery et al. [100] who show that a basepair's influence on conformation can extend until its next-nearest neighbor.

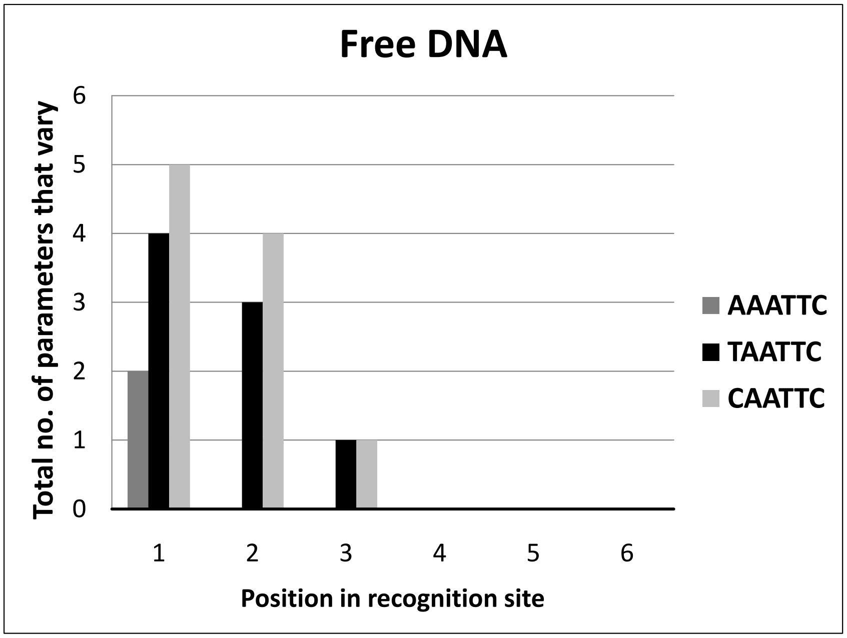


Figure 5-1. Total number of conformational parameters that vary for each of the pseudo-specific DNA from the specific DNA in the free form.

Substitution of G with A/T/C in the first basepair of the recognition sequence alters the conformation of the second and third basepairs in the recognition sequence. One must note that the subsitiution of G with a pyrimidine (T or C) results in a larger number of variations in the conformation than with a subsitiution of A, a purine. This is consistent with the expectations from the previous studies [32, 53] which show that a pyrimidine-purine step is more flexible than a purine-purine step. The flexibility introduced upon substitution of G with T or C results in a larger number of conformational variations. Further, upon dissecting the origins of the difference in the conformations to basepair- and basepair step-level, we note that, in general, the differences arise at the basepair *step* level. For example, for the most-varying noncognate sequence, CAATTC, of the total 10 conformational parameters that vary, only 3 are basepair level changes. Similarly, for the TAATTC pseudo-specific sequence, of the total 8 varying conformational parameters, only 1 is a basepair parameter.

5.3.3 Protein Environment Alters DNA Conformation at Basepair Level in a Sequence-dependent Fashion

Having identified that the protein-free DNA sequences show sequence-dependent preferred conformations resulting from the change in the first basepair of the recognition site, we now investigate the conformations adopted by these sequences upon protein-binding (i.e., in the protein-bound form) (see Figures D-4, D-5 and D-6). In Figure 5-2 we show the number of deviations of the conformations of the DNA in the noncognate complexes from the DNA conformation in the cognate complex as a function of the position in the recognition site (i.e., the number of parameters that vary beyond one standard error of the mean values). One immediately observes that the substitution at the first position in the recognition site has caused dramatic changes in the DNA conformation *throughout* the recognition site unlike in the free DNA where the differences tended to die off with subsequent basepair steps.

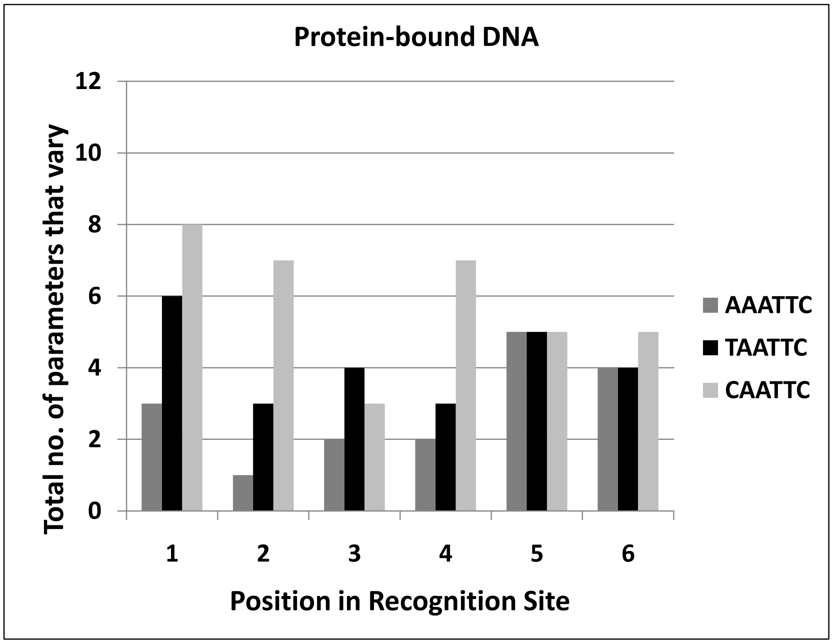


Figure 5-2. Total number of conformational parameters that vary for each of the pseudo-specific DNA from the specific DNA in the protein-bound form.

In Figure 5-3, we show the correlation coefficients of each of the helecoidal parameter in the specific complex as a function of the position with respect to the first base in the recognition site and the correlation coefficients of the corresponding free DNA. In the protein-bound form, the rise is correlated until the 4th position and twist is correlated until the 5th position. In contrast, in the free DNA, the variables are not correlated beyond the 3rd position. Thus, in general, the conformational parameters are correlated significantly longer in the protein-bound DNA sequence than in the free DNA. Interestingly, we also observe that the correlation between the conformational parameters of the bases in the recognition site with respect to the immediately upstream base in the specific complex is (see Figure 5-4) more or less the same as that of a free DNA. Thus, it is clear that protein binding induces long-range correlations within the DNA it is bound to.

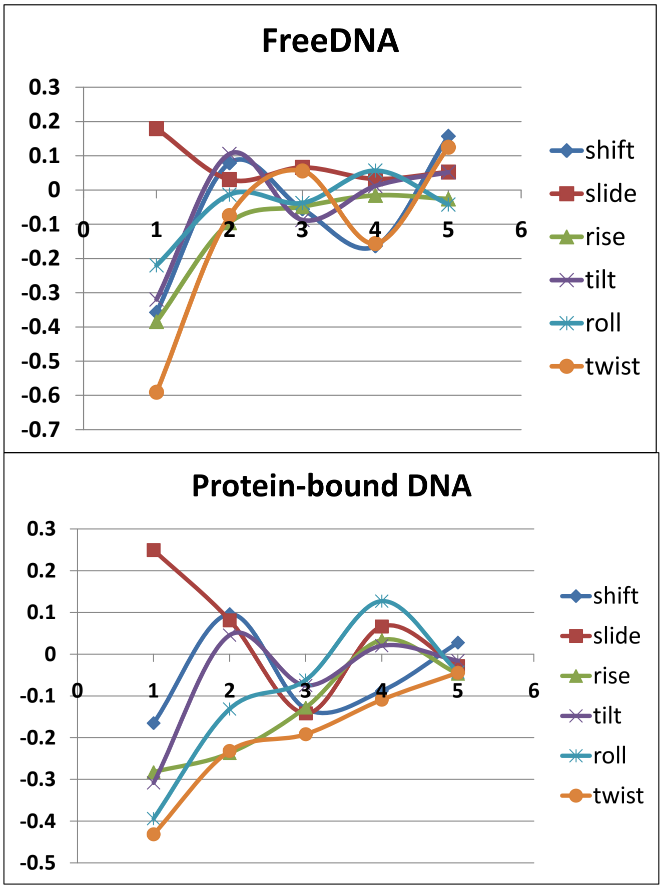


Figure 5-3. Comparison of Correlation coefficients based on GC3

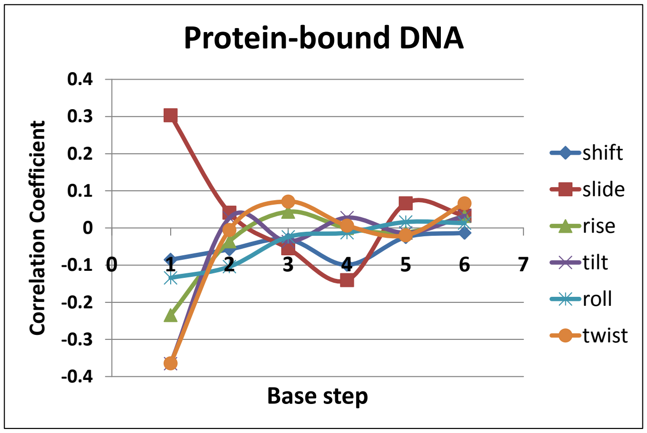


Figure 5-4. Comparison of Correlation coefficients based on CG2

Next, to elucidate the influence of the protein on the DNA conformation, we compare the number of deviations (of the preferred conformations) from the cognate sequence. We see that in the protein-bound form the sequences show greater variation than in the protein-free form (See Figure 5-1 and Figure 5-2). For example, for the CAATTC sequence, in Position 1, there are 5 parameters that vary from the specific sequence in the free form while there are 8 parameters that differ from the specific complex for the same position in the protein-bound form. Likewise, in all positions, for all of the pseudo-specific cases, the numbers of parameters that vary from the cognate sequence (i.e., the protein-free form) are less than the number of parameters that vary from the cognate complex (i.e. the protein-bound form).

To further dissect the origin of the difference, we analyze the basepair and base step parameters separately. Figure 5-5 and Figure 5-6 compare the number of basepair parameters and basepair step parameters respectively, that differ for each of the noncognate sequences from the cognate sequence in the free and protein-bound forms. In the protein-free forms, the basepair parameters only slightly vary. However, in the protein-bound complexes, the number of basepair parameters that vary from the cognate complex at a position may be as high as four (of the six parameters). Similarly, in the protein-bound complexes, the numbers of basepair step parameters that vary are greater when compared with the number of basepair step parameters that vary in the protein-free forms.

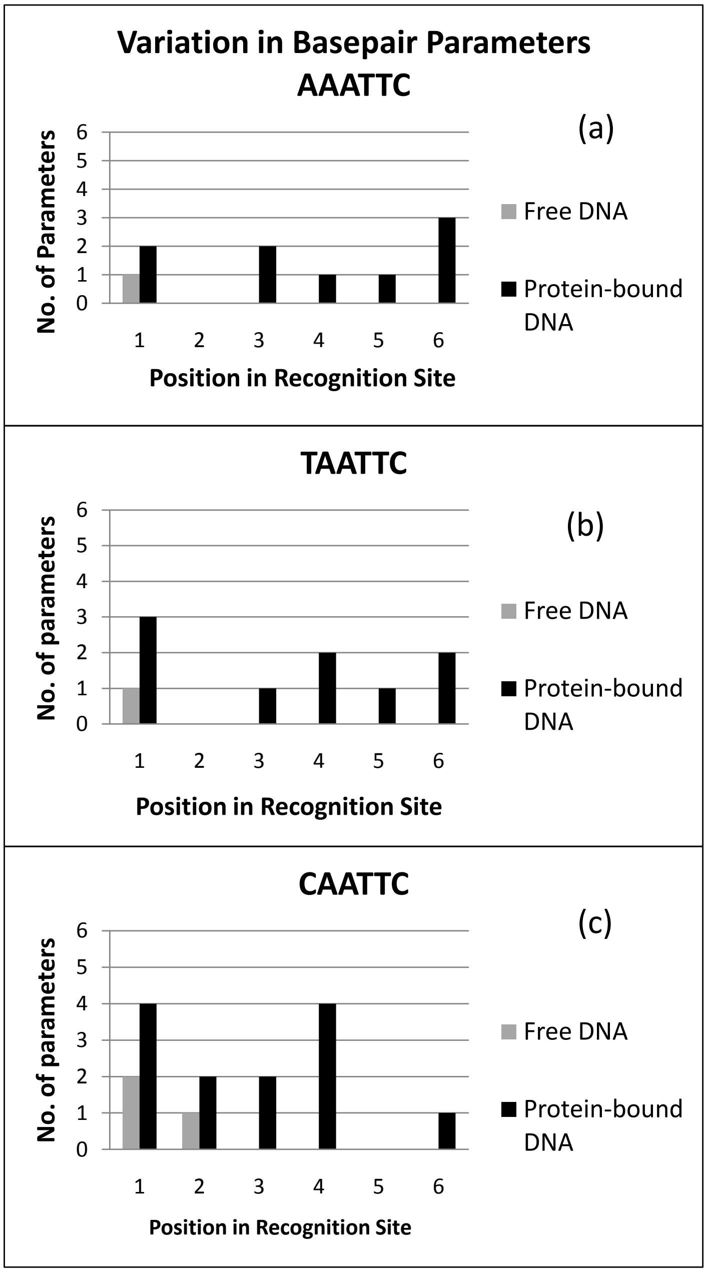


Figure 5-5. Comparison of the number of *basepair parameters* that vary for free and EcoRI-bound DNA sequences shows that in the protein-bound form the variation is high. (a) Comparison of free and protein-bound AAATTC, (b) comparison of free and protein-bound TAATTC (c) comparison of free and protein-bound CAATTC.

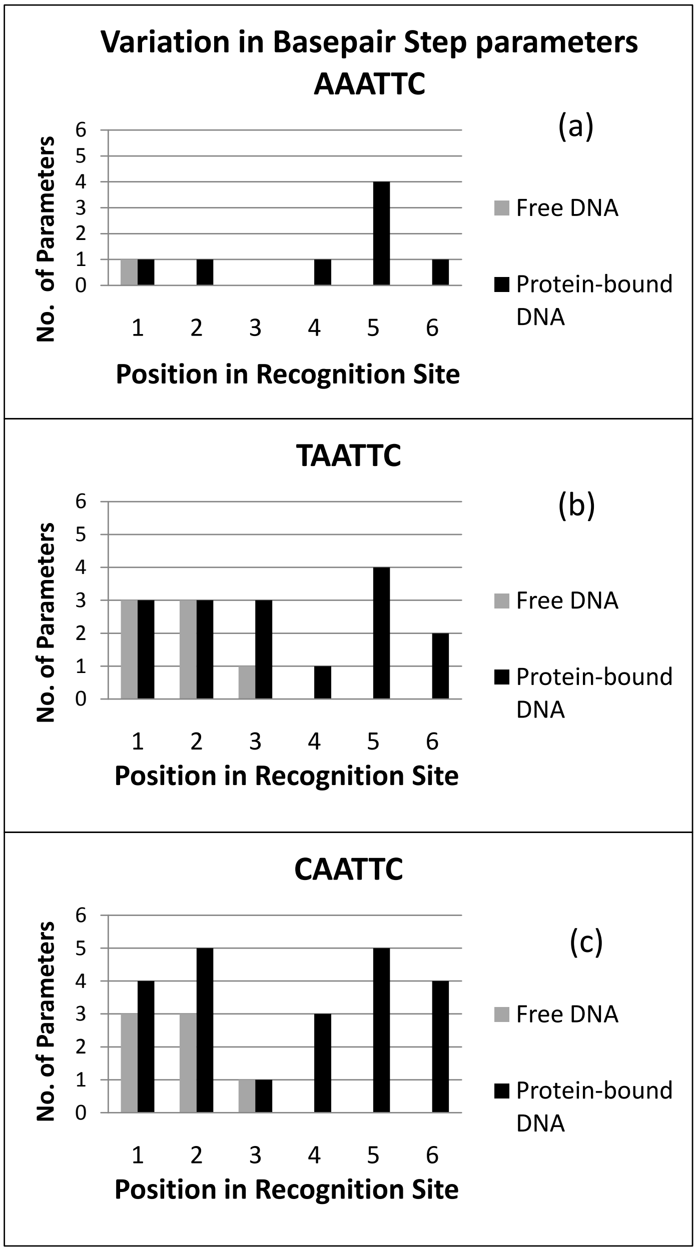


Figure 5-6. Comparison of the number of *basepair step parameters* that vary for free and EcoRI-bound DNA sequences shows that in the protein-bound form the variation is high. (a) Comparison of free and protein-bound AAATTC, (b) comparison of free and protein-bound TAATTC (c) comparison of free and protein-bound CAATTC.

Thus, in general, one observes that both the basepair parameters and the basepair step parameters have varied greatly for the sequences in the protein-bound state compared to the protein-free state. The results demonstrate that in all the cases, upon protein binding, the most preferred conformations of the basepair parameters and the basepair step parameters vary, indicating that the conformation of the DNA varies at the basepair level upon protein binding, i.e., the basepair is no longer a rigid plane. Furthermore, the variations in the average values are sequence-dependent.

5.3.4 Fluctuations in the Conformational Variables

The fluctuations in the conformational variables are indicators of the deformability, or flexibility, of the DNA. Having assessed the influence of the protein on the average conformations of the DNA, we now assess the influence of the protein on the flexibility of the DNA. However, one must note that the changes introduced in the DNA we assess here are from an already bound protein-DNA complex. This may or may not be different from the flexibility in DNA that may be induced *as* the protein binds to the DNA. In Figures D-7, D-8, D-9 and D-10, we show a comparison of the magnitude of the fluctuations at each position for the protein-free and the protein-bound cases. One observes that there is a complex trend when one compares the fluctuations in protein-bound and protein-free cases. In the case of stretch, the central residues AAT show higher fluctuations in the protein-bound form. In the case of stagger, the protein-bound DNA shows larger fluctuations throughout the recognition site. Buckle shows a complex behavior with the free DNA being more flexible than the protein-bound DNA in two cases (AAATTC and GAATTC) and the protein-bound DNA being more flexible in the other two. Propeller and opening show complex behavior, in that, some of the residues show large fluctuations in the protein-free and others show larger fluctuations in the protein-bound DNA. In the case of the basepair step parameters, shift and slide show distinctly reduced fluctuations in the protein-bound form. Rise, in contrast, shows larger fluctuations in the protein-bound form in general. While tilt and roll show complex behavior in their nature of fluctuations in the protein-bound and protein-free forms, twist, in general, has large fluctuations in the protein-free forms. Overall, one may note that the basepair parameters, in general, have become more flexible upon protein binding, while the basepair step parameters have become less flexible upon protein binding.

5.3.5 Implications of Protein-induced Sequence-dependent DNA Conformational Differences for Protein-DNA Recognition

The results so far show that a protein alters the average conformations of the DNA at the basepair level in a sequence-dependent manner and introduces long-range correlations in the DNA motions. The natural doubt that arises then in one’s mind is whether the difference in the ‘average’ DNA conformations will have a *functional* implication. Under the plausible assumption that changes in the helecoidal parameters will result in the formation or non-formation of specific hydrogen bonds, we investigated the patterns of hydrogen bonds between the protein and the DNA. Hydrogen bonds between the protein and the DNA have been known to play important roles in governing the specificity. Overall, the total number of hydrogen bonds varies as 143 for cognate, 126 for AAATTC, 139 for CAATTC and 148 for TAATTC complexes (see Table D-1 for a list of protein-DNA hydrogen bonds and their propensity). One also notes the weakening of certain hydrogen bonds and the strengthening of a few. These indicate that the changes in the average conformational parameter indeed give rise to differences in the hydrogen bonding. One must note that our purpose to analyze the hydrogen bonds is only to show that the altered conformational parameters lead to altered patterns of hydrogen bonds and not the stability of the complex. The stability of the complex depends on a number of factors, including the strength of each of the hydrogen bonds, water-mediated interactions, hydrophobic associations etc. Moreover, in the case of EcoRI, it has been observed that there is no direct correlation between the strength of binding and the rate of catalysis [156]. We also investigated the distance between the phosphorus atom and the key amino acids in the cleavage site. In Table 5-1, we show the mean distance (and its standard error) of the phosphorus atom (between G and A) and the center of mass of the aminoacid residues that have been hypothesized to be involved in the catalytic process [157]. One can see that, upon substitution, the distances between the phosphorus atom and the key amino acids [157] increase. Since the precise positioning of active site residues with respect to the phosphodiester bond are quintessential for catalysis (see [158]) the changes in the distances will not result in catalysis of the sequences.

Table 5-1. Mean distance (± standard error) of the phosphorus atom in the cleavage site and the aminoacid residues hypothesized to be involved in catalysis [157]. Distances are in nm.

	GAATTC	AAATTC	CAATTC	TAATTC
Asp91	0.60 ± 0.01	0.61 ± 0.02	0.75 ± 0.01	0.69 ± 0.01
Glu111	0.75 ± 0.01	0.76 ± 0.01	0.82 ± 0.01	0.78 ± 0.01
Lys113	0.59 ± 0.01	0.58 ± 0.01	0.60 ± 0.01	0.63 ± 0.01

The implications of these observations on the DNA conformation are easily extrapolated to specificity in protein-DNA binding. Each DNA sequence, even if different by just 1 bp, adopts different conformation. This DNA sequence-dependent difference gets amplified by the introduction of conformational difference at the *basepair* *level* upon protein binding. Since, in a protein-bound DNA, the conformations are correlated significantly longer, the amplified conformational difference *extends over the entire recognition site in the presence of the protein*. Furthermore, these conformational differences also lead to altered hydrogen bonding patterns between the protein and the DNA. Thus, for a given protein, only sequences whose conformation is altered *appropriately*, or amenable to alteration to achieve the required conformational complementarity, are 'recognized' by the protein. The other sequences, even if different by just 1 bp, do not achieve the required conformation and are not 'recognized' by the protein. We suggest that this is one of the mechanisms by which the protein achieves stringent discrimination as it slides over a sea of random sequences and chances upon and recognizes its cognate sequence. Figure 5-7 presents a schematic illustration of the above suggested mechanism of introduction of DNA conformational changes at the basepair level upon protein binding.

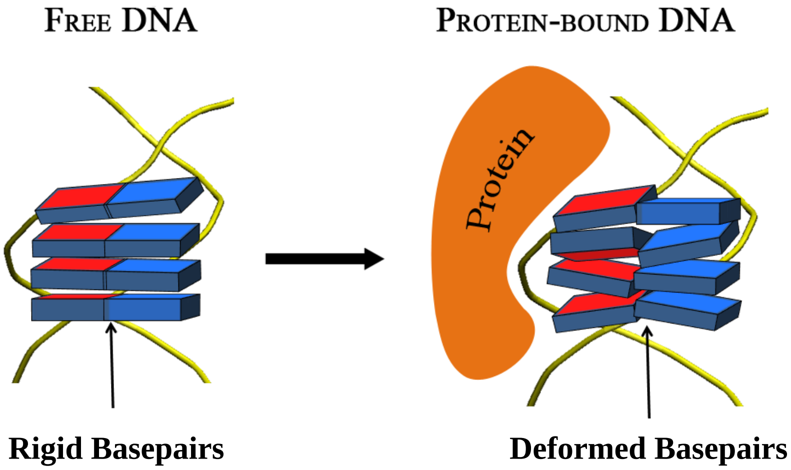


Figure 5-7. An illustration showing the flexibility in the DNA introduced at the basepair level upon protein binding.

5.4 Concluding Remarks

Our investigation here of the interactions between EcoRI and its noncognate sequences that differ at the first position from the cognate sequence shows that a difference in the first base position of the EcoRI recognition site could lead to dramatic changes in the average DNA conformation in the EcoRI-bound state, especially at the basepair level. We attribute this to the influence of the protein environment on the sequence-dependent conformation of the DNA and the long-range correlations introduced in the motions of the DNA. We suggest that while the intrinsic sequence-dependent DNA conformation provide cues for sequence discrimination to an extent, protein binding ‘amplifies’ the difference in the DNA conformation of sequences, resulting in the non-recognition of sequences that differ even by 1bp.

The unique electronic properties of each of the nucleotides give rise to sequence-specific characteristics to the DNA. Protein binding could alter the electronic properties of the nucleotides. This, in turn, would affect the properties of the DNA sequence. For example, it is very well known that proteins neutralize the negative charge on the phosphate groups of the DNA. It has even been thought to be one of the mechanisms by which proteins bend DNA [159, 160]. Would neutralization of phosphate charges and the resulting change in the charge distribution contribute to the basepair flexibility? If so, to what extent neutralization would disturb basepair rigidity? Why would different sequences exhibit different levels of flexibility for the same protein? Such questions probing the cause of the change in the basepair rigidity and detailed quantitative analyses of the effects are important for reengineering of DNA-binding specificity of proteins. Reengineering of DNA-binding specificity not only serves as a test of our current understanding of the mechanisms of protein-DNA interactions but also has immense practical relevance for gene targeting and regulation [35, 161-163]. Detailed examination of the DNA conformational mechanics in the protein environment will hasten the efforts in reengineering enzyme specificities.

6 CONCLUSIONS AND FUTURE DIRECTIONS

"*There is a mask of theory over the whole face of nature*."

- William Whewell

Mankind's journey that started as a quest to seek a theory for the inheritance of traits has traversed many diverse routes. Yet, the destination is not in sight primarily because of the complexities in delineating molecular-level phenomena. Of the many different molecular-level phenomena that are crucial to our understanding of the inheritance of traits, protein-DNA interactions are one of the first and fundamental biomolecular interactions that were identified as critical in passing on the stored genetic information and have taken large strides. However, there are still very many questions unanswered. While this is partly because of the inherent complexity of protein-DNA interactions, it is also because of the lack of systematic studies elucidating the range of structural/dynamic responses and attendant changes for a particular enzyme as it binds to cognate and noncognate sequences. Thus the unifying scope of this work is to elucidate the structural/dynamic responses to and attendant changes in a protein when it binds to cognate and noncognate sequences. Here, we have chosen EcoRI, a restriction endonuclease that cleaves the DNA between guanine and adenine in the DNA sequence $(GAATTC)_2$ with very high specificity. EcoRI restructures minimally upon binding to the DNA and thereby enabling us to study the issues of interest relatively unclouded by any folding/unfolding dynamics that occur in many of the DNA-binding proteins. We have focused on three aspects, namely, protein dynamics, water dynamics and the role of protein in DNA conformation.

6.1 An Overview of Major Conclusions

Proteins (and other biomolecules) are dynamic entities and not static as seen in crystallographic structures. In this study, we characterized the protein dynamics, using principal component analysis, of EcoRI bound to the cognate sequence and a noncognate sequence, as presented in Chapter 3. In addition, the DNA conformation and the structural relaxation of the arms were also monitored. The results showed a change from a coordinated, twisting-type dynamics in EcoRI bound to cognate sequence to a relatively scrambled dynamics when EcoRI is bound to a noncognate sequence. This difference extends all the way to the interface, with the interfacial residues showing a constricting motion in the cognate sequence. Further, the arms and the DNA showed structural relaxation when the protein is bound to a noncognate sequence.

We then investigated the dynamics and thermodynamics of water around EcoRI bound to cognate or noncognate sequences (Chapter 4). The regions around the protein-DNA complex were divided into intercalating and interfacial regions based on whether or not the water molecules were bound simultaneously to both the protein and the DNA. We first compared the number of water molecules associated with each of the complexes. The results showed a higher hydration of the noncognate complex. Then the dynamics of water was characterized using the dipole reorientational correlation function and mean-squared displacements. The entropy was calculated using the two-phase thermodynamic theory. The results revealed that while the reorientational dynamics of water in the interfacial region was essentially the same in the two complexes, the reorientational dynamics of intercalating waters in the cognate complex was significantly slowed. In particular, the slowest relaxation timescale corresponding to the tumbling motion was greatly reduced in the cognate complex. The faster reorientational dynamics of water in the intercalating regions of the noncognate complex also leads to faster break up of hydrogen bonds with the protein/DNA complex.

In essence, the studies presented in Chapter 3 and Chapter 4, collectively, indicate that as the protein slides over the DNA and chances upon the cognate sequence, specific changes in the protein dynamics occur accompanied by a slowing down of water dynamics (particularly, rotational) and by decreased inter-arm distances. These changes, together, alter the DNA conformation and promote the formation of a stable complex. The next question that arises is, whether the protein, as it slides over the DNA, just exploits the sequence-dependent free DNA conformational properties or whether it plays an active role in bringing about sequence-dependent DNA conformation (Chapter 5). For this, we compared the conformation of free DNA sequences and protein-bound DNA sequences. The DNA sequences differed in the first position of the recognition site. A comparison of the differences in the conformation the DNA sequences in the free and protein-bound cases reveals that proteins play an active role in altering the DNA conformation at the basepair-level in contrast to changes observed predominantly only at the basestep-level in free DNA sequences. Thus, collectively, the results indicate that as the protein slides along the DNA, it not only undergoes dynamical changes in specific regions (along with changes in intercalating water dynamics), but also influences the DNA conformation at the basepair level, thus amplifying the conformational differences between the cognate and noncognate sequences.

6.2 Recommendations for Further Studies

The study presented in this thesis has provided new insights into protein-DNA interactions from the point of dynamics of protein and water and the active role played by proteins towards sequence-dependent properties of DNA. The insights have also led to many further questions, primarily centered on dissecting and providing mechanistic insights into the interplay of these factors (protein dynamics, hydration and DNA conformation). These questions are outlined here as recommendations for further studies in this section.

6.2.1 DNA Sequence-dependent Protein Dynamics to Cause DNA Conformational Changes?

In Chapter 3, we observed that even minimal perturbations to the DNA sequence can cause changes in protein dynamics that extend to the interfacial region. These changes are accompanied by structural relaxations of the arms and DNA conformation. The question that now arises is whether the dynamics of the protein will be different when it is bound to different noncognate sequences. Say, for example, EcoRI is bound to a noncognate sequence that is doubly mutated or bound to a random sequence. Will the dynamics observed in these cases be similar to that observed when the protein is bound to the minimally mutated sequence? In essence, the question is whether there is sequence-dependent protein dynamics. Further, since there is a marked change in the protein dynamics upon binding to the minimally mutated DNA sequences and is associated with DNA conformational changes, a natural question that arises is whether dynamics plays a role in bringing DNA conformational changes.

To answer the above questions, we suggest the following. Simulations of EcoRI-DNA complexes with the following recognition site sequences shall be performed:

1. CAATTC – According to Lesser et al. [51], EcoRI has the least preference for this sequence as the sequence results in the loss of a hydrogen bond *and* appositional interactions.

2. AAATTA – This doubly mutated site (first and last sites of the recognition sequence) has been shown *not* to be cleaved by EcoRI [51].

3. CTTAAG – This is the inverse of the recognition sequence GAATTC and has also been shown experimentally not to be cleaved by EcoRI [51].

From the above-mentioned simulations of EcoRI with the maximally mutated (CAATTC), doubly mutated (AAATTA) and inverted DNA sequences (CTTAAG), the dynamics of EcoRI in each case shall be characterized and compared to reveal sequence-dependent dynamical changes, if any. Further, the role of dynamics in bringing about DNA conformational changes can be investigated by using appropriate position-restrained simulations as described below.

One starts with the equilibrium structure (or the crystal structure) of EcoRI bound to the cognate sequence, mutates the cognate sequence to a noncognate sequence, say, TAATTC, and performs three series of simulations with position-restraining (a) the whole protein (b) the interfacial residues or (c) the arms in each case. Since we already know that a non-position-restrained simulation in such a case will lead to the relaxation of the DNA, position-restraining (which is equivalent to absence of dynamics) will reveal if dynamics is essential to relaxing the DNA conformation. Position-restraining the specific regions (the interface and the arms) will further reveal if the dynamics of these specific regions are alone sufficient to cause DNA relaxation. Similar types of analysis can be done for the doubly mutated and inverted DNA sequences as well.

6.2.2 The Role of Dehydration in DNA Conformational Changes

In the previous section, we outlined how one might study the presence of DNA sequence-dependent protein dynamics and elucidate the role of dynamics of specific regions of the protein in preserving the DNA's kinked structure even in a noncognate sequence. In this section, we propose a study to delineate the role of dehydration of the DNA surface when the protein approaches the DNA. Dehydration of the DNA is often associated with conformational changes [164]. Hence, it is plausible that as the protein approaches the DNA, the displacement of water molecules from its surface can itself cause conformational changes in the DNA. However, it is not clear how only certain sequences manage to undergo conformational changes (like the DNA kinking observed in the cognate complex and the absence of DNA kinking in the noncognate complex). We propose an examination of the effect of a protein displacing water as it approaches the DNA. For this, it should be sufficient to represent the protein as a flat surface with charges (see Figure 6-1). We believe that despite the simplistic representation, the model can sufficiently capture the underlying physics. The "approach" of the protein can be modeled by varying the distances of the surfaces from the DNA. By comparing position-restrained and non-position-restrained DNA cases (using molecular dynamics simulations), one can elucidate the dynamical and thermodynamic aspects of water in the various regions (minor and major grooves) surrounding the DNA. That is, we propose to study the dynamics and thermodynamics of water around the DNA as functions of the distances from the flat surfaces. Another simulation where the DNA is not position-restrained can be performed where the changes in the DNA conformation as the distance between the flat surfaces decreases is then monitored along with the changes in the dynamics of water. These studies will provide insights into the role of dehydration of the DNA surface by the protein. Further, one can also vary the charge density and/or use heterogeneous charges on the flat surfaces. Heterogenously charged surfaces will better represent a protein surface.

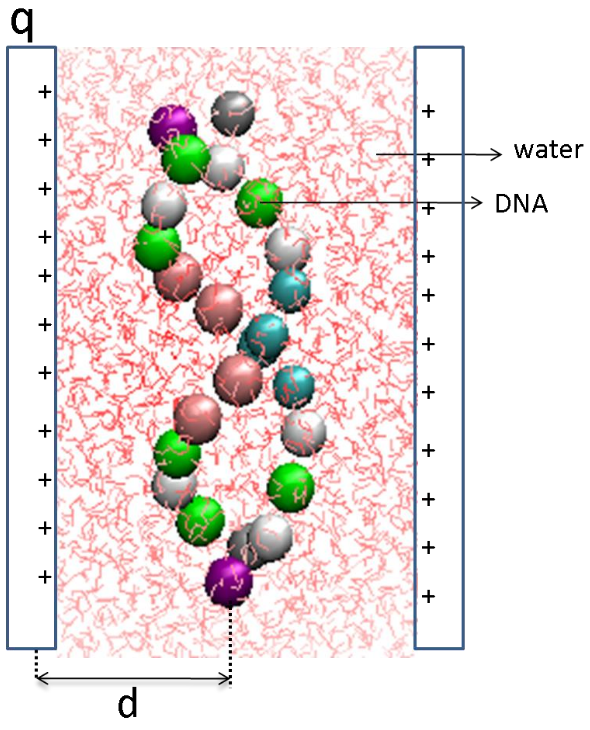


Figure 6-1. A schematic representation of the proposed work to delineate the role of dehydration of the DNA surface by the protein surface.

6.2.3 Effect of Osmolytes on Protein-DNA Interaction

Osmolytes such as glycerol are commonly used during the storage of proteins. When such proteins are then used in molecular biological applications they interfere with the process under investigation. For example, during restriction enzyme digestion of DNA, if the glycerol concentration exceeds 5% v/v it is known to induce “star” activity, where the window of specificity of the enzyme is increased [165]. The altered specificity has been found to be strongly correlated to change in osmotic pressure, which is, in turn, related to the water activity [166, 167]. Conlan et al. [168] have showed that it is possible to use neutral detergents to manipulate restriction endonuclease reaction rates and specificities. However, due to the lack of molecular-level details on how star-activity is induced, it has not been possible to rationally modulate the protein-DNA interactions and, in turn, modulate specificity. We propose that an investigation of the effect of addition of glycerol (at various concentrations) to the TAATTC-noncognate complex using molecular dynamics simulations would be useful in this regard. The objective here will be to monitor specifically the changes in hydration, dynamics of water (particularly in the intercalating regions) and the dynamics of protein. These observations along with those in the cognate complex will shed light on the molecular mechanisms of the “star” activity of EcoRI in the presence of glycerol in addition to elucidating the role of water. (The forcefield parameters including charges and the Lennard-Jones parameters σ and ε are available from the literature [169].) A clear understanding of the molecular-level details will enable us to devise rational ways to modulate the interactions between proteins and DNA and, in turn, modulate specificity in a rational way.

6.2.4 Role of Phosphate Neutralization on DNA Conformation

The unique electronic distribution on each nucleotide and the sequence of the nucleotides determine the conformation and deformability of a given stretch of DNA. Any perturbation of the electronic distributions on a nucleotide, in turn, is likely to affect its conformation and flexibility. Our investigation presented in Chapter 5 on the interactions between EcoRI and the DNA sequences that differ at the first position from the cognate sequence showed that a difference in the first base position of the EcoRI recognition site could lead to dramatic changes at the basepair level in the average DNA conformation in the protein-bound state. In contrast, a free DNA showed only basestep-level changes. We attributed this to the influence of the protein environment on the sequence-dependent conformation of the DNA and the long-range correlations introduced in the motions of the DNA. We propose an investigation of how protein alters the conformation and/or flexibility of the nucleotides by using quantum-mechanical calculations. In the next paragraph, we explain the rationale and our approach to the study.

As a protein approaches the DNA, it makes contacts with the phosphate backbone of the DNA. It has been proposed that proteins deform DNA by neutralizing the phosphate in the DNA backbone [170]. Several studies have confirmed this hypothesis. Strauss and Maher [159] showed by electrophoretic experiments and methylphosphonate substitutions that asymmetric phosphate neutralization of DNA induced DNA bending toward the neutralized surface. Hamelberg et al. [171] showed by molecular dynamics simulations that the minor groove width significantly narrowed upon the neutralization of the phosphate backbone in a sequence-dependent manner. Okonogi et al [172] showed by continuous-wave electron paramagnetic resonance that the local flexibility of duplex DNA when methylphosphonate substitutions are made increased up to 40%. These studies clearly demonstrate that phosphate neutralization causes increased flexibility of the DNA duplex and can also result in a “bent” DNA. However, its effect on basepair step and basepair deforming remains relatively less studied. It is known that proteins (such as EcoRI) make backbone contacts with the DNA and that backbone contacts mainly involve contacts with the DNA phosphate groups and their neutralization. Hence it is tempting to hypothesize that neutralization of the phosphate can result in conformational differences at the basepair level. We propose an investigation, using quantum mechanical calculations, of the effect of phosphate neutralization on basepairs with a focus on the strength of the hydrogen bonds between them and the structural changes. Phosphate neutralization is generally mimicked by methylphosphonate substitutions as depicted in Figure 6-2 [173].

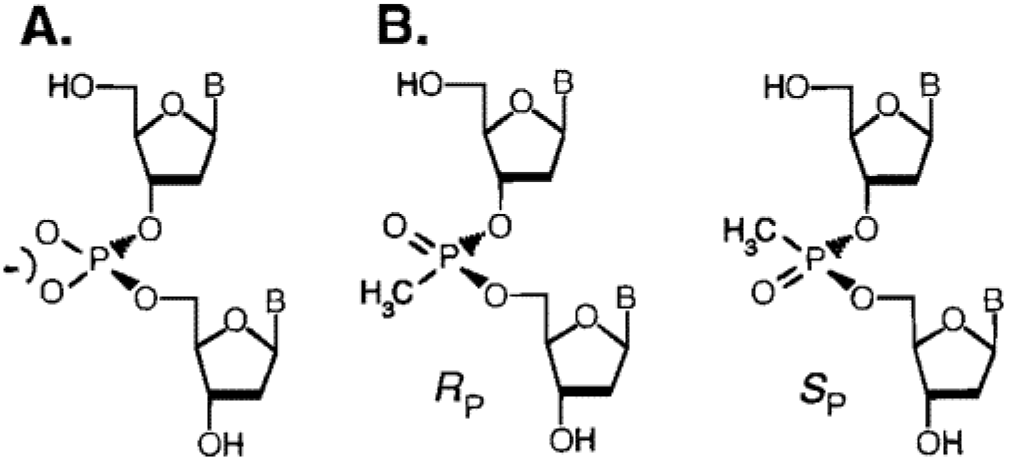


Figure 6-2. Phosphate linkages between the bases (A) are generally substituted with methylphosphonate to mimick neutralization (B). Picture adopted from [173]

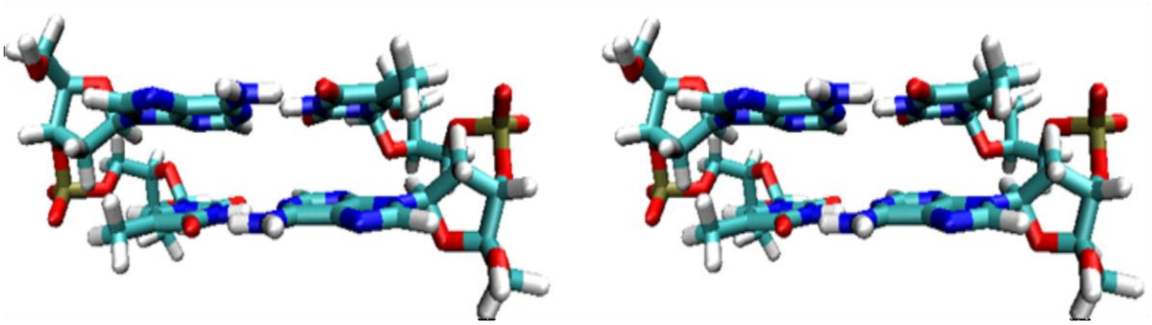


Figure 6-3. Stereoscopic view of a typical basepair step.

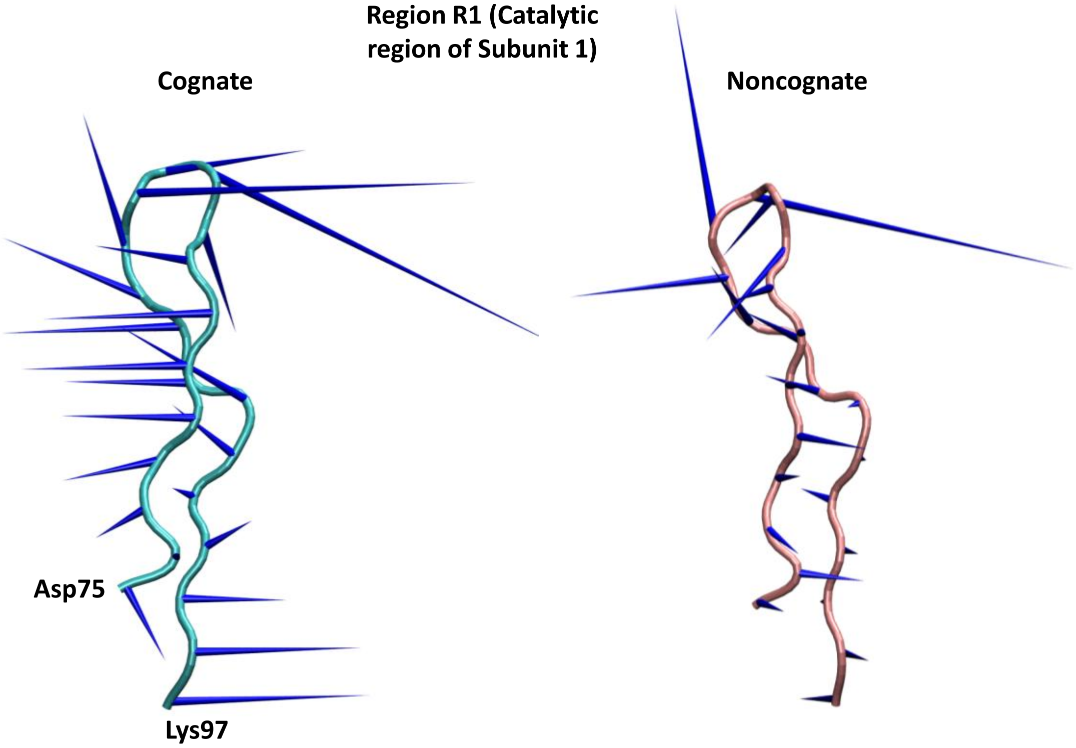
Thus, in basepair steps such as those depicted in Figure 6-3, methylphosphonate substitutions can be done. The resulting structures can be subjected to single-point energy calculations and geometry optimizations using the Gaussian09 suite of programs[4]. Density functional theory calculations using the M062X exchange correlation functional [174] has been shown to perform well to capture the noncovalent interactions in biomolecules [175, 176]. Hence, the M062X functional shall be used for all these calculations. A comparison of hydrogen bonding energies and conformations with and without methylphosphonate substitutions will then reveal possible influence of protein via phosphate neutralization. Further, the same analysis with different basepair step sequences will reveal sequence-dependent behavior. Thus, the study will provide an explanation for the sequence-dependent changes observed in the work presented in the thesis.

The studies recommended here primarily focus on dissecting and investigating each of the parameters, namely, water dynamics, protein dynamics and the influence of the protein on DNA. This thesis work, recommendations described and other ongoing studies on the diffusion of proteins along the DNA and the kinetics of their transition-states that are being carried out by researchers all over the world, will lead to a better understanding of the mechanisms of protein-DNA recognition.

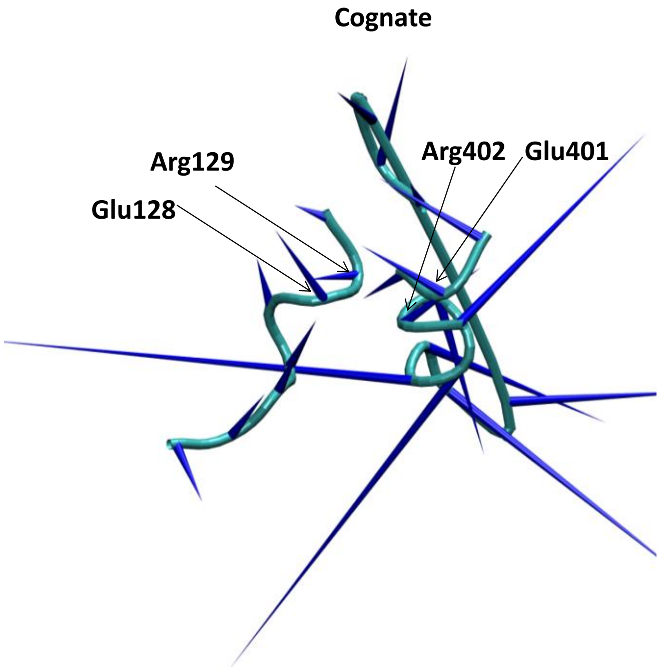
[4] http://www.gaussian.com/index.htm (Hyperlink checked as of 22 December 2011)

Appendix A: Porcupine Plots of various regions in EcoRI

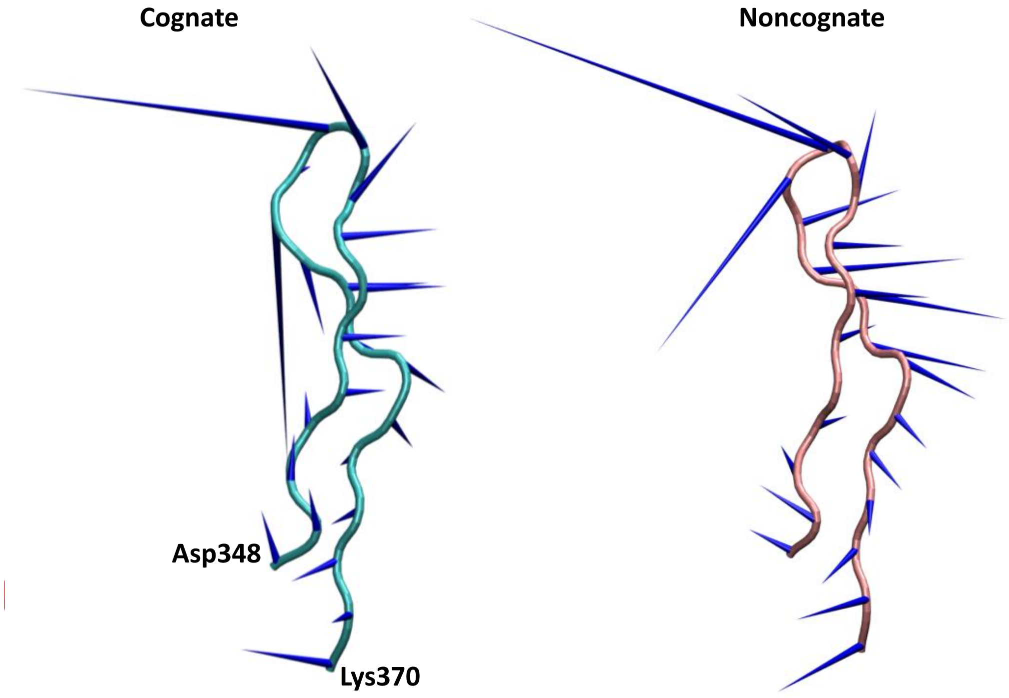
Figure A-1. Porcupine Plots showing the motion of Regions R1, R2, R3, R5 and R6 along the first principal component. The porcupine plot showing the motion of region R4 is presented in Figure 3-6.

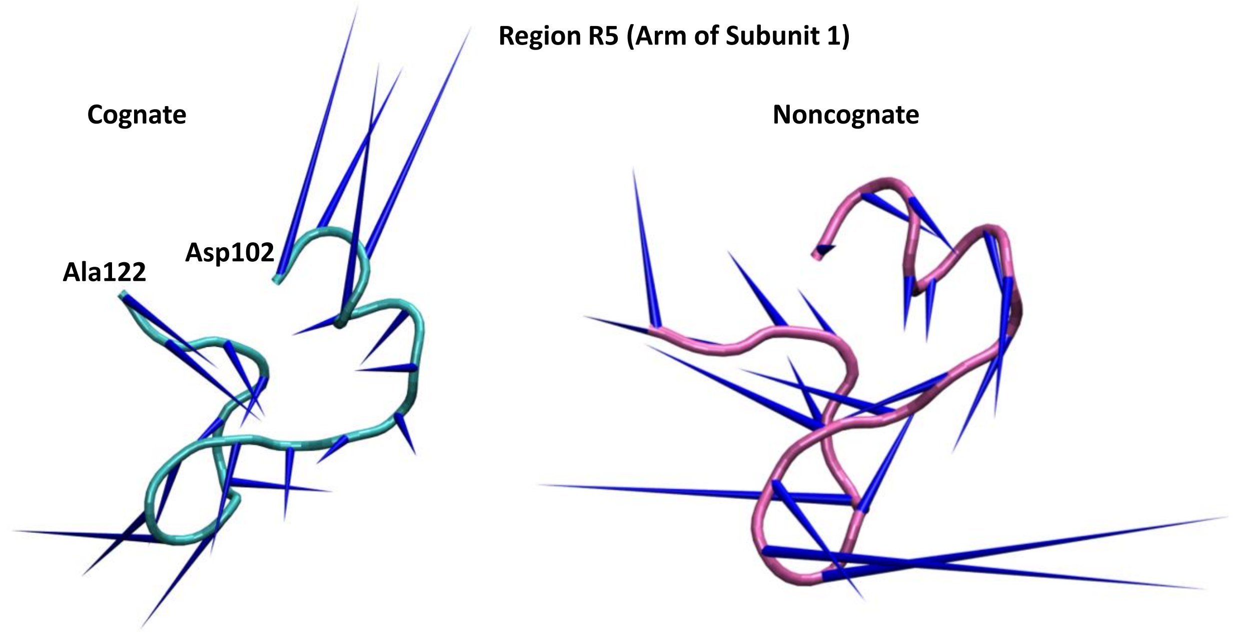


Region R2 (Crosstalk Region)



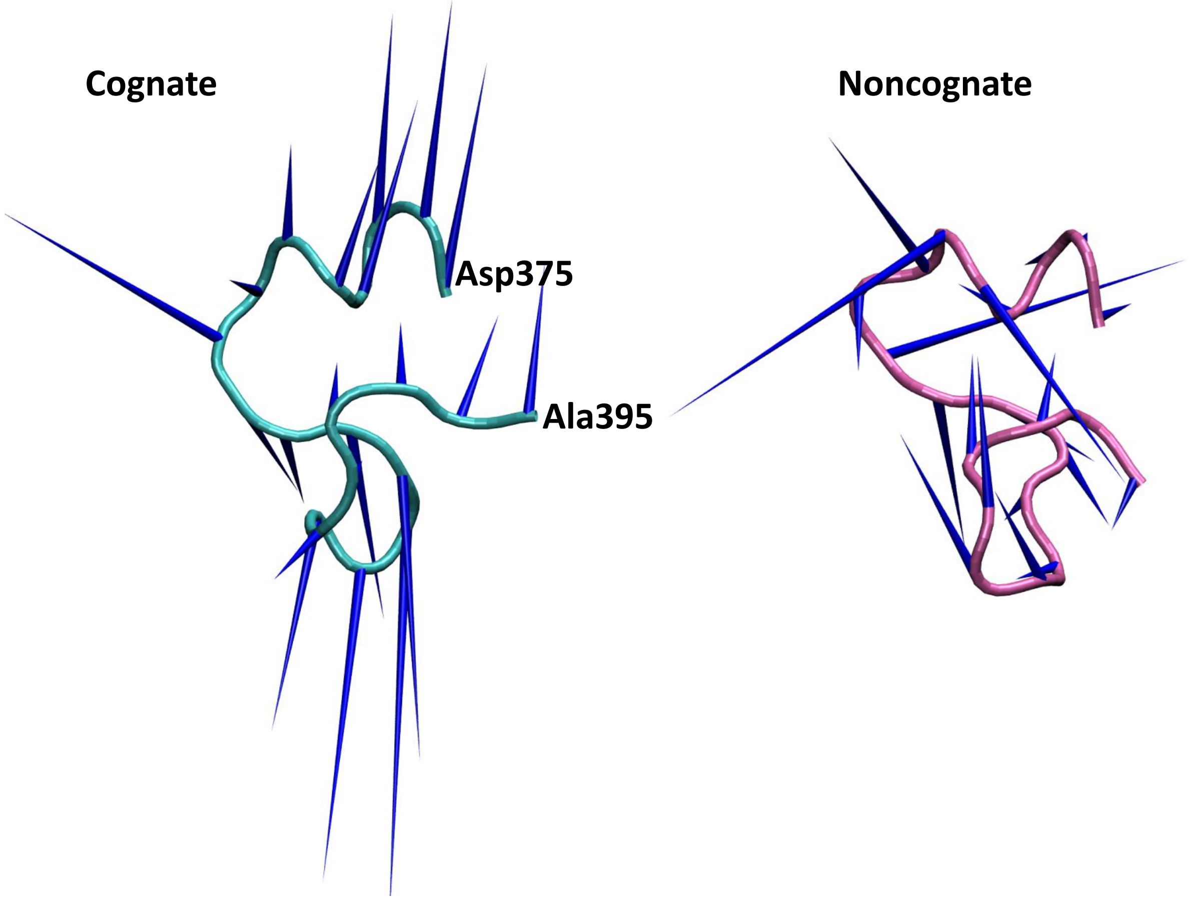
Region R3 (Catalytic Region of Subunit 2)





Region R5 (Arm of Subunit 1)

Region R6 (Arm of Subunit 2)



Appendix B: The Two-phase Thermodynamic (2PT) Theory

Molecular simulations for the numerical determination of the thermodynamic, energetic, structural, and dynamic properties of a mathematical model of a molecular assembly [177] have proved helpful for verifying theoretical models and/or in the interpretation of experimental observations based on theoretical models. Diverse chemical and biomolecular systems such as water [178, 179], protein [180, 181], nucleic acids [53, 182], and membranes ([86, 183] have been investigated using molecular simulations providing several useful insights. Since, according to the Second Law of Thermodynamics, natural systems achieve, at equilibrium or seek out from nonequilibrium position, a state of minimum free energy, calculation of free energy forms a central component in comparing theory and experiment [177] in addition to providing a conceptual framework for understanding the physico-(bio)chemical process. Thus it has been of outstanding interest to calculate free energy from molecular simulations which serve as a “bridge” between theory and experiment. However, the problem in calculating free energy from molecular simulations stems from the difficulty in obtaining accurate estimates of entropy. Conventional simulation methods (Monte Carlo or molecular dynamics) sample only small part of the entire configurational space that is accessible to a molecule [184, 185]. Thus, for reliable estimation of entropy of macromolecules, longer simulations exploring the entire configurational space are necessary, or, is sometimes intractable. Several methods have been proposed and investigated to estimate entropy reliably from molecular simulations. Gō and Scheraga introduced the harmonic approximation method of calculating the entropy to biomolecules [186, 187]. This was then followed by the quasi-harmonic (QH) approximation method introduced by Karplus and Kushick [188] in which the covariances (of the atomic coordinates) are assumed to be resulting from a harmonic energy surface rather than the actual anharmonic energy surface, and the “quasiharmonic” force constants are calculated on this basis. These force constants are then used to compute the thermodynamic properties such as entropy and free energy [189]. The accuracy of the quasi-harmonic approximation method has been thoroughly evaluated recently [189]. The entropy estimated based on the quasi-harmonic approximation constitutes the upper bound for entropy because of the neglect of correlations higher than quadratic, and QH method ignores anharmonic contributions and hence not suitable for diffusive systems such as water [190].

In the case of diffusive systems, particularly water, several methods have been investigated to calculate the absolute entropy of water from molecular dynamics simulations. For example, the hypothetical scanning method proposed by White and Meirovitch [191, 192] determines the absolute entropy and free energy from the Boltzmann probability distribution. Use of molecular pair correlation functions to obtain entropy was proposed by Lizaridis and Karplus [193] and was further improvised by Wang et al. [194]. Tyka et al. [195] proposed the confinement method to determine the absolute entropy using thermodynamic integration from a hypothetical harmonic state to the liquid state. Based on cell theory, Henchman [179] estimated entropy within the harmonic approximation in the potential surface. Lin et al. [144] claim that these methods have been quite successful in calculating the entropy and free energy of water but only under limited conditions and that the harmonic approximations in some of the methods questions the reliability because anharmonic effects are important in diffusive systems.

Accounting Anharmonicity in Entropy Calculations: Two-phase Thermodynamic Theory (2PT Theory)

To overcome the difficulty in incorporating the anharmonic effects due to diffusive motion in the entropy calculations of liquids, Lin et al. [143] proposed a “two-phase” thermodynamic model (2PT model based on separating the diffusional contributions from the vibrational contributions. One can see from the typical density of state distribution (defined as the density of normal modes of vibrations of a system [196];) for gas, liquid and a solid (Fig.B-1) that the

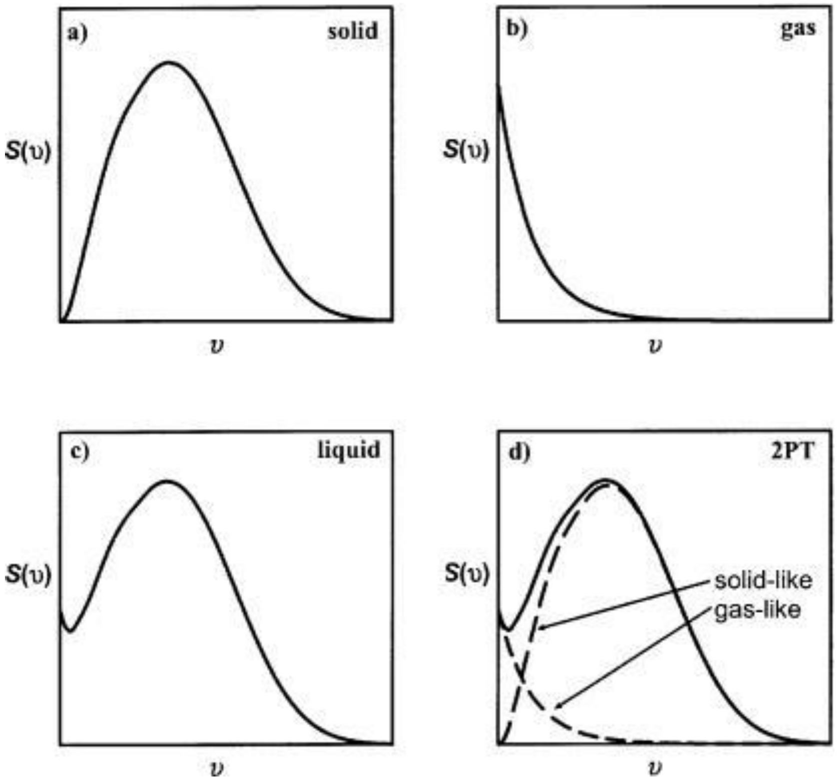


Figure B-1. Typical density of states (denoted here as $S(\upsilon)$) for solid (a), gas (b) and liquid (c). (d) shows 2PT model. (Figure adopted from [143]).

the density of states of a liquid can be supposed to be the sum of the gas-like and solid-like components (see Fig. B-1d) with appropriate weighing of these components. The gas-like component accounts for the diffusive motions and the solid-like component accounts for the vibrational motions. Lin et al. [143] show that the entropy of a liquid can be accurately calculated from the density of states thus constructed.

B.1 Canonical Partition Function, Density of States and the Thermodynamic Variables: The Overall Conceptual Framework

The total canonical partition function Q is related to the velocity autocorrelation function via the density of states (defined in the next section). The velocity autocorrelation functions are straightforward to obtain from the molecular dynamics trajectories. In the following sections we show how the total canonical partition function, the density of states, and thermodynamic variables are related. This is followed by a discussion of how the density of states can be constructed from the velocity autocorrelation function and a discussion of the 2PT theory, which prescribes a method to account for the fluidicity effects in the construction of the density of states of liquids.

B.2 The Canonical Partition Function and the Density of States

In this section we show how the canonical partition function and the density of states are related. Consider a system of N atoms linked by harmonic potentials (or, a system of N harmonic oscillators). The collective motion of the atoms (or the N harmonic oscillators) can be decomposed into independent constituents. These independent constituents are called the normal modes of vibration of the system (and the corresponding frequency is called the normal frequency). The total canonical partition Q for the system can then be expressed in terms of the partition functions q_j of the individual normal modes as

$$Q = \prod_{j=1}^{3N} q_j \qquad [7\text{-}1]$$

Or, taking logarithm gives

$$\ln Q = \sum_{j=1}^{3N} \ln q_j \qquad [7\text{-}2]$$

If the normal frequencies are continuously distributed, then one can write the above in terms of the integral

$$\ln Q = \int_0^{\infty} g(\upsilon) d\upsilon \ln q(\upsilon) \qquad [7\text{-}3]$$

where $g(\upsilon)$ is the density of the normal modes with frequency υ and is called the density of states (DoS) or the spectral density.

The thermodynamic variables (Internal energy E, Entropy S and Helmholtz free energy A) can then be obtained for a canonical ensemble from the above partition function as:

$$E = V_0 + \beta^{-1}\left(\frac{\partial \ln Q}{\partial T}\right)_{N,V} = V_0 + \beta^{-1}\int_0^{\infty} g(\upsilon) W_E(\upsilon) d\upsilon \qquad [7\text{-}4]$$

$$S = k \ln Q + \beta^{-1}\left(\frac{\partial \ln Q}{\partial T}\right)_{N,V} = k\int_0^{\infty} g(\upsilon) W_S(\upsilon) d\upsilon \qquad [7\text{-}5]$$

$$A = V_0 + \beta^{-1} \ln Q = V_0 + \beta^{-1} \int_0^{\infty} g(\upsilon) W_A(\upsilon) d\upsilon \qquad [7\text{-}6]$$

where W$_X$ (X = E, S, or A) are the weighting functions, $\beta = 1/kT$, k is the Boltzmann constant and T is the temperature. Substituting in these equations the quantum harmonic partition function (q_{HO}^{Q})

$$q_{HO}^{Q}(\upsilon) = \frac{\exp\left(\frac{-\beta h\upsilon}{2}\right)}{1 - \exp\left(\frac{-\beta h\upsilon}{2}\right)} \qquad [7\text{-}7]$$

where h is the Planck's constant, gives the quantum weighting functions [143] $W_X^Q (X = E, S \text{ or } A)$:

$$W_E^Q(\upsilon) = \frac{\beta h\upsilon}{2} + \frac{\beta h\upsilon}{\exp(\beta h\upsilon) - 1} \qquad [7\text{-}8]$$

$$W_s^Q(\upsilon) = \frac{\beta h\upsilon}{\exp(\beta h\upsilon) - 1} - \ln[1 - \exp(-\beta h\upsilon)] \qquad [7\text{-}9]$$

$$W_A^Q(\upsilon) = \ln \frac{1 - \exp(\beta h\upsilon)}{\exp(-\frac{\beta h\upsilon}{2})} \qquad [7\text{-}10]$$

In the next section, we discuss how to construct the density of states.

B.3 Constructing the Density of States

As mentioned earlier, from Equation [7-3], the density of states is the density of normal modes of vibration of a system. The density of normal modes is equal to the velocity spectrum[196]. By velocity spectrum one refers to the probability distribution of the various velocities, and this is given by the square of the Fourier transformation of the velocity time history.

For instance, the spectral density (or the density of states), $g_j^k(\upsilon)$ of atom *j* in the *k*th coordinate is given as

$$
\begin{aligned}
g_j^k(\upsilon) &= \lim_{T\to\infty} \frac{\left| \int_{-T}^{T} v_j^k(t) e^{-i2\pi\upsilon t} dt \right|^2}{\int_{-T}^{T} dt} \\
&= \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} v_j^k(t) e^{-i2\pi\upsilon t} dt \right|^2 \qquad [7\text{-}11] \\
&= \lim_{T\to\infty} \frac{1}{2T} \left| F(\upsilon) \right|^2
\end{aligned}
$$

where $v_j^k(t)$ is the *k*th velocity component of atom *j* at time *t*.

Although the spectral density can, in principle, be calculated from the velocity time history, a more convenient approach is to obtain the spectral density from the velocity autocorrelation functions. For a given stochastic process, one can show that the spectral density is equal to the Fourier transform of the velocity autocorrelation function as follows [197].

We start with the definition of the velocity autocorrelation function:

$$
c_T(\tau) = \frac{1}{2T} \int_{-T}^{T} x_T(t) x_T(t+\tau) dt \qquad [7\text{-}12]
$$

so that

$$c(\tau) = \lim_{T \to \infty} c_T(\tau) \qquad [7\text{-}13]$$

$x_T(t)$ is the coordinate of a molecule at time t. The Fourier transform of $c_T(\tau)$ is

$$\begin{aligned}
\int_{-\infty}^{\infty} c_T(\tau) e^{-i2\pi\upsilon\tau} d\tau &= \frac{1}{2T} \int_{-\infty}^{\infty} e^{-i2\pi\upsilon\tau} \int_{-\infty}^{\infty} x_T(t) x_T(t+\tau) dt\, d\tau \\
&= \frac{1}{2T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_T(t) x_T(t+\tau) e^{-i2\pi\upsilon(t+\tau)} e^{i2\pi\upsilon t} dt\, d\tau \\
&= \frac{1}{2T} \int_{-\infty}^{\infty} x_T(t) e^{i2\pi\upsilon\tau} \int_{-\infty}^{\infty} x_T(t+\tau) e^{-i2\pi\upsilon(t+\tau)} dt\, d\tau \qquad [7\text{-}14] \\
&= \frac{1}{2T} F(\upsilon) F(-\upsilon) \\
&= \frac{1}{2T} \left| F(\upsilon) \right|^2
\end{aligned}$$

Passing to the limit $T \to \infty$ and using Equation [7-11] gives,

$$\int_{-\infty}^{\infty} c(t) e^{-i2\pi\upsilon\tau} d\tau = g_j^k(\upsilon) \qquad [7\text{-}15]$$

The density of states distribution of the whole system at a particular frequency, $g(\upsilon)$, is then given as the sum of the contributions from all atoms in the system. That is,

$$g(\upsilon) = \frac{2}{kT} \sum_{j=1}^{N} \sum_{k=1}^{3} m_j g_j^k(\upsilon) \qquad [7\text{-}16]$$

where m_j is the mass of atom j. Substituting Equation [7-15] in Equation [7-16],

$$\begin{aligned} g(\upsilon) &= \frac{2}{kT}\int_{-\infty}^{\infty}\sum_{j=1}^{N}\sum_{k=1}^{3} m_j c_j^k(t) e^{-i2\pi\upsilon t} dt \\ &= \frac{2}{kT}\int_{-\infty}^{\infty} C(t) e^{-2\pi i\upsilon t} dt \end{aligned} \qquad [7\text{-}17]$$

where $c_j^k(t)$ is the velocity autocorrelation function of atom j in the k th coordinate and

$C(t) = \sum_{j=1}^{N}\sum_{k=1}^{3} m_j c_j^k(t)$.

Thus, once the velocity spectrum or the velocity autocorrelation function is calculated, one can calculate the partition function of the system. The thermodynamic properties of the canonical ensemble can then be obtained from Equations [7-4], [7-5] and [7-6].

The zero-frequency density of state value $g(\upsilon = 0)$ is related to the diffusion coefficient as follows:

$$D = \frac{1}{3}\int_{0}^{\infty} c(t)dt = \frac{1}{6}\int_{-\infty}^{\infty} c(t)dt = \frac{1}{6mN}\int_{-\infty}^{\infty} C(t)dt \qquad [7\text{-}18]$$

where N is the number of particles and m is the mass of the particle. By setting the frequency υ to zero in Equation [7-17] and using Equation [7-18] we get

$$g(0) = \frac{2}{kT}\int_{-\infty}^{\infty} C(t)dt = \frac{12mND}{kT} \qquad [7\text{-}19]$$

B.4 The So-called Two-Phase Theory to Account for the Fluidicity of Liquids in DoS calculation

The above formulation describes how one can calculate the density of states from the velocity autocorrelation function and hence the thermodynamic variables, assuming the system to be a set of harmonic oscillators. Extending the above formulation for diffusive systems is non-trivial for the following reasons. The typical density of state distribution of a solid has the form $g(0) = 0$ with $g(\upsilon)$ going through a maximum at finite υ and then decaying at higher frequencies. For a gas, the DoS is nonzero at zero frequency (i.e. $g(0) > 0$) and decays monotonically. The DoS of a liquid also is nonzero at zero frequency ($g(0) > 0$) leading to a local minimum at low frequency and a maximum at a finite υ (similar to the DoS of a solid) and then it decays for further higher frequencies. Due to the nonzero zero-frequency value of density of state for liquids and gases, use of the quantum weighting function of harmonic oscillators for entropy (Equation [7-9]) will result in infinite entropy. Moreover, since the low-frequency vibrations are anharmonic, the harmonic approximation is not valid [143]. These properties of the fluids, i.e., the nonzero $g(0)$ and the anharmonicities, are generally referred to as the fluidicity effects [143] and limit the use of density of states approach to calculating the thermodynamic properties of fluids. To account for the fluidicity effects in the density of states approach, Lin et al.[143] proposed a model in which the density of states of the intermediate, “liquid-like” system can be partitioned into a gas-like ($g^{g}(\upsilon)$) and a solid-like component ($g^{s}(\upsilon)$). That is,

$$g(\upsilon) = g^{g}(\upsilon) + g^{s}(\upsilon) \qquad [7\text{-}20]$$

The gas-like component, contributing mostly at the low frequency range, contains all the fluidic effects and the solid-like component, contributing mostly at higher frequencies, incorporates the quantum effects [143]. The thermodynamic properties P of the system can then be determined by using the appropriate quantum weighting functions (W_P^s and W_P^g) for each of the component as:

$$P = \int_0^\infty g^s(\upsilon) W_P^s(\upsilon) d\upsilon + \int_0^\infty g^g(\upsilon) W_P^g(\upsilon) d\upsilon \qquad [7\text{-}21]$$

The task now is to construct the gas-like and the solid-like components of the density of states (the intermediate case can be then calculated by using Equation [7-20]). In the 2PT model, the gas-like component of the density of states is constructed by taking the gas to be a hard-sphere fluid. In the next section we outline how the gas-like component is constructed.

B.5 Constructing the Gas-like Component of the Density of States

As mentioned above, in the 2PT model, the gas-like component is taken to be a hard-sphere fluid. The density of states of the hard-sphere fluid is then given by the Fourier cosine transformation of the velocity autocorrelation function. The velocity autocorrelation function $c^{HS}(t)$ of a hard-sphere gas is given as [197]

$$c^{HS}(t) = c^{HS}(0)\exp(-\alpha t) = \frac{3kT}{m}\exp(-\alpha t) \qquad [7\text{-}22]$$

where α is the Enskog friction constant and m is the mass of the molecule. The Fourier cosine transformation of Equation [7-22], which is the density of states distribution of the gas, is

$$\begin{aligned} g^{g}(\upsilon) = g^{HS}(\upsilon) &= \frac{4}{kT}\int_{0}^{\infty}\sum_{j=1}^{N^{g}}\sum_{k=1}^{3} m_{j} c_{j}^{HS,k}(t)\cos(2\pi\upsilon t)\,dt \\ &= \frac{4}{kT}\int_{0}^{\infty} 3N^{g}kT\,\exp(-\alpha t)\cos(2\pi\upsilon t) \qquad [7\text{-}23] \\ &= \frac{12N^{g}\alpha}{\alpha^{2}+4\pi^{2}\upsilon^{2}} \end{aligned}$$

where $N^{g} = fN$ is the effective number of hard-sphere particles in the system and f is the fraction of hard-sphere component in the overall system. The fraction f is a measure of the fluidicity of a system (and called the fluidicity factor) and depends on both the temperature (T) and density (ρ). At zero frequency, from Equation [7-23] we have

$$g^{HS}(0) = g_{0} = \frac{12\,fN}{\alpha} \qquad [7\text{-}24]$$

Using Equation [7-24], Equation [7-23] can be rewritten as

$$g^{HS}(\upsilon) = \frac{g_{0}}{1+\left[\dfrac{\pi g_{0}\upsilon}{6\,fN}\right]^{2}} \qquad [7\text{-}25]$$

Taking g_{0} as the zero-frequency DoS value of the *system* guarantees that the solid component has no contribution to the diffusivity. Now, the only remaining parameter to be determined is f, the fluidicity factor, which should satisfy the following two limiting conditions to represent the conceptual partitioning of the system into solid and gas components:

(1) At high temperatures or low density, $f = 1$ That is, there is no solid component.

(2) At high density, (i.e. when the system is a solid), $f = 0$.

Lin et al. [143] define f as proportional to the diffusion coefficients so that the above two conditions are satisfied. Thus we have

$$f = \frac{D(T,\rho)}{D_0^{HS}(T,\rho;\sigma_{HS})} \qquad [7\text{-}26]$$

where D is the self-diffusion coefficient of the *system* (determined from the zero-frequency value of the density of states) and D_0^{HS} is the diffusion coefficient of the hard-sphere gas determined in the zero-pressure limit (Chapman-Enskog theory) [197] (σ_{HS} is the diameter of the hard sphere particle) and is given as:

$$D_0^{HS}(T,\rho;\sigma_{HS}) = \frac{3}{8}\frac{1}{\rho\sigma_{HS}^2}\left(\frac{kT}{\pi m}\right)^{1/2} \qquad [7\text{-}27]$$

At this stage, one needs only to define σ_{HS} to construct the density of states distribution of the gas-like component. Lin et al. [143] bypass defining σ_{HS} by developing a "universal equation" starting with the Enskog theory that predicts the deviation of diffusivity of a dense hard-sphere fluid from its zero-pressure limit. According to the Enskog theory, the deviation of diffusivity of a dense hard-sphere fluid from its zero-pressure limit is given as

$$D^{HS}(T, f\rho) = D_0^{HS}(T, f\rho; \sigma^{HS}) \frac{4fy}{z(fy)-1} \qquad [7\text{-}28]$$

where z is the compressibility obtained from the Carnahan-Starling Equation of State for a hard-sphere fluid [198] as

$$z(y) = \frac{1+y+y^2-y^3}{(1-y)^3} \qquad [7\text{-}29]$$

In the above equation y is the hard-sphere packing fraction defined as $y = (\pi/6)\rho\sigma_{HS}^3$.

Here, in order to simplify the notations, we define the rescaled volume fraction ϕ as

$$\phi = fy = f((\pi/6)\rho\sigma_{HS}^3) \qquad [7\text{-}30]$$

From the velocity autocorrelation function in Equation [7-22] one can obtain the diffusivity of the hard sphere fluid at Temperature T and density $f\rho$ as

$$D^{HS}(T, f\rho) = \frac{1}{3}\int_0^\infty c^{HS}(t) = \frac{kT}{m\alpha} = \frac{kTg_0}{12mfN} \qquad [7\text{-}31]$$

where Equation [7-24] has been used for α .

Combining Equations [7-19], [7-25] and [7-31], we obtain $D^{HS}(T, f\rho) = D(T,\rho)/f$.

Further, since in the zero-pressure limit $D_0^{HS}(T, f\rho) = D_0^{HS}(T,\rho)/f$, from Equations [7-22], [7-28] to [7-30], one gets a cubic equation for f in terms of ϕ :

$$2\phi^3 - 6\phi^2 + 6\phi + (2-\phi)f - 2 = 0 \qquad [7\text{-}32]$$

$$\Rightarrow\ 2(\phi-1)^3+(2-\phi)f=0$$

$$\Rightarrow\ f=\frac{2(\phi-1)^3}{\phi-2} \qquad [7\text{-}33]$$

with $f\to 1$ as $\phi\to 0$ (no solid component) and $f\to 0$ as $\phi\to 1$ (no gas component). Also, by substituting Equation [7-27] in Equation [7-26] and rewriting the resulting equation in terms of ϕ, one gets

$$f^{5/3}=\Delta\phi^{2/3}\quad(\Rightarrow\ \phi=f^{5/2}\Delta^{-3/2}) \qquad [7\text{-}34]$$

Or

$$f=\Delta^{3/5}\phi^{2/5} \qquad [7\text{-}35]$$

with

$$\Delta=\Delta(T,\rho,m,g_0)=\frac{2g_0}{9N}\left(\frac{\pi kT}{m}\right)^{1/2}\rho^{1/3}\left(\frac{6}{\pi}\right)^{2/3} \qquad [7\text{-}36]$$

The normalized diffusivity Δ is proportional to the system diffusivity which in turn includes the effects of temperature, density and different material properties. Substituting Equation [7-35] in Equation [7-32] one obtains a universal expression for f in terms of Δ:

$$2\phi^3-6\phi^2+6\phi+(2-\phi)\Delta^{3/5}\phi^{2/5}-2=0 \qquad [7\text{-}37]$$

For a given Δ, Equation [7-37] gives the effective volume fraction ϕ, which, in combination with Equation [7-35], gives f.

Alternatively, one can use Equation [7-35] written for ϕ in terms of f, i.e.,

$$\phi = f^{5/2}\Delta^{-3/2} \qquad [7\text{-}38]$$

in Equation [7-32] to get the universal expression for f in terms of Δ developed by Lin et al. [143]

$$2\Delta^{-9/2} f^{15/2} - 6\Delta^{-3} f^{5} + 6\Delta^{-3/2} f^{5/2} - \Delta^{-3/2} f^{7/2} + 2f - 2 = 0 \qquad [7\text{-}39]$$

From Equation [7-37], one sees that as $\Delta \to 0,\ \phi = 1$ and as $\Delta \to \infty,\ \phi = 0 \text{ or } \phi = 2$ (physically unacceptable). Therefore, from Equation [7-35], one has $\Delta \to 0, \phi \to 1 \Rightarrow f \to 0$ and $\Delta \to \infty, \phi \to 0 \Rightarrow f \to 1$. Having obtained f, one can now construct the density of states accounting for the fluidicity effects. Once the density of states is constructed, one can use Equations [7-4] to [7-6] to obtain the thermodynamic variables. The weighting functions for the gas component (hard-sphere fluid) are:

$$W_E^{g}(\upsilon) = W_E^{HS}(\upsilon) = 0.5 \qquad [7\text{-}40]$$

$$W_S^{g}(\upsilon) = W_S^{HS}(\upsilon) = \frac{1}{3}\frac{S^{HS}}{k} \qquad [7\text{-}41]$$

$$W_A^{g}(\upsilon) = W_A^{HS}(\upsilon) = W_E^{HS}(\upsilon) - W_S^{HS}(\upsilon) \qquad [7\text{-}42]$$

where S^{HS} is the hard-sphere entropy and is given as

$$\frac{S^{HS}}{k} = \frac{5}{2} + \ln\left[\left(\frac{2\pi m k T}{h^2}\right)^{3/2} \frac{V}{fN} z(\phi)\right] + \frac{y(3\phi - 4)}{(1-\phi)^2} \qquad [7\text{-}43]$$

where ϕ is obtained from Equation [7-38] and $z(\phi)$ is the compressibility factor from the Carnahan-Starling equation of state of hard-sphere gases [198] given in Equation [7-29].

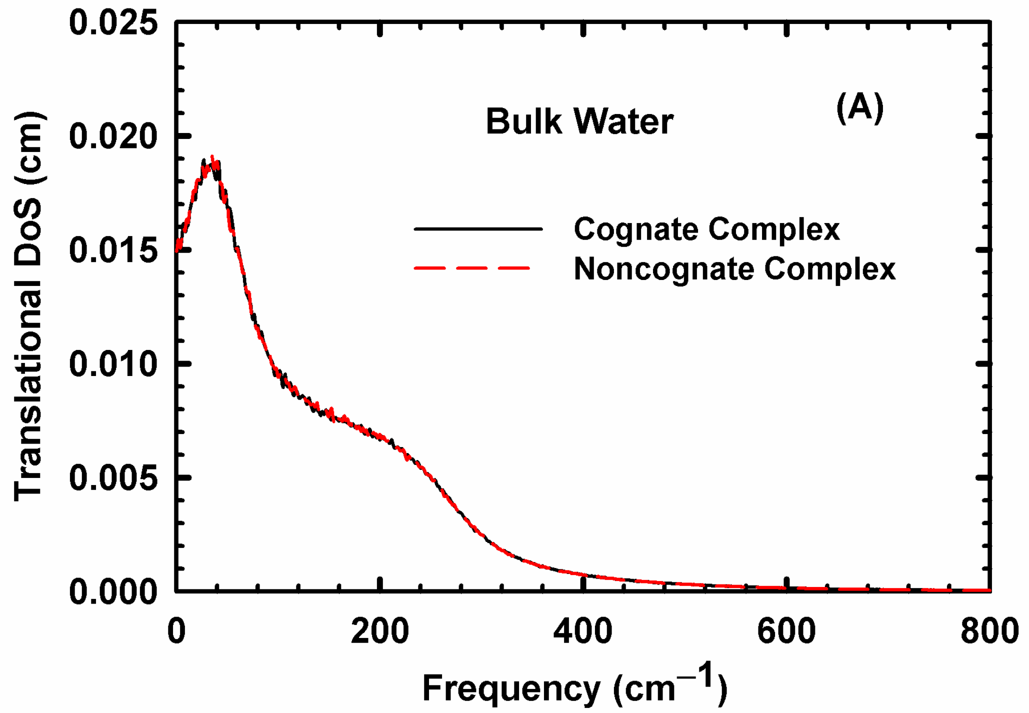
Thus the thermodynamic variables can be estimated from the velocity autocorrelation functions obtained from molecular dynamic trajectories using the 2PT model.

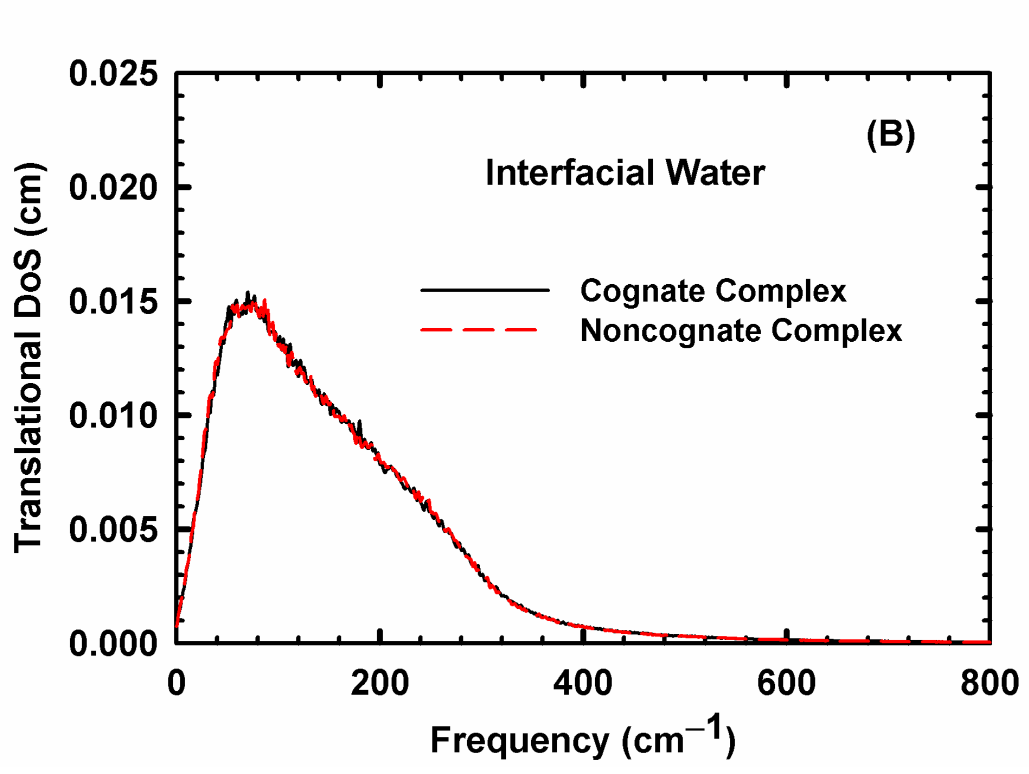
In summary, the 2PT theory can be summarized as follows:

1. Calculate the velocity autocorrelation function for the molecules of interest.
2. Fourier transform the velocity autocorrelation to obtain the density of states distribution and obtain the zero-frequency value, g_0.
3. Calculate the fluidicity factor (Equations [7-36] and [7-39]).
4. Obtain the gas component of the density of states distribution (Equation [7-25]).
5. Obtain the solid component of the density of states distribution (Equation [7-20]).
6. Using appropriate weighting functions, calculate the needed thermodynamic variables (Equations [7-4] to [7-6], [7-8] to [7-10] and [7-40] to [7-42]).

Appendix C: Comparison of Density of States Spectra of Water in Cognate and Noncognate Complexes

Figure C-1. Comparison of the translational density of states spectrum of bulk (A), interface (B) and intercalating waters (C) in the cognate and noncognate complexes.




(B)
Interfacial Water
Cognate Complex
Noncognate Complex
Translational DoS (cm)
0.025
0.020
0.015
0.010
0.005
0.000
0
200
400
600
800
Frequency (cm^{-1})

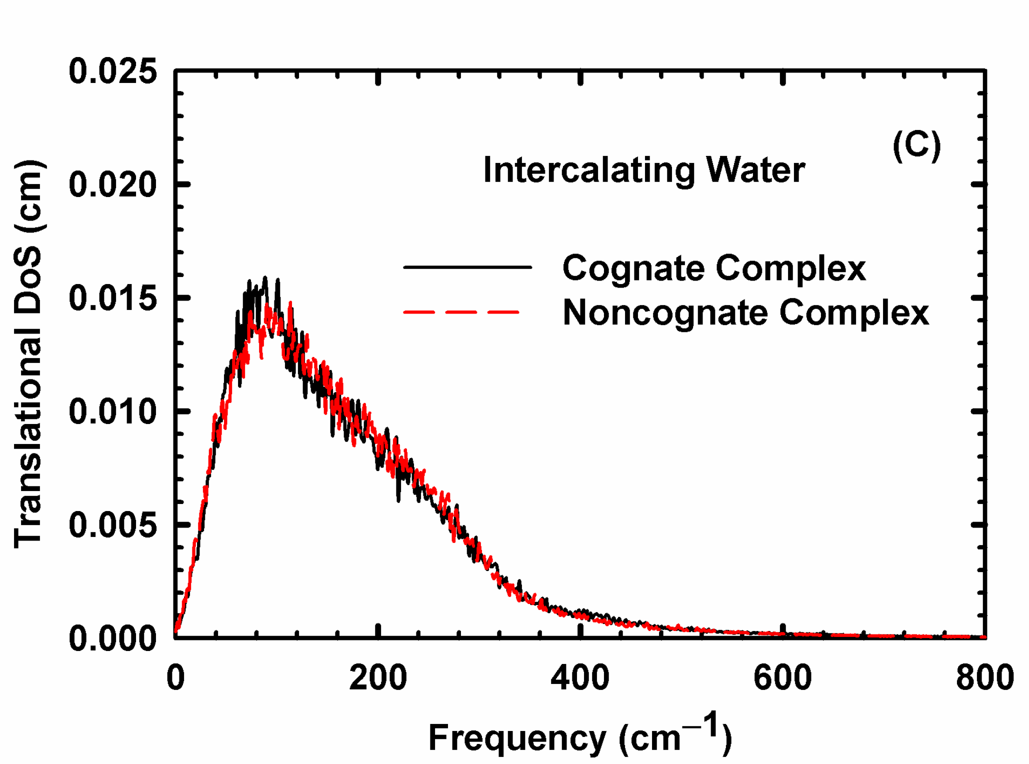
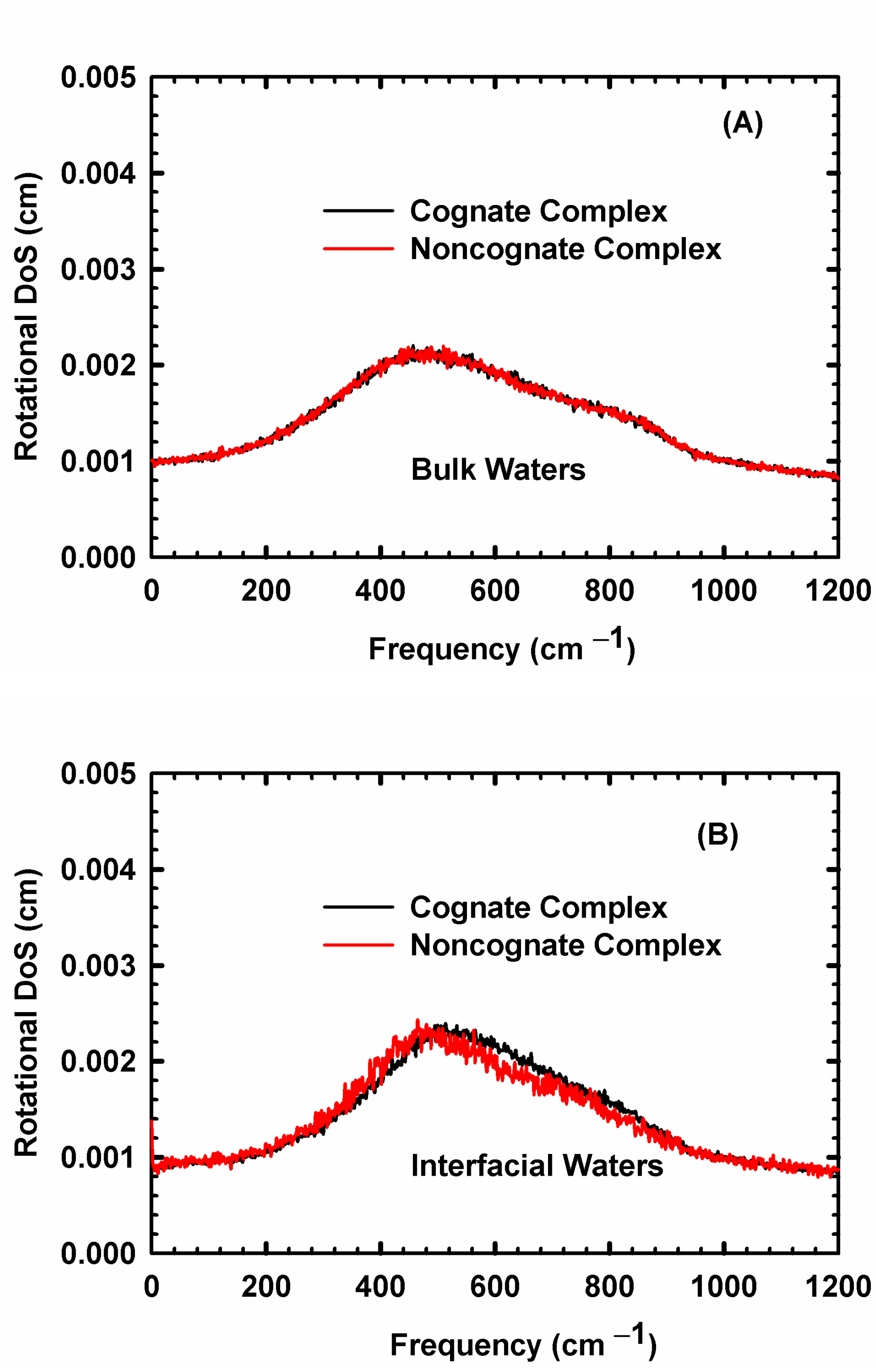
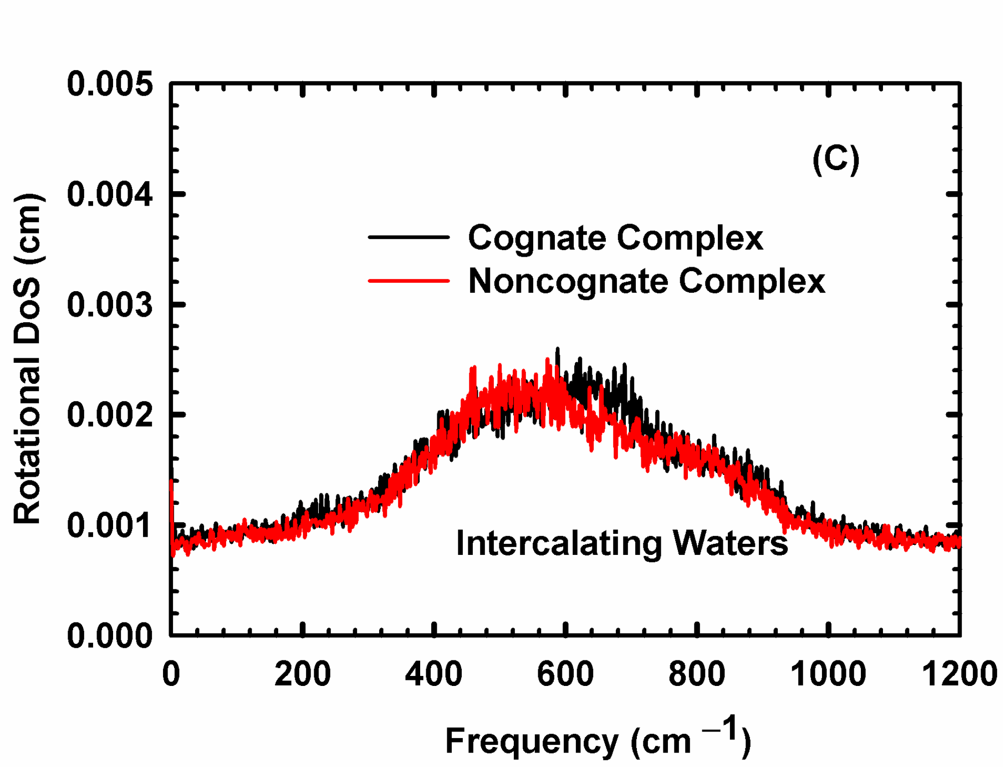

(C)
Intercalating Water
Cognate Complex
Noncognate Complex
Translational DoS (cm)
0.025
0.020
0.015
0.010
0.005
0.000
0
200
400
600
800
Frequency (cm^{-1})

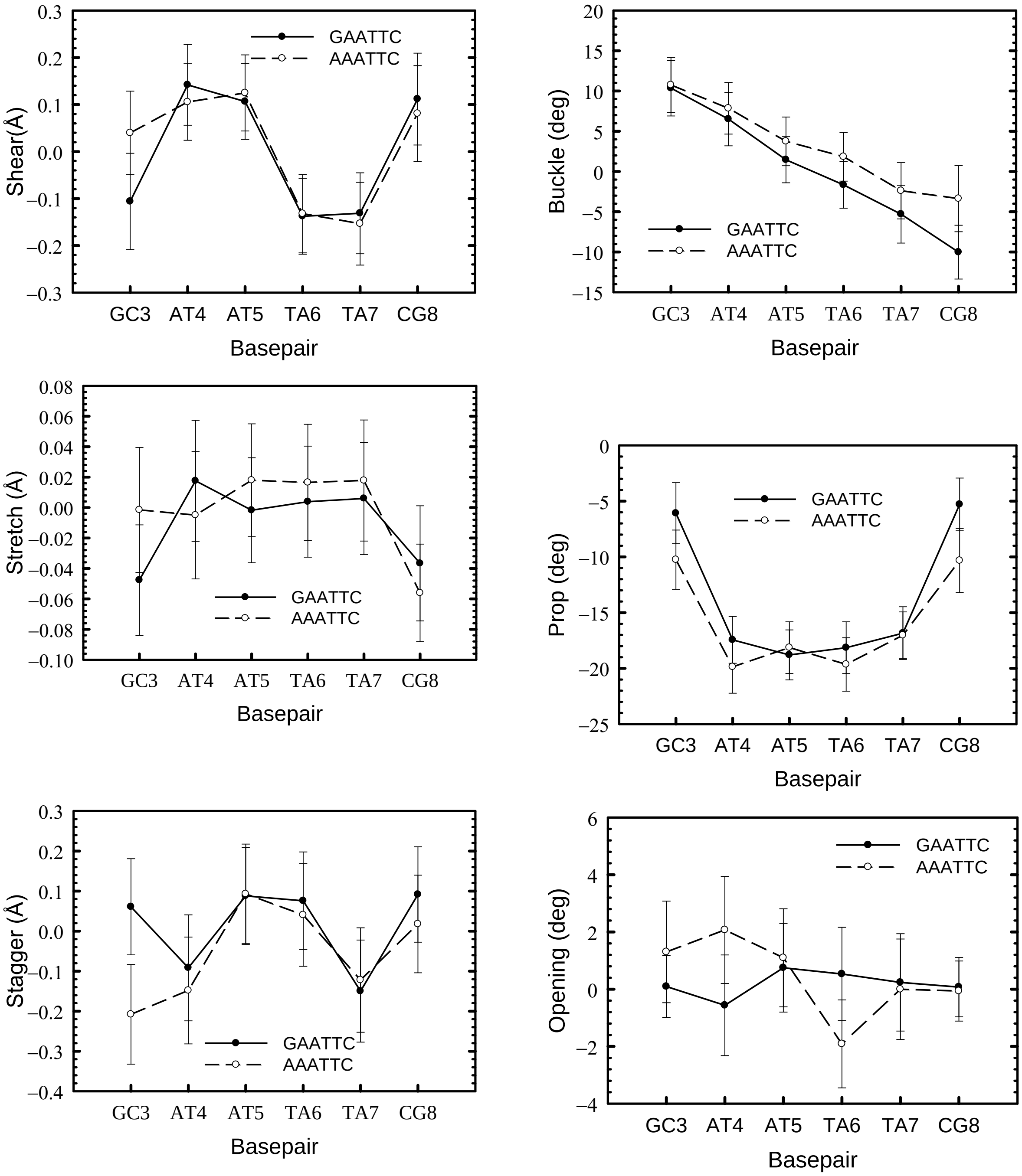
Figure C-2. Comparison of the rotational density of states spectrum of bulk (A), interface (B) and intercalating waters (C) in the cognate and noncognate complexes.

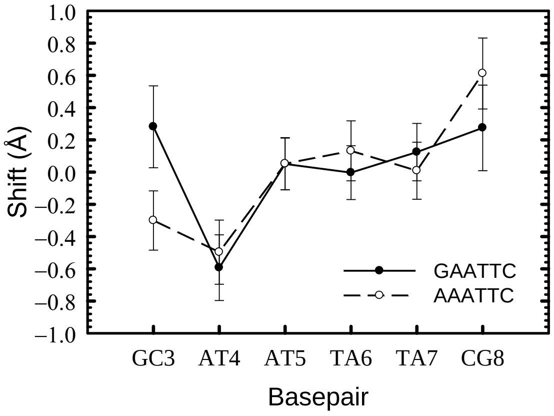


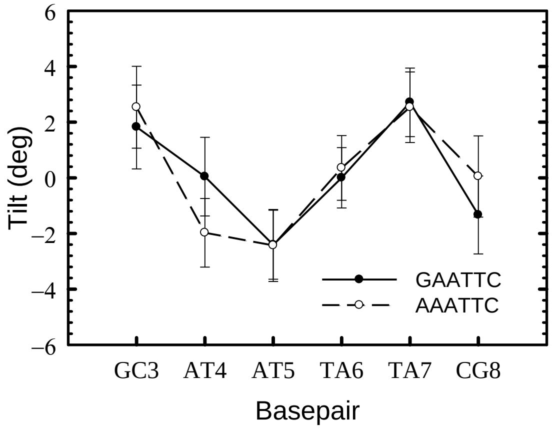

0.005
0.004
0.003
0.002
0.001
0.000
Rotational DoS (cm)
0
200
400
600
800
1000
1200
Frequency (cm $^{-1}$)
(C)
Cognate Complex
Noncognate Complex
Intercalating Waters

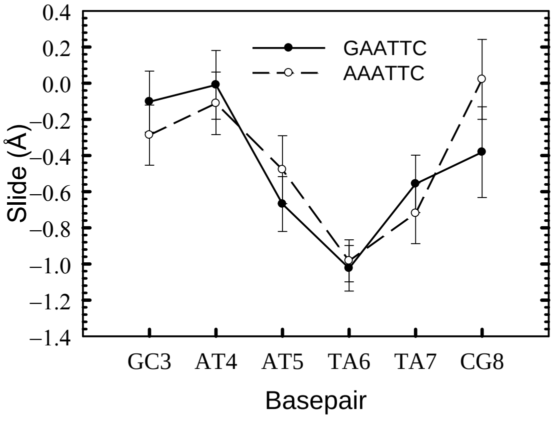
Appendix D: Helecoidal Parameters of Free and Protein-bound DNA

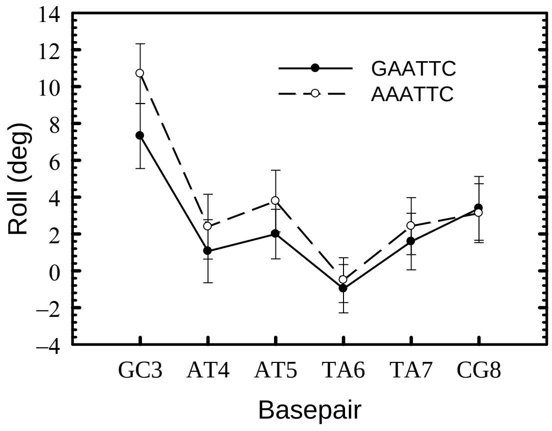
Figure D-1: Comparison of the helecoidal parameters of protein-free GAATTC and protein-free AAATTC sequences.

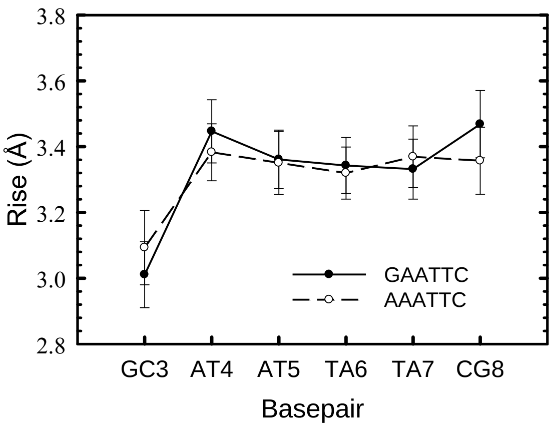



Shift (Å)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt (deg)
6
4
2
0
−2
−4
−6
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide (Å)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll (deg)
14
12
10
8
6
4
2
0
−2
−4
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Rise (Å)
3.8
3.6
3.4
3.2
3.0
2.8
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

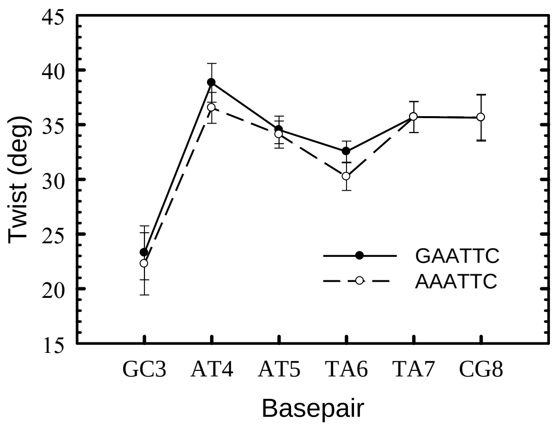
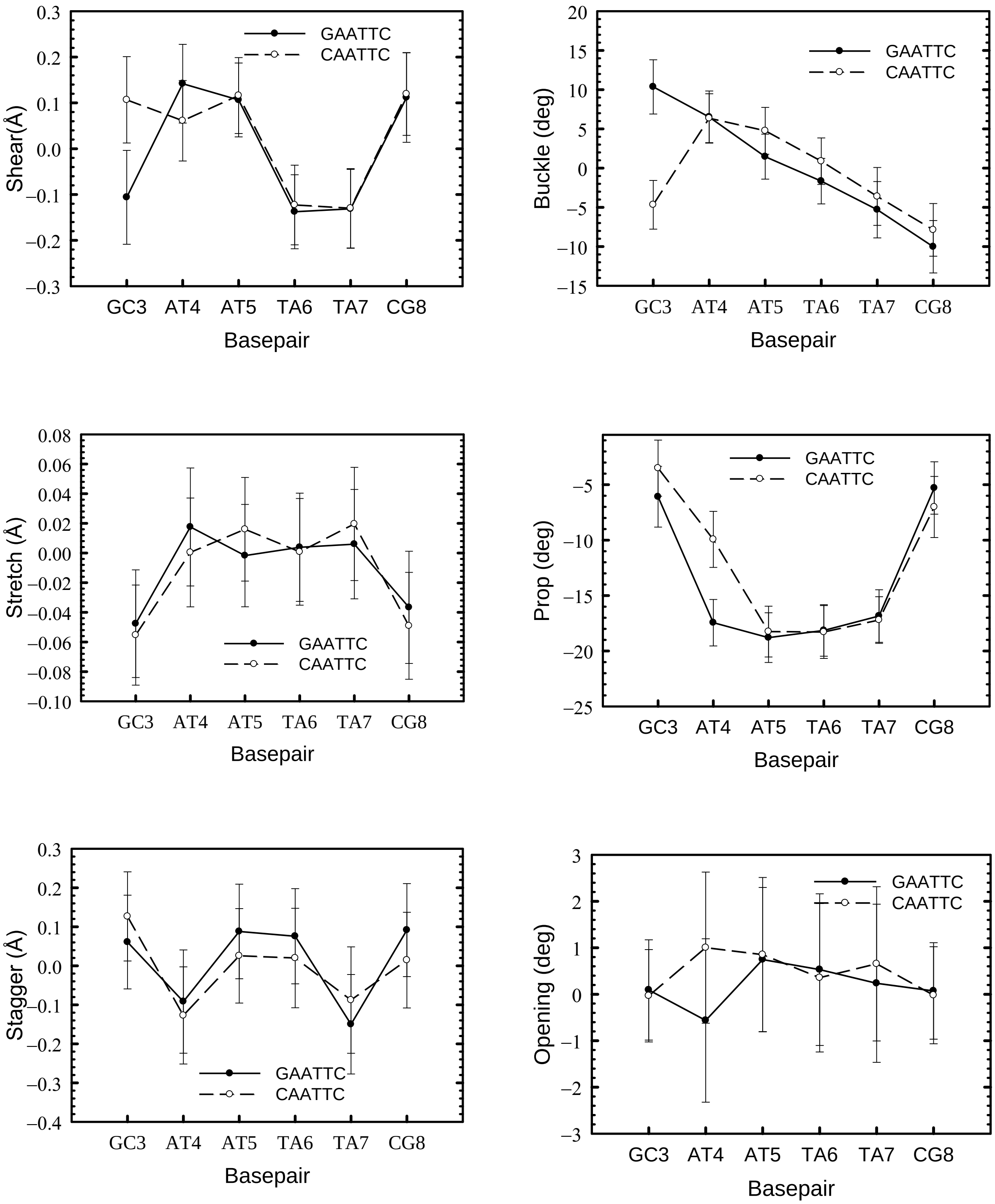
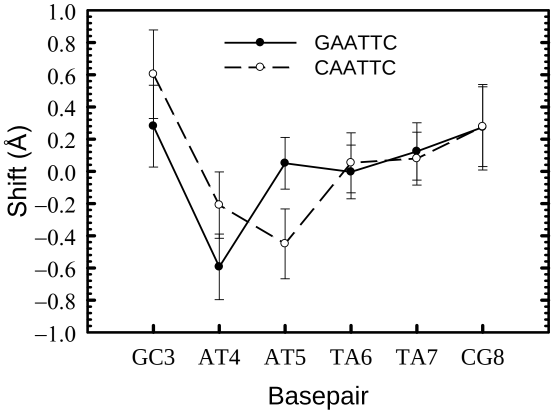
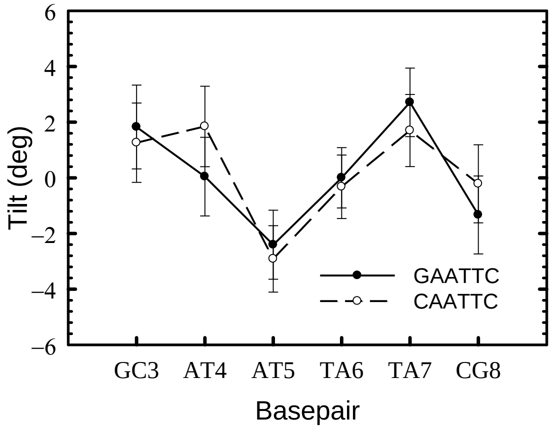
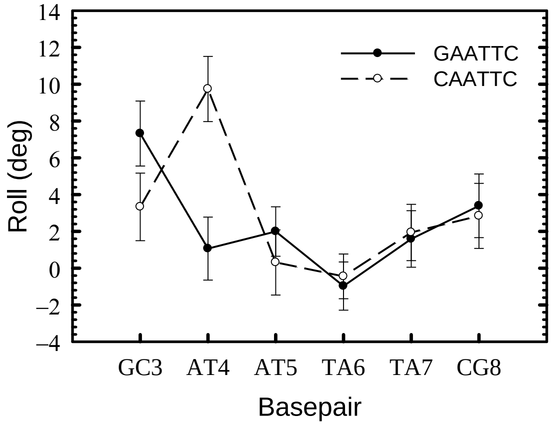

Twist (deg)
45
40
35
30
25
20
15
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

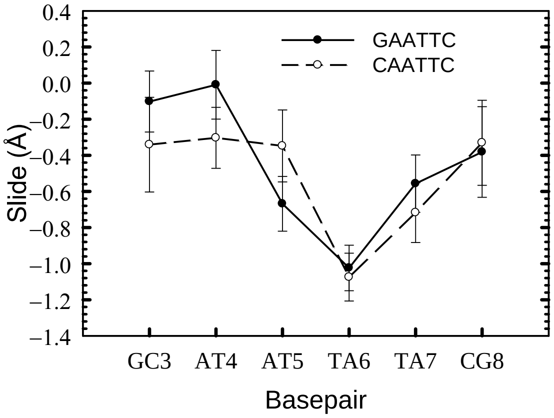
Figure D-2: Comparison of the helecoidal parameters of protein-free GAATTC and protein-free CAATTC sequences.

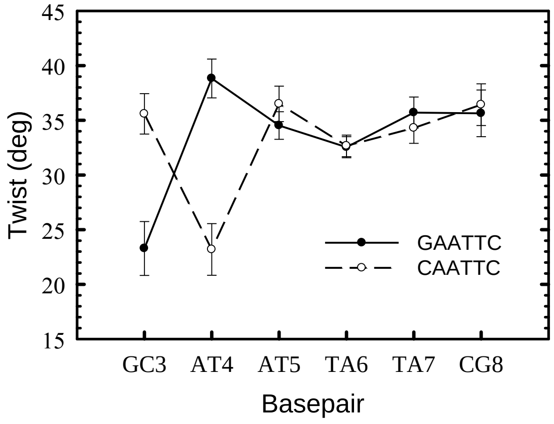



Shift (Å)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt (deg)
6
4
2
0
−2
−4
−6
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll (deg)
14
12
10
8
6
4
2
0
−2
−4
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide (Å)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Twist (deg)
45
40
35
30
25
20
15
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

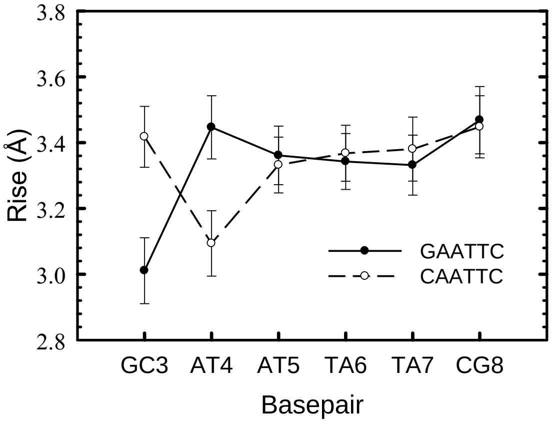
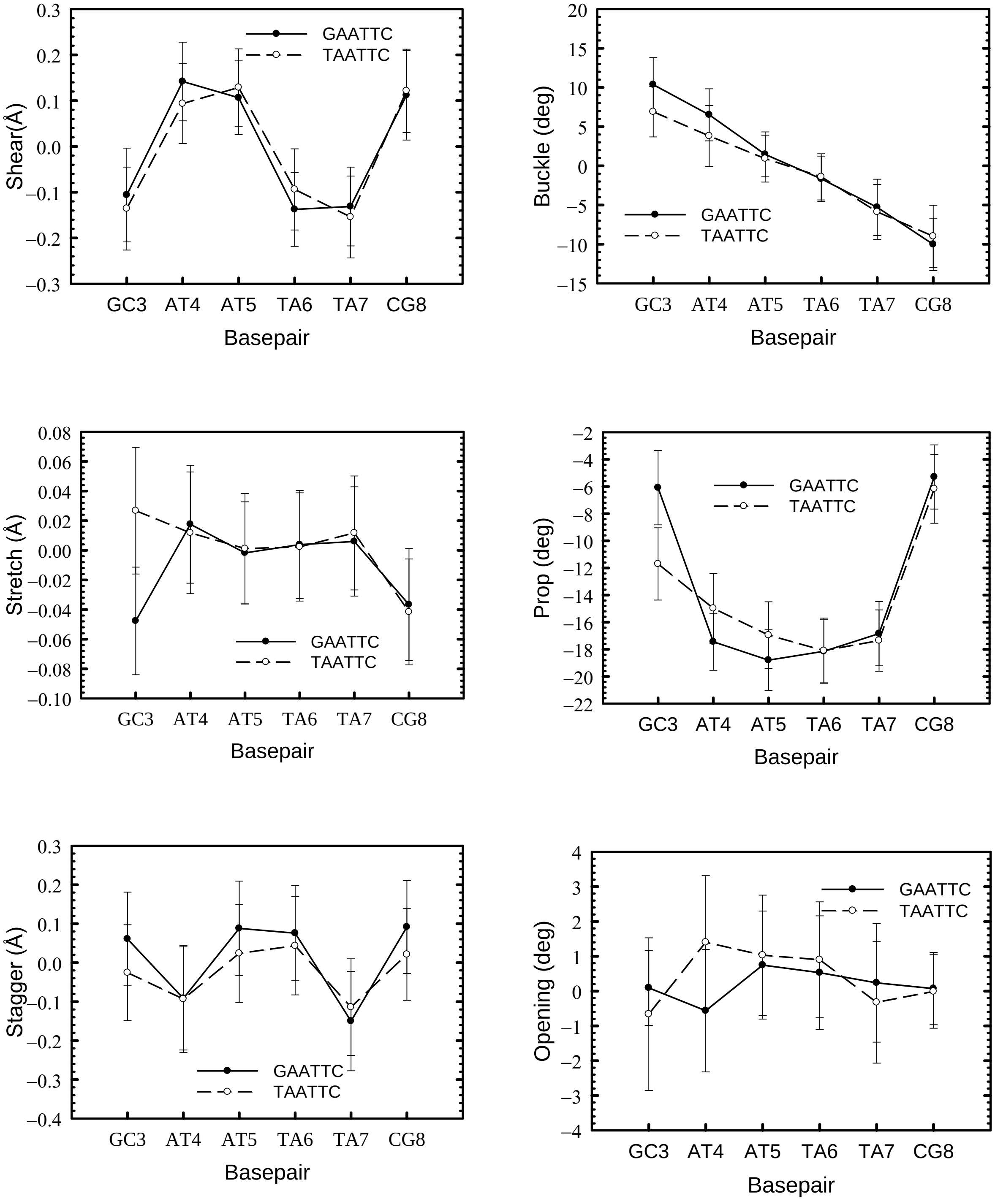
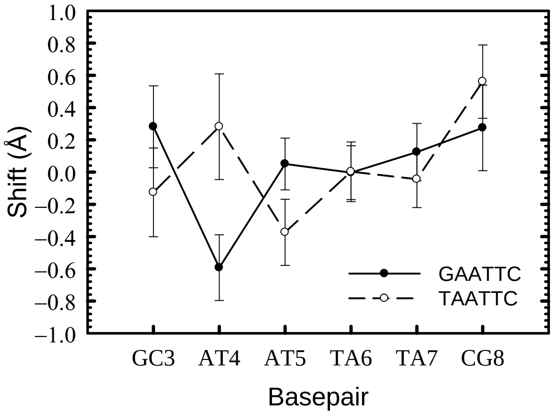
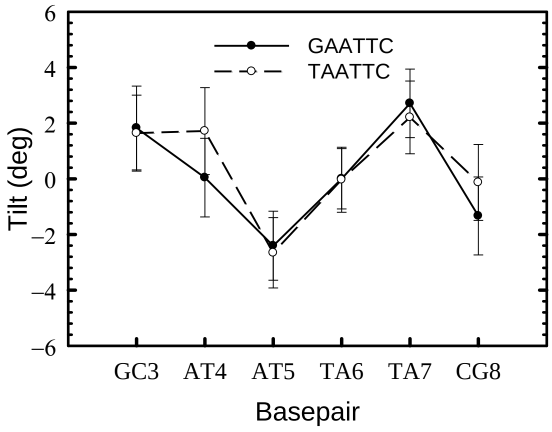
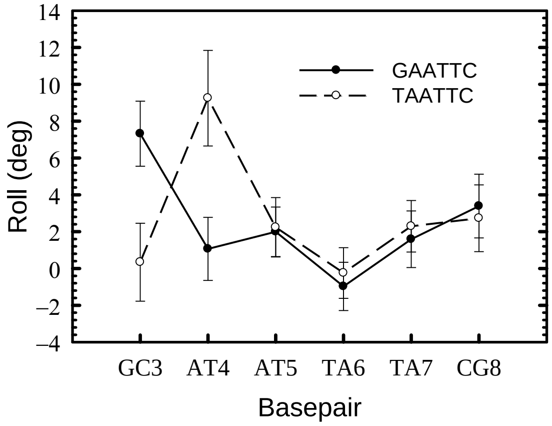

Rise (Å)
3.8
3.6
3.4
3.2
3.0
2.8
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

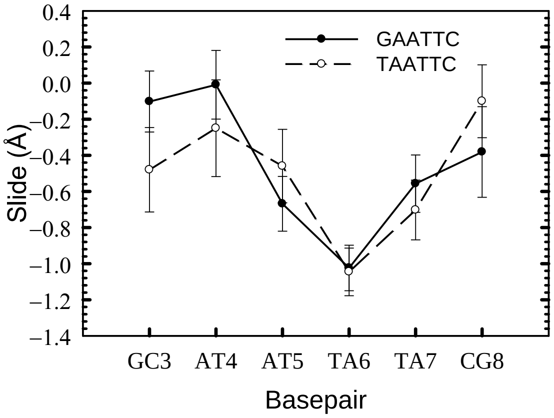
Figure D-3: Comparison of the helecoidal parameters of protein-free GAATTC and protein-free TAATTC sequences.

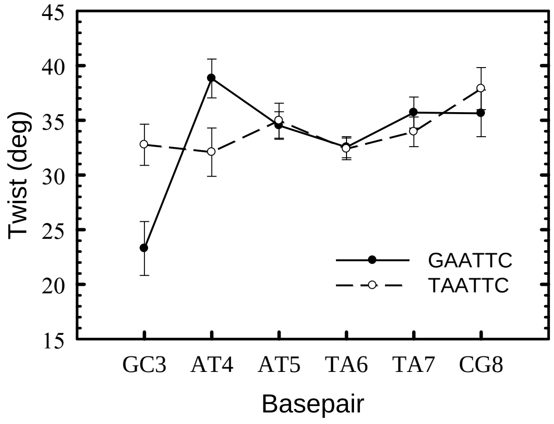



1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
Shift (Å)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


6
4
2
0
−2
−4
−6
Tilt (deg)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


14
12
10
8
6
4
2
0
−2
−4
Roll (deg)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
Slide (Å)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


45
40
35
30
25
20
15
Twist (deg)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

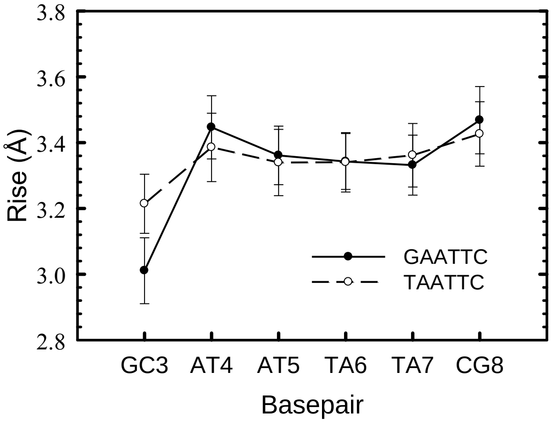
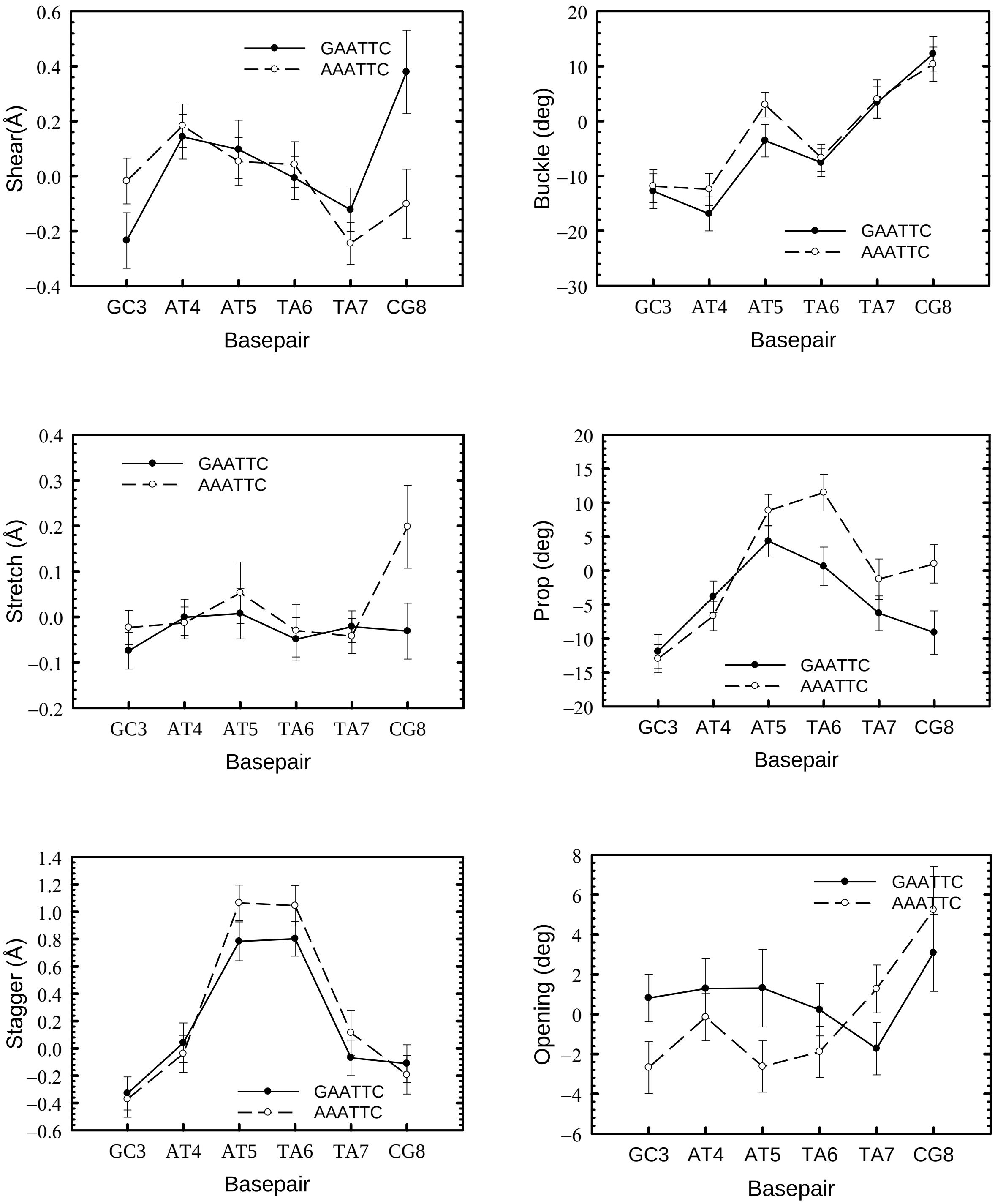
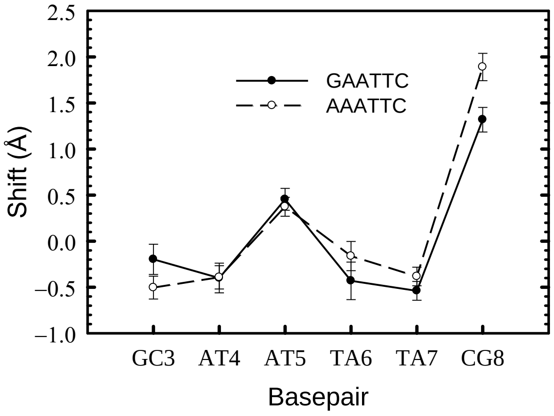
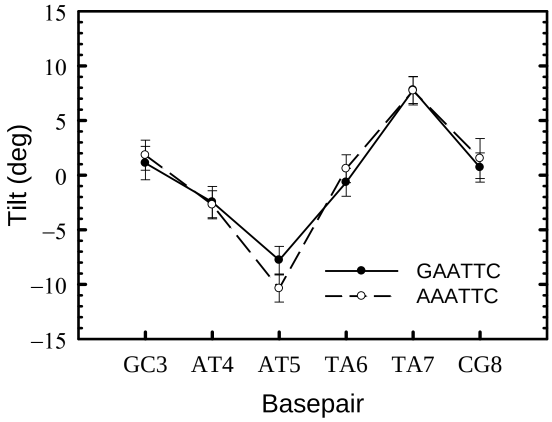
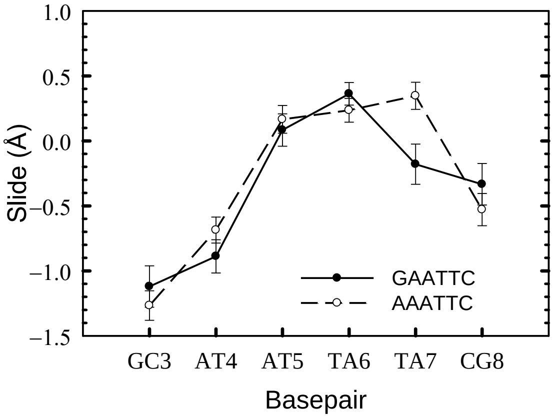

3.8
3.6
3.4
3.2
3.0
2.8
Rise (Å)
GAATTC
TAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

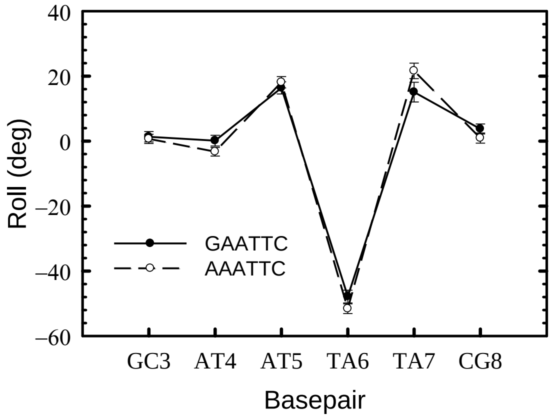
Figure D-4: Comparison of the helecoidal parameters of protein-bound GAATTC and protein-bound AAATTC sequences.

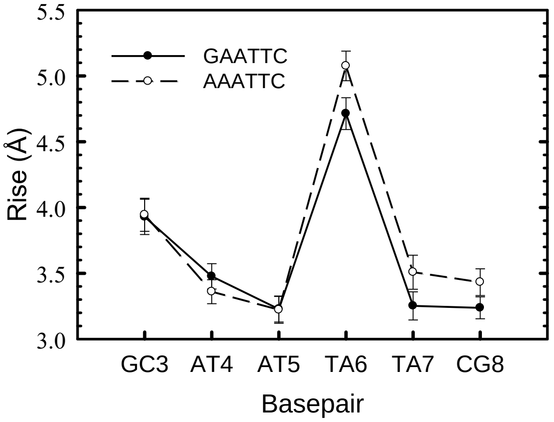



Shift (Å)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt (deg)
15
10
5
0
−5
−10
−15
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide (Å)
1.0
0.5
0.0
−0.5
−1.0
−1.5
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll (deg)
40
20
0
−20
−40
−60
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Rise (Å)
5.5
5.0
4.5
4.0
3.5
3.0
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

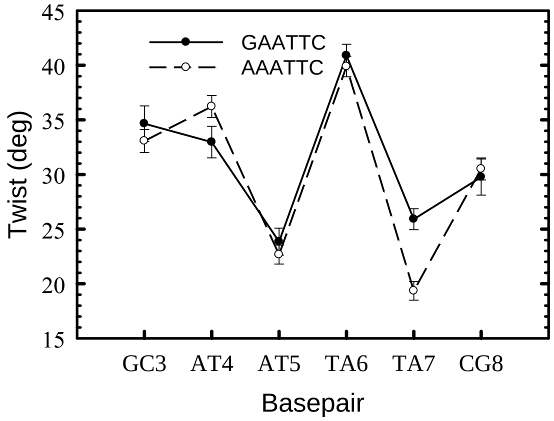
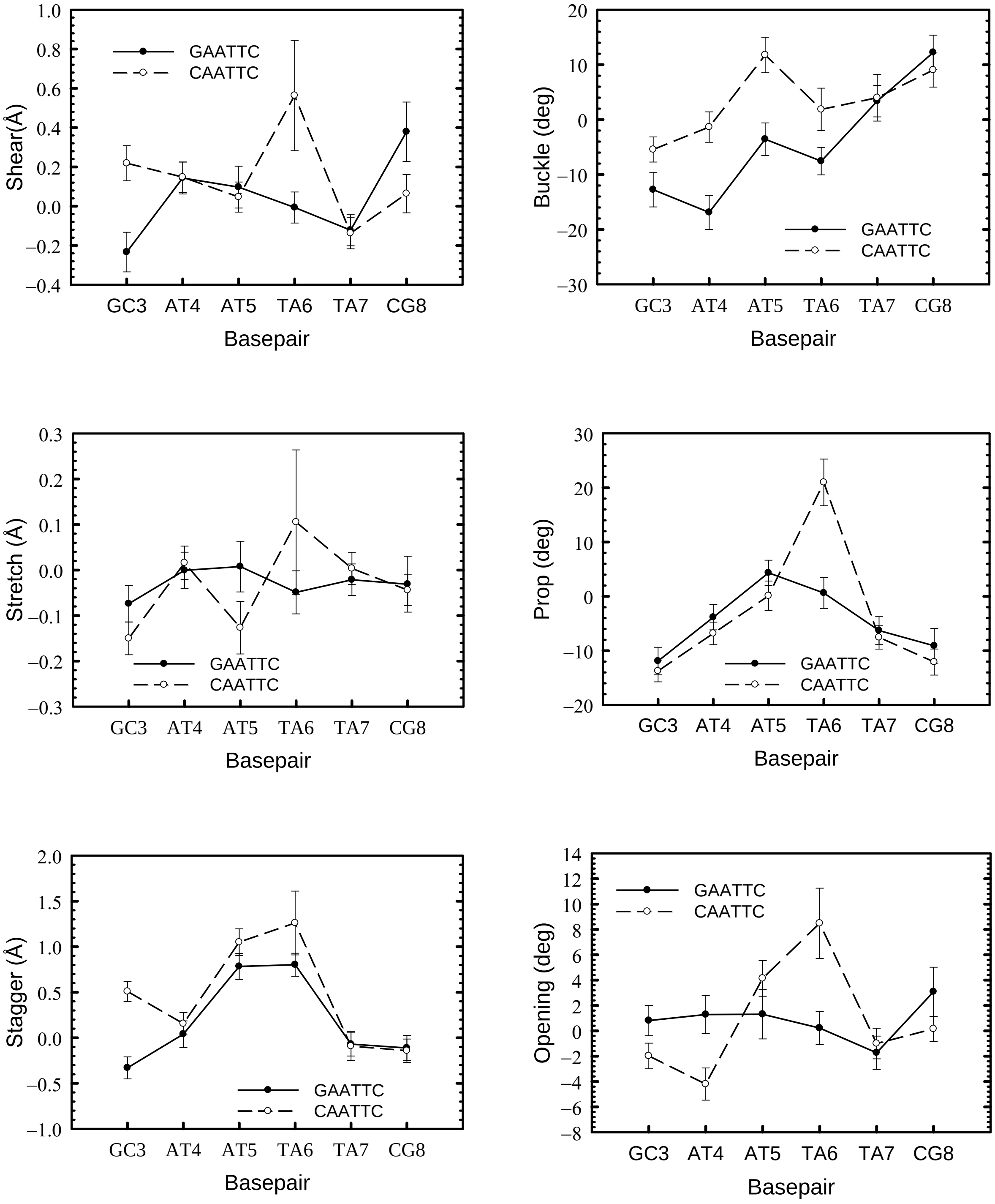
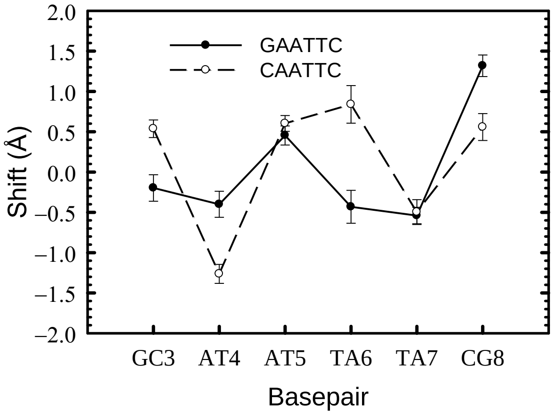
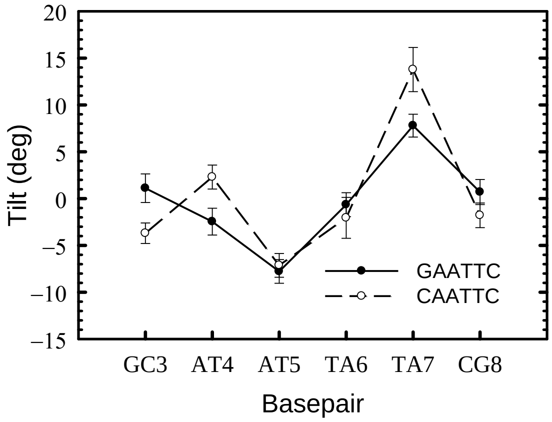
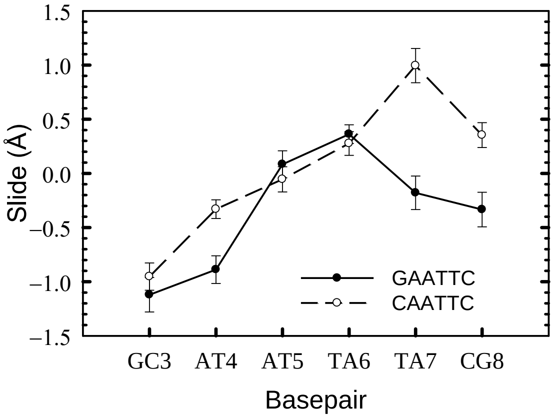

Twist (deg)
45
40
35
30
25
20
15
GAATTC
AAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

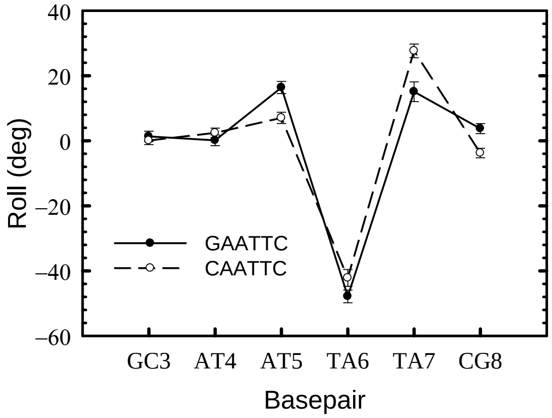
Figure D-5: Comparison of the helecoidal parameters of protein-bound GAATTC and protein-bound CAATTC sequences.

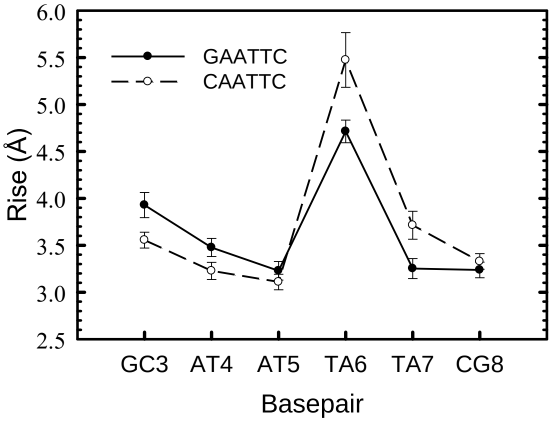



GAATTC
CAATTC
Shift (Å)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt (deg)
20
15
10
5
0
−5
−10
−15
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide (Å)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll (deg)
40
20
0
−20
−40
−60
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


GAATTC
CAATTC
Rise (Å)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

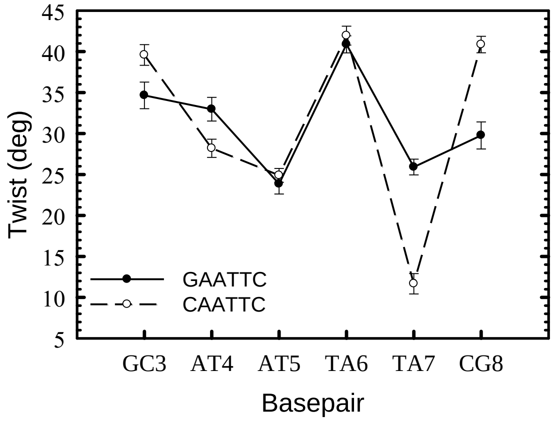
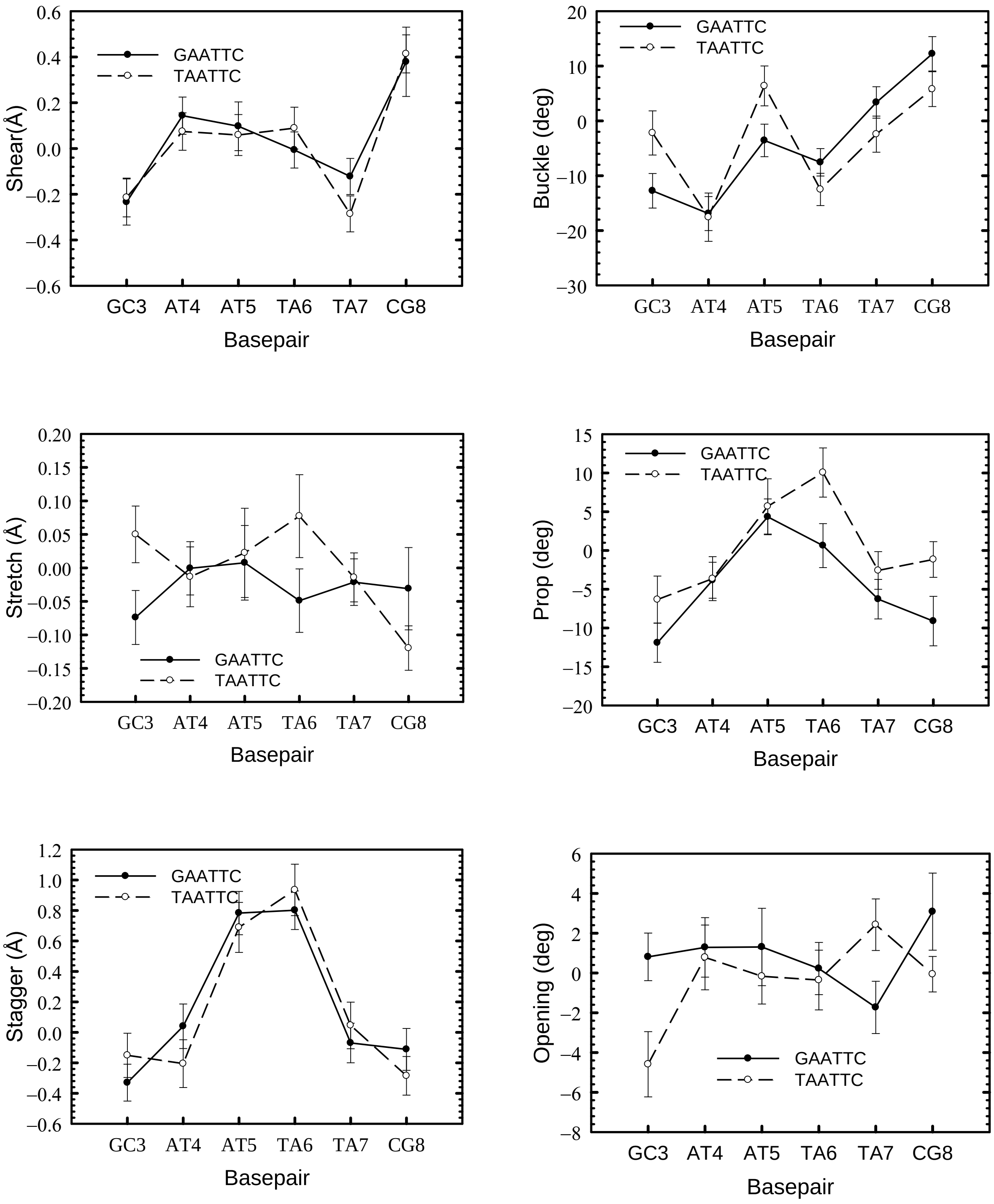
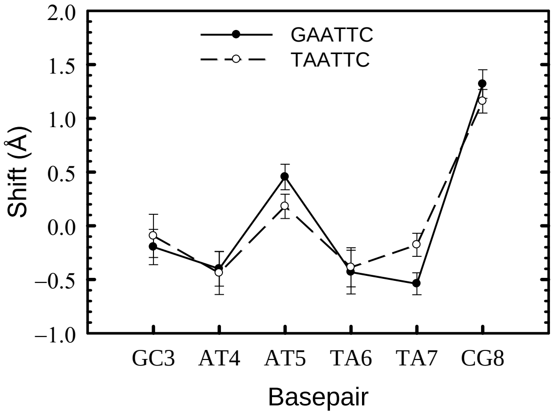
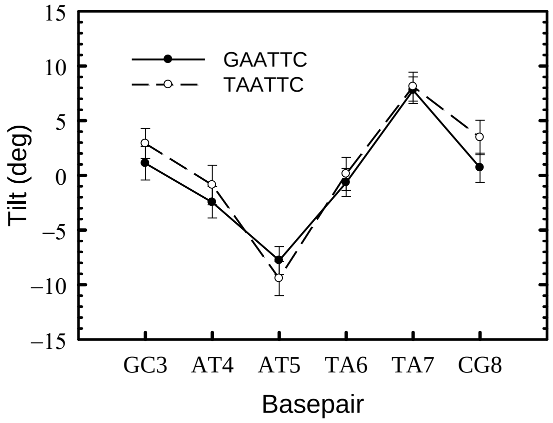
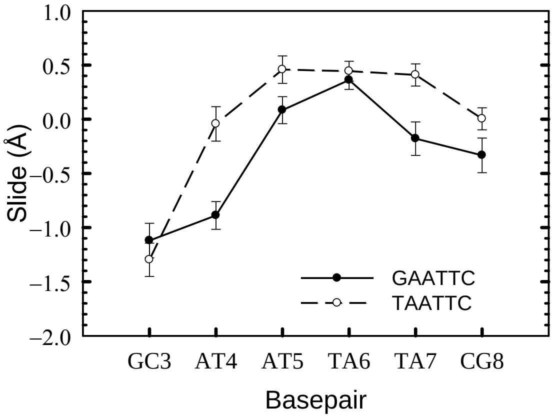

Twist (deg)
45
40
35
30
25
20
15
10
5
GAATTC
CAATTC
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

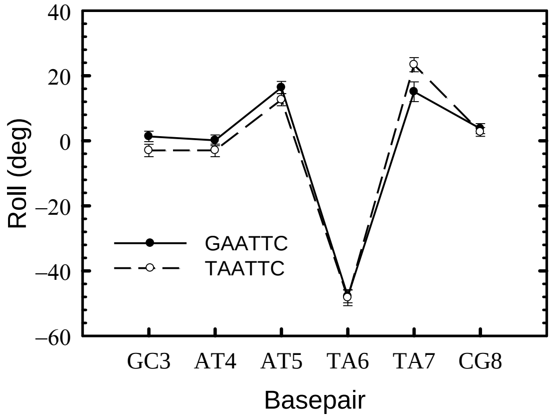
Figure D-6: Comparison of the helecoidal parameters of protein-bound GAATTC and protein-bound TAATTC sequences.

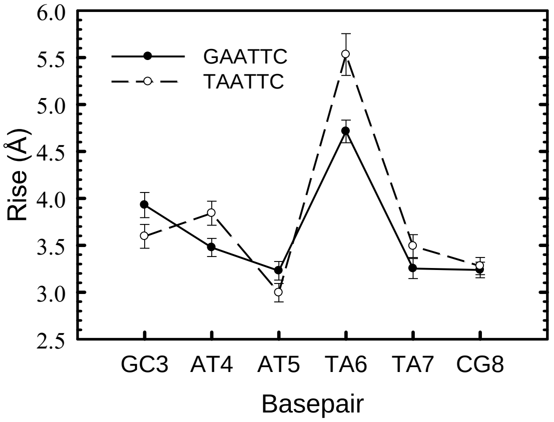



GAATTC
TAATTC
Shift (Å)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


GAATTC
TAATTC
Tilt (deg)
15
10
5
0
−5
−10
−15
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


GAATTC
TAATTC
Slide (Å)
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


GAATTC
TAATTC
Roll (deg)
40
20
0
−20
−40
−60
GC3
AT4
AT5
TA6
TA7
CG8
Basepair


GAATTC
TAATTC
Rise (Å)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

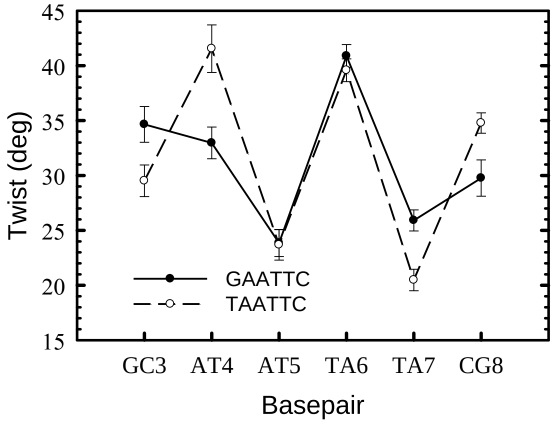
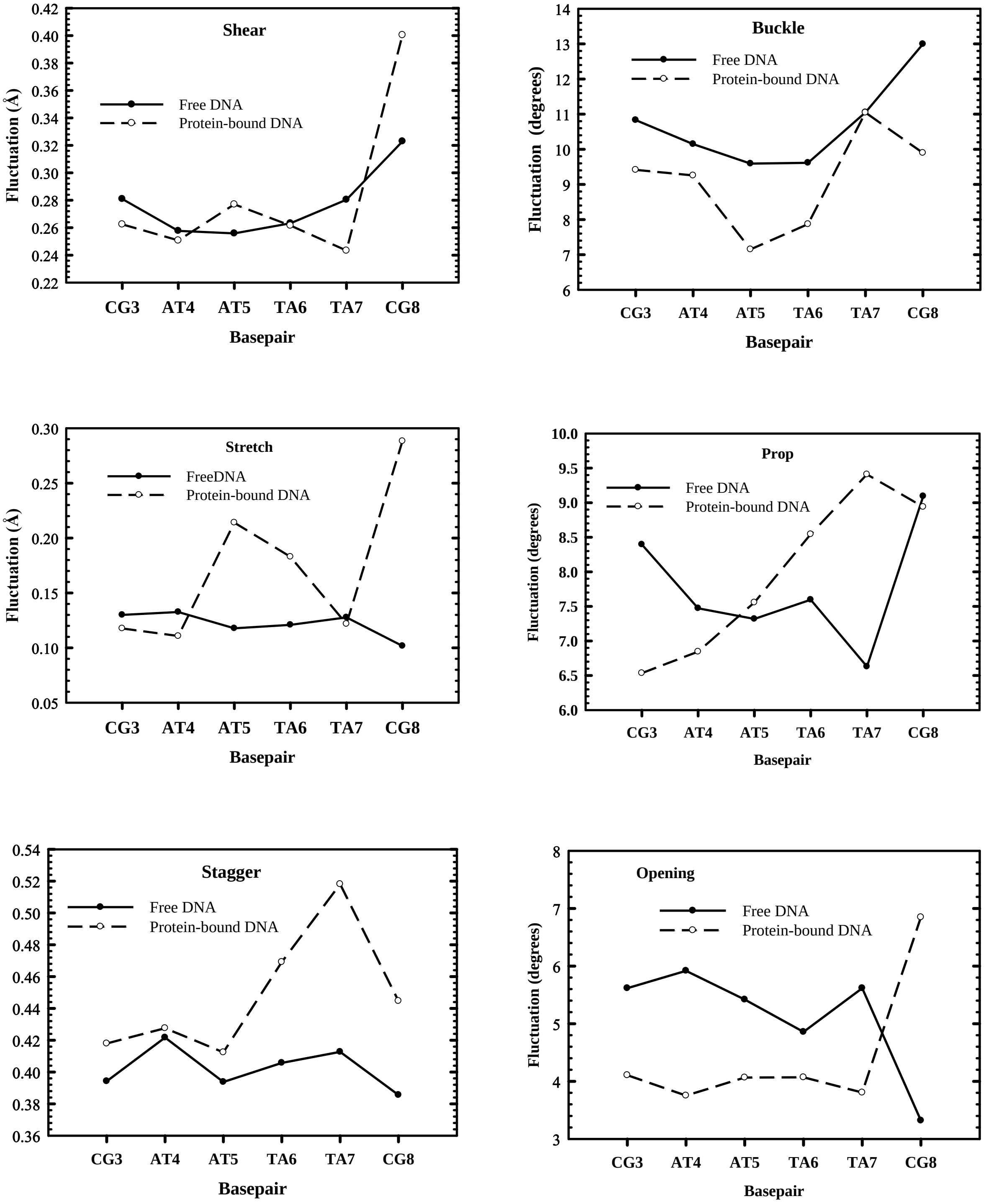
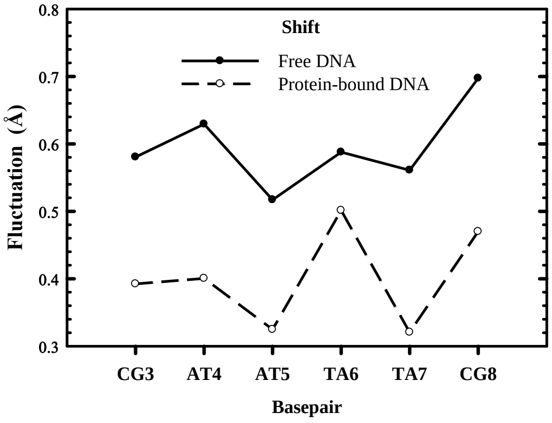
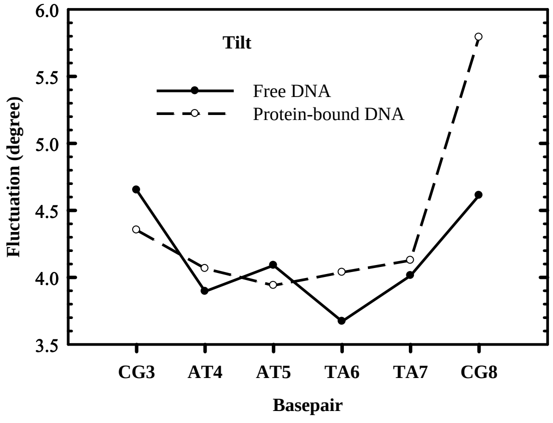
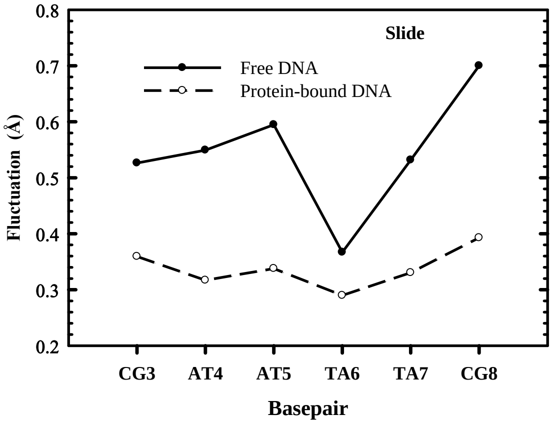

GAATTC
TAATTC
Twist (deg)
45
40
35
30
25
20
15
GC3
AT4
AT5
TA6
TA7
CG8
Basepair

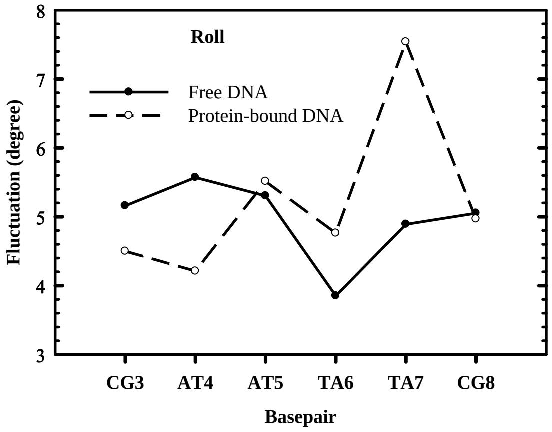
Figure D-7: Comparison of fluctuations in free and protein-bound AAATTC sequences.

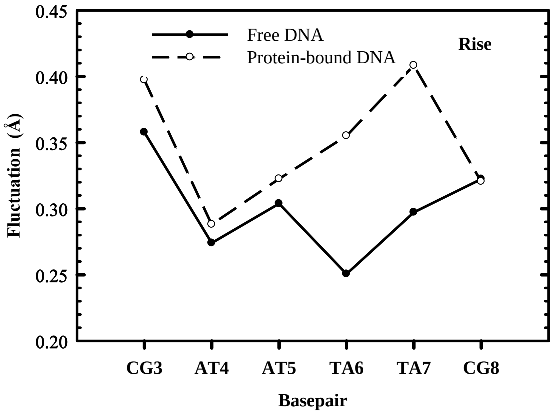



Shift
Free DNA
Protein-bound DNA
Fluctuation (Å)
0.8
0.7
0.6
0.5
0.4
0.3
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt
Free DNA
Protein-bound DNA
Fluctuation (degree)
6.0
5.5
5.0
4.5
4.0
3.5
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide
Free DNA
Protein-bound DNA
Fluctuation (Å)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll
Free DNA
Protein-bound DNA
Fluctuation (degree)
8
7
6
5
4
3
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Rise
Free DNA
Protein-bound DNA
Fluctuation (Å)
0.45
0.40
0.35
0.30
0.25
0.20
CG3
AT4
AT5
TA6
TA7
CG8
Basepair

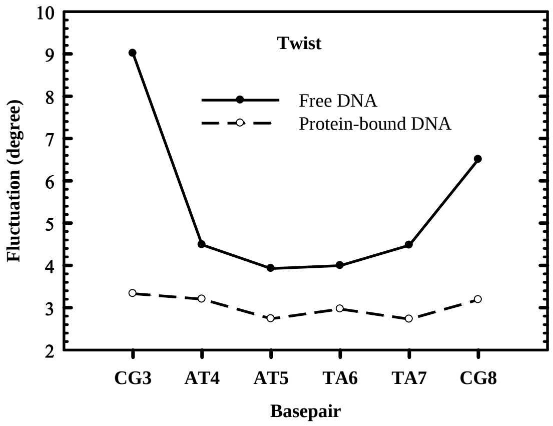
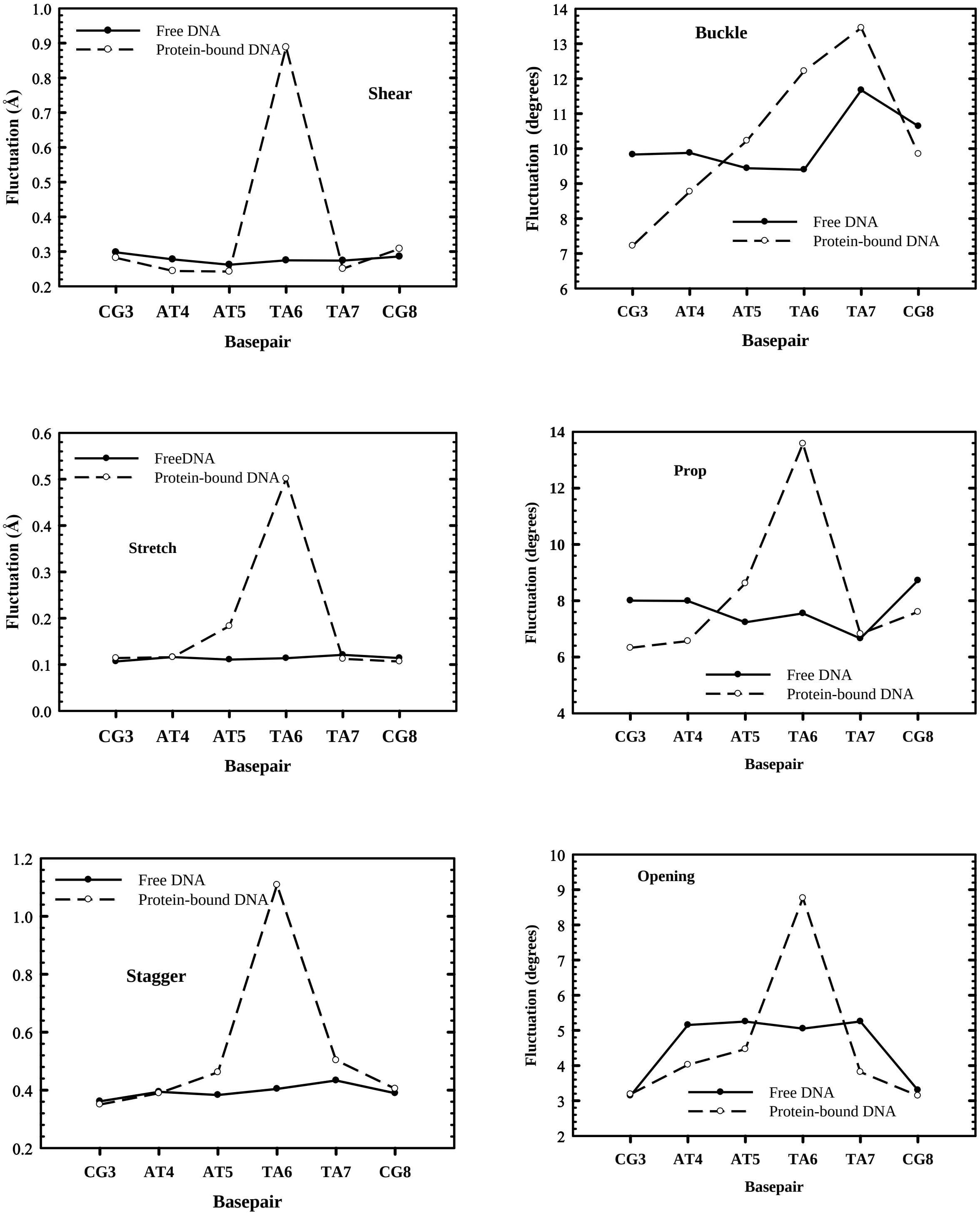
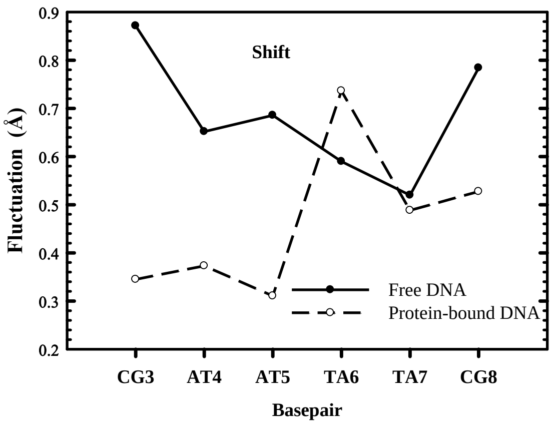
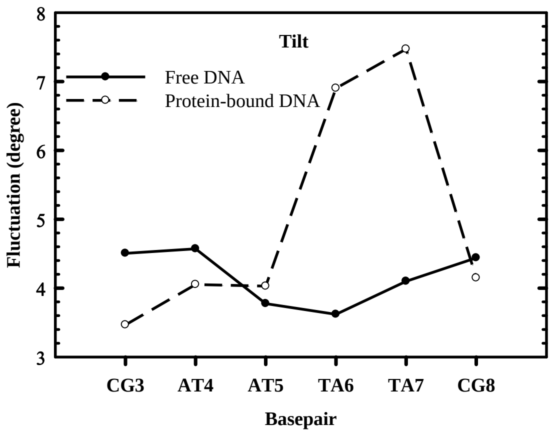
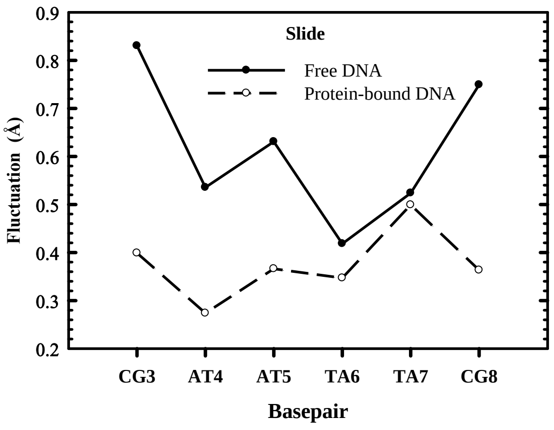

Twist
Free DNA
Protein-bound DNA
Fluctuation (degree)
10
9
8
7
6
5
4
3
2
CG3
AT4
AT5
TA6
TA7
CG8
Basepair

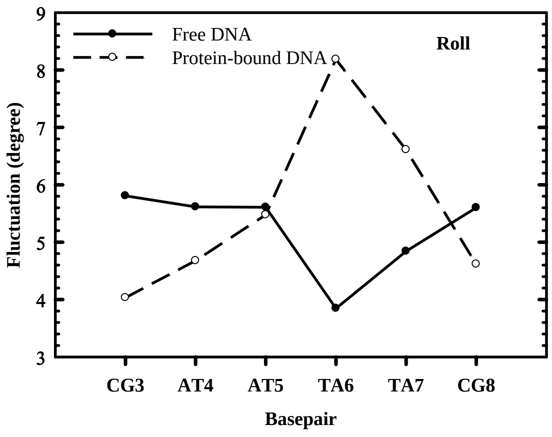
Figure D-8: Comparison of fluctuations in free and protein-bound CAATTC sequence.

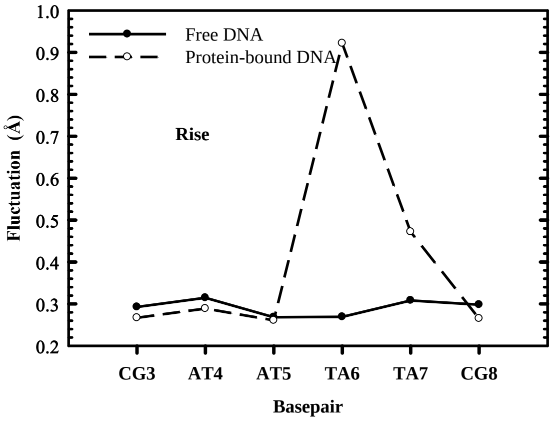



Shift
Fluctuation (Å)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt
Fluctuation (degree)
8
7
6
5
4
3
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide
Fluctuation (Å)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Roll
Fluctuation (degree)
9
8
7
6
5
4
3
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair


Rise
Fluctuation (Å)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair

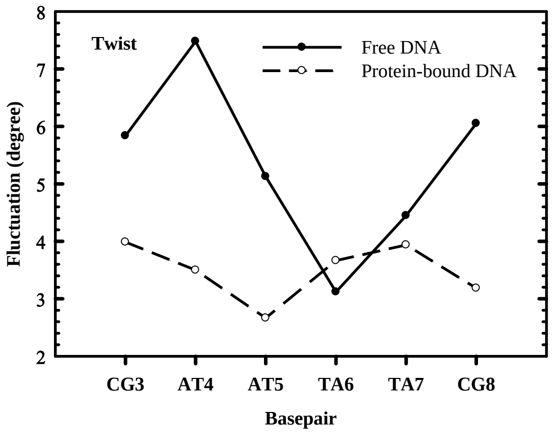
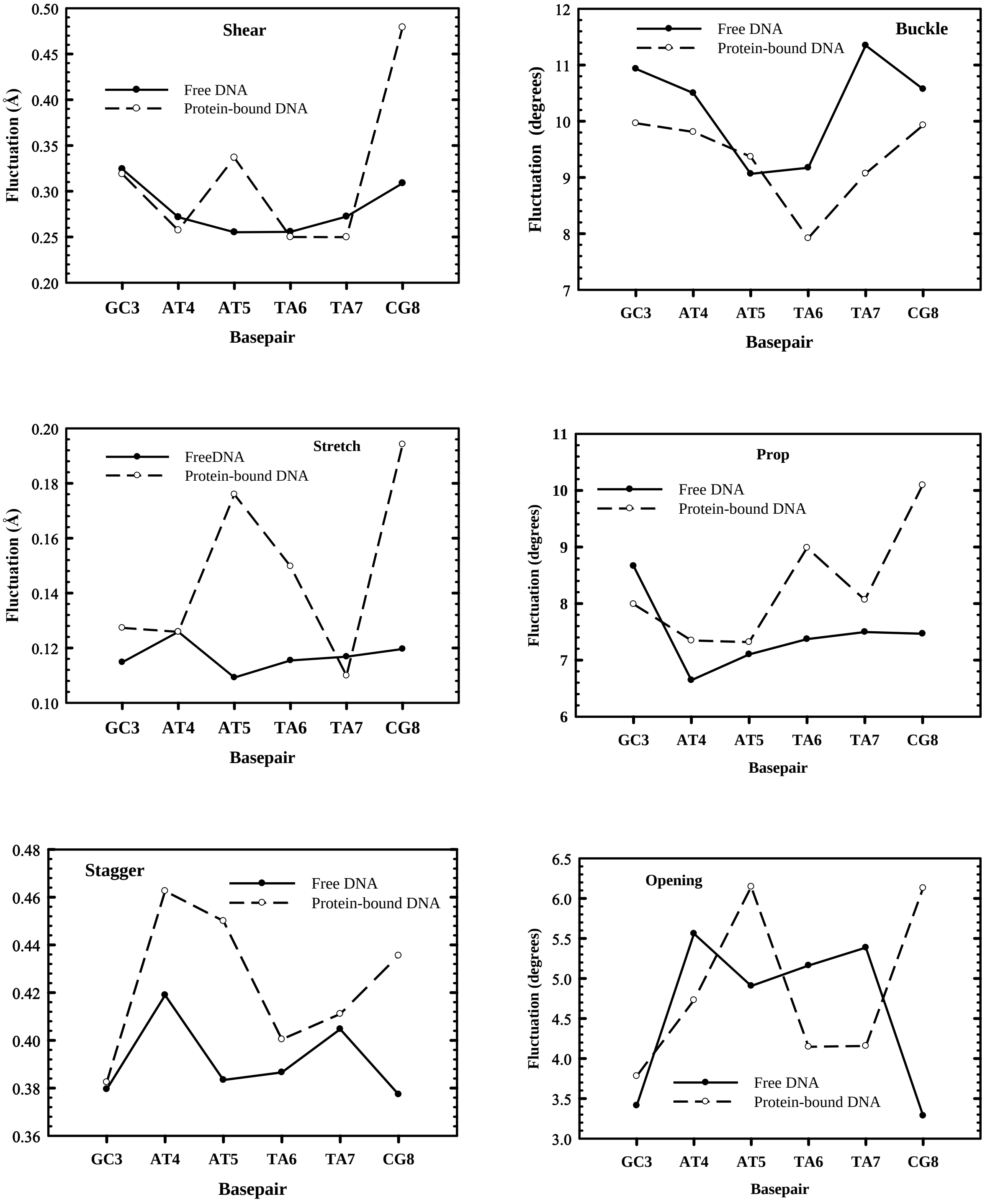
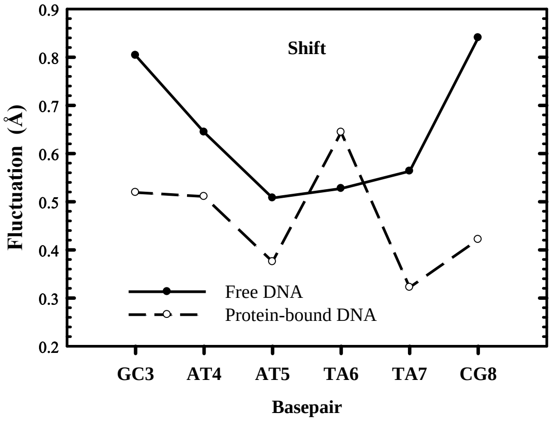
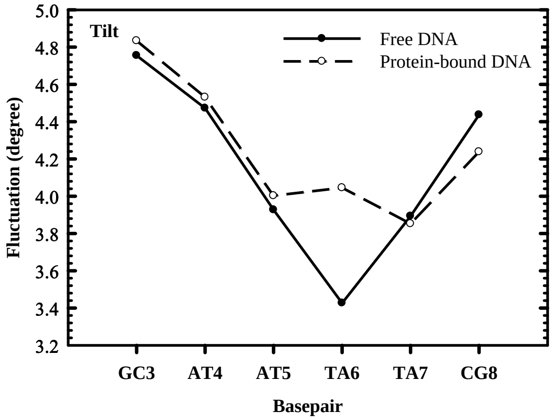
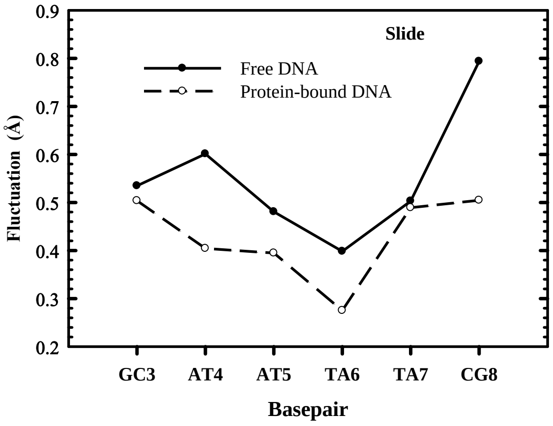

Twist
Fluctuation (degree)
8
7
6
5
4
3
2
Free DNA
Protein-bound DNA
CG3
AT4
AT5
TA6
TA7
CG8
Basepair

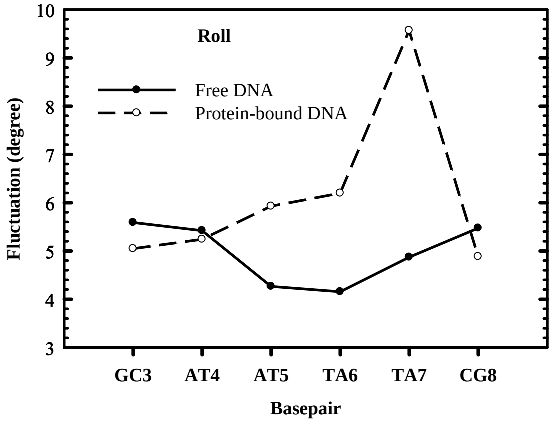
Figure D-9: Comparison of fluctuations in free and protein-bound GAATTC sequence

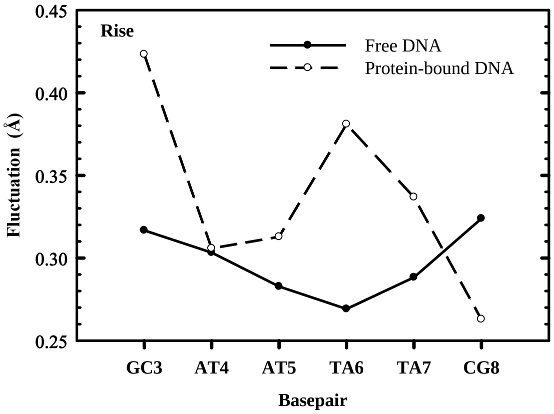



Shift
Fluctuation (Å)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA


Tilt
Fluctuation (degree)
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA


Slide
Fluctuation (Å)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA


Roll
Fluctuation (degree)
3
4
5
6
7
8
9
10
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA


Rise
Fluctuation (Å)
0.25
0.30
0.35
0.40
0.45
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA

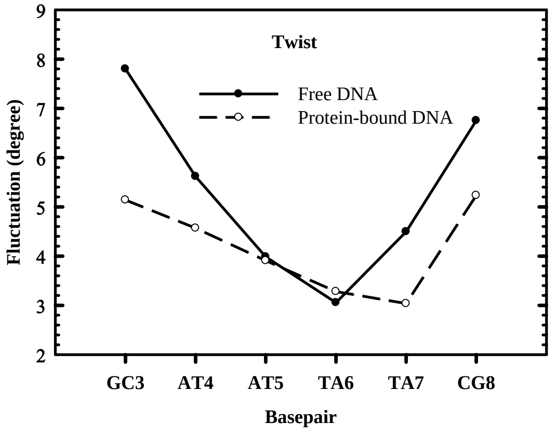
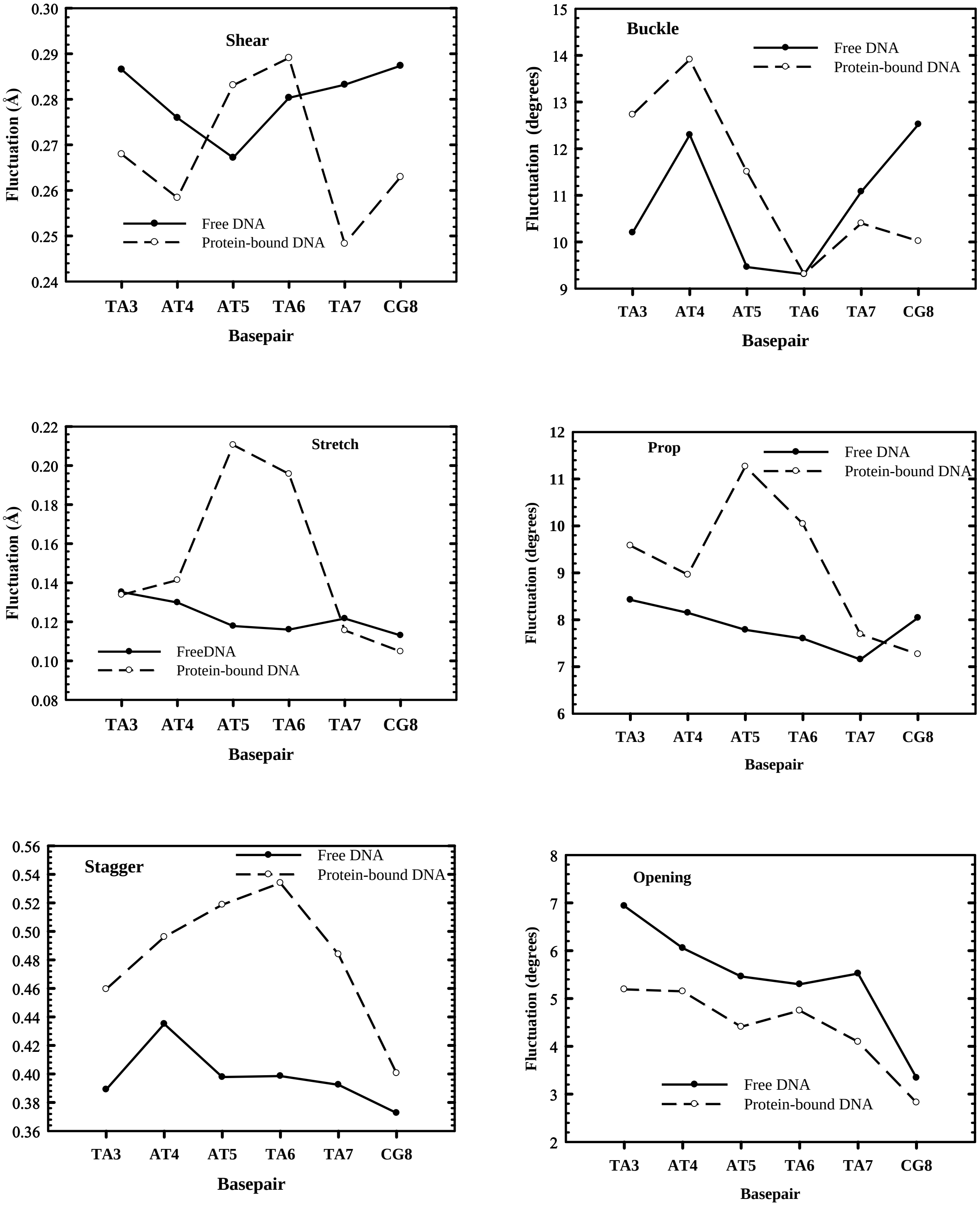
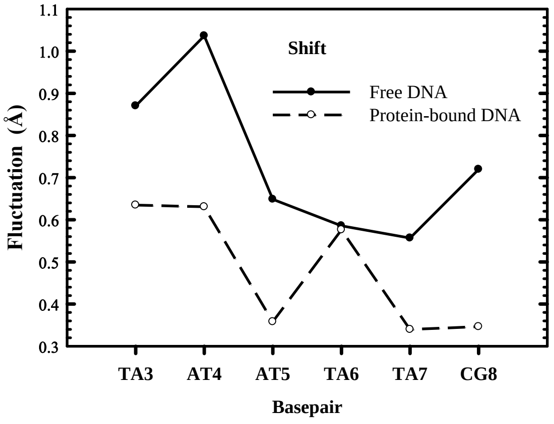
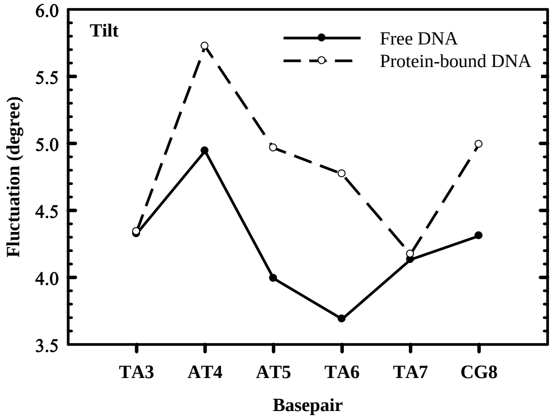
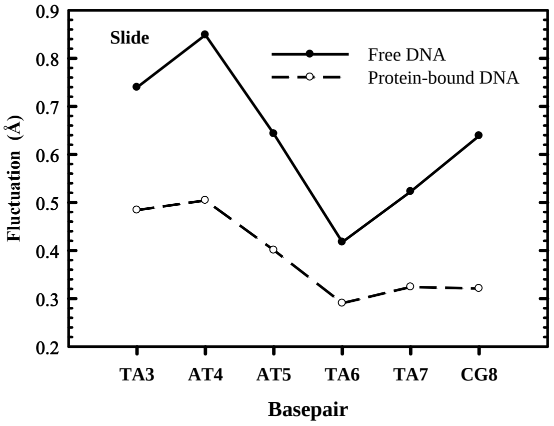

Twist
Fluctuation (degree)
2
3
4
5
6
7
8
9
GC3
AT4
AT5
TA6
TA7
CG8
Basepair
Free DNA
Protein-bound DNA

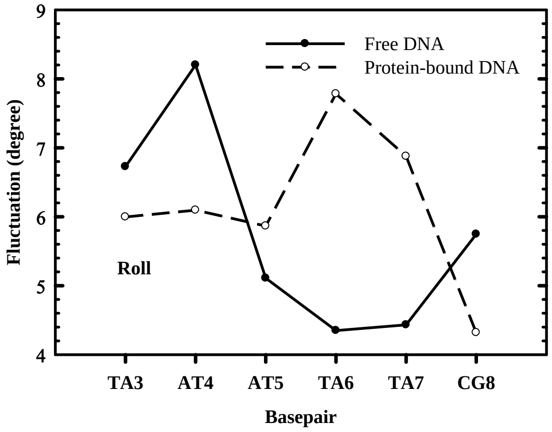
Figure D-10: Comparison of fluctuations in free and protein-bound TAATTC

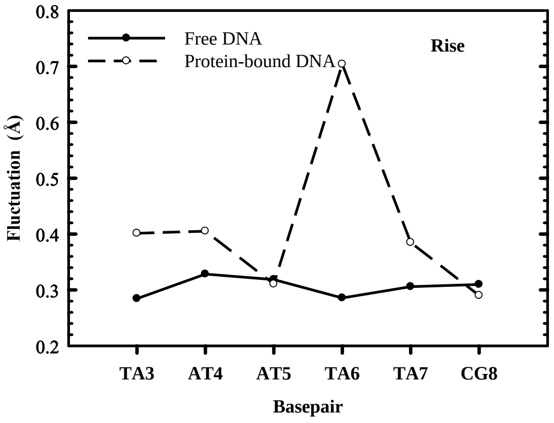



Shift
Free DNA
Protein-bound DNA
Fluctuation (Å)
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
TA3
AT4
AT5
TA6
TA7
CG8
Basepair


Tilt
Free DNA
Protein-bound DNA
Fluctuation (degree)
6.0
5.5
5.0
4.5
4.0
3.5
TA3
AT4
AT5
TA6
TA7
CG8
Basepair


Slide
Free DNA
Protein-bound DNA
Fluctuation (Å)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
TA3
AT4
AT5
TA6
TA7
CG8
Basepair


Free DNA
Protein-bound DNA
Roll
Fluctuation (degree)
9
8
7
6
5
4
TA3
AT4
AT5
TA6
TA7
CG8
Basepair


Free DNA
Protein-bound DNA
Rise
Fluctuation (Å)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
TA3
AT4
AT5
TA6
TA7
CG8
Basepair

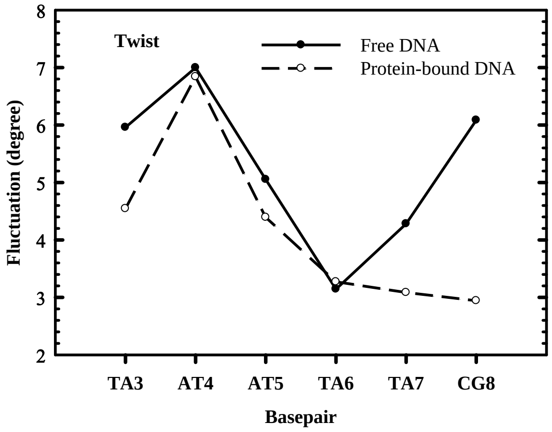

Twist
Free DNA
Protein-bound DNA
Fluctuation (degree)
8
7
6
5
4
3
2
TA3
AT4
AT5
TA6
TA7
CG8
Basepair

Table D-1: List of protein-DNA hydrogen bonds along with their propensity values as defined in section 5.2.4

S.No.	Donor Residue	Donor Atom	Acceptor Atom	Acceptor Residue	Cognate	AAATTC	CAATTC	TAATTC
1	100GLY	N	O1P	269DT	0.97	0.95	0.00	0.96
2	100GLY	N	O2P	269DT	0.00	0.01	0.00	0.00
3	101LYS	N	O1P	269DT	0.49	0.00	0.00	0.00
4	101LYS	N	O2P	269DT	0.00	0.00	0.00	0.00
5	101LYS	NZ	O1P	269DT	0.00	0.15	0.03	0.08
6	101LYS	NZ	O1P	270DT	0.14	0.49	0.00	0.61
7	101LYS	NZ	O2P	268DA	0.00	0.00	0.02	0.00
8	101LYS	NZ	O2P	269DT	0.00	0.23	0.15	0.04
9	101LYS	NZ	O2P	270DT	0.57	0.01	0.00	0.03
10	101LYS	NZ	O3'	269DT	0.02	0.00	0.00	0.01
11	101LYS	NZ	O5'	268DA	0.00	0.00	0.00	0.00
12	101LYS	NZ	O5'	269DT	0.00	0.42	0.00	0.54
13	114LYS	N	O1P	271DC	0.00	0.00	0.00	0.00
14	114LYS	NZ	N4	271DC	0.00	0.00	0.03	0.00
15	114LYS	NZ	N4	273DC	0.03	0.00	0.05	0.03
16	114LYS	NZ	N4	537DC	0.00	0.00	0.00	0.00
17	114LYS	NZ	N4	539DC	0.00	0.00	0.01	0.00
18	114LYS	NZ	N7	272DG	0.27	0.01	0.21	0.11
19	114LYS	NZ	N7	538DG	0.01	0.00	0.02	0.00

20	114LYS	NZ	O1P	271DC	0.00	0.01	0.00	0.00
21	114LYS	NZ	O1P	272DG	0.02	0.28	0.00	0.19
22	114LYS	NZ	O2P	271DC	0.00	0.14	0.00	0.02
23	114LYS	NZ	O2P	272DG	0.00	0.21	0.00	0.04
24	114LYS	NZ	O3'	271DC	0.00	0.01	0.00	0.01
25	114LYS	NZ	O5'	271DC	0.00	0.01	0.00	0.00
26	114LYS	NZ	O6	272DG	0.42	0.00	0.31	0.19
27	114LYS	NZ	O6	274DG3	0.00	0.00	0.00	0.00
28	114LYS	NZ	O6	538DG	0.27	0.00	0.26	0.05
29	115ARG	NE	O1P	537DC	0.00	0.00	0.00	0.00
30	115ARG	NE	O5'	536DT5	0.00	0.00	0.00	0.01
31	115ARG	NH1	O1P	272DG	0.03	0.00	0.00	0.03
32	115ARG	NH1	O1P	537DC	0.01	0.00	0.00	0.01
33	115ARG	NH1	O2	536DT5	0.00	0.00	0.00	0.01
34	115ARG	NH1	O2P	537DC	0.00	0.00	0.00	0.00
35	115ARG	NH1	O3'	271DC	0.00	0.00	0.00	0.00
36	115ARG	NH1	O4	536DT5	0.00	0.00	0.01	0.00
37	115ARG	NH1	O5'	536DT5	0.04	0.00	0.00	0.05
38	115ARG	NH2	O1P	272DG	0.03	0.00	0.00	0.02
39	115ARG	NH2	O1P	537DC	0.03	0.00	0.00	0.01
40	115ARG	NH2	O2	536DT5	0.01	0.00	0.00	0.01
41	115ARG	NH2	O4	536DT5	0.00	0.00	0.01	0.00
42	115ARG	NH2	O5'	536DT5	0.03	0.00	0.00	0.02
43	118GLN	NE2	O1P	537DC	0.11	0.00	0.00	0.03

44	118GLN	NE2	O2P	537DC	0.00	0.00	0.00	0.00
45	125ASN	ND2	N6	542DA	0.00	0.00	0.00	0.00
46	125ASN	ND2	N7	540DG	0.23	0.17	0.02	0.02
47	125ASN	ND2	N7	541DA	0.91	0.94	0.96	0.59
48	125ASN	ND2	N9	540DG	0.00	0.01	0.00	0.00
49	125ASN	N	N6	541DA	0.00	0.00	0.00	0.00
50	125ASN	N	O4	269DT	0.48	0.67	0.17	0.36
51	126ALA	N	O4	269DT	0.52	0.61	0.32	0.66
52	129ARG	NE	N6	268DA	0.00	0.00	0.00	0.00
53	129ARG	NE	N7	268DA	0.51	0.66	0.31	0.54
54	129ARG	NH1	O1P	267DA	0.00	0.06	0.75	0.37
55	129ARG	NH2	N7	268DA	0.52	0.39	0.09	0.76
56	129ARG	NH2	O1P	267DA	0.24	0.74	0.95	0.99
57	129ARG	NH2	O5'	267DA	0.00	0.00	0.09	0.03
58	132LYS	NZ	O1P	266DA	0.00	0.28	0.00	0.00
59	132LYS	NZ	O1P	266DC	0.00	0.00	0.72	0.00
60	132LYS	NZ	O1P	266DG	0.01	0.00	0.00	0.00
61	132LYS	NZ	O1P	266DT	0.00	0.00	0.00	0.06
62	132LYS	NZ	O2P	266DC	0.00	0.00	0.00	0.00
63	132LYS	NZ	O2P	266DT	0.00	0.00	0.00	0.01
64	133ASN	ND2	O1P	266DA	0.00	0.00	0.00	0.00
65	133ASN	ND2	O2P	266DA	0.00	0.12	0.00	0.00
66	133ASN	ND2	O2P	266DT	0.00	0.00	0.00	0.00
67	179SER	OG	O1P	538DG	0.00	0.00	0.01	0.00

68	179SER	OG	O2P	538DG	0.00	0.00	0.00	0.00
69	180GLY	N	O1P	538DG	0.02	0.02	0.07	0.00
70	180GLY	N	O2P	538DG	0.33	0.66	0.58	0.00
71	180GLY	N	O5'	538DG	0.00	0.00	0.00	0.00
72	181ILE	N	O1P	538DG	0.95	0.96	0.89	0.01
73	184ARG	NE	O1P	539DC	0.00	0.00	0.00	0.00
74	184ARG	NH1	O1P	539DC	0.83	0.00	0.79	0.00
75	184ARG	NH1	O3'	538DG	0.07	0.00	0.06	0.00
76	184ARG	NH1	O6	540DG	0.00	0.50	0.00	0.00
77	184ARG	NH2	O1P	539DC	0.97	0.00	0.97	0.00
78	184ARG	NH2	O3'	538DG	0.01	0.00	0.01	0.00
79	187ARG	NE	O1P	539DC	0.00	0.66	0.00	0.72
80	187ARG	NE	O2P	539DC	0.00	0.00	0.00	0.00
81	187ARG	NE	O5'	539DC	0.00	0.00	0.00	0.00
82	187ARG	NH1	N7	540DG	0.00	0.94	0.00	0.01
83	187ARG	NH1	O1P	540DG	0.97	0.00	0.89	0.00
84	187ARG	NH2	N7	540DG	0.00	0.59	0.00	0.00
85	187ARG	NH2	O1P	539DC	0.06	0.98	0.07	0.84
86	187ARG	NH2	O1P	540DG	0.01	0.00	0.00	0.00
87	187ARG	NH2	O2P	539DC	0.00	0.00	0.17	0.00
88	187ARG	NH2	O3'	538DG	0.00	0.00	0.00	0.22
89	187ARG	NH2	O5'	539DC	0.00	0.03	0.00	0.01
90	264DG	N2	NZ	73LYS	0.00	0.00	0.00	0.00
91	265DC	N4	O	397ALA	0.00	0.00	0.00	0.00

92	267DA	N6	ND2	399ASN	0.14	0.09	0.02	0.09
93	267DA	N6	OD1	399ASN	0.72	0.88	0.79	0.73
94	268DA	N6	NE	129ARG	0.00	0.02	0.04	0.00
95	268DA	N6	OD1	399ASN	0.25	0.23	0.14	0.13
96	271DC	N4	O	122ALA	0.39	0.45	0.03	0.92
97	271DC	N4	O	123ALA	0.15	0.11	0.00	0.00
98	274DG3	O3'	OD1	317ASP	0.00	0.00	0.00	0.00
99	274DG3	O3'	OD2	317ASP	0.00	0.00	0.00	0.00
100	343ASN	ND2	O1P	538DG	0.00	0.00	0.00	0.00
101	343ASN	ND2	O2P	537DC	0.00	0.01	0.00	0.00
102	343ASN	ND2	O2P	538DG	0.56	0.02	0.26	0.37
103	343ASN	ND2	O3'	537DC	0.02	0.02	0.05	0.00
104	343ASN	ND2	O5'	536DT5	0.00	0.00	0.01	0.00
105	343ASN	ND2	O5'	538DG	0.00	0.00	0.00	0.56
106	345SER	N	O2P	539DC	0.60	0.99	0.92	0.98
107	345SER	OG	O2P	539DC	0.00	0.00	0.00	0.00
108	347LYS	N	O2P	540DG	0.98	0.93	0.94	0.91
109	347LYS	N	O3'	539DC	0.09	0.04	0.26	0.26
110	347LYS	NZ	N2	272DG	0.00	0.00	0.00	0.00
111	347LYS	NZ	N2	538DG	0.00	0.00	0.00	0.00
112	347LYS	NZ	N3	274DG3	0.00	0.00	0.00	0.00
113	347LYS	NZ	O2	273DC	0.00	0.00	0.00	0.00
114	347LYS	NZ	O2	539DC	0.41	0.00	0.05	0.04
115	347LYS	NZ	O2P	274DG3	0.00	0.00	0.00	0.00

116	347LYS	NZ	O3'	274DG3	0.00	0.00	0.00	0.02
117	347LYS	NZ	O4'	274DG3	0.01	0.00	0.01	0.01
118	347LYS	NZ	O4'	539DC	0.01	0.00	0.02	0.01
119	347LYS	NZ	O4'	540DG	0.00	0.00	0.00	0.00
120	371LYS	NZ	O1P	541DA	0.02	0.03	0.01	0.28
121	371LYS	NZ	O2P	541DA	0.56	0.91	0.90	0.71
122	371LYS	NZ	O3'	540DG	0.00	0.00	0.00	0.01
123	372HIE	NE2	O1P	543DT	0.01	0.00	0.00	0.00
124	372HIE	NE2	O2P	543DT	0.34	0.61	0.00	0.70
125	372HIE	N	O1P	542DA	0.02	0.00	0.00	0.00
126	374GLY	N	O1P	543DT	0.47	0.97	0.87	0.26
127	374GLY	N	O2P	543DT	0.00	0.01	0.02	0.49
128	374GLY	N	O3'	542DA	0.00	0.00	0.00	0.00
129	375LYS	N	O1P	543DT	0.00	0.00	0.00	0.02
130	375LYS	N	O2P	543DT	0.00	0.00	0.00	0.00
131	375LYS	NZ	O1P	543DT	0.06	0.21	0.00	0.00
132	375LYS	NZ	O1P	544DT	0.22	0.79	0.11	0.01
133	375LYS	NZ	O2P	542DA	0.00	0.00	0.00	0.00
134	375LYS	NZ	O2P	543DT	0.14	0.00	0.43	0.42
135	375LYS	NZ	O2P	544DT	0.01	0.02	0.00	0.00
136	375LYS	NZ	O3'	543DT	0.01	0.01	0.00	0.00
137	375LYS	NZ	O5'	542DA	0.00	0.00	0.00	0.00
138	375LYS	NZ	O5'	543DT	0.20	0.75	0.01	0.04
139	388LYS	NZ	N4	263DC	0.00	0.00	0.00	0.00

140	388LYS	NZ	N4	265DC	0.00	0.00	0.00	0.00
141	388LYS	NZ	N4	547DC	0.00	0.00	0.06	0.00
142	388LYS	NZ	N7	264DG	0.00	0.00	0.00	0.00
143	388LYS	NZ	N7	545DA	0.00	0.00	0.00	0.02
144	388LYS	NZ	N7	545DG	0.00	0.00	0.01	0.00
145	388LYS	NZ	N7	546DG	0.00	0.00	0.44	0.12
146	388LYS	NZ	O1P	545DA	0.00	0.00	0.00	0.13
147	388LYS	NZ	O1P	545DC	0.12	0.00	0.00	0.00
148	388LYS	NZ	O1P	545DT	0.00	0.02	0.00	0.00
149	388LYS	NZ	O1P	546DG	0.01	0.02	0.00	0.03
150	388LYS	NZ	O2	262DT5	0.00	0.00	0.00	0.00
151	388LYS	NZ	O2P	545DA	0.00	0.00	0.00	0.01
152	388LYS	NZ	O2P	545DC	0.02	0.00	0.00	0.00
153	388LYS	NZ	O2P	545DT	0.00	0.04	0.00	0.00
154	388LYS	NZ	O2P	546DG	0.00	0.00	0.00	0.00
155	388LYS	NZ	O3'	544DT	0.00	0.00	0.00	0.00
156	388LYS	NZ	O3'	545DA	0.00	0.00	0.00	0.01
157	388LYS	NZ	O3'	545DT	0.00	0.00	0.00	0.00
158	388LYS	NZ	O4	262DT5	0.00	0.00	0.00	0.01
159	388LYS	NZ	O5'	262DT5	0.01	0.00	0.00	0.00
160	388LYS	NZ	O5'	545DA	0.00	0.00	0.00	0.02
161	388LYS	NZ	O5'	545DC	0.01	0.00	0.00	0.00
162	388LYS	NZ	O5'	545DT	0.00	0.01	0.00	0.00
163	388LYS	NZ	O6	264DG	0.00	0.00	0.20	0.02

164	388LYS	NZ	O6	545DG	0.00	0.00	0.00	0.00
165	388LYS	NZ	O6	546DG	0.00	0.00	0.38	0.10
166	388LYS	NZ	O6	548DG3	0.00	0.01	0.00	0.00
167	389ARG	NE	O1P	546DG	0.00	0.00	0.00	0.00
168	389ARG	NE	O4	262DT5	0.00	0.00	0.00	0.00
169	389ARG	NH1	O1P	546DG	0.00	0.00	0.12	0.00
170	389ARG	NH1	O2P	546DG	0.00	0.00	0.00	0.00
171	389ARG	NH1	O3'	545DG	0.00	0.00	0.00	0.00
172	389ARG	NH1	O4	262DT5	0.00	0.00	0.00	0.00
173	389ARG	NH1	O5'	262DT5	0.00	0.00	0.00	0.00
174	389ARG	NH2	O1P	546DG	0.00	0.00	0.10	0.00
175	389ARG	NH2	O2	262DT5	0.00	0.00	0.00	0.00
176	389ARG	NH2	O2P	546DG	0.00	0.00	0.01	0.00
177	389ARG	NH2	O4	262DT5	0.00	0.00	0.00	0.00
178	389ARG	NH2	O4'	262DT5	0.00	0.00	0.00	0.00
179	389ARG	NH2	O6	548DG3	0.00	0.00	0.00	0.00
180	392GLN	NE2	O5'	262DT5	0.00	0.00	0.00	0.00
181	399ASN	ND2	N7	266DA	0.00	0.05	0.00	0.00
182	399ASN	ND2	N7	266DG	0.07	0.00	0.00	0.00
183	399ASN	ND2	N7	267DA	0.97	0.93	0.93	0.93
184	399ASN	N	N6	267DA	0.00	0.00	0.00	0.00
185	399ASN	N	O4	543DT	0.39	0.70	0.60	0.60
186	399ASN	N	O4	544DT	0.00	0.00	0.00	0.00
187	400ALA	N	O4	543DT	0.37	0.60	0.61	0.29

188	403ARG	NE	N6	542DA	0.00	0.00	0.01	0.00
189	403ARG	NE	N7	542DA	0.32	0.74	0.28	0.05
190	403ARG	NH1	O1P	541DA	0.68	0.00	0.01	0.69
191	403ARG	NH2	N7	542DA	0.28	0.82	0.95	0.91
192	403ARG	NH2	O1P	541DA	0.98	0.92	1.00	0.97
193	403ARG	NH2	O5'	541DA	0.06	0.00	0.00	0.00
194	406LYS	NZ	O1P	540DG	0.83	0.91	0.76	0.06
195	406LYS	NZ	O2P	540DG	0.01	0.00	0.00	0.00
196	407ASN	ND2	O2P	540DG	0.02	0.92	0.94	0.00
197	453SER	OG	O1P	264DG	0.00	0.00	0.26	0.00
198	453SER	OG	O2P	264DG	0.00	0.00	0.00	0.00
199	454GLY	N	O1P	264DG	0.00	0.00	0.02	0.00
200	454GLY	N	O2P	264DG	0.34	0.63	0.41	0.00
201	455ILE	N	O1P	264DG	0.93	0.94	0.89	0.01
202	455ILE	N	O2P	264DG	0.00	0.00	0.00	0.00
203	458ARG	NE	O1P	265DC	0.00	0.00	0.00	0.00
204	458ARG	NH1	O1P	265DC	0.00	0.00	0.84	0.00
205	458ARG	NH1	O3'	264DG	0.00	0.00	0.00	0.00
206	458ARG	NH1	O4	266DT	0.00	0.00	0.00	0.02
207	458ARG	NH2	O1P	265DC	0.00	0.01	0.97	0.00
208	458ARG	NH2	O3'	264DG	0.00	0.00	0.03	0.00
209	461ARG	NE	O1P	265DC	0.42	0.42	0.00	0.00
210	461ARG	NE	O2P	265DC	0.00	0.00	0.00	0.00
211	461ARG	NH1	O1P	266DC	0.00	0.00	0.84	0.00

212	461ARG	NH1	O1P	266DT	0.00	0.00	0.00	0.00
213	461ARG	NH1	O2P	265DC	0.00	0.00	0.00	0.00
214	461ARG	NH2	O1P	265DC	1.00	1.00	0.08	0.03
215	461ARG	NH2	O2P	265DC	0.00	0.00	0.00	0.15
216	461ARG	NH2	O5'	265DC	0.01	0.00	0.01	0.07
217	46LYS	NZ	O2P	263DC	0.00	0.00	0.00	0.00
218	46LYS	NZ	O5'	262DT5	0.00	0.00	0.00	0.00
219	47LYS	NZ	O2P	265DC	0.00	0.00	0.00	0.00
220	536DT5	N3	NH1	115ARG	0.00	0.00	0.00	0.00
221	538DG	N2	NZ	347LYS	0.00	0.00	0.00	0.00
222	541DA	N6	ND2	125ASN	0.17	0.12	0.13	0.02
223	541DA	N6	OD1	125ASN	0.81	0.84	0.96	0.89
224	542DA	N6	OD1	125ASN	0.20	0.24	0.04	0.14
225	545DA	N6	O	396ALA	0.00	0.00	0.00	0.00
226	545DA	N6	O	397ALA	0.00	0.00	0.00	0.00
227	545DC	N4	O	397ALA	0.00	0.00	0.00	0.00
228	69ASN	ND2	O1P	263DC	0.00	0.00	0.01	0.00
229	69ASN	ND2	O1P	264DG	0.00	0.00	0.00	0.00
230	69ASN	ND2	O1P	265DC	0.00	0.00	0.00	0.01
231	69ASN	ND2	O2P	263DC	0.00	0.00	0.00	0.00
232	69ASN	ND2	O2P	264DG	0.44	0.01	0.86	0.09
233	69ASN	ND2	O2P	265DC	0.00	0.00	0.00	0.00
234	69ASN	ND2	O3'	263DC	0.01	0.04	0.00	0.00
235	69ASN	ND2	O3'	264DG	0.00	0.00	0.00	0.02

236	69ASN	ND2	O4'	264DG	0.00	0.00	0.00	0.00
237	69ASN	ND2	O5'	263DC	0.00	0.00	0.00	0.00
238	69ASN	ND2	O5'	264DG	0.00	0.00	0.00	0.01
239	71SER	N	O1P	265DC	0.00	0.00	0.00	0.00
240	71SER	N	O2P	265DC	0.91	0.97	0.87	0.73
241	71SER	OG	O2P	265DC	0.00	0.00	0.00	0.46
242	73LYS	N	O2P	266DA	0.00	0.95	0.00	0.00
243	73LYS	N	O2P	266DC	0.00	0.00	0.99	0.00
244	73LYS	N	O2P	266DG	0.99	0.00	0.00	0.00
245	73LYS	N	O2P	266DT	0.00	0.00	0.00	0.97
246	73LYS	N	O3'	265DC	0.01	0.04	0.08	0.03
247	73LYS	NZ	N3	548DG3	0.01	0.00	0.00	0.00
248	73LYS	NZ	O2	265DC	0.00	0.00	0.07	0.00
249	73LYS	NZ	O2	547DC	0.00	0.00	0.00	0.00
250	73LYS	NZ	O2P	548DG3	0.13	0.09	0.00	0.00
251	73LYS	NZ	O3'	266DG	0.00	0.00	0.00	0.00
252	73LYS	NZ	O3'	547DC	0.00	0.01	0.00	0.00
253	73LYS	NZ	O3'	548DG3	0.00	0.01	0.00	0.00
254	73LYS	NZ	O4'	265DC	0.00	0.00	0.02	0.00
255	73LYS	NZ	O4'	266DA	0.00	0.00	0.00	0.00
256	73LYS	NZ	O4'	266DC	0.00	0.00	0.00	0.00
257	73LYS	NZ	O4'	548DG3	0.02	0.00	0.04	0.00
258	73LYS	NZ	O5'	262DT5	0.00	0.00	0.00	0.00
259	97LYS	NZ	O1P	267DA	0.66	0.28	0.01	0.04

260	97LYS	NZ	O2P	267DA	0.22	0.58	0.66	0.56
261	97LYS	NZ	O3'	266DA	0.00	0.00	0.00	0.00
262	98HIE	NE2	O1P	269DT	0.02	0.00	0.00	0.02
263	98HIE	NE2	O2P	268DA	0.00	0.00	0.00	0.00
264	98HIE	NE2	O2P	269DT	0.51	0.12	0.00	0.71
265	98HIE	N	O1P	268DA	0.08	0.01	0.00	0.04

APPENDIX E: PUBLICATIONS & PRESENTATIONS

1. Vigneshwar Ramakrishnan, Srivatsan Jagannathan, Abdul Rajjak Shaikh and Raj Rajagopalan. Dynamic and Structural Changes in the Minimally Restructuring EcoRI Bound to a Minimally Mutated DNA Chain. Journal of Biomolecular Structure and Dynamics, 2012. 29(4)

2. Dhawal Shah, Aik Lee Tan, Vigneshwar Ramakrishnan, Jiang Jianwen and Raj Rajagopalan. Effect of Polydisperse Crowders on Aggregation Reactions: A Molecular Thermodynamic Analysis. Journal of Chemical Physics, 2011, 134, 064704

3. Karthik Harve, S, Vigneshwar Ramakrishnan, Raj Rajagopalan and Michael Raghunath. Macromolecular Crowding In Vitro as Means of Emulating Cellular Interiors: When Less Might be More. Proceedings of the National Academy of Sciences (PNAS), 2008: 105 (51):E119-E119

4. Vigneshwar Ramakrishnan and Raj Rajagopalan. Dynamics and Thermodynamics of Water around EcoRI Bound to a Minimally Mutated DNA Chain. *Submitted.*

CONFERENCE ORAL PRESENTATION

Vigneshwar Ramakrishnan, Soren Enemark and Raj Rajagopalan. *DNA basepair flexibility induced upon protein binding: Implications for protein-DNA specificity*. World Congress on Biomechanics, 2010.

REFERENCES

1. Moore, J.A., *Science as a Way of Knowing: The Foundations of Modern Biology*. 1999, Cambridge: Harvard University Press
2. Aristotle, *Generation of Animals. (Loeb Classical Library)*. 1943, Cambridge: Harvard University Press
3. Castle, W.E., Mendel's Law of Heredity. Science, 1903. 18(456): p. 396-406.
4. Hershey, A.D. and Chase, M., Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. The Journal of General Physiology, 1952. 36(1): p. 39-56.
5. Meselson, M. and Stahl, F.W., The replication of DNA in Escherichia coli. Proceedings of the National Academy of Sciences, 1958. 44(7): p. 671-682.
6. Tanford, C. and Reynolds, J., *Nature's Robots: A History of Proteins*. 2004, New York, USA: Oxford University Press
7. Hunter, G.K., *History of Protein Chemistry*, in *Encyclopedia of Life Sciences*. 2001, John Wiley & Sons, Ltd.
8. Jacob, F. and Monod, J., Genetic Regulatory Mechanisms in the Synthesis of Proteins. Journal of Molecular Biology, 1961. 3(3): p. 318-356.
9. Neidle, S., *Nucleic Acid Structure and Recognition*. 2002, Oxford: Oxford University Press
10. McPherson, J.D., Marco Marra, et al., A Physical Map of the Human Genome. Nature, 2001. 409(6822): p. 934-941.
11. Nirenberg, M., Leder, P., et al., RNA Codewords and Protein Synthesis, VII. On the General Nature of the RNA Code. Proceedings of the National Academy of Sciences, 1965. 53(5): p. 1161-1168.
12. Ptashne, M., Specific Binding of the lambda Phage Repressor to lambda DNA. Nature, 1967. 214(5085): p. 232-234.
13. Gilbert, W. and Müller-Hill, B., The lac Operator is DNA. Proceedings of the National Academy of Sciences, 1967. 58(6): p. 2415-2421.
14. Riggs, A.D., Bourgeois, S., and Cohn, M., The lac Represser-Operator Interaction: III. Kinetic Studies. Journal of Molecular Biology, 1970. 53(3): p. 401-417.
15. Winter, R.B., Berg, O.G., and Von Hippel, P.H., Diffusion-driven Mechanisms of Protein Translocation on Nucleic Acids. 3. The Escherichia coli lac Repressor-Operator Interaction: Kinetic Measurements and Conclusions. Biochemistry, 1981. 20(24): p. 6961-6977.
16. Berg, O.G., Winter, R.B., and Von Hippel, P.H., Diffusion-driven Mechanisms of Protein Translocation on Nucleic acids. 1. Models and Theory. Biochemistry, 1981. 20(24): p. 6929-6948.
17. Berg, O.G. and Ehrenberg, M., Association Kinetics with Coupled Three- and One-dimensional Diffusion : Chain-length Dependence of the Association Rate to Specific DNA Sites. Biophysical Chemistry, 1982. 15(1): p. 41-51.
18. von Hippel, P.H. and Berg, O.G., Facilitated Target Location in Biological Systems. Journal of Biological Chemistry, 1989. 264(2): p. 675-678.

19. Gorman, J. and Greene, E.C., Visualizing One-dimensional Diffusion of Proteins Along DNA. Nature, 2008. 15(8): p. 768-774.

20. Kochaniak, A.B., Habuchi, S., et al., Proliferating Cell Nuclear Antigen Uses Two Distinct Modes to Move along DNA. Journal of Biological Chemistry, 2009. 284(26): p. 17700-17710.

21. Jeltsch, A., Alves, J., et al., Pausing of the Restriction Endonuclease EcoRI during Linear Diffusion on DNA. Biochemistry, 1994. 33(34): p. 10215-10219.

22. Biebricher, A., Wende, W., et al., Tracking of Single Quantum Dot Labeled EcoRV Sliding along DNA Manipulated by Double Optical Tweezers. Biophysical Journal, 2009. 96(8): p. L50-L52.

23. Bonnet, I., Biebricher, A., et al., Sliding and Jumping of Single EcoRV Restriction Enzymes on Non-cognate DNA. Nucleic Acids Research, 2008. 36(12): p. 4118-4127.

24. Granéli, A., Yeykal, C.C., et al., Long-distance Lateral Diffusion of Human Rad51 on Double-stranded DNA. Proceedings of the National Academy of Sciences, 2006. 103(5): p. 1221-1226.

25. Gorman, J., Plys, A.J., et al., Visualizing One-dimensional Diffusion of Eukaryotic DNA Repair Factors Along a Chromatin Lattice. Nature Structural & Molecular Biology, 2010. 17(8): p. 932-938.

26. Ragunathan, K., Joo, C., and Ha, T., Real-Time Observation of Strand Exchange Reaction with High Spatiotemporal Resolution. Structure, 2011. 19(8): p. 1064-1073.

27. Slutsky, M. and Mirny, L.A., Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. Biophysical journal, 2004. 87(6): p. 4021-4035.

28. Stanford, N.P., Szczelkun, M.D., et al., One- and Three-dimensional Pathways for Proteins to Reach Specific DNA Sites. The EMBO Journal, 2000. 19(23): p. 6546-6557.

29. Adler, K., Beyreuther, K., et al., How lac Repressor Binds to DNA. Nature, 1972. 237(5354): p. 322-327.

30. Seeman, N.C., Rosenberg, J.M., and Rich, A., Sequence-specific Recognition of Double Helical Nucleic Acids by Proteins. Proceedings of the National Academy of Sciences, 1976. 73(3): p. 804-808.

31. Matthews, B.W., Protein-DNA Interaction: No Code for Recognition. Nature, 1988. 335(6188): p. 294-295.

32. Olson, W.K., Gorin, A.A., et al., DNA Sequence-dependent Deformability Deduced from Protein-DNA Crystal Complexes. Proceedings of the National Academy of Sciences, 1998. 95(19): p. 11163-11168.

33. Csermely, P., Palotai, R., and Nussinov, R., Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. Trends in Biochemical Sciences, 2010. 35(10): p. 539-546.

34. Boehr, D.D., Nussinov, R., and Wright, P.E., The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. Nature Chemical Biology, 2009. 5(11): p. 789-796.

35. Uil, T.G., Haisma, H.J., and Rots, M.G., Therapeutic Modulation of Endogenous Gene Function by Agents with Designed DNA-sequence Specificities. Nucleic Acids Research, 2003. 31: p. 6064-6078.
36. Arya, D.P. and Bruice, T.C., Triple-helix Formation of DNA Oligomers with Methylthiourea-linked Nucleosides (DNmt): A Kinetic and Thermodynamic Analysis. Proceedings of the National Academy of Sciences, 1999. 96(8): p. 4384-4389.
37. Escudé, C., Giovannangeli, C., et al., Stable Triple Helices Formed by Oligonucleotide N3'-->P5' Phosphoramidates Inhibit Transcription Elongation. Proceedings of the National Academy of Sciences, 1996. 93(9): p. 4365-4369.
38. Dagle, J.M. and Weeks, D.L., Positively Charged Oligonucleotides Overcome Potassium-Mediated Inhibition of Triplex DNA Formation. Nucleic Acids Research, 1996. 24(11): p. 2143-2149.
39. Ehrenmann, F., Vasseur, J., and Debart, F., Alpha-oligonucleotides with Anionic Phosphodiester and Cationic Phosphoramidate Linkages Enhanced Stability of DNA Triple Helix. Nucleosides Nucleotides Nucleic Acids, 2001. 20(4-7): p. 797-9.
40. Shimizu, M., Konishi, A., et al., Oligo(2'-O-methyl)ribonucleotides Effective Probes for Duplex DNA. FEBS Letters, 1992. 302(2): p. 155-158.
41. Escudé C, Sun JS, et al., Stable Triple Helices are Formed upon Binding of RNA Oligonucleotides and their 2'-O-methyl Derivatives to Double-helical DNA. C R Acad Sci III, 1992. 315: p. 521-525.
42. Torigoe, H., Hari, Y., et al., 2'-O,4'-C-Methylene Bridged Nucleic Acid Modification Promotes Pyrimidine Motif Triplex DNA Formation at Physiological pH. Journal of Biological Chemistry, 2001. 276(4): p. 2354-2360.
43. Lacroix, L., Arimondo, P.B., et al., Pyrimidine Morpholino Oligonucleotides Form a Stable Triple Helix in the Absence of Magnesium Ions. Biochemical and Biophysical Research Communications, 2000. 270(2): p. 363-369.
44. Lee, J.S., Woodsworth, M.L., et al., Poly(pyrimidine) poly(purine) Synthetic DNAs Containing 5-methylcytosine form Stable Triplexes at Neutral pH. Nucleic Acids Research, 1984. 12(16): p. 6603-6614.
45. Miller, P.S., Bi, G., et al., Triplex Formation by a Psoralen-Conjugated Oligodeoxyribonucleotide Containing the Base Analog 8-Oxo-Adenine. Nucleic Acids Research, 1996. 24(4): p. 730-736.
46. Cassidy, S.A., Slickers, P., et al., Recognition of GC Base Pairs by Triplex Forming Oligonucleotides Containing Nucleosides Derived from 2-Aminopyridine. Nucleic Acids Research, 1997. 25(24): p. 4891-4898.
47. Kers, I. and Dervan, P.B., Search for the Optimal Linker in Tandem Hairpin Polyamides. Bioorganic & Medicinal Chemistry, 2002. 10(10): p. 3339-3349.
48. Simon, M.D. and Shokat, K.M., Adaptability at a Protein-DNA Interface:Re-engineering the Engrailed Homeodomain to Recognize an Unnatural Nucleotide. Journal of the American Chemical Society, 2004. 126(26): p. 8078-8079.
49. Hall, B.M., Vaughn, E.E., et al., Reengineering Cro Protein Functional Specificity with an Evolutionary Code. Journal of Molecular Biology. 413(5): p. 914-928.
50. Ashworth, J., Havranek, J.J., et al., Computational Redesign of Endonuclease DNA Binding and Cleavage Specificity. Nature, 2006. 441(7093): p. 656-659.

51. Lesser, D.R., Kurpiewski, M.R., and Jen-Jacobson, L., The Energetic Basis of Specificity in the *EcoRI* Endonuclease-DNA Interactions. Science, 1990. 250(4982): p. 776-786.

52. Lesser, D.R., Kurpiewski, M.R., et al., Facilitated Distortion of the DNA Site Enhances *EcoRI* Endonuclease-DNA Recognition. Proceedings of the National Academy of Sciences, 1993. 90(16): p. 7548-7552.

53. Lankaš, F., Šponer, J., et al., DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. Biophysical journal, 2003. 85(5): p. 2872-2883.

54. Fujii, S., Kono, H., et al., Sequence-dependent DNA Deformability Studied Using Molecular Dynamics Simulations. Nucleic Acids Research, 2007. 35(18): p. 6063-6074.

55. Lesser, D.R., Grajkowski, A., et al., Stereoselective Interaction with Chiral Phosphorothioates at the Central DNA Kink of the *EcoRI* Endonuclease-GAATTC Complex. Journal of Biological Chemistry, 1992. 267(34): p. 24810-24818.

56. Heitman, J., How the EcoRI Endonuclease Recognizes and Cleaves DNA. BioEssays, 1992. 14(7): p. 445-454.

57. Rohs, R., Jin, X., et al., Origins of Specificity in Protein-DNA Recognition. Annual Review of Biochemistry, 2010. 79: p. 233-269.

58. Sarai, A. and Kono, H., Protein-DNA Recognition Patterns and Predictions. Annual Review of Biophysics and Biomolecular Structure, 2005. 34: p. 379-398.

59. Ball, P., Water as a Biomolecule. Chemical Physics and Physical Chemistry, 2008. 9(18): p. 2677-2685.

60. John WR, S., The Role of Water in Protein-DNA Interactions. Current opinion in Structural Biology, 1997. 7(1): p. 126-134.

61. Jayaram, B. and Jain, T., The Role of Water in Protein-DNA Recognition. Annual Review of Biophysics and Biomolecular Structure, 2004. 33(1): p. 343-361.

62. McClarin, J.A., Frederick, C.A., et al., Structure of the DNA-*EcoRI* Endonuclease Recognition Complex at 3 Å Resolution. Science, 1986. 234(4783): p. 1526-1542.

63. Heitman, J. and Model, P., Substrate Recognition by the *EcoRI* Endonuclease. Proteins: Structure, Function, and Genetics, 1990. 7(2): p. 185-197.

64. Frederick, C.A., Grable, J., et al., Kinked DNA in Crystalline Complex with *EcoRI* Endonuclease. Nature, 1984. 309(5966): p. 327-331.

65. Eisenmesser, E.Z., Bosco, D.A., et al., Enzyme Dynamics During Catalysis. Science, 2002. 295(5559): p. 1520-1523.

66. Wang, L., Pang, Y., et al., Functional Dynamics in the Active Site of the Ribonuclease Binase. Proceedings of the National Academy of Sciences, 2001. 98(14): p. 7684-7689.

67. Su, J.G., Xu, X.J., et al., An Analysis of the Influence of Protein Intrinsic Dynamical Properties on its Thermal Unfolding Behavior. Journal of Biomolecular Structure & Dynamics, 2011. 29(1): p. 105-121.

68. Chouard, T., Breaking the Protein Rules. Nature, 2011. 471: p. 151-153.

69. Martinez, R., Schwaneberg, U., and Roccatano, D., Temperature Effects on Structure and Dynamics of the Psychrophilic Protease Subtilisin S41 and its

Thermostable Mutants in Solution. Protein Engineering Design and Selection, 2011. 24(7): p. 533-544.

70. Kalodimos, C.G., Biris, N., et al., Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. Science, 2004. 305(5682): p. 386-389.

71. Cave, J.W., Kremer, W., and Wemmer, D.E., Backbone Dynamics of Sequence Specific Recognition and Binding by the Yeast Pho4 bHLH Domain Probed by NMR. Protein Science, 2000. 9(12): p. 2354-2365.

72. Brown, C., Campos-León, K., et al., Protein Flexibility Directs DNA Recognition by the Papillomavirus E2 Proteins. Nucleic Acids Research, 2011. 39(7): p. 2969-2980.

73. Doruker, P., Nilsson, L., and Kurkcuoglu, O., Collective Dynamics of *EcoRI*-DNA Complex by Elastic Network Model and Molecular Dynamics Simulations. Journal of Biomolecular Structure & Dynamics, 2006. 24(1): p. 1-15.

74. Uyar, A., Kurkcuoglu, O., et al., The Elastic Network Model Reveals a Consistent Picture on Intrinsic Functional Dynamics of Type II Restriction Endonucleases Physical Biology, 2011. 8: p. 056001.

75. Fuxreiter, M., Simon, I., and Bondos, S., Dynamic Protein-DNA Recognition: Beyond What can be Seen. Trends in Biochemical Sciences, 2011. 36(8): p. 415-423.

76. Papoian, G.A., Ulander, J., et al., Water in Protein Structure Prediction. Proceedings of the National Academy of Sciences, 2004. 101(10): p. 3352-3357.

77. Adkar, B.V., Jana, B., and Bagchi, B., Role of Water in the Enzymatic Catalysis: Study of ATP + AMP --> 2ADP Conversion by Adenylate Kinase. Journal of Physical Chemistry A, 2011. 115(16): p. 3691-3697.

78. Rhee, Y.M., Sorin, E.J., et al., Simulations of the Role of Water in the Protein-folding Mechanism. Proceedings of the National Academy of Sciences, 2004. 101(17): p. 6456-6461.

79. Kasson, P.M., Lindahl, E., and Pande, V.S., Water Ordering at Membrane Interfaces Controls Fusion Dynamics. Journal of Americal Chemical Society, 2011. 133(11): p. 3812-3815.

80. Gnanasekaran, R., Xu, Y., and Leitner, D.M., Dynamics of Water Clusters Confined in Proteins: A Molecular Dynamics Simulation Study of Interfacial Waters in a Dimeric Hemoglobin. Journal of Physical Chemistry B, 2010. 114(50): p. 16989-16996.

81. Ben-Naim, A., Molecular Recognition - Viewed Through the Eyes of the Solvent. Biophysical Chemistry, 2002. 101: p. 309-319.

82. Swaminathan, C.P., Nandi, A., et al., Thermodynamic Analyses Reveal Role of Water Release in Epitope Recognition by a Monoclonal Antibody against the Human Guanylyl Cyclase C Receptor. Journal of Biological Chemistry, 1999. 274(44): p. 31272-31278.

83. Swaminathan, C.P., Surolia, N., and Surolia, A., Role of Water in the Specific Binding of Mannose and Mannooligosaccharides to Concanavalin A. Journal of Americal Chemical Society, 1998. 120(21): p. 5153-5159.

84. Ahmad, M., Gu, W., et al., Adhesive Water Networks Facilitate Binding of Protein Interfaces. Nature Communications, 2011. 2: p. 261.

85. Jana, B., Pal, S., et al., Entropy of Water in the Hydration Layer of Major and Minor Grooves of DNA. Journal of Physical Chemistry B, 2006. 110(39): p. 19611-19618.
86. Venable, R., Zhang, Y., et al., Molecular dynamics simulations of a lipid bilayer and of hexadecane: an investigation of membrane fluidity. Science, 1993. 262(5131): p. 223-226.
87. Lin, S.-T., Maiti, P.K., and Goddard, W.A., Dynamics and Thermodynamics of Water in PAMAM Dendrimers at Subnanosecond Time Scales. Journal of Physical Chemistry B, 2005. 109(18): p. 8663-8672.
88. Reddy, C.K., Das, A., and Jayaram, B., Do Water Molecules Mediate Protein-DNA Recognition? Journal of Molecular Biology, 2001. 314(3): p. 619-632.
89. Otwinowski, Z., Schevitz, R.W., et al., Crystal Structure of *trp* Represser/Operator Complex at Atomic Resolution. Nature, 1988. 335(6188): p. 321-329.
90. Joachimiak, A., T E Haran, and Sigler, P.B., Mutagenesis supports water mediated recognition in the trp repressor-operator system. EMBO J., 1994. 13: p. 321.
91. Newman, M., Strzelecka, T., et al., Structure of Bam HI Endonuclease Bound to DNA: Partial Folding and Unfolding on DNA Binding. Science, 1995. 269(5224): p. 656-663.
92. Sidorova, N.Y. and Rau, D.C., *The Role of Water in EcoRI-DNA Binding*, in *Restriction Endonucleases*, A. Pingoud, Editor. 2004, Springer: Berlin. p. 319-337.
93. Luisi, B.F., Xu, W.X., et al., Crystallographic Analysis of the Interaction of the Glucocorticoid Receptor with DNA. Nature, 1991. 352(6335): p. 497-505.
94. Gewirth, D.T. and Sigler, P.B., The Basis for Half-site Specificity Explored Through a Non-cognate Steroid Receptor-DNA Complex. Nature Structural & Molecular Biology, 1995. 2(5): p. 386-394.
95. Sidorova, N.Y. and Rau, D.C., Differences in Water Release for the Binding of EcoRI to Specific and Nonspecific DNA Sequences. Proceedings of the National Academy of Sciences, 1996. 93(22): p. 12272-12277.
96. Sidorova, N.Y. and Rau, D.C., The Dissociation Rate of the EcoRI-DNA-Specific Complex is Linked to Water Activity. Biopolymers, 2000. 53: p. 363-368.
97. Duan, Y., Wilkosz, P., and Rosenberg, J.M., Dynamic Contributions to the DNA Binding Entropy of the *EcoRI* and *EcoRV* Restriction Endonucleases. Journal of Molecular Biology, 1996. 264(3): p. 546-555.
98. Jayaram, B., McConnell, K.J., et al., Free Energy analysis of Protein-DNA binding: The *EcoRI* endonuclease-DNA complex. Journal of Computational Physics, 1999. 151: p. 333-357.
99. Sen, S. and Nilsson, L., Structure, Interaction, Dynamics and Solvent Effects on the DNA-*EcoRI* Complex in Aqueous Solution from Molecular Dynamics Simulation. Biophysical Journal, 1999. 77: p. 1782-1800.
100. Lavery, R., Zakrzewska, K., et al., A Systematic Molecular Dynamics Study of Nearest-neighbor Effects on Base Pair and Base Pair Step Conformations and Fluctuations in B-DNA. Nucleic Acids Research, 2010. 38(1): p. 299-313.

101. Lankas, F., Sponer, J., et al., DNA deformability at the basepair level. Journal of the American Chemical Society, 2004. 126(13): p. 4124-4125.

102. Olmez, E.O. and Alakent, B., Alpha7 Helix Plays an Important Role in the Conformational Stability of PTP1B. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 675-693.

103. Xu, X., Su, J., et al., Thermal Stability and Unfolding Pathways of Sso7d and its Mutant F31A: Insight from Molecular Dynamics Simulation. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 717-727.

104. Zhou, Z.-L., Zhao, J.-H., et al., The Possible Structural Models for Polyglutamine Aggregation: A Molecular Dynamics Simulations Study. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 743-758.

105. Jani, V., Sonavane, U.B., and Joshi, R., Microsecond Scale Replica Exchange Molecular Dynamic Simulation of Villin Headpiece: An Insight into the Folding Landscape. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 845-860.

106. Zhang, J. and Li, D.D.W., Molecular Dynamics Studies on the Structural Stability of Wild-type Dog Prion Protein. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 861-869.

107. Zhuohang, M., Ji, L., and Hongwei, Y., Modeling of Transition State by Molecular Dynamics. Prediction of Catalytic Efficiency of the Mutants of Mandelate Racemase. Journal of Biomolecular Structure & Dynamics, 2011. 28: p. 871-879.

108. Varughese, J.F. and Li, Y., Molecular Dynamics and Docking Studies on Cardiac Troponin C. Journal of Biomolecular Structure & Dynamics, 2011. 29: p. 123-135.

109. Purohit, R., Rajendran, V., and Sethumadhavan, R., Studies on Adaptability of Binding Residues and Flap Region of TMC-114 Resistance HIV-1 Protease Mutants. Journal of Biomolecular Structure & Dynamics, 2011. 29: p. 137-152

110. Li, P., Liu, M., et al., Insight into the Inhibitory Mechanism and Binding Mode Between D77 and HIV-1 Integrase by Molecular Modeling Methods. Journal of Biomolecular Structure & Dynamics, 2011. 29: p. 311-323.

111. Behmard, E., Abdolmaleki1, P., et al., Prevalent Mutations of Human Prion Protein: A Molecular Modeling and Molecular Dynamics Study. Journal of Biomolecular Structure & Dynamics, 2011. 29: p. 379-389.

112. Wiesner, J., Kriz, Z., et al., Influence of the Acetylcholinesterase Active Site Protonation on Omega Loop and Active Site Dynamics. Journal of Biomolecular Structure & Dynamics, 2010. 28: p. 393-403.

113. Tao, Y., Rao, Z.-H., and Liu, S.-Q., Insight Derived from Molecular Dynamics Simulation into Substrate-Induced Changes in Protein Motions of Proteinase K. Journal of Biomolecular Structure & Dynamics, 2010. 28: p. 143-157.

114. Zhang, J., Chalmers, M.J., et al., DNA Binding Alters Coactivator Interaction Surfaces of the Intact VDR-RXR Complex. Nature Structural & Molecular Biology, 2011. 18(5): p. 556-563.

115. Lesser, D., Kurpiewski, M., and Jen-Jacobson, L., The Energetic Basis of Specificity in the Eco RI Endonuclease-DNA Interaction. Science., 1990. 250(4982): p. 776-786.

116. Kim, Y., Grable, J.C., et al., Refinement of EcoRI Endonuclease Crystal Structure: A Revised Protein Chain Tracing. Science., 1990. 249(4974): p. 1307-1309.
117. Rosenberg, J.M., Structure and Function of Restriction Endonucleases. Current Opinion in Structural Biology, 1991. 1: p. 104-113.
118. Guex, N. and Peitsch, M.C., SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. . Electrophoresis, 1997. 18: p. 2714-2723.
119. Hess, B., Kutzner, C., et al., GROMACS 4: Algorithms for Highly Efficient, Load-balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation, 2008. 4(3): p. 435-447.
120. Pérez, A., Marchán, I., et al., Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of [alpha]/[gamma] Conformers. Biophysical Journal, 2007. 92(11): p. 3817-3829.
121. Duan, Y., Wu, C., et al., A Point-charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-phase Quantum Mechanical Calculations. Journal of Computational Chemistry, 2003. 24(16): p. 1999-2012.
122. Jorgensen, W.L., Chandrasekhar, J., et al., Comparison of Simple Potential Functions for Simulating Liquid Water. Journal of Chemical Physics, 1983. 79: p. 926-935.
123. Joung, I.S. and Cheatham, T.E., Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. Journal of Physical Chemistry B, 2008. 112(30): p. 9020-9041.
124. Kurpiewski, M.R., Engler, L.E., et al., Mechanisms of Coupling between DNA Recognition Specificity and Catalysis in EcoRI Endonuclease. Structure, 2004. 12(10): p. 1775-1788.
125. Darden, T., York, D., and Pedersen, L., Particle mesh Ewald: An N-log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics, 1993. 98(12): p. 10089-10092.
126. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H.J.C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-alkanes. Journal of Computational Physics, 1977. 23(3): p. 327-341.
127. Huitema, H. and Liere, R.V. *Interactive Visualization of Protein Dynamics*. in *11th IEEE Visualization 2000 Proceedings*. 2000: IEEE Computer Society.
128. Barrett, C.P., Hall, B.A., and Noble, M.E.M., Dynamite: A Simple Way to Gain Insight into Protein Motions. Acta Crystallographica D, 2004. 60(12 Part 1): p. 2280-2287.
129. Humphrey, W., Dalke, A., and Schulten, K., VMD - Visual Molecular Dynamics. Journal of Molecular Graphics, 1996. 14: p. 33-38.
130. Lu, X.-J. and Olson, W.K., 3DNA: A Versatile, Integrated Software System for the Analysis, Rebuilding and Visualization of Three-dimensional Nucleic-acid Structures. Nature Protocols, 2008. 3(7): p. 1213-1227.
131. Lu, X.-J. and Olson, W.K., 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-dimensional Nucleic Acid Structures. Nucleic Acids Research, 2003. 31(17): p. 5108-5121.

132. Ramakrishnan, V., Jagannathan, S., et al., Dynamic and Structural Changes in the Minimally Restructuring EcoRI Bound to a Minimally Mutated DNA Chain. Journal of Biomolecular Structure & Dynamics, 2012. 29(4): p. 743-756.

133. Kurpiewski, M.R., Engler, L.E., et al., Mechanisms of Coupling between DNA Recognition Specificity and Catalysis in EcoRI Endonuclease. Structure., 2004. 12(10): p. 1775-1788.

134. Jen-Jacobson, L., Lesser, D., and Kurpiewski, M., The enfolding arms of EcoRI endonuclease: Role in DNA binding and cleavage. Cell, 1986. 45(4): p. 619-629.

135. Tarek, M. and Tobias, D.J., Role of Protein-Water Hydrogen Bond Dynamics in the Protein Dynamical Transition. Physical Review Letters, 2002. 88(13): p. 138101.

136. Grossman, M., Born, B., et al., Correlated Structural Kinetics and Retarded Solvent Dynamics at the Metalloprotease Active Site. Nature Structural & Molecular Biology, 2011. 18: p. 1102-1108.

137. Sinha, S.K. and Bandyopadhyay, S., Dynamic Properties of Water Around a Protein-DNA Complex from Molecular Dynamics Simulations. Journal of Chemical Physics, 2011. 135(13): p. 135101.

138. Ramakrishnan, V., Jagannathan, S., et al., Dynamic and Structural Changes in the Minimally Restructuring EcoRI Bound to a Minimally Mutated DNA Chain. Journal of Biomolecular Structure & Dynamics, 2011. 29(4).

139. Guex, N. and Peitsch, M.C., SWISS-MODEL and the Swiss-Pdb Viewer: An Environment for Comparative Protein Modeling. Electrophoresis, 1997. 18(15): p. 2714-2723.

140. Mukherjee, B., Maiti, P.K., et al., Jump Reorientation of Water Molecules Confined in Narrow Carbon Nanotubes. Journal of Physical Chemistry B, 2009. 113(30): p. 10322-10330.

141. Luzar, A., Resolving the Hydrogen Bond Dynamics Conundrum. Journal of Chemical Physics, 2000. 113(23): p. 10663.

142. Luzar, A. and Chandler, D., Hydrogen-bond Kinetics in Liquid Water. Nature, 1996. 379(6560): p. 55-57.

143. Lin, S.-T., Blanco, M., and Goddard, W.A., The Two-phase Model for Calculating Thermodynamic Properties of Liquids from Molecular Dynamics: Validation for the Phase Diagram of Lennard-Jones Fluids. Journal of Chemical Physics, 2003. 119(22): p. 11792-11805.

144. Lin, S.-T., Maiti, P.K., and Goddard, W.A., Two-Phase Thermodynamic Model for Efficient and Accurate Absolute Entropy of Water from Molecular Dynamics Simulations. Journal of Physical Chemistry B, 2010. 114(24): p. 8191-8198.

145. Sidorova, N.Y. and Rau, D.C., Removing Water from an EcoRI-Noncognate DNA Complex with Osmotic Stress. Journal of Biomolecular Structure & Dynamics, 1999. 17: p. 19-31.

146. Bhide, S.Y. and Berkowitz, M.L., The Behavior of Reorientational Correlation Functions of Water at the Water-lipid Bilayer Interface. Journal of Chemical Physics, 2006. 125(9): p. 094713.

147. Tan, H., Piletic, I.R., and Fayer, M.D., Orientational Dynamics of Water Confined on a Nanometer Length Scale in Reverse Micelles. Journal of Chemical Physics, 2005. 122(17): p. 174501.

148. Laage, D., Stirnemann, G., et al., Reorientation and Allied Dynamics in Water and Aqueous Solutions. Annual Review of Physical Chemistry, 2011. 62(1): p. 395-416.
149. Walrafen, G.E., Chu, Y.C., and Piermarini, G.J., Low-Frequency Raman Scattering from Water at High Pressures and High Temperatures. Journal of Physical Chemistry A, 1996. 100(24): p. 10363-10372.
150. Walrafen, G.E. and Chu, Y.C., Linearity between Structural Correlation Length and Correlated-Proton Raman Intensity from Amorphous Ice and Supercooled Water up to Dense Supercritical Steam. Journal of Physical Chemistry, 1995. 99(28): p. 11225-11229.
151. Debnath, A., Mukherjee, B., et al., Entropy and Dynamics of Water in Hydration Layers of a Bilayer. Journal of Chemical Physics, 2010. 133(17): p. 174704.
152. Williams, S.L., Parkhurst, L.K., and Parkhurst, L.J., Changes in DNA Bending and Flexing due to Tethered Cations Detected by Fluorescence Resonance Energy Transfer. Nucleic Acids Research, 2006. 34(3): p. 1028-1035.
153. DePaul, A.J., Thompson, E.J., et al., Equilibrium Conformational Dynamics in an RNA Tetraloop from Massively Parallel Molecular Dynamics. Nucleic Acids Research, 2010. 38(14): p. 4856-4867.
154. Sorin, E.J. and Pande, V.S., Exploring the Helix-coil Transition via All-atom Equilibrium Ensemble Simulations. Biophysical Journal, 2005. 88(4): p. 2472-2493.
155. Aaqvist, J., Ion-water Interaction Potentials Derived from Free Energy Perturbation Simulations. The Journal of Physical Chemistry, 1990. 94(21): p. 8021-8024.
156. Thielking, V., Alves, J., et al., Accuracy of the EcoRI Restriction Endonuclease: Binding and Cleavage Studies with Oligodeoxynucleotide Substrates Containing Degenerate Recognition Sequences. Biochemistry, 1990. 29(19): p. 4682-4691.
157. Jeltsch, A., Alves, J., et al., On the Catalytic Mechanism of EcoRI and EcoRV: A Detailed Proposal Based on Biochemical Results, Structural Data and Molecular Modelling. FEBS Letters, 1992. 304(1): p. 4-8.
158. Imhof, P., Fischer, S., and Smith, J.C., Catalytic Mechanism of DNA Backbone Cleavage by the Restriction Enzyme EcoRV: A Quantum Mechanical/Molecular Mechanical Analysis. Biochemistry, 2009. 48(38): p. 9061-9075.
159. Strauss, J.K. and MaherIII, L.J., DNA Bending by Asymmetric Phosphate Neutralization. Science, 1994. 266(5192): p. 1829-1834.
160. Strauss, J.K., Prakash, T.P., et al., DNA Bending by a Phantom Protein. Chemistry & Biology, 1996. 3(8): p. 671-678.
161. Urnov, F.D., Miller, J.C., et al., Highly Efficient Endogenous Human Gene Correction Using Designed Zinc-finger Nucleases. Nature, 2005. 435(7042): p. 646-651.
162. Porteus, M.H. and Baltimore, D., Chimeric Nucleases Stimulate Gene Targeting in Human Cells. Science, 2003. 300(5620): p. 763.
163. Bibikova, M., Carroll, D., et al., Stimulation of Homologous Recombination Through Targeted Cleavage by Chimeric Nucleases. Molecular and Cellular Biology, 2001. 21(1): p. 289-297.

164. Pastor, N., The B- to A-DNA Transition and the Reorganization of Solvent at the DNA Surface. Biophysical Journal, 2005. 88(5): p. 3262-3275.

165. Tikchonenko, T.I., Karamov, E.V., et al., EcoRI* Activity: Enzyme Modification or Activation of Accompanying Endonuclease? Gene, 1978. 4(3): p. 195-212.

166. Robinson, C.R. and Sligar, S.G., Molecular Recognition Mediated by Bound Water : A Mechanism for Star Activity of the Restriction Endonuclease EcoRI. Journal of Molecular Biology, 1993. 234(2): p. 302-306.

167. Sidorova, N.Y. and Rau, D.C., Differences Between *EcoRI* Nonspecific and "Star" Sequence Complexes Revealed by Osmotic Stress. Biophysical Journal, 2004. 87: p. 2564-2576.

168. Conlan, L.H., José, T.J., et al., Modulating Restriction Endonuclease Activities and Specificities Using Neutral Detergents. BioTechniques, 1999. 27(5): p. 955-960.

169. Chelli, R., Procacci, P., et al., Glycerol condensed phases Part I. A Molecular Dynamics Study. Physical Chemistry Chemical Physics, 1999. 1(5): p. 871-877.

170. Manning, G.S., Is a Small Number of Charge Neutralizations Sufficient to Bend Nucleosome Core DNA onto Its Superhelical Ramp? Journal of the American Chemical Society, 2003. 125(49): p. 15087-15092.

171. Hamelberg, D., Williams, L.D., and Wilson, W.D., Effect of a Neutralized Phosphate Backbone on the Minor Groove of B-DNA: Molecular Dynamics Simulation Studies. Nucleic Acids Research, 2002. 30(16): p. 3615-3623.

172. Okonogi, T.M., Alley, S.C., et al., Phosphate Backbone Neutralization Increases Duplex DNA Flexibility: A Model for Protein Binding. Proceedings of the National Academy of Sciences, 2002. 99(7): p. 4156-4160.

173. Strauss-Soukup, J.K., Vaghefi, M.M., et al., Effects of Neutralization Pattern and Stereochemistry on DNA Bending by Methylphosphonate Substitutions. Biochemistry, 1997. 36(29): p. 8692-8698.

174. Zhao, Y. and Truhlar, D., The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-class Functionals and 12 other Functionals. Theoretical Chemistry Accounts: Theory, Computation, and Modeling, 2008. 120(1): p. 215-241.

175. Zhao, Y. and Truhlar, D.G., Density Functionals with Broad Applicability in Chemistry. Accounts of Chemical Research, 2008. 41(2): p. 157-167.

176. Zhao, Y. and Truhlar, D.G., Density Functionals for Noncovalent Interaction Energies of Biological Importance. Journal of Chemical Theory and Computation, 2006. 3(1): p. 289-300.

177. Beveridge, D.L. and DiCapua, F.M., Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. Annual Review of Biophysics and Biophysical Chemistry, 1989. 18(1): p. 431-492.

178. Berens, P.H., Mackay, D.H.J., et al., Thermodynamics and quantum corrections from molecular dynamics for liquid water. Journal of Chemical Physics, 1983. 79: p. 2375-2389.

179. Henchman, R.H., Free Energy of Liquid Water from a Computer Simulation via Cell Theory. Journal of Chemical Physics, 2007. 126: p. 064504.

180. Karplus, M. and Kuriyan, J., Molecular dynamics and protein function. Proceedings of the National Academy of Sciences of the United States of America, 2005. 102(19): p. 6679-6685.
181. Shaw, D.E., Maragakis, P., et al., Atomic-Level Characterization of the Structural Dynamics of Proteins. Science. 330(6002): p. 341-346.
182. Pérez, A., Lankas, F., et al., Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Research, 2008. 36(7): p. 2379-2394.
183. Brandt, E.G. and Edholm, O., Dynamic Structure Factors from Lipid Membrane Molecular Dynamics Simulations. Biophysical journal, 2009. 96(5): p. 1828-1838.
184. Zuckerman, D.M., Equilibrium Sampling in Biomolecular Simulations. Annual Review of Biophysics, 2011. 40(1): p. 41-62.
185. Yamashita, H., Endo, S., et al., Sampling efficiency of molecular dynamics and Monte Carlo method in protein simulation. Chemical Physics Letters, 2001. 342: p. 382-386.
186. Gō, N. and Scheraga, H.A., *Analysis of the Contribution of Internal Vibrations to the Statistical Weights of Equilibrium Conformations of Macromolecules*. Vol. 51. 1969: AIP. 4751-4767
187. Gō, N. and Scheraga, H.A., On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation. Macromolecules, 1976. 9(4): p. 535-542.
188. Karplus, M. and Kushick, J.N., Method for estimating the configurational entropy of macromolecules. Macromolecules, 1981. 14(2): p. 325-332.
189. Chang, C.-E., Chen, W., and Gilson, M.K., Evaluating the Accuracy of the Quasiharmonic Approximation. Journal of Chemical Theory and Computation, 2005. 1(5): p. 1017-1028.
190. Hagai, M., Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. Current Opinion in Structural Biology, 2007. 17(2): p. 181-186.
191. White, R.P. and Meirovitch, H., Lower and Upper Bounds for the Absolute Free Energy by the Hypothetical Scanning Monte Carlo Method: Application to Liquid Argon and Water. Journal of Chemical Physics, 2004. 121: p. 10889-10904.
192. White, R.P. and Meirovitch, H., A simulation method for calculating the absolute entropy and free energy of fluids: Application to liquid argon and water. Proceedings of the National Academy of Sciences, 2004. 101(25): p. 9235-9240.
193. Lazaridis, T. and Karplus, M., Orientational Correlations and Entropy in Liquid Water. Journal of Chemical Physics, 1996. 105: p. 4294-4316.
194. Wang, L., Abel, R., et al., Thermodynamic Properties of Liquid Water: An Application of a Nonparametric Approach to Computing the Entropy of a Neat Fluid. Journal of Chemical Theory and Computation, 2009. 5(6): p. 1462-1473.
195. Tyka, M.D., Sessions, R.B., and Clarke, A.R., Absolute Free-Energy Calculations of Liquids Using a Harmonic Reference State. Journal of Physical Chemistry B, 2007. 111(32): p. 9571-9580.
196. Berens, P.H., Mackay, D.H.J., et al., Thermodynamics and Quantum Corrections from Molecular Dynamics for Liquid Water. The Journal of Chemical Physics, 1983. 79(5): p. 2375-2389.

197. McQuarrie, D.A., *Statistical Mechanics*. 2000: University Science Books, California

198. Carnahan, N.F. and Starling, K.E., Thermodynamic Properties of a Rigid-Sphere Fluid. Journal of Chemical Physics, 1970. 53: p. 600-603.