GDAŃSK UNIVERSITY
OF TECHNOLOGY
FACULTY OF CHEMISTRY

The author of the PhD dissertation: Kazi Amirul Hossain
Scientific discipline: Chemical Sciences

DOCTORAL DISSERTATION

Title of PhD dissertation: “Unraveling the Interplay between DNA and Proteins: A Computational Exploration of Sequence and Structure-Specific Recognition Mechanisms”

Title of PhD dissertation (in Polish): “Badanie molekularnych czynników wpływających na zdolność białek do rozpoznawania określonych sekwencji i struktur DNA: podejście obliczeniowe”

Supervisor *signature*
dr hab. inż. Jacek Czub, prof. uczelni

Gdańsk, 2023

GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF CHEMISTRY

STATEMENT

The author of the PhD dissertation: Kazi Amirul Hossain

I, the undersigned, agree/~~do not agree*~~ that my PhD dissertation entitled: "Unraveling the Interplay between DNA and Proteins: A Computational Exploration of Sequence and Structure-Specific Recognition Mechanisms" may be used for scientific or didactic purposes.[1]

Gdańsk,..... .....

signature of the PhD student

Aware of criminal liability for violations of the Act of 4th February 1994 on Copyright and Related Rights (Journal of Laws 2006, No. 90, item 631) and disciplinary actions set out in the Law on Higher Education (Journal of Laws 2012, item 572 with later amendments),[2] as well as civil liability, I declare, that the submitted PhD dissertation is my own work.

I declare, that the submitted PhD dissertation is my own work performed under and in cooperation with the supervision of dr hab. inż. Jacek Czub, ~~the second supervision of <name of the second supervisor>, the auxiliary supervision of <name of the auxiliary supervisor>, the cosupervision of <name of the cosupervisor>*~~.

This submitted PhD dissertation has never before been the basis of an official procedure associated with the awarding of a PhD degree.

All the information contained in the above thesis which is derived from written and electronic sources is documented in a list of relevant literature in accordance with art. 34 of the Copyright and Related Rights Act.

I confirm that this PhD dissertation is identical to the attached electronic version.

Gdańsk,..... .....

signature of the PhD student

I, the undersigned, agree/~~do not agree*~~ to include an electronic version of the above PhD dissertation in the open, institutional, digital repository of Gdańsk University of Technology, Pomeranian Digital Library, and for it to be submitted to the processes of verification and protection against misappropriation of authorship.

Gdańsk,..... .....

signature of the PhD student

*) delete where appropriate.

[1] Decree of Rector of Gdansk University of Technology No. 34/2009 of 9th November 2009, TUG archive instruction addendum No. 8.

[2] Act of 27th July 2005, Law on Higher Education: Chapter 7, Criminal responsibility of PhD students, Article 226.

GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF CHEMISTRY

DESCRIPTION OF DOCTORAL DISSERTATION

The Author of the PhD dissertation: Kazi Amirul Hossain

Title of PhD dissertation: “Unraveling the Interplay between DNA and Proteins: A Computational Exploration of Sequence and Structure-Specific Recognition Mechanisms”

Title of PhD dissertation in Polish: “Badanie molekularnych czynników wpływających na zdolność białek do rozpoznawania określonych sekwencji i struktur DNA: podejście obliczeniowe”

Language of PhD dissertation: English

Supervision: dr hab. inż. Jacek Czub

~~**Second supervision***: <first name, surname >~~

~~**Auxiliary supervision***: <first name, surname >~~

~~**Cosupervision***: <first name, surname >~~

Date of doctoral defense:

Keywords of PhD dissertation in Polish: Rozpoznanie DNA-białko, czynniki transkrypcyjne, symulacje dynamiki molekularnej, EXOG, G-kwadrupleks, DHX36

Keywords of PhD dissertation in English: DNA–protein recognition, transcription factors, MD simulations, EXOG, G-quadruplex, DHX36

Summary of PhD dissertation in Polish: Moja praca doktorska skupiona była na mechanizmach rozpoznawania sekwencji i konformacji DNA przez białka. Odkryłem, że kwasowe reszty aminokwasowe odgrywają ważną rolę w rozpoznaniu sekwencji DNA. Ich wkład w wiązanie zależy od obecności cytozyn, która równoważy odpychanie elektrostatyczne od szkieletu DNA poprzez oddziaływania specyficzne. Reszty kwasowe działają jako tzw. selektory negatywne, obniżając powinowactwo do sekwencji nie zawierających cytozyny, ale mogą także promować wiązanie do sekwencji docelowych, szczególnie gdy występuje w nich większa liczba cytozyn. Pokazałem także, że brak preferencji reszt kwasowych do adeniny wynika z elektrostatycznego odpychania z atomem N7 tej reszty. W części poświęconej rozpoznaniu konformacji DNA, odkryłem, że mitochondrialne białko EXOG preferuje formę A-DNA i selektywnie rozpoznaje dupleksy chimeryczne RNA/DNA. W rozpoznaniu tym główną rolę odgrywają specyficzne reszty argininowe. Ponadto zbadałem helikazę DHX36, która rozpoznaje DNA G-kwadrupleksy (G4) poprzez swoje subdomeny DSM i OB. Wiążą się one specyficznie i silnie do dwóch charakterystycznych cech strukturalnych równoległych G4: płaskiej powierzchni G-tetrady oraz konformacji szkieletu G-traktu. Co ważne, rozpoznawanie przez DSM odbywa się za pośrednictwem kontaktów van der Waalsa i oddziaływań hydrofobowych, wykazując preferencję dla dostępnej strony 5' struktury G4.

Summary of PhD dissertation in English: My PhD dissertation focused on DNA-protein interactions and the recognition of specific DNA sequences and structures. I discovered that acidic amino acid residues (Asp/Glu) play a crucial role by exhibiting a preference for cytosine. Their contribution to binding affinity depends on nearby cytosines, balancing electrostatic repulsion with specific interactions. Acidic residues act as negative selectors, discouraging non-cytosine binding, but can be favorable with increasing proximal cytosine count. They exclusively recognize cytosine

due to electrostatic repulsion with adenine's N7 atom and stronger hydrogen bonding. In another aspect of my research, I explored conformation-specific DNA recognition. I found that the EXOG protein prefers A-DNA and selectively recognizes RNA/DNA chimeric duplexes. Specific arginine residues induce the A-DNA conformation when EXOG binds to DNA/DNA duplexes, providing insights into mitochondrial replication and base excision repair. Furthermore, I investigated the DHX36 helicase, which recognizes G-quadruplexes (G4s) through its DSM and OB subdomains. The planar face of a G-tetrad and the specific backbone conformation of a G-tract are critical features in this interaction. The DSM and OB subdomains cooperatively recognize these distinctive features of parallel G4s. Importantly, the recognition by DSM is mediated through van der Waals contacts and hydrophobic interactions, exhibiting a preference for the accessible 5'-side of the G4 structure.

~~**Summary of PhD dissertation in language, in which it was written****: <summary, up to 1400 characters >~~

~~**Keywords of PhD dissertation in language, in which it was written****: <keywords>*~~

*) delete where appropriate.

**) applies to doctoral dissertations written in other languages, than Polish or English.

"DNA-protein interactions are the key to understanding how genetic information is translated into the diverse array of biological processes that shape our world. "

Sydney Brenner

Abstract

In my PhD dissertation, I explored the fascinating realm of DNA-protein interactions through computational studies. The research focuses on elucidating the mechanisms by which DNA-binding proteins recognize specific DNA sequences or structures, shedding light on important biological processes. One significant finding is the role of acidic amino acid residues (Asp/Glu) in DNA-protein recognition. Here I found that these residues exhibit a preference for cytosine and fulfill a diverse and context-dependent role in binding affinity. The contribution of acidic residues to DNA-binding affinity is delicately balanced between electrostatic repulsion from the DNA backbone and specific interactions with the cytosine nucleobase. Acidic residues function as negative selectors at non-cytosine sequences, discouraging binding. However, their contributions vary from negligible to significantly favorable as the number of cytosines in proximity increases. Additionally, my study revealed the reason why acidic residues exclusively readout cytosine and not adenine. This observation is explained by electrostatic repulsion with adenine's N7 atom and the stronger hydrogen bonding ability of cytosine, along with adenine's tendency to adopt the BII conformation. Another focal point of my research was understanding the molecular mechanisms of EXOG protein in recognizing specific DNA conformations. I observed that EXOG demonstrates a higher preference for A-DNA, as it selectively recognizes RNA/DNA chimeric duplexes. The interaction modes of specific arginine residues provided insights into the induction of the A-DNA conformation when EXOG binds to DNA/DNA duplex substrates. These findings contribute to our understanding of EXOG's role in mitochondrial replication and base excision repair processes. Furthermore, my investigation extended to the exploration of non-canonical DNA structures, particularly G-quadruplexes (G4s), and their recognition by the DHX36 helicase. In this context, I identified the critical participation of the DSM and OB subdomains in the recognition of G4s. Notably, the exposed planar face of a G-tetrad and the specific backbone conformation of a G-tract emerged as key features in this interaction. The DSM and OB subdomains cooperatively recognize these two distinctive features of parallel G4s, exhibiting high G4 affinity. Importantly, my findings highlight the significance of non-polar and van der Waals contacts in DNA recognition. I observed extensive van der Waals contacts of the GXXXG motifs and hydrophobic residues of DSM, specifically interacting with a flat guanine plane. Consequently, DSM exhibits a preference for binding to the more accessible 5′-side of the G4 structure.

Streszczenie

W mojej pracy doktorskiej zajmowałem się fascynującą tematyką oddziaływań białek z DNA, wykorzystując w tym celu podejście obliczeniowe. Moje badania były skoncentrowane na wyjaśnieniu mechanizmów, dzięki którym białka rozpoznają specyficzne sekwencje lub konformacje DNA. Jednym z moich odkryć jest ustalenie roli kwasowych reszt aminokwasowych (Asp/Glu) w rozpoznawaniu sekwencji DNA. Odkryłem, że reszty te wykazują preferencję dla cytozyny i mają zróżnicowaną i zależną od kontekstu rolę w powinowactwie do DNA. Wkład tych reszt kwasowych do energii swobodnej wiązania do DNA jest wynikiem delikatnej równowagi pomiędzy odpychaniem elektrostatycznym od szkieletu DNA a specyficznymi oddziływaniami z cytozyną. Reszty kwasowe działają jako tzw. ujemne selektory obniżając powinowactwo do sekwencji niecytozynowych. Z kolei w miejscach zawierających cytozynę korzystny wkład Asp/Glu nie polega jedynie na tworzeniu pojedynczego wiązania wodorowego, ale wymaga obecności pozytywnego potencjału generowanego przez większą liczbę cytozyn. Ponadto moje badania ujawniły powód, dla którego reszty kwasowe wykazuja silną preferencję do cytozynę a nie do adeniny. Wynika to z odpychania elektrostatycznego z atomem N7 adeniny oraz tendencji dinukleotydów purynowych do przyjmowania konformacji, w której grupa fosforanowa znajduje się bliżej środka dużego rowka (tzw. konformacja BII). Kolejnym szczegółowym problemem podjętym w pracy było zbadanie molekularnych mechanizmów rozpoznania substratów przez egzonukleazę EXOG, która pełni istotną funkcję w replikacji genomu mitochondrialnego. Zaobserwowałem, że EXOG wykazuje wyższą preferencję do A-DNA, co umożliwia jej specyficzne rozpoznanie chimerycznych dupleksów RNA/DNA. Zbadałem także szczegółowo oddziaływania leżące u podstaw tej preferencji mediowane przez określone reszty argininowe. Odkrycia te przyczyniają się do lepszego zrozumienia roli EXOG w replikacji mitochondrialnej i szlakach naprawy DNA przebiegających przez wycięcie zasady. Zakres badanych oddziaływań DNA-białko rozszerzyłem także na niekanoniczne struktury DNA, w szczególności tzw. G-kwadrupleksy (G4s) i ich rozpoznanie przez helikazę DHX36. W tym kontekście ustaliłem, że krytyczny udział w specyficznym wiązaniu G4 odgrywają subdomeny DSM i OB. Subdomeny te wspólnie rozpoznają te dwie charakterystyczne cechy równoległych G4: odsłonięta płaską powierzchnię tetrady guaninowej oraz specyficzną konformacją szkieletu traktów guaninowych. Moje odkrycia podkreślają znaczenie kontaktów niepolarnych i van der Waalsa w rozpoznawaniu DNA. Zaobserwowałem rozległe kontakty van der Waalsa motywów GXXXG i hydrofobowych reszt DSM, szczególnie oddziałujących z płaską płaszczyzną guaniny. W konsekwencji, DSM wykazuje preferencję do wiązania się z bardziej dostępną stroną 5′ struktury G4.

Acknowledgements

I am filled with immense gratitude as I reflect on my journey towards completing this PhD thesis, and I wish to express my deepest appreciation to those who have played a significant role in this remarkable chapter of my life.

First and foremost, I extend my heartfelt appreciation to my supervisor, Prof. Jacek Czub. Your guidance and support have been invaluable throughout this doctoral journey. Your constructive criticism, delivered with care and understanding, has fueled my growth and development as a researcher. I am truly grateful for the opportunity to work under your guidance, and I am profoundly thankful for the knowledge and wisdom you have imparted to me.

To my colleagues and fellow researchers at Gdańsk University of Technology, I extend my heartfelt appreciation. Our collaboration, discussions, and shared intellectual contributions have enriched my research experience. The stimulating scientific environment we shared has fostered my growth and provided invaluable insights and perspectives.

To my friends and family, I cannot express enough gratitude for their unwavering love, understanding, and encouragement throughout this challenging yet immensely rewarding endeavor. Their unwavering belief in my abilities and their constant presence have been my anchor, providing motivation and strength when I needed it most.

I would like to acknowledge the doctoral school at Gdańsk University of Technology and the SONATA BIS (2017/26/E/NZ2/00472) grant for their generous financial support, which has made my research and studies possible. Their investment in my education has been instrumental in the successful completion of this PhD thesis. I also extend my gratitude to the PL-Grid Infrastructure, TASK (Gdańsk), WCSS (Wrocław), and ICM (Warsaw) Centers for providing computational resources that have supported my research.

Kazi Amirul Hossain

Contents

List of Figures

List of Tables

List of Abbreviations

AIMD	**A**b **I**nitio **M**olecular **D**ynamics
BAR	**B**ennett **A**cceptance **R**atio
BER	**B**ase **E**xcision **R**epair
CSVR	**C**anonical **S**ampling through **V**elocity **R**escaling
CV	**C**ollective **V**ariable
DBD	**D**NA-**B**inding **D**omain
DBP	**D**NA-**B**inding **P**rotein
DNA	**D**eoxyribo **N**ucleic **A**cid
dsDNA	**d**ouble-**s**tranded **D**eoxyribo **N**ucleic **A**cid
EndoG	**E**ndonuclease **G**
EXOG	**E**xoonuclease **G**
FES	**F**ree **E**nergy **S**surface
ffG4	**f**ully-**f**olded **G**-quadruplex
GFI	**G**lobal **F**lexibility **I**ndex
HTH	**H**elix-**t**urn-**h**elix
LSP	**L**ight **S**trand **P**romoter
MD	**M**olecular **D**ynamics
MGME1	**M**itochondrial **G**enome **M**aintenance **E**xonuclease **1**
mtDNA	**m**itochondrial **D**eoxyribo **N**ucleic **A**cid
NHEJ	**N**on-**H**omologous **E**nd **J**oining
PBC	**P**eriodic **B**oundery **C**ondition
PCC	**P**earson **C**orrelation **C**oefficient
PDB	**P**rotein **D**ata **B**ank
PES	**P**otential **E**nergy **S**urface
PME	**P**article of **M**esh **E**wald
PMF	**P**otential of **M**ean **F**orce
puG4	**p**artially-**u**nfolded **G**-quadruplex
QM	**Q**uantum **M**echanics
RMSF	**R**oot **M**ean **S**quare **F**luctuations
RNA	**R**ibo **N**ucleic **A**cid
SMD	**S**teered **M**olecular **D**ynamics
SPME	**S**mooth **P**article of **M**esh **E**wald
ssDNA	**s**ingle-**s**tranded **D**eoxyribo **N**ucleic **A**cid
TBP	**T**ATA-**B**inding **P**rotein
TERRA	**TE**lomeric **R**epeat-containing **R**ibonucleic **A**cid
TF	**T**ranscription **F**actor
TFBS	**T**ranscription **F**actor-**B**inding **S**ite
TFO	**T**riplex **F**orming **O**ligonucleotide
TI	**T**hermodynamic **I**ntegration
US	**U**mbrella **S**ampling
WHAM	**W**eighted **H**istogram **A**nalysis **M**ethod
WT	**W**ild-**T**ype

Chapter 1

Scope and Goals

The study of DNA-protein recognition has a rich history that spans several decades. The discovery of the double helical structure of DNA by Watson and Crick in 1953 was a major milestone in the field, and set the stage for further research into the mechanisms of DNA-protein interactions. Exploring DNA-protein recognition is essential for understanding the fundamental mechanisms of life, including the regulation of gene expression, DNA replication, and repair. Continuous research in the field already achieved widespread success from the agricultural revolution to personalize medicine, and especially campaigns of vaccination which eradicated many life-threatening pathogenic diseases. For example, the identification of DNA-binding proteins in plants has led to the development of genetically modified crops that are resistant to pests and diseases. In medicine, DNA-protein interactions play a critical role in disease diagnosis, drug development, and personalized medicine.

Recent advancements in structure determination technologies, including cryo-EM, have greatly improved our understanding of DNA-protein interactions. For instance, the Protein Data Bank (PDB) contains nearly 5,000 experimentally determined structures of DNA-protein complexes to date, providing a valuable resource for researchers to study the structural and functional aspects of these interactions. Moreover, computational approaches have emerged as powerful tools for the analysis and interpretation of DNA-protein complex structures. These techniques leverage bioinformatic and molecular simulations, as well as machine learning methods to extract meaningful information from complex data sets, revealing new insights into the underlying mechanisms of DNA-protein interactions. As a result, we are now able to better understand how DNA-protein interactions contribute to the regulation of gene expression, genome stability, and other biological functions. However, we are still far from a complete understanding of how DNA-binding proteins discriminate their target binding sites from a vast excess of off-target sites.

For several decades now, extensive research has demonstrated that the binding between a target DNA substrate and its partner protein is a complex process that depends on several key features. One of the most critical features is the electrostatic attraction between the negatively charged sugar-phosphate backbone of DNA and positively charged residues at the binding site of the protein. This electrostatic interaction can provide general affinity, strong enough to bring the two molecules into close proximity, facilitating other interactions to occur. Moreover, entropic factors, such as the release of water molecules and the reduction of the degrees of freedom upon binding, also contribute to the energetics of protein-DNA interactions.

However, the repetitive nature of the sugar-phosphate backbone in DNA presents a challenge for site-specific DNA binding proteins in distinguishing between different DNA sequences. As a result, these proteins rely on differences in the structure and properties of the DNA bases themselves. The distinct patterns of hydrogen bond donors and acceptors within stacked arrays of DNA bases create recognizable patterns in the DNA grooves, which DNA-binding proteins selectively recognize through direct readout mechanism.

Furthermore, different arrays of bases can adopt varying DNA conformations or shapes and differences in desolvation-free energies, influencing the overall conformation and flexibility of the double helix. These specific shapes are also selectively recognized by DNA-binding proteins through indirect or shape readout mechanisms. For example, some proteins display a preference for the A-DNA conformation in RNA-DNA hybrids, which exhibits a subtle deviation from the canonical B-DNA structure. In another case, specialized helicases selectively bind to unique non-canonical DNA structures like G-quadruplexes (G4s).

Despite previous efforts, quantitatively determining the contributions of different factors, such as specific sequences and conformations, to the overall affinity of DNA-binding proteins remains challenging. It remains unclear how specific amino acid-DNA base contacts contribute to target site discrimination and how the direct readout efficiently targets a given DNA sequence. Moreover, the mechanisms underlying the recognition of specific DNA shapes, including non-canonical structures, by specific proteins have not been fully explored.

My dissertation is centered around an extensive computational analysis focused on specific aspects of DNA-protein recognition, with a particular emphasis on the direct and shape readout mechanisms. These mechanisms rely on distinct DNA sequences and structures and hold key significance in understanding the intricate process of DNA-protein interactions. Through my research, I aim to uncover novel deterministic factors that play pivotal roles in this multifaceted phenomenon.

To achieve this overarching goal, my research focuses on the following specific objectives:

1. Investigate the role of acidic amino acid residues (Asp/Glu) in DNA sequence specificity among different transcription factors. I explore the preferential interactions of these residues with specific nucleobases, contributing to the recognition of specific DNA sequences through the direct readout mechanism. This objective stems from my deep interest in understanding the fundamental principles underlying protein engineering.

2. Investigate human mitochondrial exonuclease-G (EXOG) and its ability to detect subtle conformational differences in the DNA double helix, with a preference for the A-DNA conformation commonly found in RNA-DNA hybrids. Unlike relying solely on a specific DNA sequence, EXOG recognizes a specific conformation, involving a chimeric shape of the DNA duplex where a portion adopts the canonical B-conformation and another portion adopts the A-conformation. This objective aims to unravel the mechanisms by which DNA-binding proteins recognize their target sites through shape recognition mechanisms.

3. Investigate the molecular mechanisms underlying the recognition of G-quadruplex (G4) structures by DHX36 helicase. G4 structures frequently occur at promoter

regions and serve as important regulatory genomic elements. This objective addresses the recognition mechanisms beyond the conventional DNA double helix and explores how partner proteins selectively recognize unique non-canonical shapes or structures of DNA.

While my research is primarily computational, I made an effort to connect it to existing experimental observations and data. Additionally, some of my results, particularly those related to the mechanism of EXOG, were confirmed through collaboration with experimentalists. By combining my findings with the current understanding of the field, I aim to offer valuable insights that can guide future research.

In this dissertation, Chapter 2 presents an in-depth review of the existing literature on specific and non-specific DNA-protein interactions, highlighting their importance in biological systems. It also explores various non-canonical DNA structures and examines the mechanisms by which DNA-binding proteins recognize and interact with them. By establishing this comprehensive background, Chapter 2 lays the groundwork for the subsequent discussion of my own research in the dissertation.

Chapter 3 not only provides an inventory of the computational methods employed in my doctoral research but also delves into the theoretical foundations that underpin the work of a computational biophysicist. It offers a thorough exploration of the principles and techniques essential to understanding the computational aspects of the field, setting the stage for the subsequent chapters.

Finally, Chapter 4 presents the detailed results of my research. It encompasses the findings, analyses, and interpretations derived from my computational investigations into DNA-protein interactions. The chapter offers a comprehensive exploration of the data obtained and discusses their implications within the broader context of the field.

Through this dissertation, I aspire to make a positive impact on the disciplines of computational biophysics, molecular biology, and bioengineering. By presenting a detailed account of my research and providing a comprehensive review of the literature, I believe this work can serve as a valuable resource for aspiring young scientists embarking on similar paths in these dynamic and rapidly evolving fields. It is my sincere hope that the knowledge and insights shared in this dissertation will contribute to the advancement of scientific understanding and inspire further research and innovation in the fascinating realm of DNA-protein interactions.

Chapter 2

Background and Introduction

2.1 Structural Biology of DNA-Protein Recognition

2.1.1 Canonical Structures of DNA

The double-helix structure of DNA, discovered by James Watson and Francis Crick in 1953, is a well-known and widely studied aspect of DNA[1]. However, DNA can adopt different structures depending on its sequence and environmental conditions [2]. The canonical B-DNA structure is the most common and well-known structure, with a right-handed double helix composed of antiparallel nucleotide strands. The base pairs are oriented perpendicular to the axis of the helix, with hydrogen bonds linking complementary nucleotides [2]. B-DNA exhibits remarkable stability owing to the presence of multiple hydrogen bonds between complementary base pairs, as well as the π-stacking interactions among adjacent base pairs. These bonds and interactions help to keep the two strands of the double helix firmly bound together, making B-DNA resistant to denaturation in challenging environmental conditions [3]. The major and minor grooves result from the asymmetric distribution of atoms along the helix axis and play an important role in DNA-protein interactions [4].

The structural parameters of DNA can be broadly classified into two categories: intra-base pair parameters and inter-base pair parameters [6, 7]. Intra-base pair parameters, including buckle, propeller, opening, shear, stretch, and stagger, describe the rotation and displacement within a base pair (see Fig. 2.1A), going from the complementary base to the reference one [7]. These parameters are important for understanding the local structural variations within a single base pair and how they contribute to the overall conformation of the DNA molecule[6, 7]. On the other hand, inter-base pair parameters: tilt, roll, twist, shift, slide, and rise capture the rotation and displacement from one base pair to the next one (see Fig. 2.1B) [7], following the 5′ to 3′ direction of the reference strand. These parameters are important for understanding the global structural variations along the DNA molecule and how they affect the interactions between DNA and other molecules such as proteins [8].

The global flexibility index (GFI) of a DNA molecule in a particular conformation can be determined by analyzing its structural helical parameters (Fig. 2.1) [9]. This measure quantifies the overall stiffness or flexibility of the DNA molecule, and can be used to understand the behavior of DNA in various contexts, including DNA-protein interactions. Studies have shown that the GFI of DNA is strongly dependent on the DNA sequence, with GC base pairs being generally less deformable than AT base pairs when considering intra-base pair coordinates such as internal translations and rotations [10]. Conversely, when examining inter-base pair parameters, it has been observed that AT base pairs, particularly the central AT,AT step, exhibit greater

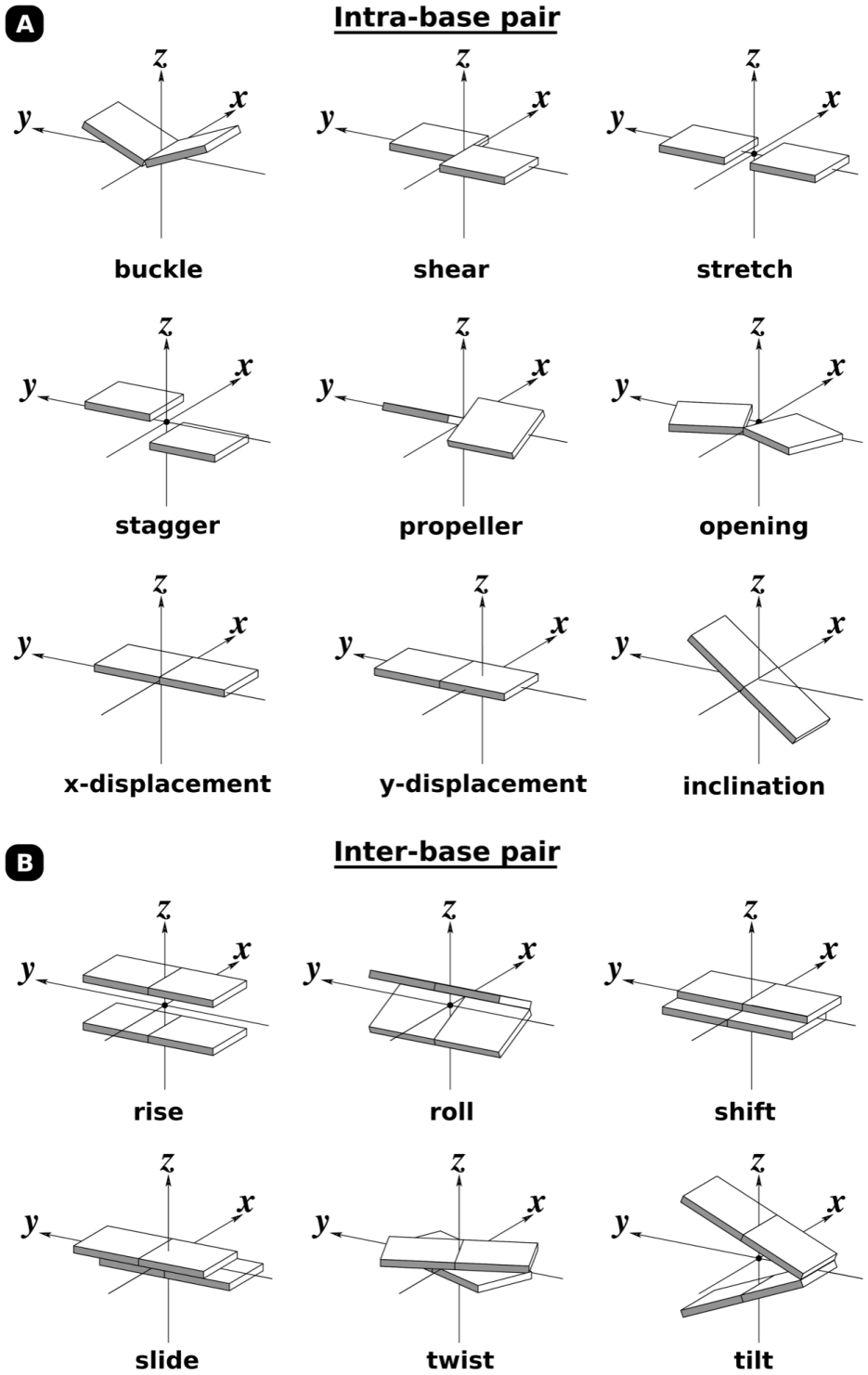

FIGURE 2.1: A schematic representation of DNA **(A)** intra-base and **(B)** inter-base pair parameters. This figure is adapted from reference [5].

rigidity compared to the CG,CG step. This rigidity is attributed to both translational movements, primarily slide, and rotational movements, mainly twist and roll [10].

The conformational freedom of the DNA backbone is primarily governed by three major elements: (i) sugar puckering, (ii) rotations around ζ/ϵ torsions, and (iii) rotations around α/γ torsions [10]. The torsions associated with the dinucleotide steps in the sugar-phosphate backbone of DNA are defined as follows: α (O3′—P—O5′—C5′), β (P—O5′—C5′—C4′), γ (O5′—C5′—C4′—C3′), δ (C5′—C4′—C3′—C2′), ϵ (C4′—C3′—O3′—P) and ζ (C3′—O3′—P—O5′).

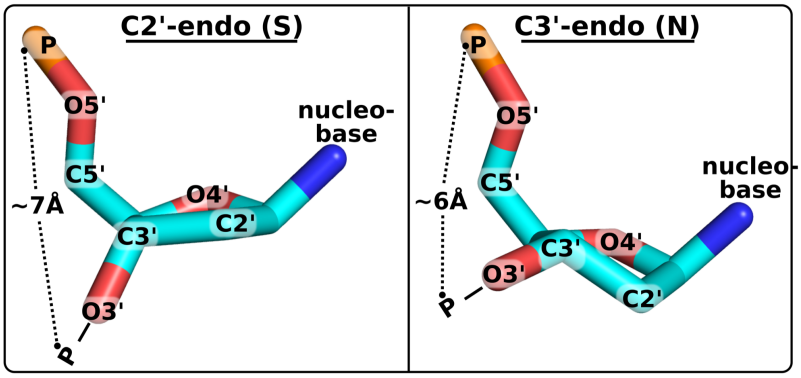

FIGURE 2.2: Structural representation of C2′-endo (South, S) and C3′-endo (North, N) sugar puckering of DNA.

Sugar puckering refers to the distortion of the sugar ring from its relatively flat, planar conformation. The most common puckering conformations are C2′-endo (South, S) and C3′-endo (North, N), which are characterized by the conformational changes within the sugar ring itself (see Fig. 2.2), involving the rotation of specific angles. These puckering conformations are primarily determined by the internal properties of the sugar ring, but they can also be slightly influenced by the rotation of the glycosidic bond between the sugar and the nucleobase [11]. This conformational change can affect the position and orientation of the neighboring bases and, consequently, the overall structure of the DNA molecule. In B-DNA, the majority of C2′-deoxyriboses are in the South conformation (see Fig. 2.2 *left*), with North puckering occurring only rarely. Studies have also shown that North conformers are more prevalent in pyrimidine nucleosides, particularly in cytidine, than in purines [10]. The transition from the South to North conformation can significantly affect the groove topology of the DNA molecule and, consequently, influence the binding of proteins. When multiple North puckers occur simultaneously, there is a moderate increase in the width of the minor groove, which becomes particularly evident [12].

B-DNA also exhibits another significant backbone structural polymorphism arising from its ability to adopt two distinct conformations, known as BI and BII, resulting from the concerted rotation around the ζ/ϵ torsions (see Fig. 2.3). The BI conformation features a trans/gauche-(t/g-) conformation, while the BII state has a gauche-/trans (g-/t) conformation [13]. The population of BI and BII conformations is sequence-dependent and purine–purine steps have a much higher propensity to adopt the BII conformation. The difference between the BI and BII conformations is subtle and does not produce significant distortions in the DNA duplex. Nonetheless, the ability of DNA to sample different backbone conformations may have important implications for its interactions with proteins and other molecules. In fact, this polymorphism holds particular relevance for the subsequent chapter, as one of the discussed results, pertaining to research objective 1, delves into the mechanism

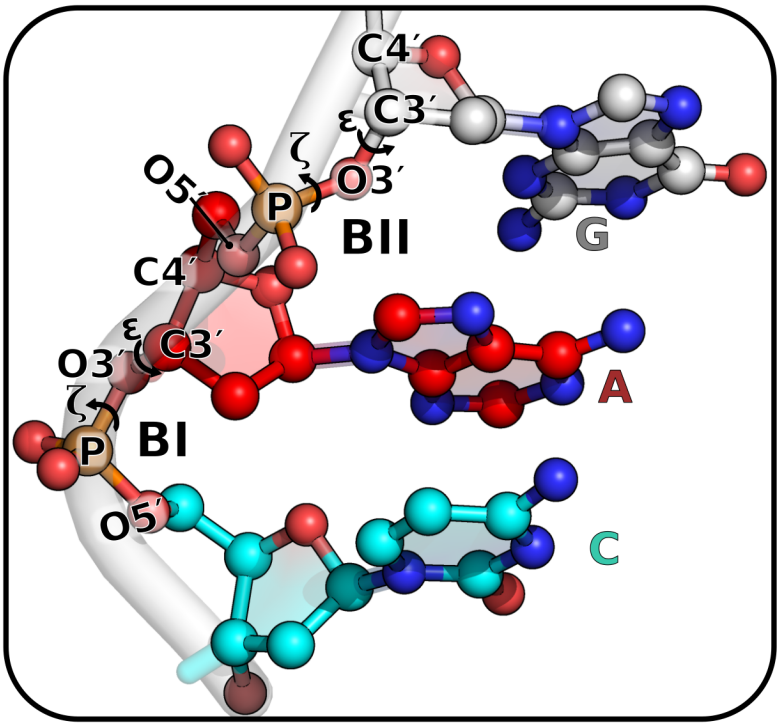

FIGURE 2.3: Structural representation of the BI and BII conformations of the DNA backbone. BI/BII conformations are typically distinguished using the difference between the ϵ (C4′—C3′—O3′—P) and ζ (C3′—O3′—P—O5′) torsion angles in a given dinucleotide step, as the defining feature. Specifically, $\epsilon - \zeta < 0^\circ$ and $\epsilon - \zeta > 0^\circ$ correspond to BI and BII, respectively.

by which acidic amino acid residues exhibit a preferential interaction with cytosine rather than adenine.

Conformational changes around the α/γ torsions are less common, occurring only 1.5% of the time [14], but they may play an important role in certain cases of DNA-protein interactions. The probability of deviations from the canonical α/γ torsions is also sequence-dependent, with a lower likelihood of such deviations in the central AATT tetramer, for instance [10]. The presence of non-canonical α/γ torsions has a significant impact on twist, but it has little effect on other helical parameters [15].

A-DNA Conformation and its Implication in RNA/DNA Hybrid Structures

The A-DNA conformation, which is considered to be one of the two canonical DNA conformations, is typically induced by the specific geometry of the ribose sugar and the nucleobase in nucleic acids [18]. This conformation is characterized by a right-handed anti-parallel helix but with a narrower major groove and a wider minor groove than the more common B-DNA conformation (see Fig. 2.4). The diameter and pitch of the A-DNA helix are approximately 26 and 28 Å , respectively, while for B-DNA 20 and 33 Å , respectively [19]. Additionally, the base pairs in A-DNA are tilted at an angle of approximately 20° with respect to the helix axis, resulting in a more compact and twisted structure (see Fig. 2.4, *top view*) [19]. The major factor determining the A-DNA conformation is the sugar puckering, i.e., C3′-endo, while in B-DNA, it is in the C2′-endo conformation (see Fig. 2.2). As a result, A-DNA has

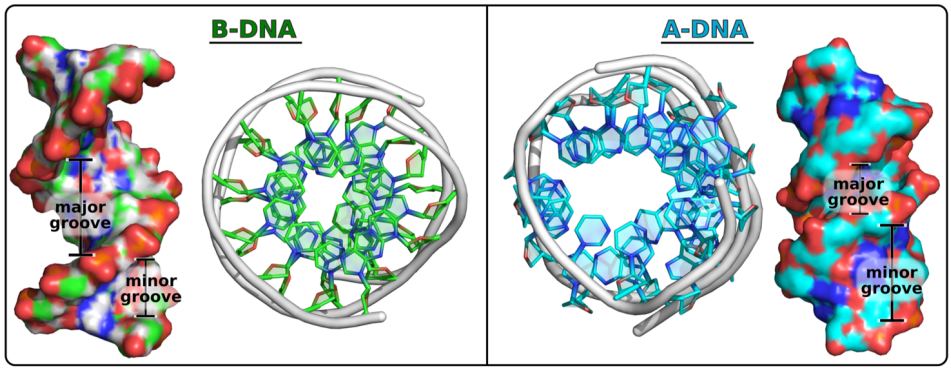

FIGURE 2.4: Structural representation of the A- and B-DNA conformations, shown in side view as surface and top view as cartoon-stick, highlighting the major and minor grooves by labeling. The A-DNA and B-DNA structures were taken from crystallographic data with PDB ID: 3QK4 [16] and 5T5C [17], respectively.

a shorter distance (by approximately 1 Å) between phosphate groups than B-DNA, which can lead to a strong repulsion between the phosphate groups in the backbone under low salt conditions, making this conformation less stable than B-DNA. In addition, the narrower major groove also leads to repulsion in the backbone see Fig. 2.4, *side view*). However, at high salt conditions, the repulsion is screened by interactions with ions, making the A-DNA conformation more stable [20, 21].

Although the B-DNA conformation is more prevalent in biological systems, certain DNA-binding proteins have been found to recognize the subtle differences in the A-DNA conformation and preferentially form specific complexes with it. For example, some minor groove binding transcription factors, polyamides, and nucleases exhibit a preference for the A-DNA conformation [22–24]. Therefore, A-DNA is believed to play a role in DNA replication, gene regulation, and DNA damage recognition [25].

The A-DNA conformation is frequently observed in RNA/DNA hybrids, which arise when an RNA strand pairs with a complementary DNA strand. Evidence from X-ray fiber diffraction and CD spectroscopy confirms the existence of RNA/DNA hybrids in important biological processes such as transcription (DNA-mediated RNA synthesis) and replication (particularly in lagging strands, where Okazaki fragments are temporarily stabilized through the formation of RNA/DNA hybrids) [26]. Furthermore, RNA/DNA hybrids are known to occur as intermediates during the reverse transcription process, making them an attractive target for small molecule therapies aimed at treating diseases caused by HIV and other retroviruses [27].

Chimeric RNA/DNA hybrids are formed when a helix of RNA/DNA hybrid is joined to double-stranded DNA. Several studies have confirmed the presence of a bend that joins two distinctly unique conformations within the polymers, as well as in transcription-replication machinery [26, 28]. The hybrid region of the duplex generally adopts the C3′-endo conformation in the ribose sugars and therefore tends to be in the A-conformation, while the rest of the DNA duplex exists mainly in the B-form. As a result, chimeric hybrids are polymorphic and contain contributions from each segment that are conformationally distinct.

It is also noteworthy that both RNA/DNA hybrids and chimeric hybrids are present during the replication of human mitochondrial DNA (mtDNA), where RNA/DNA hybrids serve as primers [17]. This fact is of particular relevance to the later chapter, as one of the discussed results, pertaining to research objective 2, focuses on the recognition mechanism of the A-DNA conformation formed by RNA/DNA chimeric hybrids by human mitochondrial exonuclease-G (EXOG).

2.1.2 Non-canonical Structures of DNA

It is fascinating to note that every mystery of nature is a unique and intriguing puzzle to solve. This holds true for the structures of genetic entities such as DNA and RNA as well, which do not always follow the Watson-Crick canon, leading to the identification of exceptional structures known as non-canonical structures. One such exceptional structure was discovered by Karst Hoogsteen [29], which involved a different base-pairing mechanism named "Hoogsteen base-pairs". In this type of pairing, adenine (A) and guanine (G) bases flip upside down to their *syn* conformation instead of the conventional *anti* conformation in Watson-Crick pairing.

Although non-canonical structures of DNA are less common than the conventional B-form in the cell, there is evidence of the existence of more than 20 types of non-canonical structures, including some popular ones such as G-quadruplex (G4), i-motif, triplex, hairpins, and cruciforms, etc [30]. It is worth noting that these structures, despite being less frequent than duplex in the cell, have a profound impact on gene expression, and thus are associated with various genetic disorders, including cancer [31]. Therefore, it is essential to study and understand these non-canonical structures of DNA to comprehend their role in gene regulation and their potential implications in disease pathology.

In 1910, there was a hint that a concentrated solution of guanylic acid can form a gel at low pH [32]. However, it took almost 50 years and the help of X-ray diffraction to identify the planar hydrogen-bonded arrangement of four guanine residues, now known as a G-quartet [33]. Later, the structure of a G-quadruplex, consisting of a stack of several G-quartets, was discovered in G-rich oligonucleotides, and it was proposed that they could have biological relevance based on their location in the genome [34].

Another type of four-stranded structure of DNA, known as i-motif, was discovered, which forms in a cytosine (C) enriched sequence through a stack of intercalating hemiprotonated C-C^+ pairs in an acidic pH environment. Like G-quadruplexes (discussed below in detail, section 2.1.2), i-motifs have been found in telomeres, centromeres, and promoter regions of proto-oncogenes [35, 36].

In 1953, Pauling and co-workers predicted the existence of DNA-triplex, where a third strand, called a triplex forming oligonucleotide (TFO), binds to the duplex [37]. Four years later, Rich et al. experimentally demonstrated the formation of DNA triplexes [38]. These structures are typically formed by homopurine-homopyrimidine sequences. To form the triad, TFO binds to the major groove of a duplex and is stabilized by Hoogsteen-base pairing. DNA triplexes have been implicated in a variety of biological processes, including gene regulation, recombination, and genome instability. They have also been considered potential therapeutic targets for a range of diseases [39].

Palindromic sequences, which contain inverted repeats, have internal symmetry, and as a result, have also the ability to form hairpins or cruciform structures by switching inter- and intra-strand base pairings. Hairpins are formed when a single DNA strand folds back on itself, while cruciform structures occur when both strands fold back on themselves. These structures play an important role in various biological processes, such as recombination [40].

Given the high biological significance of non-canonical structures of DNA, they have received increasing attention over the past several years, especially since the 2000s. Much has been learned about the structural behavior of these structures, including their formation and stability. However, there is still much to be understood about their mechanistic interactions with partner proteins.

Further research is needed to fully apprehend the complexity of these interactions and their role in biological processes such as gene regulation, recombination, and genome stability. Developing a more comprehensive understanding of these interactions could provide insights into potential therapeutic strategies for a range of diseases, including cancer and genetic disorders. As such, continued research in this field is crucial for advancing our understanding of DNA structure and function, and its potential applications in medicine.

G-quadruplex: A Prominent Non-Canonical Nucleic Acid Structure

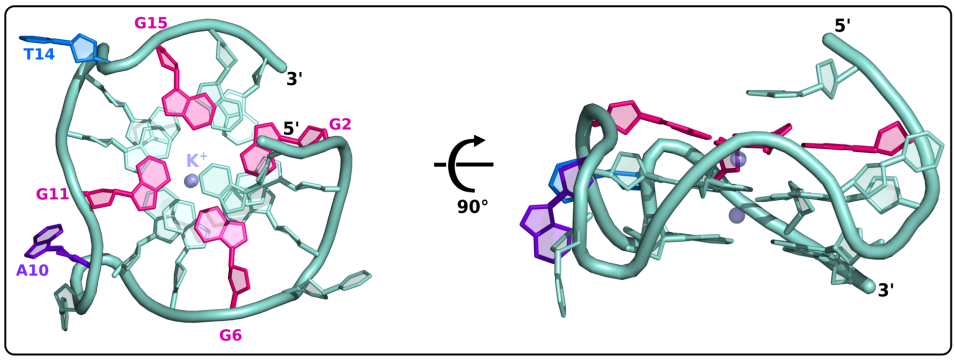

FIGURE 2.5: Structural representation of DNA^{myc}-G4, highlighting a G-quartet plane (top or 5′) in pink. The structure is shown in both side and top views, with potassium ions (K^+) represented as semi-transparent spheres. The structure was taken from PDB ID: 1XAV [41].

G-quadruplexes (G4s) are non-canonical secondary structures formed by nucleic acids, where four guanine bases associate via Hoogsteen hydrogen bonding to form a planar G-quartet. The stacking of a few G-quartets on top of each other in the presence of monovalent cations (preferably K^+) forms the G4 core, which is stabilized by a complex network of hydrogen bonding and stacking interactions (see Fig. 2.5) [34]. The topology of G4 structures is primarily described based on the direction of the strands, such as parallel (if all strands proceed in the same direction), antiparallel, and hybrid-type (mixed parallel/antiparallel), as shown in Fig. 2.6A [42].

The diversity in topology is driven by several intrinsic properties, such as the base composition of the intervening sequences (termed “loops”, see Fig. 2.6B) and

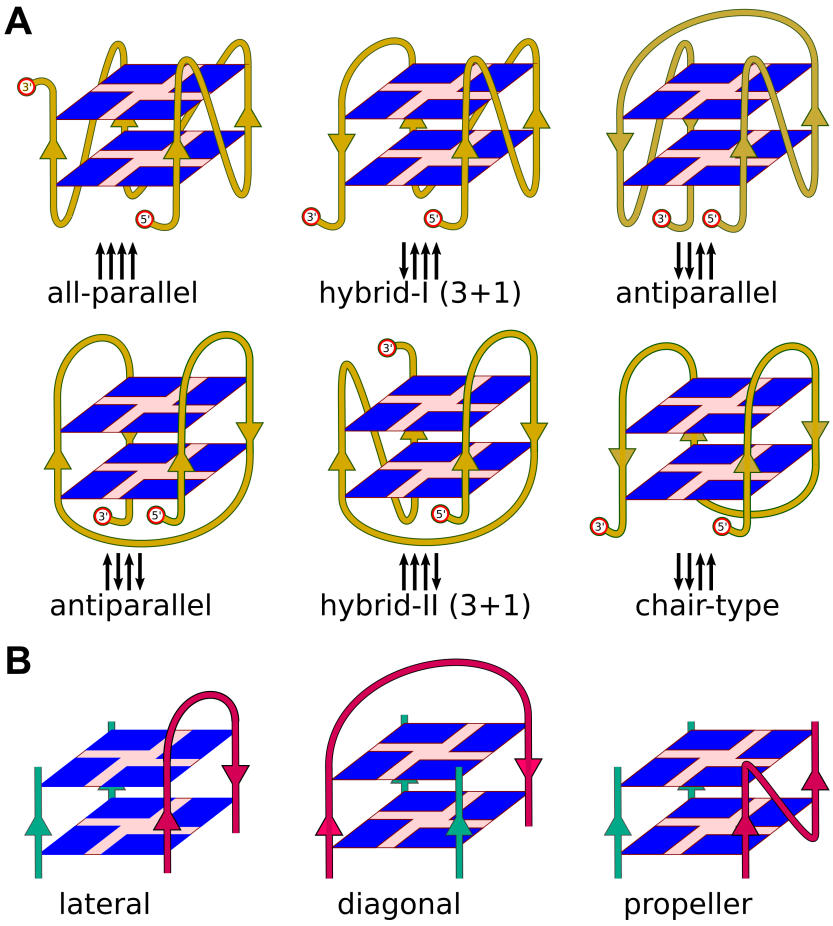

FIGURE 2.6: **(A)** Schematic representation of the different types of topologies observed in G4 DNA. Rectangular planes depict the G-quartets, and the arrows indicate the direction of the strand. Loops are also shown for each topology. **(B)** Illustration of the various loop types (highlighted in magenta) observed in G4 structures.

flanking nucleotides [43, 44], as well as environmental conditions, such as the type and concentration of ions [45] and molecular crowding. DNA G-quadruplexes are known to be polymorphic, while RNA G-quadruplexes are limited to a single dominant all-parallel topology [46].

Several methods have been developed to identify the presence of G4 structures in the human genome. A consensus sequence, initially proposed as d(G_3+N_{1-7})4 (where N represents any base), has been widely used for bioinformatic searches to identify potential G4-forming sequences across the genome [47]. However, it is important to note that this consensus sequence does not encompass all possible G4 structures, including those with two layers. Additionally, there are known G4 structures with loops longer than 7 bases [48], which are not captured by this consensus sequence. Experimental studies have revealed that relying solely on this consensus sequence can lead to a significant number of false positives [31]. Therefore, various computational algorithms, such as G4Hunter, have been developed to improve the accuracy of G4 prediction [49]. Biophysical methods, such as circular dichroism (CD) spectroscopy, can be used to determine G4 topology based on the positive or negative CD signal at specific wavelengths [50]. Biochemical techniques, such

as G4-specific antibody recognition or selective stabilization of G4 structures using ligands, can also be used for G4 detection [51]. Additionally, cell imaging methods, such as single-molecule fluorescence resonance energy transfer (smFRET) microscopy, have been employed to study G4 structures in live cells [52]. Moreover, G4-seq, a sequencing-based method, has been developed to identify G4-forming regions of the genome with high resolution [48]. Despite the availability of these methods, the prediction of G4-forming regions remains challenging, and the accuracy of each method varies depending on the specific experimental conditions and sample preparation methods.

G-quadruplexes can exist as multi-molecular structures, but in cells, they predominantly form as unimolecular/intramolecular structures, which are formed by a single guanine-rich strand [53]. These structures are enriched in telomeric and regulatory regions of the genome [54]. However, recent advancements in G-quadruplex identification techniques have revealed that there are around 700,000 possible G-quadruplexes present in the human genome [48]. A recent genome-wide detection study [55] showed that G-quadruplexes are present in more than 60% of gene promoters, particularly at the transcription start site, and in approximately 70% of genes. G-quadruplexes have been found at the promoters of some proto-oncogenes, ribosomal DNA, and immunoglobulin switch regions [56, 57], as well as in replication origins [58]. While their biological significance is not yet fully understood, guanine-rich RNA sequences, such as telomeric RNA repeats (TERRA) and the 3′- and 5′-untranslated regions of mRNA (UTRs), are also known to form stable G-quadruplexes [59, 60].

Since G4s are primarily found in critical genomic regions, there is mounting evidence that maintaining a delicate balance between folded and unfolded G4 motifs plays an important role in DNA replication and controlling of gene expression [61]. Indeed, it has been found that, on the one hand, G4s present intrinsic obstacles to DNA and RNA synthesis by promoting polymerase halts [62], but on the other, their formation might be involved in the activation of some DNA replication origins as well as in recruiting transcription factors to promoters [58].

2.2 Specificity in DNA-Protein Recognition

2.2.1 Sequence Specific DNA-protein Recognition

DNA is the blueprint of life, containing the genetic instructions that guide the development and function of all living organisms. The expression of these genes is precisely controlled by a complex interplay of regulatory proteins. Within this tightly regulated machinery of gene expression, sequence-specific DNA-protein recognition acts as the key to unlocking the secrets of how proteins orchestrate the symphony of life. As a consequence, sequence-specific DNA–protein interactions play a crucial role in a wide range of cellular functions and are essential for regulating virtually all DNA-templated processes [63–65]. Without these interactions, the integrity of the genome would be compromised, leading to severe consequences for the cell and the organism as a whole [66, 67]. Therefore, understanding the mechanisms of sequence-specific DNA–protein interactions is of great importance for unraveling the complex workings of the cell and developing new approaches for treating diseases.

Advances in high-throughput sequencing technologies have enabled genome-wide analysis of sequence-specific DNA-protein interactions. One widely used method is chromatin immunoprecipitation followed by sequencing (ChIP-seq), which allows the identification of protein-binding sites across the genome [68–70]. Using ChIP-seq, recent studies have revealed the complexity and diversity of protein-DNA interactions in different cellular contexts [68, 71]. For example, a ChIP-seq study of the transcription factor Oct4 in mouse embryonic stem cells identified thousands of binding sites, many of which are associated with key pluripotency genes [72, 73]. Therefore, for DNA-binding proteins, it can be a formidable task to distinguish subtle differences and accurately discriminate the target DNA sequence from the vast number of off-target sequences.

Numerous studies have elucidated that the binding of a target DNA substrate and its partner protein is a complex process that depends on several key features [74, 75]. One of the most crucial features is the electrostatic attraction between the negatively charged sugar-phosphate backbone of DNA and positively charged residues at the binding site of the protein [76, 77]. This electrostatic interaction is fundamental in providing general affinity between the two molecules, which is strong enough to bring them into close proximity and facilitate other specific interactions to occur [78, 79].

DNA bases possess unique patterns of functional groups, such as hydrogen bond (H-bond) donor amino groups from cytosine and adenine, and H-bond acceptor carbonyl groups from guanine and thymine in the major groove (see Fig. 2.7). Complementary to these patterns, specific proteins contain polar residues in their binding sites [80–82]. Thus, the specific nature of the binding interaction between protein and DNA is primarily determined by polar interactions such as hydrogen bonding between specific amino acids and DNA bases, as discussed in detail below. Additionally, van der Waals interactions, hydrophobic interactions, and shape complementarity also play important roles in determining binding specificity [83–87]. In addition to these interactions, entropic factors also contribute to the binding affinity between DNA and protein. For instance, the reduction of the degrees of freedom of both the protein and DNA upon binding results in decreasing the binding affinity to a varying extent [88, 89]. Conformational entropy can exhibit a wide range of effects, from favorable to unfavorable, and therefore has the potential, although infrequently, to modulate DNA-protein binding specificity. The intricate interplay between these various factors, both attractive and repulsive, ultimately determines the specificity and strength of protein-DNA interactions [90, 91]. Therefore, a comprehensive understanding of these factors is vital for elucidating the mechanisms underlying protein-DNA interactions.

Direct Readout

Many DNA-binding proteins have evolved binding surfaces that are predominantly composed of polar residues [79, 82, 92]. These polar residues are complementary to the unique pattern of functional groups on a specific base sequence (see Fig. 2.7). This mode of recognition is called direct (or base) readout, as it relies primarily on the formation of direct or water-mediated hydrogen bonds with DNA bases [80, 81]. The formation of these hydrogen bonds occurs between specific polar residues on the protein and the functional groups on the DNA bases. For example, the amino acid arginine has a positively charged (at physiological pH) guanidinium group that can form hydrogen bonds with the negatively charged carbonyl group

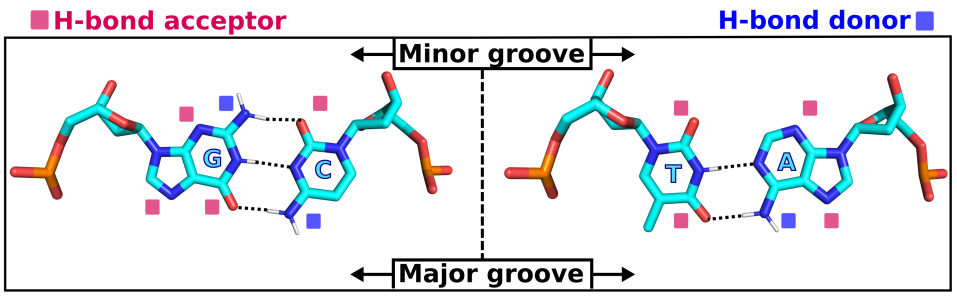

FIGURE 2.7: A schematic representation of possible hydrogen bond (H-bond) donors and acceptors at the surface of functional groups of GC or AT base pairs in the major and minor groove.

of guanine (G) or thymine (T) exposed in the major groove (see Fig. 2.7). Additionally, the positively charged (arginine and lysine) and other polar residues (e.g., asparagine and glutamine) can form hydrogen bonds with the sugar-phosphate backbone of DNA [93, 94].

There is ample evidence to suggest that the sequence context of DNA surrounding the binding site can modulate the specificity and strength of direct readout in DNA-protein interactions, highlighting the vital role of this mechanism in sequence-specific interactions [95, 96]. Moreover, recent research has revealed that the post-translational modifications, such as phosphorylation and acetylation, can regulate direct readout interactions and alter the binding specificity and affinity of proteins for their target DNA sequences [97].

The direct readout is typically carried out in the major groove of the DNA helix [77]. The major groove is wider than the minor groove and provides more space for protein residues to interact with the nucleobases. In fact, the functional group edges of base pairs are more accessible to protein residues in the major groove. This accessibility allows for more specific and accurate recognition of DNA sequences by proteins [98, 99]. Moreover, it has also been found that the direct readout mechanism involves the recognition of DNA by proteins through the utilization of shape complementarity between the protein's sequence-reading domain/motif and the major groove of the DNA. In this mechanism, a specific sequence cleft is formed by the methyl groups of thymines located on the surface of the major groove, which perfectly matches the shape of the recognition motif of the protein [100].

While direct readout is an important mechanism for DNA-binding proteins to recognize specific DNA sequences, it has become clear that it is not always sufficient to achieve the level of sequence specificity required for many biological processes [101]. Therefore, additional mechanisms are often necessary to ensure accurate recognition of specific DNA sequences [102].

Indirect Readout

The indirect readout is an alternative mechanism for protein-DNA recognition that does not rely on direct contact or H-bond between protein and DNA. Instead, the protein recognizes the overall shape and conformational flexibility of the DNA molecule, which is influenced by the specific sequence of nucleotides — thus, this phenomenon is also known as shape readout [103, 104]. The indirect readout plays

an important role in DNA-protein interactions, especially in eukaryotes, where the complexity of chromatin structure and the involvement of numerous proteins pose significant challenges for direct readout [105].

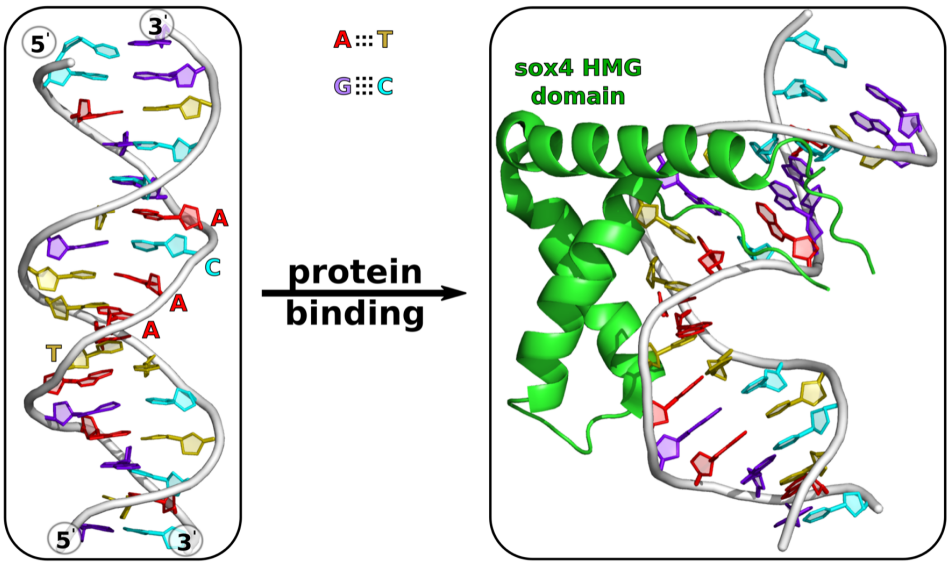

FIGURE 2.8: An example of DNA-protein interaction via indirect readout: binding of the HMG (high-mobility group) domain of Sox4 (Sry (sex-determining region on the Y chromosome)-related HMG box 4) protein facilitates the deformation in target DNA motif (5′-ACAAT-3′).

The indirect readout is a complex mechanism that exploits the intricate connections between various factors in DNA, such as the sequence-dependent conformation of the DNA helix, its ability to bend and twist (deformation), the amount of water molecules surrounding it (hydration), and the electrostatic charges on the surface of the molecule [106, 107]. These interdependent factors enable proteins to recognize and bind to specific DNA sequences without even directly contacting the DNA bases. Instead, the protein relies on either the unique shape of the DNA molecule or the energy required for the transition into the necessary conformation [108, 109]. More specifically, the flexibility of the DNA helix and the amount of energy required to bend it (such as local bending) have been found to vary depending on the specific base sequence [101, 110]. This suggests that the ability of many DNA-binding proteins to bind to their specific DNA sites may depend on how easily the DNA helix can be distorted into the necessary conformation, enabling the protein to form stabilizing interactions with the DNA (see Fig. 2.8) [111, 112].

The indirect readout is particularly important for proteins that bind to the minor groove of the DNA helix, which has a narrower width and shallower depth compared to the major groove [113]. Because the minor groove has less space than the major groove, direct contact with the nucleotide bases is limited [114]. However, the minor groove can provide more subtle, but informative, signals about the DNA sequence due to its ability to deform and twist in response to specific base sequences [115, 116]. Additionally, the properties of the surrounding water molecules and electrostatic charges on the surface of the DNA molecule also affect the shape of the minor groove [113, 117]. Therefore, proteins that bind to the minor groove rely

more heavily on indirect readout to recognize specific DNA sequences [118]. The 434 repressor, TATA-binding protein, and trp repressor were some of the earliest examples of proteins that were found to recognize specific DNA sequences indirectly [103, 119, 120]. Additionally, the transcription factor SOX4 is well-known for its indirect readout mechanism, in which its HMG domain induces a bend in the target DNA sequence (see Fig. 2.8) [121]. DNA-binding proteins that utilize the indirect readout mechanism for sequence recognition have been identified across various organisms, spanning the entire tree of life. Examples of such proteins include Met repressor, IHF/Hbb, c-myb, MarA, Papillomavirus E2 protein, estrogen receptors, CAP, HincII restriction endonuclease, P22R, and Ndt80, among others [100].

Although major groove-binding proteins primarily utilize the direct readout mechanism and minor groove binding proteins primarily use the indirect readout mechanism, in reality, these mechanisms often work in conjunction with each other. As a result, the majority of DNA-binding proteins use a combination of these mechanisms to locate their target DNA sequences [82, 122].

2.2.2 Sequence Non-specific DNA-Protein Recognition

Early studies of DNA-protein interactions focused primarily on the mechanism of sequence-specific DNA-protein recognition, where proteins recognize specific DNA sequences through direct and indirect readout mechanisms. However, studies in the 1970s and 1980s revealed that some DNA-binding proteins interact with DNA in a sequence non-specific manner [123, 124]. One of the earliest examples of such proteins were histones, which are known to package DNA into chromatin [125]. Histones package DNA into chromatin by forming nucleosome complexes, where DNA is wrapped around a histone octamer consisting of two copies of four histone proteins [125]. It was observed that histones interact with DNA without a well-defined sequence preference [126].

In the 1990s, the discovery of the Ku70/Ku80 heterodimer shed light on the importance of sequence non-specific DNA-protein recognition in DNA repair [127]. Ku70/Ku80 is a DNA repair protein that recognizes DNA non-specifically and plays an essential role in maintaining genome stability [128]. The Ku70/Ku80 is involved in the non-homologous end joining (NHEJ) pathway, which repairs DNA double-strand breaks. Ku70/Ku80 binds to the DNA ends through non-specific interactions with the DNA backbone and recruits other DNA repair proteins to the site of DNA damage [129]. These proteins recognize double-strand breaks (DSB) including blunt ends, 5′ and 3′ overhangs, and DNA hairpins in a variety of DNA sequences [130].

Moreover, studies on transcription factors revealed that some transcription factors, such as TATA-binding protein (TBP), interact with DNA non-specifically to initiate the transcription process [131]. TBP recognizes the TATA box, a short DNA sequence located upstream of the transcription start site, through a combination of sequence-specific and non-specific interactions with the DNA [132]. The non-specific interactions between TBP and DNA are believed to help stabilize the binding of TBP to the TATA box and facilitate the recruitment of other transcription factors and RNA polymerase to the site of transcription initiation [133].

Even though sequence non-specific DNA-protein recognition can occur through various mechanisms, it is believed that non-specific DNA-protein interactions are

driven primarily by electrostatic forces. The negatively charged sugar-phosphate backbone of DNA interacts with positively charged amino acid residues of proteins through electrostatic interactions, allowing proteins to bind to DNA in a non-specific manner [134, 135].

The strength of electrostatic interactions depends on the distance between the positively charged residues on the protein and the negatively charged phosphate groups on the DNA backbone, as well as the ionic strength of the surrounding environment [136, 137]. Thus, shape complementarity or shape recognition is also believed to play a crucial role in the mechanism of sequence non-specific DNA-protein interactions, by ensuring a proper fit between the binding partners [138]. For example, histones recognize the overall or global shape of canonical DNA double helix and bind to it non-specifically, allowing them to package DNA into chromatin [139].

In addition, studies have shown that non-specific DNA-protein interactions can also occur cooperatively, for instance, the *Bacillus subtilis* ParB protein displays apparent positive cooperativity, which is associated with the formation of larger, poorly defined nucleoprotein complexes [140]. This phenomenon is thought to be mediated by the formation of bridging interactions between multiple ParB dimers and non-specific DNA regions, leading to the formation of large protein-DNA complexes with undefined boundaries [140]. Such interactions have been observed in a variety of bacterial species and are thought to be essential for the proper segregation of chromosomes during cell division [141].

2.3 DNA-binding Proteins

2.3.1 Transcription Factor: A Key Regulator of Gene Expression

Transcription factors (TFs) are essential proteins that play a crucial role in regulating gene expression by binding to specific DNA sequences and either initiating or repressing the transcription process. The interaction between transcription factors and DNA is a complex process that involves multiple stages, including recognition, binding, and regulation of gene expression [142]. TFs recognize and bind to specific DNA sequences known as transcription factor binding sites (TFBSs) in the regulatory regions of target genes [143]. Once TFs have been recognized and bound to their respective TFBSs, they can either activate or repress the transcription of target genes. Transcriptional activation is achieved through the recruitment of co-activators, which enhance the activity of RNA polymerase II and facilitate the formation of the transcriptional pre-initiation complex [144]. Conversely, transcriptional repression is achieved through the recruitment of co-repressors, which inhibit the activity of RNA polymerase II and prevent the formation of the transcriptional pre-initiation complex [145].

The recognition of TFBSs by TFs is dependent on the specific sequence and structural features of the DNA-binding domain (DBD) of the transcription factor. The DBD of transcription factors is highly conserved and contains structural motifs, such as the basic helix-loop-helix, zinc-finger, leucine zipper, and helix-turn-helix, that facilitate specific interactions with DNA as discussed below [146, 147].

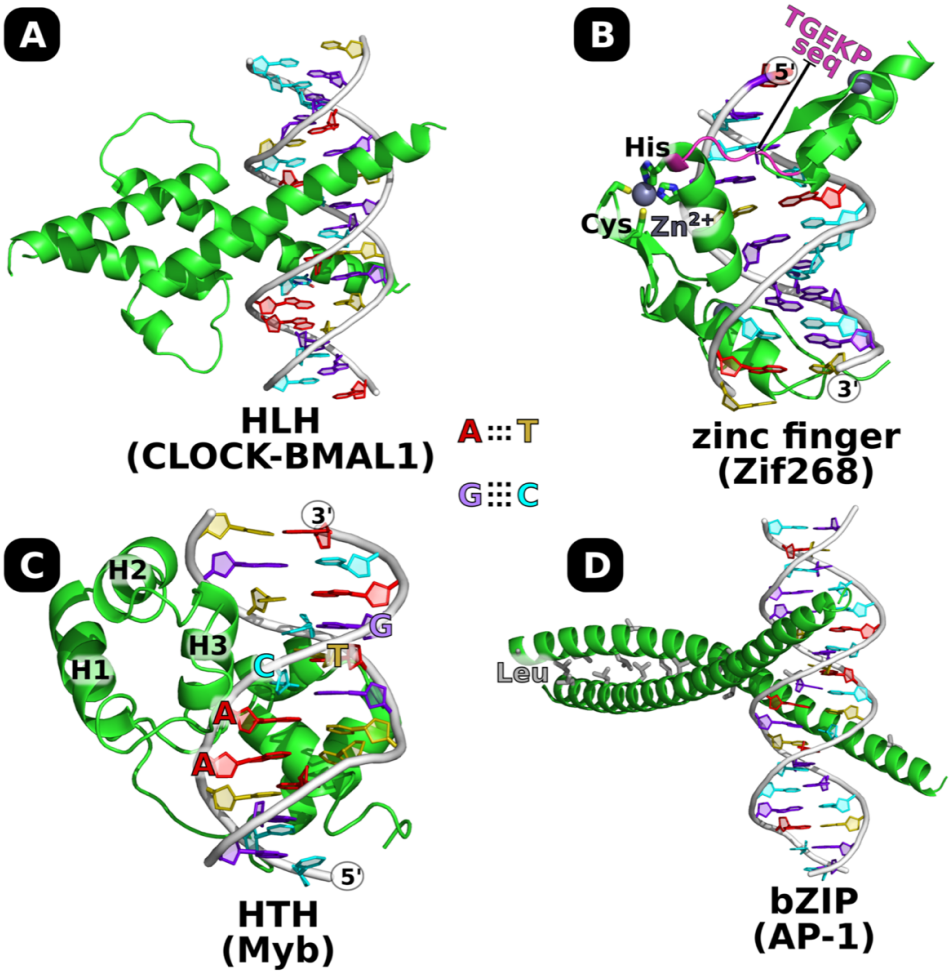

FIGURE 2.9: Structural representation of DNA-protein complexes highlighting different structural motifs of transcription factors involved in sequence-specific DNA-protein interactions. **(A)** CLOCK (Circadian locomotor output cycles kaput) and BMAL1 (brain and muscle ARNT-like 1) (PDB id: 4H10) [148] as an example of basic helix–loop–helix (bHLH) motif.**(B)** The zinc finger motif in Zif268 (PDB id: 1ZAA) [149] is displayed, with conserved Cys2His2 (coordinating the Zn ion) highlighted in stick representation, and the connecting loop containing the conserved sequence (TGEKP) in magenta. **(C)** The helix-turn-helix (HTH) motif in the tri-helical DNA-binding domain of Myb (PDB id: 1MSE) [150] is presented, bound to the consensus sequence AACTG. **(D)** The basic leucine zipper (bZIP) motif in the Activator protein-1 (AP-1) FosB/JunD domains (PDB id: 5VPE) [151] is shown, with leucine residues from this heterodimer that act as a "zipper" highlighted in grey sticks representation.

Helix-loop-helix

The basic helix-loop-helix (bHLH) motif is a common structural motif found in many eukaryotic transcription factors and is characterized by two alpha helices connected by a loop region [148, 152]. The first helix, known as the DNA-binding helix, contains basic amino acids that interact with the DNA groove region and facilitate specific binding to DNA. The second helix, known as the dimerization helix, interacts with another bHLH-containing protein to form a stable dimeric complex (see

Fig. 2.9A). The specificity of bHLH-containing transcription factors for DNA binding is largely determined by the sequence at the region of the basic amino acids containing-domain in the DNA-binding helix. These basic amino acids interact with the negatively charged phosphate groups in the DNA, as well as specific sequences, and determine the binding specificity of the protein [148]. For example, MyoD contains clusters of basic residues, which confers a preference for DNA sequences containing a specific E-box consensus sequence [153].

In fact, the specificity of bHLH-mediated DNA-protein interactions is determined by several factors, including the sequence and spacing of E-boxes within target gene regulatory regions and the combinatorial interactions between bHLH-containing transcription factors [152]. The spacing between E-boxes is critical for the optimal binding of bHLH-containing transcription factors, with optimal spacing between E-boxes ranging from 5 to 7 base pairs [154]. The combinatorial interactions between bHLH-containing transcription factors can result in the formation of homo- or heterodimers that bind to unique DNA sequences and regulate the expression of specific target genes [155]. On a similar note, the dimerization of bHLH-containing transcription factors is also critical for their function. The dimerization helix interacts with a similar helix in another bHLH-containing protein to form a stable dimeric complex that binds to DNA with high specificity [156].

Zinc Finger

The zinc finger structural motif is a highly conserved domain that is found in many transcription factors, and it has been recognized as a repeated zinc-binding motif since it was first identified in *Xenopus* transcription factor IIIA (TFIIIA) [157]. The zinc finger domain is characterized by the presence of conserved cysteine (Cys) and histidine (His) residues that coordinate a single zinc ion (see Fig. 2.9B). This motif consists of approximately 30 amino acids, and the repeated occurrence of this domain in many transcription factors allows for a diverse range of DNA-binding specificities. The folding of the zinc finger motif in the presence of zinc ions results in the formation of a compact $\beta\beta\alpha$ fold domain. The zinc ion is tetrahedrally coordinated between two cysteine residues at one end of the β-sheet and two histidine residues in the C-terminal portion of the α-helix (see Fig. 2.9B). The α-helical portion of each finger fits into the major groove of the DNA, and the binding of successive fingers causes the protein to wrap around the DNA. Each finger has a similar DNA-binding mode and contacts an overlapping four-base pair subsite, although the majority of base contacts occur in three base pair segments along one strand of the DNA (primary strand) [158].

One of the most well-studied zinc finger transcription factors, also considered in my research, is Zif268, which contains three zinc fingers [149]. The unique structure of zinc fingers allows for the recognition of specific DNA sequences, as each finger interacts with a specific set of nucleotides. The highly conserved linker of sequence TGEKP connects adjacent fingers (see Fig. 2.9B), which can enhance the specificity and stability of the protein-DNA interaction [149]. The DNA conformation of the zinc finger-DNA complex is generally similar to that of B-form DNA, but the major groove is wider and deeper than normal, resulting in a distinctive DNA conformation. This enlarged major groove is a common feature in the structures of most other zinc finger-DNA complexes and occurs in a number of other protein-DNA complexes [159].

The specificity of the zinc finger-DNA interaction is determined by both the DNA sequence and the arrangement of amino acid residues in the zinc finger motif [160]. The zinc finger motif is highly versatile, with the ability to recognize a wide range of DNA sequences, and has important implications for many biological processes, including gene regulation, DNA replication, and repair [158].

Helix-turn-helix

The helix-turn-helix (HTH) motif is another common structural motif found in transcription factors that enables sequence-specific DNA-protein interactions. One example of a transcription factor with an HTH motif is Myb, which plays an important role in regulating gene expression in a variety of organisms, including humans [161].

The HTH motif consists of multiple α-helices (tri- and tetra helical mostly) connected by a short turn; with the second or third helix acting as a recognition helix that binds to specific bases in the DNA major groove [150]. The first helix serves to stabilize the structure and provide additional contact with the DNA backbone. The specificity of the interaction between the transcription factor and the DNA is determined by the amino acid sequence of the recognition helix, which can vary among different HTH-containing proteins [162]. Myb contains two consecutive HTH motifs, each consisting of a pair of helices separated by a short loop (see Fig. 2.9C). The third helix (H3) of each motif acts as a recognition helix, binding to specific DNA sequences with high affinity and specificity. Myb is known to bind to the sequence AAC(G/T)G (see Fig. 2.9C), which is present in the regulatory regions of a number of target genes [163]. Also, structural studies have shown that the Myb protein recognizes its DNA target through a combination of base-specific contacts with the recognition helix and non-specific interactions with the DNA backbone [162].

In addition to Myb, a number of other transcription factors contain HTH motifs, including the bacterial repressor protein LacI and the eukaryotic factors Engrailed and Pit-1 [119]. The HTH motif is a versatile structural motif that has evolved to provide a mechanism for sequence-specific DNA-protein interactions in a wide range of organisms and biological contexts [161].

Leucine Zipper

The basic leucine zipper (bZIP) structural motif is a common feature among transcription factors involved in sequence-specific DNA-protein interactions. The bZIP domain consists of two structural features located on a contiguous α-helix [151]. The first feature is a basic region of approximately 16 amino acid residues, consisting of basic residues arginine and lysine, that contact the DNA [164]. The second feature is a heptad repeat of leucines or other bulky hydrophobic amino acids, creating an amphipathic helix and facilitating dimerization (homo- or heterodimers), this is commonly referred to as the ”zipper“ (see Fig. 2.9D).

The binding of bZIP transcription factors to DNA is highly sequence-specific, with the basic region of the bZIP domain recognizing and binding to a specific DNA sequence [151]. The specificity of the interaction is determined by the amino acid sequence of the basic region, which is highly conserved among bZIP transcription factors [164]. One well-studied example of bZIP is the activator protein-1 (AP-1) transcription factor, which is composed of a dimer of Fos and Jun family members,

including Fosb and JunD [165]. The DNA sequences that AP-1 binds to are known as AP-1 response elements (AREs), and they typically contain the consensus sequence 5′-TGANTCA-3′. The basic region of the AP-1 protein binds to the DNA in the major groove, specifically recognizing the GANTCA motif [166].

In addition to Fos and Jun family members, other bZIP transcription factors include C/EBP, ATF/CREB, and GCN4 [167]. These proteins have been shown to regulate the expression of genes involved in various biological processes, including metabolism, circadian rhythm, and stress response [164].

2.3.2 EXOG: A Specialized Mitochondrial Enzyme for RNA-DNA Chimeric Duplexes

Mitochondrial DNA (mtDNA) replication is a complex process that involves an interplay between transcription and replication. The content of different types of free deoxynucleotide triphosphates (dNTPs) in mitochondria is highly inhomogeneous, with dGTP being the most abundant [168]. Due to the high amount of guanine (G) available for mtDNA synthesis, the complementary strands of mtDNA are referred to as Heavy (H)-strand (abundant in G) and the Light (L)-strand [169]. Both of them are synthesized continuously with a single priming event, according to the strand displacement model [170].

The DNA polymerase γ (POLγ) is responsible for mtDNA replication and it requires an RNA primer to elongate each strand. At the light strand promoter (LSP), mitochondrial RNA polymerase (POLRMT) starts to transcribe DNA into RNA until it reaches the G-quadruplex structure at conserved sequence block 2 (CSB2; located $\sim$120 nucleotide downstream LSP) [171]. this RNA is then utilized by POLγ as a primer to synthesize the H-strand.

After initiation at the H-strand replication origin (OriH), DNA synthesis occurs in a unidirectional way without simultaneous DNA synthesis on the L-strand. Once H-strand synthesis reaches two-thirds of the genome, it encounters a small (30 base-pairs) non-coding DNA region containing the L-strand origin (OriL) [172]. After the replication machinery passes OriL, this origin becomes single-stranded and adopts a stem-loop conformation [173]. POLRMT specifically recognizes this loop and synthesizes a short ($\sim$25 nucleotide) RNA primer which can be elongated by POLγ to generate the L-strand in the opposite direction [174].

Before ligation can occur during DNA replication, it is essential to remove the short RNA primer on the L-strand and the longer one on the H-strand. Incomplete removal of the RNA primer can lead to defects in ligation since DNA ligase 3 (LIG3) has a preference for DNA over RNA [175]. The main role in the removal of RNA primer is played by Ribonuclease H1 (RNase H1) [176–178]. However, based on its mechanism, RNase H1 recognizes four consecutive ribonucleotides and cleaves in between the second and third, leaving two ribonucleotides behind and remained attached to the newly synthesized DNA [176–178]. Therefore, 5′-end processing is necessary to remove the remaining two ribonucleotides and ensure proper ligation.

Based on insights gained from the nucleus, researchers have proposed several candidates for the role of removing these two ribonucleotides. These candidates include flap structure-specific endonuclease 1 (FEN1), DNA replication helicase/nuclease 2 (DNA2), and mitochondrial genome maintenance exonuclease 1 (MGME1). However, experiments confirmed that, unlike in the nucleus, these proteins have limited

activity in mitochondria, since the removal of these proteins have shown little to no impact on mitochondrial genome integrity [179]. Furthermore, in-vitro studies have demonstrated that the addition of these proteins in the presence of POLγ, LIG3, and RNase H1 did not improve ligation [180, 181].

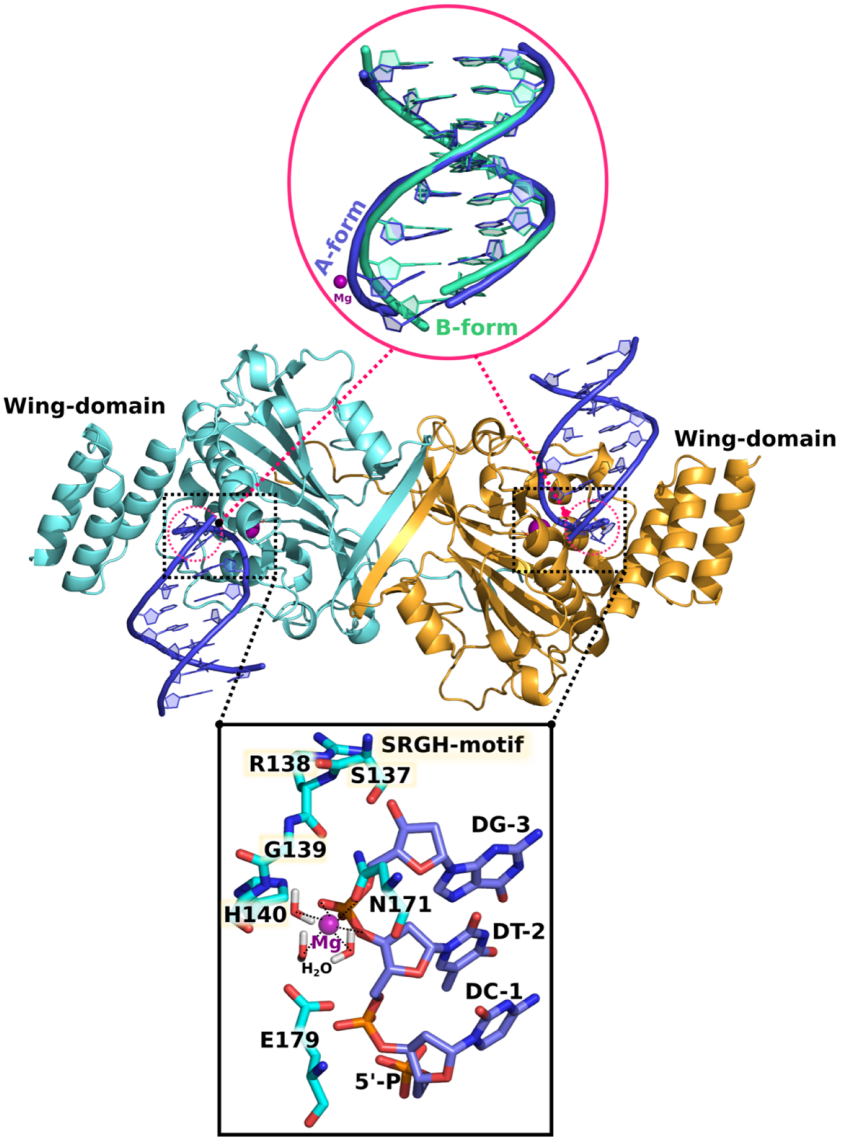

FIGURE 2.10: Crystal structure of the homodimeric human mitochondrial EXOG in complex with DNA-duplex [17]. (*lower inset*) The catalytic SRGH motif is shown in the stick representation, with an Mg^{2+} ion depicted as a purple sphere bound by three crystallographic water molecules. Additional residues at the catalytic site are also labeled for clarity. The upper inset highlights the ExoG-induced DNA transition from the canonical B-form to the A-form, specifically for the first 2–3 nucleotides from the 5′-end.

In 2008, the discovery of Exonuclease G (EXOG) [182], a paralog of endonuclease G (EndoG), finally provided insight into mitochondrial 5′-end processing. EXOG is specifically found in mitochondria and exhibits flap-independent 5′-3′ exonuclease activity. The recent crystal structure of EXOG confirms its flap-independent activity

and its unique ability to accommodate two 5′-end nucleotides in its deep substrate-binding groove (see Fig. 2.10), enabling it to precisely cleave dinucleotides (and not mono-nucleotides) from the 5′-blunt-end of double-stranded DNA (dsDNA) [17]. Interestingly, this structure also revealed a conformational transition of the bound DNA duplex from the natural B-form to the A-form (see Fig. 2.10, *upper inset*), but only in the portion present in the substrate-binding groove. This suggests that EXOG prefers RNA-DNA chimeric duplexes, specifically where only the first (5′) two nucleotides are from RNA, further confirming its role in 5′-end processing during mitochondrial replication.

A subsequent study by Wu et al. [183] provided additional evidence, as crystal structures of EXOG bound to different substrates demonstrated that EXOG shows higher affinity for and produced more product from RNA-DNA chimeric duplex compared to DNA-DNA or RNA-DNA duplexes. These findings support its role in removing the residual RNA primer left by RNase H1 during mtDNA replication.

Nevertheless, it is still unclear what mechanism underlies EXOG's high specificity for the RNA-DNA chimeric duplex, which adopts a hybrid A/B conformation, and the B-to-A conformational transition that occurs when it binds to a canonical DNA duplex has not been explored. Since its paralog, EndoG, does not contain the wing domain and is DNA-conformation nonspecific [182], it is not known whether only the wing domain is responsible for this specificity and conformational transition, or whether residues from the substrate-binding groove of the core domain also participate in this process and cooperatively facilitate the transition.

2.3.3 DHX36-Helicase: A Specialized G-Quadruplex Resolvase

The temporal control of the folding and unfolding of G-quadruplexes (G4s) is critical for cell cycle progression since G4s serve as key regulatory elements (see section 2.1.2) [184]. G4s possess high thermodynamic stability and slow unfolding rates, making it necessary for specialized helicases to evolve in cells to disrupt DNA and RNA G4s in an ATP-dependent manner [185, 186]. Several helicases, including RecQ family helicases like human BLM [187], XPD family enzymes like FANCJ [188], Pif1 family [189], and RHAU or DEAH-box family helicases like DHX36 [190], are known for their DNA unwinding abilities. Among them, DHX36 is particularly noteworthy in G4-related research because of its strong preference for G4 over DNA duplexes [185]. As a result, DHX36 plays a role in transcriptional regulation [191] and is involved in various essential cellular processes such as hematopoiesis [192] and cancer biogenesis, making it a potential target for therapeutic development [193].

Significant effort has been devoted to determining the structures of DHX36 in a bound state with G4 by several research teams. Previously, an NMR solution structure of a small portion (18 amino acids long) of the DHX36-specific motif (DSM) α1, which is part of the N-terminal region, was solved and shown to fold into an α-helix and have a high affinity for the 5′-end of G4 [195]. Recently, the X-ray crystal structures of bovine DHX36 co-crystallized with parallel G4s were reported [194], as well as the crystal structures of *Drosophila* (fly) and mouse DHX36 [196, 197]. The structural core region, which is highly conserved among all DEAH helicases [190], primarily consists of two RecA domains (RecA1 and RecA2) that are responsible for ATP binding and hydrolysis, as well as the oligonucleotide and oligosaccharide-binding fold-like (OB), degenerate winged helix (WH), and ratchet-like (RL) subdomains, collectively forming the C-terminal domain (see Fig. 2.11). The N-terminal extension is comprised of a glycine-rich loop, followed by the most important motif

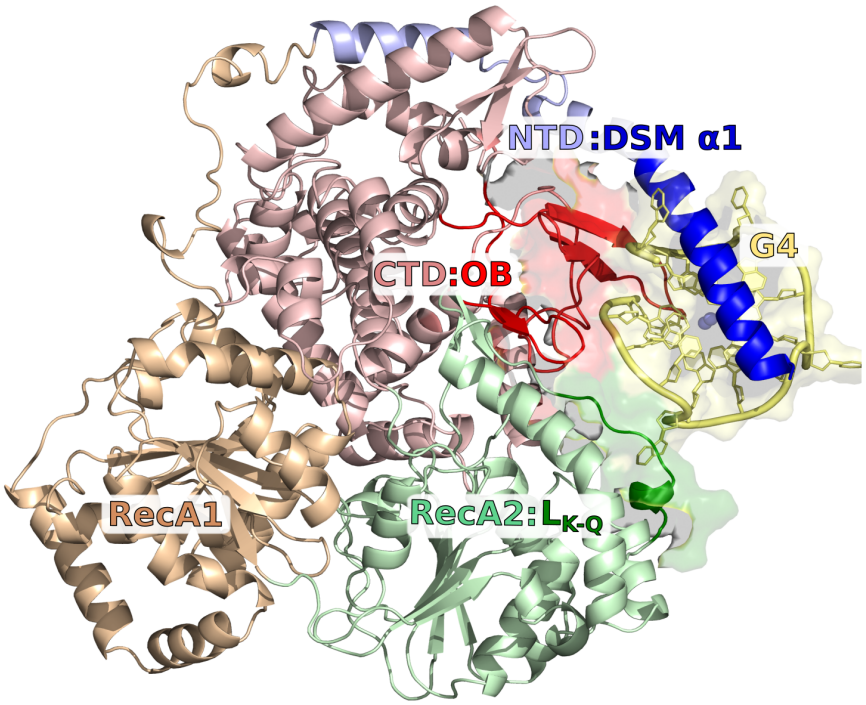

FIGURE 2.11: Structural representation of the DHX36 helicase in complex with its preferred parallel-stranded G-quadruplex (DNA^{myc}-G4), taken from the X-ray crystal structure [194]. Different domains of DHX36 are marked with various colors, where NTD and CTD refer to the N- and C-terminal domains, respectively. The darker shades indicate the binding interfaces with the G-quadruplex, and a new interface loop identified in my work is labeled as $L_{K\text{-}Q}$, which was missing in the original crystal structure.

for G4 recognition, i.e., DSM α1 (see Fig. 2.11). The 3′-single-stranded DNA portion fits into a narrow cleft between the RecA and C-terminal domains, which is enriched with basic amino acid residues [194].

Although the exact role of ATP hydrolysis in DHX36-driven remodeling of G4 is not very clear and remains disputed, Chen et al. [196] observed that the G4 bound to their solved structure (Fig. 2.11) was partially unfolded. They proposed, in agreement with previous biochemical studies [198], that passive destabilization of the G4 occurs upon its binding to DHX36, followed by active unfolding during the major conformational change of the helicase driven by ATP hydrolysis. Alternatively, it has been suggested that DHX36 undergoes ATP-independent conformational transition that exerts force on the 3′-tail of the bound parallel G4, leading to its unwinding one nucleotide at a time [194]. After ATP hydrolysis, the partially destabilized G-quadruplex is then released from the helicase [199].

Initially, it was believed that DSM α1 might be the only crucial unit for the recognition process, but the full-length DHX36 helicase is necessary to achieve a much

higher affinity for G4 than is observed for the isolated DSM α1 peptide. The dissociation constant K_d for binding of an isolated DSM α1 peptide is 310 nM, whereas K_d for the whole helicase is <10 pM [194, 195]. This raises a question about the significance of other domains, especially the OB domain, which is located in close proximity to the binding site (see Fig. 2.11), in G4 recognition. The length of DSM α1 peptide has also proven to be important because in the previously reported NMR structure [195], where the DSM α1 peptide is comparatively four residues shorter, the subsequent biochemical assay showed that it has a much lower affinity ($\sim 1\ \mu$M) towards G4.

The crystal structure presented in Figure 2.11 [194] provides important insights into the recognition process of parallel-stranded G4s by DHX36, corroborating previous biochemical findings [195, 197]. Specifically, the DSM α1 helix binds to the solvent-exposed flat surface formed by the G-tetrad at the 5′-end of the G4, in agreement with earlier NMR studies on truncated DSM-G4 complexes [195].

However, the exact mechanism by which DHX36 recognizes parallel-stranded G4s is not fully understood. It is unclear whether the DSM binding mode and therefore affinity depend on the G4 state, and whether DSM acts solely as an anchor for G4s or actively participates in the unfolding process. The strong preference of DSM to bind to the 5′-terminal side of parallel G4s also requires clarification at the molecular level. Additionally, the relative importance of DSM and OB in G4 recognition and whether G4 binding to both these subdomains is independent are still unknown. Finally, the role of individual protein residues in the recognition process and their relative contribution to the binding energy remains to be established.

Chapter 3

Theory and Methodology

3.1 Overview of Molecular Mechanics in Biomolecular Simulations

3.1.1 Classical Representation of Molecules

To better explain molecular mechanics, it is important to first understand classical mechanics as it provides the framework for understanding the behavior of molecules and their interactions with each other. In classical mechanics, a molecule is represented as a collection of atoms, each described by its position, velocity, and mass. The position of an atom is defined by its Cartesian coordinates, which specify its location in space. The velocity of an atom is defined as the rate of change of its position with respect to time, while its mass is a fundamental property that determines its response to external forces.

To represent the behavior of a molecule in classical mechanics, we need to describe the interactions between atoms. The most common method to do this is through a mathematical function known as a potential energy function. The potential energy function is a function of the positions of all atoms in the molecule and provides a measure of the total potential energy of the system. The potential energy function is typically composed of different terms that describe the different types of interactions between atoms, such as covalent bonds, electrostatic interactions, and van der Waals forces (see below for detail).

Molecular mechanics is often employed to predict macroscopic observables based on statistical mechanics (described in Section 3.2), which involves following the system in time at the molecular level. While this can be done more accurately using quantum mechanics, it is computationally expensive for larger systems, even when treating nuclei classically within the Born-Oppenheimer approximation. Therefore, a simplified description that interpolates the Born-Oppenheimer surface is needed, which is where the force field comes into play. Force fields (described below in detail) are often parametrized using quantum chemistry to provide a reasonable approximation of the potential energy surface of a system, allowing for efficient simulations of molecular motion and interaction.

Bond Terms

In molecular mechanics, a bond between two atoms in a molecule is typically modeled as a spring-like connection, where the potential energy stored in the bond is a function of the distance between the atoms. This approach allows for the simulation of molecular systems by considering the interatomic forces and energy contributions.

Traditionally, the bond term in molecular mechanics has been described using a parabolic potential energy curve, based on Hooke's law and the concept of a harmonic oscillator. According to Hooke's law, the force required to extend or compress a bond is proportional to the displacement from the equilibrium position. The proportionality constant is known as the spring constant, denoted as k_b. Mathematically, Hooke's law for a bond can be expressed as:

$$F = -k_b(r - r_e) \tag{3.1}$$

where F represents the restoring force, r denotes the current length of the bond, and r_e corresponds to the equilibrium bond length. The negative sign signifies that the force acts in the opposite direction of the displacement, always tending to restore the bond to its equilibrium length.

The harmonic bond potential energy, derived from Hooke's law, can be obtained by integrating the force equation:

$$V(r) = \frac{1}{2}k_b(r - r_e)^2 \tag{3.2}$$

This potential energy curve describes the bond as a simple harmonic oscillator, assuming that the bond stretching behavior is purely harmonic and neglecting any anharmonic effects.

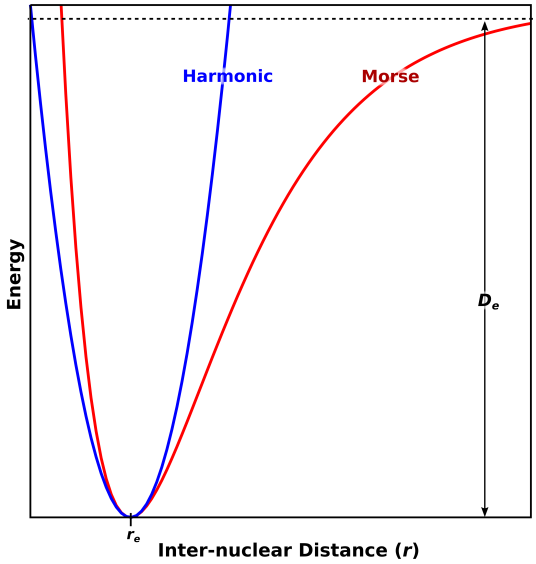

FIGURE 3.1: Comparison of the Morse potential (red) and harmonic oscillator potential (blue). The Morse potential considers the effects of bond breaking and anharmonicity, resulting in an asymmetric curve. The harmonic potential assumes a simple parabolic curve and is a good approximation when the molecule is not highly excited or stretched.

While the harmonic bond potential provides a convenient and computationally efficient model, it has certain limitations when it comes to accurately describing bond breaking and the dissociation of atoms. These limitations arise from the assumption of pure harmonic behavior and neglect of anharmonicity effects (see Fig. 3.1).

In reality, chemical bonds can exhibit anharmonic behavior, especially under large displacements from the equilibrium position. Anharmonicity refers to the deviation from the simple harmonic motion assumption, where the restoring force is no longer strictly proportional to the displacement. Anharmonic effects become significant as the bond is stretched or compressed beyond small deviations from the equilibrium bond length.

To address the limitations of the harmonic bond potential, more realistic models have been developed. One commonly used model is the Morse potential function, which provides a better approximation for the potential energy of a bond by considering anharmonicity effects. The Morse potential function is given by:

$$V(r) = D_e \left[1 - e^{-\alpha(r-r_e)}\right]^2 \tag{3.3}$$

where $V(r)$ is the potential energy as a function of the bond length r, D_e is the dissociation energy of the bond, α is a parameter related to the force constant of the bond, and r_e is the equilibrium bond length.

The Morse potential captures the anharmonicity of bond stretching behavior, allowing for a more accurate representation of bond breaking and dissociation processes. The potential energy curve generated by the Morse potential function exhibits a characteristic "well" shape with a minimum energy at the equilibrium bond length r_e (see Fig. 3.1).

Angle

The angle between three atoms is an important parameter for characterizing the vibrational behavior of a molecular structure, similar to the bond length. The angle, denoted by θ, can be calculated using the dot product of the position vectors of the atoms (as described below). The position vector of an atom is defined as the vector connecting the origin of the coordinate system to the atom's position in space.

Consider three atoms labeled as i, j, and k, with position vectors $\vec{r}_i$, $\vec{r}_j$, and $\vec{r}_k$, respectively. The angle between atoms i, j, and k is given by the following equation:

$$\theta_{ijk} = \cos^{-1}\left(\frac{\vec{r}_{ij} \cdot \vec{r}_{kj}}{|\vec{r}_{ij}||\vec{r}_{kj}|}\right) \tag{3.4}$$

where $\vec{r_{ij}} = \vec{r}_j - \vec{r}_i$ and $\vec{r_{kj}} = \vec{r}_j - \vec{r}_k$ are the position vectors connecting atom i to atom j and atom k to atom j, respectively.

The potential energy associated with the angle between three atoms can be described by a harmonic potential, which is commonly used to represent bond angles in molecular mechanics simulations. The harmonic potential energy for an angle deviation from its equilibrium value, θ_e, is given by:

$$V(\theta) = \frac{1}{2} k_\theta (\theta - \theta_e)^2 \tag{3.5}$$

where k_θ is the force constant, which determines the strength of the bond angle potential. The potential energy surface for a harmonic angle potential is parabolic, with a minimum at the equilibrium angle, θ_e.

However, this simple harmonic potential does not always accurately capture the vibrational behavior of the molecular structure. To improve the representation of bond angles, a correction term called the Urey-Bradley (UB) potential was introduced and added to the harmonic potential on the angle (Eq. 3.5). The Urey-Bradley

potential describes the interaction between atoms i and k that are bonded to a common atom j and is given by the following equation:

$$V_{\mathrm{UB}}(r_{ik}) = \frac{k_{\mathrm{UB}}}{2}(r_{ik} - r_{\mathrm{UB}})^2 \tag{3.6}$$

where r_{ik} is the distance between i and k, k_{UB} is Urey-Bradley force constant, and r_{UB} is the equilibrium distance between atoms i and k when they are bonded to atom j.

Dihedral

In molecular mechanics, another important type of bonded interaction is the dihedral potential, which describes the energy associated with the rotation around a bond. The rotation is measured by the dihedral angle, which represents the angle between two planes defined by specific atoms in the molecule. There are two main types of dihedrals: proper dihedrals and improper dihedrals.

The proper dihedral angle, denoted as ϕ, is defined as the angle between two planes formed by four consecutive atoms, labeled as i, j, k, and l, in the molecule. It quantifies the spatial orientation of these atoms. Proper dihedrals are typically represented using a periodic potential, which accounts for the periodicity of the potential energy surface as the dihedral angle rotates through a full 360° (2π). This periodicity arises from the fact that the dihedral potential energy is invariant under a rotation by 360°. The potential energy associated with a proper dihedral angle can be described by a periodic function, commonly expressed as a Fourier series:

$$V(\phi) = \sum_{n=1}^{N} V_n \left[1 + \cos(n\phi - \gamma)\right] \tag{3.7}$$

where V_n represents the amplitude of the nth term, and γ represents the phase shift. The number of terms in the Fourier series, denoted as N, determines the level of detail in the description of the dihedral potential.

While proper dihedrals adequately describe the torsional behavior of most molecules, there are cases where additional terms are required to accurately represent the molecular system. Improper dihedrals, also known as out-of-plane or planarity terms, are introduced to maintain the planarity of specific groups within a molecule or to prevent molecules from inverting to their mirror images. Improper dihedrals are typically employed in molecules that contain planar groups, such as aromatic rings or conjugated systems.

The energetics of improper dihedrals are typically described using a harmonic potential, where the deviation from the equilibrium value, ψ_e, is penalized. The potential energy for an improper dihedral can be expressed as:

$$V(\psi) = \frac{1}{2}k_\psi(\psi - \psi_e)^2 \tag{3.8}$$

where k_ψ is the force constant that determines the strength of the improper dihedral potential, and ψ is the dihedral angle. The potential energy for an improper dihedral is a parabolic function with a minimum at the equilibrium dihedral angle, ψ_e.

In biochemistry convention [200], dihedrals are often used to describe the conformation of biopolymers such as proteins and nucleic acids. The dihedral angles

between the atoms in these molecules can greatly influence their structure and function.

Van der Waals Interactions

The Lennard-Jones potential is a commonly used non-bonded interaction term in molecular mechanics. It describes the energy between two atoms as a function of their separation distance and is often used to model van der Waals forces. The Lennard-Jones potential can be written as:

$$V_{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{3.9}$$

where r is the distance between the two atoms, ε is the depth of the potential well, and σ is the distance at which the potential is zero. The term $(\sigma/r)^{12}$ represents the repulsive term of the potential, which arises due to the overlap of the closed-shell electron clouds of the two atoms, while the $(\sigma/r)^6$ term represents the attractive term, which arises due to dispersion forces between the two atoms.

However, the Lennard-Jones potential has some limitations, particularly in accurately describing the repulsion at short distances. To address this issue, the Buckingham potential was introduced as an alternative to the Lennard-Jones potential. The Buckingham potential uses an exponential repulsion term that provides a more realistic description of the interaction at short distances. The Buckingham potential is defined as follows:

$$V_{bh}(r) = Ae^{-Br} - \frac{C}{r^6} \tag{3.10}$$

where A, B, and C are parameters (depending on the atom types) that determine the strength of the attractive and repulsive forces. The first term represents the repulsive force, which decreases exponentially with distance, while the second term represents the attractive force, which decreases with distance as r^{-6}. The Buckingham potential offers a more accurate representation of repulsion compared to the Lennard-Jones potential. However, it comes with increased computational complexity due to the presence of an exponential term.

Coulombic Term

Coulombic interactions, also known as electrostatic interactions, are another important type of non-bonded interaction in classical molecular representations. These interactions arise due to the presence of charged entities in the system, such as ions or molecules with partial charges.

The Coulombic interaction energy between two charged entities, denoted as i and j, follows Coulomb's law, which states that the force between two point charges is directly proportional to the product of their charges and inversely proportional to the square of the distance separating them. Coulomb's law applies to both point charges and the interaction between induced or permanent dipole moments in polar molecules.

In classical force fields, the Coulombic interaction energy between two charged entities can be described by the equation:

$$V_{Coul}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{3.11}$$

where q_i and q_j represent the charges of entities i and j, respectively. The term r_{ij} represents the distance between the charged entities, and ϵ_0 is the vacuum permittivity.

To expedite the computation of Coulombic interactions, a common practice involves implementing a cutoff distance (r_c), beyond which the interactions are truncated. Employing a cutoff distance enables faster calculations, especially since the computational cost of evaluating interactions between all particle pairs scales as $O(N^2)$, where N represents the number of particles in the system. Nonetheless, employing a cutoff distance may introduce simulation artifacts, necessitating the use of corrective measures like long-range correction schemes or alternative methods such as particle mesh Ewald (PME) or smoothed particle mesh Ewald (SPME).

3.1.2 Algorithmic Implementations

Force Field Models

After understanding the individual terms such as bonds, angles, Coulombic interactions, etc. (as described in above section 3.1.1), it is essential to recognize the pivotal role of a force field as an algorithmic implementation that brings these components together. A force field encompasses a set of mathematical functions and parameters that dictate how these interactions are calculated and integrated into a cohesive model. A force field is derived from a combination of experimental data and quantum mechanical calculations, aiming to reproduce various experimental observables and properties of molecules. It incorporates empirical potential energy functions, molecular mechanics, and other approximations to describe the behavior of atoms and molecules in a computationally efficient manner. By capturing the essence of molecular interactions, a force field enables the simulation of complex molecular systems and their evolution over time.

The force field provides a comprehensive framework for evaluating the potential energy and forces associated with bonded and non-bonded interactions. Bonded interactions encompass covalent bonds, bond angles, and torsional angles (dihedrals), which govern the structural properties and flexibility of molecules. Non-bonded interactions include van der Waals interactions and electrostatic interactions, which account for the attractive and repulsive forces between atoms or charged entities.

The total potential energy of a molecular system can be expressed using a general force-field equation. This equation combines the contributions from various types of interactions, and it is given by:

$$\begin{aligned} E &= E_{bonded} + E_{non-bonded} \\ &= \sum_{bonds} K_b(r - r_e)^2 + \sum_{angles} K_\theta(\theta - \theta_e)^2 \\ &+ \sum_{dihedrals} K_\phi[1 + \cos(n\phi - \gamma)] + \sum_{non-bonded} (E_{LJ} + E_{Coulomb}) \end{aligned} \tag{3.12}$$

The form of potential energy presented in the above equation is employed in widely used force fields such as standard versions of OPLS [201], GROMOS [202], AMBER [203] (which is used in this thesis), and CHARMM [204]. These general functional forms can be then modified or overridden with correction terms. For instance, some popular force fields such as AMBER scale the charge-charge and Lennard-Jones interactions between the 1-4 atom pairs (i.e., atoms i and l in an $ijkl$

sequence connected by three consecutive bonds) by a specific factor (here 0.8333 and 0.5, respectively) to account for their spatial proximity [205].

There is ongoing debate among scientists about whether force fields should be based on empirical models with corrections, or whether they should be based on more unified, physics-based models. While empirical models with corrections have led to significant improvements in the accuracy of molecular simulations, some argue that this approach does not address the underlying shortcomings. However, recent developments in polarizable force fields and constant-pH molecular dynamics show promise in addressing these shortcomings and delivering more accurate simulations.

Periodic Boundary Conditions

Periodic boundary conditions (PBCs) are commonly used in molecular simulations that allow the study of systems in the thermodynamic limit, mimicking an infinite system and avoiding the boundary effects caused by finite-size simulation boxes. Consider, for example, a molecule solvated in water. If we add a limited number of water molecules, there will be an effect of surface tension at the edge with air/vacuum. However, with PBCs, the simulation box is replicated periodically in all three dimensions, creating an infinite lattice of identical boxes. When an atom or molecule leaves the simulation box, it re-enters from the opposite side, effectively maintaining the continuity of the system. This means that an atom that drifts away from one side of the box appears on the other side as its periodic image. In practice, only a single image of each atom needs to be tracked, while the lattice vectors define the periodicity of the system. Mathematically, the potential energy of a system under PBCs can be expressed as:

$$V(\vec{r_1}, \vec{r_2}, ..., \vec{r_N}) = \frac{1}{2} \sum_{i \neq j}^{N} \sum_{n_x, n_y, n_z} v\left(\vec{r_{ij}} + n_x \vec{L_x} + n_y \vec{L_y} + n_z \vec{L_z}\right) \tag{3.13}$$

where v is the pairwise potential energy function between two particles i and j, $\vec{r_{ij}}$ is the vector separation between them, and $\vec{L_x}$, $\vec{L_y}$, and $\vec{L_z}$ are the length vectors defining the simulation box in the x, y, and z directions, respectively. The sum over n_x, n_y, and n_z is taken over all periodic images of the system. PBCs are implemented by adding a minimum image convention to the calculation of intermolecular distances and forces. The minimum image convention takes into account the shortest distance between two atoms or molecules, considering the periodic images of the simulation box. This ensures that the interaction energy between two molecules in adjacent boxes is correctly calculated.

However, the use of PBCs requires caution, as the periodic images can cause artifacts in the simulation results, such as the appearance of spurious correlations due to the periodicity of the system. The most common artifact is the so-called "image charge" effect, which arises from the interaction between an atom or molecule and its periodic images. This artifact can be minimized by using larger simulation boxes to reduce the frequency of image interactions. It is worth noting that the overall impact of periodic image artifacts is generally small compared to the truncation of electrostatic interactions at a certain cutoff distance.

PBCs are widely used in molecular dynamics simulations of liquids, solutions, and crystals. They enable the simulation of larger systems than would be feasible with a finite-size box, and they provide a realistic representation of the bulk properties of the system.

Energy Minimization

Energy minimization is an essential step in molecular simulations, as it allows the system to reach a stable state by minimizing its potential energy. This is typically done before running any simulations to ensure that the starting structure is energetically reasonable, with corrected bond length and free from clashes. The most common method used for energy minimization is the steepest descent algorithm, which iteratively updates the atomic coordinates in the direction of the negative gradient of the potential energy until a minimum is reached. The steepest descent algorithm can be described by the following equation:

$$\vec{r}_i^{n+1} = \vec{r}_i^{n} - \alpha \frac{\partial V}{\partial \vec{r}_i^{n}} \tag{3.14}$$

where $\vec{r}_i^{n}$ is the position of atom i at iteration n, α is the step size, and $\frac{\partial V}{\partial \vec{r}i^{n}}$ is the gradient of the potential energy with respect to the position of atom i at iteration n. The step size is chosen to be small enough to ensure convergence while avoiding overshooting the minimum.

Other algorithms, such as conjugate gradient and quasi-Newton methods, can also be used for energy minimization. These algorithms typically converge faster than the steepest descent algorithm, but they require more computational resources.

Energy minimization can be performed with different termination criteria, such as a maximum number of iterations, a minimum change in energy, or a minimum change in atomic positions. It is important to note that energy minimization only finds a local minimum and not necessarily the global minimum. Therefore, it is recommended to perform multiple energy minimization runs with different initial coordinates to explore the potential energy surface and identify the global minimum.

In addition, energy minimization can be affected by the choice of force field and initial coordinates. Some force fields may have parameters that are not well-suited for certain types of systems, leading to unrealistic structures. Therefore, it is important to choose a force field that has been validated for the type of system being studied and to perform a thorough validation of the initial structure before proceeding with simulations.

Integration of the Equations of Motion in Molecular Dynamics

Molecular dynamics (MD) simulations provide a means to connect the microscopic world of molecules and atoms to macroscopic observables through statistical mechanics (see section 3.2). By following the time evolution of a system, one can calculate various thermodynamic properties and gain insights into the behavior of the system at different conditions. When dealing with larger systems, it becomes computationally expensive to treat the nuclei quantum mechanically. Thus, classical mechanics is used to describe the motion of atoms, and force fields (as described above) are used to approximate the potential energy surface of the system.

The equations of motion describe how the positions and velocities of atoms in a system change over time in response to the forces acting on them. In molecular dynamics, the motion is described by Newton's second law, which states that the force acting on an object is equal to its mass times its acceleration:

$$\vec{F} = m\vec{a} \tag{3.15}$$

where $\vec{F}$ is the force acting on an atom, m is its mass, and $\vec{a}$ is its acceleration

To numerically integrate the equations of motion, we need to discretize time into small time steps Δt. The positions and velocities of the atoms are then updated at each time step using algorithms such as the Verlet algorithm or the leapfrog algorithm.

The Verlet algorithm is a widely used method for integrating equations of motion in molecular dynamics simulations. The algorithm updates the positions and velocities of atoms at each time step based on the forces acting on them. The positions are updated using the following equation:

$$\vec{r}_i(t + \Delta t) = 2\vec{r}_i(t) - \vec{r}_i(t - \Delta t) + \frac{\vec{F}_i(t)}{m_i}\Delta t^2 \tag{3.16}$$

where $\vec{r}_i(t)$ and $\vec{r}_i(t + \Delta t)$ are the positions of atom i at time t and $t + \Delta t$, respectively. $\vec{F}_i(t)$ is the force acting on atom i at time t, and m_i is its mass. The velocities of the atoms can be updated using the following equation:

$$\vec{v}_i(t + \Delta t/2) = \vec{v}_i(t - \Delta t/2) + \frac{\vec{F}_i(t)}{2mi}\Delta t \tag{3.17}$$

where $\vec{v}_i(t)$ and $\vec{v}_i(t + \Delta t/2)$ are the velocities of atom i at time t and $t + \Delta t/2$, respectively.

The leapfrog algorithm is another commonly used method for integrating equations of motion in molecular dynamics simulations. This algorithm updates the positions and velocities of atoms at half-integer time steps. The positions are updated using the following equation:

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i(t + \Delta t/2)\Delta t \tag{3.18}$$

where $\vec{v}_i(t + \Delta t/2)$ is the velocity of atom i at time $t + \Delta t/2$. The velocities can be updated using the following equation:

$$\vec{v}_i(t + \Delta t/2) = \vec{v}_i(t - \Delta t/2) + \frac{\vec{F}_i(t)}{m_i}\Delta t \tag{3.19}$$

where $\vec{v}_i(t - \Delta t/2)$ is the velocity of atom i at time $t - \Delta t/2$.

The numerical integration of the equations of motion in molecular dynamics simulations is subject to a degree of inaccuracy, which can accumulate over time and lead to errors in the results. However, the integration itself is subject to very little variation between individual codes, as the equations and algorithms used are well-established and widely accepted in the field.

To enhance integration performance, one approach is to implement a multiple time-step algorithm. This method involves using different time-step sizes for integrating the equations of motion and other computations, such as the calculation of long-range forces. For instance, the r-RESPA algorithm [206] utilizes multiple time steps to separate the integration of bonded and nonbonded interactions. By doing so, simulation efficiency can be improved while maintaining accuracy.

Constraint

Constraints are commonly employed in molecular simulations to enforce specific geometrical relationships within the system. These constraints play a crucial role in improving the accuracy and stability of simulations by preventing unphysical behavior.

One common type of constraint utilized in molecular simulations is the bond length constraint. This constraint restricts the bond length between two atoms to a fixed value, thereby preventing excessive oscillations and rapid fluctuations in bond length during the simulation. By imposing bond length constraints, the system's stability is preserved, and unphysical vibrations inherent in classical molecular dynamics simulations are eliminated. Consequently, larger time steps, typically around 2 fs, can be employed without compromising accuracy or stability.

The bond length constraint is implemented by incorporating a term into the potential energy function that enforces the desired bond length. The potential energy equation with a bond length constraint can be expressed as follows:

$$E_{\text{total}} = \sum_i \frac{1}{2} k_i (r_i - r_i^e)^2 + \sum_{i<j} v(r_{ij}) + \sum_i E_{\text{constraint}}(\vec{r}_i) \tag{3.20}$$

where k_i is the spring constant for bond i, r_i is the current bond length, r_i^e is the equilibrium bond length, and $E_{\text{constraint}}(\vec{r}_i)$ is the constraint energy for atom i. The second term represents the energy due to non-bonded interactions between atoms i and j, which is typically modeled using a pairwise potential function $v(r_{ij})$.

Another type of constraint commonly used in molecular simulations is the angle constraint. This constraint restricts the dihedral angle between four atoms to a fixed value, which can be employed for various purposes. For instance, angle constraints are often employed to restrict certain torsional motions in molecules or to maintain the chirality of specific molecular configurations. Maintaining chirality is particularly crucial when simulating biological molecules like proteins, which possess distinct chirality in their structure. The implementation of angle constraints follows a similar principle to that of bond length constraints. A term is added to the potential energy function, preventing deviations from the fixed angle and ensuring its maintenance throughout the simulation. By applying angle constraints, the system's behavior can be controlled and specific geometric relationships can be preserved.

The most common algorithm used for handling constraints in molecular simulations is the SHAKE algorithm. The SHAKE algorithm uses a Lagrange multiplier approach to iteratively solve for the positions of the constrained atoms while satisfying the constraints. The SHAKE algorithm updates the Lagrange multipliers at each time step to ensure that the constraints are satisfied within the specified tolerance.

Constraints can greatly improve the accuracy of molecular simulations by preventing unphysical behavior and maintaining the geometric and energetic properties of the system. However, constraints can also increase the computational cost of simulations, as the SHAKE algorithm requires additional calculations at each time step. It is important to choose appropriate constraints and to balance the benefits of constraints with the increased computational cost.

Thermostat

The temperature of the system is related to the kinetic energy of the particles, and therefore, when potential energy drops due to, e.g., a highly exoenergetic binding event, the excess kinetic energy is released, increasing the temperature of the system. It is often necessary to maintain a constant temperature throughout the simulation in order to simulate realistic physical systems. This is achieved by coupling the simulated system to a heat bath, or a thermostat, which modifies the distribution of kinetic energy so that it corresponds to a specified temperature.

The Andersen thermostat, introduced early in the development of molecular dynamics simulations, draws the velocity of a randomly selected particle from the correct velocity distribution. However, this approach can result in the loss of correlations in the system, limiting its usefulness. The Berendsen thermostat, a more popular approach, uniformly rescales all velocities in the system at each step to bring the system towards a desired temperature. However, this approach can lead to overdamped temperature fluctuations and does not sample the correct thermodynamic ensemble.

More recently, the Berendsen thermostat has been corrected by Bussi through the introduction of the Canonical Sampling through Velocity Rescaling (CSVR) method. Rather than targeting a fixed temperature, CSVR aims to bring the system towards a randomly fluctuating temperature. Another popular thermostat is the Nosé-Hoover thermostat, which is based on an extended Lagrangian approach and introduces a scaling factor for momenta and a fictitious mass-like term to account for the coupling inertia.

Barostat

Barostats, like thermostats, are used to maintain the system at desired conditions during molecular simulations. In particular, barostats are used to control the pressure of the simulated system. The pressure is influenced by various factors, including the volume, the number of particles present in the system, and the interactions between the particles.

To determine the pressure in molecular simulations, the pressure tensor is calculated based on the virial tensor, which is obtained from the positions and velocities of the atoms. The virial tensor provides information about the forces exerted by particles on each other, enabling the estimation of pressure. Barostats work by adjusting the volume of the simulation box to achieve the desired pressure. They ensure that the pressure is maintained within the specified range throughout the simulation.

There are two primary types of barostats: isotropic and anisotropic. Isotropic barostats employ a scalar reference pressure and uniformly scale the system size in all directions, maintaining isotropic conditions. On the other hand, anisotropic barostats allow independent scaling along individual directions, accommodating anisotropic systems where different directions may require different pressures.

One of the earliest and most commonly used barostats is the Berendsen barostat, which was introduced in 1984. The Berendsen barostat rescales the size of the simulation box at each time step to reach the desired pressure with a relaxation time constant, that controls the rate of volume change. However, it has been found that the Berendsen barostat leads to non-physical behavior, such as incorrect densities and surface tensions. Many barostat designs, such as Berendsen, share similarities with thermostats, and are classified as weak coupling algorithms. These algorithms

are crucial in allowing the system to fluctuate and relax, while achieving the desired temperature and pressure.

The Parrinello-Rahman barostat is another popular choice, which uses a dynamical matrix to couple the box dimensions to the pressure, and is particularly useful for simulating anisotropic systems. Specifically, in this algorithm, the simulation cell dimensions are considered as dynamic variables, which evolve according to the equations of motion.

In addition to the commonly used Berendsen and Parrinello-Rahman barostats, various other algorithms have been proposed in the literature. One such algorithm is the Martyna-Tuckerman-Tobias-Klein (MTTK) barostat, which is based on the Nosé-Hoover thermostat and maintains the pressure by coupling the system to a fictitious piston. The MTTK barostat allows for accurate sampling of the isothermal-isobaric ensemble and provides an effective approach for simulating systems with flexible geometries or significant changes in volume.

3.2 Application of Statistical Thermodynamics in Biophysics

3.2.1 Statistical Ensembles: Boltzmann Distribution

The preceding discussion has primarily focused on accurately modeling the temporal evolution of molecular systems using the classical approach or molecular mechanics. However, when making predictions about experimentally observable quantities, researchers are often less interested in the precise finite-length trajectory of a single system realization and more interested in the macroscopic properties that emerge from complex behaviors at the atomic level. To address this challenge, statistical thermodynamics provides a unifying framework by introducing the concept of ensembles.

In statistical thermodynamics, the study of many-particle systems in thermodynamic equilibrium relies on ensembles, which provide a framework for understanding their behavior. An ensemble represents a collection of replicas of a system, each in a different microscopic state (referred to as a microstate), but sharing the same macroscopic properties. In thermodynamic equilibrium, the system's thermodynamic variables reach a steady state and remain constant over time. The macroscopic properties of an ensemble, such as temperature, pressure, and volume, are determined by the experimental or simulation conditions.

Within this context, three main ensembles are commonly employed: the microcanonical ensemble, the canonical ensemble, and the isobaric-isothermal ensemble, which will be discussed later in this section.

A fundamental concept in statistical thermodynamics is the Boltzmann distribution, named after Austrian physicist Ludwig Boltzmann, who first introduced the idea in the 19th century. The Boltzmann distribution describes the probability distribution of a system in thermal equilibrium with its surroundings. The Boltzmann distribution arises from the fundamental postulate of statistical mechanics, which states that at thermodynamic equilibrium, all accessible microstates of an isolated system are equally probable. This postulate applies to systems with a constant energy, which are described by the microcanonical ensemble.

To derive the Boltzmann distribution, consider a large isolated system consisting of the system of interest and its surroundings. This system is assumed to be

in thermal equilibrium and maintained at a constant temperature. By applying the principles of statistical mechanics, which involve considering the probabilities of different microstates, the Boltzmann distribution emerges as a probabilistic description of the system. In statistical mechanics, the system is viewed as an ensemble of particles or molecules, each possessing its own energy state. The Boltzmann distribution arises from the assumption that, at thermal equilibrium, all accessible microstates of the system are equally probable. This assumption is known as the fundamental postulate of statistical mechanics.

At thermal equilibrium, the probability of the system being in a particular microstate, denoted as P_i, is proportional to the Boltzmann factor:

$$P_i = \frac{1}{Z} e^{-\beta E_i} \tag{3.21}$$

where E_i represents the energy of the i-th microstate, $\beta = \frac{1}{k_B T}$ is the inverse temperature, k_B is the Boltzmann constant, and Z is the partition function. The partition function serves as a normalization constant, ensuring that the probabilities sum to one over all possible microstates i of the system. It is defined as:

$$Z = \sum_i e^{-\beta E_i} \tag{3.22}$$

The Boltzmann factor, determined by the energy of a microstate and the temperature of the system, quantifies the probability of the system occupying a specific microstate at thermal equilibrium. However, it is crucial to recognize that the Boltzmann factor alone does not provide an absolute measure of likelihood or occupancy, but rather serves as a comparative measure among different microstates.

At higher temperatures, the Boltzmann distribution becomes flatter, meaning that the probabilities of occupying different microstates become more similar. This indicates a higher level of randomness and greater availability of energy states for the system to explore. As the temperature decreases, the Boltzmann distribution becomes more peaked, with certain microstates having higher probabilities than others. In this case, the system exhibits a preference for specific energy configurations and becomes more ordered.

It is important to note that the Boltzmann distribution assumes a system in thermal equilibrium, where all microscopic states are equally accessible. Deviations from equilibrium, such as non-equilibrium processes or transient states, may require alternative statistical approaches beyond the scope of the Boltzmann distribution.

From the perspective of molecular dynamics simulations, three important ensembles are commonly used: microcanonical, canonical, and isobaric-isothermal ensembles.

The microcanonical ensemble, also known as the NVE ensemble, describes a closed system with a fixed number of particles, volume, and total energy. This ensemble is particularly useful for studying isolated systems, such as a gas in a container with fixed energy, volume, and the number of particles. In this case, the probability of the system being in a particular microstate i is given by:

$$P_i = \frac{1}{\Omega(E)} \delta(E - E_i) \tag{3.23}$$

Here, E_i is the energy of microstate i, $\delta(E - E_i)$ is the Dirac delta function which ensures that the energy of the system is fixed at E, and $\Omega(E)$ is the total number

of microstates of the system with energy E, known as the microcanonical partition function.

While the total energy of the system is constant, fluctuations in the potential energy (due to interactions between particles) can lead to changes in the kinetic energy, resulting in a change in the temperature. To address this issue, the canonical ensemble (NVT) was introduced, which allows for energy exchange between the system and its surroundings while keeping the temperature fixed.

The canonical ensemble is particularly useful for studying systems that are in thermal contact with a heat bath at a constant temperature (T). In this ensemble, the probability of the system being in a particular microstate i is given by the Boltzmann distribution and can be expressed as Eq. **??**. In the canonical ensemble, the system is allowed to exchange energy with the surroundings, and the total energy of the system is not fixed. Instead, the system is characterized by a fixed temperature and the average energy of the system fluctuates around a well-defined value. This ensemble is particularly useful for studying systems in contact with a heat bath, such as a protein in solution, where the surroundings can exchange energy with the system and maintain a fixed temperature.

While the canonical ensemble is useful for studying systems in contact with a heat bath at a constant temperature, it may not be appropriate for systems in which pressure changes are important. For these systems, the isobaric-isothermal ensemble (NPT) was introduced, which allows for energy and particle exchange between the system and its surroundings, while keeping the temperature and pressure fixed.

The isobaric-isothermal ensemble is particularly useful for studying systems that are in contact with a heat bath and a piston at a constant temperature and pressure. In this ensemble, the probability of a system being in a particular microstate is given by:

$$P = \frac{1}{\Xi} e^{-\beta(E+PV)} \tag{3.24}$$

where Ξ represents the isobaric-isothermal partition function, which is defined as:

$$\Xi = \int e^{-\beta(E(\vec{r})+PV)} d\vec{r} dV \tag{3.25}$$

where N denotes the number of particles in the system, $E(\vec{r})$ represents the total potential energy of the system expressed in terms of the positions $\vec{r}$ of the particles, and the integral is taken over all possible configurations and volumes of the system.

It is worth noting that, in condensed-phase systems such as biopolymers in liquid water under ambient conditions relevant to biology, the NVT and NPT ensembles typically yield similar results. This is because liquid water is nearly incompressible, and only a small range of volumes is physically plausible under constant pressure.

3.2.2 Thermodynamic Properties: Entropy, Enthalpy, and Free Energy

The study of biological systems at the molecular level requires an understanding of the fundamental principles of thermodynamics. In particular, the thermodynamic properties of entropy, enthalpy, and free energy play a crucial role in describing the energetics of biological processes. These properties are intimately related to the behavior of molecules and the interactions between them, and can be used to predict the spontaneous direction and yield of chemical reactions in living systems.

Entropy is a fundamental concept in thermodynamics that quantifies the level of randomness or disorder within a system. It plays a significant role in various fields, including biophysics, where it is crucial for determining the spontaneity of processes.

One area where entropy is particularly relevant is in protein folding. During the folding process, the formation of a folded protein structure generally leads to a decrease in conformational entropy. However, this decrease is offset by an increase in the entropy of the surroundings, typically represented by the heat bath. Heat is released as interactions, such as hydrogen bonds, form during folding, contributing to an increase in the entropy of the surroundings. The hydrophobic effect, a key driving force in protein folding, involves both enthalpic and entropic contributions. When nonpolar amino acids aggregate in the protein core, their conformational entropy decreases, resulting in enhanced protein stability. Simultaneously, water molecules that were initially located at the hydrophobic interface are released into the solvent, thereby increasing the entropy of the solvent. Consequently, protein folding is driven by the intricate interplay between protein conformational entropy, the entropy of the surroundings, and solvent entropy. Together, these factors contribute to an overall increase in entropy throughout the folding process.

The Boltzmann distribution (described in section 3.2.1) and the concept of entropy are closely interconnected. The Boltzmann distribution provides the probability of finding a system in a specific microstate, while the entropy of the system can be derived from the probability distribution associated with a given statistical ensemble. The calculation of entropy is based on the equation:

$$S = -k_B \sum_i p_i \ln p_i \tag{3.26}$$

where S represents the entropy, k_B is the Boltzmann constant, p_i denotes the probability of the system being in microstate i, and the summation encompasses all microstates in a given ensemble. This equation holds true for any statistical ensemble.

In the canonical ensemble, the probability of a microstate is given by $p_i = \frac{e^{-\beta E_i}}{Z}$, where $\beta = \frac{1}{k_B T}$ is the inverse temperature, E_i is the energy of microstate i, and Z is the partition function. Using this expression for Z, we can relate the entropy to the internal energy of the system (U) as:

$$S = k_B \ln Z + \frac{U}{T} \tag{3.27}$$

In addition to protein folding, entropy also plays a significant role in other biological processes, including DNA double helix formation. The association of the hydrophobic regions of DNA, where nonpolar molecules come together, is thermodynamically favorable due to the favorable change in entropy associated with the surrounding solvent. When the DNA strands come together, water molecules that were previously ordered around the individual strands are released into the solvent, resulting in an increase in the entropy of the solvent. Furthermore, the formation of hydrogen bonds between complementary bases and stacking interactions in the DNA double helix releases heat to the surrounding environment. This heat transfer contributes to an increase in the entropy of the heat bath.

Enthalpy is related to the internal energy of a system, which includes the energy stored in the bonds and interactions between molecules, as well as the kinetic energy due to thermal motion. In thermodynamics, enthalpy (H) is often used to describe

heat flow in a system at constant pressure, and is defined as the sum of the internal energy (U) and the product of pressure (P) and volume (V) i.e., $H = U + PV$. In biophysics, along with entropy, enthalpy is used to describe processes such as DNA-protein binding. The binding of DNA to a protein involves the release or absorption of heat, which results in a change in enthalpy (ΔH). This change in enthalpy can provide valuable information about the strength of the interaction between the DNA and the protein, as well as the degree of cooperativity in the binding process.

Enthalpy can also be expressed in terms of the partition function as:

$$H = -k_B T^2 \frac{\partial \ln Z}{\partial T} \tag{3.28}$$

Free energy is a thermodynamic property that combines the effects of entropy and enthalpy and can be expressed as $G = H - TS$. The change in free energy (ΔG) of a system is a measure of the maximum amount of work that can be extracted from the system when it undergoes a change under constant temperature and pressure conditions. A negative value of ΔG indicates that a reaction or process is thermodynamically favorable and will proceed spontaneously in the forward direction. In biological systems, free energy plays a crucial role in many processes such as specific DNA-protein binding, where the difference in free energy between the bound and unbound states provides information about the strength and specificity of the interaction. The standard free energy change of the reaction determines the extent to which it will proceed, and the equilibrium constant of the reaction is related to the free energy change according to the following equation:

$$\Delta G = -k_B T \ln K_{eq} \tag{3.29}$$

where k_B is the Boltzmann constant, T is the temperature, and K_{eq} is the equilibrium constant of the reaction. The negative sign in the equation indicates that a decrease in free energy corresponds to an increase in the equilibrium constant, and thus to a more favorable reaction.

The partition function (Z) can be used to calculate the free energy of a system, as well as the probability of a system being in a particular state, as per the following equation:

$$G = -k_B T \ln Z \tag{3.30}$$

where G is the free energy, and T is the temperature. This equation shows that the free energy is related to the logarithm of the partition function. However, it's important to note that this equation gives an absolute free energy value and not its difference between two states. To calculate the free energy difference between two states, one needs to calculate the partition function for each state separately and then take the difference. In practice, this is often done through simulations, where the partition function is approximated to extract macroscopic properties such as ΔG and ΔH.

3.2.3 Potential of Mean Force and Its Relation to Free Energy

Computing free energy differences between different states of a system is crucial in biophysics as biological macromolecules often undergo conformational changes,

binding events, and other dynamic processes. Conventional or equilibrium molecular dynamics (MD) simulations can be used to determine the free energy of binding or conformational changes; however, these simulations require a long time to converge to equilibrium, making them computationally expensive. Additionally, in many cases, the reaction or process of interest may not occur within a reasonable time scale.

Hence, applying an external force and subsequently calculating the work done by that force along a reaction coordinate or collective variable (CV) is a powerful technique to obtain the potential of mean force (PMF) or the free energy of a system as a function of a given reaction coordinate(s). By integrating the mean force from an ensemble of configurations, the PMF provides valuable insights into the relationship between microscopic forces and ensemble statistics. This connection between microscopic forces and ensemble statistics is an invaluable tool for interpreting experimental results, and recent efforts have aimed to narrow the gap between experimental and simulation data.

Consider a system with coordinates $\vec{r} = (r_1, r_2, ..., r_N)$, where N is the total number of particles in the system. The potential energy function of the system is denoted as $V(\vec{r})$. The potential of mean force (PMF) along a reaction coordinate or collective variable (CV), $s(\vec{r})$, is a measure of the system's potential energy when the variable s is restrained to a specific value. The remaining degrees of freedom are sampled according to the equilibrium distribution $P(\vec{r})$. Mathematically, the PMF is expressed as:

$$W(s) = -k_B T \ln P(s) \tag{3.31}$$

where k_B is the Boltzmann constant, T is the temperature, and $P(s)$ is the probability density of finding the system with the variable s. The negative logarithm of the probability distribution is proportional to the free energy change, hence the PMF is related to the free energy, ΔG, associated with the process of interest through the following equation:

$$\Delta G(s) = -k_B T \ln \left(\frac{P(s)}{P_0} \right) \tag{3.32}$$

where P_0 is the probability of a reference state and ΔG denotes the free energy difference with respect to this reference state.

The potential of mean force provides a means to calculate the free energy difference between two states, denoted as A and B. This can be expressed using the equation:

$$\Delta G_{AB} = -k_B T \ln \left(\frac{\int_{s_A} e^{-\beta W(s)} ds}{\int_{s_B} e^{-\beta W(s)} ds} \right) \tag{3.33}$$

The integral in the numerator represents the average work required to move the system from state A to state B along the reaction coordinate (s), and the denominator is the normalization constant. In practice, the PMF is calculated by performing a series of biased simulations in which the variable of interest is constrained to different values.

The PMF provides valuable information about the energetic barriers and stability of the system, allowing for the calculation of free energy differences associated with the transition between states A and B.

However, it should be noted that the selection of the CV must take into consideration the experimental setup. In other words, an adequate forward model of the process of interest must be available. While the construction of proper CVs that simultaneously ensure rapid convergence of the simulation statistics, correspond to experimental observables, and account for all nuances of the process can be an art in itself, the potential benefits of PMF calculations make it a worthwhile endeavor in biophysical research.

3.2.4 Methods for Free Energy Calculation

Umbrella Sampling

Thanks to the direct connection between free energies and probability densities, it is theoretically possible to calculate free energy profiles through the Boltzmann inversion by multiplying the logarithm of probability by $-k_BT$. However, this logarithmic relationship can complicate matters in practice. For instance, a free energy difference of 10 kcal/mol at ambient temperature corresponds to an almost 20 million-fold ratio of probabilities. Consequently, obtaining a single sample from the high-energy state would require tens of millions of independent samples from the low-energy state.

As a result, specialized enhanced sampling techniques have been developed to calculate free energies accurately and efficiently. Umbrella sampling is one such technique that has found extensive use in biophysics due to its simplicity and reliability.

Umbrella sampling is a simulation-based method that can be used to calculate the potential of mean force (PMF) along a reaction coordinate. The idea behind this technique is to add a biasing potential, typically harmonic, to the Hamiltonian to restrain the system along the reaction coordinate. The harmonic potential acts as an "umbrella" and the simulation samples the probability distribution of the system at different positions along the reaction coordinate. These samples are then re-weighted using the weighted histogram analysis method (WHAM) to obtain the PMF.

The Hamiltonian for the umbrella sampling simulation includes the kinetic energy of the particles in the system, the inter-particle potential energy, and the harmonic potential energy, and can be written as follows:

$$\hat{H}(\vec{r}) = \sum_i \frac{\vec{p}_i^{\,2}}{2m_i} + V(\vec{r})) + \frac{1}{2}k_{\text{umb}}(s(\vec{r}) - s_0)^2 \tag{3.34}$$

where $\vec{r}$ represents the coordinates of all the particles in the system, $\vec{p}_i$ represents the momentum of particle i, m_i represents its mass, $V(\vec{r})$ represents the inter-particle potential energy, k_{umb} represents the force constant of the harmonic potential, $s(\vec{r})$ represents the reaction coordinate and s_0 is the reference position of the harmonic potential. The probability distribution of the system along the reaction coordinate can be written as (see also eq. 3.32):

$$P(s) = \frac{1}{Z}\int d\vec{r}\,\delta(s - s(\vec{r}))e^{-\beta H(\vec{r})} \tag{3.35}$$

In this equation, we integrate over all possible configurations of the system, represented by $d\vec{r}$, but the Dirac delta function $\delta(s - s(\vec{r}))$ restricts the integration to only those configurations that have a specific value of the reaction coordinate, s. The term

$e^{-\beta H(\vec{r})}$ is the Boltzmann factor, which weights each configuration by its energy, and Z is the partition function, which normalizes the probability distribution.

To obtain the PMF, simulations are performed at different positions along the reaction coordinate with different harmonic potentials, creating a series of windows. The umbrella potential is set up at the initial position and remains fixed throughout the simulation. The equilibrium distribution of each window is obtained by running simulations for a sufficient amount of time.

The weighted histogram analysis method (WHAM) is a powerful post-simulation analysis technique that is commonly used to calculate the potential of mean force (PMF) from data generated by umbrella sampling simulations. The fundamental concept behind WHAM is to merge data from multiple umbrella sampling simulations to achieve a more accurate PMF. WHAM's success is primarily due to the reweighting of probability distributions obtained from each umbrella sampling simulation to generate the unbiased distribution of the reaction coordinate across the entire coordinate range from the US calculations, among others.

The unbiased distribution of the reaction coordinate, denoted as $P(s)$, is determined using the following equation:

$$P(s) = \frac{\sum_{i=1}^{N_w} g_i^{-1} h_i(s)}{\sum_{j=1}^{N_w} n_j g_j^{-1} e^{-\beta(w_j(s)-f_j)}} \tag{3.36}$$

where N_w represents the number of umbrella sampling simulations, g_i is given by $g_i = 1 + 2\tau_i$, where τ_i is the autocorrelation time of the umbrella window i (in units of the simulation frame time step). The term $h_i(s)$ represents the histogram of observed samples in the i-th simulation at reaction coordinate s, and n_j is a normalization factor that ensures proper scaling of the histograms.

The term $w_j(s)$ represents the bias potential associated with simulation j, and f_j is the free energy offset for that simulation. The coefficients f_j are determined through the following equation:

$$e^{-\beta f_j} = \int ds, e^{-\beta w_j(s)} P(s) \tag{3.37}$$

This integral equation ensures the proper normalization of the probability distribution $P(s)$. Note that the coefficients f_j are determined self-consistently. Initially, an estimate of $P(s)$ is used to calculate f_j using Eq. 3.37. Then, Eq. 3.36 is solved iteratively, updating $P(s)$ and re-calculating f_j until convergence or self-consistency is achieved. This iterative process ensures that the probability distribution $P(s)$ and the associated free energy offsets f_j are determined accurately.

It is important to consider that the accuracy of the results obtained by umbrella sampling simulations may depend on several factors, such as the choice of reaction coordinate, the number of windows used, the size and shape of the umbrella potential, and the number of sampling points within each window. Therefore, careful selection and optimization of these parameters are essential for obtaining accurate results. Moreover, it is worth noting that umbrella sampling simulations can be computationally expensive, as simulations need to be performed at multiple positions along the reaction coordinate. This can result in long simulation times and large computational resources required for the analysis. However, the use of advanced sampling techniques, such as replica exchange umbrella sampling, can improve the

efficiency of the sampling process and reduce the computational cost of the simulations.

Metadynamics

Metadynamics is another enhanced sampling technique that is widely used in computational chemistry or biology to explore the free energy landscape of complex biomolecular systems. One of the key advantages of metadynamics over other enhanced sampling techniques, such as umbrella sampling, is its ability to explore higher dimensional free energy surfaces without suffering from the curse of dimensionality.

In metadynamics, an external history-dependent bias potential is applied as a function of the collective variables (CVs), which represent a selected number of degrees of freedom of the system. The fundamental idea behind this technique is to bias the system's potential energy to push it out of local minima and to make all values of the collective variable equiprobable. Once this is achieved, the negative biasing potential becomes exactly equal to the free energy.

The biasing potential in metadynamics is deposited on-the-fly as a sum of Gaussian-shaped increments along the system trajectory in the CVs space to discourage the system from revisiting the configurations that have already been sampled. The width of these increments, σ, needs to be fixed for each collective variable. The bias potential at time t can be expressed as follows:

$$V(\vec{s},t) = \sum_{k\tau<t} W(k\tau) e^{-\sum_{i=1}^{d} \frac{(s_i - s_0^i(k\tau))^2}{2\sigma_i^2}} \tag{3.38}$$

where $\vec{s}$ represents the collective variables, $W(k\tau)$ is the height of the Gaussian deposited at time $k\tau$, $s_0^i(k\tau)$ is the value of the i-th component of the collective variables at the k-th deposition time τ, and σ_i is the width of the Gaussian in the i-th CV. The sum over k and τ indicates that the potential energy is the sum of Gaussians deposited at different times up to time t. The bias potential is accumulated over time as the simulation progresses, leading to a flat distribution of the collective variables and an accurate estimate of the free energy surface.

Despite its benefits, metadynamics has two major drawbacks that should be considered. The first drawback is related to the fact that, in a single run, the bias potential V does not converge, but rather oscillates around it. As a consequence, the bias potential overfills the underlying free energy surface (FES), and it can be challenging to determine when to stop the simulations. However, as a general rule, if metadynamics is used to find the closest saddle point, the simulation should be stopped as soon as the system exits from the minimum. The second major drawback of metadynamics is the challenge of identifying a set of appropriate CVs that can accurately describe complex processes. This task is far from trivial and requires careful consideration and expertise.

Well-tempered metadynamics provides a solution to the convergence problem in traditional metadynamics. In this approach, the bias deposition rate decreases over the course of the simulation, which is achieved by using a modified expression for

the bias potential:

$$V(S,t) = \sum_{t'=0,\tau_G,2\tau_G,\ldots}^{t'<t} We^{-V(S(q(t')),t')/\Delta T} \exp\left(-\sum_{i=1}^{d} \frac{(s_i(q) - s_i(q(t')))^2}{2\sigma_i^2}\right) \tag{3.39}$$

where the term $We^{-V(S(q(t')),t')/\Delta T}$ represents the effective height of the Gaussians deposited at each time step. The height decreases over time as the simulation progresses and the bias potential increases. The parameter τ_G represents the deposition time, $S(q)$ represents the collective variables, $V(S(q(t')),t')$ is the bias potential at the previous deposition times, ΔT is an input parameter with the dimension of temperature, and σ_i is the width of the Gaussian in the i-th collective variable. The histogram of the S variables collected during the simulation is used to calculate the bias potential at each time step. The time derivative of the bias potential, $V'(S,t)$, is employed to adjust the deposition rate in real-time during the simulation. This approach ensures that the bias potential converges to the true free energy surface, avoids overfilling, and might save computational time. Moreover, it enables the estimation of the free energy surface without the need for *a priori* knowledge of the system, which is particularly useful when exploring complex processes.

Conformation flooding introduced by Grubmuller [207], is a very similar method to metadynamics, where the biasing potential is added in the form of Gaussians to escape from free energy basins. Although it was proposed before metadynamics, it did not gain as much attention from computational chemists/biologists as Parrinello's [208] metadynamics did.

3.2.5 Alchemical Transformations and Free Energy Calculations

Precisely estimating relative free energy differences plays a crucial role in various processes, such as understanding the binding of small molecules to specific protein targets. Alchemical free energy calculations provide an elegant and efficient approach to tackle this challenge. By smoothly transforming the chemical identity of atoms or molecules, alchemical methods enable the calculation of free energy differences between different states without requiring extensive sampling or costly experimental measurements. These calculations offer valuable insights into the thermodynamics of molecular systems, including phenomena like binding, and facilitate the prediction of important properties.

In the realm of enhanced sampling methods, alchemical free energy calculations offer distinct advantages through the concept of morphing chemically distinct species. This approach involves defining two end states, commonly referred to as "state A" and "state B", using different molecular topologies to represent the initial and final states of a chemical process. This gives rise to two distinct Hamiltonians, $\hat{H}_A$ and $\hat{H}_B$. To connect these two end states in a continuous manner, a coupling parameter λ is introduced, which allows for the interpolation between the two Hamiltonians. The Hamiltonian at a given λ value is constructed using linear interpolation as follows:

$$\hat{H}(\lambda) = (1-\lambda)\hat{H}_A + \lambda\hat{H}_B \tag{3.40}$$

All properties of the molecule, such as charges, bond lengths, reference angles, dihedrals, and effective radii, are generally interpolated linearly from the initial state

(A) to the final state (B). However, this method introduces some numerical instabilities near the endpoints, such as disappearing atoms into non-interacting dummies. The reason for this instability is that numerical integration uses finite-length steps, and during one such step, the algorithm can rapidly transition from a region where forces are negligible to a region near a singularity, especially when charges are still present on the vanishing particle. To address this issue, soft-core potentials have been developed as a solution to avoid singularities that occur at integer values of λ. These soft-core potentials modify the Lennard-Jones and Coulombic interaction terms and are used to make the alchemical transformation more stable and computationally feasible.

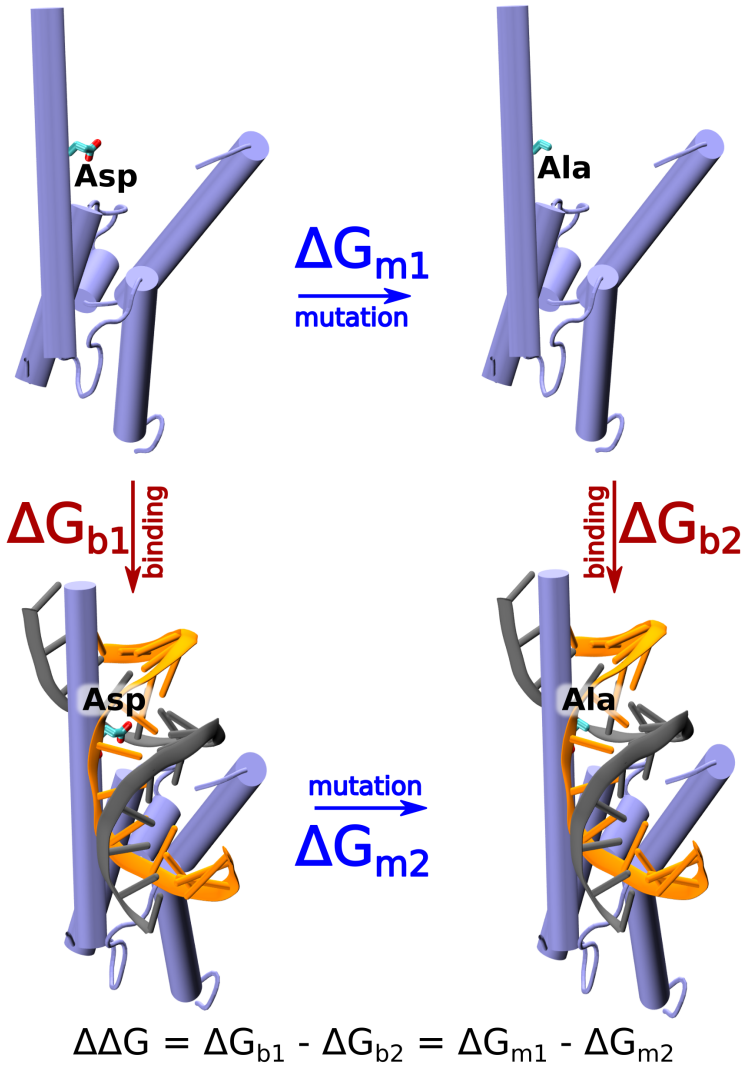

FIGURE 3.2: Example of a thermodynamic cycle used for computing the effect of single-point mutation in a protein on its DNA binding affinity. By exploiting the state function property (i.e., the fact that the sum of increments along a cycle is 0), the relative binding free energy ($\Delta\Delta G$) is obtained by subtracting the free energies of "alchemically" transforming Asp into Ala in the absence and in the presence of DNA (ΔG_{m1} and ΔG_{m2}, respectively).

The thermodynamic integration (TI) approach is a method used to calculate the change in free energy ΔG resulting from an alchemical transformation. This involves estimating the ensemble averages of $\frac{\partial \hat{H}}{\partial \lambda}$ at constant values of λ and then numerically integrating the resulting equation. The formula is as follows:

$$\Delta G = \int_0^1 d\lambda' \left\langle \frac{\partial \hat{H}}{\partial \lambda} \right\rangle_{\lambda=\lambda'} \tag{3.41}$$

Although TI has been a popular method in the past, the Bennett acceptance ratio (BAR) method [209] or its more recent multistate version (mBAR) [210] are now often used as an alternative to estimate ΔG resulting from alchemical transformations. Both of these methods are based on Zwanzig's [211] fundamental free energy perturbation equation:

$$\Delta G_{A\to B} = -k_B T \log \left\langle e^{-\frac{\hat{H}_B - \hat{H}_A}{k_B T}} \right\rangle_A \tag{3.42}$$

Here, k_B is the Boltzmann constant and T is the temperature. The expectation value $\left\langle e^{-\frac{\hat{H}_B - \hat{H}_A}{k_B T}} \right\rangle_A$ is evaluated in the initial state A and represents the average value of the exponential of the energy difference between the final state B and initial state A over all configurations of the initial state. BAR and mBAR are statistically efficient methods and have been shown to produce accurate results for a wide range of alchemical transformations.

The alchemical free energy method is generally used for calculating relative free energies. For example, it can be used to determine the pK_a shift in two separate environments, such as the difference in the protonation/deprotonation free energy of an amino acid between water and a specific environment, such as the DNA interface. More complex cases, such as determining the contribution of an amino acid to DNA affinity in terms of the difference in mutational free energy of that amino acid in different environments, can be accomplished by constructing an appropriate thermodynamic cycle (see Fig. 3.2).

3.3 Methodology Employed in the Study

3.3.1 Preparation of Molecular Systems

DNA-Protein Complexes with Asp/Glu Residues at the Interface

In order to investigate the role of acidic residues in base readout, as outlined in objective 1 (of Chapter 1), I analyzed four high-resolution structures of B-DNA duplexes bound by different sequence-specific transcription factors containing Asp/Glu residues that interact with cytosine in the major groove (see Fig. 3.3). These structures were: (1) the bHLH domain of CLOCK and BMAL1 (PDB id: 4H10) [148], (2) Zif268 zinc-finger (PDB id: 1ZAA) [149], (3) the DNA-binding domain of Myb (PDB id: 1MSE) [150], and (4) the erythroblast transformation-specific domain of ERG3 (PDB id: 5YBD) [212]. Additionally, I also included the telomeric protein TRF1 (PDB id: 1W0T [213]) in the study, as the contribution of an Asp residue in this protein has been previously assessed [214].

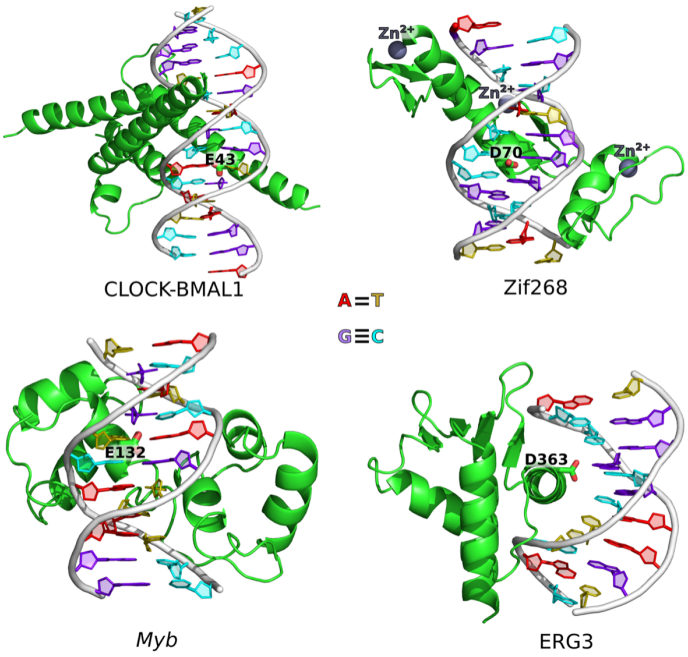

FIGURE 3.3: Structural representation of the reference DNA-protein complexes used in this work to assess the role of Asp/Glu residues in DNA binding.

To explore the sequence space, I generated 4–5 variants for each of the reference DNA/protein complexes by substituting the cytosine that directly H-bonds to Asp/Glu with all possible canonical bases, including thymine (C2T), adenine (C2A), and guanine (C2G), as well as by mutating all the nucleobases within 5 Å of Asp/Glu (all-5 Å) through transitions (purine-to-purine and pyrimidine-to-pyrimidine substitutions) or transversions (purine-to-pyrimidine and vice versa) (Fig. 3.4). The substitutions were carried out using the X3DNA package [5]. As for TRF1, considered only one off-target variant, which was an inverse telomeric sequence, in accordance with the previous study [214].

To explore the reasons behind the preference of acidic residues for cytosine over adenine, I utilized a model system where a single propionic acid anion interacts with a B-DNA decamer in the major groove. To ensure equal accessibility of cytosine and adenine to the propionate at the center of the decamer, I employed a DNA sequence: 5′-C·A·T·G·T·**C**·**A**·A·T·C-3′.

Furthermore, to investigate the effect of non-canonical BII backbone conformation of B-DNA on this preferential interaction, I also used a second sequence (5′-G·A·T·T·G·**A**·**C**·A·T·G-3′), where the GpA dinucleotide step involving the central adenine (A6) had a strong tendency to adopt the BII conformation [13, 215]. The DNA decamers were constructed using the X3DNA package [5].

To directly assess the impact of the adenine N7 atom on cytosine/adenine preferences, I created an additional variant of both sequences, wherein the partial charge on N7 of the central adenine was modified from -0.62 to -0.02 (aromatic -CH group). To avoid artificial attraction of the propionate, the compensating charge of 0.6 was evenly distributed over the remaining atoms of the modified adenine (denoted as A*), making it electrically neutral.

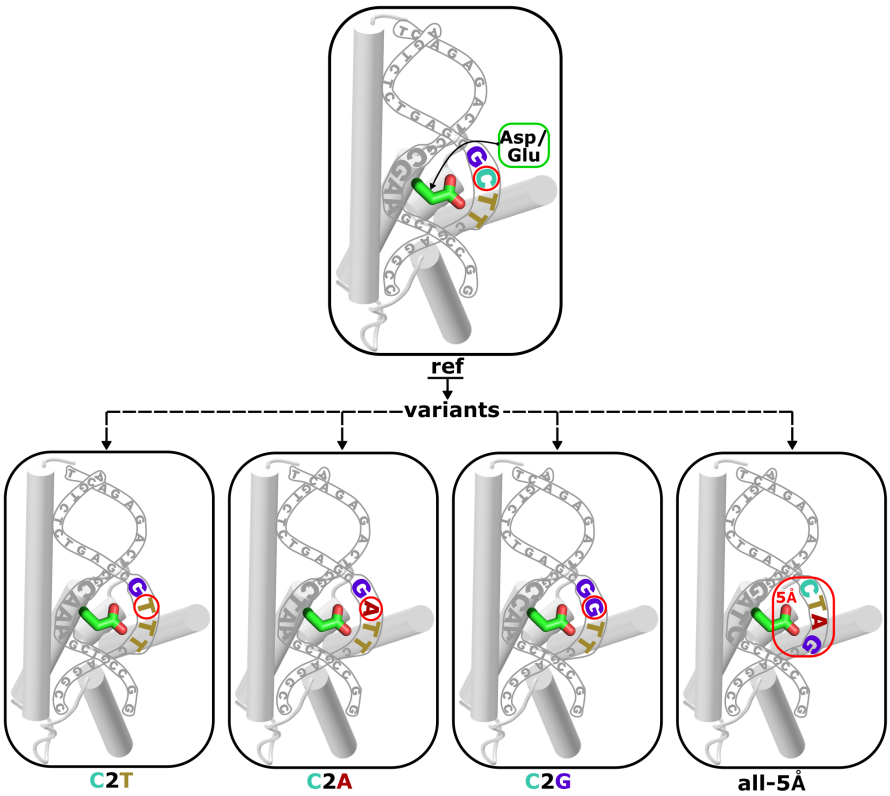

FIGURE 3.4: Schema outlining the methodology for sampling DNA sequence space. DNA sequence variants were generated from the selected experimental protein/DNA complexes (ref) containing Asp/Glu residues that directly interact with cytosine in the major groove of DNA. The variants were obtained either by substituting the cytosine directly H-bonded to Asp/Glu with thymine (C2T), adenine (C2A), or guanine (C2G), or by mutating all nucleobases within 5 Å of Asp/Glu (all-5 Å) through either transitions or transversions.

The solvation of all aforementioned molecular systems was accomplished by incorporating TIP3P water molecules [216] into a dodecahedron box, which was constructed such that the solute and the box edges were at least 1 nm apart. To attain a physiological salt concentration of 0.15 M and balance the charge of the system, K^+ and Cl^- ions were introduced. In the case of the Zif268 zinc finger, Zn^{2+} ions that existed in the crystal structure were retained and bound to the coordinating residues by utilizing the bonding parameters from the Zinc Amber Force Field (ZAFF) [217].

EXOG and its substrates

To address objective 2 (as discussed in Chapter 1) and unravel the mechanism by which human mitochondrial EXOG recognizes and induces the A-DNA conformation, I initiated my study by examining the homodimeric structure of EXOG co-crystallized with a DNA duplex obtained from the Protein Data Bank (PDB id: 5T5C) [17]. Additionally, I analyzed another structure (PDB id: 6IID) where EXOG is bound to an RNA-DNA chimeric duplex (R2-DNA/DNA, where the first two nucleotides are from RNA) [183] to investigate the substrate specificity of EXOG. To confirm EXOG's role in driving the B to A conformational transition of the DNA substrate,

I hypothesized that the DNA duplex substrate should revert back to its native B-form upon detachment of EXOG, whereas the R2-DNA/DNA should remain in the A-form, regardless of its bound or unbound state. Accordingly, I prepared isolated forms of the DNA duplex and R2-DNA/DNA using the crystallographic data mentioned above for simulations. To identify the key residues in the core domain of EXOG that might be responsible for the conformational transition, I analyzed the binding interface of EXOG and generated variant systems by mutating residues located in the vicinity of the 5′-end or conformational transition site. These mutations were selected based on theoretical considerations that they could induce conformational transitions in the DNA substrate. To examine the role of the wing domain in the B to A conformational transition, I created a system in which Arg314, the only residue located near the 5′-end of the DNA substrate strand in the wing domain, was mutated.

All the prepared systems for EXOG simulations were solvated with TIP3P water molecules in a dodecahedron box. The dimensions of the box were chosen to ensure a minimum distance of 1.2 nm between the solute and the box edges. The crystallographic water molecules that were bound with Mg^{2+} ion in each catalytic site of the homodimer were preserved. Additional K^+ and Cl^- ions were added to adjust the salt concentration to 0.15 M and neutralize the net charge of the system. A structural representation of the water molecules bound with Mg^{2+} ion is shown in Figure 2.10.

G-quadruplex and DHX36

To address objective 2 (as discussed in Chapter 1) and explore the recognition mechanism of all-parallel G-quadruplex (G4) structures by DHX36 helicase, I acquired the co-crystal structure of bovine DHX36 helicase bound to a G4 with a 3′ single-stranded (ss) DNA tail from the Protein Data Bank (PDB id: 5VHE) [194]. Although there are other solved structures of DHX36, such as mouse and drosophila [196, 197], they are not appropriate to use as a starting point for the analysis of the recognition process because they do not contain a bound G4. Using Modeller [218], I built the missing portions in 5VHE, a 20-residue linker between the DHX36-specific motif (DSM) and RecA1, and a 13-residue loop in RecA2 (see Fig. 2.11), to obtain the full-length DHX36 structure. The 3′-ssDNA tail was discarded for several reasons: (1) my primary interest was to explore the recognition process of only all-parallel-G4, (2) experimental evidence shows that removal of the 3′-ssDNA tail does not substantially affect the binding of G4 to the helicase [196, 198], (3) the flexible ssDNA tail can impede enforced G4 dissociation, which slows down the convergence of free energy computations and prevents from drawing firm conclusions. Therefore, I kept the first 17 nucleotides, which form the canonical DNA^{myc}-G4 sequence: $[A(G)_3T(G)_3TA(G)_3T(G)_3]$. To understand the role of different domains and possible cooperativity in the recognition process, I simulated two variants of the helicase: i) the complete structure involving all interfaces with G4, and ii) the structure with the DSM α1 helix discarded along with the entire N-terminal linker region and several initial residues of RecA1 forming an overhanging loop (see Fig. 2.11) in MD–refinement (residues Pro57–Ile200).

To better understand the contribution of the DSM α1 helix to the recognition and binding of the parallel G4 fold, and to investigate the preference of the helix for the 5′-G-tetrad, I prepared three additional systems with truncated DSM α1 helix (22 residues, Pro57–Lys78) and DNA^{myc}-G4 in solution. These systems were designed

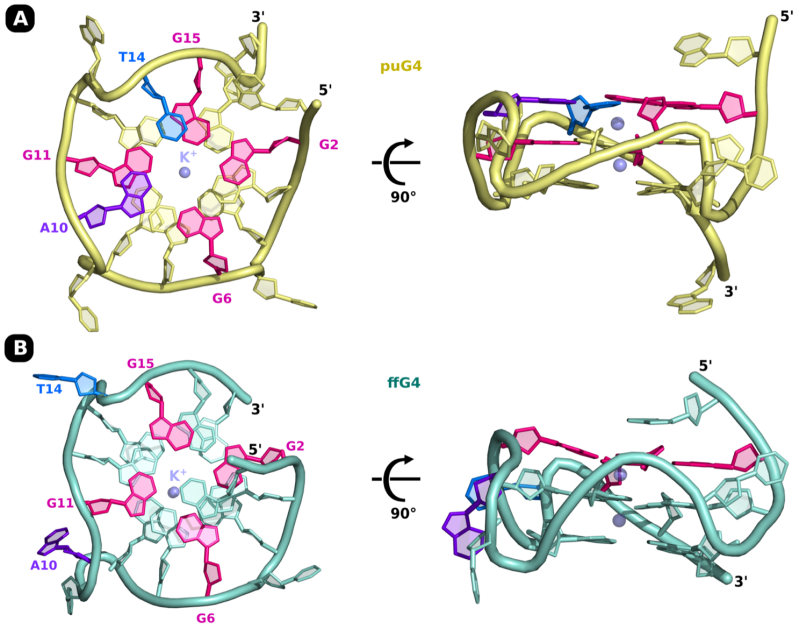

FIGURE 3.5: Structural comparison of partially unfolded (puG4) and fully folded (ffG4) states of DNA^{myc}-G4 (**A** and **B**, respectively); differences are highlighted by color labeling. Potassium ions (K^+) are depicted in semi-transparent spheres.

to facilitate better free energy convergence and to investigate the binding properties of DSM α1 helix to different G4 conformations.

The three individual DSM/G4 systems were prepared as follows: (a) DSM α1 helix bound to the register-shifted partially unfolded G4 (puG4; see Fig. 3.5A) captured in the co-crystal structure (PDB id: 5VHE); (b) DSM α1 helix bound to the 5′-G-tetrad of the native fully folded DNA^{myc}-G4 (ffG4; Fig. 3.5B) taken from its solution NMR structure (PDB id:1XAV); and (c) DSM α1 helix bound to the 3′-G-tetrad of ffG4. The first system (a) was prepared directly from the crystallographic structure 5VHE, while systems (b) and (c) were prepared by fitting the NMR structure (PDB id:1XAV) [41] of ffG4 to puG4 of system (a), using hydrogen-suppressed backbone atoms of the G-tetrad. This ensured that the initial orientation and positioning of the DSM α1 helix with respect to the respective G-quadruplexes were the same in all three systems.

Similar to the EXOG systems, all of the prepared systems were placed in a dodecahedral box with dimensions chosen such that the solute was at least 1.2 nm away from the edges of the box, and were then solvated with TIP3P water molecules. To ensure the stability of the G-quadruplexes, two K^+ ions were kept intact in the central channel of each G-quadruplex. To achieve a physiological salt concentration of 0.15 M and to neutralize the net charge of the system, additional K^+ and Cl^- ions were added.

3.3.2 Simulation details

All molecular dynamics (MD) simulations were carried out in the isothermal-isobaric (NPT) ensemble using Gromacs 2018.8 [219] in conjunction with the PLUMED 2.2.3 plugin [220] and the Amber-parmbsc1 force field [221]. The temperature was

maintained at a constant value of 300 K using the v-rescale thermostat [222] with a time constant of 0.1 ps, and the pressure was kept at 1 bar using the isotropic Parrinello-Rahman barostat [223]. To ensure periodicity in all directions, applied periodic boundary conditions were applied in three dimensions, and computed long-range electrostatic interactions using the particle mesh Ewald (PME) method [224]. A real-space cutoff of 1.2 nm and a Fourier grid spacing of 0.12 nm were used for the calculations. The Lennard-Jones potential with a cutoff of 1.2 nm and a switching distance of 1 nm was used to describe van der Waals interactions. To avoid singularity points in all alchemical free energy simulations, I used the default Gromacs soft-core potentials. P-LINCS [225] was used to constrain the bond lengths of protein and DNA molecules, while SETTLE [226] was used to maintain the geometry of water molecules. The leap-frog algorithm with a time step of 2 fs was employed for the integration of the equations of motion. Prior to all production simulations, I equilibrated all systems for at least 100 ns.

3.3.3 Free energy simulations

DNA Binding Affinity of Asp/Glu Residues: Alchemical Free Energy Calculations

The contribution of Asp/Glu residues to the DNA-binding affinity of selected transcription factors (as described in Section 3.3.1) was investigated by calculating the binding free energy difference ($\Delta\Delta G$) between the wild-type protein and its mutant in which a specific Asp/Glu was replaced with alanine. The interfacial Asp/Glu residues chosen for each protein are shown in Fig. 3.3. A thermodynamic cycle (depicted in Fig. 3.2) was employed to determine $\Delta\Delta G$, which involved computing and subtracting the free energies associated with the "alchemical" transformation of Asp/Glu to alanine in the absence (ΔG_{m1}) and presence (ΔG_{m2}) of DNA bound to the protein. This was achieved by independently simulating the system for a range of scaling parameter values (λ), which vary between 0 and 1, and interpolate linearly between the potential energy functions of the physical end states.

To expedite free energy convergence, neighboring λ-windows were permitted to exchange their configurations every 0.5 ps based on the Metropolis criterion. The values of λ were optimized to achieve an acceptance rate of at least 10%, using an in-house script (`https://gitlab.com/KomBioMol/converge_lambdas`). The pmx web-server [227] was used to generate the hybrid topology for the Asp/Glu $\rightarrow$ Ala mutations. Each system was simulated for at least 300 ns in each λ-window until a reasonable convergence of $\Delta\Delta G$ was attained (see Fig. A.1). To obtain the free energy changes resulting from the alchemical simulations, I employed the Bennett acceptance ratio (BAR) method, which was performed using the Gromacs utility gmx bar.

Due to a partial dissociation of the C2A variant of the Zif268-DNA complex, I observed a deviation from the well-defined bound state. To avoid any potential artifacts in $\Delta\Delta G$, I implemented a harmonic restraint to maintain the initial center-of-mass distance between Zif268 and DNA. The restraint was set with a force constant of 119.61 kcal/(mol·nm^2). For the analyses, I only used data obtained for the restrained complex to ensure the reliability of the results.

Umbrella Sampling Free Energy Calculation of B to A-DNA Transition by EXOG

The objective of this study was to investigate the B to A transition of DNA substrate bound by EXOG, and to evaluate the impact of the wing domain and core domain residues on this process, which is related to objective 2 described in Chapter 1.

To accomplish this objective, I employed the umbrella sampling (US) method [228] to calculate relevant free energy profiles. To verify the preferred state of DNA when immersed in the deep substrate-binding groove of ExoG, I performed free energy calculations using the wild-type (WT) ExoG. Specifically, I conducted steered molecular dynamics (SMD) simulations, in which the A-form of DNA (as observed in the crystallographic structure 5T5C, see Fig. 2.10) was forced to transition to the B-form over a 200 ns simulation time. Given that only the first 2–3 nucleotides acquire the A-form (see Fig. 2.10), I applied a moving harmonic potential with a force constant of 358.85 kcal/(mol·nm^2) to the center-of-mass coordinates of the phosphate-group-atoms between the 2^{nd} and 3^{rd} nucleotides as the separation distance, and pulled the system from 0.55 nm (representing A-form) to 0.80 nm (representing B-form). From the resulting SMD trajectory, I extracted six uniformly distributed, 0.05 nm-separated frames as initial configurations for the US windows. Each window was then simulated for 300 ns (or until convergence was achieved) using a harmonic potential with a force constant of 239.23 kcal/(mol·nm^2) to restrain the system along the reaction coordinate r.

To analyze the respective roles of wing and core domain residues in facilitating the B to A transition of the bound-substrate-DNA, I computed the free energy from two model systems. In the first system, Arg314, which is the only residue situated near the 5′-end of the DNA substrate strand in the wing domain, was mutated to Ala. In the second system, Arg109 from the core domain, which can intimately interact with the 2^{nd} and 3^{rd} nucleotides, was mutated to Ala as well, as described in section 3.3.1. To ensure that the effect of mutations was properly captured without perturbing the systems differently, these mutations were introduced in the initial configurations of each US window of the WT system. This approach offers two advantages: (1) there is no need for any additional SMD simulations for these mutation models, and (2) it guarantees that all three systems (i.e., WT, R314A, and R109A mutants) started from the same point. The US windows of these mutated systems were similarly simulated for 300 ns using the same restraining protocols mentioned above (for the WT system). To examine the natural state of the substrates and verify EXOG's preference for substrate conformations, I calculated another set of free energy profiles using the US approach on the isolated DNA duplex (from PDB id: 5T5C) and the R2-DNA/DNA chimeric duplex (PDB id: 6IID) in the absence of ExoG.

Binding Affinity of DHX36 Subdomains for G4

To explore the interplay between DHX36 subdomains in recognizing G4, I conducted a series of US simulations to calculate the relevant free energy profiles. In order to dissociate the complexes of puG4 with DHX36, I employed SMD simulations, pulling puG4 away from the helicase during 500-ns-long runs. Two systems were considered: one with DSM (system (i)) and one without DSM (system (ii)), as described in the previous section (3.3.1). The separation distance (r) between the centers of mass (COM) of the heavy atoms of the guanine core and the Cα atoms of the OB subdomain was used as the reaction coordinate for the SMD simulations. A moving harmonic potential with a force constant of 179.42 and 119.61 kcal/(mol·nm^2) was applied to system (i) and (ii), respectively. I extracted initial configurations for the US simulations from the SMD trajectories. To cover the range of r from 1.7 to 5.0nm, I divided the system into 23 uniformly distributed windows with a separation distance of 0.15 nm. In each window, the system was simulated for 500 ns with

a harmonic potential using a similar force constant to the one applied during the SMD simulations to restrain the system along the reaction coordinate r.

To dissociate the DSM α1 helix from the three complexes it formed with G4 [puG4, ffG4-5′-G-tetrad, and ffG4-3′-G-tetrad; systems (a), (b), and (c), respectively (as describe in section. 3.3.1)], I conducted SMD simulations in which DSM was pulled away from G4 during a 500-ns run. The coordinate defining the separation distance (r) between the COMs of the guanine-core heavy atoms and the Cα atoms of DSM was subjected to a moving harmonic potential with a force constant of 119.61 kcal/(mol·nm^2). From these forced dissociation trajectories, I extracted initial configurations for the US simulations. I used 22 uniformly distributed US windows with a separation of 0.15 nm in the r range of 0.85 to 4.0 nm and simulated each system for 500 ns. A force constant of 56.80 kcal/(mol·nm^2) was applied to restrain the system along the reaction coordinate r. To preserve the DSM helicity in all simulated systems, I applied a force constant of 59.8 kcal/mol·nm^2 to the ALPHARMSD coordinate [229], as the DSM α1 helix has a tendency to spontaneously unfold in the absence of G4.

To determine the free energy profiles, I used the standard weighted histogram analysis method (WHAM) [230] after discarding the first 10% of the trajectories. Uncertainties were estimated using bootstrap error analysis, taking into account the correlation in the analyzed time series. For interaction analysis, I reweighted the original biased umbrella sampling data to recover the unbiased probabilities, using weights of the form $e^{\left(\frac{U_i(r)-F_i}{k_BT}\right)}$, where $U_i(r)$ is the applied biasing potential, F_i is the free energy constant associated with the bias as calculated by the WHAM algorithm in the i-th US window, and k_B is the Boltzmann constant.

3.3.4 Methods of Analysis

Analysis of Base Readout in DNA-Protein Complexes

In order to examine the occurrence frequency of specific amino acids with H-bond-forming side chains at the DNA-protein interface and to verify amino acid-nucleobase contact preferences as part of the direct readout mechanism, an extensive structural dataset was compiled. This dataset included all 4623 protein/DNA complexes available in the PDB database as of 16-Oct-2019. DNAproDB-tool [231] was utilized to obtain information on interfacial interactions such as H-bonds, and custom python scripts were employed to extract relevant statistics.

Determinants of Asp/Glu Residues Contribution to DNA Binding: Regression Analysis

To investigate the relationship between various structural features and the $\Delta\Delta G$ values (described in section 3.3.3) of Asp/Glu residues binding to different DNA sequences, Pearson correlation coefficients were calculated. The features considered were the average number of H-bonds formed by Asp/Glu with two possible partners i.e., cytosine (#Hb-C) and adenine (#Hb-A), as well as the number of different nucleobases (#A, #C, #G, and #T) in the vicinity of Asp/Glu. As the aim was to study the involvement of Asp/Glu in the direct-readout of base functional groups, the latter features (#A, #C, #G, and #T) represent the number of exocyclic functional groups exposed in the major groove (i.e., N4-amino group of C, N6-amino group of A, O6-carbonyl group of G, and O4-carbonyl group of T) within a 5 Å cutoff of

Asp/Glu, averaged over the trajectory. To account for the effect of partial dehydration of Asp/Glu at the interface and competing salt-bridges with neighboring basic residues, the number of H-bonds with water molecules (#Hb-H_2O) and the number of contacts with Arg/Lys (#Arg/Lys) were used as additional features. An H-bond was considered to be formed when the donor-acceptor distance was $<$3.5 Å and the proton-donor-acceptor angle was $<30°$. The specific values of these features can be found in Table. A.1.

To validate the results obtained from Pearson correlation analysis and capture possible non-linear correlations between the structural features and $\Delta\Delta G$ values, a hierarchical random forest-based approach was used. To avoid masking the importance of informative features by correlated features with higher predictive power, a repetitive procedure was employed in which the best predictor of $\Delta\Delta G$ was selected in each iteration and removed from the global pool. In each iteration, an ensemble of random forest regressors was trained for all possible subsets of four features selected from the current global pool, and the best subset was selected based on the statistical significance of the associated model. The features in this subset were then ranked according to their Shapley values (discussed below) in predicting $\Delta\Delta G$. The Shapley value for the top-ranked feature was considered along with its importance, and the feature was removed from the pool. This iterative process was continued until only four features were left in the pool, and the importances of the four features with the highest predictive power were identified in this way.

Feature Importance Using Shapley Values

Assessing the importance of features in the prediction of $\Delta\Delta G$ was performed using Shapley values, which is a game theoretic concept utilized to determine the contribution of each player to the payoff of a cooperative game. In the context of model explainability, input features are considered as the players and model prediction as the payoff [232]. The Shapley value represents the average marginal contribution of a given feature across all possible coalitions of the remaining features, indicating the extent to which a particular feature contributes to the predicted value with respect to the baseline (i.e., average) prediction. Typically, the importance of a feature is determined as the average absolute Shapley value over all data points.

Decomposition of Free Energy in Asp/Glu Contribution to DNA Binding

Systematic decomposition of $\Delta\Delta G$ contributions was performed by leveraging the additive property of force fields. Trajectories were divided into subsystems, each consisting of the acidic residue of interest with (a) individual DNA nucleobases (A, C, G, and T), (b) DNA backbone (BB), (c) solvent (water and ions), and (d) intra-protein regions. Reruns were then executed on each sub-trajectory to calculate the average changes in interaction energy (ΔE) and thus the enthalpic contributions.

Structural characterization of DNA

The X3DNA analysis tool [5] was utilized to determine the structural properties of DNA in this study. Specifically, the dinucleotide steps CpA and GpA from the B-DNA decamers (analyzed in section 3.3.1) were evaluated to confirm their BI/BII population ratios. To accomplish this, the difference between the ϵ and ζ torsion

angles (defined in Fig. 2.3) were computed. Furthermore, the distributions of important helical parameters, including twist, roll, and x-disp, which characterize the BI/BII equilibrium, were also calculated.

Chapter 4

Results and Discussion

4.1 Role of Asp/Glu in Determining DNA Sequence Specificity

A primary focus of my doctoral research (as outlined in objective 1 in Chapter 1) was to investigate the role of acidic amino acid residues at the DNA-protein interface in determining DNA sequence specificity. To achieve this, I first conducted a structural bioinformatic analysis to examine the prevalence of acidic amino acids at the interface relative to other polar/charged amino acid residues. Additionally, I analyzed the interaction preferences between specific amino acid residues and nucleobases, with a particular emphasis on elucidating the molecular mechanisms by which acidic amino acid residues might contribute to DNA sequence specificity, specifically through the direct readout mechanism.

The examination of base-amino acid contact preferences in all high-resolution structures of DNA-protein complexes provides valuable qualitative insights into the energetics of direct readout (Fig. 4.1). As expected, basic amino acids, such as Arg and Lys, which are positively charged at physiological pH, are frequently found interacting with nucleobases in the major groove (Fig.4.1A, *top*). These amino acids provide both the electrostatic attraction to the negatively charged DNA backbone and preferential hydrogen bonding to the guanine base, contributing to the recognition of G-containing sequences (Fig. 4.1A, *bottom*). Polar residues, including Gln, Asn, Ser, and Thr, are also commonly found in the major groove, and they act as both proton donors and acceptors, making them versatile in their interactions with nucleobases. Asn and Gln show quite a strong preference for the adenine base.

Surprisingly, acidic residues, Asp and Glu, are found almost as frequently as other polar residues at the interface (Fig. 4.1A, *top*), despite the unfavorable interaction with the negatively charged DNA backbone. Asp and Glu are found to directly recognize the cytosine base in the major groove via hydrogen bonding with the N4-amino group (Fig. 4.1A, *bottom*, and B). In fact, Fig. 4.1B revealed that cytosine is almost exclusively recognized by Asp/Glu. This preference may explain why Asp and Glu are found frequently in DNA-protein complexes.

However, this raises the question of why adenine, which also exposes its N6-amino group in the major groove (Fig.4.1C), is not recognized by Asp and Glu through hydrogen bonding with the amino group.

In a previous study on the telomeric repeat-binding factor 1 protein (TRF1), it was found that Asp had little net impact on the binding affinity for the target telomeric sequence containing three consecutive cytosines [214]. Instead, the presence of Asp

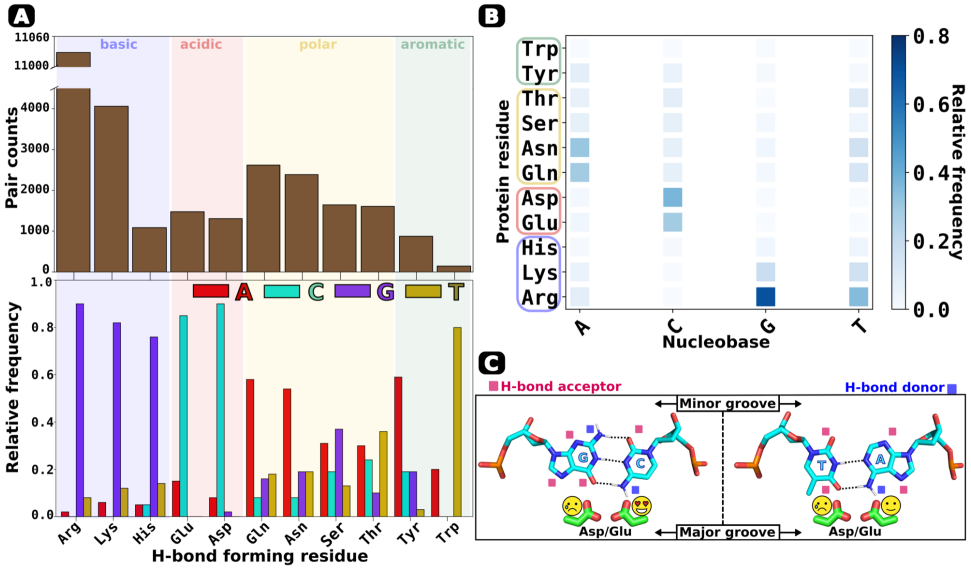

FIGURE 4.1: Amino acid-nucleobase preferences in the DNA major groove were analyzed using data from 4623 DNA-protein structures in the PDB database. Hydrogen bonding and contact information were obtained from the DNAproDB database [231]. **A.** (*top*) The total number of amino acid residues with side chains capable of forming hydrogen bonds within 4.5 Å of any nucleobase in the major groove, across all DNA-protein complexes. (*bottom*) The relative frequencies with which each amino acid side chain forms a hydrogen bond with one of the four nucleobases in the major groove. **B.** The relative frequencies with which each nucleobase forms a hydrogen bond with specific amino acid side chains in the major groove. **C.** Schematic representation of the possible modes of interaction between the acidic residues (Asp/Glu) and the GC or AT base pairs in the major groove.

prevented off-target binding to non-telomeric sequences lacking cytosine. Therefore, it can be hypothesized that acidic residues primarily act as "negative" selectors, sensing the absence of cytosine and preventing off-target binding, unlike basic (and other polar) residues that provide "positive" selection by significantly increasing affinity for cognate sites more than non-cognate sites.

4.1.1 Acidic Amino Acid Residues as Negative Selectors: preventing binding to cytosine-poor sequences

To investigate if the negative selection mechanism is involved in direct readout mediated by acidic amino acid residues, an evaluation of the contributions of these residues to the DNA-binding affinity for a set of five protein/DNA complexes (as mentioned in section 3.3.1, and see Table 4.1) was conducted. The objective was to determine the difference in the DNA-binding free energy, $\Delta\Delta G$, between the wild-type protein and its mutant in which a specific Asp or Glu residue was substituted with alanine (see Fig. 3.2) using "alchemical" transformations. A thermodynamic cycle was used to calculate the contributions, which involved transforming an acidic residue into alanine in the presence or absence of bound DNA. To determine the mutational free energy changes corresponding to these transformations (ΔG_m), Hamiltonian-replica exchange molecular dynamics simulations were performed of the examined protein/DNA complexes. The sequence space was sampled for each

protein	variant	sequence
CLOCK-BMAL (Glu43)	ref	5′-AGGAACACGTGACCC-3′
	C2T	5′-AGGAATACGTGACCC-3′
	C2A	5′-AGGAAAACGTGACCC-3′
	C2G	5′-AGGAAGACGTGACCC-3′
	all-5Å$_1$	5′-AGGACACAGTGACCC-3′
	all-5Å$_2$	5′-AGGATGTGGTGACCC-3′
Zif268 (Asp70)	ref	5′-TACGCCCACGC-3′
	C2T	5′-TACGTCCACGC-3′
	C2A	5′-TACGAACACGC-3′
	C2G	5′-TACGGGCACGC-3′
	all-5Å$_1$	5′-TACATTTACGC-3′
	all-5Å$_2$	5′-TATATTTACGC-3′
Myb (Glu132)	ref	5′-CCTAACTGACA-3′
	C2T	5′-CCTAATTGACA-3′
	C2A	5′-CCTAAATTACA-3′
	C2G	5′-CCTAAGTCACA-3′
	all-5Å$_1$	5′-CCTAATCAACA-3′
ERG3 (Asp363)	ref	5′-CACTTCCGGT-3′
	C2T	5′-CACTTTTGGT-3′
	C2A	5′-CACTTAATTT-3′
	C2G	5′-CACTTGGCCT-3′
	all-5Å$_1$	5′-CACTTTTAAT-3′
TRF1 (Asp422)	ref	5′-TTAGGG-3′
	C2T	5′-CCCTAA-3′

TABLE 4.1: DNA sequences sampled for the free energy simulations.

of the examined complexes by creating DNA variants that substituted either a directly H-bonded cytosine only or all bases within 5 Å of Asp or Glu (see schematic in Fig. 3.4). This approach allowed us to capture the dependence of $\Delta\Delta G$ on local DNA sequence. In total, 24 independent DNA/protein systems were produced (see Table 4.1) using this systematic approach, and the calculated $\Delta\Delta G$ values are presented below in Table 4.2.

According to the $\Delta\Delta G$ values shown in Table 4.2, the contributions of Asp and Glu residues to DNA affinity differ depending on whether they directly interact with cytosine or other nucleobases. The results indicate that they reduce the binding to non-C sequences, with an average $\Delta\Delta G$ of 1.10 ± 0.74 kcal/mol and 0.65 ± 0.36 kcal/mol for A and G/T sequences respectively (see Fig. 4.2), while slightly increasing the affinity for C sites, with an average $\Delta\Delta G$ of -1.41 ± 0.60 kcal/mol. These observations align with the known nucleobase propensities, including the preference for cytosine over adenine (Fig. 4.1). This finding also supports the notion of a negative selection mechanism, whereby the acidic residues prevent the protein from binding to non-C sequences.

Although we have drawn general conclusions, it is important to note that the contribution of Asp or Glu to the affinity of C sequences can vary widely, ranging from negligible to highly favorable. Despite forming a single H-bond with the nearest

	variant	int. base	$\Delta\Delta G$	$\Delta\Delta G_{model}$	residual
CLOCK-BMAL (Glu43)	ref	C	-1.01	-1.28	0.27
	C2T	G/T	1.74	1.57	0.17
	C2A	A	2.13	1.90	0.23
	C2G	G/T	0.80	0.08	0.72
	all-5Å$_1$	C	-4.53	-2.84	1.69
	all-5Å$_2$	C	-2.09	-1.79	0.30
Zif268 (Asp70)	ref	C	-0.95	-1.56	0.61
	C2T	G/T	-1.60	-1.26	0.34
	C2A	A	-1.80	-1.39	0.41
	C2G	G/T	1.41	0.83	0.58
	all-5Å$_1$	C	1.44	-0.36	1.80
	all-5Å$_2$	G/T	1.57	1.25	0.32
Myb (Glu132)	ref	C	-4.15	-2.93	1.22
	C2T	G/T	-0.75	-0.24	0.51
	C2A	A	1.57	1.45	0.12
	C2G	G/T	0.53	0.73	0.20
	all-5Å$_1$	G/T	0.02	-0.42	0.44
ERG3 (Asp363)	ref	C	0.06	-0.41	0.47
	C2T	C	-0.50	0.25	0.75
	C2A	A	0.37	0.58	0.21
	C2G	G/T	2.33	1.85	0.48
	all-5Å$_1$	G/T	0.43	0.58	0.15
TRF1 (Asp422)	ref	C	-1.00	-1.35	0.35
	C2T	A	3.00	1.59	1.41

TABLE 4.2: Contribution of Asp/Glu residues to the binding free energy of studied proteins for various DNA sequence variants, measured as $\Delta\Delta G$. The predicted values were obtained by the best random forest model (see statistical measures in Table. A.2) that used four top-ranked structural features, denoted as $\Delta\Delta G_{model}$. The difference between the predicted and calculated values is shown in the 'residual' column. The DNA bases that Asp/Glu directly interact with in each sequence variant are listed in the 'int. base' column.

cytosine, the contribution can be as low as $\sim$0 kcal/mol or as high as -4 kcal/mol. Unexpectedly, we also found negative $\Delta\Delta G$ values for a few non-C sequences (e.g., C2T and C2A variants of Zif268 in Table 4.2). These observations suggest that the base readout by Asp and Glu is influenced by the local sequence context and other structural characteristics that affect direct interaction.

To better understand the relationship between the simulation-derived $\Delta\Delta G$ contributions and relevant structural features of the DNA-protein complexes, I calculated Pearson correlation coefficients (shown in Fig. 4.3A). The analysis focused on the importance of direct H-bonds and local sequence composition, which were reflected in several features we considered (also described in section 3.3.4). Specifically, I looked at the average number of H-bonds formed by Asp/Glu with possible partners i.e., cytosine (#Hb-C) and adenine (#Hb-A), as well as the number of different

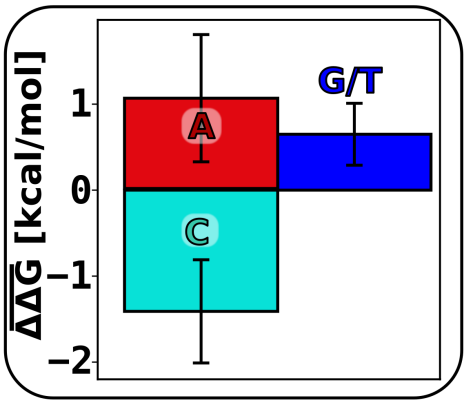

FIGURE 4.2: The $\Delta\Delta G$ values, which were obtained from simulations, were averaged over the DNA sites where Asp/Glu interacted directly with cytosine (C), adenine (A), or the remaining nucleobases (represented by G/T sequences).

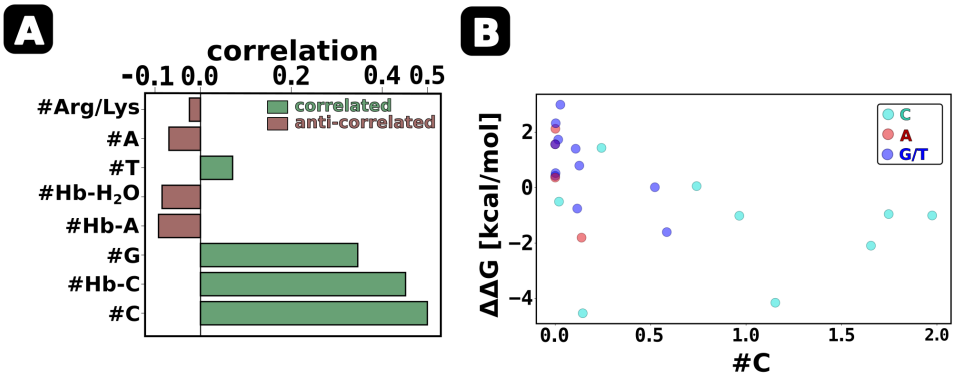

FIGURE 4.3: **A.** The Pearson correlation coefficients between the simulation-derived $\Delta\Delta G$ values and relevant structural features of DNA-protein complexes, as defined in the text. A positive correlation suggests that an increase in the feature's value leads to a more favorable contribution of Asp/Glu to the binding affinity (i.e., a more negative $\Delta\Delta G$ value). **B.** The relationship between the simulation-derived $\Delta\Delta G$ values and the number of cytosine residues in the vicinity of Asp/Glu (#C).

nucleobases (i.e., #A, #C, #G, and #T) in the vicinity of Asp/Glu. To measure the involvement of Asp/Glu in direct readout of base functional groups, the later features (i.e., #A, #C, #G, and #T) were obtained as the number of exposed exocyclic functional groups in the major groove (i.e., N4-amino group of C, N6-amino group of A, O6-carbonyl group of G, and O4-carbonyl group of T) within a 5 Å cutoff of Asp/Glu, averaged over the simulation trajectory. The effect of partial dehydration of Asp/Glu at the interface and competing salt-bridges with neighboring basic residues were also considered by including the number of H-bonds with water molecules (#Hb-H_2O) and the number of contacts with Arg/Lys (#Arg/Lys) as additional features.

Fig. 4.3A shows the linear correlations between the simulation-derived $\Delta\Delta G$ values and the structural features of the DNA-protein complexes. Surprisingly, I found that the number of cytosine residues in the local sequence of Asp/Glu (#C) is an even

better predictor of $\Delta\Delta G$ than the number of H-bonds with cytosine (#Hb-C), indicating that cytosine content plays a significant role in determining the contribution of Asp/Glu to DNA binding affinity (Fig. 4.3B).

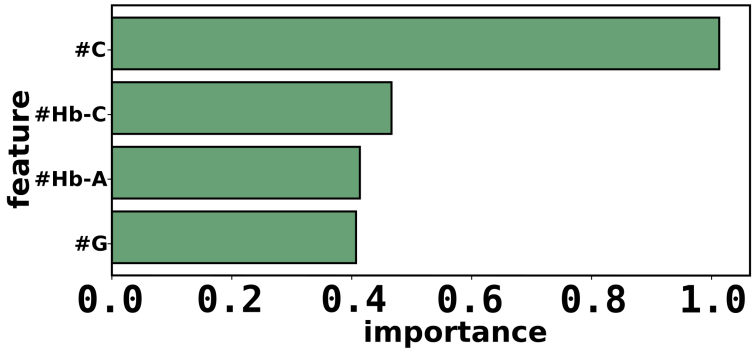

FIGURE 4.4: A hierarchical random forest-based approach (see methods in section 3.3.4) was used to determine the global importance of the top four features in predicting $\Delta\Delta G$. Importance was measured in terms of the mean absolute Shapley value, which represents the average contribution (in units of kcal/mol) of each feature to the predicted $\Delta\Delta G$ compared to the baseline prediction.

FIGURE 4.5: Comparison between the observed number of GC pairs (#GC) within 8 Å of the Asp/Glu residues in the experimentally solved DNA-protein complexes and the number expected if GC and AT pairs were equally probable (ref). The term #base-pair refers to the total number of base pairs within 8 Å of the Asp/Glu residues, regardless of their type. The average number of base pairs found within this distance cutoff in the DNA/protein complexes is 3.2,±,1.3.

To validate the relationship between the number of cytosine residues and the contribution of Asp/Glu to DNA binding affinity, a hierarchical random forest-based approach was employed that accounted for non-linear correlations between $\Delta\Delta G$ and used set of features. Table 4.2 shows the predicted values of $\Delta\Delta G_{model}$ obtained from the best random-forest model, along with their corresponding residuals. The residuals are calculated as the difference between the predicted values and the original simulation-derived $\Delta\Delta G$ values. I used Shapley values to rank the importance of the features (see Methods for details in section 3.3.4). The results, shown in Fig. 4.4,

confirmed that #C is the most important feature for predicting $\Delta\Delta G$. In fact, #C had twice the predictive power of #Hb-C, demonstrating that cytosine recognition by acidic residues is not solely dependent on the formation of a single hydrogen bond but is also influenced by the number of cytosine N4-amino groups in the local vicinity of Asp/Glu (Fig. 4.3B). This finding is consistent with a significant overrepresentation of cytosine-rich sequences among the DNA sites recognized by acidic residues in experimentally solved DNA-protein complexes (Fig. 4.5).

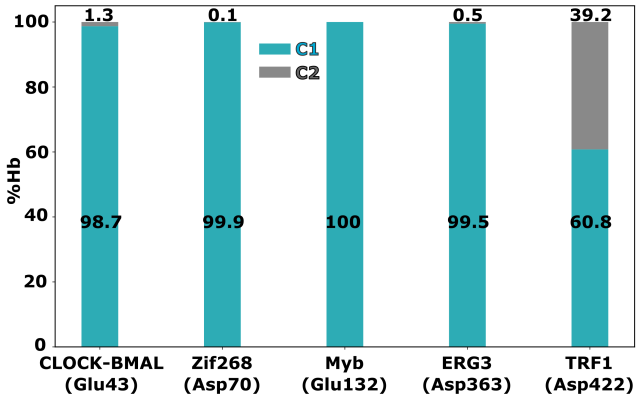

FIGURE 4.6: The percentage of H-bonds formed by Asp/Glu with individual cytosines when there are multiple cytosines in close proximity to Asp/Glu. In such cases, due to steric constraints, Asp/Glu can only form H-bonds with two different cytosines (C1 and C2). The data presented is an average over all multi-cytosine variants for a given protein. The dominant interaction mode for each protein is shown, indicating that except for TRF1, the acidic residues do not have a general tendency to dynamically switch between different cytosine H-bonding partners.

It is possible that the cumulative effect of cytosine on the binding affinity of Asp/Glu is due to either the cooperative effect of multiple hydrogen bonds formed simultaneously with two adjacent N4-amino groups or the ability to switch dynamically between these groups, resulting in a smaller entropic penalty upon DNA binding. However, my results do not support this hypothesis as I found that Asp/Glu can only form one bond with the N4-amino group, even at multi-cytosine sites where more than one cytosine is present in the immediate vicinity of Asp/Glu (see Fig. 4.6). Additionally, my analysis showed that the acidic residues exhibit a single dominant binding mode and do not dynamically switch between different cytosine hydrogen bond donors, except for TRF1 (Fig. 4.6). Therefore, the conclusion is that the cumulative effect of cytosine can be attributed to longer-range attractive electrostatic interactions between Asp/Glu and the N4-amino groups of cytosine, which is discussed further below (section 4.1.2).

Out of all other considered features, only the number of guanines close to Asp/Glu (#G) showed a noticeable correlation with $\Delta\Delta G$ (Fig. 4.3A), which can be attributed to its Watson-Crick pairing with cytosine (see Fig. A.2). Other bases, such as #A and #T, do not seem to have any significant impact on explaining $\Delta\Delta G$, consistent with their known nucleobase propensities (Fig. 4.1). The number of H-bonds with adenine (#Hb-A) was found to be significantly less correlated with $\Delta\Delta G$ than #Hb-C, reflecting the preference for cytosine readout by acidic residues.

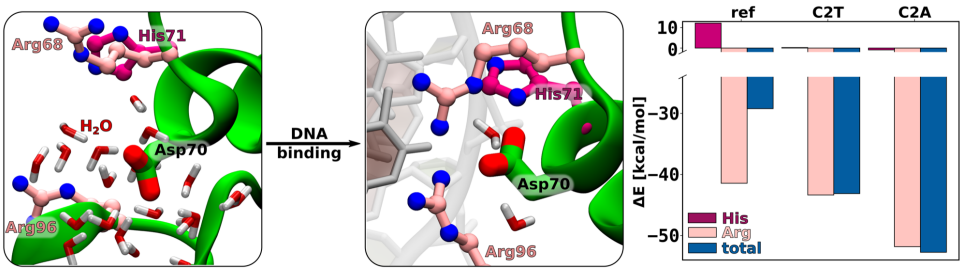

FIGURE 4.7: The network of basic residues around Asp70 in Zif268 (Arg68, Arg96, and His71) undergoes a rearrangement upon DNA binding, resulting in a stronger electrostatic attraction between them. The right panel shows the enthalpic contributions to the $\Delta\Delta G$ values for the ref, C2T, and C2A variants of Zif268 due to the binding-induced change in the interaction energy (ΔE) between Asp70 and the basic residues: Arg68/Arg96 (summed) and His71. The blue bar represents the total change in interaction energy between Asp70 and the basic residues upon DNA binding.

Although most of the other features showed little predictive power over the entire set of DNA-protein complexes, some of them were useful in interpreting the outliers. For example, in the case of the C2T and C2A variants of Zif268, unexpectedly negative $\Delta\Delta G$ values (-1.6 and -1.8,kcal/mol, respectively, see Table 4.2) could be explained in terms of the formation of salt bridges with neighboring basic residues (#Arg/Lys). Notably, for Zif268, we observed that the favorable interaction between the Asp residue and the neighboring basic residues increased significantly upon DNA binding, with this increase being even more pronounced for non-C sequences (Fig. 4.7). These observations suggest that our limited data set might not be sufficient to capture all the subtle factors that affect the contribution of Asp/Glu to DNA affinity.

4.1.2 Importance of Positive Potential Accumulation by Cytosine for Asp/Glu Binding

To explain the cumulative effect of cytosine on the binding free energy of Asp/Glu to DNA, the enthalpic contribution to the obtained $\Delta\Delta G$ values was evaluated and dissected down into interactions between the acidic residue of interest and the other components of the system (Fig. 4.8A). It can be observed that among the four canonical bases, cytosine is the only one that offsets the strong electrostatic repulsion between Asp/Glu and the negatively charged sugar-phosphate backbone, leading to an average enthalpic stabilization of approximately 10 kcal/mol per base present in the immediate vicinity of the acidic residue.

Furthermore, this stabilization increases with the number of cytosine amino groups in the local sequence within 8 Å of Asp/Glu (Fig. 4.8B). The interaction between Asp/Glu and cytosine is particularly strong when three or more N4-amino groups are present in a tract. It should be noted that this interaction energy is averaged over all systems satisfying the cutoff criterion, including those in which Asp/Glu does not form a close contact with the cytosine base. Therefore, the increments associated with each additional cytosine are smaller than the per-base enthalpic contribution calculated for direct Asp/Glu-cytosine interaction in panel A of Fig. 4.8. Therefore, it can be concluded that the observed preference of Asp/Glu for cytosine-rich sites is due to the long-range electrostatic attraction to the exocyclic cytosine amino groups.

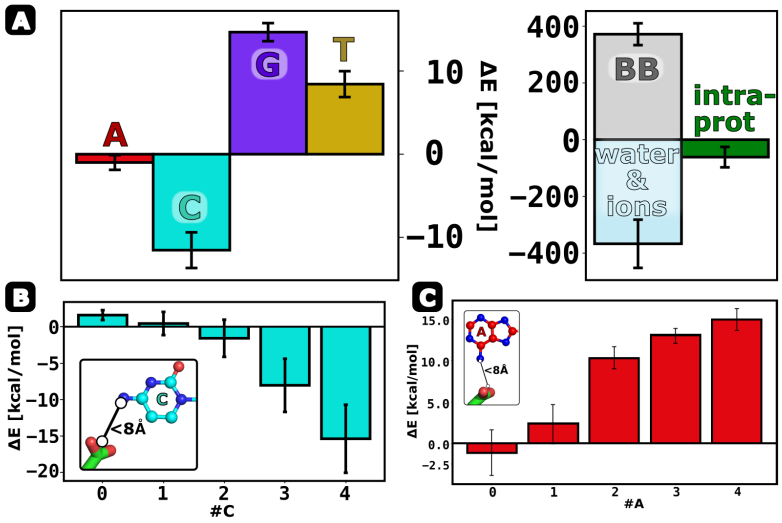

FIGURE 4.8: **A.** Enthalpic contributions to the binding free energy were calculated by determining the average changes in the interaction energy (ΔE) between the acidic residues under examination and other constituents of the system, such as DNA nucleobases (A, C, G, and T), DNA backbone (BB), solvent (water and ions), and the rest of the protein (intra-prot). The contributions from A, C, G, and T were computed based on the average interaction energy with the nucleobase of a given type present within 4.5 Å of Asp/Glu. **B.** and **C.** The interaction energy between Asp/Glu and cytosine, and between Asp/Glu and adenine, respectively, were evaluated as a function of the number of cytosine N4-amino and adenine N6-amino groups located within 8 Å of the acidic residue.

In Fig. 4.8A, it is evident that guanine and thymine, due to the negatively charged carbonyl oxygens exposed to the major groove, significantly decrease the favorable binding of acidic residues (by 15 and 8 kcal/mol, respectively). Notably, despite adenine having an amino group in the major groove, it shows only a negligible attractive interaction with Asp/Glu when it is in direct contact with it (-1 kcal/mol). Also, the dependence of the interaction energy on the number of adenine residues (#A) around Asp/Glu indicates that longer-range electrostatic interactions with adenine are actually net repulsive (Fig. 4.8C).

The enthalpic contributions of the four nucleobases to the binding free energy of acidic residues (Asp/Glu) correlate with their known propensities (Fig. 4.1), which may provide a molecular-level explanation for the negative selection mediated by the acidic residues. In addition to the nucleobases, the interaction of Asp/Glu with the solvent and the rest of the protein also promotes binding to DNA, as shown in Fig. 4.8A, (*right*). The former contribution is due to the accumulation of K^+ counterions at the DNA surface (see Fig. A.3), while the latter originates from the stabilization of salt-bridges between Asp/Glu and abundant basic residues (Arg, Lys) as a result of partial dehydration at the DNA-protein interface (see Fig. 4.7).

4.1.3 Factors contributing to the low affinity of Asp/Glu for adenine

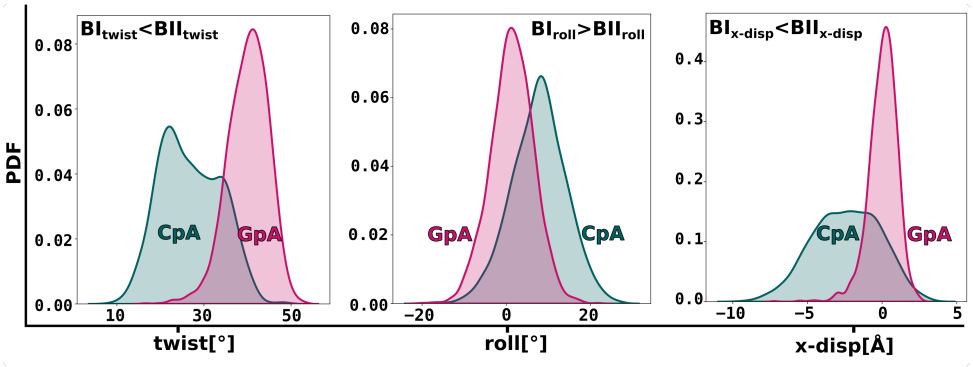

FIGURE 4.9: A comparison of three helical parameters, namely twist, roll, and x-disp, is shown for the dinucleotides CpA and GpA analyzed in my study. These parameters are defined in section 2.1.1 (see Fig. 2.1), and are used to evaluate the intrinsic flexibility of dinucleotide steps. The TRX scale developed by Heddi and Laaksonen [233] utilizes these parameters and characterizes the conformational preference of dinucleotide steps. The BI conformation is associated with higher values of roll and lower values of twist and x-disp compared to the BII conformation.

The main objective of this investigation was to understand the molecular basis for the stronger affinity of acidic residues for cytosine as compared to adenine, despite both bases having an exposed amino group in the major groove (as shown in Fig. 4.1 and Fig. 4.8A). To address this, I conducted simulations aimed to calculate the difference in the binding free energy of Asp/Glu to cytosine and adenine. The simulations were carried out using a model system consisting of a single propionic acid molecule mimicking an acidic amino acid side-chain, competing for interaction with cytosine and adenine on adjacent sites in a canonical B-DNA decamer (see section 3.3.1 for details). To ensure that both nucleobases were equally accessible in the major groove, I used the 5′-GTCAAT-3′ sequence in the middle of the decamer, as shown in Fig. 4.10A.

Previous studies have shown that purine-purine dinucleotide steps tend to favor the BII conformation over the canonical BI phosphate conformation, more so than pyrimidine-pyrimidine and pyrimidine-purine steps [13]. This conformational preference has been implicated in the specificity of protein-DNA interactions [234]. To investigate whether BI/BII conformational dynamics influences the Asp/Glu base preferences, I utilized a different DNA sequence (5′-TGACAT-3′) in which cytosine and adenine are still equally accessible at the center of the decamer, but the GpA dinucleotide step involving adenine is known to favor the BII conformation [13, 215] (see Fig. 4.10A and section 3.3.1 provide additional details).

The systems were prepared and subjected to conventional MD simulations. During the simulations, the propionate anion was held close to the DNA surface using a flat-bottom harmonic potential, and its distribution in the major groove was obtained. After running simulations for each system for approximately 500 ns, the propionate distributions were observed to have reached equilibrium (see Fig. A.4). Based on this distribution, the relative free energy of propionate binding to cytosine and adenine ($\Delta\Delta G_{ca}$) was calculated using the equation $-RT \ln(p_a/p_c)$, where p_a

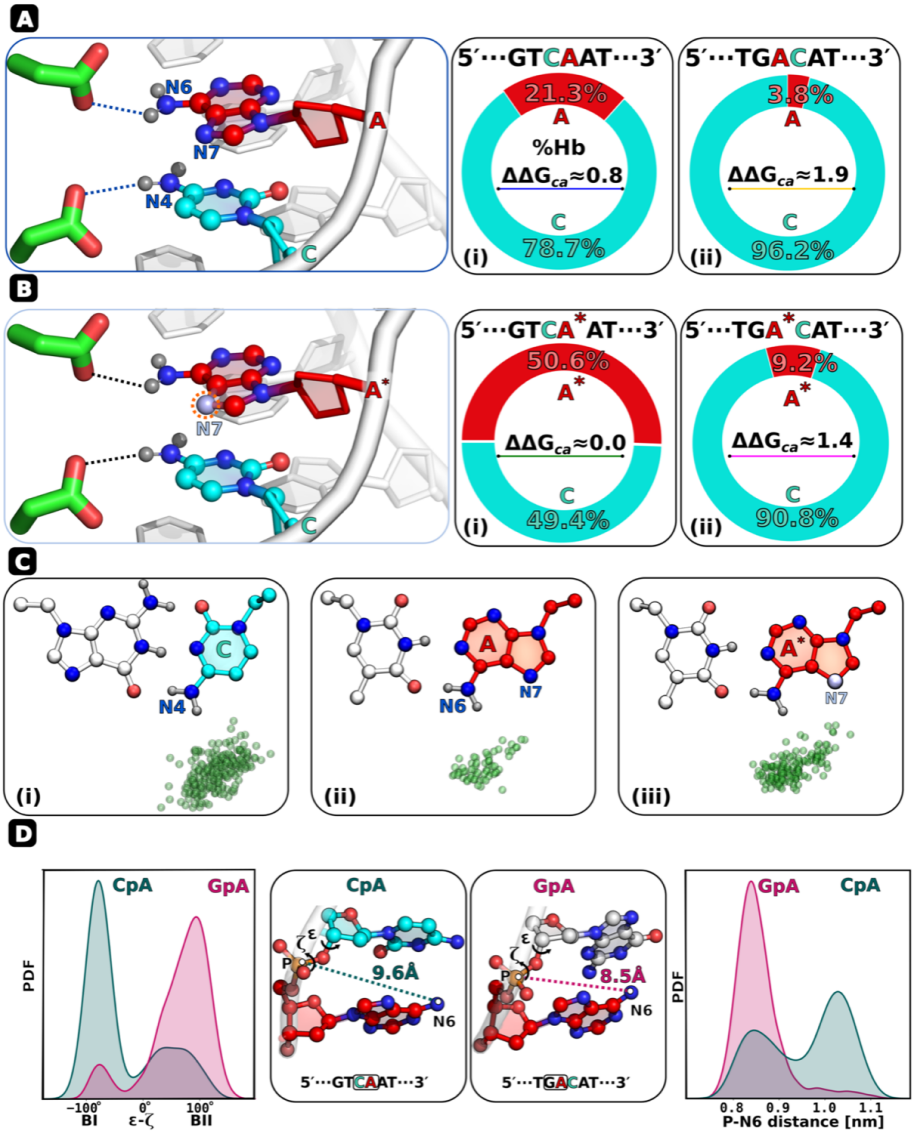

FIGURE 4.10: **A.** Comparison of the percentage of propionate ion hydrogen bonds with the cytosine N4-amino group (cyan) and the adenine N6-amino group (red) in two different sequence contexts, as shown in subpanels (i) and (ii). The corresponding differences in the free energy of propionate binding to cytosine and adenine, $\Delta\Delta G_{ca}$ (in kcal/mol), were calculated using Boltzmann inversion from the equilibrium hydrogen bond populations. **B.** Similar to **A**, but with the negatively charged N7 atom in the adenine imidazole ring made electrically neutral in both sequences (the modified adenine is denoted as A*). **C.** Spatial distribution of the propionate ion hydrogen-bonded to the N4-amino group of cytosine (C; subpanel i), N6-amino group of adenine (A; subpanel ii) and modified adenine (A*; subpanel iii). Green spheres represent the carboxylic carbon atoms, with their numbers being proportional to the equilibrium H-bond populations. **D.** (*left*) BI/BII population ratio for the CpA and GpA dinucleotide steps containing the central adenine in both studied sequences, shown as the distributions of the difference between ϵ and ζ torsion angles (for definitions, see Fig 2.3). (*right*) Distribution of the distance between the phosphate group and the adenine amino group (P–N6) for the two respective dinucleotide steps. Structural representations of the two steps are shown in the middle panel along with the average P–N6 distances. The distributions of relevant helical parameters characterizing the BI/BII equilibrium (twist, roll, and x-disp) are shown in Fig 4.9.

and p_c represent the probabilities of forming an H-bond with the central adenine and cytosine, respectively.

In Fig. 4.10A, the free energy of propionate binding to cytosine was found to be more favorable than to adenine, which is consistent with our previous structural data (Fig. 4.1) and interaction analysis (Fig. 4.8A). Interestingly, the second sequence showed a stronger preference for cytosine binding (Fig. 4.10B), which may be attributed to the tendency of adenine to adopt the BII conformation, which is supported by the so-called TRX scale of helical parameters shown in Fig. 4.9.

To further investigate the origin of these differences, the spatial distributions of propionate around its binding partners were compared as shown in Fig. 4.10C. The propionate was found to avoid close contact with the negatively charged N7 nitrogen in the adenine imidazole ring, which suggests that repulsive interactions between propionate and N7 might be a major factor responsible for the lower affinity of propionate for adenine.

To test this, the partial charge of the adenine N7 was neutralized and the $\Delta\Delta G_{ca}$ values were recalculated for both sequences using the same MD approach (see Methods section 3.3.1 for details). As shown in Fig. 4.10B, neutralizing N7 resulted in equally probable hydrogen bonding with cytosine and adenine in the first sequence context, confirming the repulsion from the imidazole ring's role in the preference for cytosine. However, in the sequence with the inverted BI/BII ratio, although $\Delta\Delta G_{ca}$ decreased by 0.5 to 1.4 kcal/mol, the interaction with cytosine remained preferred (subpanel (ii) in Fig. 4.10B). This may be due to the disfavored propionate-adenine H-bond caused by the GpA step phosphate group, which approaches the N6-amino group of adenine by almost 2 Å compared to the BI conformation (Fig. 4.10D).

The effect of the phosphate conformation on base preferences was further explored by applying external restraints to change the GpA* step conformation in the 5′-TGA*CAT-3′ sequence from BII to BI, as shown in Fig. 4.11A. This resulted in a further decrease in $\Delta\Delta G_{ca}$ from 1.4 to 0.5 kcal/mol, indicating that BII contributes 0.9 kcal/mol to the preference for cytosine in this particular sequence. Similarly, when the CpA* step in the 5′-GTCA*AT-3′ sequence was restrained to the BII conformation, the preference for cytosine increased by 0.4 kcal/mol, demonstrating that the effect of phosphate conformation on base preferences is sequence-context dependent (Fig. 4.11B).

This finding suggests that both direct base sensing and indirect effects, which rely on sequence-dependent polymorphism, are interdependent and work together to fine-tune the affinity for a particular binding site.

4.1.4 Confirmation of Asp/Glu preference for cytosine over adenine using quantum chemical calculations

To overcome the limitations of simple force field-based models that do not consider electronic polarization and may provide inadequate description of H-bonding interactions, my colleagues (J. Słabońska and S. Sappati) and I collaborated to investigate the basis of base preferences of acidic residues through quantum chemical calculations. Specifically, we utilized hybrid quantum/classical (QM/MM) ab initio molecular dynamics (AIMD) coupled with umbrella sampling to compute the free energy profiles for the binding of a single propionate ion, which serves as a proxy for the side chain of Asp/Glu, to cytosine or adenine in the major groove of a B-DNA decamer (detailed in the Methods section 3.3.1). The QM region was described using

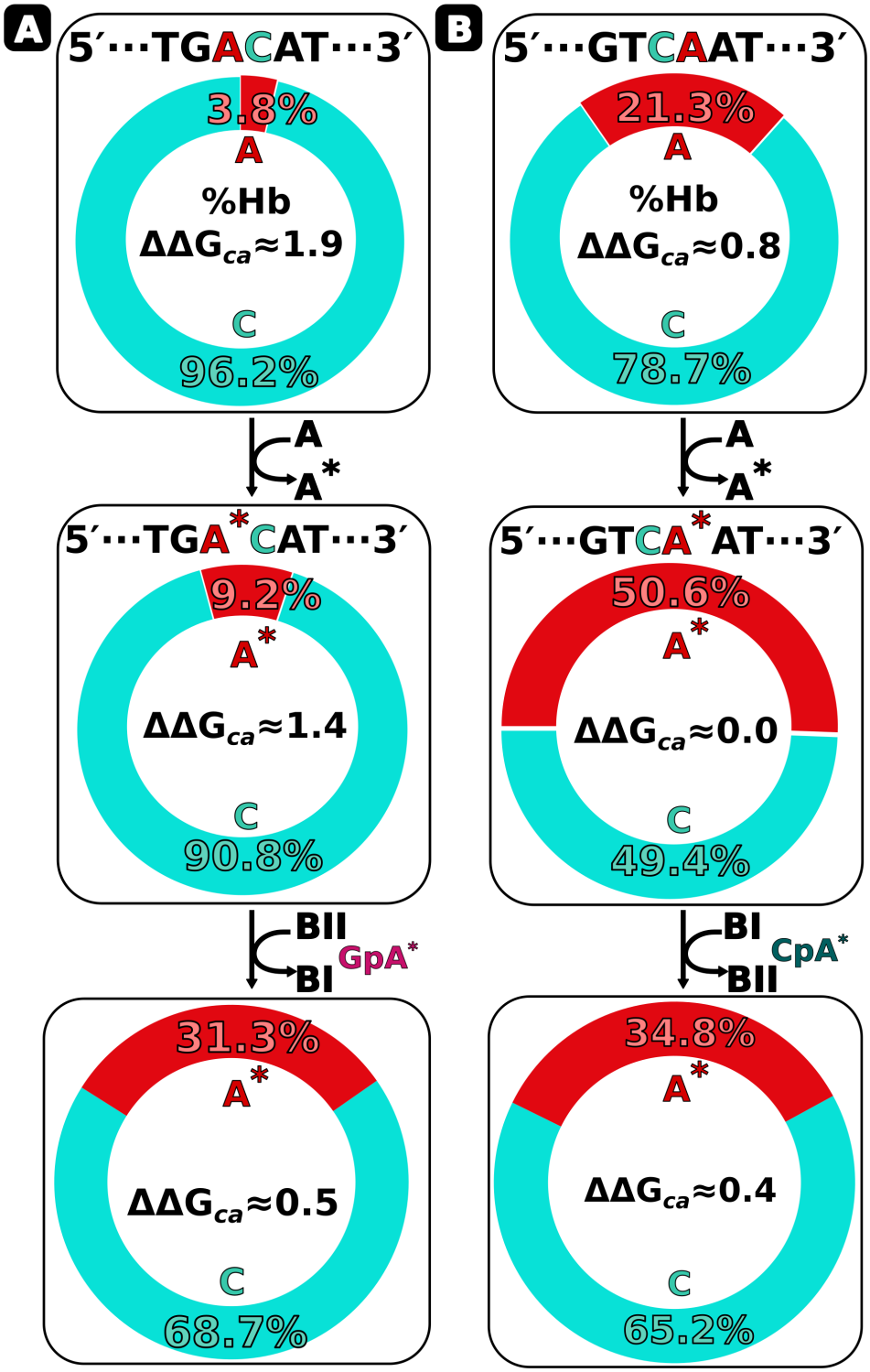

FIGURE 4.11: The differences in free energy of propionate binding, $\Delta\Delta G_{ca}$ (in kcal/mol), for cytosine and adenine in two different sequences, **A.** and **B.**. The computations were carried out three times: i) for the original DNA duplexes (*top*), ii) after neutralizing the adenine N7 atom in the original sequences (i.e., the modified adenine, labeled as A*; *middle*), and iii) after changing the conformation of the GpA* step in 5′-TGA*CAT-3′ from BII to BI and of the CpA* step in 5′-GTCA*AT-3′ from BI to BII (*bottom*). The different conformations were enforced by applying a harmonic potential with a spring constant of 500 kJ/(mol·rad^2) to the difference in dihedral angles ($\epsilon - \zeta$).

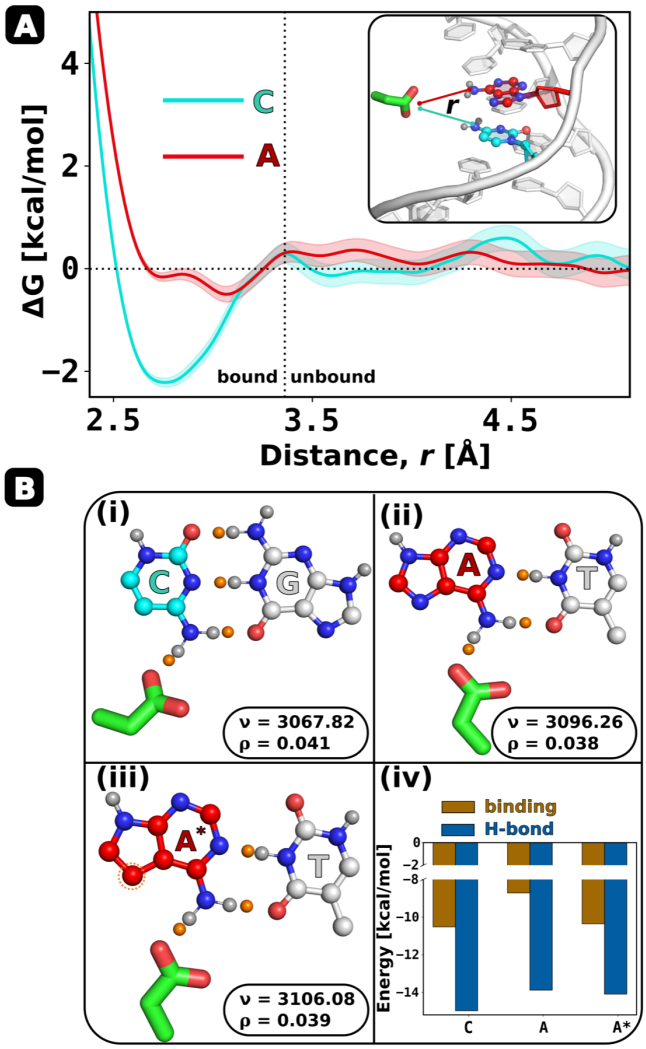

FIGURE 4.12: **A.** The free energy profiles obtained from QM/MM ab-initio molecular dynamics simulations for the binding of propionic acid to either cytosine (C) or adenine (A) in the major groove of a B-DNA decamer. **B.** DFT-optimized structures of the propionic acid complexes with three different base pairs are shown: GC (i), AT (ii), and A*T (iii) (where A* refers to 7-deazaadenine, which is an adenine analog with a -CH group replacing N7 of the imidazole ring). The hydrogen bond critical points are denoted by orange spheres, and the electron densities (ρ in e/bohr^3) at these points as well as the N–H stretching vibrational frequencies (ν in cm^{-1}) that characterize the H-bonds to the propionate ion, are displayed in the insets next to the structures. (iv) Hydrogen bond energies and binding energies (in aqueous solution) for each of the complexes were calculated at the B3LYP/def2TZVP level.

density functional theory (DFT) at the TPSS/def2SVP level and consisted of the propionate, the central base pair with cytosine or adenine, two adjacent bases above and below the selected base, and a number of water molecules between the propionate and DNA (see Fig. A.5). The distance between the propionate oxygen atom and either the nitrogen N4 in cytosine or N6 in adenine was used as the reaction coordinate, r (Fig. 4.12A).

The free energy profile generated from the quantum chemical calculations (Fig. 4.12A) shows a well-defined bound-state minimum (at r <3.3Å) of 2 kcal/mol for cytosine and a shallow minimum for adenine, which supports the observation that cytosine is preferred over adenine by acidic residues (Fig. 4.1 and 4.10). To explain this preference, simpler subsystems comprising the propionate bound to either GC or AT base pair were extracted from the bound-state ensembles obtained from AIMD simulations, and they were optimized using the B3LYP/def2TZVP model chemistry (Fig. 4.12B). Using the continuum solvation model for water, we calculated the binding energies for propionate binding to cytosine in the GC pair and to A in the AT pair. The results (Fig. 4.12B, subpanel iv) showed that the binding of propionate to cytosine in the GC pair is 1.8 kcal/mol more energetically favorable than to A in the AT pair, which is consistent with our previous findings (Fig. 4.10).

Nevertheless, the observed difference in the binding energies of propionate to cytosine and adenine cannot be fully accounted for by the strength of the hydrogen bonds formed between the propionate and the amino groups of the two bases. The hydrogen bond energies estimated from the electron densities at the H-bond critical points are only able to explain about half of the observed difference, as shown in Fig. 4.12B. A similar conclusion can be drawn from the red shifts of the N–H stretching modes upon complex formation with the propionate, which indicate the strength of the hydrogen bonds. The red shifts are quite similar for cytosine and adenine (336.6 and 209.9cm^{-1}, respectively), further supporting our prediction (Fig. 4.10) that repulsion from the imidazole ring is also a contributing factor to the lower affinity of adenine.

In order to further investigate the prediction of the destabilizing effect of repulsion from the imidazole ring on adenine binding affinity, we conducted an additional simulation. Specifically, we replaced adenine in the AT base pair with 7-deazaadenine, which has a -CH group substituted for the N7 atom in the imidazole ring. We observed a significant increase in propionate binding energy after this modification, almost reaching the value computed for cytosine. Furthermore, the hydrogen bond energy remained almost unchanged compared to adenine (Fig. 4.12B), providing further support for our hypothesis that unfavorable interaction with N7 is a primary factor contributing to lower binding affinity for adenine.

4.2 Mechanism of conformational transition induced by EXOG and its preference for A-DNA

In the other significant aspect of my doctoral research, which corresponds to objective 2 discussed in Chapter 1, I delved into the exploration of EXOG and its ability to recognize subtle variations in DNA conformation. Specifically, I directed my efforts towards investigating the intricate molecular details within the substrate binding region, particularly focusing on the first 2–3 nucleotides adjacent to the 5′-blunt-end (see Fig. 2.10). By closely examining this region, I aimed to unravel the mechanisms by which EXOG achieves conformational recognition and, more intriguingly, to elucidate the process through which EXOG induces a transition from the canonical B-DNA conformation to the A-DNA conformation when bound to DNA/DNA duplexes. This investigation was crucial in gaining a deeper understanding of the EXOG's pivotal role in mitochondrial DNA processing and replication.

Initially, to determine the ideal conformation of the first 2–3 nucleotides in the DNA/DNA duplex and RNA-DNA chimeric duplex (R2-DNA/DNA, where the first two nucleotides are from RNA) substrates for recognition by EXOG, I employed the umbrella sampling approach to calculate the conformational free energy profile for the isolated (EXOG-unbound) substrates. The reaction coordinate used in this calculation was the center-of-mass (COM) distance between the phosphate group of the 2nd and 3rd nucleotides, as shown in the inset of Fig. 4.13. This specific reaction coordinate was chosen because it captures very well the conformational transition between B and A DNA forms. The distance between these two phosphate groups is greater than 6 Å in B-DNA, whereas it is shortened to around 5 Å in A-DNA.

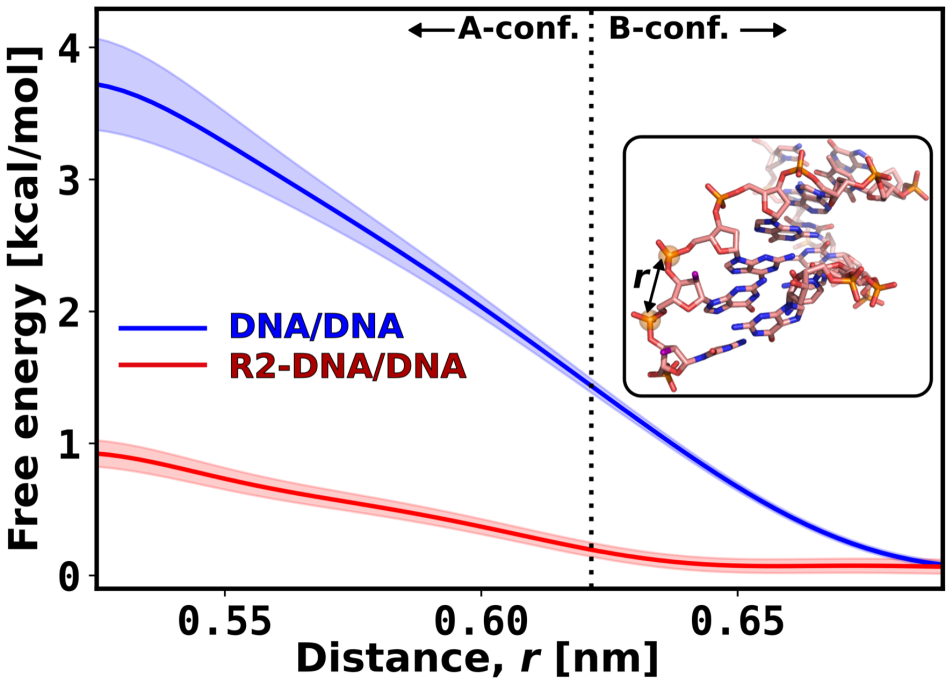

FIGURE 4.13: Free energy profiles for the A to B conformational transition of unbound DNA/DNA (blue) and R2-DNA/DNA (red) duplexes in the absence of EXOG. The free energy profiles were aligned such that the minimum energy state of the B-conformation (B-conf.) corresponds to 0 kcal/mol.

The steep free energy curve observed for the DNA/DNA duplex (Fig. 4.13) suggests that the B-conformation is the only preferred state for this duplex substrate, and that the transition to the A-conformational state is energetically costly (approximately 3.5 kcal/mol). In contrast, although the B-conformation is still preferred for the R2-DNA/DNA chimeric duplex, the flat free energy profile (Fig.4.13) indicates a negligible thermodynamic cost (approximately 1 kcal/mol) for the transition from the B to A state.

Based on these results, it can be concluded that when EXOG is bound to DNA/DNA duplex, it needs to do extra work to induce the transition and convert it to the preferred A-conformation. However, for the R2-DNA/DNA chimeric duplex, almost no extra work is required as it can easily transit between the two conformational states. These findings provide insight into the role of EXOG in 5′-end processing in mitochondrial replication, where it cleaves the dinucleotide of RNA left by RNAaseH1, as discussed in section 2.3.2.

4.2.1 Insight into EXOG's A-DNA conformation preference from free energy simulations

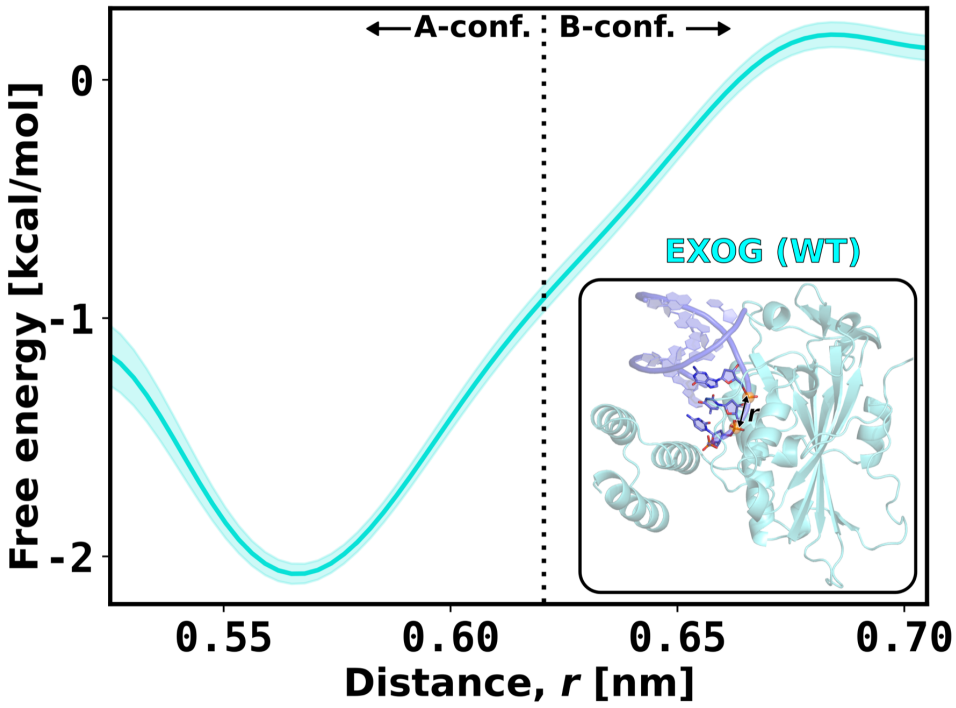

FIGURE 4.14: Free energy profile of the A to B conformational transition of wild-type EXOG. The free energy profiles were aligned by setting the minimum energy state of the B-conformation (B-conf.) to 0 kcal/mol.

Given the indirect insight provided by the available structural data [17, 183], I undertook a comprehensive investigation into the mechanism underlying the recognition of A-DNA conformation by EXOG. In this regard, I initiated my study by exploring the preferred state of the 5′-end region of EXOG bound DNA/DNA duplex and the transition cost to the alternative conformational state. Using the same umbrella sampling approach as described above for isolated substrates, I computed the free energy profile for the B-to-A transition when DNA is bound to EXOG, with a

reaction coordinate of COM distance between the phosphate groups of the 2nd and 3rd nucleotides.

The free energy profile computed (Fig. 4.14) shows a distinct, deep minimum at approximately 0.57 nm, indicating a tightly bound state of EXOG to the A-conformation of the DNA/DNA duplex substrate. However, the transition to the alternative B-conformational state incurs a penalty of approximately 2 kcal/mol. These results support the notion that EXOG preferentially recognizes and binds to the A-DNA conformation, which is consistent with its role in 5′-end processing in mitochondrial replication [183]. The observed difference in binding energy (i.e., 2 kcal/mol) between the A- and B-DNA conformation states highlights the importance of substrate conformation in the recognition and binding of EXOG to DNA, providing further insight into the mechanism of action of this protein.

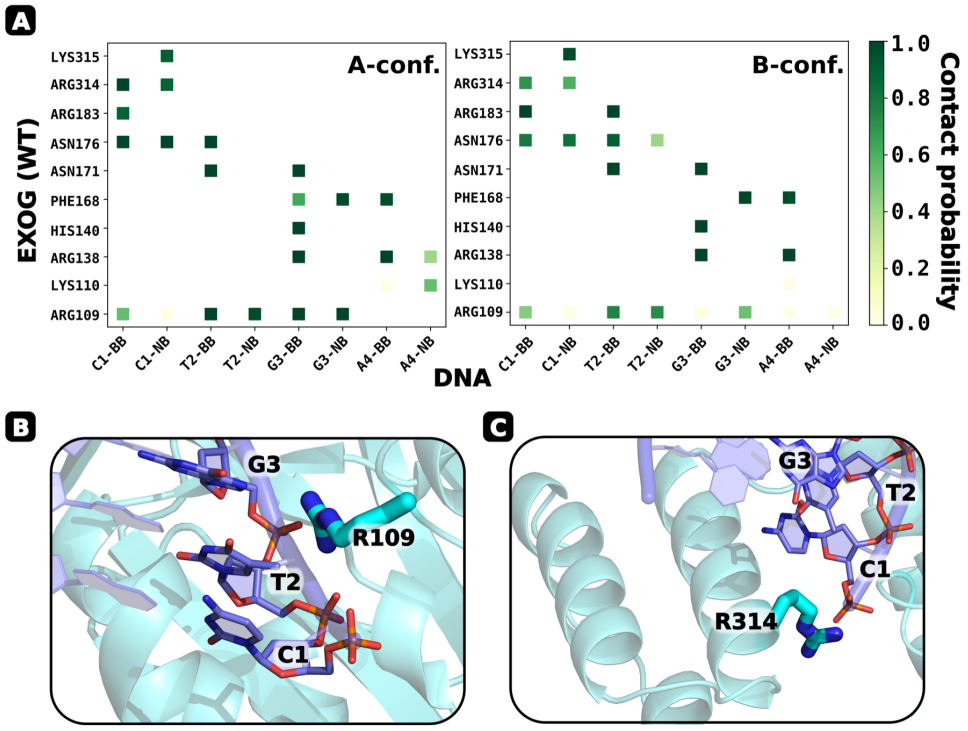

FIGURE 4.15: **A.** Comparison of equilibrium probabilities of residue-residue contacts (minimum distance <0.4 nm) across the interface of EXOG/DNA complex for A- and B-conformational states. **B.** Representative structure of EXOG highlighting the interactions of Arg109 from the core domain and DNA. **(C.)** Representative structure of EXOG highlighting the interactions of Arg314 from the wing domain and DNA.

In order to gain a deeper understanding of the molecular mechanisms underlying EXOG's preference for A-DNA conformation and to identify the interfacial residues that may be involved in inducing the transition from the native B-conformational state of DNA/DNA duplex (as observed in Fig. 4.13) to the A-conformation, a detailed analysis was conducted on the interactions between the substrate binding region of EXOG and bound DNA. Specifically, I calculated the probabilities of forming contacts between these regions, using a minimal distance cutoff of 0.4 nm. To obtain

an unbiased equilibrium distribution, the original biased umbrella sampling data were reweighted, as detailed in the Methods section (3.3.3).

The contact matrices shown in Fig. 4.15A demonstrate that the substrate binding region of EXOG is predominantly composed of polar residues that form strong and stable contacts with the nucleobase (NB) and/or sugar-phosphate moieties (BB) of the bound DNA. These matrices also reveal that the catalytic residue, His140, and other residues that are known to play a role in the catalytic process [17] form robust contacts with DNA, as expected. However, when the 5′-end region of DNA adopts the B-conformation (Fig. 4.15A, *right*), the calculated equilibrium contact probabilities are generally lower than those for the A-conformation state (Fig. 4.15A, *left*). This systematic decrease in contact stabilities implies that the preference of EXOG for A-conformation results from more favorable interactions with DNA.

From a detailed inspection of the contact matrices, I observed a notable difference in the contact pattern of Arg109 between A- and B-conformation states. The contact matrices in Fig. 4.15A show that Arg109 forms stable contacts with the sugar-phosphate moiety of the 2nd and 3rd nucleotides (thymine and guanine, respectively) when DNA adopts the A-conformation state. However, when the DNA adopts the B-conformation state, these contacts were highly destabilized. This suggests that the observed preference of EXOG for A-conformation can be attributed to the more favorable interactions of Arg109 with the DNA.

The interaction mode of Arg109 (Fig. 4.15B) provides a plausible mechanism for its potential role in inducing the transition from B-to-A conformation. Specifically, when EXOG is bound to a B-DNA substrate, Arg109 can attract the negatively charged sugar-phosphate backbone due to its positive electrical charge. This attraction can result in altered sugar-puckering and subsequent backbone conformation, ultimately leading to the induction of the A-DNA conformation. Notably, Arg has been known to exhibit this behavior in previous studies [235, 236]. Therefore, the significant difference in the contact pattern of Arg109 between A- and B-conformation states suggests that Arg109 can be a potential candidate for inducing the B-to-A transition in DNA.

It has been previously reported that EXOG has structural differences compared to its paralog EndoG, which lacks a wing domain and exhibits substrate conformation non-specificity [182]. Therefore, it is reasonable to hypothesize that the wing domain of EXOG may play a crucial role in the recognition of and transition to A-DNA conformation. The contact matrices presented in Fig. 4.15A provide additional insights into this hypothesis, as they reveal that, among the wing domain residues (residues 300–356), only Arg314 has significant differences in the stability of contacts between A- and B-conformation states. Notably, when the DNA substrate is in B-conformation, there is a pronounced destabilization of Arg314's contacts. Consistent with previous biochemical studies [17], the interaction mode of Arg314 (Fig. 4.15C) suggests that it can tightly bind the 5′-blunt-end of the substrate DNA and thus may play a vital role in recognizing the A-DNA conformation.

4.2.2 Role of Arg109 and Arg314 in inducing B-to-A conformational transition of DNA/DNA duplex by EXOG

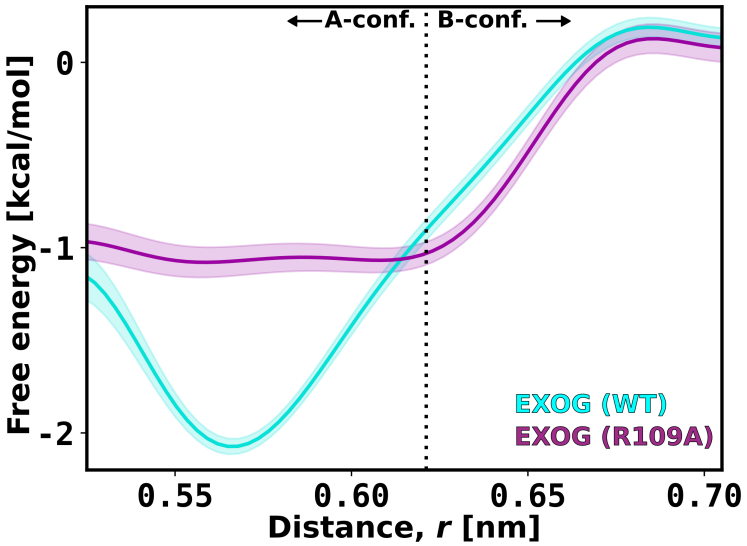

FIGURE 4.16: Comparison of free energy profiles of A to B conformational transition in the 5-′-end region of DNA/DNA duplex, in the presence of wild-type (cyan) and arginine mutant (R109A, magenta) EXOG.

Prompted by the above analysis, I sought to investigate the role of Arg109 in inducing the B-to-A conformational transition of the DNA/DNA duplex. To achieve this, I introduced a mutation in EXOG where Arg109 was replaced with alanine (R109A) and compared the free energy profiles of the wild-type and R109A variant for the transition from A- to B-DNA conformation in the 5′-end region. To calculate the free energy profiles, I utilized the same reaction coordinate and umbrella sampling approach as described previously.

The free energy profiles presented in Fig. 4.16 demonstrate that, similar to the wild-type EXOG, R109A-EXOG exhibits a preference for the A-conformation, as evidenced by the flat global minimum at distances < 0.6 nm. However, there is a significant difference in the free energy difference between the A and B-conformation states for the wild-type and R109A-EXOG. While the free energy difference for the wild-type is around 2 kcal/mol, it is reduced by half to approximately 1 kcal/mol for R109A, indicating that the transition to the alternative B-conformation state is less energetically costly for R109A.

Therefore, these results suggest that Arg109 plays a crucial role in inducing the B-to-A conformational transition when EXOG is bound to a DNA/DNA duplex, as EXOG also participates in base-excision repair [17, 237]. Nonetheless, it is worth noting that Arg109 is not the sole factor contributing to this conformational transition, and Arg314 from the wing domain may also play a role in this process, as discussed earlier.

Given that the wing domain plays a critical role in substrate conformation specificity for EXOG [17, 182], I investigated the contribution of Arg314 in the B-to-A conformational transition. This was motivated by the contact analysis in Fig. 4.15, which shows a notable difference in the contact stability of wing domain residue Arg314 between A- and B-conformation states of the DNA/DNA duplex. To accomplish this, I generated an R314A variant of EXOG and utilized the same umbrella sampling approach to calculate the free energy profile.

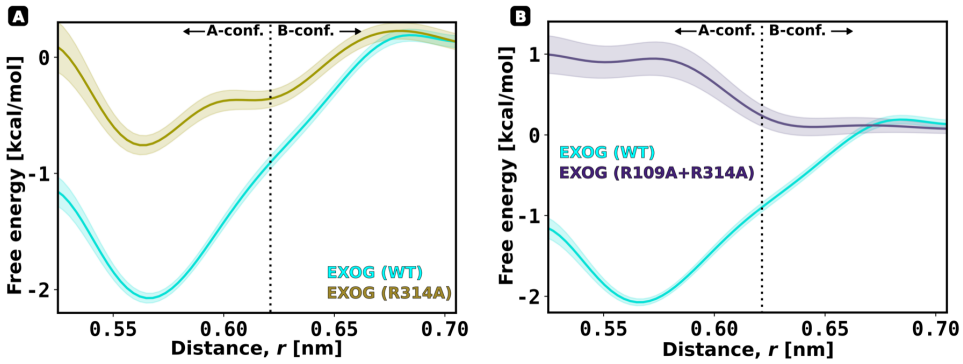

FIGURE 4.17: **A.** Comparison of the free energy profiles for the A to B conformational transition in the 5′-end region of DNA/DNA duplex in the presence of wild-type EXOG (cyan) and R314A mutant EXOG (yellow), and (**B.**) for wild-type and double mutant (R109A and R314A) EXOG.

The free energy profile comparison between the R314A variant of EXOG and the wild-type (Fig. 4.17A) revealed a similar trend to that observed for the R109A variant (Fig. 4.16). The A-conformation is still favored, as evidenced by the global minimum at approximately 0.56 nm. However, the energy required for the transition to the B-conformation state is reduced by approximately a half (i.e., $\sim$ 1 kcal/mol) compared to wild-type (around 2 kcal/mol), similar to the R109A variant.

To investigate whether Arg109 from the core and Arg314 from the wing domain act cooperatively or additively, a double mutant variant of EXOG was created where both arginines were substituted by alanine. The free energy profile for this variant was calculated and compared with the wild-type (Fig. 4.17B). The result shows that for this double mutant (R109A+R314A) variant of EXOG, the B-conformation is now more preferable (by $\sim$ 1 kcal/mol) than the A-conformation, as indicated by the global shallow minimum around 0.65 nm. This finding confirms that both arginines from their respective domains collectively perform the B-to-A conformational transition.

These results suggest that Arg109 from the core and Arg314 from the wing domain act cooperatively. In the wild-type EXOG, the preference for the A-conformation is approximately 2 kcal/mol. However, for the single mutants (R109A and R314A), the preference for A-conformation is reduced to approximately 1 kcal/mol each. Remarkably, in the case of the double mutant, the preference for the B-conformation is approximately 1 kcal/mol, instead of an equal preference for A and B conformations.

The possible explanation for this cooperative effect is that Arg109 might induce the alteration of sugar-puckering by making intimate contact with the sugar-phosphate moieties of the 2nd and 3rd nucleotides, as shown in Fig. 4.15A and B. On the other hand, Arg314 generally interacts with the 5′-blunt-end and thus might assist in widening the minor groove, as shown in Fig. 4.15A and C.

4.3 Recognition Mechanism of Parallel G-Quadruplexes by DHX36 Helicase

G-quadruplexes have gained significant attention in recent years due to their prevalent occurrence in genes and regulatory functions. A genome-wide detection in living cells has revealed that these structures are present in over 60% of gene promoters, especially at the transcription start site, and in approximately 70% of genes [55]. Despite their abundance, the molecular mechanisms underlying the recognition of these unique structures by specialized helicases such as DHX36 remain poorly understood. Therefore, in objective 3 (discussed in Chapter 1) of my doctoral research, I aimed to investigate the molecular mechanisms involved in the recognition of G-quadruplex structures by DHX36 helicase. This objective extended beyond the scope of sequence-specificity and the recognition of subtle conformational changes in conventional DNA double helices, as discussed in the preceding section (Section 4.2).

The crystal structure of DHX36/G4 complex [194] provided limited insights into the recognition of G4 structures by DHX36 helicase. This structure only captured a partially destabilized G4 with a 5′-tetrad composed of two guanines, adenine, and thymine (G·G·A·T) (Fig. 3.5A), which did not fully address the question of how DHX36 recognizes and unfolds DNA and RNA G4s with high specificity. Furthermore, the crystal structure lacked two critical fragments that are located in the proximity of the G4 binding interface, a 20-residue linker connecting DHX36-specific motif (DSM) and RecA1, and a 13-residue loop in RecA2 (highlighted in Fig. A.6; also, see Fig. 2.11), which are essential for interactions with G4 at the binding interface. Therefore, this structure is incomplete in providing a comprehensive understanding of the recognition process of G4s by DHX36.

To address these gaps in knowledge, I employed all-atom molecular dynamics simulations to investigate the recognition of parallel-type G-quadruplexes by DHX36. I used a set of model systems (described in section 3.3.1) based on the DHX36/G4 crystal structure and calculated the relative contributions of DSM and OB (see Fig. 2.11) recognition subdomains to the binding free energy of DHX36 to G4s. This allowed me to gain a better understanding of the specific G4 structural motifs recognized by DHX36 and the molecular basis for its high specificity in binding G4s with parallel topology.

4.3.1 The role of DSM in recognizing parallel G-quadruplexes

To investigate the role of the DSM as the primary recognition motif of DHX36 in G4 recognition and to understand how the base composition of the G4 plane affects its structure and recognition process by the DSM, I examined the effect of a helicase-mediated one-nucleotide register shift in the G-quadruplex on the binding of DSM. To this end, I computed free energy profiles for the binding of the isolated DSM α1 helix (residues Pro57–Lys78) to the G-quadruplex from the $c-myc$ promoter (DNA^{myc}-G4) in both its native fully folded state (ffG4; Fig. 3.5B) [41] and the register-shifted partially unfolded state extracted from the co-crystallized complex with DHX36 [194] (puG4; Fig.3.5A), using the umbrella sampling approach. The center-of-mass distance between the DSM α1 helix and the guanine cores of the respective G-quadruplexes was used as a reaction coordinate (see Fig. 4.18 *inset*). To maintain the extended α-helical conformation of DSM observed in the DHX36/G4 complex, the helicity of DSM during the free energy calculations was restrained, as detailed in the Methods section 3.3.3.

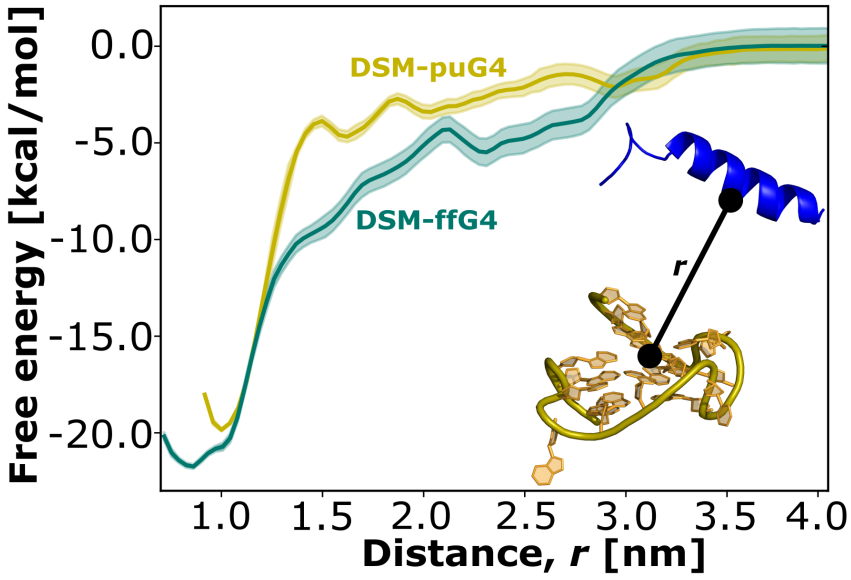

FIGURE 4.18: Binding free energy profiles of the DSM α1 helix to partially unfolded (puG4, in yellow) and fully folded (ffG4, in cyan) states of parallel G4.

The free energy curves shown in Fig. 4.18 exhibit a well-defined minimum with a depth of approximately 20 kcal/mol, which corresponds to the tightly bound state of DSM at the G-quadruplex surface (Fig. 4.19A and B). This deep minimum corresponding to the strongly bound state is consistent with the high affinity of DSM for parallel G4s, as previously measured using quantitative gel electrophoresis[194, 195]. These findings support the view that DSM plays a primary role in the recognition process[238]. However, it is important to note that the accuracy of my results is limited by force field inaccuracies and undersampling of partially bound and unbound states in the umbrella sampling approach. Therefore, the obtained free energy profiles should be interpreted with caution, particularly with respect to the absolute binding free energies, which may be overestimated for DHX36/G4 complexes. Comparative analysis is therefore recommended.

Importantly, the results presented in Fig. 4.18 demonstrate that the strength of DSM binding to G4 is not significantly influenced by the base composition of the outer tetrad, as the binding free energies for ffG4 and puG4 are predicted to be similar (with less than 1.5 kcal/mol difference). This suggests that DSM is well-suited for recognizing all-parallel G4 structures that have easily accessible flat surfaces. This observation is consistent with previous findings by Srinivasan et al. [197], who reported that DSM exhibits a preference for G4 DNA structures over the canonical double helix. Moreover, the high affinities for parallel G4s, as indicated by the deep bound state minimum in the free energy curves, support the experimental evidence that DSM is capable of promoting G-quadruplex remodeling towards the all-parallel fold [197, 198, 239]. However, it is important to note that computing the binding free energy for other G4 forms is necessary to make definitive conclusions.

In the free energy profile shown in Fig. 4.18, the distance range of 1.3–2.5 nm represents non-native DSM/G4 complexes, which are partially formed intermediates. In these intermediates, the DSM helix interacts with the 5′-tetrad, but in a less favorable

manner, further highlighting the specialization of DSM in recognizing flat molecular surfaces. Specifically, a local minimum at around 1.6 nm in the case of puG4 corresponds to a state where the DSM helix is shifted along the tetrad, and its Tyr69 and Ile65 residues interact favorably with the groove region of the G-quadruplex, as illustrated in Fig. A.7.

4.3.2 Recognition Mechanism of DSM for Parallel G4s

To gain a deeper understanding of the molecular factors behind the strong binding affinity of DSM for G4, I conducted further calculations to determine the probabilities of forming van der Waals (vdW) contacts and hydrogen bonds between the DSM α1 helix and the two G4 folds in the bound state minima (Fig. 4.19A and B). The minimal distance cutoff for vdW contacts was set to 0.4 nm. The original US data was reweighted to obtain an unbiased equilibrium distribution (see Methods section 3.3.3).

The contact matrices presented in Fig. 4.19C demonstrate that several DSM residues (more than 10) make stable and extensive contacts (with probability approximately 1) with the sugar-phosphate and nucleobase moieties of G4, especially with those nucleotides that form the 5′-tetrad (G2·G6·G11(A10)·G15(T14)). The presence of two adjacent GXXXG motifs in DSM creates a smooth surface on one side of the helix, allowing it to associate closely with the flat surface of the tetrad and resulting in a large contact area between DSM and G4 (Fig. 4.19A and B). Remarkably, four DSM glycine residues (58, 62, 66, and 74) directly form a stable vdW contact with G4. This intimate interaction enables other mostly hydrophobic residues to also establish contacts with the 5′-tetrad, such as Tyr69, which is involved in π-stacking interactions with G2 and G6 (with probability >90,%), Ala70, which interacts with A1 and G2 (with probability >70,%), Leu67 with A1 (with probability >70,%), and Ile65 with G6 (with probability >90,%).

Notably, the contact patterns observed for ffG4 and puG4 in Fig. 4.19C are very similar, indicating that DSM aligns itself in the same manner on both G4 surfaces (see Fig. 4.20). The differences in contact patterns are mainly due to DHX36-induced sequence shifts in the 5′-tetrad (G11 -> A10 and G15 -> T14; G2 and G6 stay intact). Additionally, in both cases, the A1 base from the 5′-overhang wraps around the DSM helix (see Fig. 4.19AB, *right*) and makes multiple stable contacts. Given the similar binding affinities observed in Fig. 4.18, these matching contact patterns suggest that the identified flat-surface-to-flat-surface binding mode is essential for DSM's high specificity in recognizing and binding parallel G4s, as previously reported[195, 238].

To investigate if hydrogen bonds also play a role in the stability of DSM-G4 complexes, I also conducted an analysis of hydrogen bond probabilities across the DSM-G4 interface. The results are presented in Fig. 4.19C. The H-bond matrices show that only a few hydrogen bonds are formed in the DSM-G4 complexes. Specifically, in the case of puG4, only one significant hydrogen bond (>40,% prob.) was observed between Asn77 and the G3 phosphate group (Fig. 4.19A). The lack of significant hydrogen bonds suggests that the high binding free energy observed for the DSM/G4 complexes arises mainly from the attractive van der Waals and hydrophobic forces between the two extended apolar surfaces. In the case of ffG4, which has a longer second loop, a few additional hydrogen bonds were formed between the phosphate groups of G11 and G6 and basic groups of DSM (Lys61, Arg63, Lys76 and N-terminal Pro57) with probabilities exceeding 50%. These additional hydrogen bonds might

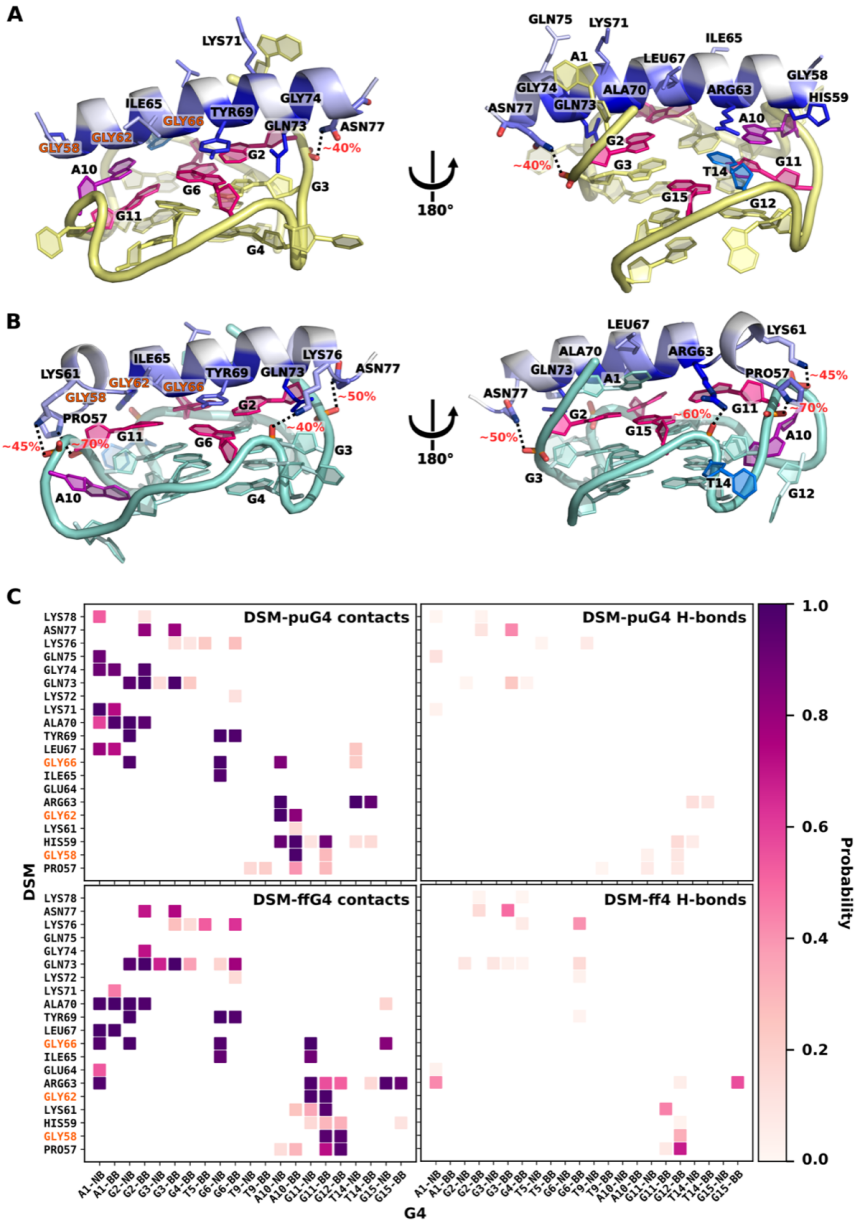

FIGURE 4.19: **A., B.** Representative structures of DSM α1 helix in complex with puG4 or ffG4, respectively, corresponding to the 0.8–1.1 nm range in Fig. 4.18. The equilibrium probabilities of DSM residues making contact with the G4 are indicated by a blue color scale. Black dotted lines denote the most probable hydrogen bonds and their formation probabilities. **C.** Equilibrium probabilities of contact formation (minimum distance <0.4 nm ; left) or hydrogen bond formation (right) between residue pairs across the DSM/G4 binding interface for complexes involving puG4 (top) or ffG4 (bottom). Nucleotide residues are shown separately for the nucleobase (NB) and sugar-phosphate moieties (BB). The glycine residues in the two adjacent GXXXG motifs in DSM are highlighted in orange.

account for the slightly higher affinity predicted for ffG4 in our simulations (see Fig. 4.18).

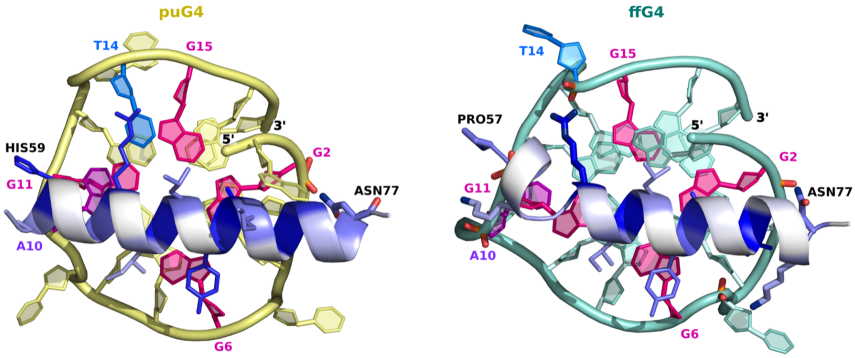

FIGURE 4.20: Representative structures of DSM α1 helix in complex with puG4 or ffG4, viewed from the top. The structures correspond to the 0.8–1.1 nm range shown in Fig. 4.18. The probability of DSM residue making contact with G4 is depicted by a blue color scale.

4.3.3 DSM preferentially binds to 5′-G-tetrad due to tighter surface contact

To understand why DSM prefers binding to the 5′-G-tetrad of parallel G4s, which has been previously reported [194, 195], I conducted a further investigation using the same US method as described earlier. This time, I obtained the free energy profile for DSM α1 helix binding to the DNA^{myc}-G4 alternative binding site, the 3′-G-tetrad. The resulting profile (3′ffG4) was compared to the original one describing the 5′-G-tetrad binding (5′ffG4) in Fig. 4.21A. The binding of DSM to the 3′-G-tetrad of G4 (3′ffG4) was also characterized by a well-defined global minimum indicative of a stable complex (Fig. 4.21BC). However, the affinity of DSM for the 3′-G-tetrad was predicted to be lower (by 5 kcal/mol) than for the 5′-G-tetrad, consistent with previous NMR results that demonstrated the binding of DSM to the 3′-G-tetrad only occurs at higher peptide concentrations or when it is enforced by G4 dimerization through the 5′-G-tetrads [195].

Since DSM plays a key role in G4 recognition, its strong preference for the 5′-G-tetrad may be critical in determining the overall binding mode and properly orienting G4 for interactions with other DHX36 subdomains involved in the unfolding process. Therefore, the molecular basis of DSM's recognition preference may have important implications in understanding the G4 recognition mechanism by DHX36.

In order to further understand why DSM has a preference for binding to the 5′-G-tetrad, I analyzed the binding mode of the 3′-G-tetrad (3′ffG4) by examining vdW contacts and hydrogen bonds, using the same methodology as for the 5′ffG4 mode. As shown in the contact matrix in Fig. 4.21D, DSM still forms extensive contacts with the G-tetrad at the 3′-end, indicating a relatively high affinity for this binding site. The orientation of the DSM helix with its flat side towards the 3′-G-tetrad (Fig. 4.21BC) is the same as for the 5′-G-tetrad, and the contact pattern is essentially identical, except for the nucleotides involved (G4·G8·G13·G17 for 3′-end).

However, the equilibrium contact probabilities calculated for the 3′ffG4 complex are lower than those for the 5′ffG4 complex (Fig. 4.19C), suggesting that the preference for the 5′-end is due to more favorable vdW and hydrophobic interactions. A detailed analysis of the binding mode revealed that the weaker vdW attraction to the 3′-G-tetrad is caused by a slightly more protruding sugar-phosphate backbone that sterically prevents the bulky DSM helix from optimally adsorbing at the flat surface

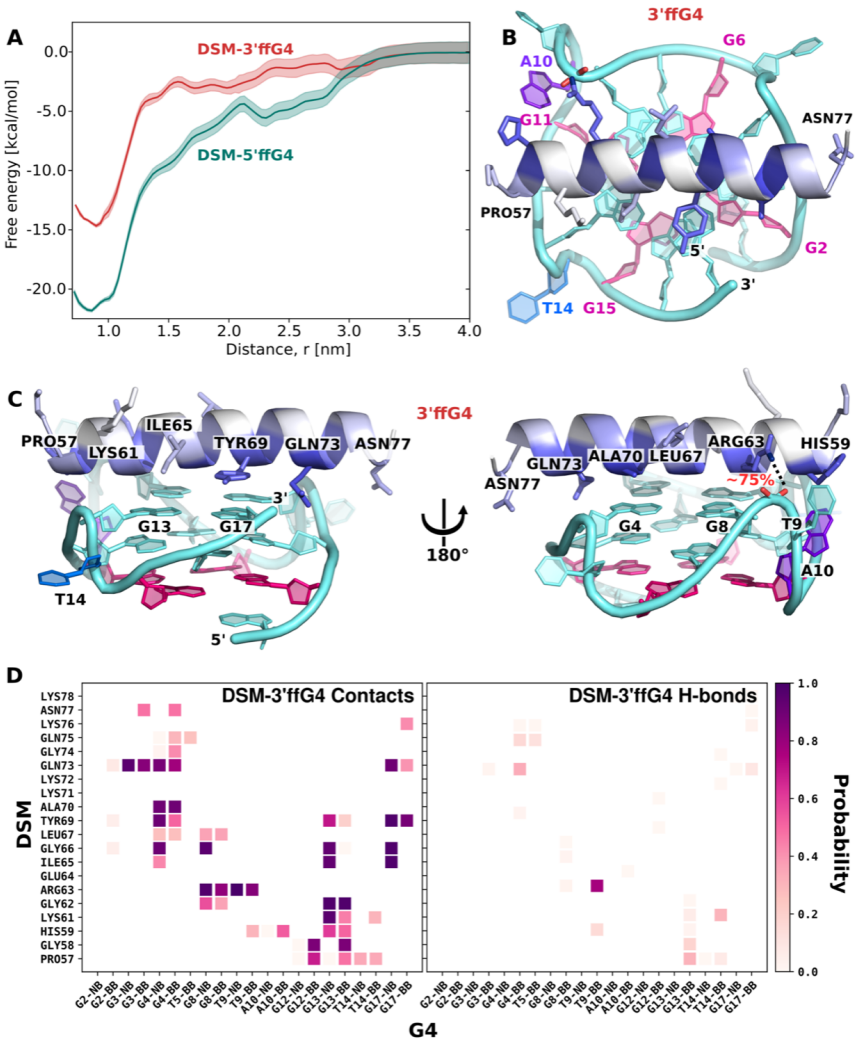

FIGURE 4.21: **A.** Comparison of the free energy profiles for DSM α1 helix binding to 3′- (3′ffG4, red) and 5′-G-tetrad (5′ffG4, cyan) of ffG4. **B, C.** Representative structures of a tightly bound complex of the DSM α1 helix and 3′ffG4, corresponding to the 0.8–1.1nm range in panel A. Equilibrium probabilities of DSM residue contact with the G4 are color-coded using a blue scale. Black dotted lines show the most probable hydrogen bonds along with their formation probabilities. **D.** Equilibrium probabilities of contact (minimum distance < 0.4 nm; *left*) or hydrogen bond (*right*) formation between pairs of residues across the DSM/G4 binding interface for complexes involving 3′ffG4. For nucleotide residues, the nucleobase (NB) and sugar-phosphate moieties (BB) are shown separately.

of the guanine plane. These differences in the arrangement of the sugar-phosphate backbone between the 3′- and 5′-ends of parallel G-quadruplexes arise from DNA strand polarity and have been discussed previously in the context of G4 binding preferences [240, 241].

Moreover, the H-bond matrix in Fig.4.21D shows only one strong hydrogen bond between Arg63 and T9 phosphate group, further indicating that polar interactions play only a marginal role in DSM binding to G4.

4.3.4 OB enhances G4 binding affinity through polar interactions with DNA backbone

Previous studies conducted both *in vitro* and *in vivo* have demonstrated that while the DSM motif is necessary for high-affinity G-quadruplex binding by DHX36, it is not sufficient on its own [197, 238, 242]. Deletion of the DSM motif reduces the binding affinity, although it has been reported that DHX36 can still bind to G4 with lower affinity [198]. Given the recent crystal structure that suggests the involvement of the OB subdomain in G4 interaction as well [194], I aimed to investigate the contribution of the OB subdomain to G4 recognition. I obtained a model of the (puG4) G-quadruplex bound to DHX36 with the DSM motif removed (see Methods section 3.3.3) and determined the free energy profile of G4 binding to the OB subdomain only, using the center-of-mass distance between the OB Cα atoms and the guanine core of G4 as a reaction coordinate (Fig. 4.22A, *inset (bottom)*).

The free energy profile (Fig. 4.22A) for the binding of the puG4 to the OB subdomain showed a deep free energy minimum of approximately 16 kcal/mol, which was found at 1.8-2.1 nm and corresponds to the OB/G4 bound mode that is identical to that in the full complex of DHX36/G4 (Fig. 4.22B, C). Other local minima at 2.5 and 3.1 nm were also noticeable and characterized loosely-bound intermediates of OB/puG4 complexes. This indicates the important role of the OB subdomain in G4 recognition. However, the obtained profile suggests that the OB contribution to the binding free energy is markedly lower than that of DSM, which is approximately 20 kcal/mol as indicated by the deep global minimum in Fig. 4.18.

Further, to investigate whether the binding of G4 to OB is cooperative with the binding of DSM to the 5′-G-tetrad, I performed additional computations. Specifically, I computed the free energy of OB/G4 binding in the context of full-length DHX36, where DSM is already bound to the 5′-G-tetrad (+DSM in the (*top*) inset in Fig. 4.22A). To ensure that DSM interacts optimally with G4 irrespective of the distance between OB and G4, I restricted the range of the reaction coordinate to 1.8–3.1 nm and analyzed the local response of the free energy profile. The DSM-G4 contacts and hydrogen bonds observed in this simulation (Fig. A.8) were compared to those found in the DSM/puG4 complex (Fig. 4.19C), which confirmed that DSM remained stably bound to the 5′-G-tetrad throughout the range of the reaction coordinate.

After comparing the two curves in the inset of Fig. 4.22A, it is evident that the presence of DSM does not alter the overall character of the binding free energy profile of OB. However, the minimum corresponding to the native complex is over 4 kcal/mol more stable, indicating a significant enhancement of OB affinity for G4 in the presence of the 5′-G-tetrad-bound DSM. This finding suggests that both DHX36 subdomains cooperate to bind to G4. Structurally, this can be attributed to the formation of a continuous binding interface that encompasses both sides of the G-quadruplex by the DSM α1 helix and OI loop (Asn851–Lys860) of the OB subdomain (see Fig. 4.23A). The free energy profile for the simultaneous binding of G4 to both G4-interacting surfaces in the full-length DXH36 (Fig. 4.23B) also supports this conclusion. Although this profile is affected by similar bound state overstabilization as those discussed earlier, the free energy gain due to G4 binding to both sides at the same time (-48kcal/mol) is significantly higher than the sum of individual gains (-20 and -16 kcal/mol for isolated DSM and OB, respectively), indicating a synergistic effect.

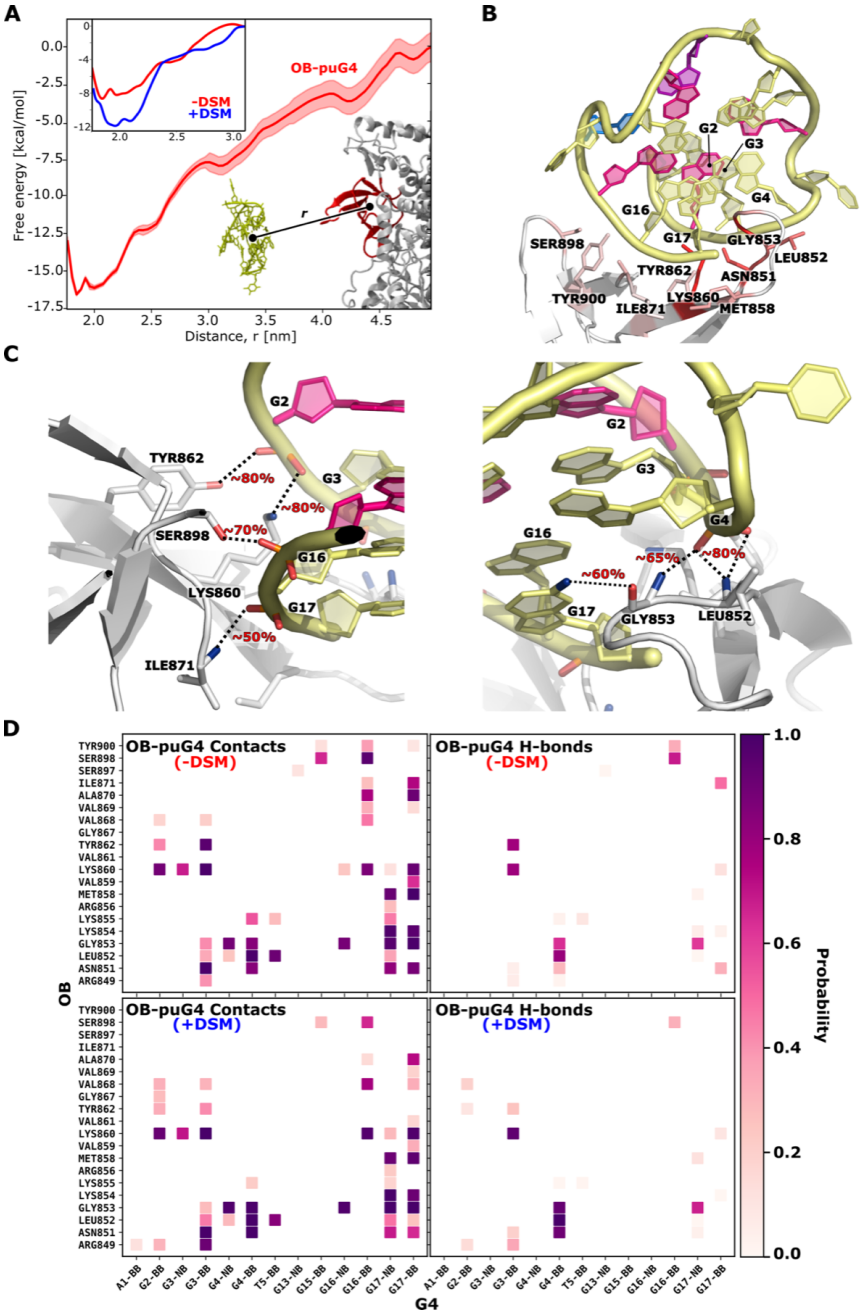

FIGURE 4.22: **A.** Free energy profiles depicting the binding of parallel G-quadruplex (puG4) to the OB subdomain of DHX36, in the absence of DSM (−DSM). The inset compares the profile in the bound-state minimum region to the profile obtained in the presence of DSM bound to the 5′-G-tetrad (+DSM), indicating a deepening of the global minimum by approximately 4 kcal/mol in the presence of DSM. **B.** Representative structure of tightly-bound puG4 to the OB subdomain, within the 1.8–2.1 nm range of r in panel A, with color-coded (in red scale) equilibrium probabilities of a given OB residue making contact with the G4. **C.** Representative structure of tightly-bound native complex of puG4 with the OB subdomain displaying the most probable complex-stabilizing hydrogen bonds (as dotted lines), along with their formation probabilities (two views of the same snapshot). **D.** The equilibrium probabilities of forming a contact (minimum distance < 0.4 nm; *left*) or a hydrogen bond (*right*) between pairs of residues across the puG4/OB binding interface in the absence (*top*) and presence (*bottom*) of the DSM subdomain. For nucleotide residues, the nucleobase (NB) and sugar-phosphate moieties (BB) are shown separately.

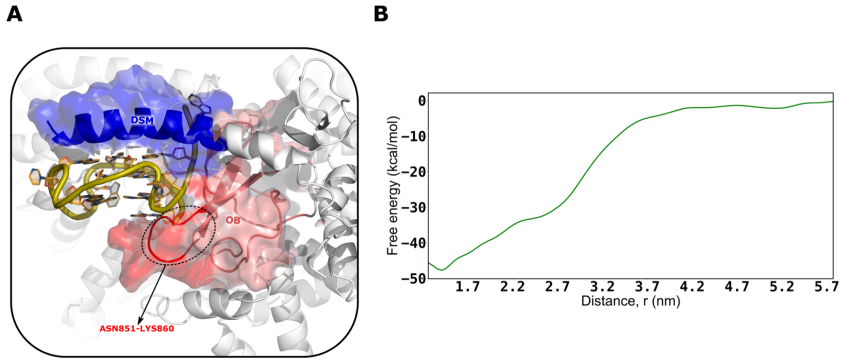

FIGURE 4.23: **A.** Representative structure of the parallel G-quadruplex (puG4) bound to the full-length DHX36. The DSM α1 helix and the OI loop (Asn851–Lys860) of the OB subdomain cooperate to form a continuous binding interface that wraps around both sides of the G-quadruplex. **B.** Free energy profiles representing the binding of the parallel G-quadruplex (puG4) to the full-length DHX36. The reaction coordinate is defined as the center-of-mass distance between the heavy atoms of the guanine core and the Cα atoms of both the DSM and OB subdomains combined.

In order to gain a better understanding of the energetics of the OB/G4 binding and how it depends on the presence of DSM, I also conducted an analysis of the vdW contacts and hydrogen bonds that stabilize the complex, using the same approach as described above. The contact matrices presented in Fig. 4.22D demonstrate that the OB/G4 interaction is mediated by several residues located throughout the OB domain, with the majority being part of the OI loop, which spans from Asn851 to Lys860 (as shown in the bottom portion of the matrix in Fig. 4.22D and in Fig. 4.23A). As seen in Fig.4.22B, the OI loop residues form stable contacts with the first three 5′-terminal guanine nucleotides (G2, G3 and G4) as well as with G4's 3′-end, specifically G16 and G17. This contact pattern arises due to the OI loop binding to the 3′-end of the G-quadruplex while also aligning itself parallel to the 5′-terminal part of the sugar-phosphate backbone (of the first G-tract), forming several specific polar interactions with it.

Furthermore, the H-bond matrices shown in Fig. 4.22D indicate that, unlike DSM, the OB subdomain forms strong hydrogen bonds with G4. Specifically, Leu852, Gly853, and Asn851 were found to form high-probability H-bonds with the phosphate group of G4 (as seen in Fig. 4.22C, *right*), while Lys860 and Tyr862 interact stably with the G3 phosphate (Fig. 4.22C, *left*). Additionally, Ser898, Gly853, and Ile871 form weaker H-bonds with the 3′-terminal guanine nucleotides, G16 and G17.

The comparison of the contact and hydrogen bond matrices between the +DSM and −DSM systems (shown in Fig. 4.22D) indicates that the binding of the DSM domain to the 5′-G-tetrad does not alter the overall pattern of OB/G4 contacts and hydrogen bonds. However, it does lead to a noticeable strengthening of the hydrogen bonds, particularly those involving the 5′-end G-tract. This strengthening of hydrogen bonds is accompanied by significant conformational changes in the side chains that participate in these interactions (see Fig. A.9).

4.3.5 Role of the Flexible Loop in the RecA2 Domain in Anchoring a G-Quadruplex

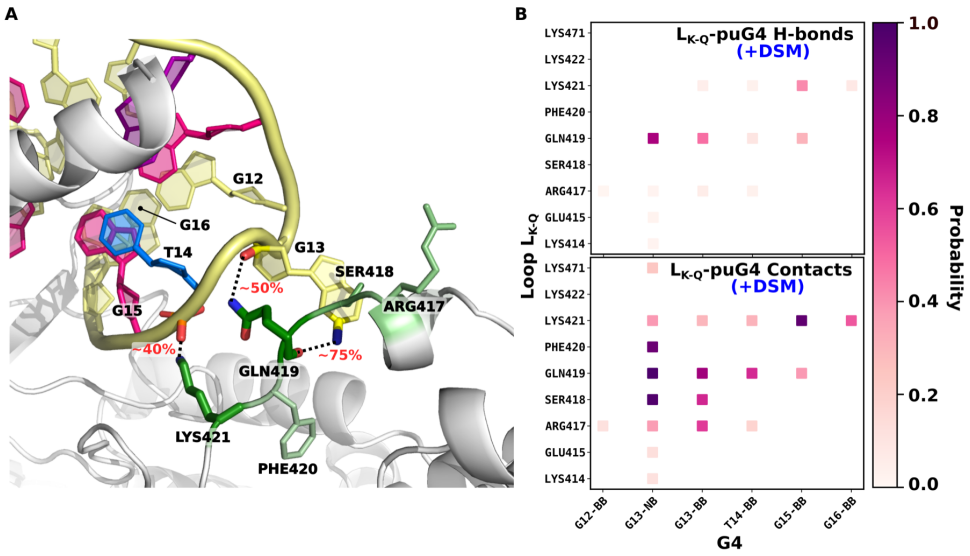

FIGURE 4.24: **A.** Representative structure showing the binding interface between the flexible $L_{K\text{-}Q}$ loop (Lys414–Gln426) and puG4, as identified in simulations of the full DHX36/puG4 complex. The equilibrium probabilities of each $L_{K\text{-}Q}$ residue making contact with the G4 are shown in green scale, with darker shades indicating higher probabilities. The most probable hydrogen bonds are indicated by black dotted lines, with their formation probabilities. **B.** Equilibrium probabilities for contact formation (minimum distance < 0.4 nm ; *top*) and hydrogen bond formation (*bottom*) between pairs of residues across the $L_{K\text{-}Q}$/G4 binding interface in the full-length DHX36 complex. For nucleotide residues, the nucleobase (NB) and sugar-phosphate moieties (BB) are displayed separately.

During the MD-based refinement of the DHX36/G4 complex, I observed that the flexible 13-residue loop of the RecA2 domain (Lys414–Gln426; referred to as $L_{K\text{-}Q}$ in Fig. 2.11) exhibited a strong interaction tendency for the G-quadruplex loop. This observation suggests that the $L_{K\text{-}Q}$ loop may form a third interface involved in G4 recognition, a finding consistent with shorter MD simulations of a parallel G4/DHX36 complex from *D. melangaster* [196]. To investigate this interaction further, I conducted an analysis of vdW contacts and h-bonds, using the same method as described previously, on the 1 μs-long unbiased trajectory of the DHX36/G4 complex.

In Fig. 4.24, the contact and h-bond probabilities indicate that the interaction between $L_{K\text{-}Q}$ and G-quadruplex is predominantly polar in nature. The most stable hydrogen bonds are formed between Gln419 and the nucleobase and backbone moieties of G13 in the third loop of G-quadruplex (Fig. 4.24A). As DHX36-mediated unfolding pulls this guanine nucleotide out of the 5′-G-tetrad, it could be speculated that the interaction with Gln419 is crucial for stabilizing the partially unfolded state of the G-quadruplex. The other residues of $L_{K\text{-}Q}$ that contribute to the interaction with G-quadruplex include Lys421 and Arg417, which form moderately stable h-bond-enhanced ion pairs with the phosphate groups of G15 and G13, respectively.

It is worth noting that the h-bond stability is low enough to allow the loop to remain flexible, as indicated by the root mean square fluctuations in Fig. 4.25A, making it challenging to resolve this loop structure through x-ray crystallography.

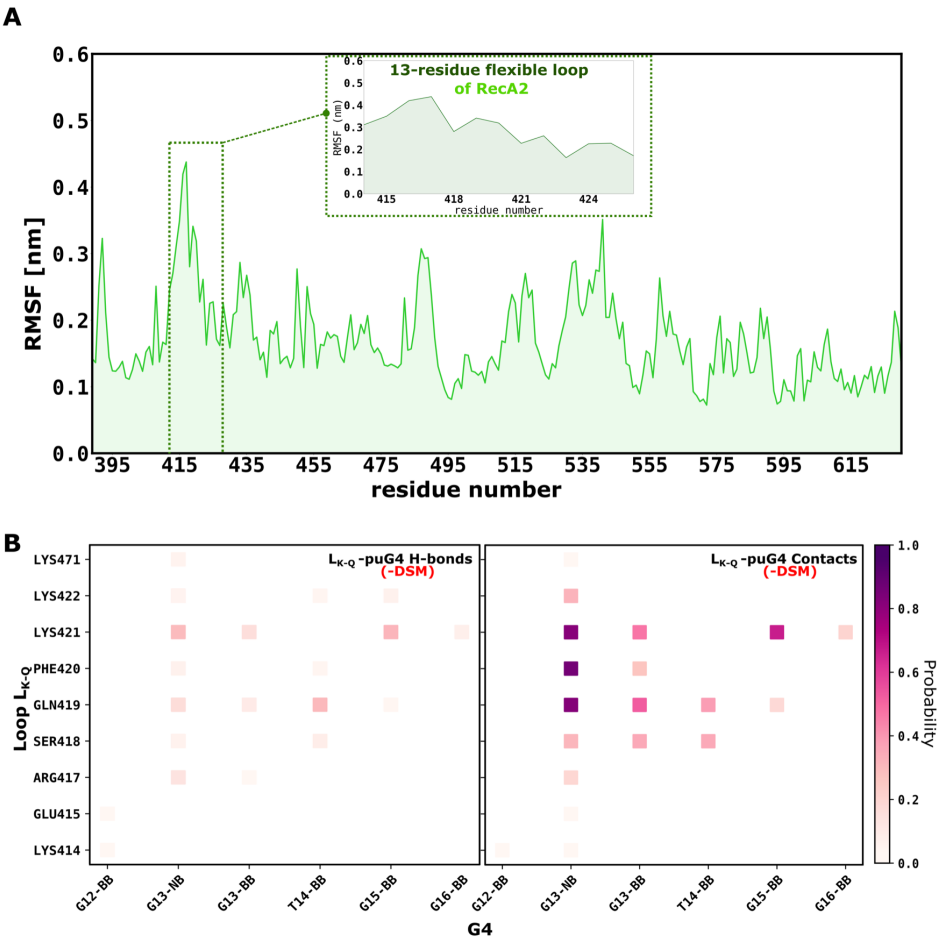

FIGURE 4.25: **A.** Residue-wise root mean square fluctuations (RMSF) of the RecA2 domain obtained from the 1 μs-long unbiased molecular dynamics (MD) simulation of the DHX36/G4 complex, using all heavy atoms for the calculation. **B.** Equilibrium probabilities of contact formation (minimum distance < 0.4 nm ; *left*) or hydrogen bond formation (*right*) between residue pairs across the $L_{K\text{-}Q}$/G4 binding interface in the DHX36 complex lacking the DSM subdomain.

In order to investigate whether the $L_{K\text{-}Q}$-mediated interaction with G-quadruplex exhibits cooperativity with other subdomains of DHX36, I conducted the same interaction analysis using an unbiased 1 μs trajectory of the DHX36/G4 complex, but with the entire DSM region removed (Fig. 4.25B). A comparison of Fig. 4.24B and Fig. 4.25B revealed that the removal of DSM resulted in the destabilization of virtually all $L_{K\text{-}Q}$-G4 hydrogen bonds. Specifically, the probabilities of h-bonds formed by Gln419 were reduced by about 7-fold, while those formed by Lys421- and Arg417 were less affected. This observation strongly suggests that G4 binding by DHX36 is cooperative in nature.

Chapter 5

Conclusions

My doctoral research employed a diverse array of computational methods, with a focus on molecular dynamics-based free energy calculations, to address fundamental questions in molecular biology. Specifically, my research aimed to elucidate the mechanisms by which DNA binding proteins locate their target DNA sequences or specific DNA structures. The findings obtained through this comprehensive computational approach have contributed to our understanding of the molecular interactions involved in DNA-protein recognition, shedding new light on these important biological processes.

One of the key aspects of my research, outlined in objective 1 of Chapter 1, revolved around investigating the direct readout mechanism. Specifically, I explored the role of acidic amino acid residues in determining DNA sequence specificity. Using structural bioinformatic analysis, I demonstrated that acidic amino acid residues (Asp/Glu) are highly prevalent in the DNA-protein interface, and that they exclusively readout cytosine. By employing free energy calculations, I investigated the role of Asp/Glu in sequence-specific DNA-protein recognition. Through computation of changes in binding free energy ($\Delta\Delta$G) of selected transcription factors upon mutation of Asp/Glu to alanine against various DNA sequences, I discovered that the contribution of acidic residues to DNA-binding affinity is delicately balanced between electrostatic repulsion from the DNA backbone and specific interactions with nucleobases. Notably, I observed that at non-cytosine sequences, where there are no significant attractive forces, the acidic residues generally disfavor binding ($\Delta\Delta G > 0$), thus acting as negative selectors to avoid these off-target sites. On the other hand, at cytosine-containing sequences, the contribution of acidic residues varies from negligible ($\Delta\Delta G \approx 0$) to favorable ($\Delta\Delta G < 0$), with the favorable contribution increasing with the number of cytosines in close proximity to Asp/Glu, as revealed by my energetic analysis. This cumulative effect is consistent with the presence of cytosine-rich sequences at the DNA sites recognized by Asp/Glu. This effect is primarily attributed to the long-range electrostatic attraction between the acidic residues and cytosines in the major groove of DNA, rather than solely relying on local hydrogen bonding interactions with the amino group. This finding suggests that acidic residues play a diverse and context-dependent role in DNA-protein recognition, with their impact on binding affinity being influenced by the specific nucleotide context. Moreover, I hypothesize that the long-range nature of this interaction could facilitate target search by destabilizing transient binding complexes at off-target sites. This may provide an evolutionary mechanism for tuning binding kinetics as target sequences become sparser in larger genomes.

Furthermore, my analysis of a model system involving the propionate ion interacting with a DNA duplex has provided insights into the strong preference of Asp/Glu for cytosine over adenine binding. Classical molecular dynamics (MD)

simulations revealed that forming a hydrogen bond with adenine is disfavored due to electrostatic repulsion with the N7 atom of the imidazole ring, and the tendency of purine-purine dinucleotides to adopt a BII backbone conformation, wherein the phosphate group approaches the amino group of adenine, making its interaction with negatively charged residues less favorable. Quantum chemical calculations reproduced the substantial difference in affinity between propionate for cytosine and adenine ($\Delta\Delta G_{ca}$ $\approx$2 kcal/mol), with roughly half of the difference attributed to repulsion from the N7 atom, and the remaining half resulting from differences in the strength of hydrogen bonds, with cytosine forming stronger hydrogen bonds by approximately 1 kcal/mol. These findings provide mechanistic insights into the preferential binding of Asp/Glu to cytosine over adenine, shedding light on the energetic factors governing this selectivity.

In another significant aspect of my doctoral research, I focused on the concept of shape readout, which involves proteins recognizing specific DNA conformations or structures.

Specifically, in the second part of my research (corresponding to objective 2 discussed in Chapter 1), I investigated the molecular mechanisms by which EXOG recognizes a particular chimeric conformation in a conventional duplex, rather than having a preference for a specific DNA sequence. Firstly, using free energy calculations of isolated DNA/DNA and RNA/DNA chimeric duplex (R2-DNA/DNA, where the first two nucleotides are from RNA) substrates, I demonstrated that unlike DNA/DNA duplex, the R2-DNA/DNA substrate can easily transition between A- and B-conformation states. Further, analysis of wild-type EXOG-bound DNA/DNA duplex confirms that EXOG has a higher preference for A-DNA in its substrate binding groove and prefers a chimeric conformation like R2-DNA/DNA, as the transition from A- to B-conformation state was energetically costly (by $\sim$ 2 kcal/mol) in my free energy calculation. This finding is consistent with EXOG's role in 5′-end processing in mitochondrial replication.

Furthermore, my analysis of the interactions between the substrate-binding region of EXOG and the bound DNA revealed that the substrate-binding region of EXOG is predominantly composed of polar residues that form strong and stable contacts with the nucleobase and/or sugar-phosphate moieties of the bound DNA. Notably, I found that the preference of EXOG for the A-conformation results from more favorable interactions with DNA and that the observed preference of EXOG for A-conformation can be attributed to the more favorable interactions of Arg109 from the core domain and Arg314 from the wing domain with DNA. Additionally, the interaction mode of these arginine residues provided a plausible mechanism for their potential role in inducing the transition from B-to-A conformation.

Further, free energy analysis of A- to B-conformational transition for R109A, R314A, and double mutant (R109A+R314A) variants of EXOG revealed that both arginines act cooperatively to make the B-to-A conformational transition when EXOG is bound to DNA/DNA duplex, which is the possible substrate for EXOG as it is also known to participate in the base excision repair pathway. Specifically, when EXOG is bound to a DNA/DNA substrate, Arg109 can attract the negatively charged sugar-phosphate backbone due to its positive electrical charge. This attraction can result in altered sugar-puckering and subsequent backbone conformation, ultimately facilitating the induction of the A-DNA conformation. On the other hand, tight contact of Arg314 with the 5′-end facilitates widening of the minor groove width.

These findings provide unprecedented insight into the mechanisms of EXOG's unique DNA conformation specificity and how it induces the transition when bound

to conventional B-DNA duplex regardless of the sequence composition. This understanding will be important for deciphering the replication and transcription machinery of mitochondria, which is less explored compared to the nucleus. Moreover, this knowledge might in the future facilitate drug discovery efforts against mitochondrial-related diseases.

In the third part of my research (corresponding to objective 3 discussed in Chapter 1), I delved into an extensive investigation of the molecular mechanisms underlying the selectivity of the DHX36 helicase for a specific DNA secondary structure known as G-quadruplexes (G4s), which go beyond the conventional duplex structure. My findings indicated that the DSM and OB subdomains of DHX36 are involved in recognizing G4s and have a high affinity for parallel-stranded G4s. This is consistent with previous experimental studies that reported high stability of the DHX36/G4 complex.

Going beyond the spatial resolution of the experiments, I also discovered that the DSM and OB subdomains are optimized for sensing two distinctive features of parallel G-quadruplexes: the exposed planar face of a G-tetrad and the specific backbone conformation of a G-tract, respectively. The DSM α1 helix interacts with the 5′-G-tetrad and contributes significantly to the binding free energy. The binding affinity does not depend on a particular base composition but rather relies on extensive contacts between adjacent GXXXG motifs and hydrophobic residues of DSM with the guanine plane. OB, on the other hand, binds to G4 mostly through polar interactions, and its contribution to the binding free energy is lower than that of DSM.

My simulations also suggested the existence of a third DHX36/G4 interaction site, the $L_{K\text{-}Q}$ region in the RecA2 domain, which was missing in the crystal structure. This region may participate in both the initial cooperative anchoring of G4 and the stabilization of its partially unfolded conformation. These findings provide important insights into the mechanism of DHX36's specificity for G-quadruplexes and the cooperative nature of the G4 recognition process.

Machine learning models are indeed poised to play a prominent role in protein design, benefiting from increasingly sophisticated heuristics [243, 244]. However, it is crucial to recognize that research focusing on the fundamental recognition rules of molecular structures, such as G-quadruplexes (G4s), retains its significance in various domains, including understanding evolution, disease mechanisms, and molecular engineering. By investigating these basic recognition rules, my research offers valuable insights that can augment the field of rational drug design and antibody engineering. It is worth emphasizing that while machine learning models offer powerful tools, understanding the actual mechanisms underlying molecular recognition remains essential for informing rational design strategies and fostering general scientific curiosity.

Furthermore, the study of mitochondrial replication and transcription machinery, such as the EXOG protein, is essential for understanding the complex interplay of molecules involved in these vital processes. Similarly, research on G4s is critical for developing new diagnostic and therapeutic strategies in various medical fields. For instance, G4s have been implicated in cancer, and understanding their roles in the regulation of gene expression and telomere maintenance could lead to new treatments for this disease. Overall, these research areas are of great importance

in advancing our understanding of the molecular basis of biological processes and in developing new tools to improve human health.

Appendix A

Supporting Information

A.1 Role of Asp/Glu in Determining DNA Sequence Specificity

	variant	feature							
		#C	#A	#G	#T	#Hb-C	#Hb-A	#Hb-H_2O	#Arg/Lys
CLOCK-BMAL (Glu43)	ref	0.96	1.15	0.35	0.92	0.81	0.01	4.43	18.94
	C2T	0.02	1.22	0.04	1.64	0.00	0.00	4.58	22.84
	C2A	0.00	2.58	0.05	1.61	0.00	0.72	4.72	0.27
	C2G	0.13	0.87	0.95	0.60	0.00	0.06	4.75	21.96
	all-5Å$_1$	0.14	0.06	0.00	0.63	0.00	0.00	5.70	15.32
	all-5Å$_2$	1.65	0.95	1.00	1.65	0.15	0.60	5.00	0.32
Zif268 (Asp70)	ref	1.75	0.00	0.78	0.00	0.79	0.00	2.22	17.79
	C2T	0.58	0.01	0.54	0.67	0.00	0.00	2.37	16.53
	C2A$_u$	0.14	0.09	0.03	0.03	0.00	0.01	3.84	12.58
	C2A	1.39	0.66	0.85	0.00	0.00	0.01	2.29	15.87
	C2G	0.11	0.00	0.14	0.00	0.00	0.03	2.92	16.48
	all-5Å$_1$	0.24	0.20	0.00	0.14	0.00	0.00	3.30	11.45
	all-5Å$_2$	0.00	0.36	0.00	1.15	0.00	0.00	3.50	12.16
Myb (Glu132)	ref	1.15	1.40	0.85	0.76	1.07	0.00	5.12	2.13
	C2T	0.12	0.40	0.00	0.95	0.00	0.01	4.89	5.90
	C2A	0.00	1.07	0.00	0.52	0.00	0.14	5.27	6.23
	C2G	0.00	0.72	0.94	0.48	0.00	0.00	4.84	14.25
	all-5Å$_1$	0.52	0.39	0.39	0.95	0.04	0.00	4.63	6.84
ERG3 (Asp363)	ref	0.74	0.00	0.07	0.00	0.03	0.00	6.84	7.30
	C2T	0.02	0.00	0.00	0.00	0.00	0.01	5.76	11.13
	C2A	0.00	0.00	0.00	0.02	0.00	0.00	6.76	4.99
	C2G	0.00	0.00	0.00	0.01	0.00	0.04	6.68	6.51
	all-5Å$_1$	0.00	0.00	0.00	0.08	0.00	0.00	6.46	2.47
TRF1 (Asp422)	ref	1.98	0.04	1.93	0.00	1.18	0.00	1.60	22.09
	C2T	0.03	0.01	0.01	0.00	0.00	0.00	4.30	15.19

TABLE A.1: Features used for evaluating associations between the simulation-derived $\Delta\Delta G$ values and the relevant structural properties of the studied DNA/protein complexes. Feature values were calculated as averages over MD trajectories of the respective complexes.

$^{i}R^2$	$^{j}R^2_a$	^{k}MAE	^{l}PRESS	$^{m}LOO_{MAE}$	$^{n}LPO_{MAE}$	^{o}k-Fold$_{MAE}$	^{p}Rep-k Fold$_{MAE}$
0.84	0.81	0.57	13.14	1.61	1.55	1.49	1.64

TABLE A.2: Statistical properties of the best random forest model involving four top-ranked structural features for predicting the contribution of Asp/Glu to the DNA-binding free energy, ΔΔG.
^{i}determination coefficient (goodness-of-fit), ^{j}adjusted R^2, ^{k} mean absolute error, The cross-validation parameters: ^{l}predicted residual error sum of squares, ^{m}leave one out error, ^{n}leave P out (P=3) error, ^{o}k-Fold error, ^{p}repeated k-Fold (#repeats=3). Errors are expressed in the free energy units (kcal/mol).

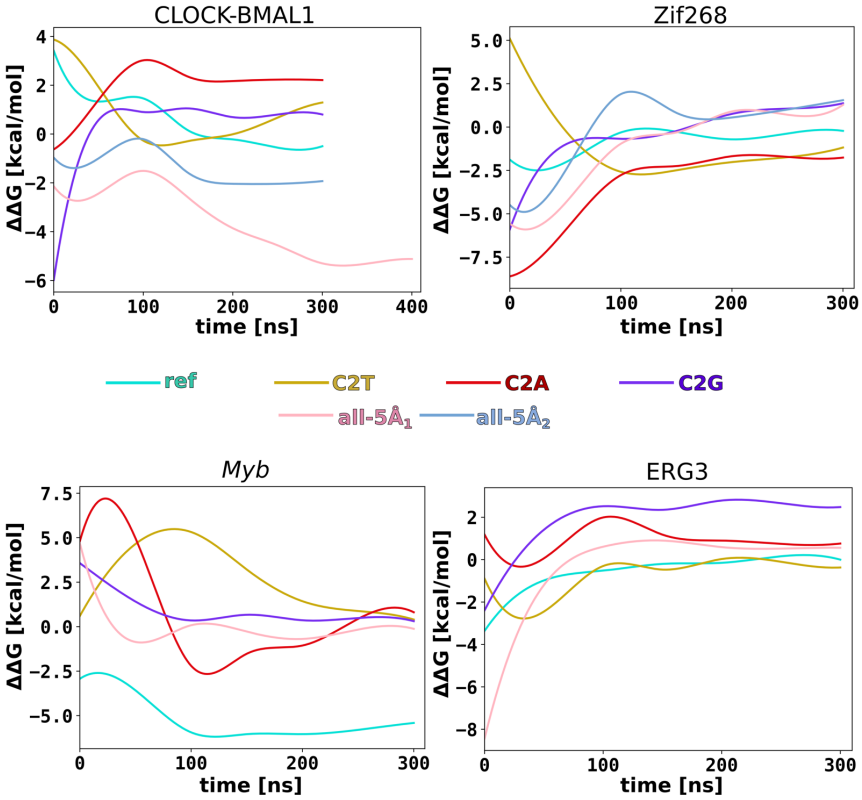

FIGURE A.1: The convergence of the difference in DNA-binding free energy (ΔΔG) between the wild-type protein and its alanine mutants.

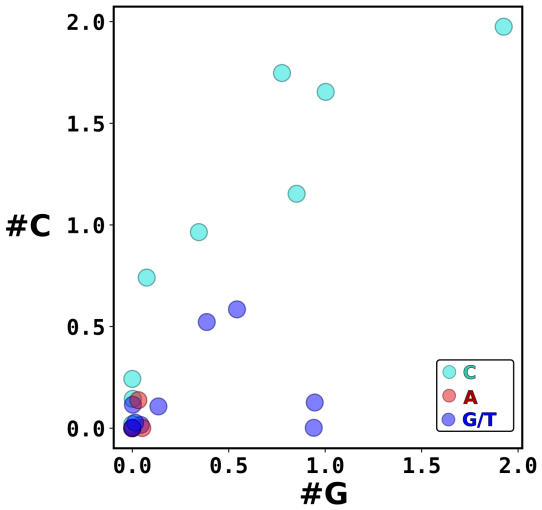

FIGURE A.2: Correlation between the number of cytosine (#C) and guanine (#G) residues in the local vicinity of Asp/Glu.

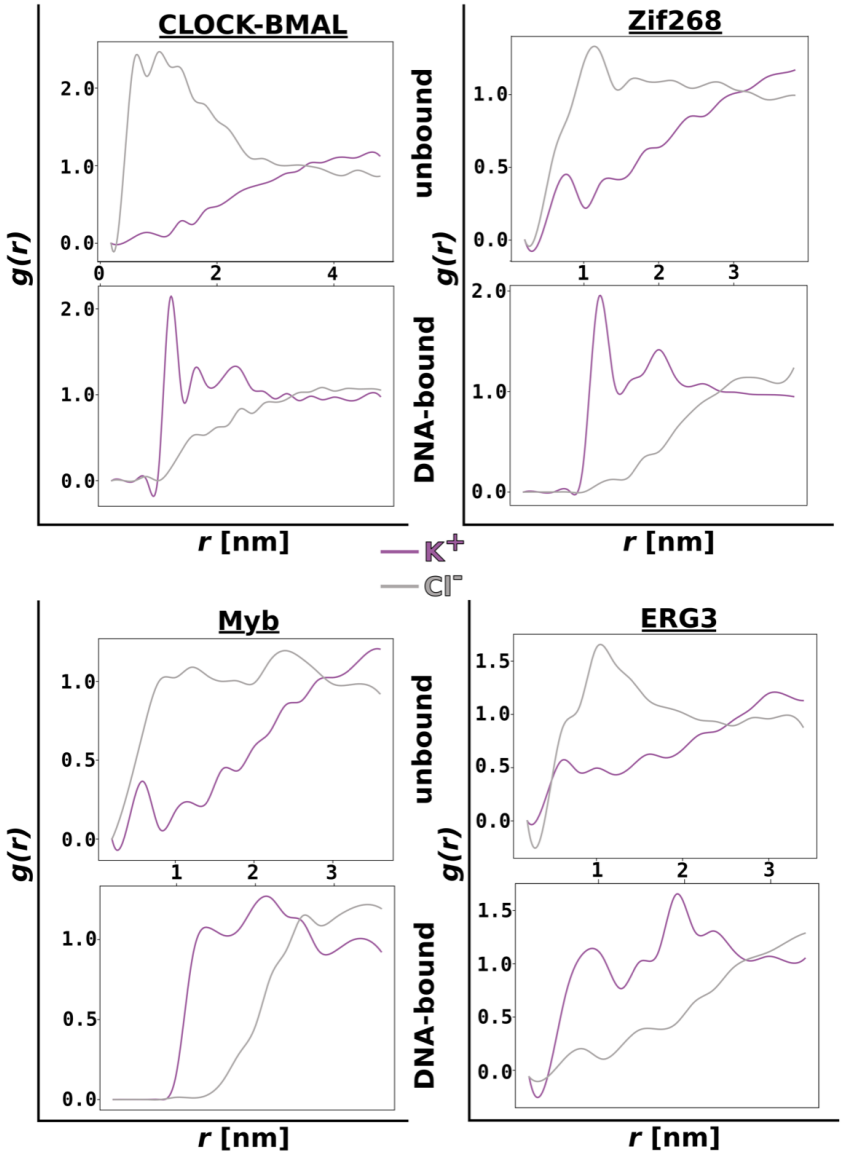

FIGURE A.3: Comparison of the radial distribution functions of K^+ and Cl^- ions around the carboxylic carbon atom of the investigated Asp/Glu residues (see Fig. 3.3) in the DNA-bound and unbound states.

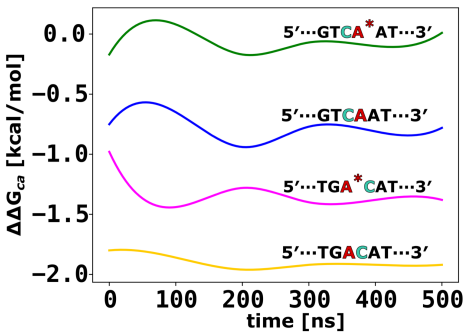

FIGURE A.4: Convergence of the $\Delta\Delta G_{ca}$ values, characterizing the propionic acid preference for cytosine vs. adenine, in the four considered sequence contexts. A* denotes adenine modified by neutralization of the partial charge on the N7 atom.

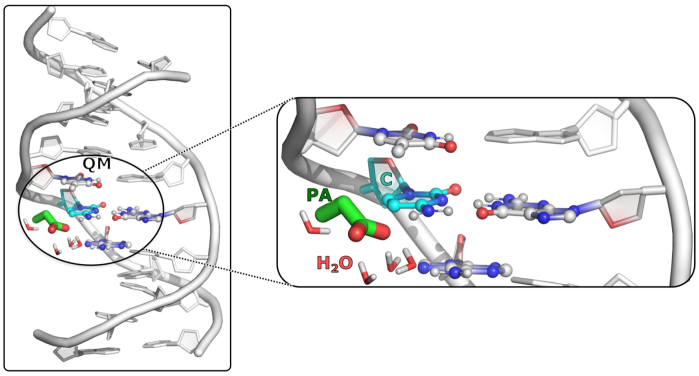

FIGURE A.5: A structural representation of the complex between the propionate ion and cytosine in a B-DNA decamer, utilized in our ab initio molecular dynamics (MD) simulations. The inset highlights the atoms from the DNA bases (ball representation), propionate ion, and water molecules that were part of the quantum mechanical (QM) region.

A.2 Recognition Mechanism of Parallel G-Quadruplexes by DHX36 Helicase

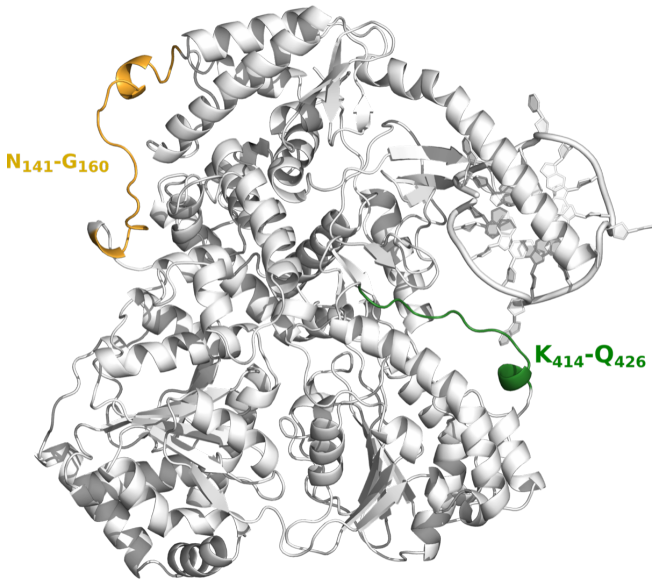

FIGURE A.6: The DHX36 crystal structure was completed using Modeller [245] to add missing segments: an orange-colored 20-residue linker between DSM and RecA1, and a green-colored 13-residue flexible loop of RecA2.

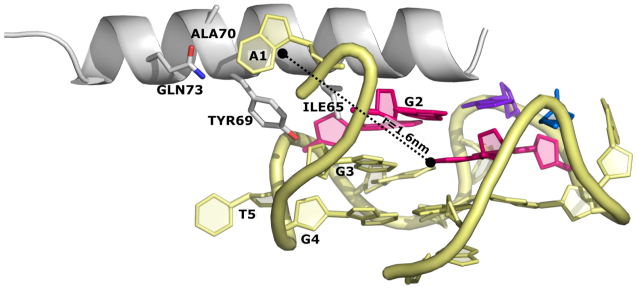

FIGURE A.7: A representative structure of the DSM/puG4 complex is shown, where the center-of-mass distance between the binding partners is approximately 1.6 nm.

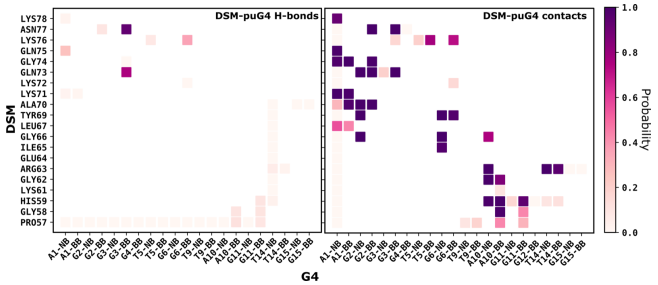

FIGURE A.8: The average probabilities of hydrogen bond formation (*left*) and residue contacts (minimum distance <0.4 nm; *right*) are depicted across the binding interface of DSM/G4. These probabilities were obtained from umbrella sampling simulations conducted to compute the free energy profile of G4 binding to the OB subdomain in the full-length DHX36 protein. The nucleotide residues are shown separately for the nucleobase (NB) and sugar-phosphate moieties (BB).

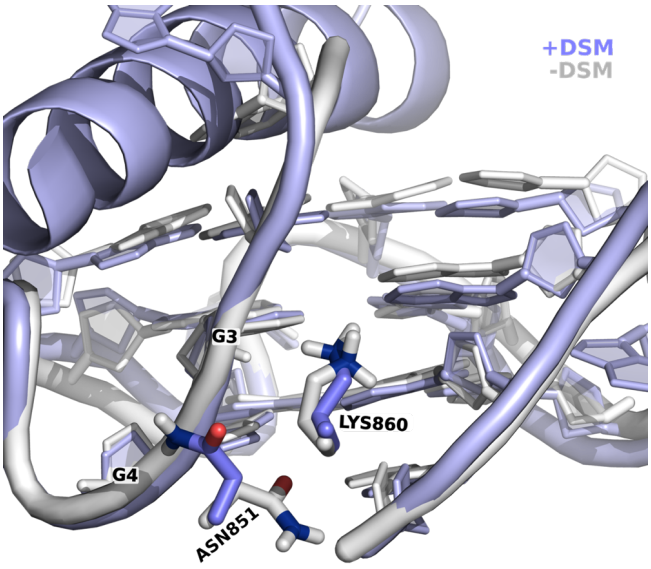

FIGURE A.9: The superposition of the OB/G4 interface is shown in the presence (+DSM) and absence (−DSM) of the DSM subdomain. Binding of DSM leads to conformational changes in Asn851 and, to a lesser extent, Lys860 of the OB subdomain, resulting in enhanced interactions with the 5′-end G-tract of the G-quadruplex (specifically residues G3 and G4).

Appendix B

Participation in Other Research Projects

During the course of my doctoral training, I actively engaged in additional research projects that were partly related to the topic of my thesis. One notable project involved the construction of membrane-embedded models of F1Fo-ATP synthase and the investigation of the molecular basis of the rotational unidirectionality of the c-ring subunit from the Fo part of ATP synthase. This collaboration resulted in a publication where I contributed as a co-author.

Furthermore, I had the opportunity to undertake a scientific internship at the Department of Bioengineering, University of California, Riverside, USA, under the supervision of Dr. Giulia Palermo (https://palermolab.com/). During this internship, I participated in various research projects within the lab. One particularly challenging project focused on elucidating the mechanism of R-loop formation by the cas9 protein in the CRISPR-Cas9 gene editing system. The formation of R-loops, where the single-stranded guide RNA displaces one of the DNA strands in a double-stranded region, is a crucial step in the targeting and recognition process of the cas9 protein. By investigating this mechanism, we aimed to deepen our understanding of the intricate molecular interactions involved in the CRISPR-Cas9 system and its potential implications for gene editing applications.

Additionally, I investigated the mechanism of broader PAM selection by xCas9, a novel variant of spCas9, which aligns with my research interests and runs parallel to the methodologies employed in my doctoral thesis. Through this ongoing collaboration, I have already obtained interesting results, such as uncovering the role of a glutamic acid residue in indirectly controlling the PAM recognition ability of the cas9 protein. I am hopeful that these projects will yield further fascinating findings, leading to scientific publications in reputable journals.

Appendix C

Scientific Achievements

Scientific Publications:

1. **K.A. Hossain**, M. Kogut, J. Słabońska, S. Sappati, M. Wieczór, J. Czub. How acidic amino acid residues facilitate dna target site selection., *Proceedings of the National Academy of Sciences, USA.* volume 120, page e2212501120. National Acad Sciences, 2023.

2. A. Marciniak, P. Chodnicki, **K.A. Hossain**, J. Slabonska, J. Czub. Determinants of directionality and efficiency of the atp synthase fo motor at atomic resolution. *The journal of physical chemistry letters*, volume 13, pages 387–392. ACS Publications, 2022.

3. **K.A. Hossain**, M. Jurkowski, J. Czub, M. Kogut. Mechanism of recognition of parallel g-quadruplexes by deah/rhau helicase dhx36 explored by molecular dynamics simulations. *Computational and Structural Biotechnology Journal*, volume 19, pages 2526–2536. Elsevier, 2021.

Conference Reports:

1. **K.A. Hossain**, Ł. Nierzwicki, P.R. Arantes, J. Czub, G. Palermo. Mechanism of broader PAM recognition by xCas9: explored by molecular dynamics simulations. *22nd Annual University of California Systemwide Symposium on Bioengineering and Biotechnology Industry Showcase.* August 8–10, 2022, Corwin Pavilion, University of California Santa Barbara, CA, USA.

2. **K.A. Hossain**, M. Kogut, J. Słabońska, S. Sappati, M. Wieczór, J. Czub. Role of Acidic Amino Acid Residues in Sequence-specific DNA-protein Interactions. *Multiscale simulations of DNA from electrons to nucleosomes: 22 years of the Ascona B-DNA Consortium.* April 17–21, 2023, Stefano Franscini Conference Center Monte Verità, Strada Collina 84 CH-6612 Ascona, Switzerland. (*Recipient of the Best Poster Award and selected for an oral presentation*).

Bibliography

1. Watson, J. D. & Crick, F. H. *The structure of DNA* in *Cold Spring Harbor symposia on quantitative biology* **18** (1953), 123–131.
2. Saenger, W. & Saenger, W. Dna structure. *Principles of Nucleic Acid Structure*, 253–282 (1984).
3. Baker, E. S. & Bowers, M. T. B-DNA helix stability in a solvent-free environment. *Journal of the American Society for Mass Spectrometry* **18,** 1188–1195 (2007).
4. Boutonnet, N., Hui, X. & Zakrzewska, K. Looking into the grooves of DNA. *Biopolymers: Original Research on Biomolecules* **33,** 479–490 (1993).
5. Lu, X.-J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research* **31,** 5108–5121 (2003).
6. Babcock, M. S., Pednault, E. P. & Olson, W. K. Nucleic acid structure analysis: mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *Journal of molecular biology* **237,** 125–156 (1994).
7. Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D & Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic acids research* **37,** 5917–5929 (2009).
8. Dršata, T. & Lankaš, F. Theoretical models of DNA flexibility. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3,** 355–363 (2013).
9. Poppleton, E. *et al.* Design, optimization and analysis of large DNA and RNA nanostructures through interactive visualization, editing and molecular simulation. *Nucleic acids research* **48,** e72–e72 (2020).
10. Pérez, A., Luque, F. J. & Orozco, M. Dynamics of B-DNA on the microsecond time scale. *Journal of the American Chemical Society* **129,** 14739–14745 (2007).
11. Kowiel, M., Brzezinski, D., Gilski, M. & Jaskolski, M. Conformation-dependent restraints for polynucleotides: the sugar moiety. *Nucleic Acids Research* **48,** 962–973 (2020).
12. Calladine, C. & Drew, H. A base-centred explanation of the B-to-A transition in DNA. *Journal of molecular biology* **178,** 773–782 (1984).
13. Balaceanu, A. *et al.* The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *The journal of physical chemistry letters* **8,** 21–28 (2017).
14. Várnai, P., Djuranovic, D., Lavery, R. & Hartmann, B. α/γ Transitions in the B-DNA backbone. *Nucleic acids research* **30,** 5398–5406 (2002).
15. Khan, N., Kolimi, N. & Rathinavelan, T. Twisting right to left: A... A mismatch in a CAG trinucleotide repeat overexpansion provokes left-handed Z-DNA conformation. *PLoS computational biology* **11,** e1004162 (2015).

16. Venkadesh, S, Mandal, P. & Gautham, N. The structure of a full turn of an A-DNA duplex d (CGCGGGTACCCGCG) 2. *Biochemical and Biophysical Research Communications* **407,** 307–312 (2011).

17. Szymanski, M. R. *et al.* A domain in human EXOG converts apoptotic endonuclease to DNA-repair exonuclease. *Nature communications* **8,** 14959 (2017).

18. Šponer, J. *et al.* The DNA and RNA sugar–phosphate backbone emerges as the key player. An overview of quantum-chemical, structural biology and simulation studies. *Physical Chemistry Chemical Physics* **14,** 15257–15277 (2012).

19. Dickerson, R. E. & Klug, A. Base sequence and helix structure variation in B and A DNA. *Journal of molecular biology* **166,** 419–441 (1983).

20. Kulkarni, M. & Mukherjee, A. Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition. *Progress in biophysics and molecular biology* **128,** 63–73 (2017).

21. Kosikov, K. M., Gorin, A. A., Zhurkin, V. B. & Olson, W. K. DNA stretching and compression: large-scale simulations of double helical structures. *Journal of molecular biology* **289,** 1301–1326 (1999).

22. Flatters, D. & Lavery, R. Sequence-dependent dynamics of TATA-box binding sites. *Biophysical journal* **75,** 372–381 (1998).

23. Pastor, N., Pardo, L. & Weinstein, H. Does TATA matter? A structural exploration of the selectivity determinants in its complexes with TATA box-binding protein. *Biophysical journal* **73,** 640–652 (1997).

24. Chenoweth, D. M., Meier, J. L. & Dervan, P. B. Pyrrole-imidazole polyamides distinguish between double-helical DNA and RNA. *Angewandte Chemie International Edition* **52,** 415–418 (2013).

25. Lohani, N., Narayan Singh, H., Agarwal, S., Mehrotra, R. & Rajeswari, M. R. Interaction of adriamycin with a regulatory element of hmgb1: spectroscopic and calorimetric approach. *Journal of Biomolecular Structure and Dynamics* **33,** 1612–1623 (2015).

26. Shaw, N. N. & Arya, D. P. Recognition of the unique structure of DNA: RNA hybrids. *Biochimie* **90,** 1026–1039 (2008).

27. Ren, J., Qu, X., Dattagupta, N. & Chaires, J. B. Molecular recognition of a RNA: DNA hybrid structure. *Journal of the American Chemical Society* **123,** 6742–6743 (2001).

28. Baranovskiy, A. G. *et al.* Activity and fidelity of human DNA polymerase α depend on primer structure. *Journal of Biological Chemistry* **293,** 6824–6843 (2018).

29. Hoogsteen, K. The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta crystallographica* **12,** 822–823 (1959).

30. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome research* **28,** 1264–1271 (2018).

31. Saini, N., Zhang, Y., Usdin, K. & Lobachev, K. S. When secondary comes first–the importance of non-canonical DNA structures. *Biochimie* **95,** 117–123 (2013).

32. Bang, I. Untersuchungen über die Guanylsäure. *Biochem. Z* **26,** 293–311 (1910).

33. Gellert, M., Lipsett, M. N. & Davies, D. R. Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences of the United States of America* **48,** 2013 (1962).
34. Sen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *nature* **334,** 364–366 (1988).
35. Gehring, K., Leroy, J.-L. & Gueron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **363,** 561–565 (1993).
36. Zeraati, M. *et al.* I-motif DNA structures are formed in the nuclei of human cells. *Nature chemistry* **10,** 631–637 (2018).
37. Pauling, L. & Corey, R. B. A proposed structure for the nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America* **39,** 84 (1953).
38. Felsenfeld, G, Davies, D. R. & Rich, A. Formation of a three-stranded polynucleotide molecule. *Journal of the American Chemical Society* **79,** 2023–2024 (1957).
39. Bacolla, A., Wang, G. & Vasquez, K. M. New perspectives on DNA and RNA triplexes as effectors of biological activity. *PLoS genetics* **11,** e1005696 (2015).
40. Mikheikin, A. L., Lushnikov, A. Y. & Lyubchenko, Y. L. Effect of DNA supercoiling on the geometry of holliday junctions. *Biochemistry* **45,** 12998–13006 (2006).
41. Ambrus, A., Chen, D., Dai, J., Jones, R. A. & Yang, D. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry* **44,** 2048–2058 (2005).
42. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic acids research* **34,** 5402–5415 (2006).
43. Hazel, P., Huppert, J., Balasubramanian, S. & Neidle, S. Loop-length-dependent folding of G-quadruplexes. *Journal of the American Chemical Society* **126,** 16405–16415 (2004).
44. Hatzakis, E., Okamoto, K. & Yang, D. Thermodynamic stability and folding kinetics of the major G-quadruplex and its loop isomers formed in the nuclease hypersensitive element in the human c-Myc promoter: effect of loops and flanking segments on the stability of parallel-stranded intramolecular G-quadruplexes. *Biochemistry* **49,** 9152–9160 (2010).
45. Bhattacharyya, D., Mirihana Arachchilage, G. & Basu, S. Metal cations in G-quadruplex folding and stability. *Frontiers in chemistry* **4,** 38 (2016).
46. Phan, A. T. Human telomeric G-quadruplex: structures of DNA and RNA sequences. *The FEBS journal* **277,** 1107–1117 (2010).
47. Maizels, N. Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nature structural & molecular biology* **13,** 1055–1059 (2006).
48. Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature biotechnology* **33,** 877–881 (2015).
49. Brázda, V. *et al.* G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* **35,** 3493–3495 (2019).
50. Del Villar-Guerra, R., Trent, J. O. & Chaires, J. B. G-quadruplex secondary structure obtained from circular dichroism spectroscopy. *Angewandte Chemie* **130,** 7289–7293 (2018).

51. Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic acids research* **42,** 860–869 (2013).
52. Di Antonio, M. *et al.* Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nature chemistry* **12,** 832–837 (2020).
53. Lipps, H. J. & Rhodes, D. G-quadruplex structures: in vivo evidence and function. *Trends in cell biology* **19,** 414–422 (2009).
54. Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nature reviews Molecular cell biology* **18,** 279–284 (2017).
55. Zheng, K.-w. *et al.* Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic acids research* **48,** 11706–11720 (2020).
56. Simonsson, T., Kubista, M. & Pecinka, P. DNA tetraplex formation in the control region of c-myc. *Nucleic acids research* **26,** 1167–1172 (1998).
57. Dai, J. *et al.* An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *Journal of the American Chemical Society* **128,** 1096–1098 (2006).
58. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nature communications* **10,** 1–16 (2019).
59. Kumari, S., Bugaut, A., Huppert, J. L. & Balasubramanian, S. An RNA G-quadruplex in the 5 UTR of the NRAS proto-oncogene modulates translation. *Nature chemical biology* **3,** 218–221 (2007).
60. Collie, G. W., Haider, S. M., Neidle, S. & Parkinson, G. N. A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex. *Nucleic acids research* **38,** 5569–5580 (2010).
61. Arora, A. *et al.* Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *Rna* **14,** 1290–1296 (2008).
62. Singh, S., Mathur, T., Gupta, K. & Garg, R. in *Legume Genomics* 261–268 (Springer, 2020).
63. Dehé, P.-M. & Gaillard, P.-H. L. Control of structure-specific endonucleases to maintain genome stability. *Nature Reviews Molecular Cell Biology* **18,** 315–330 (2017).
64. Ang, C. E. & Wernig, M. Profiling DNA–transcription factor interactions. *Nature Biotechnology* **36,** 501–502 (2018).
65. Hörberg, J., Moreau, K., Tamás, M. J. & Reymer, A. Sequence-specific dynamics of DNA response elements and their flanking sites regulate the recognition by AP-1 transcription factors. *Nucleic Acids Research* **49,** 9280–9293 (2021).
66. Ferguson, L. R. *et al. Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition* in *Seminars in cancer biology* **35** (2015), S5–S24.
67. Monk, D., Mackay, D. J., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics* **20,** 235–248 (2019).
68. Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* **13,** 840–852 (2012).

69. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research* **36,** 5221–5231 (2008).

70. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods* **5,** 829–834 (2008).

71. Zhu, L., Guo, W.-L., Deng, S.-P. & Huang, D.-S. ChIP-PIT: Enhancing the analysis of ChIP-Seq data using convex-relaxed pair-wise interaction tensor decomposition. *IEEE/ACM transactions on computational biology and bioinformatics* **13,** 55–63 (2015).

72. Ouyang, Z., Zhou, Q. & Wong, W. H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences* **106,** 21521–21526 (2009).

73. Orlov, Y. *et al.* Computer and statistical analysis of transcription factor binding and chromatin modifications by ChIP-seq data in embryonic stem cell. *Journal of Integrative bioinformatics* **9,** 88–100 (2012).

74. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annual Review of Cell and Developmental Biology* **35,** 357 (2019).

75. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics Development* **43,** 110–119 (2017).

76. Stasyuk, O. A., Jakubec, D., Vondrasek, J. & Hobza, P. Noncovalent interactions in specific recognition motifs of protein–DNA complexes. *Journal of Chemical Theory and Computation* **13,** 877–885 (2017).

77. Schleif, R. DNA binding by proteins. *Science* **241,** 1182–1187 (1988).

78. Vuzman, D., Polonsky, M. & Levy, Y. Facilitated DNA search by multidomain transcription factors: cross talk via a flexible linker. *Biophysical Journal* **99,** 1202–1211 (2010).

79. Koudelka, G. B. & Carlson, P. DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 represser. *Nature* **355,** 89–91 (1992).

80. Cheng, A. C., Chen, W. W., Fuhrmann, C. N. & Frankel, A. D. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *Journal of Molecular Biology* **327,** 781–796 (2003).

81. Kalodimos, C. G. *et al.* Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305,** 386–389 (2004).

82. Sarai, A. & Kono, H. Protein-DNA recognition patterns and predictions. *Annual Review of Biophysics and Biomolecular structure* **34,** 379 (2005).

83. Scipioni, A., Anselmi, C., Zuccheri, G., Samori, B. & De Santis, P. Sequence-dependent DNA curvature and flexibility from scanning force microscopy images. *Biophysical Journal* **83,** 2408–2418 (2002).

84. Perez, A., Lankas, F., Luque, F. J. & Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Research* **36,** 2379–2394 (2008).

85. Fuxreiter, M., Simon, I. & Bondos, S. Dynamic protein–DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences* **36,** 415–423 (2011).

86. Öztürk, M. A., Pachov, G. V., Wade, R. C. & Cojocaru, V. Conformational selection and dynamic adaptation upon linker histone binding to the nucleosome. *Nucleic Acids Research* **44,** 6599–6613 (2016).

87. Jaiswal, A. K. & Krishnamachari, A. Physicochemical property based computational scheme for classifying DNA sequence elements of Saccharomyces cerevisiae. *Computational Biology and Chemistry* **79,** 193–201 (2019).

88. Murphy, K. P., Xie, D., Thompson, K. S., Amzel, L. M. & Freire, E. Entropy in biological binding processes: estimation of translational entropy loss. *Proteins: Structure, Function, and Bioinformatics* **18,** 63–67 (1994).

89. Thorpe, I. F. & Brooks III, C. L. Molecular evolution of affinity and flexibility in the immune system. *Proceedings of the National Academy of Sciences* **104,** 8821–8826 (2007).

90. Afek, A. & Lukatsky, D. B. Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophysical journal* **105,** 1653–1660 (2013).

91. Campagne, S, Gervais, V & Milon, A. Nuclear magnetic resonance analysis of protein–DNA interactions. *Journal of the Royal Society Interface* **8,** 1065–1078 (2011).

92. Ahmad, S., Keskin, O., Sarai, A. & Nussinov, R. Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Research* **36,** 5922–5932 (2008).

93. Jones, S., Van Heyningen, P., Berman, H. M. & Thornton, J. M. Protein-DNA interactions: a structural analysis. *Journal of molecular biology* **287,** 877–896 (1999).

94. Hoffman, M. M. *et al.* AANT: The amino acid–nucleotide interaction database. *Nucleic acids research* **32,** D174–D181 (2004).

95. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences* **39,** 381–399 (2014).

96. Isbel, L., Grand, R. S. & Schübeler, D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nature Reviews Genetics* **23,** 728–740 (2022).

97. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature structural & molecular biology* **14,** 1025–1040 (2007).

98. Escudé, C. & Sun, J.-S. DNA major groove binders: triple helix-forming oligonucleotides, triple helix-specific DNA ligands and cleaving agents. *DNA Binders and Related Subjects,* 109–148 (2005).

99. Poddar, S., Chakravarty, D. & Chakrabarti, P. Structural changes in DNA-binding proteins on complexation. *Nucleic Acids Research* **46,** 3298–3308 (2018).

100. Watkins, D., Hsiao, C., Woods, K. K., Koudelka, G. B. & Williams, L. D. P22 c2 Repressor- Operator complex: Mechanisms of direct and indirect readout. *Biochemistry* **47,** 2325–2338 (2008).

101. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annual review of biochemistry* **79,** 233–269 (2010).

102. Gromiha, M. M., Siebers, J. G., Selvaraj, S., Kono, H. & Sarai, A. Intermolecular and intramolecular readout mechanisms in protein–DNA recognition. *Journal of molecular biology* **337,** 285–294 (2004).

103. Koudelka, G. B., Mauro, S. A. & Ciubotaru, M. Indirect Readout of DNA Sequence by Proteins: The Roles of DNA Sequence-Dependent Intrinsic and Extrinsic Forces. *Progress in Nucleic Acid Research and Molecular Biology* **81,** 143–177 (2006).

104. Bosch, D., Campillo, M. & Pardo, L. Binding of proteins to the minor groove of DNA: what are the structural and energetic determinants for kinking a base-pair step? *Journal of Computational Chemistry* **24,** 682–691 (2003).

105. Mondal, M., Yang, L., Cai, Z., Patra, P. & Gao, Y. Q. A perspective on the molecular simulation of DNA from structural and functional aspects. *Chemical Science* **12,** 5390–5409 (2021).

106. Alniss, H. Y. Thermodynamics of DNA minor groove binders: perspective. *Journal of Medicinal Chemistry* **62,** 385–402 (2018).

107. Battistini, F. *et al.* How B-DNA dynamics decipher sequence-selective protein recognition. *Journal of Molecular Biology* **431,** 3845–3859 (2019).

108. Chen, S. *et al.* Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *Journal of Molecular Biology* **314,** 75–82 (2001).

109. Cheatham III, T. E. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Current Opinion in Structural Biology* **14,** 360–367 (2004).

110. Harris, L.-A., Watkins, D., Williams, L. D. & Koudelka, G. B. Indirect readout of DNA sequence by p22 repressor: roles of DNA and protein functional groups in modulating DNA conformation. *Journal of Molecular Biology* **425,** 133–143 (2013).

111. Cloutier, T. E. & Widom, J. DNA twisting flexibility and the formation of sharply looped protein–DNA complexes. *Proceedings of the National Academy of Sciences* **102,** 3645–3650 (2005).

112. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome biology* **1,** 1–37 (2000).

113. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461,** 1248–1253 (2009).

114. Van de Wetering, M. & Clevers, H. Sequence-specific interaction of the HMG box proteins TCF-1 and SRY occurs within the minor groove of a Watson-Crick double helix. *The EMBO journal* **11,** 3039–3044 (1992).

115. Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences* **95,** 11163–11168 (1998).

116. Rohs, R., West, S. M., Liu, P. & Honig, B. Nuance in the double-helix and its role in protein–DNA recognition. *Current opinion in structural biology* **19,** 171–177 (2009).

117. Nguyen, B., Neidle, S. & Wilson, W. D. A role for water molecules in DNA-ligand minor groove recognition. *Accounts of chemical research* **42,** 11–21 (2009).

118. Zhang, Y., Xi, Z., Hegde, R. S., Shakked, Z. & Crothers, D. M. Predicting indirect readout effects in protein–DNA interactions. *Proceedings of the National Academy of Sciences* **101,** 8337–8341 (2004).

119. Garvie, C. W. & Wolberger, C. Recognition of specific DNA sequences. *Molecular cell* **8,** 937–946 (2001).

120. Paillard, G. & Lavery, R. Analyzing protein-DNA recognition mechanisms. *Structure* **12,** 113–122 (2004).

121. Jauch, R., Ng, C. K., Narasimhan, K. & Kolatkar, P. R. The crystal structure of the Sox4 HMG domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochemical Journal* **443,** 39–47 (2012).

122. Fujii, S., Kono, H., Takenaka, S., Go, N. & Sarai, A. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Research* **35,** 6063–6074 (2007).

123. Segall, A. M., Goodman, S. & Nash, H. Architectural elements in nucleoprotein complexes: interchangeability of specific and non-specific DNA binding proteins. *The EMBO journal* **13,** 4536–4548 (1994).

124. Verma, S. C., Harned, A., Narayan, K. & Adhya, S. Non-specific and specific DNA binding modes of bacterial histone, HU, separately regulate distinct physiological processes through different mechanisms. *Molecular Microbiology* (2023).

125. Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Current Biology* **14,** R546–R551 (2004).

126. Henikoff, S. & Smith, M. M. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology* **7,** a019364 (2015).

127. De Almeida, L. C., Calil, F. A., Machado-Neto, J. A. & Costa-Lotufo, L. V. DNA damaging agents and DNA repair: From carcinogenesis to cancer therapy. *Cancer Genetics* **252,** 6–24 (2021).

128. Mukherjee, S., Chakraborty, P. & Saha, P. Phosphorylation of Ku70 subunit by cell cycle kinases modulates the replication related function of Ku heterodimer. *Nucleic Acids Research* **44,** 7755–7765 (2016).

129. Kragelund, B. B., Weterings, E., Hartmann-Petersen, R. & Keijzers, G. The Ku70/80 ring in non-homologous end-joining: easy to slip on, hard to remove. *Frontiers in Bioscience-Landmark* **21,** 514–527 (2016).

130. Doherty, A. J., Jackson, S. P. & Weller, G. R. Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS letters* **500,** 186–188 (2001).

131. Nedialkov, Y. A. & Triezenberg, S. J. Quantitative assessment of in vitro interactions implicates TATA-binding protein as a target of the VP16C transcriptional activation region. *Archives of biochemistry and biophysics* **425,** 77–86 (2004).

132. Hieb, A. R., Gansen, A., Böhm, V. & Langowski, J. The conformational state of the nucleosome entry–exit site modulates TATA box-specific TBP binding. *Nucleic acids research* **42,** 7561–7576 (2014).

133. Wu, S.-Y. & Chiang, C.-M. TATA-binding protein-associated factors enhance the recruitment of RNA polymerase II by transcriptional activators. *Journal of biological chemistry* **276,** 34235–34243 (2001).

134. Frank, D. E. *et al.* Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator: effects of converting a consensus site to a non-specific site. *Journal of molecular biology* **267,** 1186–1206 (1997).

135. Her, C. *et al.* Molecular interactions underlying the phase separation of HP1α: role of phosphorylation, ligand and nucleic acid binding. *Nucleic Acids Research* **50,** 12702–12722 (2022).

136. Tomac, S. *et al.* Ionic effects on the stability and conformation of peptide nucleic acid complexes. *Journal of the American Chemical Society* **118,** 5544–5552 (1996).

137. Dragan, A. I. *et al.* DNA binding of a non-sequence-specific HMG-D protein is entropy driven with a substantial non-electrostatic contribution. *Journal of molecular biology* **331,** 795–813 (2003).

138. Gao, M. & Skolnick, J. From nonspecific DNA–protein encounter complexes to the prediction of DNA–protein interactions. *Plos Computational Biology* **5,** e1000341 (2009).

139. Fazary, A. E., Ju, Y.-H. & Abd-Rabboh, H. S. How does chromatin package DNA within nucleus and regulate gene expression? *International journal of biological macromolecules* **101,** 862–881 (2017).

140. Taylor, J. A. *et al.* Specific and non-specific interactions of ParB with DNA: implications for chromosome segregation. *Nucleic acids research* **43,** 719–731 (2015).

141. Jalal, A. S. & Le, T. B. Bacterial chromosome segregation by the ParABS system. *Open biology* **10,** 200097 (2020).

142. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10,** 252–263 (2009).

143. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic acids research* **34,** D95–D97 (2006).

144. Baek, I., Friedman, L. J., Gelles, J. & Buratowski, S. Single-molecule studies reveal branched pathways for activator-dependent assembly of RNA polymerase II pre-initiation complexes. *Molecular cell* **81,** 3576–3588 (2021).

145. Gaston, K. & Jayaraman, P.-S. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cellular and Molecular Life Sciences CMLS* **60,** 721–741 (2003).

146. Reid, J. E., Evans, K. J., Dyer, N., Wernisch, L. & Ott, S. Variable structure motifs for transcription factor binding sites. *BMC genomics* **11,** 1–18 (2010).

147. Pabo, C. O. & Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry* **61,** 1053–1095 (1992).

148. Wang, Z., Wu, Y., Li, L. & Su, X.-D. Intermolecular recognition revealed by the complex structure of human CLOCK-BMAL1 basic helix-loop-helix domains with E-box DNA. *Cell research* **23,** 213–224 (2013).

149. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252,** 809–817 (1991).

150. Ogata, K. *et al.* Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79,** 639–648 (1994).

151. Yin, Z., Machius, M., Nestler, E. J. & Rudenko, G. Activator Protein-1: redox switch controlling structure and DNA-binding. *Nucleic acids research* (2017).

152. Heim, M. A. *et al.* The basic helix–loop–helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Molecular biology and evolution* **20,** 735–747 (2003).

153. Shklover, J. *et al.* MyoD uses overlapping but distinct elements to bind E-box and tetraplex structures of regulatory sequences of muscle-specific genes. *Nucleic acids research* **35,** 7087–7095 (2007).

154. Chang, A. T. *et al.* An evolutionarily conserved DNA architecture determines target specificity of the TWIST family bHLH transcription factors. *Genes & development* **29,** 603–616 (2015).

155. Buck, M. J. & Atchley, W. R. Phylogenetic analysis of plant basic helix-loop-helix proteins. *Journal of molecular evolution* **56,** 742–750 (2003).

156. Longo, A., Guanga, G. P. & Rose, R. B. Crystal Structure of E47- NeuroD1/Beta2 bHLH Domain- DNA Complex: Heterodimer Selectivity and DNA Recognition. *Biochemistry* **47,** 218–229 (2008).

157. Miller, J, McLachlan, A. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal* **4,** 1609–1614 (1985).

158. Layat, E., Probst, A. V. & Tourmente, S. Structure, function and regulation of transcription factor IIIA: from Xenopus to Arabidopsis. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1829,** 274–282 (2013).

159. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* **29,** 183–212 (2000).

160. Bonchuk, A. *et al.* Structural basis of diversity and homodimerization specificity of zinc-finger-associated domains in Drosophila. *Nucleic Acids Research* **49,** 2375–2389 (2021).

161. Liu, C., Hao, J., Qiu, M., Pan, J. & He, Y. Genome-wide identification and expression analysis of the MYB transcription factor in Japanese plum (Prunus salicina). *Genomics* **112,** 4875–4886 (2020).

162. Fernandez-Lopez, R. *et al.* Structural basis of direct and inverted DNA sequence repeat recognition by helix–turn–helix transcription factors. *Nucleic Acids Research* **50,** 11938–11947 (2022).

163. Wu, Q. *et al.* The MYB transcription factor MYB103 acts upstream of TRICHOME BIREFRINGENCE-LIKE27 in regulating aluminum sensitivity by modulating the O-acetylation level of cell wall xyloglucan in Arabidopsis thaliana. *The Plant Journal* **111,** 529–545 (2022).

164. Wang, Z. *et al.* Genome-wide analysis of the basic leucine zipper (bZIP) transcription factor gene family in six legume genomes. *BMC genomics* **16,** 1–15 (2015).

165. Agarwal, S. K. *et al.* Menin interacts with the AP1 transcription factor JunD and represses JunD-activated transcription. *Cell* **96,** 143–152 (1999).

166. Ji, Z. *et al.* The forkhead transcription factor FOXK2 promotes AP-1-mediated transcriptional regulation. *Molecular and cellular biology* **32,** 385–398 (2012).

167. Koldin, B., Suckow, M., Seydel, A., von Wilcken-Bergmann, B. & Müller-Hill, B. A comparison of the difterent DNA binding specificities of the bZip proteins C/EBP and GCN4. *Nucleic acids research* **23,** 4162–4169 (1995).

168. Song, S. *et al.* DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proceedings of the National Academy of Sciences* **102,** 4990–4995 (2005).

169. Berk, A. J. & Clayton, D. A. Mechanism of mitochondrial DNA replication in mouse L-cells: asynchronous replication of strands, segregation of circular daughter molecules, aspects of topology and turnover of an initiation sequence. *Journal of molecular biology* **86,** 801–824 (1974).

170. Robberson, D. L., Kasamatsu, H. & Vinograd, J. Replication of mitochondrial DNA. Circular replicative intermediates in mouse L cells. *Proceedings of the National Academy of Sciences of the United States of America* **69,** 737 (1972).

171. Pham, X. H. *et al.* Conserved sequence box II directs transcription termination and primer formation in mitochondria. *Journal of Biological Chemistry* **281,** 24647–24652 (2006).

172. Kang, D., Miyako, K., Kai, Y., Irie, T. & Takeshige, K. In vivo determination of replication origins of human mitochondrial DNA by ligation-mediated polymerase chain reaction. *Journal of Biological Chemistry* **272,** 15275–15279 (1997).

173. Fusté, J. M. *et al.* Mitochondrial RNA polymerase is needed for activation of the origin of light-strand DNA replication. *Molecular cell* **37,** 67–78 (2010).

174. Jiang, M. *et al.* The mitochondrial single-stranded DNA binding protein is essential for initiation of mtDNA replication. *Science Advances* **7,** eabf8631 (2021).

175. Uhler, J. P. & Falkenberg, M. Primer removal during mammalian mitochondrial DNA replication. *DNA repair* **34,** 28–38 (2015).

176. Nowotny, M. *et al.* Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. *Molecular cell* **28,** 264–276 (2007).

177. Lima, W. F. *et al.* Human RNase H1 discriminates between subtle variations in the structure of the heteroduplex substrate. *Molecular pharmacology* **71,** 83–91 (2007).

178. Cotner-Gohara, E. *et al.* Human DNA ligase III recognizes DNA ends by dynamic switching between two DNA-bound states. *Biochemistry* **49,** 6165–6176 (2010).

179. Tann, A. W. *et al.* Apoptosis induced by persistent single-strand breaks in mitochondrial genome: critical role of EXOG (5-EXO/endonuclease) in their repair. *Journal of Biological Chemistry* **286,** 31975–31983 (2011).

180. Zheng, L. *et al.* Human DNA2 is a mitochondrial nuclease/helicase for efficient processing of DNA replication and repair intermediates. *Molecular cell* **32,** 325–336 (2008).

181. Al-Behadili, A. *et al.* A two-nuclease pathway involving RNase H1 is required for primer removal at human mitochondrial OriL. *Nucleic acids research* **46,** 9471–9483 (2018).

182. Cymerman, I. A., Chung, I., Beckmann, B. M., Bujnicki, J. M. & Meiss, G. EXOG, a novel paralog of Endonuclease G in higher eukaryotes. *Nucleic acids research* **36,** 1369–1379 (2008).

183. Wu, C.-C., Lin, J. L. J., Yang-Yen, H.-F. & Yuan, H. S. A unique exonuclease ExoG cleaves between RNA and DNA in mitochondrial DNA replication. *Nucleic acids research* **47,** 5405–5419 (2019).

184. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic acids research* **43,** 8627–8637 (2015).

185. Creacy, S. D. *et al.* G4 resolvase 1 binds both DNA and RNA tetramolecular quadruplex with high affinity and is the major source of tetramolecular quadruplex G4-DNA and G4-RNA resolving activity in HeLa cell lysates. *Journal of Biological Chemistry* **283,** 34626–34634 (2008).

186. Paul, T. *et al.* E. coli Rep helicase and RecA recombinase unwind G4 DNA and are important for resistance to G4-stabilizing ligands. *Nucleic acids research* **48,** 6640–6653 (2020).

187. Huber, M. D., Duquette, M. L., Shiels, J. C. & Maizels, N. A conserved G4 DNA binding domain in RecQ family helicases. *Journal of molecular biology* **358,** 1071–1080 (2006).

188. Rudolf, J., Makrantoni, V., Ingledew, W. J., Stark, M. J. & White, M. F. The DNA repair helicases XPD and FancJ have essential iron-sulfur domains. *Molecular cell* **23,** 801–808 (2006).

189. Paeschke, K. *et al.* Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **497,** 458–462 (2013).

190. He, Y., Andersen, G. R. & Nielsen, K. H. Structural basis for the function of DEAH helicases. *EMBO reports* **11,** 180–186 (2010).

191. Iwamoto, F., Stadler, M., Chalupníková, K., Oakeley, E. & Nagamine, Y. Transcription-dependent nucleolar cap localization and possible nuclear function of DExH RNA helicase RHAU. *Experimental cell research* **314,** 1378–1391 (2008).

192. Sauer, M. *et al.* DHX36 prevents the accumulation of translationally inactive mRNAs with G4-structures in untranslated regions. *Nature communications* **10,** 1–15 (2019).

193. Liu, G. *et al.* RNA G-quadruplex regulates microRNA-26a biogenesis and function. *Journal of hepatology* **73,** 371–382 (2020).

194. Chen, M. C. *et al.* Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558,** 465–469 (2018).

195. Heddi, B., Cheong, V. V., Martadinata, H. & Phan, A. T. Insights into G-quadruplex specific recognition by the DEAH-box helicase RHAU: Solution structure of a peptide–quadruplex complex. *Proceedings of the National Academy of Sciences* **112,** 9608–9613 (2015).

196. Chen, W.-F. *et al.* Molecular mechanistic insights into Drosophila DHX36-mediated G-quadruplex unfolding: a structure-based model. *Structure* **26,** 403–415 (2018).

197. Srinivasan, S., Liu, Z., Chuenchor, W., Xiao, T. S. & Jankowsky, E. Function of auxiliary domains of the DEAH/RHA helicase DHX36 in RNA remodeling. *Journal of molecular biology* **432,** 2217–2231 (2020).

198. You, H., Lattmann, S., Rhodes, D. & Yan, J. RHAU helicase stabilizes G4 in its nucleotide-free state and destabilizes G4 upon ATP hydrolysis. *Nucleic acids research* **45,** 206–214 (2017).

199. Chen, M. C., Murat, P., Abecassis, K., Ferré-D'Amaré, A. R. & Balasubramanian, S. Insights into the mechanism of a G-quadruplex-unwinding DEAH-box helicase. *Nucleic acids research* **43,** 2223–2231 (2015).

200. Comm, I.-I. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry* **9,** 3471–3479 (1970).

201. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **118,** 11225–11236 (1996).

202. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **25,** 1656–1676 (2004).

203. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **117,** 5179–5197 (1995).

204. Brooks, B. R. *et al.* CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **4,** 187–217 (1983).

205. Case, D. A. *et al.* The Amber biomolecular simulation programs. *Journal of computational chemistry* **26,** 1668–1688 (2005).

206. Humphreys, D. D., Friesner, R. A. & Berne, B. J. A multiple-time-step molecular dynamics algorithm for macromolecules. *The Journal of Physical Chemistry* **98,** 6885–6892 (1994).

207. Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E* **52,** 2893 (1995).

208. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **100,** 020603 (2008).

209. Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22,** 245–268 (1976).

210. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* **129,** 124105 (2008).

211. Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *The Journal of Chemical Physics* **22,** 1420–1426 (1954).

212. Sharma, R., Gangwar, S. P. & Saxena, A. K. Comparative structure analysis of the ETSi domain of ERG3 and its complex with the E74 promoter DNA sequence. *Acta Crystallographica Section F: Structural Biology Communications* **74,** 656–663 (2018).

213. Court, R., Chapman, L., Fairall, L. & Rhodes, D. How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. *EMBO reports* **6,** 39–45 (2005).

214. Wieczór, M. & Czub, J. How proteins bind to DNA: target discrimination and dynamic sequence search by the telomeric protein TRF1. *Nucleic Acids Research* **45,** 7643–7654 (2017).

215. Derreumaux, S., Chaoui, M., Tevanian, G. & Fermandjian, S. Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Research* **29,** 2314–2326 (2001).

216. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79,** 926–935 (1983).

217. Peters, M. B. *et al.* Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *Journal of Chemical Theory and Computation* **6,** 2935–2947 (2010).

218. Webb, B. & Sali, A. in *Functional genomics* 39–54 (Springer, 2017).

219. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1,** 19–25 (2015).

220. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **185,** 604–613 (2014).

221. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nature methods* **13,** 55–58 (2016).

222. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126,** 014101 (2007).

223. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52,** 7182–7190 (1981).

224. York, D. M., Darden, T. A. & Pedersen, L. G. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *The Journal of Chemical Physics* **99,** 8345–8348 (1993).

225. Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **4,** 116–122 (2008).

226. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13,** 952–962 (1992).

227. Gapsys, V. & de Groot, B. L. pmx Webserver: a user friendly interface for alchemistry. *Journal of Chemical Information and Modeling* **57,** 109–114 (2017).

228. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23,** 187–199 (1977).

229. Shaffer, P., Valsson, O. & Parrinello, M. Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proceedings of the National Academy of Sciences* **113,** 1150–1155 (2016).

230. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **16,** 1339–1350 (1995).

231. Sagendorf, J. M., Berman, H. M. & Rohs, R. DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic Acids Research* **45,** W89–W97 (2017).

232. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *Proceedings of the 31st international conference on neural information processing systems* (2017), 4768–4777.

233. Heddi, B., Oguey, C., Lavelle, C., Foloppe, N. & Hartmann, B. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Research* **38,** 1034–1047 (2010).

234. Tisné, C., Delepierre, M. & Hartmann, B. How NF-κB can be attracted by its cognate DNA. *Journal of Molecular Biology* **293,** 139–150 (1999).

235. Czaplewski, L., North, A., Smith, M., Baumberg, S & Stockley, P. Purification and initial characterization of AhrC: the regulator of arginine metabolism genes in Bacillus subtilis. *Molecular microbiology* **6,** 267–275 (1992).

236. Weaver, T. M. *et al.* Mechanism of nucleotide discrimination by the translesion synthesis polymerase Rev1. *Nature Communications* **13,** 2876 (2022).

237. Szymanski, M. R., Karlowicz, A., Herrmann, G. K., Cen, Y. & Yin, Y. W. Human EXOG Possesses Strong AP Hydrolysis Activity: Implication on Mitochondrial DNA Base Excision Repair. *Journal of the American Chemical Society* (2022).

238. Lattmann, S., Giri, B., Vaughn, J. P., Akman, S. A. & Nagamine, Y. Role of the amino terminal RHAU-specific motif in the recognition and resolution of guanine quadruplex-RNA by the DEAH-box RNA helicase RHAU. *Nucleic acids research* **38,** 6219–6233 (2010).

239. Chang-Gu, B., Bradburna, D., Yangyuorub, P. M. & Russell, R. The DHX36-specific-motif (DSM) enhances specificity by accelerating recruitment of DNA G-quadruplex structures. *Biological Chemistry,* 000010151520200302 (2021).

240. Parkinson, G. N., Lee, M. P. & Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417,** 876–880 (2002).

241. Kogut, M., Kleist, C. & Czub, J. Why do G-quadruplexes dimerize through the 5'-ends? Driving forces for G4 DNA dimerization examined in atomic detail. *PLoS computational biology* **15,** e1007383 (2019).

242. Chalupníková, K. *et al.* Recruitment of the RNA helicase RHAU to stress granules via a unique RNA-binding domain. *Journal of Biological Chemistry* **283,** 35186–35198 (2008).

243. Volk, M. J. *et al.* Biosystems design by machine learning. *ACS synthetic biology* **9,** 1514–1533 (2020).

244. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

245. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Current protocols in bioinformatics* **54,** 5–6 (2016).