Functions of bio-molecules deviating from canonical structures

Thesis submitted for the partial fulfillment of the requirements for the degree Doctor of Philosophy in Science

by

Kanika Kole

Department of Chemistry

Faculty of Natural and Mathematical Sciences

Presidency University,

Kolkata, India

2025

Functions of bio-molecules deviating from canonical structures

Thesis submitted for the partial fulfillment of the requirements for the degree Doctor of Philosophy in Science

by

Kanika Kole

Under the Supervision of

Prof. Jaydeb Chakrabarti
Department of Physics of complex systems
S. N. Bose National Centre for Basic Sciences
Kolkata, India

Department of Chemistry
Faculty of Natural and Mathematical Sciences
Presidency University,
Kolkata, India

2025

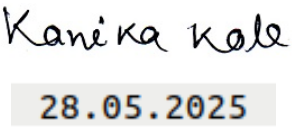
Thesis Title: *Functions of bio-molecules deviating from canonical structures*

Name of the Candidate: *Kanika Kole*

Registration Number: *RC001-22RS205210349*

Date of Registration: *12th August, 2022*

Department: *Department of Chemistry*



Signature of the Candidate with date

Dedication

To my beloved family— my mother, father and brother — for their constant love, support and inspiration.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Jaydeb Chakrabarti, for his unwavering support and invaluable guidance over the years, which not only enhanced my understanding of the subject but also helped me grow as a more capable and competent scientist. His confidence in me allowed me to pursue diverse problems simultaneously from my background in chemistry to molecular-level research which was very important for my development.

I also thank my thesis committee members, Dr. Suman Chakrabarty and Prof. Priya Mahadevan of S. N. Bose National Centre for Basic Sciences (SNBNCBS) and Prof. Shubhra Ghosh Dastidar from Bose Institute for their very insightful comments and suggestions throughout my evaluations during the PhD tenure. I would also like to express my gratitude to the reviewers for taking the time to review this thesis. Their valuable comments will be greatly beneficial to my work.

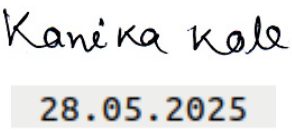
I am grateful to acknowledge my group members — Dr. Sasthi Charan Mandal, Dr. Edwine Tendong, Dr. Abhik Ghosh Moulick, Dr. Aayatti Mallick Gupta, Dr. Rahul Karmakar, Dr. Suravi Pal, Dr. Sutapa Dutta, Anirban Paul, Avik Sasmal, Anusree Sen and Sabuj Mandal — for their unique contributions and support during different phases of my work. I also sincerely thank Dr. Srabani Chakrabarty, Dr. Sourav Ghosh, Dr. Sankar Das, Ms. Mita Maity, Ms. Camelia Chowdhury and all my friends for their inspiration and encouragement.

I gratefully acknowledge SNBNCBS for providing excellent research facilities and a supportive environment throughout my academic journey. I am grateful to all the faculty members of the Centre for their sincere cooperation. I also extend my heartfelt thanks to the DR Academic, Ms. Nibedita Konar and all the academic and non-academic staff of the institute for their efficient support in providing the official documents I required from time to time. I gratefully acknowledge the Computer Section of SNBNCBS for their support and access to high-speed computing resources. I also sincerely thank the canteen and security staff at SNBNCBS for their dedicated service.

My journey would not have come this far without the unwavering love and support of my family. I am deeply grateful to my parents, Namita and

Kachiram Kole, whose dedication and sacrifices ensured that I received a quality education. I also thank my brother Kanchan for his constant support and inspiration throughout my studies. This thesis would not have taken shape without his unwavering encouragement.

Finally, I extend my sincere gratitude to the University Grants Commission (UGC) of India for awarding me the research fellowship that supported my work throughout this study.



Signature of the Candidate with date

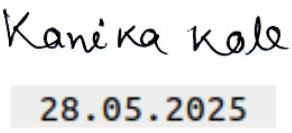
Declaration

I hereby declare that this thesis contains original research work carried out by me under the guidance of Prof. Jaydeb Chakrabarti, Senior Professor, Department of Physics of Complex Systems, S. N. Bose National Centre for Basic Sciences(SNBNCBS), Kolkata, India as part of the PhD programme.

All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

I also declare that, this work has not been submitted for any degree either in part or in full to any other institute or University before.



Signature of the Candidate with date

CENTRE FOR BASIC SCIENCES 1986
S.N. BOSE NATIONAL
1894-1974
विज्ञानेन परिपश्यन्ति धीराः

सत्येन्द्र नाथ बसु राष्ट्रीय मौलिक विज्ञान केन्द्र
SATYENDRA NATH BOSE NATIONAL
CENTRE FOR BASIC SCIENCES
সত্যেন্দ্র নাথ বসু জাতীয় মৌল বিজ্ঞান কেন্দ্র

Certificate

This is to certify that the thesis entitled " Functions of bio-molecules deviating from canonical structures " submitted by Kanika Kole, Registration Number RC001-22RS205210349 and date of registration 12th August, 2022, in partial fulfillment of the requirements for the award of "Doctor of Philosophy", is a record of bonafide research work carried out by her under my supervision.

Neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Signature of the Supervisor with date and official stamp

Jaydeb Chakrabarti 21/5/25

प्रो. जयदेब चक्रवर्ती / Prof. Jaydeb Chakrabarti
वरिष्ठ प्रोफेसर / Senior Professor
जटिल प्रणाली भौतिकी विभाग
Department of Physics of Complex Systems
सत्येंद्र नाथ बसु राष्ट्रीय मौलिक विज्ञान केंन्द्र
S. N. Bose National Centre for Basic Sciences
ब्लॉक - जे. डी., सेक्टर-III, सॉल्ट लेक, कोलकाता-106
Block - JD, Sector-III, Salt Lake, Kolkata-106

ब्लॉक – जे.डी., सेक्टर –III, सॉल्ट लेक,कोलकाता– 700 106 / Block - JD, Sector - III, Salt Lake, Kolkata - 700 106
दूरभाष / Phones : (00) 91 - (0) 33 - 2335 5706-8, 2335 3057 / 61, 2335 0312 / 1313
टेलीफैक्स / TELEFAX : +91 -33-2335 3477 / 2335 1364 / 2335 9176
वेबसाइट/ Website:http://www.bose.res.in
भारत सरकार के विज्ञान एवं प्रौद्योगिकी विभाग के अंतर्गत एक स्वायत्त संस्थान
AN AUTONOMOUS INSTITUTE UNDER DEPARTMENT OF SCIENCE & TECHNOLOGY, GOVERNMENT OF INDIA

Abstract

Non-canonical structures in biomolecules, such as DNA, RNA and proteins, play crucial roles in various cellular processes. Non-canonical Hoogsteen (HG) base pairing is an alternative of the canonical Watson-Crick (WC) base pairing in DNA double helices. Similarly, certain proteins, such as amyloid-beta (Aβ) and α-synuclein (αS), lack a well-defined three-dimensional structure even in physiological conditions. These proteins are known as intrinsically disordered proteins (IDPs). Despite their biological significance, non-canonical biomolecular structures remain less understood in the literature. We use molecular dynamics (MD) simulations to investigate the stability of non-canonical biomolecules from various perspectives at the microscopic level.

Experimental and simulation studies have shown that the majority of HG bps are stabilized in DNA–protein complexes. The molecular mechanism underlying this stabilization is not well known. We use conformational thermodynamics calculation to examine the conformational free energy and entropy changes due to the presence of a HG bp in the DNA sequences of the PDB id: 1K61 system rather than all WC bps. We observe that the presence of the homeodomain proteins in this system stabilizes and orders more the HG bp as well as the whole DNA duplex. This stabilization is primarily driven by reduced fluctuations in sugar-phosphate, sugar-base and sugar-pucker torsion angles due to protein interactions.

We compute the binding energy (B_E) of proteins with both HG-DNA and WC-DNA across various DNA–protein systems, including 1K61 using Umbrella sampling technique and correlate these B_E values with conformational thermodynamics data at the protein–DNA interface. We observe that a strong positive correlation is present there. Additionally, we use the Well-tempered metadynamics technique to compute the free energy surface for the WC $\leftrightarrow$ HG bp transition in the DNA duplex, both in the absence and presence of proteins. We find that in the absence of proteins, the WC bp is approximately 3 kcal/mol more stable than the HG bp, whereas in the presence of proteins, the HG bp becomes about 3 kcal/mol more stable than the WC form. The transition between these bp states involves energy barriers significantly higher than thermal energy.

We next examine proteins deviating from their native structures. α-Synuclein (αS) is a well-known intrinsically disordered protein (IDP) and its aggregation is a hallmark of Parkinson's disease. The NACore segment (68GAVVTGVTAVA78) within the non-amyloid-β component (NAC) region (residues 61–95) is a key driver of fibril formation. Recent experimental studies suggest that zinc oxide nanoparticles (ZnONPs) reduce the rate of amyloid formation of αS. We observe from our simulation that the conformational fluctuations of NACore residues decrease in the presence of ZnONPs. Using the XGBoost machine learning technique, we find the most essential coordinate governing the system dynamics is located far from the NACore region in the presence of ZnONPs. These suggest that ZnONPs may reduce the aggregation rate of αS by decreasing the conformational fluctuations of the NACore residues. To check the aggregation phenomena require at least two αS molecules to simulate for longer time. To achieve this we need coarse-grained (CG) simulation. Current CG model can't capture structural changes. In our polymer bead-spring model the polymer beads are taken to have additional degrees of freedom corresponding to backbone dihedral angles. We perform Monte Carlo (MC) simulation, where the interaction between single αS and ZnONP are based on data from our all-atom system. We observe that our CG model captures the protein structural information in agreement with the all-atom MD results.

Our findings on non-canonical bps and IDP protein provide valuable insights for designing stable non-canonical bimolecular complexes which have major implication in various cellular processes.

List of publications

[1] **Kanika Kole**, Aayatti Mallick Gupta and Jaydeb Chakrabarti, Conformational stability and order of Hoogsteen base pair induced by protein binding. Biophys. Chem. **2023**, *301*, 107079.
(https://doi.org/10.1016/j.bpc.2023.107079)

[2] **Kanika Kole** and Jaydeb Chakrabarti, Binding of Homeodomain proteins with Hoogsteen base pair. J. Phys. Chem. B **2025**, *129*, 1544-1554.
(https://doi.org/10.1021/acs.jpcb.4c08054)

[3] **Kanika Kole**, Aayatti Mallick Gupta and Jaydeb Chakrabarti, The effect of protein binding on the conformational stability and order of DNA duplex with Hoogsteen base pairing. AIP conf. proc. **2025**, *3198*, 020005.
(https://doi.org/10.1063/5.0248431)

[4] **Kanika Kole** and Jaydeb Chakrabarti, Free energy landscape for Watson-Crick to Hoogsteen base pair transition. (Under preparation)

[5] **Kanika Kole** and Jaydeb Chakrabarti, Conformational stability of α-synuclein in presence of zinc oxide nanoparticles. (Under preparation)

[6] **Kanika Kole**, Abhik Ghosh Moulick and Jaydeb Chakrabarti, Coarse-grained model of protein with structural information. (Under preparation)

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Non-canonical structures in biomolecules, such as DNA, RNA and proteins, play crucial roles in various cellular processes. Non-canonical Hoogsteen (HG) base pairing is an alternative to the canonical Watson-Crick (WC) base pairing in DNA double helices[1] (Fig. 1.1). Similarly, certain proteins, such as amyloid-beta (Aβ) and α-synuclein (αS), lack a well-defined three-dimensional structure even in physiological conditions. These proteins are known as intrinsically disordered proteins (IDPs)[2] (Fig. 1.2). Despite their biological significance, non-canonical biomolecular structures remain less understood in the literature.

A specifying feature of DNA is the canonical WC base pairing for adenine(A)-thymine(T) and guanine(G)-cytosine(C). However, these bases can adopt an alternative base pairing mode known as HG base pairing, in which the purine base adopts a syn conformation instead of the anti conformation seen in WC pairs, leading to hydrogen bonding through the HG edge rather than the WC edge[3] (Fig. 1.1). This structural shift results in distinct base pair (bp) geometries (Fig. 1.1), which have been observed in DNA duplexes under physiological conditions.[4] HG bps have significant biological implications, particularly in DNA-protein interactions,[5] gene regulation,[6] DNA repair[5] and drug-DNA interactions.[7,8] Experimental and earlier simulation studies have shown that HG bps are stabilized in DNA-protein complexes.[9] However, the microscopic mechanisms underlying the protein induced stability of HG bps have not yet been investigated.

A protein in its native form typically has a fixed well-defined three-dimensional structure, called globular protein. Instead of adopting a single well-defined conformation like globular proteins, IDPs exist as dynamic ensembles of conformations, allowing them to be highly flexible and functionally versatile.[10,11] This flexibility allows them to play crucial roles in cell signaling, regulation

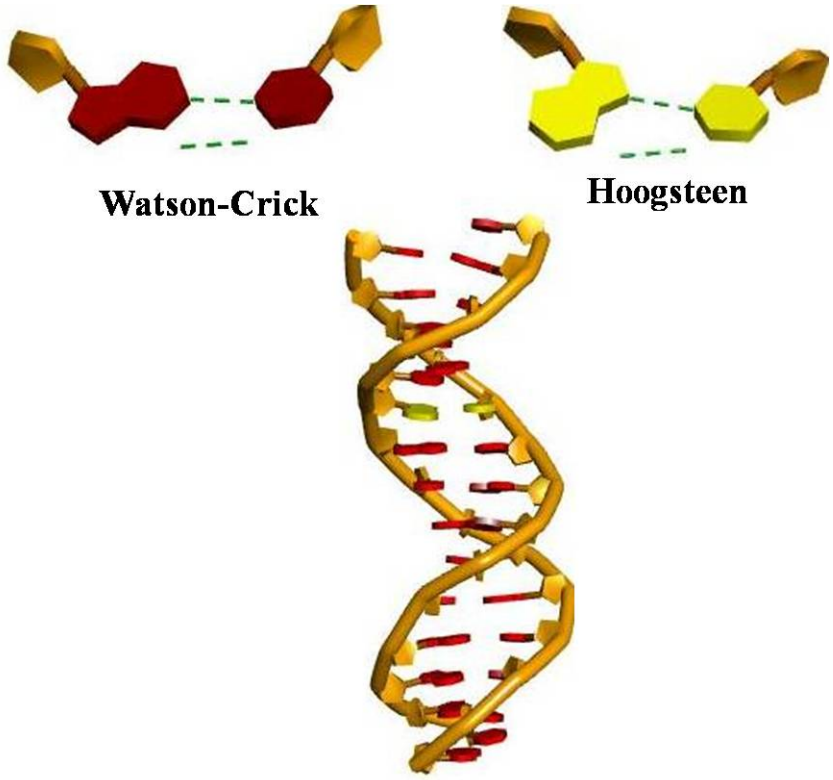


Figure 1.1: Watson-crick and Hoogsteen base pairs within a DNA double helix. The Watson-Crick base pairs are shown in red and Hoogsteen base pairs are shown in yellow.

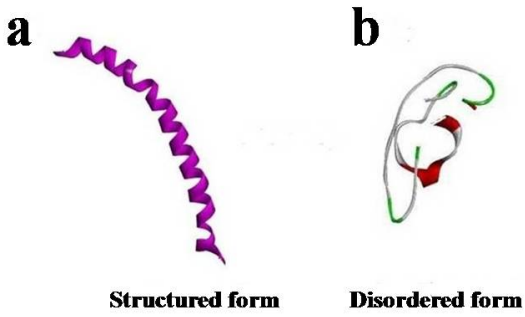


Figure 1.2: Cartoon representations of (a) the structured form and (b) the disordered form of a small fragment of the intrinsically disordered protein (IDP) α-synuclein.

and molecular recognition.[11] IDPs are not inherently toxic. Sometimes, their dynamic and flexible nature make them prone to misfolding and aggregation leading to pathological conditions.[12,13]

The primary objective of this thesis is to provide a microscopic understanding of DNA bps and proteins beyond their standard canonical forms. We employ MD simulations in all of our studies. We examine the structural characteristics of WC and HG bp as well as the conformational thermodynamics changes due

to the changing of a base pairing mode from WC to HG in both naked and protein-DNA duplexes (PDB ID: 1K61),[14] with focus of their interactions with homeodomain proteins. We observe that the presence of proteins stabilized and ordered more the HG bp as well as the whole DNA duplex in the 1K61 system with respect to the all WC bps containing naked DNA system. Then, we examine how conformational stability and order at the protein-DNA interfaces alter the binding energy (B_E) of proteins with DNA duplexes. We observe that higher the conformational stability and order at the interfaces increase the B_E.[15] We investigate the bp transitions between WC and HG pairing modes in both naked and protein-bound DNA duplexes and compute the free energy differences as well as the energy barriers associated with these transformations. We observe that, in the absence of proteins, the WC base pairing region is energetically more favorable than the HG base pairing region, whereas in the presence of proteins, the HG base pairing region becomes energetically more favorable than the WC base pairing region. These transformations require overcoming an energy barrier, the height of which is much greater than the thermal energy.

Now we consider proteins deviating from native structure. α-synuclein (αS) is a well known IDP and its aggregation is a hallmark of Parkinson's disease.[16] Non-Amyloid-β Component (NAC) region (residues 61–95) of αS is the key driver of fibril formation in the Parkinson's disease. Recent an experimental study suggest that ZnO nanoparticle (ZnONP) may alters the conformation of αS and reduce the rate of amyloid formation.[17] We analyze MD simulation trajectories to investigate how the initial conformation of the NACore residues (68GAVVTGVTAVA78) evolves over time. We observe that the conformational fluctuations of NACore residues decrease in the presence of ZnONPs. This suggests that ZnONPs may reduce the aggregation rate of αS by decreasing the conformational fluctuations of the NACore residues.

Fully microscopic simulation calculations using the AA simulation technique are quite expensive on IDP aggregation relevant to pathogenic behavior. This demands for coarse-grained (CG) description of systems. Several CG techniques have been reported in the literature. In most CG protein models reported in the literature, each amino acid's backbone and side chain are represented by some heavy atoms. Most coarse-grained (CG) models[18–22] do not accurately capture the structural information of proteins. Backbone dihedral angles are particularly important for providing structural information about the residues.[23] Motivated by this, we model the system using a bead-spring CG approach, representing the center of mass of each protein residue as a polymer bead and the oxygen

atom of each water molecule as a solvent bead. Each bead is taken to have internal degrees of freedom corresponding to the dihedral angles. Harmonic bond-stretching and angle-bending interactions are used for bonded interactions, while screened Coulomb and Lennard-Jones potentials are applied to describe non-bonded interactions between polymer bead particles. We consider only non-bonded interactions polymer and solvent, as well as among solvent beads. We simulate the system for a longer time using the Monte Carlo (MC) method.[24] This CG method incorporates all the non-bonded, bonded interactions, along with dihedral angle coupling information derived from equilibrated AA data. We apply our CG model to the small proteins GB3 (56 residues, PDB ID: 2OED).[25] Our CG simulation data show strong structural agreement with both the crystal structure and all atom MD data. We now apply this CG approach to the αS protein to examine how ZnONP reduces the fluctuation in the NACore residues of αS. We see that our CG data matches well with the AA MD simulations data.

The rest of the Chapter is organized as follows: In Section 1.1, we describe how the presence of homeodomain proteins influences the conformational stability and order of Hoogsteen base pair, as well as the overall DNA duplex. In Section 1.2, we examine how the binding energy of homeodomain proteins correlates with conformational thermodynamic data at the protein–DNA interfaces, considering both Watson-Crick and Hoogsteen DNA systems. Section 1.3 describes the free energy landscape associated with the transition between Watson–Crick and Hoogsteen base pairing in both naked and protein-bound DNA complexes. In Section 1.4, we explore the effect of Zinc oxide nanoparticles on the conformational changes of the α-synuclein. Section 1.5 describes the coarse-grained representation of proteins, including structural information.

1.1 Conformational stability and order of Hoogsteen base pair induced by homeodomain proteins binding

Several experimental and simulation studies have shown that HG bp mostly occur in regions of DNA that are highly distorted by bound protein.[14,15] In addition, a HG bp has also been observed within undistorted regions of DNA in the presence of homeodomain proteins.[9] Previous experimental and simulation studies have shown that protein binding stabilizes the HG bp.[9] The molecular mechanism

underlying this stabilization is not well known. It is in general difficult to analyze HG bp in protein-bound DNA structures by X-ray crystallography due to unclear electron density and by nuclear magnetic resonance (NMR) spectroscopy due to inadequate chemical shift dispersion. In this context, MD simulation is a powerful tool for investigating the stability of HG bps within protein-bound DNA complexes.

With this backdrop we investigate the stability of the HG bp in an undistorted DNA duplex of PDB id: 1K61 using AA MD simulations, both in the absence and presence of of specific and non-specific homeodomain proteins. Equilibrated MD snapshots of naked DNA containing all WC bps, the duplex DNA without proteins containing one HG bp and the remaining WC bps and DNA duplex with homeodomain proteins, containing one HG bp and the remaining WC bps (PDB id: 1K61) are shown in Fig. 1.3 (a), (b) and (c), respectively. We used a conformational thermodynamics calculation[14,15,26–31] to examine the conformational stability of HG bp with respect to WC bp within duplex DNA both in the absence and presence of proteins at the molecular level. The changes in conformational stability and order in bio-molecular systems are quantified in terms of conformational free energy and entropy changes derived from the fluctuations of microscopic conformational variables. In our study, the microscopic conformational variables are all bp parameters, intra and inter, sugar-phosphate, sugar-base and sugar-pucker dihedral angles. We also calculate the changes in conformational free energy and entropy of the entire DNA duplex containing a HG bp, both in the absence and presence of proteins, relative to the naked DNA duplex containing all WC bps, using equilibrated MD trajectories.

We observe that bound proteins enhance the stability and order of the HG bp, as well as the whole DNA duplex of the 1K61 system.[14] The sugar-phosphate, sugar-base and sugar-pucker torsion angles play crucial roles in stabilizing and ordering the HG bp, as well as the entire DNA duplex in the protein-bound 1K61 complex, by reducing fluctuations.[14]

1.2 Binding Energy of Homeodomain Proteins to DNA containing Hoogsteen Base Pair

In the previous section, we check the conformational stability and order of the HG bp as well as the whole DNA duplex in the 1K61 system. There, we do not focus on the binding energy (B_E) of both specific and non-specific homeodomain

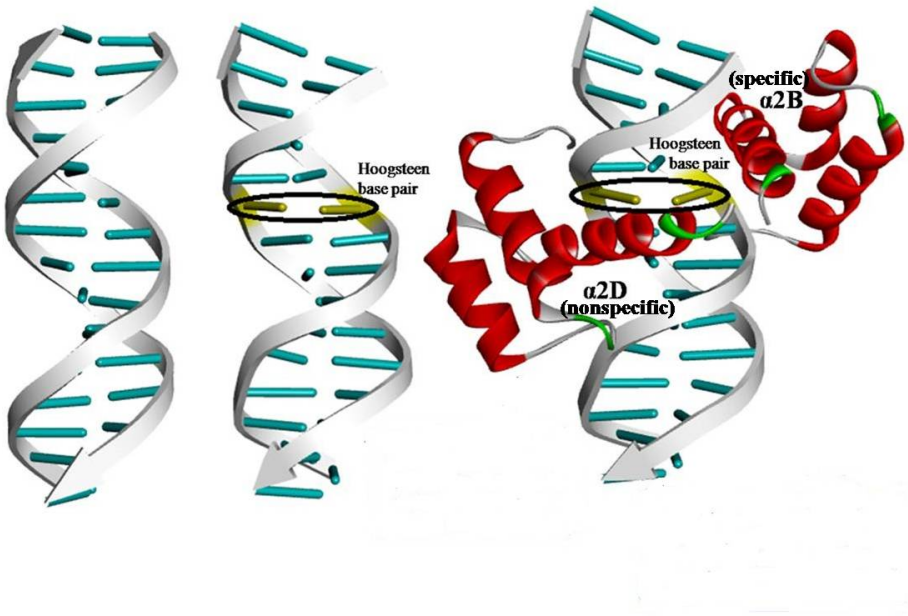


Figure 1.3: Equilibrated MD snapshots of (a) naked DNA containing all Watson-Crick (WC) bps, (b) the duplex DNA without proteins containing one HG bp and the remaining WC bps and (c) DNA duplex with homeodomain proteins, containing one HG bp and the remaining WC bps. System (a) is taken from the model structure, while (b) and (c) are taken from PDB ID: 1K61.

proteins, nor on the conformational stability and order of these proteins in the DNA-bound state compared to the unbound state. The B_E of proteins with WC-DNA is reported,[32,33] but the B_E of proteins with HG-DNA is not reported so far.

We compute the B_E of both the homeodomain proteins in this 1K61 system using the Umbrella Sampling (US) technique.[15,34–36] We also compute the conformational free energy and entropy changes at the DNA-proteins interfaces. The overall study of the protein-bound 1K61 system in this section is illustrated by a cartoon representation in Fig. 1.4. We find that the B_E of both specific and non-specific proteins are almost equal. Both proteins bind more strongly when the DNA duplex contains a HG bp in its sequence, compared to when it contains all WC bps. The neutral mutation in the N-terminal basic arm of the nonspecific protein significantly changes the B_E between nonspecific protein only, while the acidic mutation significantly changes the B_E of both the specific and nonspecific proteins with the HG bp containing DNA duplex of this 1K61 system.

The significant variation in the B_E of the homeodomains within DNA–protein complexes can be ascribed to the differences in the number of intermolecular contacts between DNA bases and protein residues. We compute the conformational free energy and entropy at the protein–DNA interfaces using the microscopic

Figure 1.4: A cartoon representation of the overall study on the protein-bound 1K61 system in the second section. Stabilized and ordered residues are highlighted in green in the protein-bound state.

degrees of freedom ϕ, ψ, and $\chi 1$ dihedral angles of protein residues and all microscopic degrees of freedom of the DNA bases, as described in the previous section. Our conformational thermodynamics calculations, based on fluctuations of microscopic conformational variables at the interface, reveal that enhanced conformational stability and order at the interface are correlated with stronger binding of the homeodomain protein.

1.3 Free energy landscape of transitions between Watson–Crick and Hoogsteen base pairing

DNA breathing are local conformational fluctuations spontaneously occurring in double-stranded DNA duplexes.[37] In particular, individual bps in duplex DNA can flip between alternative pairing modes, WC and HG on biologically relevant timescales. The transition is driven between anti (WC) and syn (HG) conformational changes of the glycosidic torsion angle [χ (O4'-C1'-N9-C4)] of the purine base.[15] DNA breathing in the presence of proteins has not been studied so far.

We examine DNA breathing in both naked DNA and protein-bound DNA systems. A cartoon representation of the DNA-breathing study, involving both naked DNA and DNA-protein systems, is shown in Fig. 1.5. We first examine DNA breathing in naked DNA containing all WC bps (WC-DNA). We use Well-Tempered Metadynamics (WT-MetaD)[38] calculations, employing χ as the collective variable (CV). We construct the free energy landscape to calculate the

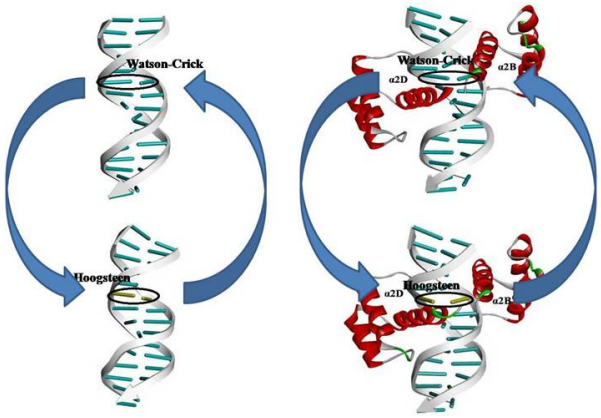


Figure 1.5: A cartoon representation of the DNA-breathing study involving naked DNA and DNA-protein systems.

relative free energy differences between the two base pairing modes and the height of the barrier for this transition. These results match well with previous experimental studies[39] and simulation studies.[37]

Next, we examine the HG to WC bp transition in the 1K61 system in the absence of proteins. We observe that the WC bp has a relatively lower free energy than the HG bp, similar to the transition observed in WC-DNA, although the barrier height decreases. Then, we investigate the transition from WC to HG base pairing in WC-DNA in the presence of both specific and non-specific homeodomain proteins. We observe a striking difference compared to the absence of proteins, the HG bp exhibits a lower free energy than the WC bp, and the barrier height decreases. Next, we explore the HG to WC bp transitions in the presence of both homeodomain proteins of the 1K61 system. Here, the HG base pairing resides at the global minimum, while the WC base pairing is at a metastable minimum. The barrier height for the HG to WC bp transition increases in the absence of proteins. Strong interfacial interactions between the HG base and the non-specific protein likely stabilize the HG bp, resulting in a higher energy barrier for the HG to WC transition in the 1K61 system. Our study may complement and guide future high-resolution experiments addressing many outstanding issues of duplex DNA breathing in protein-DNA complexes.

1.4 Impact of zinc oxide nanoparticles on the conformational stability of the α-synuclein

The aggregation and fibril formation of the intrinsically disordered protein α-synuclein are associated with Parkinson's disease.[16,17] Within its non-amyloid β component (NAC) region, spanning residues 61–95, an 11-residue segment known as NACore (68GAVVTGVTAVA78) plays a crucial role in fibril as well as amyloid formation and cytotoxicity.[40] Several studies have explored effective molecules for inhibiting harmful amyloidogenesis.[17] However, the detailed mechanism of such inhibition is yet largely unknown.

We investigate the conformational changes of α-synuclein, particularly in the NACore region, in the presence of zinc oxide nanoparticle (ZnONPs) and their impact on cytotoxicity using AA MD simulations. Experimentally it is reported that the N-terminus is rich in lysine residues and a part of the consensus hexameric (KTKEGV) motif and could be responsible for establishing strong electrostatic interactions between α-synuclein and ZnONP interface.[17] The majority of the interface-forming residues are consistent with those identified through experimental studies. All the NACore residues adopt helix conformation in the crystal. Most of the NACore residues retain their helix secondary structure[41] more effectively throughout the simulation trajectory in the presence of ZnONPs, which help stabilize and order this region with respect to the absence of ZnONPs. Fig. 1.6 is the representation of the secondary structure of the αS molecule in absence and in presence of neutral ZnONP. Our essential coordinate (EC) analysis reveals that, in the presence of ZnONPs, the most essential coordinate is located away from the NACore region.

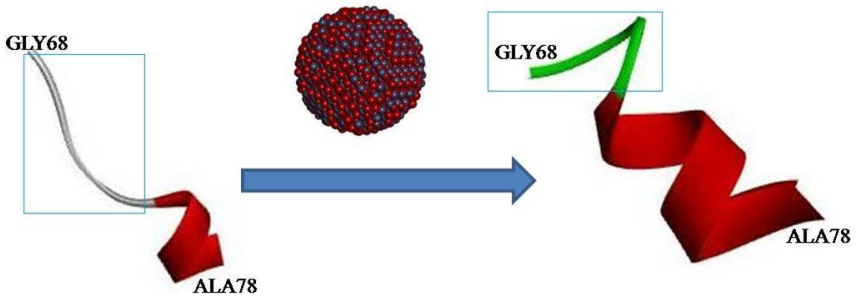


Figure 1.6: A representation of the secondary structure of the 68GAVVTGVTAVA78 (NACore) residues of α-synuclein in the absence and presence of a neutral zinc oxide nanoparticle.

1.5 Coarse-grained representation of proteins, including structural information

Many biologically relevant processes occur time scales that are beyond the reach of atomistic simulations.[22,42] These processes include, in particular, the self-assembly of bio-molecules.[20] Lowering the degrees of freedom from AA to CG model shows possibility to study systems involving large numbers of bio-molecules where typically a group of chemical moieties are represented by a single entity.[18] CG molecular modeling allows computer simulations on length and time scales 2–3 orders of magnitude larger than atomistic simulations, bridging the gap between atomistic and mesoscopic scales. It is known that protein functionality largely depends on its structure.[42] However, structural information has not been properly addressed in the most of the literature reported CG models of proteins.[18–22]

We develop a simple CG protein model representing center of mass of each residue by a polymer bead and water oxygen as a solvent bead. Each polymer bead incorporates two additional degrees of freedom corresponding to the backbone dihedral angles. An all-atom to coarse-grained representation of protein amino acid residues is shown in Fig. 1.7. In the model, bead–bead interactions are described by bonded terms: harmonic potentials for bond stretching and angle bending and nonbonded terms, including screened Coulomb and Lennard-Jones interactions. All the bonded and non-bonded interaction parameters are computed from the equilibrated AA MD trajectories. The energy cost of dihedral fluctuations is obtained from the negative logarithmic of the joint distribution of the fluctuations of the backbone dihedral in AA simulations. We include only non-bonded interactions between the polymer and solvent and among solvent beads. Monte Carlo (MC) method is used to explore the phase space. The initial coordinates and dihedral angles for each polymer bead are taken from the protein crystal structure data. When constructing a MC move, we calculate the energy cost due to different interactions like bonded, non-bonded and solvent contributions. We also change the dihedral degrees of freedom and calculate the associated energy cost using potentials derived from AA simulations.

We observe that using such model we can reproduce well the protein structure of GB3 compared to the crystal structure and AA data. We use all the interaction parameters and dihedral coupling interactions of GB3 for other proteins to check transferability of the model. We find that the protein homeodomain

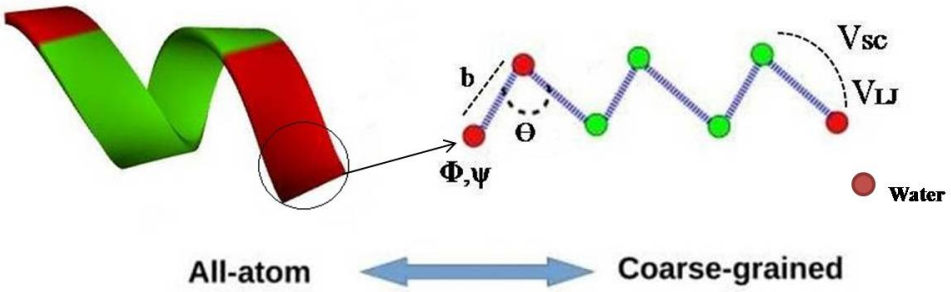


Figure 1.7: All atom to coarse-grained representation of protein amino acid residues. Red corresponds to hydrophilic and green corresponds to hydrophobic amino acid. Beads are connected via harmonic spring, corresponding the bond stretching (b) and bond angle (θ). V_{SC} and V_{LJ} are corresponding screen-Coulomb and Leneard-Jones type non-bonded interactions between polymer beads. ϕ and ψ correspond to the backbone dihedral angles of polymer bead. As hydrophilic amino acids favour the water, so, both the oxygen atom of water and hydrophilic amino acid are represented in red.

reproduces the structural information well, both with AA simulation data and crystal structures. We apply our CG simulations to αS. Here, we use lower bonded force constants, while the other interaction parameters and dihedral coupling interactions are taken from the GB3 AA simulation data. We model the interaction of αS with neutral ZnONP by immobilizing the binding residues of the protein with the nanoparticle obtained from AA trajectory. We observe that the majority of the NACore residues are stabilized and ordered in presence of ZnONP, matching well with the AA simulation data.

The organization of the rest of the thesis is as follows: In Chapter 2, we discuss the conformational stability and ordering of the Hoogsteen base pair-containing DNA-protein system from a microscopic perspective. Chapter 3 explores how conformational stability and order influence the binding energy between proteins and DNA, considering both Hoogsteen and Watson-Crick DNA structures. Chapter 4 provides a detailed study of DNA breathing in both naked and protein-bound Watson-Crick and Hoogsteen DNA systems. In Chapter 5, we examine how zinc oxide nanoparticles stabilize and order the aggregation core of the α-synuclein molecule. Coarse-grained model of protein along with structural information in terms of dihedral angle is discussed in Chapter 6. Finally, Chapter 7 contains the conclusion and outlook of the thesis.

Bibliography

[1] Yang, C.; Kim, E.; Pak, Y. Free energy landscape and transition pathways from Watson–Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Res.* **2015**, *43*, 7769-7778.

[2] Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.

[3] Hoogsteen, K. Crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.* **1963**, *16*, 907–916.

[4] Nikolova, E. N.; Zhou, H.; Gottardo, F. L.; Alvey, H. S.; Kimsey, I. J.; Al-Hashimi, H. M. A historical perspective on the role of Hoogsteen base pairs in nucleic acids. Biopolymers **2011**, *99*, 955–968.

[5] Shi, X.; Cui, W.; Pei, H. Hoogsteen base pairing in DNA-protein interactions and its role in genetic regulation. *Nat. Commun.* **2018**, *9*, 2361.

[6] Zhou, H.; Kimsey, I. J.; Nikolova, E. N.; Al-Hashimi, H. M. Hoogsteen base pairs: Expanding the repertoire of DNA recognition. *Chem. Rev.* **2019**, *119*, 7766–7786.

[7] Xu, Y.; McSally, J.; Andricioaei, I.; Al-Hashimi, H. M. Modulation of Hoogsteen dynamics on DNA recognition. *Nat. Commun.* **2018**, *9*, 1473.

[8] Nikan, M.; Shen, W.; Ma, H. Overcoming challenges in drugging structured nucleic acids: Hoogsteen base pairs as novel drug targets. *Trends Pharmacol. Sci.* **2016** *37*, 538–551.

[9] Aishima, J.; Gitti, R. K.; Noah, J. E.; Gan, H. H.; Schlick, T.; Wolberger, C. A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.* **2002**, *30*, 5244-5252.

[10] Oldfield, C. J.; Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **2014**, *83*, 553-584.

[11] Wright, P. E; Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18-29.

[12] Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **2011**, *21*, 432-440.

[13] Chiti, F.; Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **2017**, *86*, 27-68.

[14] Kole, K.; Gupta, A. M.; Chakrabarti, J. Conformational stability and order of Hoogsteen base pair induced by protein binding. *Biophys. Chem.* **2023**, *301*, 107079.

[15] Kole, K.; Chakrabarti, J. Binding of Homeodomain Proteins to DNA with Hoogsteen Base Pair. *J. Phys. Chem. B* **2025**, *129*, 1544-1554.

[16] Uversky, V. N. Alpha-synuclein misfolding and neurodegenerative diseases. *Curr. Protein Pept. Sci.* **2008**, *9*, 507–540.

[17] Asthana, S.; Bhattacharyya, D.; Kumari, S.; Nayak, P. S.; Saleem, M.; Bhunia, A.; Jha, S. Interaction with zinc oxide nanoparticle kinetically traps α-synuclein fibrillation into off-pathway non-toxic intermediates. *Int. J. Biol. Macromol.* **2020**, *150*, 68-79.

[18] Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819-834.

[19] Trylska, J. Coarse-grained models to study dynamics of nanoscale biomolecules and their applications to the ribosome. J. Condens. Matter Phys. **2010**, *22*, 453101.

[20] Saunders, M. G.; Voth, G. A. Coarse-graining of multiprotein assemblies. *Curr. Opin. Struct. Biol.* **2012**, *22*, 144-150.

[21] Mukherjee, A.; Bagchi, B. Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. *J. Chem. Phys.* **2003** *118*, 4733-4747.

[22] Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-grained protein models and their applications. *Chem. Rev.* **2016**, *116*, 7898-7936.

[23] Detcheverry, F. A.; Pike, D. Q.; Nealey, P. F.; Müller, M.; de Pablo, J. J. Monte Carlo simulation of coarse grain polymeric systems. *PRL* **2009**, *102*, 197801.

[24] Mooney, C. Z. Monte carlo simulation. Sage **1997**, *116*.

[25] Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J. Am. Chem. Soc.* **2003**, *125*, 9179-9191.

[26] Das, A.; Chakrabarti, J.; Ghosh, M. Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation. *Biophys. J.* **2013**, *104*, 1274-1284.

[27] Das, A.; Chakrabarti, J.; Ghosh, M. Conformational thermodynamics of metal-ion binding to a protein. *Chem. Phys. Lett.* **2013**, **581**, 91-95.

[28] Das, A.; Chakrabarti, J.; Ghosh, M. Thermodynamics of interfacial changes in a protein–protein complex. *Molecular Biosyst.* **2014**, *10*, 437-445.

[29] Sikdar, S.; Chakrabarti, J.; Ghosh, M. A microscopic insight from conformational thermodynamics to functional ligand binding in proteins. *Molecular Biosyst.* **2014**, *10*, 3280-3289.

[30] Mandal, S. C.; Maganti, L.; Mondal, M.; Chakrabarti, J. Microscopic insight to specificity of metal ion cofactor in DNA cleavage by restriction endonuclease EcoRV. *Biopolymers* **2020**, *111*, e23396.

[31] Moulick, A. G.; Chakrabarti, J. Fluctuation-Dominated Ligand Binding in Molten Globule Protein. J. Chem. Inf. Model. **2023**, **63**, 5583-5591.

[32] Yonetani, Y.; Kono, H. Dissociation free-energy profiles of specific and nonspecific DNA–protein complexes. J. Phys. Chem. B **2013**, *117*, 7535-7545.

[33] Jakubec, D.; Vondrasek, J. Can All-Atom Molecular Dynamics Simulations Quantitatively Describe Homeodomain–DNA Binding Equilibria? J. Chem. Theory Comput. **2019**, *15*, 2635-2648.

[34] K¨astner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 932-942.

[35] Hub, J. S.; De Groot, B. L.; Van Der Spoel, D. g-wham-A free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713-3720.

[36] Lemkul, J. A.; Bevan, D. R. Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics. *J. Phys. Chem. B* **2010**, *114*, 1652-1660.

[37] Yang, C. Kim, E. Lim, M. Pak, Y. Computational probing of Watson–crick/hoogsteen breathing in a DNA duplex containing N1-methylated adenine. J. Chem. Theory Comput. **2018**, *15*, 751-761.

[38] Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *PRL* **2008**, *100*, 020603.

[39] Nikolova, E. N. Kim, E. Wise, A. A. O'Brien, P. J. Andricioaei, I. Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. Nature, **2011**, *470*, 498-502.

[40] Hashemi, M.; Lyubchenko, Y. L. Hybrid resolution molecular dynamics simulations of amyloid proteins interacting with membranes. *Methods* **2022**, *197*, 89-96.

[41] Cino, E. A.; Choy, W. Y.; Karttunen, M. Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations. J. Chem. Theory Comput. **2012**, *8*, 2725-2740.

[42] Moulick, A. G. Chakrabarti, J. Microscopic insights to relaxation phenomena in proteins. University of Calcutta, CUE-A18258-T14408, **2024**.

Chapter 2 Homeodomain Proteins as Modulators of Hoogsteen Base Pair Stability *

2.1 Introduction

In double helical DNAs, bp often adopts other geometries than the conventional Watson-Crick (WC) base pairing. Hoogsteen (HG) base pairing is an alternative base-pairing scheme for DNA double helices[1,2] where the purine is flipped $'upside-down'$ such that the 5-ring of the purine forms a hydrogen bond (H-bond) to the pyrimidine, rather than the 6-ring (Fig. 2.1). Non-canonical DNA structures can arise in multiple regions of the human genome.[3] The switch from the canonical WC to the non-canonical HG orientation substantially modifies the chemical environment around the bp, which has major implications for DNA-protein recognition,[4,5] damage repair,[6,7] and replication.[8,9] It is in general difficult to analyze HG bp in protein-bound DNA structures by X-ray crystallography due to unclear electron density and by nuclear magnetic resonance (NMR) spectroscopy due to inadequate chemical shift dispersion. As a result, the stability of HG bp in the presence of particular proteins or ligands is far from understood.

HG bp occurs in regions of DNA that are highly distorted by bound protein.[10] The X-ray structure of TATA-binding protein (TBP) bound to DNA reveals a protonated G-C+ HG bp in the region of DNA under winding and intercalation by TBP's phenylalanine side chain.[11] Furthermore, the X-ray crystal structure of DNA in contact with the p53 tumour suppressor proteins exhibits two consecutive A-T HG bps. Wolberger and co-workers report a single A-T HG bp within

*Based on the publication: Kanika Kole, Aayatti Mallick Gupta and Jaydeb Chakrabarti. Conformational stability and order of Hoogsteen base pair induced by protein binding. Biophysical Chemistry 301, 2023, 107079. (https://doi.org/10.1016/j.bpc.2023.107079)

an undistorted B-form DNA duplex (PDB id-1K61) in the X-ray structure of MATα2 homeodomain bound to DNA.[12] Homeodomain is a 60 amino acid long DNA-binding domain that can bind to both WC-DNA and HG-DNA. The protein consists of three α helices folded around a hydrophobic core. The flexible N-terminal arm of the homeodomain gets ordered due to the binding to DNA.[13] Homeodomains are responsible for sequence-specific recognition of DNA.[14] The homeodomain family of transcription factors regulate cell- and tissue-specific gene expression and play crucial roles in patterning,[15] in particular the Hox genes.[14] Most of them are involved in nervous system development.[14] The crystal structure contains 21 fragments of duplex DNA attached to four MATα2 homeodomains. The DNA shows two binding sites for the four MATα2 homeodomains, two homeodomains, α2A and α2B bind at the specific binding sites, while the other two homeodomains, α2C and α2D bind to DNA non-specifically. Earlier MD simulation studies show that the formation of HG bp in the 1K61 system is more strongly influenced by the non-specifically bound protein α2D.[12] However, there has yet not been any attempt to understand the microscopic mechanism underlying the protein induced stability of the HG bp. Here we study this system to focus on the microscopic mechanisms to stabilize the HG bp in the 1K61 system in the presence of proteins. This system is particularly simple since there is no additional distortion in the DNA structure.

The changes in conformational stability and order in bio-molecular systems are quantified in terms of conformational free energy and entropy changes derived from the fluctuations of microscopic conformational variables.[16–25] Negative changes in conformational free energy and entropy indicate structural stability and ordering respectively in a given structure with respect to a reference structure, while positive changes indicate conformational destabilization and disorder respectively. Although originally developed for proteins, the conformational thermodynamics calculations are extended to DNA-protein systems as well.[17] The fluctuations of microscopic conformational variables can be captured via long molecular dynamics (MD) simulation of the DNA duplex. The microscopic conformational variables of DNA, described in SI, are: (i) inter-bp step parameters: shift (D_x), slide (D_y), rise (D_z), tilt (τ), roll (ρ) and twist (ω) (ii) intra-bp parameters: stagger (S_x), shear (S_y), stretch (S_z), buckle (κ), propeller (π) and opening (σ) (iii) sugar-phosphate and sugar-base backbone torsion angles: α, β, γ, δ, ϵ, ζ and χ, and (iv) sugar-puckers: $\nu0$, $\nu1$, $\nu2$, $\nu3$ and $\nu4$. We study three different systems: (1) WC, a model system of DNA, (2) HG, HG bp containing DNA, and (3) HGP, protein-bound HG bp containing DNA. We compute conformational

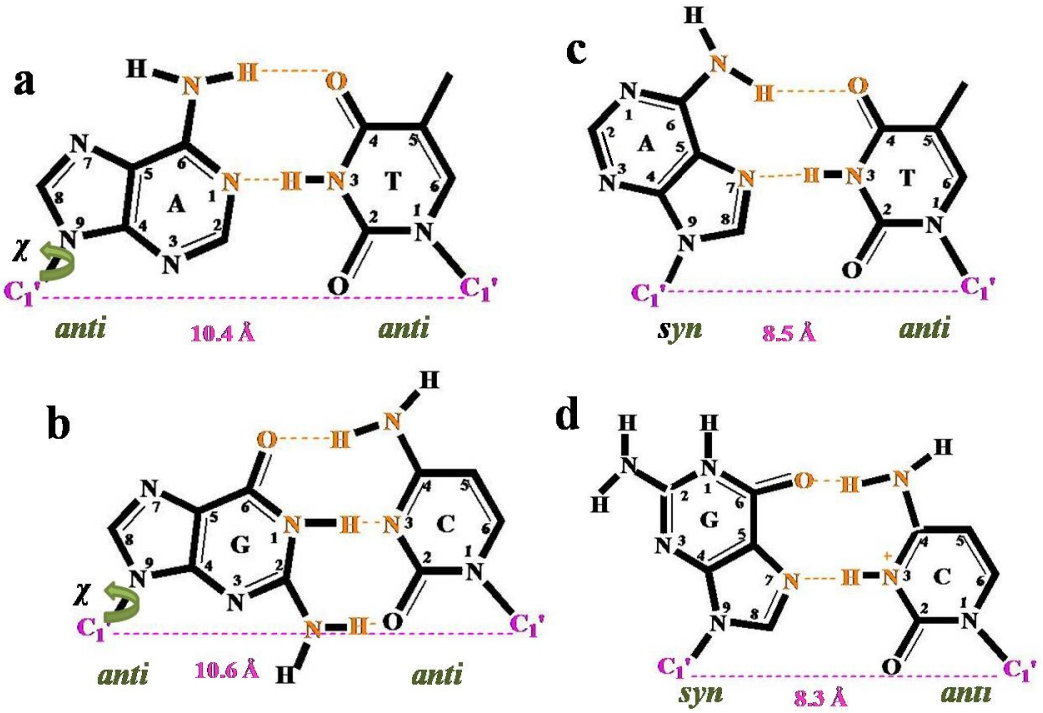


Figure 2.1: Chemical structure of (a) A-T and (b) G-C WC bp, (c) A-T and (d) G-C$^+$ HG bp. Heavy atoms involved in the WC and HG type H-bond are highlighted by orange. Average C1′- C1′ distance for each bp type is highlighted by pink. The χ angle is highlighted in green.

thermodynamics data of the DNA: (i) in the HG system and (ii) in the protein-bound HGP system, both with respect to the WC system. We calculate the total changes in conformational free energy and entropy for entire DNA systems. We observe that the flexibility of the bare DNA duplex is increased by the HG bp. In the presence of a specific protein α2B and a non-specific protein α2D, the whole DNA duplex becomes stabilized and ordered. The backbone torsion angles of bps fluctuate less to stabilize and organize the HG bp as well as the entire DNA duplex in the protein-DNA complex.

2.2 Methods

2.2.1 System preparation

The HGP system used for all-atom MD simulations is 15 bp DNA fragments or DNA with one specifically bound, α2B and one nonspecifically bound, α2D homeodomain taken from PDB ID 1K61. The numbering of phosphates, sugars, and bases as taken in our study is shown in Fig. 2.2. The HG system is the

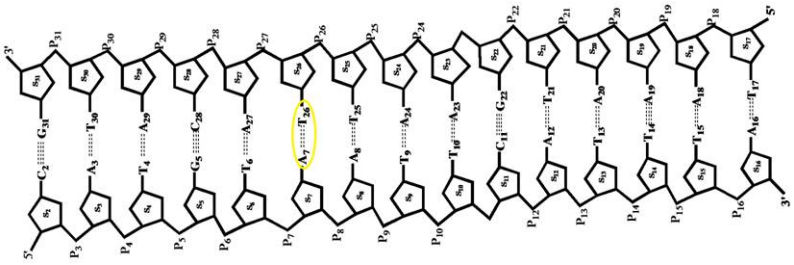


Figure 2.2: Notations and numbering of DNA bps. The HG bp is shown in yellow.

DNA fragment, knocking out all the proteins in 1K61. The A_7-T_{26} bp in the molecular structure is a HG bp. In the HGP system, this HG bp is found in the contact overlapping sites of a specifically (α2B) and a non-specifically (α2D) bound homeodomain. So, the DNA base sequence for each structure is $5'C2ATGT\underline{A}ATTCATTTA163'$ in the $5' - 3'$ direction, where the underlined A7 forms HG bp with the complementary base T26 in the $3' - 5'$ direction. The rest of the bases form WC pairing with the complementary sequence. The model structure (WC) of the DNA double helix for the same sequence containing all WC bps has been generated from the fiber-diffraction model of B-DNA[26] using the SCFBio server (`http://www.scfbio-iitd.res.in`).

2.2.2 MD simulations

We use all atom molecular dynamics (MD) simulation approach to simulate each system. We use Leap Frog algorithm to integrate the equation of motion (Appendix 2.5.1). We take the force-field parameters (Appendix 2.5.2) of the Amberff14SB_OL15[27,28] for DNA and proteins. The GROMACS 2018.6[29] package standard protocol is used to prepare the systems. The system for simulation is solvated in a cubic water box with a box size of $8.1 \times 8.1 \times 8.1$ nm^3, containing TIP3P water molecules. The periodic boundary conditions (PBC) (Appendix 2.5.3) are imposed in all directions. To minimize the end effect, the $3'$ end of the DNA strand is connected to its periodic image along the z axis (image patch), to mimic an infinite DNA chain.[30] The system is neutralized by adding the required number of sodium (Na^+) and chloride (Cl^-) ions. Subsequently, the system is energy minimized using the steepest descent method.[31] Then all atom MD simulation is carried out at 300 K and 1 atmospheric pressure in an isothermal-isobaric (NPT) ensemble starting from the energy minimized structure. Berendsen thermostat[32] (Appendix 2.5.4) is used to maintain a constant temperature, and the pressure is controlled by Parrinello-Rahman barostat[33]

(Appendix 2.5.5). The Lennard-Jones and the short-range electrostatic interactions are truncated at 10 Å. We employ the particle mesh Ewald (PME) (Appendix 2.5.6) summation method for long-range electrostatic interaction with 1 Ågrid spacing and 10^{-6} convergence criterion. The LINCS constraints[34] are applied to all bonds involving hydrogen atoms. The simulation time step of 2 fs is used for all calculations. We perform MD simulations for the systems (a) WC, (b) HG and (c) HGP, for 1μs. To make the ensembles comparable, we adjust the total number of water molecules so that the numbers of total atoms are the same (N=51068) in all the cases. Various analyses are carried out on the equilibrated portion of the trajectories with tools in GROMACS to examine the system properties. The NUPARM with BPFIND software is used to compute various structural parameters of DNA at intervals of 10 ps.[35]

2.2.3 Structural parameters of the DNA bps

We compute the DNA structural parameters (as described in Appendix 2.5.7) for each of the bp for each system. We divide the equilibrated part of the trajectory into N_w (=10) windows with an equal number of conformations from equilibrated conformations and calculate the averaged value of each parameter, θ over each of the windows. We denote the mean value of a parameter, θ in the S-th system (S = WC, HG and HGP) by $\langle\theta^S\rangle$. The error ($\pm 0.5\Delta\theta^S$) in the mean is given by $\pm 0.5\sigma_m/\sqrt{(N_w)}$, where σ_m is the standard deviation of the mean values over the windows. The pseudo rotational phase angle P, which is calculated, based on endocyclic torsion angles ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$) inside the ribose sugar ring,[36] is described in the Appendix 2.5.7 section. All the structural parameters of bps are calculated, ignoring the terminal bps at each end of the DNA duplex.

2.2.4 Root mean square fluctuation

We calculate the root mean square fluctuation (RMSF) of the atomic positions of A_7 and T_{26} bases within the equilibrated portion of the trajectories utilizing the RMSF calculation module of GROMACS.[37]

2.2.5 Conformational thermodynamics

We compute the histograms of all the microscopic variables using the equilibrated trajectory of WC, HG and HGP systems. The normalized probability distribution of any conformational variable, θ in WC, HG, and HGP systems, are given by $H^{HG}(\theta)$, $H^{WC}(\theta)$ and $H^{HGP}(\theta)$, respectively. A detailed description of the

histogram-based method (HBM) for calculating the conformational thermodynamics is reported.[16,38] The changes in free energy of any degree of freedom, θ for each bp of the HG system with regard to its corresponding bp in the WC system and each bp of the HGP system with respect to each bp of WC system are defined as,

(i) $\Delta G^{HG/WC}$ (θ) = -k_BT$\ln\frac{H^{HG}_{max}(\theta)}{H^{WC}_{max}(\theta)}$

(ii) $\Delta G^{HGP/WC}$ (θ) = -k_BT$\ln\frac{H^{HGP}_{max}(\theta)}{H^{WC}_{max}(\theta)}$

Where $'max'$ represents the maximum of the histogram and k_B is the Boltzmann constant. The change in conformational entropy of a given conformational variable, θ is evaluated from Gibbs formula,

(i) T$\Delta S^{HG/WC}$ (θ) = -k_BT[$\sum_i$ $H_i^{HG}(\theta)$ln$H_i^{HG}(\theta)$ - $\sum_i$ $H_i^{WC}(\theta)$ln$H_i^{WC}(\theta)$]

(ii) T$\Delta S^{HGP/WC}$ (θ) = -k_BT[$\sum_i$ $H_i^{HGP}(\theta)$ln$H_i^{HGP}(\theta)$ - $\sum_i$ $H_i^{WC}(\theta)$ln$H_i^{WC}(\theta)$]

Here H_i^S (θ) stands for the i^{th} bin of the histogram for θ in the S^{th} case (= WC, HG, HGP).

2.2.6 Interfacial interactions in the HGP system

The interface is identified when two atoms, one from DNA bases and the other from protein (α2D or α2B) residues, are within 0.6 nm of each other. The calculation is performed over the equilibrated portion of the HGP simulated ensemble's trajectory.

Distance and angle criteria are used to characterize the H-bonds. For H-bonds, the distance between the donor (D) and acceptor (A) atoms is $\leq$ 0.35 nm and donor-hydrogen-acceptor ($D-\hat{H}-A$) antecedent angle cut off is $\geq 150^{\circ}$. The distance between any of the nitrogen atoms of basic residues of protein and any of the oxygen atoms of the phosphate group of the $5'-3'$ strand is used to compute the salt bridge and the cutoff angle conditions are the same as for hydrogen bond interactions. Similarly, the distance and angle between any nitrogen atom from a protein basic residue and any oxygen atom from the phosphate group of the $3'-5'$ strand are used to explore the salt bridge interaction for the $3'-5'$ strand. If the distance is $\leq$ 0.4 nm and the angle is $\geq 150^{\circ}$, the pair is considered a salt bridge. Ionic species belonging to the DNA base and the protein residue at the interface are considered for electrostatic interactions (attractive charge). When the distance between oppositely charged atoms is less than 0.56 nm, the pair is counted for electrostatic interaction. We utilize GROMACS H-bond analysis module[39] for H-bonds and Discovery studio software[40] for salt bridge and electrostatic

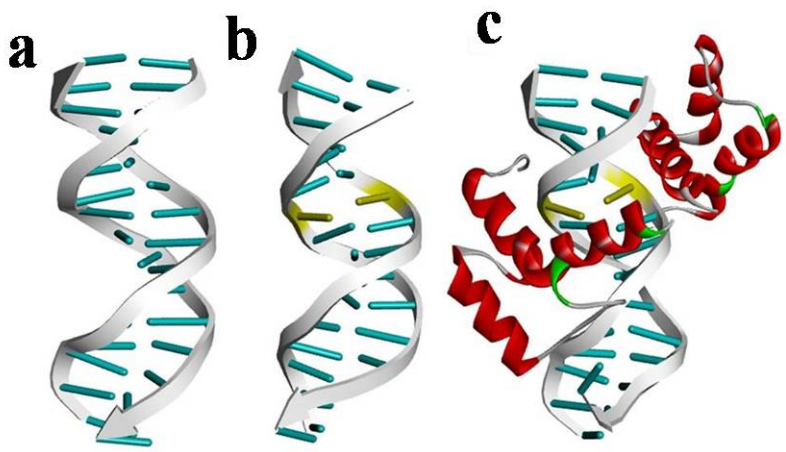


Figure 2.3: Snapshot at 500 ns of (a) WC (b) HG, and (c) HGP systems. HG bp is shown in yellow.

interaction to define the cutoff criterion.

2.3 Results and Discussions

The WC and HG systems contain 15 bps of DNA alone, whereas the HGP system contains DNA with a specifically bound α2B and a nonspecifically bound α2D homeodomain proteins. A_7- T_{26} forms a HG base pairing in both HG and HGP systems. The equilibration of each system is ensured by the saturation of the root mean square deviation (RMSD) with time. The equilibrated snapshots at 500 ns time span of the WC, HG, and HGP systems are shown in Figs. 2.3(a), (b), and (c), respectively. We consider the equilibrated part of the trajectory, 200 ns-1μs for further analysis.

2.3.1 Conformations in different systems

First, we check the DNA conformations in various cases. Tables 2.1- 2.3 and 2.4-2.6 show the averages with the errors for six bp step parameters and six intra-bp parameters for each bp in each system. These results show that the average value of each parameter corresponds well to the B-DNA data.[42] The mean value and the error of the mean value of the sugar phase angle, P, for all systems are tabulated in Table 2.7. We observe that the sugar conformation of each base in each system is predominantly C2′-endo like ($137°$-$194°$),[42] though some transition to C3′-exo ($194°$-$216°$) pucker is also found at the sugars linking to A_7 and T_{26} bases of the HG and the T_9 base of the HGP system but the frequencies of these are quite small.

System	steps	$\langle D_x^{WC} \pm 0.5\Delta D_x^{WC} \rangle$	$\langle D_y^{WC} \pm 0.5\Delta D_y^{WC} \rangle$	$\langle D_z^{WC} \pm 0.5\Delta D_z^{WC} \rangle$	$\langle \tau^{WC} \pm 0.5\Delta \tau^{WC} \rangle$	$\langle \rho^{WC} \pm 0.5\Delta \rho^{WC} \rangle$	$\langle \omega^{WC} \pm 0.5\Delta \omega^{WC} \rangle$
WC	$T_4 - A_{29}/G_5 - C_{28}$	0.40 ± 0.008	-0.17 ± 0.01	3.37 ± 0.001	0.47 ± 0.04	7.19 ± 0.11	37.68 ± 0.04
	$G_5 - C_{28}/T_6 - A_{27}$	-0.43 ± 0.01	-0.19 ± 0.003	3.29 ± 0.002	0.11 ± 0.02	0.30 ± 0.04	31.35 ± 0.05
	$T_6 - A_{27}/A_7 - T_{26}$	0.49 ± 0.03	0.24 ± 0.01	3.43 ± 0.002	1.82 ± 0.08	4.24 ± 0.11	32.65 ± 0.12
	$A_7 - T_{26}/A_8 - T_{25}$	-0.62 ± 0.01	-0.02 ± 0.01	3.28 ± 0.001	-2.35 ± 0.02	0.59 ± 0.11	36.84 ± 0.03
	$A_8 - T_{25}/T_9 - A_{24}$	0.02 ± 0.003	-0.69 ± 0.002	3.10 ± 0.003	0.02 ± 0.01	-0.73 ± 0.10	35.97 ± 0.06
	$T_9 - A_{24}/T_{10} - A_{23}$	0.32 ± 0.005	-0.23 ± 0.01	3.28 ± 0.001	1.80 ± 0.02	1.35 ± 0.10	37.34 ± 0.04
	$T_{10} - A_{23}/C_{11} - G_{22}$	0.50 ± 0.008	0.17 ± 0.01	3.39 ± 0.002	0.78 ± 0.02	2.66 ± 0.10	32.45 ± 0.10

Table 2.1: The mean with the error in the mean of each step in the WC system for the bp step parameters (D_x, D_y, D_z, τ, ρ and ω). D_x, D_y, D_z are in Å. τ, ρ and ω are in degree.

System	steps	$\langle D_x^{HG} \pm 0.5\Delta D_x^{HG} \rangle$	$\langle D_y^{HG} \pm 0.5\Delta D_y^{HG} \rangle$	$\langle D_z^{HG} \pm 0.5\Delta D_z^{HG} \rangle$	$\langle \tau^{HG} \pm 0.5\Delta \tau^{HG} \rangle$	$\langle \rho^{HG} \pm 0.5\Delta \rho^{HG} \rangle$	$\langle \omega^{HG} \pm 0.5\Delta \omega^{HG} \rangle$
HG	$T_4 - A_{29}/G_5 - C_{28}$	0.06 ± 0.01	-0.24 ± 0.01	3.33 ± 0.002	0.62 ± 0.03	7.41 ± 0.05	38.91 ± 0.05
	$G_5 - C_{28}/T_6 - A_{27}$	-0.23 ± 0.01	-0.36 ± 0.01	3.19 ± 0.002	0.80 ± 0.01	0.35 ± 0.03	33.44 ± 0.05
	$T_6 - A_{27}/A_7 - T_{26}$	-0.48 ± 0.02	-1.40 ± 0.01	3.30 ± 0.004	-3.13 ± 0.04	2.75 ± 0.08	49.29 ± 0.08
	$A_7 - T_{26}/A_8 - T_{25}$	1.20 ± 0.01	0.46 ± 0.01	3.61 ± 0.003	5.25 ± 0.06	3.75 ± 0.06	21.88 ± 0.08
	$A_8 - T_{25}/T_9 - A_{24}$	-0.04 ± 0.005	-0.66 ± 0.004	3.15 ± 0.003	-0.34 ± 0.01	-1.61 ± 0.05	34.68 ± 0.08
	$T_9 - A_{24}/T_{10} - A_{23}$	0.20 ± 0.01	-0.37 ± 0.01	3.28 ± 0.002	1.70 ± 0.02	1.46 ± 0.03	37.01 ± 0.03
	$T_{10} - A_{23}/C_{11} - G_{22}$	0.56 ± 0.01	0.18 ± 0.01	3.38 ± 0.001	0.97 ± 0.02	2.57 ± 0.02	33.22 ± 0.04

Table 2.2: The mean with the error in the mean of each step in the HG system for the bp step parameters (D_x, D_y, D_z, τ, ρ and ω). D_x, D_y, D_z are in Å. τ, ρ and ω are in degree.

System	steps	$\langle D_x^{HGP} \pm 0.5\Delta D_x^{HGP} \rangle$	$\langle D_y^{HGP} \pm 0.5\Delta D_y^{HGP} \rangle$	$\langle D_z^{HGP} \pm 0.5\Delta D_z^{HGP} \rangle$	$\langle \tau^{HGP} \pm 0.5\Delta \tau^{HGP} \rangle$	$\langle \rho^{HGP} \pm 0.5\Delta \rho^{HGP} \rangle$	$\langle \omega^{HGP} \pm 0.5\Delta \omega^{HGP} \rangle$
HGP	$T_4 - A_{29}/G_5 - C_{28}$	0.39 ± 0.03	-0.35 ± 0.05	3.27 ± 0.01	2.00 ± 0.14	8.58 ± 0.19	42.53 ± 0.44
	$G_5 - C_{28}/T_6 - A_{27}$	0.04 ± 0.005	-0.9 ± 0.01	3.16 ± 0.003	0.69 ± 0.05	4.63 ± 0.15	26.32 ± 0.13
	$T_6 - A_{27}/A_7 - T_{26}$	-1.09 ± 0.03	-2.06 ± 0.01	3.64 ± 0.002	1.81 ± 0.27	8.66 ± 0.30	55.32 ± 0.17
	$A_7 - T_{26}/A_8 - T_{25}$	1.29 ± 0.01	0.80 ± 0.01	3.43 ± 0.01	-0.66 ± 0.08	1.86 ± 0.05	18.09 ± 0.06
	$A_8 - T_{25}/T_9 - A_{24}$	0.29 ± 0.01	-1.04 ± 0.02	3.08 ± 0.003	1.42 ± 0.09	2.79 ± 0.13	34.61 ± 0.08
	$T_9 - A_{24}/T_{10} - A_{23}$	0.40 ± 0.03	-0.07 ± 0.04	3.32 ± 0.003	2.97 ± 0.10	1.54 ± 0.24	32.65 ± 0.46
	$T_{10} - A_{23}/C_{11} - G_{22}$	-0.00 ± 0.04	0.39 ± 0.04	3.34 ± 0.003	0.05 ± 0.16	1.05 ± 0.43	33.12 ± 0.72

Table 2.3: The mean with the error in the mean of each step in the HGP system for the bp step parameters (D_x, D_y, D_z, τ, ρ and ω). D_x, D_y, D_z are in Å. τ, ρ and ω are in degree.

System	steps	$\langle S_x^{WC} \pm 0.5\Delta S_x^{WC} \rangle$	$\langle S_y^{WC} \pm 0.5\Delta Sy^{WC} \rangle$	$\langle S_z^{WC} \pm 0.5\Delta S_z^{WC} \rangle$	$\langle \kappa^{WC} \pm 0.5\Delta \kappa^{WC} \rangle$	$\langle \pi^{WC} \pm 0.5\Delta \pi^{WC} \rangle$	$\langle \sigma^{WC} \pm 0.5\Delta \sigma^{WC} \rangle$
WC	$T_4 - A_{29}$	-0.12 ± 0.004	-0.11 ± 0.002	2.90 ± 0.001	2.71 ± 0.12	-11.80 ± 0.0	1.24 ± 0.05
	$G_5 - C_{28}$	0.08 ± 0.003	-0.04 ± 0.002	2.92 ± 0.001	-9.41 ± 0.10	-6.60 ± 0.10	-0.49 ± 0.01
	$T_6 - A_{27}$	0.01 ± 0.004	-0.16 ± 0.002	2.92 ± 0.001	2.74 ± 0.14	-6.11 ± 0.10	-0.34 ± 0.07
	$A_7 - T_{26}$	-0.30 ± 0.002	0.15 ± 0.001	2.88 ± 0.001	3.16 ± 0.08	-13.80 ± 0.08	1.81 ± 0.04
	$A_8 - T_{25}$	-0.21 ±0.004	0.19 ± 0.003	2.91 ± 0.001	-1.41 ± 0.03	-19.69 ± 0.12	0.17 ± 0.04
	$T_9 - A_{24}$	-0.21 ± 0.003	-0.18 ± 0.004	2.89 ± 0.001	2.12 ± 0.07	-19.41 ± 0.09	0.50 ± 0.03
	$T_{10} - A_{23}$	-0.34 ± 0.003	-0.15 ± 0.002	2.89 ± 0.001	-3.21 ± 0.07	-14.99 ± 0.04	0.66 ± 0.02

Table 2.4: The mean with the error in the mean of each step in the WC system for the bp step parameters (S_x, S_y, S_z, κ, Π and σ). S_x, S_y, S_z are in Å. κ, π and σ are in degree.

System	steps	$\langle S_x^{HG} \pm 0.5\Delta S_x^{HG} \rangle$	$\langle S_y^{HG} \pm 0.5\Delta Sy^{HG} \rangle$	$\langle S_z^{HG} \pm 0.5\Delta S_z^{HG} \rangle$	$\langle \kappa^{HG} \pm 0.5\Delta \kappa^{HG} \rangle$	$\langle \pi^{HG} \pm 0.5\Delta \pi^{HG} \rangle$	$\langle \sigma^{HG} \pm 0.5\Delta \sigma^{HG} \rangle$
HG	$T_4 - A_{29}$	0.23 ± 0.003	-0.07 ± 0.001	2.89 ± 0.001	8.38 ± 0.09	-1.32 ± 0.06	1.30 ± 0.02
	$G_5 - C_{28}$	-0.05 ± 0.004	-0.10 ± 0.003	2.91 ± 0.001	-3.94 ± 0.12	-12.78 ± 0.11	0.04 ± 0.03
	$T_6 - A_{27}$	-0.19 ± 0.003	-0.19 ± 0.002	2.86 ± 0.001	1.84 ± 0.15	-13.31 ± 0.09	3.65 ± 0.06
	$A_7 - T_{26}$	-0.27 ± 0.005	-0.09 ± 0.002	2.86 ± 0.001	1.84 ± 0.15	-13.31 ± 0.09	3.65 ± 0.06
	$A_8 - T_{25}$	-0.29 ±0.006	0.16 ± 0.005	2.87 ± 0.001	-1.47 ± 0.10	-15.36 ± 0.14	2.29 ± 0.02
	$T_9 - A_{24}$	-0.20 ± 0.006	-0.17 ± 0.005	2.89 ± 0.001	2.41 ± 0.09	-17.66 ± 0.11	0.85 ± 0.03
	$T_{10} - A_{23}$	-0.34 ± 0.004	-0.15 ± 0.001	2.89 ± 0.001	-2.87 ± 0.10	-14.95 ± 0.06	0.73 ± 0.04

Table 2.5: The mean with the error in the mean of each step in the HG system for the bp step parameters (S_x, S_y, S_z, κ, Π and σ). S_x, S_y, S_z are in Å. κ, π and σ are in degree.

System	steps	$\langle S_x^{HGP} \pm 0.5\Delta S_x^{HGP} \rangle$	$\langle S_y^{HGP} \pm 0.5\Delta Sy^{HGP} \rangle$	$\langle S_z^{HGP} \pm 0.5\Delta S_z^{HGP} \rangle$	$\langle \kappa^{HGP} \pm 0.5\Delta \kappa^{HGP} \rangle$	$\langle \pi^{HGP} \pm 0.5\Delta \pi^{HGP} \rangle$	$\langle \sigma^{HGP} \pm 0.5\Delta \sigma^{HGP} \rangle$
HGP	$T_4 - A_{29}$	0.14 ± 0.01	-0.0 ± 0.01	2.90 ± 0.001	12.00 ± 0.20	-1.87 ± 0.22	0.87 ± 0.11
	$G_5 - C_{28}$	0.01 ± 0.01	-0.48 ± 0.02	2.92 ± 0.005	-3.67 ± 0.53	-15.75 ± 0.25	2.84 ± 0.28
	$T_6 - A_{27}$	-0.27 ± 0.02	-0.11 ± 0.005	2.86 ± 0.006	-0.10 ± 0.49	-18.92 ± 0.16	11.45 ± 0.17
	$A_7 - T_{26}$	0.01 ± 0.01	-0.01 ± 0.005	2.97 ± 0.004	-16.97 ± 0.53	-11.73 ± 0.28	-6.84 ± 0.14
	$A_8 - T_{25}$	-0.18 ±0.01	0.16 ± 0.006	2.82 ± 0.002	-13.58 ± 0.38	-20.21 ± 0.19	8.69 ± 0.13
	$T_9 - A_{24}$	-0.11 ± 0.01	-0.15 ± 0.006	2.89 ± 0.002	-9.58 ± 0.38	-18.12 ± 0.30	2.19 ± 0.31
	$T_{10} - A_{23}$	-0.34 ± 0.01	-0.13 ± 0.004	2.91 ± 0.003	-5.23 ± 0.54	-11.20 ± 0.27	0.28 ± 0.16

Table 2.6: The mean with the error in the mean of each step in the HGP system for the bp step parameters (S_x, S_y, S_z, κ, Π and σ). S_x, S_y, S_z are in Å. κ, π and σ are in degree.

Base	$\langle P^{WC} \pm 0.5\Delta P^{WC} \rangle$	$\langle P^{HG} \pm 0.5\Delta P^{HG} \rangle$	$\langle P^{HGP} \pm 0.5\Delta P^{HGP} \rangle$
T_4	185.55 ± 0.06	182.35 ± 0.48	188.67 ± 1.18
G_5	186.72 ± 0.09	182.82 ± 0.25	178.78 ± 0.76
T_6	185.29 ± 0.12	181.24 ± 0.13	176.66 ± 0.38
A_7	176.46 ± 0.06	222.53 ± 0.42	177.76 ± 0.05
A_8	179.07 ± 0.07	179.31 ± 0.07	179.52 ± 0.39
T_9	187.96 ± 0.21	188.24 ± 0.30	202.81 ± 0.87
T_{10}	191.64 ± 0.27	189.96 ± 0.22	183.67 ± 0.64
A_{23}	177.58 ± 0.03	177.95 ± 0.09	178.78 ± 0.76
A_{24}	179.84 ± 0.06	181.41 ± 0.15	179.54 ± 0.24
T_{25}	185.64 ± 0.31	185.52 ± 0.33	182.17 ± 0.49
T_{26}	189.99 ± 0.51	201.37 ± 0.41	188.44 ± 0.36
A_{27}	180.55 ± 0.13	178.67 ± 0.07	178.64 ± 0.07
C_{28}	181.23 ± 0.16	181.55 ± 0.17	181.13 ± 0.51
A_{29}	179.78 ± 0.05	182.30 ± 0.21	181.91 ± 0.20

Table 2.7: The mean and error of the mean value of phase angle, P in degree of each base in WC, HG, and HGP systems.

System	$\langle d_{N7-N3} \pm 0.5\Delta d_{N7-N3} \rangle$	$\langle \theta_{N7-\hat{H}-N3} \pm 0.5\Delta\theta_{N7-\hat{H}-N3} \rangle$	$\langle d_{N1-N3} \pm 0.5\Delta d_{N1-N3} \rangle$	$\langle \theta_{N1-\hat{H}-N3} \pm 0.5\Delta\theta_{N1-\hat{H}-N3} \rangle$
WC	0.64 ± 0.0001	159.09 ± 0.01	0.29 ± 0.0001	169.83 ± 0.02
HG	0.31 ± 0.0009	168.70 ± 0.14	0.59 ± 0.0001	145.19 ± 0.03
HGP	0.30 ± 0.0004	169.09 ± 0.04	0.59 ± 0.0002	144.67 ± 0.10

Table 2.8: Mean and error of the mean of d_{H-A} and $\theta_{D-\hat{H}-A}$ in A_7-T_{26} bp. d_{H-A} in nm and $\theta_{D-\hat{H}-A}$ in degree.

Available data suggest that the transition from the WC to HG geometry requires about a 180° rotation of the base along the bond connecting the base to the sugar, known as the glycosidic bond. The WC N6-H–O4 H-bond is retained in A-T HG bp whereas the other WC N1–H-N3 H-bond is replaced by N7–H-N3 H-bond. In addition, the formation of HG-type H-bonds requires that the two bases come into closer proximity, thus shortening the $C1' - C1'$ distance by about 2 Årelative to WC bp. We check the HG base pairing criteria and compare them with the literature values.[43–47] We consider specific hydrogen bonding in A_7-T_{26}. Table 2.8 displays the mean values along with the error of the mean of these parameters, $\mathrm{d}D - A$ and $\theta_{D-\hat{H}-A}$ for all systems. We observe that H-bonds form between $N1A_7$ (donor) and $N3T_{26}$ (acceptor) in the WC system for A_7-T_{26} bp which is consistent with the known hydrogen bonding pattern. On the other hand, H-bonds are observed between $N7A_7$ (donor) and $N3T_{26}$ (acceptor) in both HG and HGP systems.

The mean and the error associated with the mean of $C1' - C1'$ distance

System	$\langle d_{C1'}^{T_6:A_{27}} \pm 0.5\Delta d_{C1'}^{T_6:A_{27}} \rangle$	$\langle d_{C1'}^{A_7:T_{26}} \pm 0.5\Delta d_{C1'}^{A_7:T_{26}} \rangle$	$\langle d_{C1'}^{A_8:T_{25}} \pm 0.5\Delta d_{C1'}^{A_8:T_{25}} \rangle$	$\langle d_{P_7-P_{38}} \pm 0.5\Delta d_{P_7-P_{38}} \rangle$	$\langle d_{P_8-P_{37}} \pm 0.5\Delta d_{P_8-P_{37}} \rangle$
WC	1.05 ± 0.001	1.05 ± 0.0002	1.05 ± 0.0002	1.76 ± 0.0003	1.77 ± 0.0002
HG	1.03 ± 0.0004	0.90 ± 0.0003	1.03 ± 0.0001	1.70 ± 0.0002	1.71 ± 0.0004
HGP	1.00 ± 0.001	0.89 ± 0.002	1.00 ± 0.001	1.62 ± 0.003	1.70 ± 0.002

Table 2.9: Mean and error of the mean of value $d_{C1'}$ for T_6-A_{27}, A_7-T_{26}, and A_8-T_{25} bps, and $d_{P_7-P_{38}}$, and $d_{P_8-P_{37}}$ for all systems. All are in nm.

System	$\langle \chi_{T_6} \pm 0.5\Delta\chi_{T_6} \rangle$	$\langle \chi_{A_7} \pm 0.5\Delta\chi_{A_7} \rangle$	$\langle \chi_{A_8} \pm 0.5\Delta\chi_{A_8} \rangle$	$\langle \chi_{T_{26}} \pm 0.5\Delta\chi_{T_{26}} \rangle$ (degree)
WC	-111.14 ± 0.08	-98.47 ± 0.13	-112.77 ± 0.06	-111.24 ± 0.19
HG	-112.43 ± 0.05	54.73 ± 0.12	-98.64 ± 0.10	-120.03 ± 0.13
HGP	-127.44 ± 0.27	60.36 ± 0.23	-96.88 ± 0.28	-105.75 ± 0.17

Table 2.10: Mean and error of the mean of value χ for T_6, A_7, A_8 and T_{26} bases. All are in degrre.

($d_{C1'-C1'}$) for T_6-A_{27}, A_7-T_{26}, and A_8-T_{25} bps for all the systems are shown in Table 2.9. The HGP system shows the shortest $d_{C1'-C1'}$ in A_7-T_{26}. Apart from A_7-T_{26} bp, the $d_{C1'-C1'}$ at nearby T_6-A_{27} and A_8-T_{25} bps of HG and HGP systems differ from the WC system. No significant changes in $d_{C1'-C1'}$ are observed for the remaining bps. Table 2.9 shows the mean of the neighboring P-P distance (d_p) and the error associated with the mean. d_p close to A_7-T_{26} bp is shortened in both HG and HGP systems (Table 2.9).

The glycosidic torsion angle χ, which characterizes the relative base/sugar orientation in DNA duplexes is defined by the torsion angles $O4' - C1' - N9 - C4$ and $O4' - C1' - N1 - C2$ in purines and pyrimidines, respectively. We calculate the mean and the error of the mean of χ for each base. The HG and HGP systems show syn χ torsion values in the A_7 base which are consistent with the HG base pairing. The value of χ at complementary T_{26} and adjoining A_8 bases shows minor variations in HG and HGP systems compared to the WC system (Table 2.10). In addition to T_{26} and A_8 bases, the χ value for the adjoining T_6 base in the HGP complex shows a little deviation from the WC conformer. The χ torsion angle does not differ much between systems in the remaining bases. We also observe that α and γ torsions are in the usual gauche$-$/gauche$+$[48] conformation about base A_7 for WC and HG systems (Table 2.11). However, in HGP, they follow the unusual gauche$+$/gauche$-$ conformation. α and γ torsion angles of the complementary T_{26} base remain unchanged for all systems. This result is in agreement with the experimental observations.[12]

In the HGP system, for each DNA base, we identify the residues of the protein that form strong contact with DNA bases, as detailed in Table 2.12. As, A_7-T_{26} bp creates HG bp, we examine in detail the interface interaction for A_7-T_{26} bp. Fig. 2.4 is a snapshot of active protein residues at the interface of HG bp. We

System	$\langle\alpha_{A_7} \pm 0.5\Delta\alpha_{A_7}\rangle$	$\langle\gamma_{A_7} \pm 0.5\Delta\gamma_{A_7}\rangle$	$\langle\alpha_{T_{26}} \pm 0.5\Delta\alpha_{T_{26}}\rangle$	$\langle\gamma_{T_{26}} \pm 0.5\Delta\gamma_{T_{26}}\rangle$ (degree)
WC	-70.51 $\pm$ 0.13	53.48 $\pm$ 0.05	-66.54 $\pm$ 0.14	55.27 $\pm$ 0.14
HG	-56.51 $\pm$ 0.57	44.14 $\pm$ 0.54	-66.80 $\pm$ 0.31	54.41 $\pm$ 0.29
HGP	63.16 $\pm$ 0.23	-65.94 $\pm$ 0.22	-77.80 $\pm$ 0.21	59.95 $\pm$ 0.12

Table 2.11: Mean and error of the mean of value of α and γ torsion angles in A_7, and T_{26} bases. All are in degrre.

Base	Interface forming residues	
DNA	Protein α2B	Protein α2D
T_4		
G_5	ARG185	GLY133, ASN182
T_6	ARG185	GLY133, HIS134, ARG135, PHE136
A_7		ARG132, GLY133, GLN175, ASN182
A_8		ASN178
T_9	ARG135	
T_{10}	ALA132	
A_{23}		ARG132
A_{24}	ASN178	
T_{25}	GLN175	
T_{26}	ARG135, PHE136, HIS134, ASN182	
A_{27}	ARG135, LYS186	
C_{28}		ARG185, ASN178
A_{29}	TYR156, SER181, ARG184	

Table 2.12: The protein residues that strongly interact with DNA bps in the HGP system.

note that protein α2D residues ARG132, GLY133 and GLN175 create interface with base A_7 of DNA (Table 2.13). These three residues form H-bonds with A_7 at the interface. Moreover, residue ARG132 is involved in forming salt bridge as well as electrostatic contact between the N atom of ARG132 and the O1P atom of A_7. On the other hand, the specific protein α2B residues HIS134, PHE136, and TRP179 create H-bonds with T_{26} (Table 2.14). In the $3' - 5'$ strand, there is no salt bridge or electrostatic contact between T_{26} and the residues of α2B.

System	Nucleotide	Protein	H-bond	Salt bridge	Electrostatic
HGP	A_7	α2D	ARG132 (0.65)	ARG132 (0.26)	ARG132 (0.93)
			GLY133 (0.92)		
			GLN175 (0.93)		

Table 2.13: The ratio of the number of frames in each non-bonded interaction to the total number of frames, for each interface forming α2D protein residue, considering the interfacial contact between A_7 and α2D residues.

System	Nucleotide	Protein	H-bond
HGP	T_{26}	α2B	HIS134 (0.61)
			PHE136 (0.99)
			TRP179 (0.59)

Table 2.14: The ratio of the number of frames in each non-bonded interaction to the total number of frames, for each interface forming α2B protein residue, considering the interfacial contact between T_{26} and α2B residues.

Figure 2.4: Snapshot of active protein residues at the interface in HGP system. Protein interface binding residues with HG bp are shown in yellow.

2.3.2 Fluctuations in conformational variables

The conformational fluctuations are quantified in terms of the distribution, H^S (θ_R), (S = WC, HG and HGP) of a microscopic conformational variable, θ pertaining to base, R over the equilibrated trajectory. Each peak of the histogram represents the most probable value for the given variable. The larger width of the histogram peaks denotes more conformational flexibility. We illustrate below a few cases (χ, ϵ, $nu0$, and $nu3$) of microscopic conformational degrees of freedom that exhibit major changes in the A_7-T_{26} (HG) bp forming region of each system. Fig. 2.5 shows the distributions of sugar-phosphate (ϵ), sugar-base (χ) and sugar-pucker ($nu0$, $nu3$) torsion angles for A_7 and T_{26} bases. H^{WC} (ϵ_{A_7}) and H^{WC} ($\epsilon_{T_{26}}$) are multi-peaked, as shown in Figs. 2.5 (a) and (b). On the other hand, H^{HG} (ϵ_{A_7}), H^{HGP} (ϵ_{A_7}), H^{HG} ($\epsilon_{T_{26}}$) and H^{HGP} ($\epsilon_{T_{26}}$) are all double peaked, showing a decrease in flexibility of the HG and HGP systems compared to the WC system for ϵ. Fig. 2.5 (c) shows that H^{HG} (χ_{A_7}) is distributed over a different range compared to H^{WC} (χ_{A_7}). The height of the peak in H^{HGP} (χ_{A_7}) is higher than that of H^{WC} (χ_{A_7}). Again, the distributions of H^{WC} ($\chi_{T_{26}}$) and H^{HG} ($\chi_{T_{26}}$) are almost identical Fig. 2.5 (d) but H^{HGP} ($\chi_{T_{26}}$) has a sharper peak.

Now we consider the distributions of the sugar torsion angles $\nu0$ and $\nu3$.

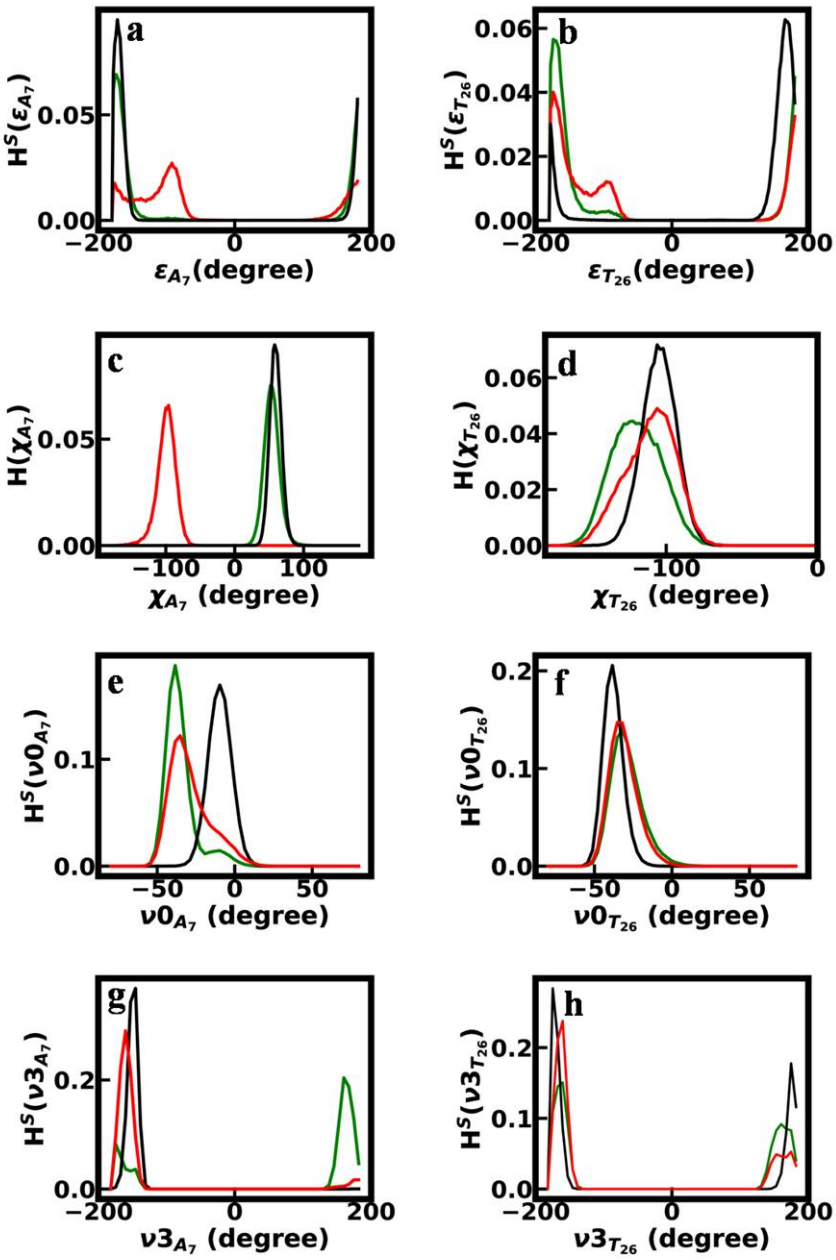


Figure 2.5: Histograms of torsion angles (a) ϵ for A_7, (b) ϵ for T_{26}, (c) χ for A_7, (d) χ for T_{26} (e) $\nu 0$ of base A_7, (f) $\nu 0$ of base T_{26}, (g) $\nu 3$ for base A_7, and (h) $\nu 3$ for base T_{26}, for three systems in equilibrated trajectories. S = WC (red), HG (green), and HGP (black).

Let us consider A_7. In Fig. 2.5 (e), H^{WC} ($\nu 0_{A_7}$) shows a single peak. The single peak changes to double peaks in H^{HG} ($\nu 0_{A_7}$), and reverts back to a single peak in H^{HGP} ($\nu 0_{(A_7)}$). But, H^{WC} ($\nu 0_{T_{26}}$), H^{HG} ($\nu 0_{(T_{T_{26}}}$) and H^{HGP} ($\nu 0_{T_{26}}$) are single peaked (Fig. 2.5 (f)), the peak being the highest H^{HGP} ($\nu 0_{T_{26}}$). Again, we see, at A_7, H^{WC} ($\nu 3_{A_7}$) has one peak, but H^{HG} ($\nu 3_{A_7}$) has two peaks, as illustrated in Fig. 2.5 (g). However,H^{HGP} ($\nu 3_{A_7}$) has a sharp peak higher than H^{WC} ($\nu 3_{A_7}$) for A_7. H^{WC} ($\nu 3_{T_{26}}$), H^{HG} ($\nu 3_{T_{26}}$) and H^{HGP} ($\nu 3_{T_{26}}$) are all double peaked, each with a sharp peak (Fig. 2.5 (h)). The peak in H^{HGP} ($\nu 3_{T_{26}}$) is the highest. Overall, the HG system shows multi-peaked or broad distribution in most of the microscopic conformational degrees of freedom compared to the WC system for A_7-T_{26} bp. This suggests that HG bp is more flexible than WC bp in naked DNA. On the other hand, in the majority of cases, the sharper peaks of the HGP complex indicate that the presence of proteins reduce the flexibility of the HG bp.

2.3.3 Conformational thermodynamics

The changes in the distribution of the microscopic conformation variables lead to the changes in conformational free energy and entropy of DNA in HG and HGP systems with reference to the WC system. By adding the changes in the individual microscopic degrees of freedom, we compute the overall changes in conformational thermodynamics for the complete DNA duplex.

Conformational thermodynamics changes of the HG system with respect to the WC system: We first consider the changes in conformational thermodynamics in the inter-bp step parameters. For a given step parameter, θ we denote the conformational free energy and entropy changes in the HG system with respect to the WC system by $\Delta G^{HG,inter}_{i:i+1}$ (θ) and $T\Delta S^{HG,inter}_{i:i+1}$ (θ) for the step $i : i+1$, where i and $i+1$ are the bases in $5' - 3'$ direction along with their complementary bases in the opposite strand. Figs. 2.6(a) and (b) show the data. We find that $\Delta G^{HG,inter}_{i:i+1}$ (D_x), $\Delta G^{HG,inter}_{i:i+1}$ (D_y), $\Delta G^{HG,inter}_{i:i+1}$ (τ) and $\Delta G^{HG,inter}_{i:i+1}$ (ρ) do not depend significantly on the bp steps in Fig. 2.6(a). On the other hand, $\Delta G^{HG,inter}_{i:i+1}$ (D_z) has maximum destabilization at the step A_7 :A_8. $\Delta G^{HG,inter}_{i:i+1}$ (ω) also contributes the most to destabilizing the step A_7:A_8. Let us now consider the changes in entropy. In Fig. 2.6(b), we see that $T\Delta S^{HG,inter}_{i:i+1}$ (D_y) and $T\Delta S^{HG,inter}_{i:i+1}$ (ρ) are not sensitive to bp steps. The step T_6:A_7 show the maximum order by $T\Delta S^{HG,inter}_{i:i+1}$ (D_x), whereas the step A_7:A_8 is most disordered by $T\Delta S^{HG,inter}_{i:i+1}$ (τ). $T\Delta S^{HG,inter}_{i:i+1}$ (D_z) shows maximum disorder at the step T_6:A_7. However, $T\Delta S^{HG,inter}_{i:i+1}$ (ω) shows maximum order at the step T_6:A_7 and maximum disorder at the step A_7:A_8. We

observe that due to inter-bp step parameters, the total changes in conformational free energy for a specific step, $\Delta G^{HG,inter}_{i:i+1}$ (= $\sum_\theta \Delta G^{HG,inter}_{i:i+1}$ (θ), where θ includes all of the step parameters) exhibits maximum destabilization at A_7:A_8 and total changes in conformation entropy, $T\Delta S^{HG,inter}_{i:i+1}$ (= $\sum_\theta T\Delta S^{HG,inter}_{i:i+1}$ (θ) shows maximum conformational order at step T_6:A_7 and maximum conformational disorder at step A_7:A_8.

Now we consider the changes in conformational thermodynamics due to intra-bp parameters. The data for the changes in the conformational free energy and entropy, denoted by $\Delta G^{HG,intra}_i$ (θ) and T$\Delta S^{HG,intra}_i$ (θ), respectively, for the intra-bp parameter θ for ith bp of the HG system with respect to the WC system are shown in Figs. 2.6(c) and (d), respectively. Here,$\Delta G^{HG,intra}_i$ (S_x), $\Delta G^{HG,intra}_i$ (κ) and $\Delta G^{HG,intra}_i$ (σ) are not sensitive to the bp (Fig. 2.6(c)). However, $\Delta G^{HG,intra}_i$ (S_y) has maximum destabilization at A_8:T_{25}. Again, $\Delta G^{HG,intra}_i$ (S_z) exhibits maximum destabilization at A_7:T_{26} and maximum stabilization for T_6:A_{27}. However, $\Delta G^{HG,intra}_i$ (π) undergoes the maximum stabilization at A_8:T_{25}. Again, $T\Delta S^{HG,intra}_i$ (S_x), $T\Delta S^{HG,intra}_i$ (S_y) and $T\Delta S^{HG,intra}_i$ (κ) do not depend significantly on the bps (Fig. 2.6(d)). However,$T\Delta S^{HG,intra}_i$ (σ) has maximum disorder at A_7:T_{26}. $T\Delta S^{HG,intra}_i$ (π) exhibits maximum order at G_5:C_{28}. Again, $T\Delta S^{HG,intra}_i$ (S_z) shows maximum disorder at A_7:T_{26} and maximum order at T_6:A_{27}. Finally, we compute the total changes in conformational free energy, $\Delta G^{HG,intra}_i$ (= $\sum_\theta \Delta G^{HG,intra}_i$ (θ), θ running over the intra-bp parameters) and conformational entropy, $T\Delta S^{HG,intra}_i$ (= $\sum_\theta T\Delta S^{HG,intra}_i$ (θ), θ running over the intra-bp parameters) for each bp for intra-bp parameters. At A_7:T_{26}, $\Delta G^{HG,intra}_i$ exhibits the most conformational destabilization, whereas $T\Delta S^{HG,intra}_i$ exhibits the most conformational disorder.

Next, we consider the conformational thermodynamic changes caused by backbone torsion angles. We compute strand specific conformational thermodynamics due to sugar-phosphate, sugar-base and sugar-puckers. The conformational free energy, $\Delta G^{HG,SP}$ (θ) and entropy, T$\Delta S^{HG,SP}$ (θ) changes data for sugar-phosphate and sugar-base torsions for the ith base of $5' - 3'$ strand of the HG system with respect to the WC system are shown in Figs. 2.7(a) and (b), respectively, where θ indicates sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ. We find that $\Delta G^{HG,SP}_i$ (α), $\Delta G^{HG,SP}_i$ (β), $\Delta G^{HG,SP}_i$ (χ),$\Delta G^{HG,SP}_i$ (ϵ), $\Delta G^{HG,SP}_i$ (γ) and $\Delta G^{HG,SP}_i$ (ζ) do not contribute significantly to bases in Fig. 2.7 (a). $\Delta G^{HG,SP}_i$ (δ) exhibits maximum destabilization and stabilization at A_7 and G_5, respectively. In Fig. 2.7 (b), we see that $T\Delta S^{HG,SP}_i$ (α), $T\Delta S^{HG,SP}_i$ (χ) and $T\Delta S^{HG,SP}_i$ (γ) are not sensitive to the

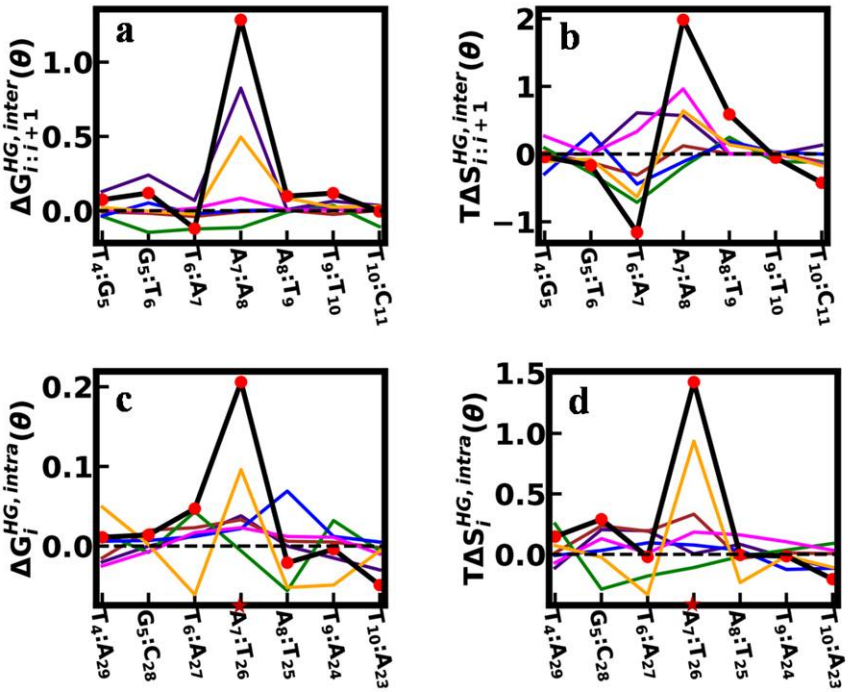


Figure 2.6: (a) $\Delta G^{HG,inter}_{i:i+1}(\theta)$ and (b) $T\Delta S^{HG,inter}_{i:i+1}(\theta)$ due to inter-bp step parameters, (c) $\Delta G^{HG,intra}_{i}(\theta)$ and (d) $T\Delta S^{HG,intra}_{i}(\theta)$ by intra-bp parameters, for the HG system with respect to the WC system. In panels (a) and (b), D_x is in green, D_y in blue, D_z in indigo, τ in magenta, ρ in brown, and ω in orange. In panels (c) and (d), S_x is in magenta, S_y in blue, S_z in orange, κ in indigo, σ in brown, and π in green. The red circles in panels (a) and (b), joined by a solid black line, reflect the overall changes for each step. Similarly, in panels (c) and (d), the red circles joined by a solid black line represent the overall changes for each bp. HG bp is marked by the maroon star in all panels. All quantities are in kJ/mol.

bases. However, $T\Delta S^{HG,SP}_{i}(\epsilon)$ contributes the most to ordering the A_7. On the other hand, $T\Delta S^{HG,SP}_{i}(\delta)$ has maximum disorder at A_7 while $T\Delta S^{HG,SP}_{i}(\zeta)$ and $T\Delta S^{HG,SP}_{i}(\beta)$ have maximum order at A_7 and A_8, respectively. We compute the total changes in conformational free energy, $\Delta G^{HG,SP}_{i}$ (= $\sum_\theta \Delta G^{HG,SP}_{i}(\theta)$, θ stands for sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ) and conformational entropy, $T\Delta S^{HG,SP}_{i}$ (= $\sum_\theta T\Delta S^{HG,SP}_{i}(\theta)$, θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ). $\Delta G^{HG,SP}_{i}$ shows maximum destabilization at A_7 while $T\Delta S^{HG,SP}_{i}$ shows maximum order at A_7. For the $3' - 5'$ strand, the conformational free energy, $\Delta G^{HG,SP}_{i_c}(\theta)$ and entropy, $T\Delta S^{HG,SP}_{i_c}(\theta)$ changes data for sugar-phosphate and sugar-base torsions for the i_c^{th} base of the HG system with respect to the WC system are shown in Figs. 2.7 (c) and (d), respectively. In Fig. 2.7 (c), $\Delta G^{HG,SP}_{i_c}(\alpha)$, $\Delta G^{HG,SP}_{i_c}(\beta)$ and $\Delta G^{HG,SP}_{i_c}(\gamma)$ are not sensitive to

bases. In contrast, $\Delta G_{i_c}^{HG,SP}$ (χ) has maximum stability at C_{28} while $\Delta G_{i_c}^{HG,SP}$ (ϵ) and $\Delta G_{i_c}^{HG,SP}$ (ζ) have maximum stability at T_{26}. $\Delta G_{i_c}^{HG,SP}$ (δ) shows maximum destabilization and stabilization at A_{29} and T_{25}, respectively. On the other hand, in Fig. 2.7 (d), $T\Delta S_{i_c}^{HG,SP}$ (α), $T\Delta S_{i_c}^{HG,SP}$ (χ) and $T\Delta S_{i_c}^{HG,SP}$ (γ) are not sensitive to bases. $T\Delta S_{i_c}^{HG,SP}$ (β) has maximum order at A_{27}. $T\Delta S_{i_c}^{HG,SP}$ (ϵ) and $T\Delta S_{i_c}^{HG,SP}$ (ζ) show maximum order at T_{26} while $T\Delta S_{i_c}^{HG,SP}$ (δ) shows maximum disorder at T_{26}.

We evaluate the total changes in conformational free energy, $\Delta G_{i_c}^{HG,SP}$ (= $\sum_\theta \Delta G_{i_c}^{HG,SP}$ (θ), θ stands for sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ) and conformational entropy, $T\Delta S_{i_c}^{HG,SP}$ (= $\sum_\theta T\Delta S_{i_c}^{HG,SP}$ (θ), θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ) for $3' - 5'$ strand. Here, $\Delta G_{i_c}^{HG,SP}$ shows maximum stabilization at T_{26} while $T\Delta S_{i_c}^{HG,SP}$ shows maximum order at T_{26}.

Let us now consider the changes in conformational thermodynamics due to sugar-puckers. In the $5' - 3'$ strand, the changes in conformational free energy, $\Delta G_i^{HG,S}$ (θ) and entropy, $T\Delta S_i^{HG,S}$ (θ) due to the sugar torsion angles ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$) of each of i^{th} base in the HG system with respect to the WC system are shown in Figs. 2.7 (e) and (f), respectively. We note in Fig. 2.7 (e) that $\Delta G_i^{HG,S}$ ($\nu 1$) is not sensitive to bases. At G_5 and A_7, $\Delta G_i^{HG,S}$ ($\nu 0$) exhibits the most destabilization and stabilization, respectively. On the other hand, $\Delta G_i^{HG,S}$ ($\nu 2$) contributes the most to destabilizing the A_7 and to stabilizing the T_6. However, $\Delta G_i^{HG,S}$ ($\nu 3$) has maximum destabilization for A_7. Again, $\Delta G_i^{HG,S}$ ($\nu 4$) contributes maximum to destabilize A_8. We find that in Fig. 2.7 (f), $T\Delta S_i^{HG,S}$ ($\nu 0$) and $T\Delta S_i^{HG,S}$ ($\nu 4$) have maximum order to A_7. $T\Delta S_i^{HG,S}$ ($\nu 1$) contributes the most to disorder the A_8. $T\Delta S_i^{HG,S}$ ($\nu 2$) shows maximum disorder at A_7. However, $T\Delta S_i^{HG,S}$ ($\nu 3$) contributes the most to order the T_6 and to disorder the A_7. We compute the total changes in conformational free energy, $\Delta G_i^{HG,S}$ (= $\sum_i \Delta G_i^{HG,S}$ (θ), θ runs over all sugar-puckers) and conformational entropy, $T\Delta S_i^{HG,S}$ (=$\sum_\theta T\Delta S_i^{HG,S}$ (θ), θ runs over all sugar-puckers). $\Delta G_i^{HG,S}$ exhibits maximum stabilization at T_6 while $T\Delta S_i^{HG,S}$ shows maximum disorder at A_8.

In the $3' - 5'$ strand, the changes in conformational free energy, $\Delta G_{i_c}^{HG,S}$ (θ) and entropy, $T\Delta S_{i_c}^{HG,S}$ (θ) due to the sugar torsion angles of each of i_c^{th} base in the HG system with regard to the WC system are shown in Figs. 2.7 (g) and (h), respectively. According to Fig. 2.7 (g), $\Delta G_{i_c}^{HG,S}$ ($\nu 3$) is not sensitive to bases. $\Delta G_{i_c}^{HG,S}$ ($\nu 0$), $\Delta G_{i_c}^{HG,S}$ ($\nu 1$) and $\Delta G_{i_c}^{HG,S}$ ($\nu 2$) show maximum destabilization at A_{24}, A_{27} and T_{26}, respectively. Again, $\Delta G_{i_c}^{HG,S}$ ($\nu 4$) shows maximum stabilization at A_{24}. In Fig. 2.7 (h), $T\Delta S_{i_c}^{HG,S}$ ($\nu 0$), $T\Delta S_{i_c}^{HG,S}$ ($\nu 1$), $T\Delta S_{i_c}^{HG,S}$ ($\nu 2$), $T\Delta S_{i_c}^{HG,S}$

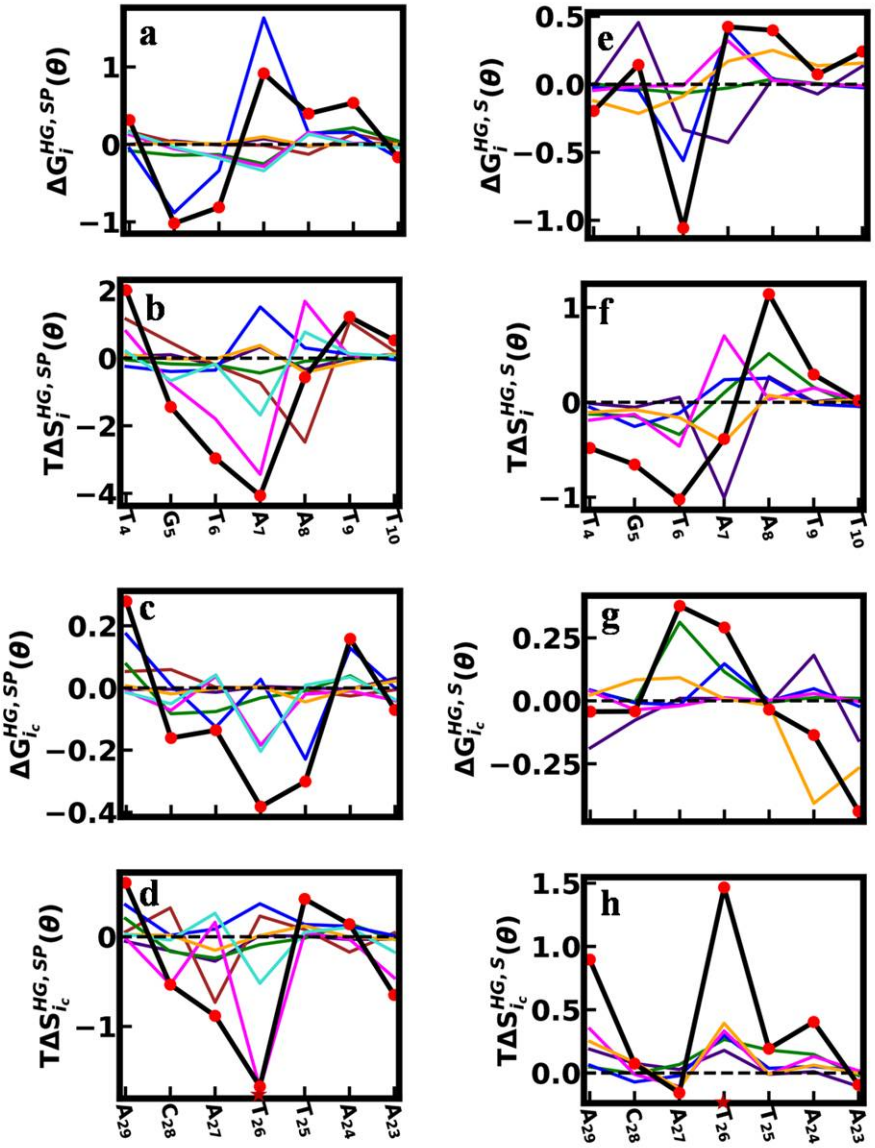


Figure 2.7: For sugar-phosphate and sugar-base torsion angles, (a) $\Delta G_i^{HG,SP}(\theta)$ and (b) $T\Delta S_i^{HG,SP}(\theta)$ for the $5'-3'$ strand, (c) $\Delta G_{i_c}^{HG,SP}(\theta)$ and (d) $T\Delta S_{i_c}^{HG,SP}(\theta)$ for the $3'-5'$ strand, in the HG system with respect to the WC system. Similarly, for sugar-puckers, (e) $\Delta G_i^{HG,S}(\theta)$ and (f) $T\Delta S_i^{HG,SP}(\theta)$ for the $5'-3'$ strand, (g) $\Delta G_{i_c}^{HG,S}(\theta)$ and (h) $T\Delta S_{i_c}^{HG,S}(\theta)$ for the $3'-5'$ strand, in the HG system with regard to the WC system. α is in indigo, β in brown, χ in green, δ in blue, ϵ in magenta, γ in orange, and ζ in turquoise in panels (a)-(d), $\nu 0$ is in indigo, $\nu 1$ in green, $\nu 2$ in blue, $\nu 3$ in magenta, and $\nu 4$ in orange in panels (e)-(h). The red circles joined by a solid black line in each panel reflect the overall changes for each base. HG bp region is marked by the maroon star in all panels. All quantities are in kJ/mol.

($\nu 3$) and $T\Delta S_{i_c}^{HG,S}$ ($\nu 4$) show maximum disorder at T_{26}. We calculate the total changes in conformational free energy, $\Delta G_{i_c}^{HG,S}$ (= $\sum_\theta \Delta G_{i_c}^{HG,S}$ (θ), θ runs over all sugar-puckers) and conformational entropy, $T\Delta S_{i_c}^{HG,S}$ (=$\sum_\theta T\Delta S_{i_c}^{HG,S}$ (θ), θ runs over all sugar-puckers) for the complementary $3'-5'$ strand. $\Delta G_{i_c}^{HG,S}$ shows maximum stabilization at A_{23} while $T\Delta S_{i_c}^{HG,S}$ shows maximum disorder at T_{26}.

Conformational thermodynamics changes of the HGP system with respect to the WC system: We first focus on the changes in conformational thermodynamics due to inter-bp step parameters. The conformational free energy and entropy changes of the HGP system in comparison to the WC system for $i : i+1$ step are denoted by $\Delta G_{i:i+1}^{HGP,inter}$ (θ) and $T\Delta S_{i:i+1}^{HGP,inter}$ (θ) as shown in Figs. 2.8 (a) and (b), respectively. Fig. 2.8 (a) shows that $\Delta G_{i:i+1}^{HGP,inter}$ (D_y), $\Delta G_{i:i+1}^{HGP,inter}$ (ρ), $\Delta G_{i:i+1}^{HGP,inter}$ (τ) and $\Delta G_{i:i+1}^{HGP,inter}$ (ω) are not sensitive to the bp steps. $\Delta G_{i:i+1}^{HGP,inter}$ (D_x) contributes the most to stabilizing the step A_7:A_8. Again, $\Delta G_{i:i+1}^{HGP,inter}$ (D_z) has maximum destabilization at the step T_6:A_7. Fig. 2.8 (b) shows that $T\Delta S_{i:i+1}^{HGP,inter}$ (D_z) and $T\Delta S_{i:i+1}^{HGP,inter}$ (τ) do not contribute significantly to the bp steps. $T\Delta S_{i:i+1}^{HGP,inter}$ (D_x), $T\Delta S_{i:i+1}^{HGP,inter}$ (D_y) and $T\Delta S_{i:i+1}^{HGP,inter}$ (ω) have maximum contributions to order the step G_5:T_6 to A_7:A_8. Again, $T\Delta S_{i:i+1}^{HGP,inter}$ (ω) and $T\Delta S_{i:i+1}^{HGP,inter}$ (ρ) contribute maximum to disorder in the step T_{10}:C_{11}. The total changes in conformational free energy for every step, $\Delta G_{i:i+1}^{HGP,inter}$ (= $\sum_\theta \Delta G_{i:i+1}^{HGP,inter}$ (θ), θ running over all bp step parameters) exhibit maximum stabilization at the step A_7:A_8. The total changes in conformation entropy for individual step, $T\Delta S_{i:i+1}^{HGP,inter}$ (=$\sum_\theta T\Delta S_{i:i+1}^{HGP,inter}$ (θ)) are shown in Fig. 2.8 (b). The data show that there is significant ordering in the bp steps T_4:G_5 to A_7:A_8, with the maximum ordering at T_6:A_7. On the contrary, step A_7:A_8 exhibits maximum destabilization and disorder in $\Delta G_{i:i+1}^{HGP,inter}$ and $T\Delta S_{i:i+1}^{HGP,inter}$, respectively.

Next we consider the intra-bp parameters. The changes in the conformational free energy and entropy for the HGP system with respect to the WC system, denoted by $\Delta G_i^{HGP,intra}$ (θ) and $T\Delta S_i^{HGP,intra}$ (θ) , respectively, are shown in Figs. **??** (c) and (d), respectively. In Fig. **??** (c), we see that $\Delta G_i^{HGP,intra}$ (κ), $\Delta G_i^{HGP,intra}$ (σ) and $\Delta G_i^{HGP,intra}$ (S_y) do not depend significantly to the bps. $\Delta G_i^{HGP,intra}$ (S_z) and $\Delta G_i^{HGP,intra}$ (π) contribute maximum to stabilize the T_6:A_{27} and A_8:T_{25}, respectively. Again, $\Delta G_i^{HGP,intra}$ (S_x) has maximum destabilization and stabilization for G_5:C_{28} and T_{10}:A_{23}, respectively. In Fig. **??** (d), we see that $T\Delta S_i^{HGP,intra}$ (S_x) and $T\Delta S_i^{HGP,intra}$ (S_y) do not contribute significantly on the bps. $T\Delta S_i^{HGP,intra}$ (S_z) has maximum disorder at G_5:C_{28} and maximum order at A_8:T_{25}. However,$T\Delta S_i^{HGP,intra}$ (σ) contributes maximum disorder to the step G_5:C_{28}. $T\Delta S_i^{HGP,intra}$ (π) shows maximum order at G_5:C_{28}. Here, the total

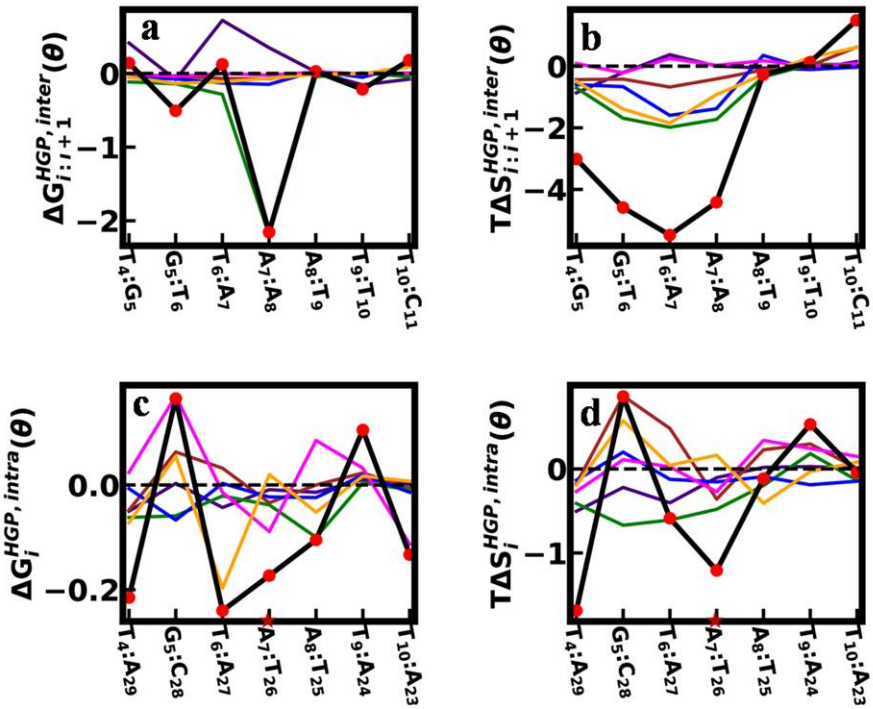


Figure 2.8: (a) $\Delta G^{HGP,inter}_{i:i+1}$ (θ) and (b) $T\Delta S^{HGP,inter}_{i:i+1}$ (θ) due to inter-bp step parameters, (c) $\Delta G^{HGP,intra}_{i}$ (θ) and (d) $T\Delta S^{HGP,intra}_{i}$ (θ) by intra-bp parameters, for the HG system with respect to the WC system. In panels (a) and (b), D_x is in green, D_y in blue, D_z in indigo, τ in magenta, ρ in brown, and ω in orange. In panels (c) and (d), S_x is in magenta, S_y in blue, S_z in orange, κ in indigo, σ in brown, and π in green. The red circles in panels (a) and (b), joined by a solid black line, reflect the overall changes for each step. Similarly, in panels (c) and (d), the red circles joined by a solid black line represent the overall changes for each bp. HG bp is marked by the maroon star in all panels. All quantities are in kJ/mol.

changes in conformational free energy for intra-bp parameters, $\Delta G^{HGP,intra}_{i}$ (= $\sum_\theta \Delta G^{HGP,intra}_{i}$ (θ),where, θ runs over the intra-bp parameters) exhibit maximum conformational stabilization at T_6:A_{27} and the total changes in conformational entropy for a specific bp, $T\Delta S^{HGP,intra}_{i}$ (=$\sum_\theta T\Delta S^{HGP,intra}_{i}$ (θ), θ runs over the intra-bp parameters) exhibits maximum order at T_4:A_{29}. In the HG system, $\Delta G^{HGP,intra}_{i}$ and $T\Delta S^{HGP,intra}_{i}$ exhibit maximum conformational destabilization and disorder respectively at A_7:T_{26}.

Next we compute strand specific conformational thermodynamics due to sugar-phosphate, sugar-base and sugar-puckers torsion angles. For sugar-phosphate and sugar-base torsion angles, the conformational free energy and entropy, $\Delta G^{HGP,SP}_{i}$ (θ), and $T\Delta S^{HGP,SP}_{i}$ (θ) for i^{th} base along $5' - 3'$ strand in the HGP system with regard to the WC system are shown in Figs. 2.9 (a) and (b), re-

spectively. In Fig. 2.9 (a), we observe that $\Delta G_i^{HGP,SP}$ (α), $\Delta G_i^{HGP,SP}$ (β) and $\Delta G_i^{HGP,SP}$ (γ) are not sensitive to the bases. $\Delta G_i^{HGP,SP}$ (χ) contributes the most to stabilize the G_5. $\Delta G_i^{HGP,SP}$ (ϵ) and $\Delta G_i^{HGP,SP}$ (ζ) show maximum stabilization at A_7. Again, $\Delta G_i^{HGP,SP}$ (χ) contributes the most to stabilizing the A_7. However, $\Delta G_i^{HGP,SP}$ (δ) has maximum contribution to stabilizing T_6. We find in Fig. 2.9 (b) that $T\Delta S_i^{HGP,SP}$ (α), $T\Delta S_i^{HGP,SP}$ (χ), $T\Delta S_i^{HGP,SP}$ (δ) and $T\Delta S_i^{HGP,SP}$ (γ) do not contribute significantly to the bases. Otherwise, $T\Delta S_i^{HGP,SP}$ (β) has maximum order at A_8. $T\Delta S_i^{HGP,SP}$ (ϵ) contributes maximum to order the A_7, here. Again, $T\Delta S_i^{HGP,SP}$ (ζ) has maximum order and disorder at A_7 and T_{10}, respectively. The total changes in conformational free energy, $\Delta G_i^{HGP,SP}$ (= $\sum_\theta \Delta G_i^{HGP,SP}$ (θ), θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ, and ζ) shows stabilization from G_5 to A_7 with a maximum value at A_7 and the total changes in conformational entropy, $T\Delta S_i^{HGP,SP}$ (= $\sum_\theta T\Delta S_i^{HGP,SP}$ (θ), θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ and ζ) shows strong ordering from G_5 to A_7, with a maximum at A_7. A_7 shows maximum destabilization in $\Delta G_i^{HGP,SP}$ in the HG system.

In the complementary $3' - 5'$ strand, the conformational free energy and entropy changes for the HGP system with respect to the WC system caused by sugar-phosphate and sugar-base torsion angles, $\Delta G_{i_c}^{HGP,SP}$ (θ) and $T\Delta S_{i_c}^{HGP,SP}$ (θ) for the base i_c^{th} are shown in Figs. 2.9 (c) and (d), respectively. $\Delta G_{i_c}^{HGP,SP}$ (α), $\Delta G_{i_c}^{HGP,SP}$ (β), $\Delta G_{i_c}^{HGP,SP}$ (γ) and $\Delta G_{i_c}^{HGP,SP}$ (ϵ) do not show sensitivity to the bases in Fig. 2.9 (c). On the other hand, $\Delta G_{i_c}^{HGP,SP}$ (χ) has maximum destabilization and stabilization at A_{29} and A_{24}, respectively. Again, $\Delta G_{i_c}^{HGP,SP}$ (δ) shows maximum destabilization and stabilization at A_{29} and T_{25}, respectively. At T_{26}, $\Delta G_{i_c}^{HGP,SP}$ (ζ) shows maximum stabilization. In Fig. 2.9 (d), $T\Delta S_{i_c}^{HGP,SP}$ (α) and $T\Delta S_{i_c}^{HGP,SP}$ (γ) are not sensitive to bases. However, $T\Delta S_{i_c}^{HGP,SP}$ (β) contributes the most to ordering the C_{28}. On the other hand,$T\Delta S_{i_c}^{HGP,SP}$ (χ), $T\Delta S_{i_c}^{HGP,SP}$ (δ) and $T\Delta S_{i_c}^{HGP,SP}$ (ζ) contribute maximum to order the T_{26}. Again, $T\Delta S_{i_c}^{HGP,SP}$ (ϵ) contributes the most to organizing the T_{26}.

Thus, for the complementary strand, the total changes in conformational free energy, $\Delta G_{i_c}^{HGP,SP}$ (= $\sum_\theta \Delta G_{i_c}^{HGP,SP}$ (θ), θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ, and ζ) demonstrate the largest stability at T_{26}. On the other hand, the total changes in conformational entropy, $T\Delta S_{i_c}^{HGP,SP}$ (= $\sum_\theta T\Delta S_{i_c}^{HGP,SP}$, θ runs over sugar-base torsion angle χ along with sugar-phosphate backbone torsion angles α, β, δ, ϵ, γ, and ζ) show significant ordering from C_{28} to T_{26}, with maximum value at T_{26}.

T_{26} is stabilized and ordered more here compared to the HG system.

For sugar-puckers, the changes in conformational free energy, $\Delta G_i^{HGP,S}(\theta)$ and entropy, $T\Delta S_i^{HGP,S}(\theta)$, in the HGP system with regard to the WC system, of i^{th} base of $5' - 3'$ strand are shown in Figs. 2.9 (e) and (f), respectively. In Fig. 2.9 (e), $\Delta G_i^{HGP,S}(\nu 1)$, $\Delta G_i^{HGP,S}(\nu 2)$ and $\Delta G_i^{HGP,S}(\nu 3)$ exhibit maximum stabilization at T_6. $\Delta G_i^{HGP,S}(\nu 0)$ has maximum contribution to stabilize the G_5 and A_8. $\Delta G_i^{HGP,S}(\nu 4)$ has maximum destabilization at T_{10}. We find that in Fig. 2.9 (f), $T\Delta S_i^{HGP,S}(\nu 1)$ and $T\Delta S_i^{HGP,S}(\nu 2)$ do not sensitive to bases. But, $T\Delta S_i^{HGP,S}(\nu 0)$ and $T\Delta S_i^{HGP,S}(\nu 4)$ contribute the most to ordering the A_7. Again, $T\Delta S_i^{HGP,S}(\nu 3)$ has maximum contribution to order the T_6. The total changes in conformational free energy, $\Delta G_i^{HGP,S}$ ($= \sum_\theta \Delta G_i^{HGP,S}(\theta)$, θ runs over all sugar-puckers) exhibit significant stabilization at T_6 and A_8, while the total changes in conformational entropy, $T\Delta S_i^{HGP,S}$ ($= \sum_\theta T\Delta S_i^{HGP,S}(\theta)$, θ runs over all sugar-puckers) show maximum order at T_6 . Compared to the HG system, T_6 is more stabilized and ordered here.

For sugar pucker angles, the changes in conformational free energy, $\Delta G_{i_c}^{HGP,S}(\theta)$ and entropy, $T\Delta S_{i_c}^{HGP,S}(\theta)$ of i_c^{th} base of the complementary $3' - 5'$ strand of the HGP system with regard to the WC system are shown in Figs. 2.9 (g) and (h), respectively. In Fig. 2.9 (g), $\Delta G_{i_c}^{HGP,S}(\nu 0)$ and $\Delta G_{i_c}^{HGP,S}(\nu 3)$ do not contribute significantly to bases. However, $\Delta G_{i_c}^{HGP,S}(\nu 1)$ and $\Delta G_{i_c}^{HGP,S}(\nu 2)$ show maximum stabilization at T_{26}. Again, $\Delta G_{i_c}^{HGP,S}(\nu 4)$ exhibits the highest stability at A_{27}. In Fig. 2.9 (h), $T\Delta S_{i_c}^{HGP,S}(\nu 0)$, $T\Delta S_{i_c}^{HGP,S}(\nu 1)$ and $T\Delta S_{i_c}^{HGP,S}(\nu 2)$ are not sensitive to the bases. Only, $T\Delta S_{i_c}^{HGP,S}(\nu 3)$ and $T\Delta S_{i_c}^{HGP,S}(\nu 4)$ have maximum ordering at T_{26}. Here, the total changes in conformational free energy, $\Delta G_{i_c}^{HGP,S}$ ($= \sum_\theta \Delta G_{i_c}^{HGP,S}(\theta)$, θ runs over all sugar-puckers) show the most stabilization at T_{26} while the total changes in conformational entropy, $T\Delta S_{i_c}^{HGP,S}$ ($= \sum_\theta T\Delta S_{i_c}^{HGP,S}(\theta)$, θ runs over all sugar-puckers) exhibit significant ordering from C_{28} to T_{26}, with a maximum value at T_{26}. In the HG system, T_{26} shows destabilization and disorder in $\Delta G_{i_c}^{HGP,S}$ and $T\Delta S_{i_c}^{HGP,S}$, respectively. Here, sugar-phosphate, sugar-base, and sugar-pucker torsion angles contribute most to the changes in conformational thermodynamics.

We use cartoon presentations to represent the stable and flexible regions of the HG and HGP systems with respect to the WC system. This representation takes into account the combined effects of the sugar-phosphate, sugar-base, and sugar-pucker torsion angles. The stable and flexible zones of the HG and HGP systems with regard to the WC system are shown in a color-coded cartoon in Fig. 2.10. The bp T_4-A_{29} in the HG system exhibits slightly disordered while

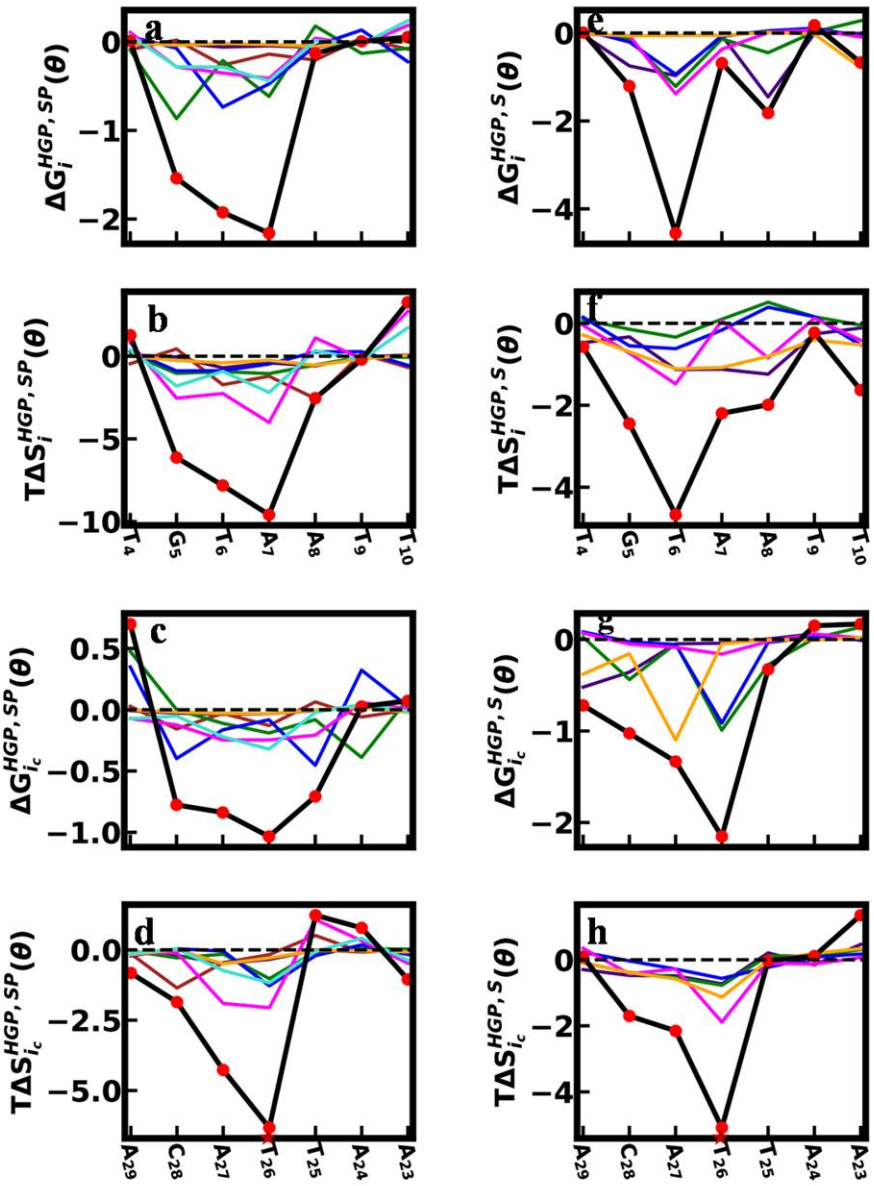


Figure 2.9: For sugar-phosphate and sugar-base torsion angles, (a) $\Delta G_i^{HG,SP}(\theta)$ and (b) $T\Delta S_i^{HG,SP}(\theta)$ for the $5'-3'$ strand, (c) $\Delta G_{i_c}^{HG,SP}(\theta)$ and (d) $T\Delta S_{i_c}^{HG,SP}(\theta)$ for the $3'-5'$ strand, in the HG system with respect to the WC system. Similarly, for sugar-puckers, (e) $\Delta G_i^{HG,S}(\theta)$ and (f) $T\Delta S_i^{HG,SP}(\theta)$ for the $5'-3'$ strand, (g) $\Delta G_{i_c}^{HG,S}(\theta)$ and (h) $T\Delta S_{i_c}^{HG,S}(\theta)$ for the $3'-5'$ strand, in the HG system with regard to the WC system. α is in indigo, β in brown, χ in green, δ in blue, ϵ in magenta, γ in orange, and ζ in turquoise in panels (a)-(d), $\nu 0$ is in indigo, $\nu 1$ in green, $\nu 2$ in blue, $\nu 3$ in magenta, and $\nu 4$ in orange in panels (e)-(h). The red circles joined by a solid black line in each panel reflect the overall changes for each base. HG bp region is marked by the maroon star in all panels. All quantities are in kJ/mol.

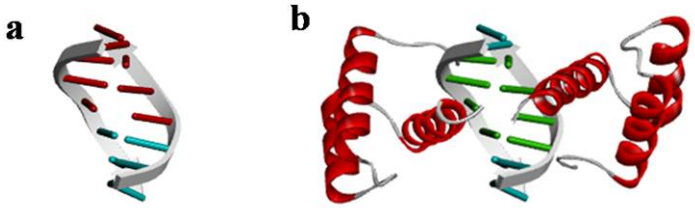


Figure 2.10: Color-coded cartoon presentation of the (a) HG and (b) HGP systems, with respect to the WC system. Compared to the WC system, slightly disordered or ordered bps of the HG system are shown in red, whereas significantly stabilized and ordered bps of the HGP system are shown in green.

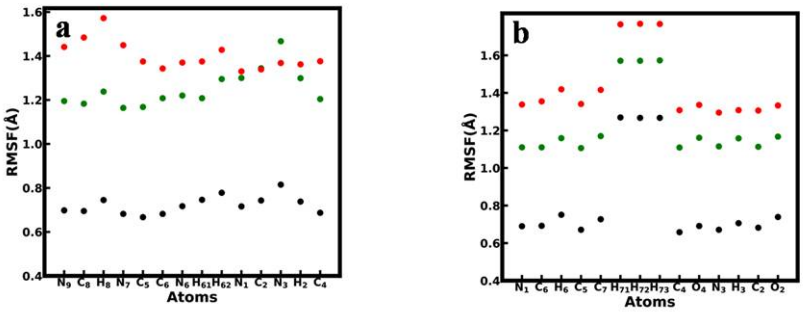


Figure 2.11: Root-mean square fluctuation in Åof atomic positions of (a) A_7 and (b) T_{26} for three different systems on equilibrated trajectories. In each panel WC data are in red, HG in green and HGP in black.

bps G_5-C_{28} to A_7-T_{26} exhibit slightly ordered. In the HGP system, bps G_5-C_{28} to A_8-T_{25} exhibit significant stabilization and order, with A_7-T_{26} being the most stabilized and ordered.

The RMSF of the atomic positions of the A_7 and T_{26} bases is shown in Fig. 2.11. In Fig. 2.11 (a), the atomic RMSF for A_7 for WC, HG, and HGP systems. Similarly, Fig. 2.11 (b) shows analogous data for T_{26}. We observe that the WC and HG systems do not significantly differ from one another, but the atoms in the HGP system exhibit much less fluctuations than the WC and HG systems. These fluctuations data are qualitatively consistent with the conformational thermodynamics data: The HG system is slightly destabilize than the WC system while the HGP system is significantly stabilized and ordered than the WC system in the A_7-T_{26} bp.

We compute the changes in conformational free energy and entropy, $\Delta G^{HG,inter}$ ($= \sum_i \Delta G^{HG,inter}_{i:i+1}$) and $T\Delta S^{HG,inter}$ ($= \sum_i T\Delta S^{HG,inter}_{i:i+1}$) for inter-bp parameters, $\Delta G^{HG,intra}$ ($= \sum_i \Delta G^{HG,intra}_{i}$) and $T\Delta S^{HG,intra}$ ($= \sum_i T\Delta S^{HG,intra}_{i}$) for intra-bp pa-

$\Delta G^{HG,inter}$	$T\Delta S^{HG,inter}$	$\Delta G^{HG,intra}$	$T\Delta S^{HG,intra}$	$\Delta G_d^{HG,SP}$	$\Delta G_c^{HG,SP}$	$T\Delta S_d^{HG,SP}$	$T\Delta S_c^{HG,SP}$	$\Delta G_d^{HG,S}$	$\Delta G_c^{HG,S}$	$T\Delta S_d^{HG,S}$	$T\Delta S_c^{HG,S}$	ΔG_{total}^{HG}	$T\Delta S_{total}^{HG}$
1.57	0.73	0.20	1.62	0.17	-0.61	-5.27	-2.57	0.02	-0.03	-1.1	2.78	1.32	-3.81
$\Delta G^{HGP,inter}$	$T\Delta S^{HGP,inter}$	$\Delta G^{HGP,intra}$	$T\Delta S^{HGP,intra}$	$\Delta G_d^{HGP,SP}$	$\Delta G_c^{HGP,SP}$	$T\Delta S_d^{HGP,SP}$	$T\Delta S_c^{HGP,SP}$	$\Delta G_d^{HGP,S}$	$\Delta G_c^{HGP,S}$	$T\Delta S_d^{HGP,S}$	$T\Delta S_c^{HGP,S}$	ΔG_{total}^{HGP}	$T\Delta S_{total}^{HGP}$
-2.41	-16.16	-0.60	-2.25	-5.69	-2.56	-21.74	-12.31	-8.74	-5.24	-13.74	-7.42	-25.23	-73.62

Table 2.15: The changes (kJ/mol) in conformational thermodynamics of HG and HGP systems in comparison to the WC system.

rameters, $\Delta G_d^{HG,SP}$ (= $\sum_i \Delta G_i^{HG,SP}$), $\Delta G_c^{HG,SP}$ (= $\sum_i \Delta G_{i_c}^{HG,SP}$), $T\Delta S_d^{HG,SP}$ (= $\sum_i T\Delta S_i^{HG,SP}$) and $T\Delta S_c^{HG,SP}$ (= $\sum_i T\Delta S_{i_c}^{HG,SP}$) for sugar-base and sugar-phosphate backbone torsion angles, $\Delta G_d^{HG,S}$ (= $\sum_i \Delta G_i^{HG,S}$), $\Delta G_c^{HG,S}$ (= $\sum_i \Delta G_{i_c}^{HG,S}$), $T\Delta S_d^{HG,S}$ (= $\sum_i T\Delta S_i^{HG,S}$) and $T\Delta S_c^{HG,S}$ (= $\sum_i T\Delta S_{i_c}^{HG,S}$) due to sugar-puckers, in the HG system with respect to the WC system. Here, ″d″ refers to the 5′-3′ strand and ″c″ refers 3′-5′ strand. The data are shown in Table 2.15. We find that most of the changes in conformational free energy and entropy are insignificant compared to the thermal energy ($\approx$ 2.5 kJ/mol) at room temperature. Only the conformational entropy, $T\Delta S_d^{HG,SP}$ and $T\Delta S_c^{HG,SP}$ show ordering due to sugar-base and sugar-phosphate torsion angles. Similarly, we show the analogous conformational thermodynamic quantities $\Delta G^{HGP,inter}$ (= $\sum_i \Delta G_{i:i+1}^{HGP,inter}$) and $T\Delta S^{HGP,inter}$ (= $\sum_i T\Delta S_{i:i+1}^{HG,inter}$), $\Delta G^{HGP,intra}$ (= $\sum_i \Delta G_i^{HGP,intra}$) and $T\Delta S^{HGP,intra}$ (= $\sum_i T\Delta S_i^{HGP,intra}$), $\Delta G_d^{HGP,SP}$ (= $\sum_i \Delta G_i^{HGP,SP}$), $\Delta G_c^{HGP,SP}$ (= $\sum_i \Delta G_{i_c}^{HGP,SP}$), $T\Delta S_d^{HGP,SP}$ (= $\sum_i T\Delta S_i^{HGP,SP}$) and $T\Delta S_c^{HGP,SP}$ (= $\sum_i T\Delta S_{i_c}^{HGP,SP}$), $\Delta G_d^{HGP,S}$ (= $\sum_i \Delta G_i^{HGP,S}$), $\Delta G_c^{HGP,S}$ (= $\sum_i \Delta G_{i_c}^{HGP,S}$), $T\Delta S_d^{HGP,S}$ (= $\sum_i T\Delta S_i^{HGP,S}$) and $T\Delta S_c^{HGP,S}$ (= $\sum_i T\Delta S_{i_c}^{HGP,S}$)in the HGP system with respect to the WC system in Table 2.15. We observe significant stabilization in $\Delta G_d^{HGP,SP}$, $\Delta G_d^{HGP,S}$ and $\Delta G_c^{HGP,S}$ and ordering in $T\Delta S^{HGP,inter}$, $T\Delta S_d^{HGP,S}$, $T\Delta S_c^{HGP,S}$, $T\Delta S_d^{HGP,S}$ and $\Delta G_c^{HGP,S}$ in the HGP system with respect to the WC system.

Finally, we compute ΔG_{total}^{HG} (=$\Delta G^{HG,inter}$ + $\Delta G^{HG,intra}$ + $\Delta G_d^{HG,SP}$ + $\Delta G_c^{HG,SP}$ + $\Delta G_d^{HG,S}$ + $\Delta G_c^{HG,S}$) and $T\Delta S_{total}^{HG}$ (=$T\Delta S^{HG,inter}$ + $T\Delta S^{HG,intra}$ + $T\Delta S_d^{HG,SP}$ + $T\Delta S_c^{HG,SP}$ + $T\Delta S_d^{HG,S}$ + $T\Delta S_c^{HG,S}$) values, for the entire HG system with respect to the WC system are 1.33 kJ/mol and -3.81 kJ/mol, respectively (Table 2.15.), while values of ΔG_{total}^{HGP} (=$\Delta G^{HGP,inter}$ + $\Delta G^{HGP,intra}$ + $\Delta G_d^{HGP,SP}$ + $\Delta G_c^{HGP,SP}$ + $\Delta G_d^{HGP,S}$ + $\Delta G_c^{HGP,S}$) and $T\Delta S_{total}^{HGP}$ (=$T\Delta S^{HGP,inter}$ + $T\Delta S^{HGP,intra}$ + $T\Delta S_d^{HGP,SP}$ + $T\Delta S_c^{HGP,SP}$ + $T\Delta S_d^{HGP,S}$ + $T\Delta S_c^{HGP,S}$) for the whole HGP system in comparison to the WC system are -25.23 kJ/mol and -73.62 kJ/mol, respectively (Table 2.15.). Overall, we observe from the total changes of conformational thermodynamics of the entire DNA system that the HG system is slightly destabilized but ordered that the WC system, while the HGP system is significantly stabilized and ordered in comparison to the WC system. The inter-bp step parameters are the primary factors to destabilize the HG system compared to the WC system, while sugar-

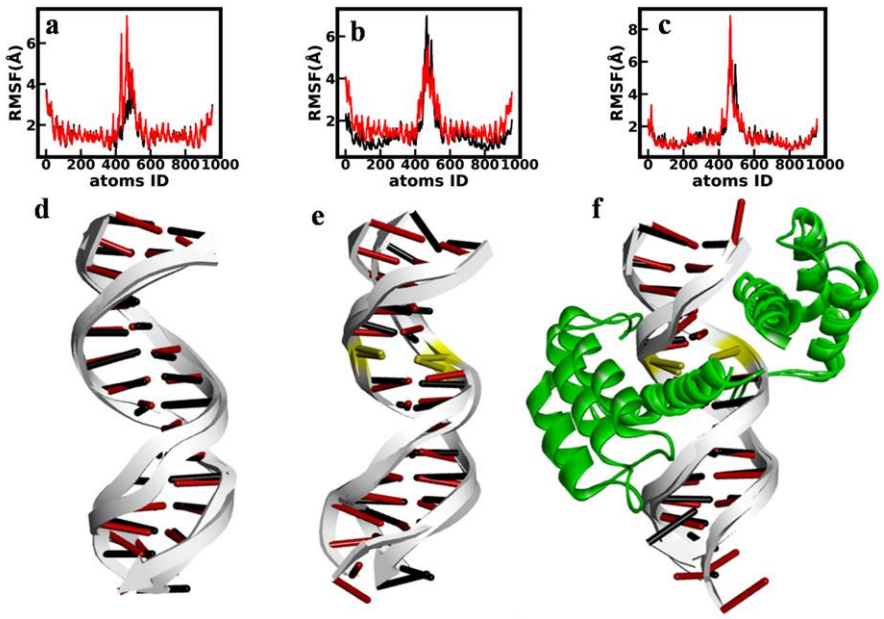


Figure 2.12: Root mean square fluctuation in Åof atomic positions of (a) WC, (b) HG, and (c) HGP systems and snapshot at 1μs time span of (d) WC, (b) HG, and (d) HGP systems, for two force fields OL15 and bsc1. OL15 is in black, and bsc1 is in red in all panels. The HG bp is marked by yellow in panel (e) and (f).

puckers are the main factors to stabilize the HGP system over the WC system.

We observe in Table 2.15. that for the 5′-3′ strand, $\Delta G_d^{HGP,SP}$, $\Delta G_d^{HGP,S}$, $T\Delta S_d^{HGP,SP}$ and $T\Delta S_d^{HGP,S}$ are more negative compared to $\Delta G_c^{HGP,SP}$, $\Delta G_c^{HGP,S}$, $T\Delta S_c^{HGP,SP}$ and $T\Delta S_c^{HGP,S}$, respectively, of the complementary 3′-5′ strand. This shows protein α2D has a greater impact on stabilizing and ordering the HGP system than protein α2B. Based on the fluctuation of the dihedral angle χ alone in the presence and absence of proteins, the earlier MD simulation on this system reveals that the non-specifically bound protein α2D has a stronger influence on the formation of HG bp.[12] In addition to χ, here we observe that the HG bp in the HGP system stabilized and ordered more for the microscopic degrees of freedom D_x, χ, δ, ϵ, ζ, $\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$. This gives a more complete microscopic picture of the stability at the HG bp induced by the bound protein. We repeat our simulations with another force field. We use the Amberff14SB_bsc1 force field. We compare the RMSF of the DNA atomic positions and 1 μs superimposed snapshots for two distinct force fields, Amberff14SB_OL15 and Amberff14SB_bsc1 for all systems Fig. 2.12. We observe no significant differences between Amberff14SB_OL15 and Amberff14SB_bsc1 results.

2.4 Conclusion

To summarize, the fluctuations of microscopic conformational variables reveal that the HG bp forming region as well as the entire DNA duplex become stabilized and ordered in the presence of a non-specifically bound protein α2D and a specifically bound protein α2B in the 1K61 system. Protein α2D is more effective than protein α2B at stabilizing and organizing the HGP system. Such study based on the probability distribution of proper conformational variable is a new approach to understanding the stability of such variable bp geometry due to non-canonical HG base pairing. Furthermore, the modulation of DNA-protein recognition due to non-canonical base pairing can also be implemented on the stability and order of tumor suppressor proteins like p53 or DNA lesion repair proteins interfered with by HG base pairing. It is highly interesting to analyze their effect on signal transduction and mismatch repair systems effectively in the future. Our study can also lead to a focus on anti-cancer drug target therapeutics acting through DNA mediated HG base pairing.

2.5 Appendix

2.5.1 Molecular dynamics(MD) simulation algorithm

MD simulation[49] is an essential tool for obtaining the time-dependent trajectory of a particle numerically. Consider a system of N particles in three dimensions, where $\vec{r_i} = \vec{r_1}, \vec{r_2}, \ldots, \vec{r_N}$ and $\vec{p_i} = \vec{p_1}, \vec{p_2}, \ldots, \vec{p_N}$ denote the position and momentum of particles $i = 1, 2, \ldots, N$. The mass of each particle is $m_i = m_1, m_2, \ldots, m_N$, and the total force acting on a particle is $\vec{F_i}$. Particles are assumed to interact via a conservative pair potential $V(r_{ij})$, where $r_{ij} = |\vec{r_i} - \vec{r_j}|$ depends only on the pairwise separation.

The force $\vec{F_i}$ acting on the i^{th} particle due to all other j^{th} particles is given by the gradient of $V(r_{ij})$:

$$\vec{F_i} = \sum_{j=1}^{N} -\nabla V(r_{ij}).$$

Thus, the time-dependent trajectory of a particle can be obtained using Newton's second law of motion:

$$\vec{F_i} = m_i \vec{a_i},$$

where a_i is the acceleration of the i^{th} particle.

In general, the Verlet algorithm[49] is widely used and is based on a central difference scheme. In this algorithm, the position of a particle, $\vec{r}(t + \Delta t)$, at a later time $t + \Delta t$, can be obtained from its position and acceleration at time t, as well as its position from the previous step, $\vec{r}(t - \Delta t)$.

Using Taylor series expansion,

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \Delta t \cdot \vec{v}(t) + \frac{1}{2}\Delta t^2 \cdot \vec{a}(t) + \dots \tag{2.1}$$

$$\vec{r}(t - \Delta t) = \vec{r}(t) - \Delta t \cdot \vec{v}(t) + \frac{1}{2}\Delta t^2 \cdot \vec{a}(t) + \dots \tag{2.2}$$

Finally, combining these equations gives,

$$\vec{r}(t + \Delta t) = 2\vec{r}(t) - \vec{r}(t - \Delta t) + \Delta t^2 \cdot \vec{a}(t) \tag{2.3}$$

The velocity can be obtained using the formula,

$$\vec{v}(t) = \frac{\vec{r}(t + \Delta t) - \vec{r}(t - \Delta t)}{2\Delta t} \tag{2.4}$$

Several algorithms are equivalent to the Verlet scheme, with the simplest being the Leap Frog algorithm. The algorithm computes the velocities at half-integer time steps and utilizes these velocities to compute the new positions. In these algorithm defining the velocities at half-integer time steps as follows:

$$\mathbf{v}(t - \Delta t/2) = \frac{\mathbf{r}(t) - \mathbf{r}(t - \Delta t)}{\Delta t} \tag{2.5}$$

and

$$\mathbf{v}(t + \Delta t/2) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t)}{\Delta t}. \tag{2.6}$$

From the equation 1.6 , we immediately obtain an expression for the new positions as follows:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \Delta t/2). \tag{2.7}$$

To update the velocities we use the following expression:

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t - \Delta t/2) + \Delta t \frac{\mathbf{f}(t)}{m}. \tag{2.8}$$

Since the Leapfrog algorithm is derived from the Verlet algorithm, it produces

identical trajectories. The trajectories obtained from MD simulations are useful for calculating both static and dynamic properties of a system.

2.5.2 Force-field used in bio-molecular simulation

In our study, we use GROMACS simulation package which use parallel computation for biomolecular simulation. Here, we perform all-atom MD simulations for naked DNA and DNA-protein systems. In GROMACS package, interactions between biomolecules, as well as their interactions with solvents or other biomolecules, are governed by well-established force fields[49] such as AMBER and GROMOS etc. The parameters of a force field are derived either from semi-empirical quantum mechanical calculations or by fitting experimental data such as X-ray diffraction, neutron diffraction, electron diffraction and NMR spectroscopy etc. The accuracy of simulation results depends on the appropriate selection of force fields. The force field contains both bonded and non-bonded interactions. Bonded interactions include bond stretching, bond rotations and torsional dihedrals using simple harmonic oscillations. Non-bonded interactions include Lennard-Jones (LJ) and Coulombic interactions. The form of the potential energy is:

$$\begin{aligned} V = & \sum_{\text{bonds}} k_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 \\ & + \sum_{\text{torsions}} k_\phi[1 + \cos(n\phi - \delta)] + \sum_{\text{improper}} k_w(w - w_0)^2 \\ & + \sum_{\text{LJ}} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \\ & + \sum_{\text{elec}} \frac{q_i q_j}{4\pi\epsilon_r \epsilon_0 r_{ij}} \end{aligned} \tag{2.9}$$

Here, the first term relates to the energy cost due to bond stretching, where k_b is the bond force constant and $(r - r_0)$ represents the deviation of the bond length from its equilibrium position r_0. Similarly, the second term accounts for changes in bond angles from the equilibrium value θ_0 with force constant k_θ. The third term represents dihedral interactions, where k_ϕ is the dihedral force constant, n is the multiplicity number, and δ is the phase shift. The fourth term corresponds to improper dihedral interactions, i.e., deviations of the out-of-plane bending from its equilibrium position w_0. The last two terms represent non-bonded interactions between pairs of atoms. The LJ potential is characterized

by ϵ, which defines the depth of the potential, and σ_{ij}, which determines the distance at which the intermolecular potential between two particles is zero. The last term accounts for Coulombic interactions, where q_i and q_j are the charges of the i^{th} and j^{th} particles. The distance between these particles is given by $r_{ij} = |r_i - r_j|$. Additionally, ϵ_0 and ϵ_r denote the permittivity of the vacuum and the relative permittivity, respectively.

2.5.3 Periodic boundary condition and minimum image convention

In all atom MD simulations, the system is initially placed in a central box. Now, in order to mimic truly infinite bulk system, periodic boundary conditions (PBC)[49] are applied in all directions. The finite size of the simulation box, L, can influence surface effects, causing particles near the boundaries to experience different forces compared to those in the bulk. To minimize this, the central simulation box is surrounded by infinite replicas. As a result, if an atom leaves the simulation box, its periodic image will enter through the opposite face. During the simulation, PBC is applied with the minimum image convention, ensuring that each particle interacts only with the closest periodic image of another atom among all the boxes. A cutoff distance ($r_c \approx L/2$) is introduced for truncating long-range interactions to avoid the interaction between an atom in central box with its mirror image in other replica boxes. This method reduces the computational cost of the simulation.

2.5.4 Berendsen thermostat

Berendsen thermostat[31] is widely used method for temperature control in NPT ensemble simulations. The Berendsen thermostat employs a weak coupling approach, where the system is subjected to first-order kinetics with an external heat bath at temperature T_0. This method gradually adjusts the system's temperature T according to:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.10}$$

where τ represents a time constant that governs the rate of temperature relaxation. A deviation from T_0 decays exponentially with time, and the coupling strength can be tuned based on user requirements. For rapid equilibration, a short coupling time (e.g., 0.01 ps) can be chosen, while for equilibrium simulations, a longer time constant (e.g., 0.5 ps) is preferred to minimally affect the system's conservative dynamics.

2.5.5 Parrinello-Rahman barostat

The Parrinello-Rahman barostat[33] is widely used to maintain constant pressure in NPT ensemble simulations. In the Parrinello-Rahman barostat, the box vectors, represented by the matrix, obey the matrix equation of motion:

$$\frac{d\mathbf{b}^2}{dt^2} = \mathbf{V}\mathbf{W}^{-1}\mathbf{b}^{-1}(\mathbf{P} - \mathbf{P}_{\text{ref}}). \tag{2.11}$$

The volume of the box is denoted as V, and $\mathbf{W}$ is a matrix parameter that determines the strength of the coupling. The matrices $\mathbf{P}$ and $\mathbf{P}_{\text{ref}}$ represent the current and reference pressures, respectively.

The conserved Hamiltonian is given by

$$E_{\text{pot}} + E_{\text{kin}} + \sum_i P_{ii}V + \sum_{i,j} \frac{1}{2} W_{ij} \left(\frac{db_{ij}}{dt} \right)^2. \tag{2.12}$$

The equations of motion for the atoms, derived from the Hamiltonian, are:

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M}\frac{d\mathbf{r}_i}{dt}, \tag{2.13}$$

where

$$\mathbf{M} = \mathbf{b}^{-1} \left[\mathbf{b}\frac{d\mathbf{b}}{dt} + \frac{d\mathbf{b}}{dt}\mathbf{b}^{-1} \right] \mathbf{b}^{-1}. \tag{2.14}$$

This additional term appearance as a friction.

The inverse mass parameter matrix, $\mathbf{W}^{-1}$, determines the strength of the coupling and how the box can deform. The box restriction is automatically fulfilled if the corresponding elements of $\mathbf{W}^{-1}$ are zero. Since the coupling strength depends on the box size, it is determined automatically when the isothermal compressibilities β_{ij} and the pressure time constant τ_p are provided as inputs:

$$(\mathbf{W}^{-1})_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L}. \tag{2.15}$$

2.5.6 Particle Mesh Ewald (PME) methods

Long-range interactions decays r^{-n}, with $n \leq 4$. We use the Particle Mesh Ewald (PME)[49] method to account for the long-range contributions of electrostatic interactions. In this method, the interaction potential is split into two parts: the long-range interaction is computed in Fourier space, while the short-range part is evaluated in real space. The discrete Fast Fourier Transform (FFT) is employed to

approximate the reciprocal-space term of the standard Ewald summation using a discrete convolution on an interpolating grid.

Electrostatic interactions between atoms or molecules are usually modeled by the Coulomb potential, which is given by:

$$U_{\text{electrostatic}}(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}$$

where ϵ_0 is the dielectric constant for vacuum, q_i and q_j are the partial charges on particles i and j, and r_{ij} is the distance between these two charges. Summation of long-range electrostatic potential for infinite neighboring atoms and periodic images are not convergent. To handle this, the charge distribution in the central simulation box is separated into two parts:

$$\rho_i(r) = \rho_i^S(r) + \rho_i^L(r) \tag{2.16}$$

where

$$\rho_i^S(r) = q_i\delta(r - r_i) - q_i G_\sigma(r - r_i), \quad \rho_i^L(r) = q_i G_\sigma(r - r_i).$$

Here, q_i is the charge at position r_i, and $G_\sigma(r - r_i) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{|r|^2}{2\sigma^2}}$ represents a Gaussian fictitious screening charge distribution. The summation for the short-range component is evaluated in real space, while the long-range component is approximated in Fourier space. In practice, the central simulation box is partitioned into a grid, and the point charges are spread across the nearest grid cells via interpolation. This results in a grid of uniformly spaced charges, which allows us to perform the FFT of the periodic images. The inverse fast Fourier transform is then performed to obtain the long-range contribution in real space.

2.5.7 Microscopic conformational variables of DNA

(I) DNA step parameters

There are six DNA bp step parameters, three are rotational parameters (tilt (τ), roll (ρ), and twist (ω)) and the other three are translational parameters (shift (D_x), slide (D_y), and rise (D_z)).[51] These are describing the position and orientation of one bp relative to the next bp. Shift is defined as the displacement of one bp with respect to the next along the X axis, while slide is the displacement about the Y axis. Rise is described as the displacement along the Z axis. Tilt, roll, and twist are defined as the rotation of the bp around shift, slide, and rise axes. The $5' - 3'$ direction of strand 1 is used as the positive Z direction. The long axis

(Y-axis) can be along the C6-C8 direction and passes through the bp center. The bp center is defined as the midpoint of C6 and C8 atoms of pyrimidines and purines respectively. The short axis (X-axis) points towards the major groove. The DNA bp step parameters are calculated using the formula:

$$\tau = -\sin^{-1}(\bar{Z}_m \cdot \bar{X}_1)$$

$$\rho = \sin^{-1}(\bar{Z}_m \cdot \bar{Y}_1)$$

$$\omega = \cos^{-1}\left(\frac{(\bar{X}_1 \times \bar{Z}_m) \cdot (\bar{X}_2 \times \bar{Z}_m)}{|(\bar{X}_1 \times \bar{Z}_m)||(\bar{X}_2 \times \bar{Z}_m)|}\right)$$

$$D_x = \bar{M} \cdot \frac{\bar{X}_1 + \bar{X}_2}{|\bar{X}_1 + \bar{X}_2|}$$

$$D_y = \bar{M} \cdot \frac{\bar{Y}_1 + \bar{Y}_2}{|\bar{Y}_1 + \bar{Y}_2|}$$

$$D_z = \bar{M} \cdot \bar{Z}_m$$

where $\bar{Z}_m$ is defined as:

$$\bar{Z}_m = \frac{(\bar{X}_1 + \bar{X}_2) \times (\bar{Y}_1 + \bar{Y}_2)}{|(\bar{X}_1 + \bar{X}_2)| \cdot |(\bar{Y}_1 + \bar{Y}_2)|}$$

Here, 1 and 2 denote two successive bp planes. $\bar{M}$ is the vector joining the bp centers of two consecutive base pairs.

(II) DNA intra-base pair parameters

There are six DNA base pair parameters; three are rotational parameters (buckle (κ), propeller (π), and opening (σ)) and three are translational parameters (stagger (S_x), shear (S_y), and stretch (S_z)).[52] These are describing the position and orientation of one base relative to other bases in a bp. The axis systems for the two bases are in a pair following the hydrogen bonding edge of the bases involved.[52] The base's X- axis is fixed perpendicular to the best mean plane across the base ring atoms, which can be along or against the $5' - 3'$ strand direction. Each base in a pair's Y-axis is defined by two hydrogen-bonding heavy atoms of the specific edge forming the hydrogen bonds. A base's Z-axis is set perpendicular to both the X- and Y-axes in a right-handed axis system, and it passes roughly

parallel to the hydrogen bonds formed in the pair. The mathematical expression used in calculating buckle is analogous to that used for calculating tilt, while the expressions for opening, propeller, shear, stagger and stretch are identical to those used for calculating roll, twist, slide, shift and rise, respectively, but with the axes for the bases rather than the bps. The DNA bp parameters are calculated using the formula:

$$\kappa = -2\sin^{-1}(\bar{Z}_m \cdot \bar{Y}_1)$$

$$\sigma = -2\sin^{-1}(\bar{Z}_m \cdot \bar{X}_1)$$

$$\pi = \cos^{-1}\left((\bar{X}_1 \times \bar{Z}_m) \cdot (\bar{X}_2 \times \bar{Z}_m)\right)$$

$$S_x = -\bar{Y}_m \times \bar{M}$$

$$S_y = \bar{X}_m \cdot \bar{M}$$

$$S_z = \bar{Z}_m \cdot \bar{M}$$

where X_1, Y_1 and Z_1 are unit vectors along the axes of the first base, X_2, Y_2 and Z_2 are those for the second base. The vector $\bar{M}$ is created by combining two base atoms, one from each base in the pair and chosen based on the base's specific hydrogen-bonding edge. The components of the mean unit vector X_m, Y_m and Z_m are calculated as follows:

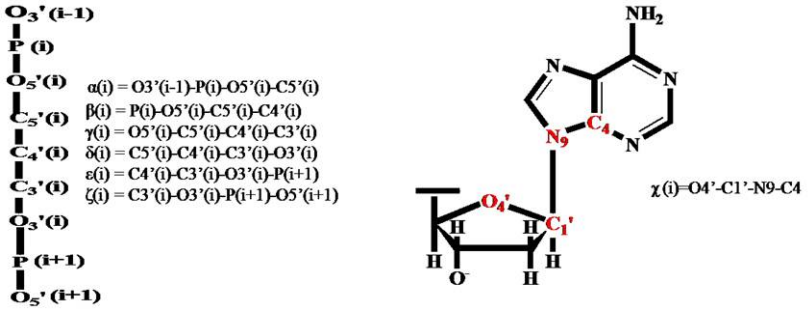
$$\bar{X}_m = \frac{\bar{X}_1 + \bar{X}_2}{|\bar{X}_1 + \bar{X}_2|}$$

$$\bar{Y}_m = \frac{\bar{Y}_1 + \bar{Y}_2}{|\bar{Y}_1 + \bar{Y}_2|}$$

$$\bar{Z}_m = \frac{(\bar{X}_1 + \bar{X}_2) \times (\bar{Y}_1 + \bar{Y}_2)}{|(\bar{X}_1 + \bar{X}_2) \times (\bar{Y}_1 + \bar{Y}_2)|}$$

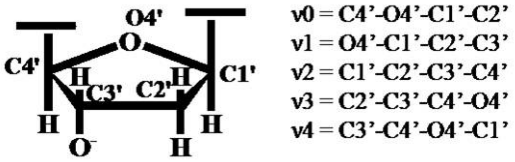
(III) Sugar-phosphate and sugar-base torsion parameters

In DNA, there are six types of sugar-phosphates (alpha (α), beta (β), gamma (γ), delta (δ), epsilon (ϵ) and zeta (ζ)) and one type of sugar-base (chi (χ)) torsion angles.[53] They are as follows:



(IV) Sugar-pucker angles

For each sugar in DNA, there are five torsion angles ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$).[54] They are as follows:



(V) Pseudo rotation phase angle (P)

The five sugar pucker angles ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$) are used to determine the pseudo rotation phase angle (P) for a sugar.[54] It is defined as:

$$P = \tan^{-1}\left(\frac{(v_2 + v_4) - (v_1 + v_3)}{2v_0(\sin 36^\circ + \sin 72^\circ)}\right)$$

P is widely different in various forms of DNA, such as A-DNA, B-DNA, and Z-DNA.[55] P is classified as C2'-endo when it falls between 137° and 194° and as C3'-exo when it falls between 195° and 216°. In B-DNA the sugar conformation is generally C2'-endo.

Bibliography

[1] Hoogsteen, K. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. Acta Crystallogr. **1963**, *16*, 907-916.

[2] Abrescia, N. G.; Thompson, A.; Huynh-Dinh, T.; Subirana, J. A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. PNAS **2002**, *99*, 2806-2811.

[3] Panczyk, T. Nieszporek, K. Wolski, P. Stability and Existence of Noncanonical I-motif DNA Structures in Computer Simulations Based on Atomistic and Coarse-Grained Force Fields, Mol. **2022**, *27*, 4915.

[4] Nikolova, E. N. Kim, E. Wise, A. A. O'Brien, P. J. Andricioaei, I. Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. Nature **2011**, *470*, 498-502.

[5] Rice, P. A. Yang, S. W. Mizuuchi, K. Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell **1996**, *87*, 1295-1306.

[6] Nikolova, E. N. Zhou, H. Gottardo, F. L. Alvey, H. S. Kimsey, I. J. Al-Hashimi, H. M. A historical account of hoogsteen base-pairs in duplex DNA. Biopolymers **2013**, *99*, 955-968.

[7] Alvey, H. S. Gottardo, F. L. Nikolova, E. N. Al-Hashimi, H. M. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. Nat. commun. **2014**, *5*, 4786.

[8] Nair, D. T. Johnson, R. E. Prakash, S. Prakash, L. Aggarwal, A. K. Replication by human DNA polymerase-l occurs by Hoogsteen base-pairing. Nature **2004**, *430*, 377-380.

[9] Portugal, J. Do Hoogsteen base pairs occur in DNA?, Trends Biochem. Sci. **1989**, *14*, 127-130.

[10] Kitayner, M. Rozenberg, H. Rohs, R. Suad, O. Rabinovich, D. Honig, B. Shakked, Z. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs, Nat. Struct. Mol. Biol. **2010**, *17*, 423-429.

[11] Patikoglou, G. A. Kim, J. L. Sun, L. Yang, S. H. Kodadek, T. Burley, S. K. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution, Genes Dev. **1999**, *13*, 3217-3230.

[12] Aishima, J. Gitti, R. K. Noah, J. E. Gan, H. H. Schlick, T. Wolberger, C. A Hoogsteen base pair embedded in undistorted B-DNA, Nucleic Acids Res. **2002**, , 5244-5252.

[13] Wolberger, C. Homeodomain interactions. Curr. Opin. Struct. Biol. **1996**, *6*, 62-68.

[14] Bürglin, T. R.; Affolter, M. Homeodomain proteins: an update. Chromosoma **2016**, *125*, 497-521.

[15] Banerjee-Basu, S. Baxevanis, A. D. Molecular evolution of the homeodomain family of transcription factors. Nucleic Acids Res. **2001**, *29*, 3258-3269.

[16] Das, A. Chakrabarti, J. Ghosh, M. Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation. Biophys. J. **2013**, *104*, 1274-1284.

[17] Mandal, S. C. Maganti, L. Mondal, M. Chakrabarti, J. Microscopic insight to specificity of metal ion cofactor in DNA cleavage by restriction endonuclease EcoRV. Biopolymers **2020**, *111*, e23396.

[18] Das, A. Chakrabarti, J. Ghosh, M. Conformational thermodynamics of metal-ion binding to a protein. Chem. Phys. Lett. **2013**, *581*, 91-95.

[19] Sikdar, S. Chakrabarti, J. Ghosh, M. A microscopic insight from conformational thermodynamics to functional ligand binding in proteins. Molecular Biosyst. **2014**, *10*, 3280-3289.

[20] Maganti, L. Dutta, S. Ghosh, M. Chakrabarti, J. Allostery in Orai1 binding to calmodulin revealed from conformational thermodynamics. J. Biomol. Struct. **2019**, *37*, 493-502.

[21] Sikdar, S. Chakrabarti, J. Ghosh, M. Conformational thermodynamics guided structural reconstruction of biomolecular fragments. Molecular BioSyst. **2016**, *12*, 444-453.

[22] Patra, P. Ghosh, M. Banerjee, R. Chakrabarti, J. Anion induced conformational preference of CαNN motif residues in functional proteins. Proteins: Struct. Funct. Gent. **2017**, *85*, 2179-2190.

[23] Das, A. Chakrabarti, J. Ghosh, M. Thermodynamics of interfacial changes in a protein-protein complex, Molecular Biosyst. **2016**, *10*, 437-445.

[24] Mallick Gupta, A. Mandal, S. Mandal, S. Chakrabarti, J. Immune escape facilitation by mutations of epitope residues in RdRp of SARS-CoV-2. J. Biomol. Struct. **2023**, *41*, 3542-3552.

[25] Dutta, S. Ghosh, M. Chakrabarti, J. In-silico studies on conformational stability of flagellin-receptor complexes. J. Biomol. Struct. **2020**, *38*, 2240-2252.

[26] Arnott, S. Campbell-Smith, P. J. Chandrasekaran, R. In Handbook of Biochemistry and Molecular Biology, 3rd ed. Nucleic Acids–Volume II, G.P. Fasman, Ed. Cleveland: CRC Press, **1976**, *17*, 411-422.

[27] Maier, J. A. Martinez, C. Kasavajhala, K. Wickstrom, L. Hauser, K. E. Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, J. Chem. Theory Comput. **2015**, *11*, 3696-3713.

[28] Galindo-Murillo, R. Robertson, J. C. Zgarbovic, M. Sponer, J. Otyepka, M. Jureska, P. Cheatham, T. E. Assessing the current state of amber force field modifications for DNA. J. Chem. Theory Comput. **2016**, *12*, 4114-4127.

[29] Abraham, M. J. van der Spoel, D. Lindahl, E. Hess B. and the GROMACS development team, GROMACS User Manual version 2018, `www.gromacs.org` **2018**.

[30] Dai, L. Mu, Y. Nordenskiöld, L. van der Maarel, J. R. Molecular dynamics simulation of multivalent-ion mediated attraction between DNA molecules. PRL **2008**, *100*, 118301.

[31] Petrova S. S. Solov'ev, A. D. The Origin of the Method of Steepest Descent Historia Mathematica **1997**, *24*, 361.

[32] Lemak, A. S. Balabaev, N. K. On the Berendsen thermostat, Mol. Simul. **1994**, *13*, 177-187.

[33] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. **1997**, *18*, 1463-1472.

[34] Saito, H. Nagao, H. Nishikawa, K. Kinugawa, K. Molecular collective dynamics in solid para-hydrogen and ortho-deuterium: The Parrinello-Rahman-type path integral centroid molecular dynamics approach. J. Chem. Phys. **2003**, *119*, 953-963.

[35] Mukherjee, S. Bansal, M. Bhattacharyya, D. Conformational specificity of non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis. J. Comput. Aided Mol. Des. **2006**, *20*, 629-645.

[36] Altona, C. T. Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. JACS **1972**, *94*, 8205-8212.

[37] Singh, H. Raja, A. Prakash, A. Medhi, B. Gmx qk: An Automated Protein/Protein-Ligand Complex Simulation Workflow Bridged to MM/PBSA, Based on Gromacs and Zenity-Dependent GUI for Beginners in MD Simulation Study. J. Chem. Inf. Model. **2023**, *63*, 2603-2608.

[38] `https://github.com/snbsoftmatter/confthermo`

[39] `https://manual.gromacs.org/documentation/5.1/onlinehelp/gmx-hbond.html`

[40] Biovia, D. S. Discovery studio modeling environment **2017**.

[41] Pingali, P. K. Halder, S. Mukherjee, D. Basu, S. Banerjee, R. Choudhury, D. Bhattacharyya, D. Analysis of stacking overlap in nucleic acid structures: algorithm and application, J. Comput. Aided Mol. Des. **2014**, *28*, 851-867.

[42] Samanta, S. Mukherjee, S. Chakrabarti, J. Bhattacharyya, D. Structural properties of polymeric DNA from molecular dynamics simulations. J. Chem. Phys. **2009**, *130*, 03B614.

[43] Tabernero, L. Bella, J. Alemán, C. Hydrogen bond geometry in DNA-minor groove binding drug complexes. Nucleic Acids Res. **1996**, *24*, 3458-3466.

[44] Pérez de Alba Ortíz, A. Vreede, J. Ensing, B. Sequence dependence of transient Hoogsteen base pairing in DNA, PLoS Comput. Biol. **2022**, *18*, e1010113.

[45] Zhou, H. Hintze, B. J. Kimsey, I. J. Sathyamoorthy, B. Yang, S. Richardson, J. S. Al-Hashimi, H. M. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey, Nucleic Acids Res. **2015**, *43*, 3420-3433.

[46] Yang, C. Kim, E. Lim, M. Pak, Y. Computational probing of Watson-crick/hoogsteen breathing in a DNA duplex containing N1-methylated adenine. J. Chem. Theory Comput. **2018**, *15*, 751-761.

[47] Yang, C. Kim, E. Pak, Y. Free energy landscape and transition pathways from Watson-Crick to Hoogsteen base pairing in free duplex DNA, Nucleic Acids Res. **2015**, *43*, 7769-7778.

[48] Várnai, P. Djuranovic, D. Lavery, R. Hartmann, B. α/γ Transitions in the B-DNA backbone, Nucleic Acids Res. **2002**, *30*, 5398-5406.

[49] Frenkel, D. Smit, B. Understanding molecular simulation: from algorithms to applications. Elsevier **2023**.

[50] `https://manual.gromacs.org/2025.0/index.html`

[51] Bansal, M. Bhattacharyya D. Ravi, B. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures, Bioinformatics **1995**, *11*, 281-287.

[52] Mukherjee, S. Bansal, M. Bhattacharyya, D. Conformational specificity of non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis. J. Comput. Aided Mol. Des. **2006**, *20*, 629-645.

[53] `https://iupac.qmul.ac.uk/misc/pnuc2.html`

[54] `https://x3dna.org/highlights/torsion-angles-of-nucleic-acid-struc`

[55] Dickerson, R. E. [5] DNA structure from A to Z, Methods Enzymol. **1992**, *211*, 67-111.

Chapter 3 Binding energy of homeodomain proteins to Hoogsteen base pair *

3.1 Introduction

In the previous chapter we show that homeodomain protein give conformational stability of Hoogsteen (HG) base pair (bp) relative to Watson-Crick (WC) bp. Understanding the binding energy (B_E) of homeodomain proteins with DNA is crucial for quantitatively evaluating their stability in presence of DNA. The B_E of homeodomain proteins with DNA are not well documented in the literature.

The B_E of a protein with DNA is the work done to separate the protein from DNA to an infinite distance away. This energy determines the overall stability of the complex and is a crucial parameter in understanding biological processes such as transcription,[1,2] replication[3,4] and DNA repair.[5,6] Experimentally, Isothermal Titration Calorimetry (ITC)[7,8] and Nuclear Magnetic Resonance (NMR)[9] spectroscopy are widely used techniques for measuring the binding energy (B_E) of protein with DNA. Computationally, Molecular Dynamics (MD)[10] simulations provide a powerful approach for accurately estimating the B_E.

Binding energy (B_E) of protein to WC-DNA has been reported. B_E of Lac repressor protein in a DNA-Lac repressor complex, wherein Lac repressor protein bound to a specific site of WC-DNA has been measured experimentally[11] . In simulations, the B_E between protein and DNA is calculated using Adaptive Biasing Force (ABF),[12] Umbrella Sampling (US)[13] etc enhanced sampling techniques.

*Based on the publication: Kanika Kole and Jaydeb Chakrabarti. Binding of Homeodomain Proteins to DNA with Hoogsteen Base Pair. The Journal of Physical Chemistry B 129, 2025, 1544-1554. (https://doi.org/10.1021/acs.jpcb.4c08054)

Based on ABF method, the B_E of proteins in specific and non-specific structures of WC-DNA-Lac repressor complexes has been computed.[12] Approximately 5 kcal/mol difference in B_E between specific and nonspecific complexes has been observed mainly due to the differences in protein conformations, particularly in the C-terminal tail. Recently, B_E data of disordered and globular homeodomain proteins with WC-DNA in the model structure of Nkx2.5 HD-Specific double stranded DNA complexes have also been reported[13] using US method. Significant differences has been observed between the binding of disordered and globular proteins due to differences in the numbers of inter- and intramolecular contacts. However, the B_E of proteins binding to HG-DNA has not been reported so far.

Here we consider several systems: α2D and α2B bound to HG and WC-DNA. Further we consider mutated α2D protein. We examine the interfacial interactions between α2D and α2B proteins with HG-DNA (PDB id-1K61)[14] and WC-DNA using conventional MD trajectories and the binding energies of the proteins using the US method.[15–23] Our primary objective is to find how the interfacial stability and order affect the binding energy in all these cases. We observe the decrease in interfacial interaction for WC-DNA compared to HG-DNA. Mutations in the N-terminal arm of the α2D protein affect the interfacial interactions between proteins and HG-DNA. We compute the B_E of α2D and α2B proteins with HG-DNA, where one of the proteins remains bound to HG-DNA. We observe that both proteins have almost similar B_E (approximately 37 kcal/mol) to HG-DNA, in the presence of the other protein. Further, the B_E of both the protein decreases in the presence of the WC-DNA bound to the other protein. We also consider the cases where only one protein binds to HG-DNA in the absence of the other protein. The B_E of the protein is lowered suggesting α2D and α2B bind cooperatively to HG-DNA. Next we check the B_E data of neutral and acidic mutated α2D protein with HG-DNA where α2B remains bound to the HG-DNA. Both the neutral and acidic mutations in the basic N-terminal arm of the α2D protein decreases the B_E of the α2D protein. We further examine the conformational stability and order of both the α2D and α2B proteins at the interface with the DNA with respect to their free states using conformational thermodynamics calculation[24–28] from the conventional MD trajectories. Conformational thermodynamics data show that most of the residues of both the proteins, α2D and α2B, stabilize and order in the HG-DNA bound state with respect to their free states. We calculate the conformational free energy and entropy at the interface by summing the conformational thermodynamics[24–30] data of the interface forming protein residues and the corresponding DNA regions. Considering different cases here,

we find that the B_E positively correlates with the conformational thermodynamics at the interface.

3.2 Methods and Analysis

3.2.1 Methods

(a) System preparation

We study the following systems: (1) The HGP system consists of 15 bp DNA fragments, with one specifically bound α2B and one non-specifically bound α2D homeodomain proteins taken from PDB ID 1K61. DNA follows the sequence $5'C_2$ATGT<u>A</u>ATTCA TTTA$_{16}3'$ in the $5'$-$3'$ direction. The underlined A_7 forms HG base pairing with the complementary base T_{26}. There are fifty-nine (132-190) residues in the specific protein α2B, while fifty-eight (132-189) residues in the non-specific protein α2D. (2) The free-WTD contains only the wild-type protein α2D. (3) The free-WTB contains the wild-type protein α2B only. (4) In the WCP system, the DNA sequence, and the protein residues are identical to the HGP system, but the DNA bases have WC type base pairing built by Discovery Studio software.[31] (5) We mutate the basic ARG132 residue by neutral ASN132 (ARG132ASN) using PyMOL software[32] in the α2D protein of the HGP system to form the NMUTP system. (6) Similarly, using PyMOL,[32] in another case, we mutate the basic ARG132 of the α2D protein in the HGP system by acidic residue ASP132 (ARG132ASP) to create the AMUTP system.

(b) Simulations

(I) MD simulations

The GROMACS[33] 2018.6 package[34] with Amberff14SB_OL15[35,36] force field (Appendix 2.5.2) is used for our simulation. Initially, the structure of the protein-DNA complex or protein is solvated in a cubic water box with the dimension of $8.1 \times 8.1 \times 8.1$ nm^3 containing TIP3P water molecules. The periodic boundary condition (Appendix 2.5.3) is used in all three spatial dimensions. The system is electrically neutralized by adding the required number of sodium (Na^+) and chloride (Cl^-) ions. The potential energy of the system is minimized using a steepest descent algorithm.[37] Then all-atom conventional molecular dynamics (MD) simulation (Appendix 2.5.1) is performed at 300K temperature and 1 atmosphere pressure maintaining an isothermal-isobaric (NPT) ensemble for

1μs for the systems: free-WTD, free-WTB, WCP, NMUTP, and AMUTP systems. We use the MD trajectory of the HGP system from our earlier study.[30] During MD, we use a Berendsen thermostat[38] (Appendix 2.5.4) and a Parrinello-Rahman barostat[39] (Appendix 2.5.5) to maintain constant temperature and pressure, respectively. The Lennard-Jones (LJ) and short-range electrostatic interactions are terminated at 10 Å. The Particle-Mesh Ewald[40] (PME) method (Appendix 2.5.6) is used to calculate the long-range electrostatic interactions. The LINCS constraints[41] are applied to all bonds involving hydrogen atoms. 2 fs time step is used for our simulation. To create a similar ensemble, we adjust the total number of water molecules to ensure that the number of total atoms (N=51068) remains the same in all the cases. In MD simulations, the equilibration of the system is confirmed by the saturation of the root mean square deviation (RMSD) with time. We consider the equilibrated part of the trajectory (200ns-1μs) for analysis.

(II) Steered MD and US simulations

For the US simulation (Appendix 3.5.1), the outputs of the steered molecular dynamics (SMD) run are required. Therefore, we run the SMD first. We use GROMACS[20,33] software to run SMD[42] and US. In SMD, initially, the system is placed in a box of dimensions $8.1 \times 19.1 \times 8.1$ nm^3. The prepared box is solvated, neutralized, and then energy minimized using the steepest descent minimization method. The box is further equilibrated at a particular temperature of 300K, and at a particular pressure of 1 bar. An external harmonic potential is used to pull the center of mass of the protein from center of mass of the remaining part of the system toward the bulk solvent along the Y axis (perpendicular to the central axis of the DNA double helix). DNA is fixed and aligned along the Z-axis while the protein is pulled. The final center of mass distance between the protein and the remaining part of the system is approximately 7 nm is achieved. A total of 500 configurations are generated from the output of the one SMD run.

From the SMD trajectory, snapshots are taken to generate the starting configurations for the US windows. We use the symmetric distribution of the sampling windows for the US with the spacing 0.2 nm.[28] The distance between the center of mass of the protein and the remaining part of the system is used as the reaction coordinate for US. In the US simulation, the sampling of intermediate states is enforced by applying an additional biasing potential limiting the sampling of the phase space to the specified reaction coordinate range.[15] The biasing potential is then removed when the potential mean force (PMF) is computed using the

Weighted Histogram Analysis Method (WHAM) (Appendix 3.5.2).[16,17] The B_E is extracted from the difference between the highest and lowest values of the PMF curve.[18–23] First, each US window is independently equilibrated for 100 ps, maintaining the NPT ensemble.[20,28,43] Next, MD is performed for each individual window. Then compute PMF from US simulation data utilizing GROMACS. We perform three independent US runs utilizing the configurations of three SMD runs for each of the cases, maintaining the same protocol.

We fix the parameters for SMD and US simulation considering B_E calculation of the α2D protein with HG-DNA. In SMD, we use an external harmonic potential of the form k (ζ_D - ζ_{D0})^2 to pull the protein α2D from the rest of the HGP system along the Y-axis. ζ_D (nm) is the reaction coordinate that represents the center of mass distance between α2D and rest of the HGP system. $\zeta_D 0$ (nm) is the center of the harmonic potential which is the initial center of mass distance between the α2D and the rest of the HGP system. The spring constant, k, of 1000 kJ/mol nm[2][28,44,45] and a pull rate of 0.01 nm/ps.[20,43] Using a slower pulling rate (0.005 nm/ps) results in the production of similar force vs time curve (Fig. 3.1(a)).[21] We also repeat our calculation using the spring constant at higher value, 1200 kJ/mol nm^2, and at a lower value, 800 kJ/mol nm^2, respectively from the spring constant value 1000 kJ/mol nm^2 fixing the pull rate of 0.01 nm/ps. Fig. 3.1(b) shows the force vs. time curve in SMD for various spring constants at the pull rate 0.01 nm/ps. We get an almost similar type of force vs time curve for the spring constant of 1000 and 800 kJ/mol nm^2. For the spring constant of 1200 kJ/mol nm^2, the force vs. time curve is different from the curve of 1000 and 800 kJ/mol nm^2 spring constant. Therefore, in SMD, we take the faster pulling rate 0.01 nm/ps[43] and the spring constant 1000 kJ/mol nm^2[43] for expedite data collection preserving reliability of the results following GROMACS tutorial.[20]

Then we run US simulation to compute the PMF as well as the B_E of the α2D protein with the rest of the HGP system. In the US, configurations from the SMD are used to set up the initial configurations, with the same reaction coordinate, ζ_D. We run three independent US simulations maintaining the same protocol using the outputs of the three SMD runs having the pulling rate of 0.01 nm/ps and spring constant 1000 kJ/mol nm^2. The histogram for three independent US simulations in case of PMF calculations are shown Figs. 3.2 (a) - (c), respectively.

The systems simulated in conventional MD and SMD/US simulations are listed in Tables 3.1 and 3.2, respectively.

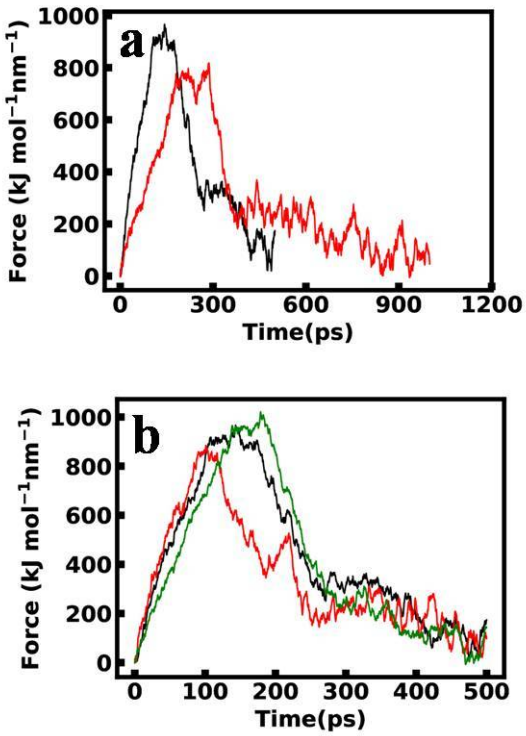


Figure 3.1: (a)Force vs time plot using different the pulling rates during pulling of the α2D protein from the remaining part of the HGP system in SMD. v = 0.01 nm/ps is in black and v = 0.005 nm/ps is in red. (b) Force vs time plot in SMD using different force constants during pulling of the α2D protein of the HGP system. k = 1000 kJ/mol nm^2 is in black, k = 1200 kJ/mol nm^2 is in red, and k = 800 kJ/mol nm^2 is in green.

Sl. No.	System
1	HGP
2	free-WTD
3	free-WTB
4	WCP
5	NMUTP
6	AMUTP

Table 3.1: Systems used for conventional MD simulations.

3.2.2 Analysis

We calculate the following quantities over equilibrated trajectories.

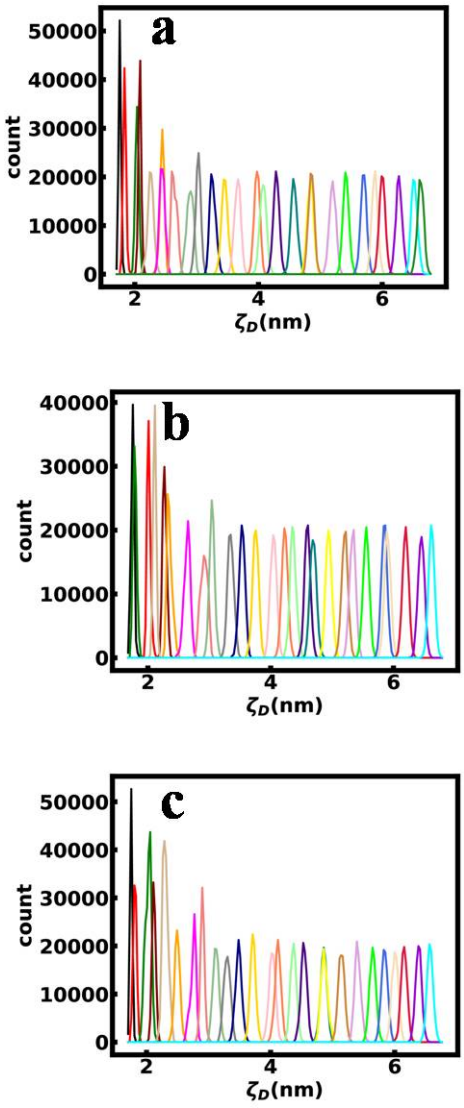


Figure 3.2: Representative histogram for the PMF calculation in the US simulation of the α2D protein with HG-DNA in the presence of the α2B protein, (a) for US-1, (b) for US-2 and (c) for US-3.

(a) Interfacial interactions

When two atoms, one from protein (α2B or α2D) residues and the other from DNA bases, are within 0.6 nm of one another, the interface is formed.[30] We use the equilibrated portion of conventional MD trajectories for this calculation. Here, we consider H-bond,[46] electrostatic,[47] and salt bridge[48] type interactions like our previous paper.[30] Distance and angle criteria are used to characterize the H-bond. For H-bond the distance between donor (D) and acceptor (A) atoms is $\leq$ 0.35

Sl. No.	System
1	Protein α2D with remaining part of the HGP system
2	Protein α2B with rest of the HGP system
3	Protein α2D with remaining portion of the WCP system
4	Protein α2B with remaining portion of the WCP system
5	Protein α2D with HG-DNA
6	Protein α2B with HG-DNA
7	Protein α2D with remaining part of the NMUTP system
8	Protein α2B with rest of the NMUTP system
9	Protein α2D with remaining portion of the AMUTP system
10	Protein α2B with rest of the AMUTP system

Table 3.2: Systems used for SMD/US-MD simulations.

nm, and donor-hydrogen-acceptor (D-$\hat{H}$-A) antecedent angle cut off is $\geq 150^{\circ}$. When the distance between oppositely charged atoms of DNA base and protein residue is less than 0.56 nm, the pair is counted for electrostatic interaction. A salt bridge is identified when a hydrogen bond and an electrostatic interaction occur simultaneously between oppositely charged groups of a DNA base and a protein residue. This requires a distance of within 0.4 nm between the groups and an angle $\geq 150^{\circ}$. We utilize the GROMACS H-bond analysis module for H-bond and Discovery Studio software for salt bridge and electrostatic interaction to define the cutoff criterion.[30] We divide the equilibrated part of the trajectory into N_w (=5) windows with an equal number of conformations. For each window, we calculate the average ratio of the number of frames in each non-bonded interaction to the total number of frames for each residue involved in forming the interface. The error in the mean is given by $\pm\, 0.5\sigma_m/\sqrt{N_w}$, where σ_m is the standard deviation of the mean value over the windows.

(b) Potential mean force and binding energy

The PMF is computed using the weighted histogram analysis method (WHAM)[16–19,43] available through the GROMACS[20] utilizing the outputs from the US simulation. The difference between the highest and lowest values of the PMF curve is the binding energy, B_E.[18–23,43] For each of the B_E calculations, we compute the average PMF with error from three US simulations using the outputs of three independent SMD runs. The error in the average PMF is computed by $\pm\, 0.5\sigma_m/\sqrt{N_w}$, where σ_m is the standard deviation of the average PMF value over the three US trajectories and N_w (=3) is the number of independent US trajectories. Then we computed the B_E from the difference between the highest and lowest value of the average

PMF curve, with the associated error calculated by summing the errors at these highest and lowest points. We performed a total of 5.25 μs of US simulations for each B_E calculation, with 1.75 μs of simulation for each individual window.

(c) Protein dihedral angles

Two main chain dihedral angles, phi (ϕ),[49] and psi (ψ)[49]and one side chain dihedral angle chi1 ($\chi 1$),[49] primarily describes the conformation of the protein backbone. We compute these dihedral angles per residue per frame using our in-house program. These are used as the microscopic conformational variable, θ, in the conformational thermodynamics calculation of protein residue.

(d) Structural parameters of the DNA bases

DNA consists of one sugar-base[30] (χ), six sugar-phosphate (α, β, γ, δ, ϵ, and ζ), and five sugar-puckers[30] ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$) backbone torsion angles for each of the bases. Each of these torsion angles acts as a microscopic conformational variable (θ) in conformational thermodynamics calculation of DNA bases. Each backbone torsion angle is computed using GROMACS dihedral module.

(e) Conformational thermodynamics

The changes in the free energy and entropy due to the conformational changes are computed based on the normalized probability distribution of each microscopic conformational variable, θ. A detailed description of this histogram-based method for the calculation of conformational thermodynamics is reported earlier[24–30] (`https://sites.google.com/view/softmatter-snbncbs/code`). We checked the conformational stability and order of the $\alpha 2$D and $\alpha 2$B proteins separately in the presence of HG-DNA with respect to free state. The conformational free energy and entropy cost for an individual protein residue can be calculated by taking the sum of the contribution of free energy and entropy of each dihedral in that residue. For both the $\alpha 2$D and $\alpha 2$B proteins, initially, we randomly select 48,000 number of frames out of 50,000 frames from the equilibrated part of the trajectory, and divide them into N_w (=12) windows. Each window contains an equal number of frames. Initially, we compute the conformational thermodynamics data of protein residues for every window. Then, evaluate the average value of the conformational thermodynamics with error over the windows. The error for each mean is given by $\pm\ 0.5\sigma_m/\sqrt{N_w}$, where σ_m is the standard deviation of the mean value over the windows, and N_w

is the number of windows. The conformational free energy change associated with any microscopic conformational variable, θ for two different conformational states, A and B, is defined as

ΔG_{conf} = -k_BT ln$\frac{H^A_{max}(\theta)}{H^B_{max}(\theta)}$

Here, $H^X(\theta)$ is the normalized probability distribution of θ. "max" represents the maximum of the histogram. k_B is the Boltzmann constant. The change in conformational entropy for a specific conformational variable, θ is evaluated from the Gibbs entropy formula.

$S(\theta)$ = -k_B $\sum_i$ $H_i(\theta) \ln H_i(\theta)$

Here, the sum is considered over the histogram bins i. Therefore, the change in conformational entropy of any degree of freedom θ for two different conformational states A and B is given by

ΔS_{conf} = $S_A(\theta)$ - $S_B(\theta)$

We use our in-house program for this calculation. The positive (+ve) and negative (-ve) values of ΔG_{conf} represent destabilization and stabilization, respectively, whereas the positive (+ve) and negative (-ve) values of ΔS_{conf} indicate disorder and order, respectively. Here, terminal residues of proteins are not taken into account. We perform this conformational thermodynamics calculation to check the stability and order of A_7 and T_{26} bases in each of the DNA-protein systems with respect to the naked WC system. We have taken the data of the WC system from our earlier study.[30]

(f) Pearson correlation coefficient

We use the Pearson correlation coefficient,[50–52] r_{xy}, to measure the strength and direction of linear correlation of conformational thermodynamics data, x at the interface between proteins (α2D and α2B) and DNA with the binding energy value, y, of proteins. r_{xy} is given as follows

$$r_{xy} = \frac{(n\sum_i x_i y_i - \sum_i x_i \sum_i y_i)}{(\sqrt{(n\sum_i x_i^2 - (\sum_i x_i)^2)}\sqrt{(n\sum_i y_i^2 - (\sum_i y_i)^2)})}$$

3.3 Results & Discussions

We characterize the structural aspects using equilibrated conventional MD trajectories. Figs. 3.3 (a)-(f) show the equilibrium snapshot at 1μs of HGP, free-WTD, free-WTB, WCP, NMUTP, and AMUTP systems, respectively. The major structural difference in the HGP system is the shorter C1′-C1′ (d_{c1}) (Fig. 3.4

(a)) distance at the A_7-T_{26} bp than in the WCP system. In the HGP system, A_7 maintains its syn ($\langle\chi\rangle$ = 60.36 °) conformation and HG type H-bonds[30] ($\langle d_{N7-N3}\rangle$ = 0.30 nm) with T_{26} throughout the equilibrated trajectory. The time evolution of χ for A_7 base in the equilibrated trajectory of the HGP system is shown in Fig. 3.4 (b). Similarly, the time evolution of d_{N1-N3} and d_{N7-N3} for A_7-T_{26} bp in the HGP system during the equilibrated trajectory is shown in Fig. 3.4 (c) and (d), respectively.

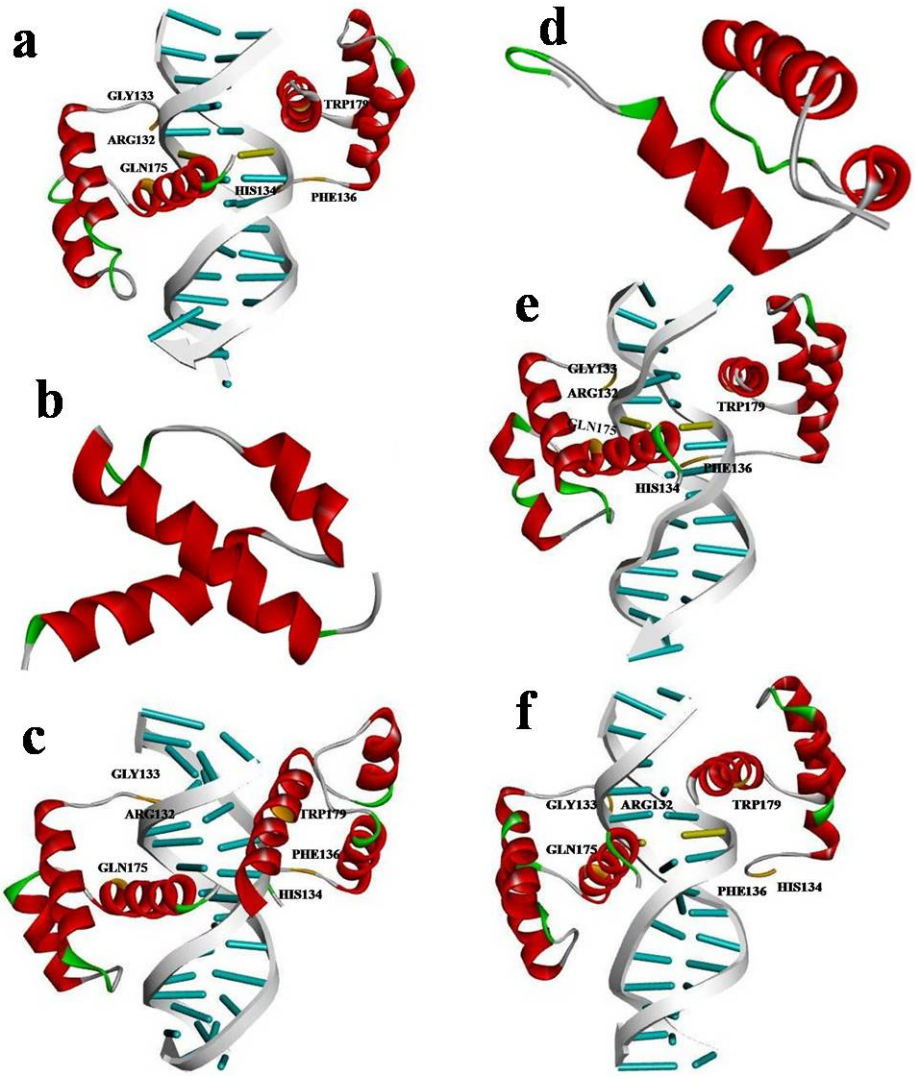


Figure 3.3: Snapshots at 1μs time : (a) HGP, (b)free-WTD, (c) free-WTB, (d) WCP, (e) NMUTP and (f) AMUTP systems. The HG bp in HGP, NMUTP and AMUTP systems are highlighted in yellow. The protein residues at the protein-DNA interface are highlighted in orange.

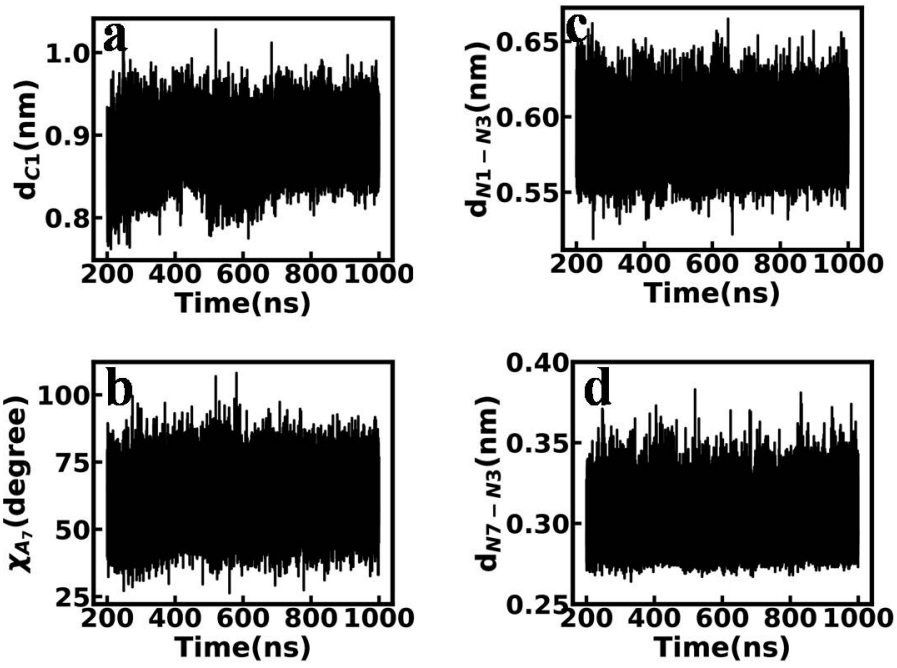


Figure 3.4: The time evolution of (a) d_{c1} for A_7-T_{26} bp, (b) glycosidic torsion angle χ for A_7, (c) d_{N1-N3} of A_7-T_{26} bp and (d) d_{N7-N3} of A_7-T_{26} bp in the equilibrated trajectory of the HGP system.

3.3.1 Protein-DNA interfaces

Earlier studies suggest the importance of interfaces in the binding energy.[13] Consequently, we focus on the interfacial interactions between proteins and DNA obtained from our conventional MD trajectories. Let us consider the HGP system first. The residues ARG132, GLY133, and GLN175 form H-bonds with base A_7 only. In addition to the H-bond, the ARG132 forms an electrostatic interaction as well as a salt-bridge with the O1P atom of A_7 (Table 3.3). The residue ARG132 forms H-bond, electrostatic interaction, and salt-bridge with the same atom O1P of A_7 maintains the criteria of H-bond, electrostatic interaction, and salt-bridge, so some frames are mutually inclusive in computing the fraction of frames form H-bond, electrostatic interaction and salt-bridge in Table 3.3. On the other hand, the specific protein α2B residues HIS134, PHE136, and TRP179 create H-bonds with T_{26} only (Table 3.3). The interfacial interactions are similar to those in our earlier study.[30] Additionally, we find that there are H-bonds between the atoms of residues ARG185 and ALA189 of α2D with the residues THR189 and ALA190 of α2B (Table 3.4).

Conventional MD trajectory of the WCP system shows that the residues

System	Nucleotide	Protein	H-bond	Electrostatic	Salt-bridge
HGP	A_7	α2D	ARG132 (0.65 ± 0.07)	ARG132 (0.93± 0.01)	ARG132 (0.26± 0.04)
			GLY133 (0.92± 0.01)		
			GLN175 (0.93± 0.01)		
WCP			ARG132 (0.11± 0.01)	ARG132 (0.05± 0.01)	ARG132 (0.05± 0.01)
			GLY133 (0.85± 0.01)		
			GLN175 (0.66± 0.01)		
NMUTP			ASN132 (0.25± 0.01)	ASN132 (0.96± 0.004)	ASN132 (0.22± 0.01)
			GLY133 (0.52± 0.02)		
			GLN175 (0.89± 0.01)		
AMUTP			ASP132 (0.46± 0.01)	ASP132 (0.97± 0.001)	ASP132 (0.23± 0.01)
			GLY133 (0.89± 0.02)		
			GLN175 (0.75± 0.01)		
HGP	T_{26}	α2B	HIS134 (0.61± 0.01)		
			PHE136 (0.99± 0.01)		
			TRP179 (0.59± 0.04)		
WCP			HIS134 (0.42± 0.02)		
			PHE136 (0.95± 0.01)		
			TRP179 (0.55± 0.01)		
NMUTP			HIS134 (0.59± 0.03)		
			PHE136 (0.97± 0.003)		
			TRP179 (0.58± 0.04)		
AMUTP			HIS134 (0.76± 0.01)		
			PHE136 (0.94± 0.03)		
			TRP179 (0.69± 0.04)		

Table 3.3: The ratio of the number of frames in each non-bonded interaction to the total number of frames (average with error) for each interface forming protein residues of α2D and α2B, considering the interfacial contact between A_7 and α2D and between T_{26} and α2B, respectively.

System	Residue (α2D)	Residue (α2B)	Number of frames (H-bond) /Number of frames (Total)
HGP	ARG185	THR189	0.07
	ALA189	THR189	0.004
	ALA189	ALA190	0.17

Table 3.4: The ratio of the number of frames containing H-bonds between α2D and α2B proteins to the total number of frames in the equilibrated trajectory of the HGP system.

ARG132, GLY133 and GLN175 of α2D form H-bonds with the base A_7 of DNA, the residue ARG132 of the WCP system form electrostatic interaction as well as salt-bridge with A_7 in addition to the H-bond (Table 3.3), just like the HGP system. Thus, some frames are also mutually inclusive here when calculating the fraction of frames that form H-bond, electrostatic interaction and salt-bridge between ARG132 of α2D and base A_7 of DNA. However, the number of H-bond, electrostatic interaction, and salt-bridge all the interface interactions between the α2D residues and A_7 base decrease in the WCP system as compared to the HGP system (Table 3.3). For α2B, like the HGP system, the residues HIS134, PHE136 and TRP179 of the WCP system form interfaces with the base T_{26} (Table 3.3). In contrast to the HGP system, the WCP system contains significantly less number

System	Residue (α2D)	Residue (α2B)	Number of frames (H-bond) /Number of frames (Total)
WCP	ALA189	THR189	0.04
	ALA189	ALA190	0.004

Table 3.5: The ratio of the number of frames containing H-bonds between α2D and α2B proteins to the total number of frames in the equilibrated trajectory of the WCP system.

of H-bonds for the interface forming residues HIS134 and PHE136 of the α2B protein as indicated by the error bars (Table 3.3). Table 3.5 further shows the inter-protein H-bonds decrease in the WCP system compared to the HGP system.

The NMUTP system shows the intermolecular interactions (H-bond and salt-bridge) between α2D and A_7 decrease (Table 3.3), while the intermolecular contacts between α2B and T_{26} almost remains unchanged (Table 3.3) compared to the HGP system. The intermolecular contacts (H-bond and salt-bridge) between the atoms of the α2D and A_7 (Table 3.3) decreases while the inter-molecular contacts (H-bond) between the atoms of the α2B and T_{26} (Table 3.3) increases in the AMUTP system compared to the HGP system. Additionally, while taking into account interface interactions between ASN132 and ASP132 residues with A_7 base, there are some frames for the NMUTP and AMUTP systems that are mutually inclusive also.

3.3.2 Binding energy

(a) Wild type protein

We compute the B_E for different systems. First, we consider the B_E calculation for the α2D protein with remaining portion of the HGP system. We run three independent US simulations to compute the average PMF curve. The PMF curves for each of the independent US simulations are shown in Fig. 3.5. The average PMF curve and associated errors are computed using the results of the three independent US datasets as shown in Fig. 3.6 (a). The inset figure in Fig. 3.6 (a) shows the magnified values of $V_{PMF}(\zeta_D)$ for low ζ_D. The $V_{PMF}(\zeta_D)$ data show a minimum of around 1.7 ζ_D value and saturates after approximately 5 ζ_D value. To calculate the B_E, we take the difference between the highest and lowest values of the average PMF curve. We compute the B_E to be approximately -36.57 $\pm$ 0.78 kcal/mol in case of Fig. 3.6 (a). The error $\pm$ 0.78 associated with the B_E is computed by summing the errors of the highest and lowest value of the average PMF curve. The similar protocol is followed to compute the B_E for the other cases.

Next, we compute the B_E of the α2B protein with the HG-DNA with α2D bound to HG-DNA. ζ_B, the center of mass distance between the α2B and the

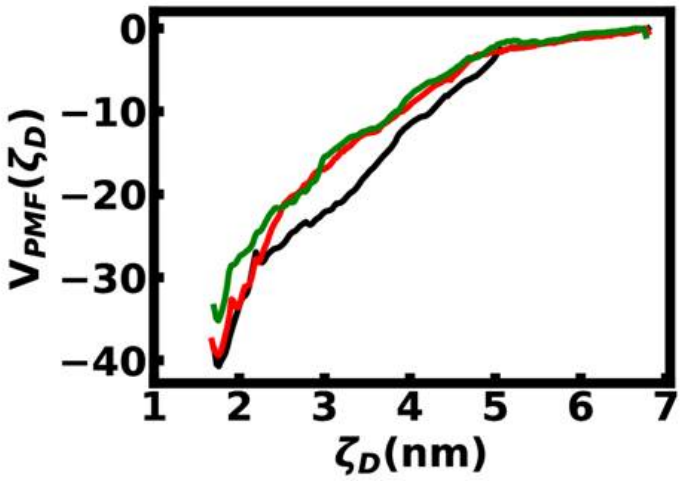


Figure 3.5: Three PMF curves computed from the three independent US simulations for the α2D protein with the remaining part of the HGP system using same force constant 1000 kJ/mol nm^2 and same pulling rate 0.01 nm/ps. Black for US-1, red for US-2 and green for US-3.

remaining part, is the reaction coordinate. The average PMF curve with error bars of α2B is shown in Fig. 3.6 (b). $V_{PMF}(\zeta_B)$ at low ζ_B values is shown in the inset figure of Fig. 3.6 (b). The B_E of α2B protein is around -37.00 $\pm$ 0.56 kcal/mol. Therefore, in the HGP system, the B_E of the non-specific protein, α2D, and the specific protein, α2B, with HG-DNA in the presence of the other protein are quite close.

In our earlier study,[30] we observed that the HGP system is stabilized and ordered than the naked WC system. Here, we check the B_E of both the α2D and α2B proteins in the WCP system where the HG bp is replaced by a WC bp . Fig. 3.6 (c) shows the average PMF plot with error bars of α2D protein with the remaining part of the WCP system. ζ'_D is the reaction coordinate being the distance between the center of mass of the α2D and the rest of the system. The inset is the magnified $V_{PMF}(\zeta'_D)$ at low ζ'_D. The B_E of the α2D protein is approximately -26.53 $\pm$ 0.86 kcal/mol in Fig. 3.6 (c), which is much lower than the B_E of wild-type α2D protein in the HGP system. Similarly, we compute the B_E of the α2B protein using ζ'_B is the reaction coordinate, which is the center of mass distance between α2B and the remaining part of the system. Fig. 3.6 (d) is the average PMF plot with error bars of α2B protein with remaining part of the WCP system. The magnified $V_{PMF}(\zeta'_B)$ for low ζ'_B is shown in the inset. The B_E of the α2B in Fig. 3.6 (d) is -33.23 $\pm$ 0.50 kcal/mol, which is also lower

than the B_E of the wild-type α2B protein in the HGP system. Therefore, both the proteins show stronger binding in the HGP system than the WCP system. The interfacial interactions of both the proteins with DNA also decrease in the WCP system compared to the HGP system.

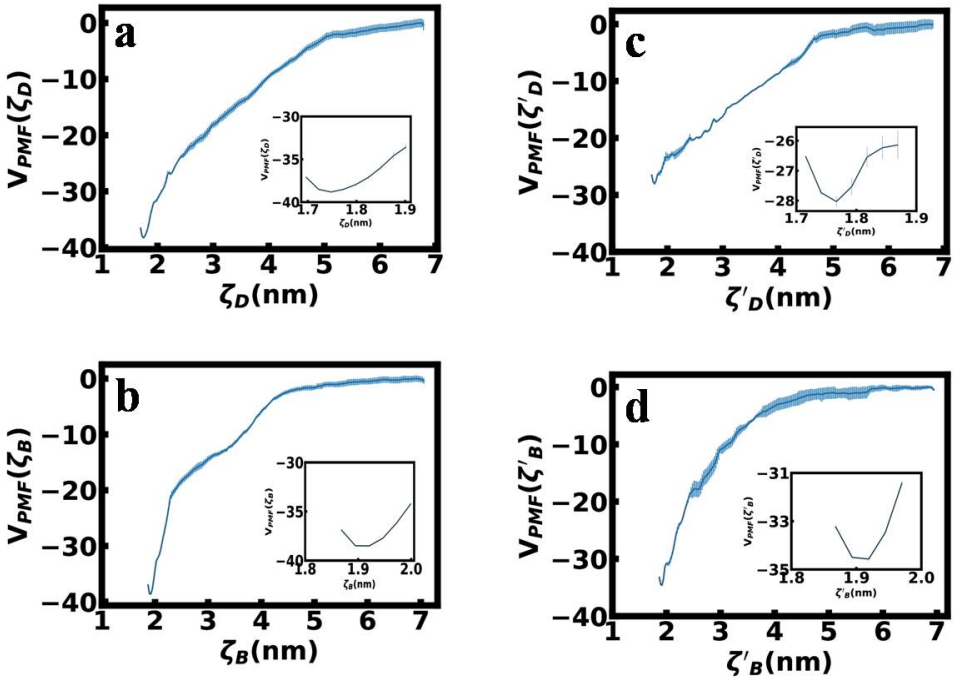


Figure 3.6: Average PMF curves with error bars for (a) α2D and (b) α2B proteins, respectively with the HG-DNA duplex in the presence of other protein obtained from the US method for the HGP system. The inset Figs. in (a) and (b) represent the magnifying values of $V_{PMF}(\zeta_D)$ and $V_{PMF}(\zeta_B)$ at low ζ_D and ζ_B values, respectively. Similarly, (c) average PMF curves with error bars for α2D and (d) for α2B proteins, respectively, with the remaining part of the WCP system. The magnified values of $V_{PMF}(\zeta'_D)$ and $V_{PMF}(\zeta'_B)$ at low ζ'_D and ζ'_B values are shown in inset Figs. (c) and (d), respectively.

Next we consider a situation where HG-DNA is bound to one of the proteins, either α2D or α2B in the absence of the other protein. Let us consider the case of α2D. In Fig. 3.7 (a), average $V_{PMF}(\zeta''_D)$ is plotted as a function of the reaction coordinate, ζ''_D being the distance between the center of mass of α2D and HG-DNA. The inset shows magnified $V_{PMF}(\zeta''_D)$ for low ζ''_D. The B_E of α2D with HG-DNA in Fig. 3.7 (a) is nearly -33.48 $\pm$ 1.23 kcal/mol less than that in the presence of α2B. Then consider the case of α2B. ζ''_B, the distance between the center of mass of α2B and DNA duplex is the reaction coordinate. In Fig. 3.7 (b), we show average $V_{PMF}(\zeta''_B)$ with ζ''_B. The inset shows the magnified values of $V_{PMF}(\zeta''_B)$ at low ζ''_B values. The B_E of α2B with HG-DNA in Fig. 3.7 (b) is -34.77

$\pm$ 0.83 kcal/mol which is lower than that of Fig. 3.6 (b). Thus, in the absence of other protein, the B_E of the α2D and α2B proteins with HG-DNA decrease. This result suggests that two proteins cooperate in their binding. Table 3.4 shows that there are some H-bonds between two proteins in the conventional MD trajectory of the HGP system. These H-bonds are likely to impact the cooperativity between these two proteins while binding to HG-DNA.

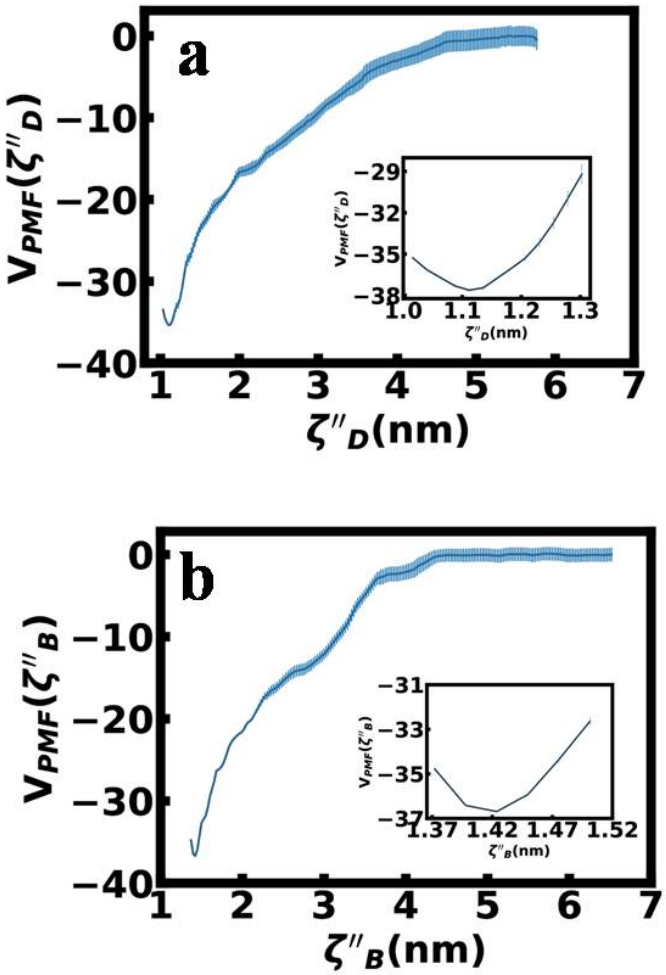


Figure 3.7: (a) Average PMF curve with error bars for α2D protein with HG-DNA duplex in the absence of α2B protein. Similarly, (b) average PMF curve with error bars for α2B protein with HG-DNA duplex in the absence of α2D protein, respectively obtained from the US method. The magnified values of $V_{PMF}(\zeta''_D)$ and $V_{PMF}(\zeta''_B)$ at low ζ''_D and ζ''_B are inserted in Figs. (a) and (b), respectively.

(b) Mutated protein

The interfacial interaction calculation shows that ARG132 is the N-terminal arm of the non-specific protein, α2D, forms the maximum number of interfacial contacts with the HG base A_7.[30] We mutate this basic residue ARG132 in α2D in the HGP system by neutral ASN132 (NMUTP system). Here we also check the B_E of the α2D and α2B proteins separately with the remaining part of the system. First we compute the B_E of the α2D protein. $\tilde{\zeta_D}$, the distance between the center of mass of α2D and the rest of the system is the reaction coordinate. The average $V_{PMF}(\tilde{\zeta_D})$ curve of the α2D protein for the NMUTP system is shown in Figs. **??** (a) with the magnified $V_{PMF}(\tilde{\zeta_D})$ at low $\tilde{\zeta_D}$ value in the inset. The B_E of α2D protein in Fig. 3.8 (a) is approximately -30.29 $\pm$ 0.53 kcal/mol while in Fig. 3.6 (a), the B_E of the wild type α2D protein in the HGP system is -36.57 $\pm$ 0.78 kcal/mol. Similarly, average $V_{PMF}(\tilde{\zeta_B})$ curve of the α2B protein for the NMUTP system is shown in Fig. 3.8 (b). The inset is the magnified $V_{PMF}(\tilde{\zeta_B})$ at low $\tilde{\zeta_B}$ value. $\tilde{\zeta_B}$ is the reaction coordinate here, which is the center of mass distance between α2B and the remaining part of the NMUTP system. The B_E of α2B protein in Fig. 3.8 (b) is approximately -36.25 $\pm$ 0.50 kcal/mol, while in Fig. 3.6 (b), the B_E of the wild-type α2B protein in the HGP system is -37.00 $\pm$ 0.56 kcal/mol. So, in the NMUTP system, the B_E of the mutated α2D protein decreases ($\sim$ 6 kcal/mol), while the B_E of α2B protein almost remains the same relative to the HGP system following the trends of the interfacial interactions.

We further mutate the basic residue ARG132 of HGP system by acidic residue ASP132 (AMUTP system). Then we compute the average $V_{PMF}(\tilde{\tilde{\zeta_D}})$ plot of α2D protein in the AMUTP system in Fig. 3.8 (c) with magnified view at low $\tilde{\tilde{\zeta_D}}$ value in the inset. $\tilde{\tilde{\zeta_D}}$, the distance between the α2D protein and the rest of the AMUTP system is the reaction coordinate. The B_E value is nearly -29.60 $\pm$ 0.77 kcal/mol. Similarly, we compute the B_E of the α2B protein using the reaction coordinate $\tilde{\tilde{\zeta_B}}$ which represents the center of mass distance between α2B and the remaining part of the AMUTP system. Fig. 3.8 (d) shows average $V_{PMF}(\tilde{\tilde{\zeta_B}})$ curve of α2B protein in the AMUTP system. The magnified view at low $\tilde{\tilde{\zeta_B}}$ value is in the inset. The B_E of the α2B is nearly -51.42 $\pm$ 0.60 kcal/mol. Therefore, the acidic mutation of ARG132 by ASP132 in the α2D protein decreases the B_E of the α2D protein nearly 7 kcal/mol as compared to the wild-type α2D protein while increasing the B_E of the α2B protein nearly 13 kcal/mol as compared to the wild-type α2B protein of the HGP system. Conventional MD simulation trajectory of the AMUTP system reveals that the decrease in B_E of α2D is coupled

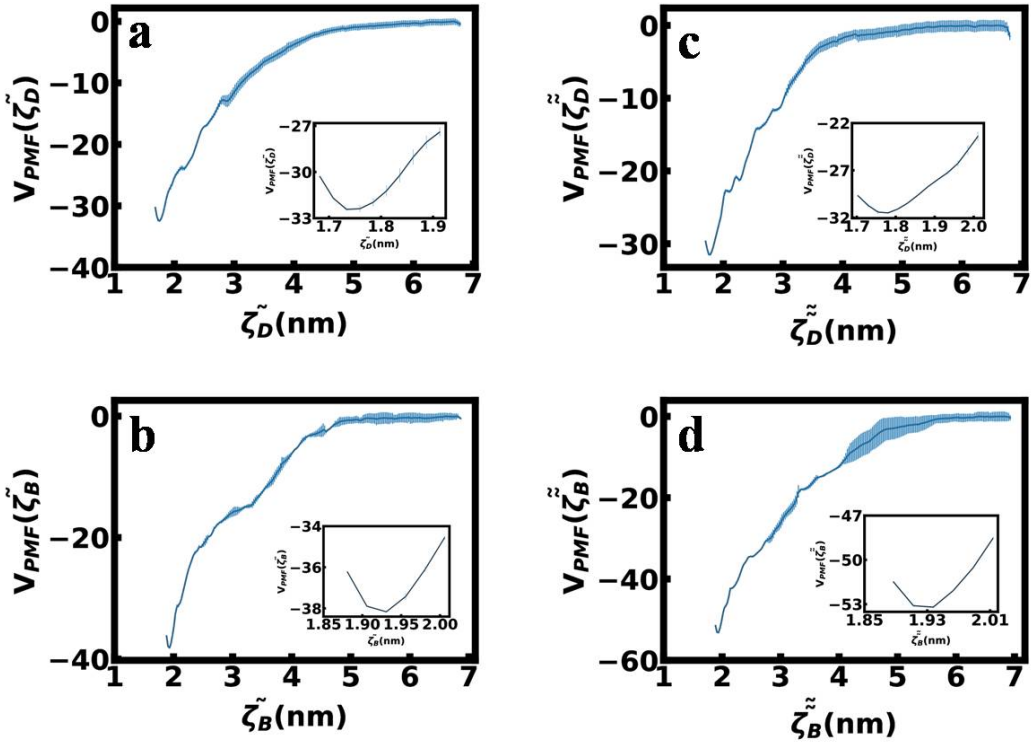


Figure 3.8: Average PMF curves with error bars for (a) α2D and (b) α2B proteins, respectively with the remaining part of the NMUTP system. Magnified $V_{PMF}(\tilde{\zeta_D})$ and $V_{PMF}(\tilde{\zeta_B})$ at low $\tilde{\zeta_D}$ and $\tilde{\zeta_B}$, respectively, are shown in the inset figures. (c) Average PMF curves with error bars for α2D and (d) α2B proteins, respectively, with the remaining part of the AMUTP system. The inset figures show magnified $V_{PMF}(\tilde{\tilde{\zeta_D}})$ and $V_{PMF}(\tilde{\tilde{\zeta_B}})$ at low $\tilde{\tilde{\zeta_D}}$ and $\tilde{\tilde{\zeta_B}}$, respectively.

to a decrease in the intermolecular contacts between the α2D and A_7 (Table 3.3). The increase in the B_E of α2B is due to the increase in the intermolecular contacts between the α2B and T_{26} (Table 3.3), respectively.

3.3.3 Conformational stability and order at the interface

We further quantify the relation between the binding energy and interfacial stability and order. In order to understand the stability and order at the interface, we compute conformational free energy and entropy from the fluctuations of the microscopic conformation variables in terms of the distribution, $H_P^S(\theta_R)$ of a microscopic conformational variable, θ of residue, R of protein, P in a given system, S over the equilibrated trajectory. The peak of the histogram defines the equilibrium value of the relevant variable. We illustrate a few cases of microscopic conformational variables for the interface forming residues of α2D

and α2B proteins with HG-DNA that exhibit major changes in their histograms between HGP and free systems. Figs. 3.9 (a) and (b) show the distributions of ϕ and ψ dihedral angles at G133 of α2D. $H^{free}_{\alpha 2D}$ (ϕ_{G133}) and $H^{free}_{\alpha 2D}$ (ψ_{G133}) are multi-peaked and double- peaked, respectively, while $H^{HGP}_{\alpha 2D}$ (ϕ_{G133}) and $H^{HGP}_{\alpha 2D}$ (ψ_{G133}) are single -peaked, showing the decrease in flexibility of the α2D in the HGP system compared to the free state for both the ϕ_{G133} and ψ_{G133}. Similarly, Figs. 3.9 (c) and (d) show the distributions of ϕ and $\chi 1$ dihedral angles at F136 of α2B. $H^{free}_{\alpha 2B}$ (ϕ_{F136}) shows a broad single peak while $H^{HGP}_{\alpha 2B}$ (ϕ_{F136}) shows a sharp single peak in Fig. 3.9 (c). In Fig. 3.9 (d), both $H^{HGP}_{\alpha 2B}$ ($\chi 1_{F136}$) and $H^{free}_{\alpha 2B}$ ($\chi 1_{F136}$) are double-peaked, though in one peak the height of the peak in the free state is slightly higher, and in another peak the height of the peak in the DNA bound state is highest. Therefore, Figs. 3.9 (c) and (d) suggest that the $\alpha 2B$ protein is more flexible in the free form compared to the HGP system. Overall, the presence of the HG bp reduces the flexibility of both the homeodomain proteins, α2D and α2B in the HGP system. We perform similar calculations for proteins in case of other systems and for interface forming bases of DNA for all the DNA-protein systems with respect to the free WC system.

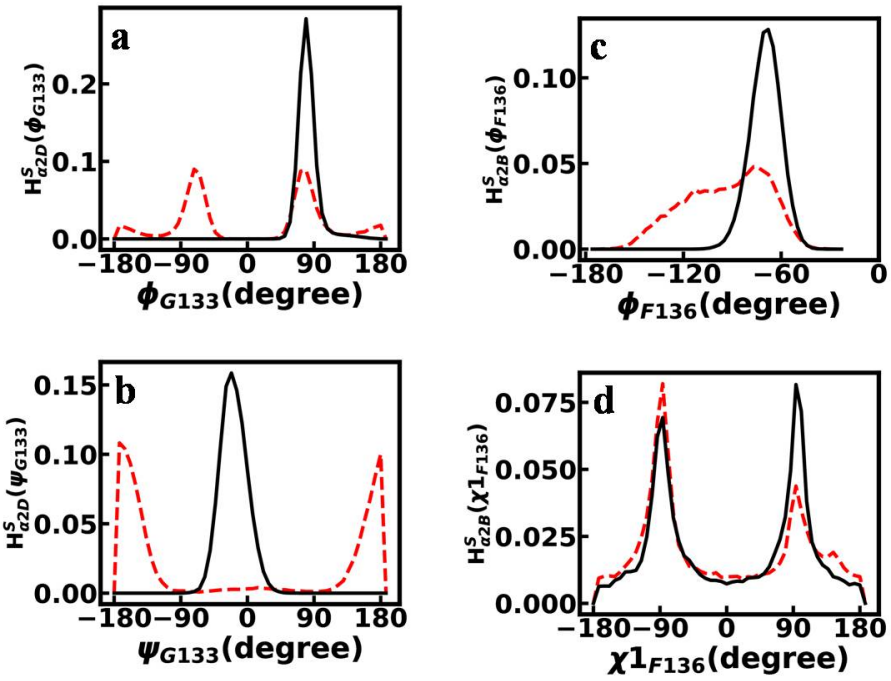


Figure 3.9: Histograms of dihedral angle (a) ϕ and (b) ψ at G133 of α2D in the free (red, dashed line) and DNA bound HGP (black, solid line) systems. Similarly, histograms of dihedral angle (c) ϕ and (d) $\chi 1$ at the F136 of α2B in the free (red, dashed line) and DNA bound HGP (black, solid line) systems.

The changes in the distribution of dihedral angles lead to the changes in the conformational free energy and entropy data of α2D and α2B proteins in the DNA bound state with respect to their free state. The total changes of conformational free energy and entropy of individual protein residue are calculated by adding the changes due to each microscopic degree of freedom of that residue. The residue-wise conformational thermodynamics data of α2D in the HGP system with respect to the free state are shown in Figs. 3.10 (a) and (b). Similarly, conformational thermodynamics data of the α2B in the HGP system with respect to the free state are shown in Figs. 3.11 (a) and (b). In Figs. 3.10 and 3.11, the conformational thermodynamics data of interface forming residues with DNA are marked by blue star symbol. All the interfaces forming residues except GLN175 of α2D protein are stabilized and ordered in the DNA bound state compared to the free state (Figs. 3.10 and 3.11). We perform conformational thermodynamics calculations in the case of proteins for all other systems. Similarly, we compute conformational thermodynamics data for interface forming bases of each of the DNA-protein systems with respect to the free WC system.

Next, we compute the conformational thermodynamics data at the interface by adding the data of the protein residues as in Ref. 21 and DNA bases as in Ref. 23 at the interfaces. Fig. 3.12(a) shows the linear correlation between conformational free energy, ΔG_{conf}, at the interface and binding energy, B_E, of proteins for different DNA-protein systems in Fig. 3.12(a). The Pearson correlation coefficient r_{xy} is 0.72, suggesting strong correlation between stability of the interface and B_E. The positive correlation between conformational entropy, $T\Delta S_{conf}$ at the interface and B_E of proteins is shown in Fig. 3.12(b). Here, the correlation coefficient is 0.66, indicating moderate positive correlation between $T\Delta S_{conf}$ and B_E. Therefore, Figs. 3.12(a) and (b) show the B_E of proteins increases with an increase in the conformational stability and order of protein residues and DNA bases at the interface.

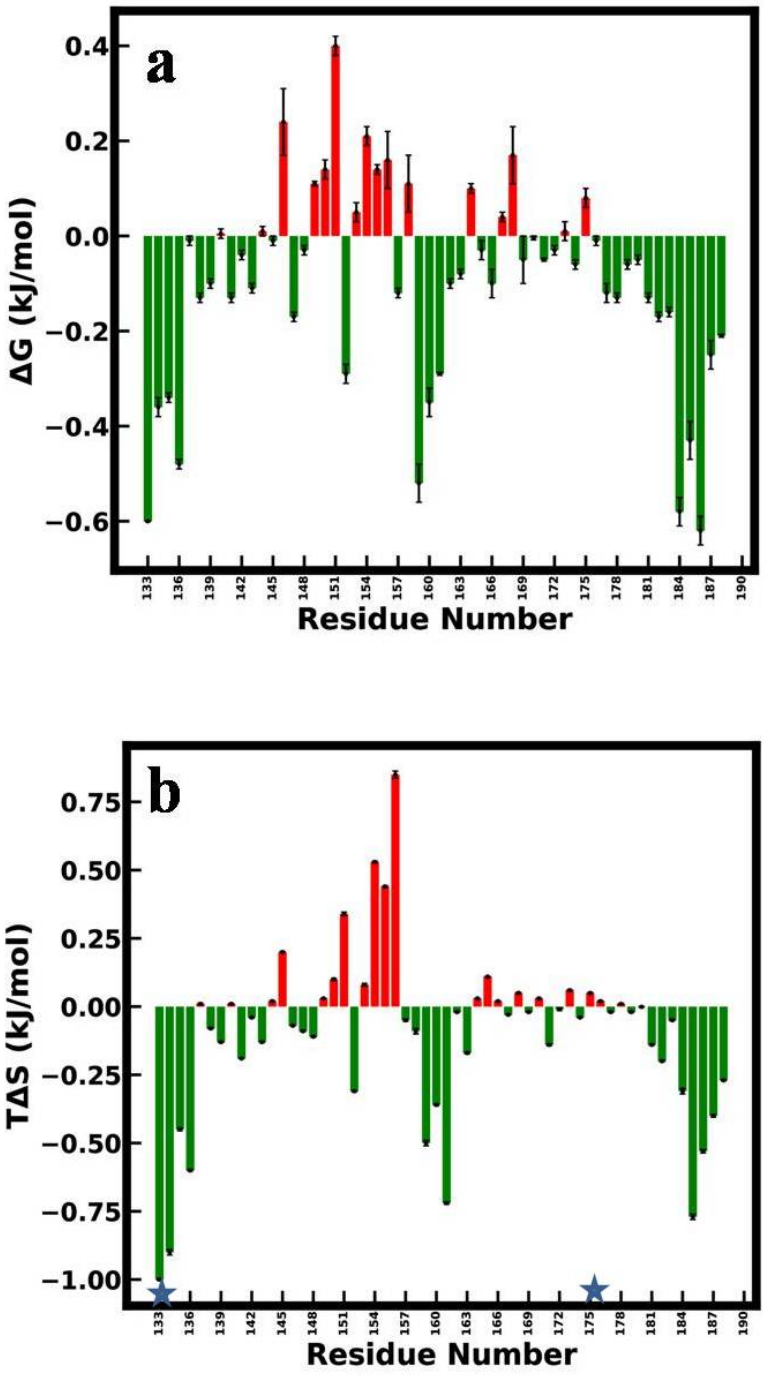


Figure 3.10: Residue-wise (a) ΔG (b) TΔS values of the α2D protein in the HGP system with respect to the free form of α2D. All the values are in kJ/mol. The ΔG and TΔS values of interfaces forming residues with DNA indicated by blue star symbol. Stabilized or ordered residues are represented by green. Destabilized or disordered residues are represented by red.

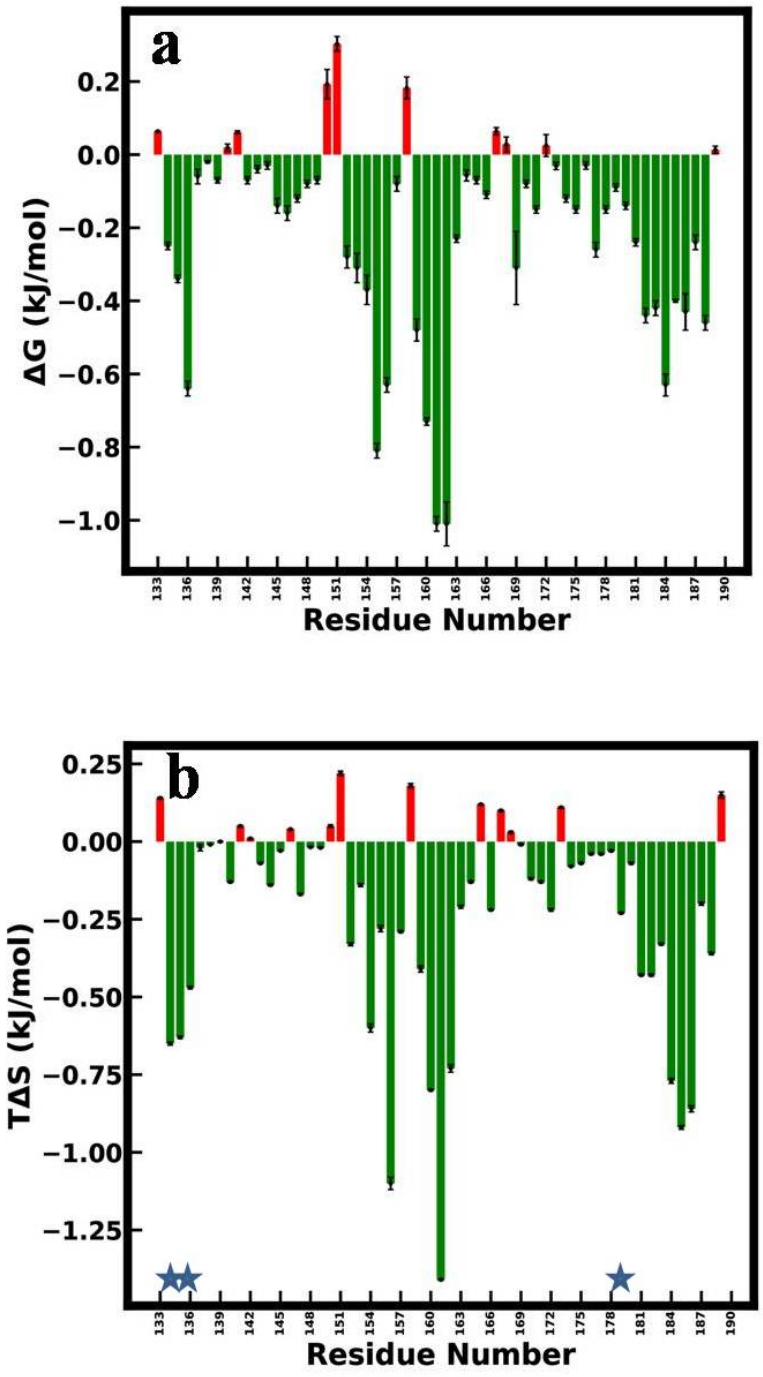


Figure 3.11: Residue-wise (a) ΔG (b) TΔS values of the α2B protein in the HGP system with respect to the free form of α2B. All the values are in kJ/mol. The ΔG and TΔS values of interfaces forming residues with DNA indicated by blue star symbol. Stabilized or ordered residues are represented by green. Destabilized or disordered residues are represented by red.

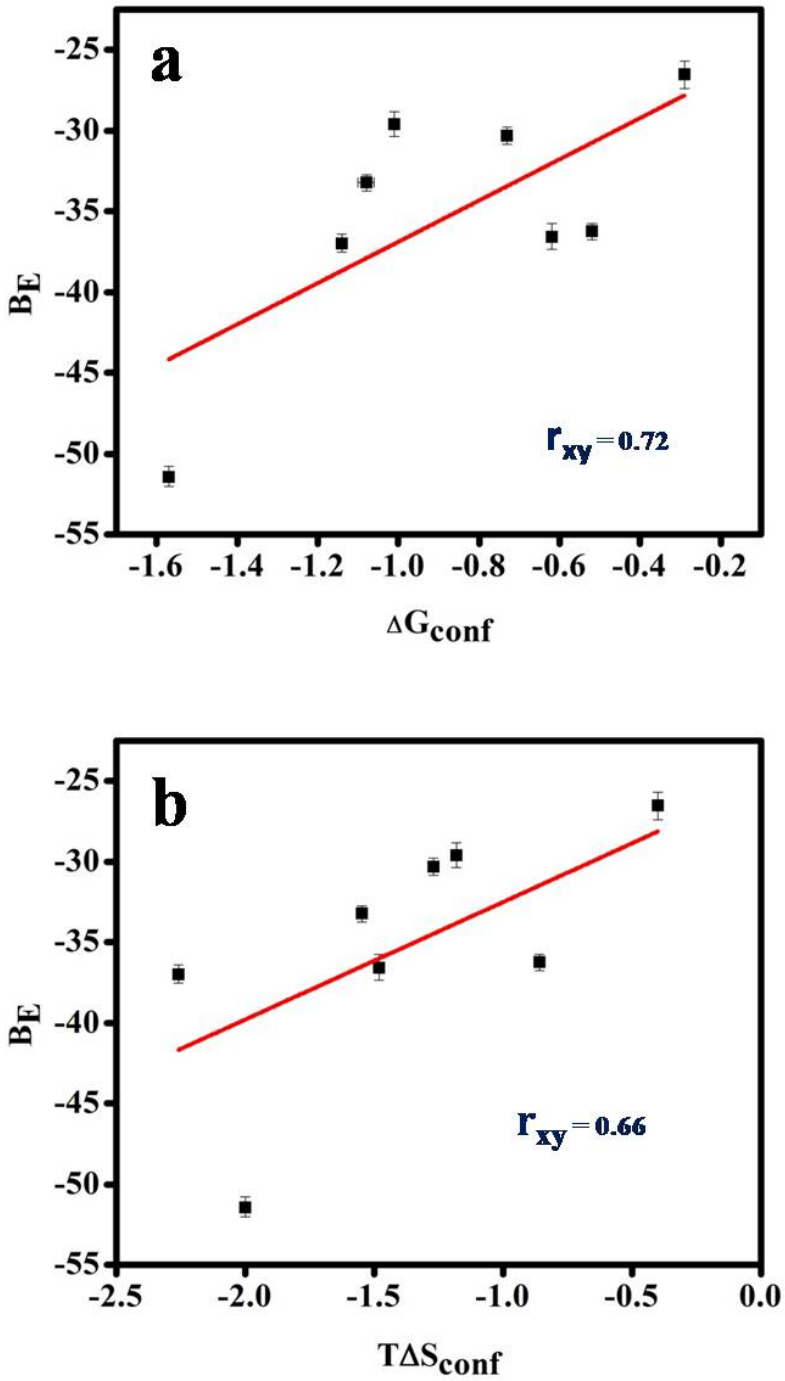


Figure 3.12: (a) Correlation plot between conformational free energy, ΔG_{conf}, of interface forming residues and binding energy B_E with error bars of proteins, $\alpha 2D$ and $\alpha 2B$. (b) Correlation plot between conformational entropy, $T\Delta S_{conf}$, with error of interface forming residues and binding energy B_E with error bars of $\alpha 2D$ and $\alpha 2B$ proteins. All data in kcal/mol. Black square symbols represent the data points and the red dotted line represents linear fitting of the data points.

System	B_E (kcal/mol)	System	B_E (kcal/mol)
α2D with rest of the HGP	-36.57 $\pm$ 0.78	α2B with rest of the HGP	-37.00 $\pm$ 0.56
α2D with rest of the WCP	-26.53 $\pm$ 0.86	α2B with rest of the WCP	-33.23 $\pm$ 0.50
α2D with HG-DNA	-33.48 $\pm$ 1.23	α2B with HG-DNA	-34.77 $\pm$ 0.83
α2D with rest of the NMUTP	-30.29 $\pm$ 0.53	α2B with rest of the NMUTP	-36.25 $\pm$ 0.50
α2D with rest of the AMUTP	-29.60 $\pm$ 0.77	α2B with rest of the AMUTP	-51.42 $\pm$ 0.60

Table 3.6: Summarize and compare the binding energy of different computed systems with the wild-type system HGP.

The strong correlation between conformational thermodynamics data at the interface and the binding energy between the homeodomain protein and DNA in Fig. 3.12, suggests that changes in conformational thermodynamics at the interface play a crucial role in protein-DNA binding. Experimentally, binding entropy and binding free energy can be measured using isothermal titration calorimetry (ITC),[24,53] while NMR relaxation experiments[26,54] can assess the changes in the conformational entropy in individual binding regions. Conformational states of small biomolecules can also be explored via UV resonance Raman measurements,[55] where free energy differences between states are derived from population ratios. Additionally, B_E can be determined through spectroscopic techniques such as fluorescence[56] and NMR.[57] Thus, the correlation between conformational thermodynamics and B_E can be experimentally verified. Such a plot may be useful as a calibration of protein binding to DNA.

In Table 3.6, we summarize and compare the B_E for different systems with respect to the HGP system. The B_E of both α2D and α2B proteins decreases in the WCP system when the HG bp is replaced with the WC bp. A similar decrease in B_E is observed in the absence of the other protein, indicating the cooperative binding of α2D and α2B in the HGP system. In the NMUTP system, the neutral mutation in the α2D protein reduces the B_E of both α2D and α2B. However, the B_E of α2D decreases and the B_E of α2B increases due to the acidic mutation of the α2D protein in the AMUTP system. Such trends can be understood from the change in interfacial interactions revealed from our conformational thermodynamics data.

3.4 Conclusion

In summary, the B_E of the non-specific protein, α2D, and the specific protein, α2B, are almost equal to HG-DNA with the other protein remaining bound. In the WCP system, the B_E and interfacial interactions of both the non-specific

and specific proteins, α2D and α2B, decrease significantly compared to the HGP system. The B_E of both the protein decrease in the absence of the other. Thus, α2D and α2B proteins mutually cooperate in their binding. A neutral mutation in the basic N-terminal arm of the non-specific protein alters significantly the interfacial contacts as well as the B_E of the α2D only in the NMUTP system while an acidic mutation in the AMUTP system changes the interfacial interactions as well as B_E for both the α2D and α2B proteins significantly. The significant differences in the B_E of proteins in different systems could be explained by comparing the number of intermolecular contacts and correlate well to conformational thermodynamics data at the interfaces of proteins and DNA. This chapter, the detailed thermodynamic information at the interfaces with the B_E of proteins would be helpful to design this type of HG-DNA-homeodomain-protein system, which may have major implications in DNA replication, recognition, damage repair, and other cellular processes.

3.5 Appendix

3.5.1 Steered molecular dynamics & umbrella sampling

Conventional molecular dynamics (MD) fails to efficiently sample the phase space of complex biological or chemical systems in the presence of high-energy barriers. To address this sampling challenge, biased molecular dynamics or enhanced sampling techniques can be employed. Steered Molecular Dynamics (SMD) and Umbrella Sampling (US)[15] are widely used methods in complex biological and chemical systems to investigate the kinetics of processes such as protein-DNA, protein-ligand binding etc. This method requires an appropriate reaction coordinate (ζ) that effectively represents the progress of the phenomenon.

Here, in DNA-protein systems the distance between the center of mass of the protein and remaining part of the system is the reaction coordinate. First, the SMD technique is used to pull the protein away from the rest of the system by applying a constant force. The position of the protein is maintained using a bias potential. Thus, SMD generates a series of configurations along the reaction coordinate. These configurations at specific intervals are used as initial conformations for US windows. Here, the bias potential applied on the system is V_i (ζ) which is only function of ζ. Now, the biased energy of the system for window i is

$$U_i^b(r) = U_i(r) + V_i(\xi) \tag{3.1}$$

One can get the biased probability distribution along the reaction coordinate, $P_i^b(\xi)$, as follows

$$P_i^b(\xi) = \frac{\int \delta(\xi - \xi_i) \exp[-\beta U_i(r) + V_i(\xi)] dr}{\int \exp[-\beta U_i(r) + V_i(\xi)] dr} \tag{3.2}$$

Since, bias potential depends only on ζ, therefore,

$$P_i^b(\xi) = \exp(-\beta V_i(\xi)) \frac{\int \delta(\xi - \xi_i) \exp[-\beta U_i(r)] dr}{\int \exp[-\beta U_i(r) + V_i(\xi)] dr} \tag{3.3}$$

One can obtain the unbiased probability distributions $(P_i^u(\xi))$,

$$P_i^u(\xi) = \frac{\int \delta(\xi - \xi_i) \exp[-\beta U_i(r)] dr}{\int \exp[-\beta U_i(r)] dr} \tag{3.4}$$

Now, from equation 1.3 and 1.4 one can write,

$$P_i^u(\xi) = P_i^b(\xi) \exp[\beta V_i(\xi)] \langle \exp[-\beta V_i(\xi)] \rangle \tag{3.5}$$

So, one can compute the free energy for window i as follows,

$$G_i(\xi) = -\frac{1}{\beta} \ln(P_i^b(\xi)) - V_i(\xi) + F_i \tag{3.6}$$

where,

$$F_i = -\frac{1}{\beta} \ln \langle \exp[-\beta V_i(\xi)] \rangle \tag{3.7}$$

depends on the applied bias potential, which needs to be determined.

To obtain the global potential mean force (PMF) curve, one must combine the free energy from each window. Therefore, the calculation of F_i is necessary to construct the global PMF curve. The Weighted Histogram Analysis Method (WHAM) is a widely used technique for computing the global PMF curve from individual US windows.

3.5.2 Weighted Histogram Analysis Method (WHAM)

WHAM is most popular approach to obtain F_i. It aims to minimize the statistical error to compute the global distribution, $P^u(\xi)$.

The global distribution, $P^u(\xi)$ is computed from a weighted average of the distributions of the individual windows:

$$P^u(\xi) = \sum_{i=1}^{N} p_i(\xi) P_i^u(\xi) \tag{3.8}$$

The weights p_i are chosen in order to minimize the statistical error of P^u.

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0 \tag{3.9}$$

under the condition $\sum p_i = 1$. This leads to

$$p_i = \frac{a_i}{\sum_j a_j}, \quad a_i(\xi) = N_i \exp[-\beta\omega_i(\xi) + \beta F_i] \tag{3.10}$$

with N_i being the total number of steps sampled for window i. The F_i are calculated by Eq. (1.7):

$$F_i = -\frac{1}{\beta} \ln\langle e^{-\beta V_i(\xi)} \rangle \tag{3.11}$$

$$-\beta F_i = \ln\langle e^{-\beta V_i(\xi)} \rangle \tag{3.12}$$

$$e^{-\beta F_i} = \int P^u(\xi) \exp[-\beta\omega_i(\xi)] d\xi. \tag{3.13}$$

Because P^u enters Eq. (3.13) and F_i enters Eq. (3.8) via Eq. (3.10), these have to be iterated until convergence.

Bibliography

[1] Sobell, H. M. Actinomycin and DNA transcription. PNAS **1985**, *82*, 5328-5331.

[2] Clayton, D. A. Transcription and replication of mitochondrial DNA. Hum. Reprod. **2000**, *15(suppl_2)*, 11-17.

[3] Leonard, A. C. Méchali, M. DNA replication origins. Cold Spring Harb. Perspect. Biol. **2013**, *5*, a010116.

[4] Kornberg, A. Baker, T. A. DNA replication. University Science Books **2005**.

[5] Lindahl, T. Wood, R. D. Quality control by DNA repair. Science **1999**, *286*, 1897-1905.

[6] Wood, R. D. Mitchell, M. Sgouros, J. Lindahl, T. Human DNA repair genes. Science **2001**, *291*, 1284-1289.

[7] Jen-Jacobson, L. Engler, L. E. Ames, J. T. Kurpiewski, M. R. Grigorescu, A. Thermodynamic parameters of specific and nonspecific protein-DNA binding. Supramol. Chem. **2000**, *12*, 143-160.

[8] Liang, Y. Applications of isothermal titration calorimetry in protein science. ABBS **2008**, *40*, 565-576.

[9] Russu, I. M. Studying DNA-protein interactions using NMR. Trends Biotechnol **1991**, *9*, 96-104..

[10] Tereshkin, E. V. Tereshkina, K. B. Krupyanskii, Y. F. Predicting Binding Free Energies for DPS Protein-DNA Complexes and Crystals Using Molecular Dynamics. JSFI **2022**, *9*, 33-45.

[11] Kalodimos, C. G.; Boelens, R.; Kaptein, R. A residue-specific view of the association and dissociation pathway in protein–DNA recognition. Nat. Struct. Biol. **2002**, *9*, 193-197.

[12] Yonetani, Y.; Kono, H. Dissociation free-energy profiles of specific and nonspecific DNA–protein complexes. J. Phys. Chem. B **2013**, *117*, 7535-7545.

[13] Jakubec, D.; Vondrasek, J. Can All-Atom Molecular Dynamics Simulations Quantitatively Describe Homeodomain–DNA Binding Equilibria? J. Chem. Theory Comput. **2019**, *15*, 2635-2648.

[14] Aishima, J.; Gitti, R. K.; Noah, J. E.; Gan, H. H.; Schlick, T.; Wolberger, C. A Hoogsteen base pair embedded in undistorted B-DNA. Nucleic Acids Res. **2002**, *30*, 5244-5252.

[15] Kästner, J. Umbrella sampling. Wiley Interdiscip. Rev. Comput. Mol. Sci. **2011**, *1*, 932-942.

[16] Hub, J. S.; De Groot, B. L.; Van Der Spoel, D. g-wham-A free weighted histogram analysis implementation including robust error and autocorrelation estimates. J. Chem. Theory Comput. **2010**, *6*, 3713-3720.

[17] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free energy calculations on biomolecules. I. The method. J. Comput. Chem. **1992**, *13*, 1011-1021.

[18] Kumar, N.; Garg, P. Probing the molecular basis of cofactor affinity and conformational dynamics of mycobacterium tuberculosis elongation factor Tu: an integrated approach employing steered molecular dynamics and umbrella sampling simulations. J. Phys. Chem. B **2022**, *126*, 1447-1461.

[19] Ngo, S. T.; Vu, K. B.; Bui, L. M.; Vu, V. V. Effective estimation of ligand-binding affinity using biased sampling method. ACS omega **2019**, *4*, 3887-3893.

[20] Umbrella Sampling - GROMACS Tutorial - MD Tutorials (`http://www.mdtutorials.com/gmx/umbrella/index.html`)

[21] Chen, J.; Wu, M.; Gong, L.; Zhang, J.; Yan, B.; Liu, J.; Zhang, H.; Thundat, T.; Zeng, H. Mechanistic understanding and nanomechanics of multiple hydrogen-bonding interactions in aqueous environment. J. Phys. Chem. C **2019**, *123*, 4540-4548.

[22] Liu, B.; Li, J.; Qi, C.; Li, X.; Mai, T.; Zhang, J. Mechanism of asphaltene aggregation induced by supercritical CO 2: insights from molecular dynamics simulation. RSC Adv. **2017**, *7*, 50786-50793.

[23] Kumar, P.; Bhardwaj, V. K.; Purohit, R. Unveiling the intricate supramolecular chemistry of Y cyclodextrin epigallocatechin gallate inclusion complexes. Ind. Eng. Chem. Res. **2024**, *63*, 2544-2554.

[24] Das, A.; Chakrabarti, J.; Ghosh, M. Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation. Biophys. J. **2013**, *104*, 1274-1284.

[25] Das, A.; Chakrabarti, J.; Ghosh, M. Conformational thermodynamics of metal-ion binding to a protein. Chem. Phys. Lett. **2013**, *581*, 91-95.

[26] Das, A.; Chakrabarti, J.; Ghosh, M. Thermodynamics of interfacial changes in a protein–protein complex. Molecular Biosyst. **2014**, *10*, 437-445.

[27] Sikdar, S.; Chakrabarti, J.; Ghosh, M. A microscopic insight from conformational thermodynamics to functional ligand binding in proteins. Molecular Biosyst. **2014**, *10*, 3280-3289.

[28] Moulick A. G.; Chakrabarti, J. Fluctuation-dominated ligand binding in molten globule protein. J. Chem. Inf. Model. **2023**, *63*, 5583-5591.

[29] Mandal, S. C.; Maganti, L.; Mondal, M.; Chakrabarti, J. Microscopic insight to specificity of metal ion cofactor in DNA cleavage by restriction endonuclease EcoRV. Biopolymers **2020**, *111*, e23396.

[30] Kole, K.; Gupta, A. M.; Chakrabarti, J. Conformational stability and order of Hoogsteen base pair induced by protein binding. Biophys. Chem. **2023**, *301*, 107079.

[31] Biovia, D. S. Discovery studio modeling environment **2017**.

[32] DeLano, W. L.; Bromberg, S. PyMOL user's guide. DeLano Scientific LLC **2004**, 629.

[33] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to super-computers. SoftwareX **2015**,1-2,19-25.

[34] Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess B. The GROMACS development team. GROMACS User Manual (**2018**).

[35] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J. Chem. Theory Comput.**2015**, *11*, 3696-3713.

[36] Galindo-Murillo, R.; Robertson, J. C.; Zgarbovic, M.; Sponer, J.; Otyepka, M.; Jureska, P.; Cheatham, T. E. Assessing the current state of amber force field modifications for DNA. J. Chem. Theory Comput. **2016**, *12*, 4114-4127.

[37] Petrova S. S.; Solov'ev, A. D. The Origin of the Method of Steepest Descent. Historia Mathematica **1997**, *24*, 361.

[38] Lemak, A. S.; Balabaev, N. K. On the Berendsen thermostat. Mol. Simul. **1994**, *13*, 177- 187.

[39] Saito, H.; Nagao, H.; Nishikawa, K.; Kinugawa, K. Molecular collective dynamics in solid para-hydrogen and ortho-deuterium: The Parrinello–Rahman-type path integral centroid molecular dynamics approach. J. Chem. Phys. **2003**, *119*, 953-963.

[40] Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. J. Chem. Phys. **1995**, *103*, 3668-3679.

[41] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. **1997**, *18*, 1463-1472.

[42] Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu' H.; Molnar, F.; Wriggers, W.; Schulten, K. Steered molecular dynamics, In Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling. Berlin, May 21–24 (**1997**) 39-65, Springer Berlin Heidelberg.

[43] Lemkul, J. A.; Bevan, D. R. Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics. J. Phys. Chem. B **2010**, *114*, 1652-1660.

[44] Gotzias, A.; Tocci, E.; Sapalidis, A. Solvent-Assisted Graphene Exfoliation from Graphite Using Umbrella Sampling Simulations. Langmuir **2023**, *39*, 18437-18446.

[45] Aho, N.; Groenhof, G.; Buslaev, P. Do all paths lead to Rome? How reliable is umbrella sampling along a single path?. J. Chem. Theory Comput. **2024**, *20*, 6674-6686.

[46] Jeffrey, G. A.; Saenger, W. Hydrogen bonding in biological structures. Springer Science & Business Media **2012**.

[47] Warshel, A.; Russell, S. T. Calculations of electrostatic interactions in biological systems and in solutions. Q. Rev. Biophys. **1984**, *17*, 283-422.

[48] Sagarik, K.; Chaiyapongs, S. Structures and stability of salt-bridge in aqueous solution. Biophys. Chem. **2005**, *117*, 119-140.

[49] Chakrabarti, P.; Pal, D. Main-chain conformational features at different conformations of the side-chains in proteins. Protein Eng. **1998**, *11*, 631-647.

[50] Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. Noise reduction in speech processing (**2009**) 1-4.

[51] Moulick, A. G.; Chakrabarti, J. Correlated dipolar and dihedral fluctuations in a protein. Chem. Phy. Lett. **2022**, *797*, 139574.

[52] Ratner, B. The correlation coefficient: Its values range between +1/-1, or do they? Journal of targeting, measurement and analysis for marketing **2009**, *17*, 139-142.

[53] Brokx, R. D.; Lopez, M. M.; Vogel, H. J.; Makhatadze, G. I. Energetics of target peptide binding by calmodulin reveals different modes of binding. J. Biol. Chem. 2001, 276, 14083-14091.

[54] Akke, M.; Brueschweiler, R.; Palmer III, A. G. NMR order parameters and free energy: an analytical approach and its application to cooperative calcium (2+) binding by calbindin D9k. J. Am. Chem. Soc. **1993**, *115*, 9832-9833.

[55] Ma, L.; Ahmed, Z.; Mikhonin, A. V.; Asher, S. A. UV resonance Raman measurements of poly-L-lysine's conformational energy landscapes: dependence on perchlorate concentration and temperature. J. Phys. Chem. B. **2007**, *111*, 7675-7680.

[56] Barbana, C.; Pérez, M.; Sánchez, L.; Dalgalarrondo, M.; Chobert, J.-M.; Haertlé, T.; Calvo, M. Interaction of bovine α-lactalbumin with fatty acids as determined by partition equilibrium and fluorescence spectroscopy. Int. Dairy J. **2006**, *16*, 18– 25,

[57] Hobbs, B.; Drant, J.; Williamson, M. P. The measurement of binding affinities by NMR chemical shift perturbation. J. Bimol. NMR, **2022**, *76*, 153-163.

Chapter 4 Free Energy Landscape between Watson-Crick and Hoogsteen Base Pairing Transitions

4.1 Introduction

Experimental studies reveal dynamic equilibrium between Watson-Crick (WC) and Hoogsteen (HG) base pairs (bps)[1–25] in naked DNA in solution. The significant impact of this fluctuation, known as "DNA breathing" on DNA recognition, binding and damage repair has driven mechanistic studies of the WC $\leftrightarrow$ HG transition.[1,26–28] Although the presence of protein gives conformational stability to the HG bp,[24] the WC $\leftrightarrow$ HG bp transition in protein-bound DNA has not been well reported in the literature.

The life time of HG bps is around millisecond (ms). The transition between the two base pairing modes occurs on timescales beyond the reach of conventional molecular dynamics (MD) simulations with current computational resources. Various enhanced sampling methods, including umbrella sampling,[1,29] transition path sampling,[1,30] metadynamics[1,31–33] and Markov state modeling[1,31] have been used to investigate the transitions between these two base-pairing modes. These computational studies provide an atomic-level insight into the process, revealing a complex mechanism driven by hydrogen bond (H-bond) breaking and formation between bps.[1]

In free DNA, the WC $\leftrightarrow$ HG bp transition via purine base flipping has been investigated using the aforementioned computational methods.[29–33] These studies examine the relative free energy difference between the two base pairing modes and identified multiple transition pathways, involving base flipping into either the major or minor groove, along with possible clockwise or counterclockwise rotations of the glycosidic angle, χ.[29,32] The small base opening pathways ($\sim$

0°) are shortcuts, facilitating the bp transition within the DNA duplex, while large base-opening pathways cause significant distortion, with the target base flipping out of the double helix.[29] Energetically, the transition is more likely to follow the small base-opening pathway.[29]

In Chapter 2, we show that binding of homeodomain proteins enhances the stability of the HG bp relative to the WC bp in the 1K61 system. In Chapter 3, we observe that both homeodomain proteins interact more strongly with the HG bp than with the WC bp in this system. The WC $\leftrightarrow$ HG bp transition in presence of the protein has not been reported in these Chapters. The HG $\leftrightarrow$ WC base transition in presence of protein (PDB id: 6P1M) using restrain-free-energy perturbation-release (R-FEP-R) method is reported.[1] This R-FEP-R calculations reveal that syn HG base is more stable than the anti WC base by 2.25 $\pm$ 0.05 kcal/mol.[1] However, this method does not compute the barrier height between HG and WC bp.[1] We use Well-Tempered Metadynamics (WT-MetaD) to compute the relative free energy difference between the two base pairing modes in homeodomain protein bound DNA in order to highlight the free energy barrier between the two states. To simplify the calculation, we employ a single collective variable (CV), the glycosidic torsion angle (χ) to see the lowest energy transition. In the presence of the protein (PDB ID: 1K61), the HG bp becomes approximately 3.5–4 kcal/mol more stable than the WC bp. Additionally, the transition energy barrier decreases to 8.4 $\pm$ 0.6 kcal/mol for the WC to HG transition and increases to 11.5 $\pm$ 0.4 kcal/mol for the HG to WC transition.

4.2 Methods

4.2.1 Systems

We perform simulations on four systems: (1) WC: a model of free WC-DNA, (2) HG: a free HG-DNA system, (3) WCP: a protein-bound WC-DNA system and (4) HGP: a protein-bound HG-DNA system. The HG and HGP systems are derived from the PDB structure 1K61 and all systems contain 15 bp DNA fragments as detailed in Chapters 2 and 3. Initial structures of the WC, HG, WCP and HGP systems as per the initial structures are shown in Figs. 4.1 (a)-(d), respectively.

4.2.2 Simulation setup

PLUMED 2.8 (Feb 22, 2022)[34] plugin with GROMACS 2021.4 package[35] with Amber14sb_OL15 force field (ff) (Appendix 2.5.2)[24] is used for our simulations.

Leapfrog algorithm (Appendix 2.5.1) is used to integrate the equations of motion. The TIP3P water model is used as the solvent. Periodic boundary conditions (Appendix 2.5.3) are applied in all three dimensions. The system is electrically neutralized by adding required number of sodium (Na^+) and chloride (Cl^-) ions. The potential energy is minimized using the steepest descent algorithm.[36] Then simulation is performed at 300K temperature and 1 atmosphere pressure maintaining an isothermal-isobaric (NPT) ensemble. We use the Berendsen thermostat (Appendix 2.5.4)[37] to maintain temperature and the Parrinello-Rahman barostat (Appendix 2.5.5)[38] to maintain constant pressure. The Lennard-Jones (LJ) and short-range electrostatic interactions are terminated at 10 Å. We use the Particle-Mesh Ewald (PME)[39] method (Appendix 2.5.6) to compute the long-range electrostatic interactions. LINCS[40] constraints are applied to all bonds involving hydrogen atoms. The system is then used for the production run.

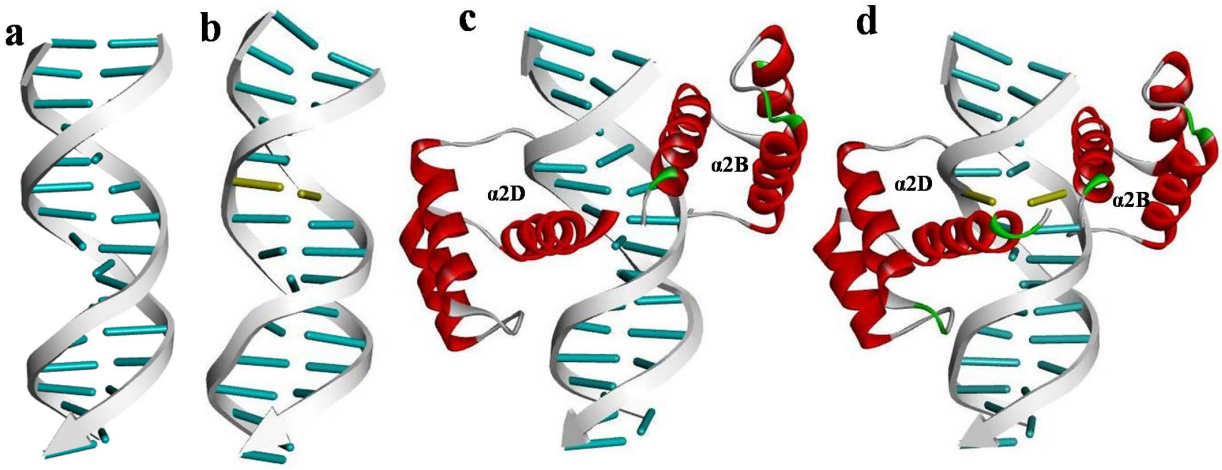


Figure 4.1: Initial snapshots of the (a) WC, (b) HG, (c) WCP and (d) HGP systems, respectively. HG bps are highlighted in yellow.

4.2.3 Well-Tempered Metadynamics (WT-MetaD)

WT-MetaD[41] (detailed in Appendix 4.6.1) is employed to compute the free energy landscape (FEL), using the glycosidic angle (χ) as the collective variable (CV). For the target nucleobase A7, χ characterizes the syn/anti glycosidic rotation. In standard WT-MetaD, a history-dependent bias potential is applied to enhance sampling during MD simulations. This external bias is constructed on the fly as a sum of Gaussians deposited along the CV coordinate (see detailed in Appendix 4.6.1). This external potential hinders sampling of frequently visited states and encourages exploration of less-sampled regions. As a result, conformational

sampling along the 1D CV surface (χ) is accelerated through more frequent barrier-crossing events.

In this study, WT-MetaD simulations are performed at constant temperature and pressure conditions of 300 K and 1 atm, respectively. In WT-MetaD, we set the Gaussian hill height to 1.2 kJ/mol, with a width of 0.35 rad. To maintain a well-tempered scheme, Gaussian hills with a bias factor of 15 are deposited every 1.0 ps onto the CV space. Five independent simulations (seeds) were performed under identical initial conditions to compute the average FES for bp transitions. The error associated with the average FES is computed by $\pm\ 0.5\sigma_m/\sqrt{N_w}$, where σ_m is the standard deviation of the average value and N_w (=5) is the number of independent FES. Each seed converged within 500 ns, resulting in a total of 2.5 μs of simulation time for computing the bp transition. We excluded the first 100 ns of the trajectory and used the final 400 ns for further analysis. We constructed the 1D FES based on CV, χ value through a reweighting process, as guided by PLUMED documentation.[42]

For nucleic acids, such as DNA and RNA, the most commonly used nonpolarizable force fields belong to the AMBER family, including Amber14sb_bsc1,[23] Amber99sb_bsc0[23] and Amber14sb_OL15.[24] We use the latest developed Amber14sb_OL15 force field[24] for our study. Additionally, we validated our WC $\leftrightarrow$ HG bp transition results in naked DNA using two other older force fields: Amber14SB_bsc1 and Amber99sb_bsc0.

4.2.4 Unbiased molecular dynamics simulation

Structures of the WC and HG base pairing conformations are extracted from WT-MetaD trajectories based on hydrogen bond (H-bond) donor–acceptor distances. A total of eight structures from the converged WT-MetaD trajectories of four different systems are used as starting points for 100 ns unbiased MD simulations. Coordinates are saved every 10 ps and H-bond donor–acceptor distances between N1(A)–N3(T) and N7(A)–N3(T) and C1′(A)-C1′(T) are continuously monitored to ensure the stability of WC and HG conformations throughout the simulation. Five 50 ns trajectory segments are extracted from the last 90 ns of each trajectory using overlapping time windows: 10–60 ns, 20–70 ns, 30–80 ns, 40–90 ns and 50–100 ns similar to ref 23.

4.2.5 Structural parameters of DNA base pairs and bases

DNA structure of bp is described using six bp step parameters and six intra bp parameters. The step parameters characterize the position and orientation of one bp relative to the next, while the intra bp parameters define the position and orientation of one base with respect to its complementary base within the same pair.[24] We computed these parameters using the NUPARM with the BPFIND software, as described in detail in Chapter 2. Each DNA base consists a sugar-base torsion angle (χ),[24] along with six torsion angles associated with the sugar-phosphate backbone: α, β, γ, δ, ϵ, and ζ. Additionally, five torsional angles characterize the sugar ring puckering: $\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$, and $\nu 4$.[24] These parameters act as microscopic conformational variables (θ) to compute the conformational thermodynamic data of DNA bases. Each backbone torsion angle is computed using GROMACS dihedral module.

4.2.6 Conformational thermodynamics

We utilize equilibrated trajectories from unbiased MD simulations to evaluate the conformational thermodynamics[2,24,43,44] (`https://sites.google.com/view/softmatter-snbncbs/code`) of the A7–T26 bp in the HG form relative to the WC form for naked DNA, following the same approach described in Chapters 2 and 3. Additionally, we computed the conformational thermodynamics of the A7–T26 bp in the HG form relative to the WC form in the presence of bound proteins.

4.3 Results

We use WT-MetaD trajectories to compute the free energy surface (FES) for bp transition. In this study, we primarily focus on the free energy pathway for the WC $\leftrightarrow$ HG bp transition. We begin with the transitions in absence of the homeodomain proteins in order to standardize our protocol with single glycosidic angle (χ) as CV to the values reported in literature. Next we apply similar protocol to WC $\leftrightarrow$ HG bp transition in presence of homeodomain proteins.

4.3.1 Free DNA

Let us consider WC $\rightarrow$ HG bp transition in free DNA, WC system. When the glycosidic torsion angle (χ) of the purine base A7 lies within the range

$-180° \leq \chi \leq -90°$ (anti),[29] it forms WC base pairing with T26, whereas within the range $0° \leq \chi \leq 90°$ (syn), it adopts HG base pairing.[45] We perform rotation of nucleobase A7 which is initially WC type H-bonded to T26 on the complementary strand of DNA in WT-MetaD simulations. In the case of the HG bp, the N7A7–N3T26 distance, rather than the N1A7–N3T26 distance, is ≤ 0.35 nm, facilitating H-bond formation.[45] In the HG bp, the distance between the C1′ atoms of the purine and pyrimidine is reduced to $\sim$ 0.85 nm, compared to $\sim$ 1.05 nm in the WC bp.[45]

To verify the WC and HG conformations throughout the WT-MetaD simulation, we monitor the H-bond donor–acceptor distances between N1(A)–N3(T) and N7(A)–N3(T), as well as the distance between the C1′ atoms of the purine and pyrimidine bases. Figs. 4.2(a)-(c) show different regions of WT-MetaD trajectory, confirming transitions between WC and HG conformations in the WC system. Then we compute the 1D FES illustrating the WC to HG bp transition in the WC system is shown in Fig. 4.3 (a). WC base pairing is formed near $\chi \sim -98°$, while HG base pairing occurs around $\chi \sim 60°$. The free energy of the WC bp is lower than HG bp. The free energy difference between the WC and HG base pairing modes ($\Delta G_{WC \rightarrow HG}$) is approximately 3.3 $\pm$ 0.2 kcal/mol (Fig. 4.3 (a)), in good agreement with the experimental value of 3.0–3.5 kcal/mol reported in,[1] derived from thermodynamic analysis of NMR relaxation dispersion spectroscopy data. The WC base pairing mode exhibits greater stability than the HG base pairing mode.

This result is also comparable to literature reported simulation values obtained using umbrella sampling (modified parmbsc0 force field),[29] mWTMetaD (bsc1_vdW, a variant of the Amberbsc1 force field),[32] TIS (Amber03 force field)[46] and meta-eABF[23] (CHARMM36 force field). The transition between these states involves an energy barrier near $\chi \sim 0°$. The barrier height is 10 $\pm$ 0.3 kcal/mol relative to the A7·T26 WC bp, consistent with the lowest energy barriers reported from simulations.[29,32] So, our Amber14sb_OL15 force field results are in good agreement with previously reported data. In contrast, the barrier estimated from NMR data is slightly higher, around 15 kcal/mol.[29] The discrepancy between simulation and experiment may arise from limitations in force field parameterization.

Next, we consider HG $\rightarrow$ WC transition in the same bp, A7-T26 in free DNA with HG bp, HG system. Here, we rotate the nucleobase A7 in WT-MetaD, which has initially HG type H-bonded to T26 on the complementary strand of DNA. We use the same Amber14sb_OL15 force field and WT-MetaD parameters as in

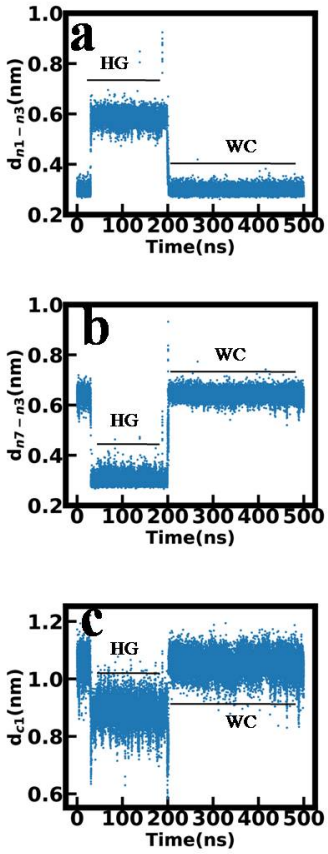


Figure 4.2: Time evolution of structural parameters in the WT-MetaD trajectory during the WC to HG base pairing transition in the WC system: (a) A7(N1)–T26(N3) distance, (b) A7(N7)–T26(N3) distance, and (c) C1'–C1' distance, highlighting transitions between WC and HG base pairing modes.

the WC $\rightarrow$ HG bp transition. The FES for the HG $\rightarrow$ WC bp transition is shown in Fig. 4.3 (b). The FES looks quite similar to $WC \rightarrow HG$ bp transition (Fig. 4.3(a)). The WC base pairing forms near $\chi \sim -100°$, while HG base pairing forms around $\chi \sim 60°$. The computed free energy difference between the HG and WC base pairing modes ($\Delta G_{HG \rightarrow WC}$) is approximately 3.3 $\pm$ 0.2 kcal/mol (Fig. 4.3 (b)) with the greater stability of the WC base pairing mode. The energy barrier for this bp transition is also located near $\chi \sim 0°$. The barrier height is about 7.0 $\pm$ 0.3 kcal/mol (Fig. 4.3 (b)), less compared to the WC to HG bp transition barrier observed in the WC system (Fig. 4.3 (a)).

We further examine the bp transition using two additional force fields, Amber14sb_bsc1 and Amber99sb_bsc0, as shown in Figs. 4.4(a)-(d). These correspond to the WC to HG transition in the A7–T26 bp of the WC system (Figs. 4.4(a)- (b)) and the HG to WC transition in the same bp of the HG system (Figs.

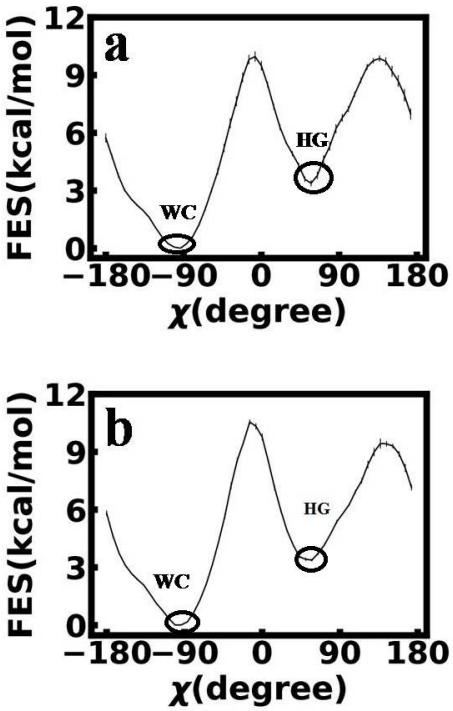


Figure 4.3: Free energy surfaces (FES) illustrating (a) the $WC \rightarrow HG$ bp transition in the WC system and (b) the $HG \rightarrow WC$ bp transition in the HG system, computed using the Amber14SB_OL15 force field. The minimum are marked in the plots.

4.4(c)-(d)), respectively. In Fig. 4.4(a), using the Amber14sb_bsc1 force field, the WC base pairing is observed near a χ value of approximately $-101°$, while the HG base pairing appears around $\chi \sim 70°$. The WC base pairing is 3.5 ± 0.3 kcal/mol more stable than HG base pairing. An energy barrier is located near $\chi \sim 0°$, with a barrier height of approximately 11 kcal/mol. In Fig. 4.4(b), using the Amber99sb_bsc0 force field, the WC base pairing is observed around $\chi \sim -94°$, while the HG base pairing forms near $\chi \sim 50°$. The free energy difference ($\Delta G_{\mathrm{WC}\rightarrow\mathrm{HG}}$) is 2.6 $\pm$ 0.3 kcal/mol with the higher stability of the WC base pairing. An energy barrier is present near $\chi \sim 0°$, with a height of about 5 kcal/mol.

For HG to WC base pairing transition in Fig. 4.4(c) for Amber14sb_bsc1 force field, the WC base pairing appears around $\chi \sim -94°$, while the HG base pairing forms near $\chi \sim 65°$. The corresponding free energy difference ($\Delta G_{\mathrm{HG}\rightarrow\mathrm{WC}}$) is 2.5 $\pm$ 0.3 kcal/mol, where the WC bp is more stable than HG bp. The transition involves an energy barrier located near $\chi \sim 0°$, with an estimated height of about 12 kcal/mol. In the case of the Amber99sb_bsc0 force field, as shown in Fig. 4.4(d), the WC base pairing forms near $\chi \sim -108°$, while the HG base pairing forms near $\chi \sim 50°$. The calculated free energy difference for the HG to WC

transition ($\Delta G_{\mathrm{HG}\rightarrow\mathrm{WC}}$) is 3 ± 0.1 kcal/mol. WC base pairing is observed to be more stable than HG base pairing. An energy barrier is present close to $\chi \sim 0°$, with an approximate height of 5 kcal/mol. As the recently developed Amber14sb_OL15 force field yields results that are more consistent with both previously reported simulations and experimental data, it is used for all further calculations.

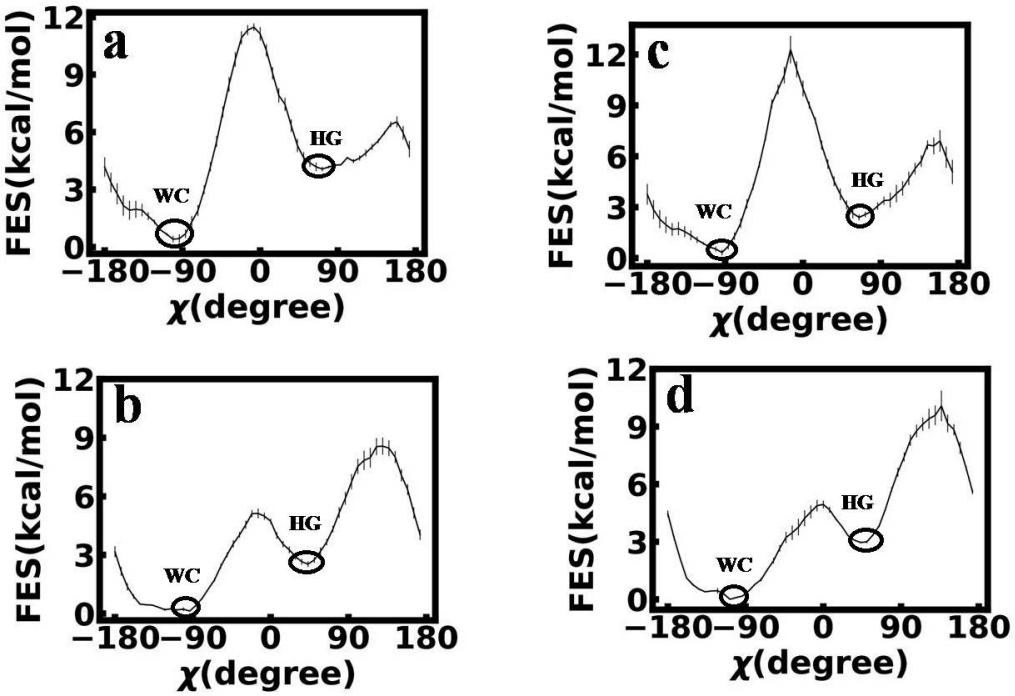


Figure 4.4: Free energy surfaces (FES) illustrating the $WC \rightarrow HG$ bp transition in the WC system for (a) Amber14sb_bsc1 and (b) Amber99sb_bsc0 force fields. Similarly, the $HG \rightarrow WC$ transition in the HG system is shown for (c) Amber14sb_bsc1 and (d) Amber99sb_bsc0 force fields. The minima are marked in each plot.

4.3.2 Protein bound DNA

Let us now examine how the presence of a protein modifies the transition between HG and WC base pairing. Here also we consider the A7:T26 bp to study the transition between WC $\leftrightarrow$ HG base pairing modes in the presence of homeodomain protein. We use the same parameters as those applied to the bp transition in naked DNA during the WT-MetaD simulation using the Amber14SB_OL15 force field.

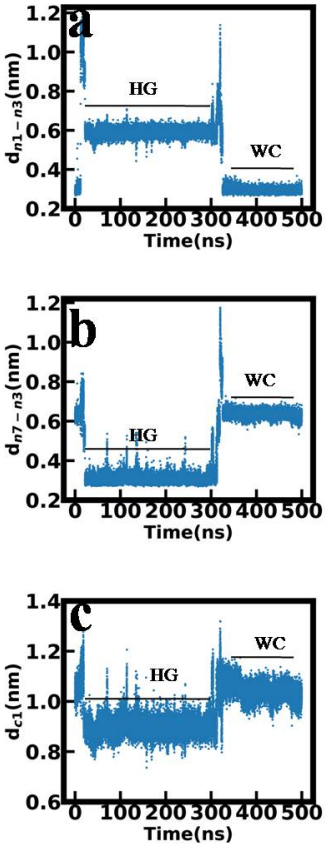


Figure 4.5: Time evolution of structural parameters in the WT-MetaD trajectory of the WCP system during the WC to HG base pairing transition: (a) A7(N1)–T26(N3) distance, (b) A7(N7)–T26(N3) distance and (c) C1'–C1' distance, highlighting conformational transitions between WC and HG base pairing modes.

We examine the WC $\rightarrow$ HG bp transition in the WCP system. In Figs. 4.5 (a)-(c), the regions in the WT-MetaD simulation are identified as the transition between WC and HG bp in the WCP system based on the H-bond donor–acceptor distances between N1(A)–N3(T) and N7(A)–N3(T), as well as the distance between the C1' atoms of the purine and pyrimidine bases respectively. Next, we compute the 1D FES for the WC to HG bp transition in the WCP system, as shown in Fig. 4.6 (a). In Fig. 4.6(a), the WC base pairing is observed near $\chi \sim -90^{\circ}$, while the HG base pairing is located around $\chi \sim 65^{\circ}$. The global minimum lies in the region corresponding to the HG bp formation, whereas a metastable minimum is observed near the WC base pairing region. This is in stark contrast to the free DNA case. We compute the relative free energy difference between the two base pairing modes and find HG bp is approximate 3.5 ± 0.6 kcal/mol more stable than WC bp, which is the opposite of what is observed in the naked

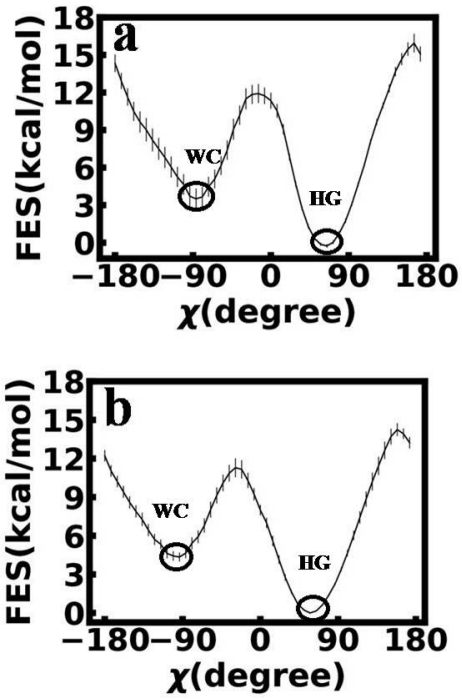


Figure 4.6: Free energy surfaces (FES) illustrate (a) the $WC \rightarrow HG$ bp transition in the WCP system and (b) the $HG \rightarrow WC$ transition in the HGP system, both in the presence of homeodomain proteins. Simulations use the Amber14sb_OL15 force field. All the minima are marked in each plot.

DNA, WC system. An energy barrier appears near $\chi \sim 0°$, with a height of approximately 8.4 $\pm$ 0.6 kcal/mol, which is lower than that observed in naked DNA. This indicates that, in the presence of the protein, the energetic cost for the WC $\rightarrow$ HG bp transition is reduced.

We, also check, the FES for HG to WC bp transition in the HGP system is shown in Fig. 4.6 (b). The FES is similar to $WC \rightarrow HG$ bp transition of Fig. 4.6(a). In Fig. 4.6(b), the WC base pairing is observed at a χ value of approximately $-94°$, while the HG base pairing appears near $\chi \sim 60°$. Similar to the WCP system, the global minimum lies in the HG bp region. We also compute the relative free energy difference between these two base pairing modes and find that the HG bp is more stable than the WC bp by approximately 4.0 ± 0.3 kcal/mol, with a transition barrier height of around 11.5 ± 0.4 kcal/mol. This scenario is completely different from the HG $\rightarrow$ WC bp transition in the naked DNA, where the barrier height is lower. Here, the interactions with proteins lead to an increased energy barrier.

4.4 Discussions

The time evolution of the χ angle throughout the simulation trajectory of the WC and WCP systems are shown Figs. 4.7(a) and (b), respectively. We observe that some structures lacking the necessary hydrogen bonds for bp formation can still be classified into one of the two categories, WC or HG based on the χ torsion angle values of Fig. 4.7, in both naked and protein-bound DNA cases. These observations suggest that χ torsion angle is not a reliable parameter for distinguishing between these two base pairing modes.

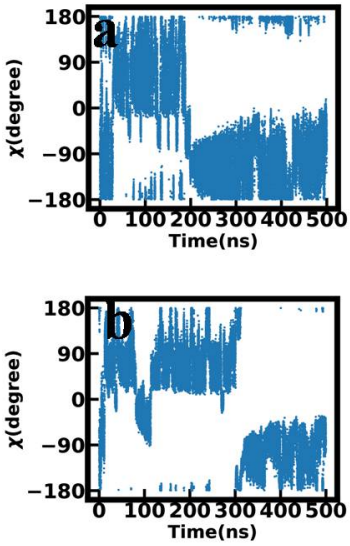


Figure 4.7: Time evolution of the χ angle during the WC to HG base pairing transition in the (a) WC and (b) WCP systems, respectively.

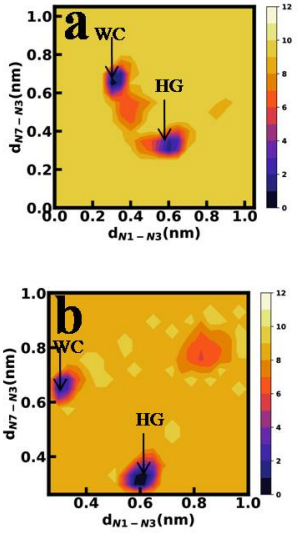


Figure 4.8: Free energy landscape of the conformational switching between WC and HG bp configurations plotted based on two H-bond donor–acceptor distances for (a) WC and (b) WCP systems, respectively.

We compute the free energy landscapes of Figs. 4.2 (a) and 4.6 (a) in Figs. 4.8 (a) and (b), respectively, in the H-bond donor–acceptor distances space. To compute the 2D FES based on H-bond donor-acceptor distances, we calculated the joint normalized probability distribution, P of distances between the N1 and N3 atoms (d_{N1-N3}) and between the N7 and N3 atoms (d_{N7-N3}). Then computed the 2D FES using the formula, G(d_{N1-N3}, d_{N7-N3}) = -k_BT lnP(d_{N1-N3}, d_{N7-N3}), where k_b is the Boltzmann constant and T is the simulation temperature. Figs. 4.8 (a) and (b) clearly distinguish the WC and HG base pairing states from distinct free energy minima within the configurational space. The relative free energy difference between the two base pairing states is approximately 3.00 kcal/mol and WC base pairing mode is more stable than HG base pairing in Fig. 4.8(a), which is in close agreement with previously reported experimental and simulation results and our results of Fig. 4.2(a) using χ as the CV. In Fig. 4.8(b), the HG base pairing is approximately 3.00 kcal/mol more stable than the WC base pairing in the presence of proteins, which is in close agreement with our results in Fig. 4.6(a) using χ as the CV. Therefore, H-bond donor–acceptor distance-based CVs are well-suited for distinguishing between WC and HG conformations during the post-processing of MD trajectories both in presence and absence of DNA binding proteins. This has already been reported for free DNA.[23] It is important to note that distance-based coordinates may not be the optimal choice as CVs for

applying a biasing force, since the transition between two base pairing modes primarily involves changes in the glycosidic torsion angle, χ. Therefore, the χ torsion angle is required as a CV to drive the bp transition and hydrogen bond donor–acceptor distance-based CVs are useful for identifying WC and HG bps during transitions. Although biased trajectories are used to construct the FES shown in Fig. 4.8, accurate calculation requires reweighting.

We perform unbiased simulations starting from randomly selected structures from WT-MetaD trajectories to ascertain that the bias protocols result in the transitions observed in our WT-MetaD simulations. We start from randomly selected configurations in which the HG conformation forms in WT-MetaD trajectories for free DNA, WC system. Throughout the equilibrated simulation trajectory, we monitor the averaged N1(A)–N3(T), N7(A)–N3(T) and C1'-C1' distances over windows, shown in Fig. 4.9 (a). In Fig. 4.9 (a) from N7(A)–N3(T) distance with time plot, we frequently observe bp opening, but no transition between the two base pairing modes. Figs. 4.9(b) and (c) show equilibrated snapshots at 10 ns and 13 ns, respectively, corresponding to cases where bp opening does not occur and does occur. Similarly, by performing unbiased simulations initiated from WT-MetaD trajectories in which the WC conformation forms in naked DNA, we monitored the averaged distances of N1 (A) - N3 (T), N7 (A) - N3 (T), and C1'-C1' over windows, shown in Fig. 4.10 (a). In Fig. 4.10 (a), we do not observe any bp opening or transition. Fig. 4.10(b) shows an equilibrated structure captured at 10 ns.

We also perform unbiased simulations on randomly selected structures from the WT-MetaD trajectories where the WC to HG bp transition occurs in the A7-T26 bp of the WCP system. We monitor the averaged N1(A)–N3(T), N7(A)–N3(T) and C1'-C1' distances over defined time windows, as shown in Fig. 4.11 (a). In the unbiased simulation, neither bp opening nor transition occurs, as shown in Fig. 4.11 (a). An equilibrated snapshot at 10 ns is shown in Fig. 4.11(b). Similarly, while performing an unbiased simulation starting from the WC bp configurations in the HGP system, we monitor the averaged N1(A)–N3(T), N7(A)–N3(T) and C1'–C1' distances over defined time windows throughout the equilibrated trajectory, as shown in Fig. 4.12(a). In some cases, the N1(A)–N3(T) distance fluctuates to slightly higher values, but no bp opening or transitions are observed. Figs. 4.12(b) and (c) show equilibrated snapshots at 10 ns and 11 ns, respectively, where the N1(A)–N3(T) distances are 0.28 nm and 0.35 nm. Despite these variations, WC base pairing remains intact.

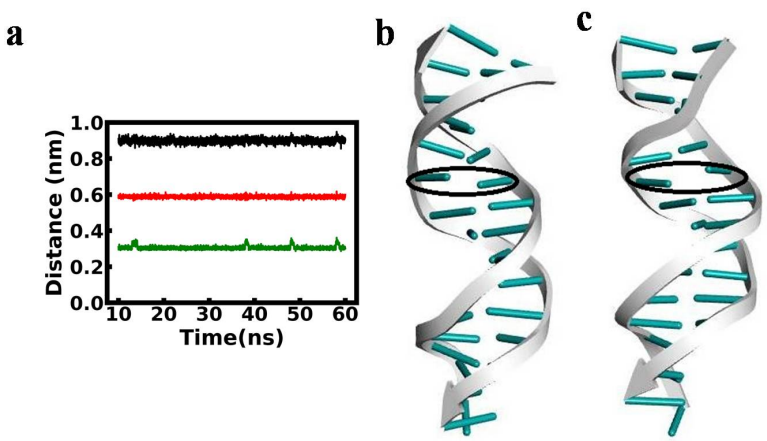


Figure 4.9: (a) Time-averaged C1′–C1′ (black), N1–N3 (red), and N7–N3 (green) distances computed over sliding windows for the A7–T26 bp in an unbiased simulation initiated from the HG base pairing conformation. (b) an equilibrated snapshot at 10 ns showing a stable bp with no opening, and (c) an equilibrated snapshot at 13 ns where bp opening is observed.

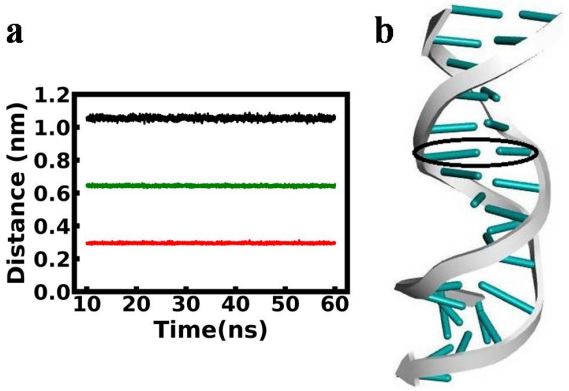


Figure 4.10: (a) Time-averaged C1′–C1′ (black), N1–N3 (red), and N7–N3 (green) distances computed over sliding windows for the A7–T26 bp in an unbiased simulation initiated from the WC base pairing conformation and (b) an equilibrated snapshot at 10 ns.

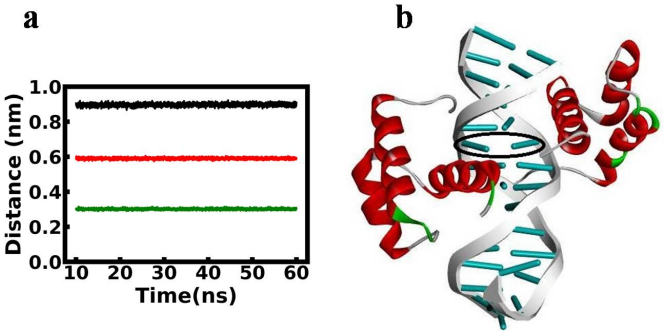


Figure 4.11: (a) The averaged C1′-C1′ (black), N1-N3 (red), and N7-N3 (green) distances over sliding windows for the A7-T26 bp in the unbiased simulation, starting from the HG base pairing mode of the A7-T26 in the WCP system and (b) an equilibrated snapshot at 10 ns.

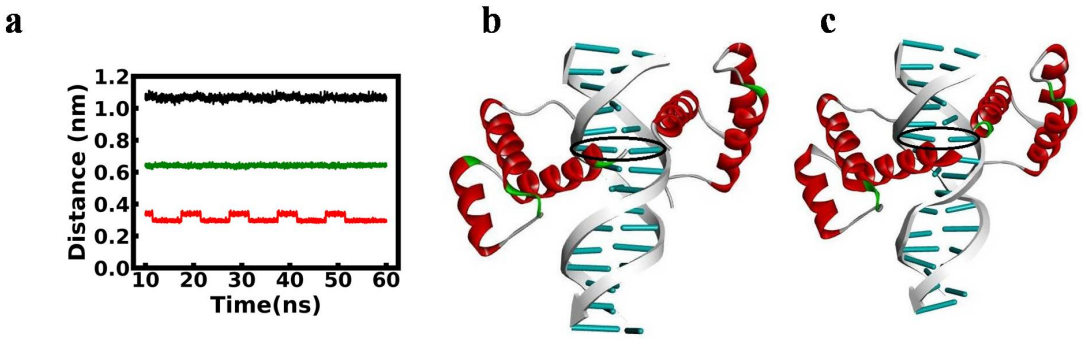


Figure 4.12: (a) The averaged C1′-C1′ (black), N1-N3 (red), and N7-N3 (green) distances over sliding windows for the A7-T26 bp in the unbiased simulation, starting from the WC base pairing mode of the A7-T26 in the HGP system and (b) and (c) are the equilibrated snapshots at 10 and 11 ns, respectively.

Our findings are very similar to the previously observed the WC $\leftrightarrow$ HG base transition in the presence of proteins using the R-FEP-R method.[1] However, their study did not report the transition barrier. The WT-MetaD method enables us to estimate the lowest energy cost associated with the WC $\leftrightarrow$ HG bp transitions in both naked and protein-bound DNA.

Previously in ref. 8 reported that the N-terminal arm, ARG132 of the non specifically bound protein α2D makes van der Waals contacts with the A7 base, as well as with the sugar-phosphate backbone of bases T6 and A7. This van der Waals interaction may stabilize the HG bp by restricting the sugar-phosphate backbone from adopting the *gauche-*/*gauche+* conformation of the α and γ torsion angles.[8] When the A7–T26 bp adopts the WC conformation, several unfavorable steric clashes occur between the α2D protein and the A7 base.[8] These steric clashes may favor the HG conformation of the A7–T26 bp upon protein binding.[8] This may explain the increased stability of the HG bp in this system. This is discussed in detail in Chapter 2.

We compute the conformational free energy and entropy for the A7–T26 bp in the HG form relative to the WC form in naked DNA, obtained from equilibrated unbiased simulations. The HG conformation is slightly destabilized ($\Delta G \sim 0.69 \pm 0.004$ kJ/mol) and ordered ($T\Delta S \sim -9.5 \pm 0.4$ kJ/mol) compared to the WC form, shown in Table 4.1. The conformational free energy is qualitatively consistent with the WT-MetaD results of Fig. 4.3. We also examine the conformational thermodynamics data of the A7–T26 bp in the HG form relative to the WC form in the presence of proteins, based on the unbiased simulation trajectories. As shown in Table 4.1, the HG form of the A7–T26 bp is significantly stabilized ($\Delta G \sim$ –5.3 $\pm$ 0.5 kJ/mol) and ordered ($T\Delta S \sim$ –10.28 $\pm$ 0.6 kJ/mol) compared to the WC form.

$\Delta G^{HG/WC}$	$T\Delta S^{HG/WC}$	$\Delta G^{HGP/WCP}$	$T\Delta S^{HGP/WCP}$
0.69 ± 0.004	-9.5 ± 0.4	–5.3 ± 0.5	–10.28 ± 0.6

Table 4.1: The changes (kJ/mol) in conformational thermodynamics data of A7-T26 bp in HG and HGP systems with respect to the WC and WCP systems.

This finding is qualitatively consistent with our WT-MetaD results in Fig. 4.6.

4.5 Conclusion

In summary, we perform WT-MetaD simulations to compute the relative free energy difference between WC and HG forms, as well as the transition barrier heights for the interconversion of these base-pairing forms in a protein–DNA complex. In naked DNA, the WC bp is 3.3 ± 0.2 kcal/mol more stable than the HG bp. The free energy barrier for the WC to HG transition is about 10.0 ± 0.3 kcal/mol, while the reverse HG to WC transition requires 7.0 ± 0.3 kcal/mol. In contrast, in the presence of proteins, this trend is reversed : the HG form is about 3.5–4.0 kcal/mol more stable than the WC form. Additionally, the energy barrier for HG to WC transition increases to 11.5 ± 0.4 kcal/mol, while the WC to HG transition barrier decreases to 8.4 ± 0.6 kcal/mol. Our studies provide insights into DNA breathing dynamics in the presence of homeodomain proteins and useful in DNA based technology.

4.6 Appendix

4.6.1 Well-tempered metadynamics (WT-MetaD)

In metadynamics,[47,48] a time-dependent external bias potential is applied within the space defined by a set of selected collective variables (CVs), denoted as $\vec{s}(q)$, which are functions of the system's microscopic coordinates. This bias potential is constructed as a summation of Gaussian functions deposited periodically along the trajectory in CV space:

$$V(\vec{s},t) = \sum_{k\tau<t} W(k\tau) \exp\left(-\sum_{i=1}^{d} \frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2}\right), \tag{4.1}$$

where τ is the time interval between the depositions of Gaussians, σ_i is the width of the Gaussian along the ith CV and $W(k\tau)$ represents the height of the Gaussian at time $k\tau$. The bias encourages the system to helps local minima and

explore new regions in phase space. In the long-time limit, the bias potential converges to minus the free energy as a function of the CVs:

$$V(\vec{s}, t \to \infty) = -F(\vec{s}) + C. \tag{4.2}$$

In standard metadynamics, Gaussian kernels with fixed height are continuously added throughout the simulation. The system is driven to explore high free-energy regions, and the free energy estimated from the bias potential fluctuates around the true value. In well-tempered metadynamics (WT-MetaD),[41,47] the height of the Gaussian is decreased with simulation time according to:

$$W(k\tau) = W_0 \exp\left(-\frac{V(\vec{s}(q(k\tau)), k\tau)}{k_B \Delta T}\right), \tag{4.3}$$

Here, W_0 denotes the initial height of the Gaussian kernel, ΔT is a user-defined parameter with units of temperature, and k_B is the Boltzmann constant. Due to this scaling of Gaussian heights, the bias potential converges gradually over time, but it does not fully compensate the underlying free energy. Instead, the long-time behavior of the bias potential is given by:

$$V(\vec{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\vec{s}) + C, \tag{4.4}$$

where T represents the system temperature. In the long time limit, the CVs effectively sample configurations from a system at a temperature, $T + \Delta T$. The parameter ΔT controls the extent of exploration in free energy space: when $\Delta T = 0$ corresponds to standard molecular dynamics; $\Delta T \to \infty$ to standard metadynamics.

In the context of well-tempered metadynamics, the term "bias factor" γ is often introduced, defined as the ratio between the effective temperature of the CVs ($T + \Delta T$) and the actual system temperature (T):

$$\gamma = \frac{T + \Delta T}{T}. \tag{4.5}$$

Selecting an appropriate bias factor γ is essential to ensure that the simulation can efficiently overcome free-energy barriers within the practical timescale of the simulation.

Bibliography

[1] Geronimo, I.; Vivo, M. D. Alchemical free-energy calculations of Watson–Crick and Hoogsteen base pairing interconversion in DNA. J. Chem. Theory Comput. **2022**, *18*, 6966-6973.

[2] Kole, K.; Chakrabarti, J. Binding of Homeodomain Proteins to DNA with Hoogsteen Base Pair. J. Phys. Chem. B **2025**, *129*, 1544-1554.

[3] Hoogsteen, K. The Crystal and Molecular Structure of a Hydrogen-Bonded Complex between 1- Methylthymine and 9-Methyladenine. Acta Crystallogr. **1963**, *16*, 907-916.

[4] Nair, D. T.; Johnson, R. E.; Prakash, S.; Prakash, L.; Aggarwal, A. K. Replication by Human DNA Polymerase-l Occurs by Hoogsteen Base-Pairing. Nature **2004**, *430*, 377-380.

[5] Nair, D. T.; Johnson, R. E.; Prakash, L.; Prakash, S.; Aggarwal, A. K. Human DNA Polymerase-l Incorporates DCTP Opposite Template G via a G.C+ Hoogsteen Base Pair. Structure **2005**, *13*, 1569-1577.

[6] Kitayner, M.; Rozenberg, H.; Rohs, R.; Suad, O.; Rabinovich, D.; Honig, B.; Shakked, Z. Diversity in DNA Recognition by p53 Revealed by Crystal Structures with Hoogsteen Base Pairs. Nat. Struct. Mol. Biol. **2010**, *17*, 423-429.

[7] Golovenko, D.; Bräuning, B.; Vyas, P.; Haran, T. E.; Rozenberg, H.; Shakked, Z. New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. Structure **2018**, *26*, 1237-1250.

[8] Aishima, J.; Gitti, R. K.; Noah, J. E.; Gan, H. H.; Schlick, T.; Wolberger, C. A Hoogsteen Base Pair Embedded in Undistorted B- DNA. Nucleic Acids Res. **2002**, *30*, 5244-5252.

[9] Patikoglou, G. A.; Kim, J. L.; Sun, L.; Yang, S.-H.; Kodadek, T.; Burley, S. K. TATA Element Recognition by the TATA Box-Binding Protein Has Been Conserved throughout Evolution. Genes Dev. **1999**, *13*, 3217-3230.

[10] Lu, L.; Yi, C.; Jian, X.; Zheng, G.; He, C. Structure Determination of DNA Methylation Lesions N 1 -MeA and N 3 -MeC in Duplex DNA Using a Cross-Linked Protein-DNA System. Nucleic Acids Res. **2010**, *38*, 4415-4425.

[11] Ling, H.; Boudsocq, F.; Plosky, B. S.; Woodgate, R.; Yang, W. Replication of a Cis-Syn Thymine Dimer at Atomic Resolution. Nature **2003**, *424*, 1083-1087.

[12] Genna, V.; Donati, E.; De Vivo, M. The Catalytic Mechanism of DNA and RNA Polymerases. ACS Catal. **2018**, *8*, 11103-11118.

[13] Geronimo, I.; Vidossich, P.; Donati, E.; De Vivo, M. Computational Investigations of Polymerase Enzymes: Structure, Function, Inhibition, and Biotechnology. WIREs Comput. Mol. Sci. **2021**, *11*, No. e1534.

[14] Ughetto, G.; Wang, A. H. J.; Quigley, G. J.; van der Marel, G. A.; van Boom, J. H.; Rich, A. A Comparison of the Structure of Echinomycin and Triostin A Complexed to a DNA Fragment. Nucleic Acids Res. **1985**, *13*, 2305-2323.

[15] Hoogsteen, K. IUCr,The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. Acta Crystallogr. **1959**, *12*, 822-823.

[16] Patel, D. J.; Shapiro, L.; Kozlowski, S. A.; Gaffney, B. L.; Jones, R. A. Covalent carcinogenic O6-methylguanosine lesions in DNA structural studies of the O6meG.A and O6meG.G interactions in dodecanucleotide duplexes. J. Mol. Biol. **1986**, *188*, 677-692.

[17] Singh, U. S.; Moe, J. G.; Reddy, G. R.; Weisenseel, J. P.; Marnett, L. J.; Stone, M. P. Proton NMR of an oligodeoxynucleotide containing a propanodeoxyguanosine adduct positioned in a (CG)3 frameshift hotspot of Salmonella typhimurium hisD3052: Hoogsteen base-pairing at pH 5.8. Chem. Res. Toxicol. **1993**, *6*, 825-836.

[18] Yang, H.; Zhan, Y.; Fenn, D.; Chi, L. M.; Lam, S. L. Effect of 1- methyladenine on double-helical DNA structures. FEBS Lett. **2008**, *582*, 1629-1633.

[19] Golovenko, D.; Bräuning, B.; Vyas, P.; Haran, T. E.; Rozenberg, H.; Shakked, Z. New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. Structure **2018**, *26*, 1237-1250.

[20] Sathyamoorthy, B.; Shi, H.; Zhou, H.; Xue, Y.; Rangadurai, A.; Merriman, D. K.; Al-Hashimi, H. M. Insights into Watson-Crick/ Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m1A. Nucleic Acids Res. **2017**, *45*, 5586-5601.

[21] Rice, P. A.; Yang, S.-w.; Mizuuchi, K.; Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U- Turn. Cell **1996**, *87*, 1295-1306.

[22] Bunting, K. A.; Roe, S. M.; Headley, A.; Brown, T.; Savva, R.; Pearl, L. H. Crystal structure of the Escherichia coli dcm very-short- patch DNA repair endonuclease bound to its reaction product-site in a DNA superhelix. Nucleic Acids Res. **2003**, *31*, 1633-1639.

[23] Stone, S. E.; Ray, D.; Andricioaei, I. Force-field-dependent DNA breathing dynamics: a case study of Hoogsteen Base pairing in A6-DNA. J. Chem. Inf. Model. **2022**, *62*, 6749-6761.

[24] Kole, K.; Gupta, A. M.; Chakrabarti, J. Conformational stability and order of Hoogsteen base pair induced by protein binding. Biophys. Chem. **2023**, *301*, 107079.

[25] Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen Base Pairs in Canonical Duplex DNA. Nature **2011**, *470*, 498-502.

[26] Frank-Kamenetskii, M. D. DNA Breathes Hoogsteen. Artif. DNA PNA XNA **2011**, *2*, 1-3.

[27] Nikolova, E. N.; Zhou, H.; Gottardo, F. L.; Alvey, H. S.; Kimsey, I. J.; Al-Hashimi, H. M. A Historical Account of Hoogsteen Base-Pairs in Duplex DNA. Biopolymers **2013**, *99*, 955-968.

[28] Shi, H.; Kimsey, I. J.; Gu, S.; Liu, H.-F.; Pham, U.; Schumacher, M. A.; Al-Hashimi, H. M. Revealing A-T and G-C Hoogsteen Base Pairs in Stressed Protein-Bound Duplex DNA. Nucleic Acids Res. **2021**, *49*, 12540-12555.

[29] Yang, C.; Kim, E.; Pak, Y. Free Energy Landscape and Transition Pathways from Watson-Crick to Hoogsteen Base Pairing in Free Duplex DNA. Nucleic Acids Res. **2015**, *43*, 7769-7778.

[30] Vreede, J.; Pérez de Alba Ortíz, A.; Bolhuis, P. G.; Swenson, D. W. H. Atomistic Insight into the Kinetic Pathways for Watson-Crick to Hoogsteen Transitions in DNA. Nucleic Acids Res. **2019**, *47*, 11069-11076.

[31] Ray, D.; Andricioaei, I. Free Energy Landscape and Conforma- tional Kinetics of Hoogsteen Base Pairing in DNA vs. RNA. Biophys. J. **2020**, *119*, 1568-1579.

[32] Kim, H.; Yang, C.; Pak, Y. Free-Energy Landscape of a pH- Modulated G·C Base Pair Transition from Watson-Crick to Hoogsteen State in Duplex DNA. J. Chem. Theory Comput. **2021**, *17*, 2556-2565.

[33] Genna, V.; Carloni, P.; De Vivo, M. A Strategically Located Arg/Lys Residue Promotes Correct Base Paring during Nucleic Acid Biosynthesis in Polymerases. J. Am. Chem. Soc. **2018**, *140*, 3312-3321.

[34] `https://www.plumed.org/doc-v2.9/user-doc/html/_c_h_a_n_g_e_s-2-8.html`

[35] `https://manual.gromacs.org/2021.4/index.html`

[36] Petrova, S. S.; Solov'ev, A. D. The Origin of the Method of Steepest Descent. Hist. Math. **1997**, *24*, 361.

[37] Lemak, A. S.; Balabaev, N. K. On the Berendsen thermostat. Mol. Simul. **1994**, *13*, 177-187.

[38] Saito, H.; Nagao, H.; Nishikawa, K.; Kinugawa, K. Molecular collective dynamics in solid para-hydrogen and ortho-deuterium: The Parrinello-Rahman-type path integral centroid molecular dynamics approach. J. Chem. Phys. **2003**, *119*, 953-963.

[39] Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. J. Chem. Phys. **1995**, **103**, 3668-3679.

[40] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. **1997**, *18*, 1463-1472.

[41] Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. PRL **2008**, *100*, 020603.

[42] https://www.plumed.org/doc-v2.9/user-doc/html/lugano-3.html

[43] Das, A.; Chakrabarti, J.; Ghosh, M. Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation. Biophys. J. **2013**, *104*, 1274-1284.

[44] Mandal, S. C.; Maganti, L.; Mondal, M.; Chakrabarti, J. Microscopic insight to specificity of metal ion cofactor in DNA cleavage by restriction endonuclease EcoRV. Biopolymers **2020**, *111*, e23396.

[45] Zhou, H.; Hintze, B. J.; Kimsey, I. J.; Sathyamoorthy, B.; Yang, S.; Richardson, J. S.; Al-Hashimi, H. M. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. Nucleic Acids Res. **2015**, *43*, 3420-3433.

[46] Vreede, J.; Pérez de Alba Ortíz, A.; Bolhuis, P. G.; Swenson, D. W. H. Atomistic Insight into the Kinetic Pathways for Watson-Crick to Hoogsteen Transitions in DNA. Nucleic Acids Res. **2019**, *47*, 11069-11076.

[47] Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. WIREs: Comput. Mol. Sci. **2011**, *1*, 826-843.

[48] https://www.plumed.org/doc-v2.9/user-doc/html/belfast-6.html

Chapter 5 Conformational stability of α-synuclein in presence of ZnO nanoparticles

5.1 Introduction

Proteins typically adopt well-defined, stable three-dimensional (3D) structures under physiological conditions, folded into specific secondary and tertiary conformations. In contrast, some proteins lack a stable 3D structure under physiological conditions, these are known as intrinsically disordered proteins (IDPs). They exist as dynamic ensembles of conformations and are typically unfolded or only partially folded. α-synuclein (αS) is an intrinsically disordered protein (IDP), consisting of 140 amino acids and abundantly found in the presynaptic areas of neurons.[1] αS is the major component of Lewy bodies[2,3] and Lewy neurites.[4] However, the aggregation and amyloid fibril formation of αS are associated with Parkinson's disease (PD),[2,3,5] a progressive neuro-degenerative disorder characterized by the loss of dopaminergic neurons in the substantia nigra, leading to dopamine deficiency and motor impairments.[5,6]

The amino-acid sequence of αS is typically divided into three regions[4] based on physicochemical properties: (i) amphipathic N-terminal domain (residues 1-60) containing seven imperfect repeats of 11 amino acids each, mainly consisting of a conserved amphipathic KA/TKE/QGV hexameric motif,[7,8] (ii) the highly hydrophobic central domain (residues 61-95), which is referred to as the non-amyloid-β component (NAC)[9] and (iii) the acidic C-terminal domain (residues 96-140) consisting mainly acidic and proline residues,[4] adopts unstructured conformation.

The NAC region of αS is essential for its aggregation and amyloid fibril formation[4,10–20] and responsible for cytotoxicity.[12,15,16,20] In the early stages of

fibrillation, this NAC region undergoes a structural transition from helix to extended states, followed by the formation of β-sheets.[21] Various fragments within the NAC domain have been studied to identify an essential region responsible for amyloid formation and cytotoxicity. These studies reveal that while some fragments within the NAC domain exhibit a strong propensity for amyloid formation and cytotoxicity, others do not.[4,13,15,16,19] Some groups[4,16,22] mentioned that among the residues in the NAC region, a segment spanning residues 68 to 78 of αS termed NACore (68GAVVTGVTAVA78) plays a crucial role in both amyloid fibril formation and cytotoxicity. They showed the dimerization of the NACore peptide is the initial process of the aggregation and fibrillation processes and NACore dimer has both intramolecular and intermolecular β-bridges.

Recently, several studies have shown that various nanoparticles have directly effect on αS aggregation, either by preventing fibrillation or moderating the fibrillation process.[23] Gold and silica nanoparticles moderate αS fibrillation.[24,25] It is reported that Zinc ions (Zn^{2+}) are essential for the human body, acting as cofactors in central nervous systems.[26] Some studies shown the higher concentration of Zn^{2+} in PD affected brains, compared to the unaffected brains.[27,28] Experimental studies[7] reveal that αS binds to zinc oxide nanoparticles (ZnONPs). Using zeta potential analysis show that the nanoparticles have a net negative surface potential at physiological pH. Therefore, for effective adsorption of the protein onto the nanoparticle surface, the positively charged and/or neutral polar regions of αS interact with the negatively charged nanoparticle interface, while the negatively charged C-terminus flanking away from the nanoparticle core. The N-terminus of αS is rich in lysine residues, which is a part of the consensus hexameric (KTKEGV) motif, may be responsible for establishing strong electrostatic interactions with ZnONP interface. Although the studies[7] show that ZnONP can moderate the fibrillation rate of αS and may serve as a potential therapeutic agent against its aggregation, the conformations of αS induced by ZnONP has not been explored.

Motivated by these findings, we investigate the effect of neutral and oxygen-rich ZnONPs on αS conformations. We identify the key regions contributing the most essential coordinate for the changing the conformation of αS, both in the absence and presence of ZnONPs using the XGBoost machine learning technique. We observe that in the absence of ZnONP, the most essential coordinate falls within the NACore region. However, in the presence of both neutral and oxygen-rich ZnONPs, the most essential coordinate shifts away from the NACore region. We investigate whether the NACore region is more stabilized and ordered

in the presence of neutral or oxygen-rich ZnONP with respect to absence of ZnONP using conformational thermodynamics calculations. Our conformational thermodynamics calculations reveal that all except one residue in the NACore region become stabilized and ordered in the presence of both types of ZnONPs compared to their absence. However, the stability and order are greater in the presence of neutral ZnONP than oxygen-rich ZnONP. We compare the conformational preferences of NACore residues between absence and presence of neutral and oxygen-rich ZnONPs throughout the equilibrated AA MD trajectories. We observe that most of the NACore residues retain their helix conformation in the presence of both types of ZnONPs than their absence, with a stronger effect observed for neutral ZnONP.

5.2 Methods

5.2.1 Quantum mechanical calculations

We optimize the bulk ZnO (CIF id: mp-2133) using quantum mechanical (QM) calculation. QM calculations are carried out within the framework of density functional theory (DFT),[29] using the projector augmented-wave (PAW) method[30,31] as implemented in the Vienna Ab initio Simulation Package (VASP).[32] The generalized gradient approximation (GGA) is used with the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional. We use a plane-wave cutoff energy of 500 eV and the total energies are converged to within 10^{-4} eV per atom. The optimization of bulk ZnO is performed using an $8 \times 8 \times 4$ k-point mesh, ensuring that the Γ point is included. The optimized structure is then used to build the nanoparticle.

5.2.2 All atom molecular dynamics simulations

The GROMACS 2018.6 package,[33] along with the GROMOS54a7 force field (Appendix 2.5.2) (ff),[34] is used to perform AA simulations. Additionally, we incorporated the van der Waals nonbonded interaction parameters (c6 or σ and c12 or ϵ) from the ref. 35 into this force field. The valence charges used for zinc and oxygen are $+$ 2 and $-$ 2, respectively. The leapfrog algorithm[36] (Appendix 2.5.1) is used to integrate the equations of motion, and the SPC water model[37] is used as the solvent. Periodic boundary conditions are applied in all three dimensions. The system is neutralized by adding the required

number of sodium (Na^+) and chloride (Cl^-) ions. Periodic boundary conditions (Appendix 2.5.3) are applied in all three dimensions. The system is electrically neutralized by adding the require number of sodium (Na^+) and chloride (Cl^-) ions. The potential energy of the system is minimized using the steepest descent algorithm.[38] Then AA MD simulations are carried out at 300 K and 1 atm pressure under an isothermal–isobaric (NPT) ensemble. Temperature is controlled using the Berendsen thermostat[39] (Appendix 2.5.4), while pressure is controlled with the Parrinello–Rahman barostat[40] (Appendix 2.5.5). Lennard-Jones (LJ) and short-range electrostatic interactions are truncated at a cutoff distance of 10 Å. Long-range electrostatic interactions are calculated using the Particle-Mesh Ewald (PME) method[41] (Appendix 2.5.6). LINCS constraints[42] are applied to all bonds involving hydrogen atoms. A time step of 0.5 femtosecond (fs) is used for integration. System equilibration is confirmed by the saturation of the root mean square deviation (RMSD) over time. The equilibrated portion of the trajectory is used for further analysis.

5.3 Analysis

5.3.1 Root mean square deviation (RMSD)

The root mean square deviation (RMSD) is calculated to quantify the structural differences between two protein conformations—typically between an initial (reference) structure and a frame from the simulation trajectory. It is used to judge the equilibration of the system. RMSD calculations are performed using the gmx rms[43] utility in GROMACS.

5.3.2 Radius of gyration

We compute the radius of gyration (R_g) of the αS molecule using GROMACS module.[44]

5.3.3 Root mean square fluctuation (RMSF)

We measure the average fluctuation of each residue of αS throughout the simulation. The root mean square fluctuation (RMSF) is calculated using the rmsf module[45] implemented in GROMACS.

5.3.4 Interfacial interactions

An interface is considered to be formed when any atom from an αS residue is within 0.6 nm of any atom from the ZnONPs. Here, we consider H-bonds[46] and electrostatic interactions (attractive).[47] H-bonds are characterized based on distance and angle criteria, the distance between donor (D) and acceptor (A) atoms must be ≤ 0.35 nm and the donor–hydrogen–acceptor (D–$\hat{\text{H}}$–A) angle must be $\geq 150^{\circ}$. Electrostatic interactions are identified when the distance between oppositely charged atoms of αS residues and ZnONPs is less than 0.56 nm. We use the GROMACS H-bond analysis module[48] to identify H-bonds, while electrostatic interactions are characterized using Discovery Studio,[49] based on the defined cutoff criteria like our previous Chapters 2 and 3.

5.4 Dihedral angles

We evaluate the backbone (ϕ and ψ) and side chain ($\chi 1$) dihedral angles[50] for each residue of the protein using our in-house program.

5.4.1 Dihedral principal component analysis

We perform principal component analysis (PCA) on the backbone dihedral angles, ϕ and ψ, using the dPCA+ method.[51–53] The conventional dihedral PCA (dPCA),[51] which often faces problems with the proper projection of periodic angular data onto principal subspaces. The dPCA+ addresses this challenge by minimizing the residual projection error by transforming the data to shift the maximal gap of the sampling to the periodic boundary. The method involves constructing a histogram of the angular data using a bin width of five degrees and identifying the bin center with the lowest population as the maximal gap. When multiple bins have equally low populations, by summing the populations of their respective neighboring bins, select the one with lowest overall population. The data is then transformed to shift the low-density region to the periodic boundary. Once the data is successfully transform into a linear data, the covariance matrix is constructed and standard eigen decomposition is applied to extract the principal components.

5.4.2 Density based clustering

We perform a robust, density-based geometrical cluster analysis[54,55] over the conformational landscape defined in the hyperspace spanned by the dihedral PCs to generate microstates. Each frame in the trajectory is represented by its corresponding PCs values. For each frame, we count the number of neighboring frames located within a fixed radius R inside the hypersphere. This produces local population (P_R) around each frame. To normalize, the local population of each frame is divided by the maximum local population (P_R^{max}) observed across all frames. Then based on the normalize quantity the relative free energy is calculated using the expression $\Delta G = -ln((P_R)/(P_R^{max}))$, that ensures that min($\Delta G$) equal to zero. This provides a local estimate of the free energy for each frame of a given trajectory. Then continue the clustering procedure by screening the free energy landscape. First, we fix a cutoff value at a relatively low free energy (F) value ($F < 0.1 k_B T$). Select all frames below this cutoff; the remaining frames are ignored. Frames are combined into clusters, if they are geometrically connected within the lumping distance, d_{lump} = 2σ. σ is the standard deviation of the nearest neighbor distances. For the apo system, d_{lump} is 0.594, for the holo-neutral system, it is 0.863; and for the holo-oxygen-rich system, it is 0.692. for The energy cut-off increases gradually at a step of 0.1 $k_B T$, all frames with lower energy than the current energy cut off are lumped into clusters. Two sets of structures are considered disjoint and form distinct clusters if no structure from one set lies within the lumping threshold distance of any structure from the other set. At higher free energy more microstates appear and eventually be lumped. Finally, at the highest value of energy cut off all frames get a specific cluster membership. This step results in a tree-based representation of the free energy landscape, with leaf nodes corresponding to local basins defined by selected microstates that discretize the input trajectory. States are then defined based on the identified clusters. A cluster is accepted as node if it

(i) geometrically disconnected to all accepted clusters

(ii) is not at highest free energy

(iii) has a population higher than the minimal population P_{min}

The idea is to identify the local minima in the free energy landscape. P_{min} is usually given $4 \times 10^4 - 0.2$ percent of the total number of frames.

However, a small fraction of the frames (approximately 0.1%) are geometrically isolated, that is, they are effectively disconnected from the majority and are defined as noise.[55] In the next step of the geometrical clustering process, the noise

frames are assigned to one of the existing microstates to construct a complete microstate trajectory. Starting from the unassigned frame with the lowest free energy, the algorithm identifies the nearest microstate and assigns the frame to its population. We are facing poor sampling on the barriers. Therefore, the projection artifact frames are assigned in incorrect states. To avoid this problem, we reassigned those frames dynamically instead of geometrically. So, at the end of this step, we get some states of distinct geometry separated by large barriers.

5.4.3 Dynamical clustering

Dynamical clustering combine MD frames which are close in time evolution rather than close in geometry. Here, we use the most probable path algorithm (MPP) developed by Jain et al.[56] First, given a set of microstates, the corresponding transition matrix is computed. If the self-transition probability of a given state falls below a specified metastability threshold $Q_{\text{min}} \in (0, 1]$, the state is merged with the one to which it has the highest transition probability. This procedure is reiterated until, for a given Q_{min}, here are no more transitions. Thus, MPP dynamical clustering produces a set of energetically well-separated metastable states, along with an assignment of each trajectory frame to one of these states. However, the metastable states found by dynamical clustering show internal fluctuations.[54] Near energy barriers, such fluctuations can be misinterpreted as transitions, leading to inaccurate transition rates in the resulting model. To correct these errors, we employ dynamic coring,[55] This approach defines core regions by requiring that, following a transition, the trajectory remains in the new state for at least a minimum duration t_{min}. If this condition is not satisfied, the trajectory points are reassigned to the previously visited state.

5.4.4 Essential coordinates

We use supervised machine learning technique to identify internal essential coordinates. We follow xtreme gradient boosting (XGBoost) algorithm.[52,55,57] The algorithm construct a machine learning model using a trajectory of MD coordinates based on dihedral angles ϕ and ψ and metastable states obtained from clustering, in order to minimize the loss function. The overall accuracy of the model is constructed by dividing the dataset into training and test sets. The importance of a dihedral is given by the gain of the loss function value in the model. A dihedral that produces a higher gain in the loss function is considered more important for characterizing the state. Once the model is trained, all

dihedral angles are sorted according to their importance. Discarding a non-essential dihedral does not significantly impact the model's overall accuracy. Therefore, this allows to identify the essential coordinates that are essential to explicitly discriminate between the states. Here, all XGBoost parameters are taken from ref. 57.

5.4.5 Secondary structure

The secondary structure[58] of individual NACore residues is computed for each system based on equilibrated trajectories using Visual Molecular Dynamics (VMD) software.[59] We classified the secondary structure into three categories: helix, β-sheet and unstructured (corresponding to elements other than helices or β-sheets). The frequency of each secondary structure is calculated by counting the number of frames in which it occurs and normalizing by the total number of frames in the equilibrated trajectory. These normalized values are presented by bar plots for clear comparison.

5.4.6 Conformational thermodynamics

We use the conformational thermodynamics method[60–64] (`https://sites.google.com/view/softmatter-snbncbs/code`) to compute the changes in conformational free energy and entropy of NACore residues of the αS in the holo-neutral and holo-oxygen-rich forms separately with respect to the apo form, following the same approach, described in the Chapters 2 and 3.

5.4.7 Structural persistence

We compute the structural persistence (S_P)[52,65] parameter per residue using the formula:

$S_P = \frac{1}{N}\sum_{i=1}^{N} e^{(\Delta\phi_i/\Delta\phi_{max})}.e^{(\Delta\psi_i/\Delta\psi_{max})}$

Here, N denotes the total number of frames. $\Delta\phi_i$ and $\Delta\psi_i$ represent the absolute deviations in the dihedral angles ϕ and ψ of the residue in frame i relative to the reference frame. $\Delta\phi_{\text{max}}$ and $\Delta\psi_{\text{max}}$ correspond to the maximum possible variations observed in the Ramachandran plot. A value of $S_P = 1$ indicates no conformational change, while lower values of S_P reflect greater deviations from the reference structure.

5.5 Results

We study the following systems in our simulations: (1) apo-αS consists of residues 1–95 of αS, taken from the crystal structure with PDB ID: 1XQ8.[31] Fig. 5.1(a) shows the initial structure, which has a diameter of approximately 2.4 nm. We exclude the disordered C-terminal region (residues 96-140) from the initial structure, as it is unstructured from the crystal. (2) holo-neutral consists of αS with neutral ZnONP. We construct the ZnONP from the quantum mechanically energy-minimized bulk ZnO structure using PyMOL, containing an equal number of zinc and oxygen atoms (1356 each). The initial structure of neutral ZnONP is shown in Fig. 5.1 (b) and (3) holo-oxygen-rich consists of αS with an oxygen-rich ZnONP containing 1354 zinc and 1358 oxygen atoms. This nanoparticle is constructed in the same way as the neutral ZnONP, using PyMOL. Fig. 5.1 (c) shows the initial structure of this oxygen-rich ZnONP.

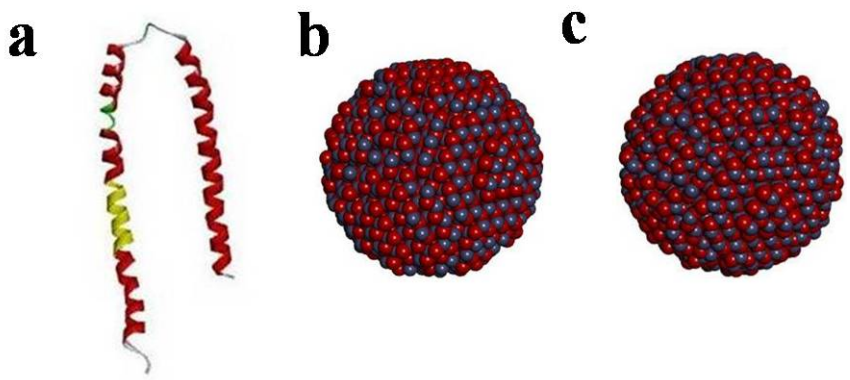


Figure 5.1: (a) The crystal structure of αS, with the NACore region highlighted in yellow, (b) the initial structure of neutral ZnONP and (c) the initial structure of oxygen-rich ZnONP. Zinc atoms are shown in blue and oxygen atoms in red in figures (b) and (c).

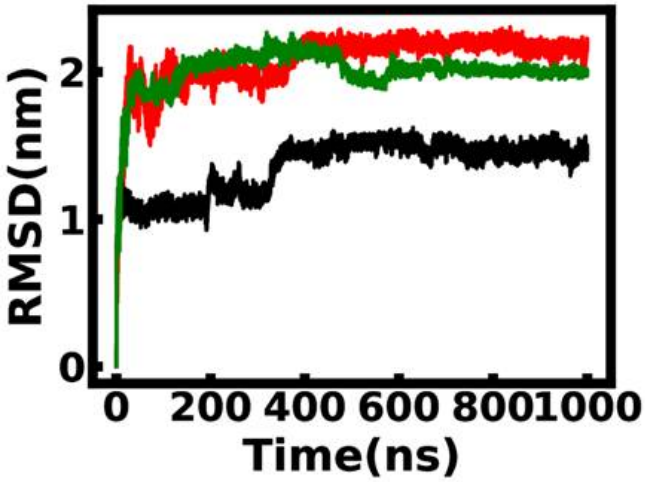


Figure 5.2: RMSD of αS in different systems as the function of simulation time. Apo, holo-neutral, and holo-oxygen-rich are shown in red, black and green, respectively.

We check the equilibration of the system from the saturation of the RMSD plot. The RMSD of αS in different systems as a function of simulation time is shown in Fig. 5.2. The systems reach equilibrium after approximately 600ns, as indicated by the onset of plateau in the RMSD curves. Fig. 5.3 (a)–(c) show the equilibrated snapshots of the apo, holo-neutral, and holo-oxygen-rich systems, respectively. Lower RMSD values of αS in the presence of ZnONPs, compared to the apo form indicates that ZnONPs enhance the structural stability of αS. We use the equilibrated trajectories (600ns-1μs) for further analysis.

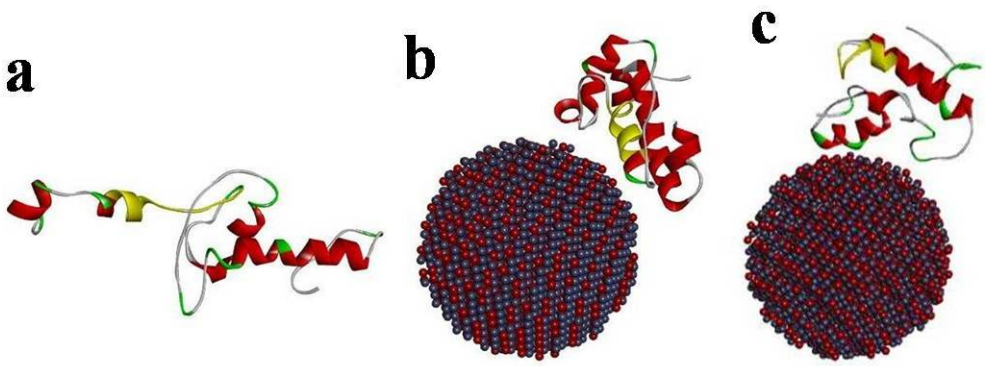


Figure 5.3: Equilibrated snapshots of (a) apo, (b) holo-neutral and (c) holo-oxygen-rich systems at 1 μs simulation time. In each snapshot, the NACore residues are highlighted in yellow.

We compute the radius of gyration (R_g) of αS throughout the trajectory in each system, as shown in Fig. 5.4. R_g decreases in the presence of ZnONPs compared to the system without ZnONPs. The decrease is more significant in the holo-oxygen-rich system than in the holo-neutral system. These results indicate that αS adopts a more compact, folded conformation in the presence of ZnONPs, the compactness being enhanced by the oxygen rich ZnONP.

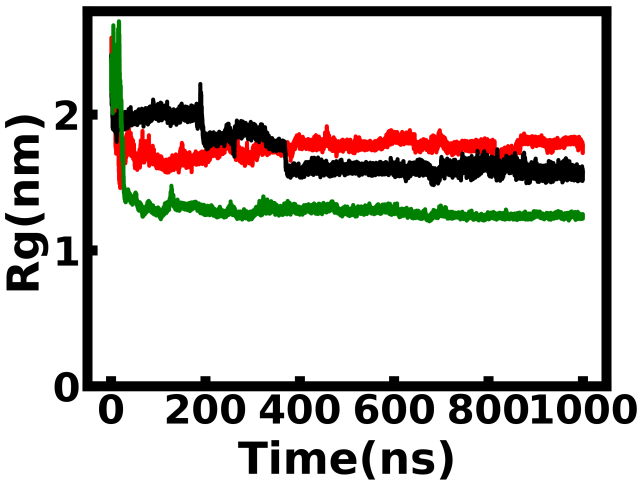


Figure 5.4: Radius of gyration R_g of αS in different systems. Apo, holo-neutral and holo-oxygen-rich are shown in red, black and green, respectively.

Then we check the RMSF of αS residues in different systems are shown in Fig. 5.5. We observe that the RMSF of most of the residues decreases in presence of neutral ZnONP in the holo-neutral system fro the apo system. The fluctuations are reduced more for the NACore residues and their neighboring residues.

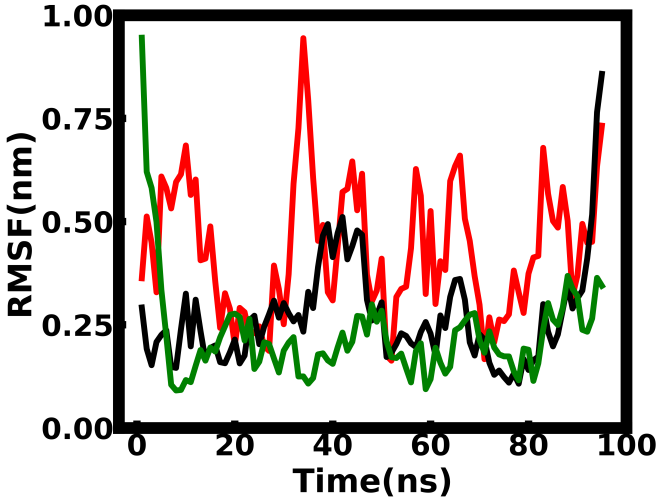


Figure 5.5: RMSF of αS per residue in different systems. Apo, holo-neutral and holo-oxygen-rich are shown in red, black and green, respectively.

5.5.1 Nano-particle protein interface

We identify the interface-forming residues of αS with ZnONPs. First, we focus on the holo-neutral system. The residues GLU13, GLU20, LYS21, LYS43, GLU57, LYS60, GLU61 and LYS80 of αS form stable interfaces with the neutral ZnONP surface. In Fig. 5.6(a), these residues are highlighted in yellow. Among them, LYS21 and LYS43 are consistent with experimentally identified interface forming residues. All GLU residues participate in electrostatic interactions, whereas the remaining LYS residues are mainly involved in hydrogen bonding.

Next, we identify the interface-forming residues of αS with the oxygen-rich ZnONP in the holo-oxygen-rich system. We observe that the residues GLU20, LYS21, LYS23, LYS32, LYS34, LYS43 and LYS45 form a stable interface with the nanoparticle. These interface forming residues are highlighted in yellow in Fig. 5.6(b). Most of the residues, LYS21, LYS23, LYS32, LYS34, LYS43 and LYS45 correspond to experimentally identified interface residues, indicating that the holo-oxygen-rich system more accurately reproduces experimentally observed interface interactions than the holo-neutral system. Here, GLU involves in electrostatic interactions with ZnONP, whereas all LYS are involved in hydrogen bonding with ZnONP like the holo-neutral case.

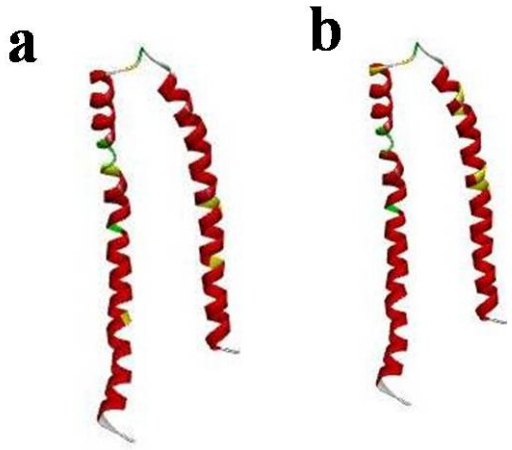


Figure 5.6: Snapshots of αS in the (a) holo-neutral and (b) holo-oxygen-rich systems, with interface-forming residues highlighted in yellow.

5.5.2 Meta-stability in the various forms of αS

Next, we evaluate the dihedral angle fluctuations of αS across different systems using the dPCA+ method on the simulated trajectories. One dimensional free energy landscape for PC1 to PC5 are shown in Figs. 5.7(a)-(e), respectively. The free energy landscapes along PC1 and PC2 show some metastability in both the apo and holo-neutral systems. No significant metastability are observed beyond PC2. Similar trend is observed in the holo-oxygen-rich system (Figs. 5.7(a)-(e)). Among the three, the apo system exhibits slightly higher metastability along PC1 compared to both holo systems (Fig. 5.7(a)). Along PC2 (Fig. 5.7(b)), the apo and holo-oxygen-rich systems exhibit a similar level of metastability, whereas the holo-neutral system shows lower in metastability.

Next, we perform density based geometrical clustering analysis in the hyperspace defined by PC1-PC5 to identify microstates. Here, we fix P_{min} (see the Methods section) equal to 24 which provides 33, 43 and 79 number of microstates for apo, holo-neutral and holo-oxygen-rich systems, respectively as discussed in ref. 52. Next, we use the most probable path algorithm (MPP) (see the Methods section) to reduce microstates in the set of the metastable states. In the MPP method, dynamically identified the more stable states from the less stable states. The most important parameter for MPP is the lagtime, τ_{mpp}, which is chosen in the time unit by the number of frames multiplied by the time step of the underlying simulation. We use the τ_{mpp} value 20 ps for each of the system. Because we take 40000 frames and 0.5 fs simulation time step in simulation. Q_{min} represents minimum metastability value. All states with a stability less then the given Q_{min}

value will be lumped according to their most probable state path.

For the apo system, we identified 32 metastable conformational states using a lag time of $\tau_{mpp} = 20$ ps and a minimum metastability of $Q_{\text{min}} = 0.800$, eliminating spurious transitions in the vicinity of the energy barrier by coring as discussed in Nagel et al.[55] For the holo-neutral system, we obtained 39 metastable states at $Q_{\text{min}} = 0.900$, using the same lag time of $\tau_{mpp} = 20$ ps. The holo-oxygen-rich system generates 32 metastable states at the end of the MPP step, using a Q_{min} value of 0.970 and the same lag time of $\tau_{mpp} = 20$ ps. In each case, the first few states carry about 50% of the total population.

Given the metastable states, we use the XGBoost model (see the Methods section) to identify the relative importance of the dihedral angles in the IDP state. The accuracy of the model is measured at each step by discarding one out of the 180 coordinates. The accuracy decreases sharply for most of the states by discarding the last 10 coordinates. These coordinates, shown in Table 5.1 as per decreasing importance with the secondary structure element in the crystal structure, are the essential coordinates (ECs) for all the three systems.

apo			holo-neutral			holo-oxygen-rich		
EC	**Residue name**	**Crystal structure**	**EC**	**Residue name**	**Crystal structure**	**EC**	**Residue name**	**Crystal structure**
$\psi 74$	VAL74	helix	$\phi 31$	GLY31	helix	$\psi 88$	ILE88	helix
$\phi 6$	LYS6	helix	$\psi 87$	SER87	helix	$\psi 72$	THR72	helix
$\psi 92$	THR92	helix	$\psi 94$	PHE94	unstructured	$\phi 21$	LYS21	helix
$\psi 65$	ASN65	helix	$\psi 40$	VAL40	unstructured	$\psi 93$	GLY93	unstructured
$\phi 4$	PHE4	helix	$\psi 38$	LEU38	unstructured	$\phi 94$	PHE94	unstructured
$\phi 93$	GLY93	unstructured	$\psi 39$	TYR39	unstructured	$\psi 66$	VAL66	helix
$\phi 85$	ALA85	helix	$\phi 42$	SER42	unstructured	$\psi 86$	GLY86	helix
$\psi 71$	VAL71	helix	$\phi 94$	PHE94	unstructured	$\psi 92$	THR72	helix
$\psi 91$	ALA91	helix	$\phi 41$	GLY41	unstructured	$\phi 47$	GLY47	helix
$\psi 82$	VAL82	helix	$\psi 90$	ALA90	helix	$\phi 67$	GLY67	helix

Table 5.1: Residues that possess ECs. The secondary structure of each residue in the initial crystal structure is mentioned.

The locations of the 10 essential coordinates are highlighted on the crystal structure of αS for all three systems, as shown in Figs. 5.7(f)–(h). The most essential coordinate is $\psi 74$ for apo system as it is obtained at the final iteration of XGBoost, removing all other coordinates. Most EC in the apo system (Table 5.1), fall within the NACore region. For the holo-neutral system, the most essential coordinate is $\phi 31$ (Table 5.1), located in the N-terminal region. In case of holo-oxygen-rich system, the most essential coordinate is $\psi 88$ (Table 5.1), located within the NAC region, although it is far from the NACore region. Thus, in the presence of both type of ZnONPs, the location of the the most essential

coordinate shifts away from the NACore segment.

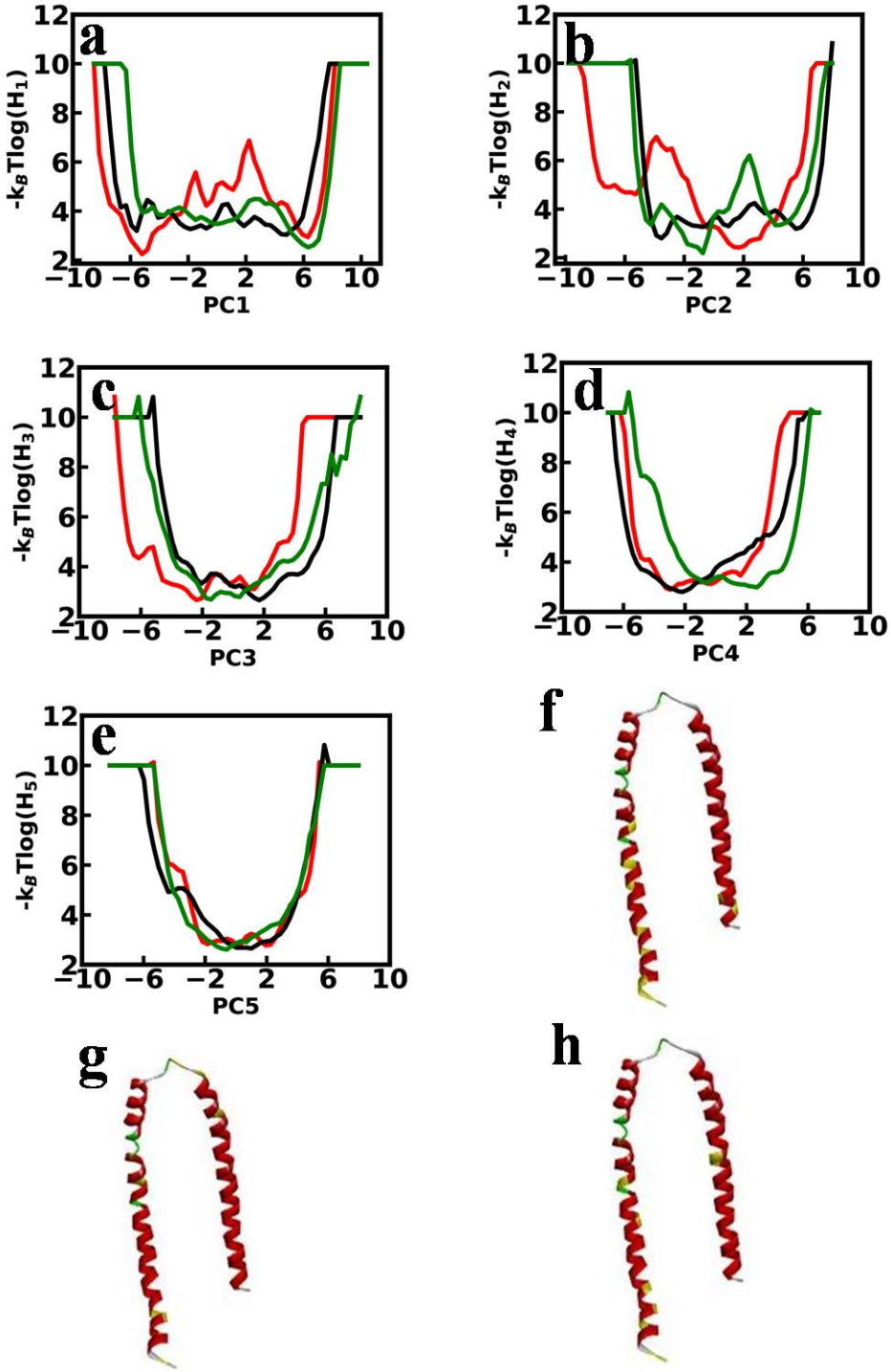


Figure 5.7: Free energy landscape for αS obtained from dPCA+ for (a) PC1, (b) PC2, (c) PC3, (d) PC4, and (e) PC5. The apo form is shown in red, holo-neutral in black and holo-oxygen-rich in green. (f)-(h) crystal structure of αS showing the essential coordinates for apo, holo-neutral and holo-oxygen-rich systems, respectively, in yellow.

5.5.3 Conformational thermodynamics of the NACore residues

In the crystal structure, all the NACore residues adopt a helical secondary structure. We analyze their preferred secondary structures throughout the equilibrated trajectories across the three different systems. In the apo system (Fig. 5.8(a)), GLY68 primarily adopts β-sheet and unstructured conformations, whereas in the holo-neutral system (Fig. 5.8(b)), it favors unstructured and helical states. Residues ALA69 to GLY73 predominantly adopt unstructured conformations with a minor β-sheet contribution in the apo system, while in the holo-neutral system they favor helical conformations with a small fraction of unstructured states. For residues VAL74 to ALA78, both helical and unstructured conformations are observed in the apo system, whereas in the holo-neutral system, they consistently adopt helical structures. Overall, the NACore residues show a greater loss of stable helical secondary structure in the apo system compared to the holo-neutral system.

On the other hand, GLY68 prefers unstructured conformations in the holo-oxygen-rich system (Fig. 5.8(c)), whereas in the apo system (Fig. 5.8(a)) it favors both β-sheet and unstructured conformations. ALA69 and VAL70 predominantly adopt unstructured and β-sheet conformations in the apo system, while favoring unstructured and helical conformations in the holo-oxygen-rich system. In the apo form, VAL71 primarily adopts an unstructured conformation with minor β-sheet content, whereas in the holo-oxygen-rich system, it predominantly adopts a β-sheet conformation with a small fraction of unstructured form. THR72 adopts a combination of unstructured and β-sheet conformations in both the apo and holo-oxygen-rich systems. No significant differences are observed in the conformational preferences of GLY73, which remains predominantly unstructured in both systems. In the apo system (Fig. 5.8(a)), residues VAL74 to ALA78 adopt both helical and unstructured conformations, while in the holo-oxygen-rich system (Fig. 5.8(c)), they adopt only helical states. Thus, except for VAL71, all NACore residues exhibit a greater loss of stable helical secondary structure in the apo system compared to the holo-oxygen-rich system.

We check the conformational stability and order of each residues of NACore segment in the holo systems with respect to the apo system in Figs. 5.9. First, we check for holo-neutral system with respect to the apo system (Figs. 5.9 (a) and (b)). Except for THR72, all residues of the NACore region are stabilized and ordered in the holo-neutral system. THR72 is destabilized but ordered. Then, we check the conformational stability and order of all residues of NACore region

in the holo-oxygen-rich system with respect to the apo system in Figs. 5.9 (c) and (d), respectively. We find that THR72 is destabilized and ordered, while all other residues are stabilized and ordered similar to holo-neutral case. However, the conformational stability and order are greater in the holo-neutral system compared to the holo-oxygen-rich system for the NACore segment.

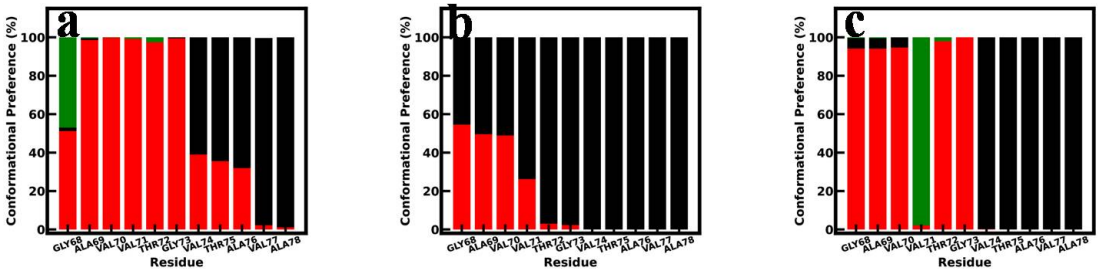


Figure 5.8: Conformational preferences of NACore residues in (a) apo, (b) holo-neutral and (c) holo-oxygen-rich systems. Helices are shown in black, β-sheets in green and unstructured in red.

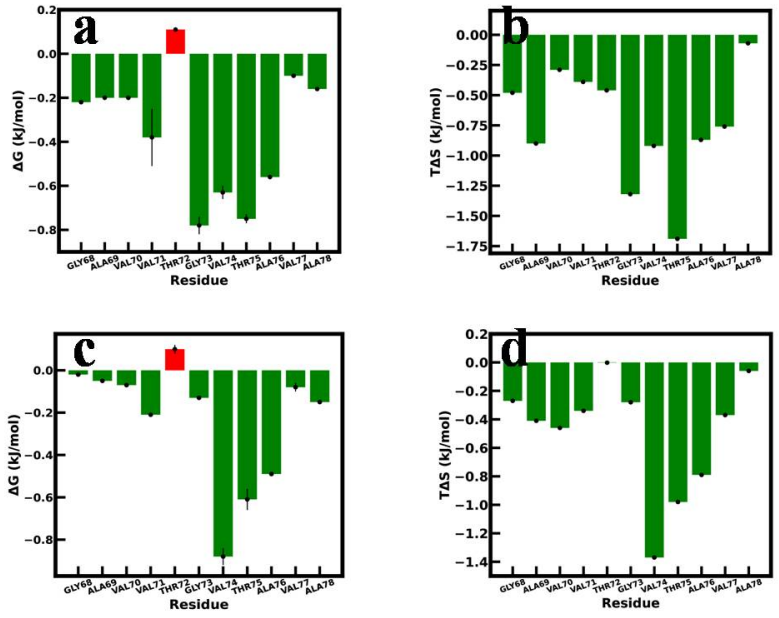


Figure 5.9: Conformational (a) free energy and (b) entropy changes of NACore residues in the holo-neutral system with respect to the apo system. Similarly, the conformational free energy and entropy changes of these NACore residues in the holo-oxygen-rich system with respect to the apo system is shown in (c) and (d), respectively. Conformational stabilized and ordered residues are shown green, whereas, destabilized and disordered residues are shown in red.

We also examine the structural persistence (S_P) values for all residues in the NACore region across the three systems, as shown in Fig. 5.10. We observe that all residues exhibit lower S_P values in the apo system compared to the holo-neutral system. Furthermore, except for GLY68, VAL71, THR72 and GLY73, the remaining residues also show lower S_P values in the apo system relative to the holo-oxygen-rich system. These results indicate that conformational changes occur more frequently in the apo system than in the holo-neutral system for all the residues of NACore segment. Similarly, the residues in the NACore region of the holo-oxygen-rich system show reduced conformational fluctuations compared to the apo system, with the exception of GLY68, VAL71, THR72 and GLY73. This trend qualitatively agrees for most of the residues with our observations from the conformational preference analysis performed on the equilibrated trajectories.

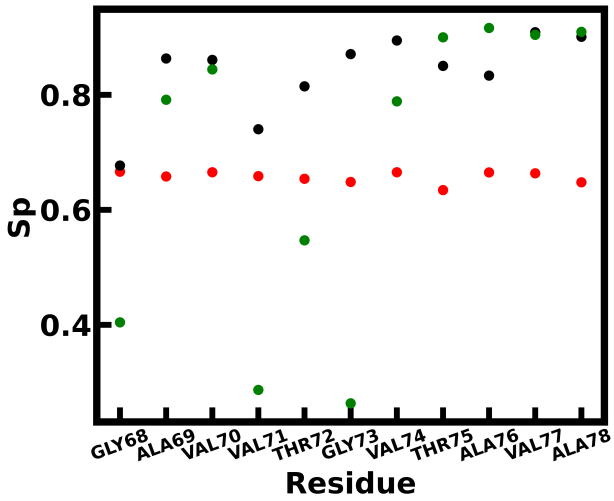


Figure 5.10: Structural persistence (S_P) values of NACore residues in apo (red), holo-neutral (black) and holo-oxygen-rich (green) systems.

5.6 Discussions

Experimentally, it has been reported that the interfaces of αS and ZnONP exhibit a net negative potential at physiological pH, as determined by zeta potential analysis. To promote protein adsorption onto the nanoparticle interface, the positively charged and/or neutral polar regions of the protein must interact with the negatively charged nanoparticle surface. The N-terminus, rich in lysine

residues and part of the consensus hexameric (KTKEGV) motif, plays a key role in establishing strong electrostatic interactions between αS and the ZnONP interface. In addition to electrostatic interactions, other forces such as salt bridges and van der Waals interactions contribute to the formation of the nanoparticle-protein complex. Our simulation studies on both holo systems reveal similar types of observations. The binding residues identified from the simulation data of the holo-oxygen-rich system match more closely with the experimentally identified binding residues than those from the holo-neutral system.

apo			holo-neutral			holo-oxygen-rich		
Residue name	Conformational preference	Crystal structure	Residue name	Conformational preference	Crystal structure	Residue name	Conformational preference	Crystal structure
GLY68	U, S & H	H	GLY68	U & H	H	GLY68	U & H	H
ALA69	U & S	H	ALA69	U & H	H	ALA69	U & H	H
VAL70	U & S	H	VAL70	U & H	H	VAL70	U & H	H
VAL71	U & S	H	VAL71	H & U	H	VAL71	S & U	H
THR72	U & S	H	THR72	H & U	H	THR72	U & S	H
GLY73	U & H	H	GLY73	H & U	H	GLY73	U	H
VAL74	U & H	H	VAL74	H	H	VAL74	H	H
THR75	U & H	H	THR75	H	H	THR75	H	H
ALA76	U & H	H	ALA76	H	H	ALA76	H	H
VAL77	H & U	H	VAL77	H	H	VAL77	H	H
ALA78	H & U	H	ALA78	H	H	ALA78	H	H

Table 5.2: Conformational preferences of NACore residues across equilibrated trajectories for three systems: *apo*, *holo*-neutral, and *holo*-oxygen-rich. Residue conformations are classified as Unstructured (U), Helix (H), or β-sheet (S) based on secondary structure assignments.

Since, the β-sheet content in the residues of the NACore region is almost zero in the presence of neutral ZnONP in the holo-neutral system, shown in Table 5.2. They retain their initial helical conformation more prominently in the holo-neutral system compared to the apo system. The stability of secondary structures follows the order: α-helix > β-sheet > unstructured. The NACore residues in the holo-neutral system predominantly adopt the most stable helical conformation (Table 5.2), leading to greater conformational stabilization and order with respect to those in the apo system in conformational thermodynamics calculations.

In the holo-oxygen-rich system, residues VAL71 and THR72 exhibit some β-sheet conformations, while in the apo system, a segment of residues GLY68 to THR72 adopts at least some β-sheet conformations (Table 5.2). In contrast, most NACore residues in the holo-oxygen-rich system predominantly adopt the more stable helical conformation, leading to conformational stability and order with respect to the apo system. VAL71 in the holo-oxygen-rich system favors the more stable β-sheet conformation, whereas in the apo system, it adopts the less stable unstructured form (Table 5.2). This preference for the more stable β-sheet conformation in the holo-oxygen-rich system contributes to its conformational

stabilization and ordering in the conformational thermodynamics data compared to the apo system.

Therefore, the NACore region maintains its helical secondary structure more effectively in the holo-neutral system than in the holo-oxygen-rich system. The lower conformational fluctuations of the NACore segment lead to greater stability and order in the holo-neutral system compared to the holo-oxygen-rich system. These observations suggest that ZnONPs may reduce the fibrillation rate of α-synuclein, with the effect being more pronounced in the presence of neutral ZnONPs compared to oxygen-rich ZnONPs. In contrast, the extended β-sheet segment in the apo system may facilitate the aggregation of monomeric αS.

5.7 Conclusion

We observe that in the presence of ZnONPs, the radius of gyration (R_g) of αS decreases and the most essential coordinate of αS, mostly responsible for its conformational fluctuations is located far from the NACore region. The NACore residues retain their initial helical conformation more throughout the equilibrated trajectory in the holo-neutral systems compared to the apo system. Most residues of the NACore region become stabilized and ordered in both holo systems compared to the apo system, with greater stability and ordering observed in the holo-neutral system. This study provides insights into the factors contributing to the stability of NACore residues in the presence of neutral and oxygen-rich ZnONPs. These findings may aid in understanding the aggregation mechanism and highlight the potential of ZnONPs as a therapeutic agent against α-synuclein aggregation. Studying aggregation phenomena requires simulating at least two αS molecules over extended timescales, which is difficult to achieve with conventional MD simulations. Coarse-grained (CG) simulations provide a possible approach to address this challenge. We have developed a simple coarse-grained model of the protein based on structural information. In the initial phase, we model our all-atom system, one protein interact with one ZnONP and attempt to reproduce all-atom MD simulation data to validate our model. A detailed study of our coarse-grained model is presented in the next chapter, Chapter 6.

Bibliography

[1] Maroteaux, L. Campanelli, J. T. Scheller, R. H. Synuclein: a neuron-specific protein localized to the nucleus and presynaptic nerve terminal. J. Neurosci. **1988**, *8*, 2804-2815.

[2] Spillantini, M. G. Crowther, R. A. Jakes, R. Hasegawa, M. Goedert, M. α-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. Proc. Natl. Acad. Sci. **1988**, *95*, 6469-6473.

[3] Baba, M. Nakajo, S. Tu, P. H. Tomita, T. Nakaya, K. Lee, V. M. Trojanowski, J. Q. Iwatsubo, T. Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. The Am. J. Pathol. **1998**, *152*, 879.

[4] Yamauchi, M. Okumura, H. Dimerization of α-Synuclein fragments studied by isothermal–isobaric replica-permutation molecular dynamics simulation. J. Chem. Inf. Model. **2021**, *61*, 1307-1321.

[5] Dauer, W. Przedborski, S. Parkinson's disease: mechanisms and models. Neuron **2003**, *39*, 889-909.

[6] Giasson, B. I. Jakes, R. Goedert, M. Duda, J. E. Leight, S. Trojanowski, J. Q. Lee, V. M. A panel of epitope-specific antibodies detects protein domains distributed throughout human α-synuclein in lewy bodies of Parkinson's disease. J. Neurosci. Res. **2000**, *59*, 528–533.

[7] Asthana, S. Bhattacharyya, D. Kumari, S. Nayak, P. S. Saleem, M. Bhunia, A. Jha, S. Interaction with zinc oxide nanoparticle kinetically traps α-synuclein fibrillation into off-pathway non-toxic intermediates. Int. J. Biol. Macromol. **2020**, *150*, 68-79.

[8] Clayton, D. F. George, J. M. The synucleins: a family of proteins involved in synaptic function, plasticity, neurodegeneration and disease, Trends Neurosci. **1998**, *21*, 249–254.

[9] Ueda, K. Fukushima, H. Masliah, E. Xia, Y. Iwai, A. Yoshimoto, M. Otero, D. A. Kondo, J. Ihara, Y. Saitoh, T. Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease. Proc. Natl. Acad. Sci. U. S. A. **1993**, *90*, 11282-11286.

[10] L Han, H. Weinreb, P. H. Lansbury, P. T. The core Alzheimer's peptide NAC forms amyloid fibrils which seed and are seeded by β- amyloid: is NAC a common trigger or target in neurodegenerative disease? Chem. Biol. **1995**, *2*, 163-169.

[11] Iwai, A. Yoshimoto, M. Masliah, E. Saitoh, T. Non-Aβ Component of Alzheimer's Disease Amyloid (NAC) is Amyloido-genic. Biochemistry **1995**, *34*, 10139-10145.

[12] El-Agnaf, O. M. A. Jakes, R. Curran, M. D. Middleton, D. Ingenito, R. Bianchi, E. Pessi, A. Neill, D. Wallace, A. Aggregates from mutant and wild-type α-synuclein proteins and NAC peptide induce apoptotic cell death in human neuroblastoma cells by formation of β-sheet and amyloid-like filaments. FEBS Lett. **1998**, *440*, 71-75.

[13] El-Agnaf, O. M. A. Bodles, A. M. Guthrie, D. J. S. Harriott, P. Irvine, G. B. The N-terminal region of non-Aβ component of Alzheimer's Disease amyloid is responsible for its tendency to assume beta-sheet and aggregate to form fibrils. Eur. J. Biochem. **1998**, *258*, 157-163.

[14] Biere, A. L. Wood, S. J. Wypych, J. Steavenson, S. Jiang, Y. Anafi, D. Jacobsen, F. W. Jarosinski, M. A.; Wu, G.-M. Louis, J.-C. Martin, F. Narhi, L. O. Citron, M. Parkinson's Disease-associated α- Synuclein Is More Fibrillogenic than β- and γ-Synuclein and Cannot Cross-seed Its Homologs. J. Biol. Chem. **2000**, **275**, 34574-34579.

[15] Bodles, A. M. Guthrie, D. J. S. Harriott, P. Campbell, P. Irvine, G. B. Toxicity of non-Aβ component of Alzheimer's disease amyloid, and N-terminal fragments thereof, correlates to formation of β-sheet structure and fibrils. Eur. J. Biochem. **2000**, *267*, 2186-2194.

[16] Bodles, A. M. Guthrie, D. J. S. Greer, B. Irvine, G. B. Identification of the region of non-Aβ component (NAC) of Alzheimer's disease amyloid responsible for its aggregation and toxicity. J. Neurochem. **2001**, *78*, 384-395.

[17] Giasson, B. I. Murray, I. V. J. Trojanowski, J. Q. Lee, V. M. Y. A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of α-Synuclein Is Essential for Filament Assembly. J. Biol. Chem. **2001**, *276*, 2380-2386.

[18] Li, H. T. Du, H.-N. Tang, L. Hu, J. Hu, H. Y. Structural transformation and aggregation of human α-synuclein in trifluor-oethanol: Non-amyloid component sequence is essential and β-sheet formation is prerequisite to aggregation. Biopolymers **2002**, *64*, 221-226.

[19] Du, H. N. Tang, L. Luo, X.-Y. Li, H.-T. Hu, J.; Zhou, J. W. Hu, H. Y. A Peptide Motif Consisting of Glycine, Alanine, and Valine Is Required for the Fibrillization and Cytotoxicity of Human α-Synuclein. Biochemistry **2003**, *42*, 8870-8878.

[20] Periquet, M. Fulga, T. Myllykangas, L. Schlossmacher, M. G. Feany, M. B. Aggregated α-Synuclein Mediates Dopaminergic Neurotoxicity In Vivo. J. Neurosci. **2007**, *27*, 3338-3346.

[21] Khammari, A. Arab, S. S. Ejtehadi, M. R. The hot sites of α-synuclein in amyloid fibril formation. Sci. Rep. **2020**, *10*, 12175.

[22] Rodriguez, J. A. et al. Structure of the toxic core of α-synuclein from invisible crystals. Nature **2015**, *525*, 486-490.

[23] Asthana, S. Hazarika, Z. Nayak, P. Z. Roy, J. Jha, A. N. Mallick, B. Jha, S. Insulin adsorp- tion onto zinc oxide nanoparticle mediates conformational rearrangement into amyloid-prone structure with enhanced cytotoxic propensity, Biochimica et Biophysica Acta (BBA)-General Subjects **2019**, *1863*, 153–166.

[24] Alvarez, Y. D. Fauerbach, J. A. Pellegrotti, J. Jovin, T. M. Jares-Erijman, E. A. Stefani, F. D. Influence of gold nanoparticles on the kinetics of α-synuclein aggregation, Nano Lett. **2013**, *13*, 6156-6163.

[25] Xie, H. Wu, J. Silica nanoparticles induce alpha-synuclein induction and aggregation in PC12-cells. Chem. Biol. Interact. **2016**, *258*, 197–204.

[26] Frederickson, C. J. Suh, S. W. Silva, D. Frederickson, C. J. Thompson, R. B. Importance of zinc in the central nervous system: the zinc-containing neuron, J. Nutr. **2000**, *130*, 1471S–1483S.

[27] Sawaya, M. R. Sambashivan, S. Nelson, R. Ivanova, M. I. Sievers, S. A. Apostol, M. I. Thompson, M. J. Balbirnie, M. Wiltzius, J. J. McFarlane, H. T. Atomic structures of amyloid cross-[beta] spines reveal varied steric zippers. Nature **2007**, *447*, 453.

[28] Coelho-Cerqueira, E. Carmo-Gonçalves, P. Pinheiro, A. Cortines, J. Follmer, C. α- Synuclein as an intrinsically disordered monomer–fact or artefact? FEBS J. **2013**, *280*, 4915–4927.

[29] Mandal, S. C. Chakrabarti, J. Surface specific adsorption of glucose to ZnO. Phys. Chem. Chem. Phys. **2023**, *25*, 7805-7814.

[30] Kresse G. Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys. Rev. B: Condens. Matter Mater. Phys. **1999**, *59*, 1758.

[31] Blo chl, P. E. Projector augmented-wave method, Phys. Rev. B: Condens. Matter Mater. Phys. **1994**, *50*, 17953.

[32] Kresse G. Furthmuller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B: Condens. Matter Mater. Phys. **1996**, *54*, 11169.

[33] Abraham, M. J. van der Spoel, D. Lindahl, E. Hess, B. TheGROMACS Development Team. GROMACS User Manual **2018.6**.

[34] Schmid, N. Eichenberger, A. P. Choutko, A. Riniker, S. Winger, M. Mark, A. E. van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions 54A7 and 54B7, Eur. Biophys. J. **2011**, *40*, 843–856.

[35] Raymand, D. Van Duin, A. C. Baudin, V. M. Hermansson, K. A reactive force field (ReaxFF) for zinc oxide, Surf. Sci. **2008**, *602*, 1020–1031.

[36] Frenkel, D. Smit, B. Understanding molecular simulation: from algorithms to applications. Elsevier **2023**.

[37] Mark, P. Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J. Phys. Chem. A **2001**, **105**, 9954-9960.

[38] Petrova, S. S. Solov'ev, A. D. The Origin of the Method of Steepest Descent. Hist. Math. **1997**, *24*, 361.

[39] Lemak, A. S. Balabaev, N. K. On the Berendsen thermostat. Mol. Simul. **1994**, *13*, 177-187.

[40] Saito, H. Nagao, H. Nishikawa, K. Kinugawa, K. Molecular collective dynamics in solid para-hydrogen and ortho-deuterium: The Parrinello-Rahman-type path integral centroid molecular dynamics approach. J. Chem. Phys. **2003**, *119*, 953-963.

[41] Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. J. Chem. Phys. **1995**, **103**, 3668-3679.

[42] Hess, B. Bekker, H. Berendsen, H. J. Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. **1997**, *18*, 1463-1472.

[43] `https://manual.gromacs.org/current/onlinehelp/gmx-rms.html`

[44] `https://manual.gromacs.org/current/onlinehelp/gmx-gyrate.html`

[45] `https://manual.gromacs.org/current/onlinehelp/gmx-rmsf.html`

[46] Jeffrey, G. A. Saenger, W. Hydrogen bonding in biological structures. Springer Science & Business Media **2012**.

[47] Warshel, A. Russell, S. T. Calculations of electrostatic interactions in biological systems and in solutions. Q. Rev. Biophys. **1984**, *17*, 283-422.

[48] `https://manual.gromacs.org/current/onlinehelp/gmx-hbond.html`

[49] Biovia, D. S. Discovery studio modeling environment **2017**.

[50] Chakrabarti, P. Pal, D. Main-chain conformational features at different conformations of the side-chains in proteins. Protein Eng. **1998**, *11*, 631-647.

[51] Sittel, F. Filk, T. Stock, G. Principal component analysis on a torus: Theory and application to protein dynamics. J. Chem. Phys. **2017**, *147*.

[52] Moulick, A. G. Chakrabarti, J. Conformational fluctuations in the molten globule state of α-lactalbumin. Phys. Chem. Chem. Phys. **2022**, *24*, 21348-21357.

[53] Moulick, A. G. Chakrabarti, J. Fluctuation-dominated ligand binding in molten globule protein. J. Chem. Inf. Model. **2023**, *63*, 5583-5591.

[54] Sittel, F. Stock, G. Robust density-based clustering to identify metastable conformational states of proteins. J. Chem. Theory Comput. **2016**, *12*, 2426-2435.

[55] Nagel, D. Weber, A. Lickert, B. Stock, G. Dynamical coring of Markov state models. J. Chem. Phys. **2019**, *150*.

[56] Jain, A. Stock, G. Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering. J. Phys. Chem. B **2014**, *118*, 7750-7760.

[57] Brandt, S. Sittel, F. Ernst, M. Stock, G. Machine learning of biomolecular reaction coordinates. J. Phys. Chem. Lett. **2018**, *9*, 2144-2150.

[58] Pelton, J. T. McLean, L. R. Spectroscopic methods for analysis of protein secondary structure. Anal. Biochem. **2000**, *277*, 167-176.

[59] Humphrey, W. Dalke, A. Schulten, K. VMD: visual molecular dynamics. J. Mol. Graph. **1996**, *14*, 33-38.

[60] Das, A. Chakrabarti, J. Ghosh, M. Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation. Biophys. J. **2013**, *104*, 1274-1284.

[61] Das, A. Chakrabarti, J. Ghosh, M. Conformational thermodynamics of metal-ion binding to a protein. Chem. Phys. Lett. **2013**, *581*, 91-95.

[62] Das, A. Chakrabarti, J. Ghosh, M. Thermodynamics of interfacial changes in a protein–protein complex. Molecular Biosyst. **2014**, *10*, 437-445.

[63] Sikdar, S. Chakrabarti, J. Ghosh, M. A microscopic insight from conformational thermodynamics to functional ligand binding in proteins. Molecular Biosyst. **2014**, *10*, 3280-3289.

[64] Kole, K. Chakrabarti. J. Binding of Homeodomain Proteins to DNA with Hoogsteen Base Pair. **2025**, *129*, 1544–1554.

[65] Chatterjee, P. Bagchi, S. Sengupta, N. The non-uniform early structural response of globular proteins to cold denaturing conditions: a case study with Yfh1. J. Chem. Phys. **2014**, *141*.

Chapter 6 Coarse-grained model of protein with structural information

6.1 Introduction

Coarse-grained (CG) models play increasingly important role in computational science to describe complex systems and are now becoming as essential as atomically detailed models.[1–5] By grouping of atoms into a unit, CG models capture targeted features of complex systems.[6] The main purpose of those studies is to reduce the number of degrees of freedom but keeping the important features of the system in tact. This provides significant computational and conceptual advantages compared to more detailed models, allowing the exploration of temporal and spatial evolutions of a system at much larger scale, far beyond the capabilities of traditional fully microscopic all atom (AA) models.[7]

A wide variety of CG models has been introduced to capture specific aspects for systems, including proteins,[8] nucleic acids,[8] lipid membranes,[9] carbohydrates,[10] water[11] and so on. Coarse-graining approaches range from often qualitative solvent-free models to more realistic explicit solvation models including chemical specificity.[7] Models of the latter category are typically parameterized by comparison with atomistic simulations.[7] The parameterization of coarse-grained (CG) models typically follows one of two main strategies: a bottom-up approach, in which the model is derived to reproduce microscopic features based on more detailed theoretical models such as atomistic or quantum mechanical simulations; or a top-down approach, where the model is built in such a way that it can reproduce a set of experimental macroscopic properties such as density and surface tensions.[12,13] The Martini force field is one such well-established CG force fields for biomolecular simulations,[10,14,15] like lipids and surfactant

molecules, cholesterol and all amino acids and nucleotides as well as for a variety of sugars, polymers and nanoparticles. MARTINI uses a bottom-up approach for parameterizing bonded interactions, while a top-down strategy is used for non-bonded interactions. SIRAH is another coarse-grained force field developed to simulate biological systems, using a bottom-up approach for its parameterization.[13] Different CG protein models usually adopt different levels of simplified polypeptide representation.[16] The main chain is represented by either all heavy atoms or one to two united atoms per residue, while the side chain is typically replaced by one or two united atoms.[16,17] In simple lattice protein-like HP models,[18–21] each amino acid is either represented as solvophobic or polar entities which ignore protein backbone information. This model is widely used to understand folding pathway of the protein by using Molecular Dynamics (MD) and/or Monte Carlo (MC) methods.[16]

It is known that protein structural and functional aspects could be understood in terms of protein dihedral angles. Many years ago, Goddard III introduced a coarse-grained (CG) protein model in which the protein structure is represented only by the Cα coordinates[22,23] and then using probability grid MC method to build first backbone conformation and then side chain. This method modifies protein conformation one residue at a time, by choosing either new backbone (ϕ and ψ) or side chain (χ) dihedral angles from probability matrices computed using the crystal structure data. However, they do not account for various bonded and non-bonded interactions in constructing the model. Various interaction models for CG representations have been reported. A more challenging approach involves 'physics-based' derivations of CG force fields, which originates from classical AA interaction models and are translated into united atom potentials.[16] The main drawback of such CG model is these kind of analysis that the models do not include the dihedral angles and hence, cannot provide any structural information.

There are many cases where the protein while in function may change conformations, for instance, in molten globule (MG) state[24,25] and intrinsically disordered protein (IDP).[26] In MG proteins, some regions of the protein lose secondary structure, keeping the overall structural integrity in tact and encountered often in extreme pH and temperature conditions. The protein in MG state only can bind to cytotoxic factor, like oleic acid.[24] In IDP some region of the molecule undergo changes in conformation. Both MG proteins and IDP are functionally active and have tendency to aggregate in solution phase[24] which are relevant to understand various pathological conditions.[24] In such proteins the structural aspects are extremely important to understand their behavior in solution and

their functions. The appropriate CG models to describe such protein need to include structural information via the backbone dihedral angles.

Motivated by this, we built a simple polymer model based on bottom-up approach to make a CG description of protein that which is capable of incorporating main chain backbone dihedral angles (ϕ and ψ). A bead represents an amino-acid and each bead is assigned with additional degrees of freedom corresponding to the backbone dihedral angles. We represent each water oxygen atom as a solvent bead. The bead radius is determined from the minimum of the free energy surface, which is based on the probability distribution of the radius of gyration (Rg) of all protein residues over an equilibrated trajectory. Rg is calculated as the square root of the mean distance of all heavy atoms from the center of mass of the residue. Different interactions are as follows: (1) bead-bead bonded interactions, like the bond stretching, bond bending and dihedral fluctuations; (2) bead-bead non-bonded interactions, like bead-bead repulsion and charge interactions; (3) bead-solvent interactions and (4) solvent-solvent interactions. Here, we follow the bead-spring CG model. All interaction parameters are derived from equilibrated AA MD simulation data to model CG interactions. We use the MC simulations based on the Metropolis sampling to generate conformations as per the energy costs of the model interactions. We consider protein GB3 which have 56 amino acid residues. GB3 is widely used as a model structure[27] in CG simulations due to its small size, well-defined structure and well-documented folding mechanism. Its fast and cooperative folding makes it an excellent benchmark for testing force fields and folding algorithms. Additionally, GB3 has been extensively studied using X-ray crystallography, NMR and single-molecule experiments, providing high-resolution structural and dynamic data. These experimental insights enable direct validation of CG models, ensuring their accuracy in capturing protein folding and stability. Our simulation study for GB3 shows good structural agreement with the crystal structure data and the AA MD simulation data. We also check whether one can reproduce structure of other protein based on our CG model generated for GB3. We test this on another small, well-defined structural protein, the homeodomain[28] and find similarity in structural elements with crystal data and AA data.

6.2 Methods

6.2.1 Systems

We study the following systems in both AA and CG simulations: (1) GB3 (PDB ID: 2OED),[29] a small globular protein consisting of 56 residues, (2) homeodomain protein, another small globular protein containing 58 residues, the initial structure is taken without DNA form the PDB structure 1K61.[30] The crystal structure snapshots corresponding to PDB IDs: 2OED and 1K61 are shown in Figs. 6.1(a) and (b), respectively. (3) α-synuclein molecules (residues: 1-95) (apo) and (4) α-synuclein molecule (residues: 1-95) with neutral ZnONP (holo-neutral). The details of systems 3 and 4 are discussed in an earlier Chapter, Chapter 5.

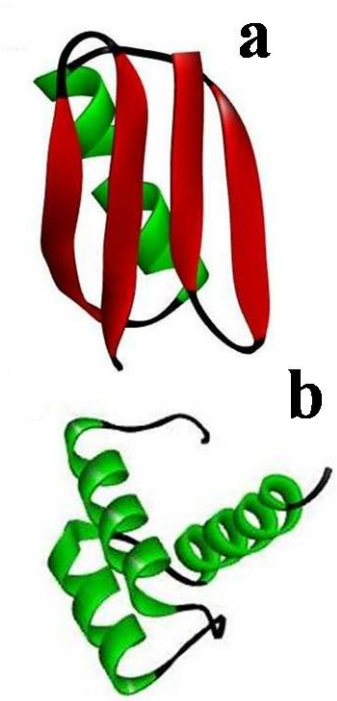


Figure 6.1: Snapshot of crystal structure: (a) GB3 (PDB id:2OED) and (b) homeodomain (PDB id: 1K61). Helix in green, sheet in red and loop in black.

6.3 All atom (AA) molecular dynamics (MD) simulations

The GROMACS[31] 2018.6 package[32] with Amber99sb force field (ff) (Appendix 2.5.2)[33] is used for our AA simulation. Leapfrog algorithm (Appendix 2.5.1) is

used to integrate the equations of motion. The TIP3P water model is used as the solvent. Periodic boundary conditions (Appendix 2.5.3) are applied in all three dimensions. The system is electrically neutralized by adding the require number of sodium ($Na+$) and chloride ($Cl-$) ions. The potential energy is minimized using the steepest descent algorithm.[34] Then AA MD simulation is performed at 300K temperature and 1 atmosphere pressure maintaining an isothermal-isobaric (NPT) ensemble. We use the Berendsen thermostat (Appendix 2.5.4)[35] to maintain temperature and the Parrinello-Rahman barostat (Appendix 2.5.5)[36] to maintain constant pressure. The Lennard-Jones (LJ) and short-range electrostatic interactions are terminated at 10 Å. We use the Particle-Mesh Ewald (PME)[37] method (Appendix 2.5.6) to compute the long-range electrostatic interactions. LINCS[38] constraints are applied to all bonds involving hydrogen atoms. We use 2 fs time step for integration. The equilibration of the system is confirmed by the saturation of the root mean square deviation (RMSD) with time. We consider the equilibrated part of the trajectory for further analysis.

6.4 Coarse-grained (CG) simulations

We build up the CG force field from the AA trajectory. We use GB3 as an illustrative case. We take a 100 mM NaCl solution to neutralize the system in AA simulation. The beads are represented by the center of mass (COM) of the amino acid residues in the crystal structure and the bead radius from the distribution of the radius of gyration of the heavier atoms in the residue. The bond stretching and bending interactions are calculated from the distributions of the bond distance between COM of two successive residues and the angle between bond vectors between COM of three successive residues. The dihedral interaction is calculated based on the coupling between dihedral angles of both intra- and inter-residue pairs within the protein. The bead-solvent interaction is computed via the distribution of water oxygen atoms around the COM of the residues. The solvent-solvent interaction is calculated via the radial distribution function between oxygen atoms of the water molecules. The model system is simulated using the MC Metropolis algorithm (detailed in Appendix 6.9.1). We perform 100,000 MC steps. Equilibration of the system is determined by its potential energy. We use equilibrated trajectories to calculate various quantities, such as solvent distribution around bead particles and secondary structure analysis. We run multiple independent trajectories with identical initial conditions to average the analysis data.

6.5 Analysis

We calculate the following quantities over equilibrated trajectories of the AA MD and CG MC simulations.

6.5.1 Solvent distribution function

We compute the solvent distribution function,[39] $\rho(r)$ to understand the arrangement of solvent beads around the solvophilic and solvophobic beads of the polymer. The $\rho(r)$ is computed using the following function:

$$\rho(r) = \frac{1}{\langle \rho_N \rangle} \frac{1}{N1} \sum_{i}^{N} \sum_{j}^{N1} \frac{\delta(r_{ij} - r)}{4\pi r^2}$$

Here, $\langle \rho_N \rangle$ represents averaged number density of solvent beads, N is the total number of solvent beads and $N1$ is the total number of solvophilic or solvophobic polymer beads.

6.5.2 Dihedral angle distribution

We compute the distribution of the phi (ϕ)[40] and psi (ψ)[38] dihedral angles per residue per frame using our in-house program.

6.5.3 Structural persistence

We calculate the structural persistence (S_P)[24,41] parameter per residue using the formula:

$$S_P = \frac{1}{N} \sum_{i=1}^{N} e^{(\Delta\phi_i/\Delta\phi_{max})} \cdot e^{(\Delta\psi_i/\Delta\psi_{max})}$$

Where, N denotes the total number of frames. $\Delta\phi_i$ and $\Delta\psi_i$ are the absolute values of the changes in dihedral angles ϕ and ψ of the residue in the frame i from the reference frame, $\Delta\phi_{max}$ and $\Delta\psi_{max}$ are the maximum alterations possible in the Ramachandran diagram. $S_P = 1$ indicates no conformational change, whereas low S_P represents greater deviation from the reference structure.

6.5.4 Ramachandran plot

We generate the Ramachandran plot[42] using the averaged ϕ and ψ dihedral angles over the equilibrated trajectory.

6.5.5 Dihedral principal component analysis

We perform principal component analysis on dihedral angles, ϕ and ψ using dPCA+ method.[24,25,43] Traditional principal component analysis based on dihedral angles often faces problems in accurately projecting data onto principal subspaces. In dPCA+, the residual projection error is minimized by transforming the data to shift the maximal gap of the sampling to the periodic boundary. In this method, compute a histogram of the data with a bin width of five degrees and select the bin center with the lowest population as the maximal gap. If multiple bins have equally low populations, sum the populations over their respective neighboring bins and select the one with lowest overall population in the neighborhood. The data is then transformed to reduce density at the periodic boundary. After successfully transforming the data into linear form, the covariance matrix is computed and eigen decomposition is performed in the standard way.

6.5.6 Conformational thermodynamics

We calculate the changes in conformational free energy and entropy of the aggregation-prone NACore region (68GAVVTGVTAVA78) of αS in presence of ZnONP with respect to its absence using the conformational thermodynamics approach described in Chapter 3. Here, dihedral angles ϕ and ψ are used as the microscopic conformational variables.

6.6 Results

6.6.1 Building of the CG force field parameters

We take the bottom-up approach, namely, use the AA data to build up the CG model. We illustrate this for a small protein GB3. The RMSD over the course of the AA simulations, shown in Fig. 6.2 (a). The data shows equilibration of the system. An equilibrium snapshot of the GB3 protein is shown in Fig. 6.2 (b). We use the equilibrated trajectories to compute different quantities. We compute H(Rg), the probability distribution of the radius of gyration Rg of all the protein residues based on the distances of their heavy atoms from the COM at the given temperature T. We assign energy corresponding to the distribution, $F = -RT \ln H(Rg)$, as shown in Fig. 6.3 (a), where, R is the ideal gas constant. We also compute the distribution H(d) of the bond distance, d

between center of mass of two consecutive amino acid residues. The stretching energy, $F = -RT \ln H(d)$, shown in Fig. 6.3 (b). Fig. 6.3 (c) shows the bending energy, $F = -RT \ln H(\theta)$, where $H(\theta)$ represents the distribution of the bond-angle between COM of three consecutive residues in AA data. We also compute the distribution of water oxygen atoms $\rho(r)$ where r is the distance of oxygen and the COM of the residue. $\rho(r)$ data are averaged over all the solvophilic and solvophobic residues separately.

The data -$RT \ln(\rho(r))$ versus r for these two cases are shown in Figs. 6.3 (d) and 6.3 (e), respectively. The oxygen-oxygen pair correlation function is also calculated and the -$RT \ln(g(r))$ versus r plot is shown in Fig. 6.3(f).

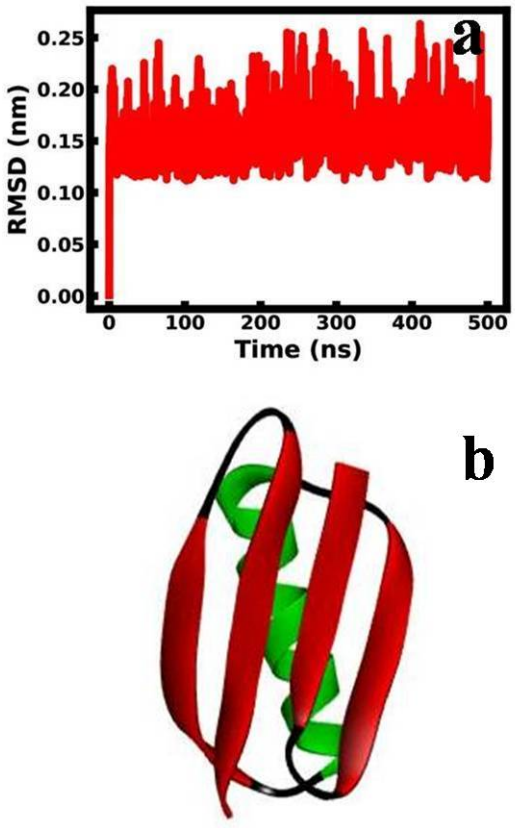


Figure 6.2: (a) RMSD over the course of the simulation and (b) an equilibrated snapshot of the protein GB3 at 500 ns time span in AA simulation. Helix in green, sheet in red and coil in black.

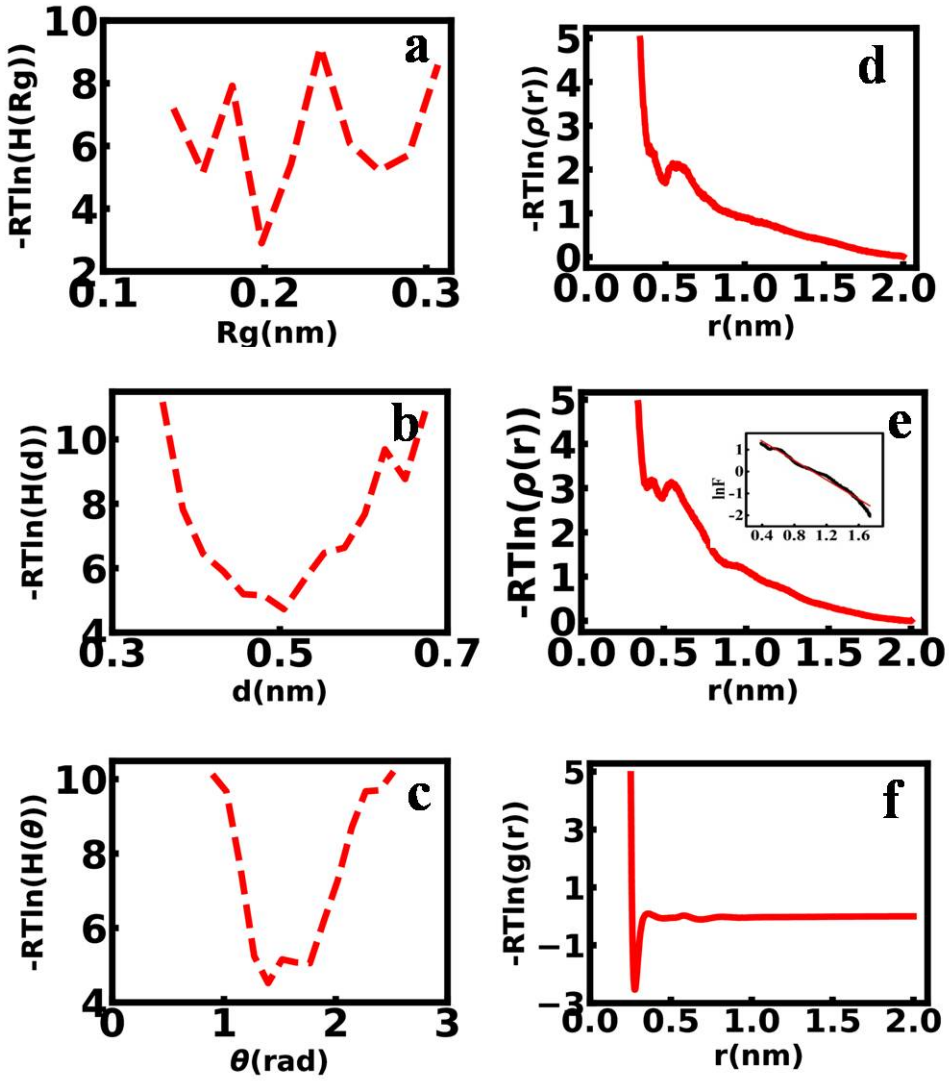


Figure 6.3: (a) Free energy profile, $-RTlnH(R_g)$ vs radius of gyration of polymer bead, R_g, (b) Free energy profile, $-RTlnH(d)$ vs distance, d between two consecutive polymer beads, (c) Free energy profile, $-RTlnH(\theta)$ vs angle, θ between three consecutive polymer beads, (d) Solvent density-derived free energy profile, -RTln $\rho(r)$ vs distance r, for solvent, water around hydrophilic residues, (e) Solvent density-derived free energy profile, $-RTln\rho(r)$ vs distance r, for solvent, water around hydrophobic residues, inset: Logarithmic free energy, $\ln F = -RT \ln \rho(r)$, as a function of r for water around hydrophobic residues and (f) Free energy profile, $-RTlng(r)$, as a function of the radial distribution function, $g(r)$, of solvent, water molecules derived from AA MD simulations data of GB3 protein.

CG model beads are arranged with the same primary sequence as the wild-type protein, the center of mass of a residue being the center of the CG model protein bead. We assign the bead radius (r_b) as 0.2 nm, corresponding to the minimum of the energy $F = -RT \ln H(R_g)$ (Fig. 6.3a)). The bead diameter (σ_b) is

0.4 nm. The model force field parameters are estimated as described below:

I. Harmonic bond stretching potential: The harmonic bond stretching potential between two polymer beads i-j is modeled as: $V_{bond}(r_{ij}) = \frac{1}{2}k_b(r_{ij} - b)^2$, where r_{ij} is the bond distance between two atoms i and j, b is the equilibrium bond distance given by the minimum of $F = -RT \ln H(d)$, shown in Fig. 6.3 (b). The stretching force constant k_b is calculated from the second-order derivative around the minimum of the stretching energy. Here, we find $b = 0.5$ nm and $k_b = 2292$ kJ/mol/nm^2.

II. Bending potential: The bending potential between three consecutive polymer beads i-j-k is represented by a harmonic potential: $V_{bend} = \frac{1}{2}k_a(\theta - \theta_0)^2$, The equilibrium bond angle, θ_0, is set to 1.4 rad where F has a minimum. The bending force constant, k_a (87 kJ/mol/rad^2), is calculated from the second-order derivative around the minimum of $-RT \ln H(\theta)$ vs θ in Fig. 6.3 (c).

III. Non-bonded interaction: Non-bonded interactions between polymer beads are of two types: (i) Repulsive part of the Lennard-Jones (LJ), $V^{bb}_{rep}(r_{ij})$ (= 4 $\epsilon_{bb}\left(\frac{\sigma}{r_{ij}}\right)^{12}$), so that the beads cannot overlap. We set the ϵ_{bb} value to 2.5 kJ/mol. (ii) Screened Coulomb interactions, $V_{SC}(r_{ij})$ (= $\frac{q_i q_j}{4\pi D\epsilon_0 r_{ij}} \exp\left(-\frac{r_{ij}}{\lambda}\right)$), where q_i and q_j are the net charges of the residues i and j, respectively, as calculated using the pdb2pqr server (`https://server.poissonboltzmann.org/pdb2pqr`) by uploading the PDB file and specifying the desired pH value. ϵ_0 is the permittivity of vacuum and D is the dielectric constant of water (set equal to 80).

We set the Debye screening length, λ (= $\left(\frac{D\epsilon_0 k_B T}{\sum_i z_i^2 c_i e^2 N_A}\right)^{\frac{1}{2}}$) equal to 0.03 nm computed from the AA data. Here, e is the elementary charge, z_i is the valency of each type of ion i present in the solution, and c_i is the ionic concentration in mol/L.

IV. Bead-solvent interaction: We treat the oxygen atom of the water as a solvent bead. We take polymer beads of two types: solvophilic, which attract the model solvent, and solvophobic, which repel the model solvent.

The solvophilic beads interact via the Lennard-Jones (LJ) 12-6 potential with the solvent bead: $V^{LJ}_{sopl\text{-}solv}(r) = 4\ \epsilon_{sb}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$. Here, r is the distance between the polymer bead and the solvent bead. The interaction parameter ϵ_{sb} is equal to 1.71 kJ/mol, which corresponds to the minimum of the energy, $F = -RT \ln \rho(r)$ (Fig. 6.3 (d)).

The solvophobic beads interact with the solvent via the potential: $V_{sopb\text{-}solv}(r) = 13.0 \exp(-2.4r)$, as shown in the inset of Fig. 6.3 (e). This is computed by fitting the semilog plot of the free energy, $F = -RT \ln \rho(r)$ vs. r, as shown in Fig. 6.3 (e).

V. Solvent-solvent bead interactions: For solvent-solvent beads, only the non-

bonded Lennard-Jones (LJ) 12-6 potential is considered: $V^{LJ}_{solv\text{-}solv}(r) = 4\,\epsilon_{ss}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$. The interaction parameter ϵ_{ss}, equal to 2.5 kJ/mol, corresponds to the minimum value of the energy, $F = -RT\ln g(r)$ (Fig. 6.3(f)).

The bead-solvent and solvent-solvent interactions are truncated at a cutoff distance of 0.625 nm.

6.6.2 Dihedral interaction parameters

All-atom MD simulation trajectories are used to generate the free energy profile of the conformational variables ϕ_i and ψ_i dihedral angles of the i^{th} residue of the protein.[44] The ϕ_i and ψ_i dihedral angles are calculated from the atomic coordinates of the residues. The intra-residue coupling is quantified by calculating the joint probability distribution $P(\Gamma_i, \Gamma_j)$, which describes the probabilities of the simultaneous occurrence of Γ_i and Γ_j for the i^{th} and j^{th} residue. We obtain the free energy profile as: $G(\Gamma_i, \Gamma_j) = -k_B T \ln P(\Gamma_i, \Gamma_j)$.

We consider data based on the hydrophilic or hydrophobic properties of the residues in neutral pH conditions, rather than considering each individual residue pair. We consider only backbone dihedral angles so that $\Gamma_i = (\phi_i, \psi_i)$ for the i^{th} residue.

Let us first consider the intra-residual ($i = j$) dihedral coupling. We group the data into two distinct energy profiles: one, considering all the hydrophilic amino acid residues, and the other, considering all the hydrophobic residues. We construct a two-dimensional grid of data, shown in Figs. 6.4 (a) and (b), respectively.

A similar method is used to calculate the free energy profile for inter-residual ($i \neq j$) dihedral coupling. The dihedral angles of the corresponding residues are subsequently classified according to the hydrophilic-hydrophilic, hydrophobic-hydrophobic, and hydrophobic-hydrophilic residue pairs. For each case, all possible dihedral angle combinations are considered and the negative logarithm of each joint probability distribution yields the corresponding free energy profile. The free energy landscape (FEL) plot for various combinations of dihedral angles of two different residues is shown in Figs. 6.5 (a)-(l).

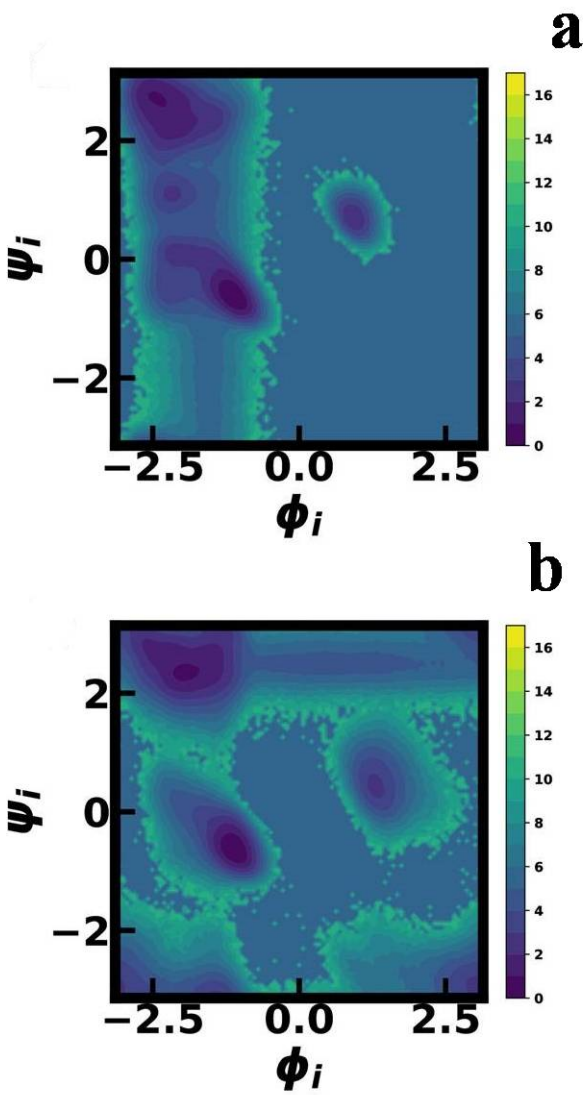


Figure 6.4: Free energy landscape for intra residual dihedral coupling, considering all (a) hydropjhilic residues and (b) hydrophobic residues.

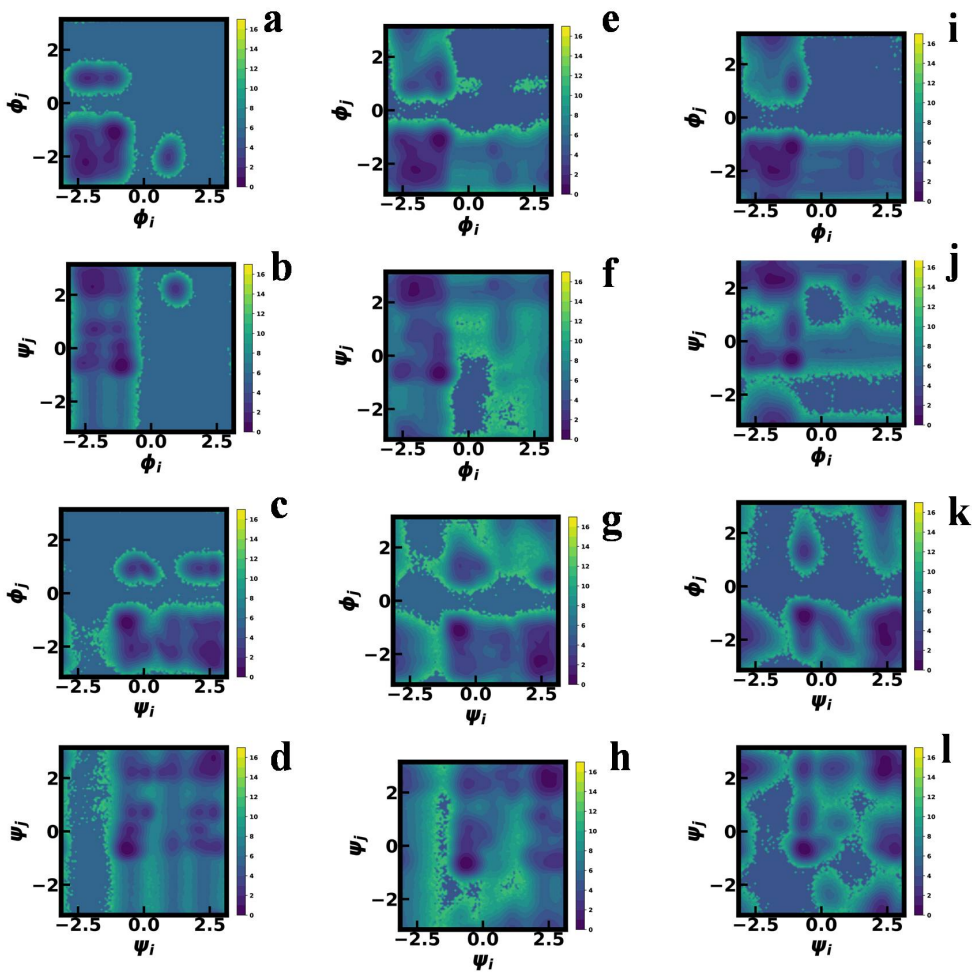


Figure 6.5: Free energy landscape for inter residual dihedral coupling considering (a-d) hydropjhilic-hydropjhilic residues, (e-h) hydropjhilic-hydrophobic/hydrophobic-hydropjhilic residues and (i-l) hydrophobic-hydrophobic residues.

6.6.3 Monte Carlo for CG model of GB3

We define the sequence of polymer model beads as the same amino-acid sequence as in the AA and crystal structure. The beads are taken into two classes, namely, solvophilic and solvophobic which are respectively the hydrophilic and hydrophobic residues in the primary sequence. Table 6.1 shows the hydrophilic and hydrophobic classification of GB3 residues. We set the diameter of the solvent bead as that of oxygen atom (σ_s) to be 0.25 nm. We consider σ_s as the unit length and thermal energy k_B T as the unit energy in our simulations. We fix the solvent bead density at $\sim$ 1 gm/cm^3 .

We take all the parameters in the model Hamiltonian as determined from the GB3 AA trajectory. The Hamiltonian takes the following form:

Res	Type	Res	Type
MET1	hydrophobic	ALA29	hydrophobic
GLN2	hydrophilic	PHE30	hydrophobic
TYR3	hydrophobic	LYS31	hydrophilic
LYS4	hydrophilic	GLN32	hydrophilic
LEU5	hydrophobic	TYR33	hydrophobic
VAL6	hydrophobic	ALA34	hydrophobic
ILE7	hydrophobic	ASN35	hydrophilic
ASN8	hydrophilic	ASP36	hydrophilic
GLY9	hydrophobic	ASN37	hydrophilic
LYS10	hydrophilic	GLY38	hydrophobic
THR11	hydrophilic	VAL39	hydrophobic
LEU12	hydrophobic	ASP40	hydrophilic
LYS13	hydrophilic	GLY41	hydrophobic
GLY14	hydrophobic	VAL42	hydrophobic
GLU15	hydrophilic	TRP43	hydrophobic
THR16	hydrophilic	THR44	hydrophilic
THR17	hydrophilic	TYR45	hydrophobic
THR18	hydrophilic	ASP46	hydrophilic
LYS19	hydrophilic	ASP47	hydrophilic
ALA20	hydrophobic	ALA48	hydrophobic
VAL21	hydrophobic	THR49	hydrophilic
ASP22	hydrophilic	LYS50	hydrophilic
ALA23	hydrophobic	THR51	hydrophilic
GLU24	hydrophilic	PHE52	hydrophobic
THR25	hydrophilic	THR53	hydrophilic
ALA26	hydrophobic	VAL54	hydrophobic
GLU27	hydrophilic	THR55	hydrophilic
LYS28	hydrophilic	GLU56	hydrophilic

Table 6.1: Classification of GB3 residues as hydrophilic or hydrophobic in the crystal structure.

$$H = \sum_{ij}{}' V_{\text{bond}}(r_{ij}) + \sum_{ijk}{}'' V_{\text{bend}}(\theta_{ijk}) + \sum_{ij} V_{\text{rep}}(r_{ij}) + \sum_{i}{}'(\sum_{j} V_{\text{SC}}(r_{ij}) + \sum_{j} V_{\text{dih}}(\Gamma_i, \Gamma_j)) + \sum_{\alpha} V_{\text{LJ}}^{\text{bead-solv}}(|\vec{r}_i - \vec{R}_\alpha|) + \sum_{\alpha,\beta}{}' V_{\text{LJ}}^{\text{solv-solv}}(|\vec{R}_\alpha - \vec{R}_\beta|) \tag{6.1}$$

Here, $r_{ij} = |\vec{r}_i - \vec{r}_j|$ is the distance between two polymer beads with coordinates $\vec{r}_i$ and $\vec{r}_j$, and Γ_i and Γ_j are the dihedral angles associated with the i$^{\text{th}}$ and j$^{\text{th}}$ bead, respectively. $\vec{R}_\alpha$ is the coordinate of the α^{th} solvent bead. The

prime over the bonding term indicates that only two consecutive residues COM and the double prime over the bending energy indicates that three consecutive residues are to be taken. The bead-solvent interaction is taken as per solvophilic or solvophobic bead type. The prime over the solvent-solvent interaction indicates the case with α = β to be excluded.

We use the MC method to generate conformations based on model interactions maintaining the canonical (NVT) ensemble, where each bead is assigned three position variables and two dihedral angles. We set the system temperature to room temperature (298 K). The initial dihedral angles of each bead are taken from the protein crystal structure data. MC simulations are carried out in a cubic box of length, L=5.6 nm with periodic boundary conditions in all directions (x,y and z). When we construct the MC moves, we compute the energy cost arising from random particle displacements and changes in the dihedral angles using the Hamiltonian H defined above. The intra-bead (i=j) interaction between dihedral of the beads are calculated by linear interpolation from the free energy profile tabulated in two-dimensional grids. We choose the type of grid depending on if the bead is solvophilic or solvophobic. The inter-bead (i $\neq$ j) dihedral coupling depends on the solvophobic or solvophilic character of the pair of beads. All the non-bonded interactions, including the bead-solvent and solvent-solvent interactions are truncated at a cutoff distance of 0.625 nm. For the inter-bead dihedral contribution, we consider only those beads which are within purpose, a 7Å around each bead in every frame. All the parameters used in CG simulation are tabulated in Table 6.2.

Parameter	**Value**
σ_b	0.4 nm
σ_s	0.25 nm
b	0.5 nm
k_b	2292 kJ/mol/nm^2
θ_0	1.5 rad
k_a	87 kJ/mol/rad^2
ϵ_{bb}	2.5 kJ/mol
ϵ_{sb}	1.25 kJ/mol
ϵ_{ss}	2.5 kJ/mol

Table 6.2: Table of parameters and their values.

We calculate the solvent bead distributions around solvophilic and solvophobic polymer beads (Fig. 6.6 (a)) and compare then with AA simulation data. Fig. 6.6 (a) shows that the first peak in the solvent distribution is higher

around solvophilic beads compared to solvophobic beads, consistent with AA simulation data, shown inset figure of Fig. 6.6 (a). We further compute S_P from the equilibrated CG MC and AA MD trajectories, as shown in Fig. 6.6 (b). The CG and AA simulation data for S_P exhibit good agreement.

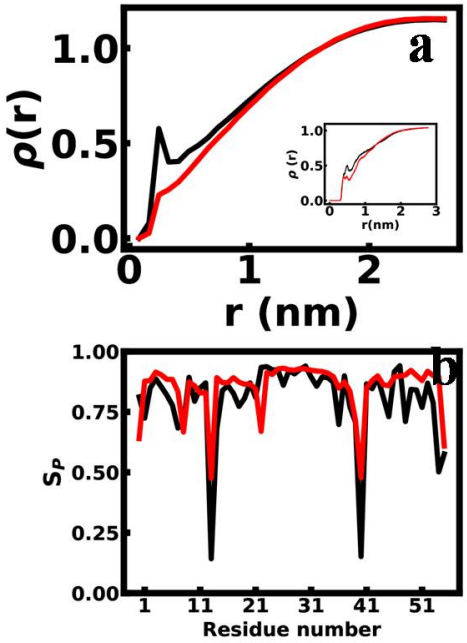


Figure 6.6: (a) Solvent distribution around solvophilic (black) and solvophobic (red) beads in the equilibrated CG MC trajectory of GB3. The inset shows the solvent distribution around hydrophilic (black) and hydrophobic (red) residues in the equilibrated AA MD trajectory and (b) Structural persistence (S_P) per residue in equilibrated CG simulations (black) and AA MD simulations (red).

We compare the secondary structural elements of each residue from the CG MC simulation with AA MD data and the initial crystal structure. We compute the secondary structure preferences of each residue over the equilibrated CG trajectory and over the equilibrated AA MD trajectory computing the ϕ and ψ dihedral angle distributions. Table 6.3 shows the structural preferences of each residue over equilibrated CG and AA data, along with their secondary structure in the crystal structure. We classify secondary structures as helix (H), β-sheet (S) and unstructured (U) based on the dihedral angles, ϕ and ψ, within known ranges. The unstructured (U) category includes all conformations that do not

fall into the helix or β-sheet. We find that secondary structural elements remain unchanged for 82% residues. We also observe our CG MC simulation more accurately captures helical (H) secondary structure elements (93.8%) compared

Res	Crystal Structure	AA Structure (MD)	CG Structure (MC)	Res	Crystal Structure	AA Structure (MD)	CG Structure (MC)
MET1	U	S	U	ALA29	H	H	H
GLN2	S	S	S	PHE30	H	H	H
TYR3	S	S	S	LYS31	H	H	H
LYS4	S	S	S	GLN32	H	H	H
LEU5	S	S	S	TYR33	H	H	H
VAL6	S	S	S	ALA34	H	H	H
ILE7	S	S	S	ASN35	H	H	H
ASN8	S	S	S	ASP36	H	H	H
GLY9	U	U	S	ASN37	H	H	H
LYS10	U	U	U	GLY38	U	U	U
THR11	U	U	U	VAL39	U	U	S
LEU12	U	U	S	ASP40	U	U	S
LYS13	S	S	S	GLY41	U	U	S
GLY14	S	S	U	VAL42	S	S	S
GLU15	S	S	S	TRP43	S	S	S
THR16	S	S	S	THR44	S	S	S
THR17	S	S	S	TYR45	S	S	S
THR18	S	S	S	ASP46	S	S	S
LYS19	S	S	S	ASP47	U	U	U
ALA20	U	S	S	ALA48	U	U	U
VAL21	U	U	U	THR49	U	U	U
ASP22	H	U	S	LYS50	U	S	U
ALA23	H	H	H	THR51	S	S	S
GLU24	H	H	H	PHE52	S	S	S
THR25	H	H	H	THR53	S	S	S
ALA26	H	H	H	VAL54	S	S	S
GLU27	H	H	H	THR55	S	S	U
LYS28	H	H	H	GLU56	S	S	U

Table 6.3: Comparison of secondary structure preferences in GB3 protein. MD denotes trajectory of all atom molecular dynamics simulations and MC denotes conformations based on monte carlo simulations. 'H' corresponds Helix, 'S' corresponds sheet and 'U' corresponds to other than element helix or sheet i.e. loop/coil/turn/bend region of the protein.

to β-sheet (S) (88%) and unstructured (U) (60%) regions, relative to the crystal structure. Next, we compare the secondary structural elements obtained from the MC simulation with those from the crystal structure. We observe that the secondary structure remains same for approximately 80% of residues.

We further analyze the ϕ-ψ Ramachandran (RC) plot to compare the average structure from CG simulations with the crystal structure and the average structure from AA simulations of GB3 protein. The RC plot of the crystal structure and the average structures obtained from AA and MC simulations is shown in Fig. 6.7 (a). The secondary structure is consistent with both the crystal structure and

the average structure obtained from AA MD simulations of the GB3 protein, as shown in Fig. 6.7(a).

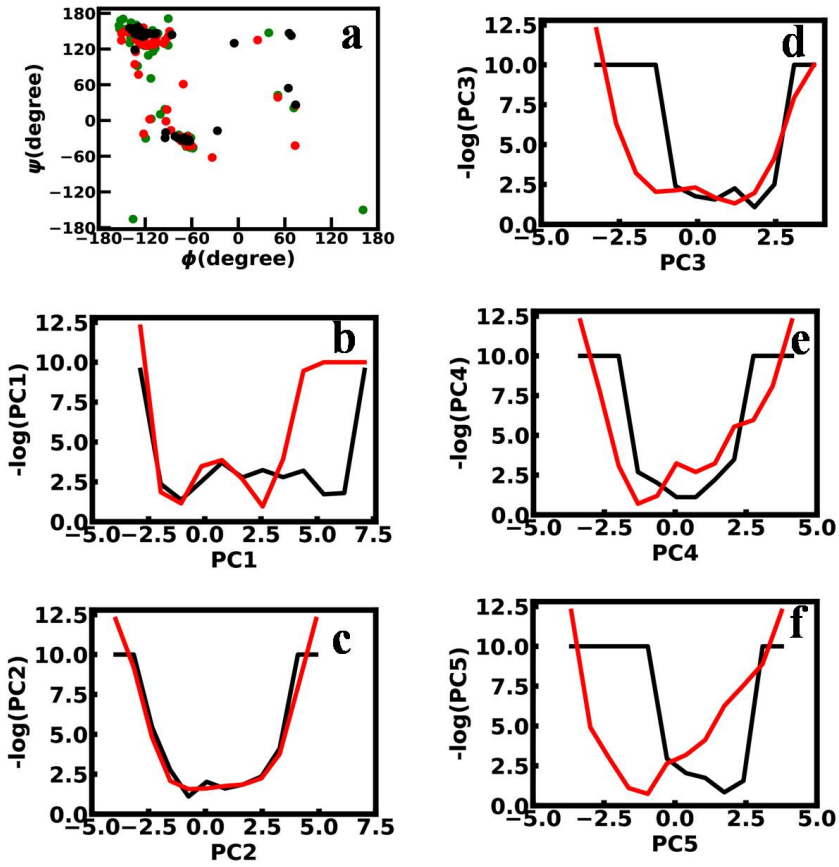


Figure 6.7: (a) Comparison of Ramachandran plot for different structure obtained from crystal structure (green), average structure based on AA MD simulations (red) and average structure (black) based on CG MC simulations for protein GB3. FEL obtained from dPCA+ along (b) PC1, (c) PC2, (d) PC3, (e) PC4 and (f) PC5 over equilibrated CG MC (black) and AA MD (red) data of GB3. Y-axis represents the negative log of population of PCs.

Since the RC plot is based on the mean values of the dihedral angles, we examine their fluctuations over the equilibrium MC conformations. We perform to this end the dihedral principal component (PC) analysis using the dPCA+ method, based on the dihedral angles ϕ_i and ψ_i for each residue over the equilibrated CG data and compare with AA data. One dimensional FEL for PC1-PC5 are shown in Fig. 6.7 (b-f). We find that for PC1-PC4 FEL of the CG data agree well with the AA data.

6.6.4 Transferability of the coarse-grained model to other proteins

Next, we check the transferability of the CG model parameters by applying to the homeodomain protein. First, the CG model of the homeodomain protein is constructed based on its amino acid sequence, following the solvophobic and solvophilic classification like GB3. The bonded, non-bonded, and dihedral interactions are the same as those in the GB3 protein. We compare the solvent distributions around solvophilic and solvophobic beads in the equilibrated CG MC simulation (Fig. 6.8 (a)) with the corresponding AA data for the homeodomain protein, shown in the inset of Fig. 6.8 (a). The CG data agrees well to the AA data. Table 6.4 shows the secondary structural elements of each residue for the crystal structure, equilibrated AA MD, and CG MC simulations of homeodomain protein. In the CG MC simulation, 74% of the residues preserve their secondary structural elements from the initial crystal structure, as shown in Table 6.4. Here, also we observe that our CG MC simulation more accurately captures helical (H) secondary structure elements (97.5%) compared unstructured (U) regions (22%) of the homeodomain, relative to the crystal structure. However, the secondary structure of 76% residues remain consistent with the AA MD data as shown in Table 6.4. Fig. 6.8 (b) shows comparison of RC plot for homeodomain. RC plot for homeodomain protein suggests that average structure obtained from CG MC is comparable with crystal structure as well as average structure generated from equilibared AA MD simulations. Thus, the force field parameters generated for GB3 can also be used to generate conformations for other proteins, such as homeodomains, which exhibit high structural similarity with their crystal structures and AA descriptions. It highlights the transferability of model effectively.

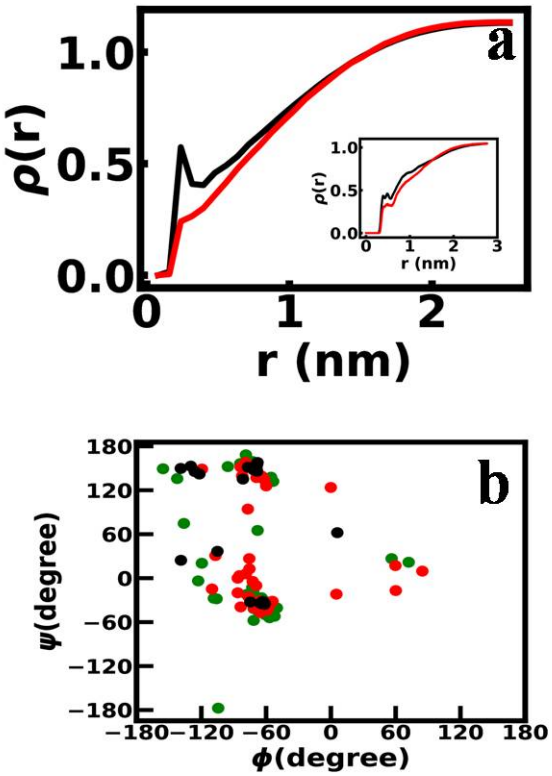


Figure 6.8: (a) Solvent distribution around solvophilic (black) and solvophobic (red) beads in the equilibrated CG MC trajectory of homeodomain protein. The inset shows the solvent distribution around hydrophilic (black) and hydrophobic (red) residues in the equilibrated AA MD trajectory and (b) Comparison of Ramachandran plot for different structure obtained from crystal structure (green), average structure based on AA MD simulations (red) and average structure (black) based on CG MC simulations for protein homeodomain.

Res	Crystal Structure	AA Structure (MD)	CG Structure (MC)	Res	Crystal Structure	AA Structure (MD)	CG Structure (MC)
ARG1	U	U	U	GLY30	H	H	U
GLY2	U	U	U	LEU31	H	H	H
HIS3	U	U	U	GLU32	H	H	H
ARG4	U	H	U	ASN33	H	H	H
PHE5	U	U	U	LEU34	H	H	H
THR6	U	U	U	MET35	H	H	H
ALA7	H	H	H	LYS36	H	H	H
GLU8	H	H	H	ASN37	H	H	H
ASN9	H	H	H	THR38	H	H	H
VAL10	H	H	H	SER39	U	U	H
ARG11	H	H	H	LEU40	U	U	U
ILE12	H	H	H	SER41	U	U	U
LEU13	H	H	H	ARG42	H	H	H
GLU14	H	H	H	ILE43	H	H	H
SER15	H	H	H	GLN44	H	H	H
TRP16	H	H	H	ILE45	H	H	H
PHE17	H	H	H	LYS46	H	H	H
ALA18	H	H	H	ASN47	H	H	H
ALA19	H	H	H	TRP48	H	H	H
ASN20	H	H	H	VAL49	H	H	H
ILE21	U	H	H	SER50	H	H	H
ALA22	U	H	H	ASN51	H	H	H
ASN23	U	U	U	ARG52	H	H	H
PRO24	U	U	U	ARG53	H	U	H
TYR25	U	U	U	ARG54	H	U	H
LEU26	U	U	U	LYS55	H	U	H
ASP27	U	U	U	GLU56	H	U	H
THR28	H	U	H	ALA57	U	U	H
LYS29	H	U	H	ALA58	U	U	H

Table 6.4: Comparison of secondary structure preferences for homeodomain protein. MD denotes trajectory of all atom molecular dynamics simulations and MC denotes conformations based on monte carlo simulations. 'H' corresponds Helix, 'S' corresponds sheet and 'U' corresponds to other than element helix or sheet i.e. loop/coil/turn/bend region of the protein.

6.6.5 CG model for αS

We model the αS molecule in its apo form in one case, and in another, its interaction with neutral ZnONP molecules in the holo-neutral form for CG simulations. We consider the similar types of bonded, non-bonded and dihedral coupling interactions like GB3 and homeodomain proteins simulations. We used the same polymer and solvent bead diameters as in the CG simulations of the GB3 protein. We compute the bonding interaction parameters from the AA equilibrated trajectories of αS. We get a lower bond stretching force constant (K_b = 748 kJ/mol·nm^2) and a slightly higher equilibrium bond length (b = 0.5 nm) compared to the values used in the GB3 simulation. We also get a lower angle force constant (K_a = 30 kJ/mol·rad^2) and a higher equilibrium angle (θ_0 = 1.6 rad) using a similar approach as for the GB3 protein. We simulate the CG model of the holo-neutral system where the interface-forming residues identified from the AA simulation (as detailed in the previous Chapter) are held fixed in both their positions and dihedral angels values. We then use the equilibrated MC trajectories for further analysis. We calculate the conformational thermodynamics of NACore residues in the holo-neutral system with respect to the apo system and compare the results with AA data. We observe that most of the NACore residues become stabilized and ordered in the presence of ZnONP, as shown in Fig. 6.9(a) and (b), where they are colored in green. The conformational free energy and entropy data of most of the residues qualitatively match with the AA data, shown in Fig. 6.9 (c) and (d). This indicates that our simplified CG approach captures the AA simulation data of IDP protein αS also reasonably well.

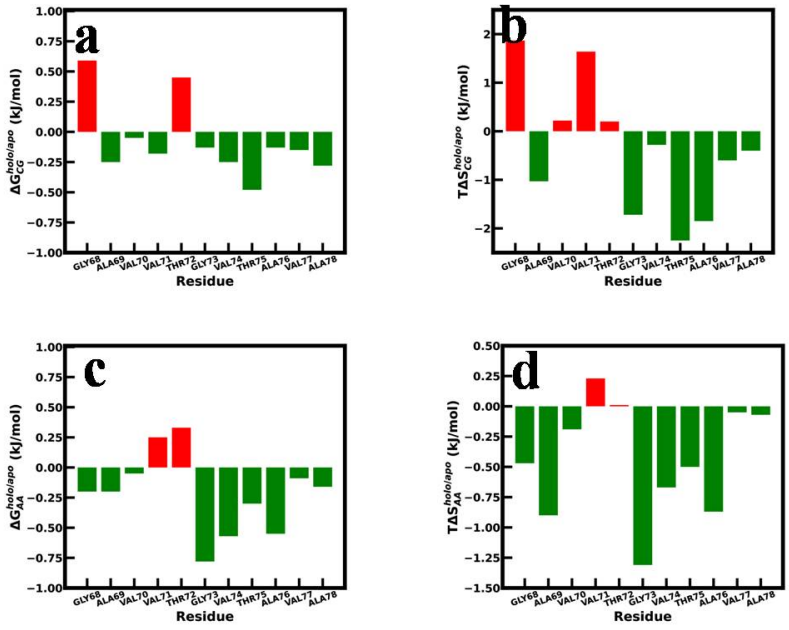


Figure 6.9: Conformational (a) free energy and (b) entropy of NACore residues in the holo-neutral system with respect to the apo system, computed from CG MC simulation data. Similarly, conformational (c) free energy and (d) entropy computed from AA MD simulation data.

6.7 Discussion

We accurately capture the structural information of protein residues using our CG model, in which each residue is represented by a single bead. In contrast, the widely used MARTINI model employs a four-to-one mapping scheme, where, on average, four heavy atoms are represented by a single interaction center, except for ring-like structures. It categorizes interaction sites into four main types: polar (P), nonpolar (N), apolar (C), and charged (Q). Within these types, subtypes are further distinguished based on hydrogen-bonding (d=donor, a=acceptor, da=both, 0=none) by a numerical value indicating polarity (ranging from 1 for lower polarity to 5 for higher polarity).[45] The most significant difference between our model and the MARTINI model is the way dihedral angles are accounted for. All the bonded and non-bonded interactions are treated using almost similar approach. The MARTINI model considers dihedral angles only when all four interacting beads maintain the same secondary structure, either helix or extended.[45] As a result, conformational transitions of protein secondary structures are not adequately captured.

In another widely used CG model, SIRAH,[46] the coarse-grained mapping of proteins is based on the positions of heavy atoms namely nitrogen (N), alpha carbon (Cα), and oxygen (O) with side chains represented using multiple interaction sites. Equilibrium distances and angles for bonded parameters between all backbone beads are adopted from the minimum energy conformation of fine-grained glycine tripeptides using the AMBER99 force field.[46] Φ, Ψ and Ω dihedral angles are considered to define the secondary structure. By polynomial fitting these dihedral angles forcing the existence of minima in the two more stable conformations, α-helices (Ψ = -57° and Ω = -47°) and β-strands (Ψ = 150° and Ω = -80°).

In contrast of these, our CG model adopts a simpler, single-bead representation, where each amino acid is modeled as a single interaction site centered at its COM. hese beads are connected through bond-stretching and angle-bending interactions, while nonbonded interactions are modeled using LJ and screened Coulomb potentials, similar to those employed in widely used CG models such as MARTINI and SIRAH. Our model is simulated in explicit solvent, incorporating polymer–solvent interactions. We explicitly incorporate backbone dihedral angles, specifically the ϕ and ψ angles of each amino acid bead and update them during simulations. As a result, our bead-spring simple CG polymer model, which uses a single bead per protein residue, more accurately captures protein secondary structures and shows good agreement with AA MD data.

Our CG model, based on GB3 protein parameters, is effectively transferable to other proteins. This transferability arises due to a couple of reasons: (I) The backbone interaction parameters are not so sensitive and given primarily via the heavy atoms like carbon, nitrogen and oxygen in all the proteins. (II) The protein conformation is driven by the hydrophilic and hydrophobic nature of the side chain and not so much on the actual chemistry. The model as it stands, however, does not consider the side chain dihedrals and hence, can't account for binding with ligands which are primarily via side chains.

6.8 Conclusion

In conclusion, we build up a CG representation of protein where the center of mass of amino acids is treated as polymer beads with dihedral angles as additional degrees of freedom. We treat the oxygen atom of the water molecules as the solvent beads. We observe structural similarity for most protein residues with the crystal structure and all atom descriptions. The transferability of the

model is validated by applying all parameters and the FEL obtained from AA MD data of the GB3 protein to other proteins, such as the homeodomain. Our simple CG model also reproduces AA simulations data of IDP protein αS also. Therefore, this simplified CG approach can be used to simulate long time scales of αS molecules, enabling the investigation of aggregation phenomena between αS molecules. This can open the way for the CG simulation of protein with structural information and may be useful for studying phenomena involving multiple protein molecules, such as protein aggregation for other IDP protein also.

6.9 Appendix

6.9.1 Monte Carlo (MC) simulation

MC[44] is a popular method of molecular simulation which is used to obtain consequence of stochastic process using random number generation and probabilistic statistics. The method of simulation is closely related to random experiment where outcome is not known apriori. In this method random walk algorithm is used to perform equilibrium sampling over the statistical ensemble. For a system of N particles, the partition function Q is defined as:

$$Q = c \int H(r^N, p^N) \exp\left(-\frac{H(r^N, p^N)}{k_B T}\right) dp^N \, dr^N \tag{6.2}$$

where q and p signify the coordinates and momenta of each particle. The Hamiltonian $H(r^N, p^N)$ represents the total energy of the system, k_B is Boltzmann's constant, and T is the temperature. The pre-factor c is defined as:

$$c = \frac{h^{3N}}{N!} \tag{6.3}$$

The expectation value of any variable A is defined as:

$$\langle A \rangle = \frac{\int dp^N dr^N A(r^N, p^N) \exp\left(-\frac{H(r^N, p^N)}{k_B T}\right)}{\int dp^N dr^N \exp\left(-\frac{H(r^N, p^N)}{k_B T}\right)} \tag{6.4}$$

Since kinetic energy depends solely on momentum, the above equation can be solved analytically only for the momentum term. Instead of directly

computing the integral

$$\int dr^N A(r^N) \exp\left(-\frac{u(r^N)}{k_B T}\right) \tag{6.5}$$

it is computed as:

$$\frac{\int dr^N A(r^N) \exp\left(-\frac{u(r^N)}{k_B T}\right)}{\int dr^N \exp\left(-\frac{u(r^N)}{k_B T}\right)} \tag{6.6}$$

The term

$$P(r^N) = \frac{\exp\left(-\frac{u(r^N)}{k_B T}\right)}{\int dr^N \exp\left(-\frac{u(r^N)}{k_B T}\right)} \tag{6.7}$$

is the probability density of finding the system in a configuration space around r^N. Here, the relative probability of visiting different points of the configuration space is considered rather than the absolute probability.

Sampling around probable configuration space is performed using the Metropolis algorithm. First, a random particle at a random position is chosen, and the corresponding potential energy $u(r^N)$ is computed. A random displacement Δ is then applied so that the new position becomes $r' = r + \Delta$, with new potential energy $u(r'^N)$. If the change in energy $\Delta U = u(r'^N) - u(r^N)$ is negative, the new configuration is accepted.

The probability of finding the particle in the new position is:

$$p_n \propto \exp\left(-\frac{u(r'^N)}{k_B T}\right) \tag{6.8}$$

and in the old position:

$$p_o \propto \exp\left(-\frac{u(r^N)}{k_B T}\right) \tag{6.9}$$

If the ratio $\frac{p_n}{p_o}$ is greater than a randomly generated number between 0 and 1, the new position is updated. This process is repeated for all particles in the system.

Periodic boundary conditions and the minimum image convention are applied in all directions. A cutoff is used for interaction truncation. If the random displacement is too small, all moves will be accepted, leading to poor sampling

of the configuration space. Conversely, if the displacement is too large, all moves will be rejected.

If the total number of trial moves is N_{trial} and the number of accepted Monte Carlo moves is N_{accept}, the displacement is adjusted to maintain an optimal acceptance ratio:

$$\frac{N_{\text{accept}}}{N_{\text{trial}}} = 0.5 \tag{6.10}$$

This ensures proper sampling of equilibrium phase space, allowing computation of various thermodynamic quantities over the conformations.

Bibliography

[1] Klein, M. L.; Shinoda, W. Large-Scale Molecular Dynamics Simulations of Self-Assembling Systems. Science **2008**, *321*, 798-800.

[2] Voth, G. A. Coarse-graining of condensed phase and biomolecular systems. CRC Press: **2008**.

[3] Noid, W. G. Perspective: coarse-grained models for biomolecular systems. J. Chem. Phys. **2013**, *139*, No. 090901.

[4] Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodriguez-Ropero, F.; van der Vegt, N. F. A. Systematic coarse-graining methods for soft matter simulations - a review. Soft Matter **2013**, *9*, 2108-2119.

[5] Ingo lfsson, H. I.; Lo pez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The Power of Coarse Graining in Biomolecular Simulations. WIREs Comput. Mol. Sci. **2014**, *4*, 225-248.

[6] Alessandri, R.; Souza, P. C.; Thallmair, S.; Melo, M. N.; De Vries, A. H.; Marrink, S. J. Pitfalls of the Martini model. J. Chem. Theory Comput. **2019**, *15*, 5448-5460.

[7] De Jong, D. H.; Singh, G.; Bennett, W. D.; Arnarez, C.; Wassenaar, T. A.; Schafer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved parameters for the martini coarse-grained protein force field. J. Chem. Theory Comput. **2013** *9*, 687-697.

[8] Takada, S. Coarse-grained molecular simulations of large biomolecules. Curr. Opin. Struct. Biol. **2012** *22*, 130-137.

[9] Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. J. Phys. Chem. B **2007**, *111*, 7812-7824.

[10] Lopez, C. A.; Rzepiela, A. J.; de Vries, A. H.; Dijkhuizen, L.; Hunenberger, P. H.; Marrink, S. J. Martini coarse-grained force field: extension to carbohydrates. J. Chem. Theory Comput. **2009**, *5*, 3195-3210.

[11] Hadley, K. R.; McCabe, C. Coarse-grained molecular models of water: a review. Molecular simulation **2012**, *38*, 671-681.

[12] Borges-Araújo, L.; Patmanidis, I.; Singh, A. P.; Santos, L. H.; Sieradzan, A. K.; Vanni, S.; ... Souza, P. C. (Pragmatic coarse-graining of proteins: models and applications. J. Chem. Theory Comput. **2023**, *19*, 7112-7135.

[13] Ghosh Moulick, A.; Patel, R.; Onyema, A.; Loverde, S. M. Unveiling nucleosome dynamics: A comparative study using all-atom and coarse-grained simulations enhanced by principal component analysis. J. Chem. Phys. **2025**, *162*.

[14] Kharche, S.; Yadav, M.; Hande, V.; Prakash, S.; Sengupta, D. Improved protein dynamics and hydration in the Martini3 coarse-grain model. J. Chem. Inf. Model. **2024**, *64*, 837-850.

[15] `https://cgmartini.nl/`

[16] Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-grained protein models and their applications. Chem. Rev. **2016**, *116*, 7898-7936.

[17] Mukherjee, A.; Bagchi, B. Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. J. Chem. Phys. **2003** *118*, 4733-4747.

[18] Dill, K. A. Theory for the folding and stability of globular proteins. Biochemistry **1985**, *24*, 1501-1509.

[19] Lau, K. F.; Dill, K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules **1989**, *22*, 3986-3997.

[20] Dill, K. A.; Fiebig, K. M.; Chan, H. S. Cooperativity in protein-folding kinetics. PNAS **1993**, *90*, 1942-1946.

[21] Farris, A. C.; Shi, G.; Wüst, T.; Landau, D. P. The role of chain-stiffness in lattice protein models: A replica-exchange Wang-Landau study. J. Chem. Phys. **2018**, *149*, 125101.

[22] Mathiowetz, A. M.; Goddard III, W. A. Building proteins from Cα coordinates using the dihedral probability grid Monte Carlo method. Protein Sci. **1995**, *4*, 1217-1232.

[23] Evans, J. S.; Chan, S. I.; Mathiowetz, A. M.; Goddard III, W. A. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. Protein Sci. **1995**, *4*, 1203-1216.

[24] Moulick, A. G.; Chakrabarti, J. Conformational fluctuations in the molten globule state of α-lactalbumin. Phys. Chem. Chem. Phys. **2022**, *24*, 21348-21357.

[25] Moulick, A. G.; Chakrabarti, J. Fluctuation-dominated ligand binding in molten globule protein. J. Chem. Inf. Model. **2023**, *63*, 5583-5591.

[26] Dunker, A. K. et al. Intrinsically disordered protein. J. Mol. Graph. & Mod. **2001**, **19**, 26-59.

[27] Kociurzynski, R.; Makshakova, O. N.; Knecht, V. Romer, W. Multiscale molecular dynamics studies reveal different modes of receptor clustering by Gb3-binding lectins. J. Chem. Theory Comput. **2021**, *17*, 2488-2501.

[28] Wolberger, C. Homeodomain interactions. Curr. Opin. Struct. Biol. **1996**, *6*, 62-68.

[29] Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J. Am. Chem. Soc. **2003**, *125*, 9179-9191.

[30] Aishima, J.; Gitti, R. K.; Noah, J. E.; Gan, H. H.; Schlick, T.; Wolberger, C. A Hoogsteen base pair embedded in undistorted B-DNA. NAR **2002**, *30*, 5244-5252.

[31] Abraham, M. J.; Murtola, T.; Schulz, R.; Pall, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX **2015**, *1*, 19-25.

[32] Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B. The GROMACS Development Team. GROMACS User Manual **2018**.

[33] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved

protein backbone parameters. Proteins: Structure, Function, and Bioinformatics **2006**, *65*, 712-725.

[34] Petrova, S. S.; Solov'ev, A. D. The Origin of the Method of Steepest Descent. Hist. Math. **1997**, *24*, 361.

[35] Lemak, A. S.; Balabaev, N. K. On the Berendsen thermostat. Mol. Simul. **1994**, *13*, 177-187.

[36] Saito, H.; Nagao, H.; Nishikawa, K.; Kinugawa, K. Molecular collective dynamics in solid para-hydrogen and ortho-deuterium: The Parrinello-Rahman-type path integral centroid molecular dynamics approach. J. Chem. Phys. **2003**, *119*, 953-963.

[37] Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. J. Chem. Phys. **1995**, *103*, 3668-3679.

[38] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. **1997**, *18*, 1463-1472.

[39] `https://manual.gromacs.org/2024.0/reference-manual/analysis/radial-distribution-function.html`

[40] Chakrabarti, P.; Pal, D. Main-chain conformational features at different conformations of the side-chains in proteins. Protein Eng. **1998**, *11*, 631-647.

[41] Chatterjee, P.; Bagchi, S.; Sengupta, N. The non-uniform early structural response of globular proteins to cold denaturing conditions: a case study with Yfh1. J. Chem. Phys. **2014**, *141* 205103.

[42] Hollingsworth, S. A.; Karplus, P. A. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. **2010**.

[43] Sittel, F.; Filk, T.; Stock, G. Principal component analysis on a torus: Theory and application to protein dynamics. J. Chem. Phys. **2017**, *147*, 244101.

[44] Moulick, A. G.; Chakrabarti, J. Microscopic insights to relaxation phenomena in proteins. University of Calcutta, CUE-A18258-T14408, **2024**.

[45] Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: extension to proteins. J. Chem. Theory Comput. **2008**, *4*, 819-834.

[46] Darré, L.; Machado, M. R.; Brandner, A. F.; González, H. C.; Ferreira, S.; Pantano, S. SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. J. Chem. Theory Comput. **2015**, *11*, 723-739.

Chapter 7 Conclusion and outlook

In this thesis, we present detailed studies on the stability of non-canonical biomolecules, like DNA and proteins from various perspectives.

First, we investigate DNA systems containing a Hoogsteen (HG) base pair (bp) along with Watson–Crick (WC) bps, both in the absence and presence of protein bound at the HG bp. We examine the conformational stability and order of each bp in HG containing naked DNA compared to that in fully WC bps containing naked DNA. Additionally, we analyze the conformational stability and order of each bp in HG bp in the presence of homeodomain proteins. We observe that protein binding enhances the stability and order of both the HG bp and the entire DNA duplex by reducing fluctuations in the sugar-phosphate, sugar-base and sugar-pucker torsion angles. Furthermore, we compute the binding energy (B_E) of proteins with both HG-DNA and WC-DNA across various DNA–protein systems and correlate these values with conformational thermodynamic data at the protein–DNA interface. The correlation between conformational thermodynamics and B_E can be experimentally verified. Such a plot could be valuable for calibrating of protein binding to DNA. We also evaluate the free energy differences between WC and HG bps, as well as the energy barrier associated with the WC $\leftrightarrow$ HG bp transition in presence of proteins.

Next, we examine the conformational stability and structural order of the intrinsically disordered protein (IDP), α-synuclein (αS), in the presence of zinc oxide nanoparticles (ZnONPs) using all-atom molecular dynamics (MD) and coarse-grained (CG) Monte Carlo (MC) simulations. Our simulations show that the region responsible for conformational variability gains structural stability in the nanoparticle bound state.

The studies on the HG bp containing DNA system may help to model stable

HG-DNA-protein system, which may have significant implications in DNA replication, recognition, damage repair and other cellular processes. On the otherhand, our detailed study of α-synuclein in the presence of zinc oxide nanoparticles (ZnONPs) suggests that reduced fluctuations in the region responsible for aggregation may be responsible for reduced fibrillation rate of αS in the presence of ZnONPs. This establishes their potential as therapeutic agents against amyloid fibrillation. Furthermore, our CG approach may prove useful for investigating aggregation phenomena in proteins in general lacking canonical structures. We have the stability criterion on the conformational thermodynamic qualities of individual base pairs or residues to assess their conformational order. These results could be experimentally validated through nuclear magnetic resonance (NMR) spectroscopy.