ON THE MOLECULAR UNDERSTANDING OF PROTEIN STABILITY IN WATER, IN CHEMICAL CHAPERONE, AND ITS INTERCALATION TO DNA

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE

REQUIREMENTS

OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

REMAN KUMAR SINGH

20123193

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF SCIENCE EDUCATION AND

RESEARCH, PUNE – 411 008

2018

Dedicated to

My Family

Specially

My Sister,

Aarti

&

Dr Arnab Mukherjee

भारतीय बिज्ञान शिक्षा एबं अनुसंधान सँस्थान, पुणे
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH (IISER), PUNE
(An Autonomous Institution, Ministry of Human Resource Development, Govt. of India)
Dr Homi Bhabha Rd, Ward No. 8, NCL Colony, Pashan, Pune, Maharashtra 411008, India.

CERTIFICATE

Certified that, the work incorporated in the thesis entitled, "**On the Molecular Understanding of Protein Stability in Water, in Chemical Chaperone, and its Intercalation to DNA**" submitted by ***Reman Kumar Singh*** was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.

Date: 31th August 2018
Pune (MH), India

Dr Arnab Mukherjee
(Thesis Supervisor)

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 31th August 2018 **Reman Kumar Singh**

Pune (MH), India **Reg: No: 20123193**

ACKNOWLEDGMENTS

I would like to acknowledge Indian Institute of Science Education and Research (IISER), Pune for providing the amazing research environment. I would like to acknowledge the University Grand Commission (UGC) for fellowship. I am grateful to **Prof. K. N. Ganesh**, former Director and **Prof. Jayant B. Udgaonkar**, current Director of IISER Pune.

I would like to thank all those who have helped me in my PhD life. First and foremost, I take this opportunity to express my sincere gratitude to my supervisor, **Dr. Arnab Mukherjee** for his valuable guidance, suggestions, encouragement and support given throughout my PhD life.

I would like to thank my Research Advisory Committee (RAC) members **Dr. Durba Sengupta** and **Dr Sai Krishnan Kayarat** for their helpful suggestions during my annual evaluations.

I would like to thank **Dr. Partha Hazra, Dr. Suman Chakraborty, Dr. Sayan Bagchi, Dr. Kausik Chakraborty, Dr. Jeetender Chugh and Dr. Amrita Hazara** for all the collaborative projects. I have learnt so many new concepts from these works, and have learnt how computational methods can be useful to study and modulate the system at molecular levels.

I would like to thank my labmates **Wilbee, Mandar, Debasis, Hridya, Pallavi, Amitosh**, for the discussions. I deeply thank **Dr. Willbee D. S.** for teaching me during my initial PhD life. I would like to thank all my NCL and IISER friends (Meghna, Plawan, Avdhoot, Sonashree, Monika, Mehak, Vikas, Praveen, Veer, Mahesh, Yashwant, Pal, and many more..) for the fun and discussion, we had together. I deeply thank my best friend **Meghna Manae** for her continuous support, care, help and patience to make my life much easier at IISER

I would like to thank all the staff members of IISER for their help and support. I would like thank the IT department of IISER Pune for all the help related to the computers.

I would like to thank YUVA and CSIR-4PI for the computational facility.

Lastly, I would like to thank my family especially my sister for all the support and encouragement.

List of Publications

Peer-reviewed

1) Sagar Sapathi, **Reman Kr. Singh**, Mukherjee, Arnab & Partha Hazara;Controlling an Anticancer Drug Mediated Gquadruplex Formation and Stabilization by a Molecular Container. ***Phys. Chem. Chem. Phys.***, **2018**,20, 7808-7818.
2) **Reman Kr. Singh**, Neharika G. Chamach, Suman Chakrabarty & Arnab Mukherjee, Mechanism of Prion Protein Unfolding. ***J. Phys. Chem. B*, 2017**, *121* (3), pp 550–564.

   (Discussed in Chapter 3)
3) **Reman Kr. Singh**, Wilbee D S & Arnab Mukherjee, Molecular Origin of DNA Kinking by Transcription factors. ***J. Phys. Chem. B*, 2015**, *119* (35), pp 11590–11596.

   (Discussed in Chapter 5)
4) Abhigyan Sengupta, **Reman Kr. Singh**, Krishna Gavvala, Raj Kumar Koninti, Arnab Mukherjee, & Partha Hazra**,** Urea induced unfolding dynamics of flavin adenine dinucleotide (FAD): spectroscopic and molecular dynamics simulation studies from femto-second to nanosecond regime. ***J. Phys. Chem. B*, 2014**, *118* (7), pp 1881–1890.

Manuscript under preparation

1) **Reman Kr. Singh**, Kausik Chakraborty & Arnab Mukherjee, Effect of monomer amino acid on ubiquitin unfolding. (Manuscript is under preparation)

   (Discussed in Chapter 4)
2) **Reman Kr. Singh**, & Arnab Mukherjee, Mechanism of DNA-Protein interaction: study of SOX4-DNA and Sac7d-DNA complex. (Manuscript is under preparation)

   (Discussed in Chapter 6)
3) **Reman Kr. Singh**, & Arnab Mukherjee, Mechanism of DNA-Protein interaction: Sac7d-DNA complex. (Manuscript is under preparation)

   (Discussed in Chapter 7)
4) **Reman Kr. Singh,** Sayan Bagchi & Arnab Mukherjee, Effect of Ions on amide bond frequency. (Manuscript is under preparation)

CONTENTS

Synopsis

Title: On the Molecular Understanding of Protein Stability in Water, in Chemical Chaperone, and its Intercalation to DNA

Protein and DNA are two very important macromolecules in living organisms. Both have different kind of functions. DNA encodes the genetic information whereas proteins catalyse the biological reaction. To perform the biological function, proteins have to be in a folded native state. The conformational change or misfolding of protein causes disease in living organisms. For example, misfolding of prion protein causes neurodegenerative disease in all mammals. Prion protein is a glycoprotein, located at brain cell. It has three helices and an anti-parallel β-sheet. The cellular prion protein, termed as PrP^{C}, converts to a toxic, more β-rich conformation (PrP^{Sc}). Various studies, including experimental and computational, have been performed to find the mechanism of prion protein conversion. Yet the mechanism is unclear. In this thesis, we have used the molecular dynamic simulation with enhanced sampling techniques to understand the mechanism of full prion unfolding for the first time in molecular detail. Our finding shows that the cellular prion protein is an intrinsically disordered protein and it is in a metastable state in which any small perturbation would lead to a more stable unfolded state.

As mention above, protein conformational change would cause the fatal disease in living organisms. The protein folding, unfolding, or conformational changes are extremely dependent on the environment of proteins. It has been proposed that the inorganic salts (e.g. Na_2So_4, $CaCl_2$) perturb the protein folding/unfolding. Similarly, isolated amino acids (AA) also act as chemical chaperones and perturb the folding and stability of the protein. The unique feature of amino acids is that they can exist in zwitterion and this way they would interact with the positive (ARG and LYS) and negative (ASP and GLU) residues of a protein.

In this thesis we have investigated the effect of glycine and other AA on protein folding. We have observed that this AA interact with the charged residues of the protein. Moreover, among these charged residues they preferentially bind to the side chain of the ARG and LYS in protein using the carboxyl group.

Proteins also bind to DNA to perform some biological functions. The protein binding is sometimes very specific. The binding of certain proteins such as transcription factors (TFs) involves a conformational change in DNA e.g. bending and kinking. It has been shown that the conformation change in DNA is the reason for specific DNA-protein complex formation. The crystal structure of all these complexes shows a partial intercalation of AA at the kinked site. Here, we have investigated the role of intercalation in DNA kinking. We have observed that the bending alone does not kink the DNA, but a partial intercalation of AA does. Further, we have investigated the mechanism of AA intercalation into DNA using two different systems (SOX4-DNA and Sac7d-DNA) and observed that bending, kinking and an intercalation is a simultaneous event.

In summary, the current thesis contains my findings on the unfolding mechanism of prion protein (a neurodegenerative disease causing protein), the roll of amino acids as chemical chaperons in protein folding, and the molecular mechanism of protein intercalation into DNA and its kinking.

Chapter 1: Introduction. This chapter gives an overview on proteins, its conformation changes, and implication of its misfolding for the disease. We also discussed the effect of environment on protein stability. Further, we have discussed the protein-DNA interactions and their functional implications.

Chapter 2: Methods. All the process we have studied in this thesis happens in the microsecond to millisecond timescale. Therefore, regular molecular dynamics simulation will take a very long time to sample the conformational space. Therefore, we have used

accelerated sampling techniques, such as umbrella sampling, well-tempered metadynamics, etc. These techniques bias the system along particular reaction coordinate (RC). In this thesis we have used Native contact (Nc), Radius of gyration (Rg), distances as RCs. We also designed some RCs on our own to study some processes. All these methods are discussed in this chapter.

Chapter 3: Mechanism of Prion Protein Unfolding. In this chapter, we have addressed the long-standing question of prion conversion from cellular prion (PrP^{C}) to another scrapy form (PrP^{Sc}) using multiple microseconds well-tempered metadynamics simulation to capture the misfolding of the prion protein. Interestingly, we have found that the native cellular prion protein (PrP^{C}) is an intrinsically disordered protein and a small perturbation would deform the protein to a more stable less structured protein. Further, we have not observed any β-rich conformation along the misfolding pathways. Thus, we believed that more β-rich structure is not possible in the monomer. It may be possible only in a dimer or higher oligomers. Free energetic studies on isolated helices of prior protein show that individual helices prefer to be in a random coil or in β-rich conformation instead of helices.

Chapter 4. Mechanism of Ubiquitin Unfolding in Chemical Chaperon. Protein folding-unfolding is dependent on the environment of proteins. For example, the presence of inorganic salts (e.g., $CaCl_2$, Na_2So_4) are known to perturb the protein folding-unfolding equilibrium. Similarly, amino acids also show the properties of chemical chaperons that perturb the protein folding rate. The uniqueness of an amino acid (AA) as chemical chaperon, unlike organic salts, is that they exist in zwitterionic form. Therefore, they interact with both positive and negative residues of a protein. This chapter describes the effect of monomeric amino acids on protein stability using extensive all-atom well-tempered metadynamics simulations. We have studied the unfolding of ubiquitin protein in the presence and absence of different amino acids (GLY, C-terminal capped GLY and N-ter ASP). We found that all

the amino acids stabilize the protein in its native state. Further analyses show that these AAs mostly interact with charged residues of protein and specifically with ARG and LYS in protein due to the presence of carboxylate group in the AA.

Chapter 5: Molecular Origin of DNA Kinking by Transcription Factors. Some functions of proteins depend on its binding to DNA. Among all possible modes of binding, intercalation is a special type where amino acid inserts between a pair of basepairs of DNA. Protein intercalation into DNA is observed in for transcription factor proteins which, upon binding, cause DNA bending (a global change in curvature) and kinking (a local sharp change in curvature). In this chapter, we investigated the origin of DNA kink observed in TF caused by the intercalation of AA. We have used four different systems where a particular protein's AA is intercalated into DNA. We studied the free energy of the intercalation of the AA while its part of the protein and in isolation. We found that bending of the DNA alone does not kink the DNA. The partial intercalation of the AA is the cause of DNA kinking. However, it would not have produced the free energetically stable kinked state without the rest of the proteins.

Chapter 6: Molecular Mechanism of Transcription Factor Protein Intercalation into DNA Inducing Bends and Kinks. In the previous chapter, we studied how bending and kinking happens when protein intercalates partially into DNA. However, the mechanism of protein intercalation is not known in molecular detail. In this chapter and the next, we studied the mechanism of protein intercalation into DNA. In this chapter, we studied the mechanism of SOX-4 protein binding to DNA using molecular dynamics (MD) and well-tempered metadynamics simulations. We created the nonspecific bound states by performing normal MD simulations. From the distribution of the associated states, we performed well-tempered metadynamics simulation using some chosen reaction coordinate and performed well-tempered metadynamics simulation to form the intercalated native state and free state. We

have observed that differently associated complexes have different paths for the association. In that, we identified the minimum free energy path (MFEP) and analyzed the molecular events along that path which revealed that bending, kinking and intercalation are simultaneous events. Interestingly, we found that the de-intercalation path follows the MFEP of the intercalation path.

Chapter 7: Molecular Mechanism of Sac7d Intercalation into DNA Inducing Bends and Kinks. In this chapter, we took another example of protein-DNA intercalation involving Sac7d-DNA system, present in hyperthomiphilc family. These complex are very stable to heat and acidic environment. Unlike the previous system, the present one has a nonspecific interaction where two amino acids (VAL26 and MET29) intercalate into the DNA. The primary objective of the present study is to understand the interdependency of the two AAs in the intercalation process. We found that, like the previous study, the AA intercalation is accompanied with bending and kinking. Analyses of the intercalation pathway show that the both AAs (VAL26 and MET29) intercalate simultaneously. Also, dissociation of one of the intercalated AA leads to the spontaneous dissociation of the other.

Chapter 1
Introduction

1.1 Proteins

Proteins are polymers with amino acids as their monomeric units. There are 20 naturally occurring amino acids (AA), which are linked by peptide bonds to form a protein[1]. These AA chains fold into a three-dimensional conformation, which is referred to as the native structure of a protein[1]. Protein folding is a spontaneous process and highly dependent on its environment and on the AA sequence. In water, hydrophobic parts of the protein get destabilized and collapse inwards into compact structures in such a manner that the hydrophobic core is embedded inside the protein whereas the hydrophilic groups of the protein turn out towards the surface and interact with water and ions [2-3].

(I) Function of proteins: Proteins perform a variety of functions in living organisms. They act as antibodies, hormones, add structural integrity to the cell, transport small molecules, and most important of all, proteins act as enzymes [4-5]. Some specific examples of protein function are transporting oxygen throughout the body (haemoglobin and myoglobin)[6], help in digestion (trypsin and pepsin), regulating sugar levels in the body (insulin)[7], maintaining the structure of tissues (Keratin and collagen), etc.

The aforementioned protein functionalities strongly depend on the protein conformation. Generally, a correctly folded protein is functional whereas an incorrectly folded or misfolded protein is non-functional. For example, insulin loses its activity at high temperatures because of unfolding.[7] Overall, several factors such as, pH, salt concentration, temperature affect protein stability. Generally, acidic media and high temperatures cause protein denaturation.

Protein folding and its functionality is a delicate balance. As mentioned above, for a protein to serve a particular function, it needs to be folded in a very specific conformation. Sometimes, a more ordered folding of the protein is also termed as misfolding as it will no longer function the way it is supposed to [8-10]. The misfolded conformation is usually infectious and causes serious diseases in mammals. This phenomena of protein misfolding is called "proteopathy".

Proteopathy causes a class of diseases in which a protein folds to an abnormal form and hence affects the cell or other biological processes[8-10]. The Prion disease[11-15], Alzheimer's disease[16], Parkinson's disease[17], Amyloidsis[18-20] and other disorders are examples of proteopathies diseases[21-22]. In most, not all, proteopathies, the structural change in the native state of the protein increases its tendency to bind to itself and cause aggregation[23-24]. This aggregated form of the protein is resistant to proteasome cleavage and gets accumulated in cells. For example, prion and tau protein aggregate to form scrapie and amyloids. It is believed that these proteins convert to a more β-sheet rich conformation which is responsible for aggregation. There are many variables which can trigger protein misfolding and cause proteopathies diseases[23, 25-26]. While there are several diseases which are related to protein misfolding, here we focus on one of them, the Prion disease.

1.2 Prion disease: Prion disease is a neurodegenerative disorder, which affects most mammals[27]. The Prion disease has a different name depending on the mammal it affects. It is known as Creutzfeldt-Jacob disease, Familial Insomnia and Kuru in humans, Chronic Wasting disease in deer and elk, scrapie in sheep, Bovine spongiform encephalopathy (also known as mad cow disease) in cattle[28]. These diseases are caused by misfolding or conformational changes of the prion protein. It is characterized by the accumulation of an abnormal form of cellular prion protein on the brain cell surface. In all these diseases, the cellular form of the prion protein (PrP^{C}) gets converted to a protease enzyme resistant, infectious, pathogenic scrapie isoform (PrP^{Sc}) [29].

(I) Cellular form of prion protein: Prion protein is a glycoprotein, mostly located on the surface of brain cells, in the central nervous system and is also found in several peripheral tissues and leukocytes[30 31]. Its structure has been completely solved using NMR, as summarised in Fig 1. The PrP^C of humans consists of an unstructured, flexible N-terminal (residue no. 23-124) and a well-structured C-terminal core (residue no. 125-228). The core region is composed of three helices (H1, H2, and H3, see Fig 1) and an anti-parallel β-sheet composed of three residues and there is a disulphide bond between residues N181 and N197.[32-33]

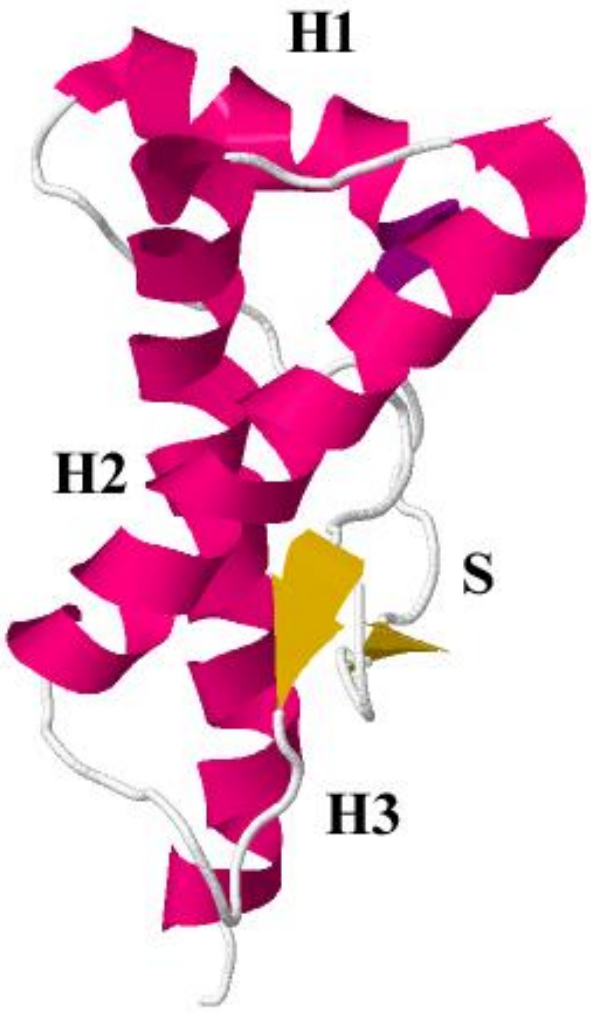

Figure 1: NMR structure of prion protein (C-terminal region only). H1, H2 and H3 are the three helixes of protein and S represents the anti-parallel beta sheet of protein. PDB ID 1HJM. [34]

(II) Function of cellular prion protein: The function of cellular prion (PrP^{C}) is still not fully understood. So far, it has been found that it helps in copper transport across cell membranes, signal transduction, myelin repair, immune cell activation, cell adhesion, and some other processes [35-40] [41-43] [44-49]. The cellular form of the prion is non-infectious, but it has been reported that its absence PrP^{C} causes serious conditions such as demyleniation of cell. However, PrP^{C} converts to infectious conformation "PrP^{Sc}". The structure of PrP^{C} has been solved completely, and it is known to have 43% helix and 3% beta sheet. While the structure of PrP^{Sc} is not completely known, CD and FRET spectroscopy shows that it has 30% helix and 30% beta sheet.

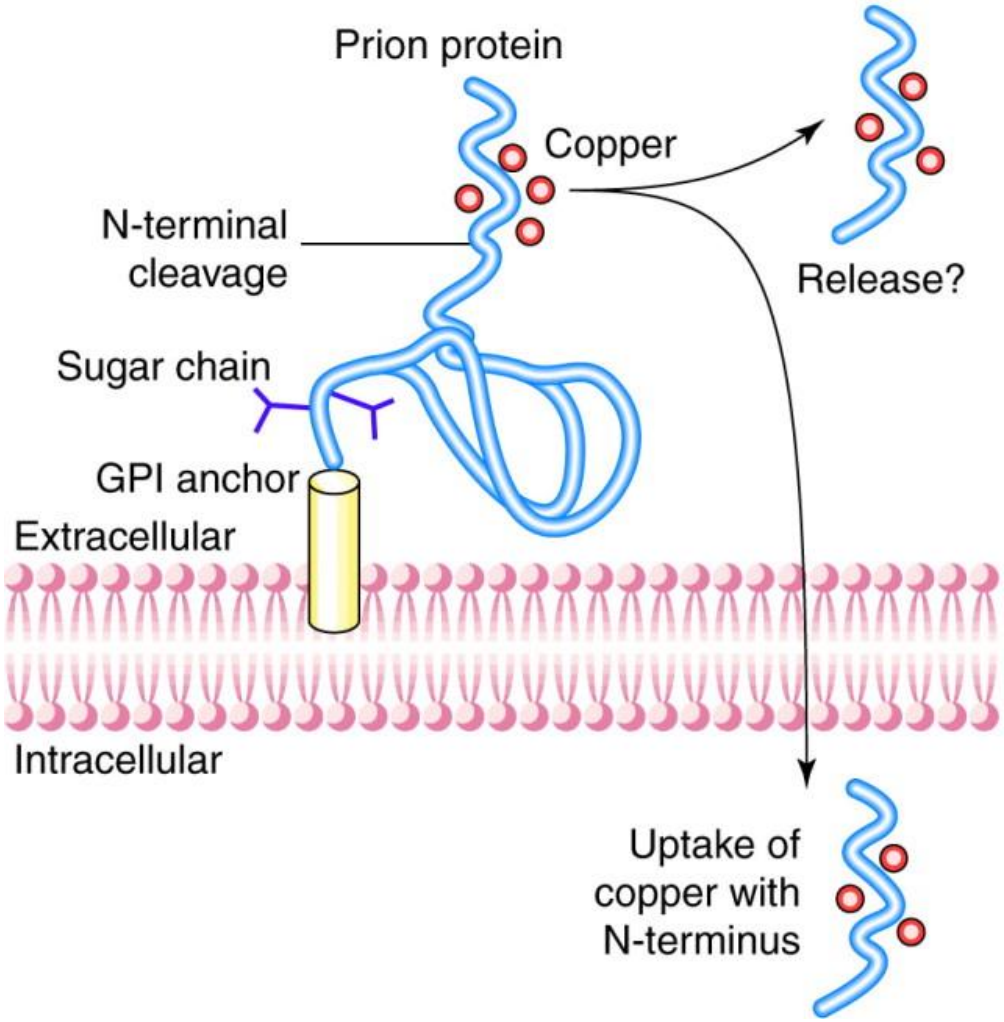

Figure 2: Prion protein attached to the brain cell membrane. This figure shows the transportation of copper across brain cell. Fig is Adapted from "Prion and prejudice: normal protein and the synapse." by David R Brown. Trend in Neurosciences.

(III) Conformational change of PrP^{C} to PrP^{Sc}:

It has been shown that the conformational change of native α-helix rich PrP^{C} to the more infectious β-sheet rich PrP^{Sc} is the reason for the accumulation of PrP^{Sc} aggregates[50] and Prion diseases. However, the exact mechanism of this conversion is still unsolved, even though many theories have been put forward. Among them, the most popular and widely accepted is the "protein only" hypothesis which states that the protein itself can replicate without the use of any nucleic acid. This was surprising in itself. According to the "protein only" hypothesis,

prion protein, the proteinaceous infectious agent, propagates by changing its cellular part into an infectious form. The PrP^{C} and PrP^{Sc} are chemically similar, with the only difference between them being their secondary structure. PrP^{C} has more helix content while PrP^{Sc} has more β-sheet proportion. Therefore, based on the secondary structure content and also from other experiments and molecular simulations, researchers have suggested that one or more helices of PrP^{C} convert to the β-sheet. This is followed by aggregation to form PrP^{Sc}. Two mechanisms have been proposed for this aggregation step: (i) the "Refolding" model which hypothesizes that some part of PrP^{C} unfolds and again refolds in the presence of PrP^{Sc} and then aggregates form and (ii) the "Seeding" model that suggests that the PrP^{C} is in equilibrium with the infectious PrP^{Sc} and this equilibrium structure acts as the seed for further growth of PrP^{Sc}[51-53]. Both models have been shown schematically in Fig 3.

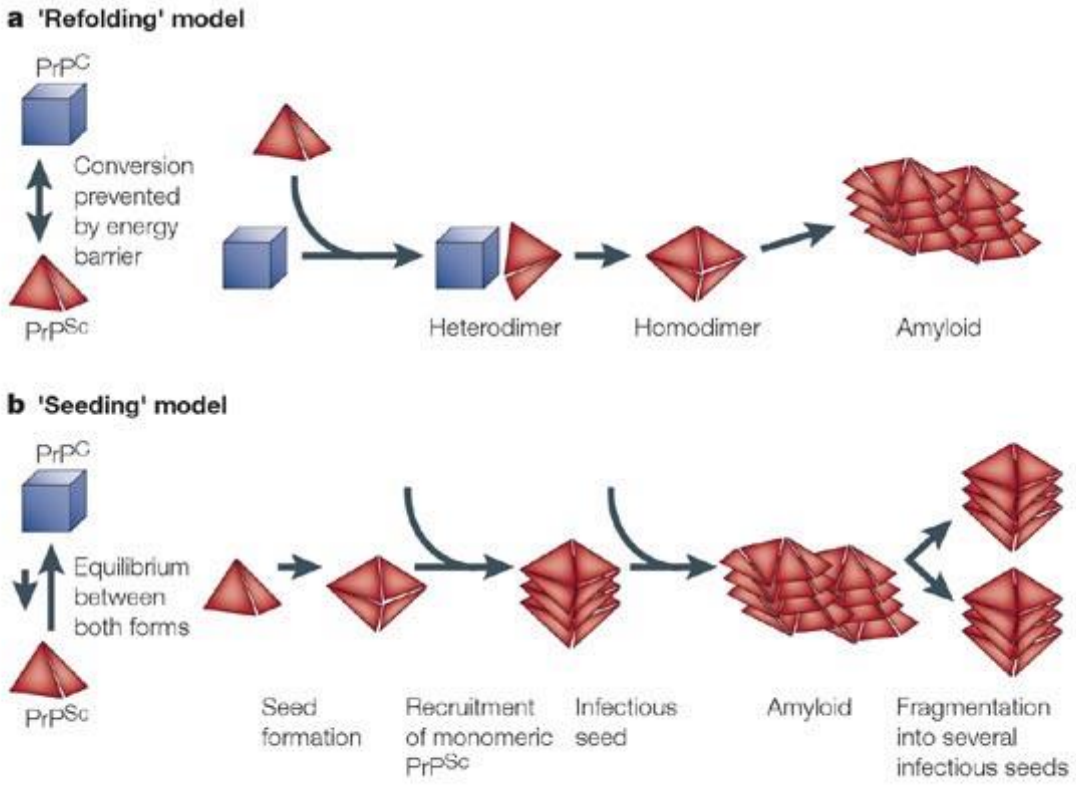

Figure 3: Schematics of proposed model for prion propagation. (a) Refolding model (b) Seeding model. Fig is Adapted from "Prions: health scare and biological challenge." by Aguzi et al., 2001. Nature Reviews Molecular Cell Biology 2, p.118-126. Copyright 2001 by Macmillan Publishers limited.

However, the complete mechanism of prion conversion is still unclear. In this thesis, we have used all-atom molecule dynamic simulations along with well-tempered metadynamics, to study the prion conversion.

1.3 Protein stability in different environments

Protein misfolding, associated with many diseases, is a consequence of an equilibrium between internal interaction of protein AA and their interaction with the environment.[54-55-57] For example, protein folding in a dehydrated environment has many local minima in its folding pathway, where the protein can be trapped and remain unfolded extended timescales. However, the same protein in a completely hydrated medium has a smooth free energy surface and the

folding is much more efficient. In addition to water, ions in the solution also affect protein stability. In 1988, Hofmeister showed that some ions like F^-, $SO4^2$, $HPO4^{2-}$, COO^- stabilize the protein, while some other ions like $ClO3^-$, I^-,$ClO4^-$, SCN^- destabilize proteins.[58-59] Following these observations, several attempts have been made to explain the effect of ions on protein stability. However, due to large number of variables such as protein structure, charge on the protein surface, concentration of ions, and identity of ions, the exact mechanism is not known,[60-62 63] though researchers have established two general mechanisms by which ions could affect proteins. (i) The first mechanism statest that ions directly bind to the protein residues and backbone.[60, 64-71] This type of mechanism is termed as "direct mechanism". (ii) In the second mechanism, ions are argued to perturb the solvation structure around the protein which eventually affects the protein stability. This type of mechanism is referred to as "indirect mechanism"[68, 72 73]. The big ions like perchlorate, acetate, and citrate have less charge density, therefore they hardly disturb the ordered water structure around proteins. These ions directly interact with the protein[64, 74]. However, small ions like SO_4^{2-}, Cl^- , K^+ have high charged density, and therefore they perturbed the water structure around protein strongly, which eventually affects protein stability[64, 74].

Later in 1995, it was found that yeast uses Trehalose molecule as a protecting agent against thermal denaturation of protein. More studies showed that small organic osmolytes and amino acids also perturbed protein stability.[75-77] The cosolvent which stabilizes the protein are termed as chemical chaperon e.g. free amino acid, carbohydrates and methylamines,[78-80] while the cosolvents which destabilizes the proteins are referred to as denaturing cosolvent e. g urea and guanidium ion.[76, 81] It has been reported that the protecting cosolvent (chemical chaperon) destabilizes the unfolded state without affecting the native state, which in turn favours the protein folding.[82-87] Alternatively, the denuarating cosolvent stabilizes the unfolded state, which favours protein unfolding[85-86]. Experimentally, scientists have correlated the transfer

free energy (ΔG_{tr}) of protein with protein stability.[86, 88-90] Transfer free energy is a measure of free energy change when the protein is transferred from water to a 1M cosolvent solution.[86, 88-90] The negative ΔG_{tr} signifies that the protein is more stable in cosolvent solution while positive ΔG_{tr} signifies the opposite. The ΔG_{tr} is negatively correlated with the cosolvent polar surface area. This negative correlation implies that the interaction between protein backbone and the polar group of the cosolvents is more favourable than with non-polar groups. Based on these observations, Rose *et al.* have proposed a theoretical model where they have constructed a solvation model in which backbone-solvent interaction energy is a function of solvent polarity, and the number of energetically equivalent ways of realizing a given interaction is a function of interacting surface area. Based on this model, they predicted that protecting/denaturing osmolytes would be preferentially excluded/accumulated around the protein backbone[77].

Here, in this thesis we have studied the ubiquitin protein stability in presence of amino acid (AA) and modified AA as cosolvents. The advantage of AA over other osmolytes is that the free AA in its zwitterionic form has both a positive terminal (NH_3^+) and a negative terminal (COO^-). Due to the presence of both charge groups, AA can interact with both positive (LYS, ARG) and negative (ASP, GLU) amino acids of protein. Therefore, the preferential effects of charges on the stability of proteins could also be captured using AA as cosolvent.

1.4 DNA-Protein Interaction

A folded protein performs many functions in a cell. For instance, it binds with different macromolecules/drug[91-96] to initiate several biological processes. Among the vast variety of functions proteins fulfil, the specific binding with DNA to initiate replication, transcription, DNA repair etc. is very important.[97-98] Therefore, the study of DNA-protein complex has important biological significance. There are many DNA protein complexes present in literature

and several studies have been performed over the last five decades on these complexes, but a complete understanding on how protein finds its specific binding site in DNA is still missing.

The structural motif of proteins which binds to DNA is called DNA binding domain (DBD). For example, leucine zipper proteins use helix rich regions for binding to DNA. Generally, the DBD of proteins have defined secondary structure in many DNA-protein complexes. However, in some complexes, a flexible segment far from the structured core of protein is used to make a specific complex as seen in case of λ-repressor protein which uses its disorderd N-terminal to make a specific complex.

The DBD is grouped into several categories depending on their structure at binding domain of protein. For example, a helix-rich domain has many α helices at DBD of protein as observed in the Lac repressor protein[99] (Fig 4a) whereas, a β-rich domain has more β-sheet at its binding site as seen in TATA box binding protein (Fig 4b). [100] Conversely, there are some DBDs in proteins which have both helix and β-sheet structure at binding, an example of which is found in Zinc finger protein which uses one α-helix and two β-sheet structural units at its binding site (Fig. 4c). [101] Apart from the above categories, some proteins have more than one DBDs and they simultaneously use all their domains to bind with DNA.

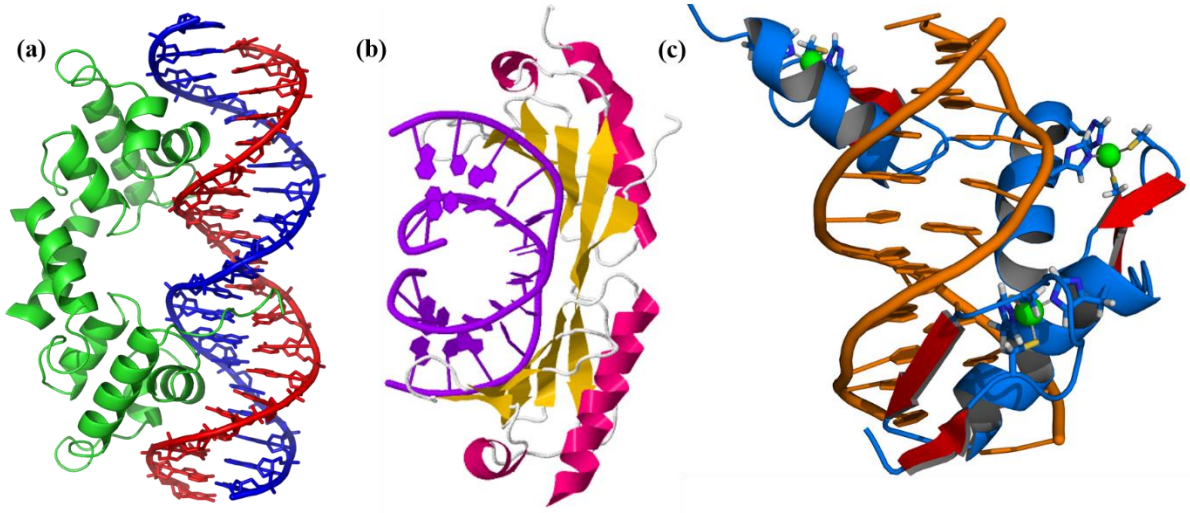

Figure 4: Some example of protein bound DNA complexes. (a) α-helix rich DBD of protein (Lambda repressor, PDB ID 1LMB) bound to DNA (b) β-sheet rich DBD of protein (TATA binding protein bound to the DNA)[100] (c) Mixed α/β rich domain (Zinc finger binding protein bound to the DNA).[101]

(I) Specific binding mechanism

The specific binding mechanism of DBD of protein with DNA is broadly divided into two categories: base readout and shape readout mechanisms. (a) In base readout mechanism, proteins read the bases of DNA and finds a specific interaction with a certain base pair. These interactions are mainly governed by the hydrophobic interactions and hydrogen bonding with bases. (b) in the shape readout mechanism found in some protein-DNA complexes, the specificity cannot be explained by hydrophobic contact or by hydrogen binding with DNA bases. Other factors, like conformation of DNA, also play equal role in DNA binding. It has been observed that the many transcription factor proteins deform the DNA conformation upon binding. However, the shape of DNA which would be preferred by the protein is dependent on the complex. Naturally, the question that arises in these "shape readout" mechanisms is that whether protein binds to preformed-deformed DNA conformation (conformational selection) or protein creates the deformation upon binding (induced fit mechanism).

The shape readout is further divided into two major categories. (i) The first one is called global shape readout where proteins recognise the overall shape of DNA, like bending in DNA, specific conformation of DNA such as B-DNA, A-DNA or Z-DNA. (ii) The second one is called local shapes read out where proteins recognize the local structural shape of DNA. For instance, some proteins bind to the minor groove of DNA because of its narrow size. Example of minor groove binding proteins are TBP (TATA binding protein), SRY (sex determining protein). LEF-1 (Lymphoid Enhancer-Binding Factor), IHF (integration host factor protein). Mostly these proteins use the arginine residue, which is present in their N-terminus, to bind to the minor groove of DNA. Like minor groove binding protein, there are some proteins which bind specifically to the major groove of DNA. However, the binding to the major groove is only favourable when some part of the protein first binds to the minor groove of DNA resulting in a wider minor groove and narrow major groove. [102] An example of major groove binding protein is the hRFX1 (Helix Regulatory Factor X-box) protein, which inserts a helix into the minor groove such that it widens the minor groove and narrows the major groove, and then uses β-ribbon motif to bind to the major groove of DNA.

Along the same lines, it has been reported that some proteins recognize the local abrupt change in DNA e.g. a kink in DNA. The kink in a DNA is produced by bending the DNA beyond a certain angle or by intercalating one or more amino acids to a specific base step [103] [104] [105-106] [107] [108]. The kink state is characterized by the high roll angle at kink base steps (see next section for detail). The binding of EcorV to DNA is an excellent example of the role of DNA local deformation into binding. EcorV bind to the bend deformed DNA only, does not shows any activity with the straight non-deformed DNA. It has been reported that the kinked state disrupts the base pair stacking and destabilizes DNA. So to maintain the stability, protein inserts an AA to the kinked base step [103] [104] [105-106] [107] [108]. Nature is full of DNA-protein complexes where the DNA is deformed locally by the partial intercalation of AA, e.g. SOX-4-DNA complex where

MET intercalates between CpA base step, IHF-DNA complex where protein intercalates between C7pA8 base step, SRY-DNA complex, where isoleucine intercalates between C9pC8 base step. Mostly, all transcription factor complexes are found to have AA intercalated to DNA producing a kinked state of DNA. However, there are some protein-DNA complexes, where kink is present in DNA without any intercalation. The catabolite activator protein is a classic example of a protein-DNA complex where DNA gets deformed by kinking at two places without intercalating with any AA. [109]

The kinking is not only observed in DNA-protein complexes, but it has also been found during the intercalation process of daunomycine and proflavine drug with DNA. Willbee *et. al.* observed in the intercalation of proflavine into DNA that at the transition state, when the drug is partially intercalated, DNA shows a sharp bending and kinking at the intercalating base step. From this observation, we hypothesized that it may be the partial intercalation of the AA that causes the kink state in the DNA. We will show in later chapters that that is indeed the case. What is the role of the whole protein in kink formation, then? We will discuss it in more detail in chapter xx. The studies mentioned above on DNA-protein complexes are from static X-ray structures. Therefore, the dynamic information and mechanism of formation of such complexes are not known and subject of the part of the thesis.

(II) DNA structure upon protein binding:

Since DNA is made of two chain of nucleotides which adopt a helical structure, binding of protein often disrupts the DNA helical structure (See Fig 4, 5).

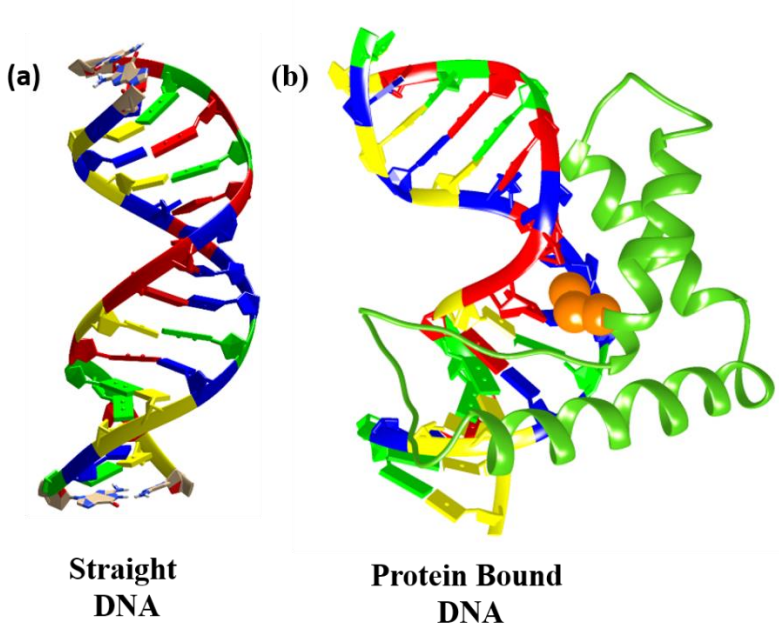

Figure 5: Straight DNA (a) and protein bound bend DNA. (b) Orange color in VdW representation is an intercalated AA.

There are several DNA parameters which define the change in DNA conformation. These are divided into two categories (i) local change in DNA base pairs (Fig 6a) and (ii) global change, (bending) of the whole DNA (Fig 6b). The local DNA parameter are broadly divided in two broad categories (i) translational parameter such as shift, rise and slide (ii) rotation parameter such as tilt, roll and twist.[110] Among these parameters, the roll and rise are being used in literature to define the intercalation. Briefly the rise define the distance between the two base step and the roll angle is the opening angle of the base pairs (Fig 6a). The intercalation of AA would increase the rise value of DNA base pairs. The bending of DNA is defined as how much DNA helical axis is bend form the regular DNA helical axis (Fig 6b).

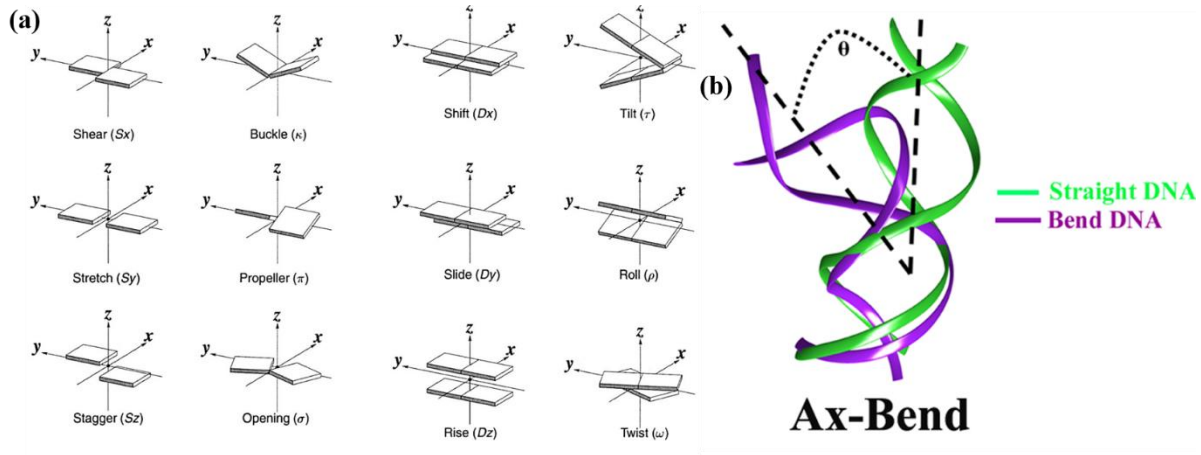

Figure 6: DNA parameters (a) Local change parameter. [110] Fig is adapted from the "3DNA server" [111] (b) Global change parameter

1.7 Outline of the Thesis

In this thesis, we have studied the unfolding mechanism of the prion protein, the effect of environment on protein stability and how protein bind to the DNA. We have used molecular dynamic with some accelerated method such as umbrella sampling and metadynamics, to studies the all the above process. A chapter-wise summary of the thesis is provided below.

Chapter 2 describe the method and the reaction coordinate used for the study.

Chapter 3 describes the mechanism of prion protein unfolding. We have performed multiple simulations and also simulations with a different force field. We have found that the prion protein is not stable in its native state, an unfolded state is more stable. Moreover, we have not observed any beta-sheet formation along unfolding pathways.

Chapter 4 describes the effect of free amino acid and modified AA on ubiquitin protein stability. Our long well-reprise metadynamics simulation shows that the free AA stabilizes the ubiquitin in its native state.

Chapter 5 describes the mechanism of DNA kinking by the transcription factors. Our studies show that the only partial intercalation of AA kink the DNA. However, the kink state is not stable hence protein is there to stabilize the kink state.

Chapter 6 describes the mechanism of DNA-Protein intercalation. We have studied the SOX4-DNA complex, in which one AA intercalate. Our study shows that the protein rotates around DNA to form a specific intercalative-native state. We have also observed that the bending and kinking happened when the AA partially intercalated. Binding alone does not bend and kink the DNA.

Chapter 7 describes the mechanism of Sac7d-DNA complex formation. The Sac7d-DNA complex has two intercalating AAs. Our study shows that the mechanism of protein association is same as observed in chapter 6. Moreover, we have also observed that the bending and kinking happened when the both AAs partially intercalated.

Chapter 8 describe two collaborative studies (Addendum I and II) with Dr Partha hazara and Dr Abhigyan Sengupta.

Addendum I describe the computational study of the FAD dynamic in water and in urea solution (8M)

Addendum II describe the computational study of G-quadruplex stability in presence and absence of TPT drug.

Reference

1. Butterworth, P. J., Lehninger: principles of biochemistry (4th edn) D. L. Nelson and M. C. Cox, W. H. Freeman & Co., New York, 1119 pp (plus 17 pp glossary), ISBN 0-7167-4339-6 (2004). *Cell Biochemistry and Function* **2005,** *23* (4), 293-294.
2. Brovchenko, I.; Oleinikova, A., Which Properties of a Spanning Network of Hydration Water Enable Biological Functions? *ChemPhysChem* **2008,** *9* (18), 2695-2702.
3. Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K., Forces contributing to the conformational stability of proteins. *The FASEB Journal* **1996,** *10* (1), 75-83.
4. Radzicka, A.; Wolfenden, R., A proficient enzyme. *Science* **1995,** *267* (5194), 90.
5. Furnham, N.; Holliday, G. L.; de Beer, T. A. P.; Jacobsen, J. O. B.; Pearson, W. R.; Thornton, J. M., The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research* **2014,** *42* (Database issue), D485-D489.
6. Muirhead, H.; Perutz, M. F., Structure Of Hæmoglobin: A Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin at 5.5 Å Resolution. *Nature* **1963,** *199*, 633.
7. Sanger, F.; Tuppy, H., The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **1951,** *49* (4), 463-481.
8. Walker, L. C.; LeVine, H., The cerebral proteopathies. *Molecular Neurobiology* **2000,** *21* (1), 83-95.
9. Walker, L. C.; LeVine, H., The cerebral proteopathies. *Neurobiology of Aging* **2000,** *21* (4), 559-561.
10. Walker, L. C.; LeVine, H., Corruption and Spread of Pathogenic Proteins in Neurodegenerative Diseases. *The Journal of Biological Chemistry* **2012,** *287* (40), 33109-33115.
11. Stanley, B. P., 1 An Introduction to Prion Biology and Diseases. *Cold Spring Harbor Monograph Archive; Volume 41 (2004): Prion Biology and Diseases, 2nd Ed.* **2004**.
12. Stanley, B. P., 2 Development of the Prion Concept. *Cold Spring Harbor Monograph Archive; Volume 41 (2004): Prion Biology and Diseases, 2nd Ed.* **2004**.
13. Stanley, B. P.; Elizabeth, W.; Jean-Louis, L.; Morikazu, S., 11 Scrapie, Chronic Wasting Disease, and Transmissible Mink Encephalopathy. *Cold Spring Harbor Monograph Archive; Volume 41 (2004): Prion Biology and Diseases, 2nd Ed.* **2004**.
14. Stephen, J. D.; James, W. I.; Essia, B.-B.; David, P.; Jan, R. F., 15 Neuropathology of Prion Diseases. *Cold Spring Harbor Monograph Archive; Volume 41 (2004): Prion Biology and Diseases, 2nd Ed.* **2004**.
15. Prusiner, S. B., Biology and Genetics of Prions Causing Neurodegeneration. *Annual review of genetics* **2013,** *47*, 601-623.
16. Walker, L. C., Proteopathic Strains and the Heterogeneity of Neurodegenerative Diseases. *Annual Review of Genetics* **2016,** *50* (1), 329-346.
17. Del Tredici, K.; Braak, H., Lewy pathology and neurodegeneration in premotor Parkinson's disease. *Movement Disorders* **2012,** *27* (5), 597-607.
18. Mrak, R. E.; Griffin, W. S. T., Dementia with Lewy bodies: Definition, diagnosis, and pathogenic relationship to Alzheimer's disease. *Neuropsychiatric Disease and Treatment* **2007,** *3* (5), 619-625.
19. Walker, L. C.; LeVine, H., III; Mattson, M. P.; Jucker, M., Inducible proteopathies. *Trends in Neurosciences* **2006,** *29* (8), 438-443.
20. Meyer-Luehmann, M.; Coomaraswamy, J.; Bolmont, T.; Kaeser, S.; Schaefer, C.; Kilger, E.; Neuenschwander, A.; Abramowski, D.; Frey, P.; Jaton, A. L.; Vigouret, J.-M.; Paganetti, P.; Walsh, D. M.; Mathews, P. M.; Ghiso, J.; Staufenbiel, M.; Walker, L. C.; Jucker,

M., Exogenous Induction of Cerebral ß-Amyloidogenesis Is Governed by Agent and Host. *Science* **2006,** *313* (5794), 1781.
21. Westermark, G. T.; Fändrich, M.; Lundmark, K.; Westermark, P., Noncerebral Amyloidoses: Aspects on Seeding, Cross-Seeding, and Transmission. *Cold Spring Harbor Perspectives in Medicine* **2017**.
22. Chiti, F.; Dobson, C. M., Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry* **2006,** *75* (1), 333-366.
23. Dobson, C. M., Protein misfolding, evolution and disease. *Trends in Biochemical Sciences* **1999,** *24* (9), 329-332.
24. Carrell, R. W.; Lomas, D. A., Conformational disease. *The Lancet* **1997,** *350* (9071), 134-138.
25. Selkoe, D., *Selkoe, D.J. Folding proteins in fatal ways. Nature 426, 900-904*. 2004; Vol. 426, p 900-4.
26. Eisenberg, D.; Jucker, M., The amyloid state of proteins in human diseases. *Cell* **2012,** *148* (6), 1188-1203.
27. Prusiner, S. B., Molecular biology of prion diseases. *Science* **1991,** *252* (5012), 1515.
28. Prusiner, S. B., Prions. *Proceedings of the National Academy of Sciences of the United States of America* **1998,** *95* (23), 13363-13383.
29. Prusiner, S. B.; Woerman, A. L.; Mordes, D. A.; Watts, J. C.; Rampersaud, R.; Berry, D. B.; Patel, S.; Oehler, A.; Lowe, J. K.; Kravitz, S. N.; Geschwind, D. H.; Glidden, D. V.; Halliday, G. M.; Middleton, L. T.; Gentleman, S. M.; Grinberg, L. T.; Giles, K., Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proceedings of the National Academy of Sciences of the United States of America* **2015,** *112* (38), E5308-E5317.
30. Hegde, R. S.; Mastrianni, J. A.; Scott, M. R.; DeFea, K. A.; Tremblay, P.; Torchia, M.; DeArmond, S. J.; Prusiner, S. B.; Lingappa, V. R., A Transmembrane Form of the Prion Protein in Neurodegenerative Disease. *Science* **1998,** *279* (5352), 827.
31. Stahl, N.; Borchelt, D. R.; Hsiao, K.; Prusiner, S. B., Scrapie prion protein contains a phosphatidylinositol glycolipid. *Cell* **1987,** *51* (2), 229-240.
32. Zahn, R.; Liu, A.; Luhrs, T.; Riek, R.; von Schroetter, C.; Lopez Garcia, F.; Billeter, M.; Calzolai, L.; Wider, G.; Wuthrich, K., NMR solution structure of the human prion protein. *Proc Natl Acad Sci U S A* **2000,** *97* (1), 145-50.
33. Apostol, M. I.; Perry, K.; Surewicz, W. K., Crystal structure of a human prion protein fragment reveals a motif for oligomer formation. *J Am Chem Soc* **2013,** *135* (28), 10202-5.
34. Calzolai, L.; Zahn, R., Influence of pH on NMR Structure and Stability of the Human Prion Protein Globular Domain. *Journal of Biological Chemistry* **2003,** *278* (37), 35592-35596.
35. Nishida, N.; Tremblay, P.; Sugimoto, T.; Shigematsu, K.; Shirabe, S.; Petromilli, C.; P Erpel, S.; Nakaoke, R.; Atarashi, R.; Houtani, T.; Torchia, M.; Sakaguchi, S.; J DeArmond, S.; B Prusiner, S.; Katamine, S., *A mouse prion protein transgene rescues mice deficient for the prion protein gene from Purkinje cell degeneration and demyelination*. 1999; Vol. 79, p 689-97.
36. Taylor, D. R.; Hooper, N. M., The prion protein and lipid rafts. *Mol Membr Biol* **2006,** *23* (1), 89-99.
37. Spielhaupter, C.; Schatzl, H. M., PrPC directly interacts with proteins involved in signaling pathways. *J Biol Chem* **2001,** *276* (48), 44604-12.
38. Loertscher, R.; Lavery, P., The role of glycosyl phosphatidyl inositol (GPI)-anchored cell surface proteins in T-cell activation. *Transplant Immunology* **2002,** *9* (2–4), 93-96.
39. Schneider, B.; Mutel, V.; Pietri, M.; Ermonval, M.; Mouillet-Richard, S.; Kellermann, O., NADPH oxidase and extracellular regulated kinases 1/2 are targets of prion protein

signaling in neuronal and nonneuronal cells. *Proc Natl Acad Sci U S A* **2003,** *100* (23), 13326-31.
40. Mouillet-Richard, S.; Ermonval, M.; Chebassier, C.; Laplanche, J. L.; Lehmann, S.; Launay, J. M.; Kellermann, O., Signal transduction through prion protein. *Science* **2000,** *289* (5486), 1925-1928.
41. Schmitt-Ulms, G.; Legname, G.; Baldwin, M. A.; Ball, H. L.; Bradon, N.; Bosque, P. J.; Crossin, K. L.; Edelman, G. M.; DeArmond, S. J.; Cohen, F. E.; Prusiner, S. B., Binding of neural cell adhesion molecules (N-CAMs) to the cellular prion protein. *J Mol Biol* **2001,** *314* (5), 1209-25.
42. Graner, E.; Mercadante, A. F.; Zanata, S. M.; Forlenza, O. V.; Cabral, A. L.; Veiga, S. S.; Juliano, M. A.; Roesler, R.; Walz, R.; Minetti, A.; Izquierdo, I.; Martins, V. R.; Brentani, R. R., Cellular prion protein binds laminin and mediates neuritogenesis. *Brain Res Mol Brain Res* **2000,** *76* (1), 85-92.
43. Graner, E.; Mercadante, A. F.; Zanata, S. M.; Martins, V. R.; Jay, D. G.; Brentani, R. R., Laminin-induced PC-12 cell differentiation is inhibited following laser inactivation of cellular prion protein. *Febs Letters* **2000,** *482* (3), 257-260.
44. Jackson, G. S.; Murray, I.; Hosszu, L. L.; Gibbs, N.; Waltho, J. P.; Clarke, A. R.; Collinge, J., Location and properties of metal-binding sites on the human prion protein. *Proc Natl Acad Sci U S A* **2001,** *98* (15), 8531-5.
45. Brown, D. R.; Qin, K.; Herms, J. W.; Madlung, A.; Manson, J.; Strome, R.; Fraser, P. E.; Kruck, T.; von Bohlen, A.; Schulz-Schaeffer, W.; Giese, A.; Westaway, D.; Kretzschmar, H., The cellular prion protein binds copper in vivo. *Nature* **1997,** *390* (6661), 684-7.
46. Kramer, M. L.; Kratzin, H. D.; Schmidt, B.; Romer, A.; Windl, O.; Liemann, S.; Hornemann, S.; Kretzschmar, H., Prion protein binds copper within the physiological concentration range. *Journal of Biological Chemistry* **2001,** *276* (20), 16711-16719.
47. Stockel, J.; Safar, J.; Wallace, A. C.; Cohen, F. E.; Prusiner, S. B., Prion protein selectively binds copper(II) ions. *Biochemistry* **1998,** *37* (20), 7185-93.
48. Jones, C. E.; Abdelraheim, S. R.; Brown, D. R.; Viles, J. H., Preferential Cu2+ coordination by His96 and His111 induces beta-sheet formation in the unstructured amyloidogenic region of the prion protein. *J Biol Chem* **2004,** *279* (31), 32018-27.
49. Leclerc, E.; Serban, H.; Prusiner, S. B.; Burton, D. R.; Williamson, R. A., Copper induces conformational changes in the N-terminal part of cell-surface PrPC. *Arch Virol* **2006,** *151* (11), 2103-9.
50. Aguzzi, A.; Sigurdson, C.; Heikenwaelder, M., Molecular Mechanisms of Prion Pathogenesis. *Annual Review of Pathology: Mechanisms of Disease* **2008,** *3* (1), 11-40.
51. Kocisko, D. A.; Come, J. H.; Priola, S. A.; Chesebro, B.; Raymond, G. J.; Lansbury, P. T.; Caughey, B., Cell-free formation of protease-resistant prion protein. *Nature* **1994,** *370*, 471.
52. Jarrett, J. T.; Lansbury, P. T., Seeding "one-dimensional crystallization" of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* **1993,** *73* (6), 1055-1058.
53. Kocisko, D. A.; Priola, S. A.; Raymond, G. J.; Chesebro, B.; Lansbury, P. T.; Caughey, B., Species specificity in the cell-free conversion of prion protein to protease-resistant forms: a model for the scrapie species barrier. *Proceedings of the National Academy of Sciences of the United States of America* **1995,** *92* (9), 3923-3927.
54. Fujita, Y.; Miyanaga, A.; Noda, Y., Effect of Alcohols on the Thermal Denaturation of Lysozyme as Measured by Differential Scanning Calorimetry. *Bulletin of the Chemical Society of Japan* **1979,** *52* (12), 3659-3662.
55. Fujita, Y.; Noda, Y., The Effect of Organic Solvents on the Thermal Denaturation of Lysozyme as Measured by Differential Scanning Calorimetry. *Bulletin of the Chemical Society of Japan* **1983,** *56* (1), 233-237.

56. Velicelebi, G.; Sturtevant, J. M., Thermodynamics of the denaturation of lysozyme in alcohol-water mixtures. *Biochemistry* **1979,** *18* (7), 1180-1186.
57. Fu, L.; Freire, E., On the origin of the enthalpy and entropy convergence temperatures in protein folding. *Proceedings of the National Academy of Sciences* **1992,** *89* (19), 9335-9338.
58. Yang, Z., Hofmeister effects: an explanation for the impact of ionic liquids on biocatalysis. *Journal of Biotechnology* **2009,** *144* (1), 12-22.
59. Nishimura, J. S.; Narayanasami, R.; Miller, R. T.; Roman, L. J.; Panda, S.; Masters, B. S. S., The Stimulatory Effects of Hofmeister Ions on the Activities of Neuronal Nitric-oxide Synthase: APPARENT SUBSTRATE INHIBITION BY l-ARGININE IS OVERCOME IN THE PRESENCE OF PROTEIN-DESTABILIZING AGENTS. *Journal of Biological Chemistry* **1999,** *274* (9), 5399-5406.
60. Bauduin, P.; Renoncourt, A.; Touraud, D.; Kunz, W.; Ninham, B. W., Hofmeister effect on enzymatic catalysis and colloidal structures. *Current Opinion in Colloid & Interface Science* **2004,** *9* (1–2), 43-47.
61. Gokarn, Y. R.; Fesinmeyer, R. M.; Saluja, A.; Razinkov, V.; Chase, S. F.; Laue, T. M.; Brems, D. N., Effective charge measurements reveal selective and preferential accumulation of anions, but not cations, at the protein surface in dilute salt solutions. *Protein Science* **2011,** *20* (3), 580-587.
62. Zhao, H., Effect of ions and other compatible solutes on enzyme activity, and its implication for biocatalysis using ionic liquids. *Journal of Molecular Catalysis B: Enzymatic* **2005,** *37* (1–6), 16-25.
63. Zhang, Y.; Cremer, P. S., The inverse and direct Hofmeister series for lysozyme. *Proceedings of the National Academy of Sciences* **2009,** *106* (36), 15249-15253.
64. Baldwin, R. L., How Hofmeister ion interactions affect protein stability. *Biophysical Journal* **1996,** *71* (4), 2056-2063.
65. Arakawa, T.; Timasheff, S. N., Mechanism of protein salting in and salting out by divalent cation salts: balance between hydration and salt binding. *Biochemistry* **1984,** *23* (25), 5912-5923.
66. Collins, K. D., Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process. *Methods* **2004,** *34* (3), 300-311.
67. Frank, H. S.; Wen, W.-Y., Ion-solvent interaction. Structural aspects of ion-solvent interaction in aqueous solutions: a suggested picture of water structure. *Discussions of the Faraday Society* **1957,** *24* (0), 133-140.
68. Wiggins, P. M., Hydrophobic hydration, hydrophobic forces and protein folding. *Physica A: Statistical Mechanics and its Applications* **1997,** *238* (1), 113-128.
69. Schellman, J. A., Protein Stability in Mixed Solvents: A Balance of Contact Interaction and Excluded Volume. *Biophysical Journal* **2003,** *85* (1), 108-125.
70. Davis-Searles, P. R.; Saunders, A. J.; Erie, D. A.; Winzor, D. J.; Pielak, G. J., Interpreting the Effects of Small Uncharged Solutes on Protein-Folding Equilibria. *Annual Review of Biophysics and Biomolecular Structure* **2001,** *30* (1), 271-306.
71. Minton, A. P., Excluded volume as a determinant of macromolecular structure and reactivity. *Biopolymers* **1981,** *20* (10), 2093-2120.
72. Tanford, C., The hydrophobic effect and the organization of living matter. *Science* **1978,** *200* (4345), 1012.
73. Perutz, M. F., Electrostatic effects in proteins. *Science* **1978,** *201* (4362), 1187.
74. Boström, M.; Williams, D. R. M.; Ninham, B. W., Why the properties of proteins in salt solutions follow a Hofmeister series. *Current Opinion in Colloid & Interface Science* **2004,** *9* (1–2), 48-52.

75. Mirsky, A. E.; Pauling, L., On the Structure of Native, Denatured, and Coagulated Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1936,** *22* (7), 439-447.
76. Schellman, J. A., Fifty years of solvent denaturation. *Biophysical Chemistry* **2002,** *96* (2–3), 91-101.
77. Street, T. O.; Bolen, D. W.; Rose, G. D., A molecular mechanism for osmolyte-induced protein stability. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (38), 13997-14002.
78. Record Jr, M. T.; Courtenay, E. S.; Cayley, D. S.; Guttman, H. J., Responses of E. coli to osmotic stress: large changes in amounts of cytoplasmic solutes and water. *Trends in Biochemical Sciences* **1998,** *23* (4), 143-148.
79. Record Jr, M. T.; Courtenay, E. S.; Cayley, S.; Guttman, H. J., Biophysical compensation mechanisms buffering E. coli protein–nucleic acid interactions against changing environments. *Trends in Biochemical Sciences* **1998,** *23* (5), 190-194.
80. Yancey, P. H.; Clark, M. E.; Hand, S. C.; Bowlus, R. D.; Somero, G. N., Living with water stress: evolution of osmolyte systems. *Science* **1982,** *217* (4566), 1214.
81. Greene, R. F.; Pace, C. N., Urea and Guanidine Hydrochloride Denaturation of Ribonuclease, Lysozyme, α-Chymotrypsin, and β-Lactoglobulin. *Journal of Biological Chemistry* **1974,** *249* (17), 5388-5393.
82. Auton, M.; Bolen, D. W., Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (42), 15065-15068.
83. Lee, J. C.; Timasheff, S. N., The stabilization of proteins by sucrose. *Journal of Biological Chemistry* **1981,** *256* (14), 7193-7201.
84. Makhatadze, G. I.; Privalov, P. L., Protein interactions with urea and guanidinium chloride. *Journal of Molecular Biology* **1992,** *226* (2), 491-505.
85. Bolen, D. W.; Baskakov, I. V., The osmophobic effect: natural selection of a thermodynamic force in protein folding1. *Journal of Molecular Biology* **2001,** *310* (5), 955-963.
86. Auton, M.; Bolen, D. W., Additive Transfer Free Energies of the Peptide Backbone Unit That Are Independent of the Model Compound and the Choice of Concentration Scale. *Biochemistry* **2004,** *43* (5), 1329-1342.
87. Hong, J.; Capp, M. W.; Anderson, C. F.; Saecker, R. M.; Felitsky, D. J.; Anderson, M. W.; Record, M. T., Preferential Interactions of Glycine Betaine and of Urea with DNA: Implications for DNA Hydration and for Effects of These Solutes on DNA Stability. *Biochemistry* **2004,** *43* (46), 14744-14758.
88. Abbas, S. A.; Sharma, V. K.; Patapoff, T. W.; Kalonia, D. S., Solubilities and Transfer Free Energies of Hydrophobic Amino Acids in Polyol Solutions: Importance of the Hydrophobicity of Polyols. *Journal of Pharmaceutical Sciences* **2011,** *100* (8), 3096-3104.
89. Gu, W.; Rahi, S. J.; Helms, V., Solvation Free Energies and Transfer Free Energies for Amino Acids from Hydrophobic Solution to Water Solution from a Very Simple Residue Model. *The Journal of Physical Chemistry B* **2004,** *108* (18), 5806-5814.
90. Serafin, J. M., Transfer Free Energy and the Hydrophobic Effect. *Journal of Chemical Education* **2003,** *80* (10), 1194.
91. Walter Filgueira de, A., Jr.; Rafael Andrade, C.; Ivani, P.; Luis Fernando, S. M. T.; Guy Barros, B.; Kelen Beiestorf, R.; Milena Botelho Pereira, S., Protein-Drug Interaction Studies for Development of Drugs Against Plasmodium falciparum. *Current Drug Targets* **2009,** *10* (3), 271-278.

92. Kalra, P.; Reddy, T. V.; Jayaram, B., Free Energy Component Analysis for Drug Design: A Case Study of HIV-1 Protease−Inhibitor Binding. *Journal of Medicinal Chemistry* **2001,** *44* (25), 4325-4338.
93. Jones, S.; Thornton, J. M., Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **1996,** *93* (1), 13-20.
94. Brandt, M. E.; Vickery, L. E., Charge pair interactions stabilizing ferredoxin-ferredoxin reductase complexes. Identification by complementary site-specific mutations. *Journal of Biological Chemistry* **1993,** *268* (23), 17126-17130.
95. Hanukoglu, I., Conservation of the Enzyme–Coenzyme Interfaces in FAD and NADP Binding Adrenodoxin Reductase—A Ubiquitous Enzyme. *Journal of Molecular Evolution* **2017,** *85* (5), 205-218.
96. Fredriksson, R.; Schiöth, H. B., The Repertoire of G-Protein–Coupled Receptors in Fully Sequenced Genomes. *Molecular Pharmacology* **2005,** *67* (5), 1414.
97. Ito, T., Nucleosome Assembly and Remodeling. In *Protein Complexes that Modify Chromatin*, Workman, J. L., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2003; pp 1-22.
98. Iftode, C.; Daniely, Y.; Borowiec, J. A., Replication Protein A (RPA): The Eukaryotic SSB. *Critical Reviews in Biochemistry and Molecular Biology* **1999,** *34* (3), 141-180.
99. Lewis, M.; Chang, G.; Horton, N. C.; Kercher, M. A.; Pace, H. C.; Schumacher, M. A.; Brennan, R. G.; Lu, P., Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* **1996,** *271* (5253), 1247.
100. Kornberg, R. D., The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (32), 12955-12961.
101. Christy, B.; Nathans, D., DNA binding site of the growth factor-inducible protein Zif268. *Proceedings of the National Academy of Sciences of the United States of America* **1989,** *86* (22), 8737-8741.
102. Gajiwala, K. S.; Chen, H.; Cornille, F.; Roques, B. P.; Reith, W.; Mach, B.; Burley, S. K., Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* **2000,** *403*, 916.
103. Robinson, H.; Gao, Y.-G.; McCrary, B. S.; Edmondson, S. P.; Shriver, J. W.; Wang, A. H. J., The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature* **1998,** *392* (6672), 202-205.
104. Ralf Jauch, C. K. L. N., Kamesh Narasimhan and Prasanna R. Kolatkar, The crystal structure of the Sox4 HMG domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochem. J.* **2012,** *443*, 39-47.
105. Peters, R.; King, C.-Y.; Ukiyama, E.; Falsafi, S.; Donahoe, P. K.; Weiss, M. A., An SRY Mutation Causing Human Sex Reversal Resolves a General Mechanism of Structure-Specific DNA Recognition: Application to the Four-Way DNA Junction. *Biochemistry* **1995,** *34* (14), 4569-4576.
106. Murphy, E. C.; Zhurkin, V. B.; Louis, J. M.; Cornilescu, G.; Clore, G. M., Structural Basis for SRY-dependent 46-X,Y Sex Reversal: Modulation of DNA Bending by a Naturally Occurring Point Mutation. *J. Mol. Biol.* **2001,** *312* (3), 481-499.
107. 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* (7), 1295-1306.
108. 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* (24), 3217-3230.
109. Lawson, C. L.; Swigon, D.; Murakami, K. S.; Darst, S. A.; Berman, H. M.; Ebright, R. H., Catabolite activator protein (CAP): DNA binding and transcription activation. *Current opinion in structural biology* **2004,** *14* (1), 10-20.

110. Lescrinier, E.; Froeyen, M.; Herdewijn, P., SURVEY AND SUMMARY: Difference in conformational diversity between nucleic acids with a six-membered 'sugar' unit and natural 'furanose' nucleic acids. *Nucleic Acids Research* **2003,** *31* (12), 2975-2989.
111. Lu, X.-J.; Olson, W. K., 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research* **2003,** *31* (17), 5108-5121.

Chapter 2
Methodology

2.1 Molecular Dynamics simulations

Here, in this thesis, all the studies were performed using molecular dynamic (MD) simulations. The MD simulation is an important tool to understand the structural and dynamical properties of any system at the atomistic detail. In classical MD, we generally solve Newton's equation of motion, where forces on atoms are calculated from a pre-parameterized potential function (*Eq. 1*). Forces on each atom lead to displacement with time, known as trajectory. The potential function consists of both bonded (bond, angle and dihedral between atoms) and non-bonded interactions (Lennard-Jones and Coulomb), as shown in Eq. 1.

$$U = \sum_{bonds} \frac{k_{bond}}{2}\left(r - r_{eq}\right)^2 + \sum_{angle} \frac{k_{angle}}{2}\left(\theta - \theta_{eq}\right)^2 + \sum_{torsions} \frac{k_{dihedral}}{2}[1 + \cos(n\emptyset - \gamma)] + \sum_{i>j}\left[\left\{\frac{A_{ij}}{{R_{ij}}^{12}} - \frac{B_{ij}}{{R_{ij}}^{6}}\right\} + \frac{q_i q_j}{\varepsilon R_{ij}}\right], \qquad Eq.1$$

The first three terms of potential function represent the bonded potential and last term is the non-bonded potential. The $k_{bond}, k_{angle}, k_{dihedral}$ are force constants for bonding, angle and dihedral interactions, respectively. n is multiplicity and γ is phase angle for dihedral angles. The equilibrium values of bonds and angle are denoted by r_{eq} and θ_{eq}, respectively. The fourth term of Eq 1. represents the non-bonded interaction present in the system, which is a combination of Lennard-Jones potential (shown in curly bracket) and the Coulomb potential. The, A_{ij}, B_{ij} are the constant in Lennard-Jones potential, q_i, q_j are partial charges on atoms, and ε is the permittivity of medium.

The timescale of the biological process varies from femtosecond (electron transfer in photosynthesis) to seconds (protein folding and DNA protein binding) to weeks (lifetime of the cell). Therefore, to capture the experimental observations in simulations, the MD

simulation should be long enough to match the time scale of the studied system. However, MD simulation uses femtosecond timestep for integration. Therefore, for a realistic system such as protein or DNA, the computational cost for performing long simulations is enormous. The longest timescale that has been simulated for proteins till now is a millisecond, achieved using GPU technique and Anton supercomputer dedicated to run such calculations[1]. However, with the improvement of computational technologies, most of the computational facilities can manage to perform only up to a microsecond long simulation. Hence, researchers have developed some enhanced sampling techniques like umbrella sampling, metadynamics, parallel tempering, etc. to investigate high barrier processes like protein folding, ligand binding, drug/protein intercalation in DNA, etc.

In this thesis, we have investigated several high barrier processes such as protein folding, the effect of a small molecule in protein folding, the role of amino acid in DNA kinking, and binding mechanism of protein to DNA. All these are barrier processes and therefore, we have enhanced sampling technique like an umbrella sampling and metadynamics methods to explore the phase space and calculate the free energy surface.

2.2 Enhanced Sampling Techniques

(I) Umbrella sampling simulations

By principle, the free energy of the states can be obtained by the counting population of that state (*Eq. 2*)

$$G = -k_b T \ln p_i \qquad Eq. 2$$

Where p_i is the population of the i^{th} state.

However, in some process, the free energy states are separated by a big barrier, which is less likely to be sampled in normal MD simulation. Therefore, we have used accelerated methods such as umbrella[2] and metadynamic[3-4] to enhance the sampling.

In these accelerated methods we apply a bias potential along the reaction coordinate. For example, in umbrella sampling, we apply a harmonic bias potential (*Eq 3*) along with a reaction coordinate (assuming that the other degree of freedom is relaxed). So more precisely, the reaction coordinate of the process is divided into several segments (called "windows") and a harmonic bias potential is applied on the reaction coordinate to go from one window to another window. The schematic of umbrella sampling is shown in Fig 1. The applied a harmonic potential has a form of Eq. 3 along a chosen reaction coordinate.

$$V_i^{umb}(s(r)) = \frac{k_i}{2}\left(s(r) - s(r)_i^{ref}\right)^2, \qquad Eq.3$$

where, k_i is force constant for i^{th} window, $s(r)_i^{ref}$ is the constant reaction coordinate value of i^{th} window and $s(r)$ is the dynamical changing reaction coordinate.

Further, through umbrella sampling, we get the population of reaction coordinate along the umbrella window but this population is biased by the applied potential. Hence, we need to unbias the population to obtain the correct free energy of the i^{th} window using the following equation.

$$\tilde{F}_i(s) = -\frac{1}{\beta} ln\, N_i(s) - V_i^{umb}(s) \qquad Eq.4$$

where N_i is population of the i$^{\text{th}}$ state.

Finally, we used the weighted histogram analysis method (WHAM) [5] to join the piecewise free energy values to obtain the free energy surface of the whole process.

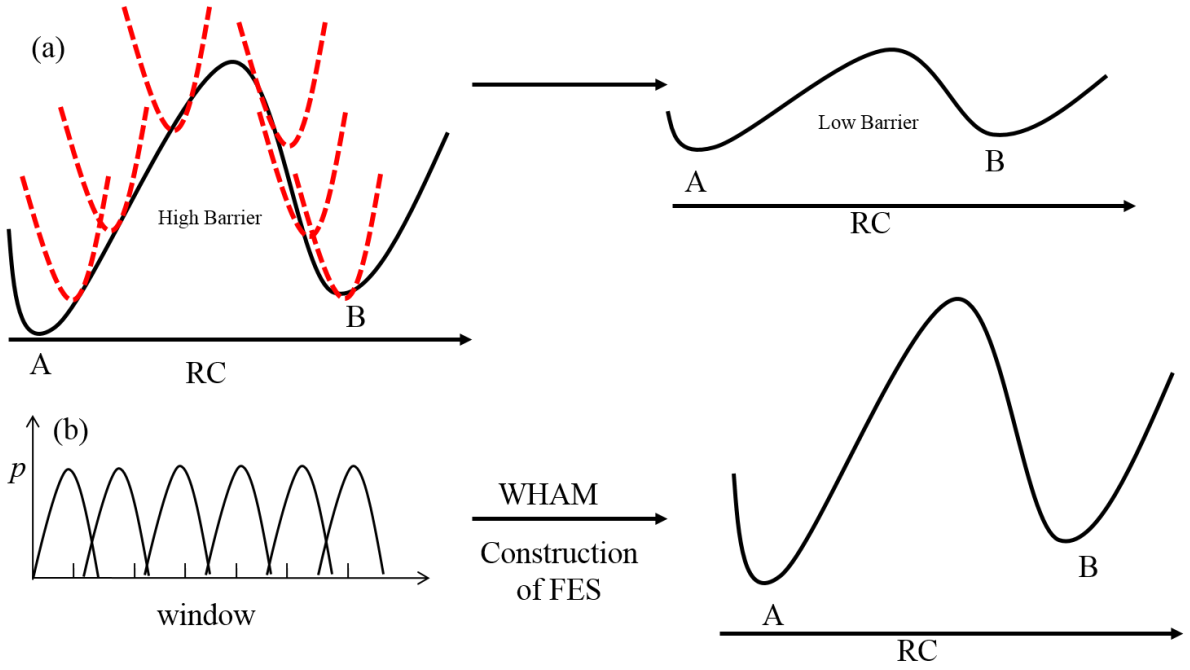

Figure 1. Schematic representation of umbrella sampling. (a) black color smooth line represents the FES (which we don't know priorly) for process A to B. The red dashed curved line represents the applied umbrella bias potential for each window. Right side of Fig 1a shows the FES after applying the bias potential. The applied potential decrease the FE barrier. (b) the left side of Fig shows the population of each window and the right side shows the FES constructed by using the WHAM.

(II) Metadynamics Simulation.

Umbrella sampling, although a rigorous and accurate enhanced sampling approach, it is computationally demanding for multidimensional free energy calculation. Metadynamics is a relatively recent advanced enhanced sampling technique to calculate multidimensional free energy surfaces.[3] Unlike umbrella sampling, metadynamics is a non-equilibrium, non-directional sampling technique. However, similar to umbrella sampling, metadynamics simulations also require reaction coordinate to bias the trajectory. We have discussed about the reaction coordinate in next section. The primary aim of metadynamics is to discourage the system to visit the same region of phase space once a second time once it had been covered earlier. This helps the system explore the free energy surface (FES) efficiently. The way it is achieved is by depositing a history dependent Gaussian potential at certain time interval to the

reaction coordinate. The potential is added until the FES becomes flat and the system is free to move anywhere in FES. We can control the speed of sampling by changing the width and height of deposited Gaussian potential as well as by changing the deposition rate.

$$V_G(s(x),t) = w \sum_{t'=t_G} exp\left(\frac{(s(x) - s(x_G(t'))^2}{2\partial s^2}\right), \quad Eq.5$$

where w is the height of the Gaussian, ∂s is width of Gaussian, t_G is frequency at which the Gaussian potential is deposited and $t < t_G$. $s(x)$ is the conformation space of the collective variables (CVs). $s(x_G(t')$ is the value of CV at time t' and t is the total sampling time.

The free energy is calculated by the formula

$$F(s) \sim -\lim_{t\to\infty} V_G(s(x),t) \qquad Eq.6$$

(III) Well-Tempered metadynamics Simulation.

Despite the popularity and success of metadynamics, there are a few drawbacks to this method, one of them being the convergence of free energy. In metadynamics, a Gaussian potential of constant height is added. As a result, the estimated free energy oscillates around the real value and does not converge to the exact value leading to an error in the free energy calculation. However, this problem has been solved by applying a Gaussian potential in which the height of potential is scaled (more precisely decreased) with time. Thus, in the case of well-tempered metadynamics[4], one applies a Gaussian potential whose height is scaled by the following equation.

$$w = w_0\, e^{-\frac{V_G(s(x),t)}{k_B \Delta T}} \qquad Eq.7$$

Here w_0 is the initial Gaussian height, k_B is the Boltzmann's constant, ΔT is the parameter with the dimension of temperature.

The FES is constructed by the following formula.

$$\tilde{F}(s(x),t) = -\frac{T + \Delta T}{\Delta T} V_G(s(x),t), \qquad Eq.8$$

where $\frac{T+\Delta T}{\Delta T}$ is CV temperature and $\frac{T+\Delta T}{\Delta T} = \gamma$, refferead as bias factor, T is simulation temperature.

2.3 Analyses Techniques

(I) Minimum Free Energy Path Calculation

A reaction or a process in general always follows the lowest free energy path (MFEP) on the multidimensional FES. Hence, it is very important to compute the MFEP to understand the mechanism of the process. Here, in this thesis, we have used an algorithm proposed by Ensing *et al*. algorithm[6] to compute the MFEP on the calculated FES. The algorithm is briefly discussed below.

Consider an FES (shown in Fig) made of two collective variables (CVs) S1 and S2. We would like to calculate the MFEP from reactant well (denoted by A in Fig) to the product well (denoted by B and C). There are many possible paths which would connect A to B. However, to calculate the MFEP, the first step is the identification of the most stable reactant and product points on the FES. This is done with the golden search method. [6] Once the minima are defined at both wells (reactant and product), we would start the minimization of potential V (Sα) from the first minima by choosing a range between Sα-δSα and Sα+δSα such that the $V(S_\alpha) < V(S_\alpha - \delta S_\alpha)$ and $V(S_\alpha) < V(S_\alpha + \delta S_\alpha)$. Thereafter, the golden search method is used to find a new point S'_α between the $S_\alpha$and the bracket $S_\alpha \pm \delta S_\alpha$end such that the potential should be lowest.

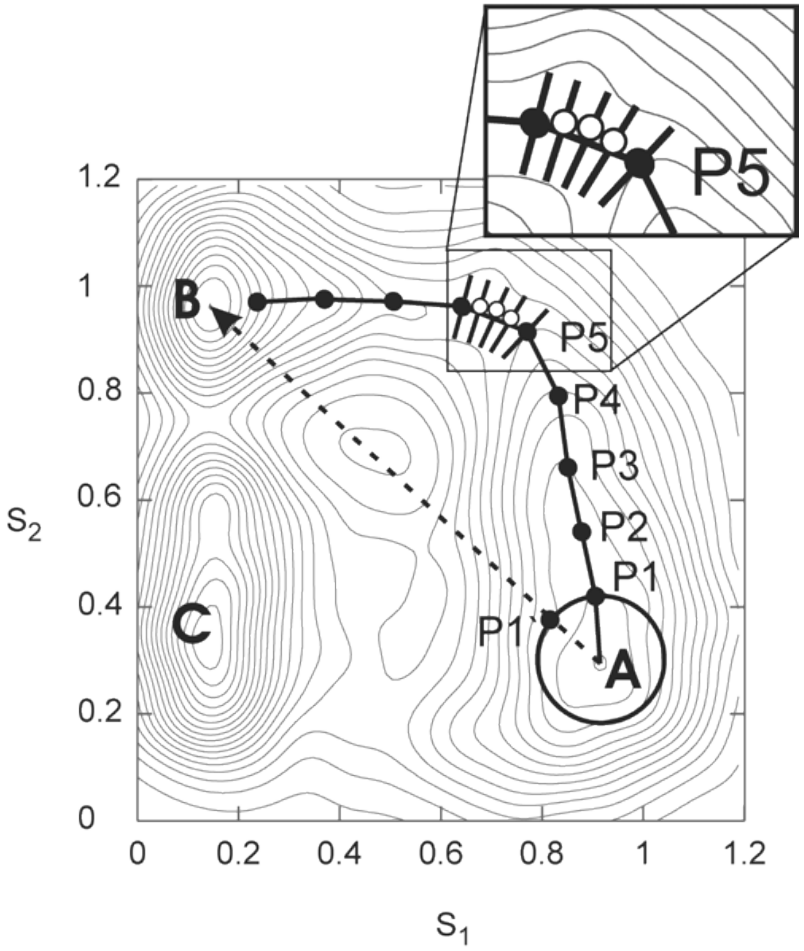

Figure 2: The free energy surface made from two CV S1 and S2. Calculation of MFEP on this 2D FES. Copyright@ 2014 American Chemical Society. [6]

Computing the MFEP from A to B happens in two steps. (i) A coarse path is constructed from A to B (shown as P1, P2, P3 in Fig 2) by using the golden search method. The step size for points should be greater than the local bumpiness in free energy surface. Thereafter, the path is reconstructed by minimizing the path by the method described above. This new path would be the desired minimum free energy path.

2.4 Reaction Coordinate (RC)

In the accelerated techniques we apply the bias potential along a particular reaction coordinate (RC), which defines the extent of reaction and depicts the mechanism of the process. Choosing a reaction coordinate is completely dependent on the process. For example, to dissociate a NaCl, a distance

between Na and Cl would be best RC. Another example is that for cis-trans conversion of a molecule, a dihedral angle would be the best RC.

For a simple process such as dissociation of NaCl, choosing an RC is simple, but for a complex process such as protein unfolding, DNA-protein binding choosing an RC is very difficult and sometimes more than one reaction coordinates are needed to describe the process completely.

We have used several reaction coordinates in the studies. We will briefly discuss the about the RCs here. Details will be given in the corresponding chapters.

(I) Native Contact (Nc):

Native contact is a commonly used collective variable in protein folding studies. It is defined by the through space proximity of groups of atoms in the native state. The native contact between the one groups ($g1$) of atom to the other group ($g2$) of atoms is defined as,

$$N_c = \sum_{i \in g1} \sum_{j \in g2} s_{ij}\,, \qquad Eq.9$$

where, s_{ij} is given by,

$$s_{ij} = \begin{cases} 1 & r_{ij} \leq 0 \\ \dfrac{1-\left(\frac{r_{ij}}{r_0}\right)^n}{1-\left(\frac{r_{ij}}{r_0}\right)^m} & r_{ij} > 0 \end{cases} \qquad Eq.10$$

and $r_{ij} = |r_i - r_j| - d_0$. The user defined parameters were chosen to be n=6, m=12, r_0=4.5 Å and d_0=0 Å. Above equation ensures the variation of s_{ij} is continous and differentiable. Schematic of Nc is shown in Fig 3.

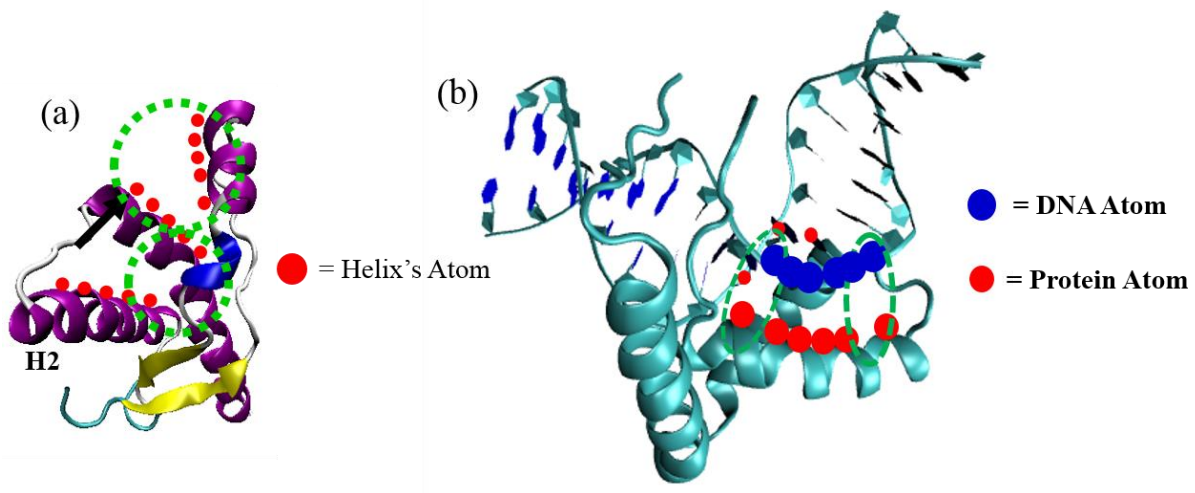

Figure 3. Shows the schematic representation of native contact (Nc). The color circled represents the atom corresponding to a segment. The green color represent the distance circle on which we count the contact. (a) represents that how the native contact define in protein. The red color dot represents the atom of helixes. When the atom of one helix is in within 4.5Å (dotted green circle) of other helix atom, we considered it a contact. Same way (b) represents the contact define in DNA-Protein complex. The blue dot color represents the atom of DNA and red dot color represents the atom of protein. When an atom pair (DNA and protein atoms) is within 4.5A (green dot circle), we count as a contact.

(II) Radius of gyration (R_g). Radius of gyration is another measure of the extent of folding of proteins. A folded protein typically has lower R_g (Fig. 4) defined as, $R_g = \sqrt{\sum_i \frac{m_i(\vec{r_i}-\vec{r_c})^2}{M}}$. m_i is mass and $\vec{r_i}$ is the position vector of the i^{th} atom. M is the total mass ($= \sum_i m_i$) and $\vec{r_c}$ is the center of mass.

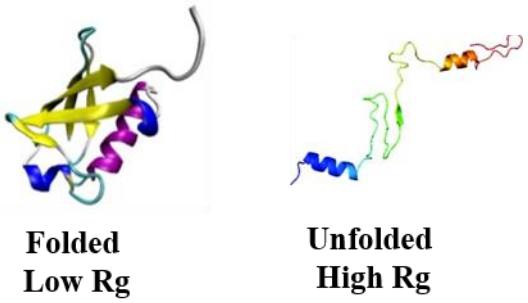

Figure 4. Schematic representation of Rg. The Folded protein will have less Rg and the unfolded protein will have more Rg.

(III) Average Psi angle (ψ_{ave}). The conformation of protein is largely dependent on two backbone dihedral angles φ (C-N-Cα-C) and ψ (N-Cα-C-N). While ψ angle for α-helix has -40°, it is 120° for a

β-sheet, with φ being almost same in both (Fig 5). Since in a particular protein, there are many ψ angles (one for each amino acid), we considered average ψ angle, $\psi_{ave} = 1/N \sum_i \psi_i$, as a reaction coordinate to monitor α-helix to β-sheet conversion.

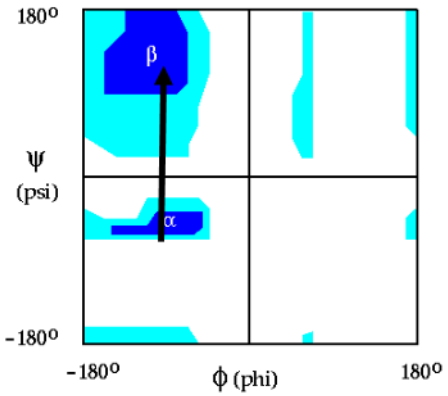

Figure 5. The Ramachandran plot of protein backbone dihedral angle. The ψ angle drives the secondary change in the protein. Since there are more AA in protein, therefore, we have taken the average ψ for probing the secondary structure of the protein.

(IV) Distance: To separate any fragment from another, we have used the distance between the center of masses of two different fragments.

(v) Distance along a vector: Distance is a radial coordinate and does not differentiate between different binding modes of protein/drug to DNA. We have defined a new RC, which measures the distance of one group along a particular body fixed vector. This allows directional approach of a particular group (such as amino acid) to another (such as DNA base pairs). More detail of this reaction coordinate is given in chapter 5).

References

1. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S. etc. Anton, a special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th annual international symposium on Computer architecture*, ACM: San Diego, California, USA, 2007; pp 1-12.
2. Kästner, J., Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011,** *1* (6), 932-942.
3. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99* (20), 12562-12566.
4. Barducci, A.; Bussi, G.; Parrinello, M., Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **2008,** *100* (2), 020603.
5. 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. *Journal of Computational Chemistry* **1992,** *13* (8), 1011-1021.
6. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press: 2007; p 1256.
7. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009,** *180* (10), 1961-1972.

Chapter 3
Mechanism of Prion Protein Unfolding

3.1 Overview

Misfolding or disorder in prion protein causes serious neurodegenerative diseases. Thus, it is very important to understand the mechanism of prion protein misfolding and aggregation. Researchers have speculated that first the cellular prion misfolds to a state PrP^{C*} and then aggregates to form PrP^{Sc}. Here, we have studied the misfolding of prion protein using the well-tempered metadynamics technique. With the help of multiple microsecond simulations we show that the unfolded (less structured prion protein) state is more stable than the native state and there are several pathways for unfolding of the prion protein.

So, our study's main point is that the native PrP^{C} is an intrinsically disorder protein (IDP) (with more secondary structure content compared to other IDPs) and a small perturbation (like, change in pH, a point mutation, change in salt concentration, etc.) would lead to a more stable unfolded conformation. We believe that this is the unfolded structure is somewhat related to the PrP* state, an intermediate state prior to the formation of PrP^{Sc}. Moreover, we have not observed any toxic β-rich conformation during the unfolding pathway of the PrP^{C}. Thus, we believe that the PrP^{Sc} with parallel β-sheets must involve from dimer or higher oligomers. Further, simulation over isolated helices from PrP^{C} shows that the individual helices especially (H2 and H3) prefer to be in a random coil or in β-sheet form rather than being in helices. So, based on our analysis we conclude that the secondary form of helices in prion protein is stabilized due to the tertiary contact between the helixes when they are together.

This study bridges the gap between earlier theoretical predictions to the current experimental endeavour to understand the origin of prion diseases.

3.2 Introduction

Prion protein is a glycosylated, glycosylphosphatidylinositol-anchored component of the extracellular surface of neurons and helps in signal transduction.[1] It is the most abundant protein in neurons of the brain and spinal cord.[2] Apart from the brain, it is also sparingly present in glial cells of the central nerve system (CNS) and peripheral cells types.[2-3] Most of the prion proteins are found at lipid bilayer of the cell surface.[4-5]

The physiological function of the prion protein is still unknown. It was found that the cleavage of prion proteins activates the process of myelin repair in Schwann Cells, whereas lack of prion protein causes demyelination of those cells.[6] Apart from these, prion also protects cells from the oxidative stress,[7-8] helps in copper transportation,[9-14] cell adhesion,[15-17] and plays a significant role in transmembrane signaling.[18-22]

Conversion of a normal, cellular isoform (PrP^{C}) into a conformationally different "scrapie" isoform (PrP^{Sc}) results in prion-related neurodegenerative diseases, especially when nucleic acids are not present. PrP^{C} converts to PrP^{Sc} by an endosomic post-translational process.[23-25] Chemical means of probing the difference between PrP^{C} and PrP^{Sc} were unsuccessful because the two forms are chemically identical.[26-27]

The "prion hypothesis" states that the infection relies solely on a self-poisoning mechanism involving protein's conformation. According to this hypothesis, the α-helix rich PrP^{C} state is converted to the infectious PrP^{Sc} isoform characterized by its high content of β sheet.[25, 28] The FT-IR and CD spectroscopy have shown that PrP^{Sc} has a high (43%) β-sheet content and a significant (30%) α-helix content. In contrast, PrP^{C} has only 3% β-sheet and 42% α-helix as measured by FTIR.[23]

PrP^{Sc} differs from PrP^{C} in many ways: (i) PrP^{Sc} is insoluble in detergents, whereas PrP^{C} is readily soluble.[29] (ii) While PrP^{C} gets completely degraded by protease, PrP^{Sc} is proteases resistant.[30] (iii) PrP^{Sc} aggregates at the brain surface, whereas PrP^{C} does not.[31] (iv) The accumulation of PrP^{Sc} in the brain forms a distinct pattern[32] compared to the distribution of PrP^{C}.

The three-dimensional structure of cellular prion protein (PrP^{C}) has been solved for many species using both X-ray crystallography and solution NMR.[33-34] The human Prion protein (huPrP) contains a unstructured N-terminus (Nter) region containing amino acids (AA) 23 to 124, and a well structured C-terminal (Cter) region (AA 125−228) composed of three α-helices. Helix-1 (H1) contains AA from 144 to 156, Helix-2 (H2) includes AA from 174 to 194, and Helix-3 (H3) comprises AA from 200−228. It also has a small β-sheet region containing AA from 128−131 and 161−164.[33,35] Residues C179 and C214 and two N-glycosylation sites (N181 and N197) form disulphide bonds in prion protein. The N-terminal residues 51−91 consist of an unstructured random coil, but acquire secondary structure in the presence of Cu^{++} and other metal ions.[22, 36] It has been found that the N-terminal of prion protein also helps in binding with the cell membrane.[37] The extensive sequence homology shows that the sequence of PrP^{C} is highly conserved amongst mammals, and only differs slightly in birds, reptiles, and amphibian.[38]

Evidence of formation of PrP^{Sc} trimeric assembly was observed in electron microscopy, which provided some structural constraints for the PrP^{Sc} monomer structure.[39] Because of insolubility and flexibility of PrP^{Sc}, the crystal structure of PrP^{Sc} is still unsolved. However, based on the threading[40] and Raman optical[41] experiment data, two different structural models of PrP^{Sc} emerged. The first model comprises two α-helices and a multi-stranded antiparallel β-sheet, which grows on a prior short β-sheet of the PrP^{C}.[42-43] However, these models have a left-handed parallel α-helix structure.[40, 44] In spite of these proposed models a clear understanding of prion conversion still eludes us.[45]

Since PrP^{Sc} has a significant β-sheet region and PrP^{C} has a predominant α-helical region, in PrP^{C} to PrP^{Sc} transformation, some α-helical region must convert to β-sheet. Therefore, the stability of the helical region and its propensity to switch to β-sheet are the key events in prion protein conversion.

Eghain *et al.* attached antibodies with the helices to monitor the conformational change of protein using NMR. They observed that H2 and H3 were stable, whereas the H1 was the one that was prone to β-sheet formation.[46] However, Thirumalai *et al.* have used a combination of sequence analysis and molecular dynamic simulation to show that H1 is unusually stable due to salt bridge formation between the *i-th* and *(i+4)-th* residues whereas H2 and H3 have a reduced helical propensity due to their unusual sequence.[47] Chakroun *et al.* also performed molecular dynamic simulation and showed that the H2 and H3 are the ones that get converted into the β-sheet rich conformation whereas H1 remains in its native conformation.[48] They also found that the unfolded or the β sheet rich conformation of H2 and H3 is the intermediate for the oligomerization of the protein. Molecular dynamic simulations at high temperature and low pH also show that the H2 and H3 unfold and convert to the β-sheet structure initiating oligomerization while H1 remains in its native state.[48-49] The idea of H2 and H3 being the initiation sites of prion misfolding has been proved experimentally by Udgaonkar *et al.* and other groups where they have shown that the increase in the helical stability of H2 prevents the misfolding and oligomerization in prions leading to a slowdown in the amyloid formation.[50-52] Honda *et al.* also have shown that the partially unfolded equilibrium states, where H3 is unfolded, are the precursors for the aggregation of the protein.[53]

Interestingly, Thirumalai and coworkers suggested that there exists an intermediate conformational state PrP^{C}*, which might be even more stable than the PrP^{C} form due to the inherent structural frustrations in the H2 and H3 regions.[54] Using steered MD simulations they calculated an unfolding barrier of >20 kcal/mol. This high barrier may be the reason for absence of a spontaneous misfolding under natural conditions.[55] Experimental studies by Cohen and coworkers[56] render support to this possibility that the native α-helical conformation is formed under kinetic control, whereas there might exist other

thermodynamically stable and possibly β-rich isoforms. Therefore, the pathway of conversion of PrP^{C} to PrP^{Sc} form can be quite complicated with multiple intermediate states with varying thermodynamic/kinetic stability. The unfolding mechanism of prion (PrP^{C}) and the source of the stability in predominantly helical form is still not clearly understood.

The conversion of PrP^{C} into PrP^{Sc} and aggregation, believed to be the cause of prion propagation,[23-24] is very much dependent on the environmental conditions,[57-59] e.g. pH, salt concentration, lipid binding, metal ions binding, etc. PrP^{C} forms β-rich misfolded oligomers at low pH and its tendency to form misfolded conformation increases with the decrease in pH.[50, 60] Diogo *et al.* have performed constant pH molecular dynamic simulations and showed that the protein forms a stable misfolded conformation at pH 2 and some of the conformations are reversible.[49] Campos *et al*. also showed that H2 and H3 are unstable at low pH.[61]

Similarly, it has been observed that the conformation of PrP^{C} changes to a β-rich structure in the pH range of 4.4–6.[62-64] Thus, low pH may can initiate the conformational change of PrP^{C} to PrP^{Sc}.

Conformational changes in prions due to changes in pH have also been studied by molecular dynamic simulation.[42, 63, 65-66] Protonation of His, Asp, and Glu in different numbers have been used to emulate a particular pH range indirectly. Garrec *et al.* used one such set up to run microsecond long simulations which highlighted the effect of electrostatic repulsion experienced by the protonated imidazole ring of H187 due to the nearby guanidinium group of R136.[67] The aforementioned effect led to mouse PrP^{C} misfolding which was argued to follow any one of the two pathways: (i) the unravelling of H2 (C-terminal) alone or the unravelling of H2 with simultaneous elongation of β sheet.[68] An acidic pH mimicked by protonation of histidine residues in human prion protein was also studied by Langella *et al.* using MD simulations. According to this study, His-155 and His-187 prevent copper binding and are directly involved in conformational rearrangements.[63] Diogo Vila-Viçosa *et al.* suggested that the pH-dependent conformational changes usually surface in the C-terminus core while the N-terminus region,

even though flexible and transiently forming β-structure, is less influenced by pH. In a later investigation, they also proposed various conformational intermediates that could be formed during PrP^C misfolding with decreasing pH along with a general propensity for reversibility regarding refolding to its cellular form with an increase in pH.[49, 61]

Prion conversion is very much affected by the point mutation. The familial prion diseases such as GSS (mutation at AA 102, 105, 117, 145, 198, and 217), fatal familial insomnia, and Creutzfeldt-Jacob (mutation at AA 178, 200, 210) diseases are related to distinct point mutants within the human gene of PrP^C.[69-73] The experiments with Ala117Val mutation showed an increasing tendency of the β-sheet region.[74] On the same line Chabaro *et al.* show that T183A mutation accelerates the prion conversion in both monomer and dimer forms[75]. In contrast, Barducci *et al*. have shown using molecular dynamics simulation that D178N mutant prevents anti-parallel β-sheet formation further compared to the wild type.[29] This destabilization was attributed to the disruption of hydrogen bonding network of the mutated residue with Arg164 and Tyr128 side chains.[66, 76] Therefore, such networks may be responsible for stabilizing the antiparallel β-sheet in the PrP^C. These findings suggest that the pathogenic conditions may favour disruption against stabilization and growth of the antiparallel β-sheet.[53]

Given the above discussion, it is quite evident that the conformational energy landscape of prions would be highly complex. Based on the molecular simulations and experiments, some mechanistic models have been proposed for the oligomerization. Among all the models, three are the ones which fit on the experimental observation: (i) the β sheet model [26, 39-40, 58], (ii) the spiral model[43], and the parallel in register β sheet model.[77-79] None of these models, however, has been able to provide a wholesome picture of the actual mechanism of prion oligomer formation or structure of PrP^{Sc} form satisfactorily.[45]

Understanding the complete conformational free energy landscape of prions would give us a direct picture of the thermodynamic and kinetic stabilities of various isoforms. Thus, the objective of the present study

is to find out the free energy surfaces of the unfolding of the prion protein (wild type) under normal conditions. Moreover, we also investigate the molecular origin of the inherent instability of prions and their propensity to misfold in terms of the stability of the secondary structural elements present in the PrP^{C} form. We show that neither the helices individually nor the prion protein as a whole is the stable state in solution and there exists a relatively compact unfolded state with lower free energy than even the native PrP^{C} form. This intermediate unfolded state might corresponds to the widely discussed PrP^{C*} state, which may act as the precursor to further oligomerization into the PrP^{Sc} and fibrillar forms.

3.3 Design of the Study & Methodology

To understand the mechanism of unfolding, we have carried out six well-tempered metadynamics (WTMD) simulations using three collective variables and three different force fields -- AMBER99SB[80] (four simulation), CHARMM27[81] (one simulation), and AMBER03[82] (one simulation). We discuss the design, methodologies and analyses protocols below.

(I) Preparation of Initial Configurations. The crystal structures of the human prion protein (PDB ID 1HJM)[34] were taken from the Protein Data Bank (PDB).[83-84] The topology and initial coordinates were generated by the GROMACS software.[85-86]

(II) Equilibration. The prepared system was put in a large cubic box of length 75 Å. TIP3P water model[87] was used to solvate the system, and ions were added to neutralize and mimic physiological (150 mM) salt concentration. The energy minimization using steepest descent algorithm[88] was performed, followed by heating up to 300 K using the Berendsen thermostat[89] with a coupling constant of 0.2 ps. During the heating process, the heavy atoms of protein were restrained with the harmonic potential of 25 kcal/mol/$Å^2$. Following this, a series of equilibration and energy minimization were carried out, where the restrain was reduced to 0.25 kcal/mol/$Å^2$ in six steps. Finally, an unrestrained 1 ns constant temperature (at 300 K) and pressure (1 bar) equilibration was performed using the Nose-Hoover

thermostat[90-91] and Parrinello-Rahman barostat[92], respectively, with coupling constant 0.4 ps for each. We used Particle Mesh Ewald (PME)[93] method for the electrostatic interaction with a long-range cutoff of 10 Å while the van der Waals interactions were treated with 10 Å cutoff. Integration time step for all simulations was kept at 2 fs.

(III) Metadynamics Simulation. Following equilibration, we have performed multiple well-tempered metadynamics[94] simulations (with different initial conditions) to explore the free energy surfaces of the Prion protein unfolding using AMBER99SB force field[80]. In the first simulation (SIM1), we have performed a long 1.8 μs metadynamics simulation (referred to as SIM1 hereafter) with hill height 0.2 kJ/mol, bias factor 10, with the hills deposition rate as 2 ps. Second simulation (SIM2) was performed for 800 ns with bias factor 15, hill height 0.2 kJ/mol and the hills deposition rate 3 ps, the third simulation (SIM3) was performed for 600 ns using bias factor 15, hill height 0.2 kJ/mol and the hills deposition rate 4 ps. The fourth simulation (SIM4) was performed for 400 ns using bias factor 15, hill height 0.2 kJ/mol and the hills deposition rate 3 ps and with a different initial velocity distribution. Gaussian widths for N_c, R_g, and ψ_{ave} were taken to be 5, 0.7 Å, and 0.05 radian, respectively, for all the three simulations. Along with the above four well-temprered metadynamic simulations, we have performed two more WTMD simulations with two different force fields: (i) a 900 ns simulation with CHARMM27 force field[95] using hill height 0.2 kJ/mol, bias factor 15 and disposition rate 3 ps, and (ii) a 300 ns simulation with amber03 force field[96] using hill height 0.2 kJ/mol, bias factor 15 and deposition rate 1 ps. All the simulations were carried out using GROMACS[85-86] (version 4.5.4) molecular dynamics program. For free energy calculations, we used the PLUMED[97] (version 1.3) with in-house modifications to incorporate a new collective variable used here.

3.4 Definition of the reaction coordinate. The unfolding of protein occurs in an immensely complex multi-dimensional energy landscape. Therefore, the choice of a limited number of reaction coordinates, known as collective variable (CV) in metadynamics simulation, is based on heuristic understanding of the

system. After trying out several different CV's, we have used the three following CV's to study the unfolding process: (I) number of native contacts (N_c), (II) radius of gyration (R_g), and (III) average psi angle (ψ_{ave}). Definitions of the CV's and the reason for their choices are mentioned below.

(I) Native Contact (N_c). The definition of N_c is given in section 2.4 of chapter 2. Here, we have chosen secondary structures (H1, H2, H3 and β-sheet) of protein for native state calculation. So, if an atom belonging to one secondary structure segment comes within 4.5 Å of another atom of a different secondary structure segment, a contact will be formed. N_c is helpful to separate the helices and β-sheet from each other's contact to drive unfolding.

(II) Radius of gyration (R_g). Radius of gyration is another measure of protein folding. A folded protein typically has lower R_g defined as, $R_g = \sqrt{\sum_i \frac{m_i(\vec{r_i}-\vec{r_c})^2}{M}}$. m_i is the mass and $\vec{r_i}$ is the position vector of the i^{th} atom. M is the total mass ($= \sum_i m_i$) and $\vec{r_c}$ is the center of mass.

(III) Average Psi angle (ψ_{ave}). The conformation of protein is largely dependent on backbone dihedral angles φ (C-N-Cα-C) and ψ (N-Cα-C-N). While ψ angle for α-helix has -40°, it is 120° for a β-sheet, with φ being almost same in both. Since in a particular protein, there are many ψ angles (one for each amino acid), we considered average ψ angle, $\psi_{ave} = 1/N \sum_i \psi_i$, as a reaction coordinate to monitor α-helix to β-sheet conversion.

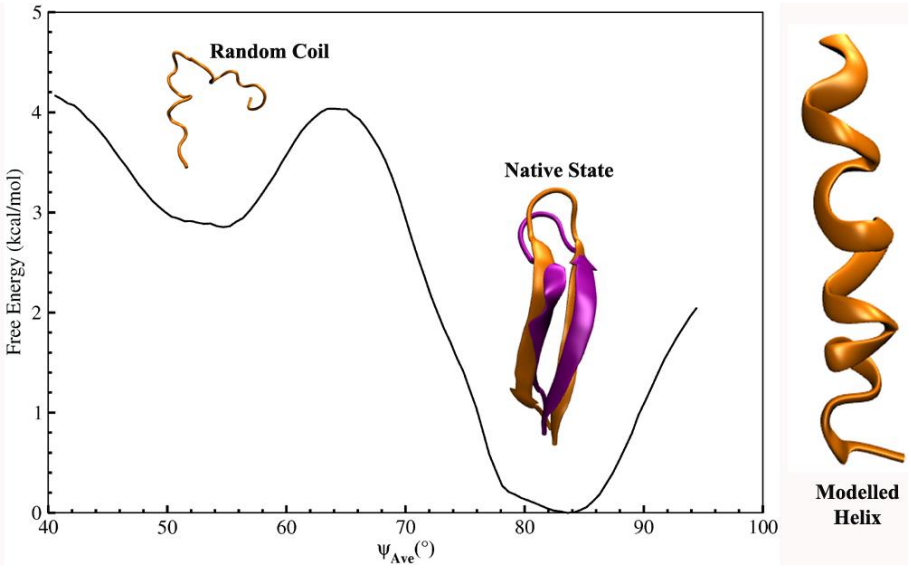

Figure 1. Free energy profile of a GB1 protein using ψ_{ave} as a reaction coordinate. We modeled the sequence initially in the helical form (right side). The normal MD simulation resulted in a random coil structure corresponding to ψ_{ave}~54˚. Thereafter, we have used $\psi_{ave}$and performed umbrella sampling and obtained the free energy profile. The most stable state (purple color) corresponds to a beta structure in close agreement with the native state (orange color). The RMSD of the stable state found here is 1 Å from the native state. The value of the reaction coordinate at the free energy minimum (83˚) is close to that of the native state (84˚).

Before using ψ_{ave} as the reaction coordinate, we tested the capability of this coordinate to create β-sheet conformation. For this we chose the β-hairpin region of the GB1 protein (PDB ID 1pgb).[83-84] We have modelled the initial structure of the protein in the helical conformation, put it in a box of 55 Å, solvated using TIP3P water model[87] and added ions to create a physiological ion concentration (150 mM). Then we performed a normal molecular dynamics simulation at 300K and 1 bar pressure Nose-Hoover thermostat[90-91] and Parrinello-Rahman barostat[92] for 2ns after usual energy minimization and equilibration steps using GROMACS software.[85-86] Simulation details not mentioned are similar to the other systems. Normal MD itself provided a random coil structure corresponding to ψ_{ave}~45˚. Then we performed a series (24 simulations of 4 ns each) of umbrella sampling[98] simulations using ψ_{ave} and calculated free energy profiles using weighted histogram analysis method (WHAM)[99]. The free energy profile is shown

in Fig. 1. The lowest free energy corresponds to a β-hairpin structure which has remarkable similarity with the native state (1 Å RMSD). The free energy of unfolding of β-hairpin found using ψ_{ave} (4 kcal/mol) is similar to that obtained using other reaction coordinate (3 kcal/mol)[100-101]. This initial test indicates that ψ_{ave} is capable of finding the β-sheet if it represents a low free energy state.

(IV) Minimum free energy path (MFEP). We calculated the MFEP using the algorithm of Ensing, *et al.*[102] We gathered the structures that fall within a small deviation (N_c, R_g, and ψ_{ave}) around each point, spaced to avoid overlap, along the MFEP. Then we performed cluster analysis of the structures belonging to each point using a root mean square deviation (RMSD) cutoff of 1 Å. All subsequent calculations of average properties were performed using the members of the biggest cluster to avoid spurious contributions from unnatural fluctuations.

3.5 Study of the isolated helical fragments. As mentioned above, prion protein has three helices and one antiparallel β-sheet. To check the stability of the helices, we have isolated each helical fragment H1, H2, and H3 keeping the terminals in zwitter-ionic form and standard protonation states of the amino acids. Initial configurations were obtained from the crystal structure. Each fragment was solvated with water molecules using TIP3P model[87] in a cubic box of length 65 Å. Ions were added to neutralize and maintain physiological (150 mM) ions concentration. All the simulation parameters (unless stated) are same as mentioned for the whole protein. After energy minimization, position restraint simulation, and equilibration for 1ns, we have performed WTMD simulation of each fragment using ψ_{ave} as the collective variable to find the free energy profile of each. The hill height, deposition rate, and bias factor for metadynamics run were taken to be 0.05 kJ/mol, 3 ps, and 4ps, respectively. The Gaussian width chosen was 0.05 radian. We have performed 500 ns, 310 ns and 600 ns metadynamics runs for H1, H2, and H3, respectively.

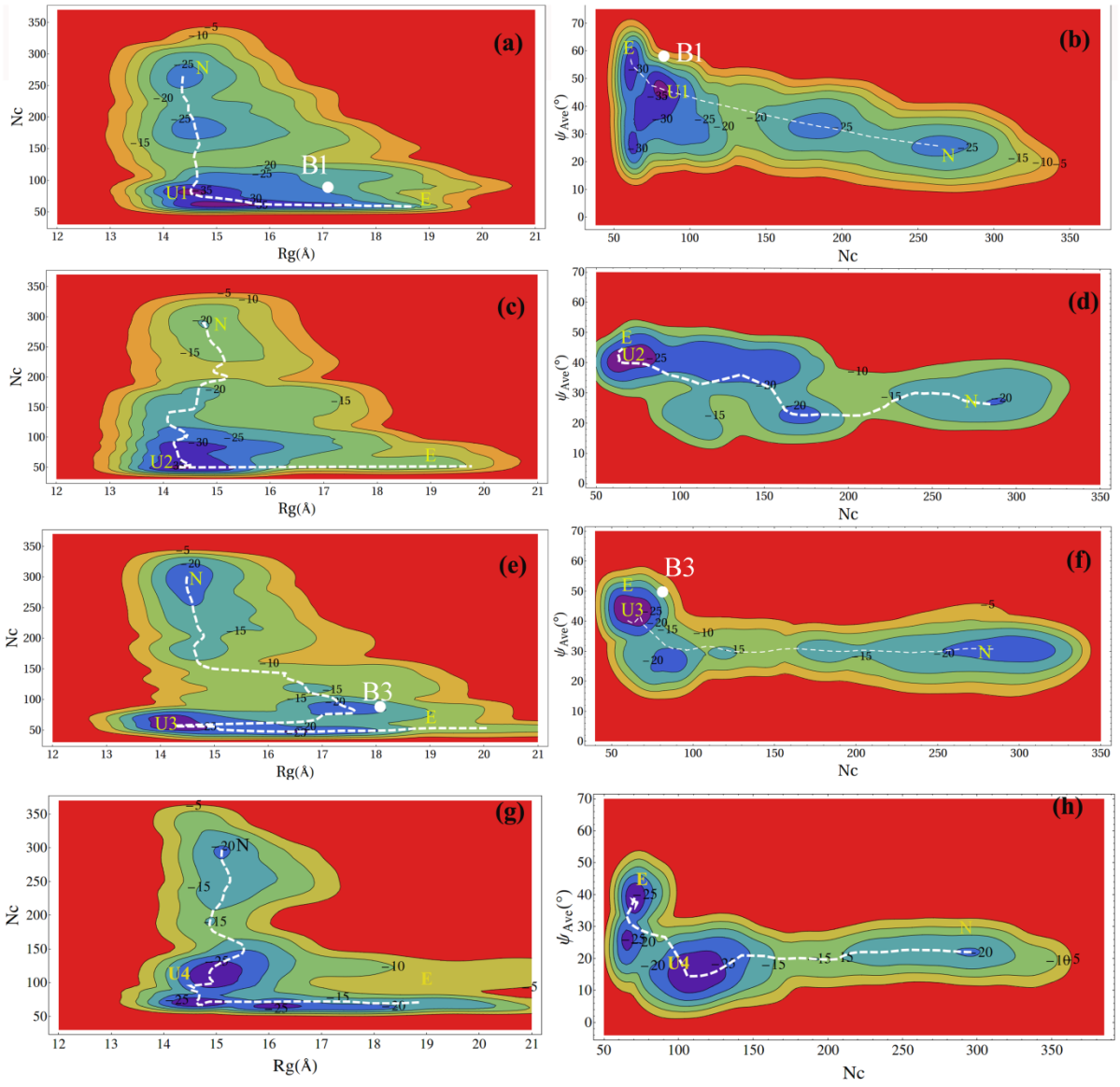

Figure 2. Free energy surfaces (FES's) of prion protein unfolding. Left column shows the FES against N_c and R_g, with the third CV (ψ_{ave}) being Boltzmann averaged for (a) SIM1, (c) SIM2, (e) SIM3, and (g) SIM4. Right column shows the FES against ψ_{ave} and N_c, with the third CV (R_g) being Boltzmann averaged for (b) SIM1, (d) SIM2, (f) SIM3, and (h) SIM4. White lines represent the calculated minimum free energy path in each case. The native and extended states are labeled as "N" and "E", respectively. Unfolded states are marked by "U1", "U2", "U3","U4"for SIM1, SIM2, SIM3, and SIM4, respectively. "B1" and "B3" denote the location from where β-sheet conformations (shown in Fig. 4). The numbers in each contour is free energy in kcal/mol.

3.6 Result and Discussion

(I) The unfolding of the full prion protein. Among the six well-tempered metadynamics[94] simulations performed here, we choose to discuss the results from the four simulations (referred to as SIM1, SIM2,SIM3, and SIM4) carried out using AMBER99SB force field for consistency. The free energy surfaces (FES's) obtained along three collective variables are shown in Fig. 2. For simplicity, we have shown in each row two cross-sections of the three dimensional (3D) FES as contour plots. The first column depicts the FES against N_c and R_g, while the second column shows the FES against ψ_{ave} and N_c. FESs using other force fields, CHARMM27[81] and AMBER03[82], are shown in Fig. 3. FES's from all the different simulations show three minima. White dashed line is the calculated minimum free energy path (MFEP) connecting the minima.

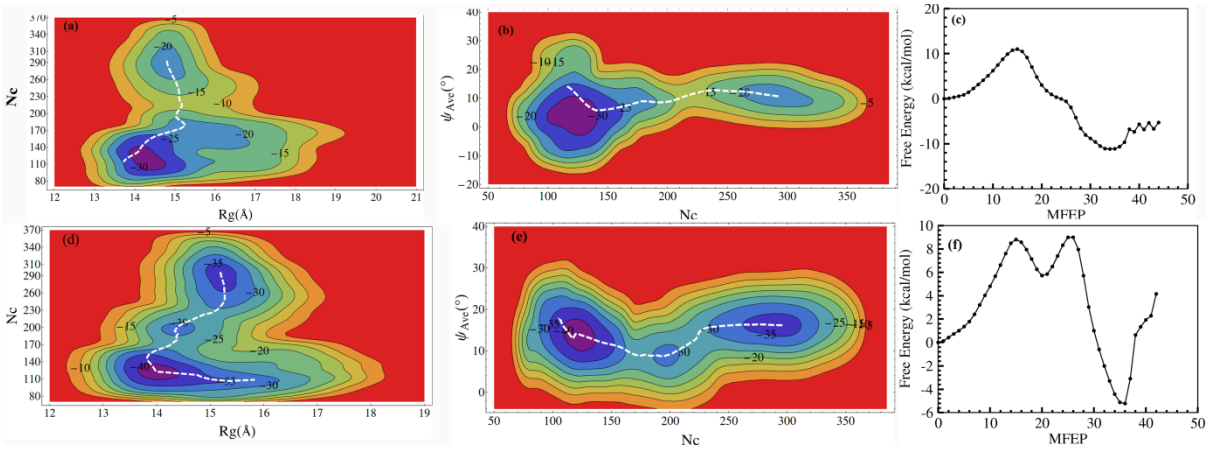

Figure 3. Free energy surfaces (FES's) of human prion protein unfolding using CHARM27 force field (top row a,b,c) and AMBER03 force-field (bottom row d,e,f). (a) and (d) FES against$N_c$and R_g (b) and (d) FES against$\psi_{ave}$and N_c. In each case, the third coordinate is Boltzmann averaged. White lines represent the calculated minimum free energy path in each case. (c) and (f) Free energy along the minimum free energy path. The numbers in each contour is free energy in kcal/mol

Among the multiple corrugations, each FES shows two major minima. The first one corresponds to the folded native state (N), whereN_c is 275, $R_g$is 14.3 Å, and ψ_{ave} is 25°. The minimum at the native configuration is present for all the four metadynamics simulations. This configuration is same as the crystal structure. Note that the ψ value of β-sheet is 120°, while it is -40° for the α-helix. The low value of ψ_{ave} in the native state indicates that most of the residues are either in α-helix or in the random coil configuration. The crystal structure is also characterized by the high native contact and low R_g. Note that, the native state does not have the highest N_c because of the way it is constructed (see method section).

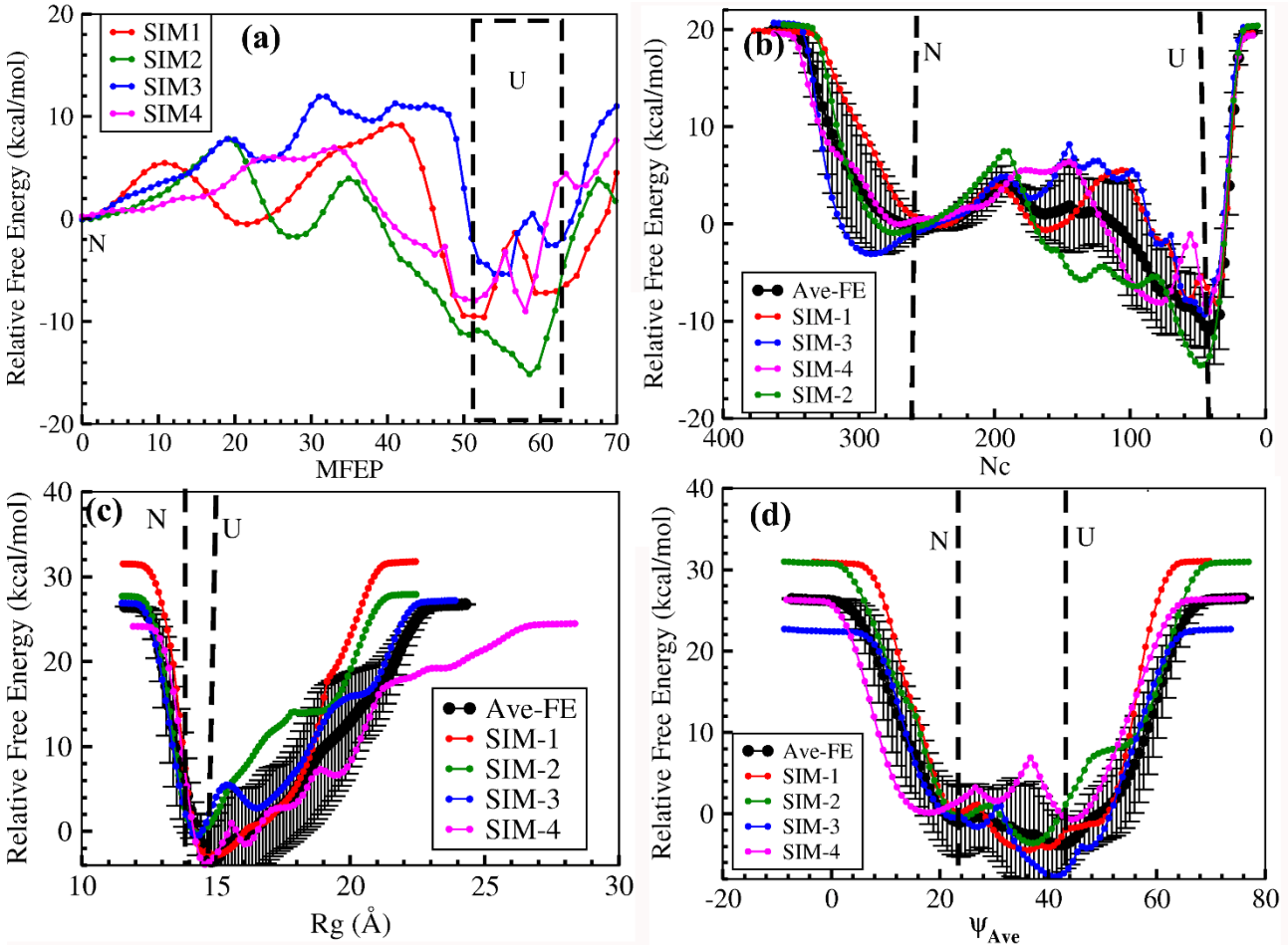

Figure 4. Relative free energy profiles with respect to native state (taken to be zero free energy) from four WTMD simulations. Free energy profiles along (a) MFEP. (b) native contact N_c (R_g and ψ_{ave} being Boltzmann averaged). (c) R_g (ψ_{ave} and N_c is being Boltzmann averaged), and (d) ψ_{ave} (N_c and R_g being Boltzmann averaged). Vertical lines in b, c, and d indicate errors in free energy. Note that both $R_g$and ψ_{ave} are unable to distinguish the native state and unfolded state. Average profile along MFEP is not meaningful because the same set of collective variables do not represent same MFEP index in different simulations.

The second minimum corresponds to N_c 60, R_g 14.1 Å, and ψ_{ave} 43° for SIM1, while it appears at N_c 60, R_g 14.2 Å, and ψ_{ave} 31° for SIM2 and at N_c 65, R_g 14.1 Å, and ψ_{ave} 35° for SIM3, at N_c 70, R_g 14.7 Å, and ψ_{ave} 28° for SIM4. We recognize these states as unfolded states and designated them by U1, U2, U3, and U4 for SIM1, SIM2, SIM3, and SIM4, respectively (see Fig. 2). In all the four independent metadynamics simulations, this unfolded state has always

lower free energy compared to the native state, although the pathways and the characteristics of these unfolded states are very different. This may be due to the lack of any structural order among the unfolded state. In fact, we see a reflection of the characteristics of the folding funnel in these four metadynamics simulations where a single native state may take different unfolding pathways to different unfolded configurations. Probably many more such paths exit and it will be almost impossible to enumerate all of them. However, from the free energy values, we can predict what would be the dominant unfolded state as discussed below.

To understand the unfolding events in each case, we compared the free energy profiles along the minimum free energy pathway (MFEP) for all the four simulations in Fig. 4a. MFEP is indexed with zero for the native state and an increasing number along the direction of unfolded state. All the four simulations show minima corresponding to three states, native state, an intermediate state, an unfolded state and an extended state. In two cases (SIM1 and SIM3), this intermediate minimum has similar free energetic stability as the native state, while for SIM2 and SIM4, the stability is much less. However, in all cases, the unfolded states have the lowest free energy and are denoted by U1, U2, U3, and U4 for SIM1, SIM2, SIM3, and SIM4, respectively. The extended state is characterizes as when the Rg of protein is high. At the extended state protein opens up and AAs of protein are completely exposed to the solvent. Therefore the extended state is unstable compare to native.

Table 1. Comparison of free energy barriers from multiple metadynamics simulations using different force fields.

Force field	Simulation	Native State FE	Barrier FE (kcal/mol)	Unfolded state FE (kcal/mol)
AMBER99SB	SIM1	0.0	9.0	-9.7
	SIM2	0.0	8.0	-15.0
	SIM3	0.0	11.1	-5.3
	SIM4	0.0	7.6	-8.9
CHARMM27	SIM5	0.0	10.7	-11.0
AMBER03	SIM6	0.0	9.1	-5.0
Average	All	0.0	9.3±1.4	-9.2±3.7

Since MFEP index is just a measure of steps taken along the unfolding pathway, same MFEP index does not indicate same set of CV's in different simulations. Therefore, although the free energy profiles along MFEP for four simulations look very different, the free energy values for a given value of the reaction coordinate should, in principle, be similar. Therefore, we plot the free energy profile along each of the reaction coordinate separately (with Boltzmann averaging of the rest two) in Fig. 4b to 4d for all the four simulations along with their average and error. We can see that the unfolding minima is captured well within the error bar, which is of ~3.7 kcal/mol. *Although a more precise estimate of free energy would require use of some recent biasing approaches such as "on the fly free-energy parameterization",[103], the present results capture the essential stability aspects of the prion protein through six different, long, and rigorous WTMD simulation with three different force fields.* We have summarized the free energies from the all different simulation in Table 1.

Fig. 4 shows that the stability of the unfolded state differs in different simulations. The reason, as shown later, lies in the different structural features of these unfolded states sampled by different simulations. Unfolded state in SIM2, denoted as U2, turns out to be the most stable minimum with 15 kcal/mol more stability than the native state, followed by U1, U4, and U3 with 9.7 kcal/mol, 8.9 kcal/mol, and 5.3 kcal/mol stability compared to the native state. Thirumalai and co-workers proposed that the prion aggregation goes through an intermediate conformation PrP^{C*} (unfolded state) which is more stable than the native state.[47] The present observation is consistent with that proposal.

These unfolded states, however, are separated from the native state with a barrier. Table 1 shows that the unfolding barrier is similar (~9.3±1.4 kcal/mol) for all four simulations. The spontaneous conversion from native to the unfolded state is probably hindered due to this barrier. Also, the present study is performed using an infinitely dilute protein concentration unlike experiments where aggregation or attachment to membrane may increase barrier towards misfolding.

Another interesting feature of the free energy profiles shown in Fig. 2 and 4a is that U1, U3, and U4 show two minima (basins) separated by a small (~4-8 kcal/mol) barrier, while U2, the deepest minimum, shows only one basin. The reason for that may lie in the details of structural changes that appear during different unfolding paths and will be discussed later. Before that, we investigated the origin of higher stability of the unfolded states through the total energy of the system and the total intramolecular hydrogen bonds (HBs) of protein along MFEP for each simulation, as shown in Fig. 5. Since the total energy of the system is enormous with a high fluctuation, we considered a subsystem comprising the protein and its first hydration shell. Then we performed the energy calculation of the structures along the MFEP (see method). Fig. 5 shows the average total energy of the subsystem. The averages were carried out on the biggest

cluster to represent the statistically dominant contributions to a particular point in the configuration space.

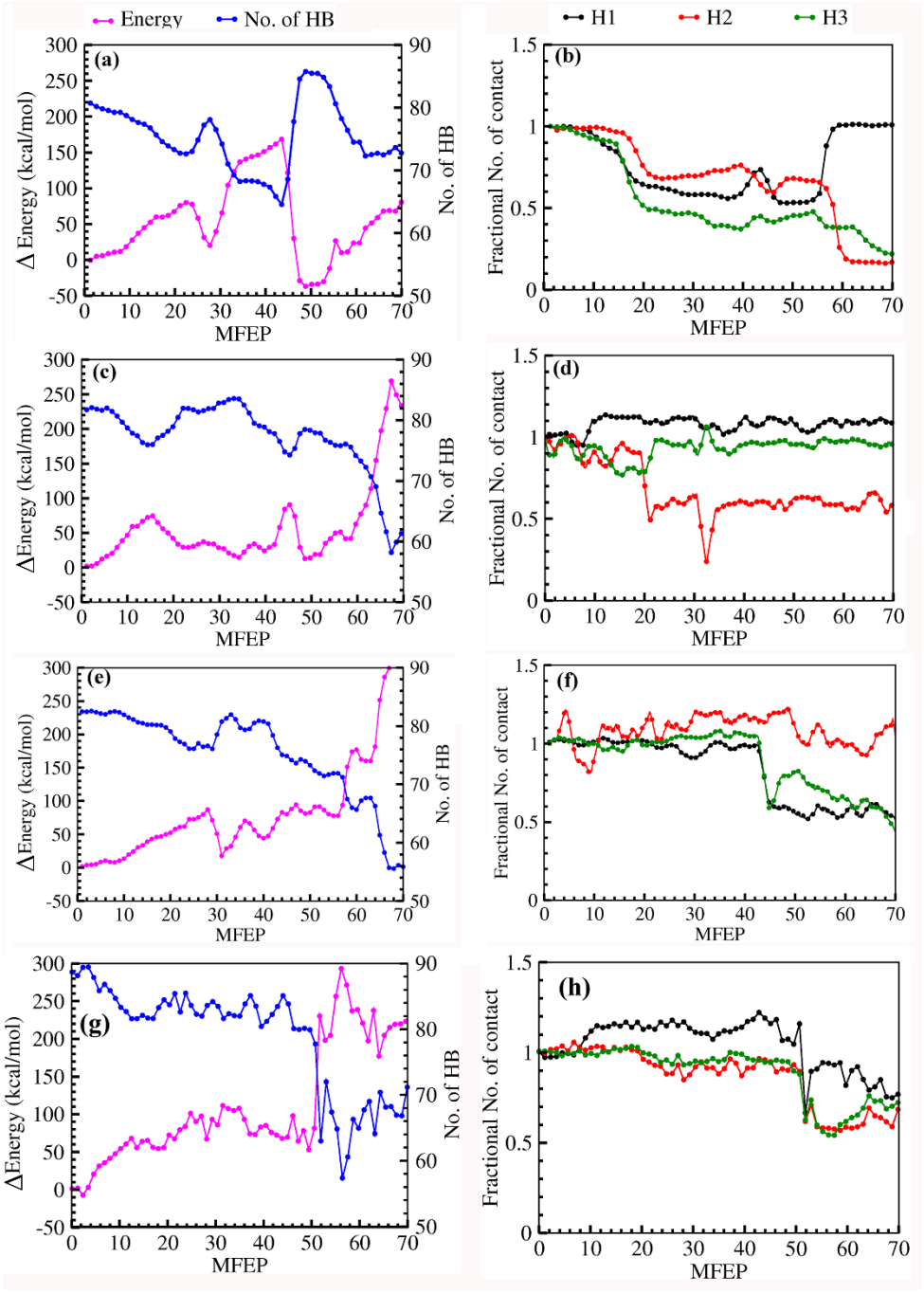

Figure 5. Left column shows the relative change in the total energy with respect to the native and intramolecular hydrogen bonds along MFEP for (a) SIM1 (c) SIM2, (e) SIM3 and (g) SIM4. Right column shows the change in the relative fraction of intramolecular native contact with respect to the native state for (b) SIM1, (d) SIM2,(f) SIM3, and (h) SIM4.

In the first simulation (SIM1), the energy increases (destabilization) up to MFEP index 41 as the protein unfolds (Fig. 5a). However, when protein is unfolded further (U1) at around MFEP index 50, the energy decreases (stabilizes) below the native state. This shows that the unfolded state (U1) is enthalpically more stable compared to the native state. The total number of HBs mirrors the energy profile and it becomes maximum (~85) for the unfolded state. Therefore, the origin of the energetic stabilization lies in the capacity of more HB formation. To investigate further, we have decomposed the internal energy in its components, Coulomb and LJ (see Fig. 6) and we found that the short-ranged Coulomb interaction, present in HB type interaction, plays a dominant role in the behavior of the total energy. Unlike SIM1, results of the other simulations (SIM2, SIM3, and SIM4) in Figs. 5b-d show an increase in the energy, which is again correlated with a decrease in HB number.

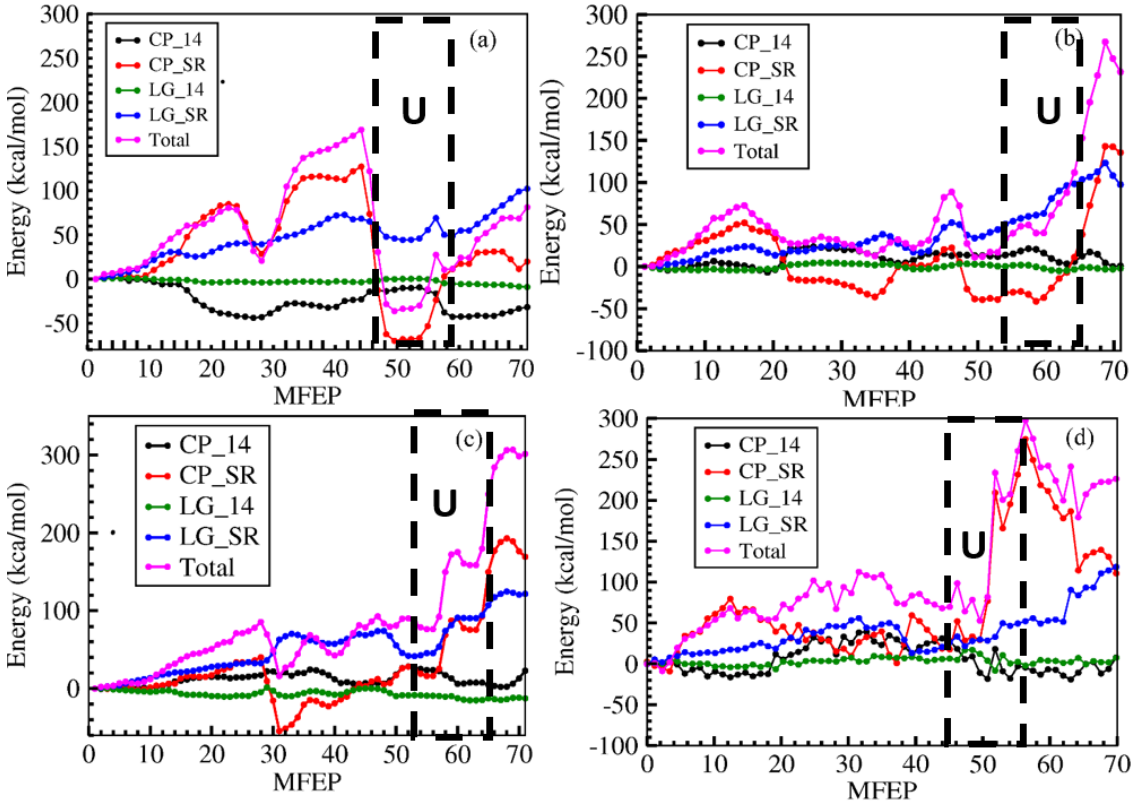

Figure 6. Decomposition of the internal energy (whole prion protein) into different components. CP_14, Coulomb 1-4 interaction (black color); CP_SR, Coulomb short-range interaction(red color); LG_14, Lennard Jones 1-4 interaction(green color); LG_SR, Lennard Jones Short Range interaction(blue color). Results are shown for (a) SIM1, (b) SIM2, and (c) SIM3 (d) SIM4. U in each graph represents the unfolded state region.

Further decomposition of the total energy shows that the water-protein interaction energies of SIM2, SIM3, and SIM4 are lower for the unfolded state (Fig. 7), whereas it is higher in SIM1 (Fig. 7). The high intramolecular HB in SIM1 (blue color of Fig 4a) compared to SIM2, SIM3, and SIM4 indicates that U1(blue color of Fig 4b-d), is more compact than the unfolded states U2, U3 and U4. The compactness of U1 is reflected in the low solvent-accessible surface area (SASA), U3, and U4 have higher SASA, on the contrary (not shown here). Therefore, we believe that the origin of the free energetic stability for U2, U3, and U4 may be entropic while for U1 it is enthalpic.

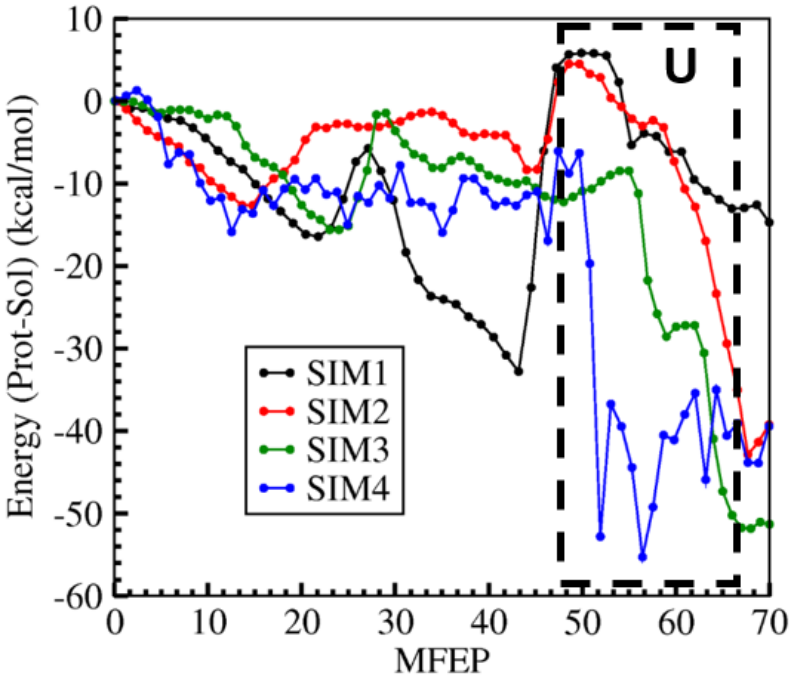

Figure 7. Interaction energy of the prion protein with solvent along MFEP in four different simulations. Different color represents different WTMD simulations using AMBER99SB forcefield. U in this graph represents the unfolded state region.

Although we have not studied the dimerization of protein, the previous study had shown that the prion oligomers are rich in β-sheet and more stable than the monomeric native state of prion protein. None of the previous study, however, shows the complete unfolding of monomeric prion[104-106]. Also, there is no study showing the state of prion protein before the oligomerization. However, we have observed that the monomeric prion unfolds with more stability of unfolded state. *Baskov et al.* have shown that the conformational transition of α-rich structure to β-rich structure goes via a high barrier -- associated with unfolding and a higher ordered kinetic process akin to the oligomerization. While α-rich prion was argued to be kinetically stable and the β-rich structure was thought to be thermodynamically more stable.[56] However we have not found here any β-rich state but found the unfolded prion protein to be more stable than the native state with high barrier of unfolding. So, we can speculate that the

unfolded states could initiate the oligomerization of prion and finally aggregate to form the amyloid structure.

Stability of the three helices of huPrP was often stated as one of the reasons for the propensity of misfolding of prion. It has been reported that H2 and H3 are more prone to unfolding compared to H1.[48-49,107] To characterize the nature of the unfolding process, we have calculated the fraction of intramolecular native contact present within each helical fragment along the MFEP. The intramolecular contact is defined to form if the distance between a specific atom (C, N, and O) of the i^{th} residue is within 4.5 Å from the same of the $(i+4)^{th}$ residue. Fig. 5(b), 5(d), 5(f), and 5(h) show the variation of the intramolecular contact within the three helical segments of SIM1, SIM2, SIM3 and SIM4 along the MFEP.

In SIM1, all helices start melting -- starting with H3 and subsequently followed by H1 and H2. However, H1 reformed while H2 and H3 melted away in the unfolded (U1) state. SIM2 shows that H2 melts while H1 and H3 remain intact. In SIM3, on the other hand, H1 and H3 melt partially, and H2 remains intact. SIM4 shows that H2 and H3 melt partially.

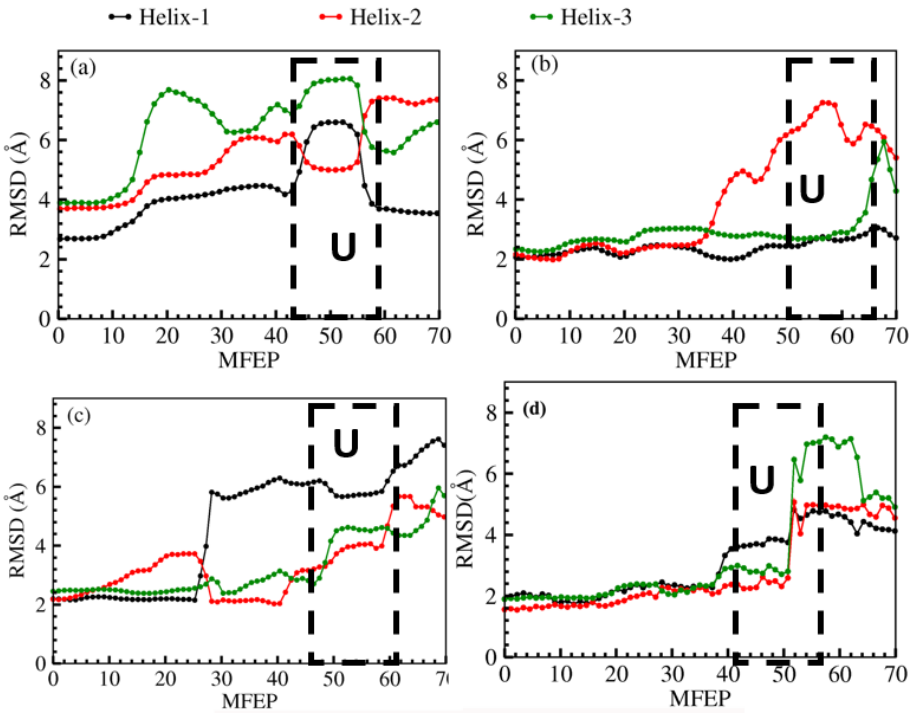

Figure 8. RMSD of helices from all three simulations: (a) SIM1, (b) SIM2, (c) SIM3 (d) SIM4. Black color represents the RMSD of H1. Red color represents the RMSD of H2. Green color represents the RMSD of H3.U in each graph represents the unfolded state region.

Stabilities of the helices observed above through contact are also reflected in the root mean square deviation (Fig. 8) and intramolecular hydrogen bonding (not shown here). Since the unfolded state in SIM2, U2, is most stable in free energy and both SIM1 and SIM2 show melting of H2, we can say that unfolding of H2 is most favorable. This is also consistent with the experimental study by Udgaonkar and co-workers who demonstrated that the stability of H2 is intimately connected to the misfolding and aggregation inprion.[50]

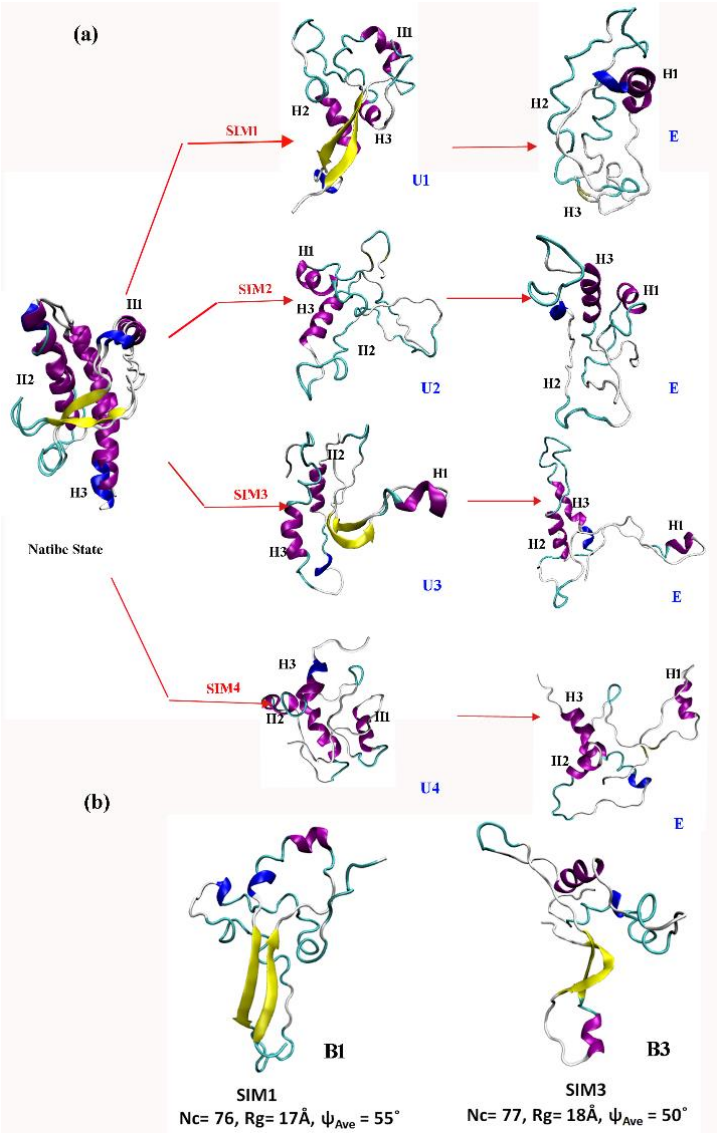

Figure 9. (a) Structural changes along the MFEP for four WTMD simulations. Note that unfolded state has some residual secondary structures. (b) Structures corresponding to "B1" and "B3" are shown in Fig. 1. Native state is superimposed on the equilibrated state to show that equilibrated structure is similar to the native state.

The structures along the MFEP of different simulations are shown in Fig. 9a. From the same native state, each simulation reaches to the free energetically different unfolded states and finally to the unstable extended configurations. We observe that H2 is entirely melted in U2, while it is partially melted in U1, U3 and U4. Note that the CV used to monitor the unfolding event did not include any intrafragmental contact. Therefore, breaking of contacts among the different fragments of the protein (contact between H1 & H3, H3 & H2, etc.) led to a

spontaneous loss of the secondary structure. This result indicates that the secondary structures of the fragments materialize due to the tertiary contacts and not due to the inherent sequence content of the protein in these fragments. This issue will be revisited in a subsequent section.

Even though we have used ψ_{ave}, shown to work for proteins with stable β sheet, we have not observed any β rich structure along MFEP. We only see the melting of helix in the unfolded state. This study however brings in the complete unfolding path and therefore shows a more stable unfolded structure with no β-sheet.

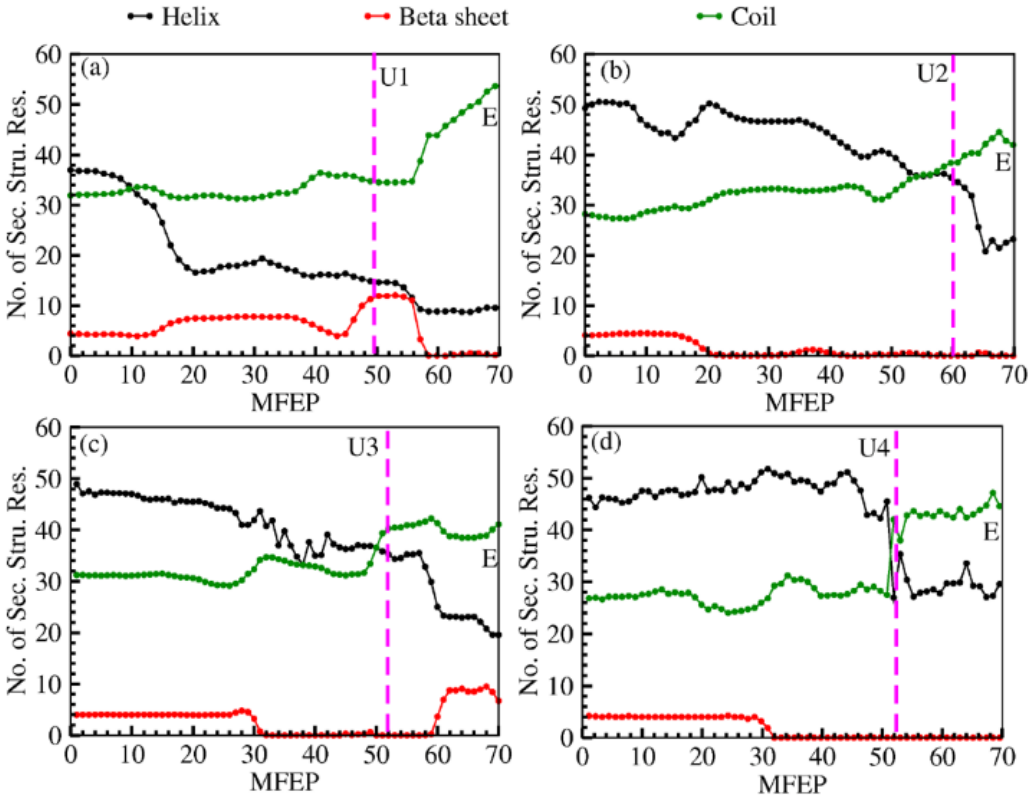

Figure 10. Number of amino acid residues in a particular secondary structure conformation (α-helix, β-sheet, and coil) for (a) SIM1 (b) SIM2 (c) SIM3, and (d) SIM4. Magenta line in each graph represents the unfolded state for the corresponding simulation.

All the different unfolding paths discussed above depict a loss of secondary structures. To quantitatively investigate the effect, we have calculated the number of residues belonging to a particular secondary structure motif (α-helix, β-sheet and coil conformation) along the MFEP

in Fig. 10. Figure 10(a) shows that during unfolding process of SIM1, primarily α-helix gets converted to random coil whereas β-sheet remains intact. The study by Parinello, *et al.*[29] showed that β-sheet is the most unstable region of prion protein. Here, we find it to be very stable. Only when the extended state starts forming around MFEP 56, β-sheet starts breaking into random coil. Another interesting observation is that there is no extension of β-sheet observed during the unfolding process, as is normally assumed for the PrP^{SC} form.[29] Therefore, we believe that the unfolding event observed in our study is just the formation of PrP^{C}*. SIM2, SIM3 and SIM4 also show increase of coil formation at the loss of both α-helical and β-sheet conformation. However, the loss of β-sheet happens much earlier than SIM1.

(II) Stability analysis of the helical fragments. The unfolding study of the whole prion protein showed us that when the tertiary native contacts are broken, the secondary structure elements, primarily the helical fragments, get converted to the coil form. Therefore, we tried to investigate the secondary structural preferences of these helical fragments H1, H2, and H3 by three different approaches: (i) by calculating the free energy against average ψ angle (ψ_{ave}), and (ii) by bioinformatic analyses of the sequence content of these helical fragments.

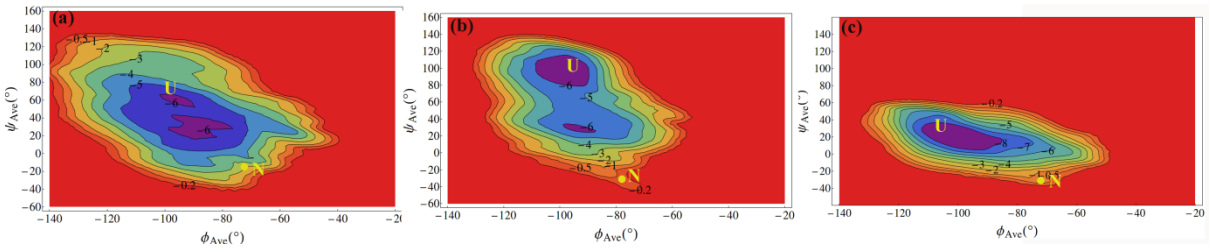

Figure 11. 2D free energy surfacesof chopped helices against $\varphi_{ave}$and ψ_{ave}. FES of (a) H1.(b) H2, and (c) H3. N and U in the FES represents the native and unfolded state. The numbers in each contour is free energy in kcal/mol

Fig 12 shows the free energy profiles of three helical fragments against ψ_{ave}, calculated using two ways: by performing one dimensional (1D) WTMD along ψ_{ave}, and a two dimensional

(2D) WTMD along φ_{ave} and ψ_{ave} follwed by projection on ψ_{ave}. The 2D FES's are shown in Fig. 11. If these fragments are in helical conformation, ψ_{ave} would be low whereas ψ_{ave} would be high if most of the residues adopt β-sheet conformation. The initial equilibrated structures from the crystal structures of H1, H2, and H3 were mostly in helical conformation with the ψ_{ave} at -16˚, -35˚, and -37˚, respectively for H1, H2, and H3. However, the lowest free energy (Fig. 6) for these segments are found to be at the random coil region with ψ_{ave} in the range of 20˚ to 40˚. The stability difference between the random coil and helical configuration found in the crystal structure is ~5 kcal/mol, ~7 kcal/mol and ~8.2 kcal/mol for H1, H2 and H3, respectively. Therefore, the tertiary configuration of the prion protein must overcome these barriers to keep the helical structure of these fragments in the prion protein. Fig. 11 also shows the structures of the fragments in the crystal configuration and in the β-rich state (high ψ_{ave}, not stable state) when in isolation.

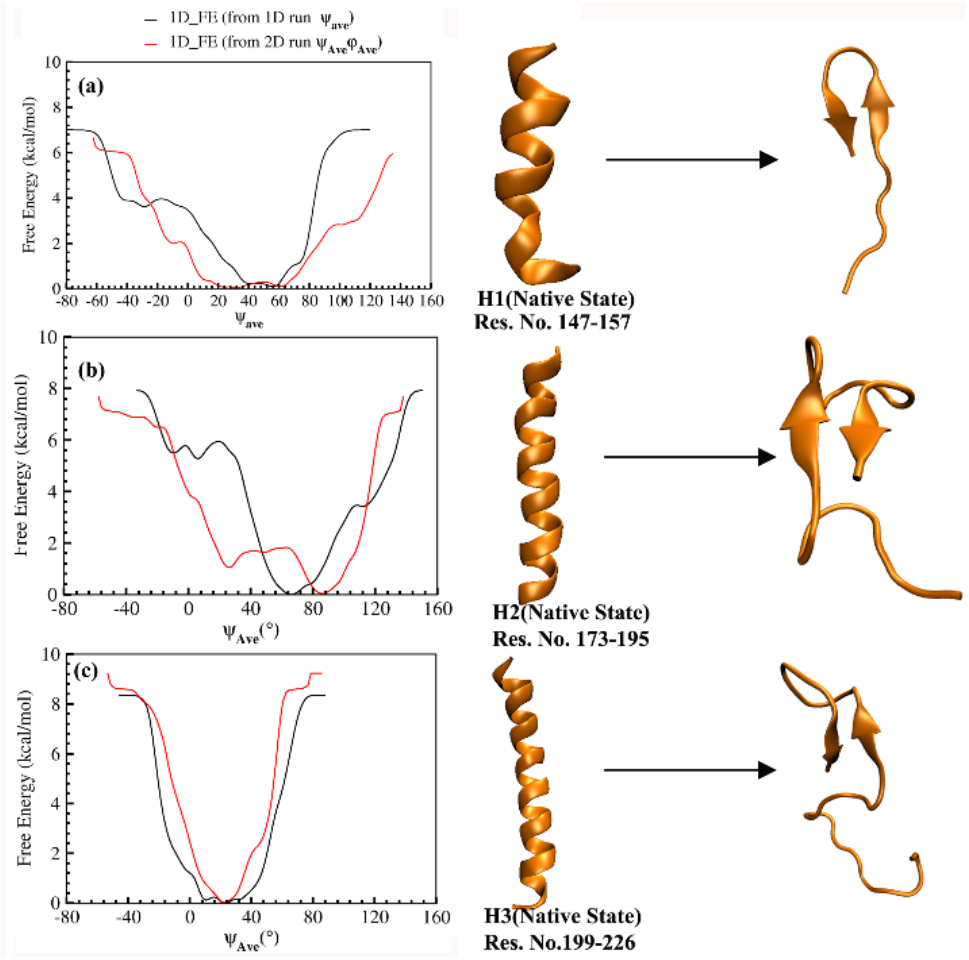

Figure 12. Free energy profiles for each fragment against ψ_{ave}: (a) H1, (b) H2, and (c) H3. H1, H2, and H3 have ψ_{ave} 16˚, -35˚, and -37˚at the native state. Black lines show the free energy profiles from 1D metadynamics simulation performed using ψ_{ave} while red lines show the free energy profile from 2D metadynamics simulation by Boltzmann averaging of φ_{ave}. Structures of the fragments in crystal environment (left) and also in the more β-rich states (right) are shown.

Initially, more residues were in helix (including 3_{10} helix) form and only a few were in the coil form. The number of residues in the helical form decreased with time as they convert to the random coil form (Fig. 13). The stable non-native structures of H1 and H3 are mostly in coil form and partially in 3_{10} helix structure (Fig. 13a,c), whereas all residues are in coil form for H2 (Fig. 13b). We have not observed any β-sheet formation over the whole trajectory for H1 and H3. For H2, five residues formed β-sheet (Fig. 13b) in the stable state.

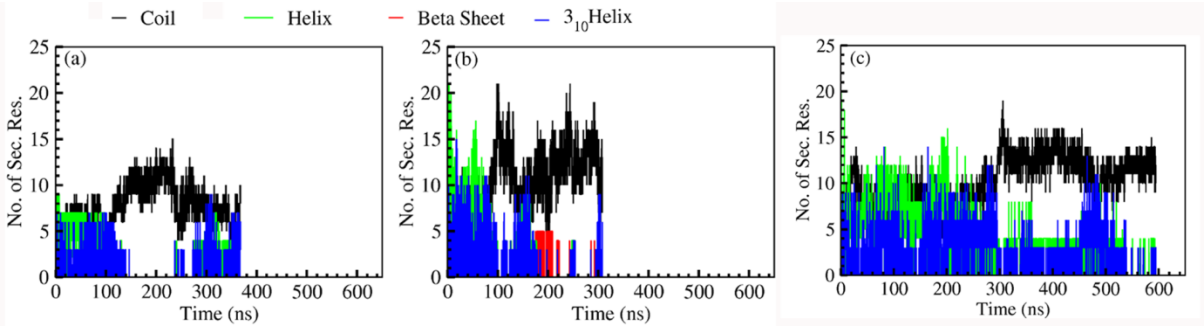

Figure 13. Number of secondary structure residue with time for the excised helices (a) H1 (b) H2 (c) H3. Different color represents different secondary structure as shown in the legend above the figure.

To understand the sequence-wise propensity towards specific secondary structure in these helices, we have performed statistical sequence analysis of the secondary structure propensity towards α-helix, β-sheet, and coil formation. NetSurfP (ver. 1.1)[108] was used for sequence-based structure prediction. Fig. 14 shows that H2 and to some extent H3 (C-terminal region) show a higher tendency of β-sheet formation compared to H1 based on their sequential information. Further, we have calculate the charge versus hydrophobity (C-H) of entire protein, H1, H2 and H3. The hydrophobicity of protein is calculated by the method discussed in ref 111. Fig 15 shows the C-H graph. The dashed magenta line show the boundary for the ordered to disordered transition. Below to the dashed line is the characteristics of intrinsically disordered protein (IDP). Helix-1 and Helix-3 falls in this region. The Helix-2 sit on the boundary. Interestingly, the entire Prion protein fall in this region (green color dot in Fig 15), which suggest that the Prion protein could be an intrinsically disorder protein.

The above analyses on the sequence, structure, and dynamics of the chopped helices highlight the higher intrinsic stability for H1 than H2 and H3 in agreement with prior theoretical studies by Thirumalai and co-workers.[47] Moreover, the inherent sequence of the H2 and H3 helical fragments of the prion protein render these segments unstable and they are not meant to be in a helical configuration. These structural elements are weakly stabilized by the intersegmental

tertiary contacts and formation of a hydrophobic core. This leads to the inherent metastable nature of the cellular PrP^C form, where breaking the tertiary contacts through the N_c collective variable results in the disruption of the secondary structures as well, and finally leading to the global minima in the FES.

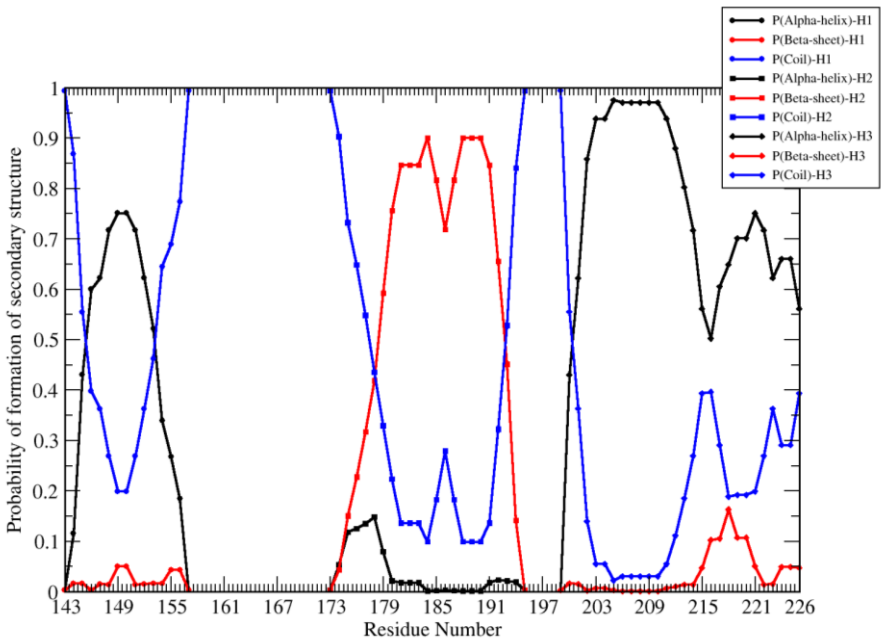

Figure 14. Bioinformatics analysis of the structural propensities of the sequences in each chopped helices (H1, H2, H3) to be forα-helix, β-sheet and coil formation.

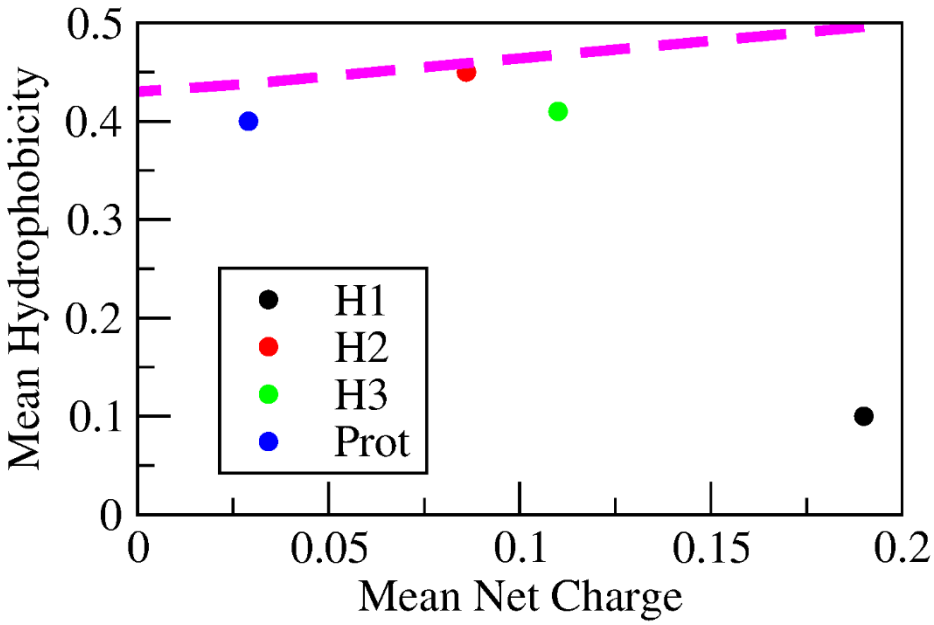

Figure 15. Net charge Vs hydrophobicity of protein. Black dot represent the helix-1. Red dot represent the helix-2. Green dot represent the helix-3 and blue dot represents the whole protein. Magenta dashed line represent the boundary for order to disordered transition.

3.7 Conclusion

This work presents an extensive Metadynamics simulation study of the conformational free energy landscape of a human Prion protein. To the best of our knowledge, this is the first theoretical proof that the PrP^{C} isoform is metastable in nature, and there exist another compact yet partially unfolded isoform with considerably higher thermodynamic stability. This state may be similar to the intermediate PrP^{C*} state that has been suggested to appear during the PrP^{C} to PrP^{Sc} transition. Estimation of free energies using two different, both rigorous, methods also shows that prion protein has characteristics of an intrinsically disordered protein, since the secondary structural elements (H2 and H3) are inherently unstable. The PrP^{C} structure is only weakly held together by tertiary contacts and solvation forces, and breakage of these contacts involve crossing of a free energy barrier (9-12 kcal/mol). The system reaches a

thermodynamically more stable state upon crossing this barrier, which might be further induced in pathogenic conditions like lower pH, higher temperatures, mutations, etc.

While prior simulation studies using REMD, higher temperatures, lower pH, etc. indicate towards formation of β-rich isoforms as hallmark of formation of PrP^{Sc} like structures, we do not observe frequent occurrences of such states in our metadynamics simulations in the unfolding reaction coordinate. We argue that the PrP^{Sc} like structures that would lead to eventual fibril formation must involve parallel β-sheet formation. However, most of the earlier MD simulation studies observe formation of anti-parallel β-sheets. Thus, we feel that formation of PrP^{Sc} with parallel β-sheets must involve significant unfolding of the inherent secondary structure present in PrP^{C} leading to formation of inter-molecular β-sheet formation. Thus, the anti-parallel β-rich states might be off-pathway as far as the formation of toxic oligomer is concerned. Interestingly, similar suggestions has been put forward by some of the recent experimental studies as well,[106] and future directions should involve understanding the mechanism of oligomerization of these structure-less unfolded states to continue the search for PrP^{Sc} structure.

Although we used three different force fields for proteins and obtained similar result, we used only TIP3P model for water. Recent studies show that this water model has a tendency to overstabilize the hydrophobic core of the protein.[109,110] Thus the choice of water model might affect the relative stability of the unfolded compact states reported here to some extent. However, these effects require further systematic investigation in future.

Reference

1. Mouillet-Richard, S.; Baudry, A.; Launay, J. M.; Kellermann, O. Micrornas and Depression. *Neurobiol. Dis.* 2012, *46*, 272-8.
2. Moser, M.; Colello, R. J.; Pott, U.; Oesch, B. Developmental Expression of the Prion Protein Gene in Glial Cells. *Neuron* 1995, *14*, 509-17.
3. Ford, M. J.; Burton, L. J.; Morris, R. J.; Hall, S. M. Selective Expression of Prion Protein in Peripheral Tissues of the Adult Mouse. *Neuroscience* 2002, *113*, 177-92.
4. Stahl, N.; Borchelt, D. R.; Hsiao, K.; Prusiner, S. B. Scrapie Prion Protein Contains a Phosphatidylinositol Glycolipid. *Cell* 1987, *51*, 229-40.
5. Lee, S.; Antony, L.; Hartmann, R.; Knaus, K. J.; Surewicz, K.; Surewicz, W. K.; Yee, V. C. Conformational Diversity in Prion Protein Variants Influences Intermolecular Beta-Sheet Formation. *EMBO J.* 2010, *29*, 251-62.
6. Nishida, N.; Tremblay, P.; Sugimoto, T.; Shigematsu, K.; Shirabe, S.; Petromilli, C.; Erpel, S. P.; Nakaoke, R.; Atarashi, R.; Houtani, T. *et al.* A Mouse Prion Protein Transgene Rescues Mice Deficient for the Prion Protein Gene from Purkinje Cell Degeneration and Demyelination. *Lab. Invest.* 1999, *79*, 689-97.
7. Brown, D. R.; Schulz-Schaeffer, W. J.; Schmidt, B.; Kretzschmar, H. A. Prion Protein-Deficient Cells Show Altered Response to Oxidative Stress Due to Decreased Sod-1 Activity. *Exp. Neurol.* 1997, *146*, 104-112.
8. Brown, D. R.; Nicholas, R. S. J.; Canevari, L. Lack of Prion Protein Expression Results in a Neuronal Phenotype Sensitive to Stress. *J. Neurosci. Res.* 2002, *67*, 211-224.
9. Jackson, G. S.; Murray, I.; Hosszu, L. L.; Gibbs, N.; Waltho, J. P.; Clarke, A. R.; Collinge, J. Location and Properties of Metal-Binding Sites on the Human Prion Protein. *Proc. Natl. Acad. Sci. U. S. A.* 2001, *98*, 8531-5.
10. Brown, D. R.; Qin, K.; Herms, J. W.; Madlung, A.; Manson, J.; Strome, R.; Fraser, P. E.; Kruck, T.; von Bohlen, A.; Schulz-Schaeffer, W. et, al. The Cellular Prion Protein Binds Copper in Vivo. *Nature* 1997, *390*, 684-7.

11. Kramer, M. L.; Kratzin, H. D.; Schmidt, B.; Romer, A.; Windl, O.; Liemann, S.; Hornemann, S.; Kretzschmar, H. Prion Protein Binds Copper within the Physiological Concentration Range. *J. Biol. Chem.* 2001, *276*, 16711-16719.

12. Stockel, J.; Safar, J.; Wallace, A. C.; Cohen, F. E.; Prusiner, S. B. Prion Protein Selectively Binds Copper(II) Ions. *Biochemistry* 1998, *37*, 7185-93.

13. Jones, C. E.; Abdelraheim, S. R.; Brown, D. R.; Viles, J. H. Preferential Cu2+ Coordination by His96 and His111 Induces Beta-Sheet Formation in the Unstructured Amyloidogenic Region of the Prion Protein. *J. Biol. Chem.* 2004, *279*, 32018-27.

14. Leclerc, E.; Serban, H.; Prusiner, S. B.; Burton, D. R.; Williamson, R. A. Copper Induces Conformational Changes in the N-Terminal Part of Cell-Surface Prpc. *Arch. Virol.* 2006, *151*, 2103-9.

15. Schmitt-Ulms, G.; Legname, G.; Baldwin, M. A.; Ball, H. L.; Bradon, N.; Bosque, P. J.; Crossin, K. L.; Edelman, G. M.; De Armond, S. J.; Cohen, F. E. et, al. Binding of Neural Cell Adhesion Molecules (N-Cams) to the Cellular Prion Protein. *J. Mol. Biol.* 2001, *314*, 1209-25.

16. Graner, E.; Mercadante, A. F.; Zanata, S. M.; Forlenza, O. V.; Cabral, A. L.; Veiga, S. S.; Juliano, M. A.; Roesler, R.; Walz, R.; Minetti, A. et, al. Cellular Prion Protein Binds Laminin and Mediates Neuritogenesis. *Brain Res. Mol. Brain Res.* 2000, *76*, 85-92.

17. Graner, E.; Mercadante, A. F.; Zanata, S. M.; Martins, V. R.; Jay, D. G.; Brentani, R. R. Laminin-Induced Pc-12 Cell Differentiation Is Inhibited Following Laser Inactivation of Cellular Prion Protein. *FEBS Lett.* 2000, *482*, 257-260.

18. Taylor, D. R.; Hooper, N. M. The Prion Protein and Lipid Rafts. *Mol. Membr. Biol.* 2006, *23*, 89-99.

19. Spielhaupter, C.; Schatzl, H. M. PrP^C Directly Interacts with Proteins Involved in Signaling Pathways. *J. Biol. Chem.* 2001, *276*, 44604-12.

20. Loertscher, R.; Lavery, P. The Role of Glycosyl Phosphatidyl Inositol (Gpi)-Anchored Cell Surface Proteins in T-Cell Activation. *Transpl. Immunol.* 2002, *9*, 93-96.

21. Schneider, B.; Mutel, V.; Pietri, M.; Ermonval, M.; Mouillet-Richard, S.; Kellermann, O. Nadph Oxidase and Extracellular Regulated Kinases 1/2 Are Targets of Prion Protein Signaling in Neuronal and Nonneuronal Cells. *Proc. Natl. Acad. Sci. U. S. A.* 2003, *100*, 13326-31.

22. Mouillet-Richard, S.; Ermonval, M.; Chebassier, C.; Laplanche, J. L.; Lehmann, S.; Launay, J. M.; Kellermann, O. Signal Transduction through Prion Protein. *Science* 2000, *289*, 1925-1928.

23. Pan, K. M.; Baldwin, M.; Nguyen, J.; Gasset, M.; Serban, A.; Groth, D.; Mehlhorn, I.; Huang, Z.; Fletterick, R. J.; Cohen, F. E. Conversion of Alpha-Helices into Beta-Sheets Features in the Formation of the Scrapie Prion Proteins. *Proc. Natl. Acad. Sci. U. S. A.* 1993, *90*, 10962-10966.

24. Chakroun, N.; Prigent, S.; Dreiss, C. A.; Noinville, S.; Chapuis, C.; Fraternali, F.; Rezaei, H. The Oligomerization Properties of Prion Protein Are Restricted to the H2h3 Domain. *FASEB J.* 2010, *24*, 3222-31.

25. Riek, R.; Hornemann, S.; Wider, G.; Billeter, M.; Glockshuber, R.; Wüthrich, K. Nmr Structure of the Mouse Prion Protein Domain Prp(121-231). *Nature* 1996, *382*, 180-182.

26. Stahl, N.; Baldwin, M. A.; Teplow, D. B.; Hood, L.; Gibson, B. W.; Burlingame, A. L.; Prusiner, S. B. Structural Studies of the Scrapie Prion Protein Using Mass-Spectrometry and Amino-Acid Sequencing. *Biochemistry* 1993, *32*, 1991-2002.

27. Basler, K.; Oesch, B.; Scott, M.; Westaway, D.; Walchli, M.; Groth, D. F.; McKinley, M. P.; Prusiner, S. B.; Weissmann, C. Scrapie and Cellular PrPIsoforms are Encoded by the Same Chromosomal Gene. *Cell* 1986, *46*, 417-28.

28. Prusiner, S. B. Prions. *Proc. Natl. Acad. Sci. U. S. A.* 1998, *95*, 13363-83.
29. Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V.; Gervasio, F. L.; Parrinello, M. Metadynamics Simulation of Prion Protein: Beta-Structure Stability and the Early Stages of Misfolding. *J. Am. Chem. Soc.* 2006, *128*, 2705-10.
30. Prusiner, S. B. Creutzfeldt-Jakob Disease and Scrapie Prions. *Alzheimer Dis. Assoc. Disord.* 1989, *3*, 52-78.
31. Stahl, N.; Prusiner, S. B. Prions and Prion Proteins. *FASEB J.* 1991, *5*, 2799-807.
32. Westergard, L.; Christensen, H. M.; Harris, D. A. The Cellular Prion Protein PrP^{C}: Its Physiological Function and Role in Disease. *Biochim. Biophys. Acta* 2007,*1772*, 629-644.
33. Zahn, R.; Liu, A.; Luhrs, T.; Riek, R.; von Schroetter, C.; Lopez Garcia, F.; Billeter, M.; Calzolai, L.; Wider, G.; Wuthrich, K. NMR Solution Structure of the Human Prion Protein. *Proc. Natl. Acad. Sci. U. S. A.* 2000, *97*, 145-50.
34. Calzolai, L.; Zahn, R. Influence of Ph on Nmr Structure and Stability of the Human Prion Protein Globular Domain. *J. Biol. Chem.* 2003, *278*, 35592-6.
35. Apostol, M. I.; Perry, K.; Surewicz, W. K. Crystal Structure of a Human Prion Protein Fragment Reveals a Motif for Oligomer Formation. *J. Am. Chem. Soc.* 2013,*135*, 10202-5.
36. Lehmann, S. Metal Ions and Prion Diseases. *Curr. Opin. Chem. Biol.* 2002, *6*, 187-92.
37. Sanghera, N.; Pinheiro, T. J. T. Binding of Prion Protein to Lipid Membranes and Implications for Prion Conversion1. *J. Mol. Biol.* 2002, *315*, 1241-1256.
38. Milhavet, O.; Lehmann, S. Oxidative Stress and the Prion Protein in Transmissible Spongiform Encephalopathies. *Brain Res. Brain Res. Rev.* 2002, *38*, 328-39.
39. Wille, H.; Michelitsch, M. D.; Guenebaut, V.; Supattapone, S.; Serban, A.; Cohen, F. E.; Agard, D. A.; Prusiner, S. B. Structural Studies of the Scrapie Prion Protein by Electron Crystallography. *Proc. Natl. Acad. Sci. U. S. A.* 2002, *99*, 3563-8.
40. Govaerts, C.; Wille, H.; Prusiner, S. B.; Cohen, F. E. Evidence for Assembly of Prions with Left-Handed B-Helices into Trimers. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101*, 8342-8347.
41. McColl, I. H.; Blanch, E. W.; Gill, A. C.; Rhie, A. G.; Ritchie, M. A.; Hecht, L.; Nielsen, K.; Barron, L. D. A New Perspective on Beta-Sheet Structures Using Vibrational Raman Optical Activity: From Poly(L-Lysine) to the Prion Protein. *J. Am. Chem. Soc.* 2003, *125*, 10019-26.
42. Alonso, D. O.; DeArmond, S. J.; Cohen, F. E.; Daggett, V. Mapping the Early Steps in the Ph-Induced Conformational Conversion of the Prion Protein. *Proc. Natl. Acad. Sci. U. S. A.* 2001, *98*, 2985-9.
43. DeMarco, M. L.; Daggett, V. From Conversion to Aggregation: Protofibril Formation of the Prion Protein. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101*, 2293-8.
44. Der-Sarkissian, A.; Jao, C. C.; Chen, J.; Langen, R. Structural Organization of Alpha-Synuclein Fibrils Studied by Site-Directed Spin Labeling. *J. Biol. Chem.* 2003, *278*, 37530-5.
45. Singh, J.; Udgaonkar, J. B. Molecular Mechanism of the Misfolding and Oligomerization of the Prion Protein: Current Understanding and Its Implications. *Biochemistry* 2015, *54*, 4431-4442.
46. Eghiaian, F.; Grosclaude, J.; Lesceu, S.; Debey, P.; Doublet, B.; Treguer, E.; Rezaei, H.; Knossow, M. Insight into the PrP^{C} -> PrP^{Sc} Conversion from the Structures of Antibody-Bound Ovine Prion Scrapie-Susceptibility Variants. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101*, 10254-10259.
47. Dima, R. I.; Thirumalai, D. Probing the Instabilities in the Dynamics of Helical Fragments from Mouse Prpc. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101*, 15335-15340.
48. Chakroun, N.; Fornili, A.; Prigent, S.; Kleinjung, J.; Dreiss, C. A.; Rezaei, H.; Fraternali, F. Decrypting Prion Protein Conversion into a β-Rich Conformer by Molecular Dynamics. *J. Chem. Theory Comput.* 2013, *9*, 2455-2465.

49. Vila-Viçosa, D.; Campos, S. R. R.; Baptista, A. M.; Machuqueiro, M. Reversibility of Prion Misfolding: Insights from Constant-pH Molecular Dynamics Simulations. *J. Phys. Chem. B* 2012, *116*, 8812-8821.
50. Singh, J.; Kumar, H.; Sabareesan, A. T.; Udgaonkar, J. B. Rational Stabilization of Helix 2 of the Prion Protein Prevents Its Misfolding and Oligomerization. *J. Am. Chem. Soc.* 2014, *136*, 16704-16707.
51. Camilloni, C.; Schaal, D.; Schweimer, K.; Schwarzinger, S.; De Simone, A. Energy Landscape of the Prion Protein Helix 1 Probed by Metadynamics and NMR. *Biophys. J.* 2012, *102*, 158-167.
52. Morrissey, M. P.; Shakhnovich, E. I. Evidence for the Role of Prp^{C} Helix 1 in the Hydrophilic Seeding of Prion Aggregates. *Proc. Natl. Acad. Sci. U. S. A.* 1999, *96*, 11293-11298.
53. Honda, Ryo P.; Xu, M.; Yamaguchi, K.-i.; Roder, H.; Kuwata, K. A Native-Like Intermediate Serves as a Branching Point between the Folding and Aggregation Pathways of the Mouse Prion Protein. *Structure* 2015, *23*, 1735-1742.
54. Thirumalai, D.; Klimov, D. K.; Dima, R. I. Emerging Ideas on the Molecular Basis of Protein and Peptide Aggregation. *Curr. Opin. Struct. Biol.* 2003, *13*, 146-59.
55. Chen, J.; Thirumalai, D. Helices 2 and 3 Are the Initiation Sites in the $PrP^{C} \rightarrow PrP^{Sc}$ Transition. *Biochemistry* 2013, *52*, 310-319.
56. Baskakov, I. V.; Legname, G.; Prusiner, S. B.; Cohen, F. E. Folding of Prion Protein to Its Native α-Helical Conformation Is under Kinetic Control. *J. Biol. Chem.* 2001,*276*, 19687-19690.
57. Baskakov, I. V.; Legname, G.; Baldwin, M. A.; Prusiner, S. B.; Cohen, F. E. Pathway Complexity of Prion Protein Assembly into Amyloid. *J. Biol. Chem.* 2002, *277*, 21140-21148.
58. Jain, S.; Udgaonkar, J. B. Evidence for Stepwise Formation of Amyloid Fibrils by the Mouse Prion Protein. *J. Mol. Biol.* 2008,*382*, 1228-1241.
59. Jain, S.; Udgaonkar, J. B. Prion Protein Aggregation. *Curr. Sci.* 2011, *101*, 1311-1327.
60. Bjorndahl, T. C.; Zhou, G.-P.; Liu, X.; Perez-Pineiro, R.; Semenchenko, V.; Saleem, F.; Acharya, S.; Bujold, A.; Sobsey, C. A.; Wishart, D. S. Detailed Biophysical Characterization of the Acid-Induced PrP^{C} to PrP β Conversion Process. *Biochemistry* 2011, *50*, 1162-1173.
61. Campos, S. R. R.; Machuqueiro, M.; Baptista, A. M. Constant-pH Molecular Dynamics Simulations Reveal a β-Rich Form of the Human Prion Protein. *J. Phys. Chem. B* 2010, *114*, 12692-12700.
62. Hornemann, S.; Glockshuber, R. A Scrapie-Like Unfolding Intermediate of the Prion Protein Domain PrP(121–231) Induced by Acidic pH. *Proc. Natl. Acad. Sci. U. S. A.* 1998, *95*, 6010-6014.
63. Langella, E.; Improta, R.; Barone, V. Checking the Ph-Induced Conformational Transition of Prion Protein by Molecular Dynamics Simulations: Effect of Protonation of Histidine Residues. *Biophys. J.* 2004, *87*, 3623-3632.
64. Alonso, D. O. V.; An, C.; Daggett, V. Simulations of Biomolecules: Characterization of the Early Steps in the Ph-Induced Conformational Conversion of the Hamster, Bovine and Human Forms of the Prion Protein. *Philos. Transact. A Math. Phys. Eng. Sci.* 2002, *360*, 1165-1178.
65. Cheng, C. J.; Daggett, V. Molecular Dynamics Simulations Capture the Misfolding of the Bovine Prion Protein at Acidic pH. *Biomolecules* 2014, *4*, 181-201.
66. Sekijima, M.; Motono, C.; Yamasaki, S.; Kaneko, K.; Akiyama, Y. Molecular Dynamics Simulation of Dimeric and Monomeric Forms of Human Prion Protein: Insight into Dynamics and Properties. *Biophys. J.* 2003, *85*, 1176-1185.

67. Langella, E.; Improta, R.; Crescenzi, O.; Barone, V. Assessing the Acid–Base and Conformational Properties of Histidine Residues in Human Prion Protein (125–228) by Means of pKa Calculations and Molecular Dynamics Simulations. *Proteins: Struct., Funct., Bioinf.* 2006, *64*, 167-177.

68. Garrec, J.; Tavernelli, I.; Rothlisberger, U. Two Misfolding Routes for the Prion Protein around pH 4.5. *PLOS Comput. Biol.* 2013,*9*, e1003057.

69. Hsiao, K.; Baker, H. F.; Crow, T. J.; Poulter, M.; Owen, F.; Terwilliger, J. D.; Westaway, D.; Ott, J.; Prusiner, S. B. Linkage of a Prion Protein Missense Variant to Gerstmann-Sträussler Syndrome. *Nature* 1989, *338*, 342-345.

70. Kretzschmar, H. A. Human Prion Diseases (Spongiform Encephalopathies). In *Unconventional Agents and Unclassified Viruses*, Kaaden, O. R.; Eichhorn, W. Czerny, C. P., Eds. Springer Vienna: 1993; Vol. 7, pp 261-293.

71. Mastrianni, J. A.; Curtis, M. T.; Oberholtzer, J. C.; Da Costa, M. M.; DeArmond, S.; Prusiner, S. B.; Garbern, J. Y. Prion Disease (PrP-A117V) Presenting with Ataxia Instead of Dementia. *Neurology* 1995,*45*, 2042-2050.

72. Tateishi, J.; Kitamoto, T.; Doh-Ura, K.; Sakaki, Y.; Steinmetz, G.; Tranchant, C.; Warter, J. M.; Heldt, N. Immunochemical, Molecular Genetic, and Transmission Studies on a Case of Gerstmann-Straussler-Scheinker Syndrome. *Neurology* 1990, *40*, 1578-1578.

73. Hsiao, K. K.; Cass, C.; Schellenberg, G. D.; Bird, T.; Devine-Gage, E.; Wisniewski, H.; Prusiner, S. B. A Prion Protein Variant in a Family with the Telencephalic Form of Gerstmann-Sträussler-Scheinker Syndrome. *Neurology* 1991, *41*, 681-684.

74. Okimoto, N.; Yamanaka, K.; Suenaga, A.; Hata, M.; Hoshino, T. Computational Studies on Prion Proteins: Effect of Ala117→Val Mutation. *Biophys. J.* 2002, *82*, 2746-2757.

75. Chebaro, Y.; Derreumaux, P. The Conversion of Helix H2 to B-Sheet Is Accelerated in the Monomer and Dimer of the Prion Protein Upon T183A Mutation. *J. Phys. Chem. B* 2009, *113*, 6942-6948.

76. Gsponer, J.; Ferrara, P.; Caflisch, A. Flexibility of the Murine Prion Protein and Its Asp178Asn Mutant Investigated by Molecular Dynamics Simulations. *J. Mol. Graph. Model.* 2001, *20*, 169-182.

77. Cobb, N. J.; Sönnichsen, F. D.; McHaourab, H.; Surewicz, W. K. Molecular Architecture of Human Prion Protein Amyloid: A Parallel, in-Register β-Structure. *Proc. Natl. Acad. Sci. U. S. A.* 2007, *104*, 18946-18951.

78. Tycko, R.; Savtchenko, R.; Ostapchenko, V. G.; Makarava, N.; Baskakov, I. V. The α-Helical C-Terminal Domain of Full-Length Recombinant PrP Converts to an in-Register Parallel β-Sheet Structure in PrP Fibrils: Evidence from Solid State Nmr. *Biochemistry* 2010, *49*, 9488-9497.

79. Groveman, B. R.; Dolan, M. A.; Taubner, L. M.; Kraus, A.; Wickner, R. B.; Caughey, B. Parallel in-Register Intermolecular B-Sheet Architectures for Prion-Seeded Prion Protein (PrP) Amyloids. *J. Biol. Chem.* 2014, *289*, 24129-24142.

80. Wang, J. M.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (Resp) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* 2000, *21*, 1049-1074.

81. Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. Importance of the Cmap Correction to the Charmm22 Protein Force Field: Dynamics of Hen Lysozyme. *Biophys. J.* 2006, *90*, L36-L38.

82. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* 2003, *24*, 1999-2012.

83. Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlić, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D. The Rcsb Protein Data Bank: Redesigned Web Site and Web Services. *Nucleic Acids Res.* 2011, *39*, D392-D401.

84. Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z. The Rcsb Protein Data Bank: A Redesigned Query System and Relational Database Based on the Mmcif Schema. *Nucleic Acids Res.* 2005, *33*, D233-D237.

85. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. et, al. Gromacs 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* 2013, *29*, 845-54.

86. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* 2008, *4*, 435-447.

87. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 1983, *79*, 926-935.

88. Press, W. H. Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge university press: 2007.

89. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* 1984, *81*, 3684-3690.

90. NosÉ, S. I. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* 2002, *100*, 191-198.

91. Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* 1984, *52*, 255-268.

92. Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* 1981, *52*, 7182-7190.

93. Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N· Log (N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 1993, *98*, 10089-10092.

94. Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 2008, *100*, 020603.

95. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et, al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* 1998, *102*, 3586-3616.

96. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* 2003, *24*, 1999-2012.

97. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A. Plumed: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.* 2009, *180*, 1961-1972.

98. Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* 1977,*23*, 187-199.

99. Souaille, M.; Roux, B. t. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* 2001, *135*, 40-57.

100. Pietrucci, F.; Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to Sh3 and Gb1. *J. Chem. Theory Comput.* 2009,*5*, 2197-2201.

101. Zagrovic, B.; Sorin, E. J.; Pande, V. Beta-Hairpin Folding Simulations in Atomistic Detail Using an Implicit Solvent Model. *J. Mol. Biol.* 2001, *313*, 151-69.

102. Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. A Recipe for the Computation of the Free Energy Barrier and the Lowest Free Energy Path of Concerted Reactions. *J. Phys. Chem. B* 2005, *109*, 6676-6687.

103. Paz, S. A.; Abrams, C. F. Free Energy and Hidden Barriers of the B-Sheet Structure of Prion Protein. *J. Chem. Theory Comput.* 2015, *11*, 5024-5034.

104. Chamachi, N. G.; Chakrabarty, S. Replica Exchange Molecular Dynamics Study of Dimerization in Prion Protein: Multiple Modes of Interaction and Stabilization. *J. Phys. Chem. B* 2016, *120*, 7332-7345.

105. Mukhopadhyay, S.; Krishnan, R.; Lemke, E. A.; Lindquist, S.; Deniz, A. A. A Natively Unfolded Yeast Prion Monomer Adopts an Ensemble of Collapsed and Rapidly Fluctuating Structures. *Proc. Natl. Acad. Sci. U. S. A.* 2007, *104*, 2649-2654.

106. Dalal, V.; Arya, S.; Bhattacharya, M.; Mukhopadhyay, S. Conformational Switching and Nanoscale Assembly of Human Prion Protein into Polymorphic Amyloids Via Structurally Labile Oligomers. *Biochemistry* 2015, *54*, 7505-13.

107. Dima, R. I.; Thirumalai, D. Exploring the Propensities of Helices in PrP^C to Form β-Sheet Using NMR Structures and Sequence Alignments. *Biophys. J. 83*, 1268-1280.

108. Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions. *BMC Struct. Biol.* 2009, *9*, 1-10.

109. Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* 2015, *119*, 5113-5123.

110. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell Jr, A. D. Charmm36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Meth* 2016, (doi:10.1038/nmeth.4067).

111. Kyte J.; Doolittle RF.; A Simple Method for Displaying the Hydrophathic Character of Protein. *J.Mol.Biol.*1982, 157,105-132

Chapter 4

Mechanism of Ubiquitin Unfolding in Chemical Chaperon

4.1 Overview

Protein stability is very important for its function. It has been shown that the protein environment such as ions, the small organic molecules have strong influence on protein stability, but their mechanism of interaction is still unclear. Here we have studied the stability of ubiquitin in presence of various amino acids (AA) and modified AA solution. Contrary to previous studies, our microsecond long well-tempered metadynamics simulation shows that these free AA and modified AA stabilize the native state of the protein. Moreover, they directly interact with charged (ARG, LYS, ASP and GLU) and polar AA of the protein. Among charged AAs, ARG shows maximum interaction with the chaperone containing free acetate groups (e.g., zwitterion GLY, Nter capped GLY and Nter capped ASP) and the reason for such interaction is because ARG can make a bidented interaction with acetate group.

Abbreviations: GLY (Glycine), ASP (Aspartic Acid), GLY (Glutamic Acid), ARG (Arginine), LYS (Lysine), GLN (Glutamine), SER (Serine), THR (Threonine), ASN (Asparagine), AA (Amino acid)

Note: The present work is in collaboration with **Kausik Chakraborty group** (IGIB, Delhi). They have performed the experiment to study the effect of chemical chaperone on protein stability and we have used molecular dynamics simulation to study the same. But here, we have discussed the observation from the molecular dynamic simulation only.

4.2 Introduction

We have shown in the previous chapter that a non-native state is more stable than the reported native structure of prion protein in water. The stability of the protein depends not only on its own internal interaction, the environment also plays an important role in its stability. Naturally, addition of any cosolvent would affect the protein stability.[1-4] It is shown that small organic molecules as cosolvent make a dramatic influence on the protein folding reaction without making any chemical modification. [5-10] Cosolvents have been used in vivo to stabilize and dissolve the proteins[11]. They are also used in many pharmaceutical and biotechnology industries to dissolve insoluble proteins and increase the life of protein[12-14].

The cosolvents which stabilize the proteins are known as chemical chaperon.[15-17] For example, yeast uses trehalose as cosolvent to protect its proteome against thermal denaturation and aggregation.[18-19] Also, some organisms inhabiting in deep sea use TMAO as protecting molecule against the denaturing effect of urea. [20-21] Some studies show that these cosolvents accumulate around the protein surface to stabilize the protein.[18-19] [20-21] Other studies show that these cosolvents destabilize the unfolded state of the protein and increase the folding rate. They do not interact with the folded protein.[22-24] Several mechanisms have been proposed to explain the effect of cosolvents on protein stability. These mechanisms are divided in two categories: (i) indirect mechanism in which cosolvent stabilizes the protein by ordering the water structure around protein, and (ii) direct mechanism in which cosolvent directly binds to the protein. For example, some studie have shown that the cosolvent such as TMAO increases the water ordering around the protein (indirect interaction), while another study on TMAO shows that it binds itself to the protein surface (direct interaction).

Bandhopadhya *et al.* have studied the effect of various chemical chaperone on folding of maltose binding protein through single-molecule F*ö*rster resonance energy transfer (smFRET).[25] Based on their observation, they have classified the chemical chaperons in two major classes: (1) chemical chaperons which

affect the refolding intermediates, and (2) chaperons that affect the transition state of protein folding.[25] For example, chemical chaperone like trimethylamine-N-oxide (TMAO), trehalose, etc. decrease the flexibility of refolding intermediates (RIs) and thus reduce the entropic barriers of folding. They do not interact with native state of the protein. These types of chaperons belong to class 1. On other hand, class 2 chaperons such as proline, serine, etc. make native like contact at the transition state which eventually decrease the transition barrier.[16] Therefore, the mechanism of how the cosolvent affect the protein stability is still not clear. Some observations suggest that the chapaerons get accumulated around protein surface, while others suggest that they destabilize the unfolded and transition state. Therefore, to add some more insight into the understanding of the mechanism of chaperone on protein stability, we have studied the effect of free amino acid and modified amino acid (GLY, Ctee caped GLY, Nter capped GLY and Nter capped ASP) on protein stability. To do so we have taken ubiquitin as a model protein. Ubiquitin is a small protein made of 76 amino acid (AA) residue and present in all eukaryotic organism. Ubiquitin has 4 ARG, 7LYS (positively charged), 5 ASP, and 6GLU (negatively charged) charged amino acid on its surface. The presence of both positive and negative AAs on protein surface enable the protein to interact with the free zwitterion AAs. In this chapter, we have performed the unfolding simulation of ubiquitin in presence of some natural and modified AA solutions. (see below section for detail)

4.3 Systems and Methods

(I) Preparation of Initial Configurations. The crystal structures of Ubiquitin protein (PDB ID 1UBQ) [26] were taken from the Protein Data Bank. The topology and initial coordinates were generated by the GROMACS software. [27-28] The initial structure of chaperones; GLY[25], Cter capped GLY (referred hereafter as CGLY), Nter capped GLY (referred hereafter as NGLY) and Nter capped Aspartic acid (referred hereafter NASP) were created by ambertools. [29] Fig 1 shows the structure of these amino acids.

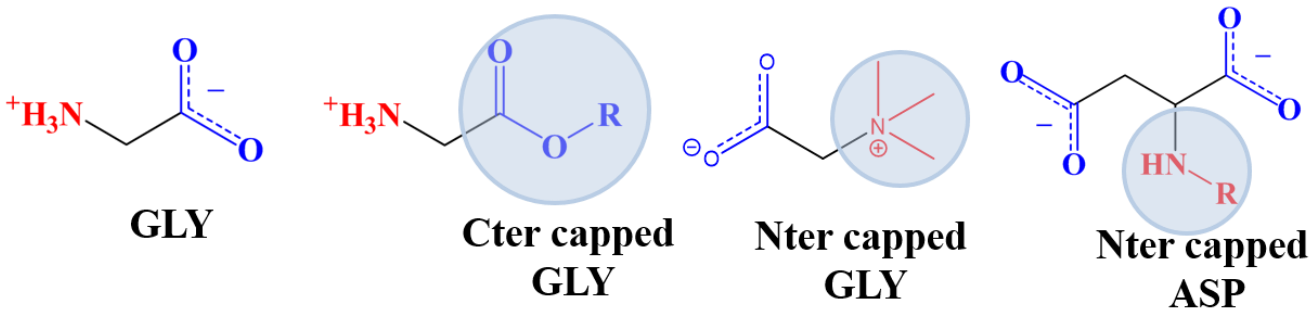

Figure 1: The chemical structures of chemical chaperone (GLY, Cter capped GLY, Nter capped GLY and Nter capped ASP) used in the present study.

(ii) Simulation Details. We have prepared five systems to study: (i) ubiquitin in water (SIM1), (ii) ubiquitin in 250 mM GLY solution (referred to as SIM2), (iii) ubiquitin in 250 mM CGLY solution (SIM3), (iv) ubiquitin in 250 mM NGLY solution (SIM4), (iv) ubiquitin in 250mM NASP solution(SIM5). The reason for choosing 250 mM concentration for chaperon is that at this concentration the chaperon activity is noticeable. The CGLY was capped by methyl, NGLY was capped by tri-methyl and NASP was capped by methyl. To create this system, the initial configurations of protein were put in a large cubic box of length 130 Å. It was then solvated with TIP4P water model[30], 150 mM ions (NaCl) were added maintain the physiological ion concentration and also to neutralize the system. Finally, the different AAs were added to created different systems. For each system, thereafter, the energy minimization was carried out by using steepest descent algorithm[31], followed by heating up to 300 K using the Berendsen thermostat[32] with coupling constant of 0.2 ps. During the heating process, the heavy atoms of the protein were restrained with the harmonic potential of 25 kcal/molÅ^{-2}. After heating, we annealed the solvent (water, ions, and chaperon) upto 400K with restraint on the protein. After that, a 20 ns long simulation at 300K was performed by restraining the protein atoms. Following this, a series of equilibration and energy minimization was carried out, where the restraint was reduced to 0.25 kcal/molÅ^{-2} in six steps. Finally, an unrestraint 20ns equilibration were performed at constant temperature 300 K and constant pressure 1 bar using the Nose-Hoover thermostat[33] and Parrinello-Rahman barostat[34], respectively, with

coupling constant 0.4 ps. For electrostatics, Particle Mesh Ewald (PME) [35] method was used with a long-range cutoff of 10 Å and van der Waals was treated with 10 Å cutoff. Integration time step for all simulations was kept at 2 fs.

After equilibration, we have performed Well-tempered metadynamics to explore the free energy surface of the ubiquitin unfolding in different chaperone solution. Simulation lengths for this final metadynamics simulation varied from 250 ns to 700 ns depending on the requirement of the convergence of the result. To capture the unfolding of protein we have used two reaction coordinates: (i) native contact (Nc) of the protein and (ii) radius of gyration (Rg). The details of the reaction coordinate is given in next section. For performing the well-tempered metadynamic[36] simulation, we used a bias factor of 15, hill height of 0.2 kJ/mol and the hills deposition rate was set at 3 ps. Gaussian widths for the native contact and radius of gyration were chosen to be 5 and 0.7 Å, respectively.

All the simulations were carried out using the GROMACS molecular dynamics program. For free energy calculations, we used the PLUMED patch to GROMACS[27-28].

4.4 Reaction Coordinate:

(I) Native Contact (N_c): The definition of native contact is same as discussed in chapter 3. We have considered all the contact between helix and β-sheet atoms and between all β-sheet atoms, if they are within 4.5 Å distance. N_c is helpful to separate the helices and β-sheet from each other's contact to drive unfolding.

(II) Radius of gyration (R_g): Radius of gyration is another measure of protein folding. A folded protein typically has lower R_g defined as, $R_g = \sqrt{\sum_i \frac{m_i(\vec{r_i}-\vec{r_c})^2}{M}}$. m_i is the mass and $\vec{r_i}$ is the position vector of the i^{th} atom. M is the total mass ($\sum_i m_i$) and $\vec{r_c}$ is the center of mass.

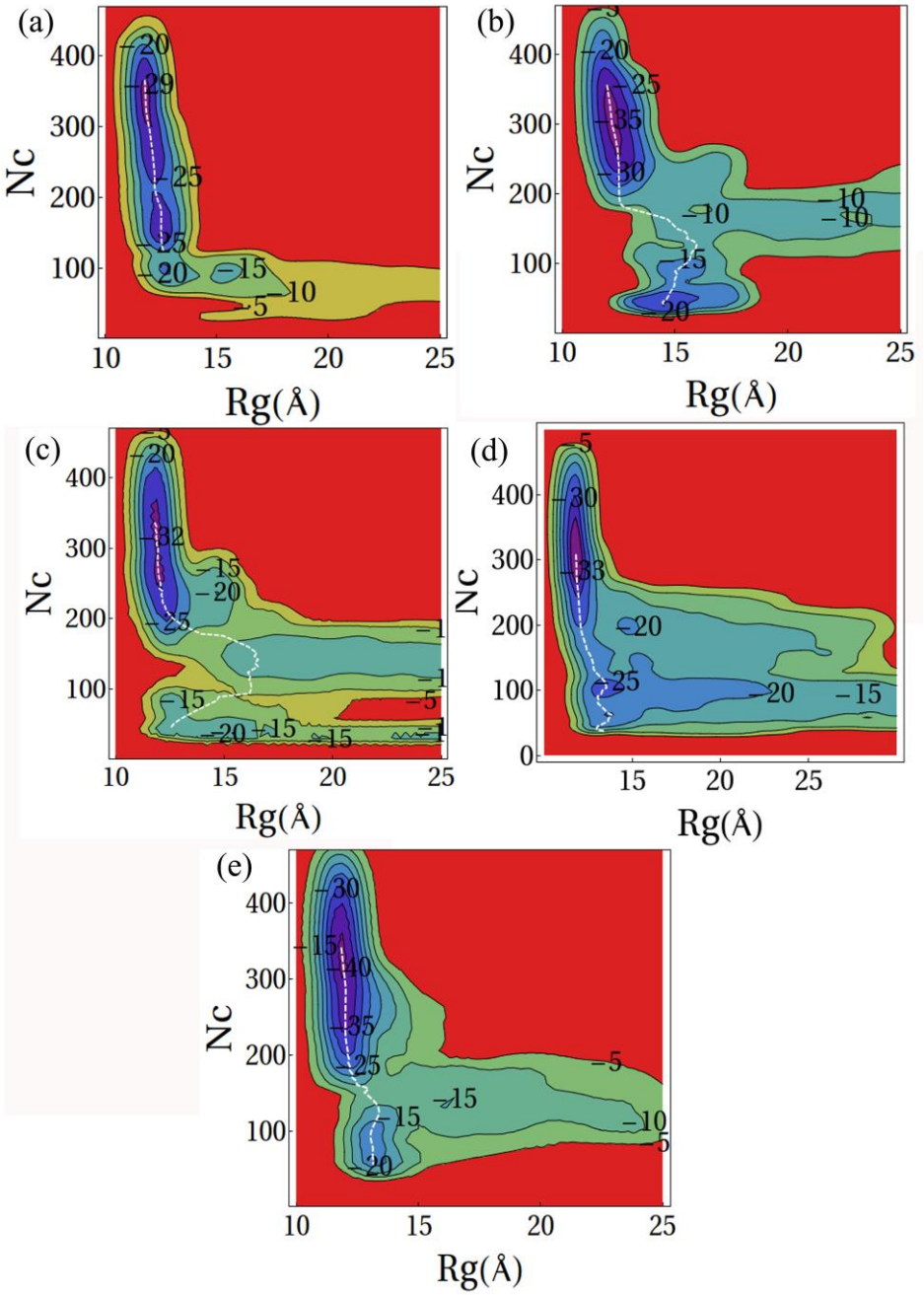

Figure 2: Free energy surfaces (FES) of ubiquitin unfolding in water and in different chaperone solution. All FESs are plotted against native contact (Nc) and R_g. (a) FES in water (b) FES in GLY solution. (c) FES in CGLY solution. (d) FES in NGLY solution. (e) FES in NASP solution. The white line on each surface shows the minimum free energy path.

RESULT AND DISCUSSION

4.5 Effect of Chaperone on Protein Stability

Figure 2 shows the free energy surfaces (FES) of ubiquitin unfolding in water and in different chaperone solution. The white line in each FES shows the minimum free energy path (MFEP) for unfolding. The native state of protein has 330 contacts and Rg of 12 Å. In all the simulations, the native state is the most stable state (lowest in free energy). FESs in Fig. 2 show that ubiquitin unfolding mechanism mostly follow the breaking of native contact followed by increase in R_g in almost all cases. All the system follows a different path of unfolding. Like, in pure water there are three major minima, first is at when Nc is 330 and Rg is 12 Å, second is at when Nc is 120 and Rg is 12.5 Å and third is at when Nc is 80 and Rg is 13 Å. In the GLY solution, there are only two major minima, one at when Nc is 330 and Rg is 12 Å and second is at when Nc is 50 and Rg is 15 Å. In the CGLY solution, there are three major minima, first is at when Nc is 330 and Rg is 12 Å, second minim is at when Nc is 300 and Rg is 15 Å and third minima is at when Nc is 50 and Rg is 12 Å. In NGLY and NASP solution, there are two major minima, one at the native state and other is when Nc is 100 and Rg is 12 Å in NGLY solution and Nc is 70 and Rg is 12 Å in NASP solution.

To compare the free energetic stability, we plotted the free energy profiles along these two reactions coordinate and also along the minimum free energy path (MFEP) in Fig. 3. FES shows that all the chaperons stabilize the native state of protein compared to pure water. While the native state stability is -30 kcal/mol for ubiquitin in water, it is -36 kcal/mol, -33 kcal/mol -34 kcal/mol and -42 kcal/mol for ubiquitin in GLY, CGLY, NGLY and NASP solutions, respectively. The reason for higher stability of ubiquitin in free GLY can be attributed to the fact that free GLY has one negatively charged acetate group (COO^-)

and one positively charged amine group (NH_3^+), which would interact with both the positive and negative AAs of the protein. However, the capped glycine NGLY and CGLY have only one free positive (amine group) and negative (acetate group) terminal, respectively. Therefore, they are limited to interact with either negative or positive AAs of the protein. The NASP, on the other hand, has two negatively charged acetate groups and it provides an increased native state stability of ~12 kcal/mol compared to protein in water. We have discussed the possible interaction of these chaperones in a later section.

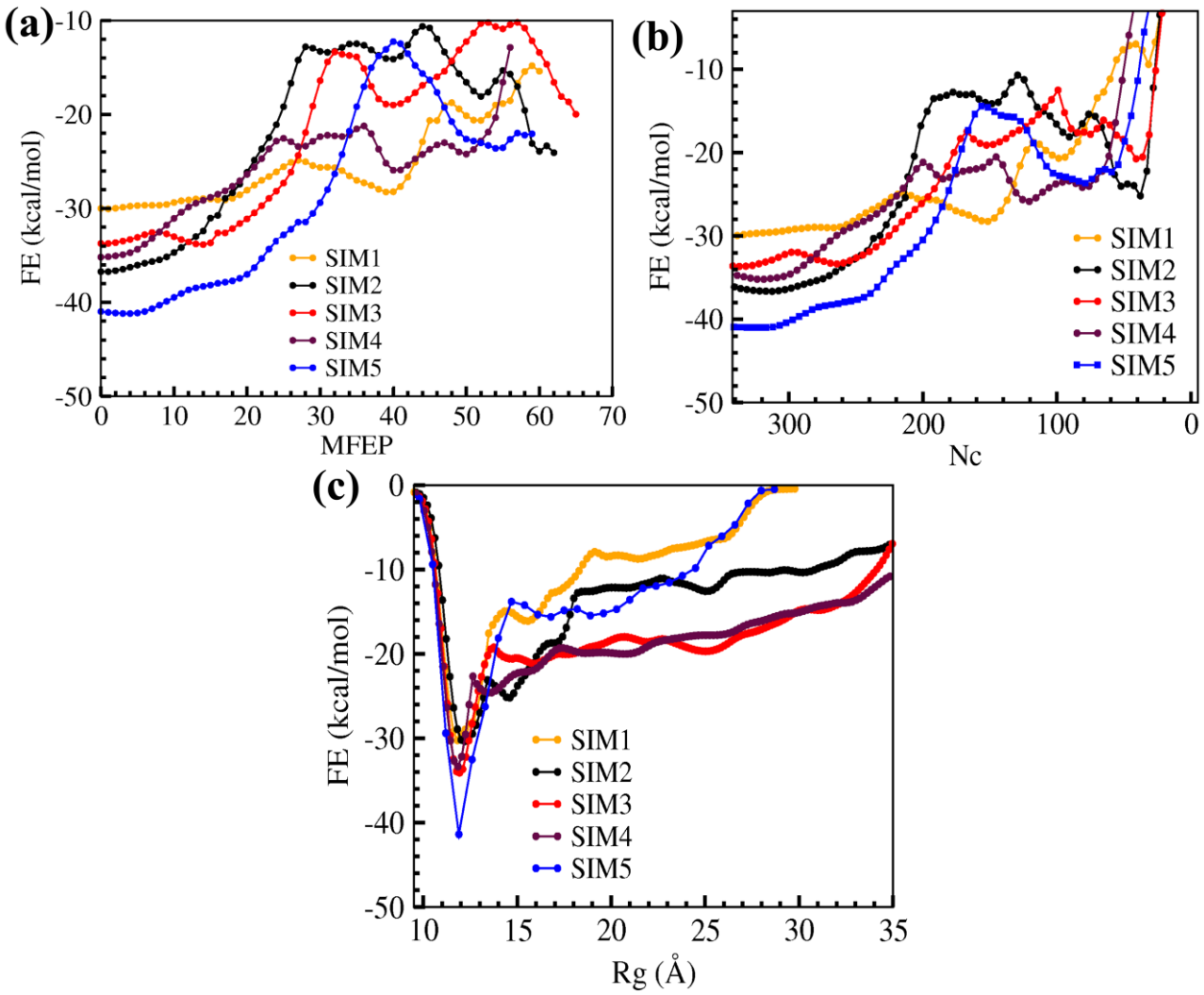

Figure 3: Free energy profile plotted against (a) MFEP, (b) Nc and (c) R_g.

First, we will discuss the stability of the native state and then we discussed the de-stability of unfolded state. We have considered all the frame which has Nc more than 300 and Rg is ~12.5 Å as native state. All the analysis is done over those frames only.

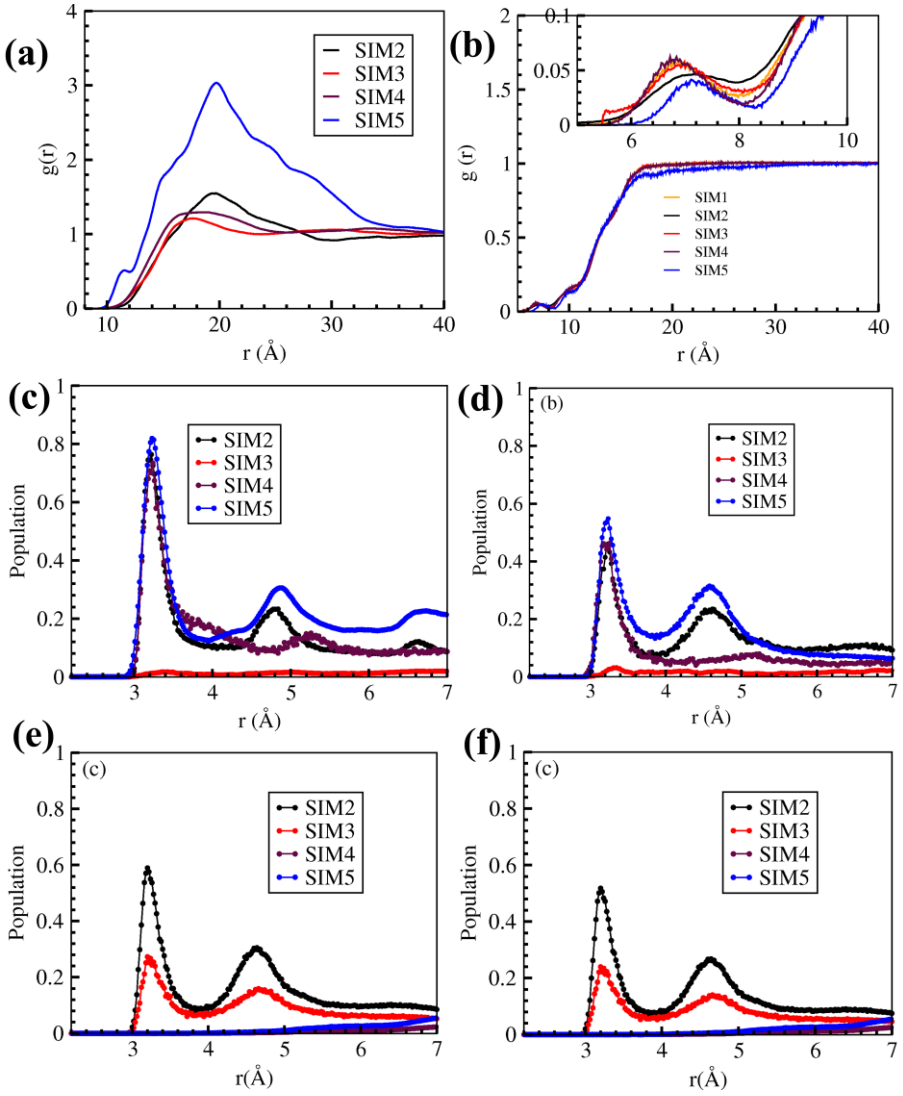

Figure 4: RDF of the (a) chaperone and (b) water from COM of the protein. The distance distribution of acetate oxygen atoms of the chaperone to the nitrogen atoms of all (c) ARG residues and (d) LYS residues of the protein. The distance distribution of amine nitrogen atoms of the chaperone to the oxygen atoms of (e) all ASP residues and (f) all GLU residues of the protein.

(I) Stablization of the Native State

To see the distribution of chaperone around the protein, we have calculated the radial distribution function (RDF) of chaperone from protein COM (Fig 4a). RDF signifies the average probability of finding a chaperone around the protein at a particular distance compared to that of an ideal gas system. The RDF shows that the chaperone is preferentially accumulated around protein surface (The peak height is greater than 1). We can observe that the peak of the RDF is around 20 Å for NASP and GLY while is a slightly smaller, around 17 Å for the CGLY and NGLY. However, the height of RDF is highest for NASP indicating that the number of NASP (blue color in Fig. 4a) is much more around the protein compared to other chaperons. It is followed by GLY, NGLY and CGLY (black, maroon and red color of Fig 4a). Similar result is also reflected from the RDF of water around the protein. The accumulation of chaperone around protein surface decreases the water content around the protein. Fig 4b shows the RDF of water from the COM of the protein. Because of the large size of the protein, RDF does not show any prominent peak. However, the inset of 4b shows that there are peaks at smaller value (~7 Å). This indicates that there are some pockets near the COM of the protein that are exposed to water. Except for NGLY, all the other AA has lower RDF than the pure water system (SIM1). The RDF of water is lowest amongst all. This indicates that NASP replaces water molecules at this level. Also, around 20 Å (Fig. 4b), the RDF of water is lowest. For other AA, it is not distinguishable at this distance because of smaller difference due to larger accessible volume. All the chaperons have charged residues – either positive, negative, or both. These free negative and positive termini of chaperone can interact with positive (ARG and LYS) and negative (ASP and GLU) AA of protein, receptively. To test this hypothesis, we have calculated the distribution of distance between the acetate oxygen and amine nitrogen atoms of chaperons to the same in the charged residues (ARG, LYS, ASP and GLU) and plotted in Fig 4c-f. Fig. 4c and 4d show the distance distribution of the chaperone acetate oxygen atom with the amine nitrogen atoms of ARG and LYS, respectively. Fig. 4e and 4f show the distribution of chaperons' nitrogen atoms with the oxygen atoms of ASP and GLU, respectively.

In all the distributions, the first peak is around 3.5Å (HB distance), which suggest that the chaperons are hydrogen bonded with protein AA. As mentioned above chaperone GLY (SIM2) has both positive and negative termini. Hence, it interacts with both the positive (ARG and LYS, black color in fig 4c-d) and negative AA (ASP and GLY, black color in Fig 4e-f). However, the chaperone CGLY (SIM3) has only positive terminal. Therefore it interacts with the negative AA (red color of fig 4e-f). The chaperone NGLY (SIM4) has only negative terminal, and it interacts with only positive AA (maroon color of fig 4c-d). Since the NASP has two free negative termini, it interacts with the positive AA (blue color in fig 4c-d) of protein much more strongly (80% time it is hydrogen bonded) than CGLY or NGLY. Generally, from the peak height of the distributions in Figs. 4c-f., we note that the chaperons interact more strongly to the positively charged AA residues of the protein compared to negatively charged ones. This suggests that the ARG and LYS of the protein play a major role in the chaperonic activity. Fig. 4c shows that there is significant probability (20%) to find NASP chaperone even 7 Å away from the ARG residues. The distance distribution from the ARG to chaperone shows the long-distance interaction upto 7Å. There is finite probability for other chaperons also. Therefore, we have calculated the number of chaperons within 7Å from the protein surface. Moreover, we have classified the number of chaperons around positive AA (ARG and LYS), negative AA (ASP and GLU) and polar AA (GLN, THR, SER and ASN) of protein. Since chaperons are charged, we have not considered the hydrophobic amino acid.

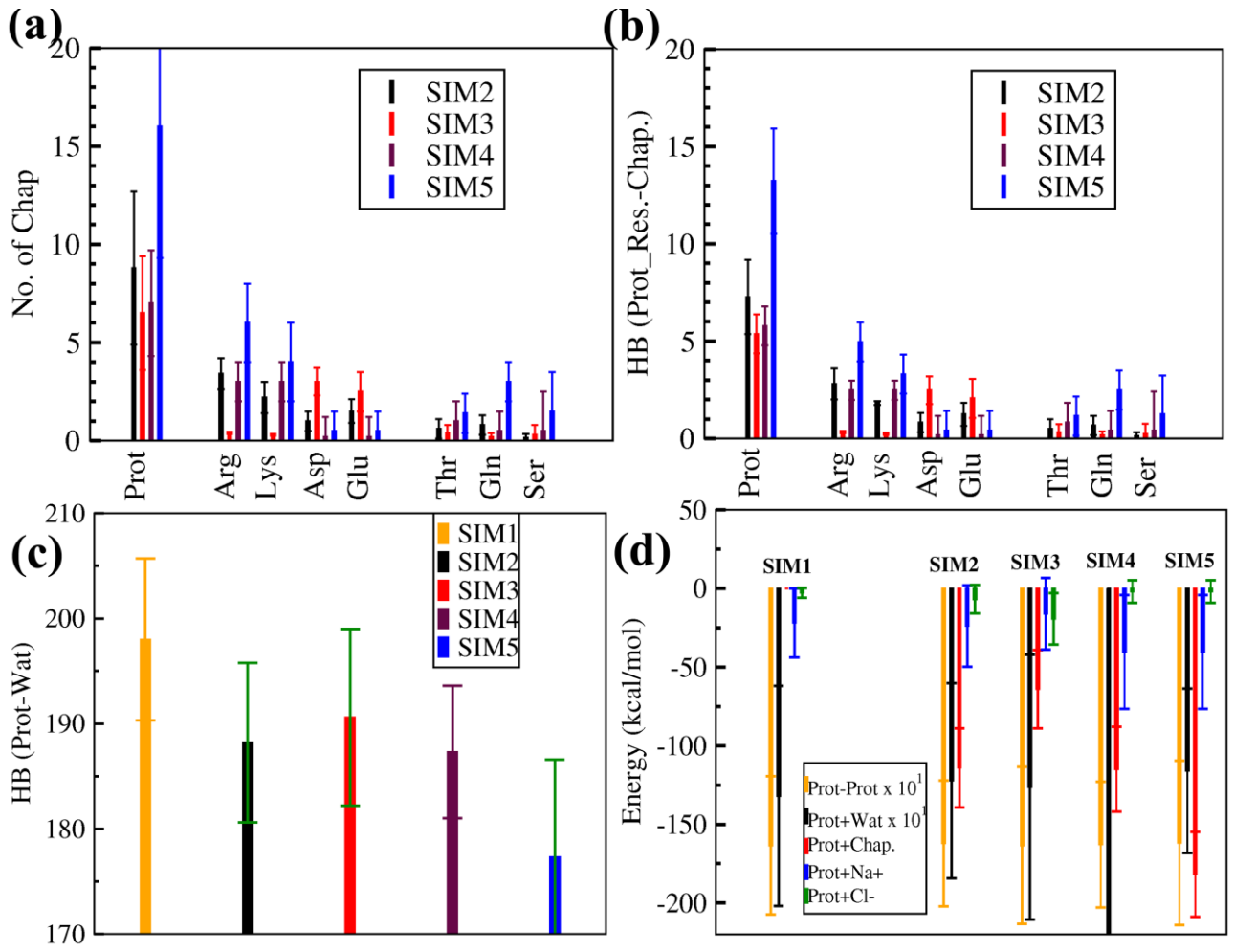

Figure 5: (a) Number of chaperons around protein and various other amino acids. (b) Number of hydrogen bonds between protein and chaperon. (c) Number of protein between protein and water in pure water and in chaperone solution. (d) Interaction energy between protein and water, chaperone and ions in pure water and in chaperone solution.

In Figure 5a, we show the number of different chaperones around protein and various other AA. Results from each chaperone is marked with different colors. Figure 5a shows that out of total 393 number of chaperones, there are 9 GLY (black bar), 7 CGLY (red bar), 8 NGLY (maroon color) and 16 NASP (blue bar) molecules around found within 7 Å of the protein surface. Figure 5b shows the number of hydrogen bond (HB) formed between the chaperone and proteins. We observe that the HB numbers follow the same pattern as the number of the chaperons (Fig. 5b). This indicates that the majority of chaperons around protein are hydrogen bonded with protein and they interact primarily with the polar and charged AA of

protein. The chaperone GLY majorly forms HBs with the positively AAs ARG and LYS (the black color of Fig. 5a and 5b) of protein and it has very less HBs with polar (THR, SER and GLN) and negatively charged AA (ASP and GLU black color of Fig. 5a, b) of the protein. On the contrary, the chaperone CGLY (due to the capping of Cter) does not make HB with ARG and LYS (red color of Fig. 5a, b). Instead, due to the presence of amine group, they form HB with the ASP and GLU AA of the protein (red color of fig 5a, b). However, this HB is less compared to the chaperone GLY. Hence, the stabilization due to the chaperone CGLY is less compared to chaperone GLY. In NGLY, Nter is capped. Therefore, it does not interact with ASP and GLU of protein, but it interacts with ARG and LYS. Therefore, it less HB compared to the chaperone GLY. Therefore, the stabilization due to the NGLY is also less compared to uncapped GLY. The NASP has no free amine group. Naturally, it does not form HB with the negative AAs of protein (blue color of fig 5a,b). But due to the presence of two free acetate groups, it forms strong HB with the ARG and LYS AA of protein (blue color of Fig 5a,b). It also makes significant number of HB with the other polar residue (THR, GLN, and SER, blue color of fig 5a,b), compared to other chaperons. Formation of more number of HBs between NASP and ARG, LYS and polar AA of protein may lead to a more stable native state of protein compared to another chaperon. The interaction of chaperons with protein causes a decrement in water number around the protein as mentioned earlier. Here, we have plotted the number of HBs formed between water molecules and proteins as shown in Fig. 5c, where we find that the number of HB between protein and water is less in chaperone solution compared to pure water system (Fig 5c). Since NASP has strongest interaction (in terms of number of HBs) with the protein, the number of water is least in the NASP system (Fig 5c). The trend is similar for other chaperones also stronger the interaction, lesser is the amount of water found around the protein. This indicates that chaperons replaces the water and interacts directly with the protein.

Next, we calculated the internal energy of the protein and interaction energy of the protein with water, ion and chaperons as shown in Fig. 5d and also summarized in table 1. Since the number of water molecules

is huge, the interaction energy between protein and water is a highly fluctuating quantity. Therefore, to calculate this particular interaction, we have considered a subsystem comprising the protein and water molecules (~400 water molecule), ions and chaperone within 10Å distance from the protein surface. Fig 5d shows that the chemical chaperone interacts strongly with the protein. The chaperone NASP has the lowest interaction energy (~ -181 kcal/mol), followed by GLY (~ -114 kcal/mol) and both capped GLY (~ -64 kcal/mol) (Fig 5d red color and Table 1). The interaction energies of chaperones with protein are majorly electrostatic. Interestingly, the interaction of chaperone with protein decrease the protein internal energy, because the AA which was initially interacting with other AA of protein are now interacting with chaperons. The greater the interaction with chaperone lesser the protein internal energy (orange color of Fig 5d and Table 1). The chaperone NASP and GLY interact more with protein therefore, they less protein internal energy whereas, the chaperone NGLY and CGLY interact less with protein and therefore decrease less protein internal energy. The accumulation of chaperone around protein decrees the water no. which result in less interaction energy with water in chaperone solution compare to pure water (blue in Fig 6d). Interestingly, when we add all the energies (Prot-Prot, Prot-Wat-Ion, Prot-Chap.) and subtract from the pure protein-water system energy, we get the 8.7 kcal/mol, 4.6 kcal/mol, 5.4 kcal/mol and 17.9 kcal/mol stabilization for GLY, CGLY, NGLY and NASP system compare to the pure water system. This stabilization energy is approximately the same as free energy stabilization of native state in chaperone solution.

Table 1: Interaction energy of protein with itself and with surrounding species of different chaperone solution. All enegy units are in kcal/mol.

System	SIM1	SIM2	SIM3	SIM4	SIM5
$E_{Prot-Prot}$	-1634.1	-1621.2	-1630	-1628	-1618.5
$E_{Prot-Chap}$	--	-113.9	-63.9	-64.0	-181.9
$E_{Prot-Wat-Ions}$	-1344.5	-1250.2	-1289.3	-1292.0	-1196.1
E_{Total}	-2978.6	-2985.4	-2983.2	-2984.0	-2996.5
ΔE_{Total} (Wat-Chap) Sys.	0.0	-8.8	-4.6	-5.4	-17.9

In summary, our observation suggests that chaperones get accumulated around protein and stabilize the native state by making hydrogen bonds and electrostatic interactions. The chaperone with more free acetate groups has more effect on protein stability compared to the ones with free amine group.

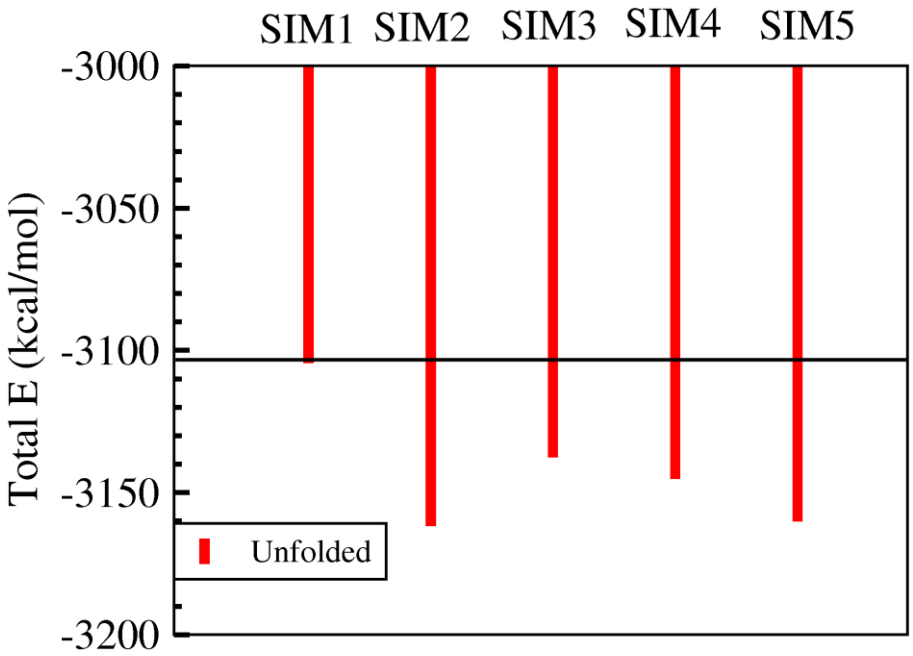

Figure 6: Total interaction of protein with surrounding in pure water, GLY, CGLY, NGLY and NASP solution. The black line shows the total interaction energy in the pure water system. It is drawn to compare the stability of unfolded state compare to the pure water system.

(II) Destabilization of the Unfolded State:

Free energy of protein folding shows that the chaperone stabilizes the protein native state and simultaneously destabilizes the unfolded state except for the NASP system. The previous discussion shows that the stabilization of the native state is due to the direct binding of chaperones to surface AA of the protein. Next, we have investigated the reason for the destabilization of the unfolded state. The unfolded state is not the same in all systems. Therefore, to compare amongst different systems, we have considered all the frames less than 120 Nc as an unfolded state. Note that, typically radius of gyration is used as a measure of the unfolded state. However, we observe that change in the native contact happens even before the change in R_g. Therefore, we used N_c as a measure of the unfolded state. It can also be regarded as molten globule state. All the analyses for the unfolded state are done over all these frames. First, we have calculated the total energy (protein + surrounding) of the unfolded state for each system. Figure 6 shows that

the unfolded state is more stable in chaperone solution compared to the pure water system. Therefore, it is not the energetic factor, rather the entropic one that may govern the reason for de-stability of unfolded state in chaperone solution. In fact, it has been shown that the unfolded state is stable to due to entropy and the native state enthalpically stable. [37-38] However, it is very computationally difficult to calculate the entropy of proteins because of a large number of the degree of freedom. Therefore, we have tried to provide an alternate estimate of entropy of proteins by calculating the flexibility of the amino acids in the unfolded state for each system. To do so, we have calculated the NMR order parameter (S^2), which measures the rotation of N→H vector of the protein backbone. It varies from 1, a perfectly rigid residue, to 0, a completely flexible residue. Fig. 7 shows the order parameter of the unfolded state in water and in different chaperone solution.

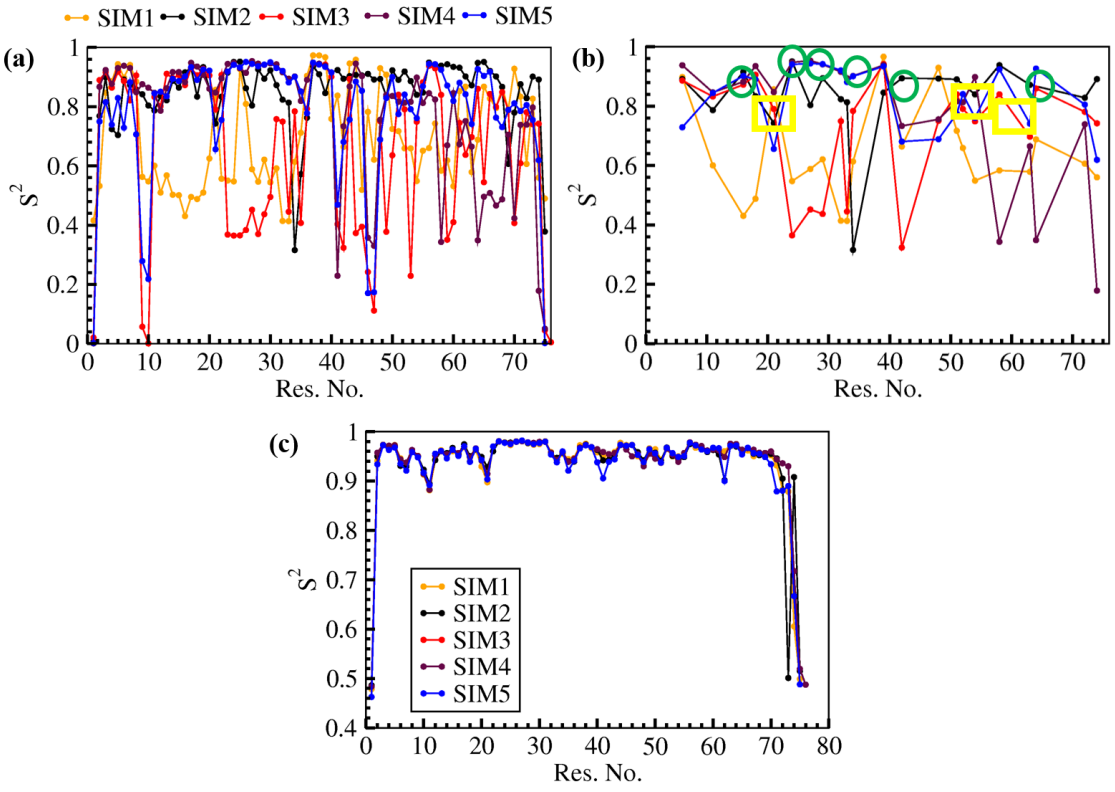

Figure 7: NMR order parameter of protein amino acids in the unfolded state in water and in different chaperones: (a) The order parameter for (a) all AA, (b) for charged AA, and (c) for the folded state. Green colored circles represent the order parameters of ARG and LYS and yellow boxes represent the order parameter for ASP and GLU.

Fig 7 shows the order parameter of AA in different chaperone solution and in pure water. The S^2 value of protein AA is less in pure water (protein AA are more flexible), but in chaperone solution, the S^2 of many AA (mostly the charged AA of protein) is high (~ 0.9) except for a few residues. Further analysis shows that charged AA which was flexible in pure water, become rigid in chaperone solution. Fig 7b shows the S^2 values for charged residues (ARG, LYS, ASP and GLU). In NASP and NGLY solutions, ARG and LYS AAs are very rigid (green circle in fig 7b) compared to that pure water (orange color in fig 7b). Similarly, in CGLY, ASP and GLU AAs are rigid. In the GLY solution, both positive and negative AAs have lesser flexibility (Fig 7b). The high value of S^2 in chaperone solution shows that the protein AA is less flexible in the unfolded state and hence may have the lesser entropy. The less flexibility of AA in chaperone solution would be the reason for destability of unfolded state. Note that, the except for the terminal residues, the S^2 of folded state is same in all solution including pure water (Fig. 7c). Therefore, we can conclude that chaperons do not effect flexibility of the folded state.

4.6 Conclusion:

Protein's stability highly depends on the surrounding environment. Here we have observed that the charged monomeric AA stabilize the protein in its native state and thus act as chemical chaperones. These chaperones directly bind (through HB and electrostatic) to the protein surface and stabilize the native state. Moreover, we have observed that primarily the charged and polar AAs in the protein surface interact with the chaperone. The interaction depends on the type of chaperone. Since GLY has both negative and positive charge, it interacts with both positive (ARG, LYS) and negatively charged AA (ASP, GLU) of protein, whereas the NASP (Nter capped ASP) only interacts with the ARG and LYS AA of protein. Further, the stabilization of the native state depends on the extent of interaction of chaperon. Stronger interaction of the protein with chaperone leads to a more the native state of a protein is stable. The NASP has the highest interaction with protein followed by GLY, NGLY and CGLY and hence the protein is

more stable in NASP solution followed by GLY, NGLY and CGLY solutions. Interestingly, we have observed the among the all charged and polar AA of protein surface, ARG has more interaction with chaperone. The reason could be that the ARG can make bidentate kind interaction with free acetate group of chaperon. Therefore, chaperones with free carboxylate groups, e.g. NASP, have stronger chaperonic effect on the proteins.

Reference:

1. Fujita, Y.; Miyanaga, A.; Noda, Y., Effect of Alcohols on the Thermal Denaturation of Lysozyme as Measured by Differential Scanning Calorimetry. *Bulletin of the Chemical Society of Japan* **1979,** *52* (12), 3659-3662.
2. Fujita, Y.; Noda, Y., The Effect of Organic Solvents on the Thermal Denaturation of Lysozyme as Measured by Differential Scanning Calorimetry. *Bulletin of the Chemical Society of Japan* **1983,** *56* (1), 233-237.
3. Velicelebi, G.; Sturtevant, J. M., Thermodynamics of the denaturation of lysozyme in alcohol-water mixtures. *Biochemistry* **1979,** *18* (7), 1180-1186.
4. Fu, L.; Freire, E., On the origin of the enthalpy and entropy convergence temperatures in protein folding. *Proceedings of the National Academy of Sciences* **1992,** *89* (19), 9335-9338.
5. Auton, M.; Bolen, D. W., Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (42), 15065-15068.
6. Lee, J. C.; Timasheff, S. N., The stabilization of proteins by sucrose. *Journal of Biological Chemistry* **1981,** *256* (14), 7193-7201.
7. Makhatadze, G. I.; Privalov, P. L., Protein interactions with urea and guanidinium chloride. *Journal of Molecular Biology* **1992,** *226* (2), 491-505.
8. Bolen, D. W.; Baskakov, I. V., The osmophobic effect: natural selection of a thermodynamic force in protein folding1. *Journal of Molecular Biology* **2001,** *310* (5), 955-963.
9. Auton, M.; Bolen, D. W., Additive Transfer Free Energies of the Peptide Backbone Unit That Are Independent of the Model Compound and the Choice of Concentration Scale. *Biochemistry* **2004,** *43* (5), 1329-1342.
10. Hong, J.; Capp, M. W.; Anderson, C. F.; Saecker, R. M.; Felitsky, D. J.; Anderson, M. W.; Record, M. T., Preferential Interactions of Glycine Betaine and of Urea with DNA: Implications for DNA Hydration and for Effects of These Solutes on DNA Stability. *Biochemistry* **2004,** *43* (46), 14744-14758.
11. Fu, L.; Freire, E., On the origin of the enthalpy and entropy convergence temperatures in protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **1992,** *89* (19), 9335-9338.
12. Brown, C. R.; Hong-Brown, L. Q.; Biwersi, J.; Verkman, A. S.; Welch, W. J., Chemical chaperones correct the mutant phenotype of the ΔF508 cystic fibrosis transmembrane conductance regulator protein. *Cell Stress & Chaperones* **1996,** *1* (2), 117-125.
13. Brown, C. R.; Hong-Brown, L. Q.; Welch, W. J., Correcting temperature-sensitive protein folding defects. *The Journal of Clinical Investigation 99* (6), 1432-1444.
14. Tatzelt, J.; Prusiner, S. B.; Welch, W. J., Chemical chaperones interfere with the formation of scrapie prion protein. *The EMBO Journal* **1996,** *15* (23), 6363-6373.
15. Record Jr, M. T.; Courtenay, E. S.; Cayley, D. S.; Guttman, H. J., Responses of E. coli to osmotic stress: large changes in amounts of cytoplasmic solutes and water. *Trends in Biochemical Sciences* **1998,** *23* (4), 143-148.
16. Record Jr, M. T.; Courtenay, E. S.; Cayley, S.; Guttman, H. J., Biophysical compensation mechanisms buffering E. coli protein–nucleic acid interactions against changing environments. *Trends in Biochemical Sciences* **1998,** *23* (5), 190-194.
17. Yancey, P. H.; Clark, M. E.; Hand, S. C.; Bowlus, R. D.; Somero, G. N., Living with water stress: evolution of osmolyte systems. *Science* **1982,** *217* (4566), 1214.

18. Felix Carla, F.; Moreira Clarisse, C.; Oliveira Mylene, S.; Sola-Penna, M.; Meyer-Fernandes José, R.; Scofano Helena, M.; Ferreira-Pereira, A., Protection against thermal denaturation by trehalose on the plasma membrane H+-ATPase from yeast. *European Journal of Biochemistry* **2001,** *266* (2), 660-664.
19. Sola-Penna, M.; Meyer-Fernandes José, R., Protective Role of Trehalose in Thermal Denaturation of Yeast Pyrophosphatase. In *Zeitschrift für Naturforschung C*, 1994; Vol. 49, p 327.
20. Yancey Paul, H.; Somero George, N., Methylamine osmoregulatory solutes of elasmobranch fishes counteract urea inhibition of enzymes. *Journal of Experimental Zoology* **1980,** *212* (2), 205-213.
21. Zou, Q.; Bennion, B. J.; Daggett, V.; Murphy, K. P., The Molecular Mechanism of Stabilization of Proteins by TMAO and Its Ability to Counteract the Effects of Urea. *Journal of the American Chemical Society* **2002,** *124* (7), 1192-1202.
22. Bolen, D. W.; Baskakov, I. V., The osmophobic effect: natural selection of a thermodynamic force in protein folding11Edited by D. Draper. *Journal of Molecular Biology* **2001,** *310* (5), 955-963.
23. Macchi, F.; Eisenkolb, M.; Kiefer, H.; Otzen, D. E., The Effect of Osmolytes on Protein Fibrillation. *International Journal of Molecular Sciences* **2012,** *13* (3), 3801-3819.
24. Street, T. O.; Bolen, D. W.; Rose, G. D., A molecular mechanism for osmolyte-induced protein stability. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (38), 13997-14002.
25. Dandage, R.; Bandyopadhyay, A.; Jayaraj, G. G.; Saxena, K.; Dalal, V.; Das, A.; Chakraborty, K., Classification of Chemical Chaperones Based on Their Effect on Protein Folding Landscapes. *ACS Chemical Biology* **2015,** *10* (3), 813-820.
26. Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J., Structure of ubiquitin refined at 1.8Åresolution. *Journal of Molecular Biology* **1987,** *194* (3), 531-544.
27. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.
28. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **2008,** *4* (3), 435-447.
29. Case, D. A.; Cheatham, T. E.; Darden, T. O. M.; Gohlke, H.; Luo, R. A. Y.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005,** *26* (16), 1668-1688.
30. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.
31. Tribbey, W., Numerical Recipes: The Art of Scientific Computing (3rd Edition) is written by William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, and published by Cambridge University Press, © 2007, hardback, ISBN 978-0-521-88068-8, 1235 pp. In *Modern database systems*, Won, K., Ed. ACM Press/Addison-Wesley Publishing Co.: 1995; pp 30-31.
32. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
33. NosÉ, S., A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* **2002,** *100* (1), 191-198.
34. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981,** *52* (12), 7182-7190.
35. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.
36. Barducci, A.; Bussi, G.; Parrinello, M., Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical review letters* **2008,** *100* (2), 020603.

37. Privalov, P. L.; Makhatadze, G. I., Contribution of Hydration to Protein Folding Thermodynamics: II. The Entropy and Gibbs Energy of Hydration. *Journal of Molecular Biology* **1993,** *232* (2), 660-679.
38. Sturtevant, J. M., Heat capacity and entropy changes in processes involving proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1977,** *74* (6), 2236-2240.

Chapter 5

Molecular Origin of DNA Kinking by Transcription Factors

5.1 Overview

The binding of transcription factor (TF) proteins with DNA causes bending in DNA, in most cases along with bending a kinking is also present in DNA. Here, we investigate the molecular origin of the DNA kinks observed in the TF-DNA complexes using small molecule intercalation pathway, crystallographic analysis, and free energy calculations involving four different transcription factor (TF) protein-DNA complexes. We find that although protein binding may bend the DNA, bending alone is not sufficient to kink the DNA. We show that partial, not complete, intercalation is required to form the kink at a particular place in the DNA. It turns out that while amino acid alone can induce the desired kink through partial intercalation, protein provides thermodynamic stabilization of the kinked state in TF-DNA complexes.

Abbreviations: IAA (Intercalating Amino Acid), IBP (Intercalating Base Pair), AA (Amino Acid), TF (Transcription factor)

5.2 Introduction

The conformational flexibility of DNA is very important for essential biological functions like transcription, replication, DNA repair and many more. Many factors evoke DNA bending like sequence composition (PollyA are more flexible and pollyC are more rigid), the environment around DNA (multivalent salts bend the DNA) or in many cases a protein binding. Bending induced by the proteins is very important for a certain function. For example, RNA polymerase requires a collection of proteins called transcription factors (TF) to activate or repress the transcription process in eukaryotes.[1] These TF's bind to DNA and bend/kink the DNA to reduce the distance between the distal parts of DNA to regulate co-operatively the binding of RNA polymerase[2-3]. See Fig 1. For more detail.

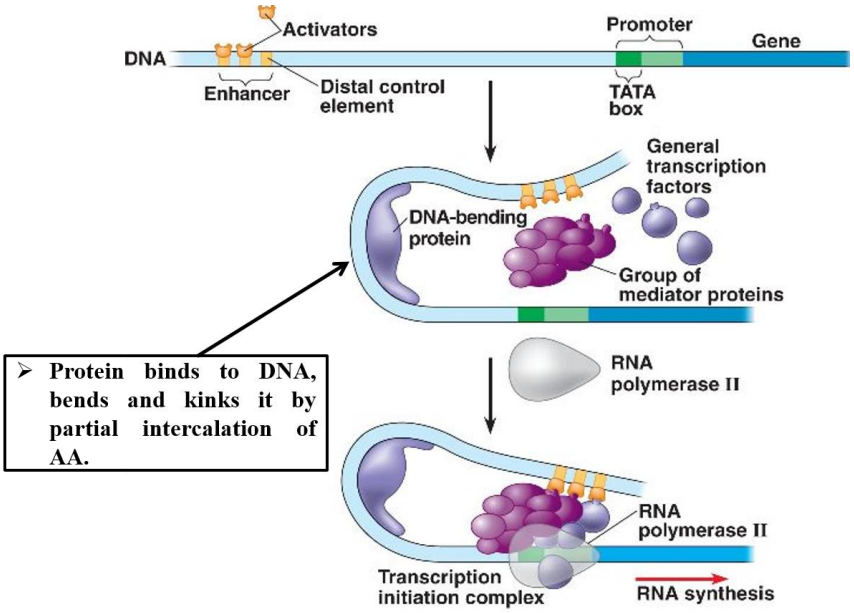

Figure 1: The role of DNA bending and kinking in a biological process. The long genome DNA sequence has the promoters and activators region, which has to come close to start replication, here is the advantage of DNA bending a group of protein (called transcription factors) bind to DNA and bend it by an intercalation of AA. The image is adopted from the "Advanced Transcription Regulation" by Landry Shipman.

DNA bending by proteins is broadly classified in two class[4]. In class 1, DNA bends away from the protein whereas in class 2, it bends towards the protein. In class 1 bending, an intercalation of the hydrophobic side chain of the binding protein has been observed which bend the DNA away from binding protein[3], whereas, in class 2, the positively charged amino acids present in the protein neutralize the negative charge of DNA backbone which eventually bends the DNA toward the protein.[5] Bending of DNA are happened by small but added changes in the Roll angle over all the DNA BP, which give a particular curvature of its helical axis. It can also happen through a sudden change of Roll angle at a particular base pair[6-7], defined as kink[8]. It was predicted[6] and shown through statistical analysis[7] that the kinking is an easier route to bend the DNA and kinking is unfavourable .[9-11] A recent computational study indicated that the change in Roll angle by 30° requires more than 4 kcal/mol free energy cost.[12] Kinks are very common in DNA-Protein complexes[13], where the stabilization is believed to come from the intercalation of AA[3]. In the absence of intercalation, kink can also form by disruption of base pairing.[14-15]

The focus of the present paper is to understand the origin of the formation kinks caused by protein-DNA intercalation. Nature is widespread with examples where protein-associated DNA bending and kinking are observed[7, 16], primarily in specific protein-DNA interactions[17]. Some examples of kinked protein-DNA complexes are Sac7d-DNA[18], SOX4-DNA[19], SRY-DNA[20-21], IHF-DNA[22], and NF-Y-DNA[23], TATA box[24], where DNA is bend and an intercalation of AA is present which cause high roll in DNA BP. Computational study of daunomycin and proflavin intercalation showed that at near the transition state, when the drug intercalates partially, the intercalating base pairs of DNA acquire high Roll angle (~40^0) and DNA bends significantly (10%).[25-26] Interestingly, DNA becomes straight with normal Roll angle when the drug completely intercalates. Therefore, partial intercalation, not the complete one, of daunomycin and proflavin

induces the kink in DNA. Experimental biophysical studies of small molecule induced DNA bending has shown that the intercalation of phenazine derivative aposafranine[27] and ruthenium complexes[28] cause kink in DNA. However, only partial intercalation was argued to be not enough to cause DNA kink in case of enantiospecific ruthenium complexes.[29]

So, with this introduction it is clear that bending and kinking are very important for TF's to function, the molecular origin of kink formation is not clear. Based on the observation that even a small molecule can kink a DNA and the fact that many protein-DNA complexes possess an intercalated AA at the kinked site, we hypothesize that partial intercalation of AA alone may be responsible for DNA kinking in protein-DNA complexes using TF proteins. Using four such protein-DNA complexes and free energetic study, we show it is indeed the case. We will also show that the protein, and not intercalating AA, helps in stabilizing the kinked state of the DNA.

5.3 Design of the Study

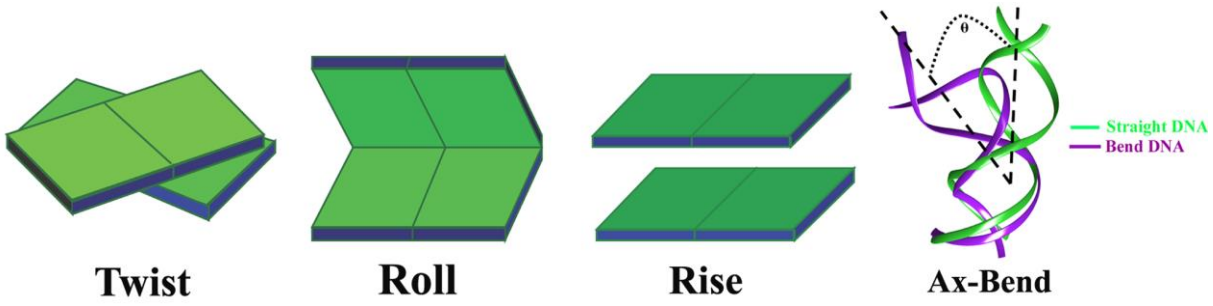

Figure 2: Schematic representation of Roll angle, Rise, Twist, and Ax-bend angle.

The definition of partial intercalation, however, is ambiguous and may depend on the intercalating agent. Since it has been shown that the Rise of the intercalating base pairs increases monotonically along the intercalation process[30], here we have considered Rise as the progress of intercalation. Roll is already established as a measure of kink in literature[8, 31]. Fig. 2 shows the schematic picture

of these DNA parameters. To see the relation between rise and roll angle in the intercalation process, we have analyzed the trajectory of metadynamics simulation of proflavine intercalation into DNA through minor groove reported recently by us[30], in Fig. 3 (a) we plot the Roll angle against the corresponding Rise of intercalating base pairs over whole metadynamic trajectory.

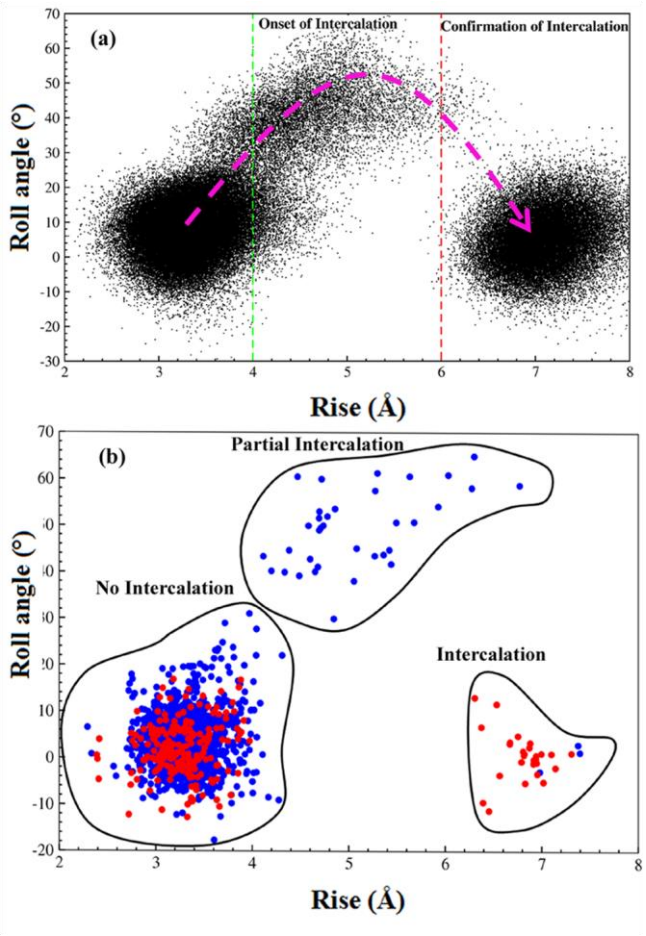

Figure 3: (a) Roll angle of the intercalating base pairs C6pG7 with the corresponding Rise value during a metadynamics simulation of proflavine intercalation into a twelve base pair DNA[30]. The curve arrow shows the overall intercalation pathway. The vertical dashed lines indicate the intermediate region is where the drug is partially intercalated. (b) Analysis of base pair step parameters from 38 protein-DNA (blue color) and 15 drug-DNA crystal structures (red color). Three different regions are encircled.

In Fig. 3a Magenta arrow shows the progress of roll and rise over trajectory, which starts with a normal DNA parameter (low Roll and Rise), goes through a kinked DNA state (high Roll and

medium Rise) ,which happened by partially intercalation of proflavin, to the final straight DNA and unkinked intercalated state (high Rise, low Roll). The two dashed line in Fig 2a shows the intermediate Rise value (4-6 Å), which correspond to Roll angle greater than 30°, beyond the standard distribution of Roll angle in regular B-DNA[32]. Therefore, we define Roll angle of a base pair step higher than 30° as kink and the corresponding Rise value, 4-6 Å, as the region of partial intercalation.

Fig 3a is observation over an MD trajectory, therefore to compare the MD result with the experiment, we have analyzed the Rise and Roll of 38 DNA-Protein systems (blue color) and 15 DNA-ligand systems (red color). These data are shown in Fig. 3(b) and table 1 shows the list of all the DNA-protein/drug system with their corresponding rise and roll value. Fig.3 (b) mirrors the observation obtained from an MD simulation. The difference, however, is that while Fig. 3(a) shows a dynamic process, Fig. 3(b) is the static crystal structure.

Table 1(a). DNA-protein complex (partial intercalation of AA)

Systems	**Systems (PDB ID only)**	**Rise (Å)**	**Roll (°)**
1	1AZP	5.3	61.3
2	1BNZ	5.9	54.1
3	1CDW	4.7	53.0
4	1CKT	5.6	60.1
5	1GDT	4.4	40.0
6	1IHF	6.7	59.0
7	1JFI	4.7	40.07
8	1JGG	3.7	40.9
9	1QN4	4.7	50.0
10	1QRV	4.7	51.6
11	1VOL	4.2	40.2

12	1WTO	5.3	57.5
13	1WTQ	6.3	58.1
14	1YTF	4.8	52.0
15	2LEF	3.8	36
16	3U2B	5.7	56.7
17	4AWL	4.4	44.7
18	4EUW	5.0	49.5
19	2GES	4.6	50.0
20	1AIS	5.2	43.2
21	1BD4	5.5	49.0
22	1GDT	4.3	40.0
23	1GTO	5.4	45.1
24	1JFI	4.1	43.4
25	1K3W	7.3	45.0
26	1MSW	5.9	31.0
27	1NG9	6.3	58.1
28	1NGM	4.8	51.7
29	1NH2	5.1	49.0
30	1NVP	5.0	51.7
31	1O3T	4.8	40.1
32	1OH5	4.7	53.0
33	1OWF	6.9	62.0
34	1QQB	5.2	49.1
35	1AZ0	4.5	50.0

(b) DNA-Protein complex (full intercalation of AA)

Systems	Systems (PDB ID only)	Rise (Å)	Roll (°)
1	2C7O	7.0	-2.0
2	1EWN	7.4	-1.0
3	2HR1	7.4	1.3

(c) DNA-drug/ligand

Systems	Systems (PDB ID only)	Rise (Å)	Roll (°)
1	1C9Z	7.0	1.1
2	1DZ1	6.8	-0.6
3	1G3X	7.1	1.0
4	1IMR	7.0	-5.6
5	1K9G	7.0	1.4
6	1MTG	6.4	6.7
7	1MXK	6.3	13.0
8	1NAB	7.3	1.1
9	1P96	6.0	-4.1
10	1R68	7.1	-2.0
11	1X95	6.5	-11
12	1Z3F	6.9	1.4
13	2MG8	6.3	-9.1
14	367D	6.8	1.8
15	4BZU	6.9	-3.3

Table 1(c) shows that the DNA-ligand/drug systems have normal Roll angle in full intercalation state which represents that the kink has been disappeared upon full intercalation. The complete intercalation of protein-DNA complexes are not available in TF's (as it would not probably have

served its purpose then), but it has been observed in some other protein-DNA complexes (Table 1 (b)) resulting in high Rise and low Roll (Fig. 1(b)). Like the DNA-Drugs, a complete intercalation is not possible because in DNA-Protein complex, intercalating AA is connected with the backbone of protein and it is impossible to intercalate completely. However, in some DNA-Protein complexes, a complete intercalation of AA in DNA is accomplished via flipping out of a base pair.

Extending the above results to TF-DNA systems, we hypothesize that the partial intercalation of AA may be responsible for kinks observed in TF-DNA complexes. We also wanted to understand whether partial intercalation stabilizes the kink in DNA. To address this, we chose to study four TF-DNA complexes: (i) SOX4-DNA[19], (ii) SRY-DNA[20-21], (iii) IHF-DNA[22], and (iv) NF-Y-DNA[23]. SRY is gene regulatory protein responsible for male sex determination in humans.[20-21] SOX-4 is related to embryonic development and determination of cell fate.[19] IHF protein is a member of bacteria type II binding proteins, which is essential for site-specific recombination, DNA replication, and transcription[33]. NF-Y is involved in cellular response to stimuli such as stress, free radical, ultraviolet irradiation, etc.[23] Fig. 4 shows the initial configuration of all the four systems.

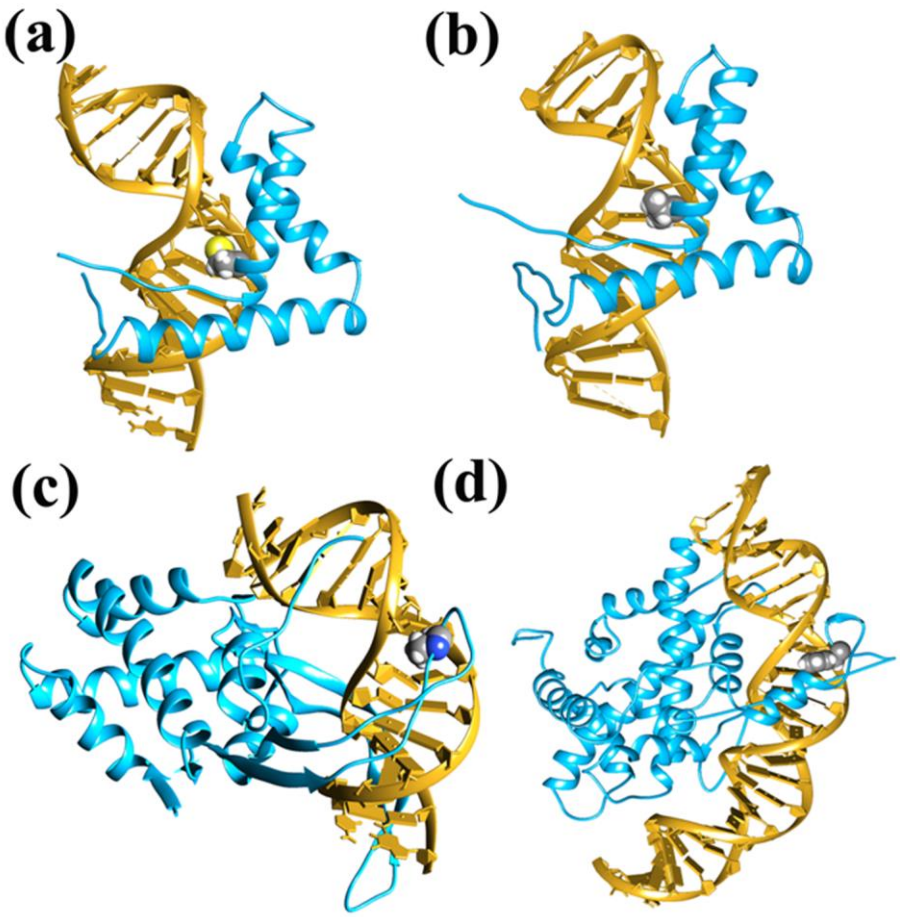

Figure 4: Initial configuration of protein-DNA complexes. (a) SOX4-DNA complex with intercalating amino acid MET (b) SRY-DNA complex with intercalating AA ILE (c) IHF-DNA complex with intercalating AA PRO (d) NF-Y-DNA complex with intercalating AA PHE. Protein and DNA are shown in a ribbon configuration, and the AA is shown using vdW representation.

All the four systems discussed above belong to class 1 as each of them has one AA intercalated into DNA from the minor groove side. However, in two complexes, the majority of the protein is bound in the major groove (IHF and NF-Y, Fig 4c,d) of DNA (but intercalation happens through minor groove only) while the other two proteins (SOX-4 and SRY Fig 4a,b) are bound to the minor groove. Table 2 provides detail on the binding characteristics. Note that, although all the four systems have different amino acid and different intercalating base pairs, all the intercalating amino

acids are hydrophobic in nature and the Rise value of intercalating base pair is in the region of partial intercalation as defined above.

Table 2: Details of the Protein-DNA complexes studied here. The intercalating AA and the position of the intercalating base pairs are mentioned along with the value of Rise and Roll for the intercalating base pair.

Complex	Intercalating AA	Kink Position	Bend Angle	Roll Angle	Rise (Å)	Binding Site on DNA
SOX4-DNA[19]	MET	T8pT9	66°	46°	5.67	Minor Groove
SRY-DNA[20-21]	ILE	A8pA9	50°	43°	4.50	Minor Groove
IHF-DNA[22]	PRO	T29pG30	160°	60°	6.42	Major Groove
NF-Y-DNA[23]	PHE	C9pA10	90°	48°	4.38	Major Groove

To test our hypothesis, we created eight systems corresponding to the above four TF-DNA complexes (see method section for detail) in which only the intercalating AA was restained at the initial crystal structure configuration. So, total we have all eight systems. With these eight systems, we calculated the free energy profile using umbrella sampling[34] for further intercalation or deintercalation of intercalating AA along the reaction coordinate *X* (Fig. 5). Details of the methods involved are given in next section.

5.4 Systems and Method

(I)Preparation of Initial Configurations. The crystal structures of IHF-DNA (PDB ID: 2IIE)[35], SRY-DNA (PDB ID: 1J46) [38], SOX4-DNA (PDB ID: 3U2B)[19] and NF-Y-DNA (PDB ID: 4AWL)[23] complexes were taken from the Protein Data Bank (PDB). Two types of initial configurations were prepared: (i) DNA-AA system, where all amino acids except the partially intercalated AA were removed and (ii) DNA-protein system. For DNA-AA systems, intercalating AA was made after capping the N-terminal and C-terminal with acetyl and N-methyl groups,

respectively using AMBERTools software[36]. The crystal configuration of the intercalating AA was retained. In all cases, amber99[37] forcefield with parmbsc0[38] modification was used for DNA and amberff99SB force-filed was used for proteins and amino acids[39]. All the simulations were carried out using GROMACS[40] molecular dynamics program. For free energy calculations, we used the PLUMED[41] patch to GROMACS. The topology and coordinates generated from AMBERTools were converted to GROMACS format by amb2gmx.pl program[42]. Analyses of DNA structures were carried out using curves+[43].

(II)Simulation Details. In all four systems, the prepared complexes were put in large cubic boxes keeping the surface of the DNA-AA/protein complex 10 Å away from the box edges. TIP3P water model[44] was used to solvate the systems, and ions were added to neutralize the charges on DNA and protein/AA. The energy minimization was done on all systems using steepest descent algorithm, followed by heating up to 300 K using the Berendsen thermostat[45] with a coupling constant of 0.2 ps. During the heating process, the heavy atoms of DNA and Protein/AA were restrained with the harmonic potential of 25-50 kcal/molÅ^{-2} depending upon the complexes. Further, a series of equilibration and energy minimization were carried out for all the systems, where the restraint was reduced. The force constant for position restraint was selected to ensure that the partially intercalated AA retains its crystal configuration. Finally, all the systems were equilibrated for 1 ns restraining the heavy atom at constant temperature 300 K and constant pressure 1 bar using the Nose-Hoover thermostat[46] and Parrinello-Rahman barostat[47], respectively, with coupling constant 0.2 ps. For electrostatics, Particle Mesh Ewald (PME) [48] method was used with a long-range cutoff of 10 Å and van der Waals was treated with 10 Å cutoff. Integration time step for all simulations was kept at 2 fs.

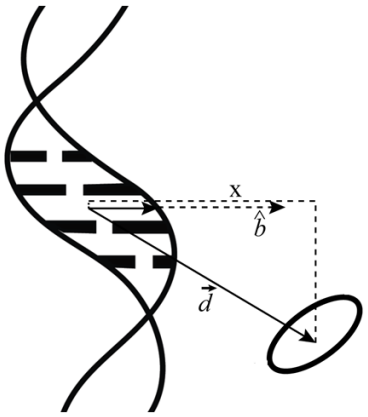

Figure 5. Reaction coordinate X along which free energy is calculated. $X = \hat{b}.\vec{d}$, where $\vec{d}$ is the vector from the center of mass (COM) of four bases involved in intercalation to the COM of the intercalating AA. $\hat{b}$ is constructed from COM of four bases to the COM of two sugar groups lying in 3' direction of the intercalating nucleotides. Therefore, X denotes the displacement of the AA perpendicular to the DNA. It is positive towards the minor groove. Note that, displacement along the DNA will not change X.

(III)Definition of Reaction Coordinates Used in Umbrella Sampling. Umbrella sampling simulation requires a reaction coordinate. As with the small molecule study[25, 30], we used the distance of the intercalating AA along the minor groove direction as a reaction coordinate,X, for intercalation and deintercalation of the amino acid. $X = \hat{b}.\vec{d}$, where $\hat{b}$ is a unit vector starting from the center of mass (COM) of four bases of intercalating base pair (T8pT9 in SOX-4, A8pA9 in SRY, T8pT9 in IHF, and C9pA10 in NF-Y) to the COM of two 3' sugars groups attached with these base pairs. $\vec{d}$ represents the vector from COM of intercalating base pairs to COM of side the chain of AA. Fig 5 shows the schematic representation of the reaction coordinate, X. The reaction coordinate is chosen such that it is positive in the minor groove side region and negative in major groove side region.

We divided the reaction coordinate, X into 25 to 30 windows and for each we performed simulations for 4 ns to get the complete free energy profile. Force constant was varied between 24

kcal/mol to 12 kcal/mol. Finally, the free energy for each of the process was calculated using Weighted Histogram Analysis Method (WHAM)[49].

(IV)Metadynamics Details for DNA Bending study. Well tempered metadynamics method[50] was employed to study the bending of the DNA. In this case, reaction coordinate was taken to be end-to-end distance calculated from the COM of the C1' atoms of T2C3 to the same of T14C15. Simulation was continued for 5 ns using 0.7 Å width and 0.2 kJ/mol height for Gaussian potential. The deposition rate was taken to be 2 ps and the bias factor of 10 was used. All the other details are same as used in umbrella sampling simulations mentioned above. Simulation was extended to 10 ns to check for the convergence. Results are plotted using the first 5ns trajectory.

(V)Umbrella Sampling Details for DNA Bending Study. The reaction coordinate was taken to be the end-to-end distance calculated from the COM of the C1' atoms of T2C3 to the COM of T14C15. We divided the reaction coordinate into 18 windows and for each we performed simulations for 2 ns to get the complete free energy profile. Force constant was varied between 1.2 kcal/mol to 2.1 kcal/mol. Finally, the free energy for each of the process was calculated using Weighted Histogram Analysis Method (WHAM)[49].

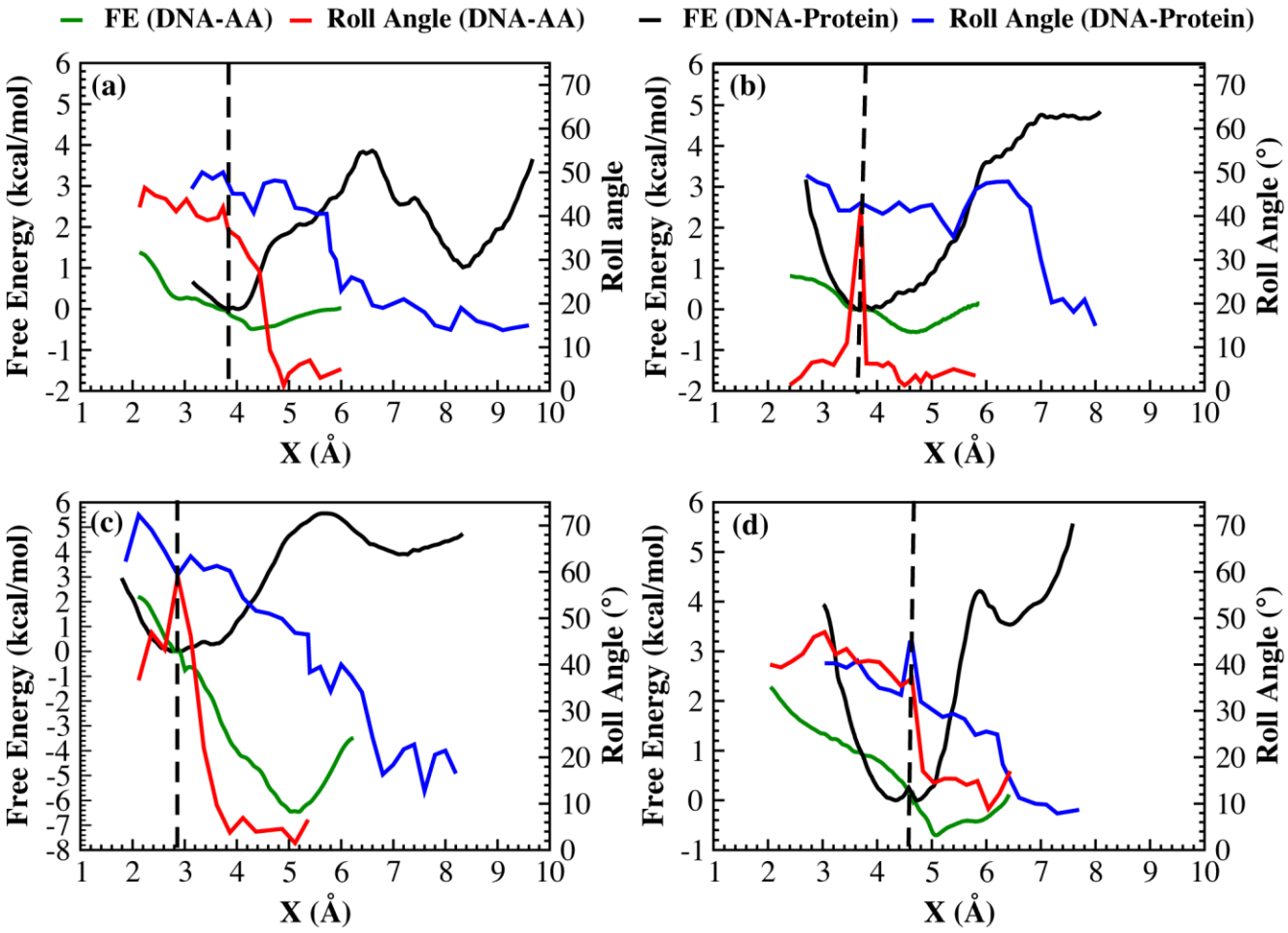

Figure 6. Free energy profile along the reaction coordinate X for DNA-Protein and DNA-AA systems. (a) SOX4 (b) SRY (c) IHF (d) NF-Y. Change in the roll angle of the respective intercalating residue is also shown along X. Dashed vertical line denotes the position of the AA at crystal configuration.

5.5 Results and Discussion

(I) Free Energy Profile of DNA-AA and DNA-Protein Systems

Fig. 6 shows the free energy profile for all the eight systems. The black dashed line in each graph denotes the position of intercalating AA at the initial configuration, i.e., at crystal structure configuration. Therefore, $X < X_{CSC}$ would represents the intercalation while $X > X_{CSC}$ would denotes deintercalation of the intercalating AA.

(II)Intercalation and Deintercalation of AA from DNA-AA system

Fig 6 (green and black color) shows the free energy corresponds to the intercalation and deintercaltion of AA. The deintercalation of the AA in DNA-AA system leads to a lowering of the free energy profile, results in the formation of a more stable state. Moreover, the free energy continues to decrease till the AA goes to the minor groove, beyond which free energy increases again (from there AA start to dissociate). This trend is common for all the four systems (Fig 6). Also, the free energy profile is akin to what was observed for small molecule intercalation process.[25-26, 30] However, unlike small molecule/drug, intercalation of the AA in these DNA-AA systems results in the increase in free energy, due to hindrance from the backbone of AA (here AA is capped).

(III)Intercalation and Deintercalation of AA from DNA-Protein system

Unlike to DNA-AA system, the DNA-protein system, shows the increase in free energy upon intercalation and deintercalation of AA, indicating expectedly that the crystal structure is the most stable configuration (Fig 6 black color). However, this stability does not come from partial intercalation. Otherwise, DNA-AA free energy would have reflected the same. Since we deintercalate only the AA in DNA-Protein system, most of the DNA-Protein interactions are unchanged. Therefore, comparing the protein-DNA and AA-DNA systems, we deduce that the decreased stability due to partial intercalation must be compensated by other factors such as entropy as was found in the case of bent protein-DNA complexes.[51] Except for SRY, the deintercalations in DNA-protein system go through a small barrier to reach a local minimum. We have not studied the complete dissociation of the protein as the objective was to see the effect of the intercalating residue (The complete dissociation of protein is studied in next chapter). Fig. 7 shows the initial and final configurations after further intercalation and deintercalation of

intercalating AA for all the eight systems. Complete SRY-DNA interaction was studied by Lavery and coworkers who showed a stability of 11.5 kcal/mol[52] having two different association pathways[53]

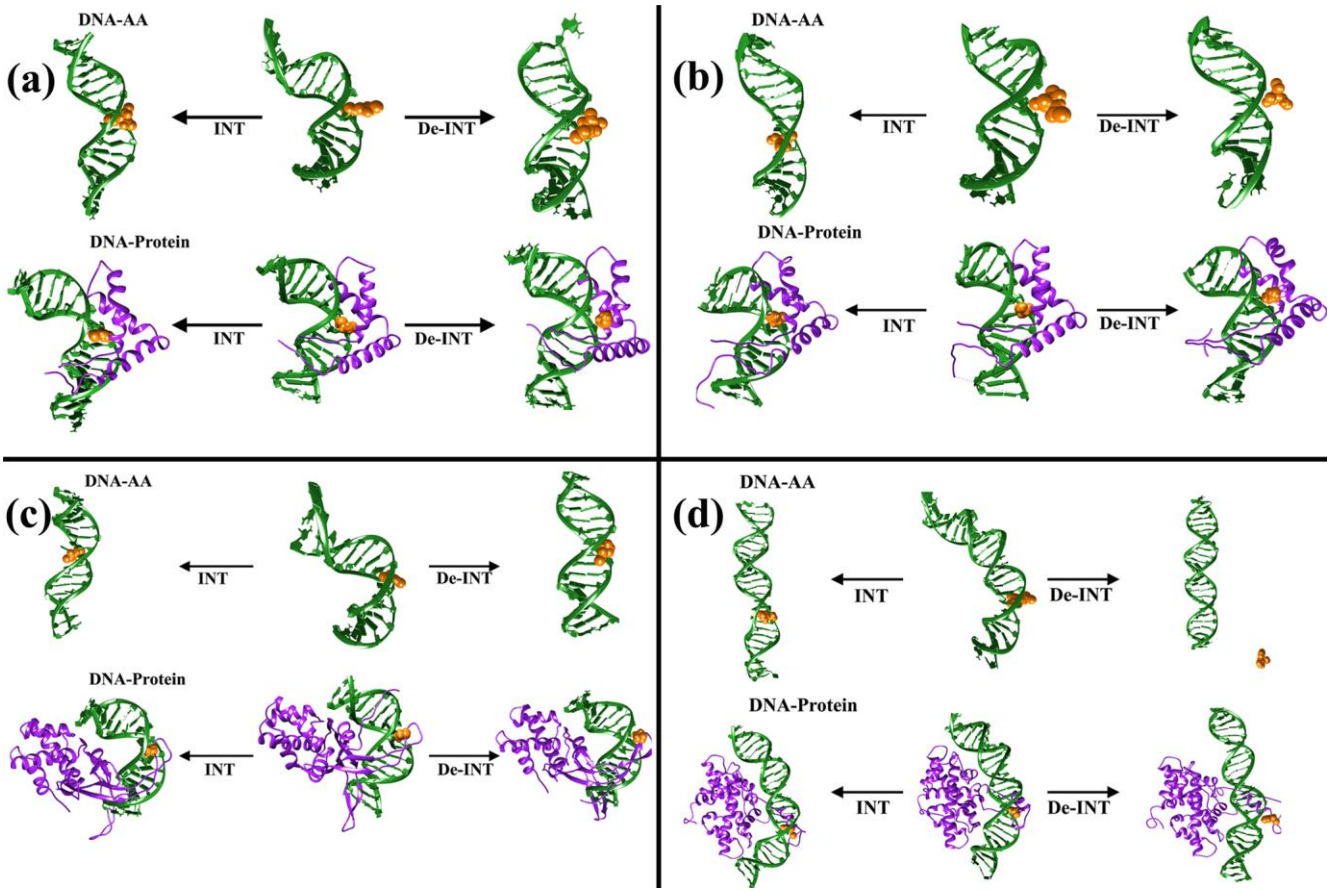

Figure 7. Initial and final configurations of all the eight systems: (a) SOX4-DNA and MET-DNA, (b) SRY-DNA and ILE-DNA, (c) IHF-DNA and PRO-DNA, (d) NF-Y-DNA and PHE-DNA. DNA (green) and protein (violet) are shown using ribbon representation and the intercalating AA are shown in the solid sphere representation.

(IV)Convergence of Free Energy

We have checked that the performed umbrella sampling for intercalation and de-intercalation of AA is converged or not. Fig. 8 which shows that all the free energy is converged within 1kcal/mol.

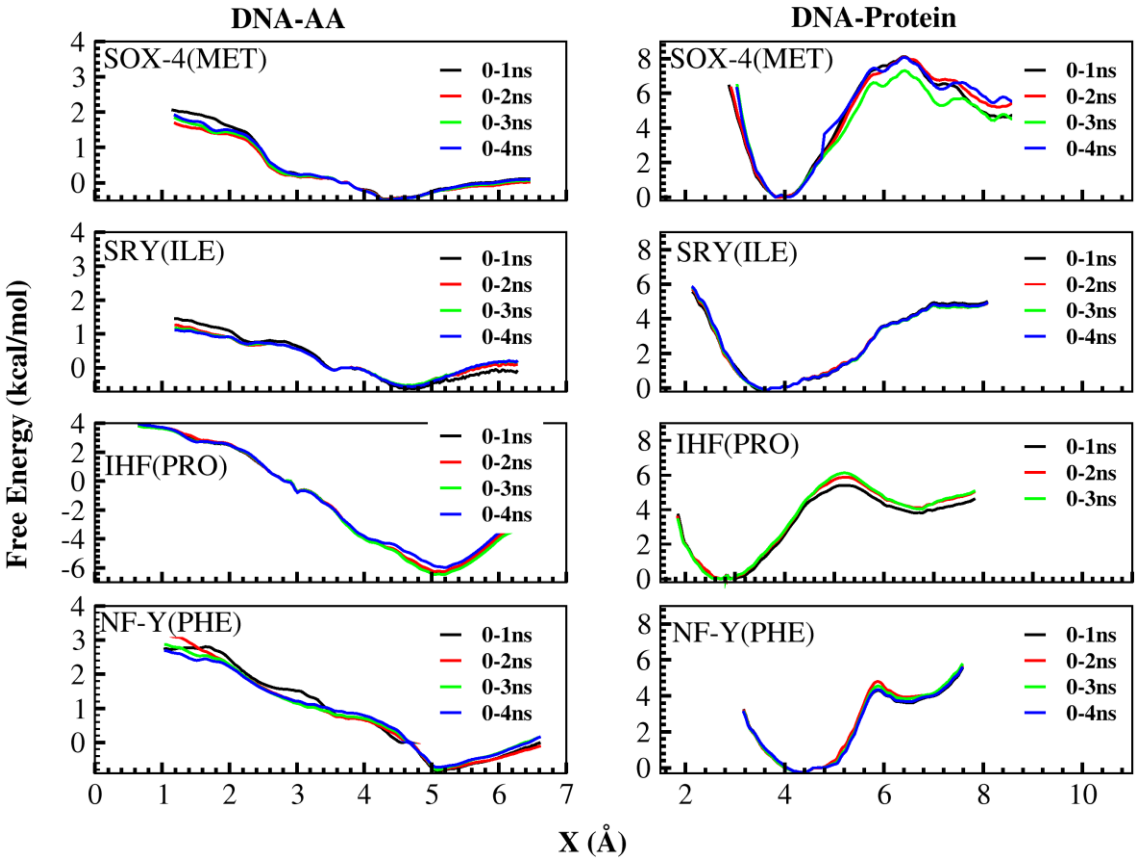

Figure 8. Free energy convergence. Left panel represents the free energy of DNA-AA systems, and right panel shows free energy profiles of DNA-Protein system.

Further, we have also calculated the error in free energy with two different methods (i) Bootstrap (Fig. 9) and (ii) Block average (Not Shown here). The dotted line in each free energy shows the error in free energy.

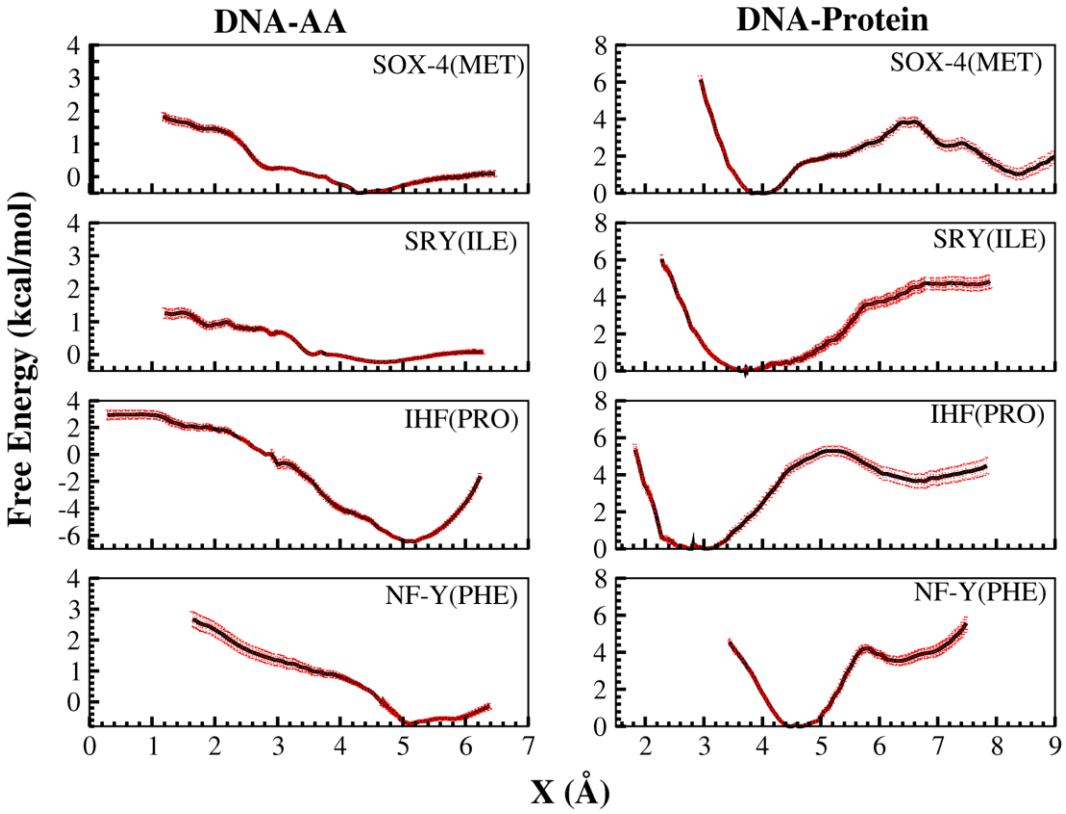

Figure 9. Error in free energy calculation (bootstrap method). Black line represents the free energy, and the red dots represent errors associated with free energy. Left panel represents errors in free energy of DNA-AA systems, and right panel shows error in free energy profiles of DNA-protein system.

(V)Change in Roll angle

Till now we have discussed how the free energy change upon intercalation and deintercalation of AA in DNA-protein/AA system. Next we have study how the intercalation and deintercalation of AA would affect the DNA topology, basically kinking. To do so we plot the average Roll angle along *X* in Fig. 6 These averages are calculated from a 4 ns free simulation with *X* constrained at a particular value. At crystal structure configuration (dashed line in Fig. 6), both the DNA-AA and DNA-protein systems have approximately the same average Roll angle.

This shows that AA alone is sufficient to inflict the kink. Deintercalation of IAA leads to a decrease of Roll to normal B-DNA values in both systems. At large separation of IAA from the

IBP, Roll angle is normal for all the eight systems, while DNA remains bent for DNA-proteins systems (see late discussion for DNA bending).

We have also extended the simulation for one DNA-AA system (SOX-4) to 100ns to check the convergence of simulation.

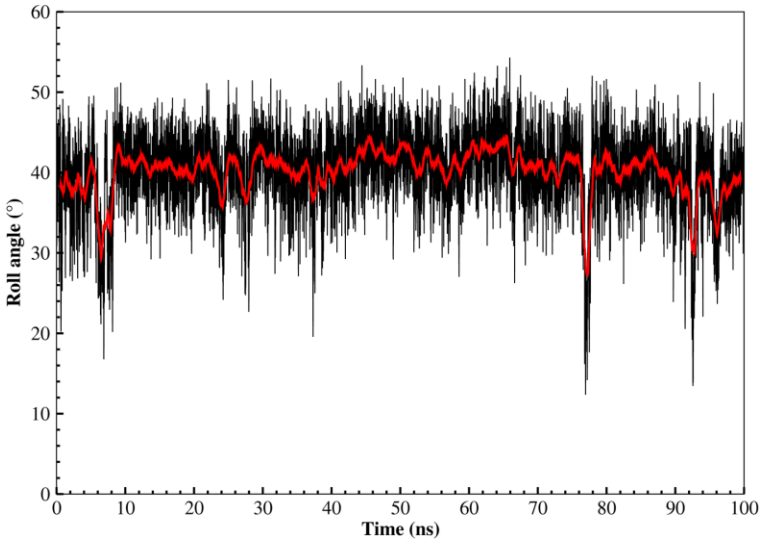

Figure 10. Variation of roll angle with time for DNA-AA system (SOX4-DNA). Red color represents the running average. Average value is 40° (±4.7), which is within the standard deviation from the average value 42° obtained from 4ns simulation.

Fig. 10 shows that average Roll angle does not change appreciably with the longer simulation. Therofore suggest that the 4ns sampling is enough. Further, we have mutated the intercalating amino acid (AA) with glycine and monitored the change in Roll angle by simulated annealing (to supply enough energy to overcome the barrier). Glycine (GLY) is small AA and does not intercalate, so we believed that mutating intercalating AA with GLY would remove the kink in DNA. Indeed we have observed the same, Fig 11 shows that Roll angle decreased to normal B-DNA value.

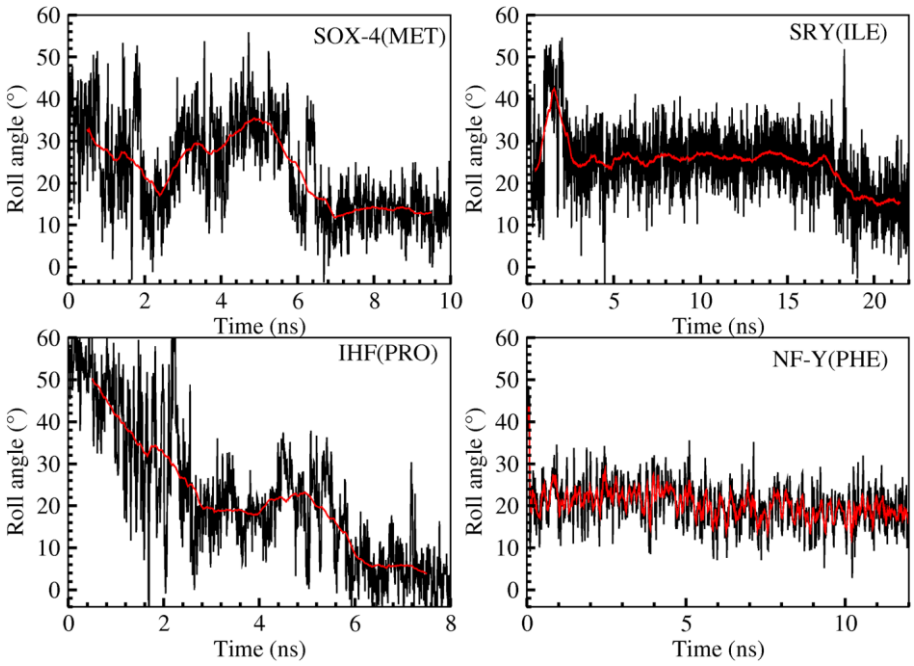

Figure 11. Variation of roll angle with a time of all four systems after mutating the intercalated original intercalated residue with glycine). We performed simulated annealing by heating the system to 500 K through 1 ns, equilibrated at 500 K for 2 ns, and then cooled down the system to 300 K in 1 ns. We further equilibrated the system at 300 K for a longer period. Change in the Roll angle over the entire period (including heating and cooling parts) are shown to show how Roll angle varied in the entire period of simulated annealing.

Since we have shown in the previous section that Rise measure the extent of intercalation, therefore Rise decreases monotonically (Blue dot line in Fig. 12) with X, and the Rise and Roll together shows the inverted parabolic behavior (blue dot line in Fig. 13) as shown earlier (Fig 1), supports that partial intercalation is the cause for kink formation.

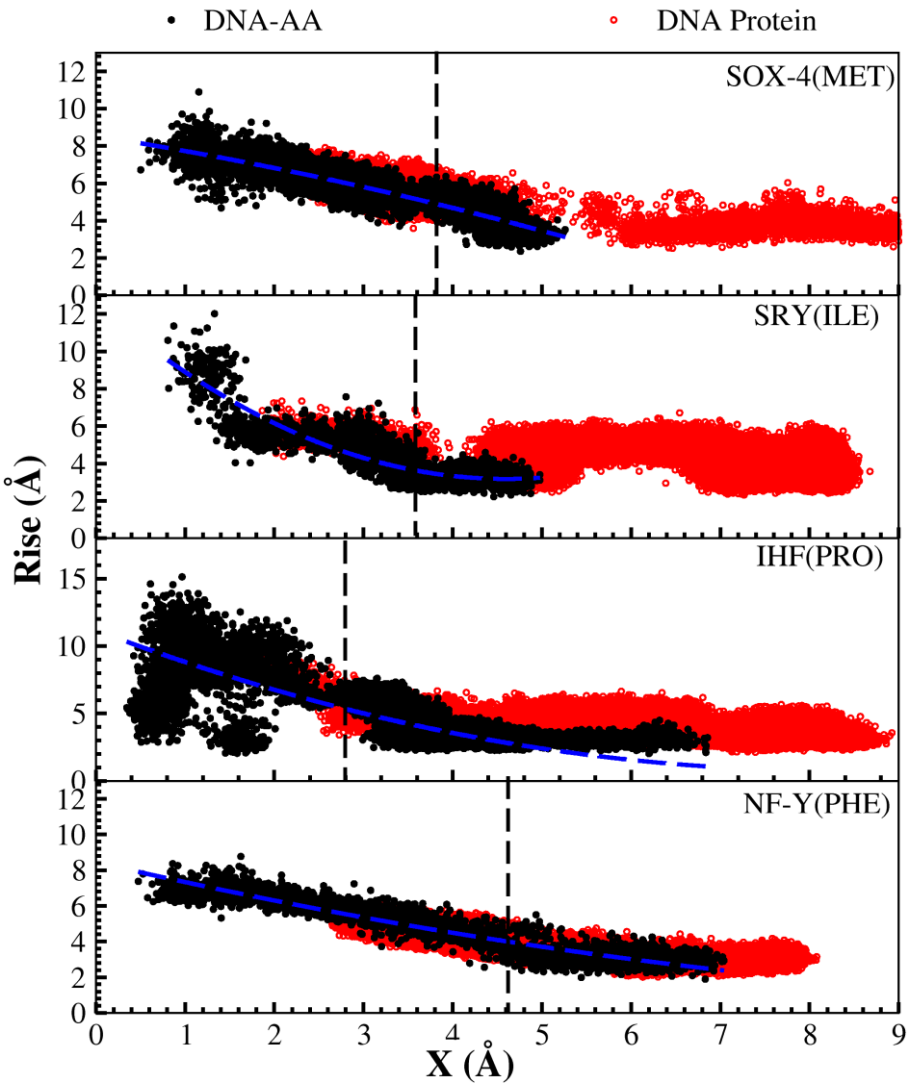

Figure 12. Variation of Rise of the intercalating base pair with X for all eight systems. Red and black dots represent the values for DNA-Protein and DNA-AA systems, respectively. Note the monotonic increase of Rise with X beyond a certain small value of X.

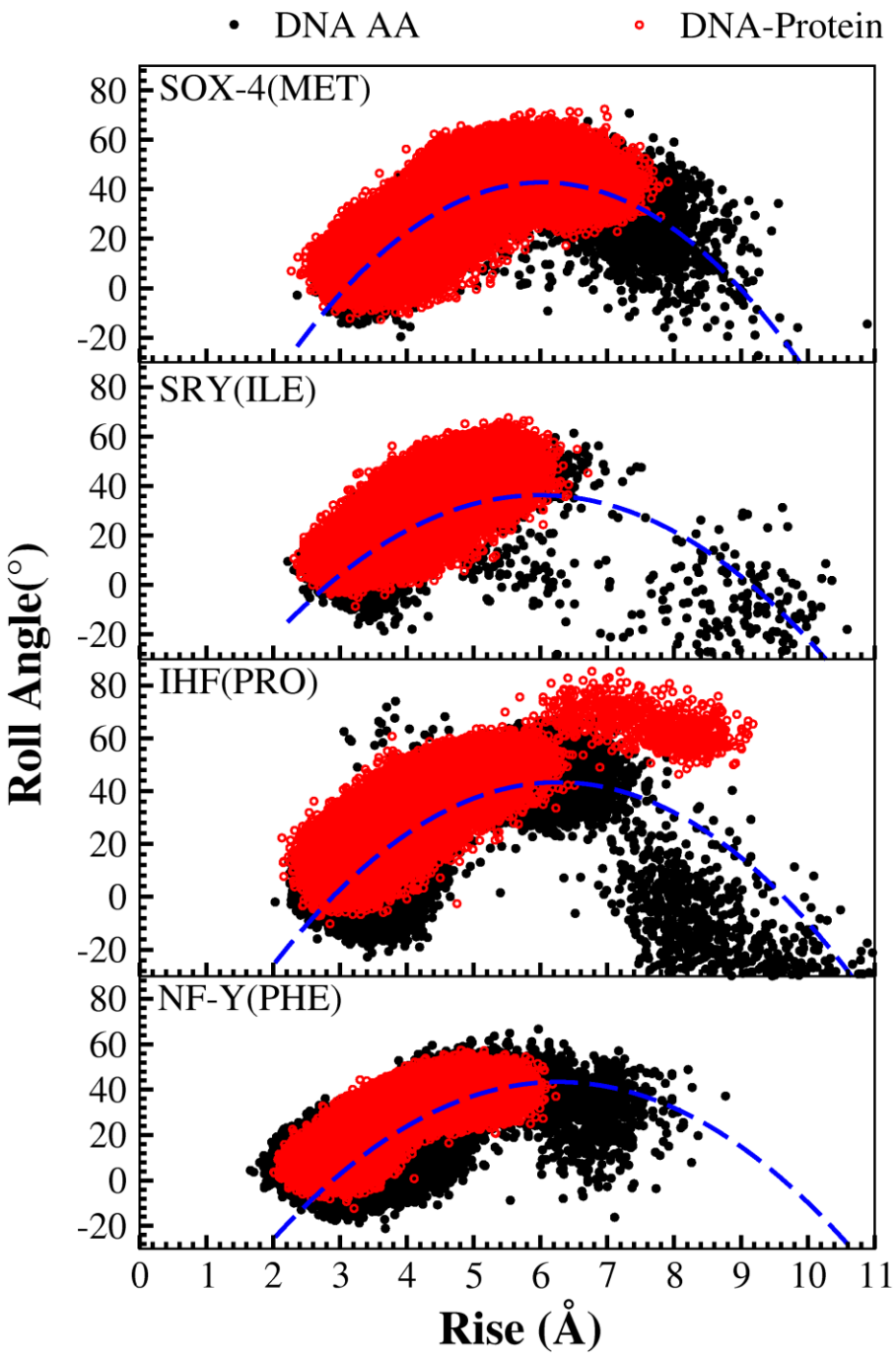

Figure 13. Variation of Roll angle with Rise of all eight systems. Red and black dots represent the results for DNA-Protein and DNA-AA systems. Note the inverted parabolic behaviour.

(VI) AA-DNA simulation (without capping): SOX4-DNA complex (Intercalating AA MET)

As shown in Fig 1(a) the complete intercalation of AA into the intercalating base pair does not increase the Roll angle as observed for a small molecule (Fig. 1a), here in DNA-AA system the backbone of AA (AA is capped) may be hindered complete intercalation and cause increase in roll angle. To reinstate this, we have performed a set of free energy simulations of DNA-AA system (using SOX4-DNA complex), where the backbone was removed from the AA. Figure 14 shows the free energy and Roll angle against the reaction coordinate. Roll angle behaves in similar manner as shown in Fig. 1 where an increase due to partial intercalation followed by a decrease upon complete intercalation is observed.

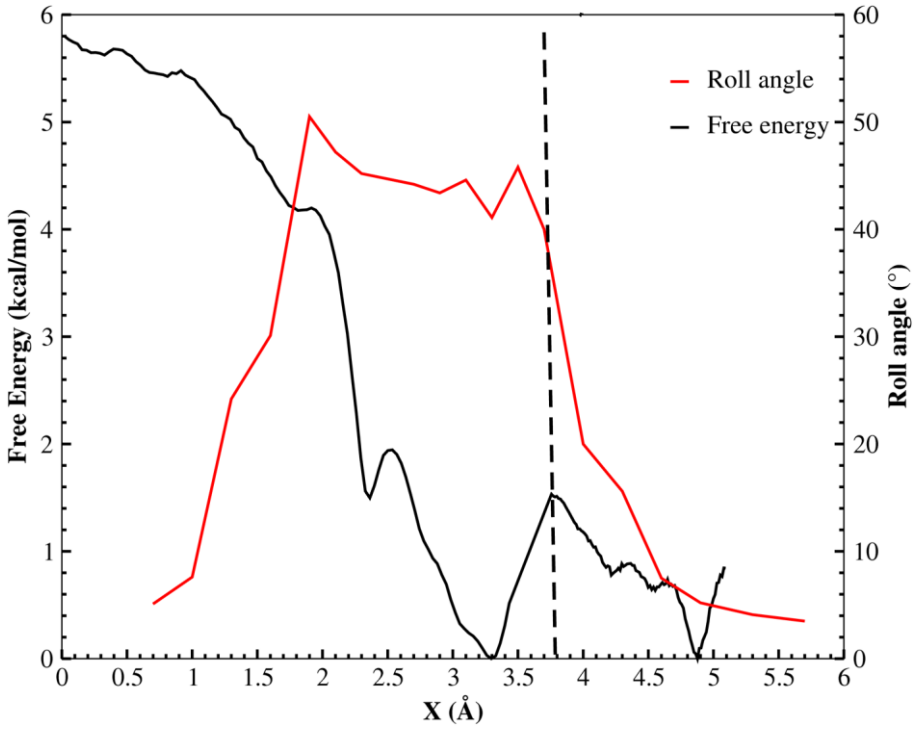

Figure 14. Change in free energy and roll angel with X. Black color represents the free energy with respect X and red color represents the Roll angle with respect to X.

Fig. 15 shows the configurations as a result of full intercalation and deintercalation of the uncapped AA. Free energy, however, did not decrease on full intercalation as observed for small

molecule. This may be due to the fact that the AA is a long chain molecule (MET) and it disfavors intercalation.

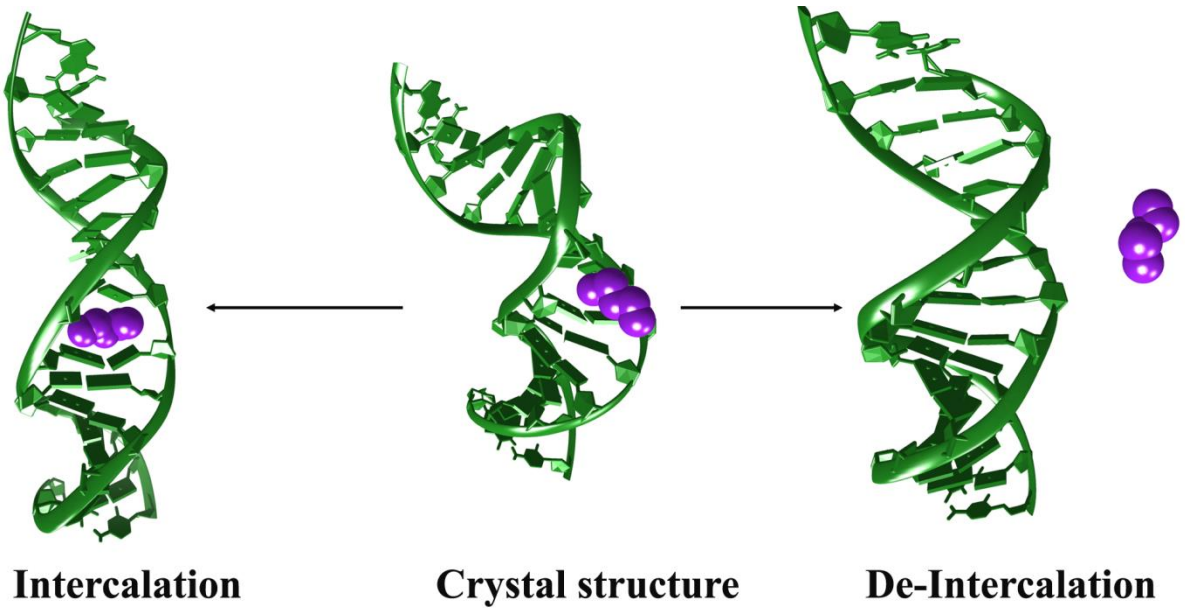

Figure 15. Change in DNA configuration during intercalation and deintercalation of AA (without capping). DNA structures are shown in green color, and the intercalating AA are shown in sphere model.

(VII)Change in Bend Angel

We did not observed a change in bending of DNA upon deintercalation and intercalation of AA. The DNA kinking is a local change (happened at BP level), therefore we thought that the intercalation and deintercalation would be affecting the nearby BP of intercalation BP. Therefore, we have calculated the changes in the local axis-bend angle (±4 base pairs from intercalating base pairs) of DNA along *X* are (shown in Fig. 16). This shows that the for DNA-AA system, bending decreases in all cases. However, a decrease in bending for protein-DNA complexes are much less, as is also evident from Fig. 7.

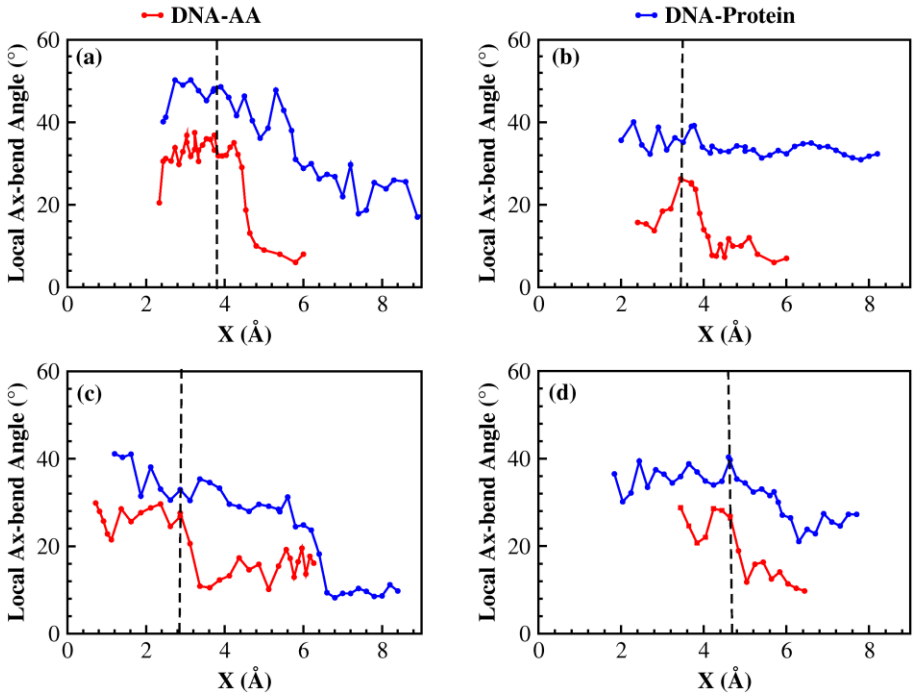

Figure 16. Local bending angle (±4 base pairs from IBP) of all eight systems. Red color represents the DNA-AA system, and blue color represents the DNA-Protein system. (a) SOX4-DNA and MET-DNA (b) SRY-DNA and ILE-DNA (c) IHF-DNA and PRO-DNA (d) NF-Y-DNA and PHE-DNA complex with intercalating AA PHE. The axis bend angle is calculated using Curves+.[1]

(VIII)Change in DNA Parameter for all DNA BP

Since DNA bending is the consequence of change in roll and twist angles, we have calculated the Roll and Twist angles of all base pairs (Table 3 and 4), which show that the change in Roll and Twist angle of all other base pair are responsible for preserving the bend of unkinked DNA in DNA-protein system.

Table 3: Roll angle of all base pairs for all four systems (DNA-protein complex only).

(a) SOX4-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Roll angle range (°))	Crystal configuration (Roll angle (°))	De-intercalation (Large X) (Roll angle range (°))
BP-1	2±1	6	3±1
BP-2	0±1	1	1±1
BP3	6±1	7	2±1

BP-4	4±3	3	2±1
BP5	7±2	3	4±2
BP-6	23±3	9	24±5
BP-7	10±2	5	9±2
BP-8	**50±2**	**46**	**10±2**
BP-9	11±1	9	9±2
BP-10	3±4	4	8±1
BP-11	3±3	2	3±2
BP-12	0±3	2	2±1
BP-13	2±3	4	0±1
BP-14	8±3	16	9±2
BP15	4±2	4	7±1

(b) SRY-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Roll angle range (°))	Crystal configuration (Roll angle (°))	De-intercalation (Large X) (Roll angle range (°))
BP-1	6±1	6	1±2
BP-2	1±1	9	10±2
BP3	7±2	10	1±2
BP-4	1±3	-3	1±3
BP5	8±1	1	4±1
BP-6	9±2	6	6±2
BP-7	10±3	5	6±2
BP-8	**31±3**	**43**	**10±1**
BP-9	12±1	10	23±5
BP-10	16±2	17	16±5
BP-11	4±1	5	4±1
BP-12	0±1	-3	1±4
BP-13	5±1	1	2±4

(c) IHF-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Roll angle range (°))	Crystal configuration (Roll angle (°))	De-intercalation (Large X) (Roll angle range (°))
BP-1	7±2	5.2	2±1
BP-2	1±2	4	2±1
BP3	0±4	-5	0±1
BP-4	13±5	13	21±3

BP-5	5±2	7	16±2
BP-6	7±1	8	10±2
BP-7	8±1	2	7±2
BP-8	**59±3**	**60**	**10±2**
BP-9	12±2	14	24±4
BP-10	3±2	6	16±7
BP-11	1±1	-2.6	1±1
BP-12	1±2	1	5±4
BP-13	3±2	8	3±2
BP-14	8±5	13	2±1
BP15	-3±1	-4	2±4

(d) NF-Y-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Roll angle range (°))	Crystal configuration (Roll angle (°))	De-intercalation (Large X) (Roll angle range (°))
BP-1	2±3	5	2±2
BP-2	1±3	-1	1±2
BP-3	1±4	0	0±1
BP-4	6±5	10	6±2
BP-5	1±2	-4	0±1
BP-6	1±2	6	0±1
BP-7	0±1	3	-5±2
BP-8	14±2	14	20±2
BP-9	**34±5**	**48**	**7±2**
BP-10	4±2	9	7±1
BP-11	4±2	10	7±2
BP-12	-5±1	-7	-4±1
BP-13	1±1	-6	0±1
BP-14	4±2	2	3±1
BP-15	2±4	-6	-5±1
BP16	-5±3	-1	-7±2
BP-17	6±3	11	3±1
BP-18	6±2	6	9±1
BP-19	5±2	5	0±1
BP-20	0±1	2	0±3
BP-21	3±2	2	3±1
BP-22	1±2	4	4±1
BP-23	3±1	3	8±2
BP-24	6±4	11	4±2

Table 4: Twist angle of all base pairs for all four systems.

(a) SOX4-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Twist angle range (°))	Crystal configuration (Twist angle (°))	De-intercalation (Large X) (Twist angle range (°))
BP-1	29±1	26	24±1
BP-2	38±2	38	35±1
BP3	29±4	37	30±1
BP-4	31±3	35	32±1
BP-5	36±2	41	40±1
BP-6	8±2	18	13±1
BP-7	19±1	22	24±1
BP-8	**21±1**	**22**	**26±1**
BP-9	39±2	32	35±3
BP-10	27±2	31	29±1
BP-11	36±1	37	36±2
BP-12	29±3	33	35±3
BP-13	37±1	41	39±1
BP-14	27±5	27	22±1
BP15	34±2	37	38±2

(b) SRY-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Twist angle range (°))	Crystal configuration (Twist angle (°))	De-intercalation (Twist angle range (°))
BP-1	32±1	36	43±
BP-2	3±1	4	-15±1
BP3	46±2	45	43±2
BP-4	29±2	38	27±1
BP5	35±1	36	35±1
BP-6	24±1	28	27±2
BP-7	39±1	31	35±2
BP-8	**19±2**	**22**	**21±1**
BP-9	22±	20	28±1
BP-10	16±1	16	14±2
BP-11	41±2	41	40±1
BP-12	31±1	31	30±1
BP-13	28±3	27	3±2

(c) IHF-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Twist angle range (°))	Crystal configuration (Twist angle (°))	De-intercalation (Large X) (Twist angle range (°))
BP-1	43±2	43	27±5
BP-2	35±2	20	35±4
BP3	35±2	36	35±4
BP-4	28±2	29	30±2
BP-5	35±4	31	30±2
BP-6	36±1	32	32±2
BP-7	20±1	23	27±1
BP-8	**28±3**	**22**	**25±2**
BP-9	25±2	15	33±4
BP-10	36±2	39	28±2
BP-11	27±1	28	20±2
BP-12	44±2	35	43±1
BP-13	35±2	29	22±4
BP-14	42±2	21	36±2
BP15	31±2	35	26±3

(d) NF-Y-DNA complex (Intercalating base pair are shown in bold font)

Base pair's position	Intercalation (Small X) (Twist angle range (°))	Crystal configuration (Twist angle (°))	De-intercalation (Large X) (Twist angle range (°))
BP-1	32±2	31	28±1
BP-2	33±1	43	35±3
BP-3	32±1	29	29±2
BP-4	31±2	28	31±1
BP-5	42±1	43	38±1
BP-6	33±1	34	33±1
BP-7	32±1	35	37±1
BP-8	32±1	33	28±1
BP-9	**18±2**	**18**	**30±2**
BP-10	36±3	39	38±2
BP-11	30 ±1	32	29±1
BP-12	41±1	43	41±1
BP-13	40 ±1	42	40±1
BP-14	31 ±1	32	29±2
BP-15	37±1	36	38±4
BP16	42±2	45	40±2
BP-17	32±1	27	18±2

BP-18	33±3	37	37±2
BP-19	37±1	35	35±2
BP-20	32±1	36	31±1
BP-21	35±1	30	30 ±1
BP-22	34±2	40	32±4
BP-23	34±2	34	28±1
BP-24	30±1	30	30±3

5.6 Kink through Bends.

(II) Metadynamic and umbrella sampling simulation on DNA (SOX-4-DNA Complex DNA sequence): We have seen that partial intercalation of an AA can cause kinks in the DNA. However, proteins may also bend the DNA and produce kink as it can also form through strong bending[54]. However, simulations showed that bending induces type II kink where two or three base pairs are affected.[55] To understand whether bending of DNA by proteins would cause the kink, we took the SOX4-DNA complex, removed the protein and calculated free energy of DNA bending using the end-to-end distance between COMs of terminal base pairs (dl) as the reaction coordinate (see definition of the reaction coordinate). When the DNA is straight, but not stretched, dl is equal to its contour length dc, i.e., dl/dc ~ 1.0. The bent DNA will have a lower value of dl/dc ratio.

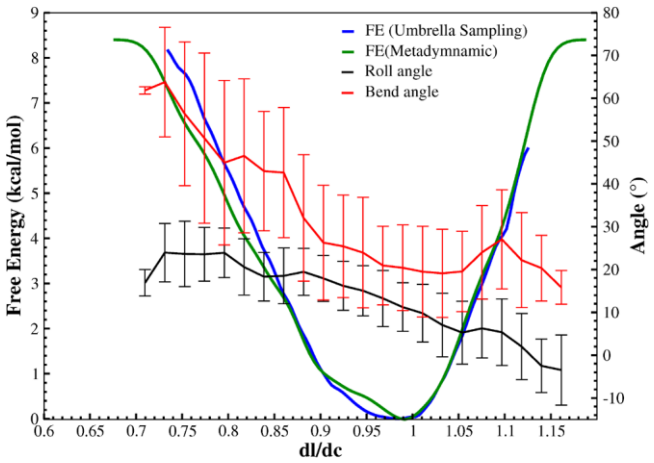

Figure 17. Free energy profile for the bending of DNA using both metadynamics and umbrella sampling method. The average values of Roll angle of T6A7 and ax-bend angle of the DNA calculated from the metadynamics trajectory are also shown, along with the error bars.

We have performed both umbrella sampling and metadynamics simulation to investigate the effect of DNA bending in kink formation in bare DNA (see simulation detail for more information). Fig. 17 shows the free energy profile against dl/dc obtained from both the different approaches. Convergence of free energy profile is shown in Fig. 18 and the errors in umbrella sampling (calculated by two different methods, bootstrap and block average (not shown here)) are shown in Fig. 19. Free energy is minimum when dl ~ dc, i.e., for straight DNA. Free energy increases quadratically with bending, in agreement with the earlier study.[55]

(II) Convergence and error in free energy

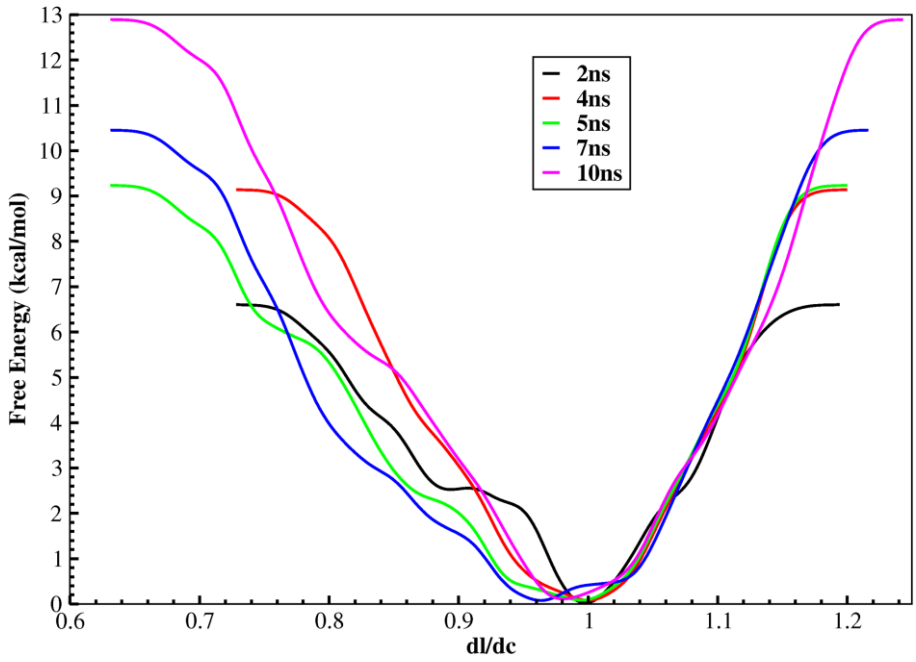

Figure 18. Showing the convergence of the free energy profile. Note that 5ns and 7ns are very similar. With more simulation, the range of collective variable increases as observed for dl/dc>1.0, i.e. for DNA stretching. Once the range is saturated, as observed for dl/dc<1.0 where DNA is bent and bending has its limit, free energy is overestimated (compare the free energy at 0.65 dl/dc for 5ns, 7ns and 10ns sampling). The dl/dc > 1.0 is DNA stretching and therefore, free energy obtained from 5ns, 7ns and 10ns overlap on each other. With more sampling the range is

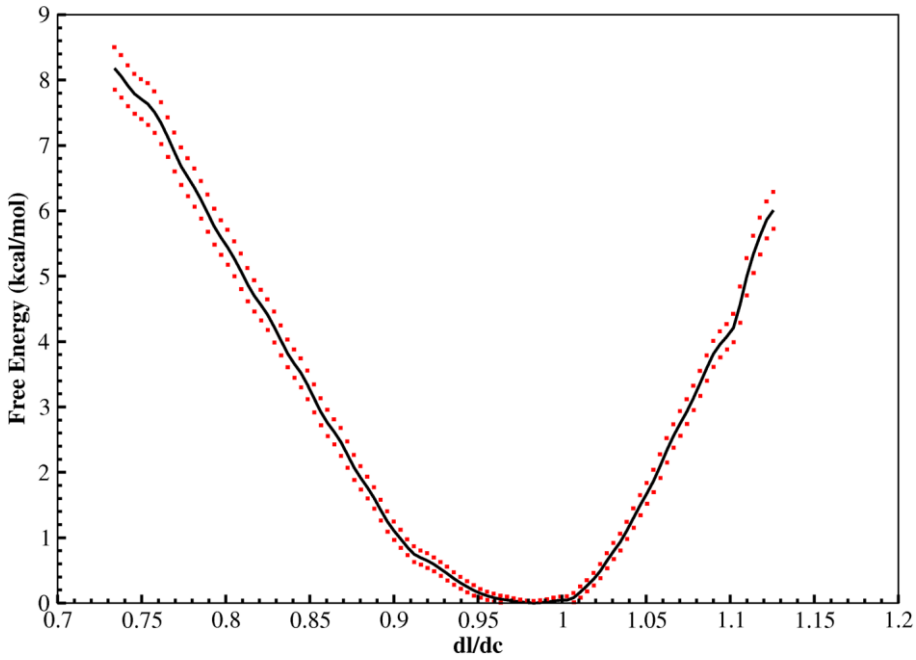

Figure 19. Error in free energy calculation. Black color represents the free energy, and red dots represents the error in free energy.

Fig. 20 shows the variation of Roll angle (of all base pairs) and Ax-bend angle with time. We found that T6A7 showed highest Roll angle. Therefore, we plotted the average Roll angle for T6A7 and global axis bending angle of the DNA along the dl/dc. Fig. 16 indicates that a 25% deviation from the contour length of DNA corresponds to 60° bending angle at the cost of 6 kcal/mol free energy. However, Roll angle is ~24° even at such a high free energy range. The correlation between Roll angle and Ax-bend also indicate a small change in Roll against a significant change in bend angle (Fig. 21). The maximum value of Roll angle observed is ~45°, occurring transiently (Fig. 22). This result indicates that bending of DNA alone does not produce the kink in the DNA.

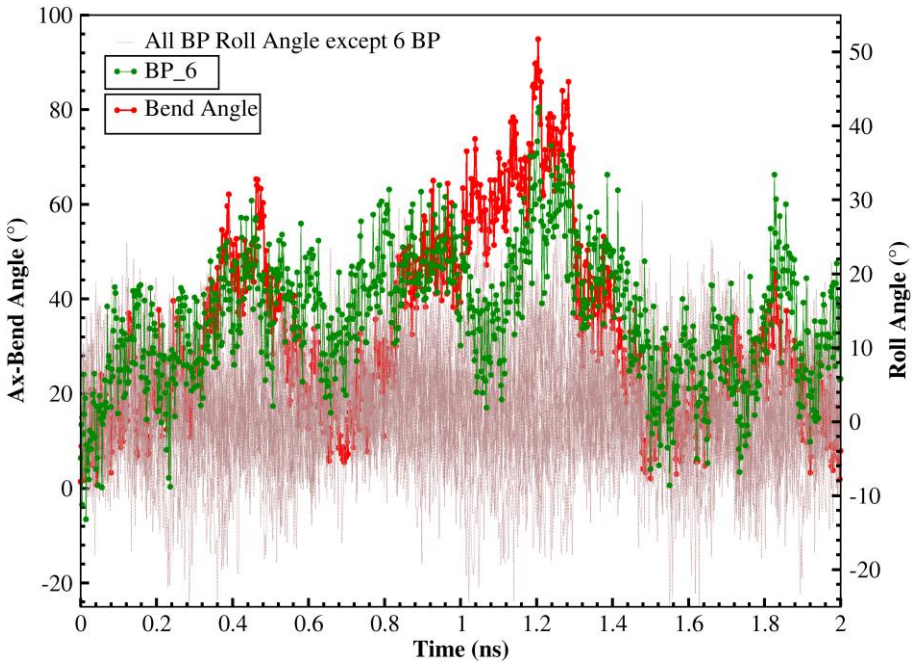

Figure 20. Roll angle for all the base pairs against time in metadynamics simulation. Note that only the T6A7 base pair step has high Roll angle. Corresponding ax-bend angle for the T6A7 base pair step is also shown.

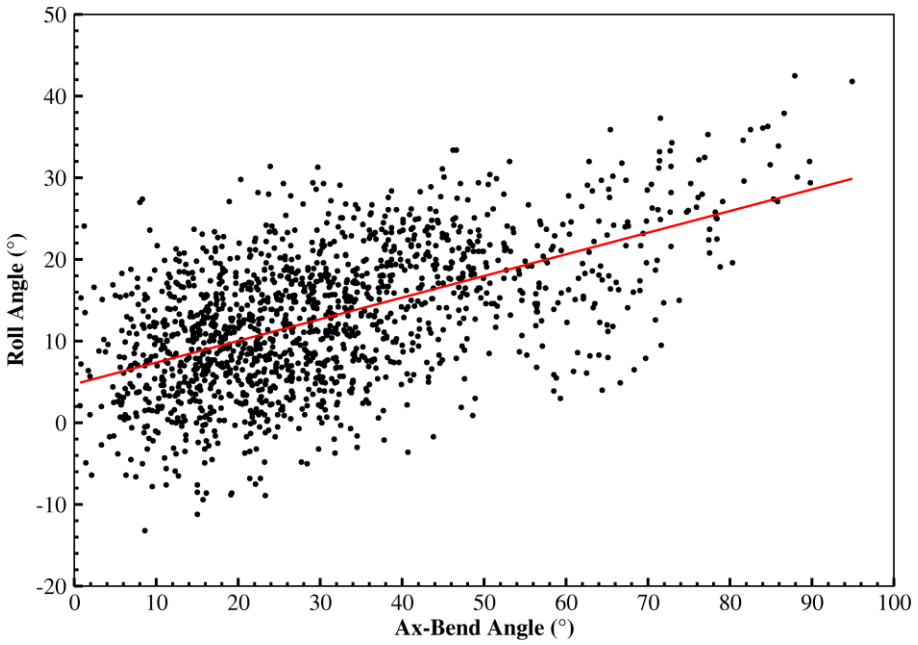

Figure 21. Correlation between Roll angle and ax-bend angle for the T6A7 base pair step.

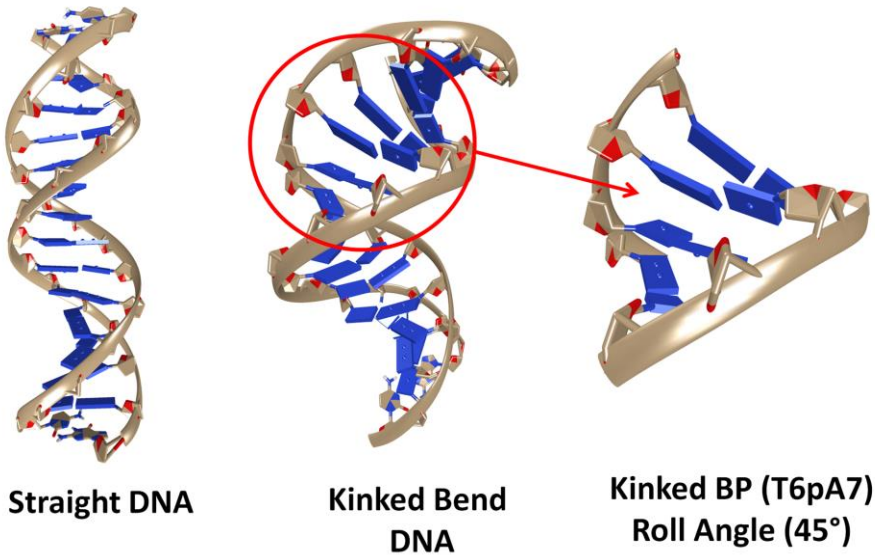

Figure 22. DNA Structure corresponding to highest Roll angle (45°) is shown.

5.7 Conclusion

In summary, we have shown using a correlation with small molecule study, crystal database analysis, and free energy calculation involving four TF-DNA complexes that partial intercalation, not complete, is responsible for kink formation both in small molecule and TF-DNA complexes. The kink observed in TF-DNA is due to the intercalation of AA alone. However, whole protein is required to make the kinked state stable. We also show that the intercalation of AA does not stabilize the kink as is commonly believed. This study leads to further questions as to why nature chooses different intercalating AA for different proteins. While our study shows that protein stabilizes the kinked DNA, the reason for the stabilization is still not clear. Partial intercalation alone does not stabilize the system (DNA-AA study). Also, deintercalation of AA in DNA-protein system leads to destabilization. Therefore, further work is required to explore the exact origin of stabilization of the kinked DNA-TF complex.

Reference

1. Hahn, S., Structure and mechanism of the rna polymerase ii transcription machinery. *Nat. Struct. Mol. Biol.* **2004,** *11*, 394-403.
2. Pérez-Martín, J.; de Lorenzo, V., Clues and consequences of DNA bending in transcription.*Annu. Rev. Microbiol.* **1997,** *51*, 593-628.
3. Werner, M. H.; Gronenborn, A. M.; Clore, G. M., Intercalation, DNA kinking, and the control of transcription. *Science* **1996,** *271*, 778-784.
4. Maher, L. J., Mechanisms of DNA bending. *Curr. Opin. Chem. Biol. 1998***,** *2*, 688-694.
5. Strauss, J. K.; Maher, L. J., 3rd, DNA bending by asymmetric phosphate neutralization. *Science* **1994,** *266*, 1829-34.
6. Crick, F. H. C.; Klug, A., Kinky helix. *Nature* **1975,** *255*, 530-533.
7. Dickerson, R. E., DNA bending: The prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.* **1998,** *26*, 1906-1926.
8. Lankaš, F.; Lavery, R.; Maddocks, J. H., Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure* **2006,** *14*, 1527-1534.
9. Vologodskii, A.; D. Frank-Kamenetskii, M., Strong bending of the DNA double helix. *Nucleic Acids Res.* **2013,** *41*, 6785-6792.
10. Pardo, L.; Campillo, M.; Bosch, D.; Pastor, N.; Weinstein, H., Binding mechanisms of tata box-binding proteins: DNA kinking is stabilized by specific hydrogen bonds. *Biophys. J.* **2000,** *78*, 1988-1996.
11. Spiriti, J.; van der Vaart, A., DNA bending through roll angles is independent of adjacent base pairs. *J. Phys. Chem. Lett.* **2012,** *3*, 3029-3033.
12. Karolak, A.; van der Vaart, A., Enhanced sampling simulations of DNA step parameters. *J. Comput. Chem.* **2014,** *35*, 2297-2304.
13. Suzuki, M.; Yagi, N., Stereochemical basis of DNA bending by transcription factors. *Nucleic Acids Res.* **1995,** *23*, 2083-2091.
14. Yuan, C.; Rhoades, E.; Lou, X. W.; Archer, L. A., Spontaneous sharp bending of DNA: Role of melting bubbles. *Nucleic Acids Res.* **2006,** *34*, 4554-4560.
15. Yan, J.; Marko, J. F., Localized single-stranded bubble mechanism for cyclization of short double helix DNA. *Phys. Rev. Lett.* **2004,** *93*, 108108.
16. Rohs, R.; Jin, X. S.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S., Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **2010,** *79*, 233-269.
17. Kalodimos, C. G.; Biris, N.; Bonvin, A. M. J. J.; Levandoski, M. M.; Guennuegues, M.; Boelens, R.; Kaptein, R., Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **2004,** *305*, 386-389.
18. Robinson, H.; Gao, Y.-G.; McCrary, B. S.; Edmondson, S. P.; Shriver, J. W.; Wang, A. H. J., The hyperthermophile chromosomal protein sac7d sharply kinks DNA. *Nature* **1998,** *392*, 202-205.
19. Ralf Jauch, C. K. L. N., Kamesh Narasimhan and Prasanna R. Kolatkar, The crystal structure of the sox4 hmg domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochem. J.* **2012,** *443*, 39-47.
20. Peters, R.; King, C.-Y.; Ukiyama, E.; Falsafi, S.; Donahoe, P. K.; Weiss, M. A., An sry mutation causing human sex reversal resolves a general mechanism of structure-specific DNA recognition: Application to the four-way DNA junction. *Biochemistry* **1995,** *34*, 4569-4576.

21. Murphy, E. C.; Zhurkin, V. B.; Louis, J. M.; Cornilescu, G.; Clore, G. M., Structural basis for sry-dependent 46-x,y sex reversal: Modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.* **2001,** *312*, 481-499.

22. 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.

23. Nardini, M.; Gnesutta, N.; Donati, G.; Gatta, R.; Forni, C.; Fossati, A.; Vonrhein, C.; Moras, D.; Romier, C.; Bolognesi, M., et al., Sequence-specific transcription factor nf-y displays histone-like DNA binding and h2b-like ubiquitination. *Cell* **2013,** *152*, 132-143.

24. 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.

25. Mukherjee, A.; Lavery, R.; Bagchi, B.; Hynes, J. T., On the molecular mechanism of drug intercalation into DNA: A simulation study of the intercalation pathway, free energy, and DNA structural changes. *J. Am. Chem. Soc.* **2008,** *130*, 9747-9755.

26. Wilhelm, M.; Mukherjee, A.; Bouvier, B.; Zakrzewska, K.; Hynes, J. T.; Lavery, R., Multistep drug intercalation: Molecular dynamics and free energy studies of the binding of daunomycin to DNA. *J. Am. Chem. Soc.* **2012,** *134*, 8588-8596.

27. Kleinwächter V, B. Z., Reinert KE, Triebel H, Grezes JR, Bär H, A study of drug binding to DNA by partial intercalation using a model phenazine derivative. *Gen. Physiol. Biophys.* **1986,** *4*, 423-432.

28. Lincoln, P.; Norden, B., DNA binding geometries of ruthenium(ii) complexes with 1,10-phenanthroline and 2,2 '-bipyridine ligands studied with linear dichroism spectroscopy. Borderline cases of intercalation. *J. Phys. Chem. B* **1998,** *102*, 9583-9594.

29. Reymer, A.; Norden, B., Enantiospecific kinking of DNA by a partially intercalating metal complex. *Chem. Commun.* **2012,** *48*, 4941-4943.

30. Sasikala, W. D.; Mukherjee, A., Molecular mechanism of direct proflavine–DNA intercalation: Evidence for drug-induced minimum base-stacking penalty pathway. *J. Phys. Chem. B* **2012,** *116*, 12208-12212.

31. Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B., DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95*, 11163-11168.

32. Beveridge, D. L.; Barreiro, G.; Byun, K. S.; Case, D. A.; Cheatham, T. E.; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H., et al., Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(c(p)g) steps. *Biophys. J.* **2004,** *87*, 3799-3813.

33. Goodrich, J. A.; Schwartz, M. L.; McClure, W. R., Searching for and predicting the activity of sites for DNA binding proteins: Compilation and analysis of the binding sites for escherichia coli integration host factor (ihf). *Nucleic Acids Res.* **1990,** *18*, 4993-5000.

34. Torrie, G. M.; Valleau, J. P., Non-physical sampling distributions in monte-carlo free-energy estimation - umbrella sampling. *J. Comput. Phys.* **1977,** *23*, 187-199.

35. Bao, Q.; Chen, H.; Liu, Y.; Yan, J.; Dröge, P.; Davey, C. A., A divalent metal-mediated switch controlling protein-induced DNA bending. *J. Mol. Biol.* **2007,** *367*, 731-740.

36. Case, D. A.; Cheatham, T. E.; Darden, T. O. M.; Gohlke, H.; Luo, R. A. Y.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The amber biomolecular simulation programs. *J. Comput. Chem.* **2005,** *26*, 1668-1688.

37. Wang, J.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000,** *21*, 1049-1074.

38. Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M., Refinement of the amber force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* **2007,** *92*, 3817-3829.

39. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins* **2010,** *78*, 1950-1958.

40. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008,** *4*, 435-447.

41. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R., Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys.* **2009,** *180*, 1961-1972.

42. Sorin, E. J.; Pande, V. S., Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **2005,** *88*, 2472-2493.

43. Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K., Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **2009,** *37*, 5917-5929.

44. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79*, 926-935.

45. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys* **1984,** *81*, 3684-3690.

46. Nosé, S., A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984,** *52*, 255-268.

47. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981,** *52*, 7182-7190.

48. Darden, T.; York, D.; Pedersen, L., Particle mesh ewald: An n·log(n) method for ewald sums in large systems. *J. Chem. Phys* **1993,** *98*, 10089-10092.

49. Souaille, M.; Roux, B. t., Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations. *Comput. Phys.* **2001,** *135*, 40-57.

50. Barducci, A.; Bussi, G.; Parrinello, M., Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008,** *100*, 020603.

51. Jen-Jacobson, L.; Engler, L. E.; Jacobson, L. A., Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure* **2000,** *8*, 1015-1023.

52. Bouvier, B.; Lavery, R., A free energy pathway for the interaction of the sry protein with its binding site on DNA from atomistic simulations. *J. Am. Chem. Soc.* **2009,** *131*, 9864-9865.

53. Bouvier, B.; Zakrzewska, K.; Lavery, R., Protein-DNA recognition triggered by a DNA conformational switch. *Angew. Chem. Int. Ed.* **2011,** *50*, 6516-6518.

54. Du, Q.; Kotlyar, A.; Vologodskii, A., Kinking the double helix by bending deformation. *Nucleic Acids Res.* **2008,** *36*, 1120-1128.

55. Curuksu, J.; Zacharias, M.; Lavery, R.; Zakrzewska, K., Local and global effects of strong DNA bending induced during molecular dynamics simulations. *Nucleic Acids Res.* **2009,** *37*, 3766-3773.

Chapter 6

Molecular Mechanism of Transcription Factor Protein Intercalation into DNA Inducing Bends and Kinks

6.1 Overview

DNA protein interaction is an important and interesting process in biological systems, whose binding mechanism is still unclear. Here, we have studied one system, SOX4-DNA which possess specific protein-DNA interaction where one AA of the protein intercalate into DNA base pairs. Our primary objective is to uncover the molecular mechanism of protein intercalating to the DNA and investigate at what stage DNA kinking and bending take place. Our long well-tempered metadynamics and MD simulations show that the protein primarily binds to the backbone of DNA and rotates over DNA to form an intercalative native state. Moreover, unlike what is reported in the literature, we have not observed significant bending during the binding of the protein. According to present observations bending and kinking happen simultaneously when AA start intercalating between DNA base pair.

6.2 Introduction

In chapter 5, we have found out that a partial AA is required to kink and bend the DNA. Moreover, the kinked bend DNA is not stable, hence protein is there to stabilize the kink DNA. However, we have not investigated the complete mechanism of protein binding. The focus of this paper is to understand the when and how protein choose to intercalate to archived kinked state.

In all organisms, protein-DNA interaction facilitates important biological processes such as DNA replication, transcription, DNA repair, etc. Some proteins, known as transcription factors due to its role in the transcription process, intercalate to a specific pair of base pairs of DNA. The resulting protein-DNA complex produces bent DNA. In chapter 5, we have shown that partial intercalation of the AA is the reason for formation of the bent and kinked state. However, it was not clear how protein finds the specific base pair to intercalate to or at what stage of protein's binding, the DNA would start bending. Therefore, in the present chapter, we would like to investigate the complete mechanism of protein intercalation into DNA and try to address the questions mentioned above.

It is intriguing, however, to realize that proteins find their specific target site given that the target sequence (typically ~16 base pairs) is much smaller than the length of the genome DNA. The current understanding is that proteins first bind nonspecifically to DNA and then search along the DNA, known as one-dimensional (1D) search [1]. Three types of 1D searches have been reported in the literature: (i) sliding[2] (ii) hopping,[3] and (iii) intersegmental transformation[3]. However, these mechanisms depend on the length of DNA and the environment of DNA. In short DNA (typical ~16 BP), these mechanisms are not possible (especially sliding and intersegmental transfer) since the protein itself is large enough that there is no room for sliding. However, hopping is still possible; the protein can break contact partially from DNA and rebind with proper orientation. Along with this mechanism, it has been observed that the protein binds nonspecifically and then spins around DNA to form a specific complex in some systems.[4-5]

There are two major ways by which specific binding happens.[6-9] One is called base readout mechanism where a particular set of DNA base pairs make specific interaction with the AAs via hydrogen bonding, van der Waals contact, etc.[10-12] The other mechanism is known as shape readout mechanism where DNA shape is recognized by the protein. This category is further divided into local shape readout, in which protein recognizes the local deformation such as kinks, nicks, etc. in DNA[13-15] and global shape readout in which proteins recognize globally deformed DNA (like the overall bent structure of the DNA).[16-17]

X-ray diffraction, NMR spectroscopy, and molecular simulations studies on DNA-protein complex have shown that first the protein binds non-specifically to DNA and then searches for specific sites[18-19]. While electrostatic interaction is believed to guide the nonspecific binding[1], hydrogen bonding, van der Waals interaction are responsible for specific interaction.[20] However, a clear molecular understanding of the transition from nonspecifically bound state to specific binding is lacking. In search of the specific site, it has been suggested that the DNA gets deformed to form a specific complex in some complexes. However, when the DNA deformation happens is still unclear. Some theories suggest that the protein recognizes the already deformed DNA to form a specific complex,[13-15, 21] while other hypotheses suggest that the protein binds to DNA and then deforms the DNA[22-23].

There are some specific protein-DNA complexes where a particular amino acid intercalates into the DNA. It induces kinks in DNA (local deformation) which leads to further bending in DNA. The partial intercalation is also thought to be the reason for DNA specificity. The first study of the role of intercalation on specificity was done by Lavery *et. al.* They studied the dissociation of SRY protein from SRY-DNA complex, which has an intercalated AA, ILE13. They suggested that the DNA deformation plays a major role in specific binding. Moreover, they have collected all paths for dissociation and proposed the mechanism of association. They have proposed two path for protein association: one path have a kinked and bend DNA while the other does not show DNA deformation. Because the first path is more populated and matches with experiment, they concluded that the bend and kink determine the DNA specificity. However, they have not shown whether the bending and kinking happen simultaneously or one leads to the other

during the process of protein binding to DNA. Further, they have not studied the association (or binding) process. But with the information obtained from the dissociation of protein from the DNA, they argued that the protein binding and DNA bending happen in a stepwise manner. [24-25] On the same line Kuznetsov *et. al.* also showed that in the IHF-DNA complex the binding and bending happened separately. [22-23]

Even though DNA deformation by protein is a common phenomenon in biology, a clear understanding of the mechanism at the molecular level of details is not available. The questions we are asking here are: (a) when does DNA deform? (b) Are kinking, bending and binding of a protein related? In this work, we try to provide some answers to these questions by studying two systems: SOX4-DNA (PDB ID 3U2B) [26] and Sac7d-DNA (PDB ID 1WD0) [27-28] complex. First, we discuss the observation of SOX4-DNA complex formation and later (in chapter 7) we discuss the Sac7d-DNA complex.

(I) SOX4-DNA Complex: Importance of SOX-4

The SOX4 protein belongs to the HMG family.[29] It is a transcription regulatory protein. These proteins are involved in many biological functions and have a compact DNA binding domain (DBD). Hence, it is a good model for studying protein-DNA recognition. It binds to the minor groove of DNA specifically at the TTCC base pair and also intercalates DNA T8pT9 base pair with MET39 residue, causing a very sharp kink of 45° at that base step. This kinking also induces an overall bending of 69˚ in the DNA. Anatoly *et al.* have experimentally studied the energetics of binding of SOX-4 to DNA[30]. They found that in non-specific complexes, electrostatic interaction dominates, whereas extensive van der Waals contact exists in the interface between a non-polar group of DNA and protein in a specific interaction. Further, by calorimetric experiments they estimated the binding constant between the protein and DNA as ~14 kcal/mol. However, the mechanism of protein binding and intercalation has not been addressed.[30 31 32 33].

To investigate the mechanism of SOX-4 protein binding to DNA is still unknown, we have used molecular dynamics (MD) and well-tempered metadynamics simulations. We have found that protein first binds

nonspecifically to DNA and proceeds towards specific binding mode through different free energy pathway. However, as expected, the unbinding process follows the path that is the most probable binding path. Our investigations here show that the bending and kinking are correlated and happen simultaneously.

6.3 System Preparation

(I) Preparation of Initial Configurations. The crystal structures of SOX4-DNA complexes were taken from the Protein Data Bank (PDB) (PDB ID 3U2B). [26] We have used amber99SB force field with parmbsc0 modification for DNA[34] and amberff-99SB force-filed for proteins[35].

(II) Simulation Details. We have put DNA-Protein complex in a large cubic box of length 130 Å. We solvated the system with TIP3P water model[36], and added salts to neutralize the system and also maintain the physiological ion concentration. Thereafter, we minimized the system with steepest descent algorithm[37], followed by heating up to 300 K using the Berendsen thermostat[38] with coupling constant of 0.2 ps. During the heating process, the heavy atoms of DNA and protein were restrained with the harmonic potential of 25 kcal/molÅ^{-2}. Next, we have removed the restraint in six steps, where each step involves equilibration followed by energy minimization. Finally, we have performed 2 ns unrestraint equilibration at constant temperature of 300 K and constant pressure of 1 bar using the Nose-Hoover thermostat[39] and Parrinello-Rahman barostat[40], respectively, with coupling constants of 0.2 ps for each. Electrostatic interactions were treated with particle mesh Ewald (PME) [41] method with a long-range cutoff of 10 Å. We used van der Waals potential for other nonbonding interaction with 10 Å cutoff. Integration time step for all simulations was kept at 2 fs. All the simulation were performed by using GROMACS-4.5.5.[42] For free energy calculations, we patched the GROMACS with plumed[43]. Analyses of DNA structures were carried out using curves+ [44].

After the equilibration step, well-tempered metadynamics method[45] was employed to study the DNA protein complex formation. We have used 0.7 Å Gaussian width for distance and X, and 5 for native contact and 0.2 kJ/mol height for Gaussian potential. The deposition rate was taken to be 2 ps and the bias factor

of 10 was used. Simulation length was chosen based on the convergence criteria and was different for different processes. However, a total of 1.2 microsecond simulation was performed to accomplish the study.

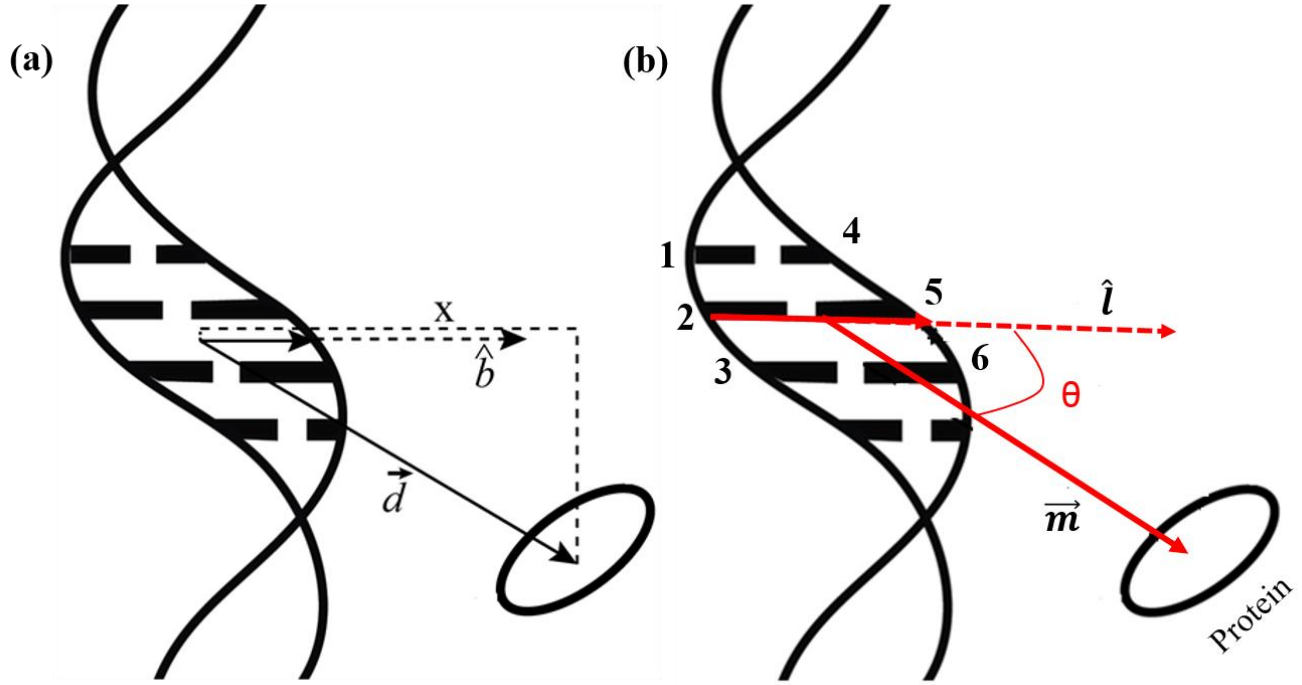

Figure 1. (a) Reaction coordinate X defined as, $X = \hat{b}.\vec{d}$, where $\vec{d}$ is the vector from the center of mass (COM) of four bases involved in intercalation to the COM of the intercalating amino acid. $\hat{b}$ is constructed from COM of four bases to the COM of two sugar groups lying in 3' direction of the intercalating nucleotides. Therefore, X denotes the displacement of the AA perpendicular to the DNA. It is positive towards the minor groove. Note that, displacement along the DNA will not change X. (b) Definition of rotation angle used for analysis. $\hat{l}$ is the vector from COM of 1-3 base step to COM of 4-6 base step and $\vec{m}$ is vector from COM of 1-6 base pair to COM of protein. θ is defined as, $\theta = \cos^{-1}\frac{\hat{l}.\vec{m}}{|\hat{l}||\vec{m}|}$. The number 1-6 are schematic.

(III) Description of the Reaction Coordinate. To study the mechanism of intercalation of proteins, we have performed both association of DNA-protein complex from separated state and dissociation of the same from the native complex. We used different set of reaction coordinates for the above processes in well-tempered metadynamics. They are mentioned below. While we used native contact N_c and vectorial distance X for the association and intercalation/deintercalation processes, we used N_c and distance d between the protein and DNA for the dissociation of the protein from the DNA. The definition of N_c is

given in section 2.4 of chapter 2 earlier. To calculate N_c, we used a distance cutoff of 4.5 Å between the protein and DNA heavy atoms. The value of N_c in the crystal structure using this method was found to be 235. X is defined as (Fig. 1a), $X = \hat{b}.\vec{d}$, where $\hat{b}$ is a unit vector starting from the center of mass (COM) of four bases of intercalating base pair (T8pT9 and A23pA24) to the COM of two 3' sugars (T8 and A24) groups attached with these base pairs. $\vec{d}$ represents the vector from COM of intercalating base pairs to COM of side the chain of amino acid. The reaction coordinate is chosen such that X is positive in the minor groove side region and negative in major groove side region. By construction, X gives a measure of distance of the amino acids perpendicular to the local DNA axis.

We have also defined a rotation angle to estimate the position of protein around DNA. The angle is defined as, $\theta = \cos^{-1}\frac{\hat{l}.\vec{m}}{|\hat{l}||\vec{m}|}$ where $\hat{l}$ is a vector starting from the center of mass (COM) of first strand of intercalating base step (T8pT9 in DNA sequence) to the COM of second strand intercalating base step (A24pA23 in DNA sequence). $\vec{m}$ represents the vector from COM of intercalating base pairs to COM of protein. Fig. 1b shows the schematic representation of angle definition. The angle is defined such that it will we have ~90˚ when it will bind to minor groove and ~0˚ when it binds to DNA backbone.

6.4 Design of the Study

The formation of specific protein-DNA complex is difficult to study computationally since there are numerous non-specific interactions that the system need to sample before it can find specific interactions. Current models suggest that protein first binds to DNA at any random location, and then it starts searching over DNA to make specific complex.[1] Based on these models, we hypothesized that the free energy surface for protein intercalation would possess at least four states as shown in the schematic FES in Fig. 2. (i) The first state is when protein is very far from DNA (there is no interaction between DNA and protein). This state is termed as free state (F). (ii) Next, protein would probably bind and make a non-specific associatiation, the first minimum labelled as 'A' in the schematic FES (Fig. 3). (iii) Following this, a structural/conformational change in DNA/Protein would give rise to another state, possibly more stable

than the previous state. This is termed as "A*" in Fig. 3. (iv) Finally, the system would form the intercalative native complex, labeled as 'I' in Fig. 3.

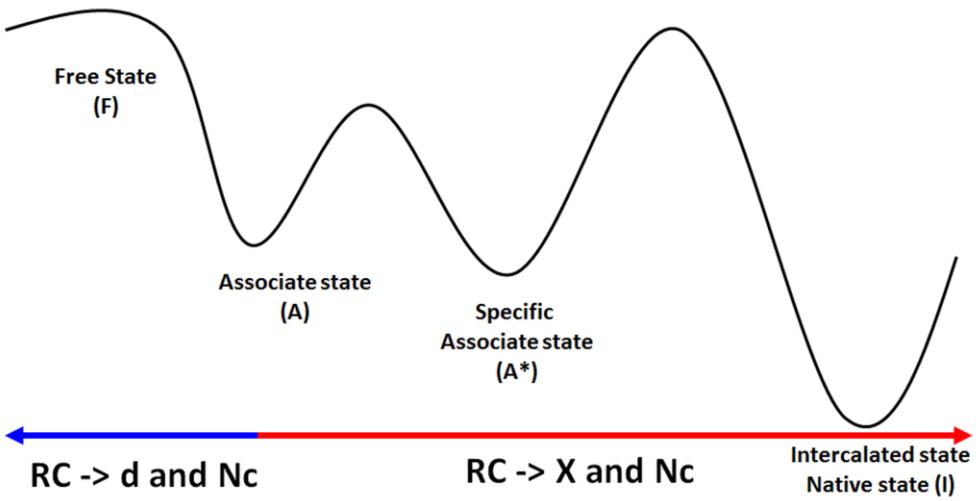

Figure 2: Schematic FE profile of protein intercalation into. We define four states: Free State ("F"), non-specific complex ("A"), specific complex ("A*"), and the native or intercalate state ("I"). RC denotes the reaction coordinate used for association and dissociation of protein.

6.5 Results and Discussion

(I) Getting non-specific Complexes

As described in the schematic FES in Fig. 3, the first problem is to find the most probable non-specific protein-DNA complex from the free protein and DNA. It is difficult, however, to identify the most probable non-specific complex because any binding other than the specific one would result in a non-specific binding. Therefore, there will be several possible non-specific binding modes present between the protein and the DNA.

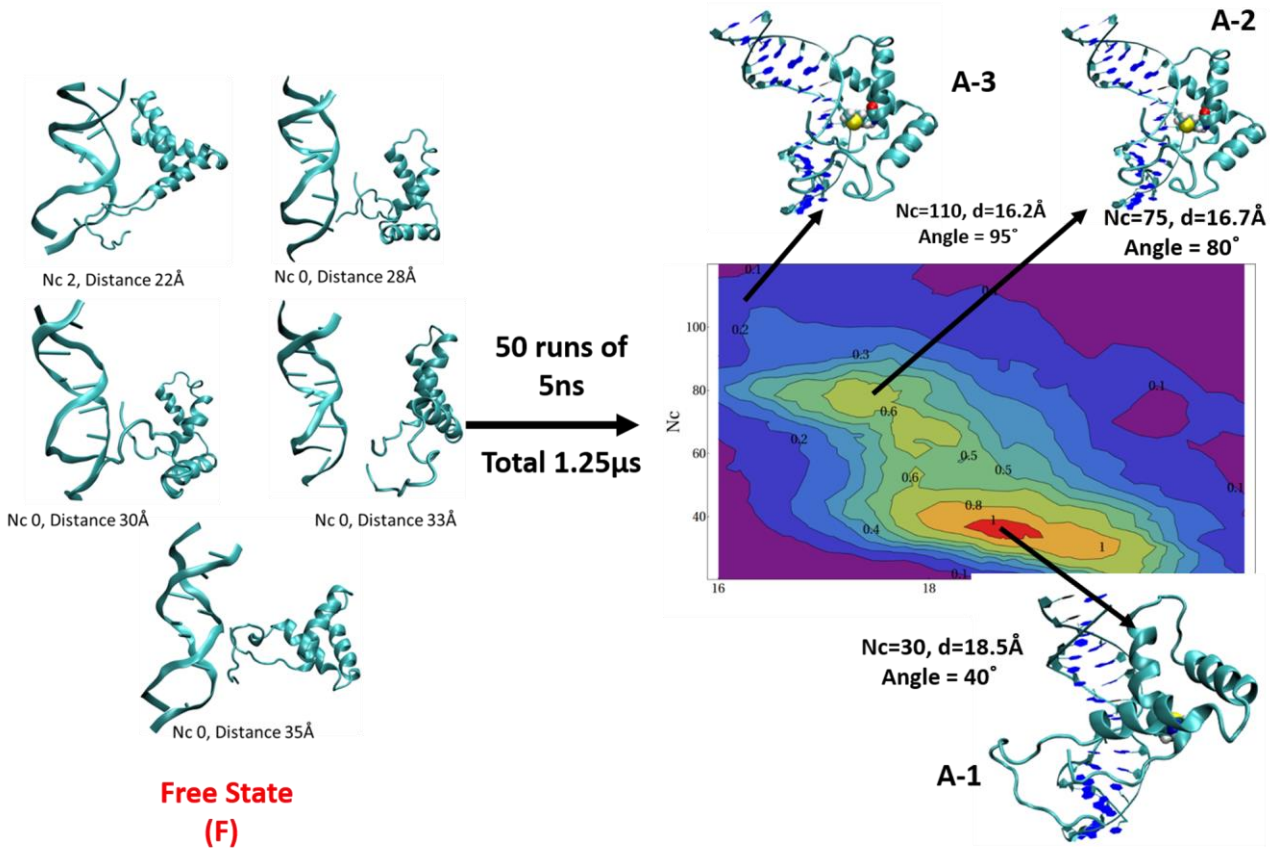

Figure 3. (Left panel) The ensemble of Free States used to obtain the nonspecific complexes. (Right panel) The distribution of N_c and d between the DNA and the protein. A-1, A-2 and A-3 are the non-specific complexes that represent the most probable configuration of that region (see text). The value just beside the complexes are the configurations of those complex.

Therefore, to sample the ensemble of non-specific complexes, we have used the following procedure. First, we have constructed a collection of 5 conformations where DNA and proteins are separated from each other (left side of Fig 4). Then, we equilibrated these configurations 10 ns with restraints on the distance and contact between DNA and protein. At this stage, DNA has a straight (bending angle is ~10°) configraution. After equilibration, we performed 50 short simulations of 5 ns each (called free run) from all the separated conformation (50 simulations from each of these configurations), i.e., a total of 1.25 μs simulation. Various nonspecific DNA-protein complexes produced from these simulations were analyzed from the last 2 ns trajectory. For that, we calculated the native contact, N_c , and distance, d, between the DNA and the protein. The probability distribution of these coordinates is plotted in Fig. 3, which shows

that the nonspecific binding is not uniform indicating that even in the non-specific mode of binding, protein has a preference to bind to DNA at a particular location. There are two major peaks in the distribution (i) when N_c is 30 and distance is 18.5 Å (the most probable non-specific conformation) and (ii) when N_c is 75 and distance is 16.5 Å (second most probable non-specific conformation). This ensemble of non-specific conformation may be regarded as associated complex "A" mentioned in the schematic free energy profile in Fig. 2. Next, we studied the formation of the specific complex from these non-specific complexes. For that, we have selected three complexes from the distribution. The idea was to choose the most probable conformation, second probable conformation and the least probable structure from the distribution. To do so, first we have collected structure around (i) when N_c is ~30 and distance is 18.5 Å (most probable distribution, referred from now on as A-1 system) (ii) when N_c is ~75 and distance is 16.5 Å (second most probable distribution, referred from now on as A-2 system), and (iii) when N_c is 110 and distance is ~16 Å (least probable distribution, referred from now on as A-3 system). After that, we have performed cluster analysis for each of the distribution using Gromos method[46], and chose the middle structure from biggest cluster from each of the distribution. The biggest cluster represents the characteristics of the distribution and middle structure is the best representative of a cluster.[47] The structures obtained from each cluster is shown in Fig 3. All the complexes are very different from each other (see Fig. 3).

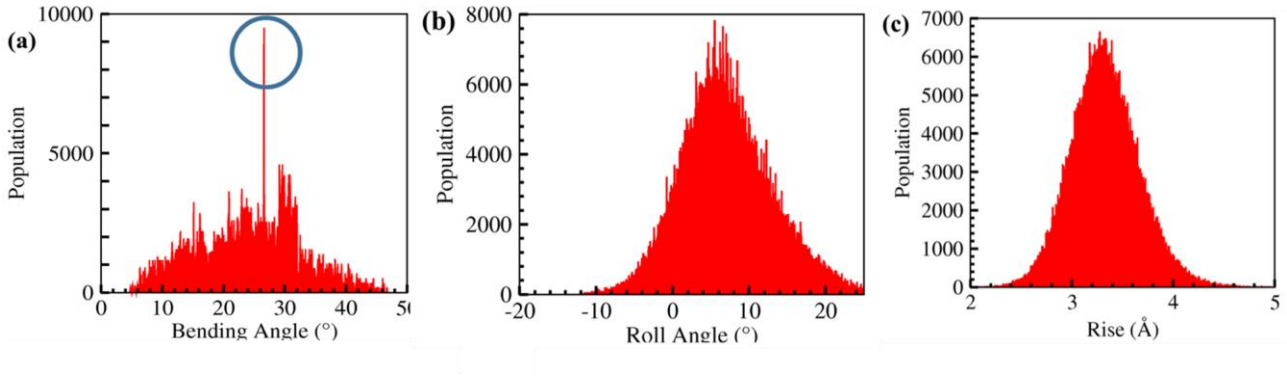

Figure 4: Distribution of bending angle, Roll angle, and rise calculated over the last 2 ns of all free rum trajectory. Circle represent the most probable bending value. Roll and rise are peaked at the canonical value of the DNA.

We also calculated the distribution of bending angle, Roll angle and rise of the DNA over the last 2 ns of all trajectories (Fig 4). The bending distribution has a peak at ~26°, which corresponds to A-1 state. The bending angle of A-2 and A-3 is ~32°. In summary, the bending angle distribution shows that protein binding (non-specific binding) bend the DNA by a small amount. The Roll angle and rise distribution (as discussed in Fig. 6 of chapter 1 and Fig 2 of chapter 5) show that the non-specific protein binding does not change the Roll and rise value.

(II) Characterization of the non-specific Protein-DNA Complexes. After obtaining the possible non-specific complexes, we investigated the nature of the adduct. For that, we calculated the rotation angle (Fig. 1b) of protein around DNA for the selected three non-specific complexes (A-1, A-2 and A-3). The rotation angle of protein in the native protein-DNA complex, where protein is in the minor groove, is 106˚. The rotation angle is ~40˚ for A-1 complex, which indicates that the protein is bound near to the backbone of DNA (perpendicular to the direction of the minor groove). However, in A-2 and A-3 complex, the angles are 80˚ and 90˚ (closer to native state), respectively, indicating that in these two complexes, protein is bound to the minor groove of DNA. The backbone of DNA is negatively charged which interacts strongly with the positively charged AAs of the protein. Therefore, the strong electrostatic interaction between negatively charged DNA and protein's positively charged AA could make the A-1 configuration as the most stable amongst the other non-specific complexes. We will discuss this in terms of interaction energy later (Sec. 6.5.V). This observation is consistent with literature that the non-specific complex is formed by electrostatic interaction.[1]

Further, we have calculated the percentage similarity of all these non-specific complexes from native state. To calculate the percentage similarity, we have calculated the percentage difference between the native contact (N_c), distance (d) and angle (θ) from the native state. To do so, we have calculated the RMSD difference between the associated state and native state along the N_c, d and θ. Figure 5 shows the percentage similarity of all three complexes from native state.

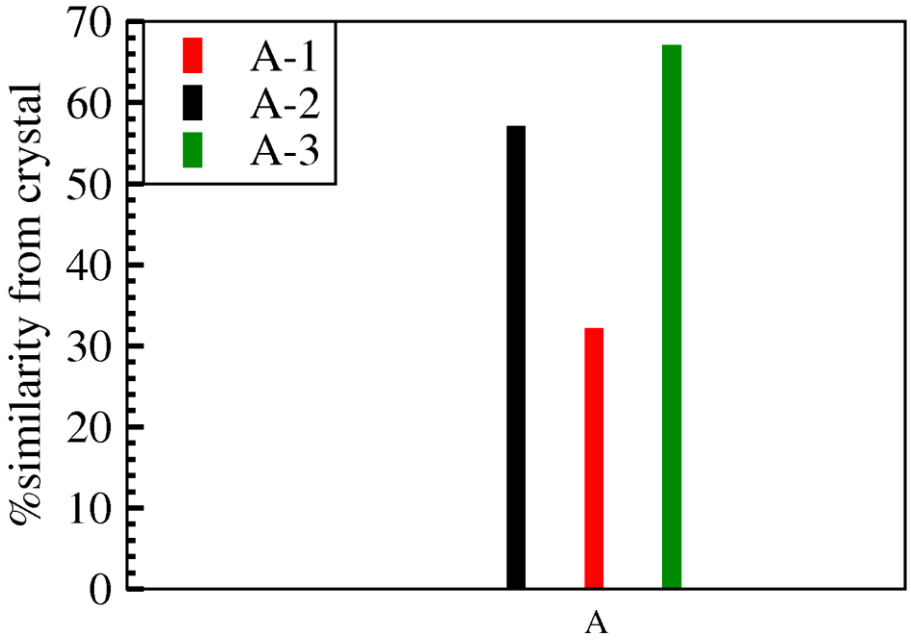

Figure 5: percentage similarity of "A" state (first non-specific complex) from the native state. Averaged over all the three parameters of Nc, distance and angle.

The A-1 complex, where protein bind to the backbone of DNA, has 30% similarity to the native state, whereas the A-2 and A-3, where protein binds to the minor groove of DNA, have 53% and 67% similarity to native state, respectively.

(III) Free Energy for Association and Dissociation of Protein: After getting the non-specific complexes (A-1, A-2 and A-3), we have calculated the free energy for association (formation of the intercalative native state) and dissociation (free unbound state) of the protein from these three non-specific complexes using two sets of well-tempered metadynamics simulations. Since A-1 is less similar to the native state; we hypothesized that A-1 state would have more barriers for the association of protein compared to A-2 and A-3 states. For the formation of the intercalative-native complex, we have used X and N_c as the reaction coordinate (RC) while for dissociation of the protein, we used d and N_c as RC. The reason for the choice of the reaction coordinate is governed by the fact that intercalation is an extremely directional process and X was shown earlier to successfully capture the intercalation mechanism for small organic molecules. [48-49] N_c is important to achieve the specificity present the complex.

Fig 6 shows the FES of association and dissociation for all three complexes (A-1, A-2 and A-3). The white line in each FES shows the minimum free energy path (MFEP) of the process and blue dot line represents the configurational restraint used to stop the sampling those areas. We have put restraint far enough that it should not affect the FE of the minima.

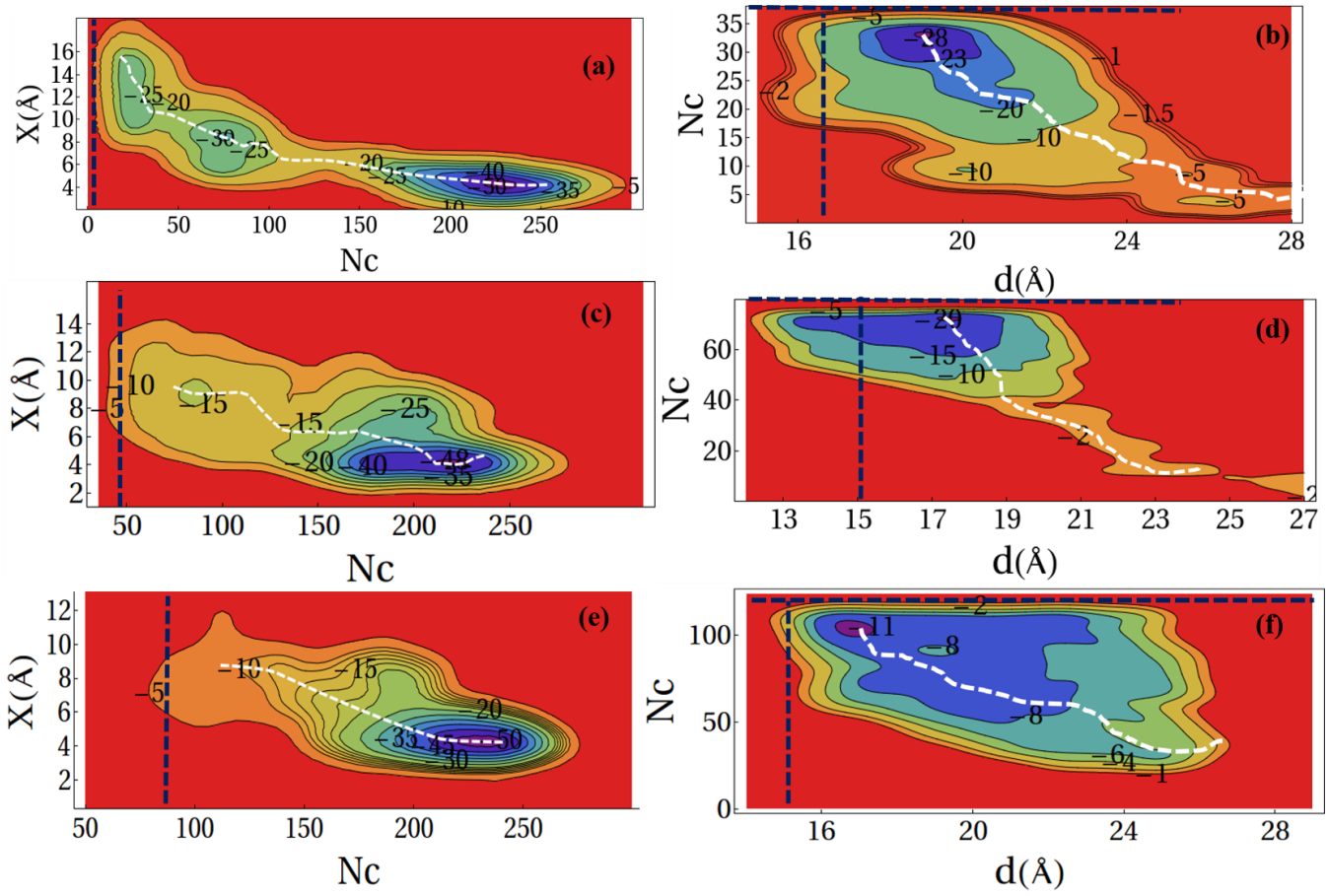

Figure 6: FES for association and dissociation for each complex. (a) The association FES for complex (a) A-1, (c) A-2 and (e) A-3. The dissociation FES of (b) A-1, (d) A-2 and (f) A-3. White dashed line in each FES shows the MFEP for the corresponding process. Blue dotted line represents the conformational restraint used to stop the sampling those areas. Some contour values of the FES are also shown, which indicate that A-1 is most stable amongst the other configurations (A-1 and A-3). Also, all the systems reach to the final native state in the associate FES.

After constructing the FES and subsequent estimation of the FE profile along the MFEP for association and dissociation processes for each non-specific complex, we have combined the FE profiles of both association and dissociation process to get a complete FE for protein binding for each of the complexes,

as shown in Fig. 7. All the complexes go through different paths, but converge to a native state conformation (N_c is 225, d is 14 Å and X is 4 Å), which is global minima for this process.

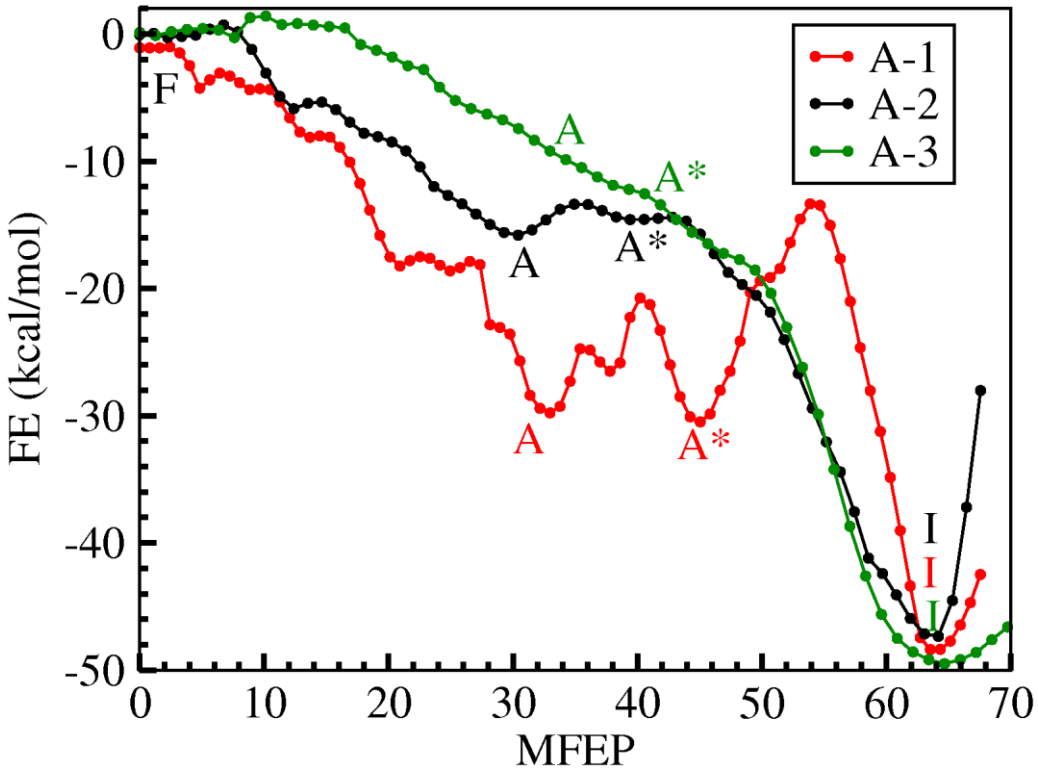

Figure 7: Complete Free energy along MFEP for all the complexes. "F" represents the Free State, where DNA and protein are not interacting. A represents the first associated state (non-specific complex). "A*" represents the next non-specific complex and we represent the intercalative native state.

The initial conformations of the non-specific complexes were different: A-1 was very far from the native state whereas, A-2 and A-3 were closer in comparison. Therefore, all the complexes follow different paths to form the intercalative-native state. However, all the paths do converge to a same native state, characterized by N_c as 225, d as 14 Å and X as 4 Å, with approximately same free energy value (~50 kcal/mol, Fig. 7). For the paths of the A-1 configurations, there are three minima: (i) one where N_c is 40 and X is 13 Å, corresponds to first non-specific associated state (termed as "A"), the second where N_c is 65 and X is 8Å, correspond to the less non-specific (or more specific) associated state (termed as "A*"), and (iii) the last where N_c is 230 and X is 4Å, corresponding to the intercalative native state (termed as I) (red color in Fig. 7). On the contrary, in A-2 and A-3 configurations, these minima are not significant. Therefore, the A-2 and A-3 systems follow somewhat barrier-less paths to form the intercalative native state.

This was excepted since A-2 and A-3 were very close to the native state hence they easily from the native state.

Table 1: Summary of the process and the corresponding associated barrier for all the complexes.

Complex	Transition	Barrier
A-1	A to A*	~9 kcal/mol
	A* to I	~15 kcal/mol
A-2	A to I	~4 kcal/mol
A-3	A to I	~0 kcal/mol

Figure 7, therefore, shows the complete pathway for protein intercalating to DNA from the separated state. The paths led by the most probable state (A-1) correspond to our previous hypothesis that the protein comes from free state ("F" state) and bind to the DNA non-specifically, followed by a rearrangement in the configuration through a ~9 kcal/mol transition barrier, leads to a more stable (-36 kcal/mol) and more specific state, denoted as "A*". Finally, "A*" state rearranges to the native state with ~15 kcal/mol transition barrier. The stability of intercalative-native is ~ -50 kcal/mol. The A-2 also follows the same mechanism (black color Fig 8) but the transition barrier is very less and A-3 does not have any barrier. Therefore, we can state that both A-2 and A-3 already form a more specific "A*" state upon DNA binding and from there directly goes to the "I" state. We have summarized the energies and barrier in Table 1 for all the complexes. Since, A-1, A-2 and A-3 are very different from each other, the barrier and mechanism of the protein binding are different for all the non-specific complexes. The A-1 complex is very much different from the native state. Hence it has to go through higher barrier for intercalative, whereas, A-2 and A-3 complexes are similar to the native state, therefore, they have lesser barrier for the transition. However, A-1 state has most stable associated complex. Therefore, it is more likely that the system would

follow the path of A-1. To understand that, we have also traced the de-intercalation pathway of the same. We will discuss that below.

(IV) How Protein Finds its Target Site:

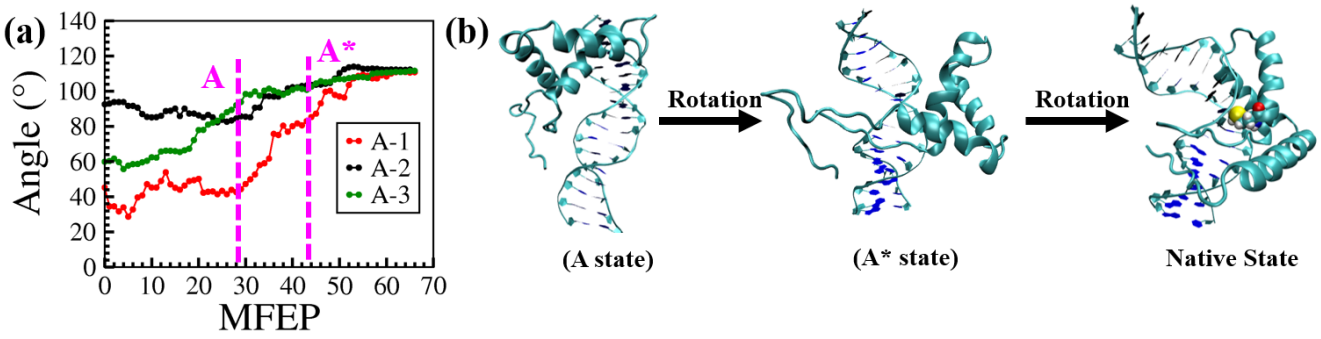

Figure 8: Orientation angle of protein around DNA during the process of binding and AA intercalation. Magenta dashed lines show the minima for "A" and "A*" along the MFEP. (b) Structures at "A", "A*" and native state for A-1 complex. Other complexes have less rotation and they are less probable.

Till now we have calculated the complete free energy profile for protein binding from the free state. Also we have found that different non-specific complexes follow different paths to from the intercalative native state. Next, we have investigated how these non-specific complexes find the intercalative native state. As mentioned before, three mechanisms have been reported in the literature for the formation of a specific complex from a non-specific complex: (i) sliding, ii) hopping, and (iii) intersegment transformation**.** However, in our study, the DNA is too short (~only 16BP) to observe any of these mechanisms.

In the free state (MFEP at 0), the rotation angle of A-1, A-2 and A-3 were 40°, 60°, and 92°. During the free simulation, the angle increased. However, in the nonspecific complex, the rotation angle of A-1 was still less (~40˚) whereas it was ~83˚ and ~95˚ for A-2 and A-3.Therefore, in the non-specific complex A-1, the protein is bound to the backbone of DNA, whereas A-2 and A-3 are in the minor groove. In the intercalative native state, protein is in the minor groove of DNA. Therefore, the protein must rotate around the DNA to make the specific complex. Therefore, in Fig. 8a, we have plotted the rotational angle of protein around DNA. (see method section for angle definition) along the MFEP.

From "A" to "A*" the rotation angle of A-1 rapidly increases. In the "A*" state, the angle of A-1 increases to ~80° and it comes close to the rotation angle of the other two complexes A-2 (100°) and A-3 (100°). As seen in Fig. 8, the rotation angle of all the complexes beyond the "A*" state follow in the same manner. Finally, all the complexes form intercalation with ~106° rotation angle.

(V) Why Specific Binding: non-specific to Specific

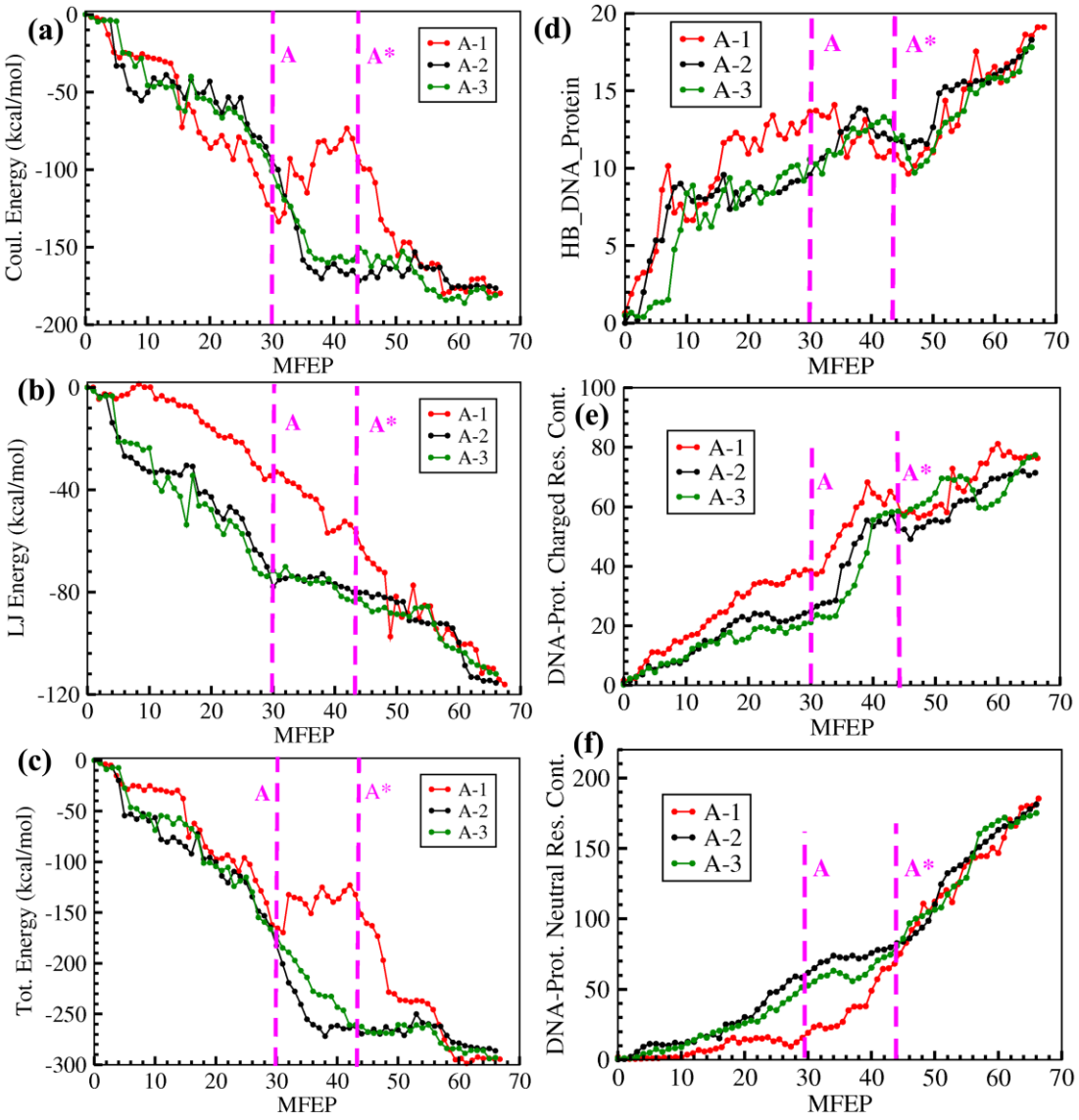

Figure 9: Interaction energy between DNA and protein along the protein binding pathways. (a) Electrostatic interaction energy (b) van der Waals (LJ) interaction energy and (c) the total energy. (e) Total number of hydrogen bonds between DNA and protein. (f) Number of contact between charged AAs (ARG, LYS) of the protein and DNA atom. (g) Number of contact between protein's neutral AAs and DNA. Magenta dashed lines show the MFEP values for "A" and "A*"

It has been reported in the literature that the specific DNA-protein complexes have more van der Waals and hydrogen bonded interaction with DNA, whereas the non-specific complexes have more electrostatic interactions. Therefore, we have calculated the electrostatic and van der Waals interaction between the DNA and protein along the MFEPs of the association process and plotted in Fig. 9. Magenta dashed lines indicate the positions of "A" and "A*" along the MFEP. A-1 system has more electrostatic interaction (~ -125 kcal/mol) compared to A-2 (~ -100 kcal/mol) and A-3 (~ -100 kcal/mol) at the "A" state. This high coulomb interaction is due to the interaction of protein's charged AA to the negatively charged backbone of the DNA. On the contrary, A-1 system has very less van der Waals interaction (~ -38 kcal/mol) at the "A" state while A-2 and A-3 have very strong van der Walls interaction (vale ~ -80kcal/mol) since the protein in A-2 and A-3 system at the "A" state is bound to the minor groove of DNA, thus having stronger interaction with DNA base pairs. Therefore, they have more van der Waals interaction compared to the A-1 system. The contact between DNA and protein's charged residues (ARG and LYS) is also have a high value (~ 55) at the "A" state for the A-1 system compared to the A-2 (~22 contacts) and A-3 (~22) system as shown in Fig. 9e. On the other hand, the contact between DNA and protein's neutral residue is high for A-2 (~ 52) and A-3 (~ 52) system compare to A-1 (~20) system at the "A" state as shown in Fig. 9f.

As protein rearranges (moves to A*" state), the electrostatic interaction energy remains approximately constant in A-2 and A-3 (Fig. 9a) system, whereas it decrease to ~ -65 kcal/mol in A-1 system. This decrease in energy arises due to the rotation of protein around the DNA as mentioned before. The rotation of protein involves the breaking of hydrogen bond and contacts between the protein and DNA backbone (Fig 9d and 9e magenta dot line), which results in the increases in energy. However, as a compensation, the rotation increases van der Waals interaction because the protein now makes new contacts with DNA base pairs in minor groove side. In the other complexes, A-2 and A-3, the change in electrostatic and van der Waals energy remain approximately same due to much less structural rearrangement between "A" and "A*" states in case of A-2 and A-3 complexes.

The total energy between DNA and protein is shown in Fig 9c. The total energy in A-2 and A-3 system at the "A" state is less compared to A-1 state and does not change significantly upon protein binding (Fig 9c magenta dot line). For the A-1 system, the total energy is more at the "A" state compared to the other systems. However, the energy decreases upon further binding. However, at the intercalative native state, all the system have the similar energy (MFEP 66 in Fig 9c).

(VI) Hydrogen Bonding (HB):

Fig 9d shows the total no. of HBs between protein and DNA. At the "A" state, A-1 system has more number of HBs (~14 HB no.), because here protein's charged AA residues (ARG, LYS) bind to the back-bone of DNA. A-2 and A-3 systems have lesser HB (~10 HB no.). When protein moves to the "A*" state, it breaks HB with DNA backbone, resulting in a decrease from 14 to 10 at the "A*" state. (Fig 9d). For the other systems, HB does not change much going from "A" to the "A*" system. The intercalative native state has the maximum number of HBs (~18) compared to any other non-specific complex, which indi-cates that the specificity of protein DNA complex is governed by HB and van der Waals interactions.

(VII) Change in DNA Parameter:

Next, we wanted to investigate the changes in DNA along the path of protein binding. In the native struc-ture of SOX-4-DNA complex, DNA is kinked with a Roll angle of 45°, bent (bending angle 69°) and has a twist of 14°. In the kinked site, MET39 is partially intercalated. Therefore, we looked into the DNA parameters such as Roll angle, rise, bending angle and twist angle to find out how protein binding affects DNA structure and at which state of the intercalation/binding process, DNA starts bending and kinking. Fig 10 shows the changes in the above-mentioned DNA parameters along the MFEP of protein binding for all system.

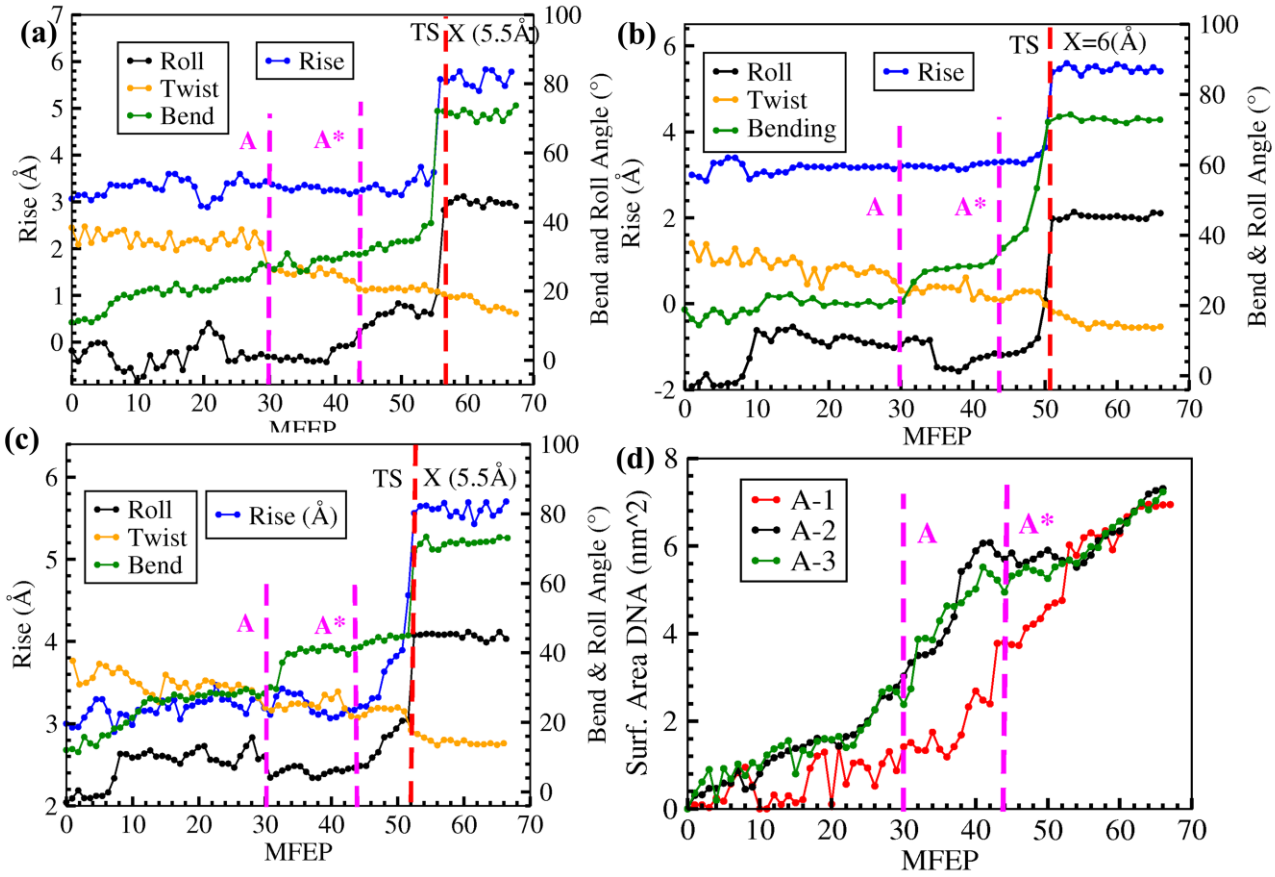

Figure 10: Change in the DNA parameters along MFEP. (a) A-1 system (b) A-2 system and (c) A-3 system. Dashed lines in magenta color shows the positions of "A" and "A*" states. Red dot line shows the intercalative native state. TS represents the transition state. (d) The surface area of DNA used by protein for binding.

When protein is unbound, all DNA parameters have normal B-DNA value (see at the zero of MFEP of Fig 10). Even when protein makes a non-specific complex ("A" state at MFEP 30), rise and Roll have normal values, whereas twist and bend angle has changed to ~26° in the A-1 system, ~30° in the A-2 system and ~30° in the A-3 system. Twist angle decreased to in range of ~21 to 23°. The reason for the small bending is that when protein binds to the DNA, it neutralizes the negative charge of DNA (by making contact between DNA backbone and positively charged AA of protein). This neutralization of charge could lead to small bending in DNA. Also, a bent DNA has more surface area for protein to bind. In Fig. 10d, the surface area (SA) of the DNA is plotted. An increase in SA could be observed at the "A" state for all the systems (Fig 10d). However, the induced bending by protein at the "A" state is not the same as that of the native state. When protein moves from "A" to "A*" state, the Roll and rise do not

change significantly, but bending further increases to ~35° (not to the native state value yet) and twist angle decreases to ~20°. Simultaneously, the SA also increases, allowing protein to bind strongly and make more contacts to DNA (Fig 10d). Therefore, DNA bending helps in protein binding. Finally, when protein moves to intercalative native state from "A*", MET39 starts to intercalate and all the Roll angle, rise, bending, and twist angle start changing simultaneously (red dot line in Fig 10a-c). At the intercalative-native state, the Roll angle is 45°, bending is 68°, rise is 6 Å and twist angle is 18°. The further bending of the DNA results in more SA of the DNA.

The previous study by Lavery *et al.* on SRY-DNA complex[24-25] showed that the binding and bending are two separate processes. However, our present study shows that the binding and bending of DNA happens in a concerted manner and simultaneously. We have observed that the non-specific binding of protein bends the DNA by a small amount, due to the charge neutralization by protein. However, more bending happens when MET39 intercalates into the DNA. Therefore, this results corroborates to our previous observation in chapter 5 that partial intercalation is responsible for the observed kinks and bends present in the transcription factor-DNA complexes. Although bending of DNA is possible just by binding, the amount of bending observed in these systems could not have been possible without partial intercalation.

(VIII) Deintercalation and Dissociation of the Protein from DNA:

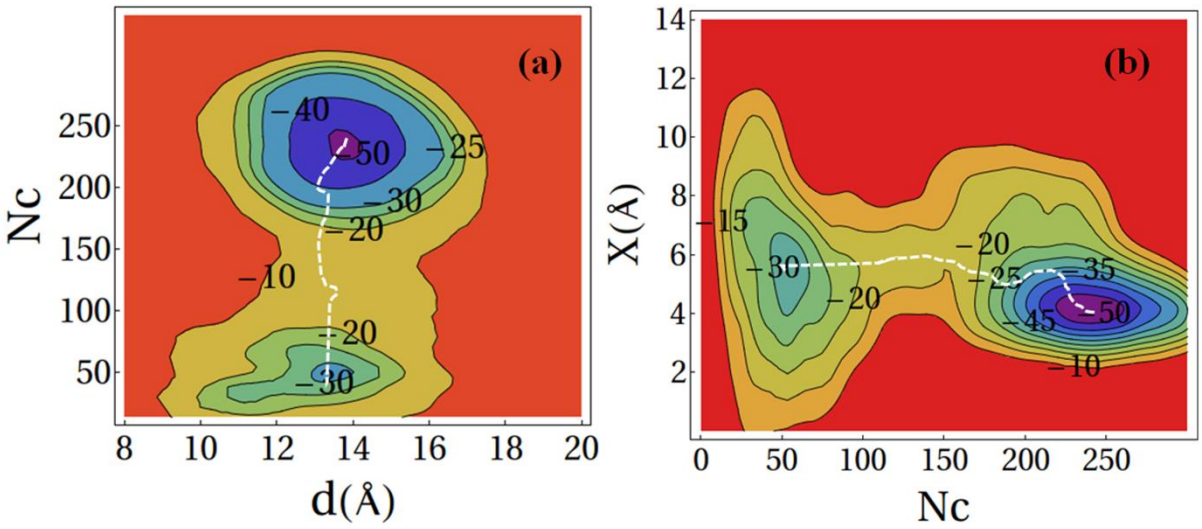

Figure 11: Free Energy surface for deintercalation and dissociation of protein from DNA. (a) Plotted along N_c and distance. (b) Plotted along X and N_c. The white line in each FES shows the MFEP.

We have performed the study of intercalation of protein using three different initial configurations of non-specific complexes. However, the most favorable path for intercalation should be same as that of the de-intercalation pathway by the nature of reversibility. Therefore, we have studied the deintercalation followed by dissociation of the protein from the DNA also ("I" to "F" state). Fig 11 shows the FES for the dissociation of protein and Fig 12a shows the free energy profile along MFEP. The deintercalation free energy profile also shows a barrier of 36 kcal/mol to reach another stable minimum, which we identified as the more specific DNA-bound complex and characterized it as "A*" state. The structure of the native state and the "A*" state superimposed on the native state are shown in Fig. 12a. Figure 12b shows the variation in the DNA parameters (Roll angle, Bend angle, rise) and the reaction coordinate for measuring the depth of intercalation X. Fig. 12b shows that as the X increases, the bending angle and the Roll angle and rise decreases. At the MFEP value of 28, the AA comes out of the DNA and X attains a value of 6 Å. Roll (~25°) and bend angle (~38° value) of the DNA becomes almost normal value when the AA comes out the DNA completely at around MFEP 40.

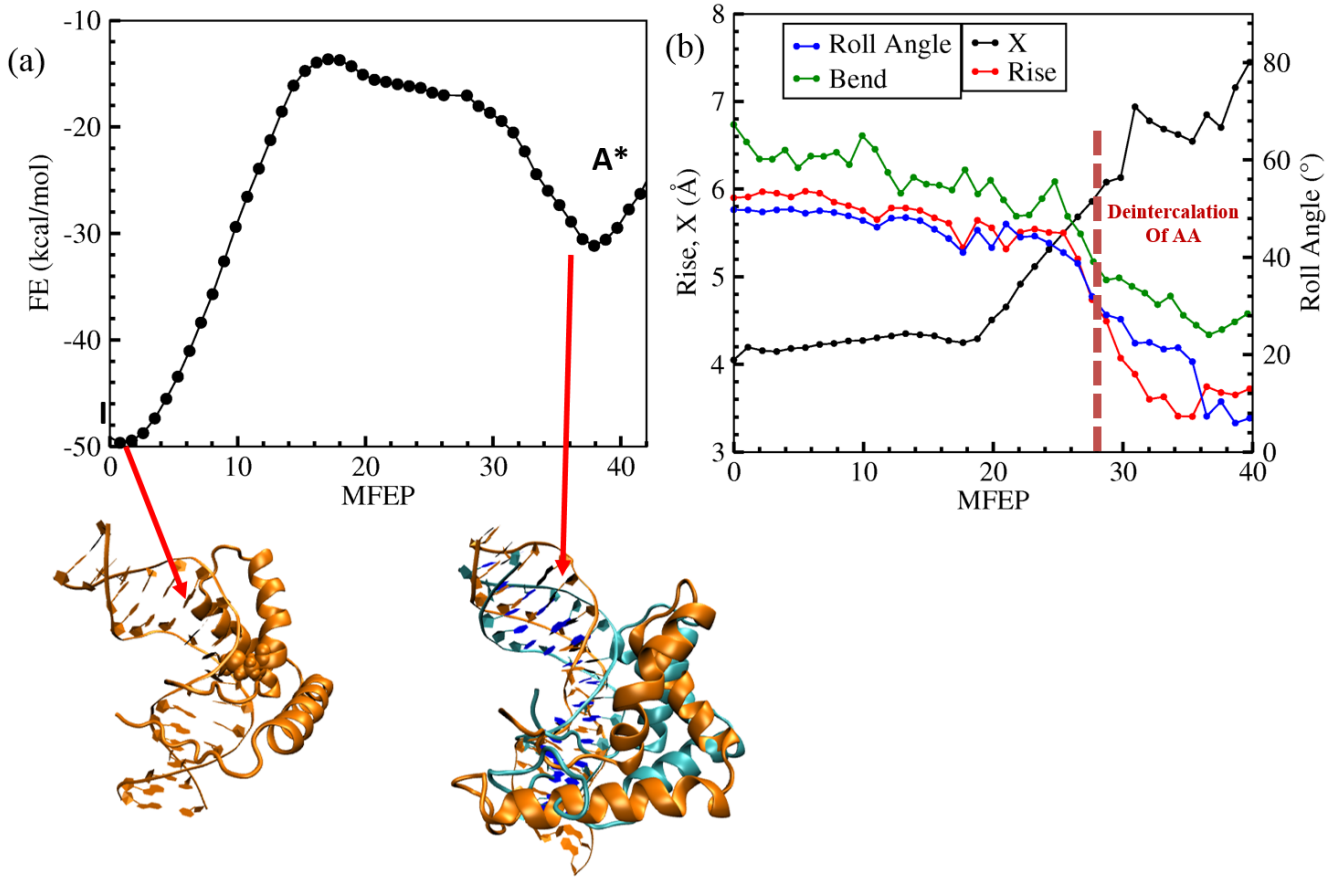

Figure 12: (a) The FE profile along the MFEP for the deintercalation process. The native state (I) structure and "A*" state structure superimposed on native state are shown. (b) Change in DNA parameter along protein dissociation. Brown vertical line shows the point where intercalating AA, deintercalate.

Finally, we compared between the intercalation and deintercalation process of protein-DNA complexes in Fig. 13. We have superimposed the FE profile along N_c for the deintercalation process with the intercalation free energy profile of A-1 system (the most probable path).

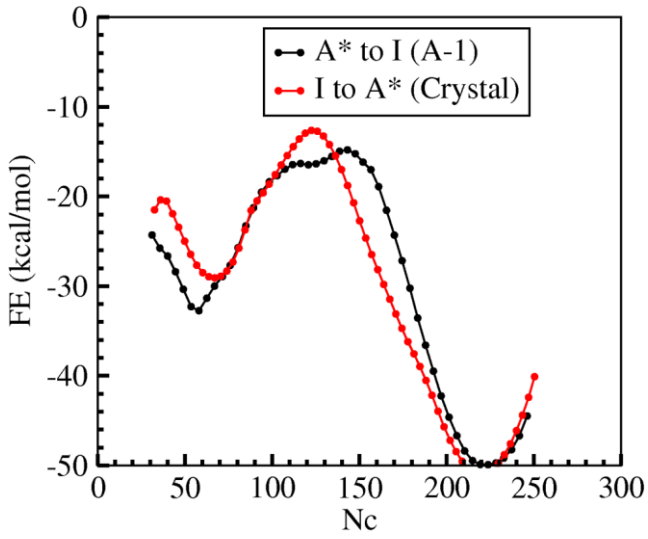

Figure 13: Comparison of free energy along N_c. Black color represents the free energy along N_c for dissociation and red color represents the free energy along N_c for association (for A-2 complex).

Both the profiles are very similar in terms of the barrier height and stability of the "A*" and the native state. This indicates that A-1, indeed, is the right pathway for the protein binding and intercalation. Also, it confirms the existence of a pre-complex formation of protein-DNA before it forms the intercalative native state.

6.6 Conclusion:

We have studied the mechanism of intercalation and deintercalation of a transcription factor protein, SOX-4 to the DNA. We found multiple pathways for intercalation depending on the initial, nonspecific binding. However, there is only one path for the deintercalation process. We showed that the most probable path for the intercalation matches well in the free energy profile with the deintercalation path. The most probable path is formed due to long range electrostatic interaction of the protein's charged residues with the DNA backbone. Thereafter the protein rotates and forms a more specific complex and then it finally intercalates. Even in the non-specific complex, the DNA is found to be bent to a smaller degree, probably due to partial charge neutralization of the DNA. Bending increases the accessible surface area of the DNA to facilitate protein binding. More significant bending and kinking of the DNA happens when amino acid

starts intercalating into the DNA. We have shown that the bending, kinking, and intercalation happen simultaneously.

Reference

1. Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S., Origins of specificity in protein-DNA recognition. *Annual review of biochemistry* **2010,** *79*, 233-269.
2. Riggs, A. D.; Bourgeois, S.; Cohn, M., The lac represser-operator interaction: III. Kinetic studies. *Journal of molecular biology* **1970,** *53* (3), 401-417.
3. Tafvizi, A.; Mirny, L. A.; van Oijen, A. M., Dancing on DNA: kinetic aspects of search processes on DNA. *Chemphyschem* **2011,** *12* (8), 1481-1489.
4. Bagchi, B.; Blainey, P. C.; Xie, X. S., Diffusion constant of a nonspecifically bound protein undergoing curvilinear motion along DNA. *The Journal of Physical Chemistry B* **2008,** *112* (19), 6282-6284.
5. Blainey, P. C.; Luo, G.; Kou, S. C.; Mangel, W. F.; Verdine, G. L.; Bagchi, B.; Xie, X. S., Nonspecifically bound proteins spin while diffusing along DNA. *Nature structural & molecular biology* **2009,** *16* (12), 1224-1229.
6. Joshi, R.; Passner, J. M.; Rohs, R.; Jain, R.; Sosinsky, A.; Crickmore, M. A.; Jacob, V.; Aggarwal, A. K.; Honig, B.; Mann, R. S., Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **2007,** *131* (3), 530-543.
7. Cherney, L. T.; Cherney, M. M.; Garen, C. R.; James, M. N. G., The structure of the arginine repressor from Mycobacterium tuberculosis bound with its DNA operator and Co-repressor, L-arginine. *Journal of molecular biology* **2009,** *388* (1), 85-97.
8. Siggers, T. W.; Silkov, A.; Honig, B., Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *Journal of molecular biology* **2005,** *345* (5), 1027-1045.
9. Luisi, B. F.; Xu, W. X.; Otwinowski, Z.; Freedman, L. P.; Yamamoto, K. R.; Sigler, P. B., Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **1991,** *352* (6335), 497-505.
10. Seeman, N. C.; Rosenberg, J. M.; Rich, A., Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences* **1976,** *73* (3), 804-808.
11. Pabo, C. O.; Sauer, R. T., Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry* **1992,** *61* (1), 1053-1095.
12. Watkins, D.; Hsiao, C.; Woods, K. K.; Koudelka, G. B.; Williams, L. D., P22 c2 Repressor− Operator Complex: Mechanisms of Direct and Indirect Readout. *Biochemistry* **2008,** *47* (8), 2325-2338.
13. Fairall, L.; Schwabe, J. W. R.; Chapman, L.; Finch, J. T.; Rhodes, D., The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **1993,** *366* (6454), 483-487.
14. Horton, N. C.; Dorner, L. F.; Perona, J. J., Sequence selectivity and degeneracy of a restriction endonuclease mediated by DNA intercalation. *Nature Structural & Molecular Biology* **2002,** *9* (1), 42-47.
15. Sierk, M. L.; Zhao, Q.; Rastinejad, F., DNA deformability as a recognition feature in the reverb response element. *Biochemistry* **2001,** *40* (43), 12833-12843.
16. Hizver, J.; Rozenberg, H.; Frolow, F.; Rabinovich, D.; Shakked, Z., DNA bending by an adenine–thymine tract and its role in gene regulation. *Proceedings of the National Academy of Sciences* **2001,** *98* (15), 8490-8495.
17. Rohs, R.; Sklenar, H.; Shakked, Z., Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* **2005,** *13* (10), 1499-1509.
18. Viadiu, H.; Aggarwal, A. K., Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Molecular cell* **2000,** *5* (5), 889-895.
19. Givaty, O.; Levy, Y., Protein sliding along DNA: dynamics and structural characterization. *Journal of molecular biology* **2009,** *385* (4), 1087-1097.

20. von Hippel, P. H., Protein-DNA recognition: new perspectives and underlying themes. *Science* **1994,** *263* (5148), 769-771.
21. Leger, J. F.; Robert, J.; Bourdieu, L.; Chatenay, D.; Marko, J. F., RecA binding to a single double-stranded DNA molecule: A possible role of DNA conformational fluctuations. *Proceedings of the National Academy of Sciences of the United States of America* **1998,** *95* (21), 12295-12299.
22. Slutsky, M.; Mirny, L. A., Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical journal* **2004,** *87* (6), 4021-4035.
23. Kuznetsov, S. V.; Sugimura, S.; Vivas, P.; Crothers, D. M.; Ansari, A., Direct observation of DNA bending/unbending kinetics in complex with DNA-bending protein IHF. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (49), 18515-18520.
24. Bouvier, B.; Lavery, R., A Free Energy Pathway for the Interaction of the SRY Protein with Its Binding Site on DNA from Atomistic Simulations. *Journal of the American Chemical Society* **2009,** *131* (29), 9864-9865.
25. Bouvier, B.; Zakrzewska, K.; Lavery, R., Protein–DNA Recognition Triggered by a DNA Conformational Switch. *Angewandte Chemie International Edition* **2011,** *50* (29), 6516-6518.
26. Jauch, R.; Ng, Calista K. L.; Narasimhan, K.; Kolatkar, Prasanna R., The crystal structure of the Sox4 HMG domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochemical Journal* **2012,** *443* (1), 39.
27. Ko, T.-P.; Chu, H.-M.; Chen, C.-Y.; Chou, C.-C.; Wang, A. H. J., Structures of the hyperthermophilic chromosomal protein Sac7d in complex with DNA decamers. *Acta Crystallographica Section D* **2004,** *60* (8), 1381-1387.
28. Krueger, J. K.; McCrary, B. S.; Wang, A. H. J.; Shriver, J. W.; Trewhella, J.; Edmondson, S. P., The Solution Structure of the Sac7d/DNA Complex: A Small-Angle X-ray Scattering Study. *Biochemistry* **1999,** *38* (32), 10247-10255.
29. Smith, E.; Sigvardsson, M., The roles of transcription factors in B lymphocyte commitment, development, and transformation. *Journal of Leukocyte Biology* **2004,** *75* (6), 973-981.
30. Dragan, A. I.; Read, C. M.; Makeyeva, E. N.; Milgotina, E. I.; Churchill, M. E. A.; Crane-Robinson, C.; Privalov, P. L., DNA binding and bending by HMG boxes: energetic determinants of specificity. *Journal of molecular biology* **2004,** *343* (2), 371-393.
31. Privalov, P. L.; Jelesarov, I.; Read, C. M.; Dragan, A. I.; Crane-Robinson, C., The energetics of HMG box interactions with DNA: thermodynamics of the DNA binding of the HMG box from mouse sox-5. *Journal of molecular biology* **1999,** *294* (4), 997-1013.
32. Crane-Robinson, C.; Read, C. M.; Cary, P. D.; Driscoll, P. C.; Dragan, A. I.; Privalov, P. L., The energetics of HMG box interactions with DNA. Thermodynamic description of the box from mouse Sox-5. *Journal of molecular biology* **1998,** *281* (4), 705-717.
33. Dragan, A. I.; Klass, J.; Read, C.; Churchill, M. E. A.; Crane-Robinson, C.; Privalov, P. L., DNA binding of a non-sequence-specific HMG-D protein is entropy driven with a substantial non-electrostatic contribution. *Journal of molecular biology* **2003,** *331* (4), 795-813.
34. Case, D. A.; Cheatham, T. E.; Darden, T. O. M.; Gohlke, H.; Luo, R. A. Y.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-1688.
35. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-1958.
36. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79*, 926-935.
37. Press, W. H., *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press: 2007.
38. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys* **1984,** *81* (8), 3684-3690.

39. Nosé, S., A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984,** *52* (2), 255-268.
40. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981,** *52* (12), 7182-7190.
41. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys* **1993,** *98* (12), 10089-10092.
42. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E., GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **2008,** *4* (3), 435-447.
43. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009,** *180* (10), 1961-1972.
44. Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K., Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Research* **2009,** *37* (17), 5917-5929.
45. Barducci, A.; Bussi, G.; Parrinello, M., Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008,** *100* (2), 020603.
46. Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L., Conformational and functional analysis of molecular dynamics trajectories by Self-Organising Maps. *BMC Bioinformatics* **2011,** *12*, 158-158.
47. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E., Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* **2004,** *38* (1-2), 236-240.
48. Mukherjee, A.; Lavery, R.; Bagchi, B.; Hynes, J. T., On the molecular mechanism of drug intercalation into DNA: A simulation study of the intercalation pathway, free energy, and DNA structural changes. *J. Am. Chem. Soc.* **2008,** *130* (30), 9747-9755.
49. Sasikala, W. D.; Mukherjee, A., Molecular Mechanism of Direct Proflavine–DNA Intercalation: Evidence for Drug-Induced Minimum Base-Stacking Penalty Pathway. *J. Phys. Chem. B* **2012,** *116* (40), 12208-12212.

Chapter 7

Molecular Mechanism of Sac7d Intercalation into DNA Inducing Bends and Kinks

7.1 Introduction

In chapter 6, we have investigated the mechanism of SOX4-DNA complex formation. In SOX-4-DNA complex, only one amino acid intercalates into the DNA. In this chapter, we have investigated the mechanism a double intercalated AAs DNA-Protein complex, Sca7d-DNA. Sac7d belongs to the class of chromosomal proteins from the hyperthermophilic areachean. Proteins in this class are very stable at high temperature and harsh condition such as high acidic medium.[1-2] These proteins bind to a DNA and thereby increase the stability of the DNA by 20°C.[3] The crystal structure of Sac7d-DNA complex has been solved completely. Sac7d binds to the minor groove of DNA by utilizing the β-sheet motif of the protein. The crystal structure shows a sharp kink of ~60° in DNA base step T5pA6 where two AAs, MET and VAL, are found intercalated i(Fig. 1).[4]

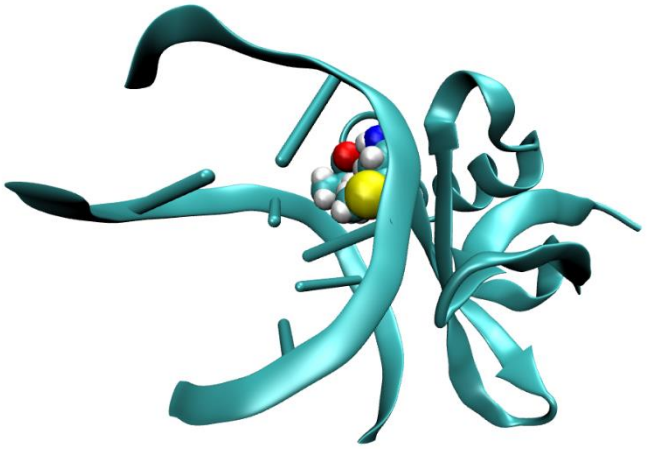

Figure 1. Crystal structure of Sca7-DNA complex. The two intercalated AAs are shown in vdw representation while protein and DNA are shown in ribbon representation.

James *et. al.* studied the binding of Sac7d with DNA in various salt concentration (upto 50 mM). They found the binding free energy is ~10 kca/mol at 50 mM ion concentration.[5] Chin-Yu-Chan *et. al.* muted the intercalating AAs with alanine (ALA) and phelynamine (PHE). They found that when both intercalating AA were muted with ALA, kink disappeared. However, mutation of the AA with PHE retained the kinking. Moreover, mutating any one of the intercalating AA did not change the kinking.[6]

Unlike to the system studied in the previous chapter, studying the present system is more challenging since we have to investigate the intercalation of both the amino acids instead of one. Also, we wanted to understand the nature of the intercalation of the two AA. More precisely, we wanted to investigate whether the intercalation of the AA is simultaneous or concerted.

7.2 Preparation of Initial Configurations. The crystal structures of Sac7d-DNA complexes were taken from the Protein Data Bank (PDB) (PDB ID 1WD0).[4] We used amber99SB force field with parmbsc0 modification for DNA[7] and amberff99SB force-filed for proteins[8].

(I) Simulation Details. We have put DNA-Protein complex in a large cubic box of size 90 Å. We solvated the system with TIP3P water model[9], and physiological plus extra ions were added to neutralize the system. All other simulation details are same as discussed in chapter 6.

(II) **Design of the Study.** Here we have performed the following processes: (i) We have performed free simulations to obtain the most probable non-native state. Then we performed well-tempered metadynamics simulation to obtain the intercalative-native complex. We also performed deintercalation of the protein from the DNA using metadynamics. Moreover, we investigated the deintercalation of the amino acid (not the whole protein) individually to see the effect on the other intercalating AA.

(a) Metadynamics Detail for protein intercalation and deintercalation: To understand the mechanism of Sa7d-DNA complex formation, we have used the same protocol as discussed in chapter 6. However, we have performed two independent metadynamics simulation where we probed one of the intercalating residues to see how it affects the other AA. For this we used native contact and X as the reaction coordinate. Note that, X here measures the distance to one of the intercalating residues.

(b) Umbrella Sampling Detail for Deintercalation of AAs: To study the deintercalation mechanism of the AA, we have used X as reaction coordinate (defined in chapter 5) to deintercalate the AAs. We have divided the reaction coordinate, X, into 20 to 30 windows and for each window, we performed 4 ns simulation to get the free energy profile. Force constant was varied between 20 kcal/mol to 12 kcal/mol. Finally, the free energy for each of the process was calculated using Weighted Histogram Analysis Method (WHAM)[10].

(c) Metadynamic Detail for deintercalation of AA only: The above study performed using umbrella sampling was also performed with metadynamics. However, we used two separate X coordinate, one for VAL and another for MET to obtain a 2D free energy surface. This helped us to compare the intercalation and deintercalation profile of the AA independently. The hill height and width of the Gaussian was taken as 0.2 kJ/mol and 0.7 Å, respectively. The deposition rate was 2 ps. We checked for convergence of free energy to determine the length of the simulation.

7.3 Results and Discussion

(I) Distribution of the nonspecific associated complex: Similar to chapter 6, here also we performed free simulation of protein and DNA from a set of separated protein and DNA (left

panel of Fig. 2). Thereafter, we have performed a total of 50 free simulations of 5 ns for each of the separated complexes. Various nonspecific DNA-protein complexes produced from these simulations were analyzed from the last 2 ns of the trajectory. For that, we calculated the native contact, N_c , and distance, d, between the DNA and the protein. The probability distribution of these coordinates is plotted in Fig. 2 (right panel).

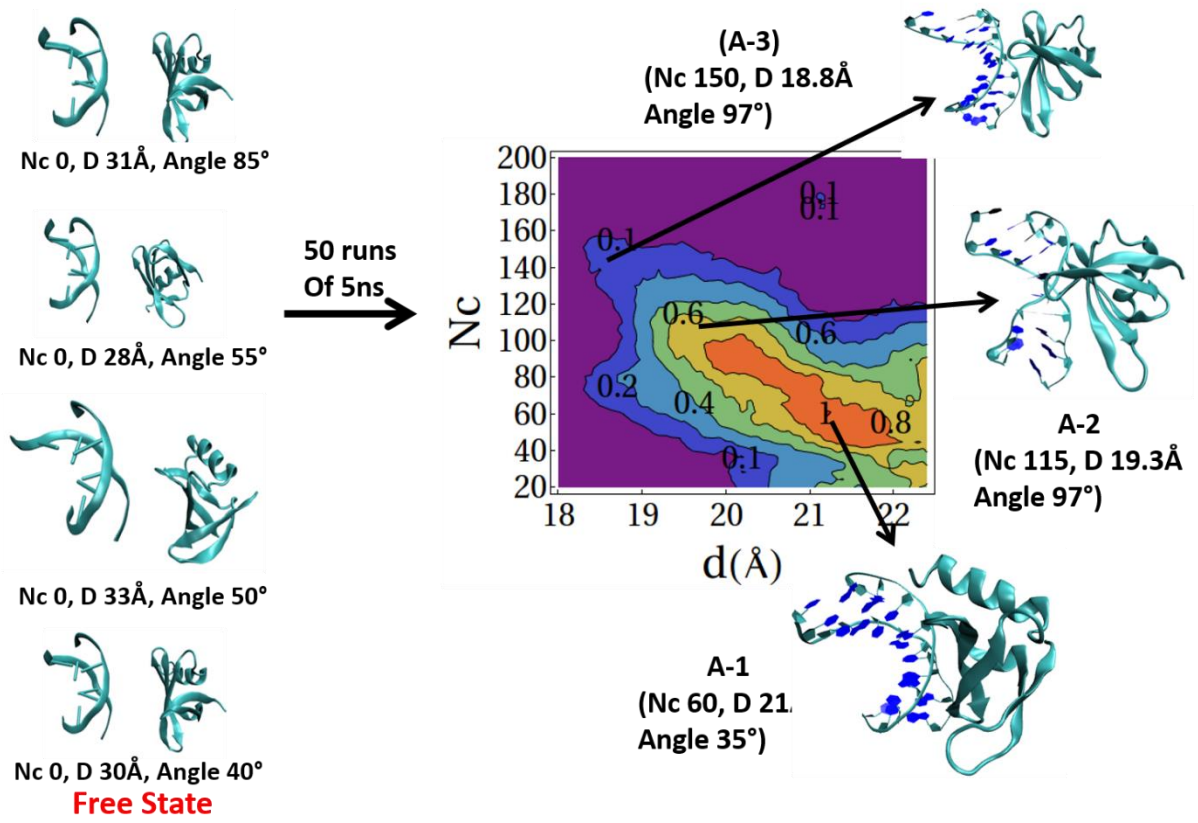

Figure 2: (Left panel) Separated Sac7d-DNA complexes. (Right panel) The distribution of N_c and distance between the DNA and protein calculated over the last 2 ns of free simulation trajectories. The three structures of nonspecific complexes are shown in which A-1 is the most probable nonspecific complex.

Similar to SOX4-DNA system, Fig. 2 also shows the distribution of various nonspecific complex formation between the protein and the DNA. Based on the distribution, we identified A-1 complex as the most probable one. However, we also show a few other (A-2 and A-3) nonspecific complexes in Fig. 2.

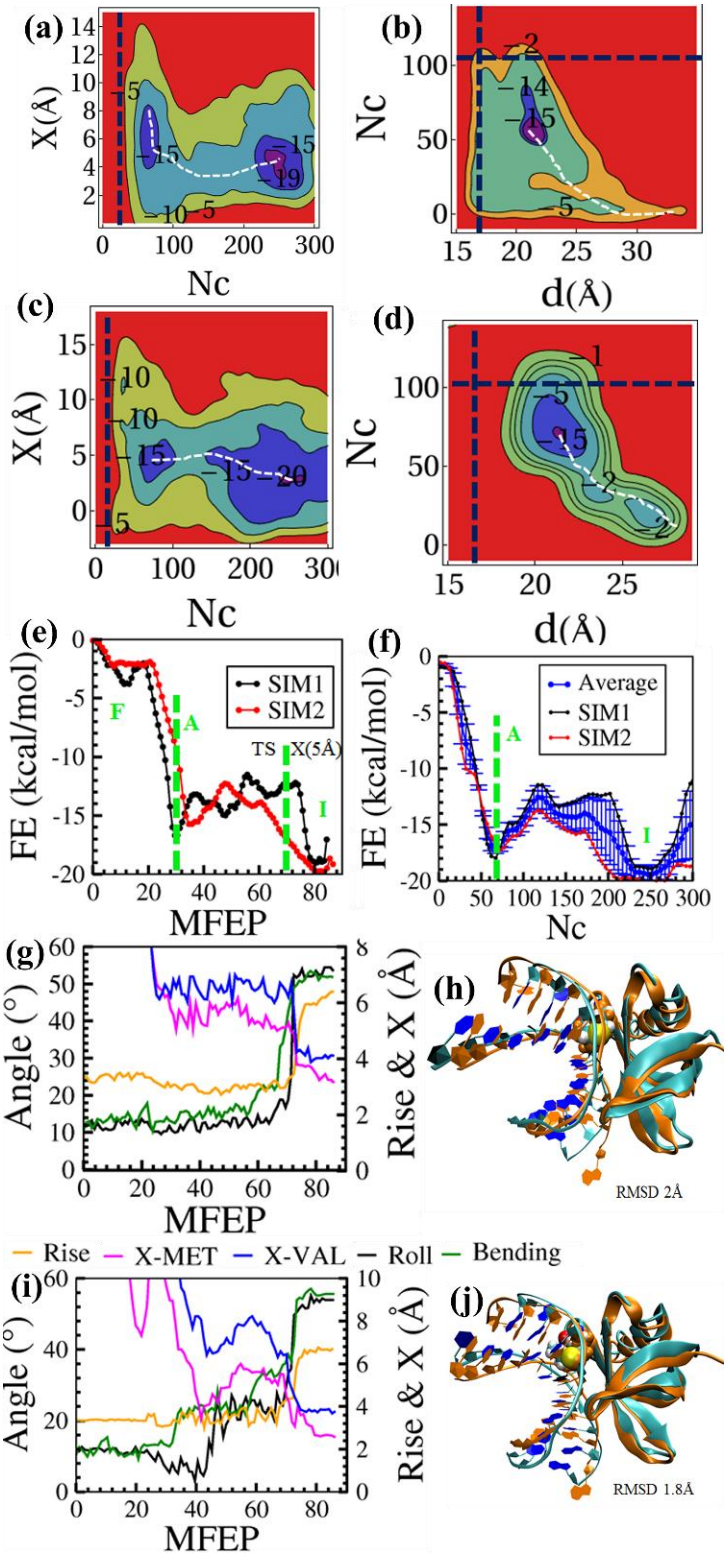

Figure 3. FES for the association and dissociation of protein to DNA. (a) Association of protein when MET is made to intercalate (b) Dissociation of protein. (c) Association of protein when VAL is made to intercalate (b) Dissociation of protein. Note that, for dissociation, SIM1 and SIM2 have no difference since *X* is not used as the reaction coordinate. White dashed line in each FES shows the MFEP of the process. Blue dashed lines show the

configurational restraints applied to avoid the sampling of those areas. (e) The free energy along MFEP and (f) the free energy along Nc for both SIM1, SIM2 and their averages. Green dashed line represents the "A" and transition state. (g) Change in DNA parameter for SIM1. (i) Change in DNA parameters for SM2. (h) and (j) show the native state structure (light green color) (superimposed on the crystal structure, orange color) obtained from SIM1 and SIM2.

(II) Formation of the Intercalative-native State from the nonspecific Complex (A-1): After obtaining the most probable nonspecific complexes (A-1), we have investigated the formation of the intercalative-native complex. Since we showed in chapter 6 that the most probable intercalation mechanism follows by the most probable nonspecific associated state, we only chose A-1 structure here to carry out the intercalation mechanism. We used X and Nc as the reaction coordinate to carry out the intercalation process from A-1 ("A" → "I") and used Nc and d to carry out the dissociation of the protein from A-1 ("A" → "F").

However, the present system poses a challenge as to which AA we should chose to perform the intercalation. Therefore, we performed two independent metadynamics simulations from the A-1 state. In one simulation, we have used X with MET (referred as SIM1) while in the other, we used X with VAL (referred to as SIM2). This indicates that in SIM1, we did not apply any metadynamics bias to VAL and similarly in SIM2, we did not apply any bias for the intercalation of MET.

Figures 3a and 3b show the free energy surface (FES) for the intercalation and deintercalation process for SIM1 while Figs. 3c and 3d show the FES of association and dissociation for SIM2. White dashed lines show the MFEP in each FES. Fig. 3e, we combine the free energy profile for intercalation and dissociation along the MFEP and plot that for both SIM1 and SIM2. We observe that free energy goes downhill to form A-1 with free energy of ~ -15 kcal/mol (Fig. 3e). However, from A-1, the barrier for intercalation was found to be smaller (~4 kcal/mol). The barrier for intercalation is similar for SIM1 and SIM2 indicating that either of the AA can

be used to intercalate the other because in the final intercalated state is same for both SIM1 and SIM2 showing similar stability for the intercalative-native state (~ -20 kcal/mol).

We have compared the free energy along N_c and calculated the average free energy for the association process. This observation suggests that the intercalating anyone AA would eventually force other AA to intercalate. To confirm this observation, we have calculated the DNA parameter and X for both the AA along SIM1 and SIM2. Figures 3g and 3i show the change in DNA parameter along MFEP of protein binding for SIM1 and SIM2 respectively. At the Free State (MFEP 0), the DNA is straight and no kink is present. When protein binds nonspecifically to DNA (MFEP ~25), the DNA conformation does not change and also the AA does not intercalate. Fig. 3g shows that when the MET starts to intercalate to DNA BP at around MFEP 76, rise, roll, and bending angle change simultaneously (orange, black and green color of Fig. 3g). Interestingly, intercalating MET, make other AA (VAL) also intercalate (blue color of Fig. 3g). Similarly, in the other simulation (SIM2) also, DNA parameters rise, roll and bending happen when VAL intercalate. Intercalation of VAL make MET also to intercalate (magenta color in Fig. 3i). This observation suggests that the intercalation of both AA happen together.

(III) How protein finds its target site:

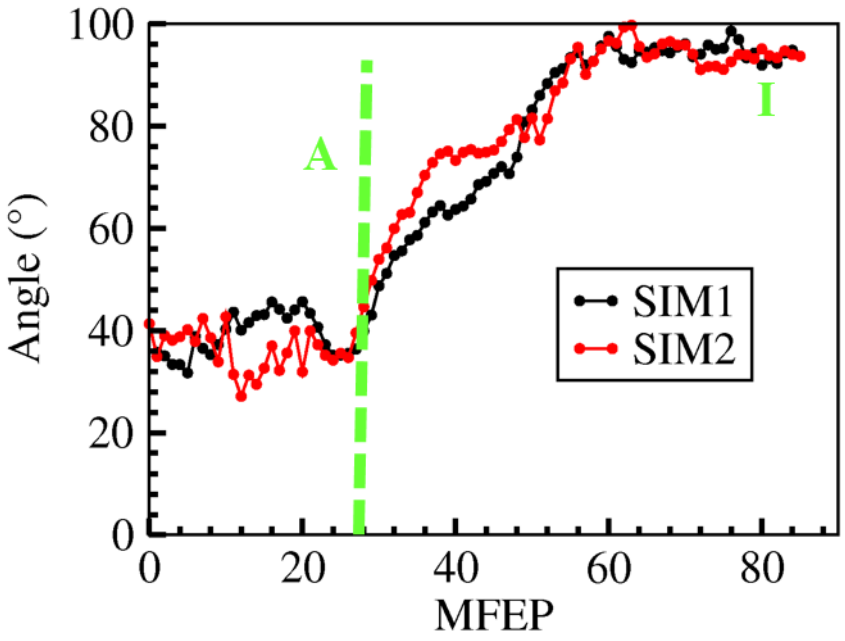

Figure 4: Rotation angle of protein around DNA for SIM1 and SIM2. Note that for both the processes, the change in the angle is similar Green dashed line represents the "A" state.

We have discussed above the complete free energy profile for the protein binding and intercalation from the free state. We have found that intercalating either of the AAs leads to the intercalation of another AA to form an intercalative native state. Now, we would like to understand how the nonspecific complexes form the intercalated native state. To do so, we have calculated the rotation angle of protein around DNA. Figure 4 shows that at the free state rotation angle is ~40°. In the nonspecific complex ("A" state) angle is still ~40°. We observed that as the intercalation proceeds (along the MFEP), protein rotates around the DNA. Interestingly, the amount of rotation is same in both SIM1 and SIM2. (Fig 4). This type of rotation was also found in case of SOX4-DNA complex discussed in the previous chapter. Therefore, we may attribute this kind of rotation as a common mechanism of protein's search for target binding site in DNA.

(IV) Dissociation of Protein from the Intercalative Native State: After the study of the association of protein, we have studied the dissociation of protein from its intercalative native state. To do so, we have used N_c and distance as RC. Fig 5a shows the FES for dissociation of protein against two reaction coordinate, and Fig 5b shows the free energy profile along MFEP. In Fig. 5c, we compare the free energy profile for the complete intercalation process shown earlier with the complete deintercalation process along N_c, Both profiles are remarkably similar, although the intercalation was carried out in two steps (but we have compared with the average free energy) (dissociation from "A" state and intercalation from "A" state).

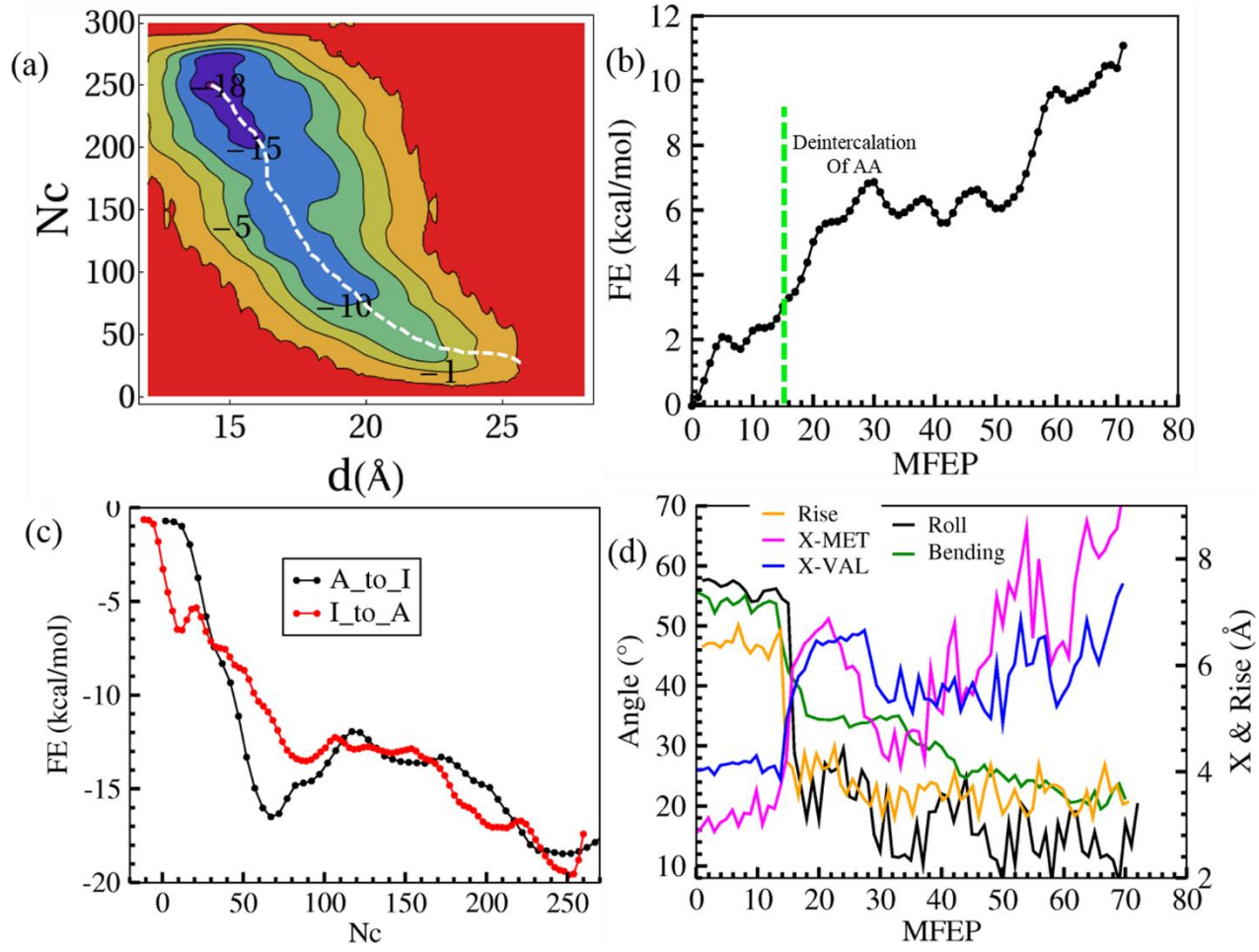

Figure 5: Observation for the dissociation of protein. (a) FES for dissociation of protein from intercalated-native state. (b) free energy along MFEP of dissociation. (c) Comparison of free energy for association and dissociation of protein along Nc. (c) Change in DNA parameter and intercalation of MET, VAL along protein dissociation.

Next, we have investigated that how the deintercalation of AA affect the DNA structure. Fig 5d shows the change in the DNA parameters during deintercalation of AAs. Fig. 5d shows that upto MFEP 14, both AAs are intercalated (rise is 6.3Å) and DNA is bent (angle 55°) and kinked (roll angle is 58° and). Once the AA deintercalate (MFEP 15 on words), kinking disappears (roll angle is 20° and rise is 3.5Å at MFEP >15) and bending angle decreases to 30° (~50%). Roll angle and rise do not change much after deintercalation of AA (i.e., after MFEP 15), whereas bending angle keeps decreasing. DNA becomes straight when protein is completely dissociated (after MFEP 65). Note that, in the deintercalation and dissociation mentioned above, we used native contact and distance as RC. Therefore, we have not biased the deintercalation of the individual AA. Yet, we observe that both AAs deintercalate (MET and VAL) simultaneously (magenta and blue color of Fig 5d). Moreover, due to deintercalation of both AAs together, the roll and rise values of intercalating BP decrease. Therefore, both the deintercalation and intercalation follow the same mechanism that when one of the AA intercalates or deintercalates, other AA follows suit.

7.4 Deintercalation of AA only

Now that we obtained the free energy profile of intercalation and deintercalation of the entire protein, we further investigated the role of the individual AA in the intercalation process. More specifically, we wanted to understand which AA out of two leads the intercalation and deintercalation process.

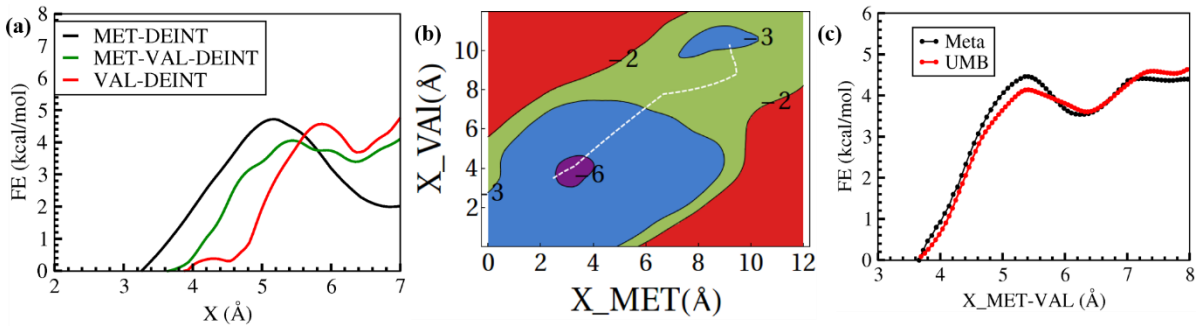

Figure 6: Deintercalation of both the AA by two methods: umbrella sampling and metadynamics. (a) The free energy for deintercalation of MET (black color), VAL (red color) and MET-VAL together (green color). (b) The FES for deintercalation of MET and VAL; white dashed line represents the MFEP. (c) Comparison of free energy profiles obtained from both umbrella sampling (red color) and metadynamics (black color) along X (MET and VAL deintercalation together).

The observation of protein association showed that, when we intercalated only one AA (MET or VAL), other AA also intercalate. Also, dissociation of protein showed that both AA deintercalate together. However, in the above studies, we have used the whole protein (the native contact) to perform the study, which could have influenced the intercalation and deintercalation of the AA simultaneously. Therefore we have performed umbrella sampling and well-tempered metadynamics simulations on the intercalating AA only (keeping the whole protein as it is) to understand the role of these two AA in the intercalation process We performd umbrella sampling simulation using X as reaction coordinate to study the deintercalation of each AA separately and both AA together. Fig 6a shows the free energy of the deintercalation of single AA (MET and VAL separately) and both AAs together. We observed the free energy barrier for deintercalation of single AA (MET or VAL) has the same energy barrier of ~4.2 kcal/mol (black and red color of Fig 6a) as that of the barrier when both are deintercalated. Therefore, deintercalation of only the AAs also leads to the same conclusion that somehow both the AAs are tied together (working like a rigid body). Whether we use one or the two AAs, the observation always leads to a simultaneous intercalation/deintercalation process of both the AAs.

Further, deintercalating of either AA leads to deintercalation of other AA. We have also performed well-tempered metadynamics simulations in two dimensions using X as the RC for both the AA. The FES is shown in Fig. 6b. The free energy profile obtained from this FES along X ois compared with the umbrella sampling free energy profile in Fig. 6c. The remarkable similarity in these two free energy profile confirms again that for the sac7d system, these two AAs work together in the intercalation and deintercalation process.

7.5 Conclusion:

In this chapter, we have studied the mechanism of Sac7d-DNA complex formation. Unlike SOX4-DNA complex (chapter 6), Sac7d –DNA complex has two intercalating AA. We show here that overall mechanism of protein binding to DNA is similar. First, protein binds to the backbone of DNA (nonspecifically) and then rotates around the DNA to form a specific intercalated-native state. The DNA bending and kinking happen simultaneously, when AA starts to intercalate between DNA base pairs (same as we have observed in chapter 6). Moreover, we have observed that the both AA intercalate/deintercalate simultaneously.

Reference

1. McAfee, J. G.; Edmondson, S. P.; Datta, P. K.; Shriver, J. W.; Gupta, R., Gene cloning, expression, and characterization of the Sac7 proteins from the hyperthermophile Sulfolobus acidocaldarius. *Biochemistry* **1995,** *34* (31), 10063-10077.
2. Choli, T.; Wittmann-Liebold, B.; Reinhardt, R., Microsequence analysis of DNA-binding proteins 7a, 7b, and 7e from the archaebacterium Sulfolobus acidocaldarius. *Journal of Biological Chemistry* **1988,** *263* (15), 7087-7093.
3. McAfee, J. G.; Edmondson, S. P.; Zegar, I.; Shriver, J. W., Equilibrium DNA binding of Sac7d protein from the hyperthermophile Sulfolobus acidocaldarius: fluorescence and circular dichroism studies. *Biochemistry* **1996,** *35* (13), 4034-4045.
4. Krueger, J. K.; McCrary, B. S.; Wang, A. H. J.; Shriver, J. W.; Trewhella, J.; Edmondson, S. P., The Solution Structure of the Sac7d/DNA Complex: A Small-Angle X-ray Scattering Study. *Biochemistry* **1999,** *38* (32), 10247-10255.
5. McAfee, J. G.; Edmondson, S. P.; Zegar, I.; Shriver, J. W., Equilibrium DNA Binding of Sac7d Protein from the Hyperthermophile Sulfolobus acidocaldarius: Fluorescence and Circular Dichroism Studies. *Biochemistry* **1996,** *35* (13), 4034-4045.
6. Robinson, H.; Gao, Y.-G.; McCrary, B. S.; Edmondson, S. P.; Shriver, J. W.; Wang, A. H. J., The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature* **1998,** *392*, 202.
7. Case, D. A.; Cheatham, T. E.; Darden, T. O. M.; Gohlke, H.; Luo, R. A. Y.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-1688.
8. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-1958.
9. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79*, 926-935.
10. Souaille, M.; Roux, B. t., Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys.* **2001,** *135* (1), 40-57.

Addendum I

Conformational study of FAD in water and 8M water-urea mixture

AI.1 Overview

The effect of urea over unfolding dynamics of FAD in solution is explored through time resolved spectroscopic and molecular dynamics (MD) simulation methods. Steady state results indicate the possibility of urea induced unfolding of FAD, inferred from increasing emission intensity of FAD with urea. The TCSPC and up-conversion techniques suggest stack-unstack dynamics of FAD is severely getting affected as the unstuck contribution of FAD enhances (from 15% in pure water to 40% in 12 M urea) in expense of stack contribution. To offer a clear insight of FAD and urea interaction up to molecular level, MD simulation is employed. MD simulation depicts that urea replaces many of the water molecules around adenine and isoalloxazine rings of FAD (though total number of hydrogen bonds remain same in water and urea-water mixture). Moreover, a major fraction of urea molecules (~30% to 40%) are detected to stack with both of the adenine and isoalloxazine rings which hinders the intramolecular stacking between isoalloxazine and adenine and thereby enhances the contribution of unstack conformer of FAD with increasing urea concentration.

AI.2 Introduction

The conformations of molecules in solution differ from their crystalline configurations. In solution, molecules exist in several configurations in equilibrium with one another. The well-known example is the FAD (Flavin adenine dinucleotide) which shows the conformational variability in solution. FAD is a member of flavin

cofactor family.Flavin cofactors arean active electron carrier in flavoprotein.They are the nucleic acid derivative of riboflavin. [1-2]Flavin is widely used to study the dynamical behaviour of bio macromolecule because of its fluorescence property.Among others members (namely flavin mononucleotide (FMN) and riboflavin (RF)) FAD shares a special importance as a photoreceptor due to its existence in a wide variety of photoactive flavoproteins such as *DNA-photolyase, BLUF*, and *Cryptocromes.* [3-4]All these flavocofactors contain the fluorescent isoalloxazine ring and ribityl side chain. Unlike others, FAD along with isoaloxazine ring and ribityl side chain, there is an adenosine moiety linked by the phsophodiester group, which acts as a quencher for isoalloxazine fluorescence. The simultaneous existence of the flouroscence and the quencher moiety makes FAD an important molecule to study the conformational behaviour in water and other solvents.

Fluorescence and other studies have shown that FAD exists in two conformations in solution -- the fluorescent open (unstack) conformation and non-fluorescent closed (stack) conformations. [5]In stack conformation, there is a π-π interaction between the adenine ring and the isoalloxazine ring.Fluorescence quenching of FAD is mainly determined by the coplanar stacking of flavin and adenine ring system, most likely through a mechanism of photo induced electron transfer. [6] An intramolecular ground-state complex between the isoalloxazine ring and the adenine moiety was proposed to prevail in aqueous solution resulting in the formation of non-fluorescent complex. [7] Although the existence of an open and closed conformation of the FAD molecule is now generally accepted, little is known about the structural details and conformational dynamics in solution. Molecular dynamic simulation and picoseconds time resolved (TCSPC) and femtosecond fluorescence up-conversion techniqueshave been used to study the conformational dynamics of the FAD in aqueous solution.

Recently it has been observed that the FAD conformation largely depends on the surrounding environment. It was observed that in solutions of lower dielectric constant like ethanol-water mixture, lowering of dielectric constant stabilizes the open conformation and increases the fluorescence lifetime[10]. The glycerol-water mixture, which has high viscosity compare to water and ethanol-water, shows the increased lifetime of the open conformation. [8] Gautman *et al.*showed through experiment and molecular dynamic simulation that water-methanol mixture destabilizes the folded state of FAD, enhancing the contribution of the open conformation. [2]Overall, it appears that the FAD conformation largely depends on the solute-solvent interaction[1]. However, it has been difficult to figure out whether the stacking interaction between the adenine and isoalloxazine or the hydrogen bonding is responsible for FAD dynamics. Recent simulation study showed that the hydrogen bond between the phosphodiester and the ribityl group makes the stack conformation more stable. [7] In this study, we exploited the role of urea in FAD conformation using molecular dynamics simulation of FAD in water and water urea mixture to investigate the role of stacking and hydrogen bonding on FAD dynamics. Urea has been chosen because it is widely accepted as chaotrope for creating chaos in hydrogen networked systems. Also urea is well-known for unfolding protein structures. [9] [10]

AI.3 Simulation Method

FAD was optimized quantum mechanically using HF/6-31G* basis set in GAUSSIAN03 software.[11] Further, restricted electrostatic potential (resp) charges on the atoms of FAD were calculated using ANTECHAMBER module of AMBER11 software.[12-13] The general amber force field (GAFF)[14] for flavin adenine dinucleotide (FAD) and urea were constructed using AmberTools software.[12] The topology and co-

ordinates generated from AmberTools were converted into GROMACS format using a perl program amb2gmx.pl.[15] Follwoign two systems were created for subsequent study: (i) FAD in water only and (ii) FAD in water-urea mixture. To create the FAD-water system, FAD was solvated with 2581 TIP3P water molecules[16] in a cubic box of dimension 4.28 Å. To create FAD-water-urea system, 372 urea molecules and 2581 TIP3P water molecules were added to FAD to make 8 M urea solution. Additionally, two sodium ions were added to neutralize the system.

Both the systems were taken through usual equilibration processes. All the simulations were performed using molecular dynamics software GROMACS.[17] First, each system was minimized using the steepest descent[18] method, followed by heating to 300 K in 100 ps by using Berendson thermostat[19] with coupling constant of 0.2 ps. During the heating process, a harmonic restraint of 25 kcal/mol was applied to the heavy atoms of FAD. The harmonic restraint was gradually reduced to 0.25 kcal/mol in six steps. In each step, 100 ps equilibration was carried out at constant temperature (300 K) and pressure (1 bar) using Berendsen thermostat and barostat[19] with coupling constants of 0.2 ps each. This was then followed by energy minimization using steepest descent method.[18] All bonds were constrained using LINCS algorithm.[20] Particle Mesh Ewald (PME) method[21] was used for the electrostatics with a 10 Å cut-off for the long range interaction. Cut-off for the van der Waals (vdW) interaction was kept at 10 Å. Simulations were performed by 1 fs time step. At the final equilibration step, we performed the 1 ns unrestrained equilibration by using the Berendsen thermostat and barostat[21] with coupling constant of 0.2 ps each.

After the equilibration steps, a final unrestrained simulation of 200 ns was performed at constant temperature 300 K and constant pressure 1 bar using the Nose-Hooverthermostat[22] and Parrinello-Rahman barostat[23], respectively, with 0.2 ps

coupling constant for each. Results reported here were obtained from this final simulation.

AI.4 Results and discussion

(I) Definition of the analysis parameters. Simulation trajectories of 200 ns for each system were used to analyze the conformational sampling of FAD in pure water and water-urea mixture. The following three parameters were used to understand the conformational variations of FAD: (i) distance, *d*, between the center of masses (COMs) of the adenine (Ade) and isoalloxazine (Iso) rings (magnitude of the vector $\vec{e}$, Fig. 1), (ii) angle, θ, between the plane of the Ade and Iso rings (angle between the vectors $\hat{g}$ and $\hat{f}$ in Fig. 1), and (iii) angle, φ between the vector normal to the Iso ring and the vector from the COM of Iso to the COM of Ade (angle between two vectors $\hat{f}$ and $\hat{e}$ Fig. 1). Stack configuration is defined when d ≤ 6.0 Å[1], and both angles (θ and φ) are either less than 40º or greater than 140º, i.e., 40 º ≥ {θ, φ} ≥ 140º, as shown schematically in Fig. 1a.

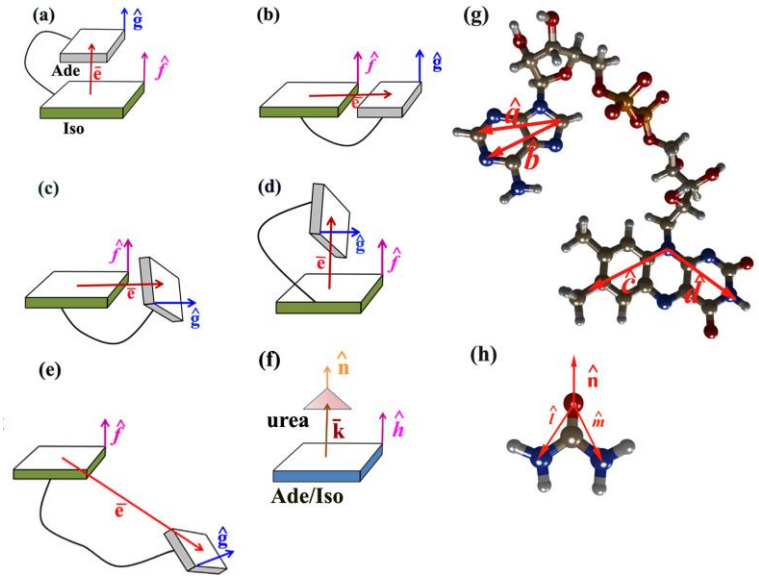

Figure 1: Schematic representation of the stack and unstack conformation. (a) represents the stack conformation, where d ≤ 6.0 Å, and θ, φ = 0º. (b)represents the unstack conformation, where d ≤ 6.0 Å, θ = 0º, φ = 90º. (c) represents the unstack conformation, where d ≤ 6.0 Å, θ = 90º, φ = 90º. (d) represents the unstack conformation, where d ≤ 6.0 Å, θ = 90º, φ = 0º. (e) represents the unstack conformation, where d ≥10 Å. (f) represents the stack conformation of Ade/Iso with urea when k ≤ 5.2 Å, 40 ° ≤ { α, β} ≤ 140 °. $\vec{e}$ represents the vector from COM of Iso ring to Ade ring. $\hat{g}$ represents the vector normal to the plane of Ade ring, which is a cross product of two vectors, $\hat{a}$ and $\hat{b}$ (Fig. 1g). $\hat{f}$ represents the vector normal to the plane of Iso ring, which is a cross product of two vectors, $\hat{c}$ and $\hat{d}$ (Fig. 1g). $\hat{n}$ represents the vector normal to the urea molecule, which is a cross product of two vectors, $\hat{l}$ and $\hat{m}$ (Fig. 1h).

We also define the two different types of unstack conformations: (i) complete unstack conformation (U4) where d>10 Å and (ii) partially unstack where d<10 Å. Based on our observation of the geometry of FAD, we further divide the partially unstack into three categories: (i) U1, when d< 6.0 Å and 40° ≤ {θ, φ} ≤ 140° (ii) U2, when 6 Å ≤ d 10 Å and 40° ≥ {θ, φ} ≥ 140° (iii) U3, 6 Å ≤ d 10 Å and 40° ≤ {θ, φ} ≤ 140°. Therefore, U1 denotes the partial unstack configuration where Ade and Iso ring are close to each other, but not in the preferably stacked configuration. U2 and U3 denote intermediate distance separation. However, in U2 the angle between the planes conform to stacked geometry, but the rings are far away (>6 Å) to have perfect stacking interaction.

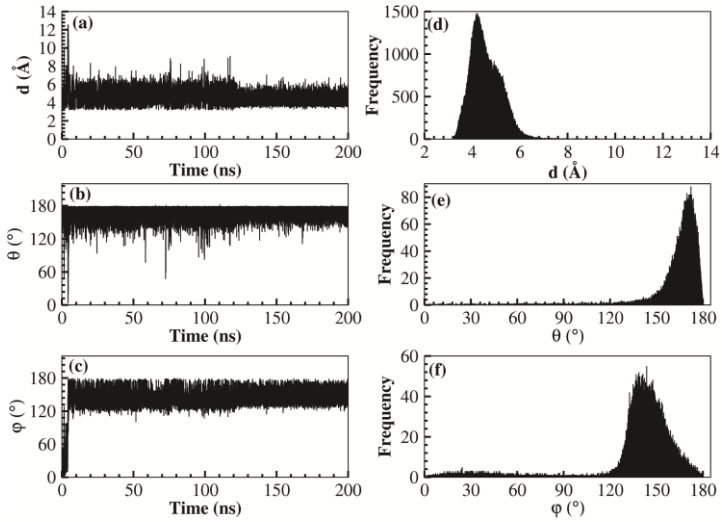

Figure 2: (a), (b) and (c) show the variation of d, θ, φ, respectively with time in water. (d), (e) and (f) are the distribution of the three parameters (d, θ and φ) in water solvent. When $d \leq 6Å^1$, and the angles is $40^{o} \geq \{\theta, \varphi\} \geq 140^{o}$, we defined these conformation as stack conformation.

Fig. 2 shows the variations of the above three parameters for FAD in water. Fig. 2a shows the time variation of the distance between Iso and Ade rings in water while their distribution is shown in Fig. 2d. Distribution is peaked at 4.2 Å (with 0.6 Å standard deviation). This indicates that most of the time the two rings of FAD are in close proximity. Time variation of the two angles (θ, φ) are shown in Figs. 2b and 2c while their corresponding distributions are shown in Figs. 2e and 2f, respectively. It is clear from the distribution that θ is peaked at 172.1º, while φ is peaked at 143.5º. Therefore, according to the stacking criteria (as mentioned above), FAD remains in the stacked configuration most of the time in water.

Similar analysis of FAD is carried out for FAD in water-urea mixture (8M) and the results are shown in Fig. 4. A dramatic fluctuation in FAD conformation is observed here. It is clear from the time-variation of the distance and angles that FAD prefers to exist in unstack coformation most of the time in the mixture. However, FAD does not remain always in the unstack conformation and stack-unstack equilibria continues.

Red boxes in Fig. 4 indicates the time zones where stack conformation appeared during the simulation. Distribution of the distance (*d*) between the two rings exhibits (Fig. 4d) distinct peaks at 4 Å, 8 Å, 11 Å and 15 Å. Interestingly, distribution of two angles, θ and φ are almost uniform which indicates that the preference for the stack conformation present in pure water is affected by urea.

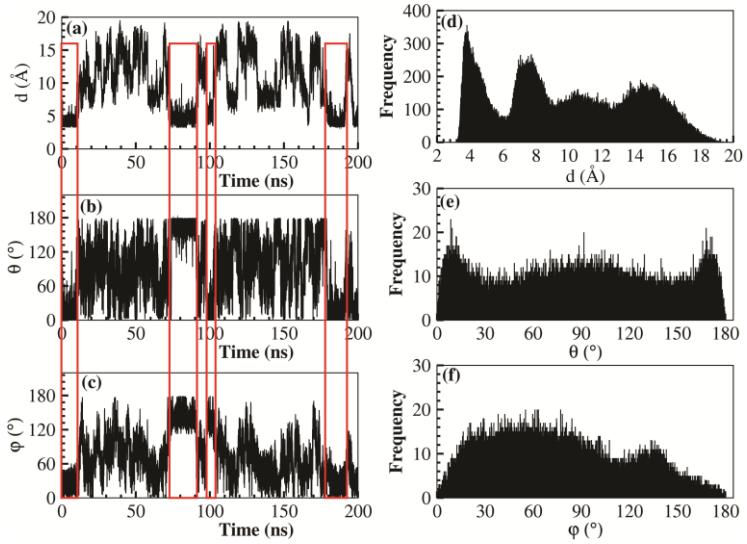

Figure 4: (a), (b) and (c) is Variation of the three parameters over trajectory. (d) ,(e) and (f) are the distribution of the three parameters (d, θ and φ) in water-urea mixture. When d ≤ 6Å and the angles is 40 º ≥ {θ, φ} ≥ 140º we defined these conformation as stack conformation. Red boxes are showing the stacking conformation present during the simulation.

The stack-unstack equilibrium is analyzed to find the probability of each conformation in both water and water-urea mixture. Table 1 shows that in water 64% of FAD conformation are in stack form and rest of them are in unstack form, which is in close agreement with the experimental observation (Table1). The unstack form in water is dominated mostly by partial unstack conformation U3 (35.6%), where the distance between two rings lies between 6 Å and 10 Å and the angles θ and φ are between 40º and 140º. This indicates that in water, FAD never actually goes to the

fully unstack configuration. Rather the rings adopt different orientations to get out of stack interaction.

Astonishingly, in the water-urea mixture, only 14.4% FAD molecules exist in the stack conformation and remaining 85% FAD molecules prevail in the unstack conformation. Among the total population of unstack conformation, the majority of the structure exists in the completely unstack conformation U4 (47.3%), where distance between the two rings are more than 10Å. The partially unstack conformation contain all the three possibilities (U1, U2 and U3) discussed previously. 18.4% structures contain two rings close to each other with unfavorable stacking orientation (U1), while 5.4% are of U2 type, where angles are of stacking type, but the distance between the rings does not allow stacking interaction. Rest 14.5% exist in U3 conformation, where both the distance and the angles do not allow stacking to occur. Note that, in water most of the unstack conformation are of U3 type. Fig. 5 shows some representative configuration of urea around FAD in stack and unstack configuration.

Table 1: Percentage of stack and unstack conformation of FAD in water and water-urea mixture calculated over the whole trajectory.

Solvent	Stack conformation	Unstack conformation			
		U1	U2	U3	U4
water	64%	0.1%	0.2%	35.6%	0.07%
water-urea mixture(8M)	14.4%	18.4%	5.4%	14.5%	47.3%

To understand the effect of water and urea on FAD conformation, we analyzed hydrogen bonds (HB's) between different parts of the FAD (Ade, Ribose (Rib), Iso rings and Phosphate groups (Phos)), and with the solvent molecules (water and urea). HB was determined by defining a distance and an angle cut-off. Distance cut-off between the donor and acceptor was taken as 3.5 Å, while the cut-off for the angle formed by the donor, hydrogen, and acceptor was taken to be 30º.[19] Average HB's between different parts of FAD and with solvents are shown in Fig. 6.

HB's between different parts of FAD (Ade, Iso, Rib and Phos) (intramolecular HB's) is almost non-existent in water, which is consistent with the simulation results reported by Radoszkowicz et al.[(ref)] In urea-water mixture, number of HB's between different parts of FAD is also negligibly small. The number of HB's between different parts of FAD and the solvent molecules in general is similar in case of pure water and the mixture (Fig. 6). Interestingly, in the mixture, some of the water molecules get replaced by urea for making HB's with different parts of FAD as shown in Fig. 6. Phosphate (Phos) makes maximum HBs with water in pure solvent (11.3). However, in the urea-water mixture, 5.5 water molecule and 8.5 urea molecules makes HBs with the phosphate group. Similarly, average number of HB's of Ade and Iso with water in pure water is 4.7 and 5.1, respectively and this changes to 2.7 and 2.9 in the mixture; whereas average HB's of Ade and Iso with urea in the mixture is 2.4 and 3.4, respectively. Therefore, our simulation results suggest that hydrogen bond is not the major driving factor towards the stack to unstack dominance in presence of urea.

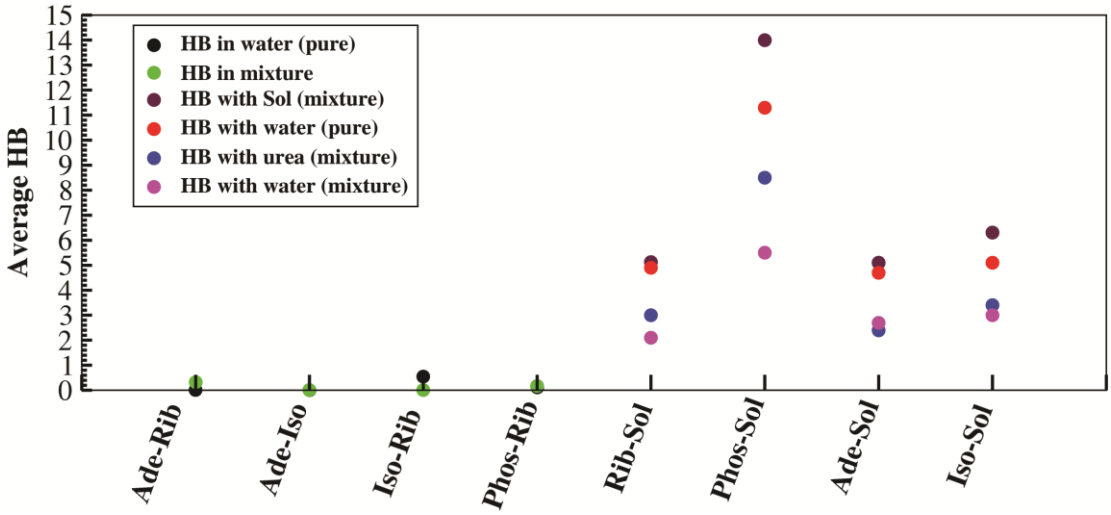

Figure 6: Average Hydrogen bonding (HB) of the different part of FAD and with the solvent (water and water-urea mixture). The black point is the average HB in the water solvent. The green point is the average HB in the water-urea mixture (8M).

To see the effect of urea on the FAD conformation, we have monitored the stacking and unstacking behaviour of urea with Ade and Iso ring by using three parameters: (i) distance k, between the COM of Ade/Iso to the urea molecule (magnitude of vector $\vec{k}$ in Fig. 1f), (ii) angle α, between the vector normal to the Ade or Iso ring and the vector normal to the urea molecule (angle between the unit vector $\hat{n}$ and $\hat{h}$ in Fig. 1f) (iii) angle β, between the vector normal to the Ade or Iso ring and the vector from the COM of Ade or Iso to the urea molecule (angle between the vector $\vec{k}$ and $\hat{h}$ in Fig. 1). To define the normal vector of the urea, we took the cross product of the two vectors $\hat{l}$ and $\hat{m}$ as shown in Fig. 1h.

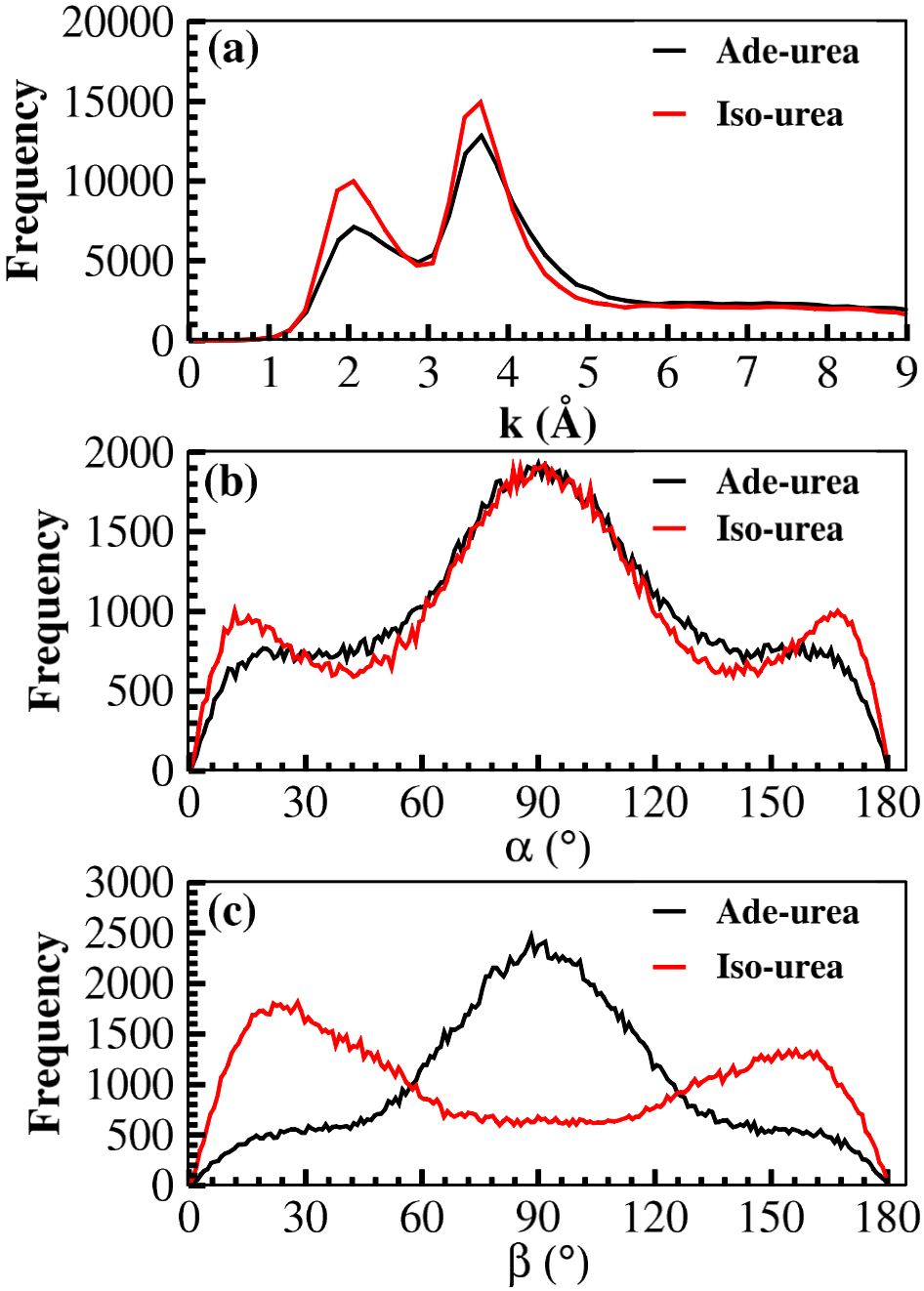

Figure 7: distribution of k, α and β for Ade/Iso with urea. (a) distribution of distance (k) between the COMs of Ade and urea, Iso and urea. (b) distribution of angle (α) between the Ade and urea, Iso and urea. (c) distribution of angle (β) between Ade and urea, Iso and urea.

Fig. 7 shows the distribution of the above distance and angles of urea with the Ade/Iso rings in the urea-water mixture. Distribution of urea molecules around the rings exhibits two peaks. We defined the stack conformation, when $k \leq 5.2$ Å and $40° \leq \{\alpha, \beta\} \leq 140°$. Rest of the structural arrangements between the urea and Ade/Iso are

defined as unstack conformation. We classified the unstack conformation into two types: (i) Un1 for 40 ° ≤ α ≤140 ° and 40 ° ≥ β ≥ 140 °. (iii) Un2 for 40 ° ≥ α ≥ 140 ° and any β value. From the above definition, we can see that Un1 denotes urea configuration where the plane of the urea molecule is perpendicular to the ring plane. Un2 denotes the configuration, where urea molecules surround the ring in the periphery.

Table 2: shows the percentage of stack and unstack conformation of urea with both ring Ade and Iso.

	Cut-off Distance (Å)	Number of urea molecule	% stack	% unstack	
				Un1	Un2
Ade	5.2	2.6	28.4%	15.0%	56.4%
Iso	5.2	3.0	39.0%	11.7%	48.5%

From the distribution of the distance of urea from both the above rings, we assign the cut-off distance at 5.2 Å for direct interaction of urea with the rings. Subsequently, we analysed the conformation of the urea in proximity of the rings from the above mentioned criteria. Table 2 shows the number of urea molecules around the Ade and Iso in both stacked and unstacked configuration. We found that on an average, 2.6 urea molecules are within 5.2 Å distance from the Ade ring and 3 urea molecules are within 5.2 Å distance from the Iso ring. Orientation of the urea molecules around the Ade and Iso rings are different as mentioned in Table 2. Urea molecules around the Iso ring form 39% stack conformation, whereas in case of Ade, 28.4% stack conformation is observed. In unstack conformation, majority of the conformation

exists in Un2 form for both the rings, and these urea molecules participate in hydrogen bond interactions with rings (see Table 1).

AI.5 Conclusion

From our analysis of the structural arrangement or urea around FAD, we observe that in the mixture, urea replaces water to form interaction with different parts of the ring. Although the overall number of hydrogen bonds between FAD and the solvent do not change from the pure water to the mixture, urea replaces some of the water molecules from both the rings to participate in stacking and hydrogen bonding interaction with Ade and Iso ring. Moreover, participation of urea around Ade and Iso is more (the values of urea around stack and unstack Aed: 3, Iso: 3.5), when FAD is in unstack conformation. This participation comes partially (~30-40%) from the stacking interaction of urea with Ade/Iso rings. This decreases the probability of FAD in the stacked configuration in presence of urea, thereby shifting the equilibrium towards more unstack conformation.

Refrence

1. Kao, Y.-T.; Saxena, C.; He, T.-F.; Guo, L.; Wang, L.; Sancar, A.; Zhong, D., Ultrafast Dynamics of Flavins in Five Redox States. *Journal of the American Chemical Society* **2008,** *130* (39), 13132-13139.
2. Radoszkowicz, L.; Huppert, D.; Nachliel, E.; Gutman, M., Sampling the Conformation Space of FAD in Water−Methanol Mixtures through Molecular Dynamics and Fluorescence Measurements. *The Journal of Physical Chemistry A* **2009,** *114* (2), 1017-1022.
3. Sancar, A., Structure and Function of DNA Photolyase and Cryptochrome Blue-Light Photoreceptors. *Chemical Reviews* **2003,** *103* (6), 2203-2238.
4. Begley, T. P., Photoenzymes: A Novel Class of Biological Catalysts. *Accounts of Chemical Research* **1994,** *27* (12), 394-401.
5. van den Berg, P. A. W.; Feenstra, K. A.; Mark, A. E.; Berendsen, H. J. C.; Visser, A. J. W. G., Dynamic Conformations of Flavin Adenine Dinucleotide: Simulated Molecular Dynamics of the Flavin Cofactor Related to the Time-Resolved Fluorescence Characteristics. *The Journal of Physical Chemistry B* **2002,** *106* (34), 8858-8869.
6. Walsh, C., Flavin coenzymes: at the crossroads of biological redox chemistry. *Accounts of Chemical Research* **1980,** *13* (5), 148-155.
7. Weber, G.; Tanaka, F.; Okamoto, B. Y.; Drickamer, H. G., The Effect of Pressure on the Molecular Complex of Isoalloxazine and Adenine. *Proceedings of the National Academy of Sciences* **1974,** *71* (4), 1264-1266.
8. Nakabayashi, T.; Islam, M. S.; Ohta, N., Fluorescence Decay Dynamics of Flavin Adenine Dinucleotide in a Mixture of Alcohol and Water in the Femtosecond and Nanosecond Time Range. *The Journal of Physical Chemistry B* **2010,** *114* (46), 15254-15260.
9. Hoccart, X.; Turrell, G., Raman spectroscopic investigation of the dynamics of urea–water complexes. *The Journal of Chemical Physics* **1993,** *99* (11), 8498-8503.
10. Finer, E. G.; Franks, F.; Tait, M. J., Nuclear magnetic resonance studies of aqueous urea solutions. *Journal of the American Chemical Society* **1972,** *94* (13), 4424-4429.
11. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, etc. *Gaussian, Inc.,Wallingford CT,* **2003**.
12. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005,** *26* (16), 1668-1688.
13. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A., Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993,** *115* (21), 9620-9631.
14. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** *25* (9), 1157-1174.
15. Sorin, E. J.; Pande, V. S., Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophysical Journal* **2005,** *88* (4), 2472-2493.
16. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.

17. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E., GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **2008,** *4* (3), 435-447.
18. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press: 2007; p 1256.
19. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.
20. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997,** *18* (12), 1463-1472.
21. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.
22. Nosé, S., A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics* **1984,** *52* (2), 255-268.
23. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981,** *52* (12), 7182-7190.

Addendum II
Controlling an Anticancer Drug Mediated G-quadruplex Formation

AII.1 Overview

The G-quadruplex (GQ) DNA structure depend on the environment. It has been known that the small molecule stabilize the GQ. Here, we have demonstrated that topotecan (TPT), a potential anticancer drug, can instigate the formation and stabilization of GQ-DNA (H24→GQ-DNA) in the absence of Na^+/K^+ ions through molecular dynamics (MD) simulation studies. Our multiple long MD simulation shows that the TPT prefer for bind to the terminal of GQ rather than the grove of GQ. Moreover, the well-tempered metadynamic also shows more binding energy for the terminal of GQ compare to grove of GQ.

AII.2 Introduction

The human telomeric DNA have the repeats of guanine rich sequences (i.e. d(TTAGGG)). This guanine rich sequences adopt a quadraplex structure in presence of ions (Na^+, K^+). The formation of GQ, stop the telomerase enzyme.[6, 7] Thus, stabilization of GQ structures by small organic molecules or ligands have been regarded as one of the potential active fields for anticancer research.[1, 8-10] Herein, we have studied the effect of TPT on the GQ stabilization.

AII.3 Simulation Methods

The structure of TPT were generated in GVIEW software[39] and optimization and electron densities were calculated quantum mechanically using HF theory with 6-31G* basis set in GAUSSIAN03 software.[39] Using this, we calculated RESP charges of TPT atoms using ANTECHAMBER module of AMBER11. For all other force-field parameters for TPT, we followed GAFF force-field[40, 41]. The topology and co-ordinates generated using AmberTools[40, 41] were converted into GROMACS format by using a Perl program amb2gmx.pl.[42]

The starting structure of hybrid G-quadruplex (3+1) was taken from protein data bank (PDB ID 2GKU). The topology and coordinates of GQ were generated by using GROMACS[43]. Amber94 force-field was used for GQ. We have created three systems: (i) GQ in water (ii) 1:2 GQ:TPT complex in water, and (iii) 1:1 GQ:TPT complex in water. Each system was solvated by ~23000 TIP3P water molecules[44] in a cubic box of dimension 90 Å. We have added physiological concentration (150 mM) of K^+ and Cl^- ions plus extra K^+ ion to neutralize the system.

All the simulations were performed using molecular dynamics software GROMACS-4.5.5.[43] Initially, we have minimized the system using the steepest descent method,[45] followed by heating it to 300 K in 100 ps using Berendsen thermostat[46] with coupling constant of 0.2 ps. We have applied restraint of 25 kcal/mol/Å^2 on heavy atoms of GQ during the heating process. The harmonic restraint was gradually reduced to 0.25 kcal/mol in six steps following the protocol used in standard DNA simulations. In each step, 100 ps equilibration was carried out at constant temperature (300 K) and pressure (1 bar) using Berendsen thermostat and barostat[46] with coupling constants of 0.2 ps each, followed by energy minimization using the steepest descent method.[45] During the simulation, LINCS algorithm[47] was used to constrain the all bonds. Particle Mesh Ewald (PME) method[48] was used for electrostatics. The distance cut-offs for the van der Waals

(vdW) and electrostatic long-range interaction was kept at 10 Å. The time step for each simulation was taken to be 2 fs. At the final equilibration step, we have performed the 2 ns unrestrained equilibration by using the Berendsen thermostat and barostat[46] with coupling constant of 0.2 ps. After the equilibration steps, we have performed multiple normal molecular dynamics simulations to see the binding of TPT with GQ. We have created three random configurations of 1:1 GQ/TPT complexes where TPT molecule was placed far away from the GQ. For each of these three different configurations, we performed three 100 ns simulations. Fig. 2a shows the initial structure and the most probable structures after the simulation. (ii) For 1:2 GQ/TPT complex, we have performed three simulations of 100 ns with different velocity distribution from a configuration where both the TPT molecules are far away from the GQ. Fig. 2b shows the initial structure for this system. Thus, we have performed a total of 1.2 μs simulation to study the TPT-GQ binding. During this final set of simulations, temperature was set at 300 K and pressure was set at 1 bar using the Nose-Hoover thermostat[49, 50] and Parrinello-Rahman barostat,[51] respectively, with 0.2 ps coupling constant.

(I) Definition of Angle Used in Analysis

Schematic representation of angle between GQ and TPT is shown in Fig. 1a. We have used two vectors: the vector $\hat{a}$ was constructed from COM of G16, G10, G4, G22 to COM of G17, G9, G3, G21 and vector $\hat{b}$ was constructed from COM of GQ to the COM of the drug. The angle of defined as, $\theta = \cos^{-1} \hat{a} . \hat{b}$. Vector $\hat{a}$ roughly shows the GQ axis.

(II) Metadynamics Simulation Details

To calculate the binding free energy of TPT and GQ, we have performed well-tempered metadynamic simulation along X (reaction coordinate). This reaction coordinate was similar to what was defined for DNA-Drug binding/unbinding mechanism by wilbee *et al.*[52] created to measure the distance between the drug and DNA perpendicular to the DNA helical axis. Here also, X approximately measures the distance of TPT perpendicular to the GQ axis denoted as vector $\hat{a}$ in Fig. 1a. Fig 1b shows the construction of the reaction coordinate X schematically. We have performed metadynamic simulations to perform two unbinding studies: (i) unbinding from terminal-bound complex. (ii) Unbinding from the groove-bound complex. To check the reproducibility of the well-tempered metadynamic simulation, we have repeated the same metadynamic simulation with different parameters. We have used Gaussian potential of 0.2 kJ/mol height and 0.7 Å width for both simulations. The potential disposition rate was 2 ps for the first simulation and 1 ps for the second simulation. The bias factor was set to 10 for both simulations. We have stopped the simulation when unbinding is happened. It took a total of 30 ns for the first and 25 ns for the second metadynamic simulation. Thus a total of 55 ns simulation was performed for the metadynamic simulation.

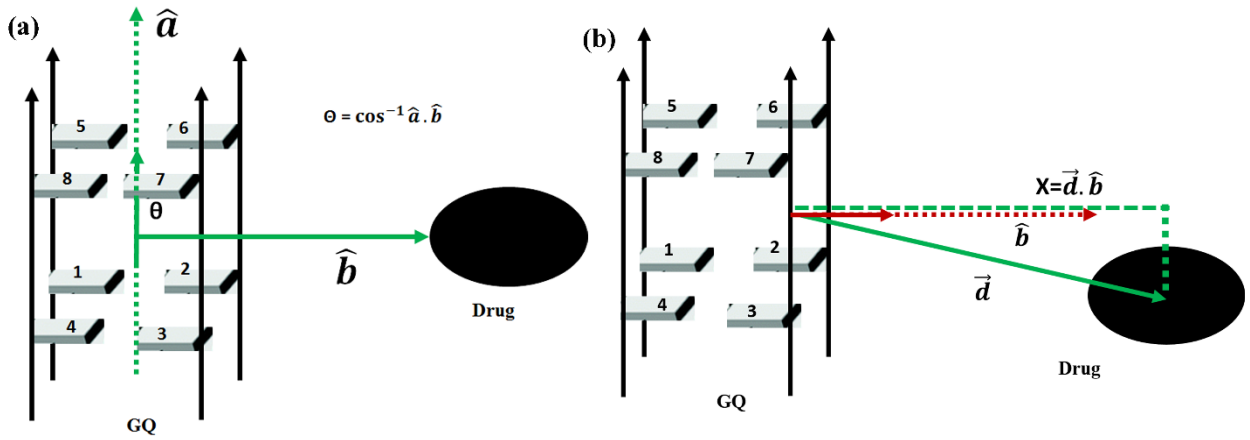

Figure 1. Definition of angle and reaction coordinate X using the schematic drawing of the GQ. (a) It is angle between two vectors $\hat{a}$ and $\hat{b}$. Vector $\hat{a}$ (GQ axis) is constructed from COM of base 1-2-3-4 to COM of base 5-6-7-8 and Vector $\hat{b}$ is constructed from COM of 1-2-3-4-5-6-7-8 to COM of drug. No. 1-8 are just for schematic representation. When drug is bound to groove of GQ (side binding), the angle would be ~90˚ and when drug bound to termini of GQ, the angle would be ~0˚ or ~180˚.(b) Definition of X (reaction coordinate) using the schematic drawing of the GQ and TPT. $X = \hat{b}.\vec{d}$. Vector $\vec{d}$ is constructed from COM of bases 2-3-6-7 to COM of drug, and vector $\hat{b}$ is constructed from COM of 2-7 to COM of 2-3-6-7 bases. Note that the numbers 1-8 are just for schematic representation and does not reflect the actual base numbering in the GQ. Same way we have defined for terminal bound state, where we have considered 1-2-3-4 bases for definition. X would be ~0, when the drug is bound to either groove or the terminal site of GQ.

AII.4 Results and Discussions

Binding mode of TPT to GQ has been explored by utilizing the molecular dynamic simulation. As discussed in the Method section, we have created two systems to explore the binding of TPT to GQ: one using same concentration of GQ and TPT (1:1 GQ:TPT complex) and other with the higher concentration of TPT. Starting from the separated configuration of GQ and TPT for both type of complexes, we performed multiple simulations with different initial conditions (see simulation detail). A total of 9 simulations were performed for 1:1 complex whereas 3 simulations were performed for 1:2 complex. The initial geometries used for the simulation are shown in Fig 1.

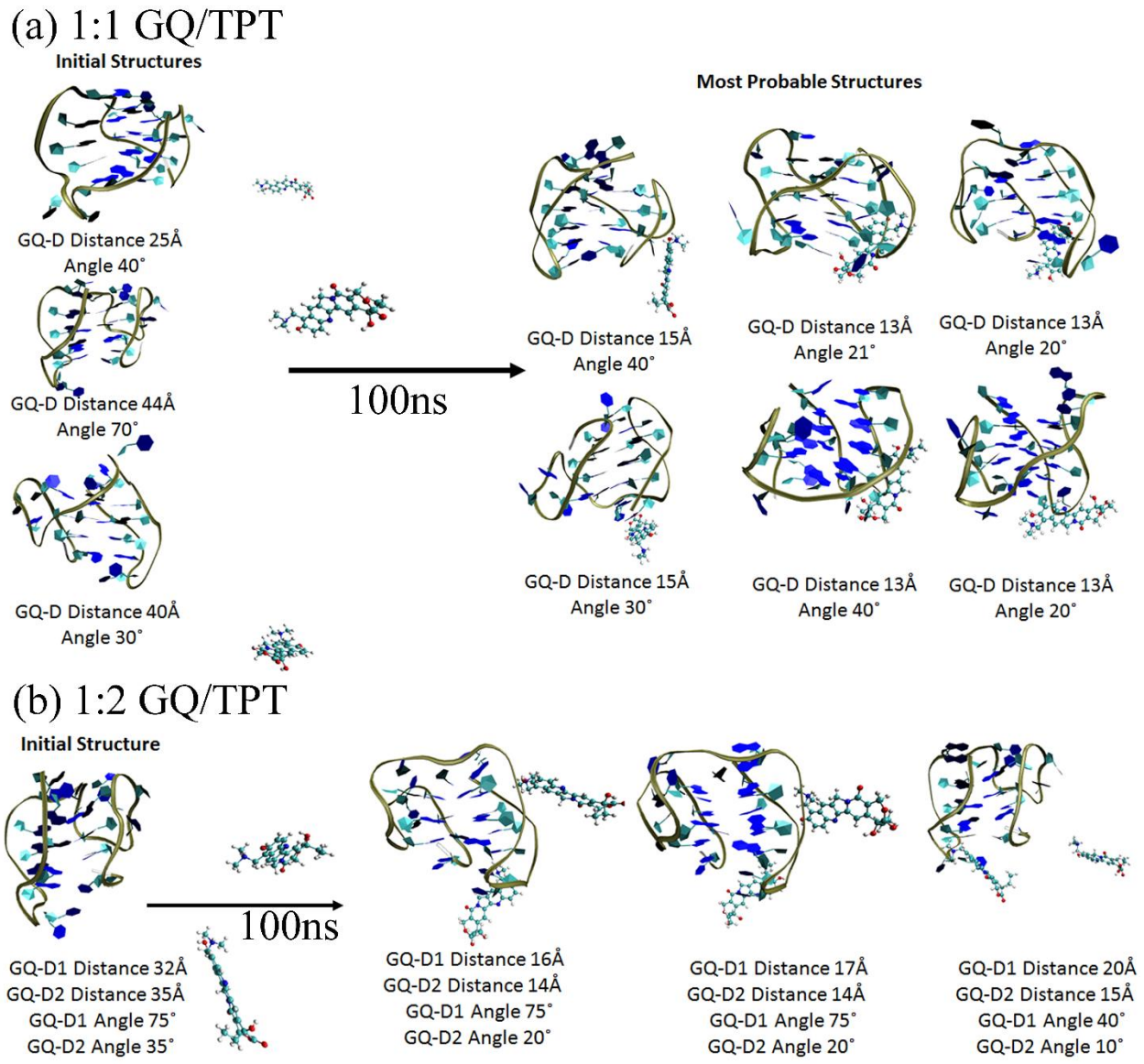

Figure. 2 Initial structure and most probable structures obtained from normal MD (a) 1:1 GQ/TPT (b) 1:2 GQ/TPT.

Analyses were carried out over the last 80 ns of each trajectory. Two binding modes of TPT are possible for 1:1 GQ-TPT system. One is a groove (side bind) binding mode where the TPT interacts with the groove of GQ and the other is the terminal binding mode where TPT molecule stacks at the termini of the GQ (Fig. 2a).

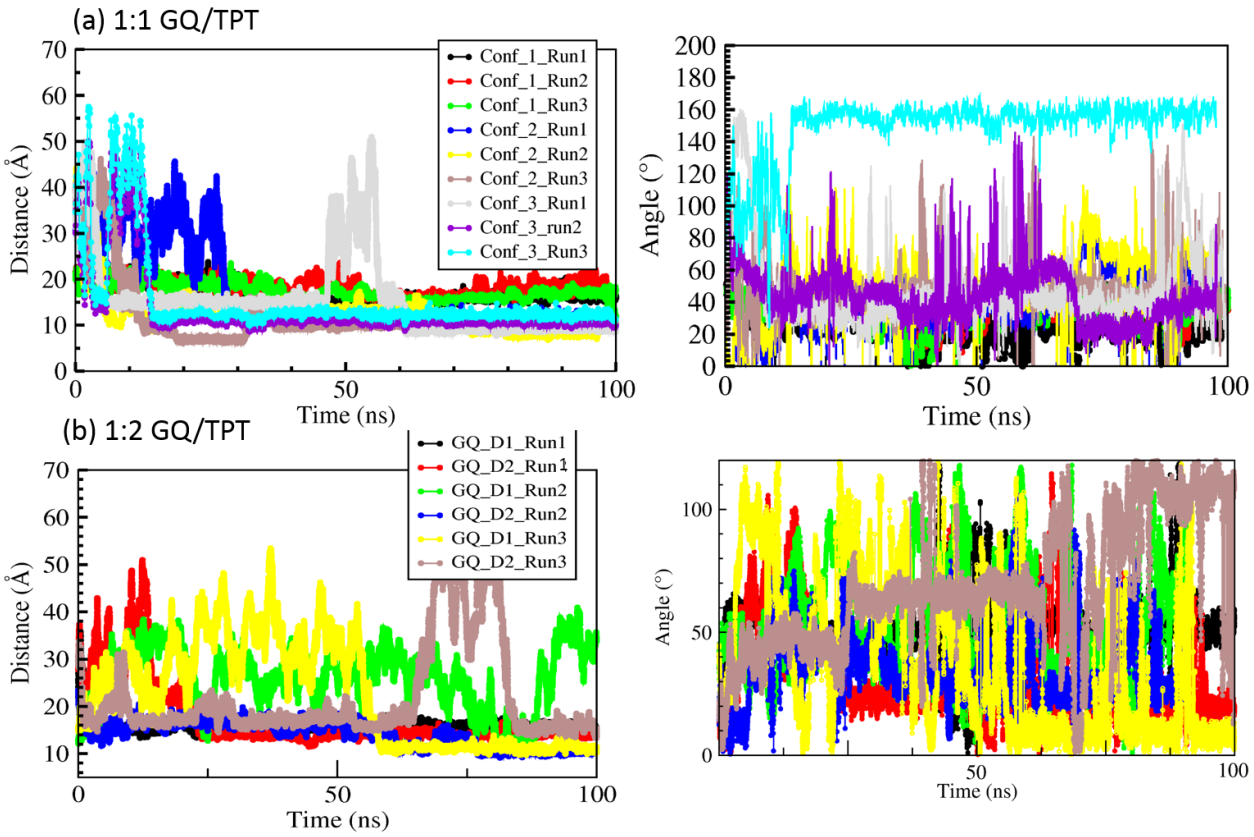

Figure 3. Variation of distance and angle with time for (a) 1:1 GQ/TPT complex (b) 1:2 GQ/TPT complex. Left column shows the variation of distances and right side column shows the change in angle. Color code is same for distance and angle in 1:1 GQ/TPT complex and 1:2 GQ/TPT complex.

The time variation of distances and angels for all the complexes (1:1 GQ/TPT and 1:2 GQ/TPT complex) are shown in Fig. 3. To characterize the preferable binding mode, we have calculated the distribution of distance (from the COM of GQ to COM of TPT) and angle (between COM of GQ to COM of TPT vector and GQ axis) for TPT around GQ for 1:1 and 1:2 GQ/TPT systems as shown in Fig. 4b and 4c. There are two peaks for 1:1 GQ/TPT complex: the first at 15 Å distance and ~30˚ angle, which represents the terminal bound state (Fig. 4b and Fig 2) and the other 14 Å distance and 40˚ angle, which is also a termini bound state but TPT is more close to the backbone of the terminal base pairs. The most probable structures for 1:1 GQ/TPT complex are shown in Fig 2. We have obtained

all possible termini bound structures. Further, very less population around 90˚ angle clearly indicates that TPT does not prefer to bind to the groove of GQ. However, for 1:2 GQ/TPT complex, the distribution of distance and angle exhibit two maxima at a distance of 13.8 Å and angle of 25˚ which represents the terminal bound state, and at the distance of 15.9 Å and angle of 79˚ representing the groove bound state (Fig. 4c and Fig 2b). Two different binding modes in case of 1:2 GQ/TPT indicates that groove-binding is the second preferable choice after the terminal position of the GQ is occupied by the first TPT.

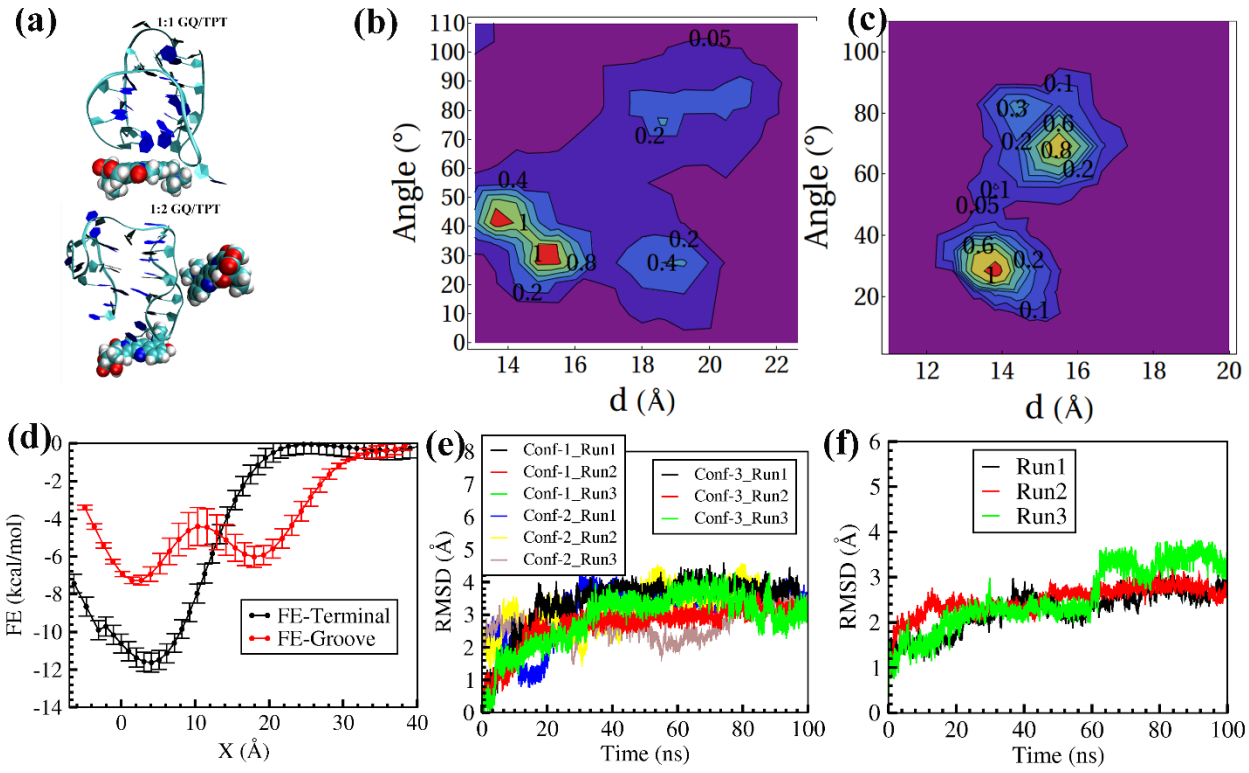

Figure 4. (a) The representative binding mode of 1:1 and 1:2 GQ/TPT complexes. Distribution of distance and angle (b) in 1:1 GQ/TPT complex and (c) in 1:2 GQ/TPT complex. (d) Average free energy (FE) of binding to the terminal (red) and groove of GQ from two independent well-tempered metadynamics simulations for 1:1 GQ/TPT complex. The RMSD of GQ (e) in 1:1 GQ/TPT complex, (f) in 1:2 GQ/TPT complex.

We have also analyzed the stabilities of 1:1 and 1:2 GQ/TPT systems in terms of their root-mean-square deviation (RMSD) (Fig. 3e and 3f) of GQ from the crystal structure geometry. Both systems show a less change in RMSD (<5Å), indicating the stability of GQ in presence of TPT. Further, for control we have simulated the GQ in the absence of TPT molecule. The higher RMSD values of GQ in absence of TPT molecule indicates that GQ is not stable in water in the absence of ions. The RMSD for this system is shown in Fig. 5. Thus, these RMSD results indicate that the binding of TPT to GQ is responsible for their GQ stability in absence of any ion in solution.

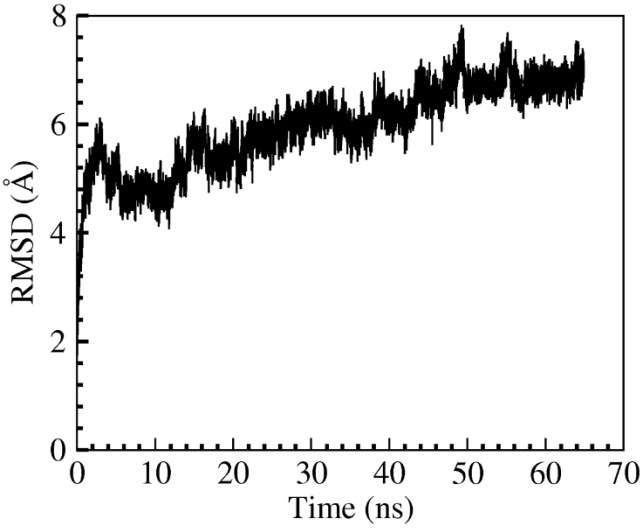

Figure 5. RMSD of GQ in water simulation.

Since TPT did not unbind from the GQ either from the terminal bound state or groove-bound state in 100 ns, the barrier for unbinding barrier for dissociation of TPT from GQ must be high and computationally extremely demanding using normal MD simulation. Therefore, we performed two independent well-tempered metadynamics simulations with different initial conditions to calculate the free energy of binding of TPT to both the terminal and groove of GQ using 1:1 GQ/TPT system. Fig. S8 in SI shows that two independent metadynamics simulations provide similar free energy profiles from both the

terminal and groove-bound systems. Fig. 4d shows that binding free energy to the terminal of GQ is 4 kcal/mol stronger than binding to the groove (Fig. 4d). This observation is consistent with the result of the normal MD simulation where more often the TPT molecule bound to the terminal than groove of GQ starting with the separated state for 1:1 GQ/TPT complex and terminal bound state was followed by groove-bound state in case 1:2 complex.

To analyze further the reason for the preference of TPT towards the terminal site of GQ, we calculated the total binding energy between the two species and also decomposed into both van der Waals (dispersion interaction) and electrostatic components.

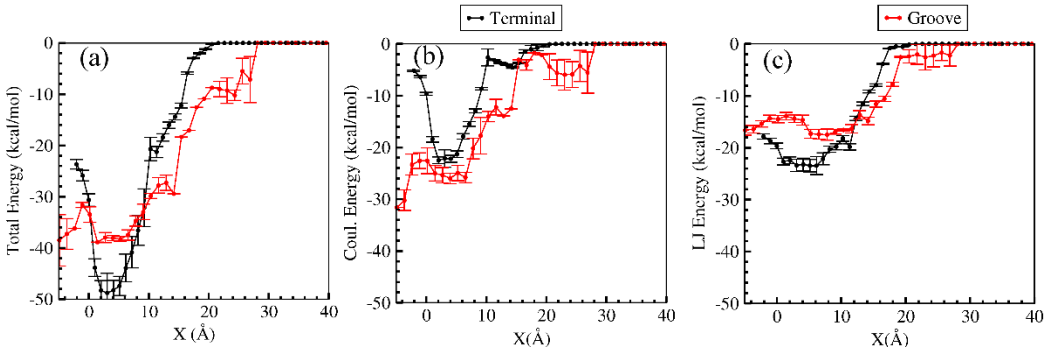

Figure 6. Interaction energy between TPT and GQ along unbinding pathways for terminal (black) and groove-bound (red) states. (a) Total interaction energy between TPT and GQ and its decomposition into (b) electrostatic energy (Coulomb interaction) and (c) van der Waals components. The interaction energy is averaged over two metadynamics simulations. Error bars are shown.

Fig. 6 shows the average total energy obtained from the two metadynamics simulation for both the terminal and groove bound state for 1:1 complex and their decomposition into dispersive and electrostatic interaction. We can see that for the terminal bound state has stronger dispersive interaction compared to the groove-bound state (Fig 6b) which, however, has a stronger electrostatic interaction than the terminal bound state (Fig 6c). Although both the states are stabilized by both van der Waals and electrostatic energy components, the stabilization stemming from the dispersive interaction due to stacking of

TPT on GQ outweighs the electrostatic stabilization for the groove-bound state such that the total energy is more favorable for the terminal bound state (Fig. 6).

AII.5 Conclusion

We have shown here that topotecan can induce the GQ formation and stabilization even in the absence of Na^+/K^+ ions. MD simulation studies show that TPT preferably binds to the terminal of GQ, and the second preferable binding mode is the groove of GQ. The more rigorous free energy calculation using well-tempered metadynamics shows that binding to the terminal is 4 kcal/mol stronger than the free energy of binding to the groove. Although both the electrostatic and van der Waals interaction contribute to both these binding mode, the former is prevalent for the terminal binding whereas the latter is stronger for the groove-binding. The overall total energy is, however, more negative for the terminal binding mode.

References

1. S. Balasubramanian, L. H. Hurley and S. Neidle, *Nat Rev Drug Discov*, 2011, **10**, 261-275.
2. A. M. Zahler, J. R. Williamson, T. R. Cech and D. M. Prescott, *Nature*, 1991, **350**, 718-720.
3. T.-m. Ou, Y.-j. Lu, J.-h. Tan, Z.-s. Huang, K.-Y. Wong and L.-q. Gu, *ChemMedChem*, 2008, **3**, 690-713.
4. R. E. Verdun and J. Karlseder, *Nature*, 2007, **447**, 924-931.
5. N. W. Kim, M. A. Piatyszek, K. R. Prowse, C. B. Harley, M. D. West, P. L. Ho, G. M. Coviello, W. E. Wright, S. L. Weinrich and J. W. Shay, *Science*, 1994, **266**, 2011-2015.
6. J. L. Huppert, *Chem. Soc. Rev.*, 2008, **37**, 1375-1384.
7. G. N. Parkinson, M. P. H. Lee and S. Neidle, *Nature*, 2002, **417**, 876-880.
8. A. M. Burger, F. Dai, C. M. Schultes, A. P. Reszka, M. J. Moore, J. A. Double and S. Neidle, *Cancer Res.*, 2005, **65**, 1489-1496.
9. J. Mohanty, N. Barooah, V. Dhamodharan, S. Harikrishna, P. I. Pradeepkumar and A. C. Bhasikuttan, *J. Am. Chem. Soc.*, 2013, **135**, 367-376.
10. R. Rodriguez, G. D. Pantoş, D. P. N. Gonçalves, J. K. M. Sanders and S. Balasubramanian, *Angew. Chem. Int. Ed.*, 2007, **46**, 5405-5407.
11. O. Mendoza, A. Bourdoncle, J.-B. Boulé, R. M. Brosh and J.-L. Mergny, *Nucleic Acids Res.*, 2016, DOI: 10.1093/nar/gkw079.
12. K. Paeschke, M. L. Bochman, P. D. Garcia, P. Cejka, K. L. Friedman, S. C. Kowalczykowski and V. A. Zakian, *Nature*, 2013, **497**, 458-462.
13. Y. Wu, K. Shin-ya and R. M. Brosh, *Mol. Cell. Biol.*, 2008, **28**, 4116-4128.
14. S. Ogasawara and M. Maeda, *Angew. Chem. Int. Ed.*, 2009, **48**, 6671-6674.
15. D. Monchaud, P. Yang, L. Lacroix, M.-P. Teulade-Fichou and J.-L. Mergny, *Angew. Chem. Int. Ed.*, 2008, **47**, 4858-4861.
16. Y. Shi, H. Sun, J. Xiang, H. Chen, S. Zhang, A. Guan, Q. Li, S. Xu and Y. Tang, *Chem. Commun.*, 2016, **52**, 7302-7305.
17. J.-L. Yao, X. Gao, W. Sun, X.-Z. Fan, S. Shi and T.-M. Yao, *Inorganic Chemistry*, 2012, **51**, 12591-12593.
18. F. Li, Y. Xu, H. Li, C. Wang, A. Lu and S. Sun, *New J. Chem.*, 2014, **38**, 1396-1400.
19. P. Kundu, S. Ghosh, S. Das and N. Chattopadhyay, *Phys. Chem. Chem. Phys.*, 2016, **18**, 3685-3693.
20. D. V. Berdnikova, T. M. Aliyeu, T. Paululat, Y. V. Fedorov, O. A. Fedorova and H. Ihmels, *Chem. Commun.*, 2015, **51**, 4906-4909.
21. X. Yang, D. He, J. Cao, X. He, K. Wang and Z. Zou, *RSC Adv.*, 2015, **5**, 84553-84559.
22. P. Galer, B. Wang, P. Šket and J. Plavec, *Angew. Chem. Int. Ed.*, 2016, **55**, 1993-1997.
23. C. Riccardi, I. Russo Krauss, D. Musumeci, F. Morvan, A. Meyer, J.-J. Vasseur, L. Paduano and D. Montesarchio, *ACS Appl. Mater. Interfaces*, 2017, **9**, 35574-35587.
24. D. M. Engelhard, R. Pievo and G. H. Clever, *Angew. Chem. Int. Ed.*, 2013, **52**, 12843-12847.
25. T. Tian, Y. Song, L. Wei, J. Wang, B. Fu, Z. He, X.-R. Yang, F. Wu, G. Xu, S.-M. Liu, C. Li, S. Wang and X. Zhou, *Nucleic Acids Res.*, 2017, **45**, 2283-2293.
26. D. Mirchandani, H. Hochster, A. Hamilton, L. Liebes, H. Yee, J. P. Curtin, S. Lee, J. Sorich, C. Dellenbaugh and F. M. Muggia, *Am. Assoc. Cancer Res.*, 2005, **11**, 5912-5919.

27. Y. Pommier, *Nat. Rev. Cancer*, 2006, **6**, 789-802.
28. C. Leonetti, M. Scarsella, G. Riggio, A. Rizzo, E. Salvati, M. Incalci, L. Staszewsky, R. Frapolli, M. F. Stevens, A. Stoppacciaro, M. Mottolese, B. Antoniani, E. Gilson, G. Zupi and A. Biroccio, *Clin. Cancer Res.*, 2008, **14**, 7284.
29. F. Li, J. Zhou, M. Xu and G. Yuan, *Int. J. Biol. Macromol.*, 2017, DOI: https://doi.org/10.1016/j.ijbiomac.2017.10.010.
30. K. Gavvala, A. Sengupta, R. K. Koninti and P. Hazra, *ChemPhysChem*, 2013, **14**, 3375-3383.
31. M. R. di Nunzio, Y. Wang and A. Douhal, *J. Photochem. Photobiol., A: Chem*, 2013, **266**, 12-21.
32. R. K. Koninti, K. Gavvala, A. Sengupta and P. Hazra, *J. Phys. Chem. B*, 2015, **119**, 2363-2371.
33. H. Joshi, A. Sengupta, K. Gavvala and P. Hazra, *RSC Adv.*, 2014, **4**, 1015-1024.
34. S. A. Strel'tsov, A. L. Mikheikin, S. L. Grokhovsky, V. A. Oleinikov and A. L. Zhuze, *Mol. Biol.*, 2002, **36**, 400-411.
35. S. Streltsov, V. Oleinikov, M. Ermishov, K. Mochalov, A. Sukhanova, Y. Nechipurenko, S. Grokhovsky, A. Zhuze, M. Pluot and I. Nabiev, *Biopolymers*, 2003, **72**, 442-454.
36. W. J. M. Underberg, R. M. J. Goossen, B. R. Smith and J. H. Beijnen, *J. Pharm. Biomed. Anal.*, 1990, **8**, 681-683.
37. M. R. di Nunzio, Y. Wang and A. Douhal, *J. Phys. Chem. B*, 2012, **116**, 7522-7530.
38. M. R. di Nunzio, Y. Wang and A. Douhal, *J. Phys. Chem. B*, 2012, **116**, 8182-8190.
39. M. J. Frisch, G. W. Trucks, H. B. Schlegel, etc., *Journal*, 2016.
40. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *J. Comput. Chem.*, 2005, **26**, 1668-1688.
41. W. D. Cornell, P. Cieplak, C. I. Bayly and P. A. Kollmann, *J. Am. Chem. Soc.*, 1993, **115**, 9620-9631.
42. E. J. Sorin and V. S. Pande, *Biophys. J.*, 2005, **88**, 2472-2493.
43. B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435-447.
44. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157-1174.
45. W. H. T. Press, S. A.; Vetterling, W. T.; Flannery, B. P. , *Numerical Recipes in FORTRAN; The Art of Scientific Computing*, Cambridge University Press, New York, 3rd edition edn., 1992.
46. H. J. C. Berendsen, J. P. M. Postma, W. F. v. Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684-3690.
47. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463-1472.
48. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089-10092.
49. S. Nosé, *Mol. Phys.*, 1984, **52**, 255-268.
50. W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695-1697.
51. M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182-7190.
52. W. D. Sasikala and A. Mukherjee, *J. Phys. Chem. B*, 2012, **116**, 12208-12212.
53. E. I. Cucolea, C. Tablet, H.-J. Buschmann and L. Mutihac, *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, 2015, **83**, 103-110.

54. Y. Huang, S.-F. Xue, Q.-J. Zhu and T. Zhu, *Supramolecular Chemistry*, 2008, **20**, 279-287.
55. N. S. Venkataramanan and A. Suvitha, *J. Phys. Chem. B*, 2017, **121**, 4733-4744.
56. K. Hyz, R. Kawęcki, E. Bednarek, W. Bocian, J. Sitkowski and L. Kozerski, *Magn. Reson. Chem.*, 2010, **48**, 575-584.