

**THEORETICAL STUDY OF B-DNA TO A-DNA
TRANSITION: FROM LOCAL DINUCLEOTIDE
STEP TO GLOBAL HELICAL CONVERSION**

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS

OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

MANDAR KULKARNI

20113112



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF SCIENCE EDUCATION AND
RESEARCH, PUNE – 411 008**

Dedicated to
Wonderful Teachers in my life
Specially
Dr. Sujata Kale,
Prof. G. C. Kulkarni
&
Dr. Arnab Mukherjee



भारतीय बिज्ञान शिक्षा ँबं 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, “*Theoretical Study Of B-DNA to A-DNA Transition: From Local Dinucleotide Step to Global Helical Conversion*” submitted by Mandar V. Kulkarni 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: 8th April 2017

Pune (MH), India

A handwritten signature in black ink, reading 'Arnab Mukherjee'.

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: 8th April 2017
Pune (MH), India



Mandar Kulkarni
Reg: No: 20113112

ACKNOWLEDGMENTS

I acknowledge Indian Institute of Science Education and Research, Pune for providing high quality research environment and graduate fellowship. I am grateful to Prof. K. N. Ganesh, Director, IISER Pune. I would like to thank my doctoral advisor, Dr. Arnab Mukherjee for his valuable guidance and most importantly for teaching a balanced and scientific way of thinking. I am really thankful to him for providing a healthy work environment, necessary freedom and multiple encouraging and enlightening discussions during my Ph.D. life. Also, I would like to thank my Research advisory Committee (RAC) members Dr. S. G. Srivatsan and Dr. Suman Chakrabarty for their helpful suggestions during annual evaluations. I am especially indebted to Dr. Suman Chakrabarty for the valuable and insightful discussions while exploring nucleation driven B-DNA $\rightarrow$ A-DNA transition in long DNA helices. Also, I am grateful to Dr. Srivatsan for his continuous encouragement to understand DNA biochemistry from the perspective of practical applications.

I would like to thank Dr. Pinaki Talukadar and Dr. Partha Hazra for providing an opportunity to work with them during collaborative projects, where I learn how computational chemistry can be useful to provide explanation for molecular interactions. I am also thankful to Sopan Shinde and Sagar Satpathi for helpful and insightful discussions during collaborative projects. I am especially indebted to Dr. Partha Hazra and Sagar Satpahti for providing experimental support for a theoretical model developed to predict A-DNA or B-DNA conformation.

Also, I consider myself fortunate enough for having an opportunity to interact with Dr. Arun Venkatnathan, Dr. Srabanti Chaudhary and Dr. Anirban Hazra during my Ph.D. studies, allowing me to understand the various aspects of computational chemistry and Ph.D. life as well.

I thank my labmates Wilbee, Reman, Debasis, Hridya, Pallavi, Amitosh, Darnesh for the stimulating discussions, and for all the fun we have had in the last five years. Also I am thankful to my friends especially Pravin, Rakesh, Avdhoot, Mahesh, Prabhat, Meghna, Bappa, Subbu, Niharika with whom I have discussed innumerable things from science to friendly chitchats.

I would like to thank my lovely wife Ramya for her strong support and faith in me. The enormous discussions with her always make me feel that it's good to have a

life partner with same career goals. I am thankful to my friend Abhijit Pardeshi and his family for all the encouragement and constant faith in me.

I would like to extend my thanks specially to Mr. Pratap Marathe and family, Dr. Abhay Sontakke and Mrs. Sangeeta Sontakke, Dr. Sujata Kale ma'am (Garware College), Prof. G. C. Kulkarni (N. Wadia College), and Prof. Arvind Natu Sir for their continuous encouragement at various points of my life.

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

CONTENTS

Declaration	I
Acknowledgement	II
List of Publications	VIII
Abstract	IX

1. Introduction

1.1 The Basic DNA Framework	1
1.2 A-DNA and B-DNA	5
1.3 Structural features of A-DNA and B-DNA	6
1.4 Biological significance of “dry” A-DNA in cellular environment	9
1.5 The Local B $\rightarrow$ A DNA transition: Conformational Preferences and Consequences	10
1.6 The Global B-A transition: Observations and Cooperative transition	14
1.7 Outline of the thesis	17

2. Methodology

2.1 Molecular Dynamics simulations	22
2.2 Umbrella sampling simulations	23
2.3 Reaction coordinate	25
2.4 Analysis Procedures	30
2.4.1 Cluster analysis	30
2.4.2. Structure analysis	30
2.4.3. Thermodynamic analysis	31
2.5 Summary	32

3. Sequence Dependent Free Energy Profiles of Localized B- to A-Form Transition of DNA in Water

3.1. Introduction	34
3.2 Computational details	35
3.2.1. Construction of the initial configuration	35
3.2.2 Force-field and equilibration	36
3.2.3. Umbrella Sampling Simulation Procedures	36

3.3 Results and Discussion	37
3.3.1. Free energy, Energy and entropy	37
3.3.2. Energetic Analysis and Entropy interpretation	39
3.3.3. Structure analysis and mechanism	43
3.4 Comparison with experiment	46
3.5 Conclusion	48
4. B-form /A-form Junction Free Energy in DNA	
4.1 Introduction	52
4.1.1 B/A junction and its biological significance	52
4.1.2 Contribution of junction energy in $\Delta G(Z'_p)$	53
4.2 Computational Details	54
4.2.1 Reaction coordinate: Z'_p and $\overline{Z'_p}$	54
4.2.2 The design of DNA sequences.	54
4.2.3 Simulation Details	57
4.3 Results and Discussion	58
4.3.1 Free energy profiles along Z'_p and $\overline{Z'_p}$.	58
4.3.2. Absolute and junction energies of homo-nucleotide steps	59
4.3.3 Absolute free energies of hetero dinucleotide steps	61
4.3.4 B-philicity of AA/TT step.	62
4.3.5 The role of solvation energy and stacking interactions in B-A transition	64
4.4 Conclusions	64
5. Prediction of A-DNA or B-DNA Conformational Preference of Sequence: Application to Crystal Structures and <i>Xenopus tropicalis</i> Genome	
5.1 Introduction	68
5.2 Material and Methods	
5.2.1 Brief Details of Previous Free energy calculations	70
5.2.2 Predicting A-DNA formation propensity	71
5.2.3 Prediction of A-DNA promoter elements in <i>Xenopus</i> genome	72
5.2.4 CD spectroscopy of potential A-DNA promoter elements	73
5.3 Results and Discussion	74
5.3.1. Prediction of a Crystal Structures as DNA sequence.	74
5.3.2. Comparison with Previous studies.	76

5.3.3 <i>Xenopus</i> Genome analysis and observed predictions	77
5.4. Conclusion	79
6. B-DNA to A-DNA Transition in Aqueous Ethanol Solution: Evidence of Multistep Non-classical Nucleation Pathway	
6.1 Introduction	82
6.1.1. Hypothesis of nucleation driven B→A transition	82
6.1.2. Classical Nucleation theory and its limitations	82
6.1.3. Experimental evidence of nucleation driven B→A transition	84
6.2. Methodology	
6.2.1 System Preparation	85
6.2.2 Choice of force field	85
6.2.3 Equilibration	86
6.2.4 Order Parameter to study the nucleation	86
6.2.5 Definition of product state <i>i.e.</i> A-DNA in solution	87
6.2.6 Trajectory Path Sampling Simulations	88
6.2.7 Attachment/detachment rate calculations	89
6.3. Results and Discussion	
6.3.1 Comment on cooperative length	90
6.3.2 Transition path sampling simulations.	91
6.3.3 Attachment/ detachment rate at critical nucleus size	94
6.3.4. Variance of largest cluster size near critical N_{Large} values	95
6.3.5. Diffusion at critical nucleus in poly(AT) and poly(GG)	97
6.4. Conclusion	98
7. Summary and Future Directions	
7.1. Summary	102
7.2. Future Directions:	104
7.2.1. 2'-5' linkage in DNA, A-form and pre-biotic RNA world	104
7.2.2. DNA nanoworm based on Scrunchworm hypothesis	108
8. DNA- Ionic Liquid interactions	
Addendum I : Ionic Liquid Induced Stabilization of G-quadruplex	114
Addendum II: Ionic Liquid Prolongs DNA Translocation through Graphene Nanopore	126

Annexure 3	156
Annexure 4	169
Annexure 5	173

List of Publications

- 1) **Kulkarni, Mandar** & Mukherjee, Arnab, Sequence dependent free energy profiles of localized B- to A-form transition of DNA in water. *J Chem Phys*, **2013**, 139, 155102.
- 2) **Kulkarni, Mandar** & Mukherjee, Arnab, Computational Approach to Explore the B/A Junction Free Energy in DNA. *Chemphyschem*, **2016**, 17, 147-54.
- 3) Shinde, Sopan, **Kulkarni, Mandar** & Talukdar, Pinaki, Helical supramolecular organization of a 1,2-diol appended naphthalene diimide organogelator via an extended intermolecular H-bonding network. *RSC Advances*, **2016**, 6, 30690-30694.
- 4) **Kulkarni, Mandar*** & Mukherjee, Arnab*, Ionic Liquid Prolongs DNA Translocation through Graphene Nanopore. *RSC Advances*, **2016**, 6, 46019 – 46029.
- 5) Satpathi, Sagar, **Kulkarni, Mandar**, Mukherjee, Arnab & Hazra, Partha, Ionic liquid induced G-quadruplex formation and stabilization: spectroscopic and simulation studies *Phys. Chem. Chem. Phys.*, **2016**, 18, 29740-29746.
- 6) Saha, Debasis*, **Kulkarni, Mandar** & Mukherjee, Arnab* Water modulates the ultraslow dynamics of hydrated ionic liquids near CG rich DNA: consequences for DNA stability *Phys. Chem. Chem. Phys.*, **2016**, 18, 32107-32115.
- 7) **Kulkarni, Mandar** and Arnab Mukherjee, The B-DNA to A-DNA transition in right-handed DNA helix: Understanding a Local to Global Mechanism (*An invited review article submitted*)
- 8) **Kulkarni, Mandar** & Mukherjee, Arnab, A method to predict conformational preference for right-handed DNA conformation. (*Manuscript under preparation*).
- 9) **Kulkarni, Mandar**, Chakrabarty, Suman and Mukherjee, Arnab, Nucleation driven intramolecular B-DNA to A-DNA transition in poly(AT) and poly(GG) DNA: A Non-classical Nucleation Perspective. (*Manuscript under preparation*).

ABSTRACT

The biological polyanionic polymer, DNA adopts canonical B-form under physiological conditions. However, it adopts a short and compact A-DNA form under dehydrating conditions. The proteins such as polymerases, endonucleases, etc. often interact with a small segment of DNA causing local A-DNA formation. The effective mechanism of B-DNA to A-DNA transition is sequence dependent and arises from multiple dinucleotide steps interacting with each other. We have employed enhanced sampling simulations to explore the thermodynamics and mechanistic details of 10 unique dinucleotide steps undergoing $B \rightarrow A$ transition in aqueous environment. In our work, the heterogeneous $B \rightarrow A$ transition mechanism of dinucleotide steps and role of solvent portrayed sequence dependency of the transition. The study is further extended to calculate the B-form/A-form interface or junction free energy and eventually, the “absolute” free energy values, which are the free energy estimates of $B \rightarrow A$ transition of a dinucleotide step uninfluenced by the flanking base pairs. A predictive model for the conformational preference of a nucleotide sequence towards A- or B-form is developed based on the absolute free energy values for 10 unique dinucleotide steps. There are only a few models existing with such predictive capability and our model shows better accuracy than the existing ones. We applied the model successfully to discover unknown A-DNA promoter sequences (APS) in genome. The reliability of our proposed method is established through experimental verification. Finally, we studied the global (whole DNA) $B \rightarrow A$ DNA transition of a single DNA helix in solution which showed the existence of a multistep non-classical nucleation mechanism for $B \rightarrow A$ transition for the first time.

Chapter 1

Introduction

1.1. The Basic DNA Framework

The basic building block of DNA- a nucleotide unit is composed of three major chemical fragments mainly a heteroaromatic nitrogenous base, a deoxyribose sugar molecule and a phosphate group. The nitrogenous bases are categorized into purine bases (adenine (dA) and guanine (dG)) and pyrimidine bases (cytosine (dC) and thymine (dT) See Fig. 1.1). These bases interact with each other via non-covalent interactions such as hydrogen bonding, stacking interactions to form a helical DNA structure.

In general, adenine forms 2 hydrogen bonds with thymine and guanine forms 3 hydrogen bonds with cytosine as shown in Fig. 1.1, thus forming rungs of the helical DNA ladder. The exocyclic amine group and –NH ring group adjacent to carbonyl group acts as a hydrogen bond donor and the exocyclic carbonyl oxygen atoms and ring nitrogen atoms act as hydrogen bond acceptors. These standard hydrogen bonding pattern of A:T and G:C base pairs are termed as Watson-Crick base pairs. The hydrogen bonding stabilization is expected to be -0.25 kcal/mol per base pair [1-4] and is cooperative in nature when multiple hydrogen bonds are considered.

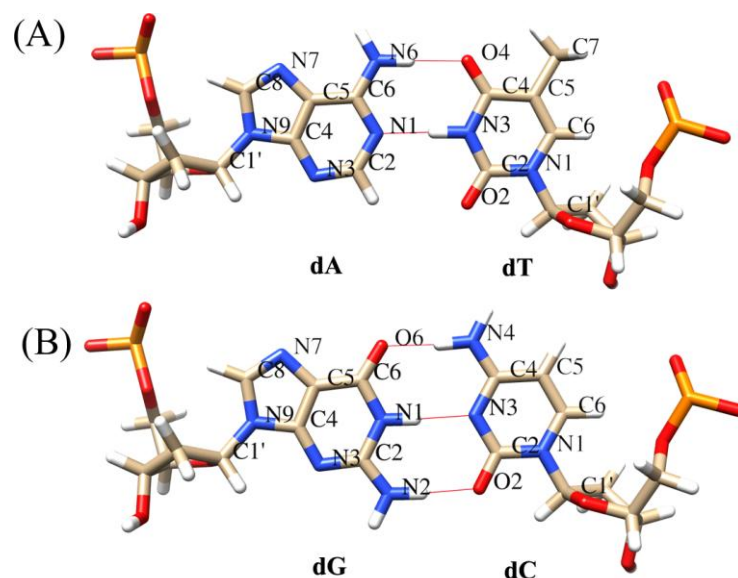


Figure 1.1. Figure showing specific atom types and Watson-Crick hydrogen bonding for (a) A:T base pair and (b) G:C base pair.

The aromatic base pairs have a tendency to stack over each other albeit in offset position leading to a conventional DNA helical geometry. The stacking arrangement of base pairs arises due to dispersion interactions between them as well as constraints imposed by sugar-phosphate backbone. The sugar-phosphate backbone holds the DNA framework stable. The flexible and hydrophilic sugar-phosphate backbone can potentially exhibit multiple conformations. For consecutive bases $i-1$, i , and $i+1$, these conformations can be observed using six backbone torsional angles $O3'-[i-1]-P-O5'-C5'$ (α), $P-O5'-C5'-C4'$ (β), $O5'-C5'-C4'-C3'$ (γ), $C5'-C4'-C3'-O3'$ (δ), $C4'-C3'-O3'-P[i+1]$ (ϵ), $C3'-O3'-P[i+1]-O5'[i+1]$ (ζ) and a dihedral angle χ around the glycosidic bond ($N1/N9-C1'$). These angles are shown in Fig. 1.2(a). For purine and pyrimidine bases, dihedral angle χ is defined as $O4'-C1'-N9-C4$ and $O4'-C1'-N1-C2$, respectively.

The five membered furanose ring (deoxyribose ring) is not flat and one of the $C2'$ or $C3'$ atom always remain out-of-plane defined by rest of the 4 atoms, leading to $C2'$ -endo and $C3'$ -endo puckering, respectively (See Fig. 1.2(b)). Similarly, sometimes other sugar ring atoms such as $C4'$ ($C4'$ -exo) or $C1'$ ($C1'$ -exo) adopt out-of-plane position leading to diverse conformational map for B-DNA conformation. This sugar puckering reduces the steric clashes between nucleobases and the sugar ring. Apart from the backbone torsional angles mentioned above, five torsional angles ν_0 , ν_1 , ν_2 , ν_3 , and ν_4 , each around five bonds of deoxyribose sugar are useful to

understand the sugar pucker configurations in terms of Phase angle and amplitude. These torsion angles are shown in Fig. 1.2(b). These values are useful to calculate the Phase angle (P) and its amplitude (τ_m)[5] to distinguish between multiple conformations of sugar molecule. The phase (P) and amplitude (τ_m) are defined as,

$$P = \frac{(\nu_2 + \nu_4) - (\nu_1 + \nu_3)}{2\nu_0(\sin 36^\circ + \sin 72^\circ)} \quad \text{Eq. 1.1}$$

$$\tau_m = \frac{\nu_2}{\cos P} \quad \text{Eq. 1.2}$$

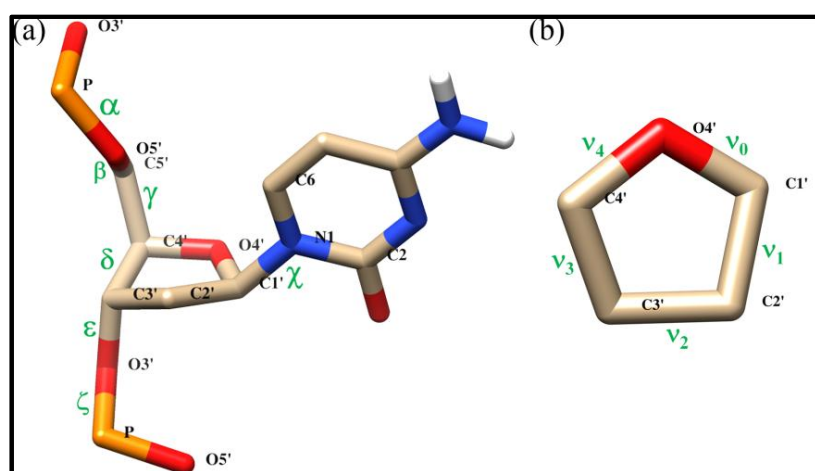


Figure 1.2. Figure showing various (a) backbone torsion angles and (b) sugar torsional angles.

The glycosidic bond (N1/N9-C1') which connects a hydrophobic base to sugar plays an important role in defining conformational map of a certain nucleotide. The purine bases Adenine and Guanine can display both syn [6-8] ($50^\circ \leq \chi \leq 80^\circ$) and anti ($180^\circ \leq \chi \leq 280^\circ$) conformations. However, the pyrimidine bases Cytosine and Thymine exhibit only syn conformations due to steric clashes between base and sugar moiety. Thus, in a DNA helix, these various interactions such as hydrogen bonding, stacking interactions, steric repulsions and environmental parameters such as backbone-ions and backbone-water interactions eventually decide the fate of the DNA conformation.

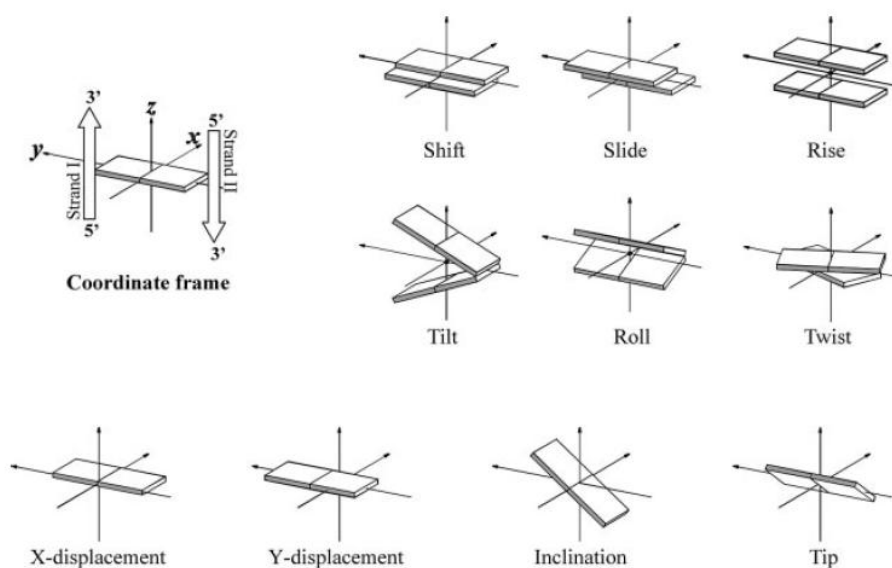


Figure 1.3. Figure showing the structural parameters of a dinucleotide step based on 3 translational motions (Shift, Slide and Rise) and 3 rotational motions (Tilt, Roll and Twist) according to block representation as proposed by Lu and Olson[9]. The figure is reprinted with permission from Ref. [10].

The bases Adenine-Thymine and Guanine-Cytosine interact via strong noncovalent hydrogen bonding interactions, thus forming a base pair. The base pair parameters such as X-displacement, Inclination display distinguishing values[9] for A-form (X-displacement = $-4.81 \text{ \AA}$, Inclination = 19.8°) and B-form (X-displacement = $0 \text{ \AA}$, Inclination = 0°). The DNA conformation depends on the various interactions present between consecutive base pairs and hence in the context of understanding DNA conformational map, a pair of base pair i.e. a dinucleotide step or a base-pair step represents the smallest possible unit. The chemical heterogeneity of each base step impose the minuscule variations in the structural properties of dinucleotide steps. These variations are recognizable using base-pair step parameters based on three translational parameters (Slide, Rise, Shift) and three rotational parameters (Roll, Twist, Tilt) as shown in Fig. 1.3. The knowledge of backbone torsional angles and sugar pucker along with these step parameters help to understand a complete conformational map of a dinucleotide step. Such meticulous information helps to anatomize a conformational diversity of whole DNA molecule into its corresponding individual dinucleotide steps, thus providing greater details to establish structure-function relationship.

1.2. A-DNA and B-DNA

The negatively charged exterior hydrophilic surface of DNA is sensitive to the subtle changes in the surrounding environment. The DNA responds via suitable conformational variations to the environmental changes such as reduced water activity, the concentration of ions, and the presence of small molecules or proteins. The reduction in relative humidity (RH) below 75% induces A-DNA conformation, whereas under physiological and at high RH conditions DNA exhibits B-DNA conformation. Under dehydrating conditions, the reduction of water molecules around DNA induces protrusion of sugar-phosphate backbone. This allows DNA to remain solvated in less amount of water by bridging water molecules with the phosphate backbone of DNA through hydrogen bonding, as proposed by “economy of hydration” principle[11], which is a typical characteristic of A-DNA.

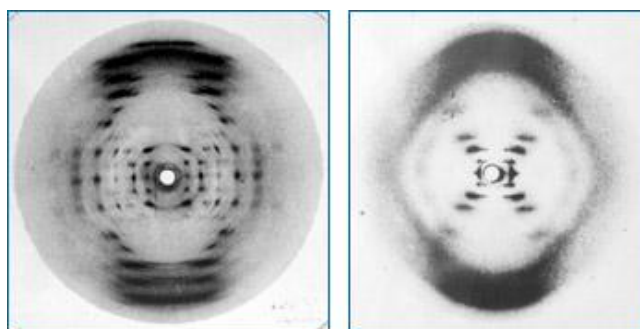


Figure 1.4. X-ray diffraction patterns for A-DNA (Left) and B-DNA (Right) studied by Franklin and co-workers. (Source: Adapted from http://undsci.berkeley.edu/article/0_0_0/dna_06)

In 1953, Franklin and Gosling[12] coined the terms “A-DNA” and “B-DNA” while studying “dry” and “wet” forms of DNA crystals, respectively. The X-ray diffraction patterns of A-DNA and B-DNA are shown in Fig. 1.4, where ‘X’ shape pattern (Fig. 1.4 (right)) is due to helicity of B-DNA and each spot on this ‘X’ pattern belongs to each helix and the separation between each spot is 34 Å. The X-ray crystallographic studies by Franklin and Gosling observed the presence of “biological”[13] water molecules clinging to the outer sugar-phosphate backbone in A-DNA and B-DNA. This seminal work laid the foundation for the important aspects of DNA structure, mainly noting that (a) phosphate backbone (including sugars) lies on the outside of DNA structure, (b) the relative humidity affects DNA conformations

and (c) B-DNA and A-DNA are inter-convertible by controlling relative humidity[12]. Franklin and Gosling speculated the plausible role of “hydrophilic” phosphate groups in protein-DNA interactions. Eventually, further studies observed that indeed proteins[14] like polymerases, endonucleases, interact with DNA to induce B→A transition at the local level; creating substantial interactions with DNA phosphate backbone. In the next section, structural features to recognize B-DNA and A-DNA are presented.

1.3. Structural features of A-DNA and B-DNA.

A-DNA is characterized as the short, compact structure whereas B-DNA is the commonly observed long structure of DNA. A-DNA possesses 11 base pairs (BP) per turn whereas there are 10.5 BP per turn of B-DNA. The bird's-eye view of A-DNA shows the characteristic hollow circle along its principal axis due to enhanced Roll and Slide values, whereas this hollow tunnel is absent in B-DNA (See Fig. 1.5(a)). The B-DNA possesses wide major groove compared to narrow and deep major groove of A-DNA (See MGw region in Fig.1.5(b)). The minor groove in B-DNA is narrow and deep unlike shallow and broad minor groove of A-DNA (See mGw Fig. 1.5(b)). In general, the atoms N7, C6, N6, O6, and C4 in A-T base pair (See Fig. 1(A)) face towards major groove, whereas in G-C base pair, atoms N7, C6, O6, N4 and C4 face major groove.

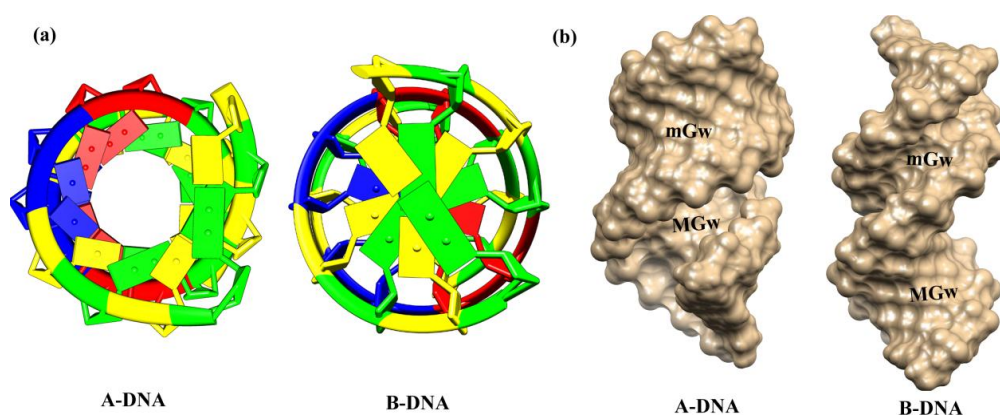


Figure 1.5. (a) Bird's-eye view of A-DNA and B-DNA (from 5' to 3' direction). The Nucleic Acid Database color convention is used indicating Cytosine (Yellow), Guanine (Green), Adenine (Red), Thymine (Blue). (b) major (MGw) and minor (mGw) groove presentation of A-DNA and B-DNA. The figures are plotted using UCSF CHIMERA visualization software [15].

The protein induced B-form→A-form transition at the local level creates shallow minor groove and this enables the protein to access buried base atoms (proton acceptors) such as exocyclic O2 and ring atom N3 as well as sugar O3' atoms.[14,16] However, the choice of the groove for protein-DNA interactions often depends on the protein family. The water dynamics plays a significant role in B→A transition. The presence of locally ordered water molecules in the major groove of crystal structures of A-DNA are observed, which are due to strong hydrogen bonding interactions with major groove atoms. In an interesting simulation “experiment” [17], the restraints on water molecules near major groove lead to A-DNA like conformation.

The A-DNA and B-DNA conformations in long DNA stretches are easily recognizable visually. However, the local A-form or B-form is hard to observe visually and thus are recognized using individual base step structural parameters such as Rise, Slide, Roll and Twist values as shown in Figs. 1.6(a)-(d)

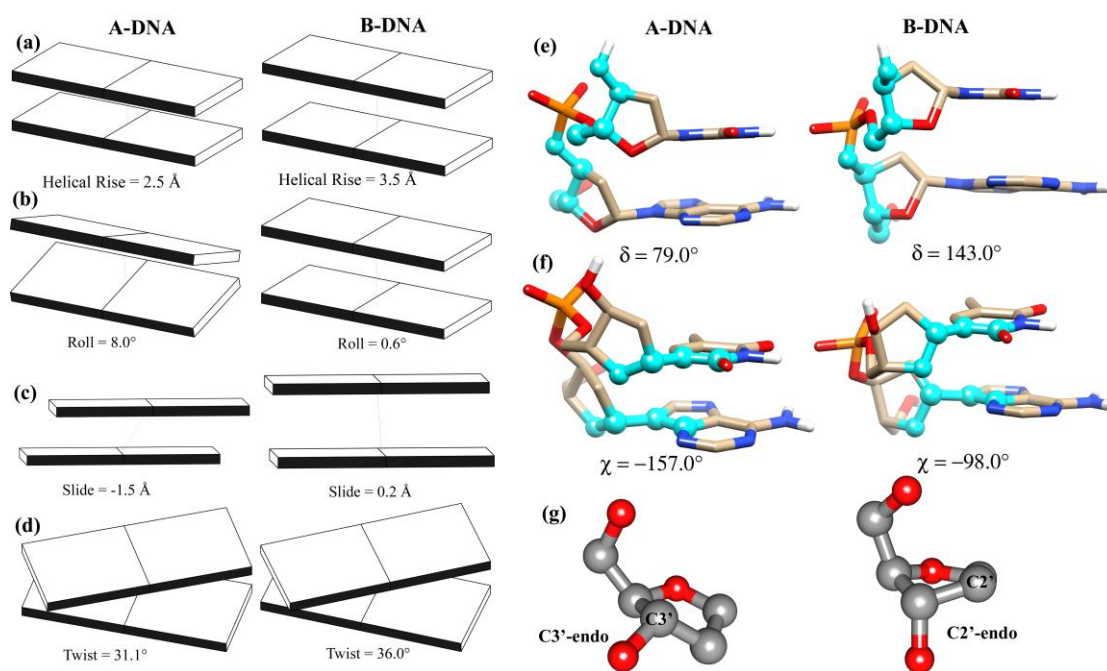


Figure 1.6. The principal structural parameters to distinguish A-form and B-form of DNA; base stacking parameters such as (a) Helical Rise, (b) Roll, (c) Slide and (d) Twist as well as torsional angles (e) δ , (f) χ and, (g) sugar pucker angles.

A pair of base pair i.e. a dinucleotide step is the basic smallest possible unit to understand B-form or A-form at the local level. The distance between two consecutive base pairs, i.e. Rise value in B-DNA is 3.4 Å as compared to 2.5 Å in A-DNA. (See Fig. 1.6(a)). The reduced distance between bases affects the non-covalent

dispersion interactions between bases, and hence stacked base pairs show enhanced Slide (see Fig. 1.6(c)) and Roll (see Fig. 1.6(b)) values in A-form to balance these interactions. The Twist values are not significantly different in A-form and B-form (see Fig. 1.6(d)). The sugar-phosphate backbone constraints the movement of bases and holds the DNA framework stable. The two major torsion angles involved in B→A transition are χ and δ (See Fig. 1.6 (e) and (f)). The δ angle affects the sugar pucker values whereas χ angle involves the glycosidic bond and hence correlates well with the motions of bases and sugar-phosphate backbone during the transition. During B-form → A-form transition, the sugar pucker shows the major conformational difference (See Fig. 1.3(g)). In A-form and B-form, deoxyribose sugar adopts C3'-endo and C2'-endo conformation, respectively.

The structural parameters Roll, Twist, and Slide, possess distinct values for A-form and B-form. However, the distribution of these parameter values obtained from protein-DNA complexes showed the overlapping region (See Fig. 1.7(a)) and created ambiguity to assume the base step conformation as either A-form or B-form.

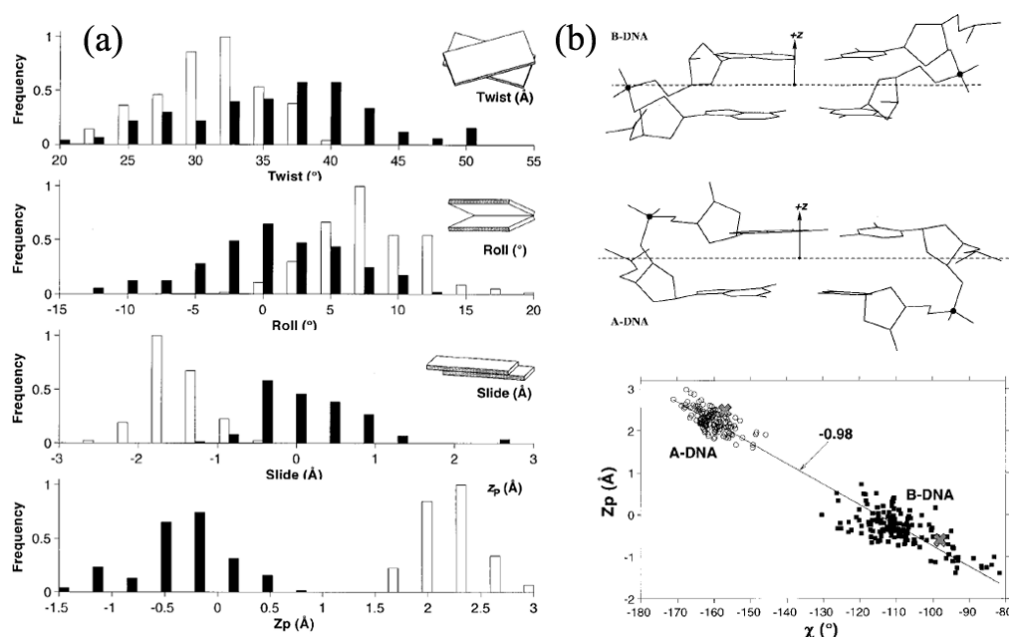


Figure 1.7. (a) The discrimination of A-form and B-form using Twist, Roll, Slide and Z_p parameter and (b) Definition of Z_p and distinct A-form or B-form recognition by Z_p and χ angle. The figures are reprinted with permission from ref. [14]

Hence, Lu *et al.*[14] proposed the Z_p parameter – a displacement of phosphorous atoms from the virtual middle plane of a dinucleotide step (See Fig. 1.7(b)). It can be seen from the Fig. 1.7(a), Z_p can distinguish local A-form and B-form more efficiently. Z_p calculates the antisense to sense strand phosphate inclination along the vertical DNA axis from middle base pair reference frame. In A-form, the phosphate groups of a dinucleotide steps are away from the virtual middle plane of a dinucleotide step and hence possess higher Z_p value of 1.50 Å (Fig. 1.7(b) lower base step), while in the case of B-form value is less than 0.8 Å (Fig. 1.7(b) upper base step). The base steps having Z_p values between 0.80 Å and 1.50 Å are considered as intermediate structures. In the next section, the role of A-form in biological process is described.

1.4 Biological significance of “dry” A-DNA in cellular environment:

Even after 63 years of A-DNA discovery, the novel intriguing biological roles of A-DNA have resurfaced, showing the involvement of this conformation in DNA packaging inside cells[18] and survival at harsh conditions[19]. The analysis of numerous protein-DNA complexes by Olson and co-workers[20] observed that the presence of hydrogen bond donating atoms of amino acid residues around the local A-form stabilizes it.

A-DNA conformation has an important biological role in the context of cellular defense mechanisms under harsh conditions[21]. The A-form of DNA protects *Bacillus Subtillis* spores from UV damage[22]. Recently, Whelan and co-workers have employed FTIR spectroscopy to show fully reversible B-DNA→A-DNA transition in living bacterial cells on desiccation and rehydration[23]. Extremophiles like SIRV2 virus (*Sulfolobus islandicus rod-shaped virus 2*) survive at extreme temperatures of 80°C and acidity of pH 3 by adopting the complete DNA in the A-form and thereby aids protein to encapsidate DNA[19].

Apart from its involvement in defence mechanisms of cells, the remarkable presence of the A-DNA is noted in many biological processes. Certain DNA-protein interactions involve direct recognition processes that require the sugar-phosphate backbone of DNA to be exposed. Proteins such as polymerases, endonucleases, etc.

perform cutting and sealing operations and cause B-DNA to A-DNA transition at the local level (See Fig. 1.8).

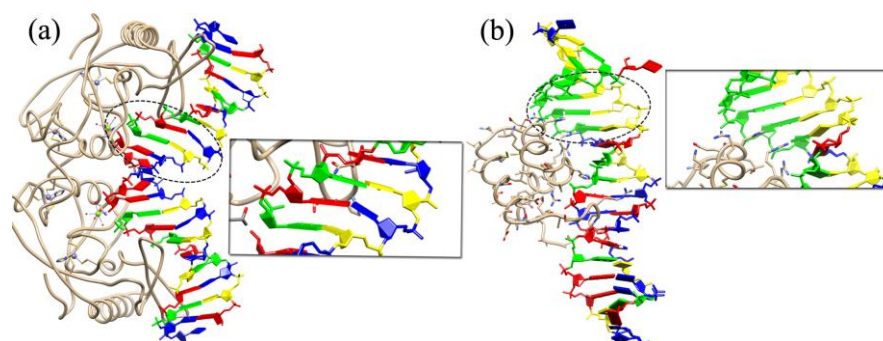


Figure 1.8. Illustrations of protein-DNA interactions for (a) DNA-endonuclease complex [24] (PDB ID: 1A73; inset: backbone interactions with local A-form), (b) crystal structure of Tc3 transposase of *C. elegans* [25] (PDB ID: 1TC3; inset: backbone interactions with local A-form).

The B→A transition involves a change in the major and minor groove widths, hence making buried parts of DNA available for interactions. The A-DNA participates in both direct and indirect protein-DNA interactions, and both mechanisms are sequence dependent. The sequence dependence arises from interactions of all dinucleotide steps involved in the local B-form to A-form transition.

The therapeutically important modified nucleotides such as 2'-O-methyl or 2'-O-methoxyethyl, 2'-Fluorine RNA or locked nucleic acids possess A-form and binds strongly to a specific nucleotide sequence. This strong binding opposes DNA/RNA cleavage by enzymes like endonuclease and prevents termination of further biological processes [26,27]. This example represents the relevance of A-DNA conformation in the present nucleotide chemistry.

The B-philicity or A-philicity has been studied previously for dinucleotide or trinucleotide steps. We will discuss this aspect in the next section.

1.5 The Local B→A DNA transition: Conformational Preferences and Consequences

The possible combinations of all four nucleotides, adenine, guanine, cytosine and thymine, lead to 16 possible dinucleotide steps. However, after removing redundant combinations (e.g. TT for AA, CT for AG and so on), only ten appear as

unique dinucleotide steps. Similarly, there are only 32 unique trinucleotide steps. All 10 unique dinucleotide steps have variable tendencies to adopt the A-form, affecting the kinetics and dynamics of the overall transition of a DNA from B- to A-form. Hence, understanding B-form to A-form transition at the local level in atomistic details may unveil the possible rationale behind the DNA sequence dependence of a specific biological process adopted by nature.

Various experimental and simulation studies attempted to understand dynamics and structural behavior of dinucleotide or trinucleotide steps to predict conformational preferences of a DNA sequence. Basham and co-workers developed “A-DNA triplet code”[28] based on solvation free energies (SFE) of trinucleotide steps to predict DNA conformational preference towards B-DNA or A-DNA conformation under crystallizing conditions. In this method, the modelled canonical A-DNA and B-DNA structures of the same sequence utilized to calculate SFEs. The SFEs are used to derive A-DNA propensity energy (APE) scale. The study provided more than 90% accuracy for a limited dataset of DNA sequences. However, APE values are not available for some trinucleotide steps like CAG, CTG, CCT, CTT, GTT, AGC, AAC, AGG, AAG, ACA, ACT, AGT, TGT and GCT. Hence, it is not possible to apply APE free energy code to predict conformational preferences in all nucleotide sequences.

Tolstorukov and co-workers[29] developed dimeric(dinucleotide) and trimeric(trinucleotide) free energy scales based on the experimental data of B-DNA to A-DNA transition. The analysis of experimental data based on one-dimensional Ising model provided free energy values for all unique 10 dinucleotide steps and 32 trinucleotide steps. This study proposed three models for dinucleotide free energy viz. D-10, D-10/12, D-12, and one model, T-32, for the trinucleotide free energy. The study was able to predict A-philic and B-philic dinucleotide steps quite well based on D-10/12 and D-12 models but was unable to predict the correct midpoint of B→A transition. These models predicted AG, AT, and CG as B-philic steps. The free energy model T-32 divides trinucleotide steps into three categories: A-philic, intermediate, and B-philic steps. T-32 model predicts AAA, AAC, AAG, AAT, and ATC as B-philic steps, ACC, ACT, ATG, CCC/GGG, and GAG/CTC as A-philic steps, and the rest of trinucleotide steps as intermediate steps. Although T- 32 is more accurate than other models, it still cannot be applied for DNA sequences having TAA/TTA step,

thus limiting its application. Apart from thermodynamic studies, other studies have attempted to capture local B-form or A-form conformational preference based on the structural features as discussed below.

The steric hindrance present in the hydrophobic cluster of thymine methyl (-CH₃) group, sugar's -CH₂ group, and thymine's C6 atom resists local B-form to A-form transition in AA/TT step. Nelson *et al.*[30] (1986) suggested that the ability to form cross-strand strong hydrogen bonds between bases also stabilizes the A-form. This observation can explain the B-philicity of CA/TG step over AC/GT and AG/CT over GA/TC. The CA/TG step can form bifurcated hydrogen bonds in both the major and minor grooves and hence it is more B-philic than the AC step, which is capable of having only a single cross-strand hydrogen bond in the major groove.

To understand the intrinsic tendency of dinucleotide step towards A- or B-form, Svozil, and co-workers[31] studied 8000 dinucleotide steps from 447 DNA crystal structures. The conformational space of these steps is divided into six conformational classes mainly BI, BII, AI, AII, B/A and A/B. This classification was based on the torsional angles' derived multidimensional space. The torsion angles α (O3'[i-1]-P-O5'-C5'), β (P-O5'-C5'-C4'), γ (O5'-C5'-C4'-C3'), δ (C5'-C4'-C3'-O3'), ϵ (C4'-C3'-O3'-P[i+1]), and ζ (C3'-O3'-P[i+1]-O5' [i+1]) angles were considered. The analysis shows a strong BI conformational preference for all the dinucleotide steps. The local influence such as the presence of protein observed as a plausible cause for dinucleotide step to adopt a broad range of conformations. The analysis shows a complete absence of AA step in the A-form indicating an extremely rigid B-philicity. Also, GG steps showed the significant population in the A-form portraying its strong A-philicity.

Based on the purine (R: A, G) and pyrimidine (Y: C, T) base classification, YR steps have displayed an abundant population in the BII and A/B families, and RY in B/A families. The YR steps such as CA, TA, CG, and TG are strongly observed at the bent and kinked regions in DNA. The pyrimidine-pyrimidine steps such as CC, TT, CT, TC are absent in BII conformation indicating their less-pronounced role in the protein-DNA recognition. The CG (YR) step can adopt AI, AII, BI, BII as well as A/B conformation. Its conformational preference, however, depends on the flanking as well as distant steps[31]. In general, AA, TT, and GA steps prefer BI conformation preference, and CA, TG base pairs show a preference for BII forms. All RR and YY steps are observed in BI conformation except GG, which is predominantly seen in the

AII form. YR steps are found to prefer AII conformation in the naked A-DNA and B-DNA crystal structures. The classification of all ten dinucleotide steps is presented in Fig. 1.6.

	Y (C,T)	R (A,G)
Y (C,T)	CC: AI, BI CT: BI TC: BI, B/A TT: BI	CA: AI, BII CG: AI, AII, BI, BII, A/B TA: AI, AII, BI, BII TG: AI, BII
R (A,G)	AC: AI AT: BI, B/A GC: AI, BI, BII, B/A GT: AI	AA: BI, B/A GG: AI, AII GA: BI, BII AG: BI

Figure 1.6. The schematic diagram showing conformational preferences for all ten unique dinucleotide steps[16].

Marathe and coworkers[32] studied the structural features of dinucleotide steps and classified them as A-phobic i.e. B-philic and A-philic in the context of protein-DNA interactions. The study confirmed B-philicity of AA and A-philicity of GG based in Z_p and Slide values. The study found AG and GA as B-philic steps, AT, GC, CA, TA, and CG as both A-philic and B-philic steps, and GG as a strongly A-philic step. The study supported Tolstorukov's assumption[29] that except AA and GG, rest of the dinucleotide steps can assume A- or B-form conformation depending on the flanking base steps.

Both the experiment-based analytical models and solvation free energies based models provided a free energy scale to predict A-DNA or B-DNA conformation from the knowledge of sequence alone. However, their effective application for structure prediction is limited by the lack of free energy values for certain trinucleotide steps. Most of the earlier computational and experimental free energy studies focused on the complete B-DNA to A-DNA transition. Therefore, all-atom molecular dynamics simulation study to characterize the behavior of each unique dinucleotide step was lacking. Hence further investigation was needed to show how the sequence dependent local B→A transition affects the overall global transition. Such effects are discussed in the next section.

1.6 The Global B-A transition: Observations and Cooperative transition

The understanding of structure and dynamics of local to B-form to A-form transition can be used to explore the global transition in a DNA helix. Experimental studies explored the sequence dependence and cooperative nature of B→A transition using various approaches[33]. These studies were performed mainly by titration of DNA sequences in ethanol to monitor the change in the Circular Dichroism (CD) signatures corresponding to A-DNA or B-DNA conformations. The aggregation of DNA molecules at a higher alcohol concentration is a serious limitation for the experimental B→A transition studies. Some studies had shown that aggregation has no effect on the cooperative B-DNA to A-DNA transition in long DNA helices[34,35]. During B→A transition, the DNA molecule gets subdivided into alternating sections of the A and B conformations. At the B→A transition midpoint, the average length of these A-DNA sections reflects the degree of cooperativity and termed as cooperative length of B→A transition. The cooperative length for the B→A transition was found to be in the order of 10-20 base pairs using one-dimensional Ising model[36]. The CD spectroscopy predicted the free energy of B→A transition at pH 7.0 as 21.2 kcal/mol and 19.8 kcal/mol at 0.2 mM NaCl and 0.5 mM NaCl, respectively. This free energy corresponds to the free energy required to convert a cooperative length of DNA (~20 base pairs). Hence, the cost per dinucleotide step for B→A conversion has been predicted to be 1 kcal/mol[37].

The presence of strongly A-philic short guanine tract (G-tract) in dodecamer DNA sequence d(CATGGGCCCATG)₂ allowed exploring intermediate conformations between A-DNA and B-DNA[38]. Independent theoretical studies have predicted the free energy barrier (less than 0.2 kcal/mol) between A- and B-forms of a G-tract to be below thermal fluctuations for CC/GG step[39,40]. The enhanced electrostatic repulsion in stacked G:C base pairs cause them to adopt structure close to A-form. Therefore, it is suggested that sequences with more %GC content easily undergoes B→A transition by reducing the cooperative length[41].

The stacking energies[42] for the dinucleotide steps of the intermediate conformation of d(CATGGGCCCATG)₂ were found to be close to the standard and crystal B-DNA geometries using *ab initio* calculations. Dickerson and co-workers[43]

have experimentally captured the trajectory of B-DNA to A-DNA transition by capturing the crystal structures of intermediate states either by methylating or by brominating the cytosine bases in the sequence d(GCCGGC)₂. The pathway of B→A transition traced the cooperative change of various parameters such as sugar puckers, Slide, and Z_p values. In intermediate states a mixed C2'-endo and C3'-endo sugar pucker was observed. Authors proposed that for a DNA sequence transforming from B-DNA to A-DNA, the intermediate structures might maintain structural features of the starting or ending states by noting the common occurrences of B-form conformation with A-like features during the B→A transition.

The simulation study of B→A transition using internal coordinate molecular dynamics simulations[36] demonstrated the reversible and cooperative nature of B→A transition. The cooperative nature suggests that the concerted motion of DNA bases at a certain fragment of DNA drives B→A transition. The cooperative length was proposed to be 20 base pairs long earlier but revised to 10 base pair recently[36]. This also suggests that DNA less than 10 BP would not exhibit B→A transition in solution. This suggestion was demonstrated using hexamer sequences which either remained in the B-DNA conformation or deviated from it marginally, never achieving the A-form.

The C2'-endo sugar conformation in B-form is slightly more stable than C3'-endo sugar conformation in the A-form with a barrier of 2.0 to 3.0 kcal/mol[44]. As stated earlier, the effective barrier for the B→A transition is around 19.0 to 20.0 kcal/mol. Hence, origin of the cooperative nature was expected due to switching of the sugar pucker values from C2'-endo to C3'-endo conformation. However, a novel computational study based on the reversible B→A transition[36] observed that sugar puckers changed in the last step of the B→A transition. This study also revealed that electrostatic interactions between DNA backbone and counterions present around the major groove stabilize the A-DNA conformation. This study also suggested that changes in the sugar pucker during B→A transition were a consequence and not the mechanism of transition. The electric field jump experiments were performed[45] on poly[d(A-T)] sequence as it exhibits high cooperativity in B→A transition. A fast process of ~10 μs timescale and a slow process of ~100 μs timescale was observed during this study of global B→A DNA transition. The slow process was attributed to the nucleation driven B→A transition in an ethanol solution. The linear Ising model

was implemented[33] to understand thermodynamics and kinetics of the cooperative B→A transition. The nucleation parameter (σ) for the B→A transition was estimated between 10^{-2} and 10^{-3} , indicating a low cooperative length of 10 -30 base pairs for B→A transition, as it is related to the nucleation parameter as $1/\sqrt{\sigma}$.

The simulation study[46] of B-DNA and A-DNA in water and water-ethanol mixtures was carried out to understand limitations of the Cornell *et al.* [47] force field. The study found that B-DNA stability is overestimated in low water activity system such as water-ethanol mixture. Further improvements of α/γ torsion angles of the same force field lead to improved description (parmbsc0 force field[48]) of A-DNA → B-DNA transition in water as well as accurate description of DNA:RNA hybrid conformations. Recently developed PARMBSC1[49] force field improves overall description of B-DNA geometry and the microsecond long simulations of B-DNA observed twist values close to experimental values. The simulation study[50] of 24-mer sequence d(CTGTATCTATATAAAACGGGCTGG)₂ was performed using CHARMM22 force field, which has a bias to form A-DNA conformation irrespective of the sequence. This 24-mer sequence provided a detailed insight into the mechanism as it contains a combination of A-philic G-tract, an alternating TA-tract, and B-philic A-tract. This study observed that water molecules achieve an economy of hydration in the A-DNA by forming water bridges between phosphate oxygen atoms because of the reduced intra-strand phosphate-phosphate distance. The strong B-philicity of AA/TT step was expected due to the spine of hydration in the minor groove and the presence of bifurcated hydrogen bonds in the major groove[29], whereas GC base pairs easily adopt the A-form due to the ability of cytosine to achieve C3'-endo conformation. The strong interactions of sodium ions with DNA phosphate and major groove atoms stabilize the A-form. The study by Banavali[51] calculated the free energy difference between the A- and B-forms for the sequence d(CTCGAG)₂ as ~2.8 kcal/ mol in the aqueous solution. The study found no intermediates along the pathway and no well-defined minimum for the A-form.

The sugar pucker changes faster to the C3'-endo conformation in guanine and cytosine bases compared to adenine and thymine. It was found that the complete B→A transition requires 11.4 kcal/mol barrier for a short hexamer sequence[52]. It was noted that although a slight deformation pushes B-DNA close to a structure similar to A-DNA, attaining the complete A-DNA structure in solution was difficult.

Therefore, complete A-DNA geometry in water is possible only in the presence of proteins or other molecules.

The experimental and simulation study indicated the sequential mechanism rather than all-or-none mechanism as expected earlier. The A→B transition in water indicated changes in sugar pucker followed by changes in a base stacking arrangement[17]. The solvent-accessible surface area (SASA) is less in A-form compared to B-form. The magnitude of reduction in SASA is significantly different for AT and GC base pairs. The SASA of the major groove region in AT base pairs is greatly reduced and hence more water rearrangement was observed in AT base pairs during A-DNA to B-DNA transition. The Raman spectroscopic observations found strong binding of water molecules in the major groove in A-form of DNA. In case of B-DNA, water molecules bind strongly to the narrow A.T minor groove to make a spine of hydration and thus high A.T content resists B→A transition.

The above discussion indicates that the previous studies lack a detailed atomistic picture of local as well as global B→A transition in an aqueous environment in the context of protein-DNA interactions. These observations necessitate exploration of thermodynamics as well as mechanistic details at the atomistic level of B-DNA to A-DNA transition at the local level (partial B → A transition in DNA) as well as on a global scale (complete B→A conversion in DNA sequence). This thesis aims to bridge this gap. A general outline of the thesis is described in the next section.

1.7 Outline of the thesis

In this thesis, we performed MD simulations along with umbrella sampling technique to understand the A-philicity or B-philicity of 10 unique dinucleotide steps. The study provides an insight on the complex B→A transition mechanism at the local level. It explores the effect of dinucleotide sequence, the role of water and the flanking base pairs in this local B → A transition. The knowledge of sequence dependent free energies is implemented to develop a model for *a priori* prediction of A-DNA or B-DNA conformation merely from the knowledge of sequence. The study is extended to understand another unexplored aspect of nucleation mechanism in global B→A transition of long DNA helices. A chapter-wise summary of the thesis is provided below.

Chapter 2 describes a novel reaction coordinate Z'_p and details of methodology implemented to explore the one-dimensional free energy profiles of all 10 unique dinucleotide steps embedded in $d(\text{CGCGAXXTCGCG})_2$ in an aqueous environment. The details of cluster analysis, thermodynamic and mechanistic analyses are also described.

Chapter 3 describes the results of the umbrella sampling simulations performed along Z'_p and thereby comments on the A-form or B-form conformational preferences of all 10 dinucleotide steps, role of different interactions such as stacking and solvation energy of dinucleotide step under study as well as the heterogeneity in B-form $\rightarrow$ A-form transition mechanism at a local level.

Chapter 4 describes the effect of flanking base steps on the local B-form to A-form transition and derives the B/A junction free energies for AA/TT and GG/CC steps. The *absolute free energies* gained after removal of B/A junction energies are also discussed thereby.

Chapter 5 describes the method to predict a B-DNA or A-DNA conformational preference of a DNA sequence merely from the knowledge of its sequence. This chapter demonstrates the prediction accuracy of the method when applied to crystal structures and *Xenopus tropicalis* genome.

Chapter 6 describes the study of the global (whole DNA) B-DNA $\rightarrow$ A-DNA transition of a single DNA helix in solution, which showed the existence of a multistep non-classical nucleation mechanism for B $\rightarrow$ A transition for the first time.

Chapter 7 describes the salient conclusions of the local as well as global B-DNA $\rightarrow$ A-DNA transition observed from various approaches. This chapter also comments on the possible future directions such as possible simulation study to understand the role of 2'-5' phosphodiester linkages during prebiotic world and possible designing of DNA nanodevices exploiting continuous B-form $\rightarrow$ A-form $\rightarrow$ B-form transition based on Scrunchworm hypothesis.

Chapter 8 describes two studies based on the DNA – ionic liquid interactions mainly (a)) a collaborative project based on G-quadruplex ionic liquid interactions and (b) the DNA translocation study as Addendum I and Addendum II respectively.

Addendum I discusses the computational study to understand the possible interactions between ionic liquid molecules and G-quadruplex to understand the stabilization mechanism by specific ionic liquid.

Addendum II comments on the computational efforts to understand the possible application of DNA-IL interactions to resolve the present limitations of DNA-Graphene nanopore technology for sequencing

References:

- (1) Kool, E. T. *Annu Rev Biophys Biomol Struct* **2001**, 30, 1.
- (2) Freier, S. M.; Sugimoto, N.; Sinclair, A.; Alkema, D.; Neilson, T.; Kierzek, R.; Caruthers, M. H.; Turner, D. H. *Biochemistry* **1986**, 25, 3214.
- (3) Serra, M. J.; Turner, D. H. In *Methods in Enzymology*; Academic Press: 1995; Vol. Volume 259, p 242.
- (4) SantaLucia, J.; Kierzek, R.; Turner, D. *Science* **1992**, 256, 217.
- (5) Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc* **1972**, 94, 8205.
- (6) Blackburn, G. M.; Chemistry, R. S. o. *Nucleic Acids in Chemistry and Biology*; RSC Pub., **2006**.
- (7) Sychrovsky, V.; Foldynova-Trantirkova, S.; Spackova, N.; Robeyns, K.; Van Meervelt, L.; Blankenfeldt, W.; Vokacova, Z.; Sponer, J.; Trantirek, L. *Nucleic Acids Res* **2009**, 37, 7321.
- (8) Karplus, P. A.; Shapovalov, M. V.; Dunbrack, R. L., Jr; Berkholtz, D. S. *Acta Crystallographica Section D* **2008**, 64, 335.
- (9) Lu, X. J.; Olson, W. K. *Nucleic Acids Res* **2003**, 31, 5108.
- (10) Lescrinier, E.; Froeyen, M.; Herdewijn, P. *Nucleic Acids Res.* **2003**, 31, 2975.
- (11) Saenger, W.; Hunter, W. N.; Kennard, O. *Nature* **1986**, 324, 385.
- (12) Franklin, R. E.; Gosling, R. G. *Nature* **1953**, 171, 740.
- (13) Bagchi, B. *Water in Biological and Chemical Processes*; Cambridge Molecular Science, **2014**.
- (14) Lu, X.-J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, 300, 819.
- (15) Couch, G. S.; Hendrix, D. K.; Ferrin, T. E. *Nucleic Acids Res.* **2006**, 34, 29.
- (16) Kiefer, J. R.; Mao, C.; Braman, J. C.; Beese, L. S. *Nature* **1998**, 391, 304.
- (17) Knee, K. M.; Dixit, S. B.; Aitken, C. E.; Ponomarev, S.; Beveridge, D. L.; Mukerji, I. *Biophys. J.* **2008**, 95, 257.
- (18) Harvey, S. C. *J. Struct. Biol.* **2015**, 189, 1.
- (19) DiMaio, F.; Yu, X.; Rensen, E.; Krupovic, M.; Prangishvili, D.; Egelman, E. H. *Science* **2015**, 348, 914.
- (20) Andrew V. Colasanti, Y. L., Gautam Singh, Xiang-Jun Lu, Wilma Olson In *Biomolecular Forms and Functions, A Celebration of 50 years of the Ramachandran Map*; Srinivasan, M. B. a. N., Ed.; IISc Press-WSPC Publishers: IISc, 2013, p 230.
- (21) Lee, K. S.; Bumbaca, D.; Kosman, J.; Setlow, P.; Jedrzejewski, M. J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, 105, 2806.
- (22) Mohr, S. C.; Sokolov, N. V.; He, C. M.; Setlow, P. *Proc Natl Acad Sci U S A* **1991**, 88, 77.
- (23) Whelan, D. R.; Hiscox, T. J.; Rood, J. I.; Bambery, K. R.; McNaughton, D.; Wood, B. R. *J. R. Soc. Interface* **2014**, 11, 20140454.
- (24) Flick, K. E.; Jurica, M. S.; Monnat, R. J.; Stoddard, B. L. *Nature* **1998**, 394, 96.
- (25) van Pouderooyen, G.; Ketting, R. F.; Perrakis, A.; Plasterk, R. H.; Sixma, T. K. *EMBO J* **1997**, 16, 6044.
- (26) Kawasaki, A. M.; Casper, M. D.; Freier, S. M.; Lesnik, E. A.; Zounes, M. C.; Cummins, L. L.; Gonzalez, C.; Cook, P. D. *J. Med. Chem.* **1993**, 36, 831.
- (27) Moreno, P. M.; Pego, A. P. *Front. Chem.* **2014**, 2, 87.

- (28) Basham, B.; Schroth, G. P.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, 92, 6464.
- (29) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, 81, 3409.
- (30) Nelson, H. C. M.; Finch, J. T.; Luisi, B. F.; Klug, A. *Nature* **1987**, 330, 221.
- (31) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. *Nucleic Acids Res.* **2008**, 36, 3690.
- (32) Marathe, A.; Karandur, D.; Bansal, M. *BMC Struct. Biol.* **2009**, 9.
- (33) Jose, D.; Porschke, D. *J. Am. Chem. Soc.* **2005**, 127, 16120.
- (34) Potaman, V. N.; Bannikov, Y. A.; Shlyachtenko, L. S. *Nucleic Acids Res* **1980**, 8, 635.
- (35) Zavriev, S. K.; Minchenkova, L. E.; Frank-Kamenetskii, M. D.; Ivanov, V. I. *Nucleic Acids Res* **1978**, 5, 2657.
- (36) Mazur, A. K. *J. Am. Chem. Soc.* **2003**, 125, 7849.
- (37) Gekko, K.; Sakaki, T.; Shindo, H. *Polym J* **1988**, 20, 751.
- (38) Ng, H. L.; Kopka, M. L.; Dickerson, R. E. *Proc Natl Acad Sci U S A* **2000**, 97, 2035.
- (39) Minchenkova, L. E.; Schyolkina, A. K.; Chernov, B. K.; Ivanov, V. I. *J Biomol Struct Dyn* **1986**, 4, 463.
- (40) Foloppe, N.; MacKerell Jr, A. D. *Biophys. J.* **1999**, 76, 3206.
- (41) Sponer, J.; Gabb, H. A.; Leszczynski, J.; Hobza, P. *Biophys J* **1997**, 73, 76.
- (42) Sponer, J.; Florián, J.; Ng, H.-L.; Sponer, J. E.; Spacková, N. a. *Nucleic Acids Res* **2000**, 28, 4893.
- (43) Dickerson, R. E.; Ng, H.-L. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, 98, 6986.
- (44) Foloppe, N.; Nilsson, L.; MacKerell, A. D. *Biopolymers* **2001**, 61, 61.
- (45) Jose, D.; Porschke, D. *Nucleic Acids Res.* **2004**, 32, 2251.
- (46) Sproun, D.; Young, M. A.; Beveridge, D. L. *J. Phys. Chem. B* **1998**, 102, 4658.
- (47) Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc* **1995**, 117, 5179.
- (48) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. *Biophys J* **2007**, 92, 3817.
- (49) Ivani, I.; Dans, P. D.; Noy, A.; Perez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goni, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpi, J. L.; Gonzalez, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. *Nat Methods* **2016**, 13, 55.
- (50) Pastor, N. *Biophys. J.* **2005**, 88, 3262.
- (51) Banavali, N. K.; Roux, B. *J. Am. Chem. Soc.* **2005**, 127, 6866.
- (52) Noy, A.; Pérez, A.; Laughton, C. A.; Orozco, M. *Nucleic Acids Res.* **2007**, 35, 3330.

Chapter 2

Methodology

2.1 Molecular Dynamics simulations

Classical MD simulations play an important role in determining the structural and dynamical properties of different systems at the atomistic level. MD simulations employ the force fields that define the inter-atomic interactions via harmonic potential functions. The Eq. 2.1 represents general form of force field,

$$U = \sum_{bonds} \frac{k_{bond}}{2} (r - r_{eq})^2 + \sum_{angle} \frac{k_{angle}}{2} (\theta - \theta_{eq})^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i>j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right], \quad Eq. 2.1$$

Where, k_{bond} , k_{angle} , V_n are force constants, n is multiplicity and γ is phase angle for dihedral angles. r_{eq} and θ_{eq} represent equilibrium values for bond and angle parameters. The last term of equation Eq. 2.1 represents all non-bonded interactions - a combination of Lenard-Jones 6-12 potential and electrostatic interactions. The force field consists of constant values for k_{bond} , k_{angle} , V_n , n , γ , r_{eq} , θ_{eq} , A_{ij} , B_{ij} , q_i , q_j , ϵ whereas the variables r , θ , ϕ , R_{ij} are updated during simulation.

The biological processes span a range of timescales[1] from picoseconds (photosynthetic electron transfer) to microseconds (helix-coil transitions, ion channel gating) to milliseconds (protein folding, ligand binding) involving single or multiple steps. However, at present, the timescales accessible to MD simulations are in the order of microsecond with the advent of GPU units and special hardware architecture[2]. The sufficient phase space sampling is necessary to capture the reliable sampling of phase space of such long timescale processes. Hence, enhanced sampling techniques like umbrella sampling, metadynamics, parallel tempering, are often employed to explore such long timescale processes.

As discussed below, the preliminary information of B→A transition also points towards the existence of a barrier for the local B-form to A-form transition. The sequence dependence of B→A transition arises from the complex interactions between various base steps simultaneously undergoing global B-DNA to A-DNA

transition in solution. The cooperativity affects the barrier of global B→A transition. As discussed in Chapter 1, the free energy of B→A transition is estimated to be ~21.2 kcal/mol at a physiological salt concentration in ethanol solutions.[3] This free energy is required to convert the structure of DNA of length equal to the cooperative length of DNA, which is roughly 20 base pairs long. This indicates that 1 kcal/mol is roughly the cost per dinucleotide step for B→A conversion[3]. The B→A transition of a short hexamer in an aqueous solution possess free energy barrier of 13.5 kcal/mol [4], indicating the free energy per dinucleotide step is ~ 2.0 kcal/mol($>3k_B T$) in aqueous solutions. The conversion of a single dinucleotide step to A-form in a long B-DNA sequence in water is an extremely rare event (one A-form step in 1000 B-form steps[5]). Hence, in this thesis, umbrella sampling technique has been employed to determine the possible barriers involved during the local B-form to A-form transition.

2.2 Umbrella sampling simulations

The free energy between two thermodynamic states – reactant and the product can be obtained by performing umbrella sampling simulations along one or multiple reaction coordinates. The reaction coordinate can also be an order parameter which is capable of distinguishing the reactant and product states. Since the sampling along the reaction coordinate is adequate when the system is in equilibrium with all the other degrees of freedom, it is advisable to choose the slowest degrees of freedom as the reaction coordinate. In practice, however, one often uses heuristic explanation to choose the reaction coordinate.

Along reaction coordinate, a process of interest progresses via close equally spaced “windows” in a quasi-static manner using a harmonic bias potential. The term “window” refers to the position of the biased potential along the reaction coordinate. The position of the bias potential allows sampling of that particular region of the reaction coordinate. Such multiple overlapping closely spaced windows provide a well-converged free energy profile and allow reliable sampling of a conformational space along the reaction coordinate. The unbiased free energy profile can be obtained by removing biases and combining the free energy obtained from each “window” using methods such as weighted histogram analysis method (WHAM), [6,7] umbrella integration [8], etc. The details of the umbrella sampling technique and application of

WHAM to obtain the unbiased free energy surface are discussed below as explained in Reference [6,7].

Consider a reaction progresses from a reactant to a product state along the reaction coordinate ξ . For umbrella sampling simulations, ξ will be divided into equal windows of fixed values of the reaction coordinate ξ_i^{ref} ($i= 1, 2 \dots N$). The harmonic bias potential is applied over each window to obtain a distribution of the reaction coordinate ξ_i .

The bias potential $\omega_i(\xi)$ can be expressed as,

$$\omega_i(\xi) = \frac{k_i}{2} (\xi - \xi_i^{ref})^2. \quad \text{Eq. 2.2}$$

This harmonic bias provides the biased distribution of ξ (indicated by $p_i^b(\xi)$). The unbiased distribution $p_i^u(\xi)$ for a certain window is given by the equation below:

$$p_i^u(\xi) = \frac{\int \exp[-\beta E(r)] \delta[\xi(r) - \xi] d^N r}{\int \exp[-\beta E(r)] d^N r} \quad \text{Eq. 2.3}$$

where, ξ is a fixed value at the centre of the window.

The biased distribution of ξ under the influence of external harmonic bias potential $\omega_i(\xi)$ is expressed as

$$p_i^b(\xi) = \frac{\int \exp\{-\beta[E(r) + \omega_i(\xi(r))]\} \delta[\xi(r) - \xi] d^N r}{\int \exp\{-\beta[E(r) + \omega_i(\xi(r))]\} d^N r}, \quad \text{Eq. 2.4}$$

As the bias $\omega_i(\xi)$ is present on the reaction coordinate ξ and integration in the numerator is performed over all degrees of freedom except ξ , the Eq. 2.4 can be presented as

$$p_i^b(\xi) = \frac{\exp[-\beta \omega_i(\xi)] \int \exp\{-\beta[E(r)]\} \delta[\xi(r) - \xi] d^N r}{\int \exp\{-\beta[E(r) + \omega_i(\xi(r))]\} d^N r}, \quad \text{Eq. 2.5}$$

Thus, from Eq. 2.3 and Eq. 2.5, we get

$$p_i^u(\xi) = p_i^b(\xi) \exp[\beta \omega_i(\xi)] \langle \exp[-\beta \omega_i(\xi)] \rangle, \quad \text{Eq. 2.6}$$

The combined unbiased distribution $P_i^u(\xi)$ is given as,

$$P^u(\xi) = \sum_{i=1}^{windows} W_i \cdot p_i^u(\xi), \quad \text{Eq. 2.7}$$

Where, W_i are weights calculated in WHAM method to reduce statistical error while combining all windows.

$$\sum_{i=1}^{windows} W_i = 1 \quad \text{Eq. 2.8}$$

$$W_i = \frac{a_i(\xi)}{\sum a_i(\xi)}; a_i(\xi) = N_i \exp[-\beta \omega_i(\xi) + \beta F_i], \quad \text{Eq. 2.9}$$

Where, N_i is number of steps in window i .

$$\exp(-\beta F_i) = \int P^u(\xi) \exp[-\beta \omega_i(\xi)] d\xi, \quad \text{Eq. 2.10}$$

The terms $P^u(\xi)$ (from Eq. 2.7) enters Eq. 2.10; and F_i enters Eq. 2.10 via Eq. 2.9

The free energy $A(\xi)$ is obtained by reiterating above equations until convergence is achieved and free energy $A(\xi)$ is obtained as

$$A(\xi) = -\frac{1}{\beta} \ln P^u(\xi), \quad \text{Eq. 2.11}$$

The details of the reaction coordinate have been described in the next section.

2.3 Reaction coordinate

The choice of the reaction coordinate is often based on the heuristic knowledge. However, an important criterion for the choice is that the coordinate should be able to capture both the reactant and product states unequivocally. Also, the coordinate should be the slowest one[9,10] so that all the other degrees of freedom can easily equilibrate along the reaction coordinate. Previous free energy studies on DNA conformational transition focussed on the global B-DNA $\rightarrow$ A-DNA transition in water[4] as well as in aqueous ethanol solution[11]. The reaction coordinates implemented in these studies were torsion angle δ and root mean square deviation

(RMSD) or the difference between RMSDs (Δ RMSD) from A-DNA and B-DNA. However, these coordinates were suitable only for global B-DNA to A-DNA transition. Thus, we explored the possible reaction coordinates to drive local B-form to A-form transition in water of a middle dinucleotide step in a 12-mer DNA sequence.

The order parameters that efficiently distinguish B-form and A-form at the local level would be the choice as a reaction coordinate. One such parameter- Z_p [12,13], was discussed in section 1.3 of Chapter 1. In the calculation of Z_p for a dinucleotide step, first each DNA base is aligned with standard coordinate frame of reference, so as to find the virtual middle plane for a pair of base pairs. Then, Z_p parameter is calculated from the positions of antisense strand's phosphorous ($P_{\text{antisense}}$) atom and sense strand's phosphorus atom (P_{sense}) from middle base pair reference frame.

Z_p distinguishes efficiently between A-form and B-forms of static (e.g., crystal) structures as well as dynamic structures obtained from simulations[14]. However, Z_p can't be used in MD simulation because the force calculation is difficult using the midframe convention used in Z_p calculation. Also, use of standard geometry in Z_p calculation will not distinguish various fluctuations in atomistic configuration in MD simulation.

Therefore, in this study, we have introduced a new coordinate Z'_p , based on a relatively rigid sugar framework, similar to Z_p to induce localized (dinucleotide step) transition from B- to A-form. To do that, we created a local DNA triad axis system (Fig. 2.1) following the qualitative guideline by Dickerson *et al.*[15], subsequently adopted in Tsukuba convention[16].

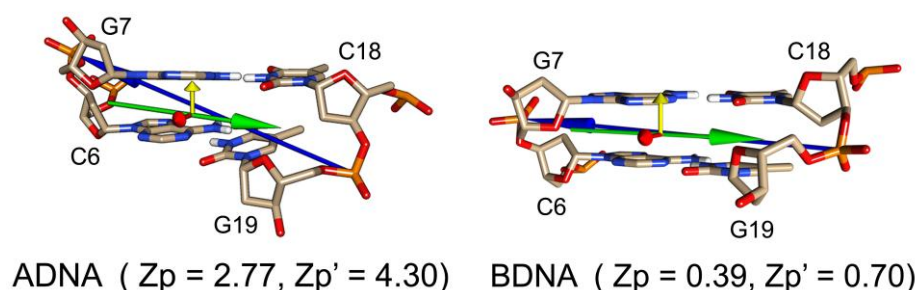


Figure 2.1 Schematic presentation of Z'_p for a middle base step C6G7/C18G19 of the DNA sequence d(CGCGACCGTCGCG)₂.

Figure 2.1 illustrates the definition of Z'_p for the middle base pairs as C6G7/C18G19. The red vector ($\vec{x}$) in Fig. 2.1 is created from the center of mass (COM) of four sugar groups (C6, G7, C18, G19) to the COM of two sugar groups on the 3' ends (sugars of G7 and G19). This vector thus points towards the minor groove. Green vector ($\vec{y}$) is constructed from the COM of two C1' atoms of the sense strand (6C1' and 7C1') to the COM of the two C1' atoms of the antisense strand (18C1' and 19C1'). The unit vector along DNA axis, $\hat{z}$, is constructed by the cross product of unit vectors $\hat{x}$ and $\hat{y}$, $\hat{z} = \hat{x} \times \hat{y}$, shown in yellow. Finally, a vector $\vec{p}$ is created from the phosphorous atoms joining the antisense strand bases (19P) to that of the sense strand bases (7P), shown in blue in Fig. 2.1. Z'_p is defined as the projection of the vector $\vec{p}$ on $\hat{z}$, i.e., $Z'_p = \vec{p} \cdot \hat{z}$

Derivation of Z'_p :

The derivation of the forces to implement Z'_p coordinate in umbrella sampling has been briefly discussed. Umbrella sampling simulations are carried out with external harmonic potential around the value of reaction coordinate where sampling is intended. The following potential can be written as,

$$U = \frac{1}{2} K (Z'_p - Z'_{p,ref})^2 \quad \text{Eq. 2.12}$$

Where, K (kcal/mol $\text{\AA}^{-2}$) is the force constant and $Z'_{p,ref}$ ($\text{\AA}$) is value around which sampling is done in a particular simulation. Z'_p is defined as,

$$Z'_p = \vec{p} \cdot \frac{\vec{x} \times \vec{y}}{|\vec{x} \times \vec{y}|} \quad \text{Eq. 2.13}$$

Each of the above vectors is connected between a group of atoms. To derive the force due to this external potential (Eq. 2.13), we express below the vectors using the notation from Allen and Tildesley[17].

$$\vec{x} = \vec{d}_a = \vec{r}_a - \vec{r}_{a-1}$$

$$\vec{y} = \vec{d}_b = \vec{r}_b - \vec{r}_{b-1}$$

$$\vec{p} = \vec{d}_c = \vec{r}_c - \vec{r}_{c-1}$$

Therefore **Eq. 2.12** can be written as,

$$Z'_p = \vec{d}_c \cdot \frac{\vec{d}_a \times \vec{d}_b}{|\vec{d}_a \times \vec{d}_b|}$$

Forces on a particular atom a (or group of atoms) can be calculated from the negative gradient of the potential on the particular atom as shown below.

$$\vec{f}_a = -\vec{\nabla}_{r_a} U = -K(Z'_p - c)\vec{\nabla}_{r_a} Z'_p$$

Gradient on a particular atom “a” can be written as,

$$\vec{\nabla}_{r_a} Z'_p = \vec{\nabla}_{r_a} \left[\vec{d}_c \cdot \frac{\vec{d}_a \times \vec{d}_b}{|\vec{d}_a \times \vec{d}_b|} \right]$$

Simplifying further the cross products, we get the following relations.

$$\begin{aligned} |\vec{d}_a \times \vec{d}_b| &= \sqrt{(\vec{d}_a \times \vec{d}_b) \cdot (\vec{d}_a \times \vec{d}_b)} \\ &= \sqrt{(\vec{d}_a \cdot \vec{d}_a)(\vec{d}_b \cdot \vec{d}_b) - (\vec{d}_a \cdot \vec{d}_b)(\vec{d}_a \cdot \vec{d}_b)} \\ &= \sqrt{C_{aa}C_{bb} - C_{ab}^2} \\ &= D_{ab}^{1/2} \end{aligned}$$

Here, $C_{ij} = \vec{d}_i \cdot \vec{d}_j$. Similarly, $C_{ijk} = \vec{d}_i \cdot (\vec{d}_j \times \vec{d}_k)$.

After some algebraic manipulations, gradients on each of the atoms (or group of atoms) involved in the Z'_p coordinate can be written as,

$$\begin{aligned} \vec{\nabla}_{r_a} Z'_p &= -C_{abc} D_{ab}^{-\frac{3}{2}} (C_{bb} \vec{d}_a - C_{ab} \vec{d}_b) + D_{ab}^{-\frac{1}{2}} (\vec{d}_b \times \vec{d}_c) \\ \vec{\nabla}_{r_b} Z'_p &= -C_{abc} D_{ab}^{-\frac{3}{2}} (C_{aa} \vec{d}_b - C_{ab} \vec{d}_a) + D_{ab}^{-\frac{1}{2}} (\vec{d}_c \times \vec{d}_a) \\ \vec{\nabla}_{r_c} Z'_p &= D_{ab}^{-\frac{1}{2}} (\vec{d}_a \times \vec{d}_b) \\ \vec{\nabla}_{r_{a-1}} Z'_p &= C_{abc} D_{ab}^{-\frac{3}{2}} (C_{bb} \vec{d}_a - C_{ab} \vec{d}_b) - D_{ab}^{-\frac{1}{2}} (\vec{d}_b \times \vec{d}_c) \\ \vec{\nabla}_{r_{b-1}} Z'_p &= C_{abc} D_{ab}^{-\frac{3}{2}} (C_{aa} \vec{d}_b - C_{ab} \vec{d}_a) - D_{ab}^{-\frac{1}{2}} (\vec{d}_c \times \vec{d}_a) \\ \vec{\nabla}_{r_{c-1}} Z'_p &= -D_{ab}^{-\frac{1}{2}} (\vec{d}_a \times \vec{d}_b) \end{aligned}$$

The capability of Z'_p to detect A-DNA and B-DNA is analyzed by plotting both Z_p and Z'_p for several A- and B-DNA crystal structures used by Olson *et al.*[12]. Figure 2.2 shows an excellent correlation between Z_p and Z'_p (correlation coefficient is 0.96).

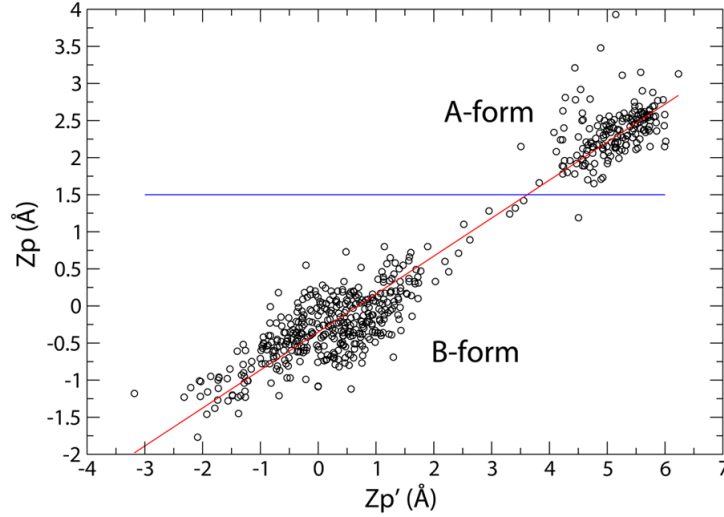


Figure 2.2 The correlation between Z_p and Z_p' values for all base steps of crystal structures of A- and B-DNA taken from Ref. [12].

Typically, $Z_p \geq 1.5$ Å was used by Olson *et al.* to ascribe B-form to A-form transition [12]. From this correlation plot, we ascribe the corresponding value of Z_p' to be 3.6 Å to define A-form. However, the correlation between static structures may not hold for small dynamic deformations during simulations. Therefore, even though Z_p' is implemented as a reaction coordinate, Z_p coordinate is used as an order parameter i.e., the parameter that indicates that the transition has taken place beyond its certain value. We use average Z_p value belonging to certain cluster as 1.5 Å to ascribe B- to A-form transition following Olson and coworkers' criteria [12]. Some studies used different values of Z_p , such as 1.3 Å by Bansal[18,19], 1.0 Å by Orozco[11], etc. to identify A-form occurred in B-A transition, due to the gap in Z_p values between A- and B-forms in various crystal structures. The value of Z_p in canonical A-DNA is 2.7 Å; the average value of Z_p in crystal structures of A-DNA is 2.2 Å[12]. Therefore, we have chosen average Z_p value as 1.5 Å to ascertain the B- to A-form transition amongst different prescriptions, which will provide the upper bound of the free energy cost for $B \rightarrow A$ transition. The analysis procedure carried out in this thesis is described in the next section.

2.4 Analysis Procedures

2.4.1. Cluster Analysis:

The calculations of the average DNA parameters, total energy, and other analyses were performed on the members of the biggest cluster of structures within a certain value of Z'_p to avoid large unphysical fluctuations. The schematic presentation of cluster analysis is shown in Fig. 2.3. First, the structures were collected within a particular value of Z'_p with 1 Å deviation (Green circles in Fig. 2.3). Subsequently, cluster analysis was performed using gromos[20] method implemented in the g_cluster utility of GROMACS[21] on the collected structures with 0.8 Å root mean square deviation (RMSD) cut-off such that the member of the biggest cluster is approximately 80% of the total number (Red circles in last step of Fig. 2.3). The values reported for the properties against Z'_p are the averages on the members of the biggest cluster.

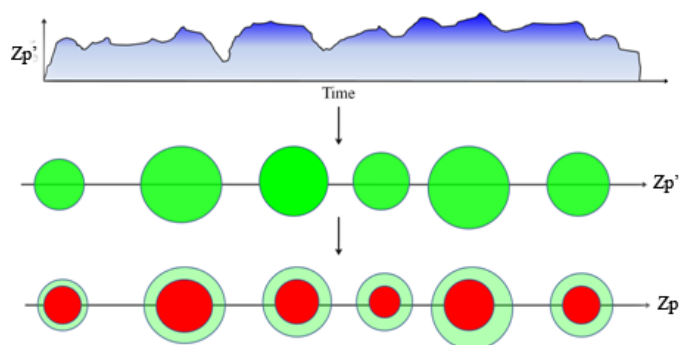


Figure 2.3 Schematic presentation of cluster analysis along reaction coordinate Z'_p .

2.4.2. Structure analysis:

We calculated changes in some of those DNA parameters as they vary along Z'_p . We analyzed both the base parameters (such as Slide, Roll, Twist) and backbone parameters (such as average δ_{avg} , χ_{avg} , and P_{avg}) of the sugar groups for the dinucleotide step undergoing transition. Earlier studies pointed towards the sequential nature of the transition through base parameters, Slide and Roll[22-24]. However, Roll has already been shown to be uncorrelated[18] with the transition. To probe into the role of both the backbone and base parameters, we introduce here a normalization function (**Eq. 2.11**) to capture the change in both the parameters together. For a property X, we have calculated the normalized value as X_{norm} ,

$$X_{norm}(Z'_p) = \frac{X(Z'_p) - X_{B-DNA}}{X_{A-DNA} - X_{B-DNA}}, \quad \text{Eq. 2.11}$$

where, $X(Z'_p)$ is a certain DNA parameter at a particular value of Z'_p . X_{A-DNA} and X_{B-DNA} denote the values of the DNA parameter at canonical A- and B-form, respectively. Therefore, any of the parameter in A-DNA will attain 1. At B-DNA, it will be 0 and between 0 and 1 during the transition. However, the values may go above one and below zero due to the unconstrained nature of the parameters.

2.4.3. Thermodynamic analysis:

We have calculated various interaction energies viz. stacking interactions, dinucleotide step-water interactions, and total potential energy along with the free energy. The total potential energy of the subsystem comprises the interaction between dinucleotide step and water molecules in the first hydration shell around it. Each dinucleotide has a different number of water molecules in their hydration shell. Therefore, to preserve the number of atoms in the subsystem, we took a maximum number of water molecules (46) found (through proximal radial distribution function[25]) for any dinucleotide step within 3.25 Å. (See Fig. 2.4)

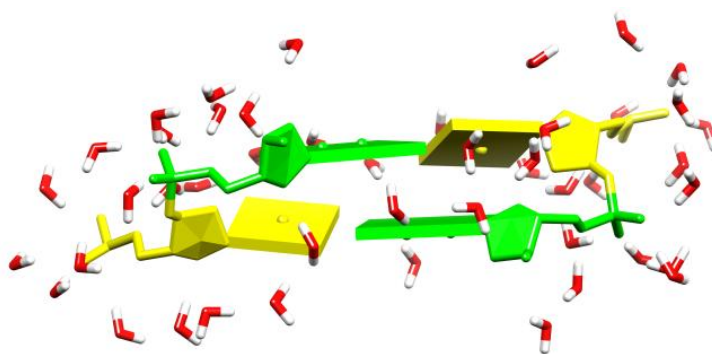


Figure 2.4 The dinucleotide step and first solvation shell to calculate the total potential energy.

The rationale for calculating the total energy in this way is two-fold. First, the fluctuation is much larger for the total energy of the system if all the atoms are considered. Hence, subsystem of interest was chosen. Secondly, this choice is based on the knowledge that each system is identical and unchanged except for the dinucleotide step. Therefore, the relative energy contribution from the complement of the subsystem is assumed to cancel out in the difference in the energy for a long

simulation. We have collected the configurations around a particular Z'_p and performed cluster analysis to further minimize the error.

2.5 Summary

We have discussed the brief details of umbrella sampling technique and development of reaction coordinate Z'_p . Z'_p correlates well with the order parameter Z_p and it successfully discriminates between A-form and B-form for static crystal structures. Thus, Z'_p is shown to be a reasonable reaction coordinate to drive local B-form to A-form transition in aqueous solution. The evolution of structural parameters or thermodynamic quantities is calculated along clusters defined around Z'_p values. The gromos method was implemented to find largest cluster and remove fluctuations. The free energy of all 10 unique dinucleotide steps along Z'_p are derived and presented in next chapter.

References:

- (1) Shamir, M.; Bar-On, Y.; Phillips, R.; Milo, R. *Cell* **2016**, *164*, 1302.
- (2) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. In *Proceedings of the 34th annual international symposium on Computer architecture*; ACM: San Diego, California, USA, 2007, p 1.
- (3) Gekko, K.; Sakaki, T.; Shindo, H. *Polym J* **1988**, *20*, 751.
- (4) Banavali, N. K.; Roux, B. *J. Am. Chem. Soc.* **2005**, *127*, 6866.
- (5) Ivanov, V. I.; Minchenkova, L. E.; Minyat, E. E.; Frank-Kamenetskii, M. D.; Schyolkina, A. K. *J. Mol. Biol.* **1974**, *87*, 817.
- (6) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275.
- (7) Kästner, J. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932.
- (8) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.
- (9) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2004**, *121*, 9186.
- (10) Krivov, S. V. *J. Phys. Chem. B* **2011**, *115*, 11382.
- (11) Noy, A.; Pérez, A.; Laughton, C. A.; Orozco, M. *Nucleic Acids Res.* **2007**, *35*, 3330.
- (12) Lu, X. J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819.
- (13) Lu, X. J.; Olson, W. K. *Nucleic Acids Res* **2003**, *31*, 5108.
- (14) Colasanti, A. V.; Lu, X. J.; Olson, W. K. *J Vis Exp* **2013**, e4401.
- (15) Dickerson, R. E. *J. Biomol. Struct. Dyn.* **1989**, *6*, 627.
- (16) Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X. J.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C. S.; Westhof, E.; Wolberger, C.; Berman, H. M. *J. Mol. Biol.* **2001**, *313*, 229.
- (17) Allen, P.; Tildesley, D. J. *Computer simulation of liquids*; Clarendon Press, **1987**.
- (18) Marathe, A.; Karandur, D.; Bansal, M. *BMC Struct. Biol.* **2009**, *9*, 24.
- (19) Marathe, A.; Bansal, M. *BMC Struct. Biol.* **2011**, *11*, 1.
- (20) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem. Int. Ed.* **1999**, *38*, 236.
- (21) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (22) Vargason, J. M.; Henderson, K.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7265.
- (23) Pastor, N. *Biophys. J.* **2005**, *88*, 3262.
- (24) Knee, K. M.; Dixit, S. B.; Aitken, C. E.; Ponomarev, S.; Beveridge, D. L.; Mukerji, I. *Biophys. J.* **2008**, *95*, 257.
- (25) Mukherjee, A. *J. Phys. Chem. Lett.* **2011**, *2*, 3021.

Chapter 3

Sequence-Dependent Free Energy Profiles of Localized B- to A-Form Transition of DNA in Water

3.1. Introduction

The 10 unique dinucleotide steps AA, AT, AC, AG, CA, CG, TA, GA, GC, and GG, possess heterogeneity at an intrinsic level. It arises from the variability of stacking energy, exposure of base atoms on major and minor groove as well as interactions with the solvent molecules. This intrinsic heterogeneity is a source of particular conformational preferences towards A-form or B-form in these steps. Previous studies have shown that, in general, AA, AT are B-philic steps and GG, AC, AG is A-philic steps [1,2].

The previous attempts to understand the sequence specificity at local level focused on the free energy scale of 10 unique dinucleotide steps or 32 trinucleotide steps. However, a comprehensive study of a reaction pathway of a local dinucleotide step undergoing B $\rightarrow$ A transition in water was not addressed.

The development of problem-specific reaction coordinate Z'_p (Section 2.3, Chapter 2), based on the unique order parameter Z_p to drive B- to A-form transition locally enabled us to calculate free energy profiles for all 10 unique dinucleotide steps embedded in a 12-mer DNA. This study is performed in aqueous environment so as to relate the results with biological process and to understand the role of water. We investigated the structural mechanism and thermodynamics of local B $\rightarrow$ A transition in water for various DNA sequences at a local dinucleotide step level. The results explain the possible role of water entropy in B-form stabilization in water and mechanistic heterogeneity of all 10 dinucleotide steps while transforming from B-form to A-form.

3.2 Computational details

3.2.1 Construction of the initial configuration

The basic DNA sequence is taken as $d(\text{CGCGA}\underline{\text{XY}}\text{TCGCG})_2$ based on Dickerson's dodecamer $d(\text{CGCGA}\underline{\text{ATT}}\text{CGCG})_2$ [3]. The middle base pair (BP) "XY" is changed to have the ten possible dinucleotide steps (AA, AC, AT, AG, CA, CG, TA, GA, GC, GG) to understand the transition at dinucleotide level. For example, to study "AA" we have used $d(\text{CGCGA}\underline{\text{AA}}\text{TCGCG})_2$ sequence. The reason for considering Dickerson dodecamer is that this sequence is essentially B-DNA. Therefore, we should get an upper bound for the conversion to A-form. Secondly, the sequence is capped with CG at the terminals, therefore, reducing the chance of fraying for reasonably long simulations[4]. We calculated Z'_p (and Z_p) values for the canonical A-form and B-form created using NAB[5]. The average values of Z_p and Z'_p for the equilibrated B-form structures calculated from 1 ns unrestrained simulations are presented in Fig. 3.1. Fig. 3.1 indicates that GG has higher Z_p/Z'_p value for equilibrated structure compared to other dinucleotide steps, indicating its A-philicity. Canonical A-form has fixed Z_p (2.73 Å) and Z'_p (4.31 Å) values. Canonical B-form has $Z_p=0.39$ Å and $Z'_p=0.7$ Å. While canonical A- and B-forms have fixed Z_p and Z'_p values, equilibrated B-form has different values for different dinucleotide steps, which reflect differential propensity towards A-form in the equilibrated B-form itself.

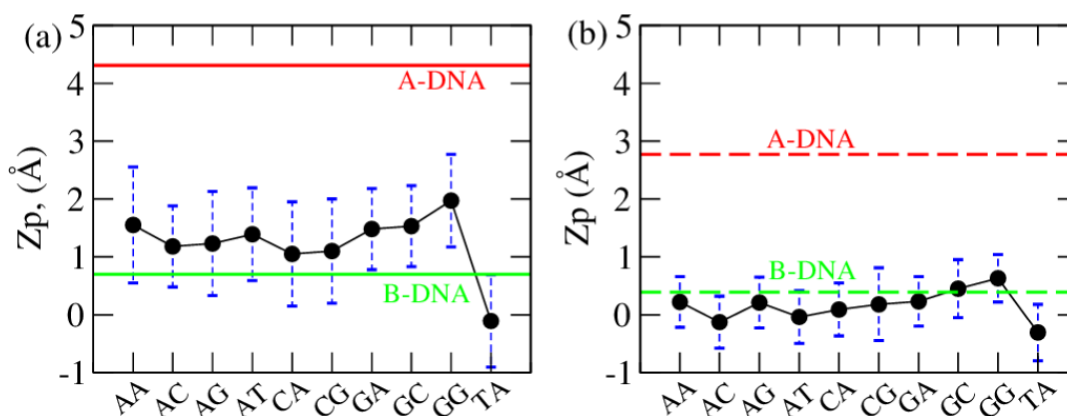


Figure 3.1 The values of (a) Z'_p and (b) Z_p for the ten different dinucleotide steps for equilibrated B-DNA. The values of Z'_p (solid lines) and (b) Z_p (dashed line) for canonical A-form and canonical B-form are presented by red line and green line, respectively.

3.2.2 Force-field and Equilibration.

All the simulations were carried out using GROMACS[6] by employing PBC conditions. Particle Mesh Ewald[7] was used for the coulombic interaction with long range cutoff at 10 Å. We used amber99 force-field[8] with parmbsc0 [9] modification, a well-established force field for B-DNA simulation[10]. The initial configuration of the canonical B-DNA was placed in a cubic box of side length 60 Å, subsequently solvated by TIP3P[11] water molecules, and 22 randomly placed Na^+ ions were used to neutralize DNA. The final system contained 756, 757 or 758 DNA atoms (depending on the sequence), 6847 water molecules, and 22 Na^+ ions. The final density of the system was 1006 kg/m^3 .

The canonical B-form was used as a starting structure for all the simulations. A series of standard minimization, position restraints, and equilibration step[4] was performed. Initially, each solvated DNA structure was minimized by applying 50 kcal/mol/Å^2 harmonic restraint potential on heavy atoms of DNA. The system at constant pressure was heated to 300 K over 100 ps using Berendsen thermostat and barostat[12], with a coupling constant of 0.4 ps, with a time step of 2 fs. The restraints were slowly reduced to zero during a series of energy minimizations and 50 ps equilibration steps. The final step was 1 ns NPT equilibration using Noose-Hoover thermostat[13] and Parrinello-Rehman barostat[14].

3.2.3 Umbrella Sampling Simulation Procedures.

With this equilibrated structures, a series of Umbrella Sampling[15] simulations was carried out for all ten DNA structures using GROMACS[6] with PLUMED software[16] after incorporating Z'_p reaction coordinate in PLUMED. The forces on the atoms generated from the harmonic potential on this coordinate are outlined in Section 2.4 of Chapter 2. The umbrella sampling simulations were carried out by varying Z'_p within the range of 0.7 Å (for canonical B-DNA) and 4.3 Å (for canonical A-DNA) initially with a step of 0.2 Å and extended whenever necessary. These simulations required approximately 25 simulations for each DNA structure, where each simulation was performed for 3 ns with a time step of 2 fs. Thus, the complete set of calculations amounts to $\sim 0.75 \mu\text{s}$ long. The length of the simulation was determined based on the convergence of the free energy profiles as discussed later.

3.3 Results and Discussion

3.3.1 Free energy, Energy and Entropy.

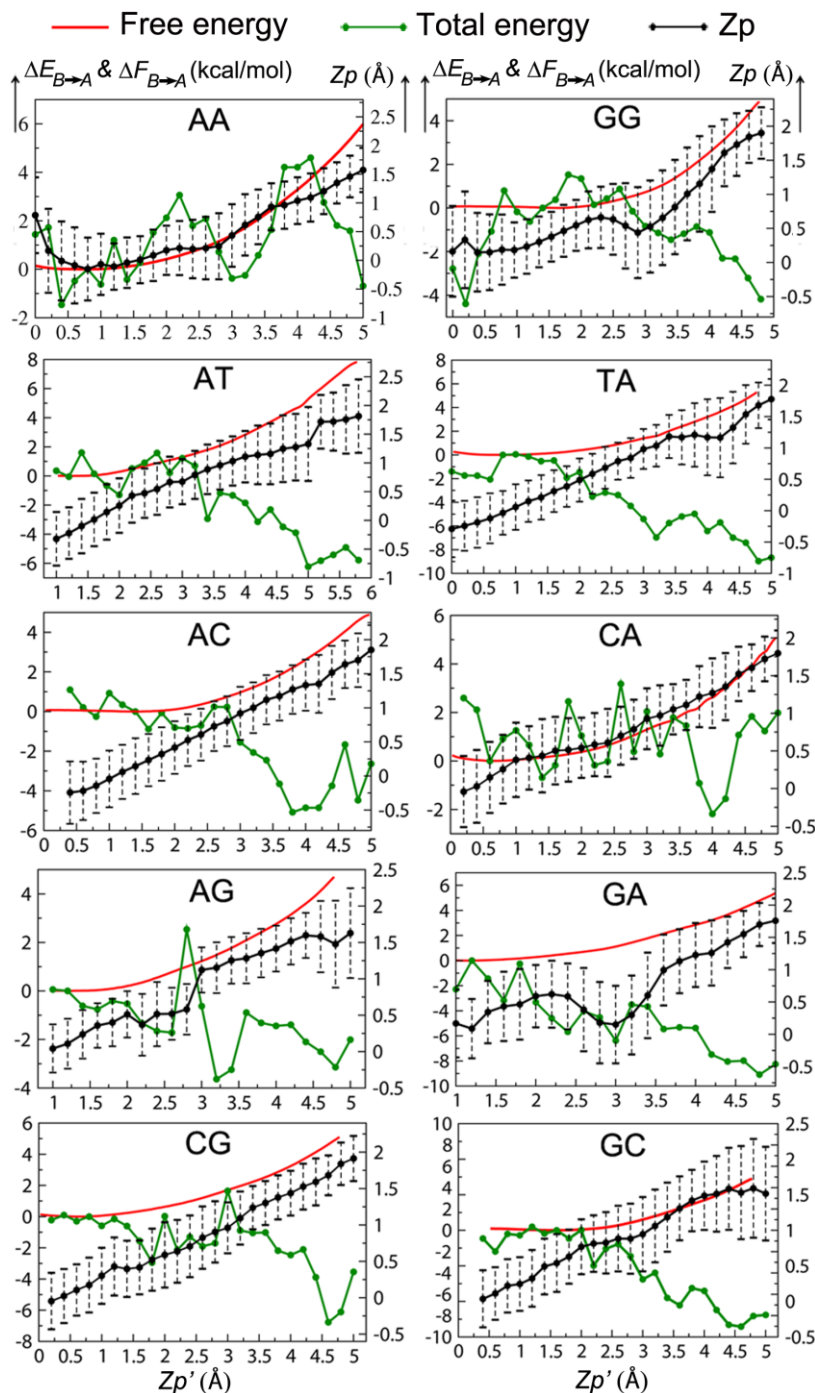


Figure 3.2 Potential of mean force (red line) for conversion from B- to A-form for all dinucleotide steps (shown in the legend) against Z_p' . The average values of the Z_p [17] (with its standard deviation) along Z_p' is also shown (black line and dots). Total potential energy variation for the subsystem containing a particular dinucleotide step with 46 closest water molecules is also shown. The error for total energy (green) is significant (15-25 kcal/mol). However, average values maintain a continuous

trend along Z'_p . For energy and Z_p calculations, cluster analysis was performed by collecting structures around each Z'_p . The values plotted here are averages over the members of the biggest cluster.

We performed umbrella sampling simulations (~25 simulations of 3 ns each for every system) to obtain free energy profiles for $B \rightarrow A$ conversion at the local dinucleotide level along Z'_p , shown in Fig. 3.2, for each of the ten dinucleotide steps embedded in the twelve base pair DNA. In the absence of any stable state in the A-form region, we ascribe the cost of $B \rightarrow A$ transition by the free energy at which average Z_p value attains 1.5 Å. This choice is governed by the reasons discussed before. Table 3.1 shows the free energy cost for all the dinucleotide steps for $B \rightarrow A$ transition. Free energy costs for all the dinucleotide steps follow the trend observed in experiments that AA and AT are B-philic whereas AC, GG, and GC are A-philic. GA appears to be intermediate rather than a B-philic step[1]. Therefore, even at the dinucleotide level, “A” and “B”-philicities are preserved. Protein-bound DNA systems show a significant proportion of CA step[1], which is A-philic according to calculation (relatively lower free energy cost).

Table 3.1: Free energy cost (kcal/mol) and standard deviations for $B \rightarrow A$ transition for ten dinucleotide steps measured at Z'_p value at which average Z_p is 1.5 Å.

Dinucleotide step	$Z'_{pZ_p=1.50}$	$\Delta F_{B \rightarrow A}$
AA	4.9	5.7 ± 1.2
AT	5.1	5.6 ± 1.3
AC	4.4	3.5 ± 0.9
AG	4.2	3.1 ± 1.1
CG	4.2	3.9 ± 0.9
GC	4.1	3.0 ± 0.8
GA	4.5	3.9 ± 1.5
CA	4.4	3.5 ± 1.4
TA	4.6	4.7 ± 1.1
GG	4.3	3.3 ± 0.8

Each free energy profile has converged within two ns indicating adequate sampling (see Fig. A-3.1 in Annexure 3). Note that; all the ten systems are identical except for the difference in the pair of base pairs

sequence of the DNA in the middle. Therefore, the free energy profiles obtained here should reflect the free energy associated with the B $\rightarrow$ A transition at the local dinucleotide level.

Since A-form DNA is not stable in water, the free energy profiles do not show any stable region for any of the dinucleotide steps. However, all the free energy profiles attain a stable minimum at the B-form (see Fig.3.2). To understand the free energy cost for A-form formation, we also plot the average Z_p values, calculated using 3DNA software[18] along Z'_p in Fig. 3.1. Although Z'_p correlates well with Z_p for static structures (Fig. 2), the correlation is not expected to be good where structural deformation takes place (note, that Z_p uses rigid canonical base pair geometry[18]). For Purine-Purine steps (e.g., AA, GG, GA, AG), Z_p does not grow linearly in the beginning. However, the growth is linear for all the dinucleotide steps at higher values of Z'_p . This observation ascertains the use of Z'_p as a reaction coordinate for studying the local deformation in the DNA.

3.3.2 Energetic Analysis and Entropy Interpretation

Along with the free energy, we calculated the total potential energy of the subsystem comprising the dinucleotide step and water molecules in the first hydration shell of the dinucleotide step. Each dinucleotide has a different number of water molecules in their hydration shell. Therefore, to preserve the number of atoms in the subsystem, we took a maximum number of water molecules (46) found (through proximal radial distribution function[19]) for any dinucleotide step within 3.25 Å. This approach allows us better comparison of total energy as each system is identical except for the dinucleotide step. Fig. 3.2 shows the change in the relative total energy of the above subsystem against Z'_p . Therefore, for a long enough simulation, the relative energy contribution from the complement of the subsystem is assumed to cancel out in the difference of the energy, giving rise to an energy profile similar to that captured in Fig. 3.2 through the energy of the subsystem. To further minimize the error, we have collected the systems around a particular Z'_p and performed cluster analysis (see Section 2.4.1 of Chapter 2). The plot of the subsystem's total potential energy in Fig. 3.2 shows mild change up to a certain value of Z'_p beyond which energy decreases for all dinucleotide steps. This indicates that the localized A-form created using Z'_p is energetically more stable than the corresponding B-form in water.

This result is unprecedented as the interpretation of Fig. 3.2 indicates that B-form in water is stable perhaps due to entropy. Previous studies primarily pointed towards the contribution of DNA electrostatics[20] and counter-ion condensation[21] for B-DNA stability.

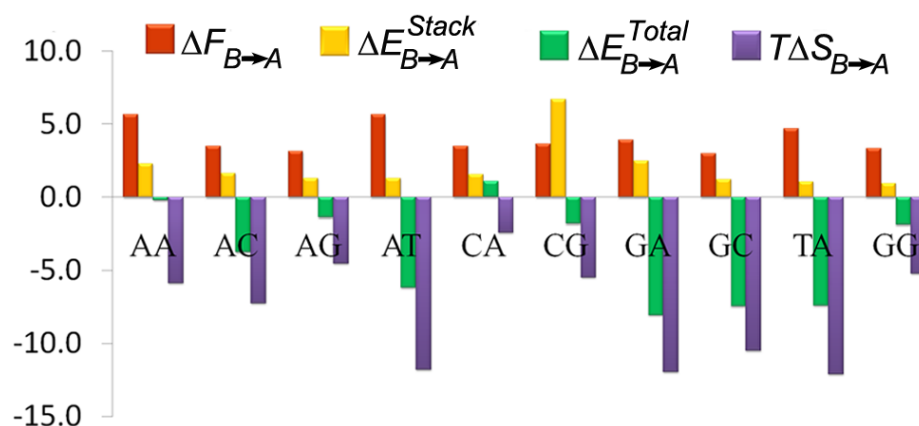


Figure 3.3. Free energy, total (potential) energy, stacking energy, and entropy (multiplied by 300K) difference (kcal/mol) for conversion of B $\rightarrow$ A (A-form minus B-form) for all ten different dinucleotide steps. Data obtained by subtracting the respective values at average $Z_p=1.5$ (A-form) from the values where free energy is minimum (stable B-form). Total energy denotes the energy of the subsystem consisting of dinucleotide steps and 46 closest water molecules. Entropy is estimated by taking the difference of free energy from the total energy.

The result is further summarized in Fig. 3.3 where we plot the difference in free energy, total energy, entropy, and stacking energy (interaction energy between two different base pairs including sugar-phosphate[22,23]) from B $\rightarrow$ A to understand the contribution of each in the process. Due to less number of counter-ions, we did not include their energetic contribution in the total energy. However, earlier studies show that counter-ions make A-form more energetically stabilized[20]. Therefore, including ions will further strengthen the trend observed here.

Note that, previous studies mostly focused on the DNA energetic, which showed that the transition from B $\rightarrow$ A is opposed by DNA electrostatics[20]. It is evident from Fig. 3.3 that indeed the stacking energy opposes the transition, as the cost is positive for all the ten dinucleotide steps. CG has the largest stacking energy penalty.

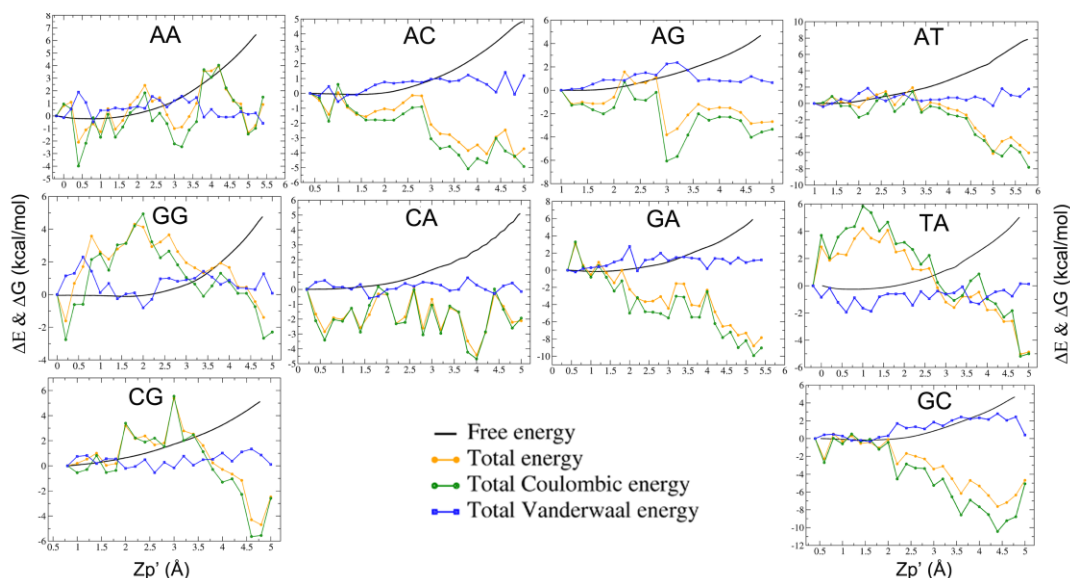


Figure 3.4. Electrostatic (green) and van der Waals (blue) components of the total energy (orange), along with the free energy plot (black). Note that the electrostatic contribution mainly drives the total energy change. The general range of error for total energy is 15-25 kcal/mol, van der Waals energy (10-15 kcal/mol) and electrostatic energy (15-25 kcal/mol)

Decomposition of the total energy shows that 90% of the contribution is electrostatic in nature (see Fig. 3.4). However, the total potential energy comes out to be favorable to form the local A-form, thus implying a negative entropy contribution for the transition as shown in Fig. 3.3. Entropy contribution is calculated by subtracting free energy from the total energy. Note that; any energy calculation is associated with a significant error[24]. Here, cluster analysis helps to reduce error. Yet, the error is large enough to debate about entropy contributions presented in Fig. 3.3. However, variation in the average energy along Z_p' is rather smooth. The trend in energy and entropy is unavoidable as it appears in every dinucleotide step. Therefore, for the localized transformations from A- to B-form in water may be driven by entropy.

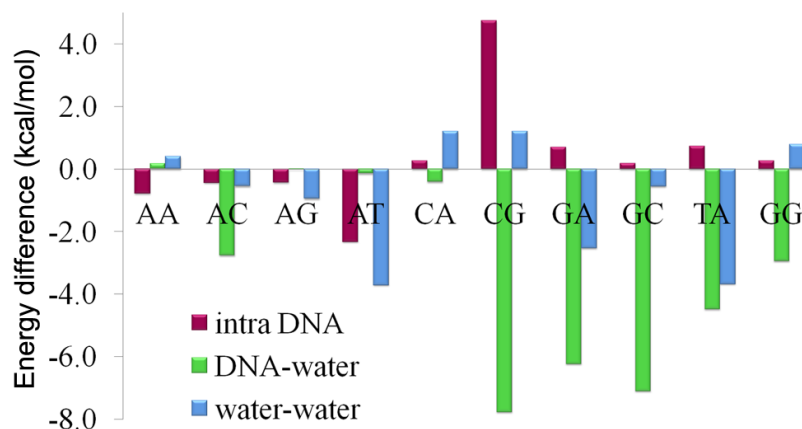


Figure 3.5 Decomposition of the total energy difference for B → A transition for all dinucleotide steps into various components of differences in intra-DNA, DNA-water, and water-water interactions.

To understand the origin of the energetic stability of the A-form, we decomposed the total energy difference into DNA-water, water-water and intra-DNA contributions as shown in Fig. 3.5. Intra-DNA interactions consist of self-interactions (both van der Waals and electrostatic) of base pairs involved in the transition (base pair 6 and 7) including the interactions between base pairs (i.e. stacking energy). DNA-water interaction consists of the interaction of middle base pair under study and closest water molecules. For A-tract dinucleotide step (AA, AC, AG, AT), the intra-DNA interaction (part of which is stacking energy) is stabilizing, whereas it is destabilizing for rest of the steps. Surprisingly, for all the dinucleotide steps, either DNA-water or water-water interaction is stabilizing in the A-form pointing to the source of an overall negative energy change towards A-form in water (Fig. 3.5). CG is also an outlier here showing significant intra-DNA energy cost for the formation of A-form compensated, however, by larger DNA-water stabilization. Variation of different energy components along Z'_p is shown in Fig. A-3.2 of Annexure 3.

Therefore, as expected, water is probably an important driving factor for the transition, *albeit* in an unintuitive way as our study shows that water energetically stabilizes the A-form. Therefore, water may contribute to the entropic stabilization of the B-form.

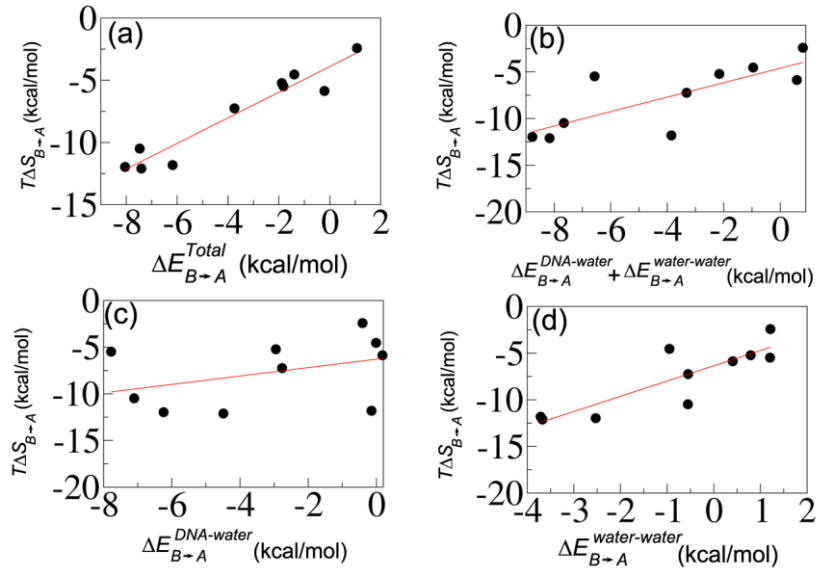


Figure 3.6. Correlation between entropy difference and different energy components for B- to A-form transition (a) correlation with total energy. Correlation coefficient (CC) is 0.96 (b) correlation with solvent energy (DNA-water and water-water) contribution (CC=0.78) (c) correlation with DNA-water interaction energy (CC=0.39) (d) Correlation with water self-interaction energy (CC=0.86).

To ascertain the role of water in the overall entropy change, we first plot the entropy and total energy changes in Fig. 3.6(a) for all dinucleotide steps. It shows an excellent correlation, which indicates that the enthalpy-entropy compensation prevails in the present system of DNA structural changes as well. After that, we plot the entropy cost with only solvation energy cost (involving DNA-water and water-water interaction energy) in Fig. 3.6(b). Apart from two dinucleotide steps (CG and AT), correlation is still acceptable. Further decomposition of solvent energy to DNA-water (Fig. 3.6(c)) and water-water (Fig. 3.6(d)) shows that the water-water interaction energy is highly correlated with the total change in entropy. This correlation ascertains that water may play an important part in the entropic destabilization of A-form compared to the B-form. An indirect corroboration is found from the recent study, where restraining of some water molecules in the major groove was shown to lead the transition of DNA from B $\rightarrow$ A [25].

3.3.3 Structural Analysis and Mechanism.

We have demonstrated above that Z'_p as the reaction coordinate successfully captures the B- to A-form transition at the local dinucleotide level measured in terms of the standard order parameter Z_p . However, other conventional markers show significant differences in the above two forms of the DNA. We have normalized the

structural parameters as discussed in Section 2.4.2 of Chapter 2. In general, we have used following Eq. 3.1 so that all values become 0 for B-form and 1 for A-form.

For a property X , we have calculated the normalized value X_{norm} defined as,

$$X_{norm}(Z'_p) = \frac{X(Z'_p) - X_{B-DNA}}{X_{A-DNA} - X_{B-DNA}}, \quad \text{Eq. 3.1}$$

where $X(Z'_p)$ is a certain DNA parameter at a particular value of Z'_p . X_{A-DNA} and X_{B-DNA} denote the values of the DNA parameter at canonical A- and B-form, respectively.

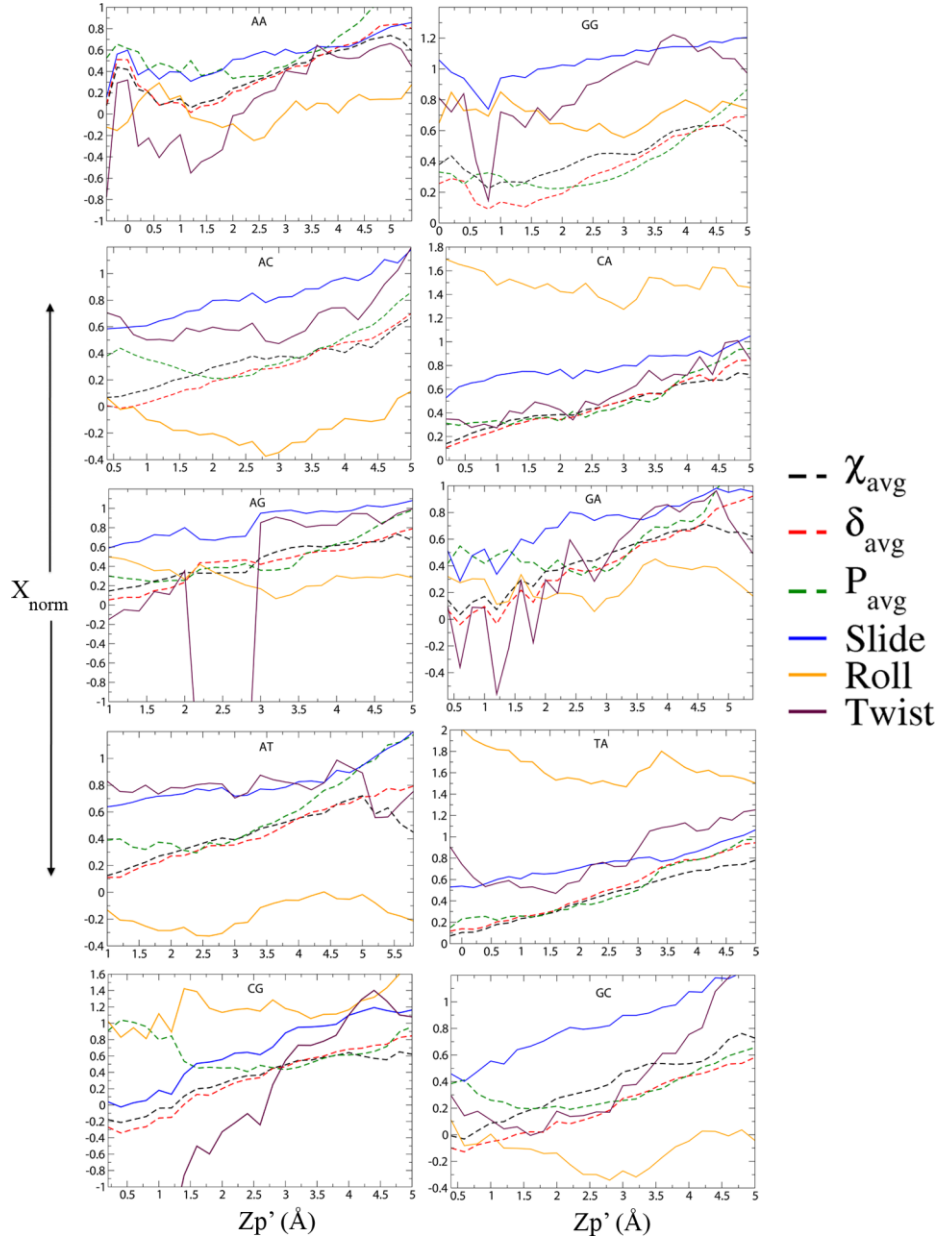


Figure 3.7. Normalized DNA structure parameters for ten dinucleotide steps from B $\rightarrow$ A. Backbone parameters are drawn using dashed lines while base parameters are drawn using solid lines. Each dinucleotide step is indicated in the legend.

Figure 3.7 shows the variation in the normalized values of the DNA parameters for all the ten different dinucleotide steps, which suggests that there is no general trend for conversion from B $\rightarrow$ A at the dinucleotide level. RR steps (AA, GG, GA, AG) show that Slide, Twist, and Pseudorotation angle (P_{avg}) are involved in B $\rightarrow$ A transition. For RY steps (AC, AT, GC), except for AT, Slide and Twist achieve A-form values. Surprisingly, AT shows that only Slide and P_{avg} achieve A-form values. The individual variations in torsional parameters χ , δ , Pseudorotation angle (P), and base step parameters Roll, Slide, and Twist are provided in Fig. A-3.3 to A-3.8 of Annexure 3. In general, we can observe that motions of backbone parameters of residue 6 and 19 are correlated with base step parameters of these residues.

For a given Z'_p value, Slide, in general, is larger than other parameters (See Fig. 3.7) indicating that Slide may be the first parameter to change in all the ten dinucleotide steps. For all base pair steps, except for AA, Slide values are between the canonical B- and A-forms in equilibrated B-DNA itself (Fig. A-3.7 in Annexure 3). For AA step Slide is close to the canonical B-form, while for GG (most “A”-philic) Slide is close to canonical A-form. YR steps (CA, CG, TA), on the other hand, maintain high Roll angle values from B $\rightarrow$ A transition indicating Slide and Twist-mediated B $\rightarrow$ A transition. This observation supports Olson’s suggestion[26] that YR steps are more easily deformable. Backbone parameters (except P_{avg} for certain dinucleotide steps), on the other hand, show almost linear variation between canonical B- and A-forms.

To understand whether the local deformation caused at the central dinucleotide step propagate to other base steps, we analyzed all base steps except terminal CG’s for Z_p and Slide along our reaction coordinate Z'_p . Figure A-3.9 and A-3.10 in Annexure 3 show the variation of Slide and Z_p for two flanking base pairs along Z'_p . We observed that only two base steps are at most affected due to the deformation. This observation is not surprising as Fig. 3.2 shows that A-form for all dinucleotide steps is free energetically unstable compared to B-form. Therefore, natural propagation to a complete A-form in water is not to be expected.

In general, we observe that the concerted changes in P_{avg} , χ_{avg} , and δ_{avg} along with Slide cause B $\rightarrow$ A transition. Even though in some cases P_{avg} , χ_{avg} , and δ_{avg} do not achieve A-form values, their continuous variation towards it indicates a

concerted mechanism at least at the dinucleotide level. Calladine and Drew[27] observed that the backbone-switching during B-A transition accommodates a continuum of base step conformations in terms of Slide, Roll and helical Twist. El Hassan and Calladine [28] studied the conformational map of dinucleotide steps and divided them into Rigid (AA, AT, GA), Bistable (GG, GC, CG, AC) and Flexible (CA, TA) categories according to their Roll, Slide and helical Twist values. They observed that homogeneous G/C steps are bistable and their interactions cause a base step to adopt low slide values, which in turn cause backbone to attain A-DNA conformation. Our study again reinforces the point mentioned by Marathe *et al.*[1] that mainly Z_p and to some extent Slide and Pseudorotation phase angles, δ , and χ are better indicators of A- and B- conformations at the dinucleotide level .

3.4 Comparison with experiment

Free energy cost per dinucleotide step obtained here may seem to be higher than that normally observed for the same estimated from whole DNA conversion either theoretically[29], or experimentally[2,30]. Tolstorukov and coworkers[2] proposed the following model to derive the free energy cost for B $\rightarrow$ A transition for each dinucleotide step from the experimental estimates of several DNA sequences[2],

$$\Delta G_{BA} = \frac{RT}{Q} n(a - a_t) = \sum_{I \in \text{all steps}} \Delta G_{BA}(I) + 2\Delta G^s, \quad \text{Eq. 3.2}$$

where, ΔG_{BA} and $\Delta G_{BA}(I)$ are the free energy costs for whole DNA and each base pair conversion. R is universal gas constant, T is the temperature, n is the number of base pairs in the DNA, a is the water activity and a_t is the apparent water activity (AWA) where free energy cost is zero. Q is a coefficient based on B $\rightarrow$ A transition width (and also depends on cooperative length[31]) and taken to be 14.5% relative humidity (RH). By calculating AWA for various DNA sequences, ΔG_{BA} was estimated and subsequently fitted to obtain the free energy cost for each dinucleotide steps $\Delta G_{BA}(I)$. ΔG^s is the cost for the terminal base pair due to interaction with solvent and taken to be fixed value 1.5 kcal/mol.

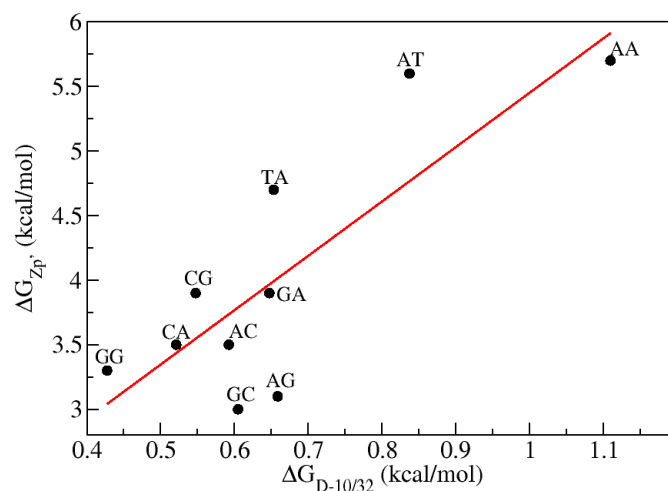


Figure 3.8. Correlation between experimentally derived D-10/32 model[2] and the present estimate of the free energy cost for B- to A-form transition for dinucleotide steps. The correlation coefficient is 0.81.

The values obtained through above procedure provided the free energy cost within the range of 1.39 to 0.97 kcal/mol for AA to 0.89 to 0.19 kcal/mol for GG steps (and similar ranges for other dinucleotide steps). Although there are minor deviations in different fitting models, the values are much smaller than our theoretical estimate. However, the correlation between these experimental estimates and our calculation is very high as shown in Fig. 3.8, where our results are compared with one of the best experimental models D-10/32 in Ref. 2. Therefore, our calculations capture the trend in the free energy cost of each dinucleotide at par with experimental estimates, *albeit* with a higher estimate.

To test whether the origin of the discrepancy comes from the choice of our collective variable, we have also tried another local coordinate such as average δ angle to convert a certain dinucleotide step to A-form. However, even at a free energy cost 20 kcal/mol, the conversion to A-form was only 15% as discussed in Annexure 3 and shown in Fig. A-3.11 in Annexure 3. Therefore, the choice of Z_p' for localized B $\rightarrow$ A transition seems reasonable.

The discrepancy in absolute estimate, however, can be resolved based on the following observations. The first thing to note is that B $\rightarrow$ A transition is cooperative in nature[25,32]. Therefore, conversion of the first basic unit would cost the most. However, the cooperative unit is estimated to be around 10 base pairs[33], making our estimate huge for the conversion of a dodecamer. Therefore, cooperativity may not be a sufficient criterion. To find the other reasons, we noted that the significant difference between experimental estimate mentioned above and our theoretical

approach is that we created a A-form unit in the middle of a B-DNA. This A-form, in turn, creates one B/A and one A/B junction on both sides. Experimental estimates of B/A junction vary from 1.4 kcal/mol to 2.1 kcal/mol.[31] If we subtract 2.1 kcal/mol as the cost of each B/A (or A/B) junction from our free energy calculation, it comes to 1.5 kcal/mol. Taking B/A junction for GG to be 1.4 kcal/mol, the estimate for GG transition comes to 0.5 kcal/mol. Both the above estimates are close to experimental values. The higher estimate may be obtained due to various other factors such as the choice of the high value of the order parameter (1.5 Å), selection of the DNA sequence (we used Dickerson dodecamer which is known to be strongly B-philic[3]), minimal ion concentration, etc. However, with a variety of assumptions in experimental model and approximations in theoretical force-field, results are still in excellent agreement with each other. Note that the free energy cost estimated here will be more appropriate for the cost of protein-DNA and drug-DNA interaction causing local A-form[17], which will invariably suffer from the extra cost of B/A junction.

3.5 Conclusion

This study captures the B- to A-form transition in water through the most basic unit of DNA – a dinucleotide step. It depicts the sequence dependent effect of the transition in molecular details through rigorous free energetic method carried out for all the ten dinucleotide steps using a new reaction coordinate Z'_p , capable of driving local B $\rightarrow$ A transition, and unique local order parameter Z_p . High correlation with experimentally obtained values for the free energy costs of different dinucleotide indicates that “A” or “B” philicity is preserved even at the dinucleotide level. The higher estimate for the free energy cost compared to experiment is attributed to the formation of B-form/A-form junctions in addition to the formation of a local A-form. Our current estimate agrees well with experiments after subtracting the cost of B/A junction. Therefore, our calculation of the free energy cost is similar to that caused by local deformation of the DNA to A-form by proteins and drugs[17].

Prevalence of dinucleotide sequences of lower free energy cost (such as GG, AG, AC) is often observed in those regions[17]. We also find that the overall energy is favorable towards the local A-form, although intra-DNA electrostatics opposes the transition from B $\rightarrow$ A as shown earlier[20]. This A-DNA stabilization in water and

positive free energy indicates that entropy opposes $B \rightarrow A$ transition. In other words, the origin of B-DNA stability in water seems to be dictated by entropy. This entropic stabilization shows that at a lower temperature or in the crowded environment, A-form will be more preferred, as demonstrated by a Raman spectroscopic study on poly(dG).poly(dC)[34]. We found that water-water interaction energy correlate with the entropy change, thereby indicating the role of water in the entropy reduction in $B \rightarrow A$ transition. However, this needs further investigation.

We also proposed a new analysis of the mechanism of transition including both backbone and base parameters. The general mechanism of $B \rightarrow A$ transition at local level involves concerted motions of backbone parameter, however, triggered by base parameter Slide. In the light of protein-DNA interactions where local, base pair level transitions occur in DNA, this study provides detailed molecular insight of the sequence-dependent effect of those local transitions in the DNA. The free energy values obtained here are influenced by the formation of B-form/A-form junctions on both 5' and 3' side. The details of these B/A junction effects are discussed in next chapter.

References:

- (1) Marathe, A.; Karandur, D.; Bansal, M. *BMC Struct. Biol.* **2009**, *9*, 24.
- (2) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, *81*, 3409.
- (3) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 2179.
- (4) 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.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. *Biophys. J.* **2004**, *87*, 3799.
- (5) Macke, T. J.; Case, D. A. *Molecular Modeling of Nucleic Acids* **1998**, 682, 379.
- (6) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (7) Tom Darden, D. Y., Lee Pedersen *J. Chem. Phys.* **1993**, *98*, 10089.
- (8) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (9) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.
- (10) Beveridge, D.; Cheatham, T., III; Mezei, M. *J. Biosci.* **2012**, *37*, 379.
- (11) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (12) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. V.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (13) Nose, S. *Mol. Phys.* **1984**, *52*, 255.
- (14) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182.
- (15) Torrie, G. M.; Valleau, J. P. *J. Comp. Phys.* **1977**, *23*, 187.
- (16) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. *Comput. Phys. Commun.* **2009**, *180*, 1961.
- (17) Lu, X. J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819.
- (18) Lu, X. J.; Olson, W. K. *Nucleic Acids Res* **2003**, *31*, 5108.
- (19) Mukherjee, A. *J. Phys. Chem. Lett.* **2011**, *2*, 3021.
- (20) Pastor, N. *Biophys. J.* **2005**, *88*, 3262.
- (21) Jayaram, B.; Sprous, D.; Young, M. A.; Beveridge, D. L. *J. Am. Chem. Soc.* **1998**, *120*, 10629.
- (22) Sasikala, W. D.; Mukherjee, A. *J. Phys. Chem. B* **2012**, *116*, 12208.
- (23) Šponer, J.; Berger, I.; Špačková, N. a.; Leszczynski, J.; Hobza, P. *J. Biomol. Struct. Dyn.* **2000**, *17*, 1.
- (24) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.
- (25) Knee, K. M.; Dixit, S. B.; Aitken, C. E.; Ponomarev, S.; Beveridge, D. L.; Mukerji, I. *Biophys. J.* **2008**, *95*, 257.
- (26) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163.
- (27) Calladine, C. R.; Drew, H. R. *Journal of Molecular Biology* **1984**, *178*, 773.
- (28) ElHassan, M. A.; Calladine, C. R. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences* **1997**, *355*, 43.
- (29) Banavali, N. K.; Roux, B. *J. Am. Chem. Soc.* **2005**, *127*, 6866.

- (30) Basham, B.; Schroth, G. P.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6464.
- (31) Ivanov, V. I.; Minchenkova, L. E.; Burckhardt, G.; Birch-Hirschfeld, E.; Fritzsche, H.; Zimmer, C. *Biophys. J.* **1996**, *71*, 3344.
- (32) Malenkov, G.; Minchenkova, L.; Minyat, E.; Schyolkina, A.; Ivanov, V. *FEBS Lett.* **1975**, *51*, 38.
- (33) Ivanov, V. I.; Krylov, D. Y.; Minyat, E. E.; Minchenkova, L. E. *J. Biomol. Struct. Dyn.* **1983**, *1*, 453.
- (34) Nishimura, Y.; Torigoe, C.; Tsuboi, M. *Nucleic Acids Res.* **1986**, *14*, 2737.

Chapter 4

B-form/A-form Junction Free Energy in DNA

4.1 Introduction

4.1.1 B/A junction and its biological significance:

The protein-DNA interactions induce B-form to A-form transition at the local level in DNA. This local A-form induction has a penalty of free energy for the junction formation at B-form/A-form interface. In 1993, Jacobo-Molina *et al.*[1] observed the A-form/B-form junctions (A/B junction or B/A junction) in HIV-1 reverse transcriptase and DNA complex. A significant bend of $40^\circ - 45^\circ$ was observed at A/B junction, and hence authors hypothesized the plausible role of A/B junction in reverse transcriptase fidelity, translocation, and other mechanisms. A crystallographic analysis study[2] observed the presence of B/A junctions commonly involved in local B $\rightarrow$ A transition.

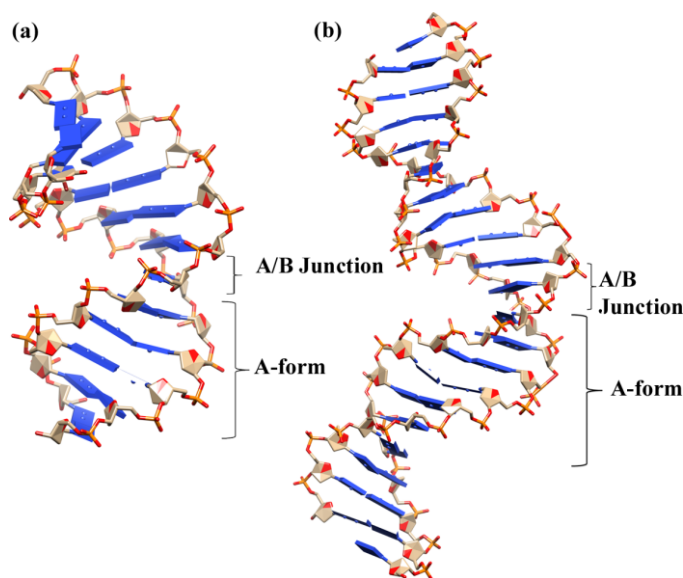


Figure 4.1. The local A-form and A-form/B-form junction in case of (a) cisplatin-DNA complex (PDB ID: 1AIO) and (b) DNA-endonuclease complex (PDB ID: 1IPP).

The bending at B/A junctions depends on the binding protein or drug molecule. For example, cisplatin-DNA complex [3] involves a smooth global bend at A/B junction and affects the two dinucleotide steps below A-form region (See Fig. 4.1(a)).

The I-PpoI homing endonuclease [4,5] interacts with DNA at two binding sites and creates a superhelical bend of 25° at each location, thus increasing overall global bend up to 55° due to two B/A junctions. Hence, the presence of B/A junctions causes local deformations, beneficial for protein binding via exposure of base atoms or minor groove widening [2]. A previous study[6] indicated that the B/A junction in contiguous DNA segment influences a B→A transition by affecting hydrodynamic properties of a sequence at the junction. An experimental study on 5S Ribosomal RNA gene proves that the sequence specific behavior of DNA causes B/A junction formation at a small fragment of CCTT[7] and prolongs B→A transition on B-side of the junction. Ivanov and co-workers experimentally detected stable B/A junction and its free energy and estimated the junction energy to be ~2.1 kcal/mol at CT/AG step[8].

4.1.2 Contribution of junction energy in $\Delta G(Z'_p)$

Recently, we studied local B-form to A-form transition in water using a designed reaction coordinate Z'_p (inspired from structural parameter Z_p [2]). In our previous study[9], we performed umbrella sampling simulations and calculated free energy profiles for B-form → A-form conversion for all 10 different dinucleotide steps embedded in a DNA. Although our study captured the B- and A-philicity of the dinucleotide steps in right trend as obtained through experimental model of Tolstorukov and co-workers[10], our calculated values were higher. This led us to realize that creating A-form locally in a B-DNA also creates two B/A junctions. Experimental estimate of the B/A junction energy can vary from 1.4 to 2.1 kcal/mol depending on the base steps. [8,11-13] Therefore, creating an A-form locally in a B-DNA (as typically done by proteins[2]) is more costly.

In this study, we attempted to calculate the B/A junction free energy and its conformational preference based on the rigorous free energy estimates for B→A conversion. By careful choice of DNA sequences as mentioned in computational details, we estimated the absolute free energy of a dinucleotide base step, which does not include the junction free energy cost. We show that our absolute values can capture the known “A” and “B” philicity of dinucleotide steps. The junction energy values also predict the expected cooperative length of B-A transition. The brief computational details are described in the next section.

4.2 Computational Details

4.2.1 Reaction coordinate: Z'_p and $\overline{Z'_p}$

We learnt from our previous study[9] that free energy along Z_p is indeed a better coordinate compared to other available choices[14]. Fig. 4.2 shows the construction of Z'_p again, which measures the distance between phosphorous atoms of two strands along the local DNA axis. The Z'_p coordinate, however, is designed for a single dinucleotide step. This limits use of Z'_p to study one-dimensional free energy profiles for multiple dinucleotide steps (such as trinucleotide steps used here). Therefore, we used average Z'_p ($\overline{Z'_p}$) in the systems where more than one dinucleotide steps are converted from B- to A-form.

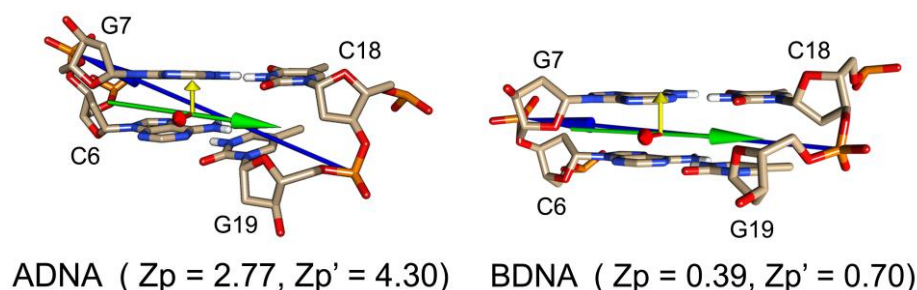


Figure 4.2. Schematic presentation of Z'_p indicating vectors. Red vector ($\vec{x}$) is from the centre of mass (COM) of four sugar rings to the COM of two sugars at the 3'-end. Green vector ($\vec{y}$) is from the COM of two C1' atoms of the sense strand to the COM of the two C1' atoms of the antisense strand. Cross product of $\vec{x}$ and $\vec{y}$ gives the yellow vector ($\vec{z}$) along the DNA axis. Finally, a vector $\vec{p}$ is created from the phosphorous atoms joining the antisense strand bases to that of the sense strand bases Z'_p is the projection of $\vec{p}$ on the unit vector $\hat{z}$. Figure is plotted using UCSF Chimera[15].

4.2.2 The design of DNA sequences.

With four possible nucleotides, there can be ten unique dinucleotide steps and 32 unique trinucleotide steps. Therefore, the design of the sequences in this study is motivated to minimize the number of computationally expensive free energy calculations to obtain B→A conversion free energy of all the ten dinucleotides and associated junction values. Based on the general Dickerson-Drew dodecamer[16] sequence d(CGCGAATTCGCG)₂, we have designed 12 DNA sequences of 13 base pairs each so that we get 10 different dinucleotide steps with only homonucleotide junctions. The general sequence can be described in general by d(CGCGXYYCYCGCG)₂, where X/Y can be either A, T, C, or G. Reason for keeping the CG sequences in both termini is to reduce the chances of fraying at the

ends[17]. Table 4.1 lists all the sequences. The corresponding dinucleotide (bold and underlined) and trinucleotide (underlined) steps that are transformed from B→A are underlined. The second column lists dinucleotide step studied, and the third column lists trinucleotide steps studied in this chapter. We will subsequently explain how this design helped us to obtain the absolute free energy values (without junction effect) for all ten unique dinucleotide steps. Note that, in our previous study[9] all ten dinucleotide steps were flanked by different steps. Thus, we also have a chance to compare the same dinucleotide free energy for B→A conversion in a different environment.

Table 4.1. List of DNA sequences used to perform Umbrella sampling simulation.

Sequence	Dinucleotide step	Trinucleotide step
CGCG <u>AAAA</u> CGCG	AA	AAA
CGCG <u>GGGG</u> CGCG	GG	GGG
CGCG <u>TTTT</u> CGCG	TT	TTT
CGCG <u>CCCC</u> CGCG	CC	CCC
CGCG AA <u>ACC</u> CGCG	AC	----
CGCG AA <u>AGG</u> CGCG	AG	----
CGCG AA <u>ATT</u> CGCG	AT	----
CGCG G <u>GAAA</u> CGCG	GA	----
CGCG C <u>CGGG</u> CGCG	CG	----
CGCG G <u>GCCC</u> CGCG	GC	----
CGCG T <u>TAAA</u> CGCG	TA	----
CGCG C <u>CAAA</u> CGCG	CA	----

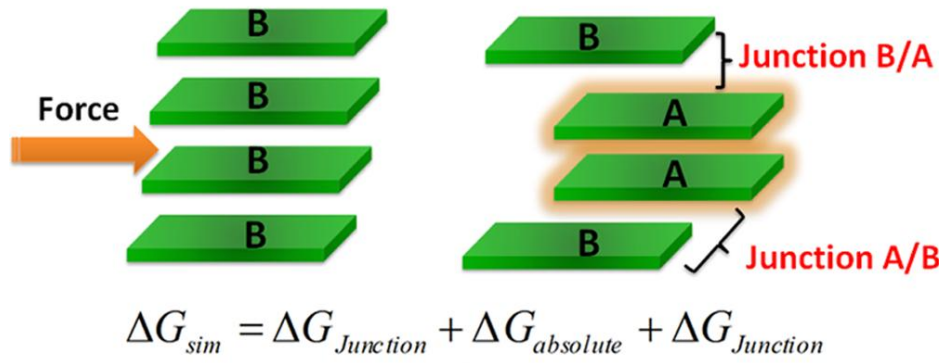


Figure 4.3. Schematic presentation of junction free energy.

The free energy obtained along Z'_p for the conversion of XY step to A-form in the sequence $d(\text{CGCGX}\underline{\text{XYY}}\text{CGCG})_2$ can be decomposed into junction free energy and absolute energy (See Fig. 4.3) as shown in Eq. 4.1 below,

$$\Delta G_{sim}(XY) = \Delta G_j(XX) + \Delta G_a(XY) + \Delta G_j(YY), \quad \text{Eq. 4.1}$$

Where, $\Delta G_{sim}(XY)$ is obtained from free energy simulations. $\Delta G_j(XX)$ and $\Delta G_j(YY)$ are the free energy costs for creating a B/A junctions created on either side of a local A-form in a B-DNA. $\Delta G_a(XY)$ is the absolute estimate of the free energy for conversion of a local XY dinucleotide step. For the conversion of the whole DNA from B- to A-form, there will be no junction free energy cost. To estimate $\Delta \Delta G_a$ for all ten different dinucleotide steps, we performed four sets of free energy calculations for homo dinucleotide steps AA, TT, GG, and CC for DNA sequences $d(\text{CGCGX}\underline{\text{XXXX}}\text{CGCG})_2$ (X= A/ C/ G/ T). The pair of base pairs converted to A-form using Z'_p is shown with underline. The free energy cost here can be decomposed as,

$$\Delta G_{sim}(XX) = \Delta G_j(XX) + \Delta G_a(XX) + \Delta G_j(XX). \quad \text{Eq. 4.2}$$

Subsequently, we converted two adjacent dinucleotide steps (i.e., a trinucleotide step), underlined in the sequence $d(\text{CGCGX}\underline{\text{XXXX}}\text{CGCG})_2$, to A-form using average $Z'_p(\overline{Z'_p})$ coordinate. The free energy decomposition here can be written as,

$$\Delta G_{sim}(XXX) = \Delta G_j(XX) + 2 \Delta G_a(XX) + \Delta G_j(XX) \quad \text{Eq. 4.3}$$

Subtracting Eq. 4.2 from Eq. 4.3, we obtain the absolute free energy for homo dinucleotide steps AA, TT, GG, and CC ($\Delta G_a(XX)$).

$$\Delta G_a(XX) = \Delta G_{sim}(XXX) - \Delta G_{sim}(XX). \quad \text{Eq. 4.4}$$

Now, we can get the junction free energy cost from Eq. 4.2 and Eq. 4.3 as,

$$\Delta G_j(XX) = \Delta G_{sim}(XX) - \frac{1}{2} \Delta G_{sim}(XXX). \quad \text{Eq. 4.5}$$

Similar simulations are performed along Z'_p for the rest of the heterodimers to get the absolute free energy. Since we know junction free energies $\Delta G_j(XX)$ and $\Delta G_j(YY)$ for XX and YY steps separately, we can easily calculate $\Delta G_a(XY)$ using Eq. 4.6 as,

$$\Delta G_a(XY) = \Delta G_{sim}(XY) - \Delta G_j(XX) - \Delta G_j(YY) \quad \text{Eq. 4.6}$$

Note that; there are multiple ways one can construct free energy equations along with different set of simulations to obtain absolute free energy values for the conversion of B- to A-form. However, the scheme presented here minimizes the number of free energy simulations. The simulation details are presented in the next section.

4.2.3 Simulation Details.

All initial configurations are modeled as the canonical B-DNA using NAB module of AMBERTOOLS.[18] For simulations. The Amber99 force-field[19] with parmbsc0[20] correction was used which is capable of reliable conformational space of B-DNA as well as to depict fast conformational transitions such as A-DNA $\rightarrow$ B-DNA transition in water at nanosecond timescale. The reaction coordinate Z'_p [9], and $\overline{Z'_p}$ were incorporated by us in PLUMED[21] and it was patched with GROMACS[22] before implementation. The DNA was solvated in TIP3P water model. Simulations were carried out at physiological ion concentration. Umbrella Sampling simulations were performed to obtain PMF along the reaction coordinate. PMFs were constructed using weighed histogram analysis method (WHAM) [23] using the program by Allen Grossfield[24].

4.3 Results and Discussion

4.3.1 Free energy profiles along Z'_p and $\overline{Z'_p}$.

Figure 4.4 shows the free energy profiles for ten dinucleotide steps against Z'_p . It also shows the free energy profiles for four homo trinucleotide steps against $\overline{Z'_p}$. As expected, we found a stable state for all the nucleotides steps in the B-form region (lower value of Z'_p). Since there was no metastable state found for the A-form, we tried to estimate the minimum free energy cost for A-form conversion by ascertaining the average Z_p parameter as discussed previously.

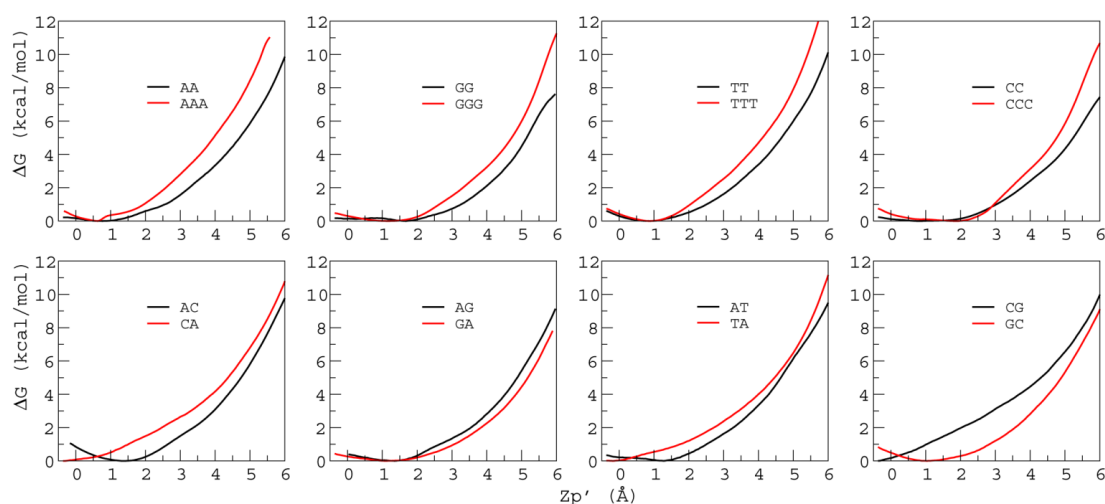


Figure 4.4. Free energy along Z'_p (black line) and $\overline{Z'_p}$ (red line) for dinucleotide and trinucleotide steps, respectively.

Thus, $Z'_{p,cut}$ obtained for each free energy profile was used to calculate the minimum free energy required for B→A conversion. Table 4.2 lists the free energy cost for B→A transition for homo dinucleotide steps and trinucleotide steps. The values were decided based on the free energy at $Z'_{p,cut}$, the values of Z'_p at which average Z_p is greater than 1.5 Å. $Z'_{p,cut}$ is also listed in Table 4.2. Note that, AA and TT have comparable total free (ΔG_{sim}) cost, so do the GG and CC steps. Even the homo trinucleotide steps have similar free energy for complimentary base triplets (AAA and TTT, GGG and CCC). This indicates that AA/TT and GG/CC are really unique dinucleotide steps in terms of total free energy of B→A conversion.

Table 4.2. Free energy values (kcal/mol) along Z'_p for homonucleotides steps AA, GG, TT, and CC. Errors are indicated in parentheses.

Sequences	$Z'_{p,cut}$ (Å)	PMF (error) (kcal/mol)
Simulations along Z'_p	$\Delta G_{sim}(XX)$	
CGCG <u>AAAA</u> CGCG	4.80	5.20 (0.79)
CGCG <u>GGGG</u> CGCG	3.80	1.77 (0.64)
CGCG <u>TTTT</u> CGCG	5.00	5.80 (0.72)
CGCG <u>CCCC</u> CGCG	3.80	2.06 (0.55)
Simulations along $\overline{Z'_p}$	$\Delta G_{sim}(XX)$	
CGCG <u>AAAA</u> CGCG	5.00	8.57 (0.39)
CGCG <u>GGGG</u> CGCG	3.80	2.84 (0.65)
CGCG <u>TTTT</u> CGCG	4.80	7.10 (1.09)
CGCG <u>CCCC</u> CGCG	3.80	2.71 (0.60)

4.3.2. Absolute and junction energies of homo-nucleotide steps:

We further decomposed the total free energy (ΔG_{sim}) cost for homo dinucleotide steps into junction (ΔG_j) and absolute free energy (ΔG_a) following Eq. 4.4 and 4.5 and shown in Table 4.3.

Table 4.3. Absolute and Junction values (kcal/mol) for homo-nucleotide steps AA, GG, TT, and CC.

Dinucleotide step	$\Delta G_a(XX)$	$\Delta G_j(XX)$	$\overline{\Delta G_a(XX)}$	$\overline{\Delta G_j(XX)}$
AA	3.37	0.92	2.34	1.59
TT	1.30	2.25		
GG	1.07	0.32	0.86	0.52
CC	0.65	0.71		

This decomposition gives a different picture. Regarding the absolute free energy cost, AA is more B-philic than a TT step, and GG is more B-philic than a CC step. However, junction free energy cost is higher for TT and CC compared to AA and GG, respectively. This makes the overall free energy cost similar. It is important to understand that the junction free energy values are important in the context of local B→A transition, which happens during the interaction of DNA base pairs with proteins and other biomolecules[2,8]

During B→A transition, a DNA sequence possesses contiguous segments of both A-form and B-form. The junction free energy[8] is related to mean cooperative length (ϑ) of such A and B stretches by the following equation..

$$\Delta G_j = RT \ln \vartheta . \quad \text{Eq. 4.7}$$

Hence, substituting our values of ΔG_j for AA and GG steps in Eq. 4.7 we get the mean cooperative length of 14.4 and 2.4, respectively. Hence, the presence of A-philic steps like GG will reduce the cooperative length for nucleation and will expedite B→A transition as expected. We connote the expected experimental cooperative length between 10-20 base pairs as a combined effect of possible A-/B-DNA patches in a sequence.

Therefore, the high junction free energy cost for TT indicates that it will oppose B→A transition when it is present as a flanking base pair. GG has the lowest junction energy; it is even lower than the thermal energy. This result supports Minchenkova and co-workers' observation that GG/CC facilitates B/A transition[25]. The A-philic GG/CC step occurs abundantly ($\sim 40\%$) in A-DNA helices[2] suggesting that proteins take advantage of the low free energy cost of GG/CC step to convert base step towards A-form. The PMF values for remaining heterodinucleotide steps are listed in Table 4.4.

Table 4.4. Free energy values (kcal/mol) from simulations along Z'_p for remaining heteronucleotide steps (underlined in the sequence). Errors are indicated in parentheses.

Sequence	$Z'_{p,cut}$ (Å)	PMF($\Delta G_{sim}(XY)$) (kcal/mol)
CGCG AA <u>A</u> CC CGCG	4.40	4.02 (0.22)
CGCG AA <u>A</u> GG CGCG	4.00	2.78 (0.70)
CGCG AA <u>A</u> TT CGCG	4.80	5.47 (0.73)
CGCG G <u>G</u> AAA CGCG	4.40	2.95 (0.54)
CGCG C <u>C</u> GCG CGCG	3.80	4.10 (0.87)
CGCG G <u>G</u> CCC CGCG	3.80	2.37 (0.65)
CGCG T <u>T</u> AAA CGCG	4.40	4.77 (0.73)
CGCG C <u>C</u> AAA CGCG	4.20	4.51 (0.77)

It also shows the $Z'_{p,cut}$ values used to obtain the free energy (ΔG_{sim}). The $Z'_{p,cut}$ values for A-philic sequences GC, GG, and CC are small (~ 3.80 Å), whereas $Z'_{p,cut}$ values for B-philic AA, AT, and TT steps are high (~ 4.80 Å). It is interesting to see that A-philic steps AC and AG have slightly higher $Z'_{p,cut}$ values because of the flanking B-philic AA base steps as junctions. As mentioned before, $\Delta G_{sim}(XX)$ for AA and TT as well as for GG and CC steps are comparable even though they have different absolute and junction values. Therefore, we used average of absolute and junction values for AA/TT and GG/CC for the calculation of the absolute free energy values for rest of the hetero dinucleotide steps. Once we obtained absolute $\Delta G_a(XX)$ and junction $\Delta G_j(XX)$ energy for homonucleotide steps AA/TT and GG/CC, we calculated absolute free energy values for the remaining steps using Eq. 4.6. The absolute values $\Delta G_a(XY)$ are provided in Table 4.5.

4.3.3 Absolute free energies of hetero dinucleotide steps:

Once the junction and absolute free energy values are available for homo dinucleotide steps, we used Eq. 4.6 to calculate the absolute free energy of all the hetero-dinucleotides and listed them in Table 5. Table 5 shows that A-philic steps GG and AG have lower absolute free energy values, and B-philic steps AA, CA, AT, and CG have higher absolute free energy values. Surprisingly, absolute free energy values suggest that CG is the most B-philic step. This observation is very similar to the Tolstorukov's D-12 model[10] where CG has the highest free energy value. It is known from quantum mechanical calculations that CG has the most stabilizing stacking energy[26] and it opposes the B→A transition at the local level[9]. Experimental study of Frank-Kamenetskii[27] also suggested that B-DNA is primarily stabilized by stacking interactions and not by the hydrogen bonding between the bases. Hence, we may attribute high absolute free energy value of CG obtained here to its stacking energy as shown later.

Table 4.5. List of absolute energy values (ΔG_a) (kcal/mol) for all 10 possible dinucleotide steps

Dinucleotide Steps	ΔG_a (kcal/mol)
AA/TT	2.34
GG/CC	0.86
AC/GT	1.91
CA/TG	2.40
AT/AT	2.29
TA/TA	1.59
AG/CT	0.67
GA/TC	0.84
CG/CG	3.06
GC/GC	1.33

The absolute energy values ($\Delta G_a(XY)$) for A-philic AC step is high and for B-philic GA step is low, whereas Tolstorukov's results correctly predicted conformational preference for AC and GA steps. To understand this, we analyze the trinucleotide patch in which these dinucleotide steps are embedded. GA step is a part of 5'-GGA-3' trinucleotide step and AC step is a part of 5'-AAC-3' trinucleotide step. Heys and co-workers[28] observed that AAC and GGA containing DNA sequences were crystallized as B-DNA and A-DNA respectively. Low $\Delta G_{sim}(GA)$ and $\Delta G_a(GA)$ values can also be attributed to less thermodynamic stability of GGA triplet as observed earlier[29]. The intermediate $\Delta G_a(AC)$ may arise here due to the 5' flanking AA step. This clearly indicates that conformational preference of a dinucleotide step is significantly affected by flanking base pairs.

4.3.4 B-philicity of AA/TT step.

Here we discuss the rationale behind resistance of AA/TT step to adopt A-form. The strong B-philicity in AA/TT arises from the steric hindrance caused by the interaction of hydrophobic cluster of C2' methylene of sugar ring, the exocyclic methyl group of thymine, and C6 ring atom of thymine (see Figure 4.5(a) and (b))[10]. In B-form, exocyclic methyl groups of thymine are 4.76 Å away, and this distance reduces to ~3.8 Å in A-form.

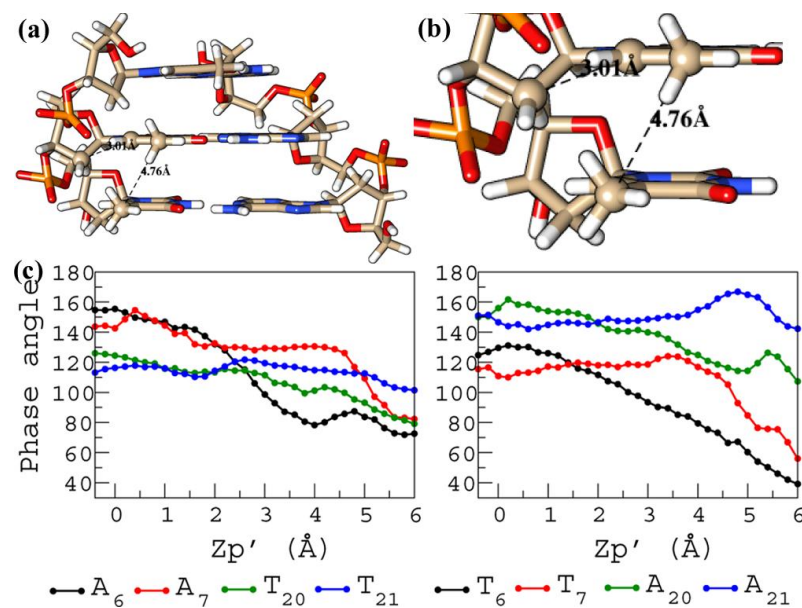


Figure 4.5. (a) The ball and stick representation of AAA.TTT trinucleotide step in the canonical A-form to indicate possible clashes between different groups. (b) Close-up of same steric clash (c) Variation of Sugar pucker along Zp' for A_6A_7 and T_6T_7 steps. Sense strand bases Base 6 (black line), base 7 (red line) and antisense strand bases base 20 (green line), base 21 (blue line) are shown in the figure.

Given that the van der Waals radius of carbon is 1.7 Å, these groups cause a significant steric hindrance in A-form. We intended to examine the effect of such steric clashes on B→A transition when TT step is a part of the antisense strand (AA/TT) and a part of the sense strand (TT/AA). The plot (See Fig. 4.5(c)) of phase angle values for $A_6T_7/T_{20}T_{21}$ and $T_6T_7/A_{20}A_{21}$ suggests all pucker values are going close to C3'-endo value in case of AA step, whereas in the case of TT only sense strand pucker values only achieve C3'-endo values. Schneider *et al.*[30] have studied conformational preferences for dinucleotide steps from crystal structures of A-DNA and B-DNA. The study categorized all 16 dinucleotide steps based on δ angle values. It suggests that TT step has an equal population in A/B and B/A conformation, whereas AA step has pronounced population in B/A conformation, indicating the different behavior of AA and TT step in crystal structures. Hence, it suggests that our reaction coordinate and method are capable of capturing such heterogeneity in unique dinucleotide AA/TT step, and it is not variation induced by the method.

4.3.5 The role of solvation energy and stacking interactions in B-A transition.

We have calculated solvation energy and stacking interactions along the reaction coordinate during B→A transition for the base steps of interest (Fig. A-4.3 in Annexure 4). The results are summarized in Fig. 4.5, where we have plotted energy difference for B→A transition for PMF, solvation energy, and stacking energy. The results indicate that the solvation energy stabilizes A-form for most cases while stacking interactions resists the transition, consistent with our previous study[9]. In the case of CG and TT steps, the retarding effect of stacking energy is pronounced. Surprisingly, solvation energy opposes B-form to A-form transition in case of TT, GG and AT steps. In our previous study, for the GG step solvation energy was found to be stabilizing B→A transition. This discrepancy could arise from the flanking base pairs around GG, which may have a strong effect on the solvation energy.

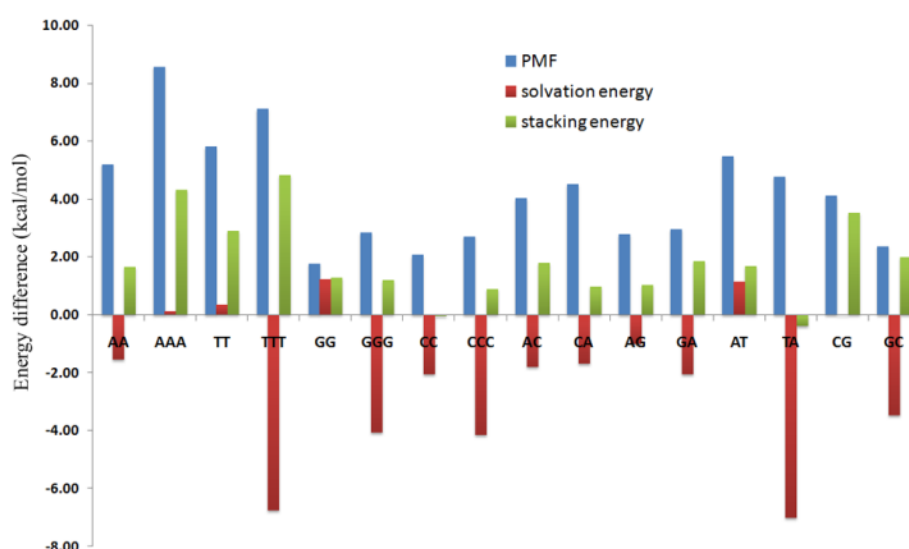


Figure 4.6. The energy difference between PMF, Base solvation energy and stacking energy.

4.4 Conclusions

In this study, we have developed a method to calculate the free energy at the junction of two different conformations of DNA -- B-form, and A-form. Please note that the junction free energy values during protein or drug induced B-A transition may differ from the values obtained in the present study being affected by interactions with this protein/drug molecule. Nevertheless, the junction free energy values obtained here are in close agreement with the experimental values, as both, the present study and previous experimental study[8] , investigated B-form/A-form junction formation

in the absence of biomolecules. Using the junction free energy values, we could extract the absolute free energy associated with the change of B- to A-form at a dinucleotide level. These absolute free energies, ΔG_a , devoid of any junction influence, are termed as absolute free energy. The ΔG_a values indicate AG, GG, TA, and GC as A-philic steps and CA, AT, AA and CG as B-philic steps. The mean cooperative length derived from junction free energies indirectly shows the effect of A-philic DNA fragment to shorten B $\rightarrow$ A transition time. The absolute free energies obtained here may be used to predict the conformation of DNA, enabling one to understand the possible functional implications of certain segments of DNA for their propensity to adopt a particular conformation. The predictive model based on absolute energies is described in the next chapter.

References:

- (1) Jacobo-Molina, A.; Ding, J.; Nanni, R. G.; Clark, A. D.; Lu, X.; Tantillo, C.; Williams, R. L.; Kamer, G.; Ferris, A. L.; Clark, P. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 6320.
- (2) Lu, X.-J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819.
- (3) Takahara, P. M.; Frederick, C. A.; Lippard, S. J. *J. Am. Chem. Soc.* **1996**, *118*, 12309.
- (4) Flick, K. E.; Jurica, M. S.; Monnat, R. J.; Stoddard, B. L. *Nature* **1998**, *394*, 96.
- (5) Galburt, E. A.; Chevalier, B.; Tang, W. L.; Jurica, M. S.; Flick, K. E.; Monnat, R. J.; Stoddard, B. L. *Nature Struct. Biol.* **1999**, *6*, 1096.
- (6) Krylov, D. Y.; Makarov, V. L.; Ivanov, V. I. *Nucleic Acids Res.* **1990**, *18*, 759.
- (7) Becker, M. M.; Wang, Z. *J. Biol. Chem.* **1989**, *264*, 4163.
- (8) Ivanov, V. I.; Minchenkova, L. E.; Burckhardt, G.; Birch-Hirschfeld, E.; Fritzsche, H.; Zimmer, C. *Biophys. J.* **1996**, *71*, 3344.
- (9) Kulkarni, M.; Mukherjee, A. *J. Chem. Phys.* **2013**, *139*, 155102.
- (10) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, *81*, 3409.
- (11) Minyat, E. E.; Ivanov, V. I.; Kritzyn, A. M.; Minchenkova, L. E.; Schyolkina, A. K. *J. Mol. Biol.* **1979**, *128*, 397.
- (12) Minchenkova, L.; Zimmer, C. *Biopolymers* **1980**, *19*, 823.
- (13) Ivanov, V. I.; Krylov, D. Y.; Minyat, E. E. *J. Biomol. Struct. Dyn.* **1985**, *3*, 43.
- (14) Banavali, N. K.; Roux, B. *J. Am. Chem. Soc.* **2005**, *127*, 6866.
- (15) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605.
- (16) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 2179.
- (17) Beveridge, D. L.; Barreiro, G.; Suzie Byun, K.; Case, D. A.; Cheatham III, T. E.; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. *Biophys. J.* **2004**, *87*, 3799.
- (18) D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolosváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman; Version 14 ed. University of California, San Francisco, 2014.
- (19) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (20) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.
- (21) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. *Comput. Phys. Commun.* **2009**, *180*, 1961.
- (22) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (23) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 1011.
- (24) Grossfield, A., "WHAM: the weighted histogram analysis method", version 2.0.8, webpage: <http://membrane.urmc.rochester.edu/content/wham>
- (25) Minchenkova, L. E.; Schyolkina, A. K.; Chernov, B. K.; Ivanov, V. I. *J. Biomol. Struct. Dyn.* **1986**, *4*, 463.

- (26) Sponer, J.; Jurecka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. *Chemistry* **2006**, *12*, 2854.
- (27) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. *Nucleic Acids Res.* **2006**, *34*, 564.
- (28) Hays, F. A.; Teegarden, A.; Jones, Z. J. R.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7157.
- (29) Huertas, D.; Azorín, F. *Biochemistry* **1996**, *35*, 13125.
- (30) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. *Nucleic Acids Res.* **2008**, *36*, 3690.

Chapter 5

Prediction of A-DNA or B-DNA Conformational Preference of Sequence: Application to Crystal Structures and *Xenopus tropicalis* Genome

5.1 Introduction

The prediction of a possible DNA conformation from the mere knowledge of its sequence presents an opportunity to presume its role in certain biological processes. The biological processes such as direct and indirect readout mechanism in protein-DNA interactions exploit the conformational flexibility exhibited by DNA. The reduction in relative humidity[1] around DNA due to the presence of other solvents like ethanol or presence of protein molecules causes B-DNA to A-DNA transition[2]. The A-DNA conformation is shorter and more compact compared to B-DNA. During B→A transition, the phosphate groups protrude out, and minor groove becomes broad and shallow to make more water bridges in agreement with the theory of economy of hydration proposed by Saenger *et al.*[3]

The protein molecules such as transposase, endonuclease, and polymerase interact with B-DNA locally and convert a few dinucleotide steps to A-form in a whole DNA[2]. The indirect recognition involves A-DNA conformation, where proteins take advantage of A-philic DNA segments in the genome to commence transcription processes. The A-form also participates in the protection of bacterial cells under extreme UV exposure[4]. Recently, Whelan and co-workers have shown fully reversible B→A-DNA transition in living bacterial cells on desiccation and rehydration using FTIR spectroscopy[5]. Extremophiles like SIRV2 virus (*Sulfolobus islandicus rod-shaped virus 2*) survives at extreme temperatures of 80°C and acidity of pH 3 by adopting complete DNA in A-form, and aids protein to encapsidate DNA[6]. Thus, it has become clear that A-DNA is merely not a non-biological

conformation of DNA but rather presents a mechanism to protect DNA under harsh conditions. Thus, it necessitates the understanding of sequence specificity of B-form $\rightarrow$ A-form transition and *a priori* detection of the A-philic segment in the genome to unveil the possible hotspots of certain biological processes in specific genes of organisms. We have developed a method to realize a priori prediction of conformational preference of a DNA sequence towards A-form or B-form.

The polymorphic nature of DNA makes the DNA conformation's prediction a challenging task. The local or partial B-form to A-form transition of a small segment of DNA sequence always possesses the penalty of B-form/A-form junction formation on both 5' and 3' ends of newly formed A-form DNA segment in a whole sequence. Considering this fact, we previously[7] performed rigorous umbrella sampling simulations to calculate this junction free energy values and characteristic local B-form to A-form free energy values for all 10 unique dinucleotide steps. The free energy values obtained therein are termed as "absolute free energy" values (ΔG_a) as they are devoid of any effects from flanking base pairs. A method is developed in this chapter to predict A-DNA or B-DNA conformational preference of a DNA sequence based on these absolute free energy values.

Earlier attempts to predict DNA conformation from its sequence include Basham's A-DNA Propensity Energy (APE) triplet code[8] and Tolsturokov's work[9]. Basham and co-workers derived A-DNA propensity energy (APE)[8] based on the solvation free energy to determine DNA structural preferences. However, APEs are unavailable for certain trinucleotide steps, thereby making this method inapplicable in some cases. In a different approach, Tolsturokov and co-workers[9] derived free energy model for all 10 unique dinucleotide steps (D-12 model) and also 32 unique trinucleotide steps (T-32 model) from experimental data of midpoints in B $\rightarrow$ A-DNA transition studied earlier by others[10,11]. The T-32 model was found more accurate than the D-12 model due to an intrinsic consideration of stereochemical effects present along the transition. However, the absence of the TAA/TTA free energy values limits the application of this dataset for a DNA structure prediction.

Hence, this chapter focuses on the development of a general and more accurate method based on the knowledge of B $\rightarrow$ A transition free energy values of 10 unique dinucleotide steps to predict the conformational preference of a DNA. We show that the method based on our absolute free energy values can capture the known

“A” and “B” philicity of DNA sequences, previously crystallized as A-DNA or B-DNA structure. The better structure prediction capability encouraged us to predict A-DNA promoter sequence (APS) in *Xenopus tropicalis* genome using recently developed APTE method[12]. The present method predicted 358 potential APS, out of which four are experimentally verified to adopt A-DNA conformation under physiological conditions. In next section, we provide brief details of the free energy calculations to elaborate the significance of absolute free energy values.

5.2 Material and Methods

5.2.1. Brief details of previous free energy calculations

The free energy values of local B-form to A-form transition were calculated using umbrella sampling simulations along a novel reaction coordinate Z'_p and average Z'_p ($\overline{Z'_p}$)[7] for 10 unique dinucleotide steps and certain trinucleotide steps embedded in the 13-mer DNA sequence. These sequences in general can be presented as d(CGCGXXYYCGCG)₂, where X/Y can be either A, T, C, or G. The presence of CG sequences on both termini reduces the possibility of base pairs fraying at the ends[13]. The free energy obtained here for the dinucleotide step XY (underlined in 13-mer sequence) from simulation can be decomposed as,

$$\Delta G_{sim}(XY) = \Delta G_j(XX) + \Delta G_a(XY) + \Delta G_j(YY). \quad \text{Eq. 5.1}$$

At this stage, we are only aware of $\Delta G_{sim}(XY)$ value. We then performed simulations on di- and tri- homonucleotide sequences d(CGCGXXXXXCGCG)₂ to find the junction and absolute free energy values for homo-dinucleotide steps. The free energy cost to convert XX step along Z'_p in sequence d(CGCGXXXXXCGCG)₂ can be decomposed as

$$\Delta G_{sim}(XX) = \Delta G_j(XX) + \Delta G_a(XX) + \Delta G_j(XX). \quad \text{Eq. 5.2}$$

Also, the free energy cost to convert XXX step in the same sequence d(CGCGXXXXXCGCG)₂ using average Z'_p ($\overline{Z'_p}$) can be decomposed as,

$$\Delta G_{sim}(XXX) = \Delta G_j(XX) + 2\Delta G_a(XX) + \Delta G_j(XX) \quad \text{Eq. 5.3}$$

Subtracting Equation 5.2 from Equation 5.3, one can obtain absolute free energy value $\Delta G_a(XX)$ and eventually junction free energy values for homo dinucleotide steps AA, TT, GG, and CC. Table 5.1 below lists these absolute and junction free energies. Using this junction free energy values ($\Delta G_j(XX)$ or $\Delta G_j(YY)$) one can calculate these absolute free energy values ($\Delta G_a(XY)$) for rest of the hetero-dinucleotide steps. These values are also listed in Table 5.1 below. Please note that, there are multiple ways one can construct free energy equations to obtain absolute free energy values for the conversion of B- to A-form. However, the scheme implemented minimizes the number of free energy simulations.

Table 5.1. List of absolute energy values (ΔG_a) for all 10 possible dinucleotide steps

Dinucleotide steps	$\Delta G_a(\text{kcal/mol})$	$\Delta G_j(\text{kcal/mol})$
AA/TT	2.34	1.59
GG/CC	0.86	0.52
AC/GT	1.91	NA*
CA/TG	2.40	NA
AT/AT	2.29	NA
TA/TA	1.59	NA
AG/CT	0.67	NA
GA/TC	0.84	NA
CG/CG	3.06	NA
GC/GC	1.33	NA

* Please note, ΔG_j values were calculated only for homonucleotide steps and not heteronucleotide steps.

5.2.2 Predicting A-DNA formation propensity

Here, we hypothesize that the absolute values obtained (Table 5.1) are additive and can be used to predict the conformation of a given DNA sequence. We used average absolute free energy per dinucleotide step (ΔG_{avFE}) to predict conformational preferences. It is calculated as,

$$\Delta G_{avFE} = \frac{1}{N} \sum_{i=1}^N \Delta G_a(XY), \quad \text{Eq. 5.4}$$

Where, N is the number of dinucleotide steps in a DNA sequence. For a DNA sequence with N basepairs, there will be $N - 1$ dinucleotide steps. Note that the free energy value obtained this way is only for the purpose of structure prediction, and may not be used to calculate the free energy for global B→A transition as it is not a simple additive function due to cooperative effects. To predict whether a given sequence will crystallize as A-DNA or B-DNA, we calculated ΔG_{avFE} for the training set to define A-DNA and B-DNA conformational boundary limits of ΔG_{avFE} values. The training set for A-DNA consists of approximately 50% octamer sequences. Octamer sequences are known to adopt A-form due to crystal packing effects exerted by one sequence on the adjacent sequence. However, to allow comparison on the same grounds with the APE triplet code, we have used the same training set.

5.2.3 Prediction of A-DNA promoter elements in *Xenopus* genome

The transcription process exploits the presence of TATA box, A-DNA in promoter region near transcription start site of the genes. Recently, Whitley and co-workers[12] applied Basham's APE triplet code to predict A-DNA promoter sequence (APS) in *Xenopus tropicalis* genome as authors found an occurrence of APS in 3.2% cases of the *Xenopus tropicalis* genes. The authors studied the role of A-DNA in the indirect readout transcription process of *Xenopus* genome to gain insight about gene regulations. The study observed that transcription factors *ilf3* and *gata2* require the presence of A-philic sequence towards 5' side of certain direct read-out motif sequences such as CCAAT, GGGCGG, AGATA, and TGATA. The Perl code (APTE) was developed by Whitley *et al.*[12] to find the potential A-form promoter sequence from *Xenopus tropicalis* JGI 4.2 genome assembly. In each gene, APTE code searches for the A-Philic sequence of length greater than 12 base pairs preceding the motif sequence or its complementary sequence up to 500 base pairs. The Circular dichroism study reliably detects A-form structure for DNA lengths above 12 base pairs. In the original APTE code, the A-philicity of the sequence was defined using Basham's A-DNA triplet code. However, due to the lack of APE values for certain trinucleotide steps certain A-DNA promoter elements remain unregistered. Hence, we used our $\Delta G_a(XY)$ values to predict A-philic DNA sequences in *Xenopus* genome. $\Delta G_a(XY)$ values have been developed for dinucleotide steps whereas APTE code used Basham's APE values for trinucleotide steps. For that, we have calculated

the ΔG_{avFE} values for all possible trinucleotide steps and used them to predict A-DNA promoter sequence. These values are provided in Table A-5.3 of Annexure 5. We have maintained the same search conditions as in Ref. [12] except we decided cut-off value of $\Delta G_{avFE} \leq 0.96$ to define A-DNA conformation rather than the choice of regular cut-off of 1.80. The cut-off of 1.80 provided more than 9000 predictions of A-DNA sequence in *Xenopus* genome. Hence, we have decided a stricter cut-off based on the average ΔG_{avFE} value calculated from known 86 A-form promoter sequences already predicted by Whitley *et al.* This cutoff provides more reliability and assurance in the prediction.

5.2.4 CD spectroscopy of potential A-DNA promoter elements

The ΔG_{avFE} method developed here has predicted 358 sequences as potential A-DNA promoter elements from *Xenopus tropicalis* genome. We have listed all these 358 sequences in the Table A-5.5 in Annexure. We have randomly chosen 4 candidates belonging to CCAAT, GGGCGG, AGATA and TGATA motif to verify the A-DNA conformation using CD spectroscopy.

Table 5.2. List of DNA sequences

Motif	Predicted A-DNA promoter sequence	ΔG_{avFE}	Predicted conformation	Observed conformation
GGGCGG	GGG GGG GGG GGG GGG CGG (Oligo-1)	0.86	A-DNA	A-DNA
AGATA	GGG GGG GGG GGG GGG TGT ATC T (Oligo-2)	0.86	A-DNA	A-DNA
CCAAT	AGA GAG AGA GGG GGG GTG ATT GG (Oligo-3)	0.79	A-DNA	A-DNA
TGATA	AGG GAG GGG AGA TAT CA (Oligo-4)	0.81	A-DNA	A-DNA

All purified DNA sequences, 5'-GGGGGGGGGGGGGGGGCGG-3'(Oligo-1, ϵ_{260} 169600 L.mole⁻¹.cm⁻¹), 5'-GGGGGGGGGGGGGGGGTGTATCT-3'(Oligo-2, ϵ_{260} 217600 L.mole⁻¹.cm⁻¹), 5'-AGAGAGAGAGGGGGGGGTGATTGG-3'(Oligo-3, ϵ_{260}

247200 L.mole⁻¹.cm⁻¹), 5'-AGGGAGGGGAGATATCA-3'(Oligo-4, ϵ_{260} 186600 L.mole⁻¹.cm⁻¹) were purchased from Integrated DNA Technology (IDT) and used without any purification. The concentrations of all DNA sequences were measured using the respective molar extinction coefficient at 260 nm. DNA was annealed for about 5 minutes by heating DNA samples at 90°C. The sample was then gradually cooled down to room temperature. All the DNA samples were prepared using 100 mM KF salt in 10 mM phosphate buffer (pH = 7.4). We are thankful to Dr. Partha Hazra and Sagar Satpathi for providing experimental support with CD spectroscopy experiments and helpful discussions.

5.3 Results and Discussion

5.3.1. Prediction of a Crystal Structures as DNA sequence.

As described earlier, we have calculated ΔG_{avFE} values for the “training” DNA sequence set used by Basham *et al.* to define the critical value of ΔG_{avFE} to distinguish A-DNA and B-DNA form. This set of sequences will be termed as the training set. Table A-5.1 in Annexure 5 provides the sequence used in the training set and the predictions for each sequence from the different methods. This allowed us to compare our ΔG_{avFE} with APE values of Basham and co-workers (see Table A-5.1 in Annexure 5 for APE values of sequences used in the training set). We observed that B-DNA sequences have an average ΔG_{avFE} of 2.08 (± 0.13) and A-DNA sequences have an average ΔG_{avFE} of 1.37 (± 0.36). Figure 5.1 shows the comparison of our calculated average absolute free energy values with Basham’s APE values for the training set.

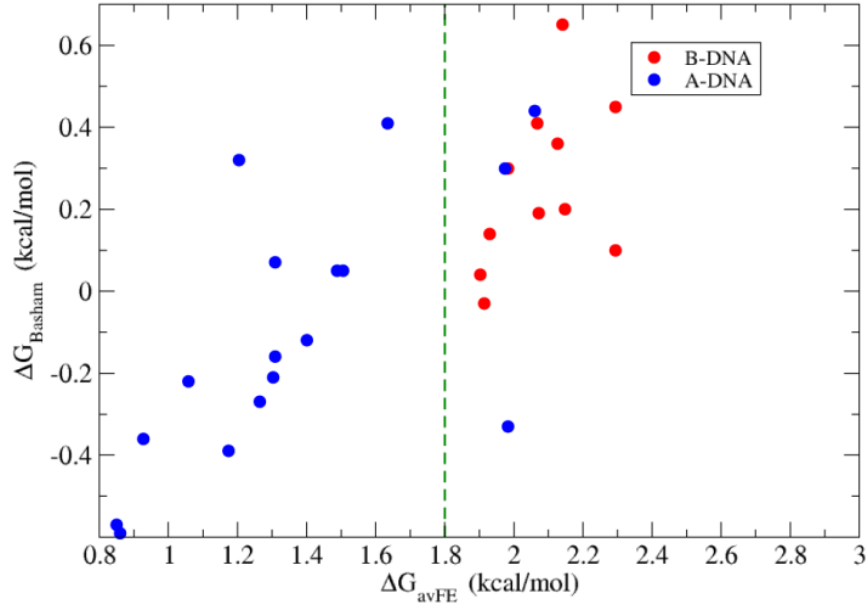


Figure 5.1. Plot of A-DNA and B-DNA crystal structures using Basham's APE values and (ΔG_{avFE}) values for the sequences in training set (Table A-5.1 in Annexure 5). The dashed line indicates the border region defined at $\Delta G_{avFE}=1.80$.

Figure 5.1 shows that ΔG_{avFE} calculated here are capable of distinguishing A-DNA and B-DNA conformations. The border between the two conformations are defined here to lie at $\Delta G_{avFE}=1.8$. We chose this point since it is almost in the midway between the lowest value of among all B-DNA structure (1.9) and highest value among all those A-DNA structures below which no B-DNA structure is present. Therefore, for a given sequence, $\Delta G_{avFE} > 1.80$ will indicate crystallization to B-form while $\Delta G_{avFE} \leq 1.80$ will indicate that the sequence may be crystallized as A-DNA. For example, for GCCGGC sequence, ΔG_{avFE} is calculated to be (From Equation 5.4) 1.05, which rightly predicts the above sequence as A-DNA[14]. To confirm the ability of our model, we tested our method on more nucleotide sequences from NDB dataset [15-17] and referred as the testing set. We have also calculated APE values for all these sequences to understand the efficiency of both methods at predicting conformational preferences. The palindromic sequences GTACGTAC, CCGTACGTACGG and GCGTACGTACGC exhibit high ΔG_{avFE} values and appear in B-DNA region. This may be due to higher proportion of AC and/or CG steps as explained later.

5.3.2. Comparison with Previous studies.

We have compared APE triplet code and our ΔG_{avFE} methods for DNA structure prediction. The training and testing set (see Table A-5.1 and A-5.2 in Annexure 5) suggests that both models have limited success to predict correct conformation for a DNA sequence. However, APE triplet code cannot be used for sequences as they are not available for some trinucleotide steps like CAG, CTG, CCT, CTT, GTT, AGC, AAC, AGG, AAG, ACA, ACT, AGT, TGT and GCT. In the testing set, there are 5 A-DNA and 14 B-DNA sequences where APE triplet code is not applicable.

Combining the training and testing sets, we have used 59 sequences (35 A-DNA and 24 B-DNA sequences), excluding sequences for which APE values are not available, to calculate the accuracy of both models. Our method correctly predicts the conformation for 49 DNA sequences (indicating the accuracy of 83.1%), and Basham's APE triplet code correctly predicted conformation for 36 DNA sequences (accuracy 61.0 %). Table A-5.2 in Annexure 5 lists the conformation predicted for all sequences using 3 methods discussed here including sequences having no APE values. Figure 5.2 shows the comparison of the efficiency of different ways in predicting the right conformation.

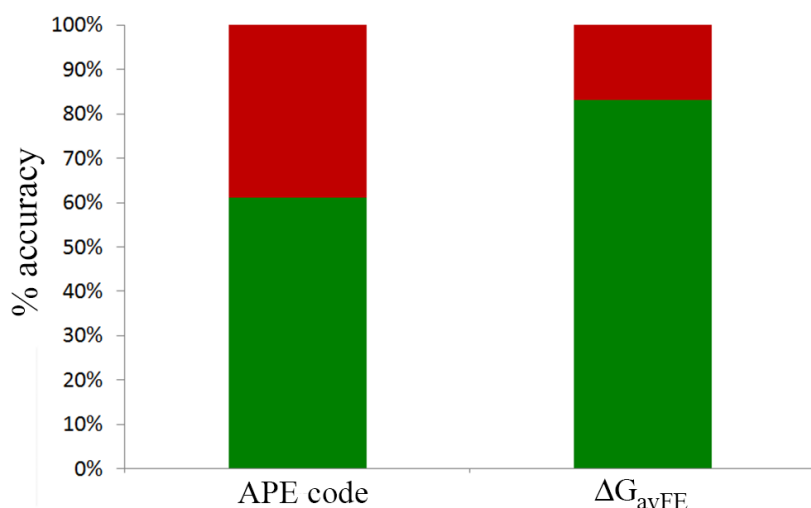


Figure 5.2. Percentage of accuracy achieved using Basham's APE method and our ΔG_{avFE} values (green colour for correct predictions and red colour for wrong predictions). Only those sequences from the combined set (training plus testing) are considered where both methods are applicable.

To understand the origin of failure of our method for certain DNA sequences, we analyzed the distribution of all 10 unique dinucleotide steps present in the wrong

predictions and presented in Fig. 5.3. The analysis based on distribution helps to understand the limitations or inapplicability of both the methods. Figure 5.3 indicates that our method might predict the wrong conformation for DNA sequences containing a higher proportion (more than 30%) of AC and/or CG dinucleotide steps. Similarly, a greater proportion of CG, GG steps (ca. 46.1%) in the case of Basham's model may result in the wrong conformational prediction.

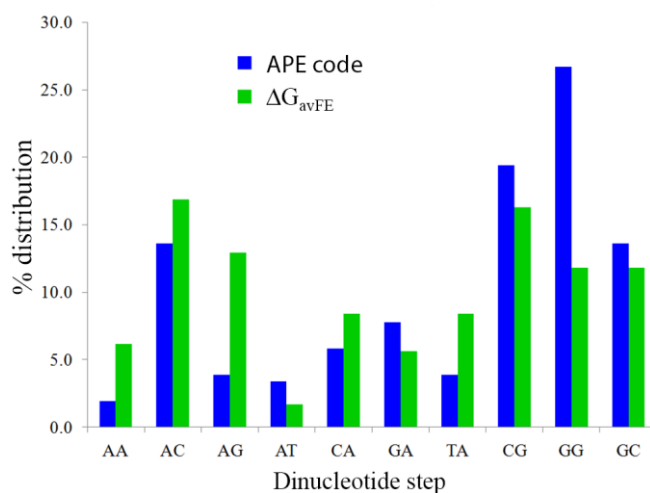


Figure 5.3. Percentage distribution of dinucleotide steps in the sequences for which wrong DNA conformation is predicted using Basham's APE code and our ΔG_{avFE} values.

5.3.3 *Xenopus* Genome analysis and observed predictions

We have predicted 358 A-DNA promoter sequences (APS) using modified APTE Perl code[12]. The only modification incorporated was our free energy values and new criteria to define A-DNA segment in *Xenopus tropicalis* genome. First, we confirmed that whether our method has also predicts 86 putative A-DNA promoter sequences predicted by Whitley *et al.*[12] or not. The analysis in Table A-5.4 in Annexure 5 shows that our method has predicted 76 APS out of 86 predicted by Ref. [12]. Some of these APS predictions belong to the gene aliases of certain genes. For example, for GGGCGG motif, Whitley *et al.* found that certain APS belong to c5orf4 gene however we have found that same APS prediction belongs to faxdc2 gene using ΔG_{avFE} method, which is a gene alias of the same c5orf4 gene. All these 76 sequences which are studied as training set have been listed in Table A-5.4 in Annexure. The average value of ΔG_{avFE} for these sequences is found to be 0.96 and hence we have chosen $\Delta G_{avFE} \leq 0.96$ as a condition to define APS. Using this condition in modified APTE code, 358 potential A-DNA promoter sequences are predicted. Out of these 358 sequences, we have confirmed A-DNA conformation of four random sequences

belonging to each motif (See Table 5.2) using CD spectroscopy, which is the mostly used and reliable spectroscopic technique to understand DNA conformation. The different structures of DNA causes distinctly different ellipticity pattern of DNA. Conventional B form of DNA exhibits a positive band around 280 nm for π - π base stacking and a negative band around 245 nm for helicity[18], while the A-DNA shows the positive band around 260 nm and a negative band around 240 nm[12,19]. Different DNAs (Oligo-1, oligo-2, oligo-3, oligo-4) corresponds to the A-form of DNA in 100 mM KF salt of 10 mM phosphate buffer (pH = 7.4) solution exhibiting similar CD spectral features like a positive peak around 260 nm and a negative band around 240 nm (Figure 4). Absence of any significant positive peak around 280 nm and negative peak around 250 nm indicates the absence of B-DNA in the CD spectrum of different DNA sequences (Figure 4). In between different DNA sequences, G rich sequences (Oligo-1, Oligo-2, and Oligo-3) found to have intense ellipticity pattern over Oligo-4, which has less number of guanine (G) bases in the sequence. This observation supports the A-DNA conformational preference of guanine rich sequence as observed earlier.

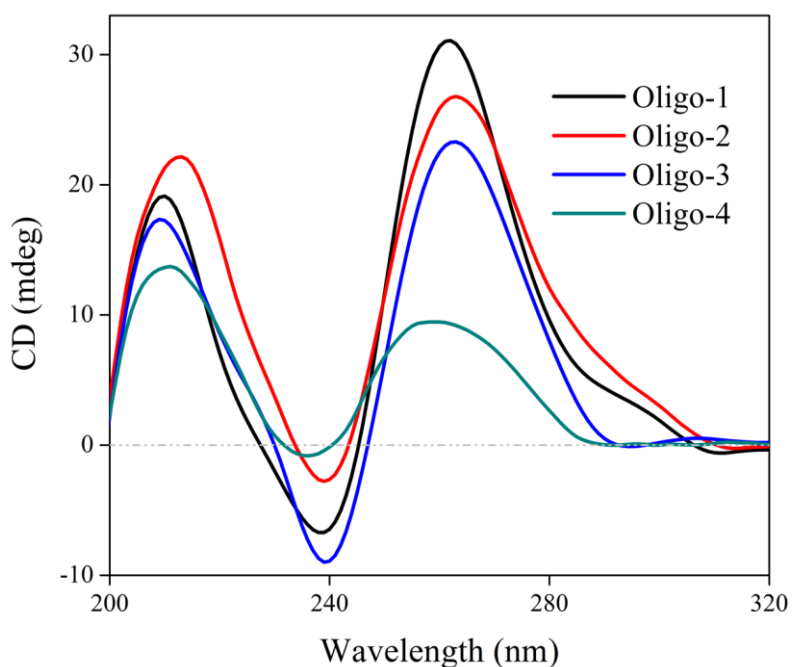


Figure 5.4. Circular dichroism spectra of different DNA sequences ($\sim 4 \mu\text{M}$ of Oligo-1, $\sim 4 \mu\text{M}$ of Oligo-2, $\sim 3.6 \mu\text{M}$ of Oligo-3, $\sim 4.5 \mu\text{M}$ of Oligo-4) in 100 mM KF salt of 10 mM phosphate buffer (pH = 7.4) solution.

The A-DNA promoter sequences predicted here were not recognized previously. However, further experimental support will be needed to ascertain the role of these predicted sequences in the transcription mechanism.

5.4. Conclusion

In Chapter 4, we had calculated the free energy values of all 10 unique dinucleotide steps, which are free from effect of flanking base steps and hence termed as “absolute” free energy (ΔG_a) of a step. The ΔG_a values indicate AG, AC, GG, TA, and GC as A-philic steps and CA, AT, AA and CG as B-philic steps. In this study, based on these ΔG_a values we have developed a simple free energy method to predict the preference of a right-handed DNA helix towards A-DNA or B-DNA conformation.

We have applied this method to 59 DNA sequences listed in NDB dataset (35 A-DNA and 24 B-DNA sequences) to predict A-DNA or B-DNA conformational preference and observed ~83% accuracy. The limitation of our method is that it may predict the wrong conformation for DNA sequences having a high percentage of AC or CG steps, the reason of which is currently under investigation. We have implemented our method to predict the A-DNA promoter sequences in *Xenopus tropicalis* genome as recently studied. We have modified the APTE Perl code accordingly to access our free energy values and ΔG_{avFE} criteria. We have successfully predicted the 76 APS out of 86 APS predicted by Ref. [12]. These 76 APS also include the five sequences that are experimentally verified by Whitley *et al* to have A-DNA conformation (See Table A-5.6 in Annexure 5 for details of these five sequences). We have applied ΔG_{avFE} method on the *Xenopus tropicalis* JGI 4.2 genome assembly and found new 358 A-DNA promoter elements belonging to different gene IDs. We believe that the proposed model based on ΔG_{avFE} value can be implemented on other genomes to find unknown A-DNA promoter elements *a priori* and further study is underway to understand its application to eukaryotic genome analysis as well as to the genome of organisms that survive under stringent conditions using A-form of DNA.

This study thus indicates that the conformational preference of a DNA lies in the fundamental free energetic driving force at a local dinucleotide level. Most of the DNA sequences used here, however, are short. Therefore, the cooperative effect may

play a role in the case of longer DNA sequences, and an effort is underway to understand this in the next chapter.

References:

- (1) Franklin, R. E.; Gosling, R. G. *Nature* **1953**, *171*, 740.
- (2) Lu, X.-J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819.
- (3) Saenger, W.; Hunter, W. N.; Kennard, O. *Nature* **1986**, *324*, 385.
- (4) Mohr, S. C.; Sokolov, N. V.; He, C. M.; Setlow, P. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 77.
- (5) Whelan, D. R.; Hiscox, T. J.; Rood, J. I.; Bambery, K. R.; McNaughton, D.; Wood, B. R. *J. R. Soc. Interface* **2014**, *11*, 20140454.
- (6) DiMaio, F.; Yu, X.; Rensen, E.; Krupovic, M.; Prangishvili, D.; Egelman, E. H. *Science* **2015**, *348*, 914.
- (7) Kulkarni, M.; Mukherjee, A. *ChemPhysChem* **2016**, *17*, 147.
- (8) Basham, B.; Schroth, G. P.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6464.
- (9) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, *81*, 3409.
- (10) Minchenkova, L. E.; Schyolkina, A. K.; Chernov, B. K.; Ivanov, V. I. *J. Biomol. Struct. Dyn.* **1986**, *4*, 463.
- (11) Ivanov, V. I.; Minchenkova, L. E.; Minyat, E. E.; Schyolkina, A. K. *Cold Spring Harb Symp Quant Biol* **1983**, *47*, 243.
- (12) Whitley, D. C.; Runfola, V.; Cary, P.; Nazlamova, L.; Guille, M.; Scarlett, G. *BMC Bioinformatics* **2014**, *15*, 288.
- (13) Beveridge, D. L.; Barreiro, G.; Suzie Byun, K.; Case, D. A.; Cheatham III, T. E.; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. *Biophys. J.* **2004**, *87*, 3799.
- (14) Mooers, B. H. M.; Schroth, G. P.; Baxter, W. W.; Ho, S. P. *J. Mol. Biol.* **1995**, *249*, 772.
- (15) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. *Biophys. J.* **1992**, *63*, 751.
- (16) A-DNA and B-DNA Dataset I, webpage: http://home.ccr.cancer.gov/lecb/tolstorukov/prot-dna/naked_DNA_structure_list.htm
- (17) A-DNA and B-DNA Dataset II, webpage: http://rutchem.rutgers.edu/~olson/ligands_AB_DNA.html
- (18) Evdokimov, Y. M.; Pyatigorskaya, T. L.; Polyvtsev, O. F.; Akimenko, N. M.; Kadykov, V. A.; Tsvankin, D. Y.; Varshavsky, Y. M. *Nucleic Acids Res.* **1976**, *3*, 2353.
- (19) Vorlícková, M.; Sedláček, P.; Kypr, J.; Sponar, J. *Nucleic Acids Res.* **1982**, *10*, 6969.

Chapter 6

B-DNA to A-DNA Transition in Aqueous Ethanol Solution: Evidence of Multistep Non-classical Nucleation Pathway

6.1 Introduction

6.1.1. Hypothesis of nucleation driven B $\rightarrow$ A transition

In previous chapters, we have seen that the conformational preferences towards A-form and B-form are preserved at local level. The global B $\rightarrow$ A transition (i.e. transition in whole DNA length) involves complex interactions between dinucleotide steps converting to A-form. The differential preferences of each dinucleotide step towards A-form give rise to the complexity in global B $\rightarrow$ A mechanism. As discussed in Chapter 1, The cooperative length for B $\rightarrow$ A transition is of the order of 10-20 base pairs according to one-dimensional Ising model[1]. The free energy barrier of B $\rightarrow$ A transition under physiological conditions is *ca.* 21.0 kcal/mol. This free energy corresponds to the conversion a cooperative length of DNA (20 base pairs) indicating $\sim$ 1.0 kcal/mol as the cost per dinucleotide step undergoing B $\rightarrow$ A form[2]. This scenario is similar to the theoretical framework of classical nucleation theory (CNT)[3-6], which states that the actual barrier of nucleation-driven process is the free energy required to convert a small part of it.

6.1.2. Classical Nucleation theory and its limitations

According to CNT, the first-order nucleation-driven phase transition involves the competition between an evolution of stable product phase and opposing surface effects. These competitive forces eventually lead to the formation of *a stable cluster* due to stochastic fluctuations in the system when intramolecular potential overcomes surface effects. The characteristics of a stable cluster are assumed similar to the

product phase. This formation of a stable cluster is an activated process, and the barrier to form this nucleus is the overall free energy barrier.

The nature has adapted nucleation mechanism as an efficient way to carry out global structural transitions in long polymers like DNA and proteins. Some illustrative examples include formation of cruciform-like DNA structure induced by negatively supercoiled DNA[7], plectoneme formation in supercoiled DNA under stress[8], toroid formation in single DNA molecules[9], nucleation of denaturation bubbles[10] in non superhelical DNA. The gas or solute condensation, melt crystallization in water, n-butanol presents typical examples of classical nucleation theory[3]. The understanding of crystallization process in the synthesis of many industrial products such as pigments, drug molecules [11] play an important role. The crystallization can be understood in terms of nucleation theory and hence its accurate description is necessary for modulation and enhanced efficiency of the crystallization process.

However, the application of CNT to real life problems like polymeric phase transition, solid-solid transitions is often limited. This limitation arises from certain assumptions of CNT[12]:

- (i) CNT assumes presence of free energy barrier for nucleation process. However, when spinodal region plays a role in the nucleation mechanism, barrier vanishes (for e.g. in case of nucleation processes involved in superheating of liquids or solids [12]).
- (ii) CNT assumes the presence of a sharp interface between the nucleus and the surrounding phase, which is always not possible and hence nucleation rate derived using CNT may deviate from the experimental values.

The deviation from CNT due to poor description of kinetics of the nucleation process leads to the different mechanisms, and are often termed as “non-classical nucleation”. The polymers like DNA and proteins tend to proceed via non-classical nucleation pathways due to restrained degrees of freedom and deviate from the typical spherical critical nucleus formation pathway as expected by the classical nucleation theory[3-6]. The examples of non-classical nucleation-driven transitions include DNA melting transitions[13], the B-DNA to Z-DNA transition in d(CG)_n sequences flanked by rigid A-tracts[14,15], protein crystallization[16] as well as protein folding[17].

6.1.3. Experimental evidence of nucleation driven B→A transition

This chapter focuses on the plausible nucleation driven process during global B→A transition. The B-DNA to A-DNA transition is considered as a first-order transition due to the coexistence of mixed B/A conformation along B-A transition pathway[18,19]. Sanyal[20] implemented Landau theory to study phase transition between A-DNA and B-DNA and speculated it as a first-order B-A transition. Leal *et al.* [21] have experimentally studied hydration of DNA-Amphiphile complex. They have discussed that A→B transition should exhibit first-order transition which is indicated by a jump in the number of water molecules at the constant water activity. But this behavior deviates due to inhomogeneities and various kinetic effects present in the system. However, in general, B→A transition is considered as a first-order transition. Many first-order phase transitions involve nucleation mechanism and hence there is a possibility of nucleation mechanism in B→A transition.

Jose and Porschke studied the B→A transition in poly(AT)[22] and natural DNA helices[23] using electric field jump experiments. The B→A transition was observed to have two time constants. The slower one (100 μ s) was speculated to be the nucleation process while the faster time constant (10 μ s) was attributed to the growth of helical regions in DNA segments. The nucleation mechanism was accounted for cooperative length of 20-30 base pairs of B-form →A-form transition. The experiments were performed on long (1600 base pairs) as well as short (70 base pairs) DNA sequences and the results were found to be independent of the chain length. These experimental findings showed the presence of multiple steps during B→A transition and hence provided the first indirect evidence of multistep nucleation in global B→A transition. Hu *et al.*[24] have shown that the intramolecular polymer crystallization is a multiple nucleation process and can be divided into primary (formation of metastable state) and secondary nucleation (chain growth). Hence, it deviates from the classical nucleation theory and represents Ostwald's step rule. Thus, B-DNA being a polymer, the same mechanism may exist in B-DNA to A-DNA transition in solution.

The linear Ising model was implemented[23] to understand thermodynamics and kinetics of the cooperative B→A transition. The nucleation parameter (σ) for the B→A transition was estimated to be 10^{-2} to 10^{-3} , hence indicating a low cooperativity for the B→A transition.

In this chapter, we have investigated (i) the possible underlying nucleation mechanism in global B-DNA to A-DNA transition, (ii) the possibility of applying CNT framework, and (iii) sequence dependence of this mechanism.

We have studied 40 base pair B-philic poly(AT) and A-philic poly(GG) DNA helices in 85% (v/v) aqueous ethanol solution using rigorous trajectory path sampling. The study comments on the cooperative length, the nature of A-DNA in solution (i.e. product state), the nucleation barrier and effect of diffusion of conformational transition via flanking base steps around A-form nucleus(rate of attachment or detachment of other base steps) at critical nucleus size.

6.2. Methodology:

6.2.1 System Preparation:

The canonical A-DNA and B-DNA geometry of 40 base pairs of poly(AT) and poly(GG) were generated using NAB[25] module of AMBERTOOLS[26]. The DNAs were placed in the center of rectangular box having dimension of approximately 8.5 nm x 8.5 nm x 17.2 nm box and aligned along the longer length of the box. The DNA was presented by CHARMM27 force field. Sprous *et al.*[27] have shown that the well equilibrated solvent structure around DNA is necessary to observe reliable A-DNA dynamics in ethanol solutions. Hence, the DNA was solvated with pre-equilibrated box of 85% ethanol solution. The final system contained approximately 6780 TIP3P water molecules and 10162 OPLS-AA ethanol molecules, constituting 85% (v/v) aq. ethanol. The system was neutralized by randomly adding 78 Na⁺ ions in the system. The final system consists of more than 114000 atoms in both cases of poly(AT) and poly(GG).

6.2.2 Choice of force field:

We have tested AMBER99 parmbsc0[28,29], recently developed parmbsc0- χ OL[30], and CHARMM27[31,32] parameters to find the reliable force field for the present study. We have observed the distribution of DNA dinucleotide step conformations of 40 base pairs poly (AT) sequence. The potential of mean force along Z_p parameter indicates that CHARMM27 force field stabilizes most of the dinucleotide steps in A-DNA conformation as expected under experimental condition. We can see that the minimum in PMF lies at $Z_p \approx 2.00$ (which indicates A-form)

while for the AMBER99 parmbsc0 and parmbsc0- χ OL, the minimum lies between $Z_p = 1.00$ and 1.20 , i.e., at an intermediate conformation between B-DNA and A-DNA.

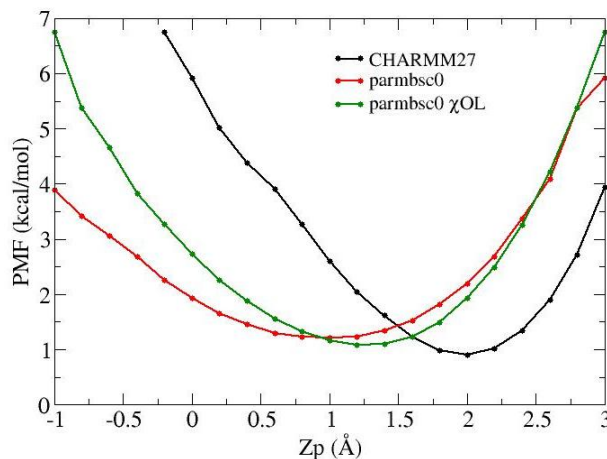


Figure 6.1. Comparison of the effect of force field on A-form stability in 85% (v/v) ethanol. PMF is plotted against Z_p parameter for all dinucleotide steps of the 40 base pair poly(AT) (excluding terminal base steps) for CHARMM27 (black), AMBER99/ parmbsc0 (red line) and parmbsc0 χ OL (green line). Figures show that PMF is most stable for higher Z_p value in CHARMM27 force field.

6.2.3 Equilibration:

The system was annealed first to assure proper mixing of water and ethanol molecules around DNA while keeping DNA heavy atoms restrained. Then a restrained equilibration run was performed for 2 ns using $5 \text{ kcal/mol/\AA}^2$ force on heavy atoms of DNA, followed by energy minimization using the steepest descent method. This procedure was followed to avoid the significant deviation from initial structure of DNA geometry from canonical A-DNA or B-DNA geometry.

6.2.4 Order Parameter to study the nucleation:

The study of nucleation process using deterministic molecular dynamics simulations is difficult due to stochastic nature of nucleation process and barriers involved to form the critical nucleus. We have implemented trajectory path sampling (TPS) method, which allows exploring reliable realistic routes of the nucleation process once converged[33].

We need a good order parameter capable of distinguishing between the starting state (state A) to the end state (state B) efficiently, to observe the nucleation barrier along pathway. Auer and Frenkel[34-36] suggested that the investigation of nucleation

mechanism should not depend on any order parameter or reaction coordinate. Although, the choice of the order parameter may result in different shapes of the nucleation free energy landscape, ideally the barrier height will remain unchanged. Also, the absolute nucleation rate should not depend on choice of an order parameter. Hence, we have implemented the usual largest nucleus size as a reaction coordinate i.e. size of largest A-form cluster (N_{Large})[37-39].

6.2.5 Definition of product state i.e. A-DNA in solution:

Unlike crystallizing conditions, it is not feasible to have all the base pairs in A-form for the DNA sequences with chain length above the cooperative length (10-20 BP) in a thermal environment. Hence, we have calculated the distribution of the percentage of A-form present in the DNA sequences from long simulation runs starting from A-DNA conformation. We have used Z_p to define A-form of DNA. As described in section 1.3 of Chapter 1, Z_p indicates the displacement of a phosphorous atoms (P) of phosphate groups of a dinucleotide step relative to virtual middle plane passing through a base step[40].

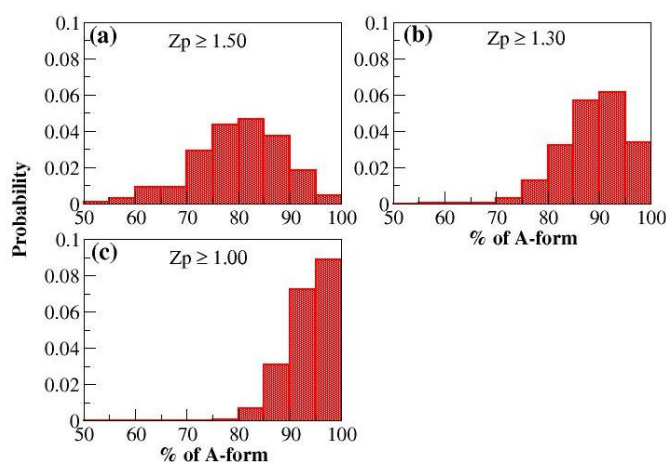


Figure 6.2. Figure showing percentage of A-form using various Z_p conditions such as (a) $Z_p \geq 1.50$, (b) $Z_p \geq 1.30$, and (c) $Z_p \geq 1.00$.

Previous studies have used $Z_p \geq 1.00$ Å [41], $Z_p \geq 1.30$ Å [42,43] or $Z_p \geq 1.50$ [40] to define A-form due to the gap in Z_p values between A- and B-forms in various crystal structures. For poly(AT), from 80 ns of A-DNA simulations, we have observed that for $Z_p \geq 1.50$ condition, 80% of steps are in A-form (See Fig. 6.2(a)), whereas for $Z_p \geq 1.30$ and $Z_p \geq 1.00$, 90% of steps are in A-form (See Fig. 6.2(b)).

Hence, we have defined a continuous stretch of dinucleotide steps in A-form as a largest A-DNA cluster (N_{Large}) if it contains 80% or 90% A-form steps for $Z_p \geq 1.50$ and for $Z_p \geq 1.30$, $Z_p \geq 1.00$ criteria, respectively. We have also verified the percentage of A-form in poly(GG) sequence (See Fig. 6.3(b)) using $Z_p \geq 1.30$ condition and found that more than 90% of steps exist in A-form. Hence, $Z_p \geq 1.30$ and 90% A-form is the condition decided for the product state for poly(AT) as well as poly(GG).

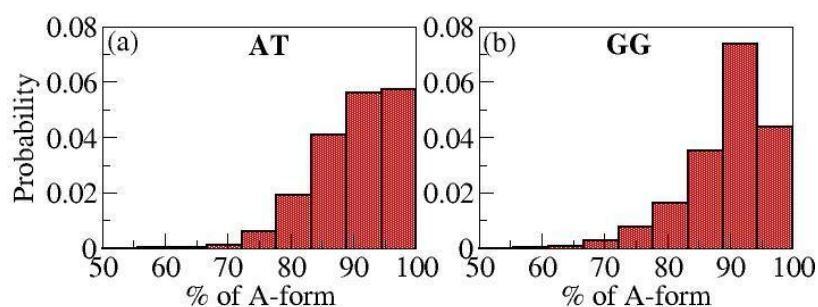


Figure 6.3. Probability of observing percentage of A-form ($Z_p \geq 1.30$) for (a) poly(AT)₄₀ (b) poly(GG)₄₀.

Hu *et al.*[24] observed that there are 20% disordered units in the intramolecular nucleation of long polymers. Hence, the absence of some dinucleotide steps in A-form may be considered as a usual observation in line with the intramolecular phase transitions of polymers. Note that, here we study an intramolecular B→A transition. In solution, at a certain instant, it is possible to observe multiple clusters of A-DNA but our order parameter is the largest cluster of A-DNA keeping in mind of our objective to look for a possible nucleation driven B→A transition[37].

6.2.6 Trajectory Path Sampling Simulations:

We have randomly chosen largest nucleus cluster sizes for trajectory path sampling simulations. The trajectory path sampling simulations refer to the unrestrained simulations starting from different points in the configuration space. We have performed 10 ns simulations starting from various N_{Large} values listed below for poly(AT) and poly(GG) DNA.

Table 6.1. The list of the largest nucleus size (N_{Large}) in poly(AT) and poly(GG) case using two different condition to define A-form.

Poly(AT)			N_{Large} (Largest nucleus size)					
$Z_p \geq 1.30 \& 100\%$	1	2	7	12	17	33	37	---
$Z_p \geq 1.30 \& 90\%$	1	2	10	14	21	37	37	---
Poly(GG)			N_{Large} (Largest nucleus size)					
$Z_p \geq 1.30 \& 100\%$	5	9	14	13	25	30	32	37
$Z_p \geq 1.30 \& 90\%$	5	10	15	20	36	37	37	37

The “largest cluster size” changes according to the definition, and hence we have provided value of N_{Large} according to two different conditions of percentage of A-form. We observe that 90% A-form with $Z_p \geq 1.30$ condition suggests that for $N_{Large} \geq 30$ base pairs shows 37 as largest nucleus size. However, 100% A-form with the $Z_p \geq 1.30$ condition give rise to different sizes for the largest nucleus. Hence, all these simulations ($Z_p \geq 1.30 \& 100\%$) were considered for the calculation of free energy along N_{Large} .

The free energy along the largest nucleus size was calculated using WHAM method[44]. For each nucleus size, three different simulations were performed having different initial velocity to assure an unbiased evolution of the A-form cluster. Also, the convergence was confirmed (data not shown)

6.2.7 Attachment/detachment rate calculations:

The nucleation free energy indicates the presence of the 4 critical nuclei sizes in case of poly(AT) and 3 critical nuclei sizes for poly(GG) sequence. We have calculated the fluctuation of these largest clusters to see the evolution of the cluster. We have also calculated the attachment/detachment rate using similar method as in Ref. [45]. The attachment rate is the frequency at which the neighboring A-form steps attach to the preformed critical A-DNA nucleus evolving to bigger nucleus and vice versa for a detachment rate. Along with fluctuation we also have calculated the diffusion of these largest clusters using mean square displacement formalism. To understand the rate of attachment or detachment of base steps to these critical nuclei, we have performed 20 simulations (for 1 ns each) starting from each critical nucleus size. For each simulation, different initial velocity was implemented. Then, the time-dependent mean square fluctuation ($\delta n^2 = (n(t) - \langle n(t) \rangle)^2$) was calculated, which

was averaged over 20 trajectories to find average mean square fluctuations $\langle(\delta n)^2\rangle$. The diffusivity (D) was calculated by a linear fit to the MSD of the cluster size evolution, averaged over 20 trajectories as discussed in Result and Discussion.

I am thankful to Dr. Suman Chakrabarty from National Chemical Laboratory, Pune for his numerous helpful discussions and guidance to understand the nucleation driven $B \rightarrow A$ transition. I am also grateful to my friend and colleague Mr. Reman Kumar Singh for his help to carry out multiple trajectory path simulations required to calculate attachment/detachment rate and diffusivity at critical nuclei.

6.3. Results and Discussion

6.3.1 Comment on cooperative length.

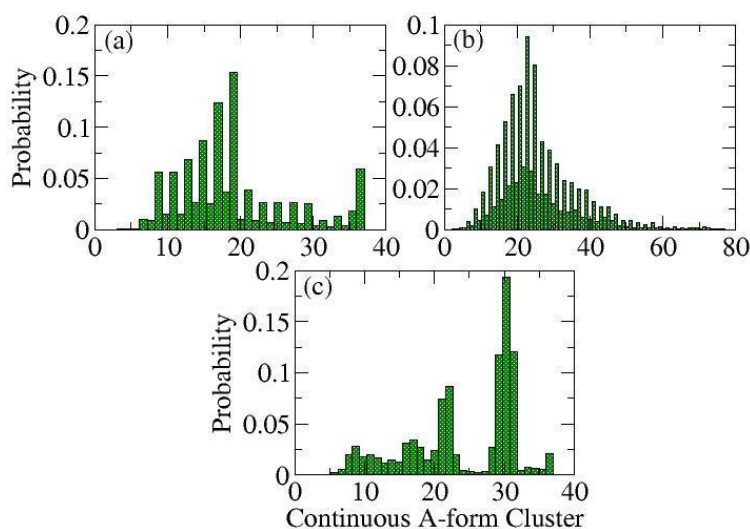


Figure 6.4. Probability of observing continuous stretch of A-form dinucleotide steps ($Z_p > 1.30$) for (a) poly(AT)₄₀ (b) poly(AT)₈₀ and (c) poly(GG)₄₀.

We have performed unrestrained 10 ns simulations with A-DNA as a starting conformation in 85% aq. ethanol solutions for 40 BP poly(AT) and poly(GG) sequences. We also studied the cooperativity with 80 BP poly(AT) DNA to find out the whether the cooperative length was dependent on the DNA length. Figure 6.4 shows the probability distribution of continuous stretch of A-form in three systems above. The cooperative length was defined as the largest continuous stretch of dinucleotide steps having the A-form. Fig 6.4 shows that the probability is maximum for ~20 bp. Fig 6.4c shows maximum at ~30 base pairs along with at 20 bp. Therefore, we realize that cooperative length to be between 20 to 30 base pairs. Previous studies have observed the cooperative length of $B \rightarrow A$ transition between 10

to 30 base pairs depending on the sequence. Hence, our result is in agreement with experiment.

6.3.2 Transition path sampling simulations.

The time evolution of N_{Large} for all the 3 runs of poly(AT) and poly(GG) is presented in block averages of 200 ps each in Fig. 6.5. In some cases, the end portion of trajectory is absent due to unacceptable Z_p values, arising due to periodic boundary conditions. In case of poly(AT), we can see that above 14 BP, the trajectory approaches global A-form and seems like one of the critical nucleus size. In case of poly(GG), simulations starting from 20 BP initially approaches a lower N_{Large} value, but eventually sample the reaction coordinate space around 20 BP again. This can be observed through the presence of plateau region around 20 BP in PMF (See Fig. 6.6).

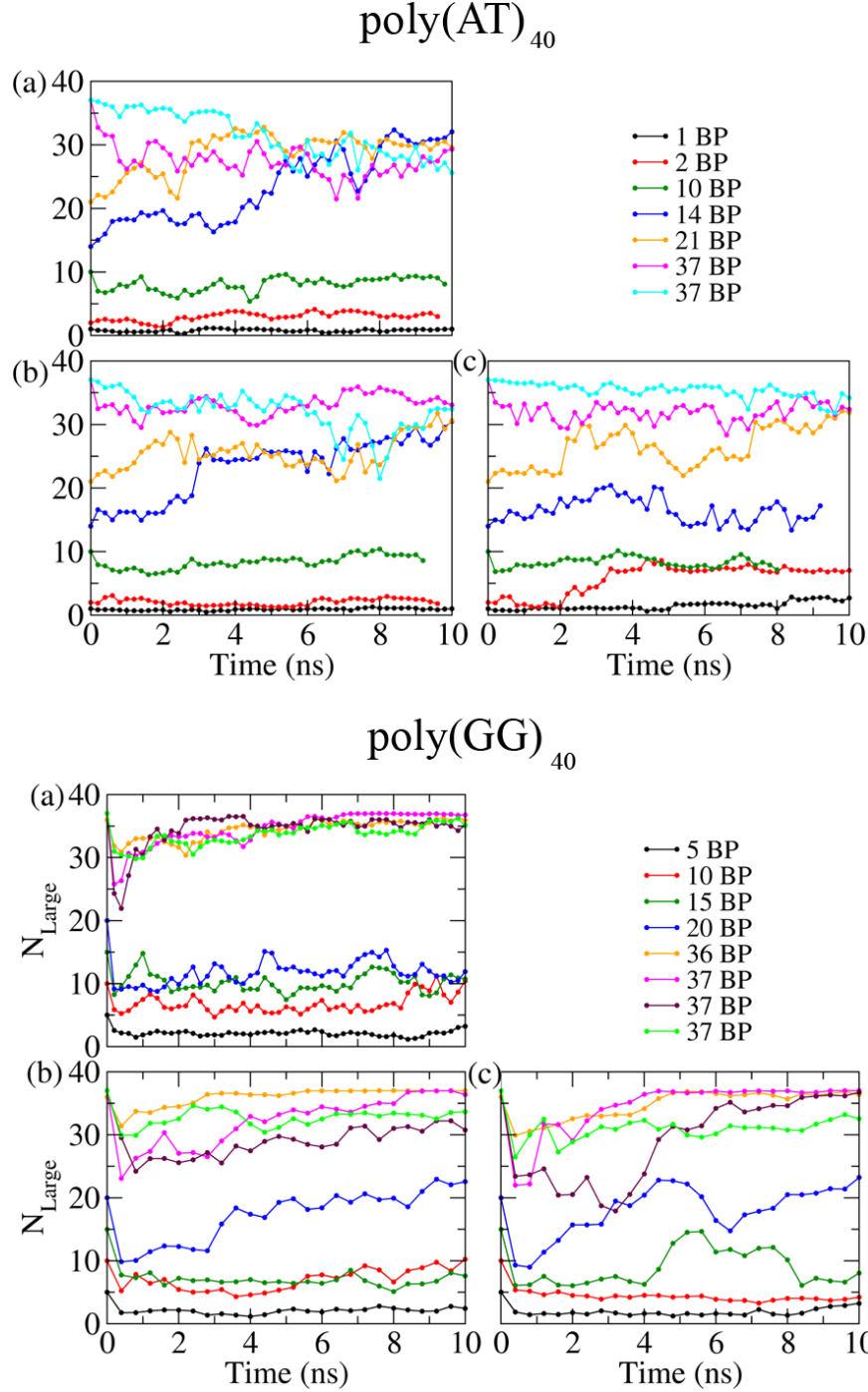


Figure 6.5. Time evolution of N_{Large} for all the 3 runs of poly(AT) and poly(GG) for (a) Run 1, (b) Run2 and (c) Run 3.

The PMF along N_{Large} was obtained using data from all 3 runs and WHAM method. The error bar is calculated using bootstrap error analysis method[46]. As described in Ref. [37], The initial dynamics of the nucleation process involves contribution of all small nuclei. The coalescence of these small nuclei leads to the bigger critical nucleus and hence their contribution is considered. And hence, please

note that, for the initial small N_{Large} cluster sizes i.e. 2 BP, 7 BP, 10 BP of poly(AT) and 5 BP, 10 BP of poly(GG), all nucleus sizes are considered while calculating PMF. Figure 6.6 shows the free energy along N_{Large} and indicates the presence of a multistep non-classical nucleation in long DNA helices poly(AT) and poly(GG). The nucleation is termed as non-classical when nucleation free energy surface is presented by either multiple-step mechanism or may be presented by different nucleation free energy landscape that what is expected from CNT[47]. A multi-step non-classical nucleation mechanism represents a special case of nucleation following Ostwald's step rule[33].

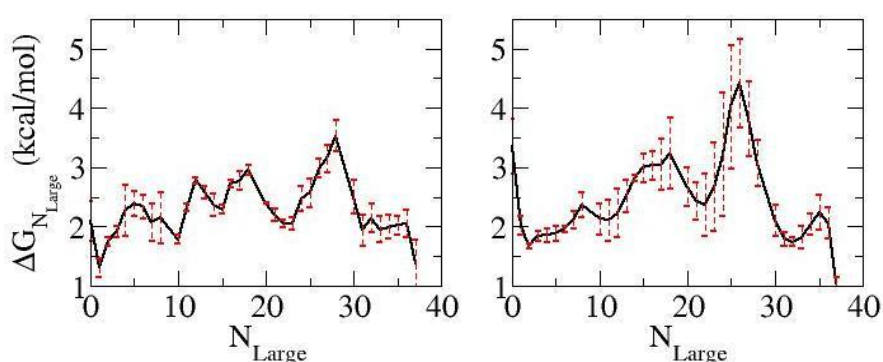


Figure 6.6. Nucleation free energy along largest cluster size for (a) poly(AT)₄₀ (b) poly(GG)₄₀.

Fig. 6.6(a) and (b) show that formation of 2 BP nucleus starting from a B-DNA conformation is a downhill process. However, there are multiple metastable states are involved in the formation of largest 39 BP cluster. This indicates the multiple step nucleation process similar to polymer nucleation multistep process.

The multi-step nucleation processes in polymers involve a primary nucleation and a secondary nucleation step. The primary nucleation step is referred to the formation of a small nucleus, which is a 2 BP nucleus in the present study for both poly(AT) and poly(GG) sequences. These primary clusters[47] are suggested to work as hotspots for further heterogeneous nucleation or secondary nucleation. The secondary nucleation refers to a growth of A-form in the DNA chain and can be considered as the limiting step for further nucleation.

The free energy minimum in each case can be observed around 10, 20, and 30 base pairs, thereby indicating that B→A transition may occur in steps of helical repeats (~10 BP is one helical repeat in DNA). However, small barriers of 1-2 kcal/mol for each nucleation step in the present case would not be able to explain the

experimental microsecond timescale observed by Jose and Porschke[23] using transition state theory. Therefore, by analogy with the kinetically controlled intramolecular nucleation present in protein folding[48], we have hypothesized here that the barrier of intramolecular nucleation driven B→A transition may be diffusion controlled.

6.3.3 Attachment/ detachment rate at critical nucleus size

The free energy barriers for multi-step nucleation are found to be low, unable to justify experimental observed microsecond timescale. Zahn[47] reviewed the classical and non-classical nucleation as well as the role of prenucleation clusters in nucleation mechanism. Author suggested that each step in multiple step nucleation free energy profile can be individually treated as classical nucleation step.

According to CNT[49], for the liquid to solid transition the nucleation rate per unit volume (I) is given by

$$I = \Gamma \exp \left[-\frac{16\pi}{3k_B T} \frac{\gamma^3}{(\rho_s |\Delta\mu|)^2} \right] \quad \text{Eq. 6.1}$$

Where, Γ is prefactor related to growth of critical nucleus, γ is the solid/liquid surface free energy, ρ_s is the number density of bulk solid, and $\Delta\mu$ is the chemical potential difference between solid and liquid. Auer and Frenkel[49] in their seminal work assumed that the barrier crossing in nucleation free energy surface is diffusive. The prefactor (Γ) depends on the rate at which nuclei grow and shrink at critical nucleus size in this diffusive region and hence it is expressed as

$$\Gamma = Z \rho_l f_{n_c}^+, \quad \text{Eq. 6.2}$$

where, Z is the Zeldovich factor which arises due to the inability of all clusters at the critical size (n_c) to proceed to the product state.

$$Z = \sqrt{|\Delta\mu| / (6 \pi k_B T n_c)}. \quad \text{Eq. 6.3}$$

$f_{n_c}^+$ is the attachment or detachment rate at the critical nucleus size and it is related to the mean square fluctuations of critical nucleus size (n_c) obtained from multiple short trajectories. $f_{n_c}^+$ is give by

$$f_{n_c}^+ = \frac{1}{2} \frac{\langle \Delta n_c^2(t) \rangle}{t} \quad \text{Eq. 6.4}$$

f_{nc}^+ is often called[49] as effective diffusion coefficient of attachment of other nuclei with critical nucleus at the top of the barrier. Thus, we have calculated the detachment or attachment rate of base steps at the critical nuclei sizes of poly(AT) and poly(GG) from the time dependent mean square fluctuations of largest nucleus size as shown in Eq. 6.4.

We have calculated attachment rate at a largest critical nucleus size from the mean square fluctuations of the nucleus size according to Ref. [45] and according to Eq. 6.5 and Eq. 6.6 provided below:

$$\langle(\delta n)^2\rangle = \frac{1}{N} \sum_{i=1}^N (n_i(t) - \langle n_i(t) \rangle)^2 \quad \text{Eq. 6.5}$$

where, N = number of short trajectories

$$f_{nc}^+ = \frac{1}{2} \frac{\langle(\delta n)^2\rangle}{t} \quad \text{Eq. 6.6}$$

6.3.4. Variance of largest cluster size near critical N_{Large} values

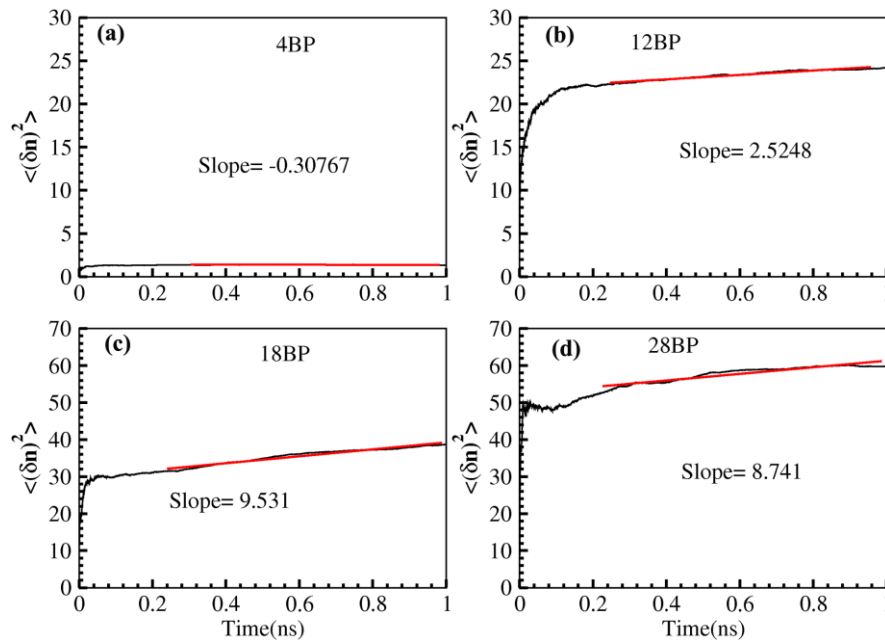


Figure 6.7. The time-dependent mean square fluctuations of starting from four critical nuclei sizes viz. 4 bp, 12 bp, 18 bp, 28 bp in case of poly(AT)₄₀ sequence. The red line indicates the linear fit to calculate the effective attachment/detachment rate of nucleus.

Fig 6.7 and Fig 6.8 shows the fluctuation of the largest cluster for poly(AT) and poly(GG) respectively. Note that, the attachment/detachment of the largest cluster obtained by a linear fit to the time dependent mean square fluctuations of nucleus size ($n(t)$), these fluctuations have been averaged over 20 independent simulations. Slope

of the linear fit represents the attachment/detachment rate to the largest cluster nucleus. Positive slope shows the growth of the cluster (attachment to the largest cluster) and negative slope represents the shrinking (detachment from the largest cluster) of the cluster. In case of poly(AT), we have analyzed the attachment/detachment at 4 critical nuclei sizes viz. 4 BP, 12 BP, 18 BP and 28 BP and for poly(GG) at 9 BP, 18 BP and 26 BP critical nucleus sizes. Fig 6.7 shows that in poly(AT) DNA, all the largest cluster have positive slope except 4 BP cluster. 4 BP poly(AT) has -0.30767 slope which shows that 4 BP cluster prefer to break whereas all other clusters (12BP, 18BP and 28BP) with positive slope, prefer to grow.

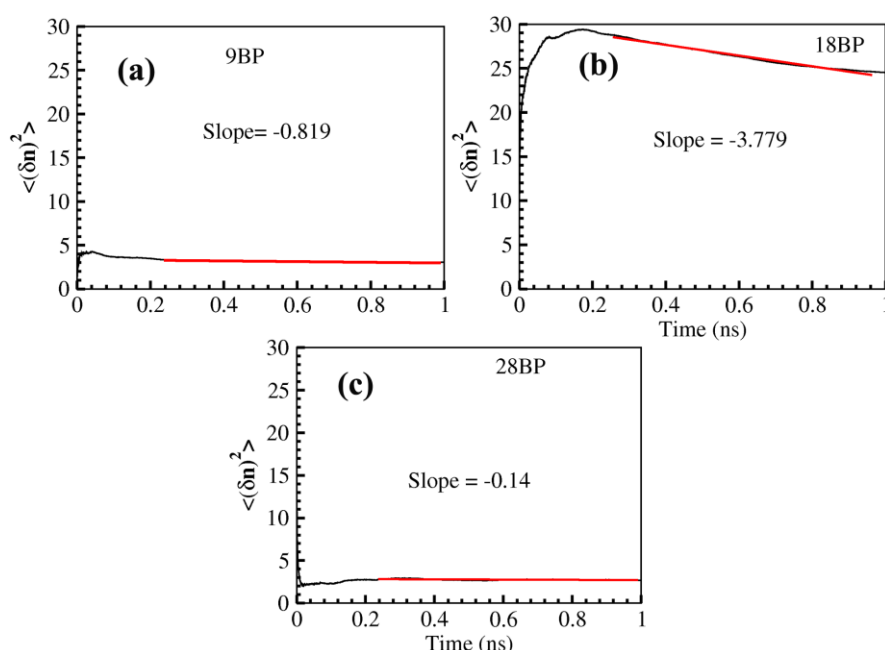


Figure 6.8. The time dependent mean square fluctuations of starting from three critical nuclei sizes viz. 9bp, 18 bp, 26 bp in case of poly(GG)₄₀ sequence. The red line indicates the linear fit to calculate effective diffusion or attachment/detachment rate of nucleus.

Contrary to poly(AT), all clusters in poly(GG) show a negative slope (Fig 6.8). Critical nucleus sizes 9 BP and 28 BP have less negative slope compare to 18 BP. So, any cluster of poly(GG) prefer to break. The rate is different because of different slopes in all clusters.

On the same line, we have calculated the diffusion of the largest cluster at critical nucleus sizes over 20 independent trajectories by linear fit to the MSD. Fig 6.9 and Fig 6.10 show the diffusion of poly(AT) and poly(GG), respectively. Slope of the linear fit represents the diffusion of the clusters. 9 BP cluster of poly(GG) has lesser diffusion (0.16 steps/ns²) compared to the other clusters. On contrary to the poly(AT),

poly(GG) clusters have very less diffusion coefficient (Fig 6.10 and table 6.2). These small diffusion coefficients suggest that even though the barrier is less, origin of the microsecond timescale in the B→A transition is probably lies in the low diffusivity. The values of diffusion coefficients are tabulated in table 6.2.

6.3.5 Diffusion at critical nucleus in poly(AT) and poly(GG)

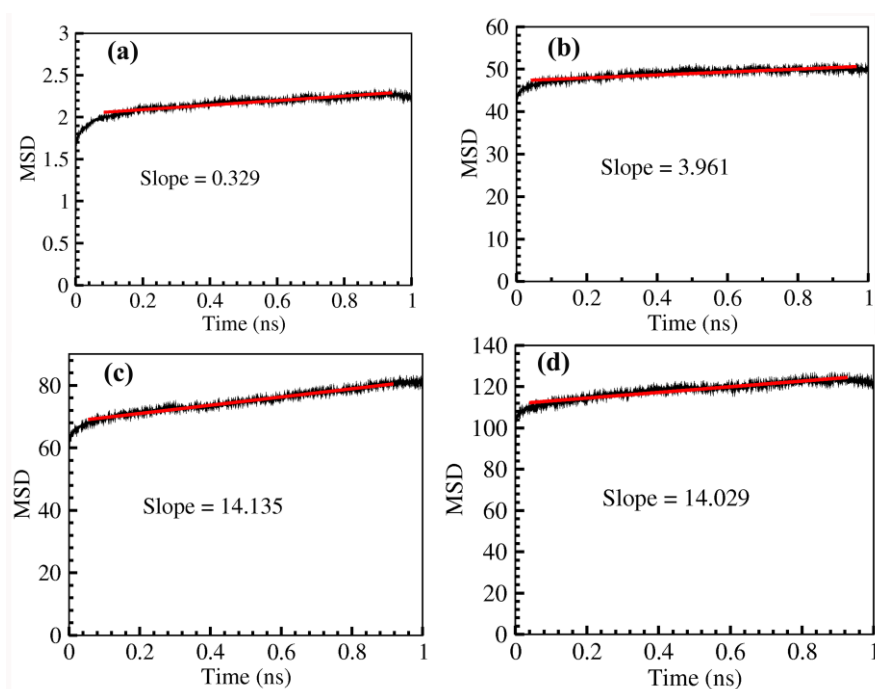


Figure 6.9. MSD of each largest nucleus of poly(AT) DNA. (a) 4 BP (b) 12 BP (c) 18 BP (d) 28 BP. Red line in each graph represents the slop of MSD. Here, reported MSDs are an average over 20 trajectories.

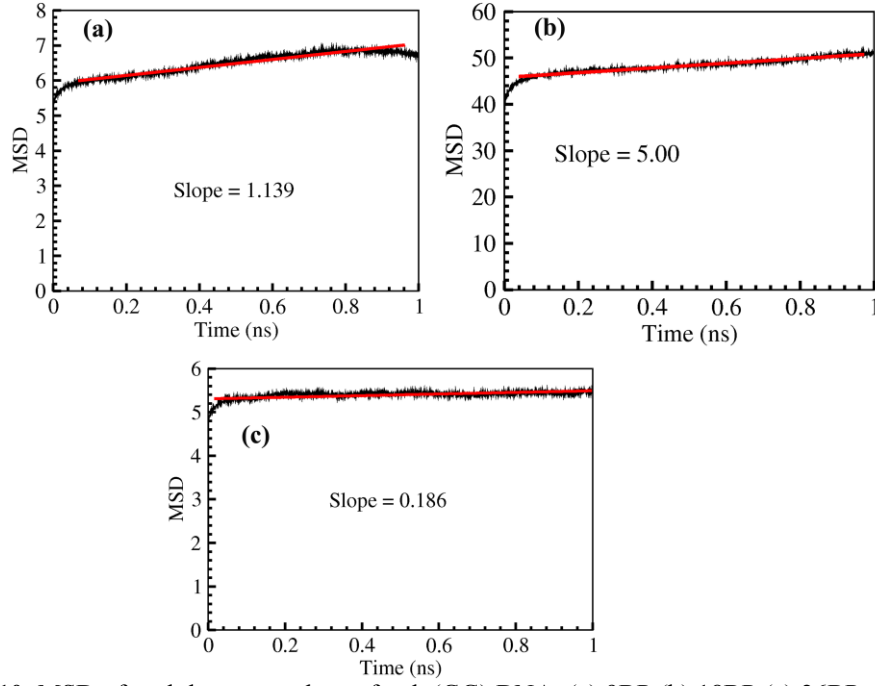


Figure 6.10. MSD of each largest nucleus of poly(GG) DNA. (a) 9BP (b) 18BP (c) 26BP. Red line in each graph represents the slop of MSD. Here, reported MSD has been averaged over 20 trajectories.

Table 6.2. List of diffusivity along critical nucleus sizes of poly(AT) and poly(GG)

poly(AT)	4 BP	12 BP	18 BP	28 BP
Diffusivity (D) (base steps) ² /ns	0.16	1.98	7.07	7.01
poly(GG)	9 BP	18 BP	26 BP	----
Diffusivity (D) (base steps) ² /ns	0.57	2.5	0.09	----

Auer and Frenkel[49] observed that the short time diffusion, studied here, is a non-Markovian process. The short time diffusion (f_{nc}^+) always differs from the diffusion calculated from long runs (D_s^L). The difference between f_{nc}^+ and D_s^L values arises due to contribution from diffusion over the barrier.

6.4. Conclusion

In this Chapter, we have explored the plausible nucleation driven $B \rightarrow A$ DNA transition with 40 BP sequences in 85% aq. ethanol solution. The length of the sequences studied is more than the cooperative length of the $B \rightarrow A$ transition. The

nucleation is studied on strongly B-philic poly(AT) sequence as strongly A-philic poly(GG) sequence. This approach allowed us to understand the sequence specific nucleation driven mechanism.

The nucleation study can be affected by the choice of force field and hence, test calculations helped us to realize that CHARMM27 force field could be the appropriate choice. To further assure force field choice, we have studied cooperative length for B $\rightarrow$ A transition in aq. ethanol and it came out to be close to experimental value of 10 to 30 BP, depending on the sequence length.

We have implemented the approach of trajectory path sampling simulations as explained in Section 6.2. The PMF along cluster size, surprisingly, indicated the presence of multistep non-classical diffusion limited nucleation in long poly(AT) and poly(GG) helices. However, the barriers failed to explain the microsecond timescale barrier involved in the global B $\rightarrow$ A transition. This made us hypothesize the presence of diffusion limited barrier for B $\rightarrow$ A transition.

Thus, we have calculated the mean square fluctuation and diffusion around each critical nuclei for both poly(AT) and poly(GG). The attachment/detachment rate indicates that for B-philic poly(AT) sequence, the detachment of other A-form steps is prominent at critical nucleus size of 4 BP and hence it retards the further growth of A-form nucleus, thus enhancing B $\rightarrow$ A transition timescale. However, in case of poly(GG), the detachment is prominent at each critical nuclei, indicating its tendency to break into shorter A-form segments. The low diffusivity values near critical nuclei for both poly(AT) and poly(GG) (See Table 6.2) indicates that the nucleation driven B $\rightarrow$ A transition (if present) is most probably limited by the diffusion and the barrier is not thermodynamic one.

Thus, the present simulation study speculates the presence of multistep non-classical diffusion limited nucleation in long poly(AT) and poly(GG) helices. The relevance of these observations will be enriched after analyzing them in case of natural DNA helices with varying GC content. However, the correlated fluctuations of different A-form domains in long DNA helices do not necessarily indicate a nucleation phenomenon. Hence, the further investigations to assure the nucleation phenomenon in DNA are in still progress.

References:

- (1) Mazur, A. K. *J. Am. Chem. Soc.* **2003**, 125, 7849.
- (2) Gekko, K.; Sakaki, T.; Shindo, H. *Polym J* **1988**, 20, 751.
- (3) Laaksonen, A.; Talanquer, V.; Oxtoby, D. W. *Annu. Rev. Phys. Chem.* **1995**, 46, 489.
- (4) Frenkel, J. *Kinetic Theory of Liquids*; Peter Smith Publisher, Incorporated, **1984**.
- (5) Debenedetti, P. G. *Metastable Liquids: Concepts and Principles*; Princeton University Press, **1996**.
- (6) Chaikin, P. M.; Lubensky, T. C. *Principles of Condensed Matter Physics*; Cambridge University Press, **2000**.
- (7) Holkers, M.; de Vries, A. A. F.; Gonçalves, M. A. F. V. *Nucleic Acids Res* **2012**, 40, 1984.
- (8) Daniels, B. C.; Sethna, J. P. *Phys. Rev. E* **2011**, 83, 041924.
- (9) Hud, N. V.; Vilfan, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, 34, 295.
- (10) Sicard, F.; Destainville, N.; Manghi, M. *J. Chem. Phys.* **2015**, 142, 034903.
- (11) Lee, E. H. *Asian Journal of Pharmaceutical Sciences* **2014**, 9, 163.
- (12) Oxtoby, D. W. *Annu Rev Mater Res.* **2002**, 32, 39.
- (13) Santos, A. T.; Klein, W. J. *Phys. A.* **2013**, 46, 415002.
- (14) Reich, Z.; Friedman, P.; Smadar, I.-Z.; Minsky, A. *J. Biol. Chem.*, 268, 8261.
- (15) Moradi, M.; Babin, V.; Roland, C.; Sagui, C. *Nucleic Acids Res* **2013**, 41, 33.
- (16) Wolde, P. R. t.; Frenkel, D. *Science* **1997**, 277, 1975.
- (17) Fersht, A. R. *Curr. Opin. Struct. Biol.* **1997**, 7, 3.
- (18) Ng, H.-L.; Kopka, M. L.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, 97, 2035.
- (19) Vargason, J. M.; Henderson, K.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, 98, 7265.
- (20) Sanyal, D. *Phys. Rev. E* **2010**, 81, 049902.
- (21) Leal, C.; Wadsö, L.; Olofsson, G.; Miguel, M.; Wennerström, H. *J. Phys. Chem. B* **2004**, 108, 3044.
- (22) Jose, D.; Porschke, D. *Nucleic Acids Res* **2004**, 32, 2251.
- (23) Jose, D.; Porschke, D. *J. Am. Chem. Soc* **2005**, 127, 16120.
- (24) Hu, W.; Frenkel, D.; Mathot, V. B. F. *Macromolecules* **2003**, 36, 8178.
- (25) Macke, T. J.; Case, D. A. *Molecular Modeling of Nucleic Acids* **1998**, 682, 379.
- (26) D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman; Version 14 ed. University of California, San Francisco, 2014.
- (27) Sprous, D.; Young, M. A.; Beveridge, D. L. *J. Phys. Chem. B* **1998**, 102, 4658.

- (28) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (29) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.
- (30) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886.
- (31) MacKerell, A. D.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105.
- (32) Foloppe, N.; MacKerell, J. A. D. *J. Comput. Chem.* **2000**, *21*, 86.
- (33) Anwar, J.; Zahn, D. *Angew. Chem. Int. Ed.* **2011**, *50*, 1996.
- (34) Auer, S.; Frenkel, D. *Nature* **2001**, *409*, 1020.
- (35) Auer, S.; Frenkel, D. *Annu. Rev. Phys. Chem.* **2004**, *55*, 333.
- (36) Auer, S.; Frenkel, D. *J. Chem. Phys.* **2004**, *120*, 3015.
- (37) Bhimalapuram, P.; Chakrabarty, S.; Bagchi, B. *Phys. Rev. Lett.* **2007**, *98*, 206104.
- (38) Pan, A. C.; Chandler, D. *J. Phys. Chem. B* **2004**, *108*, 19681.
- (39) ten Wolde, P. R.; Frenkel, D. *J. Chem. Phys.* **1998**, *109*, 9901.
- (40) Lu, X.-J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819.
- (41) Noy, A.; Perez, A.; Laughton, C. A.; Orozco, M. *Nucleic Acids Res* **2007**, *35*, 3330.
- (42) Marathe, A.; Karandur, D.; Bansal, M. *BMC Struct. Biol.* **2009**, *9*, 24.
- (43) Marathe, A.; Bansal, M. *BMC Struct. Biol.* **2011**, *11*, 1.
- (44) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275.
- (45) Knott, B. C.; Molinero, V.; Doherty, M. F.; Peters, B. *J. Am. Chem. Soc* **2012**, *134*, 19544.
- (46) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Taylor & Francis, **1994**.
- (47) Zahn, D. *Chemphyschem* **2015**, *16*, 2069.
- (48) Larini, L.; Leporini, D. *J. Chem. Phys.* **2005**, *123*, 144907.
- (49) Auer, S.; Frenkel, D. In *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*; Dr. Holm, C., Prof. Dr. Kremer, K., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2005, p 149.

Chapter 7

Summary and Future Directions

7.1. Summary

The thesis entitled “*Theoretical Study of B-DNA to A-DNA Transition: from Local Dinucleotide step to Global Helical Conversion*” discusses the thermodynamics and mechanistic details of 10 unique dinucleotide steps undergoing B $\rightarrow$ A transition in an aqueous environment. Here, we have implemented umbrella sampling simulations along the novel reaction coordinate Z'_p . The simulations are performed in aqueous solution in order to understand local B-form to A-form transition in the biological context. The simulation methodology and definition of A-form ($Z_p \geq 1.50$) allowed us to get upper bound estimate of free energy in the absence of stable A-form minima. This methodology allow us to trace the pathway of local B $\rightarrow$ A transition for all 10 unique dinucleotide steps AA, AG, AC, AT, GG, GC, GA, CA, CG and TA embedded in 12-mer DNA sequence, as described in **Chapter 3**.

The free energy values ($\Delta G(Z'_p)$) of all 10 unique dinucleotide steps correctly captured the A-philicity and B-philicity of these steps. The steps AA, AT are observed to be B-philic and AC, GG, GC are observed to be A-philic steps. Thus, it is interesting to note that conformational preferences are even maintained at smallest dinucleotide level. We have captured clusters of structures along intermediate values of Z'_p . This ensemble of structures were analyzed to understand thermodynamic and structural details along pathway.

It was observed that stacking energy opposes the transition as indicated by its positive value for all the ten dinucleotide steps. CG has the largest stacking energy penalty. The electrostatic interactions dominated the contribution to the total energy. The total potential energy is observed to stabilize the local A-form. The positive free energy and negative stabilizing total energy indicates that the localized transformations from A- to B-form in water may be driven by entropy. In conclusion, our study shows that water energetically stabilizes the A-form. Therefore, the stability of B-form in water might be originated from water entropy surrounding it. However, further detailed study is necessary to confirm this hypothesis.

We have used the normalized quantity $X_{norm}(Z'_p)$ (please see equation 2.11, Section 2.4.2, **Chapter 2**) to compare various structural parameters like Slide, Roll, Twist and torsional parameters. This unitless quantity allows us to compare all parameters simultaneously. There is no general trend for B→A transition at local level. All the steps involve Slide as a general parameter undergoing transition. However, AA, GG, AG, GA (Purine-Purine steps) involve Twist and sugar pucker along with Slide. AC and GC steps couple Slide-Twist motions to adopt A-form. In B-philic AT step, motions of sugar pucker and slide causes the local transition. The CA, CG, TA steps intrinsically possess high Roll angle and hence also involve Slide-Twist mechanism like Purine-Purine steps. Thus, B→A transition mainly involves Slide as a prominent parameter rather than Slide and Roll as observed for global transitions. The effect of local B→A transition on the flanking base steps at most propagates to two base steps in aqueous environment.

The free energy values of this study ($\Delta G(Z'_p)$) correlated well with experiment based D-10/32 free energies of Tolstorukov *et al.* [1] *albeit* with higher estimate. The free energy values are more appropriate as the local A-DNA formation cost during protein-DNA and drug-DNA interactions. The local A-DNA formation suffers the cost of B/A junction formation on both 5' and 3' side. The implementation of astute strategy regarding choice of DNA sequences in **Chapter 4** allowed us to get junction energy values for AA/TT and GG/CC steps. The junction energies for extreme B-philic AA/TT step and A-philic GG/CC steps are 1.59 kcal/mol and 0.52 kcal/mol, respectively. The mean cooperative length and junction energy are related to each other as

$$\Delta G_j = RT \ln \vartheta \quad \text{Eq. 7.1}$$

And hence, it corroborates that low junction energy of GG/CC actually reduces mean cooperative length to 2-3 base steps, eventually leading to easy B→A transition in aqueous environment.

The removal of junction energies from $\Delta G(Z'_p)$ values of rest of the dinucleotide steps provides absolute energy values (ΔG_a). These values are called absolute energies as they are devoid of any influence from flanking base steps. The ΔG_a values indicate AG, GG, TA, and GC as A-philic steps and CA, AT, AA and CG as B-philic steps.

The ΔG_a values developed in **Chapter 4** are implemented to develop a simple free energy method to predict A-DNA or B-DNA preference of a DNA merely from the knowledge from its sequence. In **Chapter 5**, the novel parameter ΔG_{avFE} based on absolute free energy values is applied on A-DNA and B-DNA crystal structures. The 59 DNA sequences listed in NDB dataset (35 A-DNA and 24 B-DNA sequences) were analyzed and accuracy of ~83% is observed. However, the method may fail for DNA sequences with a high percentage of AC or CG steps. We have implemented ΔG_{avFE} method to predict the A-DNA promoter sequences in *Xenopus tropicalis* genome as recently studied and predicted new potential 358 A-DNA promoter elements. The reliability of this method is further established by confirming A-DNA conformation of 4 randomly selected DNA sequences by CD spectroscopy. This present work is extended to find unknown A-philic DNA sequence around transcription sites in other genomes and further study is underway.

We have further investigated the possibility of sequence dependent nucleation driven mechanism in **Chapter 6** for 40 base paired B-philic poly(AT) and A-philic poly(GG) sequences. The *nucleation process* here refers to a partial transformation of a small segment in long DNA from B $\rightarrow$ A form, which eventually leads to B $\rightarrow$ A transition in a whole DNA sequence. The results found that global B-DNA to A-DNA transition in solution is a multistep non-classical nucleation process and the barrier is diffusion controlled.

7.2. Future Directions:

7.2.1. 2'-5' linkage in DNA, A-form and pre-biotic RNA world:

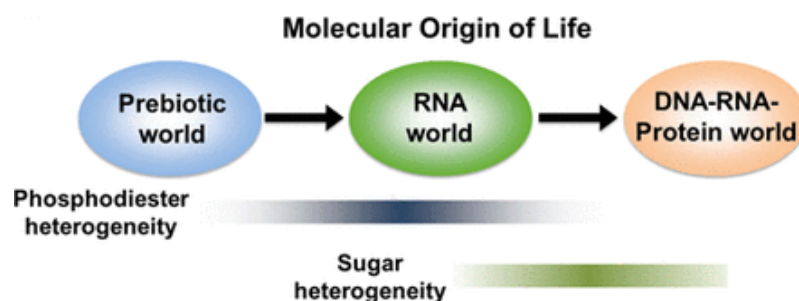


Figure 7.1. The schematic figure with proposed role of backbone and sugar modifications during evolution from prebiotic world. The figure is reproduced from reference [2] with permission.

The mechanism involved in the formation of nucleic acids and various amino acids from primordial soup of elements is still an open question. The researchers are attempting to understand the evolution of DNA and protein molecules in the quest to understand Life's origin. The seminal discussions by Woese [3], Crick [4] and Orgel [5] pointed towards the possibility of existence of an organism, where all the biological functions would involve RNA alone due to absence of protein and/or DNA molecules. This phase of the evolution is termed as RNA world. The RNA enzymes [6] or ribozymes can synthesize RNA from RNA template and catalyze the same process and hence term "RNA world" was coined by W. Gilbert [7] while speculating only RNA-based primordial world.

During the evolution, the protein and DNA molecules could have been emerged as an alternative to perform different biological functions, rather than complete RNA based system. However, the transition from RNA world to DNA-protein world and the obscurity regarding first emergence of either protein molecules or DNA molecules in this transition pose the problem similar to a "chicken and egg problem" [6].

One of the major postulates of prebiotic RNA world is the presence of 2'-5' phosphodiester linkage in the oligoribonucleotides. This rationale [6] supporting this assumption is explained below:

- (1) The linking of nucleotides needs the chemical activators such as ATP or imidazole based molecules, the existence of which is impossible in prebiotic world.
- (2) There is a possibility of an unknown molecule or metal catalyst that could have assisted linking of these nucleotides.
- (3) If there were a plausible activation of phosphate groups, then it would lead to mixture of 3 polymer products mainly having 2'-5' phosphodiester linkages, 5'-5' pyrophosphate linkage and 3'-5' phosphodiester linkages.

The presence of RNA molecules with combinations of 2'-5' and 3'-5' linkages is theorized to play an important role in RNA world as mentioned in the above postulates.

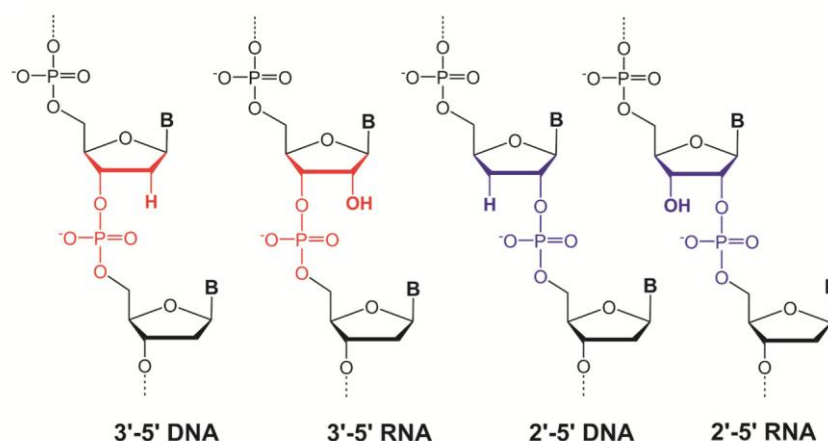


Figure 7.2. The figures showing 2'-5' and 3'-5' linkages in DNA and RNA. The figure is reproduced from reference [8] with permissions.

Robinson and coworkers [2] studied the 2'-5' connected octamer DNA helix and observed that it adopts A-DNA conformation. It was observed that the sugar pucker values of alternate bases are in C2'-endo and C3'-endo conformations. Thus, the presence of linkage at 2' position in DNA forces it to adopt A-DNA conformation and hence such 2'-5' linkage in DNA is speculated to have connection to the RNA world.

However, it is well-known that 2'-5' linkages in DNA or RNA destabilizes[9] the double helical structure. The presence of a single 2'-5' linkage reduces the thermal stability by 8.6° C in a 10-mer RNA duplex. The addition of more 2'-5' linkages in RNA duplex partially unfolds it and effectively reduces the thermal stability. However, the experimental and MD simulations study of Sheng *et al.* [9] observed that even though 2'-5' linkages reduces thermal stability, the overall structural perturbation in whole RNA sequence is very less. The neighboring base steps around 2'-5' linkage adopts according to the perturbation and screens the spread over global level.

The various crystallographic and molecular modeling studies have attempted to unveil the atomistic details for 2'-5' induced thermal instability [8,9]. However, the effect of such modifications on the structural perturbations at the local level is a missing piece of a puzzle. The 2'-5' linkages are still observed in the RNA and are of biological importance. The presence of these linkages do not disturb the RNA template dependent biological machinery and hence, presence of 2'-5' linkages are

thought to be the fossil imprints from RNA world. Recently, Xu *et al.* [8] observed that RNA polymerase II enzyme can tolerate a sugar heterogeneity such as presence of ribose sugar in genomic DNA but its transcription speed gets reduced by ~ 5400 -fold in the presence of 2'-5' phosphodiester linkage in ribose sugar.

It is well known that protein-DNA interactions involve long range correlations in structural parameters. Here we hypothesize that the 2'-5' phosphodiester linkages are unable to induce such long range correlated dynamic motions and hence are eventually omitted from transcription process during the evolution.

The study will focus on 2 major questions:

(a) The possible long range correlations of structural parameters in both RNA and DNA in the absence and presence of 2'-5' linkages and

(b) The influence of 2'-5' linkages on the flexibility of neighboring base steps in both DNA and RNA and

The study will further focus on the role of sequence specificity by implementing 2'-5' linkage using adenine, cytosine, guanine and uracil base in DNA or RNA duplex. In the present case, we can calculate the time averaged correlation coefficients and cross-correlation coefficients between various structural parameters such as Roll, Slide, and Twist etc. to realize the long range correlation, if present. Also, the time correlation functions between structural parameters at 2'-5' linkage site and rest of the base steps will help to realize the origin of long range correlation functions.

To understand the effect on the flexibility of DNA/RNA helix, we can derive the stiffness matrix of 3 translational (Slide, Rise, Shift) and 3 rotational (Roll, Twist, Tilt) structural parameters as first demonstrated by Olson *et al.* [10]. This approach is later adapted by Lankas *et al.* [11] to study the stiffness matrix of DNA basepair step deformability. This methodology is based on the assumption that the structural fluctuations in DNA or RNA are harmonic and occur within linear elasticity framework, thus obeying Hooke's law. The distribution of the structural parameters due to these fluctuations is assumed to be Gaussian.

If p_1 and p_2 are structural parameters and $\langle p_1 p_2 \rangle$ is their correlation matrix then the inverse of the stiffness matrix [12] (A^{-1}) as:

$$\langle p_1 p_2 \rangle = k_B T A^{-1} \quad \text{Eq. 7.2}$$

The diagonal elements of this stiffness matrix are force constants for structural parameters and off-diagonal elements are force constants for coupled motions between these parameters. Thus, in general, this study will help to understand the possible reasons of moving from 2'-5' linkages to 3'-5' linkages in modern nucleic acid chemistry in terms of structural flexibility.

7.2.1. DNA nanoworm based on Scrunchworm hypothesis

A) Introduction:

The concept of “DNA naoworm” is based on the combined concepts of DNA walkers and alternate B-form/A-form motions as proposed by Scrunchworm hypothesis[13]. DNA walkers are DNA based molecular nanomachines capable of “*walking*” along a predefined nucleic acid track. The sliding motions of DNA play an important role in biological motors such as DNA or RNA polymerase, kinesin or myosin motors and many others. These motors mainly convert chemical energy into mechanical energy to drive necessary motions. These ATP driven motors use mechanochemical forces of the order of 10^{-12} N to impel DNA motions.

Recently Harvey[13] proposed a DNA-centered Scrunchworm hypothesis, where continuous B-form $\rightarrow$ A-form $\rightarrow$ B-form transition causes worm like slithering motions. The Scrunchworm hypothesis assumes that protein dehydrates the sections of DNA (in multiples of 10 base pairs approximately), converting it from B-DNA to A-DNA and again rehydrates it to B-DNA. These repeating B-A-B motions cause the propagation of DNA in the forward directions by 8.5 Å. This hypothesis is presented schematically below:

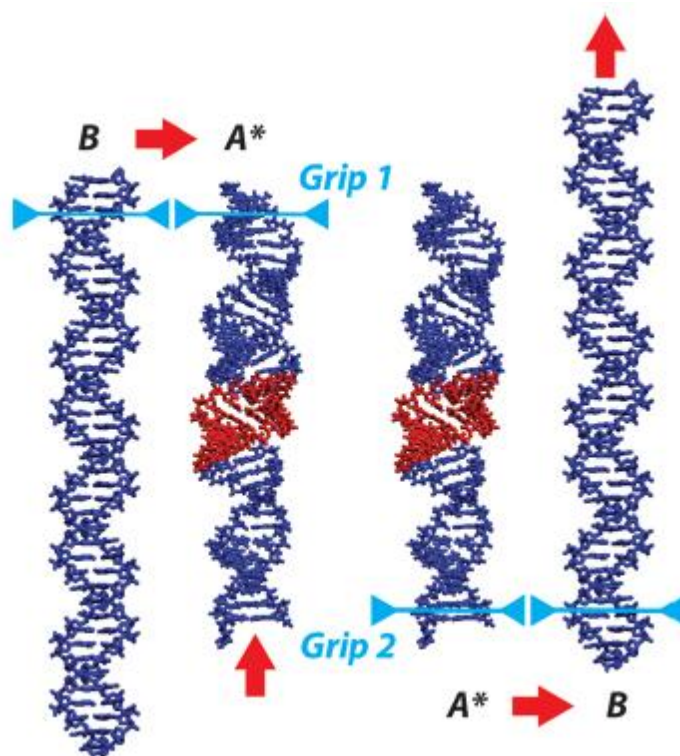


Figure 7.3. The figure showing B-form $\rightarrow$ A-form $\rightarrow$ B-form transition according to Scrunchworm hypothesis. **Step 1:** The top segment is held by grip 1 (possibly a protein moiety) and DNA undergoes B-form to A-form transition, thus pulling bottom part of DNA upwards. **Step 2:** The bottom grip 2 holds the DNA now and DNA is again hydrated thus elongating top portion. This propagates DNA in forward direction by approximately 0.85 nm. The figure is reprinted with permission from ref. [14]

B) Limitations of present synthetic DNA walkers:

The synthetic DNA walkers are powered by conformational transitions, fuel strand hybridizations, strong DNA metal interactions or photochemistry. for example, the continuous DNA walker movement was achieved by a repeated annealing/melting reaction of the walker strands by an alternating UV/vis illumination [15]. Such DNA walkers possess potential applications in cancer cell identification, therapeutics delivery, nanorobotics, DNA computing and possibly in the low-cost diagnostics of heavy metal contamination in water. However this technique is in nascent stage and limited by multiple factors such as precise control over multiple DNA walkers, small number of walking steps per fueling cycle, low fidelity and mainly the speed of the walker.

The current speeds of DNA walkers range from 0.02 to 0.10 nm/sec which are presumably slow for many applications listed above. The above speed translates to approximately 3 years to travel 1 cm distance. Recently Yehl et al. [16] developed a

rolling DNA walker to achieve better speed and achieved a speed of 1 cm per week showing significant improvement. Here, we propose a model for a *DNA nanoworm* which may cross distances faster than the present DNA walkers, albeit theoretically.

C) Proposed Project of DNA naoworm:

DNA nanoworm is the DNA walker on the grapheme like surface based on the Scrunchworm hypothesis. The schematic presentation of DNA nanoworm is presented below:

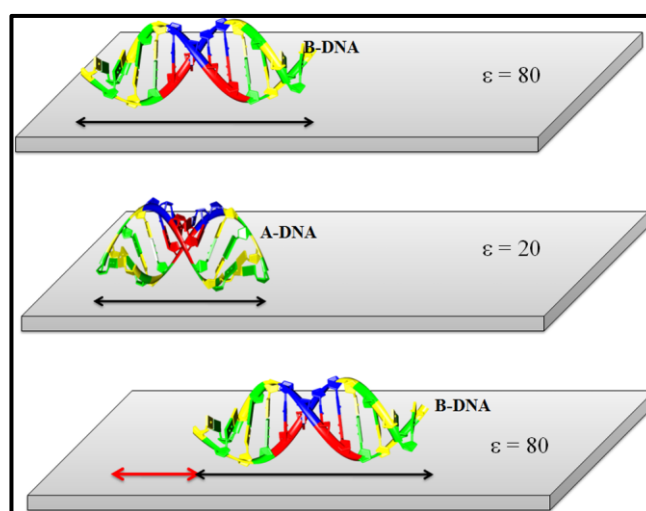


Figure 7.4. The schematic figure showing DNA nanoworm. The DNA adsorbed on the graphene like surface (indicated by grey color) can undergo continuous B-A-B transition by frequent alterations between low and high dielectric constant of the surrounding medium by controlling alternate Electric field and variation in environmental humidity.

There design of the DNA nanoworm is dependent on 2 major factors viz. (a) the B-philicity of DNA sequence and (b) the interaction between surface and DNA. transition. As recently indicated by Waters *et al.*[17], the force exerted during Scrunchworm motions is proportional to the free energy of B-A transition. Thus, more B-philic sequence will give more thrust during A to B transition allowing fast movement. Also, the hydrophobic and aromatic bases can strongly interact with the surface like graphene and hence choice of surface material will also affect the DNA nanoworm speed.

The continuous B-form $\rightarrow$ A-form $\rightarrow$ B-form can be made possible with continuous change in dielectric of the surrounding environment using alternate Electric field and variation in environmental humidity (See Fig. 7.4). Also, the

electric field of the low frequency should be implemented so as to allow the changes in polarizability of the surrounding environment.

Here, we explain the possible efficiency or speed of such nanoworm. If we assume that the DNA completes its single B-form $\rightarrow$ A-form $\rightarrow$ B-form cycle in 1 μ s to 100 μ s, thus covering 0.8 nm according to Scrunchworm hypothesis. This assumption theoretically translates that a DNA nanoworm can cross approximately 1 cm distance per 3 hours, much faster than the present fastest DNA walker suggested by Yahl *et al.* [16]. These DNA nanoworms might be a fast alternative to present models even with 5% to 10% of efficiency in real life scenario. The factors such as modulation frequency of dielectric constant of surrounding medium, choice of DNA sequence as well as choice of the surface and practical experimental setup will surely reduce the theoretically expected distance. Also, the Scrunchworm hypothesis expects the presence of protein grips which will hold DNA during the B-A-B cycle; however, in the present case the frictional forces due to surface interacting a single side of DNA would act a grip allowing propagation of DNA. The concerted motions of such multiple DNA nanoworm might be useful to perform some mechanical tasks. Thus, a DNA nanoworm can act as a potential fastest DNA walker, albeit theoretically.

References:

- (1) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, *81*, 3409.
- (2) Robinson, H.; Jung, K. E.; Switzer, C.; Wang, A. H. J. *J. Am. Chem. Soc.* **1995**, *117*, 837.
- (3) Woese, C. R. *The genetic code: the molecular basis for genetic expression*; Harper & Row, **1967**.
- (4) Crick, F. H. C. *J. Mol. Biol.* **1968**, *38*, 367.
- (5) Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381.
- (6) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99.
- (7) Gilbert, W. *Nature* **1986**, *319*, 618.
- (8) Xu, L.; Wang, W.; Zhang, L.; Chong, J.; Huang, X.; Wang, D. *Nucleic Acids Res* **2015**, *43*, 2232.
- (9) Sheng, J.; Li, L.; Engelhart, A. E.; Gan, J.; Wang, J.; Szostak, J. W. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3050.
- (10) Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163.
- (11) Lankas, F.; Sponer, J.; Langowski, J.; Cheatham, T. E., 3rd *Biophys J* **2003**, *85*, 2872.
- (12) Landau, L. D.; Lifshitz, E. M. *Statistical Physics, Part 1*; Butterworth-Heinemann, Oxford, UK: Oxford, UK, **1980**.
- (13) Harvey, S. C. *J Struct Biol* **2015**, *189*, 1.
- (14) Waters, J. T.; Kim, H. D.; Gumbart, J. C.; Lu, X. J.; Harvey, S. C. *J Phys Chem B* **2016**, *120*, 6200.
- (15) You, M.; Huang, F.; Chen, Z.; Wang, R.-W.; Tan, W. *ACS Nano* **2012**, *6*, 7935.
- (16) Yehl, K.; Mugler, A.; Vivek, S.; Liu, Y.; Zhang, Y.; Fan, M.; Weeks, E. R.; Salaita, K. *Nat Nano* **2016**, *11*, 184.
- (17) Waters, J. T.; Lu, X. J.; Galindo-Murillo, R.; Gumbart, J. C.; Kim, H. D.; Cheatham, T. E., 3rd; Harvey, S. C. *J Phys Chem B* **2016**, *120*, 8449.

Chapter 8

DNA-Ionic Liquid Interactions

Abstract:

Chapter 8 describes the studies based on DNA- Ionic liquid transitions. The chapter contains 2 addendums as described below:

Addendum I discusses the collaborative work with Dr. Partha Hazra and Sagar Satpahti. In this work, stabilization of G-quadruplex by ionic liquid molecules is explored using molecular dynamics simulation.

The CD spectroscopy and thermal melting studies by Dr. Partha Hazra and Sagar Satpahti observed that ionic liquid guanidinium tris(pentafluoroethyl)trifluorophosphate (Gua-IL) forms as well as stabilize the Human Telomeric G-quadruplex. However, the ionic liquid 1-Butylpyridinium chloride (BPyCl) was unable to stabilize G-quadruplex structure.

Thus, we have performed the molecular dynamics simulation studies to understand the interactions between G-quadruplex molecule and ionic liquid. The study observes that in presence of Gua-IL, the G-quadruplex-water interactions, G-quadruplex-Gua⁺ interactions as well as spatial distribution of Gua⁺ ions overall enhances the G-quadruplex stabilization. In general, the simulation results are observed in agreement with experimental observations.

Addendum II discusses the possible effect of ionic liquid on the DNA translocation via graphene nanopore. The practical implementation of nanopore based technology for DNA sequencing has been limited by fast translocation speed of DNA.

The previous experimental study has shown that ionic liquids strongly bind to double stranded DNA, thus increasing the translocation time of DNA. In this study, we have performed simulations to understand the plausible mechanism of ionic liquids to reduce the DNA translocation speed. The major findings of the study suggest that bulky BMIM⁺ ions accumulate near the nanopore, enhancing the effective positive charge near it. The strong electrostatic interactions between the negatively charged DNA backbone and accumulated positive BMIM⁺ ions reduce the translocation speed.

Addendum I

Ionic Liquid Induced Stabilization of G-quadruplex

AI.1. Introduction

The conformational flexibility of single stranded DNA allows it to adopt non-canonical structures such as triplex, cruciform, hairpin and G-quadruplex structures apart from canonical A-form, B-form and Z-form. The guanine rich sequences form a planar G-quartet structure where 4 guanine bases form a plane with Hoogsteen base pairs. The 2 to 3 stacked arrangements of these G-quartets constitute a G-quadruplex (GQ) structure. These structures are often stabilized by the presence of sodium or potassium ion in the core channel of the G-quadruplex. These monovalent ions interact with electronegative carbonyl oxygen atoms making an overall stable G-quadruplex structure. The flexibility of Guanine rich single stranded DNA (ssDNA) provides the broad conformational landscape for GQ structures under various conditions. Some of the important polymorphs include parallel, anti-parallel, basket type, [3+1] hybrid structures.

The biological significance exhibited by these numerous topologies of GQ structure is still unclear. The enormous efforts are involved to decode the possible role of these structures in translation, genomic instability, replication and regulation of transcription process [1]. It is observed that the location of GQ forming sequences in genome is non-random and they are conserved and present in the functional regions of genome.[1] These sequences are found abundantly in telomere regions (end regions of the chromosomes). The G-quadruplex formed by 50 to 600 nucleotides in the telomere region suppresses the telomerase activity, thus avoiding telomere extension. [2-4] The cancer cells generally exhibit a high telomerase activity and hence small binding molecules which can enhance the stability of GQ structures are considered as potential anticancer molecules[5]. The absence of methods to predict the possible 3D structure of GQ and the selective stabilization of GQ over duplex DNA pose a serious challenge of the small molecule stabilizers for GQ. The ionic liquid assisted stabilization of nucleic acids and proteins to enhance the shelf-life is recently developed as an efficient way for small volume applications. [6] The hydrated ionic

liquids such as choline di-hydrogen phosphate and deep eutectic solvents (choline chloride: urea 1:2) are known to stabilize nucleic acids. However, the underlying mechanism of such long term stabilization of biomolecules is not clear yet. Thus various experimental and theoretical studies are attempting to understand the possible interaction between ionic liquids and nucleic acids so as to modulate them.

AI.2 Simulation Methodology:

AI.2.1. Force field:

The Gua-IL induced GQ structure shows similar ellipticity pattern with that of [3+1] hybrid structure observed in case of K^+ ion, hence we have used the starting geometry of [3+1] hybrid G-quadruplex (PDB ID 2GKU). The G-quadruplex structure is presented by AMBER94 force field[7] parameters. The simulations are performed to understand stability and interactions of two ionic liquids Gua-IL and BPyCl. Also, to understand role of cations the simulations in the presence of GuaHCl are also performed.

All simulations are performed using GROMACS 4.5.3 version[8]. Force field for all IL cations and anions is adopted from previous literatures as they displayed reliable structure and dynamics of these ions. The parameters are derived from ref. [9] for Guanidinium, ref. [10] for FEP, and ref. [11] for BPy. The LJ parameters for chloride ion (Cl^-) are obtained from AMBER94[7] force field.

AI.2.2 System Preparation:

The cubic box having each side length of 5.2 nm is created and G-quadruplex molecule is kept in the center of the box. The length is chosen such as G-quadruplex molecule will get covered by 3 layers of water solvation shells and periodic boundary images would not interact with each other. First, 61 cation molecules (Gua/BPy) will be added randomly in the box and later, 61 anion molecules (FEP / Cl^-) were added randomly. Then, 3385 TIP3P water molecules were added to the box and 23 potassium ions were added randomly to neutralize the charge on DNA. This configuration is energy minimized using the steepest descent method.[12] The ratio of IL pairs to water molecules is similar to the experimental conditions.

AI.2.3 Equilibration protocol:

The system consists of G-quadruplex molecule, TIP3P water, corresponding IL pairs and K^+ ions. K^+ ions were added to neutralize the G-quadruplex and it was observed that they do not interfere in the stabilization mechanism of G-quadruplex molecule. The heavy atoms of G-quadruplex were restrained during equilibration and these restraining forces were slowly reduced along with energy minimization at each step using NPT ensemble. This helps biomolecule to properly soak into the surrounding aqueous environment. The Berendsen thermostat and barostat[13] is used to maintain temperature and pressure of the system respectively. The equilibration steps are as follows:

- i. The system was heated slowly within 100 ps to desired temperature with restraining force of $25 \text{ kcal/mol/\AA}^2$ on heavy atoms of DNA and then energy minimization was performed using steepest descent method.
- ii. Then 50 ps equilibrations were performed with restraining forces on heavy atoms of G-quadruplex DNA. The restraining force was reduced as $5 \text{ kcal/mol/\AA}^2$, $4 \text{ kcal/mol/\AA}^2$, $3 \text{ kcal/mol/\AA}^2$, $2 \text{ kcal/mol/\AA}^2$, $1 \text{ kcal/mol/\AA}^2$, $0.5 \text{ kcal/mol/\AA}^2$ with energy minimization steps in between each restrained equilibration run.
- iii. Finally, 1 ns unrestrained equilibration run with Nose-Hoover thermostat[14] and Parrinello-Rahman barostat[15] was performed. It was assured that RMSD does deviate much from the starting structure.
- iv. Every simulation was performed with periodic boundary conditions. The particle mesh Ewald[16] method was used for Columbic interactions with a $10 \text{ \AA}$ cut-off. Center of mass of the whole system was removed after every 5 ps to avoid flying ice cube effect[17].

AI.2.4 Production runs protocol

The production simulations of 300 ns were carried out in the cases where G-quadruplex forms stable structure. For rest of the simulation, the production runs were performed for $\sim 200 \text{ ns}$ or till the simulation time when quartet core of stacked guanines was disrupted.

AI.2.5 Analyses:

a) Radial distribution function (RDF):

We have calculated the RDF using `g_rdf` utility of GROMACS 4.5.3[8]. We have considered the central carbon atom (CZ) of Gua⁺ cation, positive nitrogen atom (N1) of Bpy⁺ to calculate RDF with respect to surface atoms of G-quadruplex. Similarly, oxygen atoms of water molecules (OW) were used to study RDF of water molecules around G-quadruplex.

b) Spatial distribution function (SDF):

The SDF's of cations and anions around G-quadruplex molecule in each case of GuaIL, BPyCl and GuaHCl were calculated using TRAVIS software.[18] The distribution was calculated considering CZ atom for Gua⁺ and N1 atom of Bpy⁺ molecules. The DNA molecule was aligned with respect to the starting structure before calculating SDF. The SDF is visualized using UCSF Chimera Software.[19] In case of Gua-IL, the SDF was calculated from 130 ns onwards as Gua⁺ ions get equilibrated around GQ.

AI.3 Results and discussion:

To test the above idea, we have performed MD simulations of GQ structure with Gua-IL and BPyCl keeping the ratio of the number of ionic liquid pairs and the water molecules same as used in the experimental conditions.

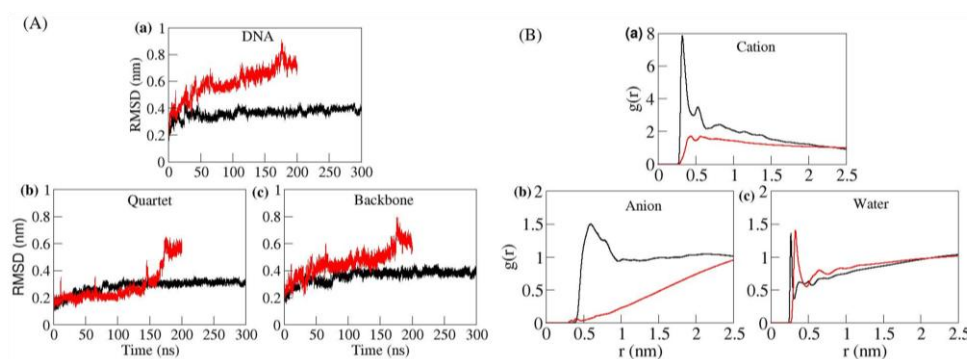


Figure AI.1. (A) Root mean square deviations of the heavy atoms of (a) G-quadruplex, (b) quartet region of G-quadruplex and (c) backbone region of G-quadruplex in presence of Gua-IL (black line) and BPyCl (Red Line). (B) Radial distribution function (RDF) of (a) cations (Gua⁺ or BPy⁺) with GQ ; (b) anions (FEP⁻ or Cl⁻) and (c) Oxygen atoms of water molecules around G-quadruplex in presence of Gua-IL (black line) and BPyCl (red line).

Figure AI.1 shows the root mean square deviations (RMSDs) of heavy atoms of GQ, core guanine quartet structure and phosphate backbone of GQ (O5'-C5'-P-O1P-O2P-O3'-C3') from the starting structure ($t = 0$ ps). All three RMSDs clearly show relatively less fluctuation for Gua-IL than that of with BPyCl. RMSD of backbone deviates continuously with time, where the quartet region suddenly distorts beyond 150 ns for BPyCl. Interestingly, we also noted that the quartet region got highly destabilized around 150 ns for BPyCl (see Fig. AI.1(A)) leading to the disintegration of GQ. This continuous fluctuation in backbone eventually leads to the disruption of the quartet region. RMSDs for Gua-IL, on the contrary, exhibit a stable structure up to 300 ns.

Along with the RMSD calculation, we have also carried out the radial distribution function (RDF) calculations of both these ILs and water molecules around surface atoms of GQ (See Fig. AI.1(B)). These results clearly indicate the presence of Gua^+ in the first solvation shell (within 0.5 nm) of GQ. In comparison with Gua^+ , the presence of BPy^+ ions around GQ is very less. Interestingly, GQ surface and its surrounding (within 0.5 nm) contains higher numbers of FEP^- ions (anion of Gua-IL) compared to chloride ions (anion of BPyCl). The RDF of water molecules shows the first solvation peak for Gua-IL at a shorter distance than BPyCl.

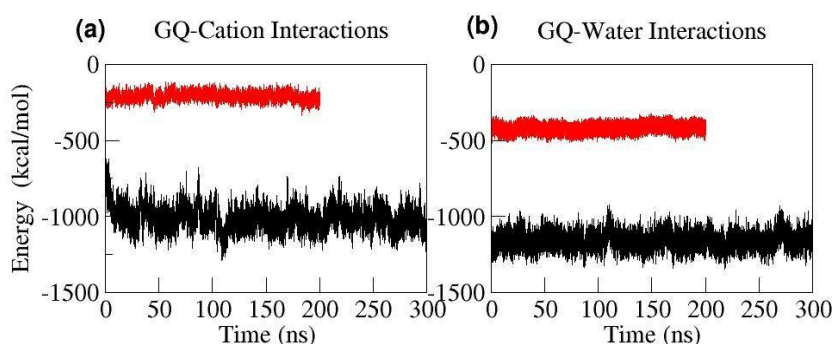


Figure AI.2. The figure showing interactions energy between (a) G-quadruplex and first solvation shell cation molecule and (b) G-quadruplex and first solvation shell water molecules in case of Gua-IL (black line) and BPyCl (red line).

The Interaction energy between GQ and cations (Gua^+ or BPy^+) in the first solvation shell indicates the strong stabilization for Gua^+ ions and convergence of energy after 50 ns (Figure AI.2(A)). Similarly, the interaction energy between GQ and water molecules within 3.1 Å distance (first solvation

shell) show approximately 3 times stronger stabilization for Gua-IL than that of for BPy-Cl (Figure AI.2(B)). This indicates that not only cation-GQ interactions but specific arrangements of water molecules are also a factor for the stabilization in case of Gua-IL. Moreover, the results imply that the GQ structure is stabilized by ion-pair rather than cation or anion alone. It is evident from the experimental and simulation results that the formation and stabilization of GQ structure by IL does not follow a general principle, rather it is a very much specific for the ion pair constituting the IL. It is interesting to inspect the individual role of cations and anions towards the formation of GQ structure.

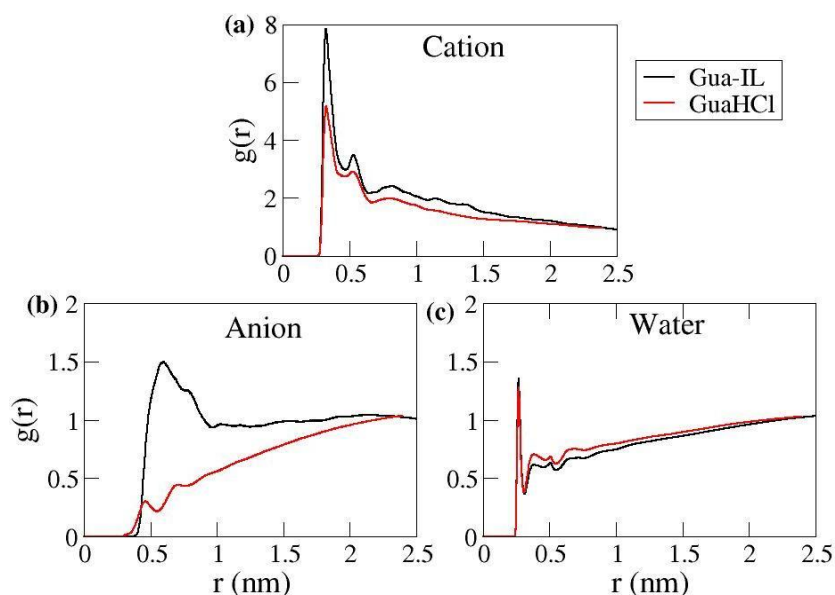


Figure AI.3. The RDF of various atoms around surface of G-quadruplex in case of Gua-IL (black line) and GuaHCl (red line) for (a) cations i.e. Gua⁺ ions ; (b) anions i.e. FEP⁻ for Gua-IL and Cl⁻ for GuaHCl and (c) water oxygen (OW) atoms.

To understand the role of ions in more detail, we have plotted RDF of ions and water molecules with the surface atoms of GQ in case of Gua-IL and GuaHCl simulations (Figure AI.3). The RDF of Gua⁺ ions indicate a lesser probability of finding cations in case of GuaHCl as well as for Cl⁻ ions in first solvation shell. This clearly indicates the lack of ion pairs for stabilizing the GQ structure, which accounts for the reduced stability of GQ in GuaHCl solutions.

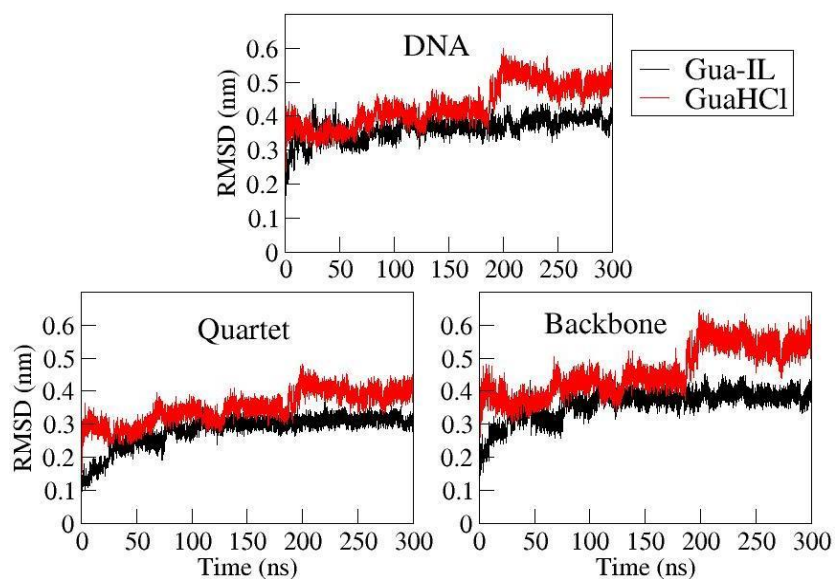


Figure AI.4. The RMSD of heavy atoms of G-quadruplex, heavy atoms of Quartet and heavy atoms of backbone in case of Gua-IL (black line) and GuaHCl (red line).

Figure AI.4 also shows that the RMSD sharply increases beyond 170 ns. Hence we assume that with further sampling one can observe the disruption of GQ. This suggests that even though GQ is not disrupted completely within 300 ns, it may eventually be destabilized.

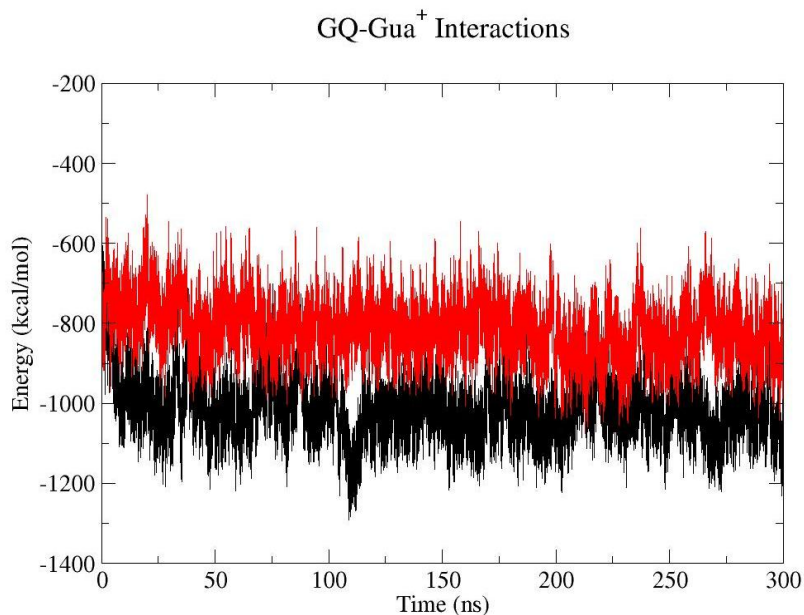


Figure AI.5. The interaction energy of Gua^+ ions is stronger in Gua-IL than GuaHCl.

Moreover, energy analysis suggests that Gua^+ is comparatively less stabilized in GuaHCl than in Gua-IL (Figure AI.5).

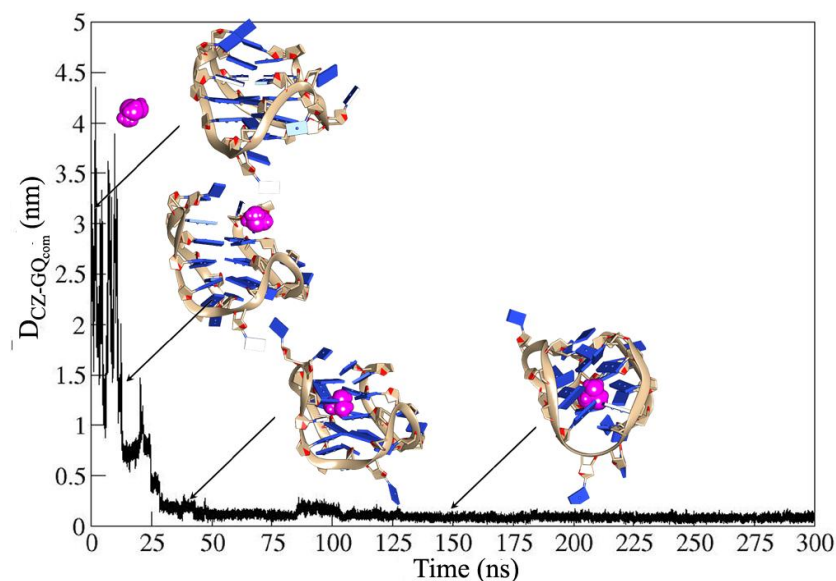


Figure AI.6. Distance between carbon atom (CZ) of Guanidinium ion (Gua-1) and Center of the mass of G-quadruplex quartet region along time. The Guanidinium ion is shown in magenta color. The snapshots belonging to certain timeframes such as 0 ns (initial frame), 21 ns (Gua⁺ reaches close to GQ), 30 ns (Insertion of Gua⁺ inside GQ) and 150 ns (stabilization of GQ by single Gua⁺ ions) are shown.

We have visualized the trajectory of Gua⁺ entering into the GQ core in Gua-IL simulations at 300 K (Figure AI.6). These simulations have been performed multiple times with different initial velocity distributions, ensuring that the observations are statistically reliable. These simulations suggest that initially Gua⁺ ions of Gua-IL interact strongly with GQ at terminal regions and in the grooves. Subsequently, one of the Gua⁺ ions enters the G-tetrad core and forms hydrogen bonds with guanine residues. These hydrogen bonds are transient. We also attribute the stability towards the strongly bound, long-lived certain Gua⁺ ions in the groove of GQ.

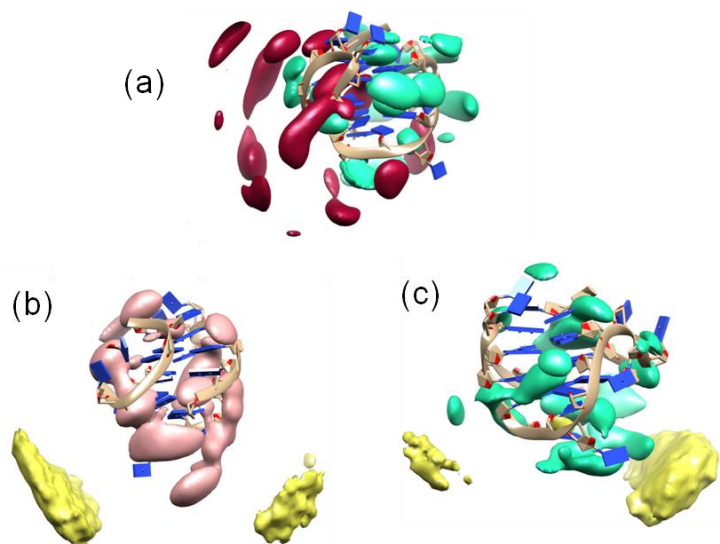


Figure AI.7. The Figure showing spatial distribution function of cations and anions around G-quadruplex as shown in (a) side view presentation of Gua⁺ (Green) and FEP⁻ (dark red) in case of Gua-IL; (b) BPy⁺ (light pink) and Cl⁻ (yellow) in case of BPyCl and (c) Gua⁺ (Green) and Cl⁻ (yellow) in case of GuaHCl.

The strong interactions between Gua⁺ and FEP⁻ (anion of Gua-IL) create a strong ion-pair distribution around GQ exerting Cage-like effect as indicated by spatial distribution of ions around GQ (Figure AI.7(a)), which is lacking in BPyCl and GuaHCl (Figure AI.7(b) and (c)).

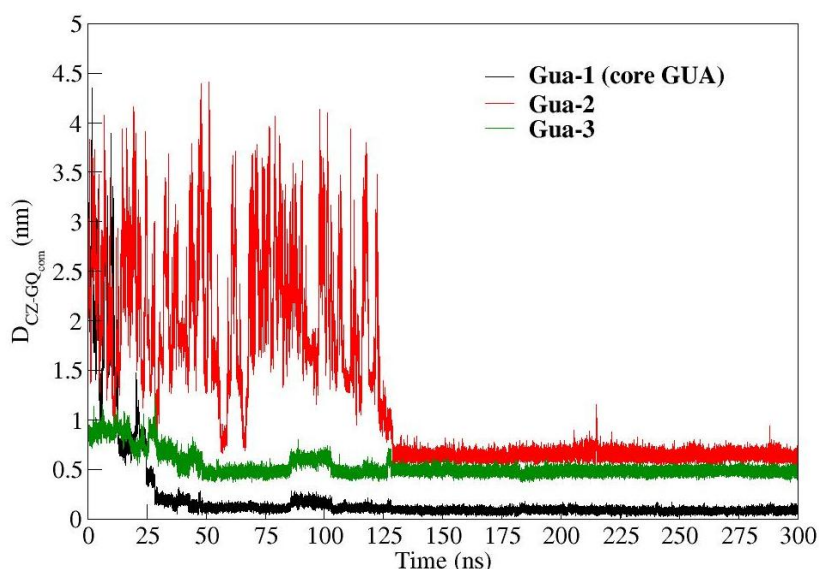


Figure AI.8 The figure showing distance between closest Gua⁺ residues such as Gua-1 (Gua⁺ present in the GQ core, black line), Gua-2 and Gua-3 (strongly bound from outer side, red and green line) along time. The carbon atom of Gua⁺ (CZ) and CoM of quadruplex quartet is considered for calculations.

We have plotted the distance between the central carbon atom of Gua⁺ (CZ) and the center of mass (CoM) for Guanine quartet region to demonstrate the trajectory of Guanidinium ions (Figure AI.8). We have considered 3 Gua⁺ ions mainly the one which translocate in the G-quartet core (Gua-1) and two others (Gua-2 and Gua-3) that are strongly bound on the outer side of G-quartet core. Figure AI.8 shows that one of the Gua⁺ (shown as Gua-1) is initially 2.6 nm away from the core region. It then samples the conformational space and eventually lands into the core region around 40 ns and remains strongly bound there. Surprisingly, this process seems spontaneous in repeated simulations.

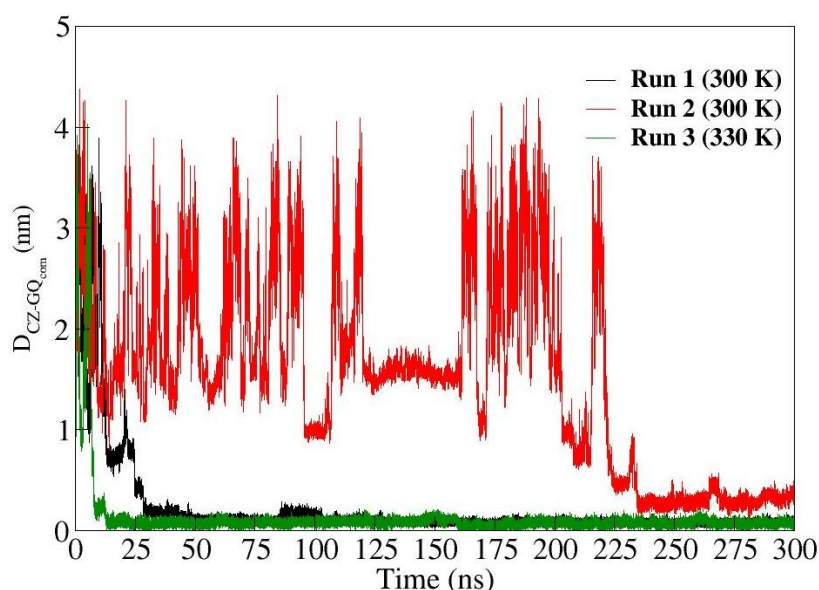


Figure AI.9. Distance between central Gua⁺ (carbon atom of Gua⁺ (CZ)) and CoM of G-quadruplex quartet region along time for Run 1 (black line) and Run 2 (red line) at 300K as well as for Run 3 (green line) performed at 330 K.

We have repeated Gua-IL/GQ simulations three times, two simulations at 300 K using different initial velocities as mentioned above and a third one at 330 K. The distance between CZ and the CoM of GQ is plotted for all these three simulations (Figure AI.9) which shows that at each time, Gua⁺ enters into the core of GQ.

In our study, we notice that Gua-IL is the only IL among three (Gua-IL, BPyCl, and GuaHCl) to instigate and stabilize the GQ structure. Interestingly, Gua-IL comprises a smaller cation (Gua⁺) and bulky anion (FEP⁻) while BPyCl contains a bigger cation (BPy⁺) and a smaller anion (Cl⁻). GuaHCl, on the other hand, consists of same cation as that of Gua-IL, but with a smaller anion i.e. Cl⁻. Above observations

suggest that a smaller cation with a bulky anion pair may have some edge over other ion pairs for GQ formation and stabilization. Based on the experimental and theoretical results, we can hypothesize that IL ion-pair constituting smaller cations can penetrate into the G-tetrad region if bulky anions support it by surface binding. In this study, the combined effect of Gua⁺ and FEP is majorly responsible for GQ formation and stabilization that is lacking in BPyCl and GuaHCl.

AI.4 Conclusion

In this study, we have performed molecular dynamics simulations to understand the difference in the interactions between stabilizing and destabilizing ionic liquids with G-quadruplex molecule. The simulations observe that rather than individual cation or anion molecule, the combined effect of ion pair decides the stability of G-quadruplex molecule. The stabilizing ion pair Gua-IL in this case is also observed to make water – GQ interactions more stable as compared to other ion pairs. The current observations are based on the long unrestrained simulations and hence to understand in detail the stabilization mechanism by ionic liquids enhanced sampling techniques are necessary.

References:

- (1) Rhodes, D.; Lipps, H. J. *Nucleic Acids Res.* **2015**, *43*, 8627.
- (2) Colgin, L. M.; Reddel, R. R. *Current Opinion in Genetics & Development* **1999**, *9*, 97.
- (3) Wright, W. E.; Tesmer, V. M.; Huffman, K. E.; Levene, S. D.; Shay, J. W. *Genes & Development* **1997**, *11*, 2801.
- (4) Sfeir, A. J.; Chai, W.; Shay, J. W.; Wright, W. E. *Molecular Cell* **2005**, *18*, 131.
- (5) Chen, Y.; Yang, D. In *Current Protocols in Nucleic Acid Chemistry*; John Wiley & Sons, Inc.: 2001.
- (6) Tateishi-Karimata, H.; Sugimoto, N. *Nucleic Acids Res* **2014**, *42*, 8831.
- (7) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2309.
- (8) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (9) Wernersson, E.; Heyda, J.; Vazdar, M.; Lund, M.; Mason, P. E.; Jungwirth, P. *J. Phys. Chem. B* **2011**, *115*, 12521.
- (10) Zhang, X.; Huo, F.; Liu, Z.; Wang, W.; Shi, W.; Maginn, E. J. *J. Chem. Phys. B* **2009**, *113*, 7591.
- (11) Sambasivarao, S. V.; Acevedo, O. *J. Chem. Theory Comput* **2009**, *5*, 1038.
- (12) Arfken, G. In *Mathematical Methods for Physicists (Third Edition)*; Arfken, G., Ed.; Academic Press: 1985, p 428.
- (13) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (14) Nose, S. *Mol. Phys.* **1984**, *52*, 255.
- (15) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182.
- (16) Tom Darden, D. Y., Lee Pedersen *J. Chem. Phys.* **1993**, *98*, 10089.
- (17) Harvey, S. C.; Tan, R. K. Z.; Cheatham, T. E. *J Comput Chem* **1998**, *19*, 726.
- (18) Brehm, M.; Kirchner, B. *J. Chem. Inf. Model* **2011**, *51*, 2007.
- (19) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J Comput Chem* **2004**, *25*, 1605.

Addendum II

Ionic Liquid Prolongs DNA Translocation through Graphene Nanopore

AII.1. Introduction

The detection of DNA molecules at a single base resolution by employing nanopore-based technology has paved the way for high throughput, label-free and low-cost personalized genome sequencing techniques. The method is based on a simple principle of measuring the reduction in ionic current while DNA passes through a nanopore under the influence of an external voltage. The increasing attempts to initiate practical application of the nanopore-based technique for genome sequencing has not reached National Human Genome Research Institute's (NHGRI) proposed target price of \$1000[1] due to various factors.

The biological pores like α -Hemolysin, *Mycobacteria smegmatis* porin A (MspA), as well as solid-state nanopores are proved suitable materials for translocation studies with their intrinsic advantages and disadvantages[2]. The application of biological pores is limited due to less mechanical strength and pronounced sensitivity towards pH, temperature, and salt concentration. Nanopores grafted in solid-state materials such as single-layer of graphene, MoS₂, or Hafnium oxide are observed to be efficient and alternative materials to biological nanopores. Solid-state nanopores allow calculation of tunneling current as well as ionic blockade current and provide better accuracy to achieve single base detection.

During translocation experiment, DNA is divided into fragments of ~50000 base pairs and passed repetitively via nanopore to improve single base reading. Under the influence of the voltage in the order of 10^{-3} volts, the translocation speed is approximately 2 μ s per base. The generation of pico-ampere level ionic currents require high bandwidth amplifier for detection and are associated with high magnitude pink noise in graphene nanopores. Thus, DNA sequencing using solid-

state nanopore is still challenged by fast translocation speed and possesses large error in single base detection of a DNA sequence.

The translocation speed of a DNA depends on various factors such as diameter of the nanopore (d), temperature, viscosity of a solution, applied electric field, pressure, and interaction of DNA with electrolyte ions. The relationship between viscosity of a solution (η) and DNA translocation[3,4] speed (v) is given by,

$$v = \varepsilon \frac{\zeta_d - \zeta_w}{\eta} E, \quad \text{Eq. AII. 1}$$

where, ε is the dielectric constant of the electrolyte. ζ_d and ζ_w are zeta potentials of DNA and nanopore surface, respectively. E is the applied electric field. Hence, DNA translocation speed can be reduced (a) by increasing viscosity of electrolyte, (b) by applying low electric fields or (c) by reducing zeta potential difference between the DNA and nanopore ($\zeta_d - \zeta_w$).

Fologea, *et al.*[5] used glycerol-water solutions to slow down DNA translocation through solid-state silicon nitride nanopores. The study suggested that combined use of reduced temperature, low bias voltage, and high glycerol-water (50%) ratio would effectively slow down DNA translocation speed. The same study indicated that there exists an interplay between the dwell time (t_d) and the amplitude of blockade ionic current or current drop, when one attempts to modulate other parameters like temperature, viscosity, and salt concentration. The dwell time is the mean translocation time for which the open pore current gets reduced. Reduction in the translocation speed also reduces the ionic conductivity, leading to a decreased blockade signal.

The DNA molecule is recognized by both dwell time and current drop (ΔI_b). The current drop (ΔI_b) is related to the thickness of nanopore (H), the solution conductivity (σ), hydrodynamic radius of the translocating molecule (R_{DNA}) and the applied voltage as,

$$\Delta I_b = \frac{\sigma V R_{DNA}}{H}, \quad \text{Eq. AII. 2}$$

The dwell time depends on the viscosity of electrolyte as,

$$t_d = K \frac{\eta L_{DNA}}{\lambda V}, \quad \text{Eq. AII. 3}$$

where, η is viscosity of solution. λ and L_{DNA} are linear charge density and length of DNA molecule, respectively. V is applied voltage and K is a constant for

complex possible conditions. The increased interaction between the nanopore surface and DNA governs the translocation speed[6]. The experimental and MD simulation study performed by Mirsaidov *et al.*[7] showed that the DNA translocation speed can be reduced up to 1 bp / 2 ms by implementing reduced pore diameter (~2 nm) and by modulating voltage and salt concentration (< 200 mV and 100 mM KCl). With such reduced speed, enhanced base-pair discrimination between C-G and G-C base pairs was observed. Akahori and coworkers[8] reduced nanopore diameter to enhance graphene-ssDNA interactions, which reduced translocation time per base up to 1.4 μ s.

Wanunu and coworkers[9] have shown the effect of salt gradient on DNA translocation motions. The “funnelling effect” exerted by salt gradient displays increased current rates and improves translocation times. Molecular dynamics simulation studies[10] also corroborated these effects and suggested that nanopore surface charges play an important role in trapping and translocation of polymers like dsDNA.

The nanopore created by high intensity electron beam using transmission electron microscope (TEM) generally assumes negative charge. The electric field bias is needed to maintain electro-osmotic flow of ions near pore surface. Crowding of ions near the pore reduces translocation speed[11]. Negatively charged surfaces create competition in hydrodynamic coupling between the electro-osmotic flow of ions and DNA motions[3], thus retarding DNA translocation through the nanopore. The simulation study[12] also revealed that the reversal of DNA’s electrophoretic motion is caused by the strong binding of large organic cations like spermine (+4) and spermidine (+3).

The translocation speed can also be reduced by lowering the temperature[5,13] or using small diameter nanopores and high pressures[14,15]. In a theoretical study of pressure driven translocation, it is shown that the current reversal happens only in the presence of trivalent or higher valency cations. Kowalczyk, *et al.* used LiCl solution[16] to reduce DNA translocation speed. This study indicated that strong binding of Li⁺ ions in the minor groove, primarily with O4’ atoms as well as N3 (of Cytosine and Thymine) and O2 (Adenine and Guanine) atoms of bases, reduces DNA translocation.

The increased viscosity and strong DNA-ion interactions can also slow down DNA translocation. Recently, room temperature ionic liquid BMIM-Cl[17] was used

to retard DNA speed while translocating through mutated α -hemolysin nanopore. Both short-lived (~ 50 - $100 \mu\text{s}$) and long-lived ($\sim \text{msec}$) translocation events were observed, albeit with a decreased current. The low current was attributed to the lower conductivity of 1 M BMIM-Cl (49.9 mS/cm) compared to 1 M NaCl solution (81.6 mS/cm)[17]. The study showed that there exists a strong interaction between BMIM cations and DNA, which hinders the DNA to squeeze through the nanopore.

The thickness of a single layer graphene sheet is 0.34 nm , similar to the DNA base pair separation. This makes the single layer of graphene a suitable material to achieve single base resolution through ratcheting motions of DNA bases. However, the effective use of graphene nanopores is limited by low (< 3) signal to noise ratio (SNR). Also, the sticking of DNA to graphene surface and nanopore restricts its effective use[18,19]. Now-a-days different materials like MOS_2 [20,21] and hafnium oxide[22] are implemented to achieve enhanced SNR (> 15). In case of MOS_2 , the hydrophilic-hydrophobic-hydrophilic (S/Mo/S) arrangement of atoms reduces noise in the ionic current. Also, the problem of DNA sticking to pore surface is absent in MOS_2 nanopores[20,23]. Recently, Ionic liquid (BMIM- PF_6) induced viscosity gradient was used with MOS_2 nanopore to reduce translocation speed and achieved a high SNR (> 10)[21]. HfO_2 membranes with less than 2.0 nm nanopore are also proved as an effective material for DNA translocation[22].

Modi and coworkers[24] studied translocation behavior of BMIM-Cl in biological nanopore. Attaining the right orientation of BMIM^+ molecule near nanopore's constriction region was observed to be the rate-limiting step in its translocation. This orientation of BMIM^+ was speculated as an important factor to slow down antibiotic translocation[24]. The passage of room temperature ionic liquid (RTIL) bulky ions via nanopore channel creates ionic clouds near nanopore surface. This accumulation of bulky ions alters the effective electric field and affects translocation of molecules[25].

The single base resolution efficiency of a single layered graphene nanopore is reported in the literature but its application is still limited mainly by the fast translocation speed as well as graphene attachment of DNA on trans side during the translocation process. The prolonged dsDNA translocation in a mutated α -Hemolysin nanopore was achieved[17] using BMIM-Cl as an electrolyte. However, the possible

application of ionic liquids to reduce translocation speed in solid-state nanopore like graphene was not demonstrated.

In the present study, we have investigated possible role of BMIM-Cl in reducing DNA translocation speed through a single graphene layer. Various simulations were performed to explore the effect of nanopore diameter and voltage on the translocation behavior of several homo-nucleotide sequences dA₂₀, dC₂₀, dG₂₀ and dT₂₀. Comparisons are drawn with normal salt such as KCl for d(ACGT)₅ sequence based system. The results indicate that single base resolution can be improved using smaller nanopore at smaller voltage gradient in the presence of ionic liquids. The translocation time increases as compared to KCl solutions. The increase in translocation time is due to slow passage of strong DNA-BMIM⁺ complex via smaller nanopore.

AII.2. Simulation details

AII.2.1. System setup

We designed five DNA sequences, each containing 20 base pairs, as dA₂₀, dC₂₀, dG₂₀, dT₂₀, and d(ACGT)₅. The length of the DNA ensures that there are two helical repeats. Canonical B-form structures of DNA for the above sequence were constructed using NAB[26] module of Ambertools[27]. The CHARMM27 force field parameters were chosen to perform the simulation. Armchair graphene sheet of approximately 9.8 nm x 9.8 nm was created using Nanotube builder in VMD[28]. The atoms within the required radius from the center atom were removed to create the pore. The ionic liquid 1-butyl-3-methylimidazolium chloride (BMIM-Cl) was used as an electrolyte. The BMIM optimized geometry was obtained using hf/6-31G*. The parameters were derived from Ref. [29]. The potassium and chlorine parameters are chosen from CHARMM27 force field for ions. GROMACS 4.5.3[30] version was used to perform simulations. Each DNA-ionic liquid system was solvated with CHARMM27 TIP3P[31] water. Ions were added randomly to make 1 M concentration for both BMIM-Cl and KCl solutions.

The interaction between DNA and graphene nanopore is a crucial parameter to achieve base specific recognition. Hence, the choice regarding nanopore size is critical, as the large diameter pores possess fewer interactions with translocating DNA sequence and small pores can completely stop electro-osmotic flow of ions.

Generally, a nanopore of radius 1.2 nm was used as it is close to the radius of a dsDNA in B-form. In case of BMIM-Cl, however, it was observed that use of such a small nanopore allowed only dsDNA to pass and not bulky molecules of BMIM⁺ ions (data not shown). This observation led us to choose nanopore radius by considering hydrodynamic radius of BMIM⁺ cations.

The hydrodynamics diameter of BMIM is 0.57 nm[32]. We have considered two possibilities assuming dsDNA as a cylinder: (a) BMIM⁺ molecules binding preferentially to either major or minor groove sides of a DNA (see Fig. AII.1 (a)) and (b) BMIM molecules binding to both grooves (see Fig. AII.1 (b)).

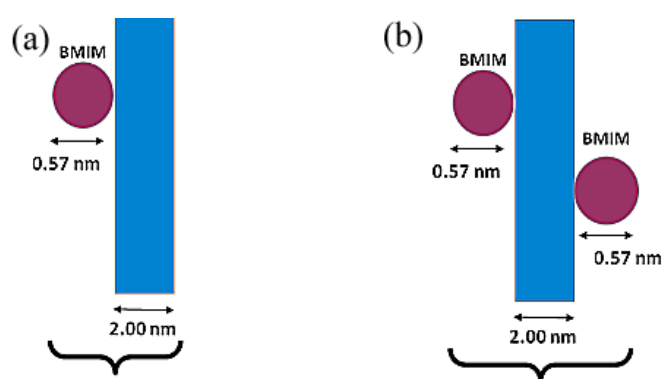


Figure AII.1. Schematic diagram showing choice of pore radius based on DNA and BMIM ion size.

In the first case, the equivalent diameter of DNA and attached BMIM would be 2.57 nm, thus requiring a nanopore of 2.6 nm diameter. In the second case, the effective diameter would be 3.14 nm. Thus, the DNA translocation simulations were performed for nanopores with diameters 2.6 nm (smaller nanopore) and 3.1 nm (bigger nanopore). For each nanopore, we performed the calculations at two different voltage gradients, 0.05 V/nm (smaller voltage gradient) and 0.10 V/nm (higher voltage gradient). These simulations are listed in Table AII.1.

Table AII.1. Table showing details of the systems and simulations performed in this study.

Simulation ID	Sequence	Electrolyte	Voltage (V/nm)	Pore Diameter (nm)	Translocation time (ns)
1	dA ₂₀	BMIM-Cl	0.05	2.60	5.00
2	dC ₂₀	BMIM-Cl	0.05	2.60	5.00
3	dG ₂₀	BMIM-Cl	0.05	2.60	5.00
4	dT ₂₀	BMIM-Cl	0.05	2.60	5.00
5	dA ₂₀	BMIM-Cl	0.05	3.10	2.50
6	dC ₂₀	BMIM-Cl	0.05	3.10	2.50
7	dG ₂₀	BMIM-Cl	0.05	3.10	2.50
8	dT ₂₀	BMIM-Cl	0.05	3.10	2.50
9	dA ₂₀	BMIM-Cl	0.10	2.60	1.50
10	dC ₂₀	BMIM-Cl	0.10	2.60	1.50
11	dG ₂₀	BMIM-Cl	0.10	2.60	1.50
12	dT ₂₀	BMIM-Cl	0.10	2.60	1.50
13	dA ₂₀	BMIM-Cl	0.10	3.10	1.00
14	dC ₂₀	BMIM-Cl	0.10	3.10	1.00
15	dG ₂₀	BMIM-Cl	0.10	3.10	1.00
16	dT ₂₀	BMIM-Cl	0.10	3.10	1.00
17	d(ACGT) ₅	BMIM-Cl	0.05	2.60	3.50
18	d(ACGT) ₅	KCl	0.05	2.60	2.00
19	d(ACGT) ₅	BMIM-Cl	0.05	3.10	1.70
20	d(ACGT) ₅	KCl	0.05	3.10	1.10

Open pore current simulations were performed to recognize mean ionic current value and its fluctuations. This study was extended for pore diameter of 2.60 nm at different voltages to observe I-V relationship in case of BMIM-Cl based system. Also, the relationship between resistance and diameter of the nanopore is studied using various open pore simulations. These simulations were performed with different pore diameters (2.40 nm, 2.60 nm, 2.80 nm, 3.00 nm, 3.10 nm, 3.40 nm, 3.80 nm and 4.6 nm) at the higher voltage gradient (VG). These simulations were performed for 5 ns each.

The graphene sheet was aligned along XY plane and the dsDNA was aligned along Z-direction. Two terminal base pairs of DNA were inserted inside pore to avoid unwarranted DNA diffusion in the solvent in search of a nanopore (See Fig. AII.2). The end part of graphene sheet was restrained. The graphene sheet was modelled as a continuous periodic molecule to avoid end effects. Periodic boundary condition was

applied along XYZ directions. The final box size of the system is approximately 9.8 nm x 9.8 nm x 19.8 nm.

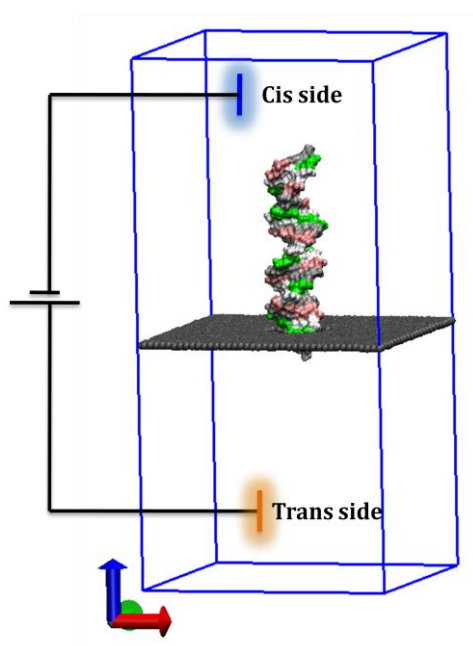


Figure AII.2. Snapshots showing initial configuration. Bottom panel is a zoomed view of the top panel.

Particle Mesh Ewald method[33] was used to calculate electrostatic interactions with a cut-off value of 1.3 nm. We took 1.2 nm cut-off radius for vander Waals interactions. Bonds involving hydrogen atoms were constrained during simulation using LINCS algorithm. We have carried out 20 simulations with DNA that involved four homo-nucleotide sequences at different conditions and four other simulations using the hetero-nucleotide sequence d(ACGT)₅. Each system contains approximately 184000 atoms. Each system was simulated three times, starting with a different initial velocity, to obtain the error estimate.

AII.2.2 Equilibration and Data Collection

Each system was heated slowly up to 300 K followed by a 50 ps equilibration using Berendsen barostat and thermostat[34]. The production runs were performed using NVT ensemble and by implementing V-rescale[35] as a thermostat. A recent study showed V-rescale thermostat maintains canonical ensemble and reduces ionic current fluctuations[36]. Apart from 20 simulations monitoring DNA translocation under various conditions, four open pore simulations were performed for nanopore diameters 2.6 nm and 3.1 nm at both 0.05 and 0.10 V/nm. For the smaller nanopore, ten open pore simulations were performed at ten different voltages to confirm that the

system obeys Ohm's law and to study I-V relationship in ionic liquid-based solid-state nanopore.

AII.2.3. Analysis

(a) DNA-graphene COM movement:

Since graphene sheet is placed in the X-Y plane, movement of the DNA and its translocation was determined by the change in the Z component of the distance between the center of masses of the DNA and the graphene sheet. The Z-component of DNA movement (D_Z) is normalized by its value at $t = 0(D_0)$. We have calculated normalized movement (D_Z/D_0) in all cases.

(b) Ionic blockade current:

As the electric field was applied along Z-direction, time-dependent ionic current was calculated from the Z-component of ions using the following equation[37],

$$I(t) = \frac{1}{\Delta t L_z} \sum_{i=1}^N q_i [z_i(t + \Delta t) - z_i(t)], \quad \text{Eq. AII. 4}$$

where, L_z is the box length along Z-axis. Δt is the time lag, which depends on the electrolyte of the system. It was suggested that Δt should be chosen such that signal to noise ratio would be larger than five[38]. q_i is the charge on the i -th atom. In case of BMIM-Cl, all atoms of BMIM^+ and Cl^- ions were considered to calculate the ionic current. Similarly, for KCl solution based systems, current due to K^+ and Cl^- were considered.

(c) Radial distribution Function:

We have studied radial distribution function of cations (BMIM^+ or K^+) around DNA phosphate groups. In case of d(ACGT)_5 , we have also monitored the distribution of Cl^- ions.

(d) Charge distribution calculations:

The charge distribution due to cations and anions for DNA translocation simulations in KCl and BMIM-Cl systems was calculated. The charge distribution along the Z-axis was calculated by dividing the simulation box into 200 slices perpendicular to Z-axis and using `g_potential` program in GROMACS[30]. The number of charged atoms belonging to a particular slice provides the effective charge

distribution caused by applied electric field. This potential is averaged over the number of frames and plotted along length of a box in Z-direction.

AII.3. Results and Discussion

AII.3.1. Open pore current analysis

(a) Signal to noise ratio (SNR):

The success of nanopore sequencing technique depends on slow translocation speed of DNA with a high signal to noise ratio (SNR) in blockade current measurement. The open pore current calculations provide an opportunity to understand SNR and conductance behavior of a simulation system under study. These studies help reveal the effect of time lag on SNR ratio and I-V curves of the system specific to simulation parameters like force field, thermostat, and other factors. Hence, the reliability of MD simulations to explore DNA translocation can be ascertained from open pore current calculations.

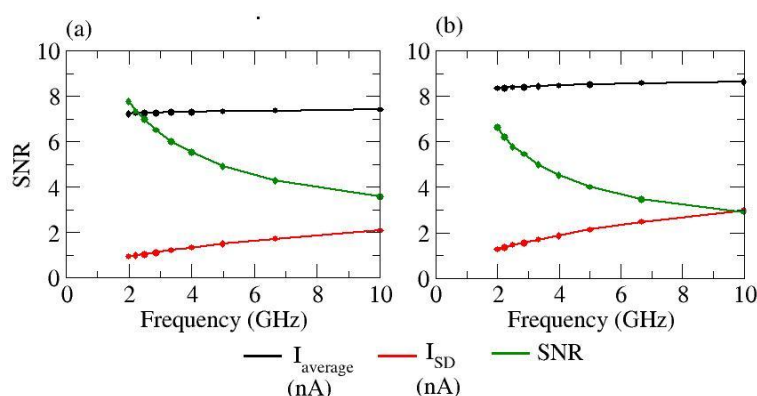


Figure AII.3. SNR for diameter (D) 2.6 nm at voltage gradients (VG) at (a) $V = 0.05$ V/nm (b) $V = 0.10$ V/nm

The ionic current due to the passage of BMIM-CL molecules through the graphene nanopore can be calculated using Eq. AII.4. The ionic current calculations are sensitive to the time lag (Δt), the choice of which in turn dictates SNR. We have studied SNR dependence on Δt for BMIM-Cl based systems to find the appropriate value of Δt , so that SNR remains above five. We have estimated SNR using open pore current, which is the current, estimated in the absence of the DNA containing only 1 M BMIM-CL solution and the nanopore of a given size. Open pore simulations were performed for two different nanopore diameters, 2.6 nm and 3.1 nm, at two different voltages of 0.05 V/nm and 0.10 V/nm. SNR was then calculated using Eq. AII.5.

Figure AII.3 suggests that the recommended SNR value (above 5) is observed at 4 GHz (equivalent to 250 ps time lag). Hence, all further simulations to estimate currents in case of BMIM-CL based systems were performed using $\Delta t = 250$ ps. This value is almost five times higher than the suggested value of 50 ps in case of KCL solutions[36]. The SNR [39] is defined as,

$$SNR = \frac{I_{avg}}{I_{SD}}, \quad \text{Eq. AII. 5}$$

Here, I_{avg} is average open pore ionic current and I_{SD} is a standard deviation of ionic current when a particular Δt is used. The noise in spectral density of a signal is inversely proportional to the frequency (f) and possesses a stochastic origin; such noise is called 1/f noise. At high frequencies, 1/f noise is comparable to the magnitude of the signal and hence provides a poor detection. For a single conical nanopore, the use of ILs like BMIM-TFSI and EMIM-SCN provide better SNR, due to crystal-like self-assembled structures[40]. The eventual slow dynamics of the translocation of ionic liquids exhibit jammed-like[41] motions which enhances the SNR[40]. Hence, use of ionic liquid may help to enhance SNR in solid-state nanopores.

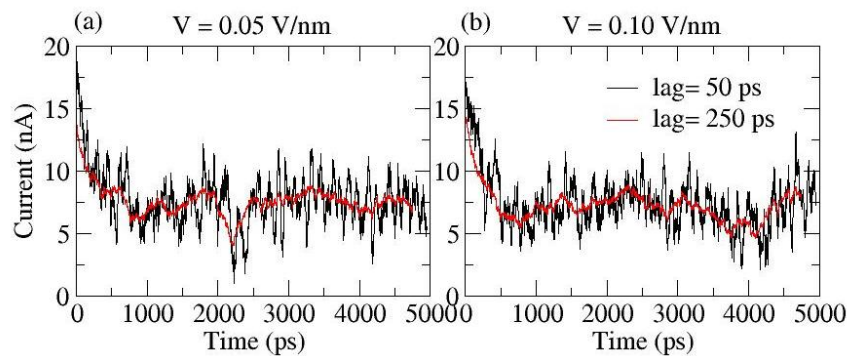


Figure AII.4. Open pore current behavior for nanopore of diameter 2.6 nm at (a) 0.05 V/nm and (b) 0.10 V/nm.

Figure AII.4 indicates that for the smaller nanopore, the average current values are 7.39 nA and 7.70 nA for the smaller and larger voltage gradients, respectively. Similarly, the open pore current values for the bigger pore are 8.28 nA and 9.92 nA for the smaller and larger voltage gradients, respectively. Figure AII.4(a) and AII.4(b) indicate that the average current values remain almost same in both cases. The choice of Δt is independent of voltage applied on the above systems. Also, it indicates that one needs high frequency apparatus to detect currents of such low values, eventually increasing the chance of instrumental errors. Such errors can be reduced by using

Al_2O_3 nanolaminated graphene membranes as shown earlier[42]. It is also shown that nanopores with larger diameter possess better SNR in low salt regions[39].

AII.3.2. Variation of current against voltage:

Since we used a non-conventional electrolyte, we wanted to verify if BMIM-CL follows Ohm's laws using voltage and current relationship. For this, we chose the smaller nanopore and performed simulations at ten different voltages ranging from 0.01 to 0.20 V/nm. Each simulation was performed for 5 ns. The current was calculated in each case using Eq. 3 for two different time lags (50 ps and 250 ps).

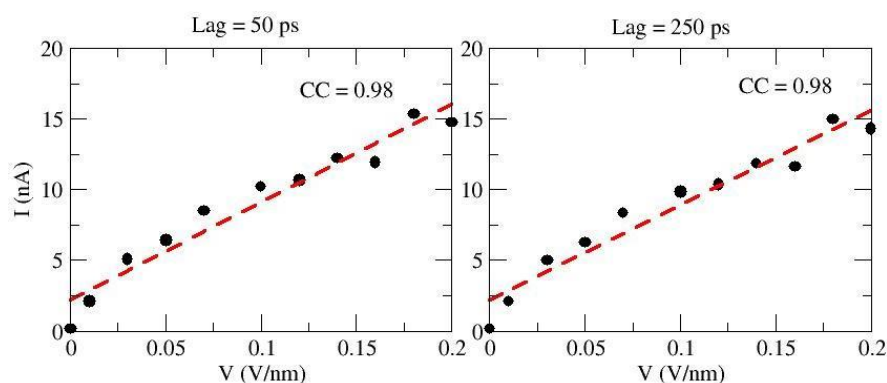


Figure AII.5. I-V characteristic of the system.

The I-V plot (Fig. AII.5) shows linear relationship between the voltage and current as expected by Ohm's law. However, it possesses a nonzero intercept on the ordinate. The simulation at 0 V, performed to check the values of ionic current and thermal fluctuations in the system, showed the presence of 0.01 nA current. The open pore current and graphene pore conductance were calculated to determine the accuracy of MD simulations. Open pore current study predicts higher bulk conductivity for 1 M BMIM-Cl than experimental bulk conductivity value as shown below. The bulk conductivity of BMIM-Cl solution is calculated using the method implemented by Aksimentiev et al[37] in the previous study. The slope of I-V plot in Fig. AII.5 provides a direct measurement of the bulk conductivity. The bulk conductivity is found to be 67.1 S/m, which is higher than experimental value 4.62 S/m. The study by Modi *et al.*[24] also observed a slightly enhanced conductivity for BMIM-Cl using TIP3P water model. Hence, the unusual high conductivity observed in this study may be attributed to the BMIM-Cl force field parameters and partly due to the TIP3P water model.

The accuracy of electrochemical properties of IL depends on force field of IL as well as water model. We would like to stress that current IL force field model is chosen on the basis of its capability to present accurate DNA-IL interactions,[43]at the cost of loosing accuracy in the electrochemical property. We believe that even though the bulk conductivity value is high, the qualitative trends in translocation will remain unchanged.

AII.3.3. Behavior of current:

Figure AII.4 shows open pore current fluctuations through the smaller nanopore at two different voltage gradients. Open pore current fluctuations are significant in both the cases, may be due to the fluctuation in the effective unoccupied area of a nanopore by bulky BMIM ions, as demonstrated earlier[44]. Also, we can see a dip around 2200 ps in case of smaller voltage gradient (Fig. AII.4(a)). Hence, bulky molecules like BMIM^+ can introduce a current blockade. The average current in both cases is similar -- 7.39 nA for 0.05 V/nm and 7.70 nA for 0.10 V/nm.

AII.3.4 Resistance and pore diameter relationship:

We have observed the dependence of resistance on the pore diameter, which was found to follow a nonlinear relationship. The nanopore size can affect the conductance of bulky BMIM^+ cations due to electrostatic interactions[45] with graphene surface leading to a different arrangement of cations near the graphene interface. The conductivity of ionic liquids depends strongly on the size of ions and pore diameter. Experimentally low conductivity was observed in case of pores having diameters less than 20 nm according to the hindered transport theory[46,47]. If thickness of the electric double layer is comparable with the pore diameter, then the ionic current rectification is observed. This rectification generates asymmetry in the I-V curve. The experimental[48] and previous theoretical study[18] based on the KCL electrolyte observed that the resistance varies inversely with the square of the diameter of a nanopore.

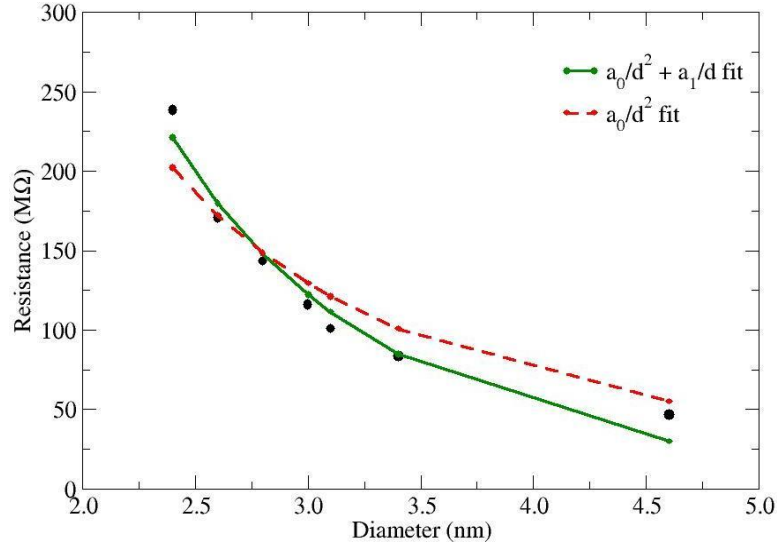


Figure.AII.6. Dependence of resistance on the pore diameter.

Surprisingly, Fig. AII.6 shows that for BMIM-Cl based system the data from simulation fits well with the Hall's equation for total resistance of a nanopore as shown below[49]

$$R = a_0/d^2 + a_1/d, \quad \text{Eq.AII. 6}$$

where, $a_0 = 4L/\pi \sigma$ and $a_1 = 1/\sigma$, L is length of the nanopore, σ is the bulk conductivity of the BMIM-Cl solution, d is the diameter of nanopore. The calculated values obtained here are $a_0 = 1976.41$, $a_1 = -292.17$ with correlation coefficient of 0.98 and $\chi^2 = 834.4$. A good fit to Eq. AII.6 indicates the presence of an enhanced nonlinear dependence of resistance of the system on the pore diameter. The first term of Eq. AII.6 represents the channel resistance and second term of Eq. AII.6 represents the access resistance. The access resistance arises from a resistance of the medium outside nanopore and causes significant reduction in potential near pore mouth. The higher values of both a_0 and a_1 indicate that the channel resistance as well as access resistance significantly contributes to the total pore resistance.

The current study does not allow to comment on the negative value of a_1 . However, based on Eq. AII.6, we would like to speculate possible reasons for the negative value of coefficient a_1 . First, we assume that conductivity of a BMIM-Cl solution is different in the bulk solution ($\sigma_{channel}$) and near graphene nanopore (σ_{access}). The negative conductivity for the access resistance can be assumed due to reversal of the current near the small region of nanopore. At the microscopic level, some BMIM⁺ molecules might be dragged against the direction of the potential

gradient due to strong binding with DNA molecule. These bound BMIM⁺ cations may get stripped off near the nanopore surface while translocating and creates negative access resistance near the nanopore. However, we strongly believe that further study is required to verify this effect and to understand whether this phenomenon is only a local effect or an overall signature for these particular types of IL-nanopore systems.

AII.3.5. Polynucleotide translocation time

The long DNA sequences threading through the small aperture are discriminated based on signature translocation time and ionic current blockades of individual bases. To understand the effect of IL on the DNA translocation, we have performed a series of simulations with different DNA base pairs, different sizes of nanopores and different voltage gradients.

The initial structure of a simulation contains the DNA partially inserted in the nanopore. Hence, the time required for a DNA sequence to completely pass through the pore may be considered as the “translocation time” for the DNA. The translocation time is decided on the basis the center of mass (COM) position along the Z-axis, also confirmed by visual inspection of trajectories.

For the smaller nanopore at both voltages, DNA sticks to the graphene surface after translocation on trans side causing partial blockades (See Fig. AII.7). In such cases, the translocation time is estimated when last atom of DNA passes through the pore and no partial blockades are observed.

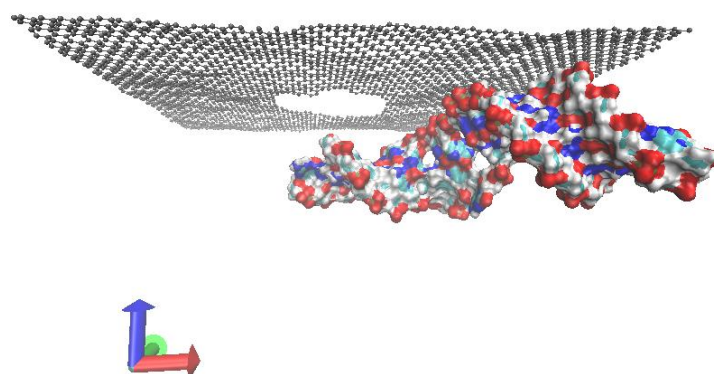


Figure AII.7. Configuration showing dG₂₀ stuck to graphene surface for nanopore of 2.6 nm diameter and voltage of 0.05 V/nm.

The DNA COM movement projected on Z-axis is normalized in Fig. AII.8. It helps to compare the behavior of translocation time under different conditions, as D_0 value for different sequences varies slightly due to inherent difference in COM of the sequences. Previous studies[39,50] have shown that smaller nanopores and low electric fields enhance the interaction between solid -state nanopore and DNA, eventually providing better discrimination of the bases. Figure AII.8 indicates that smaller nanopore at smaller voltage gradient (VG) and, surprisingly, bigger nanopore at higher VG, individual bases can be discriminated on the basis of the translocation time.

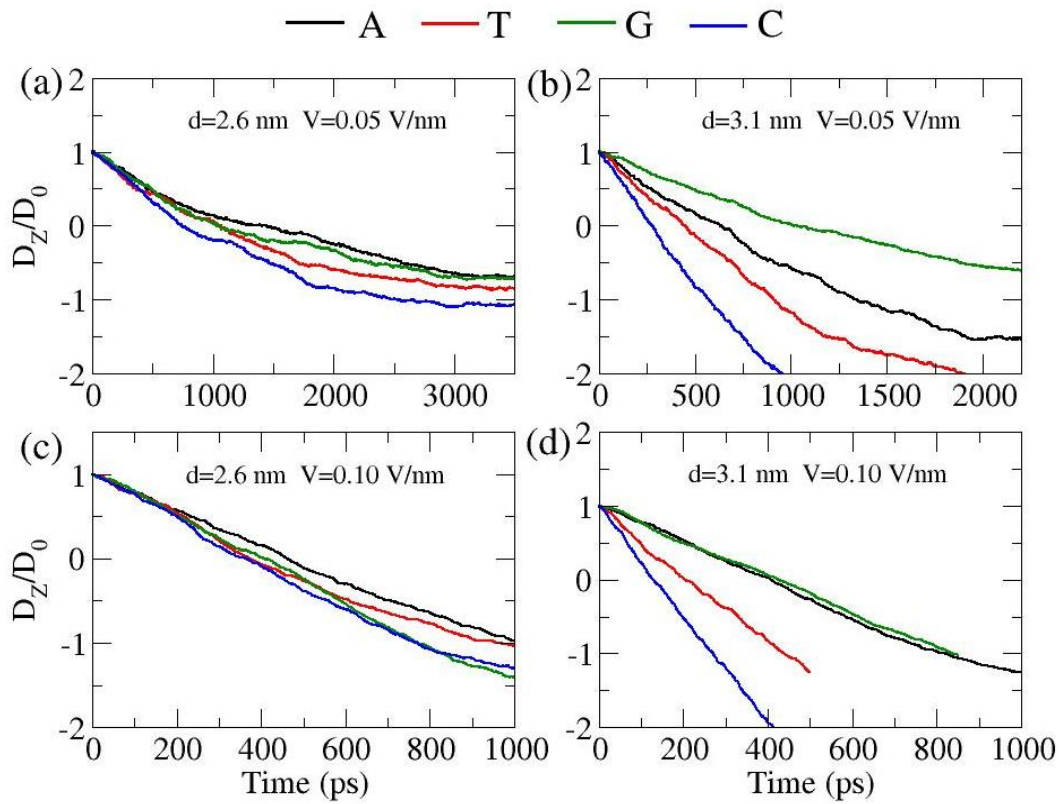


Figure AII.8. The Z-component of DNA movement (D_Z) is normalized by its value at $t=0$ (D_0). We have plotted normalized movement (D_Z/D_0) in all cases. Normalized DNA COM movements for polynucleotide systems (a) Diameter (d) = 2.6 nm, $V=0.05$ V/nm; (b) $d=3.1$ nm, $V=0.05$ V/nm; (c) $d=2.6$ nm, $V=0.10$ V/nm; (d) $d=3.1$ nm, $V=0.10$ V/nm. Color indicators: A- black, T- red, G- green, C- blue.

The normalized DNA COM movement through the smaller nanopore at smaller VG (see Fig. AII.8 (a)) indicates that all bases exhibit translocation time for more than 3000 ps. Such slow translocation is always affected by strong interaction between DNA-graphene interactions, which eventually causes sticking of DNA to the nanopore surface. This effect is observed in Fig. AII.8 (a), where D_Z/D_0 values are

converged beyond 2500 ps. This effect is absent in the larger nanopore of diameter 3.1 nm, where we can see a sharp COM movement indicating less interaction between hydrophobic graphene surface and DNA. The steep slope of COM along time in case of the large nanopore indicates fast translocation process and hence less interaction with graphene. It is known that DNA-graphene interaction affects DNA translocation speed[6].

The normalized DNA COM movement through larger nanopore at the lower VG shows the signature translocation velocities for each type of nucleotide. Hence, dA₂₀, dC₂₀, dG₂₀, and dT₂₀ possess significantly different electrophoretic mobilities under this condition. The increased electric field enhances the electrophoretic force on DNA in a nano-channel. It was shown that current blockades are caused by reduction in ion mobility due to strong interaction between DNA grooves and surface atoms[51]. Also, the avalanche of BMIM⁺ cations moving in the opposite direction provides the desired sequence-dependent translocation time discrimination of the bases.

The similar translocation time for all polynucleotide sequences in case of the smaller nanopore at high VG indicates the absence of distinctive DNA-graphene interactions due to fast translocation and hence it possesses lower resolution for the blockade current. Hence, Fig. AII.9 suggests that effective parameters such as pore diameter of either 2.6 nm or 3.1 nm can be implemented at 0.05 V/nm and 0.10 V/nm voltage gradients, respectively to achieve translocation time to distinguish between nucleotide types.

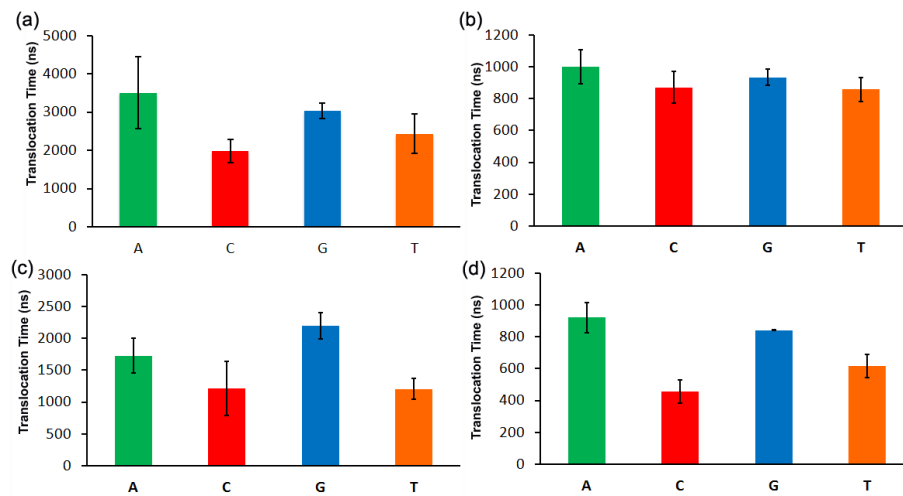


Figure AII.9. Translocation time for (a) smaller nanopore (diameter 2.6 nm) at smaller (0.05 V/nm) voltage gradient (VG); (b) smaller nanopore at higher (0.10 V/nm) VG; (c) Bigger nanopore (diameter 3.1 nm) at smaller VG; (d) Bigger nanopore at higher VG.

Hence, Fig. AII.9 suggests that effective parameters such as pore diameter of either 2.6 nm or 3.1 nm can be implemented at 0.05 V/nm and 0.10 V/nm voltage gradients, respectively to achieve translocation time to distinguish between nucleotide types.

The normalized DNA COM behaviour for above-mentioned effective parameters is almost same for dA₂₀ and dG₂₀ sequences. We assume that this behavior results due to similar cross-sectional area of guanine (56 Å²) and adenine (50 Å²)[52]. The details of molecular interactions which play important role in sequence-dependent behavior are discussed in section AII.3.7. Molecular dynamics studies showed that it is difficult to trace long-lived and short-lived signals for pore diameters of less than 3 nm[53].

The effective force on a charged polymer like DNA can be roughly estimated as $q * E$, where q is the effective charge on DNA and E is the applied electric field. Previous studies have shown that the reduction in the effective charge on DNA is caused by the hydrodynamic drag[54] of electro-osmotic flow of ions. The electro-osmotic flow depends on the local concentration of ionic species as well as the viscosity of the medium.

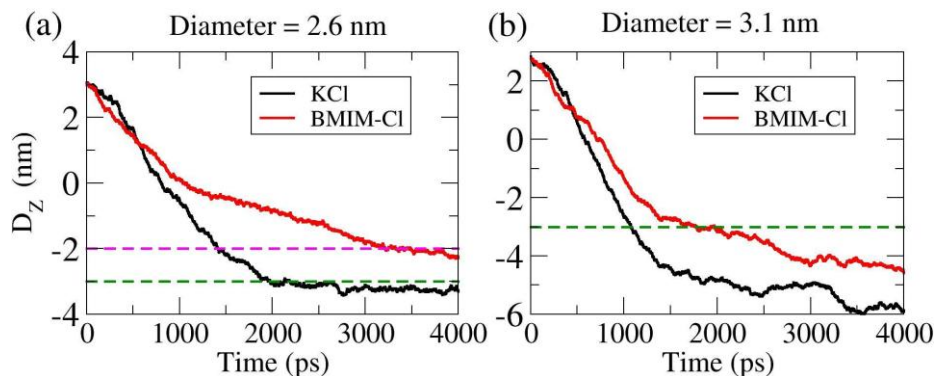


Figure AII.10. Center of mass (COM) displacement of ACGT system in BMIM-CL and KCL at 0.05 V/nm for (a) diameter 2.6 nm and (b) diameter 3.1 nm.

To consider the effect of these conditions on a heterogeneous sequence, we have studied the translocation of d(ACGT)₅ sequence in 1 M KCL and 1 M BMIM-Cl solutions. Figure AII.10 depicts the slowing down of DNA translocation in IL solution, prolonging the total translocation time. The effect is pronounced for the smaller pore. For the smaller pore, translocation in KCl solution was completed around 1.5 ns, whereas the same sequence required approximately 3 ns to complete translocation in BMIM-Cl solution. This indicates two-fold increase in translocation time in case of heterogeneous DNA sequence. The convergence in DNA COM

movement after complete translocation is due to the adherence of the DNA on the trans side at low voltage of 0.05 V/nm. Even the larger nanopore shows the reduced DNA speed in the presence of the IL compared to salt solution. However, the difference is less significant.

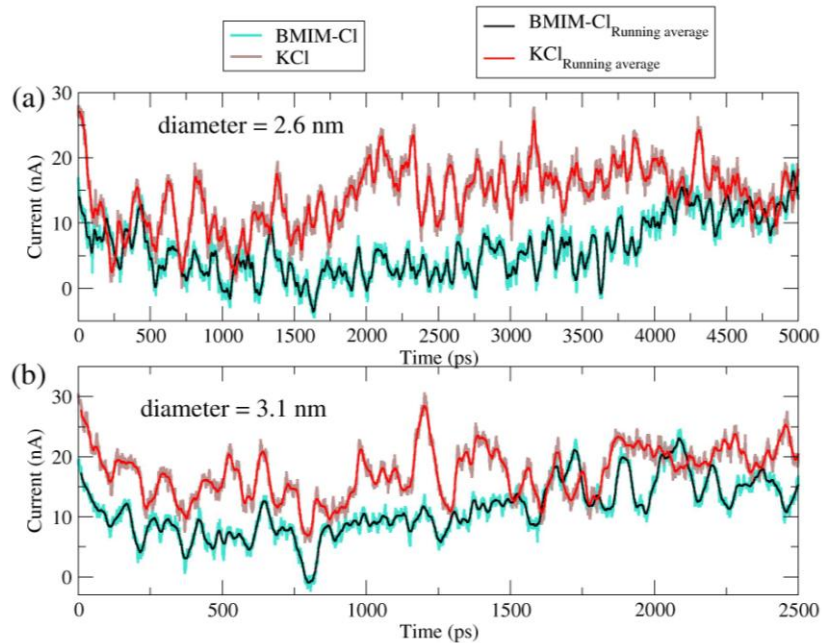


Figure AII.11. The current values for DNA translocation in case of (a) smaller nanopore ($d = 2.6$ nm) and (b) larger nanopore ($d = 3.1$ nm) using lag value of 50 ps in the presence of KCl (brown line) and BMIM-Cl (turquoise line). The running average are plotted for KCl (Red line) and BMIM-Cl simulations (Black line) to visually observe number of blockade events.

The enhanced translocation time allows DNA molecules to interact strongly with the nanopore. This results into a increased number of blockade events and hence provides a better single base resolution. We have demonstrated an increase in blockade events by plotting the current for smaller and bigger nanopores at 0.05 V/nm using 50 ns lag for both the simulations (See Fig. AII.11). This analysis helps to compare the number of blockade events presented by frequent dips in the current values in above cases. We can see in figure AII.11(a) that the number of blockade events are considerably higher in case of BMIM-Cl than KCl, providing a better single base resolution.

AII.3.6. Blockade current during polynucleotide translocation

Since the DNA is already partially inserted in the nanopore at the onset of simulation, the current observed till the complete translocation can be considered as the blockade current.

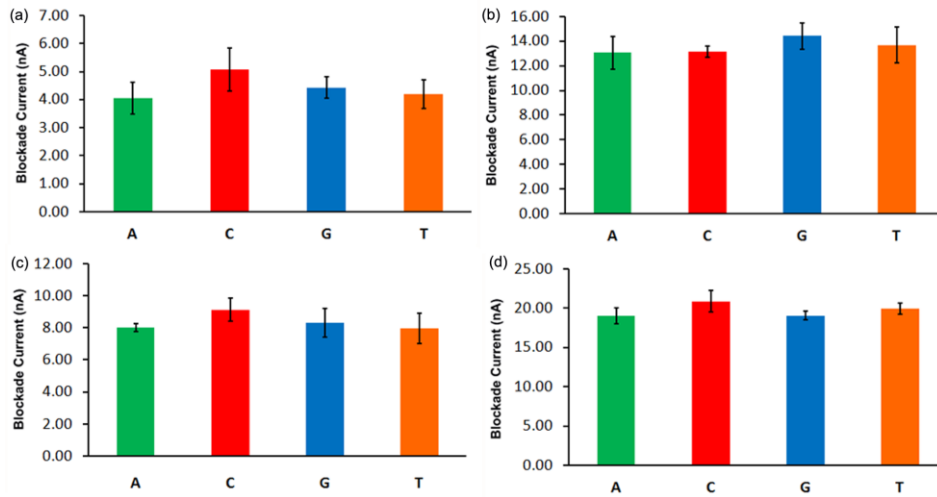


Figure AII.12. Ionic blockade current (calculated using Eq. A.6.4 with 250 ps time lag) for (a) smaller nanopore at smaller VG; (b) smaller nanopore at higher VG; (c) bigger nanopore at smaller VG; (d) bigger nanopore at bigger VG. See Fig. 5 for other details.

Figure AII.12 shows the blockade current averaged over 3 simulations. The ionic blockades currents for the smaller nanopore at smaller VG are less distinguishable for A₂₀, C₂₀, G₂₀ and T₂₀. Also, they are low as blockade current is proportional to the applied voltage. At the higher VG, the difference in blockade current among different nucleotides is insignificant. Hence, at the lower voltage, individual nucleotides might be distinguishable using both translocation time and ionic blockade current. A major challenge to detect significant ionic blockades, however, will remain as the ionic currents in IL's are small compared to KCl or NaCl based electrolytes.

It can be noticed that the blockade current values for bigger nanopore is higher than the smaller nanopore at the same voltage gradient, indicating an enhanced flow of electrolytes in the former. In case of bigger nanopore at smaller VG (see Fig. AII.12(c)), the blockades current are high and these values are comparatively more distinguishable. The adenine and cytosine possess significantly different values, whereas thymine and guanine have almost the same blockade current. The presence of DNA in a nanopore reduces effective cross sectional area of the pore and alters the local ion concentration near the pore. Both these factors cause ionic current blockade.

The ionic current blockade signatures for different nucleotides are sensitive to the orientation of nucleotides passing via nanopore[19]. Hence, ionic current fluctuations are not same for all three simulations (data not shown). We have then carried out the translocation of the hetero-nucleotide sequence, d(ACGT)₅ using smaller nanopore and smaller VG (since higher VG will result into faster

translocation). For these hetero-nucleotide sequences, currents obtained for the IL electrolyte is less than KCL for both nanopores at the smaller voltage gradient (see Fig. AII.13).

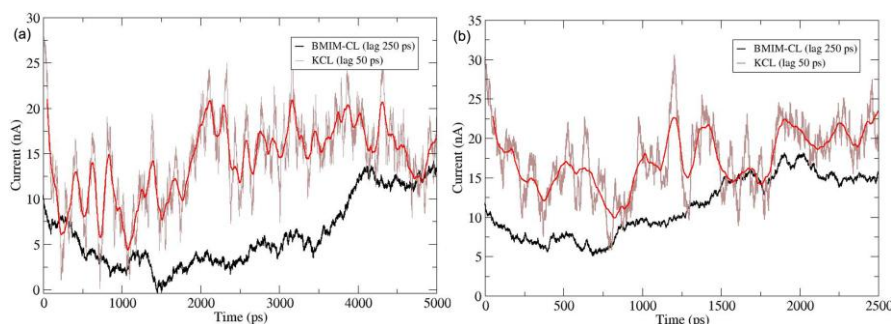


Figure AII.13. The Ionic current for d(ACGT)₅ system in BMIM-CL and KCL at 0.05 V/nm for (a) smaller nanopore, and (b) bigger nanopore. The running average of ionic current for KCl systems is shown by red line.

This behavior can directly be correlated with difference in conductivities of KCl and BMIM-Cl solutions. Experimentally, it is observed that current blockades are inversely proportional to the viscosity of solution. However, the present study is unable to capture the behavior as seen from Fig. AII.13.

The running average for KCL currents (red line in Fig. AII.13) is plotted to visually observe the trend in the currents. We can see that the ionic current blockades are prominent in KCl based solution contrary to the experimental observation. We attribute this discrepancy to the difference in the time lag used to calculate ionic currents. Computationally, the short time lag (50 ps) used for KCl systems to calculate ionic current captures the fluctuations of the current, whereas in case of BMIM-Cl, the bigger time lag of 250 ps averages out these fluctuations indicating a lesser presence of blockade events. The translocation in KCl system completes within 2500 ps and still Fig. AII.13(a) shows a fluctuation in current, possibly due to variation in the number of K⁺ and Cl⁻. For BMIM-Cl based system, however, a distribution of small blockade events is present till 4500 ps.

When simulations were performed for both systems using the bigger nanopore, the reduction in the number translocation events was observed for both systems. Hence, we propose that even though the bigger nanopore can discriminate nucleotides based on the translocation time, it will fail to achieve better resolution in case of heterogeneous sequence due to the reduced translocation events.

AII.3.7. Ion distribution around DNA

The anchored BMIM⁺ ions in DNA effectively increase its width and reduce translocation speed of DNA. Hence, we have studied the radial distribution (RDF) of cations and anions around DNA. Note that the electric field is applied along the length of the DNA. Therefore, RDF of ions around DNA should not be affected by the electric field[55]. BMIM⁺ ions are observed to be strongly accumulated around negatively charged phosphate groups of DNA in previous study[43]. These interactions are mainly electrostatic in nature.

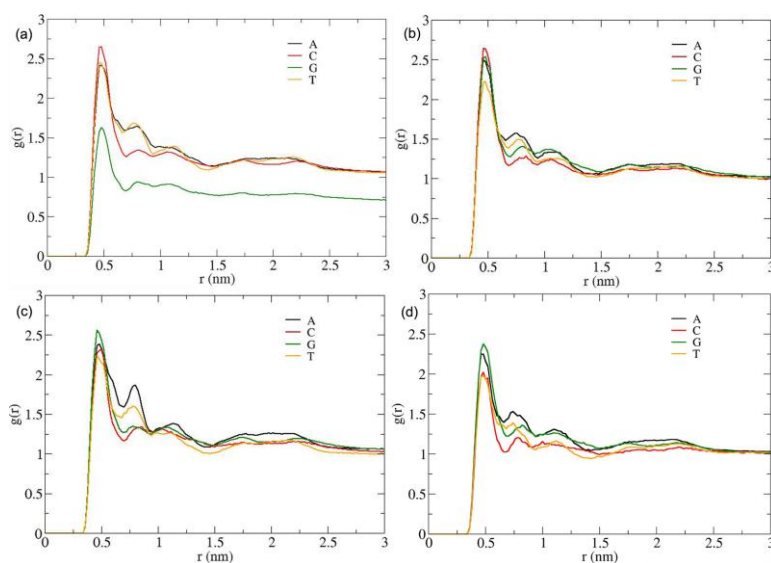


Figure AII.14. Radial distribution function of N1 atom of BMIM⁺ around phosphorous atom of DNA for (a) smaller nanopore at smaller VG; (b) smaller nanopore at higher VG; (c) bigger nanopore at smaller VG; (d) bigger nanopore at higher VG.

The study showed that BMIM-Cl has a binding free energy of ~ 8.96 kcal/mol, indicating a strong binding with DNA[43]. It is suggested that BMIM binds strongly to the minor groove due to hydrogen bond interactions between planar imidazolium ring and exocyclic moieties of nucleotide bases[43]. We have not considered RDF around grooves as their width and depth change along the course of translocation due to stretching of the DNA duplex. Figure AII.14 shows the distribution of N1 atoms of BMIM⁺ cation (Fig. AII.15 for N1 atom) around phosphate backbone of DNA under different conditions.

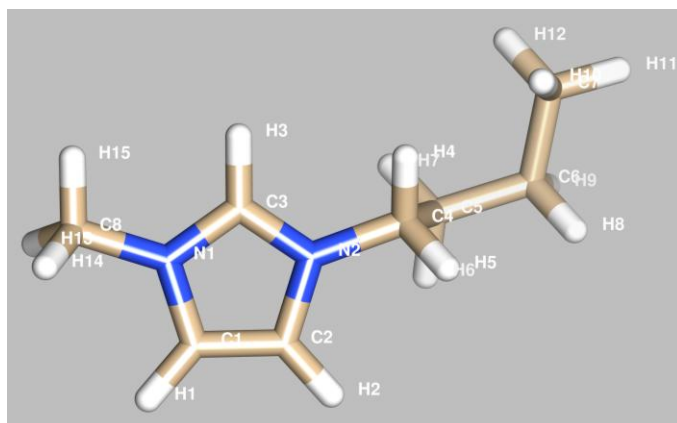


Figure AII.15.Strucutre of BMIM⁺ cation

We can see that in case of the smaller nanopore at the lower VG, the ion distribution is different in each case. The distribution of BMIM⁺ is almost similar in case of dA₂₀ and dT₂₀. Highest number of ions is observed around dC₂₀ and lowest around dG₂₀. We see that even though dC₂₀ is strongly surrounded by bulky BMIM⁺ cations, it possesses lowest translocation time but highest blockade current under these conditions.

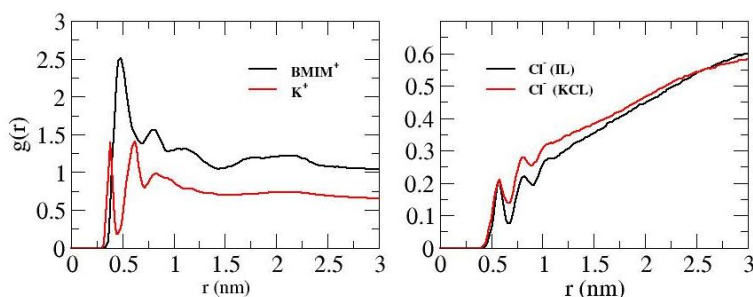


Figure AII.16. (a) Radial distribution function of N1 of BMIM and K⁺ around P of DNA for d(ACGT)₅ and (b) Radial distribution function of Cl⁻ around P of DNA for d(ACGT)₅ in both cases IL and KCL.

This suggests that DNA-BMIM⁺ behave as a combined complex during the translocation and hence it results in large current blockades. However, the number of attached BMIM⁺ ions is not sufficient to reduce translocation speed of the DNA and hence dC₂₀ possess smaller translocation time. For the smaller nanopore, dC₂₀ is always surrounded by large number of BMIM⁺ ions, whereas dG₂₀ is always surrounded by higher number of BMIM⁺ ions for the bigger nanopore. This indicates that pore size affects the distribution around DNA. The dA₂₀ sequence displays a significant second solvation shell under all four conditions suggesting the presence of long-range interactions between BMIM⁺ and dA₂₀. Such strong interactions may

increase the translocation time almost in all cases except for the bigger nanopore at the smaller VG.

We assume that if more BMIM⁺ cations are present close to a DNA sequence then it will be difficult to strip them off as they interact strongly with the DNA via non-covalent interactions. To understand this, we have studied RDF for d(ACGT)₅ simulations carried out for the smaller nanopore at the lower VG. The RDF in Fig. AII.16 indicates that BMIM⁺ interacts strongly with the DNA and has broader first layer of solvation compared to K⁺ ions in KCL solution. The effect of BMIM⁺ cations is long-ranged compared to K⁺ ions. Hence, it supports our hypothesis that removal of strongly bound BMIM⁺ ions from DNA increases translocation time. It is observed for BMIM-PF6 IL that near uncharged graphite surface well ordered structure of IL exists and first adsorption layer of BMIM⁺ is arranged parallel to graphite surface till distance of 3.5 Å[56]. The enhanced RDF of BMIM⁺ ions can be attributed to such structured arrangement also. RDF's of Cl⁻ ions in both the cases display almost same distribution and, hence, shows less effect of anion species on the translocation process.

AII.3.8. Charge distribution along nanopore

We have calculated the charge distribution of cations and anions along the Z-axis of the system. Figure AII.17 shows the difference between the charge distribution of K⁺ and BMIM⁺ cations near the nanopore. It indicates that bulky BMIM⁺ ions accumulate near the nanopore, enhancing the effective positive charge near it. The strong electrostatic interactions between negatively charged DNA backbone and accumulated positive BMIM⁺ ions reduce the translocation speed.

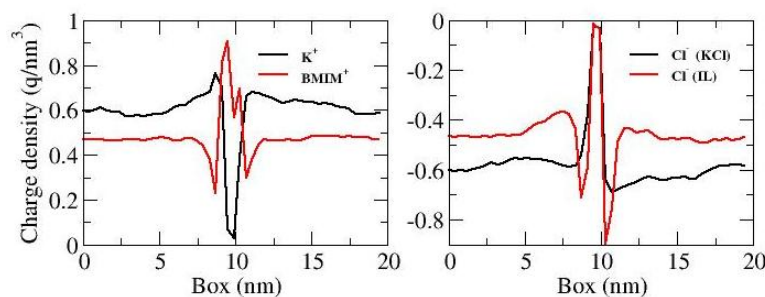


Figure AII.17. Charge distribution along Z-axis perpendicular to graphene surface for heteronucleotide DNA sequence d(ACGT)₅ in both BMIM-Cl and KCl systems.

This phenomenon is similar to the flow reversal at the interface of silicon nano-channel due to immobilization of cations on pore surface.[57] The charge reversal is not possible in the current case due to the short length of nanopore. Luan et al.[55] observed that bulky cations like spermidine(+4) and spermidine(+3) induced reversal in electrophoretic mobility. This reversal was attributed to the strong interaction between these ions and the minor groove of DNA. It was suggested that electric charge inversion or charge screening effect can be observed through RDF distribution of ions[55]. Hence, we propose a presence of strong electrostatic attractions between DNA and well-structured BMIM⁺ ions near graphene surface as main driving force to reduce DNA translocation speed.

AII.4. Conclusion

The current study is focused on different aspects of the translocation process such as open pore current behaviour, dependence of resistance on diameter of the nanopore, individual blockade current as well as translocation time for polynucleotide sequences. The study is extended to understand the effect of BMIM-Cl on the translocation of heteronucleotide sequence like d(ACGT)₅ as compared to KCl solutions. The open pore current study indicates the linear current-voltage relationship for BMIM-Cl based system. The higher bulk conductivity is observed for 1 M BMIM-Cl solution compared to experimental value, which might be attributed to the combined effect of the simulation parameters for BMIM₊ cation in the presence of TIP3P water model implemented in this study. We have also investigated dependence of Signal-to-Noise ratio (SNR) for IL-based system. The study suggests the requirement of high bandwidth amplifiers (of 250 GHz range) to maintain SNR > 5 due to low conductance of BMIM-Cl solution.

A series of simulation with different nanopore diameter at a particular voltage gradient is performed to understand the dependence of resistance of the system on pore diameter. The resistance exhibits nonlinear dependence on diameter of the nanopore. In addition, the results indicates that total resistance depends on the contribution from both channel and access resistances.

To understand the discrimination capability between individual nucleobases, we have the translocation simulations of homo polynucleotides DNA sequences dA₂₀, dC₂₀, dG₂₀, and dT₂₀ in the presence of 1M BMIM-Cl ionic liquid solution. The

simulations were performed at two different pore diameters 2.6 nm and 3.1 nm. For each of them, two different voltage gradients (0.05 V/nm and 0.10 V/nm) were used. The study suggests that the use of smaller nanopore and low VG can help to distinguish different nucleotides on the basis of translocation times in the presence of BMIM⁺ ionic liquid. The blockade currents were less distinguishable for different bases under all conditions used. The electrophoretic mobility for different homo-nucleotide sequences was found to be different for the bigger nanopore at low VG (See Fig.AII.8(b)). This suggests bigger nanopore can also be used to in case of electrolytes of bigger cations like BMIM⁺.

To understand the difference between translocation mechanism of common KCl solutions and BMIM-Cl solutions, we have compared the translocation of hetero-nucleotide sequence such as d(ACGT)₅ in both 1M KCl and 1M BMIM-Cl solutions. The results suggest that the presence of enhanced positive charge near the nanopore (see Fig. AII.17) reduces DNA translocation speed by strong interaction with DNA. The radial distribution function shows preferential strong interaction between DNA and BMIM⁺ ions compared to K⁺ ions as the underlying mechanism for the reduced translocation time in IL.

This study also suggests that viscous BMIM-Cl or any other suitable ionic liquid can be utilized to reduce the DNA translocation speed. However, the better resolution can be achieved using solid-state nanopores like MOS₂ or HfO₂ having inherently high SNR. In addition, DNA soaked in ionic liquid in the cis chamber and KCl solutions in trans chamber will provide better single base resolution in DNA sequencing, as shown by recent study[21].

References:

- (1) Porcu, E.; Sanna, S.; Fuchsberger, C.; Fritsche, L. G. In *Current Protocols in Human Genetics*; John Wiley & Sons, Inc.: 2013; Vol. 78, p 1.25:1.
- (2) Feng, Y.; Zhang, Y.; Ying, C.; Wang, D.; Du, C. *Genomics Proteomics Bioinformatics* **2015**, *13*, 4.
- (3) Ghosal, S. *Phys. Rev. Lett.* **2007**, *98*, 238104.
- (4) Kejian, D.; Sun, W. M.; Zhang, H. Y.; Peng, X. L.; Hu, H. G. *Appl. Phys. Lett.* **2009**, *94*, 014101.
- (5) Fologea, D.; Uplinger, J.; Thomas, B.; McNabb, D. S.; Li, J. L. *Nano Lett.* **2005**, *5*, 1734.
- (6) Wanunu, M.; Sutin, J.; McNally, B.; Chow, A.; Meller, A. *Biophys. J.* **2008**, *95*, 4716.
- (7) Mirsaidov, U.; Comer, J.; Dimitrov, V.; Aksimentiev, A.; Timp, G. *Nanotechnology* **2010**, *21*, 395501.
- (8) Akahori, R.; Haga, T.; Hatano, T.; Yanagi, I.; Ohura, T.; Hamamura, H.; Iwasaki, T.; Yokoi, T.; Anazawa, T. *Nanotechnology* **2014**, *25*, 277501.
- (9) Wanunu, M.; Morrison, W.; Rabin, Y.; Grosberg, A. Y.; Meller, A. *Nat. Nanotechnol.* **2010**, *5*, 160.
- (10) He, Y. H.; Tsutsui, M.; Scheicher, R. H.; Fan, C.; Taniguchi, M.; Kawai, T. *Biophys. J.* **2013**, *105*, 776.
- (11) He, Y. H.; Tsutsui, M.; Fan, C.; Taniguchi, M.; Kawai, T. *ACS Nano* **2011**, *5*, 5509.
- (12) Luan, B. Q.; Aksimentiev, A. *Soft Matter* **2010**, *6*, 243.
- (13) Verschueren, D. V.; Jonsson, M. P.; Dekker, C. *Nanotechnology* **2015**, *26*, 234004.
- (14) Lu, B.; Hoogerheide, D. P.; Zhao, Q.; Zhang, H.; Tang, Z.; Yu, D.; Golovchenko, J. A. *Nano Lett.* **2013**, *13*, 3048.
- (15) Buyukdagli, S.; Blossey, R.; Ala-Nissila, T. *Phys. Rev. Lett.* **2015**, *114*, 088303.
- (16) Kowalczyk, S. W.; Wells, D. B.; Aksimentiev, A.; Dekker, C. *Nano Lett.* **2012**, *12*, 1038.
- (17) de Zoysa, R. S. S.; Jayawardhana, D. A.; Zhao, Q. T.; Wang, D. Q.; Armstrong, D. W.; Guan, X. Y. *J. Phys. Chem. B* **2009**, *113*, 13332.
- (18) Sathe, C.; Zou, X. Q.; Leburton, J. P.; Schulten, K. *ACS Nano* **2011**, *5*, 8842.
- (19) Wells, D. B.; Belkin, M.; Comer, J.; Aksimentiev, A. *Nano Lett.* **2012**, *12*, 4117.
- (20) Farimani, A. B.; Min, K.; Aluru, N. R. *ACS Nano* **2014**, *8*, 7914.
- (21) Feng, J.; Liu, K.; Bulushev, R. D.; Khlybov, S.; Dumcenco, D.; Kis, A.; Radenovic, A. *Nat Nano* **2015**, *10*, 1070.
- (22) Larkin, J.; Henley, R.; Bell, D. C.; Cohen-Karni, T.; Rosenstein, J. K.; Wanunu, M. *ACS Nano* **2013**, *7*, 10121.
- (23) Liu, K.; Feng, J. D.; Kis, A.; Radenovic, A. *ACS Nano* **2014**, *8*, 2504.
- (24) Modi, N.; Singh, P. R.; Mahendran, K. R.; Schulz, R.; Winterhalter, M.; Kleinekathofer, U. *J. Phys. Chem. Lett.* **2011**, *2*, 2331.
- (25) Jiang, X. K.; Huang, J. S.; Sumpter, B. G.; Qiao, R. *J. Phys. Chem. Lett.* **2013**, *4*, 3120.
- (26) Macke, T. J.; Case, D. A. *Molecular Modeling of Nucleic Acids* **1998**, *682*, 379.

- (27) D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman AmberTools15, webpage: <http://ambermd.org/>
- (28) Humphrey, W.; Dalke A Fau - Schulten, K.; Schulten, K. *J. Molec. Graphics* **1996**, *14*, 33.
- (29) Sambasivarao, S. V.; Acevedo, O. *J. Chem. Theory Comput.* **2009**, *5*, 1038.
- (30) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (32) Bester-Rogac, M.; Stoppa, A.; Hunger, J.; Hefter, G.; Buchner, R. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17588.
- (33) Tom Darden, D. Y., Lee Pedersen *J. Chem. Phys.* **1993**, *98*, 10089.
- (34) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (35) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
- (36) Lv, W.; Liu S Fau - Li, X.; Li X Fau - Wu, R. a.; Wu, R. *Electrophoresis* **2014**, *35*, 1144.
- (37) Aksimentiev, A.; Schulten, K. *Biophys. J.* **2005**, *88*, 3745.
- (38) Rosenstein, J. K.; Wanunu, M.; Merchant, C. A.; Drndic, M.; Shepard, K. L. *Nat. Methods* **2012**, *9*, 487.
- (39) Smeets, R. M. M.; Keyser, U. F.; Dekker, N. H.; Dekker, C. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 417.
- (40) Tasserit, C.; Koutsioubas, a.; Lairez, D.; Zalczer, G.; Clochard, M. C. *Phys. Rev. Lett.* **2010**, *105*, 1.
- (41) Biroli, G. *Nat. Phys.* **2007**, *3*, 222.
- (42) Venkatesan, B. M.; Estrada, D.; Banerjee, S.; Jin, X.; Dorgan, V. E.; Bae, M.-H.; Aluru, N. R.; Pop, E.; Bashir, R. *ACS Nano* **2012**, *6*, 441.
- (43) Chandran, A.; Ghoshdastidar, D.; Senapati, S. *J. Am. Chem. Soc.* **2012**, *134*, 20330.
- (44) Liang, L.; Cui, P.; Wang, Q.; Wu, T.; Agren, H.; Tu, Y. *RSC Advances* **2013**, *3*, 2445.
- (45) Davenport, M.; Rodriguez, A.; Shea, K. J.; Siwy, Z. S. *Nano Lett.* **2009**, *9*, 2125.
- (46) Han, J.; Fu, J.; Schoch, R. B. *Lab Chip* **2008**, *8*, 23.
- (47) Deen, W. M. *AIChE J.* **1987**, *33*, 1409.
- (48) Merchant, C. A.; Healy, K.; Wanunu, M.; Ray, V.; Peterman, N.; Bartel, J.; Fischbein, M. D.; Venta, K.; Luo, Z.; Johnson, A. T. C.; Drndić, M. *Nano Lett.* **2010**, *10*, 2915.
- (49) Stefan, W. K.; Alexander, Y. G.; Yitzhak, R.; Cees, D. *Nanotechnology* **2011**, *22*, 315101.
- (50) Shi, C. C.; Kong, Z.; Sun, T. Y.; Liang, L. J.; Shen, J. W.; Zhao, Z. Y.; Wang, Q.; Kang, Z. Z.; Agren, H.; Tu, Y. Q. *RSC Advances* **2015**, *5*, 9389.

- (51) Kesselheim, S.; Müller, W.; Holm, C. *Phys. Rev. Lett.* **2014**, *112*, 018101.
- (52) Li, J.; Zhang, Y.; Yang, J.; Bi, K.; Ni, Z.; Li, D.; Chen, Y. *Phys. Rev. E* **2013**, *87*, 062707.
- (53) Aksimentiev, A.; Heng, J. B.; Timp, G.; Schulten, K. *Biophys. J.* **2004**, *87*, 2086.
- (54) Luan, B.; Aksimentiev, A. *Phys. Rev. E* **2008**, *78*, 021912.
- (55) Luan, B.; Aksimentiev, A. *Soft Matter* **2010**, *6*, 243.
- (56) Kislenko, S. A.; Samoylov, I. S.; Amirov, R. H. *Phys. Chem. Chem. Phys.* **2009**, *11*, 5584.
- (57) Qiao, R.; Aluru, N. R. *Phys. Rev. Lett.* **2004**, *92*, 198301.

ANNEXURES

Annexure 3

- A.3.1** Figure showing free energy convergence for all 10 dinucleotide steps,
- A.3.2** Figure showing the variation in different energy components along Z'_p ,
- A.3.3** Figure showing variation of χ for 10 dinucleotide steps angle along Z'_p ,
- A.3.4** Figure showing variation of δ for 10 dinucleotide steps angle along Z'_p ,
- A.3.5** Figure showing variation of Phase angle for 10 dinucleotide steps angle along Z'_p ,
- A.3.6** Figure showing variation of Roll values for 10 dinucleotide steps angle along Z'_p ,
- A.3.7** Figure showing variation of Slide values for 10 dinucleotide steps angle along Z'_p ,
- A.3.8** Figure showing variation of Twist values for 10 dinucleotide steps angle along Z'_p ,
- A.3.9** Figure showing variation of base pair step parameter Slide against Z'_p for two base steps above and below the middle one containing ten different dinucleotide steps,
- A.3.10** Figure showing Variation of Z_p along reaction coordinate Z'_p for two flanking base steps including the middle one being deformed using umbrella sampling.
- A.3.11** Figure showing PMF along average δ angles for middle dinucleotide step and variation of Z_p and percentage of A-form along average δ angle.

A.3.1. Free Energy Convergence

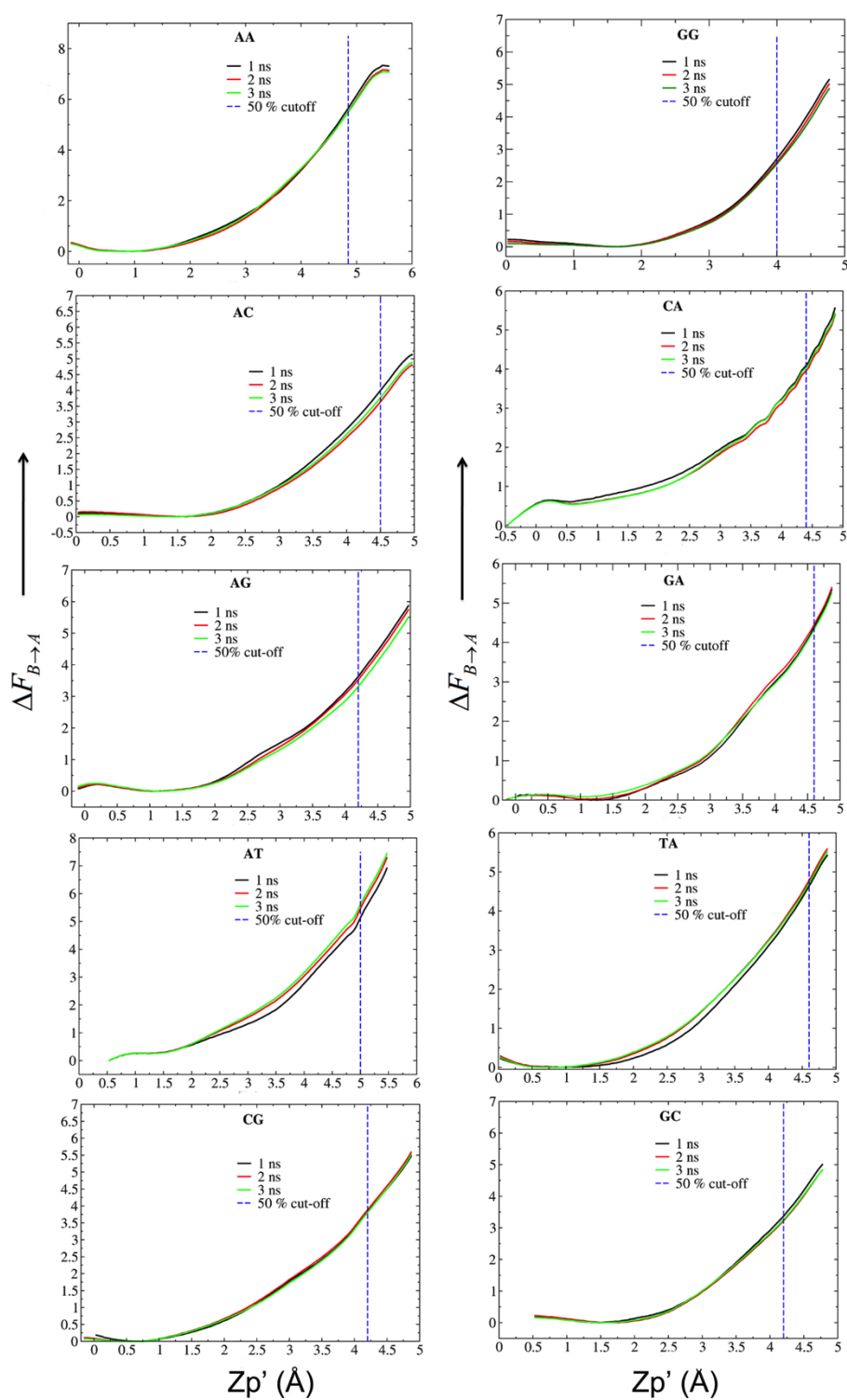


Figure A-3.1 Convergence of the potential of mean force for all ten dinucleotide steps.

A.3.2. Different Energy Contributions

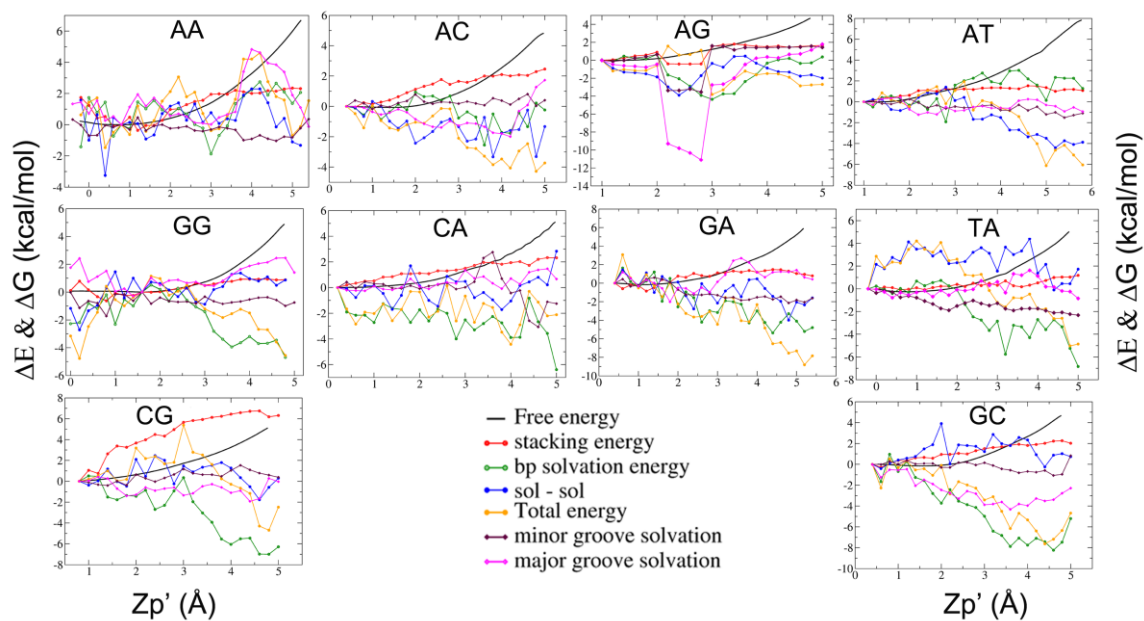


Figure A-3.2. Variation of important energy components like stacking energy (red line), middle base pair solvation energy (bp solvation energy, green line), Solvent self interaction energy (sol-sol, blue line), Total energy (orange line), minor groove and major groove solvation energies (maroon and magenta lines).

A.3.3. Variation of χ dihedral for all 10 unique dinucleotide steps along Zp' .

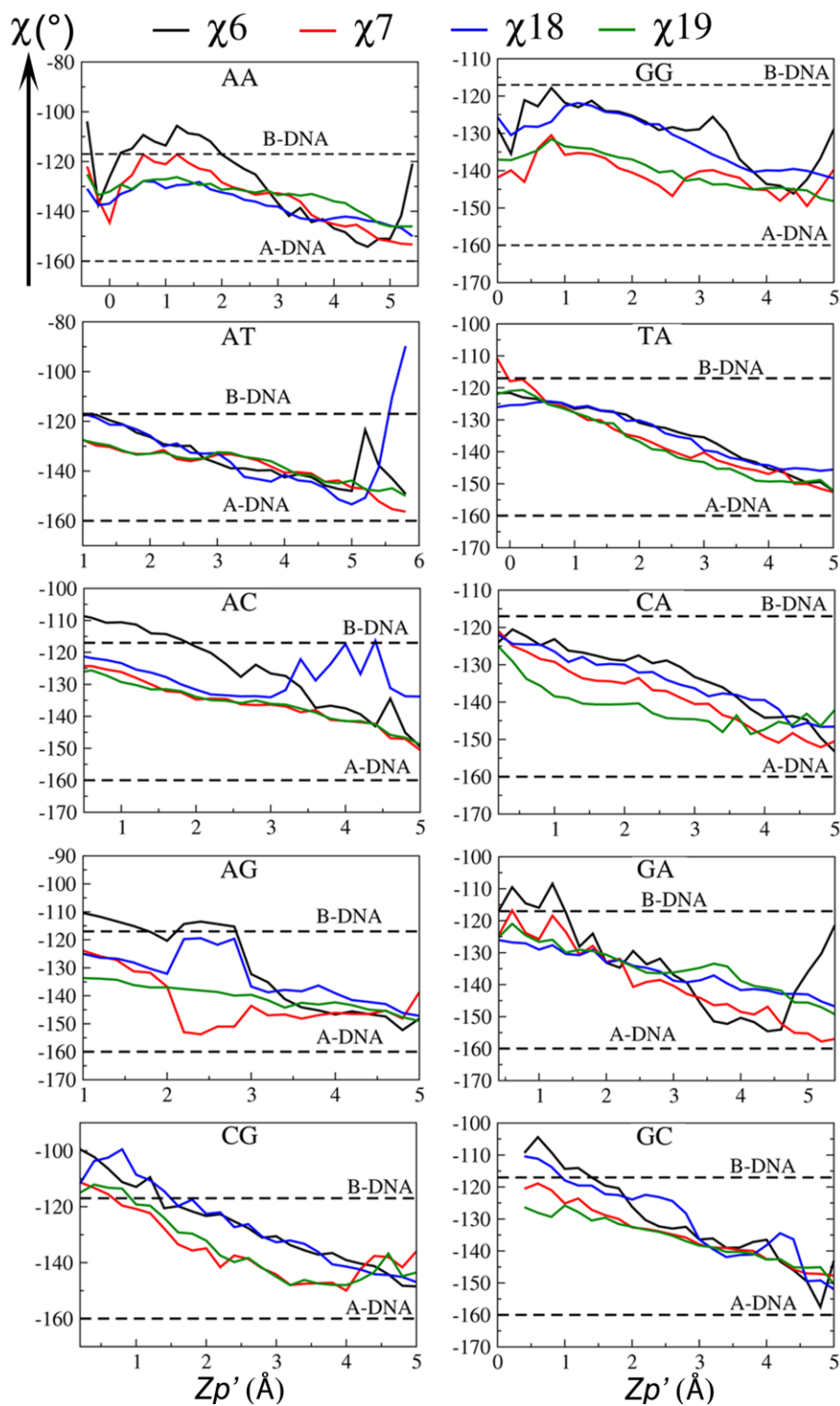


Figure A-3.3. Variation of χ ((O4'-C1'-N1-C2 for pyrimidine) and (O4'-C1'-N9-C4 for Purine)) angle along Zp' for residues 6, 7, 18 and 19.

A.3.4. Variation of δ dihedral for all 10 unique dinucleotide steps along Zp' .

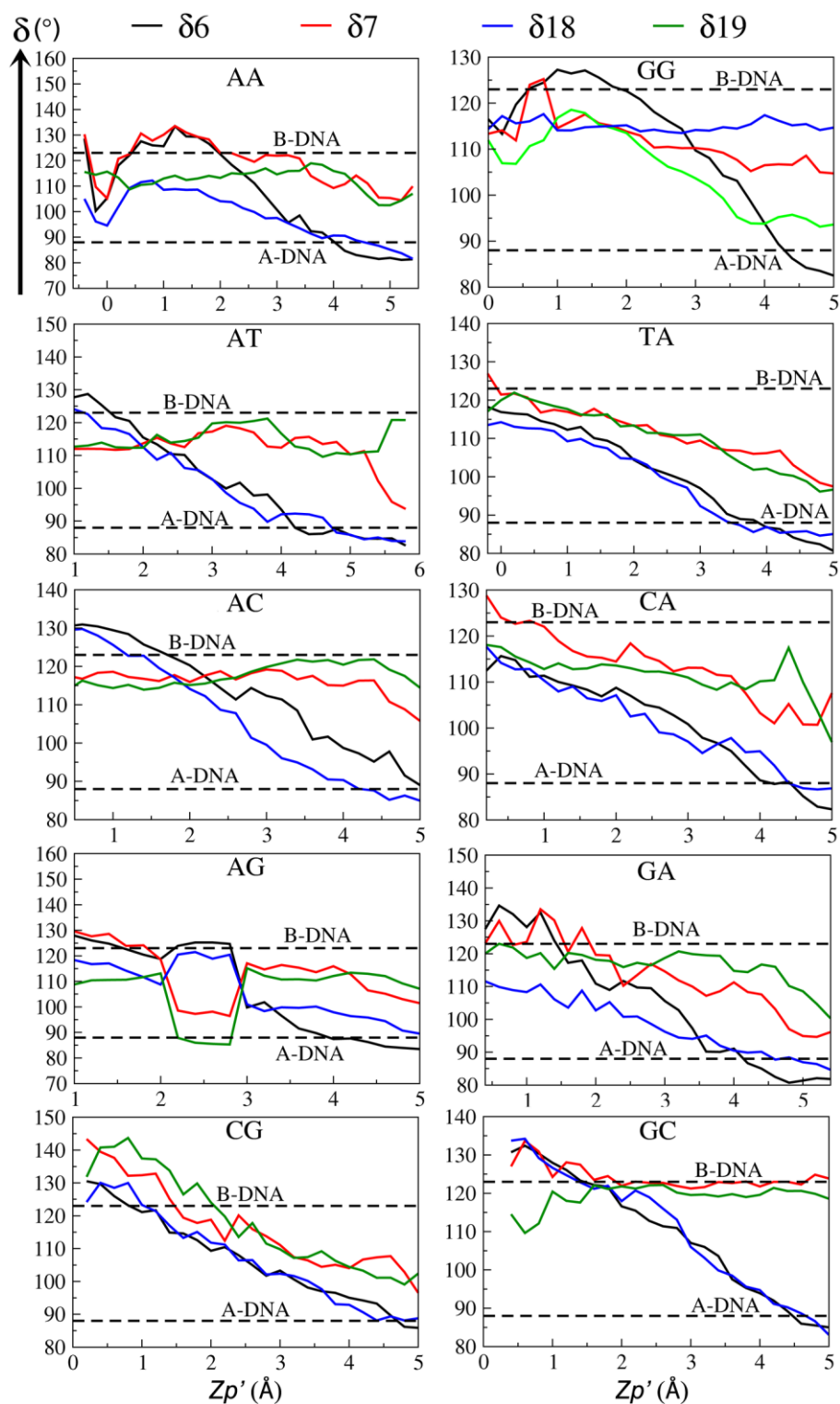


Figure A-3.4. Variation of δ ($C5'-C4'-C3'-O3'$) angle along Zp' for residues 6, 7, 18 and 19.

A.3.5. Variation of Pseudorotation angle (P) of sugars for all 10 unique dinucleotide steps along Zp'.

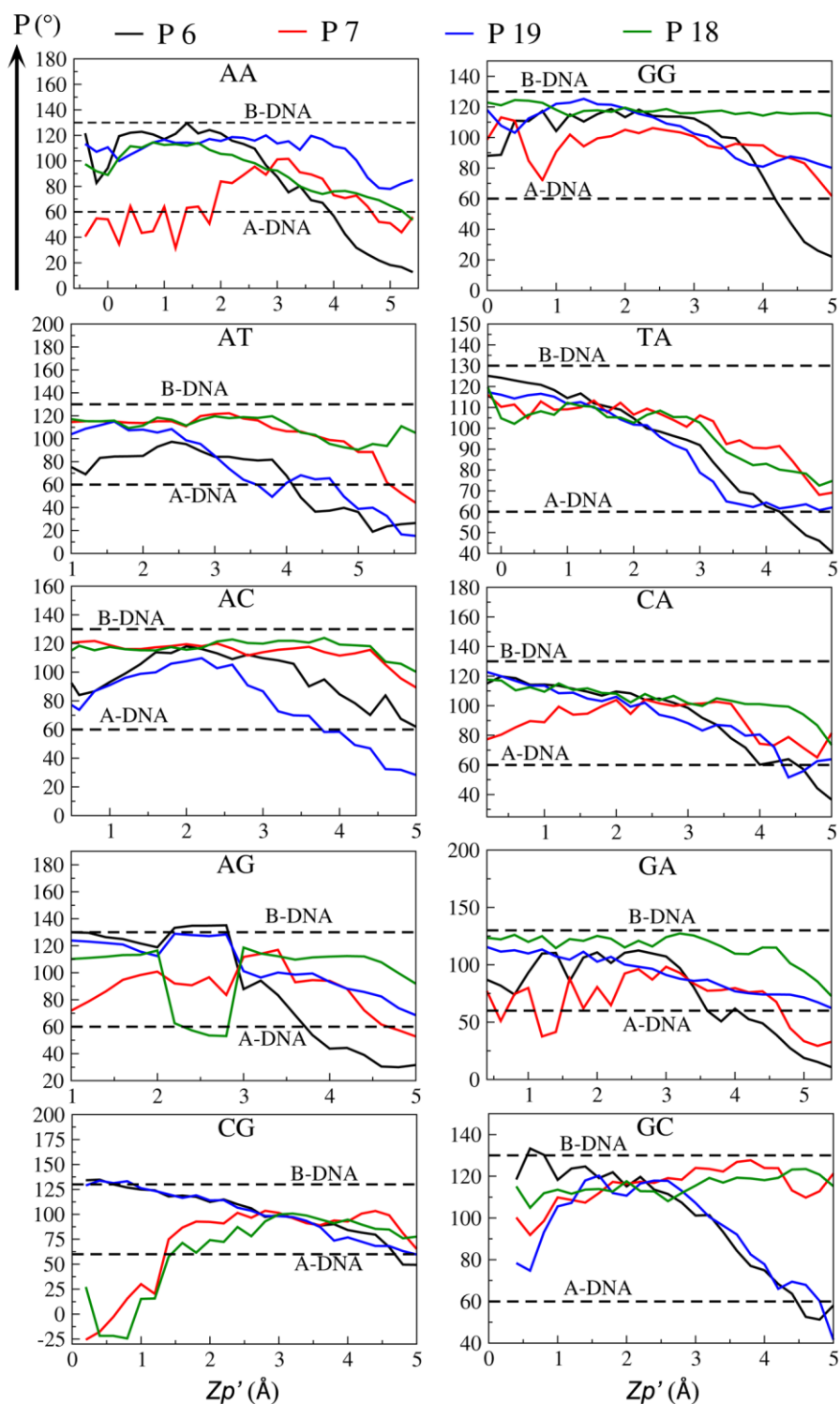


Figure A-3.5. Variation of Pseudorotation angle (P) along Zp' for residues 6, 7, 18 and 19.

A.3.6. Variation of Roll for all 10 unique dinucleotide steps along Zp' .

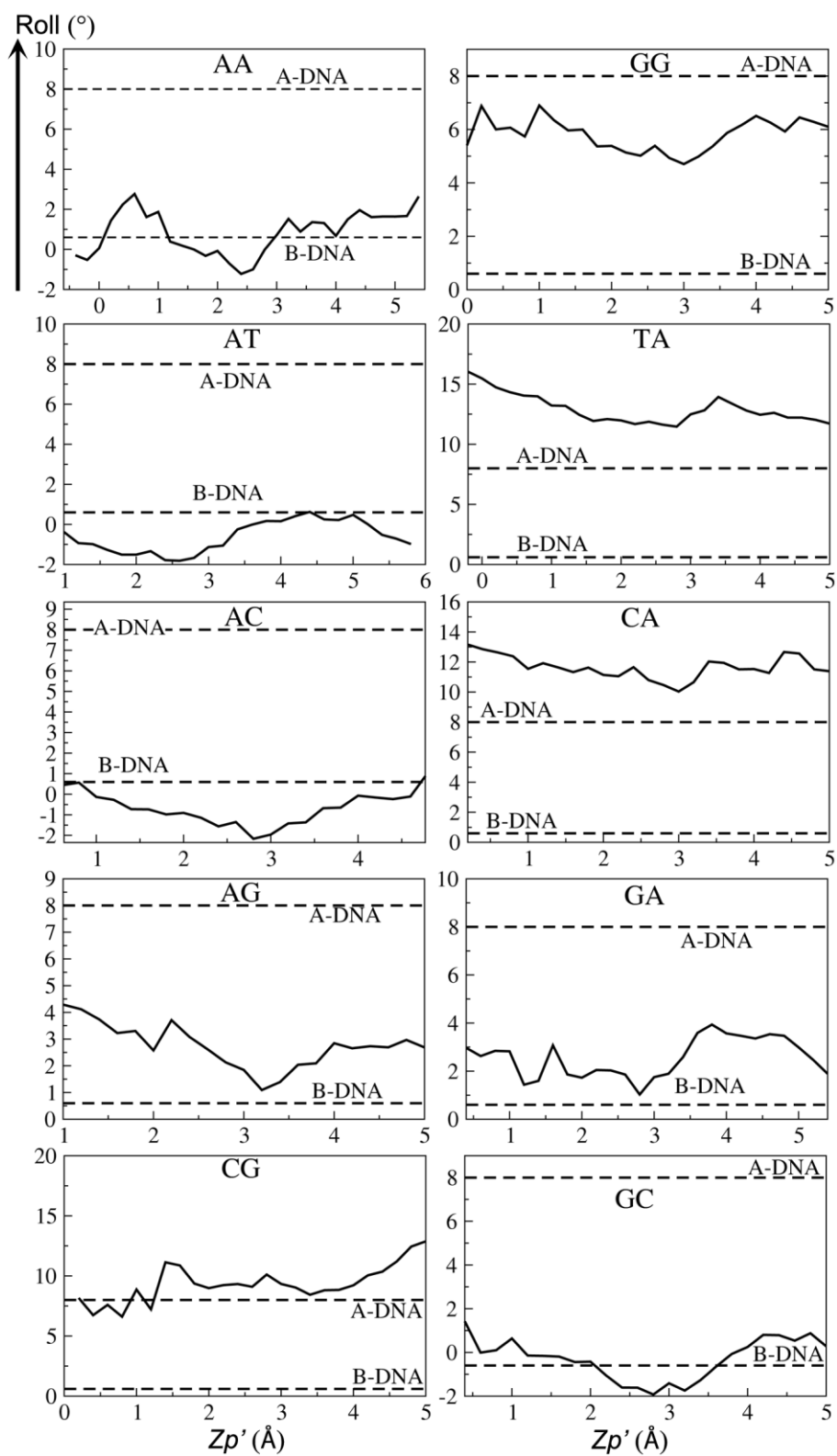


Figure A-3.6. Variation of base pair step parameter roll for all 10 dimers. CG, CA, TA and GG have higher Roll values.

A.3.7. Variation of Slide for all 10 unique dinucleotide steps along Zp' .

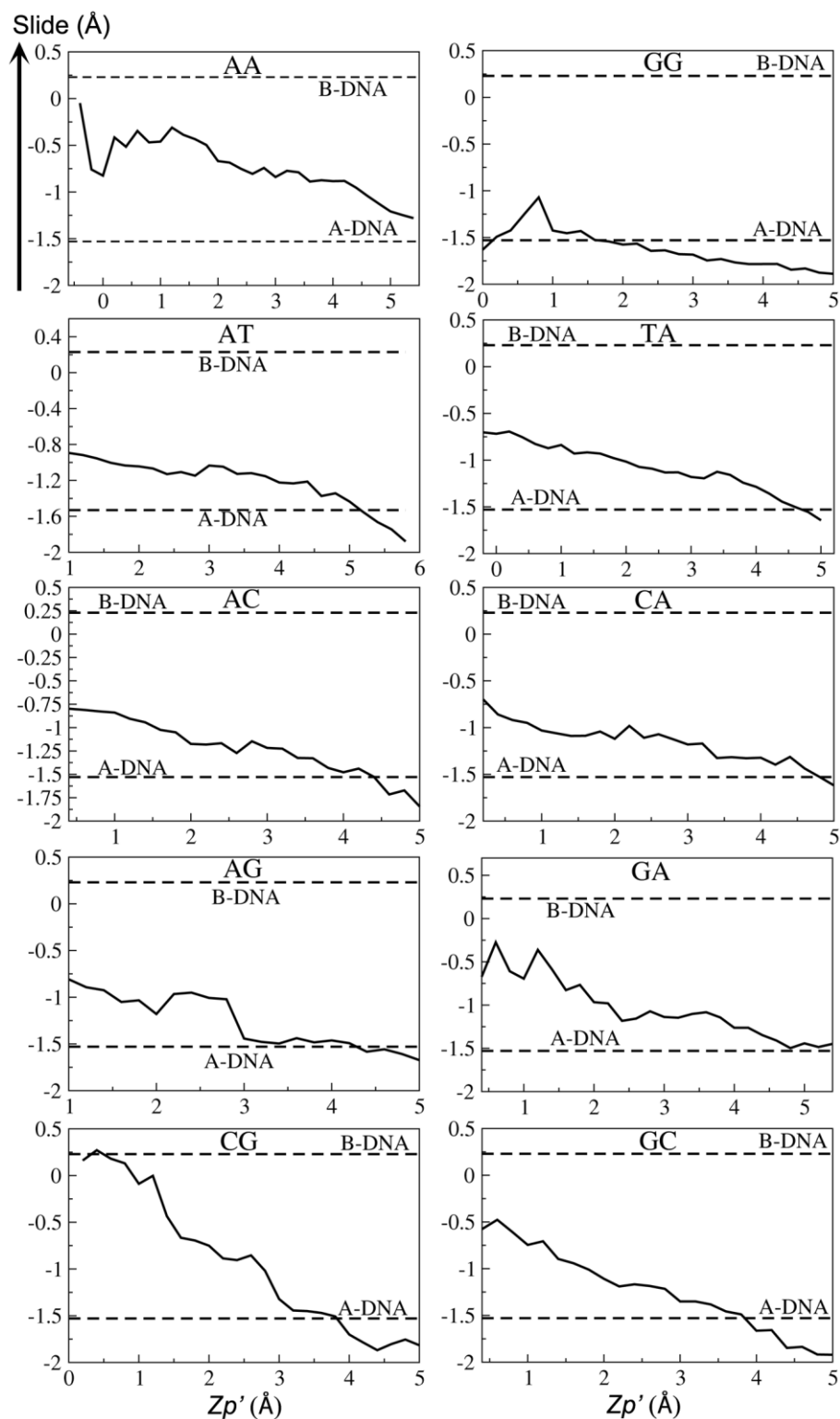


Figure A-3.7. Variation of base pair step parameter slide for all ten dinucleotide steps.

A.3.8. Variation of Twist for all 10 unique dinucleotide steps along Zp' .

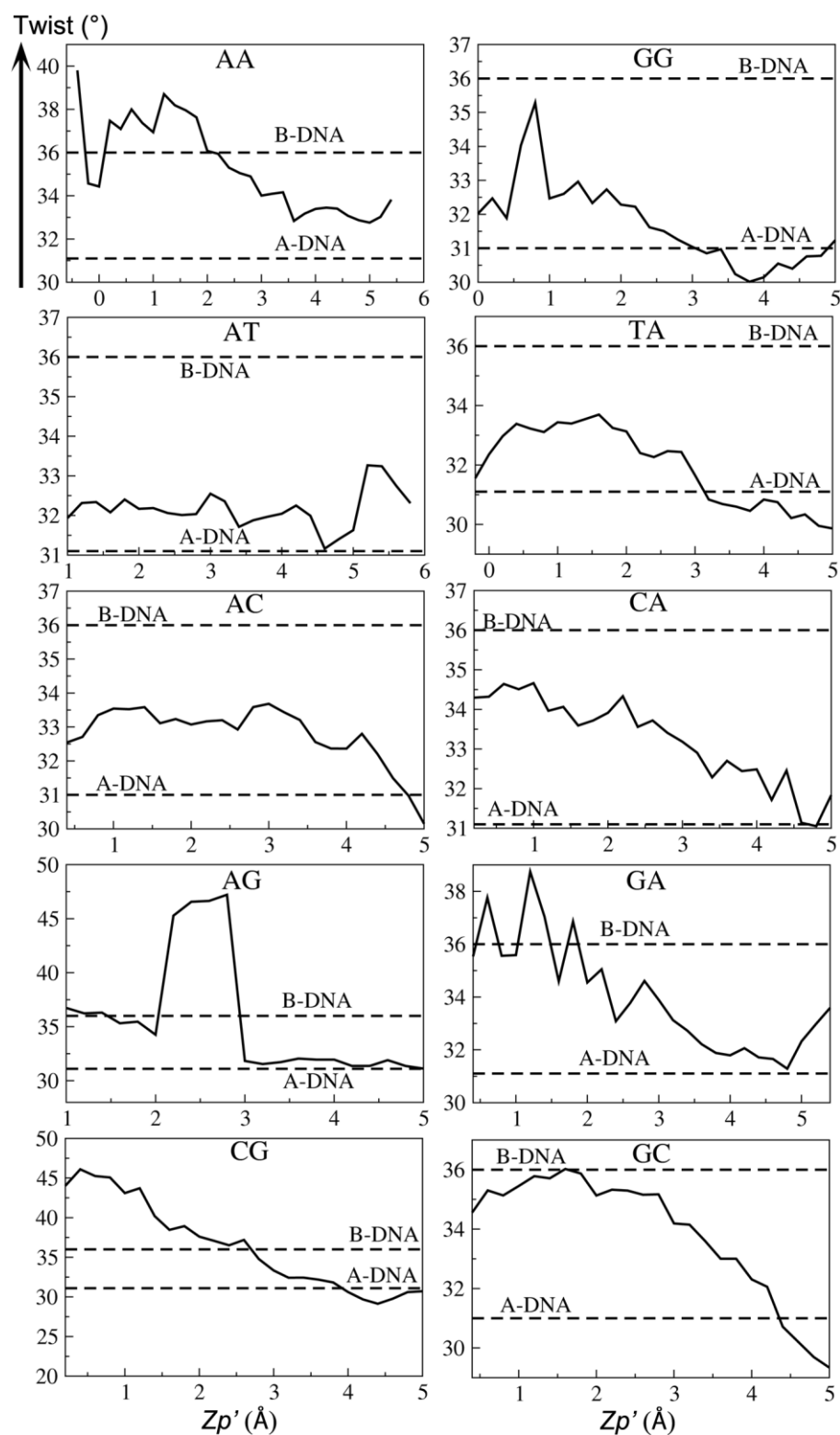


Figure A-3.8. Variation of base pair step parameter twist for all ten dimers.

A.3.9. Variation of Slide for two base steps above and below middle step of 10 dinucleotide step against Zp' .

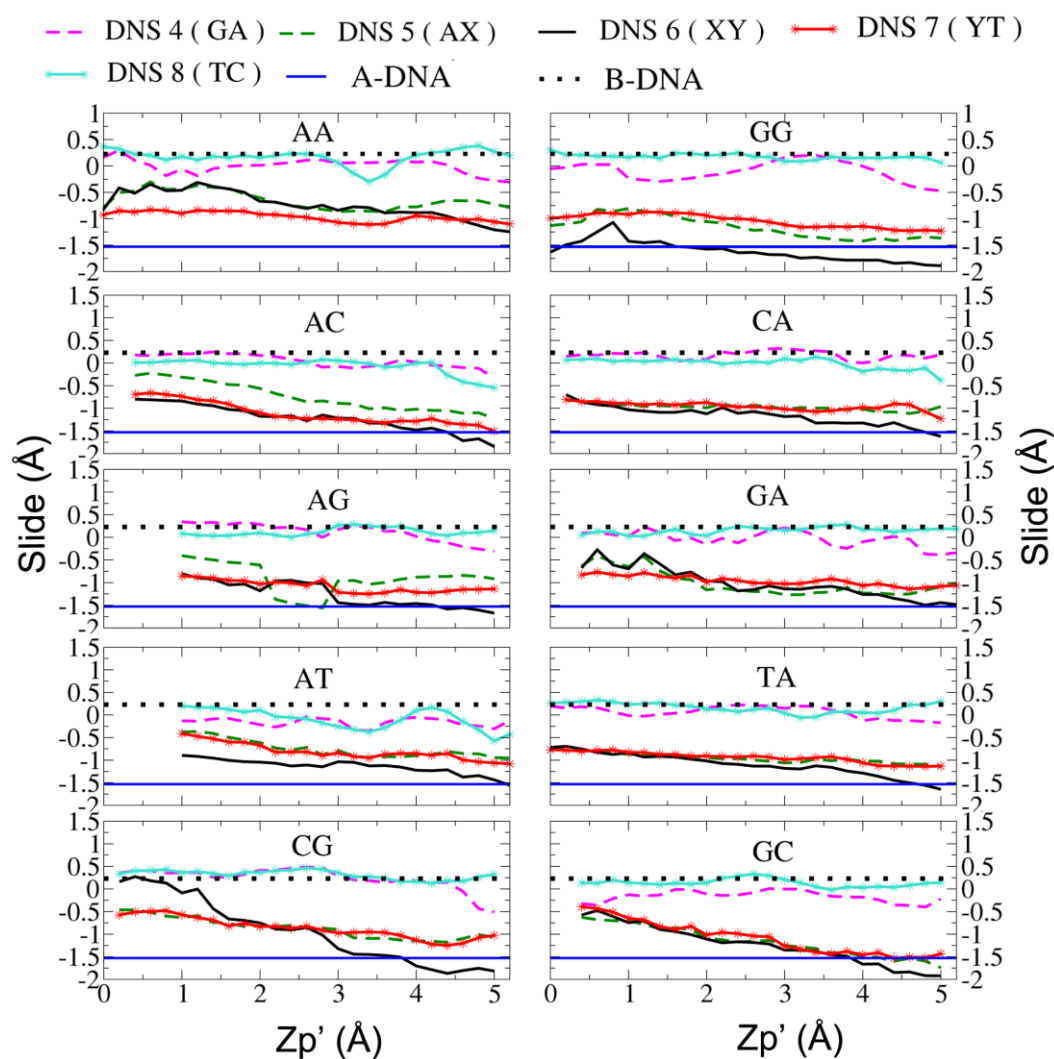


Figure A-3.9. Variation of base pair step parameter Slide against Zp' for two base steps above and below the middle one containing ten different dinucleotide steps.

A.3.10. Variation of Z_p for two base steps above and below middle step of 10 dinucleotide step against $Z_{p'}$.

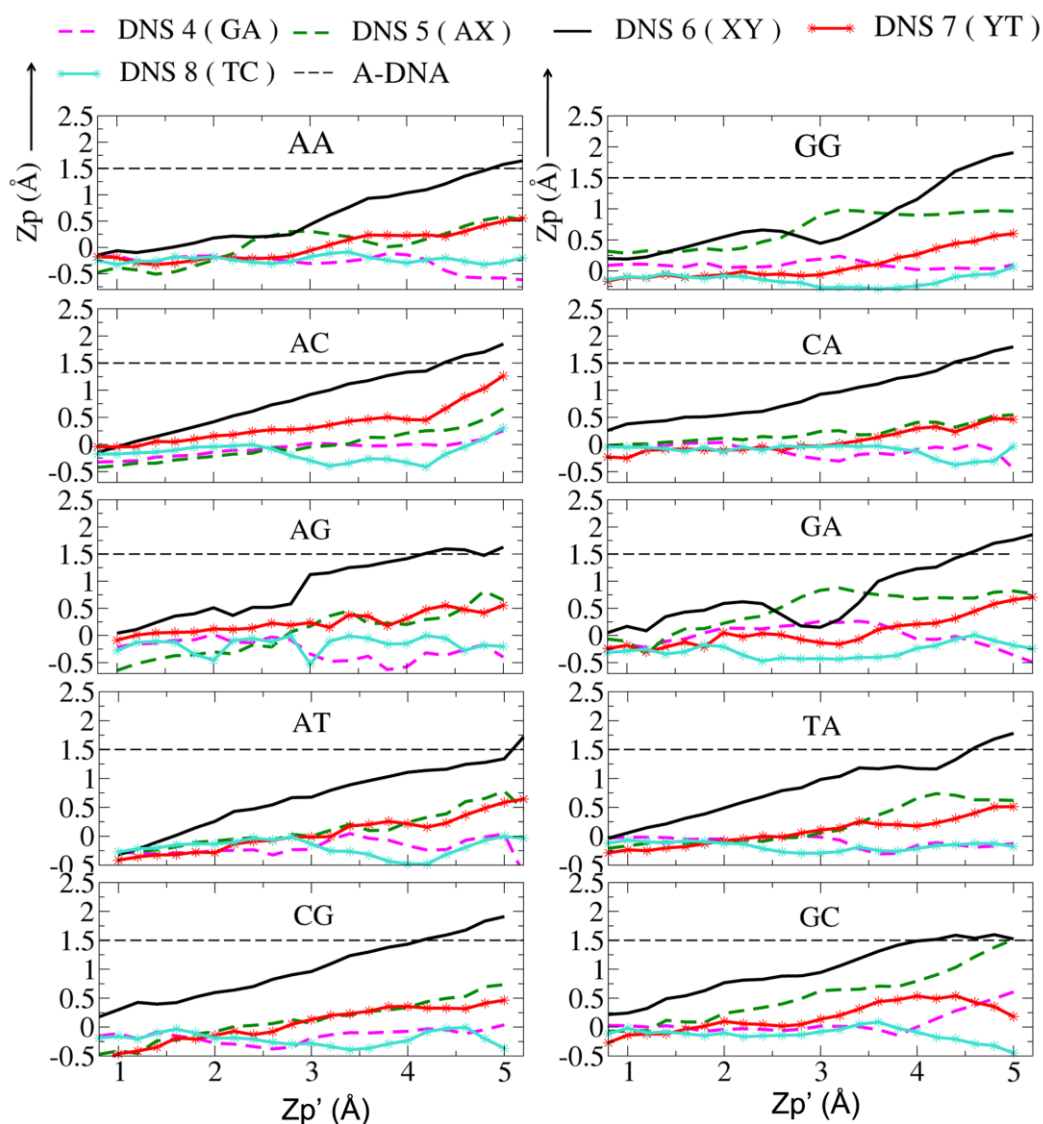


Figure A-3.10. Variation of Z_p along reaction coordinate $Z_{p'}$ for two flanking base steps including the middle one being deformed using umbrella sampling.

A.3.11. Potential of mean force for B- to A- DNA conversion against average δ angles for middle base pair

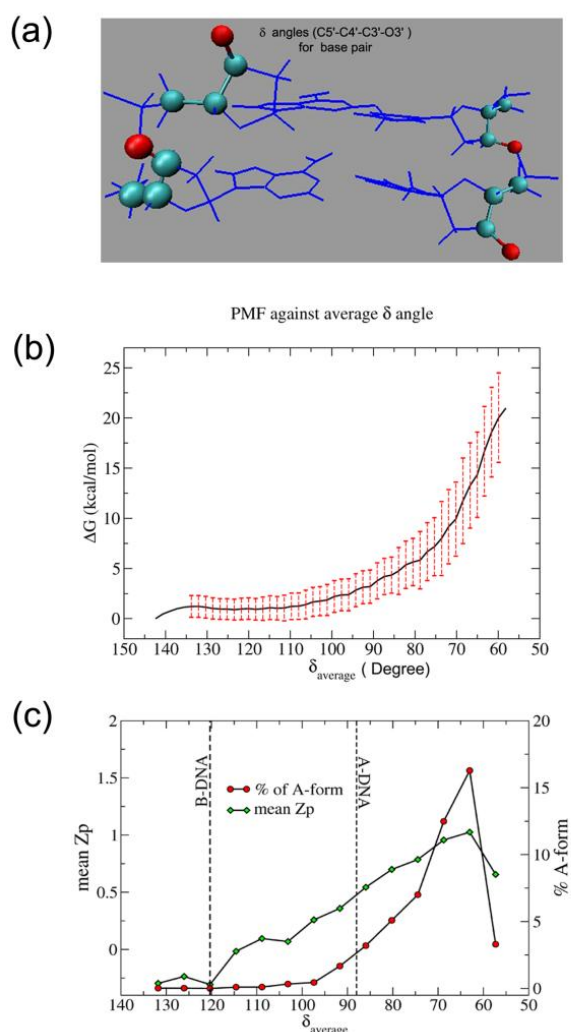


Figure A-3.11. (a) Schematic presentation of four δ dihedral angles, average of which is used as a collective variable. (b) Potential of mean force against average δ angles for middle dinucleotide step. (c) Variation of average Zp and percentage of A-form along average δ angles[1].

Umbrella Sampling simulations were performed using equilibrated structure of $d(\text{CGCGAAATCGCG})_2$, starting from average δ angle (Fig. A-3.11(a)) 120° (B-DNA region) to 88° (A-DNA region) with a step of radians. Each window was equilibrated for 4 ns. Fig. A-3.11(c) shows even though average δ angle attains A-DNA value, average Zp still remains in the intermediate region. This indirectly suggests that backbone change via δ angle does not help in A-form conversion for this dimeric step. Moreover, at high free energy values (20 kcal/mol), A-DNA

population is small (15%) and average Z_p touches only 1.0. Hence, average δ angle is not a suitable reaction coordinate to study B- to A- form transformation in water.

Annexure 4

- A.4.1** Figure showing free energy convergence for homo dinucleotide steps (AA, GG, CC, TT) and trinucleotide steps (AAA, GGG, CCC, TTT).
- A.4.2** Figure showing free energy convergence for hetero dinucleotide and hetero trinucleotide steps.
- A.4.3** Figure showing the variation in free energy, salvation energy, and stacking energy along Z'_p and $\overline{Z'_p}$.

A.4.1. Free Energy Convergence

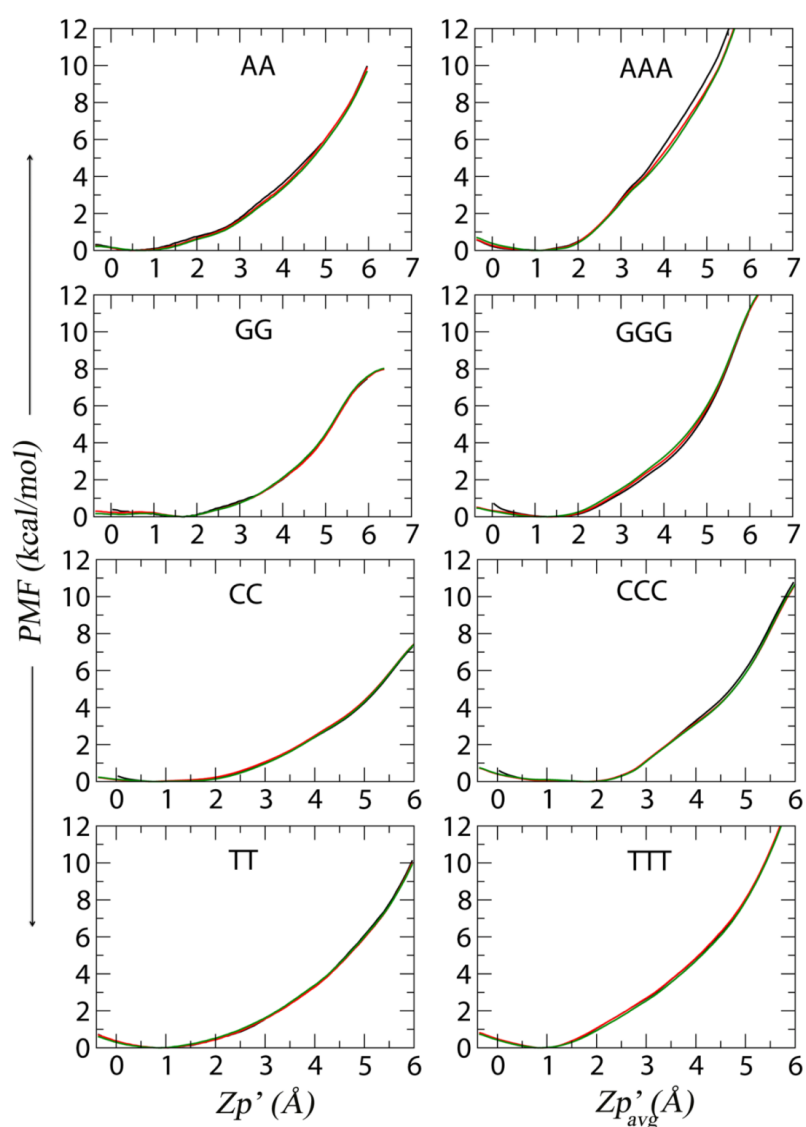


Figure A-4.1. Convergence plot for homo di- and trinucleotide (AA,GG,CC, TT, AAA, GGG, CCC, TTT) steps.

A.4.2. Free Energy Convergence for hetero dinucleotide steps and hetero trinucleotide steps

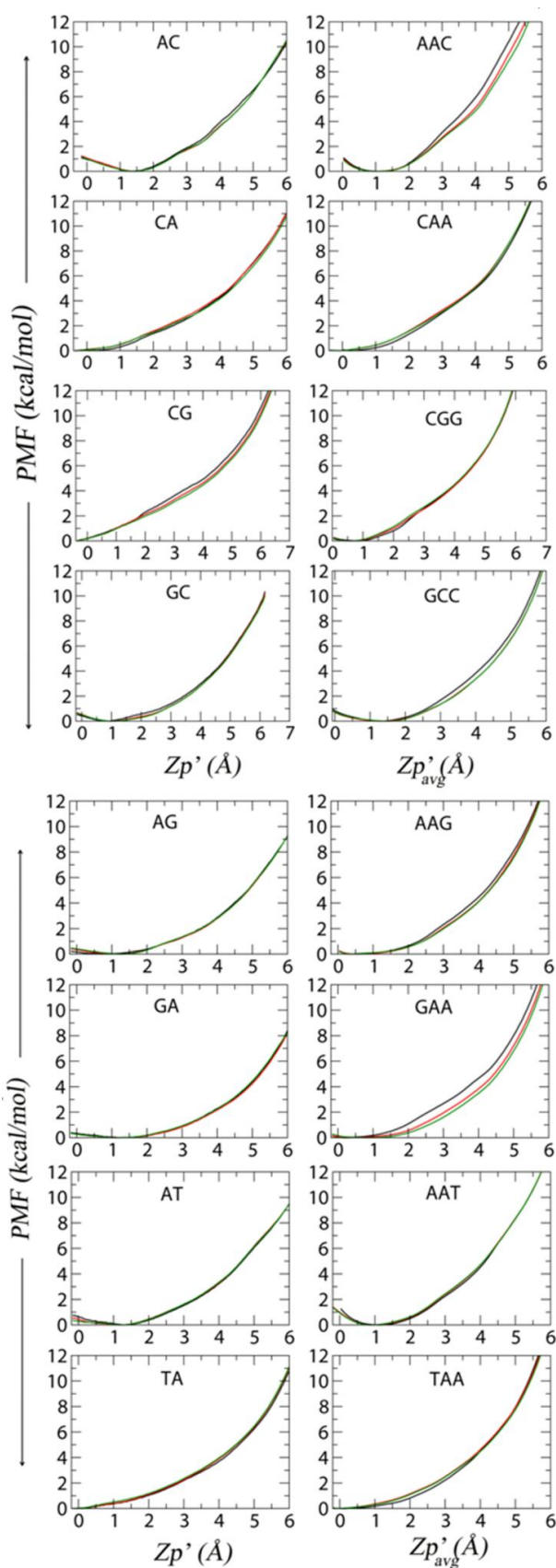


Figure A-4.2. Free energy convergence for hetero di- and trinucleotide steps.

A.4.3. various energy components along Z'_p and $\overline{Z'_p}$.

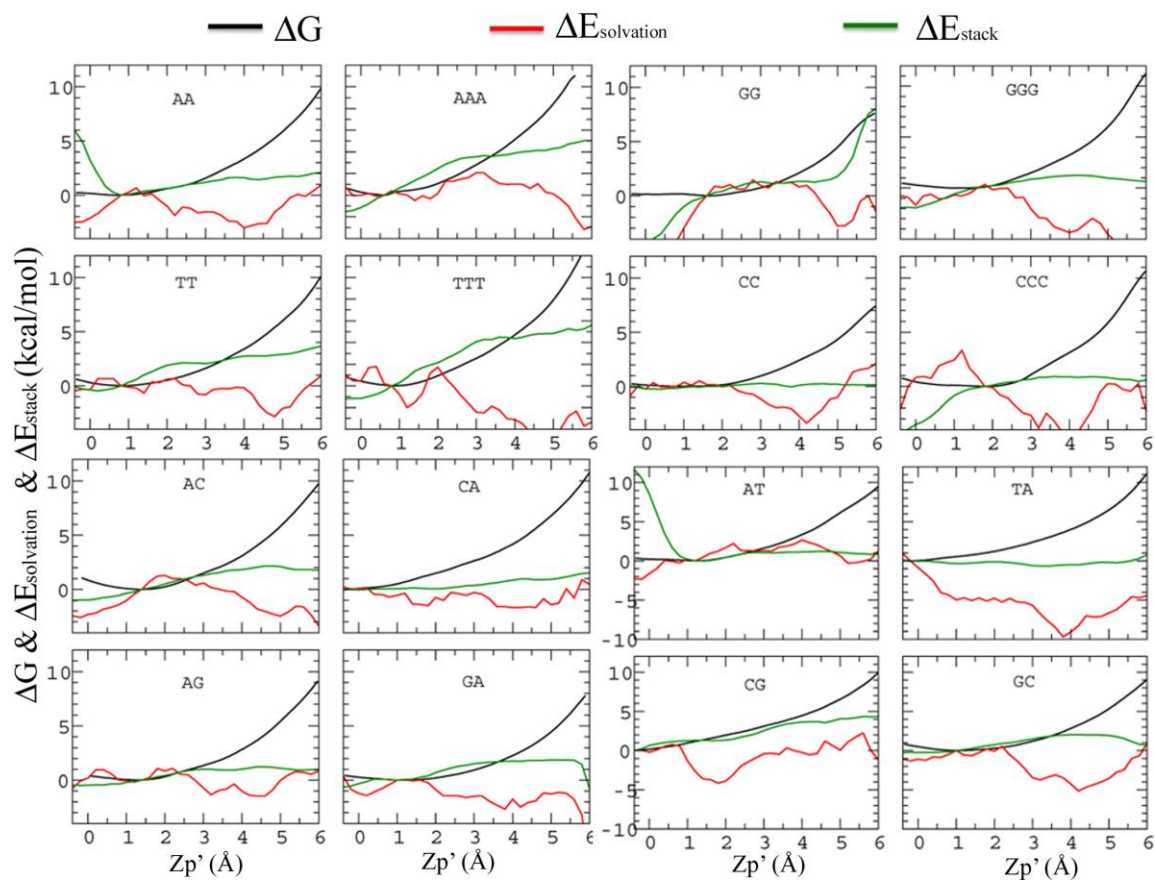


Figure A-4.3. Variation in Free energy (black line), solvation energy (red line), stacking energy (green line) along Z'_p or $\overline{Z'_p}$. The values plotted here are averages over the structures collected around certain values of Z'_p or $\overline{Z'_p}$.

Annexure 5

- A.5.1** Table listing APE and ΔG_{avFE} values for DNA sequences used as a training set.
- A.5.2** Table listing APE and ΔG_{avFE} values for DNA sequences used as a testing set.
- A.5.3** Table listing the free energy values considered for each trinucleotide step, as implemented in APTE Perl code.
- A.5.4** Table listing the ΔG_{avFE} values for 86 A-DNA promoter sequences predicted by Whitley *et al.*
- A.5.5** Table listing 358 A-DNA promoter sequences predicted using ΔG_{avFE} method in modified APTE Perl code.

Table A-5.1. Basham's APE values and present calculation for the training set. Wrong predictions are colored in red. References for all sequences are provided in third column.

sr. no.	sequence	Ref	APE (kcal/mol)	ΔG_{avFE} (kcal/mol)	Basham	Present Calculations
					Predicted Conformation	
A-DNA						
1	GCCGGC	2	0.05	1.49	B	A
2	GGGGCCCC	3	-0.36	0.93	A	A
3	GGGATCCC	4	-0.22	1.06	A	A
4	GCCCCGGG	5	-0.16	1.31	A	A
5	CCCCGGGG	6	-0.39	1.17	A	A
6	GTACGTAC	7	-0.33	1.98	A	B
7	CTCTAGAG	8	-0.57	0.85	A	A
8	GGGCGCCC	9	0.07	1.31	B	A
9	GGGTACCC	10	-0.27	1.26	A	A
10	GTCTAGAC	11	0.32	1.20	B	A
11	ACCGGCCGGT	12	0.41	1.63	B	A
12	GCGGGCCCGC	13	0.05	1.51	A	A
13	CCCGGCCGGG	14	-0.12	1.40	A	A
14	CCCCGCGGGGG	15	-0.21	1.30	B	A
15	CCGTACGTACGG	16	0.30	1.97	B	B
16	GCGTACGTACGC	16	0.44	2.06	B	B
17	GGGGGGGGGGG	17	-0.59	0.86	A	A
	Average		-0.09	1.37		
	Standard deviation		0.32	0.36		
	B-DNA					
1	CGATCGATCG	18	0.04	1.90	A	B
2	CGATTAATCG	19	0.19	2.07	B	B
3	CGATATATCG	20	0.30	1.98	B	B
4	CATGGCCATG	21	-0.03	1.91	A	B
5	CCATTAATGG	22	0.14	1.93	B	B
6	CGTGAATTCACG	23	0.36	2.13	B	B
7	CGCGAATTCGCG	24	0.65	2.14	B	B
8	CGCATATATGCG	25	0.20	2.15	B	B
9	CGCAAATTTGCG	26	0.10	2.29	B	B
10	CGCAAAAATGCG	27	0.45	2.29	B	B
11	AAAAAAAAAAAAA	28	2.06	2.34	B	B
	Average		0.41	2.08		
	Standard deviation		0.58	0.13		

Table A-5.2. Basham's APE values and our values for testing set.
The sequences for which APE values cannot be calculated are also considered in this dataset.

					Predicted Conformation	
sr. no.	sequence	ref	APE	ΔG_{avFE}	Basham	Present Calculation
A-DNA						
1	ACCCGCGGGT	29	0.41	1.63	B	A
2	ACGTACGT	30	1.08	2.19	B	B
3	ATGCGCAT	31	-0.04	2.16	A	B
4	CCCGCGGG	32	0.04	1.56	B	A
5	CCGCCGCGCG	33	0.20	1.70	B	A
6	CCGGGCCCGG	34	-0.12	1.40	A	A
7	CGCCCGCGGGCG	35	0.19	1.79	B	A
8	GACCACGTGGTC	36	0.59	1.72	B	A
9	GACCGCGGTC	37	0.73	1.63	B	A
10	GCACGCGTGC	38	0.18	2.08	B	B
11	GGATGGGAG	39	-0.45	1.20	A	A
12	GGCATGCC	40	-0.25	1.64	A	A
13	GGCCGCGGCC	41	0.23	1.51	B	A
14	GGCCGGCC	42	0.07	1.31	B	A
15	GGCGCC	43	0.40	1.49	B	A
16	GGGGCGCCCC	44	-0.09	1.21	A	A
17	GGTATACC	45 - 46	0.12	1.57	B	A
18	GTACGCGTAC	41	-0.07	2.03	A	B
19	AGGCATGCCT	47	NA* *	1.42	NA	A
20	AGGGGCCCT	48	NA	0.87	NA	A
21	CCCTAGGG	49	NA	0.91	NA	A
22	GAAGCTTC	50	NA	1.29	NA	A
23	GTGTACAC	51	NA	2.00	NA	B
					Predicted Conformation	
sr. no.	sequence		APE	ΔG_{avFE}	Basham	Present Calculation
B-DNA						
1	AAAAAAAAAAAAA A	28	2.06	2.34	B	B
2	ACCGACGTCCGT	52	0.87	1.84	B	B
3	CCGCCGCGCG	53	0.20	1.70	B	A
4	CGCAATTGCG	54	-0.39	2.28	A	B
5	CGCGATATCGCG	55	0.41	2.07	B	B
6	CGGTGG	56	0.02	1.82	B	B

7	CTCGAG	57	-1.15	1.22	A	A
8	CTCTCGAGAG	58	-1.04	1.01	A	A
9	GCGAATTCG	59	0.62	2.01	B	B
11	GCGTACGCG	60	0.52	2.05	B	B
12	GGCCAATTGG	61	-0.36	2.16	A	B
13	GGCGAATTCGCG	62	0.64	1.74	B	A
14	CAAAGAAAAG	63	NA	1.49	NA	B
15	CCAACGTTGG	64	NA	1.81	NA	B
16	CCAAGCTTGG	65	NA	2.01	NA	B
17	CCACTAGTGG	66	NA	1.54	NA	A
18	CCAGCGCTGG	67	NA	1.47	NA	A
19	CCAGGCCTGG	68	NA	1.51	NA	A
20	CCAGTACTGG	69	NA	1.21	NA	A
21	CCGCTAGCGG	70	NA	1.47	NA	A
22	CCTTTAAAGG	71	NA	1.49	NA	A
23	CGCAAAAAAGCG	25	NA	1.56	NA	A
24	CGCGAAAAAACG	22	NA	2.14	NA	B
25	CGCGTTAACGCG	72	NA	2.27	NA	B
26	CGCTAGCG	73	NA	2.27	NA	B
27	CGCTCTAGAGCG	73	NA	1.67	NA	A

**** indicates APE values cannot be calculated for a whole sequence as it possess a trinucleotide step for which APE value is not available.**

Table A-5.3. The free energy value considered for each trinucleotide step.

		Ni+1			
		C	G	A	T
Ni	C	0.860	1.960	1.630	0.765
	G	2.195	1.960	1.950	2.485
	A	2.155	1.535	2.370	2.345
	T	0.755	1.535	1.130	1.505
	N	\$ND	\$ND	\$ND	\$ND
complete C		C	C	C	C
		Ni-1			

		Ni+1			
		C	G	A	T
Ni	C	1.095	2.195	1.865	1.000
	G	1.095	0.860	0.850	1.385
	A	1.375	0.755	1.590	1.565
	T	1.375	2.155	1.750	2.125
	N	\$ND	\$ND	\$ND	\$ND
complete G		G	G	G	G
		Ni-1			

		Ni+1			
		C	G	A	T
Ni	C	1.385	2.485	2.155	1.290
	G	1.000	0.765	0.755	1.290
	A	2.125	1.505	2.340	2.315
	T	1.565	2.345	1.940	2.315
	N	\$ND	\$ND	\$ND	\$ND
complete A		A	A	A	A
		Ni-1			

		Ni+1			
		C	G	A	T
Ni	C	0.855	1.950	1.620	0.755
	G	1.865	1.950	1.620	2.155
	A	1.750	1.130	1.965	1.940
	T	1.590	2.370	1.965	2.340
	N	\$ND	\$ND	\$ND	\$ND
complete T		T	T	T	T
		Ni-1			

Table A-5.4. The ΔG_{avFE} values for 86 A-DNA promoter sequences predicted by Whitley *et al.*

Gene ID	Gene name	Motif	Presence in our analysis	Our Free ene. values
ENSXETG00000003537	plcx3	GGGCGG	Yes	0.877
ENSXETG00000008410	c5orf4	GGGCGG	Partial	0.878
ENSXETG000000030719	unknown1	GGGCGG	Yes	0.920
ENSXETG00000006282	unknown2	GGGCGG	Yes	0.880
ENSXETG00000003943	lrsam1	CCGCCC	Yes	0.900
ENSXETG000000027883	c3orf10	CCAAT	Partial	0.923
ENSXETG00000028111	unknown3	CCAAT	No	NA
ENSXETG000000016171	gata2	CCGCCC	Yes	0.942
ENSXETG00000029861	unknown4	CCAAT	No	NA
ENSXETG00000009337	gdi3	CCAAT	No	NA
ENSXETG000000012462	gtf2e1.2	CCAAT	Yes	0.925
ENSXETG000000017744	XB-GENE-5853280	CCAAT	Partial	1.043
ENSXETG000000004674	eef1d	CCAAT	Yes	1.091
ENSXETG000000004472	mcts1	CCAAT	Yes	0.937
ENSXETG000000032447	LOC100488751	CCAAT	Partial	0.946
ENSXETG000000000668	xkr5	CCGCCC	Yes	0.940
ENSXETG000000033055	unknown5	CCAAT	Yes	0.972
ENSXETG000000007609	thrsp	CCAAT	Yes	1.031
ENSXETG000000002252	unknown6	CCAAT	Yes	0.960
ENSXETG000000026459	ywhaz	TATCA	Yes	0.889
ENSXETG00000029162	unknown7	TATCA	No	NA
ENSXETG000000015053	gdpd5	TATCA	Yes	0.944
ENSXETG000000009868	tars	TATCA	Yes	0.810
ENSXETG000000010686	sepn1	TATCA	Yes	0.950

ENSXETG00000016524	LOC100493317	TATCT	No	NA
ENSXETG00000018194	fam176a	TATCT	No	NA
ENSXETG00000009404	adipor2	CCAAT	Yes	0.968
ENSXETG00000018026	sec22a	AGATA	Yes	1.075
ENSXETG00000002371	kif27	AGATA	Yes	1.075
ENSXETG00000010991	ercc4	TATCT	Yes	0.908
ENSXETG00000025304	unknown8	ATTGG	Yes	0.876
ENSXETG00000002603	gas2	TATCT	Yes	0.969
ENSXETG00000023254	zfp36l2.2	TATCA	No	NA
ENSXETG00000009124	clcn7	CCAAT	Yes	0.881
ENSXETG00000018965	crat.1	CCAAT	Yes	1.052
ENSXETG00000027013	NP_001016033.1	CCAAT	Partial	1.041
ENSXETG00000027419	a4galt	TATCA	Yes	0.993
ENSXETG00000020165	mkrr2	CCAAT	Yes	1.040
ENSXETG00000029144	unknown9	ATTGG	No	NA
ENSXETG00000030437	tnrc6a	ATTGG	Yes	0.911
ENSXETG00000018553	XB-GENE-5960869	TATCA	Partial	0.874
ENSXETG00000016062	znf184	GGGCGG	Yes	0.915
ENSXETG00000016933	ehmt1	ATTGG	Yes	0.901
ENSXETG00000014657	slc25a30	AGATA	Yes	1.142
ENSXETG00000003950	traf2	CCGCC	Yes	1.003
ENSXETG00000030164	NP_001120021.1	AGATA	Partial	0.952
ENSXETG00000030426	unknown10	TATCA	Yes	0.881
ENSXETG00000022553	fam120a	ATTGG	Yes	0.897
ENSXETG00000007987	arg2	AGATA	Yes	0.981
ENSXETG00000023393	osbp11	TGATA	Yes	0.908
ENSXETG00000017669	usp21	AGATA	Yes	0.958
ENSXETG00000013130	magi1	TATCT	Yes	1.051

ENSXETG00000023739	wrb	CCAAT	Yes	0.923
ENSXETG00000007387	bmi1	AGATA	Yes	1.120
ENSXETG00000016524	LOC100493317	ATTGG	Partial	0.978
ENSXETG00000013350	tfg	ATTGG	Yes	0.953
ENSXETG00000009412	unknown11	TATCT	Yes	0.941
ENSXETG00000000267	ccndx	CCAAT	Yes	0.881
ENSXETG00000010533	piwil2	ATTGG	Yes	0.902
ENSXETG00000007609	thrsp	TGATA	Yes	1.031
ENSXETG00000027421	HIST1H4G	TGATA	Partial	0.923
ENSXETG00000014657	slc25a30	ATTGG	Yes	1.142
ENSXETG00000014963	ctdsp1	TGATA	Yes	1.049
ENSXETG00000019650	myh11	AGATA	Yes	0.878
ENSXETG00000018194	fam176a	TATCT	Partial	0.897
ENSXETG00000029977	LOC100495404	ATTGG	Partial	0.961
ENSXETG00000008526	LOC100495179	GGGCGG	Partial	0.912
ENSXETG00000033908	UBE2U	AGATA	Yes	0.942
ENSXETG00000032885	P5F13_XENTR	ATTGG	Partial	0.848
ENSXETG00000019263	pdss2	CCAAT	No	NA
ENSXETG00000008969	rad51l3	TATCA	Partial	0.952
ENSXETG00000022325	unknown12	TATCA	Yes	0.897
ENSXETG00000020057	utp6	CCAAT	Yes	0.945
ENSXETG00000007609	thrsp	TATCT	Yes	0.942
ENSXETG00000013463	zmynd12	ATTGG	Yes	0.961
ENSXETG00000015404	shc1	TATCT	Yes	1.076
ENSXETG00000027433	otop2	ATTGG	Yes	0.967
ENSXETG00000021081	sgcg	GGGCGG	Yes	0.876

ENSXETG00000006922	ss18	TATCA	Yes	0.993
ENSXETG00000033607	asxl1	CCAAT	Yes	0.993
ENSXETG00000023477	hdhd3	ATTGG	Yes	1.052
ENSXETG00000003248	strada	TGATA	Yes	1.053
ENSXETG00000033920	F166B_XENTR	CCGCCC	Partial	0.894
ENSXETG00000010684	dnajc19	TGATA	Yes	0.989
ENSXETG00000027998	prss8	CCGCCC	Yes	1.098
ENSXETG00000010250	chrnb3	TGATA	No	NA

average value = **0.96**
standard deviation = **0.07**
maximum value = **1.14**

Table A-5.5. The list of 358 A-DNA promoter sequences predicted using ΔG_{avFE} method in modified APTE code.

Gene ID	motif	gene name	AP Sequence	ΔG_{avFE}
ENSXETG00000018957	GGGCGG	slc20a1	GGAGGGGGGGGGGGCGG	0.84
ENSXETG00000003950	GGGCGG	traf2	CCCCTCCCCCCCCCCCCAGACTGCCCGCCC	0.85
ENSXETG00000030719	GGGCGG	unknown	AGGGGGGGGGGGGGCGG	0.85
ENSXETG00000033920	GGGCGG	FAM166B	GGGGGGGGGGGGGGCCCGATAAAAATGCATGGTCCGCC C	0.86
ENSXETG00000033430	GGGCGG	unknown	AGGGGGGAGGGGGCAGTTAGGGGGGAGGGGGCGG	0.83
ENSXETG00000027998	GGGCGG	prss8	CCCCCCCCCCTTCAGTCTGTAGCTGTGCCCGCCC	0.85
ENSXETG00000024362	GGGCGG	znf568	GAGGGGGAGGAGCAGACAGGTAAAGGGGGCGG	0.81
ENSXETG00000021081	GGGCGG	sgcg	GGGGGGGGGGGGGGGGGGCAAATCTTGAGTTTAGGGGC GG	0.86
ENSXETG00000005207	GGGCGG	bhlhe40	TCTCCCCCTCCCTCCTACCCCTCCCGCCC	0.81
ENSXETG00000011510	GGGCGG	pou3f1	GGGAGGAGGGGGCGG	0.82
ENSXETG00000000265	GGGCGG	ppfia3	TCCCCCCTCCCTTACGAAGTTACCACCGCCC	0.83
ENSXETG00000020333	GGGCGG	acaa2	AGGGAGAGAGGGCGG	0.79

ENSXETG00000003619	GGGCGG	gna14	GGAGGAGGAGGAGGGTGCCTGCTCGGAACCGCCC	0.80
ENSXETG00000016681	GGGCGG	zswim6	TCCCTCCCCCTCCCCGCCC	0.83
ENSXETG00000008410	GGGCGG	faxdc2	GGGGGGGGGGGGGGGCGG	0.86
ENSXETG00000020930	GGGCGG	aldh2	CCCCCCCCCTCTCCGCTGTAGTGCCCGCCC	0.83
ENSXETG00000003971	GGGCGG	edf1	CCCTCCCCTCCCGCCC	0.82
ENSXETG00000008098	GGGCGG	rgs20	AGGGAGAGGGGGGCGG	0.81
ENSXETG00000004675	GGGCGG	maz	GGGGGGGGAGGGGGGCGG	0.84
ENSXETG00000033954	GGGCGG	unknown	CCCCCTCCCCTCTCTATACTGGGATAGGCCGCCC	0.81
ENSXETG00000003943	GGGCGG	lrsam1	TCCCCCCCCCCCCCCCCCGCCC	0.86
ENSXETG00000017813	GGGCGG	rnf219	AGGGAGAGGGGGGCGG	0.81
ENSXETG00000013207	GGGCGG	gba2	CCCCCTCCCCCAACATGTGCTATAACCGCCC	0.84
ENSXETG00000031091	GGGCGG	unknown	TCCTCCTCCTCCGCGCCAGCCCCGCCC	0.80
ENSXETG00000008526	GGGCGG	unknown	CCCCCCCCCCCCCCTTTGGCGAATTTTGGGGCGG	0.85
ENSXETG00000016062	GGGCGG	znf184	GGGGGGGGGGGGGAAACAGGGGGGCGG	0.86
ENSXETG00000016171	GGGCGG	gata2	CCCCCCCCCCCCCCCCCGCCC	0.86
ENSXETG00000000668	GGGCGG	xkr5	TCCCCCCCCCCCCCGCCC	0.86
ENSXETG00000030333	GGGCGG	unknown	TCCCCTCTCTCTCTCCGCCC	0.79
ENSXETG00000012160	GGGCGG	sreb1	AGGAGGAGAGGGGGCGG	0.80
ENSXETG00000032326	GGGCGG	sra1	AGGAGGAGGAGGGGCGG	0.80
ENSXETG00000012753	GGGCGG	XB-GENE- 5921670	GAGGGGAGGAGGGGGCGG	0.81
ENSXETG00000024302	GGGCGG	atp10a	AGGAGAGAGGAGGGCGG	0.78
ENSXETG00000007561	GGGCGG	cat.1	TCTCTCTCCCCCGCGGGGGCGG	0.80
ENSXETG00000017992	GGGCGG	smox	AGGAGGGAGAGGGGCGG	0.80
ENSXETG00000006282	GGGCGG	unknown	GGGGGGGGGGGGGCGG	0.86
ENSXETG00000016857	GGGCGG	tenm2	TCTCTCTCTCTCTCCACTGGCCGGCTCTCCGCCC	0.77
ENSXETG00000024580	GGGCGG	aqp5	AGGAGGAGGAGGAGGGCGG	0.79
ENSXETG00000011808	GGGCGG	pak2	AGGGGGGGGGGGGCGG	0.85
ENSXETG00000008142	GGGCGG	tnfaip8l2	TCTCCCTCCTCCCCAGCAGATGGGCGG	0.81
ENSXETG00000021323	GGGCGG	ube2d3	CCTCCCTCCTCCTCCGCCC	0.80
ENSXETG00000003537	GGGCGG	plcx3	GGGGGGGGGGGGGGGCGG	0.86

ENSXETG00000033314	GGGCGG	znf684	TCCCCTCCCCTCTCAGTGGTACATTCCCATGGGCGG	0.81
ENSXETG00000023496	GGGCGG	ipo5	CCCCTCCCCTCCCGCACGTAGCCCCGCC	0.82
ENSXETG00000007231	GGGCGG	fgfr4	AGAGGGAGGGGGAGCAGAGCCTGGGGGCGG	0.81
ENSXETG00000004757	GGGCGG	aldh9a1	GGGAGGGGGAGGTGGGCGG	0.82
ENSXETG00000025633	GGGCGG	dap	AGGGGGAGGAGGAGAATGGAGAAAAAAGGAAATGGG GCGG	0.80
ENSXETG00000001379	GGGCGG	pdgfa	AGAGAGAGAGAGAGGGGGGCGG	0.78
ENSXETG00000024038	GGGCGG	tecpr1	GGAGAGAGGGGGGCGG	0.80
ENSXETG00000023724	GGGCGG	ets2	CTCCTCCCCTCCTCCCTTGTGTGTAATGGGCCGCTCCGC CC	0.81
ENSXETG00000024708	TGATA	unknown	TCTCTCTCCCCCTGCTTACTTATCA	0.79
ENSXETG00000025846	TGATA	unknown	CTCCTCCCCCCCAGGTATCA	0.83
ENSXETG00000026458	TGATA	sox5	AGGGGGAGGGGGGAGTTTTTATAAATGACATTGGTATC A	0.83
ENSXETG00000031033	TGATA	unknown	AGAGGGAGGAGGAACAATATCA	0.79
ENSXETG00000018417	TGATA	sall1	CCTCCTCCCTCCCTCCCCGTGAGCCGGCCCCCTTATCA	0.81
ENSXETG00000002396	TGATA	fgfr3	TCCTCTCCCCCCCCCATAGTGTAATCTGTGATA	0.83
ENSXETG00000023393	TGATA	osbpl11	GGGGGGGGGGGGTGGAGAAAATGATA	0.86
ENSXETG00000000957	TGATA	anapc4	GGGAGGGAGAGAGACAGCACCTGGGCAGCTATCA	0.79
ENSXETG00000025981	TGATA	KIAA1328	AGGGGGGAGGGGCTAACCCTAGCCTGATA	0.83
ENSXETG00000031613	TGATA	unknown	AGGAGAGGAGGAAGAGGACACCAGGATATCA	0.79
ENSXETG00000011494	TGATA	ftcd	TCCCTCCCTCCCCCAGAGTACTGATA	0.83
ENSXETG00000018470	TGATA	vill	GGGGGGGAGGGGCAATATTATTATCA	0.84
ENSXETG00000010684	TGATA	dnajc19	TCCCCCCCCCCCCCTTTAAATGGGCCTCTGATTGATA	0.85
ENSXETG00000010686	TGATA	sepn1	GAGGGGGGGGGGATATCA	0.84
ENSXETG00000003248	TGATA	strada	CTCCCCCCCCCCCCAGAAAATCACAGGCAGAGTTGATA	0.85
ENSXETG00000007609	TGATA	thrsp	TCCCCCCCCCCCCAATGTGGTGGGTTGATA	0.86
ENSXETG00000008969	TGATA	rad51d	CCCCCCCCCCCCCAAACAGTCATGTTTATATCA	0.86
ENSXETG00000033339	TGATA	unknown	AGAGGGAGGGGGAACATGTTATTATCA	0.81
ENSXETG00000023915	TGATA	rbp5	GGGGGGGAGGGGGGGATATCA	0.84
ENSXETG00000005516	TGATA	cntnap5	GAGGGGGGGGGATATCA	0.84

ENSXETG00000015528	TGATA	bcor	AGGGGAGGGGAGGACCCTATCA	0.82
ENSXETG00000027421	TGATA	unknown	GGGGGGGGGGGATTCCAAGGTCTCTGATA	0.86
ENSXETG00000027419	TGATA	a4galt	AGGGGGGGGGGGGGGTAAATATCA	0.85
ENSXETG00000030842	TGATA	unknown	CCTCCTCCTCCTCTGAGCAGAGGGGCGCCTGATA	0.78
ENSXETG00000009239	TGATA	fgd6	AGGGAGGGGAGATATCA	0.81
ENSXETG00000013220	TGATA	ttl5	CCCCTCCCCTCTCCTCATGAAGTCTGATA	0.80
ENSXETG00000018631	TGATA	nxph3	GAGAGGGGGGAGTGATA	0.81
ENSXETG00000031859	TGATA	unknown	AGGAGAGGGGAGAGAAGGTGGAAGATATCA	0.79
ENSXETG00000005003	TGATA	gyg2	GAGAGAGAGAGAGAGAGAGAGAGAGAGGGAGAGAGAGA GAGAGGGGGGGAGGGAGTGATA	0.78
ENSXETG00000018553	TGATA	dnah14	GGGGGGGGGGGGAGGAGGGATATCA	0.84
ENSXETG00000004828	TGATA	unknown	GGGGGAGGGGGGCGAGAAGTAAAAGTGATA	0.84
ENSXETG00000008178	TGATA	immp1l	TCCCTCCTCCCCTGCCAGGATTTTCATATGATA	0.81
ENSXETG00000015216	TGATA	tmem180	GGGGGGGGAGGAAGTGAGCGGTATCA	0.84
ENSXETG00000022325	TGATA	unknown	GGGGGGGGGGGAGATGTCTGAAAAACAATATCA	0.84
ENSXETG00000022336	TGATA	olfm1	AGGAGAGAGAGAGATAGATAGATAGATGATA	0.76
ENSXETG00000009128	TGATA	ptx4	AGAGAGAGAGAGAGAGAATCTTGTATATCA	0.76
ENSXETG00000002260	TGATA	tmed1	TCCCTCCCCCTGTGGGATAATGATA	0.83
ENSXETG00000005197	TGATA	cenpa	AGGAGGAGGGGGTGAAAATTCCATTATCA	0.81
ENSXETG00000021874	TGATA	nanos3	TCTCTCCCCTCCAGACAGGTCTGTCCCTATCA	0.80
ENSXETG00000017777	TGATA	cfap44	GGGAGGAGGGGGGATTACTTGAAGGGGCTTATCA	0.82
ENSXETG00000027601	TGATA	mex3b	CCCCCCTCCTCTTTTAAGGAACAATATCA	0.81
ENSXETG00000006168	TGATA	klhl40	GAGGGGGGGGGGATATCA	0.84
ENSXETG00000030349	TGATA	unknown	GGAGGAGGGGAGTATCA	0.81
ENSXETG00000011369	TGATA	pycard	GGGGGAGGGGAGGCAAATACATCTGGTCTGATA	0.82
ENSXETG00000005076	TGATA	gfm2	AGGAGGGGGGGATGATA	0.83
ENSXETG00000019108	TGATA	ckm	TCCCCCTCTCTCTCTGCTTCACCCAGTGTATGATA	0.79
ENSXETG00000021498	TGATA	c18orf25	GAGGGGGGGGGGAAGACCTGATA	0.84
ENSXETG00000002780	TGATA	clip3	CTCCTCCCTCTCGGTATCA	0.79
ENSXETG00000003760	TGATA	setd1b	TCTCCCCCTCCCGCGGTCTCTTGTAGCCTGATA	0.82
ENSXETG00000001242	TGATA	clmp	AGAGAGAGAGAGATGATA	0.76

ENSXETG00000004791	TGATA	nipbl	TCTCCTCCCTCCTGTGATA	0.80
ENSXETG00000033070	TGATA	DLG3	CTCTCTCCCCCTCCCTATCCAGGCCCCCGTATCA	0.81
ENSXETG00000026459	TGATA	ywhaz	GGGAGGGGGGGGGGGGGGATATCA	0.85
ENSXETG00000014916	TGATA	tmbim1	AGGAGGAGGGAGGGGGGGTGTATCA	0.82
ENSXETG00000014963	TGATA	ctdsp1	GGGGGGGGGGGGGTACCGTAAGTAGCTGATA	0.86
ENSXETG00000015053	TGATA	gdpd5	GAGGGGGGGGGGGGATATCA	0.84
ENSXETG00000001773	TGATA	fitm2	CCCCCCCCCCTTGATA	0.85
ENSXETG00000019566	TGATA	fgf1	AGGGAGAGAGGAGAAAAATATTGATA	0.78
ENSXETG00000009868	TGATA	tars	GGGAGGGGGGGGGGGGATATCA	0.85
ENSXETG00000014847	TGATA	foxo3	CTCTCTCCTCCCATTGGGCAGTGGCCTGATA	0.79
ENSXETG00000019410	TGATA	pafah1b1	CTCCCTCCTCCTCCTCCTTCAGCGGTTGATA	0.79
ENSXETG00000030426	TGATA	unknown	GGGGGGGGGGGGGCAGGGGTGTATCA	0.86
ENSXETG00000006922	TGATA	ss18	GGGGGGGGGGGGGGGGGGGGGAAGGTAGAATTCAGCCC ATTATCA	0.86
ENSXETG00000002406	TGATA	sim1	TCTCTCTCTCTCTCTCTCTATCTGCCTATCA	0.76
ENSXETG00000019176	TGATA	yes1	GGGAGGGAGAGGCCTGGGGCTGGGCTAGTATTGATA	0.80
ENSXETG00000025005	TGATA	sox17a	AGAGGGAGAGAGTAAAACTGATA	0.78
ENSXETG00000030945	TGATA	unknown	GGAGGAGGAGGGTCCCTCGCAGGCATCCGTGATA	0.80
ENSXETG00000008072	TGATA	magoh	GGAGGAGGGGGGGGGGGTGTGATA	0.83
ENSXETG00000026123	TGATA	unknown	TCCTCTCCCCCCCCCTCCATCCCACCCTGATA	0.82
ENSXETG00000030228	TGATA	unknown	AGGGGAGAGAGATTTCACTGATA	0.79
ENSXETG00000017087	TGATA	tnni1	GAGGGGGAGGGGACAGAATAATCTGATA	0.82
ENSXETG00000023254	TGATA	zfp36l2.2	GGGGGGGGGGGGGCAATATCA	0.86
ENSXETG00000017177	TGATA	npc1	CTCCTCCTCTCCTATCA	0.79
ENSXETG00000032290	TGATA	clps	AGAGAGAGGGGGTAATTATATCA	0.79
ENSXETG00000001986	TGATA	ckap5	CTCTCCCTCCCCCAGACTAAATTGCACAGTATGATA	0.82
ENSXETG00000020879	AGATA	foxh1	CTCCCTCCCTCTGAGCTCTTGTGGGAGATA	0.80
ENSXETG00000014820	AGATA	lrrc16a	CCTCCCTCCTCCTATCT	0.80
ENSXETG00000015404	AGATA	shc1	CCCCCCCCCCCCCAGTCACAATGTGCACTTATCT	0.86
ENSXETG00000023190	AGATA	dgat2l6	TCCCTCTCTCCAGAAATATCT	0.80
ENSXETG00000007387	AGATA	bmi1	AGGGGGGGGGGGGTCATTTTACCAGATA	0.85

ENSXETG00000001445	AGATA	rad18	GGAGAGAGAGAGAGAGAGACTATTCGGTATGTATGAGATA	0.76
ENSXETG00000006355	AGATA	maff	TCCTCCTCCCCTCTCTAGTGTATCT	0.79
ENSXETG00000016016	AGATA	elfn2	CTCCCCCTCCCTGTTATCT	0.82
ENSXETG00000023014	AGATA	mpz11	GGAGAGGAGGGAATGAGATA	0.80
ENSXETG00000013212	AGATA	rgp1	TCCTCTCTCTCTATCT	0.77
ENSXETG00000027510	AGATA	itgb6	GGGGGGAGGGAGATA	0.82
ENSXETG00000015270	AGATA	lingo1	TCCCCTCCCCCTTACAGAACAGTAGGCCACATATCT	0.83
ENSXETG00000000057	AGATA	acot2	AGAGGGAGGAGATA	0.79
ENSXETG00000018194	AGATA	eva1a	GGGGGGGGGGGGGGGGGTGTATCT	0.86
ENSXETG00000014657	AGATA	slc25a30	AGGGGGGGGGGGGGGGGTACTGAAAGATA	0.85
ENSXETG00000021668	AGATA	xpot	GAGAGAGGGGGGGGGGAGATA	0.81
ENSXETG00000010991	AGATA	ercc4	CCCCCCCCCCCCCCCCCAAATATCT	0.86
ENSXETG00000030576	AGATA	unknown	GAGAGGGGGAGATA	0.80
ENSXETG00000032550	AGATA	unknown	GAGAGGGGGAGATA	0.80
ENSXETG00000016732	AGATA	cnn2	CCCCTCTCCTCTATCT	0.79
ENSXETG00000007609	AGATA	thrsp	TCCCCCCCCCCCCAATGTGGTGGGTGATATCT	0.86
ENSXETG00000018081	AGATA	prc1.2	GGGGGGGGGAGATA	0.84
ENSXETG00000002371	AGATA	kif27	GGGGGGGGGGGATCAGATA	0.86
ENSXETG00000014612	AGATA	psme1	CCTCCTCCTCTCCCCTCCCTCTATCT	0.79
ENSXETG00000014643	AGATA	rabggt	GGGAGAGAGGGAGATA	0.79
ENSXETG00000016448	AGATA	itpr2	AGGGAGAGGAGGGCAGATA	0.79
ENSXETG00000031859	AGATA	unknown	AGGAGAGGGGAGAGAAGGTGGAAGATA	0.79
ENSXETG00000024242	AGATA	unknown	TCTCCCCCTCCCCCAACAGATA	0.83
ENSXETG00000034039	AGATA	unknown	GGAGAGGGGGAGTCAAGCGCATCTCTGGCCTTATCT	0.81
ENSXETG00000007898	AGATA	neurog1	AGGGGGGGGAGAGTTAACTTACCATCCATATCT	0.82
ENSXETG00000017087	AGATA	tnn1	GAGGGGGAGGGGACAGAATAATCTGATAAGAGATA	0.82
ENSXETG00000003659	AGATA	unknown	TCTCCTCCCCCTCCCCAACCCTAAAGGAGATA	0.82
ENSXETG00000005163	AGATA	prrt1	AGGAGGAGGAGGAGGAGAAAAGGGAGGGGGAGAATA TCT	0.79
ENSXETG00000018294	AGATA	unknown	GGGGGGGGGAGATA	0.84
ENSXETG00000002996	AGATA	tyw3	AGAGGGGGGGGAAGGAAAGATA	0.83

ENSXETG00000031349	AGATA	unknown	AGAGAGAGGGGGAAAAGGTAGGAGGAAAGATA	0.79
ENSXETG00000027532	AGATA	unknown	GGGGGGAGGGGGAGATA	0.83
ENSXETG00000016327	AGATA	mrrf	CCTCCTCTCTCCAGATA	0.79
ENSXETG00000025992	AGATA	cdh3	AGGGGAGAGGAGATA	0.79
ENSXETG00000016048	AGATA	foxo1	TCTCCTCTCTCCCTCATTAGATTAAAACAGTTACAGATA	0.78
ENSXETG00000032579	AGATA	pnpla6	CCTCTCTCTCTCTCTCTCTCTGACAGATA	0.76
ENSXETG00000031662	AGATA	unknown	TCTCTCTCCTCTCTCTCCCCTTCCTTTCTAATTCTTCTTTA TCT	0.78
ENSXETG00000009239	AGATA	fgd6	AGGGAGGGGAGATA	0.81
ENSXETG00000005611	AGATA	krt18	TCCCTCTCTCTCTCTATCT	0.77
ENSXETG00000030510	AGATA	unknown	TCTCCTCCTCCTCCCCAACACCCTAAAGGAGATA	0.81
ENSXETG00000023914	AGATA	XB-GENE-995386	AGAGGGAGGAGAAAAGAGGTCACATATAGGTATCT	0.79
ENSXETG00000012101	AGATA	aplp1	AGAGGGGAGGAGAGATA	0.79
ENSXETG00000005808	AGATA	itgae	GAGGGAGGGGGGGGGGGGTATCT	0.83
ENSXETG00000014206	AGATA	pax2	AGGGGGGGGGGGGTATCT	0.85
ENSXETG00000018896	AGATA	wbp4	GGGGAGGGAGGAAGTGTGAGATA	0.82
ENSXETG00000016710	AGATA	unknown	AGGAGGAGAGAGAGATA	0.77
ENSXETG00000007311	AGATA	camkv	GGAGGGAGAGGGGTGGTTGAAGGTATTTTATCT	0.80
ENSXETG00000002406	AGATA	sim1	TCTCTCTCTCTCTCTCTCTCTATCT	0.76
ENSXETG00000005003	AGATA	gyg2	GAGAGAGAGAGAGAGAGAGAGAGAGAGGGGAGAGAGAGA GAGAGGGGGGGAGGGAGTGATAAATATCT	0.78
ENSXETG00000001734	AGATA	ube2ql1	CCCTCCCCCTCTTTTGTGAGGTCTGCAGATA	0.81
ENSXETG00000023132	AGATA	stx6	CCCCTCCCCCTCCAGTATCT	0.82
ENSXETG00000031950	AGATA	unknown	AGAGAGAGAGGAAGTCTTTAGTTTTGTATCT	0.77
ENSXETG00000022336	AGATA	olfm1	AGGAGAGAGAGAGATA	0.76
ENSXETG00000022336	AGATA	olfm1	AGGAGAGAGAGAGATAGATAGATA	0.76
ENSXETG00000006028	AGATA	c15orf40	AGAGAGAGGAGAGATA	0.76
ENSXETG00000033085	AGATA	rln3	GGGGAGGGGGGGGCACCGAGTAGGGTCCTTATCT	0.84
ENSXETG00000016524	AGATA	unknown	GGGGGGGGGGGGGTATCT	0.86
ENSXETG00000013759	AGATA	drc7	CCCCTCCCTCCCCTCGGATATCT	0.81
ENSXETG00000017669	AGATA	usp21	AGGGGGGGGGGGGATACACAGTCAGATA	0.85

ENSXETG00000031651	AGATA	unknown	CCCCCCCCTCCTAATCAGATTAGATA	0.83
ENSXETG00000020568	AGATA	nxnl1	CCTCCCTCCCTCTATCT	0.80
ENSXETG00000008527	AGATA	unknown	AGGGGAGGAGGAGGCATATGGAATTTAAGGGTATATCT	0.80
ENSXETG00000000727	AGATA	unknown	GGGGGGAGGGGAGATA	0.82
ENSXETG00000016619	AGATA	unknown	AGAGAGGGGGAGATA	0.79
ENSXETG00000015135	AGATA	unknown	AGGAGAGAGGAGAGAGCAGAGAGCTGCGTATCT	0.77
ENSXETG00000031157	AGATA	unknown	AGAGAGAGAGAGAGAGAGAGAGAGAAGGATAGATA	0.76
ENSXETG00000024747	AGATA	smc1a	AGAGAGAGAGAGAGAGAAATATATAGAGAGATA	0.76
ENSXETG00000020700	AGATA	slc25a15	CCCTCTCCCTCTGTTCCGTACATAGATA	0.79
ENSXETG00000019485	AGATA	sec14l5	AGGGGAGAGGGAGGGAGAGTACAAGAGATA	0.80
ENSXETG00000024253	AGATA	unknown	GGGGGGAGGGGGAGATA	0.83
ENSXETG00000007126	AGATA	unknown	GGGGGGAGGGGGAGATA	0.83
ENSXETG00000030333	AGATA	unknown	TCCCCTCTCTCTCTCCGCCCTCAGTATCT	0.79
ENSXETG00000019650	AGATA	myh11	GGGGAGGGGGGGGGGAGAAATTTGCAACTCAGATA	0.83
ENSXETG00000027960	AGATA	unknown	CCCCCCCCTCCCCGTATCT	0.84
ENSXETG00000009096	AGATA	vamp2	AGGGAGAGGGGAGAGGCAGGAAGGAAGATA	0.79
ENSXETG00000004468	AGATA	gse1	AGGGGGGGGGGGTAGGAGTATTGTATGTAGATA	0.85
ENSXETG00000005325	AGATA	cyb561	AGGAGGGGGGGGAGAGAGATA	0.80
ENSXETG00000005733	AGATA	slc25a34	GGAGGGAGGGGGGGCAGATA	0.83
ENSXETG00000009836	AGATA	nfxl1	AGGGAGAGAGGAGATA	0.78
ENSXETG00000018026	AGATA	sec22a	GGGGGGGGGGGATCAGATA	0.86
ENSXETG00000024675	AGATA	dnaaf3	CTCCCCTCCCCCATACTATCT	0.83
ENSXETG00000033908	AGATA	UBE2U	GGGGGGGGGGGGGGGGGTCTCCCTACATGAGCAGATA	0.86
ENSXETG00000017226	AGATA	ache	CTCCTCTCTCTCTCTTGCGAGGTCACAGCTGTTATCT	0.76
ENSXETG00000013130	AGATA	magi1	CTCCCCCCCCCCCCACCCTGGGCTTATCT	0.85
ENSXETG00000015072	AGATA	unknown	AGAGAGGGGGAGATA	0.79
ENSXETG00000003833	AGATA	hnnpa0	TCCCTCTCCTCCTTTCGCCACCAGATA	0.80
ENSXETG00000001668	AGATA	foxd4l1.2	CCCCCTCCCTCCTAGGTGCAGATA	0.81
ENSXETG00000018791	AGATA	tgfbli1	TCTCCCTCTCTCTCCCTATCT	0.78
ENSXETG00000030164	AGATA	unknown	TCCTCCCCCCCCCATGTACAAAGATA	0.84
ENSXETG00000012170	AGATA	hnnpa2b1	TCCCTCCTCCCCTTCTATTATATCT	0.81

ENSXETG00000023250	CCAAT	GNMT	GGAGGAGGAGGAGCCATATATTGG	0.79
ENSXETG00000022498	CCAAT	unknown	TCCCCCCCCCTCACCCCCAGCAGCCAAT	0.84
ENSXETG00000000267	CCAAT	ccndx	GGGGGGGGGGGGGCATATAAATGCCCAAT	0.86
ENSXETG00000029977	CCAAT	unknown	CCCCCCCCCCCCCGCTGCCCAGAACCTATTGG	0.86
ENSXETG00000016933	CCAAT	ehmt1	GGGGGGGGGGGGGGGACCTCCCGATTGG	0.86
ENSXETG00000019052	CCAAT	ing2	CTCCCCCTCCCTATCTTCCTTTGAACGTCCAAT	0.82
ENSXETG00000016681	CCAAT	zswim6	TCCCTCCCCCTCCCCGCCCCCCCCCAAT	0.83
ENSXETG00000013350	CCAAT	tfg	TCCCCCCCCCCCCCAACTTATAAACATTGG	0.86
ENSXETG00000022348	CCAAT	sox9	CCCTCTCCCCTCTCCAAATCCCTACAGCCAAT	0.80
ENSXETG00000033607	CCAAT	ASXL1	CCCCCCCCCCCCCACCCCCAGAGAACTAACCAAT	0.86
ENSXETG00000014391	CCAAT	kcnc1	AGGAGGGGGGGGCTATTAGCCAAT	0.83
ENSXETG00000003166	CCAAT	cnot3	CCTCCTCTCCCTAAAATGATGTAGACATTCACCAAT	0.79
ENSXETG00000033657	CCAAT	unknown	CCCCTCCCTCCCACAAGGAGCCAAT	0.82
ENSXETG00000028016	CCAAT	znf333	AGGAGGGGGAGAGCCAAT	0.80
ENSXETG0000001957	CCAAT	cul4a	AGGAGGGAGAGGAAAGGTGGATTGG	0.79
ENSXETG00000027419	CCAAT	a4galt	TCTCTCTCCCTCTCAGACGGCCAAT	0.77
ENSXETG00000012462	CCAAT	gtf2e1.2	CTCCCCCCCCCCCCCAAT	0.85
ENSXETG00000016524	CCAAT	unknown	GGGGGGGGGGGGTATCTAACACATTGG	0.86
ENSXETG00000010533	CCAAT	piwil2	GGGGGGGGGGGGGACATGCCAACACTATTGG	0.86
ENSXETG00000007609	CCAAT	thrsp	TCCCCCCCCCCCCAAT	0.86
ENSXETG00000013017	CCAAT	unknown	TCCTCTCCCCCAAT	0.82
ENSXETG00000032854	CCAAT	unknown	GGGGAGGGGAGAGGGAATTGG	0.81
ENSXETG00000033685	CCAAT	TLCD2	GGGGAGGGGAGCAGTGACTGACCAAT	0.83
ENSXETG00000011036	CCAAT	map2k4	GAGGGAGGGAGGCGGCCGGTCGCCGACAGCCAAT	0.80
ENSXETG00000030401	CCAAT	unknown	AGGAGAGAGAGGGTCCCGGCCAAT	0.78
ENSXETG00000033828	CCAAT	A2ML1	AGAGGGGAGAGAGAGAATGGGTAAACGGAGAATTGG	0.78
ENSXETG00000033603	CCAAT	unknown	GAGGGGGAGGGGATGGGTGGGAAGAAGATGATTGG	0.82
ENSXETG00000008182	CCAAT	unknown	CCTCCCCTCCCCAGAGTCCCCCAGTGTCCAAT	0.82
ENSXETG00000002298	CCAAT	znf385c	TCCCTCCCCCCCCCATAGTCATTGG	0.84
ENSXETG00000009404	CCAAT	adipor2	CCCCCCCCCCCCACCCAAT	0.86
ENSXETG00000009124	CCAAT	clcn7	CCCCCCCCCCCCGCACCAAT	0.86

ENSXETG00000016987	CCAAT	zfp3611	TCCCTCTCTCTCTCTCTCTCATACCAAT	0.77
ENSXETG00000023069	CCAAT	tdrd9	TCCCTCCCCCCCCCTCCCCTCTGCCAGATCCGCTTCTGATT GG	0.82
ENSXETG00000032885	CCAAT	unknown	GGGGGGGGGGGGGGGAGAGCGAAGGGTGCTAGATTGG	0.84
ENSXETG00000028079	CCAAT	aloxe3	GAGAGAGAGAGAGAGAGAGAGAGAACTGTCAGAGGAA ATTGG	0.76
ENSXETG00000026680	CCAAT	unknown	CCTCTCCTCTCTAATTGG	0.77
ENSXETG00000010464	CCAAT	pbx1	AGGGGGAGGGGGAGTGACCTGCCAAACTACCAAT	0.82
ENSXETG00000009858	CCAAT	TMEM8B	TCCCTCCCCCCCCGCCAAT	0.84
ENSXETG00000007545	CCAAT	alx4	GGAGGGAGAGGGGGAGGGAATTGG	0.81
ENSXETG00000007561	CCAAT	cat.1	TCCTCCTCCTCCTCCTCTCCCAAT	0.80
ENSXETG00000031251	CCAAT	RORB	AGGAGAGGGAGAGAGAAAGAGAGCCCCAAT	0.78
ENSXETG00000019330	CCAAT	r3hdm1	CCCCCCCCCCTCAAGTCACTGATTGG	0.84
ENSXETG00000012515	CCAAT	dhx9	CCCTCCTCTCTCGTCGAACGGGATTGG	0.78
ENSXETG00000021653	CCAAT	lppr3	TCTCTCCCTCCTTCCCTCCATCTCTTCTAGATTGG	0.79
ENSXETG00000032447	CCAAT	THRSP	TCCTCCCCCCCCCCCCAAT	0.84
ENSXETG00000002252	CCAAT	unknown	GGGGGGGGGGGGGGGGGGGACCAAT	0.86
ENSXETG00000008526	CCAAT	unknown	AGGGAGGAGAGGCTGCCTGGCGCCCCCAAT	0.79
ENSXETG00000006617	CCAAT	adat3	AGAGAGGAGGAGAAGATGGGAAGACTTACCTGATTGG	0.77
ENSXETG00000023638	CCAAT	cox11	GAGAGAGAGAGAGAGAGAGAGGGGGGGGATGTTGCTGT GGAGATTGG	0.78
ENSXETG00000013668	CCAAT	spcs1	GAGGAGGAGGAGCTTGACAGACATGGCAGCTGATTGG	0.79
ENSXETG00000023477	CCAAT	hdhd3	CCCCCCCCCCCCCCCCACAGGTGCCAGGGGACACATTGG	0.86
ENSXETG00000018791	CCAAT	tgfb1i1	TCCCCCCTCCCCAAT	0.84
ENSXETG00000021236	CCAAT	fth1	CCCCCCCCCCTTCTTTCTCCAAT	0.85
ENSXETG00000020165	CCAAT	mkrn2	AGGGGGGGGGGGGAAGAACCAAT	0.85
ENSXETG00000012522	CCAAT	nrbp1	TCCTCTCTCCTCCAAT	0.78
ENSXETG00000027956	CCAAT	rbm8a	GGGGGGAGGGGGAATCAGTGATCCAAT	0.84
ENSXETG00000015812	CCAAT	ppp4r3a	TCCCTCTCCTCCTGCTCCTATGAATTGG	0.80
ENSXETG00000022831	CCAAT	pabpc1l	AGGAGGGGGAGAGCCAAT	0.80
ENSXETG00000033055	CCAAT	unknown	GGGGGGGGGGGGGGGGGACCAAT	0.86

ENSXETG00000027883	CCAAT	brk1	CCCCCCCCCCCCCCCCCAAT	0.86
ENSXETG00000003876	CCAAT	slbp	TCCTCCTCCCCTGCGCTGTAATTGG	0.81
ENSXETG00000026772	CCAAT	rhob	CTCTCCTCCTCCAAT	0.79
ENSXETG00000025361	CCAAT	unknown	AGGAGGGGAGAGCCACAGTGCTTGAAGGGAATTGG	0.80
ENSXETG00000020057	CCAAT	utp6	CCCCCCCCCCCCCTGCATTACTGGGGTTCCAAT	0.85
ENSXETG00000027433	CCAAT	otop2	TCCCCCCCCCCCCCTAGTTGGGATTTGTCCATTGG	0.85
ENSXETG00000009337	CCAAT	unknown	CCCCCCCCCCCCCCCCCAAT	0.86
ENSXETG00000022694	CCAAT	uhrf1bp1l	TCCCCCCTCCTCTAACTTTGTTCATTGG	0.82
ENSXETG00000026334	CCAAT	CACNB1	CCCCCCTCTCTCTCTCACTCTCTCCAAT	0.79
ENSXETG00000010865	CCAAT	aldh1a2	CCCCCCTCCTCTCTCCCTATGTCCAAT	0.81
ENSXETG00000018708	CCAAT	axl	CCTCTCCTCCCTCCCTCCCTCTGTGGTGGGATGAACTAT TGG	0.79
ENSXETG00000013217	CCAAT	esrrb	CCTCCCTCCCCCTTTGACAGAGAGCCAAT	0.81
ENSXETG00000015270	CCAAT	lingo1	TCCCTCCCCCCCCCATTCCAAT	0.84
ENSXETG00000004472	CCAAT	mcts1	CCCCCCCCCCCCCAAT	0.86
ENSXETG00000018294	CCAAT	unknown	GGGGGGGGGAGATATTAAATGATGATTAAGTCCAAT	0.84
ENSXETG00000000887	CCAAT	rraga	TCCCCCCTCTCTTCCTATTGG	0.81
ENSXETG00000023671	CCAAT	XB-GENE- 5901104	GGGGAGGGGGGAGTTGCTGAATATTGG	0.83
ENSXETG00000001855	CCAAT	c7orf25	AGGAGGGGGAGAGCCAAT	0.80
ENSXETG00000026458	CCAAT	sox5	AGGGGGAGGGGGGAGTTTTTATAAATGACATTGG	0.83
ENSXETG00000002609	CCAAT	svep1	GGAGAGGAGAGGAGGGGGGGTGAAGGAAGGAAGATTG G	0.80
ENSXETG00000018965	CCAAT	crat.1	CCCCCCCCCCCCCAGCCCAAT	0.86
ENSXETG00000026752	CCAAT	trim72	AGGAGGGGGAGAGCCAAT	0.80
ENSXETG00000014957	CCAAT	chrne	CCCCCCTCTCTCTCATAGGGGGTCTGGGGACCCCAAT	0.80
ENSXETG00000002553	CCAAT	ccm2l	AGAGGAGAGGAGCATAGAGGGAGAATTGATTGG	0.78
ENSXETG00000022369	CCAAT	wsb2	TCCCCCCCCCTCCCAGCAGCCGATCAGCAGAACATTGG	0.84
ENSXETG00000003837	CCAAT	pura	AGGGGAGGGGGGTAAATTGG	0.83
ENSXETG00000010896	CCAAT	hsp90ab1	TCCCCCTCCCCCCCAGCAGCCAAT	0.84
ENSXETG00000014657	CCAAT	slc25a30	AGGGGGGGGGGGGGGTACTGAAAGATAAATTGG	0.85

ENSXETG00000013496	CCAAT	en2	TCTCTCTCTCTCTCTCTCTATATATATATATGTCCCATTG G	0.76
ENSXETG00000006259	CCAAT	gtf3c4	GAGGAGGAGGAGGCGGTGTTTCCATGGACCAAT	0.79
ENSXETG00000001478	CCAAT	fam78a	AGAGAGAGAGGGGGGGGTGATTGG	0.79
ENSXETG00000011510	CCAAT	pou3f1	AGGGGAGAGGAGAGCGATATTCCCAAT	0.79
ENSXETG00000015528	CCAAT	bcor	CCCCCTCCTCCTCCCAGCACCTCCTCCTTCTCCCAAT	0.81
ENSXETG00000021599	CCAAT	unkl	GGAGGAGAGGGAGGTGTGGCTACAAGACGGATTGG	0.79
ENSXETG00000023620	CCAAT	nkx1-2	CCCCCCCCCTCCACCAGCCCAAT	0.84
ENSXETG00000003951	CCAAT	fbxw5	GAGGGGGGGGGGCGCTGGGATCCGCGTGGCCCAAT	0.84
ENSXETG00000025969	CCAAT	XB-GENE- 5846549	AGAGGGGGGAGGATTGG	0.81
ENSXETG00000024103	CCAAT	alpk1	GGGGGGGGGAGGGGTATGCGGCATTGG	0.84
ENSXETG00000001322	CCAAT	syngr2	TCCCTCTCTCTCCGCTCCCCATTGG	0.78
ENSXETG00000013513	CCAAT	use1	TCCCCCCCCCCTTCTCCAGCCATGGTACGAGCCAAT	0.85
ENSXETG00000023739	CCAAT	wrb	GGGGGGGGGGGATAACAATAGATCCAAT	0.86
ENSXETG00000018417	CCAAT	sall1	CCTCCTCCCTCCCTCCCCGTGAGCCGGCCCCCTTATCAT TGG	0.81
ENSXETG00000004674	CCAAT	eef1d	CCCCCCCCCCCCCAAT	0.86
ENSXETG00000017744	CCAAT	42Sp50	CCCCCCCCCCCCCAAT	0.86
ENSXETG00000034222	CCAAT	unknown	AGGGGGGAGAGATTGG	0.81
ENSXETG00000022553	CCAAT	fam120a	CCCCCCCCCCCCCGGGTGACAGATTGG	0.86
ENSXETG00000003733	CCAAT	CEP55	CCCCTCCCCCCCACCCAAT	0.84
ENSXETG00000017315	CCAAT	adcy9	CCCCTCCTCCCCCACCACGCCACCAAT	0.83
ENSXETG00000030437	CCAAT	TNRC6A	GGGGGGGGGGGGGAAATGGAATTGG	0.86
ENSXETG00000027013	CCAAT	lcmt1	CTCCCCCCCCCAGCCCAAT	0.85

Table A-5.6. The List of experimentally verified previous five sequences (according Whitley *et al.*), using APTE code and the same predicted by our method for same genes and gene ID. **We have predicted the same sequence but with some extra nucleotides on 5'-side which are indicated by bold blue colour.**

Gene ID	Gene name	motif	Sequence From Original Paper	Sequence predicted From Our method
ENSXETG00000009337	gdi3	CCAAT	CCCCCCCCCCCCCAAT	TTCAGGTA CCCCCCCCCCCCCAAT
ENSXETG00000012462	gtf2e1.2	CCAAT	CTCCCCCCCCCCCCAAT	GGGCT CCCCCCCCCCCCCAAT
ENSXETG00000007609	thrsp	CCAAT	TCCCCCCCCCCCCAAT	ACCT TCCCCCCCCCCCCAAT
ENSXETG00000030719	unknown1	GGGCGG	GGGGGGGGGGGGCGG	AAGGGGGGGGGGGG CGG
ENSXETG00000002371	kif27	AGATA	GGGGGGGGGGGATCAGATA	GGGGGGGGGGGATCAGATA

References:

1. Kulkarni, M.; Mukherjee, A., Sequence dependent free energy profiles of localized B- to A-form transition of DNA in water. *J. Chem. Phys.* **2013**,*139*, 155102.
2. Mooers, B. H. M.; Schroth, G. P.; Baxter, W. W.; Ho, S. P., Alternating and non-alternating dG-dC hexanucleotides crystallize as canonical A-DNA. *J. Mol. Biol.* **1995**,*249*, 772-784.
3. McCall, M.; Brown, T.; Kennard, O., The crystal structure of d(G-G-G-G-C-C-C-C). A model for poly(dG).Poly(dC). *J. Mol. Biol.* **1985**,*183*, 385-96.
4. Fabian, H.; Holzer, W.; Heinemann, U.; Sklenar, H.; Welfle, H., Conformation of d(GGGATCCC)2 in crystals and in solution studied by X-ray diffraction, Raman spectroscopy and molecular modelling. *Nucleic Acids Res.* **1993**,*21*, 569-76.
5. Heinemann, U.; Lauble, H.; Frank, R.; Blöcker, H., Crystal structure analysis of an A-DNA fragment at 1.8 Å resolution: D(GCCCGGGC). *Nucleic Acids Res.* **1987**,*15*, 9531-9550.
6. Haran, T. E.; Shakked, Z.; Wang, A. H.; Rich, A., The crystal structure of d(CCCCGGGG): A new A-form variant with an extended backbone conformation. *J. Biomol. Struct. Dyn.* **1987**,*5*, 199-217.
7. Takusagawa, F., The crystal structure of d(GTACGTAC) at 2.25 Å resolution: Are the A-dna's always unwound approximately 10 degrees at the C-G steps? *J. Biomol. Struct. Dyn.* **1990**,*7*, 795-809.
8. Hunter, W. N.; D'Estaintot, B. L.; Kennard, O., Structural variation in d(CTCTAGAG). Implications for protein-DNA interactions. *Biochemistry-Us* **1989**,*28*, 2444-51.
9. Shakked, Z.; Guersteinguzikevich, G.; Eisenstein, M.; Frolow, F.; Rabinovich, D., The conformation of the DNA double helix in the crystal is dependent on its environment. *Nature* **1989**,*342*, 456-460.
10. Eisenstein, M.; Frolow, F.; Shakked, Z.; Rabinovich, D., The structure and hydration of the A-DNA fragment d(GGGTACCC) at room temperature and low temperature. *Nucleic Acids Res.* **1990**,*18*, 3185-3194.
11. Cervi, A. R.; Langlois d'Estaintot, B.; Hunter, W. N., Crystal and molecular structure of d(GTCTAGAC). *Acta Crystallogr B* **1992**,*48*, 714-719.

12. Frederick, C. A.; Quigley, G. J.; Teng, M. K.; Coll, M.; Van der Marel, G. A.; Van Boom, J. H.; Rich, A.; Wang, A. H., Molecular structure of an A-DNA decamer d(ACCGGCCGGT). *Eur. J. Biochem.* **1989**,*181*, 295-307.
13. Ramakrishnan, B.; Sundaralingam, M., High resolution crystal structure of the A-DNA decamer d(CCCGCCGGG). *J. Mol. Biol.* **1993**,*231*, 431-444.
14. Ramakrishnan, B.; Sundaralingam, M., Evidence for crystal environment dominating base sequence effects on DNA conformation - crystal-structures of the orthorhombic and hexagonal polymorphs of the A-DNA decamer d(GCGGGCCCCGC) and comparison with their isomorphous crystal-structures. *Biochemistry-Us* **1993**,*32*, 11458-11468.
15. Bingman, C. A.; Zon, G.; Sundaralingam, M., Crystal and molecular structure of the A-DNA dodecamer d(CCGTACGTACGG): Choice of fragment helical axis. *J. Mol. Biol.* **1992**,*227*, 738-756.
16. Bingman, C.; Jain, S.; Zon, G.; Sundaralingam, M., Crystal and molecular structure of the alternating dodecamer d(GCGTACGTACGC) in the A-DNA form: Comparison with the isomorphous non-alternating dodecamer d(CCGTACGTACGG). *Nucleic Acids Res.* **1992**,*20*, 6637-6647.
17. Nishimura, Y.; Torigoe, C.; Tsuboi, M., Salt induced B----A transition of poly(dG).Poly(dC) and the stabilization of a form by its methylation. *Nucleic Acids Res.* **1986**,*14*, 2737-48.
18. Grzeskowiak, K.; Yanagi, K.; Prive, G. G.; Dickerson, R. E., The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G. The effect of base pair reversals. *J. Biol. Chem.* **1991**,*266*, 8861-83.
19. Quintana, J. R.; Grzeskowiak, K.; Yanagi, K.; Dickerson, R. E., Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *J. Mol. Biol.* **1992**,*225*, 379-95.
20. Yuan, H.; Quintana, J.; Dickerson, R. E., Alternative structures for alternating poly(dA-dT) tracts: The structure of the B-DNA decamer C-G-A-T-A-T-A-T-C-G. *Biochemistry-Us* **1992**,*31*, 8009-21.
21. Goodsell, D. S.; Kopka, M. L.; Cascio, D.; Dickerson, R. E., Crystal structure of CATGGCCATG and its implications for A-tract bending models. *Proc Natl Acad Sci U S A* **1993**,*90*, 2930-4.

22. Goodsell, D. S.; Kaczorgrzeskowiak, M.; Dickerson, R. E., The crystal-structure of C-C-A-T-T-A-A-T-G-G - implications for bending of B-DNA at T-A steps. *J. Mol. Biol.* **1994**,239, 79-96.
23. Larsen, T. A.; Kopka, M. L.; Dickerson, R. E., Crystal structure analysis of the B-DNA dodecamer CGTGAATTCACG. *Biochemistry-Us* **1991**,30, 4443-9.
24. Shui, X.; Sines, C. C.; McFail-Isom, L.; VanDerveer, D.; Williams, L. D., Structure of the potassium form of CGCGAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry-Us* **1998**,37, 16877-87.
25. Yoon, C.; Prive, G. G.; Goodsell, D. S.; Dickerson, R. E., Structure of an alternating-B DNA helix and its relationship to A-tract DNA. *Proc Natl Acad Sci U S A* **1988**,85, 6332-6.
26. Tabernero, L.; Verdaguer, N.; Coll, M.; Fita, I.; van der Marel, G. A.; van Boom, J. H.; Rich, A.; Aymami, J., Molecular structure of the A-tract DNA dodecamer d(CGCAAATTTGCG) complexed with the minor groove binding drug netropsin. *Biochemistry-Us* **1993**,32, 8403-10.
27. Digabriele, A. D.; Sanderson, M. R.; Steitz, T. A., Crystal-lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. *Proc Natl Acad Sci U S A* **1989**,86, 1816-1820.
28. Chandrasekaran, R.; Radha, A.; Park, H.-S., Sodium ions and water molecules in the structure of poly(dA).Poly(dT). *Acta Crystallographica Section D* **1995**,51, 1025-1035.
29. Cheatham, T. E.; Kollman, P. A., Insight into the stabilization of A-DNA by specific ion association: Spontaneous B-DNA to A-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT]₂ in the presence of hexaamminecobalt(III). *Structure* **1997**,5, 1297-1311.
30. Wilcock, D. J.; Adams, A.; Cardin, C. J.; Wakelin, L. P., Structure of the DNA octanucleotide d(ACGTACGT)₂. *Acta crystallographica. Section D, Biological crystallography* **1996**,52, 481-5.
31. Clark, G. R.; Brown, D. G.; Sanderson, M. R.; Chwalinski, T.; Neidle, S.; Veal, J. M.; Jones, R. L.; Wilson, W. D.; Zon, G.; Garman, E., et al., Crystal and solution structures of the oligonucleotide d(ATGCGCAT)₂: A combined X-ray and NMR study. *Nucleic Acids Res.* **1990**,18, 5521-8.

32. Fernandez, L. G.; Subirana, J. A.; Verdaguer, N.; Pyshnyi, D.; Campos, L.; Malinina, L., Structural variability of A-DNA in crystals of the octamer d(pCpCpCpGpCpGpGpG). *J. Biomol. Struct. Dyn.* **1997**,*15*, 151-63.
33. Mayer-Jung, C.; Moras, D.; Timsit, Y., Hydration and recognition of methylated CpG steps in DNA. *EMBO J.* **1998**,*17*, 2709-18.
34. Tippin, D. B.; Sundaralingam, M., Nine polymorphic crystal structures of d(CCGGGCCGG), d(CCGGGCm⁵CGG), d(Cm⁵CGGGCCm⁵CGG) and d(CCGGGCC(Br)⁵CGG) in three different conformations: Effects of spermine binding and methylation on the bending and condensation of A-DNA. *J. Mol. Biol.* **1997**,*267*, 1171-85.
35. Malinina, L.; Fernandez, L. G.; Huynh-Dinh, T.; Subirana, J. A., Structure of the d(CGCCCGCGGGCG) dodecamer: A kinked A-DNA molecule showing some B-DNA features. *J. Mol. Biol.* **1999**,*285*, 1679-1690.
36. H.Raaijmakers, D. S., C.Mayer, Crystal structure analysis of the A-DNA dodecamer GACCACGTGGTC. **To be Published.**
37. Finley, J. B.; Luo, M., X-ray crystal structures of half the human papilloma virus E2 binding site: D(GACCGCGGTC). *Nucleic Acids Res.* **1998**,*26*, 5719-27.
38. Ban, C.; Sundaralingam, M., Crystal structure of the self-complementary 5'-purine start decamer d(GCACGCGTGC) in the A-DNA conformation. Ii. *Biophys. J.* **1996**,*71*, 1222-7.
39. McCall M Fau - Brown, T.; Brown T Fau - Hunter, W. N.; Hunter Wn Fau - Kennard, O.; Kennard, O., The crystal structure of d(GGATGGGAG): An essential part of the binding site for transcription factor Iiia.
40. Nunn, C. M.; Neidle, S., Structure of the A-DNA octamer d(GGCATGCC). *Acta crystallographica. Section D, Biological crystallography* **1997**,*53*, 269-73.
41. Kim, T.; Kwon, T. H.; Jung, H.; Ku, J. K.; Sundaralingam, M.; Ban, C., Crystal structures of the two isomorphous A-DNA decamers d(GTACGCGTAC) and d(GGCCGCGGCC). *Bull. Korean Chem. Soc.* **2006**,*27*, 568-572.
42. Wang, A. H. J.; Fujii, S.; Vanboom, J. H.; Rich, A., Molecular-structure of the octamer d(G-G-C-C-G-G-C-C) - modified A-DNA. *P Natl Acad Sci-Biol* **1982**,*79*, 3968-3972.
43. Urpi, L.; Navaza, J.; Subirana, J. A., Alternation of DNA and solvent layers in the A form of d(GGCGCC) obtained by ethanol crystallization. *Journal of Biomolecular Structure & Dynamics* **2000**,*18*, 363-369.

44. Savitha, G.; Viswamitra, M. A., An A-DNA structure with two independent duplexes in the asymmetric unit. *Acta crystallographica. Section D, Biological crystallography* **1999**,55, 1136-43.
45. Shakked, Z.; Rabinovich, D.; Cruse, W. B. T.; Egert, E.; Kennard, O.; Sala, G.; Salisbury, S. A.; Viswamitra, M. A., Crystalline A-DNA - the X-ray-analysis of the fragment d(G-G-T-A-T-A-C-C). *Proc R Soc Ser B-Bio* **1981**,213, 479-487.
46. Shakked Z Fau - Rabinovich, D.; Rabinovich D Fau - Cruse, W. B.; Cruse Wb Fau - Egert, E.; Egert E Fau - Kennard, O.; Kennard O Fau - Sala, G.; Sala G Fau - Salisbury, S. A.; Salisbury Sa Fau - Viswamitra, M. A.; Viswamitra, M. A., Crystalline A-dna: The X-ray analysis of the fragment d(G-G-T-A-T-A-C-C).
47. Nunn, C. M.; Neidle, S., The high resolution crystal structure of the DNA decamer d(AGGCATGCCT). *J. Mol. Biol.* **1996**,256, 340-51.
48. Gao, Y. G.; Robinson, H.; Wang, A. H., High-resolution a-DNA crystal structures of d(AGGGGCCCCT). An A-DNA model of poly(dG) x poly(dC). *Eur. J. Biochem.* **1999**,261, 413-20.
49. Tippin, D. B.; Sundaralingam, M., Structure of d(CCCTAGGG): Comparison with nine isomorphous octamer sequences reveals four distinct patterns of sequence-dependent intermolecular interactions. *Acta crystallographica. Section D, Biological crystallography* **1996**,52, 997-1003.
50. Kamitori, S.; Takusagawa, F., Multiple binding modes of anticancer drug actinomycin-D - X-ray, molecular modeling, and spectroscopic studies of d(GAAGCTTC)(2)-actinomycin-D complexes and its host DNA. *J. Am. Chem. Soc.* **1994**,116, 4154-4165.
51. Thota, N.; Li, X. H.; Bingman, C.; Sundaralingam, M., High-resolution refinement of the hexagonal A-DNA octamer d(GTGTACAC) at 1.4 angstrom. *Acta Crystallogr D* **1993**,49, 282-291.
52. Rozenberg, H.; Rabinovich, D.; Frolov, F.; Hegde, R. S.; Shakked, Z., Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc Natl Acad Sci U S A* **1998**,95, 15194-9.
53. Timsit, Y.; Moras, D., DNA self-fitting: The double helix directs the geometry of its supramolecular assembly. *EMBO J.* **1994**,13, 2737-46.
54. Nunn, C. M.; Garman, E.; Neidle, S., Crystal structure of the DNA decamer d(CGCAATTGCG) complexed with the minor groove binding drug netropsin. *Biochemistry-Us* **1997**,36, 4792-9.

55. Coll, M.; Aymami, J.; Vandermarel, G. A.; Vanboom, J. H.; Rich, A.; Wang, A. H. J., Molecular-structure of the netropsin-d(CGCGATATCGCG) complex - DNA conformation in an alternating at segment. *Biochemistry-Us* **1989**,28, 310-320.
56. Tari, L. W.; Secco, A. S., Base-pair opening and spermine binding - B-DNA features displayed in the crystal-structure of a GAL operon fragment - implications for protein-DNA recognition. *Nucleic Acids Res.* **1995**,23, 2065-2073.
57. Wahl, M. C.; Rao, S. T.; Sundaralingam, M., Crystal structure of the B-DNA hexamer d(CTCGAG): Model for an A-to-B transition. *Biophys. J.* **1996**,70, 2857-66.
58. Goodsell, D. S.; Grzeskowiak, K.; Dickerson, R. E., Crystal structure of C-T-C-T-C-G-A-G-A-G. Implications for the structure of the holliday junction. *Biochemistry-Us* **1995**,34, 1022-9.
59. Soler-Lopez, M.; Malinina, L.; Liu, J.; Huynh-Dinh, T.; Subirana, J. A., Water and ions in a high resolution structure of B-DNA. *J. Biol. Chem.* **1999**,274, 23683-6.
60. B.H.M.Mooers, P. S. H., Crystal structure of the b-DNA nonamer d(GCGTACGCG) with a novel d[G*(g.C)] base-triplet involving the minor groove. **1996 (To be Published).**
61. Vlieghe, D.; Turkenburg, J. P.; Van Meervelt, L., B-DNA at atomic resolution reveals extended hydration patterns. *Acta Crystallogr D* **1999**,55, 1495-1502.
62. Minasov, G.; Tereshko, V.; Egli, M., Atomic-resolution crystal structures of B-DNA reveal specific influences of divalent metal ions on conformation and packing. *J. Mol. Biol.* **1999**,291, 83-99.
63. Han, G. W.; Kopka, M. L.; Cascio, D.; Grzeskowiak, K.; Dickerson, R. E., Structure of a DNA analog of the primer for HIV-1 RT second strand synthesis. *J. Mol. Biol.* **1997**,269, 811-26.
64. Prive, G. G.; Yanagi, K.; Dickerson, R. E., Structure of the b-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.* **1991**,217, 177-99.
65. Grzeskowiak, K.; Goodsell, D. S.; Kaczor-Grzeskowiak, M.; Cascio, D.; Dickerson, R. E., Crystallographic analysis of C-C-A-A-G-C-T-T-G-G and its implications for bending in B-DNA. *Biochemistry-Us* **1993**,32, 8923-31.
66. Shakked, Z.; Guzikevich-Guerstein, G.; Frolow, F.; Rabinovich, D.; Joachimiak, A.; Sigler, P. B., Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature* **1994**,368, 469-73.

67. Lipscomb, L. A.; Peek, M. E.; Morningstar, M. L.; Verghis, S. M.; Miller, E. M.; Rich, A.; Essigmann, J. M.; Williams, L. D., X-ray structure of a DNA decamer containing 7,8-dihydro-8-oxoguanine. *Proc Natl Acad Sci U S A* **1995**,92, 719-23.
68. Heinemann, U.; Alings, C., Crystallographic study of one turn of G/C-rich B-DNA. *J. Mol. Biol.* **1989**,210, 369-81.
69. Kielkopf, C. L.; Ding, S.; Kuhn, P.; Rees, D. C., Conformational flexibility of B-DNA at 0.74 Å resolution: d(CCAGTACTGG)₂. *J. Mol. Biol.* **2000**,296, 787-801.
70. Eichman, B. F.; Mooers, B. H. M.; Alberti, M.; Hearst, J. E.; Ho, P. S., The crystal structures of psoralen cross-linked DNAs: Drug-dependent formation of holliday junctions. *J. Mol. Biol.* **2001**,308, 15-26.
71. Mack, D. R.; Chiu, T. K.; Dickerson, R. E., Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: A comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.* **2001**,312, 1037-1049.
72. Balendiran, K.; Rao, S. T.; Sekharudu, C. Y.; Zon, G.; Sundaralingam, M., X-ray structures of the B-DNA dodecamer d(CGCGTTAACGCG) with an inverted central tetranucleotide and its netropsin complex. *Acta crystallographica. Section D, Biological crystallography* **1995**,51, 190-8.
73. Tereshko, V.; Urpi, L.; Malinina, L.; Huynh-Dinh, T.; Subirana, J. A., Structure of the B-DNA oligomers d(CGCTAGCG) and d(CGCTCTAGAGCG) in new crystal forms. *Biochemistry-Us* **1996**,35, 11589-95.