

# **NON-CANONICAL BASE PAIRS IN RNA AND THEIR INVOLVEMENT IN DOUBLE HELIX FORMATION**

by  
**Satyabrata Maiti**

**Enrolment No.- LIFE05201504004**  
**Saha Institute of Nuclear Physics, Kolkata**

*A thesis submitted to the*  
*Board of Studies in Life Sciences*  
*In partial fulfillment of requirements*  
*for the Degree of*

**DOCTOR OF PHILOSOPHY**

*of*

**HOMI BHABHA NATIONAL INSTITUTE**



**December, 2020**

# Homi Bhabha National Institute

## Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by **Satyabrata Maiti** (HBNI Enrolment No. LIFE05201504004) entitled "**Non-canonical Base Pairs in RNA and their involvement in double helix formation**" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairman - Prof. Rahul Banerjee

*R Banerjee* 9/3/2021

Guide / Convener - Prof. Dhananjay Bhattacharyya

*Dhananjay Bhattacharyya*  
9/3/2021

Co-guide - Prof. Montu Hazra

*M Hazra* 09/03/2021

Examiner - Prof. Pradipta Bandyopadhyay, JNU, India

*Pradipta Bandyopadhyay*

09.03.21

Member 1- Prof. Subhabrata Majumder

*Subhabrata Majumder* 09/03/21

Member 2- Prof. Dulal Senapati

*D Senapati*  
09.03.21

Member 3- Prof. Ansuman Lahiri, University of Calcutta

*Ansuman Lahiri*  
9/3/21

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I/We hereby certify that I/we have read this thesis prepared under my/our direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 9/3/2021

Place: Kolkata

*M Hazra*  
09/03/2021

Signature

Co-guide (if any)

*Dhananjay Bhattacharyya*

Signature

Guide

9/3/2021

# Homi Bhabha National Institute

## Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by **Satyabrata Maiti** (HBNI Enrolment No. LIFE05201504004) entitled "**Non-canonical Base Pairs in RNA and their involvement in double helix formation**" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairman - Prof. Rahul Banerjee

*R Banerjee* 9/3/2021

Guide / Convener - Prof. Dhananjay Bhattacharyya

*Dhananjay Bhattacharyya*  
9/3/2021

Co-guide - Prof. Montu Hazra

*M Hazra* 09/03/2021

Examiner - Prof. Pradipta Bandyopadhyay, JNU, India

*Pradipta Bandyopadhyay*

09.03.21

Member 1- Prof. Subhabrata Majumder

*Subhabrata Majumder* 09/03/21

Member 2- Prof. Dulal Senapati

*Dulal Senapati*  
09.03.21

Member 3- Prof. Ansuman Lahiri, University of Calcutta

*Ansuman Lahiri* 9/3/21

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I/We hereby certify that I/we have read this thesis prepared under my/our direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 9/3/2021

*M Hazra*  
09/03/2021

Place: Kolkata

Signature

Co-guide (if any)

*Dhananjay Bhattacharyya*

Signature

Guide

9/3/2021

## STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

*Satyabrata Maiti*

-----  
Satyabrata Maiti

## DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.

*Satyabrata Maiti*

.....

Satyabrata Maiti

## List of Publications arising from the thesis

### Peer reviewed journals:

- i) **Maiti, S.;** Bhattacharyya, D. Stacking Interactions Involving Non-Watson–Crick Basepairs: Dispersion Corrected Density Functional Theory Studies. *Phys. Chem. Chem. Phys.* **2017**, *19* (42), 28718–28730
- ii) **Maiti, S.;** Mukherjee, D.; Roy, P.; Chakrabarti, J.; Bhattacharyya, D. Stacking Geometry between Two Sheared Watson-Crick Basepairs: Computational Chemistry and Bioinformatics Based Prediction. *Biochim. Biophys. Acta - Gen. Subj.* **2020**, *1864* (7), 129600.

Satyabrata Maiti

.....  
Satyabrata Maiti

**Dedicated To My**  
**“Family”**

## ACKNOWLEDGMENTS

*Yes, I am lucky enough to have excellent teachers from my School days to Research Institute, SINP.*

*Firstly I would like to express my gratitude to my supervisor, Prof. Dhananjay Bhattacharyya. His friendly supervision, expert suggestion, patience have been priceless in the journey of my research career. He also worked alongside me with so much soberness that makes my Ph.D. journey utterly beautiful.*

*I am thankful to the ex-director Prof. Ajit Kumar Mohanty and the current director Prof. Gautam Bhattacharya, of Saha Institute of Nuclear Physics, for giving me the beautiful facilities needed for my research. I am also thankful to our Dean, Prof. Partha Saha who always conveys important messages to us.*

*I am also thankful to my teachers for the Post-MSc course, especially Prof. Mantu Hazra, Prof. Samita Basu, Prof. Munna Sarkar, and Prof. Rahul Banerjee for their friendly assistance and expert advice from the beginning of SINP days. I am also thankful to my collaborator Prof. Jaydev Chakrabarti, S.N.Bose National Center, Kolkata. He gave so much excellent advice on Statistical Mechanics. I am also thankful to Dr. Gouriprasanna Roy and Dr. Subhra Ghosh Dastidar for giving me the opportunity to do quantum chemical calculation and MD simulation before joining SINP. I am also thankful to my teachers in my college days, especially Achintya Kumar Sarkar, Prof. Dipak Kumar Mandal, and Prof. Sujay Chackraborty.*

*I am fortunate to have seniors like Dr. Debasish Mukherjee, whose support, academic, and non-academic help from the beginning of my LAB make me deeply indebted to him. He first showed me the beauty of coding. I am truly fortunate to have Subhabrata Maiti as my elder brother for his*

*immense help towards learning Chemistry from the beginning of my academic career. I am grateful to my lab-senior, Sanchita Di for his excellent suggestion for my first publication. Her kindness and mental support are unforgettable. I would also like to thank my other lab-seniors, Angana Di, Kakali Di, Dr. Lakshmi Maganti, for their scientific suggestions. I am also thankful to my colleagues and seniors, especially Sayak, Tulika, Kuntal Da, Kamalendu Da, Bibhuti, Sandip, Sabyasachi Karmakar, and Sabyasachi Sen who provide me such a wonderful atmosphere in SINP that results never to have bad days here. I am blessed to have childhood friends especially, Tony, Anubhab, Kaustav, Chayan, and as well as college friends, Priyangana, Srijita, Mainak, Arghya, Srobona, Chandramouli, Abhijit, Sayantan, Tanmoy, and Debjit who always stand by me when no one else does. Scientific discussion with Saikat, Sourav Da, Tanaya, Souvik and Atanu Da helped me a lot to shape some work. I am also grateful to Bidyut Saha for his excellent guidance on learning programming.*

*My parents, elder sister, sister-in-law have always showered unconditional love and support, making me keep going in the most challenging times.*

*I would like to express my deepest gratitude to my wife, Priyanka Bhadra, who is always supportive and makes me challenging at a hard time. I would also like to thank her parents for their supportive nature towards my love of research.*

## Table of Contents

Summary.....	x
List of Figures.....	xi
List of Tables.....	xxi
Chapter- 1 Nucleic Acids: Structural Properties and Function.....	1
<b>1.1 Introduction:</b> .....	2
<b>1.2 Chemical Constituent of Nucleic Acid</b> .....	3
<i>1.2.1 Nitrogenous Bases</i> .....	3
<i>1.2.2 Sugar Phosphate backbone</i> .....	5
<i>1.2.3 Sugar conformation:</i> .....	7
<b>1.3 Canonical and Non-canonical Base pair</b> .....	10
<b>1.4 Orientation parameter in the base pair and Base pair Step</b> .....	10
<b>1.5 Base pair Nomenclature, and Local strand direction:</b> .....	12
<b>1.6 Identification of Non-canonical base pair</b> .....	16
<b>1.7 Nucleic acid Structural Polymorphism</b> .....	19
<b>1.8 Major and Minor Groove</b> .....	21
<b>1.9 Non-canonical Base-pairs in Double Helical Regions</b> .....	22
<b>1.10 Hydrogen bonding and Stacking Energy</b> .....	25
Chapter- 2 Non-canonical Base Pairs and their stabilities.....	29
<b>2.1 Introduction</b> .....	30
<b>2.2 Methods</b> .....	37
<b>2.3 Result and Discussion</b> .....	40
<i>2.3.1 Base pair in the database</i> .....	40
<i>2.3.2 Dinucleotide Step Analysis</i> .....	49
<b>2.4 Molecular Modeling of RNA double helix</b> .....	55
<b>2.5 Conclusion</b> .....	62
Chapter- 3 Stacking Interaction by most Frequent A:G base pair: Establishment of Hybrid DFT-D method.....	63
<b>3.1 Introduction</b> .....	64

<b>3.2 Material and Methods</b> .....	68
3.2.1 Crystal Structure Database Analysis:.....	68
3.2.2 Base pair and Base pair Step Modeling.....	69
3.2.3 Computational Methods.....	70
3.2.4 Hybrid Modeling.....	71
<b>3.3 Result and Discussion</b> .....	73
3.3.1 Stability of the base pairs: .....	73
3.3.2 Evaluation of the Stacking Models.....	76
3.3.3 Stacking Energy Analysis.....	79
3.3.4 Hybrid Stacking Energy Analysis .....	95
<b>3.4 Conclusion</b> .....	105
Chapter- 4 Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT-D and MD Studies .....	106
<b>4.1 Introduction</b> .....	107
<b>4.2 Materials and Methods:</b> .....	112
<b>4.3 Result and Discussion:</b> .....	118
<b>4.4 Conclusion:</b> .....	167
Chapter- 5 Molecular Modeling of SL1-helix of SARS-Corona Viral RNA .....	169
<b>5.1 Introduction</b> .....	170
<b>5.2 Methods</b> .....	175
<b>5.3 Result and Discussion</b> .....	176
5.3.1 Structure generation for Covid-19 RNA .....	180
5.3.2 Hybrid Stacking Energy Scan .....	183
<b>5.4 Conclusion</b> .....	188
Chapter- 6 Conclusion and Scope of Future .....	190
<b>6.1 Conclusion and Scope of future work (Chapter 2)</b> .....	191
<b>6.2 Conclusion and Scope of future work (Chapter 3)</b> .....	191
<b>6.3 Conclusion and Scope of future work (Chapter 4)</b> .....	192
<b>6.4 Conclusion and Scope of future work (Chapter 5)</b> .....	192
Chapter- 7 References.....	193

## List of Figures

<b>Figure 1-1</b> Molecular structure of nucleobase .....	4
<b>Figure 1-2</b> Nucleoside and Nucleotide as structural unit of nucleic acid .....	4
<b>Figure 1-3</b> Formation of phosphodiester bond.....	5
<b>Figure 1-4</b> Representation of Nucleic Acid (A) backbone torsion angle (B) pseudo torsion angle .....	6
<b>Figure 1-5</b> Two major Sugar Puckering in nucleic acid. ....	8
<b>Figure 1-6</b> syn and anti conformations of nucleosides .....	9
<b>Figure 1-7</b> A:T Hoogsteen base pair .....	10
<b>Figure 1-8</b> Standard nomenclatures of Base pair and Base pair Step parameter. ....	11
<b>Figure 1-9</b> (A) Base pairing edges for the nucleotides (B) Cis and Trans orientations of the two nucleotides and the glycosidic bonds are represented by the arrow as vectors).....	13
<b>Figure 1-10</b> Representation of possible base pairing .....	14
<b>Figure 1-11</b> (A) Secondary structure of t-RNA <sup>Phe</sup> generated by BPFIND (B) The L-shaped X-ray crystal structure of t-RNA <sup>Phe</sup> (PDB ID:1EHZ) and ball-stick model represents base triplet (Residue ID:9,12 and 23). ....	16
<b>Figure 1-12</b> Description of hydrogen bonding in a non-canonical base pair with their precursors (as used by BPFIND algorithm) .....	19
<b>Figure 1-13</b> Illustration of Zp of dinucleotide steps of A-DNA and B-DNA. Images taken from Olson et.al <sup>52</sup> .....	20
<b>Figure 1-14</b> Atomic and cartoon representation of major and minor groove .....	22

**Figure 1-15** (a) schematic and (b) by 3-D representation of base triples with non-canonocal base pairs from PDB ID 1KOG. .... 24

**Figure 1-16** (a) schematic and (b) by 3-D representation of T-loop from from PDB ID 1U96 where residues 105 and 10 make base triples with Watson-Crick base pair of 61:84. .... 24

**Figure 2-1** Base pair having residues 2868 (Chain A) and 2889 (Chain A) of 1N8R.pdb identified by ClaRNA but not by BPFIND. It shows the possibility of only one H-bond. .... 33

**Figure 2-2** Base pair having residues 608:R and 621:R of 1ASY.pdb identified by BPFIND. It shows H-bonding through 2'-OH of sugar. .... 34

**Figure 2-3** The home page of RNABPDB database, showing how to navigate from one service to another. The drop-down menu indicating other available applications appear by clicking Menu (shown by border) and the drop-down menu indicating other non-canonical base pairing patterns appear by clicking the region shown by border. Any base pair, such as A:A W:WC, can be clicked to get detail information about the base pair, its stabilizing energy, distributions of structural parameters and how it stacks on other base pairs. The base pairs shown by gray patch are impossible ones. .... 40

**Figure 2-4** Structures of two alternative forms of Cyt:Cyt base pairs involving their Watson-Crick base pairing edges in trans orientation for (a) in neutral form stabilized by two N—H...N hydrogen bonds and (b) in hemi protonated form stabilized by three hydrogen bonds. .... 41

**Figure 2-5** Histogram for change in X—H covalent bond length after H-bonding, X being primary amino group (NH—H), secondary amino group (N—H), hydroxyl group (O—H) and weak non-polar group (C—H). .... 47

**Figure 2-6** Correlation between non protonated bae pair interaction energy and their frequency ..... 49

<b>Figure 2-7</b> Representative figure showing distributions of different parameters indicating good stacking between a C:G Watson-Crick base pair and a G:U Wobble base pair. ....	51
<b>Figure 2-8</b> Correlation between stacking interaction and stacking overlap for all dinucleotide step sequence having frequency more than 5. ....	53
<b>Figure 2-9</b> Representative three dimensional view of dinucleotide platforms .....	55
<b>Figure 2-10</b> Aligned Regenerated Structure on experimental crystal structure. Ball and Stick representation indicated non-Canonical base pair. ....	58
<b>Figure 3-1</b> Possible dinucleotide step containing C:G W:WC and G:A S:HT. ....	64
<b>Figure 3-2</b> Stacking overlap area contours at different Twist value for A:U W:WC::C G:W WC are shown here. Color bar is given in Å <sup>2</sup> . Contour lines are 2.5 Å <sup>2</sup> apart. ....	78
<b>Figure 3-3</b> Model Structures for 20° Twist, -20° Roll and 1Å Slide having maximum overlap area for (a) G:C W:WC::G:A S:HT and (b) C:G W:WC::G:A S:HT are shown here indicating the short inter atomic distances. ....	79
<b>Figure 3-4</b> Intrinsic stacking energy contours considering DFT-D energy at different Twist values for A:U W:WC::C:G W:WC dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart. ....	82
<b>Figure 3-5</b> Intrinsic stacking energy contours considering DFT-D energy at different Twist values for G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart. ....	83
<b>Figure 3-6</b> Intrinsic stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart. ....	84

**Figure 3-7** Model structures of (a) G:C:W:WC::G:A S:HT for  $-20^\circ$  Roll and  $1.5\text{\AA}$  Slide and (b) C:G W:WC::G:A S:HT for  $10^\circ$  Roll and  $-1.0\text{\AA}$  Slide in their best stacking energy configurations at  $85^\circ$  Twist. The large bond lengths in sugar-phosphate group are also shown..... 88

**Figure 3-8** Intrinsic intra-strand stacking energy contours, i.e interaction energy between the two Gua residues and those between the Cyt and Ade residues (see **Figure 1** for detail), considering DFT-D energy at different Twist values for G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 91

**Figure 3-9** Intrinsic intra-strand stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 92

**Figure 3-10** Intrinsic inter strand stacking energy contours considering DFT-D energy at different Twist values for inter strand of G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 93

**Figure 3-11** Intrinsic inter strand stacking energy contours considering DFT-D energy at different Twist values for inter strand of C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 95

**Figure 3-12** Schematic representation showing that negative Slide gives better cross strand overlap and stacking interaction between two Gua bases..... 95

**Figure 3-13** Intrinsic stacking energy contours for (a, b,c,d and e) G:C W:WC::G:A S:HT and (f, g, h, i and j) C:G W:W C::G:A S:HT dinucleotide steps considering DFT-D energies and coarse grain penalty energy values at various Twist values. Color bar is presented in kcal/mol. Difference between two adjacent contour lines is 1 kcal/mol. Roll and Slide values from crystal structure

database are marked by black point. The minimum energy conformations for every Twist values are represented by red dots. .... 97

**Figure 3-14** Intrinsic stacking energy contours for A:U W:WC :: C:G W:WC dinucleotide steps considering DFT-D energies and coarse grain penalty energy values at various Twist values. Color bar is presented in kcal/mol. Difference between two adjacent contour lines lines is 1 kcal/mol. Roll and Slide values from crystal structure database are marked by black point. The minimum energy conformations for every Twist values are represented by red dots..... 98

**Figure 3-15** Stacking energy contours considering DFT-D and coarse grain penalty energy at different Twist values for G:C W:WC :: G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 102

**Figure 3-16** Stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC :: G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart..... 103

**Figure 4-1** A:A w:wC base pair forming triplet with another Ade in 16S ribosomal RNA (PDB ID. 5J5B) structure of E. coli..... 108

**Figure 4-2** (a) Negative and Positive shear configurations of A:A w:wC base pair becomes superposable upon 180° rotation about the pseudo-dyad symmetry axis shown. (b) Interaction energy of A:A w:wC base pair for different Shear values in the range between -3.00Å and +3.00Å. (c) Distribution of Shear value for the A:A w:wC base pair obtained from X-ray Crystal structure. Mean Shear for these structures is 0.37Å with 2.54 being the standard deviation. .... 110

**Figure 4-3** Representative Structures of A:A w:wC, G:U W:WC and A:U W:WC..... 118

**Figure 4-4** Distribution of stacking overlap in different system as mentioned in the figure. .... 128

**Figure 4-5** Time evolutions of Shear for A:A base pair in RNA double helices for different sequences: GAG motif for (A)  $-2.7\text{ \AA}$  and (B)  $+2.7\text{ \AA}$  Shear in initial models of the A:A w:wC base pair, UAG motif for (C)  $-2.7\text{ \AA}$  and (D)  $+2.7\text{ \AA}$  Shear in initial models of A:A w:wC (calculated by NUPARM considering s:hT base pairing pattern), GAU motif for (E)  $-2.7\text{ \AA}$  and (F)  $+2.7\text{ \AA}$  Shear in initial models of A:A w:wC..... 129

**Figure 4-6** base pairing lifetime histogram for A:A base pair in the RNA double helices for the three sequences namely GAG motif, UAG motif and GAU motif with initial  $-2.7\text{ \AA}$  and  $+2.7\text{ \AA}$  Shear in the initial structure structures of the A:A w:wC base pairs. Representative figures of A:A s:hT, A:A S:WC and A:A w:wC are shown. .... 130

**Figure 4-7** A:A base pair attains s:hT type base pair orientation in  $5'-(\text{UAG}).5'-(\text{UAG})$  motif during MD simulation of the sequence  $5' - \text{CGAUAGAGCG} - 3'$ .  
 $3' - \text{GCUGAUUCGC} - 5'$  ..... 131

**Figure 4-8** Conversion of A:A w:wC to A:A s:hT through Adenine Zipper (A:A stacked geometry) in  $5'-(\text{UAG}).5'-(\text{UAG})$  motif during MD simulation of the sequence  $5' - \text{CGAUAGAGCG} - 3'$ . The  
 $3' - \text{GCUGAUUCGC} - 5'$

A:A base pair is shown as Cartoon ring mode and G:U base pair at upstream and downstream of A:A are shown as ball and stick and remaining base pairs are shown by ladder. .... 132

**Figure 4-9** Free Energy landscape from MD simulations in Open-Shear hyperspace for the three sequence motifs with either Shear value in their initial structures. The color bar on the Figure for GAU motif with positive Shear in initial model, indicate Free Energy values (in kcal/mol) for all the systems. .... 134

**Figure 4-10** Time evolution of Shear for several base pair in GAU motif sequence during 500ns MD simulations for(A) negative Shear in initial model of the A:A w:wC base pair and (B) positive Shear in initial model of the same..... 137

**Figure 4-11** Time evolution of Open angle for several base pair in the GAU motif sequence during 500ns MD simulations with (A) negative Shear in initial model of A:A w:wC base pair and (B) positive Shear in initial model of A:A w:wC base pair. .... 139

**Figure 4-12** Time evolution of Twist angle for several base pair step in GAU motif sequence during 500ns MD simulations for (A) negative Shear in initial model of A:A w:wC base pair and (B) positive Shear in initial model of A:A w:wC base pair. .... 141

**Figure 4-13** Hybrid stacking iso-energy contours of A:A w:wC::C:G W:WC dinucleotide step sequence, with (A) negative Shear and (B) positive Shear of A:A w:wC base pair, considering MP2/cc-pVDZ and coarse-grain energy penalty for different force constant values.. Energy difference between two adjacent contour lines is 1kcal/mol. The best values of Roll and Slide for stacking are indicated by red dot. .... 143

**Figure 4-14** (A) Hybrid Stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence, (A) with negative Shear of A:A w:wC and (B) with positive Shear of A:A w:wC, considering DFT-D energy and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist values is indicated by red dot. .... 145

**Figure 4-15** Time evolution of RMSD of RNA double helix with only Watson-Crick base pairs from hypothetical B-RNA conformation with respect to initial structure, (B) RMSD of the same with respect to modeled A-RNA having same sequence considering Arnotts' fiber model. .... 146

**Figure 4-16** Time evolution Slide for different base paired dinucleotide steps of the double helix of RNA with only Watson-Crick base pairs. .... 148

**Figure 4-17** Time evolution of Zp for different base pair steps of RNA double helix with only Watson-Crick base pairs. .... 149

**Figure 4-18** Time evolution of sugar-pucker pseudo-rotation phase angle for different residues of RNA double helix with Watson-Crick base pairs..... 150

**Figure 4-19** Distribution of Shear value for the A:A w:wC base pair in A:A w:wC ::C:G W:WC dinucleotide step obtained from X-ray Crystal structure..... 151

**Figure 4-20** Intrinsic stacking energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence considering DFT-D energy at different Twist values for structures having (a-c) negative Shear value of A:A w:wC and (d-f) positive Shear value of A:A w:wC at  $\omega$ B97X-D/cc- pVDZ, M06-2X/cc-pVDZ and MP2/cc-pVDZ levels of theory, respectively. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist value is indicated by red dot. .... 153

**Figure 4-21** Structures of A:A w:wC::G:U W:WC base paired dinucleotide steps having best hybrid stacking energy with (A) negative Shear for A:A step and (B) positive Shear for the same step generated by PyMol125, indicating possible hydrogen bonds. The virtual bonds mimicking sugar- phosphate backbone are shown by Red sticks. .... 155

**Figure 4-22** Hybrid stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence, with (A) negative Shear and (B) positive Shear of A:A w:wC base pair, considering energy using MP2/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each virtual bond stretching force-constant value is indicated by red dot..... 160

**Figure 4-23** Hybrid Stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence considering water environment by using CPCM, (A) and (B) with  $\omega$ B97X-D , (C) and (D) with M06-2X and (E) and (F) with MP2 functional for mentioned Shear values of A:A w:wC, considering DFT-D energy and coarse-grain energy penalty for different force constant values.

Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist values is indicated by red dot. .... 162

**Figure 4-24** Contour plot of A:A w:wC::G:U W:WC showing difference between stacking overlap from positive Shear to Negative Shear of A:A w:wC base pair. .... 163

**Figure 4-25** Hybrid stacking energy contour of A:A w:wC :: U:G W:WC having (A) negative Shear of A:A w:wC base pair and (B) positive Shear of A:A w:wC base pair, at 30° Twist for 9.242 kcal/mol/Å<sup>2</sup> force constant representing backbone effect. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each plot is indicated by red dot. .... 164

**Figure 4-26** Schematic diagram of the two base pair dinucleotide steps highlighting two opposing forces exerted by U6 to A5 and U4 to A5 making the A2:A5 base pair as frustrated..... 166

**Figure 5-1.** 5' UTR region of RNA of Human Corona Virus ..... 170

**Figure 5-2** Three dimensional representations for protonated and non-protonated base pair. .. 172

**Figure 5-3** C:C W:+T base pair. .... 173

**Figure 5-4** Proposed model for SL1 Helix of RNA of SARS corona virus..... 175

**Figure 5-5** Reaction coordinate for the protonation of cytosine and Adenine mediated by lysine and Water molecule. .... 180

**Figure 5-6** Pictorial description of different U:U W:WC base pairs where major change in base pair is in their Shear motion..... 181

**Figure 5-7** Hybrid stacking isoenergy contours of (A) G:C W:WC :: A:C +:WC and (B) G:C W:WC :: A:C w:wC dinucleotide step sequence considering ωB97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. .... 185

**Figure 5-8** Hybrid stacking isoenergy contours of (A) A:C +:WC :: U:C W:WC and (B) A:C w:wC :: U:C W:+C dinucleotide step sequence considering  $\omega$ B97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. .... 186

**Figure 5-9** Hybrid stacking iso-energy contours of (A-B) U:C W:WC::U:U W:WC and (C-D) U:U W:WC::U:A W:WC dinucleotide step sequence considering  $\omega$ B97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. .... 187

## List of Tables

<b>Table 1-1</b> Definition for the Sugar puckering dihedrals and backbone dihedrals.....	8
<b>Table 1-2</b> Different types of base pairing schemes and associated local strand orientations of their sugar-phosphate backbone .....	14
<b>Table 1-3</b> Nomenclature system in MC-Annotate .....	17
<b>Table 1-4</b> Average values of Base pair and Base pair step parameter for A-form and B-form nucleic acid. Data is taken from olson et al <sup>53</sup> . Standard deviations are given in parenthesis. ....	21
<b>Table 1-5</b> Atomic distribution of nucleotides in the minor and major groove.....	21
<b>Table 2-1</b> List of NMR derived structures of RNA used to populate the database of canonical and non-canonical base pairs .....	35
<b>Table 2-2</b> Change in bond length for NH—H, N—H, C—H and O—H after participation in H-bonding. Optimized Bond distance in isolated free nucleotide bases: $d(\text{NH—H})_{\text{Adenine}}=1.011 \text{ \AA}$ , $d(\text{NH—H})_{\text{Guanine}}=1.011 \text{ \AA}$ , $d(\text{N—H})_{\text{Guanine/Uracil}}=1.015 \text{ \AA}$ , $d(\text{C—H})=1.094 \text{ \AA}$ , $d(\text{O—H})=0.963 \text{ \AA}$ .....	43
<b>Table 2-3</b> BSSE corrected Interaction energy and Stacking Energy for some of the important dinucleotide steps.....	52
<b>Table 2-4</b> Comparison of regenerated RNA double-helical structures with the experimentally determined structures.....	58
<b>Table 3-1</b> Mean C1'...C1' distances, calculated from X-ray crystallographic database, along with their standard deviations (in parenthesis), for the two strands of the two sequences and the calculated force-constants values for stretching of the pseudo-bonds. See <b>Figure 3-1</b> . ....	73

**Table 3-2** Geometric and energetic parameters of three base pairs obtained by different methods of modeling and optimization. .... 75

**Table 3-3** Mean and standard deviation (in parenthesis) values of base step parameters and stacking overlap values from the available crystal structures in the base paired dinucleotide steps are shown here. We did not observe any data for two sequences in the experiment. .... 77

**Table 3-4** Best stacking energy (DFT-D) and dispersion energy components in configuration having maximum stacking overlap area for all Twist values for the dinucleotide step sequences are shown below. .... 80

**Table 3-5** Geometrical and energy parameters for the structures of the two sequences having minimum DFT-D energy corresponding to each Twist value are listed. Mean values of C1'...C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns. .... 86

**Table 3-6** Geometrical and energy parameters for the structures of the dinucleotide step sequence having minimum energy (sum of DFT-D and coarse-grain energies) corresponding to each Twist value are listed. Mean values of C1'...C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns. .... 97

**Table 3-7** Geometrical and energy parameters for the structures of the two sequences having minimum energy (sum of DFT-D and coarse-grain energies) corresponding to each Twist value are listed. Mean values of C1'...C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns. .... 99

**Table 4-1** Description base pair and Base pair Step parameter for modelling of base pair steps. .... 113

**Table 4-2** Sequences of the double helices studied using molecular dynamics simulation. .... 114

<b>Table 4-3</b> (A) Base pair and (B) Base pair Step parameters for different sequence during RNA double helix generation by RNAHelix software.....	115
<b>Table 4-4</b> Base pair orientation parameter, E-value and quantum chemical Interaction energy (in kcal/mol) of A:A w:wC base pairs from (A) X-ray Crystal structures* (B) model structure for dinucleotide step sequence A:A w:wC :: G:U W:WC.....	121
<b>Table 4-5</b> Geometric and energetic parameter of step A:A w:wC::G:U W:WC base paired dinucleotide step from available X-ray crystal structures.....	123
<b>Table 4-6</b> Base pair parameters for A:A w:wC found in A:A w:wC:: U:G W:WC dinucleotide step sequence.....	124
<b>Table 4-7</b> Mean and standard deviation (in Parenthesis) of some important base pair and base pair step parameter for mentioned RNA double helix. ....	125
<b>Table 4-8</b> Geometrical and energetic parameter* of dinucleotide step sequence A:A w:wC :: G:U W:WC having (A) negative Shear value (B) positive Shear value of A:A w:wC base pair for different Twist values having minimum DFT-D energy obtained my different levels of theory. ....	153
<b>Table 4-9</b> Geometrical and energetic parameter of dinucleotide step sequence A:A w:wC :: G:U W:WC having (A) negative Shear and (B) Positive Shear value of A:A w:wC base pair for relevant Twist value having minimum DFT-D energy obtained by different levels of theory and coarse-grain energy penalty* taking different force constant values. ....	156
<b>Table 5-1</b> Different types of A:C, C:U and U:U base pairs involving different edges in either orientation along with their frequencies in the non-redundant crystal structure dataset and interaction energies. Although A:C and C:U base pairing is theoretically possible in 18 different ways but some possibilities are discarded due to impossibility of two hydrogen bonds in those	

cases, such as A:C H:HC, etc. or some base pairing being too infrequent, such as U:U H:ST (shown with Blue patch in the RNABPDB database). ..... 176

**Table 5-2** Optimized co-ordinates of atoms in transition state. .... 178

**Table 5-3** Structural comparison between generated 3-dimensional structures with possible base pair. Line representation in the figure represents central base pairs as highlighted in the table. 182

**Table 5-4** Generated model structure and the structural parameters..... 188

## **Chapter- 6 Conclusion and Scope of Future**

The complete work presented here is based on understanding non-canonical base pair from the lens of quantum mechanical calculation, MD simulation, and bioinformatics study.

### **6.1 Conclusion and Scope of future work (Chapter 2)**

All canonical and non-canonical base pair interaction energy has been calculated considering BSSE correction. The energy calculations were done for best representative structures of each type. However, often it was required to find a best representative from few examples (five or more observed crystallographic structures). This may lead to ambiguity. Geometry optimization of each base pair might improve the understanding. However, previous studies indicated that often a base pair changes its orientation, especially when the base pairing involves 2'-OH group mediated hydrogen bonding. Geometry optimizations considering BSSE corrected energies might improve the optimized geometries of the base pairs and improve correlation between frequencies of observations and interaction energy.

- How 2'-OH involves in base pairing and it improves the stability – have been highlighted from this database. The contraction of C—H bond length on hydrogen bond formation has been noted, which correlates with Blue-Shifting of C—H bond stretching frequency.
- Stacking energy is solely dependent on base pair and base pair step parameters. Multidimensional correlation analysis may reveal the proper relation between parameters and stacking energy. This work would require a substantial computational cost.

### **6.2 Conclusion and Scope of future work (Chapter 3)**

- This work shows how stacking energy scan finds the suitable configuration of dinucleotide step with non-canonical base pair.
- Stacking energy data can be useful for force field parameterization in coarse grain simulation.

### **6.3 Conclusion and Scope of future work (Chapter 4)**

- Promiscuous nature of A:A base pair gets highlighted in this chapter. The Sheared G:U base pairs induce force and A:A base pair moves depending on the resultant force's direction. It results in the adoption of unique Shear sign of A:A base pair.
- This study also points out that a similar type of behaviors might be observed for other base pairs where the bimodality of Shear is present.
- QM/MM study can reveal the activation energy for the transition between two Shear configurations.

### **6.4 Conclusion and Scope of future work (Chapter 5)**

- We have predicted the SL1 helix of RNA of SARS coronavirus, which causes deadly global pandemic nowadays considering quantum chemical calculation and bioinformatics study.
- SL1 helix contains protonated Adenine in A:C<sup>+</sup>:WC base pair.
- Complete folding of SL1 helix is incomplete without the protonation of Adenine.
- The thermodynamic integration protocol can reveal the structural change of the SL1 helix during protonation.

## **Thesis Abstract**

**Name of the Student:** *Satyabrata Maiti*

**Name of the CIOCC:** *Saha Institute of Nuclear Physics* Enrolment No.: *LIFE05201504004*

**Thesis Title:** **Non-canonical Base Pairs in RNA and their involvement in double helix formation.**

**Discipline:** *Life Sciences*

**Sub-Area of Discipline:** *RNA Structural Bioinformatics*

Though RNA has almost similar structural elements as DNA, it can regulate huge number of diverse cellular processes. Apart from the regular helical regions with Watson-Crick base pairs there exists a variety of structural motifs like, pseudo helices, hairpin structural elements with variable loop region, internal loops, multi-way junctions, bulges and many more. While proteins use twenty amino acids of various properties to perform different tasks through molecular recognition, RNA possesses only four nucleotide bases. These four bases can pair in various different ways, which are commonly referred to as non-canonical base pairs, to offer various steric and electrostatics shapes for molecular recognition. We have presented a combined network of web-servers and database (<http://hdrnas.saha.ac.in/>) for RNA structural perspective. We have additionally attempted to understand stability of these unusual contacts using Density Functional Theory based studies. (<http://hdrnas.saha.ac.in/rnabpdb>) This indicated sufficient strengths of interactions in the non-canonical base pairs and their stacks. We have found G:A S:HT base pair as one of the most frequent non-canonical base pair. We have developed a quantum chemical hybrid DFT-D based stacking energy analysis, which enables us to find the most probable configuration of the dinucleotide step containing a non-canonical base pair (G:A S:HT :: C:G W:WC). Using the above-described work as a benchmark study, we have tried to decode the promiscuous nature A:A w:wC base pair where one of the hydrogen bond is weak C-H...N mediated. Molecular dynamics simulation coupled with quantum chemical calculation and bioinformatics study ultimately dig out the hidden truth behind the promiscuous nature A:A w:wC base pair when it is stacked on sheared G:U W:WC base pair. We have later extend our work on non-canonical base pair to find the most probable configuration SL1 helix of RNA of corona virus. In the SL1 motif, either A:C +:WC or U:C W:+C base pairs are possible. Our hybrid stacking energy analysis along with transition state calculation, have supported on protonation of Adenine residue in A:C +:WC base pair.

## **Chapter- 1 Nucleic Acids: Structural Properties and Function**

## **1.1 Introduction:**

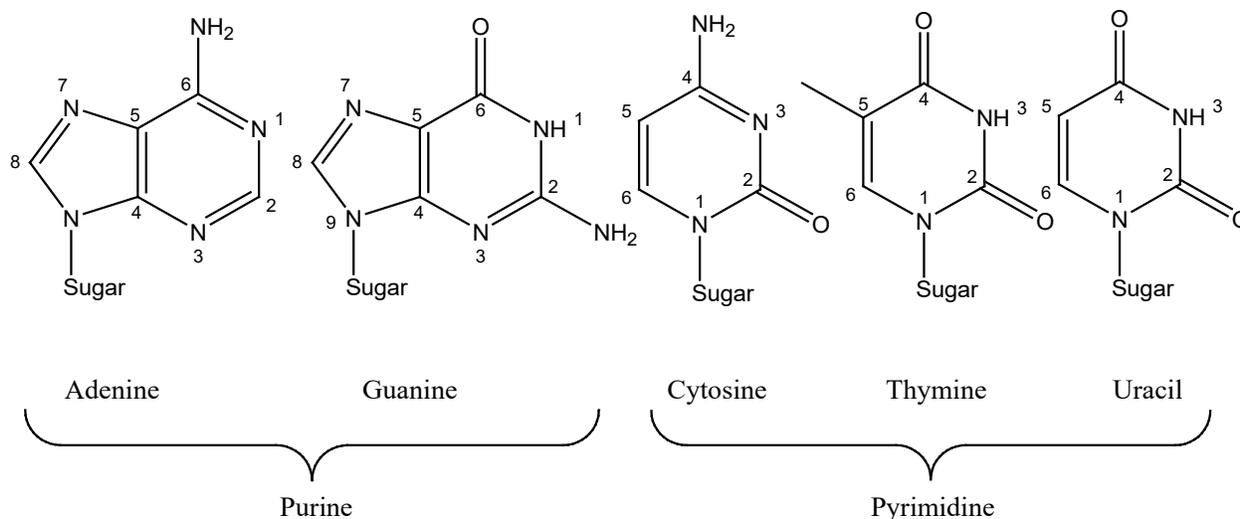
Nucleic acid is the new lens to see the hidden mystery of life. This prime macromolecule is the governing body of central dogma and controls the fate of the cell refereed by transcription and translation. Complicated molecular machinery is involved in central dogma that makes the major workers of cell i.e., protein. The diversity of protein is encrypted in the arrangement of monomers of nucleic acid polymer. Two nucleic acids are found in nature – deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Monomer units of nucleic acids are formed by a five-member sugar ring with the phosphate group and nitrogenous bases of four varieties, Adenine, Guanine, Cytosine, and Thymine (Uracil in RNA). The sugar unit of RNA has an extra hydroxyl group. The first glimpse of nucleic acid peeked through the series of experiments based on the “transforming principle” done by Fredrick Griffith. DNA as the genetic material was proved considering the above transforming principle and radio tracing study<sup>1,2</sup>. It was thought that the rudimentary secret of life engulfed into the three-dimensional structure of this unique molecule. During this time, Chargaff’s data proposed about chemical complimentary of bases<sup>3</sup>. Watson and Crick made a monumental discovery in 1953 as they ultimately solved the structure of the most beautiful molecule of life, DNA. They took Chargaff’s data and X-ray diffraction pattern and proposed DNA double-helical structure, which explains most of the hidden mysteries of life<sup>4</sup>. They also suggested DNA as a right-handed helix, which is composed of two antiparallel strands connected by hydrogen bonds between Adenine and Thymine and between Cytosine and Guanine. The concept of complementary base pair came into the world of genetics. The radio-tracing study, coupled with density gradient centrifugation technique, proved the semiconservative nature of DNA replication<sup>5</sup>. Several experiments outweighed the direct involvement of DNA in protein synthesis. The Discovery of RNA opened an era of thinking on the function of gene and protein

synthesis. Crick hypothesized “Adaptor Molecule” in protein synthesis, which was based on the outcomes of the experiments of Hoagland and Zamecnik<sup>6</sup>. Later this molecule was known as t-RNA. Many scientists were involved in the discovery of mRNA and ended the climax in the summer of 1961. Nine people, including Watson, Brenner, and Francis Jacob, were involved in the isolation of mRNA<sup>7,8</sup>. Many people argued that Nirenberg was the first person to isolate the mRNA. Later his work explored the function of mRNA. He also gave a hint of ribosomal RNA in protein synthesis<sup>9</sup>. When the catalytic activity of RNA was explored, several ribosomal RNAs (rRNA) came out as a basic unit of the ribosome. rRNA, as a helper molecule for the ribosome function, was suspected by several groups<sup>10,11</sup>. Atomistic resolution crystal structures finally proved that surmise<sup>12-14</sup>. Watson-Crick believed that the extra hydroxyl group's presence introduces van der Waals contact, which disables RNA to make a double helix. It was accepted that RNA is always single-stranded. In 1960, a DNA-RNA hybridization experiment by Rich gave an important hint on RNA double helix<sup>15</sup>. They ultimately showed RNA double helix after solving the crystal structure of it<sup>16</sup>. The Discovery of RNA double helix enhanced the scientific thinking on RNA, which led us to solve the “coding problem of RNA.” It also helps in finding RNA mediated gene regulation and microRNA.

## **1.2 Chemical Constituent of Nucleic Acid**

### *1.2.1 Nitrogenous Bases*

The chromatographic method followed by hydrolysis reaction was done to reveal correct structures of the nitrogenous bases of nucleic acid<sup>17</sup>. DNA consists of Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). RNA consists of the above three bases except for Thymine (T), which is replaced by Uracil (U) (**Figure 1-1**).



**Figure 1-1** Molecular structure of nucleobase

Adenine and Guanine have Purine rings, which are aromatic, whereas Thymine, Uracil, and Cytosine are Pyrimidine,

which are also aromatic. N9 of purine bases and N1 of pyrimidine bases are bonded

to anomeric carbon (C1') of sugar unit, resulting in adenosine, guanosine, and

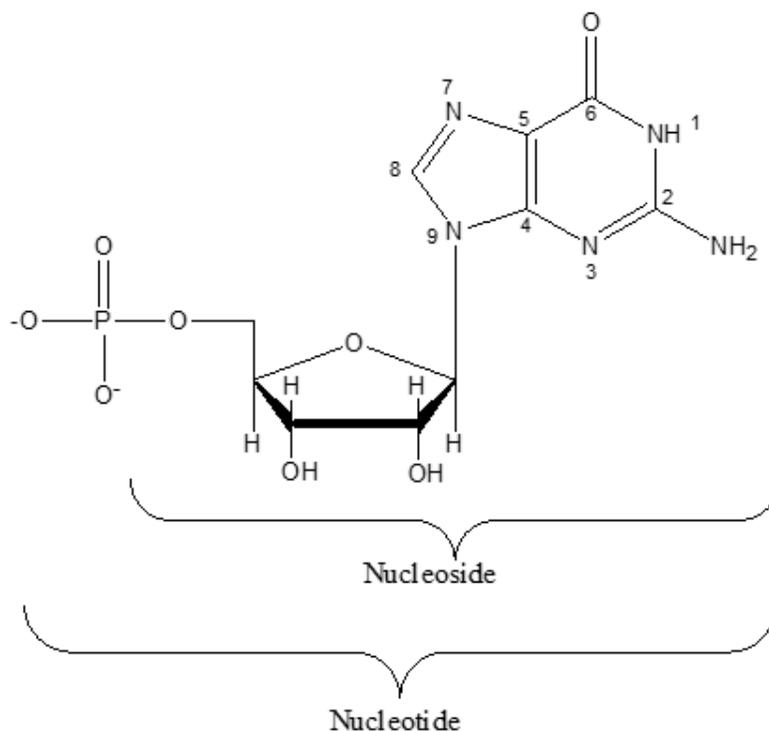
cytidine, uridine for RNA and the formation of

deoxyadenosine,

deoxyguanosine,

deoxycytidine, thymidine for **Figure 1-2** Nucleoside and Nucleotide as structural unit of nucleic acid

DNA. So Nitrogenous bases make the bond with sugar units to give the above nucleosides. The bond between N1 or N9 of nitrogenous base and C1' of sugar unit results in the glycosidic bond,



which is in  $\beta$  orientation for natural nucleic acid. It indicates that the nitrogenous base is always at the top of the plane of the sugar unit when it is viewed from the 5' hydroxyl group. The phosphate group makes a bond with the 5' hydroxyl group to form nucleotide, which is the unique building block of nucleic acid (**Figure 1-2**). The nucleic acid polymeric chain is represented by a single letter code of nucleotides from 5' to 3' direction, and it is known as a sequence. The nucleic acid chain with n number of nucleotide could have  $4^n$  number of distinct arrangement. Chemically modified bases (Dihydrouridine, Inosine, Pseudouridine, etc.) are also present in different functional RNA, such as tRNA and rRNA, etc.<sup>18</sup>.

### 1.2.2 Sugar Phosphate backbone

The sugar unit of RNA is ribose sugar, which is a cyclic aldopentose, i.e., furanoside ring.

2'-OH group is replaced by H

in the sugar unit of DNA,

which is 2-deoxyribose sugar.

Sugar units are joined together

by a phosphodiester bond.

Phosphodiester bond is

formed through a

condensation reaction

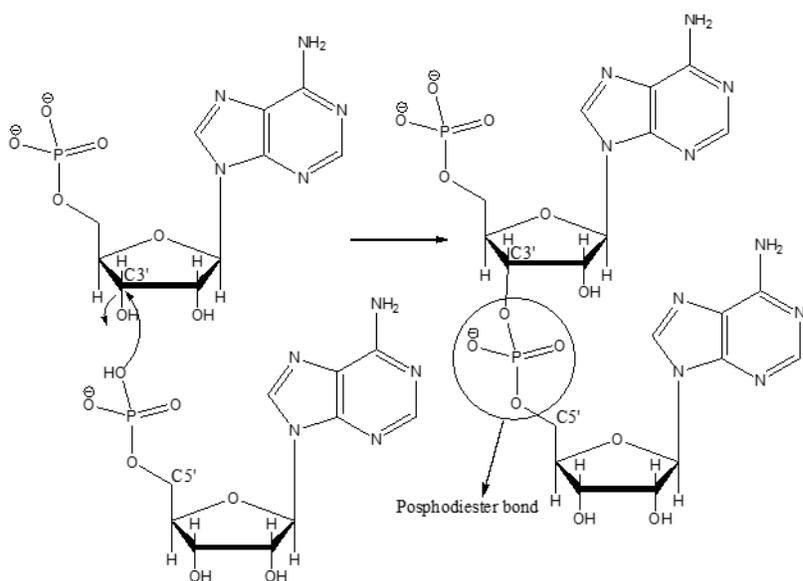
between the hydroxyl group of

two sugar units and a **Figure 1-3** Formation of phosphodiester bond

phosphate group. One water molecule is liberated in each ester bond formation. Phosphodiester

bond weaves the directionality of nucleic acid molecules. Synthesis of the phosphodiester bond

starts from 5' to 3' end (**Figure 1-3**). It results in the convention of writing nucleic acid in the



5'→3' direction. The negative charge on phosphate group makes nucleic acid polyanionic and polarity due to polyanionic feature makes this macromolecule soluble in water. Coulombic repulsion between two consecutive negative charges gets counterbalanced by the association of metal ions<sup>19</sup>.

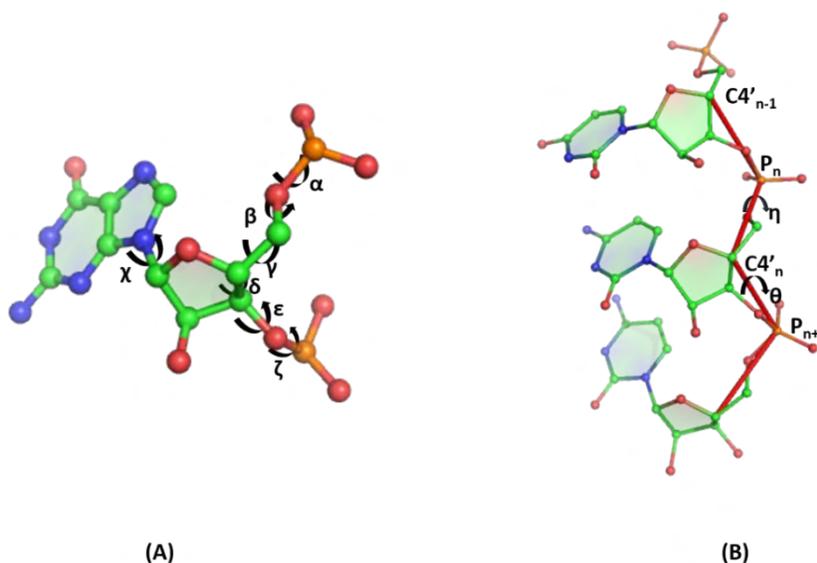
The number of possible conformations for nucleic acid is considerably higher than that of proteins. The number of variable torsion angles along single covalent bonds of the sugar-phosphate backbone is much higher than that of proteins having mainly two-variable torsion angles phi ( $\phi$ ) and psi ( $\psi$ ). IUPAC-IUB recommendation has portrayed backbone conformation in the light of torsion angles. These

torsion angles are symbolized as  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$  and  $\zeta$  representing rotations about P-O5', O5'-C5', C5'-C4', C4'-C3', C3'-O3' and O3'-P bonds respectively<sup>20</sup>

(Figure 1-4 and Table 1-1).

All conformations are never equally populated as it

depends on several electronic factors (dipole-dipole interaction, steric clash, etc.). These factors restrict some backbone conformation and weave nucleic acid flexibility<sup>21,22</sup>. Relative stabilities of backbone conformations also induce different helical structures of nucleic acid<sup>23</sup>. Backbone modeling of nucleic acid becomes a multidimensional problem as it requires 6 sets of torsional angles. Over the years, people were trying to define a



**Figure 1-4** Representation of Nucleic Acid (A) backbone torsion angle (B) pseudo torsion angle

new set of torsion angles to simplify the DNA/RNA representation. Therefore two new pseudo Torsional angles ( $\eta$  and  $\theta$ ) were described by introducing a virtual bond between P and C4'<sup>24</sup> (**Figure 1-4**). However, this concept of the torsional angle was the extension of the idea given by Malathi *et al.*<sup>25</sup>.

### 1.2.3 Sugar conformation:

Cyclic saturated hydrocarbons are never planar. They become non-planar to minimize the angular strain. Similarly, sugar unit of nucleic acid is puckered where one or two atoms go out of the plane formed by the remaining atoms. Five endocyclic torsional angles ( $\nu_0, \nu_1, \nu_2, \nu_3,$  and  $\nu_4$ ) precisely represent the conformation of sugar (**Table 1-1**). The conformation of sugar depends on the non-bonding interaction arising from neighboring substituents<sup>26</sup>. RNA and A-DNA (at hydrophobic condition) prefer C3' endo, whereas C2' endo is preferred by B-DNA (at high relative humidity) (**Figure 1-5**). These two types of conformations make dramatically different neighboring P...P distances (**Figure 1-5**). The phase angle P and the amplitude of sugar puckering,  $\nu_{max}$  determine the sugar conformation. The following equation portrays the conformation of puckered sugar<sup>27</sup>:

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2 \times \nu_2 \times (\sin 72^\circ + \sin 36^\circ)}$$

$$\nu_{max} = \frac{\nu_2}{\cos P}$$

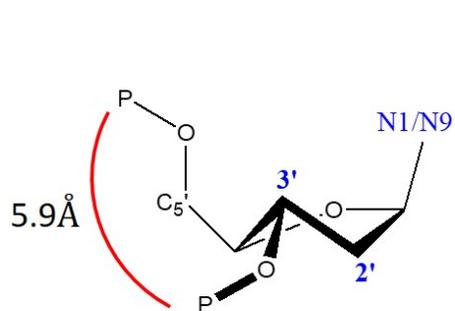
P and  $\nu_{max}$  guide the all endocyclic torsion angles by following equation<sup>28</sup>:

$$\nu_i = \nu_{max} \times \cos\{144^\circ(i - 2) + P\}$$

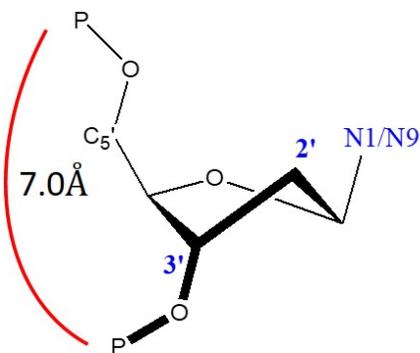
Phase angle (P) of 152° and 18° associate to C2' endo and C3' endo, respectively. Besides, there are eight significant classes of phase angle, and each of them defines a distinctly puckered sugar conformation.

**Table 1-1** Definition for the Sugar puckering dihedrals and backbone dihedrals

Definition for sugar puckering	
$\nu_0$	C4'-O4'-C1'-C2'
$\nu_1$	O4'-C1'-C2'-C3'
$\nu_2$	C1'-C2'-C4'-O4'
$\nu_3$	C2'-C3'-C4'-O4'
$\nu_4$	C3'-C4'-O4'-C1'
Definition for Sugar-Phosphate Backbone dihedral	
$\alpha$	O3' <sub>(n-1)</sub> -P <sub>(n)</sub> -O5' <sub>(n)</sub> -C5' <sub>(n)</sub>
$\beta$	P <sub>(n)</sub> -O5' <sub>(n)</sub> -C5' <sub>(n)</sub> -C4' <sub>(n)</sub>
$\gamma$	O5' <sub>(n)</sub> -C5' <sub>(n)</sub> -C4' <sub>(n)</sub> -C3' <sub>(n)</sub>
$\delta$	C5' <sub>(n)</sub> -C4' <sub>(n)</sub> -C3' <sub>(n)</sub> -O3' <sub>(n)</sub>
$\epsilon$	C4' <sub>(n)</sub> -C3' <sub>(n)</sub> -O3' <sub>(n)</sub> -P <sub>(n+1)</sub>
$\zeta$	C3' <sub>(n)</sub> -O3' <sub>(n)</sub> -P <sub>(n+1)</sub> -O5' <sub>(n+1)</sub>
$\chi$	O4' <sub>(n)</sub> -C1' <sub>(n)</sub> -N9 <sub>(n)</sub> -C4 <sub>(n)</sub> [Purine]
	O4' <sub>(n)</sub> -C1' <sub>(n)</sub> -N1 <sub>(n)</sub> -C2 <sub>(n)</sub> [Pyrimidine]
$\eta$	C4' <sub>(n-1)</sub> -P <sub>(n)</sub> -C4' <sub>(n)</sub> -P <sub>(n+1)</sub>
$\theta$	P <sub>(n)</sub> -C4' <sub>(n)</sub> -P <sub>(n+1)</sub> -C4' <sub>(n+1)</sub>



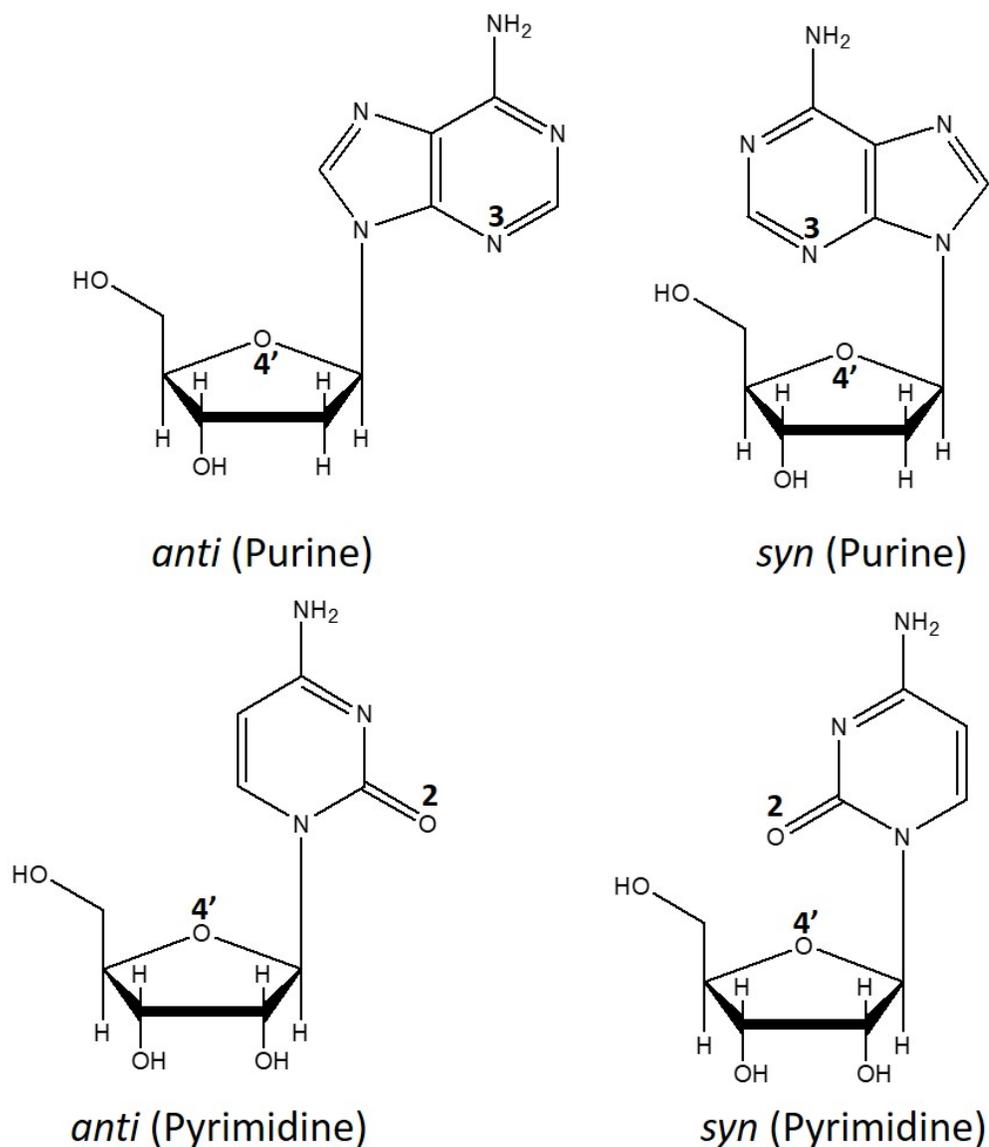
C3' endo (A-DNA like conformation)



C2' endo (B-DNA like conformation)

**Figure 1-5** Two major Sugar Puckering in nucleic acid.

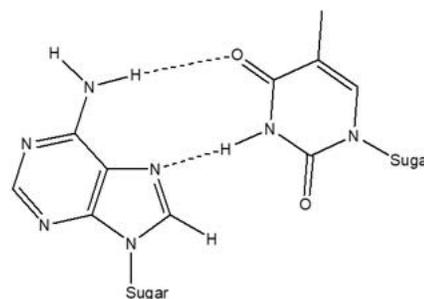
Apart from these torsional angles, another dihedral angle is required to define the relative orientation of the nitrogenous base and sugar unit. The dihedral angles about the C1'-N9 bond for purine and C1'-N1 bond for pyrimidine are named as  $\chi$  (**Table 1**). The *syn* conformation arises due to close proximity between O4' of sugar and N3 of purine or O2 of pyrimidine (**Figure 1-6**). When these two are on the opposite side, the *anti*-conformation develops.



**Figure 1-6** *syn* and *anti* conformations of nucleosides

### 1.3 Canonical and Non-canonical Base pair

In 1959, Karl Hoogsteen solved the high-resolution structure of A:T base pair using single-crystal crystallography. The N3 and O4/O2 atoms of thymine and N7 and N6 atoms of adenine were involved in the base pair described above (**Figure 1-7**), which was altogether dissimilar based on what was prescribed by Watson-Crick<sup>29</sup>. According to the alternate base pairing scheme, the hydrogen bond involving N7 atoms of purine



**Figure 1-7** A:T Hoogsteen base pair

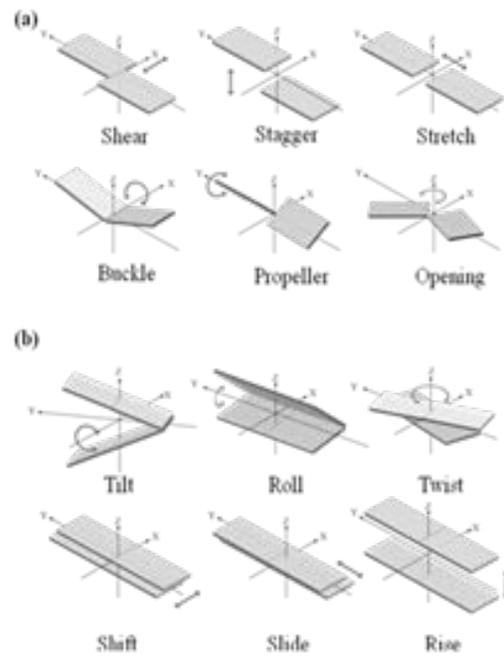
has been denoted as Hoogsteen base pair, so the above-mentioned base pair can be classified as Hoogsteen base pair. W. Guschelbauer found a high-resolution structure of G:C base pair, which was again similar to as N7 atom of Guanine base involved in base pairing. It is worth mentioning that the above-mentioned G:C Hoogsteen base pair needed protonation of cytosine, which was feasible at low pH<sup>30</sup>. Experimental data, together with NMR<sup>31</sup> and X-ray crystallographic data<sup>32</sup>, proved the existence of Watson-Crick base pairing after a decade, i.e., in the early '70s. Nearly a decade later, Richard Dickerson<sup>33</sup> and several other groups unraveled B-DNA double-helical structures with whole helical turns according to the DNA oligomer's synthetic crystal<sup>34-37</sup>. The base pairing configurations of G:C, A:U, and A:T proposed by Watson-Crick is known as a canonical base pairing, whereas other base pairing configurations, and compositions are now designated as non-canonical base pairs.

### 1.4 Orientation parameter in the base pair and Base pair Step

The Watson-Crick proposed base pairing scheme explains the key biochemical procedures, namely DNA replication, mRNA transcription, and mRNA to protein translation through base pairing between mRNA and tRNA. However, recent studies indicate that in addition to these

canonical base pairs, there are various other types of base pairs or base triples present in the high-resolution structures of tRNA, ribosome, etc., molecules. Base pairing is assisted by potential hydrogen bonding between nucleobases. The genome or transcriptome with differing structures and function comprise of base pairs with various combination of orientation. Similar kinds of base pairs show geometrical similarity with their identical C1'- C1' virtual bond distance and relative direction to the sugar-phosphate backbone<sup>38</sup>. These properties are the premise of generally speaking similar configuration in nucleic acid helices

regardless of the base sequence. Eventually, they can even have distinctive nearby structures due to the relative orientation of bases and base pairs, which induce different non-covalent interactions with protein or other nucleic acids. Relative orientation and direction of the bases inside a base pair or between two progressive base pairs in the 5'→3' direction characterizes the local structure. The known functional RNAs are not confined to just double-helical motifs. It can frame long-run



**Figure 1-8** Standard nomenclatures of *Base pair and Base pair Step parameter.*

communications through different optional auxiliary motifs and base pairing between various bases of a similar chain or other chains, much the same as proteins. It indicates the structural dependence of RNA on the diverse relative orientation of bases and base pairs, which ultimately creates complicated structural dynamics of RNA.

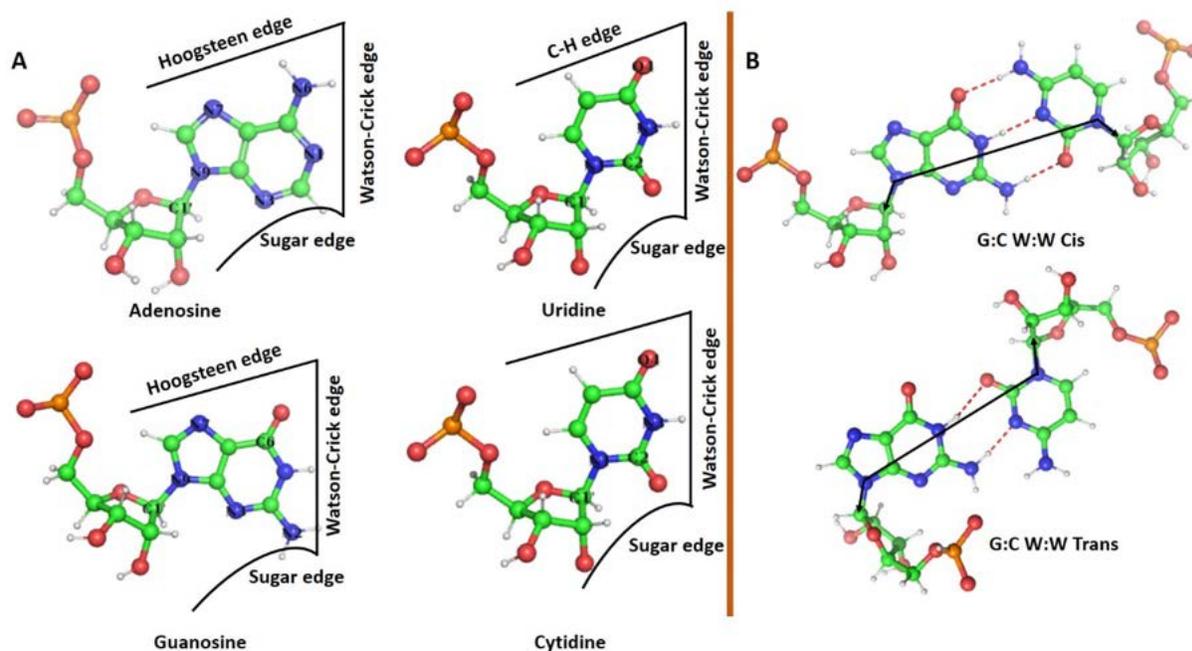
If we have two rigid blocks, we need three rotational degrees of freedom about three mutually perpendicular axes and three translational degrees of freedom alongside the three

mutually perpendicular axes to define those two rigid blocks' relative orientation. Likewise, considering each base as a rigid unit, geometry a base pair can be defined with six parameters – three translational and three rotational parameters or degrees of freedom. IUPAC suggested parameters are Propeller, Buckle, Open Angle, Stagger, Shear, and Stretch (**Figure 1-8**)<sup>39</sup>. Among six intra-base pair parameters those mentioned above, Shear, Stretch, and Open-angle describe different hydrogen-bonding pattern. In contrast, the overall non-planarity of a base pair can be quantified by the remaining three intra base pair parameters – Buckle, Propeller, and Stagger. Similarly, considering each base pair as a rigid unit, the base pair step's structural feature can be defined, using six parameters – three translational (Shift, Slide, and Rise) and three rotational (Tilt, Roll, and Twist). Curves by Richard Lavery<sup>40</sup>, 3DNA by Wilma Olson<sup>41</sup>, NUPARM by Manju Bansal<sup>42</sup>, etc., are famous publicly available software, which can calculate the above-described parameters. First two softwares compute the canonical and non-canonical base-pairs' parameters considering standard geometry of canonical Watson-Crick base pairs. Whereas the NUPARM algorithm computes the same considering the actual axis system that defines a unique base pair edge. Hence, calculated some base pair parameters by Curves or 3DNA are usually unrealistic or large in their respective intrinsically most stable configuration. Besides, base pair parameters, planarity of the two bases, and quality of hydrogen bonding calculated by NUPARM are more realistic.

### **1.5 Base pair Nomenclature, and Local strand direction:**

The nitrogenous bases of nucleic acid are planar heterocyclic system, with conjugated pi-electron cloud. Base pairing happens by multiple hydrogen bond donors and acceptors atoms which are situated around the different edges, usually labeled as W, H or S. W, H and S stands for Watson-Crick base pair, Hoogsteen base pair, or, whether the edge is adjacent to the C2'-OH group

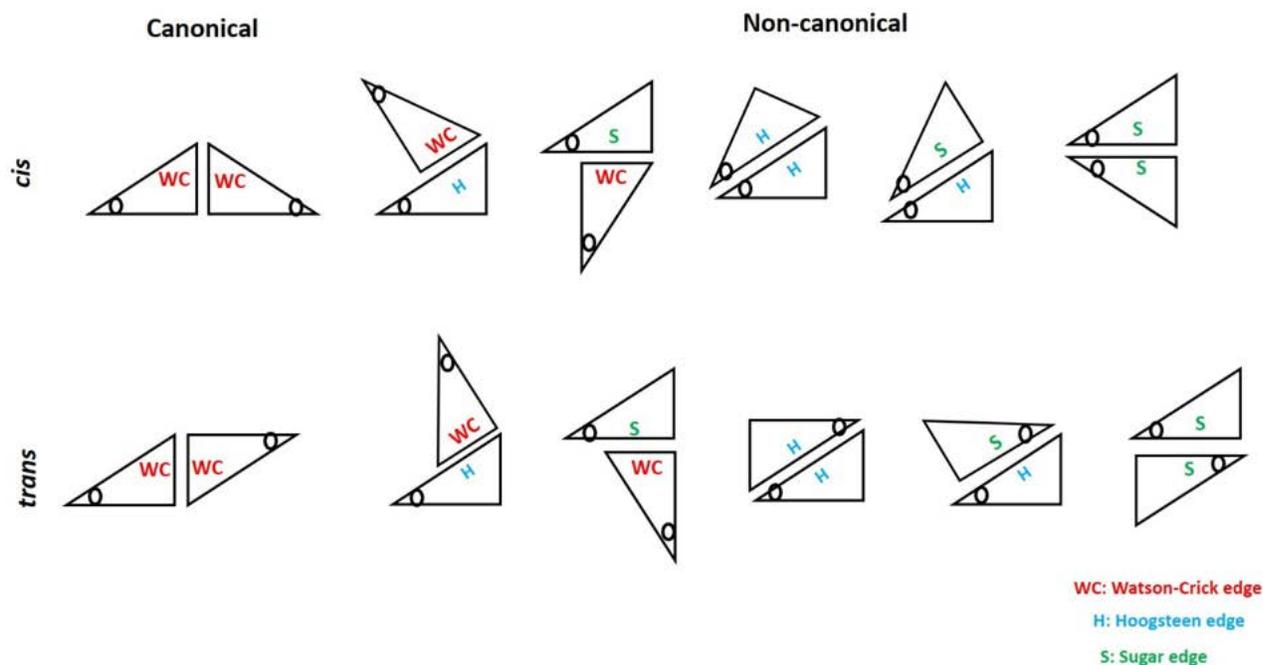
of the ribose sugar, respectively (**Figure 1-9**). Eric Westhoff and Neocles Leontis<sup>43</sup> proposed an extensively accepted nomenclature system for the base pairs which delivers a useful method towards the classification of base pair geometries with respect to interacting edges. Any polar hydrogen bond acceptor atom such as N7 is not present in Hoogsteen edge of the pyrimidine bases unlike the corresponding edges of purine bases. However, Gautam Desiraju also proposed that C—H can act as weak hydrogen bond donors in base pairing<sup>44</sup>. Hence, the Hoogsteen edge is also named as Hoogsteen/C-H edge in a suitable system for assigning equivalent positions of pyrimidines and purines. There are 6 possible base pairing edges, namely Watson-Crick/Watson-Crick (or W:W), Watson-Crick/Hoogsteen (or W:H), Watson-Crick/Sugar (or W:S), Hoogsteen/Hoogsteen (or H:H), Hoogsteen/Sugar (or H:S) and Sugar/Sugar (or S:S).



**Figure 1-9** (A) Base pairing edges for the nucleotides (B) Cis and Trans orientations of the two nucleotides and the glycosidic bonds are represented by the arrow as vectors).

When two glycosidic bonds in base pair direct same side, it is termed as *Cis* geometry. In contrast, *Trans* geometry forms when two glycosidic bonds are directed on the opposite side

(Figure 1-9B). The 6 base pairing edges coupled with *Cis* and *Trans* orientation classify 12 geometric families of base pairs (Table 1-2, Figure 1-10).



**Figure 1-10** Representation of possible base pairing

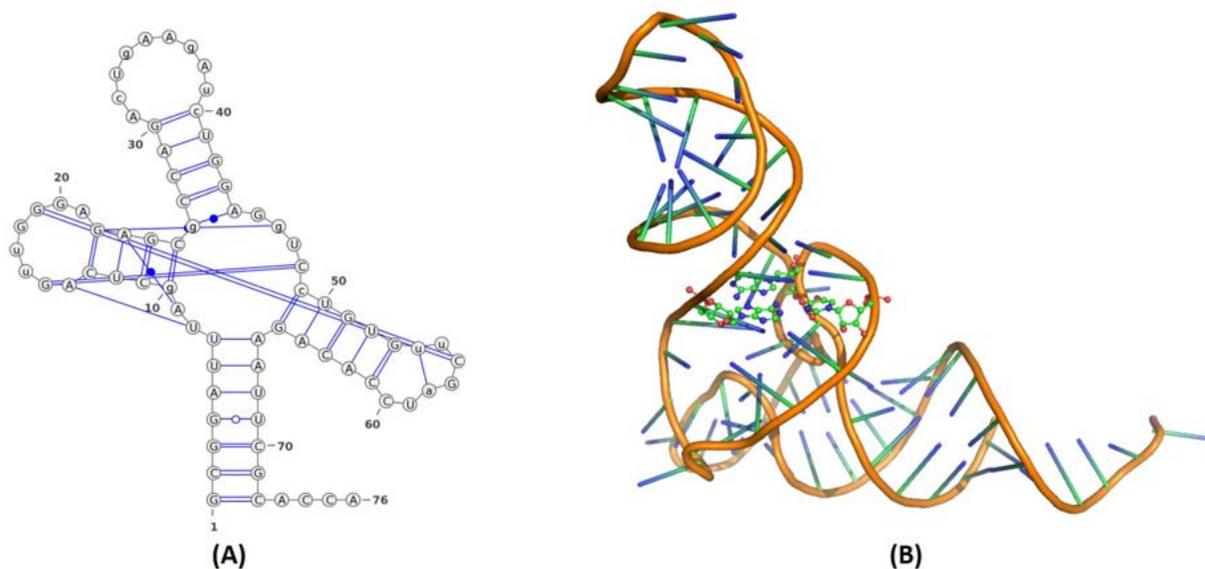
**Table 1-2** Different types of base pairing schemes and associated local strand orientations of their sugar-phosphate backbone

Interacting edges	Glycosidic bond orientation	Nomenclature	Symbolic representation	Local Strand Direction
Watson-Crick/Watson-Crick	<i>Cis</i>	cWW or <i>cis</i> Watson-Crick/Watson-Crick	●—	Antiparallel
Watson-Crick/Watson-Crick	<i>Trans</i>	tWW or <i>trans</i> Watson-Crick/Watson-Crick	○—	Parallel
Watson-Crick/Hoogsteen	<i>Cis</i>	cWH or <i>cis</i> Watson-Crick/Hoogsteen	●■	Parallel

Watson-Crick/Hoogsteen	<i>Trans</i>	tWH or <i>trans</i> Watson-Crick/Hoogsteen		Antiparallel
Watson-Crick/Sugar edge	<i>Cis</i>	cWS or <i>cis</i> Watson-Crick/Sugar edge		Antiparallel
Watson-Crick/Sugar edge	<i>Trans</i>	tWS or <i>trans</i> Watson-Crick/Sugar edge		Parallel
Hoogsteen/Hoogsteen	<i>Cis</i>	cHH or <i>cis</i> Hoogsteen/Hoogsteen		Antiparallel
Hoogsteen/Hoogsteen	<i>Trans</i>	tHH or <i>trans</i> Hoogsteen/Hoogsteen		Parallel
Hoogsteen/Sugar edge	<i>Cis</i>	cHS or <i>cis</i> Hoogsteen/Sugar edge		Parallel
Hoogsteen/Sugar edge	<i>Trans</i>	tHS or <i>trans</i> Hoogsteen/Sugar edge		Antiparallel
Sugar edge/Sugar edge	<i>Cis</i>	cSS or <i>cis</i> Sugar-edge/Sugar-edge		Antiparallel
Sugar edge/Sugar edge	<i>Trans</i>	tSS or <i>trans</i> Sugar-edge/Sugar-edge		Parallel

Like proteins, functional RNA molecules stabilize themselves in unique folding where canonical and non-canonical base pairs participate. For example, most tRNA molecules are known by their cloverleaf like two-dimensional structure, consisting of four short double-helical

segments. In contrast, the three-dimensional structure of the same is L-shape, which has several non-canonical base pairs and base triplets (**Figure 1-11**).



**Figure 1-11** (A) Secondary structure of  $t\text{-RNA}^{\text{Phe}}$  generated by BPFIND (B) The L-shaped X-ray crystal structure of  $t\text{-RNA}^{\text{Phe}}$  (PDB ID: 1EHZ) and ball-stick model represents base triplet (Residue ID: 9, 12 and 23).

### 1.6 Identification of Non-canonical base pair

If there should be an occurrence of double-helical DNA, distinguishing proof of base sets is very inconsequential utilizing atomic visualizers, for example, VMD, RasMol, PyMol, and so forth. It is, nevertheless, not all that straightforward for single-stranded folded RNA with thousands of nucleotides. A few calculations have been portrayed in the computerized algorithm to find out the base pairs from the RNA X-ray crystal structure, NMR, or different techniques. Basically, the schemes that distinguish hydrogen bonding between two bases requires their close approach and planarity before the announcement of proper base pair. Since the vast majority of the structures of RNA, accessible in PDB<sup>45</sup>, can be tackled by X-ray crystallography, the position

of hydrogen atoms is hardly stated. Henceforth, the identification of hydrogen bond turns into a non-trivial work.

The DSSR algorithm<sup>46</sup> by Lu and Wilma K. Olson believes proper base pairing when they detect at least one hydrogen bond between the bases by determining positions of the hydrogen atoms, and the perpendiculars the two bases which are almost parallel to one another. The internal coordinates such as bond length and torsional angle can determine the position of hydrogen atoms along with the position of precursor atoms. The NDB<sup>47</sup> and FR3D<sup>48</sup> databases utilize this technique to find base pairs.

Francois Major in MC-Annotate<sup>49</sup> integrated a special method for the identification of base pairs in RNA. Molecular mechanics force-fields<sup>50</sup> are utilized to determine the position of hydrogen atoms and lone pairs in this method, which helps to estimate the probabilities for hydrogen bond formation. The technique described above likewise endeavored to provide nomenclature of the base pair with other data of base pairing edges.

**Table 1-3** *Nomenclature system in MC-Annotate*

Nomenclature	Description
Ws	sugar edge corner of the Watson-Crick edge
Wh	Hoogsteen edge corner of Watson-Crick edge
Bw	bifurcated three-centered hydrogen bond

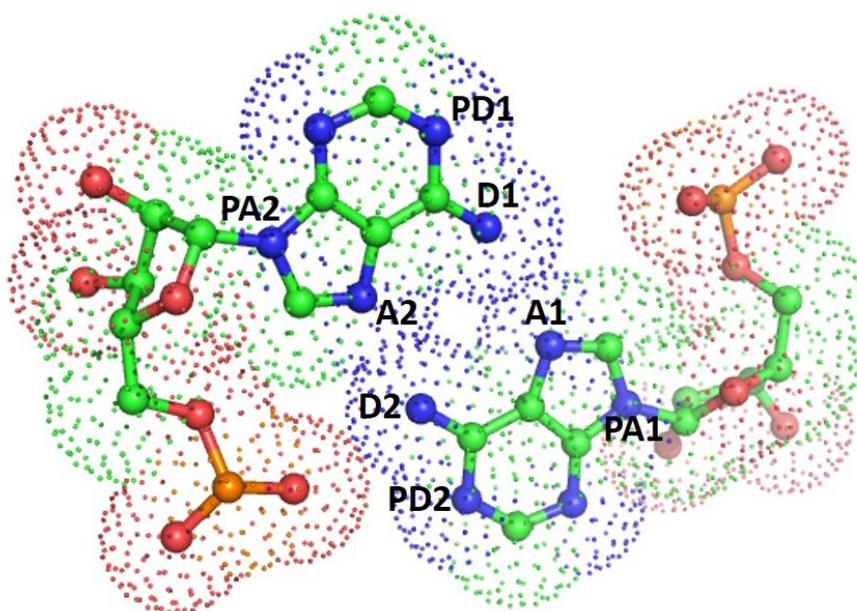
As demanded by the authors, the above-described nomenclature scheme enhances some extra highlights to Leontis-Westhof (LW)<sup>43</sup> scheme that might be alluded as the LW+ class. A significant bit of this scheme's scope covers its capacity to recognize alternate base pairing geometry, where multimodality is seen inside a LW class. Nevertheless, this technique doesn't

think about the probability of cooperation of the 2'- hydroxyl group of the ribose sugars in base pair formation.

Alternative approach, to be specific BPFIND by our group<sup>51</sup>, requires more than one hydrogen bonds utilizing particular arrangements of donor and acceptors atoms in the bases. The mentioned assumption demands the calculation of the distances between two pairs of atoms, i.e., hydrogen bond donor (D1 and D2) and acceptor (A1 and A2). It also demands the angles between appropriately picked precursor atoms (PD1, PA1, PA2, PD2) relating to the D's and A's (as shown in **Figure 1-12**). Small values of such distances coupled with large values of the angles characterized by PD1—D1—A1, D1—A1—PA1, PD2—D2—A2, D2—A2—PA2 (near 180° or  $\pi^c$ ) guarantees two geometrical features simultaneously that describe the characterization of suitable base pair:

- i) Strong and linear the hydrogen bonds and
- ii) Co-planarity of the two bases are.

It is noted that if one confines the search to find out the base pair, which gets stabilized by more than one hydrogen bonds, the above-described calculations yield a similar arrangement of base pair in various RNA structures.

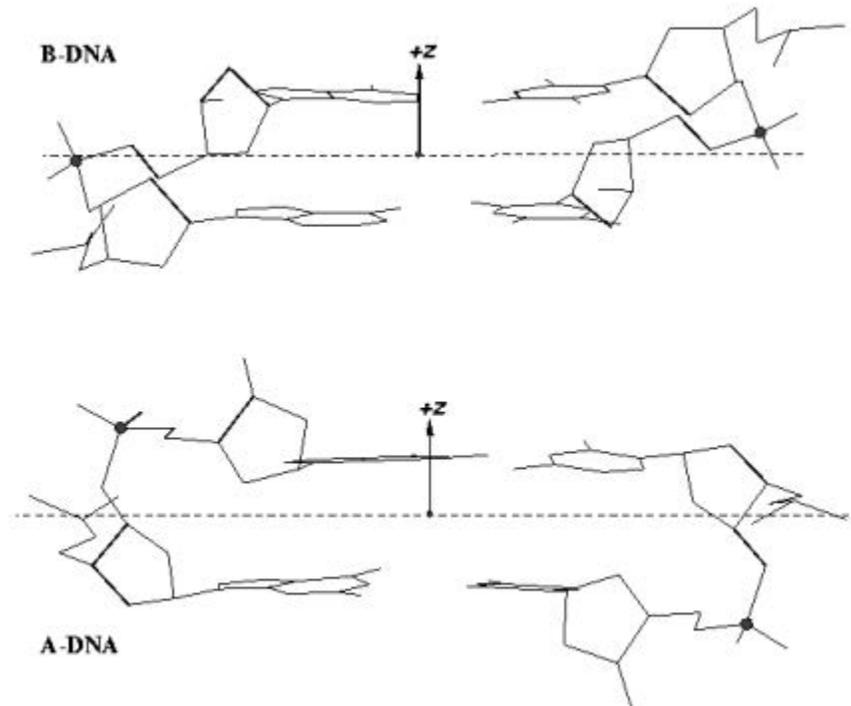


**Figure 1-12** Description of hydrogen bonding in a non-canonical base pair with their precursors (as used by BPFIND algorithm)

### 1.7 Nucleic acid Structural Polymorphism

The most well-known structure exists in most DNA at physiological conditions, is B-form, which is a classical right-handed double-helical structure. In contrast, the A-form of nucleic acid has been found in RNA duplex and DNA-RNA duplex and DNA sequences at lower humidity. A-form nucleic acid has 11 bp per turn, whereas B-form DNA has 10bp per turn. The significant variance between A-form and B-form nucleic acid arises from the corresponding sugar configuration. B-form nucleic acid prefers C2' endo sugar pucker, and A-form nucleic acid takes C3' endo sugar pucker. However, there are also some exceptions where B-DNA sugar adopts C3'-endo pucker or A-RNA sugars adopting C2'-endo pucker. The differences between A-form and B-form nucleic acid have also been reflected on base pair and base pair step parameter also (**Table 1-4**). Roll (the wedge formation between successive base pairs) values are generally found to be around zero in B-DNA, while the same is largely positive ( $5^\circ$  to  $15^\circ$ ) in A-form nucleic acid double helices. Similarly, Slide values (relative displacement of one base pair

with respect to its neighboring one along base pair long axis) are around zero in B-DNA, whereas its values are found to be around  $-1$  to  $-2\text{\AA}$  in most A-form structures). However, there are overlapping zones of Roll and Slide in A and B-form structures. Hence it is difficult to distinguish whether a structure is in A- or B-form from the parameter values. Hence, another parameter,  $Z_p$  was proposed by Olson, and it was found to be able to keep track of conversion of A-DNA to B-DNA and vice versa (**Figure 1-13**). It is noted that RNA has A-DNA like conformation. The  $Z_p$  ( $\text{\AA}$ ) tells us the mean z-coordinate of the P atom with respect to the mean reference frame of a dinucleotide step. Base planes and phosphorous atoms are two perfectly positioned entities in nucleic acid structure. The  $Z_p$  is highly robust and discriminative in order to distinguish A-DNA and B-DNA.  $Z_p$  is greater than  $1.5\text{\AA}$  for A-DNA and less than  $0.5\text{\AA}$  for B-DNA<sup>52</sup>.



**Figure 1-13** Illustration of  $Z_p$  of dinucleotide steps of A-DNA and B-DNA. Images taken from Olson et.al<sup>52</sup>.

**Table 1-4** Average values of Base pair and Base pair step parameter for A-form and B-form nucleic acid. Data is taken from olson et al<sup>53</sup>. Standard deviations are given in parenthesis.

	Parameter	A-form	B-form
Base pair Parameter	Buckle (°)	-0.1 (7.80)	0.50 (6.70)
	Propeller (°)	-11.8 (4.10)	-11.4 (5.30)
	Open (°)	0.60 (2.80)	0.60 (3.10)
	Shear (Å)	0.01(0.23)	0.00(0.23)
	Stretch (Å)	-0.18 (0.10)	-0.15 (0.12)
	Stagger (Å)	0.02 (0.25)	0.09 (0.19)
Base pair Step Parameter	Tilt (°)	0.10 (2.80)	-0.1 (2.50)
	Roll (°)	8.00 (3.90)	0.60 (5.20)
	Twist (°)	31.10 (3.70)	36.00 (6.80)
	Shift (Å)	0.00 (0.54)	-0.02 (0.45)
	Slide (Å)	-1.53 (0.34)	0.23 (0.81)
	Rise (Å)	3.32 (0.02)	3.32 (0.19)

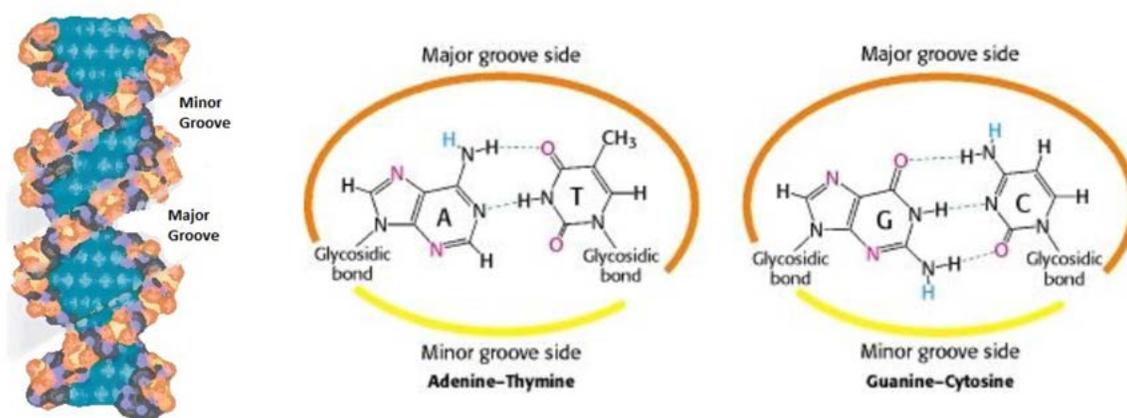
### 1.8 Major and Minor Groove

The Major groove forms when the space between two consecutive backbones is long, whereas minor groove forms when the room mentioned above is short. If glycosidic bonds of base pairs are extended to form an imaginary line, an acute and obtuse angle will be formed. Space arising from the acute angle is called minor groove, and space arising from obtuse angle is called major groove. (**Figure 1-14**) The groove functionality varies drastically with the different polar and hydrophobic groups in a sequence-specific manner. Major groove functionality changes more by sequence variation. Besides, minor groove properties are governed by imino and carbonyl groups whose electrostatic nature is quite similar (**Table 1-5**). Most of the drugs whose mode of binding with DNA are non-covalent interactions, those drugs bind at minor groove of DNA.

**Table 1-5** Atomic distribution of nucleotides in the minor and major groove.

Base name	Minor Groove	Major Groove
Adenine	N3	N6 amino, N7
Guanine	N2 amino, N3	C6 carbonyl, N7

Thymine	C2 carbonyl	C4 carbonyl, C5 methyl
Cytosine	C2 carbonyl	N4 amino
Uracil	C2 carbonyl	C4 carbonyl



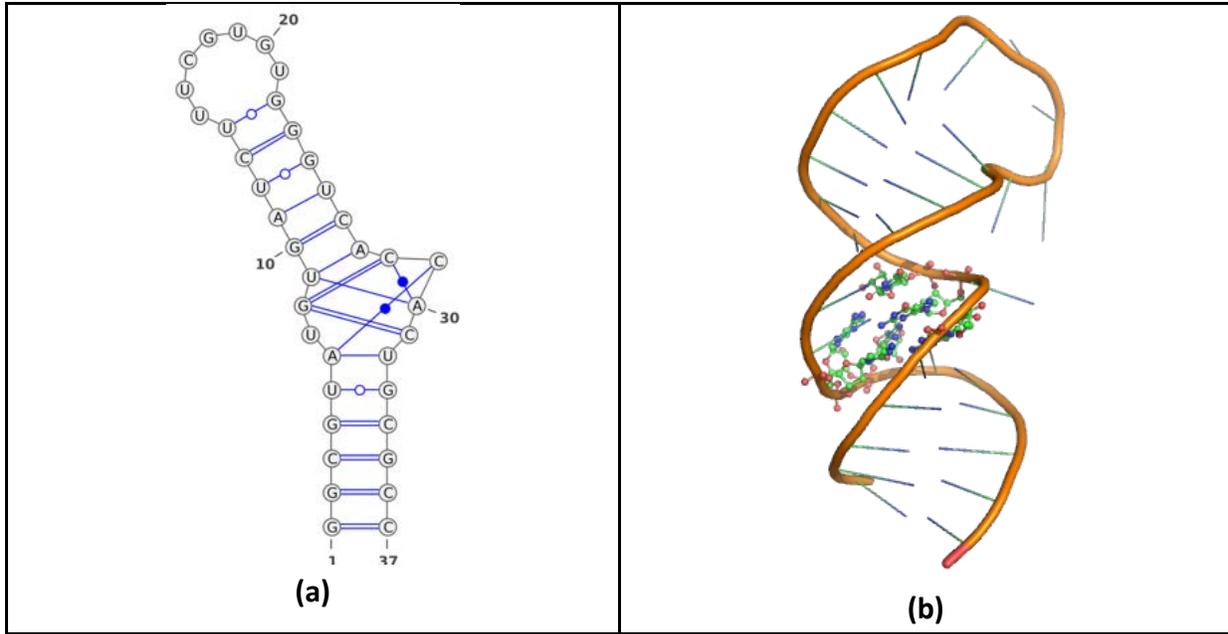
**Figure 1-14** Atomic and cartoon representation of major and minor groove

### 1.9 Non-canonical Base-pairs in Double Helical Regions

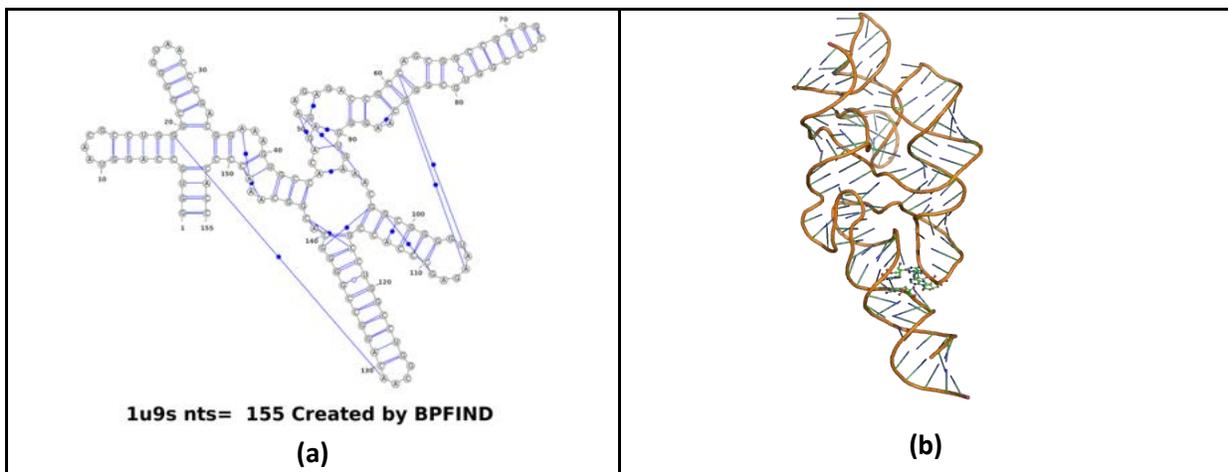
Double helical motifs of RNA often have non-canonical base pairs. The G:U cWW (or G:U W:WC) non-canonical base is nearly isosteric to the other canonical ones, and hence are often seen within the regions of double helix<sup>54-56</sup>. **Table 1-2** (Classification part) illustrates the compilation of strand direction, because of which all types of non-canonical base pairs cannot be fitted within double helical regions with *anti* glycosidic bond. Although non-canonical base pairs for example, A:G cWW or A:G W:WC, A:G tHS or A:G H:ST (trans Hoogsteen/Sugar edge), A:U tHW or A:U H:WT (trans Hoogsteen/Watson-Crick), etc. are frequently observed within symmetrical internal loops of double helical section. In recent times, various studies have been done to categorise all such cases in which 2 base pairs (canonical or non-canonical) assemble in an antiparallel way, perhaps that forms RNA double helical regions<sup>37</sup>. The stability of these base pairs enable themselves to sustain the double helical properties reasonably well. As C3'-endo sugar

pucker has anti glycosidic torsion,  $\alpha/\gamma$  around  $-60^\circ/60^\circ$ ,  $\beta/\epsilon$  around  $180^\circ$ , it indicates presence of suitable torsional angles around the residues.

When the single-stranded nucleobases of the loop region pair with a complementary sequence outside this loop of the same chain and fold back on it to form the other hairpin-loop region's stem region, a pseudoknot motif forms. In pseudoknots, several non-canonical base pairs are seen along with their unique hydrogen bonding. Structural features of these recurrent motifs have been archived in searchable databases, as The RNA FRABASE<sup>57</sup>, FR3D<sup>48</sup> etc., archive these recurrent motifs and their structural features. NASSAM<sup>58</sup> web-server also detects these kind of motifs using a PDB file. They involve in several biological function. For example, they act as capping residues in the termini of double helical segment. The G:A tSH (or G:A S:HT) base pair is one of the most commonly found base pair. GNRA (N= any nucleotide and R= Purine base) tetraloop, hammerhead ribozyme's metal binding site etc. consist of this non-canonical base pair. The GNRA tetraloop displays some flexibility and geometrical features that depend on paired G:A base pair or unpaired G:A base pair. For example, several new tetraloops motifs, such as CUYG, UNCG, YNMG and GNAC (where Y =pyrimidine and M =Adenine or Cytosine), are present in available RNA structures. However, these do not usually display the participation of non-canonical base pairing. The common hairpin loop does not form proper base pairing. Instead, they remain as unpaired where interaction between other residues and unpaired bases happen. In the C-loop motif<sup>59</sup> the bulging loop residues involve in non-canonical base pairing with the bases of double-helical regions (**Figure 1-15**). In T-loop motif<sup>59</sup>, non-canonical base pairs participate in receptor-loop interaction, as shown in **Figure 16**.



**Figure 1-15** (a) schematic and (b) by 3-D representation of base triples with non-canonical base pairs from PDB ID 1KOG.



**Figure 1-16** (a) schematic and (b) by 3-D representation of T-loop from PDB ID 1U96 where residues 105 and 10 make base triples with Watson-Crick base pair of 61:84.

Watson-Crick G:C cWW or G:C W:WC base pair is stacked by two non-canonical base pairs, namely, A:A cHS or A:A H:SC and U:A tWH or U:A W:HT base pair and these recurrent

non-canonical base pairs act as receptor of GAAA motif. It indicates the presence of non-canonical base pair in the anti-parallel RNA double-helical region. In the dinucleotide platform, base pairing between two consecutive residues is seen. The reported dinucleotide platforms are A:G, A:U, A:A base pairs where they belong to cSH or S:HC base pairing pattern. These kinds of motifs are responsible for reverse the strand direction of RNA double helix, which results in the generation kinks motif. In the Sarcin-ricin motif, dinucleotide platforms are also found where they form<sup>59</sup>.

### **1.10 Hydrogen bonding and Stacking Energy**

The canonical Watson-Crick base pair, G:C and A:T/U just as the majority of the non-canonical ones are stabilized by at least two (for example, 3 on account of G:C) hydrogen bonds. Reasonably, many studies on non-canonical base pairs have been done to benchmark their strength (interaction energies) and (geometric) stabilities with respect to those of the canonical base pairs. It is well-known here that calculations of base pair, as seen in the crystal structures, are frequently affected by a few interactions, such as metal ion – base interaction, in the crystal environment, perturbing their characteristically steady interaction emerging out of the hydrogen bonding in base pairs and stacking interaction in base pair step. It is conceivable that the experimentally detected geometries are sometimes intrinsically, not so stable and stabilized by different interactions arising from the crystal environment. Quantum chemistry revealed hydrogen bonding as the charge transfer from lone pair of H-bond acceptor to  $\sigma^*$  of H-bond donor. A few groups have endeavored to measure the interaction energies of non-canonical base pairs using quantum mechanics based methodology, for example, Density Functional Theory (DFT) or MP2 methods<sup>60-67</sup>.

There are three types of interaction energies according to optimization protocols. In the primary strategy, the geometry of base pair models, segregated from their surroundings, are completely optimized without any constraints<sup>61,62,66-68</sup>, thus giving the interaction energies for the

isolated models. This method sometimes leads to optimized geometry, which is significantly different from the initial one. Another method optimizes the base pair geometry by constraining non-hydrogen atoms (heavy atoms)<sup>63,65,67</sup>. In the third method, followed by Jiri Sponer and his group, optimization is done by constraining certain angles and dihedrals<sup>69</sup>. Interaction energies have been calculated using these optimized geometries. Generally, B3LYP functional is often used by most groups with 6-31G\*\* for optimization.  $\omega$ B97xD/cc-pVDZ, MP2/cc-pVDZ, MP2/aug-cc-pVDZ etc., level of theory has been used during the calculation of base pair interaction energies. Sponer's group generally uses RI-MP2 method for interaction energy calculation. Interaction energy calculated by the CCSD(T)/CBS method is considered gold standard, and the S22 dataset for the interaction energy has been built by Jurecka *et al*<sup>70</sup>.

Base pair stacking happens between the aromatic surfaces of the nitrogenous bases in nucleic acid. It is as significant as base pairing to get the thermodynamic stability of nucleic acid. Stacking interaction is a type of  $\pi$ - $\pi$  interaction. These types of interactions, arising from  $\pi$  orbitals, are weaker than general covalent and ionic bonds. It has been observed that systems with  $\pi$  bonds are more stable than systems without  $\pi$  bonds having the same number of atoms in both cases<sup>71</sup>. There are two types of  $\pi$ - $\pi$  interaction – face to face stacking interaction and T-shaped interaction. The interaction between closed-shell molecules whose  $\pi$  bonds are parallel to each other is called stacking interaction, and whose  $\pi$  orbitals are perpendicular to each other are called T-shaped interaction. Nucleic acid sequence and structure have a crucial link with stacking interaction between two successive base pairs<sup>72</sup>. But it is not only interaction between  $\pi$  electrons of two successive aromatic moieties, but also it includes the van der Waals' (VDW) interaction and other electrostatic terms<sup>72</sup>. Two successive base pairs often experience steric clash due to their inherent Propeller Twist, and local dinucleotide parameters (Twist, Roll, and Slide) change accordingly to

minimize this steric clash. This rule is recognized as “Calladine’s Rule,” which has been utilized for quite a long time to increase some subjective understanding on the sequence dependency of the local structure of DNA/RNA double helices<sup>73,74</sup>. These three parameters are quite useful in order to predict the configuration of base pair step with the help of stacking energy<sup>75-78</sup>.

The Hartree-Fock method takes many-electron wave functions as a linear combination of single electron wave functions. It results in the inability to incorporate electron correlation energy in the calculation. It is wholly bypassed considering electron density by Hohenberg-Kohn equation in density functional theory (DFT), which however, does not consider exchange energy term. The hybrid DFT method takes the exact exchange energy term from the Hartree-Fock method and the remaining exchange-correlation energy from other empirical methods. B3LYP is one of the most popular hybrid functional<sup>79</sup>. However, dispersion interaction is not considered by DFT, for which DFT-D is required. Dispersion effect influences base pair stacking energy. Hence, stacking energy calculation must be carried out with consideration of dispersion effect using a large basis set of atomic orbital. When two interacting units are separated by 3.3-3.4Å, atomic orbitals are used to fill this space, and it is prescribed to use a diffused polarized basis set during the calculation of stacking energy<sup>80</sup>. The easiest method to introduce the dispersion effect into DFT is to consider a type of an empirical dispersion correction to the DFT results. The empirical dispersion correction is generally founded on the asymptotic equations, which are sufficient for long-range correction. As suggested by Hobza<sup>81</sup> group,  $\omega$ B97X-D<sup>82</sup>, M06-2X<sup>83</sup>, etc., work well to calculate the stacking energy of nucleic acid bases where  $\omega$ B97X-D and M062-X include long-range (>5Å) and medium-range (<5Å) correction, respectively. Though the computationally expensive MP2 method is prevalent to estimate non-covalent interaction, substantial overestimation has been reported to assess  $\pi$ - $\pi$  stacking energy<sup>81,84,85</sup>. Though stacking energy

calculation using a large basis set needs BSSE correction by counterpoise method<sup>86</sup>, dispersion corrected density functional can generate small values of BSSE<sup>87</sup>. It decreases the computational cost of calculation.

## **Chapter- 2 Non-canonical Base Pairs and their stabilities**

## **2.1 Introduction**

The tremendous functional variability of relatively newborn non-coding RNAs has opened a door in the direction of post-transcriptional and post-translational regulation. Despite having almost the same structural elements as DNA, RNA can regulate many diverse cellular processes. Perhaps, the complicated regime of secondary and tertiary structural organization of RNA would facilitate a better understanding of its varied nature. This functional diversity is only comparable to the proteins. It is to be noted that the variety of proteins is expected with its 20 different building blocks having different charges, molecular weights, and hydrophobic properties, while RNA has only 4. The monotonous helical B-form of DNA with its wide major and narrow minor groove is replaced with a deep narrow major groove and shallow wide minor groove in RNA. This architectural change makes major groove inaccessible while leaving greater access to uniform and shallow minor groove. This variability is accompanied by a large number of non-canonical base pairs, which introduces a new variety in structural organizations of the grooves. Apart from the regular helical regions with canonical and non-canonical base pairs, there are various structural motifs like pseudo helices, hairpin structural elements with variable loop regions, and internal loops, multi-way junctions, bulges, and many more. This wide range of possibilities enables RNA to get involved in RNA-RNA, RNA-protein interactions covering a broad range of functional aspect.

Nitrogen bases are hypothesized to have three edges for interaction, namely, Watson-Crick edge(W), Hoogsteen edge(H), and Sugar edge(S), out of which DNA utilizes only Watson-Crick edge forming double helices. RNA, on the other hand, makes use of all three types of edges along with local environment-dependent protonation for base pairing. A vast majority of base pairs in RNA are of non-canonical type. This finding fueled much research on the study of non-canonical

pairing, their characterization, nomenclature, classification, and modeling. Various computational approaches tried to detect these non-canonical as well as protonated pairs. Some of these, however, considered a single hydrogen bond in the formation of base pairs, which might not be as important in stabilizing RNA secondary structure as suggested by Saenger<sup>38</sup>. Our methodology in BPFIND<sup>88</sup> for base pair detection from three-dimensional structure considers at least two hydrogen bonds of N—H···O/N or C—H···O/N types.

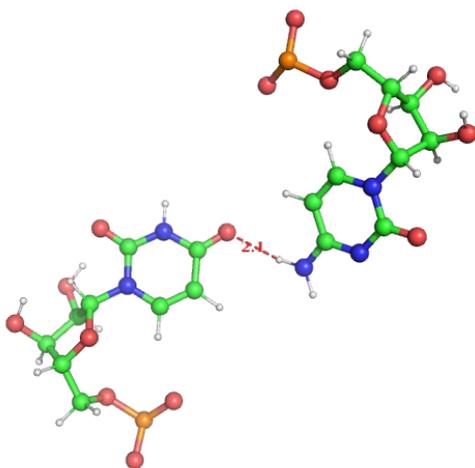
Interaction among the base-pair through hydrogen bond (HB) stabilizes a helical structure along with hydrophobic interaction of the medium. Many studies tried to figure out the interaction energies that stabilize different base pairings. Various approaches including experimental procedures like NMR, crystallography, and theoretical procedures like quantum mechanical (QM) computations of interaction and molecular dynamics (MD) simulation have already revealed many other important features of base pairing, such as pyramidalization of exocyclic amino groups, base pairs involving amino acceptor interactions, base pair involving ribose O2' as hydrogen bond donor and/or acceptor, base phosphate interactions, base pairing involving bifurcated and water-mediated hydrogen bonds, the role of protonation in base pairing, C-H···O/N hydrogen bonds in base pairing and of course stacking overlap. All of these studies emphasize the importance of the non-canonical pairs in the higher-order organization of RNA. In the perspective of structural modeling of RNA, these non-canonical pairs would prove to be indispensable. This information about the pairing is not enough for structure modeling. The rotational and translation parameters of intra and inter base pair orientations are found to be very crucial in this respect. These parameters are well studied and standardized for DNA but not explored to that extent for non-canonical pairs, which poses a potential problem towards modeling these structures. Our software, NUPARM, for analysis of these parameters using edge-specific axis system and associated

software, RNAHelix<sup>89</sup> for modeling RNA double helices containing non-canonical base pairs have successfully regenerated multiple structures having non-canonical pairs. However, this list was not exhaustive, which is the demand of the recent modeling paradigm and requires a thorough investigation and categorization of all possible non-canonical pairs.

The availability of a huge number of RNA crystal structures and their respective non-canonical base pairs demands organized structural management. Most of the currently available RNA structure databases like RNA STRAND<sup>90</sup>, FRABASE<sup>91,92</sup>, FR3D RNA motif Atlas<sup>93</sup>, SCOR<sup>94</sup>, most recently the RNA Bricks database<sup>95</sup>, concentrate on either secondary/tertiary structures or motifs and their respective classification. Few databases which provide exclusive information about non-canonical base pairs are RNABP COGEST<sup>96</sup> informing interaction energy-related information, NCIR<sup>97</sup> for non-canonical interactions and a parameter related to it, BPS<sup>98</sup> having base pair parameters but not up to date, and FR3D RNA base pair catalog available through webFR3D do provide all available pairs but not with the description of base pair parameters. These databases mainly focus on crystal occurrences of different base pair types, their structural annotation, and classification. Base pair steps and their step parameters alongside the base pairs are essential for modeling RNA structures. None of these databases, however, exhaustively emphasize this aspect. To our best knowledge, so far, no databases have ever attempted to quantify the stacking overlap of all possible base-pairs. Our group has recently developed a database of RNA base pairs (RNABPDB), showing different canonical and non-canonical base pairs' structural parameters. This database also provides structural information about the base paired dinucleotide steps, which are not found anywhere else. The base pair database addresses these issues and exhaustively paves the path for future RNA structure modeling using both canonical and non-canonical pairs. While, RNAHelix web server utilizes the available data of RNABPDB and

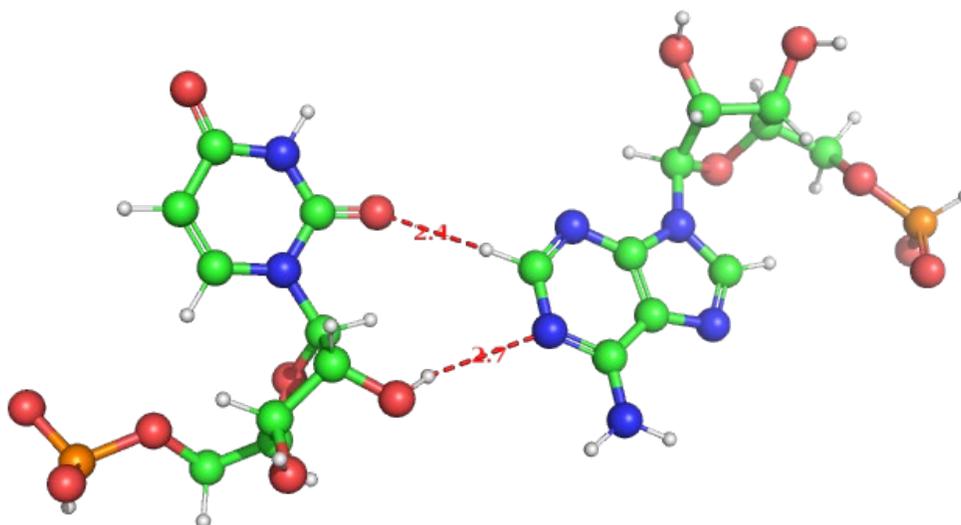
generates helical structure, helix with bulge comprising of both canonical and non-Canonical base pairs.

Analysis of structure, dynamics, and energetic stability of different non-canonical base pairs was attempted by several groups using different techniques<sup>61–65,67,99–103</sup>. Identification of different types of base pairs from the experimentally derived structures is also a non-trivial process. Several groups attempted to use different philosophies, and there are some software and servers for that, such as DSSR<sup>104</sup>, ClaRNA<sup>105</sup>, MC-Annotate<sup>106</sup>, BPFIND<sup>88</sup>, etc. Although algorithms for identifying base pairs are widely different, these programs mostly identify the same base pairs in large RNA structures with some exceptions. For example, DSSR and ClaRNA identify two bases as paired (residues 2868 and 2889 of 1N8R.pdb, for example, **Figure 2-1**) when they form a single hydrogen bond, while BPFIND expects at least two hydrogen bonds between the paired bases. (Like that there are 100 more base pairs detected by ClaRNA). Similarly, BPFIND identifies base pairs formed by hydrogen bonds involving a 2'-OH group of sugar, which is often not considered in ClaRNA (residues 608:R and 621:R of 1ASY.pdb, for example, **Figure 2**). These



**Figure 2-1** Base pair having residues 2868 (Chain A) and 2889 (Chain A) of 1N8R.pdb identified by ClaRNA but not by BPFIND. It shows the possibility of only one H-bond.

base pairs are often further stabilized by  $\pi$ - $\pi$  stacking interactions between successive base pairs, which was partially studied for a few Watson-Crick-like base pairs only<sup>107,108</sup>. Structural features of a canonical or non-canonical base pair can be best understood in IUPAC-IUB recommended orientation parameters<sup>39</sup>.



**Figure 2-2** Base pair having residues 608:R and 621:R of 1ASY.pdb identified by BPFIND. It shows H-bonding through 2'-OH of sugar.

The six intra-base pair parameters, namely Buckle, Open, Propeller, Stagger, Shear, and Stretch values for different types of A:G base pairs were studied recently<sup>109</sup> to understand the possible transition between the different types using 3DNA algorithm<sup>110</sup>. It is, however, difficult to characterize the base-pairing types from the parameter values obtained from 3DNA, particularly for non-canonical base pairs, as it gives an impression that these non-canonical base pairs are highly distorted with poor hydrogen bonds. The base pairing edge-specific axis system, as adopted in NUPARM algorithm<sup>111</sup>, allows one to understand the true deformation of a base pair from ideality. The NUPARM algorithm and software were evaluated for different DNA structures with Watson-Crick base pairs against several others, such as Curves, NEWHELIX, etc., and the performance was found suitable<sup>112,113</sup>. Furthermore, higher-order structures in RNA, including

those for base triplets, quartet<sup>114</sup>, can best be analyzed by NUPARM. Similarly, base pair stacking orientations between successive canonical or non-canonical base pairs were never analyzed in terms of their geometry and stability and can be analyzed. Furthermore, these base pair centric parameters are capable of model building RNA, which is extremely important now to understand the function of many non-coding RNA<sup>89,115</sup>.

In this work, a non-redundant RNA crystal structure dataset has been employed to study the intrinsic geometries of all canonical Watson-Crick and as well as non-canonical base pairs and the dinucleotide steps formed by them in the helical regions. We used BPFIND<sup>88</sup> to recognize the base pairs and higher-order structures in the experimentally derived RNA structures, as it is more stringent in the identification of base pairs with two or more hydrogen bonds comprising bases. We have employed NUPARM<sup>111,115,116</sup> for the calculation of all types of parameters related to nucleic acids. We have made a web server for such computation with an appropriate description. Outcomes of all the analyses for the non-redundant data are circulated through RNABPDB database. All the available RNA structures unraveled by NMR spectroscopy (**Table 2-1**) are also studied and tallied with crystal structure data. We have also computed interaction energies stabilizing each base pair, and the dinucleotide step by dispersion corrected density functional theory (DFT-D) wherever significant numbers of structures of the types are found in the database.

**Table 2-1** List of NMR derived structures of RNA used to populate the database of canonical and non-canonical base pairs

PDB IDs of NMR structures											
124D,	176D,	17RA,	1A1T,	1A3M,	1A4D,	1A4T,	1A51,	1A60,	1A9L,	1AC3,	1AFX,
1AJF,	1AJL,	1AJT,	1AJU,	1AKX,	1AL5,	1AM0,	1ANR,	1AQO,	1ARJ,	1ATO,	1ATV,
1ATW,	1AUD,	1B36,	1BAU,	1BGZ,	1BIV,	1BJ2,	1BN0,	1BVJ,	1BYJ,	1BYX,	1BZ2,
1BZ3,	1BZT,	1BZU,	1C00,	1C2Q,	1C4L,	1CQ5,	1CQL,	1CX5,	1D0T,	1D0U,	1D6K,
1DRR,	1DZ5,	1E4P,	1E95,	1EBQ,	1EBR,	1EBS,	1EFS,	1EHT,	1EI2,	1EJZ,	1EKA,
1EKD,	1EKZ,	1ELH,	1ESH,	1ESY,	1ETF,	1ETG,	1EXY,	1F5G,	1F5H,	1F5U,	1F6U,
1F6X,	1F6Z,	1F78,	1F79,	1F7F,	1F7G,	1F7H,	1F7I,	1F84,	1F85,	1F9L,	1FEQ,
1FHK,	1FJE,	1FL8,	1FMN,	1FNX,	1FQZ,	1FYO,	1FYP,	1G3A,	1G70,	1GUC,	1H0Q,
1HG9,	1HHW,	1HHX,	1HJI,	1HLX,	1HS1,	1HS2,	1HS3,	1HS4,	1HS8,	1HWQ,	1I3X,

1I3Y, 1I46, 1I4B, 1I4C, 1I9F, 1I9K, 1IDV, 1IE1, 1IE2, 1IK1, 1IKD, 1J4Y, 1JO7, 1JOX, 1JP0, 1JTJ, 1JTW, 1JU1, 1JU7, 1JUR, 1JWC, 1JZC, 1K1G, 1K2G, 1K4A, 1K4B, 1K5I, 1K6G, 1K6H, 1K8S, 1KAJ, 1KIS, 1KKA, 1KKS, 1KOC, 1KOD, 1KOS, 1KP7, 1KPD, 1KPY, 1KPZ, 1L1C, 1L1W, 1LC6, 1LDZ, 1LMV, 1LPW, 1LUU, 1LUX, 1LVJ, 1M5L, 1M82, 1ME1, 1MFJ, 1MFK, 1MFY, 1MIS, 1MNB, 1MNX, 1MT4, 1MUV, 1MV1, 1MV2, 1MV6, 1MWG, 1MY9, 1n32, 1N53, 1N66, 1N8X, 1NA2, 1NAO, 1NBK, 1NBR, 1NC0, 1NEM, 1NTQ, 1NTS, 1NTT, 1NXR, 1NYB, 1NZ1, 1O15, 1OKF, 1OO7, 1OQ0, 1OSW, 1OW9, 1P5M, 1P5N, 1P5O, 1P5P, 1PBL, 1PBM, 1PBR, 1PJJ, 1Q75, 1Q8N, 1QC8, 1QD3, 1QES, 1QET, 1QFQ, 1QWA, 1QWB, 1R2P, 1R3X, 1R4H, 1R7W, 1R7Z, 1RAU, 1RAW, 1RFR, 1RGO, 1RHT, 1RKJ, 1RNG, 1RNK, 1ROQ, 1RRD, 1RRR, 1S2F, 1S34, 1S9L, 1S9S, 1SCL, 1SLO, 1SLP, 1SY4, 1SYZ, 1SZY, 1T28, 1T2R, 1T4L, 1T4X, 1TBK, 1TFN, 1TJZ, 1TLR, 1TOB, 1TUT, 1TXS, 1U2A, 1U3K, 1U6P, 1ULL, 1UTS, 1UUD, 1UUI, 1UUU, 1VOP, 1WKS, 1WTS, 1WTT, 1WWD, 1WWE, 1WWF, 1WWG, 1XHP, 1XSG, 1XSH, 1XST, 1XSU, 1XV0, 1XV6, 1XWP, 1XWU, 1YFV, 1YG3, 1YG4, 1YLG, 1YMO, 1YN1, 1YN2, 1YNC, 1YNE, 1YNG, 1YSV, 1Z2J, 1Z30, 1Z31, 1ZBN, 1ZC5, 1ZIF, 1ZIG, 1ZIH, 219D, 28SP, 28SR, 2A9L, 2A9X, 2AD9, 2ADB, 2ADC, 2ADT, 2AHT, 2AP0, 2AP5, 2AU4, 2AWQ, 2B6G, 2B7G, 2BJ2, 2C06, 2CD1, 2CD3, 2CD5, 2CD6, 2CJK, 2D17, 2D18, 2D19, 2D1A, 2D1B, 2DD1, 2DD2, 2DD3, 2ERR, 2ES5, 2ESE, 2EUY, 2EVY, 2F4X, 2F87, 2F88, 2FDT, 2FEY, 2FY1, 2G1G, 2G1W, 2GBH, 2GIO, 2GIP, 2GM0, 2GRW, 2GV3, 2GV4, 2GVO, 2H49, 2HEM, 2HGH, 2HNS, 2HUA, 2I2Y, 2I7E, 2I7Z, 2IHx, 2IRN, 2IRO, 2IXY, 2IXZ, 2JPP, 2JQ7, 2JR4, 2JRG, 2JRQ, 2JSE, 2JSG, 2JTP, 2JUK, 2JWV, 2JXQ, 2JXS, 2JXV, 2JYF, 2JYH, 2JYJ, 2JYM, 2K3Z, 2K41, 2K5Z, 2K65, 2K66, 2K7E, 2K95, 2K96, 2KBP, 2KD4, 2KD8, 2KDQ, 2KE6, 2KEZ, 2KF0, 2Kfy, 2KG0, 2KG1, 2KGP, 2KH9, 2KHY, 2KM8, 2KMJ, 2KOC, 2KP3, 2KP4, 2KPC, 2KPD, 2KPV, 2KRP, 2KRQ, 2KRV, 2KRW, 2KRY, 2KRZ, 2KTZ, 2KU0, 2KUR, 2KUU, 2KUV, 2KUW, 2KVN, 2KWG, 2KX5, 2KX8, 2KXM, 2KXN, 2KXZ, 2KY0, 2KY1, 2KY2, 2KYD, 2KYE, 2KZL, 2L1F, 2L1V, 2L2J, 2L2K, 2L3C, 2L3E, 2L3J, 2L41, 2L5D, 2L5Z, 2L6I, 2L8C, 2L8F, 2L8H, 2L8U, 2L8W, 2L94, 2L9E, 2LA5, 2LA9, 2LAC, 2LAR, 2LB4, 2LBJ, 2LBK, 2LBL, 2LBQ, 2LBR, 2LBS, 2LC8, 2LDL, 2LDT, 2LDZ, 2LEB, 2LEC, 2LHP, 2LI4, 2LI8, 2LJJ, 2LK3, 2LKR, 2LP9, 2LPA, 2LPS, 2LPT, 2LQZ, 2LU0, 2LUB, 2LUN, 2LUP, 2LV0, 2LVY, 2LWK, 2LX1, 2M12, 2M18, 2M10, 2M1V, 2M21, 2M22, 2M23, 2M24, 2M39, 2M4Q, 2M4W, 2M57, 2M58, 2M5U, 2M8D, 2M8K, 2MB0, 2MEQ, 2MER, 2MF0, 2MF1, 2MFC, 2MFD, 2MFE, 2MFF, 2MFG, 2MFH, 2MGZ, 2MHI, 2MI0, 2MIS, 2MIY, 2MJH, 2MKI, 2MKK, 2MKN, 2MN0, 2MNC, 2MQO, 2MQP, 2MQQ, 2MQT, 2MQV, 2MS0, 2MS1, 2MS5, 2MTJ, 2MTK, 2MTV, 2MVS, 2MVY, 2MXJ, 2MXK, 2MXL, 2MXS, 2MXY, 2MZ1, 2N0J, 2N0R, 2N1Q, 2N2O, 2N2P, 2N3O, 2N3Q, 2N3R, 2N4J, 2N4L, 2N6S, 2N6T, 2N6W, 2N6X, 2N7C, 2N7M, 2N7X, 2N82, 2N8L, 2N8M, 2N8V, 2NBX, 2NBY, 2NBZ, 2NC0, 2NC1, 2NCI, 2NCQ, 2NCR, 2O32, 2O33, 2O81, 2O83, 2OJ7, 2OJ8, 2OOM, 2P89, 2PCV, 2PCW, 2PN9, 2QH2, 2QH3, 2QH4, 2RLU, 2RN1, 2RO2, 2RP0, 2RP1, 2RPK, 2RPT, 2RQC, 2RQJ, 2RRA, 2RRC, 2RS2, 2RSK, 2RU3, 2RU7, 2RVO, 2TOB, 2TPK, 2U2A, 2XC7, 2XEB, 2XFM, 2Y95, 2YH1, 3PHP, 484D, 4A4R, 4A4S, 4A4T, 4A4U, 4B8T, 4BS2, 4BY9, 4CIO, 5A17, 5A18, 5IEM, 5J0M, 5J10, 5J2W, 5KH8, 5KMZ, 5KQE, 5LSN, 5LWJ, 5M8I, 5MPG, 5MPL, 5N5C, 5N8L, 5N8M, 5OR0, 5U9B, 5UF3, 5UZT, 5UZZ, 5V16, 5V17, 5V2R, 5VH7, 5VH8, 5WQ1, 5X3Z, 5XI1, 6BY4, 6BY5, 6D2U, 6EZ0, 6GE1, 6GMY, 6MCE, 6MCF, 6MCI, 8DRH, 8PSH
---

These analyses of structure and strength of association of non-canonical base pairs are useful for understanding different modes of interactions in RNA but are also applicable for structure

modeling. We had recently developed RNAHelix<sup>89</sup> for the model building of RNA double-helical structures containing non-canonical base pairs, but its use was limited due to the unavailability of parameters for most non-canonical base pairs. It was seen that regenerated double-helical structures, from base pair parameters, superpose with the original structures generally with RMSD less than 0.2Å, indicating high accuracy of the RNAHelix algorithm. This software, as a web-server, has now been designed to obtain these parameters as mean values from the RNABPDB database. As a benchmark study, we have regenerated several double-helical RNA fragments using the consensus parameters from RNABPDB database and compared them with the original structures. There are several servers to predict and three-dimensional model structures of RNA, such as MC-Fold<sup>117</sup>, RNAPDBee<sup>118</sup>, RNAComposer<sup>119</sup>, SimRNA<sup>120</sup>, etc., but these either stitch structures of different motifs from RNA-FRABASE<sup>57</sup> or do Monte Carlo simulation to predict structures, while RNAHelix uses exact mathematical relations to regenerate. We found RMSD between original and regenerated double-helical structures are significantly smaller than those reported by the above servers. Our server, hence, is now able to do the model building of different pre-miRNA structures from their sequence and assuming different base pairing scheme for the predicted unpaired pairs of bases.

## **2.2 Methods**

We have built up a web-worker for the investigation and model building of RNA three-dimensional structures. There are three segments in this work – the BPFIND-NUPARM server, the RNABPDB database, and the RNAHelix server. The BPFIND-NUPARM server (<http://hdrnas.saha.ac.in/Tools/NUPARM>) is designed mainly for analyzing a single three-dimensional structure of RNA or DNA. The non-redundant list of RNA molecules from BGSU<sup>122</sup>(v2.122 with resolution  $\leq 3.0$  in May 2017) representing three-dimensional information

of RNA crystals were composed of PDB<sup>45</sup> and NDB<sup>47</sup> along with their biological assembly, where a single chain was present. Non-standard bases and other chemically modified bases were excluded in this work. The geometry based identification and cataloging of helical base pair structures was completed using BPFIND<sup>88</sup> program with default conditions. CIF files were taken in place of different PDB bundle files to encounter inter-chain interaction of large macromolecular assembly. NUPARM<sup>88,115,116</sup> program was employed with “-tor” “-pp” and “-ovl” options to compute six intra-base pair and six dinucleotide step parameters along with stacking overlap, C1'-C1' distance along backbone and all the backbone torsion angles in the base paired dinucleotide steps. All the base pair parameters and base pair step parameters are taken into account for the analysis after 3 $\sigma$  correction for greater reliability. Similar type of base pairs and their steps were grouped together with in-house scripts. The alternate configuration (A:U W:W C vs U:A W:W C) of the same pair were grouped together after necessary sign reversal of their respective pair parameters (Buckle and Shear for *cis* oriented pairs and Open and Stagger for *trans* oriented pairs). The equivalent dinucleotide pairs (G:A S:H T::G:C W:W C vs C:G W:W C::A:G H:S T) took the same strategy where the sign of Tilt and Shift needed mandatory adjustments. Representative structure for each of the pair type was found depending upon the E-value computed from BPFIND, and the planarity of the structure was determined from this. Besides, the representative of the dinucleotide steps was chosen based upon the step score ( $SS_{ab}$ ) that takes both the planarity as well as the stacking interaction between the pairs.

$$SS_{ab} = \frac{Overlap_{ab}}{Eval_a + Eval_b}$$

Where  $Overlap_{ab}$  denotes the stacking overlap between pair  $a$  and pair  $b$ . The representative pairs and pair steps are presented via the Jmol HTML5 plug-in<sup>123</sup>. As positions of the hydrogen atoms are usually unavailable in the best representative structures of the base pairs or base paired

dinucleotide steps, pymol<sup>124</sup> was used for the modelling purpose considering standard geometry and restrained geometry optimization of the hydrogen atoms using *opt=modredundant* option of Gaussian16<sup>125</sup> software with  $\omega$ B97X-D<sup>82</sup> functional and cc-pVDZ<sup>126,127</sup> basis set were done. Finally, Basis Set Superposition Error (BSSE)<sup>86</sup> corrected interaction energies of these base pairs, and base paired steps have been computed considering  $\omega$ B97X-D/cc-pVDZ. BSSE corrected stacking energy for the base pair steps have also been computed by considering each base pair as one monomer for obtaining total stacking energy, i.e., the sum of inter-strand and cross-strand interaction energies. MySQL maintains background database while the front end is taken care of by php. It is worth mentioning that software development and database management parts were done by my lab-mate, Dr. Debasish Mukherjee and details work is enlisted in his thesis (***Structure and Dynamics of Secondary Structural Motifs in Noncoding RNA***). The underline backbone of the RNAHelix web server is based upon RNAHelix standalone program. Base pair parameters and base pair step parameters are obtained from the database to generate the structure, while the sugar-phosphate backbone is incorporated with CHARMM<sup>128</sup>. The detail about the methodology is described earlier<sup>89</sup>. Necessary implementation of cross-talk between the database and server is done through perl script.

## 2.3 Result and Discussion

### 2.3.1 Base pair in the database

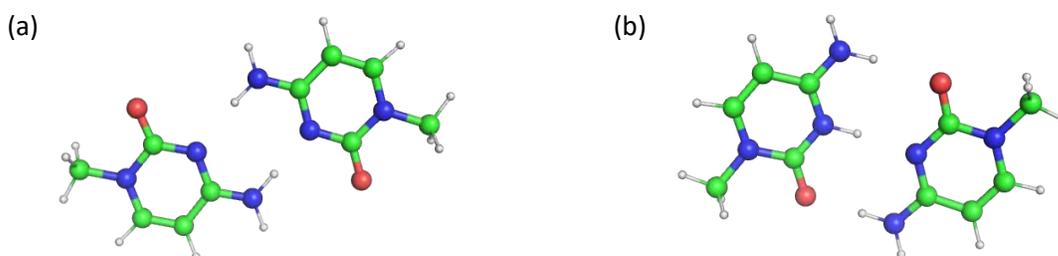
We have identified 72552 base pairs classified in 108 distinctive sorts and have analyzed their structures in terms of numerous valuable parameters. We have recognized and examined the stacking of these base pair with their neighbors in double helix in terms of their parameters. These are available in the open-access RNABPDB database (<http://hdrnas.saha.ac.in/rnabpdb>), giving each detailed information. As there are varieties within the parameters for the diversity in

The screenshot shows the RNABPDB database interface. At the top left is a 'Menu' button. The main header reads 'RNA Basepair Database'. A search bar is located at the top right. Below the header, a green banner states 'A database of RNA base pairs and their stacks, as detected by BPFIND in list of non-redundant RNA structures'. A dropdown menu is open, showing 'W:W:C' as the selected option. Below this, there are buttons for nucleobases: Ade, Gua, Cyt, Ura. A grid of buttons represents different base pair types: A:A W:W:C, A:G W:W:C, A:C W:W:C, A:U W:W:C, G:G W:W:C (highlighted in gray), G:C W:W:C, G:U W:W:C, C:C W:W:C, C:U W:W:C, and U:U W:W:C. To the right of the grid is a diagram of a base pair stack showing two Watson-Crick (W:C) pairs and one Wobble (W:W:C) pair, with 'SUG' labels indicating sugar-sugar interactions. At the bottom, there are tabs for 'Overview' and 'References', and a 'Database Overview' section with descriptive text.

**Figure 2-3** The home page of RNABPDB database, showing how to navigate from one service to another. The drop-down menu indicating other available applications appear by clicking Menu (shown by border) and the drop-down menu indicating other non-canonical base pairing patterns appear by clicking the region shown by border. Any base pair, such as A:A W:W:C, can be clicked to get detail information about the base pair, its stabilizing energy, distributions of structural parameters and how it stacks on other base pairs. The base pairs shown by gray patch are impossible ones.

base pairs and their stacks, I would talk about the striking highlights of some features in this report. The “Home page” of the database appears all potential Watson-Crick and Wobble base pair (**Figure 2-3**), which are classified as W:W:C, from where the client can select points of interest of other classes by pull-down menu. The G:G W:W:C base pair and few others are

impossible as in no way two bases can form two hydrogen bonds with each other through their corresponding edges in those circumstances. It may be noted that ClARNA illustrates many G:G contacts, however, with wide variations of their approaches, representing no stable mode of recognition between them. We tend to find a few base pairing types comprising of two dissimilar hydrogen bonding modes. As an example, C:C W:WT base pairing is feasible in two ways: (a) with protonation of one of the Cyt bases, giving rise to three hydrogen bonds (C:C W:+T or termed as W:PT in the database), that is taken into account because the nucleotide is useful for stabilizing I-motif<sup>129</sup> and (b) N4-H and N3 of two Cyt residues involve in two N—H...N hydrogen bonds that stabilize the sheared C:C W:WT base pair (**Figure 2-4**). Similarly, additional eleven types of base pairs are identified where two bases interact involving the same edges but with different sets of hydrogen bonds.



**Figure 2-4** Structures of two alternative forms of Cyt:Cyt base pairs involving their Watson-Crick base pairing edges in *trans* orientation for (a) in neutral form stabilized by two N—H...N hydrogen bonds and (b) in hemi protonated form stabilized by three hydrogen bonds.

It has been found that 64% base pairs belong to canonical type, and the remaining 36% belong to non-canonical class. Out of all non-canonical base pair sets, 52.68% are involved in *cis* orientation whereas 47.32% comprised *trans* orientation. Base pairs comprising C-H...O/N hydrogen bonds (H-bond) are also observed (20.9%). Base pairs arising from sugar edge often display the direct participation of the 2'-OH group in Hydrogen bond formation. The base pair's construction is somewhat dissimilar from the sugar edge base pairs having no direct involvement of the 2'-OH

group. 28 out of 43 sugar edge mediated pairs involve the direct participation of the 2'-OH group for Hydrogen bond formation. A:G H:ST is the most frequently found non-canonical base pair. This base pair also participate in the various biological role that will be discussed in the next chapter.

Base pairs are energetically stable due to the formation of at least two hydrogen bonds that help them be co-planar with each other and play a significant role in constructing the double helix. We have chosen a suitable representative of each base pair from the database where sufficient data are present to know the base pairs' energetic stabilization. First, we have optimized the position of hydrogen atoms by DFT-D method using Gaussian16. After that, the BSSE corrected interaction energies between the H-optimized base pairs have been computed using DFT-D, and all the values are given in the database. We detected good interaction energy originated from the H-bonding in all the non-canonical base pairs. The hydrogen bond involves in  $n \rightarrow \sigma^*$  transition, which results in elongation of N-H/O-H bonds. Likewise, The N-H bonds in base pairing, get always elongated (**Table 2-2, Figure 2-4**).

**Table 2-2** Change in bond length for NH—H, N—H, C—H and O—H after participation in H-bonding. Optimized Bond distance in isolated free nucleotide bases:  $d(\text{NH—H})_{\text{Adenine}}=1.011 \text{ \AA}$ ,  $d(\text{NH—H})_{\text{Guanine}}=1.011 \text{ \AA}$ ,  $d(\text{N—H})_{\text{Guanine/Uracil}}=1.015 \text{ \AA}$ ,  $d(\text{C—H})=1.094 \text{ \AA}$ ,  $d(\text{O—H})=0.963 \text{ \AA}$

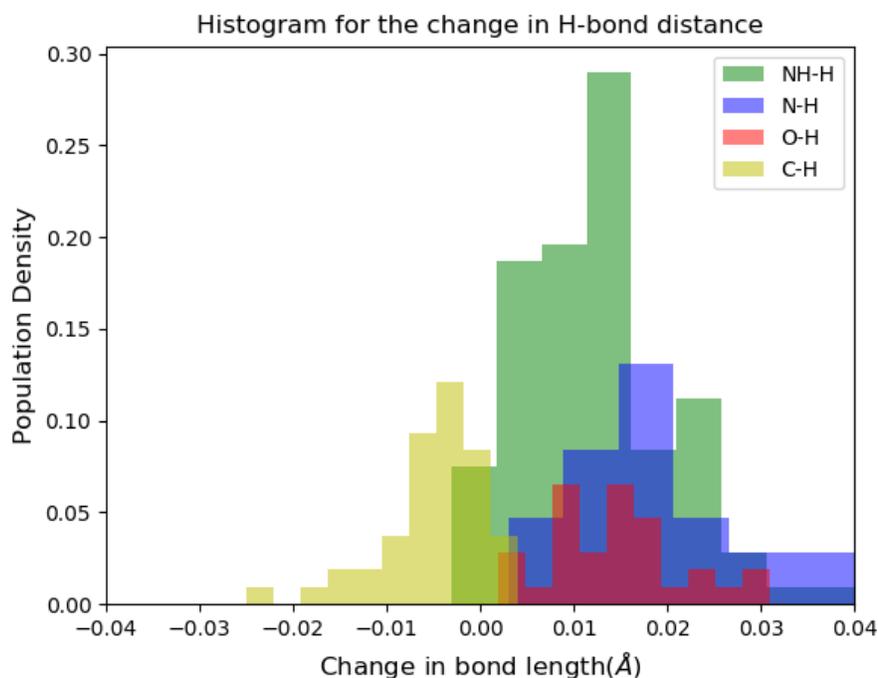
Base pair	Frequency Of Base pair	Frequency of Base pair in Stack	Base pair Interaction energy (kcal/mol)	Change in Bond length after H-bonding					
				$\Delta d(\text{NH—H}) \text{ \AA}$	$\Delta d(\text{NH—H}) \text{ \AA}$	$\Delta d(\text{N—H}) \text{ \AA}$	$\Delta d(\text{N—H}) \text{ \AA}$	$\Delta d(\text{C—H}) \text{ \AA}$	$\Delta d(\text{O—H}) \text{ \AA}$
AA hhC	14	0	-7.82	0.005	-	-	-	-0.003	-
AA HHT	633	15	-10.88	0.004	0.007	-	-	-	-
AA hsT	256	7	-5.53	0.015	-	-	-	-0.001	-
AA hwC	1	0	-1.23	-	-	-	-	-	-
AAHWT	444	8	-13.54	0.011	0.012	-	-	-	-
AASHC	117	0	-4.98	0.005	-	-	-	-	-
AAssC	182	6	-11.22	-	-	-	-	-0.003	0.029
AASWC	385	5	-15.08	0.013	-	-	-	-	0.015
AAswT	45	2	-4.4	0.015	-	-	-	-0.005	-
AAwwC	115	8	-6.3	0.016	-	-	-	-0.001	-
AAWWT	381	5	-14.31	0.018	0.015	-	-	-	-
AChsT	29	0	-14.56	-	-	-	-	-0.002	0.028
AChwC	5	0	-6.7	0.001	-	-	-	-0.001	-
ACHWT	531	5	-16.32	0.011	0.015	-	-	-	-
ACssC	1233	2	-9.92	-	-	-	-	0	0.019
ACSWC	147	0	-17.34	0.014	-	-	-	-	0.014
ACwsT	80	0	-14.56	-	-	-	-	0.002	0.031
ACWWT	87	8	-16.91	0.015	0.018	-	-	-	-
AGhhC	1	0	-8.27	0.006	-	-	-	-0.002	-
AGhwT	7	1	-9.46	0.016	-	-	-	-0.009	-
AGPHC	18	1	-45.51	0.021	0.037	-	-	-	-
AGssC	406	4	-14.83	-	-	-	-	0	0.02
AGSSC	379	3	-13.84	0.005	-	-	-	-	0.015

AGssT	1716	4	-5.4	0.011	-	-	-	-0.003	-
AGSWC	3	0	-10.06	-	-	-	0.009	-0.008	0.008
AGWST	858	4	-14.41	0.016	0.01	-	-	-	-
AGwsT	615	3	-14.29	0.002	-	-	-	0	0.014
AGWWC	810	7	-19.48	0.024	0.016	-	-	-	-
AGwwT	18	1	-11.07	0.016	-	-	-	0.002	-
AGzhC	10	0	-33.48	0.022	-	-	-	-0.002	-
AUhhC	16	0	3.03	0.012	-	-	-	-0.016	-
AUhsT	144	1	-11.22	-	-	-	-	-0.012	0.024
AUswC	111	1	-7.5	0.022	-	-	-	-0.005	-
AUswT	87	0	-10.53	0.027	-	-	-	-0.003	-
AUWSC	166	4	-15.51	0.006	-	-	-	-	0.023
AUWWT	395	4	-14.43	0.008	-	-	0.024	-	-
CAhsC	27	0	-3.55	0.011	-	-	-	-0.017	-
CAPWT	2	1	-20.21	-0.003	-	-	0.058	-	-
CASHC	64	0	-9.26	0.002	-	-	-	-	0.004
CASWC	533	4	-14.66	0.007	-	-	-	-	0.011
CAWPC	182	8	-42.6	0.025	-	-	0.033	-	-
CAwwC	58	7	-8.48	0.019	-	-	-	0.004	-
CChsT	65	2	-10.33	0.012	-	-	-	-	0.004
CChwT	15	0	-10.23	0.011	-	-	-	-0.005	-
CCshC	15	3	-11.02	-0.003	-	-	-	-0.007	-
CCSWC	35	1	-4.49	0.01	-	-	-	-	0.019
CCwhC	170	0	-9.22	0.004	-	-	-	-0.004	-
CCWPC	47	7	-42.67	0.032	-	-	0.039	-	-
CCWPT	17	2	-47.5	0.009	-	-	-	-0.025	-
CCWWT	17	2	-23.7	0.025	0.022	-	-	-	-
CGhhC	9	0	-10.36	0.005	-	-	-	-0.005	-
CGhhT	28	0	-8.07	0.002	-	-	-	-0.014	-
CGPHC	8	1	-47.49	0.027	-	-	0.042	-	-
CGPWC	90	6	-35.39	0.01	-	-	0.062	-	-
CGSSC	172	0	-15.27	0.005	-	-	-	-	0.01

CGWSC	47	1	-16.43	0.009	-	-	-	-	0.017
CGWST	95	0	-17.94	0.013	0.019	-	-	-	-
CGWzC	4	0	-40.78	0.045	-	-	0.014	-	-
CUhsT	34	1	-5.62	0.002	-	-	-	-0.006	-
CUPWC	17	0	-27.12	-	-	0.042	0.007	-	-
CUSWT	12	0	-14.05	0.021	-	-	-	-	0.016
CUwhC	6	3	-7.28	0.015	-	-	-	-0.001	-
CUWSC	80	3	-7.05	0.005	-	-	-	-	-
CUWWT	11	0	-12.06	0.015	-	-	0.03	-	-
GAHPT	15	0	-46.08	0.023	-	-	0.034	-	-
GASHT	3685	24	-17.38	0.005	0.008	-	-	-	-
GAswC	360	0	-8.33	0.016	-	-	-	0.003	0.002
GAWHC	56	0	-17.86	0.014	-	-	0.017	-	-
GAwsC	30	0	-12.06	-	-	-	0.016	-0.005	-
GCHPT	26	0	-44.11	0.028	-	-	0.033	-	-
GCWSC	52	4	-9.81	0.001	-	-	0.011	-	-
GCWST	110	0	-24.72	-	-	-	0.014	-	0.01
GCWWC	35156	55	-31.36	0.02	0.013	-	0.016	-	-
GCWWT	229	0	-15.98	0.009	0.013	-	-	-	-
GGHWC	463	3	-15.96	0.003	-	-	0.0075	-	-
GGsHT	27	3	-14.69	0.014	-	-	-	-	0.009
GGSSC	80	0	-5.34	0.008	-	-	-	-	0.018
GGsST	239	2	-13.2	0.013	0.014	-	-	-	-
GGWHT	130	2	-19.76	0.007	-	-	0.013	-	-
GGWSC	16	0	0.71	0.017	-	-	0.015	-	-
GGWWT	41	3	-26.16	0.02	-	-	0.019	-	-
GGzHT	13	2	-46.12	0.022	-	-	0.027	-	-
GUSSC	148	0	-12.12	0.004	-	-	-	-	0.009
GUWSC	65	4	-12.95	0	-	-	0.005	-	0.007
GUWWT	43	5	-16.83	-	-	0.013	0.02	-	-
UAhsT	8	0	-11.76	-	-	-	-	-0.003	0.015
UAhwC	19	1	-0.96	0.015	-	-	-	-0.011	-

UASHC	108	3	-15.1	0.004	-	-	-	-	0.013
UAssC	559	3	-10.86	-	-	-	-	-0.002	0.019
UAswT	329	3	-7.99	-	-	-	-	-0.005	0.013
UAWHC	528	2	-15.65	0.01	-	-	0.023	-	-
UAWHT	2148	20	-10.37	0.001	-	-	0.019	-	-
UAWWC	11147	53	-14.59	0.008	-	-	0.023	-	-
UCWSC	25	0	-0.72	0.022	-	-	-	-	0.014
UCWWC	24	0	-14.36	0.013	-	-	0.021	-	-
UGhsC	141	5	-12.31	0.014	-	-	-	0	-
UGhwT	346	0	-10.56	0.018	-	-	-	0.001	-
UGSWT	38	1	-14.58	0.001	-	-	0.006	-	-
UGwhC	26	2	-13.64	0.023	-	-	-	-0.005	-
UGwhT	80	0	-14.52	-	-	-	0.027	-0.004	-
UGWSC	39	2	-13.99	0.012	-	-	0.019	-	-
UGWST	5220	17	-16.35	0.011	-	-	0.018	-	-
UGWWC	6	0	-3.31	-	-	0.025	0.013	-	-
UUhwC	47	0	-8.37	-	-	-	0.017	-0.004	-
UUSWT	22	0	-14.29	-	-	-	0.013	-	0.009
UUwhT	97	0	-6.86	-	-	-	0.003	-0.008	-
UUWSC	20	0	-6.25	-	-	-	0.012	-	0.01
UUWWC	693	22	-12.42	-	-	0.016	0.02	-	-
UUWWT	103	3	-10.63	0.019	-	0.016	0.016	-	-

Interaction energies for Non-protonated and non-canonical base pairs where base pairing occurs through Watson-Crick edge such as A:G W:WC, U:U W:WC etc., have been estimated as -10 kcal/mol to -28 kcal/mol. The non-canonical G:A sheared base pair (G:A S:HT), considering the sugar unit of the Guanine, shows better stability (-17 kcal/mol) than that of G:U W:WC base pair (-16 kcal/mol) which indicates that strength of non-canonical base pair is comparable with canonical base pairs. There are 48 base pairs where one Hydrogen bond is mediated by either C-H...O or C-H...N atoms whose interaction energies are below -10 kcal/mol. These kinds of hydrogen bonds often are termed as anti-hydrogen bonds where we have observed C-H bond contraction (**Table 2-2**) as seen in this base pair set<sup>130,131</sup> (**Figure 2-5**).

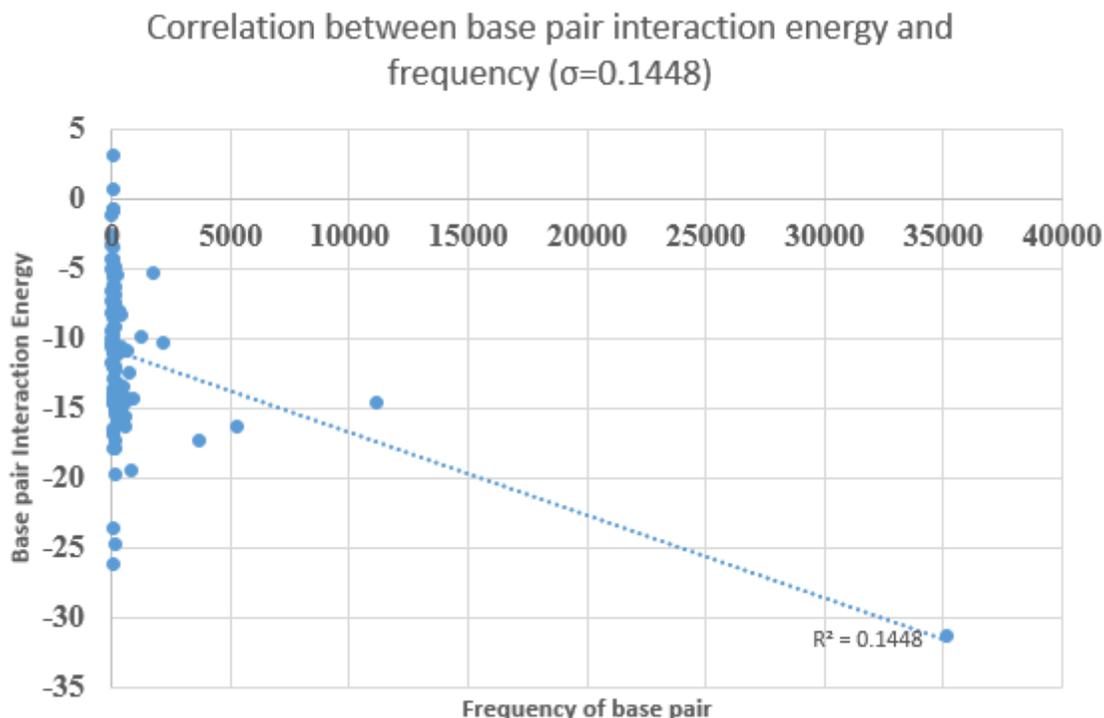


**Figure 2-5** Histogram for change in  $X-H$  covalent bond length after  $H$ -bonding,  $X$  being primary amino group ( $NH-H$ ), secondary amino group ( $N-H$ ), hydroxyl group ( $O-H$ ) and weak non-polar group ( $C-H$ ).

The 13 protonated base pairs with protons in their Watson-Crick and Sugar edges estimate strong interaction energy, which vary from -30 kcal/mole to -50 kcal/mole. An increase in O2'-H

bond length also enhances the additional proof that ribose sugar unit involves in appropriate cases (**Table 2-2, Figure 2-5**).

It is expected that the energetically highly stable base pairs would appear more frequently in the crystal and NMR structures. Earlier quantum chemical studies could not show such correlation as the number of data of different types of base pairs were not sufficient, and the earlier structural database studies were not based on a sufficiently large number of non-redundant structures. We now have enough data for most base pairs, although some base pairs are not appropriately represented. Our analysis, as shown in **Figure 2-6**, indicates that such a correlation is absent. However, the points (frequency and energy data) for the protonated base pairs are most poorly represented as their energies are quite high, but they appear quite infrequently. This is probably due to the fact that protonation is also energetically unfavorable<sup>132</sup>. Hence, we removed the points corresponding to the protonated base pairs and analyzed the data again. This gives rise to a correlation coefficient of 0.144, which is statistically significant. We expect this correlation to improve further the availability of more experimental data for all the possible non-canonical base pairs. Furthermore, we calculated the interaction energies of the best representative structures from each base pair, which may be affected by other interactions within the RNA structures.

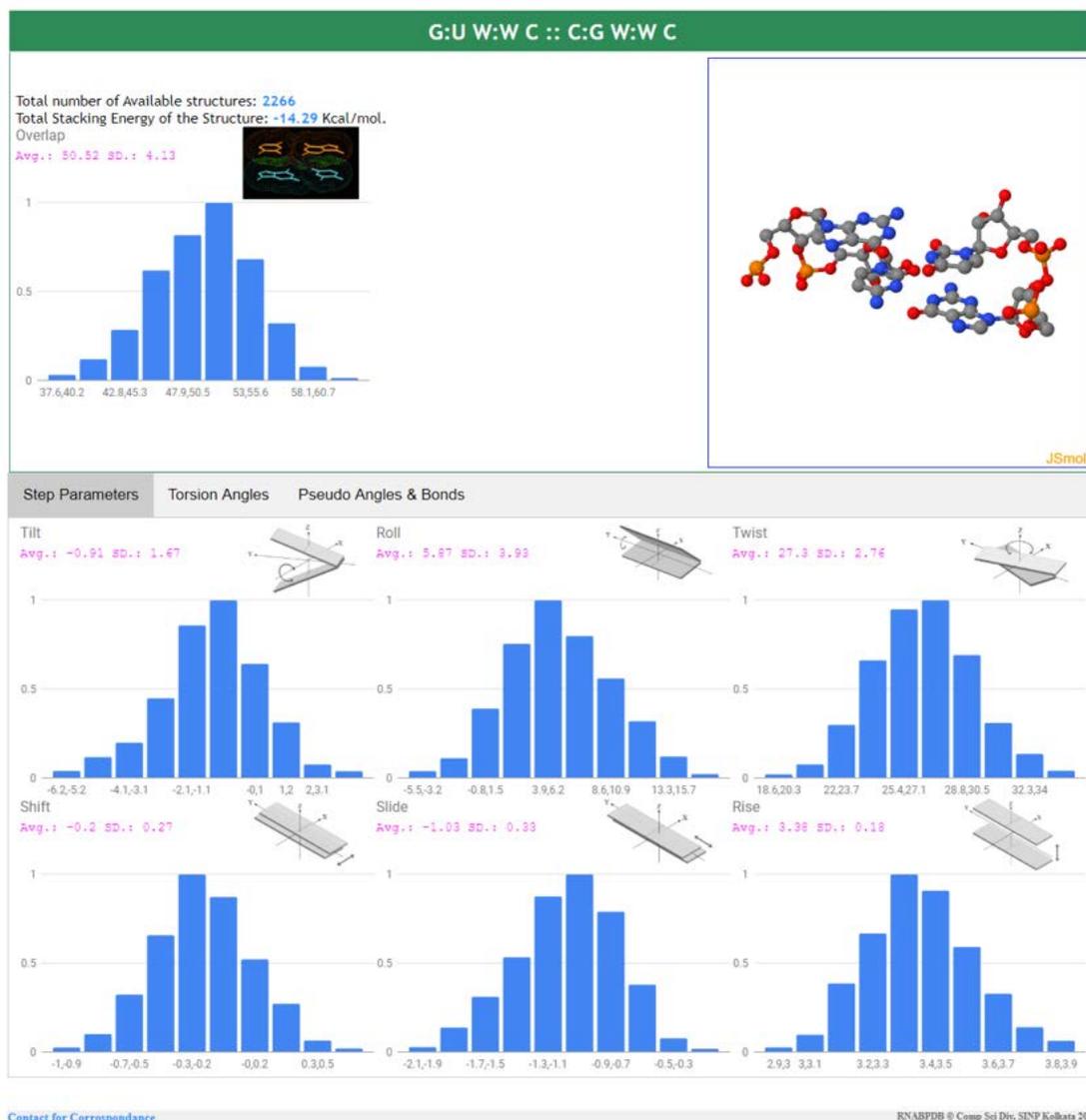


**Figure 2-6** Correlation between non protonated base pair interaction energy and their frequency

### 2.3.2 Dinucleotide Step Analysis

According to the energetics data and statistics of base pair step parameters indicating coplanarity of the base pairs, most of these base pairs can stack on other base pairs leading to the formation of RNA double-helical structure. Configuration of the dinucleotide steps formed by Watson-Crick as well as non-canonical base pairs has been identified from the available RNA structures. It has been seen that an enormous number of dinucleotide steps are formed entirely by Watson-Crick base pairs, e.g., by G:C W:WC, A:U W:WC and G:U W:WC (**Figure 2-7**). As for example, we found 15556 dinucleotide stacks consisting of 11147 A:U W:WC base pairs and G:C W:WC, G:U W:WC, and another A:U W:WC base pair, which also indicate some of the A:U W:WC base pairs are obviously stacked from both the 5'- and 3'-ends. We also detected 1083 base paired dinucleotide structures formed by A:U W:WC and a non-canonical base pair of different types. The mean dinucleotide step parameters for the Watson-Crick base paired configuration are

quite comparable to previous reports<sup>107,108</sup>. Along with the structures of the dinucleotide steps comprising of only Watson-Crick base pairs, structural features of the dinucleotide steps involving non-canonical base pairs have been also studied. The stacking overlap values between two consecutive base pairs vary mostly from 40 to 60 Å<sup>2</sup>, regardless of whether these are Watson-Crick or non-canonical base pairs, as found previously<sup>116</sup>. These structures still do not possess base pair orientation parameters comparable to those in usual A-RNA structures. Among the base pair step parameter, Slide and Twist are maximally perturbed due to presence of a non-canonical base pair. For instance, mean Twist of G:G W:HC::G:C W:WC (‘:’ indicates base pairing and ‘::’ indicates stacking) dinucleotide step is 47°, Twist of G:A S:HT::A:U H:WT is found to be 84°, Twist of A:G H:ST::G:C W:WC is found to be 7.7°. However, these unusual parameters do not specify bad configuration as stacking overlaps display high values suggesting robust interactions between the base pairs. Furthermore, when the frequency of occurrence of the dinucleotide steps is not negligible, none express bimodal distributions. It results in small standard deviations of the base pair step parameters, and hence, one can proceed with the mean values of the orientation parameters for molecular modeling of double helix comprising non-canonical base pairs.



**Figure 2-7** Representative figure showing distributions of different parameters indicating good stacking between a C:G Watson-Crick base pair and a G:U Wobble base pair.

Stabilization of RNA double-helical structure mainly originates from stacking interaction. We have initially calculated BSSE corrected interaction energy, which includes base pair interaction and stacking energy (Table 2-3). Computation of base pair step interaction energy considers each base as one component in the counterpoise method for BSSE. Consideration of

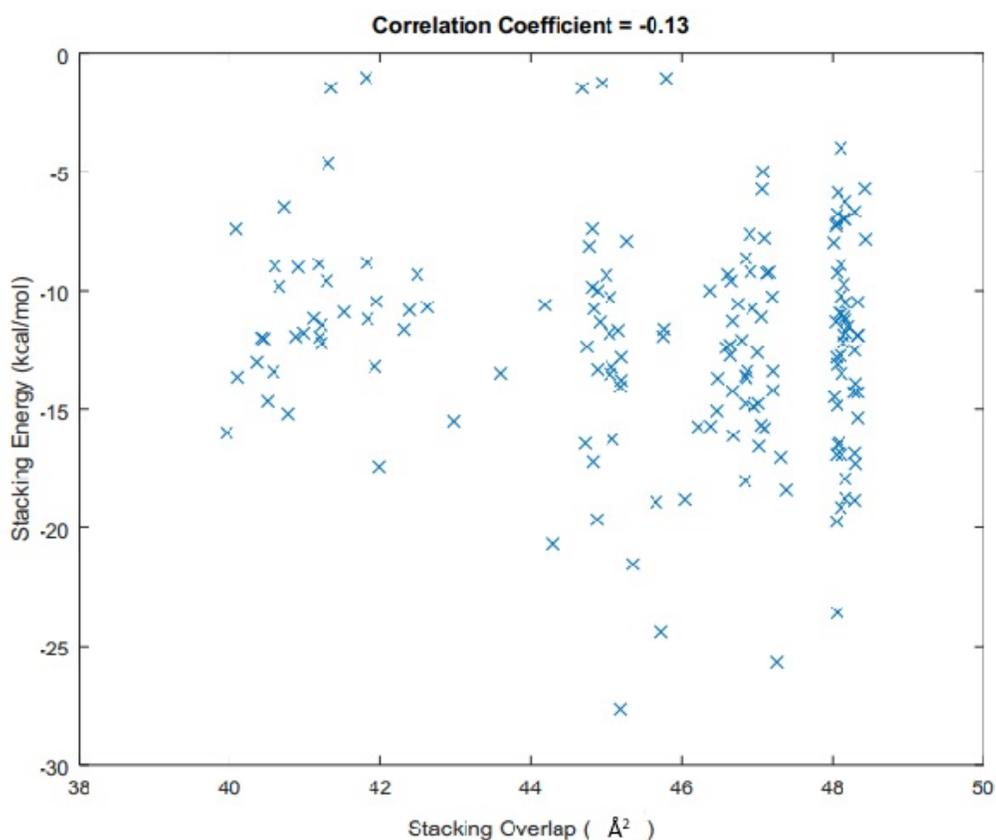
each base pair as one component in the counterpoise method for BSSE gives us only stacking energy and these are given in our database.

**Table 2-3** BSSE corrected Interaction energy and Stacking Energy for some of the important dinucleotide steps.

Dinucleotide Step Sequence	Interaction Energy (kcal/mol)	Stacking Energy (kcal/mol)
G:C W:WC::C:G W:WC	-79.94	-18.77
C:G W:WC::G:C W:WC	-75.18	-17.29
C:G W:WC::C:G W:WC	-66.58	-7.84
A:G W:WC::G:C W:WC	-66.53	-18.42
G:A W:WC::C:G W:WC	-62.14	-16.51
A:G W:WC::C:G W:WC	-60.65	-17.03
G:U W:WC::C:G W:WC	-57.17	-14.29
G:U W:WC::G:C W:WC	-56.74	-14.29
U:A W:WC::G:C W:WC	-55.41	-11.51
G:C W:WC::U:A W:WC	-55.02	-11.87
A:G H:ST::G:C W:WC	-62.84	-25.69
A:G H:ST::G:U W:WC	-38.14	-16.57
A:G H:ST::U:A W:WC	-35.84	-14.90
A:A w:wC::G:U W:WC	-29.19	-14.81
A:C +:WC::C:G W:WC	-91.02	-24.03

We have also checked the correlation between Stacking Overlap and stacking energy. This study is attributing the less correlation to the fact that the energy values are from representative structures but not optimized structures (**Figure 2-8**). Furthermore, optimization of four unlinked bases is difficult and is possible by parameter scans only. This would also strengthen your detail stacking energy study by varying relevant structural parameters. It is worth mentioning that given stacking overlap values are the average of all stacking overlap values for each distinct set of dinucleotide stacks and calculation of stacking energy considers the best representative stacks for each set of dinucleotide stacks. It may be the reason for the low correlation between stacking overlap and stacking energies. High Roll value results in close proximity between two base pair, which causes high stacking overlap. Besides, it introduces Van der Waals (VDW) repulsion, which weakens stacking interaction. It imparts a negative correlation between stacking overlap and stacking

energy. To pursue a systematic study, we categorized base pair steps into four groups – canonical, non-canonical and protonated base pair. Stacking energies of canonical base pair step dictate good stability. It is worth mentioning that canonical base pair steps U:G W:WC::G:U W:WC and G:U W:WC::U:G W:WC have stacking energies of

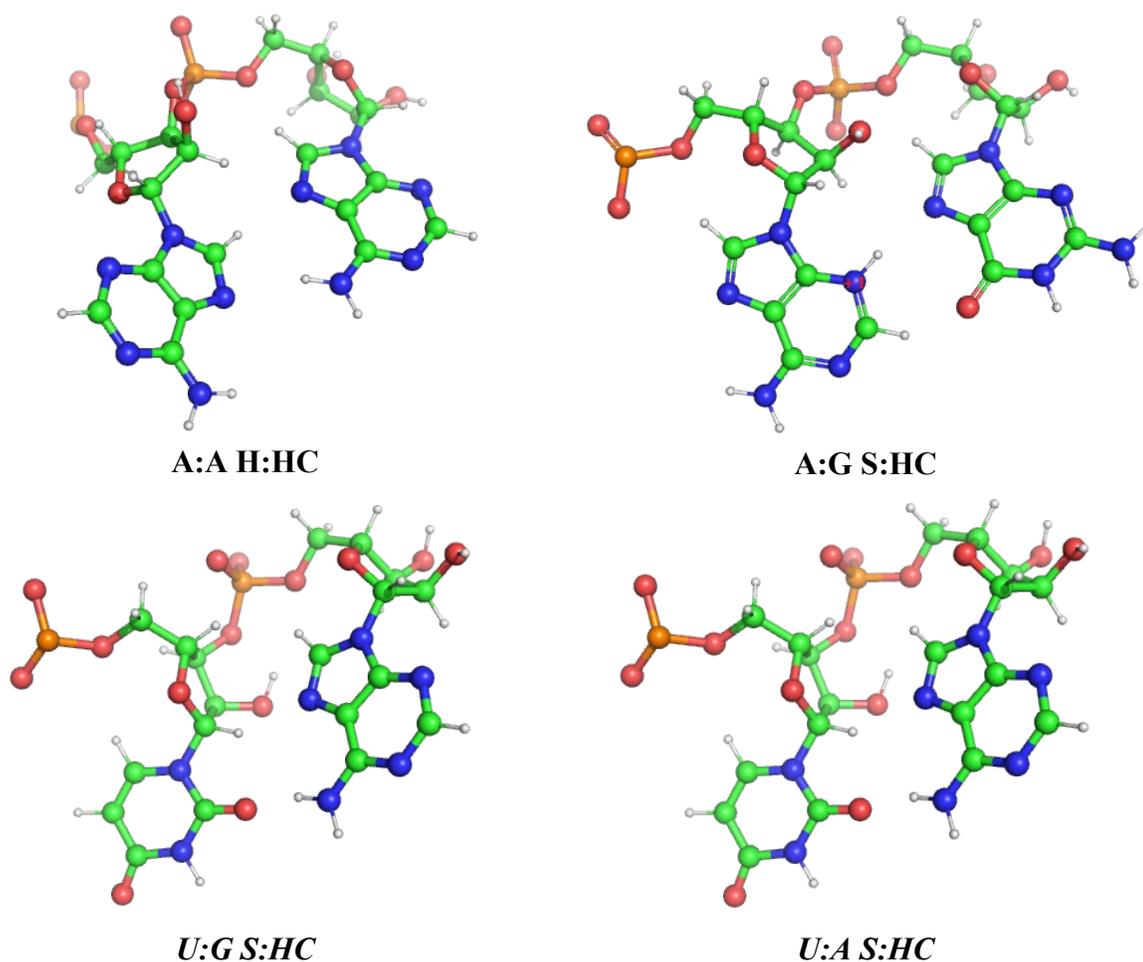


**Figure 2-8** Correlation between stacking interaction and stacking overlap for all dinucleotide step sequence having frequency more than 5.

-5.86 kcal/mol and -10.49 kcal/mol. Large difference in stacking energy is mainly due to higher stacking overlap of G:U W:WC::U:G W:WC base pair step mediated by shearing motion of G:U base pair. It has been well established from past studies that non-canonical base pair steps are playing miscellaneous function in RNA. It indicates relative stability of non-canonical and canonical base pair step. Our database also supports the fact. For example stacking interaction of dinucleotide step A:G H:ST::G:C W:WC is -25.69 kcal/mol, which is higher than that of any

canonical base pair step. Another striking observation emerges out from the A:U H:WC and A:U H:WT base pair. Occurrence of these two base pairs are 528 and 2148, respectively. Though we have sufficient A:U H:WC base pairs, none of them can stack with other base pair. A:U H:WT base pair stacks with other base pair whose orientation is only *trans*. Stacking energy analysis may reveal the reason behind A:U H:WT on *trans* base pair during stacking. Several studies have been done on protonated base pair in RNA motif in low pH environment<sup>133,134</sup>. Our database has reported 15 distinct protonated base pair steps having occurrence more than 5. For example, dinucleotide step A:C +WC::C:G W:WC has been found 81 times in X-ray crystal structures, and the best representative structure estimates -26.30 kcal/mol stacking interaction. Other protonated base pair steps also acquire high stacking interaction. Protonation in the one of the aromatic ring adds strengths to cation- $\pi$  interaction<sup>135</sup> which may be the reason behind the high stacking interaction in the protonated base pair steps.

It may be noted that all strong, stable, and planar base pairs are not probably suitable for double helix formation. We said most of the identified A:A H:HC, A:U H:SC, A:G S:HC and G:U S:HC base pairs are between two consecutive residues connected by sugar-phosphate backbone (**Figure 2-9**). Such dinucleotide platforms are an integral part of GAAA receptor motifs, as detected by FR3D<sup>136</sup> but may not form a double helix.



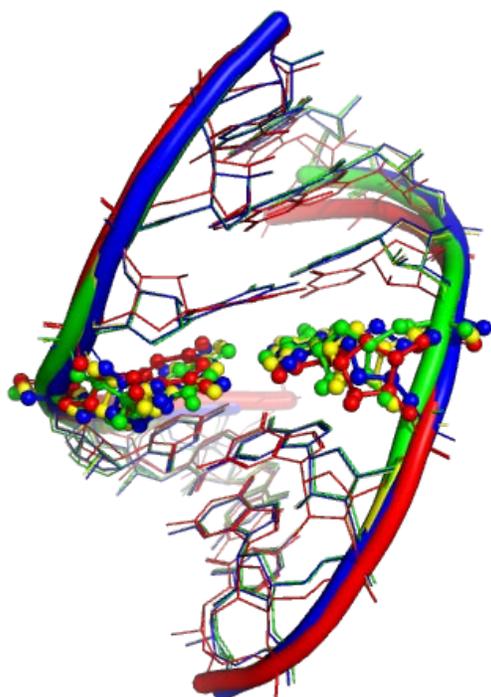
*Figure 2-9 Representative three dimensional view of dinucleotide platforms*

## 2.4 Molecular Modeling of RNA double helix

Molecular modeling of RNA structures containing non-canonical base pairs is an important aspect in understanding different biochemical processes by functional RNAs. The non-canonical base pairs' structural parameters were not available earlier, but those for most of the base pairs are now available in the RNABPDB database. These can now be used to construct molecular models of double-helical stems of RNA using RNAHelix software. We have used the server for benchmark study to regenerate several double-helical fragments whose crystal structures are available. To confirm quality of the models we have used base pairing information, as obtained from BPFIND,

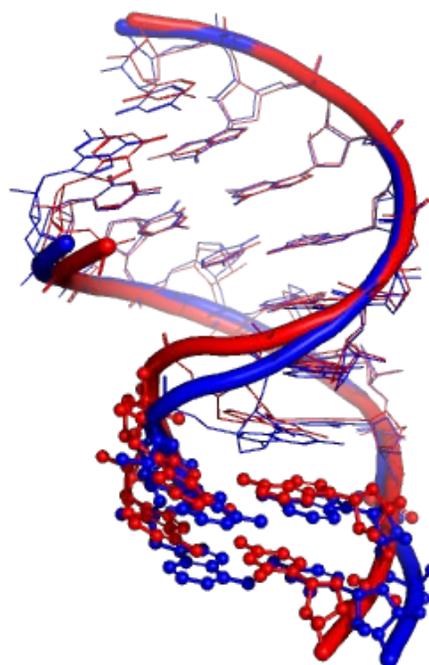
for several double-helical fragments of different length from different experimentally derived structures and regenerated the structures using RNAHelix server. These structures were superposed on the experimental ones to calculate RMSD between the two (**Table 2-4**). The double-helical fragment of PDB ID 1J5A contains a U:U cWW base pair, while the RNABPDB database shows a bimodal distribution of its Shear values. Hence, we generated three models (**Figure 2-10**) considering three Shear values and found reasonable RMSD with the original crystal structure for all the models. It may be noted that the Shear value of the U:U base pair in the original structure is 1.5Å, significantly different from either mode in the distributions (**Figure 2-10A**). It is seen that RMSD between base atoms (including C1' atoms of sugar, which also remain in the same base plane) are generally smaller than 0.5Å. In case of 4V9R\_bundle4, the mean Twist value of G:A tSH::A:A tSH (G:A S:HT::A:A s:hT) was obtained from RNABPDB as -0.71, which is unsuitable for RNAHelix. We have slightly changed this value to a small positive Twist (0.70) and regenerated, but obtained a slightly larger RMSD (**Figure 2-10D**). As the sugar-phosphate atoms are generated by restrained energy minimization, there are chances of trapping into local minima, which possibly give rise to slightly larger RMSD values. It may be noted that the other available RNA structure generation procedures generate three-dimensional structures in two steps, (i) by predicting secondary structure using dynamic programming algorithm based on free-energy data and (ii) by obtaining the three-dimensional structure of the predicted one. The dynamic programming algorithms generally do not predict secondary structures containing non-canonical base pairs. In that procedure, the servers also attempt to predict structures of turns, internal or hairpin loops, which are single-stranded and highly flexible. We do not perform the first step in our approach, but our approach is equally suitable for generating three-dimensional structures of double helices containing non-canonical base pairs. Thus, it is now possible to extend our method

with the appropriate method for predicting RNA secondary structures containing non-canonical base pairs.



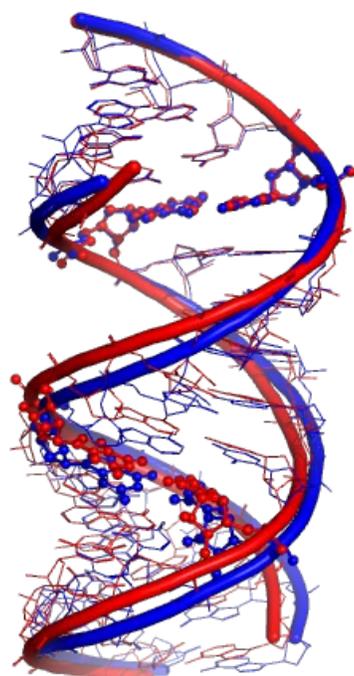
**(A) PDB: 1J5A**

Red: Experimental Crystal Structure,  
Green: Regenerated Structure with  
positive Shear of U:U W:WC, Green:  
Regenerated Structure with negative  
Shear of U:U W:WC Green: Regenerated  
Structure with Zero Shear of U:U W:WC



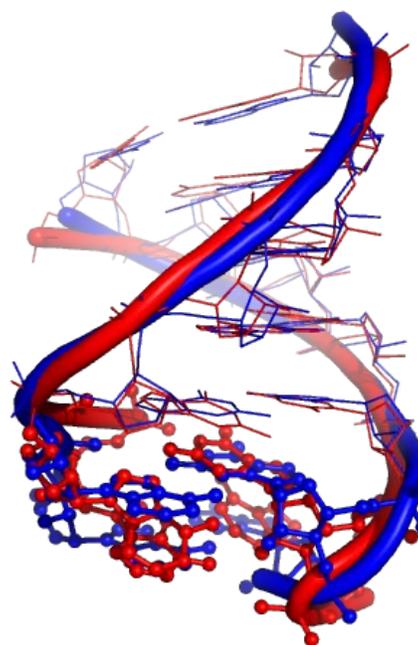
**(B) PDB: 1N32-b(bio Assembly)**

Red: Experimental Crystal Structure,  
Blue: Regenerated Structure



**(C) PDB: 3ND3**

Red: Experimental Crystal Structure,  
Blue: Regenerated Structure



**(D) PDB: 4V9R-bundle4**

Red: Experimental Crystal Structure,  
Blue: Regenerated Structure

**Figure 2-10** Aligned Regenerated Structure on experimental crystal structure. Ball and Stick representation indicated non-Canonical base pair.

**Table 2-4** Comparison of regenerated RNA double-helical structures with the experimentally determined structures.

PDB ID	Reason For Selection	Residue numbers ( Chain ID in parenthesis)	RMSD (Å) with respect to regenerated Structure	
			Base Atoms	Whole Double helix
259D	Double helix with Watson-Crick base pairs	1-8(A) 9-16(B)	0.409	0.991
2V7R	Double helix with Watson-Crick base pairs	1-7(A) 66-72(B)	0.491	1.159
2VUQ	Double helix with Watson-Crick base pairs	1-7(A) 66-72(B)	0.435	1.137
3GVN	Double helix with Watson-Crick base pairs	1-7(A) 66-72(B)	0.790	1.216

*Non-canonical base pairs and their stabilities*

3HGA- Bio Assembly-1	Double helix with Watson-Crick base pairs	3-6(X) 6-3(X)	0.363	0.918
435D	Double helix with Watson-Crick base pairs	1-7(A) 8-14(B)	0.577	1.148
4U3L	Double helix with Watson-Crick base pairs	1-8(A) 9-16(B)	0.951	1.820
5L00	Double helical regions with Watson-Crick base pairs, discarding the single stranded residues	5-12(A) 5-12(B)	0.617	1.179
5TDJ-Bio Assembly-1	Double helix with Watson-Crick base pairs	1-10(A) 1-10(A)	0.501	1.535
5UED	Double helix with Watson-Crick base pairs	4-13(A) 4-13(B)	0.497	1.181
5V0J	Double helix with Watson-Crick base pairs	5-12(A) 5-12(A)	0.567	1.178
1N33	Double helical fragment with G:U cWW base pair at center	1421–1431 (A) 1479–1469 (A)	0.584	1.260
2G3S	Double helix with G:U cWW at center	101-108(A) 109-116(B)	0.660	1.195
3ND3- Bio Assembly-1	Double helix with two G:U cWW base pairs	1-16(A) 17-32(A)	0.292	1.267
3SJ2	Double helical region with two G:U W:WC base pairs and three G:G cHW base pairs	3-19(A) 3-19(B)	0.903	1.636
4PCO	Double helix with two G:U cWW base pairs	1-10(A) 1-10(B)	0.848	1.689
1FJG	Double helical region with A:G cWW base pair at center	1409–1416 (A) 1491–1484 (A)	0.511	1.359
2Q1R	Double helix with two G:A cWW base pair at center	1001-1012(A) 1012-1001(A)	0.802	2.066

1N32-a	Double helical fragment with A:A tSH (s:hT), A:U tHW (H:WT) and A:G tHS (H:ST) base pairs at center		778–786 (A) 804–796 (A)	0.618	1.154
1J5A	Double helical fragment with U:U W:WC base pair at center	Negative Shear of U:U cWW	2066–2072 (A) 2215–2209 (A)	0.710	1.169
		Positive Shear of U:U		0.615	1.088
		Zero Shear of U:U		0.629	1.070
1N32-b	Double helical fragment with A:A tWH (W:HT) and C:A tWH (W:HT) base pair at one terminal		1241–1249 (A) 1296–1288 (A)	0.368	1.348
1XMQ	Double helical fragment with A:U tHW (H:WT), A:G tHS (H:ST) and G:A tSH (S:HT) base pairs at termini		439–448 (A) 495–486 (A)	0.405	1.001
2AZX	Double helical fragment with A:A cWS (W:SC) base pair at one terminal		510–514 (C) 525–521 (C)	0.667	1.043
3L3C	Double helical fragment with G:G tWW (W:WT) base pair at center		75-89 (B) 108-94 (B)	0.953	2.202
3R1C	Double helix with G:G cWH (W:HC) base pairs at center		1–8 (Q) 8–1 (R)	0.364	1.401
4V4Q	Double helical fragment with G:A tSH and A:G tHS base pairs at center		533–543 (B) 560–550 (B)	0.728	1.699
4V88-bundle2	Double helical fragment with G:A tSH A:G tHS and A:A tHH base pairs		1645–1656 (A) 1810–1799 (A)	0.582	1.976
4V9R-bundle4	Double helical fragment with G:A tSH and A:A tSH base pairs at one terminal		1198–1204 (A) 1247–1241 (A)	1.174	1.468

Non-canonical base pairs and their stabilities

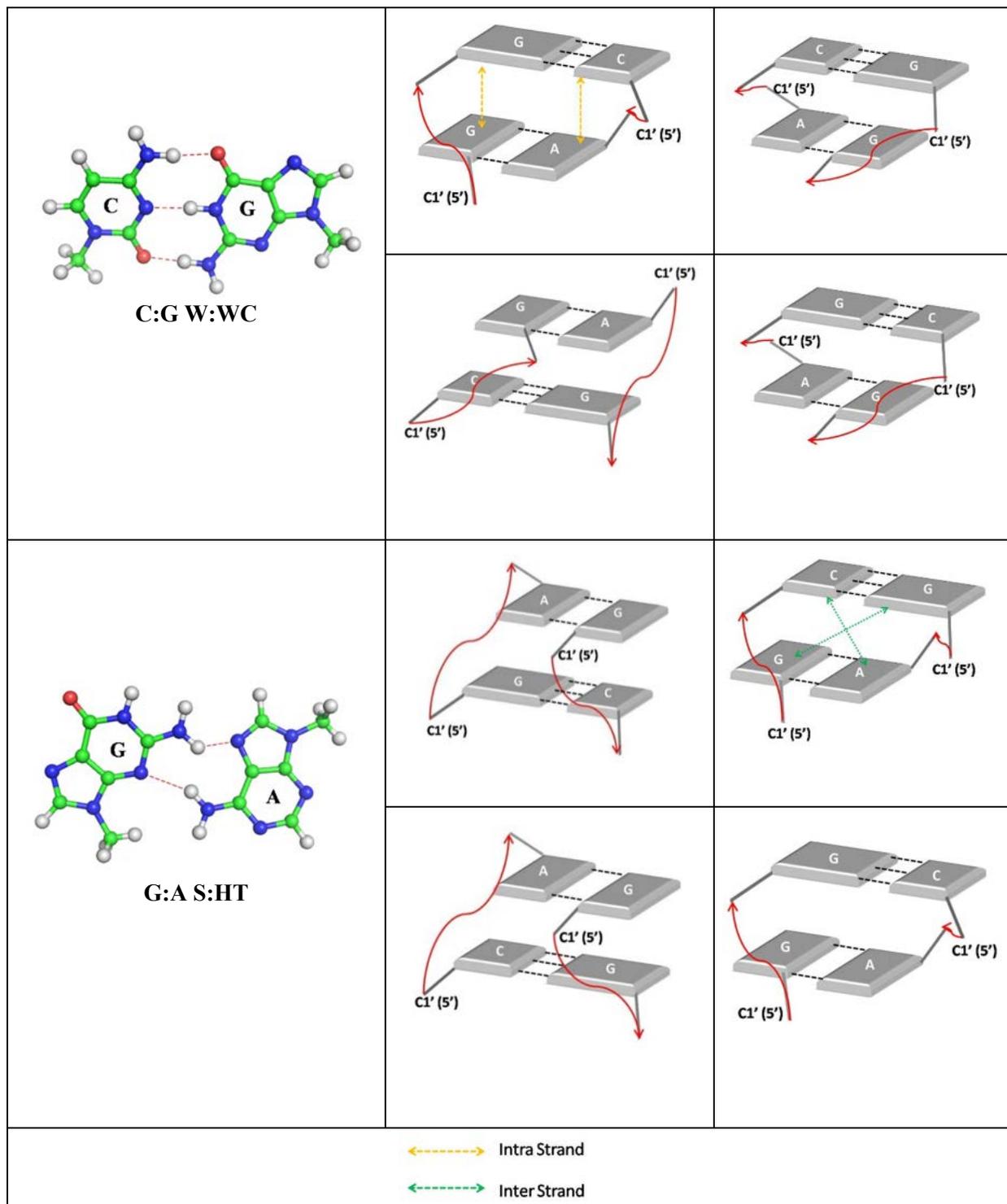
5DM6	Double helical fragment with C:A cSW (S:WC) base pair at terminal	3–13 (Y) 122–112 (Y)	0.606	1.580
5J7L-bundle3	Double helical fragment with G:A tSH, A:U tHW and A:A tHW base pairs at one terminal	150–161 (d) 176–165 (d)	0.527	1.807
3CZW-Bio Assembly-1	Double helix with two G:G cHW base pairs	3-16(X) 16-3(X)	0.412	2.444
5NXT-Bio Assembly-1	Double helix with two G:U cWW and one A:C cWW (+:WC) base pairs	30-47(U) 30-47(U)	0.514	1.729

## **2.5 Conclusion**

As the pace in structure determination of RNA is exceptionally high now, it is expected that the RNABPDB database can be updated soon with a sufficient number of statistically meaningful data for all types of non-canonical base pairs. Moreover, computational chemistry methods, which were seen to predict suitable structures of base pairs and base paired dinucleotide steps, can be used to obtain optimized values of the parameters for non-canonical base pair containing dinucleotides. This database revealed that G:A S:HT base pair as one of the most frequent non-canonical base pair. It stacks with C:G W:WC and G:C W:WC with adequate frequency. These dinucleotide step sequences may be used as a reference template to develop a method to theoretically predict the proper configuration of a dinucleotide step containing a non-canonical base pair. RNABPDB database also makes known to us about A:A w:wC base pair which has two modes of H-bonding with the same stability, which leads to a bimodal distribution of its Shear value. It prefers positive Shear when it stacks under U:G W:WC (A:A w:wC :: U:G W:WC) 36 times. Whereas the frequency of A:A w:wC base pair is found to be only 4 when it stacks under G:U W:WC (A:A w:wC :: G:U W:WC). The latter dinucleotide step prefers negative Shear. The above data indicate the promiscuous nature of A:A w:wC base pair with sheared G:U W:WC base pair. This information compels us to do a further high-level investigation about stacks of A:A w:wC base pair. It may help us understand the nature of other base pair like U:U W:WC with a bimodal distribution of Shear.

## **Chapter- 3 Stacking Interaction by most Frequent A:G base pair: Establishment of Hybrid DFT-D method**

### 3.1 Introduction



**Figure 3-1** Possible dinucleotide step containing C:G W:WC and G:A S:HT.

As indicated in the earlier chapter that the occurrence of A:G base pairs in helical termini of 90% bacterial rRNA and metal binding site of hammerhead ribozymes are also found to be very frequent<sup>137,138</sup>. Sheared A:G base pair, which involves Hoogsteen edge of A and Sugar edge of G in trans orientation regarding the hydrogen bonding direction named as A:G H:ST<sup>88</sup> (**Figure 3-1**), has been witnessed as second most frequently in the non Watson-Crick base pair family whereas G:U base pair are seen most frequently<sup>139</sup>. This unusual A:G H:ST base pair closes the GNRA tetraloop (N is any nucleotide and R stands for purine base), a distinct building block in the viewpoint of structure-function relationship in RNA because this type of hairpin loop mediates interaction with broad range of receptor.

The major groove of DNA double helices made by Watson-Crick base pairs are enriched in base sequence dependent variability to position the functional groups. Thus the major groove of B-DNA is vital for molecular recognition of DNA or base sequence reading by gene regulatory proteins. As the narrow major groove is unable to accommodate any protein, molecular recognition uses the shallow minor groove. Watson-Crick base pairs in the minor groove of double helices are unable to deliver the base sequence directed variations of functional groups<sup>140</sup>. The shallow minor groove of both A:U or U:A W:WC base pairs consist of only two symmetrically placed hydrogen bond acceptors whereas minor groove of both G:C or C:G base pairs possess a centrally located hydrogen bond donor along with two symmetric hydrogen bond acceptors. The grooves arising from the RNA double helices comprising of only Watson-Crick base pairs, therefore does not deliver adequate deviation of functional groups by various base sequence. The non Watson-Crick base pairs possibly progress variability in disposition of hydrogen bonds for proper molecular recognition. For example, the A:G H:ST base pair, allows specific recognition of RNA double helices and it appears very frequently in these double helices<sup>141</sup>.

Multiple Hydrogen bonds are required not only for the stabilization of base pair but also for the planarity of the base pairs. Hydrogen bonding interaction is the key for base pair data, is commonly electrostatic interaction which enable us to understand it using classical force-field based methods comprising of coulomb interaction between atoms. There are many studies on computation of geometry and energetics of the Watson-Crick and as well as non Watson-Crick base pairs with various quantum chemical approaches such as HF, DFT, hybrid DFT, MP2, etc representing that many of the non Watson-Crick base pairs are comparable to Watson crick base pair in order to judge their stability<sup>60,142,143</sup>. Base stacking interaction stabilizes double helical arrangements of nucleic acids. It was suggested that perhaps base stacking interaction may impart more leading role over base pairing interaction in the double helical formation<sup>144</sup>. Various investigation shows that ca,  $45\text{\AA}^2$  surface area of the base pairs are used in the stacking interactions<sup>116</sup>. Such adequate surface overlapping area between planar aromatic base in the two base pairs provide broad interactions between extended  $\pi$  orbitals of the aromatic bases. Along with the  $\pi\dots\pi$  interactions some other interactions for example, electrostatic, X-H... $\pi$  etc., are also important. As  $\pi\dots\pi$  interactions are electronic phenomena, it somewhat outweighs stacking overlap area. It results requirement of cutting-edge quantum chemical approaches for the computation of stacking energy. Extensive studies have been executed to understand the dependency of nucleic acid double helices formation on stacking energy by various approaches<sup>75,77,145-149</sup>. Estimation of stacking free energy for different base pair step from the melting temperature of double helical nucleic acid has been reasonably effective for several purposes, for example, PCR primer design<sup>150,151</sup>. Such thermodynamic investigation of double helical structures comprising of single or tandem mismatches have also been done<sup>152-154</sup>. However these experiments do not offer adequate explanation on whether the two mispaired bases make any

base pair or they persist like symmetric unpaired bulges. It is worth mentioning that Adenine and Guanine can generate 16 different configuration of base pairing stabilized through two hydrogen bonds<sup>139</sup>. These experimental investigations fail to disclose base pairing pattern i.e., which interacting edges of the bases are used in base pairing. As per above discussion, quantum chemical calculations for getting the proper electronic configuration of the bases and base pairs are needed to compute energetic preferences of stacking. Density functional theory is well-known to offer a decent estimation of energy. Still the DFT functionals are unable to compute the dispersion energy/interaction even between molecular systems<sup>155,156</sup>. It was also revealed recently that dispersion interaction, which follows  $-C_6/R^6$ , is very important to provide stabilization energy between canonical or non-canonical base pairs<sup>149</sup>.

As no previous study has been carried out on stacking energy estimation of dinucleotide sequence involving non-Watson crick base pair, we have concentrated on dinucleotide sequences formed by the most frequently observed non canonical base pair A:G H:ST and the most frequently observed canonical base pair C:G W:WC. There are eight possible ways for these two base pairs to stack on each other (**Figure 3-1**). Four out of the eight are identical to remaining four except for differences in the strand directions. Out of the four unique dinucleotide steps, two, namely C:G W:WC::G:A S:HT and G:C W:WC::G:A S:HT (‘:’ indicates base pairing and ‘::’ indicates stacking) have been found quite frequently in the available RNA structure database. We have therefore attempted to understand stacking geometry of these sequences which can be compared with experimental data. Considering RNA bases as rigid blocks, we can represent base pair geometry by six intra-base pair parameters. Similarly, considering a base pair as a rigid block, we can represent geometry of a base paired step by six inter-base pair step parameters. Thus, a complete representation of a base paired dinucleotide step requires 18 independent parameters

(base pair and base pair step parameters). Thus, stacking interaction can be analyzed by six inter-base pair step parameters - tilt, Roll, Twist, shift, Slide and rise as suggested by IUPAC-IUB convention<sup>39,73</sup>. Among these Roll, Slide and Twist only show sequence directed preferences. The three other parameters, namely tilt, shift and rise do not show significant variability and these do not depend on sequence of dinucleotide step<sup>149,157</sup>. Hence we have characterized stacking interaction for different geometries of the above mentioned dinucleotides by varying Roll, Slide and Twist. We have scanned the energy surface using dispersion corrected density functional theory (DFT-D) by  $\omega$ B97X-D<sup>82</sup> with cc-pVDZ<sup>127,158,159</sup> basis set, which is fairly acceptable for calculating  $\pi\dots\pi$  interaction<sup>160,161</sup>. The energy contours are compared with observed structural parameters from X-ray crystallographic data. A Consideration of sugar phosphate backbone by implicit and coarse grain method gives high similarity between stacking energy and experimental observation.

## **3.2 Material and Methods**

### **3.2.1 Crystal Structure Database Analysis:**

High resolution non redundant dataset of RNA crystal structures with resolution better than 3.0 Å from BGSU RNA site<sup>122</sup> (Released 2.91 database) are taken to represent experimental data points. We have found out 38393, 11671 and 3916 base pairs having C:G W:WC, A:U W:WC and G:A S:HT sequence. We have found 2687, 1168 and 647 base paired dinucleotide steps having A:U W:WC::C:G W:WC, C:G W:WC::G:A S:HT and G:C W:WC::G:A S:HT sequences, respectively, as detected by BPFIND<sup>88</sup> software. We have calculated all the base pair and base pair step parameters from the crystal dataset for the corresponding sequences using NUPARM<sup>111,115</sup>. We have also calculated distances between C1' atoms along each strand for these experimental structures of the two dinucleotide sequences. We have calculated stacking overlap<sup>116</sup> values

between the two successive base pairs by NUPARM for all the dinucleotide base pair steps of the given sequences.

### *3.2.2 Base pair and Base pair Step Modeling*

We have modeled C:G W:WC, A:U W:WC and G:A S:HT base pairs by RNAHelix<sup>89</sup> software using average base pair parameters as obtained from X-ray crystal structures. We have also modeled C:G W:WC::G:A S:HT, G:C W:WC::G:A S:HT and A:U W:WC::C:G W:WC base paired dinucleotide steps by RNAHelix<sup>89</sup> software using assigned intra-base pair and inter-base pair step parameter. In the cases of dinucleotide steps we have considered 0°, 1.12°, -15°, 0Å, -0.07Å and 2.8Å values for Buckle, Open, Propeller, Stagger, Shear and Stretch, respectively, for G:C W:WC base pair, 0°, 1.12°, -15°, 0Å, 0.07Å and 2.8Å values for Buckle, Open, Propeller, Stagger, Shear and Stretch, respectively, for C:G W: WC base pair and 0°, -13.20°, -10°, 0Å, 2.23Å and 3.10Å values for Buckle, Open, Propeller, Stagger, Shear and Stretch, respectively, for G:A S:HT base pair. These are the average values of the intra-base pair parameter as calculated by NUPARM<sup>111</sup> on all the non-redundant RNA structures with few exceptions in buckle and propeller for consistency with double helix formation. We have considered 0.00°, 0.00°, -15.00°, 0.00Å, 0.00Å and 2.80Å for Buckle, Open, Propeller, Stagger, Shear and Stretch, respectively, for A:U W:WC and C:G W:WC base pair, 0.00°, 0.00°, -10.00°, 0.00Å, 0.00Å and 2.80Å for Buckle, Open, Propeller, Stagger, Shear and Stretch, respectively, in A:U W:WC::C:G W:WC base paired dinucleotide steps. We have generated structures for each dinucleotide base pair step by changing Twist value between 5° to 85° in steps of 5°. We have generated 99 models for each Twist value by changing Roll value between -20° to +20° in steps of 5° and Slide value between -2.5Å to +2.5Å in steps of 0.5Å. We have kept constant values to the remaining insensitive parameters: tilt = 0.0°, shift = 0.0 Å, rise = 3.4 Å. We have varied Twist

between 5° and 50° for A:U W:WC::C:G W:WC dinucleotide base pair step. In all these cases, sugar-phosphate backbone was represented by methyl groups attached to N9 (of purines) or N1 (of Cytosine) atoms.

### 3.2.3 Computational Methods

We have optimized the model base pair structures in few ways: (i) free optimization of all the atoms to lowest energy and (ii) optimization of only the hydrogen atoms of the base pairs constraining the non-hydrogen atoms. We have calculated interaction energy of C:G W:WC, A:U W:WC and G:A S:HT base pairs from different conformation using  $\omega$ B97X-D<sup>158</sup> DFT-D functional with cc-pVDZ<sup>126,127,158</sup> and aug-cc-pVDZ<sup>126</sup> basis set by Gaussian09<sup>162</sup>. We have calculated total interaction energy as  $\Delta E_{\text{opt}}^{62,64,163}$  as-

$$\Delta E_{\text{opt}} = E_{\text{inter}} + E_{\text{def}} \quad \dots 1$$

For a Base paired system  $E_{\text{inter}}$  and  $E_{\text{def}}$  are calculated as –

$$E_{\text{inter}} = E_{\text{A:B}}^{\text{opt}} - E_{\text{A}}^{(\text{opt-A:B})} - E_{\text{B}}^{(\text{opt-A:B})} \quad \dots 2$$

$$E_{\text{def}} = (E_{\text{A}}^{(\text{opt-A:B})} - E_{\text{A}}^{\text{opt}}) + (E_{\text{B}}^{(\text{opt-A:B})} - E_{\text{B}}^{\text{opt}}) \quad \dots 3$$

Where  $E_{\text{A:B}}^{\text{opt}}$ ,  $E_{\text{A}}^{\text{opt}}$ ,  $E_{\text{B}}^{\text{opt}}$  are the optimized energies of the paired bases (A:B) and individual bases (A and B), respectively. Whereas  $E_{\text{A}}^{(\text{opt-A:B})}$  and  $E_{\text{B}}^{(\text{opt-A:B})}$  are the energy of isolated base A and B respectively in the optimized geometry of the A:B base pair.  $E_{\text{def}}$  is the energy required to deform the optimized isolated base to attain the geometry in paired base.

We have performed single point energy calculation for the 1683 (17×9×11) structures for each base paired dinucleotide step, containing 72 atoms, using  $\omega$ B97X-D<sup>82</sup> DFT-D functional with cc-pVDZ<sup>126,127,158</sup> basis set by Gaussian09<sup>162</sup>. In addition to total energy calculations for the base pair steps for the different structures (A:B::C:D) we have also calculated single point energies

for the two base pairs namely C:G W:WC,G:A S:HT. A:U W:WC generated using the same intra-base pair parameters as mentioned above. Stacking energy ( $E_{Int}$ ) is calculated as –

$$E_{Int} = E_{Total}(A:B::C:D) - E_{BP1}(A:B) - E_{BP2}(C:D) \quad \dots 4$$

Where A:B and C:D are H-bonded base pairs and A:B::C:D is the dinucleotide sequence. The same base paired dinucleotide step can also be represented as 5'(A-C)3'.5'(D-B)3'. We have also calculated intra-strand and inter strand stacking energies using by equation (1) where the energies are calculated using two bases (A.C), (B.D), (A.D) and (B.C), and energies of the bases.

We have also generated atomic coordinates of the sugar-phosphate backbone atoms using CHARMM<sup>128</sup> for all the systems having best stacking energy for each different Twist value. The initial structures have torsion angles as in A-RNA fiber model and are generated using IC PARAM followed by IC BUILD commands of CHARMM. The modified IC tables are from the RNAHelix<sup>89,164</sup> distribution. Energy minimizations of these structures in vacuum were performed by 1500 steps Steepest Descent, 5000 steps Conjugant gradient methods skipping electrostatic interaction and further by 1000 steps steepest descent, 5000 steps Conjugant gradient and 10000 steps adopted basis Newton-Raphson method by CHARMM considering all types of interaction using AMBER ff14SB<sup>165</sup> force field. All the coordinates of the atoms of the bases were restrained to their initial geometry by large harmonic force of 2000 kcal/mol/Å during all the minimized steps.

### 3.2.4 Hybrid Modeling

Effect of sugar-phosphate backbone has also been mimicked by distances between C1' atoms of two successive residues along the two strand. Mean C1'...C1' distances calculated from RNA crystal structure database for each dinucleotide base pair step sequence are given in

supplementary Table 1. If probability distribution of a parameter (such as C1'...C1' distance) is  $\rho(d)$ , we can write-

$$\frac{\rho(d)}{\rho(d_0)} = e^{\frac{-k(d-d_0)^2}{2k_B T}} \quad \dots 5$$

Where  $d_0$ ,  $k_B$  and  $T$  are the equilibrium value (or mean value) of corresponding parameter, Boltzmann constant and absolute temperature respectively, while  $k$  is the force constant for deviation of a parameter from its equilibrium value. Considering  $\sigma$  as standard deviation of C1'...C1' distance we can evaluate force constant value ( $k$ ) from the standard deviation (half-width at half maxima) of above distribution.

$$\frac{1}{2} = e^{\frac{-k\sigma^2}{2k_B T}} \quad \dots 6$$

$$\ln 2 = \frac{k\sigma^2}{2k_B T} \quad \dots 7$$

$$k = \frac{2k_B T \ln 2}{\sigma^2} \quad \dots 8$$

Considering standard deviation values of C1'... C1' distances and  $T = 300\text{K}$ , the force constants ( $k$ ) are calculated for the two sequences<sup>64</sup>, as listed in **Table 3-1**. Considering energy penalty arising from backbone rigidity of modeled structure, total stacking interaction energy of the dinucleotide base pair steps can be expressed as,

$$E_{\text{stacking}} = E_{\text{Total}}(\text{A:B::C:D}) - E_{\text{BP1}}(\text{A:B}) - E_{\text{BP2}}(\text{C:D}) + \frac{1}{2} k^1_d (d^1 - d^1_o)^2 + \frac{1}{2} k^2_d (d^2 - d^2_o)^2 \quad \dots 9$$

Where 3<sup>rd</sup> and 4<sup>th</sup> terms of above expression are the energy penalty values for the two strands,  $d^1_o$  and  $d^2_o$  are the mean C1'...C1' distances between two consecutive base pairs for the two strands from RNA crystal structures dataset analysis and  $d^1$  and  $d^2$  are the C1'...C1' distances along the two strands between residues A to C and D to B of our model structures.

The stacking iso-energy contours have been generated by MATLAB 2015a with 1 kcal/mol energy difference between two adjacent contours. Roll and Slide values from experimental crystal structures of the dinucleotide base pair steps are also marked on the stacking energy contour maps of our modeled dinucleotide base pair steps for comparison with experimental structures. The red dots in the figures are indicating minimum energy i.e. best stacking energy between two base pair for each Twist value and the region around the red dot is considered as “best stacking energy region” or “lowest energy region” within ( $E_{\min}+1$ ) kcal/mol.

**Table 3-1** Mean C1'...C1' distances, calculated from X-ray crystallographic database, along with their standard deviations (in parenthesis), for the two strands of the two sequences and the calculated force-constants values for stretching of the pseudo-bonds. See **Figure 3-1**.

Sequence	First Strand(C1'...C1')		Second Strand(C1'...C1')	
	Mean Distance (Å)	Force Constant (kcal/mol/Å <sup>2</sup> )	Mean Distance (Å)	Force Constant (kcal/mol/Å <sup>2</sup> )
C:G W:WC::G:A S:HT	C-G: 5.25(0.23)	15.72	A-G: 5.03(0.23)	16.00
G:C W:W:C::G:A S:HT	G-G: 5.36(0.20)	19.79	A-C: 5.13(0.25)	13.09
A:U W:WC::C:G W:WC	A-C: 5.38(0.30)	9.20	G-U: 5.63(0.32)	8.33

### 3.3 Result and Discussion

#### 3.3.1 Stability of the base pairs:

We have model built structures of two Watson-Crick base pairs, C:G W:WC and A:U W:WC and the non-canonical sheared G:A S:HT base pair using assigned average base pair parameters as obtained from analysis of X-ray crystal structures (Table 1) and have geometry optimized them using DFT-D with two basis sets. The interaction energies of the aug-cc-pVDZ optimized structures appear slightly smaller than the cc-pVDZ optimized structures. As expected the interaction energy for the G:C Watson–Crick base pair with three hydrogen bonds is the highest

(-34.03 kcal mol<sup>-1</sup> for the cc-pVDZ basis set), followed by that of the A:U base pair (-19.84 kcal mol<sup>-1</sup>) and G:A base pair (-16.55 kcal mol<sup>-1</sup>), similar to earlier results<sup>96,147,166</sup>. The base pair parameters of the two sets of optimized structures are, however, quite similar, indicating similar orientations obtained by the two methods. The optimized structures of A:U and G:C Watson-Crick base pairs, are, however, quite planar with nearly zero value of propeller as also reported earlier<sup>96</sup>. The base pairs are however, known to have significantly large negative Propeller values in the structures of DNA or RNA double helices (**Table 3-2**). The G:A S:HT base pair, however, become quite highly Propeller Twisted upon free optimization. The interaction energies of the H-opt structures are also quite similar to the free-opt cases, indicating energy wise the two sets of structures are not much different. In order to keep consistency with RNA double helical structures, we have also optimized hydrogen atomic positions of the base pairs generated with considerable amount of Propeller (-15° for the Watson-Crick base pairs and -10° for the G:A base pair), as generally found in RNA double helices. As expected the C:G base pair has consistently high interaction energy by all methods while the G:A S:HT base pair is the weakest in all the types of optimized structures (**Table 3-2**).

**Table 3-2** Geometric and energetic parameters of three base pairs obtained by different methods of modeling and optimization.

Base Pair	Method		Buckle (°)	Open (°)	Propeller (°)	Stagger (Å)	Shear (Å)	Stretch (Å)	$\Delta E_{\text{opt}}$ (kcal/mol)
C:G W:WC	H-optimized models in stacking energy calculations		0.00	1.12	-15.00	0.00	0.07	2.80	-35.04
	Free optimized	cc-pVDZ Basis Set	-3.89	-3.78	-1.44	-0.03	0.10	2.91	-34.03
		aug-cc-pVDZ Basis Set	0.04	-3.69	-0.10	0.00	0.11	2.91	-30.97
	H-optimized as in average crystal structures		5.97(9.75)	1.08(4.06)	-8.82(7.87)	-0.11(0.35)	0.06(0.45)	2.86(0.16)	-35.52
A:U W:WC	H-optimized models in stacking energy calculations		0.00	0.00	-15.00	0.00	0.00	2.80	-19.78
	Free optimized	cc-pVDZ Basis Set	-0.77	2.62	-0.42	0.00	0.07	2.80	-19.84
		aug-cc-pVDZ Basis Set	0.39	2.21	-0.11	0.00	0.08	2.82	-16.73
	H-optimized as in average crystal structures		-2.54(9.42)	3.64(5.10)	-9.12(8.12)	-0.06(0.35)	0.12(0.35)	2.80(0.15)	-20.05
G:A S:HT	H-optimized models in stacking energy calculations		-0.00	-13.20	-10.00	0.00	2.23	3.10	-13.59
	Free optimized	cc-pVDZ Basis Set	10.16	-20.23	-41.01	-0.02	1.90	3.17	-16.55
		aug-cc-pVDZ Basis Set	1.37	-20.71	-41.92	0.05	1.93	3.19	-14.40
	H-optimized as in average crystal structures		-2.82(16.01)	-13.16(5.70)	1.98(14.40)	-0.19(0.47)	2.23(0.32)	3.31(0.17)	-12.94

### 3.3.2 Evaluation of the Stacking Models

As indicated above the G:A S:HT base pairs appear very frequently in the crystallographic dataset as a non-Watson-Crick base pair<sup>139</sup>. We have found that this base pair appears quite frequently in the double helical regions also where it often stacks on top of G:C or G:U base pairs. The values of inter-base pair parameters, Tilt, Roll, Rise etc., indicate nearly parallel stacking between the non-canonical base pair with the Watson-Crick base pairs (**Table 3-3**). The mean values of Tilt and Shift for both the sequences are smaller than their corresponding standard deviation values (**Table 3-3**), which indicates that their central tendencies are zero. This also indicates symmetry of the two strands. The Roll values are generally large positive indicating that they adopt A-form RNA geometry. The Slide values are however generally low negative (**Table 3-3**) as compared to those between Watson-Crick base pairs in RNA<sup>108</sup>. We have observed that mean Twist values are 11.35° and 7.13° for the dinucleotide base pair steps for the sequence G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT, respectively. In regular A-form of RNA structures one observes Twist values around 33° with some variation due to sequence and other environmental factors. In spite of slightly large rise values, the average stacking overlap values are usually greater than 45Å<sup>2</sup> for all possible dinucleotide base pair step sequences having non canonical G:A S:HT and Watson-Crick base pairs (**Table 3-3**). These high overlap values may arise due to involvement of a larger purine-purine base pair but also indicate good stability of the stacking.

In order to understand such unusual preference of Twist, we have carried out detail model building studies of the two frequent dinucleotide base pair step sequences. We have varied the three important parameters – Roll, Twist and Slide in their widest possible range forming right handed double helix and have measured stacking overlaps of the model structures as the first

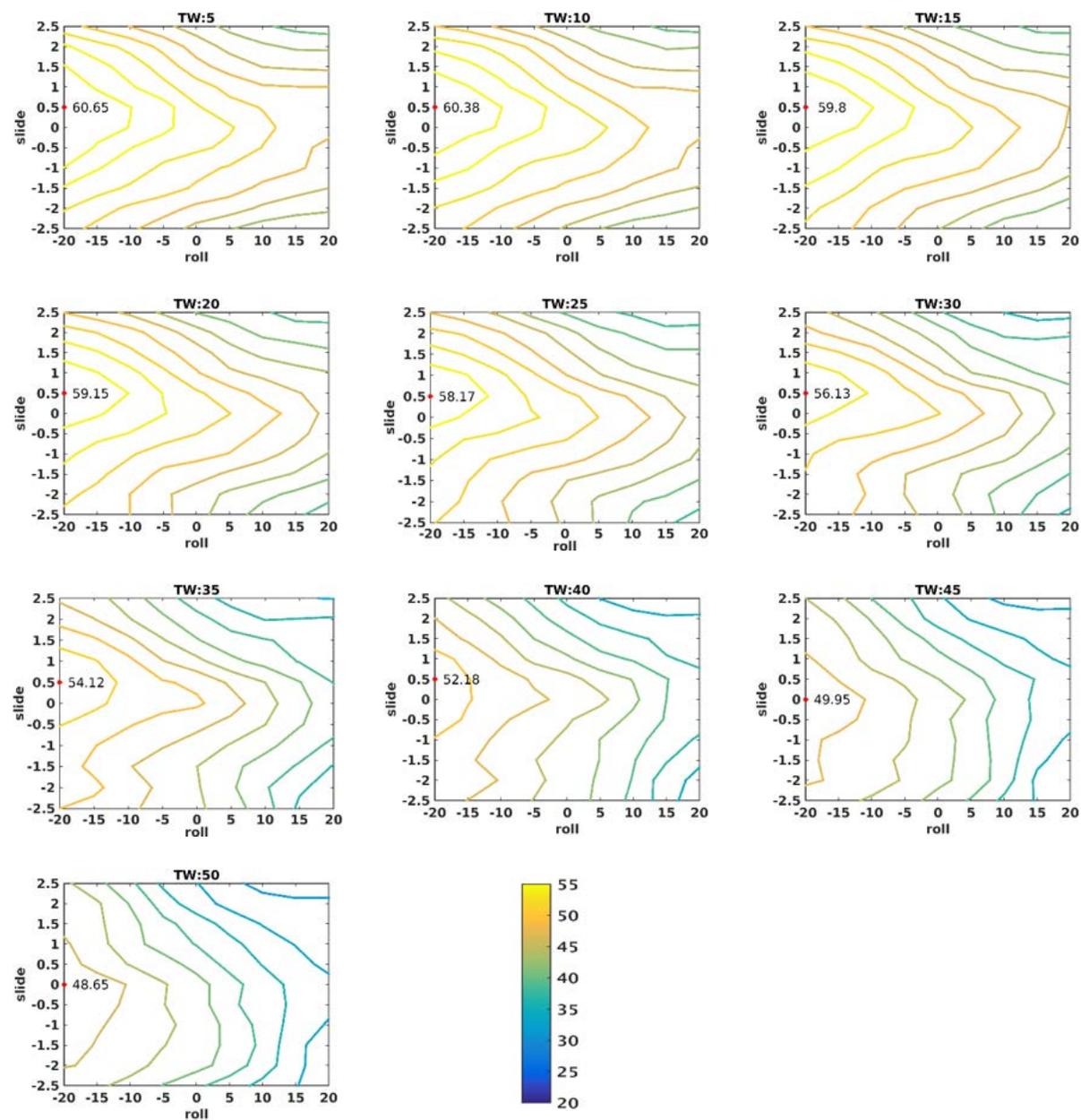
measure of stackability. It is seen that the stacking overlap values between the base pairs strongly depend on Roll and Slide for smaller Twist values ( $5^\circ$  to  $45^\circ$ ) while at larger Twist values ( $60^\circ$  to  $85^\circ$ ) stacking overlap values are smaller and they became insensitive to Roll and Slide.

**Table 3-3** Mean and standard deviation (in parenthesis) values of base step parameters and stacking overlap values from the available crystal structures in the base paired dinucleotide steps are shown here. We did not observe any data for two sequences in the experiment.

Base pair Step	No. of data points	Tilt ( $^\circ$ )	Roll ( $^\circ$ )	Twist ( $^\circ$ )	Shift ( $\text{\AA}$ )	Slide ( $\text{\AA}$ )	Rise ( $\text{\AA}$ )	Overlap ( $\text{\AA}^2$ )
C:G W:WC::G:A S:HT	1168	0.57 (1.95)	9.20 (6.92)	7.13 (6.59)	-0.22 (0.60)	-0.87 (0.46)	3.49 (0.26)	47.57 (5.89)
C:G W:WC::A:G H:ST	0	-	-	-	-	-	-	-
G:C W:WC::G:A S:HT	647	-0.86 (2.05)	9.31 (6.72)	11.34 (5.82)	0.15 (0.67)	-0.63 (0.39)	3.49 (0.18)	47.04 (5.26)
G:C W:WC::A:G H:ST	0	-	-	-	-	-	-	-
A:U W:WC::C:G W:WC	2687	0.69 (2.27)	5.44 (4.80)	31.97 (3.58)	0.20 (0.53)	-1.38 (0.50)	3.33 (0.21)	47.80 (4.37)

Highest stacking overlaps of  $50\text{\AA}^2$  and  $54\text{\AA}^2$  are found at  $20^\circ$  Twist for dinucleotide base pair step sequence G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT respectively, which appears to be even larger than those between Watson-Crick and the above base pairs in the crystal structure database<sup>116</sup> (**Table 3-3**). The models having maximum stacking overlap values are found to have large negative Roll and moderately positive Slide. This combination of Roll and Slide are, however, not suitable for A-form RNA structures and far from experimental mean values of the same. We have also analyzed stacking overlap values for A:U W:WC::C:G W:WC base pair step (**Figure 3-2**), which also show large overlap for small Twist and large negative Roll. It may be mentioned that large negative Roll and positive Slide values are quite suitable for B-DNA structure. It may be noted that van der Waals surface overlap between two atoms continues to increase when distance between the two atoms reduces. Hence for very close distance between

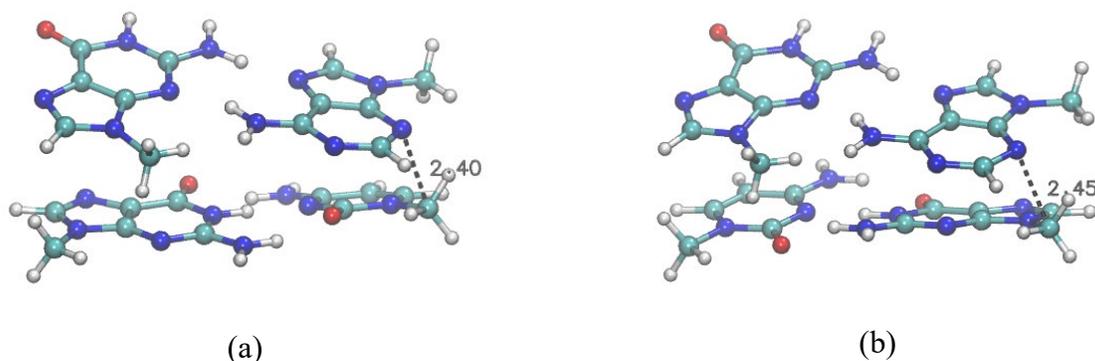
atoms of the two base pairs the stacking overlap can be high but such structures may be sterically hindered.



**Figure 3-2** Stacking overlap area contours at different Twist value for A:U W:WC::C G:W WC are shown here. Color bar is given in  $\text{\AA}^2$ . Contour lines are  $2.5 \text{\AA}^2$  apart.

In fact the structures corresponding to large negative Roll and large positive Slide having stacking overlap values above  $50 \text{\AA}^2$  are found to have very short distances around  $2.4 \text{\AA}$  between C1' of Cyt

and N3 atoms of Ade for both the sequences (**Figure 3-3**). It is important to mention that these short inter-atomic contacts may give large negative dispersion energy for configurations having maximum overlap as the dispersion energy term, which is  $-C_6/R^6$ , becomes very large. The DFT component of such configuration may be large positive though. This indicates that molecular modeling studies based on stacking overlap alone is incapable to give full picture of preferred geometry. The G:A S:HT base pairs do not appear in B-DNA structures possibly to avoid the above short-contacts.



**Figure 3-3** Model Structures for  $20^\circ$  Twist,  $-20^\circ$  Roll and  $1\text{\AA}$  Slide having maximum overlap area for (a) G:C W:WC::G:A S:HT and (b) C:G W:WC::G:A S:HT are shown here indicating the short inter atomic distances.

### 3.3.3 Stacking Energy Analysis

Stacking overlap values are nevertheless indicative of good stacking to justify water release and entropy contribution. However the enthalpy components of stacking arise from other stabilizing interactions between two base pairs. Hence we have calculated stacking energy between two base pairs using dispersion corrected density functional theory. As dispersion energy depends on  $-C_6/R^6$ , we have obtained high negative dispersion energy and high positive KS-DFT energy and combining effect is highly destabilizing for configuration having maximum overlap (**Table 3-4**). This confirms the presence of short contact between atoms quantitatively. It indicates that stacking overlap area is not a complete parameter to quantify stacking between two consecutive

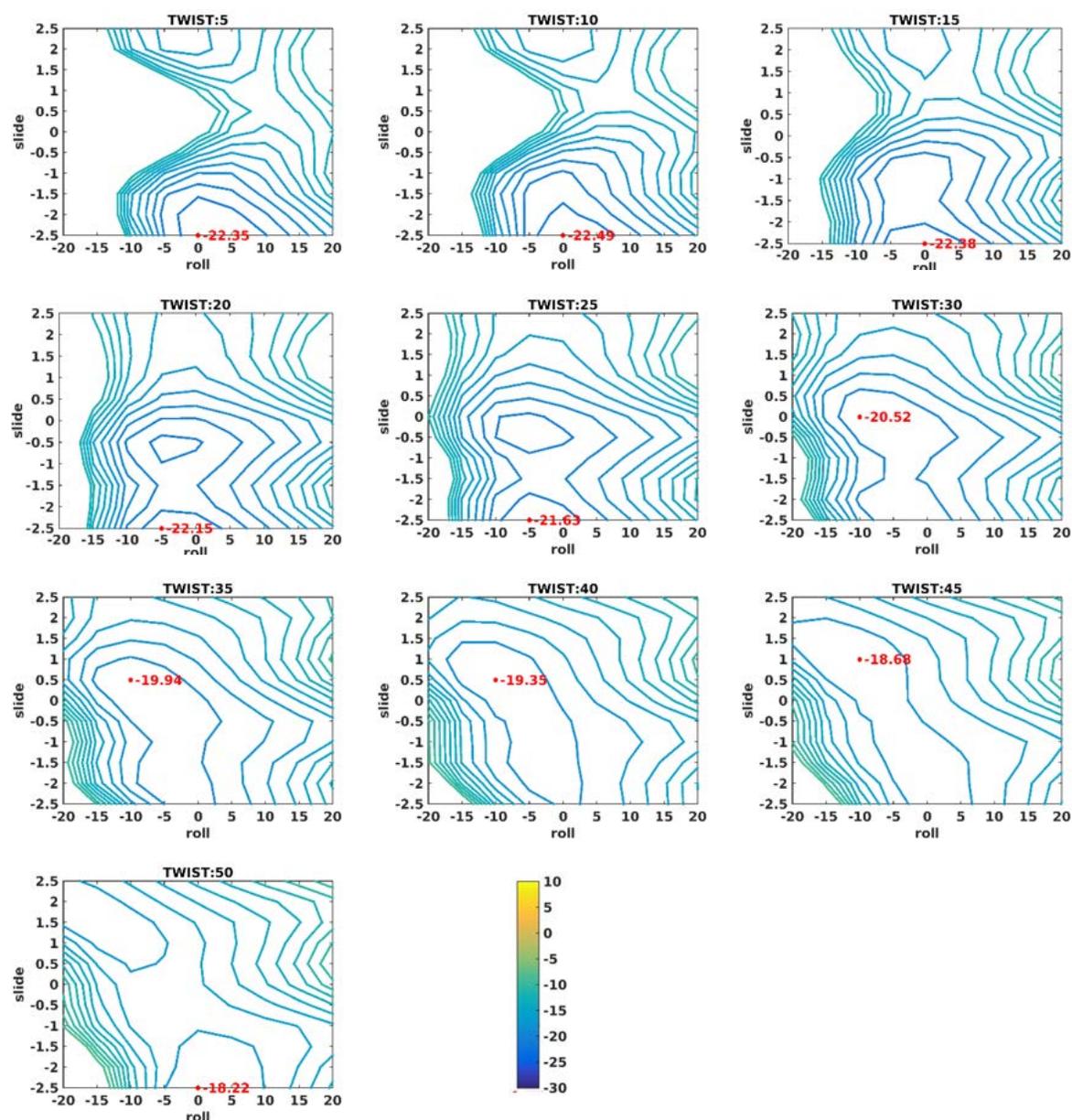
base pairs. The iso energy contours of Watson-Crick dinucleotide sequence like A:U W:WC::C:G W:WC elucidate that structure near 0° Roll and -2.5Å Slide and 10° Twist is the most stable configuration (**Figure 3-4**).

**Table 3-4** Best stacking energy (DFT-D) and dispersion energy components in configuration having maximum stacking overlap area for all Twist values for the dinucleotide step sequences are shown below.

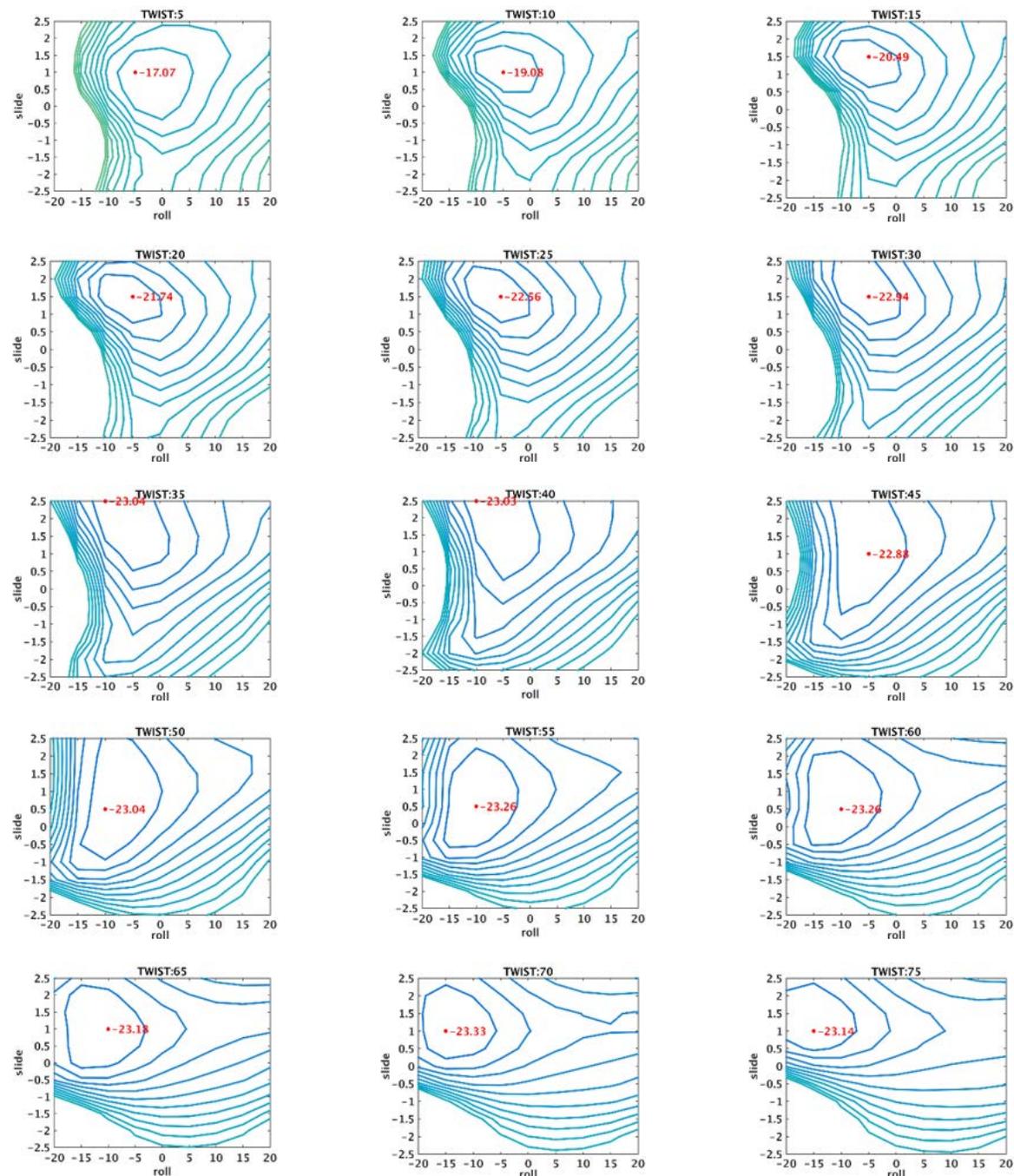
<b>a) G:C W:WC::G:A S:HT</b>					
Twist (°)	Roll (°)	Slide (Å)	Maximum Overlap Area (Å <sup>2</sup> )	Best Stacking Energy (kcal/mol)	Dispersion Energy (kcal/mol)
5	-20	1	48.13	2.23	-21.66
10	-20	1	49.42	-0.58	-21.95
15	-20	1	49.77	1.84	-22.03
20	-20	1	50.22	4.26	-21.74
25	-20	0.5	49.9	16.96	-20.91
30	-20	1	49.88	4.11	-20.81
35	-20	1	48.73	4.13	-20.41
40	-20	1	47.25	1.53	-20.06
45	-20	1	45.4	-6.46	-19.56
50	-20	2.5	44.43	-14.37	-18.97
55	-20	2.5	43.93	-16.10	-18.47
60	-20	2.5	42.77	-18.17	-17.95
65	-20	2.5	41.88	-20.13	-17.38
70	-20	2.5	40.65	-21.39	-16.78
75	-20	2.5	39.4	-21.81	-16.18
80	-20	2	37.83	-22.19	-15.55
85	-20	2.5	36.5	-21.45	-15.09
<b>b) C:G W:WC :: G:A S:HT</b>					
Twist (°)	Roll (°)	Slide (Å)	Maximum Overlap Area (Å <sup>2</sup> )	Best Stacking Energy (kcal/mol)	Dispersion Energy (kcal/mol)
5	-20	1	53.3	3.46	-24.07
10	-20	1	54.02	-0.21	-24.29
15	-20	1	54.18	1.32	-24.29
20	-20	1	54.43	3.52	-23.94
25	-20	1	53.88	4.02	-23.42
30	-20	0.5	53.08	10.25	-22.35
35	-20	0.5	51.72	11.98	-21.82
40	-20	0.5	50.93	10.46	-21.26
45	-20	0.5	49.73	8.25	-20.63
50	-20	0.5	48.92	13.32	-20.06
55	-20	0.5	47.35	25.50	-19.50
60	-20	0.5	45.88	39.83	-18.76
65	-20	1	44.33	21.08	-18.49
70	-20	2	43.77	-8.60	-18.44

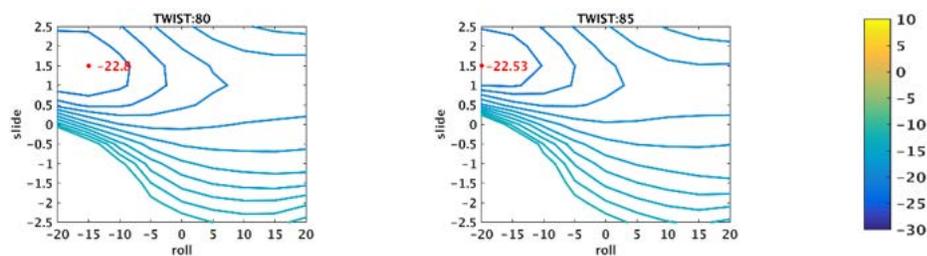
75	-20	2	43.38	-3.28	-17.87
80	-20	2.5	42.67	-8.38	-17.60
85	-20	2.5	42.8	-4.73	-17.20

The iso-energy contours of stacking energies, however, show that the structure near  $-15^\circ$  Roll,  $1.0\text{\AA}$  Slide and  $70^\circ$  Twist and  $-5^\circ$  Roll,  $1.5\text{\AA}$  Slide and  $20^\circ$  Twist are the most stable configuration for G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT sequences respectively, where stacking overlaps are also significant (**Figure 3-5** and **Figure 3-6**).

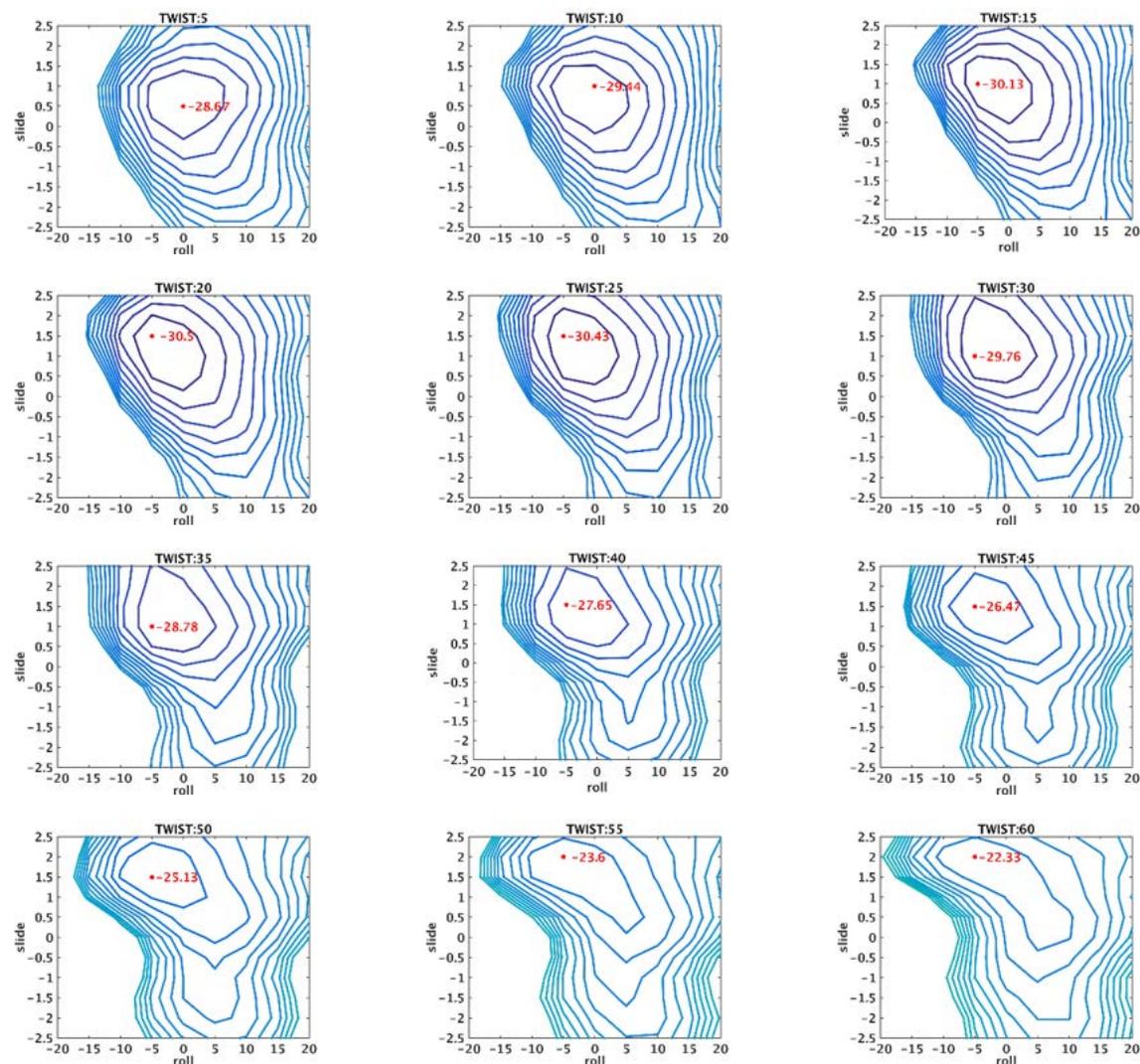


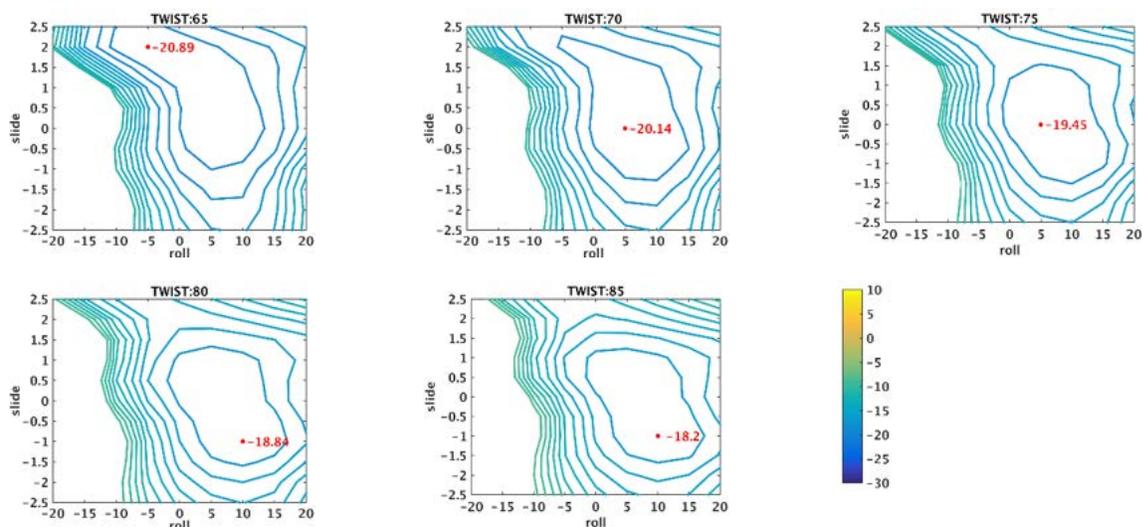
**Figure 3-4** Intrinsic stacking energy contours considering DFT-D energy at different Twist values for A:U W:WC::C:G W:WC dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.





**Figure 3-5** Intrinsic stacking energy contours considering DFT-D energy at different Twist values for G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.





**Figure 3-6** Intrinsic stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.

These Twist values for best stacking orientations are however significantly different from the average Twist value obtained from analysis of available experimental structures. Best stacking energy is thus not sufficient for formation of double helix as nucleotide bases need to be joined by sugar- phosphate backbone atoms. We have generated the sugar-phosphate backbone for each best model structure of the base pair step corresponding to each Twist value and have minimized only the atomic positions of sugar phosphate backbone, restraining the atoms of the bases. The sum of all bonding energies ( $E_{\text{strain}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}}$ ) for these structures are tabulated in **Table 3-5**. This indicates that the conformation having best stacking energy of G:C W:WC::G:A S:HT sequence for 25° Twist has least strain energy ( $E_{\text{strain}}$ ) arising from backbone connection. This structure, however, has -5° Roll and 1.5Å Slide, unsuitable for A-RNA double helix. The  $E_{\text{strain}}$  is considerably high for the best structures at 70° Twist for this sequence. In this restrained energy minimized structure at 70° Twist, sugar-phosphate backbone connection can be ruled out, as bond distances between P and O3' and between P and O5' increase to 1.8Å (**Figure 3-7**) and some bond angles involving  $sp^3$  atoms deviates to 136° or 68°. Besides sugar puckers of some of the

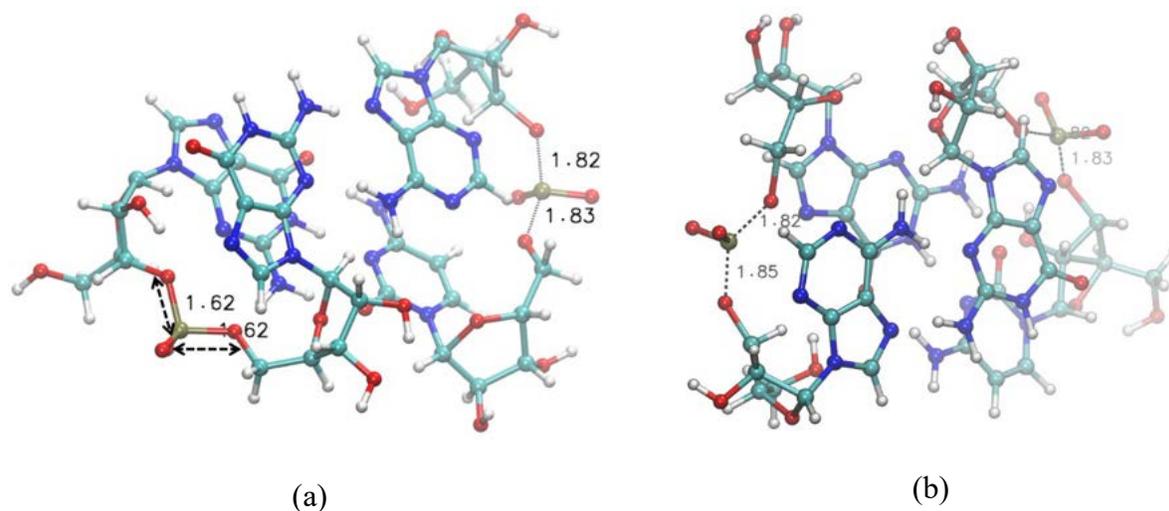
minimized structures were found to be in O4'-exo region, which is known to be unstable. These unusual geometries, with high  $E_{\text{strain}}$ , arise due to large distances (upto 10.5Å) between C1'...C1' atoms along the strands (**Figure 3-7a**).

**Table 3-5** Geometrical and energy parameters for the structures of the two sequences having minimum DFT-D energy corresponding to each Twist value are listed. Mean values of C1'...C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns.

a) G:C W:WC::G:A S:HT										
Twist (°)	Roll (°)	Slide (Å)	Best Stacking Energy (DFT-D component) (kcal/mol)	Intra Strand Energy (Kcal/mol)		Inter Strand Energy (kcal/mol)		Total Strain energy (kcal/mol)	C1'...C1' Distance(Å)	
				DFT-D component	Dispersion component	DFT-D component	Dispersion component		1 <sup>st</sup> Strand (5.36)	2 <sup>nd</sup> Strand (5.13)
5	-5	1	-17.07	-13.86	-16.16	-5.32	-2.57	134.80	4.05	3.67
10	-5	1	-19.08	-16.23	-16.48	-4.90	-2.42	129.90	3.93	3.96
15	-5	1.5	-20.49	-17.49	-16.01	-5.02	-2.63	122.19	4.07	4.33
20	-5	1.5	-21.74	-19.06	-15.98	-4.74	-2.51	121.91	4.24	4.73
25	-5	1.5	-22.56	-20.17	-15.9	-4.53	-2.4	111.59	4.43	5.14
30	-5	1.5	-22.94	-20.58	-15.74	-4.34	-2.31	119.69	4.04	5.40
35	-10	2.5	-23.04	-18.14	-14.34	-6.83	-3.68	122.00	4.27	5.81
40	-10	2.5	-23.03	-18.26	-14.04	-6.67	-3.6	120.99	5.38	6.51
45	-5	1	-22.88	-21.42	-14.94	-3.09	-1.83	143.03	5.81	7.00
50	-10	0.5	-23.04	-21.39	-15.25	-3.39	-1.77	190.75	6.06	7.42
55	-10	0.5	-23.26	-21.44	-14.53	-3.50	-1.82	206.42	3.95	3.35
60	-10	0.5	-23.26	-21.42	-13.76	-3.60	-1.91	166.12	6.31	7.83
65	-10	1	-23.18	-20.90	-13.09	-3.86	-2.11	246.84	6.25	8.08
70	-15	1	-23.34	-20.25	-13.14	-4.73	-2.41	275.94	6.40	8.40
75	-15	1	-23.14	-19.80	-12.39	-5.03	-2.58	321.46	6.64	8.78
80	-15	1.5	-22.80	-18.33	-11.5	-5.85	-3.02	347.00	6.61	9.03
85	-20	1.5	-22.53	-17.22	-11.5	-6.85	-3.46	409.99	6.76	9.31
b) C:G W:WC::G:A S:HT										
Twist (°)	Roll (°)	Slide (Å)	Best Stacking Energy (DFT-D component) (kcal/mol)	Intra Strand Energy (Kcal/mol)	Inter Strand Energy (kcal/mol)	Intra Strand Energy (Kcal/mol)	Inter Strand Energy (kcal/mol)	Total Strain energy (kcal/mol)	C1'...C1' Distance(Å)	Total Strain energy (kcal/mol)
				DFT-D component	Dispersion component	DFT-D component	Dispersion component			
5	0	0.5	-28.66	-24.13	-18.84	-4.69	-2.42	136.47	4.15	3.81

*Stacking Interaction by Most Frequent A:G base pair: Establishment of Hybrid DFT-D Method*

10	0	1	-29.44	-25.17	-19.11	-4.37	-2.09	141.72	4.10	4.05
15	-5	1	-30.13	-26.02	-19.83	-4.56	-2.22	121.05	4.01	4.37
20	-5	1.5	-30.49	-26.94	-19.62	-4.26	-2.00	122.97	4.17	4.78
25	-5	1.5	-30.43	-26.74	-19.12	-4.49	-2.09	131.33	4.62	5.28
30	-5	1	-29.76	-25.43	-18.26	-5.25	-2.62	143.21	4.84	5.71
35	-5	1	-28.78	-24.37	-17.56	-5.44	-2.83	157.64	4.79	6.04
40	-5	1.5	-27.65	-23.58	-17.10	-5.21	-2.60	170.93	5.02	6.47
45	-5	1.5	-26.47	-22.40	-16.28	-5.44	-2.84	222.47	4.30	3.59
50	-5	1.5	-25.13	-21.09	-15.40	-5.52	-3.12	215.55	5.26	6.89
55	-5	2	-23.60	-19.92	-14.65	-5.18	-3.11	200.89	5.27	7.24
60	-5	2	-22.33	-18.55	-13.77	-5.25	-3.46	218.00	5.51	7.65
65	-5	2	-20.89	-17.11	-12.89	-5.17	-3.80	231.47	5.76	8.04
70	5	0	-20.14	-10.48	-9.64	-10.49	-5.33	186.58	7.31	9.04
75	5	0	-19.45	-9.08	-9.02	-11.25	-5.68	225.34	7.54	9.41
80	10	-1	-18.84	-5.65	-7.49	-13.93	-6.90	491.99	8.50	10.21
85	10	-1	-18.20	-4.75	-7.06	-14.28	-7.26	588.97	8.70	10.54

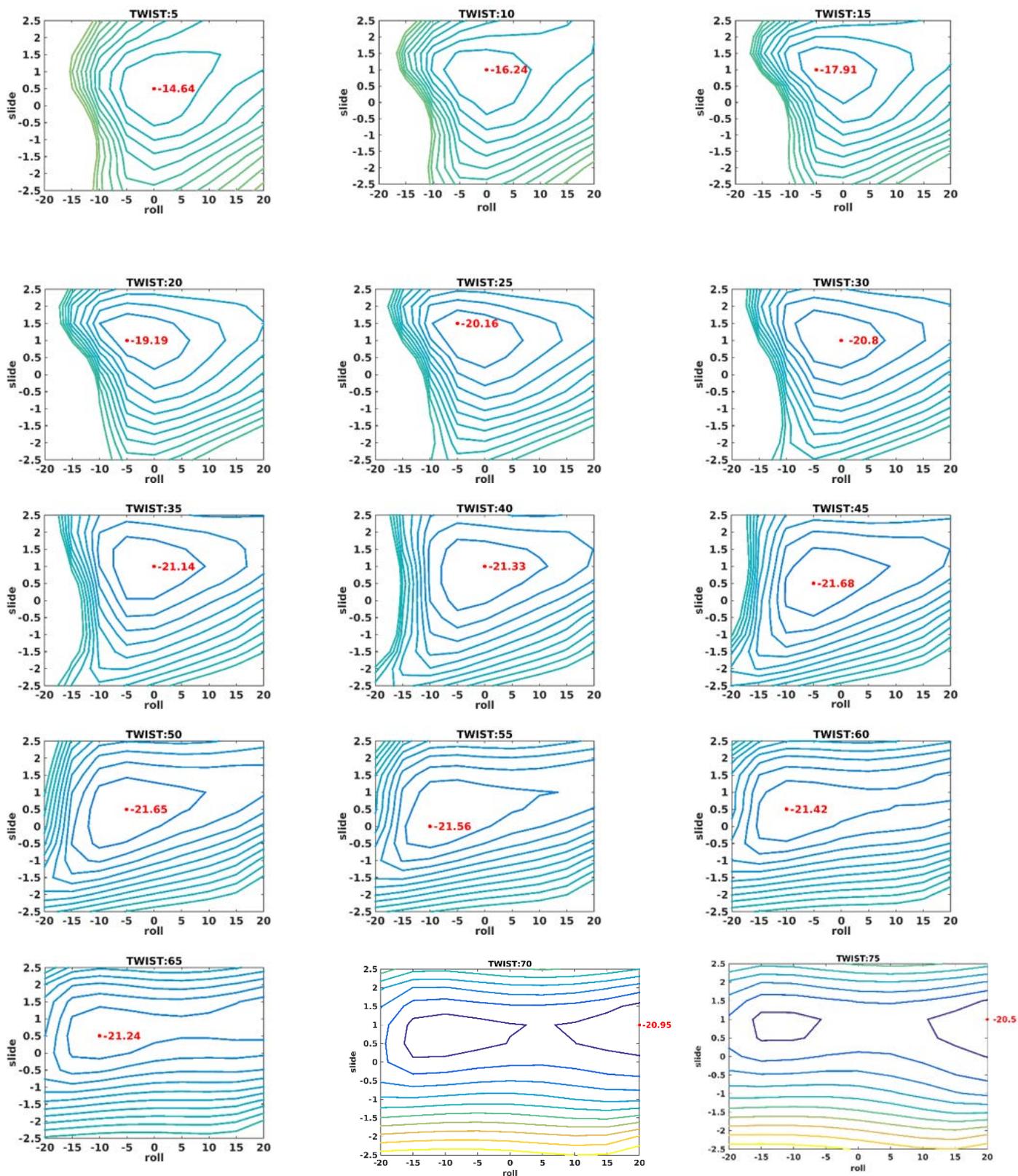


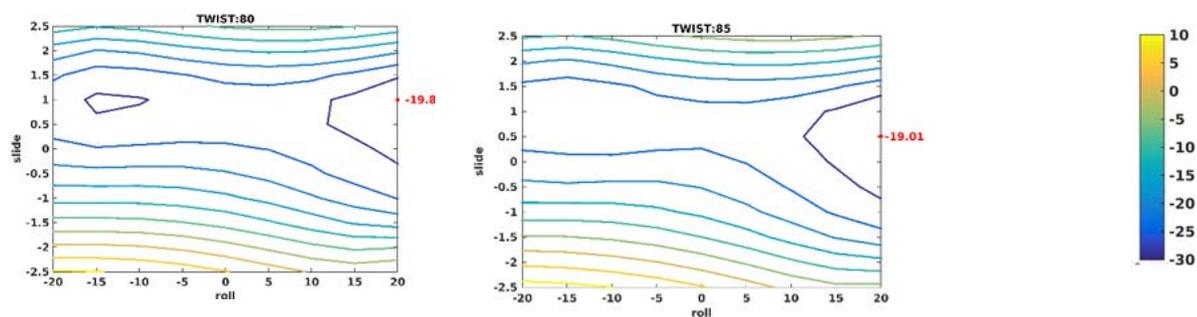
**Figure 3-7** Model structures of (a) *G:C:W:WC::G:A S:HT* for  $-20^\circ$  Roll and  $1.5\text{\AA}$  Slide and (b) *C:G W:WC::G:A S:HT* for  $10^\circ$  Roll and  $-1.0\text{\AA}$  Slide in their best stacking energy configurations at  $85^\circ$  Twist. The large bond lengths in sugar-phosphate group are also shown.

It may be noted that the mean distance between  $C1'\dots C1'$  atoms along the strands, as calculated from crystal structure database for these sequence, are around  $5.2\text{\AA}$  (**Table 3-1**). This also indicates that the structures having large  $C1'\dots C1'$  distance cannot form double helix with proper sugar phosphate backbone. As energy minimizations may get trapped in local energy minima leading to difficulty in predicting best structure on all aspects, we have adopted the coarse grained method to obtain stacking energy landscape considering energy penalty arising from stretching of  $C1'\dots C1'$  pseudo bonds. Furthermore the stacking energies obtained from DFT-D calculations are around  $-20$  kcal/mol while the backbone energies from AMBER ff14SB<sup>165</sup> force field are always more than  $+100$  kcal/mol, hence are not comparable. We have therefore added the coarse-grain energy penalty values, from  $C1'\dots C1'$  distances, to the stacking energy from DFT-D calculations to mimic the effect of backbone into our model structures. These coarse grain energy penalties vary from  $0.1$  to  $155$  kcal/mol and are quite comparable to the DFT-D energies.

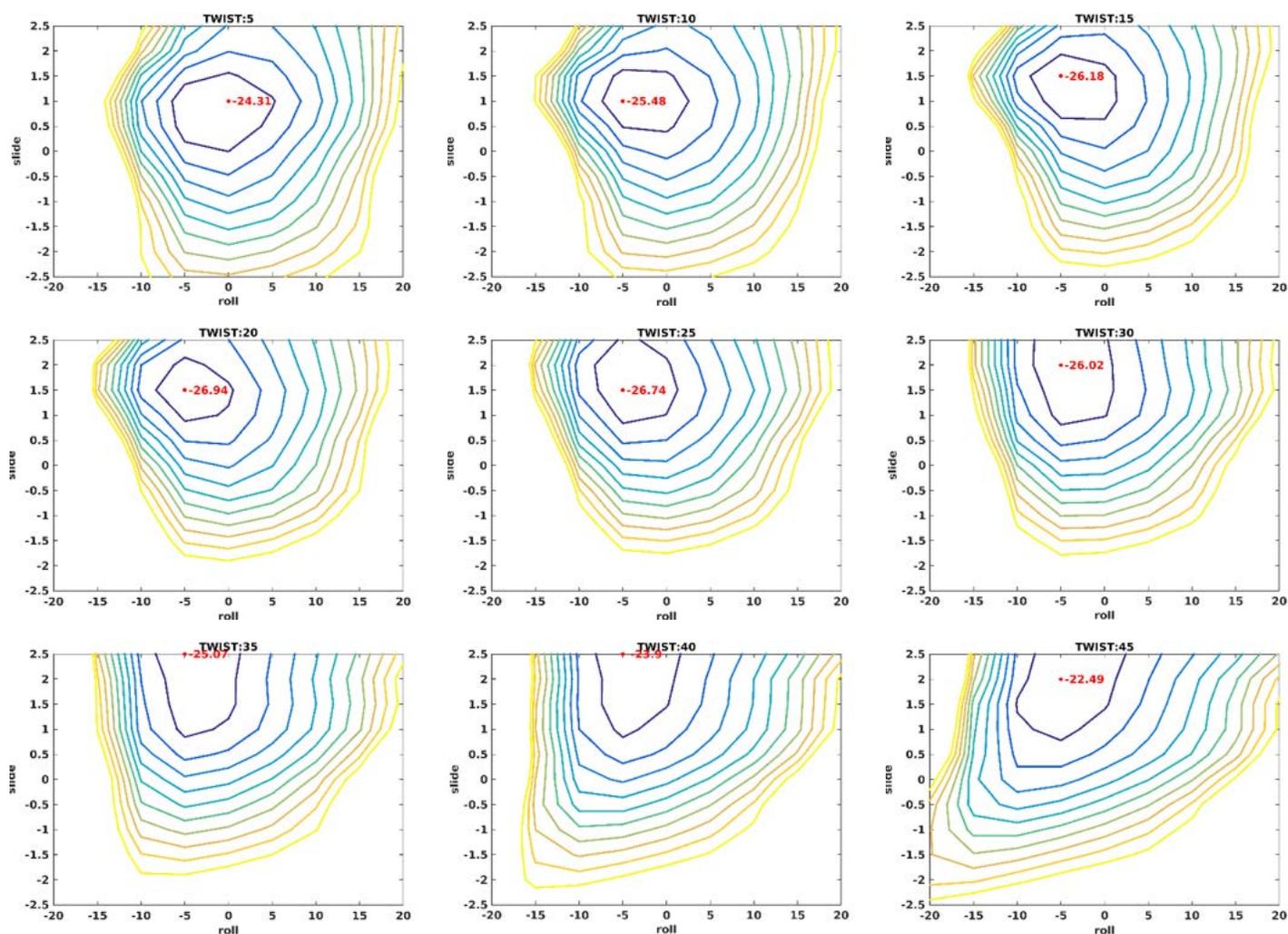
We have also calculated stacking energy as a summation of inter-strand and intra-strand stacking components and we have compared it with complete stacking energies. The sum of intra-strand and inter-strand stacking energies are found to be systematically larger than the stacking energies considering the whole base pairs, The non-additive property of the DFT energies possibly arises due to alterations of electronic structures of an isolated base and the same base in a base pair. It indicates that stacking energy calculation considering total dinucleotide base pair step is more accurate method than that of considering as a summation of inter strand and intra strand components which have earlier been done by various other groups<sup>166,167</sup>. Iso-energy contours of intra-strand stacking for C:G W:WC::G:A S:HT and G:C W:WC::G:A S:HT elucidate stabilities for stacks with positive Slide and negative Roll (**Figure 3-8** and **Figure 3-9**). These values are unsuitable for the A-form structure of RNA. Iso-energy contours of inter-strand stacking for C:G W:WC::G:A S:HT and G:C W:WC::G:A S:HT elucidate that the best stacking energies are observed at large negative Slide and large positive Slide, respectively, as inter-strand G.G and inter-strand G.A parallel stacking are possible at high negative and high positive Slide values, respectively (**Figure 3-10** and **Figure 3-11**). In both the sequences containing G:A base pair, we noticed quite high intra-strand interactions, which however give more stabilities for the stacks with positive Slide and negative Roll. These values are unsuitable for A-form structure of RNA. The inter-strand stacking interaction for C:G W:WC::G:A S:HT dinucleotide, however, strongly favors large positive Roll and negative Slide. It may be noted that at negative Slide, the two G residues of opposite strands, with large dipole moment, of this sequence approach each other (**Figure 3-12**). Such interactions presumably give extra stability to the C:G W:WC::G:A S:HT dinucleotide sequence.

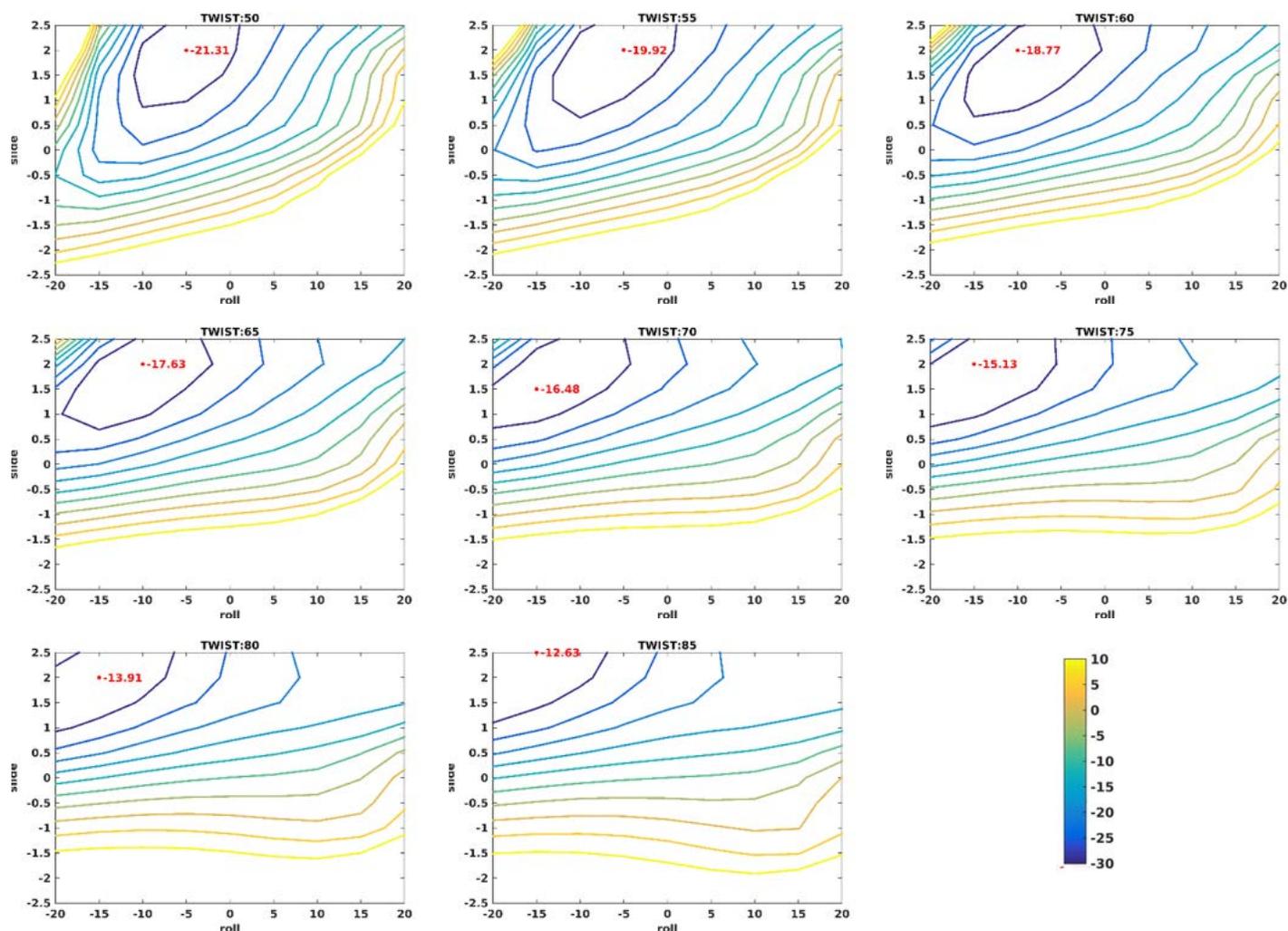
*Stacking Interaction by Most Frequent A:G base pair: Establishment of Hybrid DFT-D Method*



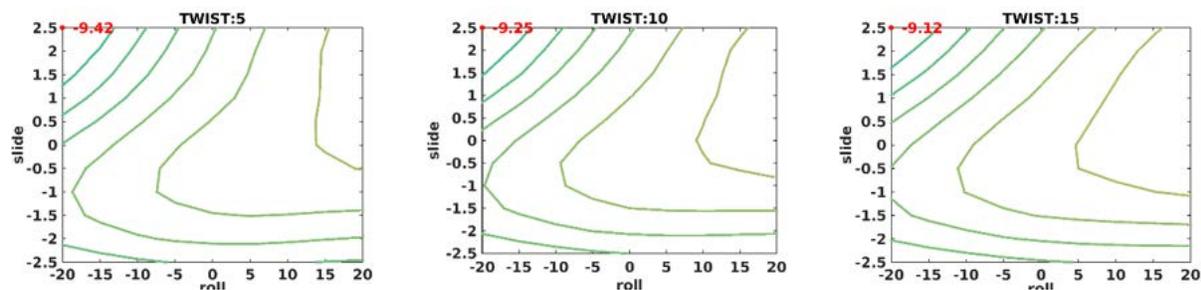


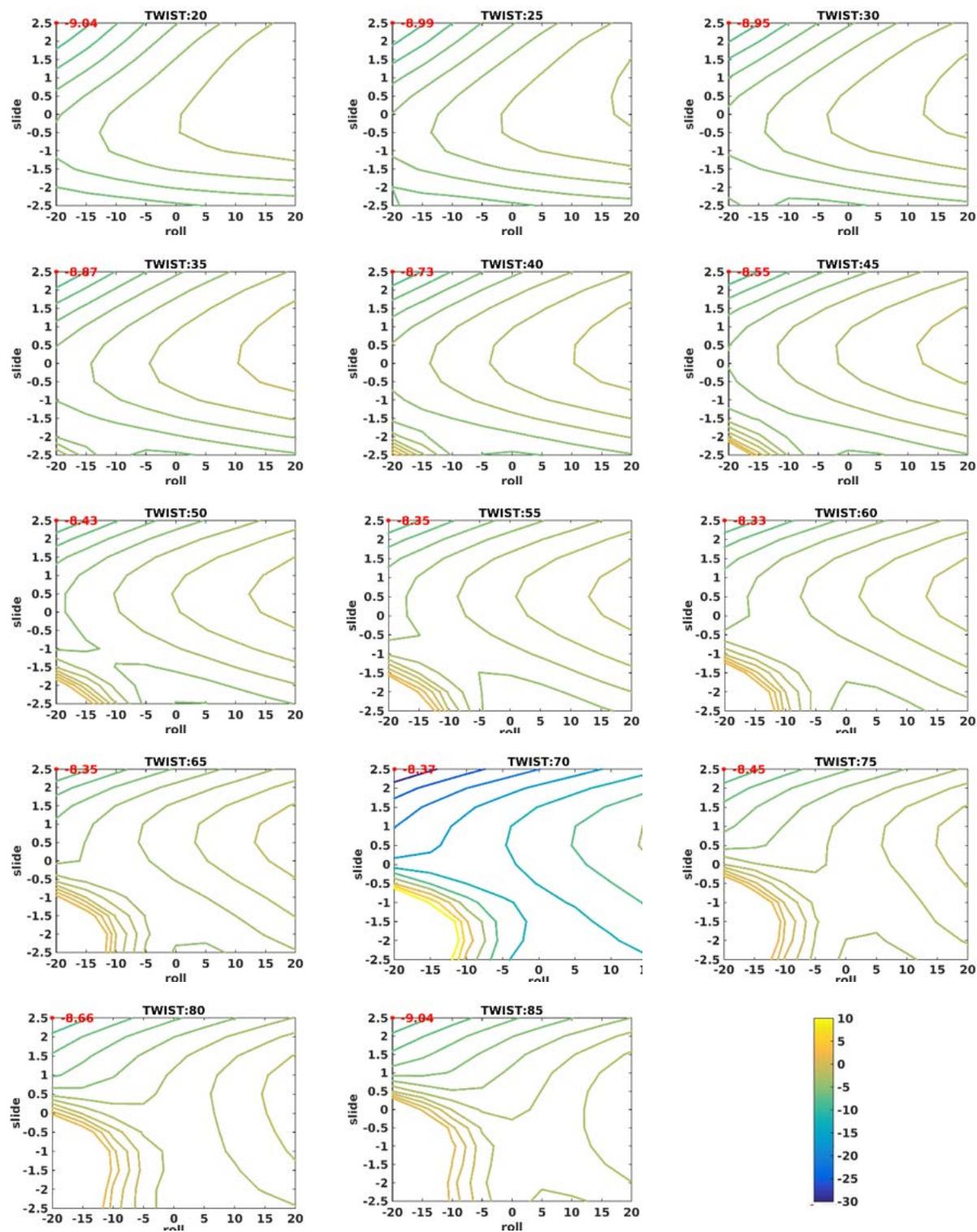
**Figure 3-8** Intrinsic intra-strand stacking energy contours, i.e interaction energy between the two Gua residues and those between the Cyt and Ade residues (see **Figure 1** for detail), considering DFT-D energy at different Twist values for G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.





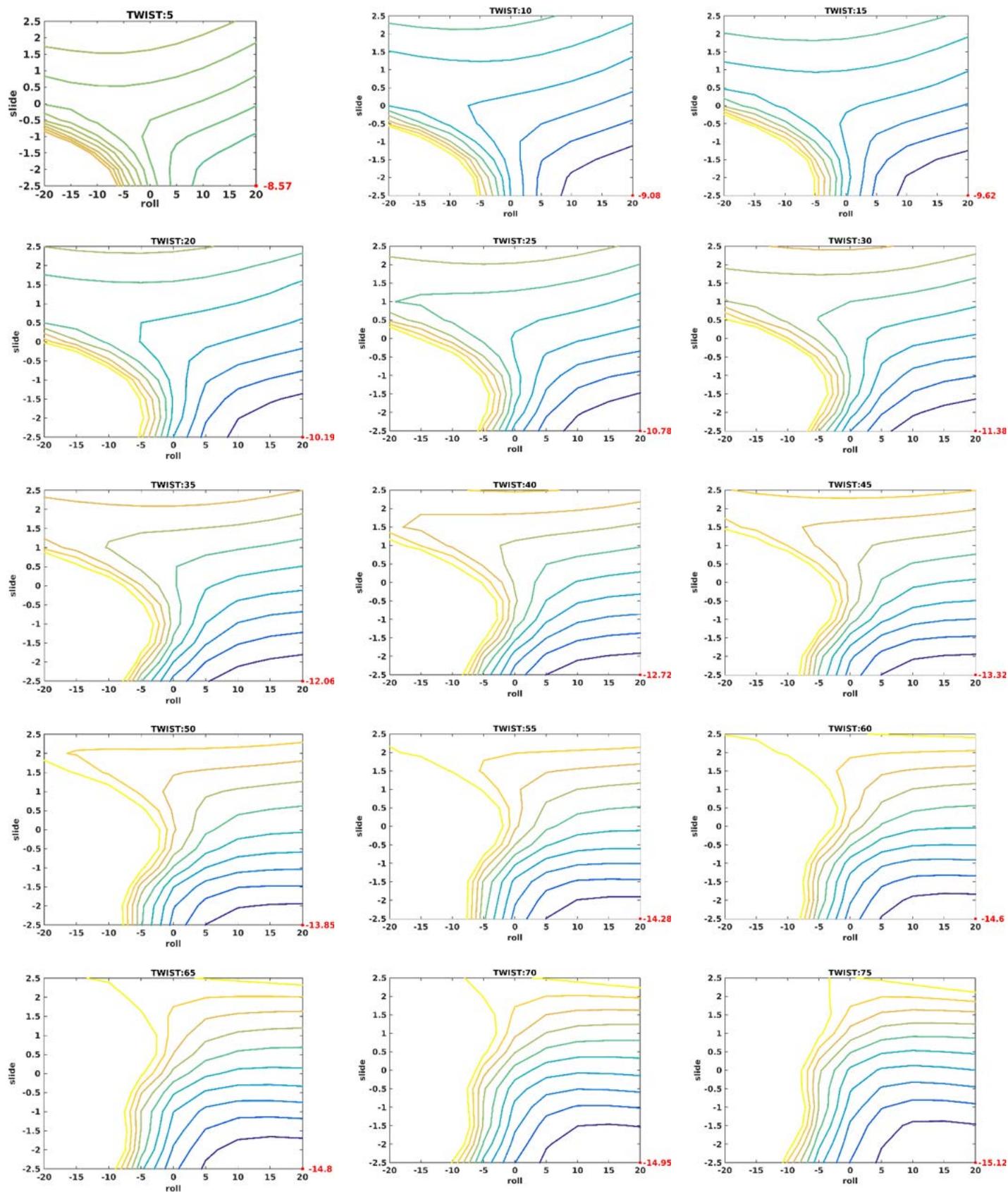
**Figure 3-9** Intrinsic intra-strand stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.

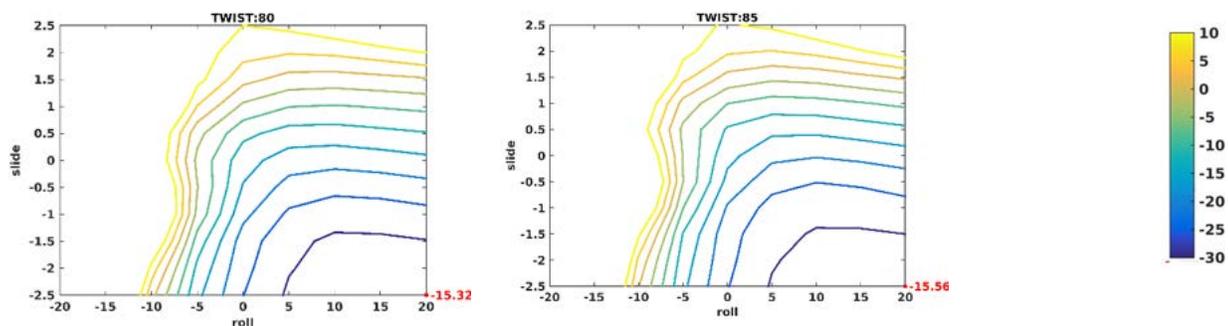




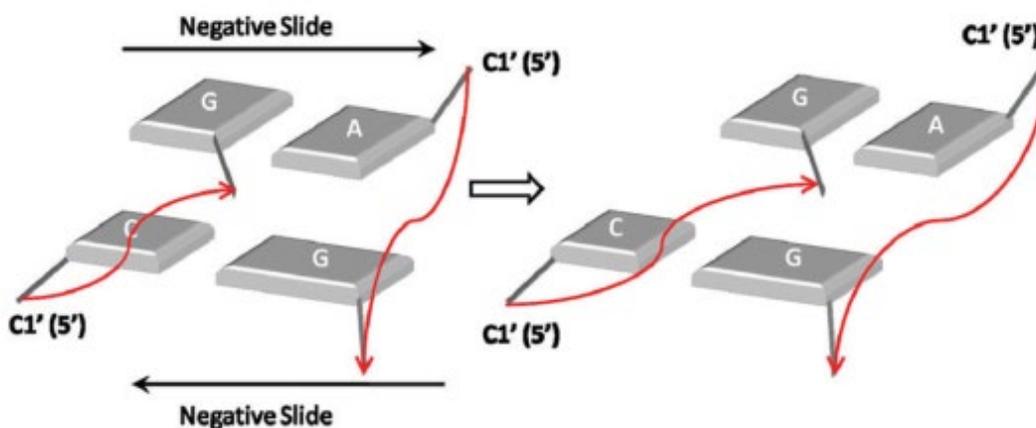
**Figure 3-10** Intrinsic inter strand stacking energy contours considering DFT-D energy at different Twist values for inter strand of G:C W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.

*Stacking Interaction by Most Frequent A:G base pair: Establishment of Hybrid DFT-D Method*





**Figure 3-11** Intrinsic inter strand stacking energy contours considering DFT-D energy at different Twist values for inter strand of C:G W:WC::G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.



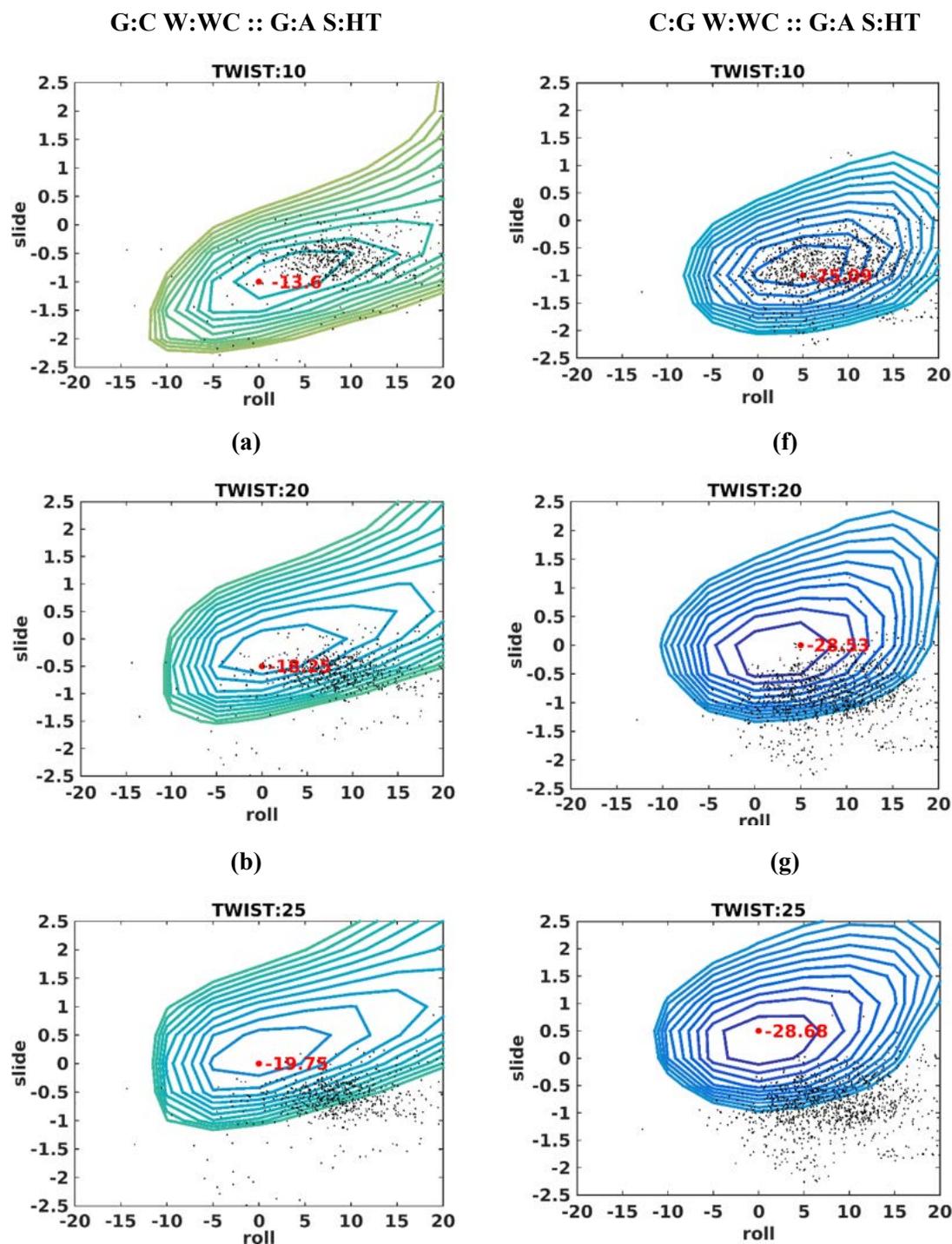
**Figure 3-12** Schematic representation showing that negative Slide gives better cross strand overlap and stacking interaction between two Gua bases.

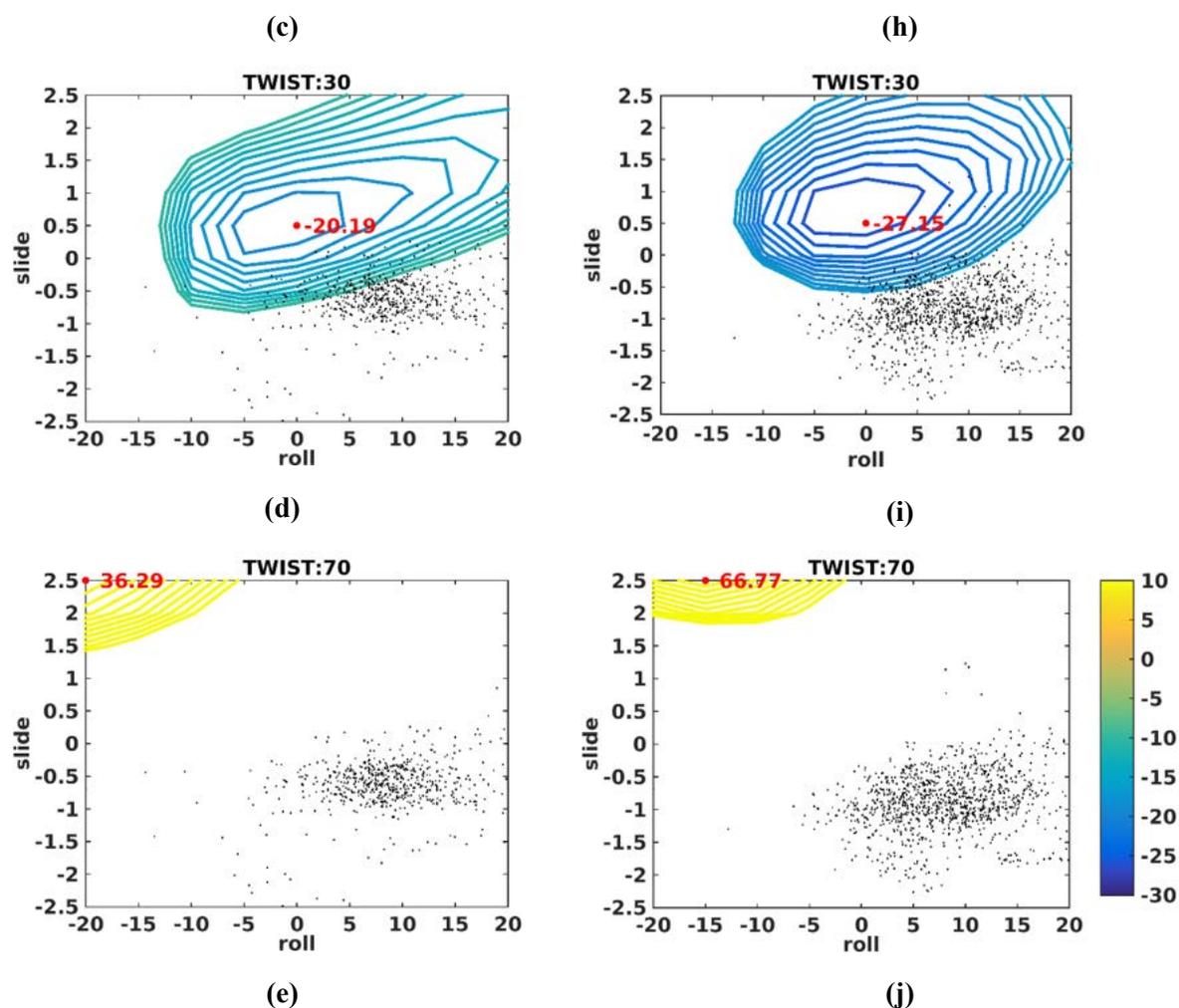
### 3.3.4 Hybrid Stacking Energy Analysis

We have generated the energy contours in Roll-Slide hyperspace for all positive Twist values considering sum total of DFT-D and coarse grain energy penalty values arising from C1'... C1' distance using **equation 9**. We have also plotted the Roll and Slide values from the crystal structures into the contour plots. Representative contour plots are shown in **Figure 3-13**. The best stacking energy for A:U W:WC::C:G W:WC reduces from -22.50 kcal/mol for 10° Twist to -21kcal/mol for 25° Twist (**Table 3-6** and **Figure 3-4** and **Figure 3-14**). The best stacking energy reduces from -23.33 kcal/mol for 70° Twist to -20.19 kcal/mol for 30° Twist for G:C W:WC::G:A

S:HT. Similarly the best stacking energy for C:G W:WC::G:A S:HT sequence reduces from -30.50 kcal/mol corresponding to Twist 20° to -28.68 kcal/mol corresponding to Twist 25° (Table 3-5,

Table 3-7, Figure 3-5, Figure 3-6, Figure 3-15 and Figure 3-16).



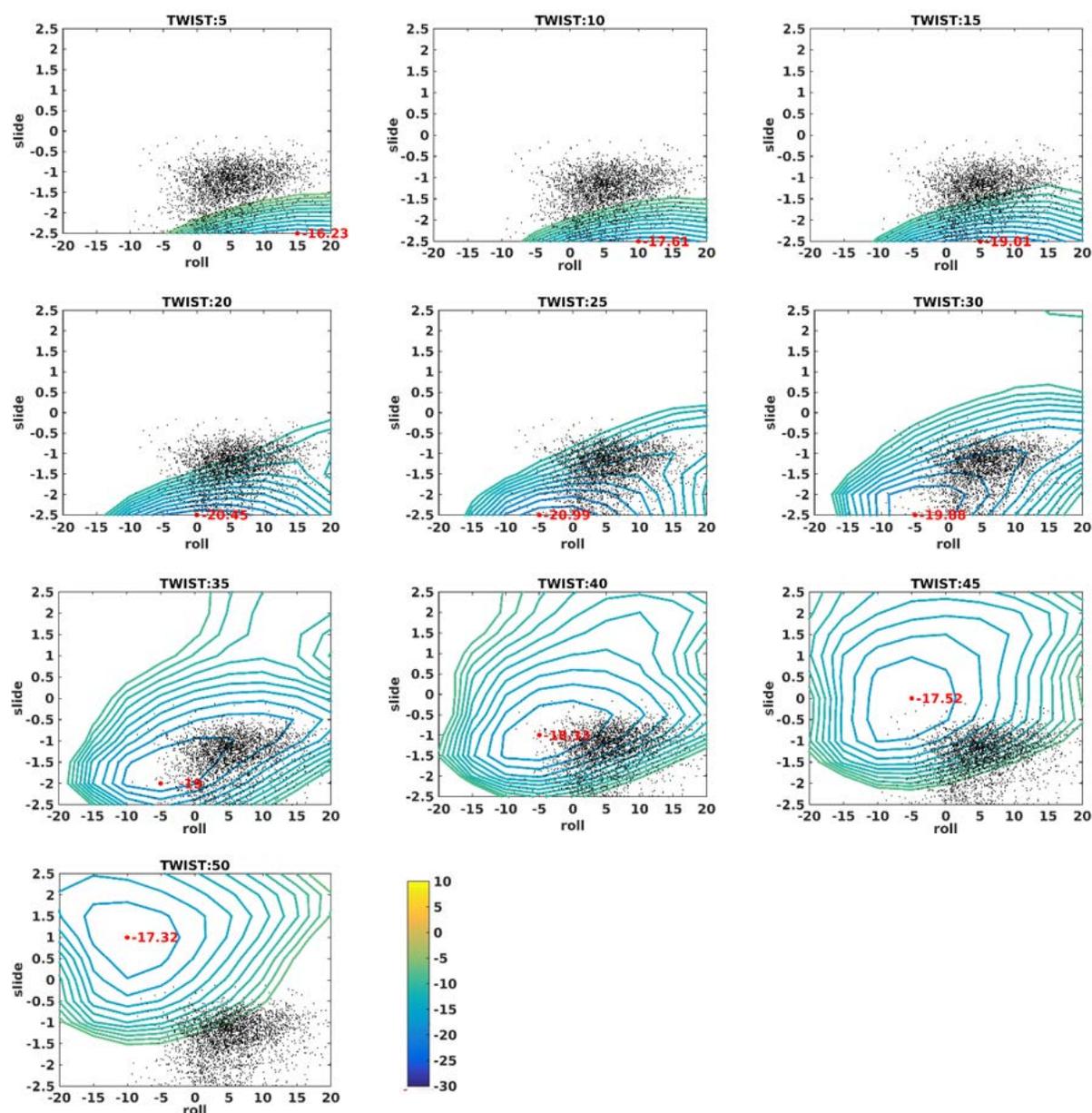


**Figure 3-13** Intrinsic stacking energy contours for (a, b,c,d and e) G:C W:WC::G:A S:HT and (f, g, h, i and j) C:G W:W C::G:A S:HT dinucleotide steps considering DFT-D energies and coarse grain penalty energy values at various Twist values. Color bar is presented in kcal/mol. Difference between two adjacent contour lines is 1 kcal/mol. Roll and Slide values from crystal structure database are marked by black point. The minimum energy conformations for every Twist values are represented by red dots.

**Table 3-6** Geometrical and energy parameters for the structures of the dinucleotide step sequence having minimum energy (sum of DFT-D and coarse-grain energies) corresponding to each Twist value are listed. Mean values of C1' ... C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns.

A:U W:WC::C:G W:WC:						
Twist (°)	Roll (°)	Slide (Å)	Best Stacking Energy (DFT-D and coarse grain penalty component) (kcal/mol)	Penalty Value(Kcal/mol)	C1'...C1' Distance(Å)	
					1 <sup>st</sup> Strand (5.38)	2 <sup>nd</sup> Strand (5.63)
5	15	-2.5	-16.23	3.69	4.96	4.80
10	10	-2.5	-17.62	3.36	4.99	4.83

15	5	-2.5	-19.01	2.59	5.07	4.92
20	0	-2.5	-20.45	1.52	5.20	5.06
25	-5	-2.5	-21.00	0.64	5.37	5.24
30	-5	-2.5	-19.88	0.53	5.72	5.58
35	-5	-2	-19.00	0.63	5.75	5.61
40	-5	-1	-18.33	0.36	5.56	5.40
45	-5	0	-17.53	0.43	5.51	5.34
50	-10	1	-17.32	0.52	5.46	5.29



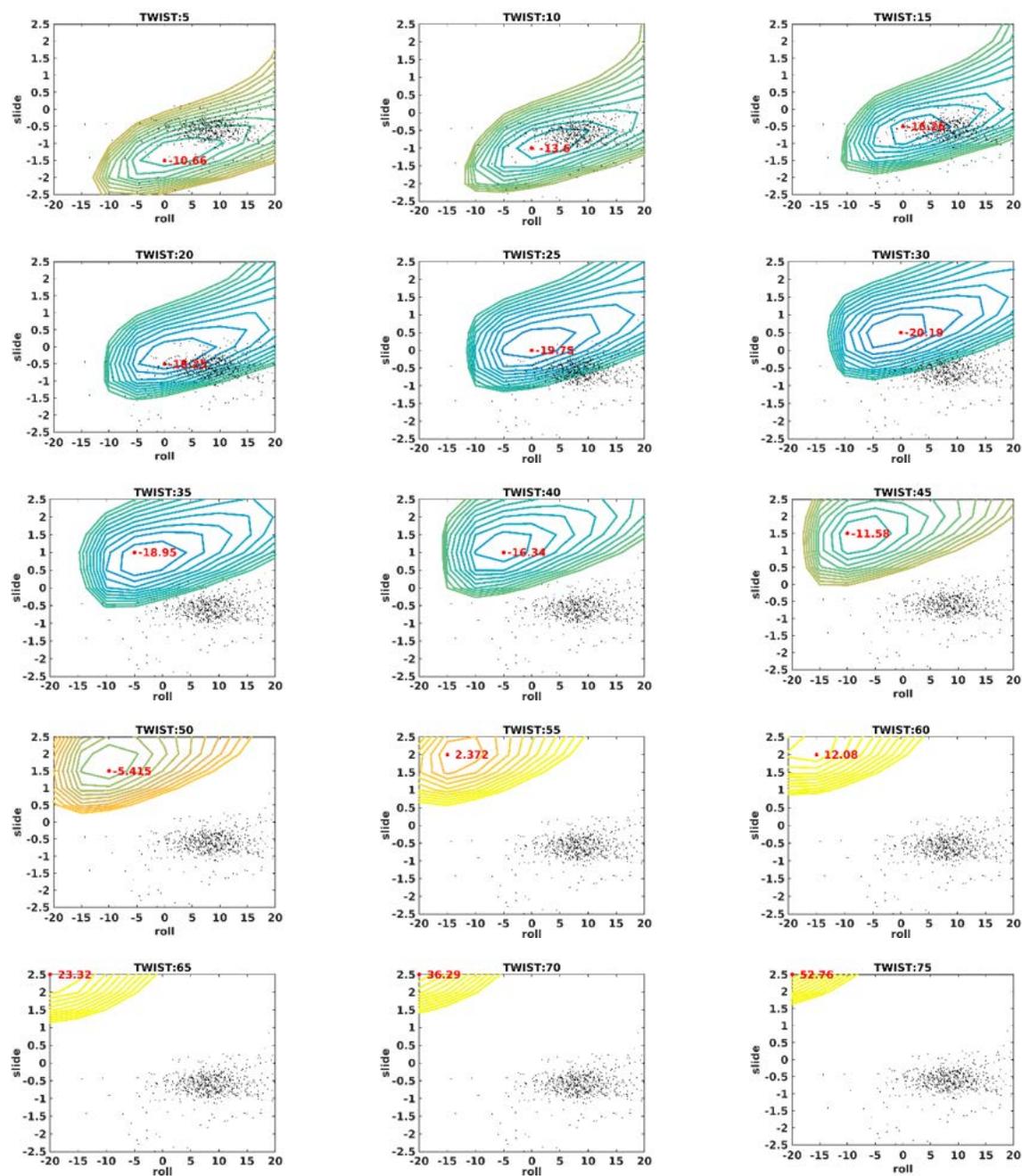
**Figure 3-14** Intrinsic stacking energy contours for A:U W:WC :: C:G W:WC dinucleotide steps considering DFT-D energies and coarse grain penalty energy values at various Twist values. Color bar is presented in kcal/mol. Difference between two adjacent contour lines is 1 kcal/mol. Roll and Slide values from crystal structure database are marked by black point. The minimum energy conformations for every Twist values are represented by red dots

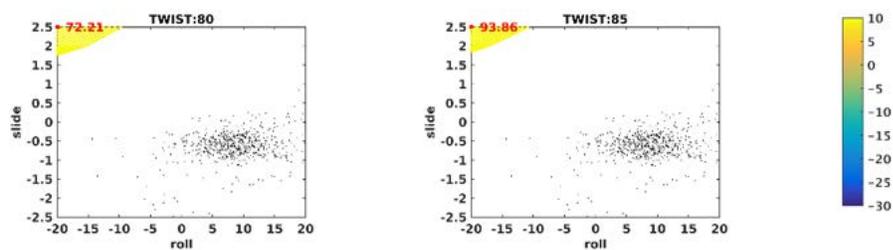
**Table 3-7** Geometrical and energy parameters for the structures of the two sequences having minimum energy (sum of DFT-D and coarse-grain energies) corresponding to each Twist value are listed. Mean values of C1'...C1' distances (Å) from X-ray crystal structure database are also shown in parenthesis in the last two columns.

(a) G:C W:WC::G:A S:HT											
Twist (°)	Roll (°)	Slide (Å)	Best Stacking Energy (DFT-D and coarse grain penalty component) (kcal/mol)	Intra Strand Energy (Kcal/mol)		Inter Strand Energy (Kcal/mol)		Total Strain Energy(kcal/mol)	Penalty Value (kcal/mol)	C1'...C1' Distance (Å)	
				DFT-D component	Dispersion component	DFT-D component	Dispersion component			1 <sup>st</sup> Strand (5.36)	2 <sup>nd</sup> Strand (5.13)
5	0	-1.5	-10.66	-11.41	-13.41	-4.46	-2.59	123.46	3.22	5.48	4.59
10	0	-1	-13.60	-13.83	-14.38	-3.77	-2.2	118.34	2.06	5.27	4.68
15	0	-0.5	-16.26	-16.21	-15.09	-3.32	-1.91	110.72	1.42	5.44	5.06
20	0	-0.5	-18.25	-17.02	-15.02	-3.07	-1.83	110.84	0.10	5.28	5.24
25	0	0	-19.75	-18.85	-15.37	-2.85	-1.68	111.00	0.14	5.15	5.47
30	0	0.5	-20.19	-20.32	-15.39	-2.83	-1.65	116.97	1.19	4.92	5.66
35	-5	1	-18.95	-21.09	-15.89	-3.48	-1.90	123.08	3.79	5.15	6.09
40	-5	1	-16.34	-21.28	-15.46	-3.25	-1.86	131.33	6.48	4.97	6.33
45	-10	1.5	-11.58	-19.84	-15.42	-4.31	-2.37	133.73	10.91	5.21	6.75
50	-10	1.5	-5.41	-20.13	-14.89	-4.23	-2.33	175.11	17.37	5.71	4.58
55	-15	2	2.37	-17.37	-14.48	-5.96	-3.14	183.36	24.01	5.09	7.02
60	-15	2	12.07	-17.99	-13.88	-5.95	-3.14	194.50	34.41	5.35	7.42
65	-20	2.5	23.32	-13.35	-13.10	-8.35	-4.28	411.44	43.45	5.27	7.70
70	-20	2.5	36.29	-14.38	-12.47	-8.37	-4.31	231.11	57.68	5.54	8.09
75	-20	2.5	52.76	-14.57	-11.81	-8.45	-4.38	272.55	74.57	5.80	8.46
80	-20	2.5	72.21	-14.30	-11.12	-8.67	-4.49	366.43	93.90	6.05	8.82
85	-20	2.5	93.86	-13.85	-10.44	-9.03	-4.65	396.70	115.31	6.30	9.17
(b) C:G W:WC::G:A S:HT											

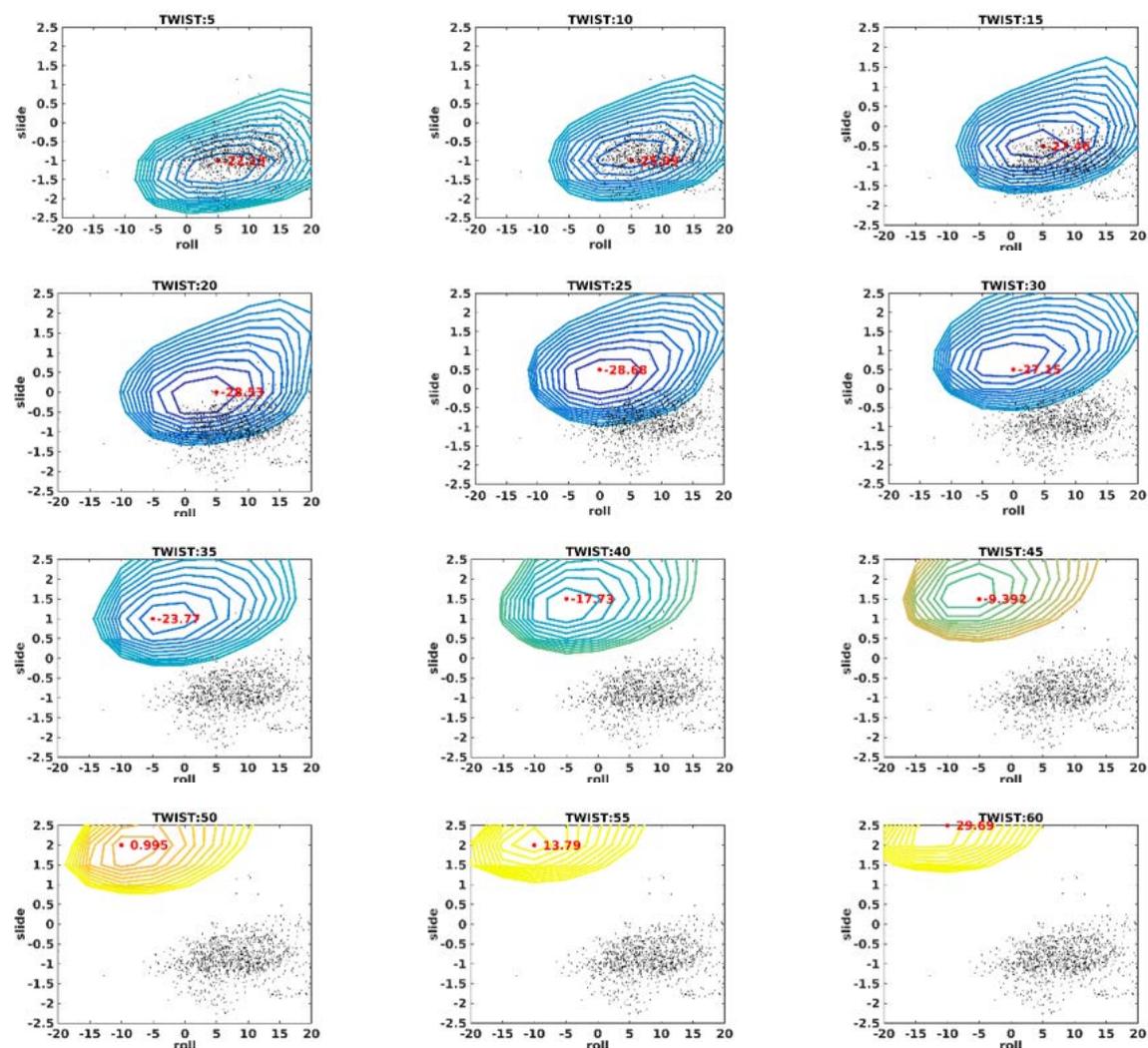
Stacking Interaction by Most Frequent A:G base pair: Establishment of Hybrid DFT-D Method

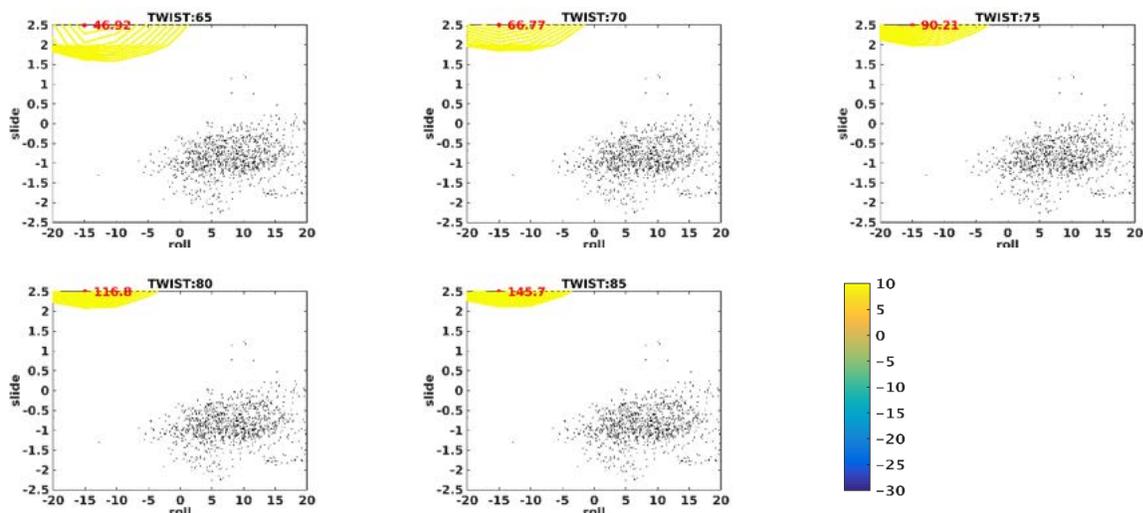
Twist (°)	Roll (°)	Slide (Å)	Best Stacking Energy (DFT-D and coarse grain penalty component) (kcal/mol)	Intra Strand Energy (kcal/mol)		Inter Strand Energy (kcal/mol)		Total Strain Energy(kcal/mol)	Penalty Value (kcal/mol)	C1'...C1' Distance (Å)	
				DFT-D component	Dispersion component	DFT-D component	Dispersion component			1 <sup>st</sup> Strand (5.25)	2 <sup>nd</sup> Strand (5.03)
5	5	-1	-22.13	-20.24	-14.5	-6.39	-4.18	128.18	4.07	5.51	4.65
10	5	-1	-25.09	-20.54	-14.31	-6.62	-4.26	134.80	1.66	5.31	4.76
15	5	-0.5	-27.46	-22.15	-15.4	-6.31	-3.63	132.88	0.60	5.15	4.94
20	5	0	-28.53	-22.99	-16.17	-6.03	-3.09	135.83	0.14	4.87	5.08
25	0	0.5	-28.68	-24.73	-17.32	-5.61	-2.79	134.68	1.16	5.07	5.50
30	0	0.5	-27.15	-23.98	-16.72	-5.84	-2.97	141.22	2.02	4.84	5.71
35	-5	1	-23.77	-24.37	-17.56	-5.44	-2.83	157.64	5.01	4.79	6.05
40	-5	1.5	-17.73	-23.58	-17.1	-5.21	-2.6	170.94	9.92	5.02	6.47
45	-5	1.5	-9.39	-22.40	-16.28	-5.44	-2.84	222.47	17.08	4.88	6.76
50	-10	2	1.00	-20.60	-16.54	-5.00	-3.02	220.54	24.86	5.13	7.17
55	-10	2	13.78	-19.69	-15.65	-5.05	-3.34	178.10	36.79	5.37	4.33
60	-10	2.5	29.69	-17.69	-14.68	-4.80	-3.52	207.89	50.38	5.18	7.54
65	-15	2.5	46.92	-14.95	-14.78	-4.74	-4.09	241.43	64.48	5.31	7.87
70	-15	2.5	66.77	-14.99	-13.87	-4.31	-4.43	221.81	83.96	5.56	8.25
75	-15	2.5	90.21	-14.42	-12.96	-3.49	-4.77	241.22	106.09	5.81	8.63
80	-15	2.5	116.84	-13.57	-12.04	-2.29	-5.19	277.49	130.56	6.06	8.99
85	-15	2.5	145.71	-12.63	-11.14	-1.00	-5.69	259.28	156.97	6.30	9.34





**Figure 3-15** Stacking energy contours considering DFT-D and coarse grain penalty energy at different Twist values for G:C W:WC :: G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.





**Figure 3-16** Stacking energy contours considering DFT-D energy at different Twist values for C:G W:WC :: G:A S:HT dinucleotide step are shown here. Color bar is given in kcal/mol. The contour lines are 1kcal/mol apart.

We have also selected the stacks corresponding to minimum stacking energy including C1'...C1' distance penalty for all values of Twist. Sugar phosphate backbone atoms were generated for these and their constrained energy minimizations were performed. The strain energy values for the best structures for each Twist values are given in

**Table 3-7.** This indicates that minimum strain energies are not really corresponding to the best structures considering DFT-D stacking energies and the coarse grain energy penalties.

Maximum number of experimental data of Roll and Slide for A:U W:WC::C:G W:WC step fall within ( $E_{\min} + 2$ ) kcal/mol contour at Twist  $35^\circ$  which is near the mean Twist value as obtained from X-ray crystallographic data. This contour is covering positive Roll region and negative Slide region which is quite similar to experimental observation (**Figure 3-14**). Although the best stacking energies for G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT sequences at  $30^\circ$  and  $25^\circ$  respectively have  $0^\circ$  Roll value, the ( $E_{\min} + 1$ ) kcal/mol iso-energy contours are seen to cover larger positive Roll region around that. The ( $E_{\min} + 2$ ) and ( $E_{\min} + 3$ ) kcal/mol contours are even more asymmetric and are extended more towards larger positive Roll. The backbone corrected stacking

energy contours show lowest energies near positive Roll and negative Slides in many cases especially for small Twist values. The distribution of Roll and Slide in experimental structures are, however, away from the lowest energy contours in both the sequences. It has been seen that maximum number of experimental data (almost 51%) of Roll and Slide fall within  $(E_{\min} + 2)$  kcal/mol for  $10^\circ$  Twist for C:G W:WC::G:A S:HT (**Figure 3-13f**). Similarly most of the Roll and Slide values (almost 72%) fall within  $(E_{\min} + 2)$  kcal/mol contour for  $10^\circ$  Twist for G:C W:WC::G:A S:HT (**Figure 3-13a**). It may be noted that these Twist, Roll and Slide values for above configurations for both dinucleotide base pair step sequences fall near the corresponding experimental mean values and these values are close to those obtained in RNA structures formed by Watson-Crick base pairs<sup>108</sup>.

The reason behind unfavorable stacking energy may be due to total bonding strain energy ( $E_{\text{strain}}$ ) of the systems. It has been seen that total classical bonding energy for sugar-phosphate backbone connection for the sequences at unfavorable stacking energy zone are quite high. We have noticed  $E_{\text{strain}}$  for G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT sequences are minimum for  $15^\circ$  and  $5^\circ$  Twist, respectively. Although the coarse grain backbone corrected minimum energy for above mentioned base pair steps are different. Total of 55% and 48% of experimental data points are found within  $(E_{\min} + 2)$  kcal/mol contour of the G:C W:WC :: G:A S:HT and C:G W:WC :: G:A S:HT sequences at  $15^\circ$  and  $5^\circ$  Twist, respectively. It is worth mentioning that mean Twist value of the G:C W:WC::G:A S:HT and C:G W:WC::G:A S:HT sequences are at  $11.3^\circ$  and  $7.1^\circ$  Twist values close to the above.

As indicated earlier, the stacking energy is stronger in the sequence C:G W:WC::G:A S:HT than that of other sequence. The C:G W:WC::G:A S:HT sequences are also found more frequently in the crystal structure database. This difference in energy is possibly due to extent of stacking

overlap area between the two successive base pairs. Maximum overlap area of C:G W:WC::G:A S:HT sequence is greater than that of G:C W:WC::G:A S:HT sequence by about  $7\text{\AA}^2$  considering configurations of best energy of the respective systems.

### **3.4 Conclusion**

Our analysis indicate that stacking energy, estimated by  $\omega$ B97X-D/cc-pVDZ augmented with coarse grain energy penalty from C1'...C1' distances along the strands, is a reasonable approach for prediction of structures of base pair stacks even containing non Watson-Crick base pairs. Similar agreement between stacking energy based prediction of base pair parameters and experimental values for dinucleotide stacks made of only Watson-Crick base pairs were shown earlier. Here we could demonstrate accuracy of the method for prediction of structures of double helices containing non Watson-Crick base pairs. Consideration of coarse grain energy penalty due to elongation of C1'...C1' distances along the strands, from their mean values, seem to have some limitations. The minimal disagreement between predicted and observed structural parameters may also arise from complete neglect of interaction of the bases and the charged phosphate groups. Consideration of such effect might improve our prediction algorithm. Interaction energy for A:U W:WC and G:A S:HT base pairs appear to be weaker than the stacking energies for the most stable conformations in all the three systems studied.

Our earlier analysis of stacking energy between Watson-Crick base pairs, and those between G:U W:WC and different Watson-Crick base pairs and the present analysis gives us enough confidence in predicting stacking orientations in double helices found by all other non-Watson-Crick base pairs. The study also rationalizes some reason behind G:A S:HT non Watson-Crick base pairs not appearing within DNA structures. We strongly feel that it would be extremely useful for prediction of secondary structures of different coding and non-coding RNA sequences.

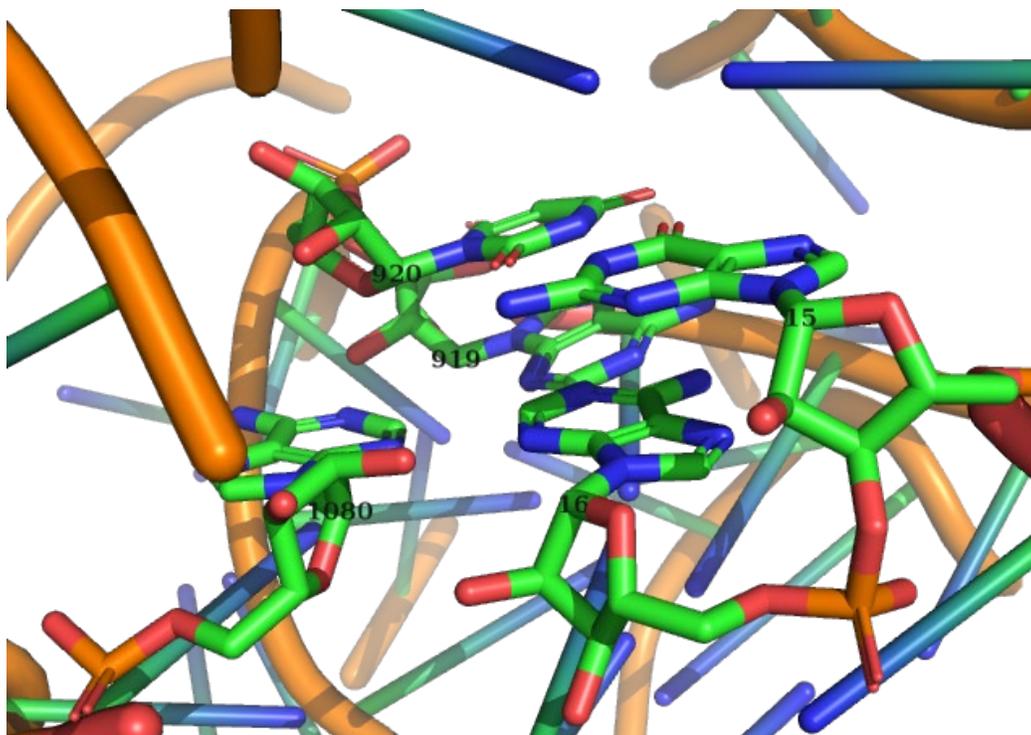
## **Chapter- 4 Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT-D and MD Studies**

## 4.1 Introduction

Structures of RNA in different functional forms, apart from messenger RNA, are composed of many double helical segments stabilized by base pairing and stacking interactions. These double helical motifs are mostly stabilized by base pairs having hydrogen bonding between the bases of complementary strands or regions by formation of G:C or A:U Watson-Crick base pairs. But unlike in DNA, many different types of non-canonical base pairs are also seen quite frequently in the functional RNA motifs. The wobble G:U base pair is very frequently found in such double helices, which have inherent stable shearing motion between the bases for formation of two hydrogen bonds<sup>168,169</sup>. As hydrogen bonding donor and acceptor atoms are present in all the bases at various locations, the bases can form various types of specific pairing by appropriate molecular recognition. Such pairing, by formation of two or more hydrogen bonds, is possible through three edges, namely Watson-Crick (W), Hoogsteen (H) and Sugar (S) edges, of each base. The base pairs, can further be in *cis* or *trans* orientations with respect to the hydrogen bonds<sup>170</sup>. This gives rise to several possible non-canonical base pairs, most of them are quite frequently observed in the available experimentally determined X-ray crystallographic or NMR derived structures of various functional RNA. Large number of reports exists focusing on structure, stability and dynamics of these non-canonical base pairs, which was compiled recently in RNABP COGEST database<sup>96</sup> and RNABPDB database (<http://hdrnas.saha.ac.in/rnabpdb/>).

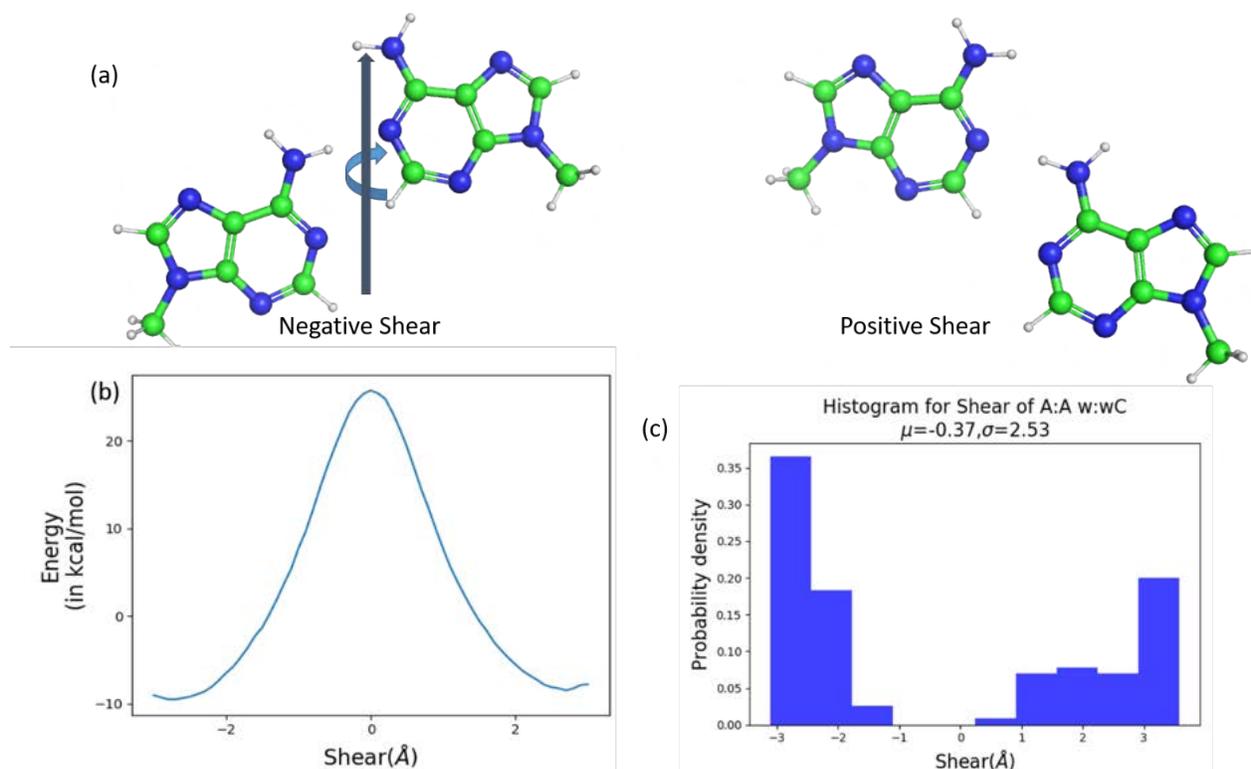
Two adenine residues can form base pairs in ten different orientations by formation of at least one polar and at most one non-polar (C-H...O/N) hydrogen bonds. Among these non-canonical base pairs, A:A w:wC (or cWW in Leontis-Westhof nomenclature<sup>171</sup>) is unique in its presence in double helical segments at vital locations in many important functional RNAs, such as Ribozyme, Ribosome, etc. Occurrence of A:A base pair in bacterial RNA, GAAA duplex of

hammerhead ribozyme and GAAA tetra-loop in Group I intron of tetrahymena ribozyme have been studied<sup>137,172–174</sup>. This base pair is known to be crucial in recognition of proper codon-anticodon pairing at the translational site of ribosome decoding center<sup>175</sup>. It is worth mentioning that this base pair often forms base triplet with another Adenine when it is stacked with G:U base pair (**Figure 4-1**). Hence it is possible that it may appear in some other important RNA structures as well, such as miRNA, which is known to have different types of internal loops<sup>176</sup>. The A:A w:wC base pair is stabilized by a polar N—H...N and a weak C—H...N hydrogen bonds and it is mostly observed as stacked on top of other base pairs. Furthermore these hydrogen bonds can form in two sheared geometries of the base pair which are related by two-fold rotational symmetry. On the other hand two adenine residues can also form stable sheared tSH (tran sugar edge / Hoogsteen edge) base pairing within various systematic internal loops<sup>177–179</sup>.



**Figure 4-1** A:A w:wC base pair forming triplet with another Ade in 16S ribosomal RNA (PDB ID. 5J5B) structure of *E. coli*.

Considering availability of large number of experimental structures of different RNA molecules, it is now possible to predict structure of an RNA fragment whose experimental data is not available. This can be achieved from mean values of relative orientation parameters of the bases of a base pair and base pairs of dinucleotide steps as suggested by IUPAC-IUB<sup>39</sup>, following bioinformatics approach. Most of the non-canonical base pairs and their dinucleotide stacks show normal distributions of their parameters in RNABPDB database (<http://hdrnas.saha.ac.in/rnabpdb/>) allowing one to model build structure using RNAHelix<sup>89</sup> X3DNA<sup>109</sup> etc. using the average values of the structural parameters. The inherent symmetry of the A:A w:wC base pair, however, lead to bimodal distribution of its shear values (**Figure 4-2**), leading to unsuitable prediction of structure containing this base pair. It is clear that the mean value of Shear is most improbable for the A:A w:wC base pair. The symmetry, however, may break when the A:A base pair stacks on some other base pair. As for example, most of the A:A w:wC stacked on C:G W:WC base pair near ribosome decoding center, have positive Shear values. Especially prominent symmetry breaking is possible when A:A w:wC base pair stacks on the sheared wobble G:U base pair. Although the frequency of stacking between A:A w:wC followed by U:G W:WC (in 5' to 3' direction from A to U), the A:A w:wC :: U:G W:WC dinucleotide ( “:” represents base pairing and “::” represents stacking) is somewhat significant in the RNABPDB database the other combination for A:A followed by G:U is quite rare. The frequencies are not however sufficient to arrive at experimental preferred geometry of the dinucleotide steps which can be used for bioinformatics driven structure prediction. Presence of non-canonical base pairs within RNA double helical regions demands good stacking interaction between them and their neighboring ones. Such stacking interactions by isolated bases are also found to stabilize terminal base pairs in Kink-turn motifs<sup>93</sup>.



**Figure 4-2** (a) Negative and Positive shear configurations of A:A w:wC base pair becomes superposable upon 180° rotation about the pseudo-dyad symmetry axis shown. (b) Interaction energy of A:A w:wC base pair for different Shear values in the range between -3.00Å and +3.00Å. (c) Distribution of Shear value for the A:A w:wC base pair obtained from X-ray Crystal structure. Mean Shear for these structures is 0.37Å with 2.54 being the standard deviation.

Similarly, a base triplet involving two consecutive bases (commonly known as dinucleotide platform) is known to stabilize two double helical regions of Sarcin-ricin domain by stacking interaction, which allow a shift in helix direction<sup>174</sup>. Calculation of base pairing energies is quite routine procedure now on availability of good computational facility and efficient DFT functionals, but proper estimation of stacking interaction requires higher level theory, which incorporates electron correlation as well as dispersion interactions. Many dispersion-corrected DFT functionals (DFT-D) are now available, which can estimate reliable stacking interaction between the polar heterocyclic bases<sup>180</sup>. Formation of double helical segment by non-canonical base pairs further require the effect of sugar-phosphate backbone as the base pairs need to be connected to the stacked ones by proper covalent bonds. Recent studies indicate that the hybrid

stacking energy from DFT-D and coarse-grain methods, for estimating effect of sugar-phosphate linkage, is highly accurate for predicting structures of base pair stacks involving Watson-Crick as well as non-canonical base pairs<sup>75,76</sup>.

We have, therefore attempted to predict the preferred geometry of the dinucleotide step formed by A:A w:wC and G:U W:WC base pairs as 5'(AG)3'.5'(UA)3' dinucleotide using theoretical calculation. As the two strands are anti-parallel, the same dinucleotide can also be written as 5'-(UA).5'-(AG), U:G W:WC::A:A w:wC or A:A w:wC::G:U W:WC. Considering context dependent preference of Shear value by the A:A w:wC base pair we have performed 500ns long MD simulations of three possible motifs constituted by A:A w:wC and G:U W:WC base pairs to confirm preferred value of Shear for A:A base pair when it is surrounded by two G:U base pairs on its either sides. These indicate that the RNA double helix containing 5'(GAG).5'(UAU) motif adopts stable structure with negative shear for the A:A base pair. The A:A w:wC base pair in the 5'(UAG).5'(UAG) and 5'(GAU).5'(GAU) appear unstable. We have further carried out energy landscape scanning using different levels of quantum chemistry methods, such as DFT-D of different types and MP2, for both stacks to understand reason behind the disorder of the sequence motif. The hybrid quantum chemical calculations indicate that A:A w:wC base pair followed by G:U W:WC base pair, always prefers negative Shear while the same A:A w:wC followed by U:G W:WC base pair demands positive Shear. It has been seen that the quantum chemical methods are consistent in predicting structures suitable for A-RNA like double helix formation. Energy consideration also predicted a structure with positive Slide, which is possible in B-DNA conformations. Previous studies by various methods indicated, however, that presence of 2'-OH groups induces A-form structure to nucleic acid double helices with large negative Slide<sup>181-183</sup>. Hence the predicted structure with positive Slide may be discarded as suitable for RNA.

Preference of two opposite Shear values for A:A w:wC :: G:U W:WC and A:A w:wC :: U:G W:WC stacks imparts tug of war situation on A:A w:wC base pair. In 5'(GAU).5'(GAU) sequence the G:U base pair at the 5' and 3' sides of the A:A base pair pull the Adenine of second strand in opposite directions resulting disordered structure. Similar structural plasticity of A:A h:sT base pair in double helical motif was studied recently where structural alterations between the major and minor forms were seen to be inter-convertible by molecular dynamics simulations<sup>184,185</sup>. Our studies, however, predict a possible stable sequence motif and two sequence motifs which may show conformational plasticity.

#### **4.2 Materials and Methods:**

All the non-redundant structures of RNA solved by X-ray crystallography with resolution better than 3.0Å were analyzed by BPFIND<sup>88</sup> to identify different base pairs and NUPARM<sup>111,116,186</sup> to obtain the intra base pair and dinucleotide orientation parameters of all the base pairs and stacking arrangements. These are used to populate the RNABPDB database (<http://hdrnas.saha.ac.in/rnabpdb/>).

Stacking interaction energies with BSSE corrections<sup>86</sup> have been calculated for the structures of all the four base pair step for A:A w:wC::G:U W:WC. We have also calculated BSSE corrected interaction energies for the H-optimized structure of each base pair in the dinucleotide steps. We have calculated stacking interaction energy using  $\omega$ B97X-D<sup>82</sup>/cc-pVDZ<sup>126,127</sup>, M06-2X<sup>187</sup>/cc-pVDZ, MP2<sup>167,188</sup>/cc-pVDZ levels of theory by Gaussian16<sup>125</sup>. The effect of solvent polarizability was also analyzed by implementation of conductor-like polarizable continuum model through united-atom topological model with  $\epsilon = 78.39$  (CPCM)<sup>189</sup>.

Model structures of the dinucleotide step sequence have been generated by varying Twist value from 5° to 60° with 5° interval.

**Table 4-1** Description base pair and Base pair Step parameter for modelling of base pair steps.

<b>A.</b>						
Base pair	Buckle (°)	Open (°)	Propeller (°)	Stagger (Å)	Shear (Å)	Stretch (Å)
A:A w:wC	-1.47	19.06	-13.13	-0.61	+2.7, -2.7	2.42
G:U W:WC	-1.48	-0.30	-9.07	-0.10	-2.25	2.81
<b>B.</b>						
Base pair Step	Tilt(°)	Roll (°)	Twist (°)	Shift (Å)	Slide (Å)	Rise (Å)
A:A w:wC :: G:U W:WC	0.00	-20 to +20	5 to 60	0.00	-2.5 to +2.5	3.16

We have varied Roll between -20° and +20° in the steps of 5° and Slide between -2.5Å and +2.5Å in step of 0.5Å, for each Twist value. Thus Total  $12 \times 9 \times 11 \times 2 = 1188 \times 2 = 2376$  structures have been generated for two Shear values for the stacking energy calculations.

We have modeled 61 structures of A:A w:wC base pair by varying Shear value between -3.0Å to +3.0Å using RNAHelix<sup>89</sup> software keeping the other base pair parameters as their mean values obtained from RNABPDB. Molecular Modeling of the A:A w:wC :: G:U W:WC, A:A w:wC :: U:G W:WC and A:A w:wC :: C:G W:WC dinucleotide step sequences have also been carried out by RNAHelix software using suggested base pair and base pair step parameter (**Table 4-1**) Model structures of the dinucleotide step sequence have been generated for 1188 conformations for each Shear value by changing Twist, Roll and Slide in wide range. DFT-D and MP2 stacking energies of these model dinucleotide steps have been calculated using the above mentioned three methods. The effect of sugar-phosphate backbone has been introduced using C1'-C1' distance of each strand and added to quantum chemically calculated stacking interactions to get the Hybrid stacking energies expressed as <sup>64,76</sup>.

$$E_{\text{hybrid stacking energy}} = E_{\text{stacking energy}} + \frac{1}{2}k_1(d_1 - d_1^0)^2 + \frac{1}{2}k_2(d_2 - d_2^0)^2$$

Where  $d_1^0$  and  $d_2^0$  are the mean  $C1' \dots C1'$  distances of the two strands. As mean  $C1' \dots C1'$  distance in Watson-Crick base pair step is  $5.50\text{\AA}$ , we have taken this value during the estimation of Hybrid stacking energy.  $d_1$  and  $d_2$  are the  $C1' \dots C1'$  distances for model dinucleotide step.

Stacking overlap between two base pairs for all the models were calculated using NUPARM software<sup>116</sup>.

We have generated B-DNA double helical structure having only canonical Watson-Crick base pairs for  $5' - \text{AAUUUGCCAAUA} - 3'$  sequence (Table 4-2) from Arnotts' X-ray fiber diffraction derived model<sup>190</sup>. Thymine residues of this DNA model were converted to Uracil and the deoxyribose sugars were also converted to ribose sugars using CHARMM<sup>128</sup>. Eventually this model gave us a structure of RNA in B-DNA conformation (or hypothetical B-RNA).

**Table 4-2** Sequences of the double helices studied using molecular dynamics simulation.

System	Sequence	Sugar type	Parameter type	Share value
1 <sup>st</sup>	$5' - \text{AAUUUGCCAAUA} - 3'$ $3' - \text{UUAAACGGUUAU} - 5'$	Ribose	B-DNA	Usual
2 <sup>nd</sup> or GAG motif	$5' - \text{CGAGAGAGCG} - 3'$ $3' - \text{GCUUAUUCGC} - 5'$	Ribose	RNA	+2.70(A:A w:wC)
3 <sup>rd</sup> or GAG motif	$5' - \text{CGAGAGAGCG} - 3'$ $3' - \text{GCUUAUUCGC} - 5'$	Ribose	RNA	-2.70(A:A w:wC)
4 <sup>th</sup> or UAG motif	$5' - \text{CGAUAGAGCG} - 3'$ $3' - \text{GCUGAUUCGC} - 5'$	Ribose	RNA	+2.70(A:A w:wC)
5 <sup>nd</sup> or UAG motif	$5' - \text{CGAUAGAGCG} - 3'$ $3' - \text{GCUGAUUCGC} - 5'$	Ribose	RNA	-2.70(A:A w:wC)
6 <sup>th</sup> or GAU motif	$5' - \text{CGAGAUAGCG} - 3'$ $3' - \text{GCUUAGUCGC} - 5'$	Ribose	RNA	+2.70(A:A w:wC)
7 <sup>th</sup> or GAU motif	$5' - \text{CGAGAUAGCG} - 3'$ $3' - \text{GCUUAGUCGC} - 5'$	Ribose	RNA	+2.70(A:A w:wC)

We have modelled structures of three RNA double helices for  $5' - \text{CGAGAGAGCG} - 3'$  sequence,  $3' - \text{GCUUAUUCGC} - 5'$

the GAG motif,  $5' - \text{CGAUAGAGCG} - 3'$  sequence, the UAG motif and  $3' - \text{GCUGAUUCGC} - 5'$

5' -CGAGAUAGCG -3'  
 3' -GCUUAGUCGC -5' sequence, the GAU motif, containing G:U W:WC base pairs on either side of A:A w:wC. The initial structures were built with both positive (+2.70) and negative Share (-2.70) for the A:A w:wC base pair (Table 4-2). We have used RNAHelix web server at (<http://hdrnas.saha.ac.in/Tools/RNAHelix/>) for building these models. This server uses average values of all the intra-base pair and inter-base pair step parameters for each base pair and base pair step from the RNABPDB database (Table 4-3) and also allows the users to change any desired parameter. All of the seven structures were processed for molecular dynamics (MD) simulations for 500ns each using ff99bsc0 with  $\chi_{OL3}$  correction<sup>191-193</sup> by GROMACS simulation package (version 5.1)<sup>194</sup>.

**Table 4-3** (A) Base pair and (B) Base pair Step parameters for different sequence during RNA double helix generation by RNAHelix software.

A.								
Sequence	Base pair	Buckle(°)	Open(°)	Propeller(°)	Stagger(Å)	Shear(Å)	Stretch(Å)	Base pair type
5' - CGAGAGAGCG -3' 3' - GCUUAGUCGC -5' System: 2 <sup>nd</sup> and 3 <sup>rd</sup>	C:G	5.85	0.92	-8.74	-0.11	0.09	2.86	W:WC
	G:C	-5.89	0.92	-8.76	-0.11	-0.08	2.87	W:WC
	A:U	-2.57	3.76	-9.65	-0.06	-0.06	2.82	W:WC
	G:U	-1.49	-0.31	-9.06	-0.10	-2.25	2.81	W:WC
	A:A	-0.93	19.07	-13.18	-0.60	-2.71, +2.71	2.50	w:wC
	G:U	-1.49	-0.29	-9.05	-0.10	-2.24	2.81	W:WC
	A:U	-2.57	3.78	-9.65	-0.06	0.16	2.82	W:WC
	G:C	-5.86	0.92	-8.76	-0.11	-0.09	2.86	W:WC
	C:G	5.87	0.94	-8.75	-0.11	0.09	2.86	W:WC
	G:C	-5.86	0.91	-8.74	-0.11	-0.09	2.86	W:WC
5' - CGAUAGAGCG -3' 3' - GCUUAGUCGC -5' System: 4 <sup>th</sup> and 5 <sup>th</sup>	C:G	5.86	0.90	-8.77	-0.11	0.09	2.86	W:WC
	G:C	-5.86	0.91	-8.76	-0.11	-0.09	2.86	W:WC
	A:U	-2.57	3.74	-9.65	-0.06	0.16	2.82	W:WC
	U:G	1.48	-0.33	-9.06	-0.10	2.25	2.81	W:WC
	A:A	-0.83	19.07	-13.18	-0.60	-2.71, +2.71	2.50	w:wC
	G:U	-1.49	-0.29	-9.05	-0.10	-2.25	2.81	W:WC
	A:U	-2.57	3.78	-9.65	-0.06	0.16	2.82	W:WC
	G:C	-5.86	0.92	-8.76	-0.11	-0.09	2.86	W:WC
	C:G	5.87	0.94	-8.76	-0.11	-0.09	2.86	W:WC

	G:C	-5.86	0.91	-8.74	-0.11	-0.09	2.86	W:WC
5' - CGAGAUAGCG -3' 3' - GCUUAGUCGC -5' System:6 <sup>th</sup> and 7 <sup>th</sup>	C:G	5.86	0.92	-8.75	-0.11	0.09	2.86	W:WC
	G:C	-5.86	0.95	-8.76	-0.11	-0.09	2.86	W:WC
	A:U	-2.58	3.76	-9.64	-0.06	0.16	2.82	W:WC
	G:U	-1.48	-0.31	-9.06	-0.10	-2.25	2.81	W:WC
	A:A	-0.83	19.07	-13.16	-0.60	-2.71, +2.71	2.50	w:wC
	U:G	1.49	-0.32	-9.07	-0.10	2.25	2.81	W:WC
	A:U	-2.57	3.78	-9.65	-0.06	0.16	2.82	W:WC
	G:C	-5.86	0.92	-8.76	-0.11	-0.09	2.86	W:WC
	C:G	5.87	0.94	-8.75	-0.11	0.09	2.86	W:WC
G:C	-5.86	0.91	-8.74	-0.11	-0.09	2.86	W:WC	
<b>B.</b>								
Sequence	Base pair Step	Tilt(°)	Roll(°)	Twist(°)	Shift(Å)	Slide(Å)	Rise (Å)	Base pair type
5' - CGAGAGAGCG -3' 3' - GCUUAUUCGC -5' System:2 <sup>nd</sup> and 3 <sup>rd</sup>	C:G	0.08	11.77	30.18	0.03	-1.85	3.08	W:WC
	G:C	0.32	5.66	31.11	-0.02	-1.65	3.29	W:WC
	A:U	-1.10	10.00	36.67	0.03	-1.84	3.15	W:WC
	G:U	0.00	4.28	28.35	-0.09	-2.16	3.07	W:WC
	A:A	-8.25	4.74	32.98	0.00	-1.61	3.03	w:wC
	G:U	-1.21	5.33	26.27	-0.36	-1.38	3.33	W:WC
	A:U	0.90	8.32	31.19	0.16	-1.59	3.24	W:WC
	G:C	-0.06	3.37	31.72	0.00	-1.56	3.36	W:WC
	C:G	0.07	11.77	31.72	0.00	-1.85	3.08	W:WC
G:C	-	-	-	-	-	-	-	W:WC
5' - CGAUAGAGCG -3' 3' - GCUGAUUCGC -5' System:4 <sup>th</sup> and 5 <sup>th</sup>	C:G	0.07	11.76	30.18	0.03	-1.85	3.08	W:WC
	G:C	0.32	5.66	31.11	-0.02	-1.65	3.29	W:WC
	A:U	0.92	4.99	27.32	0.35	-1.22	3.38	W:WC
	U:G	8.26	4.75	32.99	0.00	-1.61	3.03	W:WC
	A:A	-8.25	4.74	32.98	0.00	-1.61	3.03	w:wC
	G:U	-1.22	5.33	26.27	-0.36	-1.38	3.33	W:WC
	A:U	0.90	8.32	31.19	0.16	-1.59	3.24	W:WC
	G:C	-0.06	3.37	31.72	0.00	-1.57	3.36	W:WC
	C:G	0.07	11.77	30.18	0.03	-1.85	3.08	W:WC
G:C	-	-	-	-	-	-	-	W:WC
5' - CGAGAUAGCG -3' 3' - GCUUAGUCGC -5'	C:G	0.08	11.77	30.18	0.03	-1.85	3.08	W:WC
	G:C	0.32	5.66	31.12	-0.02	-1.65	3.29	W:WC
	A:U	-1.11	10.01	36.66	0.03	-1.84	3.15	W:WC
	G:U	0.01	4.29	28.36	-0.09	-2.16	3.07	W:WC
	A:A	-0.01	4.27	28.35	0.09	-2.16	3.07	w:wC
	U:G	1.06	14.11	36.29	-0.01	-1.75	3.11	W:WC
	A:U	0.90	8.32	31.19	0.16	-1.59	3.24	W:WC
	G:C	-0.06	3.37	31.72	0.00	-1.57	3.36	W:WC
	C:G	0.07	11.77	30.18	0.03	-1.85	3.08	W:WC

	G:C	-	-	-	-	-	-	W:WC
--	-----	---	---	---	---	---	---	------

The structures were placed in cubic boxes having at least 15Å distance from their edges in periodic boundary condition. The boxes were filled with TIP3P water molecules as explicit solvent and the systems were neutralized with randomly placed charge neutralizing sodium (Na<sup>+</sup>) ions. This produces salt concentration of about 0.1 M, equivalent to physiological concentration. Energy minimizations using steepest descent method were carried out for 50000 steps followed by NVT and NPT equilibration for 100ps each, restraining positions of the atoms of RNA. 10Å cut off was employed for Lennard-Jones and short-range Coulombic interactions. Particle Mesh Ewald summation (PME) method<sup>195</sup> was applied to evaluate long-range electrostatic interactions. Pressure was maintained at 1.0 atm by Parrinello-Rahman algorithm<sup>196</sup> with a time constant of 2ps. Constant temperature of 300K was maintained using velocity rescaling<sup>197</sup> with a time constant of 1ps. Analysis of trajectory was done using GROMACS utilities, NUPARM<sup>111,186</sup> and BPFIND<sup>88</sup>. Free energies were evaluated from population analysis of Shear and Open using Gibbs distribution formula<sup>198,199</sup>

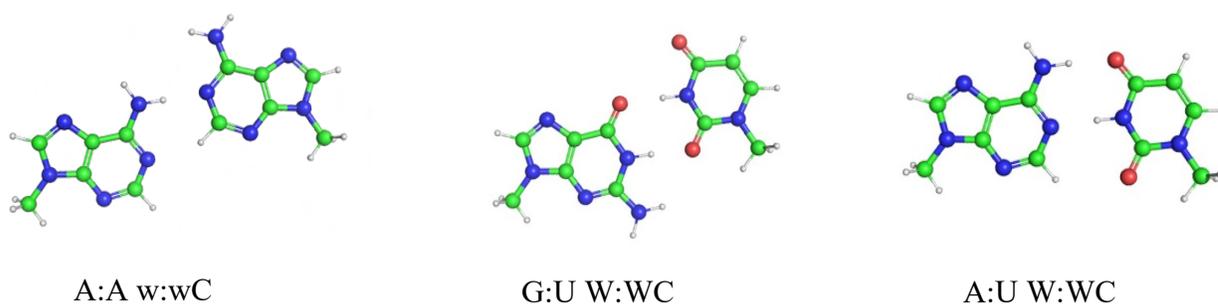
$$F = -k_B T \ln(p\{Shear, Open\})$$

$$p\{Shear, Open\} = \frac{(No.configs.Shear, Open)}{(No.ofSnapshots)}$$

Where *No.configs.Shear, Open* is no. of configuration with given Shear, within Shear+ΔShear, and Open, within Open+ΔOpen. We considered ΔShear=0.5Å, ΔOpen=6°, *k<sub>B</sub>* is Boltzmann constant, *T* is absolute temperature (300K in our calculations) and *F* is free energy.

### 4.3 Result and Discussion:

The RNABPDB database filters out 116 and 5140 base pairs for A:A w:wC and G:U W:WC, 36 dinucleotide steps for A:A w:wC :: U:G W:WC but only 4 dinucleotide steps of A:A w:wC::G:U W:WC. Two Adenine bases do not appear symmetrically face to face with their Watson-Crick edges to form hydrogen bonded base pair, as Adenine does with Thymine or Uracil (Figure 4-3).



**Figure 4-3** Representative Structures of A:A w:wC, G:U W:WC and A:U W:WC.

A significant amount of shearing motion is required for formation of two hydrogen bonds. Orientation of two bases of a base pair with respect to each other can be quantitatively analyzed by six parameters, as suggested by IUPAC-IUB<sup>39</sup> and analysis of the available structures of A:A w:wC base pair shows a distinct bimodal distribution of its Shear values with its large standard deviation (<http://hdrnas.saha.ac.in/rnabpdb/>) (Figure 4-2). The database also indicates similar large standard deviation and small mean of Shear from the structures obtained from NMR spectroscopy. In case of negative Shear the hydrogen of the N6-amino group of the Adenine of the first strand (or left strand when viewed from minor groove side) forms hydrogen bond with N1 of the second Adenine and C2-H of the second Adenine forms hydrogen bond with N1 of the first (i.e., N6-H(1)...N1(2) and C2-H(2)...N1(1) hydrogen bonds), where the Adenine of the second strand moves towards major groove (Figure 4-2). It may be noted that the wobble G:U Watson-Crick base pair (G:U W:WC) requires similar large Shear value ( $-2.25\text{\AA}\pm 0.36$  from RNABPDB) for the

formation of two hydrogen bonds between the bases where the Uracil base moves towards major groove. Such required Shear movement makes the G:U base pair non-isostreic to the U:G base pair<sup>56,168</sup>. Two hydrogen bonds are also possible with positive Shear for the A:A w:wC base pair where the Adenine of the second strand moves towards minor groove forming N6-H(2)...N1(1) and C2-H(1)...N1(2) hydrogen bonds. The base pair with positive Shear, if rotated by 180° through base pair short axis (or pseudo-dyad axis<sup>200</sup>), superposes on the base pair with negative Shear. Hence the two hydrogen bonding modes are symmetric (**Figure 4-2**). Our energy scan of the base pair with different Shear values also predicts this symmetry and bimodal distribution (**Figure 4-2**). Furthermore, Shear is a parameter that changes sign, like Tilt, Shift and Buckle, when calculated from the opposite strand<sup>39,111</sup>.

Although the two forms of A:A w:wC are energetically and otherwise symmetric, the symmetry breaks when the base pair is stacked on top of another base pair. We found the A:A w:wC base pairs stacked on C:G W:WC base pair, as A:A w:wC::C:G W:WC appears quite frequently in the non-redundant set of RNA structure. Important role of this base paired dinucleotide step has been discussed earlier by several groups<sup>175,201</sup>. The database further gives structures of two dinucleotide sequences, for A:A w:wC::G:U W:WC and A:A w:wC :: U:G W:WC with different characteristics. We noted that the A:A base pair adopts negative Shear in three out of four structures of A:A w:wC :: G:U W:WC dinucleotide step. The structure of the base pair from PDB ID: 5UNE with positive Shear is further found to be energetically unfavorable (**Table 4-4**). Stacking energy calculations however, do not indicate any instability of the odd structure (**Table 4-5**). There are 36 crystal structures for the A:A w:wC::U:G W:WC dinucleotide step in the RNABPDB database and most of them have large positive Shear (**Table 4-6**). The

bioinformatics data, however, is insufficient to arrive at specific conclusion about structural preference of the dinucleotide steps.

We have thus designed a sequence for GAG motif  $\begin{pmatrix} 5' - \text{CGAGAGAGCG} - 3' \\ 3' - \text{GCUUAUUCGC} - 5' \end{pmatrix}$  to understand preferred Shear of the A:A base pair, in between two G:U base pairs. Considering sign reversal of Shear when calculated from second strand, the A:A w:wC base pair also can have negative shear in the G:U W:WC::A:A w:wC base pair. We have carried out 500ns long MD simulations for the two models, one with positive Shear for the A:A w:wC base pair and another with negative Shear. The trajectories indicate stable RMSD for the structure with initial negative Shear for A:A w:wC base pair and significant variations of the RMSD for initial positive shear model. Both the structures appear to become similar with similar values of Roll, Slide, Twist, Open and Stretch (**Table 4-7**). Both the structures also show good stacking overlap (**Figure 4-4**). These parameters are indicative of regular RNA double helical structure formed by canonical base pairs. The Shear values of both the systems become large negative (around  $-2.5\text{\AA}$ ) with small fluctuations (**Figure 4-5**). Strength of the A:A w:wC base pair is not high as it is stabilized by one polar and a weak non-polar hydrogen bonds. Hence, breakage of the base pair is quite possible at 300K and is observed frequently, giving rise to larger values of Shear for short durations. Nevertheless, w:wC type base pairing persists in more than half of the snapshots as detected by BPFIND (**Figure 4-6**). This possibly indicate negative Shear is most likely for G:U W:WC :: A:A w:wC dinucleotide sequence.

**Table 4-4** Base pair orientation parameter, E-value and quantum chemical Interaction energy (in kcal/mol) of A:A w:wC base pairs from (A) X-ray Crystal structures\* (B) model structure for dinucleotide step sequence A:A w:wC :: G:U W:WC.

A.													
Base pair step information (Resolution in Å)	Base pair Sequence	Buckle (°)	Open (°)	Propeller (°)	Stagger (Å)	Shear (Å)	Stretch (Å)	E-Value	Method	Interaction Energy With BSSE (kcal/mol)	Interaction Energy Without BSSE (kcal/mol)	Interaction Energy of H-opt structures $\Delta E_{opt}$ (kcal/mol)	Interaction Energy of H-opt Structures with BSSE (kcal/mol)
3KFU 189-190.172-173 (3.00)	A:A w:wC 172_190	3.51	9.16	7.59	-1.52	-2.55	2.48	1.49	$\omega$ B97x-d	-1.55	-3.95	-6.74	-5.55
									M06-2X	-0.80	-3.17	-5.63	-4.25
	G:U W:WC 173_189	10.98	0.69	-16.17	-0.44	-2.09	2.94	0.11	$\omega$ B97x-d	-13.67	-17.32	-18.35	-16.04
									M06-2X	-12.61	-16.16	-17.08	-14.70
3KFU 260-261.243-244 (3.00)	A:A w:wC 243_261	36.52	9.28	7.39	-1.53	-2.56	2.48	1.52	$\omega$ B97x-d	-1.54	-3.93	-6.75	-5.51
									M06-2X	-0.79	-3.14	-5.61	-4.23
	G:U W:WC 244_260	11.51	0.30	-16.35	-0.44	-2.10	2.93	0.11	$\omega$ B97x-d	-13.62	-17.26	-18.67	-16.00
									M06-2X	-12.57	-16.11	-17.06	-14.70
5UNE 63-64.28-29 (2.90)	A:A w:wC 28_64	-0.99	-0.48	0.39	-0.07	1.97	2.58	0.60	$\omega$ B97x-d	1.75	-1.62	-3.26	-1.10
									M06-2X	2.27	-0.93	-2.68	-0.67
	G:U W:WC 29_63	-0.46	-4.63	-12.43	-0.07	-2.18	2.83	0.10	$\omega$ B97x-d	-16.13	-20.08	-19.98	-17.20
									M06-2X	-14.66	-18.50	-18.43	-15.69

Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT and MD Studies

5T5H 1029- 1030.1066- 1067 (2.54)	A:A w:wC 1066_1030		17.28	20.94	-1.80	0.12	-2.91	1.75	1.26	$\omega$ B97x-d	33.84	29.46	22.91	23.92
										M06-2X	34.58	30.38	23.61	24.43
	G:U W:WC 1067_1029		3.31	-2.67	-5.45	0.54	-2.36	2.49	0.59	$\omega$ B97x-d	-8.03	-12.62	-13.84	-11.14
										M06-2X	-6.93	-11.43	-12.82	-10.66
<b>B.</b>														
Base pair		Method		Functional/Basis set		Buckle(°)	Open(°)	Propeller(°)	Stagger(Å)	Shear(Å)	Stretch(Å)	$\Delta E_{opt}$ (Kcal mol <sup>-1</sup> )	$\Delta E_{opt}$ BSSE corrected (kcal mol <sup>-1</sup> )	
A:A w:wC	H optimized models in Stacking Interaction Calculation	Negative Shear (-2.7Å)	$\omega$ B97X-D/cc-pVDZ	-1.47	19.06	-13.12	-0.61	-2.70	2.42	-9.83	-6.66			
			M06-2X/cc-pVDZ	-1.47	19.06	-13.12	-0.61	-2.70	2.42	-8.55	-5.48			
	Free Optimized	Negative Shear (-2.7Å)	$\omega$ B97X-D/cc-pVDZ	7.56	19.04	-4.43	-0.06	-2.62	2.59	-10.60	-8.09			
			M06-2X/cc-pVDZ	7.56	19.04	-4.43	-0.06	-2.62	2.59	-9.18	-6.80			
			$\omega$ B97x-D/aug-cc-pVDZ	1.21	18.23	-0.90	-0.00	-2.63	2.61	-8.24	-7.87			
			M06-2X/aug-cc-pVDZ	2.24	18.86	-2.05	-0.01	-2.68	2.63	-7.09	-6.63			
G:U W:WC	H optimized models in Stacking Interaction Calculation	$\omega$ B97X-D/cc-pVDZ	-1.47	-0.29	-9.08	-0.10	-2.25	2.81	-20.80	-15.46				
		M06-2X/cc-pVDZ	-1.47	-0.29	-9.08	-0.10	-2.25	2.81	-19.39	-14.03				
	Free Optimized	$\omega$ B97X-D/cc-pVDZ	1.64	-0.24	-1.19	-0.07	-2.37	2.82	-20.15	-17.71				
		M06-2X/cc-pVDZ	1.64	-0.24	-1.19	-0.07	-2.37	2.82	-18.62	-16.21				
		$\omega$ B97x-D/aug-cc-pVDZ	0.52	-0.29	-1.00	-0.04	-2.37	2.82	-17.49	-17.99				
		M06-2X/aug-cc-pVDZ	0.51	-0.15	-1.33	-0.06	-2.37	2.84	-16.26	-16.61				

\*Geometry optimizations of A:A w:wC and G:U W:WC base pairs have been carried out in several ways: (i) Free unconstrained optimization (F-opt) of the base pairs using  $\omega$ B97X-D/cc-pVDZ ,  $\omega$ B97X-D/aug-cc-pVDZ, M06-2X/cc-pVDZ and M06-2X/aug-cc-pVDZ and (ii) also constraining the positions of all the non-hydrogen atoms (H-opt) using  $\omega$ B97X-D/cc-pVDZ and M06-2X/cc-pVDZ by Gaussian16. Counterpoise method of Boys and Bernardi has been imposed during the calculation of  $\Delta E_{opt}$  for modeled H-opt and free optimized base pairs and experimental base pair systems using same level of theory.

**Table 4-5** Geometric and energetic parameter of step A:A w:wC::G:U W:WC base paired dinucleotide step from available X-ray crystal structures.

PDB ID and residue numbers (Resolution in Å)	Tilt (°)	Roll (°)	Twist (°)	Shift (Å)	Slide (Å)	Rise (Å)	Overlap (Å <sup>2</sup> )	Energy Calculation Method	Stacking Energy (kcal/mol)	BSSE corrected Interaction Energy (kcal/mol)	BSSE Corrected Stacking Energy (kcal/mol)	C1'...C1' Distance (Å)	
												1 <sup>st</sup> Strand	2 <sup>nd</sup> Strand
3KFU_189_190_172_173 (3.00)	13.1	-1.22	32.63	0.14	-1.55	2.96	36.12	$\omega$ B97x-D	-16.36	-25.84	-10.62	5.75	4.89
								M06-2X	-12.19	-19.65	-6.24		
3KFU_260_261_243_244 (3.00)	12.99	-0.95	32.63	0.14	-1.55	2.97	36.12	$\omega$ B97X-D	-16.56	-25.93	-10.77	5.35	6.34
								M06-2X	-12.26	-19.69	-6.33		
5UNE_63_64_28_29 (2.90)	0.36	14.4	37.54	-1.15	-1.48	3.28	33.33	$\omega$ B97X-D	-19.57	-29.19	-14.81	5.58	5.19
								M06-2X	-14.75	-22.22	-9.83		
5T5H_1029_1030_1066_1067 (2.54)	6.91	6.71	29.04	0.63	-1.88	2.9	41.95	$\omega$ B97X-D	-19.90	11.98	-13.83	5.20	5.59
								M06-2X	-14.93	19.05	-8.6		

**Table 4-6** Base pair parameters for A:A w:wC found in A:A w:wC:: U:G W:WC dinucleotide step sequence.

PDB ID_Residue1 ID_Chain1 ID_Residue2 ID_Chain2 ID (Resolution in Å)	Buckle(°)	Open(°)	Propeller(°)	Stagger(Å)	Shear(Å)	Stretch(Å)
5IB7 17 13 897 13 (2.99)	-0.73	33.07	-23.76	-0.95	2.9	2.1
1N32 17 A 897 A (3.00)	-2.33	29.59	-20.58	-0.86	2.53	2.12
1FJG 17 A 897 A (3.00)	-5.71	28.95	-13.19	-0.71	3.11	2.11
2VQE 17 A 897 A (2.50)	-2.51	31.25	-19.37	-0.72	2.47	2.32
2UXC 17 A 897 A (2.90)	-3.48	28.21	-14.97	-0.63	2.67	2.26
2UUA 17 A 897 A (2.90)	-4.21	27.5	-15.77	-0.61	2.39	2.3
4B3T 16 A 896 A (3.00)	-4.57	28.84	-14.55	-0.52	2.56	2.26
3T1Y 12 A 892 A (2.80)	-1.97	27.12	-15.19	-0.55	2.58	2.22
4V9R 17 CA 897 CA (3.00)	-5.29	30.24	-14.3	-1.18	2.43	2.27
4V90 16 AA 896 AA (2.95)	1.69	32.35	-10.97	-0.52	2.82	2.44
4V8D 12 AA 892 AA (3.00)	-0.46	33.3	-13.3	-1.28	1.97	2.25
4V8D 12 CA 892 CA (3.00)	-3.83	30.1	-14.21	-0.89	2.6	2.35
4V8B 12 AA 892 AA (3.00)	3.78	34.49	-16.93	-0.86	2.77	2.28
4V8B 12 CA 892 CA (3.00)	-0.79	25.85	-9.67	-0.66	2.5	2.06
4V88 11 A2 1143 A2 (3.00)	-10.28	26.56	-8.44	-0.63	2.87	2.34
4V88 992 A1 1057 A1 (3.00)	6.48	14.1	-7.23	-0.23	-2.01	2.57
4V88 11 A6 1143 A6 (3.00)	-2.77	31.13	-20.07	-0.7	2.79	2.05
4V88 992 A5 1057 A5 (3.00)	4.66	22.32	-15.31	-0.46	-1.32	2.86
4U4U 11 2 1143 2 (3.00)	-7.67	27.11	-13.73	-0.84	2.49	2.53
4U4U 992 1 1057 1 (3.00)	7.89	5.66	-9.29	-0.28	-2.2	2.56
4U4U 11 6 1143 6 (3.00)	-5.35	27.54	-20.47	-0.63	2.64	2.22
4U4U 992 5 1057 5 (3.00)	8.79	13.19	-12.33	-0.55	-1.54	2.87
4U26 15 AA 918 AA (2.80)	-0.41	29.97	-10.34	-1.33	2.54	2.08
4U26 15 CA 918 CA (2.80)	-8.97	26.55	-5.33	-0.17	2.44	2.4
5T2A 11 2 1497 2 (2.90)	-8.16	35.29	-14.05	-1.05	3.06	2.52
5J7L 16 AA 919 AA (3.00)	-0.76	35.44	-15.75	-0.81	2.51	2.42
5J7L 16 BA 919 BA (3.00)	-0.51	32.92	-12.95	-0.73	2.37	2.45
5J5B 16 AA 919 AA (2.80)	-1.9	33.13	-17.92	-0.79	2.19	2.44
5J5B 16 BA 919 BA (2.80)	-2.19	29.83	-19.57	-0.68	2.28	2.45
5IBB 17 13 897 13 (2.96)	3.17	29.78	-17.89	-1.3	2.58	2.14

5IBB 17 1G 897 1G (2.96)	-4.14	23.98	-14.15	-0.75	1.96	2.45
5IB7 17 1G 897 1G (2.99)	-5.87	29.32	-9.72	-0.64	2.16	2.46
5FDU 16 1A 896 1A (2.90)	-3.15	22.55	-16.92	-0.73	2.68	2.16
5FDU 16 2A 896 2A (2.90)	-1.73	27.43	-13.02	-0.59	3.11	2.19
5E81 17 13 897 13 (2.95)	-0.26	29.86	-19.5	-1.19	2.87	1.96
5E81 17 1G 897 1G (2.95)	1.79	28.28	-18.09	-0.82	1.51	2.45

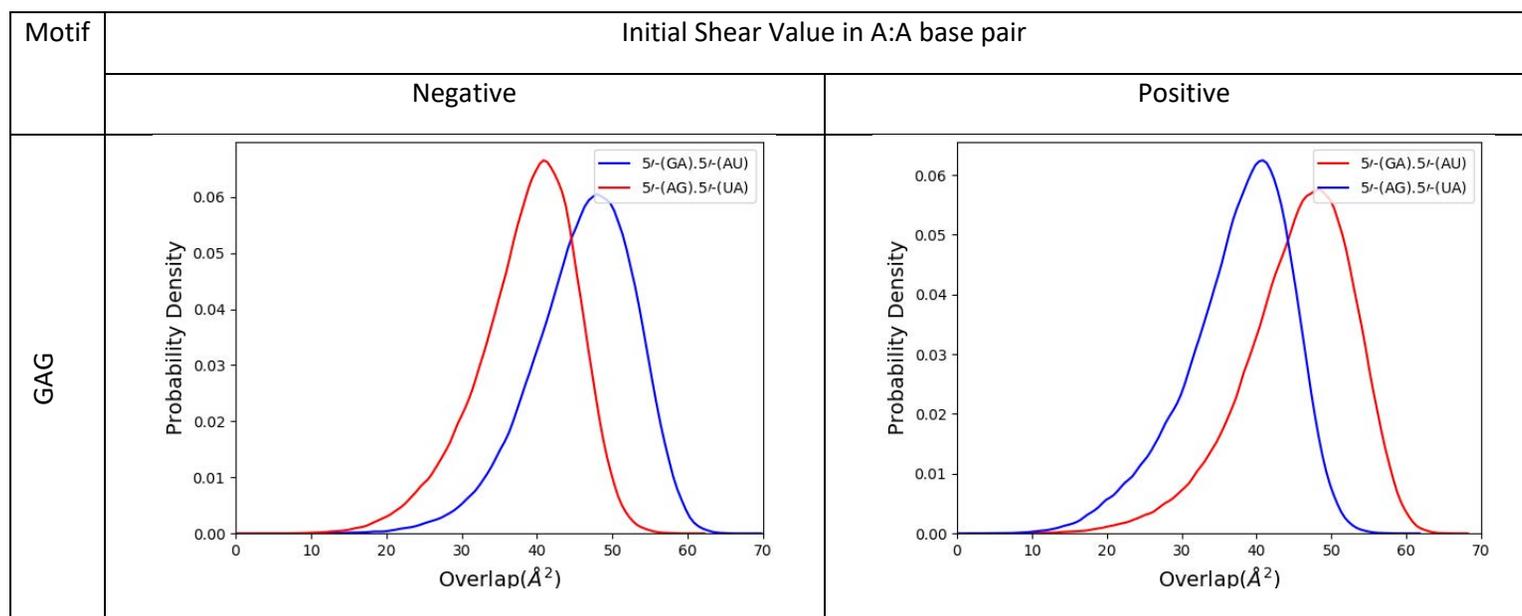
**Table 4-7** Mean and standard deviation (in Parenthesis) of some important base pair and base pair step parameter for mentioned RNA double helix.

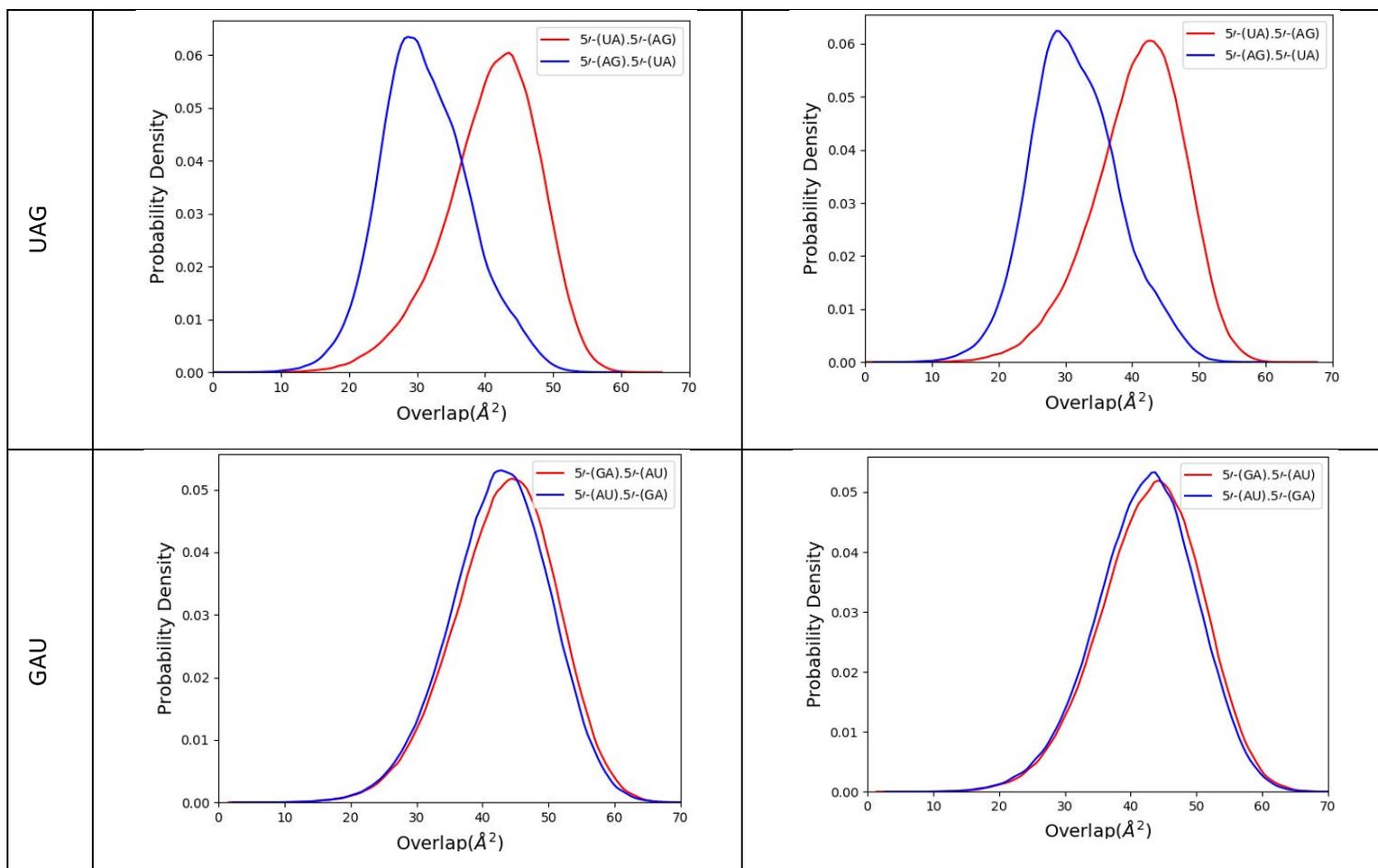
System	Base pair Step	Base pair Step Parameter			Base pair	Base pair parameter			
		Twist (°)	Roll (°)	Slide (Å)		Open (°)	Stretch (Å)	Shear (Å)	
5' - CGAGAGAGCG -3' 3' - GCUUAUUCGC -5'	Negative Shear	5'-(AG).5'-(UU)	37.67 (4.10)	13.65 (6.68)	-1.55 (0.57)	A:U W:WC	2.93 (6.68)	2.87 (0.12)	0.06 (0.61)
		5'-(GA).5'-(AU)	27.91 (6.12)	1.89 (5.60)	-1.25 (0.56)	G:U W:WC	3.30 (18.15)	3.09 (0.45)	-2.38 (0.77)
		5'-(AG).5'-(UA)	28.10 (4.46)	9.08 (6.74)	-0.83 (0.76)	A:A w:wC	9.60 (23.26)	2.78 (0.41)	-1.58 (1.96)
		5'-(GA).5'-(UU)	26.89 (5.91)	9.89 (5.92)	-1.47 (0.43)	G:U W:WC	5.31 (16.62)	3.08 (0.42)	-2.26 (0.47)
		5'-(AG).5'-(CU)	29.08 (5.04)	11.53 (6.10)	-1.75 (0.43)	A:U W:WC	4.36 (7.19)	2.84 (0.16)	-0.95 (1.94)
	Positive Shear	5'-(AG).5'-(UU)	37.17 (5.42)	13.84 (6.81)	-1.53 (0.57)	A:U W:WC	3.28 (0.34)	2.84 (0.25)	-0.48 (0.33)
		5'-(GA).5'-(AU)	27.03 (7.36)	2.40 (7.08)	-1.17 (0.64)	G:U W:WC	3.77 (1.64)	3.05 (0.66)	-2.04 (1.64)
		5'-(AG).5'-(UA)	27.79 (8.74)	9.57 (8.59)	-0.89 (0.83)	A:A w:wC	11.10 (2.92)	2.62 (0.87)	-0.97 (2.92)
		5'-(GA).5'-(UU)	25.71 (4.94)	9.64 (5.85)	-1.41 (0.43)	G:U W:WC	3.54 (0.51)	3.04 (0.38)	-2.29 (0.51)
		5'-(AG).5'-(CU)	30.10 (4.31)	11.56 (6.20)	-1.78 (0.43)	A:U W:WC	3.94 (1.48)	2.85 (0.16)	-0.48 (1.47)

5' - CGAUAAGCG -3' 3' - GCUGAUCGC -5' System:4 <sup>th</sup> and 5 <sup>th</sup>	Negative Shear	5'-(AU).5'-(GU)	23.26 (5.03)	6.13 (5.04)	-1.91 (0.57)	A:U W:WC	12.19 (17.27)	2.86 (0.26)	0.20 (0.38)
		5'-(UA).5'-(AG)	20.98 (3.77)	6.74 (5.30)	-1.36 (0.62)	U:G W:WC	20.83 (24.29)	3.19 (0.53)	2.53 (0.43)
		5'-(AG).5'-(UA)	55.62 (8.71)	8.20 (9.13)	-2.74 (0.84)	*A:A s:hT	15.78 (23.44)	-2.20 (1.14)	6.28 (0.88)
							-16.68 (23.67)	*3.08 (0.82)	*3.03 (0.92)
		5'-(GA).5'-(UU)	17.90 (3.61)	6.35 (5.38)	-1.47 (0.37)	G:U W:WC	5.87 (18.01)	2.86 (0.39)	2.63 (0.41)
	5'-(AG).5'-(CU)	28.91 (3.28)	6.56 (6.33)	-1.79 (0.37)	A:U W:WC	-1.52 (6.42)	2.95 (0.17)	0.19 (0.30)	
	Positive Shear	5'-(AU).5'-(GU)	23.48 (4.95)	6.04 (4.94)	-1.91 (0.55)	A:U W:WC	12.17 (17.44)	2.86 (0.26)	0.20 (0.39)
		5'-(UA).5'-(AG)	20.78 (3.45)	6.79 (5.22)	-1.32 (0.60)	U:G W:WC	19.76 (23.89)	3.17 (0.50)	2.55 (0.43)
		5'-(AG).5'-(UA)	56.64 (7.62)	8.58 (9.09)	-2.70 (0.79)	A:A w:wC	12.88 (20.48)	-2.37 (0.91)	6.38 (0.42)
						*A:A s:hT	-13.74(20.64)	*3.12 (0.39)	*2.54 (0.74)
5'-(GA).5'-(UU)		17.94 (3.57)	6.24 (5.33)	-1.46 (0.36)	G:U W:WC	3.85 (16.08)	2.81 (0.35)	2.63 (0.40)	
5'-(AG).5'-(CU)	28.90 (3.28)	6.34 (6.35)	-1.78 (0.37)	A:U W:WC	-1.74 (6.59)	2.96 (0.18)	0.19 (0.29)		
5' - CGAUAAGCG -3' 3' - GCUGAUCGC -5' System:6 <sup>th</sup> and 7 <sup>th</sup>	Negative Shear	5'-(AG).5'-(UU)	51.10 (6.80)	12.15 (7.36)	-3.80 (0.67)	A:U W:WC	-71.90 (7.60)	4.01 (0.35)	1.64 (0.80)
		5'-(GA).5'-(AU)	2.35 (5.53)	5.98 (5.25)	-1.14 (0.95)	G:U W:WC	99.49 (16.59)	3.31 (0.98)	4.77 (0.75)
		5'-(AU).5'-(GA)	2.46 (5.42)	7.35 (5.41)	-1.81 (1.00)	A:A w:wC	-94.75 (19.68)	5.73 (0.95)	-0.76 (2.54)
		5'-(UA).5'-(UG)	43.18 (7.61)	12.50 (7.32)	-4.15 (0.91)	U:G W:WC	-86.52 (23.87)	2.72 (1.03)	-4.86 (0.58)
		5'-(AG).5'-(CU)	50.47 (3.60)	6.94 (6.40)	-2.47 (0.43)	A:U W:WC	-71.19 (9.70)	4.09 (0.36)	1.53 (0.34)

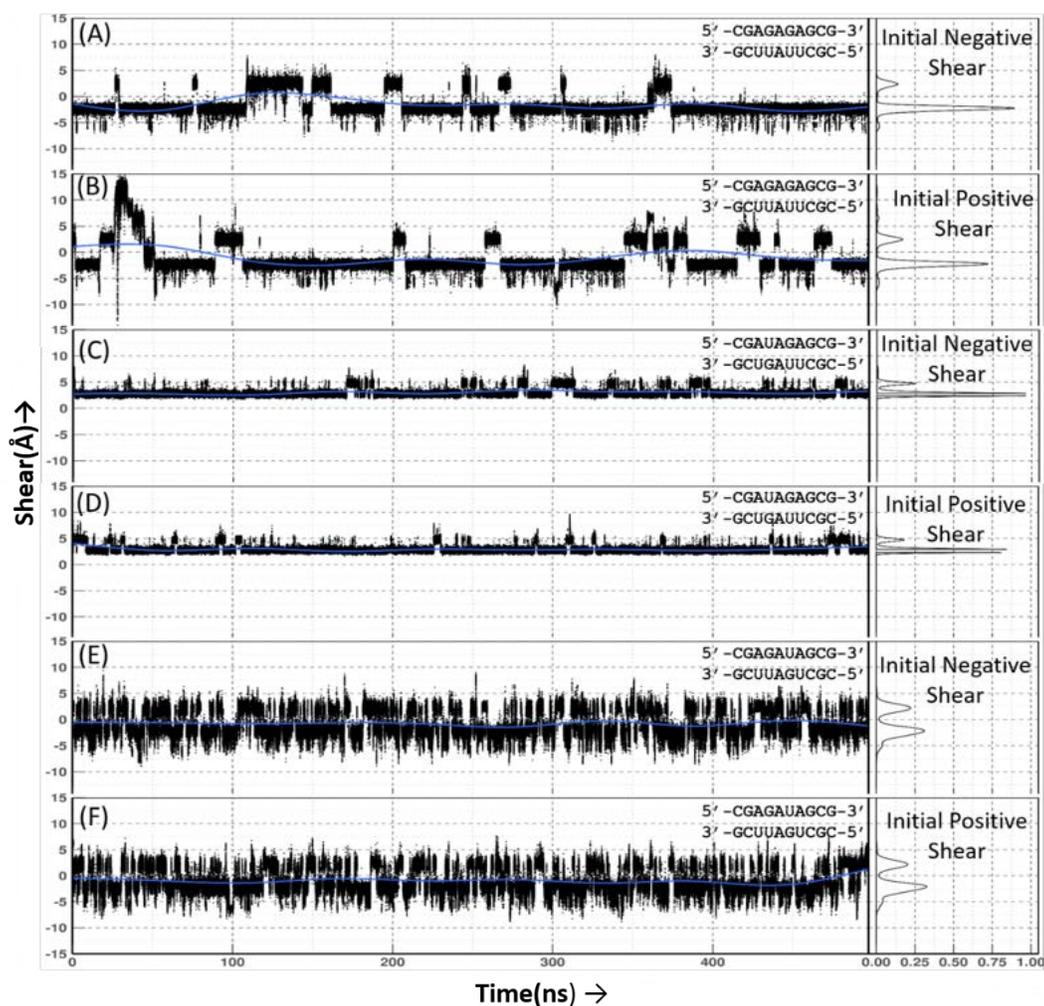
Positive Shear	5'-(AG).5'-(UU)	51.38 (6.71)	13.25 (7.27)	-3.81 (0.70)	A:U W:WC	-72.14 (8.22)	4.05 (0.27)	1.57 (0.32)
	5'-(GA).5'-(AU)	2.03 (5.41)	6.05 (5.19)	-1.07 (0.97)	G:U W:WC	-99.95 (17.71)	3.20 (1.02)	4.78 (0.67)
	5'-(AU).5'-(GA)	2.92 (5.91)	7.50 (5.41)	-1.84 (0.98)	A:A w:wC	-96.30 (20.24)	5.84 (1.06)	-0.93 (2.54)
	5'-(UA).5'-(UG)	42.78 (7.47)	12.97 (7.33)	-4.11 (0.87)	U:G W:WC	-88.09 (23.59)	2.84 (0.97)	-4.86 (0.58)
	5'-(AG).5'-(CU)	50.24 (3.97)	6.50 (6.47)	-2.48 (0.43)	A:U W:WC	-70.91 (9.73)	4.03 (0.38)	-1.52 (0.32)

\*It may be noted that the base pairing edge specific axis system adopted in NUPARM calculates Stretch value of any good and stable base pair around  $2.8\text{\AA}$ , which is calculated as  $0\text{\AA}$  by the other programs, namely X3DNA or CURVES. Deviation of Stretch from  $2.8\text{\AA}$  in either side indicates opening of the base pair as it happened for both the  $5'(\text{GAU}).5'(\text{GAU})$  and  $5'(\text{UAG}).5'(\text{UAG})$  sequence motifs.

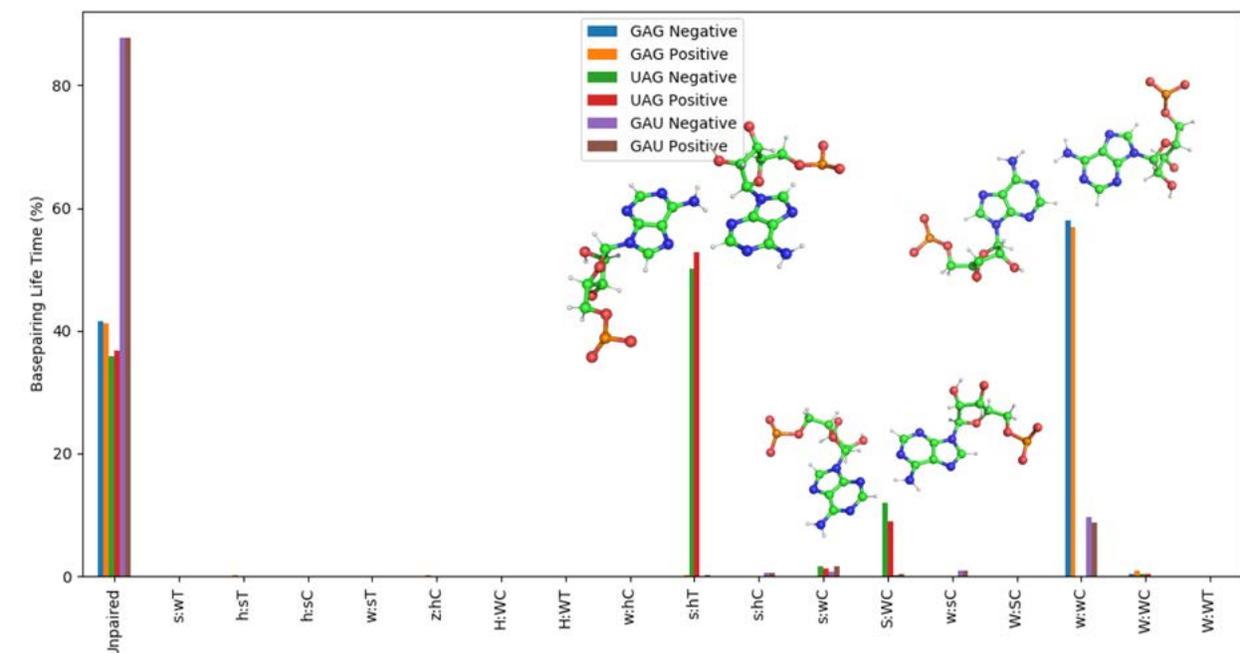




**Figure 4-4** Distribution of stacking overlap in different system as mentioned in the figure.



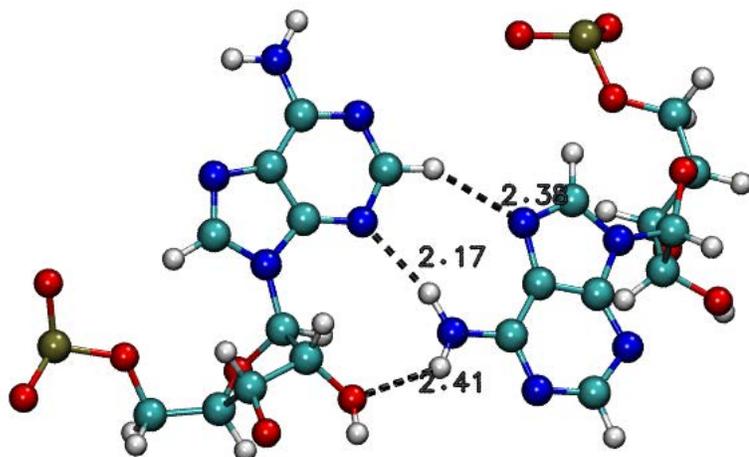
**Figure 4-5** Time evolutions of Shear for A:A base pair in RNA double helices for different sequences: GAG motif for (A)  $-2.7 \text{ \AA}$  and (B)  $+2.7 \text{ \AA}$  Shear in initial models of the A:A w:wC base pair, UAG motif for (C)  $-2.7 \text{ \AA}$  and (D)  $+2.7 \text{ \AA}$  Shear in initial models of A:A w:wC (calculated by NUPARM considering s:hT base pairing pattern), GAU motif for (E)  $-2.7 \text{ \AA}$  and (F)  $+2.7 \text{ \AA}$  Shear in initial models of A:A w:wC.



**Figure 4-6** base pairing lifetime histogram for A:A base pair in the RNA double helices for the three sequences namely GAG motif, UAG motif and GAU motif with initial  $-2.7\text{\AA}$  and  $+2.7\text{\AA}$  Shear in the initial structure structures of the A:A w:wC base pairs. Representative figures of A:A s:hT, A:A S:WC and A:A w:wC are shown.

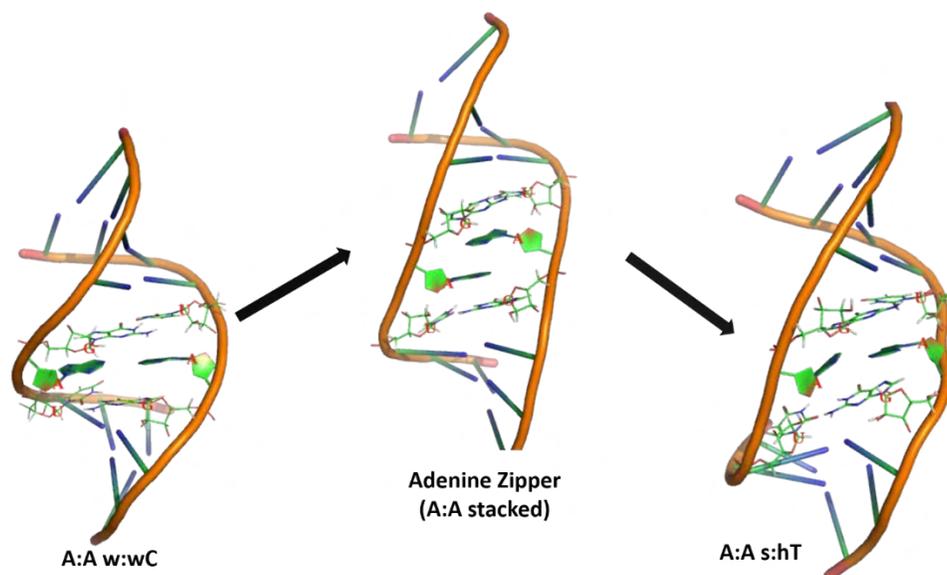
An A:A w:wC base pair can be stacked by G:U (or U:G) W:WC base pairs in two other sequence motifs, namely UAG and GAU, and we have modeled structures of these two motifs with either positive or negative Shear for the A:A w:wC base pair. The UAG motif effectively contains two consecutive A:A w:wC :: G:U W:WC dinucleotides but one in reverse way as U:G W:WC::A:A w:wC with sign reversal of Tilt, Shift, Buckle and Shear. As indicated earlier, the geometry and energetic symmetry of two configurations of A:A w:wC base pair breaks in presence of asymmetric G:U W:WC sheared base pair. We have observed major structural alteration of the A:A w:wC base pair during MD simulation. The Shear values of the A:A base pair changes to around  $5\text{\AA}$  from  $-2.7\text{\AA}$  or  $+2.7\text{\AA}$ , indicative of broken base pair. We have nevertheless tried to identify base pairing pattern, if any, using BPFIND and found that the A:A base pair breaks at very early stage of the simulation but it persists as s:hT (tSH in Leontis-Westhof nomenclature<sup>170</sup>) type

(Figure 4-7), for around 50% simulation time (Figure 4-6). During this transition, the adenine residues attains an intermediate adenine zipper like geometry (Figure 4-8) having very short life time (~74ps). We have calculated buried surface area between the Adenines (equivalent to overlap)



**Figure 4-7 A:** A base pair attains *s:hT* type base pair orientation in 5'-(UAG).5'-(UAG) motif during MD simulation of the sequence 5' -CGAUAGAGCG - 3',  
3' -GCUGAUUCGC - 5'

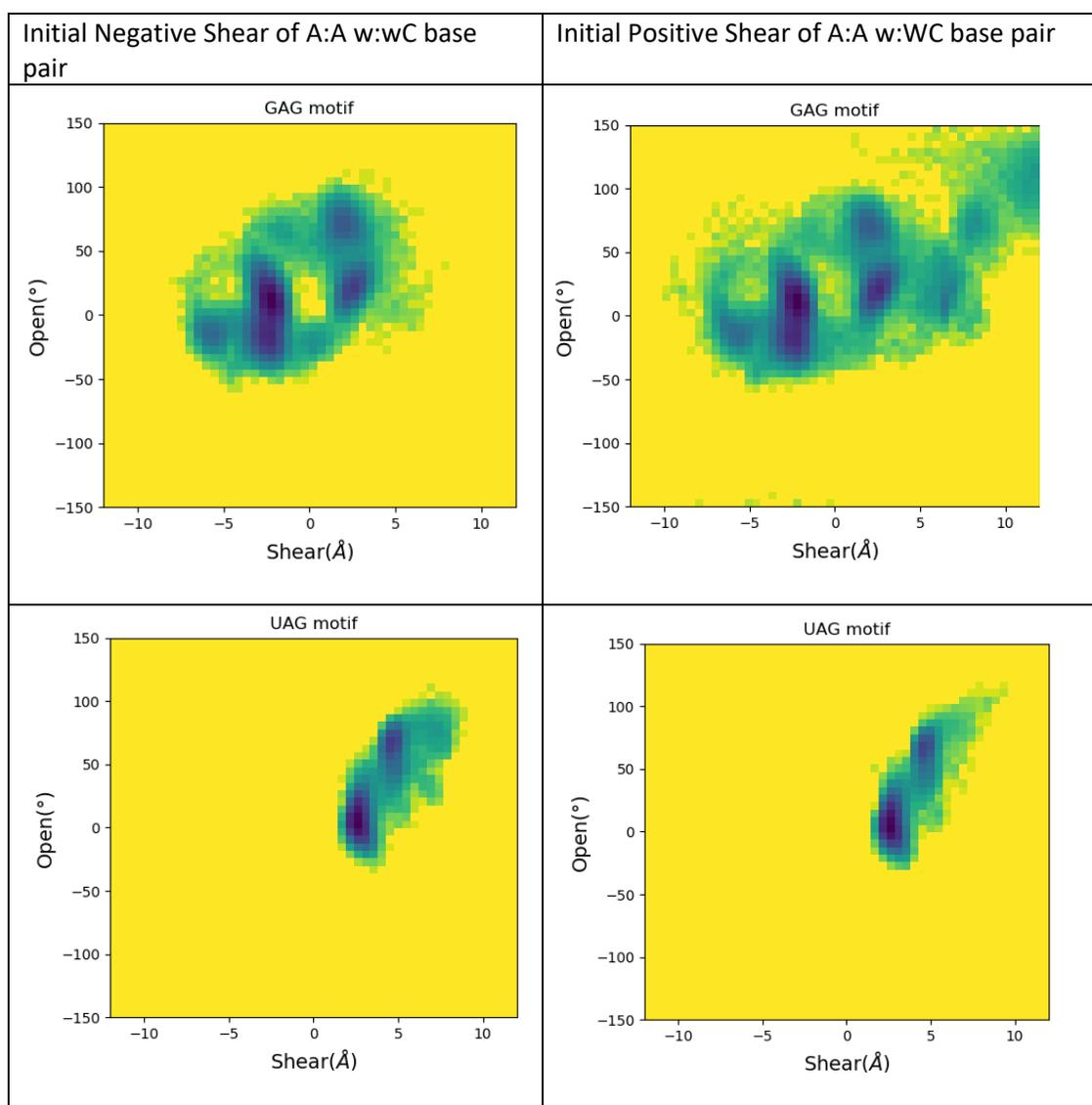
throughout the simulation time following an algorithm similar to that by Pingali et al<sup>116</sup>. The results show stacking overlap of around  $22\text{\AA}^2$  during the transition as Adenine zipper in the very beginning of MD simulation. Since then the A:A base pair maintained the *s:hT* geometry till 500ns, such transition from Watson-Crick to Hoogsteen base pairing in DNA was recently shown by combination of simulation and experiment<sup>202</sup>. It is well known that intra base pair parameters of a non-canonical base pair, particularly those involving non-Watson-Crick edges, show unusual values by CURVES<sup>203</sup> or 3DNA<sup>109</sup>. NUPARM allows one to calculate base pair parameters considering base pairing edge specific axis system<sup>111</sup> which gives small values of the most of the parameters for a base pair with proper hydrogen bond irrespective of whether it is canonical or non-canonical. We have recalculated Shear values for the A:A base pair considering *s:hT* type pairing between them.

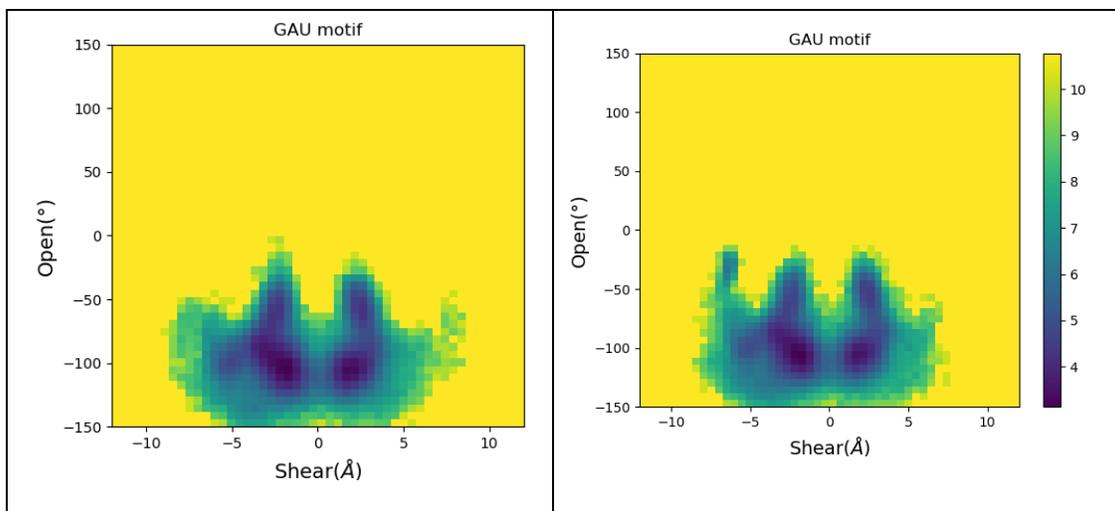


**Figure 4-8** Conversion of A:A w:wC to A:A s:hT through Adenine Zipper (A:A stacked geometry) in 5'-(UAG).5'-(UAG) motif during MD simulation of the sequence 5'-CGAUAGAGCG-3'. The A:A base pair is shown as Cartoon ring mode and G:U base pair at upstream and downstream of A:A are shown as ball and stick and remaining base pairs are shown by ladder.

Base pair parameter analysis, considering s:hT type, gives Shear value  $\sim 2.0\text{\AA}$  throughout the simulation with Stretch around  $\sim 3.0\text{\AA}$  (Std. 0.82) (Figure 4-5C and Figure 4-5D, Table 4-7). The values indicate stable structure of the A:A base pair, as found previously by A:A s:hT base pairs in symmetric internal loops<sup>178,179,204</sup>. This altered geometry of the base pair appears to be quite stable having  $-9.95\text{kcal/mol}$  BSSE corrected interaction energy between the two Adenine bases as compared to  $-6.30\text{kcal/mol}$  for w:wC geometry. Mean Twist value around  $55^\circ$  for the 5'-(AG).5'-(UA) base pair step for both Shear value also indicates s:hT base pairing of A:A (Table 4-7). We have attempted to estimate free energy of transition using Gibbs' formula from the probabilities of Shear. We found that Shear is often correlated to Open angle, especially in GAG and UAG systems with correlation coefficients of 0.81 and 0.62, respectively. Hence we have analyzed free energy landscape from two dimensional plot and found s:hT base pairs type is more favorable in the UAG system. We found  $F_{min}$  for UAG(positive)= $1.58\text{ kcal/mol}$ , for UAG(negative)= $1.63\text{ kcal/mol}$ ;  $F_{min}$

for  $GAG(\text{positive})=2.40$  kcal/mol, for  $GAG(\text{negative})=2.29$  kcal/mol and  $F_{min}$  for  $GAU(\text{positive})=3.13$  kcal/mol, for  $GAU(\text{negative})=3.24$  kcal/mol, where “positive” or “negative” indicate initial values of Shear in those systems. The conformational fluctuations also take place within narrow domain in Open-Shear space indicating restricted motion of A:A base pair in s:hT type in the UAG motif (**Figure 4-9**). This preference is possibly due to additional hydrogen bond formation involving 2'-OH group of the Adenine with the Adenine pairing using its Hoogsteen edge (**Figure 4-7**).

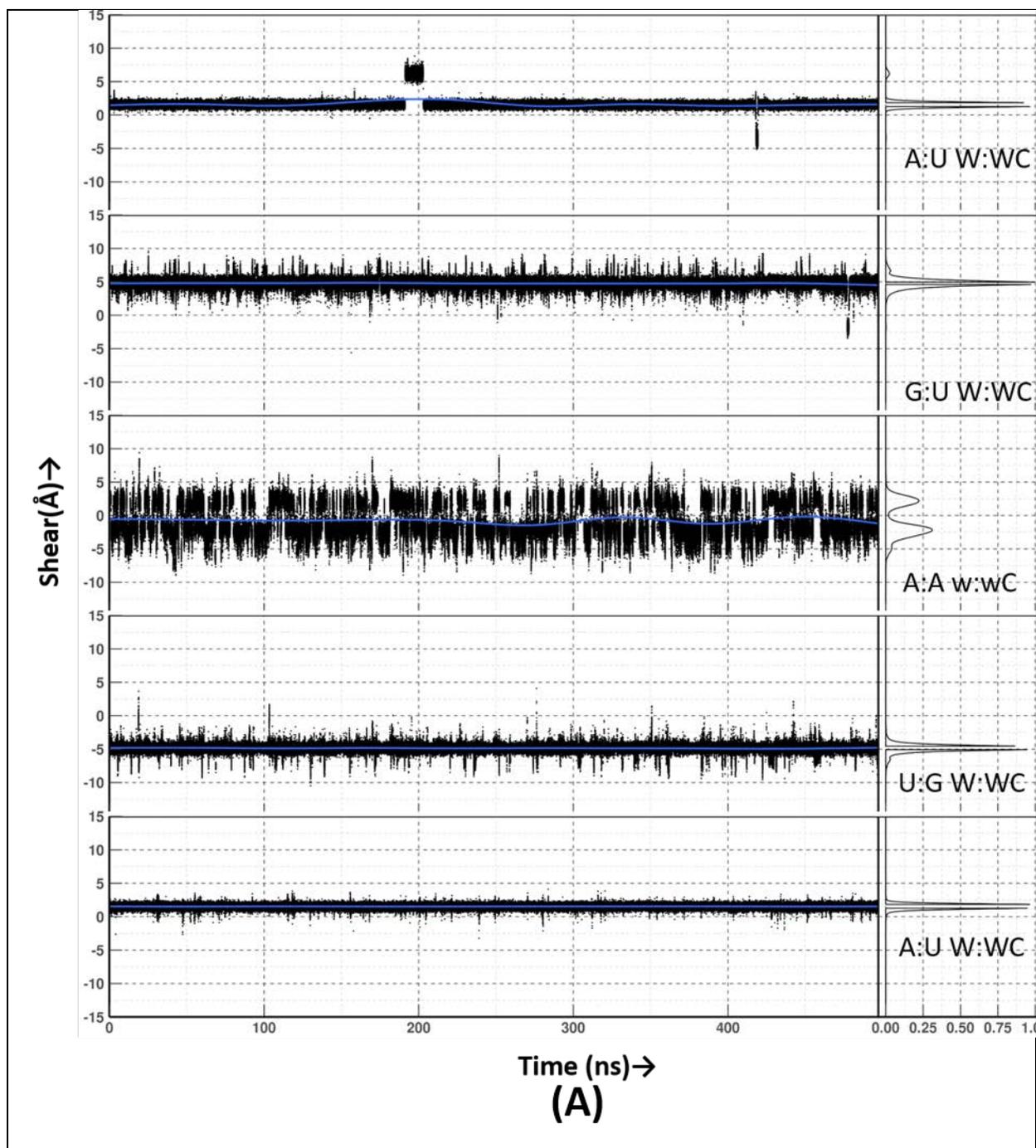


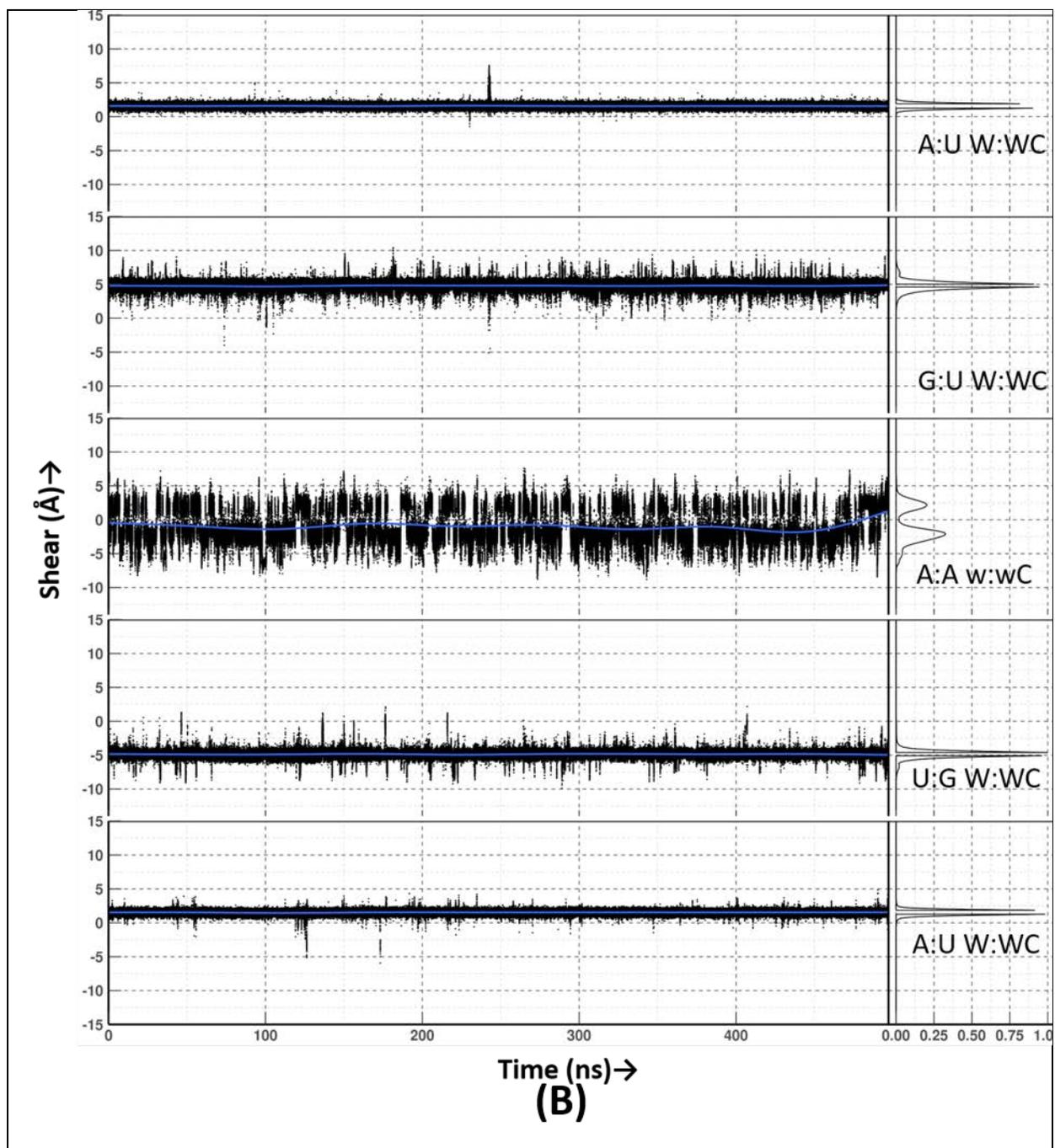


**Figure 4-9** Free Energy landscape from MD simulations in Open-Shear hyperspace for the three sequence motifs with either Shear value in their initial structures. The color bar on the Figure for GAU motif with positive Shear in initial model, indicate Free Energy values (in kcal/mol) for all the systems.

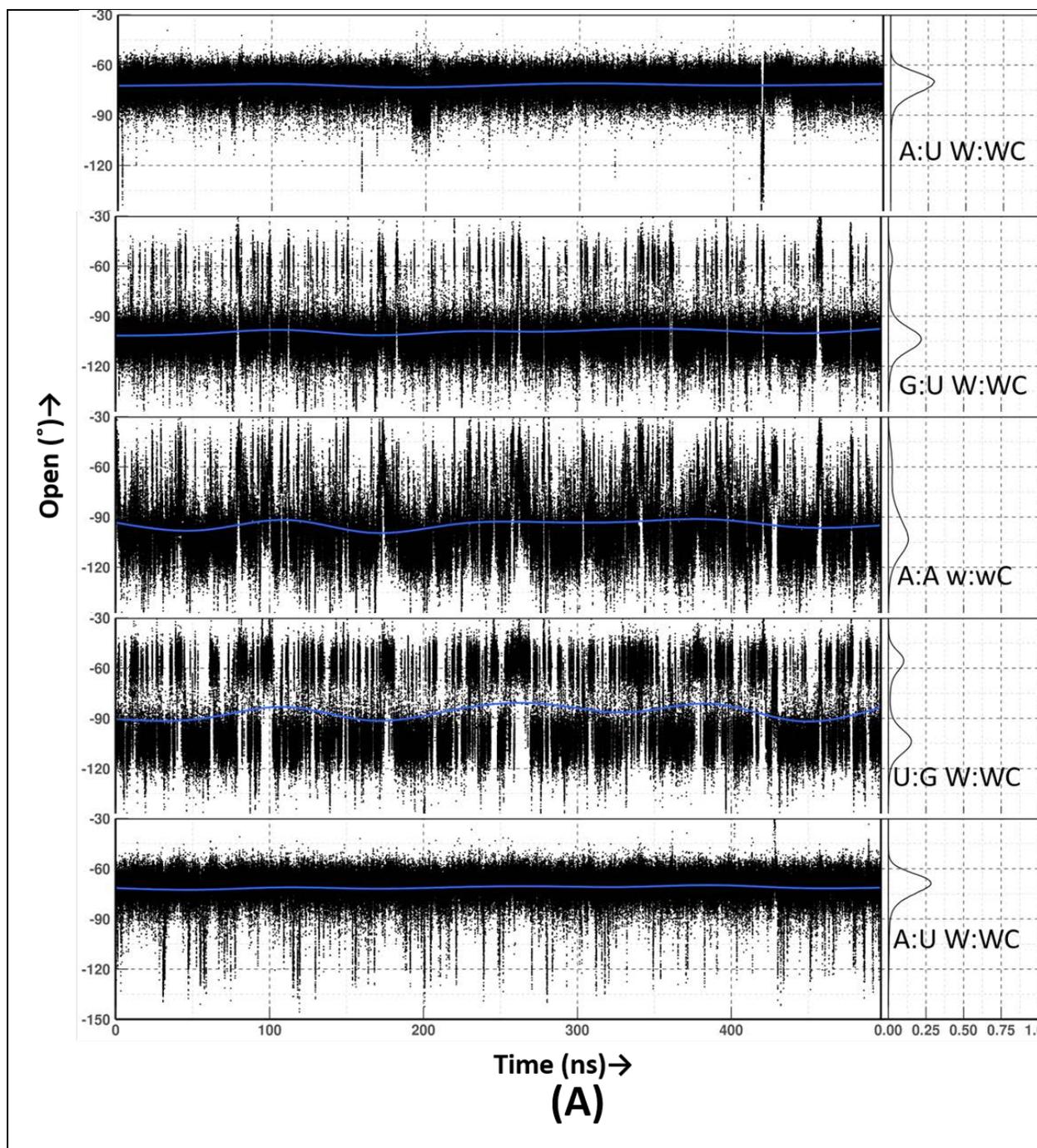
Similarly two models of GAU motif, having initial positive and negative Shear of A:A w:wC base pair, consist of two consecutive A:A w:wC::U:G W:WC stacks but in reverse way. The sandwiched A:A base pair is found to be oscillating between two extremes Shear values during MD simulation (**Figure 4-5E** and **Figure 4-5F**). This oscillation of A:A base pair strongly affects the nearby base pairs of the double helix by inducing large Shear to the stacked G:U base pairs (around  $\pm 5 \text{ \AA}$ ), and also to the A:U base pairs (around  $\pm 2 \text{ \AA}$ ) (**Figure 4-10**). The Open angles of these neighboring base pairs became very large (around  $90^\circ$ ) and the Twist values are found to be either around  $0$  or  $50^\circ$  for even the Watson-Crick base paired steps (**Figure 4-11** and **Figure 4-12**). These structures nevertheless possess good stacking overlap, indicating the Adenine residues do not come out of stacks, similar to the UAG motif structures (**Figure 4-4**). The base pairing analysis from trajectory by BPFIND reveals absence of any particular base pair type during the simulation time with occasional appearance of various other types of patterns for very short durations (**Figure 4-6**). Free energy analysis indicates the negative Shear region is marginally more stable (**Figure 4-9**). It should be remembered that the negative Shear regions also do not have proper hydrogen bonds.

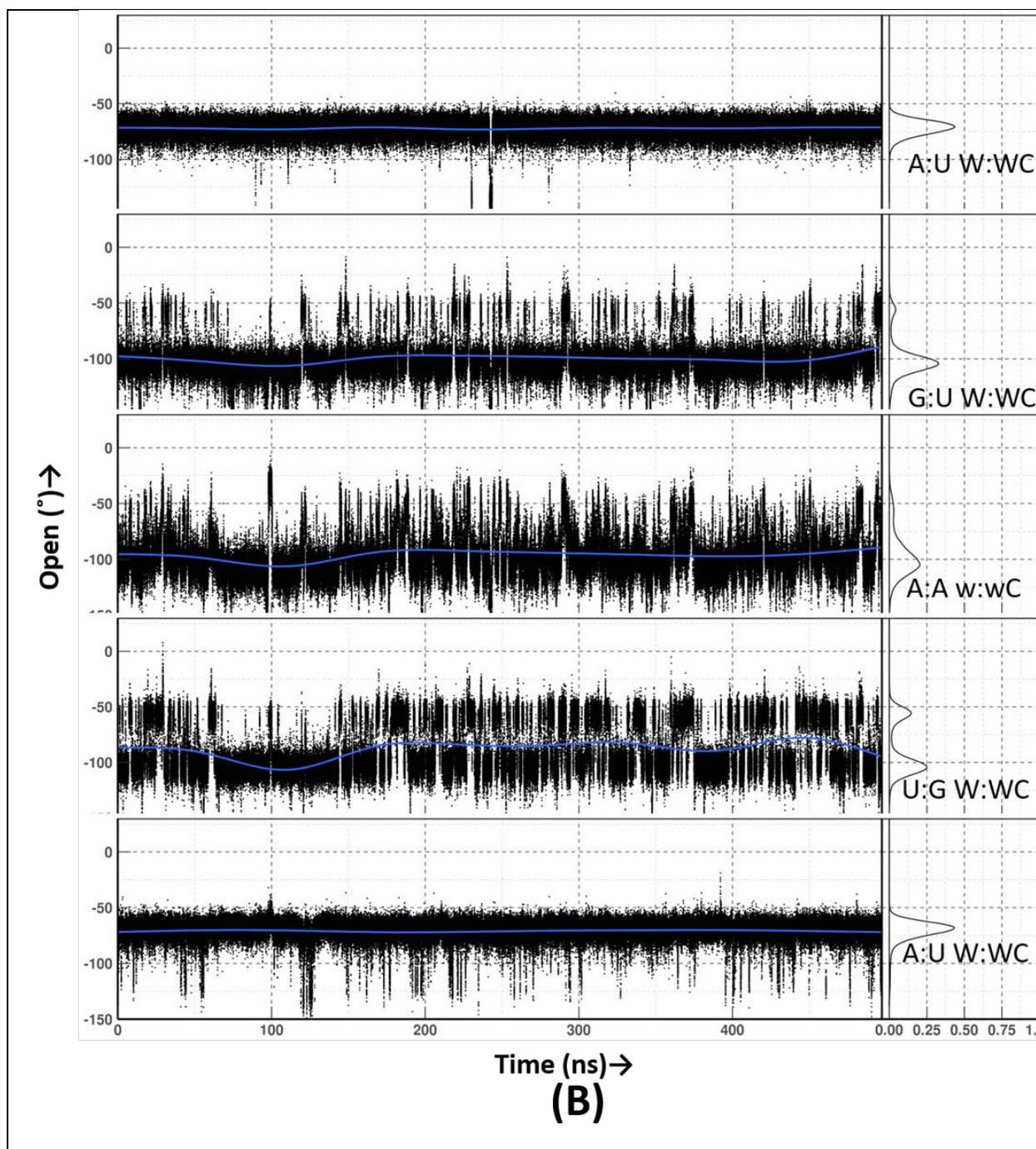
The above results from MD simulation further opened the question: Why the Adenine bases remain paired in w:wC fashion in some situation but why they undergo major structural transitions in other motifs. Recent studies indicated that molecular mechanics force-fields have a tendency to over predict stacked geometry for nucleic acid bases or base pairs which might have forced to maintain structures for an intrinsically disordered sequence<sup>205,206</sup>. Moreover the force-fields were never tuned to reproduce structure and energetics of non-canonical base pairs. Furthermore, one can never confirm that a long MD simulation sampled the full phase space at physiological temperature. Hence we have adopted *ab initio* quantum chemical calculations to understand stacking interactions between A:A w:wC base pair and G:U W:WC base pair using various methods. Our adopted methods are also seen to be compatible with the recent benchmark studies for Watson-Crick base paired dinucleotide steps<sup>180</sup>.



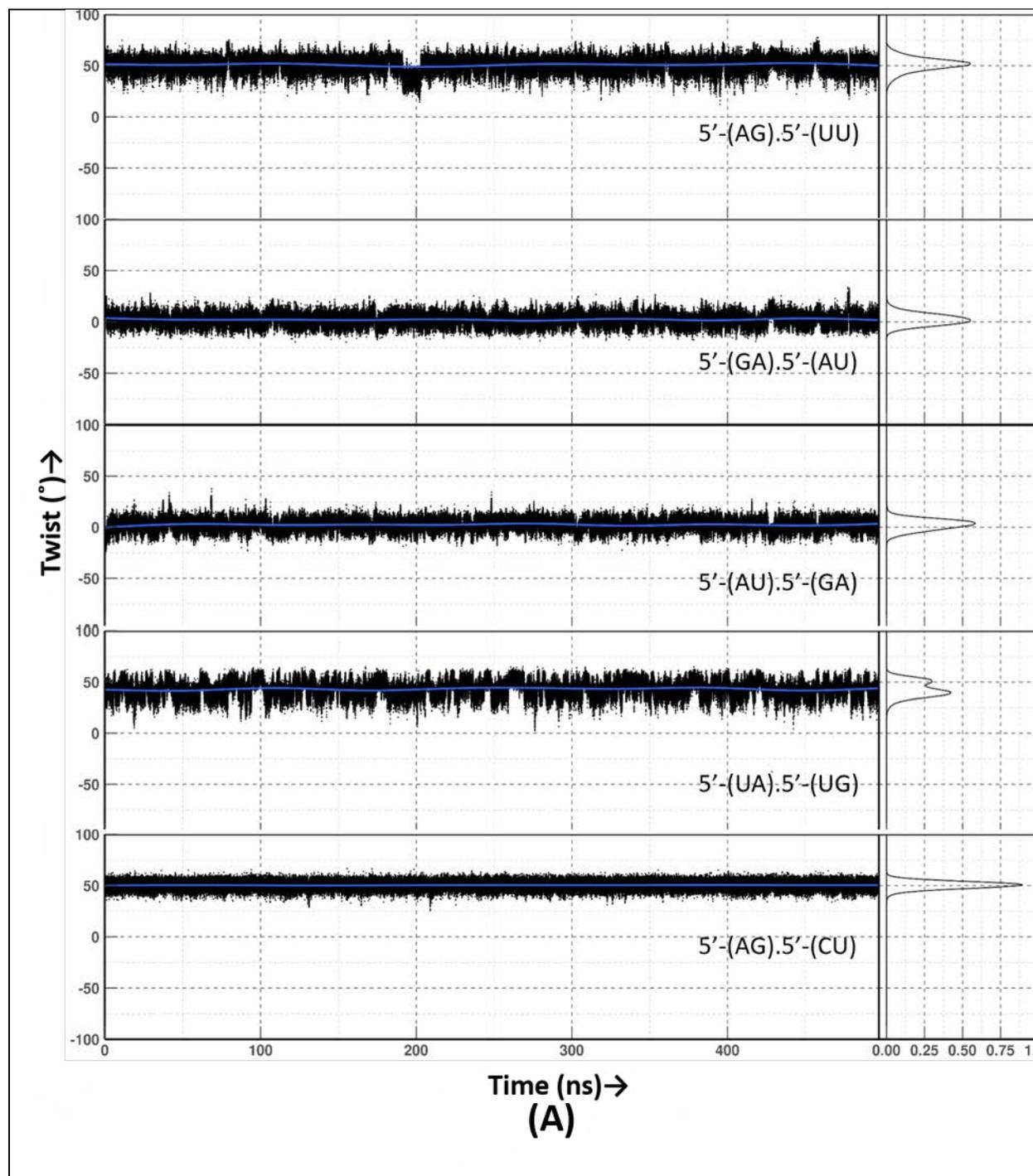


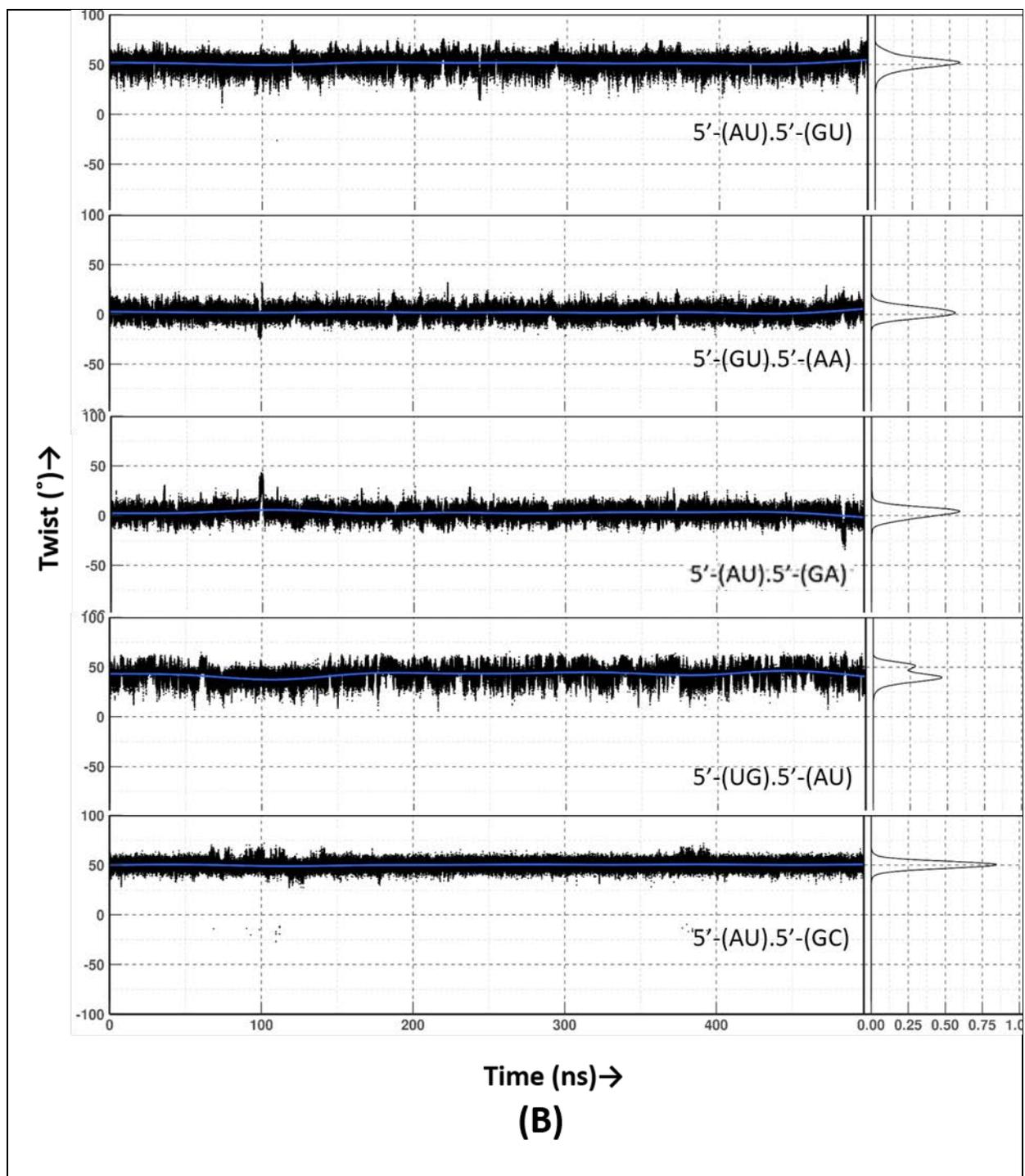
**Figure 4-10** Time evolution of Shear for several base pair in GAU motif sequence during 500ns MD simulations for (A) negative Shear in initial model of the A:A w:wC base pair and (B) positive Shear in initial model of the same.





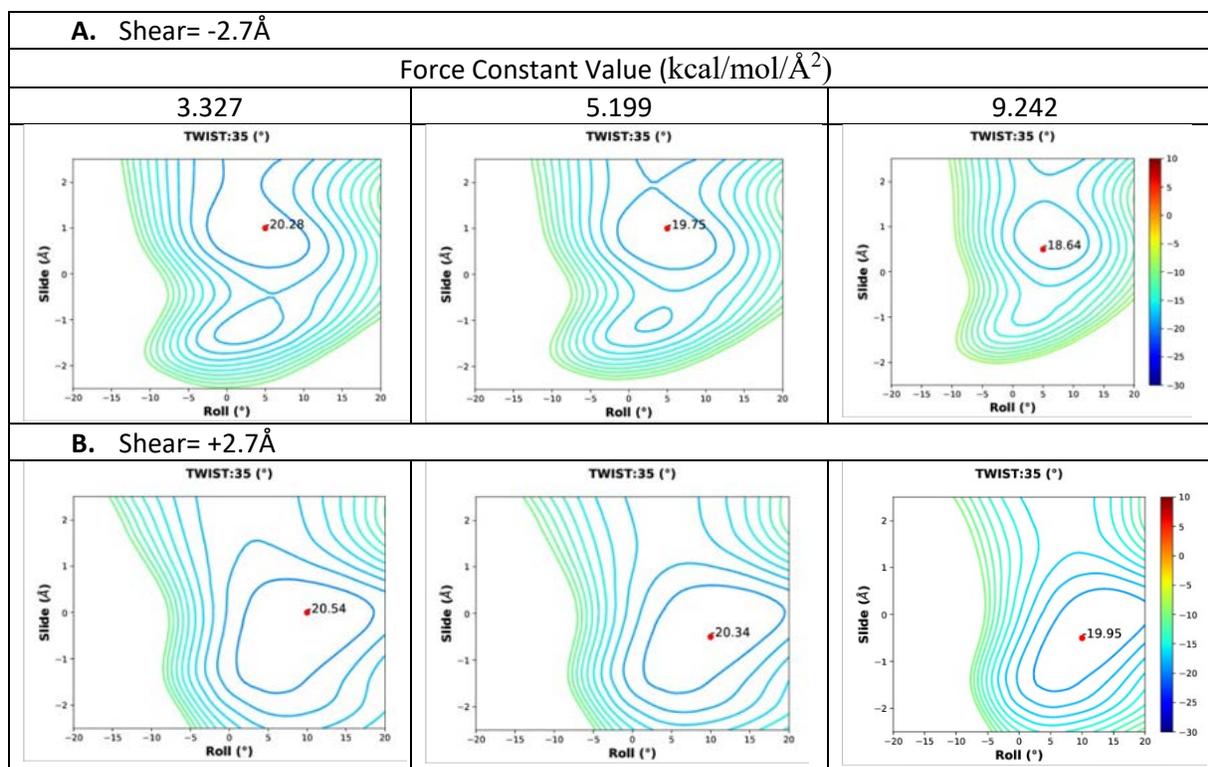
**Figure 4-11** Time evolution of Open angle for several base pair in the GAU motif sequence during 500ns MD simulations with (A) negative Shear in initial model of A:A w:wC base pair and (B) positive Shear in initial model of A:A w:wC base pair.



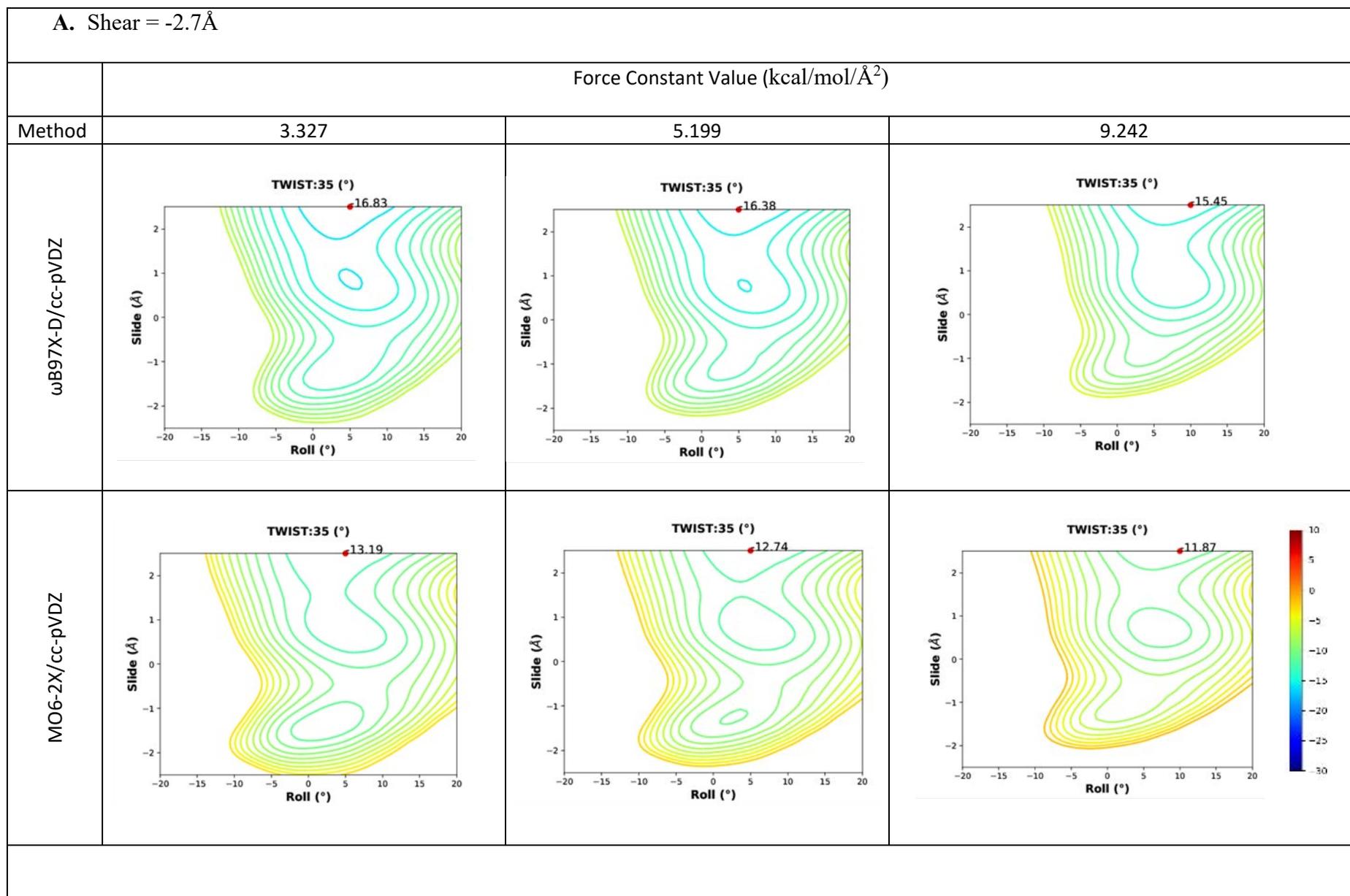


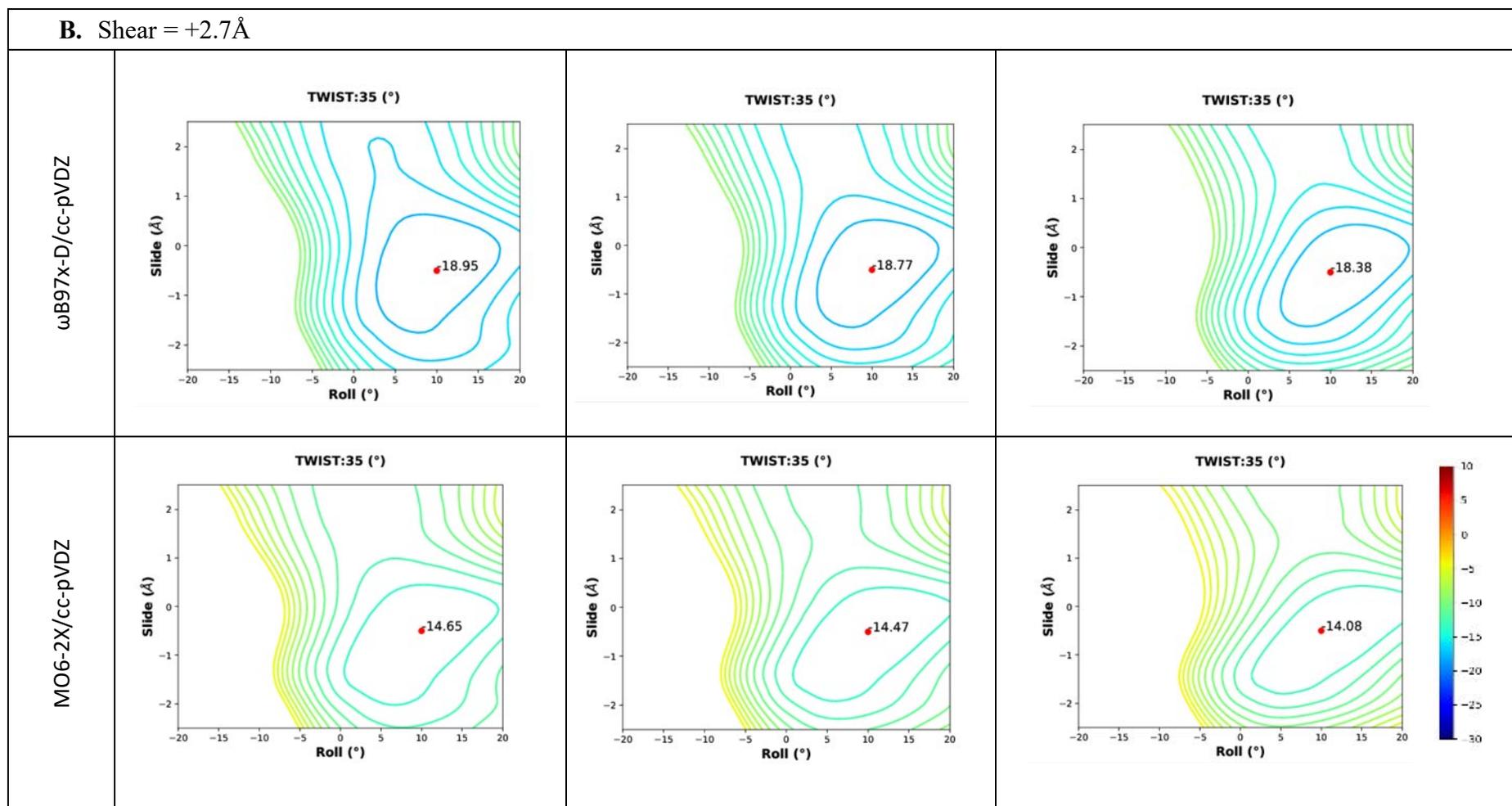
**Figure 4-12** Time evolution of Twist angle for several base pair step in GAU motif sequence during 500ns MD simulations for (A) negative Shear in initial model of A:A w:wC base pair and (B) positive Shear in initial model of A:A w:wC base pair.

We have carried out detail DFT-D based studies for base pair stacking energy to understand such dynamic behavior of the A:A w:wC base pair in different situations. As structures of the A:A w:wC::C:G W:WC dinucleotide step is seen with reasonably frequency, we have first carried out stacking energy scan in Roll-Twist-Slide space of the step considering most probable two values of Shear for the A:A w:wC base pair. Our earlier publications<sup>75,76</sup> indicate a need for addition of effect of the sugar-phosphate backbone to completely understand stacking preferences. Such hybrid energy can be calculated from variance of the observed C1'...C1' distances along a strand. Such Boltzmann inversion technique requires large number of data. Hence, we have considered different values of the force constant for calculation of total hybrid energy by assuming different values of standard deviations of C1'...C1' distance (0.1, 0.2, 0.3, 0.4, 0.5 and 0.6 and corresponding force constant k being 83.178, 20.794, 9.242, 5.199, 3.327, 2.310, 1.698 kcal/mol/Å<sup>2</sup> respectively) as obtained for all the dinucleotide steps in the RNABPDB database. The structures of the dinucleotide step with positive Shear of the A:A base pair show well defined lowest energy contour regions which is suitable for formation of A-RNA double helix with negative Slide, positive Roll (**Figure 4-13** and **Figure 4-14**) and Twist around 35°. The structures with negative Shear for the A:A w:wC base pair, however, show lowest energy region with positive Slide, which is unsuitable for A-RNA type structure formation<sup>183</sup>. The lowest energy zone with positive Shear is similar to the mean values of the corresponding parameters in the crystal structures. The energetically best structures with negative Shear have considerably less stability as compared to those with positive Shear.



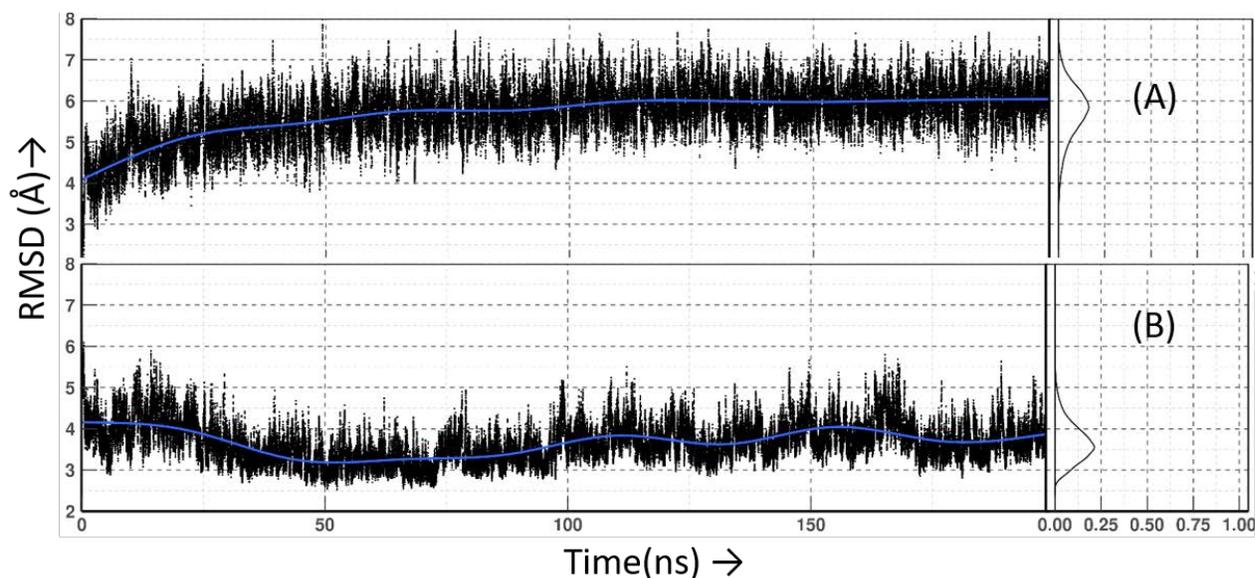
**Figure 4-13** Hybrid stacking iso-energy contours of A:A w:wC::C:G W:WC dinucleotide step sequence, with (A) negative Shear and (B) positive Shear of A:A w:wC base pair, considering MP2/cc-pVDZ and coarse-grain energy penalty for different force constant values.. Energy difference between two adjacent contour lines is 1kcal/mol. The best values of Roll and Slide for stacking are indicated by red dot.





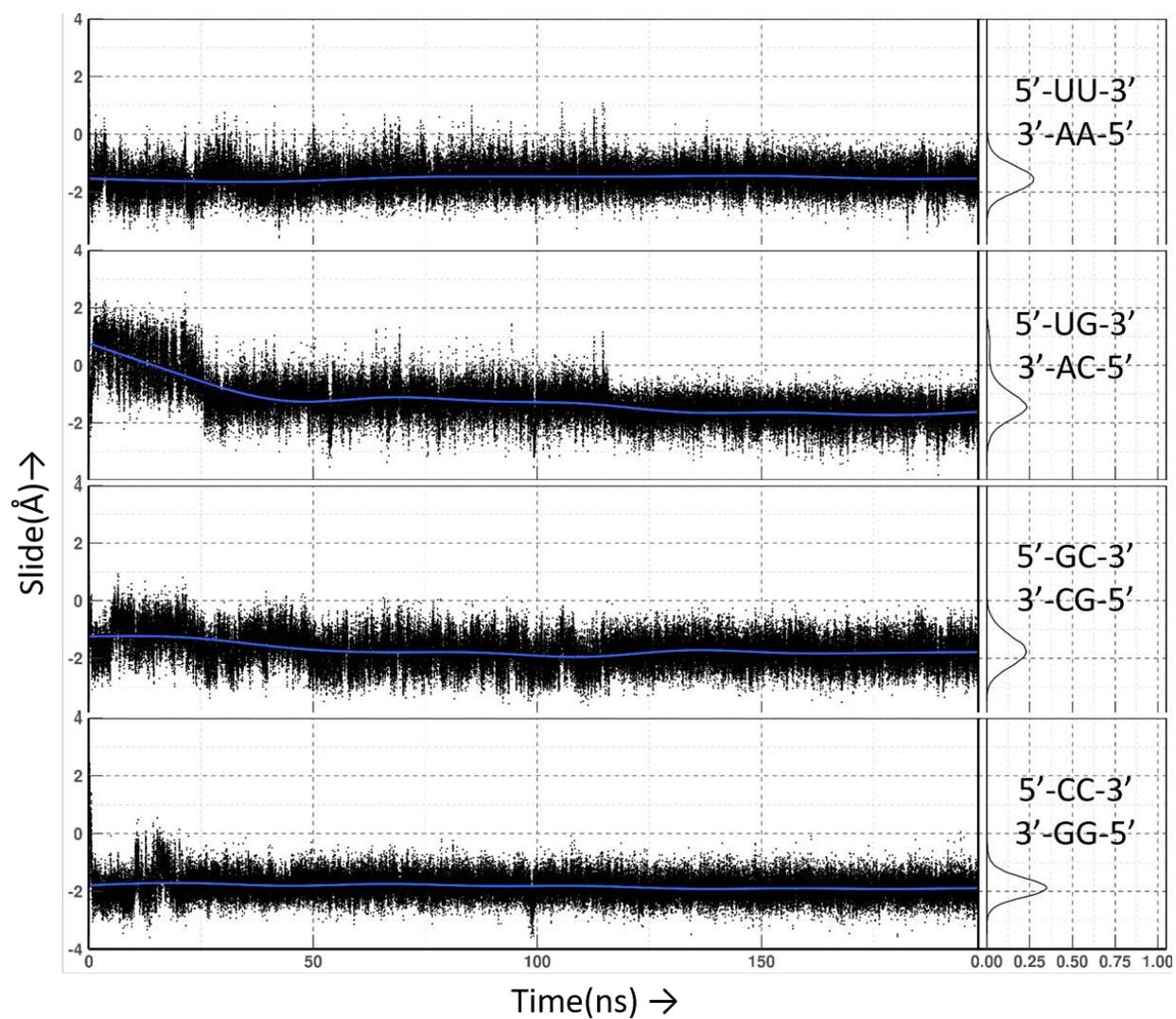
**Figure 4-14** (A) Hybrid Stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence, (A) with negative Shear of A:A w:wC and (B) with positive Shear of A:A w:wC, considering DFT-D energy and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist values is indicated by red dot.

One may argue that even positive Slide or negative Roll might be possible in RNA double helix. To further confirm our hypothesis we have performed all-atom MD simulation for a RNA double helix with canonical base pairs starting from B-DNA conformation having 0.56Å Slide and 2.21° Roll at physiological condition. The RNA backbone is found not at all stable in B-RNA like conformation as the structure changes significantly from initial model. The Root Mean Square Deviations (RMSD) between the initial structure and the trajectory snapshots continue to increase to large values around 6Å (Figure 4-15). Interestingly the RMSDs of the same snapshots with respect to regular A-RNA structure reduce within 50ns of simulation which confirms that the double helix is getting closer to A-RNA like conformation (Figure 4-15B).

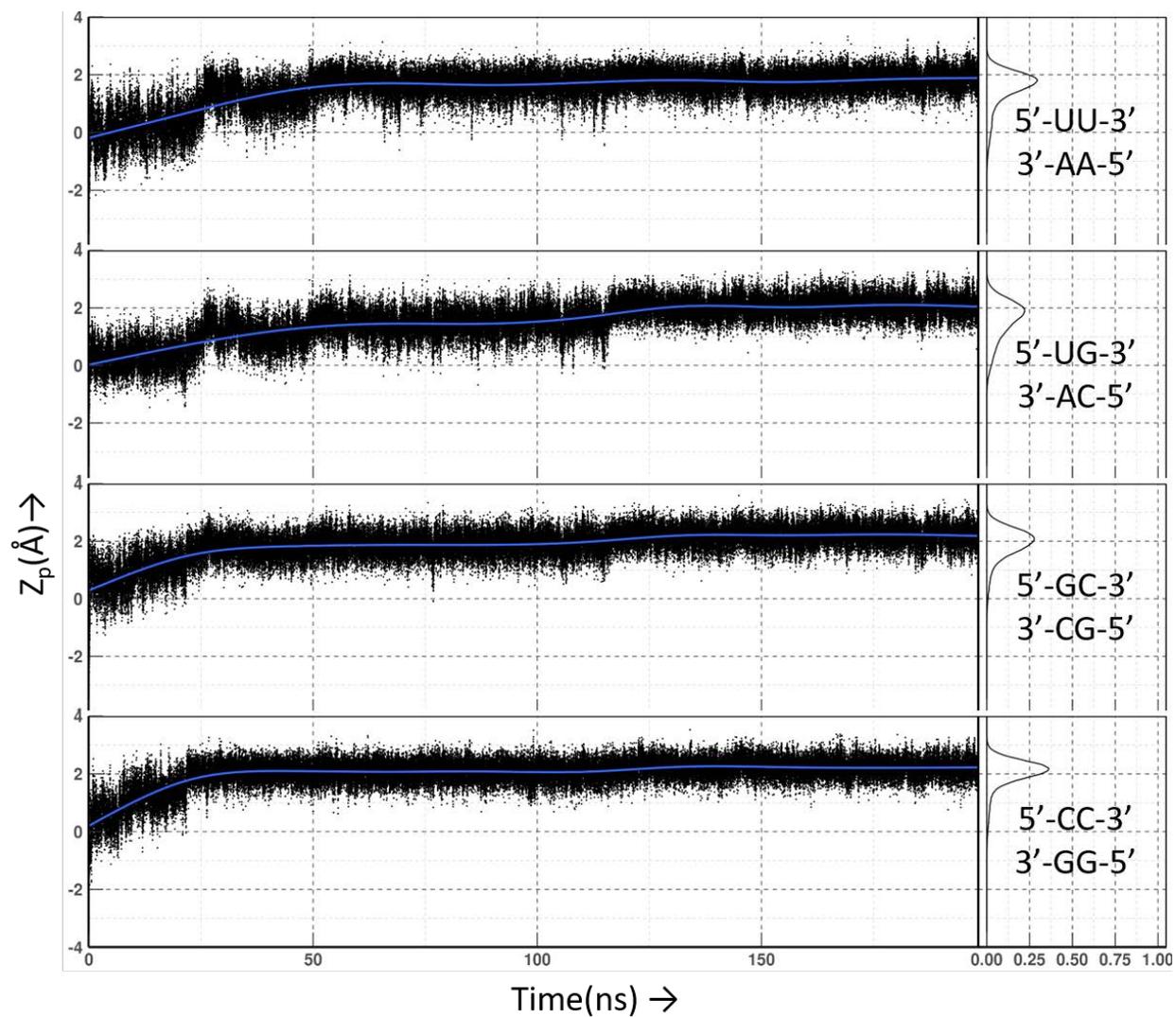


**Figure 4-15** Time evolution of RMSD of RNA double helix with only Watson-Crick base pairs from hypothetical B-RNA conformation with respect to initial structure, (B) RMSD of the same with respect to modeled A-RNA having same sequence considering Arnotts' fiber model.

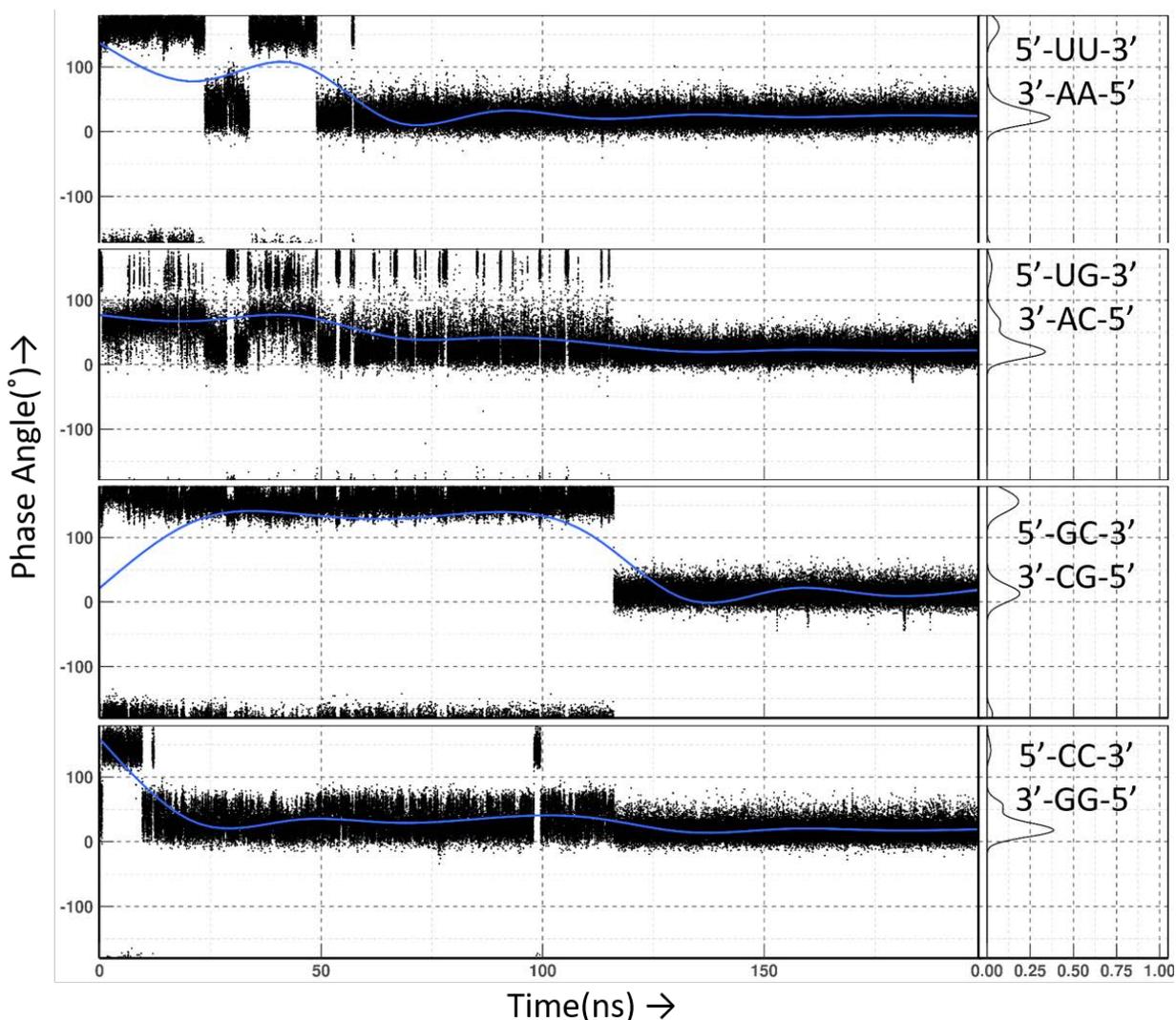
It is observed that positive Slide value, negative  $Z_p$ <sup>52</sup> and C2'-endo puckered sugars of the initial B-RNA conformation quickly attain negative and positive values with C3'-endo puckering, respectively, within 50ns<sup>183</sup> (**Figure 4-16, Figure 4-17 and Figure 4-18**). Similar feature was also noted in the crystal structures of DNA having some ribose sugars at different positions in PDB IDs 1D87, 1D88, 193D and 194D, where single ribose sugars induced A-form double helices. Take home message of the simulation is that positive Slide is impossible for RNA double helical structure.



**Figure 4-16** Time evolution Slide for different base paired dinucleotide steps of the double helix of RNA with only Watson-Crick base pairs.



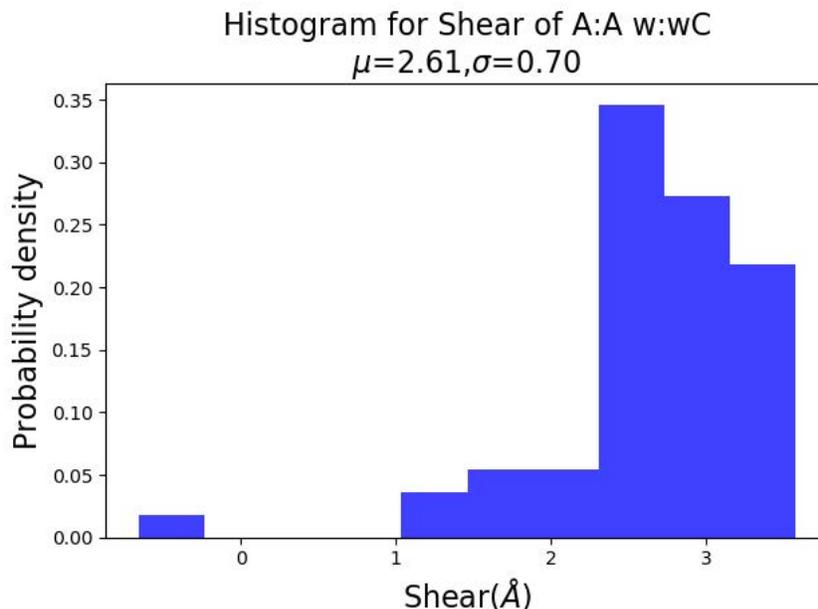
**Figure 4-17** Time evolution of  $Z_p$  for different base pair steps of RNA double helix with only Watson-Crick base pairs.



**Figure 4-18** Time evolution of sugar-pucker pseudo-rotation phase angle for different residues of RNA double helix with Watson-Crick base pairs.

In order to get support of our quantum chemical calculations we looked at the values of Shear in the experimental structures of A:A w:wC :: C:G W:WC and found these A:A base pairs almost exclusively have positive Shear (**Figure 4-19**). This analysis indicates the stacking energy analysis using the adopted method is highly accurate to predict structures of a dinucleotide step having non-canonical base pairs. A C:G base pairs is quite symmetric in terms of Shear but all the cis base pairs have some inherent asymmetry due to the sugar moieties, which breaks the symmetry

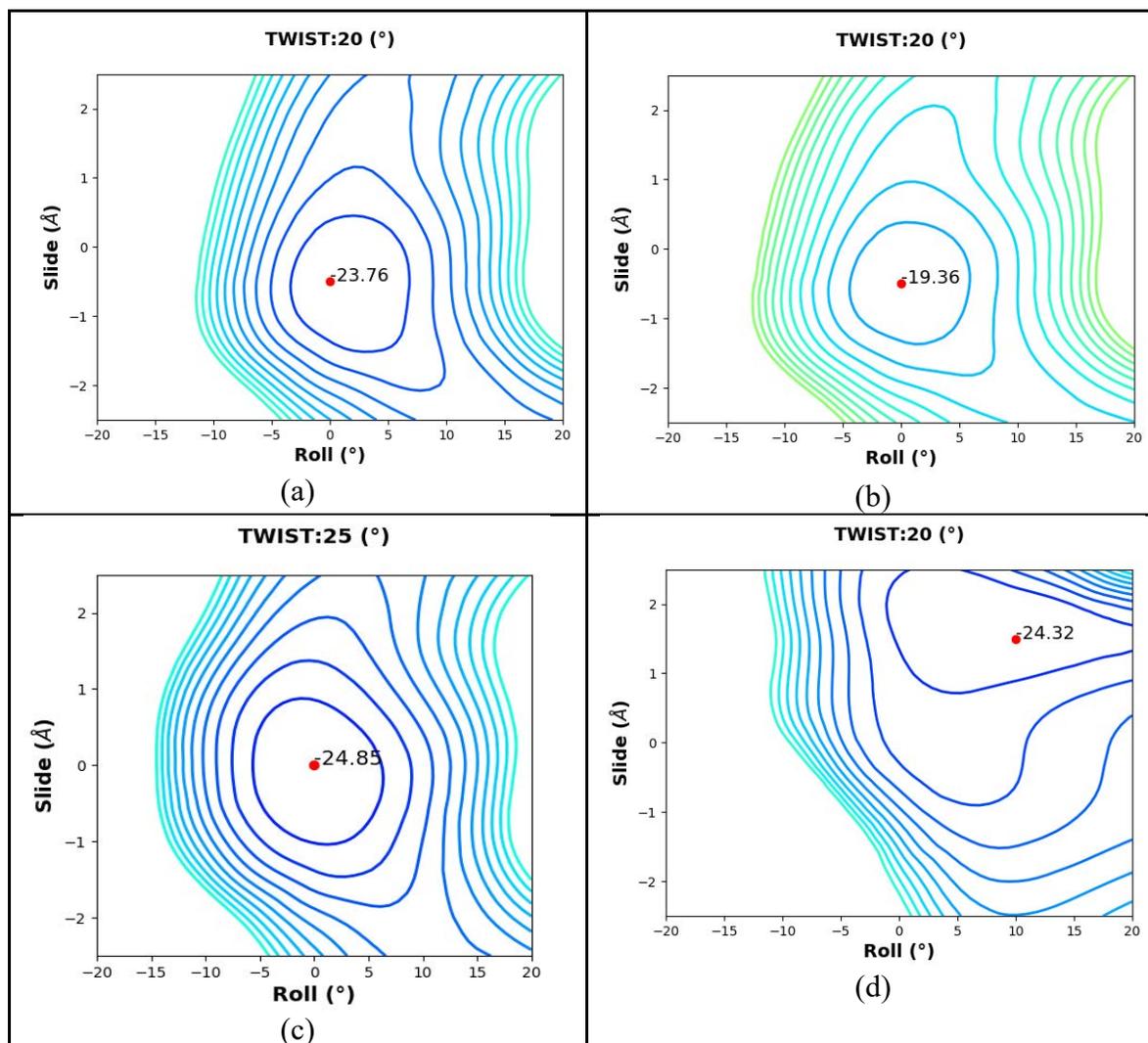
of A:A w:wC base pairs. The G:U W:WC base pairs are more asymmetric and can have the larger effect for symmetry breaking of A:A base pair.

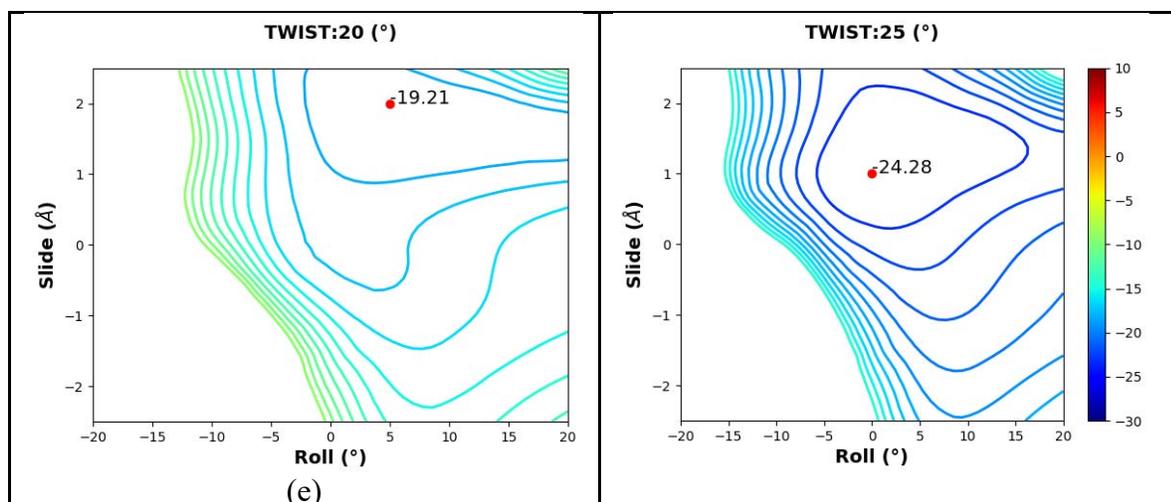


**Figure 4-19** Distribution of Shear value for the A:A w:wC base pair in A:A w:wC ::C:G W:WC dinucleotide step obtained from X-ray Crystal structure.

Hence We have carried out detail stacking energy analysis by varying the important base pair step parameters to remove the apparent contradictions of geometry of A:A w:wC base pair and of the A:A w:wC::G:U W:WC dinucleotide for both large positive and large negative Shear values. Energetically preferred stacking energy contours of the base pair stack with A:A w:wC base pair having negative Shear lie mostly in positive Roll and zero or negative Slide zone for Twist value in between  $10^\circ$  to  $35^\circ$  (**Figure 4-20** and **Table 4-8**). In contrast preferred stacking energy contours for same stack having positive Shear value of A:A w:wC base pair cover mostly positive Roll and large positive Slide zone for similar Twist values. Computationally expensive MP2 method also shows similar stacking energy trend for both types of structures, indicating our observation are not due to limitations of DFT-D (**Table 4-8** and **Figure 4-20**). We have further noted that stacking energies decrease at higher Twist values for both Shear values and the energetically

preferred contours for the same are displaced towards more negative Roll and positive Slide zone. This also indicate preferred Twist of this step is around 30°, as found in all the crystallographic structures.





**Figure 4-20** Intrinsic stacking energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence considering DFT-D energy at different Twist values for structures having (a-c) negative Shear value of A:A w:wC and (d-f) positive Shear value of A:A w:wC at  $\omega$ B97X-D/cc-pVDZ, M06-2X/cc-pVDZ and MP2/cc-pVDZ levels of theory, respectively. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist value is indicated by red dot.

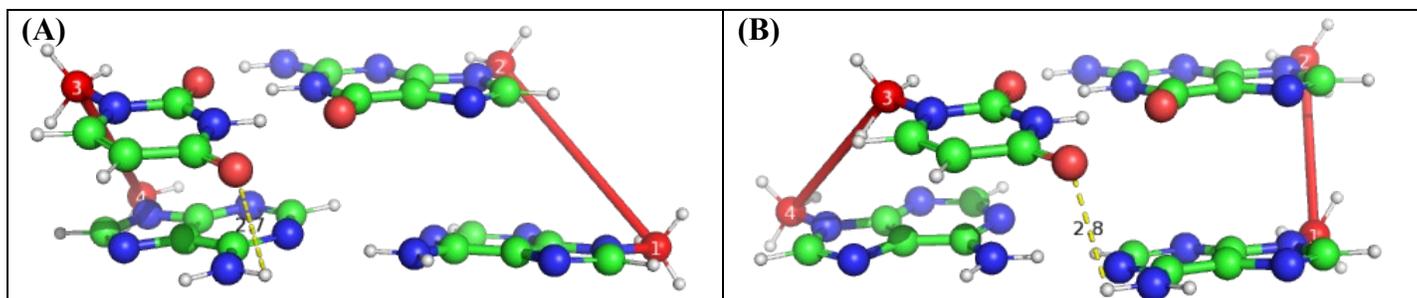
**Table 4-8** Geometrical and energetic parameter\* of dinucleotide step sequence A:A w:wC :: G:U W:WC having (A) negative Shear value (B) positive Shear value of A:A w:wC base pair for different Twist values having minimum DFT-D energy obtained by different levels of theory.

A. Shear = -2.7 Å									
Different Method	$\omega$ B97X-D/cc-pVDZ			M06-2X/cc-pVDZ			MP2/cc-pVDZ		
Twist(°)	Roll(°)	Slide(Å)	Best Stacking Energy (kcal/mol)	Roll(°)	Slide(Å)	Best Stacking Energy (kcal/mol)	Roll(°)	Slide(Å)	Best Stacking Energy (kcal/mol)
5	5	2.5	-22.62	5	2.5	-18.06	5	2.5	-24.26
10	10	-2	-22.47	5	-2	-17.68	5	2	-23.83
15	5	-1.5	-22.86	0	-0.5	-18.84	0	-0.5	-24.67
20	0	-0.5	-23.76	0	-0.5	-19.36	0	-0.5	-25.17
25	0	-0.5	-23.69	0	-0.5	-18.88	0	0	-24.85
30	0	0	-23.37	0	0	-18.33	0	0	-24.25
35	0	0	-22.93	-5	0.5	-18.25	-5	0.5	-23.93
40	-5	0.5	-22.61	-5	0.5	-17.86	-5	0.5	-23.21
45	-5	1	-22.24	-10	1	-17.29	-10	1	-22.58
50	-5	1.5	-21.33	-5	1	-16.29	-10	1.5	-21.67
55	-5	1.5	-20.39	-10	1.5	-15.43	-5	1.5	-20.45
60	-10	2	-19.07	-10	2	-14.38	-10	2	-19.27
B. Shear = +2.7 Å									
Different Method	$\omega$ B97X-D/cc-pVDZ			M06-2X/cc-pVDZ			MP2/cc-pVDZ		
Twist(°)	Roll(°)	Slide(Å)	Best Stacking Energy	Roll(°)	Slide(Å)	Best Stacking Energy	Roll(°)	Slide(Å)	Best Stacking Energy

			(kcal/mol)			(kcal/mol)			(kcal/mol)
5	10	2	-24.56	5	2	-20.10	5	2	-25.92
10	5	2	-24.73	5	2	-20.07	5	2	-25.76
15	5	2	-24.75	5	2	-19.84	5	2	-25.45
20	10	1.5	-24.32	5	2	-19.21	5	1.5	-24.86
25	5	1.5	-23.87	5	1.5	-18.66	0	1	-24.28
30	5	1.5	-23.13	0	1	-17.82	0	1	-23.68
35	0	1.5	-22.48	0	1.5	-17.10	0	1.5	-22.99
40	0	1.5	-22.00	0	1.5	-16.56	0	1.5	-22.42
45	-5	2	-21.38	-5	2	-16.19	-5	2	-21.91
50	-5	2.5	-21.36	-5	2.5	-16.32	-5	2.5	-21.92
55	-5	2.5	-20.46	-5	2.5	-15.66	-5	2.5	-21
60	0	2.5	-18.95	0	2.5	-14.52	0	2.5	-19.57

$$*E_{stacking\ energy} = E(A:B :: C:D) - E(A : B) - E(C : D)$$

The above energy contours give part of the structural preferences. As the base pairs are also covalently linked to the consecutive ones by sugar phosphate backbone, we have augmented the quantum chemical energy through coarse graining by addition of C1'...C1' virtual bond based penalty value<sup>76</sup>. A comparison of energies of the lowest hybrid stacking energy points for two Shear values indicate positive Shear is perhaps preferable (**Table 4-9**). We found some hydrogen bonds between the two base pairs possibly give extra stability to the lowest energy structures<sup>207,208</sup>. Typical hydrogen bonds are detected involving a 6-amino group of Adenine (which does not form base pairing hydrogen bond) with O4 of Uracil of same strand (**Figure 4-21**). Such hydrogen bonds are not detected in the energetically stable structures with positive Shear but some cross-strand hydrogen bonds are found to stabilize the contour zone (**Figure 4-21B**). However, the best stacking geometry for positive Shear appears in a region away from A-RNA like structural zone considering all three levels of hybrid theory and also for coarse grain hybrid energy force constants (**Table 4-9**, **Figure 4-22B**, and **Figure 4-23B**).



**Figure 4-21** Structures of *A:A w:wC::G:U W:WC* base paired dinucleotide steps having best hybrid stacking energy with (A) negative Shear for *A:A* step and (B) positive Shear for the same step generated by PyMol125, indicating possible hydrogen bonds. The virtual bonds mimicking sugar- phosphate backbone are shown by Red sticks.

**Table 4-9** Geometrical and energetic parameter of dinucleotide step sequence A:A w:wC :: G:U W:WC having (A) negative Shear and (B) Positive Shear value of A:A w:wC base pair for relevant Twist value having minimum DFT-D energy obtained by different levels of theory and coarse-grain energy penalty\* taking different force constant values.

A. Shear= -2.7Å												
Force Constant (kcal/mol/Å <sup>2</sup> )	ωB97X-D				M06-2X				MP2			
	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)
1.698	20	5	-1.5	-20.75	20	5	-1.5	-15.91	20	5	-1.5	-21.56
	25	5	-0.5	-21.43	25	5	-1.0	-16.44	25	5	-0.5	-22.14
	30	5	-0.5	-21.86	30	5	-0.5	-16.72	30	0	0.0	-22.46
	35	0	0.0	-21.94	35	0	0.0	-16.8	35	-5	0.5	-22.48
	40	0	0.5	-21.89	40	-5	0.5	-17.08	40	-5	0.5	-22.43
	45	-5	1	-21.74	45	-5	1	-16.77	45	-5	1	-22.08
	50	-5	1.5	-20.74	50	-10	1.5	-15.71	50	-10	1.5	-21.14
2.310	20	10	-2.0	-20.19	20	15	-2.5	-15.57	20	5	-1.5	-20.86
	25	5	-1.0	-20.78	25	5	-1.0	-15.87	25	5	-0.5	-21.44
	30	5	-0.5	-21.45	30	5	-0.5	-16.31	30	5	-0.5	-21.87
	35	0	0.0	-21.58	35	0	0.0	-16.44	35	0	0.0	-22.05
	40	0	0.5	-21.67	40	-5	0.5	-16.8	40	-5	0.5	-22.15
	45	-5	1	-21.56	45	-5	1	-16.59	45	-5	1	-21.9
	50	-5	1.5	-20.53	50	-10	1.5	-15.52	50	-10	1.5	-20.95
3.327	20	15	-2.5	-19.57	20	15	-2.5	-15.08	20	10	-2.0	-20.08
	25	5	-1.0	-19.83	25	5	-1.0	-14.92	25	5	-1.0	-20.41
	30	5	-0.5	-20.77	30	5	-0.5	-15.63	30	5	-0.5	-21.19
	35	5	0.0	-21	35	0	0.0	-15.85	35	0	0.0	-21.46
	40	0	0.5	-21.31	40	-5	0.5	-16.34	40	-5	0.5	-21.69
	45	-5	1	-21.26	45	-5	1	-16.29	45	-5	1	-21.6
	50	-5	1.5	-20.17	50	-10	1.5	-15.21	50	-10	1.5	-20.64
5.199	20	15	-2.5	-18.68	20	15	-2.5	-14.19	20	15	-2.5	-19.12
	25	10	-2.0	-18.33	25	5	-1.5	-13.7	25	10	-2.0	-18.87
	30	5	-0.5	-19.52	30	5	-0.5	-14.47	30	5	-1.0	-19.94
	35	5	0.0	-20.2	35	5	0.0	-14.97	35	5	-0.5	-20.48
	40	0	0.5	-20.66	40	-5	0.5	-15.48	40	-5	1	-20.85
	45	-5	1	-20.71	45	-5	1	-15.74	45	-5	1	-21.05
	50	-5	1.5	-19.52	50	-10	1.5	-14.63	50	-10	1.5	-20.06

Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT and MD Studies

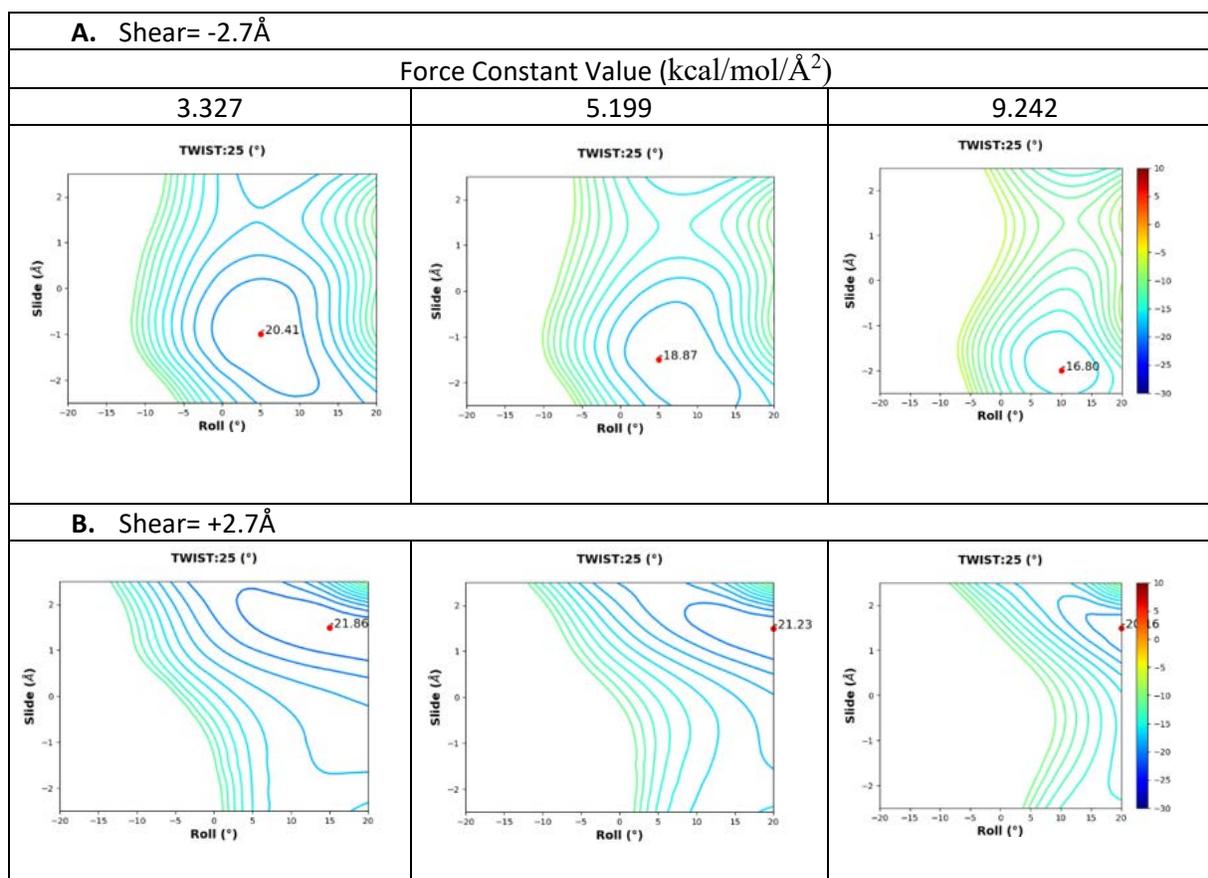
9.242	20	15	-2.5	-16.75	20	15	-2.5	-12.26	20	15	-2.5	-17.19
	25	10	-2.0	-16.34	25	10	-2.0	-11.71	25	10	-2.0	-16.8
	30	10	-1.0	-17.12	30	10	-1.0	-12.24	30	10	-0.5	-17.5
	35	10	2.5	-18.9	35	5	-0.5	-13.29	35	5	0.0	-18.75
	40	5	2.5	-19.76	40	0	0.5	-13.96	40	0	2.5	-19.53
	45	0	1.5	-19.57	45	-5	1	-14.55	45	-5	1.5	-19.98
	50	-5	1.5	-18.12	50	-10	1.5	-13.37	50	-10	1.5	-18.8
20.794	20	15	-2.5	-11.23	20	20	-2.5	-6.95	20	15	-2.5	-11.67
	25	15	-2.0	-10.74	25	15	-2.0	-6.34	25	10	-2.0	-11.1
	30	15	2.5	-14.96	30	15	2.5	-9.51	30	15	2.5	-14.76
	35	10	2.5	-17.91	35	15	2.5	-12.17	35	10	2.5	-17.5
	40	5	2.5	-19.16	40	5	2.5	-13.28	40	5	2.5	-18.82
	45	0	2.5	-18.35	45	-5	2.5	-12.69	45	-5	2.5	-18.43
	50	-10	2.5	-15.23	50	-10	2.5	-10.5	50	-10	2.5	-16.15
83.178	20	20	-2	14.44	20	20	-2.0	18.6	20	20	-2.0	13.92
	25	20	2.5	2.02	25	20	2.5	6.86	25	20	2.5	1.64
	30	20	2.5	-11.29	30	20	2.5	-6.19	30	20	2.5	-11.22
	35	15	2.5	-16.45	35	15	2.5	-10.86	35	15	2.5	-15.98
	40	10	2.5	-16.82	40	10	2.5	-10.96	40	10	2.5	-16.28
	45	0	2.5	-13.21	45	0	2.5	-7.51	45	0	2.5	-13.12
	50	-10	2.5	-4.55	50	-10	2.5	0.18	50	-10	2.5	-5.47
<b>B. Shear= +2.7Å</b>												
Force Constant (kcal/mol/Å <sup>2</sup> )	ωB97x-D				M06-2X				MP2			
	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)	Twist(°)	Roll(°)	Slide(Å)	Best hybrid Stacking Energy (kcal/mol)
1.698	20	10	2	-22.89	20	20	1.5	-18.13	20	10	2	-23.09
	25	15	1.5	-22.6	25	15	1.5	-17.61	25	10	1.5	-22.73
	30	10	1.5	-22.11	30	10	1.5	-16.89	30	5	1.5	-22.33
	35	5	1.5	-21.76	35	5	1.5	-16.38	35	0	1.5	-22.12
	40	0	1.5	-21.68	40	0	1.5	-16.24	40	0	1.5	-22.1
	45	-5	2	-21.29	45	-5	2	-16.1	45	-5	2	-21.82
	50	-5	2.5	-21.13	50	-5	2.5	-16.09	50	-5	2.5	-21.69
2.310	20	20	1.5	-22.5	20	20	1.5	-17.87	20	10	2	-22.66
	25	15	1.5	-22.32	25	20	1.5	-17.37	25	15	1.5	-22.33
	30	10	1.5	-21.85	30	10	1.5	-16.63	30	5	1.5	-21.93
	35	5	1.5	-21.56	35	5	1.5	-16.18	35	5	1.5	-21.86

Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT and MD Studies

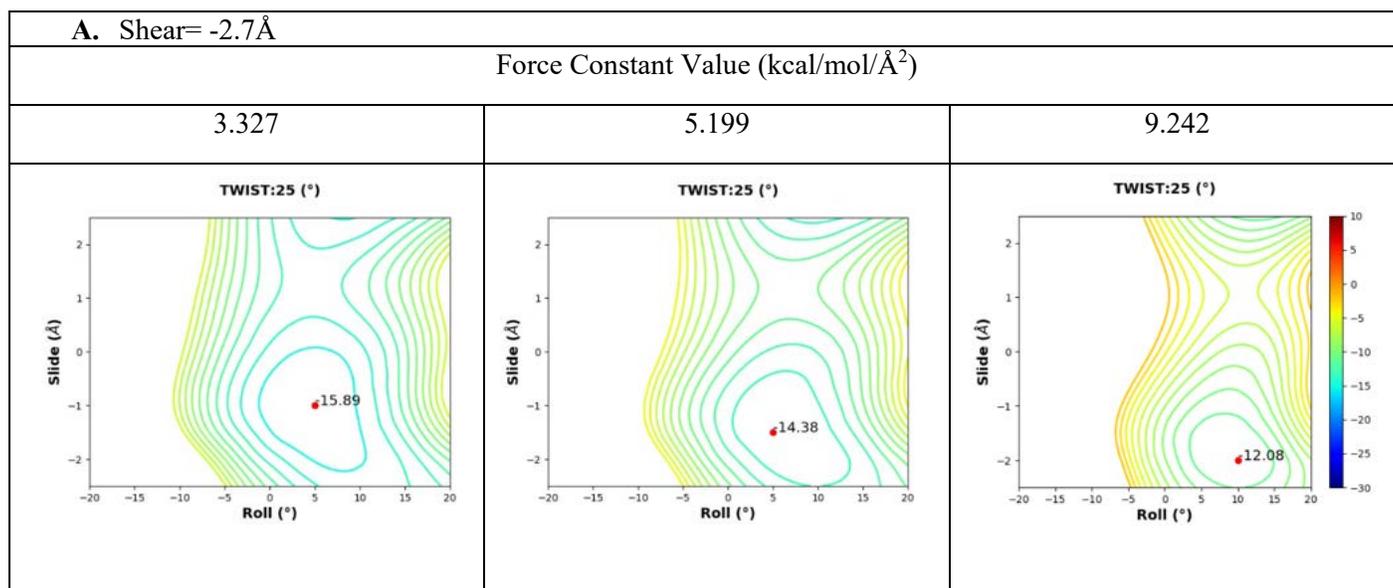
	40	0	1.5	-21.57	40	0	1.5	-16.13	40	0	1.5	-21.99
	45	-5	2	-21.26	45	-5	2	-16.07	45	-5	2	-21.79
	50	-5	2.5	-21.05	50	-5	2.5	-16.01	50	-5	2.5	-21.61
3.327	20	20	1.5	-22.05	20	20	1.5	-17.42	20	20	1.5	-22.2
	25	15	1.5	-21.85	25	20	1.5	-17.1	25	15	1.5	-21.86
	30	20	1	-21.48	30	20	1	-16.36	30	10	1.5	-21.5
	35	10	1	-21.28	35	5	1.5	-15.86	35	5	1.5	-21.54
	40	0	1.5	-21.38	40	0	1.5	-15.94	40	0	1.5	-21.8
	45	-5	2	-21.2	45	-5	2	-16.01	45	-5	2	-21.73
	50	-5	2.5	-20.91	50	-5	2.5	-15.87	50	-5	2.5	-21.47
5.199	20	20	1.5	-21.24	20	20	1.5	-16.61	20	20	1.5	-21.39
	25	20	1.5	-21.31	25	20	1.5	-16.6	25	20	1.5	-21.23
	30	20	1	-21.19	30	20	1	-16.07	30	20	1	-21.18
	35	15	1	-20.96	35	15	1	-15.66	35	10	1	-21.05
	40	5	1	-21.13	40	5	1.5	-15.62	40	0	1.5	-21.45
	45	0	1.5	-21.13	45	-5	2	-15.91	45	-5	2	-21.63
	50	-5	2.5	-20.66	50	-5	2.5	-15.62	50	-5	2.5	-21.22
9.242	20	20	1.5	-19.48	20	15	-2.5	-12.26	20	20	1.5	-19.48
	25	20	1.5	-20.24	25	10	-2.0	-11.71	25	20	1.5	-20.24
	30	20	1	-20.55	30	10	-1.0	-12.24	30	20	1	-20.55
	35	15	1	-20.57	35	5	-0.5	-13.29	35	15	1	-20.57
	40	10	1	-20.83	40	10	1	-15.38	40	10	1	-21.03
	45	0	1.5	-21.06	45	0	1.5	-15.74	45	-5	2	-21.42
	50	-5	2.5	-20.12	50	-10	2.5	-15.26	50	-10	2.5	-20.76
20.794	20	20	2	-15.68	20	20	2	-11.32	20	20	2	-15.71
	25	20	1.5	-17.16	25	20	1.5	-12.45	25	20	1.5	-17.08
	30	20	1	-18.72	30	20	1.5	-13.66	30	20	1	-18.71
	35	20	1	-20.25	35	20	1	-15.24	35	20	1	-20.21
	40	10	1	-20.46	40	10	1	-15.01	40	10	1	-20.66
	45	0	1.5	-20.85	45	0	1.5	-15.53	45	0	1.5	-21.18
	50	-5	2	-19.25	50	-10	2.5	-14.33	50	-10	2.5	-19.83
83.178	20	20	2	2.48	20	20	2	6.84	20	20	2	2.45
	25	20	2	-3.8	25	20	2	0.63	25	20	2	-3.74
	30	20	1.5	-11.38	30	20	1.5	-6.63	30	20	1.5	-11.24
	35	20	1	-18.82	35	20	1	-13.81	35	20	1	-18.78
	40	15	0.5	-20.12	40	15	1	-14.79	40	15	0.5	-20.28
	45	5	1	-20.76	45	5	1	-15.38	45	5	1	-20.98

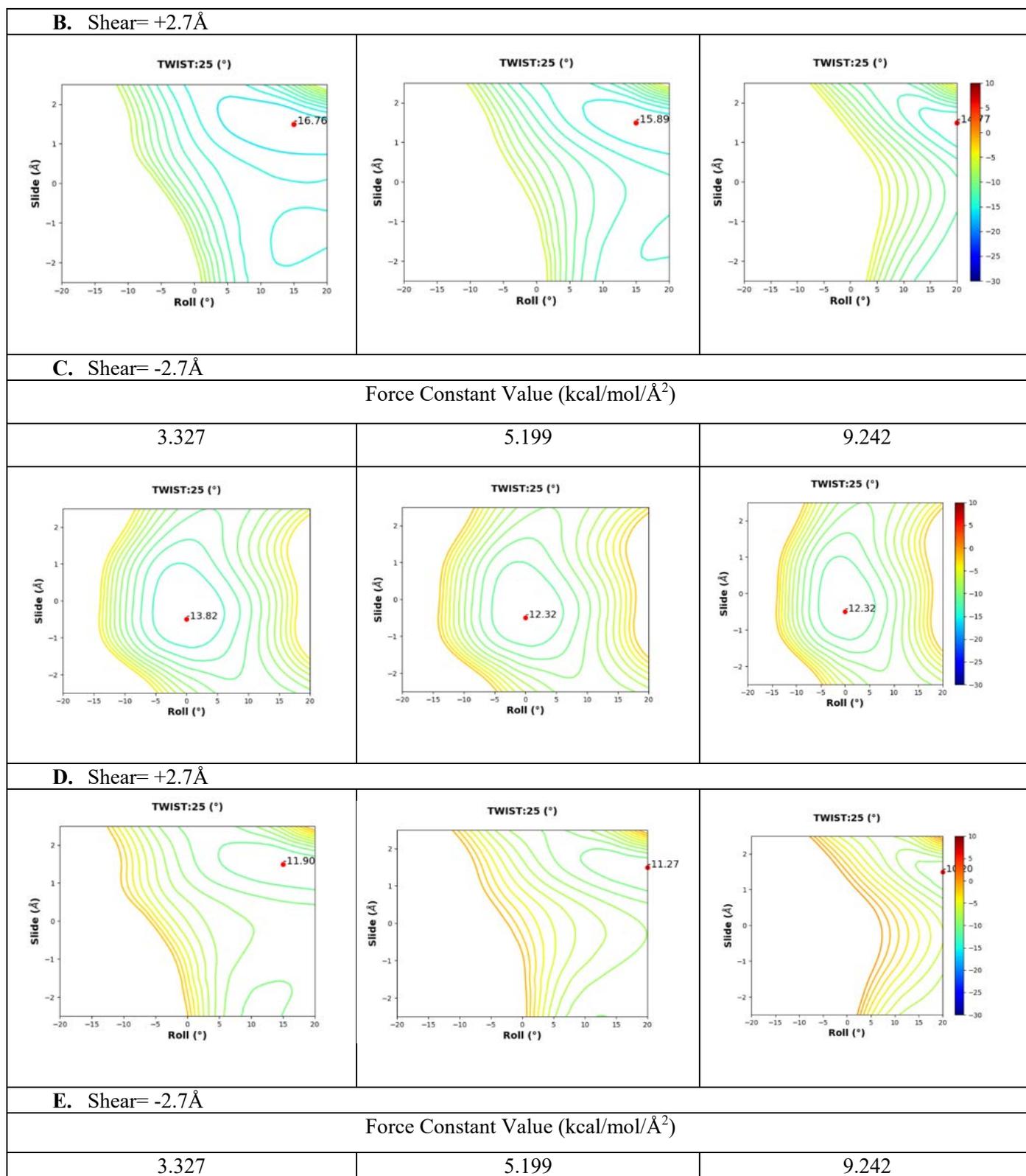
*Stacking Interaction by C-H...N Hydrogen Bonded Base Pair: Hybrid DFT and MD Studies*

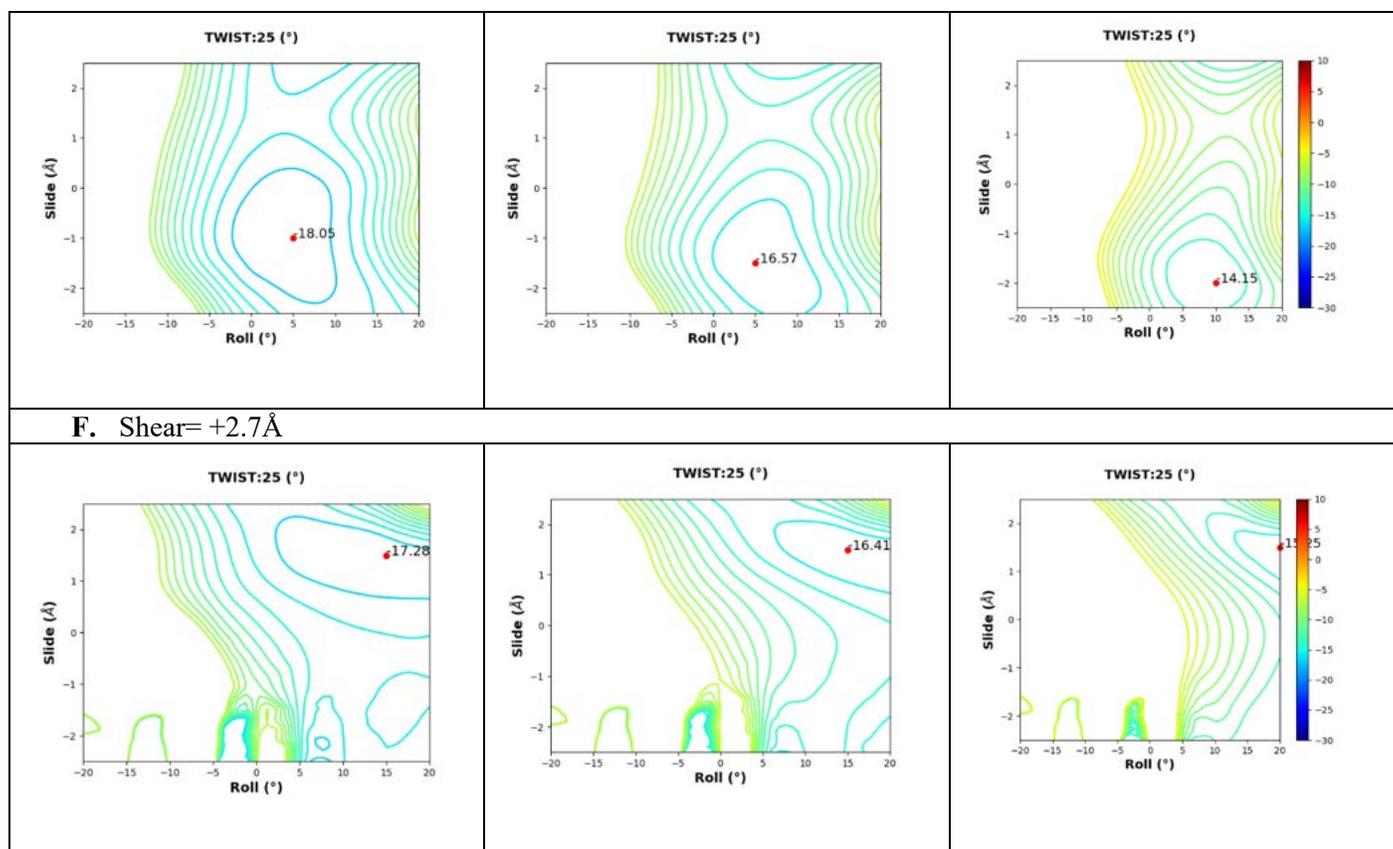
	50	-5	2	-15.9	50	-5	1.5	-11.06	50	-5	2	-16.35
--	----	----	---	-------	----	----	-----	--------	----	----	---	--------



**Figure 4-22** Hybrid stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence, with (A) negative Shear and (B) positive Shear of A:A w:wC base pair, considering energy using MP2/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each virtual bond stretching force-constant value is indicated by red dot.



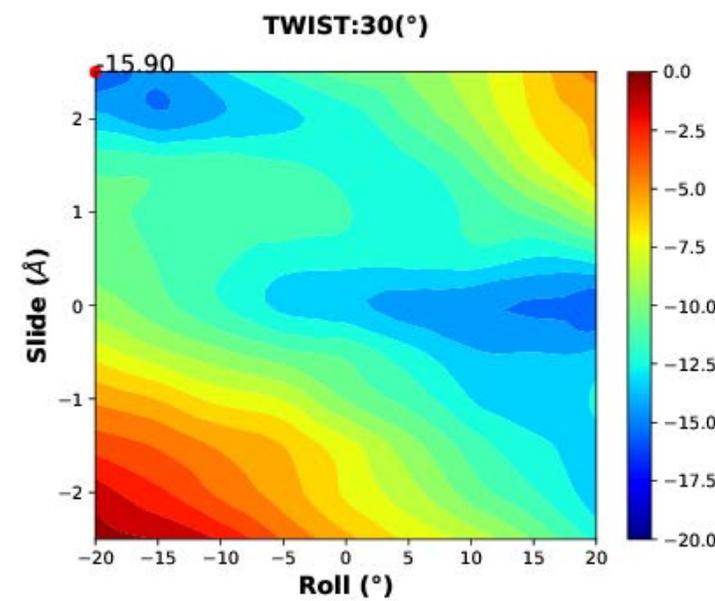




**Figure 4-23** Hybrid Stacking iso-energy contours of A:A w:wC::G:U W:WC dinucleotide step sequence considering water environment by using CPCM, (A) and (B) with  $\omega$ B97X-D, (C) and (D) with M06-2X and (E) and (F) with MP2 functional for mentioned Shear values of A:A w:wC, considering DFT-D energy and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each Twist values is indicated by red dot.

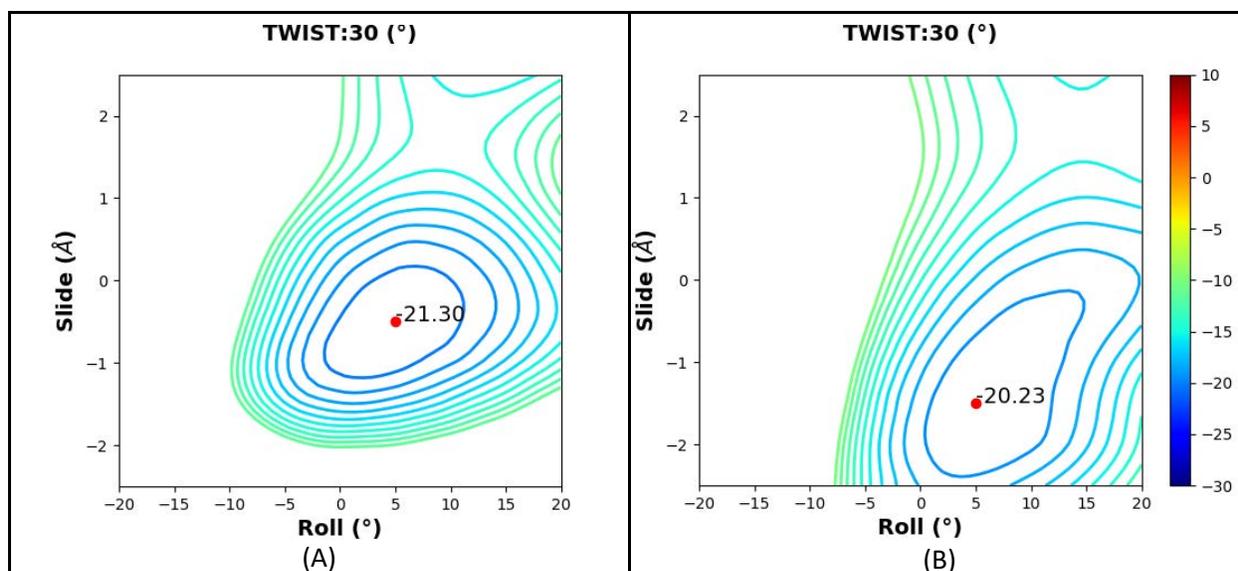
Energetically best stacking geometry for stacks having negative Shear for A:A w:wC base pair appears in the A-RNA like structural region which corresponds to positive Roll and negative Slide (**Figure 4-22, Figure 4-23** and **Table 4-8**). The base pair stacks are also stabilized by their pi-pi stacking overlap (Ovl), which also have been calculated for all the structures using NUPARM<sup>116</sup>. We have further calculated difference between stacking overlap values for the structures as  $Ovl_{(positive\ shear)} - Ovl_{(negative\ shear)}$  as shown in **Figure 4-24**. It is clear that the structures with positive Roll, negative Slide and negative Shear have always significantly larger stacking overlap. It is expected that these double helical regions remain buried within the core of long RNA folded structure, within somewhat hydrophobic environment. Else, environment along the groove

regions may be hydrophilic while the base faces remain covered by other base pairs, giving rise to a mixed environment to the stacks. Nevertheless, we have analyzed the stacking energy contours considering water environment by using CPCM method in calculations of the DFT-D and MP2 energies. The contours (**Figure 4-23**) show near identical behavior with the vacuum plots which reduces the significance of calculation in implicit water environment.



**Figure 4-24** Contour plot of A:A w:wC::G:U W:WC showing difference between stacking overlap from positive Shear to Negative Shear of A:A w:wC base pair.

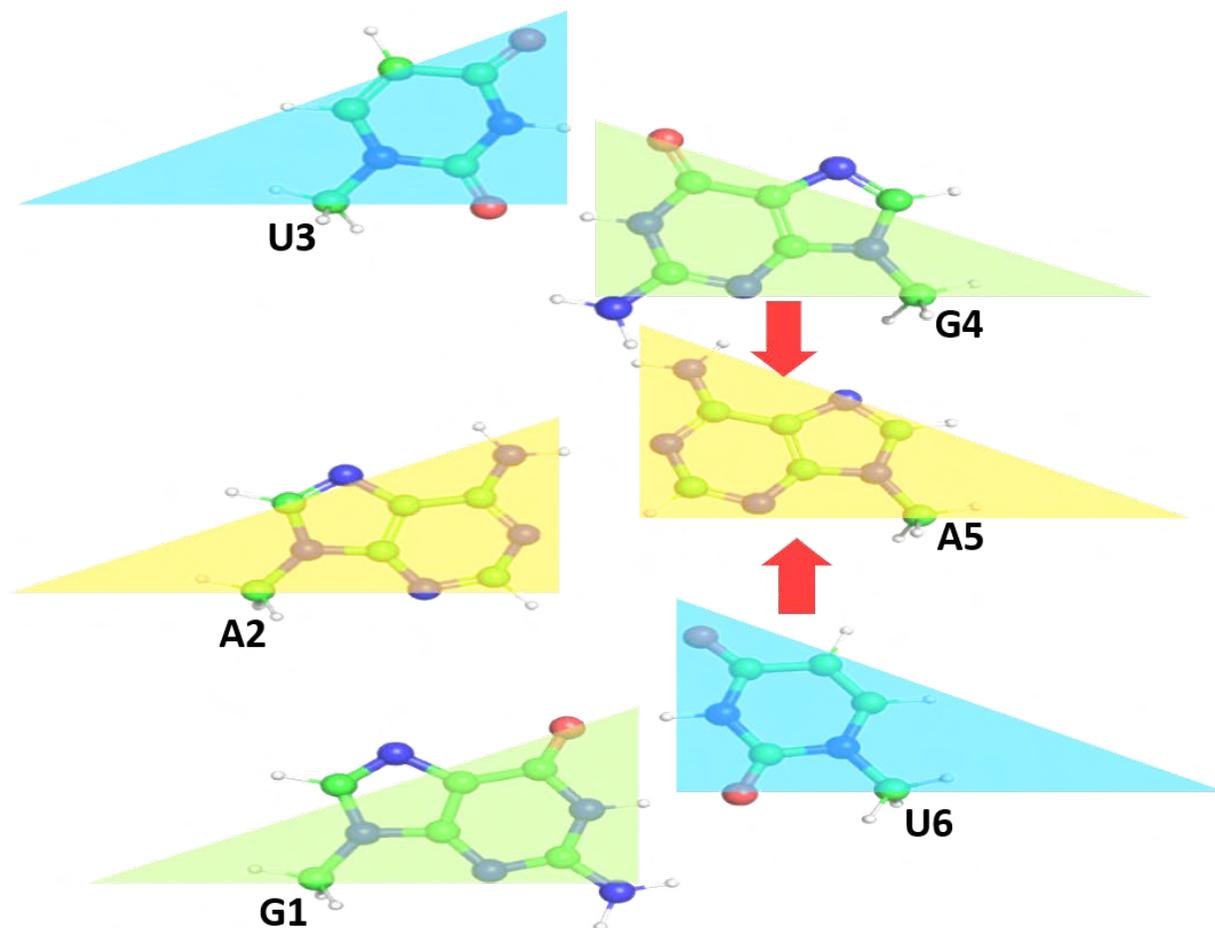
. We have also analyzed hybrid stacking energy scan in Roll-Twist-Slide space for A:A w:wC::U:G W:WC dinucleotide step using DFT-D method and found positive Shear gives lowest energy contours near A-RNA structure forming region (**Figure 4-25**). The experimental structures also mostly have positive Shear for the A:A base pairs forming this dinucleotide step (32 out of 36 structures have positive Shear in RNABPDB database).



**Figure 4-25** Hybrid stacking energy contour of A:A w:wC :: U:G W:WC having (A) negative Shear of A:A w:wC base pair and (B) positive Shear of A:A w:wC base pair, at 30° Twist for 9.242 kcal/mol/Å<sup>2</sup> force constant representing backbone effect. Energy difference between two adjacent contour lines is 1kcal/mol. The best stacking energy for each plot is indicated by red dot.

Thus the quantum chemical calculations indicate, similar to crystallographic results, that the A:A w:wC::G:U W:WC dinucleotide prefers negative Shear while the A:A w:wC::U:G W:WC dinucleotide prefers positive Shear to stabilize A-RNA like structure with positive Roll, negative Slide and Twist around 30°. In these orientations the stacking overlap between the two base pairs are also largest (**Figure 4-24**). This is possibly due to the large inherent Shear of G:U (or U:G) base pair. When the A:A w:wC base pair stacks on top of (at 3'-location) a G:U base pair, the Adenine of second strand needs to follow the Uracil (Uracil moved towards major groove) of second strand, which demands a negative Shear of the A:A base pair (**Figure 4-26**). Similarly when the A:A base pair is located at 5'-direction of (below) a G:U base pair (consider U3:G4 replaced by G3:U4 in **Figure 4-26**, where the U4 moves towards major groove), the Adenine of the second strand needs to move towards major groove, having negative Shear. Thus, the structure of the GAG sequence motif remains stable with only negative Shear of the central A:A w:wC base pair. In case of UAG sequence motif, the Guanine base below the A:A base pair of the second strand (consider G1:U6 replaced by U1:G6 in **Figure 4-26**) moves towards minor groove and the Uracil base of second

strand on top of the A:A base pair moves towards major groove (**Figure 4-26**). Neither Shear value of the A:A w:wC base pair can improve stacking between the Adenine bases with the G:U and U:G base pairs simultaneously. As a result, the central Adenine bases are exposed to solvent from either face. This is thermodynamically unstable situation, which the system wants to avoid. As a result, the Adenine residues forms an Adenine Zipper like structure during MD simulations. This structure is also not stable enough due to extra space available to them at that internal loop. Hence the bases finally adopt a different geometry involving their Hoogsteen and Sugar edges in trans orientation.



**Figure 4-26** Schematic diagram of the two base pair dinucleotide steps highlighting two opposing forces exerted by U6 to A5 and U4 to A5 making the A2:A5 base pair as frustrated.

This base pair in s:hT orientation is additionally stabilized by a 2'-OH group mediated hydrogen bond (**Figure 4-7**). In case of the GAU motif, the bases of the second strand move in opposite directions – the Uracil below the A:A base pair moves towards major groove while the Guanine on top of A:A base pair moves towards minor groove and this situation is shown in **Figure 4-26**. The Adenine bases remain buried in all situations – (i) when paired with negative Shear, (ii) when paired with positive Shear or (iii) when unpaired with broken hydrogen bonds. The bases do not have any space to readjust forming Adenine Zipper like structure also. However, the G:U base pair at 5' direction of the central A:A base pair demands negative Shear of the A:A base pair while the

U:G base pair at 3' direction of it demands positive Shear for the same A:A base pair. The A:A w:wC base pair maintains hydrogen bonds corresponding to negative Shear for some time (few ps) and transforms to another form corresponding to positive Shear. With either Shear it is unstable due to two inconsistent stacking interactions. This was possibly the reason behind conformational plasticity of the GAU motif, as observed in MD simulation.

#### **4.4 Conclusion:**

Our study using hybrid quantum chemical calculations followed by all-atom molecular dynamics simulations indicate possibility of A:A w:wC base pair with a specific type of structure in trinucleotide 5'(GAG).5'(UAU) sequence motif with the central Adenine residues paired in w:wC way. The studies additionally indicate that the trinucleotide sequences 5'(UAG).5'(UAG) or 5'(GAU).5'(GAU) cannot form stable three dimensional structure maintaining w:wC type base pairing between the central Adenine residues. The A:A w:wC base pair attempts to adopt a Shear sign depending on its nearest neighboring G:U W:WC base pair, which also has high Shear value. But its promiscuous nature allows it to choose a particular type of Shear value depending on its neighboring G:U or U:G W:WC base pair at its 5' or 3' side.

Such bimodal distribution of Shear is also noticed in U:U W:WC base pair, which also appears symmetric in all respect but is stabilized by two strong N—H...O hydrogen bonds. The U:U base pair, hence, can have similar symmetry and symmetry breaking due to stacking, leading to intrinsically disordered nature for adoption of double helical structure. Opening the U:U W:WC base pair, however may not be easy at physiological temperature due to presence of two stronger hydrogen bonds between the bases. Many U:U base pair was also seen to have stronger ion-mediated interactions due to strong electronegative sites around both major and minor grooves of U:U W:WC base pair<sup>175</sup>. The energy barrier between the two modes of A:A w:wC structure

appears to be too high when only Shear is considered as Reaction Coordinate. However, such alteration might occur affecting some other relative orientations, such as Open or Stretch also, thereby reducing the instability of the Saddle Point significantly. It is expected that the transition state for inter-conversion between two modes of U:U W:WC base pair can be much larger.

Prediction of structure of biomolecules, especially RNA using bioinformatics approach, i.e., molecular modeling using average values of different structural parameters, is quite useful and important. In the cases where bimodal distributions of such structural parameters are seen, then the knowledge based analysis using their central tendencies fail. We could successfully predict structure of GAG motif considering one of the preferred geometries of Shear for the A:A base pair. However, in all other motifs, we observe tug of war between two types of Shear, leading to conformational plasticity and intrinsic disorder of the GAU and UAG motifs.

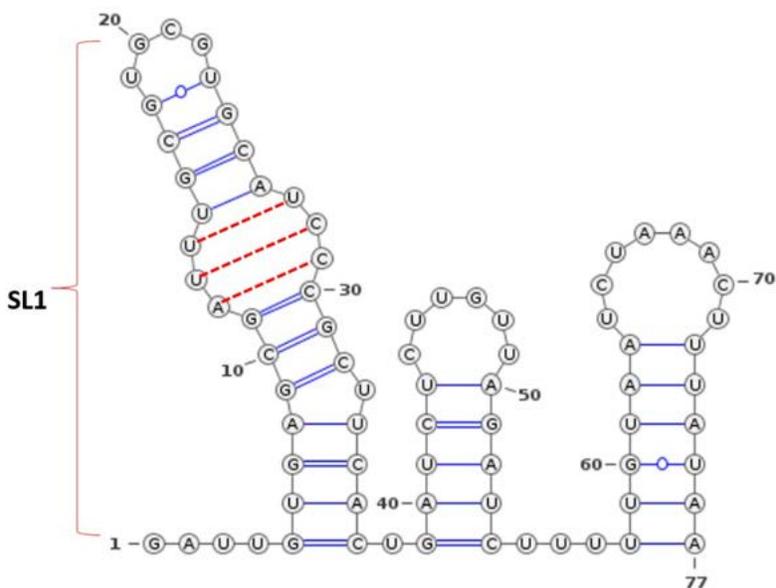
Similarly there are quite a few stable and frequently detected base pairs, such as A:U H:WC, which are not found within dinucleotide stacks. It would be interesting to detect whether some feature, as described above induces its inability to form double helix.

## **Chapter- 5 Molecular Modeling of SL1-helix of SARS-Corona Viral RNA**

## 5.1 Introduction

The 5' untranslated regions of RNA of Human SARS-Corona virus may adopt few stem loop structures and the SL1 helix among them may contain U:U, A:C and U:C non Watson-Crick base pairs(**Figure 5-1**)<sup>121</sup>. It was found from our RNABPDB database that there are several possibilities of U:U, A:C and U:C base pairing involving different edges of the bases. However, their Watson-Crick base pairing edges appeared to be most favorable for base pair formation in those mismatch systems considering the frequencies of occurrences of the base pairs as well as interaction energies.

It was shown that the two Uracil bases could form base pair in their neutral forms in two ways with either positive Shearing motion between the two bases or with negative Shearing motion. Similarly A:C base pairing is possible in two ways: (i) when the Adenine base



**Figure 5-1.** 5' UTR region of RNA of Human Corona Virus

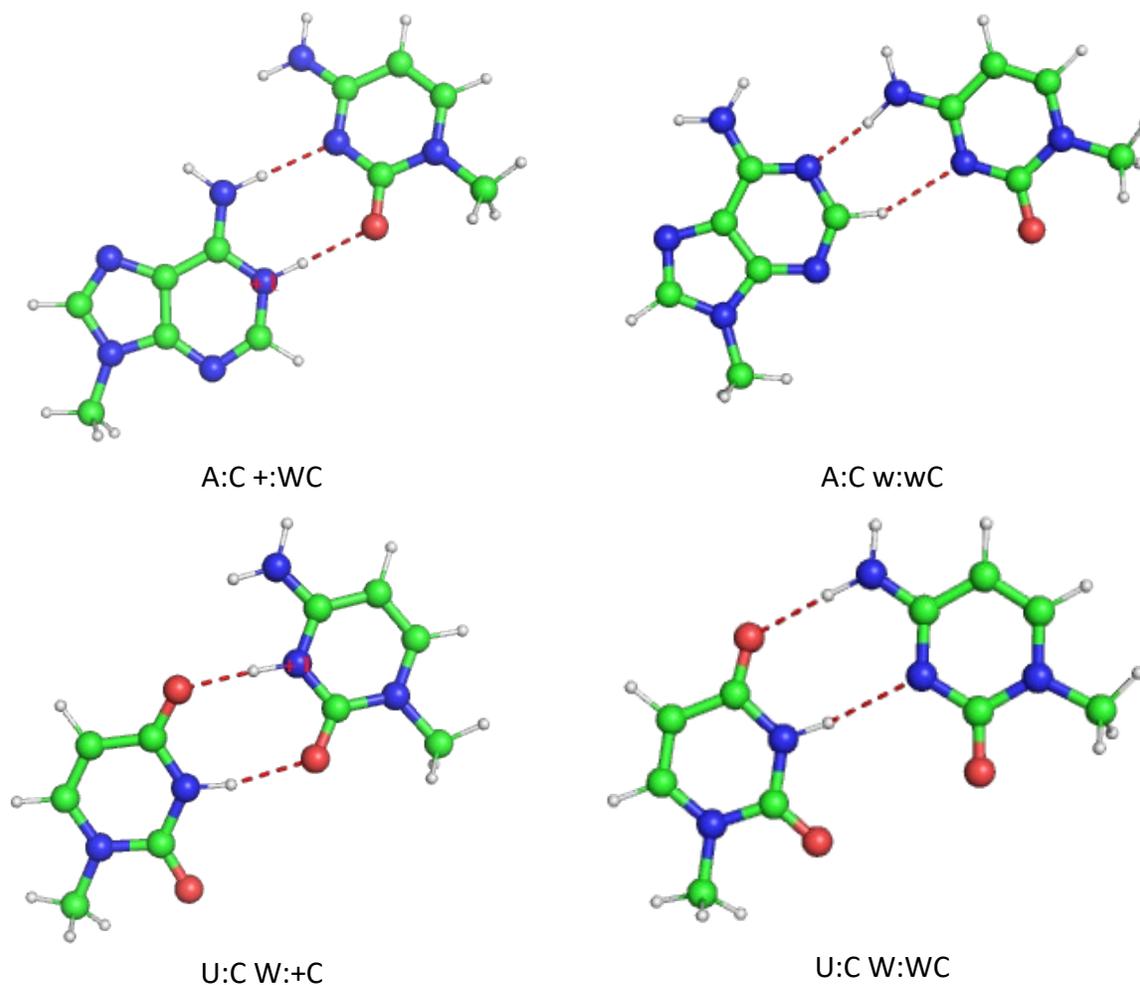
is additionally protonated at its N1 position to form two hydrogen bonds between N1(A)...O2(C) and N7(A)-N6(C) with stabilizing energy of -42.60 kcal/mol (**Figure 5-2**) and (ii) with a C—H...N mediated weak hydrogen bonding involving (A)N1(A)...N4(C) and C2(A)...N3(C) with stabilizing energy of -8.48 kcal/mol (**Figure 5-2**). The second type of hydrogen bonding does not require extra protonation although it is energetically less stable. Similarly the U:C base pairing is also possible in two ways (i) in neutral form of the two bases through N3(U)...N3(C) and O4(U)...N4(C) H-

bonds (**Figure 5-2**) and (ii) through N3(C)...O4(U) and O2(C)...N3(U) H-bonds which is possible only when N3(C) is protonated. Naturally the protonated base pair is more stable with stronger interaction energy of -27.12 kcal/mol vs. -12.31 kcal/mol of the neutral U:C base pair (**Figure 5-2**).

Thus, it appears that both Adenine and Cytosine can become protonated bases giving rise to stronger base pairs with Cytosine and Uracil, respectively, as compared to their neutral forms. However, protonation is costly at physiological pH, as shown by several groups, including those by theoretical calculations. In *in vitro* systems the pH requires to be lowered by acid but in the cellular systems such artificial pH reduction is not possible. Although pK<sub>a</sub> values of the bases are referred to quite often but I did not find any recent reference on measurement of these values with modern sophisticated instrumentation of theoretical results apart from the values quoted in Saenger's book<sup>121</sup>. Moreover, there are several proton donors, such as Lys, His, etc. amino acids in cellular environment, which can donate their protons to a nucleotide base. Hence we have studied the reaction mechanism for transition from neutral bases to protonated bases assisted by Lys residue and water molecules.

DNA or RNA nucleotide bases have several polar atoms, such as carbonyl oxygen, amino (primary or secondary) nitrogen, imino nitrogen. Some of these are fully protonated at physiological condition, such as the primary amino groups in N6 of Adenine, N2 of Guanine and N4 of Cytosine. These amino groups are generally involved in hydrogen bonding with complementary bases to form Watson-Crick base pairs. In addition, these are also involved in base sequence specific molecular recognition for binding with different proteins, which control gene regulation. There are non-protonated sites also in the bases, such as the carbonyl oxygens O6 of Guanine, O2 of Cytosine and O2 and O4 of Thymine or Uracil, N1 of Adenine, N3 of Cytosine, N7 and N3 of the purines. The nitrogen atoms generally remain in imino form and all these atoms

are also often involved in hydrogen bonding as H-bond acceptors. The sugar edge of RNA of all the nucleotides are capable to form H-bonds involving 2'-OH group, which can act both as H-bond donor as well as acceptor, which is not possible for DNA.

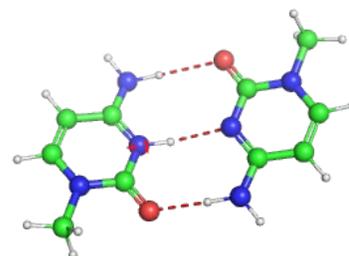


**Figure 5-2** Three dimensional representations for protonated and non-protonated base pair.

Several studies, however, reported possibility of alternate protonation of some of the bases in synthetic as well as biological contexts. As for example, C.G\*C base triple has been characterized at low pH where the Cytosine bases of the third strand remain in protonated charged form<sup>209</sup>. In this form the N3 imino nitrogen of Cytosine obtains a proton from solvent and become secondary amino group. This allows the protonated Cytosine to form a second hydrogen bond with N7 imino nitrogen atom of Guanine while the N4 amino group of Cytosine forms another H-bond

with O6 atom of Guanine. The base pairing between Guanine and Cytosine in their neutral forms in such a manner appears impossible as the two imino nitrogen atoms (N3 of Cytosine and N7 of Guanine) may repel each other in absence of the additional proton. Similar protonation at N1 position of Adenine also is possible at lower than physiological pH, which has been reported recently in crystal structure of an RNA double helix<sup>210</sup>. In this high resolution crystal structure the authors could identify positions of the extra hydrogen atoms.

Similar situation is perhaps possible in the structures of RNA where various types of base pairing between non-complementary bases are seen, which are generally called non-canonical base pairs. These structures are, however, solved at moderate resolution where the positions of the hydrogen atoms could not be identified.



**Figure 5-3** *C:C W:+T base pair.*

However, considering possibility of electrostatic repulsion between negatively charged imino nitrogen atoms or carbonyl oxygen atoms, one could hypothesize that protonation of one of the bases is perhaps mandatory. Considering this possibility, several protonated base pairs are identified by BPFIND<sup>51</sup> software, similar to DSSR<sup>104</sup> or other tools for base pair identification in RNA.

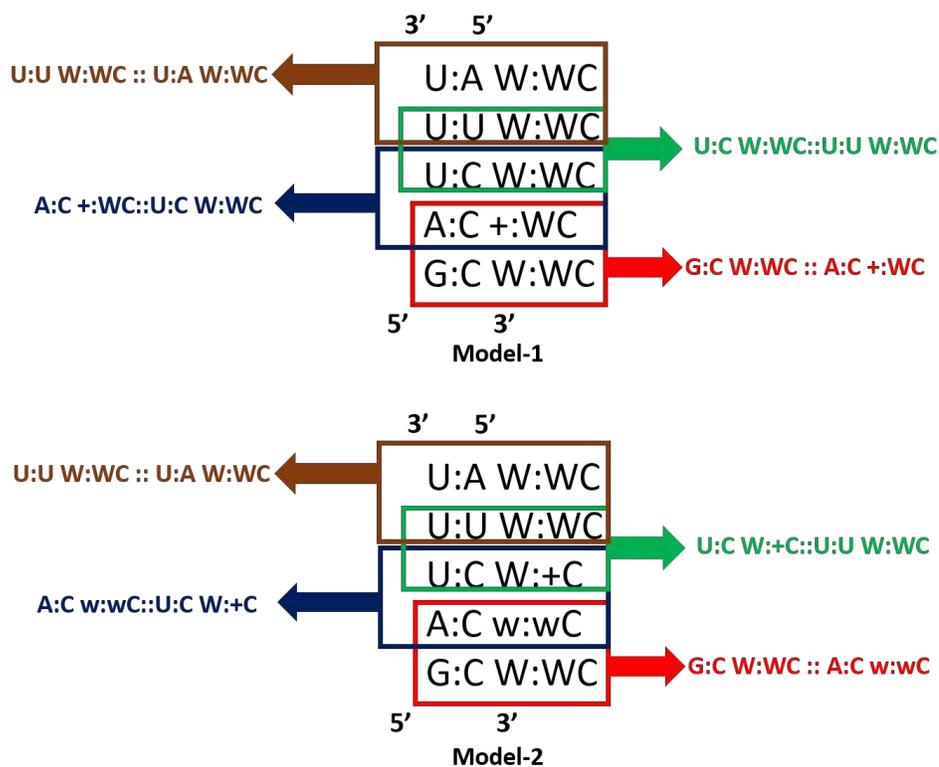
The pKa of the bases are known to be 2 to 3 units far from physiological pH (7.5), which makes the protonated bases unstable at normal condition. The protonated bases may remain stabilized once they form base pairing involving their protonated faces. It has been found that the base pairs having extra proton in positively charged conditions are generally more stable with very high interaction energies<sup>211</sup> (<http://hdrnas.saha.ac.in/rnabpdb/>). This is due to additional interaction between the positive charge of the protonated base and the charge induced dipole moment of the neutral base. Nevertheless, protonation of the nucleotide bases are costly. It was

hypothesized that the energy cost due to unfavorable protonation is compensated by the extra stabilization of the formation of the protonated base pairs<sup>212</sup>. The theoretical and experimental studies indicate Cytosine protonation is possible at close to physiological pH while Adenine protonation is possible at rather lower acidic pH. Different groups also performed experiments to show that in the primordial soup at highly acidic condition Cytosine bases are significantly more stable in its protonated form. Such N3 protonated Cytosine residues can pair with another neutral Cytosine in reverse Watson-Crick form (W:WT) to form a four stranded DNA or RNA structure known as i-motif (**Figure 5-3**)<sup>213,214</sup>.

The previous report on base protonation commented only on thermodynamic barrier of protonation process, which does not explain how the bases acquire proton from cellular environment<sup>132</sup>. Hence, I have attempted to understand the process where some amino acid may help the protonation, thereby lowering the activation energy.

As mentioned above the base pairs appear in SARS-corona viral RNA in a double helical arrangement, where stacking interaction is additional stabilizing factor. Hence, we have studied stacking interaction between U:U::U:C and U:C::A:C base paired dinucleotide sequences also considering two different shearing motions of U:U base pair, two different shearing motions of the U:C base pair due to different protonation state and two sheared A:C base pair due to two different protonation states. Considering all the models and their energy landscapes we have proposed two models (**Figure 5-4**) and we have marked the dinucleotide steps in the **Figure 5-4**. These dinucleotide steps have been considered for the stacking energy scan in the Roll-Slide hyperspace for different Twist angle. It is worth to mention that we have also considered two Shear value of U:U W:WC in each model. This complete study is expected to shade light on the structure of the

SARS-corona viral RNA to combat our fight against the deadly disease and world-wide pandemic situation.



**Figure 5-4** Proposed model for SL1 Helix of RNA of SARS corona virus.

## 5.2 Methods

We used QST3<sup>215</sup> of Gaussian16<sup>125</sup> to calculate the transition states for the protonation of Cytosine and Adenine. Transition state structures have been verified by single imaginary frequency. We have considered CH<sub>3</sub>NH<sub>3</sub><sup>+</sup> to mimic lysine in the calculation<sup>216</sup>. We have used B3LYP<sup>217</sup>/cc-pVDZ<sup>126,127</sup> level of theory during transition state calculations. As, role of water molecule/molecules are broadly highlighted in our proposed reaction mechanism for protonation, we have done all calculations in implicit solvent model. The effect of solvent polarizability has

been analyzed by implementation of conductor-like polarizable continuum model through united-atom topological model with  $\epsilon = 78.39$  (CPCM)<sup>189</sup>.

We have followed the same protocols and methods for calculations of stacking energy scan as mentioned in Chapter 3 and 4.

### 5.3 Result and Discussion

It was experimentally shown that the 5'-untranslated region of such viral RNA can adopt few stem loop structures<sup>121</sup>, SL1, SL2, etc. The SL1 helix has a symmetric internal loop containing three non-canonical base pairs, namely A:C, U:C and U:U. Our database indicates different types of base pairings between these residues are possible (Table 5-1) while base pairing through their Watson-Crick edges are most likely as these appear most frequently with high interaction energy in cWW (or W:WC) orientation. NMR and thermodynamics signatures suggested possibility of protonation of the A:C base pair at low pH with increase in melting temperature of the SL1 helix of RNA of SARS corona virus<sup>121</sup>. There is also U:C base pair in SL1 helix and we cannot rule out the possibility protonated U:C W:+C base pair, as this base pair also estimates sufficient interaction energy (Table 5-1). Kinetic study will reveal the relative order of ease of protonation for Adenine and Cytosine. It requires the determination of transition state for protonation.

**Table 5-1** Different types of A:C, C:U and U:U base pairs involving different edges in either orientation along with their frequencies in the non-redundant crystal structure dataset and interaction energies. Although A:C and C:U base pairing is theoretically possible in 18 different ways but some possibilities are discarded due to impossibility of two hydrogen bonds in those cases, such as A:C H:HC, etc. or some base pairing being too infrequent, such as U:U H:ST (shown with Blue patch in the RNABPDB database).

Base Pairing type	Frequency	Interaction energy	Remarks
A:C W:WC	240	-42.60 and -8.48	Combination of +:WC and w:wC
A:C W:WT	89	-16.91 and -20.21	Combination of W:WT and +:WT
A:C S:SC	1233	-9.92	C—H...O mediated

A:C H:WT	531	-16.32	
A:C W:SC	533	-14.66	Sugar 2'-OH mediated
A:C W:ST	80	-14.56	C—H...O mediated
A:C S:WC	147	-17.34	Sugar 2'-OH mediated
A:C H:SC	64	-9.26	Sugar 2'-OH mediated
A:C H:ST	29	-14.56	Sugar 2'-OH mediated
A:C S:HC	27	-3.55	Sugar 2'-OH mediated
C:U W:WC	158	-27.12 and -12.31	Combination of +:WC and W:WC
C:U W:WT	11	-12.06	
C:U W:HC	6	-7.28	C—H...O mediated
C:U W:SC	80	-7.05	Sugar 2'-OH mediated
C:U S:WC	24	-14.36	Sugar 2'-OH mediated
C:U S:WT	12	-14.05	Sugar 2'-OH mediated
C:U H:ST	34	-5.62	Sugar 2'-OH mediated
U:U W:WC	693	-12.42	
U:U W:WT	103	-10.63	
U:U W:HC	47	-8.37	
U:U W:HT	97	5.04	C—H...O mediated
U:U W:SC	20	-13.21	Sugar 2'-OH mediated
U:U W:ST	22	-14.29	Sugar 2'-OH mediated

We have proposed a mechanism for the protonation of bases, Adenine and Cytosine assisted by water and Lysine. One molecule of water assists to transfer proton from lysine ( $\text{CH}_3\text{NH}_3^+$ ) to cytosine. Whereas two molecules of water are required during the protonation of Adenine. Coordinates for the transition states for protonation of Adenine and Cytosine are given in **Table 5-2**.

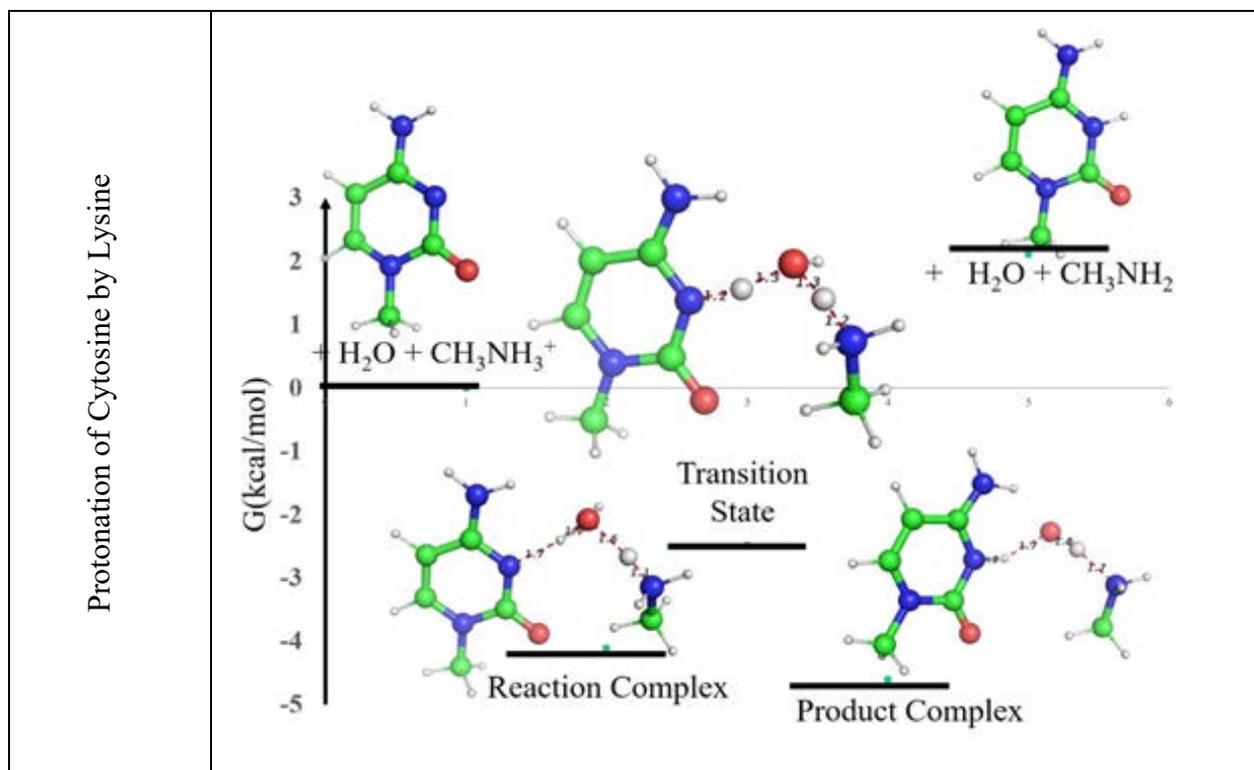
**Table 5-2** Optimized co-ordinates of atoms in transition state.

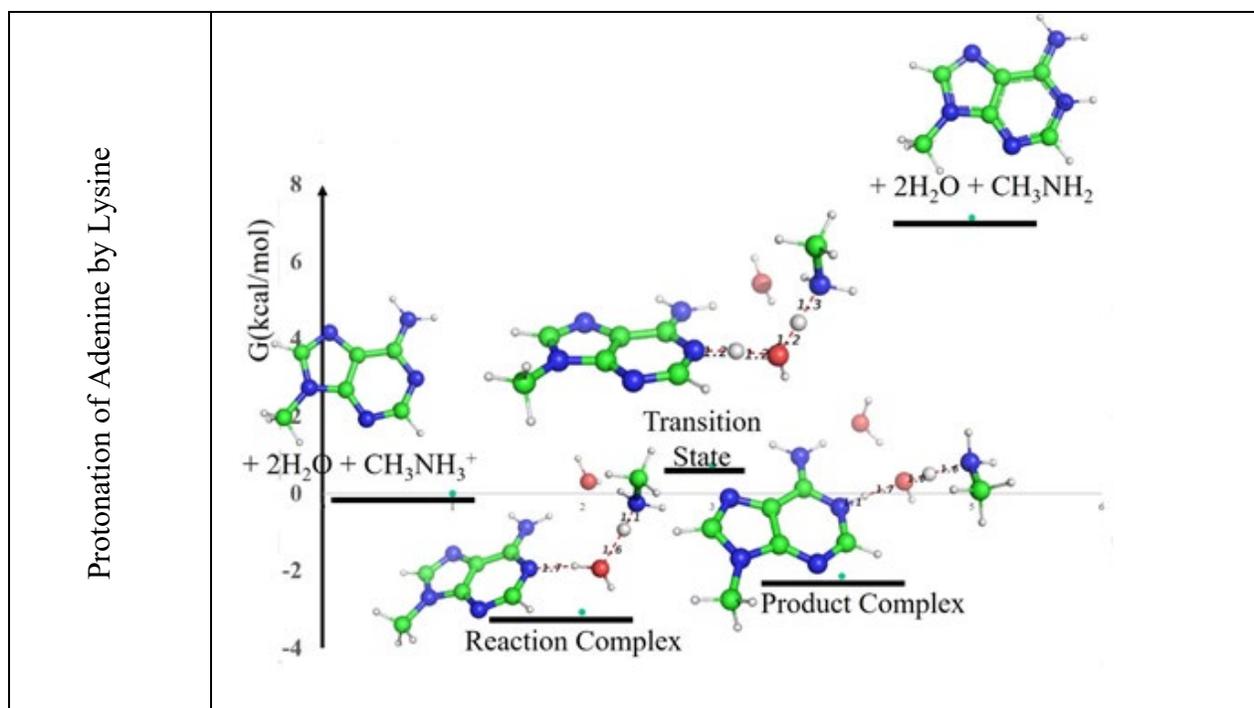
Optimized Coordinates for the transition state of Cytosine	Optimized Coordinates for the transition state of Adenine
C -2.532885 -2.367235 0.034533	H 1.457666 -0.921666 -0.774870
N -2.035067 -0.988214 0.033962	O 2.562946 -1.402431 -1.082206
C -2.862219 0.074498 0.228047	H 2.749191 -1.145341 -1.999634
H -3.915279 -0.164939 0.378419	H 3.470118 -1.046009 -0.443442
C -2.411938 1.362974 0.238243	N 4.484271 -0.556193 0.297350
H -3.090626 2.196837 0.399048	H 5.387252 -0.828167 -0.099292
C -1.016275 1.562772 0.030435	H 4.396342 0.462432 0.172963
N -0.440784 2.767049 0.022137	C 4.413838 -0.922237 1.724733
H 0.569114 2.821787 -0.116789	H 3.441697 -0.604163 2.128331
H -0.984706 3.607478 0.168632	H 4.492143 -2.014021 1.827762
N -0.211422 0.496901 -0.171086	H 5.213848 -0.450680 2.316906
C -0.654903 -0.804633 -0.169078	N 0.278579 -0.628949 -0.488384
O 0.088225 -1.766223 -0.331529	C -0.187061 0.630722 -0.260589
H -2.042208 -2.941956 0.831775	C -0.555910 -1.702678 -0.466609
H -3.615693 -2.348891 0.201171	C -1.576482 0.731853 -0.005354
H -2.311576 -2.845248 -0.929362	N 0.625650 1.688574 -0.286407
H 0.952281 0.779280 -0.370164	H -0.063395 -2.657406 -0.667632
O 2.093249 1.360886 -0.492672	N -1.854318 -1.702704 -0.236171
H 2.316239 1.375742 -1.438224	N -2.377827 1.822276 0.262895
H 2.986488 0.664074 0.095239	C -2.320406 -0.453588 -0.010574
N 3.763594 -0.122248 0.636745	H 1.642729 1.641367 -0.405096
H 4.723049 0.232765 0.596244	H 0.216806 2.594693 -0.089927
H 3.522613 -0.184144 1.630350	C -3.577260 1.295656 0.413979
C 3.653028 -1.447897 -0.017158	N -3.610433 -0.074744 0.261373
H 2.596231 -1.748855 -0.021056	H -4.484844 1.853411 0.638571
H 3.999516 -1.355113 -1.055306	C -4.769673 -0.952150 0.366410
H 4.259674 -2.203403 0.501189	H -4.625773 -1.674132 1.182450
	H -5.655703 -0.340618 0.573974
	H -4.916399 -1.498271 -0.575809
	O 3.493346 2.117651 -0.407572
	H 3.806659 2.413849 -1.278718
	H 3.637602 2.889655 0.165119

One molecule of water involves in proton transfer from Lysine to N1 of Adenine and another molecule of water probably gives some electrostatic stabilization by entering between amine group of Adenine and  $\text{CH}_3\text{NH}_3^+$ . Both system follows a “Trapeze Game” during protonation where proton is taken up by water from lysine concurrently with the transfer of proton of second water molecule to either N1 of Adenine or N3 of Cytosine. Transition state calculation

revealed that activation energy (Free Energy difference from Reaction complex to TS) for the protonation of Cytosine is lower than that of protonation of Adenine by 2kcal/mol (**Figure 5-5**). It is also found that protonation of cytosine is exothermic and protonation of Adenine is endothermic (**Figure 5-5**). It indicates that protonation of Cytosine is not only kinetically preferred but also thermodynamically preferred.

Bioinformatics study has revealed that total number of base pair with protonated cytosine is 389 whereas this number is 267 for the protonated Adenine. This observation can be explained by the difference in activation energy as activation energy for the protonation of Adenine is about twice that of Cytosine.

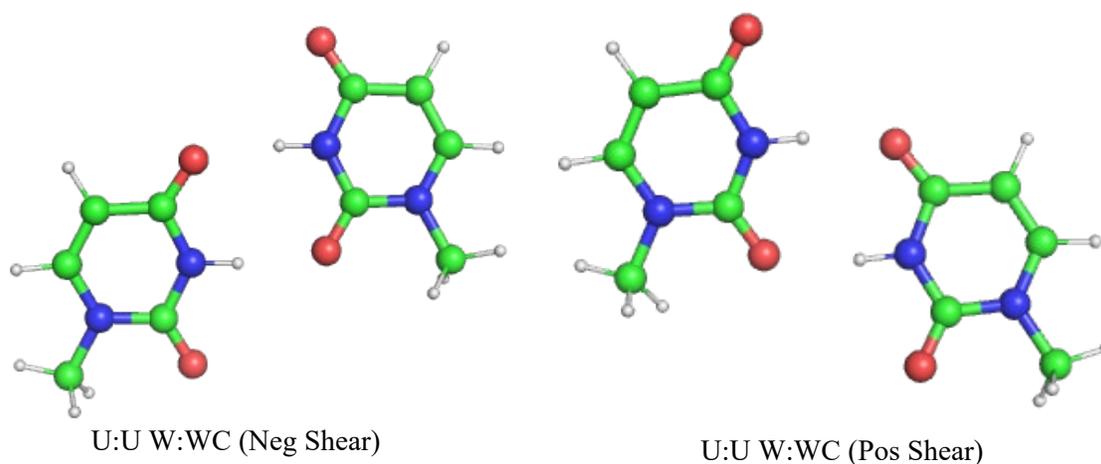




**Figure 5-5** Reaction coordinate for the protonation of cytosine and Adenine mediated by lysine and Water molecule.

### 5.3.1 Structure generation for Covid-19 RNA

As mentioned earlier A:C base pairing involving Watson-Crick edges of Ade and Cyt, is possible by two ways – in their neutral condition through C—H...N hydrogen bonding or protonated Adenine (**Figure 5-2**) while the protonated base pair is stronger. Earlier work also give the evidence of RNA double helix with A:C +:WC base pair<sup>210</sup>. Similarly the U:C base pairing is also possible in their neutral as well as protonated forms and U:U base pairing is possible in two symmetric modes (**Figure 5-6**). Thus, six distinct conformational states are possible for the 5'(GAUUU).5'(AUCCC) i.e., central region of the SL1 helix (**Figure 5-4**).

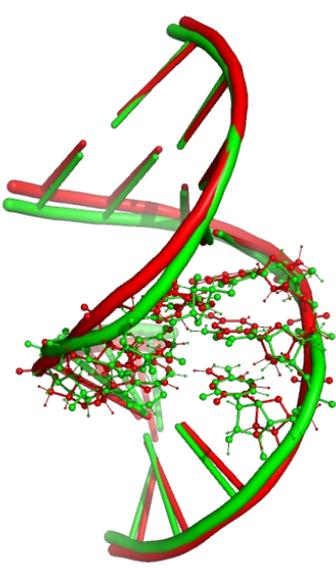


**Figure 5-6** Pictorial description of different U:U W:WC base pairs where major change in base pair is in their Shear motion.

The initial model structures based on mean values of different parameters obtained from RNABPDB database and from RNA-11 fiber model structure are in fact widely different with RMSD of 1Å or more (Table 3-3)<sup>218</sup>. The complete understanding of the most probable structure at physiological pH and low pH requires detail analysis of the double helical structures through scanning the parameter hyperspace of several base paired dinucleotide steps.

Further it is also important to resolve the confusion about whether the U:C base pair is protonated with protonated Cyt or the A:C base pair is protonated with protonated Ade. In addition to this, preference of Shear value for U:U W:WC base pair defines the most probable three dimensional configuration of SL1 helix of RNA of SARS corona virus.

**Table 5-3** Structural comparison between generated 3-dimensional structures with possible base pair. Line representation in the figure represents central base pairs as highlighted in the table.

		Tilt (°)	Roll (°)	Twist (°)	Shift (Å)	Slide (Å)	Rise (Å)	Buckle (°)	Open (°)	Propeller (°)	Stagger (Å)	Shear (Å)	Stretch (Å)	
	5'G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
	C:G W:WC	0.07	11.75	30.19	0.03	-1.86	3.08	5.91	0.99	-8.94	-0.11	0.09	2.86	
	G:C W:WC	-2.26	11.38	31.55	-0.15	-2	3.29	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
	A:C +:WC	0.22	10	30.13	0.12	-1.05	3.2	-4.28	10.09	-12.17	-0.08	-2.33	2.8	
	U:C W:WC	-1.93	13.52	35.92	0.13	-1.69	3.21	-1.93	-6.95	-18.51	-0.1	0	3.02	
	U:U W:WC	0.37	6.96	28.76	-0.38	-1.38	3.28	-0.32	-3.11	-14.73	-0.07	2.5	2.89	
	U:A W:WC	0.22	11.35	30.31	0.02	-1.63	3.18	2.45	3.75	-9.47	-0.05	-0.12	2.81	
	G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
	C:G W:WC	-0.32	12.89	35.69	0.27	-2.04	3.04	5.91	0.99	-8.94	-0.11	0.09	2.86	
	3'G:U W:WC	0	0	0	0	0	0	-1.47	-0.32	-9.05	-0.1	-2.25	2.81	
	5'G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
	C:G W:WC	0.07	11.75	30.19	0.03	-1.86	3.08	5.91	0.99	-8.94	-0.11	0.09	2.86	
	G:C W:WC	1.29	6.48	26.46	0.14	-1.43	3.21	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
	A:C w:WC	0.22	10	30.13	0.12	-1.05	3.2	-13.33	13.96	-10.31	-0.07	2.33	2.54	
	U:C W:+C	0.22	10	30.13	0.12	-1.05	3.2	-10.55	-7.17	-6.43	-0.24	-2.59	2.84	
	U:U W:WC	0.37	6.96	28.76	-0.38	-1.38	3.28	-0.32	-3.11	-14.73	-0.07	-2.5	2.89	
	U:A W:WC	0.22	11.35	30.31	0.02	-1.63	3.18	2.45	3.75	-9.47	-0.05	-0.12	2.81	
	G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86	
C:G W:WC	-0.32	12.89	35.69	0.27	-2.04	3.04	5.91	0.99	-8.94	-0.11	0.09	2.86		
3'G:U W:WC	0	0	0	0	0	0	-1.47	-0.32	-9.05	-0.1	-2.25	2.81		

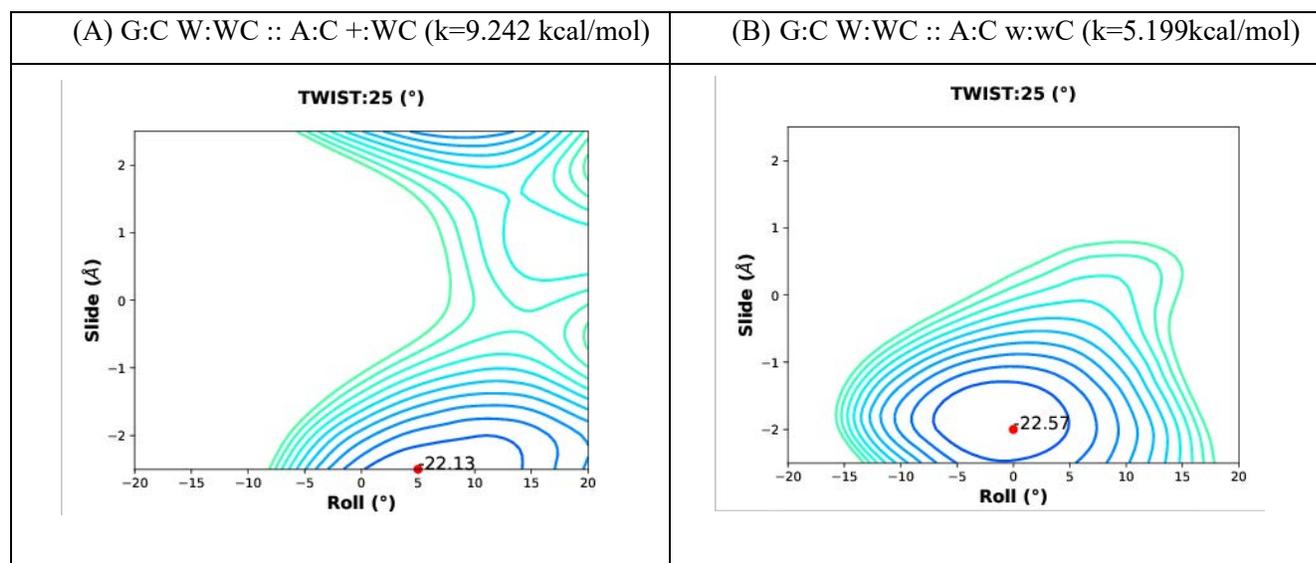
### 5.3.2 Hybrid Stacking Energy Scan

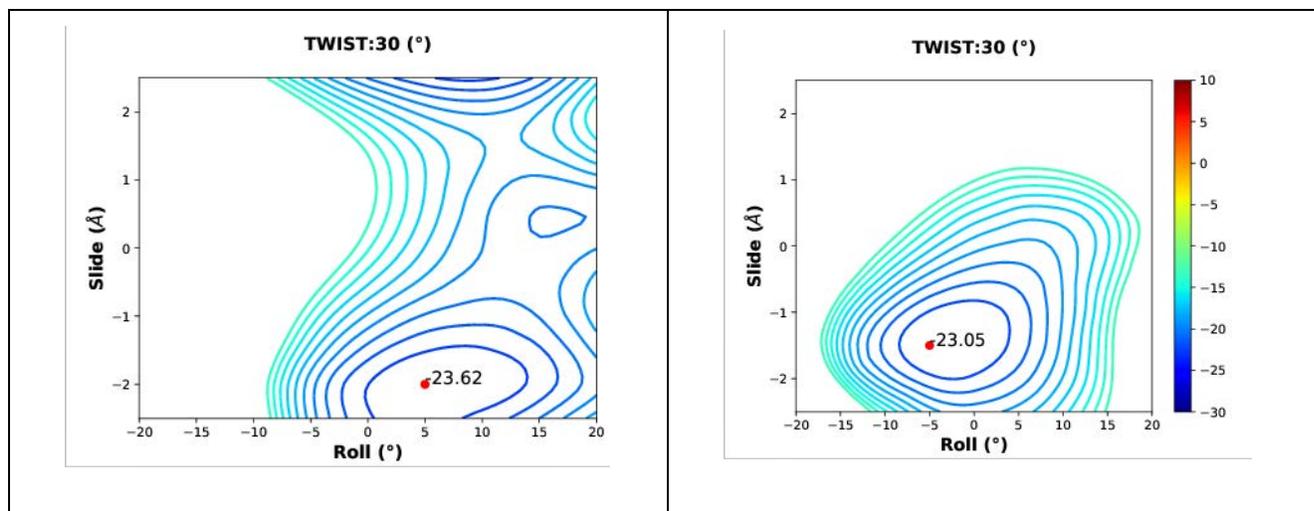
We have done stacking energy scan for the mentioned dinucleotide step as shown in **Figure 5-4** to predict the most probable model. Protonated Ade and Cyt have been considered in model-1 and model-2, respectively. As discussed in Chapter 3 and 4, we have added penalty value into DFT-D based stacking energy in order to mimic the effect of sugar-phosphate backbone. Here base pairs in dinucleotide steps are Watson-Crick type. Standard deviation for C1'...C1' distance for Watson-Crick base pair step is about 0.3 which estimates backbone force constant 9.242 kcal/mol which have been considered for the generation of hybrid stacking energy contours on Roll-Slide hyperspace for different Twist value. Whereas penalty value for dinucleotide step with one C-H...N mediated base pair has been considered 5.199 kcal/mol according to distribution of C1'...C1' virtual bond distances.

Hybrid stacking energy contours in Roll-Slide hyperspace for G:C W:WC :: A:C +:WC in model-1 indicates that best stacking energy zones span from 25° to 30° Twist with 5° Roll and -2.5 Å to -2Å Slide value where computed stacking energy is about -22 kcal/mol (**Figure 5-7**). Above estimated zone from stacking energy analysis is suitable for A-RNA like double helix. Whereas best stacking energy zones for G:C W:WC :: A:C w:wC dinucleotide step span such Roll-Slide hyperspace which does not depict A-RNA like structure. The lowest energy contours have negative Roll (**Figure 5-7B**). It may be mentioned that the earlier chapter indicates that RNA inherently prefers positive Roll and negative Slide as molecular dynamics simulation studies from B-RNA structure with small negative Roll and small positive Slide was shown to convert to A-RNA structure with large positive Roll and negative Slide. Thus we can conclude that stacking energy analysis indicates protonation of Ade base in A:C +:WC base pair.

The dinucleotide step A:C +:WC :: U:C W:WC estimates best stacking energy about -18kcal/mol which spans 30° to 35° Twist with positive Roll and negative Slide i.e. A-RNA like double helical configuration (**Figure 5-8**). When Ade is non-protonated and Cyt is protonated i.e. dinucleotide step A:C w:wC :: U:C W:+C, the best hybrid stacking energy zones span in positive Roll and positive Slide zone for all Twist value, which does not indicate A-RNA like configuration (**Figure 5-8B**). These results rule out the possibility of protonated cytosine in U:C W:+C in SL1 helix of RNA for SARS corona virus. So Model-1 is preferred over model-2.

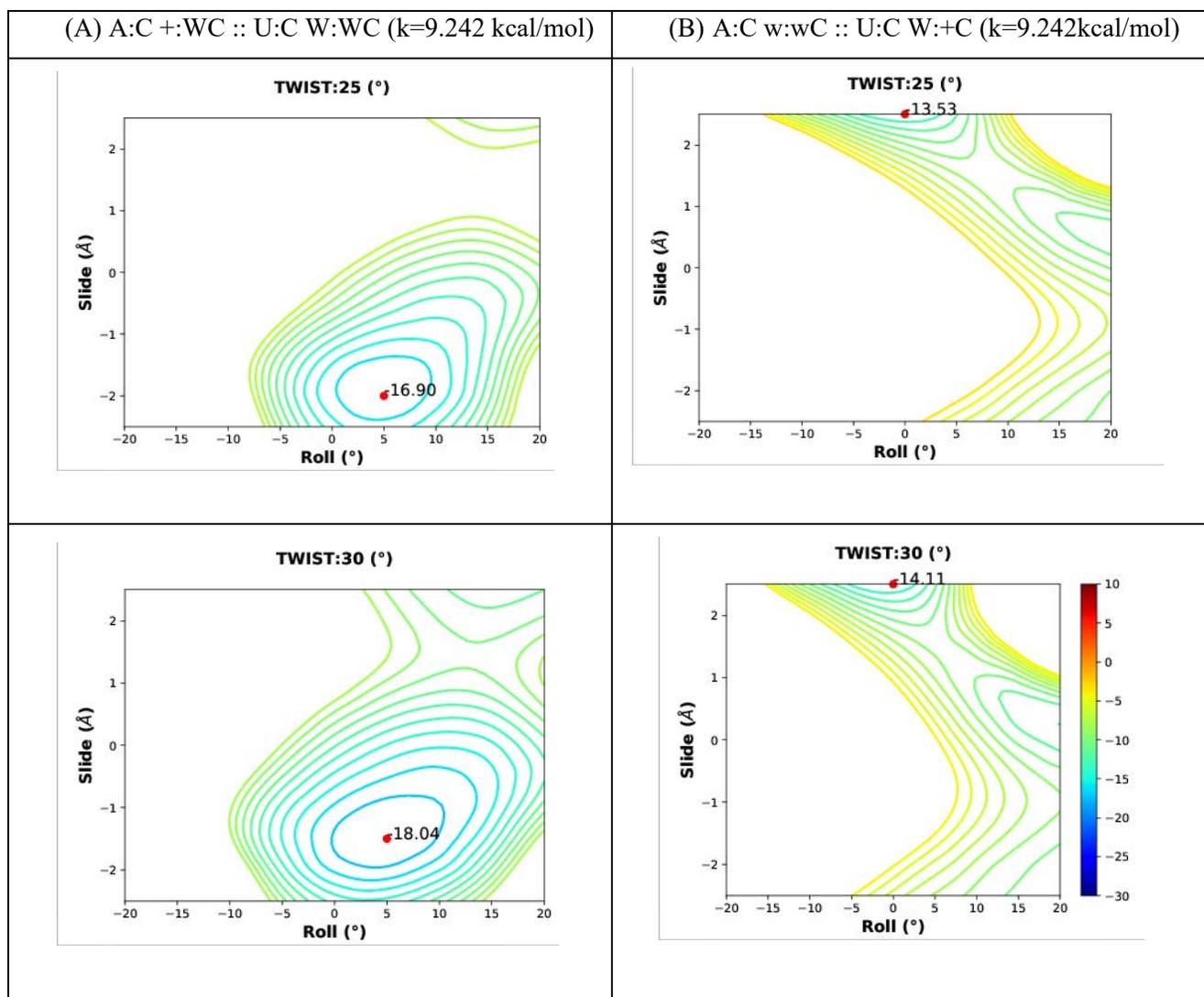
Still it is not clear about preference of Shear value in U:U W:WC base pair as it has bimodal distribution of Shear. Two hydrogen bonds are also possible with positive Shear for the U:U W:WC base pair where the Uracil of the second strand moves towards minor groove forming N3-H(2)...O2(1) and N3-H(1)...O2(2) hydrogen bonds. The base pair with positive Shear, if rotated by 180° through base pair short axis (or pseudo-dyad axis), superposes on the base pair with negative Shear.





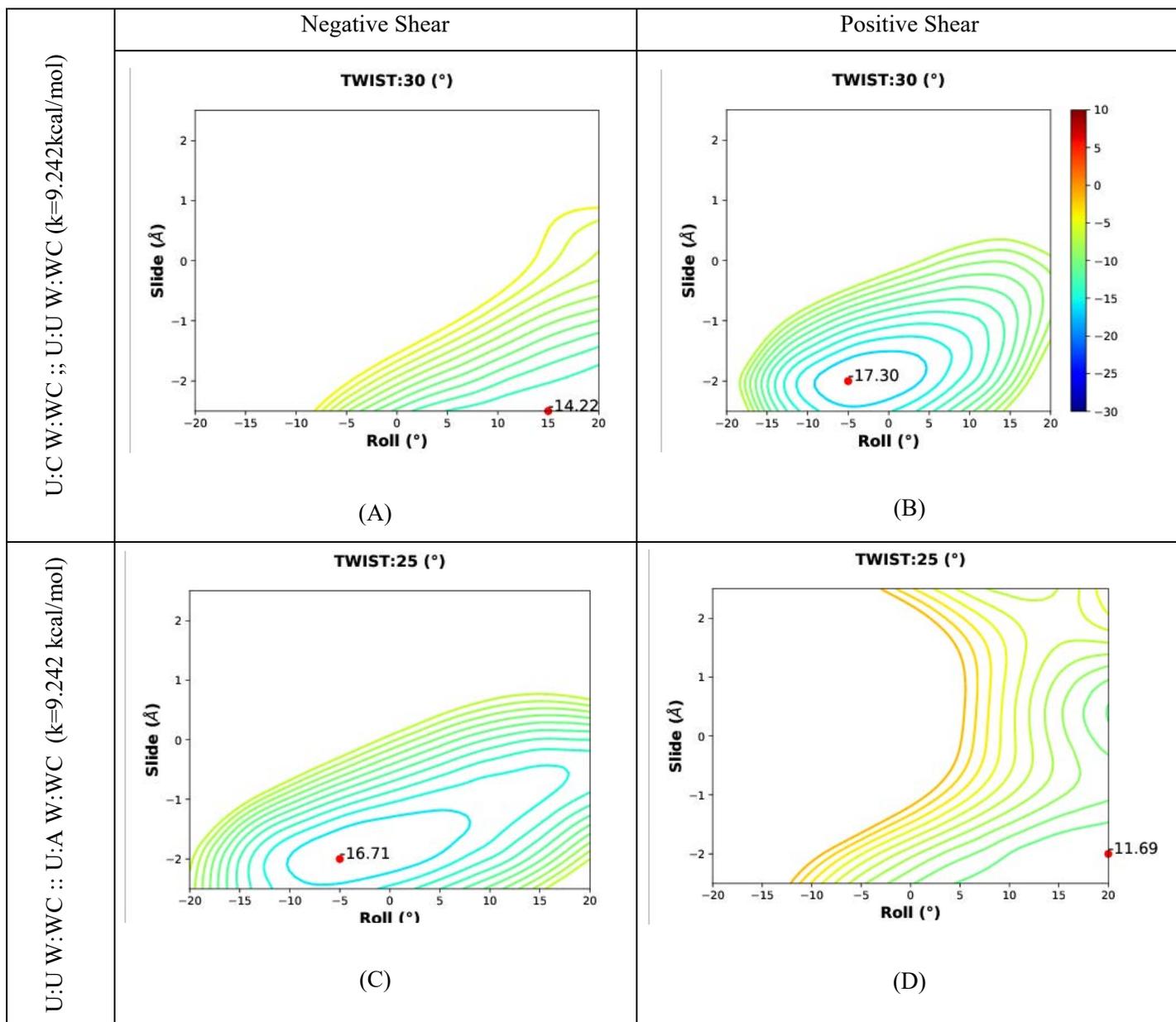
**Figure 5-7** Hybrid stacking isoenergy contours of (A) G:C W:WC :: A:C +:WC and (B) G:C W:WC :: A:C w:wC dinucleotide step sequence considering  $\omega$ B97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol.

Hence the two hydrogen bonding modes are symmetric (**Figure 5-6**). However, as indicated in the previous chapter, similar to A:A w:wC base pair, this symmetry breaks when the U:U W:WC base pair stacks on another base pair. Best hybrid stacking energy for U:C W:WC :: U:U W:WC with negative Shear of U:U W:WC base pair lies in positive Roll and negative Slide zone (**Figure 5-9**). Whereas best hybrid stacking energy for the same dinucleotide step with positive Shear of U:U W:WC base pair lies in negative Roll which does not depict A-RNA like configuration (**Figure 5-9B**). In U:U W:WC :: U:A W:WC dinucleotide step, Uracil (1) is at 5' end. Here best stacking energy zone prefers A-RNA like configuration when U:U W:WC has positive Shear (**Figure 5-9C** and **Figure 5-9D**). It points to preference of Shear for U:U W:WC base pair depends on strand direction and only one Shear gets preferred for suitable A-RNA like double helix formation. The symmetry breaking of A:A w:wC base pair was shown to be very prominent in stack with highly asymmetric sheared U:G base pair, while U:A W:WC base pair has Shear value around zero. Nevertheless, the stacking interaction appear to break rotational symmetry of the U:U W:WC base pair.



**Figure 5-8** Hybrid stacking isoenergy contours of (A) A:C +:WC :: U:C W:WC and (B) A:C w:wC :: U:C W:+C dinucleotide step sequence considering  $\omega$ B97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol.

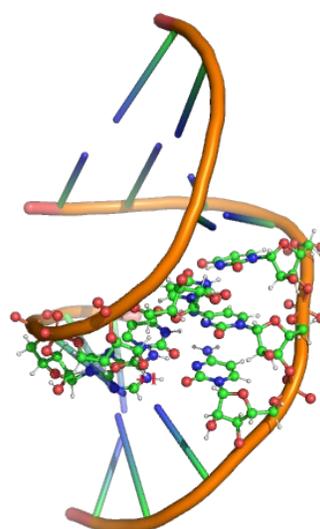
Quantum mechanical stacking energy scan and bioinformatics data thus help to generate molecular model of the SL-1 helix of RNA of SARS corona virus (Table 5-4).



**Figure 5-9** Hybrid stacking iso-energy contours of (A-B) U:C W:WC::U:U W:WC and (C-D) U:U W:WC::U:A W:WC dinucleotide step sequence considering  $\omega$ B97X-D/cc-pVDZ and coarse-grain energy penalty for different force constant values. Energy difference between two adjacent contour lines is 1kcal/mol.

**Table 5-4** Generated model structure and the structural parameters.

	Tilt (°)	Roll (°)	Twist (°)	Shift (Å)	Slide (Å)	Rise (Å)	Buckle (°)	Open (°)	Propeller (°)	Stagger (Å)	Shear (Å)	Stretch (Å)
5'G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86
C:G W:WC	0.07	11.75	30.19	0.03	-1.86	3.08	5.91	0.99	-8.94	-0.11	0.09	2.86
G:C W:WC	-2.26	11.38	31.55	-0.15	-2	3.29	-5.91	0.99	-8.94	-0.11	-0.09	2.86
A:C +:WC	0.22	10	30.13	0.12	-1.05	3.2	-4.28	10.09	-12.17	-0.08	-2.33	2.8
U:C W:WC	-1.93	13.52	35.92	0.13	-1.69	3.21	-1.93	-6.95	-18.51	-0.1	0.02	3.02
U:U W:WC	0.37	6.96	28.76	-0.38	-1.38	3.28	-0.32	-3.11	-14.73	-0.07	2.5	2.89
U:A W:WC	0.22	11.35	30.31	0.02	-1.63	3.18	2.45	3.75	-9.47	-0.05	-0.12	2.81
G:C W:WC	-0.03	3.49	31.7	0	-1.57	3.35	-5.91	0.99	-8.94	-0.11	-0.09	2.86
C:G W:WC	-0.32	12.89	35.69	0.27	-2.04	3.04	5.91	0.99	-8.94	-0.11	0.09	2.86
3'G:U W:WC	0	0	0	0	0	0	-1.47	-0.32	-9.05	-0.1	-2.25	2.81



## 5.4 Conclusion

Stacking energy analysis of several base pair step in SL1 helix of RNA of SARS corona virus have pointed out model-1 as preferred RNA double helix where Adenine is protonated in A:C +:WC base pair. Though activation energy of protonated Cytosine is less than that of Adenine, protonated A:C +:WC base pair is preferred in SL1 helix. In this context, protonation of Adenine requires either low pH or involvement of lysine. Latter is biologically important as lysine remains charged (side chain  $-(CH_2)_4-NH_3^+$ ) at physiological condition. In addition to this, promiscuous nature of U:U W:WC has been perceived as depending on the strand direction Shear value of U:U base pair gets allotted for suitable A-RNA like structure. This model might be confirmed by all

atom molecular dynamics simulation to understand how it is stabilized at physiological condition in presence of sugar-phosphate backbone, water/ion and physiological temperature. It would now be possible to design drug against this virus so as to target the SL1 helix structure, whose rational model can be generated from the base pair and dinucleotide step parameters.

## **Chapter- 6 Conclusion and Scope of Future**

The complete work presented here is based on understanding non-canonical base pair from the lens of quantum mechanical calculation, MD simulation, and bioinformatics study.

### **6.1 Conclusion and Scope of future work (Chapter 2)**

All canonical and non-canonical base pair interaction energy has been calculated considering BSSE correction. The energy calculations were done for best representative structures of each type. However, often it was required to find a best representative from few examples (five or more observed crystallographic structures). This may lead to ambiguity. Geometry optimization of each base pair might improve the understanding. However, previous studies indicated that often a base pair changes its orientation, especially when the base pairing involves 2'-OH group mediated hydrogen bonding. Geometry optimizations considering BSSE corrected energies might improve the optimized geometries of the base pairs and improve correlation between frequencies of observations and interaction energy.

- How 2'-OH involves in base pairing and it improves the stability – have been highlighted from this database. The contraction of C—H bond length on hydrogen bond formation has been noted, which correlates with Blue-Shifting of C—H bond stretching frequency.
- Stacking energy is solely dependent on base pair and base pair step parameters. Multidimensional correlation analysis may reveal the proper relation between parameters and stacking energy. This work would require a substantial computational cost.

### **6.2 Conclusion and Scope of future work (Chapter 3)**

- This work shows how stacking energy scan finds the suitable configuration of dinucleotide step with non-canonical base pair.
- Stacking energy data can be useful for force field parameterization in coarse grain simulation.

### **6.3 Conclusion and Scope of future work (Chapter 4)**

- Promiscuous nature of A:A base pair gets highlighted in this chapter. The Sheared G:U base pairs induce force and A:A base pair moves depending on the resultant force's direction. It results in the adoption of unique Shear sign of A:A base pair.
- This study also points out that a similar type of behaviors might be observed for other base pairs where the bimodality of Shear is present.
- QM/MM study can reveal the activation energy for the transition between two Shear configurations.

### **6.4 Conclusion and Scope of future work (Chapter 5)**

- We have predicted the SL1 helix of RNA of SARS coronavirus, which causes deadly global pandemic nowadays considering quantum chemical calculation and bioinformatics study.
- SL1 helix contains protonated Adenine in A:C<sup>+</sup>:WC base pair.
- Complete folding of SL1 helix is incomplete without the protonation of Adenine.
- The thermodynamic integration protocol can reveal the structural change of the SL1 helix during protonation.

## **Chapter- 7 References**

- (1) Avery, O. T.; Macleod, C. M.; McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from *Pneumococcus* Type Iii. *J. Exp. Med.* **1944**, *79* (2), 137–158.
- (2) HERSHEY, A. D.; CHASE, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J. Gen. Physiol.* **1952**, *36* (1), 39–56.
- (3) CHARGAFF, E.; ZAMENHOF, S.; GREEN, C. Human Desoxypentose Nucleic Acid: Composition of Human Desoxypentose Nucleic Acid. *Nature* **1950**, *165* (4202), 756–757.
- (4) WATSON, J. D.; CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171* (4356), 737–738.
- (5) Meselson, M.; Stahl, F. W.; Vinograd, J. EQUILIBRIUM SEDIMENTATION OF MACROMOLECULES IN DENSITY GRADIENTS. *Proc. Natl. Acad. Sci.* **1957**, *43* (7), 581–588.
- (6) Hoagland, M. B.; Zamecnik, P. C.; Stephenson, M. L. Intermediate Reactions in Protein Biosynthesis. *BBA - Biochim. Biophys. Acta* **1957**, *24* (C), 215–216.
- (7) Gros, F.; Hiatt, H.; Gilbert, W.; Kurland, C. G.; Risebrough, R. W.; Watson, J. D. Unstable Ribonucleic Acid Revealed by Pulse Labelling of *Escherichia Coli*. *Nature* **1961**, *190* (4776), 581–585.
- (8) Brenner, S.; Jacob, F.; Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **1961**, *190* (4776), 576–581.
- (9) NIRENBERG, M. W.; MATTHAEI, J. H. The Dependence of Cell-Free Protein Synthesis

- in *E. Coli* upon Naturally Occurring or Synthetic Polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47* (2), 1588–1602.
- (10) Noller, H. F. Ribosomal RNA and Translation. *Annu. Rev. Biochem.* **1991**, *60* (1), 191–227.
- (11) Watson, J. D. Involvement of RNA in the Synthesis of Proteins. *Science* (80-. ). **1963**, *140* (3562), 17–26.
- (12) Ban, N.; Freeborn, B.; Nissen, P.; Penczek, P.; Grassucci, R. A.; Sweet, R.; Frank, J.; Moore, P. B.; Steitz, T. A. A 9 Å Resolution X-Ray Crystallographic Map of the Large Ribosomal Subunit. *Cell* **1998**, *93* (7), 1105–1115.
- (13) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vornheln, C.; Hartsch, T.; Ramakrishnan, V. Structure of the 30S Ribosomal Subunit. *Nature* **2000**, *407* (6802), 327–339.
- (14) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* (80-. ). **2000**, *289* (5481), 905–920.
- (15) Rich, A. A HYBRID HELIX CONTAINING BOTH DEOXYRIBOSE AND RIBOSE POLYNUCLEOTIDES AND ITS RELATION TO THE TRANSFER OF INFORMATION BETWEEN THE NUCLEIC ACIDS. *Proc. Natl. Acad. Sci.* **1960**, *46* (8), 1044–1053.
- (16) Rosenberg, J. M.; Seeman, N. C.; Kim, J. J. P.; Suddath, F. L.; Nicholas, H. B.; Rich, A. Double Helix at Atomic Resolution. *Nature* **1973**, *243* (5403), 150–154.

- (17) DALY, M. M.; ALLFREY, V. G.; MIRSKY, A. E. Purine and Pyrimidine Contents of Some Desoxyribose Nucleic Acids. *J. Gen. Physiol.* **1950**, *33* (5), 497–510.
- (18) Harcourt, E. M.; Kietrys, A. M.; Kool, E. T. Chemical and Structural Effects of Base Modifications in Messenger RNA. *Nature*. Nature Publishing Group January 18, 2017, pp 339–346.
- (19) Manning, G. S. Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. *J. Chem. Phys.* **1969**, *51* (3), 924.
- (20) Symbols for Specifying the Conformation of Polysaccharide Chains. Recommendations 1981. *Eur. J. Biochem.* **1983**, *131* (1), 5–7.
- (21) Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. B-DNA Twisting Correlates with Base-Pair Morphology. *J. Mol. Biol.* **1995**, *247* (1), 34–48.
- (22) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (19), 11163–11168.
- (23) Lavery, R. *Multiple Aspects of DNA and RNA from Biophysics to Bioinformatics*; 2005.
- (24) Duarte, C. M.; Pyle, A. M. Stepping through an RNA Structure: A Novel Approach to Conformational Analysis. *J. Mol. Biol.* **1998**, *284* (5), 1465–1478.
- (25) Malathi, R.; Yathindra, N. Backbone Conformation in Nucleic Acids: An Analysis of Local Helicity through Heminucleotide Scheme and a Proposal for a Unified Conformational Plot. *J. Biomol. Struct. Dyn.* **1985**, *3* (1), 127–144.
- (26) Wilson, H. R.; Rahman, A. Nucleoside Conformation and Non-Bonded Interactions. *J.*

- Mol. Biol.* **1971**, *56* (1), 129–142.
- (27) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. a New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94* (23), 8205–8212.
- (28) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*; Interdisciplinary Applied Mathematics; Springer New York: New York, NY, 2010; Vol. 21.
- (29) Hoogsteen, K. The Structure of Crystals Containing a Hydrogen-Bonded Complex of 1-Methylthymine and 9-Methyladenine. *Acta Crystallogr.* **1959**, *12* (10), 822–823.
- (30) Courtois, Y.; Fromageot, P.; Guschlbauer, W. Protonated Polynucleotide Structures: 3. An Optical Rotatory Dispersion Study of the Protonation of DNA. *Eur. J. Biochem.* **1968**, *6* (4), 493–501.
- (31) Patel, D. J.; Tonelli, A. E. Assignment of the Proton Nmr Chemical Shifts of the T□N3H and G□N1H Proton Resonances in Isolated AT and GC Watson-Crick Base Pairs in Double-stranded Deoxy Oligonucleotides in Aqueous Solution. *Biopolymers* **1974**, *13* (10), 1943–1964.
- (32) Seeman, N. C.; Rosenberg, J. M.; Suddath, F. L.; Kim, J. J.; Rich, A. RNA Double-Helical Fragments at Atomic Resolution. I. The Crystal and Molecular Structure of Sodium Adenylyl-3',5'-Uridine Hexahydrate. *J. Mol. Biol.* **1976**, *104* (1), 109–144.
- (33) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA Dodecamer: Conformation and Dynamics. *Proc. Natl. Acad. Sci.*

- U. S. A.* **1981**, 78 (4), 2179–2183.
- (34) Wang, A. H. J.; Fujii, S.; Van Boom, J. H.; Van Der Marel, G. A.; Van Boeckel, S. A. A.; Rich, A. Molecular Structure of r(GCG)<sub>n</sub>d(TATACGC): A DNA-RNA Hybrid Helix Joined to Double Helical DNA. *Nature* **1982**, 299 (5884), 601–604.
- (35) Heinemann, U.; Alings, C. Crystallographic Study of One Turn of G/C-Rich B-DNA. *J. Mol. Biol.* **1989**, 210 (2), 369–381.
- (36) Dock-Bregeon, A. C.; Chevrier, B.; Podjarny, A.; Johnson, J.; de Bear, J. S.; Gough, G. R.; Gilham, P. T.; Moras, D. Crystallographic Structure of an RNA Helix: [U(UA)<sub>6</sub>A]<sub>2</sub>. *J. Mol. Biol.* **1989**, 209 (3), 459–474.
- (37) Coimbatore Narayanan, B.; Westbrook, J.; Ghosh, S.; Petrov, A. I.; Sweeney, B.; Zirbel, C. L.; Leontis, N. B.; Berman, H. M. The Nucleic Acid Database: New Features and Capabilities. *Nucleic Acids Res.* **2014**, 42 (D1), D114–D122.
- (38) Saenger, W. *Principles of Nucleic Acid Structure*; Springer Advanced Texts in Chemistry; Springer New York: New York, NY, 1984.
- (39) Dickerson, R. E. Definitions and Nomenclature of Nucleic Acid Structure Components. *Nucleic Acids Res.* **1989**, 17 (5), 1797–1803.
- (40) Blanchet, C.; Pasi, M.; Zakrzewska, K.; Lavery, R. CURVES+ Web Server for Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. *Nucleic Acids Res.* **2011**, 39 (suppl), W68–W73.
- (41) Lu, X. J.; Olson, W. K. 3DNA: A Versatile, Integrated Software System for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic-Acid Structures. *Nat. Protoc.*

- 2008**, 3 (7), 1213–1227.
- (42) NUPARM-Plus-A Program for analyzing sequence dependent variations in nucleic acids.
- (43) Leontis, N. B.; Westhof, E. Geometric Nomenclature and Classification of RNA Base Pairs. *RNA* **2001**, 7 (4), 499–512.
- (44) Desiraju, G. R. (Gautam R. .; Steiner, T. *The Weak Hydrogen Bond : In Structural Chemistry and Biology*; Oxford University Press, 2001.
- (45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28 (1), 235–242.
- (46) Lu, X. J.; Olson, W. K. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, 31 (17), 5108–5121.
- (47) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. The Nucleic Acid Database. A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.* **1992**, 63 (3), 751–759.
- (48) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *J. Math. Biol.* **2008**, 56 (1–2), 215–252.
- (49) Lemieux, S.; Major, F. RNA Canonical and Non-Canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research*. Oxford

- University Press October 1, 2002, pp 4250–4263.
- (50) Bayly, C. I.; Merz, K. M.; Ferguson, D. M.; Cornell, W. D.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; Cieplak, P.; Gould, I. R.; Spellmeyer, D. C. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.
- (51) Das, J.; Mukherjee, S.; Mitra, A.; Bhattacharyya, D. Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Biomol. Struct. Dyn.* **2006**, *24* (2).
- (52) Lu, X.-J.; Shakked, Z.; Olson, W. K. A-Form Conformational Motifs in Ligand-Bound DNA Structures. *J. Mol. Biol.* **2000**, *300* (4), 819–840.
- (53) 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.; et al. A Standard Reference Frame for the Description of Nucleic Acid Base-Pair Geometry. *J. Mol. Biol.* **2001**, *313* (1), 229–237.
- (54) Leontis, N. B. The Non-Watson-Crick Base Pairs and Their Associated Isostericity Matrices. *Nucleic Acids Res.* **2002**, *30* (16), 3497–3531.
- (55) Nasalean, L.; Stombaugh, J.; Zirbel, C. L.; Leontis, N. B. RNA 3D Structural Motifs: Definition, Identification, Annotation, and Database Searching. In *Non-Protein Coding RNAs*; Springer Berlin Heidelberg, 2008; pp 1–26.
- (56) Stombaugh, J.; Zirbel, C. L.; Westhof, E.; Leontis, N. B. Frequency and Isostericity of RNA Base Pairs. *Nucleic Acids Res.* **2009**, *37* (7), 2294–2312.

- (57) Popena, M.; Szachniuk, M.; Blazewicz, M.; Wasik, S.; Burke, E. K.; Blazewicz, J.; Adamiak, R. W. RNA FRABASE 2.0: An Advanced Web-Accessible Database with the Capacity to Search the Three-Dimensional Fragments within RNA Structures. *BMC Bioinformatics* **2010**, *11* (1), 231.
- (58) Hamdani, H. Y.; Appasamy, S. D.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. NASSAM: A Server to Search for and Annotate Tertiary Interactions and Motifs in Three-Dimensional Structures of Complex RNA Molecules. *Nucleic Acids Res.* **2012**, *40* (W1).
- (59) RNA 3D Motif Atlas.
- (60) Šponer, J. E.; Špačková, N.; Kulhánek, P.; Leszczynski, J.; Šponer, J. Non-Watson-Crick Base Pairing in RNA. Quantum Chemical Analysis of the Cis Watson-Crick/Sugar Edge Base Pair Family. *J. Phys. Chem. A* **2005**, *109* (10), 2292–2301.
- (61) Kelly, R. E. A.; Lee, Y. J.; Kantorovich, L. N. Homopairing Possibilities of the DNA Base Adenine. *J. Phys. Chem. B* **2005**, *109* (24), 11933–11939.
- (62) Oliva, R.; Cavallo, L.; Tramontano, A. Accurate Energies of Hydrogen Bonded Nucleic Acid Base Pairs and Triplets in tRNA Tertiary Interactions. *Nucleic Acids Res.* **2006**.
- (63) Bhattacharyya, D.; Koripella, S. C.; Mitra, A.; Rajendran, V. B.; Sinha, B. Theoretical Analysis of Noncanonical Base Pairing Interactions in RNA Molecules. In *Journal of Biosciences*; Springer, 2007; Vol. 32, pp 809–825.
- (64) Roy, A.; Panigrahi, S.; Bhattacharyya, M.; Bhattacharyya, D. Structure, Stability, and Dynamics of Canonical and Noncanonical Base Pairs: Quantum Chemical Studies. *J. Phys. Chem. B* **2008**, *112* (12), 3786–3796.

- (65) Sharma, P.; Šponer, J. E.; Šponer, J.; Sharma, S.; Bhattacharyya, D.; Mitra, A. On the Role of the Eis Hoogsteen: Sugar-Edge Family of Base Pairs in Platforms and Triplets-Quantum Chemical Insights into RNA Structural Biology. *J. Phys. Chem. B* **2010**, *114* (9), 3307–3320.
- (66) Brovarets, O. O.; Yurenko, Y. P.; Hovorun, D. M. Intermolecular CH $\cdots$ O/N H-Bonds in the Biologically Important Pairs of Natural Nucleobases: A Thorough Quantum-Chemical Study. *J. Biomol. Struct. Dyn.* **2014**, *32* (6), 993–1022.
- (67) Marino, T. DFT Investigation of the Mismatched Base Pairs (T-Hg-T) $_3$ , (U-Hg-U) $_3$ , d(T-Hg-T) $_2$ , and d(U-Hg-U) $_2$ . *J. Mol. Model.* **2014**, *20* (6), 2303.
- (68) Roy, A.; Panigrahi, S.; Bhattacharyya, M.; Bhattacharyya, D. Structure, Stability, and Dynamics of Canonical and Noncanonical Base Pairs: Quantum Chemical Studies. *J. Phys. Chem. B* **2008**, *112* (12).
- (69) Šponer, J. E.; Leszczynski, J.; Sychrovský, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109* (39), 18680–18689.
- (70) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.
- (71) Grimme, S. Do Special Noncovalent  $\pi$ - $\pi$  Stacking Interactions Really Exist? *Angew. Chemie Int. Ed.* **2008**, *47* (18), 3430–3434.

- (72) Hunter, C. A. Sequence-Dependent DNA Structure. *BioEssays* **1996**, *18* (2), 157–162.
- (73) Calladine, C. R. Mechanics of Sequence-Dependent Stacking of Bases in B-DNA. *Journal of Molecular Biology*. 1982, pp 343–352.
- (74) El Hassan, M. A.; Calladine, C. R. Propeller-Twisting of Base-Pairs and the Conformational Mobility of Dinucleotide Steps in DNA. *J. Mol. Biol.* **1996**, *259* (1), 95–103.
- (75) Mondal, M.; Mukherjee, S.; Halder, S.; Bhattacharyya, D. Stacking Geometry for Non-Canonical G:U Wobble Base Pair Containing Dinucleotide Sequences in RNA: Dispersion-Corrected DFT-D Study. *Biopolymers* **2015**, *103* (6), 328–338.
- (76) Maiti, S.; Bhattacharyya, D. Stacking Interactions Involving Non-Watson–Crick Basepairs: Dispersion Corrected Density Functional Theory Studies. *Phys. Chem. Chem. Phys.* **2017**, *19* (42), 28718–28730.
- (77) Mukherjee, S.; Kailasam, S.; Bansal, M.; Bhattacharyya, D. Stacking Interactions in RNA and DNA: Roll-Slide Energy Hyperspace for Ten Unique Dinucleotide Steps. *Biopolymers* **2015**, *103* (3), 134–147.
- (78) Maiti, S.; Mukherjee, D.; Roy, P.; Chakrabarti, J.; Bhattacharyya, D. Stacking Geometry between Two Sheared Watson-Crick Basepairs: Computational Chemistry and Bioinformatics Based Prediction. *Biochim. Biophys. Acta - Gen. Subj.* **2020**, *1864* (7), 129600.
- (79) Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98* (2), 1372–1377.

- (80) Šponer, J.; Shukla, M. K.; Wang, J.; Leszczynski, J. Computational Modeling of DNA and RNA Fragments. In *Handbook of Computational Chemistry*; Springer International Publishing, 2017; pp 1803–1826.
- (81) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110* (9), 5023–5063.
- (82) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10* (44), 6615–6620.
- (83) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.
- (84) Hobza, P.; Selzle, H. L.; Schlag, E. W. Potential Energy Surface for the Benzene Dimer. Results of Ab Initia CCSD(T) Calculations Show Two Nearly Isoenergetic Structures: T-Shaped and Parallel-Displaced. *J. Phys. Chem.* **1996**, *100* (48), 18790–18794.
- (85) Jaffe, R. L.; Smith, G. D. A Quantum Chemistry Study of Benzene Dimer. *J. Chem. Phys.* **1996**, *105* (7), 2780–2788.
- (86) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19* (4), 553–566.

- (87) Malenov, D. P.; Ninković, D. B.; Zarić, S. D. Stacking of Metal Chelates with Benzene: Can Dispersion-Corrected DFT Be Used to Calculate Organic-Inorganic Stacking? *ChemPhysChem* **2015**, *16* (4), 761–768.
- (88) Das, J.; Mukherjee, S.; Mitra, A.; Bhattacharyya, D. Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Biomol. Struct. Dyn.* **2006**, *24* (2), 149–161.
- (89) Bhattacharyya, D.; Halder, S.; Basu, S.; Mukherjee, D.; Kumar, P.; Bansal, M. RNAHelix: Computational Modeling of Nucleic Acid Structures with Watson-Crick and Non-Canonical Base Pairs. *J. Comput. Aided. Mol. Des.* **2017**, *31* (2), 219–235.
- (90) Andronescu, M.; Bereg, V.; Hoos, H. H.; Condon, A. RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. *BMC Bioinformatics* **2008**, *9* (1), 340.
- (91) Popena, M.; Blazewicz, M.; Szachniuk, M.; Adamiak, R. W. RNA FRABASE Version 1.0: An Engine with a Database to Search for the Three-Dimensional Fragments within RNA Structures. *Nucleic Acids Res.* **2008**, *36* (Database issue), D386-91.
- (92) Popena, M.; Szachniuk, M.; Antczak, M.; Purzycka, K. J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R. W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40* (14), e112.
- (93) Petrov, A. I.; Zirbel, C. L.; Leontis, N. B. Automated Classification of RNA 3D Motifs and the RNA 3D Motif Atlas. *RNA* **2013**, *19* (10), 1327–1340.
- (94) Klosterman, P. S.; Tamura, M.; Holbrook, S. R.; Brenner, S. E. SCOR: A Structural

- Classification of RNA Database. *Nucleic Acids Res.* **2002**, *30*, 392–394.
- (95) Chojnowski, G.; Waleń, T.; Bujnicki, J. M. RNA Bricks - A Database of RNA 3D Motifs and Their Interactions. *Nucleic Acids Res.* **2014**, *42* (D1), D123.
- (96) Bhattacharya, S.; Mittal, S.; Panigrahi, S.; Sharma, P.; S P, P.; Paul, R.; Halder, S.; Halder, A.; Bhattacharyya, D.; Mitra, A. RNABP COGEST: A Resource for Investigating Functional RNAs. *Database (Oxford)*. **2015**, *2015* (0), bav011–bav011.
- (97) Westhof, E.; Fritsch, V. RNA Folding: Beyond Watson-Crick Pairs. *Structure*. March 2000, pp R55–R65.
- (98) Nagaswamy, U.; Larios-Sanz, M.; Hury, J.; Collins, S.; Zhang, Z.; Zhao, Q.; Fox, G. E. NCIR: A Database of Non-Canonical Interactions in Known RNA Structures. *Nucleic Acids Res.* **2002**, *30* (1), 395–397.
- (99) Brovarets, O. O.; Yurenko, Y. P.; Hovorun, D. M. Intermolecular CH···O/N H-Bonds in the Biologically Important Pairs of Natural Nucleobases: A Thorough Quantum-Chemical Study. *J. Biomol. Struct. Dyn.* **2014**, *32* (6), 993–1022.
- (100) Šponer, J. E.; Leszczynski, J.; Sychrovský, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109* (39), 18680–18689.
- (101) Sharma, P.; Mitra, A.; Sharma, S.; Singh, H.; Bhattacharyya, D. Quantum Chemical Studies of Structures and Binding in Noncanonical Rna Base Pairs: The Trans Watson-Crick:Watson-Crick Family. *J. Biomol. Struct. Dyn.* **2008**, *25* (6), 709–732.
- (102) Mládek, A.; Sharma, P.; Mitra, A.; Bhattacharyya, D.; Šponer, J.; Šponer, J. E. Trans

- Hoogsteen/Sugar Edge Base Pairing in RNA. Structures, Energies, and Stabilities from Quantum Chemical Calculations. *J. Phys. Chem. B* **2009**, *113* (6), 1743–1755.
- (103) Chawla, M.; Sharma, P.; Halder, S.; Bhattacharyya, D.; Mitra, A. Protonation of Base Pairs in RNA: Context Analysis and Quantum Chemical Investigations of Their Geometries and Stabilities. *J. Phys. Chem. B* **2011**, *115* (6), 1469–1484.
- (104) Lu, X. J.; Bussemaker, H. J.; Olson, W. K. DSSR: An Integrated Software Tool for Dissecting the Spatial Structure of RNA. *Nucleic Acids Res.* **2015**, *43* (21), gkv716.
- (105) Walé, T.; Chojnowski, G.; Gierski, P.; Bujnicki, J. M. ClaRNA: A Classifier of Contacts in RNA 3D Structures Based on a Comparative Analysis of Various Classification Schemes. *Nucleic Acids Res.* **2014**, *42* (19), 151.
- (106) Gendron, P.; Lemieux, S.; Major, F. Quantitative Analysis of Nucleic Acid Three-Dimensional Structures. *J. Mol. Biol.* **2001**, *308* (5), 919–936.
- (107) Olson, W. K.; Esguerra, M.; Xin, Y.; Lu, X. J. New Information Content in RNA Base Pairing Deduced from Quantitative Analysis of High-Resolution Structures. *Methods*. March 2009, pp 177–186.
- (108) Kailasam, S.; Bhattacharyya, D.; Bansal, M. Sequence Dependent Variations in RNA Duplex Are Related to Non-Canonical Hydrogen Bond Interactions in Dinucleotide Steps. *BMC Res. Notes* **2014**, *7* (1), 83.
- (109) Lu, X.-J. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, *31* (17), 5108–5121.

- (110) Olson, W. K.; Li, S.; Kaukonen, T.; Colasanti, A. V.; Xin, Y.; Lu, X.-J. Effects of Noncanonical Base Pairing on RNA Folding: Structural Context and Spatial Arrangements of G·A Pairs. *Biochemistry* **2019**, acs.biochem.9b00122.
- (111) Mukherjee, S.; Bansal, M.; Bhattacharyya, D. Conformational Specificity of Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Comput. Aided. Mol. Des.* **2006**, *20* (10–11), 629–645.
- (112) Lu, X. J.; Babcock, M. S.; Olson, W. K. Overview of Nucleic Acid Analysis Programs. *J. Biomol. Struct. Dyn.* **1999**, *16* (4), 833–843.
- (113) Lu, X. J.; Olson, W. K. Resolving the Discrepancies among Nucleic Acid Conformational Analyses. *J. Mol. Biol.* **1999**, *285* (4), 1563–1575.
- (114) Bhattacharya, S.; Jhunjhunwala, A.; Halder, A.; Bhattacharyya, D.; Mitra, A. Going beyond Base-Pairs: Topology-Based Characterization of Base-Multiplets in RNA. *RNA* **2019**, *25* (5), 573–589.
- (115) Bansal, M.; Bhattacharyya, D.; Ravi, B. NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Bioinformatics* **1995**, *11* (3), 281–287.
- (116) Pingali, P. K.; Halder, S.; Mukherjee, D.; Basu, S.; Banerjee, R.; Choudhury, D.; Bhattacharyya, D. Analysis of Stacking Overlap in Nucleic Acid Structures: Algorithm and Application. *J. Comput. Aided. Mol. Des.* **2014**, *28* (8), 851–867.
- (117) Parisien, M.; Major, F. The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data. *Nature* **2008**, *452* (7183), 51–55.

- (118) Zok, T.; Antczak, M.; Zurkowski, M.; Popenda, M.; Blazewicz, J.; Adamiak, R. W.; Szachniuk, M. RNAppdbec 2.0: Multifunctional Tool for RNA Structure Annotation. *Nucleic Acids Res.* **2018**, *46* (W1), W30–W35.
- (119) Popenda, M.; Szachniuk, M.; Antczak, M.; Purzycka, K. J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R. W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40* (14), e112–e112.
- (120) Boniecki, M. J.; Lach, G.; Dawson, W. K.; Tomala, K.; Lukasz, P.; Soltysinski, T.; Rother, K. M.; Bujnicki, J. M. SimRNA: A Coarse-Grained Method for RNA Folding Simulations and 3D Structure Prediction. *Nucleic Acids Res.* **2015**, *44* (7), 63.
- (121) Liu, P.; Li, L.; Millership, J. J.; Kang, H.; Leibowitz, J. L.; Giedroc, D. P. A U-Turn Motif-Containing Stem-Loop in the Coronavirus 5' Untranslated Region Plays a Functional Role in Replication. *RNA* **2007**, *13* (5), 763–780.
- (122) Leontis, N. B.; Zirbel, C. L. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In *RNA 3D Structure Analysis and Prediction*; 2012; Vol. 27, pp 281–298.
- (123) Hanson, R. M. Jmol-a Paradigm Shift in Crystallographic Visualization. *J. Appl. Crystallogr.* **2010**, *43* (5 PART 2), 1250–1260.
- (124) Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version~1.8*; 2015.
- (125) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. Gaussian16 Revision A.03. Gaussian Inc., Wallingford, CT **2016**.

- (126) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96* (9), 6796–6806.
- (127) Dunning Jr, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (1989), 1007.
- (128) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
- (129) Gehring, K.; Leroy, J. L.; Guéron, M. A Tetrameric DNA Structure with Protonated Cytosine-Cytosine Base Pairs. *Nature* **1993**, *363* (6429), 561–565.
- (130) Gu, Y.; Kar, T.; Scheiner, S. Fundamental Properties of the CH  $\cdots$  O Interaction: Is It a True Hydrogen Bond? *J. Am. Chem. Soc.* **1999**, *121* (40), 9411–9422.
- (131) Cubero, E.; Orozco, M.; Hobza, P.; Luque, F. J. Hydrogen Bond versus Anti-Hydrogen Bond: A Comparative Analysis Based on the Electron Density Topology. *J. Phys. Chem. A* **1999**, *103* (32), 6394–6401.
- (132) Halder, A.; Halder, S.; Bhattacharyya, D.; Mitra, A. Feasibility of Occurrence of Different Types of Protonated Base Pairs in RNA: A Quantum Chemical Study. *Phys. Chem. Chem. Phys.* **2014**, *16* (34), 18383–18396.
- (133) Kapinos, L. E.; Operschall, B. P.; Larsen, E.; Sigel, H. Understanding the Acid-Base Properties of Adenosine: The Intrinsic Basicities of N1, N3 and N7. *Chem. - A Eur. J.* **2011**, *17* (29), 8156–8164.

- (134) Jang, S. B.; Hung, L. W.; Chi, Y. I.; Holbrook, E. L.; Carter, R. J.; Holbrook, S. R. Structure of an RNA Internal Loop Consisting of Tandem C-A<sup>+</sup> Base Pairs. *Biochemistry* **1998**, *37* (34), 11726–11731.
- (135) Ma, J. C.; Dougherty, D. A. The Cation- $\pi$  Interaction. *Chem. Rev.* **1997**, *97* (5), 1303–1324.
- (136) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *J. Math. Biol.* **2008**, *56* (1–2), 215–252.
- (137) Elgavish, T.; Cannone, J. J.; Lee, J. C.; Harvey, S. C.; Gutell, R. R. AA.AG@helix.Ends: A:A and A:G Base-Pairs at the Ends of 16 S and 23 S rRNA Helices. *J. Mol. Biol.* **2001**, *310* (4), 735–753.
- (138) Pley, H. W.; Flaherty, K. M.; McKay, D. B. Three-Dimensional Structure of a Hammerhead Ribozyme. *Nature*. 1994, pp 68–74.
- (139) Bhattacharya, S.; Mittal, S.; Panigrahi, S.; Sharma, P.; S. P., P.; Paul, R.; Halder, S.; Halder, A.; Bhattacharyya, D.; Mitra, A. RNABP COGEST: A Resource for Investigating Functional RNAs. *Database* **2015**, *2015* (0), bav011–bav011.
- (140) Halder, S.; Bhattacharyya, D. RNA Structure and Dynamics: A Base Pairing Perspective. *Prog. Biophys. Mol. Biol.* **2013**, *113* (2), 264–283.
- (141) Halder, S.; Bhattacharyya, D. Structural Stability of Tandemly Occurring Noncanonical Basepairs within Double Helical Fragments: Molecular Dynamics Studies of Functional RNA. *J. Phys. Chem. B* **2010**, *114* (44), 14028–14040.

- (142) Panigrahi, S.; Pal, R.; Bhattacharyya, D. Structure and Energy of Non-Canonical Basepairs: Comparison of Various Computational Chemistry Methods with Crystallographic Ensembles. *J. Biomol. Struct. Dyn.* **2011**, *29* (3), 541–556.
- (143) Bandyopadhyay, D.; Bhattacharyya, D. Estimation of Strength in Different Extra Watson–Crick Hydrogen Bonds in DNA Double Helices through Quantum Chemical Studies. *Biopolymers* **2006**, *83* (3), 313–325.
- (144) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix. *Nucleic Acids Res.* **2006**, *34* (2), 564–574.
- (145) Šponer, J.; Šponer, J. E.; Mládek, A.; Jurečka, P.; Banáš, P.; Otyepka, M. Nature and Magnitude of Aromatic Base Stacking in DNA and RNA: Quantum Chemistry, Molecular Mechanics, and Experiment. *Biopolymers*. 2013, pp 978–988.
- (146) Hunter, C. A.; Lu, X.-J. DNA Base-Stacking Interactions: A Comparison of Theoretical Calculations with Oligonucleotide X-Ray Crystal Structures. *J. Mol. Biol.* **1997**, *265* (5), 603–619.
- (147) Hobza, P.; Šponer, J. Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations. *Chem. Rev.* **1999**, *99* (11), 3247–3276.
- (148) Hunter, R. S.; van Mourik, T. DNA Base Stacking: The Stacked Uracil/Uracil and Thymine/Thymine Minima. *J. Comput. Chem.* **2012**, *33* (27), 2161–2172.
- (149) Mukherjee, S.; Kailasam, S.; Bansal, M.; Bhattacharyya, D. Energy Hyperspace for Stacking Interaction in AU/AU Dinucleotide Step: Dispersion-Corrected Density

- Functional Theory Study. *Biopolymers* **2014**, *101* (1), 107–120.
- (150) Xia, T.; SantaLucia, J.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson - Crick Base Pairs. *Biochemistry* **1998**, *37* (42), 14719–14735.
- (151) Marky, L. A.; Breslauer, K. J. Calorimetric Determination of Base-Stacking Enthalpies in Double-Helical DNA Molecules. *Biopolymers* **1982**, *21* (11), 2185–2194.
- (152) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Naturally Occurring RNA Single Mismatches with G-U Nearest Neighbors. *Biochemistry* **2008**, *47* (38), 10178–10187.
- (153) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Single Mismatches Found in Naturally Occurring RNA. *Biochemistry* **2007**, *46* (46), 13425–13436.
- (154) Christiansen, M. E.; Znosko, B. M. Thermodynamic Characterization of Tandem Mismatches Found in Naturally Occurring RNA. *Nucleic Acids Res.* **2009**, *37* (14), 4696–4706.
- (155) Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion Interactions in Density-Functional Theory. *J. Phys. Org. Chem.* **2009**, *22* (12), 1127–1135.
- (156) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. Assessment of the Performance of the M05–2X and M06–2X Exchange-Correlation Functionals for Noncovalent Interactions in Biomolecules. *J. Chem. Theory Comput.* **2008**, *4* (12), 1996–2000.
- (157) El Hassan, M. A.; Calladine, C. R. Conformational Characteristics of DNA: Empirical

- Classifications and a Hypothesis for the Conformational Behaviour of Dinucleotide Steps. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **1997**, 355 (1722), 43–100.
- (158) Peterson, K. A.; Kendall, R. A.; Dunning, Jr., T. H. Benchmark Calculations with Correlated Molecular Wave Functions. II. Configuration Interaction Calculations on First Row Diatomic Hydrides. *J. Chem. Phys.* **1993**, 99 (1993), 1930.
- (159) Kendall, R. a; Dunning Jr., T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, 96 (May 2013), 6796.
- (160) Huber, R. G.; Margreiter, M. A.; Fuchs, J. E.; von Grafenstein, S.; Tautermann, C. S.; Liedl, K. R.; Fox, T. Heteroaromatic  $\pi$ -Stacking Energy Landscapes. *J. Chem. Inf. Model.* **2014**, 54 (5), 1371–1379.
- (161) Aragó, J.; Sancho-García, J. C.; Ortí, E.; Beljonne, D. Ab Initio Modeling of Donor-Acceptor Interactions and Charge-Transfer Excitations in Molecular Complexes: The Case of Terthiophene- Tetracyanoquinodimethane. *J. Chem. Theory Comput.* **2011**, 7 (7), 2068–2077.
- (162) MJ Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, V Barone, B Mennucci, GA Petersson, H Nakatsuji, M Caricato, X Li, HP Hratchian, AF Izmaylov, J Bloino, G Zheng, JL Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Haseg, D. F. Gaussian09, Revision B.01. *Gaussian Inc., Wallingford, CT* **2010**.
- (163) Šponer, J.; Jurečka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, 126 (32), 10142–10151.

- (164) Bhattacharyya, D. Bioinformatics Resource Materials developed at SINP  
<http://www.saha.ac.in/biop/bioinformatics.html>.
- (165) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.
- (166) Poater, J.; Swart, M.; Bickelhaupt, F. M.; Fonseca Guerra, C.; Lewis, F. D.; Berlin, Y. A.; Ratner, M. A.; Siebbeles, L. D. A.; Nibbering, E. T. J.; Elsaesser, T. B-DNA Structure and Stability: The Role of Hydrogen Bonding,  $\pi$ - $\pi$  Stacking Interactions, Twist-Angle, and Solvation. *Org. Biomol. Chem.* **2014**, *12* (26), 4691–4700.
- (167) Šponer, J.; Leszczyński, J.; Hobza, P. Nature of Nucleic Acid–Base Stacking: Nonempirical Ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. *J. Phys. Chem.* **1996**, *100* (13), 5590–5596.
- (168) Ananth, P.; Goldsmith, G.; Yathindra, N. An Innate Twist between Crick’s Wobble and Watson-Crick Base Pairs. *RNA* **2013**, *19* (8), 1038.
- (169) Mokdad, A. Structural and Evolutionary Classification of G/U Wobble Basepairs in the Ribosome. *Nucleic Acids Res.* **2006**, *34* (5), 1326–1341.
- (170) Leontis, N. B.; Stombaugh, J.; Westhof, E. The Non-Watson-Crick Base Pairs and Their Associated Isostericity Matrices. *Nucleic Acids Res.* **2002**, *30* (16), 3497–3531.
- (171) Leontis, N. B.; Westhof, E. Geometric Nomenclature and Classification of RNA Base Pairs. *RNA* **2001**, *7* (4), 499–512.

- (172) Baeyens, K. J.; De Bondt, H. L.; Pardi, A.; Holbrook, S. R. A Curved RNA Helix Incorporating an Internal Loop with G·A and A·A Non-Watson-Crick Base Pairing. *Proc. Natl. Acad. Sci.* **1996**, *93* (23), 12851–12855.
- (173) Doudna, J. A.; Cate, J. H. RNA Structure: Crystal Clear? *Curr. Opin. Struct. Biol.* **1997**, *7* (3), 310–316.
- (174) Spackova, N.; Šponer, J. Molecular Dynamics Simulations of Sarcin-Ricin RRNA Motif. *Nucleic Acids Res.* **2006**, *34* (2), 697–708.
- (175) Kondo, J.; Westhof, E. The Bacterial and Mitochondrial Ribosomal A-Site Molecular Switches Possess Different Conformational Substates. *Nucleic Acids Res.* **2008**, *36* (8), 2654–2666.
- (176) Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. MiRBase: From MicroRNA Sequences to Function. *Nucleic Acids Res.* **2019**, *47* (D1), D155–D162.
- (177) Réblová, K.; Střelcová, Z.; Kulhánek, P.; Beššeová, I.; Mathews, D. H.; Van Nostrand, K.; Yildirim, I.; Turner, D. H.; Šponer, J. An RNA Molecular Switch: Intrinsic Flexibility of 23S RRNA Helices 40 and 68 5'-UAA/5'-GAN Internal Loops Studied by Molecular Dynamics Methods. *J. Chem. Theory Comput.* **2010**, *6* (3), 910–929.
- (178) Halder, S.; Bhattacharyya, D. Structural Variations of Single and Tandem Mismatches in RNA Duplexes: A Joint MD Simulation and Crystal Structure Database Analysis. *J. Phys. Chem. B* **2012**, *116* (39), 11845–11856.
- (179) Halder, S.; Bhattacharyya, D. Structural Stability of Tandemly Occurring Noncanonical Basepairs within Double Helical Fragments: Molecular Dynamics Studies of Functional

- RNA. *J. Phys. Chem. B* **2010**, *114* (44), 14028–14040.
- (180) Kruse, H.; Banáš, P.; Šponer, J. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.* **2019**, *15* (1), 95–115.
- (181) Egli, M.; Usman, N.; Rich, A. Conformational Influence of the Ribose 2'-Hydroxyl Group: Crystal Structures of DNA-RNA Chimeric Duplexes. *Biochemistry* **1993**, *32* (13), 3221–3237.
- (182) Ban, C.; Ramakrishnan, B.; Sundaralingam, M. A Single 2'-Hydroxyl Group Converts B-DNA to A-DNA: Crystal Structure of the DNA-RNA Chimeric Decamer Duplex d(CCGGC)r(G)d(CCGG) with a Novel Intermolecular G·C Base-Paired Quadruplet. *J. Mol. Biol.* **1994**, *236* (1), 275–285.
- (183) Lindqvist, M.; Sarkar, M.; Winqvist, A.; Rozners, E.; Strömberg, R.; Gräslund, A. Optical Spectroscopic Study of the Effects of a Single Deoxyribose Substitution in a Ribose Backbone: Implications in RNA-RNA Interaction. *Biochemistry* **2000**, *39* (7), 1693–1701.
- (184) Gang Chen, §; Scott D. Kennedy, ¶; Jing Qiao, §; Thomas R. Krugh, § and; Douglas H. Turner\*, §,⊥. An Alternating Sheared AA Pair and Elements of Stability for a Single Sheared Purine-Purine Pair Flanked by Sheared GA Pairs in RNA†,‡. **2006**.
- (185) Chakraborty, D.; Wales, D. J. Dynamics of an Adenine-Adenine RNA Conformational Switch from Discrete Path Sampling. *J. Chem. Phys.* **2019**, *150* (12), 125101.
- (186) Bansal, M.; Bhattacharyya, D.; Ravi, B. NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Comput. Appl. Biosci.*

- 1995, *11* (3), 281–287.
- (187) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.
- (188) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. A Direct MP2 Gradient Method. *Chem. Phys. Lett.* **1990**, *166* (3), 275–280.
- (189) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102* (11), 1995–2001.
- (190) Chandrasekaran, R.; Arnott, S. The Structure of *B*-DNA in Oriented Fibers. *J. Biomol. Struct. Dyn.* **1996**, *13* (6), 1015–1027.
- (191) Pérez, A.; Marchán, I.; Svozil, D.; Šponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of Alpha/Gamma Conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829.
- (192) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6* (12), 3836–3849.
- (193) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. Refinement of the Cornell et Al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*

- 2011**, 7 (9), 2886–2902.
- (194) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, 1, 19–25.
- (195) Darden, T.; Pearlman, D.; Pedersen, L. G. Ionic Charging Free Energies: Spherical versus Periodic Boundary Conditions. *J. Chem. Phys.* **1998**, 109 (24), 10921–10935.
- (196) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, 52 (12), 7182–7190.
- (197) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, 126 (1), 014101.
- (198) Das, A.; Chakrabarti, J.; Ghosh, M. Conformational Contribution to Thermodynamics of Binding in Protein-Peptide Complexes through Microscopic Simulation. *Biophys. J.* **2013**, 104 (6), 1274–1284.
- (199) Sarkar, S.; Maity, A.; Sarma Phukon, A.; Ghosh, S.; Chakrabarti, R. Salt Induced Structural Collapse, Swelling, and Signature of Aggregation of Two SsDNA Strands: Insights from Molecular Dynamics Simulation. *J. Phys. Chem. B* **2019**, 123 (1), 47–56.
- (200) Saenger, W. *Principles of Nucleic Acid Structure*; Springer Advanced Texts in Chemistry; Springer New York: New York, NY, 1984.
- (201) Romanowska, J.; McCammon, J. A.; Trylska, J. Understanding the Origins of Bacterial Resistance to Aminoglycosides through Molecular Dynamics Mutational Study of the Ribosomal A-Site. *PLoS Comput. Biol.* **2011**, 7 (7), e1002099.

- (202) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen Base Pairs in Canonical Duplex DNA. *Nature* **2011**, *470* (7335), 498–502.
- (203) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational Analysis of Nucleic Acids Revisited: Curves+. *Nucleic Acids Res.* **2009**, *37* (17), 5917–5929.
- (204) Réblová, K.; Střelcová, Z.; Kulhánek, P.; Beššeová, I.; Mathews, D. H.; Van Nostrand, K.; Yildirim, I.; Turner, D. H.; Šponer, J. An RNA Molecular Switch: Intrinsic Flexibility of 23S rRNA Helices 40 and 68 5'-UAA/5'-GAN Internal Loops Studied by Molecular Dynamics Methods. *J. Chem. Theory Comput.* **2010**, *6* (3), 910–929.
- (205) Brown, R. F.; Andrews, C. T.; Elcock, A. H. Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *J. Chem. Theory Comput.* **2015**, *11* (5), 2315–2328.
- (206) Häse, F.; Zacharias, M. Free Energy Analysis and Mechanism of Base Pair Stacking in Nicked DNA. *Nucleic Acids Res.* **2016**, *44* (15), 7100–7108.
- (207) Nelson, H. C. M.; Finch, J. T.; Luisi, B. F.; Klug, A. The Structure of an Oligo(DA)·oligo(DT) Tract and Its Biological Implications. *Nature* **1987**, *330* (6145), 221–226.
- (208) Bandyopadhyay, D.; Bhattacharyya, D. Estimation of Strength in Different Extra Watson–Crick Hydrogen Bonds in DNA Double Helices through Quantum Chemical Studies. *Biopolymers* **2006**, *83* (3), 313–325.

- (209) Frank-Kamenetskii, M. D.; Mirkin, S. M. Triplex DNA Structures. *Annu. Rev. Biochem.* **1995**, *64* (1), 65–95.
- (210) Garg, A.; Heinemann, U. A Novel Form of RNA Double Helix Based on G·U and C·A+ Wobble Base Pairing. *RNA* **2018**, *24* (2), 209–218.
- (211) Chawla, M.; Sharma, P.; Halder, S.; Bhattacharyya, D.; Mitra, A. Protonation of Base Pairs in RNA: Context Analysis and Quantum Chemical Investigations of Their Geometries and Stabilities. *J. Phys. Chem. B* **2011**, *115* (6), 1469–1484.
- (212) Halder, A.; Halder, S.; Bhattacharyya, D.; Mitra, A. Feasibility of Occurrence of Different Types of Protonated Base Pairs in RNA: A Quantum Chemical Study. *Phys. Chem. Chem. Phys.* **2014**, *16* (34), 18383–18396.
- (213) Bernhardt, H. S.; Tate, W. P. Primordial Soup or Vinaigrette: Did the RNA World Evolve at Acidic PH? *Biol. Direct* **2012**, *7* (1), 4.
- (214) Wang, B. The RNA I-Motif in the Primordial RNA World. *Orig. Life Evol. Biosph.* **2019**, *49* (1–2), 105–109.
- (215) Peng, C.; Bernhard Schlegel, H. Combining Synchronous Transit and Quasi-Newton Methods to Find Transition States. *Isr. J. Chem.* **1993**, *33* (4), 449–454.
- (216) Herzog, E.; Frigato, T.; Helms, V.; Lancaster, C. R. D. Energy Barriers of Proton Transfer Reactions between Amino Acid Side Chain Analogs and Water Fromab Initio Calculations. *J. Comput. Chem.* **2006**, *27* (13), 1534–1547.
- (217) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

- (218) Arnott, S.; Hukins, D. W. L.; Dover, S. D. Optimised Parameters for RNA Double-Helices. *Biochem. Biophys. Res. Commun.* **1972**, *48* (6), 1392–1399.

## **Chapter- 7 References**

- (1) Avery, O. T.; Macleod, C. M.; McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from *Pneumococcus* Type Iii. *J. Exp. Med.* **1944**, *79* (2), 137–158.
- (2) HERSHEY, A. D.; CHASE, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J. Gen. Physiol.* **1952**, *36* (1), 39–56.
- (3) CHARGAFF, E.; ZAMENHOF, S.; GREEN, C. Human Desoxypentose Nucleic Acid: Composition of Human Desoxypentose Nucleic Acid. *Nature* **1950**, *165* (4202), 756–757.
- (4) WATSON, J. D.; CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171* (4356), 737–738.
- (5) Meselson, M.; Stahl, F. W.; Vinograd, J. EQUILIBRIUM SEDIMENTATION OF MACROMOLECULES IN DENSITY GRADIENTS. *Proc. Natl. Acad. Sci.* **1957**, *43* (7), 581–588.
- (6) Hoagland, M. B.; Zamecnik, P. C.; Stephenson, M. L. Intermediate Reactions in Protein Biosynthesis. *BBA - Biochim. Biophys. Acta* **1957**, *24* (C), 215–216.
- (7) Gros, F.; Hiatt, H.; Gilbert, W.; Kurland, C. G.; Risebrough, R. W.; Watson, J. D. Unstable Ribonucleic Acid Revealed by Pulse Labelling of *Escherichia Coli*. *Nature* **1961**, *190* (4776), 581–585.
- (8) Brenner, S.; Jacob, F.; Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **1961**, *190* (4776), 576–581.
- (9) NIRENBERG, M. W.; MATTHAEI, J. H. The Dependence of Cell-Free Protein Synthesis

- in *E. Coli* upon Naturally Occurring or Synthetic Polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47* (2), 1588–1602.
- (10) Noller, H. F. Ribosomal RNA and Translation. *Annu. Rev. Biochem.* **1991**, *60* (1), 191–227.
- (11) Watson, J. D. Involvement of RNA in the Synthesis of Proteins. *Science* (80-. ). **1963**, *140* (3562), 17–26.
- (12) Ban, N.; Freeborn, B.; Nissen, P.; Penczek, P.; Grassucci, R. A.; Sweet, R.; Frank, J.; Moore, P. B.; Steitz, T. A. A 9 Å Resolution X-Ray Crystallographic Map of the Large Ribosomal Subunit. *Cell* **1998**, *93* (7), 1105–1115.
- (13) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vornheln, C.; Hartsch, T.; Ramakrishnan, V. Structure of the 30S Ribosomal Subunit. *Nature* **2000**, *407* (6802), 327–339.
- (14) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* (80-. ). **2000**, *289* (5481), 905–920.
- (15) Rich, A. A HYBRID HELIX CONTAINING BOTH DEOXYRIBOSE AND RIBOSE POLYNUCLEOTIDES AND ITS RELATION TO THE TRANSFER OF INFORMATION BETWEEN THE NUCLEIC ACIDS. *Proc. Natl. Acad. Sci.* **1960**, *46* (8), 1044–1053.
- (16) Rosenberg, J. M.; Seeman, N. C.; Kim, J. J. P.; Suddath, F. L.; Nicholas, H. B.; Rich, A. Double Helix at Atomic Resolution. *Nature* **1973**, *243* (5403), 150–154.

- (17) DALY, M. M.; ALLFREY, V. G.; MIRSKY, A. E. Purine and Pyrimidine Contents of Some Desoxyribose Nucleic Acids. *J. Gen. Physiol.* **1950**, *33* (5), 497–510.
- (18) Harcourt, E. M.; Kietrys, A. M.; Kool, E. T. Chemical and Structural Effects of Base Modifications in Messenger RNA. *Nature*. Nature Publishing Group January 18, 2017, pp 339–346.
- (19) Manning, G. S. Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. *J. Chem. Phys.* **1969**, *51* (3), 924.
- (20) Symbols for Specifying the Conformation of Polysaccharide Chains. Recommendations 1981. *Eur. J. Biochem.* **1983**, *131* (1), 5–7.
- (21) Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. B-DNA Twisting Correlates with Base-Pair Morphology. *J. Mol. Biol.* **1995**, *247* (1), 34–48.
- (22) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (19), 11163–11168.
- (23) Lavery, R. *Multiple Aspects of DNA and RNA from Biophysics to Bioinformatics*; 2005.
- (24) Duarte, C. M.; Pyle, A. M. Stepping through an RNA Structure: A Novel Approach to Conformational Analysis. *J. Mol. Biol.* **1998**, *284* (5), 1465–1478.
- (25) Malathi, R.; Yathindra, N. Backbone Conformation in Nucleic Acids: An Analysis of Local Helicity through Heminucleotide Scheme and a Proposal for a Unified Conformational Plot. *J. Biomol. Struct. Dyn.* **1985**, *3* (1), 127–144.
- (26) Wilson, H. R.; Rahman, A. Nucleoside Conformation and Non-Bonded Interactions. *J.*

- Mol. Biol.* **1971**, *56* (1), 129–142.
- (27) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. a New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94* (23), 8205–8212.
- (28) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*; Interdisciplinary Applied Mathematics; Springer New York: New York, NY, 2010; Vol. 21.
- (29) Hoogsteen, K. The Structure of Crystals Containing a Hydrogen-Bonded Complex of 1-Methylthymine and 9-Methyladenine. *Acta Crystallogr.* **1959**, *12* (10), 822–823.
- (30) Courtois, Y.; Fromageot, P.; Guschlbauer, W. Protonated Polynucleotide Structures: 3. An Optical Rotatory Dispersion Study of the Protonation of DNA. *Eur. J. Biochem.* **1968**, *6* (4), 493–501.
- (31) Patel, D. J.; Tonelli, A. E. Assignment of the Proton Nmr Chemical Shifts of the T□N3H and G□N1H Proton Resonances in Isolated AT and GC Watson-Crick Base Pairs in Double-stranded Deoxy Oligonucleotides in Aqueous Solution. *Biopolymers* **1974**, *13* (10), 1943–1964.
- (32) Seeman, N. C.; Rosenberg, J. M.; Suddath, F. L.; Kim, J. J.; Rich, A. RNA Double-Helical Fragments at Atomic Resolution. I. The Crystal and Molecular Structure of Sodium Adenylyl-3',5'-Uridine Hexahydrate. *J. Mol. Biol.* **1976**, *104* (1), 109–144.
- (33) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA Dodecamer: Conformation and Dynamics. *Proc. Natl. Acad. Sci.*

- U. S. A.* **1981**, 78 (4), 2179–2183.
- (34) Wang, A. H. J.; Fujii, S.; Van Boom, J. H.; Van Der Marel, G. A.; Van Boeckel, S. A. A.; Rich, A. Molecular Structure of r(GCG)<sub>n</sub>d(TATACGC): A DNA-RNA Hybrid Helix Joined to Double Helical DNA. *Nature* **1982**, 299 (5884), 601–604.
- (35) Heinemann, U.; Alings, C. Crystallographic Study of One Turn of G/C-Rich B-DNA. *J. Mol. Biol.* **1989**, 210 (2), 369–381.
- (36) Dock-Bregeon, A. C.; Chevrier, B.; Podjarny, A.; Johnson, J.; de Bear, J. S.; Gough, G. R.; Gilham, P. T.; Moras, D. Crystallographic Structure of an RNA Helix: [U(UA)<sub>6</sub>A]<sub>2</sub>. *J. Mol. Biol.* **1989**, 209 (3), 459–474.
- (37) Coimbatore Narayanan, B.; Westbrook, J.; Ghosh, S.; Petrov, A. I.; Sweeney, B.; Zirbel, C. L.; Leontis, N. B.; Berman, H. M. The Nucleic Acid Database: New Features and Capabilities. *Nucleic Acids Res.* **2014**, 42 (D1), D114–D122.
- (38) Saenger, W. *Principles of Nucleic Acid Structure*; Springer Advanced Texts in Chemistry; Springer New York: New York, NY, 1984.
- (39) Dickerson, R. E. Definitions and Nomenclature of Nucleic Acid Structure Components. *Nucleic Acids Res.* **1989**, 17 (5), 1797–1803.
- (40) Blanchet, C.; Pasi, M.; Zakrzewska, K.; Lavery, R. CURVES+ Web Server for Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. *Nucleic Acids Res.* **2011**, 39 (suppl), W68–W73.
- (41) Lu, X. J.; Olson, W. K. 3DNA: A Versatile, Integrated Software System for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic-Acid Structures. *Nat. Protoc.*

- 2008**, 3 (7), 1213–1227.
- (42) NUPARM-Plus-A Program for analyzing sequence dependent variations in nucleic acids.
- (43) Leontis, N. B.; Westhof, E. Geometric Nomenclature and Classification of RNA Base Pairs. *RNA* **2001**, 7 (4), 499–512.
- (44) Desiraju, G. R. (Gautam R. .; Steiner, T. *The Weak Hydrogen Bond : In Structural Chemistry and Biology*; Oxford University Press, 2001.
- (45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28 (1), 235–242.
- (46) Lu, X. J.; Olson, W. K. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, 31 (17), 5108–5121.
- (47) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. The Nucleic Acid Database. A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.* **1992**, 63 (3), 751–759.
- (48) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *J. Math. Biol.* **2008**, 56 (1–2), 215–252.
- (49) Lemieux, S.; Major, F. RNA Canonical and Non-Canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research*. Oxford

- University Press October 1, 2002, pp 4250–4263.
- (50) Bayly, C. I.; Merz, K. M.; Ferguson, D. M.; Cornell, W. D.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; Cieplak, P.; Gould, I. R.; Spellmeyer, D. C. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.
- (51) Das, J.; Mukherjee, S.; Mitra, A.; Bhattacharyya, D. Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Biomol. Struct. Dyn.* **2006**, *24* (2).
- (52) Lu, X.-J.; Shakked, Z.; Olson, W. K. A-Form Conformational Motifs in Ligand-Bound DNA Structures. *J. Mol. Biol.* **2000**, *300* (4), 819–840.
- (53) 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.; et al. A Standard Reference Frame for the Description of Nucleic Acid Base-Pair Geometry. *J. Mol. Biol.* **2001**, *313* (1), 229–237.
- (54) Leontis, N. B. The Non-Watson-Crick Base Pairs and Their Associated Isostericity Matrices. *Nucleic Acids Res.* **2002**, *30* (16), 3497–3531.
- (55) Nasalean, L.; Stombaugh, J.; Zirbel, C. L.; Leontis, N. B. RNA 3D Structural Motifs: Definition, Identification, Annotation, and Database Searching. In *Non-Protein Coding RNAs*; Springer Berlin Heidelberg, 2008; pp 1–26.
- (56) Stombaugh, J.; Zirbel, C. L.; Westhof, E.; Leontis, N. B. Frequency and Isostericity of RNA Base Pairs. *Nucleic Acids Res.* **2009**, *37* (7), 2294–2312.

- (57) Popena, M.; Szachniuk, M.; Blazewicz, M.; Wasik, S.; Burke, E. K.; Blazewicz, J.; Adamiak, R. W. RNA FRABASE 2.0: An Advanced Web-Accessible Database with the Capacity to Search the Three-Dimensional Fragments within RNA Structures. *BMC Bioinformatics* **2010**, *11* (1), 231.
- (58) Hamdani, H. Y.; Appasamy, S. D.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. NASSAM: A Server to Search for and Annotate Tertiary Interactions and Motifs in Three-Dimensional Structures of Complex RNA Molecules. *Nucleic Acids Res.* **2012**, *40* (W1).
- (59) RNA 3D Motif Atlas.
- (60) Šponer, J. E.; Špačková, N.; Kulhánek, P.; Leszczynski, J.; Šponer, J. Non-Watson-Crick Base Pairing in RNA. Quantum Chemical Analysis of the Cis Watson-Crick/Sugar Edge Base Pair Family. *J. Phys. Chem. A* **2005**, *109* (10), 2292–2301.
- (61) Kelly, R. E. A.; Lee, Y. J.; Kantorovich, L. N. Homopairing Possibilities of the DNA Base Adenine. *J. Phys. Chem. B* **2005**, *109* (24), 11933–11939.
- (62) Oliva, R.; Cavallo, L.; Tramontano, A. Accurate Energies of Hydrogen Bonded Nucleic Acid Base Pairs and Triplets in tRNA Tertiary Interactions. *Nucleic Acids Res.* **2006**.
- (63) Bhattacharyya, D.; Koripella, S. C.; Mitra, A.; Rajendran, V. B.; Sinha, B. Theoretical Analysis of Noncanonical Base Pairing Interactions in RNA Molecules. In *Journal of Biosciences*; Springer, 2007; Vol. 32, pp 809–825.
- (64) Roy, A.; Panigrahi, S.; Bhattacharyya, M.; Bhattacharyya, D. Structure, Stability, and Dynamics of Canonical and Noncanonical Base Pairs: Quantum Chemical Studies. *J. Phys. Chem. B* **2008**, *112* (12), 3786–3796.

- (65) Sharma, P.; Šponer, J. E.; Šponer, J.; Sharma, S.; Bhattacharyya, D.; Mitra, A. On the Role of the Eis Hoogsteen: Sugar-Edge Family of Base Pairs in Platforms and Triplets-Quantum Chemical Insights into RNA Structural Biology. *J. Phys. Chem. B* **2010**, *114* (9), 3307–3320.
- (66) Brovarets, O. O.; Yurenko, Y. P.; Hovorun, D. M. Intermolecular CH $\cdots$ O/N H-Bonds in the Biologically Important Pairs of Natural Nucleobases: A Thorough Quantum-Chemical Study. *J. Biomol. Struct. Dyn.* **2014**, *32* (6), 993–1022.
- (67) Marino, T. DFT Investigation of the Mismatched Base Pairs (T-Hg-T) $_3$ , (U-Hg-U) $_3$ , d(T-Hg-T) $_2$ , and d(U-Hg-U) $_2$ . *J. Mol. Model.* **2014**, *20* (6), 2303.
- (68) Roy, A.; Panigrahi, S.; Bhattacharyya, M.; Bhattacharyya, D. Structure, Stability, and Dynamics of Canonical and Noncanonical Base Pairs: Quantum Chemical Studies. *J. Phys. Chem. B* **2008**, *112* (12).
- (69) Šponer, J. E.; Leszczynski, J.; Sychrovský, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109* (39), 18680–18689.
- (70) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.
- (71) Grimme, S. Do Special Noncovalent  $\pi$ - $\pi$  Stacking Interactions Really Exist? *Angew. Chemie Int. Ed.* **2008**, *47* (18), 3430–3434.

- (72) Hunter, C. A. Sequence-Dependent DNA Structure. *BioEssays* **1996**, *18* (2), 157–162.
- (73) Calladine, C. R. Mechanics of Sequence-Dependent Stacking of Bases in B-DNA. *Journal of Molecular Biology*. 1982, pp 343–352.
- (74) El Hassan, M. A.; Calladine, C. R. Propeller-Twisting of Base-Pairs and the Conformational Mobility of Dinucleotide Steps in DNA. *J. Mol. Biol.* **1996**, *259* (1), 95–103.
- (75) Mondal, M.; Mukherjee, S.; Halder, S.; Bhattacharyya, D. Stacking Geometry for Non-Canonical G:U Wobble Base Pair Containing Dinucleotide Sequences in RNA: Dispersion-Corrected DFT-D Study. *Biopolymers* **2015**, *103* (6), 328–338.
- (76) Maiti, S.; Bhattacharyya, D. Stacking Interactions Involving Non-Watson–Crick Basepairs: Dispersion Corrected Density Functional Theory Studies. *Phys. Chem. Chem. Phys.* **2017**, *19* (42), 28718–28730.
- (77) Mukherjee, S.; Kailasam, S.; Bansal, M.; Bhattacharyya, D. Stacking Interactions in RNA and DNA: Roll-Slide Energy Hyperspace for Ten Unique Dinucleotide Steps. *Biopolymers* **2015**, *103* (3), 134–147.
- (78) Maiti, S.; Mukherjee, D.; Roy, P.; Chakrabarti, J.; Bhattacharyya, D. Stacking Geometry between Two Sheared Watson-Crick Basepairs: Computational Chemistry and Bioinformatics Based Prediction. *Biochim. Biophys. Acta - Gen. Subj.* **2020**, *1864* (7), 129600.
- (79) Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98* (2), 1372–1377.

- (80) Šponer, J.; Shukla, M. K.; Wang, J.; Leszczynski, J. Computational Modeling of DNA and RNA Fragments. In *Handbook of Computational Chemistry*; Springer International Publishing, 2017; pp 1803–1826.
- (81) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110* (9), 5023–5063.
- (82) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10* (44), 6615–6620.
- (83) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.
- (84) Hobza, P.; Selzle, H. L.; Schlag, E. W. Potential Energy Surface for the Benzene Dimer. Results of Ab Initia CCSD(T) Calculations Show Two Nearly Isoenergetic Structures: T-Shaped and Parallel-Displaced. *J. Phys. Chem.* **1996**, *100* (48), 18790–18794.
- (85) Jaffe, R. L.; Smith, G. D. A Quantum Chemistry Study of Benzene Dimer. *J. Chem. Phys.* **1996**, *105* (7), 2780–2788.
- (86) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19* (4), 553–566.

- (87) Malenov, D. P.; Ninković, D. B.; Zarić, S. D. Stacking of Metal Chelates with Benzene: Can Dispersion-Corrected DFT Be Used to Calculate Organic-Inorganic Stacking? *ChemPhysChem* **2015**, *16* (4), 761–768.
- (88) Das, J.; Mukherjee, S.; Mitra, A.; Bhattacharyya, D. Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Biomol. Struct. Dyn.* **2006**, *24* (2), 149–161.
- (89) Bhattacharyya, D.; Halder, S.; Basu, S.; Mukherjee, D.; Kumar, P.; Bansal, M. RNAHelix: Computational Modeling of Nucleic Acid Structures with Watson-Crick and Non-Canonical Base Pairs. *J. Comput. Aided. Mol. Des.* **2017**, *31* (2), 219–235.
- (90) Andronescu, M.; Bereg, V.; Hoos, H. H.; Condon, A. RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. *BMC Bioinformatics* **2008**, *9* (1), 340.
- (91) Popena, M.; Blazewicz, M.; Szachniuk, M.; Adamiak, R. W. RNA FRABASE Version 1.0: An Engine with a Database to Search for the Three-Dimensional Fragments within RNA Structures. *Nucleic Acids Res.* **2008**, *36* (Database issue), D386-91.
- (92) Popena, M.; Szachniuk, M.; Antczak, M.; Purzycka, K. J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R. W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40* (14), e112.
- (93) Petrov, A. I.; Zirbel, C. L.; Leontis, N. B. Automated Classification of RNA 3D Motifs and the RNA 3D Motif Atlas. *RNA* **2013**, *19* (10), 1327–1340.
- (94) Klosterman, P. S.; Tamura, M.; Holbrook, S. R.; Brenner, S. E. SCOR: A Structural

- Classification of RNA Database. *Nucleic Acids Res.* **2002**, *30*, 392–394.
- (95) Chojnowski, G.; Waleń, T.; Bujnicki, J. M. RNA Bricks - A Database of RNA 3D Motifs and Their Interactions. *Nucleic Acids Res.* **2014**, *42* (D1), D123.
- (96) Bhattacharya, S.; Mittal, S.; Panigrahi, S.; Sharma, P.; S P, P.; Paul, R.; Halder, S.; Halder, A.; Bhattacharyya, D.; Mitra, A. RNABP COGEST: A Resource for Investigating Functional RNAs. *Database (Oxford)*. **2015**, *2015* (0), bav011–bav011.
- (97) Westhof, E.; Fritsch, V. RNA Folding: Beyond Watson-Crick Pairs. *Structure*. March 2000, pp R55–R65.
- (98) Nagaswamy, U.; Larios-Sanz, M.; Hury, J.; Collins, S.; Zhang, Z.; Zhao, Q.; Fox, G. E. NCIR: A Database of Non-Canonical Interactions in Known RNA Structures. *Nucleic Acids Res.* **2002**, *30* (1), 395–397.
- (99) Brovarets, O. O.; Yurenko, Y. P.; Hovorun, D. M. Intermolecular CH···O/N H-Bonds in the Biologically Important Pairs of Natural Nucleobases: A Thorough Quantum-Chemical Study. *J. Biomol. Struct. Dyn.* **2014**, *32* (6), 993–1022.
- (100) Šponer, J. E.; Leszczynski, J.; Sychrovský, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109* (39), 18680–18689.
- (101) Sharma, P.; Mitra, A.; Sharma, S.; Singh, H.; Bhattacharyya, D. Quantum Chemical Studies of Structures and Binding in Noncanonical Rna Base Pairs: The Trans Watson-Crick:Watson-Crick Family. *J. Biomol. Struct. Dyn.* **2008**, *25* (6), 709–732.
- (102) Mládek, A.; Sharma, P.; Mitra, A.; Bhattacharyya, D.; Šponer, J.; Šponer, J. E. Trans

- Hoogsteen/Sugar Edge Base Pairing in RNA. Structures, Energies, and Stabilities from Quantum Chemical Calculations. *J. Phys. Chem. B* **2009**, *113* (6), 1743–1755.
- (103) Chawla, M.; Sharma, P.; Halder, S.; Bhattacharyya, D.; Mitra, A. Protonation of Base Pairs in RNA: Context Analysis and Quantum Chemical Investigations of Their Geometries and Stabilities. *J. Phys. Chem. B* **2011**, *115* (6), 1469–1484.
- (104) Lu, X. J.; Bussemaker, H. J.; Olson, W. K. DSSR: An Integrated Software Tool for Dissecting the Spatial Structure of RNA. *Nucleic Acids Res.* **2015**, *43* (21), gkv716.
- (105) Walé, T.; Chojnowski, G.; Gierski, P.; Bujnicki, J. M. ClaRNA: A Classifier of Contacts in RNA 3D Structures Based on a Comparative Analysis of Various Classification Schemes. *Nucleic Acids Res.* **2014**, *42* (19), 151.
- (106) Gendron, P.; Lemieux, S.; Major, F. Quantitative Analysis of Nucleic Acid Three-Dimensional Structures. *J. Mol. Biol.* **2001**, *308* (5), 919–936.
- (107) Olson, W. K.; Esguerra, M.; Xin, Y.; Lu, X. J. New Information Content in RNA Base Pairing Deduced from Quantitative Analysis of High-Resolution Structures. *Methods*. March 2009, pp 177–186.
- (108) Kailasam, S.; Bhattacharyya, D.; Bansal, M. Sequence Dependent Variations in RNA Duplex Are Related to Non-Canonical Hydrogen Bond Interactions in Dinucleotide Steps. *BMC Res. Notes* **2014**, *7* (1), 83.
- (109) Lu, X.-J. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, *31* (17), 5108–5121.

- (110) Olson, W. K.; Li, S.; Kaukonen, T.; Colasanti, A. V.; Xin, Y.; Lu, X.-J. Effects of Noncanonical Base Pairing on RNA Folding: Structural Context and Spatial Arrangements of G·A Pairs. *Biochemistry* **2019**, acs.biochem.9b00122.
- (111) Mukherjee, S.; Bansal, M.; Bhattacharyya, D. Conformational Specificity of Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *J. Comput. Aided. Mol. Des.* **2006**, *20* (10–11), 629–645.
- (112) Lu, X. J.; Babcock, M. S.; Olson, W. K. Overview of Nucleic Acid Analysis Programs. *J. Biomol. Struct. Dyn.* **1999**, *16* (4), 833–843.
- (113) Lu, X. J.; Olson, W. K. Resolving the Discrepancies among Nucleic Acid Conformational Analyses. *J. Mol. Biol.* **1999**, *285* (4), 1563–1575.
- (114) Bhattacharya, S.; Jhunjhunwala, A.; Halder, A.; Bhattacharyya, D.; Mitra, A. Going beyond Base-Pairs: Topology-Based Characterization of Base-Multiplets in RNA. *RNA* **2019**, *25* (5), 573–589.
- (115) Bansal, M.; Bhattacharyya, D.; Ravi, B. NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Bioinformatics* **1995**, *11* (3), 281–287.
- (116) Pingali, P. K.; Halder, S.; Mukherjee, D.; Basu, S.; Banerjee, R.; Choudhury, D.; Bhattacharyya, D. Analysis of Stacking Overlap in Nucleic Acid Structures: Algorithm and Application. *J. Comput. Aided. Mol. Des.* **2014**, *28* (8), 851–867.
- (117) Parisien, M.; Major, F. The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data. *Nature* **2008**, *452* (7183), 51–55.

- (118) Zok, T.; Antczak, M.; Zurkowski, M.; Popenda, M.; Blazewicz, J.; Adamiak, R. W.; Szachniuk, M. RNAPdbec 2.0: Multifunctional Tool for RNA Structure Annotation. *Nucleic Acids Res.* **2018**, *46* (W1), W30–W35.
- (119) Popenda, M.; Szachniuk, M.; Antczak, M.; Purzycka, K. J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R. W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40* (14), e112–e112.
- (120) Boniecki, M. J.; Lach, G.; Dawson, W. K.; Tomala, K.; Lukasz, P.; Soltysinski, T.; Rother, K. M.; Bujnicki, J. M. SimRNA: A Coarse-Grained Method for RNA Folding Simulations and 3D Structure Prediction. *Nucleic Acids Res.* **2015**, *44* (7), 63.
- (121) Liu, P.; Li, L.; Millership, J. J.; Kang, H.; Leibowitz, J. L.; Giedroc, D. P. A U-Turn Motif-Containing Stem-Loop in the Coronavirus 5' Untranslated Region Plays a Functional Role in Replication. *RNA* **2007**, *13* (5), 763–780.
- (122) Leontis, N. B.; Zirbel, C. L. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In *RNA 3D Structure Analysis and Prediction*; 2012; Vol. 27, pp 281–298.
- (123) Hanson, R. M. Jmol-a Paradigm Shift in Crystallographic Visualization. *J. Appl. Crystallogr.* **2010**, *43* (5 PART 2), 1250–1260.
- (124) Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version~1.8*; 2015.
- (125) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. Gaussian16 Revision A.03. Gaussian Inc., Wallingford, CT **2016**.

- (126) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96* (9), 6796–6806.
- (127) Dunning Jr, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (1989), 1007.
- (128) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
- (129) Gehring, K.; Leroy, J. L.; Guéron, M. A Tetrameric DNA Structure with Protonated Cytosine-Cytosine Base Pairs. *Nature* **1993**, *363* (6429), 561–565.
- (130) Gu, Y.; Kar, T.; Scheiner, S. Fundamental Properties of the CH  $\cdots$  O Interaction: Is It a True Hydrogen Bond? *J. Am. Chem. Soc.* **1999**, *121* (40), 9411–9422.
- (131) Cubero, E.; Orozco, M.; Hobza, P.; Luque, F. J. Hydrogen Bond versus Anti-Hydrogen Bond: A Comparative Analysis Based on the Electron Density Topology. *J. Phys. Chem. A* **1999**, *103* (32), 6394–6401.
- (132) Halder, A.; Halder, S.; Bhattacharyya, D.; Mitra, A. Feasibility of Occurrence of Different Types of Protonated Base Pairs in RNA: A Quantum Chemical Study. *Phys. Chem. Chem. Phys.* **2014**, *16* (34), 18383–18396.
- (133) Kapinos, L. E.; Operschall, B. P.; Larsen, E.; Sigel, H. Understanding the Acid-Base Properties of Adenosine: The Intrinsic Basicities of N1, N3 and N7. *Chem. - A Eur. J.* **2011**, *17* (29), 8156–8164.

- (134) Jang, S. B.; Hung, L. W.; Chi, Y. I.; Holbrook, E. L.; Carter, R. J.; Holbrook, S. R. Structure of an RNA Internal Loop Consisting of Tandem C-A<sup>+</sup> Base Pairs. *Biochemistry* **1998**, *37* (34), 11726–11731.
- (135) Ma, J. C.; Dougherty, D. A. The Cation- $\pi$  Interaction. *Chem. Rev.* **1997**, *97* (5), 1303–1324.
- (136) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *J. Math. Biol.* **2008**, *56* (1–2), 215–252.
- (137) Elgavish, T.; Cannone, J. J.; Lee, J. C.; Harvey, S. C.; Gutell, R. R. AA.AG@helix.Ends: A:A and A:G Base-Pairs at the Ends of 16 S and 23 S rRNA Helices. *J. Mol. Biol.* **2001**, *310* (4), 735–753.
- (138) Pley, H. W.; Flaherty, K. M.; McKay, D. B. Three-Dimensional Structure of a Hammerhead Ribozyme. *Nature*. 1994, pp 68–74.
- (139) Bhattacharya, S.; Mittal, S.; Panigrahi, S.; Sharma, P.; S. P., P.; Paul, R.; Halder, S.; Halder, A.; Bhattacharyya, D.; Mitra, A. RNABP COGEST: A Resource for Investigating Functional RNAs. *Database* **2015**, *2015* (0), bav011–bav011.
- (140) Halder, S.; Bhattacharyya, D. RNA Structure and Dynamics: A Base Pairing Perspective. *Prog. Biophys. Mol. Biol.* **2013**, *113* (2), 264–283.
- (141) Halder, S.; Bhattacharyya, D. Structural Stability of Tandemly Occurring Noncanonical Basepairs within Double Helical Fragments: Molecular Dynamics Studies of Functional RNA. *J. Phys. Chem. B* **2010**, *114* (44), 14028–14040.

- (142) Panigrahi, S.; Pal, R.; Bhattacharyya, D. Structure and Energy of Non-Canonical Basepairs: Comparison of Various Computational Chemistry Methods with Crystallographic Ensembles. *J. Biomol. Struct. Dyn.* **2011**, *29* (3), 541–556.
- (143) Bandyopadhyay, D.; Bhattacharyya, D. Estimation of Strength in Different Extra Watson–Crick Hydrogen Bonds in DNA Double Helices through Quantum Chemical Studies. *Biopolymers* **2006**, *83* (3), 313–325.
- (144) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix. *Nucleic Acids Res.* **2006**, *34* (2), 564–574.
- (145) Šponer, J.; Šponer, J. E.; Mládek, A.; Jurečka, P.; Banáš, P.; Otyepka, M. Nature and Magnitude of Aromatic Base Stacking in DNA and RNA: Quantum Chemistry, Molecular Mechanics, and Experiment. *Biopolymers*. 2013, pp 978–988.
- (146) Hunter, C. A.; Lu, X.-J. DNA Base-Stacking Interactions: A Comparison of Theoretical Calculations with Oligonucleotide X-Ray Crystal Structures. *J. Mol. Biol.* **1997**, *265* (5), 603–619.
- (147) Hobza, P.; Šponer, J. Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations. *Chem. Rev.* **1999**, *99* (11), 3247–3276.
- (148) Hunter, R. S.; van Mourik, T. DNA Base Stacking: The Stacked Uracil/Uracil and Thymine/Thymine Minima. *J. Comput. Chem.* **2012**, *33* (27), 2161–2172.
- (149) Mukherjee, S.; Kailasam, S.; Bansal, M.; Bhattacharyya, D. Energy Hyperspace for Stacking Interaction in AU/AU Dinucleotide Step: Dispersion-Corrected Density

- Functional Theory Study. *Biopolymers* **2014**, *101* (1), 107–120.
- (150) Xia, T.; SantaLucia, J.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson - Crick Base Pairs. *Biochemistry* **1998**, *37* (42), 14719–14735.
- (151) Marky, L. A.; Breslauer, K. J. Calorimetric Determination of Base-Stacking Enthalpies in Double-Helical DNA Molecules. *Biopolymers* **1982**, *21* (11), 2185–2194.
- (152) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Naturally Occurring RNA Single Mismatches with G-U Nearest Neighbors. *Biochemistry* **2008**, *47* (38), 10178–10187.
- (153) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Single Mismatches Found in Naturally Occurring RNA. *Biochemistry* **2007**, *46* (46), 13425–13436.
- (154) Christiansen, M. E.; Znosko, B. M. Thermodynamic Characterization of Tandem Mismatches Found in Naturally Occurring RNA. *Nucleic Acids Res.* **2009**, *37* (14), 4696–4706.
- (155) Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion Interactions in Density-Functional Theory. *J. Phys. Org. Chem.* **2009**, *22* (12), 1127–1135.
- (156) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. Assessment of the Performance of the M05–2X and M06–2X Exchange-Correlation Functionals for Noncovalent Interactions in Biomolecules. *J. Chem. Theory Comput.* **2008**, *4* (12), 1996–2000.
- (157) El Hassan, M. A.; Calladine, C. R. Conformational Characteristics of DNA: Empirical

- Classifications and a Hypothesis for the Conformational Behaviour of Dinucleotide Steps. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **1997**, 355 (1722), 43–100.
- (158) Peterson, K. A.; Kendall, R. A.; Dunning, Jr., T. H. Benchmark Calculations with Correlated Molecular Wave Functions. II. Configuration Interaction Calculations on First Row Diatomic Hydrides. *J. Chem. Phys.* **1993**, 99 (1993), 1930.
- (159) Kendall, R. a; Dunning Jr., T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, 96 (May 2013), 6796.
- (160) Huber, R. G.; Margreiter, M. A.; Fuchs, J. E.; von Grafenstein, S.; Tautermann, C. S.; Liedl, K. R.; Fox, T. Heteroaromatic  $\pi$ -Stacking Energy Landscapes. *J. Chem. Inf. Model.* **2014**, 54 (5), 1371–1379.
- (161) Aragó, J.; Sancho-García, J. C.; Ortí, E.; Beljonne, D. Ab Initio Modeling of Donor-Acceptor Interactions and Charge-Transfer Excitations in Molecular Complexes: The Case of Terthiophene- Tetracyanoquinodimethane. *J. Chem. Theory Comput.* **2011**, 7 (7), 2068–2077.
- (162) MJ Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, V Barone, B Mennucci, GA Petersson, H Nakatsuji, M Caricato, X Li, HP Hratchian, AF Izmaylov, J Bloino, G Zheng, JL Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Haseg, D. F. Gaussian09, Revision B.01. *Gaussian Inc., Wallingford, CT* **2010**.
- (163) Šponer, J.; Jurečka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, 126 (32), 10142–10151.

- (164) Bhattacharyya, D. Bioinformatics Resource Materials developed at SINP  
<http://www.saha.ac.in/biop/bioinformatics.html>.
- (165) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.
- (166) Poater, J.; Swart, M.; Bickelhaupt, F. M.; Fonseca Guerra, C.; Lewis, F. D.; Berlin, Y. A.; Ratner, M. A.; Siebbeles, L. D. A.; Nibbering, E. T. J.; Elsaesser, T. B-DNA Structure and Stability: The Role of Hydrogen Bonding,  $\pi$ - $\pi$  Stacking Interactions, Twist-Angle, and Solvation. *Org. Biomol. Chem.* **2014**, *12* (26), 4691–4700.
- (167) Šponer, J.; Leszczyński, J.; Hobza, P. Nature of Nucleic Acid–Base Stacking: Nonempirical Ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. *J. Phys. Chem.* **1996**, *100* (13), 5590–5596.
- (168) Ananth, P.; Goldsmith, G.; Yathindra, N. An Innate Twist between Crick’s Wobble and Watson-Crick Base Pairs. *RNA* **2013**, *19* (8), 1038.
- (169) Mokdad, A. Structural and Evolutionary Classification of G/U Wobble Basepairs in the Ribosome. *Nucleic Acids Res.* **2006**, *34* (5), 1326–1341.
- (170) Leontis, N. B.; Stombaugh, J.; Westhof, E. The Non-Watson-Crick Base Pairs and Their Associated Isostericity Matrices. *Nucleic Acids Res.* **2002**, *30* (16), 3497–3531.
- (171) Leontis, N. B.; Westhof, E. Geometric Nomenclature and Classification of RNA Base Pairs. *RNA* **2001**, *7* (4), 499–512.

- (172) Baeyens, K. J.; De Bondt, H. L.; Pardi, A.; Holbrook, S. R. A Curved RNA Helix Incorporating an Internal Loop with G·A and A·A Non-Watson-Crick Base Pairing. *Proc. Natl. Acad. Sci.* **1996**, *93* (23), 12851–12855.
- (173) Doudna, J. A.; Cate, J. H. RNA Structure: Crystal Clear? *Curr. Opin. Struct. Biol.* **1997**, *7* (3), 310–316.
- (174) Spackova, N.; Šponer, J. Molecular Dynamics Simulations of Sarcin-Ricin RRNA Motif. *Nucleic Acids Res.* **2006**, *34* (2), 697–708.
- (175) Kondo, J.; Westhof, E. The Bacterial and Mitochondrial Ribosomal A-Site Molecular Switches Possess Different Conformational Substates. *Nucleic Acids Res.* **2008**, *36* (8), 2654–2666.
- (176) Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. MiRBase: From MicroRNA Sequences to Function. *Nucleic Acids Res.* **2019**, *47* (D1), D155–D162.
- (177) Réblová, K.; Střelcová, Z.; Kulhánek, P.; Beššeová, I.; Mathews, D. H.; Van Nostrand, K.; Yildirim, I.; Turner, D. H.; Šponer, J. An RNA Molecular Switch: Intrinsic Flexibility of 23S RRNA Helices 40 and 68 5'-UAA/5'-GAN Internal Loops Studied by Molecular Dynamics Methods. *J. Chem. Theory Comput.* **2010**, *6* (3), 910–929.
- (178) Halder, S.; Bhattacharyya, D. Structural Variations of Single and Tandem Mismatches in RNA Duplexes: A Joint MD Simulation and Crystal Structure Database Analysis. *J. Phys. Chem. B* **2012**, *116* (39), 11845–11856.
- (179) Halder, S.; Bhattacharyya, D. Structural Stability of Tandemly Occurring Noncanonical Basepairs within Double Helical Fragments: Molecular Dynamics Studies of Functional

- RNA. *J. Phys. Chem. B* **2010**, *114* (44), 14028–14040.
- (180) Kruse, H.; Banáš, P.; Šponer, J. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.* **2019**, *15* (1), 95–115.
- (181) Egli, M.; Usman, N.; Rich, A. Conformational Influence of the Ribose 2'-Hydroxyl Group: Crystal Structures of DNA-RNA Chimeric Duplexes. *Biochemistry* **1993**, *32* (13), 3221–3237.
- (182) Ban, C.; Ramakrishnan, B.; Sundaralingam, M. A Single 2'-Hydroxyl Group Converts B-DNA to A-DNA: Crystal Structure of the DNA-RNA Chimeric Decamer Duplex d(CCGGC)r(G)d(CCGG) with a Novel Intermolecular G·C Base-Paired Quadruplet. *J. Mol. Biol.* **1994**, *236* (1), 275–285.
- (183) Lindqvist, M.; Sarkar, M.; Winqvist, A.; Rozners, E.; Strömberg, R.; Gräslund, A. Optical Spectroscopic Study of the Effects of a Single Deoxyribose Substitution in a Ribose Backbone: Implications in RNA-RNA Interaction. *Biochemistry* **2000**, *39* (7), 1693–1701.
- (184) Gang Chen, §; Scott D. Kennedy, ¶; Jing Qiao, §; Thomas R. Krugh, § and; Douglas H. Turner\*, §,⊥. An Alternating Sheared AA Pair and Elements of Stability for a Single Sheared Purine-Purine Pair Flanked by Sheared GA Pairs in RNA†,‡. **2006**.
- (185) Chakraborty, D.; Wales, D. J. Dynamics of an Adenine-Adenine RNA Conformational Switch from Discrete Path Sampling. *J. Chem. Phys.* **2019**, *150* (12), 125101.
- (186) Bansal, M.; Bhattacharyya, D.; Ravi, B. NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Comput. Appl. Biosci.*

- 1995, *11* (3), 281–287.
- (187) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.
- (188) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. A Direct MP2 Gradient Method. *Chem. Phys. Lett.* **1990**, *166* (3), 275–280.
- (189) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102* (11), 1995–2001.
- (190) Chandrasekaran, R.; Arnott, S. The Structure of *B*-DNA in Oriented Fibers. *J. Biomol. Struct. Dyn.* **1996**, *13* (6), 1015–1027.
- (191) Pérez, A.; Marchán, I.; Svozil, D.; Šponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of Alpha/Gamma Conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829.
- (192) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6* (12), 3836–3849.
- (193) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. Refinement of the Cornell et Al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*

- 2011**, 7 (9), 2886–2902.
- (194) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, 1, 19–25.
- (195) Darden, T.; Pearlman, D.; Pedersen, L. G. Ionic Charging Free Energies: Spherical versus Periodic Boundary Conditions. *J. Chem. Phys.* **1998**, 109 (24), 10921–10935.
- (196) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, 52 (12), 7182–7190.
- (197) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, 126 (1), 014101.
- (198) Das, A.; Chakrabarti, J.; Ghosh, M. Conformational Contribution to Thermodynamics of Binding in Protein-Peptide Complexes through Microscopic Simulation. *Biophys. J.* **2013**, 104 (6), 1274–1284.
- (199) Sarkar, S.; Maity, A.; Sarma Phukon, A.; Ghosh, S.; Chakrabarti, R. Salt Induced Structural Collapse, Swelling, and Signature of Aggregation of Two SsDNA Strands: Insights from Molecular Dynamics Simulation. *J. Phys. Chem. B* **2019**, 123 (1), 47–56.
- (200) Saenger, W. *Principles of Nucleic Acid Structure*; Springer Advanced Texts in Chemistry; Springer New York: New York, NY, 1984.
- (201) Romanowska, J.; McCammon, J. A.; Trylska, J. Understanding the Origins of Bacterial Resistance to Aminoglycosides through Molecular Dynamics Mutational Study of the Ribosomal A-Site. *PLoS Comput. Biol.* **2011**, 7 (7), e1002099.

- (202) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen Base Pairs in Canonical Duplex DNA. *Nature* **2011**, *470* (7335), 498–502.
- (203) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational Analysis of Nucleic Acids Revisited: Curves+. *Nucleic Acids Res.* **2009**, *37* (17), 5917–5929.
- (204) Réblová, K.; Střelcová, Z.; Kulhánek, P.; Beššeová, I.; Mathews, D. H.; Van Nostrand, K.; Yildirim, I.; Turner, D. H.; Šponer, J. An RNA Molecular Switch: Intrinsic Flexibility of 23S rRNA Helices 40 and 68 5'-UAA/5'-GAN Internal Loops Studied by Molecular Dynamics Methods. *J. Chem. Theory Comput.* **2010**, *6* (3), 910–929.
- (205) Brown, R. F.; Andrews, C. T.; Elcock, A. H. Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *J. Chem. Theory Comput.* **2015**, *11* (5), 2315–2328.
- (206) Häse, F.; Zacharias, M. Free Energy Analysis and Mechanism of Base Pair Stacking in Nicked DNA. *Nucleic Acids Res.* **2016**, *44* (15), 7100–7108.
- (207) Nelson, H. C. M.; Finch, J. T.; Luisi, B. F.; Klug, A. The Structure of an Oligo(DA)·oligo(DT) Tract and Its Biological Implications. *Nature* **1987**, *330* (6145), 221–226.
- (208) Bandyopadhyay, D.; Bhattacharyya, D. Estimation of Strength in Different Extra Watson–Crick Hydrogen Bonds in DNA Double Helices through Quantum Chemical Studies. *Biopolymers* **2006**, *83* (3), 313–325.

- (209) Frank-Kamenetskii, M. D.; Mirkin, S. M. Triplex DNA Structures. *Annu. Rev. Biochem.* **1995**, *64* (1), 65–95.
- (210) Garg, A.; Heinemann, U. A Novel Form of RNA Double Helix Based on G·U and C·A+ Wobble Base Pairing. *RNA* **2018**, *24* (2), 209–218.
- (211) Chawla, M.; Sharma, P.; Halder, S.; Bhattacharyya, D.; Mitra, A. Protonation of Base Pairs in RNA: Context Analysis and Quantum Chemical Investigations of Their Geometries and Stabilities. *J. Phys. Chem. B* **2011**, *115* (6), 1469–1484.
- (212) Halder, A.; Halder, S.; Bhattacharyya, D.; Mitra, A. Feasibility of Occurrence of Different Types of Protonated Base Pairs in RNA: A Quantum Chemical Study. *Phys. Chem. Chem. Phys.* **2014**, *16* (34), 18383–18396.
- (213) Bernhardt, H. S.; Tate, W. P. Primordial Soup or Vinaigrette: Did the RNA World Evolve at Acidic PH? *Biol. Direct* **2012**, *7* (1), 4.
- (214) Wang, B. The RNA I-Motif in the Primordial RNA World. *Orig. Life Evol. Biosph.* **2019**, *49* (1–2), 105–109.
- (215) Peng, C.; Bernhard Schlegel, H. Combining Synchronous Transit and Quasi-Newton Methods to Find Transition States. *Isr. J. Chem.* **1993**, *33* (4), 449–454.
- (216) Herzog, E.; Frigato, T.; Helms, V.; Lancaster, C. R. D. Energy Barriers of Proton Transfer Reactions between Amino Acid Side Chain Analogs and Water Fromab Initio Calculations. *J. Comput. Chem.* **2006**, *27* (13), 1534–1547.
- (217) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

- (218) Arnott, S.; Hukins, D. W. L.; Dover, S. D. Optimised Parameters for RNA Double-Helices. *Biochem. Biophys. Res. Commun.* **1972**, *48* (6), 1392–1399.

## Thesis Highlight

**Name of the Student:** *Satyabrata Maiti*

**Name of the CIOCC:** *Saha Institute of Nuclear Physics* Enrolment No.: *LIFE05201504004*

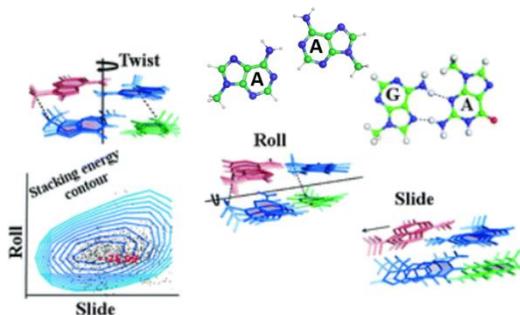
**Thesis Title:** *Non-canonical Base Pairs in RNA and their involvement in double helix formation.*

**Discipline:** *Life Sciences*

**Sub-Area of Discipline:** *RNA Structural Bioinformatics*

Though RNA has almost similar structural elements as DNA, it can regulate huge number of diverse cellular processes. Apart from the regular helical regions with Watson-Crick base pairs there exists a variety of structural motifs like, pseudo helices, hairpin structural elements with variable loop region, internal loops, multi-way junctions, bulges and many more. While proteins use twenty amino acids of various properties to perform different tasks through molecular recognition, RNA possesses only four nucleotide bases. These four bases can pair in various different ways, which are commonly referred to as non-canonical base pairs, to offer various steric and electrostatics shapes for molecular recognition.

We have presented a combined network of web-servers and database (<http://hdrnas.saha.ac.in/>) for RNA structural perspective. We have additionally attempted to understand stability of these unusual contacts using Density Functional Theory based studies. (<http://hdrnas.saha.ac.in/rnabpdb>) This indicated sufficient strengths of interactions in the



*Stacking Energy scan of dinucleotide step with one non-canonical base pair in Roll-Slide hyperspace for different Twist angle*

non-canonical base pairs and their stacks. We have found G:A S:HT base pair as one of the most frequent non-canonical base pair. We have developed a quantum chemical hybrid DFT-D based stacking energy analysis, which enables us to find the most probable configuration of the dinucleotide step containing a non-canonical base pair (G:A S:HT :: C:G W:WC). Using the above-described work as a benchmark study, we have tried to decode the promiscuous nature A:A w:wC base pair where one of the hydrogen bond is weak C-H...N mediated. Molecular dynamics simulation coupled with quantum chemical calculation and bioinformatics study ultimately dig out the hidden truth behind the promiscuous nature A:A w:wC base pair when it is stacked on sheared G:U W:WC base pair. We have later extend our work on non-canonical base pair to find the most probable configuration SL1 helix of RNA of corona virus. In the SL1 motif, either A:C +:WC or U:C W:+C base pairs are possible. Our hybrid stacking energy analysis along with transition state calculation, have supported on protonation of Adenine residue in A:C +:WC base pair.