Understanding structural features of biomolecular interactions : from classical simulations to ab initio calculations

Hartono, Yossa Dwi

2017

Hartono, Y. D. (2017). Understanding structural features of biomolecular interactions : from classical simulations to ab initio calculations. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/72892>

<https://doi.org/10.32657/10356/72892>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**UNDERSTANDING STRUCTURAL FEATURES OF
BIOMOLECULAR INTERACTIONS:
FROM CLASSICAL SIMULATIONS TO *AB INITIO*
CALCULATIONS**

YOSSA DWI HARTONO

SCHOOL OF BIOLOGICAL SCIENCES

2017

From DEPARTMENT OF BIOSCIENCES AND NUTRITION
Karolinska Institutet, Stockholm, Sweden

**UNDERSTANDING
STRUCTURAL FEATURES OF
BIOMOLECULAR INTERACTIONS:
FROM CLASSICAL SIMULATIONS
TO *AB INITIO* CALCULATIONS**

Yossa Dwi Hartono

**Karolinska
Institutet**

Stockholm 2017

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

Printed by E-Print AB 2017

© Yossa Dwi Hartono, 2017

ISBN 978-91-7676-829-7

Understanding structural features of biomolecular
interactions: from classical simulations to *ab initio*
calculations
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Yossa Dwi Hartono

Principal Supervisor:

Assoc. Prof. Alessandra Villa
Karolinska Institutet
Department of Biosciences and Nutrition

Co-supervisors:

Prof. Lennart Nilsson
Karolinska Institutet
Department of Biosciences and Nutrition

Assoc. Prof. Konstantin Pervushin
Nanyang Technological University
School of Biological Sciences
Division of Structural Biology and Biochemistry

Opponent:

Prof. Carmay Lim
Academia Sinica
Institute of Biomedical Sciences
&
National Tsing Hua University
Department of Chemistry

Examination Board:

Prof. Jayaraman Sivaraman
National University of Singapore
Department of Biological Sciences

Assoc. Prof. Curtis Alexander Davey
Nanyang Technological University
School of Biological Sciences
Division of Structural Biology and Biochemistry

Dr. Peter John Bond
Agency for Science, Technology, and Research
Bioinformatics Institute
&
National University of Singapore
Department of Biological Sciences

To knowledge

“

quaecumque sunt vera

quaecumque pudica

quaecumque iusta

quaecumque sancta

quaecumque amabilia

quaecumque bonae famae

si qua virtus

si qua laus disciplinae

haec cogitate

– Phil. 4:8

ABSTRACT

The structures of biomolecules and their interactions dictate their functions. In this thesis, five papers are presented to illustrate how the dynamics of biomolecules can be investigated and derivation of desired thermodynamic quantities obtained by utilising a diverse range of computational techniques, from simulations utilising classical mechanical descriptions to calculations employing quantum mechanical descriptions.

Classical simulation, referring to molecular dynamics simulation with atomistic force fields, has been used in every paper in this thesis. In Paper I, classical simulation and homology modelling are used to investigate the dynamics of a protein as well as that of its homologues, which have a missing region. Protein purification and production of these homologues was also attempted.

When state transitions like protonation and tautomerisation equilibria are central to the query, we employed lambda-dynamics, an extension to conventional simulation that can describe transitions between states by including coupling parameter lambda in the dynamics. In Papers II and III, protonation and tautomerisation equilibria respectively are central to the query.

In Paper II, lambda-dynamics and multiple pH regime are both used to calculate the pK shifts of cytidine in triplex nucleic acid environments. In some of the triplex nucleic acid systems, sugar modification LNA (locked nucleic acid) is present. The force field parameters of LNA have been updated to provide better descriptions for pK calculations. In Paper III, lambda-dynamics is used to describe tautomerisation equilibrium between two tautomers of pseudoisocytidine in single-stranded and triple-stranded nucleic acids in order to observe how the equilibrium shifts in different environments. *In vitro* binding assay is used to corroborate the computational results.

When greater accuracy for certain properties like electrostatics or energetics is required, we employed quantum mechanical calculations as well as hybrid methods which combine classical and quantum mechanical descriptions. In Paper IV, QM and QM/MM calculations were performed to calculate the energetic difference between two tautomers in the ribosome. In Paper V, protein-specific polarised charge, a charge update scheme that updates the atomic charges with QM and Poisson-Boltzmann calculations during classical simulation, is used for better electrostatics description of a peptide.

LIST OF SCIENTIFIC PAPERS

- I. Hartono, Y.D.; Pervushin, Konstantin. *Megavirales* homologues of translation termination factor eRF1: protein production, homology modelling, and molecular dynamics. *Manuscript*.
- II. Hartono, Y.D.*; Xu, Y*; Karshikoff, A; Nilsson, L.; Villa, A. Modeling p*K* shift in DNA triplex containing locked nucleic acids. *Manuscript*.
- III. Hartono, Y.D.; Pabon-Martinez, Y.V.; Uyar, A.; Wengel, J.; Lundin, K.E.; Zain, R.; Smith, C.E.; Nilsson, L.; Villa, A. Role of Pseudoisocytidine Tautomerization in Triplex-Forming Oligonucleotides: In Silico and in Vitro Studies. *ACS Omega*, **2017**, 2(5), 2165-2177.
[DOI: 10.1021/acsomega.7b00347](https://doi.org/10.1021/acsomega.7b00347)
- IV. Hartono, Y.D.; Ito, M.; Villa, A; Nilsson, L. A computational study of uracil tautomeric forms on the ribosome: the case of uracil and 5-oxyacetic acid uracil in the first anticodon position of tRNA. *Manuscript*.
- V. Hartono, Y.D.; Yip, Y.M.; Zhang, D. Adsorption and folding dynamics of MPER of HIV-1 gp41 in the presence of DPC micelle. *Proteins: Structure, Function, and Bioinformatics*, **2013**, 81(6), 933-944.
[DOI: 10.1002/prot.24256](https://doi.org/10.1002/prot.24256)

* These authors contributed equally

CONTENTS

1	Introduction	5
2	Biomolecular systems	6
3	Three-dimensional structure	11
3.1	X-ray crystallography	11
3.2	NMR spectroscopy	11
3.3	Cryo-electron microscopy	12
3.4	Sample preparation	12
3.5	Computationally-sourced models	13
3.5.1	From homology modelling	13
3.5.2	From <i>de novo</i> protein folding	13
4	Biophysical properties	14
4.1	Binding and adsorption	14
4.2	Protonation	15
4.3	Tautomerisation	16
5	Molecular descriptions	17
5.1	Classical molecular mechanics descriptions	17
5.2	<i>Ab initio</i> descriptions	18
5.3	Hybrid descriptions	20
5.3.1	QM/MM	20
5.3.2	Polarised protein-specific charge	20
6	Simulation techniques	22
6.1	Molecular dynamics	22
6.2	Lambda-dynamics	23
6.3	Multiple pH regime	25
6.4	Free energy calculation	26
6.4.1	Bennett acceptance ratio	26
6.4.2	Potential of mean force	27
7	Summary, conclusion, and outlook	29
7.1	Summary	29
I.	<i>Megavirales</i> homologues of translation termination factor eRF1: protein production, homology modelling, and molecular dynamics	29
II.	Modeling p <i>K</i> shift in DNA triplex containing locked nucleic acids	30
III.	Role of Pseudoisocytidine Tautomerization in Triplex-Forming Oligonucleotides: In Silico and in Vitro Studies	31
IV.	A computational study of uracil tautomeric forms on the ribosome: the case of uracil and 5-oxyacetic acid uracil in the first anticodon position of tRNA	32
V.	Adsorption and folding dynamics of MPER of HIV-1 gp41 in the presence of DPC micelle	33
7.2	Conclusion and outlook	34
8	Popular science summary	36
	Acknowledgements	39
	References	41

LIST OF ABBREVIATIONS

BAR	Bennett Acceptance Ratio
bp	basepair
Cryo-EM	Cryo-electron microscopy
DNA	Deoxyribonucleic acid
DPC	Dodecylphosphocholine
eRF1	Eukaryotic release factor 1
FEP	Free energy perturbation
HG	Hoogsteen basepair
HIV	Human immunodeficiency virus
LNA	Locked nucleic acid
MD	Molecular dynamics
mRNA	Messenger RNA
NBB	Non-Boltzmann BAR
NMR	Nuclear magnetic resonance
PMF	Potential of mean force
QM/MM	Quantum mechanics/molecular mechanics
RNA	Ribonucleic acid
TFO	Triplex-forming oligonucleotide
tRNA	Transfer RNA
WC	Watson-Crick basepair
Ψ C	Pseudoisocytidine

1 INTRODUCTION

An important principle tenet of structural biology is that structure leads to function. The core questions that comes up yet and again in this thesis, or indeed in structural biology at large, is about structure: how a residue substitution will affect the structure, how different interactions will change the structure, how the structure explains the experimental observation, and so on.

In this thesis, a diverse set of computational tools has been utilised to answer these questions. Classical simulations, referring to simulating the system with classical mechanics, are used; but if the question requires greater level of accuracy, *ab initio* calculations, which describe the system with quantum mechanics, are employed, albeit at greater computational cost. Quantum mechanics consider the electrons explicitly, thus taking into account the electronic effects. As the title of this thesis suggests, not only these two ends of the spectrum are used, but also some others in-between.

The aim of this thesis is to understand biomolecules from the perspective of their structures — which give rise to biophysical properties such as state transitions, free energy, interactions with other molecules, among others — with various computational techniques.

This thesis frame has been structured to begin from more general concepts and progress to increasingly specialised ones:

The biomolecular systems are briefly introduced in Section 2. The common starting point of a computational biophysics investigation, the three-dimensional structure, is discussed in Section 3; followed by several biophysical properties of focus, in Section 4. Different levels of molecular descriptions are described in Section 5, then the various computational techniques used to sample the phase space and to derive the quantities of interest in Section 6. Finally, summaries of the papers, conclusions, and future outlook are presented in Section 7.

In addition, the layman summary is provided in Section 8. The author strongly believes in the importance of effective communications of scientific ideas to the lay public, and hopes that this summary makes this thesis more accessible to the general audience.

2 BIOMOLECULAR SYSTEMS

At the molecular level, life is composed of many types of molecules, such as proteins, nucleic acids, lipids, carbohydrates, small organic and inorganic molecules, co-factors, and others. While lipids and carbohydrates often form repeating assemblies, proteins and nucleic acids are long polymers that fold to specific shapes depending on their monomeric composition and the environment.

Proteins are composed of diverse combinations of over 20 different amino acids with varying biochemical properties: charged, polar, hydrophobic, acidic, basic, and others. Consequently, proteins, with their vast array of structures, perform many functions: catalysis, ion transport, receptor, messenger, among others.

Figure 1. MPER peptide adsorbed to a micelle.

An example of protein is MPER (membrane proximal ectodomain region) peptide,¹ a short peptide fragment of HIV envelope protein (Figure 1). The envelope protein plays a major role in HIV cell infection process by facilitating membrane fusion.² This is the protein of interest in Paper V.

Compared to proteins, nucleic acids are composed of fewer letters (residues) with less diverse biochemistry. The two types of nucleic acids, DNA and RNA, differ in the sugar moiety by a 2' hydroxyl group, which is present in RNA and absent in DNA. DNA, typically forming double-stranded helix, mainly functions as storage of genetic information. RNA, typically single-stranded with more diverse secondary structures, has diverse functions, for example: catalyst function in ribozyme, messenger in messenger RNA (mRNA) and transfer RNA (tRNA), scaffold and catalyst in ribosomal RNA.

The ribosome is a protein-producing machinery which itself is an assembly of proteins and RNA. The protein-producing process is called translation, during which genetic information from the mRNA is “translated” by the tRNA to polypeptides, which may undergo further post-translational processing to yield the functional proteins. The information in mRNA is read one nucleotide triplet (termed codon) at a time by the tRNA, which has the triplet with complementary sequence to the codon (termed anticodon). The process of mRNA reading during elongation stage of translation happens in a specific site in the ribosome called the decoding region in the aminoacyl site. The ribosome has three sites for tRNA: aminoacyl site (A-site) where it holds the incoming tRNA with the correct anticodon binding to the mRNA codon; peptidyl site (P-site) where it holds tRNA with growing polypeptide; exit site (E-site) where it holds the exiting tRNA without its amino acid charge.

Figure 2. Subset of the ribosomal decoding region from X-ray crystal structure. tRNA ASL refers to the tRNA anticodon stem loop. Inset shows $cmo^5U:G$ basepair in Watson-Crick geometry

The ribosomal scaffold ensures the fidelity of translation by making non-complementary tRNA-mRNA interactions unfavourable. This selection is tighter for the first and second codon positions, while some mismatches are allowed for the third position.³⁻⁴ One example is G:U mismatch which typically form the wobble configuration with two hydrogen bonds, instead of three in a complementary Watson-Crick G:C match. In Paper IV, we are interested

in a particular interaction in the third position, where a modified U:G basepair is observed to have Watson-Crick geometry in the X-ray crystal structure (Figure 2).

Figure 3. The structures of human eRF1 (X-ray) and vRF1s (homology models).

Besides coding for the amino acids, the codons also signal the start and the end of translation. Codons that signal for termination of translation are called stop codons, and instead of being read by tRNA, a protein factor recognises the stop codon instead. In eukaryotes, this protein is called eukaryotic release factor 1 (eRF1).⁵ eRF1 and its homologues found in giant viruses, are the subject of interest in Paper I (Figure 3).

Although the double helix is the typical conformation of DNA, in particular circumstances there is enough steric space to accommodate a third strand to form a triplex. When a double helix with Watson-Crick basepairing is formed between a homopurine strand and a homopyrimidine strand, a third homopyrimidine strand can bind at the major groove of the double helix via Hoogsteen basepairing to form a parallel triple helix (Figure 4). With canonical DNA bases, the possible base triads are $C^+ \cdot G-C$ and $T \cdot A-T$ (‘-’ refers to Watson-Crick and ‘•’ to Hoogsteen base pair), where cytidine in the third strand needs to be protonated to form Hoogsteen hydrogen bond. The parallel triple helix system is discussed in Papers II and III. When discussing a triple helical nucleic acids system, it is common to refer

to the third strand as the triplex-forming oligonucleotides (TFO). The conformation of the duplex bound to the TFO is known to be close to that of A-form geometry.⁶⁻⁸

Figure 4. Parallel triple helix and base triads, shown with Watson-Crick (WC) and Hoogsteen (HG) hydrogen bonding.

Although there are only 4 basic letters for DNA and RNA that are used in the genetic code, there are various modifications, natural and artificial, of the bases that play important roles in various translational and epigenetical processes. The tRNA, for example, contains many modifications that modulate its codon-reading activities.⁹⁻¹⁰ A natural DNA modification of cytosine, 5-methylation,¹¹ is mentioned in Papers II and III. A natural RNA base modification of uridine at the first anticodon position, 5-oxyacetic acid uridine (cmo⁵U) (Figure 5), is mentioned in Paper IV. With this modification at first anticodon position, it can accept all four bases.¹² There are also various artificial modified bases such as pseudoisocytosine (Ψ C),¹³ (Figure 5) derived from natural modified base pseudouridine, that is incorporated in a TFO to form a more stable triplex nucleic acids, mentioned in Paper III.

Figure 5. Chemical structures of two modified bases, shown with the corresponding canonical bases. (A) Cytidine and pseudoisocytidine (B) Uridine and 5-oxyacetic acid uridine

An artificial sugar modification, locked nucleic acid¹⁴⁻¹⁵ (LNA, see Figure 6), is also mentioned in Papers II and III, where it is also used in TFO to improve duplex binding to form the triplex.¹⁶ With LNA modification, the sugar pucker is fixed to C3'-*endo* (also called North conformation), which is prevalent in A-form geometry.¹⁴

Figure 6. Chemical structures of LNA and DNA.

3 THREE-DIMENSIONAL STRUCTURE

The three-dimensional coordinates of atoms of the biomolecule of interest is the starting point for a computational biophysics investigation. These coordinates are by large resolved experimentally (while some are modelled computationally) and are typically deposited in public repositories such as the Protein Data Bank (PDB) (rcsb.org)¹⁷ and the Electron Microscopy Data Bank (EMDB) (emdatabank.org).¹⁸ The three main experimental techniques to resolve the three-dimensional structure of a biomolecule are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). The sample preparation stage needed for these techniques is also discussed, with special focus on protein production and purification.

3.1 X-RAY CRYSTALLOGRAPHY

X-ray crystallography makes use of the diffraction patterns obtained by beaming high-intensity X-ray at a crystallised biomolecule (see Ref. ¹⁹ for a short review). Diffraction occurs due to the molecules oriented in repeating manner. Back-calculation of the diffraction patterns can yield back the atomic coordinates of the molecule.

In X-ray crystallography structures, hydrogen atoms are not visible below certain resolution because of their low electron density. Also, flexible parts of the molecule often cannot be resolved since they may end up in slightly different positions in the crystal and do not give rise to good diffraction. Crystallisation condition is also highly individualised for each protein and conditions at which a protein will crystallise well are often determined by trial and error, i.e. by screening in thousands of conditions. Another problem is that due to close contact of the molecules in a crystal, the environment may be vastly different compared to the molecule in solution, which is the typical physiological environment of a soluble protein.

3.2 NMR SPECTROSCOPY

Structural determination by solution NMR spectroscopy²⁰ uses mainly measured nuclear Overhauser effect (NOE) nucleus-nucleus distances, which are used to restrain the initial model of the molecule in a narrow scope of conformations.²¹ The coordinates are therefore typically deposited as an ensemble of structures that are lowest in energy or have least restraint violations. Also, since the atomistic model is already pre-built, an NMR structure would not have missing atoms like an X-ray structure might. Other NMR measurements, such as residual dipolar coupling, can also be inputted as additional restraints in structure

calculation. The NMR data processing relies on computational tools to interpret the data and evaluate the biomolecular structure.²²

Since the sample is in solution, the biomolecule can also be measured in its close-to-native environment, although solid state NMR spectroscopy for biomolecules is now an active area of research. However, a biomolecule might only be highly soluble in certain conditions that might not be close to its native physiological environment, since NMR spectroscopy requires quite high concentration of the biomolecule to yield good signal-to-noise ratio. Another limitation is that solution NMR spectroscopy can only be typically used up to certain biomolecular size as the signal will degrade faster the bigger the size. Development of NMR pulse sequences to increase the size limit, for instance that used in transverse-relaxation optimised spectroscopy (TROSY),²³ is an ongoing effort.

3.3 CRYO-ELECTRON MICROSCOPY

In cryo-electron microscopy, the biomolecule is suspended in thin layer of vitreous ice, and electron micrographs are taken. Image sorting algorithm is then used to sort the images of the molecule in different orientations and to construct the three-dimensional model. Vitreous ice ensures that the molecule is in close to solution environment. Cryo-EM have been used for large particles such as the ribosome,²⁴⁻²⁷ that may be more challenging to obtain in good yield or to crystallise for X-ray measurement. The purity requirement is also less strict than that of X-ray crystallography since contaminants can be graphically sorted out. Cryo-EM has been on the rise in the recent years, as it is approaching atomic resolution.²⁸

3.4 SAMPLE PREPARATION

In order to perform these measurements, the biomolecule sample needs first to be prepared. For small peptides and oligonucleotides, chemical synthesis may be used. For proteins, they may be purified directly from the source, but recombinant technology may be used to increase the yield. For example, in Paper I, we cloned the gene of interest in *E. coli* and cultured it to overexpress the protein of interest. Other hosts such as yeast, insect cells, or human cells might be used as well, depending on the suitability with the protein. For example, *E. coli* may not be appropriate for proteins for which post-translational modifications are important for the investigation.

Let us consider a typical workflow of recombinant protein production and purification.²⁹ First, the gene should be cloned in an appropriate vector and sequence-verified. The host is then transfected, cultured, and induced at appropriate times to produce the protein. The cells are harvested and centrifuged, then resuspended in buffer and agitated to break the cell

membranes, then centrifuged again. For soluble proteins, the supernatant is subjected to purification stage.

For protein, it is a common practice to include an affinity tag covalently bound to the recombinant protein. Proteins containing hexahistidine tag,³⁰ for instance, can be purified in the first stage with immobilised metal affinity chromatography (IMAC) with Ni²⁺ or Co²⁺ ion column. More chromatographic techniques typically follow, depending on the criteria of purity. Size-exclusion chromatography (SEC) is often a choice due to its versatility. Other columns, for instance, ion-exchange chromatography (IEX), may need calibration trials to determine the optimal ionic strength gradient.

Protein production and purification is often challenging, because although standardised workflow exists, it needs to be customised and optimised specific to the protein of interest. Here are some considerations other than those already mentioned: sequence construct – beginning or ending amino acids may affect the overall solubility and expression of the protein; difficulty level of maintaining host culture; buffer composition and pH.

3.5 COMPUTATIONALLY-SOURCED MODELS

3.5.1 From homology modelling

As the PDB contains many deposited structures, they can be used as templates to resolve the structure of homologous sequence. For example, certain sequence might always form an α -helix, so the structure of a new (never structurally resolved) protein containing this sequence can be predicted to be an α -helix. Homology modelling has low computational cost, but the validation stage is challenging. Many homology modelling software suites and servers are available, such as MODELLER,³¹⁻³⁴ SWISS-MODEL,³⁵⁻³⁸ and I-TASSER.³⁹⁻⁴¹

3.5.2 From *de novo* protein folding

In *de novo* protein folding,⁴² one starts from linear conformation of the molecule, and let the force field and molecular dynamics simulation or other computational techniques 'fold' it into lower energy conformation, which should ideally correspond to the experimental structure.

4 BIOPHYSICAL PROPERTIES

The aim of a computational biophysics investigation is often to obtain biophysical properties, qualitatively or quantitatively, for comparison with experiments, or to gain insights into processes which are experimentally intractable. In this section three processes of interest – binding, protonation, and tautomerisation – are discussed.

4.1 BINDING AND ADSORPTION

Binding interactions are specific interactions between molecules, often designated as a ligand and a target. Fischer's lock-and-key model⁴³ is a useful metaphor to illustrate the specificity criterion of a binding event. A target molecule has specific moieties with specific orientations, and the ligand molecule has to match these in order to bind. A protein-water interaction, for example, is not binding because it is non-specific, though there are exceptions where a water molecule acts as a ligand, always keeping the same orientation — we know this as there are crystal structures with resolved water molecules.⁴⁴

In many cases, we are interested intermolecular interactions that drive binding as they are intricately linked to structure and function. For example, it is mentioned above that the ribosome performs the translation process (function) by accommodating various molecules (binding) at different stages. The ribosome architecture itself (structure) is important such that the molecules can be accommodated.

In triplex nucleic acids, we are mostly interested in the binding interactions of TFO with the double-stranded target via Hoogsteen hydrogen bond. Binding interactions of TFO with the target duplex have prospective applications in antigene therapy, where sequence-specific TFO may be designed to block certain region of the double-stranded DNA target gene to prevent its transcription. One problem of TFOs containing cytosine residue is that it needs to be protonated to form two hydrogen bonds of Hoogsteen. This limits the TFO application as the acidic environment becomes a pre-requisite. The use of pseudocytidine to substitute for cytidine is one strategy to circumvent this limitation.

An experimental technique, called electrophoretic mobility shift assay (EMSA), can be used to assess binding of biomolecules *in vitro*. In the case of duplex-TFO binding, first the target duplex is radiolabelled, and TFO is added. After allowing time for hybridisation, the sample is loaded in the polyacrylamide electrophoresis gel with non-denaturing condition. Electrophoresis is performed and the gel is then visualised. If triplex is formed, an upward

band shift would occur because the triplex would have greater molecular weight than that of the duplex (Figure 7).

Figure 7. An electrophoretic mobility shift assay (EMSA) gel of TFO binding experiment. DS is duplex; TS is triplex.

Adsorption, in contrast to binding, is non-specific. We have used the term loosely in Paper V to describe the non-specific interactions of a peptide and a micelle and quantify it simply by measuring the inter- centre-of-mass distance.

4.2 PROTONATION

Protonation/deprotonation equilibrium is ubiquitous since it involves the universal solvent water. The term pK_a , the decimal cologarithm of acid dissociation constant K_a , is a measure of how much H^+ dissociates from an acid. The more general notation pK shall be used in this thesis. Let us consider the protonation/deprotonation equilibrium of cytosine:

The equilibrium constant K and cologarithm pK are:

$$K = \frac{[cyt][H_3O^+]}{[cytH^+]} \text{ and } pK = -\log \frac{[cyt][H_3O^+]}{[cytH^+]}$$

The pK value of the cytosine is 4.4.⁴⁵ Since the equation can be rearranged to include pH , i.e. $pK - pH = \log \frac{[cyt]}{[cytH^+]}$, we may also statistically interpret $pK = 4.4$ as the pH value for which the populations of deprotonated and protonated states of cytosine are equal.

The free energy of deprotonation can be expressed in terms of pK :

$$\Delta G_{\text{deprot}} = \ln(10) k_B T (pK - pH)$$

4.3 TAUTOMERISATION

Tautomers are readily interconvertible isomers,⁴⁶ where a common form is prototropic tautomers, where the tautomers differ just by the location of a proton. Investigations of tautomerisation, particularly that of nucleic acids, are often experimentally challenging due to the fast interconversion, structural similarity of the tautomers, and low abundance of minor tautomers.⁴⁷

Figure 8. Tautomers of (A) pseudouridine (Ψ C) and (B) uridine.

In this thesis, the prototropic tautomers of pseudouridine (Ψ C), the aforementioned artificial base, which are two main amino tautomers, Ψ C(H1) and Ψ C(H3) will be mentioned; as well as those of uridine, which are one main keto tautomer and two rarer 2- and 4-enol tautomers (Figure 8).

The associated thermodynamic quantities for $\text{A} \rightleftharpoons \text{B}$ tautomerisation are the equilibrium constant $K_{\text{A} \rightleftharpoons \text{B}} = \frac{[\text{B}]}{[\text{A}]}$ and the free energy of tautomerisation $\Delta G_{\text{A} \rightleftharpoons \text{B}} = k_{\text{B}} T \ln K_{\text{A} \rightleftharpoons \text{B}}$.

5 MOLECULAR DESCRIPTIONS

Now that we have the three-dimensional coordinates of the atoms, we have to describe the biomolecule and its intra- and inter-particle interactions. We can divide molecular descriptions in two broad categories: (1) with classical mechanics, also known as molecular mechanical (MM) description; and (2) with *ab initio* or quantum mechanical (QM) description. There also exist hybrid methods which mix the two, or use both descriptions for different subsets of the system (the latter is also called multiscale modelling).

5.1 CLASSICAL MOLECULAR MECHANICS DESCRIPTIONS

The biomolecule can be represented as rigid beads with various inter-bead interactions described by classical mechanics. If a bead represents an atom each, the description is said to be atomistic or all-atom. To expedite calculations, a bead may also represent a group of atoms – this is termed coarse-grained description. In this thesis, we shall focus on the atomistic level of granularity.

The model describing the interatomic interactions is called the force field. There are various kinds of force fields, which may be designed for different purposes, or specifically for a specific class of molecules. Examples of force fields for biomolecules include OPLS,⁴⁸ GROMOS,⁴⁹⁻⁵² AMBER,⁵³⁻⁵⁵ and CHARMM;⁵⁶⁻⁶⁴ the latter two of which are used in this thesis.

The force field is formulated as a potential energy function with empirical parameters describing the atoms and their interactions. The empirical parameters are supplied by experimental values and quantum mechanical *ab initio* calculations. Let us consider an all-atom potential energy function:

$$\begin{aligned} U(\vec{R}) = & \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\phi(1 + \cos(n\phi - \delta)) \\ & + \sum_{\text{impropers}} K_\omega(\omega - \omega_0)^2 \\ & + \sum_{\text{non-bonded pairs}} \left\{ \epsilon_{ij}^{\min} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}} \right\} \end{aligned}$$

This is the potential energy function used by CHARMM⁵⁶ (omitting some correction terms for clarity). The terms comprise bonded and non-bonded terms. The three bonded terms – bonds, angles, and impropers – are described by harmonic functions with their respective

force constants K_b , K_θ , K_ω and equilibrium bond length, angle, improper values b_0 , θ_0 , ω_0 . The bonded term, dihedrals, is described by a cosine function in which K_ϕ is the amplitude; n , periodicity; δ , phase angle. Non-bonded term consists of 6-12 Lennard-Jones potential, describing van der Waals interactions, and Coulomb potential, describing electrostatic interactions. ij is the indices of the atom pair in consideration; $\epsilon_{ij}^{\min}$, well depth; r_{ij} , interatomic distance; $R_{ij}^{\min}$, the interatomic distance of minimal potential energy; q_i , point charge of atom i ; ϵ_0 , permittivity of vacuum; ϵ , permittivity relative to vacuum.

It is apparent in this formulation, the force field cannot account for certain chemical phenomena, such as polarisation, quantum effects, and chemical reactions where bonds are formed and broken.

5.2 AB INITIO DESCRIPTIONS

While we can compute the potential energy of the system with empirical parameters with force fields, *ab initio* calculations compute the energy of the system from first principles – molecular orbital theory in quantum mechanics – without empirical inputs (see reference textbook ⁶⁵).

In molecular orbital theory, the energy of the system can be computed by solving the time-independent electronic Schrödinger equation, $\mathcal{H}\Psi = E\Psi$, where evaluation of the Hamiltonian operator $\mathcal{H}$ on the wavefunction Ψ yields back the wavefunction multiplied by the electronic energy E .

The level of theory refers to the method of approximation used to solve the Schrödinger equation, since it cannot be solved exactly in many-electron system. The Hartree-Fock (notation: HF) method is a fundamental approximation which higher levels of theory (‘post-HF’) improve upon. One important assumption in HF method is the Born-Oppenheimer approximation where the nuclei are considered fixed and decoupled from electronic motions. Employing the variation theorem, which states that a better approximation to the wavefunction leads to a lower energy, HF method includes an iteration procedure to produce lower energy. For this reason, HF method is also known as self-consistent field (SCF) method.

HF method also assumes that each electron interacts with an average charge distribution due to other electrons (mean field approximation), so it does not fully account for electron correlation. Møller–Plesset perturbation theory is a post-HF method that adds the electron correlation correction terms to the HF energy. It has been mentioned that QM calculations are

used to parameterise force field. Such calculations are typically done at second-order Møller–Plesset (notation: MP2) level of theory for geometry optimisations. We will also encounter B3LYP, a hybrid functional method combining HF and density functional theory (DFT), which also provides improvement over HF result.

The basis set is the set of atomic functions used to construct the wavefunction. Although Slater-type orbitals are better representations, in practice, an approximation of these orbitals by linear combination of Gaussian-type functions are used, for ease of computation. The size of basis set to be used in a calculation depends on the system and accuracy-speed trade-off consideration. A minimal basis set contains only the functions representing all the filled orbitals. For example, in STO-3G minimal basis set, 3 Gaussian functions are used to represent a Slater-type orbital. In larger basis sets, additional functions are used represent orbital diffusion or contraction in response to the molecular environment. For example, in 3-21G basis set, 3 Gaussian functions are used to describe core orbitals, 3 Gaussian functions are also used to describe the valence orbitals – of which 2 is for contraction, 1 is for diffusion. These so-called split valence basis sets are most used in this thesis. Even larger basis sets may also include polarisation functions (*) and additional diffuse functions (+). A single * or + notation indicates that the functions are added for heavy atoms, and a double ** or ++ notation indicates that the functions are added for all atoms.

In this thesis, we will encounter notations of *ab initio* calculations specifying the level of theory and the basis set. Let us consider the following notation: MP2/6-31++G**. The notation specifies that the energy is calculated at MP2 level of theory with basis set 6-31++G**, which is a split valence basis set where 6 Gaussian functions are used to represent a Slater-type orbital (for the valence orbitals, 3 are used for contractions and 1 for diffusion) and polarisation and additional diffuse functions are included for all atoms.

If the notation contains two levels of theory and basis sets, for example: MP2/6-31++G**//MP2/3-21+G**, this signifies that geometry optimisation has been done at MP2/3-21+G**, and the energy of the optimised structure is evaluated at MP2/6-31++G**. This is a common practice whereby the more expensive geometry optimisation calculation is done with smaller basis set and the energy evaluation with a larger one, to lower computational cost.

To include bulk solvent effects, a common strategy is to use polarisable continuum model (PCM).⁶⁶ In PCM method, the solute is represented by point charges on the surface of a cavity surrounded by a polarisable medium. The charges polarise the medium, which creates

a reaction field, which polarises the solute back until equilibrium is reached. The solvation free energy calculated by PCM comprises electrostatic, dispersion-repulsion, and cavity terms.

5.3 HYBRID DESCRIPTIONS

5.3.1 QM/MM

Since QM calculation is computationally costly, it is only practical for small-molecule systems. QM/MM hybrid method represents a part of the system with QM, for which precision is needed, and the rest of the system with MM.

The system partitioning gives rise to problems associated with the interface between the MM and QM regions. Three such problems are highlighted and the available strategies briefly described.⁶⁷

First, how to add the energies. In subtractive scheme, the energy of entire system is calculated at MM level, energy of QM region is evaluated at QM level, then energy of QM region evaluated at MM level is subtracted from the total to correct for the link atoms. In additive scheme, MM calculation is only done for MM region, energy of QM region is evaluated at QM level, then energy of coupling between the two systems is added.

Second, how to treat electrostatics. In mechanical embedding, the electrostatics of the QM region is treated the same as that of MM region. In electrostatic embedding, the rigid MM point charges are included in the QM Hamiltonian. In polarised embedding, flexible MM point charges are used to allow for polarisation of MM region.

Third, how to cut covalent bond. Often, the QM/MM boundary would cut through a covalent bond. A common strategy is to add a monovalent link atom, typically hydrogen, to electronically saturate the QM region. The link atom itself introduces other problems such as additional artificial degrees of freedom, and chemical difference between the link atom and the original group. Boundary-atom scheme uses an atom that is included in both QM and MM regions. Localised-orbital scheme uses hybrid orbitals to cap the QM region instead.

The Q-Chem/CHARMM interface⁶⁸ used in this thesis (Paper IV) uses additive scheme, electrostatic embedding, and hydrogen link atoms.

5.3.2 Polarised protein-specific charge

Polarised protein-specific charge (PPC) scheme⁶⁹ is an attempt to describe protein electrostatics better, since in standard force fields, polarisation is not fully accounted for. The

scheme updates the atomic charges as the environment changes in MD simulation run, thus taking polarisation implicitly into account. The charge update itself involves fragmenting the molecule, QM calculation at B3LYP/6-31G* level, charge fitting, and linearised Poisson-Boltzmann equation — this is iterated until solvation energy and the charges converge. There also exists a variant of PPC scheme, adaptive hydrogen bond charge (AHBC),⁷⁰ where the charge update is only applied to residues involved in hydrogen bonds to expedite computation.

6 SIMULATION TECHNIQUES

Now that we have the molecular representations and ways to account for the interactions, we need to sample the phase space, which contain every possible momentum and position of every particle in the system. We shall consider first conventional molecular dynamics, used in every paper in this thesis, then an extension called λ -dynamics that is able to describe transitions between states, for example those between tautomeric forms or protonation states. Next we shall consider techniques to obtain specific thermodynamics properties of interest: multiple pH regime to obtain pK and methods to obtain free energy and PMF.

6.1 MOLECULAR DYNAMICS

Molecular dynamics (MD) simulation propagates the coordinates of the molecules in function of time by numerically integrating Newton's equation of motion, which in one dimension is:

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i}$$

Where it describes the motion of a particle i of mass m_i along coordinate x_i with force F_{x_i} acting on the particle in direction x_i . The forces, being the negative partial derivative of the potential energy, can be obtained from the force field.

Since motions of particles are coupled together, the equations of motion cannot be solved analytically and is solved numerically in discrete time steps. The integration time step is chosen to be small enough to conserve momenta of molecular motion. In the case of biomolecules, time step of 2 fs with SHAKE constraints⁷¹ on bonds involving hydrogen is typically sufficient.

MD simulation thus samples positions and velocities of the particles in the system and produces a time trajectory thereof, from which the time average of the desired observables can be obtained. Assuming ergodicity, the time average is regarded as equal to the ensemble average, from which thermodynamic properties can be derived with statistical mechanics.

To apply statistical mechanics, MD simulations have to be performed in the appropriate thermodynamics ensembles, which are typically constant NVE (microcanonical), constant NVT (canonical), and constant NPT (isothermal-isobaric) [N : number of particles; V : volume; E : internal energy; P : pressure].

To maintain temperature, simple velocity scaling⁷² can be employed, since the temperature is related to the average kinetic energy. Alternatively, Berendsen thermostat⁷³ can be used, in

which the system is coupled to an external heat bath. Nosé-Hoover thermostat⁷⁴⁻⁷⁵ is a more rigorous method where the heat bath is part of the system, represented as additional degree of freedom with fictitious mass, which controls the energy flow between the system and the heat bath. With Andersen thermostat,⁷⁶ a particle is chosen at random at intervals and its velocity randomly reassigned from Maxwell-Boltzmann distribution. With Langevin thermostat, frictional drag and random collisions are introduced and adjusted to achieve the desired temperature.⁶⁵ To maintain pressure, barostats analogous to the above thermostats may be employed where volume of the simulation cell is scaled. Berendsen barostat⁷³ analogously couple the system to a ‘pressure bath’ like the thermostat. Andersen barostat⁷⁶ is analogous to Nosé-Hoover thermostat where external variable acts as a piston to maintain pressure.

Periodic boundary conditions (PBC) are often used in order to eliminate boundary effects. Here, the simulation box is replicated in all directions. A particle that leaves the box will thus be replaced by its image entering from the opposite side. In some cases, PBC may not be an appropriate choice. For instance, in Paper IV we simulated a spherical region carved out of the ribosome, and PBC is not suitable for such inhomogeneous system. We instead used deformable stochastic boundary potential⁷⁷⁻⁷⁸ to contain the system in a sphere. Here, the process of interest is kept in the centre, surrounded by Langevin stochastic heat bath to maintain equilibrium in the central region.

To treat long-range forces, distance cutoff may be used, and as it introduces energy gap, smoothing functions are used to preserve energy continuity. Particle mesh Ewald method,⁷⁹ which more efficiently computes energy summation in Fourier space, may be used for electrostatics.

6.2 LAMBDA-DYNAMICS

One limitation of MD with molecular mechanics force field is that it cannot describe chemical transitions, where chemical bonds are broken and formed. Two examples of such chemical transition equilibria have been mentioned above: protonation/deprotonation and tautomerisation equilibria.

λ -dynamics⁸⁰⁻⁸² is an extension of molecular dynamics, where additional alchemical variable(s), typically denoted as λ , are included in the dynamics along with the atomic coordinates. λ indicates the distance along the alchemical transformation pathway. These transformations or transitions may occur between more than two forms or states – for example histidine has three tautomerisation/protonation states. They may also occur at more than one sites (multisite), for example a dipeptide containing two glutamic acids may need

descriptions of two protonation states for each titration site. In this thesis, we have used a particular formulation of λ -dynamics for CHARMM⁸³ that has been shown to work for constant pH simulation of nucleic acids in explicit solvent.⁸⁴

Let us consider a λ -dynamics simulation where a two-state transition between state A and state B occurs at a single site. The two states A and B are described and propagated by continuous variables λ_A and λ_B respectively. The potential energy function is:

$$\begin{aligned} U_{\text{tot}}(X, \{x\}, \{\lambda\}) \\ = U_{\text{env}}(X) + \lambda_A[U(X, x_A) - \Delta G_{A \rightarrow B}(\text{model})] + \lambda_B[U(X, x_B)] + F^{\text{bias}}(\lambda_A) \\ + F^{\text{bias}}(\lambda_B) \end{aligned}$$

We note that compared to a typical MD simulation, there are extra terms in the potential energy function. X is the coordinates of environment atoms, x_A and x_B are coordinates of atoms corresponding to the states involved in the transition. $U_{\text{env}}(X)$ is the potential energy of environment atoms not involved in the transition.

The variable λ_i ($i = A, B$) scales the potential energy of the corresponding state with the constraints:

$$0 \leq \lambda_i \leq 1 \text{ and } \lambda_A + \lambda_B = 1$$

Since λ is a continuous variable, it may linger in intermediate unphysical states, instead of at the end physical states which we are more interested in. To improve sampling at end states, a sampling bias term for each state is therefore introduced:

$$F^{\text{bias}}(\lambda_i) = \begin{cases} k_{\text{bias}}(\lambda_i - 0.8)^2; & \text{if } \lambda_i < 0.8 \\ 0; & \text{otherwise} \end{cases}$$

Here, $0.8 \leq \lambda_i \leq 1$ is considered as a physical end state. k_{bias} is the force constant of the biasing potential.

$\Delta G_{A \rightarrow B}(\text{model})$ is the free energy difference between the two states in a model compound, which needs to be calculated beforehand. The free energy term is summed to the potential energy term of state A inside the scaling of λ_A , such that it is equally likely to be in state A as it is to be in state B. This value needs to be offset according to what is needed.

For constant-pH simulation, let us consider deprotonated and protonated states A and A+. An additional pH-dependent term, $\ln(10) k_B T (\text{p}K_{\text{model}} - \text{pH})$, is included in λ_A scaling, such that:

$$\begin{aligned}
U_{\text{tot}}(X, \{x\}, \{\lambda\}, \text{pH}) \\
&= U_{\text{env}}(X) + \lambda_A [(U(X, x_A) - \Delta G_{A \rightarrow A+}(\text{model}) \\
&\quad + \ln(10) k_B T (\text{p}K_{\text{model}} - \text{pH}))] + \lambda_{A+} [U(X, x_{A+})] + F^{\text{bias}}(\lambda_A) + F^{\text{bias}}(\lambda_{A+})
\end{aligned}$$

The term $\ln(10) k_B T (\text{p}K_{\text{model}} - \text{pH})$ adds offset when $\text{pH} \neq \text{p}K_{\text{model}}$, where $\text{p}K_{\text{model}}$ is experimentally measured $\text{p}K$ of the model compound. Without any offset, the equation would therefore describe a constant-pH simulation at $\text{pH} = \text{p}K_{\text{model}}$.

In the case of deprotonation/protonation equilibrium, to further improve sampling, it is also possible to run simultaneous simulations at different pHs, termed pH replica exchange.⁸⁵ Since the replicas effectively just need to exchange Hamiltonians, this is a variant of Hamiltonian replica exchange.

Finally, the simulations will yield the populations of each state. In the case of deprotonation/protonation equilibrium, $\text{p}K$ of the system can be obtained by fitting to Henderson-Hasselbalch equation: $S^{\text{deprot}} = \frac{1}{1 + 10^{-n(\text{pH} - \text{p}K)}}$, where S^{deprot} is the fraction of deprotonated state; n is Hill coefficient and indicates degree of cooperativity between titratable groups.

6.3 MULTIPLE PH REGIME

Multiple pH regime⁸⁶ is an approach that specifically aims to calculate $\text{p}K$ by combining configuration sampling of MD with Poisson-Boltzmann equation for electrostatic calculation. Let us consider that we want to calculate $\text{p}K$ of residue A in a certain environment.

First, we sample A in this environment in two sets of ensembles, corresponding to deprotonated and protonated states, with MD simulation. From the two sets of configurations, we calculate the degree of deprotonation, and plot two titration curves. We then average the two titration curves and calculate the $\text{p}K$ value. The electrostatic calculation is needed when we calculate the degree of protonation, since it is a function of change in electrostatic energy, which is in turn a function of intrinsic $\text{p}K$:

$$\text{p}K_{\text{int}} = \text{p}K_{\text{mod}} + \Delta \text{p}K_{\text{Born}} + \Delta \text{p}K_{\text{cc}}$$

The intrinsic $\text{p}K$, $\text{p}K_{\text{int}}$, is the $\text{p}K$ value independent of ionisation properties of other groups in the system. $\text{p}K_{\text{mod}}$ is the standard $\text{p}K$ value for the model compound, which we need to supply. $\Delta \text{p}K_{\text{Born}}$ is the $\text{p}K$ shift due to desolvation, and $\Delta \text{p}K_{\text{cc}}$ is $\text{p}K$ shift due to charge-charge interactions.

In Paper II, we have used a variant of this approach in which we use Poisson, instead of Poisson-Boltzmann, equation. Instead of having the ions implicit as ionic strength term, we include the ions explicitly as point charges and assume that they follow Boltzmann distribution, since they are sampled by MD simulation. This is because the environment of interest is the triple helix nucleic acids, where there are many negatively charged phosphates. Since Poisson-Boltzmann equation does not take into account the finite size of the mobile ions, the negative charges may introduce artefactually high mobile ion concentration around them, so the mobile ions are used explicitly as the source of electric potential.

6.4 FREE ENERGY CALCULATION

The free energy between two states 0 and 1, $\Delta G(0 \rightarrow 1)$, can be computed with:

$$\Delta G(0 \rightarrow 1) = -k_B T \ln \langle \exp[-\frac{(U_1 - U_0)}{k_B T}] \rangle_0$$

Where k_B is Boltzmann constant, T is temperature, U is the potential energy, $\langle \rangle_i$ denotes ensemble average over state i . This equation is the basis of free energy perturbation (FEP) and is commonly called the exponential or Zwanzig equation.⁸⁷ The free energy may also be expressed in different ways, for instance as an integral of U over a parameter λ in thermodynamic integration method.⁸⁸

Bennett acceptance ratio and potential of mean force, which are used in this thesis, are described next.

6.4.1 Bennett acceptance ratio

In this thesis, we utilised an extension of FEP called Bennett acceptance ratio (BAR) method⁸⁹ to compute the free energy. The BAR equation is:

$$\Delta G_{0 \rightarrow 1} = k_B T \ln \left(\frac{\langle f(U_0 - U_1 + C) \rangle_1}{\langle f(U_1 - U_0 - C) \rangle_0} \right) + C$$

Where $f(x)$ is Fermi function, $f(x) = \frac{1}{1 + \exp(\frac{x}{k_B T})}$, C is the free energy of interest, which is to be solved in self-consistent manner.

Comparing Zwanzig and BAR equations, it is notable that ensemble averages for both states are needed in BAR method, whereas in Zwanzig equation only that for one is needed. Consequently, due to its double-sided sampling, BAR calculation is often more efficient, requiring fewer intermediate states to converge.⁹⁰

We have also utilised yet another extension of FEP called QM-Non-Boltzmann BAR (QM-NBB).⁹¹ The equation introduces QM/MM energies as reweighting terms to the BAR equation, such that:

$$\Delta G_{0 \rightarrow 1} = k_B T \ln \left(\frac{\langle f(U_0 - U_1 + C) \exp\left(\frac{V_1^b}{k_B T}\right) \rangle_{1,b} \langle \exp\left(\frac{V_0^b}{k_B T}\right) \rangle_{0,b}}{\langle f(U_1 - U_0 - C) \exp\left(\frac{V_0^b}{k_B T}\right) \rangle_{0,b} \langle \exp\left(\frac{V_1^b}{k_B T}\right) \rangle_{1,b}} \right) + C$$

Where the biasing potential V_i^b ($i = 0, 1$) is the difference between MM and QM/MM energies:

$$V_i^b = U_i^{\text{MM}} - U_i^{\text{QM/MM}}$$

For alchemical intermediate states, the QM/MM energy is obtained by linear scaling:

$$U_\lambda^{\text{QM/MM}} = (1 - \lambda)U_0^{\text{QM/MM}} + \lambda U_1^{\text{QM/MM}}$$

QM-NBB has been shown to improve free energy calculations since it includes reweighting terms in the form of QM/MM potential, which has higher level of accuracy.⁹¹

6.4.2 Potential of mean force

The potential of mean force (PMF) is the free energy surface along a defined reaction path. Since the free energy surface may include high-energy states, conventional MD sampling may be inadequate. The common way to adequately sample the energies along a reaction path is by performing umbrella sampling, where biasing potentials, typically in harmonic form, are used to sample configurations in windows along the reaction path.

The PMF is then obtained by weighted histogram analysis method (WHAM).⁹² It is a maximum likelihood statistical approach that considers the histogram of energy values, assign optimal weights, and calculate the best estimate of the unbiased probability distribution. The best estimate of the unbiased probability distribution is given by:

$$P(x) = \frac{\sum_{i=1}^N n_i(x)}{\sum_{i=1}^N M_i \exp([A_i - U_{\text{bias},i}(x)]/k_B T)}$$

Where N is number of simulations; i is simulation index; $n_i(x)$ is number of counts in histogram bin associated with x ; M_i is number of samples of simulation i ; $U_{\text{bias},i}$ is biasing

potential; A_i is unknown free energy shift and is solved together with $P(x)$ by iteration to self-consistency:

$$A_i = -k_B T \ln \left(\sum_{x_{\text{bins}}} P(x) \exp \left[-\frac{U_{\text{bias},i}(x)}{k_B T} \right] \right)$$

7 SUMMARY, CONCLUSION, AND OUTLOOK

7.1 SUMMARY

The papers have been arranged according to the computational techniques, from classical simulations to *ab initio* calculations:

Paper	Computational Technique	Experimental Technique	Keyword
I	homology modelling, MD	Protein production and purification	Giant virus, protein, eRF1
II	MD, λ -dynamics, multiple pH regime, force field parameterisation	-	nucleic acids, triplex, LNA, pK
III	MD, λ -dynamics	EMSA	Tautomer, nucleic acids, triplex, LNA, pseudoisocytidine
IV	MD, PMF, QM, BAR/NBB, QM/MM	-	Tautomer, ribosome, A-site, cmo ⁵ U
V	MD, PPC	-	Peptide, micelle, membrane, adsorption

Detailed summaries of the papers are presented next, followed by conclusion and future outlook.

I. ***Megavirales* homologues of translation termination factor eRF1: protein production, homology modelling, and molecular dynamics**

Giant viruses (*Megavirales*) are newly discovered group of viruses with large particle and genome size, which exceeds some bacteria and archaea.⁹³ Intriguingly, their genome sequencing reveals that they possess genes homologous to eRF1, which has been shown as unlikely to be pseudogenes, since there is evidence of regulation of their expression following the viral replication cycle.⁹⁴

The translation termination factor eRF1 is an important class I factor that recognises the stop codon at the end of translation process and mediates peptidyl-tRNA hydrolysis to release the translational protein product.⁹⁵ In function, eRF1 may be regarded as tRNA-mimic, since codons are typically read by tRNAs. The translation termination process is also assisted by class II release factors which may facilitate stop codon reading by class I factor, polypeptide release, and ribosome recycling.⁹⁶

The presence of eRF1 homologues in giant viruses is puzzling. First, viruses typically do not encode their own protein translational machinery. Along with eRF1 homologues, sequences corresponding to aminoacyl-tRNA synthetases were also found. Second, no sequence

homologous to class II release factors was found. Third, the gene contains two internal stop codons. To produce a gene product similar in length and sequence to eRF1, the first stop codon needs to be readthrough, the second needs +1 codon frameshifting.⁹⁷ Furthermore, this is true for many giant viruses, but there is a giant virus whose gene does not have the two internal stop codons. However, the presence of internal stop codon itself is not surprising. In some bacteria, the gene encoding class I release factor RF2 contains an internal stop codon which acts as regulation switch.⁹⁸

Our initial inquiry was to structurally compare eRF1 and its *Megavirales* homologues (we shall call it vRF1 for brevity). To achieve this, we attempted to produce and purify the *Megavirales* homologues in order to prepare samples for X-ray crystallography and/or NMR spectroscopy. We chose *E. coli* as the protein expression host, cloned the vRF1 genes, and subjected them to purification stage. The production and purification of vRF1 were challenging mostly due to poor protein yield and solubility. Although we managed to obtain sample with good purity, preliminary NMR spectroscopy showed poorly resolved spectrum and preliminary crystallographic condition screening did not yield crystal suitable for further stage (data not shown in the paper). The experimental investigation was thus paused at this stage.

We then focussed on computational investigation instead. The X-ray crystal structure of human eRF1 is available in the PDB. The individual domains are also available as NMR structures. We also built vRF1 homology models with eRF1 template with the protein structure and function prediction platform, I-TASSER. Notably, the vRF1s do not have a small flexible region of eRF1, called the minidomain, that is not fully resolved in the crystal structure. In the NMR structure, the minidomain exists in two distinct conformations.

We then performed molecular dynamics (MD) simulations and compare the dynamics of eRF1 and vRF1s. Overall, the secondary structures of eRF1 and vRF1s are conserved in the simulation, although there is large interdomain movement. For eRF1, indeed some transient and more persistent secondary structures appear in the minidomain region, though the two NMR conformations cannot be distinguished.

II. Modeling pK shift in DNA triplex containing locked nucleic acids

DNA triple helices have been shown to play important roles in cellular processes and have been used in many biotechnological and biomedical applications, particularly antigene strategy⁹⁹ where the triplex-forming oligonucleotide (TFO) can be used to target intracellular duplex DNA.

In a parallel triple helix, a homopyrimidine- homopurine duplex is bound by Watson-Crick hydrogen bonds, and a third homopyrimidine strand (the TFO) binds to the major groove of the duplex via Hoogsteen hydrogen bond to the homopurine strand. Hence, with canonical DNA bases, the possible base triads are $C^+ \bullet G-C$ and $T \bullet A-T$ (‘-’ refers to Watson-Crick and ‘•’ to Hoogsteen base pair), where cytidine in the third strand needs to be protonated to form Hoogsteen hydrogen bond.

Locked nucleic acid (LNA) is a modification that locks the sugar conformation, and is typically used to improve triplex formation. In this study, we reparameterised and validated the CHARMM force field parameters for LNA (previous parameter: Pande-Nilsson¹⁰⁰) as well as calculated pK values of cytidine in various triplex environment with two different approaches, multiple pH regime and λ -dynamics. Multiple pH regime uses MD sampling of protonated and deprotonated ensembles, and determine the pK shift by calculating electrostatics with linearised Poisson equation⁸⁶. λ -dynamics incorporates λ , the variable which is coupled to the protonation-deprotonation transition, in the dynamics.⁸⁴

The reparameterised LNA force field reproduced A-form nucleotide geometry and has good agreement with experimental structures. The two pK approaches both predict the pK of cytidine to be shifted higher than physiological pH in triplex environment. 5-methylation shifted pK even higher. Neighbouring LNAs do not seem to have large effect on pK shift. For cytidine, multiple pH regime predicts higher pK value than λ -dynamics and similar ones for 5-methylcytidine. Multiple pH regime predicts downward pK shift when ionic strength is increased, while λ -dynamics predicts no change. This is presumably due to the different treatments in the long-range interactions in both methods.

III. Role of Pseudoisocytidine Tautomerization in Triplex-Forming Oligonucleotides: In Silico and in Vitro Studies

In a parallel triple helix, cytidine in the third strand needs to be protonated to form Hoogsteen hydrogen bond. Pseudoisocytidine (ΨC) is a cytidine analogue designed to circumvent this problem. However, ΨC exists in two main tautomers, only one of which, $\Psi C(H3)$, is able to form Hoogsteen hydrogen bond to form the base triad $\Psi C(H3) \bullet G-C$.

We have used λ -dynamics⁸⁴ to model the tautomerisation between the tautomers of ΨC , $\Psi C(H1) \rightleftharpoons \Psi C(H3)$ with various sequence, single-strand, and triplex contexts. We complemented the computational prediction with electrophoretic mobility shift assay (EMSA), an *in vitro* binding experiment used to confirm triplex formation. Conventional MD simulation was also performed to characterise the structure of triplexes containing ΨC .

As there is no quantitative experimental data on this tautomerisation equilibrium, we assumed that the two tautomers exist in equal amount as nucleoside. We calculated the free energy of tautomerisation, needed as an input of λ -dynamics, and calibrated the biasing potential force needed to efficiently sample the physical end states. We then performed in various single-stranded trimers and heptamers containing Ψ C. We observed that sequence containing consecutive Ψ Cs tend to disfavour Ψ C(H3), while cytidine neighbours tend to favour Ψ C(H3) compared to thymine neighbours. LNA does not have clear influence on the tautomerisation. In triplex environment, where we have started the simulation starting from the triplex structure, Ψ C(H3) fully dominates, even when Ψ C is consecutive.

EMSA shows that sequences containing three or more consecutive Ψ C do not form triplex under intranuclear conditions at pH 7.4, even when intercalators that promote triplex formation were used. On the other hand, when we have another sequence containing non-consecutive Ψ C and LNA, it shows triplex formation. We modelled the latter triplex with conventional MD simulation and characterise the triplex structure and found that its duplex in triplex conformation is consistent with what has been previously observed,⁶ that duplex in triplex conformation has slide and twist values of A-type geometry with x-displacement value between A- and B-types.

IV. A computational study of uracil tautomeric forms on the ribosome: the case of uracil and 5-oxyacetic acid uracil in the first anticodon position of tRNA

An unexpected Watson-Crick geometry was observed for a modified U:G base pair in the ribosome decoding centre at third codon position, where wobble geometry is expected instead.¹⁰¹ With this modification of U on the tRNA, 5-oxyacetic acid, it can accept all four bases on the mRNA.¹² The authors of the X-ray crystallographic study speculated that the U enol tautomer, much rarer than the keto tautomer, is consistent with the observed Watson-Crick geometry. However, since at this resolution hydrogen atoms are not resolved, this remains an open question.

We have employed *ab initio* calculations to calculate the energy of the tautomers alone and in basepair with polarisable continuum model (PCM)⁶⁶ to account for solvent effects. We have found that the modification offers very little stabilisation to the enol tautomer, alone and in Watson-Crick basepair geometry. There is significant stabilisation if we allow the enol hydroxyl to form intramolecular hydrogen bond with the acetic acid group in the modification, but this geometry would not be consistent with the X-ray structure.

To account for the ribosomal environment, we performed MD simulation and free energy calculation with the ribosomal context. We have taken a spherical region of the decoding centre and used deformable stochastic boundary potential⁷⁷⁻⁷⁸ to contain the system. For the free energy calculation, we have utilised Bennett acceptance ratio (BAR)⁸⁹ as well as quantum mechanics-non Boltzmann BAR (QM-NBB)⁹¹ which involves reweighting factors from QM/MM calculation for greater accuracy. Free energy calculation shows that the ribosomal environment does not offer significant stabilisation of the enol tautomer compared to the aqueous environment.

MD simulation cannot maintain the geometry observed in the X-ray structure fully. First, a geometry close to Watson-Crick was achieved, but with only two hydrogen bonds maintained and the U enol hydrogen points away from the basepair partner G and does not form a hydrogen bond. Second, the modification moiety points away from the Watson-Crick edge, while it points towards the Watson-Crick edge in the X-ray structure.

We calculated the PMF along the dihedral that describes the enol hydrogen rotation and found that the orientation observed in the MD simulation was stabilised in aqueous environment by water-bridged intramolecular hydrogen bond. Since the orientation of the modification moiety in the X-ray structure cannot be maintained in MD simulation, we also propose that there may be entities not resolved by X-ray diffraction that stabilise this orientation, such as Mg^{2+} , for which we performed preliminary MD simulation.

We also revisited our initial assumptions such as the protonation state of the acetic acid in the modification moiety that we assumed to be protonated, as well as the possibility of N3-deprotonated uridine. For the latter, we performed preliminary MD simulation but still did not achieve Watson-Crick geometry.

V. Adsorption and folding dynamics of MPER of HIV-1 gp41 in the presence of DPC micelle

Membrane-proximal ectodomain region (MPER) of HIV-1 is part of its envelope glycoprotein gp41, which undergoes conformational change to facilitate the viral membrane fusion.¹⁻² This region has been known to be epitope of several monoclonal antibodies¹⁰²⁻¹⁰⁴ and due to its proximity to the membrane, it has different structural features compared to a typical soluble protein.

Protein-polarised charge (PPC)⁶⁹ scheme is a charge updating scheme based on iteration of quantum mechanical calculation at B3LYP/6-31G* and Poisson-Boltzmann linearised

equation to implicitly include polarisation by updating charges in response to the environment.

In this study, we simulated MPER peptide with starting charges from both AMBER03 force field and PPC, in the absence and presence of dodecylphosphatidylcholine (DPC) micelle. The starting structure is taken from NMR study.¹⁰⁵ Without DPC micelle, the MPER peptide is not structurally preserved well with PPC preserving it slightly better compared to with AMBER charges. In the presence of micelle, MPER peptide always adsorb to the micelle, with PPC preserving the peptide structure better.

We also attempted *de novo* folding of MPER peptide from its linear structure with adaptive hydrogen bond charge (AHBC)⁷⁰ charge updating scheme which is a PPC variant, but the charge update is only applied to residues involved in hydrogen bonding. Unlike the one-time PPC update above, AHBC is applied by locating hydrogen bonds every 5 ps. The peptide folded similar to the NMR structure while undergoing adsorption to the micelle. The peptide adsorption to the micelle appears to aid the folding process.

7.2 CONCLUSION AND OUTLOOK

In Paper I, we gained some insights in the structure of vRF1s, and the missing minidomain region may be a crucial structural feature. However, there is the caveat that these structures are homology models and the validation from experimental investigations still awaits. X-ray crystallography would probably be the most suitable experimental technique to resolve their three-dimensional structures. The minidomain region itself would probably benefit from a more careful loop modelling, which we did attempt, but with no conclusive result.

In Paper II, we updated the LNA force field parameters and predicted pK values for cytidine in triplex environment. Since experimental pK values of cytidine in triplexes are sparse, computational predictions would be useful in the context of designing a cytidine-containing TFO for antigene therapy. Multiple pH regime with explicit ions was shown to offer good predictions, so future investigations involving nucleic acids may take note. λ -dynamics also offers predictions where some are close to the experimental value, but it has greater computational cost. pK calculations may be used in other systems that would benefit from such predictions where experimental measurement may be difficult or costly.

In Paper III, we employed λ -dynamics to complement *in vitro* binding experimental result. This study illustrates how computational methods can provide insights to processes which are experimentally more intractable. It would be interesting to apply similar strategy to other

tautomeric systems. What we have known so far about pseudoisocytidine and its tautomerisation may inform applications other than cytidine substitution in TFO.

In Paper IV, we explored the possibility of U enol tautomer forming Watson-Crick basepair with G in the ribosomal decoding centre at the third position. We conclude that the energetics and what we have observed in MD simulation is not consistent with the U enol tautomer. We have offered several alternative hypotheses which can be further explored in the future: the inclusion of Mg^{2+} , the protonation state of the oxyacetic acid moiety, the N3-deprotonated form of uridine, and the possibility of enol tautomer of guanine instead of uridine. More experimental data such as higher resolution structure may also be needed.

In Paper V, we did various simulations of a peptide associated with a micelle. The electrostatics proves to be crucial, since the method with more extensive electrostatics description is able to structurally preserve or fold the peptide better. However, since the charge update scheme is computationally costly, it has not gained much widespread use. In the future, we may focus on reducing the computational cost or devising other schemes. MPER peptide is also only a small subset of the membrane fusion machinery. Other parts of the machinery and itself as a whole also need to be investigated in order to understand the membrane fusion mechanism better.

8 POPULAR SCIENCE SUMMARY

While the mechanical engineer looks at a machine and dismantles its innards to understand its workings, in the same way the structural biologist looks at life and subjects its inner clockwork to scrutiny. But the latter faces a fundamental problem: these engines are extremely tiny and frustratingly so – this is a challenge beyond van Leeuwenhoek's microscopes: no light microscope can resolve the structure of the protein molecule, for it is optically impossible; the visible light wavelength is 400-700 nm, while the ribosome, itself a complex of many pieces of proteins and nucleic acids, is 20-30 nm in dimension.

However, we have biophysical tools to tease out the secrets of nature's architecture, among them are X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy – all of them can produce 3D representations of molecules, and then some. These 3D coordinates are commonly deposited in a public depository called the Protein Data Bank.

It is worth pointing out that the construction of these models is not the same way that one typically thinks about resolving images – think of a camera: shine a light on the object, capture the light reflections, done – only electron microscopy, among the three, works like this. Even then, an assemblage of these images is taken to construct the final model. X-ray crystallography is superficially similar, initially: shine an X-ray on a crystal, but what is captured then is diffraction patterns (recall Rosalind Franklin's famous diffraction pattern of the DNA double helix) and calculations have to be made to process the patterns to reconstruct the model. NMR spectroscopy too begins with shining radiowaves on the sample, but what it records is the magnetic properties of atomic nuclei, which can be used to find out distances between atoms – these distances are inputted to the model building, which restrains the conformation narrowly to the final model.

If we step back from model building, which is sample *measurement*, there is sample *preparation*, a messy backstage work that is often unseen. In the case of proteins, a common production strategy is to genetically modify bacteria, typically *E. coli*, to coax them to produce the desired protein, as the author has done in Paper I. Proteins function in very diverse environments – aqueous to lipidous, low to high pH, dilute to concentrated – yet our production techniques greatly favour the soluble protein. NMR is still largely done in solution; crystallography requires highly solubilised protein in order to grow good crystals; cryo-EM specimen is also prepared in aqua, though at less forgiving concentration than crystallography.

Aren't these *only* models, then? — is a question the structural biologist sooner or later encounters. Yes and no. On one hand, it is true there are limitations such as missing parts not resolved by X-ray crystallography. But on the other hand, to wax philosophical, everything is a model. This text you are reading on the screen or the paper is your brain's mental construct – to be sure, *everything*: the text, the screen, the paper, your hands. The molecular modeller indeed should know the inherent limitations of the models, but it does not mean that they render the model useless.

What a computational biophysicist often does is to extract more information from these static models. X-ray and cryo-EM structures are often one single 'frame' while an NMR structure might include some 20 different conformations. An X-ray structure often does not resolve hydrogens since they do not have much electron to diffract. Paper IV opens with exactly this problem: the common placing of hydrogen in an X-ray structure does not make sense; could it be that the hydrogen is placed on a different site? Calculations were then performed to see whether there is space for this hydrogen, whether there is enough energy, and so on.

Molecular dynamics, the computational technique that is a prominent theme of this thesis, incorporates the classical mechanics, treating the biomolecules like solid charged balls having different attractions and repulsions, connected with springs of different lengths and rigidities, moving under Newton's law of motion. This is not the most accurate description, as quantum mechanics would be it, but since the latter requires heavy computation, so it is used sparingly, or compromisingly with approximations, or in hybrid conjunction with the classical mechanics description. The parameters that go to the molecular model used for molecular dynamics are obtained from real-world measurements and high-level quantum mechanical calculation, and then carefully calibrated and validated, so that they would correctly reproduce biophysical properties.

Paper I provide examples of conventional molecular dynamics simulations, wherein a protein is put amidst water molecules and ions, then simulated. In Paper I, the interest is in the dynamics of the protein, especially since the 'arms' of the T-shaped protein is known to move a lot. Proteins with similar sequence, said to be homologous, were also structurally predicted and simulated to compare their structures, especially since the homologues are missing a certain region of the template protein.

One limitation of the conventional molecular dynamics is that it cannot take into account chemical reactions, which involves bond breaking – for the models are solid balls and springs with fixed rigidities. Lambda-dynamics, used in Papers II and III, is an extension of the

technique that addresses exactly this issue, where the eponymous 'lambda' is a variable that accounts for the transition of interest. The chemical transitions of interest in this thesis are acid-base protonation (Paper II) and tautomerisation (Paper III). The added variable adds to the computation: the parameters of both states are needed, the free energy between the two states has to be calculated, and sampling has to be tuned in order for it not to linger at intermediate states, which are not of interest.

In Paper II, lambda-dynamics allows calculation of pK_a , the acid dissociation constant, which is a measure of how strong an acid is. In this case, our interest is one of nucleic acid bases, C; how its pK_a changes when put in different environments, for in the environment of interest, triplex nucleic acids, experimental measurement is a challenge.

In Paper III, lambda-dynamics was used to model transition between two tautomers of a nucleic acid base – tautomers are closely related isomers where they only differ by the position of one hydrogen atom. Here we have experimental data of triplex nucleic acid binding which was initially puzzling, but in the light of the tautomerisation data from the simulations, was better explained.

Paper IV is concerned also with tautomers, but that of the nucleic acid base U. As mentioned above, the common tautomer does not fit into the model and the other, much rarer tautomer, seems to. The environment is the ribosome, which might be special and somehow compensates the energy cost. We employed *ab initio* quantum mechanical calculations, free energy calculations, hybrid QM/MM calculations, and molecular dynamics simulation to determine whether this is so.

In Paper V, a ball of lipid molecules called a micelle is also introduced to the system to study how the protein would interact with it. The protein in question is in fact part of HIV infection machinery. Understanding it better will help to devise strategies against the virus. Hybrid method that incorporates quantum mechanical calculations to the molecular dynamics simulation was also used to better describe the electrostatics of the solvated peptide.

The structural biologist stares at life and tries to decode its molecular machinery. This thesis is concerned with just a small set of her repertoire of tools, the computational techniques. She keeps tinkering; she keeps wondering; for "What is life?" is not all a trivial question.

ACKNOWLEDGEMENTS

First and foremost, a heartfelt thanks to my supervisors. Konstantin, who saw the potential in me and accepted me as member of the group, imparting your vast NMR knowledge, and coming up with endless ideas. Lennart, who entrusted me MD projects, being a solid guide and mentor, and a constant presence. Alessandra, who was only listed as my supervisor at the end of my PhD, but has played a big role even before, by encouraging and giving the push to report weekly progress update, gently directing when the projects reached dead ends, and putting me in charge of pseudoisocytidine project – which turned out to be my first PhD paper. I aspire to be a teacher and a mentor like you.

To NTU colleagues and collaborators: Dr Mu Yuguang in Thesis Advisory Committee; Rubing, for being all-around helpful; Leo, Shubhadra and Bai Yang, for mentoring your junior; Margaret, for being constant encouragement and trusty friend; Vidhya, for being a reliable older sister – I admire how you unfathomably manage your life between work, PhD, and family; Rachel, for many advices about Sweden; the late Alistair, who passed away too soon, your presence late in the evening gave me company and motivation; Eugene, Raymond, and May for administrative support; Li Teng and Frances, for shared PhD sufferings; Yew Mun, for being helpful always and sharing your room with me; Kwok Kiong, Charlie, Andy, Peiyong for working together in Zhang group; Raymond, Celine, and Yi Ying, for our undergraduate years; Ming Han, CK, Boon, Guan Da and others from NTU Lifesavers.

To KI colleagues and collaborators: Ted, Rula and Karin; Roger and Caroline in halftime board; You Xu, for being a senior I can always questions to and a friend outside work; Eva and Arzu, for making the work environment bright; Joanna and Zohreh, for brief stay in the group but warmly coloured the atmosphere; Vladimir, a close collaborator and my faithful blog reader; Andrea, for being a kind neighbour and friend; neighbours Carsten's and Peter's groups; the administration and IT personnel – of which Monica deserves a special mention; Emma, Kerstin, and Anethe from KI Career Service as well as Natalie and fellow bloggers in KI Career and Research blog team, thank you for giving me free rein to write; Johanna from International Relations.

To friends I made in Sweden: Debora and Pieter, Nienke and Mikael, Kavitha and Jack, Tabea, Maria, Denise, Andreas, Michael, Nathan, Bethany, Andrew, Pedro, Cam and others at Immanuel Church; Ci Lia, Tante-tante, Eross and Dahlia, and others in Indonesian Fellowship; Viggo and swimmers and coaches of SSIF Swimming groups – swimming keeps me sane, really.

To friends in Singapore: Peter, Ronny, Matron Joei who always came to meet me whenever I was in Singapore; ACS schoolmates; GRII friends; Living Waters friends.

To all my teachers and mentors: Mrs Patricia Thong; Mdm Yong Lee Har; Dr Alistair Chew, who inculcated my love for chemistry; Dr Zhang Dawei, for trusting me with writing journal articles, which propelled me for a good PhD position; Prof Lars Nordenskiöld, and countless others.

To my family: Papa, Mama, Ko Yossy, Ci Fenti, Jojo and Oswald – thank you for keeping me in your prayers; Tante Vonny, Uncle Tim, Chris and Tania – I had a most pleasant Christmas in Canterbury.

To others not mentioned, but crossed my path and left a mark.

Finally, to God, from whom all blessings flow, and to whom all glories belong.

REFERENCES

- (1) Montero, M.; van Houten, N. E.; Wang, X.; Scott, J. K., The membrane-proximal external region of the human immunodeficiency virus type 1 envelope: dominant site of antibody neutralization and target for vaccine design. *Microbiol. Mol. Biol. Rev.* **2008**, *72* (1), 54-84.
- (2) Melikyan, G. B.; Markosyan, R. M.; Hemmati, H.; Delmedico, M. K.; Lambert, D. M.; Cohen, F. S., Evidence that the transition of HIV-1 gp41 into a six-helix bundle, not the bundle configuration, induces membrane fusion. *J. Cell Biol.* **2000**, *151* (2), 413-424.
- (3) Ogle, J. M.; Brodersen, D. E.; Clemons, W. M.; Tarry, M. J.; Carter, A. P.; Ramakrishnan, V., Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **2001**, *292* (5518), 897-902.
- (4) Ogle, J. M.; Murphy, F. V.; Tarry, M. J.; Ramakrishnan, V., Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell* **2002**, *111*, 721-732.
- (5) Song, H. W.; Mugnier, P.; Das, A. K.; Webb, H. M.; Evans, D. R.; Tuite, M. F.; Hemmings, B. A.; Barford, D., The crystal structure of human eukaryotic release factor eRF1 - Mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* **2000**, *100* (3), 311-321.
- (6) Esguerra, M.; Nilsson, L.; Villa, a., Triple helical DNA in a duplex context and base pair opening. *Nucleic Acids Res.* **2014**, *42*, 11329-11338.
- (7) Pabon-Martinez, Y. V.; Xu, Y.; Villa, A.; Lundin, K. E.; Geny, S.; Nguyen, C.-H.; Pedersen, E. B.; Jørgensen, P. T.; Wengel, J.; Nilsson, L.; Smith, C. I. E.; Zain, R., LNA effects on DNA binding and conformation: from single strand to duplex and triplex structures. *Scientific Reports* **2017**, *7* (1), 11043.
- (8) Hartono, Y. D.; Pabon-Martinez, Y. V.; Uyar, A.; Wengel, J.; Lundin, K. E.; Zain, R.; Smith, C. E.; Nilsson, L.; Villa, A., Role of Pseudoisocytidine Tautomerization in Triplex-Forming Oligonucleotides: In Silico and in Vitro Studies. *ACS Omega* **2017**, *2* (5), 2165-2177.
- (9) Carell, T.; Brandmayr, C.; Hienzsch, A.; Müller, M.; Pearson, D.; Reiter, V.; Thoma, I.; Thumbs, P.; Wagner, M., Structure and function of noncanonical nucleobases. *Angew. Chem. Int. Ed.* **2012**, *51* (29), 7110-7131.
- (10) Helm, M.; Alfonzo, J. D., Posttranscriptional RNA modifications: playing metabolic games in a cell's chemical Legoland. *Chem. Biol.* **2014**, *21* (2), 174-185.
- (11) Squires, J. E.; Patel, H. R.; Nusch, M.; Sibbritt, T.; Humphreys, D. T.; Parker, B. J.; Suter, C. M.; Preiss, T., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **2012**, *40* (11), 5023-5033.
- (12) Mitra, S. K.; Lustig, F.; Akesson, B.; Lagerkvist, U., Codon-anticodon recognition in the valine codon family. *J. Biol. Chem.* **1977**, *252* (2), 471-478.
- (13) Ono, A.; Ts'o, P. O. P.; Kan, L. S., Triplex formation of oligonucleotides containing 2'-O-methylpseudoisocytidine in substitution for 2'-deoxycytidine. *J. Am. Chem. Soc.* **1991**, *113*, 4032-4033.
- (14) Obika, S.; Nanbu, D.; Hari, Y.; Morio, K.-i.; In, Y.; Ishida, T.; Imanishi, T., Synthesis of 2'-O, 4'-C-methyleneuridine and-cytidine. Novel bicyclic nucleosides having a fixed C 3,-endo sugar puckering. *Tetrahedron Lett.* **1997**, *38* (50), 8735-8738.
- (15) Koshkin, A. a.; Singh, S. K.; Nielsen, P.; Rajwanshi, V. K.; Kumar, R.; Meldgaard, M.; Olsen, C. E.; Wengel, J., LNA (Locked Nucleic Acids): Synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition. *Tetrahedron* **1998**, *54*, 3607-3630.
- (16) Højland, T.; Kumar, S.; Babu, B. R.; Umemoto, T.; Albaek, N.; Sharma, P. K.; Nielsen, P.; Wengel, J., LNA (locked nucleic acid) and analogs as triplex-forming oligonucleotides. *Org. Biomol. Chem.* **2007**, *5* (15), 2375-9.
- (17) 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.
- (18) Lawson, C. L.; Baker, M. L.; Best, C.; Bi, C.; Dougherty, M.; Feng, P.; van Ginkel, G.; Devkota, B.; Lagerstedt, I.; Ludtke, S. J., EMDatabank. org: unified data resource for CryoEM. *Nucleic Acids Res.* **2010**, *39*, D456-D464.

- (19) Smyth, M.; Martin, J., x Ray crystallography. *Mol. Pathol.* **2000**, *53* (1), 8.
- (20) Wüthrich, K.; Wider, G.; Wagner, G.; Braun, W., Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* **1982**, *155* (3), 311-319.
- (21) Kumar, A.; Ernst, R.; Wüthrich, K., A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.* **1980**, *95* (1), 1-6.
- (22) Wüthrich, K., The way to NMR structures of proteins. *Nat. Struct. Mol. Biol.* **2001**, *8* (11), 923-925.
- (23) Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (23), 12366-12371.
- (24) Taylor, D.; Unbehauen, a.; Li, W.; Das, S.; Lei, J.; Liao, H. Y.; Grassucci, R. a.; Pestova, T. V.; Frank, J., Cryo-EM structure of the mammalian eukaryotic release factor eRF1-eRF3-associated termination complex. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, 1-6.
- (25) Wong, W.; Bai, X.-c.; Brown, A.; Fernandez, I. S.; Hanssen, E.; Condron, M.; Tan, Y. H.; Baum, J.; Scheres, S. H., Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **2014**, *3*, e03080.
- (26) Voorhees, R. M.; Fernández, I. S.; Scheres, S. H.; Hegde, R. S., Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell* **2014**, *157* (7), 1632-1643.
- (27) Amunts, A.; Brown, A.; Toots, J.; Scheres, S. H.; Ramakrishnan, V., The structure of the human mitochondrial ribosome. *Science* **2015**, *348* (6230), 95-98.
- (28) Kühlbrandt, W., The resolution revolution. *Science* **2014**, *343* (6178), 1443-1444.
- (29) Gräslund, S.; Nordlund, P.; Weigelt, J.; Bray, J.; Gileadi, O.; Knapp, S.; Oppermann, U.; Arrowsmith, C.; Hui, R.; Ming, J., Protein production and purification. *Nat. Methods* **2008**, *5* (2), 135-146.
- (30) Hochuli, E.; Bannwarth, W.; Döbeli, H.; Gentz, R.; Stüber, D., Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Nat. Biotechnol.* **1988**, *6* (11), 1321-1325.
- (31) Šali, A.; Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779-815.
- (32) Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Šali, A., Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29* (1), 291-325.
- (33) Webb, B.; Sali, A., *Comparative Protein Structure Modeling Using MODELLER*. John Wiley & Sons, Inc.: 2002; Vol. 5.6.1-5.6.32.
- (34) Webb, B.; Sali, A., Protein structure modeling with MODELLER. *Protein Structure Prediction* **2014**, 1-15.
- (35) Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T., The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **2006**, *22* (2), 195-201.
- (36) Kiefer, F.; Arnold, K.; Künzli, M.; Bordoli, L.; Schwede, T., The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* **2008**, *37*, D387-D392.
- (37) Guex, N.; Peitsch, M. C.; Schwede, T., Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **2009**, *30* (S1).
- (38) Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L., SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **2014**, *42* (W1), W252-W258.
- (39) Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008**, *9* (1), 40.

- (40) Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **2010**, *5*, 725–38.
- (41) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2015**, *12* (1), 7-8.
- (42) Rizzuti, B.; Daggett, V., Using simulations to provide the framework for experimental protein folding studies. *Arch. Biochem. Biophys.* **2013**, *531* (1), 128-135.
- (43) Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Eur. J. Inorg. Chem.* **1894**, *27* (3), 2985-2993.
- (44) Spyraakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E., The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**.
- (45) Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B., Calculation of pK_as in RNA: On the structural origins and functional roles of protonated nucleotides. *Journal of molecular biology* **2007**, *366* (5), 1475-1496.
- (46) McNaught, A. D., *Compendium of chemical terminology*. Blackwell Science Oxford: 1997; Vol. 1669.
- (47) Singh, V.; Fedeles, B. I.; Essigmann, J. M., Role of tautomerism in RNA biochemistry. *RNA* **2014**, *21*, 1–13.
- (48) Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657-1666.
- (49) Schuler, L. D.; Daura, X.; Van Gunsteren, W. F., An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22* (11), 1205-1218.
- (50) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F., A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656-1676.
- (51) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Kräutler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F., An improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.* **2005**, *26* (7), 725-737.
- (52) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F., Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40* (7), 843.
- (53) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179-5197.
- (54) Ponder, J. W.; Case, D. A., Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27-85.
- (55) 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.
- (56) MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586-3616.
- (57) MacKerell, A. D.; Banavali, N. K., All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **2000**, *21*, 105–120.
- (58) Foloppe, N.; MacKerell, A. D. J., All-Atom Empirical Force Field for Nucleic Acids : I . Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86–104.
- (59) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., CHARMM general force field: A force field for

- drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31* (4), 671-90.
- (60) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D., Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.* **2011**, *8*, 348–362.
- (61) Denning, E. J.; Priyakumar, U.; Nilsson, L.; Mackerell, A. D., Impact of 2'-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.* **2011**, *32* (9), 1929-1943.
- (62) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257-3273.
- (63) Xu, Y.; Vanommeslaeghe, K.; Aleksandrov, A.; MacKerell, A. D., Jr.; Nilsson, L., Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.* **2016**, *37* (10), 896-912.
- (64) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14* (1), 71-73.
- (65) Leach, A. R., *Molecular modelling: principles and applications*. Pearson education: 2001.
- (66) Miertuš, S.; Scrocco, E.; Tomasi, J., Electrostatic interaction of a solute with a continuum. A direct utilization of AB initio molecular potentials for the prediction of solvent effects. *Chem. Phys.* **1981**, *55* (1), 117-129.
- (67) Senn, H. M.; Thiel, W., QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48* (7), 1198-1229.
- (68) Woodcock, H. L.; Hodošček, M.; Gilbert, A. T.; Gill, P. M.; Schaefer, H. F.; Brooks, B. R., Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations. *J. Comput. Chem.* **2007**, *28* (9), 1485-1502.
- (69) Ji, C.; Mei, Y.; Zhang, J. Z., Developing polarized protein-specific charges for protein dynamics: MD free energy calculation of pK_a shifts for Asp 26/Asp 20 in Thioredoxin. *Biophys. J.* **2008**, *95* (3), 1080-1088.
- (70) Duan, L. L.; Mei, Y.; Zhang, D.; Zhang, Q. G.; Zhang, J. Z. H., Folding of a helix at room temperature is critically aided by electrostatic polarization of intraprotein hydrogen bonds. *J. Am. Chem. Soc.* **2010**, *132*, 11159–11164.
- (71) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327-341.
- (72) Woodcock, L.-V., Isothermal molecular dynamics calculations for liquid salts. *Chem. Phys. Lett.* **1971**, *10* (3), 257-261.
- (73) Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J., Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **1984**, *81* (8), 3684-3690.
- (74) Nosé, S., A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52* (2), 255-268.
- (75) Hoover, W. G., Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31* (3), 1695.
- (76) Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics* **1980**, *72* (4), 2384-2393.
- (77) Brooks, C. L.; Karplus, M., Deformable stochastic boundaries in molecular dynamics. *J. Chem. Phys.* **1983**, *79*, 6312.
- (78) Brünger, A.; Brooks, C. L.; Karplus, M., Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.* **1984**, *105*, 495–500.

- (79) Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089-10092.
- (80) Kong, X.; Brooks III, C. L., λ -dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105* (6), 2414-2423.
- (81) Guo, Z.; Brooks, C.; Kong, X., Efficient and flexible algorithm for free energy calculations using the λ -dynamics approach. *J. Phys. Chem. B* **1998**, *102* (11), 2032-2036.
- (82) Knight, J. L.; Brooks, C. L., λ -Dynamics free energy simulation methods. *J. Comput. Chem.* **2009**, *30* (11), 1692-1700.
- (83) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoseck, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545-614.
- (84) Goh, G. B.; Knight, J. L.; Brooks, C. L., Constant pH molecular dynamics simulations of nucleic acids in explicit solvent. *J. Chem. Theory Comput.* **2012**, *8*, 36-46.
- (85) Itoh, S. G.; Damjanović, A.; Brooks, B. R., pH replica-exchange method based on discrete protonation states. *Proteins* **2011**, *79*, 3420-36.
- (86) Nilsson, L.; Karshikoff, A., Multiple pH Regime Molecular Dynamics Simulation for p K Calculations. *PLoS ONE* **2011**, *6* (5), e20116.
- (87) Zwanzig, R. W., High-temperature equation of state by a perturbation method. I. nonpolar gases. *J. Chem. Phys.* **1954**, *22* (8), 1420-1426.
- (88) Kirkwood, J. G., Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300-313.
- (89) Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245-268.
- (90) Shirts, M. R.; Pande, V. S., Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **2005**, *122*, 144107.
- (91) König, G.; Hudson, P. S.; Boresch, S.; Woodcock, H. L., Multiscale free energy simulations: An efficient method for connecting classical MD simulations to QM or QM/MM free energies using non-Boltzmann Bennett reweighting schemes. *J. Chem. Theory Comput.* **2014**, *10*, 1406-1419.
- (92) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. a., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011-1021.
- (93) Raoult, D.; Audic, S. e.; Robert, C.; Abergel, C.; Renesto, P.; Ogata, H.; La Scola, B.; Suzan, M.; Claverie, J.-M., The 1.2-megabase genome sequence of Mimivirus. *Science* **2004**, *306* (5700), 1344-50.
- (94) Legendre, M.; Audic, S.; Poirot, O.; Hingamp, P.; Seltzer, V.; Byrne, D.; Lartigue, A.; Lescot, M.; Bernadac, A.; Poulain, J., mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.* **2010**, *20* (5), 664-674.
- (95) Frolova, L. Y.; Merkulova, T. I.; Kisselev, L. L., Translation termination in eukaryotes: polypeptide release factor eRF1 is composed of functionally and structurally distinct domains. *RNA* **2000**, *6*, 381-390.
- (96) Petry, S.; Weixlbaumer, A.; Ramakrishnan, V., The termination of translation. *Curr. Opin. Struct. Biol.* **2008**, *18*, 70-77.
- (97) Jeudy, S.; Abergel, C.; Claverie, J. M.; Legendre, M., Translation in Giant Viruses: A Unique Mixture of Bacterial and Eukaryotic Termination Schemes. *PLoS Genet.* **2012**, *8*.
- (98) Craigen, W. J.; Caskey, C. T., Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **1986**, *322*, 273-275.

- (99) Goñi, J. R.; Vaquerizas, J. M.; Dopazo, J.; Orozco, M., Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **2006**, *7* (1), 63.
- (100) Pande, V.; Nilsson, L., Insights into structure, dynamics and hydration of locked nucleic acid (LNA) strand-based duplexes from molecular dynamics simulations. *Nucleic Acids Res.* **2008**, *36* (5), 1508-16.
- (101) Weixlbaumer, A.; Murphy, F. V.; Dziergowska, A.; Malkiewicz, A.; Vendeix, F. a. P.; Agris, P. F.; Ramakrishnan, V., Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nat. Struct. Mol. Biol.* **2007**, *14* (6), 498-502.
- (102) Muster, T.; Steindl, F.; Purtscher, M.; Trkola, A.; Klima, A.; Himmler, G.; Rüker, F.; Katinger, H., A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J. Virol.* **1993**, *67* (11), 6642-6647.
- (103) Zwick, M. B.; Wang, M.; Poignard, P.; Stiegler, G.; Katinger, H.; Burton, D. R.; Parren, P. W., Neutralization synergy of human immunodeficiency virus type 1 primary isolates by cocktails of broadly neutralizing antibodies. *J. Virol.* **2001**, *75* (24), 12198-12208.
- (104) Zwick, M. B.; Labrijn, A. F.; Wang, M.; Spenlehauer, C.; Saphire, E. O.; Binley, J. M.; Moore, J. P.; Stiegler, G.; Katinger, H.; Burton, D. R., Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J. Virol.* **2001**, *75* (22), 10892-10905.
- (105) Sun, Z.-Y. J.; Oh, K. J.; Kim, M.; Yu, J.; Brusica, V.; Song, L.; Qiao, Z.; Wang, J.-h.; Wagner, G.; Reinherz, E. L., HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity* **2008**, *28* (1), 52-63.

Megavirales homologues of translation termination factor eRF1: protein production, homology modelling, and molecular dynamics

Yossa Dwi Hartono, Konstantin Pervushin

School of Biological Sciences, Nanyang Technological University, Singapore

Abstract

Eukaryotic release factor 1 (eRF1) facilitates termination of translation in eukaryotes in the ribosome. Genes homologous to eRF1 have been found in the family of giant viruses (*Megavirales*), but the genes are interrupted with internal stop codons. In this study we produced and purified viral RF1s (vRF1s). We also performed homology modelling and molecular dynamics and speculate on the structures of vRF1s. vRF1 structural models will be useful to further understand eRF1 function and mechanism, to explore molecular evolution of release factor in the context of eukaryotes and giant viruses, and to guide drug design efforts targeting release factors.

Introduction

Translation is the final step of genetic information transfer from RNA transcript to polypeptide. Several protein factors assist translation event in different phases of initiation, elongation, and termination. The termination phase is mainly facilitated by translation termination factors (also called release factors). Class I release factors recognise stop codons and mediate peptidyl-tRNA hydrolysis at the ribosomal peptidyl transferase centre [1]. Class II release factors mediate the release of class I release factor from the ribosome [2].

Due to their role in stop codon recognition and ribosomal A site occupancy, class I release factors are regarded as functional mimics of tRNA [3]. eRF1 comprises three structurally and functionally distinct domains: N-domain recognises stop codon at ribosomal A site, M-domain facilitates peptidyl-tRNA hydrolysis at peptidyl transferase centre (PTC), while C-domain interacts with class II release factor eRF3 [1, 4].

For stop codon recognition, TASNKS and YxCxxxF motifs are found to be conserved among standard-code organisms (i.e. they recognise the three stop codons in standard genetic code) [1, 4-6]. Another proposition is that eRF1 possesses three different cavities to accommodate the stop codon nucleotides. Though there are some overlaps of residues implicated in stop codon recognition, these two perspectives have not been fully resolved [7]. M-domain possesses GGQ motif that is universal to all species and substitutions of the Gly residues have been shown to abolish peptidyl-tRNA hydrolysis, which is evidence that this motif plays a central role in mediating peptidyl-tRNA hydrolysis at the PTC [8]. For C-domain, eRF1/eRF3a complex crystal structure has also been resolved, albeit not at high resolution and only domains 2 and 3 of eRF3a are included. Interaction happens exclusively between the C-domains of the two release factors.

The giant viruses (also known as *Megavirales*) were discovered about a decade ago, beginning with Mimivirus [9] and its complete genome was reported the following year [10]. It has large particle and genome size, with the genome size being larger than certain bacteria and archaea [10]. Genome sequencing also reveals that unlike other known viruses, the giant viruses

encode key components of translation machinery such as aminoacyl-tRNA synthetases and release factors [10]. The expression of these genes are regulated following the viral replication cycle, thus they are unlikely to be pseudogenes [11]. Simple horizontal gene transfer has been argued against, too, since there are structural differences in certain Mimivirus aminoacyl-tRNA synthetases but they are genuinely functional [12].

Jeudy et al. [13] took a closer look at the genomes of several giant viruses, particularly at the ORFs encoding release factors. From the whole genome sequencing [10], the gene products are homologous to the eukaryotic version of release factor 1, though it appears to be much shorter (Figure 1).

Figure 1. Comparison of full eRF1 sequence (green) and Mimivirus R726 gene (red). The eRF1 X-ray structure and the predicted structure of R726 gene product are also shown.

Jeudy et al. observed that Megavirus and Mimivirus ORFs have a large portion of 5' untranslated region. If this region is included, the gene would have two internal stop codons, the first one to be readthrough, the second to be +1 frameshifted. The revised annotations are compelling: they include the TASNIKS and YxCxxx F motifs, conserved in eukaryote and archaea, and are involved in stop codon recognition; as well as GGQ motif that is conserved in all organisms. Figure 2 shows the excellent alignments of human eRF1 and the viral gene products using the revised annotations and assuming appropriate recoding events of the two internal stop codons (henceforth vRF1s).

```

CLUSTAL W (1.83) multiple sequence alignment

human      MADDPSAADRNVEIWKKIKLIKSLEAARG--NGTSMISLIIPPKDQISRVAKMLADEFGT
mimivirus  -----MITDPELLQNLTQAISQINSTD LVTLYVPANNLWLFVEHINKELST
megavirus  -----MTNIIDCSLELTKTLELAESNINSTNLITMYLPPNNLWLTRDHINREIKT
              :      :* :.* * . *.*.:.: :*:.:. : : *:*

human      ASNIKSRVNRLSVLGAITSVQORLKLYNKVPPNGLVVYCG-----TIVTEEGKEK
mimivirus  APNIKNKQLRKSVIQNLQTINYQLKTIKNI PETGLVVCAGNYTIKKTHTTNSREILSCLS
megavirus  ANNIKNKILRKSI IEALASVNYQLKLLTDIPEYGMIIYAG-----DFRLPNNVEHPLCLN
              * **.: * *: : : : : ** ..:* *.:. . *

human      KVNIDFEPFKPINTSLYLCDNKFHTEALTALLS-DDSKFGFIVIDGSGALFGTLQGNTR
mimivirus  KNYHVFVPPNPVVKFY YKCDKKFHLEDLIKLYNDNGKKAIVLISGKVTYFYQYTVNNTK
megavirus  SNCILLRPHHTVKKFF YKCDKKFYVDDIINMFN-NYQDYAIVLISGKKTQFYTHNIHDTK
              . : * :. . * **.: : : : : . : .:.*. : * : :

human      VLHKFTVDLPKKHGRGGQ SALRFARLRMEKRHNYVRKVAETAVQLFISGDKVNVAAGLVLA
mimivirus  LLKTIQDIPNKHKTGGQSAQRFERIRDGSIKQYSKKILEIMIQLYTSNGNFDTCLILA
megavirus  LIKTLTETLPNQHTGGQSAVRFERIRGEKIHVYIKKILDNMIQFYTSNGKFRLTGIIIA
              :.:.* :*:.* ***** ** *.* . : * :*: : :*:.*.: : :*:

human      GSADFKTELSQSDMFQRLQSKVLKLVDISYGGENGFNQAIELSTEVLNS--VKFIEQEK
mimivirus  GPAEMKNIVIDHDLFS-IFKDHVSCIHNISEITDNSISQVVSMSIESIDS--DNIDSIK
megavirus  GPAELKDLVIQDDIFNTIFKQYLSKTLTISEILDNSIHKVITLSQDILDNLNNDHNLN
              *.*:* : : *.*. :. : ** :*. :. : * : :. . : :

human      KLIGRYFDEISQDTGKYCFGVEDTLKALEMGAVEILIVYENLDIMRYVLHCQGTEEEKIL
mimivirus  EFENKLQNPI--KTNLFVFGSKI VLKLFGLNRLSDIYITSDDYDIETILENRGK-----
megavirus  KLQNI IQDPR--QISLLVFGNDDIELHMNRGELKEIYVFHEYPNINYILNLNYK-----
              :. . : . . ** . : . : : : : :. :*. . .

human      YLTPEQEKDKSHFTDKETGQEHELIESMPLLEWFANNYKKFGATLEIVTDKSQEGSQFVK
mimivirus  -----CSVHYMDPILFK-
megavirus  -----TIIHVIKSKHFVN
              : . *

human      GFGGIGGILRYRVDFQGM EYQGGDEFFDLDDY-
mimivirus  KYGDI VGVKYYEYDYNDEYDNY-----
megavirus  QYGYLVGIKYYHINNEN-----
              :* : * : * . : :

```

Figure 2. CLUSTAL W alignment [14] of human eRF1 and Mimivirus/Megavirus vRF1s.

Although phylogenetically close to the eukaryotic version of class I RF, there seems to be no homologues of class II RF in Mimivirus/Megavirus genome, although in eukaryotes effective termination would involve the interactions of the two classes of RFs. Several hypotheses can be proposed to explain this anomaly: it may use another GTPase to function as class II RF (as archaea are known to do so); it may recruit the host's class II RF; it may do away with class II RF.

It is known that mutation of the TASNIKS motif can affect whether class II RF is needed to terminate at UAA and UAG stop codons [15, 16] and this motif is present in vRF1 sequences with deviation from the consensus TASNIKS (TAPNIKN for Mimivirus and TANNIKN for Megavirus). There is

significant deletion in the vRF1 sequences near the C-terminal (Figure 2). Since C-terminal domain interacts with class II RF, this may imply that vRF1 does not interact with class II RF.

It is known in some bacteria that RF2 gene contain an internal stop codon as a regulatory element [17]. In such cases, RF2 gene comprises two overlapping ORFs that need frameshift to produce full-length RF2. The feedback loop regulation is as follows: when RF2 is abundant, translation is efficiently terminated at the first ORF; truncated RF2 is thus produced, limiting the amount of full-length RF2; in turn, the termination efficiency at the internal stop codon is reduced and frameshift frequency increases; in turn, producing full-length RF2. If a similar system is being employed in the giant viruses RF1, it is also curious that the regulation involves two switches instead of only one. It could be speculated that a more delicate fine-tuning is needed for the virus since it needs to recruit the rest of the translation machinery from its host.

Nonsense mutations cause “approximately 20–40% of the individual cases of more than 240 inherited diseases” [18]. Frameshift and nonsense mutations in regulatory genes are also associated to the most common forms of a several cancers [18]. More specifically, premature stop codon (PTC) in tumour suppressor (TS) gene results in defective protein that is unable to properly inhibit cell proliferation or facilitate apoptosis [19]. TS genes have also been pinpointed as good targets for PTC suppression because they have disproportionately more nonsense mutations compared to oncogenes, which are largely rendered inactive by missense mutations [19].

The initial aim of this work is structural characterisation of vRF1s with X-ray crystallography complemented with NMR spectroscopy. However, protein production and purification of vRF1s proved to be a challenge. We managed to optimise a purification protocol, but did not manage to proceed with X-ray crystallography stage. Instead, we complemented the experimental work with computational investigations using homology modelling and molecular dynamics simulation.

Materials and methods

Cloning and plasmid construction

pUC57 cloning vectors containing Mimivirus, Megavirus and Marseillevirus vRF1 genes are ordered from ShineGene Bio-Technologies, Inc (Shanghai, China). For Mimivirus and Megavirus vRF1, the genes have been edited to remove the two internal stop codons (Trp-substitution readthrough and +1 frameshift).

The genes are subcloned into the expression vector using ligation independent cloning (LIC). Several primers are designed such that variable amino acid boundaries of the gene product constructs would be produced. Nine constructs for each vRF1 are initially designed for small-scale expression screening to assess the yield and solubility of the proteins. Additional four constructs are designed for Marseillevirus vRF1 owing to poor yield and solubility of eight out of nine initial constructs.

Expression vector used is pNIC28-Bsa4 from Structural Genomics Consortium (SGC) [20]. This pET-derived vector has N-terminal hexa-Histidine tag with TEV protease cleavage site, LIC sites, Kanamycin resistance, and T7-LacO promoter. *E. coli* strain used for expression is BL21(DE3) Rosetta T1R (Chloramphenicol resistant). The work involving subcloning and small-scale expression was done by Protein Production Platform (PPP, <http://proteins.sg>)

Glycerol stocks of *E. coli* expression strain containing the expression vector are obtained from PPP for large-scale expression. The expression vector is amplified using Miniprep Plasmid Kit from Qiagen (Hilden, Germany) following the manufacturer's protocol and sequence-verified.

Protein production

These stock solutions are made with ultrapure water and filter-sterilised: 1 M 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (HEPES) (AppliChem, Germany) with pH adjusted to 7.5 with 10 M NaOH (Merck, Germany), 4 M NaCl (Merck, Germany), 2 M imidazole (Sigma-Aldrich, USA), 1 M dithiothreitol (DTT) (AppliChem, Germany), 1 M isopropyl β -D-1-

thiogalactopyranoside (IPTG) (Gold BioTechnology, USA), 50 mg/ml Kanamycin (Kan) (Invitrogen, USA), 20 mg/ml Chloramphenicol (Cam) (AppliChem, Germany), 0.34 M / 1.44 M $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ (Merck, USA), 0.5 M ethylenediaminetetraacetic acid (EDTA) (USB, USA) and 1 M nickel (II) sulphate (AppliChem, Germany). No stock solutions of L-Arginine (Sigma-Aldrich, USA) and L-glutamic acid (AppliChem, Germany) are prepared owing to their low solubility; when preparing the buffer, the solid is added directly instead.

2 L of Terrific Broth (TB) medium is made by mixing 24 g tryptone (BD, USA), 48 g yeast extract (BD, USA), and 50 ml 50% glycerol (Promega, USA) in 1800 ml distilled water. After autoclaving, the mixture is allowed to cool to room temperature, after which 50 ml 0.44 M / 1.77 M $\text{K}_2\text{HPO}_4/\text{KH}_2\text{PO}_4$ and necessary antibiotics are added. The antibiotics used in plates and media are 50 $\mu\text{g/ml}$ Kan and 20 $\mu\text{g/ml}$ Cam.

E. coli expression strain is subjected to several rounds of colony selection to maximise protein yield. This is done by plating from glycerol stock on Luria Bertani (LB) agar plate overnight at 37°C. Several colonies are picked and inoculated each to 2 ml TB medium, then incubated at 37°C, 180 rpm. After approximately 6 hours, glycerol stock for each colony is prepared by extracting 100 μl from each culture, mixing with 100 μl sterilised 50% glycerol, then stored in -80°C. Incubation temperature is lowered to 18°C. After 1 hour, the cultures are induced with 0.5 mM IPTG, and grown at the same temperature overnight. Protein expression level from each colony is then assessed by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) using 10% or 12% SDS-polyacrylamide gel. This constitutes one round of selection. The colony which exhibits the highest protein yield is then plated again for another round of selection. Typically 3-4 rounds of colony selection are done for each vRF1.

For large scale expression, the glycerol stock is inoculated to 100 ml of TB medium in 500 ml conical flask, incubated at 37°C, 180 rpm overnight. The inoculum is inoculated to 4 L cultivation culture (2x2L TB medium in 2x5L culture flasks). After 6 hours of incubation with shaking (37°C, 180 rpm), the

temperature is lowered to 18°C. After 1 hour, the cultivation culture is induced with 0.5 mM IPTG and is harvested the next morning. Harvesting is done by spinning down the culture at 10,000x g at 4°C for 30 minutes. Supernatant is discarded and cell pellet is stored at -80°C.

Protein purification

For every 10 g of cell pellet 40 ml of lysis buffer is used to resuspend the cell pellet to make a homogeneous mixture. The composition of the lysis buffer is 20 mM HEPES, 500 mM NaCl, 10 mM imidazole, 1 mM DTT, with final pH adjusted to 8.0. Lysis is done by ultrasonication using long tip at 27% power on 1s/1s on/off cycle for 45 minutes on ice bath. For larger lysate volume, microfluidiser with pressure at 12,000 psi and 5 cycles on ice bath is used to lyse the cells. The crude lysate is centrifuged at 48,000x g for 1 hour at 4°C and the supernatant is recovered for immobilised metal affinity chromatography (IMAC).

For every 20 ml of clarified lysate, 1 ml bed volume of Ni-NTA Agarose beads (Qiagen, Germany) loaded to a 20-ml gravity flow column (Bio-Rad, USA) is used. Before loading the clarified lysate, the column is equilibrated with 4 ml of wash I buffer. The composition of the wash I buffer is 20 mM HEPES, 500 mM NaCl, 20 mM imidazole, 1 mM DTT, with final pH unadjusted (the pH of 1 M HEPES stock solution is adjusted to 7.5). The beads are incubated with the clarified lysate for 1 hour with gentle shaking in cold room. Flowthrough is discarded and the beads are washed with 25 ml wash I buffer. The beads are then further washed with 5 ml of wash II buffer. The composition of wash II buffer is the same as wash I buffer, with the NaCl concentration reduced to 250 mM and addition of 50 mM Arg and 50 mM Glu. The bound protein is eluted with 10 ml of elution buffer (20 mM HEPES, 250 mM NaCl, 250 mM imidazole, 50 mM Arg, 50 mM Glu, 1 mM DTT, final pH unadjusted). The eluate is concentrated using centrifugal concentrator (Merck Millipore, MWCO 30 kDa) by spinning down at 3,800x g at 4°C.

The beads are washed with 1 column volume of ultrapure water, 1 column volume of 0.5 M NaOH, 1 column volume of ultrapure water, and stored in 20% ethanol (Merck, Germany). When the beads have been used

several times, Ni^{2+} are stripped using 5 ml 0.1 M EDTA and regenerated with 5 ml 0.1 M nickel (II) sulphate to restore hexa-Histidine binding efficiency. 1 column volume of ultrapure water is used in between the steps to wash the excess stripping/regeneration reagents.

For size-exclusion chromatography (SEC) using HiLoad 16/60 Superdex 75 column (GE Healthcare, USA) and AKTA FPLC (GE Healthcare, USA), the sample is concentrated to less than 4 ml of volume. The composition of SEC buffer is 20 mM HEPES, 175 mM NaCl, 50 mM Arg, 50 mM Glu, 1 mM DTT, with final pH adjusted to 8.0. Before sample loading, the column is equilibrated with 1.5 column volume of SEC buffer. Marseillevirus vRF1 vc002 (~45 kDa) elutes at around 50 ml after injection. The column is washed with 2 column volume of ultrapure water and is stored in 20% ethanol.

For ion exchange chromatography (IEX) using HiTrap Q HP column (GE Healthcare, USA) and AKTA FPLC (GE Healthcare, USA), the sample is buffer-exchanged using PD-10 columns (Bio-Rad, USA) to IEX A buffer. The composition of IEX A buffer is the same as SEC buffer. The composition of IEX B buffer is the same as IEX A buffer, with NaCl concentration increased to 1 M. After sample injection, the column is washed with IEX A buffer until UV signal returns to baseline. The sample is eluted out using ionic strength gradient (0% IEX B to 100% IEX B in 150 minutes). Marseillevirus vRF1 vc002 is eluted at around 40-50% conductivity. The column is washed with more than 10 column volume of IEX B (high salt wash), then the same volume of ultrapure water, and is stored in 20% ethanol. Protein purity is assessed by SDS-PAGE. If contaminants are still observed, the sample is buffer-exchanged to IEX A to undergo another round of IEX chromatography. For long-term storage, the protein is buffer-exchanged to IEX A, concentrated, and stored in -80°C .

Molecular dynamics

For human eRF1, the X-ray crystal structure was obtained from Protein Data Bank (PDB 1DT9 [4]). Missing loops are modelled using SWISS-MODEL [21-25]. GROMACS Version 4.5.6 [26-28] *mdrun* module is used for MD production with Amber99SB*-ILDN force field [29, 30]. TIP3P water

molecules [31] were added and minimal distance of cubic box is set as 1.0 nm. Na⁺ and Cl⁻ ions are added to maintain electrostatic neutrality; NaCl salt concentration is set at 0.3 M. The protein is energy-minimised by steepest descent until the maximum force is smaller than 1000.0 kJ mol⁻¹nm⁻¹. Coulombic non-bonded cutoff distance is set at 1.0 nm. LINCS algorithm is applied to constrain all bonds [32, 33]. After minimisation and 100 ps solvent equilibration, production MD is carried out at 300 K, with time-step of 0.002 ps for 100,000 ps.

For the vRF1s, homology models are obtained with I-TASSER [34-36], with the model that uses eRF1 X-ray crystal structure as template used as starting structures for MD production. The MD production parameters are the same as eRF1.

For post-production analysis, water molecules and ions are removed from the trajectory and each frame is rotationally and translationally fitted to the starting structure for visualisation. Analysis is performed using various GROMACS modules; Dictionary of Secondary Structure of Protein (DSSP) analysis is done using *dssp* programme [37, 38]; binding pocket analysis is done using *mdpocket* in *fpocket2* package [39, 40]; plots are drawn with *xmgrace*; visualisation is done using VMD [34].

Result and discussion

Protein production and purification

Figure 3 shows nine constructs for each vRF1 (vc001-vc009) which are all the nine possible combinations of three starting amino acids and three ending amino acids. The constructs are generally designed to end with charged amino acids at the termini. For Marseillevirus vRF1, only vc002 expresses good amount of soluble protein. Four additional constructs are designed (vc010-vc013), but none gives good soluble yield. For Megavirus vRF1, vc002 and vc003 gives good soluble yield. Finally, for Mimivirus vRF1, vc001-vc006 gives good soluble yield.

Figure 3. Various constructs and small-scale expression of vRF1s by PPP (<http://proteins.sg>). The constructs have different combinations of starting and ending amino acids.

The protein yield for Marseillevirus vRF1 vc002 is then optimised by colony selection. Overexpression level is reasonably high, considering that the amount of target protein is significantly more than that of the rest of *E. coli* proteins. However, the soluble yield is not so high, because significant amount ends up in the insoluble fraction (inclusion body), as shown in Figure 4A. IMAC elution profile shows that the protein readily binds to the Ni^{2+} beads at lysis buffer condition (pH 8.0) and dissociates as expected in increasing imidazole concentration.

Figure 4. SDS-PAGE: Lysis and IMAC of Marseillevirus RF1 vc002

The protein is observed to readily precipitate at initial elution buffer condition (without Arg+Glu), especially when the sample is concentrated to small volume for SEC. Due to this, pH screening of the buffer was also attempted by buffer-exchanging concentrated IMAC eluate to several SEC buffers at constant NaCl concentration of 300 mM and varying pH values (7.0, 7.5, 8.0, 8.5, 9.0). The protein is observed to precipitate less at higher pH (8.0-9.0). pH 8.0 is chosen for subsequent purification attempts so as not to deviate too much from physiological pH.

For SEC, the chromatogram shows a single peak eluting at about 45 ml after injection. However, SDS-PAGE run shows that the target protein is still poorly separated from the contaminants in the elution fractions. With addition of Arg+Glu in the SEC buffer, more protein can be concentrated due to increased protein solubility, but separation is still poor (Figure 5).

Size-exclusion chromatogram (without Arg+Glu)

EC E1 E2 E3 E4 E5 E6 M

M E7 E8 E9 E10 E11 E12 EC

EC IMAC eluate concentrate
E1-E12 Elution fractions

A4-B4 Elution fractions

Figure 5. Size exclusion chromatogram and SDS-PAGE of Marseillevirus RF1 vc002 purification.

Since SEC is not giving good separation, IEX chromatography is attempted instead as to purify the protein after IMAC. Since the predicted pI is 5.59 and the buffer pH is 8.0, it is assumed that the protein is negatively charged at this condition and anionic IEX column (IEX Q HP) is used. Starting NaCl concentration is determined by trial-and-error. 175 mM NaCl is observed to be high enough that the protein does not readily aggregate, leading to precipitation, but low enough that the protein binds to the IEX column. After several IEX runs to refine the parameters, the run condition as detailed in Materials and Methods section was obtained. Good separation is achieved when ionic strength gradient is set longer than 60 minutes. Marseillevirus vRF1 vc002 elutes at around 45% conductivity.

IEX chromatogram (with Arg+Glu), 60-min ionic strength gradient

A3 A7 A8 A9 A10 A11 A12 M B1 B2 B3 B4 B5

A3-B5 Elution fractions

Figure 6. IEX (anionic) chromatogram and SDS-PAGE of Marseillevirus vRF1 vc002 purification.

The use of 50 mM Arg+Glu additive to improve protein solubility and stability has been described elsewhere [41]. This improvement is observed to be also true for Marseillevirus vRF1 vc002. The additive is thus added to the purification buffers to maximise the amount of soluble target protein in purification steps. However, one should not add the additive to lysis buffer where the lysate is to be loaded to IMAC column, because presence of charged amino acids presumably interferes with Ni²⁺-to-His-tag binding because no protein is eluted when this was attempted. The composition of the purification buffers is thus designed to maximise the soluble protein yield (Figure 7).

Lysis	:	500	mM NaCl, 0	mM Arg+Glu, 10	mM imidazole
Wash I	:	500	mM NaCl, 0	mM Arg+Glu, 20	mM imidazole
Wash II	:	250	mM NaCl, 50	mM Arg+Glu, 20	mM imidazole
Elution	:	250	mM NaCl, 50	mM Arg+Glu, 250	mM imidazole
IEX A	:	175	mM NaCl, 50	mM Arg+Glu, 0	mM imidazole
IEX B	:	1000	mM NaCl, 50	mM Arg+Glu, 0	mM imidazole

Figure 7. Variation in salt and additives in purification buffers

In theory, one can add the additive in the lysate and load the clarified lysate to IEX column and this was attempted (Figure 8). However, in practice there are several difficulties that were encountered with this strategy. First, the chromatogram is quite crowded and it is be difficult to see which fractions contain the target protein. Second, the fractions would be expected to still have a lot of contaminants, even with a long ionic strength gradient. Third, to remove the contaminants, then the next appropriate purification stage would be IMAC; the additive has to be removed by buffer exchange. This would negate the benefits of adding the additive in the first place; some amount of protein would precipitate out and reduce the yield.

IEX chromatogram (with Arg+Glu), 60-min ionic strength gradient

1-28 Elution fractions in increasing ionic strength

Figure 8. IEX (anionic) chromatogram and SDS-PAGE of Marseillevirus vRF1 vc002

purification. Clarified lysate was directly loaded to IEX column without IMAC intermediary.

Molecular dynamics

Cartoon representations of eRF1 X-ray crystal structure (1DT9) and I-TASSER homology models of vRF1s are shown in Figure 9. For all three vRF1s, the highest global structural similarity (TM-score) is with 1DT9; so models used for the starting structures are similar in their global structures to 1DT9. A notable difference is that the N-domains of the vRF1s seem to have longer unstructured loops.

Figure 9. Human eRF1 and vRF1 I-TASSER homology models

Another conspicuous feature of the homology models is that they are missing the minidomain located in the C-domain (Figure 9). The minidomain is not well-defined in the X-ray crystal structure but solution NMR study of the C-domain shows that the minidomain exists in two conformational states [42]. From superimposition to the crystal structure, the minidomain is located near N-domain and it was also shown that mutation at the tip of the minidomain

affects stop codon recognition (function of N-domain). Regarding the absence of the minidomain in the vRF1 homology models, indeed the sequence alignment shows that vRF1 sequences have a missing stretch of amino acids near the C-terminal, but it is not clear whether this is due to missing template (since the crystal structure is missing most of the minidomain) or they are really absent in vRF1s. This remains to be answered and resolving the structures of vRF1s will hopefully provide the answers.

Large conformation change is expected when eRF1 binds to eRF3 and subsequently, the ribosome. Comparison between several available structural models of eRF1 reveals large movement of M-domain. Figure 10 shows superimposed eRF1 structural models from various systems: eRF1 alone (PDB 1DT9 [4]), eRF1/eRF3_{2,3} (PDB 3E1Y [43]), eRF1/eRF3/ribosome (cryo-EM model PDB 3J2K [44]). Taking C-domain as superimposing reference, M-domain appears to move closer towards C-domain. If we consider the centres of mass of the three domains as three points in space (N, M, C) and the centre of mass of the whole protein as another (O), the angle between M-domain and C-domain ($\angle\text{MOC}$) decreases from 133° to 126° then 121° as the size of eRF1 complex increases, which is a significant range of motion.

The model shown in Figure 10 (purple) is an atomistic model obtained by running MD simulation on restricted conformational space defined by the cryo-EM electron density map, termed Molecular Dynamics Flexible Fitting (MDFF)¹. The long extended α -helix of M-domain appears to be bent in this model (3J2K), which is not energetically favourable, compared to mostly straight in the crystal structure (1DT9 and 3E1Y). It should be emphasised that atomistic depiction of PDB 3J2K must be carefully considered, owing to cryo-EM resolution. However, this model is consistent with the ribosomal context. The distance between ribosomal decoding site and peptidyl transferase centre is ~73 Å [45]. In 1DT9, the distance between NIKS and GGQ motifs is 98 Å (measured between NIKS Ile62 C α and GGQ Gly184 C α), while in 3E1Y, this distance is 92 Å. In 3J2K MDFF atomistic model, this distance is 74 Å, which

¹ Atomistic model (MDFF) was obtained from personal communication with the authors

is a good agreement. The helix bending observed in this model could be a possible explanation of how eRF1 accommodates itself into the ribosome, although, again, considering the cryo-EM resolution, this is still open to investigation.

Figure 10. Large domain movement in human eRF1

Long MD simulation time (100 ns) is used to investigate large movement such as domain movement occurs in longer timescale. The simulation will also be repeated three times and the average will be taken for analysis for better statistical sampling. In this report, only the first simulation is presented as the replicate simulations are ongoing. Human eRF1 is observed to reach equilibrium after about 10 ns, as seen from RMSD plot (Figure 11). N-domain RMSD is observed to fluctuate at lowest range of values compared to global RMSD and that of the other two domains. M-domain is observed to have highest RMSD, and the values also vary with greater fluctuation, indicating significant conformational dynamics. All global and domain fluctuations follow similar trend with global maximum at 57.6 ns, followed by local maxima at 37.1, 77.5, and 98.5 ns.

Figure 11. Human eRF1 RMS deviation

Figure 12. RMS atomic fluctuation (heavy atoms, 10-100 ns)

RMSF plot shows certain residues that contributes to the intradomain dynamics of eRF1 (Figure 12). In N-domain, most fluctuations are <0.5 nm. A notable maximum occurs in residue Arg-65, right after the TASNIKS motif, which is involved in stop codon recognition. The N-terminal tail is surprisingly not very dynamic, staying around 0.3 nm of fluctuation. In M-domain, there is considerable dynamics around GGQ motif which interacts with peptidyl transferase centre (PTC) in the ribosome. The maximum occurs at Arg-182, right before the GGQ motif. This is consistent with the fact that the region around GGQ motif is an unstructured loop and considerable flexibility is needed to close the distance gap to reach the ribosomal PTC (from 98 Å in crystal structure to 73 Å in the ribosome). The C-domain shows fluctuation comparable to GGQ loop, specifically in the minidomain region that is mostly not visible in the crystal structure. The C-terminal tail shows the highest fluctuation of 1.0 nm.

Figure 13. DSSP analysis of eRF1

DSSP analysis provides additional insight of protein dynamics at the secondary structure level. Figure 13 shows various secondary structures that are present throughout the simulation. Overall, most secondary structures are preserved. Loop around TASNKS motif form a transient 3-helix for the first 10 ns but reverts to unstructured coil for the rest of the simulation. Loop around GGQ motif shows the opposite behaviour – it starts as unstructured but part of the loop forms 3-helix at around 7 ns and this is maintained throughout the simulation. The minidomain shows the most variation in secondary structure dynamics throughout the simulation. There is formation of parallel sheet at 7 ns which is then maintained for the rest of the simulation. There is also appearance of transient 3-helix which is continuously broken and formed throughout the simulation. The helical structure at this region is consistent with the NMR studies, although no distinct open and closed conformers, as observed in NMR studies [42], are observed in the simulation. In the open conformer, the aromatic residues seem to stabilise the structure by stacking. In the simulation, no such interaction is observed (Figure 14).

Open (D) and closed (E) minidomain conformers from NMR structures, taken from [42]

Snapshots of minidomain during simulation

Figure 14. Conformation of minidomain

Figure 15. Angular orientation of human eRF1 domains relative to each other.

Considering the large range of movement that occurs when eRF1 binds to eRF3 and ribosome, significant interdomain dynamics of eRF1 in solution is

expected. The hinge regions between the three domains have low fluctuation compared to the rest of the protein (Figure 12), indicating that the hinges are quite rigid. To track domain movement during simulation, interdomain angles are also calculated by aforementioned method (form vectors corresponding to centres of mass of the individual domains and the whole protein, then calculate angle between the vectors), as shown in Figure 15. Large movement of M-domain is observed during simulation. Angle $\angle\text{MOC}$ is observed to decrease from 133° (0 ns) to 120° at equilibrium, occasionally even lower to about 115° . Correspondingly, this is accompanied with less pronounced increases in angles $\angle\text{NOM}$ and $\angle\text{NOC}$. The values of interdomain angles appear stable during 10-100 ns, indicating that either this is a more stable conformation at equilibrium, or it might be the case that M-domain rocks back and forth in very low frequency outside the timescale of this simulation. In the structural models, although there is also large domain movement when eRF1 binds to eRF3 and, subsequently, the ribosome, the direction of movement is quite different from the simulation. While it is consistent that M- and C-domains are moving closer, the trajectory of approach of M-domain is clearly different, as shown in Figure 16. The bending of M-domain to accommodate itself to the ribosome thus may be a conformation that is accessible only when eRF3 and/or ribosome is present, as it is not observed in equilibrium state of the simulation.

Figure 16. Various eRF1 structural models superimposed at C-domain.

Lastly, MD simulation trajectory, which contains information about conformational dynamics of a protein, can also be used to find potential ligand binding site. Often, the search for such cavity on a static snapshot of a crystal structure is a significant challenge since in its solution state, a cavity might open transiently. A previous computational search with the crystal structure was not very successful in finding an obvious cavity for ligand binding (data not shown). Figure 17 shows cavities that are conserved throughout the simulation with several plots of the cavity volume are shown. There are two big cavities, one between N- and C-domain and one near the hinge region. While big cavities are good in providing steric space, the fluctuation in volume is also great in amplitude, considering that they are interdomain space. There are also conserved small cavities in N- and M-domains that are not prominent in 0 ns (crystal structure), but gain volume as the simulation proceeds. It should be noted that in view of strong correlation between domain movement and eRF1 function, targetting the hinge region of eRF1, although challenging, should be kept in mind. Armed with the dynamics information, computational small-ligand docking might be attempted more successfully.

Figure 17. Conserved cavity in eRF1 throughout simulation time.

RMSD fluctuations of Marseillevirus and Megavirus vRF1s are at the same scale of magnitude as eRF1 (Figure 18). Marseillevirus vRF1 reaches equilibrium at around 40 ns; Megavirus vRF1, around 15 ns. Mimivirus vRF1 shows big fluctuation and only reach equilibrium at around 60 ns. In terms of domain movement, Mimivirus vRF1 seems to have larger N- and M-domain movement (Figure 19). Megavirus vRF1 has large M-domain movement but not much that of N-domain. Marseillevirus vRF1 has very little N-domain movement and moderate M-domain movement.

Figure 18. RMSD of all four MD systems

Figure 19. Snapshots of vRF1 during simulation.

Since Marseillevirus vRF1 has the highest sequence similarity with human eRF1 compared to the other two vRF1s, it is expected that it will behave similarly with eRF1. Atomic fluctuation profile of Marseillevirus vRF1 share similar features with eRF1. The greatest fluctuation comes from the loop containing the GGQ motif, with domain near N-terminal having moderate fluctuation. Fluctuation associated with minidomain in eRF1 is conspicuously absent from all the vRF1s, since vRF1s do not have a region homologous to the minidomain. Megavirus and Mimivirus vRF1s have another conspicuous peak near atomic index 1500-1700, which correspond to residues not present in eRF1 and Marseillevirus vRF1 when all four are sequence-aligned. At atomic index 4500-5000, Marseillevirus vRF1 shows low fluctuation, while Megavirus and Mimivirus vRF1 shows certain amount of fluctuation similar to eRF1 at the corresponding region. At secondary structure level, without a corresponding minidomain region, all the three vRF1s generally shows secondary structure stability, as shown in Figure 21. Megavirus vRF1 has small stretches of residues around residues 111 and 255 that show intermittent breaking and reforming.

Figure 20. RMSF of vRF1s

Figure 21. DSSP analysis of vRF1s.

Conclusion

In conclusion, we have attempted to produce and purify viral RF1s, homologs of eRF1 present in *Megavirales* genome. We offered preliminary computational investigations of eRF1 and vRF1s. The polypeptide release mechanism of eRF1 is not yet fully understood and structural models of vRF1s will aid in this direction. *Megavirales* as a novel group of viruses are also still poorly understood and it is intriguing that non-pseudogene that is homologous to eukaryotic release factor 1 should be present in their genomes. The structural models will hopefully further our knowledge of molecular evolution of the release factors. Lastly, eRF1 is a potential drug target and structural models of its homologues will provide a more complete picture of its structure and dynamics and aid in design of small-molecule binders.

References

1. Frolova, L.Y., T.I. Merkulova, and L.L. Kisselev, *Translation termination in eukaryotes: Polypeptide release factor eRF1 is composed of functionally and structurally distinct domains*. Rna-a Publication of the Rna Society, 2000. **6**(3): p. 381-390.
2. Petry, S., A. Weixlbaumer, and V. Ramakrishnan, *The termination of translation*. Current opinion in structural biology, 2008. **18**(1): p. 70-7.
3. Ito, K., et al., *Conserved motifs in prokaryotic and eukaryotic polypeptide release factors: tRNA-protein mimicry hypothesis*. Proceedings of the National Academy of Sciences, 1996. **93**(11): p. 5443-5448.
4. Song, H.W., et al., *The crystal structure of human eukaryotic release factor eRF1 - Mechanism of stop codon recognition and peptidyl-tRNA hydrolysis*. Cell, 2000. **100**(3): p. 311-321.
5. Bertram, G., et al., *Endless possibilities: translation termination and stop codon recognition*. Microbiology-Uk, 2001. **147**: p. 255-269.
6. Conard, S.E., et al., *Identification of eRF1 residues that play critical and complementary roles in stop codon recognition*. Rna-a Publication of the Rna Society, 2012. **18**(6): p. 1210-1221.
7. Wong, L.E., et al., *Selectivity of stop codon recognition in translation termination is modulated by multiple conformations of GTS loop in eRF1*. Nucleic Acids Research, 2012. **40**(12): p. 5751-65.
8. Frolova, L.Y., et al., *Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis*. Rna-a Publication of the Rna Society, 1999. **5**(8): p. 1014-1020.
9. La Scola, B., et al., *A giant virus in amoebae*. Science, 2003. **299**(5615): p. 2033-2033.
10. Raoult, D., et al., *The 1.2-megabase genome sequence of Mimivirus*. Science (New York, N.Y.), 2004. **306**(5700): p. 1344-50.
11. Legendre, M., et al., *mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus*. Genome research, 2010. **20**(5): p. 664-674.
12. Abergel, C., et al., *Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS*. Journal of virology, 2007. **81**(22): p. 12406-12417.
13. Jeudy, S., et al., *Translation in Giant Viruses: A Unique Mixture of Bacterial and Eukaryotic Termination Schemes*. PLoS genetics, 2012. **8**(12): p. e1003122.
14. Larkin, M., et al., *Clustal W and Clustal X version 2.0*. Bioinformatics, 2007. **23**(21): p. 2947-2948.
15. Fan-Minogue, H., et al., *Distinct eRF3 requirements suggest alternate eRF1 conformations mediate peptide release during eukaryotic translation termination*. Molecular cell, 2008. **30**(5): p. 599-609.
16. Betney, R., et al., *Autoregulatory systems controlling translation factor expression: Thermostat-like control of translational accuracy*. Rna, 2010. **16**(4): p. 655-663.
17. Craigen, W.J. and C.T. Caskey, *Expression of peptide chain release factor 2 requires high-efficiency frameshift*. 1986.
18. Welch, E.M., W. Wang, and S.W. Peltz, *Translation termination: it's not the end of the story*. Translational control of gene expression, 2000: p. 467-485.
19. Mort, M., et al., *A meta-analysis of nonsense mutations causing human genetic disease*. Human mutation, 2008. **29**(8): p. 1037-1047.
20. Savitsky, P., et al., *High-throughput production of human proteins for crystallization: the SGC experience*. J Struct Biol, 2010. **172**(1): p. 3-13.
21. Guex, N. and M.C. Peitsch, *SWISS - MODEL and the Swiss - Pdb Viewer: an environment for comparative protein modeling*. electrophoresis, 1997. **18**(15): p. 2714-2723.
22. Schwede, T., et al., *SWISS-MODEL: an automated protein homology-modeling server*. Nucleic acids research, 2003. **31**(13): p. 3381-3385.
23. Kopp, J. and T. Schwede, *The SWISS-MODEL Repository: new features and functionalities*. Nucleic acids research, 2006. **34**(suppl 1): p. D315-D318.
24. Kiefer, F., et al., *The SWISS-MODEL Repository and associated resources*. Nucleic acids research, 2009. **37**(suppl 1): p. D387-D392.
25. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling*. Bioinformatics, 2006. **22**(2): p. 195-201.
26. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free*. Journal of computational chemistry, 2005. **26**(16): p. 1701-1718.
27. Hess, B., et al., *GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation*. Journal of chemical theory and computation, 2008. **4**(3): p. 435-447.
28. Berendsen, H.J., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation*. Computer Physics Communications, 1995. **91**(1): p. 43-56.

29. Best, R.B. and G. Hummer, *Optimized molecular dynamics force fields applied to the helix– coil transition of polypeptides*. The Journal of Physical Chemistry B, 2009. **113**(26): p. 9004-9015.
30. Lindorff - Larsen, K., et al., *Improved side - chain torsion potentials for the Amber ff99SB protein force field*. Proteins: Structure, Function, and Bioinformatics, 2010. **78**(8): p. 1950-1958.
31. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water*. The Journal of chemical physics, 1983. **79**: p. 926.
32. Hess, B., *P-LINCS: A parallel linear constraint solver for molecular simulation*. Journal of Chemical Theory and Computation, 2008. **4**(1): p. 116-122.
33. Hess, B., et al., *LINCS: a linear constraint solver for molecular simulations*. Journal of computational chemistry, 1997. **18**(12): p. 1463-1472.
34. Roy, A., J. Yang, and Y. Zhang, *COFACTOR: an accurate comparative algorithm for structure-based protein function annotation*. Nucleic acids research, 2012. **40**(W1): p. W471-W477.
35. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction*. Nature Protocols, 2010. **5**(4): p. 725-738.
36. Zhang, Y., *I-TASSER server for protein 3D structure prediction*. BMC Bioinformatics, 2008. **9**.
37. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features*. Biopolymers, 1983. **22**(12): p. 2577-2637.
38. Joosten, R.P., et al., *A series of PDB related databases for everyday needs*. Nucleic acids research, 2011. **39**(suppl 1): p. D411-D419.
39. Schmidtke, P., et al., *MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories*. Bioinformatics, 2011. **27**(23): p. 3276-3285.
40. Le Guilloux, V., P. Schmidtke, and P. Tuffery, *Fpocket: an open source platform for ligand pocket detection*. BMC bioinformatics, 2009. **10**(1): p. 168.
41. Golovanov, A.P., et al., *A simple method for improving protein solubility and long-term stability*. J Am Chem Soc, 2004. **126**(29): p. 8933-9.
42. Mantsyzov, A.B., et al., *NMR solution structure and function of the C-terminal domain of eukaryotic class I polypeptide chain release factor*. The FEBS journal, 2010. **277**(12): p. 2611-27.
43. Cheng, Z., et al., *Structural insights into eRF3 and stop codon recognition by eRF1*. Genes & Development, 2009. **23**(9): p. 1106-1118.
44. Taylor, D., et al., *Cryo-EM structure of the mammalian eukaryotic release factor eRF1-eRF3-associated termination complex*. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(45): p. 18413-8.
45. Jackson, R.J., C.U. Hellen, and T.V. Pestova, *Termination and post-termination events in eukaryotic translation*. Adv Protein Chem Struct Biol, 2012. **86**: p. 45-93.

Modeling pK shift in DNA triplex containing locked nucleic acids

Yossa Dwi Hartono^{1,2} *, You Xu¹*, Andrey Karshikoff¹ , Lennart Nilsson¹ and Alessandra Villa¹

¹ Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden.

² Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

* These authors contributed equally to this work

Corresponding author: Alessandra Villa, alessandra.villa@ki.se

Keywords: LNA, DNA, pK, triplex, lambda dynamics

Abstract

The protonation state for nucleic acids base is experimentally difficult to assess. In the context of triplex DNA, the protonation state of cytidine in the third strand is particularly important, because it needs to be protonated in order to form Hoogsteen hydrogen bonding. A sugar modification, locked nucleic acid (LNA), is widely in triple forming oligonucleotide to target sites in the human genome. In this study, the parameters for LNA are developed in line with the CHARMM nucleic acid force field and validated towards the available structural experimental data. In conjunction, two computational methods were used to calculate the protonation state of cytidine in various DNA triplex environments: λ -dynamics and multiple pH regime. Both approaches predict pK shift greater than physiological pH when cytidine is in third strand in a triplex environment. The predicted values significantly differ with λ -dynamics values close to experimental data. Both methods show an upper shift of more 1 pK unit due to cytidine methylation, and little or no effect (≤ 0.5 pK unit) when the cytidine is flanked by LNA residues. Moreover, a down shift in pK is observed under sugar modification and in presence of protonated cytidine. The predicted pK values for cytidine in DNA triplex environment can inform the design of better-binding oligonucleotides.

Introduction

DNA triple helices play a key role in cellular processes, such as regulation of replication and transcription, and recombination.¹⁻⁵ Triplex-forming oligonucleotides (TFOs) are DNA major groove ligands which target specific DNA sequences by forming a DNA triplex⁶⁻⁷ in anti-parallel orientation or parallel orientation.⁸ Many biotechnological and biomedical applications make use of the TFOs ability to form triplex, such as isolation of specific DNA sequences by triplex affinity capture,⁹⁻¹¹ detection and capture of polymerase chain reaction products,¹² detection of DNA mutation,¹³ and site-directed mutagenesis.¹⁴⁻¹⁵

To be competitive, a TFO in such applications needs to bind with higher efficiency to the target DNA duplex than the prototypical all-DNA TFO. One approach is to use non-natural oligonucleotides, for instance with a modified sugar that restrict the conformation of the ribose moiety, as in locked nucleic acid (LNA).¹⁶⁻¹⁸ LNA is characterized by an oxymethylene group bridging the atom C2' and C4' of the sugar ring (Figure 1).

Figure 1. Chemical structures of nucleoside ribose: locked ribose in LNA and deoxyribose in DNA.

Triple-helix target sites in the human genome are abundant especially in promoter regions,¹⁹⁻²⁰ where the TFOs targeting sequences are G-rich. Thus, TFOs targeting G-rich sequences are of biomedical importance for regulating transcription as an anti-gene strategy.²⁰ To target a G-rich sequences, the TFO should be C-rich and bind the targeted duplex $C^+ \bullet G-C$ base triads ('-' refers to Watson-Crick and ' $\bullet$ ' to Hoogsteen base pair),²¹ as shown in Figure 2. The formation of such a parallel triplex is affected at physiological pH by the protonation state of cytosine (pK 4.4).²²⁻²⁶ In this article, we will use the more general term pK instead of pK_a to discuss the protonation equilibrium. Cytidine protonation provides a stable Hoogsteen base pair in the pyrimidine strand (Figure 2). This would limit the therapeutic application of TFOs

as anti-gene strategy to regulate transcription, specifically when targeting G-rich sequences.²⁷⁻
²⁸ Thus, this is important to understand if the substitution of deoxyribose by locked ribose in the TFO shifts the pK of the cytidine toward high values. This will contribute to optimize the design of TFO targeting G-rich sequence.

Figure 2. Base triad configuration in parallel triplex. (A) Optimal base triad $C^+ \bullet G-C$ configuration: the duplex is bound by Watson-Crick hydrogen bonds (black); the third strand is bound by Hoogsteen hydrogen bond (red). (B) When C in the third strand is not protonated, it may form non-canonical interactions, of which one example is shown (yellow).

The protonation state of a nucleic acid can be determined by experimental techniques such as UV absorption and CD spectroscopies.^{23, 29} NMR spectroscopy can also be used to distinguish protonation states by simply monitoring the change in chemical shift,³⁰ though more recently, relaxation dispersion was also utilized to probe short-lived and/or low abundance states.³¹⁻³²

Specifically for cytidine in triplex environment, not much experimental data is available. An NMR study found that the pK value of cytidine in an intramolecular triplex is 5.3–7.4 depending on the position and sequence context³⁰ and no data is available on environment of triplexes containing LNA.

In this study, we have calculated pK value of cytidine in triplex context using two computational methods. The methods must sufficiently take into account physiochemical factors that affect the protonation equilibrium such as solvent, ion, and neighboring residues. The first computational approach, λ -dynamics with an explicit description of the solvent,³³ offers microscopic description of the biomolecule and the solvent and accounts for conformational rearrangement in response to the changing protonation state. In the second computational approach, multiple pH regime method,³⁴ conformational ensemble is generated by MD simulation for each protonation state. The snapshots are then used to calculate the

degree of protonation of cytidine of interest as a function of pH by means of continuum electrostatic model and then averaged over the snapshots. λ -dynamics is more computationally expensive of the two approaches, with simulation needed for every titration point, while multiple pH regime only needs two sets of simulations. λ -dynamics also requires more preparation by calculating the free energy between the two states and calibrating biasing potential used as titration barrier to bias sampling at physical end states.

Both computational approaches rely on the correct force field parameters for the deprotonated and protonated states. Charges are particularly important for p*K* calculations, especially for multiple pH regime, since it relies on the point charges to describe electrostatics accurately. To evaluate the effect of LNA sugar on p*K* of cytidine, we re-parameterized LNA force field parameters based the general CHARMM strategy.³⁵⁻³⁷ The procedure is based quantum mechanics calculation of model compounds as reference and the point charges are parameterized to accurately reproduce interaction with water and the dipole moment. The current CHARMM force field for LNA (Pande-Nilsson parameters)³⁸ was initiated from analogs. While it gave reasonable results for different duplexes, as mentioned in another report of free energy calculation,³⁹ several non-optimized parameter terms cause bias on backbone torsions, which may accumulate when the system contains multiple LNAs. The re-parameterization therefore is also needed to reproduce various types of LNA-containing helical structures.

In this study, we present the re-parameterization of LNA CHARMM force field and validations in single, duplex, and triplex systems. Then, we use the updated parameters to calculate the p*K* of cytidine in the DNA/LNA triplex. In particular we focus on parallel triplex with pyrimidine motif where the pyrimidine bases of TFO hydrogen-bond to the duplex purine strand by forming Hoogsteen pairs (Figure 2). To calculate the p*K* values we use to two computational approaches. We also discuss the p*K* results against the available experimental data and the influence of LNA on the p*K* of cytidine.

Methods

Locked nucleic acid parameterization

The LNA force field parameters were developed in line with CHARMM force field philosophy.³⁷ The quantum mechanics (QM) optimization and energy calculation of model

compounds were performed in MP2/6-31G* level. The same geometries and properties were calculated in molecular mechanics (MM) using the Pande-Nilsson parameters³⁸ as the initial guess. Two model compounds were used: model I, which is a locked ribose without C5' and base, and model II which is model I with an imidazole attached on C1' position (Figure S1).

The model compound I was used for charge optimization. Each solvent accessible atom of model I was probed by a water molecule. The QM interaction energies and dipole moment were calculated at HF/6-31G* level. The water complexes were then reproduced in MM by adjusting the charge until the results were consistent with QM (Figure S2). The targeting QM energies were scaled by 1.16 and the dipole by 1.20 to account for the polarizability in the additive force field.^{36, 40} The charges of all hydrogen atoms were fixed to be 0.09 e⁻.

The optimized charges of model I were directly transferred to LNA sugar by taking off atoms H1'' and H4' and compensating charge (+0.09 e⁻) on C1' and C4' respectively, and the base and C5' parts kept the same as original charges. Atom types were chosen in line with nucleic acids force field (i.e. CN7B, CN7 and CN8B atom type was used for C2', C4' and C6', respectively) to avoid introducing extra parameter terms.

Model II was used for geometry optimization. Most bonded parameters were transferred from RNA force field and only the terms containing both O2' and C6' need to be parameterized. The equilibrium points of bond and angle were adjusted targeting on MP2/6-31G* geometry. The force constants were optimized through scanning and fitting QM potential energy surface (PES). Among those terms, the PES of covalent bonds and angles were scanned over 3 points, i.e. the equilibrium point and the points of ± 0.01 Å and $\pm 1^\circ$; the PES of ring torsions were scanned from -18° passing equilibrium point to $+18^\circ$ with an increment of 3° (Figure S3). The dihedrals (three on sugar ring and three for the oxymethylene bridge) were simultaneously parameterized to account for the correlations between each other.

Table 1. Description of experimental helical structures containing DNA (D), RNA (R) and LNA (L) residues used in force field validation.

Structure type	PDB ID	Length	Sequence type ^(a)	Method	Publication Year	Residues ^(b)
duplex	2X2Q	7	L:L	X-ray (1.9 Å)	2010 ⁴¹	A, G, T, m ⁵ C
	1H0Q	9	L:R	NMR	2004 ⁴²	A, G, T, m ⁵ C
	1HHX	9	D/L ^{2,5,7} :R	NMR	2002 ⁴³	T
	1I5W	10	D/L ⁶ :D/L ⁶	X-ray (1.4 Å)	2001 ⁴⁴	T
triplex	1W86	8	D:D:D/L ^{2,4,6}	NMR	2004 ⁴⁵	m ⁵ C

^(a) Upper case indicate the position of the LNA residues in the strand

^(b) Abbreviations are for adenine (A), guanine (G), thymine (T), 5-methylated cytidine (m⁵C)

Molecular systems

The set of structures for the validation of force field include a single strand of a tetra oligonucleotide, double helices (PDB ID: 2X2Q,⁴¹ 1HHX,⁴³ 1H0Q,⁴² and 1I5W⁴⁴) and a triple helix (PDB ID: 1W86⁴⁵) (Table 1).

The starting structures of a tetra nucleotide CAAU were built in *anti* and *syn* conformations following strategy used in the previous AMBER parameterization paper.⁴⁶ In *anti* conformation, the glycosidic torsion χ between base and sugar moiety (O4'-C1'-N1-C2 for pyrimidine or O4'-C1'-N9-C4 for purine) is $170^\circ < \chi < 330^\circ$, where in *syn* conformation $30^\circ < \chi < 90^\circ$. For NMR structures (PDB ID: 1HHX, 1H0Q, and 1W86) the first conformer was chosen as starting structure for the simulations.

The triplex structures for the pK calculations were built on T-A•T DNA fiber model using 3DNA web server⁴⁷ and the bases were modified to be the target sequence. DNA nucleotides in TFO strand were then substituted to LNA according to the sequence by patching an oxymethylene group on ribose. All cytidines with LNA sugar were 5-methylated (m⁵C). The LNA sugars were energy-minimized in 100 steps using the steepest descent method while constraining other atoms, to convert the sugar pucker. 10-mer DNA triplex have been used for multiple pH regime method while only the central 5-mers have been used for λ -dynamics, with the 7-mer duplex extended by one residue on either end (Table 2).

Table 2. Sequence of third strand (TFO) in triplexes used for p*K* calculation.

Multiple pH regime		λ -dynamics	
Name	TFO sequence	Name	TFO sequence
C-DDD-10	TTTTT C TTTT	C-DDD-5	TT C TT
C-LDL-10	<u>TTTTT</u> C <u>TTTT</u>	C-LDL-5	TT C TT
M-DDD-10	TTTTT M TTTT	M-DDD-5	TT M TT
M-LDL-10	<u>TTTTT</u> M <u>TTTT</u>	M-LDL-5	TT M TT
C-DLD-10	<u>TTTTT</u> C <u>TTTT</u>	–	–
M-DLD-10	<u>TTTTT</u> M <u>TTTT</u>	–	–
–	–	C ⁺ C-DDD-5	TC ⁺ C TT

Residue for which p*K* is calculated is **bolded**; M is 5-methylcytosine; ⁺ indicates fixed protonation; underline denotes residues with locked sugar (LNA). For some systems, the counterpart is not available in the other method; this absence is denoted by “–”.

All structures were solvated in a cubic box of water. The shortest distance between box edge and solute is at least 8 Å and the periodic boundary conditions were applied. The systems were neutralized by adding sodium ions and a salt concentration of 0.1–0.5 M NaCl was used.

Molecular dynamic simulations

For molecular dynamics simulations, the structures were energy-minimized with steepest descent and Adopted-Basis Newton-Raphson method, with a harmonic force 10 kcal/mol/Å² restrained on backbone atoms and the distances of base pair hydrogen bonds in *vacuum*. All structures were then solvated in a cubic box of TIP3P water model.⁴⁸ The particle mesh Ewald (PME) method⁴⁹ was applied for long range electrostatic interactions, with a direct space cutoff of 9 Å, and the switch function over the range 8-9 Å was used on the force of van der Waals interactions.

Molecular dynamics simulations were performed on graphical processing units with the program CHARMM⁵⁰ and the OpenMM interface.⁵¹ The CHARMM36 force field of nucleic acids^{35, 52} and modified nucleotides³⁷ were used. The simulations were performed using Langevin dynamics with a friction coefficient of 5 ps⁻¹ in the NPT ensemble. The leap-frog integrator and 2 fs time step were used. Bonds involving hydrogen atom were constrained using the SHAKE algorithm.⁵³ Before the production run all systems were equilibrated in 10 ns with harmonic restraint on hydrogen bond distances of end base pairs and backbone atoms at 298 K. The production runs are 2 μs for tetra nucleotides; 200 ns for experimental structures in parameterization validation; and 100 ns for triplexes in the p*K* calculation using multiple pH regime method. The harmonic force 10 kcal/mol/Å² with 2.94 Å equilibrium distance between heavy atoms was applied on the hydrogen bonds of end Watson-Crick base

pairs in all duplexes and triplexes. Additional Hoogsteen distance restraints distance restraints were included to preserve the triplex structures in pK calculation.

λ -dynamics simulation

Constant pH molecular dynamics, in the framework of multisite λ -dynamics (CPHMD^{MSLD}),³³ were set up for cytidine such that the deprotonated and protonated two protonation states C and C⁺ are described and propagated by continuous variables λ_C and λ_{C+} respectively at a given pH. The potential energy function at desired pH is given by:

$$\begin{aligned} U_{\text{tot}}(X, \{x\}, \{\lambda\}, \text{pH}) \\ = U_{\text{env}}(X) + \lambda_C(U(X, x_C) - \Delta G_{C \rightarrow C+}(\text{model})) + \ln(10) k_B T (\text{p}K_{\text{mod}} - \text{pH}) \\ + \lambda_{C+}(U(X, x_{C+})) + F^{\text{bias}}(\lambda_C) + F^{\text{bias}}(\lambda_{C+}) \end{aligned}$$

Where X is the coordinates of environment atoms, x_C and x_{C+} are coordinates of atoms corresponding to the deprotonated and protonated states, and:

$$F^{\text{bias}}(\lambda_{Ci}) = \begin{cases} k_{\text{bias}}(\lambda_{Ci} - 0.8)^2; & \text{if } \lambda_{Ci} < 0.8 \\ 0; & \text{otherwise} \end{cases}$$

Where Ci refers to deprotonated and protonated states λ_{Ci} scales the potential energy of the corresponding protonation state with the constraints:

$$0 \leq \lambda_{Ci} \leq 1 \text{ and } \lambda_C + \lambda_{C+} = 1$$

$\Delta G_{C \rightarrow C+}(\text{model})$ is the free energy of protonation for the cytidine model compound (deoxyribocytosine 5'-monophosphate) in aqueous solution (see below for free energy calculation details). This term is included to flatten the potential energy surface such that the two protonation states of the nucleoside in solution are equi-populated as the free energy between the two states becomes zero. This is the state when $\text{pH} = \text{p}K_{\text{mod}}$ at which population ratio protonated:deprotonated is 1:1. The term $\ln(10) k_B T (\text{p}K_{\text{mod}} - \text{pH})$ adds offset when $\text{pH} \neq \text{p}K_{\text{mod}}$, where $\text{p}K_{\text{mod}}$ is experimentally measured pK of the model compound. Cytidine was used as a model compound with $\text{p}K_{\text{mod}}$ 4.44.⁵⁴ Two harmonic biasing potentials $F^{\text{bias}}(\lambda_C)$ and $F^{\text{bias}}(\lambda_{C+})$ are included to bias sampling towards the physical end states. In this formulation, $0.8 \leq \lambda_{Ci} \leq 1$ is considered as a physical end state. k_{bias} is the force constant of the harmonic potentials. The force constant is equal for both protonation states.

λ -dynamics is performed within the CHARMM BLOCK module⁵⁵ using the Multi-Site λ -Dynamics framework (MSLD)⁵⁶. The functional form $\lambda_{Ci} = \frac{e^{5.5 \sin \theta_{Ci}}}{\sum_i e^{5.5 \sin \theta_{Ci}}}$, was used. Defining λ as a function of θ in this way has been shown to be optimal for sampling and convergence.⁵⁷ θ is assigned a fictitious mass of 12 amu·Å² (amu = atomic mass unit) and its temperature was

maintained at 298K by coupling to Langevin heat bath with a collision frequency of 10 ps⁻¹. λ is saved every 10 steps. Bond, angle, and dihedral energy terms are excluded from scaling by λ so that only geometrically relevant states are sampled. An optimized sampling bias (k_{bias}) was applied for each state. The CHARMM REPD module⁵⁵ is used to run eight titration points in parallel and attempt to exchange energies every 1000 steps (pH replica exchange); this is repeated to a total of five independent runs. Titration points for most of triplexes are pH = 5,6,7,...,12, except for triplex C⁺C-DDD-5 for which points pH=2,3,4 ...9 were used. For model compound points pH=1,2,3...8 were used. A non-bonded list cut-off of 15 Å was used with electrostatics force switch and van der Waals switch functions between 10 and 12 Å. 20-ns simulations were performed in the NVT ensemble with Langevin dynamics with collision frequency of 10 ps⁻¹, with the last 18 ns used for analysis.

To calculate pK, the fraction of deprotonated states, S^{deprot} , is calculated at each titration point and fitted to Henderson-Hasselbalch equation: $S^{\text{deprot}} = \frac{1}{1+10^{-n(pH-pK)}}$, where n is Hill coefficient and indicates degree of cooperativity between titratable groups.

The value of k_{bias} was calibrated by performing 1 ns runs of λ -dynamics at various k_{bias} values and observing the fraction of physical end states in the trajectory (FPL, fraction physical ligand) and the frequency of transitions between the two protonation states (transition rate, ns⁻¹). Initial k_{bias} values were 22.50, 25.00, 27.50, 30.00, 32.50, 35.00, 37.50 kcal/mol. Additional runs were added as needed to determine the optimal value.

For methyl-cytidine, we used as model compound 5-methyl deoxyribocytosine 5'-monophosphate and an estimated pK value of 4.60 (no experimental value is available).

Calculation of free energy

The free energy of protonation between the two states, $\Delta G_{C \rightarrow C^+}(\text{model})$, was calculated with Bennett acceptance ratio (BAR) method.⁵⁸ The hybrid model for deoxyribocytosine 5'-monophosphate used in the free energy perturbation is constructed with a deoxyribose sugar and 5' phosphate and 3' hydroxyl and two pyrimidine bases corresponding to the two protonation states (dual topology); the two bases are maintained within the same volume of space by distance restraints between all pairs of common atoms in the two protonation states. The CHARMM BLOCK module⁵⁵ is used to partition the system into three blocks: (I) environment, (II) neutral C base, and (III) protonated C base. Interactions between blocks II and III are set to null; interactions between blocks I and II and within block II are scaled with λ ; interactions between blocks I and III and within block III are scaled with 1- λ . Scaling with

λ and $1-\lambda$ is not applied to bond, angle and dihedral energy terms. Eight λ values corresponding to alchemical intermediate/end states were used ($\lambda = 0.0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 1.0$) and in each window the simulation length is 3 ns, with the last 2.9 ns used for data collections. The CHARMM REPD module⁵⁵ is used to run eight alchemical intermediate/end states in parallel and attempt to exchange energies every 1000 steps (Hamiltonian replica exchange). The calculated free energy of protonation between the two states is 52.0 ± 0.1 kcal/mol for cytidine and 45.1 ± 0.1 kcal/mol for 5-methylated cytidine.

Multiple pH regime method

Multiple pH regime method has been described in details in Ref³⁴, here we present a brief summary. The degree of deprotonation of the ionizable site was calculated with multiple hydrogen locator approach,⁵⁹ where the population of microscopic protonation state α of a titratable group i is described by:

$$p_{i\alpha}(\text{pH}) = \frac{\sum_{\{\mathbf{x}\}} \delta(x_i, \alpha) \exp(-\frac{\Delta E(\text{pH}; \mathbf{x})}{RT})}{\sum_{\{\mathbf{x}\}} \exp(-\frac{\Delta E(\text{pH}; \mathbf{x})}{RT})}$$

Where $\{\mathbf{x}\}$ is the sets of all protonation states of the triplex nucleic acids. $\mathbf{x} = (x_1, x_2, \dots, x_i, \dots, x_N)$ contains N elements corresponding to the number of titratable groups. x_i is the microscopic states of group i . $\delta(x_i, \alpha)$ is Kronecker delta where $\delta(x_i, \alpha) = 1$ if $x_i = \alpha$ and $\delta(x_i, \alpha) = 0$ if $x_i \neq \alpha$. $\mathbf{x} = (0, 0, \dots, 0)$ is the reference state of the system. $\Delta E(\mathbf{x})$ is the change of electrostatic energy from the reference state to state $\mathbf{x}$:

$$\Delta E(\mathbf{x}) = 2.3RT \sum_i^N (\text{p}K_{ix_i, \text{int}} - v_{ix_i, \text{pH}}) + \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}}^N E_{ix_i, jx_j}$$

Where i and j are the titratable sites. v_{ix_i} is the difference of the number of protons bound to site i in reference state and state x_i . E_{ix_i, jx_j} is the interaction energy between sites i and j in their respective protonation states x_i and x_j . $\text{p}K_{ix_i, \text{int}}$ is intrinsic $\text{p}K$ at site i :⁶⁰

$$\text{p}K_{ix_i, \text{int}} = \text{p}K_{ix_i, \text{mod}} + \Delta \text{p}K_{ix_i, \text{Born}} + \Delta \text{p}K_{ix_i, \text{pc}} + \Delta \text{p}K_{i, \text{ref}}$$

Where $\text{p}K_{ix_i, \text{mod}}$ is the standard $\text{p}K$ value of microscopic state x_i at site i . $\Delta \text{p}K_{ix_i, \text{Born}}$ is the contribution from desolvation; $\Delta \text{p}K_{ix_i, \text{pc}}$, from interactions with permanent charges; $\Delta \text{p}K_{i, \text{ref}}$, from charge-charge interactions between site i and other titratable groups in their reference states.

The electrostatic terms $pK_{ix_i, int}$ and E_{ix_i, jx_j} were calculated by solving linearized Poisson-Boltzmann equation using finite difference method⁶¹⁻⁶² as used elsewhere.^{59, 63} The following parameters were used: relative dielectric constant of the solvent 80, relative dielectric constant of the solute 4, ion exclusion radius 2 Å. Cytosine was used as a model compound with pK_{mod} 4.4.⁶⁴

The current calculations differ from the previous study³⁴ in the treatment of the ionic strength, since the current system is nucleic acids instead of proteins. Here the charge and the position of the mobile ions were explicitly included in the calculations of the electrostatic interactions, with the assumption that the mobile ions positions on average obey the Boltzmann distribution.

The final population of each protonated form is obtained by the arithmetic averaging over an conformational ensemble. Here conformational ensembles were collected from 100 ns MD simulation every 100 ps. For each of the triplexes (Table 2) two ensembles were generated: one with a N3-protonated cytidine in the TFO and one with a N3-deprotonated cytidine in the TFO. The pK of cytidine was calculated from the average of the titration curves calculated from the ensembles for protonated and deprotonated cytidine. The pK of an individual site is defined as the value of pH at which the corresponding site is half protonated.

Structural analysis

Snapshots, saved every 50 ps, were analyzed using CHARMM and Curves+⁶⁵ software packages. The conformations were characterized using the backbone torsions from α to ζ , base orientation, sugar pucker and base pair step parameters .

The base orientation is defined by the glycosidic torsion (χ): O4'-C1'-N1-C2 (pyrimidine) or O4'-C1'-N9-C4 (purine), and its main conformations are denoted as *anti* ($170^\circ < \chi < 330^\circ$, where $\chi < 220^\circ$ is low *anti* and $> 270^\circ$ is high *anti*) and *syn* ($30^\circ < \chi < 90^\circ$). The sugar pucker is defined by the pseudorotation phase angle (P), which is a combination of five ring torsions, and it is denoted as *north* ($-90^\circ < P \leq 90^\circ$) and *south* ($90^\circ < P \leq 270^\circ$).

To check the preservation of base pairs in the duplex and triplex, the N1-N3 distances of WC base pairs and N7-N3 of HG base pairs were monitored. A distance shorter than 3.5 Å indicates that a hydrogen bond is formed between the heavy atoms and the bases are considered to be paired.

If not specified, the backbone torsions were calculated by excluding the end base pair in each strand and the base pair geometries were calculated by excluding the last two nucleotides in each end.

Results and discussion

Validation of LNA parameters

The parameterization was performed in two parts: the charge optimization took account only the ribose and the geometries were optimized by considering the effects from base and backbone atom C5'. The new charges (Table S2) reproduced better HF/6-31G* water-ribose interaction energy and the dipole moment than Pande-Nilsson charges (Table S1). The bond and angle parameters well reproduced the optimized QM conformation (Table S3), except the dihedral O4'-C1'-C2'-H2'' and N1-C1'-C2'-H2'' (Table S3). The deviations however are tolerable since the orientation of aliphatic H2'' does not influence much the ribose and nucleotides properties. The new set of parameters (see Supporting Information for the parameter file) keeps most parameter terms consistent with canonical CHARMM36 nucleic acid force field,^{35, 52} including the backbone and glycosidic torsion terms.

The new set of LNA parameters was validated on structural properties of LNA single strand⁴⁶ and mixed LNA/DNA or LNA/RNA duplex⁴¹⁻⁴⁴ and triplex⁴⁵ helical structures, for which NMR or X-ray structural information are available.

LNA CAAU single strand

Two 2- μ s simulations were performed in NaCl solution (0.15 M) for the tetramer starting from *syn* and *anti* base orientations. The two simulations sample similar torsional space: the nucleotides are mainly in low ($\chi < 220^\circ$) and high ($> 270^\circ$) *anti* conformation; the portion of *syn* conformation is negligible (Figure S4). The tail nucleotide U4 was consistently in low *anti*, a conformation typical for A-type helix. The other nucleotides however had mixed low and high *anti*. The middle purines A2 and A3 behaved 30%–70% low *anti* while the pyrimidine C1 had >80% low *anti*. Since the 3'-end nucleotide U4 was less interactive with other nucleotides, its base orientation reflects as a mono nucleotide more than others. The first three nucleotides had more complicated base stacking and base-backbone interactions in the single strand, so the conformations kept converting to response the interactions. Particularly the smaller charge repulsion between ribose and purine makes A2 and A3 easier to keep high *anti* than C1.⁶⁶

The distances between hydrogen atoms involved in a nuclear Overhauser effect (NOE) signal were calculated to compare the sampled conformation with the NMR data.⁴⁶ An NOE signal is detectable when the distance between two sites is shorter than 5.5 Å. All the intra-nucleotide distances were within 5.5 Å (Figure S5), which implies the nucleotide conformation is in agreement with the solution NMR data. On the other hand, the inter-nucleotide distances had a wider distribution going from 3 to 12 Å, which implies the oligonucleotide strand is dynamic and samples a wide range of conformations. However, all the calculated inter nucleotide distances sampled distance shorter than 5.5 Å, in agreement with the presence of an NOE signal.

Overall, the parameters reproduce A-type conformation for nucleotides, which is consistent with NMR data, and enable conformational dynamics for single strand in reasonable agreement with NMR observations.

Duplex and triplex helices containing LNA

The simulations of duplexes and triplexes were performed in NaCl solution (0.15 M) using the new parameter sets for LNA and were compared with the experimental structure and simulations performed using Pande-Nilsson parameters. The first 20 ns of each run was excluded from the data analysis. The duplexes having both LNA strands (2X2Q), one LNA strand (1H0Q) and several LNA substitutions in a DNA or RNA strand (1HHX and 1I5W) were used (Table 1). After equilibration, all the helical structures were stable, and kept the experimental hydrogen bond pattern (Figure S6), even if some sporadic base pair openings occurred. By visual inspection, the simulated structures showed a very good agreement with the experimental one (Figure 3), except for 1I5W, where some difference in the helical structure between experiment and simulation were observed (see discussion below).

To analyze the helical conformation in more details, we calculated the base-pair step for each base pair in the double helix. In general, the simulated double helices sample similar helical conformations with the new and Pande-Nilsson parameter sets and agree with the experimental observation, as shown by base pair step parameters values in Figure S7. No differences between the force field were observed for the structures with partial LNA substitutions (1HHX, 1I5W and 1H0Q), whereas the structure of full LNA (2X2Q) showed some conformational difference between the two force fields: the new parameter set gave a better agreement with crystal structure 2X2Q, than Pande-Nilsson set, especially for the twist values, only exception are x-displacement and slide. In case of the crystal structure 1I5W (with single LNA in each strand), the x-displacement values could not be reproduced by either

force fields: experimental data show typical A-type DNA values for x-displacement (around -4 Å), but the simulation values were more shifted toward B-type conformation (-1.8 Å). We cannot rule out that the difference might be due to the special crystallization conditions. The 7-mer helix 2X2Q existed as dimer and is coordinated by multivalent ions⁴¹ in the crystal, while it is a monomer and is in monovalent ion (0.15 M NaCl) solutions in the simulations.

The triple helix 1W86 (Table 1) is formed by a DNA duplex and a DNA/LNA third strand. The simulated structure agreed with the NMR structure (Figure 3). The base pair geometries for the DNA duplex part were better reproduced by new parameter set than by Pande-Nilsson set, especially for x-displacement and slide (Figure S7). The helical conformation is between A and B type, a typical conformation previously observed for DNA triplex⁶⁷. Both parameter sets however equally failed to reproduce the propeller value. The largely negative propeller in NMR structure corresponds to the spiral-like hydrogen bonds in the alternative C and T context, where the atom O4 or N4 in TFO pyrimidine forms an additional bifurcated hydrogen bond with the atom N4 or O4 in duplex 5'-flanking pyrimidine.⁴⁵ Such hydrogen bond pattern was not observed in the simulations; as a consequence the rise and twist were slightly shifted from NMR structure.

While the new parameters better reproduced the base pair geometries for triplex and full LNA duplex than the Pande-Nilsson set but equally good for partly LNA substituted duplexes, the backbone torsions showed a large improvement with the new set. Figure 4 shows that Pande-Nilsson parameters failed to reproduce the 5'-torsions α , β and γ . This did not directly affect base pair geometry (Figure S7) nor base pair hydrogen bonding,³⁸ but it caused widely spread twist value for the steps with consecutive LNA (2X2Q and 1H0Q in Figure S7) and problematic 5'OH orientation for nucleoside.

As the simulations of single strand, duplex, and triplex show, the new parameters with better agreement with QM properties reproduced the A-form conformation (north and low *anti*) for LNA nucleotides and keep consistent structural features with experiments for LNA mixed DNA and RNA helices. The new charges reproduced better interactions with water and the dipole moment as shown by the comparison with *ab initio* data. The improvement of the LNA force field is especially necessary to properly describe the influence of LNA residues on the adjacent positional contexts in triplex structure. In the next section, LNA force field will be used to understand if the effect of LNA on the structure will also affect the pK values of close-by cytosine base.

Figure 3. The simulated structures (black and magenta) superimposed on experimental structures (green and yellow) together with the corresponding PDB ID. Duplex in black and green and TFO in magenta and yellow. The averaged structure from 200 ns simulation is reported for the simulation.

Figure 4. The backbone conformation, glycosidic torsion and sugar pucker for LNA residues. Simulation data using Pande-Nilsson in red and using new parameters in black, experimental data from PDB survey in blue.

Calculating cytidine pK in LNA context

The experimental values of pK of cytidine are crucial, since it is an input in our calculations. We tabulate these values with special attention to the presence of ribose and/or phosphate moieties since they slightly shift pK value (Table 3). No experimental value is available for deoxyribonucleoside 5'-monophosphate of 5-methylcytidine. Thus, we have used a value of 4.6 for λ -dynamics, speculating that the sugar-phosphate addition will contribute to shift a little the pK value from that of the base, like it is observed for cytidine.

Table 3. Experimental pK values for cytidine (C) and 5-methylcytidine (M) in literature. The values used in our calculations are **bolded**.

Context	C	M
Nucleobase	4.4 ⁶⁴ 4.45 ⁶⁸ 4.45 ⁶⁹ 4.5-4.6 ⁵⁴	4.6 ⁶⁹
Ribonucleoside	4.11 ⁶⁹ 4.08 ⁵⁴ 4.17 ⁷⁰	4.28 ⁶⁹
Deoxyribonucleoside	4.25 ⁶⁹	
Ribonucleoside 5'-monophosphate	4.24 ⁵⁴ 4.54 ⁷⁰	
Deoxyribonucleoside 5'-monophosphate	4.44 ⁵⁴	
Triplex, third strand CTCCTT	6.8 ³⁰	
Triplex, third strand CTCTCT	7.4 ³⁰	

λ -dynamics

We performed λ -dynamics simulations using a calibrated value for free energy of deprotonation and for the k_{bias} . The free energy of deprotonation was calculated for the isolated model compounds, 5'-monophosphate cytidine and 5-methylated cytidine, in 0.1 M NaCl solution. The calculated free energy of protonation between the two states is 52.0 ± 0.1 kcal/mol for 5'-monophosphate cytidine and 45.1 ± 0.1 kcal/mol for 5'-monophosphate 5-methylated cytidine.

The value of k_{bias} was calibrated by performing runs of λ -dynamics at various k_{bias} values in 0.1 NaCl solution and observing the fraction of physical end states in the trajectory (FPL, fraction physical ligand) and the frequency of transitions between the two protonation states (transition rate, ns^{-1}). An optimal value of k_{bias} was chosen so as to yield a high transition rate and simultaneously maintain high FPL (above 0.8).³³ The value of the optimized k_{bias} is 29.50 kcal/mol for cytidine and 29.25 kcal/mol for 5-methylated cytidine (k_{bias} calibration for cytidine is shown in Figure 5A). With this setting, transition rate between the protonation states is fairly high and sampling is sufficiently biased at end states (Figure 5B).

To check the λ -dynamics setting, first we calculated the pK value for the model compound 5'-monophosphate cytidine in 0.1 M NaCl solution using 5 independent pH-replica-exchange runs of 4-ns each going from pH 1 to pH 8. The extracted values from last 2 ns of 4-ns simulations of the model C nucleotide yielded back the pK value of 4.45 ± 0.02 , very close to the experimental pK of 4.44 (input), see Figure 6. The average values of $S_{\text{deprotonated}}$ were fitted to Henderson-Hasselbalch equation and pK values was determined by reading the value of pH off the curve at when $S_{\text{deprotonated}}=0.5$.

Figure 5. (A) Optimization of k_{bias} value. A k_{bias} value of 29.5 kcal/mol (red) is chosen to maintain the fraction of physical states (FPL, black) above 0.8, while maintaining a moderately high transition rate (blue). Error bar is standard error of the mean of five independent runs. (B) Transition of λ_C values in a 1-ns run of model system nucleoside C. Only $0.8 \leq \lambda_C \leq 1$ is counted as physical state of C, whereas $0 \leq \lambda_C \leq 0.2$ would be counted as physical state of C^+ . Transition rates: 37 ns^{-1} ; fractions of physical states: 82%.

Figure 6. Titration result λ -dynamics, fitted to Henderson-Hasselbalch equation. Error bar for each titration point is the standard error of the mean from 5 independent runs.

The value of protonation free energy and k_{bias} were then used for the larger systems, namely 5-mer triplexes (Table 2). For each 5-mer triplex 5 independent pH-replica-exchange runs of 20 ns each were used for pK calculation (the first 2 ns are discarded from data analysis). Figure 6 reported the titration curves obtained using Henderson-Hasselbalch equation and Table 4 the corresponding pK values. The standard error of mean of the five independent runs of the pK values is 0.1-0.2 pK unit for the cytidine and 0.3 pK unit for methyl-cytidine.

The pK values of cytidine in 5-mer triplex are all shifted toward values higher than 4.5 (pK calculated for free cytidine), when the cytidine is flanked by a thymine. The pK for cytidine in C-DDD-5 (triplex, third strand TTCTT) is 7.2 (Table 4), a value is remarkably close to the experimental value of 7.4 of a similar intramolecular triplex (third strand CTCTCTT, Table 3) determined by NMR spectroscopy³⁰. The methylation in 5-position shifts the pK values of 2 pK unit to a value around 9, experimentally an increase of apparent pK was also observed, but of less than 1 pK unit³⁰. We again note that the λ -dynamics calibration was not done on experimental value for methyl cytidine, so comparison should be done in more qualitative manner. Mutating the sugar from deoxyribose to locked ribose does not affect the pK value of neighboring cytidine.

When the cytidine is flanked by a protonated cytidine, like in C⁺C-DDD-5, a pK of 4.7 is observed (close to pK calculated for free cytidine). The pK value of C⁺C-DDD-5 (triplex, third strand TC⁺CTT) is markedly different from the other triplexes since the protonation state of the neighbor C⁺ is fixed. The charge-charge repulsion makes the titrating C less readily protonated, shifting the real pK value downwards. As the experimental pK value for cytidine in a similar triplex (third strand CTCCTTT)³⁰ is 6.8, this is a fair prediction.

Multiple pH regime

Two sets of 100-ns simulations corresponding to each protonation state of cytidine were performed for each 10-mer triplex (Table 2). One challenge in this approach was that TFO containing deprotonated cytidine tended not to stay in Hoogsteen configuration since one hydrogen bond is lost (Figure 7, insets). As such, we have taken care that the triplex structure is preserved, even when third strand cytidine is deprotonated and less tightly bound. The average structures of the two ensembles of C-DDD-10 are shown as an example (Figure 7). The deprotonated average structure shows slight distortion on the TFO backbone and the deprotonated C is hydrogen-bonded with only one bond to the purine strand of the duplex and occasionally flips out. Otherwise the triplex keeps its triple helical structure. We have also attempted a system containing two consecutive Cs, as a counterpart to C⁺C-DDD-5, but the deprotonation of two consecutive Cs destabilized triplex structure (data not shown).

Figure 7. The average structures from 100-ns simulations of triplex C-DDD-10 (A) when C is deprotonated and (B) protonated; with insets showing the triplet hydrogen bond configuration; side and bottom (third strand 3'-end) views. Duplex is shown in green; TFO, orange; C, red.

The difference between the pK values for deprotonated and protonated ensembles arises from the electrostatic environment of cytidine. In the case of the protonated ensemble the electrostatic environment is essentially more favorable. Hoogsteen hydrogen bonding in the protonated ensemble stabilizes the third strand cytidine, whereas the lack of it in the deprotonated ensemble makes cytidine less tightly bound. Average of titration curves of the two ensembles was taken (Figure 8).

Figure 8. Titration curve for C in C-DDD-10. Note that the pK/mid does not coincide with the average pK . pK/mid is the pH point where the average titration curve (blue) has a value protonation = 0.5. In some cases the average curve shows a wide plateau at the level of half protonation.

The calculated pK values of cytidine in 10-mer triplex are all shifted toward value higher than the free cytosine. In the triplex, the largest pK value is observed for 5-methyl deoxycytidine, followed by deoxycytidine, the lowest values for LNA cytidine. Interesting the methylation on LNA residues has less effect than on DNA residues. The presence of methylation in M-DDD-10 only causes slight increase in pK by 0.3-0.7 pK unit compared to C-DDD-10. Neighboring LNAs increases pK by another 0.5 unit, similar to unmethylated case. Again, when its own sugar is locked, pK actually shifts downwards.

Locking the sugar ring results in a slight upward pK shift by 0.5 pK unit for the neighboring cytidine. This is consistent with rigidifying effect of LNA on the triplex structure, presumably making the nucleotide less likely to flip out, exposed to the solvent and deprotonated. However, when its own sugar is locked in C-DLD-10, its pK shift is less than that of C-DDD-10.

Comparison between the two approaches

We use two approaches to calculate the pK values for cytidine and methyl cytidine embedded in a DNA triplex, in presence or in absence of LNA residues. We note that the lengths used for the triplex also differ: 5-mer for λ -dynamics and 10-mer for multiple pH regime. Moreover, not all systems have counterparts in both approaches (Table 4). Multiple pH regime does not have the counterpart of C⁺C-DDD-5 where there are consecutive Cs, since we could not maintain a stable triplex in the ensemble where both Cs are deprotonated. λ -dynamics does not have the counterparts of C-DLD-10 and M-DLD-10 where the titratable residue is itself an LNA, because the model system in λ -dynamics is a nucleotide, and not the nucleobase as in multiple pH regime. Up to now no experimental data are available for pK of LNA nucleotide.

Table 4. pK values calculated by multiple pH regime and λ -dynamics.

Multiple pH regime				λ -dynamics	
	$pK_{\text{dep}}^{(a)}$	$pK_{\text{pro}}^{(b)}$	$pK_{\text{mid}}^{(c)}$		pK
C-DDD-10	5.27	12.55	9.27	C-DDD-5	7.2
C-LDL-10	5.66	12.80	9.77	C-LDL-5	7.1
M-DDD-10	5.40	13.67	9.5 to 10	M-DDD-5	9.5 ^(d)
M-LDL-10	5.95	13.63	10 to 10.5	M-LDL-5	9.0 ^(d)
C-DLD-10	1.10	11.98	8 to 8.5	-	-
M-DLD-10	4.98	11.99	8.64	-	-
-	-	-	-	C ⁺ C-DDD-5	4.7 ^(e)
C-DDD-10 (0.5M NaCl)			8.0 to 8.5	C-DDD-5 (0.5M NaCl)	8.0 ^(f)

^(a) pK_{dep} is the pK for the deprotonated ensembles; ^(b) pK_{pro} is the pK for the protonated ensembles; ^(c) pK_{mid} is the midpoint of the averaged titration curve. ^(d) The model pK value is not based on experimental value, since it is not available. ^(e) Simulation time is 18 ns. ^(f) Simulation time is 10 ns.

Both methods agree in showing an increase of the pK value for the cytidine residues when they are embedded in a triple helical structure and an effect of the methylation on pK :

$$pK(\text{free cytidine}) < pK(\text{cytidine in triplex}) < pK(\text{methyl cytidine in triplex})$$

Locking the sugar ring results in a pK shift of maximum 0.5 pH for the neighboring cytidine, but in different directions. In C-LDL-10, there is slight upward pK shift by 0.5 pK unit, which is consistent with rigidifying effect of LNA on the triplex structure, presumably making the nucleotide less likely to flip out, exposed to the solvent and deprotonated. The neighboring LNAs in C-LDL-5 do not affect pK shift of cytidine, while in M-LDL-5 downward shift the

pK (Table 4). The smaller size of the system used λ -dynamics might be a cause to the lack of *pK* shift since the small system might already be quite rigid in the first place (due to restraint end) and LNA substitution offers only marginal increase of rigidity. All in all, LNA does not cause large *pK* shift as the shifts are less than 0.5 *pK* unit.

The *pK* value of C in C-DDD-5 (triplex, third strand TTCTT) determined by λ -dynamics (*pK* 7.51 ± 0.11 , Table 4) is remarkably close to the experimental value of 7.4 of a similar intramolecular triplex (third strand CTCTCTT, Table 3) determined by NMR spectroscopy.³⁰ Multiple pH regime method predicted greater *pK* shift as the *pK* value of C in C-DDD-10 is 9.27 (Table 4). This suggests that the values calculated from multiple pH regime are to be considered in qualitative and not quantitative manner.

Simulations were also performed at higher ionic concentration (0.5 M NaCl), to evaluate how the methods could account for change in the ionic environment. In the higher ionic concentration, the *pK* of cytidine in C-DDD-5 appears to have no significantly different, considering a rather large uncertainty at titration point at pH 8 (Figure 6). Multiple pH regime, shows marked downward *pK* shift in increasing the ionic strength from 9.27 to 8.50. The difference between the two approaches could be due to different way the methods are settled. Since λ -dynamics uses a model system with 0.1 M NaCl, we have checked if the calibration on such a model system is valid to describe the effect of higher ionic strength. We have calculated *pK* value for monophosphate cytidine (model compound) at 0.5 M NaCl, increasing the water box dimension to accommodate more ions, and it yields shifted *pK* value of 3.98 ± 0.02 . Two approaches account differently for the contribution of ions to the *pK*: multiple pH regime accounts for the contribution of all the ions, including those outside the cutoff distance (12 Å) of λ -dynamics. Finally we can not exclude that the length of system may also have an effect. A 10-mer triplex is more representatives for comparison with experiment experimental since it know that triple forming oligonucleotides shorter than 11mer hardly form an intermolecular triplex.

Conclusion

We have used λ -dynamics and multiple pH regime to evaluate the *pK* values of cytidine-type residues embedded in a triple helix. In particular, we aim to understand if the substitution of deoxyribose by locked ribose (LNA) in the triple forming nucleotide shifts the *pK* of the

cytidine toward high values promoting the formation of the Hoogsteen base pair between the third strand and the target DNA duplex.

To achieve that we have re-parameterized the LNA force field in line with the CHARMM philosophy to properly account for the interaction between the locked ribose and the other DNA residues and the water. The optimized LNA force field reproduced the A-form structure for nucleotides correctly. It achieved a good agreement with the available experiment structures, judging from base pair geometries, helix conformation and backbone torsions.

The two computational approaches predicted that cytidine in triplex environment has a high pK shift to above physiological pH, making it more likely to be protonated under this condition, although they predicted significantly different pK values. The predictions for triplex containing 5-methyl-cytidine, on the other hand, agree in both approaches.

Both approaches also suggest that 5' and 3' neighboring thymine LNAs, as well as increase in ionic strength, have little to no effect in the pK of cytidine in the third strand of triplex nucleic acids. When its own sugar is locked, the titrating cytidine/5-methyl-cytidine even undergoes downward pK shift. And when the cytidine is flanked by a protonated cytidine, a down shift to the pK is observed reaching a value close to free cytidine.

λ -dynamics seems to offer a more accurate pK prediction in comparison to experimental data, but much more computationally expensive compared to multiple pH regime. As a rough comparison, 20-ns λ -dynamics simulation of a 5-mer system at simultaneous 8 pH points with pH replica exchange (total 160 ns) takes about 5-6 times longer CPU hours compared to two 100-ns MD simulations (total 200 ns) for a 10-mer system for multiple pH regime; with similar post-processing analysis times.

Supplementary material

Supplementary Table S1-S3 and Figures S1-S7 and force field parameters

Acknowledgement

The authors thank Nanyang Technological University Research Scholarship (to Y. D. H.), and the Swedish Research Council (VT 2015-04992).

References

- (1) Veselkov, A. G.; Malkov, V. A.; Frank-Kamenetskii, M. D.; Dobrynin, V. N., Triplex model of chromosome ends. *Nature* **1993**, *364*, 496-496.
- (2) Baran, N.; Lapidot, A.; Manor, H., Formation of DNA Triplexes Accounts for Arrests of DNA Synthesis at d(TC)_n and d(GA)_n Tracts. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 507-511.
- (3) Daube, S. S.; Hippel, P. v., Functional transcription elongation complexes from synthetic RNA-DNA bubble duplexes. *Science* **1992**, *258*, 1320-1324.
- (4) Dayn, A.; Samadashwily, G. M.; Mirkin, S. M., Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 11406-11410.
- (5) Møllegaard, N. E.; Buchardt, O.; Egholm, M.; Nielsen, P. E., Peptide nucleic acid-DNA strand displacement loops as artificial transcription promoters. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 3892-3895.
- (6) Moser, H. E.; Dervan, P. B., Sequence-specific cleavage of double helical DNA by triple helix formation. *Science* **1987**, *238* (4827), 645-650.
- (7) Thuong, N. T.; Hélène, C., Sequence - specific recognition and modification of double - helical DNA by oligonucleotides. *Angew. Chem. Int. Ed.* **1993**, *32* (5), 666-690.
- (8) Frank-Kamenetskii, M. D.; Mirkin, S. M., Triplex DNA structures. *Annu. Rev. Biochem.* **1995**, *64*, 65-95.
- (9) Ito, T.; Smith, C. L.; Cantor, C. R., Sequence-specific DNA purification by triplex affinity capture. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (2), 495-498.
- (10) Ito, T.; Smith, C. L.; Cantor, C. R., Triplex affinity capture of a single copy clone from a yeast genomic library. *Nucleic Acids Res.* **1992**, *20* (13), 3524-3524.
- (11) Ito, T.; Smith, C. L.; Cantor, C. R., Affinity capture electrophoresis for sequence-specific DNA purification. *Genet. Anal.: Biomol. Eng.* **1992**, *9* (3), 96-99.
- (12) Vary, C., Triple-helical capture assay for quantification of polymerase chain reaction products. *Clin. Chem.* **1992**, *38* (5), 687-694.
- (13) Olivas, W. M.; Maher, L., Analysis of duplex DNA by triple helix formation: Application to detection of a p53 microdeletion. *Biotechniques* **1994**, *16*, 128-128.
- (14) Havre, P. A.; Glazer, P. M., Targeted mutagenesis of simian virus 40 DNA mediated by a triple helix-forming oligonucleotide. *J. Virol.* **1993**, *67* (12), 7324-7331.
- (15) Havre, P. A.; Gunther, E. J.; Gasparro, F. P.; Glazer, P. M., Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (16), 7879-7883.
- (16) Torigoe, H.; Hari, Y.; Sekiguchi, M.; Obika, S.; Imanishi, T., 2'-O,4'-C-methylene bridged nucleic acid modification promotes pyrimidine motif triplex DNA formation at physiological pH: thermodynamic and kinetic studies. *The Journal of biological chemistry* **2001**, *276* (4), 2354-2360.
- (17) Koshkin, A. A.; Singh, S. K.; Nielsen, P.; Rajwanshi, V. K.; Kumar, R.; Meldgaard, M.; Olsen, C. E.; Wengel, J., LNA (Locked Nucleic Acids): Synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition. *Tetrahedron* **1998**, *54* (14), 3607-3630.
- (18) Obika, S.; Nanbu, D.; Hari, Y.; Morio, K.-i.; In, Y.; Ishida, T.; Imanishi, T., Synthesis of 2' -O, 4' -C-methylenuridine and-cytidine. Novel bicyclic nucleosides having a fixed C 3,-endo sugar pucker. *Tetrahedron Lett.* **1997**, *38* (50), 8735-8738.
- (19) Goñi, J. R.; De La Cruz, X.; Orozco, M., Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.* **2004**, *32* (1), 354-360.
- (20) Goñi, J. R.; Vaquerizas, J. M.; Dopazo, J.; Orozco, M., Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **2006**, *7* (1), 63.
- (21) Soyfer, V. N.; Potaman, V. N., General Features of Triplex Structures. In *Triple-Helical Nucleic Acids*, Springer New York, 1996; pp 100-150.
- (22) Faucon, B.; Mergny, J. L.; Helene, C., Effect of third strand composition on the triple helix formation: purine versus pyrimidine oligodeoxynucleotides. *Nucleic Acids Res.* **1996**, *24* (16), 3181-8.
- (23) Volker, J.; Klump, H. H., Electrostatic effects in DNA triple helices. *Biochemistry* **1994**, *33* (45), 13502-8.
- (24) Lee, J. S.; Woodsworth, M. L.; Latimer, L. J.; Morgan, A. R., Poly(pyrimidine) . poly(purine) synthetic DNAs containing 5-methylcytosine form stable triplexes at neutral pH. *Nucleic Acids Res.* **1984**, *12* (16), 6603-14.
- (25) Vekhoff, P.; Ceccaldi, A.; Polverari, D.; Pylouster, J.; Pisano, C.; Arimondo, P. B., Triplex formation on DNA targets: how to choose the oligonucleotide. *Biochemistry* **2008**, *47* (47), 12277-12289.

- (26) Sugimoto, N.; Wu, P.; Hara, H.; Kawamoto, Y., pH and Cation Effects on the Properties of Parallel Pyrimidine Motif DNA Triplexes. *Biochemistry* **2001**, *40* (31), 9396-9405.
- (27) Helene, C.; Thuong, N. T.; Harel, A., Control of Gene Expression by Triple Helix-Forming Oligonucleotides. The Antigene Strategy. *Ann. N.Y. Acad. Sci.* **1992**, *660*, 27-36.
- (28) Paugh, S. W.; Coss, D. R.; Bao, J.; Lauder milk, L. T.; Grace, C. R.; Ferreira, A. M.; Waddell, M. B.; Ridout, G.; Naeve, D.; Leuze, M.; LoCascio, P. F.; Panetta, J. C.; Wilkinson, M. R.; Pui, C.-H.; Naeve, C. W.; Uberbacher, E. C.; Bonten, E. J.; Evans, W. E., MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. *PLoS Comput. Biol.* **2016**, *12*, e1004744.
- (29) Thaplyal, P.; Bevilacqua, P. C., Experimental approaches for measuring pKa's in RNA and DNA. *Methods Enzymol.* **2013**, *549*, 189-219.
- (30) Leitner, D.; Schröder, W.; Weisz, K., Influence of sequence-dependent cytosine protonation and methylation on DNA triplex stability. *Biochemistry* **2000**, *39*, 5886-5892.
- (31) Kimsey, I. J.; Petzold, K.; Sathyamoorthy, B.; Stein, Z. W.; Al-Hashimi, H. M., Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature* **2015**, *519* (7543), 315-320.
- (32) Szymanski, E. S.; Kimsey, I. J.; Al-Hashimi, H. M., Direct NMR Evidence that Transient Tautomeric and Anionic States in dG· dT Form Watson-Crick-like Base Pairs. *J. Am. Chem. Soc.* **2017**, *139* (12), 4326-4329.
- (33) Goh, G. B.; Knight, J. L.; Brooks, C. L., Constant pH molecular dynamics simulations of nucleic acids in explicit solvent. *J. Chem. Theory Comput.* **2012**, *8*, 36-46.
- (34) Nilsson, L.; Karshikoff, A., Multiple pH Regime Molecular Dynamics Simulation for p K Calculations. *PLoS one* **2011**, *6* (5), e20116.
- (35) Foloppe, N.; MacKerell, A. D., All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem* **2000**, *21* (2), 86-104.
- (36) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31* (4), 671-90.
- (37) Xu, Y.; Vanommeslaeghe, K.; Aleksandrov, A.; MacKerell, A. D., Jr.; Nilsson, L., Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.* **2016**, *37* (10), 896-912.
- (38) Pande, V.; Nilsson, L., Insights into structure, dynamics and hydration of locked nucleic acid (LNA) strand-based duplexes from molecular dynamics simulations. *Nucleic Acids Res.* **2008**, *36* (5), 1508-16.
- (39) Xu, Y.; Villa, A.; Nilsson, L., The free energy of locking a ring: Changing a deoxyribonucleoside to a locked nucleic acid. *J. Comput. Chem.* **2017**, *38* (15), 1147-1157.
- (40) Mackerell, A. D.; Karplus, M., Importance of Attractive Vanderwaals Contribution in Empirical Energy Function Models for the Heat of Vaporization of Polar Liquids. *J. Phys. Chem.* **1991**, *95* (26), 10559-10560.
- (41) Eichert, A.; Behling, K.; Betzel, C.; Erdmann, V. A.; Furste, J. P.; Forster, C., The crystal structure of an 'All Locked' nucleic acid duplex. *Nucleic Acids Res.* **2010**, *38* (19), 6729-6736.
- (42) Nielsen, K. E.; Rasmussen, J.; Kumar, R.; Wengel, J.; Jacobsen, J. P.; Petersen, M., NMR studies of fully modified locked nucleic acid (LNA) hybrids: Solution structure of an LNA : RNA hybrid and characterization of an LNA : DNA hybrid. *Bioconjugate Chem.* **2004**, *15* (3), 449-457.
- (43) Petersen, M.; Bondensgaard, K.; Wengel, J.; Jacobsen, J. P., Locked nucleic acid (LNA) recognition of RNA: NMR solution structures of LNA : RNA hybrids. *J. Am. Chem. Soc.* **2002**, *124* (21), 5974-5982.
- (44) Egli, M.; Minasov, G.; Teplova, M.; Kumar, R.; Wengel, J., X-ray crystal structure of a locked nucleic acid (LNA) duplex composed of a palindromic 10-mer DNA strand containing one LNA thymine monomer. *Chem. Commun.* **2001**, (7), 651-652.
- (45) Sorensen, J. J.; Nielsen, J. T.; Petersen, M., Solution structure of a dsDNA : LNA triplex. *Nucleic Acids Res.* **2004**, *32* (20), 6078-6085.
- (46) Condon, D. E.; Yildirim, I.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Turner, D. H., Optimization of an AMBER Force Field for the Artificial Nucleic Acid, LNA, and Benchmarking with NMR of L(CAAU). *J. Phys. Chem. B* **2014**, *118* (5), 1216-1228.
- (47) Zheng, G. H.; Lu, X. J.; Olson, W. K., Web 3DNA-a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **2009**, *37*, W240-W246.
- (48) Jørgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, 926.

- (49) Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089-10092.
- (50) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545-614.
- (51) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S., Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem* **2009**, *30* (6), 864-72.
- (52) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D., Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *J Chem Theory Comput* **2012**, *8* (1), 348-362.
- (53) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327-341.
- (54) Izatt, R. M.; Christensen, J. J.; Rytting, J. H., Sites and thermodynamic quantities associated with proton and metal ion interaction with ribonucleic acid, deoxyribonucleic acid, and their constituent bases, nucleosides, and and nucleotides. *Chemical Reviews* **1971**, *71* (5), 439-481.
- (55) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Others, CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.
- (56) Knight, J.; Brooks, C., Λ -Dynamics Free Energy Simulation Methods. *J. Comput. Chem.* **2009**, *30*, 1692-1700.
- (57) Knight, J. L.; Brooks, C. L., Applying efficient implicit nongeometric constraints in alchemical free energy simulations. *J. Comput. Chem.* **2011**, *32*, 3423-3432.
- (58) Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245-268.
- (59) Koumanov, A.; Rüterjans, H.; Karshikoff, A., Continuum electrostatic analysis of irregular ionization and proton allocation in proteins. *Proteins: Structure, Function, and Bioinformatics* **2002**, *46* (1), 85-96.
- (60) Tanford, C.; Kirkwood, J. G., Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.* **1957**, *79* (20), 5333-5339.
- (61) Warwicker, J.; Watson, H., Calculation of the electric potential in the active site cleft due to α -helix dipoles. *J. Mol. Biol.* **1982**, *157* (4), 671-679.
- (62) Nicholls, A.; Honig, B., A rapid finite difference algorithm, utilizing successive over - relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.* **1991**, *12* (4), 435-445.
- (63) Koumanov, A.; Benach, J.; Atrian, S.; González - Duarte, R.; Karshikoff, A.; Ladenstein, R., The catalytic mechanism of Drosophila alcohol dehydrogenase: evidence for a proton relay modulated by the coupled ionization of the active site lysine/tyrosine pair and a NAD⁺ ribose OH switch. *Proteins: Structure, Function, and Bioinformatics* **2003**, *51* (2), 289-298.
- (64) Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B., Calculation of pK a's in RNA: On the structural origins and functional roles of protonated nucleotides. *Journal of molecular biology* **2007**, *366* (5), 1475-1496.
- (65) 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.
- (66) Foloppe, N.; Nilsson, L., Toward a full characterization of nucleic acid components in aqueous solution: Simulations of nucleosides. *J. Phys. Chem. B* **2005**, *109* (18), 9119-9131.
- (67) Esguerra, M.; Nilsson, L.; Villa, a., Triple helical DNA in a duplex context and base pair opening. *Nucleic Acids Res.* **2014**, *42*, 11329-11338.
- (68) Close, D. M., Calculated p K a's of the DNA Base Radical Ions. *The Journal of Physical Chemistry A* **2013**, *117* (2), 473-480.
- (69) Fox, J. J.; Van Praag, D.; Wempen, I.; Doerr, I. L.; Cheong, L.; Knoll, J. E.; Eidinoff, M. L.; Bendich, A.; Brown, G. B., Thiation of Nucleosides. II. Synthesis of 5-Methyl-2'-deoxycytidine and Related Pyrimidine Nucleosides. *J. Am. Chem. Soc.* **1959**, *81* (1), 178-187.
- (70) Egli, M.; Saenger, W., *Principles of nucleic acid structure*. Springer Science & Business Media: 2013.

Supporting Information

Modeling pK shift in DNA triplex containing locked nucleic acids

Yossa Dwi Hartono*, You Xu*, Andrey Karshikoff , Lennart Nilsson and Alessandra Villa

Yossa Dwi Hartono^{1,2}*, You Xu¹*, Andrey Karshikoff¹ , Lennart Nilsson¹ and Alessandra Villa¹

¹ Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden.

² Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

* These authors contributed equally to this work

Corresponding author: Alessandra Villa, alessandra.villa@ki.se

Keywords: LNA, DNA, pK, triplex, lambda dynamics

Tables

Table S1. Statistics of calculated water interaction and dipole moment for locked ribose (model I). QM calculation is at HF/6-31G* level.

Parameters set	RMS/Max of MM – QM		Dipole moment	
	Energy (kcal/mol)	Distance (Å)	QM/MM (D)	Angle of MM – QM (degree)
Pande-Nilsson new	0.60/0.85	0.04/0.07	1.65/1.82	13.55
	0.33/0.55	0.03/0.09	1.65/2.04	6.73

Table S2. Optimized new and Pande-Nilsson atomic charges (e^-) for locked ribose (model I).

Atom	new	Pande-Nilsson	Atom	new	Pande-Nilsson
C1'	0.080	0.070	H3''	0.090	0.090
H1'	0.090	0.090	C4'	0.190	0.190
H1''	0.090	0.090	H4'	0.090	0.090
C2'	0.140	0.070	O4'	-0.500	-0.480
H2''	0.090	0.090	C6'	0.050	0.050
O2'	-0.500	-0.440	H6'	0.090	0.090
C3'	-0.180	-0.180	H6''	0.090	0.090
H3'	0.090	0.090			

Table S3. The calculated geometric terms of locked ribose (model II). Only the terms in ribose part are listed. The bond lengths are in Å; angles and torsions are in degree.

Terms	QM	MM	diff	Terms	QM	MM	diff
C1'-H1'	1.096	1.114	0.018	C4'-C5'-H5'1	110.28	109.98	-0.31
C1'-C2'	1.534	1.522	-0.012	C4'-C5'-H5'2	110.28	110.36	0.08
C1'-O4'	1.427	1.435	0.008	C4'-C5'-H5'3	109.97	110.05	0.08
C2'-H2''	1.091	1.106	0.015	H5'1-C5'-H5'2	108.86	108.97	0.12
C2'-O2'	1.440	1.44	0	H5'1-C5'-H5'3	108.49	108.6	0.11
C2'-C3'	1.522	1.496	-0.025	H5'2-C5'-H5'3	108.9	108.84	-0.07
O2'-C6'	1.452	1.445	-0.007	O4'-C1'-C2'-H2''	-167.78	-162.76	5.02
C3'-H3'	1.093	1.107	0.015	O4'-C1'-C2'-O2'	70.97	69.41	-1.56
C3'-H3''	1.093	1.109	0.016	O4'-C1'-C2'-C3'	-36.96	-37.7	-0.75
C3'-C4'	1.527	1.513	-0.014	N1-C1'-C2'-H2''	-49.72	-40.15	9.58
C4'-O4'	1.464	1.451	-0.014	N1-C1'-C2'-O2'	-170.97	-167.98	2.99
C4'-C6'	1.527	1.518	-0.008	N1-C1'-C2'-C3'	81.1	84.91	3.81
C4'-C5'	1.504	1.509	0.005	C2'-C1'-O4'-C4'	1.09	1.72	0.64
C6'-H6'	1.097	1.109	0.012	N1-C1'-O4'-C4'	-120.03	-120.31	-0.28
C6'-H6''	1.095	1.11	0.016	C1'-C2'-O2'-C6'	-70.68	-70.68	0.01
C5'-H5'1	1.093	1.111	0.018	H2''-C2'-O2'-C6'	164.09	159.13	-4.95
C5'-H5'2	1.094	1.11	0.017	C3'-C2'-O2'-C6'	35.18	33.05	-2.13
C5'-H5'3	1.093	1.111	0.019	C1'-C2'-C3'-C4'	54.56	55.99	1.42
H1'-C1'-C2'	111.46	109.56	-1.9	H2''-C2'-C3'-C4'	-177.26	-178.39	-1.13
H1'-C1'-O4'	110.33	107.37	-2.97	O2'-C2'-C3'-C4'	-55.03	-54.46	0.58
C2'-C1'-O4'	102.9	103.77	0.87	C2'-O2'-C6'-C4'	1.18	3.16	1.97
C1'-C2'-H2''	115.63	116.84	1.21	C2'-O2'-C6'-H6'	-119.28	-116.07	3.21

C1'-C2'-O2'	105.93	107.52	1.59	C2'-O2'-C6'-H6"	120.01	123.04	3.03
C1'-C2'-C3'	100.94	98.64	-2.3	C2'-C3'-C4'-O4'	-54.74	-56.49	-1.75
H2"-C2'-O2'	109.33	112.7	3.37	C2'-C3'-C4'-C6'	53.75	54.11	0.36
H2"-C2'-C3'	119.78	116.03	-3.75	C2'-C3'-C4'-C5'	-175.13	-177.54	-2.41
O2'-C2'-C3'	103.8	103.43	-0.37	C3'-C4'-O4'-C1'	35.15	34.66	-0.5
C2'-O2'-C6'	105.18	105.01	-0.17	C6'-C4'-O4'-C1'	-69.72	-68.93	0.79
C2'-C3'-H3'	114.43	115.05	0.62	C5'-C4'-O4'-C1'	163.61	161.01	-2.6
C2'-C3'-H3"	112.38	113.27	0.89	C3'-C4'-C6'-O2'	-36.54	-37.29	-0.75
C2'-C3'-C4'	91.38	93.18	1.8	C3'-C4'-C6'-H6'	81.55	79.52	-2.03
H3'-C3'-H3"	110.85	112.01	1.16	C3'-C4'-C6'-H6"	-154.35	-154.96	-0.61
H3'-C3'-C4'	112.95	111.01	-1.94	O4'-C4'-C6'-O2'	69.5	67.52	-1.98
H3"-C3'-C4'	113.66	110.9	-2.76	O4'-C4'-C6'-H6'	-172.41	-175.68	-3.26
C3'-C4'-O4'	102.27	100.95	-1.32	O4'-C4'-C6'-H6"	-48.31	-50.16	-1.84
C3'-C4'-C6'	100.68	99.22	-1.46	C5'-C4'-C6'-O2'	-169.56	-166.28	3.28
C3'-C4'-C5'	120.59	118.48	-2.12	C5'-C4'-C6'-H6'	-51.47	-49.47	2
O4'-C4'-C6'	105.37	108.09	2.72	C5'-C4'-C6'-H6"	72.63	76.05	3.42
O4'-C4'-C5'	108.49	110.76	2.27	C3'-C4'-C5'-H5'1	177.08	179.14	2.06
C6'-C4'-C5'	117.52	117.51	0	C3'-C4'-C5'-H5'2	-62.66	-60.6	2.05
C1'-O4'-C4'	106.27	105.48	-0.79	C3'-C4'-C5'-H5'3	57.47	59.53	2.07
O2'-C6'-C4'	103.26	103.94	0.68	O4'-C4'-C5'-H5'1	59.81	63.24	3.43
O2'-C6'-H6'	109.45	108.42	-1.03	O4'-C4'-C5'-H5'2	-179.93	-176.5	3.42
O2'-C6'-H6"	109.86	109.03	-0.83	O4'-C4'-C5'-H5'3	-59.8	-56.37	3.44
C4'-C6'-H6'	112.89	111.93	-0.96	C6'-C4'-C5'-H5'1	-59.49	-61.65	-2.15
C4'-C6'-H6"	111.31	112.25	0.94	C6'-C4'-C5'-H5'2	60.77	58.61	-2.16
H6'-C6'-H6"	109.86	110.94	1.08	C6'-C4'-C5'-H5'3	-179.11	178.75	2.14

Figures

Figure S1. The structures of locked ribose and deoxyribose, and the model compounds used in parameterization. Model I is used for charge optimization and model II for bonded parameters.

Figure S2. The water complex of locked ribose with possible probe positions corresponding for charge optimization. Only one water molecule was included in each calculation.

Figure S3. The potential energy surface scan of six ring torsions concerning the introduced atoms of locked ribose. The scan for each torsion was performed $\pm 30^{\circ}$ around the equilibrium position with a 3° increment. The atom numbering is in Figure S.

Figure S4. The glycosidic torsions of each member in tetra oligonucleotide CAAU as the function of simulation time. The left panel shows the anti starting structure and right one shows syn as the starting.

Figure S5. The distance measurement of tetra-nucleotide CAAU in 2 μ s simulation comparing with NOE data from previous report (). The black histograms are sampled from the simulation with CAAU-*anti* as starting structure and the red histograms are which starting from CAAU-*syn*, which were taken from ref(1). The measured distances are labeled in each panel.

Figure S6. The resistance of base pairs of simulated experimental structures as the function of simulation time. The Watson-Crick (WC) base pairs are shown for duplexes 2X2Q, 1HHX and 1i5W and WC and Hoogsteen (HG) base pairs are shown for triplex 1W86. Closed base pairs are shown as filled stripes while the opened base pairs are shown as blank. To check the preservation of base pairs in the duplex and triplex, the N1-N3 distances of WC base pairs and N7-N3 of HG base pairs were monitored. A distance shorter than 3.5 Å indicates that a hydrogen bond is formed between the heavy atoms and the bases are considered to be paired.

Figure S7. The base pair conformations reproduced using Pande-Nilsson (red) and new (black) parameters, comparing with PDB survey data. In both simulation and survey data, all the DNA and LNA nucleotides are considered. The survey data are shown in blue vertical lines for crystal or histograms for NMR structures.

Force field parameters

```
* Additive CHARMM Force field of locked nucleic acids
* Updated by You Xu, Karolinska Institutet, 2016
*

! Should be used with parent files  NA36: top_all36_na.inp and
par_all36_na.inp
!

read rtf card append
* Topologies of modified nucleotides
*
36 1

PRES LNA      -0.34      ! patch to convert 2'-D-ribose to 2'-O,4'-C-
methylene
GROUP          ! re-optimized, 2016, yxu
ATOM C4'   CN7      0.28
ATOM O4'   ON6B     -0.50
ATOM C2'   CN7B      0.15
ATOM O2'   ON6B     -0.50
ATOM C6'   CN8B      0.05
ATOM H6'   HN8       0.09
ATOM H6''  HN8       0.09
!
DELETE ATOM H2'
DELETE ATOM H4'
!
BOND O2' C6'
BOND C6' H6'
BOND C6' H6''
BOND C6' C4'
IC  H2'' C2' O2' C6'  1.09  114.23  158.19  105.57  1.43
IC  C2' O2' C6' H6'  1.43  105.57 -109.79  109.77  1.09
IC  C2' O2' C6' H6'' 1.43  105.57  127.83  113.51  1.09
IC  C2' O2' C6' C4'  1.43  105.57    1.4   40.0   1.54
! LNA PRES END

END

read param card flex append
* Parameters of modified nucleotides
*

BONDS
ON6B  CN8B      320.000    1.435  ! LNA, yxu, 12/2016

ANGLES
ON6B  CN7B  CN7    115.00    107.50 ! LNA, yxu, 12/2016
CN7B  ON6B  CN8B    84.00    111.00 ! LNA, yxu, 12/2016
ON6B  CN8B  CN7    130.00    107.00 ! LNA, yxu, 12/2016
CN8B  CN7   CN8B    51.00    112.00 ! LNA, yxu, 12/2016
ON6B  CN8B  HN8    70.00    107.30 ! LNA, from OG3C51 CG3C52 HGA2,
yxu, 12/2016
```


DIHEDRALS

CN7B	CN7	ON2	P2	0.6	5	0.0	!	LNA na36, from CN7, yxu
CN7B	CN7	ON2	P2	0.2	4	0.0	!	LNA na36, from CN7, yxu
CN7B	CN7	ON2	P2	0.0	3	180.0	!	LNA na36, from CN7, yxu
CN7B	CN7	ON2	P2	0.4	2	0.0	!	LNA na36, from CN7, yxu
CN7B	CN7	ON2	P2	1.9	1	180.0	!	LNA na36, from CN7, yxu
NG2R61	CN7B	CN7B	ON6B	3.000	3	0.0	!	LNA n2, yxu, 12/2016
NG2R51	CN7B	CN7B	ON6B	3.000	3	0.0	!	LNA n2, yxu, 12/2016
NN2B	CN7B	CN7B	ON6B	3.000	3	0.0	!	LNA n2, yxu, 12/2016
NN2	CN7B	CN7B	ON6B	3.000	3	0.0	!	LNA n2, yxu, 12/2016
ON6B	CN7B	CN7B	ON6B	2.000	1	180.0	!	LNA n2, yxu, 12/2016
ON6B	CN7B	CN7B	ON6B	0.900	6	180.0	!	LNA n2, yxu, 12/2016
ON2	CN7	CN7B	ON6B	0.000	3	0.0	!	LNA n3, yxu, 12/2016
ON5	CN7	CN7B	ON6B	0.000	3	0.0	!	LNA n3, yxu, 12/2016
HN7	CN7	CN7B	ON6B	0.500	3	180.0	!	LNA n3, yxu, 12/2016
CN7	CN7	CN7B	ON6B	0.000	2	0.0	!	LNA n3, yxu, 12/2016
CN7	CN7B	ON6B	CN8B	0.800	3	180.0	!	LNA o2, yxu, 12/2016
CN7B	CN7B	ON6B	CN8B	0.800	3	180.0	!	LNA o2, yxu, 12/2016
HN7	CN7B	ON6B	CN8B	0.300	3	180.0	!	LNA o2, yxu, 12/2016
CN7	CN8B	ON6B	CN7B	0.000	3	0.0	!	LNA co, yxu, 12/2016
HN8	CN8B	ON6B	CN7B	0.000	3	0.0	!	LNA co, yxu, 12/2016
ON6B	CN8B	CN7	ON6B	0.200	3	0.0	!	LNA c6, yxu, 12/2016
ON6B	CN8B	CN7	CN8B	0.200	3	0.0	!	LNA c6, yxu, 12/2016
ON6B	CN8B	CN7	CN7	0.200	3	0.0	!	LNA c6, yxu, 12/2016
ON5	CN8B	CN7	CN8B	0.000	3	180.0	!	LNA gam, yxu, 12/2016
ON2	CN8B	CN7	CN8B	0.000	3	180.0	!	LNA gam, yxu, 12/2016
HN8	CN8B	CN7	CN8B	0.000	3	180.0	!	LNA gam, yxu, 12/2016

END

References

1. Condon, D.E., Yildirim, I., Kennedy, S.D., Mort, B.C., Kierzek, R. and Turner, D.H. (2014) Optimization of an AMBER Force Field for the Artificial Nucleic Acid, LNA, and Benchmarking with NMR of L(CAAU). *J. Phys. Chem. B*, **118**, 1216-1228.

Role of Pseudoisocytidine Tautomerization in Triplex-Forming Oligonucleotides: In Silico and in Vitro Studies

Yossa Dwi Hartono,^{†,‡} Y. Vladimir Pabon-Martinez,[§] Arzu Uyar,^{†,‡} Jesper Wengel,^{||} Karin E. Lundin,[§] Rula Zain,^{§,⊥} C. I. Edvard Smith,[§] Lennart Nilsson,[†] and Alessandra Villa^{*,†,||}

[†]Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden

[‡]Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore

[§]Department of Laboratory Medicine, Clinical Research Center, Karolinska Institutet, SE-141 86 Huddinge, Sweden

^{||}Department of Physics, Chemistry and Pharmacy, Nucleic Acid Center, University of Southern Denmark, 5230 Odense M, Denmark

[⊥]Department of Clinical Genetics, Centre for Rare Diseases, Karolinska University Hospital, SE-171 76 Stockholm, Sweden

Supporting Information

ABSTRACT: Pseudoisocytidine (Ψ C) is a synthetic cytidine analogue that can target DNA duplex to form parallel triplex at neutral pH. Pseudoisocytidine has mainly two tautomers, of which only one is favorable for triplex formation. In this study, we investigated the effect of sequence on Ψ C tautomerization using λ -dynamics simulation, which takes into account transitions between states. We also performed in vitro binding experiments with sequences containing Ψ C and furthermore characterized the structure of the formed triplex using molecular dynamics simulation. We found that the neighboring methylated or protonated cytidine promotes the formation of the favorable tautomer, whereas the neighboring thymine or locked nucleic acid has a poor effect, and consecutive Ψ C has a negative influence. The deleterious effect of consecutive Ψ C in a triplex formation was confirmed using in vitro binding experiments. Our findings contribute to improving the design of Ψ C-containing triplex-forming oligonucleotides directed to target G-rich DNA sequences.

INTRODUCTION

The formation of DNA triple helices plays a key role in cellular processes¹ such as regulation of replication and transcription,^{2–5} chromosome folding,⁶ stabilization of telomeres, and recombination.⁷ Triplex-forming oligonucleotides (TFOs) have been used in many biotechnological and biomedical applications that make use of their ability to target the major groove of a DNA duplex. Examples are isolation of specific DNA sequences (triplex affinity capture),^{8–10} detection and capture of polymerase chain reaction products,¹¹ detection of DNA mutation,¹² and site-directed mutagenesis.^{13,14} Triplexes can form in different ways: with purine (antiparallel orientation) or pyrimidine (parallel orientation) motifs.¹⁵ In a parallel triplex, T•A–T and C⁺•G–C base triads are formed (“–” refers to a Watson–Crick base pair and “•” refers to a Hoogsteen base pair).¹⁶

Triple-helix target sites in the human genome are abundant, especially in promoter regions.^{17,18} TFOs targeting G-rich sequences are of biological importance because such regions are frequently present in promoters, which are potential targets for regulating transcription as an antigene strategy.¹⁸ However, the formation of parallel triplexes is not favorable at physiological

pH because it requires the protonation of cytosine (pK_a 4.1).¹⁹ This limits the therapeutic application of TFOs as antigene strategy to regulate transcription, specifically when targeting G-rich sequences.^{20,21}

Pseudoisocytidine (Ψ C) is an artificial pyrimidine analogue that is derived from pseudouridine.²² It has at least two relevant tautomers, Ψ C(H1) and Ψ C(H3), corresponding to the presence of a proton at N1 and N3, respectively (Figure 1). Tautomer Ψ C(H1) has the hydrogen bond donor/acceptor set for the Watson–Crick hydrogen bonding scheme to guanine, whereas tautomer Ψ C(H3) has the same set of hydrogen bond donors/acceptors as a protonated cytidine, which is favorable for Hoogsteen hydrogen bonding to guanine.²³ Tautomer Ψ C(H3) is, thus, desirable in a TFO as a substitute for C to target G-rich sequences forming a pyrimidine-motif triplex. The substitution of cytidine by analogues such as Ψ C is one strategy to target G-rich sequences at physiological pH.^{22,24,25}

Received: March 23, 2017

Accepted: May 2, 2017

Published: May 17, 2017

Figure 1. Cytidine and pseudoisocytidine. (A) Structures of protonated cytosine (C⁺), pseudoisocytidine tautomers, $\Psi\text{C(H1)}$ and $\Psi\text{C(H3)}$, with the corresponding atom numbers. R corresponds to the sugar position. (B) Base triad configurations, C⁺•G–C and $\Psi\text{C(H3)}$ •G–C, with Watson–Crick (black) and Hoogsteen (red) hydrogen bonds.

In this study, we aim to understand the ways of optimizing the design of intermolecular TFOs targeting G-rich sequences by using ΨC as a C analogue. To achieve this, we combined molecular dynamics (MD) simulations with in vitro binding experiments. In particular, we want to understand the effect of the environment (flanking nucleotides and bound/unbound states) on the tautomerization of ΨC . Experimental investigation of tautomerization is challenging because of the structural similarity, fast interconversion, and ambient aqueous condition.²⁶ Alternatively, molecular simulation methods, primarily λ -dynamics,²⁷ can be used to describe the change between tautomeric states. MD simulations have previously been successfully used to investigate DNA triple helices both in a parallel and an antiparallel fashion.^{28–30} These studies show that the DNA double-helix overslides in the negative direction to increase the major groove and to accommodate the third strand, and the resultant triple helical conformation is somewhere between A- and B-types, with base pairs remaining almost perpendicular to the helical axis.

Electrophoretic mobility shift assay (EMSA) was used to detect in vitro binding of TFOs containing ΨC under intranuclear conditions. The triplex intercalator, benzoquinoloxaline (BQQ), was used to probe for triplex formation.^{31,32} Pseudocytidines were incorporated both consecutively and nonconsecutively in the TFO sequence and combined for the first time with sugar-modified nucleotides, locked nucleic acids (LNAs). LNA has an oxy-methylene bridge locking the sugar pucker in C3'-endo rather than in C2'-endo as in DNA, which restricts the conformation of the sugar. The inclusion of LNAs in TFOs has been shown to increase triplex stability.³³ As a target sequence, we used an upstream G-rich region of the human trefoil factor (human TFF) gene close to an estrogen response element (ERE), including runs of consecutive guanines.³⁴ This sequence was previously studied

using bisLNAs, which contain a C-rich TFO part and were found to be a poor target at neutral pH.³⁵

First, we discuss the result from λ -dynamics simulations for short single-stranded and triplex DNA (trimers and 7-mers) containing ΨC in various positions and sequence contexts. Then, we compare the observed sequence effect with the result from in vitro binding experiments of six 17-mer TFOs containing ΨC . Finally, we characterize the structure of the observed triplexes using classical MD.

RESULTS AND DISCUSSION

Pyrimidine motif triplexes are unstable at physiological pH because of the need for protonation of C in the third strand. A way to solve this problem is to use C analogues such as ΨC . Because the aim of this study is to identify rules on how to incorporate ΨC in LNA containing oligonucleotides (ONs) for optimal hybridization under intracellular conditions, we started by performing simulation studies, specifically investigating how the surrounding bases might influence the tautomerization of ΨC .

Nucleoside Tautomerization: Reference State. The exact tautomeric ratio of ΨC in an aqueous solution is not known. In the crystal structure, the tautomeric ratio $\Psi\text{C(H1)}/\Psi\text{C(H3)}$ of the isocytosine base is exactly 1:1.³⁶ ¹H nuclear magnetic resonance (NMR) spectra did not show separate signals for the two ΨC tautomers in the aqueous solution, but enzymatic incorporation experiments confirmed the existence of the two tautomers in the solution for the deoxyribonucleoside ΨC .³⁷ The measured $\text{pK}_{\text{a}1}$ and $\text{pK}_{\text{a}2}$ at the two protonation sites corresponding to the two tautomeric states are highly similar [$\text{pK}_{\text{a}1}$ 3.79 and 3.69 and $\text{pK}_{\text{a}2}$ 9.36 and 9.42, for tautomers $\Psi\text{C(H1)}$ and $\Psi\text{C(H3)}$, respectively], and in the same study, ab initio calculation (Hartree–Fock) for the methylated ΨC base indicates that tautomer $\Psi\text{C(H3)}$ is favored over tautomer $\Psi\text{C(H1)}$ in a vacuum, and less so when the solvent effect is accounted for using the polarizable continuum model.³⁸

Taking these into account, we decided to set the tautomeric ratio of the model system, deoxyribonucleoside ΨC , to be 1:1 as the reference, where only the physical end states were considered. All variations observed in the tautomeric ratio should be interpreted as relative to the reference and not as absolute values. Henceforth, we will call this quantity tautomeric propensity to reflect on this point.

To set the tautomeric ratio to 1:1 in the λ -dynamics formulation, we supply a biasing potential exactly equal to the calculated free energy, $\Delta G_{\text{H1} \rightarrow \text{H3}}$, for the model system in water, 28.9 ± 0.1 kcal/mol. This results in an average population ratio of 1:1 for deoxyribonucleoside ΨC . To guarantee an optimal transition between the two tautomeric states, we calibrate k_{bias} value to yield a high fraction of physical end states in the trajectory and a high frequency of transitions between the two tautomeric states. By calibrating with a set of 1 ns λ -dynamics runs of the model system at various k_{bias} values, we found $k_{\text{bias}} = 19.5$ kcal/mol to result in >80% physical states and >60 transitions/ns, which we judged to be sufficient to ensure good sampling in simulations of length 1–10 ns (Figure 2). In practice, when this k_{bias} value is applied to other systems containing ΨC , the transition rate is 40–50 ns^{−1}, and more than 80% of the population is physical when there is only one ΨC . For two ΨC s, a fraction of 60–70% is physical, whereas for three ΨC s, 50–60% is physical. For an illustration of the

Figure 2. Optimization of k_{bias} value. A k_{bias} value of 19.5 kcal/mol (red) is chosen to maintain the fraction of physical states [fraction physical ligand (FPL), black] above 0.8, while maintaining a moderately high transition rate (blue). The error bar is the standard error of the mean of five independent runs.

fluctuation of λ values at this transition rate and the fraction of physical state, see Figure 3.

Five independent λ -dynamics simulations were performed for nucleoside ΨC (reference compound) and nucleotide ΨC . The addition of a 5' monophosphate group slightly shifts the tautomeric propensity to favor tautomer $\Psi\text{C}(\text{H1})$ [from 52 to 41% $\Psi\text{C}(\text{H3})$].

Effect of the Neighboring Bases on the Tautomerization of Pseudoisocytidine. We performed λ -dynamics simulations in single-stranded DNA trimers and 7-mers with varying sequences to investigate the effect of neighboring bases on the tautomerization of ΨC in a single strand. Besides DNA bases thymine (T) and cytosine (C) as neighboring residues, we also included protonated and/or 5-methylated C (meC) and restricted sugar moiety (LNA, denoted by underline).

Summarizing Figure 4, we observe that ΨC is 100% $\Psi\text{C}(\text{H3})$ in the triplex structures, and in the single-stranded ONs, we found the following $\Psi\text{C}(\text{H3})$ propensities (for the ΨC indicated in **bold**):

- >75%, $\text{C}^+\Psi\text{CC}^+$, $\text{meC}^+\Psi\text{CT}$, $\text{T}^+\Psi\text{CmeC}$, $\text{T}^+\Psi\text{CT}$, $\text{meC}^+\Psi\text{CmeC}$, $\text{meC}^+\Psi\text{CmeC}$
- 60–75%, $\text{C}^+\Psi\text{CC}$, $\text{meC}^+\Psi\text{CmeC}^+$, $\text{C}^+\Psi\text{CT}$, $\text{C}^+\Psi\text{CT}$, $\text{meC}^+\Psi\text{CT}$, $\text{T}^+\Psi\text{CC}^+$, $\text{T}^+\Psi\text{CC}$, ΨCTT , $\Psi\text{C}^+\Psi\text{CT}$, $\Psi\text{C}^+\Psi\text{CT}$, $\Psi\text{C}^+\Psi\text{C}^+\Psi\text{C}$, $\text{TT}^+\Psi\text{C}^+\Psi\text{CTT}$, $\text{T}^+\Psi\text{CTTT}^+\Psi\text{CT}$
- 40–59%, $\text{T}^+\Psi\text{CmeC}^+$, $\text{T}^+\Psi\text{CT}$, $\Psi\text{C}^+\Psi\text{CT}$, $\Psi\text{C}^+\Psi\text{CT}$, $\Psi\text{C}^+\Psi\text{C}^+\Psi\text{C}$, $\text{TT}^+\Psi\text{CT}^+\Psi\text{CTT}$, $\text{TT}^+\Psi\text{CTT}^+\Psi\text{CT}$, $\text{TT}^+\Psi\text{CmeC}^+\Psi\text{CTT}$
- 25–39%, $\text{TT}^+\Psi\text{CTT}^+\Psi\text{CT}$, $\text{T}^+\Psi\text{CTTT}^+\Psi\text{CT}$

- <25%, $\text{TT}^+\Psi\text{C}$, $\Psi\text{C}^+\Psi\text{C}^+\Psi\text{C}$, $\text{TT}^+\Psi\text{C}^+\Psi\text{C}^+\Psi\text{CTT}$, $\text{TT}^+\Psi\text{CT}^+\Psi\text{CTT}$, $\text{TT}^+\Psi\text{CmeC}^+\Psi\text{CTT}$

With a T or C on either side, the propensity slightly shifts to favor tautomer $\Psi\text{C}(\text{H3})$. Methylated C or protonated C neighbors also shift the propensity to favor tautomer $\Psi\text{C}(\text{H3})$, but less so when they are both methylated and protonated. Analysis of the base–base interaction energies revealed that $\Psi\text{C}(\text{H3})$ has favorable electrostatic interactions with a methylated or protonated C neighbor on its 3'-side (Figure S1).

Protonation of an unmethylated LNA-C (C) neighbor offers little to no improvement in $\Psi\text{C}(\text{H3})$ propensity, and protonation of meC disfavors tautomer $\Psi\text{C}(\text{H3})$. We would like to reiterate here that the normal commercial version of LNA-C is always methylated (meC), and we include unmethylated LNA-C in our computational study to delineate the contributions of methylation and sugar locking.

The 7-mer $\text{TT}^+\Psi\text{CmeC}^+\Psi\text{CTT}$, which contains two ΨCs next to meC , disfavors $\Psi\text{C}(\text{H3})$ compared with $\text{TT}^+\Psi\text{CT}^+\Psi\text{CTT}$, but notably the position trend is reversed: the 5' ΨC in $\text{TT}^+\Psi\text{CT}^+\Psi\text{CTT}$ favors $\Psi\text{C}(\text{H3})$ more, as generally observed in other systems (vide infra), but not in $\text{TT}^+\Psi\text{CmeC}^+\Psi\text{CTT}$ where the 3' ΨC favors $\Psi\text{C}(\text{H3})$. However, by itself, locked sugars in the neighboring LNA residues have a modest to no effect on the tautomeric propensity. There is a modest improvement going from $\text{T}^+\Psi\text{CT}$ to $\text{T}^+\Psi\text{CT}$, but little to none between $\text{meC}^+\Psi\text{CmeC}$ and $\text{meC}^+\Psi\text{CmeC}$ or $\Psi\text{C}^+\Psi\text{CT}$ and $\Psi\text{C}^+\Psi\text{CT}$.

When there is more than a single ΨC residue in the system, their tautomeric states do not appear to strongly correlate with each other. In the trimer $\Psi\text{C}^+\Psi\text{CT}$, when the first ΨC is the $\Psi\text{C}(\text{H1})$ tautomer, the second ΨC has similar tendencies to be $\Psi\text{C}(\text{H1})$ or $\Psi\text{C}(\text{H3})$; and the same is observed when the first ΨC is $\Psi\text{C}(\text{H3})$ (Table 1). However, when there are more residues in between, the ΨC nearer to the 3'-end always favors tautomer $\Psi\text{C}(\text{H1})$, as observed in 7-mers $\text{TT}^+\Psi\text{CT}^+\Psi\text{CTT}$, $\text{TT}^+\Psi\text{CTT}^+\Psi\text{CT}$, and $\text{T}^+\Psi\text{CTTT}^+\Psi\text{CT}$. In $\Psi\text{C}^+\Psi\text{C}^+\Psi\text{C}$ and $\text{TT}^+\Psi\text{C}^+\Psi\text{C}^+\Psi\text{CTT}$, when the first ΨC is $\Psi\text{C}(\text{H3})$, both the second and third ΨCs tend to be $\Psi\text{C}(\text{H1})$.

The position of ΨC in the sequence has a large effect on the tautomeric propensity. When ΨC is at the 5'-end in ΨCTT , the tautomeric propensity is 63% $\Psi\text{C}(\text{H3})$, whereas when ΨC is at the 3'-end in $\text{TT}^+\Psi\text{C}$, it is 24%. This position effect can also be clearly observed in 7-mer $\text{T}^+\Psi\text{CTTT}^+\Psi\text{CT}$, where both ΨCs are flanked by T, but their propensities are vastly different [63 and 28% $\Psi\text{C}(\text{H3})$, respectively]. We observed that when ΨC is positioned toward the 3'-end, it often forms intramolecular

Figure 3. Transition of λ_{H1} values in one run each of nucleoside ΨC (model system used as the reference; henceforth “ref”) and nucleotide ΨC . Only $0.8 \leq \lambda_{\text{H1}} \leq 1$ is counted as the physical state of tautomer $\Psi\text{C}(\text{H1})$, whereas $0 \leq \lambda_{\text{H1}} \leq 0.2$ would be counted as the physical state of tautomer $\Psi\text{C}(\text{H3})$. The base structures of the corresponding tautomers are shown. The transition rates are 60 and 40 ns^{-1} , and the fractions of physical states are 83 and 84%.

Figure 4. Tautomeric propensity [given in terms of % tautomer $\Psi\text{C}(\text{H}3)$] of pseudoisocytidine in different systems. ΨC is pseudoisocytidine; meC is 5-methylcytosine; $^+$ indicates protonation; underline denotes residues with locked sugar (LNA). When there are multiple ΨC s, data for the ΨC at the nearest 5'-end are presented first. The error bar is the standard error of mean of five independent runs.

hydrogen bonds with the preceding residues. Notably, the hydrogen bonding analyses of trimers and 7-mers show that H1 is much more frequently involved in intramolecular hydrogen bonding compared with N1, N3, and H3, and it is, to a large extent, correlated with the appearance of tautomer $\Psi\text{C}(\text{H}1)$. We select two examples from one run of trimers $\text{T}\Psi\text{CT}$ and $\text{TT}\Psi\text{C}$ to show such correlation (Figure 5). The position effect can thus be explained in terms of intramolecular hydrogen bonds: when ΨC is positioned toward the 3'-end, it has more available hydrogen bonding partners for H1, favoring the formation of associated tautomer $\Psi\text{C}(\text{H}1)$.

More detailed analyses were undertaken for 7-mer $\text{TT}\Psi\text{C}\Psi\text{C}\Psi\text{CTT}$, which is a fragment of 17-mer $\text{TFOS-DNALNA}\Psi\text{C}$ used in the triplex formation experiments. For this 7-mer sequence, we performed another set of conventional MD simulations, fixing the tautomeric states to be the most populated one (combination 311: 59%, Table 1) to exclude artifacts from the dual topology on hydrogen bond and solvent-accessible surface area (SASA) analyses.

The two protonation sites associated with the two tautomers are in similar chemical environments, except for O2 near N3/H3. In 7-mer $\text{TT}\Psi\text{C}\Psi\text{C}\Psi\text{CTT}$ (311), residues $\Psi\text{C}4$ and $\Psi\text{C}5$ have some intramolecular hydrogen bonds involving H1, whereas H3 in $\Psi\text{C}3$ has no such hydrogen bonds (Table 2). Notably, the position effect can be observed here: on average, $\Psi\text{C}5$ nearer to the 3'-end has its H1 involved in more intramolecular hydrogen bonds than $\Psi\text{C}4$, which is nearer to the 5'-end (Table 2).

The SASA for the N1 atom of Ψ in 7-mer $\text{TT}\Psi\text{C}\Psi\text{C}\Psi\text{CTT}$ (311) is lower for $\Psi4$ and $\Psi5$ than for $\Psi3$, even when considering the presence of H1 in $\Psi4$ and $\Psi5$, whereas for the N3 atom, the SASA is very similar for $\Psi3$ and $\Psi4$ even though $\Psi3$ has H3 present and $\Psi4$ does not (Figure S2). This is consistent with the observation that $\Psi(\text{H}1)$ tends to form intramolecular hydrogen bonds; thus, it tends to be less exposed to the solvent.

Consecutive ΨC lowers $\Psi\text{C}(\text{H}3)$ propensities, except for the first residue at the 5'-end. In $\Psi\text{C}\Psi\text{CT}$, the first ΨC has a moderate propensity for $\Psi\text{C}(\text{H}3)$ [65% $\Psi\text{C}(\text{H}3)$], but the second favors $\Psi\text{C}(\text{H}1)$ instead [42% $\Psi\text{C}(\text{H}3)$]. The third

Table 1. Average Population (in %) of Tautomer Combinations with Standard Error of Mean in Five Independent Runs^a

2 Ψ C								
tautomer combination	Ψ C Ψ CT	Ψ C Ψ CT	TT Ψ CT Ψ CTT	TT Ψ CTT Ψ CT	T Ψ CTTT Ψ CT	TT Ψ C ^{me} C ⁺ Ψ CTT	triplex TT Ψ CT Ψ CTT	triplex TT Ψ C ^{me} C ⁺ Ψ CTT
11	18 (2)	19 (1)	51 (9)	43 (7)	26 (4)	45 (4)	0 (0)	0 (0)
13	15 (3)	14 (3)	9 (3)	7 (1)	10 (3)	36 (5)	0 (0)	0 (0)
31	39 (7)	48 (5)	33 (9)	43 (6)	46 (6)	12 (2)	0 (0)	0 (0)
33	27 (5)	18 (4)	8 (4)	7 (2)	18 (7)	8 (1)	100 (0)	100 (0)
3 Ψ C								
tautomer combination	Ψ C Ψ C Ψ C		TT Ψ C Ψ C Ψ CTT		triplex TT Ψ C Ψ C Ψ CTT			
111	11 (3)		30 (7)		0 (0)			
113	1 (1)		3 (1)		0 (0)			
131	15 (2)		2 (1)		0 (0)			
133	2 (1)		0 (0)		0 (0)			
311	40 (9)		59 (7)		0 (0)			
313	2 (2)		4 (1)		0 (0)			
331	25 (7)		2 (1)		0 (0)			
333	4 (3)		0 (0)		100 (0)			

^aTautomer combination is shown in shorthand; for example, tautomer combination 31 means that the first Ψ C is tautomer Ψ C(H3) and the second is Ψ C(H1).

Figure 5. Intramolecular hydrogen bonding and tautomeric states of one run of T Ψ CT and TT Ψ C. Hydrogen bond label denotes hydrogen bond pairs; for example, (2 N1-X) refers to the intramolecular hydrogen bond involving N1 of residue index 2; X is any intramolecular hydrogen acceptor or donor. Only the physical states of tautomers Ψ C(H1) (red) and Ψ C(H3) (black) are shown. Snapshots of the two trimers at 4 ns are shown with the intramolecular hydrogen bonds (orange); hydrogen atoms are not shown for clarity.

Table 2. Intramolecular Hydrogen Bond Occupancies of 7-mer TT Ψ C Ψ C Ψ CTT Fixed Tautomer 311^a

	1	2	3	4	5
3 N1-X			0.3		
3 N3-X					
3 N3-H3-X					
4 N1-X					
4 N1-H1-X	0.8		0.1	0.3	
4 N3-X					
5 N1-X					
5 N1-H1-X		0.3	0.4	0.8	0.2
5 N3-X					

^aOnly the hydrogen bonds involving N1, H1, N3, and H3 of the three Ψ C residues are shown. The row label is hydrogen bond pairs; for example, (3 N1-X) refers to intramolecular hydrogen bond involving N1 of residue index 3; X is any intramolecular hydrogen acceptor or donor. The column label is the run index of five independent runs. Blank refers to zero hydrogen bond occupancy.

consecutive Ψ C has an even more pronounced shift: in Ψ C Ψ C Ψ C, 3' Ψ C has a propensity of only 9% Ψ C(H3). Likewise, in 7-mer TT Ψ C Ψ C Ψ CTT, the last Ψ C has a similar propensity [7% Ψ C(H3)], and in addition, the middle Ψ C also significantly shifts [3% Ψ C(H3)].

The low propensities for Ψ C(H3) of consecutive Ψ C are deleterious for triplex formation. Not only does the 3' Ψ C have

a low propensity for Ψ C(H3), but the population of favorable tautomer combinations [all Ψ C(H3)] is extremely low. In the trimer Ψ C Ψ C Ψ C, the all- Ψ C(H3) (333) population is only 4%; and in 7-mer TT Ψ C Ψ C Ψ CTT, it is 0% (Table 1). However, when the 7-mer is in a triplex, the all- Ψ C(H3) population becomes 100%. When the middle Ψ C in the triplex TT Ψ C Ψ C Ψ CTT is substituted with T so that the Ψ Cs are no longer consecutive, as in the triplex TT Ψ CT Ψ CTT, or substituted with ^{me}C⁺ and introducing LNA neighbors, as in triplex TT Ψ C^{me}C⁺ Ψ CTT, the all- Ψ C(H3) population is, as expected, 100%. This suggests that Hoogsteen hydrogen bonding in the triplex is strong enough to shift the tautomeric propensity to favor Ψ C(H3) and confer thermodynamic stability. The low all- Ψ C(H3) population in single-stranded systems is of concern because there is only a small amount of the “correct” population, that is, with all- Ψ C(H3), that can bind to the duplex, which may result in slow kinetics of binding.

To characterize the behavior of Ψ C(H1) in the triplex environment, we performed classical MD simulations of triplex TT Ψ C Ψ C Ψ CTT tautomer combinations 133 and 331. Although Ψ C(H3) forms Hoogsteen hydrogen bonds with G, Ψ C(H1) partially flips out and interacts with N7 or 5' phosphate of G instead (Figure 6). The residue Ψ C(H1) is not observed to completely flip out, and the triplexes stay mostly stable during 100 ns simulations (Watson–Crick and Hoogsteen hydrogen bonds during the simulations are shown in Figure S3). The

Figure 6. Observed configurations when ΨC is in TFO. (A) Canonical configuration when $\Psi\text{C(H3)}$ is in TFO with Watson–Crick (black) and Hoogsteen (red) hydrogen bonds. (B,C) $\Psi\text{C(H1)}$ partially flips out and interacts with N7 or 5' phosphate of G (purple).

average structures show helical distortions around $\Psi\text{C(H1)}$ (Figure 7).

Figure 7. Classical MD simulations of triplexes TT $\Psi\text{C}\Psi\text{C}\Psi\text{CTT}$ tautomer combinations 331, 133, and 333. Average structures from the last 50 ns of the 100 ns simulation are shown in side and top views (duplex in green and TFO in orange).

In summary, from our simulations, we have found that the neighboring residues have different effects on ΨC tautomerization. Methylated or protonated C shifts the tautomeric propensity to favor $\Psi\text{C(H3)}$; T or LNA neighbors do not affect the tautomerization equilibrium directly; ΨC itself as a neighbor affects the tautomeric propensity to disfavor $\Psi\text{C(H3)}$,

which is not desirable in the context of TFO binding in triplex formation.

Verifying the Effect of Consecutive and Nonconsecutive ΨC in TFOs for in Vitro Binding. Getting a stable TFO formation in vitro requires longer TFO sequences than the 7-mer TFO used in the simulations. To verify the effect of consecutive ΨC -residues, we thus designed TFOs as 17-mers, targeting a region in the human *TFF* gene close to an ERE. This target is a good candidate for in vitro studies of the influence of ΨC in 17-mer TFOs because it contains a majority of Gs, including stretches of consecutive Gs 5'-AGGGG-GAAGGGAAGGAG-3'.³⁴ We decided to evaluate TFOs with this size because previous in vitro studies performed with 13-mer TFOs containing ΨC bases did not show any TFO binding (unpublished experiments). Each TFO was hybridized with the double-stranded (DS) target for a period of up to 72 h at pH 7.4, and the triplex formation was analyzed using EMSA.

Pseudoisocytidines were located in a consecutive or nonconsecutive manner in the TFOs. Three different stretches of two, three, and five consecutive ΨCs were present in the sequences of TFO1-DNAfull ΨC , TFO2-DNAfull ΨC -TINA, TFO3-DNALNAfull ΨC , and TFO4-DNALNAfull ΨC -TINA, where TINA denotes twisted intercalating nucleic acid. One or two thymines (DNA or LNA) were spaced between them, and all of these sequences contained a ΨC at the 3'-end ultimate position.

Initially, the DNA containing TFO1-DNAfull ΨC was evaluated. In this sequence, all Cs were substituted by ΨC . After 72 h of incubation, no triplex formation was detected, even in the presence of the triplex-stabilizing BQQ compound

Figure 8. TFO binding of 17-mer TFO sequences containing consecutive ΨC : (a) DSS1 and electrophoretic mobility shift profile of DSS1 in the presence of (b) TFO1-DNAfull ΨC and (c) TFO3-DNALNAfull ΨC , both with 11/17 nucleotides being ΨC and in TFO3 4/6 Ts being LNA Ts. Hybridization with TFO, in the absence of and (as indicated only at the highest ratio) in the presence of BQQ, was carried out for 72 h. Triplex structures are detected as slower migrating bands. DNA duplex and triplex complexes are indicated as DS and TS, respectively.

Table 3. TFO Sequence and Triplex Formation under Intranuclear Conditions^a

Name	Evidence of triplex formation in EMSA assay (-BQQ)	Sequence
TFO1-DNAfull ^Ψ C	No	5' - T ^Ψ C ^Ψ C ^Ψ C ^Ψ C ^Ψ C T T ^Ψ C ^Ψ C ^Ψ C T T ^Ψ C ^Ψ C T ^Ψ C - 3'
TFO3-DNALNAfull ^Ψ C	No	5' - T ^Ψ C ^Ψ C ^Ψ C ^Ψ C ^Ψ C T T ^Ψ C ^Ψ C ^Ψ C T T ^Ψ C ^Ψ C T ^Ψ C - 3'
TFO5-DNALNA ^Ψ C	Yes	5' - T ^{me} C ^Ψ C ^{me} C ^Ψ C ^{me} C T T ^Ψ C ^{me} C ^Ψ C T T ^{me} C ^Ψ C T ^Ψ C - 3'
TFO6-DNALNA ^{me} C	No	5' - T ^{me} C ^{me} C ^{me} C ^{me} C ^{me} C T T ^{me} C ^{me} C ^{me} C T T ^{me} C ^{me} C T ^{me} C - 3'
TFO7-DNA ^Ψ C	No	5' - T ^Ψ C C C ^Ψ C T T ^Ψ C C ^Ψ C T T ^Ψ C C T ^Ψ C - 3'

^aTriplet sequence in which ^ΨC gives >75% ^ΨC(H3) tautomeric propensity in the λ -dynamics simulations are shaded.

Figure 10. Simulations for triplex-containing TFO5-DNALNA^ΨC with classical MD. (A) Average structure from one run (duplex in green and TFO in orange), side and top views. (B) Triplex base pair parameter distributions (excluding 2 residues at either TFO end), from the last 50 ns of four independent 100 ns runs for the triplex containing TFO5. The dashed and dotted lines represent the average values for A- and B-form DNAs, respectively.⁵¹

stabilization as measured by the melting temperature when three Cs were substituted with ^ΨC and 2'-O-methyl-^ΨC.^{44–46} Also, it has been shown that ^ΨC combined with peptide nucleic acid (PNA) at the 3'-end of the TFO in a nonconsecutive manner with every second position containing a T^{47,48} can reduce the pH sensitivity. On the basis of NMR experiments, Leitner et al. have demonstrated that for intramolecular TFO-DNA, protonation is disfavored for adjacent C or for C at the end of the triplex,⁴⁹ also arguing in favor of our in silico and in vitro results. All of these reports are in line with our conclusion that ^ΨC in a nonconsecutive manner is the best option for designing DNA/LNA mixmer TFOs containing ^ΨC. The TFO was able to target the G-rich region under intranuclear conditions when ^ΨC is flanked by ^{me}C or T, in agreement with simulation results (Table 3). To our knowledge, this is the first time that this is shown in vitro for TFO DNA/LNA containing ^ΨC.

Two different intercalators have been used in this work: BQQ and TINA. BQQ is a triplex helix-intercalating compound that can bind specifically to and stabilize the triplex structures of purine and pyrimidine motifs.⁵⁰ We have also previously used BQQ to confirm and probe for triplex formation using different TFOs (Pabon et al. unpublished result). There we demonstrated that BQQ could also stabilize triplexes formed by LNA-containing TFOs. Here, we can confirm that this is also valid for C-rich TFOs with consecutive ^ΨC. Also, this is the first study showing that BQQ can stabilize ^ΨC-containing TFO DNA/LNA. The influence of TINA positioning is not discussed in this work, but we have chosen to locate TINA at the 3'-end position based on previous work (Pabon et al. unpublished result). Surprisingly, the presence of TINA in our

^ΨC-containing TFOs seems not to increase the rate of triplex formation under the experimental conditions used here.

Structural Characterization of Triplex-Containing TFO5-DNALNA^ΨC. To characterize the 3D structure of the triplex containing TFO5-DNALNA^ΨC, we performed simulations with classical MD at fixed tautomeric states. We fixed the tautomerization state of ^ΨC to be ^ΨC(H3); based on the result in the smaller triplexes, all-^ΨC(H3) is favored in the triplex environment. We found that the triplex is stable during the course of the simulation, preserving most of the Watson–Crick and Hoogsteen hydrogen bonds (Figure 10). Out of four independent runs, one has a loss of Hoogsteen hydrogen bond towards the 3'-end, but the other three are similar, where all Hoogsteen hydrogen bonds are preserved most of the time (one such run is shown in Figure S7). The average structure shows no obvious distortion in the triplex structure that might contribute to instability (Figure 10). Upon binding of TFO5, the DNA duplex overslides in the negative direction to accommodate the third strand. The resultant helical structure has slide and twist parameters similar to A-type duplex DNA, but an α -displacement value is between those of A- and B-types (Figure 10). Replacing LNA with DNA in the bound TFO does not affect the structural feature of the triplex (Figure S8), supporting that what promotes LNA-containing TFO binding to the DNA duplex is that the TFO is indeed preorganized for major groove binding (Pabon et al. unpublished data). In conclusion, upon binding of ^ΨC-containing TFO, the triple helical conformation is between A- and B-types with base pairs remaining almost perpendicular to the helical axis, in agreement with what was observed in other DNA duplexes involved in triplex formation.²⁸

CONCLUSIONS

We have performed λ -dynamics simulations and binding experiments under intranuclear conditions to investigate the ability of pseudoisocytidine to efficiently target the major groove of a DNA duplex and form a triplex structure. In particular, we have investigated the tautomerization of $^{\Psi}\text{C}$ in different short single-stranded and triplex DNAs. In single strands, we have observed a clear influence of sequence on the tautomeric propensity of $^{\Psi}\text{C}$. The predisposition for tautomer $^{\Psi}\text{C}(\text{H3})$ is higher when the neighboring residues are cytidines (even higher when cytidine is 5-methylated) compared with when they are thymine. When the neighboring residues are $^{\Psi}\text{C}$, the propensity of tautomer $^{\Psi}\text{C}(\text{H3})$ (located between two $^{\Psi}\text{Cs}$) is low. Furthermore, the sugar modification LNA on the neighboring residues does not affect the $^{\Psi}\text{C}$ tautomerization equilibrium directly.

Once the single strand is bound to the targeted duplex, forming a triplex, nearly all $^{\Psi}\text{Cs}$ are $^{\Psi}\text{C}(\text{H3})$ tautomers, allowing hydrogen bonding with the Hoogsteen site of the guanine of the double strand, even if the propensity of $^{\Psi}\text{C}(\text{H3})$ in free, unbound TFOs was very low, as in the case of consecutive runs of $^{\Psi}\text{Cs}$. This suggests that Hoogsteen hydrogen bonding in the triplex is strong enough to shift the tautomeric predisposition to favor $^{\Psi}\text{C}(\text{H3})$ and confer thermodynamic stability.

The in vitro experiment shows that the TFOs having three or more consecutive $^{\Psi}\text{Cs}$, such as TFO-DNAfull $^{\Psi}\text{C}$ and TFO-DNALNAfull $^{\Psi}\text{C}$, were unable to form triplexes under intranuclear salt conditions at pH 7.4 (also when ligands promoting triplex formation such as BQQ and TINA were included). Only when nonconsecutive $^{\Psi}\text{Cs}$ were included in combination with alternating DNA/LNA residues (LNA residues are $^{\text{me}}\underline{\text{C}}$ and $\underline{\text{T}}$), the 17-mer TFO was able to form a triplex. In the formed triplex, the pseudoisocytidine targeted the Hoogsteen site of the guanine with two hydrogen bonds, and the duplex structure goes under conformation rearrangement with slide and twist parameters similar to A-type, but the x -displacement is between those of A- and B-forms.

We conclude, based on the combination of in silico and in vitro studies, that the inclusion of alternating $^{\Psi}\text{C}$ and the combination with alternating LNA enhances the formation of the evaluated C-rich intermolecular triplexes. Therefore, based on our results, we suggest that when designing DNA/LNA mixmer TFOs containing $^{\Psi}\text{C}$, incorporation of $^{\Psi}\text{C}$ and LNA in a nonconsecutive manner is preferable.

MATERIALS AND METHODS

Simulations. The tautomeric equilibria are influenced by chemical and physical factors, including solvent, ion concentration, and biomolecular environment. An accurate prediction should account for the small energy differences that cause shifts in the tautomeric equilibrium and the need to sample different conformational states accessible to the biomolecule. Methods based on a macroscopic description of the biomolecule and the solvent do not explicitly account for dielectric heterogeneity and response to conformational rearrangement, nor do they take into account the conformational rearrangement.⁵² We choose the λ -dynamics approach with an explicit description of the solvent⁵³ because it addresses these issues by enabling the direct coupling between tautomerization processes and conformational dynamics. Moreover, the accuracy of the

method can be improved through a fine calibration of the force field.

Theory. Multisite λ -dynamics⁵³ is set up for the $^{\Psi}\text{C}$ residue such that the two tautomeric states $^{\Psi}\text{C}(\text{H1})$ and $^{\Psi}\text{C}(\text{H3})$ are described and propagated by continuous variables λ_{H1} and λ_{H3} , respectively. The potential energy function is given by

$$U_{\text{tot}}(X, \{x\}, \{\lambda\}) = U_{\text{env}}(X) + \lambda_{\text{H1}}(U(X, x_{\text{H1}}) - \Delta G_{\text{H1} \rightarrow \text{H3}}(\text{model}) + \lambda_{\text{H3}}(U(X, x_{\text{H3}}) + F^{\text{bias}}(\lambda_{\text{H1}}) + F^{\text{bias}}(\lambda_{\text{H3}}))$$

where X is the coordinates of the environment atoms, x_{H1} and x_{H3} are the coordinates of atoms in $^{\Psi}\text{C}$, corresponding to the tautomers $^{\Psi}\text{C}(\text{H1})$ and $^{\Psi}\text{C}(\text{H3})$ respectively, and

$$F^{\text{bias}}(\lambda_{\text{Hi}}) = \begin{cases} k_{\text{bias}}(\lambda_{\text{Hi}} - 0.8)^2 & \text{if } \lambda_{\text{Hi}} < 0.8 \\ 0 & \text{otherwise} \end{cases}$$

where Hi refers to the tautomeric state $^{\Psi}\text{C}(\text{H1})$ or $^{\Psi}\text{C}(\text{H3})$. λ_{Hi} scales the potential energy of the corresponding tautomer with the constraints

$$0 \leq \lambda_{\text{Hi}} \leq 1 \text{ and } \lambda_{\text{H1}} + \lambda_{\text{H3}} = 1$$

$\Delta G_{\text{H1} \rightarrow \text{H3}}(\text{model})$ is the free energy for transforming tautomer $^{\Psi}\text{C}(\text{H1})$ into $^{\Psi}\text{C}(\text{H3})$ of the pseudoisocytidine model compound in aqueous solution. This term is included to flatten the potential energy surface such that the two tautomeric states of the nucleoside in solution are equipopulated as the free energy between the two states becomes zero. The model compound structure used in the free energy calculation is the reference state at which the tautomeric ratio $^{\Psi}\text{C}(\text{H1})/^{\Psi}\text{C}(\text{H3})$ is 1:1. In the investigated systems, the deviation from this tautomeric ratio would come from the contribution of the environment.

Two harmonic biasing potentials $F^{\text{bias}}(\lambda_{\text{H1}})$ and $F^{\text{bias}}(\lambda_{\text{H3}})$ are included to bias the sampling toward the physical end states. In this formulation, $0.8 \leq \lambda_{\text{Hi}} \leq 1$ is considered to be a physical end state. k_{bias} is the force constant of the harmonic potentials. The force constant is equal for both tautomeric states.

Calculation of Free Energy. The free energy, $\Delta G_{\text{H1} \rightarrow \text{H3}}(\text{model})$, was calculated using the Bennett acceptance ratio (BAR) method.⁵⁴ The pseudoisocytidine hybrid model compound used in the free energy perturbation calculation is constructed with deoxyribose sugar and 5' and 3' hydroxyls and two pyrimidine bases corresponding to the two tautomeric states (CHARMM dual topology); the two bases are maintained within the same volume of space by distance restraints between all pairs of common atoms in the two tautomeric states. The residue is solvated with TIP3P water molecules⁵⁵ in a cubic box with 20 Å side length. The CHARMM BLOCK module⁵⁶ is used to partition the system into three blocks: (I) environment, (II) tautomer $^{\Psi}\text{C}(\text{H1})$ base, and (III) tautomer $^{\Psi}\text{C}(\text{H3})$ base. Interactions between blocks II and III are set to null; interactions between blocks I and II and within block II are scaled with λ ; interactions between blocks I and III and within block III are scaled with $1 - \lambda$. Scaling with λ and $1 - \lambda$ is not applied to the bond, angle, and dihedral energy terms. Eight λ values corresponding to alchemical intermediate/end states were used ($\lambda = 0.0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9$, and 1.0), and each window simulation length is 1 ns, with only the last 900 ps used for the BAR

Table 4. ON Sequences^a

name	length (nt)	sequence
TFO1-DNAfull ^Ψ C	17	5'-T ^Ψ C ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ CT ^Ψ C-3'
TFO2-DNAfull ^Ψ C-TINA	17	5'-T ^Ψ C ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ CTP ^Ψ C-3'
TFO3-DNALNAfull ^Ψ C	17	5'- <u>T</u> ^Ψ C ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ CT ^Ψ C-3'
TFO4-DNALNAfull ^Ψ C-TINA	17	5'- <u>T</u> ^Ψ C ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ C ^Ψ CTT ^Ψ C ^Ψ CTP ^Ψ C-3'
TFO5-DNALNA ^Ψ C	17	5'-T ^{me} C ^{me} C ^{me} C ^{me} C ^{me} CTT ^{me} C ^{me} C ^{me} CTT ^{me} C ^{me} CT ^{me} C-3'
TFO6-DNALNA ^{me} C	17	5'-T ^{me} C ^{me} C ^{me} C ^{me} C ^{me} CTT ^{me} C ^{me} C ^{me} CTT ^{me} C ^{me} CT ^{me} C-3'
TFO7-DNA ^Ψ C	16	5'-T ^Ψ CCC ^Ψ CTT ^Ψ CC ^Ψ CTT ^Ψ CCT ^Ψ C-3'

^aDNA is indicated in capital letters and LNA is indicated in underlined capital letters; P, p-TINA; ^{me}C, 5-methyl-C; and ^ΨC, pseudoisocytidine. The commercial version of LNA C is always methylated.

calculation. The CHARMM REPD module⁵⁶ is used to run eight alchemical intermediate/end states in parallel and attempt to exchange energies every 1000 steps (Hamiltonian replica exchange). The calculated free energy is 28.9 ± 0.1 kcal/mol.

Calibration of the Biasing Potential Force Constant. The value of k_{bias} is calibrated by performing 1 ns runs of λ -dynamics at various k_{bias} values and observing the fraction of physical end states in the trajectory (FPL, fraction physical ligand) and the frequency of transitions between the two tautomeric states (transition rate, ns⁻¹). An optimal value of k_{bias} was chosen so as to yield a high transition rate and simultaneously maintain a high FPL (above 0.8).⁵³ The initial k_{bias} values are 15, 20, 25, and 30 kcal/mol. Additional runs are added as needed to determine the optimal value. The value of the optimized k_{bias} is 19.5 kcal/mol.

Simulation Settings. Simulations were performed with CHARMM (version 41a2)⁵⁶ with CHARMM36 force field for DNA,⁵⁷ TIP3P water,⁵⁵ ions,⁵⁸ and modified nucleic acids^{59,60} (for ^ΨC). Updated LNA parameters were used (Xu; Nilsson; Villa unpublished result). Initial 7-mer and 17-mer triplex structures were taken from a parallel DNA triplex fiber model from the 3DNA Web server.⁶¹ For both 7-mer and 17-mer triplexes, the duplex is longer than the TFO by 2 residues at either side. The third strand of the triplex is used as the initial structure for the single-strand trimers and 7-mers.

For single-strand monomers, trimers, 7-mers, and 7-mer triplexes, λ -dynamics was used to allow interconversion between the two ^ΨC tautomers. Several selected systems were also run with the standard MD (fixed tautomer state) to aid analysis—these are run with the same cutoff, settings, and lengths as the λ -dynamics run, but with single topology without λ scaling. For the 17-mer triplex, conventional MD simulation is used and the tautomeric state of ^ΨC is fixed to tautomer ^ΨC(H3), the tautomer involved in Hoogsteen hydrogen bonding. For all λ -dynamics simulations, five independent runs were performed. Simulation lengths were chosen so that the standard errors of the mean of the tautomeric propensities from five runs do not exceed 10%. For monomers and trimers, these are 6 ns; 7-mers, 8 ns; and 7-mer triplexes, 40 ns. For MD simulation of the 17-mer triplex, four independent 100 ns runs were performed. Simulation systems and lengths are summarized in Table S1.

The structures were minimized with the steepest descent and adopted-basis Newton–Raphson methods with large position restraints on the heavy atoms. For triplexes, additional distance restraints are added for Watson–Crick and Hoogsteen hydrogen bonds. The systems were solvated in boxes of TIP3P water molecules, with dimensions of $20 \times 20 \times 20$ Å³ for monomers, $50 \times 50 \times 50$ Å³ for trimers and 7-mers, $65 \times 45 \times 45$ Å³ for 7-mer triplexes, and $88 \times 42 \times 42$ Å³ for 17-mer

triplexes. After the addition of sodium ions to neutralize the system, additional sodium and chloride ions are included to reach an ionic concentration of approximately 0.1 M.

λ -dynamics is performed within the CHARMM BLOCK module⁵⁶ using the multisite λ -dynamics framework (MSLD).⁶²

The functional form of λ_{Hi} , $\lambda_{\text{Hi}} = \frac{e^{5.5 \sin \theta_{\text{Hi}}}}{\sum_i e^{5.5 \sin \theta_{\text{Hi}}}}$, was used. Defining

λ as a function of θ in this way has been shown to be optimal for sampling and convergence.⁶³ θ is assigned a fictitious mass of 12 amu·Å² (amu = atomic mass unit). The temperature was maintained at 298 K by coupling to Langevin heat bath with a collision frequency of 10 ps⁻¹. λ is saved every 10 steps. The bond, angle, and dihedral energy terms are excluded from scaling by λ so that only geometrically relevant states are sampled. A sampling bias was applied for each tautomeric state with a force constant of 19.5 kcal/mol. A nonbonded list cutoff of 15 Å was used with the electrostatic force switch and van der Waals switch functions between 10 and 12 Å. Simulations were performed in the NVT ensemble with Langevin dynamics with a collision frequency of 10 ps⁻¹. For 7-mer triplexes, distance restraints were applied to Watson–Crick and Hoogsteen hydrogen bonds (distance 2.9 Å and force constant 10 kcal/mol/Å²) for the base pairs at 5'- and 3'-end positions.

MD simulations for 17-mer triplexes were performed on graphics processing units (GPUs) with CHARMM⁵⁶ and the CHARMM/OpenMM⁶⁴ interface in the NVT ensemble with Langevin dynamics with a collision frequency of 5 ps⁻¹. A van der Waals force switching function was used between 8 and 9 Å. Particle mesh Ewald was used to treat electrostatic interactions with a nonbonded cutoff of 8 Å and a grid point spacing of 1.0 Å. The distance cutoff in generating the list of pairwise interactions was 17 Å. The temperature was maintained at 298 K by coupling to a Langevin heat bath with a frictional coefficient of 10 ps⁻¹. Distance restraints were applied to Watson–Crick hydrogen bonds (distance 2.9 Å and force constant 10 kcal/mol/Å²) for the base pair in the 5'- and 3'-ends and all Hoogsteen N7–N3 hydrogen bonds (same distance and force constant). After minimization, the system was equilibrated for 2 ns. During production, the distance restraints on the Hoogsteen hydrogen bonds were released.

In all simulations, the SHAKE algorithm was used to constrain bonds involving hydrogen.⁶⁵ A lookup table was used for interactions between water molecules,⁶⁶ except for GPU simulations. The leapfrog integrator was used with an integration time-step of 2 fs. The hydrogen bond, interaction energy, and SASA analyses were performed within CHARMM. Triplex base pair step parameter analyses were performed with Curves+.⁶⁷

In Vitro Binding Experiments. ONs. Standard methods were used to synthesize the ONs containing pyrimidine

Table 5. Target Sequences Used for Experiments^a

Name	Length (nt)	Sequence
DS48	48	5' - GCCACCGTGACCTTGCAGGGGAAGGGAAGGAGCTCATGAAGCTTCGAG - 3' 3' - CGGTGGCACTGGAACGTCCCCCTTCCCTTCCTCGAGTACTTCGAAGCTC - 5'*
DS51	51	5' - GGCCACCGTGACCTTGCAGGGGAAGGGAAGGAGCTCATGAAGCTTCGAGA - 3' 3' - CCGGTGGCACTGGAACGTCCCCCTTCCCTTCCTCGAGTACTTCGAAGCTCT - 5'*

^aDNA is indicated in capital letters. DS, double-stranded target sequences: DS48 and DS51. The TFO binding site is shown in a gray box with letters in bold. The star (*) indicates the strand which was radiolabelled using [γ -³²P] ATP.

analogues and TINA. TFO5 and TFO6 were synthesized at the Nucleic Acid Center at the University of Southern Denmark in Jesper Wengel Laboratory. Mixer LNA/DNA ONs were synthesized using solid-phase phosphoramidite chemistry on an automated DNA synthesizer on a 1.0 mmol synthesis scale.⁶⁸ Purification to at least 85% purity of all modified ONs was performed using reversed-phase high-performance liquid chromatography (RP-HPLC) or ion-exchange HPLC (IE-HPLC), and the composition of all synthesized ONs was verified using matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) analysis recorded using 3-hydroxy-picolinic acid as a matrix. TFO1-TFO4 and TFO7 were provided by Anapa Biotech A/S Company from Denmark. DNA target sequences were ordered from Sigma. The ONs and target sequences used here are presented in Tables 4 and 5, respectively. The ON concentrations of stock solutions were confirmed using a NanoDrop spectrophotometer (Thermo Scientific).

ON Hybridization. The double-stranded (DS) target (5.0 nM) was incubated with ONs at different concentrations (0.5, 1.0, 2.0, and 4.0 μ M, corresponding to TFO versus DS target ratios of 100, 200, 400, and 800, respectively). Many cytosine-rich ONs can potentially form an intramolecular i-motif⁶⁹ and to avoid that, the TFOs were heated before hybridization for 5 min at 65 °C, followed by cooling on ice. Hybridization was performed in an intranuclear buffer (Tris-acetate 50 mM, pH 7.4, 120 mM KCl, 5 mM NaCl, and 0.5 mM MgOAc) and in a total volume of 10 μ L at 37 °C for up to 72 h in the absence or presence of BQQ (1 μ M).

Preparation of ³²P-Labeled dsDNA Target. The pyrimidine strand of the target sequence was labeled using [γ -³²P] ATP and T4 polynucleotide kinase (Fermentas) according to the manufacturer's protocol and then purified using a QIAquick Nucleotide Removal Kit (Qiagen). The pyrimidine strand labeled ON was annealed with the unlabeled complementary strand at a 1:1 ratio. The annealing was performed by heating for 5 min at 95 °C, followed by decreasing the temperature to 40 °C at a rate of 1 grade per minute using a thermocycler.

Electrophoretic Mobility Shift Assay (EMSA). DNA complexes were analyzed using nondenaturing polyacrylamide gel electrophoresis 10% (29:1) in Tris-acetate–ethylenediaminetetraacetic acid (EDTA) (TAE) buffer (1 $\times$, pH 7.4 supplemented with 0.5 mM MgOAc and 5 mM NaCl). The gels were run at 150 V and 200 mA for 5 h with circulation water cooling and analyzed using a Molecular Imager FX system. The intensity of the gel bands was quantified using the Quantity One software (Bio-Rad). All experiments were repeated three times.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00347.

Electrostatic interaction energy between bases of various trimers from MD simulation, distributions of SASA of ¹³C residues, Watson–Crick and Hoogsteen hydrogen bonds during the simulations, TFO binding of 17-mer TFO sequences containing TINA and consecutive ¹³C, TFO binding of 16-mer TFO sequences containing no consecutive ¹³C, TFO binding of 17-mer TFO sequences containing nonconsecutive ¹³C, and simulations for triplex containing TFO5-DNALNA¹³C and another with the same TFO sequence with LNA changed to DNA using classical MD (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: alessandra.villa@ki.se. Phone: +46-8-52481081 (A.V.).

ORCID

Jesper Wengel: 0000-0001-9835-1009

Alessandra Villa: 0000-0002-9573-0326

Present Address

#(A.U.) Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, 48824, United States.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS), Colombia (Ph.D. grant resolución 02007/24122010 to Y.V.P.-M.), Nanyang Technological University Research Scholarship (Y.D.H.), and the Swedish Research Council (VT 2015-04992, K2015-68X-11247-21-3) and EU Marie Skłodowska-Curie [ITN 71613, MMBio] for support. The authors are indebted to Søren Morgenthaler Echwald, Anapa Biotech A/S, Hørsholm, Denmark, for providing TINA-containing ONs.

■ ABBREVIATIONS

BQQ, benzoquinoxaline; C, G, A, T, DNA bases (cytosine, guanine, adenine, thymine); EMSA, electrophoretic mobility shift assay; ERE, estrogen response element; human TFF, human trefoil factor; LNA, locked nucleic acid; MD, molecular dynamics; ^mC, 5-methyl-cytidine; ^mC, 5-methyl-cytidine LNA; NMR, nuclear magnetic resonance; ON, oligonucleotide; SASA, solvent-accessible surface area; T,

thymine LNA; TFO, triplex-forming oligonucleotide; TINA, p-twisted intercalating nucleic acid; ^3C , pseudoisocytidine

REFERENCES

- (1) Zain, R.; Sun, J.-S. Do natural DNA triple-helical structures occur and function in vivo? *Cell. Mol. Life Sci.* **2003**, *60*, 862–870.
- (2) Baran, N.; Lapidot, A.; Manor, H. Formation of DNA Triplexes Accounts for Arrests of DNA Synthesis at d(TC)_n and d(GA)_n Tracts. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 507–511.
- (3) Daube, S. S.; von Hippel, P. H. Functional transcription elongation complexes from synthetic RNA–DNA bubble duplexes. *Science* **1992**, *258*, 1320–1324.
- (4) Dayn, A.; Samadashwily, G. M.; Mirkin, S. M. Intramolecular DNA triplexes: Unusual sequence requirements and influence on DNA polymerization. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 11406–11410.
- (5) Møllegaard, N. E.; Buchardt, O.; Egholm, M.; Nielsen, P. E. Peptide nucleic acid/DNA strand displacement loops as artificial transcription promoters. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 3892–3895.
- (6) Hampel, K. J.; Lee, J. S. Two-dimensional pulsed-field gel electrophoresis of yeast chromosomes: Evidence for triplex-mediated DNA condensation. *Biochem. Cell Biol.* **1993**, *71*, 190–196.
- (7) Veselkov, A. G.; Malkov, V. A.; Frank-Kamenetskii, M. D.; Dobrynin, V. N. Triplex model of chromosome ends. *Nature* **1993**, *364*, 496.
- (8) Ito, T.; Smith, C. L.; Cantor, C. R. Sequence-specific DNA purification by triplex affinity capture. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 495–498.
- (9) Ito, T.; Smith, C. L.; Cantor, C. R. Triplex affinity capture of a single copy clone from a yeast genomic library. *Nucleic Acids Res.* **1992**, *20*, 3524.
- (10) Ito, T.; Smith, C. L.; Cantor, C. R. Affinity capture electrophoresis for sequence-specific DNA purification. *Genet. Anal.: Biomol. Eng.* **1992**, *9*, 96–99.
- (11) Vary, C. P. Triple-helical capture assay for quantification of polymerase chain reaction products. *Clin. Chem.* **1992**, *38*, 687–694.
- (12) Olivas, W. M.; Maher, L. J. Analysis of duplex DNA by triple helix formation: Application to detection of a p53 microdeletion. *BioTechniques* **1994**, *16*, 128–132.
- (13) Havre, P. A.; Glazer, P. M. Targeted mutagenesis of simian virus 40 DNA mediated by a triple helix-forming oligonucleotide. *J. Virol.* **1993**, *67*, 7324–7331.
- (14) Havre, P. A.; Gunther, E. J.; Gasparro, F. P.; Glazer, P. M. Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7879–7883.
- (15) Frank-Kamenetskii, M. D.; Mirkin, S. M. Triplex DNA structures. *Annu. Rev. Biochem.* **1995**, *64*, 65–95.
- (16) Soyfer, V. N.; Potaman, V. N. General Features of Triplex Structures. In *Triple-Helical Nucleic Acids*; Springer: New York, 1996; pp 100–150.
- (17) Goñi, J. R.; de la Cruz, X.; Orozco, M. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.* **2004**, *32*, 354–360.
- (18) Goñi, J. R.; Vaquerizas, J. M.; Dopazo, J.; Orozco, M. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **2006**, *7*, 63.
- (19) Beal, P. A.; Dervan, P. B. Second structural motif for recognition of DNA by oligonucleotide-directed triple-helix formation. *Science* **1991**, *251*, 1360.
- (20) Helene, C.; Thuong, N. T.; Harel, A. Control of Gene Expression by Triple Helix-Forming Oligonucleotides. The Antigenic Strategy. *Ann. N. Y. Acad. Sci.* **1992**, *660*, 27–36.
- (21) Paugh, S. W.; Coss, D. R.; Bao, J.; Lauder milk, L. T.; Grace, C. R.; Ferreira, A. M.; Waddell, M. B.; Ridout, G.; Naeve, D.; Leuze, M.; Lo Cascio, P. F.; Panetta, J. C.; Wilkinson, M. R.; Pui, C.-H.; Naeve, C. W.; Uberbacher, E. C.; Bonten, E. J.; Evans, W. E. MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. *PLoS Comput. Biol.* **2016**, *12*, No. e1004744.
- (22) Ono, A.; Ts'o, P. O. P.; Kan, L. S. Triplex formation of oligonucleotides containing 2'-O-methylpseudoisocytidine in substitution for 2'-deoxycytidine. *J. Am. Chem. Soc.* **1991**, *113*, 4032–4033.
- (23) Kan, L.-S.; Lin, W.-C.; Yadav, R. D.; Shih, J. H.; Chao, I. NMR Studies of the Tautomerism in Pseudoisocytidine. *Nucleosides Nucleotides* **1999**, *18*, 1091–1093.
- (24) Mayer, A.; Häberli, A.; Leumann, C. J. Synthesis and triplex forming properties of pyrrolidino pseudoisocytidine containing oligodeoxynucleotides. *Org. Biomol. Chem.* **2005**, *3*, 1653–1658.
- (25) Shahid, K. A.; Majumdar, A.; Alam, R.; Liu, S.-T.; Kuan, J. Y.; Sui, X.; Cuenoud, B.; Glazer, P. M.; Miller, P. S.; Seidman, M. M. Targeted cross-linking of the human β -globin gene in living cells mediated by a triple helix forming oligonucleotide. *Biochemistry* **2006**, *45*, 1970–1978.
- (26) Singh, V.; Fedeles, B. I.; Essigmann, J. M. Role of tautomerism in RNA biochemistry. *RNA* **2014**, *21*, 1–13.
- (27) Khandogin, J.; Brooks, C. L. Constant pH molecular dynamics with proton tautomerism. *Biophys. J.* **2005**, *89*, 141–157.
- (28) Esguerra, M.; Nilsson, L.; Villa, A. Triple helical DNA in a duplex context and base pair opening. *Nucleic Acids Res.* **2014**, *42*, 11329–11338.
- (29) Aviñó, A.; Cubero, E.; González, C.; Eritja, R.; Orozco, M. Antiparallel triple helices. Structural characteristics and stabilization by 8-amino derivatives. *J. Am. Chem. Soc.* **2003**, *125*, 16127–16138.
- (30) Semenyuk, A.; Darian, E.; Liu, J.; Majumdar, A.; Cuenoud, B.; Miller, P. S.; MacKerell, A. D.; Seidman, M. M. Targeting of an Interrupted Polypurine:Polypyrimidine Sequence in Mammalian Cells by a Triplex-Forming Oligonucleotide Containing a Novel Base Analogue. *Biochemistry* **2010**, *49*, 7867–7878.
- (31) Escudé, C.; Nguyen, C. H.; Kukreti, S.; Janin, Y.; Sun, J.-S.; Bisagni, E.; Garestier, T.; Hélène, C. Rational design of a triple helix-specific intercalating ligand. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 3591–3596.
- (32) Zain, R.; Marchand, C.; Sun, J.-s.; Nguyen, C. H.; Bisagni, E.; Garestier, T.; Hélène, C. Design of a triple-helix-specific cleaving reagent. *Chem. Biol.* **1999**, *6*, 771–777.
- (33) Højland, T.; Kumar, S.; Babu, B. R.; Umemoto, T.; Albaek, N.; Sharma, P. K.; Nielsen, P.; Wengel, J. LNA (locked nucleic acid) and analogs as triplex-forming oligonucleotides. *Org. Biomol. Chem.* **2007**, *5*, 2375–2379.
- (34) Klinge, C. M. Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Res.* **2001**, *29*, 2905–2919.
- (35) Geny, S.; Moreno, P. M. D.; Krzykowski, T.; Gissberg, O.; Andersen, N. K.; Isse, A. J.; El-Madani, A. M.; Lou, C.; Pabon, Y. V.; Anderson, B. A.; Zaghoul, E. M.; Zain, R.; Hrdlicka, P. J.; Jorgensen, P. T.; Nilsson, M.; Lundin, K. E.; Pedersen, E. B.; Wengel, J.; Smith, C. I. E. Next-generation bis-locked nucleic acids with stacking linker and 2'-glycylamino-LNA show enhanced DNA invasion into supercoiled duplexes. *Nucleic Acids Res.* **2016**, *44*, 2007–2019.
- (36) Sharma, B. D.; McConnell, J. F. The crystal and molecular structure of isocytosine. *Acta Crystallogr.* **1965**, *19*, 797–806.
- (37) Hirao, I.; Kimoto, M.; Yamakage, S.-i.; Ishikawa, M.; Kikuchi, J.; Yokoyama, S. A unique unnatural base pair between a C analogue, pseudoisocytosine, and an A analogue, 6-methoxypurine, in replication. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1391–1393.
- (38) Kan, L.-S.; Lin, W.-C.; Yadav, R. D.; Shih, J. H.; Chao, I. NMR Studies of the Tautomerism in Pseudoisocytidine. *Nucleosides Nucleotides* **1999**, *18*, 1091–1093.
- (39) Filichev, V. V.; Pedersen, E. B. Stable and selective formation of Hoogsteen-type triplexes and duplexes using twisted intercalating nucleic acids (TINA) prepared via postsynthetic Sonogashira solid-phase coupling reactions. *J. Am. Chem. Soc.* **2005**, *127*, 14849–14858.
- (40) Filichev, V. V.; Gaber, H.; Olsen, T. R.; Jørgensen, P. T.; Jessen, C. H.; Pedersen, E. B. Twisted intercalating nucleic acids—intercalator influence on parallel triplex Stabilities. *Eur. J. Org. Chem.* **2006**, 3960–3968.

- (41) Faucon, B.; Mergny, J.-L.; Hélène, C. Effect of third strand composition on the triple helix formation: Purine versus pyrimidine oligodeoxynucleotides. *Nucleic Acids Res.* **1996**, *24*, 3181–3188.
- (42) Lee, J. S.; Woodsworth, M. L.; Latimer, L. J. P.; Morgan, A. R. Poly(pyrimidine) poly(purine) synthetic DNAs containing 5-methylcytosine form stable triplexes at neutral pH. *Nucleic Acids Res.* **1984**, *12*, 6603–6614.
- (43) Vekhoff, P.; Ceccaldi, A.; Polverari, D.; Pylouster, J.; Pisano, C.; Arimondo, P. B. Triplex formation on DNA targets: How to choose the oligonucleotide. *Biochemistry* **2008**, *47*, 12277–12289.
- (44) Volker, J.; Klump, H. H. Electrostatic effects in DNA triple helices. *Biochemistry* **1994**, *33*, 13502–13508.
- (45) Sugimoto, N.; Wu, P.; Hara, H.; Kawamoto, Y. pH and Cation Effects on the Properties of Parallel Pyrimidine Motif DNA Triplexes. *Biochemistry* **2001**, *40*, 9396–9405.
- (46) Chin, T.-M.; Lin, S.-B.; Lee, S.-Y.; Chang, M.-L.; Cheng, A. Y.-Y.; Chang, F.-C.; Pasternack, L.; Huang, D.-H.; Kan, L.-S. “Paper-clip” type triple helix formation by 5′-d-(TC)₃Ta(CT)₃Cb(AG)₃ (*a* and *b* = 0–4) as a function of loop size with and without the pseudoisocytosine base in the Hoogsteen strand. *Biochemistry* **2000**, *39*, 12457–12464.
- (47) Egholm, M.; Christensen, L.; Deulholm, K. L.; Buchardt, O.; Coull, J.; Nielsen, P. E. Efficient pH-independent sequence-specific DNA binding by pseudoisocytosine-containing bis-PNA. *Nucleic Acids Res.* **1995**, *23*, 217–222.
- (48) Hansen, M. E.; Bentin, T.; Nielsen, P. E. High-affinity triplex targeting of double stranded DNA using chemically modified peptide nucleic acid oligomers. *Nucleic Acids Res.* **2009**, *37*, 4498–4507.
- (49) Leitner, D.; Schröder, W.; Weisz, K. Influence of sequence-dependent cytosine protonation and methylation on DNA triplex stability. *Biochemistry* **2000**, *39*, 5886–5892.
- (50) Zain, R.; Polverari, D.; Nguyen, C.-H.; Blouquit, Y.; Bisagni, E.; Garestier, T.; Grierson, D. S.; Sun, J.-S. Optimization of Triple-Helix-Directed DNA Cleavage by Benzoquinoxaline–Ethylenediaminetetraacetic Acid Conjugates. *ChemBioChem* **2003**, *4*, 856–862.
- (51) Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X.-J.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C.-S.; Westhof, E.; Wolberger, C.; Berman, H. M. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **2001**, *313*, 229–237.
- (52) Wallace, J. A.; Shen, J. K. Predicting pK_a values with continuous constant pH molecular dynamics. *Methods Enzymol.* **2009**, *466*, 455–475.
- (53) Goh, G. B.; Knight, J. L.; Brooks, C. L. Constant pH molecular dynamics simulations of nucleic acids in explicit solvent. *J. Chem. Theory Comput.* **2012**, *8*, 36–46.
- (54) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (55) Jørgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.
- (56) 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*, 1545–1614.
- (57) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.* **2012**, *8*, 348–362.
- (58) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J. Chem. Phys.* **1994**, *100*, 9050–9063.
- (59) Xu, Y.; Nilsson, L.; MacKerell, A. D. An Additive Charmm Force Field for Modified Nucleic Acids. *Biophys. J.* **2015**, *108*, 235a–236a.
- (60) MacKerell, A. D. CHARMM Force Field Files. http://mackerell.umaryland.edu/charmm_ff.shtml#charmm (accessed Feb 6, 2017).
- (61) Zheng, G.; Lu, X.-J.; Olson, W. K. Web 3DNA—A web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **2009**, *37*, W240–W246.
- (62) Knight, J. L.; Brooks, C. L. λ -Dynamics Free Energy Simulation Methods. *J. Comput. Chem.* **2009**, *30*, 1692–1700.
- (63) Knight, J. L.; Brooks, C. L. Applying efficient implicit nongeometric constraints in alchemical free energy simulations. *J. Comput. Chem.* **2011**, *32*, 3423–3432.
- (64) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* **2009**, *30*, 864–872.
- (65) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (66) Nilsson, L. Efficient Table Lookup Without Inverse Square Roots for Calculation of Pair Wise Atomic Interactions in Classical Simulations. *J. Comput. Chem.* **2009**, *30*, 1490–1498.
- (67) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **2009**, *37*, S917–S929.
- (68) Singh, S. K.; Koshkin, A. A.; Wengel, J.; Nielsen, P. LNA (locked nucleic acids): Synthesis and high-affinity nucleic acid recognition. *Chem. Commun.* **1998**, 455–456.
- (69) Mergny, J.-L.; Lacroix, L. Kinetics and thermodynamics of i-DNA formation: Phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.* **1998**, *26*, 4797–4803.

SUPPORTING INFORMATION

The role of pseudoisocytidine tautomerisation in triplex-forming oligonucleotides: *in silico* and *in vitro* studies

Yossa Dwi Hartono^{1,5}, Y. Vladimir Pabon-Martinez², Arzu Uyar^{1a}, Jesper Wengel³, Karin E. Lundin², Rula Zain^{2,4}, C. I. Edvard Smith², Lennart Nilsson¹, Alessandra Villa^{1*}

¹ Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden.

² Department of Laboratory Medicine, Clinical Research Center, Karolinska Institutet, SE-141 86 Huddinge, Sweden.

³ Department of Physics, Chemistry and Pharmacy, Nucleic Acid Center, University of Southern Denmark, 5230 Odense M, Denmark.

⁴ Department of Clinical Genetics, Centre for Rare Diseases, Karolinska University Hospital, SE-171 76 Stockholm, Sweden.

⁵ Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

*To whom correspondence should be addressed. Tel: +46-8-52481081; Email: alessandra.villa@ki.se

^a Current address: Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan (MI), 48824, United States.

Table S1. System and simulation length.

System	Simulation length
Monomers	
Nucleoside Ψ C (ref)	5 x 2 ns
Nucleotide Ψ C	5 x 6 ns
Trimers ΨC,T	
$T\Psi CT$ $\underline{T}\Psi CT$ ΨCTT $TT\Psi C$ $\Psi C\Psi CT$ $\Psi C\Psi \underline{CT}$	5 x 6 ns
Trimers ΨC,C,$\overset{me}{C}$,T	
$\underline{C}^{+\Psi}CT$ $T\Psi CC^{+}$ $\overset{me}{C}^{+\Psi}CT$ $\underline{T}\Psi C^{me}C^{+}$ $\underline{C}\Psi CT$ $\underline{T}\Psi CC$ $\overset{me}{C}\Psi CT$ $\underline{T}\Psi C^{me}C$	
7-mers	
$TT\Psi C\Psi C\Psi CTT$ $TT\Psi CT\Psi CTT$ $TT\Psi CT$ $T\Psi CT$	5 x 8 ns
$T\Psi CTT\Psi CTT$ $\underline{T}\Psi C^{me}C^{+\Psi}CTT$	
7-mer triplexes	
Triplex $TT\Psi C\Psi C\Psi CTT$	
5'- T T Ψ C Ψ C Ψ C T T -3'	
5'- G C $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ G G -3'	
3'- $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ -5'	5 x 40 ns
Triplex $TT\Psi CT\Psi CTT$	
5'- T T Ψ C T Ψ C T T -3'	
5'- G C $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ G G -3'	
3'- $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ -5'	
Triplex $\underline{T}\Psi C^{me}C^{+\Psi}CTT$	
5'- T $\underline{T}$ Ψ C $\overset{me}{C}$ Ψ C $\underline{T}$ T -3'	
5'- G C $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ G G -3'	
3'- $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ -5'	
17-mer triplex	
Triplex containing TFO5-DNALNA Ψ C	
5'- T $\overset{me}{C}$ Ψ C $\overset{me}{C}$ Ψ C $\overset{me}{C}$ T $\underline{T}$ Ψ C $\overset{me}{C}$ Ψ C $\underline{T}$ T $\overset{me}{C}$ Ψ C $\underline{T}$ Ψ C -3'	4 x 100 ns
5'- G G $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ $\overset{\bullet}{G}$ $\overset{\bullet}{A}$ $\overset{\bullet}{G}$ G G -3'	
3'- $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{T}$ $\overset{ }{C}$ $\overset{ }{C}$ $\overset{ }{C}$ -5'	
Ψ C is pseudoisocytidine $\overset{me}{C}$ is 5-methylcytosine; $^{+}$ indicates protonation; <u>underline</u> denotes residues with locked sugar (LNA).

Figure S1. Electrostatic interaction energy between bases of various trimers from MD simulation (fixed tautomer). Data are collated from 5 runs (for each system, each tautomer).

Figure S2. Distributions of solvent-accessible surface areas (SASA) of Ψ C residues in 7-mer TT Ψ C Ψ Ψ CTT (311), data from all 5 fixed-tautomer runs combined. Upper panel shows SASA for the whole base; lower panel shows SASA for only N1 and N3 atoms.

Figure S3. Watson-Crick and Hoogsteen hydrogen bonds for triplex TT^ΨC^ΨC^ΨCTT, tautomer combinations 133 (left) and 331 (right). Only one 100-ns run of simulation each is shown.

(a) DS51

5' - GGCCACCGTGACCTTGCAGGGGGAAGGGAAGGAGCTCATGAAGCTTCGAGA - 3'
 3' - CCGGTGGCACTGGAACGTCCCCCTTCCCTTCCTCGAGTACTTCGAAGCTCT - 5'

Figure S4. TFO binding of 17-mer TFO sequences containing TINA and consecutive ^ΨC: (a) DS51 and electrophoretic mobility shift profile of DS51 in the presence of (b) TFO2-DNAfull^ΨC-TINA and (c) TFO4-DNALNAfull^ΨC-TINA. Hybridisation with TFO in the absence (left side) and in the presence (right side) of BQQ was carried out for 72 h. Triplex structures are detected as slower migrating bands. DNA duplex and triplex complexes are indicated as DS and TS, respectively.

(a) DS48

5' - GCCACCGTGACCTTGCAGGGGGAAGGGAAGGAGCTCATGAAGCTTCGAG - 3'
 3' - CCGGTGGCACTGGAACGTCCCCCTTCCCTTCCTCGAGTACTTCGAAGCTC - 5'

Figure S5. TFO binding of 16-mer TFO sequences containing no consecutive ^ΨC: (a) DS51 and electrophoretic mobility shift profile of DS51 in the presence of (b) TFO7-DNA^ΨC. Hybridisation with TFO in the absence (left side) and in the presence (right side) of BQQ was carried out for 72 h. Triplex structures are detected as slower migrating bands. DNA duplex indicated as DS

(a) DS51

5' - GGCCACCGTGACCTTGC**AGGGGGAAGGGAAGGAG**CTCATGAAGCTTCGAGA - 3'
3' - CCGGTGGCACTGGAACGT**CCCCCTTCCCTTCCT**CGAGTACTTCGAAGCTCT - 5'

Figure S6. TFO binding of 17-mer TFO sequences containing non-consecutive ^{13}C in buffer 4-Morpholineethanesulfonic acid (MES) at pH 6.0: (a) DS51 and electrophoretic mobility shift profile of DS51 in the presence of (b) TFO5-DNALNA ^{13}C . Hybridisation with TFO in the absence (left side) and in the presence (right side) of BQQ was carried out up to 24 h. Triplex structures are detected as slower migrating bands. DNA duplex and triplex complexes are indicated as DS and TS, respectively.

Figure S7. Watson-Crick and Hoogsteen hydrogen bonds of triplex containing TF05-DNALNA^ΨC (1 out of 4 independent runs). The triplex sequence is shown and aligned with the hydrogen bonds (Watson-Crick: black dash, Hoogsteen: red dot).

Figure S8. Simulations for triplex containing TFO5-DNALNA^ΨC (black) and another with the same TFO sequence with LNA changed to DNA (red) with classical MD. Triplex base pair parameter distributions (excluding 2 residues at either TFO end), from the last 50 ns of independent 100 ns runs. Dashed and dotted lines represent average values for A- and B-form DNA, respectively.

CORRECTION FOR THIS THESIS

In the publication, p. 2173, the equation for potential energy function of λ -dynamics is missing a parenthesis. The equation should read (missing bracket is shaded):

$$\begin{aligned} U_{\text{tot}}(X, \{x\}, \{\lambda\}) \\ = U_{\text{env}}(X) + \lambda_{\text{H1}}(U(X, x_{\text{H1}}) - \Delta G_{\text{H1} \rightarrow \text{H3}}(\text{model})) + \lambda_{\text{H3}}U(X, x_{\text{H3}}) + F^{\text{bias}}(\lambda_{\text{H1}}) \\ + F^{\text{bias}}(\lambda_{\text{H3}}) \end{aligned}$$

A computational study of uracil tautomeric forms on the ribosome: the case of uracil and 5-oxyacetic acid uracil in the first anticodon position of tRNA

Yossa Dwi Hartono^{1,2}, Mika Ito^{1,(a)}, Alessandra Villa¹ and Lennart Nilsson^{1*}

¹ Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden.

² Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

*To whom correspondence should be addressed. Tel: +46-8-52481099; Email: Lennart.Nilsson@ki.se

^(a)Current address Department of Anatomy, Physiology and Biochemistry, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

Abstract

Tautomerism plays an important role in many biomolecular interactions, not least in RNA biology. Crystallographic studies increasingly show the possible presence of the minor tautomer form of tRNA anticodon bases on the ribosome. However, since the hydrogen positions are not resolved in the X-ray studies, this structural observation and its energetic consequences is worth investigating further. Here, we employ ab initio calculations and molecular dynamics simulations to understand if and how the minor enol form of uracil (U), or of a modified uracil with an oxyacetic acid moiety at position 5 (cmo⁵U), can be accommodated in tRNA:mRNA interactions in the ribosome decoding center. Ab initio calculations on isolated bases show that the modification affects the keto-enol equilibrium of the uracil base only slightly, and the keto form is dominant (>99.99%) in both U and cmo⁵U; other factors such as interactions with surrounding nucleotides or ions would therefore be required in order shift the equilibrium to favor the enol tautomer. Classical molecular simulations accounting for the ribosomal environment show a better agreement with X-ray structure for the enol form but free energy calculations still indicating the most stable form is the keto. In the ribosome instead of the expected three hydrogen bonds, the enol tautomer of U, as well as cmo⁵U, pairs with a guanine forming two hydrogen bonds, which do not involve the enol group. The oxyacetic acid modification does not keep the conformation observed in the X-ray structure and has minor contribution on the keto-enol equilibrium.

Introduction

The ribosome is an RNA-protein complex responsible for the translation of genetic information to protein synthesis. During translation, the transfer-RNA (tRNA) is selected by the ribosome to match the messenger-RNA (mRNA) in the A-site ¹. The ribosomal scaffold ensures the fidelity of translation by making non-complementary tRNA-mRNA interactions unfavorable. This selection is tighter for the first and second codon positions, while some mismatches are allowed for the third position ². One example is G:U mismatch which typically form the wobble configuration with two hydrogen bonds, instead of three in a complementary Watson-Crick G:C match.

Tautomerization of nucleic acid bases is known to influence RNA biochemistry (see ³ for recent review). On one hand the minor tautomers may be advantageous in the biochemical activities of RNA enzymes and aptamers; on the other hand, they may be implicated as the source of replication, transcription, and translation error. In the context of tRNA-mRNA interactions in ribosomal translation and in other relevant interactions requiring base complementarity involved in replication and transcription, the minor tautomers can be detrimental to fidelity since the tautomers would have different sets of hydrogen bond donor and acceptors, causing errors in these processes.

Since tautomerization investigation is experimentally challenging due to extremely low abundance of the minor tautomers, tautomerization studies may use, for example, a highly specialized experimental technique such as NMR relaxation dispersion ⁴, where the authors observed tautomeric populations of nucleic acids that are consistent to replication and translation errors. There are also examples of approaches which combined experiments and quantum mechanical (QM) calculation⁵. Singh et al. ^{5a} for instance attempted 2D IR, variable temperature FTIR and NMR spectroscopies, binding isotope effects (BIEs), and computational methods to identify and quantify tautomers of a ligand bound to an RNA aptamer. There are also examples of pure computational approaches⁶, which typically utilize ab initio and/or free energy calculations. Maximoff et al. ^{6a} noted a G:T mismatch with Watson-Crick geometry in DNA polymerase λ and through molecular dynamics (MD) simulations, free energy perturbation, and ab initio calculations concluded that the enol G accounts for this observation. Satpati and Åqvist ^{6b} examined G:U mismatches in first and second codon positions in ribosomal A-site and through the same set of computational tools also concluded that enol G is the most likely explanation.

In crystallographic studies of ribosome complexes ⁷, minor enol tautomers have been put forward as possible explanations for the observed Watson-Crick geometries of uracil:guanine (U:G) base pairs. For example, U:G type base pairs were observed in the first and second codon positions in the ribosome A- and P-sites ^{7a,b} for unmodified uridines, but there are also examples in the third/wobble codon position, where U is modified and is on the tRNA ^{7c,d}. The 5-oxyacetic acid (cmo⁵) modification of uridine in the first anticodon position (U34), which base pairs with third codon position, accept all four bases ⁸. Weixlbaumer et al. ^{7c} solved the X-ray crystal structures with the four possible pairings and observed the expected geometries with A/U/C, but Watson-Crick instead of wobble geometry with G. The authors proposed the enol form of cmo⁵U as the explanation for this observation with three main arguments: (1) More favorable cmo⁵U:G interaction. In Watson-

Crick geometry, there would be three hydrogen bonds, instead of two in wobble geometry. (2) The modification. Hillen et al.⁹ observed that the crystal structure of 5-methoxyuridine showed an increase in the C4-O4 bond length and a decrease in the N3-C4 bond length, indicating that the group attached to C5 affects the keto-enol equilibrium of uridine. Analogously, the similar 5-oxyacetic acid modification is expected to affect the keto-enol equilibrium of uridine. (3) The ribosomal environment. Dabkowska et al.^{6c} showed that interactions with a glycine zwitterion decreased the energy difference between the keto and enol tautomers. The ribosomal decoding center environment, particularly base pairing to G, might also stabilize the enol tautomer and lower the energy difference.

However, since the hydrogen positions are not resolved in the structures, and though the electron density seems to rule out the wobble geometry, it is not conclusive (3.1 Å resolution). In this study, we examined the hypothesis that enol form of cmo⁵U is consistent with the observed Watson-Crick geometry, taking into accounts the above arguments in favor of it.

First, we calculate the energetic difference between the keto and enol forms of U and cmo⁵U, for the base alone and base paired with a G in aqueous environments with ab initio calculations to find out whether the modification makes the enol form more favorable.

To take the ribosomal environment surrounding the basepair, we sampled the conformations of keto and enol forms of U and cmo⁵U in the ribosomal decoding center environment with molecular dynamics. In addition, we calculated the free energy of tautomerization in aqueous and ribosomal environments, using molecular mechanics (MM) description, corrected with QM/MM calculations for more accuracy.

Finally, since we found that some conformations in the crystal structure do not keep in molecular dynamics simulation, particularly the energetically favorable Watson-Crick geometry, we speculated on alternative or additional hypotheses besides the enol form of cmo⁵U.

Methods

***Ab initio* calculations**

Initial coordinates of the bases (U and cmo⁵U) and the base pairs (U:G and cmo⁵U:G) were taken from the X-ray crystal structure (PDB: 2UU9)^{7c}. Hydrogen atoms were added with the CHARMM program¹⁰ using the CHARMM36 force field for nucleic acids¹¹ and the force field created for the modified uridine¹². Sugars were substituted by hydrogen atoms. Atom numbering for U and cmo⁵U is shown in Figure S1.

Geometry optimizations were carried out using the Hartree-Fock (HF) method¹³ or the second-order Møller-Plesset (MP2) perturbation method¹⁴ with the 3-21+G** basis set. For the base pairs, the Watson-Crick geometries, in which two bases interact with their Watson-Crick edges, were targeted for geometry optimizations. For the enol form, although there are two directions of the H atom of the hydroxy group, only the conformation with the H atom directed to the Watson-Crick edge was used in this study, because the H atom is considered to be involved in the interaction between the modified uridine and guanosine in the previous experimental study^{7c}. Energies of the structures were calculated at MP2/6-31++G**//MP2/3-

21+G** level. Solvent effects were included using the polarizable continuum model (PCM)¹⁵ with a dielectric constant = 78.4. QM calculations were performed with the Gaussian 03¹⁶.

Molecular dynamics simulations

Molecular dynamics simulations were performed for the modified and unmodified uracil base alone and base paired to guanine, in solution and in the ribosome complex using the CHARMM (version 41a2) program¹⁰ and the CHARMM36 force field for nucleic acids¹¹, ions¹⁷ and modified nucleotide parameters¹².

The ribosome complex was simulated as a spherical system with radius 25 Å, centered on tRNA-U34-O4, using the X-ray structure (PDB ID: 2UU9)^{7c} as starting structure and with a spherical boundary potential¹⁸ to prevent ions and water from leaving the sphere (**Figure 1**). Atom deletion outside the sphere is done by residue to prevent residue fragmentation. Atoms that ended up outside the sphere were restrained with a force constant of 2 kcal/mol/Å² throughout all simulations and minimizations. The systems were solvated using TIP3P water molecules¹⁹. Ions were added to neutralize the system. For more details on the setup of the starting structure, see Ref. ²⁰.

Figure 1. Representation of subset of the X-ray structure used for MD simulation, 25 Å radius sphere centered at cmo⁵U-G. Shown are mRNA (blue), tRNA anticodon stem loop fragment (red), various A-site ribosomal components (gray), water and ions (dots). Inset shows cmo⁵U-G basepair configuration of the X-ray crystal structure.

For simulations of the uracil base and of the U:G basepair (with and without modifications) the solutes were solvated using TIP3P water molecules¹⁹ in cubic 25×25×25 Å³ and orthogonal 40×20×20 Å³ boxes, respectively. Ions were added to neutralize the system and additional sodium and chloride ions are included to reach an ionic concentration of approximately 0.1 M. Simulations were performed using periodic boundary condition. To

keep the base pair planar, harmonic restraints were applied on base atoms (also on O7, C8, C9 for cmo⁵U) in the z direction with force constant of 50 kcal/mol/Å², such that the base pair can only move freely in the x-y plane. The length of the simulations is 5 ns.

The structures were minimized with steepest descent and adopted-basis Newton-Raphson methods with position restraints on the heavy atoms with force constant of 85 kcal/mol/Å². A non-bonded list cut-off of 16 Å was used with electrostatics force switch and van der Waals switch functions between 10 and 12 Å. Simulations were performed in the NVT ensemble with Langevin dynamics with collision frequency of 10 ps⁻¹. In all simulations, the SHAKE algorithm was used to constrain bonds involving hydrogens²¹. A lookup table was used for non-bonded interactions²². The leapfrog integrator was used with an integration time-step of 2 fs. The length of simulations is 5ns for base and basepair and 10 ns for ribosome.

Potential of mean force

Potential of mean force was calculated along the dihedral d3 (N3-C4-O4-H4) of the enol form of uracil using umbrella sampling. For umbrella sampling, a harmonic potential was applied along the dihedral d3 (N3-C4-O4-H4) in windows separated by 5° in the range of 0°-360°. Initial structures were generated with 0.6 ps simulation for each window with force constant of 1000 kcal/mol/rad². The final structure in each window was used as the starting structure in the next. The production run for each window is 500 ps, with harmonic force constant of 500 kcal/mol/rad² at 300 K. The potential of mean force (PMF) was constructed from the dihedrals in all windows from the last 250 ps using the Weighted Histogram Analysis Method²³.

A harmonic restraint with force constant 5 kcal/mol/Å² was applied to the N1-N3 distance to keep the basepair together during the sampling. For cmo⁵U, additional harmonic restraints with force constant 5 kcal/mol/Å² were applied to dihedrals C6-C5-O7-C8 and C5-O7-C8-C9 to keep the modification moiety in the position observed in the crystal structure.

Free energy calculations

The free energy of tautomerisation between the two tautomers, $\Delta G_{\text{keto} \rightarrow \text{enol}}$ was calculated with Bennet acceptance ratio (BAR) method²⁴. Simulations were done at both states and the potential energies of coordinates for each of the two trajectories were calculated for both states. The BAR equation is:

$$\Delta A(0 \rightarrow 1) = k_B T \ln \left(\frac{\langle f(U_0 - U_1 + C) \rangle_1}{\langle f(U_1 - U_0 - C) \rangle_0} \right) + C \quad (1)$$

Where k_B is Boltzmann constant, T is temperature, U_i is the potential energy of coordinates for state i , $\langle \rangle_i$ denotes ensemble average over state i , $f(x)$ is Fermi function $f(x) = \frac{1}{1 + \exp(\frac{x}{k_B T})}$, C is the free energy, which is to be solved in self-consistent manner.

To improve of quality of results, we also calculated QM/MM potential evaluating the same MM trajectory. The QM/MM energy is then used to reweigh the MM potential. This calculation is termed QM-Non-Boltzmann BAR (QM-NBB) and has been shown to improve free energy calculations²⁵. The equation introduces reweighting terms to the BAR equation (Eqn. 1), such that:

$$\Delta A(0 \rightarrow 1) = k_B T \ln \left(\frac{\langle f(U_0 - U_1 + C) \exp\left(\frac{V_1^b}{k_B T}\right) \rangle_{1,b} \langle \exp\left(\frac{V_0^b}{k_B T}\right) \rangle_{0,b}}{\langle f(U_1 - U_0 - C) \exp\left(\frac{V_0^b}{k_B T}\right) \rangle_{0,b} \langle \exp\left(\frac{V_1^b}{k_B T}\right) \rangle_{1,b}} \right) + C \quad (2)$$

Where the biasing potential V^b is the difference between MM and QM/MM energy:

$$V^b = U^{MM} - U^{QM/MM} \quad (3)$$

QM/MM reweighting can be done just on the end states (indirect) or all states (direct). If alchemical intermediate states are to be reweighted, the QM/MM energy is obtained by linear scaling:

$$U_{\lambda}^{QM/MM} = (1 - \lambda)U_0^{QM/MM} + \lambda U_1^{QM/MM} \quad (4)$$

For free energy calculations, 10-ns MD simulations for each λ window were done with U and modified U nucleoside in gas, aqueous, and ribosome phases. In aqueous and ribosome phases, same setup as MD simulations was used. 13 λ points were used for gas and aqueous phases ($\lambda = 0.00, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.90, 1.00$) and 11 λ points were used for ribosome phase ($\lambda = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00$).

QM/MM single point calculations were performed at B3LYP/6-31G* level with Q-Chem/CHARMM interface²⁶ to treat QM/MM interface on 100 ps snapshots from last 9 ns of each λ window. The nucleobase U/cmo⁵U was designated as QM region with linking hydrogen atom between the nucleobase and the ribose C1'. We checked that this level of theory and basis set (B3LYP/6-31G*) gave good accuracy compared to the single-point ab initio calculations at MP2/6-31++G by calculating the QM relative energy of one system used in ab initio calculations and noted that the value of relative keto-enol energy only differs by 0.1 kcal/mol.

Results and Discussion

Keto-enol equilibrium in aqueous environment

We performed quantum mechanical (QM) calculations for the keto and enol forms of uracil (U) and 5-oxyacetic uracil (cmo⁵U) alone and when they are base paired with guanine (G). Uracil and guanine were used as simple models of uridine and guanosine, respectively and the solvent effect was included using the polarizable continuum model (PCM). The major tautomeric form of uridine is the diketo form, and two enol tautomers are possible. We focus on the form involving O4, since it is the form proposed by the X-ray study^{7c} and previous theoretical studies on C5 substituents of uridine show that this enol form is more stable than that involving O2^{6d}.

In the X-ray crystal structure^{7c}, the 5-oxyacetic acid modification adopts a particular conformation, with the four atoms C5, O5, C7, and C8 co-planar with the pyrimidine ring, regardless of which base (uridine, cytidine, adenosine, or guanosine) is paired to the modified uridine. The dihedral angles d1(C6-C5-O5-C7) and d2(C5-O5-C7-C8) (Figure S2) of the modified uridine bound to guanosine in the X-ray crystal structure are 178.39° and 0.80°, respectively.

To examine whether the modified uracil (cmo⁵U) is stable with the modification in this conformation, energies of the two forms of the modified uracil, cmo⁵U_{keto} and cmo⁵U_{enol}, were calculated with the dihedral angle d1 fixed at 180°, while the dihedral angle d2 was held fixed at 30° intervals from -180° to +180° (Figure S2). cmo⁵U_{keto} is more stable than cmo⁵U_{enol} at all values of d2. It is also clear that d2 = 0° is the least stable conformation for both cmo⁵U_{keto} and cmo⁵U_{enol}. In the X-ray crystal structure, this unstable conformation of the modification may be stabilized by hydrogen bonds between the modified uridine and

neighboring nucleotides between O5 of the modified uridine and O2' of a neighboring uridine, and between O8 of the modified uridine and N6 of a neighboring adenosine.

In our calculations, the dihedral angles d1 and d2 were fixed at 180° and 0°, respectively, to take into consideration the effects of the modification in this particular conformation. The results for the isolated base show that the keto form is always the most favorable form and the keto-enol equilibrium is only slightly affected by the modifications. Having the hydroxyl group pointing toward the modification stabilizes the enol structure by 5-6 kcal/mol compared to other enol forms (Figure 2).

In the geometry optimization the base pairs with the keto tautomer of U shifted from the starting Watson-Crick geometry into a classic wobble geometry – a clear indication that for this tautomer the Watson-Crick geometry is not optimal. Energy calculations for the optimized geometries show this quantitatively (Figure 2). In the base pair between the unmodified U and G, two intermolecular hydrogen bonds are formed between U_{keto} and G, and three intermolecular hydrogen bonds are formed between U_{enol} and G. Despite these, the U_{keto}:G pair is about 4 kcal/mol more stable than the U_{enol}:G pair (Figure 2). The energy difference between U_{keto}:G and U_{enol}:G is smaller than the energy difference between U_{keto} and U_{enol}, indicating that the keto-enol equilibrium is affected by the additional hydrogen bond in the base pair with G.

Figure 2. *Ab initio* optimized structures of the keto and enol forms of U and cmo⁵U. For the enol forms two possible orientations for the hydroxyl H are considered. ΔE is the

difference in energy of the optimized structure relative to that of the keto tautomer, in kcal/mol. Donor-acceptor lengths of hydrogen bonds are shown in Å. The configurations found in X-ray crystal structure (PDB ID: 2UU9)^{7c} as well as those taken from MD simulations are also shown for comparison.

Base pairing of the modified U with G is similar to that of the unmodified. The cmo⁵U_{keto}:G pair is about 5 kcal/mol more stable than the cmo⁵U_{enol}:G pair, which at room temperature corresponds to an abundance of about 5×10^{-4} for the cmo⁵U_{enol}:G form. There is no major effect of the modification on the keto-enol stability.

Our results for the keto-enol equilibrium of U and cmo⁵U using PCM solvent model are comparable to previous gas-phase results^{6c}. We found that U_{keto} is 10.87 kcal/mol (ΔE) more stable than U_{enol} in water at the MP2/6-31++G** level; and cmo⁵U_{keto} is 11.34 kcal/mol (ΔE) more stable than cmo⁵U_{enol} in water. Dabkowski et al.^{6c} showed that U_{keto} is 12 kcal/mol (ΔE) more stable than U_{enol} in gas phase at the B3LYP/6-31++G** level. Additionally, both studies showed that, even though the energy difference between U_{keto} and U_{enol} is decreased by interactions with a neighboring molecule, U_{keto} is still more stable than U_{enol}. A more recent ab initio study^{6d} also found that U_{keto} is more stable than U_{enol} by 10.8 kcal/mol in water solution (CPCM-B3LYP-GD3/6-311++G(3df,2p)//B3LYP/6-31+G(d)), the U:G interaction is stabilized by around 2 kcal/mol for the keto form of unmodified U and by 3 kcal/mol for U with an analogous modification in the 5 position (5-methoxy U).

Keto-enol equilibrium in ribosomal decoding environment

To take into account the ribosomal decoding environment, we ran MD simulations with ribosomal region centered at A-site first anticodon position. The keto form of cmo⁵U was recently parameterized in the CHARMM force field as part of a set of parameters for naturally occurring modified nucleotides¹². The parameters for the enol form of cmo⁵U and U were also included in this set precisely due to the speculation of the enol form from the structural observation.

The keto forms of U and cmo⁵U converged to the expected wobble orientation (Figure 4). However, for the enol forms we encountered difficulties in accommodating a good U:G Watson-Crick geometry in the crystallographic structure. When adding hydroxyl hydrogen atom H4 to O4 to build the enol cmo⁵U, we noticed that the positions of cmo⁵U O4 and guanine O6 in the experimental structure is such that the uracil enol group (O4-H4) and the hydrogen bond acceptor G O6 are not collinear with angle of 108.2°, which is not an optimal hydrogen bond geometry. Additionally, there might also be steric repulsion between guanine H1 and uracil H4.

We attempted various strategies to stabilize the observed crystallographic Watson-Crick geometry. We tried distance restraints based on the Watson-Crick hydrogen bond distances, but the geometry does not hold once the restraints were released. We attempted tuning the C-O-H angle parameter such that O4-H4-O6 are closer to collinearity. We also attempted doing these strategies in more reduced system of just tRNA anticodon stem loop (ASL) and mRNA. However, none of these strategies preserved the canonical Watson-Crick geometry with three hydrogen bonds. One persistent observation is that the hydroxyl H4 prefers to be oriented away from the Watson-Crick edge (dihedral $d3(C3-N4-O4-H4) = 180^\circ$). We looked

at the force field parameter for of this dihedral (Figure S3) and 180° is not the global energy minimum. We thus investigated further to determine what caused the stabilization at this hydroxyl orientation.

We performed umbrella sampling along d3 of U/cmo⁵U for isolated base as well as base paired to guanine in aqueous condition. The gas-phase PMF is consistent with the profile of the model system 5-methoxy uracil used for the parameterization as expected (data not shown). The PMF profile in aqueous solution shows that the global minimum is at 180° (Figure 3), consistent with what was observed in the MD simulation. Hydrogen bond analysis showed that at d3= 180° , there is high probability of a water-bridged hydrogen bond between N3 and O4 (Figure S4). When U is base paired with G, the global minimum is even lower in energy. Notably, the local minimum previously at 0° slightly shifted such that the minimum is at a configuration where the H4 atom is not co-planar with the rest of the base (dihedral value 0°), indicating a steric constraint of the canonical Watson-Crick geometry.

Figure 3. Potential of mean force along the dihedral C3-N4-O4-H4 of the uracil enol form in aqueous environment (U_{enol} : black and cmo^5U_{enol} : red). The dotted line is the U/ cmo^5U ; the solid line is U/ cmo^5U basepaired with G. Restraints were applied to maintain base pairing with G and the orientation of the modification moiety. Structural snapshots at indicated dihedral angles are shown for unmodified U_{enol} and cmo^5U_{enol} to the right in the upper and lower panels respectively.

From our simulations, we observed that when the enol forms of U and cmo^5U are base paired to G in the ribosome they do not stay in a typical Watson-Crick geometry with three hydrogen bonds (Figure 4). This geometry is halfway between Watson-Crick and what Sochacka et al. calls new wobble orientation [30], in which G shifts towards the major groove side instead of towards the minor groove side as in typical wobble orientation (Figure 4).

Figure 4. Snapshots at 5 ns from simulations of the uridine-guanine base pair in the ribosome (green: X-ray, gray: keto, orange: enol). (A) U-G (B) cmo⁵U-G.

Basepair configurations of U_{enol}-G and cmo⁵U_{enol}-G taken from MD simulations at 5 ns converged to the 'new wobble' orientation when geometry-optimized at the ab initio level (MP2/6-31++G**//MP2/3-21+G**) (Figure 2). In ab initio calculation, the modification moiety of cmo⁵U_{enol} was taken in the conformation observed in the in the X-ray structure. In both U and cmo⁵U, the enol form is still unfavorable compared to their respective keto forms. In addition, the enol form in new wobble geometry is still less energetically favorable compared to that in Watson-Crick geometry since the latter has one more hydrogen bond. For U the energy difference is 11.06 kcal/mol, whereas for cmo⁵U the intramolecular hydrogen bond reduces this energy gap to 2.14 kcal/mol. In the optimized geometry, the carboxyl group of the oxyacetic acid is rotated of around 90° with respect to the X-ray structure and form an intramolecular hydrogen bond with the enol group. Relative $\Delta E_{\text{keto} \rightarrow \text{enol}}$ calculated by subtracting the average of geometry-optimized QM/MM energies B3LYP/6-31G* level from the same snapshots used for NBB free energy calculation also yield values in the same order of magnitude. For U in ribosome, this is 18.16 ± 11.16 , whereas for cmo⁵U, this is 5.72 ± 11.75 .

Calculations of free energy between the keto and enol tautomers showed that the $\Delta G_{\text{keto} \rightarrow \text{enol}}$ is almost the same for aqueous and ribosomal environments for modified U (Table 1). Since the uncertainties of $\Delta \Delta G_{\text{keto} \rightarrow \text{enol}}$ between aqueous and ribosome environments include zero, we conclude that the ribosome environment is not more favorable than the aqueous environment in keto-to-enol tautomerization for both U and cmo⁵U.

Table 1. Relative free energies of U and cmo⁵U.

Base (environment)	$\Delta\Delta G_{\text{keto} \rightarrow \text{enol}}^{(a)}$		
	BAR	QM-NBB direct	QM-NBB indirect
U	-1.88±0.55	-1.00±1.35	-3.08 ±0.58
cmo ⁵ U	-0.70±0.52	0.04±1.21	0.19 ±0.35

^(a) $\Delta\Delta G_{\text{keto} \rightarrow \text{enol}} = \Delta G_{\text{keto} \rightarrow \text{enol}}(\text{ribosome}) - \Delta G_{\text{keto} \rightarrow \text{enol}}(\text{aq})$. All values are in kcal/mol. Error is estimated by calculating s.e.m. of BAR/QM-NBB energies on 3 equal parts of divided trajectories.

Alternative hypotheses

MD simulations show that the enol hydroxyl H4 does not participate in hydrogen bonding, but is oriented away from the Watson-Crick edge. This is observed both in U and cmo⁵U. The unmodified U does not have a modification moiety that might possibly interact with H4, whereas the cmo⁵U modification moiety potentially can form an intramolecular hydrogen bond with H4, but this was not observed in the MD simulation. Instead, the oxyacetic acid group prefers to point away from the Watson-Crick edge towards the solvent environment.

The observation that the modification does not maintain the orientation it has in the X-ray structure was also mentioned in a previous simulation study²⁰. The simulations showed that the hydrogen bond between U33 O2' and the modification O5, as hypothesized by the X-ray study, is not enough to maintain the orientation. We note here that we assume that the modification acetic acid group is deprotonated, as is expected at neutral pH, but the local ribosomal environment might cause a pK_a shift, which could cause the acetic group to become protonated.

We also considered that the experimental orientation might be stabilized by interaction with entities not visible at this resolution, such as ions or other ligands. For instance, a recent QM study showed how Mg²⁺ coordination affects base/basepair charge distribution and geometries²⁷. As a proof of concept that an ion could influence the conformation of the modification, we simulated base pair cmo⁵U:G with an Mg²⁺ ion close by to the carboxyl oxygen and the guanine O6. After 5 ns, the modification moiety keeps a conformation close to what observed in X-ray structure (Figure S5).

We also note that the pK_a of uridine at the N3 protonation site is 9.3²⁸ and a pK_a of 8.54 is observed for N3 of 5-methoxy U^{6d}, thus the deprotonated U/cmo⁵U at N3 may also be a plausible alternative consistent with the X-ray electron density. We attempted to construct N3-deprotonated U/cmo⁵U by redistributing the atomic partial charges without the usual validation steps of force field parameterization, for a preliminary inquiry. However, we observed that N3-deprotonated U/cmo⁵U do not form Watson-Crick geometry, either, but 'new wobble' geometry, which is consistent with a previous ab initio study that investigated this particular form of U, among others^{6d}. Both the keto and the deprotonated forms deviate more from the crystallographic structure than the enol form (Figure S6–S7).

Conclusion

We examined the hypothesis that enol form of cmo⁵U is consistent with the observed Watson-Crick geometry in the crystal structure.

First we addressed whether the modification lower the keto-enol energy difference. Our ab initio calculations showed that in aqueous environment, the modification moiety of cmo⁵U do not affect keto-enol energy difference compared to that of U. The keto-enol energy difference is indeed lowered if intramolecular hydrogen bond is formed, but in this case the three hydrogen bonds of Watson-Crick geometry cannot be maintained.

To take into account the ribosomal environment, we sampled the basepair in the ribosomal decoding environment, as well as performed free energy calculations. Free energy calculation showed that the ribosomal environment does not lower keto-enol energy difference. Even when we used molecular mechanics description corrected with QM/MM calculations, the obtained free energy do not account for the contribution from electronic energy. The keto-enol equilibrium is affected by ± 1 kcal/mol in the ribosomal compared to aqueous environment. This energy gap is too small to influence the keto-enol electronic energy difference of more than 4 kcal/mol.

Unexpectedly, our molecular dynamics simulation showed that the enol form cannot maintain Watson-Crick geometry. Instead, it formed a 'deformed Watson-Crick' configuration with only two hydrogen bonds: the enol hydroxyl is not participating in the base pair interactions as expected from analogous C:G type Watson-Crick base pairing.

We thus conclude that enol form of cmo⁵U is not consistent with the observed Watson-Crick geometry in the crystal structure. We keep in mind the inherent limitations in computational techniques such as insufficient sampling and force field accuracy. Also, we have always assumed that cmo⁵U is in enol form and its modification deprotonated (conjugate base form) as the departing point. We did not consider, for example, that the modification is in its acid form, or that the G is in enol form instead.

Finally, we offered several hypotheses for future directions of inquiry such as deprotonated form of cmo⁵U and the presence of with chemical entities not resolved in the structure that may explain the position of the modification moiety like Mg²⁺.

Supporting Information. SupplementaryFigures.pdf: Figures S1-S6

Acknowledgment

This work was supported by Nanyang Technological University Research Scholarship (Y. D. H.), and the Swedish Research Council (VT 2015-04992)

References

1. (a) Wohlgemuth, I.; Pohl, C.; Rodnina, M. V., Optimization of speed and accuracy of decoding in translation. *The EMBO journal* **2010**, 29 (21), 3701-3709; (b) Kramer, E. B.; Farabaugh, P. J., The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **2007**, 13 (1), 87-96; (c) Johansson, M.; Zhang, J.; Ehrenberg, M., Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proceedings of the National Academy of Sciences* **2012**, 109 (1), 131-136.

2. (a) Ogle, J. M.; Brodersen, D. E.; Clemons, W. M.; Tarry, M. J.; Carter, A. P.; Ramakrishnan, V., Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **2001**, 292 (5518), 897-902; (b) Ogle, J. M.; Murphy, F. V.; Tarry, M. J.; Ramakrishnan, V., Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell* **2002**, 111, 721-732.
3. Singh, V.; Fedeles, B. I.; Essigmann, J. M., Role of tautomerism in RNA biochemistry. *RNA* **2014**, 21 (1), 1-13.
4. (a) Kimsey, I. J.; Petzold, K.; Sathyamoorthy, B.; Stein, Z. W.; Al-Hashimi, H. M., Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature* **2015**, 519 (7543), 315-320; (b) Szymanski, E. S.; Kimsey, I. J.; Al-Hashimi, H. M., Direct NMR Evidence that Transient Tautomeric and Anionic States in dG· dT Form Watson-Crick-like Base Pairs. *J. Am. Chem. Soc.* **2017**, 139 (12), 4326-4329.
5. (a) Singh, V.; Peng, C. S.; Li, D.; Mitra, K.; Silvestre, K. J.; Tokmakoff, A.; Essigmann, J. M., Direct observation of multiple tautomers of oxythiamine and their recognition by the thiamine pyrophosphate riboswitch. *ACS Chem. Biol.* **2013**, 9 (1), 227-236; (b) Berdakin, M.; Féraud, G.; Dedonder-Lardeux, C.; Jouvet, C.; Pino, G. A., Excited states of protonated DNA/RNA bases. *Phys. Chem. Chem. Phys.* **2014**, 16 (22), 10643-10650; (c) Rypniewski, W.; Banaszak, K.; Kuliński, T.; Kiliszek, A., Watson-Crick-like pairs in CCUG repeats: evidence for tautomeric shifts or protonation. *RNA* **2016**, 22 (1), 22-31.
6. (a) Maximoff, S. N.; Kamerlin, S. C.; Florian, J., DNA Polymerase λ Active Site Favors a Mutagenic Mismatch Between the Enol Form of Deoxyguanosine Triphosphate Substrate and the Keto Form of Thymidine Template. A Free Energy Perturbation Study. *The Journal of Physical Chemistry B* **2017**; (b) Satpati, P.; VAAqvist, J., Why base tautomerization does not cause errors in mRNA decoding on the ribosome. *Nucleic Acids Res.* **2014**, 42, 1-9; (c) Dabkowska, I.; Gutowski, M.; Rak, J., Interaction with glycine increases stability of a mutagenic tautomer of uracil. A density functional theory study. *J. Am. Chem. Soc.* **2005**, 127 (6), 2238-48; (d) Sochacka, E.; Lodyga-Chruscinska, E.; Pawlak, J.; Cypryk, M.; Bartos, P.; Ebenryter-Olbinska, K.; Leszczynska, G.; Nawrot, B., C5-substituents of uridines and 2-thiouridines present at the wobble position of tRNA determine the formation of their keto-enol or zwitterionic forms-a factor important for accuracy of reading of guanosine at the 3'-end of the mRNA codons. *Nucleic Acids Res.* **2017**, 1-12.
7. (a) Demeshkina, N.; Jenner, L.; Westhof, E.; Yusupov, M.; Yusupova, G., A new understanding of the decoding principle on the ribosome. *Nature* **2012**, 484 (7393), 256-259; (b) Rozov, A.; Demeshkina, N.; Westhof, E.; Yusupov, M.; Yusupova, G., Structural insights into the translational infidelity mechanism. *Nat. Commun.* **2015**, 6, 7251; (c) Weixlbaumer, A.; Murphy, F. V.; Dziergowska, A.; Malkiewicz, A.; Vendeix, F. a. P.; Agris, P. F.; Ramakrishnan, V., Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nat. Struct. Mol. Biol.* **2007**, 14 (6), 498-502; (d) Kurata, S.; Weixlbaumer, A.; Ohtsuki, T.; Shimazaki, T.; Wada, T.; Kirino, Y.; Takai, K.; Watanabe, K.; Ramakrishnan, V.; Suzuki, T., Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U· G wobble pairing during decoding. *J. Biol. Chem.* **2008**, 283 (27), 18801-18811.
8. Mitra, S. K.; Lustig, F.; Akesson, B.; Lagerkvist, U., Codon-anticodon recognition in the valine codon family. *J. Biol. Chem.* **1977**, 252 (2), 471-478.
9. Hillen, W.; Lindner, E. E.; Joerg, H.; Gassen, H. G.; Vorbrueggen, H., 5-Methoxyuridine: the influence of 5-substituents on the keto-enol tautomerism of the 4-carbonyl group. *Journal of Carbohydrates, Nucleosides, Nucleotides* **1978**, 5 (1), 23-32.
10. Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Others, CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.
11. (a) MacKerell, A. D.; Banavali, N. K., All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **2000**, 21 (2), 105-120; (b) Foloppe, N.; MacKerell, A. D. J., All-Atom Empirical Force Field for Nucleic Acids : I . Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, 21 (2), 86-104; (c)

- Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D. J., Impact of 2' -Hydroxyl Sampling on the Conformational Properties of RNA : Update of the CHARMM All-Atom Additive Force Field for RNA. *J. Comput. Chem.* **2011**, 32 (9), 1929-1943.
12. Xu, Y.; Nilsson, L.; Mackerell, A. D. J., An Additive CHARMM Force Field for Modified Nucleic Acids. *Biophys. J.* **2015**, 108 (2), 235a-236a.
 13. Roothaan, C. C. J., New developments in molecular orbital theory. *Rev. Mod. Phys.* **1951**, 23 (2), 69.
 14. Møller, C.; Plesset, M. S., Note on an approximation treatment for many-electron systems. *Physical Review* **1934**, 46 (7), 618-622.
 15. (a) Miertus, S.; Tomasi, J., Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.* **1982**, 65 (2), 239-245; (b) Miertuš, S.; Scrocco, E.; Tomasi, J., Electrostatic interaction of a solute with a continuum. A direct utilization of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, 55 (1), 117-129.
 16. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery Jr, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; others, Gaussian 03; Gaussian, Inc. Wallingford, CT **2004**, 6492.
 17. (a) Beglov, D.; Roux, B., Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations. *J. Chem. Phys.* **1994**, 100 (12), 9050-9063; (b) Allnér, O.; Nilsson, L.; Villa, A., Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, 8 (4), 1493-1502.
 18. Brooks III, C. L.; Karplus, M., Deformable stochastic boundaries in molecular dynamics. *J. Chem. Phys.* **1983**, 79 (12), 6312.
 19. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, 79 (2), 926-935.
 20. Allner, O.; Nilsson, L., Nucleotide modifications and tRNA anticodon-mRNA codon interactions on the ribosome. *RNA* **2011**, 17, 2177-2188.
 21. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, 23 (3), 327-341.
 22. Nilsson, L., Efficient Table Lookup Without Inverse Square Roots for Calculation of Pair Wise Atomic Interactions in Classical Simulations. *J. Comput. Chem.* **2008**, 30 (9), 1490-1498.
 23. (a) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, 13 (JANUARY 1992), 1011-1021; (b) Boczeko, E. M.; Brooks III, C. L., Constant-temperature free energy surfaces for physical and chemical processes. *J. Phys. Chem.* **1993**, 97 (17), 4509-4513.
 24. Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, 22, 245-268.
 25. König, G.; Hudson, P. S.; Boresch, S.; Woodcock, H. L., Multiscale free energy simulations: An efficient method for connecting classical MD simulations to QM or QM/MM free energies using non-Boltzmann Bennett reweighting schemes. *J. Chem. Theory Comput.* **2014**, 10, 1406-1419.
 26. (a) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuš, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; Woodcock, H. L.; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; DiStasio, R. A.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyayev, K.; Kim, J.; Kim,

J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Van Voorhis, T.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M., Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, 113 (2), 184-215; (b) Woodcock, H. L.; Hodošček, M.; Gilbert, A. T.; Gill, P. M.; Schaefer, H. F.; Brooks, B. R., Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations. *J. Comput. Chem.* **2007**, 28 (9), 1485-1502.

27. Halder, A.; Roy, R.; Bhattacharyya, D.; Mitra, A., How Does Mg²⁺ Modulate the RNA Folding Mechanism: A Case Study of the G: C⁺ W Trans Basepair. *Biophys. J.* **2017**.

28. Egli, M.; Saenger, W., Principles of nucleic acid structure. Springer Science & Business Media: 2013.

SUPPORTING INFORMATION

A COMPUTATIONAL STUDY OF URACIL TAUTOMERIC FORMS ON THE RIBOSOME: THE CASE OF URACIL AND 5-OXYACETIC ACID URACIL IN THE FIRST ANTICODON POSITION OF tRNA

Yossa Dwi Hartono^{1,2}, Mika Ito^{1,(a)}, Alessandra Villa¹ and Lennart Nilsson^{1*}

¹ Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden.

² Division of Structural Biology and Biochemistry, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

*To whom correspondence should be addressed. Tel: +46-8-52481099; Email: Lennart.Nilsson@ki.se

^(a)Current address Department of Anatomy, Physiology and Biochemistry, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

Figure S1. Atom numbering for keto and enol tautomers of U and cmo⁵U nucleobases.

Figure S2. Relative energies of $\text{cmo}^5\text{U}_{\text{keto}}$ and $\text{cmo}^5\text{U}_{\text{enol}}$ calculated at the HF/3-21+G level. All energies are relative to the energy of $\text{cmo}^5\text{U}_{\text{keto}}$ with $d2 = 0^\circ$.**

Figure S3. Dihedral scan used for parameterization. The QM potential energy scan is reproduced by the MM force field. A model system with a smaller modification moiety (5-methoxy) was used.

Figure S4. Water-bridged hydrogen bond probability between U_{enol} N3 and O4 atoms in umbrella sampling windows.

Figure S5. Simulation of uridine-guanine base pair with Mg²⁺ ion placed near modification moiety and guanine O6. Snapshot at 5 ns is shown.

A

B

Figure S6. Simulations of uridine-guanine base pair in the ribosome. Snapshots at 5 ns are shown. (A) Keto form (B) N3-deprotonated form.

Figure S7. Snapshots at 5 ns from simulations of the uridine-guanine base pair in the ribosome (green: X-ray, gray: keto, orange: enol, red: deprotonated). (A) U-G (B) cmo⁵U-G.

Adsorption and folding dynamics of MPER of HIV-1 gp41 in the presence of DPC micelle

Yossa Dwi Hartono, Yip Yew Mun, and Dawei Zhang*

Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ABSTRACT

Membrane-proximal ectodomain region (MPER) of HIV-1 gp41 is known to have several epitopes of monoclonal antibodies. It also plays an important role in the membrane fusion process that is well-evidenced, though not well-elucidated. There are also disputes over the true structure of MPER. In this study, MPER NMR structure in the presence of dodecylphosphatidylcholine micelle is used in the molecular dynamic simulation to elucidate structural dynamics and adsorption to model MPER interaction in a membrane environment. Polarized protein-specific charge derived from its NMR structure is found to better preserve the helical structure found in the NMR structure compared to AMBER03 calculation. The preserved helical structure also adsorb to the micelle using the hydrophobic side-chains, consistent to the NMR structure. *Ab initio* folding of MPER predicts a structure quite in well agreement with the NMR structure (RMSd 3.9 Å) and shows that the micelle plays a role in the folding process.

Proteins 2013; 81:933–944.
© 2013 Wiley Periodicals, Inc.

Key words: MPER adsorption; MPER folding; DPC micelle; AHBC; charge variation.

INTRODUCTION

To gain entry to its target cell, human immunodeficiency virus Type 1 (HIV-1) uses its envelope glycoproteins (Env) to facilitate membrane fusion. Env is produced as glycosylated gp160 precursor, which is processed into two non-covalently associated subunits. gp120 would initially bind to its major cellular receptor, CD4, inducing conformational changes which exposes gp41.¹ Subsequently, gp41 undergoes a series of conformational changes. Membrane fusion is achieved when the N- and C-heptad repeats (NHR, CHR) form trimers-of-hairpin configuration, also known as six-helix bundle (6HB), the formation of which releases free energy that drives membrane fusion.²

During such sequence of conformational change of gp41, several targets for antibodies and anti-HIV drugs are accessible. Much has been said about the highly conserved hydrophobic binding pocket on NHR, targeted by antibodies D5, DN9, 8K8, and HK20,^{3–5} as well as Roche's fusion inhibitor enfuvirtide,⁶ human lysozyme, and the latter's peptide derivatives HL18 and HL9.^{7,8} Another region of gp41, the Trp-rich membrane-proximal ectodomain region (MPER), has been found to elicit responses from monoclonal antibodies 2F5, 4E10, and Z13.^{9–11} MPER is located at the base of gp41, proximal to the viral membrane and gp41 transmembrane region,

hence its name.¹² MPER is an attractive drug target not only because of the several epitopes situated there, but MPER sequence is also known to be very conserved.^{9,10,13} It is also a crucial part of the viral membrane fusion machinery, for which there are several evidences: it contributes to the activity of fusion inhibitor T-20 (synonym of enfuvirtide)^{14,15}; fusion is affected by deletion or substitution in MPER sequence^{16,17–20}; it is possibly involved in gp41 oligomerization and location in the viral membrane²¹; and peptides generated from MPER sequence associate with lipid membrane and cause membrane leakage^{22–24}—the very last point is especially an incentive to study MPER together with membrane dynamics. Though we know that MPER is crucial in the membrane fusion process, it is still not clear how exactly so. Consequently, structural and mechanism studies are paramount to contribute to our understanding of the role of MPER in the context of membrane fusion.

Grant sponsor: Nanyang Technological University AcRF Tier 1; Grant number: RG58/11

*Correspondence to: Dawei Zhang, Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore. E-mail: zhangdw@ntu.edu.sg

Received 18 August 2012; Revised 7 December 2012; Accepted 4 January 2013
Published online 23 January 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24256

MPER has a high number of hydrophobic residues, conferring poor solubility in aqueous solution and presenting a challenge for structural studies. For instance, an X-ray crystal structure of gp41 (PDB 1AIK²⁵) excludes MPER, presumably because the inclusion would not favor crystallization conditions. However, recent structural studies do manage to include MPER, though typically the part of MPER nearing C-terminus is designated as coil or not visible in the crystal structure.^{26,27}

Several crystallographic and NMR studies maintained that MPER is an α -helix.^{25,28,29} However, others found that MPER contains a β -turn instead.^{30,31} These diverse findings, complemented with low immunogenicity of MPER, may suggest that MPER undergoes structural change during fusion and the different structures are merely snapshots at different stages of viral entry. It was also found that antibodies 4E10 and Z13, though having overlapping epitopes on MPER, bind to different faces of it.³² While this presents more possibilities in terms of epitopic target, it also shows that MPER may adopt different conformations and orientations during fusion, and different parts are exposed transiently, presenting challenge in designing fusion inhibitors targeting MPER.

There are two parts in this molecular dynamics (MD) study: first is using the known structure of MPER with good resolution (PDB 2PV6),³³ putting it in various environments to study its unfolding and adsorption to the membrane environment; second is *ab initio* folding of MPER in the presence of the membrane environment to study the folding dynamics and the role of the membrane environment in folding.

For the first part, we used near-atomic resolution NMR structure in which MPER is found to be L-shaped—it consists of two helical segments joined by a central hinge.³³ The flexibility of this conformation, with backbone RMSd of 0.59 Å, is mostly attributed to the central hinge, since N- and C-terminal segments converged well when respectively aligned (backbone RMSd of 0.24 Å and 0.15 Å respectively). This L-shaped NMR structure was used as the starting structure for MD production. In the NMR study, dodecylphosphatidylcholine (DPC) was added to solubilize MPER; therefore we also added DPC micelle to the system, using the micelle to model the membrane environment as well.

MPER is positioned at a distance to study the mechanism of MPER adsorption to DPC micelle. Two different atomic charges (elaborated below) are used to study how electronic polarization affects helix stability in the presence of micelle [Fig. 1(A)]. Two orientations of MPER are used to study how different orientations of side-chains, with respect to the micelle surface, affect peptide adsorption and helix stability [Fig. 1(B)]. MD productions were also performed in the absence of DPC micelle, henceforth referred to as DPC(–), to study the stability of the peptide in purely aqueous environment. In the L-shaped NMR structure, hydrophobic side-chains W666,

W670, W672, F673, and W678 (the first three Trp are conserved residues) are buried at the micelle surface, while hydrophilic side-chains N671, N674, and N677 are exposed to the solvent.³³ Based on this, two different orientations are used: in Mode 1, the five hydrophobic side-chains are poised toward the micelle surface; while in Mode 2, they are poised away.

MD simulations of native MPER on the micelle surface will be carried out based on standard AMBER force field and newly developed polarized protein-specific charge (PPC) model respectively to compare the performance of protein dynamics on the highly polar micelle surface. PPC model is a recent attempt to more accurately model electrostatic interaction in protein.^{34,35} Standard force fields (FF) are amino-acid specific,³⁶ thus less adequately describing electrostatic interactions in protein-specific environments. Distinct from AMBER charge, PPC is derived from quantum mechanical calculation of protein in solution by using molecular fractionation with conjugate caps (MFCC) approach³⁶ incorporating the Poisson-Boltzmann (PB) equation. Solution of the PB equation is carried out first to get the induced charges on protein–solvent interface. Then the DFT calculations are performed with the induced charges taken as background charges to mimic the polarization effect from the solvent molecules. New atomic charges which are fitted to the electrostatic potential from DFT calculations will generate new induced surface charges. This self-consistent procedure will be applied until convergence is reached. The scheme together with on-the-fly charge fitting in other studies has met some success in showing secondary structure stability,^{37,38} protein folding,^{39–42} and modeling protein–ligand binding^{43,44} more accurately compared to standard FF.

In the second part of this study, we construct a typical *ab initio* folding setup, starting from the linear structure of MPER, using MD production to observe the folding process and compare the stabilized, folded protein to the NMR structure. Similar to the earlier part, DPC micelle is included in the system in order to investigate the role of membrane environment in the folding process. Following the result from earlier part where PPC scheme preserves helical structure of MPER better than AMBER03, MD production is done with PPC calculation of atomic charges of hydrogen bond donors and acceptors updated and refitted on the fly during the course of the simulation to reduce computational expense. This scheme, termed adaptive hydrogen bond charge (AHBC) scheme,⁴² has been successfully applied to study the effects of sequence mutation and the influence of environment on protein folding.^{39,40}

MATERIALS AND METHODS

The best representative conformer of MPER of gp41 NMR structure ensemble was obtained from Protein Data

Figure 1

A: Comparison of charge distribution of MPER between AMBER03 and PPC, only backbone charge is shown in electrostatic potential surface. **B:** Comparison of orientations of sidechains between Modes 1 and 2. MPER is shown in ribbon representation with Trp and Phe sidechains, DPC micelle is shown with phosphate group on the surface.

Bank (PDB ID: 2PV6).³³ DPC structure and micelle assembly was built using MAESTRO 9.2 (Schrödinger LLC). MPER linear structure was built using LEaP module of AMBER 9.0 program.⁴⁵ For Modes 1 and 2, MPER is put 10 Å away from micelle surface [Fig. 1(B)]. PMEMD module of AMBER 9.0 was used for MD production with ff03 force field^{46,47} complemented with GAFF and DPC parameter sets. TIP3P water molecules⁴⁸ were added with the minimal distance between the complex and the edge of water box set as 10 Å. No counterions were added as the system is neutral. Non-bonded cutoff was set to 12.0 Å. SHAKE algorithm⁴⁹ was applied with all bonds involving hydrogen atoms restrained. After minimization and heating up from 10 to 308 K (corresponding to the tem-

perature in the NMR study³³) in 100 ps, production MD was carried out at 308 K, with time-step of 2.0 fs for 100 ns period for Modes 1 and 2, 20,000 ps period for DPC(–), 60,000 ps period for *de novo* folding, with complex conformation collected every 1 ps. Starting atomic charges are calculated using AMBER03 FF and PPC model from the NMR structure [Fig. 1(A)]. In PPC scheme, linearized PB method is used to solve self-consistent reaction-field equation and MFCC is used in the quantum chemistry calculation to obtain electron densities. RESP method is used to fit the atomic partial charges. In this study, the calculated set of charges is used to replace the standard FF charges only in the starting structure. For AHBC scheme, hydrogen bonds were located every 5 ps using HBPlus.

For post-processing analysis, water molecules and periodic boundary information were removed from the trajectory. RMSd, root-mean-square atomic fluctuation (RMSf), distance, and hydrogen bonding analysis were carried out using ptraj module. Hydrogen bonding distance and angle cutoffs are 3.5 Å and 120°, respectively.

Electrostatic surface potential was visualized by MAESTRO 9.2 (Schrödinger LLC). All other simulation images were generated using visual molecular dynamics (VMD).⁵⁰

RESULTS AND DISCUSSIONS

Unfolding and adsorption

The differences in the orientations of Modes 1 and 2 as well as atomic charge distribution between standard AMBER03 FF (henceforth AMBER) and PPC are depicted in Figure 1. Figure 1(A) depicts how AMBER03 and PPC assign atomic charges differently. It is apparent that they are similar in assigning charges to terminal residues, but PPC calculates more pronounced intensity of polar regions in the middle of MPER, as is expected from backbone atoms. This is observable, for example, in the negative region (colored red). Accurate electrostatic distribution model is especially important in this study because α -helical structure is held by backbone hydrogen bonds, which are electrostatic in nature.

The backbone RMSd plots show that MPER stabilize within the first 20 ns of MD production (Fig. 2). Also shown in Figure 2 is the backbone RMSf by residue. The terminal residues, as expected, undergo the most fluctuations, having fewer stabilizing interactions compared to inner residues. The spike in fluctuations around residues 11–15 are due to the flexibility of the hinge region, which

is not part of the network of backbone hydrogen bonds that form the helical structure.

From RMSd plots, it is apparent that those with PPC charge tend to have less amplitude of vacillation for backbone deviation, while those with AMBER charge are marked with occasional jump and dip. Residue-wise, RMSf plots show that Mode 1 PPC fluctuates slightly more than Mode 1 AMBER, while Mode 2 PPC undergoes much less fluctuation compared to Mode 2 AMBER. Mode 2 PPC is especially of note since the fluctuations are minimal. In the absence of DPC micelle, MPER with AMBER charge show more fluctuation, while MPER with PPC charge show less, though not as low as Mode 2 PPC.

In the starting (NMR) structure, the backbone hydrogen bonds that support the α -helical structure consist of $i-i+4$ and $i-i+3$, which are shown in Figure 3. Over the course of the simulation, generally those belonging to the N-helix, mostly $i-i+4$, are preserved; while those belonging to the C-helix, all are $i-i+3$, especially those nearer to the hinge at the middle, are not preserved. This is apparent from the occupancy and distance plots (Fig. 3), where those of N-helix are mostly of high occupancy and short bond length, while those of C-helix are of low occupancy and long bond length, except the two nearest to the C-terminus. 12O–15H and 13O–16H, near the hinge, are almost always broken first within the first 10 ns.

Mode 1 AMBER follows the above general trend, except that its 2O–5H hydrogen bond is broken at around 51 ns, and 2O–6H is formed instead. It is worth noting that the 3O–7H bond length transiently lengthens around the same time, as the helix readjusts its twist. Mode 1 PPC follows the general trend, but the breaking of 12O–15H and 13O–16H near the hinge occurs faster than in Mode 1 AMBER.

Figure 2

Root-mean-square deviation (RMSd) and root-mean-square by-residue positional fluctuation (RMSf) of MPER backbone. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 3

Backbone hydrogen bonding occupancy. Those present in the NMR structure are highlighted in black ($i-i+4$) and blue ($i-i+3$), and the bond distance over the simulation is shown. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Mode 2 AMBER follows the general trend, but there are occasional spikes in bond length: 5O—9H at 30–40 ns, 4O—8H and 3O—7H at 70–80 ns, though none of these are permanent as subsequently the bonds shorten and stabilize to their previous values. Mode 2 PPC deviates from the general trend in that almost all of the C-helix $i-i+3$ backbone hydrogen bonds that are initially present are not preserved over the course of the simulation. From the

occupancy graph, it is apparent that it forms other high-occupancy $i-i+3/i-i+4$ hydrogen bonds instead: 9O—13H, 10O—13H, 10O—14H, 11O—15H, 14O—17H, and 14O—18H. The presence of these high-occupancy hydrogen bonds may explain the low positional fluctuation of this particular scheme. It is also worth noting that these hydrogen bonds are not formed in the other schemes, or at least if they are, the occupancy values are not as high.

In the absence of DPC micelle, DPC(-) AMBER follows the general trend and additionally 7O—12H near the hinge and belonging to the N-helix is broken. The other N-helix hydrogen bonds undergo constant breaking and reforming—this may be due to more accessibility of water molecules to attack the hydrogen bonds since there is no DPC micelle for MPER to adsorb to. DPC(-) PPC has noticeable deviations from the general trend: 7O—11H and 8O—12H belonging to the N-helix and near the hinge are broken, while in Modes 1 and 2 they are preserved. Other hydrogen bonds near both termini are generally stable.

In Modes 1 and 2, MPER adsorbs to DPC micelle within the first 25 ns of the simulation (Fig. 4). The average micellar radius is 20.3 Å (from trimethylammonium group of a DPC molecule to the micelle center of mass), so this value is used as a reference point for the micelle surface. MPERs in Mode 1 are faster to approach the micelle than Mode 2; and MPERs in Mode 2 move away from the micelle before approaching it. This is expected, since in Mode 1 the hydrophobic sidechains known to adsorb to the micelle are already positioned toward the micelle, while in Mode 2 they are away, thus MPER reorients itself before adsorption. The initial increase in MPER-micelle distance is also observed in the initial stage of Mode 1 PPC, though the increase is only about 5 Å, compared to about 10 Å in Mode 2 (both AMBER and PPC).

Figure 5 depicts the system in various snapshots during adsorption. In Mode 1 AMBER, uncoiling of N-helical region nearest to the hinge and the whole of C-helix are already observed as early as 80 ps. At 3120 ps, the N-terminal helix is already close to the micelle, while the C-

terminal uncoiled region is pointed away. At 6000 ps, the C-region is already adsorbed on the micelle surface. MPER continues to embed itself deeper in the micelle until about 18,000 ps. Afterwards, MPER-micelle distance fluctuates as the C-helix further uncoils. The snapshot at 20,000 ps shows how the C-region is more stretched out and adopts random coil conformation.

Similar mechanism is observed in Mode 1 PPC: the N-helix approaches the micelle first. Noticeably, full adsorption of MPER takes longer in Mode 1 PPC compared to Mode 1 AMBER—from the distance plot (Fig. 4), this is around 30 ns in PPC compared to about 10 ns in AMBER. The C-region also appears to move away from the micelle surface, resulting in the initial increase of MPER-micelle distance. The helical region nearest to the C terminus and that of N-region near the hinge is preserved in PPC, but not in AMBER. From the hydrogen bond analysis (Fig. 3), we know that the hydrogen bonds supporting these helical regions are 7O—11H, 8O—12H, 17O—20H, and 18O—21H.

For Mode 2 AMBER, initially the N-helix approaches the micelle first, as observed at 1000 ps. However, at 2390 ps, the hinge is actually closest to the micelle while the two termini are away, leading to the observation that MPER-micelle distance increases initially (Fig. 4). MPER continues to bury itself deeper in the micelle, as observed at 10,00 ps. At 16,300 ps, the N-helix starts to close in to the micelle surface, while the C-region is still pointing away; and this is still observed until 25,000 ps. Distance plot indeed shows that Mode 2 AMBER reaches stable distance only at around 30 ns.

Mode 2 PPC adsorption does not follow the same stages as other schemes. At 350 ps, it is observed that the hinge region becomes helical, while the whole C-region uncoils. At 2600 ps, the helix extends itself straight, which causes the C-terminus to point away from the micelle. At 7900 ps, the initially helical region nearest to the C-terminus loses its characteristic helical turn and stretches to random coil conformation. The C-terminal gradually closes in to the micelle as well and by 20,000 ps, the helix is approximately tangent to the micelle surface. Subsequently, MPER maintains the distance and does not adsorb as deep as Mode 2 AMBER, but approximately just below the micelle surface. This shallow adsorption is also observed in Mode 1 PPC, although it is more prominent in Mode 2 PPC. Immersion depth notwithstanding, similar sidechains are used to adsorb to the micelle, particularly those on the N-helix. This can be observed in Figure 6, shown with Trp and Phe residues—these hydrophobic sidechains point into the micelle.

Since MPER is known to cause membrane leakage,^{22–24} it is expected for MPER to disrupt or at least deform the micellar sphere. This is indeed observed, for instance refer to Figure 5 Mode 2 AMBER at 10,000 ps; however the deformation is transient and not significantly disruptive of the micelle. There may be several factors contributing to

Figure 4

The distance of MPER to DPC micelle (between centers of mass). The average DPC micelle radius, 20.3 Å, is indicated. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 5

The mechanism of MPER adsorption to DPC micelle over the first 25 ns of MD production. N-terminus is always positioned at the left. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 6

Last frames of MD simulations. Phe and Trp residues of MPER are shown. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

this. First is the limitation of DPC micelle as a model for cellular membrane. The micelle in this case does not accurately model viral/cell membrane; small aggregation like the micelle may be more resistant to deformation compared to large aggregation like the phospholipid bilayer. The cell membrane itself has a variety of membrane proteins, cholesterol, and others, that affect the fluidity of the membrane. MPER may also disrupt the membrane synergistically, either with another MPER or with another part of the fusion machinery.

Observing the last frames of the simulation in Figure 6, it is apparent that PPC schemes generally preserve the helical structure more compared to the respective AMBER schemes. This observation is important, especially when the micelle is present, because it is already known that MPER forms two-region helix in the presence of DPC micelle (from NMR structure used as the starting structure in all of the MD simulations). Mode 1 PPC preserves the helical structure reasonably well, though a part of the C-helix near the hinge uncoils; while for the corresponding AMBER scheme, the whole C-region uncoils. Mode 2 PPC is an anomaly, because the hinge region becomes helical instead, while the rest of the C-helix is uncoiled. In the absence of the micelle, MPER in AMBER scheme uncoils at C-region near the hinge; while in PPC, the helical regions are generally preserved, but the L-shape takes a more acute angle. PPC schemes, then, are generally successful in these MD productions. One caveat is that, ideally, the process of MPER folding, MPER adsorption, and micellar self-

assembly should occur simultaneously. In this study, we assume that the kinetics of micellar self-assembly is much faster than the other two processes. Another observation worthy of note is that helix uncoiling always begins from the hinge region and usually propagates toward the C-terminus. This may indicate that MPER folding process may begin from the termini toward the hinge, that is, the reverse direction of the uncoiling, by the principle of microscopic reversibility.

It is worth reiterating that the hinge seems to play an important role in the dynamics of MPER, both for adsorption and uncoiling. It serves as nucleation site of uncoiling, as mentioned; and its flexibility seems to confer certain degree of freedom for the C-helix—possibly this helps with uncoiling process, and there may be a function served in the fusion process to have the C-region flailing around while the N-helix anchors MPER to the membrane.

De novo folding and adsorption

MPER *de novo* folding from the linear structure in the presence of DPC was attempted. Only PPC scheme was used since it was shown that it preserves MPER helical structure better and it has met success particularly in secondary structure stability studies.^{37,38}

RMSd and adsorption distance (Fig. 7) show that MPER stabilizes around 32 ns, though dictionary of secondary structure of protein (DSSP) plot shows that from 35 to 60 ns the structure of the helix nearer to the N-terminus

Figure 7

Folding snapshots, RMSd, adsorption distance, and secondary structure (DSSP) of MPER. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(N-helix) alternates between 3-, 4- and 5-helices. Folded MPER reaches the RMSd of 3.9 Å at 60 ns and the lowest RMSd of 3.1 Å at 41,226 ps, both referenced to the NMR structure. Low RMSd fluctuation coincides with low adsorption-distance fluctuation (after 32 ns) and secondary structure is also stable when MPER has adsorbed to the micelle surface (after 27 ns), both of which show that micelle adsorption contributes to structural stability of MPER, possibly by limiting conformational space and forming hydrophobic interaction, which also exclude water molecules from interacting MPER face buried in the micelle surface. The orientation of MPER also largely matches the NMR structure in that out of W666, W670, W672, F673, and W678 (buried sidechains in NMR structure), all Trp sidechains are buried in the micelle, though not Phe—this is most observable in the snapshot at 31,806 ps (Fig. 7).

MPER interacts with DPC micelle even before full adsorption occurs, between its hydrophobic sidechains and DPC hydrophobic tail. This is observed with W670 at 804 ps, W672 at 2197 ps, and F673 at 5266 ps. Noticeably these residues belong to the hinge region, which is known to be flexible, as is also evident by the lack of secondary structure at this region. Whether the flexibility causes the sidechains to have higher probability of making contacts with the micelle or—the other way round—the hydrophobic interactions with the fluid micelle causes the flexibility; this remains to be concluded. These transient interactions are also noteworthy because it might aid folding by directing most of the hydrophobic sidechains to one face, for instance.

Consistent with the NMR structure, folded MPER has two helical segments. C-helix forms first transiently at 368 ps, then stabilizes after 1066 ps, while N-helix forms transiently at 938 ps and stabilizes after 8990 ps. After full adsorption at around 33,000 ps, C-helix stabilizes: the region nearer to the hinge forms 3-helix; the region nearer to the N-terminus, 4-helix. N-helix, on the other hand, fluctuates between 3-, 4-, and 5-helix. The last frame (Fig. 8) deviates from the NMR structure with an RMSd of 3.9 Å.

Hydrogen bonding occupancy plot (Fig. 8) reveals that, although NMR structure and folded MPER both consist of two helical segments, there is substantial difference between the types of helix formed. While most of the hydrogen bonds present in the NMR structure are also present with high occupancies in the folding simulation, some are completely absent, like 5O—9H, 7O—11H, and 8O—12H. The hinge region in the simulation is consistent with the NMR structure at around the middle of MPER, as indicated by the low-occupancy hydrogen bonds, though the hinge region in the simulation covers larger region than that in the NMR structure, as is also seen in the ribbon representations where the last frame helix seems to unravel more at the middle region (Fig. 8). Both helical segments appear stable, with

N-helix perhaps slightly less so since it alternates between several types of helices.

Limitations

It is then tempting to readily conclude that MPER would play a significant role in associating with the membrane and anchor the ectodomain fusion machinery to the membrane. Considering that there already exists a strong anchor, the transmembrane region, MPER anchoring role would be more transient and dynamic, which would be suitable for the significant conformational changes during fusion.

There are a few caveats that need to be taken into consideration. First, the membrane model that is used in this study is a micelle, which is certainly different from the physiological membrane bilayer in terms of composition, geometry, and curvature. The MPER peptide is a short peptide, while in reality, gp41, MPER, and the transmembrane region are continuous polypeptide. The flexibility and dynamics observed in the NMR experiment and MD may not be observed when MPER are chained at both ends with the two flanking regions. Also arising from this, the transmembrane region that extends into the membrane may also affect the membrane curvature and/or geometry, in which case the MPER/membrane interactions may look radically different.

Although our result in this study is preliminary, it opens up future research inquiries which hopefully will overcome the current limitations in this study toward understanding the mechanism of fusion.

CONCLUSION

In this study, MD simulations of the adsorption of MPER of HIV-1 gp41 on micelle surface are investigated. There are two charge schemes, AMBER and PPC; two orientations of sidechains; and with or without DPC micelle; totalling to six schemes with the NMR structure as the starting frame. Through the simulations, we found that PPC schemes preserve the helical structure better, with or without the micelle. The hinge region, known to confer flexibility to the MPER structure, serves as initial site where hydrogen bonds break, leading to helix uncoiling. Uncoiling occurs mostly on the C-region. MPER adsorbs to the micelle surface in all schemes. Generally, the N-region adsorbs first, followed by the C-region. Under AMBER schemes, MPER immerse deeper into the micelle. No significant micelle deformation was observed in the process of MPER adsorption and after.

Additionally, MPER is also folded *de novo* only using PPC scheme, implemented as on-the-fly charge update. The *de novo* folding dynamics of MPER in the presence of DPC micelle reveals folded structure mostly in agreement with the NMR structure. However, the deviation

Figure 8

Hydrogen bonding occupancy of *de novo* MPER folding. □: present in NMR structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of the hinge region is larger and the N-helix appears to be less stable, contrary to the earlier findings where N-helix is more stable during adsorption. The DPC micelle was observed to play a role in aiding the folding process; which warrants further studies, since folding

simulations of non-soluble peptides are still scarce in literature.

Overall, we have shown the importance of PPC charge update in folding simulation and possibly a glimpse of how membrane-aided folding may look like.

ACKNOWLEDGMENTS

D.W.Z. thanks NTU HPC support and resources.

REFERENCES

- Feng F, Broder CC, Kennedy PE, Berger EA. HIV-1 entry co-factor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* 1996;272:872–877.
- Melikyan GB, Markosyan RM, Hemmati H, Delmedico MK, Lambert DM, Cohen FS. Evidence that the transition of HIV-1 gp41 into a six-helix bundle, not the bundle configuration, induces membrane fusion. *J Cell Biol* 2000;151:413–423.
- Sabin C, Corti D, Buzon V, Seaman MS, Lutje Hulsik D, Hinz A, Vanzetta F, Agatic G, Silacci C, Mainetti L, Scarlatti G, Sallusto F, Weiss R, Lanzavecchia A, Weissenhorn W. Crystal structure and size-dependent neutralization properties of HK20, a human monoclonal antibody binding to the highly conserved heptad repeat 1 of gp41. *PLoS Pathog* 2010;6:e1001195.
- Luftig MA, Mattu M, Di Giovine P, Geleziunas R, Hrin R, Barbato G, Bianchi E, Miller MD, Pessi A, Carfi A. Structural basis for HIV-1 neutralization by a gp41 fusion intermediate-directed antibody. *Nat Struct Mol Biol* 2006;13:740–747.
- Hartono YD, Lazim R, Yip YM, Zhang D. Computational study of bindings of HK20 Fab and D5 Fab to HIV-1 gp41. *Bioorg Med Chem Lett* 2012;22:1695–1700.
- Liu S, Wu S, Jiang S. HIV entry inhibitors targeting gp41: from polypeptides to small-molecule compounds. *Curr Pharm Des* 2007;13:143–162.
- Lee-Huang S, Maiorov V, Huang PL, Ng A, Lee HC, Chang Y, Kalenbach N, Huang PL, Chen H. Structural and functional modeling of human lysozyme reveals a unique nonapeptide, HL9, with anti-HIV activity. *Biochemistry* 2005;44:4648–4655.
- Hartono YD, Lee AN, Lee-Huang S, Zhang D. Computational study of bindings of HL9, a nonapeptide fragment of human lysozyme, to HIV-1 fusion protein gp41. *Bioorg Med Chem Lett* 2011;21:1607–1611.
- Muster T, Steindl F, Purtscher M, Trkola A, Klima A, Himmler G, Rucker F, Katinger H. A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J Virol* 2011;67:6642–6647.
- Zwick MB, Labrijn AF, Wang M, Spencehauer C, Saphire EO, Binley JM, Moore JP, Stiegler G, Katinger H, Burton DR, Parren PW. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J Virol* 2001;75:10892–10905.
- Zwick MB, Wang M, Poignard P, Stiegler G, Katinger H, Burton DR, Parren PW. Neutralization synergy of human immunodeficiency virus type 1 primary isolates by cocktails of broadly neutralizing antibodies. *J Virol* 2001;75:12198–12208.
- Montero M, van Houten NE, Wang X, Scott JK. The membrane-proximal external region of the human immunodeficiency virus type 1 envelope: dominant site of antibody neutralization and target for vaccine design. *Microbiol Mol Biol Rev* 2008;72:54.
- Purtscher M, Trkola A, Gruber G, Buchacher A, Predl R, Steindl F, Tauer C, Berger R, Barrett N, Jungbauer A, Katinger H. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. *AIDS Res Hum Retrovir* 1994;10:1651–1658.
- Kilby JM, Hopkins S, Venetta TM, DiMassimo B, Cloud GA, Lee JY, Alldredge L, Hunter E, Lambert D, Bolognesi D, Matthews T, Johnson MR, Nowak MA, Shaw GM, Saag MS. Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nat Med* 1998;4:1302–1307.
- Lawless MK, Barney S, Guthrie KI, Bucy TB, Petteway SR, Jr, Merutka G. HIV-1 membrane fusion mechanism: structural studies of the interactions between biologically-active peptides from gp41. *Biochemistry* 1996;35:13697–13708.
- Salzwedel K, West JT, Hunter E. A conserved tryptophan-rich motif in the membrane-proximal region of the human immunodeficiency virus type 1 gp41 ectodomain is important for Env-mediated fusion and virus infectivity. *J Virol* 1999;73:2469–2480.
- Bellamy-McIntyre AK, Lay CS, Bar S, Maerz AL, Talbo GH, Drummer HE, Pombourios P. Functional links between the fusion peptide-proximal polar segment and membrane-proximal region of human immunodeficiency virus type 1 gp41 in distinct phases of membrane fusion. *J Biol Chem* 2007;282:23104–23116.
- Munoz-Barroso I, Salzwedel K, Hunter E, Blumenthal R. Role of the membrane-proximal domain in the initial stages of human immunodeficiency virus type 1 envelope glycoprotein-mediated membrane fusion. *J Virol* 1999;73:6089–6092.
- Pombourios P, Ahmar W, McPhee DA, Kemp BE. Determinants of human immunodeficiency virus type 1 envelope glycoprotein oligomeric structure. *J Virol* 1995;69:1209–1218.
- Zwick MB, Jensen R, Church S, Wang M, Stiegler G, Kunert R, Katinger H, Burton DR. Anti-human immunodeficiency virus type 1 (HIV-1) antibodies 2F5 and 4E10 require surprisingly few crucial residues in the membrane-proximal external region of glycoprotein gp41 to neutralize HIV-1. *J Virol* 2005;79:1252–1261.
- Saez-Cirion A, Gomara MJ, Agirre A, Goni FM, Nieva JL. Pre-transmembrane sequence of Ebola glycoprotein. Interfacial hydrophobicity distribution and interaction with membranes. *FEBS Lett* 2003;533:47–53.
- Moreno MR, Pascual R, Villalain J. Identification of membrane-active regions of the HIV-1 envelope glycoprotein gp41 using a 15-mer gp41-peptide scan. *Biochim Biophys Acta* 2004;1661:97–105.
- Suarez T, Gallaher WR, Agirre A, Goni FM, Nieva JL. Membrane interface-interacting sequences within the ectodomain of the human immunodeficiency virus type 1 envelope glycoprotein: putative role during viral fusion. *J Virol* 2000;74:8038–8047.
- Suarez T, Nir S, Goni FM, Saez-Cirion A, Nieva JL. The pre-transmembrane region of the human immunodeficiency virus type-1 glycoprotein: a novel fusogenic sequence. *FEBS Lett* 2000;477:145–149.
- Chan DC, Fass D, Berger JM, Kim PS. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* 1997;89:263–273.
- Buzon V, Natrajan G, Schibli D, Campelo F, Kozlov MM, Weissenhorn W. Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. *PLoS Pathog* 2010;6:e1000880.
- Shi W, Bohon J, Han DP, Habte H, Qin Y, Cho MW, Chance MR. Structural characterization of HIV gp41 with the membrane-proximal external region. *J Biol Chem* 2010;285:24290–24299.
- Tan K, Liu J, Wang J, Shen S, Liu M. Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proc Natl Acad Sci USA* 1997;94:12303–12308.
- Weissenhorn W, Dessen A, Harrison SC, Skehel JJ, Wiley DC. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* 1997;387:426–430.
- Ofek G, Tang M, Sambor A, Katinger H, Mascola JR, Wyatt R, Kwong PD. Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. *J Virol* 2004;78:10724–10737.
- Pai EF, Klein MH, Chong P, Pedyczak A. Fab-epitope complex from the HIV-1 cross-neutralizing monoclonal antibody 2F5. US patent WO-00/61618; 2000.
- Nelson JD, Brunel FM, Jensen R, Crooks ET, Cardoso RM, Wang M, Hessel A, Wilson IA, Binley JM, Dawson PE, Burton DR, Zwick MB. An affinity-enhanced neutralizing antibody against the membrane-proximal external region of human immunodeficiency virus type 1 gp41 recognizes an epitope between those of 2F5 and 4E10. *J Virol* 2007;81:4033–4043.
- Sun ZY, Oh KJ, Kim M, Yu J, Brusica V, Song L, Qiao Z, Wang JH, Wagner G, Reinherz EL. HIV-1 Broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity* 2008;28:52–63.

34. Ji CG, Mei Y, Zhang JZH. Developing polarized protein-specific charges for protein dynamics: MD free energy calculation of pK_a shifts for Asp²⁶/Asp²⁰ in Thioredoxin. *Biophys J* 2008;95:1080.
35. Mei Y, Ji CG, Zhang JZH. A new quantum method for electrostatic solvation energy of protein. *J Chem Phys* 2006;125:94906.
36. Zhang DW, Zhang JZH. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy. *J Chem Phys* 2003;119:3599.
37. Duan LL, Mei Y, Zhang QG, Zhang JZH. Intra-protein hydrogen bonding is dynamically stabilized by electronic polarization. *J Chem Phys* 2009;130:115102.
38. Lu Y, Mei Y, Zhang JZH, Zhang D. Communications: electron polarization critically stabilizes the Mg²⁺ complex in the catalytic core domain of HIV-1 integrase. *J Chem Phys* 2010;132:131101.
39. Wei C, Tung D, Yip YM, Mei Y, Zhang D. Communication: the electrostatic polarization is essential to differentiate the helical propensity in polyaniline mutants. *J Chem Phys* 2011;134:171101.
40. Xu Z, Lazim R, Sun T, Mei Y, Zhang D. Solvent effect on the folding dynamics and structure of E6-associated protein characterized from ab initio protein folding simulations. *J Chem Phys* 2012;136:135102.
41. Mei Y, Wei C, Yip YM, Ho CY, Zhang JZH, Zhang D. Folding and thermodynamic studies of Trp-cage based on polarized force field. *Theor Chem Acc* 2012;131:1168.
42. Duan LL, Mei Y, Zhang D, Zhang QG, Zhang JZH. Folding of a helix at room temperature is critically aided by electrostatic polarization of intraprotein hydrogen bonds. *J Am Chem Soc* 2010;132:11159–11164.
43. Wei C, Mei Y, Zhang D. Theoretical study on the HIV-1 integrase–5CITEP complex based on polarized force fields. *Chem Phys Lett* 2010;495:121–124.
44. Tong Y, Mei Y, Li YL, Ji CG, Zhang JZH. Electrostatic polarization makes a substantial contribution to the free energy of avidin–biotin binding. *J Am Chem Soc* 2010;132:5137–5142.
45. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman AP AMBER 9. San Francisco. University of California; 2006.
46. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 2006;65:712–725.
47. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee TJ. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Comput Chem* 2003;24:1999–2012.
48. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79:926–935.
49. Ryckaert J, Ciccotti G, Berendsen H. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-Alkanes. *Comput Phys* 1977;23:327–341.
50. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996;14:33–38,27–28.