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

Stockholm 2017

Published by Karolinska Institutet.
Printed by E-Print AB 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 p*K* shifts of cytidine in triplex nucleic acid environments. In some of the triplex nucleic acid systems, sugar modification LNA is present. The force field parameters of LNA have been updated to provider better descriptions for p*K* 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 triplex DNA. *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

IV. Hartono, Y.D.; Ito, M.; Villa, A; Nilsson, L. Keto-enol tautomerisation of modified uridine in ribosome decoding centre. *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

* These authors contributed equally

CONTENTS

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
$^{\Psi}$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. 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 molecules, 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, hydrophobic, acidic, basic, and others. Consequently, proteins, with their vast array of structures, perform many functions: catalysis, ion transport, receptor, messenger, among others.

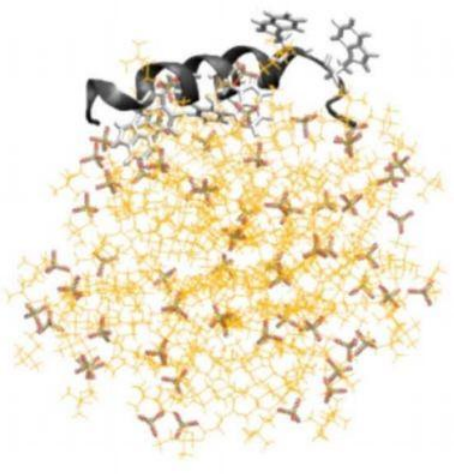


Figure 1. MPER peptide adsorbed to a micelle.

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 port-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 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.

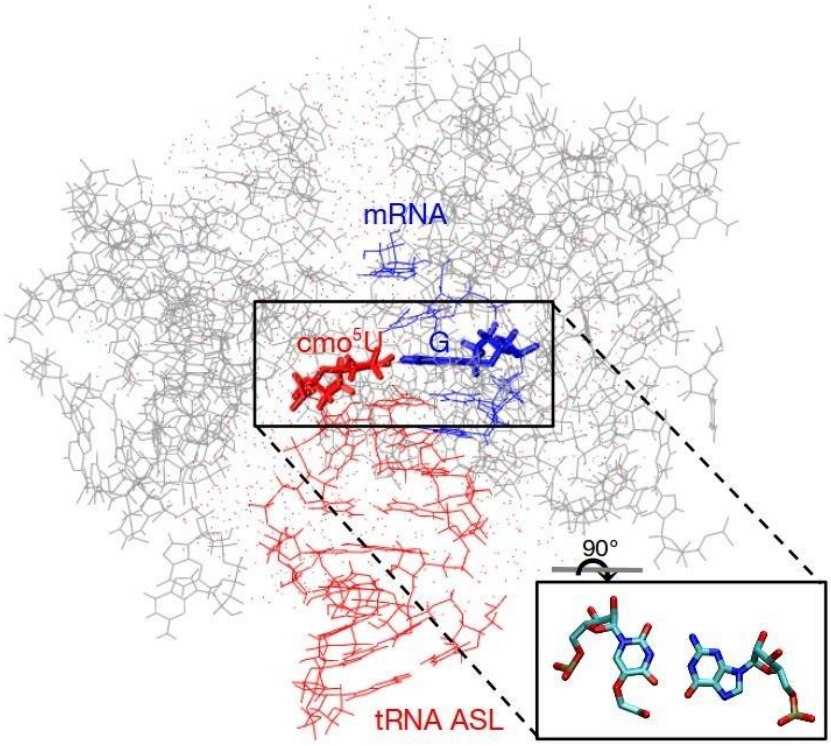


Figure 2. Subset of the ribosomal decoding region from X-ray crystal structure. tRNA ASL refers to the tRNA anticodon stem loop. Inset shows cmo^{5}U:G basepair in Watson-Crick geometry

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).

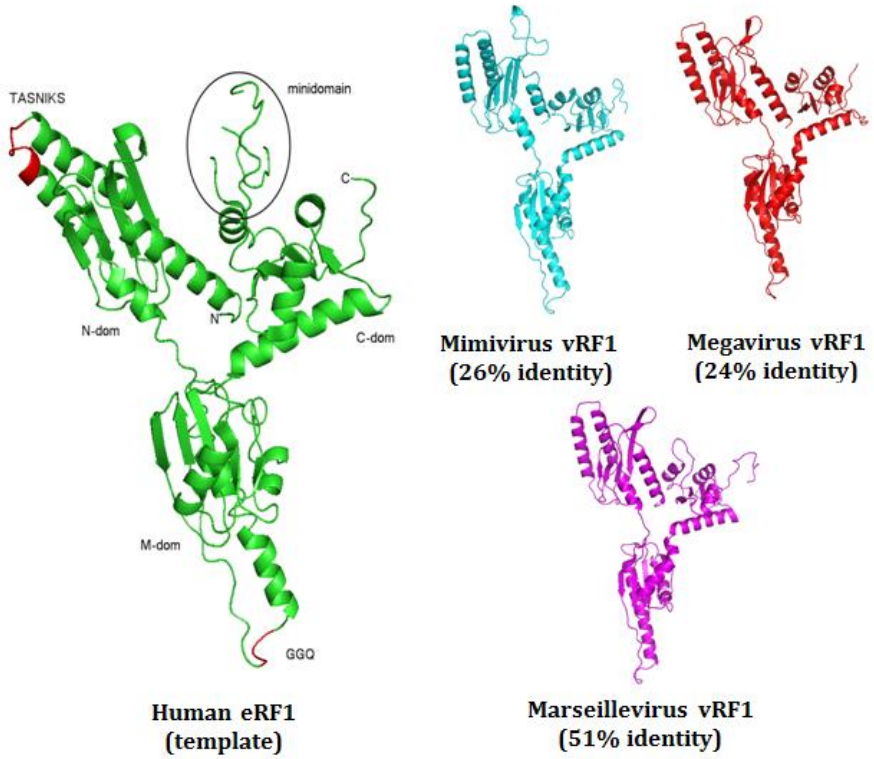


Figure 3. The structures of human eRF1 (X-ray) and vRF1s (homology models).

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 homopyrimidine strand and a homopyrimidine strand, a third homopyrimidine strand can bind at the major groove of the double helix via Hoogsteen basepairing to from a parallel triple helix (Figure 4). With canonical DNA bases, the possible base triads are C^{+}•G–C and T•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]

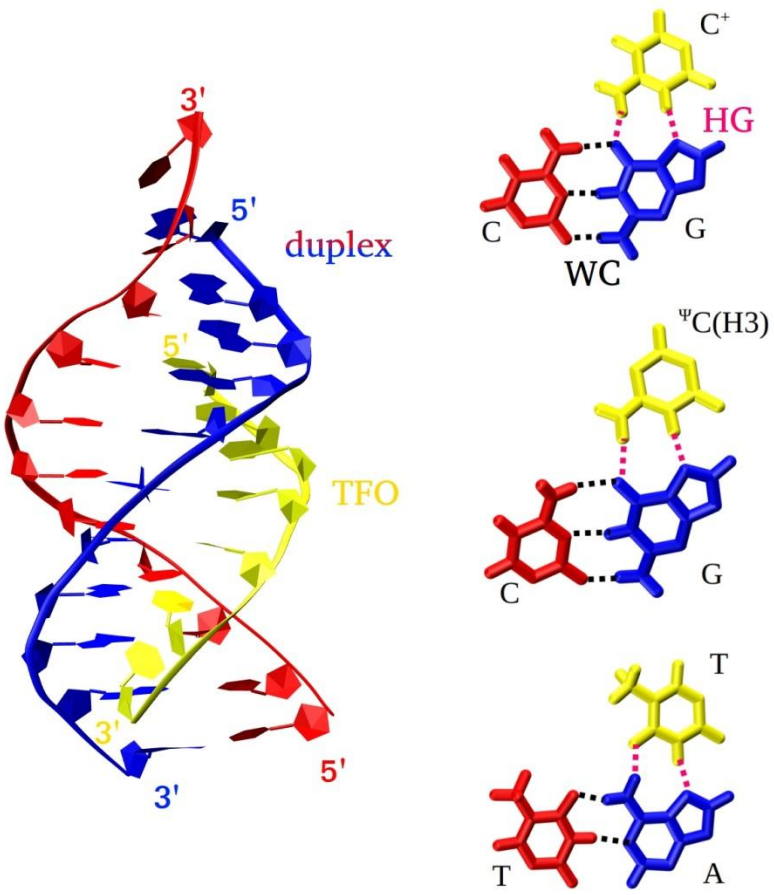


Figure 4. Parallel triple helix and base triads, shown with Watson-Crick (WC) and Hoogsteen (HG) hydrogen bonding.

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 (cmo^5U) (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 ($^{\Psi}C$),[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.

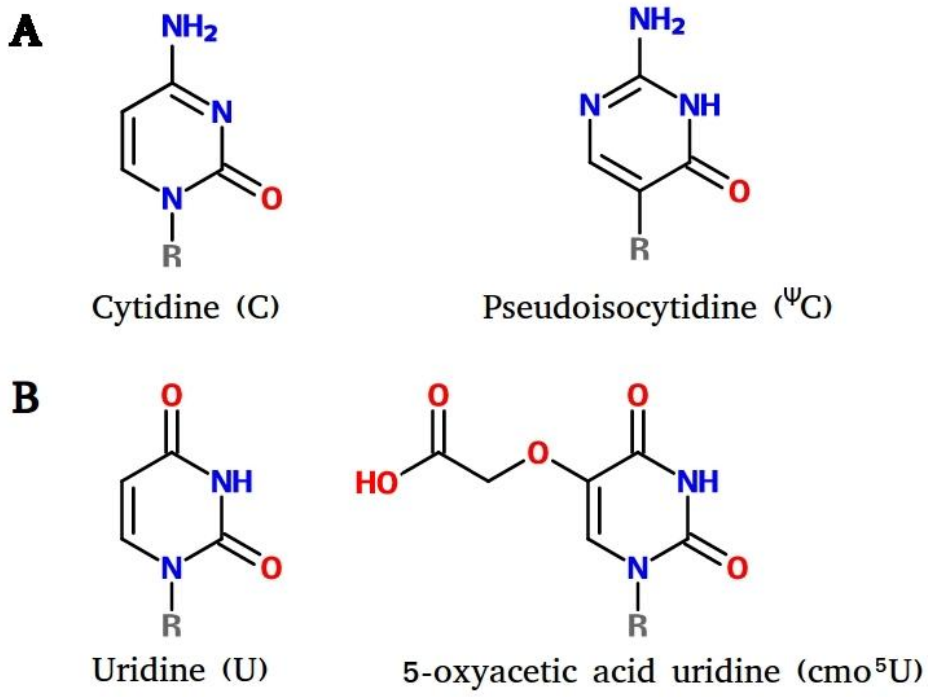


Figure 5. Chemical structures of two modified bases, shown with the corresponding canonical bases. (A) Cytidine and pseudoisocytidine (B) Uridine and 5-oxyacetic acid uridine

An artificial sugar modification, locked nucleic acid[14-15] (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]

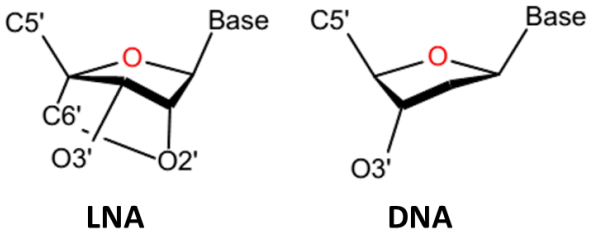


Figure 6. Chemical structures of LNA and DNA.

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.

X-ray crystallography typically cannot resolve hydrogen atoms 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 spectroscopy[20] 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 of 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 Ni^{2+} or Co^{2+} 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 model[43] 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 pseudoisocytidine 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).

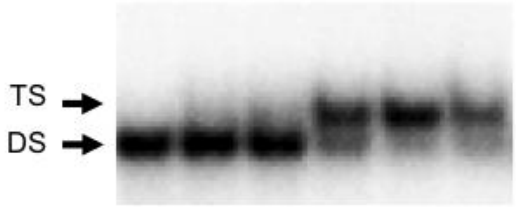
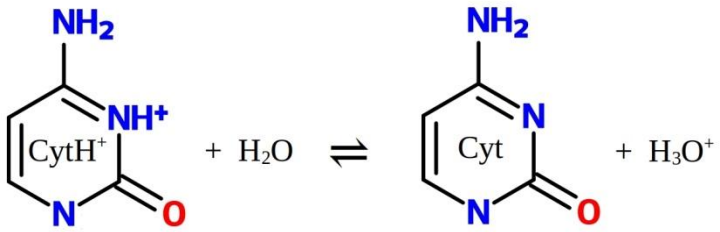


Figure 7. An electrophoretic mobility shift assay (EMSA) gel of TFO binding experiment. DS is duplex; TS is triplex.

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:

$$K = \frac{[\text{cyt}][\text{H}_3\text{O}^+]}{[\text{cytH}^+]} \text{ and } \text{p}K = -\log\frac{[\text{cyt}][\text{H}_3\text{O}^+]}{[\text{cytH}^+]}$$

The pK value of the cytosine is 4.4.[45] Since the equation can be rearranged to include pH, i.e. $\text{p}K - \text{pH} = \log\frac{[\text{cyt}]}{[\text{cytH}^+]}$, 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:

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

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]

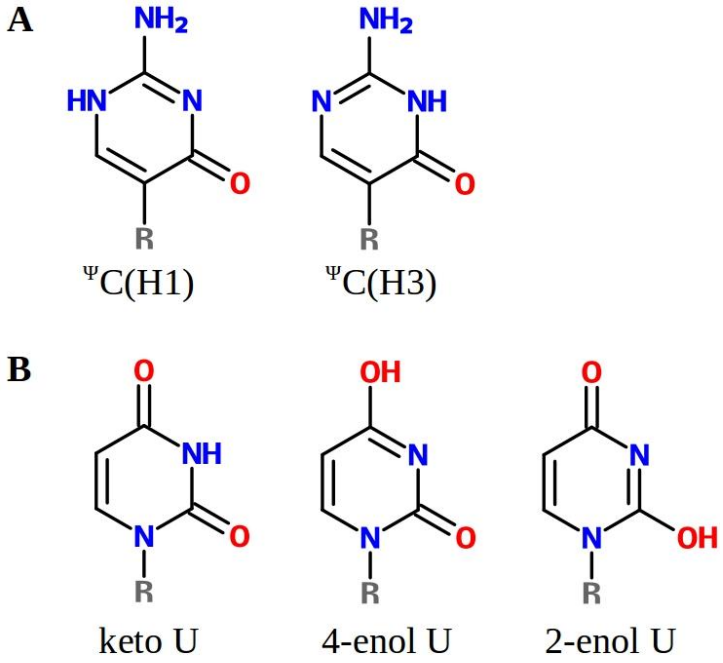


Figure 8. Tautomers of (A) pseudoisocytidine ($^{\Psi}$C) and (B) uridine.

In this thesis, the prototropic tautomers of pseudoisocytidine ($^{\Psi}$C), the aforementioned artificial base, which are two main amino tautomers, $^{\Psi}$C(H1) and $^{\Psi}$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 A⇋B tautomerisation are the equilibrium constant $K_{\mathrm{A}\leftrightharpoons\mathrm{B}} = \frac{[\mathrm{B}]}{[\mathrm{A}]}$ and the free energy of tautomerisation $\Delta G_{\mathrm{A}\leftrightharpoons\mathrm{B}} = k_{\mathrm{B}} T \ln K_{\mathrm{A}\leftrightharpoons\mathrm{B}}$.

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:

$$U(\vec{R}) = \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} K_\varphi(1+\cos(n\varphi-\delta)) + \sum_{\text{impropers}} K_\omega(\omega-\omega_0)^2 + \sum_{\text{non-bonded pairs}} \left\{ \varepsilon_{ij}^{\min} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\}$$

This is the potential energy function used by CHARMM[56] (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; $\varepsilon_{ij}^{\min}$, 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 interface[68] 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) scheme[69] 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 p*K* 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 constraints[71] 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 scaling[72] can be employed, since the temperature is related to the average kinetic energy. Alternatively, Berendsen thermostat[73] can be used, in

which the system is coupled to an external heat bath. Nosé-Hoover thermostat[74-75] 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 barostat[73] analogously couple the system to a 'pressure bath' like the thermostat. Andersen barostat[76] 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 inhomogenous system. We instead used deformable stochastic boundary potential[77-78] 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-dynamics[80-82] 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 CHARMM[83] 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:

$$U_{\mathrm{tot}}(X,\{x\},\{\lambda\}) = U_{\mathrm{env}}(X) + \lambda_{\mathrm{A}}[U(X,x_{\mathrm{A}}) - \Delta G_{\mathrm{A}\to\mathrm{B}}(\mathrm{model})] + \lambda_{\mathrm{B}}[U(X,\mathrm{x_B})] + F^{\mathrm{bias}}(\lambda_{\mathrm{A}}) + F^{\mathrm{bias}}(\lambda_{\mathrm{B}})$$

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. $U_{\mathrm{env}}(X)$ is the potential energy of environment atoms not involved in the transition.

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

$$0 \le \lambda_i \le 1 \text{ and } \lambda_{\mathrm{A}} + \lambda_{\mathrm{B}} = 1$$

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:

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

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

$\Delta G_{\mathrm{A}\to\mathrm{B}}(\mathrm{model})$ 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, $\ln(10)\, k_B T(\mathrm{p}K_{\mathrm{model}} - \mathrm{pH})$, is included in λ_{A} scaling, such that:

$$U_{\text{tot}}(X, \{x\}, \{\lambda\}, \text{pH}) = U_{\text{env}}(X) + \lambda_{\text{A}}[U(X, x_{\text{A}}) - \Delta G_{\text{A}\to\text{A+}}(\text{model}) + \ln(10)\, k_B T(\text{p}K_{\text{model}} - \text{pH})] + \lambda_{\text{A+}}[U(X, \text{x}_{\text{A+}})] + F^{\text{bias}}(\lambda_{\text{A}}) + F^{\text{bias}}(\lambda_{\text{A+}})$$

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 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 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, 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 regime[86] is an approach that specifically aims to calculate p*K* by combining configuration sampling of MD with Poisson-Boltzmann equation for electrostatic calculation. Let us consider that we want to calculate 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 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 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 p*K*, $\text{p}K_{\text{int}}$, is the p*K* value independent of ionisation properties of other groups in the system. $\text{p}K_{\text{mod}}$ is the standard p*K* value for the model compound, which we need to supply. $\Delta\text{p}K_{\text{Born}}$ is the p*K* shift due to desolvation, and $\Delta\text{p}K_{\text{cc}}$ is 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, $\langle \quad \rangle_i$ 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) method[89] 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:

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

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

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

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

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

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:

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

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; $U_{\mathrm{bias},i}$ is biasing

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

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

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, p*K*
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 p*K* shift in triplex DNA

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 strategy[99] 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 bounds, 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^+•G–C and T•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-Nilsson[100]) as well as calculated p*K* 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 p*K* shift by calculating electrostatics with linearised Poisson equation.[86] 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 p*K* approaches both predict the p*K* of cytidine to be shifted higher than physiological pH in triplex environment. 5-methylation shifted p*K* even higher. Neighbouring LNAs do not seem to have large effect on p*K* shift. For cytidine, multiple pH regime predicts higher p*K* value than λ-dynamics and similar ones for 5-methylcytidine. Multiple pH regime predicts downward p*K* 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 ($^{\Psi}$C) is a cytidine analogue designed to circumvent this problem. However, $^{\Psi}$C 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)•G–C.

We have used λ-dynamics[84] to model the tautomerisation between the tautomers of $^{\Psi}$C, $^{\Psi}\mathrm{C(H1)} \rightleftharpoons {}^{\Psi}\mathrm{C(H3)}$ 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 $^{\Psi}$C.

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 $^{\Psi}$C. We observed that sequence containing consecutive $^{\Psi}$Cs tend to disfavour $^{\Psi}$C(H3), while cytidine neighbours tend to favour $^{\Psi}$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, $^{\Psi}$C(H3) fully dominates, even when $^{\Psi}$C is consecutive.

EMSA shows that sequences containing three or more consecutive $^{\Psi}$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 $^{\Psi}$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. Keto-enol tautomerisation of modified uridine in ribosome decoding centre

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 potential[77-78] 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 to 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 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 antibodies[102-104] 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 preserve 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 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 p*K* values for cytidine in triplex environment. Since experimental p*K* 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. p*K* 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 dismantle 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 a*b 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 heartful 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, Peiying 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.; Nousch, 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.; Unbehaun, 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) Spyrakis, 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 a s 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 $\chi 1$ and $\chi 2$ 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 utilizaion of AB initio molecular potentials for the prevision 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.; 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.

(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 pK 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.; Brusic, 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.