Scuola Internazionale Superiore di Studi Avanzati (SISSA)
International School for Advanced Studies

Investigating Peptide/RNA Binding in Anti-HIV Research by Molecular Simulations:

Electrostatic Recognition and Accelerated Sampling

Thesis submitted for the degree of
Doctor Philosophiæ

Candidate:

Nhu Trang Do

Thesis Advisors:

Dr. Giovanni Bussi

Prof. Paolo Carloni

Trieste, October, 2012

Acknowledgment

The PhD-obtaining process including the writing of this thesis is so far one of the most significant scientific and academic challenges I have ever faced.

I would like to thank my first supervisor, Prof. Paolo Carloni. Paolo has been a great manager of my PhD project: promoting and insisting on the scientific subject, motivating scientific discussions, interacting and connecting involved collaborators. Without Paolo, the subject of my PhD project would not have been so exciting. His knowledge and experience has broadened my scientific awareness.

I am grateful to my second supervisor, Dr. Giovanni Bussi. I happened to be Giovanni's student in a scientific "incident". The time we have been really working together is not much, but in this short period, his guidance has a great impact on me. I am thankful to Giovanni for his patience in seeing me making mistakes, correcting my mistakes, and getting ready for correcting them another time. With Giovanni, I have learned how to pose the relevant scientific questions, how to have a non-biased mindset for solving scientific problems, and a great set of know-how skills.

I sincerely thank my two supervisors, again, for getting together to guide me and lead the project. It has been very interesting working with them both who have different personalities and different scientific backgrounds. All these make my PhD training an unforgettable experience.

I am indebted to Dr. Emiliano Ippoliti. Emiliano is like my third advisor, he guided me through the very beginning of my PhD training period when I came unprepared and had a serious lack of experience on the new area.

I am also thankful to Prof. Michele Parrinello and Prof. Gabriele Varani for their plenty of precious comments on the project.

My thanks also go to Dr. Attilio Vargiu, Dr. Giacomo Fiorin, and Dr. Adriana Pietropaolo for their great scientific help and technical support.

Prof. Alessandro Laio and Prof. Cristian Micheletti are members of the staff of the SBP sector at SISSA. From their lectures I have learned a great deal of science. Alessandro's first lectures on Metadynamics and Cristian's exam on Jarzynski's equality in my first year have been helpful to my research and so will be.

I also want to take advantage of this PhD thesis to thank again my previous advisers, Prof. Hoang Zung, Dr. Do Hoang Son, Prof. Nguyen Thi Phuong Thoa, and Prof. Mai Suan Li. I will never be able to forget their invaluable guidance not only in science but also in many other life aspects.

To the whole SBP sector: you all have been my great colleagues and your friendship has eased my stress and my anxiety at work. I wouldn't have enough space to mention all your names and what are the great things each of you have done for me. But each of you know personally how much I appreciate. Thanks to Francesco Colizzi for his cool friendship and his Panda power. To Fahimeh, Zhaleh, Shima, and Shima: you are as close as sisters to me. I know I will never be able to thank you enough for all you gave me. Especially to Fahimeh and Zhaleh, thank you girls for all sleepless nights and countless shared pizzas of PhD fighters!

My thoughts are now for anh Minh and family, Yanier - Dania and family, anh Chuong (Genarino), Ana Chiara - Attilio and family, chi Quy, chi Linh, anh Linh, anh Dinh, anh Huy Viet, anh Tuan Anh. In different periods of my life, each of you has been like my second family.

Finally, to my real family, my Dad, my Mom, little sister, Grandpa, Grandma, and my husband: you are always crazily supporting me. Grandpa and Grandma have been giving me so much care and education. With them, I always feel like I am forever a little kid. Dad and Mom are always a great source of advice, support, encouragement, cheers, and comfort. Their great courage in life inspires me everytime I have to make a decision or fight for my decisions. Little sister (even though not anymore little but will always remain so to me) is so funny and joyful. She always teaches me how to look at things with her special and interesting point of view. The last person to whom I want to express deeply my gratefulness is my husband, Rolando. He has stood right beside me through all my difficult moments in both life and work. To me, he is a colleague, a friend, an advisor, a big brother, and an unconditional lover. This thesis cannot be in its shape today without Rolando reading and correcting it I don't know how many times.

Preface

Studying protein/RNA binding is of great biological and pharmaceutical importance. In the past two decades, RNA has gained growing attention in biomedical and pharmaceutical research due to its key roles in gene replication and expression [1, 2]. From a pharmaceutical point of view, the advantages of targeting RNA over the conventional protein targets include slower drug-resistance development, more selective inhibition, and lower cytotoxicity. Targeting RNA is, however, more challenging than targeting proteins. Designing RNA-binding drugs is limited by the lack of medicinal chemistry studies on RNA and the poor understanding of ligand/RNA molecular recognition mechanisms [2, 3].

Computer-aided drug-design targeting RNA faces several difficulties including the RNA flexibility [3] and the highly charged nature of RNA molecules [4, 5]. On the one hand, rigid-body docking and Brownian dynamics simulations are insufficient to properly describe the conformational changes and relevant inter-molecular contacts upon ligand/RNA binding. On the other hand, standard molecular-dynamics simulations require unaffordable computational cost for a full binding event to be observed. For all these reasons, there are still no *clinically relevant* RNA-binding drugs successfully developed by computer-based drug design approaches [3]. However, promising prospects are coming from the development of more accurate force fields for RNA [6, 7, 8, 9] and enhanced sampling methods [10, 11, 12, 13], which allow better and faster ways of investigating ligand/RNA complex formation.

In this thesis, we study a typical case of protein/RNA binding by means of both standard molecular dynamics and bidirectional steered molecular dynamics at the atomistic level with an explicit representation of solvent and ion molecules. Such an explicit representation allows a detailed and quantitative analysis of ion and solvent effects, which are highly important for charged systems.

With a total of more than 0.5 μs of standard molecular dynamics simulations, we are able to reproduce the structural and dynamical properties of the chosen system as observed in NMR experiments. Besides providing detailed information on ion and solvent distributions upon binding, our molecular dynamics simulations also confirm that binding between an RNA and a positively-charged peptide is a spontaneous process strongly driven by electrostatic interactions. However, standard molecular dynamics in sub-μs time-scales is insufficient to study binding events of large and highly charged systems.

We therefore introduce a methodological improvement to efficiently accelerate binding/unbinding processes: a new collective variable named *Debye-Hückel energy*. As an approximation of the electrostatic free energy component, this collective variable represents closely the electrostatics of the system including the screening effect of the ionic solution, and hence can be proficiently used to accelerate the dynamics of molecules in explicit solvent. To the best of our knowledge, this is the first physics-based collective variable designed to study binding/unbinding processes.

We next perform 4.2 μs of bidirectional steered molecular dynamics simulations, in which a biasing force acts on our Debye-Hückel energy collective variable. Within the framework of bidirectional steerings, we also propose a method to reconstruct the potential of mean force as a function of any *a posteriori* chosen collective variable. This allows a flexible post-processing of simulation results. From our steering simulations, we could predict the correct binding pocket observed in NMR experiments in a completely "blind" manner, i.e., without any guidance from the NMR bound structure. Such a self-guiding feature is important since it is applicable even when experimental information in unavailable, which is the case of most computational drug designs.

Contents

List of Figures

List of Tables

List of Abbreviations

AFM	Atomic Force Microscopy, 27
AIDS	Acquired Immune Deficiency Syndrome, 3
BAR	Bennett Acceptance Ratio, 37
CV	Collective Variable, 47
DHEN	Debye-Hückel ENergy, 52
DNA	DeoxyriboNucleic Acid, 4
HIV	Human Immunodeficiency Virus, 3
LOT	Laser Optical Tweezers, 27
LTR	Long Terminal Repeat, 6
MD	Molecular Dynamics, 16
NMR	Nuclear Magnetic Resonance, 8
NOE	Nuclear Overhauser Effect, 24
NOESY	Nuclear Overhauser Effect SpectroscopY, 8
PBC	Periodic Boundary Condition, 22
PBE	Poisson-Boltzmann Equation, 48
PCA	Principle Component Analysis, 62
PMF	Potential of Mean Force, 32
RMSD	Root Mean Square Deviation, 61
RMSF	Root Mean Square Fluctuation, 62
RNA	RiboNucleic Acid, 1

SMD	Steered Molecular Dynamics, 27
TAK	Tat-Associated Kinase, 7
TAR	Trans-Activation Response, 7
Tat	Trans-activator of transcription, 6
WHAM	Weighted Histogram Analysis Method, 43
WTMetaD	Well-Tempered MetaDynamics, 96

Chapter 1

Introduction

Contents

1.1 Overview

This thesis contains an *in silico* study of how a small peptidic ligand binds an RNA molecule. The **M***olecular* **D***ynamics* (**MD**) simulations were performed at an atomistic level in explicit ionic solvent. This is crucial not only for observing the conformational changes during the binding of ligands to RNA but also for understanding the sequence-specific recognition, ion rearrangement, and the role of solvents upon binding.

In addition to standard MD, which requires extremely high computational costs to cover biologically relevant time-scales, we adopt a state-of-the-art enhanced sampling technique, namely **S***teered* **M***olecular* **D***ynamics* (**SMD**) [10, 12]. This technique is based on biasing an *a priori* chosen **C***ollective* **V***ariable* (**CV**). Although SMD and metadynamics have already been used for ligand binding studies (see, for instance, Refs [14, 15]), a proper choice of the CVs remains challenging. As a new approach for an efficient acceleration, we have designed and implemented a physics-based CV, called **D***ebye*-**H***ückel* **EN***ergy* (**DHEN**), which is an approximation of the electrostatic-interaction term of the free energy. Based on a slightly different philosophy, several authors have proposed to bias the potential energy of the system [16, 17, 18, 19, 20]. However, the potential energy cannot be interpreted as a proper CV in solvated systems where the large fluctuating contributions arise from the solvent-solvent interactions. Indeed, these methods are efficiently used in solvated systems in a spirit more related to that of parallel tempering [21], simulated tempering [22], and multicanonical sampling [23], i.e., to let the system evolve in an ensemble where effective barriers are decreased and conformational transitions are more likely to occur. In this respect, it appears fascinating to push this idea further and to use as a CV only the component of the energy which is relevant for the transition of interest such that solvent fluctuations are averaged out. To the best of our knowledge, this is the first time that an "approximation of a free energy component" is used as a CV in nonequilibrium simulations to explore the "real" free energy. This new CV stands out as a physics-based CV that takes into consideration the long range electrostatic interaction with ionic solution screening.

Besides the computational advances, the thesis also targets a growing biological interest: peptide/RNA interaction, which is extremely challenging due to the high charge and flexibility of RNAs as well as their peptidic ligands. Last but not least, from a pharmaceutical point of view, targeting viral RNAs is presumably more effective than targeting viral proteins since viral proteins can mutate more easily than RNAs to develop drug resistance. Therefore, the development of a robust protocol for ligand/RNA association could pave the way to the computer-based design of new drugs targeting viral RNAs. The typical stem-bulge-loop HIV-1 TAR RNA in complex with a small cyclic peptidic ligand (partially mimicking the sequence and structure of the HIV-1 Tat protein, see Ref. [24]) is chosen as a case study.

In the next Sections, we introduce the specific pharmaceutical relevance of our case study, followed by a detailed description of our system of choice and a review on the current stage of computational studies of protein/RNA bindings.

1.2 Pharmaceutical Relevance of Studying HIV-1 TAR RNA/ Ligand Complex

In this section, we present the basis of HIV and its replication process together with the current stage of research in seeking anti-HIV drugs − progresses as well as drawbacks. We also describe the chosen biological system for our simulations throughout the thesis and discuss its pharmaceutical implication.

1.2.1 Introduction

The family *Retroviridae* consists of several non-icosahedral, enveloped viruses [25]. Two fundamental characteristics of the replication process of the Retroviridae family include the reverse transcription of the genomic RNA into a linear double-stranded DNA and the subsequent covalent integration of this DNA into the host genome. A virion of all *retroviruses* is composed of

(i) an *envelope* made of *glycoproteins* (encoded by the *env* gene) and *lipids* (obtained from the plasma membrane of the host cell during the budding process),

(ii) a *dimer RNA* with terminal noncoding regions and the internal regions that encode virion proteins for gene expression,

(iii) *proteins* encoded by the retroviral genomes including: the *Gag polyproteins* (encoded by the *gag* gene) forming the major components of the viral capsid; the *protease* (encoded by the *pro* gene) performing proteolytic cleavages during virion maturation to make mature gag and pol proteins; the *reverse transcriptase*, *RNase H* and *integrase* (all encoded by the *pol* gene) responsible for synthesis of viral DNA and integration into host DNA after infection; and the *surface glycoprotein* and *transmembrane* proteins (encoded by the *env* gene) mediating cellular receptor binding and membrane fusion. The env protein is what enables the retrovirus to be infectious.

The **H***uman* **I***mmunodeficiency* **V***irus* (**HIV**) is a *retrovirus* that affects vital cells in the human immune system, especially the *CD4$^+$ T cells*, and causes the **A***cquired* **I***mmune* **D***eficiency* **S***yndrome* (**AIDS**) [26, 27]. The discovery of HIV in the early 1980s [28] earned Françoise Barré-Sinoussi and Luc Montagnier the Nobel Prize in Physiology and Medicine in 2008. Today, HIV has become a global pandemic. It has been reported that HIV/AIDS caused more than 27 million deaths among 34.2 million infected cases worldwide by the end of 2011 [29]. There are two types of HIV that have been characterized, namely HIV-1 and HIV-2. HIV-1 causes global infections while HIV-2 is less infective and mostly confined to West Africa [30, 31]. In this thesis we focus on HIV-1.

The six basic steps of HIV-1 replication include [32]:

Step 1: *Fusion and entry.* The glycoprotein *gp120* on the envelope of HIV-1 strongly interacts with the CD4 receptor on the surface of human T cells [33]. The binding of gp120 to CD4 receptor promotes further binding of a co-receptor resulting in a subsequent conformational change in gp120. This allows *gp41*, another viral glycoprotein embedded in the viral envelope, to unfold and insert its hydrophobic terminus into the cell membranes, facilitating the fusion of the viral and cellular membranes. Finally, the viral capsid enters the host cell and releases two viral RNA strands and three essential viral enzymes [34].

Step 2: *Reverse transcription.* Genetic information of the retrovirus HIV-1 is carried by two strands of RNA [25] while genetic material of human cell is found in DNA. Therefore, HIV-1 makes a DNA copy from its RNA through a process called reverse transcription. This task is done by the viral reverse transcriptase enzyme [35]. The new DNA produced by this process is called proviral DNA.

Step 3: *Integration.* The viral integrase enzyme cleaves two nucleotides from each 3′ end of the proviral DNA creating two *sticky* ends. Integrase then carries the cleaved proviral DNA into the nucleus of the host cell and facilitates its integration into the host cell's DNA [35, 36]. The host cell's genome now contains the genetic information of HIV-1 as well.

Step 4: *Transcription.* After the proviral DNA is integrated into the host cell's genome, the host cell's machinery accidentally induces the transcription of proviral DNA into viral messenger RNA (mRNA) [37].

Step 5: *Translation.* The mRNA containing the instruction to make new virus then migrates to the cytoplasm. Each section of the mRNA corresponds to a protein building block of the virus. As each mRNA strand is processed, a corresponding string of proteins is synthesized. After translated from viral mRNA, the long strings of proteins need to be cut into smaller proteins which are able to carry out their own functions. This process is done by the viral protease enzyme and is crucial to create an infectious virus [38, 39].

Step 6: *Assembly and maturation.* Two viral RNA strands and the enzymes then come together and other proteins assemble around them to form the viral capsid [40]. This immature viral particle buds off the host cell using the cell's membranes to make its own envelope. The virus then becomes mature and ready to infect other cells.

The above-mentioned steps are crucial for the HIV-1 life cycle and hence can be considered as targets for chemotherapeutic intervention [41]. Current anti-HIV drugs can decrease the replication capacity of HIV-1 in infected cells (see Table 1.1). Unfortunately, they cannot completely stop the replication of the virus [42].

1.2.2 Drug-Resitance Development in HIV-1

For a successful reproduction in life, it is necessary to have a mechanism that creates copies of the original genetic materials. Several enzymes performing this function are generally known as *replicases*. In the specific case of DNA copying processes, these enzymes are known as *DNA polymerases*. As it is critically important to copy the genetic material in a precise way, it is not surprising to find *error-correcting* or *proof-reading* mechanisms in most species from bacteria to human. The proof-reading mechanism is performed through

Class of drugs	First invention	Actions of drugs
Fusion/entry inhibitors	2005 (Enfuvirtide [43])	preventing HIV-1 from binding to or entering human immune cells (**step 1**)
Nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs)	1987 (Zidovudine [44])	interfering with the action of reverse transcriptase enzyme which transcribes viral RNA into proviral DNA (**step 2**)
Non-nucleoside reverse transcriptase inhibitors (NNRTIs)	1996 (Nevirapine [45])	inhibiting reverse transcriptase enzyme (**step 2**)
Integrase inhibitors	2007 (Raltegravir [46])	interfering with integrase enzyme which helps HIV-1 to insert its genetic materials into human genome (**step 3**)
Protease inhibitors	1995 (Saquinavir [47])	inhibiting protease enzyme which cleaves long strings of viral proteins into smaller functional proteins (**step 6**)

Table 1.1: Current anti-HIV drugs can inhibit the viral replication at a certain stage in the viral life cycle.

a *3′−5′ exonuclease* activity. Shortly, if during the copying process, an incorrect base has been incorporated, then this base is immediately recognized and the DNA polymerase reverses its direction by one base pair and eliminates the mismatched base. After that, the polymerase can re-insert the correct base and continue the copying process.

As we mention in the previous section, the copying enzyme for the genetic material of HIV-1 is called reverse transcriptase. However, this enzyme copies RNA into DNA. The copying process performed by the reversed transcriptase is fast but inaccurate. Unlike its bacterial or human counterparts, this copying does not possess a proof-reading mechanism and therefore the process of HIV-1 reverse transcription is extremely error-prone [48]. The resulting *mutations* can cause structural differences, i.e., each new generation of HIV-1 differs slightly from the previous one. Mutations occur randomly and are common in HIV-1. On one hand, most mutations give disadvantages to the virus itself: mutations may affect the viral functions and slow down its ability of infecting CD4$^+$ T cells. On the other hand, some mutations can actually give HIV-1 a survival advantage to escape from the control of human immune system and to fight against medical treatment: mutations can block the anti-body and anti-HIV drugs from interacting with the HIV-1 enzymes which they are designed to target. HIV-1 has been developing resistance to all current clinical drugs including fusion inhibitors, nucleoside and non-nucleoside reverse transcriptase inhibitors, integrase inhibitors, and protease inhibitors [49, 50].

HIV-1 drug resistance is one of the main reasons why *mono-therapy* treatment, i.e., using only one type of drugs, fails after used for a prolonged period of time. One of the medical solutions to decrease drug resistance is to take a combination of three or more drugs simultaneously [51].

1.2.3 Searching for New Anti-HIV Drugs

Searching for new classes of drugs could be the temporary solution of the strongly growing drug-resistance problem [52]. In this respect, in the HIV-1 replication process, the transcription step (step 4) play an extremely important role. It is required not only during the exponential growth of the virus but also, critically, during the activation of the integrated proviral genome which also facilitates drug resistance [53]. However, although several compounds blocking the viral transcription have been synthesized [54, 55, 56, 57, 58, 59, 60], neither of them have yet been approved as anti-HIV drugs.

1.2.3.1 HIV-1 Transcription Process

The transcription from an integrated proviral DNA into a viral mRNA is accidentally carried out by the human cellular transcription factors. However, this process cannot be completed without the regulation of a complex interplay among the cellular transcription factors and two key transcriptional regulatory elements of the virus itself, namely the proviral **L***ong* **T***erminal* **R***epeats* (**LTR**s), and the viral **T***rans*-**a***ctivator of* **t***ranscription* (**Tat**) protein [61]:

(i) LTRs are the two identical sequences found at the two ends of proviral DNA. They carry out two important functions [62]: *(i)* they are the sticky ends used by the HIV-1 integrase to insert the proviral DNA into human genome and *(ii)* they act as promoter/enhancer in the transcription process. The 5′ LTR normally acts as an RNA polymerase II promoter while the 3′ LTR normally takes part in the transcription termination. When integrated into the host genome, they influence the cell transcriptional machinery to change the amount of transcripts which are going to be made.

(ii) Tat is a viral protein comprising of between 86 and 101 amino acids [62, 63]. It enhances the transcriptional elongation [64].

The transcription is initiated at 5′ LTR promoter. In the absence of Tat, most of the viral transcriptions terminate prematurely, producing a short RNA molecule with 60 to 80 nucleotides [65, 66] (see Figure 1.2.1a). When Tat is present, it increases the transcription level by several thousand-fold [64, 67]. In this process, Tat binds to the viral **T***rans*-**A***ctivation* **R***esponse* element (**TAR**) [68]. TAR is a small RNA portion transcribed from the first 59 nucleotides of the proviral DNA [69]. As soon as TAR is transcribed, Tat enhances the transcription of the remainder HIV-1 genetic code (see Figure 1.2.1b). The exact mechanism of enhancement is still under debate. However, it is known that after binding to

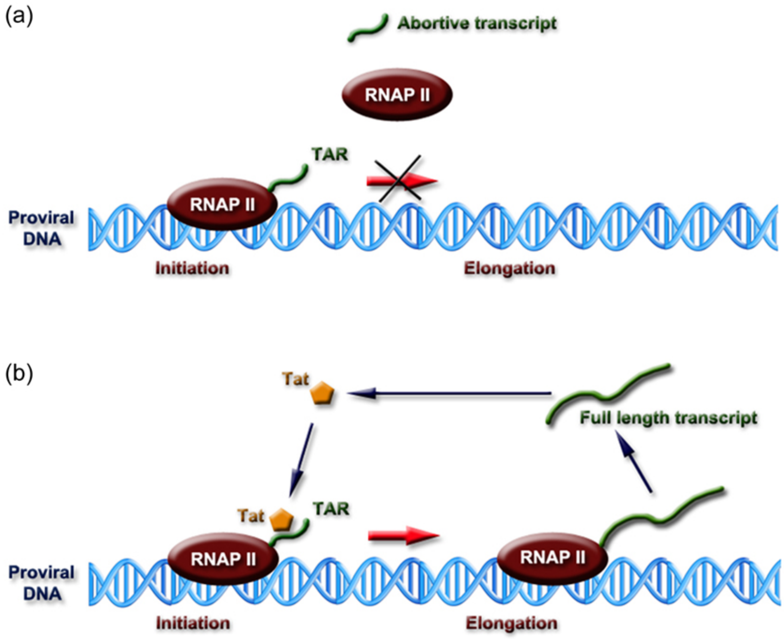

Figure 1.2.1: A schematic representation of the HIV-1 transcription process. (a) The transcription is aborted after producing 60−80 nucleotides due to the absence of Tat. (b) A full-length transcription is enhanced by Tat protein binding to TAR RNA.

TAR, Tat stimulates a specific kinase called **T***at*-**A***ssociated* **K***inase* (**TAK**) [70]. This kinase then performs a hyperphosphorilation of the cellular RNA polymerase II. Perhaps more interesting is the presence of a functional link between the cyclin T1 component of TAK and transactivation. In fact, TAK is able to form a tenary complex with TAR and cyclin T1 only when a functional loop of TAR has been generated suggesting that under these conditions, this loop region can act as a binding site for other cellular co-factors that can potentially enhance the transcription of the last part of rest of the proviral DNA [70].

1.2.3.2 Targeting Tat/TAR Interactions for Plausible Intervention

The key role played by Tat in HIV-1 transcription has made it the center of attention since its discovery in 1985 [62]. Inhibition of Tat-TAR interactions is becoming an attractive target since:

(i) TAR is the viral RNA element transcribed from 59 first nucleotides of the viral 5′ LTR promoter [69]. This promoter plays a crucial role in the replication of HIV-1. *In vitro* studies have shown that conserved nucleotides in TAR stem regions are critical for Tat binding [71] and mutations of nucleotides in TAR hairpin severely affect the formation of the viral RNA dimer [72]. Therefore, designing drugs targeting TAR

could greatly reduce the drug resistance due to viral mutations.

(ii) *in vitro* studies have shown that the apical portion of TAR (from G17 to G45) specifically binds to the arginine-rich region of Tat [73, 74, 75]. This discovery sheds light on the general features of the Tat-binding-site of TAR and hence limits the range for searching binding inhibitors. This could lead to classes of inhibitors with high affinity and specificity which are important criteria in drug design.

(iii) there is no counterpart of TAR or Tat in humans cells. Therefore, targeting Tat/TAR interaction is presumably advantageous over most of current anti-HIV therapies which are interfering with the human cell's function.

By combinatorial approaches, Hamy et al. were able to identify a peptide named CGP64222 which could compete with Tat in binding to TAR. In fact, NMR analysis has shown that CGP64222 binds to TAR at Tat binding site [76]. CGP64222 is the first antiviral compound that can selectively inhibit a protein-RNA interaction. Since then, numerous Tat-TAR interaction inhibitors have been synthesized and evaluated in the last two decades [77, 78, 54, 79, 80, 55, 56, 81, 82, 57]. However, none of these molecules have been approved for preclinical studies due to their low binding affinity or specificity, and hence, inhibiting Tat-TAR interactions still remains challenging.

1.3 Structural Features of a Tat Mimic in Complex with TAR

As a new approach to tackling Tat-TAR interaction, Davidson et al. synthesized conformationally constrained mimics of HIV-1 Tat. They discovered a family of 14-amino-acid beta-hairpin cyclic peptides able to bind to TAR with nanomolar affinity and greatly improved specificity compared with previous ligands [24]. Among 100 peptides in the investigated family, the arginine-rich sequence cyclo-RVRTRKGRRIRIPP (L22 hereafter, see Figure 1.3.1b) stands out for its potency. L22 binds to TAR with an affinity of 1 nM and exhibits a large number of intermolecular **N***uclear* **O***verhauser* **E***ffect* (**NOE**) data observed in **N***uclear* **O***verhauser* **E***ffect* **S***pectroscop***Y** (**NOESY**) spectra when in complex with TAR (pdb code: 2KDQ). NMR experiments have shown that L22 binds to the major groove of the upper RNA helix (nucleotides 26-29 and 36-39, see Figure 1.3.1c), which is also the binding pocket of Tat [73, 74, 75].

As observed in NMR experiments, the loop residue A35 was not only flipped out from the RNA but also drawn downward through a cation-π interaction with the guanidinium group of Arg11 (see Figure 1.3.2a). The hydrophobic residue Ile10 was buried in the TAR major groove and presumably facilitated the formation of the base triple U23/A27-U38 (Figure 1.3.2b). The side chain of Arg5 stacked on top of the base of U23 and thus provided

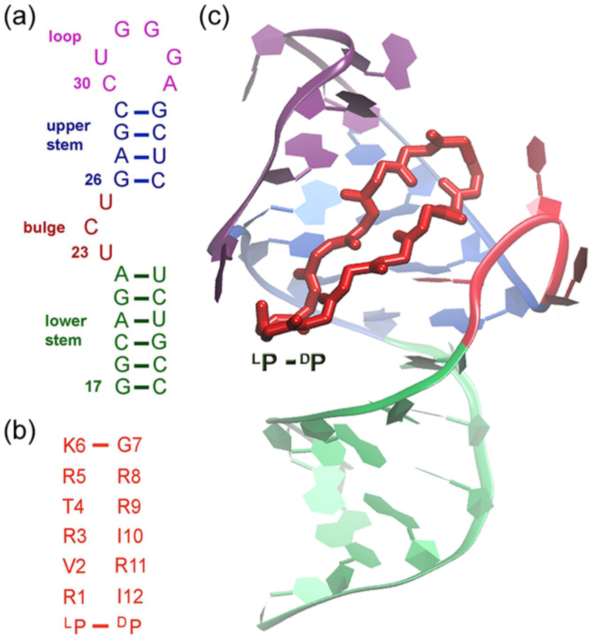

Figure 1.3.1: Sequences of TAR RNA (a) and L22 peptide (b) and the L22-TAR complex (c), whose structure has been determined by NMR experiment [24]. TAR has two well-defined double helical regions (green and blue), a bulge (red), and an apical loop (magenta). L22 has a $\beta-$hairpin loop structure, stabilized by ^{L}Pro$-^{D}$Pro.

a cation-π interaction with this nucleotide (Figure 1.3.2c). The guanidinium group of Arg5 formed hydrogen bonds with G28 (Figure 1.3.2d). Besides the above-mentioned key interactions, L22 also formed several other polar and hydrophobic interactions with TAR. They all contributed to keeping the complex well structured with a high binding affinity, i.e, $K_d = 1$ nM, in an ionic solution of 10 mM, in which K^+ was used as cation and a mixture of $HPO_4{}^{2-}$ and $H_2PO_4{}^{-}$ was used as anion [83, 24].

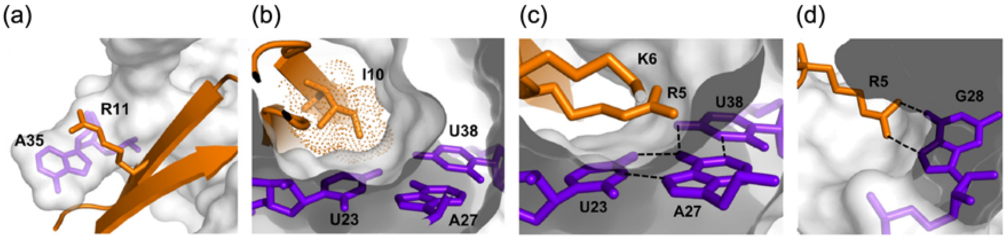

Figure 1.3.2: Key L22-TAR interactions as observed in NMR experiments. Figures are reproduced from Ref. [24].

The L22-TAR complex system is ideally suited to study protein-RNA binding for several reasons:

(i) the structures and dynamics of apo-TAR and L22-TAR complex have been characterized

by NMR experiments [84, 83, 24]; this allows a detailed comparison with molecular simulations,

(ii) since L22 is a structural mimic of Tat and binds to TAR at the same region as Tat [85] with comparable affinities ($K_d = 1$ nM versus $K_d = 10$ nM, respectively [86, 83, 24]), it is expected that several aspects of the L22-TAR binding mechanism would be shared with the much less well understood Tat-TAR interaction,

(iii) L22 is nearly as active as the antiviral drug nevirapine against a variety of clinical isolates in human lymphocytes, and therefore represents an attractive antiviral lead compound [87].

The L22-TAR complex was thus chosen as a case study for this thesis. *In silico* studies can provide important insights on structural and dynamical properties of the L22-TAR complex to gain a better understanding of the underlying molecular recognition mechanism.

1.4 Computational Studies of Peptide-RNA Bindings

In the past three decades, while docking tools have been successfully developed for predicting protein-protein interactions [88, 89], much less has been achieved for an efficient and accurate prediction of protein-RNA and peptide-RNA interactions despite great effort has been spent on improving the docking methods and scoring functions [90, 79, 91, 92, 93, 94, 95, 96, 97, 98]. The challenge for rigid-body docking methods comes from the high plasticity of RNA. Indeed, binding affinity and specificity are strongly dominated by induced-fit mechanisms [99, 100]. An understanding of molecular recognition thus requires a knowledge of RNA's conformational change upon binding.

MD simulations are currently the most suitable tools to study molecular recognition involving RNA. However MD simulations of nucleic acids seriously lagged behind those of proteins mainly due to the lack of available experimental high-resolution nucleic-acid structures which can allow a thorough validation of force fields and simulations [101]. Specifically, among MD simulations of nucleic acids, there are remarkably less simulations of RNA compared to DNA. This can be explained by the more complex structural determinants of RNA with respect to DNA. Indeed, unlike DNA which is often found in double helix structures, RNA can frequently feature non-canonical structural elements such as loops, hairpins, and bulges, which usually play an important role in the binding sites. This section contains discussions on both recent developments and current challenges faced by RNA MD simulations.

1.4.1 Achievements of MD Simulations of Protein-RNA Complexes

MD simulations on RNA have improved progressively. Before 1995, simulations were unable to provide stable RNA trajectories beyond 500-ps timescale. This was due to the limited resources in both computing power and experimentally resolved RNA structures (based on which the force field parametrization was done.) Since then, together with the increment of computing power as well as the introduction of PME treatment for long-range electrostatic forces (see Section 2.2.1.5 for the basic concepts), RNA force fields have been considerably refined (see Section 2.2.1.3 for discussions on force field improvements). A broad variety of RNA simulations (e.g., single- and double-stranded RNAs, catalytic RNAs, and increasingly large RNA-protein complexes) has been performed using AMBER force fields. In several cases, the results not only matched quantum chemical data but were also in good agreement with experiments. These simulations also provided stable structures of complex RNA systems in longer timescales (i.e., ~100 ns), shedding light on new structural and dynamical properties of RNA in aqueous ionic solution. The combination of increased computer power, more reliable experimental structures, and refined force fields makes MD simulation a promising tool for studying the structures, dynamics, and functions of RNA.

1.4.2 Challenges for MD Simulations of Protein-RNA Bindings

Highly Charged Character of RNA Molecules

The highly charged nature of RNA creates a strong solvation and ion association around the molecules [4, 5]. However, inclusion of an explicit representations of solvent and ion molecules increases significantly the number of atoms to be simulated; in a typical explicit-solvent MD simulations, more than 90% of the computational cost is spent on calculating solvent-solvent interactions [101]. More efficient alternative approaches have been recently developed with the creation of implicit-solvent representations based on Poisson-Boltzmann and Generalized-Born theories [102, 103, 3]. By using Poisson calculations in combination with a set of optimal Born atomic radii for small model compounds constituting the building blocks of nucleic acids, Banavali and Roux were able to not only provide a good agreement with solvation free energies from explicit-solvent calculations but also accurately describe the free energy associated with base pairing for both standard and mismatched nucleic-acid base pairs [104]. Regarding the Generalized-Born approach, which is an approximation of the Poisson-Boltzmann equation, application to RNA is still limited. However, Rizzo et al. found a high correlation between these two implicit-solvent approaches in representing absolute free energies of hydration for more than 500 neutral and charged compounds [105]. Despite some success has been reported for RNA systems, implicit-solvent methods still suffer from inaccuracy and high computational cost for large systems [101].

During the binding process, the displacement of water molecules and ions from the RNA structures results in an entropic contribution to the energetics of complex formation which needs to be treated carefully [106]. The residence time of solvent and ion molecules as well as the stabilizing effects of both mobile and bound solvents and ions partially determine the binding affinity and specificity [107]. Moreover, the presence of ions creates a screening effect which reduces the repellent interactions among the positively charged phosphate groups of nucleic acid backbones [108, 109]. Ions are thus crucial for stabilizing nucleic acid structures and require a careful treatment in simulations. Therefore, although time consuming, using an explicit solvent and ion representations is necessary to provide important information on the RNA structural adaptation and molecular recognition under solvation and ionic effects.

Force Field Inaccuracies

AMBER force field is one of the most used force fields for RNA simulations. Despite the satisfactory descriptions of structural and thermodynamic features of some RNA systems, several limitations have been also reported for this force field, including:

(i) the challenging description of the sugar-phosphate backbone which has multiple degrees of freedom and therefore cannot be correctly described using one set of partial atomic charges [110, 111].

(ii) the fixed charge model of the most current force fields is not able to characterize the highly polarizable feature of the phosphodiester moiety. This difficulty is expected to be overcome by a new generation of polarizable force fields [112], which has been tested only for proteins and DNA simulations.

(iii) the fixed charge model also fails to accurately describe interactions between RNA and divalent ions such as Mg^{2+} and anions such as Cl^{-} [113, 3].

(iv) the non-canonical structural elements of RNA, i.e., loops, hairpins, bulges, pseudoknots, mismatched, etc., challenges the current force fields [114]. More simulations on these structures are needed for a careful comparison with available experimental or quantum mechanical data.

Insufficient Sampling

Current computer power allows performing MD simulations up to microsecond timescale for small proteins [115, 116]. Plenty of MD simulations of RNA have also reached time length of several hundreds of nanoseconds. However, timescale of real events is still much longer than affordable simulation timescale. To overcome this problem several efforts have been spent in the last decades on developing accelerated sampling methods [117, 10, 11, 12, 13]. Longer simulations and efficient sampling are also crucial for force field testing and validation.

1.5 Thesis Organization

In Chapter 2, we present the computational methods used in this thesis, which include standard MD simulation, bidirectional SMD simulation, and free-energy reconstruction from nonequilibrium works. Here, we also propose a reweighting method to project the free energy on any *a posteriori* chosen CV.

In Chapter 3, we introduce our new electrostatic-based CV, which is an approximation of the electrostatic free energy given by the Debye-Hückel formalism.

In Chapter 4, we present results from ~0.5 μs obtained by nine standard MD simulations starting from the NMR structure of the L22-TAR complex. Besides reproducing the NMR experimental features of this system and confirming the well-known role of electrostatic interactions during peptide/RNA binding events, these simulations motivated us to design and implement the CV introduced in Chapter 3.

In Chapter 5, we present results from ~4.2 μs bidirectional SMD simulations of TAR and L22 binding/unbinding events using our new CV. Besides reproducing the reported NMR binding pose, we found a new binding pose in the same TAR pocket.

Chapter 6 contains the conclusions of the thesis and the perspectives in which we present the preliminary results of well-tempered metadynamics simulations applied on our proposed CV as a complementary approach supporting the SMD findings described in Chapter 5.

Chapter 2

Computational Methods

Contents

2.1 Overview

Computer simulation (referred to as simulation hereafter) is the science of modeling a real or theoretical system, executing the model on a computer, and analyzing the execution output. Starting from the early 1950s, simulations have grown hand-in-hand with the fast development of computer performance. Nowadays, simulations have become an important discipline in physics, chemistry, and biology.

Science is about both observation and comprehension. Science is incomplete until observations are fully comprehended [118]. Traditionally, observation is provided by experiments and comprehension is based on theories. Simulations represent a new scientific methodology, forming a triangle-like relationship with theory and experiment. On the one hand, simulations act as computer experiments to validate theories. That is when an experimental probe is out of reach; e.g., studying of truly isolated systems, extreme temperature and pressure conditions, subtle details of molecular motion and structure such as fast ion conduction or enzyme action, etc. [119, 120]. On the other hand, simulations help interpreting experimental observations. That is when theories fail to provide an exact analytical solution or derivation due to the complexity of the system (e.g., liquid, imperfect gases, macromolecules, etc.) and therefore has to rely on one or more approximation schemes [119]. Simulations have a valuable role in such cases by providing essentially exact results that can be compared with or interpret experimental results. Therefore, simulations have been intensively used in material and biophysical sciences to study dense molecular systems as important counterparts of experiments.

M*olecular* **D***ynamics* (**MD**) simulations [121] are among the most commonly used techniques in biomolecular studies. Given nowadays computer power, atomistic MD simulations can explore up to microsecond timescale for small protein systems [115, 116]; however, most important biological processes happen at the millisecond or even second timescales. With these technical limitations, current MD simulations mostly provide information around local equilibrium states of biomolecular systems. There are, however, important processes involving transition between states or conformations such as biomolecule-ligand binding/unbinding. MD simulations, in such cases, require large computational resources to bring the system out of a local equilibrium. Several enhanced sampling techniques have been introduced to accelerate the state transition such as **S***teered* **M***olecular* **D***ynamics* (**SMD**) [10, 12] and Metadynamics [11]. These methodologies are based on MD simulations and accelerate the transition by adding an external force or a bias potential to help the system escape from a local trapping free-energy minimum.

In this chapter, we will introduce the basic concept and methodology of MD simulations, followed by a description of SMD.

2.2 Molecular Dynamics Simulations

This section describes the basic concepts and theoretical background associated with MD simulations. It also contains common methods to validate the structural and dynamical properties from MD simulations against NMR data.

2.2.1 MD Methodology

2.2.1.1 MD Algorithms

Classical MD simulations are based on three approximations

(i) *Born-Oppenheimer approximation* [122] enables separating the electronic and nuclear motions. Such assumption is valid due to the much lighter weight of an electron compared to the nucleus of an atom.

(ii) *Adiabatic approximation*, originally stated by Born and Fock in form of *adiabatic theorem* [123], This approximation assumes that the electrons adjust rapidly to any nuclear displacement and thus they always remain in their instantaneous eigenstate (e.g., ground state) if the nuclear displacement is slow enough and if there is a gap between the eigenvalue and the rest of the spectrum.

(iii) Nuclei are presumably heavy enough to be considered *classical objects*, and hence their movements can be described solely by Newton's equations of motion. However, this assumption is generally not valid for hydrogen atoms. Its use is only justified by two facts: it is computationally expensive to include quantum nuclear effects and these effects are already taken care of by the *empirical force fields* which are parametrized by quantum calculations and carefully compared to experimental data (see Section 2.2.1.3 for more details).

As a consequence of these approximations, in classical MD simulations, the energy of a molecule can be considered a function of the nuclear coordinates only. Such simplification is useful when there is no quest for electronic properties. The movement of an atom is then described by numerically integrating the Newton's equation of motion

$$\mathbf{f} = m\ddot{\mathbf{r}}, \tag{2.2.1}$$

where $\mathbf{f}$ is the total force acting on the atom, m is the atomic mass, and $\ddot{\mathbf{r}}$ is the acceleration caused by the force $\mathbf{f}$. Integrating Equation (2.2.1) requires the knowledge of initial atomic coordinates and velocities.

One of the simplest algorithms used to integrate the equations of motion (2.2.1) is the so-called *Verlet algorithm* [124]. This algorithm is simply based on a Taylor expansion of coordinate $\mathbf{r}$ around time t, which yields

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - r(t-\Delta t) + \frac{\mathbf{f}(t)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4). \tag{2.2.2}$$

Velocity is then computed based on the knowledge of the trajectory

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2). \tag{2.2.3}$$

There are several other algorithms equivalent to the Verlet scheme, among which, the simplest is the so-called *leap-frog algorithm* [125]. This algorithm evaluates the velocities at half-integer time step. The new positions are then updated from these velocities

$$\mathbf{v}(t+\frac{\Delta t}{2}) = \mathbf{v}(t-\frac{\Delta t}{2}) + \frac{f(t)}{m}\Delta t, \tag{2.2.4}$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t+\frac{\Delta t}{2})\Delta t. \tag{2.2.5}$$

Leap-frog is the algorithm employed in the MD simulations of this thesis.

Verlet and leap-frog algorithms can be recovered from each other, therefore they produce identical trajectories and accuracies. These algorithms are preferred in most MD simulations since they share the same time-reversibility with the Newton's equations of motion.

2.2.1.2 Statistical Ensembles

Statistical mechanics provides a *linkage* between the *microscopic world* such as the atomistic information of a system described by the Newton's equation of motion (i.e., Equation (2.2.1)) and the *macroscopic observables* such as thermodynamic, structural, and dynamical properties of the same system. Such a linkage involves the concept of *ensemble*, which was originally introduced by Gibbs in 1876 [126]. An ensemble is an imaginary collection of systems that

(i) share the same set of macroscopic properties (e.g., total energy E, number of particles N, volume V, pressure P, temperature T, chemical potential μ, etc.) and

(ii) can be described by the same Hamiltonian or the same set of microscopic laws of motion with different initial conditions so that each system has a *unique* microscopic state at a given instant of time.

In other words, there are many microscopic configurations of a system which lead to the same macroscopic properties. Once an ensemble of these configurations is defined, the macroscopic observables are calculated by performing *ensemble averages* over all microscopic configurations. The "general" form of an ensemble average is heuristically given by [127]

$$\langle A \rangle \equiv \frac{\sum_i A_i \rho_i}{\sum_i \rho_i}. \tag{2.2.6}$$

Here $\langle A \rangle$ represents the ensemble average of the system property A, i indexes the microstate, A_i is the value of property A measured when the system is in microstate i, and ρ_i is the probability of observing the system in the microstate i. Note that the discretized phase space is used for notational convenience and continuous case can be easily generalized.

The probability distribution function ρ is dependent on the type of ensemble. The simplest and most fundamental ensemble is that of an isolated system with constant N, V, and E, which is also referred to as the NVE ensemble or the *microcanonical* ensemble. The phase space distribution of an NVE ensemble is uniform over the constant energy hypersurface E and zero otherwise.

However, NVE ensemble involves perfectly isolated systems with constant total energies. This condition cannot be achieved in real-life experiments. It is thus important to define other ensembles that are more practical and are able to reflect the common experimental setups. Those include the *canonical* ensemble (NVT), the *isothermal-isobaric* ensemble (NPT), and the *grand canonical* ensemble (μVT).

In the following, we present the probability distribution function for NVT and NPT ensembles, which represent the most commonly performed conditions in experiments. We also briefly introduce how to perform MD simulations on these two ensembles.

Canonical Ensemble The condition of canonical ensemble, or NVT ensemble, is achieved by coupling the system with an infinite external heat bath. When the system is in thermal contact with the heat bath, its energy is allowed to fluctuate such that its temperature remains unchanged. Notably, the extended system which is composed of the system and the heat bath can be considered in a microcanonical formulation.

The probability distribution function of NVT ensemble is given by the Boltzmann's distribution (or Boltzmann's law), also called Gibbs' distribution [128]

$$\rho_i^{(NVT)} = \frac{\mathrm{e}^{-\beta E_i}}{Z^{(NVT)}}. \tag{2.2.7}$$

Here the probability is characterized by the probability $\rho_i^{(NVT)}$ of finding the system in a *microstate* with energy E_i; $\beta = 1/k_B T$, where k_B is the Boltzmann constant; $Z^{(NVT)}$ is a normalization factor given by

$$Z^{(NVT)} = \sum_i \mathrm{e}^{-\beta E_i}. \tag{2.2.8}$$

$Z^{(NVT)}$ is also called the *partition function* of the canonical ensemble.

To perform MD simulation of a system characterized by an NVT ensemble, we need to mimic the effect of the heat bath. Several methods to obtain such a *thermostat* have been proposed including those by Andersen in 1980 [129], Berendsen et al. in 1984 [130], Nosé in 1984 [131], Hoover in 1985 [132], and Bussi et al. in 2007 [133].

Isothermal-Isobaric Ensemble Isothermal-isobaric ensemble, or NPT ensemble, is one of the most important ensembles due to its close reflection of the most commonly used experimental conditions. Such a condition is achieved by concurrently coupling the system with a heat bath and an imaginary "piston" to maintain a fixed temperature and pressure. Consequently, we must allow the volume of the system to fluctuate. Therefore, the probability distribution function of NPT ensemble must include volume V as its variable and is hence given by [127]

$$\rho_i^{(NPT)} = \frac{\mathrm{e}^{-\beta(E_i+PV_i)}}{Z^{(NPT)}}. \tag{2.2.9}$$

The normalization factor $Z^{(NPT)}$ is then given by

$$Z^{(NPT)} = \sum_i \mathrm{e}^{-\beta(E_i+PV_i)}. \tag{2.2.10}$$

The strategies to achieve the isobaric condition of the NPT ensemble, which are now widely referred to as *barostats*, were first introduced by Andersen [129] and later generalized by Parrinello and Rahman [134].

2.2.1.3 Force Fields

A force field is defined as a functional form and parameter sets describing the potential energy of a system of interacting particles, from which the force acting on a single atom is calculated at each time step of an MD simulation:

$$\mathbf{f}_i = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i}, \tag{2.2.11}$$

where $V(\mathbf{r})$ denotes the potential energy. The most common functional form of a force field is a simple additive four-term expression quantifying the intra- and inter-molecular interactions [135]:

$$
\begin{aligned}
V(\mathbf{r}) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) \\
+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\},
\end{aligned}
\tag{2.2.12}
$$

The first three terms in Equation (2.2.12) represents the bonded interactions in a molecular system including the bond stretching, angle bending, and bond rotating. Displacement of bond lengths and angles from their equilibrium values (i.e., the bond length l_i deviating from its equilibrium value $l_{i,0}$ in the first term and valence angle θ_i varying from the its equilibrium value $\theta_{i,0}$ in the second term) causes energetic penalties, which are modeled with a harmonic potential form. The third term is a torsional potential quantifying the energy changes associated with bond rotations. The fourth term represents the non-bonded interaction between all pairs of atoms in the same or different molecules and separated by at least three bonds. Non-bonded interaction includes electrostatic and van der Waals interactions which are commonly described by a Coulomb potential and a Lennard-Jones potential, respectively. Various constants k_i, V_n, ϵ_{ij}, σ_{ij} in Equation (2.2.12) characterize the atoms and their unique behavior when interacting with other specific atoms. Such constants of various atoms form the force field parameter sets.

Finding parameter sets for a certain functional form of force fields is not a trivial task. A "good" parametrization has to ensure that calculations can produce appropriate molecular structures and interaction energies. As reference for force field parametrization, there is a wide class of experimental and quantum mechanics computational data. Therefore, an important characteristic of MD force fields is that they are all "empirical"; there is no "correct" force fields with respect to both functional form and parameter set. Indeed, force field only provides an estimation of the true underlying interaction energy, which controls all molecular behavior. Therefore, the accuracies of MD simulations depend severely on the quality of the chosen force field. Despite a significant amount of effort has been spent on force-field development, the force fields in use today still suffer from inaccuracies.

In MD simulations, the force field is chosen according to the aim of simulation and the properties of the systems. For RNA systems, AMBER force fields are among the most widely used. In AMBER force fields, base stacking and hydrogen bonding parameters are sufficiently well described [136, 137]. Additionally, AMBER utilizes partial atomic charges to calculate the electrostatic field around the monomers, which reproduces the molecular interactions of nucleobases quite well [113]. Recently, other refinements have been made on the standard AMBER99 force field to improve the description of RNA backbone dihedral angles α, γ, and χ [6, 138, 7]. With these improvements, stability of RNA backbones in MD simulations has been confirmed in rather long timescales of hundreds of nanoseconds [139]. Despite good performance in major aspects, AMBER force field has also few significant deficiencies, including:

(i) the flexibility of RNA sugar-phosphate backbone challenges the description of electrostatic potential which is currently based on a constant-point-charge model [110, 111].

(ii) polarization effects are still neglected by AMBER force fields for RNA

(iii) as a consequence of the neglected polarization, the description of anions and divalent cations such as Mg^{2+} is also a big challenge for the force field [113, 3].

Besides these explicit deficiencies, other subtle RNA-force-field issues caused by the limitations of modeling have been also reported for: *(i)* description of single-stranded RNA hairpin segments [7] and *(ii)* dependence on the choice of salt strength [140].

2.2.1.4 Periodic Boundary Conditions

P*eriodic* **B***oundary* **C***onditions* (**PBC**s) extend a finite system to an infinite and continuous one. Therefore they can be suitably used in biomolecular simulations. Indeed, *in vivo* and *in vitro* biomolecules are surrounded by a relatively infinite water medium. However, doing simulations of biomolecules in a large water box to mimic the infinite condition is simply unaffordable and impractical. PBCs become useful in such a case. They create repeatable regions which are (small) boxes of water with molecules immersed inside. When a molecule moves in the central box, its periodic image in every one of the other boxes moves exactly in the same way. PBCs thus ensure that when an atom moves off the edge, it reappears on the other side; or in other words, there is no wall at the boundary. With PBCs, the system is infinite, continuous, and has no surface.

Besides the above-mentioned advantages, PBCs create artifacts as well. One of the most severe PBC problems is the artificial interaction between the molecules and their surrounding images. In systems of highly charged molecules such as DNAs and RNAs, this problem becomes serious because of the spurious electrostatic interaction between the charged molecules and their periodic images. To adjust this side effect, we have to “add” extra *neutralizing ions* to the simulation box. These neutralizing ions together with the *buffer ions* from experiment are distributed around the molecules and hence create a screening effect which helps reducing the spurious electrostatic interaction between periodic images.

The solution of adding more ions could help decreasing the artificial electrostatic interactions but in turn creates other side effects. In fact, these computational neutralizing ions are conceptually different from the experimental buffer ions.

2.2.1.5 Long-Range Interactions

Coulomb electrostatic interaction possesses a long-range feature that poses many challenges to current MD simulations and needs a special treatment.

Mathematical and computational problems of long-range electrostatics Let us consider a molecule containing N charged atoms in a cubic box with diameter L ($V = L^3$). PBCs are applied in all directions. The Coulomb interaction energy is then given by

$$E = \frac{1}{2} \sum_{i=1}^{N} q_i \phi(\mathbf{r}_i), \tag{2.2.13}$$

where

$$\phi(\mathbf{r}_i) = \sum_{\mathbf{n}} \sum_{j=1\,(j \neq i)}^{N} \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}L|} = \sum_{\mathbf{n}} \sum_{j=1\,(j \neq i)}^{N} \frac{q_j}{|\mathbf{r}_{ij} + \mathbf{n}L|} \tag{2.2.14}$$

is the electrostatic potential at the position of atom i due to the contributions of all other atoms and $\mathbf{n} = (n_x, n_y, n_z)$ $(n_i \in \mathbb{Z})$ is the box index vector. The sum over $\mathbf{n}$ represents the effect of periodic boundary condition, atom i interacts not only with atom j at $\mathbf{r}_j$ but also with all images of j at $\mathbf{r}_j + \mathbf{n}L$. Theoretically, Equations (2.2.13) and (2.2.14) represent a well-defined electrostatic problem: given all charges q_i and their positions $\mathbf{r}_i$, we can compute the electrostatic interactions. In practice, such computation is not trivial, due to two main problems:

Mathematical problem. The sum in Equation (2.2.14) is not an absolutely convergent series, but rather a conditionally convergent one. The result depends on the order in which we sum up the terms. A natural choice is to take boxes in roughly spherical layers. This choice leads to a slow mathematically conditional convergence.

Computational problem. Equation (2.2.13) is a sum over $N(N-1)/2$ pairs, thus scales as $\mathrm{O}(N^2)$. Biomolecules may contain from a few tens to millions of atoms. The computational costs to perform electrostatic calculations in such systems become too expensive.

Approaches to solve long-range electrostatic problem This part discusses some solutions for the above-mentioned problems.

Cut-off methods. There is a historical scheme of methods in which electrostatic interactions are simply ignored beyond a certain cut-off. However, there is a tradeoff: a long cut-off is computationally expensive, while a short cut-off gives rise to inaccuracies.

Reaction field methods [141, 142]. In these methods, each atom is surrounded by a cut-off sphere. Within this sphere, the interactions with other atoms are described explicitly.

The space outside the sphere is treated as a homogeneous dielectric medium with a certain permittivity and ionic strength. The computational cost of this approach is slightly higher than that of the cut-off methods, but the accuracies are greatly improved.

Ewald summation method [143]. This method is in the group of more reliable lattice summation techniques. Such schemes are more expensive than simple truncations and reaction field methods. They are also more advanced as they respect the long-range character of the interactions. The idea of the Ewald method is to convert the single slowly and conditionally convergent summation (i.e., Equation (2.2.13)) into two quickly convergent terms: *(i)* a short-range term which sums accurately and quickly in real space and *(ii)* a long-range term which is a smoothly varying term that sums quickly in Fourier space [144]. As the computational effort of the Ewald summation scales as $\mathcal{O}(N^{3/2})$, this approach is still expensive for large systems [145].

Particle-Mesh Ewald method [146]. This approach is an alternative of the Ewald summation method. In this approach, the distribution of all system charges is mapped on a grid by B-spline interpolation; this grid is then Fourier transformed by a single operation instead of the N^2 operations required in Ewald method. If **F***ast* **F***ourier* **T***ransform* (**FFT**) algorithms [147] are applied, the reciprocal space calculation is then reduced to $\mathcal{O}(N \log(N))$.

2.2.2 Validation of MD Simulations Against NMR Relaxation Measurements

Due to the strong dependence of MD simulations on force fields and simulation protocols, the validation of simulations against experimental data is critically important [148]. On one hand, NMR relaxation yield dipolar correlation function, from which dynamical quantities such as generalized parameter S^2 can be extracted [149, 150]. On the other hand, this quantity can be actually calculated readily from MD trajectories [119]. Additionally, NMR relaxation measurements provide **N***uclear* **O***verhauser* **E***nhancement* (**NOE**) data which are used to derive structural information such as interatomic distances of the measured system. These interatomic distances can also be measured directly from MD trajectories and compared with experimental ones. These two common comparisons, referred to as *dynamical* and *structural* validations, are discussed in the two following sections.

2.2.2.1 Dynamical Validation – Order Parameter S^2

Order parameter S^2 is a measure for the spatial restriction of motion. In this section, we briefly summarize the *model free* approach [149, 150] to extract S^2 from NMR relaxation experiments.

Experimental derivation The NMR relaxation due to dipole−dipole interaction between two nuclei, one being a heavy atom X (e.g., ^{15}N or ^{13}C) and the other being its bonded hydrogen, can be described by the correlation function [151]

$$C_{X-H}(t) = \langle P_2(\hat{\mu}_{X-H}(t) \cdot \hat{\mu}_{X-H}(0)) \rangle, \tag{2.2.15}$$

where $\hat{\mu}_{X-H}(t) = (\mathbf{r}_H(t) - \mathbf{r}_X(t))/|\mathbf{r}_H(t) - \mathbf{r}_X(t)|$ is the unit vector pointing along the X−H bond at time t; $P_2(x) = 1/2(3x^2 - 1)$ is the second order Legendre polynomial; and $\langle\ \rangle$ denotes the equilibrium average.

Model free approach assumes that the overall molecular motion is *isotropic* and hence can be *adiabatically separated* from the internal motion [149, 150]. Based on this assumption, the correlation function in Equation (2.2.15) can be factored as

$$C_{X-H}(t) = C_0(t) C^I_{X-H}(t), \tag{2.2.16}$$

where $C_0(t)$ describes the overall motion and $C^I_{X-H}(t)$ is the correlation function for internal motion. For overall isotropic motion, $C_0(t)$ is rigorously given by

$$C_0(t) = \frac{1}{5} e^{-t/\tau_c}, \tag{2.2.17}$$

with the rotational correlation time τ_c proportional to the inverse of the rotation diffusion constant.

Now if we define a generalized order parameter S^2 as $S^2 = \lim_{t\to\infty} C^I_{X-H}(t)$, then in the model free approach the internal coordination function can be truncated and written in terms of the order parameter S^2 and the effective (or internal) correlation time τ_e as

$$C^I_{X-H}(t) = S^2 + (1 - S^2) e^{-t/\tau_e}. \tag{2.2.18}$$

The value of S^2 ranges from 0 to 1. If $S^2 = 0$, the motion of the two atoms with respect to each other is not restricted in any way; and if $S^2 = 1$, the interatomic vector $\hat{\mu}_{ij}$ is rigidly fixed in the molecular frame.

Order parameter validation of MD simulations In computational studies, the internal correlation function $C^I_{X-H}(\tau_r)$ at time interval τ_r can be calculated directly from the trajectories $\mathbf{r}_i^{(t_j)}$ of coordinates $\mathbf{r}_i$ and time t_j as

$$C^I_{X-H}(\tau_r) = \langle P_2((\mathbf{r}_X^{(t_j)} - \mathbf{r}_H^{(t_j)}) \cdot (\mathbf{r}_X^{(t_j+\tau_r)} - \mathbf{r}_H^{(t_j+\tau_r)})\rangle_{t_j}. \tag{2.2.19}$$

The order parameter S^2 and its error are then calculated over the time beyond the internal correlation time scale τ_e as

$$S^2 = \langle C^I_{X-H}(\tau_r) \rangle_{\tau_r > 3\tau_e}, \tag{2.2.20}$$

and

$$\sigma^2_{S^2} = \langle (C^I_{X-H}(\tau_r) - S^2)^2 \rangle_{\tau_r > 3\tau_e}. \tag{2.2.21}$$

The condition $\tau_r > 3\tau_e$ should hold to ensure that each time interval of evaluating the internal correlation function between two bonded atoms can cover all internal motions of the bond. It follows that τ_r should be larger than 500 picoseconds to satisfy this condition for most chemical groups. In practice, the typical length of $C_{X-H}(\tau_r)$ is about several nanoseconds up to half of the simulation time length [119]. In the case of nucleic acid studies, order parameter S^2 are often calculated for the ribose C1'−H1' and the base C6−H6 or C8−H8 dipoles.

2.2.2.2 Structural Validation − NOE Data

The nuclear Overhauser effect is the primary source for solving NMR structures. It provides geometric information to build the three-dimensional structure of the measured object.

Experimental observables The spectral density is defined as the cosine Fourier transform of the correlation function in Equation (2.2.15), i.e., as

$$J(\omega) = 2\int_0^\infty C(t)\cos(\omega t)dt. \tag{2.2.22}$$

In the model free approach, $J(\omega)$ is given by

$$J(\omega) = \frac{2}{5}\Big(\frac{S^2\tau_c}{1+\tau_c^2\omega^2} + \frac{(1-S^2)\tau}{1+\tau^2\omega^2}\Big), \tag{2.2.23}$$

with $\tau^{-1} = \tau_c^{-1} + \tau_e^{-1}$. The spectral density functions are then used to compute the NMR relaxation parameters including the longitudinal (T_1) and transverse (T_2) magnetization transfer and the heteronuclear NOE between the heavy atom X and its bonded H [152] as

$$T_1 = d_{00}[3J(\omega_X) + J(\omega_{H-X}) + 6J(\omega_{X+H})] + c_{00}\omega_X^2 J(\omega_X), \tag{2.2.24}$$

$$T_2 = \frac{1}{2}d_{00}[4J(0) + 3J(\omega_X) + J(\omega_{H-X}) + J(\omega_H) + 6J(\omega_{X+H})] + \frac{1}{6}c_{00}\omega_X^2[4J(0) + 3J(\omega_X)], \tag{2.2.25}$$

and

$$\text{NOE} = 1 + \frac{\gamma_H}{\gamma_X}d_{00}T_1[6J(\omega_{X+H}) - J(\omega_{H-X})], \tag{2.2.26}$$

in which

$$d_{00} = \frac{\hbar}{20}\left(\frac{\mu_0}{4\pi}\right)^2\gamma_X^2\gamma_H^2\frac{1}{r_{XH}^6}; \tag{2.2.27}$$

$$c_{00} = \frac{1}{15}\Delta\sigma^2; \tag{2.2.28}$$

μ_0 is the vacuum permeability; $\gamma_{X,H}$ are the gyromagnetic rations; ω_X, ω_H, ω_{X+H}, and ω_{H-X} are the Larmor frequencies; and $\Delta\sigma$ is the chemical shift anisotropy of X.

In the *initial rate* approximation, NOE is proportional to the sixth root of the distance between two atoms, i.e., NOE $\propto r_{XH}^{-6}$. From this set of observables, interatomic distances are obtained and three dimensional structure is built up.

Validation of MD studies In principle, relaxation parameters T_1, T_2, and NOE can be directly calculated from simulation results and compared to those from experiments. However, in practice, checking the violation of interatomic distances in simulations with respect to the distances obtained from NOE data is a more widespread protocol.

2.3 Steered Molecular Dynamics Simulations

Single-molecule force-probe experiments (e.g., **A***tomic* **F***orce* **M***icroscopy* (**AFM**), **L***aser* **O***ptical* **T***weezers* (**LOT**), etc.) enable the characterization of biomolecules in response to mechanical force, which in turn reveal mechanical properties and functions of the biomolecules. However, in these experiments, the underlying atomistic dynamics and interactions that give rise to molecular mechanisms cannot be disclosed. This is the main motivation of reproducing these experiments *in silico* by means of atomistic simulations that are now widely referred to as **S***teered* **M***olecular* **D***ynamics* (**SMD**) simulations. SMD simulations involve not only applying external forces to manipulate biomolecules for the purpose of exploring their responses and functions, but also accelerating processes that are unaffordable by means of standard MD simulations. SMD has become a powerful *in silico* tool in guiding, complementing, and explaining *in vitro* single-molecule force-probe experiments.

In this section, we first discuss some historical facts and events related to the SMD method. We then present the common constant-velocity SMD algorithm.

2.3.1 Historical Background

2.3.1.1 Single-Molecule Experiments *In Vitro*

Force probe experiments were initiated in 1994 by Florin, Moy, and Gaub [153]. In this AFM experiment, biotin was extracted from its complex with streptavidin and the unbinding force was measured. This was reported as the first force measurement of individual ligand-receptor pairs.

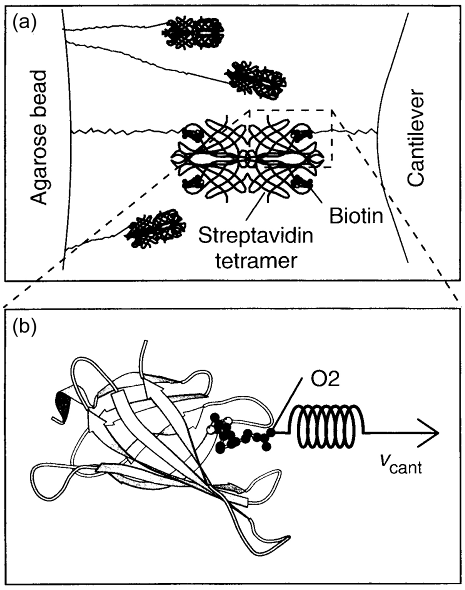

Figure 2.3.1: Setups of a single-molecule experiment *in vitro* (a) and *in silico* (b). Figures are reproduced from ref. [10].

Figure 2.3.1a illustrates the experiment by Florin et al. [153], in which the cantilever of the AFM microscope acted as a force sensor. A polymer linker connected the avidin (or streptavidin) molecules with a surface (on the left). At the same time several biotin molecules were also connected through another polymer linker with the tip of the cantilever (on the right). When the cantilever approached the surface, several biotin-streptavidin complexes formed. As the cantilever retracted, the complexes dissociated one after the other. Occasionally, one single complex remained until the very end of the experiment. In such a case, the authors measured the force required to dissociate this last complex by observing the jump of the deflection of the cantilever to zero. On a microsecond timescale, the rupture was dominated by thermal fluctuations and hence has a probabilistic feature. Consequently, repeating the experiment several hundreds of times revealed a distribution of forces, or a histogram of forces. The maximum in the histogram indicates the most probable dissociation force, which is the force required to break the binding between a ligand and a receptor, and thus represents the binding strength.

The interaction between biotin and streptavidin is one of the strongest noncovalent interactions in nature with the binding free energy of 22 kcal/mol. The force probe AFM experiment by Florin et al. was able to reproduce this binding free energy. However, the experiment was unable to reveal the underlying atomistic dynamics and interactions between biotin and streptavidin and thus the binding/unbinding mechanism remained unknown.

2.3.1.2 Single-Molecule Experiments *In Silico*

Motivated by the force probe experiment of Florin et al. [153] and the quest of molecular mechanism, the first computer simulation modeling such experimental set-up was introduced two years later (i.e., in 1996) by Grubmüller, Heymann, and Tavan [10]. In this simulation, the authors modeled the effect of the cantilever by a symbolic “spring” (see Figure 2.3.1b) which caused an additional harmonic steering potential:

$$V_{spring} = k_0[z_{O2}(t) - z_{cant}(t)]^2/2, \tag{2.3.1}$$

acting on the z coordinate of atom O2 (z_{O2}) of biotin molecule. In Equation (2.3.1), k_0 is the spring constant and $z_{cant}(t)$ denotes the cantilever position at which the spring potential is centered:

$$z_{cant}(t) = z_{cant}(0) + v_{cant}t, \tag{2.3.2}$$

here v_{cant} modeled the velocity of the cantilever. During the simulation, V_{spring} was shifted on z direction as the free end of the spring moved with velocity v_{cant}. This ensured that in the simulation atom O2 was subjected to the same force as in the experiment, in which the same atom was covalently connected to the cantilever through a polymer linker.

The *force probe simulation* (as called by the authors) allowed determining the dissociation force from the force profile[1]. Besides, this simulation provided information on atomistic interactions and how they broke during the dissociation. These insights are inaccessible from experiments. However, the simulation was performed in nanosecond timescale, which is many orders of magnitude faster than the experimental time of milliseconds.

[1] Note that in force probe experiment, the dissociation force is determined from the force histogram.

In the next year (1997), Schulten and co-workers performed a series of "steering" simulations on the same biotin-streptavidin complex system with the same spirit of modeling the cantilever as a symbolic spring [154]. The difference was that the authors modeled a spring with a varying harmonic restraint coefficient:

$$k_t = \alpha t. \tag{2.3.3}$$

Eight simulations with different values of the rate α were performed to explore the dependence of unbinding on the speed of rupture. Since the name *Steered Molecular Dynamics simulation* later came from Schulten group, it is fair, for historical reason, to mention this work as their first attempt to perform *in silico* single-molecule experiment. However, their first published work on this technique had negative results when compared to the AFM experiment by Florin et al. [153]. This work contained several severe problems, one of which was that the simulations were performed in vacuum, which, according to the authors, "did not affect the actual binding pocket" [154].

Since the first attempts to model AFM experiments, SMD has grown to be a complete computational methodology that helps not only explaining but also complementing and guiding experiments. Such expansion is based on two foundations: *(i)* the blooming development of theories on nonequilibrium process starting by the famous Jarzynski's equality in 1997 and *(ii)* the advantage of molecular simulation that allows one to apply more complex forces than can be performed in AFM experiment. Details of the first issue is discussed in section 2.4. With regard to the second issue, indeed, the force in simulation can be applied on a group of atoms if necessary instead of only one atom as in experiment; the change of directions is easily permitted in simulation while more complicated to be achieved in experiment; simulation also allows applying the force to nonlinear and subtle coordinates, or **C***ollective* **V***ariables* (**CV**s), that may be more efficient in capturing the nature of the conformational transitions than the distance used in experiment. The latter is discussed in chapter 3.

2.3.2 Steered Molecular Dynamics Algorithm

There are two common SMD protocols: constant force and constant velocity. In constant-force SMD simulation, a force is *directly* applied to one or more atoms and atomic displacement is then monitored. A variation of this scheme is to apply customized time-dependent forces. In constant-velocity SMD simulation, a moving harmonic potential modeled as a symbolic spring is used to cause atomic motion along the CV. The free end of the spring moves with a constant velocity. The atoms attached to the other end are subject to a steering force, which can be evaluated by the spring extension. Here we employ the constant-velocity protocol for our SMD simulations.

In constant-velocity SMD algorithm, the CV z acting on one atom or a group of atoms is restrained to a *symbolic point* which is initially positioned at z_0, say the *restraint point*, by a *symbolic spring* whose stiffness is k. The restraint point is then pulled in the direction of the chosen CV with a constant velocity v, acting a harmonic potential V on the CV,

$$V(t) = k(z(t) - z_{restraint}(t))^2/2 = k(z(t) - vt - z_0)^2/2. \tag{2.3.4}$$

The external force or steering force exerted on CV z can be expressed as

$$F(t) = k(z_0 + vt - z(t)). \tag{2.3.5}$$

Under this force, the CV changes its value. Through the response of the CV to the steering force at each timestep, the molecular system also adapt to a new conformation defined by the newly adopted CV value.

The cumulative external work at time t can be calculated as

$$W(t) = \sum_{t'=0}^{t} vk(z(t') - vt' - z_0)\Delta t'. \tag{2.3.6}$$

Note that the time is discretized for notational convenience. This is applicable for MD-based simulations.

2.4 Reconstruction of Free Energy from Nonequilibrium Works

Steering is a nonequilibrium process in which the system is driven away from its equilibrium states. For such a process in microscopic scale, the second law of thermodynamics states that the average work exceeds the free-energy difference between the initial and final equilibrium states,

$$\langle W \rangle \geq \triangle F, \tag{2.4.1}$$

here the bracket $\langle\ \rangle$ denotes the average over a statistical ensemble of realizations. The equality ($\langle W \rangle = \triangle F$) only holds if the steering process is *reversible*, i.e., the steering speed is *infinitely slow*. In physically realistic situations, all thermodynamic processes are irreversible, i.e., happen at finite rates. The difference $\langle W \rangle - \triangle F$ is referred to as the wasted work or dissipated work associated with the entropic increment during an irreversible process. It is not trivial to quantify such an entropic change and thus challenging to recover the free-energy difference from the nonequilibrium works over all realizations of a given thermodynamic process.

In 1997, Jarzynski discovered a renowned equality which permits the reconstruction of free energy, an equilibrium property, from the nonequilibrium works done on a system. This is the pioneering foundation for the discovery of nonequilibrium work relations and **P***otential of* **M***ean* **F***orce* (**PMF**) estimators that allow obtaining equilibrium properties of a system by observing how the system responds when driven away from equilibrium. These powerful theoretical foundations have opened big avenues for both experimental and computational applications.

In this section, we first review and rederive in a unified framework the methods for reconstructing free energy from both *in vitro* and *in silico* nonequilibrium processes, including:

(i) the nonequilibrium work relations, namely the *Jarzynski's equality* and *Crook's fluctuation theorem*. These relations allow recovering free-energy differences between states as a function of the *restrained CV* of an *extended system* (i.e., the system coupled with the external perturbation),

(ii) the **B***ennett* **A***cceptance* **R***atio* (**BAR**) that, in a similar spirit to the Crook's fluctuation theorem, estimates the free energy of the extended system utilizing the works from both forward and backward processes. However, in advance to Crook's theorem, the free energy from BAR method maximizes the chance of observing these work values.

(iii) the PMF estimators proposed by *Hummer and Szabo* for unidirectional realizations and proposed by *Minh and Adib* for bidirectional cases. These estimators can recover the PMF as a function of the *unrestrained CV* of the *unperturbed system* (i.e., in the absence of external potential).

Next we propose our reweighting method that allows projecting the free energy profile on any *a posteriori* chosen CV that is not necessary to be the steered CV.

2.4.1 Nonequilibrium Work Relations

Among the most general and widely used nonequilibrium work relations are the Jarzynski's equality [155] and Crook's fluctuation theorem [156]

(i) Jarzynski's equality is expressed as

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \mathrm{e}^{-\beta \Delta F}, \tag{2.4.2}$$

here the bracket $\langle\ \rangle$ denotes the average over a statistical ensemble of realizations; the factor $\beta = 1/k_B T$ is the inverse temperature. The critical assumptions leading to this equality include *(i)* the evolution of the system must be *Markovian* and *(ii)* the process must satisfy *detailed balance* conditions for each of the values taken by the external control parameter[2] during the switching process[3]. Jarzynski's equality basically states that the free-energy difference between two equilibrium states of a system can be accurately recovered from the exponential average of the nonequilibrium works performed on *switching* the system from one equilibrium state to the other. Independently of path and rate of this thermodynamic process, Jarzynski's equality puts a strong constraint on the distribution of the works which remains valid even if the system is steered away from its thermal equilibrium. Jarzynski's equality has a wide range of applications since it allows the determination of equilibrium free-energy difference and hence several other equilibrium properties of a system from monitoring its response in nonequilibrium processes.

(ii) Crook's fluctuation theorem is formulated as

$$\frac{\rho_F(+W)}{\rho_B(-W)} = \mathrm{e}^{\beta(W-\Delta F)}, \tag{2.4.3}$$

here $\rho_F(W)$ and $\rho_B(-W)$ denote the work distributions in the forward and backward processes respectively. It is noteworthy that a backward process can be considered the reverse of a corresponding forward process where the work takes the opposite sign. Crook's theorem implies that $\rho_F(W)$ and $\rho_B(-W)$ meet at $W = \Delta F$. The assumptions of Jarzynski's equality regarding Markovian and detailed balance conditions also hold for Crook's theorem. Moreover, Crook's theorem can be rearranged, i.e., by multiplying both sides of Equation (2.4.3) with $\rho_B(-W)\mathrm{e}^{-\beta W}$ and then integrating over W, to obtain Jarzynski's equality.

There have been a number of derivations of Jarzynsky's equality including the pioneering proofs of Jarzynski himself, e.g., the derivation for a Hamiltonian system weakly coupled to a heat bath [155], or the derivation based on a master equation approach [157], and several other approaches, see for instance references [158, 159]. However, one year after the discovery of Jarzynski's equality, Crook proved that the equality came out as a direct consequence of the critical assumptions regarding Markovian process and detailed balance condition [160].

In the following, we summarize Crook's derivation of both Jarzynski's equality and Crook's fluctuation theorem under the above-mentioned assumptions [160, 156]. For convenience, let us first clarify the notations and assumptions. The derivations follow right after that.

[2] See Section 2.4.1 for the definition of external control parameter and the formulation of detailed balance conditions.

[3] These conditions are satisfied by e.g., Hamilton equations with a time dependent Hamiltonian.

Notations and Assumptions

Consider a classical microscopic system in contact with a heat bath at a constant temperature T. Let us assume that the system can be controlled by a single external parameter λ. Manipulation of the value of λ cost an amount of work W, resulting in an exchange heat Q with the heat bath and invoking the change in the total energy ΔE and free energy ΔF of the system. We are interested in a process happening in a finite time τ in which λ is switched between an initial value λ_0 and a final value λ_τ. During the process, let λ_t be the value of the external controllable parameter, let i_t label the internal state of the system, and let $E(i_t, \lambda_t)$ denote the energy of the system at time t [4].

For a canonical ensemble, the equilibrium probability of a state i at a given value of λ can be expressed as

$$P(i|\lambda) = \frac{\mathrm{e}^{-\beta E(i,\lambda)}}{\sum_j \mathrm{e}^{-\beta E(j,\lambda)}} = \mathrm{e}^{\beta(F_\lambda - E(i,\lambda))}, \tag{2.4.4}$$

where F_λ denotes the free energy of the system at a given λ.

If the evolution of a system is assumed to be Markovian then the probability of going from state i_t to state i_{t+1} (i.e., $P(i_t \xrightarrow{\lambda} i_{t+1})$) depends only on the state at time t and not on all previous states. In other words, a Markovian process is a *memoryless* process. Under this assumption, the probability of evolving in a path from state i_0 to state i_τ given the control parameter at all time λ_t ($t = 0,\ 1,\ 2,\ \ldots,\ \tau$) can be split as

$$P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau) = P(i_0 \xrightarrow{\lambda_1} i_1)P(i_1 \xrightarrow{\lambda_2} i_2)\ldots P(i_{\tau-1} \xrightarrow{\lambda_\tau} i_\tau). \tag{2.4.5}$$

Now if every single step is assumed to be microscopically reversible then the following detailed balance condition must be satisfied

$$\frac{P(i \xrightarrow{\lambda} j)}{P(i \xleftarrow{\lambda} j)} = \frac{P(j|\lambda)}{P(i|\lambda)} \equiv \frac{\mathrm{e}^{-\beta E(j,\lambda)}}{\mathrm{e}^{-\beta E(i,\lambda)}}. \tag{2.4.6}$$

Let us now reproduce the derivation of detailed balance condition for a multiple-step process given that the process is Markovian. Using the Markovian property (Equation (2.4.5)), we can rewrite the ratio between the probabilities of a forward process and a corresponding time-reversed process as

[4] Discrete time and phase space will be used hereafter for notational convenience. The derivation presented here can be easily generalized to continuous time and phase space.

$$\frac{P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau)}{P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau)} = \frac{P(i_0 \xrightarrow{\lambda_1} i_1)P(i_1 \xrightarrow{\lambda_2} i_2)\ldots P(i_{\tau-1} \xrightarrow{\lambda_\tau} i_\tau)}{P(i_0 \xleftarrow{\lambda_1} i_1)P(i_1 \xleftarrow{\lambda_2} i_2)\ldots P(i_{\tau-1} \xleftarrow{\lambda_\tau} i_\tau)}. \tag{2.4.7}$$

Applying the detailed balance condition (Equation (2.4.6)) for every single step, we obtain

$$\begin{aligned}\frac{P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau)}{P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau)} &= \frac{\mathrm{e}^{-\beta E(i_1,\lambda_1)}\mathrm{e}^{-\beta E(i_2,\lambda_2)}\ldots\mathrm{e}^{-\beta E(i_\tau,\lambda_\tau)}}{\mathrm{e}^{-\beta E(i_0,\lambda_1)}\mathrm{e}^{-\beta E(i_1,\lambda_2)}\ldots\mathrm{e}^{-\beta E(i_{\tau-1},\lambda_\tau)}} \\ &= \mathrm{e}^{-\beta[(E(i_1,\lambda_1)-E(i_0,\lambda_1))+\cdots+(E(i_\tau,\lambda_\tau)-E(i_{\tau-1},\lambda_\tau))]} \\ &= \mathrm{e}^{-\beta Q}. \end{aligned} \tag{2.4.8}$$

Here Q is the total heat that the system exchanges with the heat bath. Equation (2.4.8) represents the detailed balance condition for a Markovian microscopically reversible system.

If both forward and backward processes start from the equilibrium initial and final states i_0 and i_τ, then by combining Equations (2.4.4) and (2.4.8), we find

$$\frac{P(i_0|\lambda_0)P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau)}{P(i_\tau|\lambda_\tau)P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau)} = \frac{\mathrm{e}^{\beta\left[F_{\lambda_0}-E(i_0,\lambda_0)\right]}}{\mathrm{e}^{\beta\left[F_{\lambda_\tau}-E(i_\tau,\lambda_\tau)\right]}}\mathrm{e}^{-\beta Q} = \mathrm{e}^{\beta(W-\Delta F)}. \tag{2.4.9}$$

Here $W = \Delta E - Q$ is the external work performed on the system.

Derivations

Below we rederive the proof that both Jarzynski's equality and Crook's fluctuation theorem arise directly from Equation (2.4.9) as a consequence of the assumptions regarding Markovian and detailed balance conditions.

(i) Derivation of Jarzynski's equality.

Relation (2.4.9) can be rearranged to yield

$$\mathrm{e}^{-\beta W} = \frac{P(i_\tau|\lambda_\tau)P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau)}{P(i_0|\lambda_0)P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau)}\mathrm{e}^{-\beta\Delta F}. \tag{2.4.10}$$

From Equation (2.4.10), the Jarzynski's equality follows directly. Indeed, if the process starts from a canonical equilibrium distribution, the ensemble average of the quantity $\mathrm{e}^{-\beta W}$ is given by

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \sum_{i_0, i_1, \ldots, i_\tau} P(i_0|\lambda_0) P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau) \mathrm{e}^{-\beta W}. \tag{2.4.11}$$

Here the ensemble average (on the left-hand side) is taken over all processes and hence the sum (on the right-hand side) runs over all paths in the discrete phase space, given a fixed sequence of the control parameter. Placing Equation (2.4.10) into (2.4.11), we can easily find

$$\left\langle \mathrm{e}^{-\beta W} \right\rangle = \mathrm{e}^{-\beta \Delta F} \sum_{i_0, i_1, \ldots, i_\tau} P(i_\tau|\lambda_\tau) P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau). \tag{2.4.12}$$

It is noteworthy that the free-energy difference ΔF is path independent. From here, Equation (2.4.2) arises directly because probabilities are normalized.

(ii) Derivation of Crook's fluctuation theorem.

The work distributions $\rho_F(+W')$ and $\rho_B(-W')$ associated with forward and backward paths are obtained by integrating over all possible discrete paths

$$\rho_F(+W') = \sum_{i_0, i_1, \ldots, i_\tau} P(i_0|\lambda_0) P(i_0 \xrightarrow{\lambda_1} i_1 \xrightarrow{\lambda_2} i_2 \rightarrow \ldots \xrightarrow{\lambda_\tau} i_\tau) \delta(W' - W). \tag{2.4.13}$$

$$\rho_B(-W') = \sum_{i_0, i_1, \ldots, i_\tau} P(i_\tau|\lambda_\tau) P(i_0 \xleftarrow{\lambda_1} i_1 \xleftarrow{\lambda_2} i_2 \leftarrow \ldots \xleftarrow{\lambda_\tau} i_\tau) \delta(W' - W). \tag{2.4.14}$$

Equation (2.4.9) can be rearranged as

$$P(i_0|\lambda_0) P(i_0 \xrightarrow{\lambda_1} i_1 \ldots \xrightarrow{\lambda_\tau} i_\tau) = P(i_\tau|\lambda_\tau) P(i_0 \xleftarrow{\lambda_1} i_1 \ldots \xleftarrow{\lambda_\tau} i_\tau) \mathrm{e}^{\beta(W - \Delta F)}. \tag{2.4.15}$$

By multiplying both sides of the above equation with $\delta(W' - W)$ and then summing over all possible paths, we obtain

$$\rho_F(+W') = \rho_B(-W')\mathrm{e}^{\beta(W'-\Delta F)}, \tag{2.4.16}$$

This relation is equivalent to Crook's fluctuation theorem (Equation (2.4.3)).

In conclusion, the free-energy difference between two equilibrium states of a system can be directly related to the nonequilibrium works required to switch the system between these two states. Jarzynski's equality allows recovering the free-energy difference from an exponential average of the works performed in unidirectional switching processes. However this estimation strongly depends on the behavior at the tails of the work distributions, which is poorly sampled with respect to the rest of the distribution especially for small sample sizes. Crook's fluctuation theorem involves the work distributions in both forward and backward processes and thus permits a better estimation compared to using information from only one direction. However, the quality of free-energy estimation by Crook's theorem is still dictated by the sample sizes. Jarzynski's equality can be recovered from Crook's fluctuation theorem. Both relations are proved to emerge as direct consequences of two essential assumptions that the evolution of the system satisfies both Markovian and detailed balance conditions.

2.4.2 Bennett Acceptance Ratio

Formulation of the Bennett Acceptance Ratio

Before the introduction of Crook's fluctuation theorem, in 1976 when examining two states of a system at *equilibrium*, Bennett already proposed to use the information of potential energy in both forward and backward distributions to improve the estimation of the free-energy difference [161]. Bennett's derivation of the so called **B***ennett* **A***cceptance* **R***atio* (**BAR**) based on *equilibrium potential energy* can be trivially generalized to the case of *nonequilibrium work* and rewritten as

$$\left\langle \frac{1}{n_B + n_F e^{\beta(W-\Delta F)}} \right\rangle_F = \left\langle \frac{1}{n_F + n_B e^{-\beta(W-\Delta F)}} \right\rangle_B, \tag{2.4.17}$$

where n_F and n_B denote the number of forward and backward processes respectively.

BAR can be considered the best *asymptotically unbiased* estimator [5] of the free energy given a set of nonequilibrium works performed in forward and backward processes. The free energy in BAR method is estimated by iteratively solving Equation (2.4.17).

[5] an asymptotically unbiased estimator is the estimator that becomes unbiased as the sample size goes to infinity

Derivation of BAR Using Maximum Likelihood Principles

Here we summarize a rederivation of BAR by Shirts et al. [162] using the maximum likelihood principles [163]. Details of the maximum likelihood can be found in Appendix A. Maximum likelihood estimators can be shown under relatively weak conditions to be asymptotically efficient, i.e., there are no other asymptotically unbiased estimator with lower variance. Therefore by deriving BAR using this estimator, we also prove that BAR is the best asymptotically unbiased estimator of free energy given a set of nonequilibrium works. This is the theoretical advantage of this derivation.

We start the derivation by rewriting the Crook's fluctuation theorem (Equation (2.4.3)) as

$$\frac{\rho(W|F)}{\rho(W|B)} = \mathrm{e}^{\beta(W-\Delta F)}. \tag{2.4.18}$$

For notational convenience, we have replaced $\rho_F(W)$ and $\rho_B(-W)$ with $\rho(W|F)$ and $\rho(-W|B)$ respectively. To further simplify the notation, we have substituted $-W$ with W without loss of generality. We now want to compute the likelihood of a free energy estimate from a given work measurements which come from either a forward or a backward process. For this purpose, we first rewrite the left hand side of Equation (2.4.18) using the properties of conditional probabilities

$$\frac{\rho(W|F)}{\rho(W|B)} = \frac{\frac{\rho(F|W)\rho(W)}{\rho(F)}}{\frac{\rho(B|W)\rho(W)}{\rho(B)}} = \frac{\rho(F|W)\rho(B)}{\rho(B|W)\rho(F)} = \frac{\rho(F|W)}{\rho(B|W)}\frac{n_B}{n_F}, \tag{2.4.19}$$

here $\rho(F|W)$ and $\rho(B|W)$ are, respectively, the conditional probabilities of a forward and backward process in which a work W is performed; and the ratio $\frac{\rho(B)}{\rho(F)}$ between the probabilities of backward and forward processes is equivalent to $\frac{n_B}{n_F}$.

From Equations (2.4.18) and (2.4.19), we have

$$\frac{\rho(F|W)}{\rho(B|W)} = \frac{n_F}{n_B}\mathrm{e}^{\beta(W-\Delta F)}. \tag{2.4.20}$$

Using the fact that $\rho(F|W) + \rho(B|W) = 1$, we can write the probability of a single measurement $\rho(F|W_i)$ and $\rho(B|W_i)$ as

$$\rho(F|W_i) = \frac{1}{1 + \frac{n_B}{n_F}\mathrm{e}^{-\beta(W_i-\Delta F)}}, \tag{2.4.21}$$

and

$$\rho(B|W_i) = \frac{1}{1 + \frac{n_F}{n_B} \mathrm{e}^{\beta(W_i - \Delta F)}}. \tag{2.4.22}$$

Given the value of ΔF, we now can write the overall likelihood $\mathcal{L}$ of obtaining the given measurements, i.e., the joint probability of forward processes at the specified work values times the joint probability of backward processes at the specified work values, meaning

$$\mathcal{L}(\Delta F) = \prod_{i=1}^{n_F} \rho(F|W_i) \prod_{j=1}^{n_B} \rho(B|W_j), \tag{2.4.23}$$

Employing Equations (2.4.21) and (2.4.22), Equation (2.4.23) becomes

$$\mathcal{L}(\Delta F) = \prod_{i=1}^{n_F} \frac{1}{1 + \frac{n_B}{n_F} \mathrm{e}^{-\beta(W_i - \Delta F)}} \prod_{j=1}^{n_B} \frac{1}{1 + \frac{n_F}{n_B} \mathrm{e}^{\beta\left(W_j - \Delta F\right)}}. \tag{2.4.24}$$

The log-likelihood is then written as

$$\ln \mathcal{L}(\Delta F) = \sum_{i=1}^{n_F} \ln \frac{1}{1 + \frac{n_B}{n_F} \mathrm{e}^{-\beta(W_i - \Delta F)}} + \sum_{j=1}^{n_B} \ln \frac{1}{1 + \frac{n_F}{n_B} \mathrm{e}^{\beta\left(W_j - \Delta F\right)}}. \tag{2.4.25}$$

The most likely value of ΔF is the value that maximizes the log-likelihood, or in other words is the solution of the following equation

$$\frac{\partial \ln \mathcal{L}(\Delta F)}{\partial \Delta F} = 0, \tag{2.4.26}$$

which is equivalent to

$$\sum_{i=1}^{n_F} \frac{-\beta}{1 + \frac{n_F}{n_B} \mathrm{e}^{\beta(W_i - \Delta F)}} + \sum_{j=1}^{n_B} \frac{\beta}{1 + \frac{n_B}{n_F} \mathrm{e}^{-\beta\left(W_j - \Delta F\right)}} = 0, \tag{2.4.27}$$

which can be further rearranged to yield

$$\sum_{i=1}^{n_F} \frac{\beta n_B}{n_B + n_F \mathrm{e}^{\beta(W_i - \Delta F)}} = \sum_{j=1}^{n_B} \frac{\beta n_F}{n_F + n_B \mathrm{e}^{-\beta\left(W_j - \Delta F\right)}}. \tag{2.4.28}$$

Dividing both sides of this equation by $\beta n_F n_B$, we obtain

$$\frac{1}{n_F} \sum_{i=1}^{n_F} \frac{1}{n_B + n_F \mathrm{e}^{\beta(W_i - \Delta F)}} = \frac{1}{n_B} \sum_{j=1}^{n_B} \frac{1}{n_F + n_B \mathrm{e}^{-\beta\left(W_j - \Delta F\right)}}. \tag{2.4.29}$$

This equation is exactly equivalent to the BAR method expressed in Equation (5.3.8).

In conclusion, given a set of nonequilibrium works measured in both forward and backward directions, the BAR estimator results in the free energy that maximizes the chance of observing these work values.

2.4.3 Potential of Mean Force Estimators

In the previous section, we presented the BAR method to estimate the most likely value of the free-energy difference given the nonequilibrium works from both forward and backward processes. It is important to notice that this is the free energy of the extended system, i.e., the system described by an extended Hamiltonian that is the sum of the external potential $V(t)$ and the Hamiltonian in the absence of the external perturbation. However, we are usually more interested in the **P***otential of* **M***ean* **F***orce* (**PMF**) of the *unperturbed system* described by the unperturbed Hamiltonian. In the following sections, we present the most commonly used PMF estimators for reconstructing the free energy of the unperturbed system from the nonequilibrium works. These estimators include *(i)* the Hummer-Szabo estimator for unidirectional processes and *(ii)* the Minh-Adib estimator for bidirectional cases.

2.4.3.1 Hummer-Szabo PMF Estimator for Unidirectional Steerings

Hummer-Szabo Formulation The Hummer-Szabo estimator is written as [158]

$$\mathrm{e}^{-\beta G_0(z)} = \frac{\sum_t \left\langle \delta(z-z_t)\mathrm{e}^{-\beta W_0^t} \right\rangle \mathrm{e}^{\beta \Delta F_t}}{\sum_t \mathrm{e}^{-\beta[V(z;t)-\Delta F_t]}}, \tag{2.4.30}$$

where $G_0(z)$ denotes the unperturbed free energy as a function of a chosen CV z; z_t is the value of the CV z at time t in a specific trajectory; $V(z;t) = k(z-\lambda_t)^2/2$ is the harmonic potential centering at λ_t acting on the CV z; W_0^t is the cumulative pulling work up to time t; $\triangle F_t$ is the free-energy difference of the extended system between the equilibrium state at time t and the initial equilibrium state. Note that the nonequilibrium works and free-energy difference satisfy Jarzynski's equality

$$\left\langle \mathrm{e}^{-\beta W_0^t} \right\rangle = \mathrm{e}^{-\beta \Delta F_t}. \tag{2.4.31}$$

The Hummer-Szabo estimator allows recovering the free-energy difference of the unperturbed system by relating it with the nonequilibrium works and the perturbing potentials.

Derivation In this section, we rederive the Hummer-Szabo estimator using the maximum likelihood principles. We first start by defining the biased equilibrium probability of observing the value z_t of the CV at time t in the trajectory i^{th} as

$$P(z_t^{(i)};t) = \frac{\mathrm{e}^{-\beta(G_0(z_t^{(i)})+V(z_t^{(i)};t))}}{Z(t)}, \tag{2.4.32}$$

here $Z(t) = \sum_{z'} \mathrm{e}^{-\beta(G_0(z')+V(z';t))}$ denotes the partition function at time t. $Z(t) \equiv e^{-\beta\Delta F_t}$ acts as a normalization factor.

Given the free energy $G_0(z)$ of the unperturbed system when the CV adopts the value z, the probability of observing a whole set of trajectories is then written as

$$P(G_0(z)) = \prod_t \prod_i P(z_t^{(i)};t) = \prod_t \prod_i \frac{\mathrm{e}^{-\beta(G_0(z_t^{(i)})+V(z_t^{(i)};t))}}{Z(t)}. \tag{2.4.33}$$

This is also defined as the likelihood $\mathcal{L}(G_0(z))$ of observing the set of trajectories given the free-energy difference $G_0(z)$. The log-likelihood is then given by

$$\ln \mathcal{L}(G_0(z)) = \sum_t \sum_i \left[-\beta(G_0(z_t^{(i)}) + V(z_t^{(i)};t)) - \ln Z(t) \right]. \tag{2.4.34}$$

The most likely value of $G_0(z)$ is the value that maximizes the log-likelihood, or in other words is the solution of the following equation

$$\frac{\partial \ln \mathcal{L}(G_0(z))}{\partial G_0(z)} = 0, \tag{2.4.35}$$

which is equivalent to

$$\sum_t \sum_i \left[-\beta\delta(z - z_t^{(i)}) + \frac{\beta \mathrm{e}^{-\beta(G_0(z)+V(z;t))}}{Z(t)} \right] = 0. \tag{2.4.36}$$

Note that trajectories with lower work values are closer to equilibrium and thus more reliable to be used when evaluating the equilibrium free energy. Therefore, a work-weighting factor $\mathrm{e}^{-\beta W_i(t)}$ together with its normalization factor $\mathrm{e}^{\beta\Delta F_t}$ should be added to each trajectory to give more weight to the low-work ones. Equation (2.4.36) then becomes

$$\sum_t \sum_i \left\{ \left[-\beta\delta(z - z_t^{(i)}) + \frac{\beta \mathrm{e}^{-\beta(G_0(z)+V(z;t))}}{Z(t)} \right] \mathrm{e}^{-\beta W_i(t)} \mathrm{e}^{\beta\Delta F_t} \right\} = 0, \tag{2.4.37}$$

which can be further rearranged to yield

$$\mathrm{e}^{-\beta G_0(z)} \sum_t \left\{ \left[\mathrm{e}^{-\beta V(z;t)} \mathrm{e}^{\beta \Delta F_t} \right] \sum_i \mathrm{e}^{-\beta W_i(t)} \mathrm{e}^{\beta \Delta F_t} \right\} = \sum_t \left\{ \left[\sum_i \delta(z - z_t^{(i)}) \mathrm{e}^{-\beta W_i(t)} \right] \mathrm{e}^{\beta \Delta F_t} \right\}, \tag{2.4.38}$$

here we use the definition of the partition function $Z(t) = \mathrm{e}^{-\beta \Delta F_t}$. Applying the Jarzynski's equality $\left\langle \mathrm{e}^{-\beta W_i(t)} \right\rangle = \mathrm{e}^{-\beta \Delta F_t}$, Equation (2.4.38) can be rewritten as

$$\mathrm{e}^{-\beta G_0(z)} = \frac{\sum_t \left\{ \left[\sum_i \delta(z - z_t^{(i)}) \mathrm{e}^{-\beta W_i(t)} \right] \mathrm{e}^{\beta \Delta F_t} \right\}}{n \sum_t \mathrm{e}^{-\beta [V(z;t) - \Delta F_t]}} = \frac{\sum_t \left\langle \delta(z - z_t) \mathrm{e}^{-\beta W_0^t} \right\rangle \mathrm{e}^{\beta \Delta F_t}}{\sum_t \mathrm{e}^{-\beta [V(z;t) - \Delta F_t]}}. \tag{2.4.39}$$

Here n denotes the total number of trajectories. Equation (2.4.39) is exactly the Hummer-Szabo estimator as shown in Equation (2.4.30).

2.4.3.2 Minh-Adib PMF Estimator for Bidirectional Steerings

Minh-Adib Formulation The Minh-Adib PMF estimator is given by [164]

$$\mathrm{e}^{-\beta G_0(z)} = \frac{\sum_t \left[\left\langle \frac{n_F \delta(z - z_t) \mathrm{e}^{-\beta W_0^t}}{n_F + n_B \mathrm{e}^{-\beta(W - \triangle F)}} \right\rangle_F + \left\langle \frac{n_B \delta(z - z_{\tau - t}) \mathrm{e}^{\beta W_{\tau - t}^{\tau}}}{n_F + n_B \mathrm{e}^{\beta(W + \triangle F)}} \right\rangle_B \right] \mathrm{e}^{\beta \triangle F_t}}{\sum_t \mathrm{e}^{-\beta [V(z;t) - \triangle F_t]}}, \tag{2.4.40}$$

where $G_0(z)$ denotes the unperturbed free energy as a function of a CV z ; τ is the total time of each pulling; z_t is the value of the CV z at time t ; $\langle\ \rangle_F$ and $\langle\ \rangle_B$ denote the averages taken over all forward and backward realizations respectively; n_F and n_B are the number of realizations in forward and backward pullings; W_0^t is the cumulative pulling work at time t; W is the total works at time τ performed in a certain forward or backward pulling; $V(z;t) = k[z - z_0(t)]^2/2$ is the harmonic potential acting on the CV at time t, where $z_0(t)$ denotes the "position" to which the CV is restrained at time t; $\triangle F_t$ is the free-energy difference between the equilibrium state at time t and the initial equilibrium state of the forward process, whose value is given by:

$$\mathrm{e}^{-\beta \triangle F_t} = \left\langle \frac{n_F \mathrm{e}^{-\beta W_0^t}}{n_F + n_B \mathrm{e}^{-\beta(W - \triangle F)}} \right\rangle_F + \left\langle \frac{n_B \mathrm{e}^{\beta W_{\tau - t}^{\tau}}}{n_F + n_B \mathrm{e}^{\beta(W + \triangle F)}} \right\rangle_B, \tag{2.4.41}$$

in which $\triangle F = \triangle F_\tau$ is the free-energy difference between the initial and final equilibrium states of the pulling, which can be calculated from the BAR method. Equation (2.4.41) can also be rearranged to give the BAR formula (Equation (2.4.17)) in the particular case where $t = \tau$.

The Minh-Adib estimator allows recovering the free-energy difference of the unperturbed system from the nonequilibrium works in both forward and backward processes.

Derivation In this section, we partially summarize the derivation of Minh-Adib estimator by Minh and Adib [164]. Notice that in their derivation, the bidirectional estimator was straightforwardly generalized from a unidirectional estimator, i.e., the Hummer-Szabo estimator, which was rederived using the **W***eighted* **H***istogram* **A***nalysis* **M***ethod* (**WHAM**) [165]. However, WHAM requires large numbers of samples to be valid. A workaround is to rederive the Hummer-Szabo estimator using maximum likelihood as we show in Section 2.4.3.1. Maximum likelihood principles are applicable even for small numbers of trajectories.

We start the derivation of Minh-Adib estimator by rewriting the Crook's fluctuation theorem (see Section 2.4.1)

$$P_F(W) = P_B(-W)\mathrm{e}^{\beta(W-\Delta F)}, \tag{2.4.42}$$

which can be rearranged to give

$$P_F(W) = \frac{n_F P_F(W) + n_B P_B(-W)}{n_F + n_B \mathrm{e}^{-\beta(W-\Delta F)}}. \tag{2.4.43}$$

The ensemble average of $f(W)$ over all forward realizations is given by

$$\begin{aligned}
\langle f(W)\rangle_F \equiv \frac{\sum_i^{n_F} f(W_i)P_F(W_i)}{\sum_i^{n_F} P_F(W_i)} &= \frac{\sum_i^{n_F} \frac{n_F f(W_i)P_F(W_i) + n_B f(W_i)P_B(-W_i)}{n_F+n_B\mathrm{e}^{-\beta(W-\Delta F)}}}{\sum_i^{n_F} P_F(W_i)} \\
&= \frac{\sum_i^{n_F} \frac{n_F f(W_i)P_F(W_i)}{n_F+n_B\mathrm{e}^{-\beta(W-\Delta F)}}}{\sum_i^{n_F} P_F(W_i)} + \frac{\sum_i^{n_F} \frac{n_B f(W_i)P_F(W_i)\mathrm{e}^{-\beta(W-\Delta F)}}{n_F+n_B\mathrm{e}^{-\beta(W-\Delta F)}}}{\sum_i^{n_F} P_F(W_i)} \\
&= \left\langle \frac{n_F f(W)}{n_F + n_B\mathrm{e}^{-\beta(W-\Delta F)}} \right\rangle_F + \left\langle \frac{n_B f(W)\mathrm{e}^{-\beta(W-\Delta F)}}{n_F + n_B\mathrm{e}^{-\beta(W-\Delta F)}} \right\rangle_F .
\end{aligned} \tag{2.4.44}$$

This is equivalent to

$$\langle f(W)\rangle_F = \left\langle \frac{n_F f(W)}{n_F + n_B\mathrm{e}^{-\beta(W-\Delta F)}} \right\rangle_F + \left\langle \frac{n_B f(-W)}{n_F + n_B\mathrm{e}^{\beta(W+\Delta F)}} \right\rangle_B . \tag{2.4.45}$$

Here we transform the forward to backward average in the second term using the Crook's path-ensemble average theorem [166], which can be rewritten as

$$\langle f(W)\rangle_F = \left\langle f(-W)\mathrm{e}^{-\beta(W+\Delta F)}\right\rangle_B . \tag{2.4.46}$$

Note that for backward trajectories, W is used instead of $-W$ due to the fact that a backward trajectory can be considered a reverse of the forward one. Now if we substitute $f(W) = \delta(z - z_t)\mathrm{e}^{-\beta W_0^t}$ into Equation (2.4.45), we get

$$\left\langle \delta(z - z_t)\mathrm{e}^{-\beta W_0^t}\right\rangle_F = \left\langle \frac{n_F\delta(z - z_t)\mathrm{e}^{-\beta W_0^t}}{n_F + n_B\mathrm{e}^{-\beta(W-\Delta F)}}\right\rangle_F + \left\langle \frac{n_B\delta(z - z_{\tau-t})\mathrm{e}^{\beta W_{\tau-t}^{\tau}}}{n_F + n_B\mathrm{e}^{\beta(W+\Delta F)}}\right\rangle_B . \tag{2.4.47}$$

Here the time and work in backward processes are “reverted”. If we further insert the right-hand side of the above equation into the following unidirectional Hummer-Szabo PMF estimator [6]

$$\mathrm{e}^{-\beta G_0(z)} = \frac{\sum_t \left\langle \delta(z - z_t)\mathrm{e}^{-\beta W_0^t}\right\rangle \mathrm{e}^{\beta\Delta F_t}}{\sum_t \mathrm{e}^{-\beta[V(z;t)-\Delta F_t]}}, \tag{2.4.48}$$

we obtain the so-called bidirectional Minh-Adib PMF estimator

$$\mathrm{e}^{-\beta G_0(z)} = \frac{\sum_t \left[\left\langle \frac{n_F\delta(z-z_t)\mathrm{e}^{-\beta W_0^t}}{n_F+n_B\mathrm{e}^{-\beta(W-\triangle F)}}\right\rangle_F + \left\langle \frac{n_B\delta(z-z_{\tau-t})\mathrm{e}^{\beta W_{\tau-t}^{\tau}}}{n_F+n_B\mathrm{e}^{\beta(W+\triangle F)}}\right\rangle_B\right]\mathrm{e}^{\beta\triangle F_t}}{\sum_t \mathrm{e}^{-\beta[V(z;t)-\triangle F_t]}}. \tag{2.4.49}$$

Here $\mathrm{e}^{\beta\triangle F_t}$ is used as a normalization factor. Its expression is given by choosing $f(W) = \mathrm{e}^{-\beta W_0^t}$ in Equation (2.4.45)

$$\mathrm{e}^{-\beta\triangle F_t} = \left\langle \frac{n_F\mathrm{e}^{-\beta W_0^t}}{n_F + n_B\mathrm{e}^{-\beta(W-\triangle F)}}\right\rangle_F + \left\langle \frac{n_B\mathrm{e}^{\beta W_{\tau-t}^{\tau}}}{n_F + n_B\mathrm{e}^{\beta(W+\triangle F)}}\right\rangle_B . \tag{2.4.50}$$

This equation can be rearranged to give BAR formula (i.e., Equation (2.4.17)) when $t = \tau$.

2.4.4 Projection of Free Energy on an *A Posteriori* Chosen CV

The PMF computed as a function of the steered CV does not necessarily provide a good picture of the investigated transition, as the steered CV is not guaranteed to properly distinguish all the relevant states. Moreover, in many cases it is instructive to look at the same result from a different perspective, i.e., computing the PMF as a function of a dif-

[6] which can be derived from maximum likelihood principles (see Section 2.4.3.1)

ferent, *a posteriori* chosen CV. Such task can be performed by employing a reweighting scheme. Suitable schemes have been proposed for other kinds of non-equilibrium simulations including metadynamics [167]. For SMD simulations, a reweighting algorithm for unidirectional pullings was introduced by some of us in a recent work, i.e., see ref. [168]. Here we extend this scheme to the case of bidirectional pullings.

The free energy as a function of an arbitrary, *a posteriori* chosen CV $\bar{z}$ is defined as

$$\mathrm{e}^{-\beta G_0(\bar{z})} = \sum_z \mathrm{e}^{-\beta G_0(z)} \delta(\bar{z} - \bar{z}(z)). \tag{2.4.51}$$

Using the Minh-Adib estimator for $G_0(z)$ we have

$$\begin{aligned} \mathrm{e}^{-\beta G_0(\bar{z})} &= \sum_z \frac{\sum_t \left[\left\langle \frac{n_F \delta(z - z_t) \mathrm{e}^{-\beta W_0^t}}{n_F + n_B \mathrm{e}^{-\beta(W - \triangle F)}} \right\rangle_F + \left\langle \frac{n_B \delta(z - z_{\tau - t}) \mathrm{e}^{\beta W_{\tau - t}^{\tau}}}{n_F + n_B \mathrm{e}^{\beta(W + \triangle F)}} \right\rangle_B \right] \mathrm{e}^{\beta \Delta F_t}}{\sum_t \mathrm{e}^{-\beta[V(z;t) - \Delta F_t]}} \delta(\bar{z} - \bar{z}(z)) \\ &= \frac{\sum_t \left[\sum_i^{n_F} \frac{\mathrm{e}^{-\beta(W_i(t) - \Delta F_t)}}{n_F + n_B \mathrm{e}^{-\beta(W_i - \triangle F)}} + \sum_j^{n_B} \frac{\mathrm{e}^{-\beta(W_i(t) - \triangle F_{\tau - t})} e^{-\beta \Delta F}}{n_B + n_F \mathrm{e}^{-\beta(W_j + \triangle F)}} \right]}{\sum_t \mathrm{e}^{-\beta[V(z_t;t) - \Delta F_t]}}. \end{aligned} \tag{2.4.52}$$

Here the free energy and work of the backward trajectories have been adjusted for a notational consistency. Now let us define the *weighting factors* of the forward and backward processes as

$$w_i^F(t) = \frac{\mathrm{e}^{-\beta(W_i(t) - \Delta F_t)}}{\sum_{t'} \mathrm{e}^{-\beta(V(z_t, t') - \triangle F_{t'})}} \times \frac{1}{n_F + n_B \mathrm{e}^{-\beta(W_i - \triangle F)}}, \tag{2.4.53}$$

and

$$w_j^B(t) = \frac{e^{-\beta(W_j(t) - \triangle F_{\tau - t})}}{\sum_{t'} e^{-\beta(V(z_t, t') - \triangle F_{t'})}} \times \frac{e^{-\beta \triangle F}}{n_B + n_F e^{-\beta(W_j + \triangle F)}}. \tag{2.4.54}$$

Based on these weights, the free energy can be estimated as a function of any a posteriori chosen CV $\bar{z}$ as:

$$\mathrm{e}^{-\beta G_0(\bar{z})} = \sum_t \left(\sum_i^{n_F} w_i^F(t) + \sum_j^{n_B} w_i^B(t) \right) \delta(\bar{z} - \bar{z}_t). \tag{2.4.55}$$

Our reweighting scheme (Equations (2.4.53), (2.4.54), and (2.4.55)) can be rearranged to give an identical expression to the Minh-Adib bidirectional PMF estimator (Equation (2.4.40)) when $\bar{z} \equiv z$. However, this approach of calculating a weighting factor at every time-frame is more general as it allows estimating the PMF as a function of a different, *a posteriori* chosen CV. Applying this algorithm, one can be flexible in projecting the PMF on the appropriate CVs for different post-processing purposes.

Chapter 3

Implementation of an Electrostatic-Based Collective Variable

Contents

3.1 Overview

Although **S***teered* **M***olecular* **D***ynamics* (**SMD**) and metadynamics simulations have been extensively used for ligand-binding studies, a proper choice of **C***ollective* **V***ariables* (**CVs**) still remains challenging and can be highly dependent on the specific problem. The center-to-center distance, a common choice of CV in binding/unbinding enhanced-sampling simulations, may disfavor the right complex formation by not taking into account the charge-charge interaction which is an important driving force in biomolecular recognition.

This thesis devises a strategy that uses a CV that is a proper approximation of the electrostatic-free-energy difference between the actual state of a biomolecular complex and a reference unbound state. This free energy can be easily computed on the fly within the Debye-Hückel formalism and can be used as a descriptor to distinguish the bound (lower free energy) and unbound (higher free energy) states.

In this chapter, we first rederive the formulation of the CV starting from the Debye-Hückel theory and the general **P***oisson*-**B***oltzmann* **E***quation* (**PBE**). We then compare the electrostatic free energies calculated by our proposed expression and by solving the non-linear PBE.

3.2 Derivation of the Electrostatic-Based Collective Variable

3.2.1 Debye-Hückel Continuum-Solvent Model

The continuum model of molecules in ionic solutions was first proposed by Debye and Hückel in 1923 for electrostatic-free-energy calculations of spherical ions [169]. Since then, it has been extended and considered an important tool to study electrostatic interactions in biochemical molecular systems.

In the original Debye-Hückel model, there is a particular ion of interest located at the region Ω_1. However, the model can be trivially extended to model a macromolecule in region Ω_1 with a dielectric constant ϵ_1 (see Figure 3.2.1). Region Ω_3 contains the solvent with a dielectric constant ϵ_3. Mobile ions also belong to this region. Region Ω_2 is called the ion-exclusion region and is described as a transition space which bears the same dielectric constant as the solvent region Ω_3, i.e., $\epsilon_2 = \epsilon_3$, but to which no mobile charges have access.

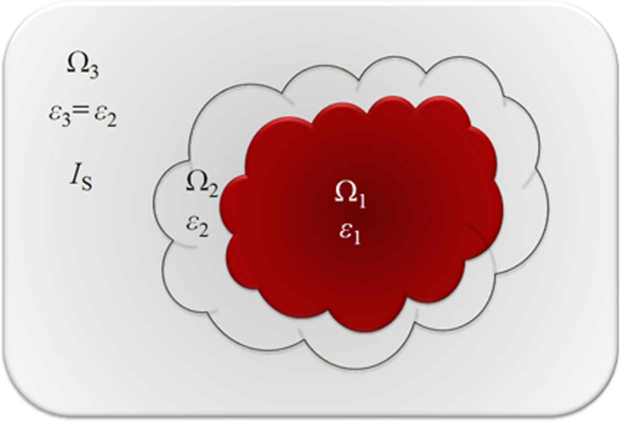

Figure 3.2.1: A two-dimensional schematic representation of the extended three-dimensional Debye-Hückel model for macrobiomolecular systems.

In the next section, we will describe the electrostatics of the model by deriving a Poisson equation for each region.

3.2.2 Nonlinear Poisson-Boltzmann Equation

The electrostatic potential in each of the region satisfies the Poisson equation which has the following form

$$-\nabla^2\Phi_k(\mathbf{r}) = \frac{4\pi}{\epsilon_k}\rho_k(\mathbf{r}), \tag{3.2.1}$$

in which $k = 1,\ 2,\ 3$ denotes the index of each region; $\Phi_k(\mathbf{r})$ is the electrostatic potential at a position $\mathbf{r}$ in region Ω_k; and ρ_k is the charge density function which depends on the charge distribution in each region Ω_k.

Suppose that the molecule is composed of N_m charges q_i at positions $\mathbf{r}_i$. These can be partial charges. The electrostatic potential in region Ω_1 is defined as

$$\Phi_1(\mathbf{r}) = \sum_{i=1}^{N_m} \frac{q_i}{\epsilon_1|\mathbf{r} - \mathbf{r}_i|}, \tag{3.2.2}$$

which consequently gives

$$-\nabla^2\Phi_1(\mathbf{r}) = \frac{4\pi}{\epsilon_1}\sum_{i=1}^{N_m} q_i\delta(\mathbf{r} - \mathbf{r}_i), \tag{3.2.3}$$

here δ is the Dirac delta function.

As exclusive to charges, the charge density function in region Ω_2 is given by $\rho_2(\mathbf{r}) = 0$. The Poisson equation for this region then becomes

$$-\nabla^2\Phi_2(\mathbf{r}) = 0. \tag{3.2.4}$$

The critical assumption in Debye-Hückel theory states that the mobile ions in region Ω_3 obey the *Boltzmann distribution law*, in other words the ratio between the ion local density and the bulk density of each type of ions is given by Boltzmann distribution law, namely

$$\frac{\rho_3^i(\mathbf{r})}{\rho_0} = \mathrm{e}^{-\beta W_i(\mathbf{r})}, \tag{3.2.5}$$

in which i denotes the ion type; $\rho_3^i(\mathbf{r})$ is the local density of the ion type i in region Ω_3; ρ_0 represents the ion bulk density; and $W_i(\mathbf{r})$ is the work required to bring one ion of type

i from infinity to the position $\mathbf{r}$.

In case of a *1:1 electrolyte*, which is applicable for most molecular simulations, we have two types of ions with opposite charges of $+e_c$ and $-e_c$. The works required to move these ions from far away to the position $\mathbf{r}$ can be written as

$$W_+(\mathbf{r}) = +e_c\Phi_3(\mathbf{r}), \tag{3.2.6}$$

and

$$W_-(\mathbf{r}) = -e_c\Phi_3(\mathbf{r}). \tag{3.2.7}$$

The charge density in region Ω_3 is then given by

$$\rho_3(\mathbf{r}) = \rho_3^+(\mathbf{r}) - \rho_3^-(\mathbf{r}) = \rho_0 e_c(\mathrm{e}^{-\beta e_c\Phi_3(\mathbf{r})} - \mathrm{e}^{\beta e_c\Phi_3(\mathbf{r})}) = -2\rho_0 e_c \sinh\Big(\frac{e_c\Phi_3(\mathbf{r})}{k_BT}\Big). \tag{3.2.8}$$

We next define $I_s = 1/2\sum_{i=1}^{N_I} c_i z_i^2$ as the ionic strength of the solvent, which is determined by N_I types of ions, each type has a charge $q_i = z_i e_c$ and a concentration c_i. The ion bulk density of a 1:1 electrolyte is then related to the ionic strength as

$$\rho_0 = \frac{N_A I_s}{1000}, \tag{3.2.9}$$

where N_A is the Avogadro's number. The Poisson equation for region Ω_3 can be then written as

$$-\nabla^2\Phi_3(\mathbf{r}) = \kappa^2\Big(\frac{k_BT}{e_c}\Big)\sinh\Big(\frac{e_c\Phi_3(\mathbf{r})}{k_BT}\Big), \tag{3.2.10}$$

here

$$\kappa = \Big(\frac{8\pi N_A e_c^2}{1000\epsilon_w k_BT}\Big)^{1/2} I_s^{1/2}, \tag{3.2.11}$$

is defined as the *Debye-Hückel parameter*.

From the Equations (3.2.3), (3.2.4), and (3.2.10), a single equation can be generalized as

$$-\nabla(\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})) = 4\pi\sum_{i=1}^{N_m} q_i\delta(\mathbf{r}-\mathbf{r}_i) - \bar{\kappa}^2(\mathbf{r})\Big(\frac{k_BT}{e_c}\Big)\sinh\Big(\frac{e_c\Phi(\mathbf{r})}{k_BT}\Big), \tag{3.2.12}$$

in which $\Phi(\mathbf{r})$ denotes the electrostatic potential at any point $\mathbf{r}$ in space; the permittivity $\epsilon(\mathbf{r})$ adopts the appropriate dielectric constant values in different regions (i.e., ϵ_1, ϵ_2, or ϵ_3). Notice the introduction of the *modified dielectric-independent Debye-Hückel parameter* $\bar{\kappa}(\mathbf{r})$ which is defined as $\bar{\kappa}(\mathbf{r}) = \sqrt{\epsilon_w}\kappa$ if the point $\mathbf{r}$ belongs to the solvent region Ω_3 and $\bar{\kappa}(\mathbf{r}) = 0$ elsewhere. Hereafter ϵ_w replaces ϵ_3 to represent the dielectric constant of water.

The **P***oisson*-**B***oltzmann* **E***quation* (**PBE**) (3.2.12) is a second-order nonlinear partial differential equation whose analytical solutions are not trivial to derive in general cases. Recently developed softwares including APBS [170] and DelPhi [171] have achieved remarkable accomplishment and improvement in providing robust numerical solutions of nonlinear PBE. However, the high computational cost makes impractical the integration of solving nonlinear PBE into MD-based simulations.

3.2.3 Linearized Poisson-Boltzmann Equation

Under the assumption of *dilute solution* such that the relation $e_c\Phi(\mathbf{r}) \ll k_BT$ holds, one can approximately keep only the first term of a linear approximation of $\sinh(x)$ and rewrite the PBE in a much simpler form:

$$-\nabla(\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})) = 4\pi\sum_{i=1}^{N_m} q_i\delta(\mathbf{r}-\mathbf{r}_i) - \bar{\kappa}^2(\mathbf{r})\Phi(\mathbf{r}). \tag{3.2.13}$$

Equation (3.2.13) is referred to as the linearized PBE or Debye-Hückel equation and its analytical solution can be explicitly constructed for the solvent region as:

$$\Phi^{DH}(\mathbf{r}) = \frac{1}{k_BT\epsilon_w}\sum_{i=1}^{N_m}\frac{q_i e^{-\kappa|\mathbf{r}-\mathbf{r}_i|}}{|\mathbf{r}-\mathbf{r}_i|}. \tag{3.2.14}$$

From the electrostatic potential in Equation (3.2.14), one can easily derive the electrostatic-interaction term in the free energy of a system consisting of two non-overlapping molecules as

$$G^{DH} = \sum_{j\in B} q_j\Phi^{DH}(\mathbf{r}_j) = \frac{1}{k_BT\epsilon_w}\sum_{j\in B}\sum_{i\in A} q_iq_j\frac{e^{-\kappa|\mathbf{r}_{ij}|}}{|\mathbf{r}_{ij}|}, \tag{3.2.15}$$

where A (B) is the set of all the atoms of the first (second) molecule; i and j are the atom indexes in the two sets A and B; and $|\mathbf{r}_{ij}| = |\mathbf{r}_i - \mathbf{r}_j|$ denotes the distance between atoms i and j.

Since Debye-Hückel equation is an approximation of nonlinear PBE in extremely dilute solution condition, the electrostatic potential $\Phi^{DH}(\mathbf{r})$ given by Equation (3.2.14) is hence an approximation. One can easily notice that $\Phi^{DH}(\mathbf{r})$ and thus G^{DH} do not account for the difference in electrostatic interactions due to different atomic sizes. Furthermore, neither of them consider the increasing in strength of electrostatic interactions close to and inside the molecular region, where the screening due to the ionic solution is smaller. The modified generalized Born model developed by Onufriev et al. [172] solved the limitations of linear Debye-Hückel equation by *(i)* introducing into the formula of electrostatic potential extra terms associated with the dielectric constant of the molecular region ϵ_p and *(ii)* modifying the distance $|\mathbf{r}_{ij}|$ to an effective distance taking into consideration atomic radii. In their model, the electrostatic potential defined at the position of each atom i is calculated as:

$$\Phi^{GB}(\mathbf{r}_i) = -\left(\frac{1}{\epsilon_p} - \frac{e^{-\kappa f_{ij}(r_{ij})}}{\epsilon_w}\right)\sum_j \frac{q_j}{f_{ij}(r_{ij})} + \sum_{j\neq i}\frac{1}{\epsilon_p}\frac{q_j}{r_{ij}}, \tag{3.2.16}$$

where $f_{ij}(r_{ij}) = \sqrt{r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)}$ is the modified distance, in which R_i and R_j are the so-called effective Born radii of atoms i and j . Effective Born radius account for the change in interaction energy in case that the atom may be surrounded by neighbors which replace the solvent. The method for estimating R_i is fully introduced in reference [172]. Undoubtedly, the Equation (3.2.16) tends to quantify better the electrostatic interaction than the Equation (3.2.14). However, its drawback involves the nontrivial estimation of effective Born radius for each atom that in turn makes it impractical to be implemented as a CV.

We thus propose to use the expression (3.2.15), from now on referred to as **D***ebye*-**H***ückel* **EN***ergy* (**DHEN**), as a CV due to its generality and computational efficiency and due to the small number of parameters needed. Here we assume that only the inter-molecular electrostatic interactions contribute significantly to the free-energy difference between the bound and unbound states of a complex, thus ignoring the intramolecular relaxation. This assumption and the dilute-solution approximation do not affect the accuracy of free-energy calculation, which is obtained from the all-atom accelerated simulations. Indeed, G^{DH} in Equation (3.2.15) is not claimed to be the electrostatic free energy of the system. It is only used as a CV for guiding the exploration of the conformational space.

3.3 Free Energy Calculations from Nonlinear versus Linearized PBEs

For a comparison of electrostatic free energy between solving nonlinear and linear PBEs, we performed electrostatic calculations on configurations featuring all relative orientations between L22 and TAR. This section presents the calculation procedure and the comparison results.

3.3.1 Calculation Procedure

To prepare for the calculation, an arbitrary coordinate system was chosen. L22 was then put at the origin and TAR was placed at 40 Å on an axis, e.g., the x-axis. This choice of distance ensured that L22 and TAR were far enough to have no inter-molecular contacts. The dipole moment of both molecules were aligned with another direction, e.g., the z-axis. We next progressively rotated both L22 and TAR molecules about the x-, y-, and z- axes. The combination of these rotations is equivalent to placing the ligand in all three-dimensional rotations everywhere on the surface of a sphere with a radius of 40 Å around the RNA.

The "angle-step" of the rotation was 1°. At every step, the electrostatic-interaction free energy was calculated by two methods: *(i)* numerically solving the nonlinear PBEs using APBS 1.3 [170] and *(ii)* using our proposed DHEN estimator (Equation (3.2.15)). These calculations provided the orientation dependence of the electrostatic interaction.

3.3.2 Results

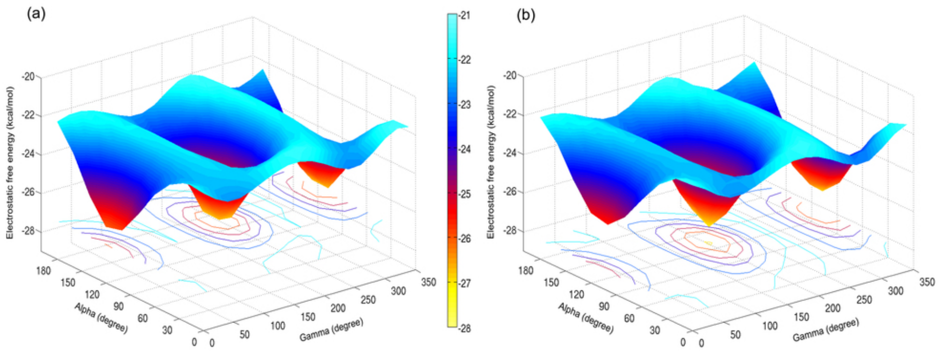

Figure 3.3.1: Electrostatic interaction free energy as a function of TAR's orientations calculated by both methods: (a) numerically solving nonlinear PBE and (b) using Equation (3.2.15) which analytically arises from the linearized PBE. In both calculations, we found two energy minima corresponding to two orientations of TAR at which L22 face both the upper and lower major groove of TAR.

The linearized PBE gives an explicit expression of the electrostatic free energy, i.e., Equation (3.2.15). Using this expression has a great computational advantage compared to solving the nonlinear PBE numerically. In fact, for every rotation, it takes only a fraction of a second to calculate the electrostatic free energy using Equation (3.2.15) while it takes $3-4$ minutes to numerically solve the Equation (3.2.12) using APBS. The results, however, are not considerably different from one another. That can be observed in Figure (3.3.1), which shows the dependence of electrostatic free energy on the two angles α and γ, which describe the rotation of TAR about the x- and y- axes. Both calculation methods agree on the two minima representing two orientations of TAR at which the electrostatic free energy of the system have the lowest values. This result provided us with more confidence when using DHEN as a CV for the accelerated simulations which will be presented in Chapter 5.

3.3.3 Qualitative Prediction of L22-TAR Binding Modes

As a further notice, both methods suggest that there are two orientations of TAR featuring the lowest electrostatic free energy of the system. Since both L22 and TAR were treated as rigid molecules, these calculations did not provide an accurate estimation of the electrostatic free energy. However, from these calculations, we can learn two important facts

(i) the electrostatic free energy DHEN is more "collective" and "selective" than the center-to-center distance when describing the system. Indeed, if we look at the distance only, we cannot tell the difference among the relative orientations of L22 and TAR. The electrostatic free energy, however, can tell us that at some orientations, the interaction becomes stronger than at the others,

(ii) there are two possible low-energy funnels for the L22-to-TAR encounter path: approaching the upper major groove from above and the lower major groove from below (as sketched in Figure 3.3.2).

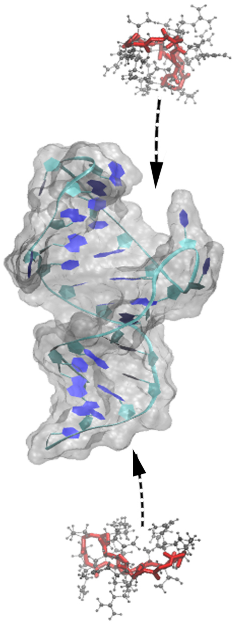

Figure 3.3.2: Possible approaches of L22 to TAR predicted by electrostatic-free-energy calculations: *(i)* upper major groove, which is as well the binding site of Tat and *(ii)* lower major groove. Solving both nonlinear and linearized PBE are in agreement on this result.

However, this is merely a qualitative assessment. We will verify it by a more quantitative methodology in the next sections.

3.4 Concluding Remarks

We propose to use DHEN, i.e., the electrostatic free energy given by Equation (3.2.15), as a CV for accelerating simulations. Although derived from the linear Debye-Hückel equation, which is obtained by a dilute-solution approximation, the crudeness of this CV does not affect the accuracy of free energy calculation. Indeed, it is only used as a CV on top of which a suitable bias is added. The computational advantage of the DHEN formulation is that besides the atomic charges which are ready in the force field and the well-defined intrinsic properties of the system including temperature, ionic strength, and solvent dielectric constant, DHEN does not require extra parametrization. The DHEN CV was implemented in an in-house version PLUMED 1.3 [167]. Therefore, it can be employed interactively with the common MD simulation engines such as GROMACS, NAMD, AMBER, etc.

An electrostatic-free-energy calculation of the L22-TAR system using DHEN CV qualitatively predicts that there are two possible low-energy funnels for L22 to approach TAR which lead to two possible binding pockets: the upper and lower major grooves. This prediction is to be quantitatively assessed in the next chapters.

Chapter 4

Molecular Dynamics Simulations of the L22-TAR Complex

Contents

4.1 Overview

Here we use **M**_olecular_ **D**_ynamics_ (**MD**) simulations to investigate changes in TAR structure and plasticity as well as ion redistribution upon L22 binding. We first validate our computational approach by comparing the simulated structural parameters and conformational fluctuations with experimental results obtained by NMR. Our calculated structural features and order parameter S^2 values are in agreement with experimental data. We then average the ion distributions during the simulations of apo-TAR and bound-TAR. The calculations are complemented by an analysis of the hydration of TAR in both free and bound states. Finally, we perform several standard MD simulations starting from different initial configurations of the unbound states of the complex system to see how the two molecules recognize each other. Our calculations show that the *encounter* between the TAR RNA and the positively-charged L22 peptide is a *spontaneous process* strongly driven by *electrostatic interaction*, which happens in a very short time-scale of no more than 5 ns. Indeed, electrostatic interaction plays a critical role in binding of proteins and small molecules to RNA [173]. This interaction is presumed to be modulated by the distributions of ions around the RNA molecule which in return are also altered during the binding process. However, this issue has not yet been fully targeted in RNA studies. Therefore we carefully investigate the rearrangement of ions during the molecular recognition events leading to the formation of the L22-TAR complex. Additionally, we calculate the ion occupancies with respect to each RNA nucleotide. Our calculations shed light on ion redistribution upon ligand binding, a feature that has yet to be examined.

In this chapter, we first introduce the biological systems to be simulated and the simulation protocols. We then present the main MD results as well as a thorough comparison with the NMR studies.

4.2 Systems and Simulation Protocols

4.2.1 Apo-L22, Apo-TAR, and L22-TAR Complex Systems

The molecular systems used for the MD simulations presented in this chapter were extracted from the NMR structure of the L22-TAR complex [24] (pdb code: 2KDQ, see Figure 4.2.1a). A total of 544 ns of MD production run were performed in nine simulations including (i) 100-ns simulation of apo-L22, (ii) 200-ns simulation of apo-TAR, (iii) 200-ns simulation of L22-TAR complex, and (iv) 44 ns of six simulations starting from six different unbound structures (referred to as L22//TAR hereafter). To prepare the initial structures of L22//TAR, we defined an arbitrary Cartesian coordinate system originating at TAR's geometric center. L22 was then placed at ±40 Å from the origin along the three axes x, y, and z, generating six starting structures (see Figure 4.2.1b). This approach is general because the positions of L22 are arbitrarily chosen due to the arbitrary definition of the coordinate system.

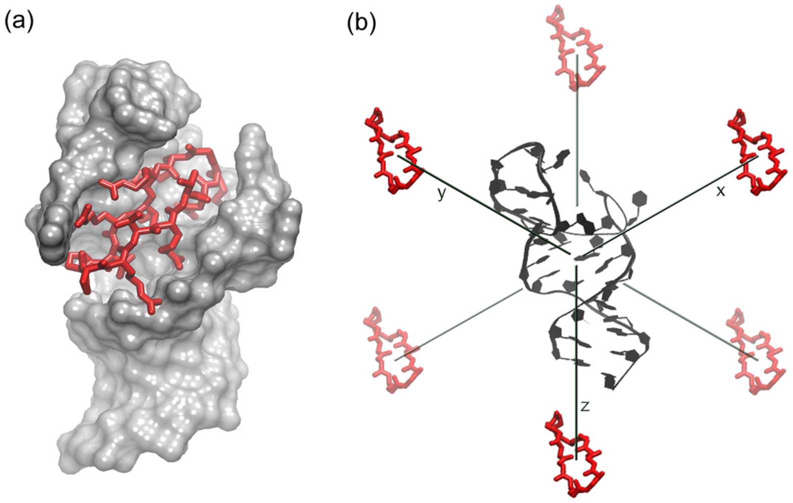

Figure 4.2.1: The starting structures of nine MD simulations. (a) The L22-TAR complex from NMR experiment, which provides the starting conformations for three simulations including apo-L22, apo-TAR, and L22-TAR complex. (b) Six starting structures (L22//TAR) for encounter simulations in which L22 is placed at 40 Å away from the TAR's center.

4.2.2 Simulation Setups

The nine starting structures, including apo-L22, apo-TAR, L22-TAR, and six structures of L22//TAR, were embedded in explicit water boxes to which the periodic boundary conditions were applied. The solutes and their images were located at a minimum distance of 24 Å. The use of a large simulation box was necessary to decrease artificial interactions between the highly charged molecules and their periodically repeated images (see Section 2.2.1.4 for a detailed discussion on this issue). A total of 3,175; 7,401; 7,374; and 31,260 water molecules were used in the simulations of apo-L22; apo-TAR; L22-TAR; and L22//TAR respectively. KCl was added to neutralize the charges and reproduce the experimental ion concentration of 10 mM in all cases [83]. The number of K^+ and Cl^- ions in each simulation along with other setup details such as box size and total number of atoms are specified in Table 4.1.

Systems	Simulation setup				
	Box size (Å^3)	Ions		No. of atoms	Force fields
		K^+	Cl^-		
Apo-L22	57 × 51 × 46	1	8	9,803	ff03
Apo-TAR	60 × 72 × 71	30	2	23,165	ff03+parmbsc0
L22-TAR	66 × 72 × 71	23	2	23,346	
L22//TAR	107 × 107 × 107	27	6	95,012	

Table 4.1: Summary of MD simulation setup information.

We employed TIP3P model [174] for water, AMBER ff03 force field [175] for L22, and ff03 with parmbsc0 reparametrization [6] for TAR (see Section 2.2.1.3 for the clarification this reparametrization). This combination of force fields has been shown to provide good results for protein/nucleic-acid complexes [176, 177].

4.2.3 Control Parameters and Simulation Protocols

Control Parameters

All-atom MD simulations were performed using the program NAMD 2.6 [178]. The Particle-Mesh Ewald method [146, 179] was used to treat the long-range electrostatic interactions with a real space cut-off of 12 Å. The same cut-off was also used for the van der Waals interactions. The simulation time-step was 1 fs and the non-bonded atom pair list was updated every 20 steps. The SHAKE algorithm [180] was applied to constrain all bonds involving hydrogen atoms. *NPT* conditions were controlled by the Langevin equation [181, 182] describing the coupling of the systems to a thermostat at 300 K with a damping coefficient of 1 ps^{-1} and a barostat at 1 atm with an oscillation period of 200 fs and a decay coefficient of 100 fs.

Simulation Protocols

MD Simulations of all systems were performed with the following protocol

Step 1: *Minimization*. A minimization procedure was conducted for 15,000−18,000 steps until the root-mean-square energy gradient reached the value of about 10^{-2} kcal/mol.

Step 2: *Solvent equilibration*. Water molecules and ions underwent a 50-ps constant-volume MD simulation, keeping restraint on each atom of the solute with the force constant of 500 kcal/mol/Å^2.

Step 3: *Heating*. The whole system was heated up to 300 K with a 200-ps constant-volume MD simulation. No restraint was applied to any atom.

Step 4: *Equilibration*. A 500-ps MD simulation in NPT condition was performed at 300 K and 1 atm. The density fluctuated around 1 g/ml. The configuration associated with the density that is closest to the average density was saved as the starting configuration for the next step.

Step 5: *Production*. A long MD simulation was carried out with the same protocol as in the equilibration step. For MD production run, we performed 100 ns for apo-L22, 200 ns for apo-TAR, another 200 ns for L22-TAR, and a total of 44 ns for six L22//TAR systems.

4.3 Results

In this section, we first validate our computational approach by comparing simulated structural parameters and conformational fluctuations with experimental results obtained by NMR studies. We then describe the distributions of water molecules and ions around TAR in the simulations of apo-TAR and L22-TAR. Finally, we present the results of the MD simulations starting from different initial conditions of the L22//TAR systems. These simulations provide insight into ion redistribution upon ligand binding.

4.3.1 General Features and Comparisons with NMR Results

4.3.1.1 Structural Features of TAR RNA in Both Apo- and Bound- States

TAR becomes more rigid and compact upon complex formation. The structural features of apo-TAR and bound-TAR, as obtained by 200-ns MD simulations, reproduce well the NMR observations [24]. As expected, TAR becomes more rigid upon complex formation.

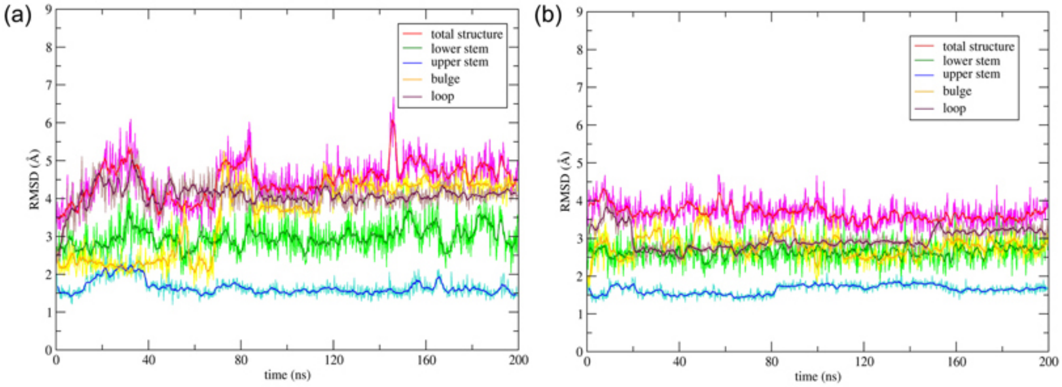

Figure 4.3.1: RMSDs and running averages with respect to the initial NMR structure of apo-TAR (a) and bound-TAR (b) in 200-ns MD simulations. The values for the entire TAR structure are shown in red. Those for specific regions, i.e., lower stem, upper stem, bulge, and loop, are shown in green, blue, yellow, and brown respectively. All regions in bound-TAR vary less than those in apo-TAR.

Figure 4.3.1 shows the **R***oot* **M***ean* **S***quare* **D***eviations* (**RMSD**s) with respect to the NMR structure (i.e., the starting structure in all simulations) of apo-TAR (panel (a)) and bound-TAR (panel (b)). All regions of TAR vary less when TAR is bound to L22. Indeed, the average RMSD of apo-TAR with respect to its initial configuration is 4.5 ± 0.6 Å, while that of bound-TAR is 3.7 ± 0.3 Å. Looking closer into the local deviations, we found that in both cases, the bulge and loop regions exhibit greater variations than the rest since they are unstructured and hence are allowed to move freely during the simulations. The upper stem, in both cases, shows the lowest deviation from its starting structure because the location of this stem is more constrained than the rest of the regions. Interestingly, the lower stem, most part of which is not structurally bound to the ligand, also experiences a decrement in flexibility upon complex formation.

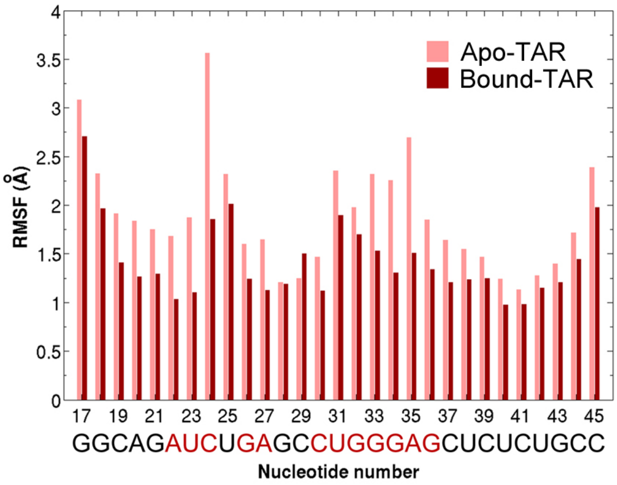

Figure 4.3.2: Average RMSFs are shown for each nucleotide of apo-TAR (in pink) and bound-TAR (in red). Below the x-axis, the letters representing the nucleotides involved in ligand binding are marked in red while the rest are shown in black. Overall, bound-TAR is more rigid than apo-TAR and the ligand-binding regions change the most upon complex formation.

A similar feature can be observed when plotting the average **R***oot* **M***ean* **S***quare* **F***luctuations* (**RMSF**s) of each nucleotide in both apo-TAR and bound-TAR (Figure 4.3.2). The bulge and loop regions have the highest flexibility in both cases. The nucleotides involved in peptide binding significantly decrease their flexibilities upon complex formation. For instance, the RMSFs of C24 (a bulge nucleotide) and A35 (a loop nucleotide) decrease from 3.6 Å to 1.9 Å and from 2.7 Å to 1.5 Å respectively.

P*rinciple* **C***omponent* **A***nalysis* (**PCA**) [183] further showed that upon complex formation, TAR decreased considerably its flexibility; this is clearer for the case of nucleotide A35 (see Figure 4.3.3). Additionally, in the bulge region, C24 becomes more rigid due to its interaction with a peptide residue (specific L22-TAR interactions are to be discussed in Section 4.3.1.2). However, U25 of the bulge region turned to be more flexible presumably because it was more exposed to the solvent in the bound state.

A closer look into the conformational changes of the hairpin loop is presented in Figure 4.3.4. In the starting NMR structure (panel (a)), all loop nucleotides except for G33 and A35 point toward the major groove, forming a compact structure. In apo-TAR, during the MD simulation, the hairpin loop is flexible: U31, G32, G33, and A35 point outward to the solvent, while C30 and G34 point toward the loop (panels (b1)−(b5)). In bound-TAR, the loop is more rigid compared to that of apo-TAR. However, the bases of C30, U31, G32, and G34 still adopt a wide range of conformations, while their backbones are structurally

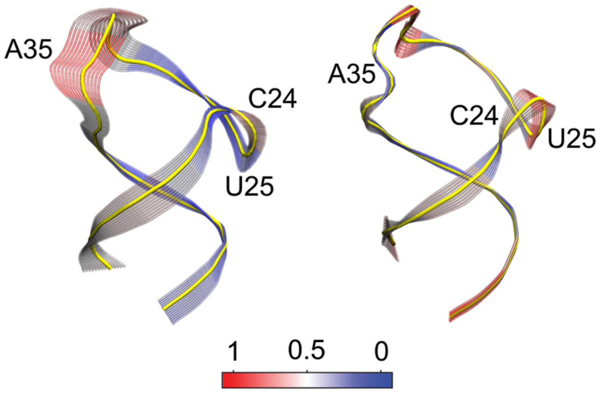

Figure 4.3.3: Superposition of the main configurations adopted by apo-TAR (left) and bound-TAR (right) from PCA calculations. Normalized degree of flexibility per nucleotide is shown in the colors ranging from blue (complete rigidity) to red (complete flexibility). Bound-TAR is more rigid than apo-TAR.

similar to those of the NMR structure (panels (c1)−(c5)). A35 is relatively rigid due to its interaction with a peptide residue.

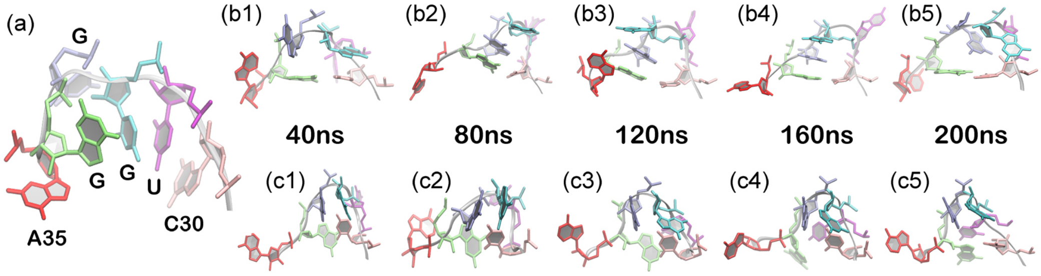

Figure 4.3.4: (a) The hairpin loop conformation in the starting NMR structure with the loop nucleotides C30-U-G-G-G-A35 presented in pink, purple, cyan, ice blue, lime, and red, respectively. (b1-5) Snapshots of the loop structure at 40 ns, 80 ns, 120 ns, 160 ns, and 200 ns respectively in simulation of apo-TAR. (c1-5) Snapshots of the loop structure at 40 ns, 80 ns, 120 ns, 160 ns, and 200 ns respectively in simulation of bound-TAR.

Besides losing flexibility, TAR becomes slightly more compact upon L22 binding as well; the average radius of gyration of apo-TAR and bound-TAR are 14.5 ± 0.5 Å and 13.7 ± 0.2 Å respectively.

Agreement between MD-derived and NMR-resulted order parameters. Our calculations reproduced well the experimentally-derived order parameter S^2 values, which were obtained from NMR analysis of the relaxation of the base C8−H8 dipolar interactions (in adenine and guanine) and C6−H6 dipolar interactions (in cytosine and uracil) for both apo-TAR and bound-TAR (see Section (2.2.2.1) for the definition and formulation of the NMR order parameter S^2). The results are presented in Figure 4.3.5. In particular, in both NMR experiments and MD simulations, the bulge and loop are the most mobile regions in apo-TAR, and remain so in the bound-TAR. The bulge nucleotide U25 and loop nucleotide U31 of apo-TAR are flexible with low order parameters, indicative of local mobility ($S^2 = 0.57$ and 0.31 respectively). Upon ligand binding, U25 becomes more solvent exposed and hence it increases the mobility; U31, on the contrary, becomes more rigid through the reorganization of the loop. There is, however, a remarkable discrepancy between simulations and experiments in the S^2 value of the loop nucleotide A35. This is probably due to the solvent-exposed property of A35 which leads to its high flexibility and hence its relaxation time is probably poorly estimated within the performed computational time. Our simulations also provide information on the S^2 values of several nucleotides (mostly in the helical regions) which are not available by NMR experiments due to spectral overlapping.

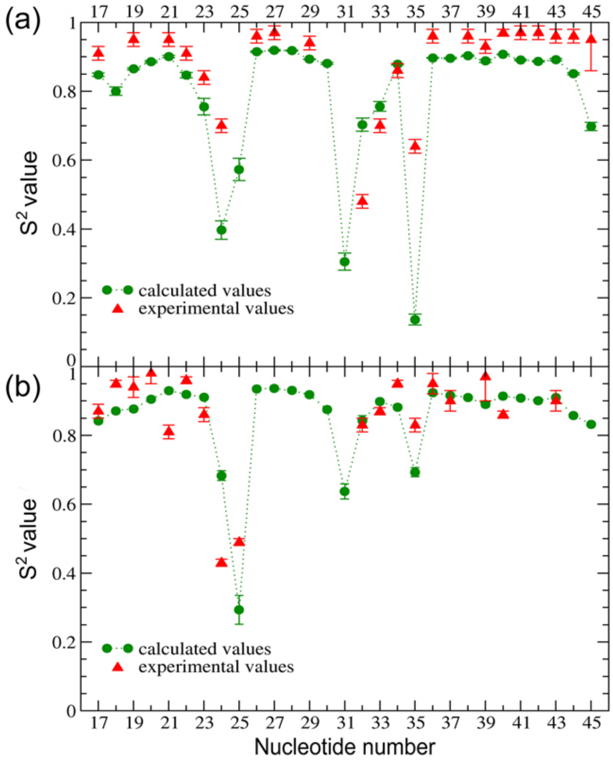

Figure 4.3.5: Per nucleotide NMR order parameter (S^2) as obtained by MD simulations (green circles) and by NMR experiments (red triangles) in apo-TAR (a) and bound-TAR (b).

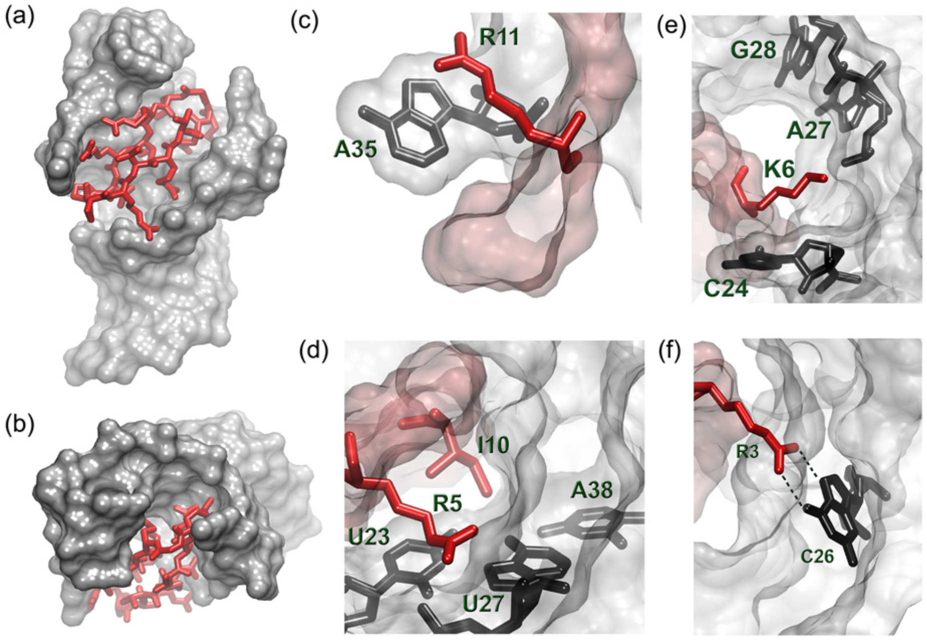

Figure 4.3.6: (a) A side view of L22-TAR NMR structure. (b) A top view of L22-TAR NMR structure. (c)-(f) Main contacts of L22 and TAR as seen in MD simulation and in agreement with NMR structure: (c) cation-π interaction between Arg11 and A35; (d) cation-π interaction between Arg5 and U23 together with the burial of the hydrophobic residue Ile 10 into the RNA backbone which facilitates the formation of the U23/U27-A38 base triple; (e) Lys6 side chain pointing to the pocket created by the backbones of C24, A27, and G28; and (f) hydrogen bonds between Arg3 and C26.

4.3.1.2 L22-TAR Interactions

All key hydrophobic and polar contacts between L22 and TAR observed by NMR experiments are reproduced in our simulations (see Figure 4.3.6). These include:

(i) the cation-π interaction between the guanidinium group of Arg11 and the loop nucleotide A35 (panel (c)). This interaction draws A35 toward the UCU bulge (panel (a)). Such a displacement in turn draws down the other loop nucleotides G32, G33 and G34, forming a cavity where the peptide is partially buried (panel (a)).

(ii) another cation-π interaction was also observed between Arg5 guanidinium group and U23 (panel (d)).

(iii) the hydrophobic interactions between the Ile10 methyls and the TAR nucleobases, including U23, A27-U38, and G28-C37 base pairs. These interactions, first observed in the related BIV Tat-TAR complex [184, 185], facilitate the formation of the U23/A27-U38 base triple (see also panel (d)).

(iv) the hydrogen bonds between the side chain of Lys6 with the phosphate groups of C24,

A27, and G28 as Lys6 points to a pocket formed by the backbone of these nucleotides (panel (e)).

(v) Arg3 forms a hydrogen bonds with the phosphate of U23, although it has been proposed to interact also with A22 [75, 186]. Arg3 also forms hydrogen bonds with G26 (panel (f)), although these did not emerge directly from the NMR data.

(vi) Arg5 forms hydrogen bonds with G28.

Other selected L22-TAR interactions whose life-time exceed 30% of the 200-ns MD simulation are reported in Table 4.2.

L22	TAR	average distance (Å)	time course (%)
Arg8 (NE)	C30 (N3)	3.0 (0.1)	95
Arg8 (NH1)	G34 (O2′)	2.9 (0.1)	93
Arg8 (NH1)	C30 (O2)	3.0 (0.2)	85
Arg5 (NH1)	G28 (O2P)	2.8 (0.1)	74
Arg5 (NH1)	A27 (O2P)	2.9 (0.2)	70
Arg3 (NH1)	G26 (N7)	3.0 (0.1)	69
Arg3 (NH2)	G26 (O6)	3.0 (0.2)	56
Arg11 (NH1)	A35 (O1P)	3.0 (0.2)	52
Arg8 (NH2)	A35 (O2P)	3.1 (0.2)	42
Arg9 (NH1)	A35 (O1P)	2.9 (0.2)	42
Arg5 (NH2)	G28 (O2P)	3.1 (0.2)	38
Arg3 (NE)	G28 (O1P)	2.9 (0.2)	38
Arg5 (NE)	U23 (O2)	3.1 (0.2)	37
Arg1 (NH1)	A22 (O2P)	2.8 (0.1)	37
Arg1 (NH1)	A22 (N7)	3.0 (0.2)	33
Arg9 (NH1)	G33 (O2′)	3.1 (0.2)	32
Arg11 (NE)	A35 (O2P)	2.9 (0.2)	31

Table 4.2: Key L22-TAR interactions with average length and time course (> 30%) as observed in the 200-ns MD simulation of L22-TAR complex.

4.3.2 Hydration and Ion Distribution around TAR

4.3.2.1 Hydration Properties of TAR in Both Apo- and Bound- States

The bulge and loop of apo-TAR are highly hydrated. The most hydrated residues are C24, U31, and A35, which are fully solvent-exposed in the absence of ligand (see Figure 4.3.7). Binding of L22 causes a decrease of hydration for these nucleotides. The numbers of water molecules within the first hydration shell around the mentioned nucleotides decrease from 36 - 37 in apo-TAR to 23 - 27 in bound-TAR. At the same time, U25 becomes the most solvent-exposed nucleotide in bound-TAR.

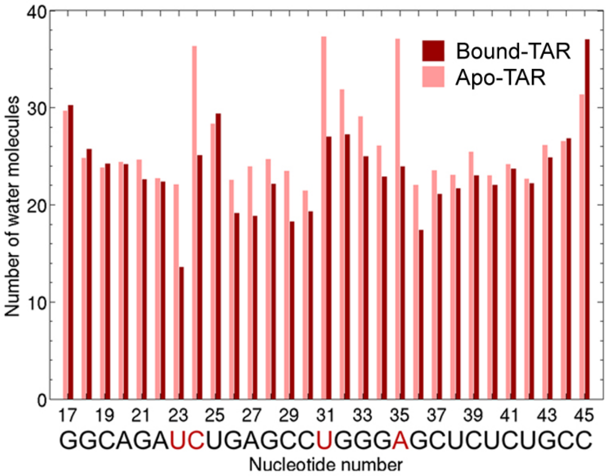

Figure 4.3.7: Number of water molecules in the first hydration shell around each nucleotide in apo-TAR (pink) and bound-TAR (red).

4.3.2.2 Electrostatic Properties of TAR and Surrounding Ion Distribution

The electrostatic potentials on the surface of TAR were calculated by solving nonlinear PBEs using APBS 1.3 [170]. The RNA major groove was found to be more electro-negative than the minor groove (see Figure 4.3.8). In general, this result supports the electrostatic feature of A-form nucleic acids. Within the major groove, the upper part (i.e., nucleotides U23-C39) involves more electro-negativity when compared to the lower part (i.e., nucleotides G17- A22 and U40- C45). This observation qualitatively explains the binding position of the positively-charged Tat's region and also of the ligand L22.

Consistently, in apo-TAR, K^+ occupancy has the following order: *upper major groove*, *lower major groove*, and minor groove (see Figure 4.3.9a). Most of the K^+ ions are found close to the nucleotides 20-23 and A27, which are located around the UCU bulge and form the crucial part for L22 and Tat binding (see Figure 4.3.10). Among those nucleotides, A22 has the highest propensity to bind K^+ ions. This result agrees with what was found in a recent MD study [187], although that simulation was considerably shorter (20 ns).

In bound-TAR, as L22 is present in the upper major groove, K^+ ions are displaced towards the lower major groove and Cl^- ions are also found to couple with K^+ ions in the lower major groove (see Figure 4.3.9b).

4.3.3 Formations of the L22-TAR Complex from Encounter Simulations

Six 5-ns MD simulations were performed for the L22//TAR systems (for details on how the systems were prepared, see Section 4.2). By doing these simulations, we aimed at

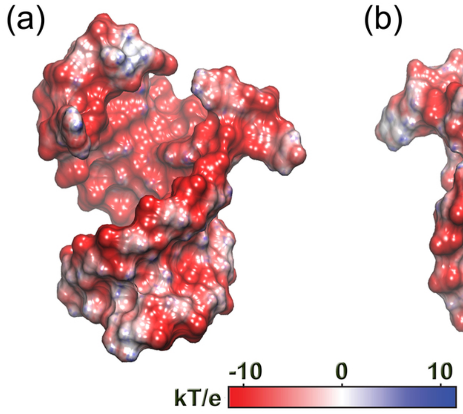

Figure 4.3.8: Electrostatic potentials on TAR's surface are shown in a continuous RWB color scale with red representing negative potentials (from -10 kT/e), white representing neutral potentials (0 kT/e), and blue representing positive potentials (up to $+10$ kT/e). The subfigures show: (a) the view from TAR's front and (b) the view from TAR's back. TAR's major groove is more electro-negative than the minor groove. Within the major groove, the upper part has more electro-negativity than the lower part. TAR's upper major groove is where Tat protein and ligand L22 bind [73, 74, 75].

investigating how L22 and TAR would encounter each other and how the ions would rearrange themselves during the L22-TAR encounter process. In all cases, L22 bound to TAR within a very short time-scale. In this section, we summarize the structural features of all six final states and proceed further into ion analysis.

4.3.3.1 Structural Properties of the Encountered Complexes

Figure 4.3.11 shows the snapshots taken at 5 ns of six MD simulations of the L22//TAR encounters. In all cases, L22 binds to TAR starting from everywhere within a distance of 40 Å. Therefore, these events are driven by electrostatic interactions. In two simulations, L22 approaches the TAR minor groove (panel (a)). In the other four cases, L22 binds to the TAR major groove (panel (b)). Among the four major-groove complexes, there are two cases in which L22 binds to the upper major groove of TAR (see the subfigures marked by the elliptic frames). These binding poses share analogous features to the NMR structure, the most important of which is that L22 binds to the same TAR pocket as seen in NMR experiment [24] and this is the Tat-binding pocket as well [73, 74, 75].

The average RMSDs of TAR with respect to the NMR structure in each simulation vary from 4.0 ± 0.4 to 4.7 ± 0.6 Å. In one of the two upper-major-groove binding poses, L22 was found along the RNA groove in a similar way as observed in the NMR structure but with the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template pointing upward instead of downward as with NMR (see the subfigure in the left top of Figure 4.3.11b). For notational convenience, we signal this

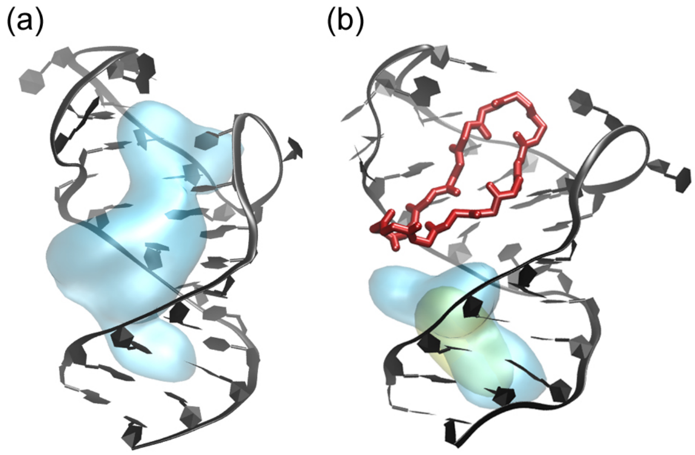

Figure 4.3.9: A schematic representation of the isosurface of ion density around TAR. (a) In apo-TAR, the K^+ ions (shown in cyan) mostly occupy the upper major groove of TAR. (b) In bound-TAR, K^+ ions are repelled by L22 and hence shift to the lower major groove; Cl^- ions (shown in yellow) are also found in this groove.

binding pose as **MD1**. In the other upper-major-groove pose (denoted as **MD2**), L22 has an orthogonal orientation thus not so similar to the NMR structure (see Figure 4.3.11b right bottom).

These two simulations, i.e, **MD1** and **MD2**, were then prolonged to 12 ns for further analysis. The RMSD of TAR in each simulation in this timescale oscillates around 4.6 ± 0.4 Å (see Figure B.0.1 in Appendix B). Salt bridges are formed mostly between the Arginine side chains of L22 and the TAR's phosphate groups or hairpin loop nucleobases. See Table 4.3 for the selected salt bridges in the last 5 ns of the simulation **MD1**. These are not the same interactions as found in NMR structure and in our previous simulation of L22-TAR complex presumably due to the short simulation timescale which does not allow the system to relax and finally reorganize itself into the correct binding pose. However, all of the stable salt bridges in **MD1** involve the L22's Arginine side chains. This confirms that Arginine residues play a critical role in molecular recognition and moreover suggests that electrostatics provides the driving force for the encounter process.

4.3.3.2 Ion Redistribution upon Complex Formation

A quantitative analysis of ion redistribution involves the calculation of ion occupancy using the so-called *proximity method* [188, 189, 190]. Each K^+ ion was assigned to the closest TAR atom within a cutoff distance of 5 Å. Analogously, each Cl^- ion was assigned to the closest L22 atom within the same cutoff. At every time frame, a summation of ion occupancies by atoms provides the ion occupancy for each nucleotide (in TAR) or residue (in L22). The K^+

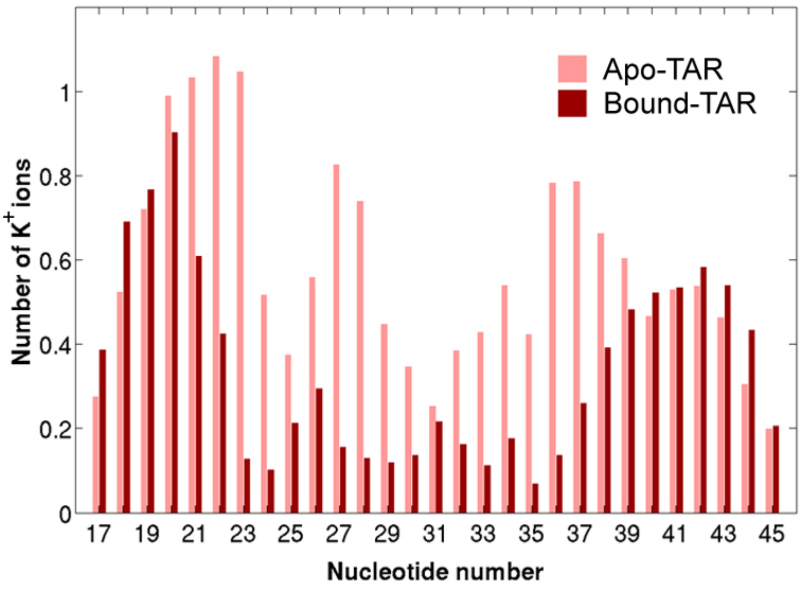

Figure 4.3.10: Number of K^+ ions within a distance of 5 Å around each nucleotide of TAR in both apo- (pink) and bound- (red) states.

and Cl^- occupancies in the simulation **MD1** were monitored as the peptide encountered TAR and compared with those of the isolated molecules in aqueous solution.

At the beginning of the simulation, K^+ ions are found in the upper major groove as also seen in apo-TAR (see Figure 4.3.12a). As L22 approaches TAR, it causes a displacement of the K^+ ions from the upper major groove, hence K^+ occupancy decreases significantly. After 5 ns, when L22 is fully bound to TAR, no ions are located inside the L22 binding pocket (see Figure 4.3.12b).

A similar and complementary picture is obtained for the Cl^- ions (see Figures 4.3.12c and d). In this case, these ions are found around L22 at the beginning of the simulation, but the approach of TAR causes a displacement of the Cl^- ions. As L22 approaches TAR, Cl^- ion occupancy decreases and in the complex, the ions are fully displaced. A very similar picture is obtained for the **MD2** simulation as well (see Figure B.0.2 in Appendix B).

It can be seen clearly that Cl^- ions are lost from L22 much earlier than K^+ ions are lost from TAR. This is due to the fact that the mass of L22 is much smaller than that of TAR. Hence, during the fast 5-ns encounter process, L22 moves quickly towards TAR while TAR is basically staying in the same place.

The total K^+/Cl^- occupancy around L22-TAR in **MD1** is shown as a function of the shortest distance between the molecular surfaces of L22 and TAR upon binding (see Figure 4.3.13). K^+ (and Cl^-) are highly distributed around TAR (and L22) when TAR and L22 are

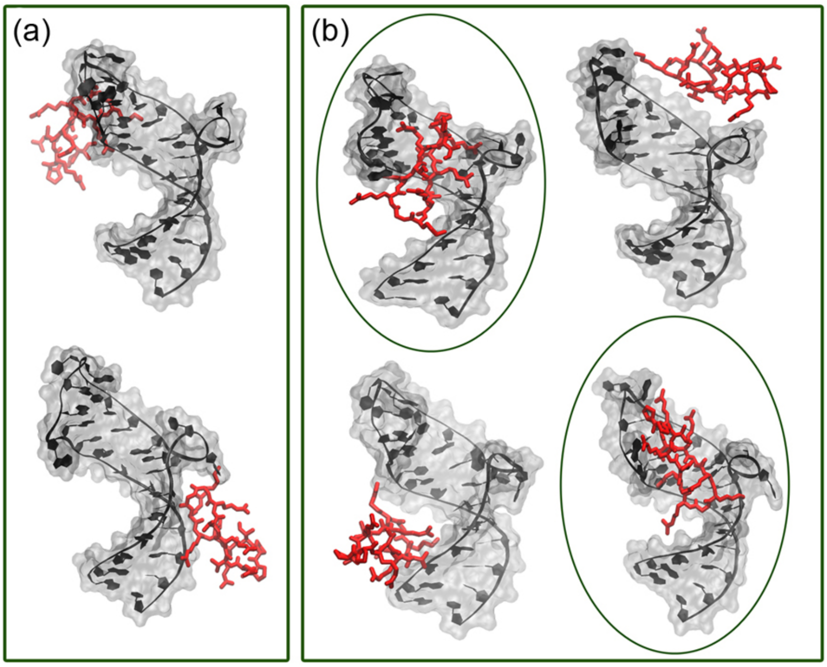

Figure 4.3.11: Final configurations of L22-TAR complex formations after 5 ns of MD simulations starting from six randomly positioned L22//TAR systems. (a) two minor-groove binding poses. (b) 4 major-groove binding poses, in which there are two poses (marked by elliptic frames) featuring the upper-major-groove binding mode, i.e., the same L22-binding pocket as observed in NMR studies and the same Tat-binding pocket as well [73, 74, 75].

separated, i.e., at surface distances larger than 8 Å. At distances of between 7.5 and 8 Å, the ion occupancies clearly decrease due to the presence of L22 in TAR's major groove.

L22	TAR	average distance (Å)	time course (%)
Arg8 (NH1)	A35 (O1P)	2.9 (0.2)	97
Arg8 (NH1)	A35 (O2P)	2.8 (0.1)	92
Arg8 (NH2)	G36 (O2P)	3.2 (0.1)	86
Arg8 (NH2)	G36 (O1P)	3.0 (0.1)	84
Arg8 (NE)	A35 (O2′)	3.1 (0.2)	76
Arg11 (NH1)	C30 (N4)	3.1 (0.1)	73
Arg11 (NH1)	G33 (O2′)	3.1 (0.2)	70
Arg11 (NH2)	G34 (O1P)	3.0 (0.2)	64
Arg11 (NH2)	G34 (O2P)	3.2 (0.2)	58
Arg11 (NE)	G34 (O6)	3.1 (0.1)	53
Arg9 (NH1)	C37 (O2P)	2.9 (0.1)	48
Arg9 (NH1)	C37 (O1P)	3.2 (0.2)	44
Arg5 (NH1)	A35 (N6)	2.8 (0.2)	35

Table 4.3: Key L22-TAR interactions in the last 5 ns of the simulation **MD1** shown with average length and time course (> 30%).

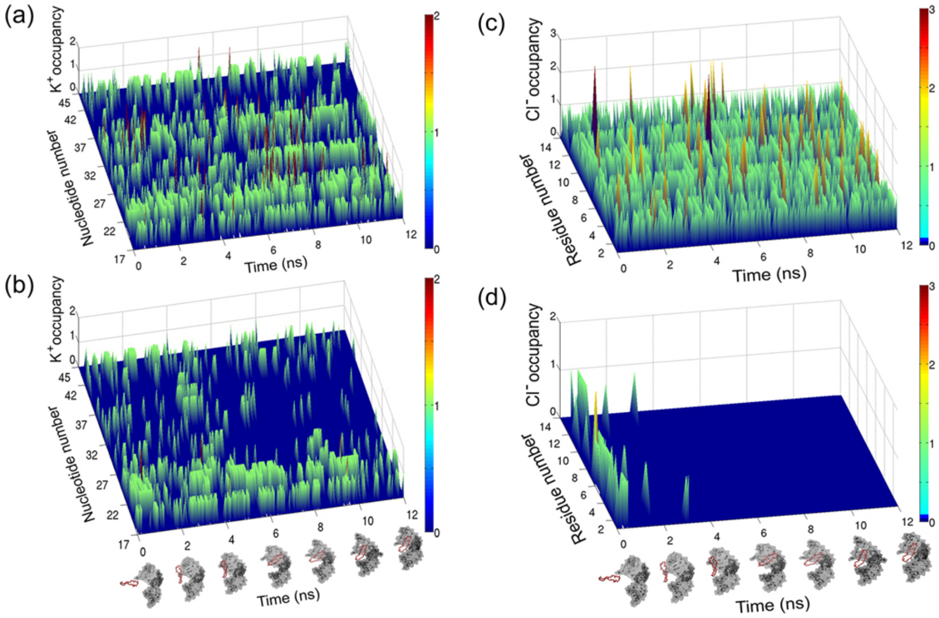

Figure 4.3.12: Ion occupancy as a function of simulation time calculated by proximity method [188, 189, 190]: K$^+$ occupancy along the RNA nucleotides in the simulations of apo-TAR (a) and **MD1** (b); Cl$^-$ occupancy along the L22 residues in the simulations of apo-L22 (c) and **MD1** (d).

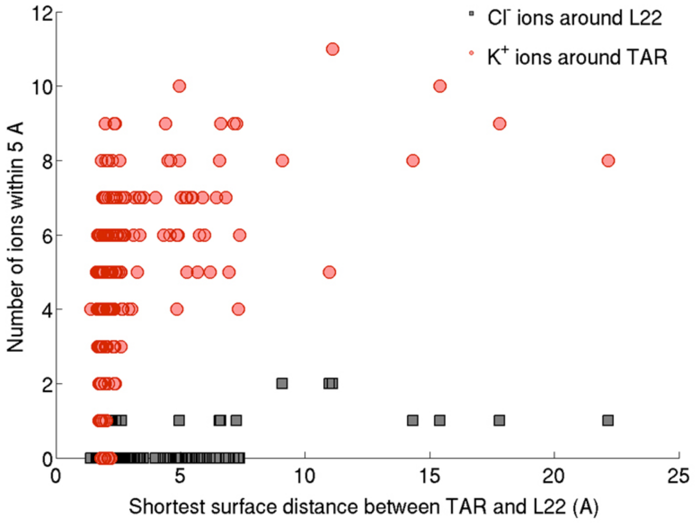

Figure 4.3.13: Ion occupancies as a function of the shortest distance between the molecular surfaces of TAR and L22 upon binding.

4.4 Discussions

4.4.1 Assessment of the Chosen RNA Force Field

The α/γ torsions in our simulations, under parmbsc0 reparametrization [6], still visit the energetically unfavorable *trans* conformation from time to time but with low probability (see Figure C.0.1 in Appendix C). Furthermore, in MD simulations of RNA based on ff99 force fields and its variants, the glycosidic χ torsion may adopt the *high-anti* conformation. After a few tens of nanosecond, this may distort RNA double helices to generate a ladder-like shape. A new parametrization has been recently developed to address this issue [191]. Our MD simulations were performed before this modification was introduced. Although such a reparametrization was not applied, the *high-anti* conformation was not observed in our simulations (in both apo-TAR and bound-TAR) (see Figure C.0.2 in Appendix C).

4.4.2 Critical Role of Electrostatic Interactions in Protein/RNA Molecular Recognition

In our study, the ligand is charged with +7*e* and the biomolecular target has a total charge of -28*e*. Despite an obvious total neutral box content, the long-range electrostatic interaction still acts as a driving force for the rapid formation of first encounter complexes. Indeed, different complex formations are found within 5 ns of several standard MD simulations. Therefore, our simulations confirm that binding between an RNA and a positively-charged peptide is a spontaneous process strongly driven by electrostatic interactions [192, 173].

Our short MD simulations only observe the intermediate complexes. They did not allow a more quantitative estimation such as the free-energy difference in the transition or how long it would take the biomolecules to rearrange themselves into the correct binding conformation. However, classical MD simulations for studying full biomolecular binding/unbinding events require considerable computational time. Therefore, enhanced sampling methods with a proper choice of CVs are preferably used to accelerate the transition process.

4.5 Concluding Remarks

Despite the limitations inherent to MD simulations which cover only the sub-μs time-scale [113], and of the known inaccuracies of force fields (see Section 2.2.1.3), the computational results presented here compare well with experimental NMR data and provide new insight into the paradigmatic TAR RNA and its complex with a lead inhibitor of viral replication.

In addition to demonstrating that our simulations satisfactorily reproduce both structural and dynamic properties of TAR and its complex with a Tat-mimic peptide, and that we obtain results in agreement with those obtained from NMR experiments, we also provide new information on the ligand binding process, as well as changes in ion distributions that occur upon complex formations. Most interestingly, ions are displaced from the two molecules even as they are at long distances from each other, and the peptide and RNA are able to spontaneously bind each other within the first few nanoseconds of simulation. The results from our short binding simulations also suggest that electrostatic interactions play an important role in molecular recognition.

These observations are the motivations for our next step: designing a CV that contains the description of electrostatic interactions (Chapter 3) and applying this CV to enhanced sampling methods.

Chapter 5

Bidirectional Steered Molecular Dynamics Simulations of the L22-TAR System

Contents

5.1 Overview

In the previous work presented in Chapter 4, we employed standard **M***olecular* **D***ynamics* (**MD**) simulations to investigate structural and dynamical properties of the L22-TAR complex. Our simulations agreed with experimental data in structural and dynamical properties of the system. Our simulations also confirmed that binding between an RNA and a positively-charged peptide is a spontaneous process strongly driven by electrostatic interactions [192, 173]. However, a well-known difficulty of atomistic MD simulations is that they can be used to follow the system dynamics on the microsecond timescale at most. The studies of slower conformational transitions in larger molecules do require some form of acceleration. Here we employed a bidirectional scheme of **S***teered* **M***olecular* **D***ynamics* (**SMD**) simulations with **D***ebye*-**H***ückel* **EN***ergy* (**DHEN**) as a **C***ollective* **V***ariables* (**CV**s). A total of 4.2 μs of SMD simulations were performed. Using this approach, we were able to make a blind prediction of the correct NMR binding pocket. Additionally, we also produced several putative complex structures.

In this chapter, we first describe the simulation protocols including the bidirectional steering scheme and the controlling algorithms as well as parameters. We then perform the free-energy reconstruction using the Hummer-Szabo and Minh-Adib **P***otential of* **M***ean* **F***orce* (**PMF**) estimator. We next describe and classify all the binding poses obtained from our binding SMD. We then conduct a thorough quantitative assessment on the two dominant poses in which the ligand is bound to TAR at the upper major groove, i.e., the same binding pocket as seen in NMR experiments.

5.2 Simulation Protocols

Here we perform constant-velocity SMD simulations (see Section 2.3.2) with DHEN (Equation (3.2.15)) chosen as a CV. The external harmonic potential used to steer the CV in this case is given by

$$V(t) = \frac{k}{2}(z(t) - z_{restraint}(t))^2 = \frac{k}{2}\left[\frac{1}{k_B T \epsilon_w}\sum_{j\in B}\sum_{i\in A} q_i q_j \frac{e^{-\kappa|\mathbf{r}_{ij}(t)|}}{|\mathbf{r}_{ij}(t)|} - vt - z_0\right]^2, \quad (5.2.1)$$

where k is the spring constant, v is the velocity of the steering, and z_0 is the initial restrained value of DHEN CV.

The cumulative work perform in such a steering is then given by

$$W(t) = \sum_{t'=0}^{t} vk(z(t') - z_{restraint}(t'))\Delta t. \tag{5.2.2}$$

5.2.1 Bidirectional SMD Scheme

A total of 4.2 microseconds of simulations have been performed in a bidirectional steering scheme. Hereafter we refer to the unbinding direction as *forward* SMD and the binding direction as *backward* SMD.

5.2.1.1 Forward Scheme

Our forward SMD scheme included:

(1) 20 ns of NPT MD simulation starting from the NMR structure of L22-TAR complex. An average DHEN (-140 kJ/mol), and its standard deviation ($\sigma = 3$ kJ/mol) were calculated from this equilibrium simulation.

(2) 64 ns of CV-restrained simulation in which the value of the DHEN was restrained to the average value of -140 kJ/mol. A spring constant $k = 0.2$ kJ mol^{-1} nm^{-2} was used here and in the following SMD[1].

(3) Configurations were then extracted every 1 ns and used as initial structures for 64 forward (unbinding) SMD simulations (25 ns each), in which the DHEN was pulled from the value of $z_0 = -140$ kJ/mol to -30 kJ/mol. This target value has been chosen large enough so that the two molecules are completely separated. Indeed, among 64 final structures of unbinding SMD simulations, the smallest center-to-center distance between L22 and TAR is about 32 Å while the smallest distance between an L22 atom and a TAR atom is about 7 Å.

5.2.1.2 Backward Scheme

Our backward SMD scheme was consisted of the following steps:

(1) Each of the structures obtained at the end of the SMD was equilibrated for 1 ns, restraining the DHEN at -30 kJ/mol.

(2) A random configuration was then extracted from each CV-restrained simulation and used as the starting structure for another set of 64 backward (binding) SMD (25 ns each), in which the DHEN CV was pulled from $z_0 = -30$ kJ/mol back to -140 kJ/mol with the same speed and spring constant as in the forward SMD simulations.

[1] Empirical rule for choosing the spring constant in SMD simulations: $k \approx k_B T/\sigma^2$

5.2.2 Additional Forward SMD Simulations from the Upper-Major-Groove Binding Pose

Among 64 bound configurations at the end of the backward SMD simulations, we found two dominant classes in which L22 binds to the TAR's upper major groove, i.e., the same binding pocket observed in NMR studies. To further quantify the difference between these two classes, for each class, we selected the structures associated with the lowest external work. We then repeated the same forward SMD procedure as described in Section 5.2.1.1, which, for each selected structure, included *(i)* 30 ns of MD equilibration for step *(1)*, *(ii)* 12 ns of CV-restraint MD for step *(2)* at the same DHEN value (-140 kJ/mol) and spring constant (0.2 kJ mol^{-1} nm^{-2}), and *(iii)* 12 forward SMD simulations (25 ns each) for step *(3)*.

5.2.3 System Preparation and Control Parameters

System Preparation

We used a truncated octahedral box of explicit water, in which L22 and TAR could reach a center-to-center distance of at least 40 Å. The box contained 11,780 water molecules. An excess ion concentration of 10 mM was set in all simulations to reproduce the experimental conditions [83], which resulted in 23 K^+ cations and 2 Cl^- anions. We employed TIP3P model [174] for water, AMBER ff99SB-ILDN force field [193] for L22, ff99SB-ILDN with parmbsc0 reparametrization [6] for TAR. When using the standard ff99SB-ILDN force field for K^+ and Cl^- ions at the ion concentration of 150 mM, we experienced the growing of ion crystallization after the first 5 ns of MD simulation (data not shown). In fact, the AMBER ff9X force fields, i.e., ff94 [194], ff98 [195], and ff99 [196, 197] and their variants, have been reported to facilitate the ion crystallization due to the incorrect parametrization which causes the imbalance between cation-anion interactions[2] [200, 201, 202]. Therefore, the ff99SB-ILDN force field corrected by new ions' reparametrization[3] [203] was used for the K^+ and Cl^- ions in our simulations.

Control Parameters

All standard MD simulations were performed using GROMACS 4.5.5 [204]. Additionally, SMD simulations with DHEN CV were performed using an in-house version of PLUMED 1.3 integrated with GROMACS 4.5.5.

The Particle-Mesh Ewald method [146, 179] was used to treat the long-range electro-

[2] The AMBER ff9X force fields mix the AMBER-adapted Åqvist parameters [198] for the cations and Dang parameters for Cl^-[199].

[3] This new reparametrization involves reoptimizing the parameters of the Lennard-Jones potential for the ions.

static interactions with a real space cut-off of 12 Å. The same cut-off was also used for the van der Waals interactions. The simulation time-step was 2 fs and the non-bonded atom pair list was updated every 10 steps.

In all simulations, temperature was kept constant at 300 K using the velocity rescaling algorithm [133]. In all equilibrium MD simulations, which were performed under NPT condition, pressure was kept constant at 1 atm using a Parrinello-Rahman barostat [205].

In all SMD and CV-restrained simulations, Equation (3.2.15) was used to determined the DHEN CV. Sets A and B contained all atoms of L22 and TAR respectively. Atomic charges were extracted from the ff99SB-ILDN force field. A value of 80 was chosen for the dielectric constant of water ϵ_w, which, together with the ionic strength $I = 10$ mM, resulted in the Debye-Hückel parameter $\kappa \simeq 0.033$ Å^{-1}.

5.3 Results

5.3.1 Free-Energy Reconstruction from SMD Simulations

5.3.1.1 Hummer-Szabo PMF Estimator for Unidirectional SMD Simulations

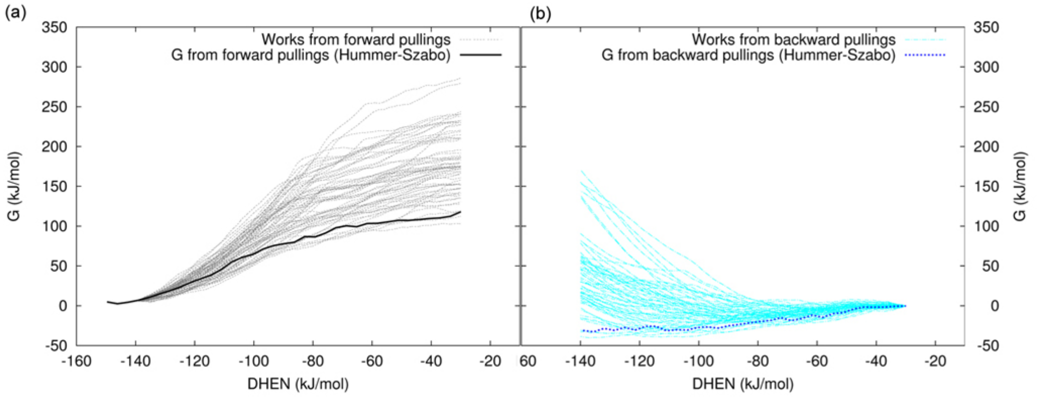

Figure 5.3.1: (a) Nonequilibrium works (gray lines) and PMF calculated by the Hummer-Szabo estimator (black line) as functions of DHEN CV from 64 forward SMD simulations. (b) Nonequilibrium works (cyan lines) and PMF calculated by the Hummer-Szabo estimator (blue line) as functions of DHEN CV from 64 backward SMD simulations. The resulted PMFs are those of the unperturbed system and are strongly governed by the low-work trajectories.

Figure 5.3.1a shows the nonequilibrium works and the PMF as a function of DHEN CV calculated by the Hummer-Szabo estimator from 64 forward SMD simulations. Similarly, those of backward SMD simulations are shown in Figure 5.3.1b. Since the Hummer-Szabo estimator is applied, the resulted PMFs are those of the unperturbed system, which is the system itself without any external perturbations. Notably, in both cases, the behavior of

PMF is strongly governed by the low-work trajectories. This is due to the property of the exponential average $\langle e^{-\beta W} \rangle$, which gives more weight to the lower work values.

5.3.1.2 Minh-Adib PME Estimator for Bidirectional SMD Simulations

The free energy of the unperturbed system as a function of DHEN can be reconstructed using the Minh-Adib PMF estimator (Equation (2.4.40)) which utilizes the information from both forward and backward pullings. Figure 5.3.2 shows a quantitative comparison between the Hummer-Szabo unidirectional and Minh-Adib bidirectional estimators.

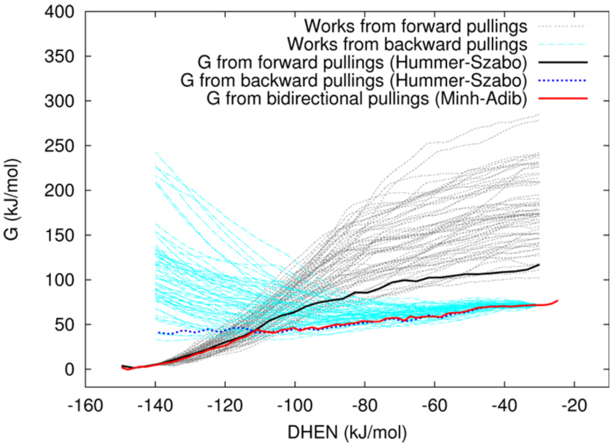

Figure 5.3.2: Reconstruction of PMF of the unperturbed system as a function of DHEN CV by the Minh-Adib estimator for bidirectional steerings (red line) in comparison with those by the Hummer-Szabo estimator for forward steerings (black line) and backward steerings (blue lines). The works from forward and backward pullings are also shown (gray lines and cyan lines respectively). Backward works are shifted by ΔF as estimated from Equation 5.3.8. The Minh-Adib estimator outperforms the Hummer-Szabo one by providing the optimal combination of trajectories from both forward and backward directions.

Our calculations confirmed that the bidirectional estimator provides a better result than the unidirectional one. Indeed, while the Hummer-Szabo method for combining unidirectional steerings is strongly biased when the system moves away from its starting equilibrium state; the Minh-Adib method provides an optimal way to combine both steering directions into the nonequilibrium path averages. Especially, when the system is still closer to the starting equilibrium state, Minh-Adib estimator still gives larger weights to the trajectories on the direction leaving this state. However, when the system is out of equilibrium and getting closer to the ending equilibrium state, Minh-Adib estimator favors the time-reversed counterparts of the trajectories on the reversed direction. Therefore, the bidirectional Minh-Adib method outperforms the unidirectional Hummer-Szabo method by optimally combining forward and backward trajectories to give the least biased PMF estimation.

5.3.1.3 Reweighting Scheme for Projecting the PMF on Center-to-Center Distance

Divergence of the PMF at the Unbound State. If we define the unbound state as the state in which the two molecules are infinitely far apart, we have the PMF diverge in the unbound state. Indeed, the further away the two molecules are, the larger conformational space the system possesses, and hence the larger the entropy becomes. However, it is nontrivial to quantify such an entropic contribution in terms of DHEN CV. On the contrary, it is easy to determine the entropic contribution as a function of the center-to-center distance d between the two molecules. Indeed, the partition function of the system is proportional to the surface area of a sphere with the radius d

$$Z(d) \sim 4\pi d^2. \tag{5.3.1}$$

The free energy of the system is thus given by

$$G(d) = E(d) - TS(d) = E(d) - Tk_B \ln Z(d) = E(d) - 2k_B T \ln d + C. \tag{5.3.2}$$

where C sums up all constant terms. When d is large enough, the change in total energy is only given by the change in electrostatic interaction. We can then rewrite the free energy as

$$G(d) = \mathrm{DHEN}(d) - 2k_B T \ln d + C. \tag{5.3.3}$$

For systems with two opposite charges like our case, during the increment of d, $\mathrm{DHEN}(d)$ increases toward 0 and $-2k_B T \ln d$ decreases. At some point, the decrement of the entropic term starts to take over the increment in the value of DHEN. As d goes to infinity, DHEN becomes 0 and $G(d)$ finally diverges due to the divergence of $-2k_B T \ln d$.

To calculate the free-energy difference between the bound and unbound states, we need first to determine the free energy value in each state. It is more advantageous if we can quantify and then compensate the entropic contribution and thus make the PMF converge to *zero* in the unbound state. DHEN may be a good CV for guiding the biased simulations, but not a convenient CV for manipulating the PMF. In this respect, the center-to-center distance d can be a better choice.

Projection of PMF on the Center-to-Center Distance Our proposed reweighting scheme allows manipulating the PMF by projecting it on any *a posteriori* chosen CV. We here apply this reweighting scheme to compute the PMF as a function of the distance between the centers of mass of the two molecules (see Equation (2.4.55)). The entropic contribution can now be easily evaluated ($-2k_B T \ln d$) and added to the PMF, and hence allows estimating the free-energy difference between the bound and unbound states, which turns out to be approximately 85 ± 5 kJ/mol (see Figure 5.3.3). This value is larger than that obtained from experiment, which was approximately 52 kJ/mol . As it will be discussed in Section 5.4, we prove that this discrepancy does not come from statistical inaccuracies, therefore it is presumably due to a combination of the inaccuracy of the force fields and the uncertainty of the experimental measure.

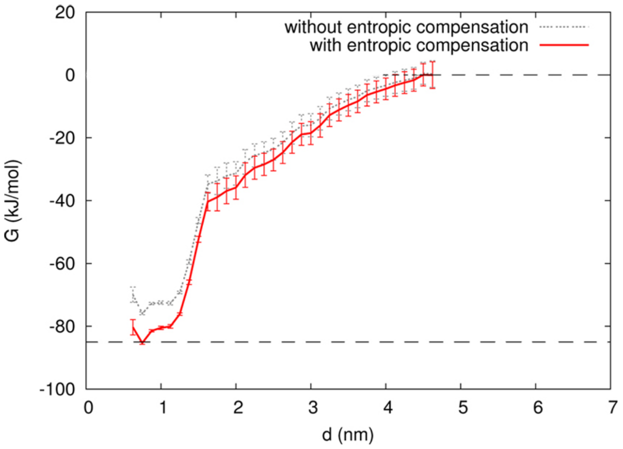

Figure 5.3.3: Free energy as a function of the geometric center-to-center distance (d) obtained by the reweighting scheme (Equation (2.4.55)) without any entropic compensation (black solid line) and with an entropic compensation of $2k_BT \log d$ (red solid line). The plots have been shifted so that free energies are aligned at $d = 4$. The projection of free-energy profile onto distance allows defining both bound and unbound states. The free-energy difference between these states is approximately 85 ± 5 kJ/mol.

5.3.2 Structural Features of L22-TAR Complexes Obtained from Backward SMD Simulations

64 binding (backward) SMD simulations, in which the DHEN CV was pulled from -30 kJ/mol to -140 kJ/mol, all ended up with L22-TAR complexes. The binding poses of L22 to TAR can be classified in the following way (see also Figure 5.3.4 and Table 5.1)

(i) L22 binds to TAR at the major groove in 51 complexes (80%), among which 33 complexes (52%) can be classified as upper-major-groove binding (i.e., the same binding pocket as of the Tat protein and as seen in NMR experiment [24, 73]), 12 complexes (19%) feature lower-major-groove binding, and 6 complexes (9%) have L22 bind to TAR at the region lying between the upper and lower major groove (referred to as middle major groove hereafter).

(ii) L22 binds to TAR at the minor groove in 10 complexes (16%), among which there are 4 complexes (7%) classified as upper-minor-groove binding and 6 complexes (9%) classified as lower-minor-groove binding.

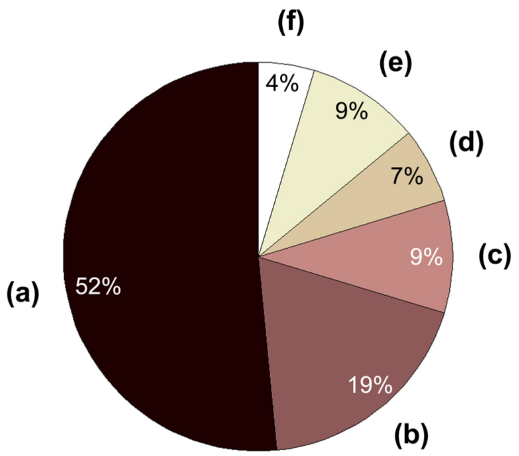

Figure 5.3.4: A pie chart representation of the L22-TAR binding poses from binding SMD simulations which can be classified as: **(a)** upper-major-groove, **(b)** lower-major-groove, **(c)** middle-major-groove, **(d)** upper-minor-groove, **(e)** lower-minor-groove, and **(f)** otherwise. More than half of the simulations (i.e., 33 in a total of 64) blindly bring L22 to the upper major groove pocket, the same binding pocket as of the Tat protein and as seen in NMR experiment [24, 73].

(iii) 3 complexes (4%) do not belong to any of the above categories.

Major groove (80%)							Minor groove (16%)		Others
Upper (52%)			Lower (19%)			Middle	Upper	Lower	
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**				
31%	13%	8%	9%	7%	3%	9%	7%	9%	4%

Table 5.1: Occurrence of L22-TAR binding poses obtained from 64 binding SMD simulations. To better classify the binding poses at the major groove, we subdivide the upper-groove and lower-groove binding poses into smaller groups. For the upper groove, **(1)** represents the pose in which the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template of L22 points outward TAR; **(2)** denotes the pose with the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template pointing inward TAR; and **(3)** contains the rest of the upper-major-groove-binding complexes that are not trivial for the determination of L22 orientation. Similarly, poses **(4)**, **(5)**, and **(6)** respectively represents the same classification criteria for the lower-major-groove binding. It is noteworthy that pose **(2)** is the binding pose observed in the NMR experiment [24].

In the upper-major-groove binding, we found 20 complexes (31%) in which the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template of L22 points outward TAR (i.e., pose **(1)** in Table 5.1 and Figure 5.3.5a) and 8 complexes (13%) in which the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template points inward TAR (pose **(2)** in Table 5.1 and Figure 5.3.5b). Similarly, in the lower-major-groove binding, we also found 6 complexes (9%) with the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template pointing outward (pose **(4)** in Table 5.1 and Figure 5.3.5c) and 4 complexes (7%) with the $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro template pointing inward (pose **(5)** in Table 5.1 and Figure 5.3.5d). Pose **(2)** represents the same binding pose as observed in the NMR experiment [24]. However, it only appears as the second dominant pose in our simulations. The first dominant pose, which has the same binding pocket but with $^{\mathrm{L}}$Pro$-^{\mathrm{D}}$Pro pointing to the opposite direction, occurs with a higher probability.

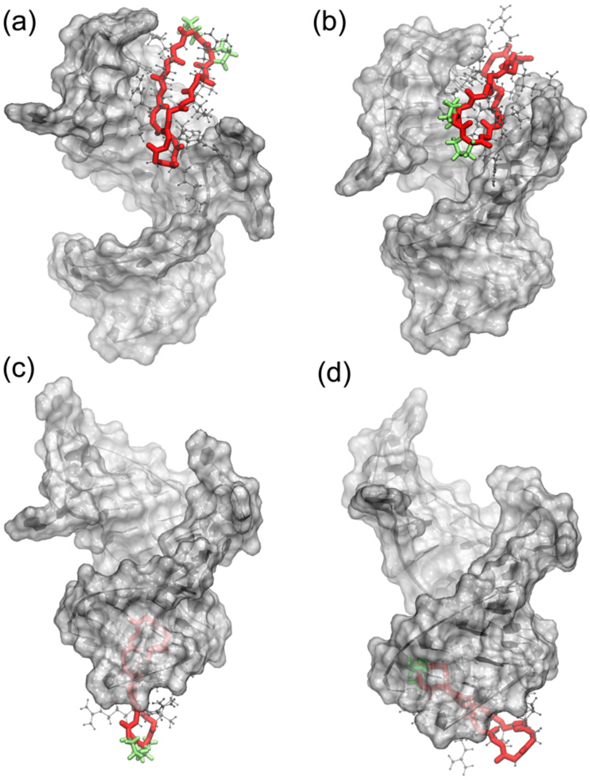

Figure 5.3.5: Structures of four dominant binding poses obtained from 64 binding SMD simulations including (a) upper-major-groove binding pose with the ^{L}Pro$-^{D}$Pro template, colored in green, points outward TAR (i.e., pose **(1)** as classified in Table 5.1), (b) upper-major-groove binding pose with ^{L}Pro$-^{D}$Pro points outward TAR (pose **(2)**), (c) lower-major-groove binding pose with ^{L}Pro$-^{D}$Pro points outward TAR (pose **(4)**), and (d) lower-major-groove binding pose with ^{L}Pro$-^{D}$Pro points inward TAR (pose **(5)**). Pose **(2)** is close to what observed in NMR experiment.

5.3.3 Quantitative Comparison in Stability of the Two Dominant Upper-Major-Groove Binding Poses

A quantitative assessment of the relative stabilities between pose **(1)** and pose **(2)** can be done by looking at the PMF.

5.3.3.1 PMF Comparison

As discussed in the previous Section, pose **(1)** in Figure 5.3.5 is at the same time the most frequent and the one for which the lowest work is performed during the binding SMD simulations. In pose **(2)**, the ligand occupy the same binding pocket in a different orientation, which is consistent with that obtained from NMR data. For a quantitative comparison between pose **(1)** and pose **(2)**, we performed 12 unbinding SMD simulations starting from the complex obtained by the lowest work in each pose (i.e., the globally lowest work in case of pose **(1)** and the sixth-lowest work in case of pose **(2)**). We then combined these 12 unbinding simulations of each pose with a selected set of the previous binding simulations which ended on the same pose (i.e., 20 binding simulations resulting in pose **(1)** and 8 simulations resulting in pose **(2)**). Equation (2.4.40) was then used to calculate the free energy as a function of the restrained DHEN CV based on such combination of each pose (see Figure (5.3.6)). Remarkably, this equation permits this calculation despite different numbers of forward and backward trajectories. The free-energy difference between the two end states associated with pose **(1)** is larger than that of pose **(2)** (79 ± 7 versus 69 ± 6 kJ/mol respectively), thus pose **(1)** can be considered more stable than pose **(2)**.

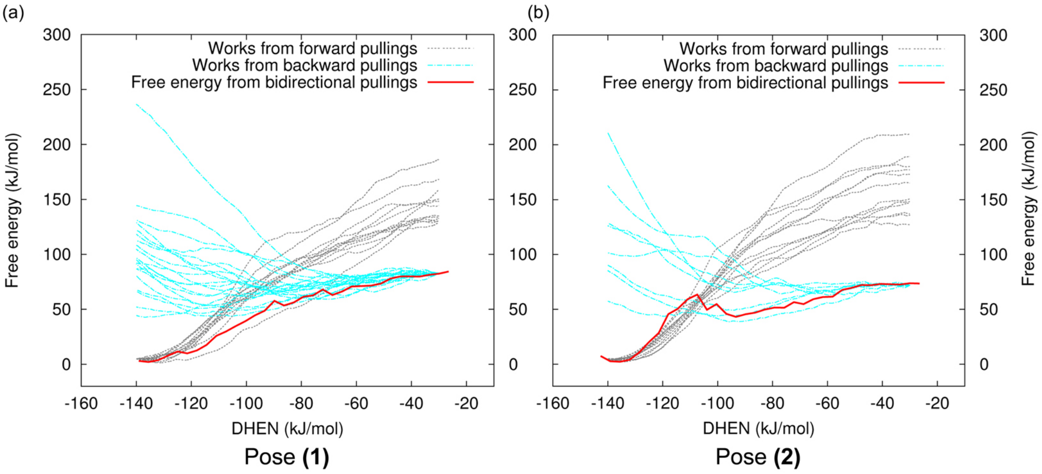

Figure 5.3.6: Reconstruction of free energy of the unperturbed system in pose **(1)** (a) and pose **(2)** (b) as a function of DHEN CV (solid red line). The works from forward and backward pullings are also shown (gray lines and cyan lines respectively).

Applying the proposed reweighting scheme on the center-to-center distance CV, we found that the free-energy differences as functions of distance in both poses **(1)** and **(2)** are comparable to that of the previous calculation shown in Figure 5.3.3 on the whole set of 64 forward (starting from NMR structure) and 64 backward pullings (see Figure 5.3.7).

5.3.3.2 Verification of the Robustness of the Comparison

To test the robustness of the comparison, we repeated 500 times of solving the BAR equation (5.3.8) (which is a special case of Equation (2.4.41) when one considers only the end states) on the randomly chosen half-set, i.e., 6 unbinding and 10 binding simulations for pose **(1)** and 6 unbinding and 4 binding simulations for pose **(2)**. The resulted free-energy difference from 500 BAR calculations for each pose can be found in Figure 5.3.8. The average values of free-energy difference are 75 ± 6 and 66 ± 6 kJ/mol for pose **(1)** and pose **(2)** respectively. Despite random choice of the half-set of simulations to be involved in BAR calculations, pose **(1)** consistently shows a higher stability than pose **(2)**.

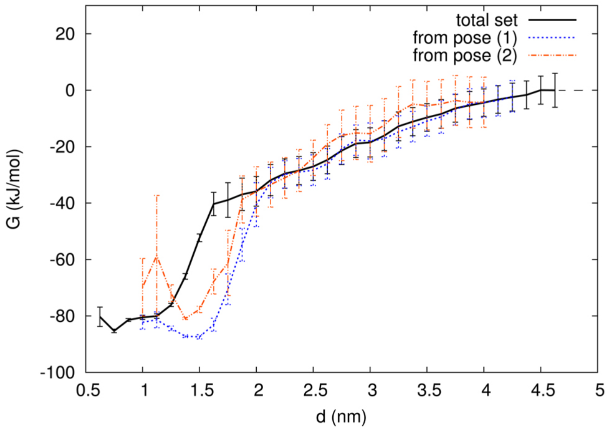

Figure 5.3.7: Free energy as a function of the center-to-center distance obtained by the reweighting scheme performed on the whole set of simulations (black solid line) and the set of simulations coming from and to pose **(1)** (blue dotted line) and pose **(2)** (red dashed line). free-energy difference between the bound and unbound states have comparable values in all cases.

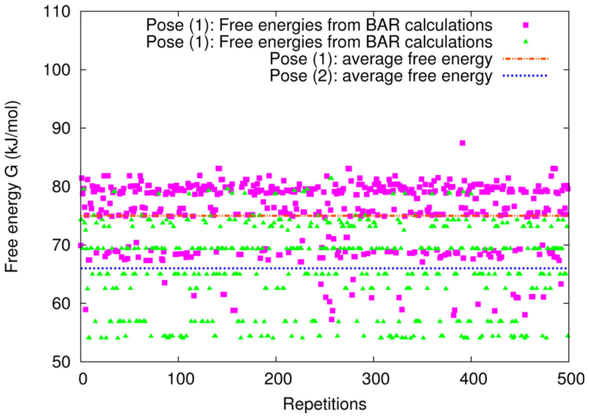

Figure 5.3.8: Results of 500 times of BAR free energy calculations on a randomly choice of half-set (i.e., 6 unbinding realizations and 4 binding realizations) for both pose **(1)** (magenta squares) and pose **(2)** (green triangles). Pose **(1)** consistently shows more stability (larger free energy) than pose **(2)**. The average free-energy difference values are 75±6 kJ/mol (red line) and 66±6 kJ/mol (blue line) respectively.

5.4 Discussions

5.4.1 Effectivity, Computational Efficiency, and Generality of the Proposed Electrostatic-Based Collective Variable

Our proposed DHEN CV (expressed in (3.2.15)) includes only the intermolecular electrostatic interactions thus does not contain the noise coming from intramolecular interactions or interactions with the ionic solvent. Our CV is a function of not only atomic coordinates but also ionic strength, temperature, solvent dielectric constant, and atomic charges accessible from the force fields. Therefore, it is expected to be more "selective" and effective than the conventional center-to-center distance CV in describing peptide/RNA binding/unbinding processes. Indeed, the distance CV may disfavor the right complex formation by not taking into account the charge-charge interaction, an important driving force in most peptide/RNA recognitions.

Besides the parameters that are trivially determined from the simulation setups, our proposed CV does not requires any other extra parameters that need to be updated during the simulations. This is the computational advantage of this CV.

In addition, our proposed CV has a general formalism that, in theory, is applicable for any binding problems. However, in practice, due to its nature of describing electrostatic interactions, we recommend using it for the systems in which binding events are dominantly driven by electrostatics.

5.4.2 Bidirectional Steering Outperforms Unidirectional Steering

Steering involves out-of-equilibrium processes. Although the unidirectional Jarzynski's equality and the Hummer-Szabo PMF estimator allow reconstructing the free energy from nonequilibrium works, the resulted free energy is strongly dominated by the low work values due to the behavior of the exponential average. These low work values are associated with the trajectories of the "rare" events which are poorly sampled. This poses a convergence challenge to unidirectional SMD simulations. Indeed the more the system departs from its initial equilibrium state, the more the unidirectional estimators tends to *overestimate* the free energy change.

The bidirectional estimators such as the one introduced by Minh-Adib and our proposed reweighting scheme provide an optimal combination of the works from both forward and backward processes. When the forward and backward processes are properly combined, the overestimation of free energy in both directions are then "averaged" out. Bidirectional estimators hence outperform unidirectional ones (see our results and detailed discussions in Figure 5.3.2).

5.4.3 Verification of the Statistical Accuracy

The procedure of analyzing a selected subset of the binding and unbinding trajectories allows us to assess the statistical accuracy of the results. Indeed, bidirectional pulling protocols can still exhibit large statistical errors when backward pullings do not properly bring the system to the correct starting point. However, by performing the pulling out of the actually reached bound pose (Section (5.3.3.2)) we can guarantee that their stability is fairly evaluated.

5.4.4 "Blind" Prediction of the Binding Pocket

Employing bidirectional SMD simulations with our proposed DHEN CV, we were able to find the binding pocket in agreement with the NMR structure, i.e., the upper major groove of TAR. Although the L22-TAR binding pose and its structural properties were well characterized in experiments, we did not use any of these experimental observations to make further bias or constraint in our simulations. Indeed, our pullings were performed in a "blind" manner, in which L22 is not constrained to bind to the known NMR binding pocket but rather free to decide its encounter paths upon increment of electrostatic interactions. In such manner, we were still able to find two dominant binding poses, in both of which L22 bound to the correct pocket as observed by experiments. This result not only confirms the assumption that electrostatics plays an important role in L22-TAR binding but also strongly justifies the use of DHEN, an approximation of the electrostatic interaction free energy, as a CV in accelerated simulations.

TAR is a rather large RNA containing 29 nucleotides with a complicated double-helix conformation featured by two stems, a bulge, and an apical loop. The apical loop partially closes the access to the upper major groove associated with the upper stem, which is also the binding pocket of Tat [73]. Any designed molecule able to bind to TAR at this pocket is a promising candidate for HIV-transcription inhibition. L22 is not a very small molecule compared to its receptor TAR (269 versus 930 atoms). Moreover, L22 has a rigid $\beta-$hairpin backbone and long side chains (i.e., mostly composed of Arginine side chains), which make it difficult to navigate and end up inside a partially closed pocket. Interestingly, in experiments, L22 was reported to bind and fit completely in this pocket. And we were also able to reproduce such a non-trivial binding mode only by using SMD simulations pulling on an electrostatic-potential-energy-based CV without any further guidance from experiments.

5.4.5 Accuracy of the Predicted Binding Affinity

As we discussed above, we are confident in the statistical accuracy of our calculations. Thus, we are convinced that the discrepancy between our estimate of the affinity and the one reported in ref. [24] should be ascribed to other causes such as the difference between the anion type we used (Cl^-) and the one used in the experiments (mixture of HPO_4^{2-} and $H_2PO_4^-$); the not so large simulation box and the resulting periodic boundary artifacts which are even more amplified in smaller simulation boxes; and probably the inaccuracies of the force fields. These latter inaccuracies could be likely associated with *(i)* the challenging description of the multi-degree-of-freedom sugar-phosphate backbone using a constant-point-charge model [110, 111], *(ii)* the difficulties in describing the RNA non-canonical structural elements [114, 7], i.e., the bulge and hairpin loop in our case, *(iii)* subtle force-field dependence on ionic strength and types [140, 206], and most importantly *(iv)* the well-known inaccuracies of the non-polarizable force fields in describing the electrostatic interaction between the RNA and the anions (i.e, Cl^- ions) as well as the strong electrostatic interaction between the highly polarizable phosphodiester moiety of RNA and the positively charged atoms [113, 3] including the cations K^+ and those of the peptidic ligand L22 in our case.

5.5 Concluding Remarks

We have performed a total of 4.2 microseconds of SMD simulations in a bidirectional scheme with the electrostatic-based DHEN CV. Using our proposed reweight method, we were able to reconstruct the PMF as a function of the distance between the L22 and TAR. This resulted in a larger free-energy difference between the unbound and bound states when compared to the experimental value. By an extensive analysis method that uses random half-sets of data, we proved that this overestimation was free from statistical inaccuracy. It was then presumably due to both the difference in anion type used in our simulations and experiments and the improper description of electrostatics by non-polarizable force fields.

Despite of these defects, our simulations were still able to blindly predict the correct binding pocket, i.e., TAR's upper major groove. Besides reproducing the NMR binding pose, we also found another pose which consistently showed more stability than the NMR one. In this new pose, the ligand L22 occupied the same binding pocket but in an opposite orientation, namely the ^{D}Pro$-^{L}$Pro pointed upward instead of downward as observed by NMR. This new pose is totally reasonable as L22 is a rather symmetric ring with Arginine residues equally distributed. There is hence no reason why the ^{L}Pro$-^{D}$Pro template has to point downward only as discovered in NMR experiment.

Chapter 6

Conclusions and Perspectives

6.1 Conclusions

Designing drugs targeting RNA is, at the same time, a promising and challenging task [1, 2, 207, 208, 106]. The main challenges come from RNA's highly charged nature and structural flexibility [3]. *In silico* studies, especially **M***olecular* **D***ynamics* (**MD**) simulations, can provide important insights on the molecular recognition and structural adaptation processes, which are highly critical for ligand/RNA binding but are difficult to understand from static X-ray and NMR structures [113, 209, 3].

Here we perform all-atom standard MD and enhanced sampling simulations on a system of HIV-1 TAR RNA in complex with the cyclic peptide inhibitor L22 (pdb code: 2KDQ [24]). L22 was designed to be a competitor inhibitor of the viral Tat protein for the binding site on the viral TAR RNA element [83, 24]. In the normal viral cycle, Tat/TAR interaction enhances the viral full-length transcription process and is thus crucial for HIV-1 replication. *In vitro* studies showed that L22 binds to TAR with a high binding affinity of 1 nM[1] at the upper part of TAR major groove, which is also the Tat-binding pocket [85]. L22 appeared as a promising anti-HIV transcriptional inhibitor; however, the details of the molecular recognition mechanism upon TAR binding were unknown.

Our first approach to studying this system is to employ all-atom standard MD simulations [139]. We performed ~0.5 μs of MD simulations starting from several configurations extracted from the NMR structure of L22-TAR complex (i.e., including apo-TAR, apo-L22, and bound as well as unbound structures of the complex). These simulations:

1. show a quantitative agreement with experiments on the NMR order parameter values, and hence on the *dynamical properties* of important TAR's nucleotides. Our simulations

[1] 8 nM in case of Tat-TAR binding [86]

also provide complementary order parameter values missing in NMR experiments,

2. confirm the *structural features* of TAR including *(i)* the flexibility decrease of TAR upon ligand binding, *(ii)* the more compact shape of bound-TAR, and *(iii)* the conformational changes at the hairpin-loop region of TAR (which is directly involved in the ligand binding site) during the time course of 200 ns,

3. reproduce all key intermolecular hydrogen bonds and hydrophobic contacts observed in NMR experiments. Our simulations also provide complementary information on the hydrogen-bond lengths and their time course, thus validate the stability of the L22-TAR complex within the time-scale of 200 ns,

4. show the difference in hydration and ion distribution around apo- and bound-TAR. Moreover, our simulations of the encounter process also give insights on ion redistribution upon ligand binding. This information is only accessible by all-atom MD simulations,

5. reproduce the NMR binding pocket (in 2 of 6 encounter simulations), which is the same as the Tat-binding pocket. Our simulations also show that *(i)* binding between L22 and TAR is a spontaneous process strongly driven by electrostatic interactions and *(ii)* arginine residues play important roles in L22-TAR molecular recognition. The former has been shown to be also the case of Tat-TAR binding [192] and generally true for binding between RNA and positively charged proteins [173].

The 200-ns-long MD simulation of the L22-TAR complex is, however, insufficient for the molecular dissociation to occur. This is a well-known limitation of atomistic MD simulations in studying slow conformational transitions of large molecules. Therefore, enhanced sampling methods with a proper choice of **C***ollective* **V***ariables* (**CV**s) are preferably used to accelerate such transition processes. Choosing a "good" CV that can describe and differentiate all relevant states is essentially a challenging task.

Motivated by the observation that binding between an RNA and a positively charged ligand is driven by electrostatic interaction and by the idea of using (potential) energy of a system as a CV [16, 17, 18, 19, 20], we here propose an electrostatic-based CV that is an approximation of the *intermolecular electrostatic component of free energy*, which is given by the Debye-Hückel formalism and is easily computed during the simulations. Our proposed CV, called **D***ebye*-**H***ückel* **EN***ergy* (**DHEN**), has the following characteristics:

1. DHEN includes only the intermolecular electrostatic interactions and is thus computationally efficient.

2. DHEN is not only a function of atomic coordinates, but also of the ionic strength, the temperature, the solvent dielectric constant, and the atomic charges defined in the

force fields. By construction, DHEN is expected to be more effective than the conventional center-to-center distance CV in describing peptide/RNA binding/unbinding processes.

3. DHEN has a general formalism that can be applicable to any binding event involving electrostatic interactions.

The DHEN CV is a general and computationally efficient CV that uses straightforward definitions for its parameters. This obviously comes at the cost of accuracy, which, however, appears to be well justified since the approximated electrostatic free energy of the system is used only as a CV for guiding the exploration of the conformational space.

We next perform bidirectional **S***teered* **M***olecular* **D***ynamics* (**SMD**) simulations of the L22-TAR complex system with the bias applied on our proposed DHEN CV [210]. In addition, we also propose a reweighting method to project the **P***otential of* **M***ean* **F***orce* (**PMF**) on any *a posteriori* chosen CV in a bidirectional steering scheme. This post-processing technique permits looking at the PMF in a different perspective which can be instructive in some cases. We find that:

1. The bidirectional steering scheme outperforms the unidirectional one. Indeed, when the forward and backward trajectories are optimally combined (i.e., by using the Minh-Adib PMF estimator or our proposed reweighting technique), the errors due to overestimation of PMF are averaged out.

2. At the end of 64 binding SMD simulations, 80% of the L22-TAR complex structures feature a major-groove binding mode. This confirms an important electrostatic feature of A-form RNA: the major groove is more electro-negative than the minor groove.

3. 52% of the complexes can be classified as upper-major-groove binding. In other words, more than a half of our binding SMD simulations are able to end up in the correct binding pocket without any guidance from the NMR information. This is an efficient self-guiding prediction protocol, which is extremely important for drug design when experimental structures are not available.

4. Among the upper-major-groove-binding structures, some complexes have a similar ligand orientation as in NMR structure, i.e., the $^{\mathrm{L}}$Pro-$^{\mathrm{D}}$Pro template of L22 points downward and inward TAR; some other complexes have the ligand in the same pocket but with opposite orientation, i.e., the $^{\mathrm{L}}$Pro-$^{\mathrm{D}}$Pro template points upward and outward TAR. A thorough PMF analysis and statistical accuracy test show that the latter pose consistently exhibits higher stability than the former pose. Since L22 is a small (14 amino acids) and rather symmetric cyclic peptide with 6 arginine residues almost equally distributed, we are convinced that this new pose is also likely to occur. In any case, both poses feature the correct binding pocket.

5. The free-energy difference computed from our bidirectional SMD simulations is, however, higher than that obtained by experiments (i.e., 85 ± 5 versus 52 kJ/mol respectively). Since we are confident in the statistical accuracy of the simulations, we would ascribe the discrepancy between the estimated PMF from simulations and experiments to other causes, probably including and not limited to *(i)* the difference in anion type, *(ii)* the insufficiently large simulation box, and *(iii)* the well-known force-field inaccuracies, especially the challenging description of strong electrostatic interactions by the non-polarizable force fields [113, 3].

6.2 Perspectives

As a complementary approach to SMD simulations, we also performed **W***ell*-**T***empered* **MetaD***ynamics* (**WTMetaD**) simulation [13], which can be applied on multiple CVs simultaneously. The simulation starts from the NMR binding pose. Besides the DHEN CV, the number of intermolecular hydrogen bonds is used as the second CV. Within the time course of 500 ns, we observe two times of complete complex dissociation. After the second dissociation, the ligand associates to TAR in the same pocket but with the opposite orientation, i.e., the dominant orientation found in our binding SMD simulations (see Figure 6.2.1). Since this WTMetaD simulation is at an early stage, in which the free energy profile has not converged and the time that the complex spends in the new binding pose has not exceeded the time it stays in the NMR pose, we cannot make judgment on the relative stability between these two poses at this stage. However, this preliminary result supports the previous SMD simulations on the observation of the new binding pose that has not yet been reported by experiments.

In theory, if the time of the WTMetaD simulation is large enough, the fluctuation of free-energy difference between any two metastable states is progressively damped to the correct value[2] [13]. Figure 6.2.2 shows the time evolution of the free-energy difference between a representative unbound state (e.g., chosen at DHEN$= -40$ kJ/mol and $d = 35$ Å correspondingly) and the bound state (i.e., corresponding to the minimum of the free energy surface in the last time frame) as a function of DHEN and distance CVs[3]. In both cases, the free-energy difference has not converged to a constant but rather fluctuates around the value of 80 kJ/mol, which is at the same order of magnitude as the free-energy difference estimated by our bidirectional SMD simulations. This result again confirms that the discrepancy in free-energy difference between our SMD simulations and experiments is due to neither the statistical insufficiency nor the simulation techniques. Therefore it is more likely due to other sources of errors as previously discussed in Sections 5.4.5 and 6.1.

[2] This is how WTMetaD is more advantageous than standard metadynamics in controlling convergence.

[3] Note that the free energy surface as a function of the distance CV is reconstructed by the reweighting technique for WTMetaD introduced in Ref. [211].

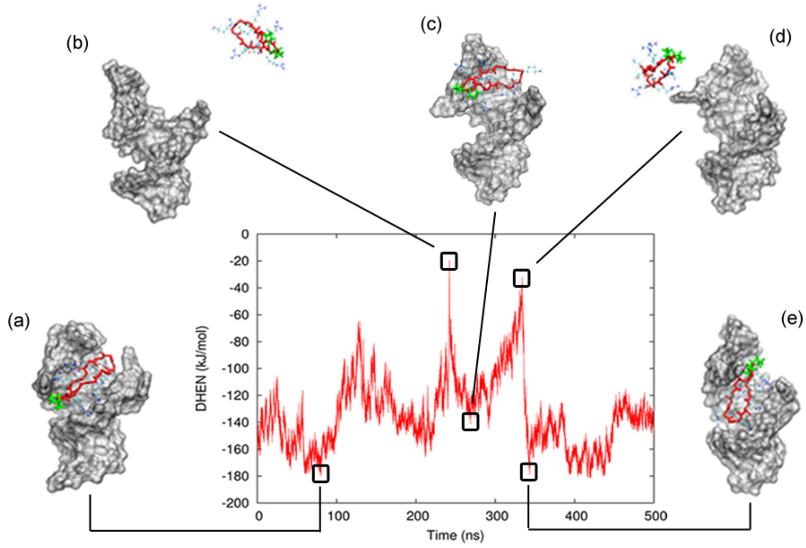

Figure 6.2.1: Time dependence of the DHEN CV during the 500-ns WTMetaD simulation. The simulation starts from the NMR binding pose (a), completely dissociates for the first time at 250 ns (b), associates back to a pose very similar to the NMR one (c), dissociates again at 340 ns (d), and finally associates at 345 ns to a new pose (e), which has the same binding pocket but opposite ligand's orientation.

In this particular case of L22-TAR binding, due to the special geometrical property of L22, our proposed DHEN CV (so as the distance and number-of-hydrogen-bond CVs[4]) is not able to distinguish the two binding poses. Indeed, these two poses belong two the same basin in the free energy profile projected on each CV (see Appendix D). However, our CV is able to guide the ligand several times to the right binding pocket, which is the same pocket as in Tat-TAR binding [85]. Therefore we are confident that our proposed DHEN CV is useful to study the molecular binding events involving electrostatic interactions.

[4] see Appendix D for the time evolution of distance and number of hydrogen bonds during the WTMetaD simulation.

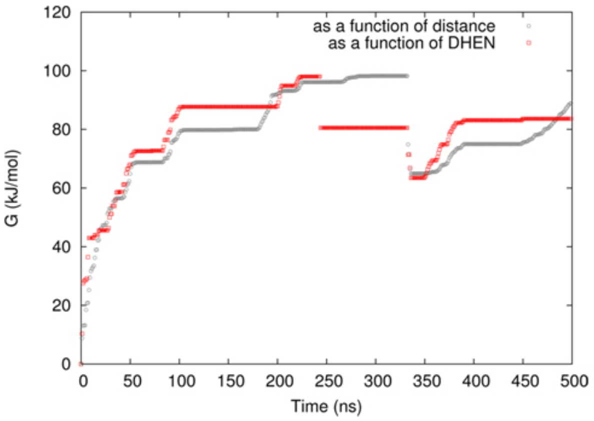

Figure 6.2.2: Time evolution of the free-energy difference between unbound and bound states projected on the DHEN CV (red squares) and distance CV (gray circles).

Appendix A

Maximum Likelihood

The method

The maximum likelihood estimation method was introduced by the English statistician and population geneticist R. A. Fisher in 1922 [163]. The making of maximum likelihood was one of the most important developments in 20th century statistics [212]. In general, the maximum likelihood method finds the estimate of a *model parameter* that *maximizes* the probability of observing the data given a *specific model* for the data.

Simple examples

Example 1

Consider tossing a coin, which has a number and a figure side. The result of every toss is registered as 1 for figure side and 0 for number side. Suppose that n tosses have been made and we obtain a series of data $x_1, x_2, \ldots, x_n$ in which x_i is either 1 or 0. This set of data defines a *specific model* of data, or a *specific distribution function*. The maximum likelihood method can help answering the question "*what is the probability of obtaining the figure side (or number side) in a single toss given the result of n tosses?*". For that purpose, maximum likelihood method first treats the probability of obtaining a specific side in a single toss as a *parameter*. Then the method involves finding the value of this parameter that maximizes the chance of observing the given data (i.e., the specific data model).

Example 2

One might be interested in the distribution of the weights of one-year-old children in a specific population but is only able to measure the weights of some children in that population. Supposed that the weights of all one-year-old children in the population are normally distributed with unknown mean and variance. The maximum likelihood method can help finding the mean and variance (treated as *model parameters*) of a distribution given only some sample of the overall population (i.e., the *given model*).

The principles

Suppose there is a sample of *independent and identically distributed* observations x_1, x_2, $\ldots$, x_n that has a probability density function ρ_0. The specific form of ρ_0 is unknown, however, we know that ρ_0 belongs to a certain family of distribution $\{\rho(\cdot|\theta),\ \theta \in \Theta\}$, in which $\theta \in \Theta$ denotes a certain parameter in the parametric space, so that $\rho_0 = \rho(\cdot|\theta_0)$. Here θ_0 is the value of the parameter that describes the probability density function of the given sample. The maximum likelihood method allows finding the parameter θ_0 given the sample of observations x_1, x_2, $\ldots$, x_n. [1]

For that purpose, we first write the density function for all observations as

$$\rho(x_1, x_2, \ldots, x_n|\theta) = \rho(x_1|\theta)\rho(x_2|\theta)\ldots\rho(x_n|\theta). \tag{A.0.1}$$

This joint density function is valid for independent and identically distributed observations. Notationally, this function represents the probability of observing the sample x_1, x_2, $\ldots$, x_n out of a given distribution characterized by the parameter θ. Now we can look at this same function from a different perspective by considering the sample x_1, x_2, $\ldots$, x_n as fixed parameters of this function and θ is the function's variable to be found so that the distribution described by θ is the same distribution determined by the sample of observations. We can then define a function to be called the *likelihood* as followed

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \rho(x_1, x_2, \ldots, x_n|\theta) = \rho(x_1|\theta)\rho(x_2|\theta)\ldots\rho(x_n|\theta). \tag{A.0.2}$$

The problem now becomes finding a value of θ that maximizes $\mathcal{L}(\theta|x_1, x_2, \ldots, x_n)$. An equivalent is finding θ that maximizes *log-likelihood* $\ln \mathcal{L}(\theta|x_1, x_2, \ldots, x_n)$ which is more convenient to solve due to the product of density functions on the right hand side. We can write the *log-likelihood* as

[1] Note that the observations x_i and parameter θ can be vectors.

$$\ln \mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln \rho(x_i|\theta). \tag{A.0.3}$$

The value of θ that maximizes the log-likelihood is the solution of the following equation

$$\frac{\partial \ln \mathcal{L}(\theta|x_1, x_2, \ldots, x_n)}{\partial \theta} = 0. \tag{A.0.4}$$

Appendix B

Additional Information on the Encounter MD Simulations

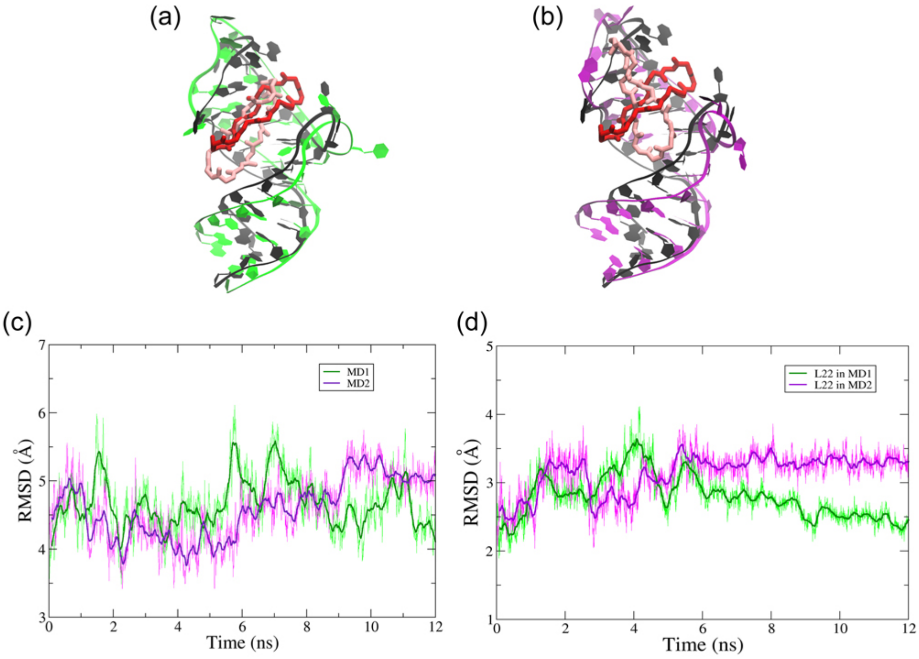

Figure B.0.1: (a) and (b): Superpositions of NMR structure and the last snapshot of the **MD1** and **MD2** trajectories respectively. The NMR TAR is shown in black, **MD1** TAR is in green, and **MD2** TAR is in purple. The NMR L22 is shown in red and both **MD1** and **MD2** L22 are shown in pink. In **MD1**, L22 is located along the upper major groove, providing a binding mode similar to the NMR structure. In **MD2**, L22 is orthogonal to the groove. (c) RMSDs and running averages of TAR (with respect to NMR structures) in **MD1** (green) and **MD2** (purple). (d) RMSDs and running averages of L22 (with respect to NMR structures) in **MD1** (green) and **MD2** (purple).

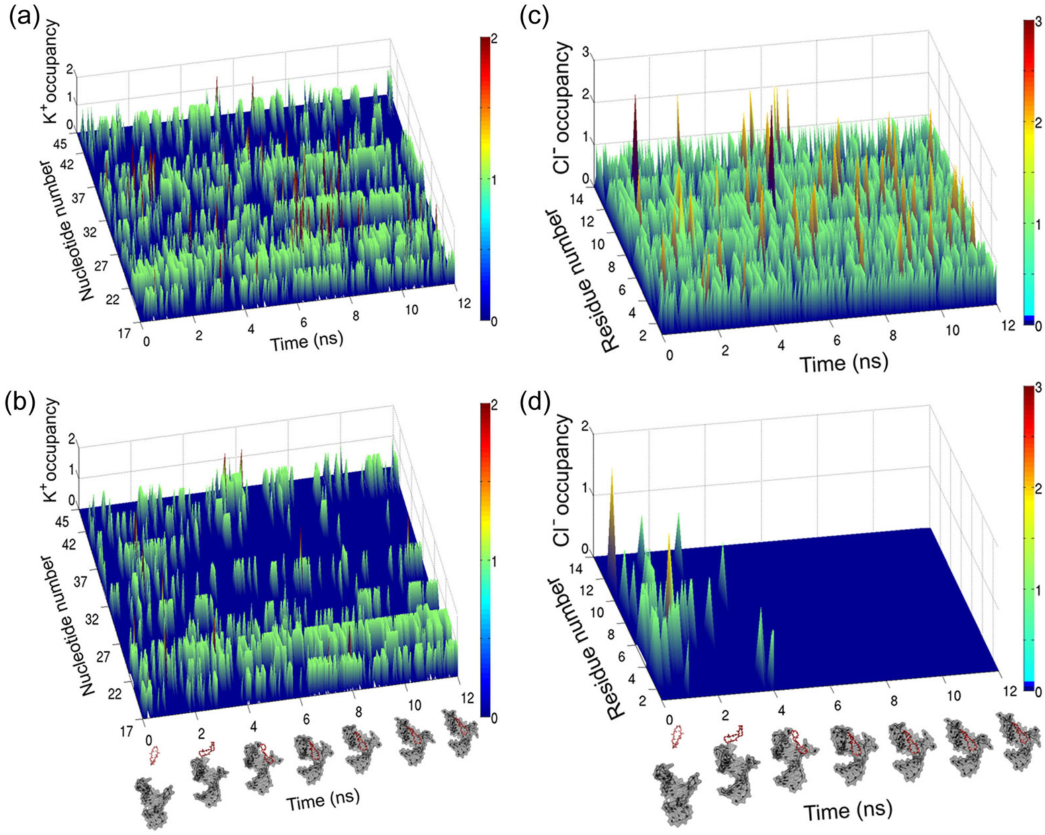

Figure B.0.2: Ion occupancy as a function of simulation time calculated by proximity method [188, 189, 190]: K^+ occupancy along the RNA nucleotides in the simulations of apo-TAR (a) and **MD2** (b); Cl^- occupancy along the L22 residues in the simulations of apo-L22 (c) and **MD2** (d).

Appendix C

Assessment of the Chosen RNA Force Field for MD simulations

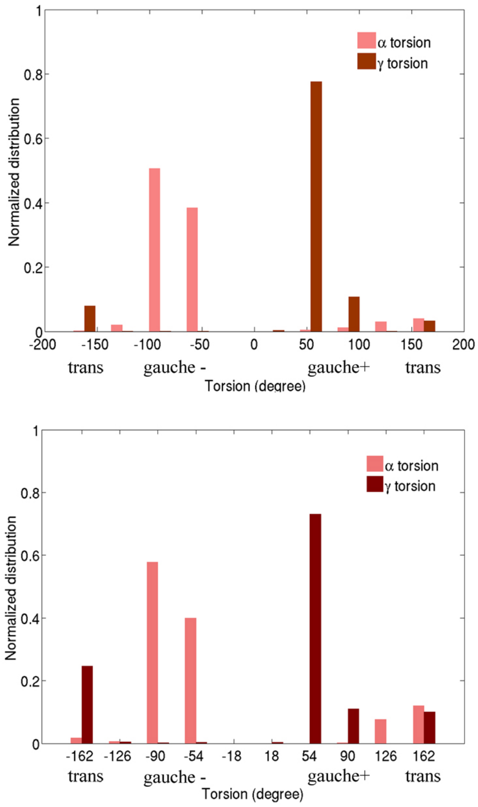

Figure C.0.1: α/γ torsion distribution of lower stem (upper panel) and upper stem (lower panel) in TAR. The distribution of these angles in the energetically unfavorable *trans* conformations is small.

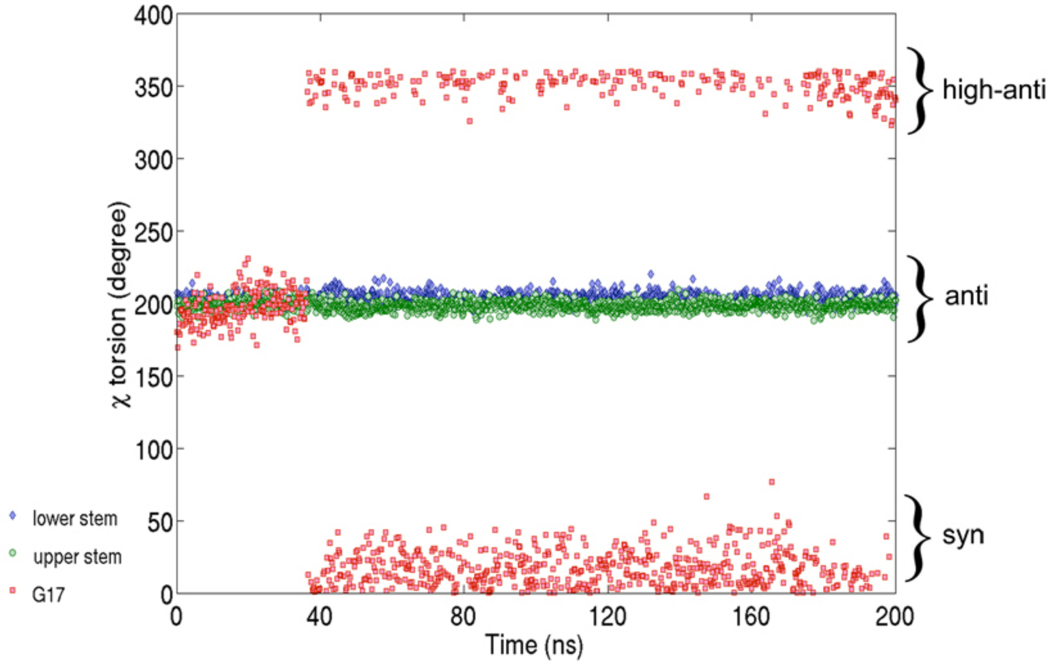

Figure C.0.2: χ torsion angles (degrees) for the helical regions of TAR plotted as a function of simulation time (ns). Average χ angles of the lower stem (blue) and upper stem (green) are distributed mostly in the *anti* region (i.e., from 180 to 250 degrees). Only in one case is a nucleotide (G17) occasionally occupying the *high-anti* conformation (i.e., form 320 - 360 degrees). However, G17 belongs to the lowest base pair of the lower stem. It is far (i.e., about 20 Å) from the L22 binding site and it is very likely to play a negligible role for L22-TAR interactions.

Appendix D

Preliminary Results of WTMetaD Simulation

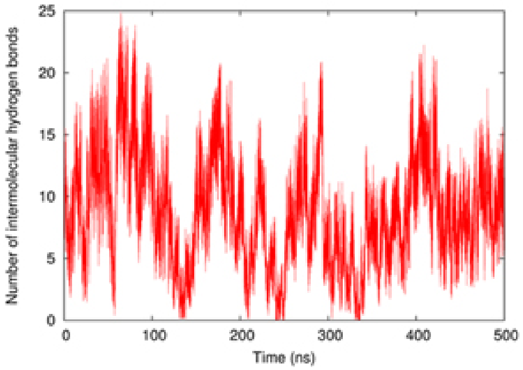

Figure D.0.1: Time dependence of the number of hydrogen bonds.

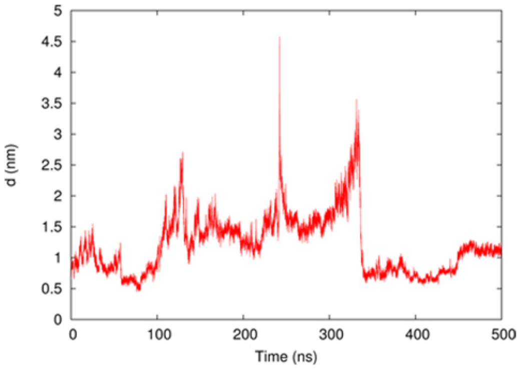

Figure D.0.2: Time dependence of the center-to-center distance.

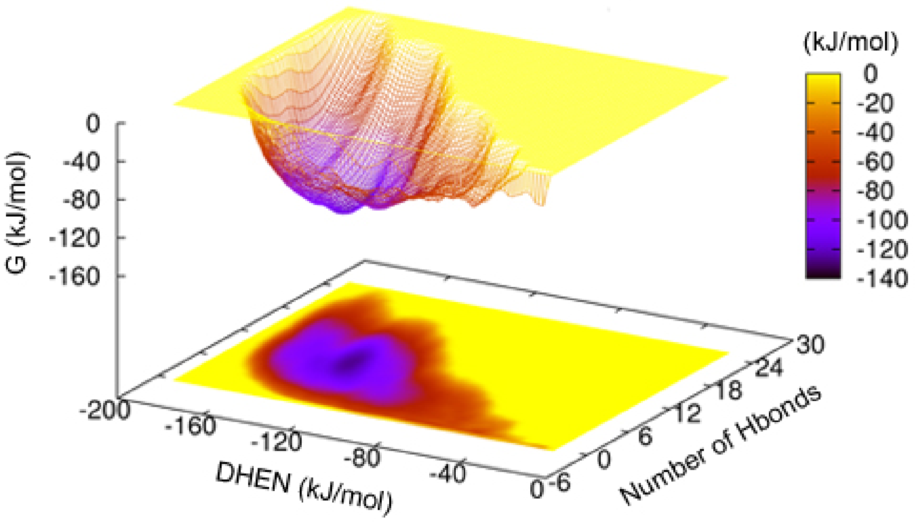

Figure D.0.3: Free energy surface as a function of the DHEN and number-of-hydrogen-bond CVs estimated from a 500-ns well-tempered metadynamics simulation.

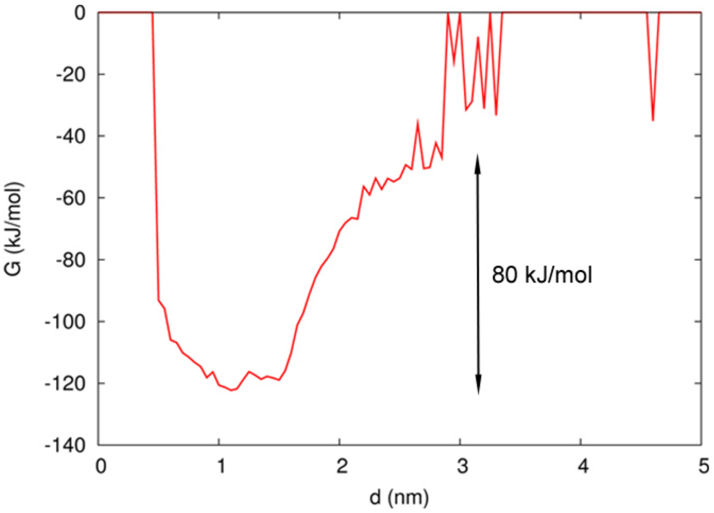

Figure D.0.4: Free energy as a function of the distance CV estimated from a 500-ns well-tempered metadynamics simulation by a reweighting technique.

Bibliography

[1] T. Hermann and E. Westhof. RNA as a drug target: Chemical, modelling, and evolutionary tools. *Curr. Opin. Biotechnol.*, 9:66–73, 1998. iii, 93

[2] J. Gallego and G. Varani. Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. *Acc. Chem. Res.*, 34:836–843, 2001. iii, 93

[3] S. Fulle and H. Gohlke. Molecular recognition of RNA: Challenges for modelling interactions and plasticity. *J. Mol. Recognit.*, 23:220–231, 2010. iii, 11, 12, 22, 91, 93, 96

[4] P. Auffinger and E. Westhof. Water and ion binding around RNA and DNA (C,G) oligomers. *J. Mol. Biol.*, 300:1133–1131, 2000. iii, 11

[5] P. Auffinger and Y. Hashem. Nucleic acid solvation: from outside to insight. *Curr. Opin. Struc. Biol.*, 17:325–333, 2007. iii, 11

[6] A. Perez, I. Marchan, D. Svozil, J. Sponer, T.E. Cheatham, C.A. Laughton, and M. Orozco. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.*, 92:3817–3829, 2007. iii, 21, 60, 74, 80

[7] P. Banas, D. Hollas, M. Zgarbova, P. Jurecka, M. Orozco, T.E.III Cheatham, J. Sponer, and M. Otyepka. Performance of molecular mechanics force fields for RNA simulations: Stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.*, 6:3836–3849, 2010. iii, 21, 22, 91

[8] E.J. Denning, U.D. Priyakumar, L. Nilsson, and A.D.Jr. Mackerell. 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.*, 32:1929–1943, 2011. iii

[9] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T.E. III Cheatham, and P. Jurecka. Refinement of the Cornell et al. nucleic acid force field based on reference quantum chemical calculations of torsion profiles of the glycosidic torsion. *J. Chem. Theory Comput.*, 7:2886–2902, 2011. iii

[10] H. Grubmüller, B. Heymann, and P. Tavan. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science*, 271:997–999, 1996. iii, 1, 12, 16, 28, 29

[11] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, 99:12562–12566, 2002. iii, 12, 16

[12] M. Sotomayor and K. Schulten. Single-molecule experiments in vitro and in silico. *Science*, 316:1144–1148, 2007. iii, 1, 13, 16

[13] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100:020603–020606, 2008. iii, 13, 96

[14] F.L. Gervasio, A. Laio, and M. Parrinello. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.*, 127:2600–2607, 2005. 1

[15] F. Colizzi, R. Perozzo, L. Scapozza, M. Recanatini, and A. Cavalli. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *J. Am. Chem. Soc.*, 132:7361–7371, 2010. 1

[16] C. Bartels and M. Karplus. Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy. *J. Phys. Chem. B*, 102:865–880, 1998. 2, 94

[17] F. Wang and D.P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, 2001. 2, 94

[18] C. Micheletti, A. Laio, and M. Parrinello. Reconstructing the density of states by history-dependent metadynamics. *Phys. Rev. Lett.*, 92:170601–170604, 2004. 2, 94

[19] C. Michel, A. Laio, and A. Milet. Tracing the entropy along a reactive pathway: The energy as a generalized reaction coordinate. *J. Chem. Theory Comput.*, 5:2193–2196, 2009. 2, 94

[20] M. Bonomi and M. Parrinello. Enhanced sampling in the well-tempered ensemble. *Phys. Rev. Lett.*, 104:190601–190604, 2010. 2, 94

[21] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999. 2

[22] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.*, 19:451–458, 1992. 2

[23] B.A. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, 68:9–12, 1992. 2

[24] A. Davidson, T.C. Leeper, Z. Athanassiou, K. Patora-Komisarska, J. Karn, J.A. Robinson, and G. Varani. Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc. Natl. Acad. Sci. U.S.A.*, 106: 11931–11936, 2009. 2, 8, 9, 10, 58, 61, 68, 84, 85, 91, 93

[25] J.M. Coffin. *The Retroviridae (Levy J.A.)*, chapter Structure and Classification of Retroviruses. New York: Plenum Press, 1 edition, 1992. 3, 4

[26] R.A. Weiss. How does HIV cause AIDS? *Science*, 260:1273–1279, 1993. 3

[27] D.C. Douek, M. Roederer, and R.A. Koup. Emerging concepts in the immunopathogenesis of AIDS. *Annu. Rev. Med.*, 60:471–484, 2009. 3

[28] F. Barré-Sinoussi, J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for Acquired Immune Deficiency Syndrome (AIDS). *Science*, 220:868–871, 1983. 3

[29] Joint United Nations Programme on HIV/AIDS (UNAIDS). Together we will end aids. 2012. 3

[30] J.D. Reeves and R.W Doms. Human Immunodeficiency Virus type 2. *J. Gen. Virol.*, 83:1253–1265, 2002. 3

[31] P.B. Gilbert, I.W. McKeague, G. Eisen, C. Mullins, A. Guéye-NDiaye, S. Mboup, and P.J. Kanki. Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Stat. Med.*, 22:573–593, 2003. 3

[32] E. De Clercq. *HIV-1 Integrase: Mechanism and Inhibitor Design (N. Neamati)*, chapter HIV life cycle: targets for anti-HIV agents, pages 1–14. John Wiley & Sons, Hoboken, N.J., U.S.A., 2011. 3

[33] R. Wyatt and J. Sodroski. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, 280:1884–1888, 1998. 3

[34] D. Chan and P. Kim. HIV entry and its inhibition. *Cell*, 93:681–684, 1998. 3

[35] Y.H. Zheng, N. Lovsin, and B.M. Peterlin. Newly identified host factors modulate HIV replication. *Immunol. Lett.*, 97:225–234, 2005. 4

[36] Y. Pommier, A.A. Johnson, and C. Marchand. Integrase inhibitors to treat HIV/AIDS. *Nature Rev. Drug Discov.*, 4:236–248, 2005. 4

[37] M. Stevens, E. De Clercq, and J. Balzarini. The regulation of HIV-1 transcription: Molecular targets for chemotherapeutic intervention. *Med. Res. Rev.*, 26:595–625, 2006. 4

[38] N.E. Kohl, E.A. Emini, W.A. Schleif, L.J. Davis, J.C. Heimbach, R.A. Dixon, E.M. Scolnick, and I.S. Sigal. Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl. Acad. Sci. U.S.A.*, 85:4686–4690, 1988. 4

[39] D.R. Davies. The structure and function of the aspartic proteinases. *Annu. Rev. Biophys. Biophys. Chem.*, 19:189–215, 1990. 4

[40] H.R. Gelderblom. Assembly and morphology of HIV: potential effect of structure on viral function. *AIDS*, 5:617–637, 1991. 4

[41] E. De Clercq. Strategies in the design of antiviral drugs. *Nature Rev. Drug Discov.*, 1: 13–25, 2002. 4

[42] L.M. Stolk and J.F. Luers. Increasing number of anti-HIV drugs but no definite cure: review of anti-HIV drugs. *Pharm. World Sci.*, 26:133–136, 2004. 4

[43] E. Poveda, V. Briz, and V. Soriano. Enfuvirtide, the first fusion inhibitor to treat HIV infection. *AIDS Rev.*, 7:139–147, 2005. 5

[44] M.A. Fischl, D.D. Richman, M.H. Grieco, M.S. Gottlieb, P.A. Volberding, O.L. Laskin, J.M. Leedom, J.E. Groopman, D. Mildvan, R.T. Schooley, G.G. Jackson, D.T. Durack, and D. King. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial. *N. Engl. J. Med.*, 317:185–191, 1987. 5

[45] Benfield P. Patel, S.S. New drug profile: nevirapine. *Clinical Immunotherapeutics*, 6: 307–317, 1996. 5

[46] R.T. Steigbigel, D.A. Cooper, and P.N. Kumar. Raltegravir with optimized background therapy for resistant HIV-1 infection. *N. Engl. J. Med.*, 359:339–354, 2008. 5

[47] C.J. la Porte. Saquinavir, the pioneer antiretroviral protease inhibitor. *Expert. Opin. Drug Metab. Toxicol.*, 5:1313–1322, 2009. 5

[48] S. Freeman and J.C. Herron. *Evolutionary Analysis - A case for evolutionary thinking: understanding HIV*. Pearson Benjamin Cummings, San Francisco, CA, 4 edition, 2007. 5

[49] B.G. Brenner, D. Turner, and M.A. Wainberg. HIV-1 drug resistance: can we overcome? *Expert Opin. Biol. Ther.*, 2:751–761, 2002. 5

[50] M.J. Kozal. Drug-resistant human immunodeficiency virus. *Clin. Microbial. Infec.*, 15: 69–73, 2009. 5

[51] P.A. Cane. New developments in HIV drug resistance. *J. Antimicrob. Chemoth.*, 64: 37–40, 2009. 6

[52] J.A. Turpin. The next generation of HIV/AIDS drugs: novel and developmental antiHIV drugs and targets. *Expert Rev. Anti. Infect. Ther.*, 1:97–128, 2003. 6

[53] M. Emerman and M.H. Malim. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science*, 280:1880–1884, 1998. 6

[54] S. Hwang, N. Tamilarasu, K. Ryan, I. Huq, S. Ritchter, W.C. Still, and T.M. Rana. Inhibition of gene expression in human cells through small molecule-RNA interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 96:12997–13002, 1999. 6, 8

[55] K.F. Blount and Y. Tor. Using pyrene-labeled HIV-1 TAR to measure RNA-small molecule binding. *Nucleic Acids Res.*, 31:5490–5500, 2003. 6, 8

[56] T.D. Bradick and J.P. Marino. Ligand-induced changes in 2-aminopurine flourescence as a probe for small molecule binding to HIV-1 TAR RNA. *RNA*, 10:1459–1468, 2004. 6, 8

[57] M. Baba. Recent status of HIV-1 gene expression inhibitors. *Antiviral Res.*, 71:301–306, 2006. 6, 8

[58] A.M. Mhashilkar, D.K. Biswas, J. LaVecchio, A.B. Pardee, and W.A. Marasco. Inhibition of human immunodeficiency virus type 1 replication in vitro by a novel combination of anti-Tat single-chain intrabodies and NF-kB antagonists. *J. Virol.*, 71: 6486–6494, 1997. 6

[59] L.M. Bedoya, M. Beltrán, R. Sancho, D.A. Olmedo, S. Sánchez-Palomino, E. del Olmo, J.L. López-Pérez, E. Muñoz, A. San Feliciano, and J. Alcami. 4-Phenylcoumarins as HIV transcription inhibitors. *Bioorg. Med. Chem. Lett.*, 15:4447–4450, 2005. 6

[60] J.A. Cook, A. August, and A.J. Henderson. Recruitment of phosphatidylinositol 3-kinase to CD28 inhibits HIV transcription by a Tat-dependent mechanism. *J. Immunol.*, 169:254–260, 2002. 6

[61] D. Daelemans, E.D. Clercq, and A.M. Vandamme. Control of RNA initiation and elongation at the HIV promoter. *AIDS Rev.*, 2:229–240, 2000. 6

[62] J. Sodroski, R. Patarca, C. Rosen, F. Wong-Staal, and W. Haseltine. Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III. *Science*, 229:74–77, 1985. 6, 7

[63] A. Mujeeb, K. Bishop, B.M. Peterlin, C. Turck, T.G. Parslow, and T.L. James. NMR structure of a biologically active peptide containing the RNA-binding domain of human immunodeficiency virus type 1 Tat. *Proc. Natl. Acad. Sci. U.S.A.*, 91:8248–52, 1994. 6

[64] S.Y. Kao, A.F. Calman, P.A. Luciw, and B.M. Peterlin. Anti-termination of transcription within the long terminal repeat of HIV-1 by Tat gene product. *Nature*, 330: 489–493, 1987. 6

[65] M.F. Laspia, A.P. Rice, and M.B. Mathews. HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation. *Cell*, 59:283–292, 1989. 6

[66] M. Pangat, T. Meier, R. Keene, and R. Landick. Transcriptional pausing at +62 of the HIV-1 nascent RNA modulates formation of the TAR RNA structure. *Mol. Cell*, 1: 1033–1042, 1998. 6

[67] C. Liang and M.A. Wainberg. The role of Tat in HIV-1 replication: an activator and/or a suppressor? *AIDS Rev.*, 4:41–49, 2002. 6

[68] T.M. Rana and K.T. Jeang. Biochemical and functional interactions between HIV-1 Tat protein and TAR RNA. *Arch. Biochem. Biophys.*, 365:175–185, 1999. 6

[69] M.A. Muesing, D.H. Smith, and D.J. Capon. Regulation of mRNA accumulation by a human immunodeficiency virus trans-activator protein. *Cell*, 48:691–701, 1987. 6, 7

[70] J. Karn. Tackling tat. *J. Mol. Biol.*, 293:235–254, 1999. 7

[71] U. Delling, L.S. Reid, R.W. Barnett, M.Y. Ma, S. Climie, M. Sumner-Smith, and N. Sonenberg. Conserved nucleotides in the TAR RNA stem of human immunodeficiency virus type 1 are critical for Tat binding and trans activationmodel for TAR RNA tertiary structure. *J. Virol.*, 66:3018–3025, 1992. 7

[72] H. Huthoff and B. Berkhout. Mutations in the TAR hairpin affect the equilibrium between alternative conformations of the HIV-1 leader RNA. *Nucleic Acids Res.*, 29: 2594–2600, 2001. 7

[73] C. Dingwall, I. Ernberg, M.J. Gait, S.M. Green, S. Heaphy, J. Karn, A.D. Lowe, M. Singh, M.A. Skinner, and R. Valerio. Human immunodeilciency virus I Tat protein binds trans-activation-responsive region (TAR) RNA in vitro. *Proc. Natl Acad. Sci. U.S.A.*, 86:6925–6929, 1989. 8, 68, 71, 84, 85, 90

[74] K.M. Weeks, C. Ampe, S.C. Schultz, T.A. Steitz, and D.M. Crothers. Fragments of the HIV-1 Tat protein specifically bind TAR RNA. *Science.*, 249:1281–1285, 1990. 8, 68, 71

[75] B.J. Calnan, B. Tidor, S. Biancalana, D. Hudson, and A.D. Frankel. Arginine-mediated RNA recognition: the arginine fork. *Science.*, 252:1167–1171, 1991. 8, 66, 68, 71

[76] F. Hamy, E.R. Felder, G. Heizmann, J. Lazdins, F. Aboul-Ela, G. Varani, J. Karn, and T. Klimkait. An inhibitor of the Tat/TAR RNA interaction that effectively suppresses HIV-1 replication. *Proc. Natl. Acad. Sci. U.S.A.*, 94:3548–3553, 1997. 8

[77] H.Y. Mei, M. Cui, A. Heldsinger, S.M. Lemrow, J.A. Loo, K.A. Sannes-Lowery, L. Sharmeen, and A.W. Czarnik. Inhibitors of protein-RNA complexation that target the RNA: Specific recognition of Human Immunodeficiency Virus Type 1 TAR RNA by small organic molecules. *Biochemistry*, 37:14204–14212, 1998. 8

[78] F. Hamy, V. Brondani, A. Florsheimer, W. Stark, M.J.J. Blommers, and T.A. Klimkait. A new class of HIV-1 Tat antagonist acting through Tat-TAR inhibition. *Biochemistry*, 37:5086–5095, 1998. 8

[79] K.E. Lind, Z. Du, K. Fujinaga, and T.L. Peterlin, B.M. amd James. Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chem. Biol.*, 9:185–193, 2002. 8, 10

[80] Z. Du, K.E. Lind, and T.L. James. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.*, 7:707–712, 2002. 8

[81] B. Davis, M. Afshar, G. Varani, A.I. Murchie, J. Karn, G. Lentzen, M. Drysdale, J. Bower, A.J. Potter, I.D. Starkey, T. Swarbrick, and F. Aboul-ela. Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic 'hot spot'. *J. Mol. Biol.*, 336:343–356, 2004. 8

[82] A.I.H. Murchie, B. Davis, C. Isel, M. Afshar, M.J. Drysdale, J. Bower, A.J. Potter, I.D. Starkey, T.M. Swarbrick, S. Mirza, C.D. Prescott, P. Vaglio, F. Aboul-ela, and J. Karn. Structure-based drug design targeting an inactive RNA conformation: Exploiting the flexibility of HIV-1 TAR RNA. *J. Mol. Biol.*, 336:625–638, 2004. 8

[83] M.F.Jr. Bardaro, Z. Shajani, K. Patora-Komisarska, J.A. Robinson, and G. Varani. How binding of a small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape. *Nucl. Acids Res.*, 37:1529–1540, 2009. 9, 10, 59, 80, 93

[84] G. Aboul-ela, J. Karn, and G. Varani. Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic Acids Res.*, 24:3974–3981, 1996. 10

[85] C. Dingwall, I. Ernberg, M.J. Gait, S.M. Green, S. Heaphy, J. Karn, A.D. Lowe, M. Singh, and M.A. Skinner. HIV-1 Tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO J.*, 9:4145–4153, 1990. 10, 93, 97

[86] J. Zhang, N. Tamilarasu, S. Hwang, M.E. Garber, I. Huq, K.A. Jones, and T.M. Rana. HIV-1 TAR RNA enhances the interaction between Tat and Cyclin T1. *J. Biol. Chem.*, 275:34314–34319, 2000. 10, 93

[87] M.S. Ladonde, M.A. Lobritz, A. Ratcliff, M. Chamanian, Z. Athanassiou, M. Tyagi, J. Wong, J.A. Robinson, J. Karn, G. Varani, and E.J. Arts. Inhibition of both HIV-1 reverse transcription and gene expression by a cyclic peptide that binds the Tat-Transactivating Response Element (TAR) RNA. *PLoS Pathog.*, 7:1–17, 2011. 10

[88] H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.*, 41:2645–2676, 2002. 10

[89] S.F. Sousa, P.A. Fernandes, and M.J. Ramos. Protein-ligand docking: current status and future challenges. *Proteins*, 65:15–26, 2006. 10

[90] F. Leclerc and M. Karplus. MCSS-based predictions of RNA binding sites. *Theor. Chem. Acc.*, 101:131–137, 1999. 10

[91] X.S. Kang, R.H. Shafer, and I.D. Kuntz. Calculation of ligand-nucleic acid binding free energies with the generalized-born model in DOCK. *Biopolymers*, 73:192–204, 2004. 10

[92] C. Detering and G. Varani. Validation of automated docking programs for docking and database screening against RNA drug targets. *J. Med. Chem.*, 47:4188–4201, 2004. 10

[93] S.D. Morley and M. Afshar. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock. *J. Comput. Aid. Mol. Des.*, 18:189–208, 2004. 10

[94] F. Barbault, L.R. Zhang, L.H. Zhang, and B.T. Fan. Parametrization of a specific free energy function for automated docking against RNA targets using neural networks. *Chemom. Intell. Lab. Syst.*, 82:269–275, 2006. 10

[95] P. Pfeffer and H. Gohlke. Drugscore$^{\text{RNA}}$ - knowledge-based scoring function to predict RNA-ligand interactions. *J. Chem. Inf. Mod.*, 47:1868–1876, 2007. 10

[96] X.Y. Zhao, X.F. Liu, Y.Y. Wang, Z. Chen, L. Kang, H.L. Zhang, X.M. Luo, W.L. Zhu, K.X. Chen, H.L. Li, X.C. Wang, and H.L. Jiang. An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA. *J. Chem. Inf. Mod.*, 48:1438–1447, 2008. 10

[97] C. Guilbert and T. James. Docking to RNA via root-mean-squaredeviation-driven energy minimization with flexible ligands and flexible targets. *J. Chem. Inf. Mod.*, 48: 1257–1268, 2008. 10

[98] P.T. Lang, S.R. Brozell, S. Mukherjee, E.F. Pettersen, E.C. Meng, V. Thomas, R.C. Rizzo, D.A. Case, T.L. James, and I.D. Kuntz. DOC 6: combining techniques to model RNA-small molecul complexes. *RNA*, 15:1219–1230, 2009. 10

[99] N. Leulliot and G. Varani. Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, 40:7947–7956, 2001. 10

[100] H.M. Al-Hashimi. Dynamics-based amplification of RNA function and its characterization by using NMR spectroscopy . *Chem. Bio. Chem.*, 6:1506–1519, 2005. 10

[101] A.D.Jr. Mackerell and L. Nilsson. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, 18:194–199, 2008. 10, 11

[102] J. Warwicker and H.C. Watson. Calculation of the electric-potential in the active-site cleft due to alpha-helix dipoles. *J. Mol. Biol.*, 157:671–679, 1982. 11

[103] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990. 11

[104] N.K. Banavali and B. Roux. Atomic radii for continuum electrostatics calculations on nucleic acids. *J. Phys. Chem. B*, 106:11026–11035, 2002. 11

[105] R.C. Rizzo, T. Aynechi, D.A. Case, and I.D. Kuntz. Estimation of absolute free energies of hydration using continuum methods: accuracy of partial, charge models and optimization of nonpolar contributions. *J. Chem. Theor. Comput.*, 2:128–139, 2006. 11

[106] Foloppe. N., N. Matassova, and F. Aboul-Ela. Towards the discovery of drug-like RNA ligands? *Drug Disc. Today*, 11:1019–1027, 2006. 12, 93

[107] H. Gouda, I.D. Kuntz, D.A. Case, and P.A. Kollman. Free energy calculations for theophylline binding to an RNA aptamer: MM-PBSA and comparison of thermodynamic integration methods. *Biopolymers*, 68:16–34, 2003. 12

[108] D.E. Draper, D. Grilley, and A.M. Soto. Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, 34:221–243, 2005. 12

[109] A.A. Chen, M. Marucho, N.A. Baker, and R.V. Pappu. Simulations of RNA interactions with monovalent ions. *Methods Enzymol.*, 469:411–432, 2009. 12

[110] C.A. Reynolds, J.W. Essex, and W.G. Richards. Atomic charges for variable molecular conformations. *J Am Chem Soc*, 114:9075–9079, 1992. 12, 22, 91

[111] P. Cieplak, W.D. Cornell, C. Bayly, and P.A. Kollman. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J Comput. Chem.*, 16:1357–1377, 1995. 12, 22, 91

[112] A. Warshel, M. Kato, and A.V. Pisliakov. Polarizable force fields: history, test cases, and prospects. *J. Chem. Theor. Comput.*, 3:2034–2045, 2007. 12

[113] S.E. McDowell, N. Spackova, J. Sponer, and N.G. Walter. Molecular dynamics simulations of RNA: An *in silico* single molecule approach. *Biopolymers*, 85:169–184, 2007. 12, 21, 22, 74, 91, 93, 96

[114] M. Orozco, A. Noy, and A. Perez. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struc. Biol.*, 18:185–193, 2008. 12, 91

[115] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341–346, 2010. 12, 16

[116] K. Lindorff-Larsen, S. Piana, R.O. Dror, and D.E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011. 12, 16

[117] M. Mezei. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *J. Comput. Phys.*, 68:237–248, 1987. 12

[118] D.C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, 2nd edition, 2004. 15

[119] M.P. Allen and D.J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, 1987. 16, 24, 26

[120] J.M. Haile. *Molecular Dynamics Simulation: Elementary Methods*. John Wiley & Sons, New York, London, Sydney, 1992. 16

[121] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and Teller. E. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953. 16

[122] M. Born and J.R. Oppenheimer. Zur quantentheoride der molekeln. *Ann. Physik*, 84: 457–484, 1927. English version: "On the Quantum Theory of Molecules" translated by S. M. Blinder with emendations by B. Sutcliffe and W. Geppert (2002). 17

[123] M. Born and V.A. Fock. Beweis des Adiabatensatzes. *Zeitschrift für Physik*, A51: 165–180, 1928. 17

[124] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98–103, 1967. 17

[125] R.W. Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, 9:136–211, 1970. 18

[126] J.W. Gibbs. On the equilibrium of heterogeneous substances. In *Transactions of the Connecticut Academy*, volume III, pages 108–248. 1876. 18

[127] M.E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, NewYork, 2010. 18, 20

[128] L.D. Landau and E.M. Lifshitz. *Statistical Physics*, volume 5. Oxford: Pergamon Press, 3 edition, 1976. Translated by J.B. Sykes and M.J. Kearsley. 19

[129] H.C. Andersen. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2394, 1980. 20

[130] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–2690, 1984. 20

[131] S. Nosé. A unified formulation of the constant-temperature molecular dynamics methods. *J. Chem. Phys.*, 81:511–519, 1984. 20

[132] W.G. Hoover. Canonical dynamics-Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985. 20

[133] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:014101–014107, 2007. 20, 81

[134] M. Parrinello and A. Rahman. Crystal-structure and pair potentials - A molecular-dynamics study. *Phys. Rev. Lett.*, 45:1196–1199, 1980. 20

[135] A.R. Leach. *Molecular Modelling: Principles and Applications*. Prentice-Hall, 2nd edition, 2001. 21

[136] J.E. Sponer, N. Spackova, J. Leszczynski, and J. Sponer. Principles of RNA base pairing: structures and energies of the trans Watson-Crick/sugar edge base pairs. *J. Phys. Chem. B*, 109:11399–11410, 2005. 21

[137] J. Sponer, P. Jurecka, I. Marchan, J. Luque, M. Orozco, and P. Hobza. Nature of base stacking: Reference quantum-chemical stacking energies in ten unique B-DNA base-pair steps. *Chem. Eur. J.*, 12:2854–2865, 2006. 21

[138] I. Yildirim, H.A. Stern, S.D. Kennedy, J.D. Tubbs, and D.H. Turner. Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for Cytidine and Uridine. *J. Chem. Theory Comput.*, 6:1520–1531, 2010. 21

[139] T.N. Do, E. Ippoliti, P. Carloni, G. Varani, and M. Parrinello. Counterion redistribution upon binding of a Tat-protein mimic to HIV-1 TAR RNA. *J. Chem. Theory Comput.*, 8:688–694, 2012. 21, 93

[140] I. Besseova, M. Otyepka, K. Reblova, and J. Sponer. Dependence of A-RNA simulations on the choice of the force field and salt strength. *Phys. Chem. Chem. Phys.*, 11: 10701–10711, 2009. 22, 91

[141] J.A. Barker and R.O. Watts. Monte Carlo studies of the dielectric properties of water-like models. *Mol. Phys.*, 26:789–792, 1973. 23

[142] R.O. Watts. Monte Carlo studies of liquid water. *Mol. Phys.*, 28:1069–1083, 1974. 23

[143] P.P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Ann. Phys.*, 64:253–287, 1921. 24

[144] C. Sagui and T.A. Darden. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.*, 28:155–179, 1999. 24

[145] D. Frenkel and B. Smit. *Understanding Molecular Simulation: from Algorithms to Applications*. Academic Press, San Diego, 2nd edition, 2002. 24

[146] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10093, 1993. 24, 60, 80

[147] M. Frigo and S.G. Johnson. The design and implementation of FFTW3. In *Proceedings of the IEEE*, volume 93, pages 216–231, 2005. 24

[148] S.A. Showalter and R. Brüschweiler. Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: applications to the AMBER99SB forcefield. *J Chem. Theory Comput.*, 3:961–975, 2007. 24

[149] G. Lipari and A. Szabo. Model free approach to the interpretation of Nuclear Magnetic Resonance relaxation in macromolecules. 1. theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4559, 1982. 24, 25

[150] G. Lipari and A. Szabo. Model free approach to the interpretation of Nuclear Magnetic Resonance relaxation in macromolecules. 2. analysis of experimental results. *J. Am. Chem. Soc.*, 104:4559–4570, 1982. 24, 25

[151] R.R. Ernst, G. Bodenhausen, and A. Wokaun. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford University Press, NewYork., 2004. 25

[152] A. Abragam. *Principles of Nuclear Magnetism*. Clarendon Press, Oxford, U. K., 1961. 26

[153] E.L. Florin, V.T. Moy, and H.E. Gaub. Adhesion forces between individual ligand-receptor pairs. *Science*, 264:415–417, 1994. 28, 29, 30

[154] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J.*, 72:1568–1581, 1997. 30

[155] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690–2693, 1997. 32, 33

[156] G.E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721–2726, 1999. 32, 33

[157] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56:5018–5035, 1997. 33

[158] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U.S.A.*, 98:3658–3661, 2001. 33, 40

[159] C. Jarzynski. Nonequilibrium work theorem for a system strongly coupled to a thermal environment. *J. Stat. Mech.: Theor. Exp.*, page P09005, 2004. 33

[160] G.E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, 90:1481–1487, 1998. 33

[161] C.H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, 22:245–268, 1976. 37

[162] M.R. Shirts, E. Bair, G. Hooker, and V.S. Pande. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.*, 91:140601–140604, 2003. 38

[163] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, 222:309–368, 1922. 38, 99

[164] D.D.L. Minh and A.B. Adib. Optimized free energies from bidirectional single-molecule force spectroscopy. *Phys. Rev. Lett.*, 100:180602–180606, 2008. 42, 43

[165] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13:1011–1021, 1992. 43

[166] G.E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61:2361–2366, 2000. 43

[167] M. Bonomi, A. Barducci, and M. Parrinello. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J. Comput. Chem.*, 30:1615–1621, 2009. 45, 55

[168] F. Colizzi and G. Bussi. RNA unwinding from reweighted pulling simulations. *J. Am. Chem. Soc.*, 134:5173–5179, 2012. 45

[169] P. Debye and E. Hückel. Zur theorie der elektrolyte. I. Gefrierpunktserniedrigung und verwandte erscheinungen (English translation: The theory of electrolytes. I. Lowering of freezing point and related phenomena). *Physikalische Zeitschrift*, 24: 185–206, 1923. 48

[170] N.A. Baker, D. Sept, S. Joseph, M.J. Holst, and J.A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, 98:10037–10041, 2001. 51, 53, 67

[171] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface for both molecules and geometric objects: Applications to the finite difference Poisson-Boltzmann method. *J. Comp. Chem.*, 23:128–137, 2002. 51

[172] A. Onufriev, D. Bashford, and D. A. Case. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B*, 104:3712–3720, 2000. 52

[173] J. R. Thomas and P.J. Hergenrother. Targeting RNA with small molecules. *Chem. Rev.*, 108:1171–1224, 2008. 58, 74, 78, 94

[174] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79: 926–935, 1983. 60, 80

[175] Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24:1999–2012, 2003. 60

[176] M. Garcia-Diaz, K. Bebenek, A.A. Larrea, J.M. Havener, L. Perera, J.M. Krahn, L.C. Pedersen, D.A. Ramsden, and T.A. Kunkel. Scrunching during DNA repair synthesis. *Nat. Struct. Mol. Biol.*, 16:967–972, 2009. 60

[177] M. Mori, F. Manetti, and M. Botta. Predicting the binding mode of known NCp7 inhibitors to facilitate the design of novel modulators. *J. Chem. Inf. Model.*, 51:446–454, 2011. 60

[178] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26:1781–1802, 2005. 60

[179] U. Essman, L. Perela, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth Particle Mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995. 60, 80

[180] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23:327–341, 1977. 60

[181] P. Langevin. Une formule fondamentale de théorie cinétique. *Ann. Chim. Phys.*, 5: 245–288, 1905. 60

[182] P. Langevin. Sur la théorie du mouvement Brownien. *Comptes rendus hebdomadaires des séances de l'academie des sciences*, 146:530–533, 1908. 60

[183] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2:559–572, 1901. 62

[184] X. Ye, R.A. Kumar, and D.J. Patel. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.*, 2:827–840, 1995. 65

[185] J.D. Puglisi, L. Chen, S. Blanchard, and A.D. Frankel. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science*, 270:1200–1203, 1995. 65

[186] J.D. Puglisi, R. Tan, B.J. Calman, A.D. Frankel, and J.R. Williamson. Conformational of the TAR RNA-arginine complex by NMR spectroscopy. *Science.*, 257:76–80, 1992. 66

[187] L. Sethaphong, A. Singh, A.E. Marlowe, and Y.G. Yingling. The sequence of HIV-1 TAR RNA helix controls cationic distribution. *J. Phys. Chem. C.*, 114:5506–5512, 2010. 67

[188] P.K. Mehrotra and D.L. Beveridge. Structural analysis of molecular solutions based on quasi-component distribution functions. application to [H2CO]aq at 25.degree.C. *J. Am. Chem. Soc.*, 102:4287–4294, 1980. 69, 72, 104

[189] M. Mezei and D.L. Beveridge. Structural chemistry of biomolecular hydration via computer simulation: the proximity criterion. *Methods Enzymol.*, 127:21–47, 1986. 69, 72, 104

[190] S.Y. Ponomarev, K.M. Thayer, and D.L. Beveridge. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 101:14771–14775, 2004. 69, 72, 104

[191] P. Sklenovsky, P. Florova, P. Banas, K. Reblova, F. Lankas, M. Otyepka, and J. Sponer. Understanding RNA flexibility using explicit solvent simulations: The ribosomal and group I intron reverse kink-turn motifs. *J. Chem. Theory Comput.*, 7:2963–2980, 2011. 74

[192] J. Tao and A.D. Frankel. Electrostatic interactions modulate the RNA-binding and transactivation specificities of the Human Immunodeficiency Virus and Simian Immunodeficiency Virus Tat proteins. *Proc. Natl Acad. Sci. U.S.A.*, 90:1571–1575, 1993. 74, 78, 94

[193] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, and D.E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78:1950–1958, 2010. 80

[194] W.D. Cornell, P. Cieplak, C.T. Bayly, I.R. Gould, K.M.Jr. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995. 80

[195] T.E. III Cheatham, P. Cieplak, and P.A. Kollman. A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, 16:845–862, 1999. 80

[196] J. Wang, P. Cieplak, and P.A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21:1049–1074, 2000. 80

[197] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–725, 2006. 80

[198] J. Åqvist. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.*, 94:8021–8024, 1990. 80

[199] L.X. Dang. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: A molecular dynamics study. *J. Am. Chem. Soc.*, 117:6954–6960, 1995. 80

[200] A. Savelyev and G.A. Papoian. Inter-DNA electrostatics from explicit solvent molecular dynamics simulations. *J. Am. Chem. Soc.*, 129:6060–6061, 2007. 80

[201] P. Auffinger, T.E. III Cheatham, and A.C. Vaiana. Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue? *J. Chem. Theory Comput.*, 3:1851–1859, 2007. 80

[202] A.A. Chen and R.V. Pappu. Parameters of monovalent ions in the AMBER-99 forcefield: Assessment of inaccuracies and proposed improvement. *J. Phys. Chem. B*, 111: 11884–11887, 2007. 80

[203] I.S. Joung and T.E. Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112:9020–9041, 2008. 80

[204] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comp.*, 4:435–447, 2008. 80

[205] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981. 81

[206] A. Noy, I. Soteras, F.J. Luque, and M. Orozco. The impact of monovalent ion force field model in nucleic acids simulations. *Phys. Chem. Chem. Phys.*, 11:10596–10607, 2009. 91

[207] M. Froeyen and P. Herdewijn. RNA as a target for drug design, the example of Tat-TAR interaction. *Curr. Top. Med. Chem.*, 2:1123–1145, 2002. 93

[208] M. Zacharias. Perspectives of drug design that targets RNA. *Curr. Med. Chem.: Anti-Infect. Agents.*, 2:161–172, 2003. 93

[209] D.D. Boehr, R. Nussinov, and P.E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5:789–796, 2009. 93

[210] T.N. Do, P. Carloni, G. Varani, and G. Bussi. RNA/peptide binding driven by electrostatics - Insight from bidirectional pulling simulations. *J. Chem. Theory Comput.*, 9: 1720–1730, 2013. 95

[211] D. Branduardi, G. Bussi, and M. Parrinello. Metadynamics with adaptive gaussians. *J. Chem. Theory Comput.*, 8:2247–2254, 2012. 96

[212] J. Aldrich. R. A. Fisher and the making of maximum likelihood. *Statist. Sci.*, 12: 162–176, 1997. 99