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

Translocation of NS3 from Hepatitis C Virus on RNA: Insights from Atomistic Molecular Simulations

PhD in Physics and Chemistry of Biological Systems

Doctoral Dissertation of:
Andrea Pérez Villa

Supervisor:
Prof. Giovanni Bussi

Trieste, October 12 2015

Abstract

Helicases are motor proteins that unwind double stranded nucleic acids and are important parts of the genetic apparatus. A notable member of this family of enzymes is the nun-structural protein NS3 from Hepatitis C Virus. NS3 helicase unwinds nucleic acids by translocating along a single strand. Single molecule experiments and X-ray crystallography suggest that NS3 follows an *inchworm-like* motion during the translocation mechanism, consuming one ATP molecule per cycle. In spite of the available experimental data, the mechanistic and chemical details of the translocation process are still unclear.

The aim of this study is to model at atomistic detail the NS3h-RNA complex at the different stages of the translocation. For this purpose, atomistic molecular dynamics simulations were performed in explicit solvent in the presence and in the absence of ATP and ADP. Simulations were initialized based on existing crystallographic structures. All the stages of translocation were considered, and their relative stabilities were analyzed by computing electrostatic interactions, relative enthalpies, and hydrogen-bond patterns. Additionally, well-tempered metadynamics and Hamiltonian replica exchange simulations were performed to characterize the free-energy landscape associated to translocation and to describe the conformational transitions.

Contents

List of Figures

List of Tables

List of Acronyms

bp base pairs 6

COM centers of mass 32

CV collective variable 18

D1 domain 1 6

D2 domain 2 6

D3 domain 3 6

ds double strand 5

EEC Enthalpy-Entropy Compensation 49

FF Force Field 14

FRET Fluorescence Resonance Energy Transfer 6

G^{DH} Debye-Hückel energy 43

HREX Hamiltonian replica exchange 22

HCV Hepatitis C virus 1

MD Molecular Dynamics 8

NS3 non-structural protein 3 4

NS3h NS3 helicase 4

NA nucleic acid 5

nt nucleotides 61

ORF Open Reading Frame 2

Pi phosphate 7

PCA Principal Component Analysis 34

RMSD Root-mean-square deviation 16

ss single strand 5

SF2 Superfamily 2 4

TMD Targeted Molecular Dynamics 16

UTR Untranslated Regions 2

WTmetaD Well-tempered metadynamics 19

CHAPTER 1

Introduction

SINCE the discovery of Hepatitis C virus (HCV) in the late-1980's, many efforts have been done towards the prevention and treatment of a silent disease that at long-term infection becomes the main cause of functional and structural damages in the liver. Hepatitis C is transmitted to roughly 2 million people every year and ends as a chronic infection for nearly 80% of patients [1, 2]. The long-term infection is the reason of serious health problems such as liver cirrhosis and hepatocellular carcinoma. Up to now there is no vaccine that can prevent the transmission, and for this reason the study of several pathways related with the virus replication has been an issue of great interest for the development of an effective therapy. The studies have revealed information about the genome, the virus structure, and have also provided some insights about the encoded proteins which play a crucial role for the virus spreading in the host organism.

1.1 HCV genome

Hepatitis C is originated by a blood-borne virus that belongs to the *flaviviridae* family having as principal target the hepatocites, *i.e*. liver cells. Until now, nearly 60 species of flaviviridae viruses have been discovered, which are classified in 4 genera: Flavivirus, Pegivirus, Pestivirus and Hepacivirus [3]. Flavivirus genus comprises Yellow fever, Dengue fever and Tick-borne encephalitis viruses to name a few. Pestiviruses and Pegiviruses comprise Bovine viral diarrhea, Classical swine fever,

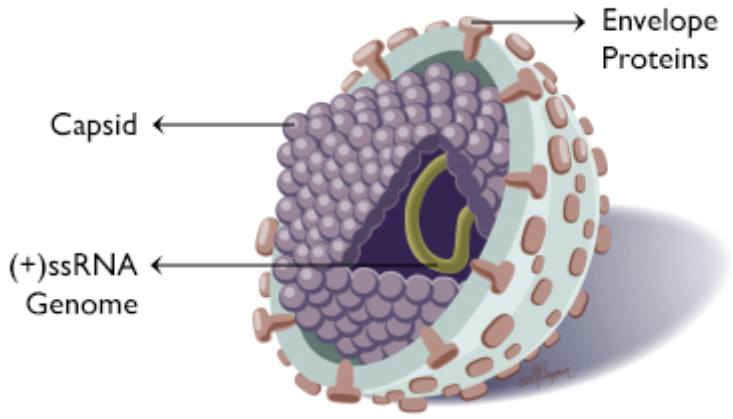


Figure 1.1: Cartoon representing a flaviviridae virus. The external structure formed by the lipid bilayer (light gray) and the envelope proteins (brown) contains the icosahedral capsid (purple). Inside the capsid there is the viral RNA genome (green). Picture adapted from Toronto Centre for Liver Disease [7].

GB viruses and so forth. Furthermore, Hepacivirus genus includes Tamarin virus and HCV, the latter being the most well-known member with roughly 6 genotypes identified and various subtypes. All viruses from the flaviviridae family have structural and biochemical common aspects [4]. Their external structure is a lipid bilayer that binds their envelope proteins. This structure protects the icosahedral capsid that encapsulates the positive single stranded RNA (Figure 1.1). The RNA genome in its Open Reading Frame (ORF), which is the sequence that translates for peptides, encodes for a long polyprotein that contains structural and non-structural proteins, fundamental for assembling and replication of the virus. In addition, sequence analysis between RNA protease-helicases and RNA polymerases from different flaviviridae viruses revealed a motif conservation in similar regions of the proteins [5, 6], suggesting common features in the mechanism of action performed by the same type of flaviviridae proteins.

In more quantitative terms, HCV is a small virus with a size of approximately 55 nm. Its entire genome has around ~9600 nucleotides (3011 codons), including the flanking 5′ and 3′ Untranslated Regions (UTR) and the long ORF which is processed co- and post-translationally by cellular and viral proteases to generate the specific viral species [6]. The non-translated regions also called 5′UTR and 3′UTR, are

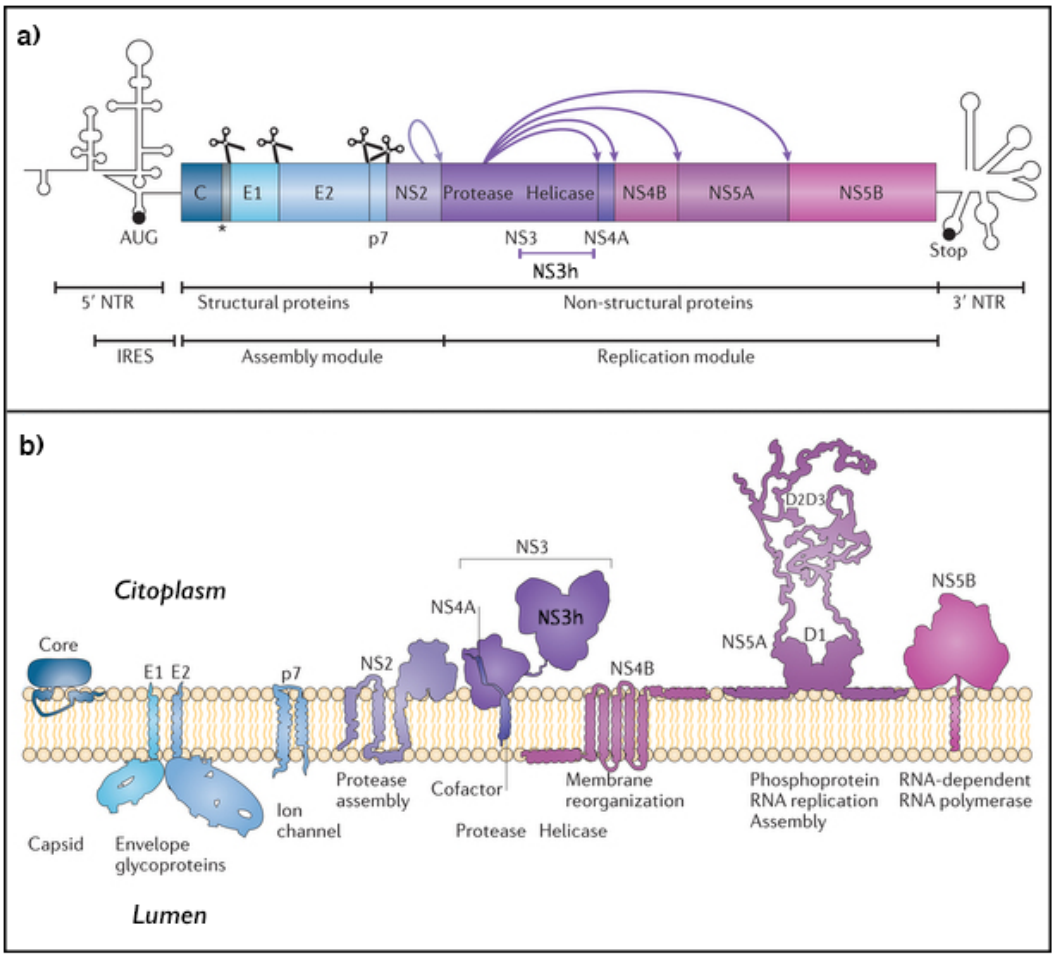


Figure 1.2: HCV genome and its encoded peptides. (a) Structure of HCV genome, showing the long open reading frame and the non-translated regions. The initiation sequence (AUG) that indicates the starting signal for the RNA translation is located in the 5′ untranslated region. (b) Cartoon representing the encoded HCV proteins and its presence in the cellular environment. Adapted from Bartenschlager et al [8] 2013.

constituted by important sequences that act as signals, essential for the regulation of RNA replication. Moreover, the 5′UTR protects an internal ribosome entry site, IRES, that leads the viral translation through a cap-independent mechanism [4]. The long ORF contains a sequence that encodes for a total of ten functional peptides corresponding to four structural proteins, which are responsible for the assembly of the external envelope and the inner icosahedral capsid, and six non-structural proteins, which perform different enzymatic functions related with the

virus replication (Figure 1.2.a). As a result of several works focused on the isolation and characterization of these peptides, now it is possible to describe and understand in a better way their biological activity in the viral infection for most of them.

1.2 Non-structural protein 3: A multifunctional enzyme

Among the encoded peptides, the non-structural protein 3 (NS3) is a molecular motor with a very versatile role, since it is constituted by a serine protease in the N-terminal region and the Superfamily 2 (SF2) helicase in the C-terminal region. The serine protease domain is a trypsin/chymotrypsin-like enzyme composed by nearly 180 amino acid residues, and has an important function for the cleavage of the HCV polyprotein. The structure is stabilized by the tetrahedral coordination complex between a Zn^{2+} cation with three cysteine residues and one water molecule, and the presence of Zn is indispensable for its activity [9]. Under biological conditions, the NS3 serine protease forms a non-covalent heterodimer complex with another encoded peptide, NS4A, which acts as a cofactor to enhance the proteolytic function. The lack of NS4A leads to the cleavage of only one site of the polyprotein (NS5A/NS5B), so the protease activity is remarkably reduced [10]. The promising results obtained in the design of a HIV protease inhibitor drugs have suggested that proteases could be good targets for a structure-based drug design. In this context, NS3 serine protease has been widely studied in order to develop a series of inhibitors that could cause the interruption of the protease activity, the processing of other NS proteins and the viral RNA replication [11]. Based on the previous consideration, some of the current treatments for chronic hepatitis C, are based in the protease inhibition where *boceprevir* and *telaprevir* are the most widespread drugs in the market, although their use is only limited to treat hepatitis C genotype 1 infection.

In addition, the C-terminal NS3 helicase (NS3h) is a DExH/D box ATPase enzyme [12] with a primary structure of nearly 440 amino acid residues, being approximately 60% of the total NS3 molecular weight. The principal function of the helicase is the unwinding and/or translocation on DNA or RNA, so it is not nucleic acid specific, and its processivity is driven in the 3′→5′ direction through a periodic stepwise mechanism [13, 14, 15]. The unwinding process is expected to happen when the

protein interacts with the nucleic acid in the single strand (ss)-double strand (ds) junction, while the translocation is intended to be the movement across the single stranded substrate [16]. The occurrence of these processes is related with several factors, such as the distance between the enzyme and ss-ds junction, and the energy barrier related to the opening of the ds in presence of the enzyme, *e.g.* if the helicase is located far away from the ss-ds junction, it will exclusively translocate along the ss nucleic acid (NA). Under certain conditions, helicases can couple the ss translocation with the ds unwinding, by means of *passive* or *active* mechanisms [17, 18]. In the active process, the helicase affects directly the dsNA stability; whereas in the passive process, the helicase mediates indirectly the unwinding by interacting with the ssNA. Within the active unwinding model, the helicase should have different NA binding sites able to interact with the ds in order to destabilize the base-pairing and with the successive ss portion. Under normal conditions the formation of the dsNA is stable, however, the thermal fluctuations allow transiently the opening in the ss-ds junction. Within the passive mechanism model, the helicase is able to translocate along the transiently formed ss and avoids the further closing to ds [16]. The consideration of these mechanistic features across the NA is essential for the understanding of the NS3h function as well as for other SF1 and SF2 helicases. However, the main biological role and the details of NS3h unwinding/translocation mechanism are still not very clear, but it is presumed to be highly involved in the initiation of the viral RNA replication. Through a periodical mechanism, the NS3h unwinds the base pairing interaction between nucleotides of the RNA duplex and this process should be performed as a previous step of the replication. A potential inhibition in the helicase process should carry out adverse consequences for the further virus replication, and that is why the NS3h emerges as a promising target for the design of novel antiviral therapies in alternative to the current protease-based ones [19].

1.3 An overview of the HCV RNA helicase

The helicase region of NS3, is a motor protein that performs the unwinding/translocation along the viral RNA, and leaves the genetic material available to be read by the HCV RNA polimerase. In spite of this premise, the structural and chemical details about its mechanism remain elusive. In this respect, several experimental

studies such as X-ray crystallography [20, 21, 22], single-molecule experiments [23, 24, 25, 26] and biochemical assays [27, 28, 29] have been done in order to understand the helicase activity. The kinetics and stoichiometry of the unwinding has been measured by means of Fluorescence Resonance Energy Transfer (FRET) and Optical Tweezers. FRET assays on the NS3-DNA complex have shown that every cycle performed along RNA has a stepwise of 3 base pairs (bp) with 3 hidden substeps. For every substep, the helicase unwinds 1 bp per 1 ATP molecule that is hydrolyzed following an inchworm-like mechanism [23]. Studies using double trap optical tweezers on the NS3-RNA complex have also revealed the unwinding step size. From these assays has been measured a resulting average step of 12 bp constitued by 4 substeps of size 3, leading to the same relationship of 1 ATP molecule per 1 bp unwound [30, 24]. Previous spectroscopic studies such as Dynamic Light Scattering assays on NS3(and NS3/NS4A)-DNA have shown the formation of dimeric and oligomeric complexes during the unwinding process [31]. However, NS3h can also act as a monomer, and this is a common feature for all the SF1 and SF2 helicases [32, 33]. Fluorescence Anisotropy experiments on NS3h-DNA, have shown that the helicase acts as an independent unit when the concentration of NA exceeds the NS3h, but when the concentration of the helicase exceeds the NA, multiple peptides tend to interact along the same NA substrate improving in this way the helicase activity. This behavior is known as a cooperative effect [34, 35]. The presence of NS3 serine protease domain in complex with the cofactor NS4A has also a role for the translocation kinetics, since it enhance the binding of the NS3h with the NA substrate through the interaction between the serine protease domain with the displaced ssNA [9]. However, the removal of the serine protease does not imply an inhibition of the helicase activity [21].

From X-ray crystallography, some intermediate snapshots related to the translocation mechanism have been reported in the literature. The structures of NS3 and NS3h have been isolated in complex with ssDNA and ssRNA, in the presence or the absence of the ATP mimic, ADP·BeF_3^- [20, 21, 22]. The helicase is constituted by three catalytic domains, where domain 1 (D1) and domain 2 (D2) are RecA-like units, and the α helical domain 3 (D3) is an accessory unit important to guide the translocation directionality. NS3h has a sequence of 436 residues where D1 comprises the residues from 189 to 324, D2 from 325 to 483 and D3 from 484 to 624. From the crystal snapshots two different conformers named as *open* and *closed* structures have been

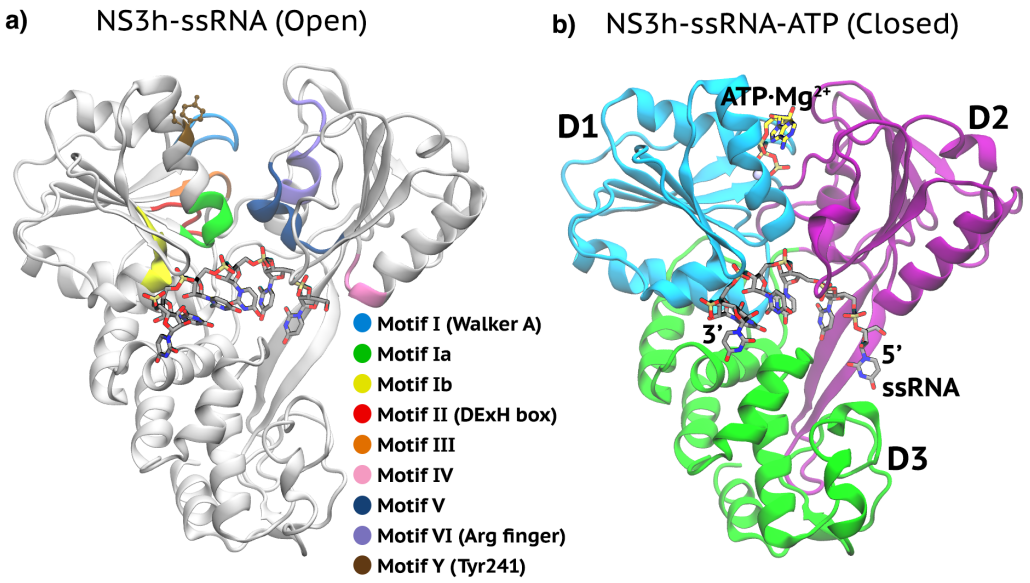


Figure 1.3: Cartoon representation of NS3h (ribbon) in complex with ssRNA and ATP·Mg^{2+} (sticks). (a) Complex NS3h-ssRNA in the open form. All the protein motifs are colored in the ribbon representation and motif Y is also depicted in the ball-and-stick representation. Adapted from PDB: 3O8C. (b) NS3h-ssRNA-ATP·Mg^{2+} ternary complex in its closed conformation showing the three catalytic domains: D1 (cyan), D2 (purple) and D3 (green). Adapted from PDB: 3O8R.

identified. The open conformation corresponds to the protein with a larger distance between D1 and D2 and has been crystallized for the NS3(h)-ssNA form, while the isolated closed conformation has been observed for the NS3(h)-ssNA in complex with the ATP mimic (Figure 1.3). Under biological conditions, when the enzyme binds the ATP, the phosphates of the nucleoside are coordinated to a magnesium ion. The NS3h as many other SF2 helicases contains different conserved motifs which are located in the RecA-like domains and are involved in the ATP and NA binding and the ATP hydrolysis. For the NS3h 9 conserved motifs have been identified (Figure 1.3.a), that corresponds to: Motif I (residues 204 to 211), motif Ia (residues 230 to 235), motif Ib (residues 268 to 272), motif II (residues 290 to 293), motif III (residues 322 to 324), motif IV (residues 365 to 372), motif V (residues 411 to 419), motif VI (residues 460 to 467) and motif Y (tyrosine 241 residue). In particular, the motif I also known as the Walker A is a phosphate binding sequence corresponding to the pattern GXXXXGK(T/S), where X means any amino acid residue. The main role of the Walker A is the ATP binding through the interaction with β phosphate (Pi)

and the nucleoside atoms [36, 37]. Another important sequence is the motif II, also known as DExH/D box, that is involved in the ATP hydrolysis according to some mutational studies performed on different SF2 helicases [12]. Additionally, the motif VI also known as *arginine finger*, is an arginine-rich sequence that is presumed to stabilize the ATP binding and the subsequent hydrolysis reaction.

Studies by means of single molecule experiments allows to estimate the kinetics and the processivity of the enzyme in the translocation mechanism, however, they do not provide details at structural level. Furthermore, in many experimental assays the forces applied during mechanical manipulation are much larger than the actual forces felt by the molecules *in vivo* [38, 39]. Additionally, from X-ray crystallography detailed static snapshots at atomistic resolution for the different forms of the NS3 complexes can be isolated, but they do not provide dynamical details of the mechanism.

The study of the HCV NS3h translocation/unwinding along the ss/dsNA, has been a challenging task from the experimental and theoretical point of view, however a detailed understanding of the NS3h mechanism could also provide some insights for the SF2 helicases. For this reason, the application of computational techniques becomes an attractive alternative in order to describe the molecular details in the mechanism of action of this motor protein. In this context, coarse-grained methods such as the elastic network models [40] have provided meaningful insights in the mechanical motion of the protein residues. Previous works done on the NS3h have described the bidirectional transition between open and closed conformations and modeled the helicase translocation through a ratcheting mechanism along the DNA [41, 42]. Although this method is a good alternative to model the mechanical motion of the peptide at low computational cost and simplicity, it does not describe the chemical details for the ATP and ADP binding with the protein, as well as the interactions with the solvent. Therefore Molecular Dynamics (MD) simulations [43] with accurate force fields become an additional option that can provide dynamical and chemical information to the available crystal structures. From previous studies performing targeted MD simulations on NS3h, the domain motions related to the conformational change from open to closed and backwards have been observed. Subsequent analysis with the elastic network model have displayed the same motion pattern obtained by targeted simulations when the first 10 modes are

considered [42]. Nevertheless, it is important to keep on mind that in targeted MD the pulling setup will determine how realistic is the final structure since it implies an entropic reduction step-by-step which is linked to the number of degrees of freedom that experience the restraint. In several cases a fast pulling protocol can lead to trajectories out of equilibrium, and the final generated structures in general could be thermodynamically unstable. Therefore, the implementation of other MD methods is important to tackle with more details all the steps involved in the mechanism of this interesting enzyme.

Additionally, to perform standard MD simulations is important to model the system with accurate descriptors for the inter and intramolecular interactions. In this context, atomic force fields have been an issue of constant development where AMBER is one of the most well-known parametrization for biomolecules. In the last decades, MD simulations have been specially addressed to model protein systems where several improvements have been done leading to a more realistic description of their interactions. Simultaneously, with less frequency some studies about nucleic acids have been reported in the literature, where simulations of DNA are surprisingly more common compared to the ones of RNA. The reason of this difference is due to several factors such as the poorer database collection for RNA structures compared with DNA and proteins, as well as the less accurate parametrization of the actual force fields for this kind of molecules. In virtue of the above, MD simulations of protein-RNA complexes is a non trivial task, for this reason only a few works about these systems have been reported in the literature (*e.g.* [44, 45]). However, recent force field improvements addressed to describe correctly the canonical and non-canonical RNA interactions, have reflected promising results compared with the experimental data, making possible a more accurate modeling of protein-RNA complexes.

1.4 Outline

In Chapter 2, the computational techniques implemented in this PhD thesis are presented, that comprise plain MD simulations, targeted MD, Well-tempered metadynamics and Hamiltonian replica exchange.

In Chapter 3, the results obtained from standard MD simulations on six NS3-ssRNA complexes are presented: In absence (named *apo* form) and presence of ADP/ATP; starting from the open and closed conformation. The structures were based on the available crystal snapshots. For all the simulations a recent version of the AMBER force field was selected and microsecond time scale MD were performed with the aim of providing statistically meaningful results. The results were complemented with targeted simulations to assess qualitatively the relative stability of the apo, ADP, and ATP structures.

In Chapter 4, the results from enhanced sampling simulations are presented in order to obtain a more quantitative estimation of the relative free-energy values between the two NS3h conformations. In the initial part of the Chapter, the modeling of NS3h conformers in complex with a ssRNA chain of 6 nucleotides in the apo, ADP and ATP form is reported. Additionally, in the second part some preliminary results obtained from the modeling of NS3h in complex with a ssRNA of 10 nucleotides in the ATP form are described.

In Chapter 5 are the final conclusions and perspectives of this PhD thesis.

CHAPTER 2

Computational Methods

In the last decades, computer simulations have represented an important scientific method for the understanding of physical systems. Their valuable role is based on the complementary relationship with experiments and theory. Simulations are built on models that allow the determination of physical and chemical properties, and lead to the development of theoretical and computational methods. The great advantage of performing simulations is the possibility to tackle different levels of complexity, going from the quantum world, to the mesoscopic scale. The constant improvements of these models have allowed a more realistic representation of nature and have provide several details that are not directly accessible through experimental assays. Among the different computational methods, MD simulations have been widely implemented in the study of biomolecules. Thanks to the advances in computer technology, nowadays it is possible to perform atomistic MD simulations at the microsecond timescale [44, 45] or in exceptional cases up to milliseconds when specialized supercomputers such as Anton machine are employed [46, 47]. However, many important processes happen at hundreds of milliseconds or even longer. With these technical limitations, standard MD simulations are mostly able to provide details around the equilibrium state. Processes such as conformational transitions, folding protein and ligand binding, requires generally many computational resources to undergo the transition to another local minima. In this context, several enhanced sampling techniques, such as TMD, metadynamics and replica exchange methods have been developed to accelerate the processes. These techniques coupled with MD simulations allow a

faster transition by adding an harmonical restraint, a bias potential or exploring conformations at different temperatures (or scaled hamiltonians) and help the system to escape from a local trapping free-energy minimum.

This chapter will be dedicated to introduce some basic concepts and the methodology of the computational techniques implemented to carry out the simulations on the NS3h-ssRNA complexes.

2.1 Molecular Dynamics

MD simulations emerged for the first time as a computational technique thanks to the works of Alder and Wainwright in the mid-1950's, where the liquid-solid phase transition for hard spheres and the velocity autocorrelations function decay in liquids were described [48, 49]. Then, in 1964, Rahman independently published the results obtained by atomistic simulations on liquid argon using a Lennard-Jones potential [50]. These works constituted the basis of the MD simulations, so in the late-1970's, the first protein trajectory was obtained for the bovine pancreatic trypsin inhibitor by McCammon et al [51]. Nowadays, due to the great advances in computer performance and algorithmic improvements, the literature has been significantly enriched by many MD simulations studies reported for solvated proteins, protein-NA complexes, lipid systems and so on, addressing different thermodynamical and structural issues and becoming a valuable tool in many areas of physics and chemistry. In general, MD simulations are based on three considerations:

1. Born-Oppenheimer approximation [52], which is based on the difference in mass between nuclei and electrons, so that electronic and nuclear motions can be separated.

2. Adiabatic theorem, introduced by Born and Fock in 1928 [53], stating that electrons will remains in its instantaneous eigenstate if a nuclear perturbation is acting slowly enough and if there is a gap between the eigenvalue and the rest of the Hamiltonian's spectrum.

3. Since the dynamics of the nuclei can be considered without taking explicitly into account the electrons, and the nuclei are heavy enough to be taken

as classical particles, this means that their movements can be described by Newton's equations.

This separation of the electronic and nuclear parts also indicates that, instead of solving the Schrödinger's equation (infeasible for biomolecules), an alternative strategy can be applied, in which the effect of the electrons on the nuclei is expressed by an empirical potential. In this regard, atomic force fields with the optimal empirical parameters can represent an optimal description of potential as well as the physics and chemistry of the systems of interest.

As a consequence of the previous considerations, MD is a computational approach based on the classical equations of motion, *i.e* Newton's second law, which are solved numerically, providing a time evolution of the positions, velocities and accelerations of the particles. Newton's equation is given by $\mathbf{f}_i = m_i \cdot \mathbf{a}_i$, where $\mathbf{f}_i$ corresponds to the force applied on the particle i, m_i to its mass and $\mathbf{a}_i$ to its acceleration. The force can also be expressed as the negative of the gradient of the potential energy, $\mathbf{f}_i = -\nabla_i V$. So the combination of both expressions leads to:

$$-\frac{dV}{d\mathbf{r}_i} = m_i \frac{d\mathbf{r}_i}{dt} \tag{2.1}$$

Therefore, the equations of motions can be related with the derivative of the potential energy V to the changes in position as a function of time. The potential energy is a function that depends of all atomic position. Due to the complicated nature of this function, there is no analytical solution to the equations of motion, so the integration in every step must be solved numerically. MD is a determinist method. This means that once the positions and velocities of all particles are known, the state of the system can be predicted at any time in the future or the past. Numerical integrators of the equations of motion, are built with the aim to find an expression that defines positions at the time $t + \Delta t$ in terms of the already known positions at time t. In this context, the Verlet algorithm [54] is one of most commonly used in MD simulations, due to its simplicity and stability. This algorithm is based on a Taylor expansion for the positions $\mathbf{r}$ around time t, as:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \frac{\mathbf{f}(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{2.2}$$

From this expression, the truncation error of the algorithm when the system is evolved by Δt is of the order of Δt^4, even if third derivatives do not appear explicitly. However, one the problems with this algorithm, is that velocities are not directly generated. Although they are not necessary for the time evolution, their knowledge is important, since they are required to compute the kinetic energy K, whose estimation is used to test the conservation of the total energy $E = K + V$, which is one of ways to verify that a MD simulation is evolving correctly. So, through the Verlet integration, velocities are calculated from the positions as:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t - \Delta t)}{2\Delta t} + O(\Delta t^2) \tag{2.3}$$

From this velocity expression, it can be appreciated that the error associated is of order Δt^2 rather than Δt^4. Additionally, from this equation some numerical errors during the simulation could be induced, hence some variants of the Verlet algorithm have been developed. They give rise to the same trajectory, and differ in which variables are stored in memory and at what times. One of these algorithms is the leap-frog integrator [55], where velocities are evaluated at half-integer time step ($\Delta t/2$) and the new positions are updated from these velocities in the following way:

$$\mathbf{v}\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}\left(t - \frac{\Delta t}{2}\right) + \frac{\mathbf{f}(t)}{m}\Delta t \tag{2.4}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{2.5}$$

Leap-frog integrator is the algorithm employed for all MD simulations performed in this PhD thesis.

2.1.1 Force Field

An atomic Force Field (FF) is defined by a functional form that describes the potential energy of a system of interacting particles, from which the force acting on a single atom is calculated every time step of the MD simulation. The typical FF implemented in the modeling of biomolecules, takes the following functional form [56]:

$$
\begin{aligned}
V = & \sum_{i}^{n_l} \frac{k_{i,l}}{2}(n_i - l_{i,eq})^2 + \sum_{i}^{n_\theta} \frac{k_{i,\theta}}{2}(\theta - \theta_{i,eq})^2 \\
& + \sum_{i}^{n_\phi} \frac{k_{i,\phi}}{2}[1 + cos(n\phi_i + \phi_{i,eq})] + \sum_{i}^{n_\xi} \frac{k_{i,\xi}}{2}(\xi - \xi_{i,eq})^2 \\
& + \sum_{i}^{vdW} \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i}^{coul.} \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}
\end{aligned}
\tag{2.6}
$$

In the first four terms the summation indices run for all bonds, angles, dihedrals and improper torsions defined by the covalent structure of the system, also known as the bonded interactions. The last two terms summation indices run over all the pairs of atoms, separated by a distance $r_{ij} = |r_i - r_j|$ corresponding to the non bonded interactions. More precisely, the first two terms describe the energies associated to the deformations of the bond lengths l_i and bond angles θ_i from their respective equilibrium values $l_{i,eq}$ and $\theta_{i,eq}$. The harmonic form of these terms (with force constants $k_{i,l}$ and $k_{i,\theta}$) represent the correct chemical structure, but prevents modeling chemical changes such as bond breaking. The third and fourth terms describe the rotations around the chemical bond, which are characterized by periodic energy terms (being n the periodicity, $\phi_{i,eq}$ the angle where the potential passes through its minimum value, and $k_{i,\phi}$ the heights of the rotational barriers) in the case of proper dihedrals, or by a harmonic potential (with a force constant $k_{i,\xi}$) in the improper ones. Whereas the fifth and sixth terms describe the van der Waals repulsive and attractive (dispersion) interatomic forces in the form of the Lennard-Jones 12-6 potential, and the Coulomb electrostatic interactions.

The problem of defining a realistic potential that would adequately mimic the true energy surfaces is not a trivial task, so this brings to a series of computational simplifications. The classical MD is based on empirical potentials with a specific functional form, that represents the physics and chemistry. A feasible parametrization has to demonstrate that calculations can produce reasonable molecular structures and interaction energies. The FF parametrization is selected in order to describe a good fit with experimental and quantum mechanics data. It is important to take into account, that atomic FF are empirical forms and there is no absolute *correct* representation to its functional expression as well as for the parameter set. The force

field only provides an estimation of the true underlying interaction energy, which controls all molecular behavior. For this reason, the accuracies of MD simulations depend directly on the quality of the chosen force field. Although the significant efforts done on FF improvements, the parametrizations used nowadays still exhibits some inaccuracies, becoming an issue of constant development.

2.2 Targeted Molecular Dynamics

Targeted Molecular Dynamics (TMD) is a computational method that induce a conformational transition, starting from an initial structure to a final one known *a priori* [57]. The simulations are carried out at ordinary temperature by applying a time-dependent geometrical constraint, where the transition is enforced independently of the height of energy barriers. Since the late-1990's, several works based on TMD calculations have been done to study the protein unfolding, the conformational changes of an enzyme, or to expel ligands from binding sites [58, 59, 60, 61]. In TMD simulations, the Root-mean-square deviation (RMSD) [62] is constrained to a target structure, and is continuously decreased during the simulation. The system is slowly forced to follow a path from its initial conformation to the final one. The conformation of a protein can be described as a vector $\mathbf{X}$ in the $3N$ dimensional space, being N is the number of atoms in the protein. The length of the difference vector between two conformations is proportional to their *RMSD*. TMD simulations are performed with an additional time-dependent holonomic constraint, which acts on the length of the difference vector between the actual conformation and a target structure:

$$\phi(\mathbf{X}) = |\mathbf{X} - \mathbf{X_f}|^2 - \rho(t)^2 = 0 \tag{2.7}$$

where $\mathbf{X_f}$ is the vector corresponding to the target structure. $\rho(t)$ starts at a value equal to the distance between the initial and the final state and is slowly reduced to zero during the simulation. The space of conformations satisfying this constraint corresponds to a hypersphere of radius $\rho(t)$ around the final conformation. A simpler implementation can be obtained using a restraining potential with the following functional form:

$$V_{\mathbf{restr}}(\mathbf{X}) = \frac{1}{2} NK \left(RMSD_{\mathbf{f}} - \rho(t) \right)^2 \tag{2.8}$$

$$RMSD_{\mathbf{f}} = \frac{|\mathbf{X} - \mathbf{X}_{\mathbf{f}}|}{\sqrt{N}} \tag{2.9}$$

where K is the force constant of the harmonic potential, N is the number of constrained atoms and $RMSD_{\mathbf{f}}$ is the Root-mean squared value between a configuration at time point t and the target configuration, and $\rho(t)$ is the reference $RMSD$ value at time t. When $\rho(t)$ is monotonically decreased, the moving structure is gradually driven toward the target structure. A quasi-static equilibrium transformation occurs in the limit when the force constant goes to 0 and the time goes to infinity. So in practice, smaller force constants and longer simulation trajectories will represent more accurate final structures, but at higher computational cost.

With TMD, the cartesian distance between two conformers is assumed to be a good control parameter of the progress along the reaction path. However, this assumption must be taken carefully, since in different molecular transitions is not valid. When a number of different pathways are possible to yield a transition, the choice of the control parameter will affect the results significantly. Indeed, it is important to take into account that $RMSD$ is not symmetric with respect to the initial and final structures since the hypersphere on which the system is restrained is centered on the final (target) conformation. This indicates that different parts of the conformational space are sampled in the forward and backward simulations. Moreover, since the initial value corresponds to the RMSD between the initial and the target conformation and is monotonically decreased, pathways that lead over structures with a larger RMSD to the target conformation cannot be sampled. So, although many TMD simulations are commonly performed in the forward and backward direction, the stability of the solutions with respect to small perturbations of the initial conditions is still a matter of discussion.

2.3 Metadynamics

Metadynamics is a computational method used to estimate the free-energy landscape of a molecular system [63]. This method was initially proposed as a coarse dynamics

in the collective variable space [64], but in the implementation of further works, the dynamics were subsequently accelerated by adding a penalty to already visited states, inspired in the taboo search method [65]. In metadynamics, a set of collective variables, $s(x)$, are chosen which are assumed to provide a coarse-grained description of the system and are function of the cartesian coordinates x. During the simulation, dynamics is driven by the free energy $F(s)$ which is biased by a history-dependent potential constructed as a sum of Gaussians centered on the explored points in the collective variable (CV) space. The bias potential at time t is expressed as:

$$V_G(s,t) = w_G \sum_{i=1}^{t/\tau_G} exp\left[-\sum_{\alpha=1}^{n_{CV}} \frac{[s_\alpha - s_\alpha(i\tau_G)]^2}{2\sigma_\alpha^2} \right] \tag{2.10}$$

where w_G and σ_α are the height and the width of the Gaussian distributions, and $1/\tau_G$ is the frequency at which the Gaussian distributions are deposited. The α index runs for all CV considered in the simulation. This means that the gaussian width must be set for every CV. These three parameters determine the accuracy and efficiency of the free energy reconstruction. If the Gaussians are large, the free energy surface will be explored at a fast pace, but the reconstructed profile will be affected by large errors. On the other hand, if the Gaussian are small, the reconstruction will be accurate, but at high computational cost.

Metadynamics disfavors the exploration of the already visited states in the CV space, and is able to provide an estimate of the free-energy surface. While more and more Gaussians are added to the bias, the system will explore different regions of the CV space. Moreover, Gaussians are mostly deposited at points where the effective total free energy $F(s) + V(s)$ is lower (in a metastable state well), and they will tend to flatten the $F(s) + V(s)$ function. For long simulation times, bias will start growing parallel to itself and will converge to $V_G(s,t) \to -F(s) + C(t)$, where $C(t)$ is an arbitrary constant [66]. Qualitatively, if the CV's are uncorrelated, the time required to reconstruct a free energy surface at a fixed accuracy scales exponentially with the number of CV's used during the simulation. This means, that the performance of the algorithm will rapidly deteriorates as the dimensionality of the CV space increases. Additionally, it is of crucial importance to bias explicitly all the *slow* variables. If an important degree of freedom is forgotten, the free energy estimation will be affected by large errors.

2.3.1 Well-tempered metadynamics

In the last years, metadynamics has become a good alternative to tackle complex molecular transitions that happen at long time scale (also named as *rare events*), and has gain more popularity among the enhanced sampling techniques, however, from the practical point of view, this method faces two problems:

1. Usually it is difficult to decide when to terminate a metadynamics run. When a simulation is performed, the free energy does not converge to a value but fluctuates around the correct one.

2. Metadynamics try to explore the whole CV space, so extending the simulation time could push the system towards thermodynamically non relevant configurations.

In this context, an alternative metadynamics scheme, named Well-tempered metadynamics (WTmetaD), has been introduced in reference [67]. This formulation is aimed to obtain an estimate of the free-energy that converges in the long time limit, and simultaneously, offer the possibility of controlling the regions of free-energy surface that are physically meaningful to explore. These improvements are based on the rescaling of the height of the Gaussian which is decreased with the simulation time according to the following expression:

$$\dot{w}_G = w_0 \, exp\left(-\frac{V_G(s,t)}{k_B \, \Delta T}\right) \tag{2.11}$$

where w_0 is the initial Gaussian height, ΔT an input parameter with the dimension of temperature and k_B the Boltzmann constant. This equation indicates that, after the initial filling, Gaussians of different height are deposited in different regions of the CV space. Specially, on top of wells, where bias has been already accumulated, the additional Gaussians will have small height, while at the border of the explored region, where the bias is still small, the additional Gaussians have larger height. With this rescaling of the Gaussian height, the bias potential smoothly converges in the long time limit, but only compensates a fraction of the free energy,

$$V_G(s, t \to \infty) = -\frac{\Delta T}{T + T\Delta T} \, F(s) + C(t) \tag{2.12}$$

In the long time limit, the CVs sample an ensemble at a temperature $T + \Delta T$ which is higher than the system at temperature T. The parameter ΔT can be set to regulate the extent of free-energy exploration. If ΔT=0, a standard MD is carried out, while if $\Delta T \rightarrow \infty$ a standard metadynamics simulation is performed. This last observation, is very important, since WTmetaD allows to limit the exploration of the CV space only to the regions of interest. So, the choice of ΔT should be done according to the height of the typical free-energy barrier for the problem under consideration, in order to avoid the exploration of regions with higher energy barriers.

2.3.2 Selecting the CV's

When a metadynamics (or WTmetaD) simulation is performed, there are different input parameters that must be set before, such as the set of CV's, the width of the deposited Gaussian potential, the energy rate at which the Gaussian potential is grown and, for the case of WTmetaD there is also the ΔT value that determines the rescaling factor for the Gaussian height along the simulation. However, among these parameters, the choice of a suitable set of CV's can become a difficult task. For this reason some considerations should be taken into account [68]:

- A previous knowledge of the system, based on exhaustive bibliographic search can help to understand the key factors related with the molecular transition of interest.
- Sometimes, a good descriptor does not imply that it might be a suitable biasing coordinate. This means, that in several cases different conformations can be characterized by a certain parameter, but if the changes on this parameter are a consequence of another complex rearrangement, it might be insufficient to drive the reaction reversibly and the results from metadynamics simulations might be quite dissapointing.
- The selected CV's must be able to produce a force in a direction that permits the system to travel between all relevant configurations.
- The lack of convergence of a metadynamics calculation can be due to an incomplete choice for the set of CV's instead of a problem with the specific variables implemented in the enhanced sampling. This means, that can be more useful to explore multiple sets of CV's rather than change them.

2.3.3 Path collective variables

The choice of various CV's allows a better understanding of the molecular transition. However, a larger number of CV implies a significantly longer simulation time in order to reconstruct the free-energy surface properly. For this reason, many efforts have been done towards the development of CV's able to describe complex conformational changes. In this context, the path collective variables introduced in reference [69] provide a good alternative to sample the transition along a hypothetical reaction tube that progresses from reactants to products. This approach focuses on the average flux of reactive trajectories, where the flux can be described as a tube in the configurational space. The tube center corresponds to the reaction coordinate that comes from the average over all reactive trajectories, and its width is related to the fluctuations around this averaged trajectory.

In the practical implementation the reactive trajectories are discretized. For this reason, it is necessary to prepare a putative path constituted by M available snapshots, $\widetilde{q}(i)$, that represent the transition between reactants and products. The frames of this path are generally extracted from high temperature MD, TMD or steered MD. So, the progress along this putative path can be defined by the variable $s(q)$ as:

$$s(q) = \frac{1}{M-1} \frac{\sum_{i=1}^{M} (i-1) e^{-\lambda \|q - \widetilde{q}(i)\|^2}}{\sum_{i=1}^{M} e^{-\lambda \|q - \widetilde{q}(i)\|^2}} \tag{2.13}$$

where λ is a parameter comparable to the inverse of the distance between contiguous snapshots, and $\|q - \widetilde{q}(i)\|$ the distance between the current MD frame and the snapshot $\widetilde{q}(i)$. This parameter gives a value that range from 0 (reactant) to 1 (product) and allows to track the conformations along this ideal path variable. Therefore, a complementary variable is also introduced, and provides the distance from the closest snapshot, $z(q)$, as:

$$z(q) = -\frac{1}{\lambda} ln \left(\sum_{i=1}^{M} e^{-\lambda \|q - \widetilde{q}(i)\|^2} \right) \tag{2.14}$$

The distance from the path, $z(q)$, is a useful parameter, since its exploration can be artificially limited avoiding to visit the regions with high distance from the putative path.

One of the greatest advantages of the path collective variable (s, z) approach, is that it allows to find a thermodynamically favored path which is unrelated to the initial and/or final guesses. Indeed, this fact has been observed in the application with enhanced sampling simulations, where conformational transitions, enzymatic catalysis [70], chemical reactions [71, 72], and protein folding [73] has been reported, providing a good description of the free-energy surface on different chemical contexts.

2.4 Hamiltonian Replica Exchange

Another alternative technique for an exhaustive sampling of the free-energy surface is the coupling of metadynamics with replica exchange MD simulations. In general, replica exchange methods are based on the simulation of a *cold* replica with unbiased statistics, plus a series of additional gradually *hotter* replicas that can explore the free-energy landscape faster. From this framework, the *hottest* replica is able to overcome easily the energy barriers, while the other intermediate replicas are simulated to take the system smoothly from the *hottest* ensemble to the *coldest* one. One of the widely used replica exchange methods is the parallel tempering [74], where the terms *cold* and *hot* are assigned respectively to the replicas with low and high canonical temperature. However, in several cases the transition rates might be caused by entropic effects, so increasing the temperature of the system does not imply that the barrier can be overcome easily [75]. More importantly, the number of required replicas scale with the size of the system and perturbing the temperature of the whole system could lead to conformational changes that are difficult to recover (e.g. protein unfolding). For this reason another replica exchange method has been introduced, were the replicas are constructed with different Hamiltonians. This method, also known as Hamiltonian replica exchange (HREX), allows the exchange between replicas with the same temperature, T, but with the Hamiltonian arbitrary rescaled. So, the resulting product ensemble corresponds to:

$$P(r_1) \times ... \times P(r_N) \propto exp\left(-\frac{1}{k_B T}\left[U_1(r_1) + ... + U_N(r_N)\right]\right) \tag{2.15}$$

where U_i and r_i are the potential energy and the coordinates of the i-th replica and N is the total number of replicas. If $U_1 = U_2 = ... = U_N$ this means that the

ensemble probability only depends on the term $U/(k_B T)$, so a double temperature is equivalent to half of the energy. One of the advantages of choosing the potential energy instead of the temperature, is the fact that energy is an extensive property, while the temperature is an intensive one. This implies that a region of the system as well as some terms of the Hamiltonian can be selected for the rescaling (*heated*). Technically the system is split into two regions, $\mathcal{H}$ (*hot*) and C (*cold*), where each atom is statistically assigned to the $\mathcal{H}$ or the C region. Then, a parametrized Hamiltonian is built as $H(\lambda)$, where $0 \leq \lambda \leq 1$, so if $\lambda = 0$ there is no interaction in the H region (equivalent to infinite temperature), while if $\lambda = 1$ there is no rescaling in H. Taking the parametrization introduced in the reference [76], the replicas are constructed by the following criteria:

- **Electrostatic interactions:** If the atoms are inside the $\mathcal{H}$ region, the atomic charges are scaled by $\sqrt{\lambda}$
- **Van der Walls:** If the atoms are inside the $\mathcal{H}$ region, the ϵ parameter from the Lennard-Jones potential is scaled by λ
- **Proper dihedrals:** If the first and fourth atoms of a proper dihedral are inside the $\mathcal{H}$ region, its potential is scaled by λ, while if either the first or the fourth atom are in the $\mathcal{H}$ region, its dihedral potential is scaled by a factor $\sqrt{\lambda}$

From these considerations, only the FF terms that contribute to the energy barriers are rescaled, so:

- Inside the $\mathcal{H}$ region the interactions are maintained at an effective temperature equal to: T/λ
- Between the $\mathcal{H}$ and C regions the interactions are kept at an intermediate temperature of $T/\sqrt{\lambda}$
- Inside the C region all interactions are maintained at a temperature T

When a metadynamics combined with HREX is performed, the free-energy landscape is simultaneously sampled in every replica, and the CV's space becomes diffusive due to the metadynamics calculation. Thus HREX is able to visit regions with lower probability and brings a better estimation of the energy barriers. Moreover, with this method it is possible to sample other degrees of freedom that are not included by the chosen CV's. After the coupling of both techniques, the acceptance

for a coordinates swap is calculated in the most general manner, and allows the exchange between replicas with different bias potential and Hamiltonian, as

$$\alpha = min\left(1, e^{-\frac{-\bar{U}_i(r_j) + \bar{U}_i(r_i)}{k_B T} + \frac{-\bar{U}_j(r_i) + \bar{U}_j(r_j)}{k_B T}}\right) \tag{2.16}$$

where $\bar{U}_i$ corresponds to sum of the FF (U_i) and bias potential from the enhanced sampling $V_{G,i}$.

CHAPTER 3

Equilibrium properties of NS3h-ssRNA complexes

During the HCV NS3h translocation mechanism, several conformational changes are induced to undergo every step of the cycle. In principle, each of these steps involves an intermediate structure, whose characterization is crucial for having a global view of the process. This can be achieved, for example, by examining details such as the contact maps, in order to find the chronological sequence of the events associated to the translocation mechanism along the NA chain. In this context, the present thesis introduces a novel approach to characterize the NS3h-ssRNA complexes based on computational simulations by means of standard MD in the microsecond time scale. In a few words, the procedure consisted in performing simulations starting from the crystal available structures and from artificial putative intermediates constructed by structural alignment. Therefore, from the MD trajectories, the equilibrium properties and the conformational space were carefully analyzed.

3.1 Building of artificial intermediates

As discussed in Section 1.3, only a few static crystallographic snapshots of the NS3h in complex with RNA have been isolated. These available structures correspond to NS3h-ssRNA (apo) and NS3h-ssRNA-ATP in the open and closed conformation,

respectively. An important step involved in the translocation mechanism is the ATP hydrolysis reaction, that produces phosphate and ADP molecules. This feature suggests that the NS3h-ssRNA-ADP complex could be a possible intermediate in every translocation cycle, although this structure has not yet been isolated for the HCV NS3. It is therefore, reasonable to consider that during the translocation/unwinding mechanism the helicase would exhibit three different stages in order to complete every cycle: **(1)Apo → (2)ATP → (3)ADP → (1)Apo → (⋯)**. In addition, the conformation associated to every step is not obvious, although the available crystal structures have been purified in a specific conformer. For this reason, the characterization of every possible scenario is useful and enlightening to describe these stages. Therefore, it is important to consider three different molecular systems corresponding to apo, with ADP·Mg^{2+} and with ATP·Mg^{2+}, starting from two conformations, *i.e.*, open and closed, for a total of six NS3h-ssRNA complexes.

Since there are only two available crystal structures, named open-apo (PDB: 3O8C) and closed-ATP (PDB: 3O8C), it is necessary to build the other four missing conformations, which correspond to open-ADP, open-ATP, closed-apo and closed-ADP. To build the closed-apo system, the ATP ligand is removed from the crystal closed-ATP. To construct the remaining three systems, open-ADP/ATP and closed-ADP, the protocol implemented is described in the following Subsections (3.1.1) and (3.1.2), and is essentially based on a series of structural alignments between different crystal snapshots.

3.1.1 Structural alignments on Dengue NS3 snapshots

Due to the lack of a crystal structure for the HCV NS3h-ssRNA-ADP complex, other NS3 helicases encoded by flaviviridae virus were considered to observe the ADP binding. Among the possible candidates that belong to this family of virus, Dengue and Yellow Fever proteins has been especially analyzed. However, looking at the protein data bank, only for the Dengue virus are available the crystal structures in the complex NS3h-ssRNA-ATP/ADP (PDB: 2JLV and 2JLZ respectively) [77]. For this purpose, the structural alignment of the *Protein backbone* between Dengue NS3 crystal snapshots were performed using the *RMSD calculator tool* of VMD program [78]. The snapshots selected for the alignment correspond to: NS3-ssRNA-ATP (ATP in ground state), NS3-ssRNA-ATP‡ (ATP in transition state) and the NS3-ssRNA-ADP.

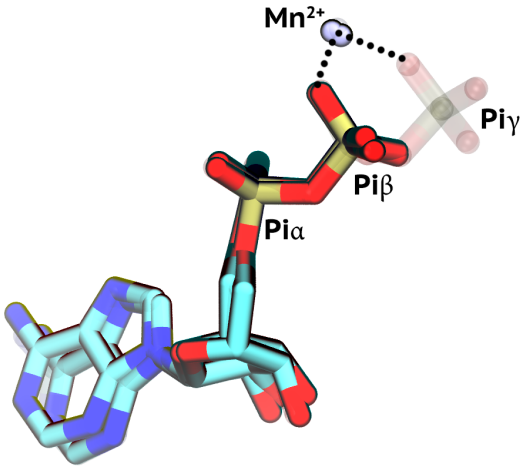


FIGURE 3.1: Ligand position after the structural alignment of the protein backbone between Dengue NS3 snapshots. Representations with solid color correspond to PDB files 2JLX (ATP‡, depicted only up to Piβ) and 2JLZ (ADP). Representation in transparent corresponds to PDB file 2JLV of Dengue NS3 with ATP in the ground state.

From the structural alignment it is observed that the ligand is positioned nearly in the same region for the three different snapshots (Figure 3.1.1). This observation is also supported by the RMSD values of the ligands after the alignment of the protein backbone. For the RMSD calculation the set of ligand atoms up to Piβ is considered. This structural alignment, indicates that ATP and ADP are nearly identically located in the 2JLV and 2JLZ PDB structures of Dengue virus. So, in order to build the HCV closed-ADP structure, the coordinates of ATP are taken from the crystal HCV closed-ATP omitting the Piγ.

TABLE 3.1: RMSD values for Dengue NS3 ligands after the structural alignment of the protein backbone. ADP* denotes the set of ligand atoms selected for the RMSD calculation that corresponds to the ones up to Piβ. The reference structure is the Dengue NS3-ATP complex in the ground state.

PDB file	**RMSD$_{ADP*}$ (Å)**
2JLX (ATP‡)	0.39
2JLZ (ADP)	0.12

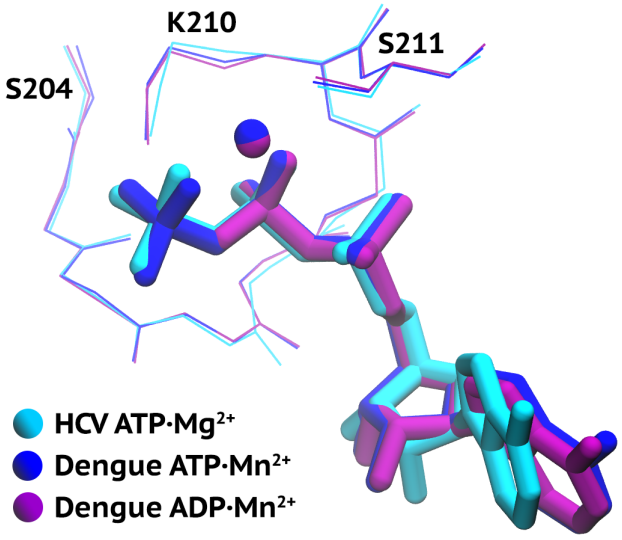


Figure 3.2: Structural alignment of Walker A residues between HCV NS3 and Dengue NS3. Walker A residues are depicted in lines representation and the ligands (ADP or ATP) coordinated to Mg^{2+}/Mn^{2+} in sticks representation.

3.1.2 Walker A: Analysis between two *flaviviridae* helicases

With the aim of complementing the previous results obtained for the Dengue NS3, an additional analysis is done on the Walker A sequence which is a well conserved motif across SF2 NS3 helicases and it is the responsible of the ligand binding. In this case, a structural alignment between HCV NS3 and Dengue NS3 was done to observe the ligand binding similarities in both flaviviridae helicases. To compute the alignments and estimate the RMSD values between the aligned structures, the *trjconv* tool from GROMACS 4.6 [79] and the *RMSD calculator* from VMD program were used. From the results, is observed that ligand placement in the Walker A is virtually identical for Dengue and Hepatitis C proteins (Figure 3.1.2 and Table 3.2).

Table 3.2: RMSD for Walker A after structural alignment between HCV and Dengue proteins. ADP* denotes the set of ligand atoms selected for the RMSD calculation, corresponding to the ones up to Piβ. The reference structure is the HCV NS3-ATP.

NS3-ligand	**RMSD (Å)**	
	Walker A	ADP*
Dengue-ATP	0.24	0.39
Dengue-ADP	0.23	0.39

3.1.3 Open conformations in presence of the ligand

To build the open-ADP/ATP structures the ligand coordinates must be placed on the available open-apo structure. For this purpose, a structural alignment between the Walker A of open-apo and closed-ATP was done in order to get the ATP coordinates on the open form. To perform the alignment and obtain the final atomic coordinates the *trjconv* tool of GROMACS 4.6 was used. Once the final open structure is generated with the ATP correctly placed in the binding pocket, the missing snapshot, open-ADP, is constructed by the Piγ removal from the artificial open-ATP.

3.2 Molecular Dynamics Simulations

3.2.1 Computational details

Every system was constructed by the NS3h peptide (436 aminoacids), a poly-uracil ssRNA chain of 6 nucleotides, and, when present, ADP/ATP·Mg^{2+}. All simulation boxes were solvated with 31058 water molecules (*i.e.* 93174 atoms), 70 Na^+ ions, corresponding to a 0.1 M concentration, and the missing Cl^- counterions to neutralize the charge of the system, for a total of ~100000 atoms. (See also Table 3.3). To perform all the plain MD simulations, GROMACS 4.6 program [79] with AMBER99sb*ILDN-parmbsc0-χ_{OL3}+AMBER99ATP/ADP force field were used. This force field is based on AMBER99 force field [80], and includes additional corrections for the protein backbone [81, 82], the protein side-chain of isoleucine, leucine, glutamate and asparagine amino acids [83], the RNA backbone [84], and the RNA glycosidic torsions (χ angle) [85] as well as the parameters for ATP and ADP [86] and Mg^{2+} [87]. The water molecules solvating the system were described by the TIP3P model [88]. All the MD simulations were performed on the isothermal-isobaric ensemble (NPT), implementing the stochastic velocity rescaling thermostat at 300 K [89] and the Berendsen barostat with an isotropic pressure coupling of 1 bar [90]. The systems were simulated with an integration time step of 2 fs, using periodic boundary conditions, the LINCS algorithm [91] to constrain bonds, and the particle-mesh Ewald method [92, 93] for the long-range electrostatics. In every system, the total simulation time was of 1 μs, and the structures obtained after 200 ns were extracted, in order to start a couple of new control simulations initializing the atomic velocities with a random seed and performing a total of 200 ns each.

Table 3.3: Total number of atoms for every simulation.

System	Peptide	ssRNA	ligand	Mg^{2+}	Na^+	Cl^-	Water	Total atoms
Apo	6528	179	- - -	- - -	70	62	93174	100012
ADP·Mg^{2+}	6528	179	39	1	70	61	93174	100051
ATP·Mg^{2+}	6528	179	43	1	70	60	93174	100054

3.2.2 Simulation protocol

Step 1 Preparation of solute molecules: From the initial crystal structures, the systems are reported as dimers of the entire NS3 (protease domain included). For the simulations only the monomer of the helicase domains is considered, so the additional residues from the PDB file are omitted as well as the crystallographic water molecules farther than 4 Å from the NS3-ssRNA. Then, the missing hydrogens are added to the solute molecules, keeping the same protonation for all the six systems.

Step 2 Energy minimization: After the protonation of the crystal structure, it is performed a minimization procedure for 2000 steps until the maximum force is smaller than 0.24 kcal/(mol Å).

Step 3 Solvation and ions addition: A dodecahedral box is defined placing the solute in the center and setting the box edge at 20 Å from the solute. Then, the box is filled with water molecules and subsequently the Na^+ and Cl^- ions are added reaching a concentration of 0.1 M and neutralizing the charge of the system.

Step 4 Energy minimization: After the solvation and the addition of Na^+ and Cl^- ions, a minimization is carried out for 50000 steps until the maximum force is smaller than 0.24 kcal/(mol Å).

Step 5 Production MD: A long plain MD simulation of 1 μs is performed for every system. The first 200 ns of the simulation are taken as an equilibration stage. The structures obtained at 200 ns are extracted to perform additional control simulations, long 200 ns each, using the same setup stated in the Subsection (3.2.1) and initializing the atomic velocities with a random seed.

3.2.3 Results

3.2.3.1 Conformational analysis

Root-mean-square deviation

With the aim of controlling the stability of the simulated structures, the RMSD is calculated after the structural alignment with a reference frame [62]. The subset of atoms used to calculate the RMSD, was selected based on the location in the protein regions mainly involved in the translocation mechanism and on the proximity to the RNA chain (around 4 Å). A detailed explanation of the subset of atoms is shown in Table 3.4 and illustrated in Figure 3.3. The total number of atoms is 101, corresponding to 84 Cα and 17 nucleotide atoms. To calculate the RMSD values the *g_rms* tool of GROMACS 4.6 was used. From the long MD, the conformation obtained at 200 ns is chosen as the reference structure for the least square fit and the RMSD calculation. This frame is extracted to perform two additional simulations, long 200 ns each, and initialized with a random seed.

Table 3.4: Subset of atoms selected to compute RMSD. The symbol "$\star$" denotes the protein residues around 4 Å of ssRNA.

Protein nodes:		α **carbon**
Motif I		204 to 211, Walker A
Motif Ia		230 to 235
Motif Ib		268 to 272
Motif II		290 to 293, DExH box
Motif III		322 to 324
Motif IV		369 to 372
Motif V		411 to 419
Motif VI		460 to 467, Arginine finger
Motif Y		Tyrosine 241
$\star$around 4Å	D1	229, 253 to 256, 273 to 277 and 296 to 298
	D2	391 to 393, 410, 431 to 434 and 447 to 450
	D3	493, 497, 499 to 503, 551, 555, 556 and 558
RNA nodes:		**P, C4' and N1**
Poly-uracil chain		U3 (5' terminus) to U8 (3' terminus)

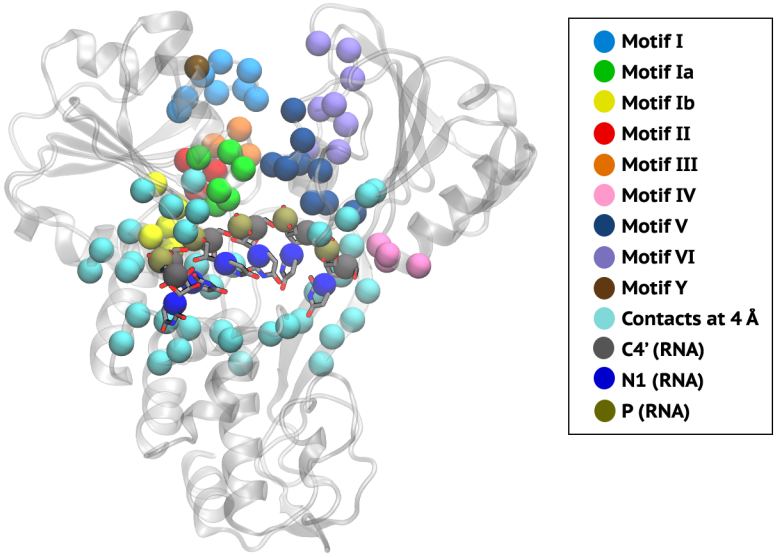


Figure 3.3: Representation for NS3h-ssRNA complex. The helicase is illustrated in ribbons and nucleic acid in sticks representation. Subset of atoms selected to compute RMSD are depicted as solid spheres.

The RMSD is subsequently computed for the long MD and the additional control simulations in order to observe consistent values compared with the long MD trajectory. In general, the RMSD values obtained for the three different simulations of every system does not show large fluctuations indicating that after 200 ns the sampled structures are equilibrated. From the plot of open-apo, a transitory jump is observed for the open-apo system (Figure 3.4.a, violet curve), due to a larger opening between D1 and D2. After that, the original intra-domain gap is recovered.

Interdomain distances

Two structural conformations have been identified in the NS3h, named open and closed. A qualitative way to differentiate the open conformation is identifying the structure with the larger gap between D1 and D2. Due to the thermal fluctuations, this qualitative differentiation could become ambiguous, so a quantitative criteria is introduced in order to differentiate both conformations and control an eventual structural perturbation. In this regard, the gap interdomain is computed as the distance between the centers of mass (COM) of D1 and D2. Then, in order to calculate COM_{D1} and COM_{D2} are considered all heavy atoms for the residues from 198 to 329 in D1 and from 325 to 433 plus from 450 to 483 in D2.The distances D1-D2

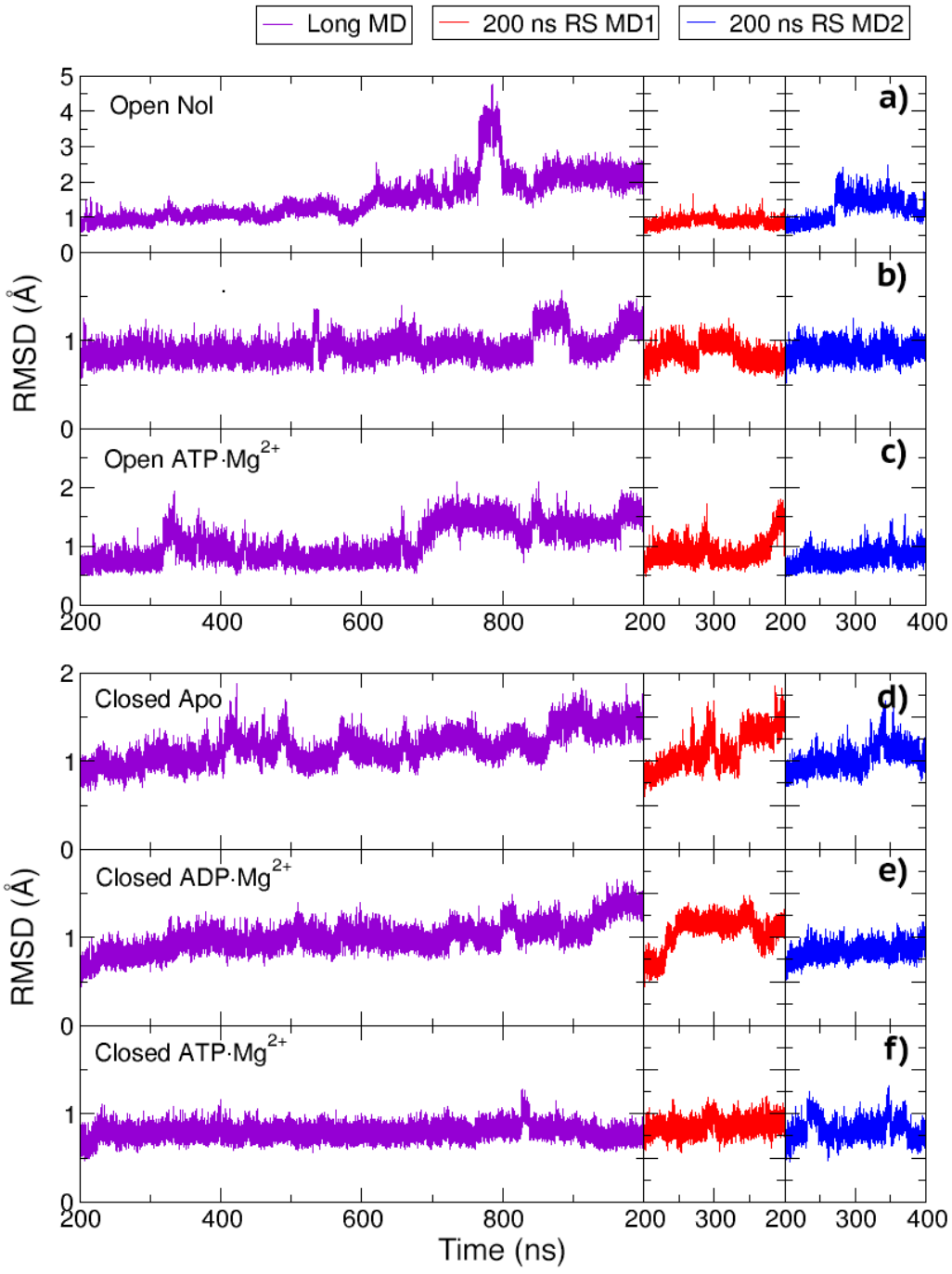


Figure 3.4: RMSD plots for the long MD trajectories and the additional control simulations. Violet curves are RMSD for the long MD going from 200 ns to 1 μs, while red and blue curves are for the control simulations. RS denotes "random seed".

Table 3.5: Gap between D1 and D2 computed as the average distance between the COM of D1 and D2. Results for first and second halves of the trajectories are also shown. Error bars were calculated by binning analysis, with a bin width of 80 ns. Errors lower that 0.05 Å are shown as "0.0".

	COM_{D1}-COM_{D2} (Å)	
Full Trajectory	**Closed**	**Open**
Apo	26.0 ± 0.0	28.9 ± 0.4
ADP·Mg^{2+}	25.8 ± 0.1	27.6 ± 0.0
ATP·Mg^{2+}	25.5 ± 0.0	28.5 ± 0.0
First Half	**Closed**	**Open**
Apo	26.1 ± 0.1	27.9 ± 0.0
ADP·Mg^{2+}	25.6 ± 0.1	27.6 ± 0.0
ATP·Mg^{2+}	25.6 ± 0.0	28.5 ± 0.0
Second Half	**Closed**	**Open**
Apo	25.9 ± 0.1	30.0 ± 0.5
ADP·Mg^{2+}	26.0 ± 0.1	27.6 ± 0.0
ATP·Mg^{2+}	25.5 ± 0.0	28.6 ± 0.1

for the whole trajectory as well as for the first and second halves of the simulation are shown in Table 3.5. In general, the open conformations have an interdomain distance 3 Å larger compared with the closed ones. During the first half of the simulation the open-apo structure exhibited a lower gap, but for the second half the interdomain distance increased and the structure presented higher fluctuations . In the open-ADP structure it is appreciated a gap slightly shorter compared with the other open systems.

Principal component analysis

With the aim of analyzing the motions of the protein during the simulation, a Principal Component Analysis (PCA) was carried out [94]. The PCA was done using a single trajectory obtained by concatenating all the 6 simulations, with a stride of 1 ns. The same subset of atoms selected to calculated the RMSD was used in this analysis. To calculate and diagonalize the covariance matrix and analyze the eigenvectors, the *g_covar* and *g_anaeig* tools of GROMACS 4.6 were used. Once the eigenvectors are extracted, every simulation is projected on the eigenvectors with

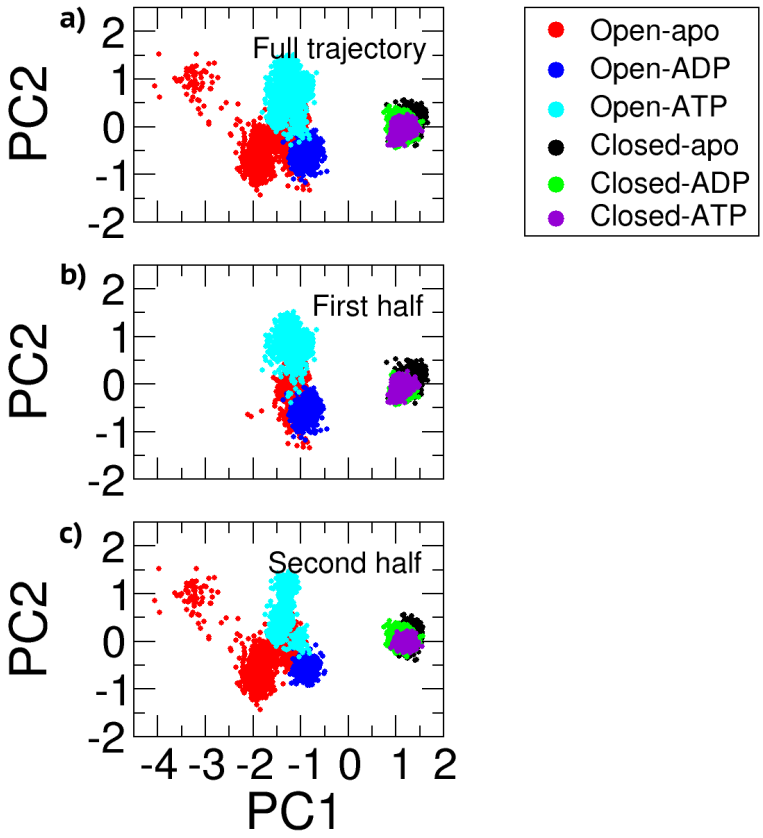


Figure 3.5: Projection of MD trajectories on the first two principal components. PCA is done on the trajectory generated after concatenate all 6 simulations. The results for the entire trajectory are shown in panel (a), while in panels (b) and (c) are shown the projections for the first and the second halves of the simulations.

the two largest eigenvalues. From this projection, it is observed that all the open conformations are more flexible compared with the closed ones. In the second part of the simulation, the open-apo structure explored a region which is very far from the closed structures, and this fact it also supported by the interdomain distance and RMSD values discussed before. However, from this analysis it is difficult to observe the shorter D1-D2 gap for the open-ADP system.

3.2.3.2 ATP/ADP binding pocket

In the protein-ligand complex, the interactions in the binding pocket are determined by the protein conformation and the kind of ligand. For this reason, it is important to analyze with deeper details all the different hydrogen bonds formed and the coor-

dination features towards the magnesium ion. Inspecting the two ATP complexes, it is observed that after 1μs of simulation the closed-ATP system is very similar to the initial crystal structure. In particular, the interactions of ATP·Mg^{2+} complex with Walker A (residues 204 to 211), DExH box (residues 290 to 293), and arginine finger (residues 460 to 467) are maintained during the whole MD simulation (see also Figure 3.6.a), and this structural stability is in agreement with the fact that closed-ATP is one of the isolated crystal structures [22]. However, for the open-ATP system it is observed a reorganization of the binding pocket after 2 nanoseconds of simulation. The triphosphate portion of the ATP exhibits a bending that leads to the coordination of the Mg^{2+} cation to α, β and γ phosphates, which is also an alternative metastable conformation of the ATP/Mg^{2+} complex [95]. Consequently, a contact between Mg^{2+} and D290 (from DExH box) is formed and the initial coordination of the Mg^{2+} cation by S211 is then mediated by a water molecule (Figure 3.6.b).

After the hydrolysis reaction, the interactions between protein and ADP present some differences compared with ATP. In particular, for the open-ADP system a pentacoordinated complex with the magnesium ion is observed in a squared base pyramidal geometry at the beginning of the simulation. After a few nanoseconds, this complex experienced a rearrangement that lead to a hexacoordinated species, where Mg^{2+} interacts with 3 water molecules, the oxygens of the α and β phosphates (Pi's) and the OH- group from S211 (Figure 3.6.d). Then, this hexacoordinated complex is maintained during the rest of the simulation. It is important to take into account that a pentacoordinated Mg^{2+} is not expected for this kind of system [96], so the force field implemented for the magnesium ion [87] is able to describe correctly the hexacoordinated species. For the open conformation some differences are observed according to the kind of ligand. Especially, for the open-ATP complex, the ligand exhibits a very poor interaction with D2, however, for the open-ADP complex a few contacts are formed as a consequence of R467 rearrangement that allows the interaction with the Piβ. Additionally, inspecting the interactions between the motif Y (Y241) and the ATP/ADP ligand, it is observed that for the closed-ADP and open-ATP systems, the stacking between the phenol ring from Y241 and the adenine from the ligand is formed for a small fraction of the time, while for closed-ATP and open-ADP this interaction is more stable. All these features mentioned above suggest that stability of the ligand is enhanced in closed-ATP relatively to the other cases.

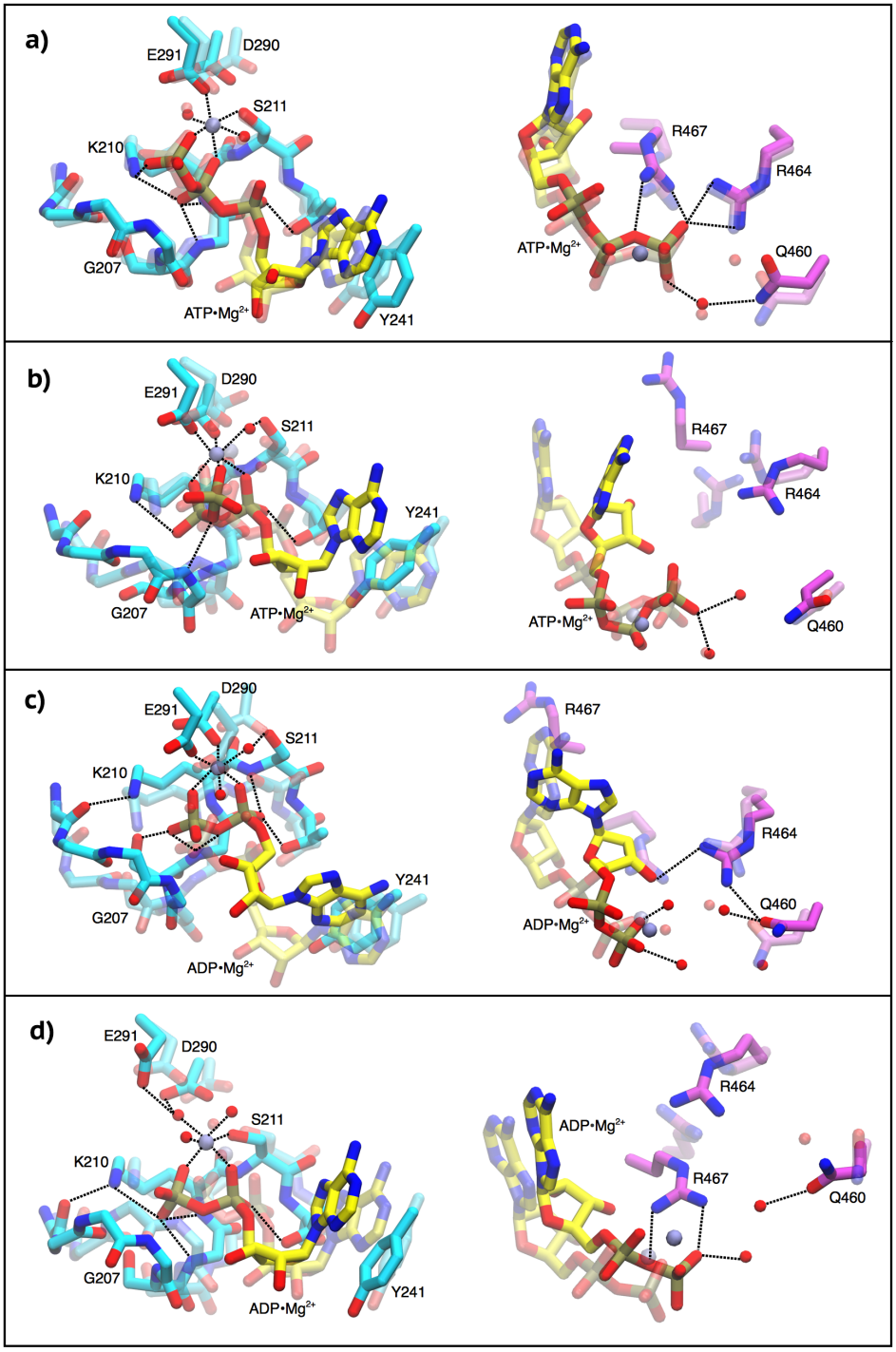


FIGURE 3.6: ATP/ADP binding site. Residues from D1(cyan) and D2 (purple) and ligand (yellow) are shown as sticks, Mg^{2+} ion is shown as a lilac sphere. Snapshots for (a) closed-ATP, (b) open-ATP (c) closed-ADP, and(d) open-ADP are shown for D1 and D2 separately. The initial and final structures (after 1μs of MD) are depicted in transparent and solid color, respectively. Coordination with Mg^{2+} and hydrogen bond interactions are shown with the black dashed line

3.2.3.3 RNA binding cleft

All nucleotides from the simulated RNA interact with the three protein domains and the contacts with the helicase residues occur with the RNA backbone. This important observation is consistent with the fact that NS3 can process the RNA independently from the sequence. The way how RNA interacts with the helicase is related with the conformation. A detailed analysis of the contacts in every domain is described below.

Domain 1: All the interactions between D1 and RNA are maintained in closed and open conformations. The main contacts occur between the side chains from K272 and T269 with the phosphate oxygens from U8 and U7 respectively, and between the V232 backbone (amide nitrogen) with U6 phosphate oxygen (Figure 3.7.a-f).

Domain 2: According to the protein conformation, some contacts between D2 and RNA are lost or shifted by one nucleotide. For the open structures, the T416, T411, and K371 side chains interact respectively with phosphate oxygens from U5, U4, and U4 (Figure 3.7.a,c,e). Conversely, for the closed structures, the contact between T416 with RNA is missing, and both T411 and K371 interact with phosphate oxygens from U5 (Figure 3.7.b,d,f). Moreover, the R393 side chain shows an interaction with phosphates from U6 and U5 in all the open structures, while in the closed-ATP structure it is observed with phosphates from U7 and U6 after ~50 ns and is maintained during the entire simulation. This interaction is missing in the ADP and apo simulations.

Domain 3: The main contacts between D3 and RNA are from W501 and N556 with the nucleobase. All the open conformations and the closed-ATP system form an hydrogen bond with the residue N556. For all the six systems, there is a stacking interaction between the indole group from W501 and the pyrimidine ring from uracil.

Another important feature is observed for the residues W501 and V432 (see also Figure 3.8), which act as gates and are located at the 3′ terminus and 5′ respectively of the simulated RNA. In the closed conformations the residue V432 is between U4 and U3, while for the open conformations is located at the 5′ terminus facing only the U3 ring. This difference could be associated to the shifting of the nucleotide

in every translocation cycle. In addition, W501 is also a crucial residue due to the stable stacking interaction that is formed with the U7 so, in order to complete every translocation cycle, this interaction should be broken to form a new one with the next nucleobase located in the 3′ → 5′ direction.

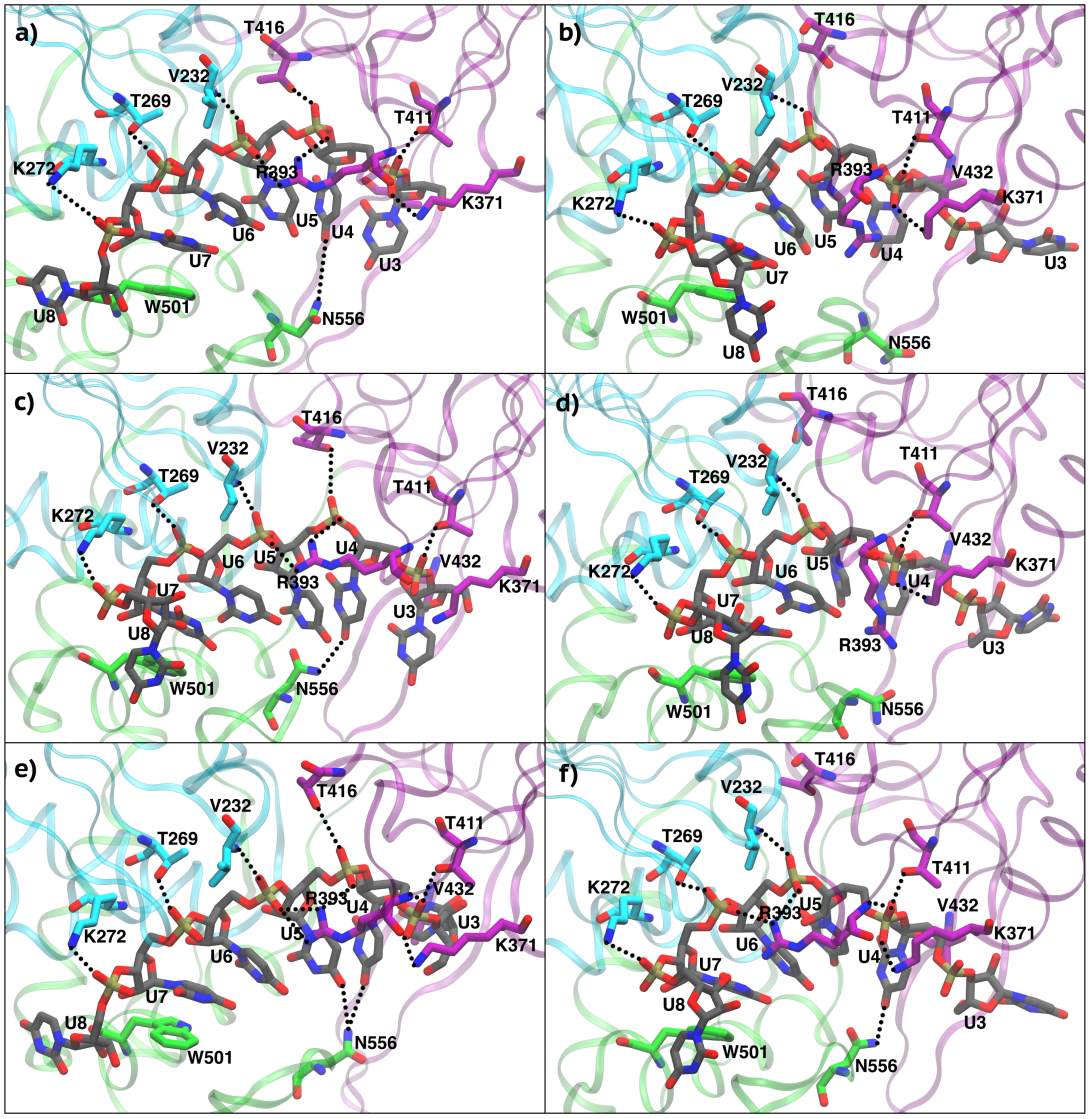


Figure 3.7: Snapshots of the RNA binding cleft. Structures depicted correspond to the RMSD centroids of trajectories extracted by the clustering algorithm discussed in [97], using a cutoff radius of 1.25 Å. RNA chain (gray) and the interacting amino acids from D1 (cyan), D2 (purple) and D3 (green) and are shown in sticks representation. Snapshots are shown respectively for open and closed conformations in the apo (Panels a and b), ADP (Panels c and d) and ATP (Panels e and f) forms.

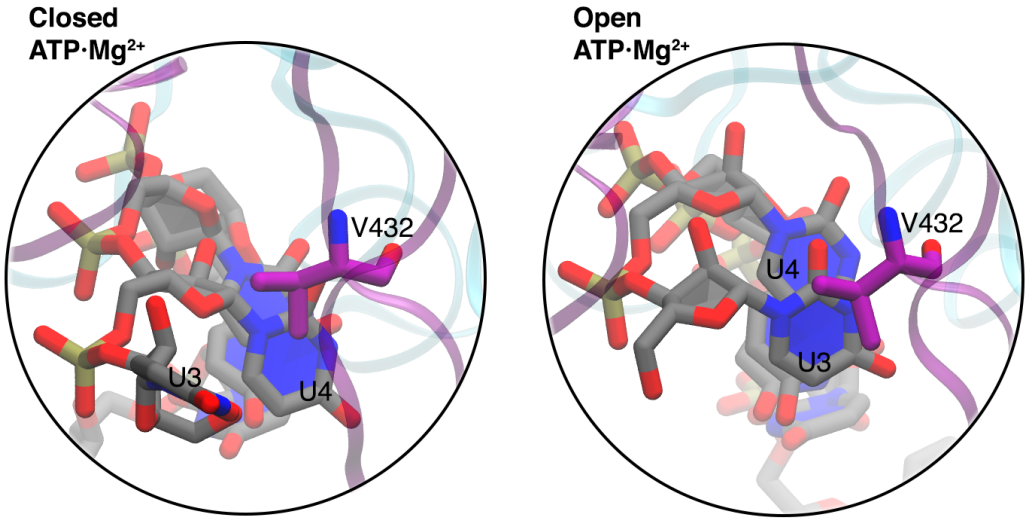


Figure 3.8: Snapshots of RNA 5′ terminus and V432 for the ATP complexes. RNA chain (gray) and V432 (purple) are depicted in stick representation.

3.2.3.4 Hydrogen bonds

Hydrogen bonds are probably one of the most important non-covalent interactions and play a crucial role for the stability and the molecular properties of biomolecules. The inspection of the hydrogen bond network between NS3h, ssRNA and, when present, ligand, could shed light on the differences between conformers and the effect induced by the presence of ATP and ADP. In this regard, all hydrogen bonds were determined using the distance-angle geometric criteria [98]. The cut-off radius selected for the distance donor-acceptor was 3.5 Å and the cut-off angle between acceptor-donor-hydrogen was set at 30°, being 0° the strongest interaction. The number of hydrogen bonds formed during the MD simulation by several groups of solute atoms, including the three protein domains, RNA and, when present, ATP/ADP were computed. The *g_hbond* tool of GROMACS 4.6 was used for the calculations, and the errors were estimated through a binning analysis. Detailed results for the average values are reported in Table 2 and include the number of hydrogen bonds for the two available crystal structures, corresponding to open-apo (PDB: 3O8C) and closed-ATP (PDB: 3O8R). The average number of contacts obtained for the first and second halves of simulation are reported in Appendix B in order to control the consistency of the results.

Table 3.6: Average number of intra-solute hydrogen bonds. Values for the crystal structures open-apo (PDB: 3O8C) and closed-ATP (PDB: 3O8R) are also reported. Errors were estimated by binning analysis with a bin width of 80 ns. Error bars lower than 0.05 are written as "0.0".

Ligand	Apo			ADP·Mg^{2+}		ATP·Mg^{2+}		
	Closed	Open	3O8C	Closed	Open	Closed	Open	3O8R
D1-D2	5.5 ± 0.3	1.6 ± 0.3	0	4.0 ± 0.3	2.3 ± 0.0	2.8 ± 0.2	0.4 ± 0.1	3
D1-D3	2.5 ± 0.1	2.0 ± 0.1	0	1.7 ± 0.1	1.8 ± 0.0	2.3 ± 0.0	1.8 ± 0.1	0
D2-D3	7.6 ± 0.2	6.3 ± 0.1	7	7.3 ± 0.2	6.6 ± 0.2	5.2 ± 0.1	6.4 ± 0.1	5
D1-D1	98.5 ± 0.4	98.7 ± 0.3	90	95.4 ± 0.4	96.6 ± 0.2	95.6 ± 0.2	94.8 ± 0.5	91
D2-D2	113.6 ± 0.5	118.7 ± 0.4	110	113.0 ± 0.7	116.1 ± 0.6	120.2 ± 0.4	114.6 ± 0.5	112
D3-D3	104.7 ± 0.4	105.4 ± 0.3	101	105.6 ± 0.5	104.2 ± 0.3	104.6 ± 0.2	104.3 ± 0.3	102
D1-ligand	- - -	- - -	- - -	7.9 ± 0.4	7.0 ± 0.0	7.7 ± 0.1	8.1 ± 0.2	6
D2-ligand	- - -	- - -	- - -	1.0 ± 0.2	2.6 ± 0.3	4.4 ± 0.1	0.6 ± 0.2	3
D3-ligand	- - -	- - -	- - -	0	0	0	0	0
Ligand-ligand	- - -	- - -	- - -	1.0 ± 0.0	0	0.1 ± 0.0	1.0 ± 0.0	0
RNA-ligand	- - -	- - -	- - -	0	0	0	0	0
D1-RNA	5.2 ± 0.1	4.1 ± 0.1	4	4.1 ± 0.2	3.9 ± 0.1	4.4 ± 0.1	3.8 ± 0.0	3
D2-RNA	7.4 ± 0.2	4.7 ± 0.3	4	7.4 ± 0.1	5.5 ± 0.1	9.3 ± 0.1	5.7 ± 0.1	3
D3-RNA	0.5 ± 0.0	4.0 ± 0.2	3	0.7 ± 0.1	2.0 ± 0.1	1.0 ± 0.0	3.5 ± 0.2	0
RNA-RNA	3.9 ± 0.1	3.3 ± 0.2	1	4.0 ± 0.1	5.0 ± 0.1	4.4 ± 0.1	3.5 ± 0.2	1
Total	349.4 ± 0.9	348.8 ± 0.8	320	353.1 ± 1.1	353.6 ± 0.8	362 ± 0.6	348.5 ± 0.9	329
ΔHB_{oc}	-0.6 ± 1.2			0.5 ± 1.4		-13.5 ± 1.1		

In general, the values obtained for the simulations starting from the experimental structures are very similar to those obtained for the crystals (see Table 3.2.3.4 columns open-apo vs 3O8C, and closed-ATP vs 3O8R). However, all the averages are slightly larger for the MD trajectories when compared to the crystal structures. In particular, this discrepancy in the open-apo case is due to the formation of some interactions between D1 and D2 during the MD, as a consequence of a slight interdomain proximity that allows the interaction between flexible loops. These contacts were identified and correspond to the hydrogen bonds between an acceptor from the arginine finger (Q460) and the donors from the DExH box motif (E291, H293). Moreover, new interactions between D1 and D3 from the simulated trajectories of both open-apo and closed-ATP structures were observed corresponding to the contacts between residues T305-R512 and S297-E493. These new contacts observed from the MD trajectories are not present in their respective crystals structures. The reason of these differences is that the distances donor-acceptor and the angles acceptor-donor-hydrogen are slightly larger from the cut-off values selected to identify hydrogen bonds, so small fluctuations of these residues lead to count these hydrogen bonds in the MD. The same explanation is also valid for the larger number of intradomain D1-D1 and D2-D2 hydrogen bonds.

Analyzing the apo and ADP forms, it is observed a very similar number of intra-solute hydrogen bonds for the open and closed conformations. In particular, the number of D1-D2 interactions missing in the open structure are compensated by an equivalent number of intradomain D2-D2 bonds. Furthermore, the protein-RNA bonds are in general maintained during the simulation, whereas some D2-RNA bonds are replaced with D3-RNA bonds upon opening (Table 3.2.3.4). Even tough, the number of hydrogen bonds is not a measurement of structural stability, it is expected to give a large contribution to the interaction energy. These results obtained for the apo and ADP forms suggest that open and closed conformers might have a comparable stability. Conversely, in the ATP complexes the closed structure exhibited nearly 15 additional hydrogen bonds in contrast with the open one. This is due to a combination of several effects, such as the formation of new D1-D2 and D2-D2 hydrogen bonds, and the interaction of the ligand with D2. The increased number of hydrogen bonds suggests that ATP stabilizes the closed structure. In the presence of the ligand, some differences are appreciated for the contacts D1/D2-ligand. In particular, for the open-ATP structure, the ligand interacts

poorly with D2, while in presence of ADP, the total counts are larger due to the rearrangement of the R467. This result is also supported by the binding pocket analysis (subsection 3.2.3.2).

3.2.3.5 Electrostatic interactions

Electrostatic interactions were computed to determine the binding of RNA and ligand with the helicase. The interactions between protein-RNA and protein-ligand were estimated based on the Debye-Hückel energy (G^{DH}) [99], using the implementation in reference [100]:

$$G^{DH} = \frac{1}{k_B T \epsilon_w} \sum_{j \in B} \sum_{i \in A} q_i q_j \frac{e^{-\kappa |r_{ij}|}}{|r_{ij}|} \tag{3.1}$$

where k_B and T denote the Boltzmann constant and the temperature of the system, ϵ_w is the water dielectric constant, A and B correspond to the set of atoms of the selected interacting molecules, i and j are the atom indexes for the sets A and B, q_i (q_j) is the charge of atom i (j), $|\mathbf{r}_{ij}| = |\mathbf{r}_i - \mathbf{r}_j|$ denotes the distance between atoms i and j and $1/\kappa$ is the screening length. All calculations were made using PLUMED plugin [101]. The ionic strength was selected at 0.1M that corresponds to a screening length of ≈ 10 Å. Debye-Hückel interaction energies were computed discarding the initial 200 ns of simulation and the associated errors were calculated through a binning analysis.

Electrostatic interaction between protein and RNA was determined as a qualitative parameter to rank the RNA binding energy for all the sampled conformations. In general, the protein-RNA interactions are slightly larger (~1 kcal/mol) in the closed conformations compared with the respective open ones, and this trend is independent from the presence and kind of ligand (Table 3.7). However, for the system closed-ATP, the protein-RNA interaction is stronger than in the other cases. To inspect the influence of the ligand in the protein-RNA interaction, the direct RNA-ligand interaction (also shown in Table 3.7) was considered separately. From these results it is observed that differences in the RNA binding are related to a rearrangement of the protein due to the presence of the ligand and not to a direct RNA-ligand electrostatic coupling. In order to determine the contribution of the individual amino acids, the G^{DH} is estimated for all the protein residues located within a distance of 6 Å from the RNA. (see Appendix A for the detailed values).

Table 3.7: Electrostatic interaction computed as Debye-Hückel energies (G^{DH}). Protein* denotes Protein-RNA complex. Errors were computed from binning analysis (bin width: 80ns). Errors lower that 0.05 kcal/mol are shown as "0.0".

	ΔG^{DH} (kcal/mol)	
Protein-RNA	**Closed**	**Open**
Apo	-4.9 ± 0.0	-3.7 ± 0.2
ADP·Mg^{2+}	-5.0 ± 0.1	-3.9 ± 0.0
ATP·Mg^{2+}	-5.5 ± 0.1	-4.7 ± 0.1
RNA-Ligand	**Closed**	**Open**
ADP·Mg^{2+}	-0.1 ± 0.1	-0.2 ± 0.0
ATP·Mg^{2+}	-0.2 ± 0.1	-0.2 ± 0.1
Protein*-Ligand	**Closed**	**Open**
ADP·Mg^{2+}	-0.4 ± 0.0	-1.1 ± 0.0
ATP·Mg^{2+}	-0.8 ± 0.0	-0.1 ± 0.0

From this analysis, it is appreciated that for the closed conformations the most interacting residue is K371 while for the open structures is R393. This result is consistent with the fact that lysine and arginine are positively charged amino acids that can interact attractively with the negatively charged RNA backbone. Whereas, an exception is observed for the closed-ATP complex where also the R393 is the most interacting residue, due to the structural rearrangement discussed previously in the RNA binding cleft analysis (subsection 3.2.3.3).

As a final analysis, the interaction between the ligand and the protein-RNA complex was computed and the results are also reported in Table 3.7. In presence of ATP, it is observed a higher electrostatic interaction of the ligand with the protein in its closed conformation, while in the ADP case, the stronger interaction is between the ligand and the protein in its open conformation. Although the G^{DH} provides qualitative values to estimate the binding affinities, the results are clearly consistent with the fact the ATP has a higher interaction with the NS3h in the closed conformation.

3.2.3.6 Enthalpy

In the isothermal-isobaric ensemble, enthalpy (H) is defined by: $H = \langle U \rangle + p\langle V \rangle$, where U is the potential energy, p the external pressure, and V the volume of the simulated box [102]. For all 6 systems, the enthalpy values were computed with the *g_energy* tool of GROMACS 4.6 program. For rational energy differences, the simulations corresponding to different conformations (open or closed) of the same system were prepared to contain exactly the same number and types of atoms (see also Table 3.3). The energy values might be affected by numerical details, so to keep uniformal conditions for all the systems, the simulations were performed using identical settings on identical machines. The first 200 ns of MD were performed using an hybrid GPUs+CPUs calculation and the remaining 800 ns using CPUs only. To control the effect of using GPUs in the calculations, the energy values obtained for GPUs+CPUs trajectories were compared with CPUs ones, without showing significant discrepancies. The total energy calculation might be affected by statistical errors usually due to fluctuations of solvent contributions. The average enthalpies were computed discarding the initial 200 ns of equilibration in every trajectory, and the errors were calculated using a binning analysis with a bin width of 80 ns.

Enthalpy differences between open and closed conformations were computed, selecting the systems with exactly the same number and types of atoms (Table 3.8). From the results, it is observed that all open conformations have a systematically lower enthalpy than their respective closed ones. However, this difference is more considerable for the ATP complexes (~54kcal/mol). This result is very puzzling since the crystal snapshot in complex with ATP has been isolated in the closed conformation. The statistical errors obtained from binning analysis are very low for all the systems. Estimation of the enthalpy values for the first and second halves of the simulations are also reported in Table 3.8 and the values are consistent with the reported ones from the entire trajectory.

3.3 Targeted Molecular Dynamics

With the aim of estimating the relative stability between closed and open conformations, a series of short TMD simulations starting from the closed structures in apo/ADP/ATP forms were performed [57], following a protocol similar to the one

Table 3.8: Enthalpy differences between open and closed conformations. The values are computed between systems with same number and kind of molecules. Error bars are computed from binning analysis (bin width: 80 ns).

Ligand	ΔH_{oc} (kcal/mol)		
	First half	**Second half**	**Full trajectory**
Apo	-15.2 ± 3.5	-4.2 ± 3.8	-9.7 ± 2.8
ADP·Mg^{2+}	-28.8 ± 4.1	-21.8 ± 3.8	-25.3 ± 3.1
ATP·Mg^{2+}	-51.6 ± 7.6	-56.5 ± 4.6	-54.1 ± 4.1

used in reference [42]. From this analysis, the relative stabilities of the closed and open structures in presence and absence of the ligand were estimated by measuring the work performed during the pulling.

3.3.1 Computational details

A time dependent harmonic restraint with a stiffness 500 kcal/(mol Å^2) was applied to the RMSD calculated with the open snapshot. RMSD were computed using all heavy atoms of protein and RNA. The restraint was moved from its initial value to zero in a simulation long 20 ns. In the first nanosecond the restraint was left at the starting value, while in the subsequent 19 ns, the restraint was moved linearly in time to zero. The RMSD values between the closed conformations after equilibration and the crystal open structure were calculated for all three cases. In order to avoid any bias, all systems started with the same RMSD values, resulting in three different pulling conditions. Every system was simulated three times, initializing the atomic velocities with a different random seed. This brings a total of 27 independent simulations in 3 different forms (apo/ADP/ATP), with 3 different pulling protocols (starting from initial RMSD of apo/ADP/ATP structures) and 3 different random seeds. All TMD simulations were performed using GROMACS 4.6 program and PLUMED plugin in its version 2.1. All set-up details are similar to the ones considered for the standard MD simulation explained in Subsection (3.2.1).

3.3.2 Analysis of the pulling simulations

The average work provides an overestimation of the free-energy change, however, a more accurate assessment could be obtained using the Jarzynski's equality [103]. In

spite of that, a very large number of simulations might be required to get a proper estimate of the free-energy change, due to the complex rearrangement involved in this conformational change. Moreover, when a strong restraint is applied to many degrees of freedom in a TMD simulation, the value of the free-energy change would be affected by the decrease in the entropy of the restrained atoms. Despite the limitations mentioned above, the average work can be used as a qualitative parameter to rank the free-energy differences for apo, ADP, and ATP structures towards the opening. From the different pulling protocols, is appreciated that the work required to open the apo form is systematically lower than the one required to open the ADP and ATP forms (Figure 3.9). This result is consistent with the binding pocket analysis, the hydrogen bonds network between solute molecules, and the electrostatic interactions discussed above, so the enthalpy values might be compensated by a large entropic change in presence of ATP. This ligand stabilization of the closed structure is also compatible with the fact that the NS3 in complex with ATP has been isolated in this conformation.

3.4 Discussion

After 1μs of MD simulation, all the systems were stable and did not experience any significant structural perturbation. This validates the protocol used to build the four artificial intermediates. The structural stability of the different conformations was analyzed by RMSD calculations, the projections on the first principal components and the interdomain distances. The ATP and ADP open structures have a lower enthalpy than their respective closed structures, indicating that they have been properly equilibrated. In addition, the closed-ATP complex, presented less fluctuations compared with the open one and with closed-apo/ADP. All open structures, including the crystal apo form, exhibited larger fluctuations. This indicates that NS3 serine protease domain might stabilize the open-apo in the experiments. From the binding pocket analysis some differences were observed according to the type of ligand and the conformation of the protein. In particular, the closed-ATP structure maintained the contacts from the crystal between the Motifs I, VI and the ligand, during whole simulation. However, the open-ATP complex presented some rearrangements due to the missing interactions with D2. This could imply an ATP-induced stabilization of the closed conformation. The modeling of ADP complexes were considered as

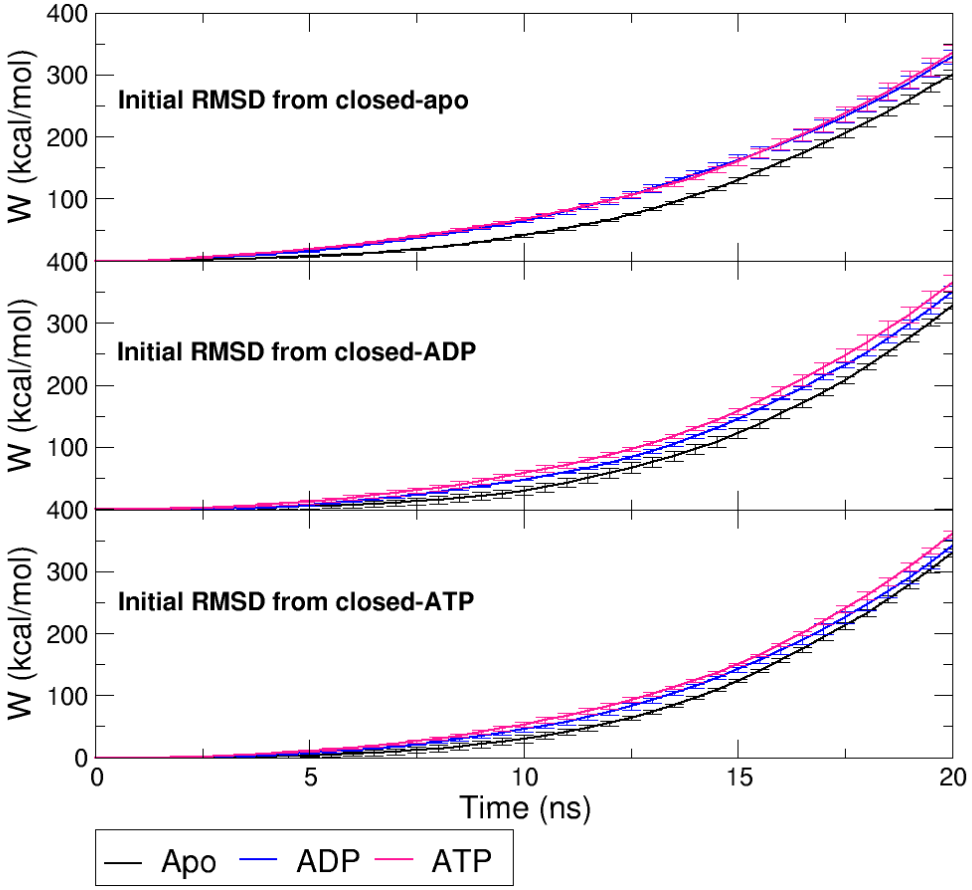


Figure 3.9: Perfomed work vs simulation time from TMD. Average work was obtained from 3 TMD simulations initialized with different random seeds. Standard deviations of the three trajectories are shown as error bars.

possible intermediates after the hydrolysis reaction. From these structures, more interactions between the ligand and D2 were observed. This suggests that the stabilizing effect of ADP on the open conformation is larger than the one induce by ATP. Thus, after hydrolysis reaction, the closed conformation might be destabilized and a conformational transition towards to open structure could be favored. From the protein-RNA interactions, it is observed that NS3 interacts mainly with the phosphate groups of the RNA backbone. This is consistent with the experimental assays where the helicase is also processing DNA [23, 21], since NS3h is not nucleic acid specific. Contacts formed between T269 (in D1) and T411 (in D2) with the phosphate oxygens present different patterns according to the protein conformation and act like hooks with the RNA during the unwinding/translocation mechanism.

The same type of interaction has been reported for other SF2 RNA-helicases such as Vasa drosophila and eIF4A, suggesting a possible common mechanism among several helicases of this superfamily [23, 104, 105]. At the same time, the contacts with residues R393 and N556 are present only in the open conformations and the closed-ATP complex, indicating an ligand-dependent allosteric effect for the latter case. The average number of hydrogen bonds for the ATP systems, showed a larger interactions between solute molecules in the closed structure when compared with the open one. This is compatible with the eventual ATP-dependent stabilization in the closed conformation mentioned before. In contrast, both apo and ADP complexes presented an equivalent number of hydrogen bonds for open and closed conformers.

All features discussed above, indicates that the free-energy difference ΔG towards the closing might be thermodynamically more favored for the ATP case. This is supported by the TMD simulations and the analysis of the electrostatic interactions estimated within the Debye-Hückel approximation. This ΔG can be expressed by the difference of the ligand affinity in the two conformations. So,

$$\Delta G_{ATP} - \Delta G_{Apo} = -k_B T \log K_d^{(closed)} + k_B T \log K_d^{(open)},$$

being k_B the Boltzmann constant, T the temperature of the system, and $K_d^{(open)}$ and $K_d^{(closed)}$ the ATP dissociation constants in the open and closed conformation. This expression indicates that the ATP affinity is larger in the closed form. Contrarily, enthalpy calculations pointed out an opposite result where open-ATP is dramatically more stable than closed-ATP. However, although enthalpy provides an important contribution to the free-energy, in several cases is not the principal factor to measure relative stabilities, and it is observed an entropic contribution that compensates the enthalpic one, also known as Enthalpy-Entropy Compensation (EEC) phenomenom [106]. From these simulations it is difficult to determine the entropic differences that not only depend on the solute flexibility but also on the solvent entropy which is connected with e.g. hydrophobic effects in the cavity between D1 and D2. For this reason, the use of enhanced sampling techniques becomes an interesting alternative to estimate the free-energy differences [107, 108]. The further implementation to the study of the NS3-ssRNA complex is discussed in the Chapter 4. Additionally, all the analysis discussed in this Chapter about the six simulates systems, is also reported in a recent publication [109].

Total simulation time

Table 3.9: Total simulation time for the calculations reported in Chapter 3

System		# runs	time/run (μs)	total time (μs)
Standard MD				
open-apo		1	1	1
closed-apo		1	1	1
open-ADP		1	1	1
closed-ADP		1	1	1
open-ATP		1	1	1
closed-ATP		1	1	1
Subtotal time				6
Control MD				
open-apo		2	0.2	0.4
closed-apo		2	0.2	0.4
open-ADP		2	0.2	0.4
closed-ADP		2	0.2	0.4
open-ATP		2	0.2	0.4
closed-ATP		2	0.2	0.4
Subotal time				2.4
TMD closed → open				
Apo	$RMSD_{apo}$	3	0.02	0.06
	$RMSD_{ADP}$	3	0.02	0.06
	$RMSD_{ATP}$	3	0.02	0.06
ADP	$RMSD_{apo}$	3	0.02	0.06
	$RMSD_{ADP}$	3	0.02	0.06
	$RMSD_{ATP}$	3	0.02	0.06
ATP	$RMSD_{apo}$	3	0.02	0.06
	$RMSD_{ADP}$	3	0.02	0.06
	$RMSD_{ATP}$	3	0.02	0.06
Subotal time				0.54
Total time:				**8,94**

CHAPTER 4

Enhanced sampling of NS3h conformational changes

The results obtained from the standard MD simulations suggest that a strong entropic compensation stabilizes the experimentally isolated closed-ATP snapshot. However, the determination of the entropy is not a trivial task, and a direct estimation of the free-energy is necessary to determine quantitatively the relative stabilities between the studied intermediates. For this reason, the use of enhanced sampling techniques emerges as a helpful strategy to reconstruct the free-energy landscape of the simulated conformations. In this particular case, the systems were sampled by means of metadynamics simulations (in its well-tempered scheme), a method that has been widely used in several chemical applications such as the ligand binding, the phase transitions and the conformational changes [110, 111, 112, 113]. The great advantages of this technique are based on the possibility to accelerate the sampling of rare events by pushing away the system from the local minima and explore different pathways when the system tends to escape the free-energy minima crossing the lowest transition state [114]. Additionally, a further approach that combines metadynamics with the HREX simulations was used. The application of these techniques accelerates efficiently the free-energy estimation, allows a wide conformational sampling and exploits the parallelization of the codes on high performance computers. As a final analysis, one of the crystallographic structures is taken to build a longer nucleotide chain and

analyze the effect of the NA length in the conformational transitions. Some of the enhanced sampling simulations reported in this Chapter are shown as preliminary results, since they need a longer computational time to converge.

4.1 Modeling NS3h in complex with 6 nt ssRNA

4.1.1 Building a set of collective variables

4.1.1.1 Linear path variables

In order to monitor the conformational transitions, a set of order parameters built as a linear combination of RMSD's from two reference structures is defined. In reference [42], a criterion to control a conformational change based on the difference between squared RMSD from the open and closed structure was introduced. A similar approach is defined in this application, where a variable S is calculated in the same way but are taken into account some additional aspects. To compute the RMSD values, is used the same subset of atoms described in Subsection (3.2.3.1). This choice is done since RMSD is a global parameter and the contribution of fluctuating loops that do not participate in the mechanism can produce some noise in the RMSD calculation. Simultaneously, a variable Z that measures the distance from a hypothetical transition path obtained as a linear interpolation between the two reference structures is defined. The expressions for (S, Z) variables take the following forms:

$$S = \frac{R_o^2 - R_c^2}{2R_{oc}} \tag{4.1}$$

$$Z = \frac{R_o^2 + R_c^2}{2} - S^2 - \frac{R_{oc}^2}{4} \tag{4.2}$$

Being R_o and R_c the RMSD of the simulation frame from the open (PDB: 3O8C) and closed (PDB:3O8R) crystal structures respectively, and R_{oc} the RMSD between the open and the closed structures. A cartoon representation that illustrates the building of both variables is also shown in Figure 4.1. From this scheme, it is observed that the difference between squared RMSD's is related to the progression variable $s(q)$ used in the path collective variables approach [69, 115].

With the aim of controlling the suitability of the linear path variables as CV's for the enhanced sampling, a projection of the 6 MD trajectories analyzed before on the

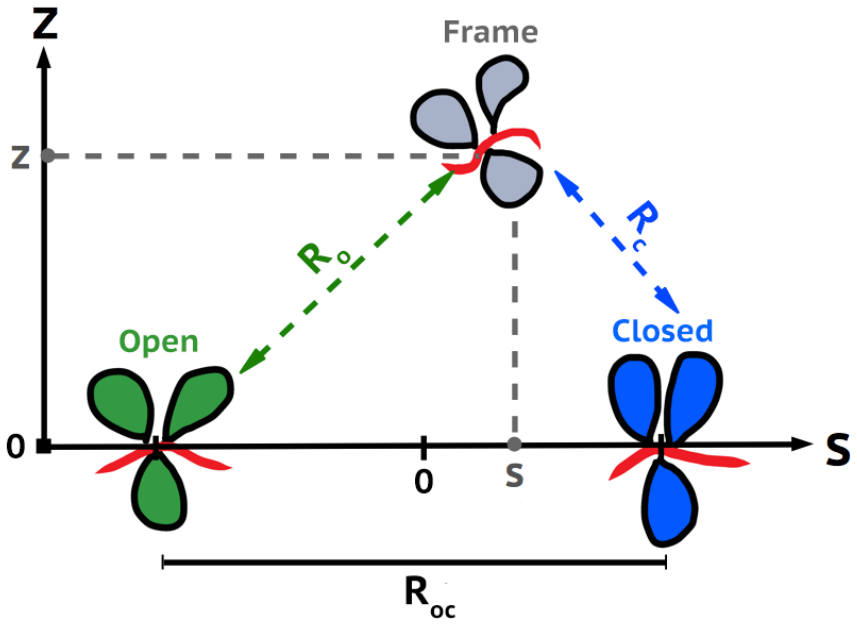


FIGURE 4.1: Cartoon diagram for the definition of linear path variables (S, Z). The two reference structures considered in this application correspond to the open and closed crystal snapshots from reference [22].

(S, Z) space was done. As discussed in the previous Chapter, all simulations were stable without undergoing any significant structural perturbation. Interestingly, this fact is also supported by the projection of the trajectories on the linear path variables, where all simulations remain near to the starting experimental structure ($Z \lesssim 2$ Å), and every conformation is well defined in the (S, Z) space since the values do not overlap between them (Figure 4.2). From the trajectories, it was also observed that protein-RNA complexes were more flexible in the open conformation than in the closed one. This fact is also reflected in the projections, due to the larger fluctuations of the Z value for the open structures. In presence of ADP and ATP, the fluctuations are partly reduced for both open and closed structures. This is consistent with the fact that these systems have explored a smaller region in the linear path space. Analyzing the distribution of the S variable, it is observed that both apo and ADP open systems explore regions that are slightly towards the closed reference structure. However, a bimodal distribution for the open-apo was observed, with one peak corresponding to a lower distance from the closed structure and the other one corresponding to the transient larger opening between D1 and D2. The projections

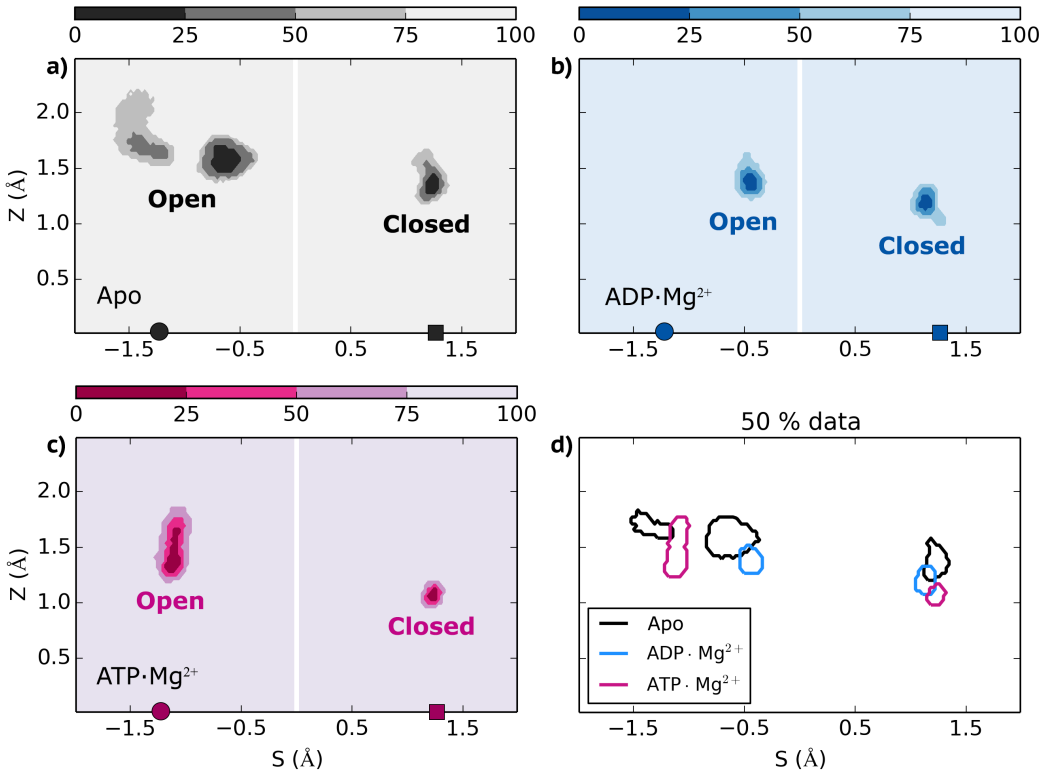


Figure 4.2: Projection of trajectories on linear path variables for apo (a), ADP (b), and ATP (c) simulations. The contours indicate a fraction of the data in the (S, Z) space for every conformation (25, 50, and 75%). Open and closed regions are well defined and the projected values do not overlap between them. Panel (d) summarizes the contour lines corresponding to 50% of the data. The S and Z values of open and closed reference structures are shown by a circle and a square, respectively

on the 6 systems indicate, that linear path variables might be a promising candidate as a CV for the enhanced sampling simulations.

4.1.2 Well-Tempered Metadynamics

To explore the free-energy landscape associated to the conformational change (*open* ↔ *closed*) in the three different intermediates considered for the translocation mechanism (apo/ATP/ADP), a series of WTmetaD simulations were performed. For this analysis, the 6 systems discussed in Chapter 3 were simulated independently and were initialized with the structures extracted at 200 ns from the plain MD. The linear path variables were selected as the set of CV's.

4.1.2.1 Computational details

Every system was constructed with the same set-up details considered for the standard MD simulations explained in Subsection (3.2.1) with a total simulation time of 600 ns for every system. The input parameters for the metadynamics were selected based on the projections of the standard MD trajectories on the linear path variables (Figure 4.2). In particular, the Gaussian widths for S and Z were set at 0.08 Å . These values were obtained as a third of the standard deviation from the trajectories projections. The initial Gaussians were 0.72 kcal/mol high, and the bias deposition stride was of 1 ps. To perform the enhanced sampling simulations, GROMACS 4.6 program and PLUMED plugin in its version 2.1 were used. For the WTmetaD formalism, an additional parameter known as *biasfactor* which is the ratio between the temperature of the CV's ($T + \Delta T$) and the temperature of the system (T) is required. The *biasfactor* selected for all simulations was equal to 15 (*i.e*, well-tempered factor ΔT = 4200 K). To avoid the exploration of other regions which are not relevant in the translocation mechanism, upper and lower walls for the linear path variables were defined in every simulation. These potential restraints acted on the system when the value of the CV's is greater (upper wall) or lower (lower wall) than a certain value a_i. The expression for the bias due to the walls (V_{wall}) is given by the following expression:

$$V_{wall} = \begin{cases} \sum_i k_i \left(\frac{x_i - a_i}{s_i} \right)^{n_i}, & \text{if } x_i > a_i \text{ (upper) or if } x_i < a_i \text{ (lower)} \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

where k_i is an energy constant, s_i is a rescaling factor and n_i is the exponent determining the power law. For the Z variable was imposed an upper wall at 5 Å , whereas, the values of S were limited between -2.5 and 2.5 Å. The value of k was equal to $k_B T$ and the exponent of the potential was set to the fourth power. Additionally, a lower wall for the RNA 3'-5' distance was set to avoid a large bending of the chain. The distance was computed between C3' in the 3' terminal and C5' in the 5' terminal nucleotides, with a minimum value equal to 14Å , using the same k and exponent selected in the other walls.

4.1.2.2 Results

For the 6 different NS3h-ssRNA complexes, a qualitative description of the WTmetaD simulations is reported below. From these enhanced sampling there is no a quantitative estimation of the free-energy, due to the lack of converge in the simulations.

Apo form

Open: During the first 300 ns, the structure fluctuates experiencing in different moments a larger gap between D1 and D2 (Figure 4.3, high Z values). Then, a significantly larger opening between D1 and D2, changes the interactions of the nucleotide in the 5' terminus with the protein producing a bending of the RNA chain. After this perturbation, the interdomain distance D1-D2 is decreased and the protein is found in the closed conformation up to 400

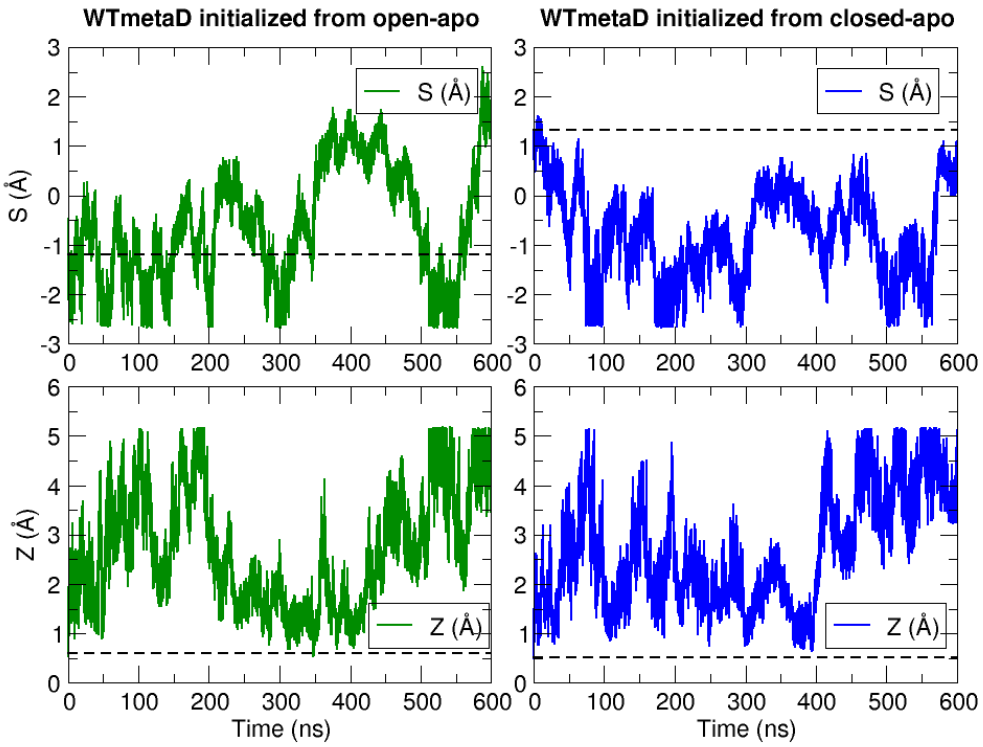


Figure 4.3: Time series for S and Z during WTmetaD, for open (green lines) and closed (blue lines) in the apo form. The black dashed line indicates the S and Z values for the initial conformation.

ns (Figures 4.3 and 4.6.a). In the following 150 ns, the structure fluctuates towards to open conformation. During the last 50 ns of WTmetaD simulation, the nucleotide in the 3′ terminus transiently moves perpendicular to the plane of the RNA chain and simultaneously the gap interdomain D1-D2 is reduced. Then, RNA is reincorporated shifting by one nucleotide the interactions with the protein.

Closed: In the first 100 ns of simulation, NS3h experiences a large opening between D1 and D2, that leads a conformational change to the open structure. In the following 470 ns the system fluctuates around the open structure and explores also regions around $S = 0$ (Figures 4.3 and 4.6.b). During the last 30 ns, the interdomain distance D1-D2 tends to decrease but the closed conformation is not recovered.

ADP FORM

Open: In the first 200 ns, NS3h fluctuates around the open conformation experiencing in different moments a larger interdomain distance D1-D2 (Figure 4.4, high Z values). Then, the structure exhibits a conformational change towards the closed structure and remains in its minima for 70 ns (Figures 4.4 and 4.6.c, low S and Z). After that, the uracil in the 3′ terminus moves perpendicular to the RNA chain, and the gap D1-D2 fluctuates towards the opening (S around 0). In the last 50 ns, the original stacking interaction between W501 and one uracil (labeled U7, Figure 3.7) is replaced by the consecutive base in the 5′→3′ direction (labeled U8, 5′ terminal nucleotide) shifting by one nucleotide all interactions protein-RNA, with a gap between D1 and D2 significantly larger compared with the original one at the beginning of the simulation.

Closed: During the first 100 ns of simulation, the helicase experiences a conformational change towards the open structure, and this conformation is maintained until 140 ns. After that, D2 fluctuates far away from D1, D3 and the RNA chain, and at 200 ns of simulation the closed conformation is recovered (Figure 4.4). Then, at 270 ns the structure suffers several fluctuations in the RNA and the interdomain distance D1-D2, and after 350 ns, the RNA chain moves away from D3 for the following 50 ns. Finally, in the last 200 ns the RNA is reincorporated with the NS3 in the open conformation and the stacking

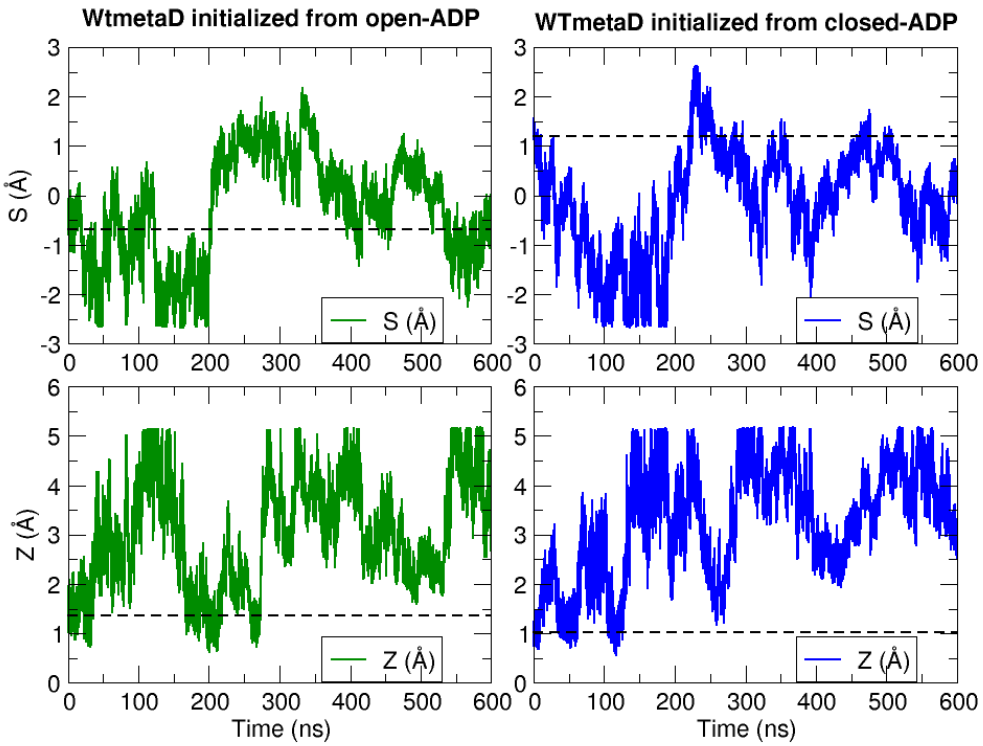


Figure 4.4: Time series for S and Z during WTmetaD, for open (green lines) and closed (blue lines) in the ADP form. The black dashed line indicates the S and Z values for the initial conformation.

interaction with W501 initially formed with U7 is replaced by U8, shifting by one nucleotide all interactions protein-RNA.

ATP form

Open: In the first 240 ns of simulation, NS3h remains in the open conformation, and the structure fluctuates to larger interdomain distances between D1 and D2 (Figure 4.5, high Z values). After that, the 3′ terminus moves away from D3 during 70 ns, and then, the RNA chain moves forward by one nucleotide in the 5′→3′ direction, shifting as well the interactions with the protein. Then, the protein experiences a conformational change towards the closed structure and the RNA chain is shifted backwards until 480 ns. In the last 120 ns of simulation, the RNA chain exhibits a bending directed far away from the protein domains. This perturbation causes that RNA is partially expelled from the NS3h.

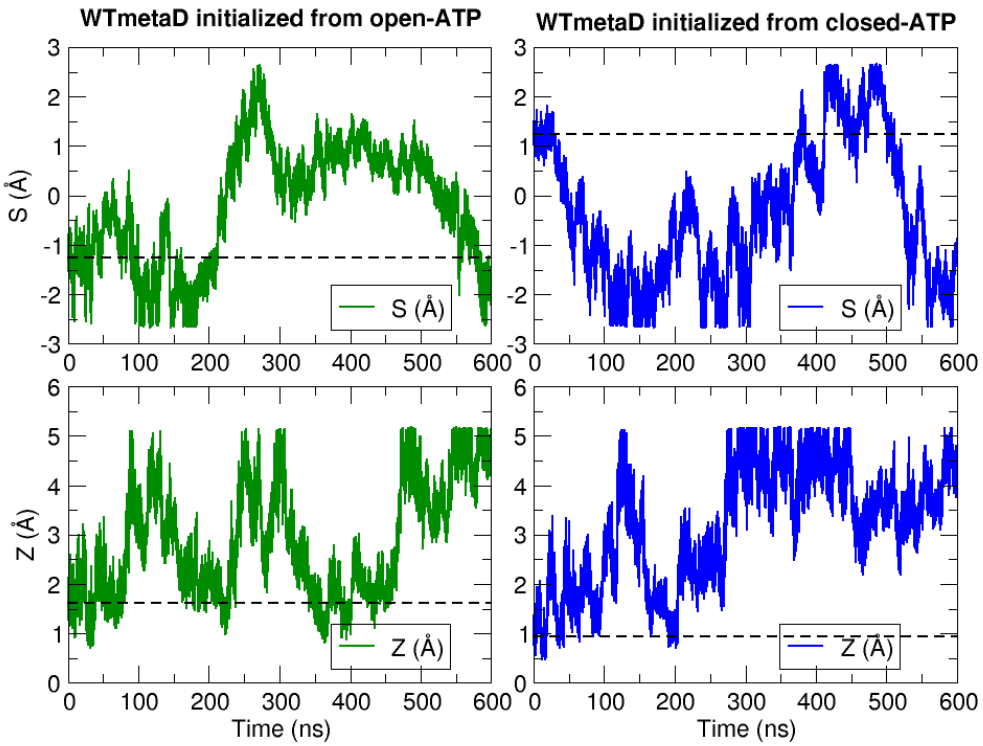


FIGURE 4.5: Time series for S and Z during WTmetaD, for open (green lines) and closed (blue lines) in the ATP form. The black dashed line indicates the S and Z values for the initial conformation.

Closed: The protein undergoes a conformational change to the open structure in the first 150 ns of simulation (Figure 4.5). This transition reveals the same pattern discussed before for the other WTmetaD simulations, where first a significantly larger opening between D1 and D2 happens to allow the reorganization of the RNA and then, the protein reaches the open minima. This conformation is maintained during the following 80 ns. For the rest of simulation, D2 moves far away from D1, D3 and RNA. This perturbation leaves specially free the RNA which experiences different bending patterns of the chain, and is gradually expelled from the complex with the helicase.

The simulations initialized from the open-apo, open/closed-ADP and open-ATP during the backward conformational change to the open structure experienced a shifting of one nucleotide between the protein-RNA interactions. These changes on the RNA contacts are very important since they are related with every cycle of

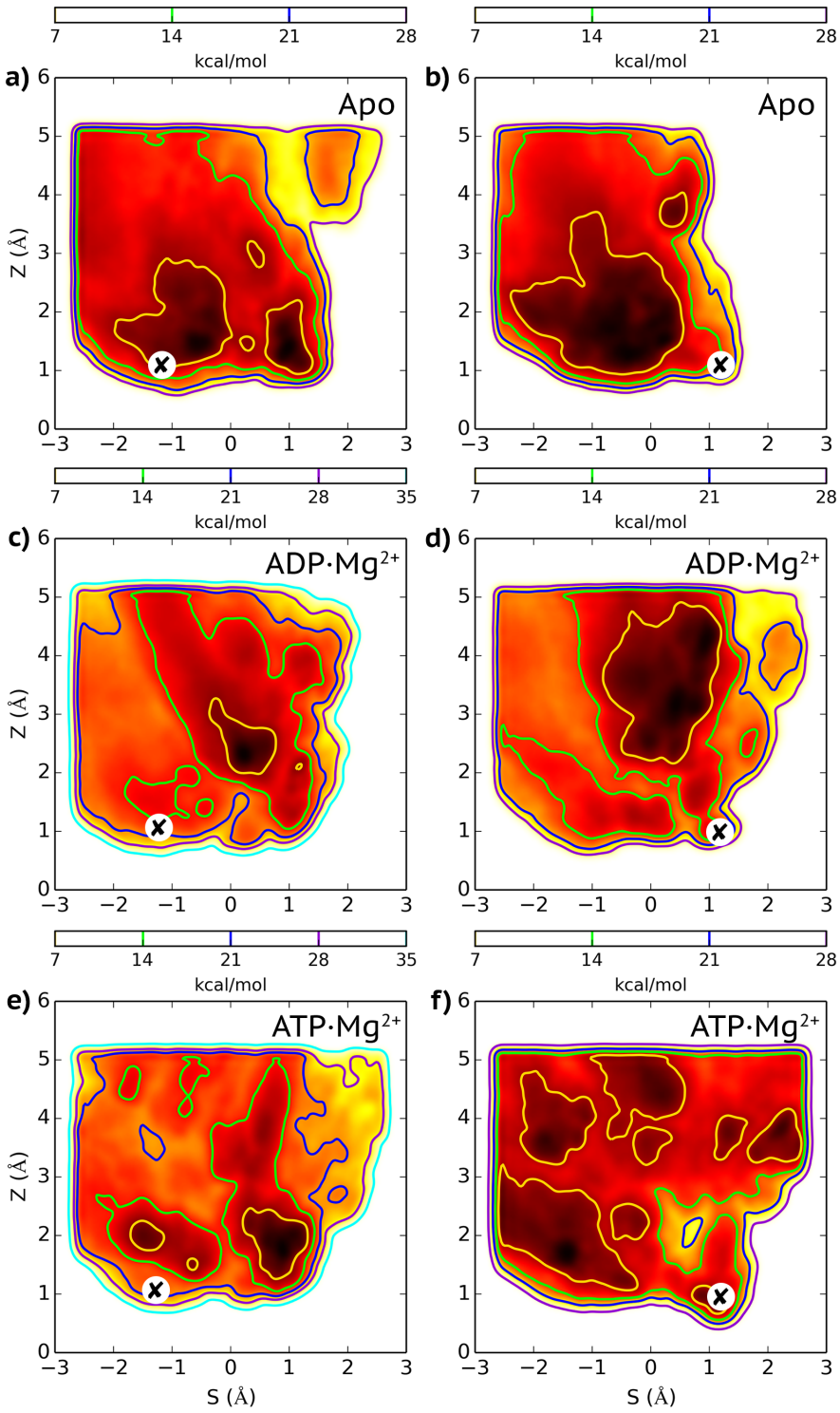


Figure 4.6: Free-energy as a function of the linear path variables. The symbol "✘" illustrates the starting structure of the simulation in the (S, Z) space. Panels (a), (c) and (e) are for the initially open conformations, while panels (b), (d) and (f) for the initially closed ones.

the translocation. However, all these transitions were observed with the opposite directionality (5′→3′ instead of 3′→5′) of the NS3h. The resulting structures from these transitions indicate that probably a lack of additional bases in the 3′ and 5′ terminus is one of the reasons related with this inconsistency. It is important to take into account that all the simulations were based on the crystallographic structures reported in reference [22]. These X-ray experiments were conducted with a polyuracil of 8 nucleotides (nt) [22], and due to a poor resolution of the first two nucleotides (labeled U1 and U2), it was not possible to properly solve them and their coordinates were not reported in the PDB. The importance of the ssNA length was also supported by previous experimental assays done on NS3-ssDNA by means of fluorimetric titration [116], where the ssDNA chains of 8 and 10 nt presented a higher binding rate (k_{on}) to the monomeric NS3 compared with the case of 6 nt. Additionally, in the open/closed-ATP simulations a disruption of the NS3h-RNA complex was observed, due to the reorganization of the NA. The final structures are very different compared with the original ones and it is not possible to undergo the translocation mechanism due to the partial expulsion of the RNA from the protein domains (see for example, open-ATP case in Figure 4.7). This is observed after several domain fluctuations, where the short RNA was more free to form new intramolecular interactions that simultaneously led to a poorer binding with the protein.

From the free-energy surfaces it is observed a energy barrier around $S \approx 0$ and low Z values (Figure 4.6). In general, the minima associated to the open conformations present a wider free-energy well respect to the closed ones, except for the cases in presence of ADP where neither the closed nor the open were very well defined and the lowest free-energy region is observed for some intermediate structures at higher Z values (above 2.5 Å). In addition, for the ATP forms, the systems explored regions at larger Z, that corresponds to the protein fluctuations and the partial expulsion of the RNA chain. All these observations reflect the necessity to model a NS3h complex with a longer ssRNA. However, there is no an available experimental structure that can provide a longer chain. To solve this limitation, the building of a longer RNA chain and the modeling for this specific application was done. The protocol implemented and the further enhanced sampling simulations are discussed with more details in the following Section.

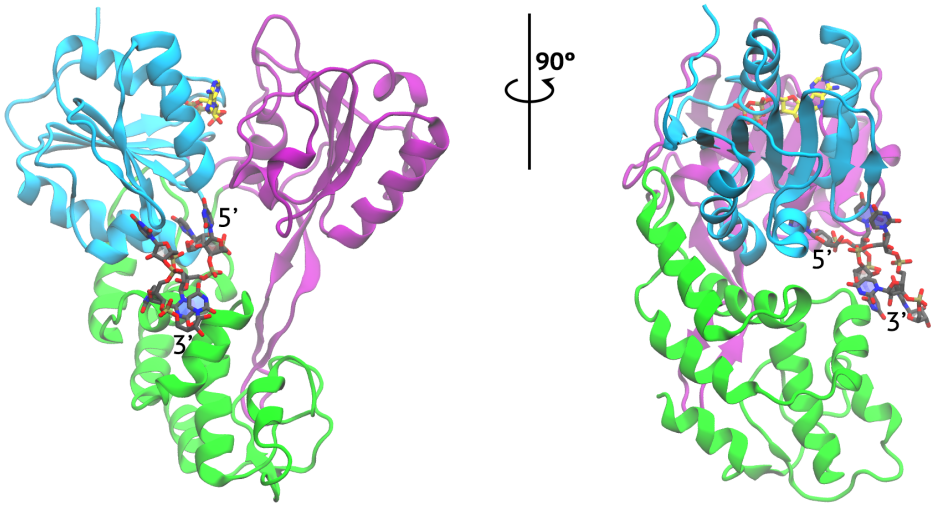


Figure 4.7: Representation of open-ATP complex after 600 ns of WTmetaD. From this structure is observed a partial expulsion of the ssRNA.

4.2 Modeling NS3h in complex with 10 nt ssRNA

Due to the lack of a crystal structure for the NS3h in complex with a longer ssRNA, The inclusion of additional nucleotides to the polyuracil chain was carried out considering a wide database collection for different RNA structures. This procedure was exclusively done in the system closed-ATP, in order to use a snapshot that was previously crystallized.

4.2.1 Building a longer RNA chain

To construct a longer polyuracil chain, a protocol based on structural alignments able to provide additional nucleotides bonded to the original polyuracil chain was implemented. A total of 1850 conformations of tri-uracil units UUU were considered for this procedure. These conformations were extracted from 342 different RNA molecules that presented the UUU triplet in different regions of the sequence. The structural alignments were done in the 3′ and 5′ terminus of the RNA using the PDB coordinates (*i.e* without protons).

The protocol implemented is described in the following steps:

1. **Addition of new nucleotides on the 3′ terminus:** Structural alignments between the sugar and phosphate atoms of the nucleotide in the 3′ terminus (labeled U8) with the UUU triplets were performed. For the structural alignment, the *RMSD calculator tool* of VMD program was used. After this procedure, 1850 systems of the NS3h in complex with a ssRNA of 8 nt were generated, placing the new uracils in the 3′ terminus.

2. **Control of overlapping atoms:** With the aim of characterizing the suitability of the new 1850 systems, a control of the atom clashes between the original system and the new uracils (labeled U9 and U10) atoms was done. For this purpose, the coordination number of the new uracils with the NS3h-ssRNA(6 nt, U3-U8) was calculated using the PLUMED 2.1 plugin. The coordination number (*CN*) is defined by discrete function as:

$$CN = \sum_{i \in A} \sum_{j \in B} CN_{ij} \tag{4.4}$$

where CN_{ij} equal to 1 correspond to the case when a contact between the atoms i and j is formed, and CN_{ij} equal to 0 when it is not formed. However, in the PLUMED implementation, the CN_{ij} can be expressed as a continuos parameter, defined by the following switching function:

$$CN_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^m} \tag{4.5}$$

where n=6 and m=12, d_0=0 Å and r_0=1 Å were set for this particular case. From these parameters, the *CN* with the new uracils was computed counting the atoms of the original system that were within 1 Å.

3. **Selecting suitable candidates:** The systems with a coordination number below 0.1 were selected as candidate structures. With this criteria, the total number of candidate structures with a RNA chain long 8 nt (from U3 to U10) was 12.

4. **Addition of new nucleotides on the 5′ terminus:** Following the same procedure described in the items (1), (2) and (3) for the 5′ terminus, a total of 4 candidates systems with a RNA chain long 8 nt (from U1 to U8) were selected.

5. **Building of 10 nt systems:** From the 4 candidates structures with a longer 5′ terminus, the nucleotides U1 and U2 (the last 2 nt in the 5′ terminus) were combined with the 12 NS3h candidates constructed with a longer 3′ terminus; for a total of 48 putative structures of NS3h-ssRNA(10 nt)-ATP.

6. **Selection of 10 nt system:** The 48 structures were protonated keeping the same number of atoms and subsequently, an energy minimization for every system was performed. This procedure was executed for 2000 steps until the maximum force was smaller than 0.24 kca/(mol Å) using GROMACS 4.6 program. The structure with the lowest potential energy was selected as the final system for the further modeling.

4.2.2 MD production

The selected system was constructed with the same set-up details considered for the standard MD explained in Subsection (3.2.1), and following a similar simulation protocol described in Subsection (3.2.2). Additionally, before the enhanced sampling simulations, a test for the best initial conformation was done, building and simulating 4 different set-ups with a rescaled Hamiltonian for the new nucleotides in the 3′ and 5′ terminus. This is done to control how the flexibility of the new nucleotides will affect the interactions with the helicase. For this purpose, the following procedure was done:

1. **Preparation of the systems:** 4 different set-ups were prepared, rescaling the force field terms for the new nucleotides (U1, U2 in the 5′ terminus and U9, U10 in the 3′ terminus) based on the HREX procedure described in Chapter 2. In this particular application, 4 different scaling factor ($\lambda = 1, 0.9, 0.8, 0.7$) were used, being the case of $\lambda = 1$ the system with the original FF.

2. **Standard MD simulations:** For every system, a plain MD simulation of 50 ns was performed. All trajectories were analyzed, and the conformations of the U3 and U8 during the simulations were compared with the original ones from the crystal.

3. **Selection of the initial guess:** The system that kept a similar conformation of U3 and U8 to the crystal, was chosen as the initial guess. In this specific case, the structure scaled by a factor λ=0.9 was selected. Subsequently, a plain MD

simulation long 200 ns with the original FF was performed as an equilibration stage and the final frame is taken to initialize the enhanced sampling.

4.2.3 WTmetaD using linear path variables

With the aim of exploring the free-energy landscape associated to the conformational change *closed* ↔ *open* for the ATP system, a WTmetaD simulation was performed. For this initial analysis, the linear path variables were selected as the set of CV's.

4.2.3.1 Computational details

In order to perform the WTmetaD simulation on the NS3-ssRNA(10 nt)-ATP complex, the same input parameters used for the systems with 6 nt in Subsection (4.1.2.1) were selected. To perform the enhanced sampling simulations, GROMACS 4.6 program and PLUMED 2.1 plugin were used. The total simulation time for this application was of 850 ns.

4.2.3.2 Results

For the NS3h-ssRNA(10 nt)-ATP complex, a qualitative description of the WTmetaD simulation is reported. From these enhanced sampling there is no a quantitative estimation of the free-energy, due to the lack of converge in the simulations.

Analysis of the trajectory:

From the WTmetaD trajectory, it is observed that after 20 ns of simulation the NS3h experienced a conformational change towards the open structure, presenting some fluctuations of the interdomain distance between D1 and D2 (Figure 4.8, high Z values) and changing the interactions of the RNA backbone with the D2 of the protein. After 580 ns, the basin for the open minima is explored while in the last 170 ns the interdomain distance tend to slightly decrease (Figure 4.8, $S \approx 0$) without recovering the initial closed conformation of simulation, and the U1 is placed inside the RNA-D3 interface forming a stacking interaction with the U3. This interaction between U1-U3 is undesirable, since it might affect the progress of the translocation mechanism. The ATP binding pocket is also analyzed observing some fluctuations of the adenine ring from the ATP, disrupting the stacking interaction with the Y241 in D1 in the different moments of the simulation. Moreover, during the simulation the Mg^{2+} is coordinated to S211 (Walker A), E291 (DExH box), two water molecules

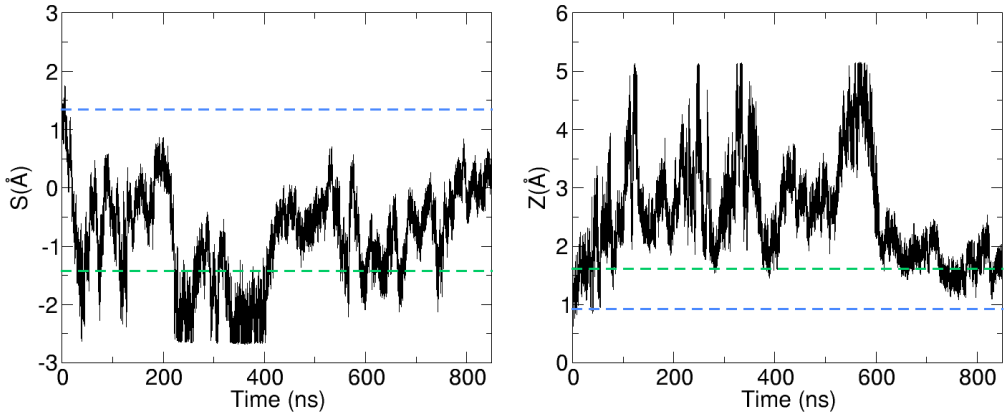


Figure 4.8: Time series of the linear path variables during the WTmetaD trajectory. The dashed lines indicates the S and Z values for the closed-ATP (blue line) and open-ATP (green line) conformations.

and the oxygens from the Piβ and Piγ. This coordination complex is maintained during the entire simulation and is the same identified from the crystal.

Free-energy surface: a qualitative interpretation

From the free-energy surface (Figure 4.9), it can be appreciated that the minima associated to the open conformation present a wider free-energy well respect to the closed one. However, it is not possible to provide an accurate description of the two free-energy minima due to the lack of converge in the enhanced sampling. During the exploration of the free-energy landscape, it is also observed an intermediate structure at higher Z values (Figure 4.9, $Z \approx 3$ Å), that corresponds to the conformation observed between 400-500 ns of simulation. This structure is formed presenting a slightly shorter gap D1-D2 compared with the open-ATP ($S \approx$ -0.5).

4.2.4 WTmetaD using path collective variables

After 850 ns of WTmetaD using the linear path variables as the set of CV's, the lack of converge of the simulation becomes a limiting factor to describe properly the free-energy surface of the NS3h-ssRNA(10 nt)-ATP conformational change. Additionally, due to the large values of the Z value, this allows the system to visit regions that are not involved in the conformational transition. For this reason,

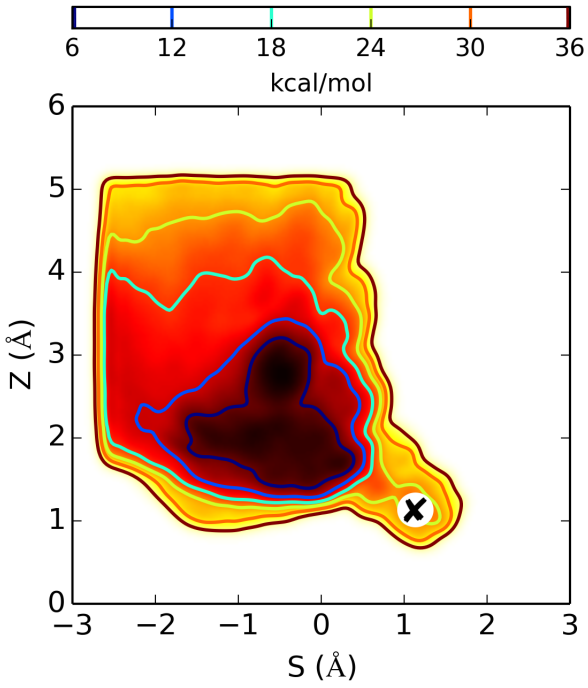


FIGURE 4.9: Free-energy as a function of the linear path variables. The symbol "✘" indicates the starting structure (closed-ATP) of the simulation in the (S, Z) space.

the initial set of CV's is changed, substituting the linear path variables by path collective variables [69], with the aim of observing a faster and better sampling of the free-energy surface.

4.2.4.1 Computational details

As discussed in Subsection (2.3.3), the path collective variables are very useful for the study of transitions between two states. In this specific case, the two states of interest are the closed and open conformations of NS3h in presence of ATP. With the aim of selecting an appropriate reactive path, the transition between the closed and open conformation of the enzyme was described using the path extracted from the TMD simulations of closed-ATP discussed in Section (3.3), selecting a frame every 2 ns for a total of 11 reference structures $\widetilde{q}(i)$. In this particular application, the distances between the simulation frame and the reference snapshots were computed as the RMSD of the subset of atoms discussed in Subsection (3.2.3.1) and the λ parameter was set to 1 $Å^{-2}$. The WTmetaD was performed only in the space of

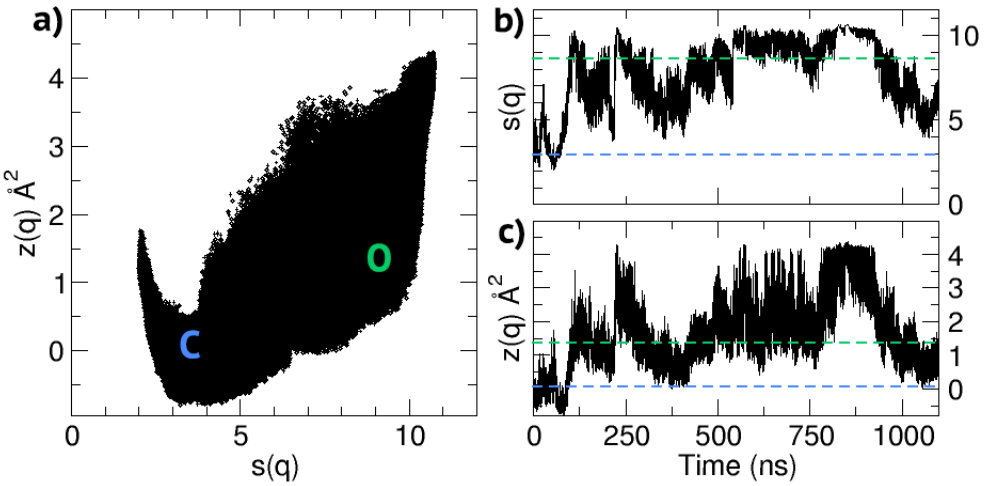


Figure 4.10: Projection of the WTmetaD trajectory on the path collective variables (Panel a) and their respective time series (Panels b and c). The dashed lines indicates the S and Z values for the closed-ATP (blue line) and open-ATP (green line) conformations and the "C" and "O" letters denote the closed and open basins respectively. In the

$s(q)$ using a Gaussian width of 0.8 Å, an initial Gaussian deposition rate of 0.72 kcal/mol per picosecond and a well-tempered factor ΔT of 4200 K. Additionally, the $z(q)$ variable was constrained using an upper wall at 4 Å^2 with a constant k equal to $k_B T$ and the exponent of the potential set to the second power. This choice avoids the possibility to the system of exploring other regions in the free-energy landscape that are not relevant for the conformational transition. To perform the enhanced sampling simulations, GROMACS 4.6 program and PLUMED 2.1 plugin were used. The total simulation time was of 1.1μs.

4.2.4.2 Results

The analysis for the WTmetaD using path collective variable to model NS3h-ssRNA(10 nt)-ATP conformational change, are reported as preliminary results, since it is necessary to prolong the simulation time to reach convergence.

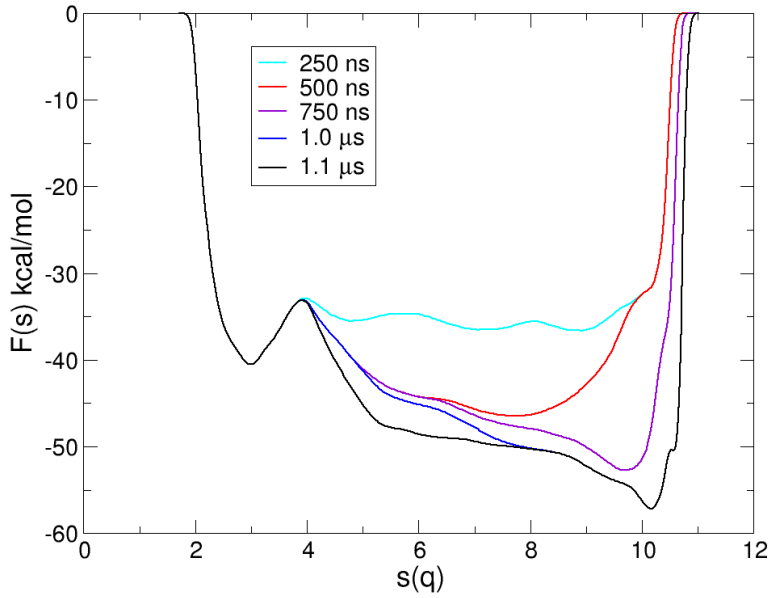


FIGURE 4.11: Free-energy as a function of the path collective variable $s(q)$

Analysis of the trajectory:

After 100 ns of simulation the helicase experienced a conformational change to the open structure (Figure 4.10) which is maintained up to 750 ns. Then, during the last 350 ns of simulation, the interdomain distance tends to decrease without recovering the closed conformation and the RNA presented some fluctuations causing a slightly bending of the chain that allows the formation of a hydrogen bond between a phosphate oxygen of U5 with the side chain of T416. Additionally analyzing the intramolecular interactions of the RNA, the stacking between U1-U3 previously detected in the WTmetaD simulation using the linear path variables, is not formed in this case. Analyzing the ATP binding pocket, it is observed that the stacking interaction between Y241 from D1 and the adenine ring from the ATP is maintained during the entire simulation, and the coordination complex formed with Mg^{2+} cation is identical to the one identified from the crystal.

Free-energy surface: preliminary results

The free-energy surface (Figure 4.11), indicates an energy barrier around $s(q) \approx 4$. Additionally, from this preliminary energy profile it is observed that the minima

associated to the open conformation present a wider free-energy well compared with the closed one, which is a feature also recognized from the previously discussed WTmetaD using the linear path variables as CV's. However, an accurate description of the two states is still imprecise since the sampling needs a longer simulation time for the correct estimation of the well shape and the relative free-energy values.

4.2.5 WTmetaD coupled with HREX simulations

As a final approach, a WTmetaD simulation combined with HREX for the NS3h-ssRNA(10 nt)-ATP system was performed, using the path collective variables as the set of CV's. This strategy is chosen, since metadynamics is a technique that accelerates the sampling of rare events overcoming the energy barriers and HREX is able to reproduce correctly the free-energy difference between two metastable states and the shape of the free-energy wells.

4.2.5.1 Computational details

For the WTmetaD, the same set-up details used in the Subsection (4.2.4) were selected in this application. A total of 16 replicas were constructed rescaling some FF terms of a specific region of the solute, named the *hot* ($\mathcal{H}$) region, and following the criteria explained in Section (2.4). The $\mathcal{H}$ region comprises the atoms discussed in Subsection (3.2.3.1) but instead of considering only the α carbons from the selected protein residues and some nodes from the ssRNA, the entire amino acid residues and all 10 nucleotides from the new RNA chain are taken. The λ_{HREX} rescaling factor was selected according to a geometric distribution between 1 and 0.5, which is equivalent to an effective temperature that goes from 300 to 600 K, even if all replicas were simulated at the same canonical temperature of 300 K. The exchange rate between replicas was set every 0.4 ps. The enhanced sampling simulations were performed using GROMACS 4.6 program and PLUMED plugin in its version 2.1. The simulation time per replica was of 200 ns, for a total time of 3.2μs.

4.2.5.2 Results

Replica exchange MD, is an expensive computational technique since the simulation cost scales proportionally with the number of replicas that are chosen for the sampling. The choice of 16 replicas scaled up to $\lambda_{HREX} = 0.5$, presented an average acceptance for every exchange attempt of ~6%. This allows that the energy barriers

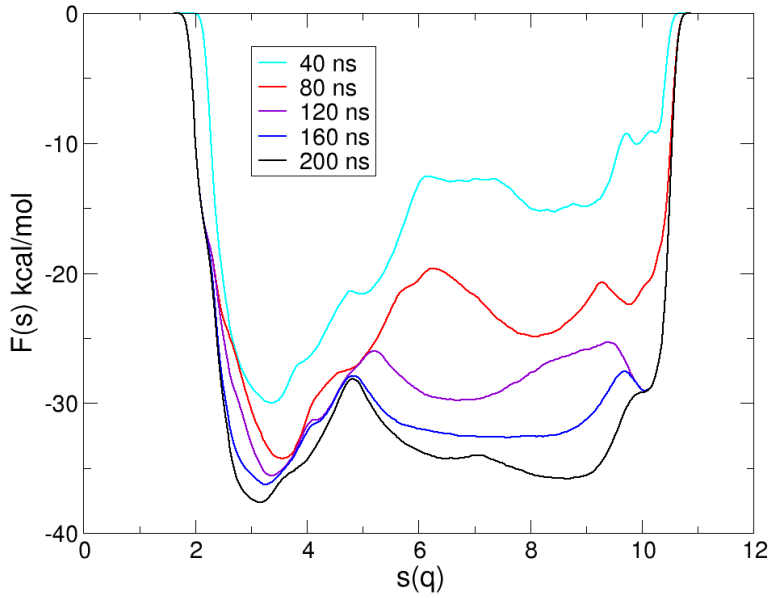


FIGURE 4.12: Free-energy as a function of the path collective variable $s(q)$ for the replica with the original Hamiltonian ($\lambda = 1$).

can be overcome easily and the replicas experience several effective exchanges during the simulation. The choice of number of replicas is crucial for an appropriate sampling, since a too small number of replicas with a large Hamiltonian rescaling (or high temperatures, for the case of parallel tempering [74]) will imply a significantly low acceptance rate in every exchange attempt, while a very large number of replicas will presuppose a very expensive simulation that in practical terms might be inconvenient. Although the average acceptance rate for this particular case is relatively low (~6%), from the simulation several roundtrips between the 16 replicas are noticed.

From the free-energy surface (Figure 4.12) obtained for the replica with the original Hamiltonian, it is observed that an energy barrier around $s(q) \approx 5$ must be overcome to undergo the conformational transition closed-ATP $\leftrightarrow$ open-ATP. This barrier is slightly shifted in the space of $s(q)$ respect to the preliminary free-energy profile obtained in the WTmetaD without HREX ($s(q) \approx 4$). In addition, this graph also indicates that the simulation have not reached convergence since the bias potential deposited in the different moments of the simulation is larger for the open

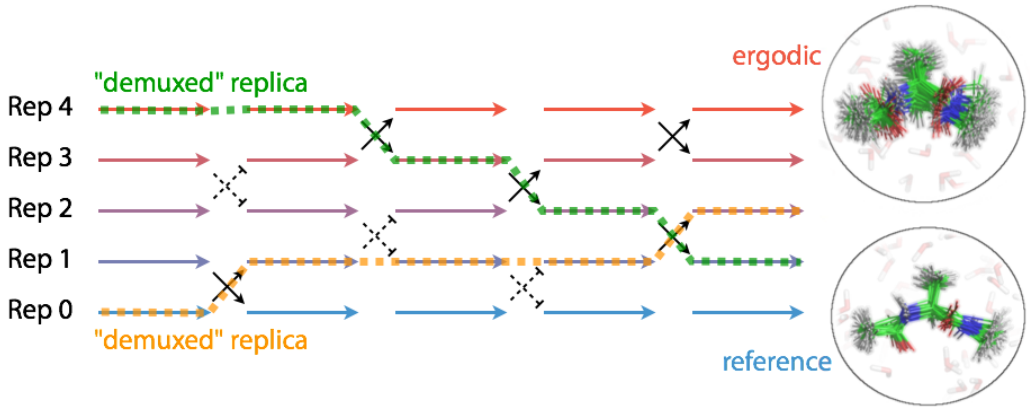


Figure 4.13: Graphical representation of the *demux* procedure implemented in GROMACS 4.6. Courtesy of Giovanni Bussi.

minima compared with the closed one. From this preliminary free-energy profile, it is observed that the open conformation present a wider free-energy well respect to the closed one, which is a common feature observed on all enhanced sampling simulations discussed in this Chapter.

In every exchange attempt, GROMACS 4.6 code implementation exchange the atomic coordinates between replicas, however, this feature generates trajectories that are not continuos, this means that after every successful exchange, a structure completely different would appear in the trajectory file. For this reason, a procedure that generates the correct time series of every replica is applied in order to analyze the trajectory with continuos coordinates. This procedure in the GROMACS 4.6 code is known as *demux*. A graphical representation that illustrates the *demux* is shown in Figure 4.13. After *demuxing* the HREX trajectories, a projection on the path collective variables is done for every replica (Figure 4.14). This is done to control how the replicas are evolving in the space of the path collective variables in order to verify the recrossing between the two conformational states. From these plots, it is observed that 8 (Reps. 0, 1, 2, 6, 10, 12, 13, 15) of the 16 replicas experienced a transition to the open conformation basin, while the other 8 are still in the closed minima or are exploring regions at higher $z(q)$ values. This projection as well as the previous analysis done in this simulation indicates the necessity to perform a longer simulation with the aim of estimating properly the relative free-energy

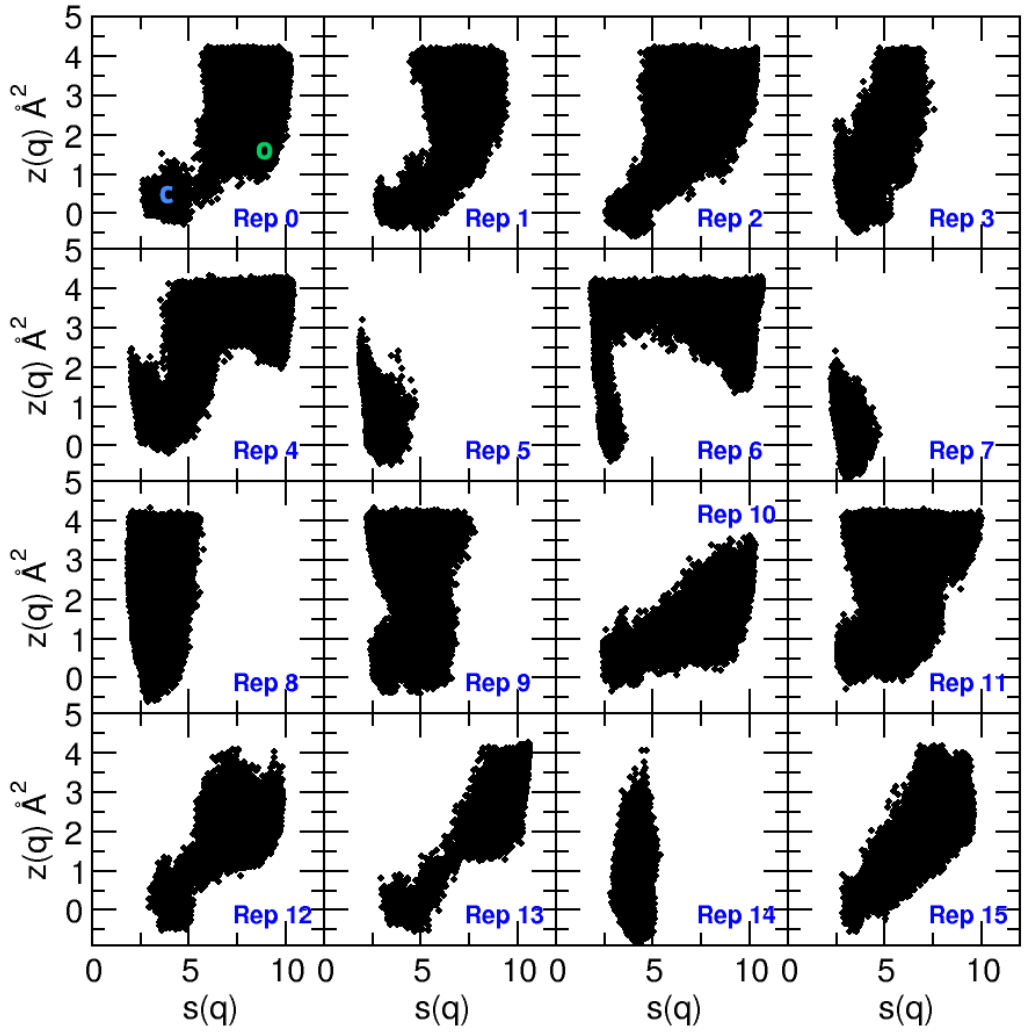


Figure 4.14: Projection of the path collective variables $s(q)$ and $z(q)$ on the *demuxed* trajectories for all 16 replicas. "C" and "O" letters denote the location in the CV space of the closed-ATP and open-ATP basins respectively.

between the NS3h conformers in presence of ATP. Additionally, analyzing the λ_{HREX} graph (Figure 4.15) for the *demuxed* replicas, it is observed that some replicas are not experiencing many exchanges. This means that not all of them are diffusing effectively in λ, indicating that it is necessary to prolong the simulation in order to fulfill this condition and estimate correctly the free-energy surface.

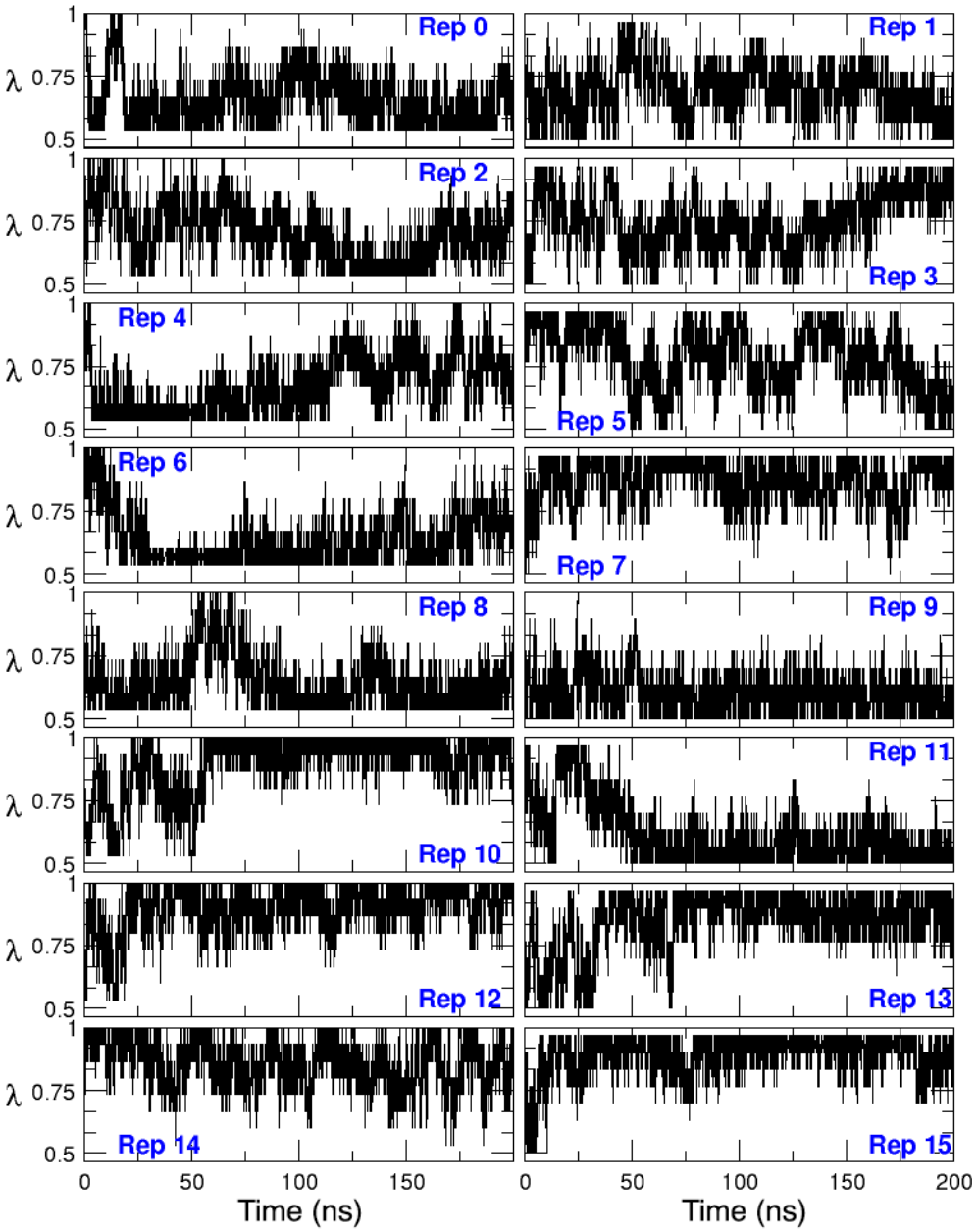


Figure 4.15: λ_{HREX} graph for the 16 *demuxed* replicas.

4.3 Discussion

From the WTmetaD simulations using linear path variables as CV on the NS3h structures in complex with a polyuracil chain of 6 nucleotides, the systems presented

different structural deviations presumably related to the length of the modeled RNA that directly interfere with the reversible conformational change and the subsequent translocation mechanism along the RNA chain. In particular, the partial expulsion of the ssRNA from the complex with the protein domains and the wrong directionality observed in four of the six simulated systems were the main indicatives that reflected the necessity to model a longer RNA chain with the aim of understanding this specific structural transition.

In this context, the building of a longer ssRNA chain based on structural alignments for the closed-ATP system was carried out, leading to a putative structure that contained a polyuracil unit of 10 nucleotides. Subsequently, a WTmetaD simulation on this new system was performed, experiencing a transition to the open conformation and indicating that the choice of linear path variables as the set of CV's might not be suitable for the modeling of the conformational transition. The limitations of the linear path variables, are mainly related with the boundaries for the Z values. In general, the path that connects both minima of NS3h conformers is curved in the space of the linear path variables. This implies that a restrained range for Z values would lead to a trajectory that is thermodynamically unfavored. However, a larger range for Z values will allow a wider exploration of regions on the free-energy landscape that are not relevant for this specific mechanism. For this reason, the implementation of an advanced set of CV's such as the path collective variables [69] was done with the aim of modeling this complex conformational transition that involves several changes for the contacts between protein and RNA residues. Consequently, a WTmetaD using path collective variables as CV was performed, presenting a conformational transition to the open structure without experiencing a recrossing towards the initial closed one. From this simulation, there is no a quantitative estimation of the relative free-energy values since it requires a longer time to reach convergence. In this regard, an additional sampling that combines WTmetaD with HREX simulations was performed, getting similar results to the previous case but showing some differences in the energy barrier that must be overcome for the conformational change of the ATP forms. The advantage of implementing WTmetaD-HREX is that it allows a more exhaustive exploration of the free-energy landscape but at the same time maintains the system closer to the equilibrium.

Total simulation time

Table 4.1: Total simulation time for the calculations reported in Chapter 4

System	# runs	time/run (μs)	total time (μs)
NS3h-ssRNA (6 nt)			
WTmetaD **CV:** linear path			
open-apo	1	0.6	0.6
closed-apo	1	0.6	0.6
open-ADP	1	0.6	0.6
closed-ADP	1	0.6	0.6
open-ATP	1	0.6	0.6
closed-ATP	1	0.6	0.6
Subtotal time			3.6
NS3h-ssRNA (10 nt)			
WTmetaD **CV:** linear path			
closed-ATP	1	0.85	0.85
WTmetaD **CV:** path collective			
closed-ATP	1	1.1	1.1
WTmetaD/HREX **CV:** path collective			
closed-ATP	1	0.2×16 reps.	3.2
Subotal time			5.15
Total time:			**8,75**

CHAPTER 5

Conclusions and Perspectives

5.1 Conclusions

A novel approach to describe the HCV NS3h translocation along ssRNA based on computer simulations was presented. From this work, the characterization of the equilibrium properties for the helicase conformers in complex with ssRNA in the apo, ADP and ATP forms, as well as a detailed analysis of the conformational transitions that are involved in the translocation mechanism were carefully described.

- **Insights from plain MD simulations.** As an initial approach of this study, the three different NS3h molecular forms (apo/ADP/ATP) in complex with a short ssRNA chain of 6 nucleotides in the open and closed conformations were modeled by means of standard MD simulations at the microsecond time scale. Only two of these simulations were initialized with an available experimental structure, while the additional four systems were constructed based on structural alignments. NS3h has been experimentally isolated in the apo[a] and ATP[b] forms [22, 21, 20], however, in this study the intermediates after the ATP hydrolysis reaction were included in order to analyze the effect of ADP in the NS3h complex. This consideration is consistent with the fact that an X-ray intermediate of the NS3 in complex with ADP has been reported

[a] In the open conformation
[b] In the closed conformation

for Dengue virus [77]. During the MD simulations, all experimental and artificial structures were stable within the microsecond time scale, suggesting that the AMBER force fields [80, 81, 82, 83, 84, 85, 86, 87] employed in this application might be suitable for the study of protein-RNA complexes. In general, the average values obtained for the different parameters analyzed in the simulations presented small statistical errors, suggesting that MD simulations were able to provide a converged conformational ensemble around the equilibrium structures. Additionally, the analysis of the contact network between solute molecules, as well as the electrostatic interactions within the Debye-Hückel approximation [100] and the TMD simulations [57] suggested an ATP-dependent stabilization for the closed conformation, which is consistent with the fact that it is one of the snapshots experimentally crystallized, but contradictory to the enthalpy values calculated between the two helicase conformers in presence of the ATP molecule, suggesting an Enthalpy-Entropy Compensation effect [106] for the closed conformation. On the contrary, the structural analysis done on the ADP complexes indicated a larger stabilizing effect in the open conformation, which is also supported by the electrostatic interaction and the enthalpy differences, suggesting that after the ATP hydrolysis reaction the NS3h might be found in the open conformation. All the exhaustive analysis discussed in this initial approach, has been also reported in a recent publication [109].

- **Enhanced sampling of the conformational transition.** Based on these results, the second part of this study was dedicated to describe the conformational changes of the NS3h by means of enhanced sampling techniques such as metadynamics and HREX, in order to provide the mechanistic features of the transition and estimate the relative free-energy values able to verify a possible Enthalpy-Entropy Compensation effect for the ATP intermediates. Six independent WTmetaD simulations initialized with the six different NS3h structures[c], evidenced the necessity to model the translocation with a longer RNA chain, due to the radical structural deviations observed in the ATP forms that impede the progress of the mechanism and the observation of the wrong helicase directionality in four of the six simulations. For this reason,

[c]Open and closed NS3h conformations in complex with ssRNA of 6 nucleotides in the apo, ADP and ATP forms.

the building of longer ssRNA based on a series of structural alignments was carried out. This procedure was exclusively done on the closed-ATP intermediate, providing a putative structure that was modeled by means of WTmetaD using the same setup implemented in the system of 6 nucleotides, but without reaching convergence and reflecting some limitations of the initially selected reaction coordinates (CV's). Metadynamics is a very robust technique that accelerates the sampling of rare events disfavoring the already visited states [114], however, the simulations sometimes face two inconvenients: (*i*) it is difficult to obtain a converged result and (*ii*) it is difficult to choose appropriate reaction coordinates. Both difficulties were encountered in this specific application, for this reason, the implementation in a further modeling of an advanced CV such as the path collective variables [69] was necessary in order to describe properly the complex conformational change of this helicase. From this simulations, it was observed a conformational transition to the open structure, without exhibiting large structural deviations in the protein domains and the RNA chain, and forming inter- and intra-molecular interactions during this process that might not dramatically interfere with the translocation process. Nevertheless, the analysis of these simulations were reported as preliminary results since the enhanced sampling requires a longer time to reach convergence. Due to this limitation, a final approach based on the coupling of HREX simulations with WTmetaD using path collective variables was implemented. Despite this is a strategy at higher computational cost, it allows a more extensive exploration of the free-energy surface and keeps the system closer to equilibrium. From this simulation, it was observed a transition to open conformation for half of the replicas (*i.e* 8 reps.), however, the analysis was also reported as preliminary results since they need a longer time to reach convergence. This work is still in progress but will hopefully be concluded in a few months so as to be submitted to a journal for publication.

From this PhD thesis, the total simulation time including all the systems and the different theoretical approaches was approximately 18 μs (see Tables 3.9 and 4.1 for a complete summary), providing a detailed overview of the equilibrium properties and the different mechanistic features of this important enzyme. Due to the extensive analysis presented in this work, an hypothetical mechanism of the NS3h translocation along the ssRNA can be proposed (see also Figure 5.1). Thus,

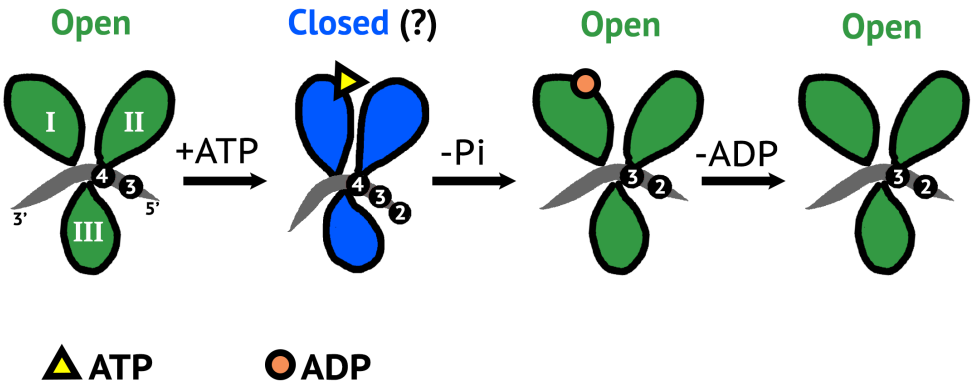


Figure 5.1: Proposed mechanism for the NS3h translocation along ssRNA. A quantitative estimation of the free-energy for the NS3h complexes in presence of ATP would confirm which conformation is thermodynamically more favorable.

the order of the events during every cycle of the mechanism should be driven in the following manner: the initially open-apo structure binds the ATP molecule inducing a conformational change to the closed conformation. Once the hydrolysis reaction is yield (producing phosphate and ADP molecules) the open conformation, this time in complex with ADP, should be recovered shifting by one nucleotide the contacts between protein and RNA in the 3′→5′ direction. Then due to the weaker[d] interaction between ADP-protein, the ADP leaves the complex leading to the open-apo structure completing in this way one translocation cycle.

5.2 Perspectives of this PhD thesis

With the aim of obtaining a quantitative estimation of the free-energy for the open and closed conformation, it is necessary to extend the simulation time of the WTmetaD/HREX using the path collective variables for the system with the RNA long 10 nucleotides. Nonetheless, some additional details must be taken into account.

Modeling the conformational change *open* ↔ *closed*, based on the available crystal snapshots, implies several changes for the contacts between D2 and the RNA

[d] compared with ATP, see also Table (3.7)

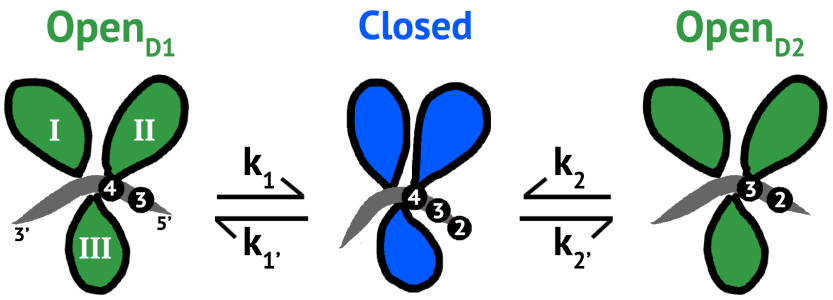


FIGURE 5.2: Cartoon representation for the conformational changes *closed* $\rightleftharpoons$ *open*. k_1 and k_2 correspond to the rate constant associated to the NS3h closing changing the contacts between D1-RNA and D2-RNA respectively.

backbone, but not for D1. However, in order to follow an *inchworm-like* mechanism during the translocation/unwinding process, it is also necessary a change of the contacts between D1 and RNA in one of the steps involved in the complete cycle. In this context, the modeling of the backward opening is crucial to understand the progress of the mechanism (Figure 5.2). For this purpose, the modeling by means of enhanced sampling using path collective variables as CV's can be applied, generating the reference frames of the path collective through TMD simulations that would have as a target the open structure with different contacts between D1 and RNA. This additional approach would be able to explain some features still unclear about the ATP forms, but it would also provide new details about the *inchworm* movement of the NS3h.

Additionally, in the last years NS3h have emerged as an alternative target to improve the current HCV therapies [19]. For this reason, the understanding of the translocation mechanism along the RNA might be able to provide the basis for the future design of a novel antiviral treatment.

Appendix A

Electrostatic interaction:

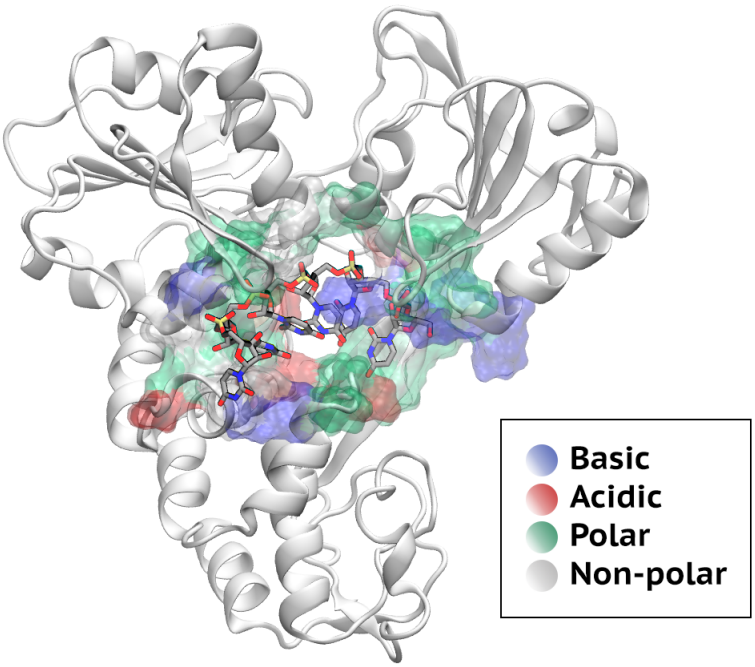


Figure 1: Protein residues within 6 Å from the RNA chain

Table 1: G^{DH} (kcal/mol) between RNA and protein residues located within 6 Å from the RNA chain. Errors are calculated from binning analysis (bin width: 80 ns) and are reported in parenthesis. Error bars lower than 0.05 kcal/mol are not shown.

Apo				**ADP·Mg^{2+}**				**ATP·Mg^{2+}**			
Closed		**Open**		**Closed**		**Open**		**Closed**		**Open**	
K371	-1.8	R393	-1.9	K371	-1.9	R393	-1.9	R393	-1.9	R393	-2.0
R393	-1.5(0.1)	K272	-1.0	R393	-1.4	K272	-1.0	K371	-1.8	K272	-1.0
K272	-1.0(0.1)	K371	-0.6(0.1)	K272	-1.0	K371	-0.8	K272	-0.9	K371	-0.9
K551	-0.4	K551	-0.5	K551	-0.4	K551	-0.4	K551	-0.4	K551	-0.5
K372	-0.4	S231	-0.2	K372	-0.4	T269	-0.2	K372	-0.4	T269	-0.2
T411	-0.2	T269	-0.2	T411	-0.2	S231	-0.2	T411	-0.2	V232	-0.2
T269	-0.2	V232	-0.2	T269	-0.2	T411	-0.2	T269	-0.2	S231	-0.2
S231	-0.2	G255	-0.2	S231	-0.2	V232	-0.2	V232	-0.2	T411	-0.2
G255	-0.2	T254	-0.2	G255	-0.2	G255	-0.2	G255	-0.2	G255	-0.2
V232	-0.2	T411	-0.1	S370	-0.2	T254	-0.1	S231	-0.2	K372	-0.2
S370	-0.2	K372	-0.1	V232	-0.2	Y392	-0.1	S370	-0.2	T254	-0.1
T254	-0.2	T416	-0.1	T254	-0.1	K372	-0.1	Y392	-0.1	Y392	-0.1
T298	-0.1	G271	-0.1	Y392	-0.1	A233	-0.1	T254	-0.1	T416	-0.1
Y392	-0.1	Y392	-0.1	H369	-0.1	N556	-0.1	H369	-0.1	A233	-0.1
G394	-0.1	A233	-0.1	T298	-0.1	G271	-0.1	Q434	-0.1	W501	-0.1
H369	-0.1	T298	-0.1	G394	-0.1	G417	-0.1	T298	-0.1	G271	-0.1
A233	-0.1	W501	-0.1	T448	-0.1	T298	-0.1	A233	-0.1	T298	-0.1
G271	-0.1	V256	-0.1	G271	-0.1	A275	-0.1	G271	-0.1	N556	-0.1
A275	-0.1	Y270	-0.1	A233	-0.1	V256	-0.1	A275	-0.1	V256	-0.1
V256	-0.1	A275	0	A275	-0.1	T416	0	N556	-0.1	A275	0
Q434	0	G417	0	V256	-0.1	Y270	0	V256	-0.1	Y270	0
Y270	0	G554	0	N556	0	G394	0	G394	0	G394	0
Y391	0	G394	0	Y270	0	L274	0	Y270	0	G417	0
L274	0	L274	0	L274	0	Q552	0	L274	0	H369	0
N556	0	N556	0	S297	0	S370	0	S297	0	L274	0
Y502	0	S297	0	Y391	0	T433	0	T448	0	S370	0
S297	0	T433	0	T450	0	T448	0	T450	0	A413	0
T448	0	S370	0	Y502	0	H369	0	Y502	0	S297	0
T450	0	T448	0	A413	0	S297	0	Q552	0	T433	0
W501	0	Y391	0	G417	0	Y502	0	T433	0	T448	0
A413	0	A497	0	Q552	0	A497	0	W501	0	T450	0
A497	0	H369	0	T416	0	A500	0	Y391	0	A497	0
A500	0	T450	0	W501	0	T450	0	A413	0	Q434	0
G417	0	Y502	0	A497	0	W501	0	A497	0	Y391	0
T416	0	A500	0	A500	0	Y391	0	A500	0	A500	0
T433	0	Q434	0	T433	0	A413	0	G417	0	Q552	0
T449	0	Q552	0	V432	0	Q434	0	V432	0	T449	0
V432	0	T449	0	T449	0	V432	0	T416	0	V432	0
Q552	0	A413	0	G554	0	G554	0	T449	0	Y502	0
G554	0	V432	0	P230	0.1	T449	0	G554	0	G554	0
P230	0.1	P230	0.1	Q434	0.1	P230	0.1	P230	0.1	P230	0
E503	0.2	E503	0.1	E503	0.2	E503	0.2	E503	0.2	E503	0.1
D412	0.5	D412	0.5	D555	0.5(0.1)	E493	0.5	D555	0.4	D555	0.4
D296	0.6	E493	0.6	D412	0.6	D412	0.5	D412	0.5	E493	0.5
E493	0.6	D555	0.7	D296	0.6	D296	0.6	D296	0.6	D412	0.5
D555	0.8(0.1)	D296	0.8	E493	0.6	D555	0.7	E493	0.6	D296	0.6

Appendix B

Hydrogen Bonds:

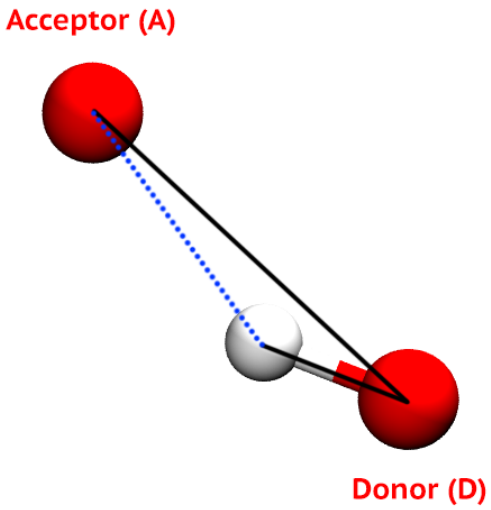


Figure 2: Representation of a hydrogen bond interaction. Black solid line indicates the angle between A-D-H, while blue dashed line shows the non-covalent interaction between A and H

Geometric criteria to count Hydrogen bonds:

- Distance D-A $\leq$ 3.5Å.
- Angle between A-D-H $\leq 30°$

Table 2: Average number of solute hydrogen bonds, for the first and second halves of the simulation. The abbreviation "lig" denotes ligand.

Ligand	Apo		ADP·Mg^{2+}		ATP·Mg^{2+}	
1^{st} half	Closed	Open	Closed	Open	Closed	Open
D1-D2	5.5 ± 0.6	2.3 ± 0.2	4.0 ± 0.3	2.3 ± 0.0	2.5 ± 0.2	0.4 ± 0.0
D1-D3	2.5 ± 0.1	1.9 ± 0.1	1.5 ± 0.1	1.7 ± 0.1	2.3 ± 0.1	1.8 ± 0.1
D2-D3	7.5 ± 0.3	6.3 ± 0.2	7.2 ± 0.2	6.9 ± 0.2	5.1 ± 0.2	6.7 ± 0.2
D1-D1	99.3 ± 0.4	98.7 ± 0.4	95.4 ± 0.7	96.4 ± 0.3	96.1 ± 0.2	95.3 ± 0.9
D2-D2	114.5 ± 0.7	118.4 ± 0.2	112.1 ± 1.1	115.8 ± 1.0	120.7 ± 0.3	113.5 ± 0.5
D3-D3	104.6 ± 0.4	105.0 ± 0.2	105.8 ± 0.8	103.9 ± 0.5	104.5 ± 0.3	105.0 ± 0.4
D1-lig	- - -	- - -	6.9 ± 0.3	7.0 ± 0.1	7.7 ± 0.1	7.8 ± 0.1
D2-lig	- - -	- - -	0.7 ± 0.2	2.7 ± 0.2	4.3 ± 0.1	1.0 ± 0.3
D3-lig	- - -	- - -	0	0	0	0
Lig-lig	- - -	- - -	1.0 ± 0.0	0	0.1 ± 0.0	1.0 ± 0.0
RNA-lig	- - -	- - -	0	0	0	0
D1-RNA	5.1 ± 0.3	3.8 ± 0.1	3.9 ± 0.1	3.8 ± 0.1	4.2 ± 0.1	3.8 ± 0.0
D2-RNA	7.0 ± 0.4	5.4 ± 0.3	7.4 ± 0.1	5.5 ± 0.0	9.1 ± 0.2	5.4 ± 0.2
D3-RNA	0.4 ± 0.1	3.7 ± 0.5	0.8 ± 0.1	2.1 ± 0.1	1.0 ± 0.1	3.0 ± 0.1
RNA-RNA	4.1 ± 0.2	3.7 ± 0.2	4.0 ± 0.1	5.1 ± 0.1	4.3 ± 0.1	3.2 ± 0.1
Total	350.5 ± 1.3	349.2 ± 0.8	350.7 ± 1.6	353.2 ± 1.2	361.9 ± 0.6	347.9 ± 1.2
ΔHB_{oc}	-1.3 ± 1.5		2.5 ± 2.0		-14.0 ± 1.4	
2^{nd} half	Closed	Open	Closed	Open	Closed	Open
D1-D2	5.5 ± 0.4	1.0 ± 0.2	4.1 ± 0.5	2.3 ± 0.0	3.1 ± 0.4	0.4 ± 0.1
D1-D3	2.5 ± 0.2	2.0 ± 0.3	1.9 ± 0.3	1.8 ± 0.1	2.2 ± 0.1	1.8 ± 0.1
D2-D3	7.8 ± 0.4	6.3 ± 0.1	7.3 ± 0.4	6.3 ± 0.2	5.2 ± 0.0	6.1 ± 0.1
D1-D1	97.7 ± 0.1	98.6 ± 0.6	95.5 ± 0.6	96.9 ± 0.2	95.1 ± 0.3	94.3 ± 0.5
D2-D2	112.8 ± 0.6	119.0 ± 0.8	113.9 ± 1.1	116.4 ± 0.7	119.7 ± 0.9	115.7 ± 0.6
D3-D3	104.8 ± 0.7	105.8 ± 0.5	105.4 ± 0.6	104.5 ± 0.3	104.8 ± 0.2	103.6 ± 0.2
D1-lig	- - -	- - -	8.9 ± 0.3	7.0 ± 0.1	7.7 ± 0.1	8.4 ± 0.3
D2-lig	- - -	- - -	1.2 ± 0.3	2.4 ± 0.6	4.5 ± 0.1	0.3 ± 0.1
D3-lig	- - -	- - -	0	0	0	0
Lig-lig	- - -	- - -	1.0 ± 0.0	0	0.1 ± 0.0	1.0 ± 0.0
RNA-lig	- - -	- - -	0	0	0	0
D1-RNA	5.2 ± 0.1	4.3 ± 0.1	4.3 ± 0.3	4.0 ± 0.1	4.5 ± 0.2	3.9 ± 0.0
D2-RNA	7.7 ± 0.3	4.0 ± 0.4	7.4 ± 0.1	5.6 ± 0.1	9.5 ± 0.1	6.0 ± 0.2
D3-RNA	0.5 ± 0.1	4.2 ± 0.1	0.7 ± 0.1	1.9 ± 0.2	1.0 ± 0.1	4.0 ± 0.2
RNA-RNA	3.6 ± 0.1	3.0 ± 0.1	3.9 ± 0.2	4.9 ± 0.1	4.5 ± 0.2	3.8 ± 0.3
Total	348.1 ± 1.2	348.2 ± 1.3	355.5 ± 1.7	354 ± 1.1	361.9 ± 1.1	349.3 ± 1.0
ΔHB_{oc}	0.1 ± 1.7		-1.5 ± 2.0		-12.6 ± 1.5	

Appendix C

Accurate Multiple Time Step in Biased Molecular Simulations:

This article presents the development of an algorithm that treats smooth biasing forces within a multiple time step (MTS) scheme. The implementation of this approach allows a speed up when a expensive set of CV's is selected to perform enhanced sampling simulations, and the computational profit can be significant when a GPU-based molecular dynamics software is employed. A theoretical framework to compute the errors is also introduced, which can be used to control the choice of the integration time step in both single and multiple time step biased simulations.

Although I was not involved in the theoretical discussions for the development of the MTS approach, I contributed in the practical application of a more realistic system like the NS3h-ssRNA complex, performing in this particular case a WTmetaD with an expensive and complex CV's as the Debye-Hückel interaction energy. From this application, it is observed a significant improvement in the computational performance without introducing large errors.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

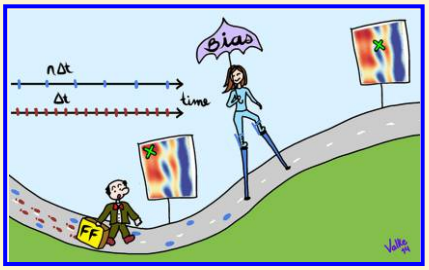
Accurate Multiple Time Step in Biased Molecular Simulations

Marco Jacopo Ferrarotti, Sandro Bottaro, Andrea Pérez-Villa, and Giovanni Bussi*

Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy

Supporting Information

ABSTRACT: Many recently introduced enhanced sampling techniques are based on biasing coarse descriptors (collective variables) of a molecular system on the fly. Sometimes the calculation of such collective variables is expensive and becomes a bottleneck in molecular dynamics simulations. An algorithm to treat smooth biasing forces within a multiple time step framework is here discussed. The implementation is simple and allows a speed up when expensive collective variables are employed. The gain can be substantial when using massively parallel or GPU-based molecular dynamics software. Moreover, a theoretical framework to assess the sampling accuracy is introduced, which can be used to assess the choice of the integration time step in both single and multiple time step biased simulations.



INTRODUCTION

Molecular simulations allow the dynamics of a system to be followed at atomistic resolution and thus can be used as virtual microscopes to investigate chemical reactions and conformational transitions in molecular systems.[1−3] The time scale that can be simulated with current computers and algorithms is on the order of a microsecond for a system of a few tens of thousands of atoms modeled with a typical atomistic empirical force field. Even though recently developed *ad hoc* hardware allowed for a sudden jump in this scale,[4] many relevant processes are still out of range for atomistic molecular dynamics (MD). A lot of effort has been made in the last decades to tackle or, at least, to alleviate this issue by means of modified algorithms known as enhanced sampling techniques. Among these, a class of methods is based on the idea of biasing a few carefully selected collective degrees of freedom, known as collective variables (CVs).[5−15] Biased enhanced-sampling simulations require these descriptors to be computed on the fly at every simulation time step. The cost of computing the CVs can in principle slow down the calculation. This is particularly true for expensive CVs, such as Steinhardt order parameters,[16,17] path CVs,[18] template-based CVs,[19] property maps,[20] social permutation−invariant coordinates,[21] sketch maps,[22] and Debye−Hückel energy,[23] just to mention a few. The calculation of the CVs and of the resulting biasing forces can be done in dedicated routines of MD software or can be implemented in external plugins.[24−26] It is very difficult to optimize the calculation of the CVs for several reasons. First, they sometime involve long-range exchange of information that makes them difficult to be compatible with domain-decomposition algorithms. Second, because their definition is highly system dependent, routines implementing these schemes should be flexible, thus paying an unavoidable price in terms of efficiency. On the opposite side, the calculation of physical force fields is becoming faster and faster thanks to the use of massively parallel and optimized software[27−29] and hardware.[4] The overall consequence of this trend is that the calculation of CVs can be expected to become a relevant bottleneck in enhanced sampling simulations, particularly in the context of massively parallel or GPU-accelerated MD.

We here introduce a scheme based on multiple time step integration (MTS) in its reference system propagator algorithm (RESPA) formulation.[30,31] More precisely, MTS is used here to split the integration of biasing forces and physical forces. The underlying principle is that often the former have a smoother dependence on atomic positions; thus, they change at a lower rate and can be integrated with a larger time step. This splitting allows for a substantial increase in the computational efficiency of biased sampling methods. Furthermore, we introduce a theoretical framework so as to assess the unavoidable time-step discretization errors. This is done by extending the concept of *effective energy*[32,33] so as to take into account the effect of biasing forces separately. This scheme allows artifacts that could be hidden when observing total-energy conservation to be highlighted. In the Method section, we introduce the methodology, first reviewing the MTS integration then introducing the scheme to assess sampling accuracy. We then present numerical tests of our algorithm on a one-dimensional model, on alanine dipeptide and on a large protein/RNA complex, using CVs of increasing complexity. In the latter case, we show that the use of MTS can greatly improve the performance without introducing significant errors.

METHODS

Multiple Time Step. We consider here a system of N_{at} atoms evolving accordingly to biased Hamilton equations in the form:

Received: August 4, 2014
Published: December 11, 2014

DOI: 10.1021/ct5007086
J. Chem. Theory Comput. 2015, 11, 139−146

Bibliography

[1] Francisco M. Averhoff, Nancy Glass, and Deborah Holtzman. Global Burden of Hepatitis C: Considerations for Healthcare Providers in the United States. *Clin. Infect. Dis.*, 55:S10–S15, 2012.

[2] Centers for Disease Control and Prevention. Hepatitis C FAQs for the Public. http://www.cdc.gov/hepatitis/hcv/, May 2015.

[3] Brett D. Lindenbach, Heinz-Jürgen Thiel, and Charles M. Rice. *Flaviviridae: The Viruses and Their Replication*. Fields Virology. Lippincott-Raven Publishers, Philadelphia, 5th edition, 2007.

[4] Seng-Lai Tan, editor. *Hepatitis C Viruses: Genomes and Molecular Biology*. Horizon Bioscience, Norfolk (UK), 1st edition, 2006.

[5] Roger H. Miller and Robert H. Purcell. Hepatitis C virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups. *Proc. Natl. Acad. Sci. U. S. A.*, 87(6):2057–2061, 1990.

[6] Qui-Lim Choo, Kathy H. Richman, Jang H. Han, Kim Berger, Cindy Lee, Christine Dong, Carol Gallegos, Doris Coit, Angelica Medina-Selby, and Philip J. Barr. Genetic organization and diversity of the Hepatitis C virus. *Proc. Natl. Acad. Sci. U. S. A.*, 88(6):2451–2455, 1991.

[7] Toronto Centre for Liver Disease. Hepatitis C - Basic Information. http://torontoliver.ca/hepatitis-c-basic-faq/, 2015.

[8] Ralf Bartenschlager, Volker Lohmann, and Francois Penin. The molecular and structural basis of advanced antiviral therapy for Hepatitis C virus infection. *Nat. Rev. Micro.*, 11(7):482–496, 2013.

[9] Vaishnavi Rajagopal, Madhura Gurjar, Mikhail K. Levin, and Smita S. Patel. The protease domain increases the translocation stepping efficiency of the Hepatitis C virus NS3-4A helicase. *J. Biol. Chem.*, 285(23):17821–17832, 2010.

[10] Ralf Bartenschlager, Ludwina Ahlborn-Laake, Jan Mous, and Helmut Jacobsen. Kinetic and structural analyses of Hepatitis C virus polyprotein processing. *J. Virol.*, 68(8):5045–5055, 1994.

[11] Raffaele De Francesco and Giovanni Migliaccio. Challenges and successes in developing new therapies for Hepatitis C. *Nature*, 436(7053):953–960, 2005.

[12] Beate Schwer. A new twist on RNA helicases: DExH/D box proteins as RNPases. *Nat. Struct. Mol. Biol.*, 8(2):113–116, 2001.

[13] Phillip S. Pang, Eckhard Jankowsky, Paul J. Planet, and Anna Marie Pyle. The Hepatitis C viral NS3 protein is a processive DNA helicase with cofactor enhanced RNA unwinding. *EMBO J.*, 21(5):1168–1176, 2002.

[14] Anna Marie Pyle. Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.*, 37(1):317–336, 2008.

[15] Steve C. Ding and Anna Marie Pyle. Molecular mechanics of RNA translocases. *Methods Enzymol.*, 511:131–147, 2012.

[16] Meredith D. Betterton and Frank Jülicher. Opening of nucleic-acid double strands by helicases: Active versus passive opening. *Phys. Rev. E*, 71:011904, 2005.

[17] Timothy M. Lohman and Keith P. Bjornson. Mechanisms of helicase-catalyzed DNA unwinding. *Annu. Rev. Biochem.*, 65(1):169–214, 1996.

[18] Peter H. von Hippel and Emmanuelle Delagoutte. A general model for nucleic acid helicases and their "coupling" within macromolecular machines. *Cell*, 104(2):177 – 190, 2001.

[19] Kevin D. Raney, Suresh D. Sharma, Ibrahim M. Moustafa, and Craig E. Cameron. Hepatitis C virus non-structural protein 3 (HCV NS3): A multifunctional antiviral target. *J. Biol. Chem.*, 285(30):22725–22731, 2010.

[20] Nanhua Yao, Paul Reichert, S. Shane Taremi, Winifred W. Prosise, and Patricia C. Weber. Molecular views of viral polyprotein processing revealed by the crystal structure of the Hepatitis C virus bifunctional protease–helicase. *Structure*, 7(11):1353 – 1363, 1999.

[21] Meigang Gu and Charles M. Rice. Three conformational snapshots of the Hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism. *Proc. Natl. Acad. Sci. U. S. A.*, 107(2):521–528, 2010.

[22] Todd C. Appleby, Robert Anderson, Olga Fedorova, Anna M. Pyle, Ruth Wang, Xiaohong Liu, Katherine M. Brendza, and John R. Somoza. Visualizing ATP-dependent RNA translocation by the NS3 helicase from HCV. *J. Mol. Biol.*, 405(5):1139 – 1153, 2011.

[23] Sua Myong and Taekjip Ha. Stepwise translocation of nucleic acid motors. *Curr. Opin. Struct. Biol.*, 20(1):121 – 127, 2010.

[24] Wei Cheng, Srikesh G. Arunajadai, Jeffrey R. Moffitt, Ignacio Tinoco, and Carlos Bustamante. Single–base pair unwinding and asynchronous RNA release by the Hepatitis C virus NS3 helicase. *Science*, 333(6050):1746–1749, 2011.

[25] Sophie Dumont, Wei Cheng, Victor Serebrov, Rudolf K. Beran, Ignacio Tinoco, Anna Marie Pyle, and Carlos Bustamante. RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature*, 439(7072):105–108, 2006.

[26] Wei Cheng, Sophie Dumont, Ignacio Tinoco, and Carlos Bustamante. NS3 helicase actively separates RNA strands and senses sequence barriers ahead of the opening fork. *Proc. Natl. Acad. Sci. U. S. A.*, 104(35):13954–13959, 2007.

[27] Mariusz Krawczyk, Anna Stankiewicz-Drogoń, Anne-Lise Haenni, and Anna Boguszewska-Chachulska. Fluorometric assay of Hepatitis C virus NS3 helicase activity. *Methods Mol. Biol.*, 587:211–221, 2010.

[28] Sourav Mukherjee, Alicia M. Hanson, William R. Shadrick, Jean Ndjomou, Noreena L. Sweeney, John J. Hernandez, Diana Bartczak, Kelin Li, Kevin J. Frankowski, Julie A. Heck, Leggy A. Arnold, Frank J. Schoenen, and David N. Frick. Identification and analysis of Hepatitis C virus NS3 helicase inhibitors using nucleic acid binding assays. *Nucleic Acids Res.*, 40(17):8607–8621, 2012.

[29] Gustavo Tavares Ventura, Emmerson Corrêa Brasil da Costa, Anne Miranda Capaccia, and Ronaldo Mohana-Borges. pH-dependent conformational changes in the HCV NS3 protein modulate its ATPase and helicase activities. *PLoS ONE*, 9(12):e115941, 2014.

[30] Srikesh G. Arunajadai. RNA unwinding by NS3 helicase: A statistical approach. *PLoS ONE*, 4(9):e6937, 2009.

[31] Thomas A. Jennings, Samuel G. Mackintosh, Melody K. Harrison, Deniz Sikora, Bartek Sikora, Bhuvanesh Dave, Alan J. Tackett, Craig E. Cameron, and Kevin D. Raney. NS3 helicase from the Hepatitis C virus can function as a monomer or oligomer depending on enzyme and substrate concentrations. *J. Biol. Chem.*, 284(8):4806–4814, 2009.

[32] Martin R. Singleton, Mark S. Dillingham, and Dale B. Wigley. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, 76(1):23–50, 2007.

[33] Eckhard Jankowsky. RNA helicases at work: binding and rearranging. *Trends Biochem. Sci.*, 36(1):19 – 29, 2011.

[34] Mikhail K. Levin, Yuh-Hwa Wang, and Smita S. Patel. The functional interaction of the Hepatitis C virus helicase molecules is responsible for unwinding processivity. *J. Biol. Chem.*, 279(25):26005–26012, 2004.

[35] Bartek Sikora, Yingfeng Chen, Cheryl F. Lichti, Melody K. Harrison, Thomas A. Jennings, Yong Tang, Alan J. Tackett, John B. Jordan, Joshua Sakon, Craig E. Cameron, and Kevin D. Raney. Hepatitis C virus NS3 helicase forms oligomeric structures that exhibit optimal DNA unwinding activity in vitro. *J. Biol. Chem.*, 283(17):11516–11525, 2008.

[36] John E. Walker, Matti Saraste, Michael J. Runswick, and Nicholas J. Gay. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, 1(8):945–951, 1982.

[37] Phyllis I. Hanson and Sidney W. Whiteheart. AAA+ proteins: have engine, will work. *Nat. Rev. Mol. Cell Biol.*, 6(7):519–529, 2005.

[38] Jan Liphardt, Sophie Dumont, Steven B. Smith, Ignacio Tinoco, and Carlos Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of jarzynski's equality. *Science*, 296(5574):1832–1835, 2002.

[39] Michael D. Brenner, Ruobo Zhou, and Taekjip Ha. Forcing a connection: Impacts of single-molecule force spectroscopy on in vivo tension sensing. *Biopolymers*, 95(5):332–344, 2011.

[40] Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.

[41] Holger Flechsig and Alexander S. Mikhailov. Tracing entire operation cycles of molecular motor Hepatitis C virus helicase in structurally resolved dynamical simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 107(49):20875–20880, 2010.

[42] Wenjun Zheng and Mustafa Tekpinar. Structure-based simulations of the translocation mechanism of the Hepatitis C virus NS3 helicase along single-stranded nucleic acid. *Biophys. J.*, 103(6):1343 – 1353, 2012.

[43] Mark E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 1st edition, 2010.

[44] Lars V. Bock, Christian Blau, Gunnar F. Schröder, Iakov I. Davydov, Niels Fischer, Holger Stark, Marina V. Rodnina, Andrea C. Vaiana, and Helmut Grubmüller. Energy barriers and driving forces in tRNA translocation through the ribosome. *Nat. Struct. Mol. Biol.*, 20(12):1390–1396, 2013.

[45] Miroslav Krepl, Marek Havrila, Petr Stadlbauer, Pavel Banáš, Michal Otyepka, Josef Pasulka, Richard Štefl, and Jiri Sponer. Can we execute stable microsecond-scale atomistic simulations of protein-RNA complexes? *J. Chem. Theory Comput.*, 11(3):1220–1243, 2015.

[46] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, 51(7):91–97, 2008.

[47] David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 39:1–39:11, New York, NY, USA, 2009.

[48] Bernie J. Alder and Thomas E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27(5):1208–1209, 1957.

[49] Bernie J. Alder and Thomas E. Wainwright. Studies in molecular dynamics. i. general method. *J. Chem. Phys.*, 31(2):459–466, 1959.

[50] Aneesur Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.*, 136:A405–A411, 1964.

[51] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[52] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys.*, 389(20):457–484, 1927.

[53] Max Born and Vladimir Fock. Beweis des adiabatensatzes. *Z. Phys.*, 51(3-4):165–180, 1928.

[54] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[55] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

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

[57] Jürgen Schlitter, Marc Engels, Peter Krüger, Edgar Jacoby, and Axel Wollmer. Targeted molecular dynamics simulation of conformational change-application to the T ↔ R transition in insulin. *Mol. Simul.*, 10(2-6):291–308, 1993.

[58] Philippe Ferrara, Joannis Apostolakis, and Amedeo Caflisch. Computer simulations of protein folding by targeted molecular dynamics. *Proteins: Struct., Funct., Bioinf.*, 39(3):252–260, 2000.

[59] Arjan van der Vaart and Martin Karplus. Simulation of conformational transitions by the restricted perturbation–targeted molecular dynamics method. *J. Chem. Phys.*, 122(11):114903, 2005.

[60] Jingwei Weng, Kangnian Fan, and Wenning Wang. The conformational transition pathways of ATP-binding cassette transporter BtuCD revealed by targeted molecular dynamics simulation. *PLoS ONE*, 7(1):e30465, 2012.

[61] Romain M. Wolf. Extracting ligands from receptors by reversed targeted molecular dynamics. *J. Comput. Aided Mol. Des.*, pages 1–10, 2015.

[62] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A*, 32(5):922–923, 1976.

[63] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99(20):12562–12566, 2002.

[64] C.W. Gear, Ioannis G. Kevrekidis, and Constantinos Theodoropoulos. 'coarse' integration/bifurcation analysis via microscopic simulators: micro-galerkin methods. *Comput. Chem. Eng.*, 26(7–8):941 – 963, 2002.

[65] Djurdje Cvijović and Jacek Klinowski. Taboo search: An approach to the multiple minima problem. *Science*, 267(5198):664–666, 1995.

[66] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96:090601, 2006.

[67] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100:020603, 2008.

[68] Giovanni Bussi and Davide Branduardi. *Free-Energy Calculations with Metadynamics: Theory and Practice*, pages 1–49. John Wiley & Sons, Inc., 2015.

[69] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From A to B in free energy space. *J. Chem. Phys.*, 126(5):–, 2007.

[70] Alessio Lodola, Davide Branduardi, Marco De Vivo, Luigi Capoferri, Marco Mor, Daniele Piomelli, and Andrea Cavalli. A catalytic mechanism for cysteine N-terminal nucleophile hydrolases, as revealed by free energy simulations. *PLoS ONE*, 7(2):e32397, 2012.

[71] Davide Branduardi, Marco De Vivo, Nadia Rega, Vincenzo Barone, and Andrea Cavalli. Methyl phosphate dianion hydrolysis in solution characterized by path collective variables coupled with DFT-based enhanced sampling simulations. *J. Chem. Theory Comput.*, 7(3):539–543, 2011.

[72] Grégoire A. Gallet, Fabio Pietrucci, and Wanda Andreoni. Bridging static and dynamical descriptions of chemical reactions: An ab initio study of CO2 interacting with water molecules. *J. Chem. Theory Comput.*, 8(11):4029–4039, 2012.

[73] Massimiliano Bonomi, Davide Branduardi, Francesco L. Gervasio, and Michele Parrinello. The unfolded ensemble and folding mechanism of the C-terminal GB1 β-hairpin. *J. Am. Chem. Soc*, 130(42):13938–13944, 2008. PMID: 18811160.

[74] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1–2):141 – 151, 1999.

[75] Hugh Nymeyer. How efficient is replica exchange molecular dynamics? an analytic approach. *J. Chem. Theory Comput.*, 4(4):626–636, 2008.

[76] Giovanni Bussi. Hamiltonian replica exchange in gromacs: a flexible implementation. *Mol. Phys.*, 112(3-4):379–384, 2014.

[77] Dahai Luo, Ting Xu, Cornelia Hunke, Gerhard Grüber, Subhash G. Vasudevan, and Julien Lescar. Crystal structure of the NS3 protease-helicase from Dengue virus. *J. Virol.*, 82(1):173–183, 2008.

[78] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

[79] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[80] Junmei Wang, Piotr Cieplak, and Peter 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(12):1049–1074, 2000.

[81] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65(3):712–725, 2006.

[82] Robert B. Best and Gerhard Hummer. Optimized molecular dynamics force fields applied to the helix coil transition of polypeptides. *J. Phys. Chem. B*, 113(26):9004, 2009.

[83] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David .E. Shaw. Improved side-chain torsion potentials for the AMBER ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.*, 78(8):1950–1958, 2010.

[84] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Sponer, Thomas E. Cheatham III, Charles A. Laughton, and Modesto Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of [alpha]/[gamma] conformers. *Biophys. J.*, 92(11):3817–3829, 2007.

[85] Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mládek, Pavel Banáš, Thomas E. Cheatham, and Petr Jurečka. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[86] Kristin L. Meagher, Luke T. Redman, and Heather A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.*, 24(9):1016–1025, 2003.

[87] Olof Allnér, Lennart Nilsson, and Alessandra Villa. Magnesium ion–water coordination and exchange in biomolecular simulations. *J. Chem. Theory Comput.*, 8(4):1493–1502, 2012.

[88] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[89] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[90] Herman J. C. Berendsen, Johan P. M. Postma, Wilfred F. van Gunsteren, Alfredo Di Nola, and Jan R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[91] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[92] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[93] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[94] Andrea Amadei, Antonius B. M. Linssen, and Herman J. C. Berendsen. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.*, 17(4):412–425, 1993.

[95] Davide Branduardi, Fabrizio Marinelli, and José D. Faraldo-Gómez. Atomic-resolution dissection of the energetics and mechanism of isomerization of hydrated ATP-Mg^{2+} through the SOMA string method. *J. Comput. Chem.*, doi: 10.1002/jcc.23991:1–10, 2015.

[96] Kun-Yi Hsin, Yi-gong Sheng, Marjorie M. Harding, Paul Taylor, and Malcolm D. Walkinshaw. MESPEUS: a database of the geometry of metal sites in proteins. *J. Appl. Crystallogr.*, 41(5):963–968, 2008.

[97] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.*, 38(1-2):236–240, 1999.

[98] Alenka Luzar and David Chandler. Hydrogen-bond kinetics in liquid water. *Nature*, 379(6560):55–57, 1996.

[99] Peter Debye and Erich Hückel. The theory of electrolytes. i. lowering of freezing point and related phenomena. *Physik. Z.*, 24:185–206, 1923.

[100] Trang N. Do, Paolo Carloni, Gabriele Varani, and Giovanni Bussi. RNA/peptide binding driven by electrostatics—insight from bidirectional pulling simulations. *J. Chem. Theory Comput.*, 9(3):1720–1730, 2013.

[101] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014.

[102] Ken A. Dill and Sarina Bromberg. *Molecular Driving Forces*. Garland Science, New York, 2nd edition, 2010.

[103] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690–2693, 1997.

[104] Toru Sengoku, Osamu Nureki, Akira Nakamura, Satoru Kobayashi, and Shigeyuki Yokoyama. Structural basis for RNA unwinding by the DEAD-Box protein Drosophila Vasa. *Cell*, 125(2):287 – 300, 2006.

[105] Christian B. F. Andersen, Lionel Ballut, Jesper S. Johansen, Hala Chamieh, Klaus H. Nielsen, Cristiano L.P. Oliveira, Jan Skov Pedersen, Bertrand Séraphin, Hervé Le Hir, and Gregers Rom Andersen. Structure of the exon junction core complex with a trapped dead-box atpase bound to rna. *Science*, 313(5795):1968–1972, 2006.

[106] John D. Chodera and David L. Mobley. Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.*, 42(1):121–142, 2013.

[107] Christoph Dellago and Gerhard Hummer. Computing equilibrium free energies using non-equilibrium molecular dynamics. *Entropy*, 16(1):41–61, 2014.

[108] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2014.

[109] Andrea Pérez-Villa, Maria Darvas, and Giovanni Bussi. ATP dependent NS3 helicase interaction with RNA: insights from molecular simulations. *Nucleic Acids Res.*, doi: 10.1093/nar/gkv872, 2015.

[110] Vittorio Limongelli, Luciana Marinelli, Sandro Cosconati, Concettina La Motta, Stefania Sartini, Laura Mugnaini, Federico Da Settimo, Ettore Novellino, and Michele Parrinello. Sampling protein motion and solvent effect during ligand binding. *Proc. Natl. Acad. Sci. U. S. A.*, 109(5):1467–1472, 2012.

[111] Jeremy C. Palmer, Roberto Car, and Pablo G. Debenedetti. The liquid-liquid transition in supercooled st2 water: a comparison between umbrella sampling and well-tempered metadynamics. *Faraday Discuss.*, 167:77–94, 2013.

[112] Elena Papaleo, Ludovico Sutto, Francesco Luigi Gervasio, and Kresten Lindorff-Larsen. Conformational changes and free energies in a proline isomerase. *J. Chem. Theory Comput.*, 10(9):4169–4174, 2014.

[113] Matteo Salvalaglio, Claudio Perego, Federico Giberti, Marco Mazzotti, and Michele Parrinello. Molecular-dynamics simulations of urea nucleation from aqueous solution. *Proc. Natl. Acad. Sci. U. S. A.*, 112(1):E6–E14, 2015.

[114] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *WIREs Comput. Mol. Sci.*, 1(5):826–843, 2011.

[115] Grisell Díaz Leines and Bernd Ensing. Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett.*, 109:020601, 2012.

[116] Mikhail K. Levin, Madhura Gurjar, and Smita S. Patel. A brownian motor mechanism of translocation and strand separation by Hepatitis C virus helicase. *Nat. Struct. Mol. Biol.*, 12(5):429–435, 2005.

[117] John Stone. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

Acknowledgments

The PhD has been the most challenging and unique experience that I have faced so far during my academical career. This process would not be the same without the presence of different professors, researchers and friends that I meet during these four years.

First of all I would like to thank to my supervisor, Prof. Giovanni Bussi because he has always been very supportive and patient with me. Giovanni taught me so many concepts that were crucial for the progress of my research, and thanks to his experience I have learned which are the relevant issues that must be take into account when you are doing research. For this reason, I feel very fortunate that I have had the opportunity to do my PhD with him.

I want to thank to Prof. Giorgio Colombo and Prof. Francesco Luigi Gervasio, for kindly accepting to be the external examiners of my research project.

I would also like to thank to the *Molecular and Statistical Biophysics group* (SBP) for all the lectures, seminars and conferences that contributed significantly to my scientific formation. I would like to thank to all SBP professors, specially to Prof. Alessandro Laio for all support that I have received from him to join the PhD in SISSA. I also want to thank to all its postdocs and students. Specially to Ana-Maria Florescu and Negar Nahali for the ice-creams and teas together, Francesca Rizzato and Edoardo Sarti for helping me when I needed it and for their friendship.

I am also very thankful to all the guys from Giovanni's lab (SRNAS), specially to Simón Poblete and Sandro Bottaro. All scientific support, explanations and suggestions that I have received from both of them were very important.

I am also very thankful to SISSA for giving me the possibility to do my PhD in this beautiful italian institution. I am also grateful for the *SISSA per le scuole* program with Simona Cerrato, because this experience taught me how to communicate science in a better way.

I am also very grateful to Catalina Reyes, Ivan Gladich, Duván Franco and Fahimeh Baftizadeh for listening to me and being so supportive in different periods of my PhD.

I am sincerely thankful to Prof. Albeiro Restrepo, my former supervisor in Colombia at Universidad de Antioquia. He is undoubtedly a great scientist and taught me the scientific basis that I needed to start my experience in SISSA.

I want to thank to my Colombian friends spread all over the world. Specially to Cristiam Santa, Felipe Zapata and Steven Roldán for always care about me beside the distance and for their invaluable friendship.

I am deeply thankful to my parents for always believing in me and for all their support. Without them I would not be able to pursue my PhD and I would not be the person that I am now. Last but not least, I have all my gratitude to Enrico Cavo for his love and unconditional support received every single day, and for always giving me a beautiful reason to smile and be happy.

Colophon

This document was written using the LaTeX typesetting system and the memoir class created by Peter Wilson. The body text is set 11pt with URW Palladio. Other fonts include ZALLMAN. Most of the molecular representations were generated with VMD program using the Tachyon rendering [78, 117].