Helicase Functional Dynamics from Low-Resolution Experimental Data and Simulation

Gaël Radou

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of Leeds
School of Molecular and Cellular Biology

October, 2015

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 contains results from publication, *Radou G, Dreyer FN, Tuma R, Paci E, Functional Dynamics of the Packaging Motor P4 Probed by Hydrogen Exchange and Simulation. Biophys J 2014, 107(4):983–990*. I performed all molecular dynamics simulations and analysis of the simulations. The paper was jointly written by all authors.

Chapter 5 involves a free-energy profile calculation method developed by Dr. Sergei Krivov that I had already applied in publication, *Radou G, Encisco M, Krivov S, Paci E (2013), Modulation of a Protein Free-Energy Landscape by Circular Permutation. J Phys Chem B 117(44): 13743-13747*.

Acknowledgements

There are a number of people behind this piece of work who deserve to be both acknowledged and thanked.

First and foremost, I would like to express my sincere gratitude to my supervisors, Emanuele Paci and Roman Tuma, for their support and commitment throughout my Ph.D.

I am also thankful to Alexander Borodavka, Amit Sharma, Anna Polyakova, Diana Montero, Girish Tampi, Gloria Comadira, James Gowdy, Kostas Papachristos, Matthew Batchelor, Nigel Taylor, Paulina Banushkina, Philip Morrison, Rebecca Thompson, Sergei Krivov, Theodoros Karamanos. All these people, one way or another, helped, supported and encouraged me over the last four years.

Finally, I would like to thank my family and Clément for their love and unconditional support.

Abstract

The biological function of large macromolecular assemblies depends on their structure and their dynamics over a broad range of timescales; for this reason its investigation poses significant challenges to conventional experimental techniques. A promising experimental technique is hydrogen-deuterium exchange detected by mass spectrometry (HDX-MS). I begin by presenting a new computational method for quantitative interpretation of deuterium exchange kinetics. The method is tested on a hexameric viral helicase φ12 P4 that pumps RNA into a virus capsid at the expense of ATP hydrolysis. Molecular dynamics simulations predict accurately the exchange kinetics of most peptide fragments and provide a residue-level interpretation of the low-resolution experimental results. This approach is also a powerful tool to probe mechanisms that cannot be observed by X-ray crystallography, or that occur over timescales longer than those that can be realistically simulated, such as the opening of the hexameric ring. Once validated, the method is applied on a homologous system, the packaging motor φ8 P4, for which RNA loading and translocation mechanisms remain elusive. Quantitative interpretation of HDX-MS data, as well as Förster resonance energy transfer (FRET) and computational observations, suggest that the C-terminal domain of the motor plays a crucial role. A new translocation model of φ8 P4 is proposed, for which the affinity between the motor and RNA is modulated by the C-termini. In the final result chapter, the amount of the structural information carried by HDX-MS data is quantitatively analysed. The impact of the averaging of the exchange over peptide fragments on the information content is investigated. The complementarity of data obtained from HDX-MS and data obtained from other techniques (such as NMR, FRET or SAXS) is also examined.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAA$^+$	ATPases associated with diverse cellular activities
ACF	Autocorrelation function
ADP	Adenosine diphosphate
AF488	Alexa Fluor 488 dye
AF4594	Alexa Fluor 594 dye
ALEX	Alternating laser excitation
AMPcPP	α, β-methyleneadenosine-5’-triphosphate
ATP	Adenosine triphosphate
CHARMM	Chemistry at HARvard Molecular Mechanics
CPU	Central processing unit
CS	Chemical shift
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
ECD	Electron-capture dissociation
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EX1	Hydrogen-deuterium exchange in regime 1
EX2	Hydrogen-deuterium exchange in regime 2
FCS	Fluorescence correlation spectroscopy
FEP	Free-energy profile
FRET	Förster resonance energy transfer
GA	Genetic algorithm
H/D	Hydrogen-deuterium

HDX	Hydrogen-deuterium exchange
HDX-MS	Hydrogen-deuterium exchange mass-spectrometry
IM	Ion-mobility
k_{int}	Intrinsic hydrogen exchange rate constant
LB	Lysogeny broth
LD	Langevin dipole
MD	Molecular dynamic
MSD	Mean square deviation
NAMD	NAnoscale Molecular Dynamics program
NH group	Amide NH group
NMR	Nuclear magnetic resonance
NTA	Nitrilotriacetic
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PD	Protein dipole
PDLD	Protein dipole – langevin dipole
Poly(A)	Ribonucleic acid chain of adenines
Pfact	Protection factor
PMSF	Phenylmethylsulfonyl fluoride
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid
SAXS	Small-angle X-ray scattering
SDS	Sodium dodecyl sulphate
SDS-PAGE	SDS polyacrylamide gel electrophoresis
SOC	Super Optimal broth with Catabolite repression
SF	Super family

ssRNA	Single-stranded ribonucleic acid
TAE	Tris-acetate-EDTA
TCEP	Tris(2-carboxyethyl)phosphine
TEMED	Tetramethylethylenediamine
TIP3P	Transferable Intermolecular Potential 3P
TIRF	Total internal reflection fluorescence
WT	Wild-type

Chapter 1: Introduction

1.1 The virus packaging motors P4

1.1.1 Genome encapsulation in viruses:

Viruses are small infectious entities that can only replicate using the machinery and metabolism of a host cell. Their genome consists of one or multiple segments, single or double stranded, of DNA or RNA. In RNA viruses, the strand that serves for replication is known as the (-)-strand and the strand used for protein synthesis is referred as the (+)-strand. Either one strand or both strands (dsRNA) can serve as a genomic RNA.

Viruses carry their genetic material inside a capsid consisting of protein shell and in some cases also a lipid bilayer. Encapsulation protects the genome and also enables attachment to specific receptors of the targeted cells. Most capsids are made of hundreds of identical protein subunits arranged with a high degree of symmetry, which can be either helical or icosahedral. For icosahedral viruses, two mechanisms exist for condensation of the viral genome inside capsid: (i) nucleation of the capsid around the nucleic acid or (ii) packaging of the genome into the preformed capsid, called the procapsid. In the first case, the virus assembles spontaneously in the host cell (1). Assembly is driven by the high affinity of the protein subunit of the capsid and packaging sites along the genome (2). Packaging into a procapsid is performed by active portals integrated into the capsid. The portals are molecular motors that pump the nucleic acid chain into the capsid at the expense of ATP hydrolysis. The tight confinement of the negatively charged nucleic acid results in a loss of entropy and in electrostatic repulsions (3). Hence, the packaging motors have to generate high forces needed to compact the genome against increasing internal forces, while translocation of ssRNA may not require high forces. For translocation of ssRNA, the packaging motor also needs to unwind the RNA strand that tends to form extensive secondary structure.

Key to understanding packaging mechanisms in RNA viruses remains partially elusive. The goal of this thesis is to elucidate the mechanisms used by packaging motors P4, found in double-stranded RNA bacterial viruses from the Cystoviruses family (φ6, φ8, φ12 and φ13).

1.1.2 Formation of the Cystovirus capsid

Bacteriophages φ6, φ8, φ12 and φ13 of the *Cystoviridae* family infect plant pathogenic bacteria (4). A virion consists of three-layers. The outer part consists of a lipid layer integrating membrane proteins that are exposed to surface binding receptors mediating fusion with the host outer membrane. Under this membrane is found the nucleocapsid. The nucleocapsid is composed of the procapsid, and a concentric shell of protein P8. The four proteins named P1, P2, P4 and P7 are sufficient to form an icosahedral procapsid of ~50 nm in diameter (Figure 1-1). P1 is the major structural component with a total of 120 copies (5). It co-assembles with ~10 copies of RNA-dependent RNA polymerases P2, ~11-12 P4 hexamers (6) and 12 copies of the assembly cofactor P7. P2 was shown to act both as a replicase and a transcriptase utilizing single- or double-stranded RNA, respectively (7, 8). The presence of P7 stabilizes capsid during RNA packaging and is essential for the activity of the virion (9, 10). Finally, P4 hexamer is the portal through which RNA is pumped and is our protein of interest. The resulting capsid is functional and can sequentially package (+)-strand RNA genomic precursors, synthesize the complementary minus strand RNA and then transcribe additional plus strands. After packaging and replication, the RNA-filled procapsid is enveloped by ~600 P8 dimers (which is missing in φ8 (11)) to give rise to a nucleocapsid of ~58 nm (12). Then the virus acquires its lipid envelope during maturation.

The viral genome is composed of three double-stranded RNA molecules: L, large (~6kb); M, middle (4kb); and S, small (~3kb) (4). Each (+)-strand is packed into the capsid in the 5' to 3' direction. Packaging is performed in a specific order determined by the *pac* sequence localized at the 5' ends (13). Since the packaging motor P4 binds all types of polynucleotides, RNA sequence selection is controlled by another component, very likely the assembled P1 (14). The viral capsid undergoes drastic conformational changes during packaging due to increase of the internal pressure. It was proposed that those changes activate and inactive RNA binding sites to control selection of the *pac* sequence (15). P4 also acts as a passive conduit for the exit of ssRNA transcripts (16). The life cycle of the virus is summarised in Figure 1-1.

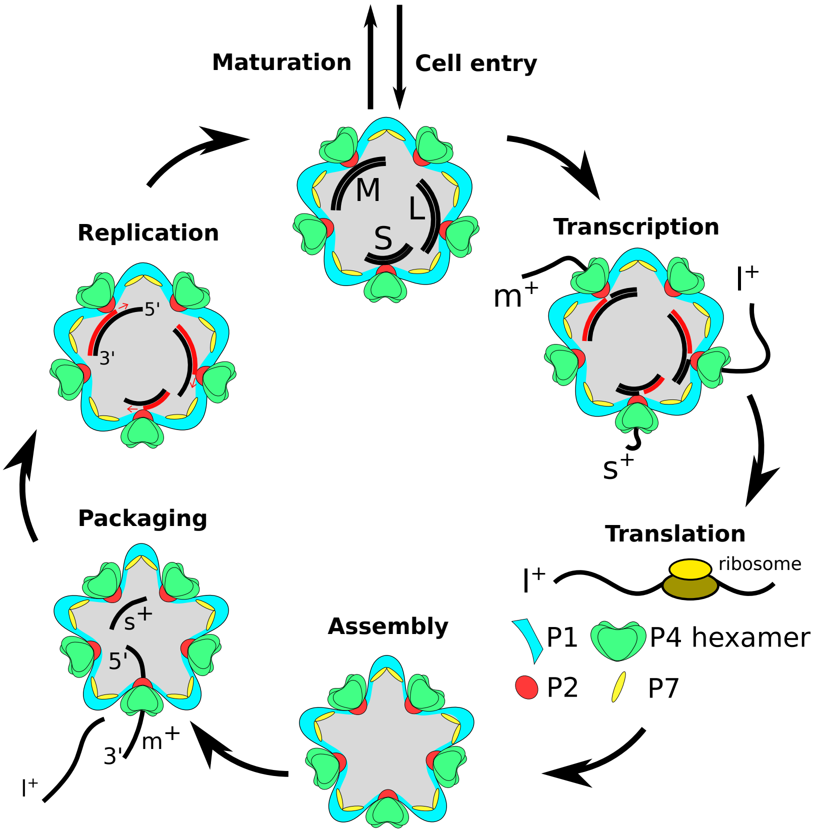


***Figure 1-1**: Scheme of the Cystoviridae core replication.*
After cell entry, transcription of (+)-RNA by P2 is activated. Upon transcription, new (+)-strand (l$^+$, m$^+$ and s$^+$) are excluded from the procapsid via P4 which acts as a passive portal. The ribosome of the host cell translates l$^+$ RNA to produce P1, P2 P4 and P7, which co-assemble to form empty procapsids. The different elements of the procapsid, P1, P2, P4 and P7, are represented in cyan, red, green and yellow, respectively. Once the procapsid is assembled, (+)-RNA are sequentially packed by the motor P4. The procapsid expands upon packaging due to the internal pressure. The (+)-RNA is replicated inside the procapsid by P2 to yield double stranded-RNA. Modified from (16). The red strands indicate RNA being synthesized by P2.

1.1.3 The packaging motors P4:

Crystal structures of the packaging motors P4 from bacteriophages φ6, φ8, φ12 and φ13 are available (17). They are similar in terms of structure to hexameric helicases, although their similarities are poor in terms of sequence (14, 18). Helicases are classified into six superfamilies (SF1-SF6) based on their sequence (19). In each superfamily, the overall sequence similarity is usually poor and confined to short sequence motifs of the AAA$^+$ or RecA-like structural domains (20). The RecA-like core converts the energy of ATP binding and hydrolysis into mechanical force to perform DNA or RNA translocation. The ATP binding site is localised at the interface between two subunits or RecA-like domains (for monomeric helicases), which is a crucial position to enable subunit communication and ATP hydrolysis coordination. SF1 and SF2 are the largest groups and their helicases exhibit the highest similarities. Contrary to SF3-SF6, SF1-SF2 helicases do not form ring structures. SF3 enzymes form hexamers or double-hexamers and share a modified AAA$^+$ core. Members of SF4 are all hexameric. They “walk” along the nucleic acid chain in 5’ to 3’ direction and have a RecA core. Although the Rho factor is closely related to SF4 helicase, due to its specific sequence, it was classified into a separate family called SF5. Finally, SF6 contains all the hexameric enzymes containing the core AAA$^+$ that do not fall into SF3.

P4 proteins belong to the Superfamily SF4, which is characterized by five well conserved motifs (H1, H1a, H2, H3 and H4) (21), as well as a structurally conserved arginine finger (Figure 1-2). Motifs H1a and H2 form a binding pocket where hydrolysis takes place. H1a binds the γ-phosphate and assists hydrolysis, whereas H2 coordinates Mg^{2+} for catalysis. H1, also known as the P-loop, interacts with the α- and β-phosphates of the nucleotide bound to the catalytic site. H3 contains the γ-phosphate sensor. H4 encompasses the α6 helix. The α6 helix is directly connected to the L2 loop which binds to RNA. X-ray crystal structures of φ12 P4 in different states during ATP hydrolysis (18) revealed that the α6 helix and L2 loop form a lever that has two positions: “up” or “down” (Figure 1-2C). In the absence of nucleotide, the lever is found in both positions, whereas binding of ATP locks the lever in the “up” position. Upon ATP hydrolysis, the nucleotide binding P-loop undergoes conformational changes, swivels up and perturbs one end of the L2 loop. It causes the lever to pivot down and switch into the “down” position (22). This motion drags down to ~6 Å the tip of the lever where is located lysine which binds RNA. The L1 loop is extremely flexible and is thought to act as a grommet that keeps the RNA at the centre of the channel (18). The arginine finger contributes to the catalytic site from the neighbouring subunit. By pointing into the binding pocket, it neutralises the negative charge of the phosphate groups of the nucleotide and stabilizes the transition state, resulting in faster ATP hydrolysis (23). The arginine finger is not well conserved among helicase sequences and can be found at different positions in the structure.

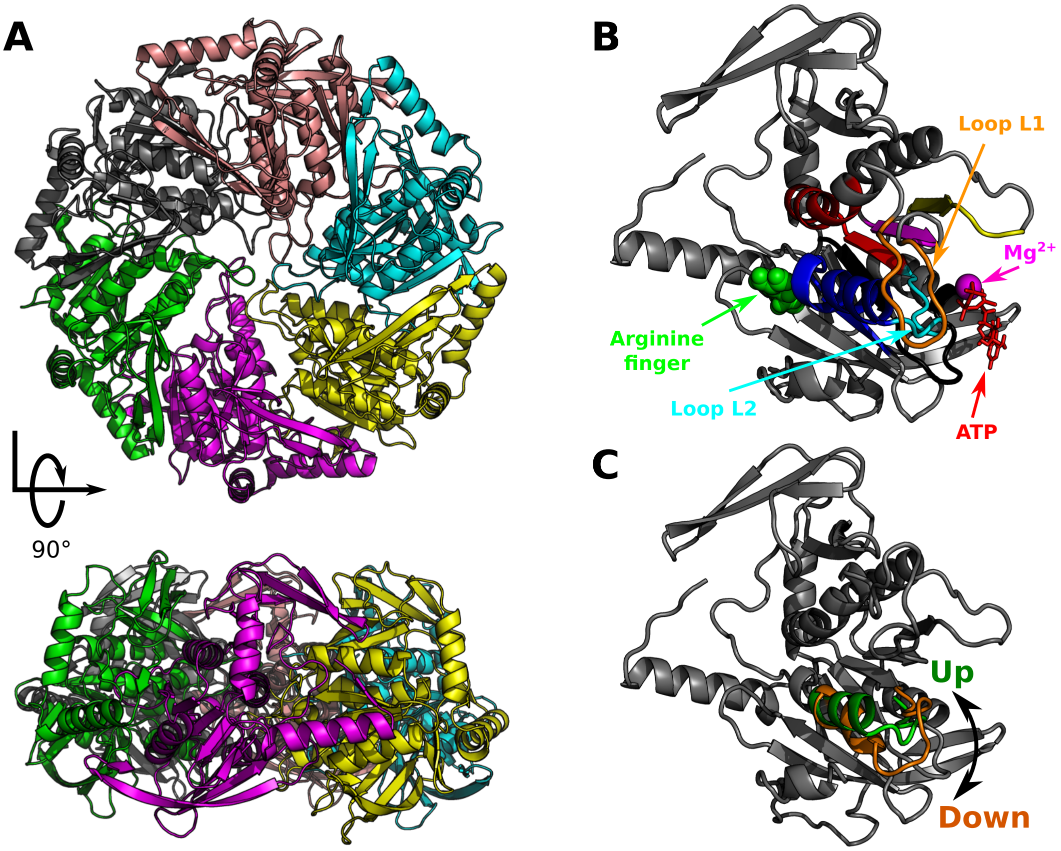


Figure 1-2*: The overall fold of P4 proteins.*
The structure of φ12 P4 is shown as an example (PDB access code 4BLR, (18)). (A) Top and lateral view of the helicase. (B) Structural conservation of the helicase motifs in P4 proteins. H1, H1a, H2, H3, H4, L1 and L2 are labelled in black, yellow, magenta, red, blue, orange and cyan, respectively. The arginine finger, the Mg^{2+} ion and the ATP analogue (AMPcPP) are highlighted in green, magenta and red, respectively. The L1 loop was modelled (C) Two positions of the lever and L2 observed in the X-ray crystal structure of φ12 P4. The “up” and “down” states are respectively depicted in green and orange.

The primary function of P4 requires the coordination of the six levers that allow P4 to walk along the strand. The cross-talk between arginine finger and the lever form the basis of cooperativity between subunits. On the other hand, RNA binding to at least three consecutive subunits is required to induce cooperativity (22). Based on kinetic and structural observations, a model was proposed to describe the coordination of ATP hydrolysis between three consecutive subunits i-1, i and i+1 (Figure 1-3). Initially, stochastic motion of L2 loop in subunit i-1 inserts the arginine finger into the catalytic site of subunit i, thus triggering ATP hydrolysis. Hydrolysis at subunit i switches down the lever, drags down RNA attached to it and triggers insertion of the arginine finger. Since RNA is also bound to the (i+1) lever, motion of the (i) lever pulls down the (i+1) $\alpha 6$ helix, which also promotes formation of the transition state at subunit (i+1).

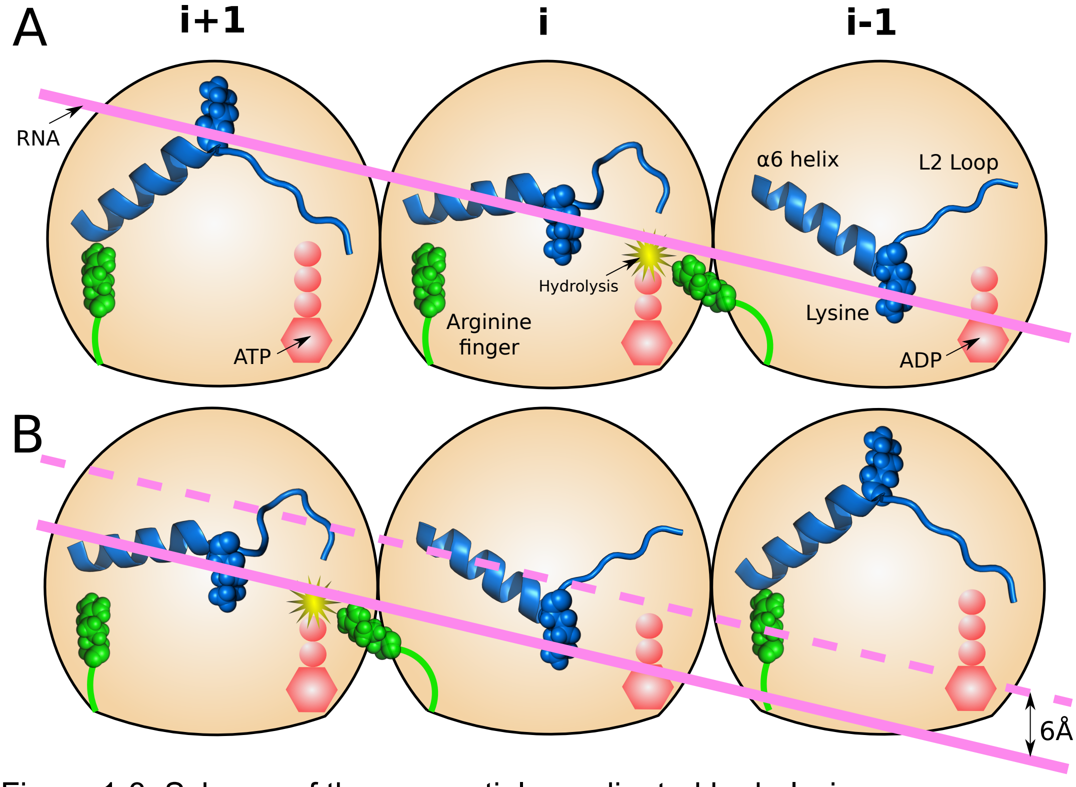


Figure 1-3: Scheme of the sequential coordinated hydrolysis.

Three consecutive subunits are viewed from within the central channel. The lever is represented in blue, the nucleotides in red, the arginine finger in green and the bound RNA in magenta. Upon hydrolysis of subunit (i-1), its lever switches to the "down" position and move the arginine finger into the catalytic site of subunit i. The RNA bound to the lysine is also dragged down, which stabilises the transition state of subunit i. (B) After the hydrolysis of subunit I (yellow star), the transition state of subunit i+1 is stabilised, while a new nucleotide can bind to subunit i-1, which is therefore switched to the "up" position. Modified from (24).

Despite the extensive characterisation of the P4 motors, many questions remain to be answered. For instance, the mechanisms of RNA loading and sequential ATP hydrolysis activation by RNA remain elusive. Comparing the X-ray crystal structures of the system in different states provides a first glimpse of P4 machinery. However, these phenomena involve rapidly interconverting species that are transiently populated or remarkably dynamic. Understanding the structural dynamics of the motor is crucial to bridge the gap between the static structure of the motor and its function. The link between the global dynamics of proteins and their function is discussed in Section 1.2.

1.2 Functional dynamics of proteins

1.2.1 Link between function and dynamics

Proteins are not static objects and their behaviour cannot be accurately described based on information from one rigid structure (25). They are intrinsically dynamic systems that undergo conformational changes driven by energy exchanges with the surrounding solvent (or ligands) (26). Protein dynamics is characterised by the time-scale of the fluctuations, as well as their amplitude and directionality (27). Local fluctuations occur at fast time-scales (ps-ns), whereas large-scale motions are slower due to their collective nature. These different dynamics are unified in the concept of the free-energy landscape of protein (Figure 1-4). Both the thermodynamic and kinetic properties of a protein can be inferred from its free-energy landscape. Briefly, the probability p_i that the protein adopts at state *i*, is directly related to its associated free-energy, F_i, through the Boltzmann distribution law:

$$p_i = \frac{e^{-F_i/k_B T}}{Z}$$

where $Z = \sum_i e^{-F_i/k_B T}$ is the partition function which normalises the probability, *T* is the temperature and k_B is the Boltzmann's constant. Hence, the relative population between two states *i* and *j* adopted by the protein depends on the free-energy difference $\Delta F_{ij} = F_i - F_j$. The averaged time, $\langle\tau\rangle$, required for the protein to move from state *i* to state *j* is related to the free-energy barrier separating the two states, $\Delta\mathrm{F}^{\ddagger}$, such as:

$$\langle\tau\rangle \propto e^{\Delta\mathrm{F}^{\ddagger}/k_B T}$$

The higher the free-energy barrier is, the slower the transition is. Proteins are high dimensional objects and their associated free-energy landscape has as much dimensions. The free-energy landscape can be projected along a one-dimensional reaction coordinate to facilitate its analysis and characterisation. However, the dimensionality reduction has to be rigorous in order to preserve the underlying properties of the protein (28, 29). More explanations about free-energy landscape of proteins can be found in Refs (30, 31).

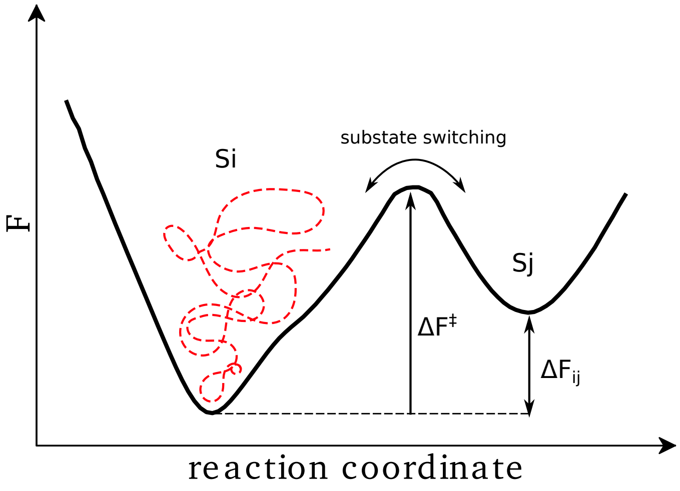


Figure 1-4: *One-dimensional schematic free-energy landscape of a protein.*
In this example, the two substates of the protein are noted S_i and S_j. A state is defined as a minimum in the energy surface, whereas a transition state corresponds to the maximum between two basins. Driven by the thermal noise, the protein diffuses on the free-energy surface. Local and fast fluctuations enable the protein to explore the basin into which it is trapped (red dashed lines). Switching between two states separated by a high free-energy barrier ($\Delta F^{\ddagger}$) is a rare event that requires larger collective motions. At equilibrium, the relative population between two states depends on the difference between their minimum of energy (ΔF_{ij}).

It is now well accepted that protein dynamics play an essential role in their function. Myoglobin is a classic example to illustrate the concept of conformational substates, i.e. that the native conformation of a protein comprises many slightly different conformers. The protein contains heme which delivers oxygen in muscle cells (Figure 1-5A). Rebinding kinetics of carbon monoxide and dioxygen at different temperatures revealed four different rebinding processes (25). It has been shown that the first rebinding process has non-exponential rebinding kinetics at low temperature. Such a kinetics implies that its associated free-energy barrier presents a spectrum of activation energies, instead of one discreet barrier. The spectrum of free-energy barriers was explained by the existence of an ensemble of conformational states for myoglobin.

Enzymes are also a good illustration of the link between structure and dynamics. It is now well established that the flexibility of the catalytic site, needed to align the catalytic groups in their correct orientations, plays an essential role in enzyme activity (32–34). It was even suggested that the intrinsic dynamics of the catalytic site was controlling the turnover rate of the reaction (35). Finally, the most intriguing example constitutes the class of intrinsically disordered proteins (Figure 1-5B). Such proteins, although disordered, fold upon binding to their biological targets or play an essential role in the assembly of molecular complexes by forming flexible linkers (36, 37).

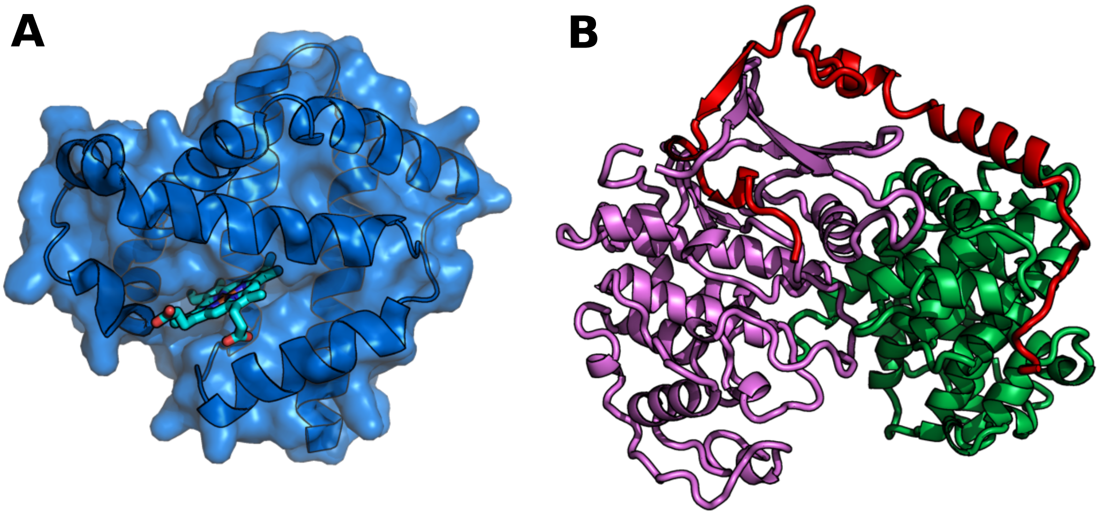


Figure 1-5: *Structure of proteins for which dynamics is important for their function.*
(A) Structure of myoglobin. Its heme is represented with sticks (PDB Ref 3RGK). (B) Structure of the intrinsically disordered protein P27 (red) bound to cyclin A (green) and Cdk2 (violet) (PDB Ref 1JSU). When unbound, P27 is mainly disordered.

Throughout this thesis, the notion of protein dynamic will refer to the ensemble of conformations adopted by the protein, i.e. the “flexibility” of the protein, and the kinetic aspects will be ignored. The protein will be treated as an ensemble of structures that will be produced by molecular dynamics simulations as explained in the next section. The notion of free-energy landscape will be used in Chapter 5 in order to characterise and compare different ensembles of structures.

1.2.2 Molecular dynamics simulations to study protein dynamics

Molecular dynamics (MD) simulations provide a link between structure and dynamics by giving a picture of the conformational space visited by the protein (38). Assuming that the system is ergodic, its conformational space can be theoretically sampled by running a molecular dynamics simulation. Whereas the first MD simulation of a protein was 9.2 ps long and was performed on a small system in vacuum (39), advances in computer hardware and software have made possible to produce millisecond trajectories of proteins in explicit solvent (28, 40, 41). Simulations provide a detailed atomistic view of the time evolution of the system that cannot be reached by any current experimental technique. However, they remain computationally expensive and in practice simulations of macromolecules (enumerating millions of atoms) rarely exceed a few hundreds nanoseconds. Such trajectories are way too short to explore the different states of large molecular-complexes, for which relevant time-scales easily extend to milliseconds.

Various methods have been developed to fill the gap between the femtosecond time-step of simulations - necessary to maintain the stability of the integration - and longer time-scales relevant for biological processes (42). One approach consists of smoothing the force-field to lower the free-energy barrier between the different states, which eventually accelerates the exploration of the conformational space (43–45). Another well-established approach, known as the string method, enables to “link” two stable conformational states by searching an optimal transition path and its associated transition state (46, 47). Other powerful techniques, such as transition path and milestoning, have been developed to study the reaction path of biological systems (48–51). Finally, a promising approach consists of coarse-graining the system, i.e. clustering groups of atoms into beads (52). The idea is to reduce the atomic-scale information into a lower resolution model that preserves the relevant physical features of the system. The new system has a less rugged free-energy landscape that can be explored faster. In addition, due to the reduction of the number of atoms, with equivalent computational resources, longer time-scales can be explored. Some of the methods mentioned above have been applied to hexameric helicases in order to investigate their mechanisms. Ma *et al.* addressed the question of how the free-energy stored in the Rho motor during hydrolysis drives the lever motion by using path sampling techniques (53). Other groups constructed coarse-grained models to investigate the translocation motion of hexameric helicases (54, 55).

Another problem for MD simulations is their accuracy. Indeed, simulations are reliant on models – known as force fields – of the physics underlying protein dynamics. The majority of force-fields are empirical models, which suffer from approximations (56–58). Hence, to date, the quality of MD simulations of proteins, especially large proteins, should always be evaluated with experimental data.

Throughout this Thesis, computation is used to push the interpretation of experimental data further. In order to reproduce the dynamics of the system as accurately as possible, full-atomistic models with explicit solvent are used. Since large macromolecular complexes are handled, limited timescales are achieved. MD simulations of a few hundreds nanoseconds are performed for each system of interest to sample their local conformational space. It is worth noting that even with timescales of two more orders of magnitude, the length of the trajectory (i.e. ~10 μs) would remain way too short to explore extensively the conformational space of the protein for which relevant timescales are at least ~1 ms.

1.3 Experimental techniques to probe structural dynamics of proteins

It is not possible to directly observe the motion of all atoms within the protein at the same time. Instead, structure and dynamics are inferred from the measure of macroscopic physical properties of the systems. As mentioned above, proteins exhibit dynamics with different amplitudes of motion involving divers spatial-scales that different experimental techniques can probe (Figure 1-6).

1.3.1 Different techniques for different spatial-scales

NMR is traditionally a powerful method to study the fast dynamics of proteins (59–62). By probing the relaxation properties of spins along the backbone and in the side chains, or the hydrogen-deuterium exchange of amide hydrogen (HDX), NMR provides information on the local environment of each residue. Time resolved X-ray crystallography is a promising technique to investigate fast reactions at high resolution (63, 64). Both Small-angle X-ray scattering (SAXS) and single-molecule Förster resonance energy transfer (FRET) are well-established techniques to obtain larger-scale information about protein structure and dynamics (65, 66). Cryo-electron-microscopy (EM) is a promising technique with an intermediate spatial-resolution that continuously improves (67) and can deal, to some degree, with protein flexibility by observing single macromolecular-complexes (68).

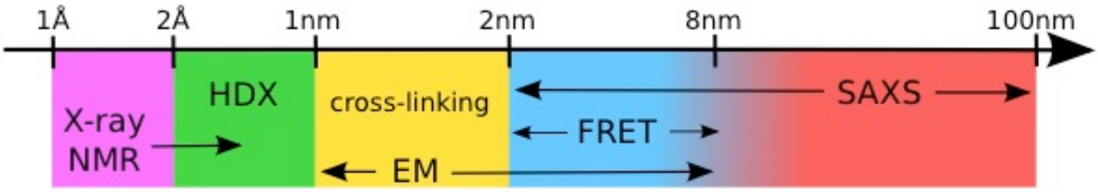


***Figure 1-6**: Resolution of the information carried by different popular experimental techniques.*

In this Thesis, mechanisms of P4 were investigated using hydrogen-exchange probed by mass-spectrometry (HDX-MS) and FRET (69, 70). These two experimental techniques are described bellow as well as the models used to interpret their macroscopic data in terms of microscopic information.

1.3.2 Hydrogen-deuterium exchange probed by mass-spectrometry

Principles of hydrogen-deuterium exchange

A powerful technique to investigate the dynamics of proteins is hydrogen-deuterium exchange. The method is based on the spontaneous exchange of the amide hydrogens of the protein with deuterium from solvent containing deuterium oxide (^2H_2O) and has been extensively used to investigate protein folding (71–74). Key to interpreting HDX kinetics is the fact that exchange occurs faster for amides that are solvent-exposed and not involved in hydrogen bonds. Deuterium incorporation has been measured using NMR with residue level resolution for small proteins (75). At neutral pH the exchange is fast for solvent exposed amides while hydrogen bonding, e.g. within helices or β-sheets, slows it down. When fully exposed, the exchange kinetics of the amide is governed by an intrinsic rate, k_{int}, that depends on the temperature, solution pH and side chains of the two neighbouring residues (see Section 2.2.2). Within a folded protein, the exchange of amide hydrogen requires local "opening" of the structure and can be approximated as a two-step process (76):

$$\boldsymbol{NH_{cl} \xleftrightarrow{k_{cl}/k_{op}} NH_{op} \xrightarrow{k_{int}} ND_{op}}$$

(1-1)

where k_{cl} and k_{op} are the local "closing" and "opening" rates. The observed deuterium uptake rate, k_{obs}, can be expressed as:

$$\boldsymbol{k_{obs}} = \frac{\boldsymbol{k_{int}k_{op}}}{\boldsymbol{k_{int}} + \boldsymbol{k_{op}} + \boldsymbol{k_{cl}}}$$

(1-2)

Two limiting regimes, called EX1 and EX2, are invoked in interpreting HDX kinetics of proteins. For both regimes, the protein is considered to be in native conditions, i.e. $k_{cl} \gg k_{op}$. In the EX1 limit $k_{int} \gg k_{cl}$ implies that the amide exchanges as soon as it becomes exposed to solvent, i.e., $k_{obs} = k_{op}$. In this regime the exchange is limited by slow conformational changes that are usually associated with global unfolding (77) or cooperative changes in quaternary structure (78). In the EX2 limit, $k_{cl} \gg k_{int}$,

$$\boldsymbol{k_{obs}} = \frac{\boldsymbol{k_{int}}}{\boldsymbol{P}}$$

(1-3)

where $P = k_{cl}/k_{op}$ is a protection factor for the particular amide. The EX2 limit governs exchange under native conditions and is sensitive to local stability. In the EX2 regime the kinetics is sensitive to pH (through k_{int}).

Structural interpretation

The link between the protection factor of a residue and its structural dynamics is not straightforward. Hence, interpretation of HDX is often assisted by computational methods. These methods are based on estimation of protection factors either by calculating the difference of free energy between the open and closed states, $\Delta G^0 = RT \ln P$ (79–81) or by relating the protection factor to the local environment of the residue (74, 82–84). Solvent accessibility is generally used to predict the exchange-competence of a residue. Although a strong correlation exists between protection factors and solvent accessibility, many residues located at the protein surface (i.e. totally solvent exposed) exhibit exchange rates much slower than their intrinsic rates (82, 85). It is now well established that exchange of amide hydrogens also requires the breaking of hydrogen bonds formed with the side chains or the protein backbone (86). In the EX2 regime, the protection factor of an amide hydrogen of residue *i* can be approximately estimated from the structure of the protein using the phenomenological equation (74, 83):

$$\ln P_i^{sim}(X) = \beta_c N_i^c(X) + \beta_h N_i^h(X)$$

(1-4)

where *X* is a particular conformation of the protein, $N_c(X)$ and $N_h(X)$ is the number of contacts between non-hydrogen atoms and the number of hydrogen bonds to the amide hydrogen, respectively. In this approximation, hydrogen exchange rate is governed primarily by the burial of the amide within the hydrophobic core or subunit interface and by participation in secondary structure. The phenomenological approximation in Equation (**1-4)** can be used to predict or attempt interpretation of experimental HDX data from a single protein structure. In doing so, however, one neglects thermal fluctuations and conformational heterogeneity that contribute to the H/D exchange (25, 87). Assuming the validity of Equation (**1-4)** protection factors should then be estimated as an ensemble average, in an equilibrium molecular dynamics simulation. If residue *i* contains an amide hydrogen, the averaged protection factor, $\langle P_i \rangle$, is defined as:

$$\boldsymbol{\langle P_i \rangle = \frac{1}{|E|} \sum_{X \in E} P_i(X)}$$

(1-5)

where *E* is the ensemble of conformers adopted by the protein in the MD simulation.

Probing the exchange kinetics by mass spectrometry

For larger proteins and their complexes, NMR cannot probe their hydrogen-deuterium exchange. Instead, detection of hydrogen-deuterium exchange by high-resolution mass spectrometry (MS) has emerged as an alternative (77, 88–90). HDX-MS relies on the measurable difference of mass between the deuterated and non-deuterated polypeptide chain. The protein is fragmented by proteolysis before analysis by mass spectrometry (Figure 1-7).

Fragmentation is made at low pH and low temperature to reduce back-exchange and preserve the isotopic pattern, even under non-native conditions. Although the exchange process is significantly slowed down under these conditions, forward exchange can occur during proteolysis, where solvent contains mostly heavy water. On the contrary, during fragment-separation – a necessary step for MS measurement – the solvent is non-deuterated and deuterated residues can back-exchange. Nevertheless, the residual forward and back-exchange is readily corrected for (91, 92) allowing determination of region specific exchange patterns (usually covering 10-20 amino acid segments) (93). The deuterium incorporated into the side chain groups is rapidly back exchanged. As a consequence, HDX-MS is only sensitive to the backbone amide hydrogen. Recent advances in mass spectrometry (e.g. electron capture dissociation (94)) and development of in-line proteolysis (95) suggest that HDX-MS/MS can be used to measure hydrogen exchange at single residue resolution. However, the required uniform coverage and resolution of isotopic envelopes may be hard to achieve for larger proteins and multi-protein assemblies (96). Monitoring of deuterium incorporation for each fragment over time yields exchange kinetics. Exchange profiles contain information about local and global stability averaged over all amide NH groups within the fragment. When the exchange is probed by MS, the information is averaged over a segment instead of being residue-specific. The first attempt of this Thesis will be to develop a new method to support the structural and dynamic interpretation of HDX-MS data (see Chapter 3). It is worth noting that no HDX-MS data were collected during the Thesis and all handled experimental HDX-MS data were already published.

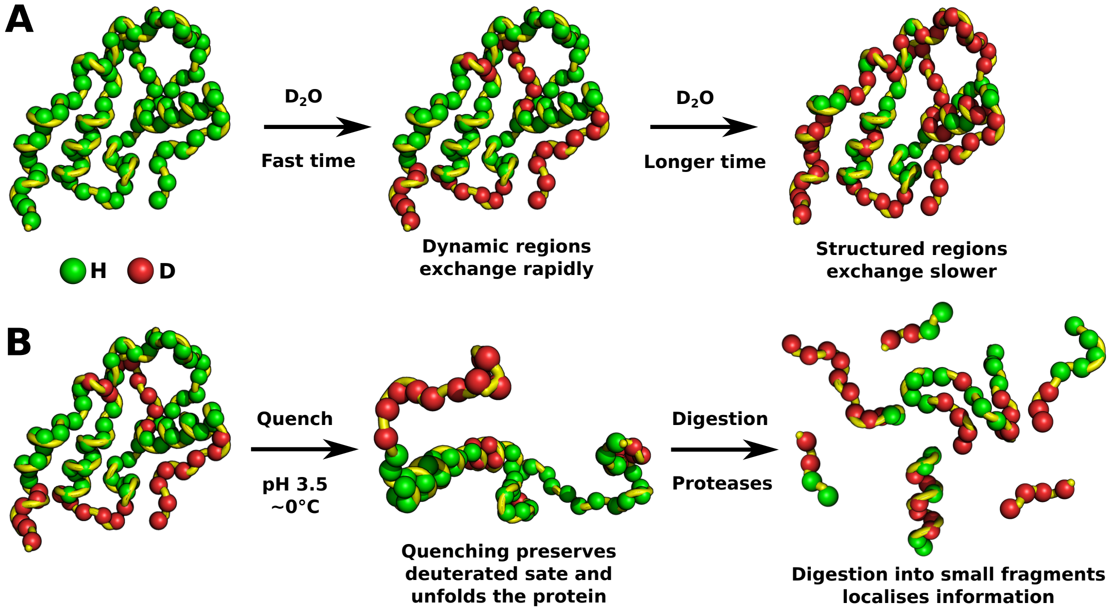

A
D_2O
Fast time
D_2O
Longer time
H
D
Dynamic regions exchange rapidly
Structured regions exchange slower
B
Quench
pH 3.5
~0°C
Digestion
Proteases
Quenching preserves deuterated sate and unfolds the protein
Digestion into small fragments localises information

Figure 1-7: *Hydrogen-deuterium exchange probed by mass-spectrometry.*
Spheres represent the amide hydrogen along the backbone. Protons are coloured in green and deuteriums in red (A) The non-deuterated protein is mixed with a deuterated buffer. The exposed and dynamics regions exchange faster than the buried and structured ones. (B) After a given time, protein is quenched at low pH and temperature to stop the exchange process.

1.3.3 Fluorescence spectroscopy

FRET is a popular technique to measure the distance and the fluctuations between two residues (97). FRET relies on the excitation of a donor, which relaxes to its ground state by transferring its energy to a nearby acceptor (Figure 1-8A and B). The energy transfer results from a dipole-dipole interaction and is therefore non-radiative. The rate of energy transfer between the donor and the acceptor is given by:

$$k_T = k_D \left(\frac{R_0}{r}\right)^6$$

(1-6)

where R_0 is the Förster distance, k_D is the radiative rate of the donor in the absence of the acceptor and r the distance between the donor and acceptor. The Förster distance depends on multiple factors:

$$R_0 = \frac{9000\,(\ln 10)\kappa^2 Q_D \mathrm{J}}{128\pi^5 n^4 N_A}$$

where *J* is the overlap integral between the donor emission ad acceptor absorption spectra (Figure 1-8C), Q_D is the donor's fluorescence quantum yield, *n* the refractive index of the medium between the dyes and N_A the Avogadro's constant. The factor κ^2 depends on the relative orientation of the chromophores:

$$\kappa = \overrightarrow{\mu_D} \,.\, \overrightarrow{\mu_A} - 3\,(\overrightarrow{u_{DA}} \,.\, \overrightarrow{\mu_D})(\overrightarrow{u_{DA}} \,.\, \overrightarrow{\mu_A})$$

where $\overrightarrow{\mu_D}$ and $\overrightarrow{\mu_A}$ are the unit vectors in the directions of the donor and acceptor dipoles, respectively. $\overrightarrow{u_{DA}}$ is a unit vector in the direction D to A. If the two dipoles are orthogonal, κ^2 is equal to 0, whereas the transfer is maximal when the dipoles are collinear ($\kappa^2 = 4$). When the orientations of the two dyes are isotropic, the averaged κ^2 is equal to 2/3. The fraction absorbed photons that are transferred, without radiation, to the acceptor is called the transfer efficiency, *E*:

$$E = \frac{k_T}{k_T + k_D}$$

(1-7)

Substituting Equations (1-6**)** and (1-7**)** yields:

$$E = \frac{R_0^6}{R_0^6 + r^6}$$

(1-8)

Equation (1-8**)** provides a direct link between efficiency and the distance between the two dyes (Figure 1-8D). Therefore FRET can be used as a "molecular ruler". It is worth noting that the variations of the efficiency are more significant around the Förster distance. Distances measurable by FRET are typically ~2-8nm. In a typical FRET experiment, the efficiency is measured by calculating the ratio between the number of photons emitted by the acceptor and the total number of photons emitted by the donor and acceptor.

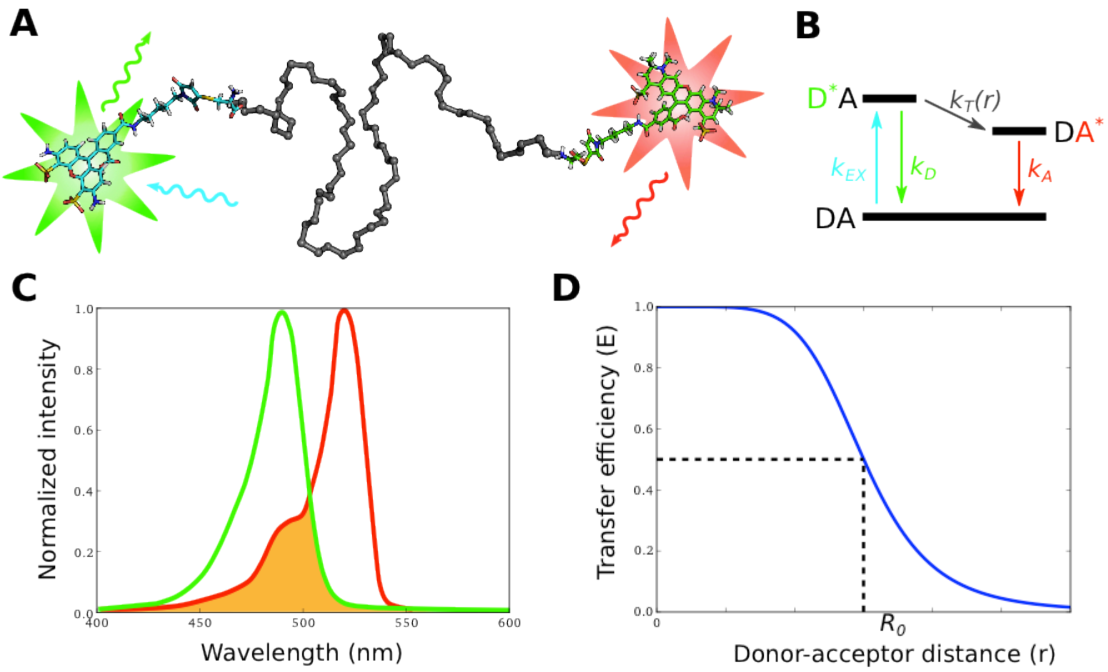


Figure 1-8: *Förster resonance energy transfer (FRET).*
(A) A molecule labelled with donor (Alexa-488) and acceptor (Alexa-594) dyes. The donor (green) is excited with a laser (cyan arrow). The excited donor relaxes by either emitting a photon (green arrow) or via dipole-dipole interaction with the acceptor (red). In the last case the acceptor will relax by emitting a photon (red arrow). (B) Simplified Jablonski diagram of FRET that illustrates the transitions between the ground and excited state of the donor (D) and acceptor (A). (C) Normalized emission (green) and absorbance (red) spectra of the donor and acceptor, respectively. The overlapping area of the two spectra is filled in orange. (D) FRET efficiency as a function of the distance between the two dyes. At high distances, E~0, whereas E~1 at short distances.

1.4 Overview and aims of the Thesis

In this Thesis, molecular dynamics simulations are combined with sparse experimental data to investigate mechanisms of the packaging motors P4. Before discussing the main finding of my research, I will provide in Chapter 2 a concise explanation of the background required to understand the methods and techniques. In Chapter 3, a new method to interpret quantitatively hydrogen-deuterium exchange probed by mass spectrometry (HDX-MS) data is presented. The method is tested with the packaging motor φ12 P4. This system is an ideal test case since its X-ray crystal structure is available, as well as previously published HDX-MS data. I show that a ~100 ns simulation is sufficient to predict accurately the experimental HDX kinetics of the system. The approach also turns out to be a valuable tool to validate the assignment of the fragments and to assess structural models. This work has been published in (98). At the end of the chapter, the limits of the model used to predict the protection factors of residues are discussed and a more accurate model is suggested. In Chapter 4, the method is applied on the packaging motor φ8 P4, for which a crystal structure was published recently. HDX-MS data of the system had also already been published. I reinterpreted quantitatively the HDX kinetics to investigate the RNA loading mechanisms of the motor. To gain information about the structural conformational changes occurring upon RNA binding, single-molecule Förster resonance energy transfer (smFRET) experiments were undertaken. Based on experimental and computational observations, I present a new model to explain the modulation of RNA affinity in φ8 P4. In Chapter 5, I investigate the structural and dynamic information carried by sparse data from different popular experimental techniques, such as HDX-MS, smFRET, but also ion-mobility cross-section, NMR and small-angle X-ray scattering. Since smFRET and HDX-MS data are found to carry too little information alone, I tested whether combining these two techniques helps to restrain the ensemble. The results suggest that the two techniques carry complementary information. In the final chapter, all the results are reviewed and future prospects are suggested.

Chapter 2: Theory, Materials and Methods

In this chapter, the theoretical background of molecular dynamics simulations is briefly introduced, as well as the tools used to predict protection factors and intrinsic exchange rates for residues. The different analytical methods used throughout the thesis are also described. The protocols used for the molecular biology work and the fluorescence spectroscopy techniques are described at the end of the chapter.

2.1 Molecular dynamics simulations

2.1.1 Integration of the empirical energy function

Molecular dynamics simulations were performed using the freeware molecular dynamics package NAMD (NAnoscale Molecular Dynamics program, (99)). Given the initial coordinates, *r(0)*, and velocities, $\dot{r}(0)$, of the biomolecular system, the detailed time-evolution of its coordinates can be calculated by solving Newton's equation of motion:

$$m\ddot{r}(t) = F(r,t) = -\nabla V(r,t)$$

where *V* is the energy function. NAMD numerically integrates the equation using the popular Verlet algorithm (100). To understand the basic idea of the Verlet integration, let's write the third-order Taylor expansions for the coordinates of the system one step forward and one step backward:

$$r(\, t + \Delta t) = r(t) + \, \dot{r}(t)\Delta t + \frac{1}{2}\ddot{r}(t)\Delta t^2 + \frac{1}{6}\dddot{r}(t)\Delta t^3 + \mathrm{O}(\Delta t^4)$$

$$r(\, t - \Delta t) = r(t) - \, \dot{r}(t)\Delta t + \frac{1}{2}\ddot{r}(t)\Delta t^2 - \frac{1}{6}\dddot{r}(t)\Delta t^3 + \mathrm{O}(\Delta t^4)$$

Adding the two expressions gives:

$$r(\, t + \Delta t) = 2r(t) - r(\, t - \Delta t) + \ddot{r}(t)\Delta t^2 + \;\mathrm{O}(\Delta t^4)$$

Hence, one can calculate the next position vector from the previous two and the instant acceleration of the system. The latter is obtained by evaluating the potential energy of the system. Although velocities are not required to propagate the positions towards the next time step, they are usually calculated in order to estimate the total energy of the system. It will be shown later that velocities are also useful for the thermostat. The potential energy was calculated using the CHARMM36 force field (101). The force field can be decomposed into two terms (102):

$$V(r) = \; V_{bonded}(r) + V_{non-bonded}(r)$$

The first term refers to the covalent interactions and is given by the following summation:

$$\begin{aligned} V_{bonded} &= E_{stretch} + E_{bend} + E_{dihedral} + E_{improper} \\ &= \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 \\ &+ \sum_{angles} k_\phi[cos(n\phi + \delta) + 1] + \sum_{bonds} k_\omega(\omega - \omega_0)^2 \end{aligned}$$

The $E_{stretch}$ term describes the potential of the harmonic vibration of a covalent bond, where $b - b_0$ is the deviation from the equilibrium bond length and k_b the bond force constant (1-2 interactions). The E_{bend} term describes the angular vibrational motion occurring between three bonded atoms (1-3 interactions). Also the E_{bend} term can be modelled by a quadratic function where θ_0 is the equilibrium angle and k_θ the angular force constant. $E_{dihedral}$ accounts for the torsional force between two atoms separated by three covalent bonds (1-4 interactions). The integer n indicates the periodicity. The improper term, $E_{improper}$, enforces the planarity of chemical groups (e.g. planarity of rings or chirality of atoms). The non-boned term of the force field is a combination of two terms:

$$\begin{aligned} V_{non-bonded} &= E_{VdW} + E_{coulomb} \\ &= \sum_{non-bonded} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{D r_{ij}} \end{aligned}$$

The E_{VdW} represents the van der Waals energy interactions and is modelled by the Lennard-Jones 12-6 potential (103). E_{VdW} is a sum of repulsive and attractive interactions. The repulsive term is due to overlapping of electron orbitals and occurs at very small distances. The attractive term captures long-range van der Waals type dispersion forces induced by instantaneous dipoles. σ_{ij} is the distance at which the potential is zero and corresponds to the sum of the van der Waals radii of atoms *i* and *j*. ϵ_{ij} is the depth of the energy well. The final term $E_{coulomb}$ represents the Coulombic potential. *D* denotes the dielectric constant. The $E_{coulomb}$ term is computed using the particle mesh Ewald method, which divides the potential into two parts:

$$E_{coulomb} = E_{short-range} + E_{long-range}$$

The short-range contribution is calculated by summing explicitly the pair interactions of local atoms, whereas the long-range contribution is calculated in the reciprocal space, where the sum converges faster. Such a method requires using boundary conditions. The advantage of this method is its computational efficiency. A drawback of this technique is that the system has to be sufficiently large to avoid artefacts when repeated periodically (5).

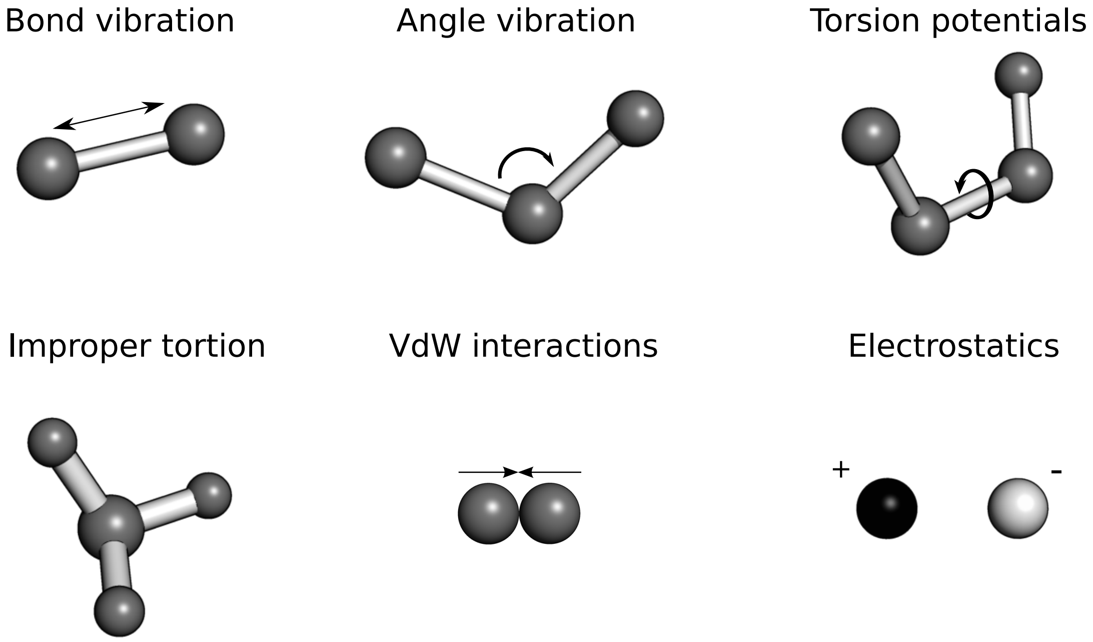


Figure 2-1: *Schematic view of force field interactions.*
Hard spheres and heavy solid lines indicate the atoms and covalent bonds, respectively.

In CHARMM36, hydrogen atoms are modelled explicitly but are simply linked to heavy atoms with springs. Parameters for bond terms are fit to crystallographic and spectroscopic data. Torsion and angle parameters, as well as σ_{ij} and ϵ_{ij} are refined using quantum- and molecular- mechanical calculations and NMR data (101, 104). All simulations were run in parallel with up to 256 CPUs.

2.1.2 Thermostat

The coupling of the system to a heat bath of a specific temperature was simulated by Langevin dynamics. In Langevin dynamics, Newton's equation of motion is modified by introducing damping and random forces:

$$m\ddot{r}(t) = \nabla V(t) - \gamma m\dot{r}(t) + f(t)$$

where γ is the friction coefficient, and f a random force which accounts for collisions between atoms of the simulated system and the virtual heat bath. The force is called random because it is assumed that $\langle \dot{r}(t), f(t) \rangle = 0, \ \forall t$. It has a Gaussian probability distribution with zero mean and variance, σ. The fluctuation dissipation theorem (105) gives:

$$\sigma = \sqrt{2m\gamma k_B T / \Delta t}$$

where T is the temperature of the system, k_B is the Boltzmann constant and Δt the time-step used in MD to integrate the equations of motion. Hence, adjusting the dispersion of the random force intensity allow the temperature to be controlled. Throughout the Thesis, the value of γ was fixed to 1 ps^{-1}.

2.1.3 Explicit solvent model

Protein hydration is crucial for their structure, dynamics and activity. Thus, it is important to model accurately the solvent. Water was modelled explicitly with the three-site model TIP3P (106). This model has positive charges on the hydrogens and a negative charge on the oxygen. Water models are parameterized to reproduce water density, self-diffusion, radial distribution function, dielectric constant and enthalpy of vaporization observed experimentally (107) . TIP3P model matches well the two last properties but reproduces poorly the density and the self-diffusion rate (108).

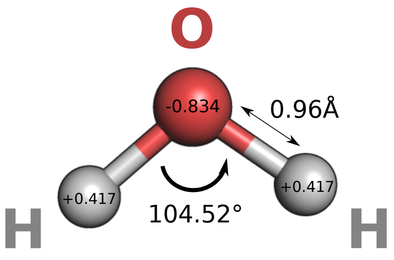


Figure 2-2 *TIP3P geometry and partial charges.*

2.1.4 Modeller

Missing segments of PDB structure were modelled as loops using MODELLER (109). Briefly, MODELLER generates an initial loop by placing uniformly the residues along the line connecting the N- and C-terminal anchor regions. Atoms are then randomly displaced to generate ~500 randomized initial structures. The structure with the lowest energy in the CHARM22 force field is selected for further optimization. The system is first relaxed by conjugate gradients minimization, followed by simulated annealing. For each structural model, 50 different models were generated with MODELLER and the structure with the best score was used as initial conditions for MD simulations.

2.2 Prediction of hydrogen-exchange kinetics

The exchange kinetic of a residue was predicted using the definition:

$$k_{obs} = \frac{k_{int}}{P}$$

where k_{int} and P are the intrinsic rate and protection factor of the residue, respectively (see Section 1.3.2). A detailed description of the calculation of these two quantities is given bellow.

2.2.1 Approximation of protection factors

The protection factor of each residue *i* in conformation *X* was computed with CHARMM (110), based on the phenomenological approximation:

$$\ln P_i^{sim}(X) = \beta_c N_i^c(X) + \beta_h N_i^h(X) \tag{2-1}$$

where N_i^c and N_i^h are the number of contacts with heavy atoms (i.e. non-hydrogen) and hydrogen bounds residue *i* is involved in. β_c and β_h parameters were calculated to simultaneously optimize prediction of a set of 7 proteins for which experimental protection factors were available (83). Cutoffs for atom contacts and hydrogen bonds are 6.5 Å and 2.4 Å, respectively. In CHARMM, the cut-off function is smoothed as such:

$$N_i^c = \sum\nolimits_{j \in H_i} \frac{1}{1 + \exp\left(5 * \left(r_{ij} - 6.5\right)\right)}$$
$$N_i^h = \sum\nolimits_{j \in O_i} \frac{1}{1 + \exp\left(10 * \left(r_{ij} - 2.4\right)\right)}$$

(2-2)

where r_{ij} is the distance between the amide hydrogen of residue *i* and atom *j* (in Angstroms). H_i is the list of heavy atoms, i.e. all atoms except protons, which are not part of residues *i-1*, *i* or *i+1*. O_i is the list of all oxygens of the system not included in residues *i-1*, *i* or *i+1*. It is worth noting that directionality of hydrogen bonds is ignored. The protection factor, *P*, of a residue was averaged over the conformational space sampled during the molecular dynamic simulation:

$$\langle P \rangle = \frac{1}{N}\sum_{i=1}^{N} P_i$$

where *N* is the total number of frames constituting the trajectory.

2.2.2 Intrinsic rate calculations

The hydrogen-exchange intrinsic rate*, k_{int}*, of a residue is the exchange rate of its amide hydrogen in a random coil. The dependency of k_{int} on temperature, pH and neighbouring side chains has been well characterized (111, 112). The exchange is catalysed primarily by water ions, leading to high dependency on pH. The intrinsic rate is expressed as:

$$k_{int} = k_{acid}\,10^{-pD} + k_{base}\,10^{pD-pK_D} + k_w$$

(2-3)

where pD is the value read on the glass-electrode pH meter incremented by 0.4 to take into account isotopic effect on the pH meter (i.e. pD=pH$_{\text{read}}$+0.4).

K_D is the D_2O dissociation constant and k_{acid} , k_{base} and k_w the second order rate constants for catalysis by D_3O^+, DO^- and D_2O, respectively.

The rate constants for catalysis have been estimated using a poly alanine peptide at 20˚C and were used as a reference. The temperature effects are accounted for by adjusting the reference rates using the Arrhenius law:

$$k_i(T) = k_i(293)exp\left(-\frac{E_a^i}{R}\left(\frac{1}{T}-\frac{1}{293}\right)\right) \tag{2-4}$$

where E_a^i is the activation energy of the catalysis by species *i (e.g. OD⁻)*, $k_i(293)$ the rate constant at 293K, *R* is the universal gas constant and T the temperature (in Kelvins).

Molday *et al.* characterised effects of the two neighbouring side chains on the exchange. They showed that their respective effects can either be positive or negative and are additive, i.e. left side chain affects exchange independently of right side chain and *vice versa*. This implies that mechanisms underpinning these effects are not steric, since competition for the same space of neighbouring side chains may not act independently (112). Instead, side chains are thought to stabilize or destabilize the charged intermediate and transition states. Side chain effects are accounted for by multiplying the reference rate constants of poly-alanine by a factor depending only on the side chain and its position (left or right):

$$k_i(L,R) = k_i(Alanine)\varphi_i(L)\rho_i(R) \tag{2-5}$$

where $k_i(Alanine)$ is the rate constant for catalysis by ionic species *i* measured for poly-alanine, $\varphi_i(L)$ is the correction factor for the left side chain, *L* and $\rho_i(R)$ then correction factor for the right side chain, *R*. Based on the same principles developed above, a homemade script was written to estimate intrinsic rates of each residue as a function of pH, temperature and amino acid sequence. All the parameters and correction factors can be found in (112).

2.3 Ensemble refinement

In Chapter 5, I present a protocol to assess the structural information content carried by a physical property of a protein, e.g. its protection factors (see Section 5.3.1). The protocol involves the refinement of an initial ensemble of structures (called the pool) such that the refined ensemble matches the physical property of the protein. In this section, I introduce the genetic algorithm I implemented the refinement procedure. The models used to back-calculate the different physical properties from an ensemble of structures are also described.

2.3.1 Genetic algorithm

The procedure described in (113) was implemented. The genetic algorithm is based on the selection of a sub-ensemble of structures from the pool. This selection is optimized over several generations. L sub-ensembles (chromosomes) were composed of *N* structures (genes) picked from the conformer pool (typically L = 50 and N = 100). In the first generation, sub-ensembles are generated by selecting randomly conformers from the pool. To generate the next generation, chromosomes are submitted to three consecutive operations: *random mutation*, *crossing-over* and *selection* (see Figure 2-3). In *random mutation*, up to 20% of the genes of a chromosome were modified, half were exchanged with others from the pool and the other half with genes from chromosomes of the same generation. The percentage of genes modified in each chromosome was fixed randomly (20% being a maximum), such that this value was sometimes low enough to allow finer optimizations. In *crossing-over*, each chromosome was paired with another chromosome chosen randomly from the same generation and their segments were swapped, with a minimum of two genes transferred to the offspring. Each crossing-over generates two new children, leading to a total of 3L chromosomes. For each chromosome, the average of the different observables (single-molecule Förster resonance energy transfer (smFRET), hydrogen exchange probed by NMR or mass-spectrometry (HDX-MS), small-angle X-ray scattering (SAXS), chemical-shifts (CS), ion-mobility cross-section) were computed and compared with the synthetic experimental data (Equation (**2-6)**). The L chromosomes with the lowest mean square deviation were selected for further evolution, typically for up to 50,000 generations. The algorithm stops prematurely if a perfect match is found, i.e. the mean square deviation reaches zero. The process was repeated 400 times to generate 400 different ensembles made of 100 structures. The weight of each structure of the pool were refined as:

$$w_s = \frac{5{,}000}{400 * 100} N_s$$

where *s* is a structure of the pool, w_s is its refined weight and N_s is the number of times structure *s* appeared in the 400 ensembles. The refined weights are normalized by 5,000/(400*100) such that $\sum w_s = 5{,}000$, as in the conformer pool. The 5,000-structure ensemble with the refined weights is called the refined ensemble.

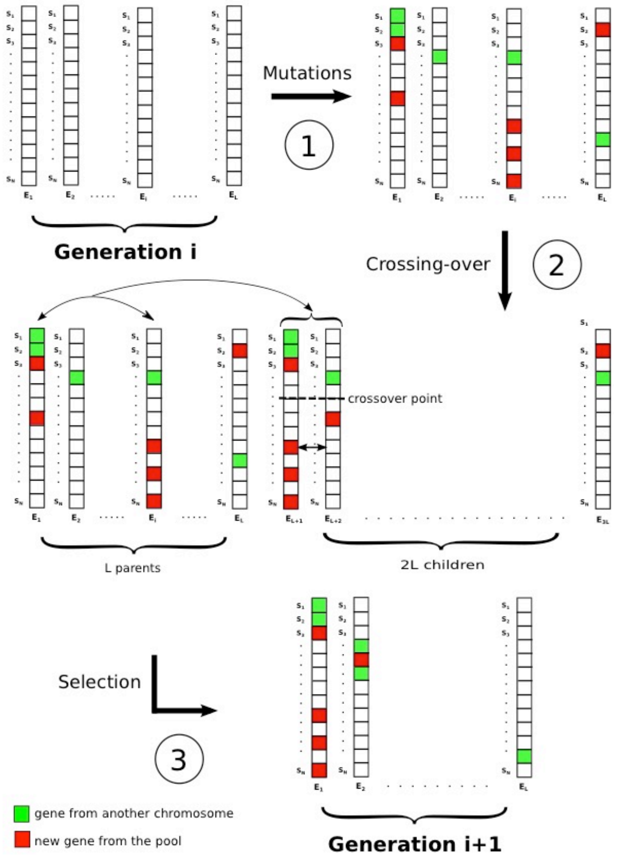


Figure 2-3: *Genetic algorithm scheme.*
The L chromosomes of generation i are first mutated by exchanging their genes with others from the pool or from other chromosomes (1). Each chromosome is then paired with another random one to be crossed-over and generate two children (2). The L chromosomes with the best fitness scores (lowest total mean square deviation) are eventually selected to produce generation i+1 (3). The green and red genes represent structures which have been replaced by a new structure from another chromosome or from the pool, respectively.

2.3.2 Computing of physical properties

HDX data

Protection factors and deuterium fraction were calculated as described in Section 2.2. The mean square deviation (MSD) of protection factors from the reference data was calculated as:

$$MSD_{HDX}(E_{refined}) = \frac{1}{N_{res}} \sum_{r=1}^{N_{res}} \left(\frac{\langle P_r \rangle (E_{reference}) - \langle P_r \rangle (E_{refined})}{\langle P_r \rangle (E_{reference})} \right)^2$$

where $\langle P_r \rangle (E)$ is the averaged protection factor of residue *r* calculated in ensemble *E* and N_{res} is the number of residue in the protein. To account for the lack of accuracy of HDX-MS data, a relative cutoff, ε, was introduced. If the relative error between the reference deuterium fraction and the refined one was bellow ε, the matching was considered to be perfect, i.e. equal to zero. Hence, the MSD of the deuterium fraction from the reference data was defined as:

$$MSD_{HDX-MS}(E_{refined}) = \frac{1}{n} \frac{1}{N_{frag}} \sum_{l=1}^{n} \sum_{j=1}^{N_{frag}} {\chi_j}^2(t_l)$$

with:

$${\chi_j}^2(t_l) = \begin{cases} \left(\frac{D_j(t_l)(E_{reference}) - D_j(t_l)(E_{refined})}{D_j(t_l)(E_{reference})} \right)^2 & if \geq \varepsilon^2 \\ 0 & else \end{cases}$$

where *n* and N_{frag} are the number of time points and fragments, respectively, and $D_j(t)(E)$ is the deuterium fraction of fragment *j* at the time point *t* in ensemble *E*.

SAXS data

If not combined to simulation, the interpretation of SAXS data would be limited to the radius of gyration of the protein. The SAXS intensity curves were computed using the program CRYSOL developed by Svergun *et al.* (114). Default parameters of the program were used. Each profile was made of N_p=51 points. Hence, the observable was a vector with N_p coordinates. The *qth* coordinate was readily computed by averaging its value over the ensemble *E*:

$$\langle I_q \rangle(E) = \frac{1}{N}\sum_{i=1}^{N} I_q^i$$

where *N* is the number of structures and I_q^i is the *qth* scattering vector point of the SAXS profile of structure *i* . The MSD from the reference ensemble is:

$$MSD_{SAXS}(W) = \frac{1}{N_p}\sum_{q=1}^{N_p}\left(\frac{\langle I_q \rangle(E_{reference}) - \langle I_q \rangle(E_{refined})}{\langle I_q \rangle(E_{reference})}\right)^2$$

Ion-mobility cross-section

In ion-mobility spectrometry, protein is ionized and accelerated by an electric field through a buffer gas that opposites the ion motion. Measuring the drift time caused by collisions with gas molecules allows the cross-section of the protein to be estimated (115). The open source script, MOBCAL, was used to calculate the cross section of a structure given its coordinates (116). The trajectory method, where the ion is treated as a collection of atoms represented by a 12-6-4 potential, was used. Charge distribution was assumed to be uniform. The observable is a scalar:

$$\langle \sigma \rangle(E) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i$$

where *N* is the number of structure in ensemble *E* and σ_i the cross-section of structure *i*. The MSD from the reference ensemble was calculated as:

$$MSD_{mobility}\left(E_{refined}\right) = \left(\frac{\langle \sigma \rangle\left(E_{reference}\right) - \langle \sigma \rangle(E_{refined})}{\langle \sigma \rangle\left(E_{reference}\right)}\right)^2$$

Single molecule FRET data

smFRET data were modelled as histograms of transfer efficiency, *E*, defined as:

$$E = \frac{1}{1 + \left(\frac{R}{R_0}\right)^6}$$

where R_0 is the Forster distance and *R* the distance between the centre-of-mass between the two residues where the donor and acceptor dyes are assumed to be attached. The Forster distance was fixed to 15 Angstroms. This value is low compared to the values encountered in practice (~60 Å), but more adapted to the size of FIP35, which is relatively small. The flexibility of the dyes was ignored in this model. *N_{fret}* random pairs of residues were selected and their histogram calculated according to the weight of each structure. The MSD was calculated by comparing the histograms of the FRET efficiency calculated from the reference ensemble and the refined one. Each histogram was made of 20 bins of equal size. The similarity between the two distributions was measured using the Jenson-Shannon divergence (117):

$$D_{JS}(X,Y) = \frac{D_{KL}(p_X, (p_X + p_Y)/2) + \ D_{KL}(p_Y, (p_X + p_Y)/2)}{2}$$

where X and Y are two distribution function and D_{KL} is the Kullback-Leibler divergence (118):

$$D_{KL}(p_X, p_Y) = \int p_X(x) log \frac{p_X(x)}{p_Y(x)} dx$$

Then, the MSD was defined as:

$$MSD_{FRET}(E_{refined}) = \frac{1}{N_{fret}} \sum_{k=1}^{N_{fret}} D_{JS}(E_{reference}, E_{refined})$$

Chemical shift data

Chemical shifts of 6 atom types C_α, H_α, C_β, C_O, N and H_N of each residue and each structure were computed using Camshift (119). In ensemble *E*, the observed chemical shift of atom type X from residue r was calculated as:

$$\langle X_r \rangle (E) = \frac{1}{N} \sum_{i=1}^{N} X_r^i$$

where N is the number of structures in ensemble *E*. The MSD from the reference ensemble was defined as:

$$MSD_{CS}\left(E_{refined}\right) = \frac{1}{6\,(N_{res}-2)} \sum_{r=2}^{N_{res}-1} \sum_{X=1}^{6} \left(\frac{\langle X_r \rangle (E_{reference}) - \langle X_r \rangle (E_{refined})}{\langle X_r \rangle (E_{reference})} \right)^2$$

where N_{res} is the total number of residues in the protein.

Total deviation from synthetic data

The total deviation of the observables from the synthetic experimental data (i.e. the observables calculated from the reference ensemble) was evaluated as:

$$MSD_{tot}\left(E_{refined}\right) = \frac{1}{\sum_{i\in Obs} \lambda_i} \sum_{i\in Obs} \lambda_i MSD_i\left(E_{refined}\right)$$

where Obs is the ensemble of the observables (smFRET, HDX(-MS), SAXS, CS, ion-mobility) and λ_i the Lagrange multiplier of observable *i*. Changing the Lagrange multipliers allows favouring the matching of one observable over another. Typically, λ_i=1 and is set to 0 if observable *i* is disregarded. To make sure that for a same Lagrange multiplier, two observables count equally during refinement, each deviation was normalized by its average value such as:

$$MSD_{tot}\left(E_{refined}\right) = \frac{1}{\sum_{i\in Obs} \lambda_i} \sum_{i\in Obs} \lambda_i \frac{MSD_i\left(E_{refined}\right)}{\langle MSD_i \rangle} \qquad (2\text{-}6)$$

$\langle MSD \rangle$ of each single observable was evaluated by averaging the mean square deviation of the observable over 1,000 ensembles , which had been generated randomly.

2.4 Analytical tools

2.4.1 Root mean square deviation and fluctuation

The structural similarity between two conformations of a protein was measured using the root mean square deviation (RMSD), defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\overrightarrow{r_i^0} - \vec{r_i}\right\|^2}$$

where $\overrightarrow{r_i^0}$ and $\vec{r_i}$ are the vector positions of the C_α atom of residue *i* in the initial conformation and the new conformation, respectively. *N* is the total number of residue in the protein. The new structure was aligned to the initial one before calculation.

For a given trajectory, the root mean square deviation of a residue *i* is defined as:

$$RMSF_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\|\overrightarrow{r_i^{ref}} - \overrightarrow{r_i^j}\right\|^2}$$

where $\overrightarrow{r_{i,j}^{ref}}$ and $\overrightarrow{r_i^j}$ are the coordinates of the C_α atom of residue *i* in the reference structure and the jth frame of the trajectory, respectively. *M* is the total number of frames in the trajectory. The reference structure was calculated by averaging the conformations of the protein over the trajectory. The trajectory was then aligned to the reference structure before calculation. Both the RMSD and RMSF were calculated using the free software Wordom (120).

2.4.2 Identification of subunit interface

Solvent accessible surface area was calculated using NACCESS (http://www.bioinf.manchester.ac.uk/naccess). Briefly, the program calculates the atomic surface area by rolling a probe of a given size around the van der Waals surface of the residue. The probe had the default radius of 1.4 Å. The variation of solvent accessible surface area from the monomeric to the hexameric form for a fragment *j*, ΔS_j, was quantified as:

$$\Delta S_j = \frac{1}{n_j} \sum_{i=m_j}^{m_j+n_j-1} \frac{S_i^{mono} - S_i^{hexa}}{S_i^{mono}}$$

where S_i^{mono} and S_i^{hexa} are the total solvent accessible surface area of the residue *i* in the monomeric and hexameric structure, respectively.

2.4.3 Verification of peak assignments in MS spectra

To verify whether mismatches between experimental and predicted exchange kinetics of some fragments were due to misassignment, we double-checked the assignment of the raw MS peaks. The mass of all possible fragments of the primary sequence between length 5 and 40 aminoacids was calculated by summing the residue masses and adding the mass of a water molecule (18.01056 atomic mass unit, u), corresponding to the adding of OH and H during the hydrolysis and formation of the new C- and N-terminus. Each mass experimentally detected was compared with the calculated ones. A threshold precision of 0.05u was used, i.e. the theoretical peak was considered to potentially match with the experimental one if $|m_{exp}-m_{th}|=\Delta m$ was lower than 0.05u (see Table A - 3).

2.4.4 Comparing ensembles

To compare two ensembles of structures of FIP35, each ensemble was projected along the optimal reaction coordinate (see Section 5.3.3), and their distributions were compared using the Jenson-Shannon divergence (117) introduced in Section 2.3.2 in the single molecule FRET description.

2.5 Molecular Biology Methods

2.5.1 His-tagging of φ8 P4 at the N-terminus

φ8 P4 was His-tagged at its N-terminus to simplify purification. The wild type φ8 P4 (φ8-WT) was cloned into the plasmid bSJ1b (pet32b(+), approximately 6900bp, (121)). To His-Tag the wild type at its N-terminus, oligonucleotide primers were designed to amplify a φ8-WT gene that contains additional extensions encoding NdeI/XhoI restriction sequences (Figure 2-4 and Table 2-1). Pet28ac(+) plasmid was digested overnight with NedI and XhoI enzymes and alkaline phosphatase to prevent religation. The DNA was purified using a PCR purification kit (New England Biolabs). The linear plasmid and the amplified gene were ligated and then transformed into XL1-blue cells, followed by colony-picking and medium-prep extraction. The resulting plasmid, φ8-NHisTag, was confirmed by sequencing (performed by GATC Biotech).

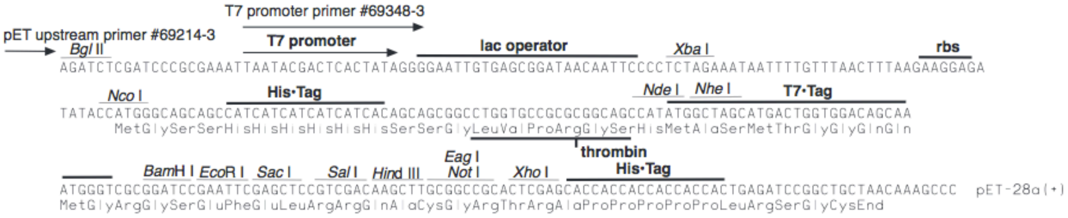

T7 promoter primer #69348-3
pET upstream primer #69214-3
Bgl II
T7 promoter
lac operator
Xba I
rbs
AGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGA
Nco I
His•Tag
Nde I
Nhe I
T7•Tag
TATACCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGGCTAGCATGACTGGTGGACAGCAA
MetGlySerSerHisHisHisHisHisHisSerSerGlyLeuValProArgGlySerHisMetAlaSerMetThrGlyGlyGlnGln
thrombin
BamH I
EcoR I
Sac I
Sal I
Hind III
Eag I
Not I
Xho I
His•Tag
ATGGGTCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCC
pET-28a(+)
MetGlyArgGlySerGluPheGluLeuArgArgGlnAlaCysGlyArgThrArgAlaProProProProProLeuArgSerGlyCysEnd

***Figure 2-4**: pET-28a-c(+) cloning/expression regions.*

Name	Sequence	Modification
f8f f8r	CTG GAG TCA TTT GTC AAC TCC TTC AAT GT CAT ATG GCT AGA AAA ACG AAA GTT ACA C	His-tag
P4C128Af P4C128Ar	CTA CGC AGC CCA GAT GGC TGC GAA AGG TCT GAA G CTT CAG ACC TTT CGC AGC CAT CTG GGC TGC GTA G	C128A
8p4A285Cf 8p4A285Cr	CGC ACC GAT GGC GTT GCA GTT GCT GAC ACC CCG CGG GGT GTC AGC AAC TGC AAC GCC ATC GGT GCG	A285C
8p4A287Cf 8p4A287Cr	GAA CAC CGC ACC GAT GCA GTT TGC GTT GCT GAC A TGT CAG CAA CGC AAA CTG CAT CGG TGC GGT GTT C	A287C
8p4A290Cf 8p4A290Cr	CAA CAA GAA TTG ATT GAA CAC GCA ACC GAT GGC GTT TGC GTT GCT AGC AAC GCA AAC GCC ATC GGT TGC GTC TTC AAT CAA TTC TTG TTG	A290C
8p4A304Cf 8p4A304Cr	TGT TTG GCC CGA GGA AGC AAC TGC CGT CCT GAC CG CGG TCA GGA CGG CAG TTG CTT CCT CGG GCC AAA CA	A304C

Table 2-1: *Sequences of primers used for amplification and site directed mutagenesis.*

2.5.2 Site directed mutagenesis

In vitro site-directed mutagenesis was used to change the position of the single cysteine (Cys128 in the wild type). Oligonucleotide primers containing the desired nucleotide changes were used to mutate φ8-NHisTag. About 50 ng of purified φ8-NHisTag plasmid was mixed with 2.5 µl of 10 µM of forward primers, 2.5 µl of 10 µM of reverse primers, 20 µl of 5x phusion buffer HF, 2 µl of 25 mM dNTP, 2 µl phusion HF polymerase (2 U/µl) and 69 µl dH_2O. Amplification was performed following the program described in Table 2-2. PCR products were incubated with DpnI (New England Biolabs) for 3 hours at 37 ˚C then the temperature was increased to 80 ˚C for 20 min to inactivate enzymes. PCR products were purified with a PCR purification kit (New England Biology). Plasmids where then transformed into XL1-Blue cells with 45s heat-shock pulse at 42 ˚C and grown overnight, followed by colony-picking and medium-prep extraction (New England Biolabs). All plasmids were confirmed by sequencing (performed by GATC Biotech).

Step	Temperature	Time
1	98˚C	30s
2 (x15)	98˚C	10s
	55˚C	1min
	72˚C	3.5min (~30s/kbp)
3 (final elongation)	72˚C	5min

Table 2-2: *Temperature cycling program for PCR*

2.5.3 Expression of φ8 P4

φ8 P4 wild type or mutants were expressed with the same protocol. 10 ng of purified DNA was transformed into 25 µl of BL21 cells with 45 s heat-shocking at 42 ˚C. Cells were mixed in 200 µl SOC and incubated for 1 h at 37 ˚C and 250 rpm. Positive transformants were selected by plating out cells on Kanamycin agar plates. A single colony was picked and grown in 10 mL LB media (10 g/L NaCl, 10 g/L trypton, 5 g/L yeast extract, 50 µg/ml Kanamycin) for 6 h at 37 ˚C and 250 rpm. The cells were then transferred in 250 mL of fresh media and incubated in the same conditions for 4 h. The inoculated media was then split over 6 flasks of 1L LB media. Once OD_{600nm} reached 0.6, cells were induced by the addition of 1 mM IPTG and incubated for 16 h at 18 ˚C and 250 rpm, then harvested by centrifugation at 4,500 rpm for 15 min at 4 ˚C. Cells were resuspended in 30 ml Buffer A (20mM Tris-HCl pH8.5, 400mM NaCl, 7.5mM $MgCl_2$) to which a cocktail of protease inhibitors, 1mM of PMSF, 1mg/mL of lysozyme and 10µg/µL DNAse was added. Cells were incubated for 30 min at 4˚C and then lysed by sonication in ice (three times 15 s with 45 s breaks). The resulting lysate was centrifuged for 45min at 40,000 rpm (50Ti rotor) and 4 ˚C to pellet insoluble material and the supernatant was collected and immediately purified.

2.5.4 Purification of φ8 P4

The whole purification was performed at 4˚C. The soluble part of lysate was loaded onto a Ni-NTA chromatography column (HisTrap FF 5ml column, GE Healthcare), which was previously equilibrated with washing buffer (75mM

Imidazole, 20mM Tris-HCl pH8.5, 500mM NaCl, 7.5mM $MgCl_2$). The column was washed with 20 column volumes of washing buffer before being equilibrated with a buffer with low salt concentration (20mM Tris-HCl pH8.5, 50mM NaCl, 7.5mM $MgCl_2$). The protein was eluted by rapidly increasing the concentration of eluting buffer (500mM Imidazole, 20mM Tris-HCl pH8.5, 50mM NaCl, 7.5mM $MgCl_2$) to 100%. Fractions containing the protein (identified by SDS-PAGE analysis) were loaded onto an anion exchange column (HiTrap Q HP 5ml column, GE Healthcare) previously equilibrated with buffer A (20mM Tris-HCl pH8.5, 50mM NaCl, 7.5mM $MgCl_2$). The column was washed with 20 column volumes of buffer A to remove all imidazole and potential contaminants. The protein was eluted with buffer B (20mM Tris-HCl pH8.5, 500mM NaCl, 7.5mM $MgCl_2$). In order to maximize the protein concentration of the elute, no gradient was used, which did not affect the purity of the elute. The final φ8 P4 concentration was about 150 μM with a total volume of approximately 7 mL, i.e. the total yield was about 12 mg (molecular weight: 35kDa, extinction coefficient: 11,920$M^{-1}cm^{-1}$). The protein was flash frozen with liquid nitrogen in 50μL aliquots and kept at -80˚C.

2.5.5 Labelling of φ8 P4 and purification from free dyes

500μL of 100μM φ8 P4 was mixed with 10μL of 1M TCEP and incubated for 15min at room temperature to reduce potential disulfide bonds between cysteines (unlike DTT and isopropanol, TCEP does not react readily with maleimides). The sample was then mixed with 3 fold excess of each dye (Alexa Fluor 488 and 594 C5 Maleimide, ThermoFisher) and incubated for 1h at room temperature. The sample was quenched by 100-fold dilution in buffer B (20mM Tris-HCl pH8.5, 0.5M NaCl and 4mM DTT). After incubation for 1h at 4˚C, proteins were separated from unreacted dyes by using a 1mL Ni-NTA chromatography column as described in 2.5.4 (except that 2mM DTT was added to the washing buffer). Fractions containing the labelled protein were collected, diluted 100 fold in Buffer B and incubated overnight at 4˚C. The labelled protein was purified using a 1mL Ni-NTA chromatography column as described above.

The degree of labelling of the protein was calculated such as:

$$D_{488} = \frac{A_{495nm}}{\varepsilon_{488}\,[P4]} \quad \text{and} \quad D_{594} = \frac{A_{590nm}}{\varepsilon_{594}\,[P4]}$$

with

$$[P4] = \frac{A_{280nm} - \gamma_{488}A_{495nm} - \gamma_{594}A_{590nm}}{\varepsilon}$$

where *A* is the absorbance, γ_{488} and γ_{594} are the correction factor of the Alexa Fluor 488 and 594. ε_{488}, ε_{594} and ε are the molar extinction coefficients of the two dyes and φ8 P4, respectivly (see Table 2-3).

Alexa Fluor	Correction factor	Extinction coefficient
AF488	0.11	71,000 $cm^{-1}.M^{-1}$
AF594	0.56	73,000 $cm^{-1}.M^{-1}$

Table 2-3: *Molar extinction coefficients and correction factors of Alexa Fluor dyes.*

2.5.6 Characterisation of φ8 P4

ATPase assays

The ATPase activity of φ8 P4 was verified using an EnzChek Phosphate Assay Kit (Molecular Probes). 1 µM φ8 P4, 1 mM ATP and 0.5 mg/ml poly(A) were mixed with 20X buffer (1M Tris-HCL pH 7.5, 20mM $MgCl_2$, 2mM sodium azide), the purine nucleoside phosphorylase (PNP) and the 2-amino-6-mercapto-7-methyl-purine riboside (MESG). Preparation is summarised in Table 2-4. The evolution of the absorbance at 360 nm over the time was measured with a micro-plate reader (POLARstar OPTIMA). The activity of mutants and/or labelled proteins was systematically compared with the activity of the wild type. Controls without RNA were performed to confirm RNA induced ATPase-activity and detect potential false positive due to contamination with other ATPase. For the linear calibration of absorbance *versus* phosphate concentration, the absorbance of different standard phosphate concentrations (0 µM, 20 µM, 50 µM, 100 µM) was systematically collected.

Product	Stock	Final concentration	Volume added
Buffer	20X	1X	10µL
MESG	1mM	200µM	40µL
PNP	100 U/mL	1 U/mL	2µL
ATP	10mM	1mM	20µL
PolyA	10mg/mL	0.5mg/mL	10µL
P4	~30µg/mL	~0.5µM	5-50µL
dH_2O	-	-	113-68 µL

***Table 2-4**: Preparation of the ATPase assays. The total volume in each well is 200 µL.*

Translocase assay

A complementary way to check the functionality of φ8 P4 is to verify whether the motor can unwind a double stranded nucleic-acid chain. Hence, translocase assays were performed. A 42-nt-long RNA strand was annealed with a 29-nt-long DNA strand. DNA was labelled with Alexa Fluor 488 dye at its 5' end. The first 8 nucleotides of the DNA strand were not complementary to the rest of the RNA such that the duplex was forming a fork at the 5' end of the RNA strand. The fork, which is about twice longer than the central channel of the helicase, i.e. ~ 90 Å, was designed to enable the helicase to bind to the duplex (see Figure 4-11). 1 nM of the duplex (RNA/AF488-DNA) was mixed with 10 µM P4 and 1 mM ATP. To avoid re-annealing of AF488-DNA after unwinding, a large excess of non-labelled DNA was added (20nM). The sample was incubated for 30 min at room temperature and analysed on a native gel.

Strand	Sequence	Extinction coefficient
RNA	CCC CCC CCC CCC CUG CCC AAG AGA AAA AGA GAA UAC CUG CCG UU 3Biodt	415,400 M-1cm-1
DNA	CGG CAG GTA TTC TCT TTT TCT CTT GGG CAT TTT TT	306,400 M-1cm-1
AF488-DNA	Alexa488 CGG CAG GTA TTC TCT TTT TCT CTT GGG CAT TTT TT	369 OD/mmol

***Table 2-5:** Sequences of RNA and DNA strands designed for the translocase assay.*

2.5.7 General analytical methods

Agarose Gel 1.5% electrophoresis

Agarose gels were routinely run to examine the results of plasmid purification. 50mL dH_2O was mixed with 0.7g agarose in a flask and heated for 1min in microwave. Once tepid, 1 mL of 50X TAE (2M Tris, 0.5M EDTA, 5.71% glacial acetic acid) and 3µL of 10mg/mL Ethidium Bromide were added. Samples were made by 2 µL DNA with 2µL 6X loading dyes and 8 µL dH_20. Gels were run for 80 min at 5V/cm, then stained with SYBR-Gold (Life Technologies) for 15 min and visualized by trans-illumination.

Sodium dodecyl sulphate polyacrylamide gel electrophoresis

Protein identification and purity was assessed by SDS-polyacrylamide gel electrophoresis using the tris-glycine buffer system (routinely 15% gels were made). Resolving and stacking gels were prepared as described in Table 2-6. TEMED was added right before pouring the gel into the cast. A few drops of isopropanol were added on top of resolving gel and removed before pouring the stacking gel. For each protein sample, 10µL of sample was mixed with 1:1 2X loading buffer (DTT free in order to detect potential dimer formation), boiled for 10min, cooled down in ice for 2min and span before being loaded on the well. 400mL of running buffer (25mM Tris, 192mM glycine, 0.1% SDS) was poured in the tank. The gel was run for 1h at 180V (constant voltage) then stained with InstantBlue (Expedeon) for 15min. The loading buffer was DTT free to detect potential dimer formation.

Reagent	Resolving (15%)	Stacking (8%)
30% Acrylamide (37:1)	7.5mL	1.34mL
1M Tris pH 8.8	5.6mL	-
1M Tris pH 6.8	-	1.25mL
dH_2O	1.85mL	6.67mL
10% SDS	150µL	100µL
10% APS	120µL	80µL
TEMED	20µL	10µL

***Table 2-6**: Preparation of 15% SDS-PAGE gels. Volumes shown are sufficient to cast two 8 cm X 10 cm mini gels using 0.75 mm spacer.*

2.5.8 Native Polyacrylamide Gel

Native gels were run to verify the hybridization of the nucleotide strands. Gel were prepared as shown in Table 2-7. The gel was run for 1-2h at 5V/cm and stained with SYBR-Gold (Life Technologies) for 15min. The gel was scanned at 473nm (400V, FITC) to visualize labelled strands.

Reagent	Volume
30% Acrylamide (29:1)	5mL
10% APS	170 µL
5X TBE buffer	3mL
TEMED	17 µL
dH_20	6.8mL

***Table 2-7**: Preparation of 10% native gels.*

2.6 Fluorescence Spectroscopy

2.6.1 Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) is a correlation analysis of the fluctuations of the fluorescence intensity in a sub-femtolitre volume. A comprehensive review about FCS can be found in Ref (122). FCS was used to estimate the hydrodynamic radius of the labelled P4. The autocorrelation of the fluctuations was fitted using the model described in Ref (123), which accounts for the triplet state component:

$$G(\tau) = \frac{1}{N} \times \frac{1 - f_t + f_t e^{-\frac{\tau}{\tau_t}}}{1 - f_t} \left(1 + \frac{\tau}{\tau_D}\right)^{-1} \left(1 + \frac{\tau}{\kappa^2 \tau_D}\right)^{-\frac{1}{2}}$$

where τ is the lag time and τ_D the characteristic residence time in the confocal volume of the fluorophore. N denotes the average number of fluorescent particles in the confocal volume. κ is the ratio of axial to radial radii of the confocal volume, f_t and τ_t are the fraction and the characteristic time constant of the triplet-dependant dynamics, respectively. Fitting was performed with a non-linear least squares method based on the popular Levenberg-Marquardt algorithm.

The Einstein –Stokes relationship gives:

$$D = \frac{k_B T}{6\pi\eta R_h}$$

where D and R_h are the diffusion constant and hydrodynamic radius of the particle, respectively. k_B is the Boltzmann constant, *T* the temperature and η the dynamic viscosity.

Furthermore:

$$D = \frac{\omega_r^2}{4\tau_D}$$

where ω_r is the radial radius of the confocal volume. Hence, knowing the radius of gyration, *R_{ref}*, and diffusion time, τ_{ref}, of a reference particle, one can estimate the radius of gyration of a second particle as:

$$R_h = R_{ref}\,\frac{\tau_D}{\tau_{ref}}$$

FCS measurements were performed on labelled P4 at 1 nM. AF-488 and AF-594 were used as references, knowing that their hydrodynamic radius is ~0.7 nm.

2.6.2 Ensemble FRET data collection

Ensemble FRET measurements were performed on dual labelled P4. 1 μM of the labelled protein was mixed with 1 mM ATP and 0.5 mg/mL Poly(A). Sample was excited at 488 nm and its emission spectrum was collected in the wavelength range of 500 nm to 700 nm.

2.6.3 Alternating laser excitation data collection

Alternating Laser EXcitation (ALEX) is a single molecule fluorescence spectroscopy technique. A comprehensive description of the method can be found in Refs (124, 125).

Based on rapid switching between excitation of the donors (e.g. AF-488) and the acceptors (e.g. AF-594) which pass through the confocal volume, the method enables the sorting of fluorescently labelled species based on the number and type of fluorophores present. As a labelled protein passes through the confocal volume, it generates a florescence burst. For each burst, the apparent FRET efficiency, is calculated as:

$$E = \frac{f_{D_{ex}}^{A_{ex}}}{f_{D_{ex}}^{D_{ex}} + f_{D_{ex}}^{A_{ex}}}$$

where $f_{D_{ex}}^{D_{ex}}$ and $f_{D_{ex}}^{A_{ex}}$ are the fluorescence intensities in the donor and acceptor channel after donor excitation. As explained in Section 1.3.3, *E* gives information about the donor and acceptor distances. An additional term, called the stoichiometry, is defined as:

$$S = \frac{f_{D_{ex}}^{D_{ex}} + f_{D_{ex}}^{A_{ex}}}{f_{D_{ex}}^{D_{ex}} + f_{D_{ex}}^{A_{ex}} + f_{A_{ex}}^{A_{ex}}}$$

where $f_{A_{ex}}^{A_{ex}}$ denotes the fluorescence intensity in the acceptor-emission channel after direct excitation of the acceptor. For donor-only-species, S ~ 1 (as $f_{A_{ex}}^{A_{ex}} \sim 0$), whereas S ~ 0 for only-acceptor-species (as $f_{D_{ex}}^{D_{ex}} + f_{D_{ex}}^{A_{ex}} \sim 1$). A species carrying both fluorophores exhibit a distinguishable stoichiometry S ~ 0.5. Various species present in the sample are identified using the two-dimensional *ES* histogram.

Data collection

Samples were diluted to a final concentration of 100 pM labelled P4 in 20 mM Tris-HCl pH8.0, 300mM NaCl, 7.5mM $MgCl_2$ and 2mM DTT. Data were collected with a custom-made invert confocal microscope setup. Briefly, donor- and acceptor- excitations were performed with a diode-pumped 488 nm laser and a He-Ne 594 nm laser, respectively. Alternation was achieved with electro-optical modulators, which were controlled with a LabView script (126). Laser beams were depolarized and their sizes were adjusted with telescopes. Beams were guided to the objective and the photo detectors (avalanche photodiodes) using a set of mirrors, dichroic mirrors, pinholes and filters. Cross-talk and gamma corrections were performed as described in (127).

Chapter 3: Functional dynamics of helicase probed by hydrogen deuterium exchange and simulation

3.1 Introduction

Proteins are highly dynamic molecular entities (128) and their conformational variability is essential to their function (38). This is particularly the case for macromolecular complexes that play essential roles in the cell such as molecular motors (129, 130). As mentioned in the Introduction (Section 1.3.2), a powerful technique to investigate the dynamics of large proteins and their complexes is hydrogen-deuterium exchange probed by mass spectrometry (HDX-MS) (89, 90). Key to interpreting HDX kinetics is the fact that exchange occurs faster for amides that are solvent-exposed and not involved in hydrogen bonds. The link between the HDX kinetics and the structural dynamics of a residue, and *a fortiori* of a chain segment, is not straightforward. Thus, HDX-MS data are usually limited to qualitative analysis, e.g., by mapping the apparent rate of exchange of different fragments on the available structure and comparing directly the kinetics of the same fragments under different conditions (78, 92, 131), although computational methods have been proposed to predict HDX of proteins from structure (79, 86, 132–134).

Hexameric packaging motors (P4 proteins, Figure 3-1A) from cystoviruses φ6, φ8, φ12 and φ13 are responsible for genome translocation into preformed capsids using energy from ATP hydrolysis (135). These proteins have been characterized extensively and many high resolution structures in different conformational states are available (17, 18, 24), making them an excellent model system for the related SF4 helicases (136). A more detailed review of the mechanisms of P4 is given in Section 1.1.3. HDX-MS kinetics have been obtained for a free hexamer and capsid-bound φ12 P4 and qualitatively interpreted previously (131). The φ12 P4 subunit is constituted of three regions: the N-terminal apical domain, the conserved RecA-like

ATPase core and the C-terminal extension (Figure 3-1B). The C-terminal extension (residues 290-331) is essential for the binding of the hexamer to the capsid (131, 135). Loops L1 (residues 196-206, partially disordered) and L2 (residues 233-238) protrude into the central channel where they contact RNA during translocation (78, 137). The loop L2 together with helix α6 constitutes a moving lever that affects the translocation power stroke (18).

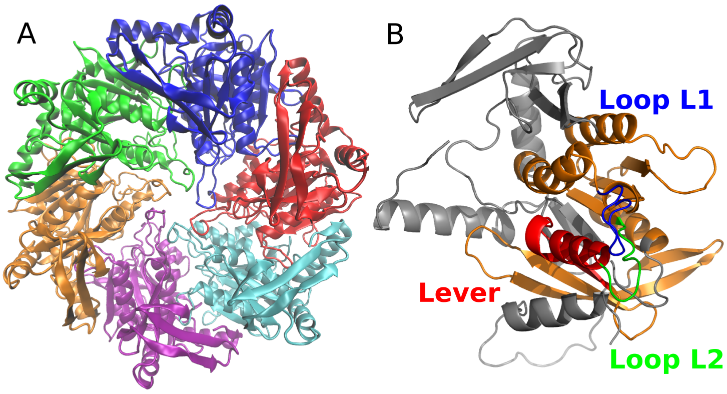


Figure 3-1: *Structure of φ12 P4.*
(A) X-ray crystal structure of the hexamer φ12 P4 in the apo state. Different subunits are shown in different colours. (B) Elements located in the channel and interacting with RNA (Loops L1, blue, and L2, green). The conserved RecA-like nucleotide-binding domain is coloured gold and the lever in red. In the depicted monomeric structure, the lever is in a "down" state. In the proposed mechanism the lever is locked in a "up" conformation in ATP-bound state and moves to the "down" configuration as a result of hydrolysis and phosphate release (133).

3.2 Overview of the chapter

In this Chapter, the deuterium fractions of any chain fragment of the packaging motor P4 from bacteriophage φ12 is estimated from molecular dynamic simulations of the native state of the P4 hexamer and monomer. Sampling the local conformational space of the protein with a ~100 ns simulation is sufficient to predict (with some instructive exceptions) the experimental exchange kinetics for times ranging from seconds to hours. Thus the simulation provides a high-resolution representation of the microscopic structures and dynamics responsible for the hydrogen-deuterium exchange over several orders of magnitude in time, which is validated by the experiment. The proposed method is also a powerful tool to validate the assignment of the fragments, to assess the structure of modelled regions missing from the crystal structure and to probe conformational variability that cannot be observed by X-ray crystallography.

3.3 Methods

3.3.1 Molecular dynamics simulations

Simulations of the φ12 P4 hexamer and monomer in the apo state were performed with NAMD using the CHARMM36 force field; 77349 TIP3P water molecules were included to ensure that at least 10 Å separate periodic images of the proteins as well as 235 Na ions and 205 Cl ions to set the ion concentration at 0.15 M. The crystal structure of the apo state φ12 P4 (17) (PDB access code: 4BLR) was used as initial conformation. The missing residues (196-206, 236-241, 299-312, 329-331) were modelled using MODELLER . Simulations were performed at 298 K and atmospheric pressure. Periodic boundary conditions were applied and long-range electrostatic interactions were calculated with the particle mesh Ewald method, with a cut-off of 12 Å and grid spacing of 1 Å. Neighbour-atom lists were constructed including all atoms being less than 14 Å away from a given atom. A 2 fs time step was used and conformations were saved every 500 time steps (1 ps). The production runs were 100 ns long preceded by a 20 ns equilibration where temperature was increased from 0 to 298 K using 20 K increments every 500 ps.

3.3.2 Predicting fragments kinetics from MD simulations

A MD simulation of the biological system was first run to sample its local conformational space. Then, the averaged protection factor of each residue was calculated as explained in Section 2.2.1. The phenomenological Equation (**1-4)** depends on parameters β_c and β_h. We used the values ($\beta_c = 0.35$ and $\beta_h = 2$) previously shown to provide the best prediction for a set of seven proteins for which residue-specific data was available (10). If residue *i* contains an amide hydrogen, the protection factor, P_i, is defined as:

$$\langle P_i \rangle = \frac{k_{int}^i}{k_{obs}^i} \tag{5}$$

where k_{int}^{i} and k_{obs}^{i} are the intrinsic and observed exchange rates of the residue *i*, respectively. Thus, the deuterium fraction of residue *i* at time *t* is:

$$\boldsymbol{D_i(t) = 1 - e^{-(k_{int}^{i}/\langle P_i \rangle)t}}$$

(3-1)

where k_{int}^{i} is the intrinsic exchange rate of the residue *i*. The intrinsic exchange rates have been estimated as described in Ref (112) (see 2.2.2 for more details). Thus, the deuterium fraction $D_j^{sim}(t)$ of the fragment *j* at time *t* was:

$$D_j^{sim}(t) = \frac{1}{n_j}\sum_{i=1}^{n_j} D_i(t) = \frac{1}{n_j}\sum_{i=m_j}^{m_j+n_j-1}\left(1 - e^{-(k_{int}^{i}/P_i^{sim})t}\right)$$

(3-2)

where n_j and m_j are the number of amide hydrogen and the index of the first residue in the fragment *j*, respectively.

3.4 Results

Protection factors for each residue of the apo φ12 P4 hexamer were calculated using Equation (**1-4)** and averaged over a 100 ns MD trajectory at room temperature (Figure 3-2A). Protection factors are generally smaller for residues exposed to solvent (Figure 3-2B) and for residues located in particularly flexible regions, i.e., characterized by larger positional fluctuations (Figure 3-2C). This is the case of loops A76-S80, L1, L2 and part of the C-terminus (S299-I312) that have a large root mean square fluctuation consistent with the fact that they are either not resolved in the crystal structure or have a large B factor. Protection factors obtained from the crystal structure are systematically higher than those obtained from the simulation (Figure 3-2A), particularly in regions exhibiting higher fluctuations. This reflects the mechanism of EX2 exchange in which local conformational fluctuations mediate instantaneous solvent accessibility.

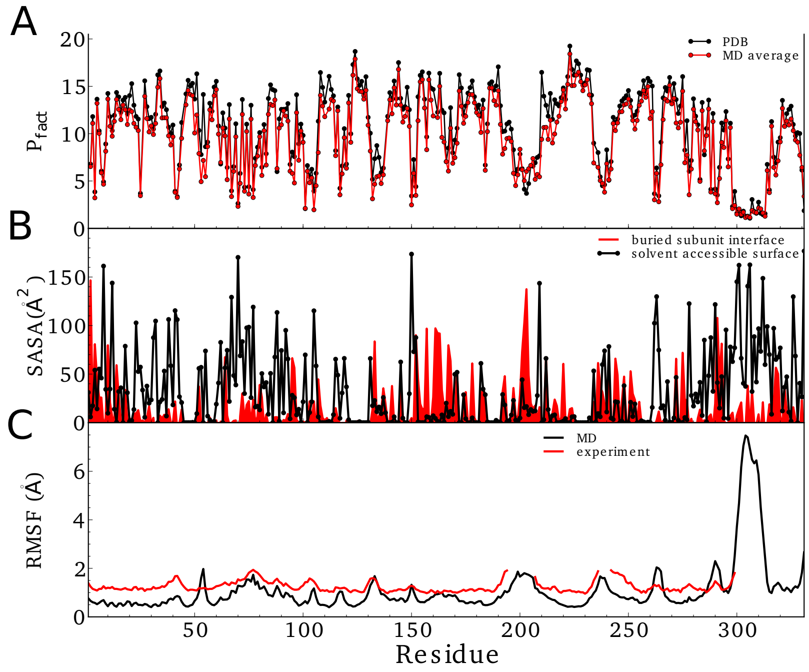


Figure 3-2: *Structural and dynamical characteristics of residues.*
(A) Protection factors calculated for the crystal structure of the hexamer of φ12 P4 (black) or time averaged over a 100 ns simulation (red). In both cases they represent the average over all monomers within the hexamer. (B) Solvent accessible surface calculated from the crystal structure for the hexamer of φ12 P4 (black line) and variation of solvent accessible surface area between the monomer and the hexamer (red line). The latter corresponds to the surface buried upon oligomerization and is shown to highlight the interfaces between monomers. (C) Root mean square fluctuation of the hexamer from its structure averaged over the 100 ns simulation or calculated from experimental Bfactors (red line).

3.4.1 Millisecond time scale kinetics is important

The time-dependent deuteration of each fragment *D(t)* was calculated using the protection factors calculated for each residue and Equation (**3-1)**. Figure 3-3A illustrates *D(t)* for selected fragments that have been analysed by MS but over a broader time interval than accessible experimentally with manual mixing (dashed vertical line in Figure 3-3A). It is evident from the plots that *D(t)* also provides valuable information over shorter timescales that require rapid mixing and quenching.

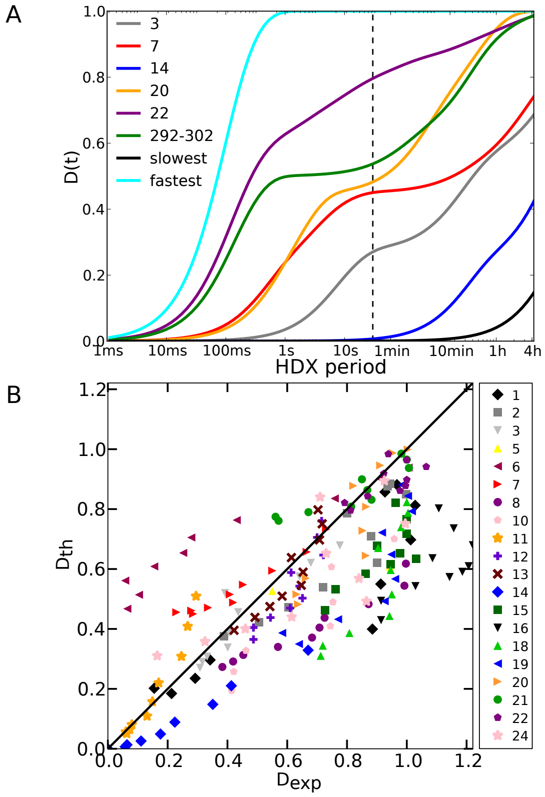


Figure 3-3*: Comparison of deuterium fraction predictions with experimental data.*
(A) Estimated deuterium fractions D(t) for selected fragments, including three that were not probed experimentally but representing the fastest (301-311) and the slowest (220-230) exchanging 11-residue fragments. Fragment 292-302 was chosen to highlight the similar kinetics to fragment 20 at long timescales while exhibiting a different one at shorter times. The vertical dashed line designates the fastest time experimentally measureable with manual mixing. (B) Estimated vs experimental deuterium fractions D(t) (including the time points t=30s, 1min, 2min, 4min, 8min, 15min, 30min, 1h, 2h and 4h) of all fragments of φ12 P4 (free in solution). Each fragment is reported with different symbol/color. Diagonal line represents perfect match. See Table S2 for assignments of each fragment.

Direct comparison between calculated and experimental *D(t)* for the 20 non-redundant fragments from Ref (131) is shown in Figure A - 2. In Figure 3B are plotted the *D(t)* values calculated from simulation (y-axis) against the experimental data (x-axis) for the free hexamer for each fragment and time-point for which experimental data is available. While points concentrate around the diagonal, the prediction is rather poor for a few fragments.

3.4.2 Peak assignment validation

One reason for poor prediction is potentially incorrect assignment, which may result from interpreting tandem mass spectra (MS/MS) of a complex mixture of primary ions. The assignment of each fragment has thus been checked. Interestingly, the monoisotopic mass of fragment 16 (originally assigned to residues 230-245) matches better that of a fragment encompassing residues 292-308 (Table A - 1 in Appendix). The predicted kinetics of the newly assigned fragment is in excellent agreement with the experimental kinetics, suggesting that the correct assignment should have been 292-308 (Figure 3-4A). However, no better assignment was found for the other fragments that exhibit discrepancy between experiment and prediction.

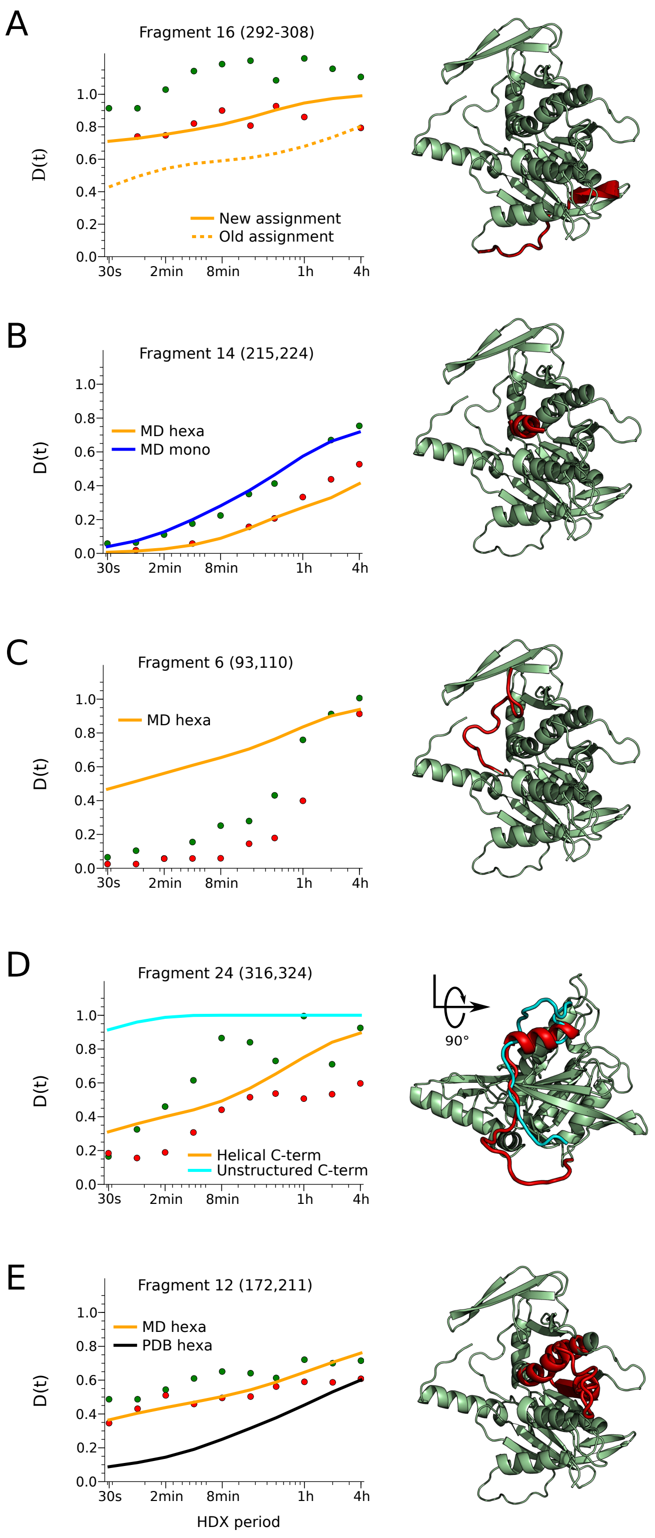


Figure 3-4: *Hydrogen-deuterium exchange kinetics of five relevant fragments.*
(A) fragment 16 prediction with the new assignment(S292-S308) and the old one (I230-L245). (B) fragment 14 (L215-S224). (C) fragment 6 (Q93-S110). (D) fragment 24 (I316-V324). (E) fragment 12 (172-211). The experimental exchange kinetics of the free and capsid-bound are shown as green and red circles, respectively. On the right, each fragment is highlighted in red within the structure of one subunit of φ12 P4. The whole C-terminal domain (K300-N331) is highlighted in (D) instead of only the fragment 24; note that the view is rotated 90 degrees wth respect to the other panels. The modelled C-terminal domain is highlighted in cyan.

3.4.3 Opening of the ring

For fragment 14 (residues L215-S224) (Figure 3-4B) we predicted slower exchange than the experimental one for free hexamers but in excellent agreement with that measured for the capsid-bound hexamer. Since this fragment is located at the subunit interface in the hexamer it is conceivable that the faster exchange is related to ring opening and is consistent with stabilisation of the hexamer by interactions with the capsid. These additional interactions prevent large conformational changes such as dissociation of subunit interfaces, therefore keeping the fragments localized at the interface buried. We thus formulated the hypothesis that the free form consists of a mixture of hexamers and open hexamers or lower order assemblies, including monomers. We simulated a single solvated monomer for 100 ns (see Figure A - 1 caption for details) and estimated the exchange kinetics of each fragment (Figure A - 2). Regions at the monomer-monomer interface or within the channel in the hexameric structure are exposed to solvent in the monomeric form and their exchange is predicted to be faster than in the hexamer while exchange kinetics remains unchanged for fragments located further from the interface (Figure A - 3 and Figure A - 4). Comparing experimental HDX kinetics of the hexamer free in solution with that of capsid-bound, indicates that exchange is significantly faster also for fragment 10, which is completely buried in the monomer-monomer interface like fragment 14. The crystal structure of φ12 P4 reveals that the fragments 10 and 14 are adjacent at the core of the monomer-monomer interface, such that fragment 10 is exposed to solvent if and only if fragment 14 is exposed as well (Figure A - 3). Since fragment 14 is helical, its secondary structure further limits the hydrogen exchange process even when exposed during ring opening. In contrast, fragment 10 lacks regular secondary structure and rapidly exchanges when exposed to solvent.

3.4.4 Structural model assessment

It is instructive here to mention the case of fragment 24 (residues I316-V324), which encompasses the C-terminus. As shown in Figure 3-4D, despite a quite large dispersion of the experimental results, the trend is well predicted by the native simulation. The C-terminus was only partially resolved (residues 301-331 disordered) in the previous crystal structure (18) (PDB access code: 1W4C). I initially performed the same simulation described in methods starting from 1W4C and modelling the C-terminal region as flexible region (138). As a result the C-terminus was quite dynamic and explored different conformations, which resulted in a large overestimation of the fraction of deuterium exchanged by fragment 24 at all times (90% is exchanged already at *t*=30s while experimentally the value is around 20%, for both the free and capsid-bound experiments), likely because the region was not correctly modelled. This finding highlights that the method can also be used to validate structural models.

For fragment 6 (residues Q93-S110) (Figure 3-4C) we predicted faster exchange than that measured experimentally. Fragment 6 encompasses residues 93-110 which are located in a loop close to the monomer-monomer interface. In simulations, the loop fluctuates and remains solvent exposed, as it is the case in the crystal structure, leading to the fast exchange kinetics prediction. One explanation is that the loop may adopt a more structured form in solution than it appears in the crystal structure and that this alternative conformation would be attained on a longer time scale than the current simulation

For fragments 18 (residues Q268-L294) and 19 (residues L270-V294) we predicted slower exchange than measured experimentally. The ring opening could not explain this mismatch since these fragments are not localized at the interface of two subunits. Hence, the kinetics of these two fragments suggests a local conformational change, which has not been captured over the 100 ns simulation. The mismatch is less pronounced for fragment 19, whereas the kinetics of fragment 20 is accurately predicted, suggesting that the conformational change occurs between residues 268 and 284, which encompasses one side of the ATP binding site and includes the highly mobile arginine finger 279. This region has been shown to be highly dynamic and responds to ATP and RNA binding in φ8 P4 (17, 78).

3.4.5 Importance of dynamics to interpret HDX data

We have seen above that a native state ensemble as sampled by a 100 ns room temperature simulation reproduces the experimentally probed hydrogen-deuterium exchange occurring on timescales ranging from 30 s to hours, except for specific regions for which we have to assume that conformational changes and large scale fluctuations, not sampled by the 100 ns simulation, may occur. In fact, a relatively fair prediction could also be obtained by neglecting the native state dynamics altogether and estimating the protection factors and *D(t)* from Equations (**1-4**) (**1-5**) and (**3-1**) using the crystal structure coordinates. Indeed, exchange would be predicted to be systematically slower (Figure 3-4E, Figure 3-5 and Figure A - 5), and this could be corrected by re-fitting the parameters β_c and/or β_h. However, the overall discrepancy between calculated and experimental *D(t)*, with all the caveats discussed above about the two different sets of experimental *D(t)*, would be larger. The importance of accounting for dynamics by estimating *D(t)* using protection factors calculated as ensemble averages is particularly evident for a few fragments, such as fragment 12, for which the fraction of deuterium exchange is seriously underestimated if calculated from the crystal structure alone (Figure 3-4E and Figure A - 5). This is also the case for other fragments encompassing a loop such as 13, 22 and 24 (Figure 3-5).

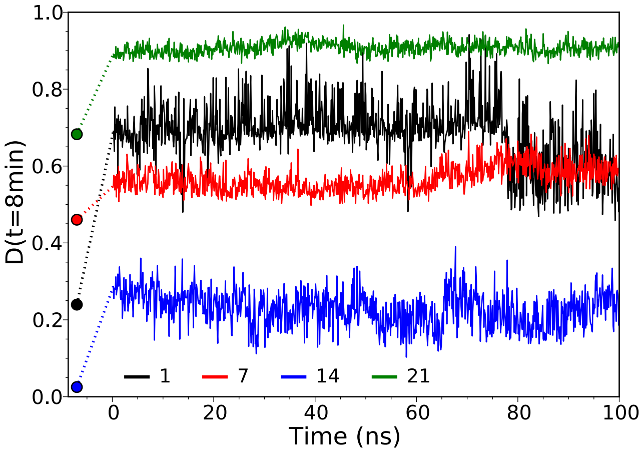


Figure 3-5: *Importance of interpreting HDX data through an ensemble of structures. Time series of the deuterium fraction for fragments 1 (black), 7 (red), 14 (blue) and 21 (green) at t=8min, D(t=8min), calculated for structures along the trajectory (computed for a single subunit within the hexamer, i.e. without averaging over the six subunits). The circles show the corresponding initial value obtained from the crystal structure.*

The importance of estimating hydrogen deuterium exchange as an average over a realistic ensemble of the relevant states appears clearly from the evolution of protection factors along the trajectory; as shown in Figure 3-5, for four fragments, the instantaneous estimated deuterium content varies significantly along the trajectory. Particularly interesting is the case of fragment 14 where the fraction of deuterium exchanged at t=8 min varies between 0.10 and 0.53 along the trajectory, with an average of 0.23, in excellent agreement with the experimental value (0.22), but considerably different from the value (0.02) obtained from the crystal structure alone.

3.5 Concluding discussion

We devised and tested a method based on detailed atomistic simulation to sample the native bound state for a large complex, such as a hexameric helicase, which allows prediction of hydrogen-deuterium exchange and facilitates direct, quantitative comparison with experimental data obtained by mass spectrometry. The results show that native-state dynamics is necessary and sufficient to predict, with some instructive exceptions, the HDX kinetics occurring over the timescale extending over six orders of magnitude. The central assumption is that the protection factor of individual residues can be estimated as an ensemble average of a function of the atomic coordinates of the protein, and that such a function can be empirically approximated as the sum of two terms, one term being proportional to the number of hydrogen bonds while the other to the number of contacts with neighbouring residues. Such an approximation has been previously proposed and shown to provide a relatively good prediction of the protection factors measured by NMR for small proteins (82). Here we use the same approximation to estimate the fraction of deuterium exchange for fragments of a large protein, as a function of time, and directly compare with HDX/MS measurements. The overall agreement with the experiment confirms the validity of the central assumption of the method. The second assumption is that HD exchange on timescales from ms to hours depends on the native state dynamics and that the ~100 ns trajectory samples accurately the ensemble of structures representing the bound native state. The method provides atomic resolution of the underlying dynamics and structural variability that is captured in the experiment over times ranging from seconds to hours.

This work has implications for refining HDX-MS methodology and for high-resolution structure validation. The first is illustrated by the discrepancy between the prediction and experiment for fragment 16 (Figure 3-4A), which was due to an incorrect assignment, an issue particularly important for larger, more complex assemblies which yield complex MS spectra. The other discrepancy reflected the wrong assumption about disorder in the C-terminal region based on the absence of electron density in the original crystal structure. Simulations that employed the more recent, higher resolution structure, in which the C-terminal region is helical, led to a slower exchange kinetics that, in turn, is in excellent agreement with the experiment (Figure 3-4D). This demonstrates that the quantitative prediction can be used as a quality check in HDX-MS experiments as well as it can complement X-ray crystallography in assessing modelled structures that are otherwise not resolved in the electron density.

The method also provides additional insights into the mechanism of the packaging motor. A quantitative comparison between the experimental and predicted kinetics for the free and capsid-bound hexamer, respectively, demonstrates that the free hexamer exists in a rapid equilibrium between closed and open conformation (Figure 3-4B). On the other hand the procapsid-bound hexamer matches well the intact hexamer prediction (Figure A - 2, fragment 14) and thus adopts the closed conformation. Since the ring opening is required for RNA loading into the hexamer, it has been proposed for the φ8 bacteriophage that the capsid bound P4 is in the open conformation (139). This is clearly not the case for φ12 P4 bound to the procapsid (i.e. capsid void of RNA).

Figure 3-6 illustrates another benefit, which the prediction brings to interpreting of HDX-MS. Although in principle possible, especially with the new ECD technology, residue-specific information is seldom obtained for large proteins and their complexes. In cases there is a good match between the fragment-specific experimental data one can assume this reflects the overall good prediction on the residue level and use the prediction to further interpret the observations. For example, the conserved P-loop (Walker A or H1 helicase motif involved in Mg and ATP coordination) fragment exhibits a biphasic kinetics (fragment 8, Figure A - 2) leading to an intermediate overall exchange rate (Figure 3-6A) while the predictions uncover great variations (Figure 3-6B). Contrary to expectations, the tip of the helix, which encompasses the conserved Thr137, is more flexible than certain parts of the loop upstream. As expected the rest of the downstream helix is buried within the core and protected. Another example of substantial and unexpected exchange rate variation is within the less conserved but essential nucleobase binding loop (Figure 3-6), which encompasses essential residues Tyr288, and Ser292 (fragment 20). The former stacks against the adenine base while the latter donates a hydrogen bond to the N7 position of the purine base, making the ATPase purine specific (17). In the apoprotein neither of the two residues is engaged in these interactions. Although both residues are part of the same beta hairpin, Tyr288 is as unprotected as the adjacent loop while Ser292 exchanges with an intermediate rate. Based on comparison with nucleotide-bound states of φ8 P4 (78) it is expected that these exchange rates will be sensitive to the nucleotide binding.

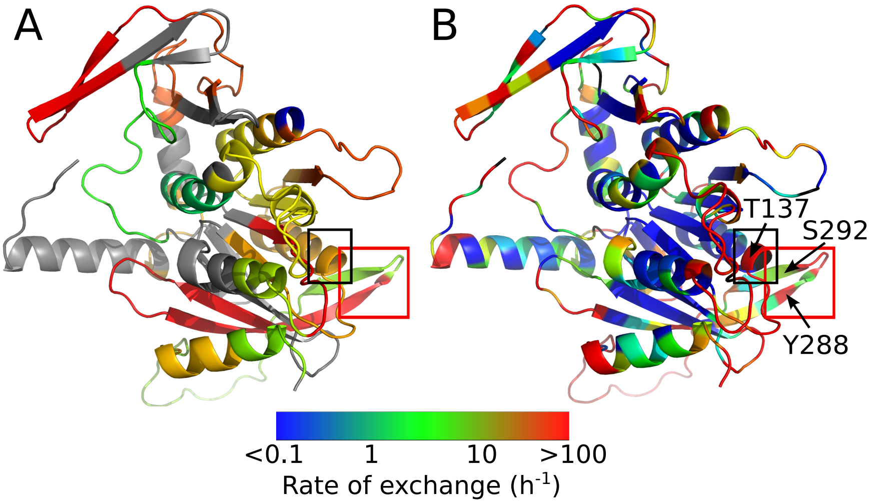


***Figure 3-6**: Comparison between the experimental, fragment specific apparent exchange rates (A) and the residue specific, predicted rates (B).*

The apparent rates of the fragments in panel A were calculated as described before (131). The predicted rates of residues in panel B were calculated using the computed protection factors and intrinsic rates. Black box delineates the P-loop. Red box highlights the nucleobase binding hairpin. Residues without amide hydrogens are represented in black in panel B.

Another important insight from the predictions is that exchange at short times provides additional, valuable information about the dynamics of the system that cannot be inferred from exchange at longer times. Most common HDX-MS experiments, such as those available for φ12 P4 studied here rely on manual mixing, and the shortest time at which the fraction of deuterium exchanged is measured is of the order of ten or more seconds. The estimation of the kinetics of deuterium exchange on subsecond times (Figure 3-3A) reveals that fragments with similar exchange kinetics on the timescale of, e.g., seconds and minutes, may have very different kinetics at shorter times. A pertinent case is the comparison between the kinetics of experimentally observed fragment 20 (residues 284-293) and the fragment 292-302. Their kinetics are almost identical in the range of time 30s-4h whereas they are clearly distinguishable on a shorter timescale. In fact a time-resolution of about 10-100 ms, accessible by conventional rapid quench flow apparatus (82), would cover the relevant exchange kinetics while little information would be obtained on shorter timescale. This time scale is also relevant to the overall turnover rate (~6 s^{-1}) of the enzyme and quantitative prediction of exchange from a population of modelled states on this time scale will be essential in making use of HDX to monitor and interpret conformational changes associated with motor action.

Most of the theoretical models interpret HDX exchange kinetics obtained by NMR at the residue level for relatively small proteins (75). Only a few methods attempted to predict deuteration measured by HDX-MS and were limited to comparison with experimental data at only one time point (82, 134). As illustrated here, reliable simulations of the entire experimental kinetics allow to extract the residue specific protection factors at different amide sites within each fragment (see e.g. Figure 3-6), thus enriching information content of the HDX-MS results and providing direct link to the sequence, e.g. by informing site-directed mutagenesis experiments.

The COREX (133) method is based on populating protein microstates in which each residue is either in fully folded (protected) or fully unfolded (exchangeable) state. Contribution of these microstates to exchange is then weighted according to their relative stability. This method, albeit computationally intensive, is effective in predicting HDX-MS. One limitation is that in its present form the COREX approach ignores regions that are not resolved in the high-resolution structure. In addition to missing predictions for such regions this omission from the model may affect exchange of the neighbouring sites. In our approach this issue is dealt with by modelling the missing regions within the context of the whole structure, using MD to relax the model and, importantly, calculate protection factors as Boltzmann averages. However, as illustrated by the C-terminal helix case here the quality of the initial model plays crucial role in the success of this approach since the relatively short duration of the MD run does not account for larger conformational changes that occur on longer timescales. An iterative approach in which different models of the missing regions are tested and the simulation results compared with experiment may yield a complete, plausible structure.

3.6 Limits of the phenomenological approximation of protection factors

In this section, the limits of the approximation are investigated and new approaches to improve the current HDX model are discussed.

3.6.1 Transferability of parameters B_c and B_h

Throughout this thesis, all protection factors are predicted using Equation (**1-4)** with the parameters ($\beta_c = 0.35$; $\beta_h = 2.0$), and averaged over an ensemble of structures produced in the CHARMM36 force field. However, these parameters were optimized for a different force field. I checked that the parameterisation of Equation (**1-4)** remains valid in CHARMM36.

Best *et al.* (83) obtained the parameters $\beta_c = 0.35$ and $\beta_h = 2.0$ by fitting a set of proteins for which experimental protection factors were well established and their structures available. The native basin of each protein was sampled by collecting 1ns MD simulations using the CHARMM19/EEF1 force-field (140). This force field treats solvation effects implicitly to limit computational costs. Although progress have been made to develop more accurate implicit solvent models (141, 142), explicit models remain less questionable. Hence, fluctuations of φ12 P4 around its native state were sampled in CHARMM36 – the current state of the art of available force-fields – for which solvent is treated explicitly.

The optimal values for the coefficients β_c and β_h obtained using a united-atom model such CHARMM19/EEF1 may not be optimal when using an all-atom model such as CHARMM36. For this reason, the coefficients β_c and β_h were re-optimized based on the experimental data of five proteins used by Best *et al.* These proteins included barnase (143), ribonuclease H (144), equine lysozyme (73), human α-lactalbumin (145) and basic pancreatic trypsin inhibitor (146). MD simulations were performed in CHARMM36 with explicit water molecules. The free-energy landscapes of biomolecules in explicit solvent is more rugged than in implicit models (147). Multiple free-energy barriers hamper sampling of the conformational space. Therefore, the produced runs were at least 100 ns long (Figure A - 6), i.e. 2 orders of magnitudes longer than simulations produced by *Best et al.* (83) The mean square deviation (see Appendix) of the predicted protection factors from experimental measurements was calculated for varying values of β_c and β_h using a grid search. Very slow or fast exchanging residues, for which accurate protection factors were not available due to limit in experimental methods, were ignored during fitting.

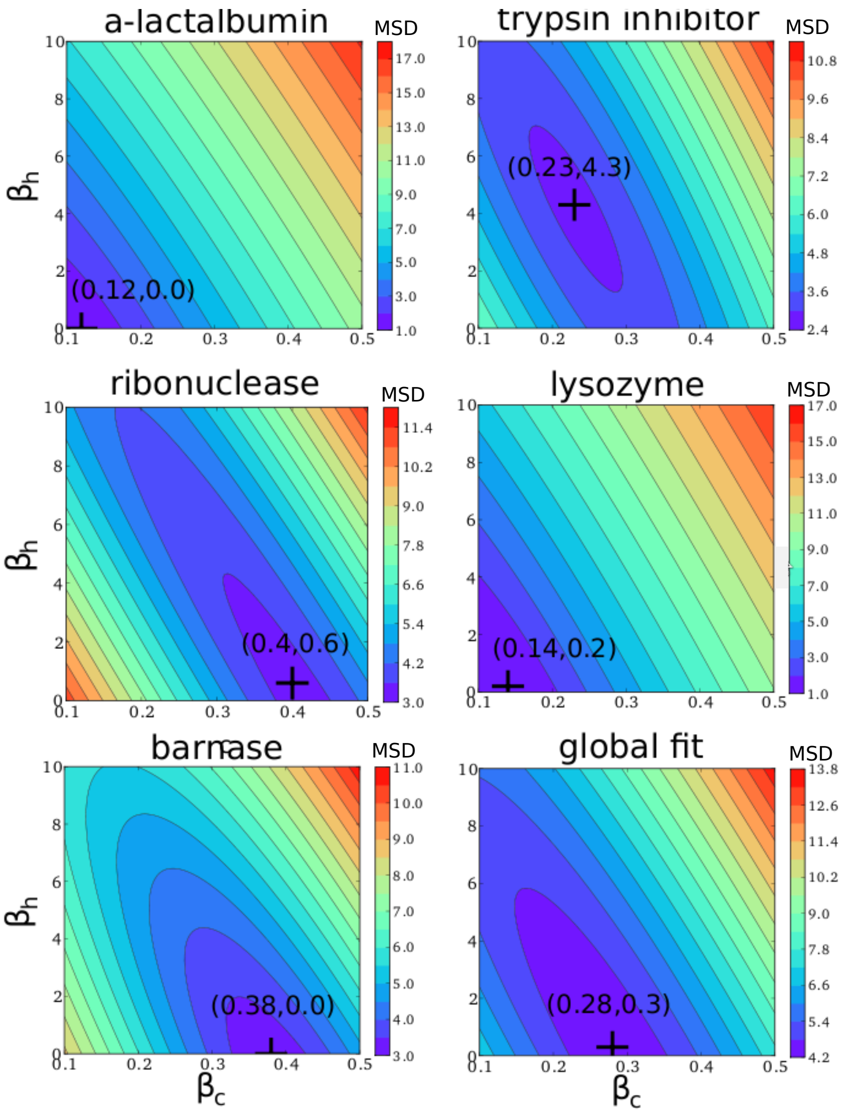


Figure 3-7 *: Fitting of the parameters $\boldsymbol{\beta_c}$ and $\boldsymbol{\beta_h}$.*
Contour plots of the mean square deviations (MSD) between experimental and predicted protection factors over the parameter space (β_c, β_h). Protection factors were averaged over the native state simulation in CHARMM36 and explicit solvent. Optimisation was performed individually for each protein, or globally for all proteins (bottom right panel). Optimal values of parameters are indicated by a black cross. Blue and red areas represent low and high MSD, respectively.

The mean square deviation over parametric space (β_c, β_h) is shown in Figure 3-7 for each protein and for the global fitting. The individual optimal parameters can be significantly different from protein to protein. The optimal β_c value can vary from 0.12 to 0.4 whereas the optimal β_h ranges from 0 to 4.3. This broad range of values leads to a global fitting with a large and shallow minimum basin (bottom right panel in Figure 3-7 and Figure 3-8B). Hence, parameterisation of Equation (**1-4**) in CHARMM36 is relatively flexible and the optimal coefficients ($\beta_c = 0.28$ and $\beta_h = 0.3$) do not give a significantly better agreement than the parameters found by Best *et al.* (Figure 3-8B and C). For this reason, throughout the thesis, protection factor calculations were always performed with the previously published coefficients $\beta_c = 0.35$ and $\beta_h = 2.0$. Another interesting feature is the "stretched" shape of the minimum basic: it indicates that the number of contacts, N_c, and hydrogen bonds, N_h, associated to β_c and β_h, respectively (see Equation (**1-4)**), are highly correlated. The correlation is such strong that optimizing Equation (**1-4**) with only one parameter, i.e. $\beta_h = 0$ (see optimization in Figure A - 10), gives predictions as good as with two parameters (Figure 3-8 B and D). Indeed, in both cases the resulting MSD is equal to 4.3.

The transferability of parameters of one force field to another is a sign of robustness. However, their transferability from one protein to another is limited (Figure 3-8 A and B). The overall agreement between prediction and experiment is good enough to detect large conformational changes such as the opening of the φ12 P4 ring (see Chapter 3). However, studying finer mechanisms involving smaller motions would require a more advanced and accurate HDX model. The next section discusses the theoretical framework that can be used to improve the model.

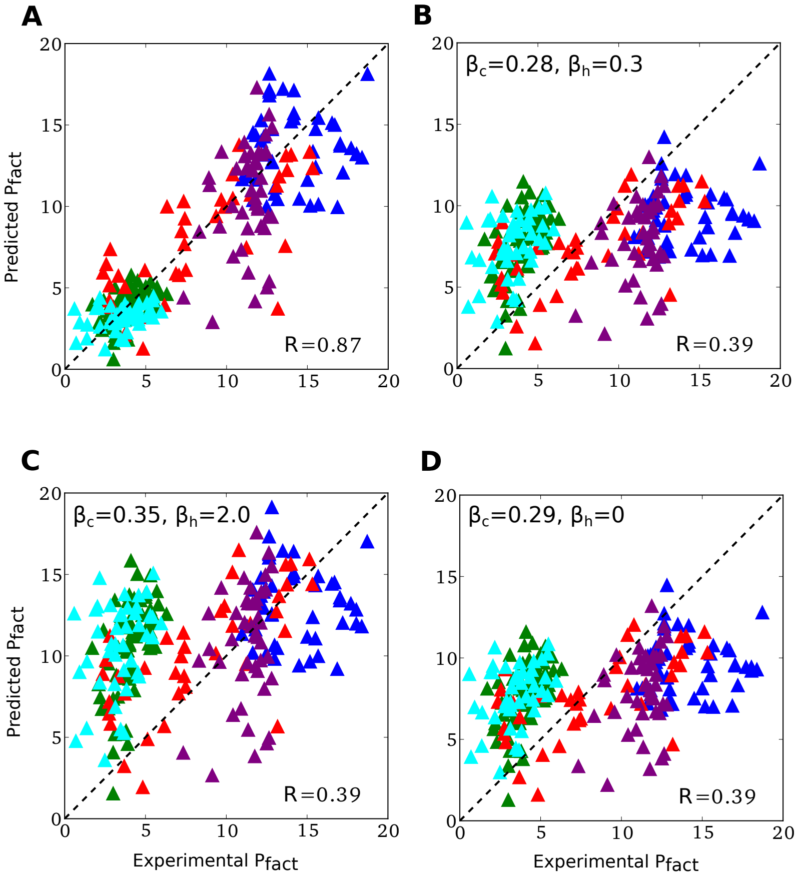


***Figure 3-8** : Comparison of measured and predicted protection factors.*
Correlation plots of a-lactalbumin, trypsin inhibitor, ribonuclease, lysozyme and barnase are displayed in blue, green, red, cyan and purple, respectively. The protection factors were calculated using (A) the optimal coefficients β_c and β_h found for each individual protein and reported in Figure 3-7 (i.e. coefficients were different for each protein), (B) the optimal coefficients for CHARMM36, (C) the coefficients found by Best et al. and (D) the optimal parameterisation when only heavy contacts are used. The black dashed line represents perfect match with experiment. The overall agreement between prediction and experiment is very similar when we use optimal parameters for CHARMM36, or parameters from Best et al. or only one paramter, with respective MSD equal to 4.3, 5.3 and 4.3.

3.6.2 Accounting for electrostatic effects in HDX mechanism

HDX occurs mainly via water ion catalysis, rather than direct exchange with water. The rate constant of the exchange, k_{ex}, can be expressed as:

$$k_{ex} = k_{H_2O} + k_{H_3O^+}[H_3O^+] + k_{OH^-}[OH^-]$$

where k_{H_2O}, $k_{H_3O^+}$ and k_{OH^-} are water, acid and base-catalysis rate constants, respectively. Catalysis by a water ion involves a charged transition state and charged intermediate state (Figure 3-9A). The local presence of a charge can stabilize or destabilize this transition state, resulting in a lower or higher free-energy barrier and therefore in an acceleration or slowing down of the exchange (Figure 3-9B). This effect is totally ignored in our model, apart from nearest-neighbouring charges via calculation of the intrinsic rate.

The electrostatic field induced by the two nearest-neighbour side chains of the NH group is well known to influence the exchange rate of the amide hydrogen (112). This phenomenon has been empirically introduced in the calculation of the exchange rate of random coil conformations. However, the electrostatic field surrounding the NH group is different in a structured protein, where other charges can locate close to the exchanging amide. Those charges can be divided into two groups: the background and the titrating charges. Background charges are partial charges carried by the peptide bonds and non-titrating polar groups, such as the hydroxyl group of serine and threonine, or charges carried by metal ions. Titrating groups are present in side chains of arginine, histidine, lysine, aspartic acid, glutamic acid or in free cysteines and termini. Through modification of the charged state of the titrating groups, pH can modify protein electrostatics. Beside its direct impact on the local electric field surrounding the NH group, it can also affect the overall stability of the protein and therefore the protein's structure which determines the local environment of the amide.

To estimate the impact of the local charges to the prediction of the protection factor, a comparison between predicted and experimentally measured protection factors of the five proteins mentioned hereinabove was carried out. The averaged error between predicted and measured protection factors was calculated as a function of the number of charged side chains surrounding the amide hydrogen. (*Figure 3-10*). The figure *Figure 3-10* shows that discrepancy tends to increase with the number of local charges. This result supports the idea that a model integrating electrostatic effects would improve the overall agreement with experiments by correcting the abnormal error for NH groups surrounded by charges.

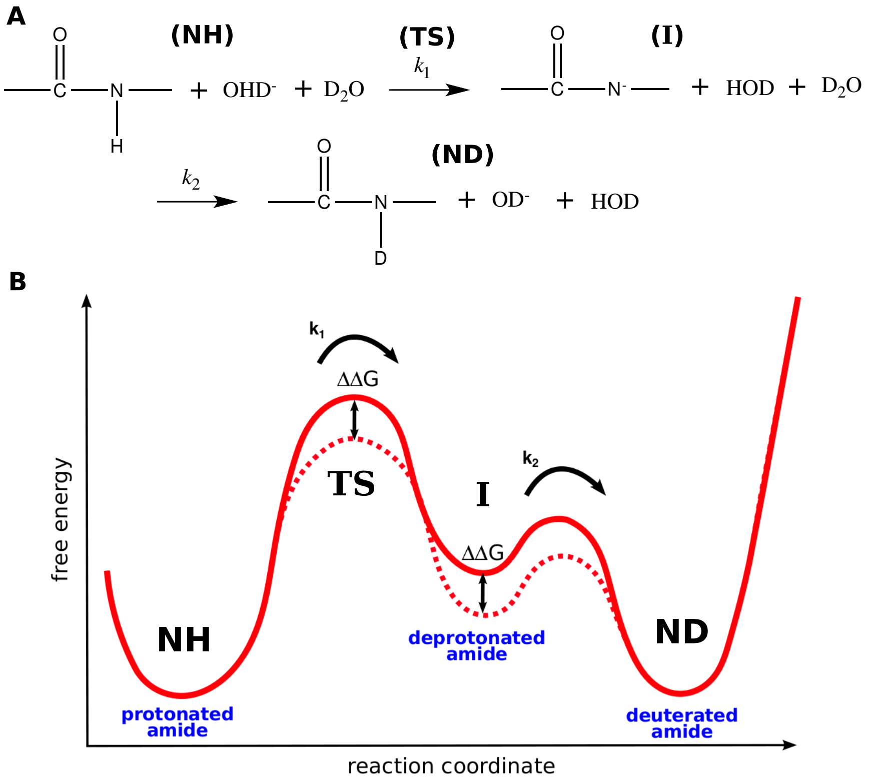


***Figure 3-9** : Electrostatic effects on the free energy profile of the catalysed hydrogen-deuterium exchange reaction.*

(A) Chemical equation of the based-catalysed hydrogen exchange reaction. (B) Free energy landscape of based-catalysed HDX in the absence of local charges (red curve) and in the presence of a positive charge (dashed red curve). Deprotonation of the amide nitrogen involves a negatively charged transition state (TS) and intermediate state (I). In the presence of a local positive charge, the free energy of TS and I are stabilized, resulting of a lower free energy barrier and therefore of an acceleration of chemical exchange. The deprotonated amide nitrogen reacts with neutral water at a diffusion-limited rate (k_2=$10^{10}$$M^{-1}s^{-1}$), making the capture of the amide hydrogen by a hydroxide ion (k_1) the limiting step of the ion-catalysed hydrogen exchange reaction. For this reason, electrostatic effects on the rate constant k_2 are ignored.

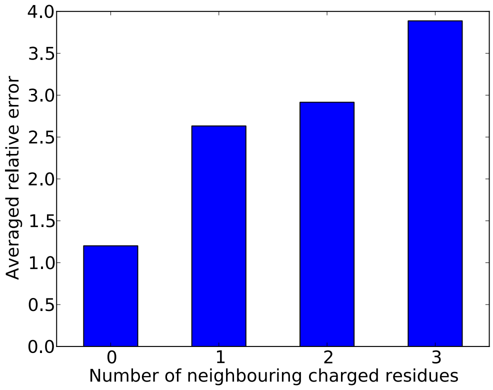


***Figure 3-10** : Averaged relative error between measured and calculated protection factors as a function of the number of local residues with a charged side chain.*

Residues having their charged side chain (arginine, lysine, aspartic acid and glutamic acid) within 4.5 Å of the amide hydrogen were considered as a charged neighbour.

Electrostatic field calculation

Before considering electrostatic effects on HDX, one needs to calculate the electrostatic field inside the protein. This task is difficult and simple models have been used. Two main approaches exist: the macroscopic models and the microscopic models. The former assumes that charges interact through a continuum medium characterized by its dielectric constant. In the latter approach, permanent dipoles (water molecules and polar side chain of amino-acids) and induced dipoles (electron cloud deformation) are treated explicitly such that they can change their orientations.

In 1924, Linderstrom-Lang proposed a protein electrostatic model with different dielectric constants inside and outside of the protein (148). In such a configuration, the Coulomb law is not valid and the Poisson-Boltzmann equation has to be solved (see Appendix). An analytical solution was provided by Linderstrom-Lang assuming that the protein was spherical and that its total charge was smeared uniformly over its surface. Tanford and Kirkwood improved the Linderstrom-Lang model to account for the protein's native structure (149). The protein is still assumed spherical, but the system (i.e. protein and solvent) is divided into two concentric spherical regions of uniform dielectric constant: a region with a low dielectric constant - representing the protein - surrounded by the solvent which has a high dielectric constant (Figure 3-11). Similarly, the system is divided into a region of zero ionic strength - corresponding to the interior of the protein with a boundary layer of solvent not accessible to ions - and a second region with an ionic strength equal to that of the bulk solvent. All pairs of atoms are placed on a sphere of radius *r* buried into the protein (see dashed circle in Figure 3-11). *r* is independent of the atom pair, i.e. that all charges are assumed to be at the same depth beneath the protein surface. The distance between the two atoms is determined according to their positions in the X-ray structure. A more advanced model, called "modified Tanford-Kirwood model", accounts for the solvent accessibility of the charges in the protein structure (150). In this new model, the electrostatic contribution of a charged transition complex at site *j*, F_{el_j}, to the total free energy of the protein, is given by (151):

$$F_{el_j} = \sum_i W_{ij}(1 - \frac{1}{2}(SA_i + SA_j))$$

where SA_i and SA_j are the static accessibility of the charged side-chains of residue *i* and *j*, respectively*.* The individual energy contributions, W_{ij}, are calculated from:

$$W_{ij} = Z_i Z_j \frac{e^2}{4\pi\varepsilon_0}\left(\frac{(A_{ij} - B_{ij})}{b} - \frac{C_{ij}}{a}\right)$$

where *a* and *b* are the radii of the spheres delimiting the areas with different ionic strength and dielectric constant, respectively (see Figure 3-11). Z_i and Z_j are the respective side chain charges of residues *i* and *j*. ε_0 is the vacuum permittivity and *e* is the charge of one electron. The expression of A_{ij}, B_{ij} and C_{ij} is given in (149).

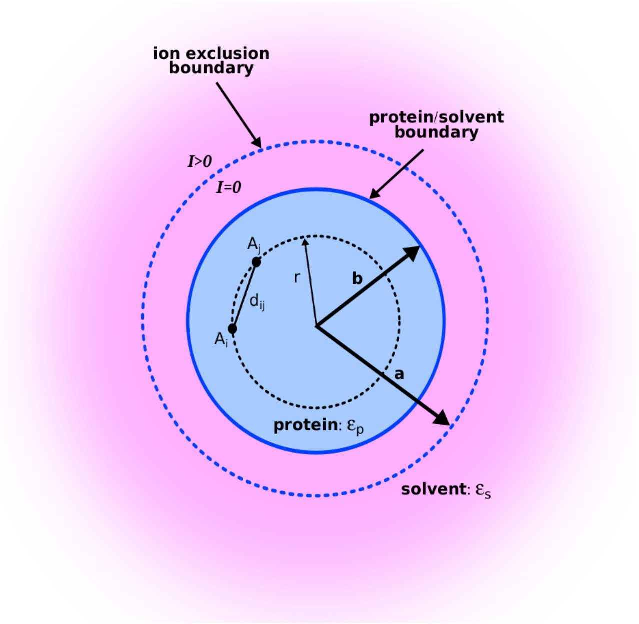


Figure 3-11 *: The modified Tanford-Kirkwood model.*
The system is divided by a first sphere (of radius b) that delimits the protein interior where the dielectric constant is low (ε_p) from the solvent where the dielectric constant is high (ε_s) and by a second sphere (of radius a) defining the ion exclusion boundary, i.e. the limit of the area accessible by ions. In this region, the ionic strength (I) is larger than zero, leading to an electrostatic shielding. The intermediate area surrounding the protein, that counter ions cannot reach, is called the water shell. Charges are distributed at the surface of a unique sphere of radius r buried into the protein. The distances between two charged points, d_{ij}, is equal to their distance in the X-ray crystal structure of the protein. The model was later improved by distributing charges on a grid according to their X-ray coordinates.

Purely macroscopic models might be inadequate for proteins because they disregard the local polarity defined largely by the net orientations of protein permanent dipoles. In the Protein Dipole Langevin Dipole model (PDLD), in addition to the Coulombic interaction, point dipoles are associated with each atom (152). The magnitude and the direction of each dipole are calculated iteratively. In a first step, the electric field on each atom is calculated using only fixed charges in the protein. Induced dipoles are initially calculated as the product of atomic polarisability and local electric field. Then, the electric field induced by these dipoles is added to the total field. The coupling between electric field and induced dipoles is refined iteratively until it converges. This Protein Dipole (PD) model is combined with a Langevin Dipole (LD) approach, which models electrostatic effects of the solvent using the Langevin equation to calculate the water dipole (see Appendix). A shell of hydration is treated explicitly and the surrounding region is covered by a grid of point dipoles, their density being equivalent to the density of bulk water molecules. The dipole amplitude at each point is approximated by a Langevin function (see Appendix), with its orientation in the direction of the local polarizing electric field. This model is more accurate than the continuum ones but comes with the cost of lower computational efficiency.

Effect of the electric field on HDX

Previous studies have already attempted to investigate the effect of the electrostatic field inside the protein on HDX. A first model was proposed by Delapierre *et al.* (153). They calculated the electrostatic field with the modified TK model (Figure 3-11). However, they ignored the impact of hydrogen bonds on the HDX kinetics. Dynamical effects on HDX were introduced in a very simplistic and empirical way. A few years later, Matthew *et al.* used a similar approach based on a model which had been developed to study pKa values of titrating groups in proteins (154). Their main conclusion was that electrostatic effects on HDX were mostly due to alteration in protein stability rather than to variations of the acid- base-catalysis rate. The most advanced model to study electrostatic effects on HDX was proposed by Le Master (155, 156). His approach integrates explicitly protein dynamics by using an ensemble of structures produced by NMR-restrained molecular dynamics. The electrostatic potential is calculated explicitly using the Delphi linearized Poisson-Boltzmann program (157). Interestingly, ensemble averaging is performed by averaging rate constants instead of their respective pKa values (or equivalently the electrostatic potentials). The method was able to predict robustly the hydrogen exchange of amides which were solvent exposed in the high-resolution X-ray crystal structures.

Integration of electrostatic effects in our model

The model of Le Master presented above was only assessed against solvent exposed amide hydrogens that do not require a local unfolding of the protein. Hence, neither the burial of the amide nor the hydrogen bonds were considered. By combining the approach presented in Section 2.2 with the approach of Le Master, one can obtain an accurate HDX model integrating simultaneously effects of temperature, protein dynamics, hydrogen bonds, electrostatics and pH. The hydrogen exchange rate of an NH group in a protein conformer *X* can be expressed as:

$$k_{ex}(T,pD,X) = \frac{k_{D_2O}(T) + k_{D_3O^+}(T,pD,X)10^{-pD} + k_{OD^-}(T,pD,X)10^{pD-\log(K_D)}}{\beta_c N_i^c(X) + \beta_h N_i^h(X)}$$

where K$_D$ is the D$_2$O dissociation constant and $k_{D_3O^+}$, k_{OD^-} and k_{D_2O} the second order rate constants for catalysis by D$_3$O$^+$, DO$^-$ and D$_2$O, respectively. N_i^c and N_i^h are the number of heavy contact and hydrogen bonds the amide hydrogen is involved in. The base- and acid-catalysis rates are corrected to account for electrostatic effects:

$$k_{D_3O^+}(T,pD,X) = k^0_{D_3O^+}(T,pD)e^{\Delta F^+/RT}$$

$$k_{OD^-}(T,pD,X) = k^0_{OD^-}(T,pD)e^{\Delta F^-/RT}$$

where $k^0_{D_3O^+}$ and $k^0_{OD^-}$ are the ion catalysis rate constants defined in Section 2.2.2 for a random coil. $\Delta F^{\pm}$ is the apparent activation energy conferred by the protein charge array on the acid- and based-catalysed step. This term can be written as a sum of two independent free-energy contributions:

$$\Delta F^{\pm} = \Delta F^{\pm}_{q} + \Delta F_{el}$$

where ΔF_{el} is the electrostatic component of the overall protein stability, and $\Delta F^{\pm}_{q}$ the free energy required to bring the catalytic ion to a particular amide in the presence of the electrostatic field. Assuming that the additional positive or negative charge present in the transition state has a small effect on the charge distribution of the rest of the protein, one has:

$$\Delta F^{+}_{q} = -\Delta F^{-}_{q} = \Delta F = q\phi$$

where q is the charged of the transition state and ϕ the local electrostatic field experienced by the transition state. The local electrostatic field can be calculated using either the software Delphi (157), based on the Poisson-Boltzmann equation, or the software MOLARIS (152), based on the PDLD model. Since $k_{D_3O^+}$ and k_{OD^-} already account for electrostatics effects on the neighbouring residue, ΔF calculation shouldn't integrate the charges from the right and left side chains of the NH group.

The high correlation between *N^c* and *N^h* suggests that the definition of a hydrogen bond used to calculate *N^h* is too extensive. In addition to the distance criteria, the donor-hydrogen-acceptor angle should be considered. A more accurate definition of the number of hydrogen bonds a residue *i* is involved in, N_i^h, would be:

$$N_i^h = \sum_{j \in O_i} \frac{w(\theta_{ij})}{1 + \exp\left(10 * \left(r_{ij} - 2.4\right)\right)}$$

where r_{ij} is the distance between the amide hydrogen of residue *i* and atom *j* (in Angstroms), O_i is the list of all oxygens not included in residues *i-1* or *i*, θ_{ij} is the nitrogen-hydrogen-oxygen angle and w is defined as:

$$w(\theta_{ij}) = \begin{cases} \cos(\theta) & if\ \theta \geq 90° \\ 0 & if\ \theta < 90° \end{cases}$$

The parameters β_c and β_h would be re-optimized as described in Section 3.6.1.

It is worthy to note that we are reaching the limits of the two-states model. The two-state model assumes that the protein is locally unfolded during the exchange, such that the local environment of the residue is equivalent to the one encountered in a random coil. However, in the method we are introducing electrostatic effects which are calculated based on the structures adopted by the protein in the native state. Very likely, the exchange occurs in an intermediate state between the random coil and the native structure.

This model could easily be tested by recalculating the protection factors for the five proteins mentioned in Section 3.6.1 and verify whether the relative errors illustrated in Figure 3-10 are reduced.

Chapter 4: Insights into helicase–RNA interactions from hydrogen exchange and fluorescence spectroscopy

4.1 Introduction

As mentioned in Chapter 1, P4 proteins are active portals that unwind and translocate single stranded RNA into a preformed virus capsid. Packaging of the RNA molecule initially requires loading of the nucleic acid chain into the central channel. A full description of how RNA is loaded into the pore of the packaging motor P4 has remained elusive. In φ8 P4, the L1 loop that encompasses the LKK motif RNA binding site (78), is localised within the central channel (Figure 4-1B). Given that RNA cannot directly reach the loop, it was proposed that loading occurs via opening of the ring (78), as has been established for other helicases (158, 159). In the previous chapter, a quantitative analysis of HDX-MS data of φ12 P4 confirms that the ring opens spontaneously when free in solution. The transient opening of the ring preceding translocation shown by HDX-MS, together with the ability of P4 to bind circular genomes (160), supports the idea that RNA loading proceeds via opening of the ring. However, a recently published X-ray crystal structure of φ8 P4 suggests a more complicated mechanism (17). Unlike other P4 helicases, the C-terminus of φ8 P4 plunges inside the pore, restricting the entrance and occluding the subunit interface through which RNA is thought to be loaded (Figure 4-1A) (17). In addition to this distinctive structural feature, it has been shown that φ8 P4 loses its activity when its C-terminus is truncated (17), suggesting that it may also play an active role in loading rather than just through passive occlusion. A tight coupling between RNA binding and ATPase activity presumably reflects a difference in RNA loading mechanisms between φ8 P4 and other P4 helicases, which can hydrolyse ATP in the absence of RNA, albeit slowly. It has been postulated that the C-terminus comes out from the pore upon RNA loading and somehow activates the enzyme. A qualitative analysis of the hydrogen-exchange kinetics of the C-terminus probed by mass spectrometry supported this hypothesis (78).

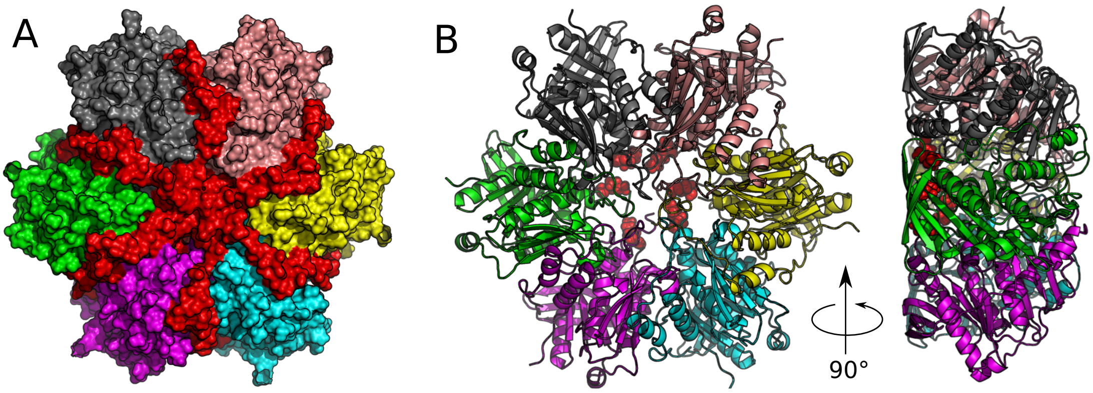


***Figure 4-1**: X-Ray crystal structure of φ8 P4.*
(A) Surface of φ8 P4. The C-terminal tail (in red) totally obstructs the entrance to the pore, as well as the subunit interface through which RNA is thought to be loaded. The C-termini may come out and bind the virus capsid. (B) Top (left) and lateral (right) ribbon representation of φ8 P4. The base of the motor is thought to sit on the capsid. Red spheres represent lysine 185 of L1 loops, which bind to RNA. Their location inside the central channel does not allow them to directly bind RNA.

4.2 Overview of the Chapter

In this Chapter, the conformational changes of the C-termini of φ8 P4 upon RNA binding are investigated. To test whether the C-termini remain inside or move outside the pore when RNA is loaded, the two scenarios are assessed using HDX data and quantitative interpretation of the kinetics using the method developed in Chapter 3. Two structural models of φ8 P4, with an RNA strand in the centre of the channel and C-termini either inside or outside the pore, are constructed. For each model ~200 ns MD simulations have been performed. Predictions for both models are compared quantitatively with available HDX-MS data (78). Our results demonstrate that only full exposure of the C-terminus to solvent can explain the fast exchange kinetics observed experimentally upon RNA binding. HDX only provides ensemble-averaged information. The dynamics of individual C-termini cannot be distinguished and it is not clear whether all C-termini come out or only some of them. To build a more detailed picture of the dynamics of the C-termini, their conformational changes are probed by single-molecule fluorescence spectroscopy. Experimental and computational observations suggest that loading and translocation mechanisms of φ8 P4 are different from those assumed for φ12 P4. I put forward a new translocation model for φ8 P4 where the C-termini become crucial for the processivity of the motor. The revised model requires that a few C-termini (probably half of them) remain inside the central channel during translocation. When the lever switches to the “down” position and drags down the RNA chain (see Section 1.1.3), the tip of the C-terminal tail competes with RNA to bind to the lever. This triggers a detachment of the RNA from the lever before the lever switches back to the “up” position. This mechanism may be necessary due to the higher affinity for RNA of φ8 P4 compared with other P4 motors (131).

4.3 Method

4.3.1 Construction of the structural models

HDX kinetics of φ8 P4 were collected both in the apo form and in complex with RNA (78). A crystal structure of φ8 P4 in the apo state is available (PDB access code: 4BWY). Missing sections (M1-D11, K185-V189, G281-I288, D301-G302, Y315-K321) were incorporated using MODELLER (Figure 4-2A). In order to add an RNA strand inside the central channel of φ8 P4, the apo structure was aligned with the X-ray crystal structure of the bacterial Rho factor (PDB access code: 3ICE). This homologous helicase was crystallised in complex with a poly(U) strand. Only six nucleotides of the strand were resolved. They formed almost one turn of a pseudo helix, narrower than a type A helix. Rho factor is distinct from other homologous helicases in the following aspect of RNA binding. Although it translocates from the 5' to 3' direction as P4 does, the motor is flipped around with respect to the RecA domain (161). The six nucleotide RNA strand was therefore rotated 180° around the axis perpendicular to the six-fold axis of the hexamer. Since HDX data of φ8 P4 were collected with poly(C), uracil bases were converted into cytosine bases. The strand was extended at both ends to come out from both sides of the pore (the six-nucleotide strand was copied and translated along the axis of the pore). For φ8 P4 in complex with RNA and with the C-termini inside the pore, the RNA strand was extended to 36 nucleotides (Figure 4-2B). Missing sections of the polypeptide chain were then incorporated with MODELLER. For the model with the C-termini outside the pore, the whole fragment F305-K321 was constructed outside the pore. To do so, for each subunit, the C-alpha atom of residue R314 was moved by 30 Å along the channel axis towards the top of the hexamer. Once residue R314 was fixed outside the channel, sections F305-R313 and R315-K321 were constructed with MODELLER. The RNA strand was slightly longer (48 nucleotides) than in the second model in order to reach the tip of the extended C-terminal domain (Figure 4-2C).

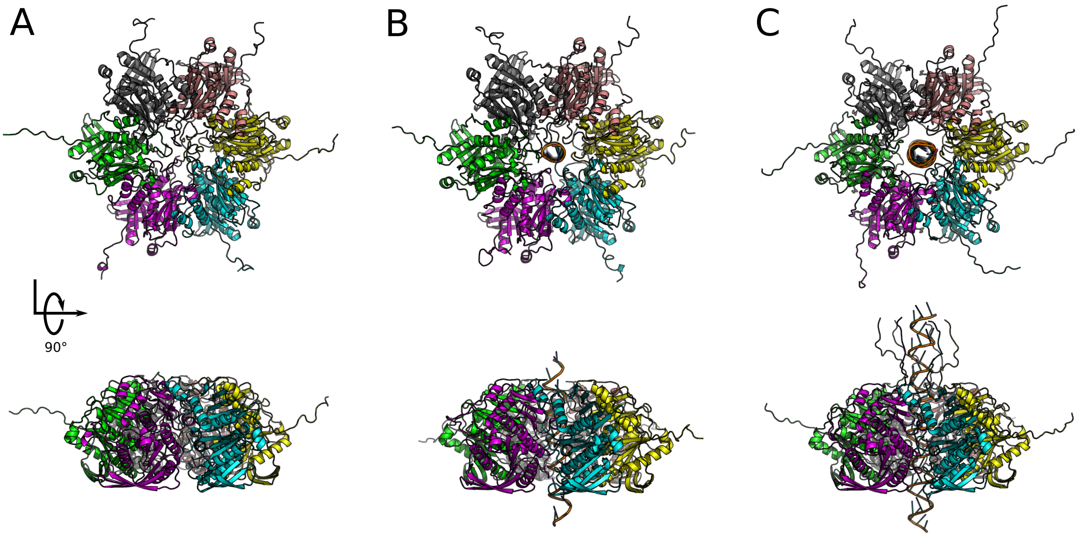


Figure 4-2: *Structures of the different models.*
Ribbon representation of the three models, with all the missing parts added. (A) Structure of the apo state of φ8 P4 (PDB Ref: 4BWY). (B) Apo state loaded on a short RNA strand with C-termini inside the pore. (C) Apo state loaded on a short RNA strand with the C-termini outside the pore.

4.3.2 MD simulations

Simulations of the φ8 P4 monomer and hexamer in the apo state or with RNA bond were performed with NAMD using the CHARMM36 force field; up to 194768 TIP3P water molecules were included to ensure that at least 10 Å separated the periodic images of the proteins as well as 506 Na^+ ions and 459 Cl^- ions to set the ion concentration to 0.15 M. The models described above were used as initial conformations (Figure 4-2). Simulations were performed at 298 K and atmospheric pressure. Periodic boundary conditions were applied and long-range electrostatic interactions were calculated with the particle mesh Ewald method, with a cut-off of 12 Å and grid spacing of 1 Å. Neighbour-atom lists were constructed including all atoms being less than 14 Å away from a given atom. A 2 fs timestep was used and conformations were saved every 500 timesteps (1 ps). The production runs were 100–280 ns long preceded by 20 ns of equilibration in which the temperature was increased from 0 to 298 K using 20 K increments every 500 ps (Figure A - 7).

4.4 Results

4.4.1 Hydrogen exchange predictions

Protection factors of φ8 P4 in the apo state were calculated using Equation (**1-4)** and averaged over the ~100 ns MD simulation. They were used to compute the time-dependant deuteration D(t) of each fragment based on Equation (**3-2)**. Although 28 fragments of φ8 P4 have been probed experimentally (Table A - 2), only the 20 non-redundant fragments, for which good quality experimental data was available, were considered. A direct comparison between calculated and experimental D(t) in the apo state is shown in Figure A - 10. In Figure 4-3 the D(t) values calculated from simulation (y-axis) are plotted against the experimental data (x-axis) of the hexamer in the apo state for each fragment and at time-points for which experimental data are available. Compared to φ12 P4, prediction for φ8 P4 is less reliable (R=0.59) with some fragments showing significant departure from the expected diagonal location (black line in Figure 4-3). However, the disagreements between the apo model and experiment appear to be very instructive. The most interesting cases are discussed below.

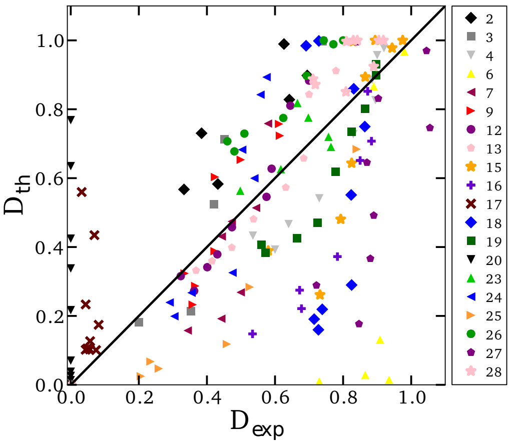


Figure 4-3: *Comparison of experimental deuterium fraction data of the apo state with prediction using the hexameric state of φ8 P4 without RNA inside the pore.*
Correlation plots for different fragments are displayed in different colours/markers. Diagonal line represents perfect match with experiment. See Table A-2 in Appendix for assignment.

For fragment 2 (R3-L24) our modelling predicted faster exchange than was found experimentally. This implies that the conformational space explored by the fragment in solution is different from that in simulation. Fragment 2 encompasses the N-terminus, for which sequence M1-D11 was not resolved in the X-ray crystal structure (PDB access code: 4BWY). We modelled the distal N-terminal portion as a flexible region. Interestingly, the mismatch is less pronounced for fragment 3 (I14-M25), the sequence of which includes that of fragment 2 except for the modelled portion. The mismatch suggests that we are underestimating the protection of the sequence M1-D11. Hence, it is likely that the distal N-terminus adopts a more structured conformation in solution than was modelled (likely a helix).

Fragment 28 encompasses the C-terminus of the protein. This fragment exhibits partial protection in the apo state and the fastest exchange within the protein upon addition of RNA (Figure 4-4). To explain this increase in exposure, El Omari *et al.* proposed a scenario where the C-termini are expelled from the central channel upon RNA binding (17). We put this scenario to the test by reinterpreting the HDX data with our quantitative method. In Figure 4-4, predictions from the three different models are compared with the two experimental data sets. The calculated exchange of the apo state (blue curve) exhibits partial protection in relatively good agreement with experiment (green dots). In the case of φ8 P4 mixed with RNA, only the model with the C-termini outside the channel (orange curves) can explain the instantaneous exchange observed experimentally (red dots). Interestingly, the presence of RNA inside the pore leads to faster exchange even when C-termini are kept buried inside the channel (black curve); but this increase is not significant enough to accurately fit with experiment (see explanation for this in discussion). Modelling allows predictions to be made for kinetics over a broader time-scale than accessible with manual mixing. It is clear that having access to millisecond to second time-scales would provide more valuable information to validate or reject the different scenarios. We also predicted the kinetics that would be expected for fragment 28 if half of the C-termini remained in the central channel upon RNA binding while the other half came out (dashed black curve). To model this intermediate scenario, we simply averaged the protection factors calculated in the presence of RNA with all C-termini either inside or outside the central channel. This illustrates further the advantage of having access to shorter time scales. Indeed, it turns out that over the time scale accessible by manual mixing (30 seconds - hours), the HDX kinetics of fragment 28 appear as fast with all C-termini out as they are with only half of them exposed. However, the two scenarios generate distinguishable kinetics on a time scale shorter than 10 seconds.

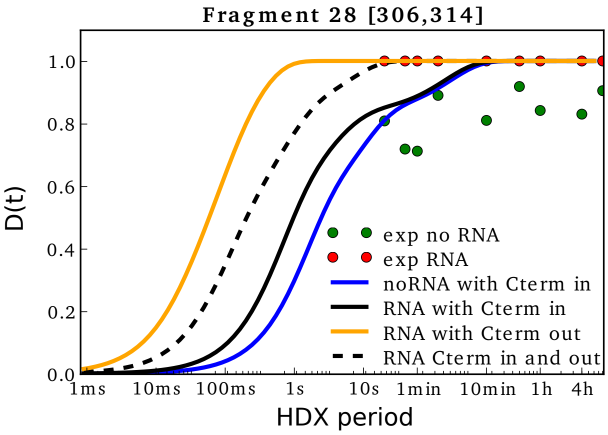


***Figure 4-4**: Assessment of the C-terminus models.*
Predicted deuterium fractions (average over the MD simulation) of fragment 28 for φ8 P4 in the apo state (blue line), or with RNA and the C-termini inside the pore (black line), or with RNA and the C-termini outside the pore (orange line). The dashed black line shows calculated exchange when half of the C-termini are assumed to remain inside the channel and the other half are assumed to come out. The experimental deuterium fractions of the helicase without and with RNA are shown as green and red dots, respectively. This fragment is the only one (out of the 28 available fragments), which encompasses the C-terminus. Calculated protection factors are shown over a broader time-scale than those accessible experimentally with manual mixing.

For fragments 6, 18 and 27 we predicted slower exchange than measured experimentally. All fragments are localized at the subunit interfaces (Figure A - 11). Contrary to other P4 helicases, φ8 P4 remains fully functional when not embedded into the virus capsid (131). Although the oligomeric state of φ8 P4 is particularly stable when isolated from the capsid, it is conceivable that its interface undergoes conformational fluctuations. These local breathings of the protein would result in the transient exposure of fragments normally buried between two subunits. In order to put this hypothesis to the test, we followed the same approach used for φ12 P4 in Chapter 3. φ8 P4 was simulated starting from the discrete state “open hexamer”, which was modelled as a single solvated monomer. Its local conformational space was sampled for 270 ns (Figure A - 9) and the exchange kinetics of each fragment were calculated (Figure A - 10).

Although fragments 6, 18 and 27 exhibit faster kinetics in the “open state”, the predicted kinetics remain much slower than the experiment. This prediction comes as no surprise knowing that fragment 6 encompasses a ß-strand, and fragments 18 and 27 contain helices; two secondary structures involving extensive hydrogen bond networks, and therefore high protection. Hence, a structural dynamic interpretation of the fast HDX kinetics of these fragments would require unfolding of their secondary structures, as well as important modifications of the motor’s quaternary structure. Moreover, exposure and disruption of the ß-sheet that is contained within fragment 6 would also further expose fragment 7, for which predictions already agree perfectly with experiment (see Figure A - 8).

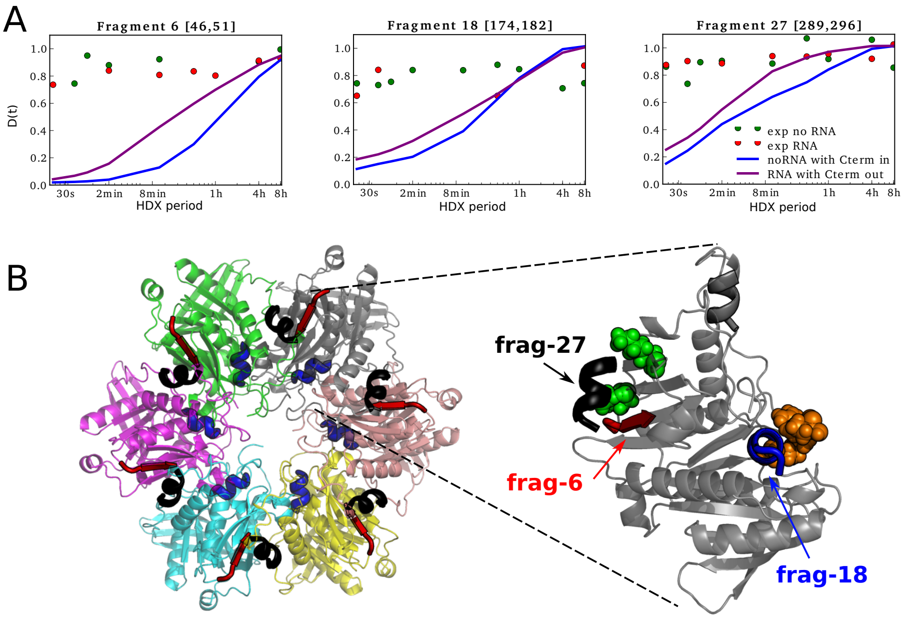


Figure 4-5: *Unexpectedly fast HDX: an electrostatic effect manifestation.*
(A) Predicted deuterium fractions (average over the MD simulation) for fragments 6, 18 and 27 in the apo state (blue line) or in the monomeric state (purple line). The experimental deuterium fractions of the helicase with and without RNA are shown as red and green dots, respectively. (B) Mapping of the fragments on φ8 P4 hexamer and subunit structures. Fragments 6, 18 and 27 are highlighted in red, blue and black, respectively. In the zoomed-in view, green and orange spheres represent, respectively, lysines and arginines in the vicinity of the fragments. The fragment 27 shown in black in the zoomed-in view belongs to the adjacent subunit. The positive charges carried by the side chains of lysine and arginine can accelerate hydroxide-based HDX catalysis.

The slower prediction of the exchange kinetics of a fragment compared to the measured kinetics can be the indication of a conformational change of the fragment in solution that is not captured by the MD simulation. However, the limitations of the HDX model should not be ignored (see Section 3.6). As mentioned in Chapter 3, the approximations for HDX mechanisms ignore electrostatic effects. Interestingly, fragments 6 and 27 are adjacent in the hexameric state of φ8 P4 (Figure 4-5) and are in the proximity to two lysines (L63 and L68). Similarly, two arginines (R174 and R191) are in the vicinity of fragment 18. HDX-MS data were collected at pH 7.5 (78). At this pH, exchange is dominated by hydroxide ion catalysis, which is accelerated in the presence of positive charges (see Section 3.6.2). Hence, mismatches observed for fragments 6, 18 and 27 may well be a manifestation of electrostatic effects on HDX.

For fragment 17 and 20 we predicted faster exchange than measured experimentally (Figure 4-6). Even after 8 hours, no exchange is observed by MS for this fragment. Both fragments encompass parts of the hydrophobic core of the motor. Our HDX model may not be reliable for predicting exchange from highly protected regions, since its parameterisation is based on NMR data, for which long exchange time scales are ignored due to the experimental method. Hence, although the model minimizes exchange for fragment 17 and 20 (their HDX are much slower than for all other fragments), the calculated exchange will never show a full protection at this time scale (i.e. ~ several hours).

Interestingly, fragment 20 displayed an EX1 kinetics signature during the transient state corresponding to RNA loading (black dots in Figure 4-6). In the EX1 regime, the refolding of the fragment is so slow that all its residues exchange before the fragment returns to the folded state. In this regime, the observed exchange rate is equal to the rate constant of the cooperative unfolding event controlling the exposure of the fragment. As the apparent exchange rate was close to the RNA loading rate, the exposure of fragment 20 was proposed to be due to the opening of the ring upon RNA loading

(78). No crystal structure of φ8 P4 was available when HDX-MS data were published. In Ref (78), they localised fragment 20 at the subunit interface only based on sequence alignment with φ12 P4. A more quantitative analysis of the kinetics of fragment 20 shows that a simple opening of the ring cannot explain its observed rate. The exchange kinetics for fragment 20 was predicted for the "open state" as explained above (purple curves in Figure 4-6). No significant increase in the exchange for fragment 20 was observed compared to the "closed state" (blue curve). This comes as no surprise since the fragment is not directly localised at the subunit interface. If only the opening of the ring controlled the exchange of fragment 20, the predicted exchange in the "open state" (purple line) should be much faster than the observed rate (black dots). However, the predicted HDX kinetics in the "open state" is slower than the observed exchange. This suggests that further conformational changes occur upon RNA loading, leading to the disruption of the tertiary structure and maybe secondary structure of the fragment. Surprisingly, this direct involvement of RNA in the exchange kinetics of fragment 20 does not affect fragment 17, which remains fully protected upon RNA loading.

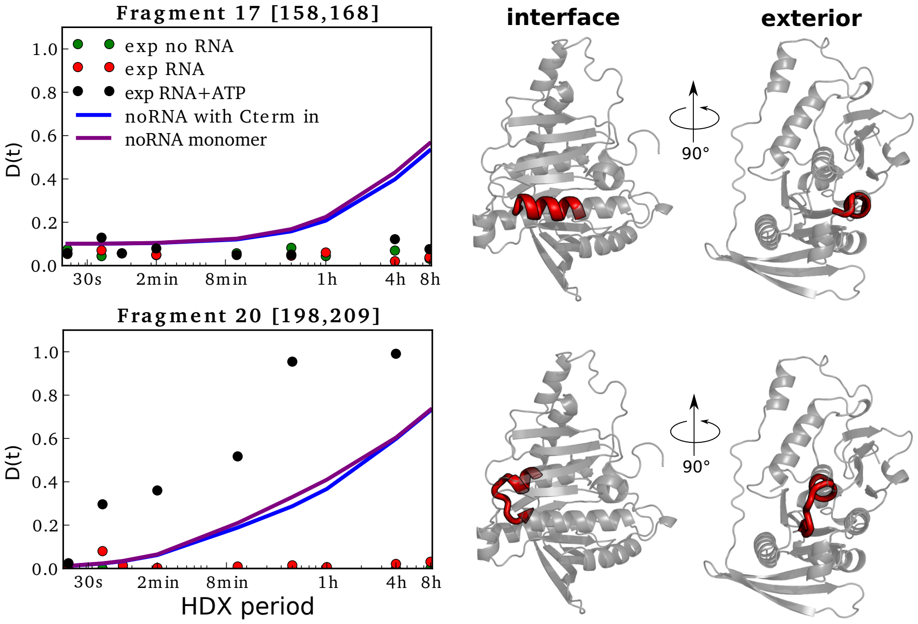


***Figure 4-6**: The fully protected fragments.*
In the left panel are depicted the predicted deuterium fractions of fragment 17 and 20 in the apo state (blue line) and in the monomeric state (purple line). The experimental deuterium fractions of the helicase with and without RNA are shown as red and green dots, respectively. Black dots represent experimental exchange when the helicase was mixed with RNA and ATP. A subunit of the hexamer is shown in the right panel. Both faces at the subunit interface and from the exterior are represented. Fragments are highlighted in red.

4.4.2 Fluorescence spectroscopy

To study further conformational changes of the C-termini upon RNA binding and to go beyond the ensemble averaged view provided by HDX, single molecule fluorescence spectroscopy was undertaken.

Design and purification of engineered P4 for fluorescent labelling

Studying φ8 P4 by single molecule FRET requires the site-specific labelling of the protein with fluorescent dyes, without interfering with its activity, *i.e.* labelled helicases should be able to translocate along RNA. To visualize the structural changes of the C-terminus upon RNA binding, φ8 P4 was labelled at different sites along the C-terminus. Alexa Fluor 488 (AF488) and 594 (AF595) were chosen as donor and acceptor dyes for their brightness, and because their excitation spectra match the 488 nm and 594 nm laser lines of the available instrumentation. Maleimide derivatives of AF488 and AF594 were used to specifically conjugate to cysteines. Each subunit of wild type φ8 P4 has only one cysteine (C128), partially exposed to solvent, that reacts with maleimide derivative dyes. Labelling of C128 has been shown to inactivate P4 (data not shown). Using site-directed mutagenesis, this cysteine was replaced with an alanine. To make purification of φ8 P4 easier, the protein was also His-tagged. Based on previous structures of φ6, φ12 and φ13 (17), the C-terminus of φ8 P4 was expected to be at the bottom of the helicase, which normally interacts with the virus capsid. For this reason, φ8 P4 was crystalized with a His-tag at its C-terminus. Surprisingly, the φ8 P4 X-ray crystal structure revealed that the C-terminus climbs along the protein and plunges inside the channel. In contrast, the N-terminus is disordered in the φ8 P4 crystal structure and HDX data show that the distal N-terminal domain is exposed to solvent. Hence, we chose to incorporate the His-tag at the N-terminus of the protein. His-tagged φ8 P4 C128A was constructed (Nterm-C128A), expressed and purified as described in Section 2.5. SDS-PAGE showed that about 70% of the Nterm-C128A stock was truncated. ATPase assays were performed for both the wild type and Nterm-C128A (Figure 4-7). Translocation of P4 along RNA can be inferred by the increase of phosphate concentration, which is produced during ATP hydrolysis. ATP hydrolysis reaction is catalysed cooperatively with the 6 active sites located at the subunit interfaces of the hexamer, resulting in non-Michaelis–Menten kinetics. The catalytic rates of both the wild type and Nterm-C128A were estimated by fitting their absorbance curves during steady states (linear portion in Figure 4-7 and Table 4-1). The constant rate of the wild type was similar to the previously published value (4). The activity of Nterm-C128A was three times lower, which was consistent with the degradation of 70% of the stock. Hence, it seems that the truncation of the construct adversely affected the functionality of the protein, rather than the addition of the His-tag or the substitution of the cysteine C128. For purifications of the next constructs, the concentration of proteases inhibitors was increased in order to limit degradation of the sample.

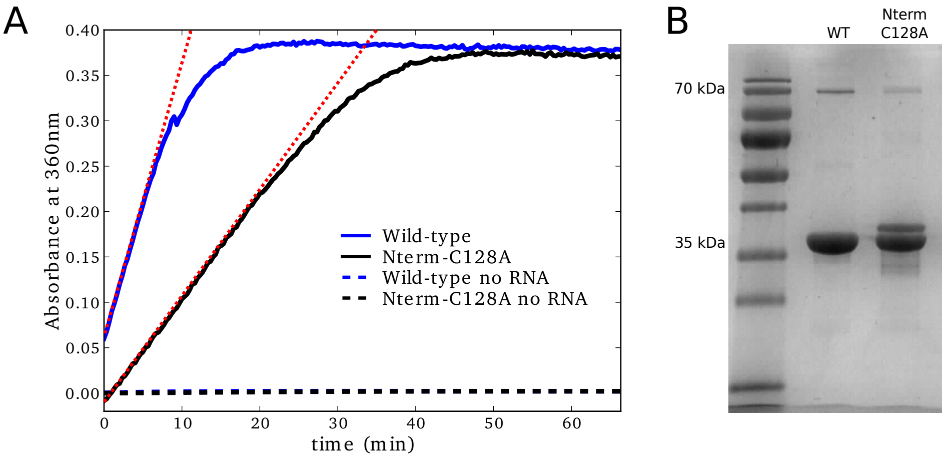


***Figure 4-7**: Functionality of φ8 P4 C128A with a His-tag at the N-terminus.*
(A) ATPase assay. Activity of the non-His-tagged wild type (WT) and the new construct (Nterm-C128A) are shown in blue and black, respectively. 0.5 µM of P4 was mixed with 0.5 mg/mL PolyA and 1 mM ATP. Controls without RNA were also run to check RNA ATPase activity induction (dashed lines). The two controls were both flat and are superimposed on the graph. Dotted red lines indicate fitting of the steady state part for estimation of k_{cat}. (B) SDS-gel of the wild type and the new construct, visualized by staining in InstantBlue®. The higher activity of WT (~2 times higher) is only due to degradation of more than 50% of the Nterm-C128A stock, as indicated by the double band on the SDS-gel.

Construct	k_{cat} (s^{-1})	
	Non-labelled	**Labelled**
WT	7.8	-
Nterm-C128A	2.6	-
285	4.0	1.7
287	1.6	0.8
290	5.6	1.2
304	7.0	7.6
Published WT	~8.0	-

***Table 4-1**: Initial rate of ATP hydrolysis reaction for labelled and non-labelled mutants. Turnovers were estimated by fitting with a straight-line absorbance at the steady state (i.e. from t = 0 s to t = 200 s). The catalytic rate constant is lower for the wild-type due to degradation of the protein stock. The published value for the WT is shown for reference (160).*

Next, selected alanine residues were substituted with a cysteine using site-directed mutagenesis either at position 285, 287, 290 or 304 along the C-terminus (Figure 4-8). All positions are readily accessible to solvent, a necessary condition for efficient labelling. 304 is located at the entrance to the channel, just before the C-terminus plunges inside. All mutants were expressed and purified as described in Section 2.5 (Figure 4-9 A and B). The purity and integrity of the mutants were confirmed by SDS-PAGE (Figure 4-9C). Partial degradation of mutant 287 was observed, even when purified with an increased concentration of protease inhibitors. Dimers were observed for all mutants when run on SDS-gel without reducing agent. Dimerization was most noticeable for mutant 304, which comes as no surprise since the cysteines in neighbouring subunits are close together in the hexameric state. The ratio between dimers and monomers increased up to 1:1 after one day at 4 ˚C (data not shown). Apart from the degraded mutant 287, all mutants exhibited an ATPase activity similar to those of the WT (Figure 4-11 and Table 4-1).

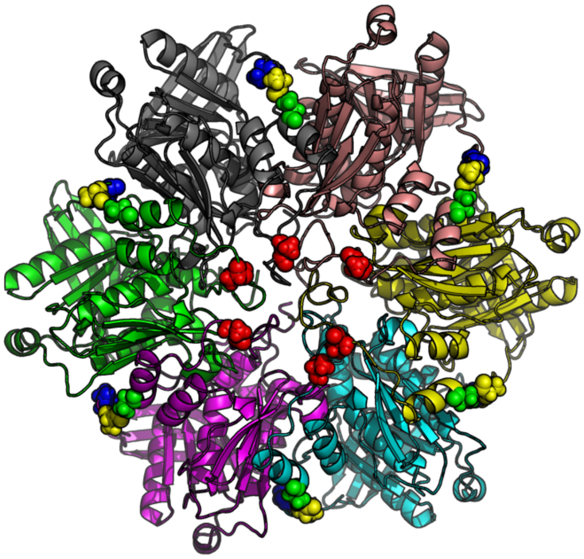


Figure 4-8: *Labelling of φ8 P4 at different positions.*
The unique cysteine of the wild type (C128) was replaced by an alanine and an alanine was substituted with a cysteine either at position 285 (blue spheres) or 287 (yellow spheres) or 290 (green spheres) or 304 (red spheres), along the C-terminus.

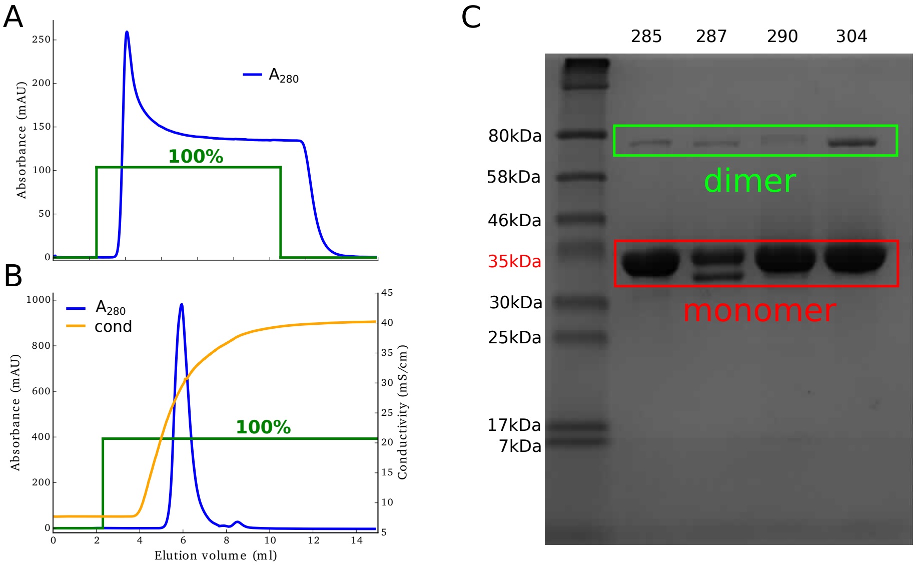


Figure 4-9: *Purification of φ8 P4.*
(A) Affinity chromatography purification of φ8 P4 (solvent A: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$, 75mM Imidazole; solvent B: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$, 1 M imidazole). Fractionation was monitored by absorbance at 280 nm (blue line). (B) Ion exchange chromatography purification of φ8 P4 (solvent A: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$; solvent B: 20 mM Tris-HCl pH 8.0, 400 mM NaCl, 7.5 mM $MgCl_2$). The conductivity is shown as orange trace. (C) SDS-PAGE of φ8 P4 mutants after purification visualized by staining in Instant Blue®. Loading buffer was DTT free so there was no prevention of disulfide bond formation between cysteines. All mutants tend to form dimers. Dimers were not observed when 2 mM DTT was added to the loading buffer.

Labelling of φ8 P4

φ8 P4 mutants were labelled with both AF488 and AF594, such that each hexamer would have on average one of the six C-termini labelled with a donor dye AF488, and another with an acceptor dye AF594. Labelling and purification were performed as described in Section 2.5.5 (Figure 4-10). The labelling step proved to be challenging. In order to study the effect of the dyes on the activity of the protein, the degree of labelling was maximized. The first attempts systematically resulted in a low yield and low degree of labelling. To improve the degree of labelling, the concentration of protein was increased up to 100 µM, with up to a 10-fold excess of dyes (500 µM AF488 and 500 µM AF594). Cysteines were reduced with 5 mM TCEP, a strong reducing reagent known not to react as readily with maleimides when compared to other common agents (DTT and BME). Since magnesium is known to interact with TCEP (162), Mg^{2+} free buffer was used. To limit oxidative dimerization, buffer was degassed and the reaction was performed in an inert atmosphere (nitrogen). Initially, the labelling reaction was performed at pH 7.5 and in 50 mM NaCl. Although φ8 P4 is soluble under these conditions, the labelled hexamers appeared to precipitate, probably due to the negative charges carried by the dyes that can significantly affect the isoelectric point of the protein (the predicted isoelectric point of non-labelled φ8 P4 is ~ 7.1). To ensure that labelled φ8 P4 remained soluble, the pH and sodium chloride concentration had to be increased up to pH 8.0 and 300 mM.

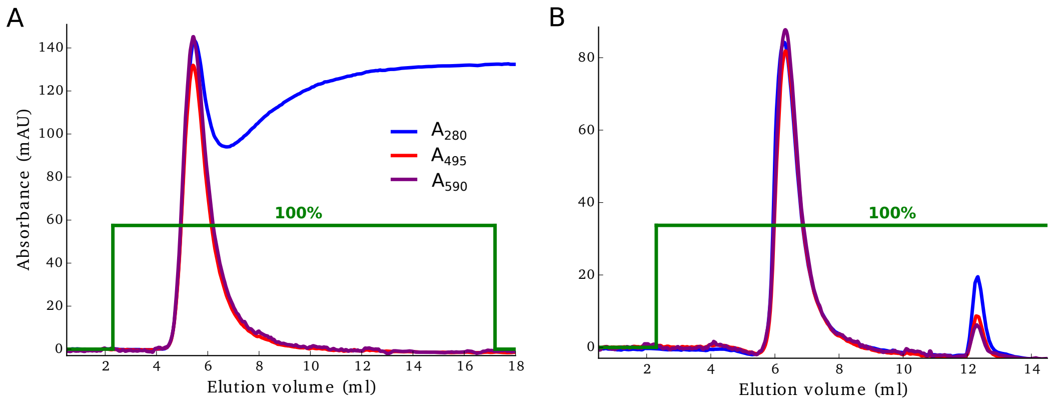


***Figure 4-10**: Purification of dual labelled φ8 P4.*
(A) Affinity chromatography purification of dual labelled φ8 P4 (solvent A: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$, 75 mM imidazole; solvent B: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$, 1 M imidazole). Fractionation was monitored by absorbance at 280 nm (blue trace), 495 nm (red trace) and 590 nm (purple trace). (B) Ion exchange chromatography purification of dual labelled φ8 P4 (solvent A: 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 7.5 mM $MgCl_2$; solvent B: 20 mM Tris-HCl pH 8.0, 400 mM NaCl, 7.5 mM $MgCl_2$).

Activity of labelled φ8 P4

The functionality of the labelled mutants was examined first. ATPase assays were performed for labelled and non-labelled mutants (Figure 4-11A). The activities of labelled mutants were quantitatively assessed by comparing their steady states with non-labelled mutants (Table 4-1). Adding a dye at position 285, 287 or 290 appeared to adversely affect the functionality of the protein. However, the ATPase activity of mutant 304 remains as high after labelling. This indicates that mutant 304 remains hexameric and functional when labelled. To confirm the functionality of mutant 304, its RNA translocase activity was also checked (Figure 4-11B). The RNA substrate was prepared by annealing an unlabelled 42-nt-long RNA strand with an AF488-labelled 29-nt-long DNA strand (see Section 2.5.6). The DNA strand was designed to target the first 21 nucleotides of the 3'-proximal region of the RNA strand as described in Figure 4-11B (78). The 5' termini of the RNA forms a single stranded overhang to which helicase can bind. The duplex was incubated with the labelled mutant 304 and ATP for 15 min at room temperature (see Section 2.5.6). If the helicase translocates along the RNA/DNA duplex, DNA would be displaced (78). A large excess of non-labelled DNA strands was also added, such that any displaced labelled DNA will not rebind due to the unfavourable competition between labelled and non-labelled DNA. The liberated labelled DNA probes were separated from the input mixture by PAGE under native conditions (163). The translocase assay shows that labelled mutant 304 can displace DNA probe from the duplex in the presence of ATP. Thus, labelled mutant 304 can translocate along RNA. As a control, a sample incubated without ATP was also run. Some displaced DNA probes were also observed in the absence of ATP although to a lesser extent. P4 was in large excess and it is very likely that multiple helicases bound to the same duplex. Hence, the helicase may liberate the probe by binding to the longer RNA strand and sterically displacing DNA (164).

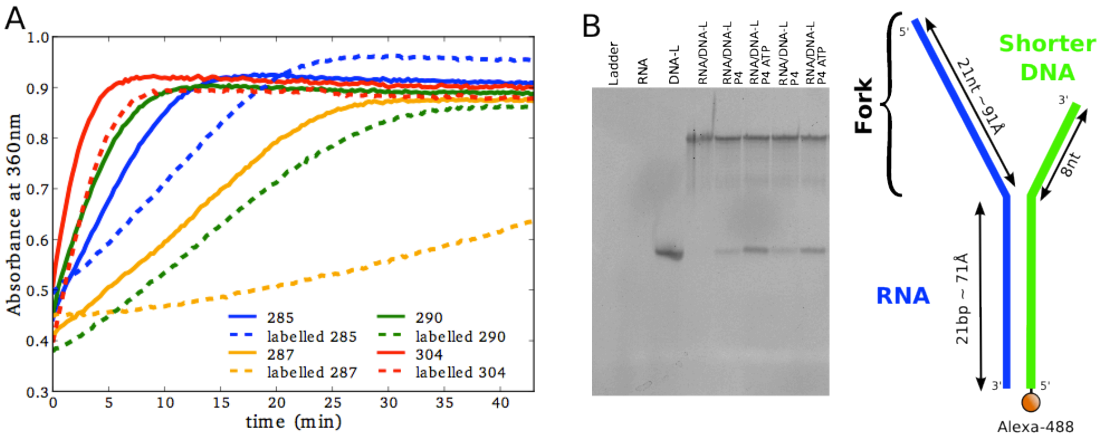


***Figure 4-11**: Activity of the labelled mutants.*
(A) ATPase activity of the mutants. ATPase assays were performed in parallel on the same plate, with 4 µM P4, 0.5 mg/mL PolyA and 1 mM ATP. The results for the mutants 285, 287, 290 and 304 are depicted in blue, orange, green and red, respectively. Activities of non-labelled mutants are depicted with solid lines and activities of dual labelled mutants with dashed lines. Apart from mutant 304, protein stops to function when a dye is attached to its cysteine. The low activity of non-labelled 287 is probably due to degradation of the stock. (B) Translocase assay of φ8 P4. All samples had an excess of unlabelled DNA strand.

Non-covalent interaction of P4 with the dyes

Initially, to separate the unreacted dyes from φ8 P4, we followed the protocol built for purification of non-labelled hexamer (Section 2.5.4). When purified, fluorescently labelled mutant 304 was run on a SDS-PAGE gel (Figure 4-12C, column “no wash”). Electrophoresis revealed a significant amount of AF488 and AF594 non-covalently bound to the protein, which became free upon denaturation in SDS and heating. AF488 and AF594 are hydrophobic and negatively charged dyes that may “stick” to P4. Indeed, P4 has a relatively high isoelectric point (~7.1) and therefore carries lot of positive charge. The purification protocol was modified to eliminate the “sticky” dyes. While labelled proteins were still loaded onto the nickel column, the column was thoroughly washed with 1 M NaCl and detergent (1% Tween20). Unfortunately, this did not help to eliminate the presence of non-covalently bound dyes and precipitation of the protein was observed.

In order to detect and eliminate the presence of free-dyes, single-molecule fluorescence correlation spectroscopy was performed. The autocorrelation function (ACF) of purified, fluorescently labelled mutant 304 was acquired and processed as described in Section 2.6.1. Previous small angle neutron scattering data have estimated R_h at about 51 Å (11). This value was compared with R_h calculated from ACF. A smaller value would indicate the presence of either free dyes or dissociated labelled-monomers, whereas a match would demonstrate that sample is clean and contains only hexamers. The apparent hydrodynamic radius, R_h, of fluorescently labelled mutant 304 was estimated either upon 488 nm or 594 nm excitation to assess the labelling for each dye. The ACF for the labelled protein was identical to the ACF of the free dye control. This confirmed that most of the dyes present in the sample were not covalently bound to the protein and also indicated that, although very “sticky”, dyes dissociate upon dilution of the protein to nanomolar concentrations. We made use of this phenomenon to improve the purification protocol. After labelling and quenching, the sample was diluted 100X into buffer A (20 mM Tris-HCl pH 8.0, 300 mM NaCl and 4 mM DTT), and kept for 1 h at 4 ˚C before being loaded onto the nickel column for purification. This new purification step is referred as the “1st wash”. As shown by the SDS-PAGE gel (Figure 4-12C, column “1st wash”), it allowed most of the free AF488 to be eliminated and reduced the quantity of free AF594. Single-molecule fluorescence correlation spectroscopy confirmed the decrease in non-covalently bound dyes (Figure 4-12 A and B, blue curves and Table 4-2). For excitation at 488 nm, R_h was now ~40 Å, instead of being identical to the R_h of free AF488 dye. This value is consistent with the hydrodynamic radius of φ8 P4 reported in (11). Hence, most of AF488 dyes are bound to the protein and P4 is mostly hexameric (with some monomers). For excitation at 594 nm, R_h was only ~27 Å, showing that a significant amount of free AF594 was still present. Therefore, a second purification step was performed, referred as the "2nd wash". Labelled proteins were diluted 100X into buffer A and kept at 4 ˚C overnight, followed by purification. No free AF594 could be detected by electrophoresis and R_h upon 594 nm excitation was now also ~40 Å (Table 4-2). Therefore, the labelled protein stock was deemed free of non-covalently bound dyes. The final degree of labelling of the protein was about 8% for AF488 and 4.6% for AF594. Hence, about 9% of the hexamers were dual labelled, *i.e.* had at least one monomer labelled with AF488 and one monomer labelled with AF594 (see Annex for calculation of proportion of dual labelled hexamers).

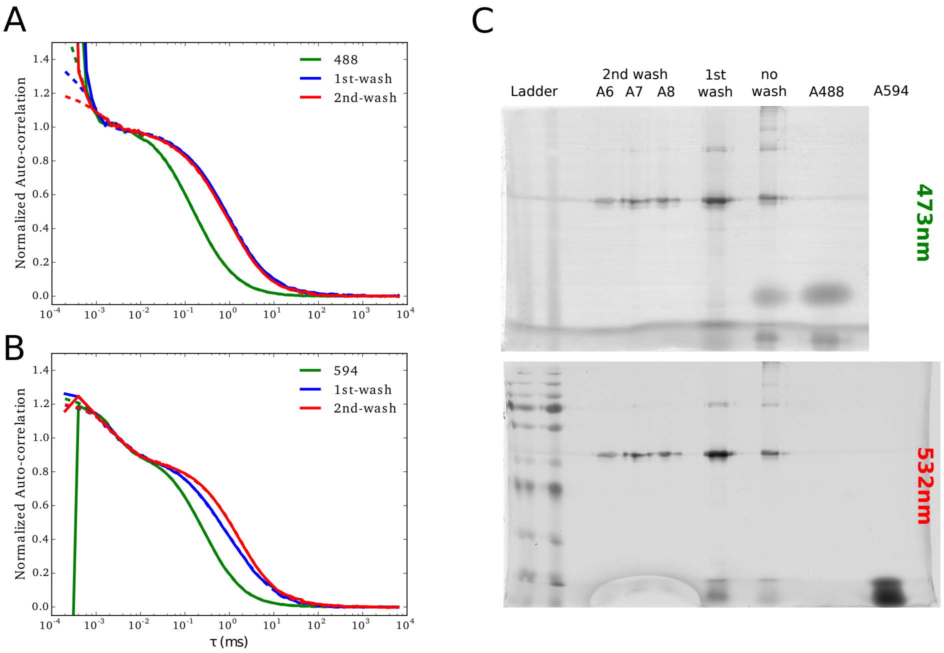


***Figure 4-12**: Presence of non-covalently bound dyes.*
Panel A and B show the normalized average ACF of dual labelled φ8 P4. Free AF488 and AF594 were used for calibration of the hydrodynamic radius, R_h (green lines). ACFs were acquired upon 488 nm (A) or 594 nm (B) excitation. Curves were normalized and fitted with a double state model, as described in section 2.6.1. ACFs after the first and second purification steps are plotted in blue and red, respectively. (C) SDS gel of dual labelled mutant 304 at different steps in the purification. The gel was visualized by excitation with 473 nm (top panel) and 532 nm (bottom panel) lasers to detect either AF488 or AF594.

Species	488 nm	594 nm
AF488	167 µs	-
AF594	-	265 µs
304 1st wash	955 µs (R_h ~ 40 Å)	1035 µs (R_h ~ 27 Å)
304 2nd wash	932 µs (R_h ~ 39 Å)	1510 µs (R_h ~ 40 Å)

***Table 4-2:** Diffusion times of dual labelled mutant 304.*
Diffusion times were estimated from the 2-state model fitting of the ACF. The apparent hydrodynamic radii, R_h, were derived from estimated diffusion times as explained in Section 2.6.1.

Ensemble FRET

The ensemble emission spectrum of dual labelled mutant 304 upon excitation at 488 nm was acquired as described in Section 2.6.2 (Figure 4-13). A peak at the acceptor emission wavelength (around 617 nm) was observed, indicating that non-radiative transfer occurs. Comparing the intensities, *I*, of the donor and acceptor emission peaks, we estimated the mean FRET efficiency, *E*, as:

$$E = \frac{I_{617nm}}{I_{525nm} + I_{617nm}}$$

A relatively high *E* = 0.58 was measured in the apo state. A high value was expected since in the X-ray crystal structure of φ8 P4 in the apo state, the distance between two 304 residues within a hexamer ranges from 10 Å to 24 Å. No change was observed when φ8 P4 was mixed either with PolyA or ATP individually. When both PolyA and ATP were mixed with P4, the energy emitted by the donor increased, while emission of the acceptor decreased, resulting in a measurably lower *E* = 0.54. This ensemble study demonstrates that residue 304 does not undergo noticeable conformational changes when φ8 P4 binds to RNA or when it is in the presence of ATP alone. Hence, the conformational change that the C-terminus undergoes upon RNA binding does not come about due to the detachment of the entire C-terminal domain.

Instead, residue 304 and the upstream part of the C-terminus remain anchored to the main structure of the motor, while the distal C-terminal part may be expelled from the central channel during RNA loading (see discussion and Figure 4-17 for more explanations). However, the decrease in FRET efficiency when φ8 P4 is mixed with PolyA and ATP indicates that the average distance between different 304 residues in the hexamer increases upon translocation. As the FRET efficiency remains relatively high, two scenarios are possible: (i) the C-termini uniformly open to a small extent across the hexamer, or (ii) processivity of the motor involves a transient intermediate state where the C-terminus undergoes a substantial conformational changes, but when averaged across the ensemble leads to a small overall change in FRET efficiency.

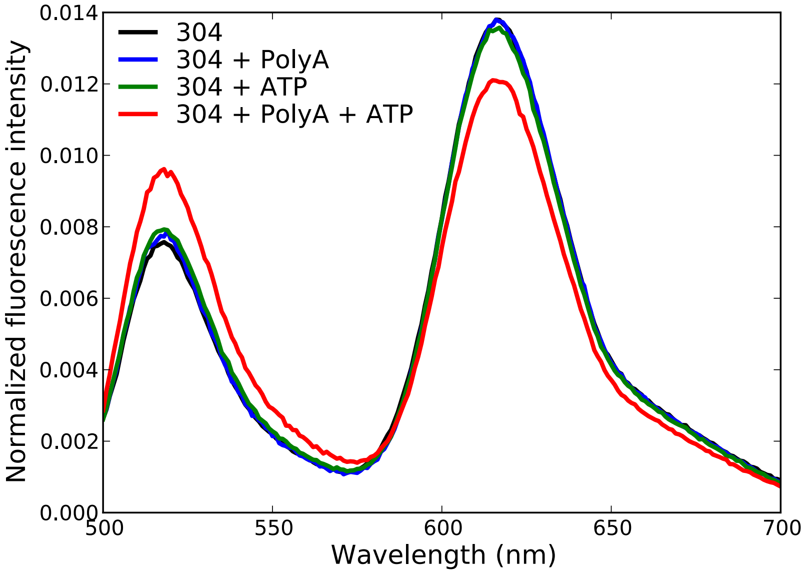


Figure 4-13: *Ensemble FRET of dual labelled mutant 304.*
Emission spectrum of the dual labelled mutant 304 upon 488 nm excitation. Ensemble FRET was measured with φ8 P4 only (black line), φ8 P4 in the presence of PolyA (blue line), φ8 P4 in the presence of ATP (green line) and φ8 in the presence of both PolyA and ATP (red line). The measurements were performed at φ8 P4 concentration ~100 nM, with 0.1 mg/ml PolyA and 1 mM ATP. All curves were normalized by their integral. The high acceptor emission (peak at 617 nm) indicates high FRET. The decrease of FRET in the presence of PolyA and ATP indicates a conformational rearrangement of the C-termini upon translocation.

Single molecule FRET:

To investigate further the relative distance between two C-termini upon translocation, we used alternative laser excitation (ALEX) to avoid ensemble averaging. A confocal microscope setup excited freely diffusing proteins passing through the confocal volume in an alternating pattern at 488 nm and 594 nm. The dual labelled 304 mutant sample was diluted down to 100 pM to observe individual proteins one by one. Single molecule FRET data were accumulated for ~1 h and processed to establish the stoichiometry (*S*) and the FRET efficiency (*E*) of each event, as described in Section 2.6.3. A two-dimensional histogram for *E* and *S* of dual labelled φ8 P4 in the apo state is shown in Figure 4-14 A. The raw averaged FRET efficiency, *E* = 0.61, was consistent with ensemble FRET data. Knowing the stoichiometry of each burst, one can identify hexamers that had only a donor (high S) or only an acceptor (low S). The FRET efficiency distribution was rebuilt keeping only bursts for which the stoichiometry was between 0.3 and 0.7 (Figure 4-14 B). The averaged FRET efficiency based on the corrected histogram is equal to 0.65. This value is more accurate since it removes from the count many of the singly-labelled hexamers; for this reason the ensemble FRET measurements will tend to underestimate the average FRET efficiency. The relatively wide distribution cannot only be due to the intrinsic conformational flexibility of the fluorescent probes. Instead, the bimodal distribution (peak at 0.7 and shoulder at 0.5) indicates the presence of multiple species (2 or more) with different conformations. A FRET efficiency of ~0.7 is in line with the expectations based on the structure of the hexamer in the apo state. The shoulder (*E*~0.5) is a less populated sate, probably corresponding to hexamers with a C-terminus already deployed. When ALEX was attempted with φ8 P4 in the presence of PolyA and ATP, PolyA caused scattering issues and the hexamer appeared to dissociate upon extreme dilution (~100 pM). Although no valuable ALEX data were collected for φ8 P4 in the presence of RNA and ATP during this work, it is worthy noting that, based on previous ensemble FRET results, one expects a shift of the FRET efficiency distribution towards lower values.

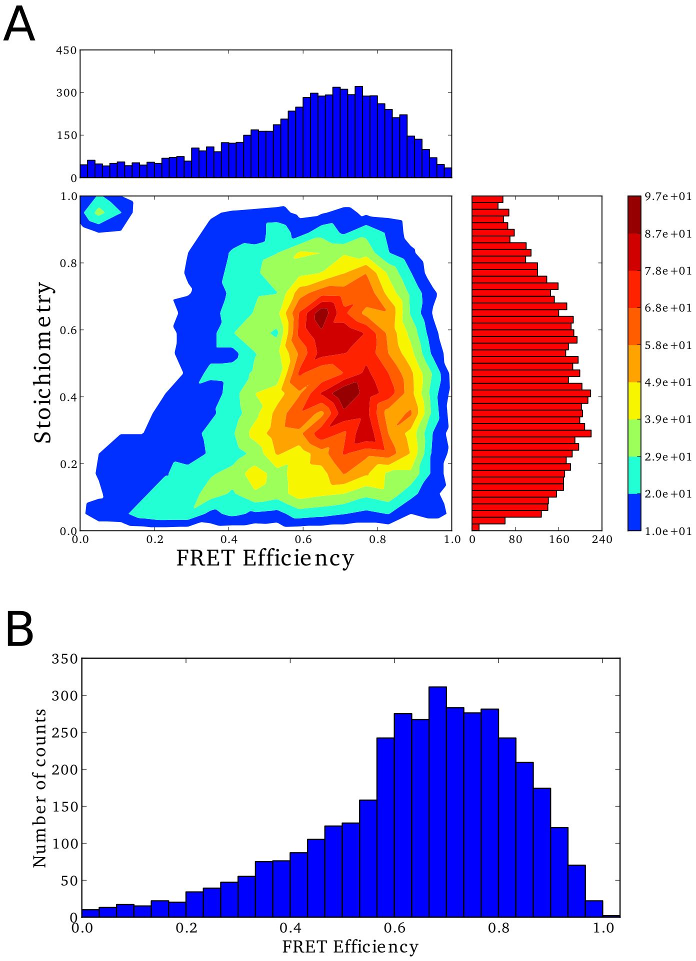


Figure 4-14: *Single-molecule FRET of dual labelled mutant 304.*
(A) Two-dimensional E-S histogram of dual labelled mutant 304. The measurement was performed at ~100 pM. (B) FRET efficiency distribution of events for which 0.3<S<0.7. One notices a peak at 0.7 and a shoulder at 0.5

4.5 Concluding discussion

By combining MD simulations, HDX-MS and fluorescence spectroscopy, we have shed some light on the conformational changes that occur in the C-terminus of φ8 P4 when it binds to RNA.

Stability of the φ8 P4 ring

Previously published HDX-MS data for φ8 P4 were re-analysed with the quantitative method introduced in Chapter 3. Unlike φ12 P4, the analysis showed that the ring of φ8 P4 remains closed in solution and opens only upon RNA loading. Overestimation of the protection factor of fragment 20 in the "open state" in our modelling suggests that dissociation of two consecutive subunits is accompanied by substantial modifications of the secondary structure of the fragments localised at the interface. Remarkably, all contacts between consecutive subunits are localised at the apical part of the motor. The strong subunit interactions at the top of the motor prevent φ8 P4 from dissociating and may explain why, unlike other P4 motors, φ8 P4 remains fully functional when not embedded into the capsid (131). Previous cryo-EM studies revealed that one of the six interfaces of φ8 P4 opens partially at the base when the hexamer is embedded into the virus capsid (139) (see Figure 4-15). The mismatch between the 6-fold symmetry of the hexamer and the 5-fold symmetry of the capsid may explain this partial opening of the ring. The base of the subunit interface encompasses the hydrophilic loop L2. By analogy with φ12 P4, L2 is thought to be essential for RNA binding. It suggests that the capsid controls RNA loading via an adjustment of the exposure of the L2 loop (15, 139). The strongly bound apical part of the motor may play the role of a collar that keeps the ring closed when the capsid opens the base of the interface.

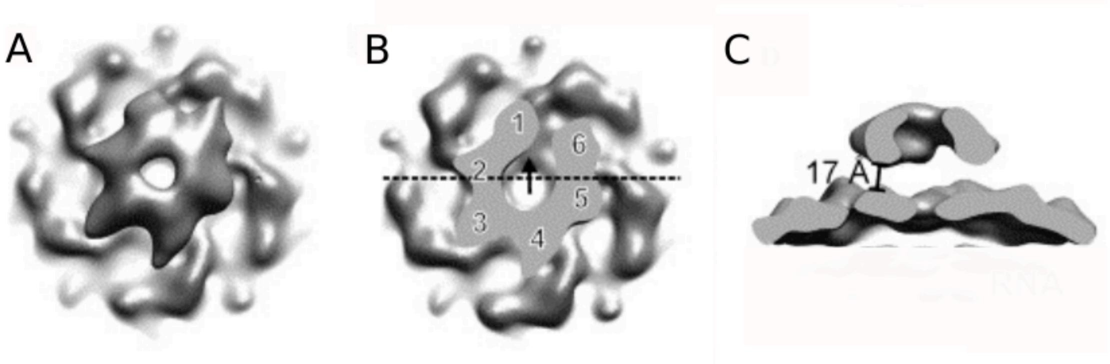


Figure 4-15 : *Capsid-associated structure of P4.*
Asymmetric reconstruction of φ8 P4 and the procapsid based on cryo-EM data (A) Top view of the motor. (B) The hexamer density of the apical dome of the motor was removed to reveal the opening of the base of one interface (black arrow). (C) Side view of the hexamer along the plan indicated by a dotted line in (B). The gap between the shell and the base of the hexamer may be filled by the N-terminus. Modified from (11).

Function of the N-terminus

In contrast to φ6, φ12 and φ13 P4, the C-terminal domain of φ8 P4 covers the apical part of the hexamer, whereas its N-terminal domain is closer to the base. It has been previously shown that φ6 P4 interacts with the P1 shell through its flexible C-termini and that its sides are in direct contact with the P8 outer layer (5). Since φ8 lacks P8, its motors are only embedded in the P1 icosahedral structure (165). The topological inversion of the C- and N-terminal domains in φ8 P4 suggests that the motor interacts with P1 through its N-terminus. The distal part of the N-terminus was not resolved in the X-ray crystal structure of φ8 P4, suggesting the N-terminus is relatively flexible. Modelling the missing sequence with an unstructured segment resulted in overestimation of the exchange kinetics. This finding suggests that the N-terminus adopts a more structured conformation in solution. According to a cryo-EM reconstruction, the closest distance between the motor and the virus capsid is 17 Å (11). Mobile and partially structured N-termini could be deployed by φ8 P4 as tentacles to span the distance and anchor it to the virus capsid.

Conformation of the C-terminus in the apo state

In the X-ray crystal structure of φ8 P4, the C-terminal domain obstructs the entrance of the channel and the interface through which RNA is thought to be loaded. This structure was obtained with a C-terminally His-tagged construct. It is possible that the presence of the C-terminal tail inside the pore was only an artefact caused by the six extra histidines added at the C-terminus and/or crystallisation. In the lab, we handled both C- and N-terminally His-tagged φ8 P4. During purification, we noticed that the former elutes at ~100 mM imidazole, whereas the latter comes off the nickel column at ~300 mM imidazole. The lower affinity of the C-terminally His-tagged hexamer for Ni^{2+} resin suggests that, in solution, at least some of the C-termini are buried into the central channel. Moreover, HDX kinetics of the C-terminus in the apo state exhibits partial protection of the C-termini, a feature that is incompatible with a scenario in which C-termini are all outside the pore. HDX data were performed with a non-His-tagged protein. It seems clear that the presence of the C-terminus inside the pore in the apo state was not an artefact caused by the His-tag or crystallisation.

Conformational changes of the C-terminus upon RNA binding

The fast HDX kinetics of the C-terminal tail suggests that at least some C-termini are expelled from the central channel upon RNA binding. On the other hand, the non-perturbed average FRET efficiency upon RNA binding indicates that the 304 residues remain very close to each other, as observed in the apo structure. Hence, expulsion of the distal part of the C-terminus from the central channel is not accompanied by a larger opening of the C-terminal domain. It is still not clear whether all the C-termini come out upon RNA binding or only a few of them. Indeed, one could imagine a scenario where only the C-terminus localized at the interface though which RNA is loaded, and maybe the two adjacent C-termini, are expelled from the central channel upon RNA loading (Figure 4-17). As shown in Figure 4-4, both scenarios lead to equally fast HDX kinetics at the time-scales accessible by manual mixing.

Interestingly, the ensemble-averaged FRET efficiency decreases when φ8 P4 is mixed with RNA and ATP, *i.e*. during translocation. Since single-molecule fluorescence spectroscopy data during translocation are not available, any structural interpretation of the FRET efficiency decrease becomes very speculative. Since the decrease is relatively low, it either indicates a transient large conformational change of the C-terminus upon RNA loading, or a limited reorganisation of the apical part of the motor during translocation.

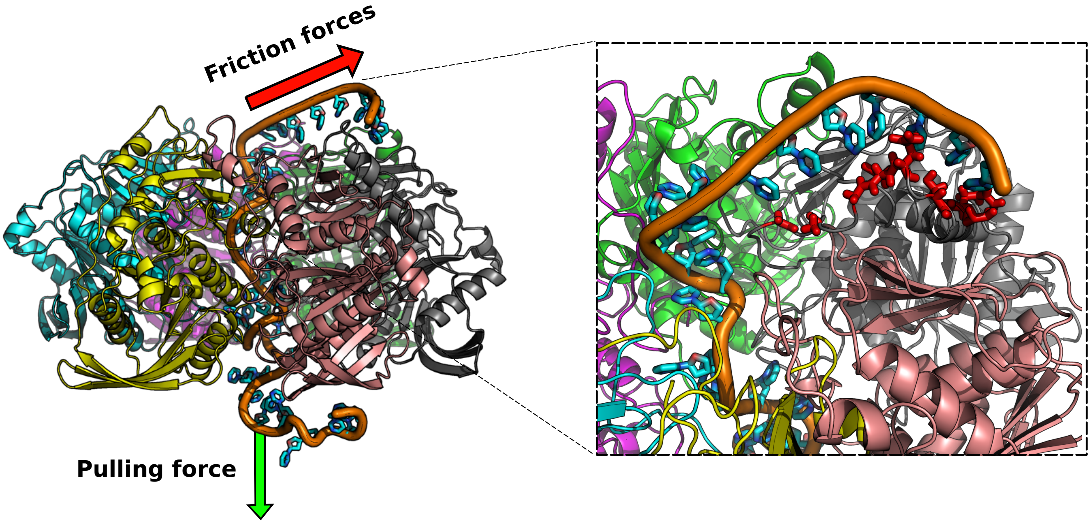


Figure 4-16: *Interaction of RNA with the top of the hexamer.*
Snapshot of a conformation adopted by the RNA strand during simulation of φ8 P4 with RNA and C-termini inside the channel. Whereas RNA is confined in an area around the axis of the channel when C-termini are spread outside the channel, RNA tends to stick to the top of the helicase (zoom in) when C-termini are kept inside. Interaction of RNA with the surface of the protein may lead to friction forces (red arrow), opposing the pulling force generated by the motor (green arrow). Residues interacting with RNA (i.e. within 4.5 Å of RNA) are highlighted in red sticks (residues 294-307).

Insights from MD simulations

MD simulations of RNA-loaded φ8 P4 with the C-termini inside the pore revealed that the poly-nucleic-acid chain interacts strongly with C-terminal domain located on the apical dome of the motor (residues 297-307, see Figure 4-16). Knowing that the C-terminus (i) restricts the interface through which RNA is loaded, (ii) is repelled from the central channel upon RNA binding, and (iii) interacts strongly with RNA, it is postulated that RNA binds to the hexamer via the exposed part of the C-terminus, leading to an expulsion of several C-terminal tails from the central channel.

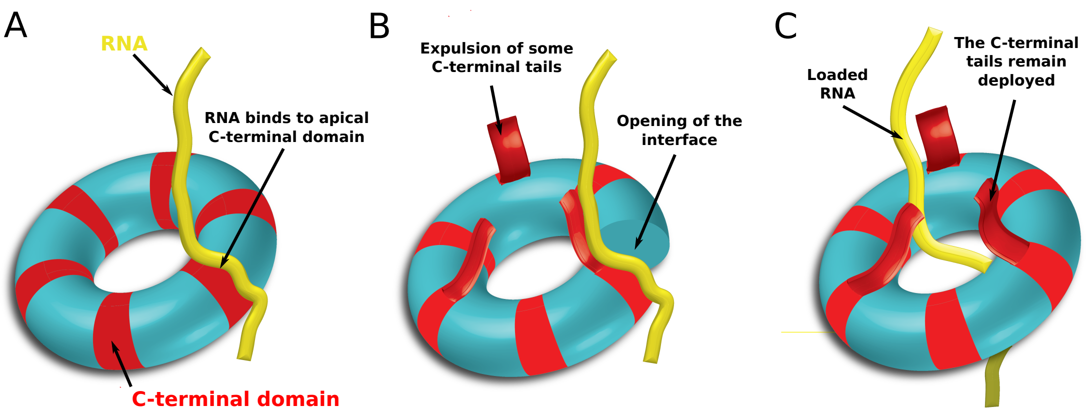

A
RNA
RNA binds to apical C-terminal domain
C-terminal domain
B
Expulsion of some C-terminal tails
Opening of the interface
C
Loaded RNA
The C-terminal tails remain deployed

Figure 4-17 : *RNA loading mechanism of φ8 P4.*
(A) An RNA strand binds to the apical C-terminal domain (residues 294-307) that occludes an interface. (B) This leads to the expulsion of some C-terminal tails and to the transient opening of the interface. (C) The RNA strand is loaded onto the central channel and the C-terminal tails remain deployed to avoid re-binding of RNA to the apical C-terminal domain during translocation.

Expulsion of the C-terminal tails outside the pore raises two questions: (i) do the C-termini disturb translocation if kept inside the pore? and (ii) what is the utility of the C-termini?

Possibly, the C-terminal tails are expelled from the central channel simply because there is not enough space to accommodate both the RNA strand and all the C-termini into the pore. The structure of φ8 P4 with both the RNA and the C-termini into the pore supports this idea. In the model constructed with MODELLER, the central channel could not accommodate six extended C-termini, the RNA molecule and the loops (Figure 4-2B). In order to squeeze everything inside the pore, MODELLER extended only three of the C-termini along the channel, and the other three were bent to fit at the entrance of the pore, with their tips pointing towards the top of the channel (Figure 4-18B). As a consequence, for bent C-termini, although the tips (F305-G307) were protected, the segment S313-G318 was exposed to the solvent instead of being buried deep inside the channel (Figure 4-18A). This explains why the calculated HDX kinetics of fragment 28 was counter-intuitively faster with the RNA and the C-termini inside the channel compared to the apo state (Figure 4-4). The evident lack of space inside the central channel leads us to believe that RNA loading requires prior expulsion of at least some of the C-termini.

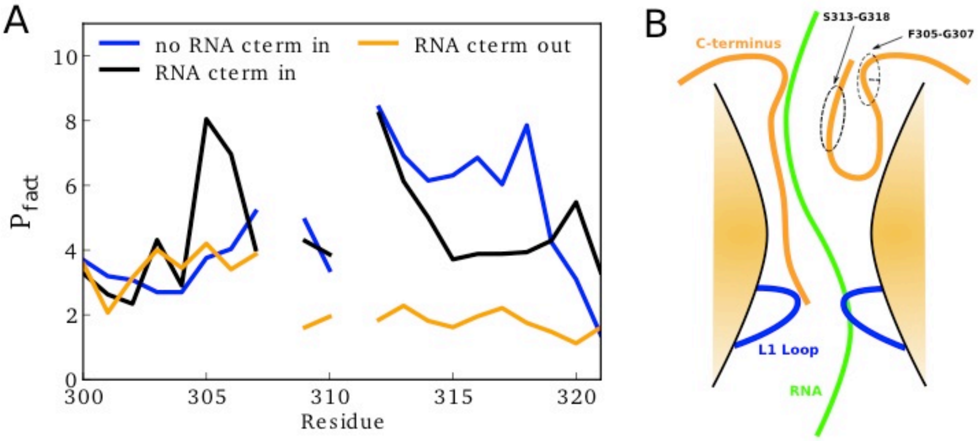


Figure 4-18: *Impact of RNA loading on protection factors of the C-terminus.*
(A) Calculated protection factors (averages over MD simulation) of residues 300 to 321 using the apo state (blue line) and the models with RNA loaded and the C-terminus either inside (black line) or outside (orange line) the pore. Gaps represent residues without amide hydrogen (prolines 308 and 311). (B) Schematic representation of the two different conformations of the C-termini inside the channel with RNA. For bent C-termini, the segment S313-G318 is overexposed, whereas segment F305-G307 is overprotected due to interaction with segment V319-K321.

Potential interference between the φ8 P4 motor and the C-termini were examined by comparing the interactions of RNA with key pieces of the motor (*i.e*. the L1 and L2 loops) when all C-termini are either inside or outside the pore. It has been shown by mutagenesis that loop L2 is essential for the functionality of the φ12 P4 motor (24, 131, 137). No RNA binding event to the equivalent L2 loop is observed in the MD simulations of φ8 P4 whether C-termini are inside or outside the pore (Table 4-3). The L2 loop is too far from the central channel to interact with RNA suggesting that either L2 has a different role in φ8 P4 or that the base of the motor undergoes large conformational changes upon RNA loading that bring L2 closer to the centre of the channel. The LKK motif that is found within the L1 loop of φ8 P4 is known to be essential for the activity of the motor (24, 78). When the C-termini are outside the central channel, the LKK motif binds to the RNA backbone in 72% of the frames on average. However, if the C-termini are kept inside, this proportion decreases to 41%. Remarkably, the aspartic acid (D320) and the lysine (K321) present at the tip of the C-terminus exhibit strong affinities to L1 and RNA, respectively (Table 4-3). Hence, the “DK motif" of the C-terminus seems to act as a competing ligand that limits interaction between the RNA and the LKK motif. It becomes clear that the C-termini strongly interfere with the core of the motor φ8 P4 when they are kept inside the pore. Given that the C-terminus is (i) crucial for helicase functionality (17), (ii) exactly long enough to reach the L1 loop at the bottom of the channel, and (iii) interacts with the RNA binding site, it suggests that the C-terminus is an important "cog in the engine" rather than a "gate" that would have to be taken away before translocation. Hence, although some C-termini have to be expelled from the central channel to make space for the RNA, a few might be required inside to ensure the good functionality of the motor.

Residues	C-termini in	C-termini out
L2 loop – RNA	0%	0%
L1 loop – RNA	41%	72%
L1 loop – D320	72%	-
K321 – RNA	26%	-

***Table 4-3**: Differences in interactions with C-termini inside or outside the channel.*
The table provides the percentage of frames two regions were found in contact. A region was considered to be in contact with another if the distance between them was smaller than 4.5 Å. Only phosphorus atoms in the RNA backbone were considered. L1 and L2 loops correspond to regions L184-K186 and D220-A225, respectively. To estimate the interaction between D320/L1 or K321/RNA, only the three C-termini that pointed towards the centre of the channel were considered (see Figure 4-18).

There may be a second benefit of keeping some C-termini outside the pore during translocation. As mentioned above, RNA tends to stick to the apical dome of the hexamer. Although it may be essential for RNA loading, this propensity to bind to the hexamer may also create friction forces opposite to the pulling force generated by the motor (Figure 4-16). Such a resistance may slow down the translocation or even lead the motor to stall. Interestingly, in the MD simulations where all C-termini are outside, RNA remains trapped between the six deployed C-termini such that it never interacts with the top of the hexamer. Hence, some of the C-termini may come out the central channel to prevent RNA from re-binding to the apical dome of φ8 P4 during translocation.

The revised model

A new model is proposed to explain the role of the C-termini. Upon RNA loading, a number of C-termini are expelled from the central channel to leave space for RNA. The deployed C-termini prevent RNA from re-binding to the dome of the motor and therefore limit friction during translocation. In the model, the rest of the C-termini remain inside the pore. Due to the pulling forces caused by translocation, the tips of the C-termini remain at the bottom of the pore, near the location occupied by the lever in the “down” position. Unlike φ12 P4 in which the lever makes use of the L2 loop, the lever of φ8 P4 utilises the L1 loop. The LKK motif of the L1 loop binds more strongly to RNA than the single arginine of the L2 loop of φ12 P4. Hence, the high affinity of the lever to RNA is modulated by the DK motif. When the L1 loop switches to the “down” position, the proximal DK motif competes with the LKK motif and RNA is detached from the lever prior its return to the “up” position. Hence, RNA is not pulled back to its initial position. This model remains very speculative and several experiments to validate this mechanism are proposed in Chapter 6.

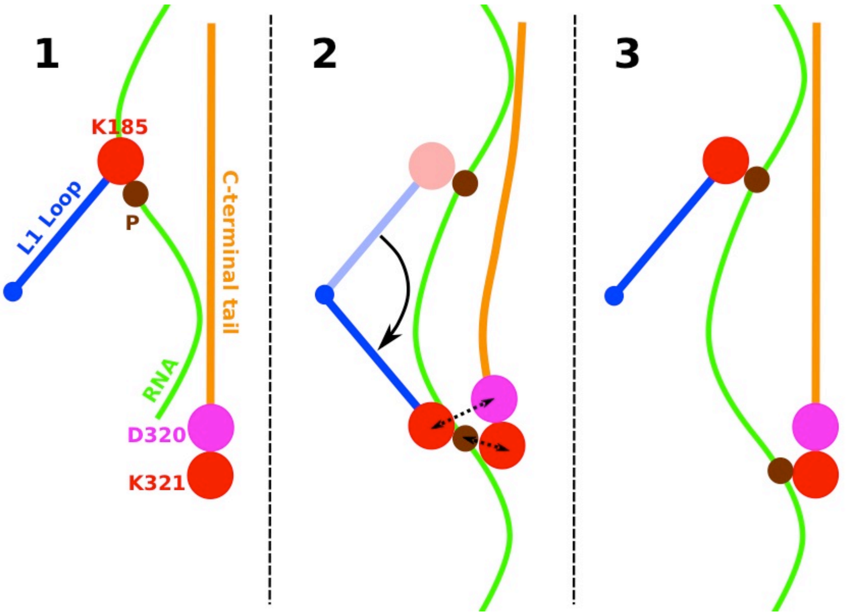


Figure 4-19: *Revised mechanisms of φ8 P4.*
Schematic representation of sequential binding of RNA during translocation. In the central channel is represented the L1 loop (blue), the RNA chain (green) and the C-terminal tail (orange). (1) The lysine of the L1 loop (K185) binds to the RNA phosphate. (2) Upon ATP hydrolysis, L1 is switched to the “down” position. The lysine carried by the C-terminus (K321) competes with the lysine of L1 to bind to RNA, whereas the aspartic acid (D320) competes with the RNA phosphate to bind to L1. (3) The RNA detaches from the L1 loop and the lever switches back to the “up” position, leaving the RNA a few Ångtroms down.

Chapter 5: High-resolution models of protein states from sparse experimental data

5.1 Introduction

The functions of proteins depend on both their structure and dynamics. Due to their intrinsic dynamics, proteins adopt different conformations. To account for the conformational heterogeneity of proteins, they are sometimes represented by an ensemble of conformations rather than one single structure (166). The number of conformers adopted by a protein in solution is potentially infinite, however, since most conformers will not be significantly populated, it is more relevant and also more useful to represent the protein by a limited ensemble of structures. An example where conformational heterogeneity is of particular importance is for intrinsically disordered proteins (IDPs). IDPs possess relatively flat energy landscapes and can thus, easily transition between diverse conformations (167). Making account for the inherent flexibility of proteins is also vital when interpreting experimental data. Indeed, interpretation of experimental data based on rigid models can be biased due to conformational averaging and the coexistence of different statistically significant conformations (168).

Despite remarkable technical advances, experimental studies cannot directly probe protein dynamics at atomic resolution over the entire range of functionally relevant timescales. On the other hand simulations still face the dilemma of increased accuracy at the cost of computational efficiency. A promising approach consists of combining experiments and simulations to generate an ensemble of structures, for which averaged computed observables agree with the available experimental data. However, the ensemble of structures may not be representative or meaningful if the amount of information provided by the experimental data is small compared to the degrees of freedom of the protein (169–173). Indeed, refining an ensemble of *M* structures requires the simultaneous estimation of ~3x*N*x*M* parameters, where *N* is the number of atoms in the protein. If the information provided by the experiments is less than this, the ensemble of structures may not represent the state probed by the experiment. Maximization of the information provided by experimental data is crucial to improve the data-to-parameter ratio. The information content of the experimental restraint is often improved by combining multiple experimental observables (174–176). This raises the question of the nature of the structural information carried by the physical properties and their complementarity. Local structural information such as hydrogen exchange (HDX) or NMR chemical-shifts (CS) and larger scale information such as that provided by small-angle X-ray scattering (SAXS) or from ion-mobility (IM) cross-sections are examples of techniques that provide complementary information.

5.2 Overview of the Chapter

In this Chapter, I present a method to assess how significantly an observable improves an ensemble. The problem is addressed via a purely conceptual approach that involves only computed data. The assessment is based on the ability of an observable to guide the construction of a conformer ensemble towards the Boltzmann ensemble of the protein (*i.e*. the observed ensemble). A genetic algorithm is used to generate an ensemble of structures that reproduces the given observable (*e.g*. protection factors). The closer the generated ensemble is to the Boltzmann ensemble, the more informative the observable. The small protein FIP35 is used as a test case. Since this thesis is mainly based on the interpretation of HDX-MS data, this chapter focuses more on this technique. The information content of HDX-MS data are briefly compared with that of HDX probed by NMR (*i.e*. protection factors), as well as with the structural dynamic information provided by other popular experimental techniques such as CS, single-molecule Förster resonance energy transfer (smFRET), SAXS or IM. The loss of information due to fragment averaging of HDX-MS kinetics is evaluated as a function of the length of the fragments. The utility of hydrogen exchange at short timescales, which has been highlighted several times in this thesis, is also investigated. Finally, the complementarity between HDX-MS and smFRET, two techniques combined in Chapter 4, is discussed.

5.3 Method

5.3.1 Overview of the method

A protocol is presented to compare the structural dynamic information content of different observables. The information content of a given observable (*e.g*. protection factors) is assessed on its ability to “drive” the reconstruction of a structure ensemble towards the Boltzmann ensemble of the protein FIF35. A molecular dynamics simulation of the protein FIP35 is used to generate an ensemble of structures, called the “reference ensemble” (section 5.3.3). In this work, the reference ensemble models the Boltzmann distribution of the protein. The observable is back-calculated from the reference ensemble and is referred to as the “synthetic experimental data” (section 2.3.2). A second ensemble made by selecting random structures of FIP35 is generated (section 5.3.4). This second ensemble, called the conformer pool, does not exhibit the Boltzmann distribution of the protein. The pool is refined in order to obtain a new ensemble for which the back-calculated observable matches the “synthetic experimental data”, *i.e*. the mean square deviation (MDS) between the two observable is close to zero. The refinement is performed with a genetic algorithm (section 5.3.2). The new ensemble is referred as the “refined ensemble”. Ideally, the refined ensemble should exhibit the Boltzmann distribution of the protein. Once the refinement procedure is completed, the reference and refined ensembles are directly compared. Usually, a cross-validation analysis is performed by comparing new observables which have not been used to guide the refinement (177). The new observables are back-calculated from the reference and refined ensemble and then compared. If the refined ensemble reproduces the new observables of the reference ensemble, the two ensembles are considered identical. In this work, I perform a more systematic cross-validation in which the free-energy profiles (FEP) of the ensembles are compared (more explanation in section 5.3.5). Greater similarity between the reference and the refined ensembles indicates improved information content from the observable. The overall method is summarised in Figure 5-1.

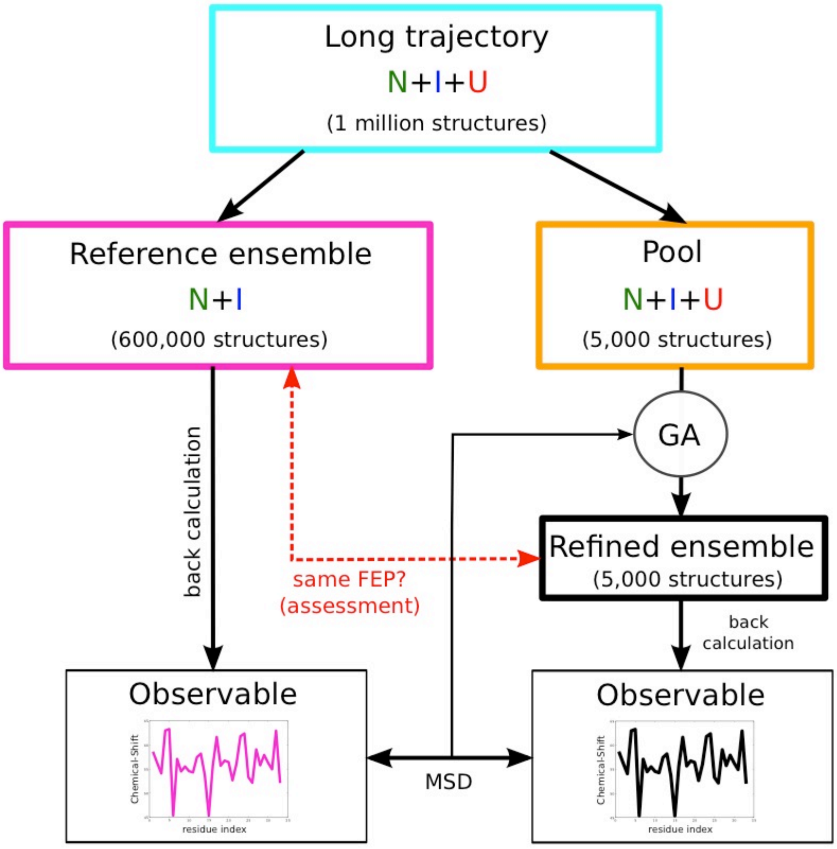


***Figure 5-1**: Method to assess the structural dynamic information of an observable.*
A large MD simulation of FIP35 is used to generate the reference ensemble and the pool. The reference ensemble models the Boltzmann ensemble of the protein and is used to calculate the synthetic experimental data of FIP35. The reference ensemble is only made of two distinct states, the native state (N) and the intermediate state (I); while the pool contains a third state, the unfolded state (U). These three states are introduced in section 5.3.3. Using a genetic algorithm (GA), the pool is refined to minimize the mean square deviation (MSD) from the synthetic experimental data of the observable back-calculated from the refined ensemble. Once the refinement procedure is finished, a cross-validation analysis is performed, in which the free-energy profiles (FEP) of the reference and refined ensembles are compared. The better the refinement, the better the information content of the observable.

5.3.2 Ensemble refinement

A number of methods have been proposed in the past to generate ensembles of structures for which computed observables match experimental data. These approaches either rely on the introduction of a restraining term in the Hamiltonian or on the selection of a subset of structures from a large sample.

Examples of the former approach are restrained-ensemble molecular dynamics simulations (83, 178, 179). In this method, multiple simulations are run in parallel with an additional energy term that acts to confine them within a conformational space that fulfils a given experimental constraint. The biasing energy term is usually a simple harmonic potential:

$$E = \frac{\alpha}{2}(\bar{q} - Q)^2$$

where α is the force constant, Q is the experimental observable to target and $\bar{q}$ is the calculated observable averaged over N different simulations. The force constant is gradually increased to improve agreement with experiment.

A second approach is based on the maximum entropy principle (65, 180–182). A given ensemble of structures (usually generated by molecular dynamics simulations) is refined to minimize the difference between an estimated and experimental property and simultaneously minimize the perturbation of the Boltzmann distribution. During the refinement, the same structures are used and only the weight associated with each structure is modified. Deviations from experimental observables can be quantified with a quadratic error function $U = (\bar{q} - Q)^2$, while the perturbation is defined as the Kullback-Leibler divergence (118) of the refined distribution, p, from the initial one, p_0, $S = -\int p(x) log\,(p(x)/p_0(x))\,dx$. By analogy with the free-energy defined in thermodynamics, U corresponds to the potential energy and S to the entropy of the system. Hence, the "least invasive" solution, *i.e.* the solution which requires fewest modifications to the pre-generated ensemble, corresponds to the distribution that minimizes the free-energy $E = U - \theta S$, where θ is a temperature-like parameter, which controls the distribution-modification-tolerance. It has been shown that restrained MD simulations and the maximum entropy technique become statistically equivalent when $\alpha \to \infty$ and $N \to \infty$ with $\alpha \gg N$ (183).

A third approach, similar to the maximum entropy method, involves the selection of a representative subset of structures from a pool of possible conformers using a genetic algorithm (113, 184). All selected structures have the same weight. Their non-uniform distribution is accounted for by including different numbers of conformers with similar shapes or the same conformer several times. Unlike the maximum entropy method, the genetic algorithm approach does not try to make as few modifications to the pool as possible during the refinement. A detailed description of this approach is given in Section 2.3.1. The three main approaches mentioned above have been applied to intrinsically disordered proteins for which the ensemble representation is particularly relevant (167, 185).

In this work, ensembles compatible with the synthetic experimental data are constructed with the genetic algorithm (for more details see Section 2.3.1). A pool of 5,000 structures of the FIP35 protein (section 5.3.4) is refined to maximise agreement of the macroscopic property between reference ensemble and sub-ensemble. How well the synthetic experimental data and the observable back-calculated from the refined ensemble match is quantified with mean square deviation, as described in Section 2.3.2. The refined ensemble is eventually compared with the reference ensemble, as described in Section 2.4.4.

5.3.3 The reference ensemble of FIP35

Recent advances in hardware and simulation methodology allow the realistic folding of relatively small proteins to be studied (28, 186). Shaw *et al*. made available to the scientific community a 200 μs MD simulation of the 35-residue protein FIP35 in explicit solvent, where 15 folding-unfolding events were observed. This provided extensive sampling of the conformational space of the protein. They saved the coordinates of the protein every 0.2 ns to yield a one-million-frame trajectory. High dimensional data such as an ensemble of structures can be projected along a one-dimensional reaction coordinate to facilitate its analysis and characterisation. This provides a clear illustration of the different states accessible by the protein (see section 1.2). However, dimensionality reduction has to be rigorous in order to correctly recapture the underlying properties of the protein (28, 29). A systematic projection method consists of finding a reaction coordinate along which the dynamics of the protein remains diffusive (29). Such a reaction coordinate is called the optimal reaction coordinate. Previously published work constructed the optimal reaction coordinate of FIP35 and characterized its free-energy profile based on the 200 µs simulation provided by Shaw *et al.* (187). This work revealed that the protein folds via a stable on-pathway intermediate state, as shown in Figure 5-2. Hence, three different states are present: the unfolded state (U), the intermediate state (I) with a first hairpin formed and the native state (N) with two hairpins. The energetic state of a conformer, X, can be identified with its optimal coordinate value, $r(X)$:

- if $r(X) \leq 19$, then X is in the native state
- if $19 < r(X) \leq 30$, then X is in the intermediate state
- if $r(X) > 30$, then X is in the unfolded state

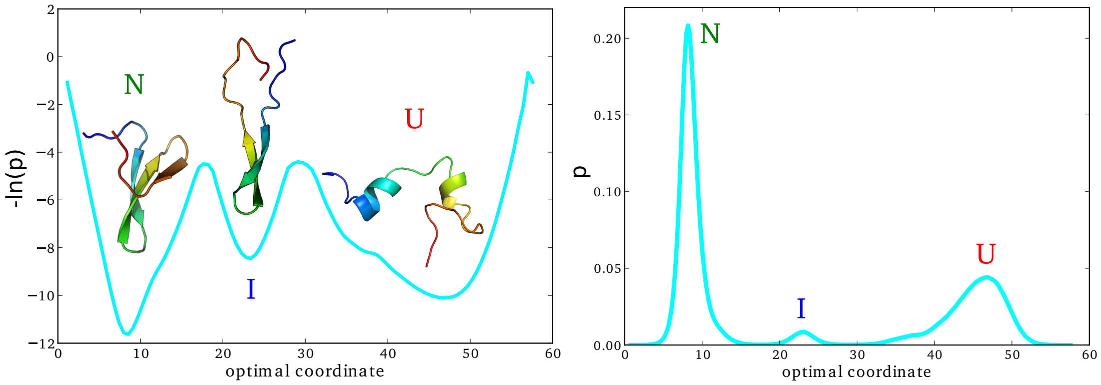


***Figure 5-2:** Free-energy profile of FIP35 along its optimal reaction coordinate.*
On the left panel, three different basins can be identified: the native basin (N), the intermediate basin (I) and the unfolded basin (U). In each basin is shown a representative structure of the state. The optimal coordinate refers to a coordinate along which dynamics is diffusive. On the right panel is shown the probability (p) of the different micro-sates projected onto the optimal coordinate. The free-energy of a micro-state is estimated as –ln(p).

Usually, observables are collected in native conditions. Hence, the reference ensemble of FIP35, used to calculate the synthetic experimental data, contains only the native and intermediate basins, *i.e.* all conformers for which their optimal reaction is lower than 30. The reference ensemble contains ~600,000 conformers, 95% of which were native structures and 5% were intermediates.

5.3.4 The conformer pool

An assumption made during ensemble refinement is that the pool contains all the statistically relevant states adopted by the protein in solution, and possibly also many of the “irrelevant structures”, *i.e*. structures that are not accessible (or very rarely accessible) by the protein. Hence, the refinement process has to (i) reject all the irrelevant structures, and (ii) correct the ratio between the different relevant states. The conformer pool will contain structures from the basins N, I and U. Throughout the rest of the chapter, the term “irrelevant structures” will refer to structures that belong to the unfolded state. On the other hand, structures from the native or intermediate basins will be referred to as “relevant structures”, since they are part of the reference ensemble. To construct the pool, 5,000 different structures were picked from the one million structures contained in the FIP35 trajectory. The pool did not contain more than 5,000 structures due to computational limits. Structures from the different states where not picked with the same probability, such that the pool ensemble had 80% of the structures belonging to the unfolded state, *i.e*. most of the structures in the pool are irrelevant. Hence, the free-energy profile (FEP) of the pool contains one more basin than the FEP of the reference ensemble (Figure 5-3). Of the 20% relevant structures, 65% are from the native state and 35% from the intermediate state, *i.e*. I/(N+I) = 0.35.

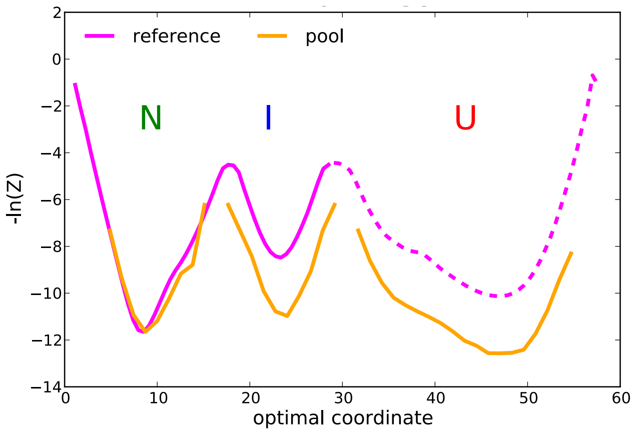


***Figure 5-3**: Free-energy profile of the reference ensemble and the conformer pool.*
The profile of the reference ensemble and the pool are depicted in magenta and orange, respectively. In the conformer pool the ratio I/(N+I) is ~35% instead of ~5% in the reference ensemble. Dotted magenta line indicates the profile when the unfolded basin is included. ~80% of the structures in the conformer pool are unfolded. An offset was added to the profile of the pool to align the bottom of the native basin with those of the reference profile.

A successful refinement of the pool should simultaneously reject all the unfolded structures (*i.e*. the irrelevant structures) and readjust the ratio of intermediate states among the relevant structures (*i.e*. I/(U+N) = 5%). In terms of free-energy profile, the unfolded basin of the refined ensemble should have its minimum energy as high as possible, whereas the difference between the energy minima of the basins N and I should match that of the reference ensemble.

5.3.5 Similarity between the reference and refined ensembles

In this work, the similarity between two ensembles of FIP35 is estimated by comparing their free-energy profiles (FEP) along the optimal coordinate. Both ensembles are projected along the optimal reaction coordinate and their distributions are compared using the Kullback-Leibler divergence (section 2.4.4). It is assumed that if the refined ensemble reproduces the FEP of the reference ensemble, then the two ensembles are equally informative.

5.4 Results

5.4.1 Illustration of the optimization procedure

The genetic algorithm (GA) was first tested with a trivial case. A 1,000-structure pool was generated. Ten structures were randomly picked from the pool to create a small reference ensemble. The synthetic experimental FRET efficiency distribution as well as the protection factors, the SAXS profile, the chemical shifts and the cross-section were averaged over the 10 structures of the reference ensemble. The GA was set to generate ensembles made of 10 structures, which were optimized over up to 50,000 generations as explained in Section 2.3.1. All observables were used to guide the optimization, *i.e*. all Lagrange multipliers were set to 1 (see Section 2.3.2). In this test case, the reference ensemble was included in the pool. Thus there existed at least one perfect solution, *i.e*. one ensemble for which the total mean square deviation (MSD_{tot}) was equal to 0. At each new generation, if a new sub-ensemble with a lower MSD_{tot} was found, this new sub-ensemble replaced the current solution. Hence, a new sub-ensemble was not necessarily produced at every single generation. The evolution of optimizing the deviation from experiment for the different observables is shown in Figure 5-4. The MSD_{tot} started at 386 and continuously decreased towards zero. Although MSDs of the different observables fluctuated, they also tended to converge towards zero. The number of structures from the reference ensemble that were contained within the refined ensemble was used to quantify the convergence of the refined ensemble towards the reference ensemble. This number tended to increase during optimisation, and this shows the algorithm converged towards the reference ensemble. After 2026 generations and 50 solutions, the GA found a perfect match and stopped. The solution proposed by the GA was identical to the reference ensemble, which validated the robustness of the method. At several points during optimisation, a perfect match with the ion-mobility cross-section was found (MSD_{ion} = 0) although the refined ensemble did not match the reference ensemble (number of correct structures < 10). This false positive illustrates the problem of using degenerate observables.

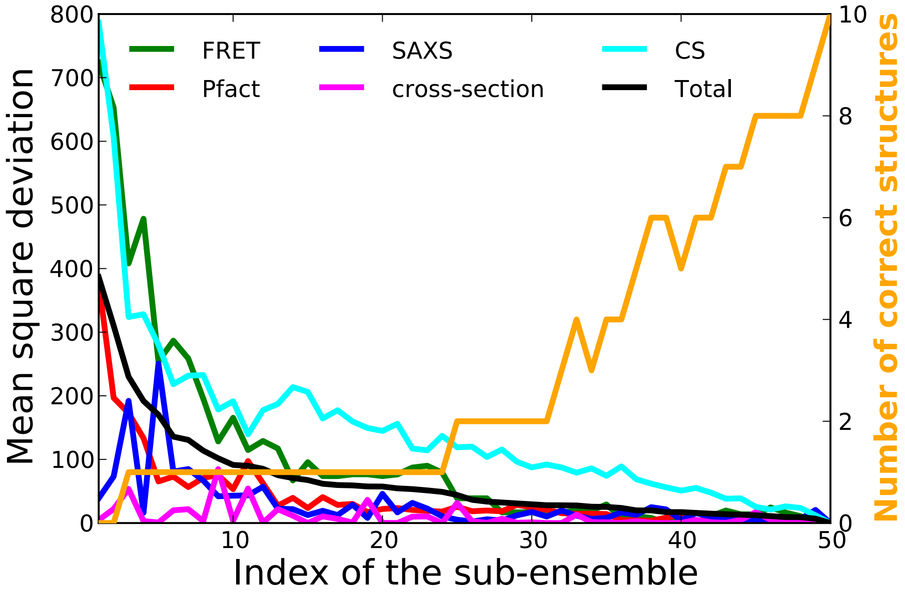


Figure 5-4*: Evolution of the deviation from synthetic experimental data during refinement. For each new optimal ensemble found by the genetic algorithm, the mean square deviations from the different observables were plotted. The total, FRET, Pfact, SAXS, ion-mobility cross-section and CS mean square deviations were depicted in green, red, blue, magenta and cyan, respectively. The number of structures of the refined ensemble that are in the reference ensemble is shown in orange. If this number is equal to 10, then the refined ensemble contains the same structures as the reference ensemble.*

5.4.2 Refinement with one observable

In this section, the 5,000-conformer pool was refined based on a single observable back-calculated from the reference ensemble (Figure 5-1). The free-energy profiles of the refined and reference ensemble were then compared.

Chemical shift

Chemical shifts (CS) of atoms along the protein backbone are highly correlated with the secondary structure of the protein (188). This complex relationship is approximated in Camshift by a polynomial function that describes the interatomic distances defining the local environment of the atoms (119). Camshift was used to compute the chemical shifts of C_α, H_α, C_β, C_O, N and H_N atoms for all residues apart from the termini. These were then averaged over the reference ensemble to generate the synthetic experimental CS. The left panel of Figure 5-5A shows the CS for C_α atoms averaged over each individual basin of FIP35. The difference between the 3 states indicates that CS are sensitive enough to distinguish each state. However, the CS of some residues show similarities between different basins. For the intermediate state, CS of residues 2-21 are identical to those of the native state, whereas CS of residues 22-34 overlap with the unfolded state. This comes as no surprise since the N- and C-terminal parts of FIP35 are structurally close to the native and unfolded state, respectively. An ensemble was generated with the GA using only the CS as a constraint. The ensemble was projected along the optimal reaction coordinate and compared with the profile of the reference ensemble (Figure 5-5A.) The refined ensemble maintained the relative population between the native and intermediate state and contained only 5% of irrelevant structures (Table 5-1). It shows that CS provides sensitive and unambiguous structural information that discriminates between the three states of the protein.

SAXS

In SAXS, the random positions and orientations of proteins results in an isotropic intensity, which is proportional to the scattering of a single particle averaged over all orientations. The SAXS profile of each structure was computed using Crysol (113). An ensemble of FIP35 structures, compatible with the synthetic SAXS data, was produced by the GA (left panel in Figure 5-5B). The resulting ensemble contained 67.8% of irrelevant structures and the ratio I/(N+I) was almost as bad as in the pool (Table 5-1). When SAXS profiles were averaged over basins N, I or U (right panel in Figure 5-5B), no significant differences were observed between each basin. Since the state of a structure cannot be clearly identified based on its SAXS profile, it comes as no surprise that the observable failed to guide the refinement procedure towards the reference ensemble.

Protection factors

Protection factors of residues for each structure were calculated based on the phenomenological approximation introduced in Section 2.3.2. In this approximation, protection factors were assumed to depend on the local environment of the residue. For example, the number of hydrogen bonds the residue possesses and its packaging density. Protection factors were averaged over the different energetic basins of the protein (right panel in Figure 5-5C). The right panel in Figure 5-5C shows that exchange clearly depends on the state of the protein. As previously noticed for CS, the N-terminal domain of the intermediate was protected to the same extent as the native state, whilst the C-terminus of the intermediate exchanged with a comparable speed to that of the unfolded state. The resulting ensemble produced by pure protection factor based refinement produced a FEP that matched as well with the reference FEP as the one obtained with CS (left panel in Figure 5-5C). The ratio of intermediate states among the relevant structures (*i.e.* I/(N+I)) was respected and only 2.3% of irrelevant conformers were retained by the GA (Table 5-1).

Cross-section

In ion mobility spectrometry, the protein is ionized and accelerated by an electric field through a buffer gas that slows down the ion motion. Measuring the drift time caused by collisions with gas molecules allows the cross-section of the protein to be estimated (115). The ensemble resulting from multiple refinements driven by the cross-section is shown in the left panel of Figure 5-5D. The free-energy profile of the refined ensemble is identical to that of the pool. Interestingly, contrary to the other observables, a perfect match (i.e. $MSD_{ion} = 0$) was systematically found. Hence, many random ensembles were compatible with the synthetic experimental cross-section, such that all conformers were equally likely to be selected. This explains why the refined ensemble is almost identical to the pool.

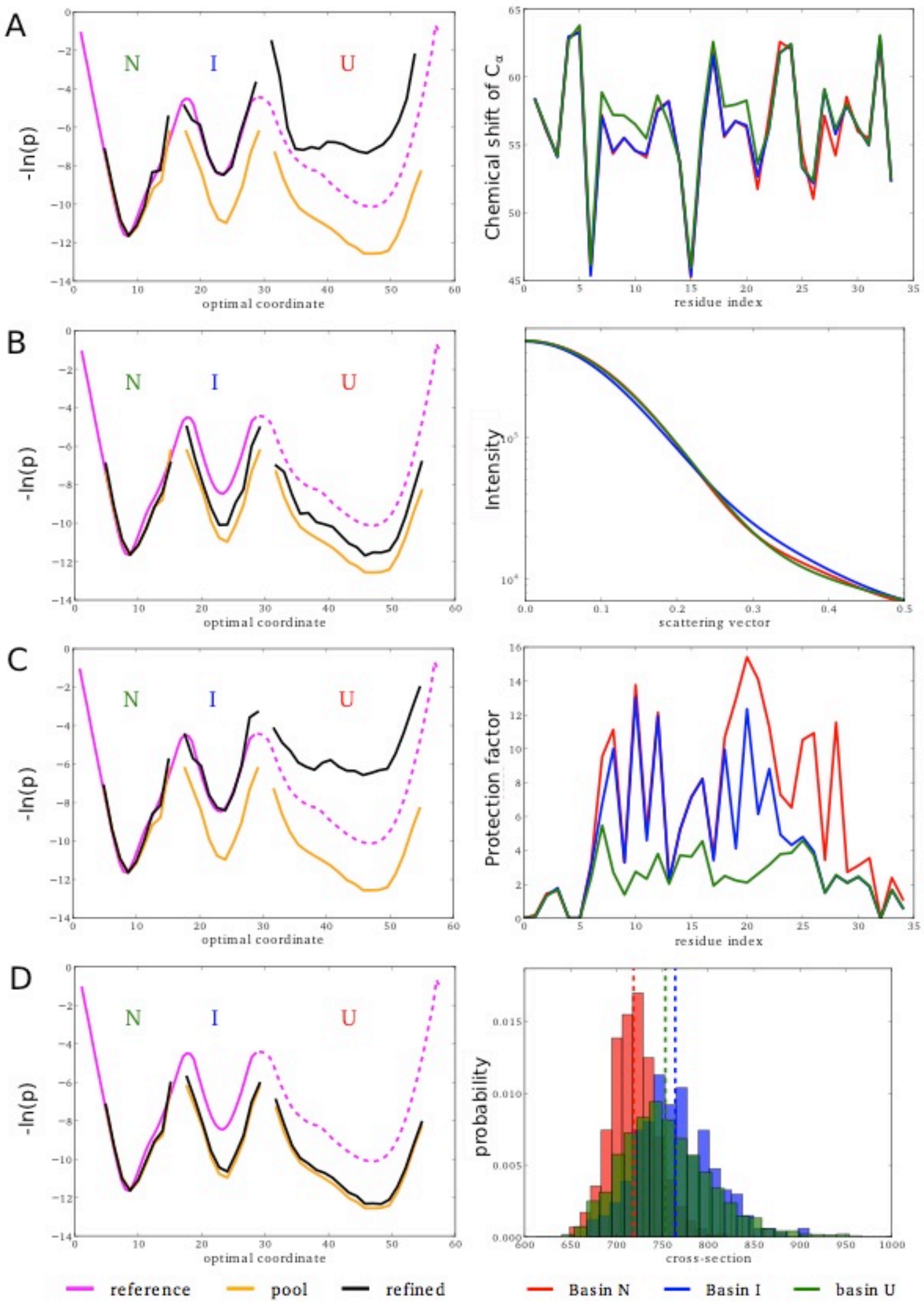


***Figure 5-5**: Free energy profiles of the ensemble refined with a single observable.*

The ensemble was refined using only the chemical shifts (A), the SAXS profile (B), the protection factors (C) or the ion-mobility cross-section (D). The left panel shows the free energy profile along the optimal coordinate of the reference ensemble (magenta), the conformer pool (orange) and the refined ensemble (black). Dotted magenta line indicates the profile when the unfolded basin is included. On the right panel is shown the observable averaged over the structures of the native basin (red), the intermediate basin (blue), or the unfolded basin (green). Only CS for C_α atoms are presented in (A). Although the distribution of the cross-section is depicted, only the averaged value (dashed vertical line) was used to refine the ensemble.

Single molecule FRET

Single molecule Förster resonance energy transfer (smFRET) probed by alternating laser excitation avoids ensemble averaging. This technique provides the FRET efficiency distribution between two labelled residues over the ensemble of structures adopted by the protein. In a way, smFRET provides the Boltzmann distribution of the ensemble projected along the end-to-end distance of the chosen pair of residues. The FRET efficiency was calculated as explained in Section 2.3.2. In practice, the Förster distance for a pair of dyes (*R_0*) is typically ~50–70 Å. However, such a *R_0* would result in FRET saturation (i.e. *E* = 1) for our small protein system even in the unfolded basin. *R_0* was assumed to be equal to 15 Å as this was a distance that was more suited to the size of the protein. This provided a broader FRET efficiency distribution for the system studied. Four different labelling positions were chosen such that their FRET efficiency distributions could differentiate either:

- none of the states (1-35)
- N+I from U (8-16)
- N from I+U (7-33)
- all the states (12-28)

FRET efficiency histograms for the different labels and in each basin are shown in Figure 5-6. As the two termini are flexible in the native state, the end-to-end distance (label 1-35) presented almost the same uniform distribution in all basins. Hence, it is unsurprising that the refined ensemble remains very similar to the pool (Figure 5-6A). For label 8-16, the FRET efficiency distribution was identical in basin N and I but very different to the unfolded basin. As a consequence, the refinement failed to correct the relative populations of N and I but it rejected the irrelevant conformers (Figure 5-6B and Table 5-1). Label 7-33 cannot distinguish the intermediate state from the unfolded one, and the GA equally decreased the ratio of intermediate and irrelevant conformers (Figure 5-6C). For label 12-28, the FRET efficiency distributions are different in all basins and unsurprisingly the refined ensemble came closer to the reference ensemble (Figure 5-6D). Though these examples demonstrate that single-molecule FRET contains useful structural information if label positions are well chosen, this information is somewhat limited. Indeed, even with label 12-28, the refined ensemble contains 45% of irrelevant conformers (Table 5-1). Note that label 12-28 and 8-16 are very complementary; while the former could reconstruct the proportion between N and I, the latter seemed much better at rejecting wrong conformers. The ensemble was further refined by using the FRET efficiency distribution of both the labels 12-28 and 8-16. The resulting ensemble perfectly captured the ratio between N and I and contained only 12% of irrelevant conformers (free-energy profile not shown, see Table 5-1). This demonstrates that two relatively poor observables can be combined to form a better restrain.

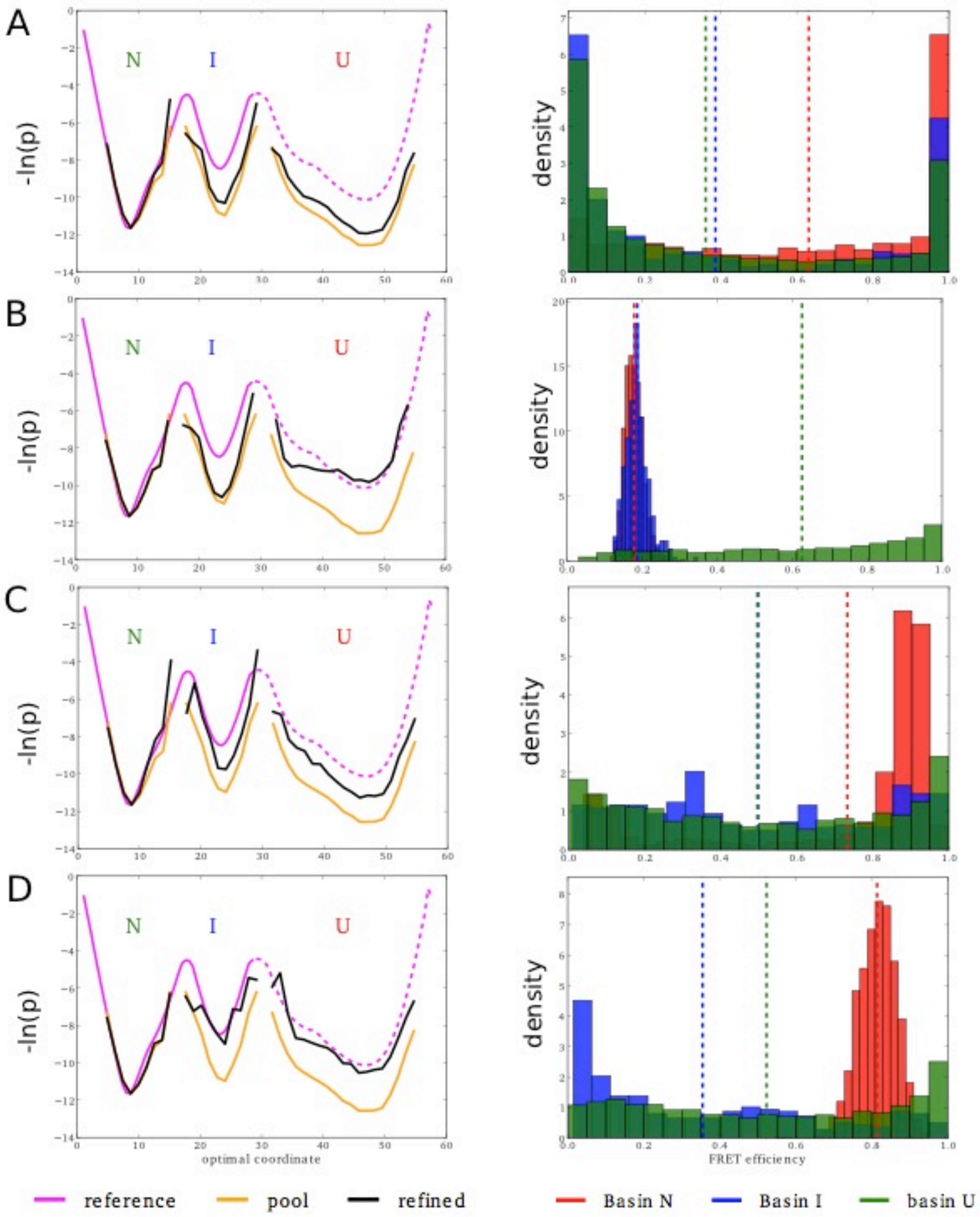


***Figure 5-6**: Free energy profiles of the refined ensemble using single-molecule FRET. The ensemble was refined using the distribution of the FRET efficiency between residues 1-35 (A), 8-16 (B), 7-33 (C) or 12-28 (D). The left panels represent the free-energy profiles along the optimal coordinate of the reference ensemble (magenta), the conformer pool (orange) and the refined ensemble (black). Dotted magenta line indicates the profile when the unfolded basin is included. The right panel shows the FRET efficiency distribution in the native basin (red), the intermediate basin (blue), and the unfolded basin (green). The averaged FRET efficiency in each basin is indicated by a vertical dotted line.*

HDX-MS

We also tested the structural information carried by the HDX kinetics of peptide fragments. It was assumed that the exchange kinetics of five peptide fragments (G1-G7, W8-R14, D15-F21, N22-S28 and Q29-G35) of equal length and covering the entire sequence of the protein is available. After having averaged the protection factors over the ensemble, the exchange kinetics of each fragment was computed as described in Section 3.3.2. The intrinsic rates were calculated at 20 ˚C and pH 8.0. Synthetic experimental deuterium fractions of the fragments were computed at the following time points: *t* = 30 s, 1 min, 2 min, 4 min, 8 min, 15 min, 30 min, 1 h, 2 h and 4 h (kinetics on the right side of the vertical black dotted line in Figure 5-7).

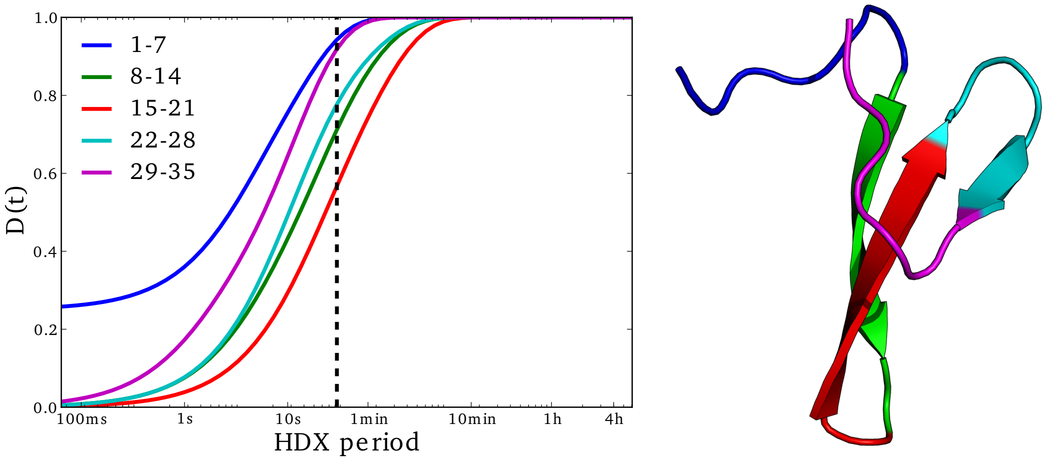


Figure 5-7*: Hydrogen-deuterium exchange kinetics of all fragments.*
The vertical dashed line designates the fastest time experimentally measureable with manual mixing. The time points on the left side of the vertical dotted line are only accessible with fast mixing. Each fragment is highlighted with its corresponding colour on the native structure of FIP35.

Incorrect assignment due to the complexity of MS-MS spectra, as well as the inevitable back and forward rate of exchange creates a systematic experimental error. This error can be significant and the quality of the information provided by the kinetics is degraded, as observed for the HDX-MS data presented in Chapter 3 and 4. Hence, the inherent HDX-MS experimental error cannot be ignored in our analysis. The inaccuracy of HDX-MS data was accounted for by introducing a cut-off within the definition of the mean square deviation. Thus, a predicted deuterium fraction *D(t)* within $\pm 10\%$ of the experimental value was considered as a perfect match (see Section 2.3.2).

The conformer pool was refined to minimize the mean square deviation from the synthetic experimental deuterium fraction *D(t)* $(30\ \mathrm{s} \leq t \leq 4\ \mathrm{h})$. Its free-energy profile was projected along the optimal coordinate and compared with that of the reference ensemble (Figure 5-8A). The refined ensemble contained ~25% irrelevant structures and the ratio I/(N+I) is equal to ~9%, instead of 5% (Table 5-1). This result is mitigated considering the quality of the previous refinement obtained with protection factors. By design, HDX kinetics provide a similar kind of information to that obtained from protection factors. However, the information is averaged over a peptide fragment instead of being residue specific. Furthermore, only time points accessible by manual mixing were considered during the refinement process. Analysis of deuterium exchange kinetics on the millisecond to second timescale (Figure 5-7) revealed that fragments with similar exchange kinetics at longer timescales (seconds to minutes) might have considerably different kinetics at shorter times.

Additional synthetic experimental data for the same fragments was generated for faster time points: *t* = 30 ms, 50 ms, 100 ms, 250 ms, 500 ms, 1 s, 2 s, 4 s, 8 s and 15 s. These time-points are readily accessible by conventional rapid quenched flow apparatus. The refinement process was carried out again after including these additional time points (Figure 5-8B). Although the relative I/(N+I) ratio was not improved upon, the ensemble contained fewer irrelevant structures (~9.6%). This clearly shows that faster time points capture structural and dynamical information not captured by slower time-scale.

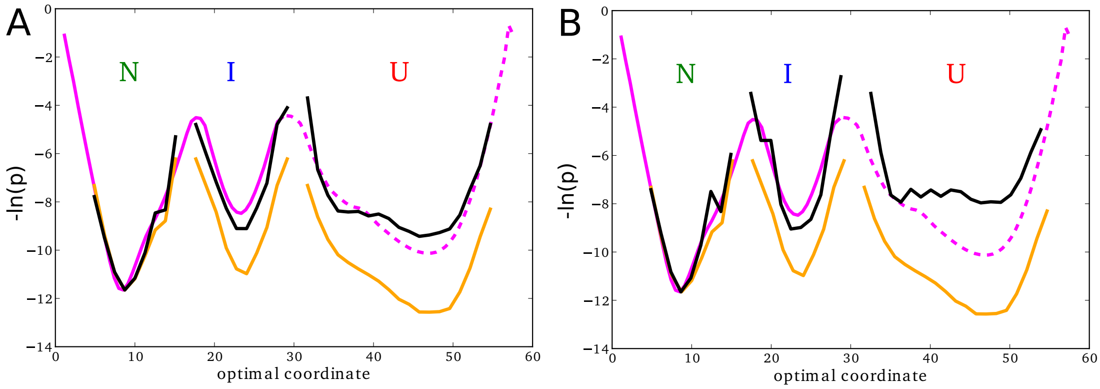


Figure 5-8: *Free-energy profile of the refined ensemble using HDX-MS.*
The refinement was guided by time points accessible by manual-mixing (A) or time points accessible by fast-mixing (B). The profile of the reference ensemble (magenta), the conformer pool (orange) and the refined ensemble (black) are projected along the optimal coordinate. Dotted magenta line indicates the profile when the unfolded basin is included.

The spatial resolution of HDX-MS data is limited by the size of the fragments produced by proteolytic cleavage of the protein. The impact of fragment averaging on the quality of the refinement was investigated. Multiple refinements were performed with different fragment lengths (Figure 5-9). As the length of fragments decreased, there was an improvement in matching between reference and refined ensembles. The refinement was performed using either only the time points accessible by manual mixing (i.e. $30\ \mathrm{s} \leq t \leq 4\ \mathrm{h}$) or the time points also accessible by fast mixing (i.e. $30\ \mathrm{ms} \leq t \leq 4\ \mathrm{h}$). Interestingly, when fast time points and short peptide fragments (1 residue per fragment) were used, the quality of refinement almost reached those obtained with the protection factors, *i.e.* the HDX probed by NMR. Hence, the two data sets carry the same structural information. The significant improvement of the refinement as the length of the fragment decreases illustrates the importance of optimizing protein digestion and peptide assignment to maximize the structural and dynamical information obtained from HDX-MS data.

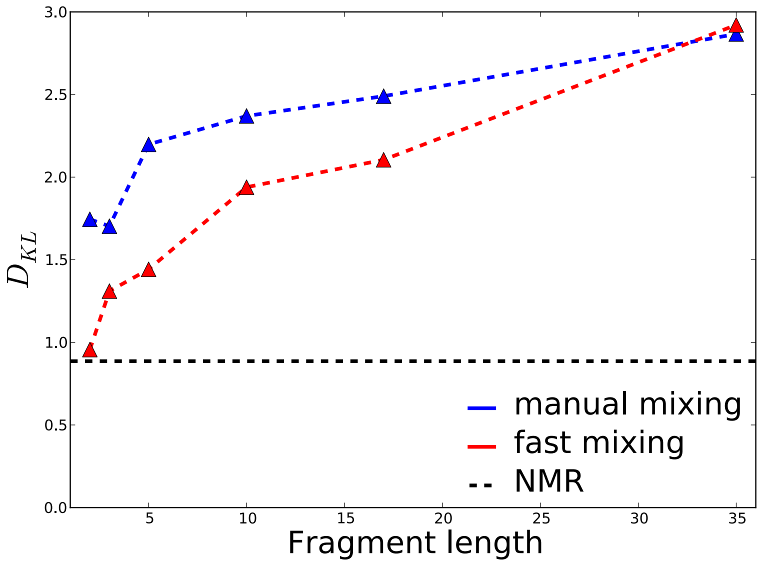


Figure 5-9: *Impact of fragment averaging on the structural information of HDX-MS data. Matching between the refined and the reference ensemble was measured as a function of fragment size. Differences between the two ensembles were quantified by comparing their free-energy profiles along the optimal reaction coordinate with the Kullback-Leibler divergence (D_{KL}). Refinement was processed using only time points accessible by manual mixing (blue points) or by also taking the time points accessible to fast-mixing (red points) into account. The horizontal dashed black line represents the deviation of the ensemble refined with the protection factors, i.e. HDX probed by NMR.*

5.4.3 Combining smFRET and HDX-MS

Macromolecules such as the P4 helicases are not suitable for NMR and therefore their protection factors or chemical shifts are not readily accessible. The study of their structural dynamics is limited to sparser data such as HDX-MS or single-molecule FRET. As demonstrated above, these observables carry more ambiguous information, which resulted in poorer refinement. Previous ensemble refinement using only HDX-MS (with manual mixing) showed that the observable is relatively efficient to reject irrelevant structures but fails to reconstruct the relative population between the states N and I. When the ensemble was refined using only the FRET efficiency distribution of the label 12-28, the relative thermodynamic stability between N and I was correct but the ensemble still contained 45% of irrelevant conformers. To take advantage of the apparent complementarity of these two observables, they were simultaneously incorporated in the procedure to drive better refinement of the ensemble. Both observables had their

Lagrange coefficient set to 1 (see Section 2.3.2). The free-energy profile of the resulting refined ensemble is shown in Figure 5-10. The ensemble contained 21% of irrelevant conformers and the ratio I/(N+I) was now correct (Table 5-1). Although the new ensemble remained a poorer match than the ensembles obtained with the CS or the protection factors, a significant improvement was observed compared to the solution obtained with only HDX-MS or smFRET. This illustrates the benefit of combining different experimental data to maximize the structural information used to restrain the ensemble.

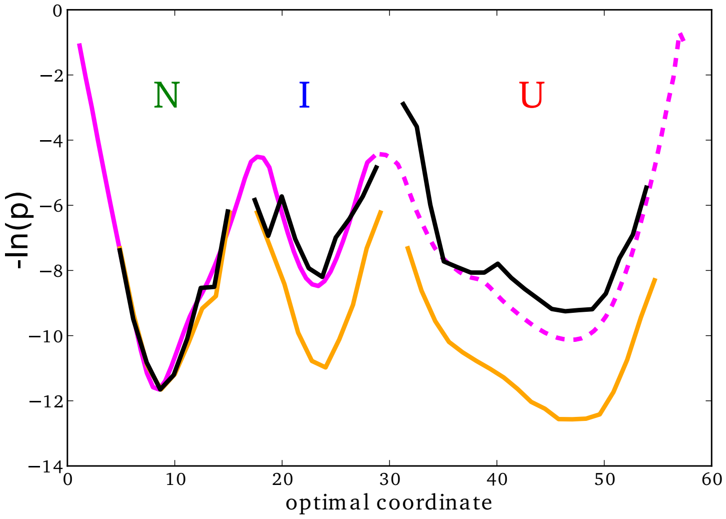


Figure 5-10: *FEP of the ensemble refined using HDX-MS combined with smFRET.*
The refinement was driven using time points accessible by manual mixing and the FRET efficiency distribution of the label 12-28. The profile of the reference ensemble (magenta), the conformer pool (orange) and the refined ensemble (black) are projected along the optimal coordinate. Dotted magenta line indicates the profile when the unfolded basin is included.

Observable	Relative population I/(N+I)	Irrelevant structures	D_{KL}	MSD
CS	5.1%	5.0%	1.1	1.81
SAXS	20.1%	68%	4.7	0.27
Pfact	4.6%	2.3%	0.9	0.10
cross-section	29.2%	79%	5.3	0.00
smFRET 1-35	24.0%	74%	5.0	8.06
smFRET 8-16	28.7%	29%	2.9	0.86
smFRET 7-33	15.2%	62%	4.3	8.48
smFRET 12-28	6.0%	45%	3.3	33.4
smFRET 12-28 and 8-16	5.5%	13%	1.5	16.8
HDX-MS manual	9.0%	25%	2.2	0.00
HDX-MS fast-mixing	9.5%	9.6%	1.4	0.00
HDX-MS and smFRET 12-28	4.8%	21%	2.0	16.7

Table 5-1*: Assessment of the refined ensembles.*

Two main criteria were used to assess the refined ensemble: the relative population between the native and intermediate states (I/(N+I)) in the refined ensemble and the percentage of irrelevant structures (i.e. structures from the unfolded state). The ratio I/(N+I) was calculated as the number of structures in energetic basin I divided by the number of structures in basins N and I. The values have to be compared with those of the reference ensemble and the pool. The reference ensemble contains 0% irrelevant structures and the ratio I/(N+I) is equal to 5.0%. In the pool, 81.4% of the structures are irrelevant and the ratio I/(N+I) is equal to 35.5%. The overall divergence between the reference and refined ensembles can also be estimated with the Kullback-Leibler divergence (D_{KL}). The last column indicates the averaged MSD of the observable back-calculated from the reference ensemble from the synthetic experimental data used to guide the refinement.

5.4.4 Concluding discussion

I examined how helpful different observables are to reconstruct a representative ensemble of a protein. As this approach was purely methodological, the framework was simplified in order to focus on a specific aspect of the refinement method, which was the structural information carried by an observable. Hence, several important simplifications were made throughout this work without affecting the pertinence of the results: (1) the conformer pool was assumed to be exhaustive enough to contain all the energetic basins visited by the protein in solution, (2) the models used to translate the experimental data into structural restraints were assumed to be accurate, (3) the variations of the experimental conditions (temperature, pH, pressure, concentration…) from one experimental technique to another were assumed to not affect the Boltzmann ensemble of the protein, (4) the experimental errors were neglected (apart from HDX-MS data). All these ideal conditions are usually not observed in practice.

Observables such as chemical shifts and protection factors were found to be sensitive enough to reconstruct the Boltzmann ensemble of the protein. To a lesser extent, smFRET also appears to be a good candidate to investigate the conformational space of the protein. As with all single-molecule techniques, smFRET affords more detailed structural and dynamical information that is not available in ensemble measurements due to averaging. However, the information provided by the distance between two specific residues, even at a single molecule level, remains limited. Several distance pairs had to be combined in order to guide the refinement towards the correct Boltzmann distribution. In comparison, SAXS and ion-mobility cross-section were shown to carry too much ambiguous structural information to effectively restrain the ensemble of structures.

It comes as no surprise that protection factors and chemical shifts, which depend on the local environment of each residue, carry more structural information than the cross-section or the SAXS profile of the protein, which provide only low-resolution data. In order to investigate the similarity of the structural information provided by two different observables, their correlation was estimated. Multiple refinements were performed in the previous section. Given two observables, the relationship between their respective mean square deviations over the different refinements (see explanations in Figure 5-11) was gathered in a scatter plot (Figure 5-12). Then, the coefficient of determination R^2 of the scatter plot was calculated to quantify the correlation between the two observables.

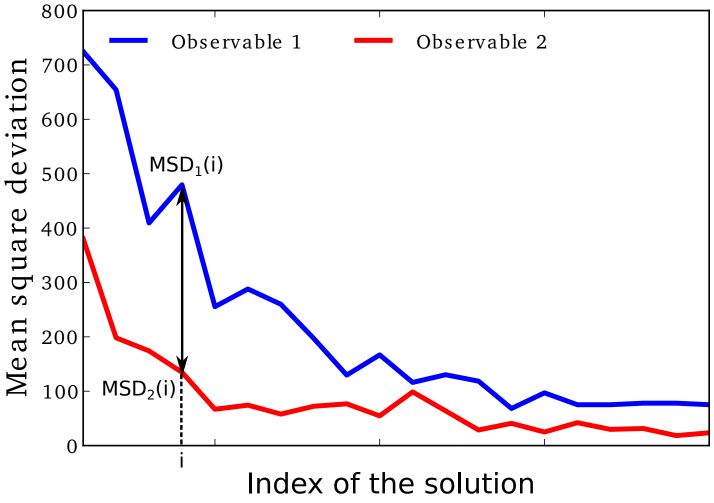


***Figure 5-11**: Construction of the scatter plots.*

Given two observables and a specific refinement, the mean square deviation of each observable were collected over the refinement. For each intermediate solution, i, saved by the genetic algorithm, the point ($MSD_1(i)$, $MSD_2(i)$) was added to the scatter plot. The process was repeated for all the refinements which had been performed in the results section.

We calculated the correlation between the protection factors and the chemical shifts (Figure 5-12A). Interestingly, the two observables are highly correlated (R^2 = 0.98). This clearly indicates that information carried by protection factors and chemical shifts are of the same kind. Consequently, the two observables are not complementary and combining them would not improve the refinement. *Per contra*, single-molecule FRET and HDX-MS data are strongly anti-correlated (Figure 5-12C). The scatter plot indicates that many ensembles compatible with HDX-MS data are not compatible with smFRET data, and *vice versa*. This explains why the combination of HDX-MS and smFRET data resulted in a better refinement. SAXS, as well as the ion-mobility cross-section, are de-correlated with all observables (see example in Figure 5-12B). Due to the low-resolution nature of their information, the observables are compatible with very diverse ensembles. It leads to a scattering of their values that eventually de-correlate them with other observables. SAXS has been used to refine the ensemble of intrinsically disordered proteins (189). The degeneracy observed during our analysis suggests that SAXS may not produce physically sound ensembles for flexible systems which are expected to have complex free-energy profiles.

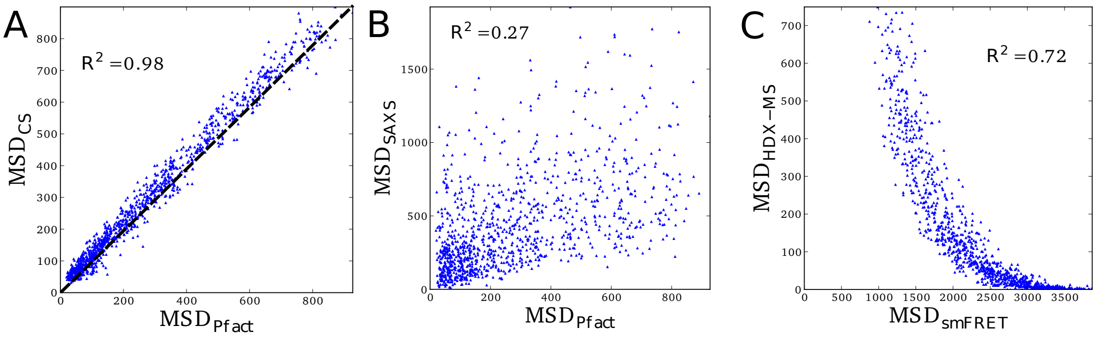


Figure 5-12: *Correlation between observables.*
The evolution of the mean square deviations (MSD) of the different observables during refinement guided by HDX-MS data are compared. (A) Comparison between MSD of protection factors and chemical shifts. (B) Comparison between MSD of SAXS and protection factors. (C) Comparison between MSD of single molecule FRET and HDX-MS.

It is worth noting that the number of structures used to calculate the synthetic experimental data (~600,000 structures) is much larger than the number of structures in the pool (5,000 structures). The disproportionate size difference between the two ensembles is also encountered on handling real experimental data, because the number of structures in the observed ensemble (*i.e.* the reference ensemble) is on the order of the Avogadro number N_A ~ 10^{23}. The ability of the refined ensemble to reproduce the free-energy profile of the protein, suggests that a limited number of structures is informative enough to illustrate the global Boltzmann ensemble of a protein.

Chapter 6: Conclusions and future perspectives

In this thesis, the mechanisms of the packaging motors P4 are investigated by combining simulations with sparse experimental data from hydrogen-exchange mass-spectrometry (HDX-MS) and fluorescence spectroscopy. The thesis is mainly focused on a new approach to quantitatively interpret deuterium labelling probed by mass-spectrometry. The overall message of the thesis is that, although data provided by HDX-MS are sparse, combining this technique with simulations enables to extract valuable structural and dynamical information. A summary of the conclusions and future prospects of each chapter is given bellow.

1) Chapter 3: functional dynamics of helicase probed by hydrogen deuterium exchange and simulation

The biological function of large macromolecular assemblies depends on their structure and their dynamics over a broad range of time- and spatial-scales. For this reason, it is challenging to investigate large complexes using conventional, high resolution experimental techniques. One of the most promising experimental techniques is hydrogen-deuterium exchange detected by mass spectrometry. In Chapter 3, a new computational method to qualitatively interpret hydrogen-deuterium exchange probed by mass-spectrometry was presented. The method was successfully tested on the packaging motor φ12 P4. This hexameric helicase unwinds and translocates single-stranded RNA into virus capsids at the expense of ATP hydrolysis. Room-temperature dynamics probed by a hundred nanoseconds of all-atom molecular dynamics simulations was sufficient to predict the exchange kinetics of most peptide fragments. The proposed method was also shown to be a powerful tool to validate the assignment of fragments and to assess structural models of polypeptide regions that were missing or disordered in the high resolution structure.

Since our methodology has proved a valuable tool for validating structural models, the approach could be used to restrain the docking of large complexes based on HDX-MS data. To enable fast calculations, the parameterisation of the phenomenological approximation used to predict the protection factors could be re-optimised for shorter conformational samplings (at the cost of lower accuracy). In the last part of Chapter 3, the limits of the hydrogen exchange model have been discussed. It appeared that the model suffers from neglecting the electrostatic effects on HDX kinetics. Different approaches previously suggested to integrate the impact of electrostatic on deuterium labelling were discussed. A more advanced HDX model, based on the model used throughout this thesis but integrating electrostatic effects, was finally proposed to guide future improvements of HDX prediction.

2) Chapter 4: Insights into helicase-RNA interaction from hydrogen exchange and fluorescence spectroscopy

The packaging motor φ8 P4 is structurally and functionally homologous to φ12 P4. When its X-ray crystal structure was finally published in 2013, surprisingly the C-termini appeared to be inside the central channel, restricting the entrance of the pore and occluding the interface through which RNA is thought to be loaded. It was suggested that the C-termini might come out upon RNA binding. To put to the test whether the C-termini remain inside the central channel or come out upon RNA loading, both scenarios were modelled. Their local conformational space was sampled for ~100-200 ns with MD simulations and the HDX kinetics of the C-terminal domain predicted for both structures. Comparison between the experimental and predicted exchange kinetics confirmed that only an exposition to the solvent of at least some C-termini could explain the fast exchange observed experimentally. The difficulties to construct a model with both the RNA and all C-termini inside the central channel also support the idea that some C-termini need to come out to make space for RNA. Ensemble FRET experiments suggest that only the C-terminal tail (residues 305-321) undergoes conformational changes upon RNA binding, while the rest of the C-terminal domain (residues 290-304) remains bound to the apical dome of the motor. Further analysis of the MD simulations revealed that the C-termini interact with the L1 loop when they are kept inside the pore. The “DK motif”, located at the very end of the C-terminus, exhibits high affinity for the LKK motif of the L1 loop, resulting in lower affinity of the L1 loop for RNA. It suggests that the C-terminus plays an essential role in translocation. The affinity of φ8 P4 for RNA is known to be higher than that of the other P4 motors, due to the pair of lysines in the L1 loop. I propose a new model for which part of the C-termini comes out the pore upon RNA binding, while the other C-termini remain inside the channel to modulate the affinity between the L1 loops and RNA during translocation. The work presented in this chapter is a nice example of how experiment and simulations can stimulate each other.

The revised model is mainly based on computational observations and therefore requires further experimental validations. Modifying the DK motif by site directed mutagenesis would enable to investigate its function. It would be interesting to know whether only the lysine K321 is essential for modulating RNA affinity or both the aspartic acid D320 and the lysine K321. A strong assumption of the model is that the C-terminus allows the detachment of RNA from the L1 loop only in the “down” position. This proposition could be verified by shortening the C-terminal such that the “DK motif” reaches the L1 loop only in the “up” position. The conformational changes of the apical dome of the motor observed by ensemble FRET upon translocation could be further investigated by smFRET. However, it would require first to address the dissociation problem of the hexamer upon extreme dilution.

To gain an understanding of the mechanisms of φ8 P4, it would be interesting to follow the translocation of the motor along RNA using single-molecule fluorescence spectroscopy. Total internal reflection fluorescence (TIRF) microscopy allows direct, time-resolved single-molecule imaging. Similar to the work of Deindl *et al.* (190), we propose to label φ8 P4 and an RNA strand in order to monitor the translocation of the motor by FRET. A schematic of the experimental setup is illustrated in Figure 6-1. The RNA strand would be labelled at its 3' end and would be completed by a complementary DNA strand. The resulting duplex would be stiff enough (lp ~ 50nm) to limit its bending, such that the position of the motor along RNA could be directly deduced from the distance between the two dyes.

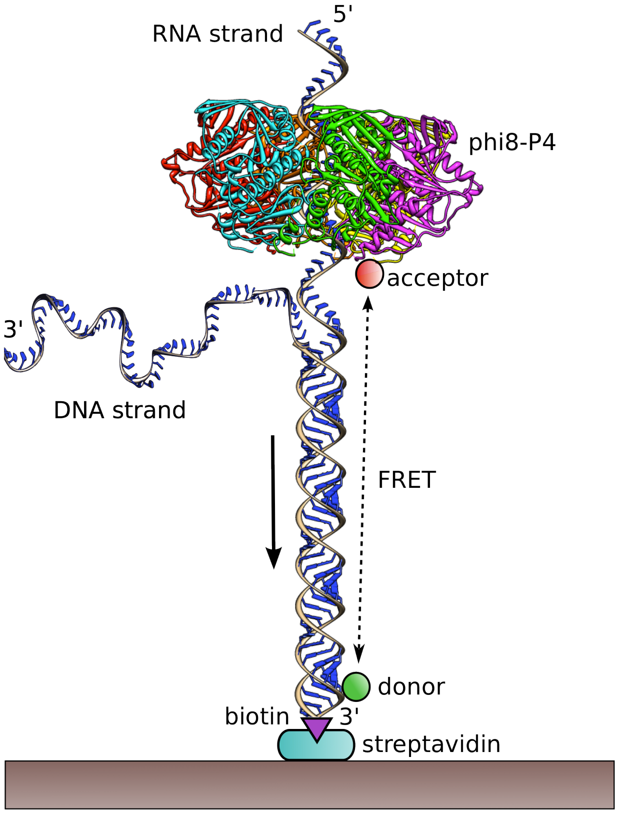


Figure 6-1: *Schematic of experimental setup to study P4 translocation by TIRF.*
The RNA strand would be immobilized at its 3' end on the surface using a streptavidin/biotin complex. A complementary DNA strand would reinforce the stiffness of the nucleic acid chain. The 3' end of the RNA strand and the top of φ8 P4 would be labelled with A488 and A594 dyes, respectively. The FRET activity would be measured by total internal fluorescence spectroscopy.

3) Chapter 5: High-resolution models of protein states from sparse experimental data

In Chapter 5, I investigated the structural information provided by sparse data such as HDX-MS kinetics, NMR chemical-shifts (CS), ion mobility cross-sections, small-angle X-ray scattering and single-molecule FRET (smFRET). A protocol to assess the information content carried by an observable or their combination was devised. The assessment was based on the ability of an observable to guide the reconstruction of the Boltzmann ensemble of the protein FIP35. The closer to the Boltzmann ensemble the refinement was, the more informative the observable was considered. Observable such as CS and HDX kinetics appeared to carry more valuable information than SAXS or ion-mobility cross-section. The information carried by the HDX kinetics of a peptide fragment was compared to the information provided by HDX kinetics when probed at a residue level (HDX-NMR). Unsurprisingly, HDX-MS data appeared to be less informative than HDX-NMR data. It was shown that decreasing the size (down to about 5 residues) of the peptide fragments and acquiring faster exchange kinetics would allow HDX-MS data to be as informative as the high resolution HDX-NMR data. The complementarity of HDX with other observables was also examined. HDX and CS were shown to provide very similar structural information, whereas smFRET data appeared complementary to HDX data. Hence, combining HDX-MS and smFRET appears as a promising way to study the structural dynamics of large macromolecular complexes.

List of References

1. Fraenkel-Conrat H, Williams RC (1955) Reconstitution of active tabacco mosaic virus from its inactive protein and nucleic acid components. *Proc Natl Acad Sci* 41(10):690–698.

2. Stockley PG, et al. (2007) A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. *J Mol Biol* 369(2):541–552.

3. Riemer SC, Bloomfield VA (1978) Packaging of DNA in bacteriophage heads: some considerations on energetics. *Biopolymers* 17(3):785–794.

4. Semancik JS, Vidaver AK, Van Etten JL (1973) Characterization of segmented double-helical RNA from bacteriophage phi6. *J Mol Biol* 78(4):617–625.

5. Butcher SJ, Dokland T, Ojala PM, Bamford DH, Fuller SD (1997) Intermediates in the assembly pathway of the double-stranded RNA virus phi6. *EMBO J* 16(14):4477–4487.

6. Pirttimaa MJ, Paatero AO, Frilander MJ, Bamford DH (2002) Nonspecific nucleoside triphosphatase P4 of double-stranded RNA bacteriophage phi6 is required for single-stranded RNA packaging and transcription. *J Virol* 76(20):10122–10127.

7. Makeyev EV, Bamford DH (2000) The polymerase subunit of a dsRNA virus plays a central role in the regulation of viral RNA metabolism. *EMBO J* 19(22):6275–6284.

8. Makeyev EV, Bamford DH (2000) Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage phi6. *EMBO J* 19(1):124–133.

9. Poranen MM, Butcher SJ, Simonov VM, Laurinmäki P, Bamford DH (2008) Roles of the minor capsid protein P7 in the assembly and replication of double-stranded RNA bacteriophage phi6. *J Mol Biol* 383(3):529–538.

10. Poranen MM, Paatero AO, Tuma R, Bamford DH (2001) Self-assembly of a viral molecular machine from purified protein and RNA constituents. *Mol Cell* 7(4):845–854.

11. Hoogstraten D, et al. (2000) Characterization of phi8, a bacteriophage containing three double-stranded RNA genomic segments and distantly related to Phi6. *Virology* 272(1):218–224.

12. Bamford DH, Romantschuk M, Somerharju PJ (1987) Membrane fusion in prokaryotes: bacteriophage phi 6 membrane fuses with the Pseudomonas syringae outer membrane. *EMBO J* 6(5):1467–1473.

13. Qiao X, Qiao J, Mindich L (1997) Stoichiometric packaging of the three genomic segments of double-stranded RNA bacteriophage phi6. *Proc Natl Acad Sci* 94(8):4074–4079.

14. Kainov DE, et al. (2003) RNA packaging device of double-stranded RNA bacteriophages, possibly as simple as hexamer of P4 protein. *J Biol Chem* 278(48):48084–48091.

15. Huiskonen JT, et al. (2006) Structure of the bacteriophage phi6 nucleocapsid suggests a mechanism for sequential RNA packaging. *Struct Lond Engl 1993* 14(6):1039–1048.

16. Kainov DE, Lísal J, Bamford DH, Tuma R (2004) Packaging motor from double-stranded RNA bacteriophage phi12 acts as an obligatory passive conduit during transcription. *Nucleic Acids Res* 32(12):3515–3521.

17. Omari K El, et al. (2013) Tracking in atomic detail the functional specializations in viral RecA helicases that occur during evolution. *Nucleic Acids Res* 41(20):9396–9410.

18. Mancini EJ, et al. (2004) Atomic snapshots of an RNA packaging motor reveal conformational changes linking ATP hydrolysis to RNA translocation. *Cell* 118(6):743–755.

19. Singleton MR, Dillingham MS, Wigley DB (2007) Structure and mechanism of helicases and nucleic acid translocases. *Annu Rev Biochem* 76:23–50.

20. Ye J, Osborne AR, Groll M, Rapoport TA (2004) RecA-like motor ATPases - lessons from structures. *Biochim Biophys Acta* 1659(1):1–18.

21. Ilyina TV, Gorbalenya AE, Koonin EV (1992) Organization and evolution of bacterial and bacteriophage primase-helicase systems. *J Mol Evol* 34(4):351–357.

22. Lísal J, Tuma R (2005) Cooperative mechanism of RNA packaging motor. *J Biol Chem* 280(24):23157–23164.

23. Kötting C, Kallenbach A, Suveyzdis Y, Wittinghofer A, Gerwert K (2008) The GAP arginine finger movement into the catalytic site of Ras increases the activation entropy. *Proc Natl Acad Sci* 105(17):6260–6265.

24. Kainov DE, et al. (2008) Structural basis of mechanochemical coupling in a hexameric molecular motor. *J Biol Chem* 283(6):3607–3617.

25. Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, Gunsalus IC (1975) Dynamics of ligand binding to myoglobin. *Biochemistry (Mosc)* 14(24):5355–5373.

26. Frauenfelder H, Fenimore PW, Chen G, McMahon BH (2006) Protein folding is slaved to solvent motions. *Proc Natl Acad Sci* 103(42):15469–15472.

27. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. *Nature* 450(7172):964–972.

28. Freddolino PL, Liu F, Gruebele M, Schulten K (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J* 94(10):75–77.

29. Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci* 101(41):14766–14770.

30. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci* 89(18):8721–8725.

31. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct Funct Bioinforma* 21(3):167–195.

32. Artymiuk PJ, et al. (1979) Crystallographic studies of the dynamic properties of lysozyme. *Nature* 280(5723):563–568.

33. Brooks III CL, Karplus M, Pettitt BM, Austin RH (2008) Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Phys Today* 43(2):120–122.

34. Daniel RM, Dunn RV, Finney JL, Smith JC (2003) The role of dynamics in enzyme activity. *Annu Rev Biophys Biomol Struct* 32:69–92.

35. Eisenmesser EZ, et al. (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438(7064):117–121.

36. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197–208.

37. Verkhivker GM, et al. (2003) Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding. *Proc Natl Acad Sci U S A* 100(9):5148–5153.

38. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. *PNAS* 102(19):6679–6685.

39. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* 267(5612):585–590.

40. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517–520.

41. Zhao G, et al. (2013) Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497(7451):643–646.

42. Elber R (2005) Long-timescale simulation methods. *Curr Opin Struct Biol* 15(2):151–156.

43. G. M. Torrie JPV (1977) Non-physical sampling distributions in monte-carlo free-energy estimation - umbrella sampling. *J Comput Phys* 23(2):187–199.

44. Mitsutake A, Sugita Y, Okamoto Y (2003) Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test. *J Chem Phys* 118(14):6664–6675.

45. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Chem Phys* 120(24):11919–11929.

46. Czerminski R, Elber R (1989) Reaction path study of conformational transitions and helix formation in a tetrapeptide. *Proc Natl Acad Sci* 86(18):6963–6967.

47. Pan AC, Sezer D, Roux B (2008) Finding transition pathways using the string method with swarms of trajectories. *J Phys Chem B* 112(11):3432–3440.

48. Mills G, Jacobsen KW, Jonsson H (1998) Nudged elastic band method for finding minimum energy paths of transitions. *Classical and Quantum Dynamics in Condensed Phase Simulations* (World Scientific).

49. Chu J-W, Trout BL, Brooks III CL (2003) A super-linear minimization scheme for the nudged elastic band method. *J Chem Phys* 119(24):12708–12717.

50. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* 53:291–318.

51. Faradjian AK, Elber R (2004) Computing time scales from reaction coordinates by milestoning. *J Chem Phys* 120(23):10880–10889.

52. Saunders MG, Voth GA (2013) Coarse-graining methods for computational biology. *Annu Rev Biophys* 42(1):73–93.

53. Ma W, Schulten K (2015) Mechanism of substrate translocation by a ring-shaped ATPase motor at millisecond resolution. *J Am Chem Soc* 137(8):3031–3040.

54. Liu H, Shi Y, Chen XS, Warshel A (2009) Simulating the electrostatic guidance of the vectorial translocations in hexameric helicases and translocases. *Proc Natl Acad Sci* 106(18):7449–7454.

55. Yoshimoto K, Arora K, Brooks III CL (2010) Hexameric helicase deconstructed: Interplay of conformational changes and substrate coupling. *Biophys J* 98(8):1449–1457.

56. Mackerell AD Jr (2004) Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem* 25(13):1584–1604.

57. Beauchamp KA, Lin Y-S, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 8(4):1409–1414.

58. Piana S, Klepeis JL, Shaw DE (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol* 24:98–105.

59. Wüthrich K, Wagner G (1978) Internal motion in globular proteins. *Trends Biochem Sci* 3(4):227–230.

60. Igumenova TI, Frederick KK, Wand AJ (2006) Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem Rev* 106(5):1672–1699.

61. Palmer AG, Massi F (2006) Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy. *Chem Rev* 106(5):1700–1719.

62. Dobson CM, Karplus M (1986) Internal motion of proteins: nuclear magnetic resonance measurements and dynamic simulations. *Methods Enzymol* 131:362–389.

63. Chapman HN, et al. (2011) Femtosecond X-ray protein nanocrystallography. *Nature* 470(7332):73–77.

64. Schotte F, Soman J, Olson JS, Wulff M, Anfinrud PA (2004) Picosecond time-resolved X-ray crystallography: probing protein function in real time. *J Struct Biol* 147(3):235–246.

65. Różycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Struct Lond Engl 1993* 19(1):109–116.

66. Roy R, Hohng S, Ha T (2008) A practical guide to single-molecule FRET. *Nat Methods* 5(6):507–516.

67. Zhou ZH (2011) Atomic resolution cryo electron microscopy of macromolecular complexes. *Adv Protein Chem Struct Biol* 82:1–35.

68. Villarreal SA, Stewart PL (2014) CryoEM and image sorting for flexible protein/DNA complexes. *J Struct Biol* 187(1):76–83.

69. Joo C, Balci H, Ishitsuka Y, Buranachai C, Ha T (2008) Advances in single-molecule fluorescence methods for molecular biology. *Annu Rev Biochem* 77:51–76.

70. Craggs TD, Kapanidis AN (2012) Six steps closer to FRET-driven structural biology. *Nat Methods* 9(12):1157–1158.

71. Clarke J, Fersht AR (1996) An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. *Fold Des* 1(4):243–254.

72. Krishna MMG, Hoang L, Lin Y, Englander SW (2004) Hydrogen exchange methods to study protein folding. *Methods San Diego Calif* 34(1):51–64.

73. Morozova LA, Haynie DT, Arico-Muendel C, Van Dael H, Dobson CM (1995) Structural basis of the stability of a lysozyme molten globule. *Nat Struct Biol* 2(10):871–875.

74. Vendruscolo M, Paci E, Dobson CM, Karplus M (2003) Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *J Am Chem Soc* 125(51):15686–15687.

75. Dempsey CE (2001) Hydrogen exchange in peptides and proteins using NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 39(2):135–170.

76. Englander SW, Sosnick TR, Englander JJ, Mayne L (1996) Mechanisms and uses of hydrogen exchange. *Curr Opin Struct Biol* 6(1):18–23.

77. Zhang Z, Smith DL (1993) Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci Publ Protein Soc* 2(4):522–531.

78. Lísal J, et al. (2005) Functional visualization of viral molecular motor by hydrogen-deuterium exchange reveals transient states. *Nat Struct Mol Biol* 12(5):460–466.

79. Hilser VJ, Freire E (1996) Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J Mol Biol* 262(5):756–772.

80. Craig PO, et al. (2011) Prediction of native-state hydrogen exchange from perfectly funneled energy landscapes. *J Am Chem Soc* 133(43):17463–17472.

81. Bahar I, Wallqvist A, Covell DG, Jernigan RL (1998) Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry (Mosc)* 37(4):1067–1075.

82. Truhlar SME, Croy CH, Torpey JW, Koeppe JR, Komives EA (2006) Solvent accessibility of protein surfaces by amide H/2H exchange MALDI-TOF mass spectrometry. *J Am Soc Mass Spectrom* 17(11):1490–1497.

83. Best RB, Vendruscolo M (2006) Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Struct Lond Engl 1993* 14(1):97–106.

84. Park I-H, et al. (2015) Estimation of hydrogen-exchange protection factors from MD simulation based on amide hydrogen bonding analysis. *J Chem Inf Model* 55(9):1914–1925.

85. Fersht A (1985) *Enzyme structure and mechanism* (Biochemical Education).

86. Skinner JJ, Lim WK, Bédard S, Black BE, Englander SW (2012) Protein dynamics viewed by hydrogen exchange. *Protein Sci* 21(7):996–1005.

87. Bai Y (2006) Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem Rev* 106(5):1757–1768.

88. Katta V, Chait BT (1993) Hydrogen/deuterium exchange electrospray ionization mass spectrometry: a method for probing protein conformational changes in solution. *J Am Chem Soc* 115(14):6317–6321.

89. Englander JJ, et al. (2003) Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proc Natl Acad Sci* 100(12):7057–7062.

90. Konermann L, Pan J, Liu Y-H (2011) Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem Soc Rev* 40(3):1224–1234.

91. Lam TT, et al. (2002) Mapping of protein:protein contact surfaces by hydrogen/deuterium exchange, followed by on-line high-performance liquid chromatography-electrospray ionization Fourier-transform ion-cyclotron-resonance mass analysis. *J Chromatogr A* 982(1):85–95.

92. Kan Z-Y, Mayne L, Chetty PS, Englander SW (2011) ExMS: data analysis for HX-MS experiments. *J Am Soc Mass Spectrom* 22(11):1906–1915.

93. Suchanova B, Tuma R (2008) Folding and assembly of large macromolecular complexes monitored by hydrogen-deuterium exchange and mass spectrometry. *Microb Cell Factories* 7:12.

94. Landgraf RR, Chalmers MJ, Griffin PR (2012) Automated hydrogen/deuterium exchange electron transfer dissociation high resolution mass spectrometry measured at single-amide resolution. *J Am Soc Mass Spectrom* 23(2):301–309.

95. Kan Z-Y, Walters BT, Mayne L, Englander SW (2013) Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc Natl Acad Sci* 110(41):16438–16443.

96. Huang RY-C, Garai K, Frieden C, Gross ML (2011) Hydrogen/deuterium exchange and electron-transfer dissociation mass spectrometry determine the interface and dynamics of apolipoprotein E oligomerization. *Biochemistry (Mosc)* 50(43):9273–9282.

97. Schuler B (2013) Single-molecule FRET of protein structure and dynamics - a primer. *J Nanobiotechnology* 11 Suppl 1:S2.

98. Radou G, Dreyer FN, Tuma R, Paci E (2014) Functional dynamics of hexameric helicase probed by hydrogen exchange and simulation. *Biophys J* 107(4):983–990.

99. Phillips JC, et al. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802.

100. Verlet L (1967) Computer "'experiments"' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys Rev* 159(1):98.

101. Best RB, et al. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. *J Chem Theory Comput* 8(9):3257–3273.

102. Ponder JW, Case DA (2003) Force fields for protein simulations. *Adv Protein Chem* 66:27–85.

103. Lennard-Jones JE (1931) Cohesion. *Proc Phys Soc* 43(5):461.

104. Mackerell AD Jr, Feig M, Brooks CL 3rd (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11):1400–1415.

105. van Kampen NG (1992) *Stochastic processes in physics and chemistry* (Elsevier).

106. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.

107. Feig Mi (2010) *Modeling solvent environments* (John Wiley & Sons, Ltd).

108. Guillot B (2002) A reappraisal of what we have learnt during three decades of computer simulations on water. *J Mol Liq* 101(1–3):219–260.

109. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. *Protein Sci Publ Protein Soc* 9(9):1753–1773.

110. Brooks BR, et al. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4(2):187–217.

111. Molday RS, Englander SW, Kallen RG (1972) Primary structure effects on peptide group hydrogen exchange. *Biochemistry (Mosc)* 11(2):150–158.

112. Bai Y, Milne JS, Mayne L, Englander SW (1993) Primary structure effects on peptide group hydrogen exchange. *Proteins* 17(1):75–86.

113. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129(17):5656–5664.

114. Svergun D, Barberato C, Koch MHJ (1995) *CRYSOL* – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* 28(6):768–773.

115. Atri V D', Porrini M, Rosu F, Gabelica V (2015) Linking molecular models with ion mobility experiments. Illustration with a rigid nucleic acid structure. *J Mass Spectrom* 50(5):711–726.

116. Ruotolo BT, Benesch JLP, Sandercock AM, Hyung S-J, Robinson CV (2008) Ion mobility–mass spectrometry analysis of large protein complexes. *Nat Protoc* 3(7):1139–1152.

117. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37:145–151.

118. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86.

119. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J Am Chem Soc* 131(39):13894–13895.

120. Seeber M, Cecchini M, Rao F, Settanni G, Valencia PA Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23(19):2625–2627.

121. Kainov DE, Butcher SJ, Bamford DH, Tuma R (2003) Conserved intermediates on the assembly pathway of double-stranded RNA bacteriophages. *J Mol Biol* 328(4):791–804.

122. Elson EL (2011) Fluorescence Correlation Spectroscopy: Past, Present, Future. *Biophys J* 101(12):2855–2870.

123. Widengren J, Mets U, Rigler R (1995) Fluorescence correlation spectroscopy of triplet states in solution: a theoretical and experimental study. *J Phys Chem* 99(36):13368–13379.

124. Hohlbein J, Craggs TD, Cordes T (2014) Alternating-laser excitation: single-molecule FRET and beyond. *Chem Soc Rev* 43(4):1156–1171.

125. Kapanidis AN, et al. (2004) Fluorescence-aided molecule sorting: Analysis of structure and interactions by alternating-laser excitation of single molecules. *Proc Natl Acad Sci* 101(24):8936–8941.

126. Lee NK, et al. (2005) Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys J* 88(4):2939–2953.

127. Sharma A, et al. (2014) Domain movements of the enhancer-dependent sigma factor drive DNA delivery into the RNA polymerase active site: insights from single molecule studies. *Nucleic Acids Res* 42(8):5177–5190.

128. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9(9):646–652.

129. Frauenfelder H, et al. (2009) A unified model of protein dynamics. *Proc Natl Acad Sci* 106(13):5129–5134.

130. Huang YJ, Montelione GT (2005) Structural biology: Proteins flex to function. *Nature* 438(7064):36–37.

131. Lísal J, et al. (2006) Interaction of packaging motor with the polymerase complex of dsRNA bacteriophage. *Virology* 351(1):73–79.

132. Lobanov MY, et al. (2013) A novel web server predicts amino acid residue protection against hydrogen–deuterium exchange. *Bioinformatics* 29(11):1375–1381.

133. Liu T, et al. (2012) Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *J Am Soc Mass Spectrom* 23(1):43–56.

134. Petruk AA, et al. (2012) Molecular dynamics simulations provide atomistic insight into hydrogen exchange mass spectrometry experiments. *J Chem Theory Comput* 9(1):658–669.

135. Paatero AO, Mindich L, Bamford DH (1998) Mutational analysis of the role of nucleoside triphosphatase P4 in the assembly of the RNA polymerase complex of bacteriophage phi6. *J Virol* 72(12):10058–10065.

136. Mancini EJ, Tuma R (2012) Mechanism of RNA packaging motor. *Adv Exp Med Biol* 726:609–629.

137. Kainov DE, Tuma R, Mancini EJ (2006) Hexameric molecular motors: P4 packaging ATPase unravels the mechanism. *Cell Mol Life Sci CMLS* 63(10):1095–1105.

138. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815.

139. Huiskonen JT, Jäälinoja HT, Briggs JAG, Fuller SD, Butcher SJ (2007) Structure of a hexameric RNA packaging motor in a viral polymerase complex. *J Struct Biol* 158(2):156–164.

140. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Proteins Struct Funct Bioinforma* 35(2):133–152.

141. Bottaro S, Lindorff-Larsen K, Best RB (2013) Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J Chem Theory Comput* 9(12):5641–5652.

142. Kleinjung J, Fraternali F (2014) Design and application of implicit solvent models in biomolecular simulations. *Curr Opin Struct Biol* 25(100):126–134.

143. Clarke J, Hounslow AM, Bycroft M, Fersht AR (1993) Local breathing and global unfolding in hydrogen exchange of barnase and its relationship to protein folding pathways. *Proc Natl Acad Sci* 90(21):9837–9841.

144. Chamberlain AK, Handel TM, Marqusee S (1996) Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat Struct Mol Biol* 3(9):782–787.

145. Schulman BA, Redfield C, Peng Z, Dobson CM, Kim PS (1995) Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin. *J Mol Biol* 253(5):651–657.

146. Kim KS, Fuchs JA, Woodward CK (1993) Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry (Mosc)* 32(37):9600–9608.

147. Wolynes PG (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci* 94(12):6170–6175.

148. Linderstrom-Lang KU (1924) On the ionisation of proteins. *CR Trav Lab Carlsberg* 15:1–29.

149. Tanford C, Kirkwood JG (1957) Theory of protein titration curves. I. General equations for impenetrable spheres. *J Am Chem Soc* 79(20):5333–5339.

150. Shire SJ, Hanania GI, Gurd FR (1974) Electrostatic effects in myoglobin. Hydrogen ion equilibria in sperm whale ferrimyoglobin. *Biochemistry (Mosc)* 13(14):2967–2974.

151. Matthew JB, Richards FM (1983) The pH dependence of hydrogen exchange in proteins. *J Biol Chem* 258(5):3039–3044.

152. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol* 103(2):227–249.

153. Delepierre M, et al. (1987) Electrostatic effects and hydrogen exchange behaviour in proteins: The pH dependence of exchange rates in lysozyme. *J Mol Biol* 197(1):111–122.

154. Matthew JB, Richards FM (1982) Anion binding and pH-dependent electrostatic effects in ribonuclease. *Biochemistry (Mosc)* 21(20):4989–4999.

155. Hernández G, Anderson JS, LeMaster DM (2009) Polarization and polarizability assessed by protein amide acidity. *Biochemistry (Mosc)* 48(27):6482–6494.

156. LeMaster DM, Anderson JS, Hernández G (2009) Peptide conformer acidity analysis of protein flexibility monitored by hydrogen exchange. *Biochemistry (Mosc)* 48(39):9256–9265.

157. Li L, et al. (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys* 5(1):9.

158. Skordalakes E, Berger JM (2003) Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell* 114(1):135–146.

159. Picha KM, Ahnert P, Patel SS (2000) DNA binding in the central channel of bacteriophage T7 helicase-primase is a multistep process. Nucleotide hydrolysis is not required. *Biochemistry (Mosc)* 39(21):6401–6409.

160. Lísal J, Kainov DE, Bamford DH, Thomas GJ Jr, Tuma R (2004) Enzymatic mechanism of RNA translocation in double-stranded RNA bacteriophages. *J Biol Chem* 279(2):1343–1350.

161. Thomsen ND, Berger JM (2009) Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell* 139(3):523–534.

162. Tan Y-W, Hanson JA, Yang H (2009) Direct Mg2+ binding activates adenylate kinase from Escherichia coli. *J Biol Chem* 284(5):3306–3313.

163. Kadaré G, Haenni AL (1997) Virus-encoded RNA helicases. *J Virol* 71(4):2583–2590.

164. Borodavka A, Ault J, Stockley PG, Tuma R (2015) Evidence that avian reovirus σNS is an RNA chaperone: implications for genome segment assortment. *Nucleic Acids Res* 109(39):15769–15774.

165. Mindich L, et al. (1999) Isolation of additional bacteriophages with genomes of segmented double-stranded RNA. *J Bacteriol* 181(15):4505–4508.

166. Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* 40(3):191–285.

167. Fisher CK, Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* 21(3):426–431.

168. Receveur-Brechot V, Durand D (2012) How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci* 13(1):55–75.

169. Habeck M (2011) Statistical mechanics analysis of sparse data. *J Struct Biol* 173(3):541–548.

170. Madl T, Gabel F, Sattler M (2011) NMR and small-angle scattering-based structural analysis of protein complexes in solution. *J Struct Biol* 173(3):472–482.

171. Ren X, et al. (2009) Hybrid structural model of the complete human ESCRT-0 complex. *Struct Lond Engl 1993* 17(3):406–416.

172. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* 38:540–544.

173. de Bakker PIW, Furnham N, Blundell TL, DePristo MA (2006) Conformer generation under restraints. *Curr Opin Struct Biol* 16(2):160–165.

174. Alber F, Förster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 77:443–477.

175. Beauchamp KA, Pande VS, Das R (2014) Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys J* 106(6):1381–1390.

176. Kozak M, Lewandowska A, Ołdziej S, Rodziewicz-Motowidło S, Liwo A (2010) Combination of SAXS and NMR techniques as a tool for the determination of peptide structure in solution. *J Phys Chem Lett* 1(20):3128–3131.

177. Richter B, Gsponer J, Várnai P, Salvatella X, Vendruscolo M (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J Biomol NMR* 37(2):117–135.

178. Best RB, Vendruscolo M (2004) Determination of protein structures consistent with NMR order parameters. *J Am Chem Soc* 126(26):8090–8091.

179. Deshmukh L, et al. (2013) Structure and dynamics of full-length HIV-1 capsid protein in solution. *J Am Chem Soc* 135(43):16133–16147.

180. Turjanski AG, Gutkind JS, Best RB, Hummer G (2008) Binding-induced folding of a natively unstructured transcription factor. *PLoS Comput Biol* 4(4).

181. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K (2014) Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* 10(2).

182. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. *Phys Rev Lett* 63(12):1195–1198.

183. Roux B, Weare J (2013) On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys* 138(8).

184. Jones G (2002) Genetic and evolutionary algorithms. *Encyclopedia of Computational Chemistry* (John Wiley & Sons, Ltd).

185. Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with Bayesian statistics. *J Am Chem Soc* 132(42):14919–14927.

186. Shaw DE, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330(6002):341–346.

187. Krivov SV (2011) The free energy landscape analysis of protein (FIP35) folding dynamics. *J Phys Chem B* 115(42):12315–12324.

188. Wishart D, Sykes B, Richards F (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. 222(2):311–333.

189. Bernadó P, Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* 8(1):151–167.

190. Deindl S, et al. (2013) ISWI remodelers slide nucleosomes with coordinated multi-base-pair entry steps and single-base-pair exit steps. *Cell* 152(3):442–452.

Appendix

Mean square deviation between calculated and measured protection factors:

Mean square deviation was calculated such as:

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N}\left(P_i^{exp} - P_i^{sim}\right)^2$$

Where P_i^{exp} and P_i^{sim} are the measured and calculated protection factors of residue *i*, respectively. N is the number of residue of the protein for which the experimental protection factor value was available.

The value of the dielectric constant:

The dielectric properties of a system can be described by the dielectric constant that reflects the reorientation of dipoles under the local electric field. The higher the dielectric constant is, the easier the local field can reorient the dipole. A high dielectric value is used to model the solvent (usually 80) due to the high mobility of water molecules, while a small one is used to model the environment inside the protein, where the permanent dipoles are virtually fixed. The value of the dielectric constant inside the protein is controversial. It seems that a small dielectric constant is more appropriate for buried residues (~4) and a higher value (~20) is necessary for residues located at the protein surface.

It is important to note that the behaviour of the dielectric constant in water and in a protein is very different since permanent dipoles have more restrictive mobility in the protein. Furthermore, using a unique average dielectric constant in a protein would lead to underestimate interactions with a charge and a fix dipole while the model would overestimate the interactions between a charge and a highly fluctuating dipole.

Relation between the exchange rate constant and the minimum pH point:

Hydrogen exchange is catalysed by water ions and its rate constant, *k*, can be expressed as:

$$k = k_{H_2O} + k_{H_3O^+}[H_3O^+] + k_{OH^-}[OH^-]$$

Or:

$$k = k_{H_2O} + k_{H_3O^+}[H_3O^+] + k_{OH^-}\frac{K_w}{[H_3O^+]}$$

where $K_w\ (0°C) = 10^{-14.94}$. Deriving the previous equation gives:

$$\frac{\partial k}{\partial [H_3O^+]} = k_{H_3O^+} - k_{OH^-}\frac{K_w}{[H_3O^+]^2}$$

At the minimum pH point, *pH_{min}*, $\frac{\partial k}{\partial [H_3O^+]} = 0$. It leads to:

$$[H_3O^+]_{min} = \sqrt{\frac{K_w k_{OH^-}}{k_{H_3O^+}}}$$

So:

$$pH_{min} = -\log\left([H_3O^+]_{min}\right) = -\frac{1}{2}\log\left(\frac{k_{OH^-}K_w}{k_{H_3O^+}}\right)$$

Poisson's equation:

The Maxwell's equations give:

$$\nabla.\mathrm{E} = \frac{\rho}{\varepsilon_0}$$

where ρ is the charge density, $\varepsilon_0$the permittivity and E the electric field. By definition, the electric field is related to the electric potential, ϕ, as such:

$$\mathrm{E} = -\nabla\phi$$

It gives the Poisson's equation:

$$\nabla.\nabla\phi = \nabla^2\phi = -\frac{\rho}{\varepsilon_0}$$

when the space is free of charge, the Poisson's equation gives the Laplace's equation:

$$\nabla^2 \phi = 0$$

The Poisson-Boltzmann equation:

When solvent is modelled implicitly, screening effect due to ions is modelled by introducing an extra term into the Poisson equation:

$$\nabla[\varepsilon(\mathrm{r})\nabla\phi(\mathrm{r})] = -4\pi\left(\rho(r) + \sum_{i=1}^{N} c_i^{bulk}\, Z_i e_0 e^{-(Z_i e_0 \phi(\mathrm{r})/RT)}\right)$$

where *N* is the number of charges particles, $\varepsilon(\mathrm{r})$ is the spatial varying dielectric constant, c_i^{bulk} is the concentration of ions i in the bulk and Z_i is their charge. It assumes that ions are distributed according to the Boltzmann distribution, hence the name of the equation. The dielectric constant allows scaling the electronic energies that is stored by the system by means of polarization, i.e. the induced dipoles in the protein and the solvent are modelled implicitly. In other words, the dielectric constant measures all the interactions that are not treated explicitly in the model, explaining why its value depends on the model and the site considered. The modified version of the Poisson equation is called the Poisson-Boltzmann (PB) equation.

For small electrostatic potentials ($e_0\phi(\mathrm{r})/RT \ll 1$), the equation can be linearized by expanding up to the linear term:

$$\nabla[\varepsilon(\mathrm{r})\nabla\phi(\mathrm{r})] = -4\pi\left(\rho(r) + \sum_{i=1}^{N} c_i^{bulk}\, Z_i e_0 - \sum_{i=1}^{N} c_i^{bulk}\, {Z_i}^2 {e_0}^2 \frac{\phi(\mathrm{r})}{RT}\right)$$

The first term is equal to zero because of the electro-neutrality of the solution. It is useful to introduce the terms:

$$I = \frac{1}{2}\sum_{i=1}^{N} c_i^{bulk}\, {Z_i}^2$$

$$\varkappa^2 = \frac{8\pi {e_0}^2 I}{RT}$$

where $\varkappa^2$ is called the Debye-Huckel screening parameter and I the ionic strength. Hence:

$$\nabla[\varepsilon(\mathrm{r})\nabla\phi(\mathrm{r})] = -4\pi\rho(r) + \varkappa^2\phi(\mathrm{r})$$

Analytical solution for the linearized Poisson-Boltzmann equation exists only for simple systems and numerical methods (finite element method) are required in most of the case. This equation poses a problem at a molecular level where the continuum assumption does not hold and the nature of the electrostatic constant is not clear in heterogeneous milieu.

The Generalized Born model:

Initially, the born model estimates the electrostatic component, $\Delta\mathrm{G}_{el}$, of the free energy of solvation for placing a charge in a spherical solvent cavity. It postulates that the solvation energy is equal to the work done to transfer the ion from vacuum to the medium:

$$\Delta\mathrm{G}_{el} = -\frac{q^2}{2a}\left(1 - \frac{1}{\varepsilon}\right)$$

where *q* is the charge of the particle, *a* its Born radii estimated from the crystal structure and ε dielectric constant of the solvent. For multiple charged points, the Coulomb interactions between each pair of charges needs to be added to the electrostatic free energy term. The resulting equation is called the Generalized Born model:

$$\Delta\mathrm{G}_{el} = \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{q_i q_j}{\varepsilon r_{ij}} - \frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\frac{{q_i}^2}{a_i}$$

By introducing the empirical function:

$$f_{ij} = \sqrt{r_{ij}^2 + a_i a_j e^{-\left(r_{ij}^2/4a_i a_j\right)}}$$

one can combine the two terms of the previous equation such as:

$$\Delta\mathrm{G}_{el} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{ij}}$$

Replacing the interatomic distance r_{ij} by f_{ij} decreases the term $\frac{q_i q_j}{r_{ij}}\left(1-\frac{1}{\varepsilon}\right)$ as r_{ij} becomes smaller. The effective dielectric screening thus increases with the interatomic distance. It is worthy to note that the same dielectric term is used for the Born energy and the charge-charge interaction terms, which can be a problematic assumption.

The Langevin equation for orientation polarization:

In a water molecule, the negatively charged oxygen and the two positively charged hydrogens have different centre of charge, leading to a dipole moment. When an electrostatic field *E* is introduced in a water bulk, it leads to an average reorientation of the dipole of water molecules, leading to polarization of water. However, thermal motion induces random rotations of water molecules and counteracts polarization effects. For this reason, one needs to consider the free energy instead of the internal energy of the system to calculate the average polarization of water.

From basic electrostatics one knows:

$$\mathrm{U}(\alpha) = -\vec{\mu}.\vec{E} = -\mu E cos(\alpha)$$

where *U* is the electrostatic energy, $\vec{\mu}$ is the dipole of the molecule, $\vec{E}$ the local electrostatic field and α the angle between the two vectors. The Boltzmann distribution of water molecules, *N*, according to their orientation to the field is:

$$\mathrm{N}(\alpha) = Ae^{-\frac{U(\alpha)}{kT}}$$

where *A* is a normalisation constant. To get the average polarization, one integrates over the whole spherical coordinates:

$$\langle\mu\rangle = \frac{\int_0^\pi \mu(\alpha)e^{-\frac{U(\alpha)}{kT}}d\Omega}{\int_0^\pi e^{-\frac{U(\alpha)}{kT}}d\Omega}$$

Knowing that: $d\Omega = 2\pi\sin(\alpha)\mathrm{d}\alpha$, it gives:

$$\langle\mu\rangle = \frac{\int_0^{\pi} \mu\cos(\alpha) e^{\frac{\mu E cos(\alpha)}{kT}} 2\pi\sin(\alpha) d\alpha}{\int_0^{\pi} e^{\frac{\mu E cos(\alpha)}{kT}} 2\pi\sin(\alpha) d\alpha}$$

Using the substitutions: $y = \frac{\mu E}{kT}$ *and* $x = \cos(\alpha)$, the integral reduces to:

$$\langle\mu\rangle = \mu \frac{\int_0^1 x e^{bx} \mathrm{dx}}{\int_0^1 e^{bx} \mathrm{dx}} = \mu \left(\coth(b) - \frac{1}{b}\right) = \mu \mathrm{L(b)}$$

L is called the Langevin function. When $\mu E \ll kT$, one obtains the Langevin-Debye equation:

$$\langle\mu\rangle = \frac{\mu^2 E}{3kT}$$

Ratio of dual labelled hexamers:

The proportion of dual labelled hexamers, *P*, was calculated as:

$$P(\mathrm{p}_{488}, \mathrm{p}_{594}) = \sum_{i=1}^{5} \sum_{j=1}^{6-i} \binom{6}{i} \binom{6-i}{j} \mathrm{p}_{488}{}^{i} \mathrm{p}_{594}{}^{j}$$

where p_{488} and p_{594} are the degree of labelling of AF488 and AF594, respectively.

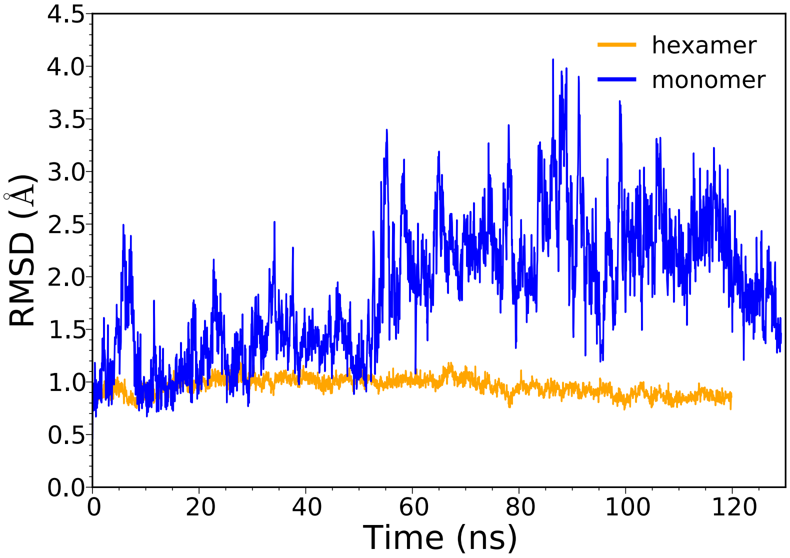


Figure A - 1: *Root mean square deviation along simulations of φ12 P4.*
Root mean square deviation from the X-crystal structure (RMSD) of the monomer (blue) and the hexamer (orange), along the molecular dynamics simulation. The RMSD of the monomer stabilises around 2.5 Å, a considerably larger value than that observed for the average monomer in the hexamer (1 Å). The simulation suggests that the monomer native state is stable in solution but slightly deformed (especially at the N-terminus) and more fluctuating relative to the monomer in the hexamer. The simulation of the monomer was performed as described for the hexamer, at the same pressure and temperature, but in a smaller water box containing 24970 TIP3P molecules.

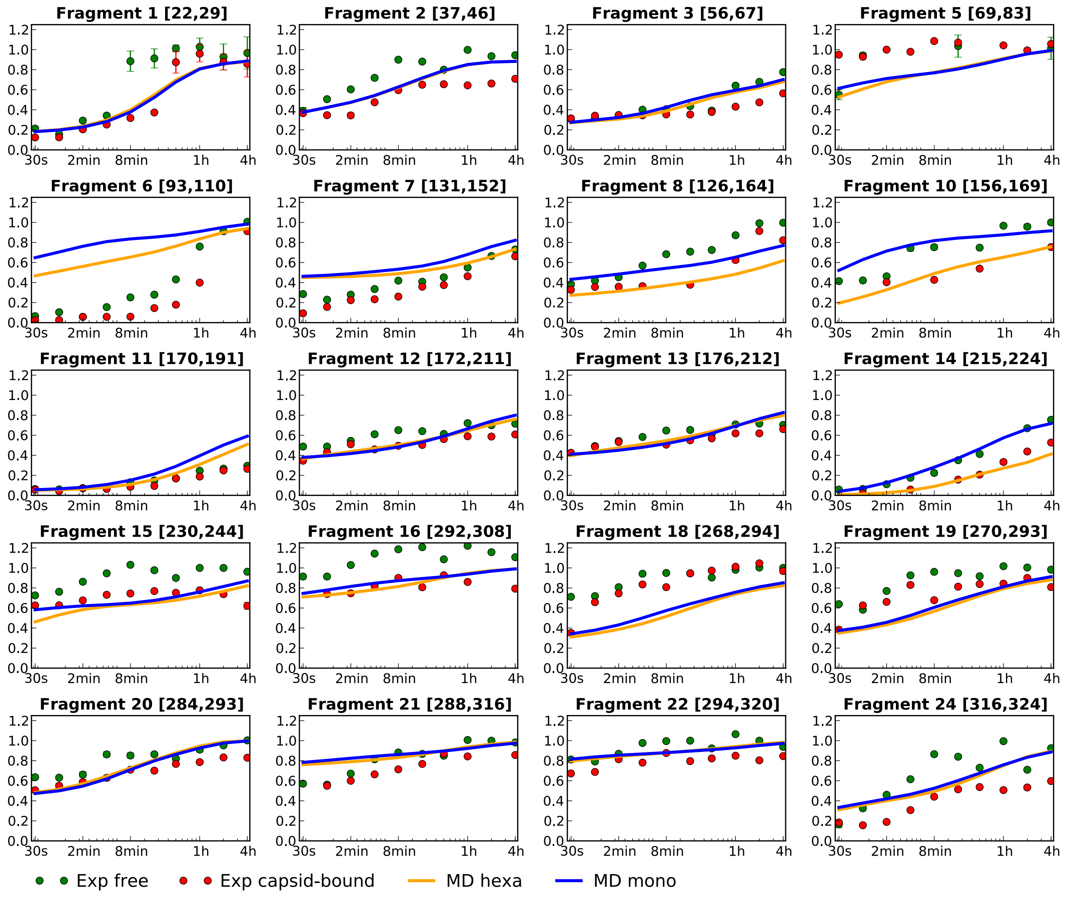


Figure A - 2: *Hydrogen exchange kinetics of all fragments of φ12 P4.*
Predicted deuterium fractions (averages over the MD simulation) for the monomer (blue line) and the hexamer (orange line). The experimental deuterium fractions of the free hexamer and the hexamer assembled with the procapsid are shown as green and red dots, respectively. Experimental error bars are shown when larger than the symbols.

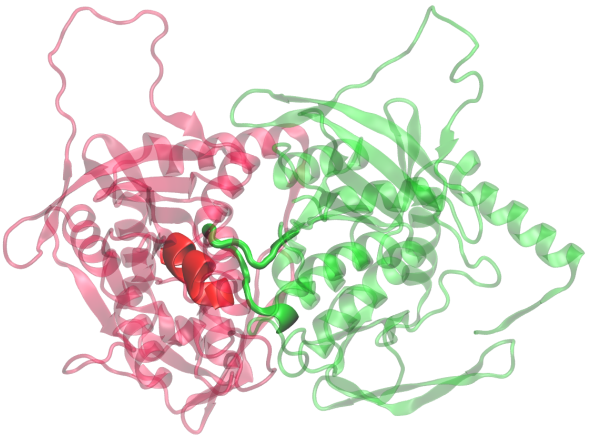


Figure A - 3: *Structure of fragments of φ12 P4 localised at the subunit interface.*
Cartoon representation of the interface between two neighbouring subunits of the hexamer. The fragments 14 and 10 are highlighted in red and green, respectively. For better clarity, the subunits are depicted up-side-down compare to Figures 1 and 4.

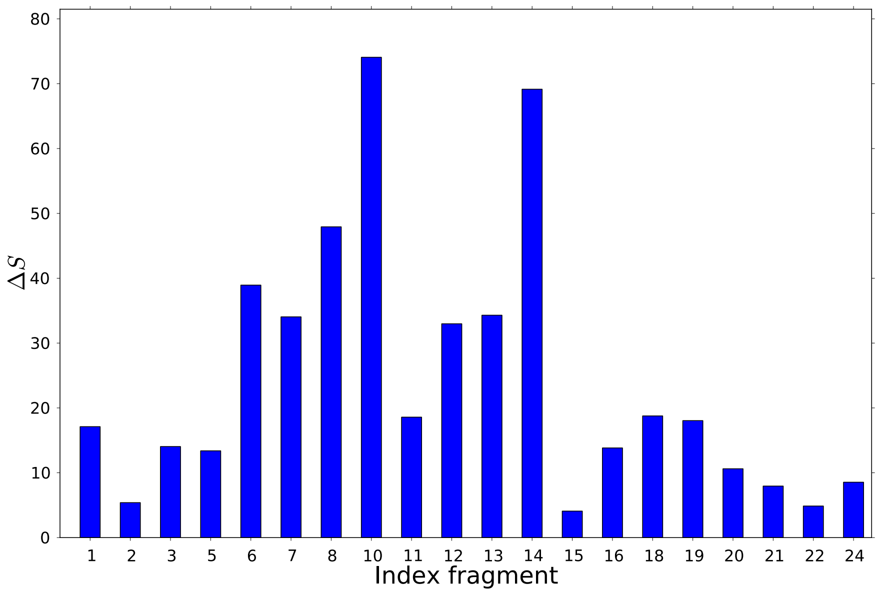


Figure A - 4: *Change of the solvent accessible surface area of each fragment between the monomer and hexamer in the crystal structure of φ12 P4.*

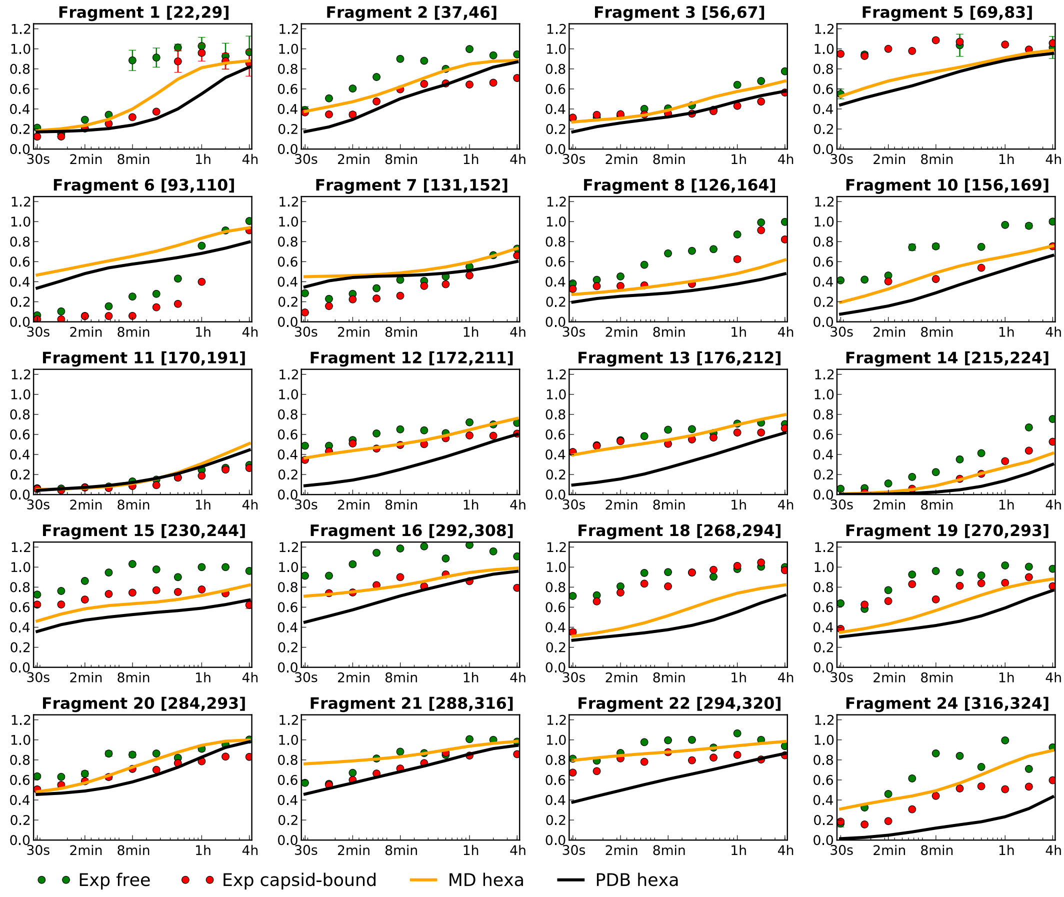


Figure A - 5: *Hydrogen exchange kinetics of all fragments of φ12 P4.*
Exchange kinetics of the hexamer with (orange line) or without (black line) dynamics predicted from the MD simulations. The experimental fraction of the free hexamer and the hexamer assembled with the procapsid are represented as green and red dots, respectively. Experimental error bars are shown when larger than the symbols.

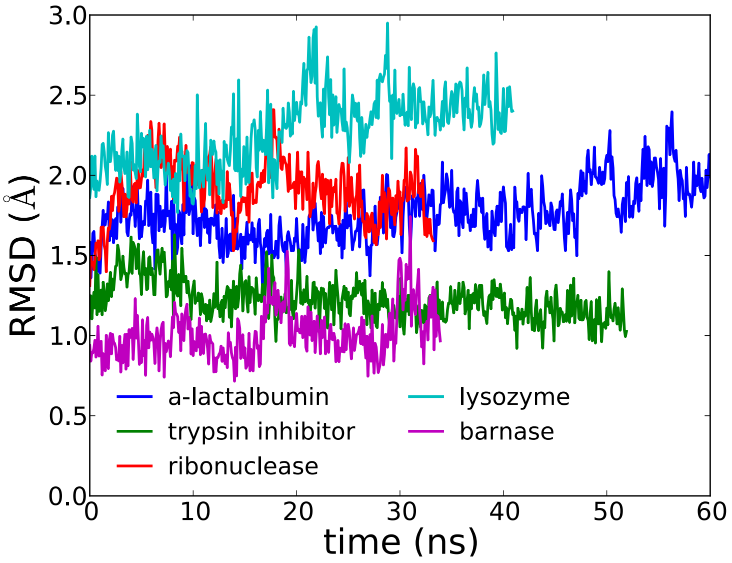


***Figure A - 6**: Root mean square deviation along simulations of φ8 P4.*
Root mean square deviation of a-lactalbumin, trypsin inhibitor, ribonuclease, lysozyme and barnase from their PDB structure along the MD simulations. The native state of each basin was sampled for at least 30 ns in CHARMM36 and explicit solvent. Proteins were relaxed for 10ns before equilibration (data not shown).

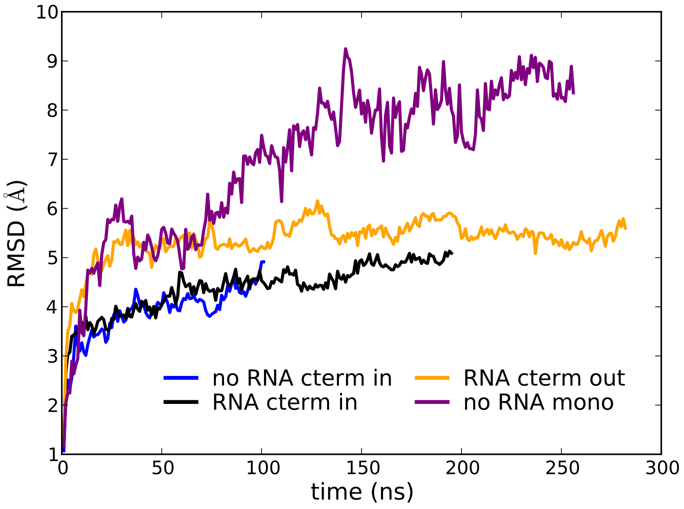


***Figure A - 7** Root mean square deviations of φ8 P4 simulations from their initial structures. The apo state, the monomeric state and the models with RNA and the C-terminus either inside or outside the pore are represented in blue, purple, black and orange, respectively. The significant increase of the monomer RMSD is mainly due to an unfolding of the C-terminus.*

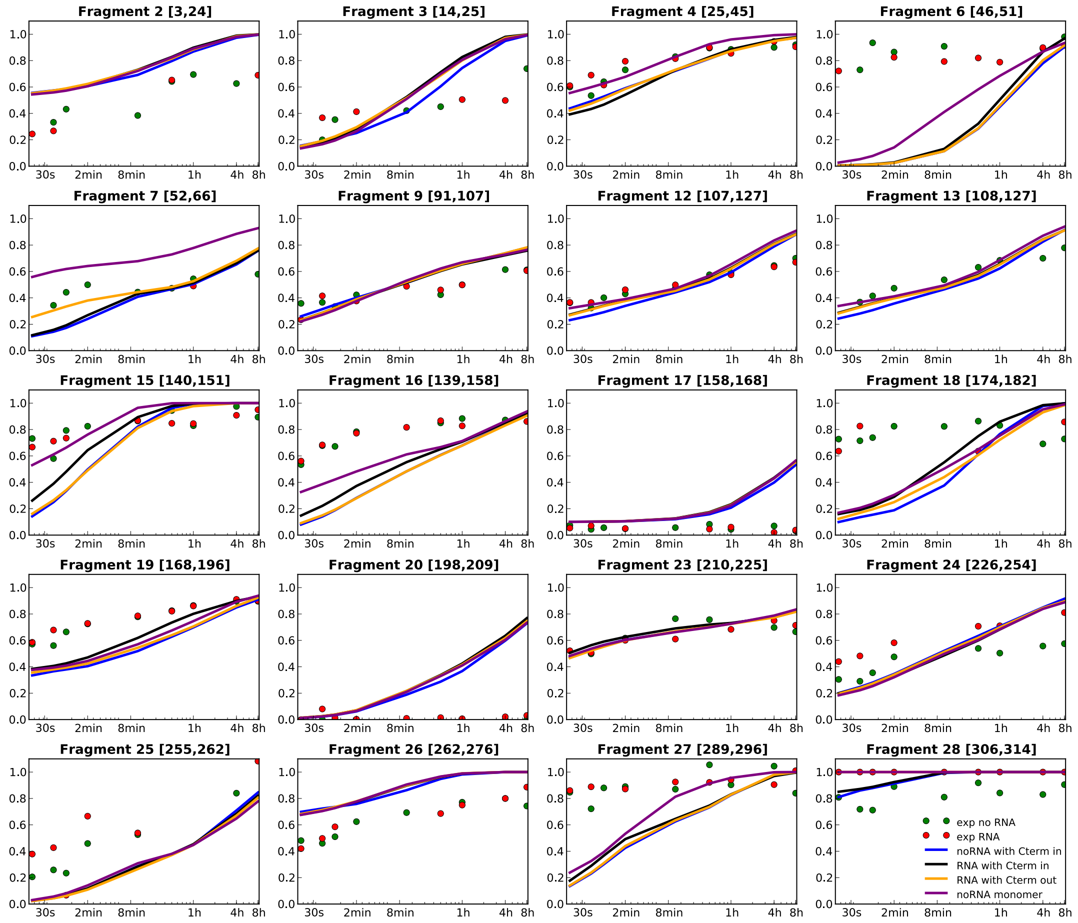


Figure A - 8: *Hydrogen exchange kinetics of all fragments of φ8 P4.*
Predicted deuterium fractions (average over the MD simulation) for φ8 P4 in the apo state (blue line), or in the monomeric state (purple line), or with RNA and the C-terminus inside the pore (black line), or with RNA and the C-terminus outside the pore (orange line). The experimental deuterium fractions of the helicase without and with RNA are shown as green and red dots, respectively.). Only the 20 non-redundant fragments with good experimental data quality were kept.

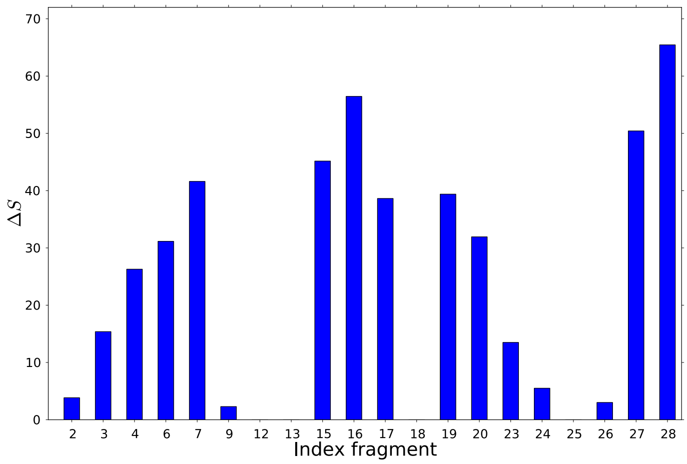


Figure A - 9: *Change of the solvent accessible surface area of each fragment between the monomer and hexamer in the crystal structure of φ8 P4.*

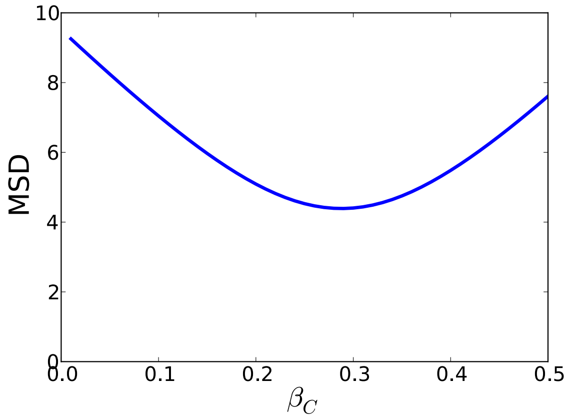
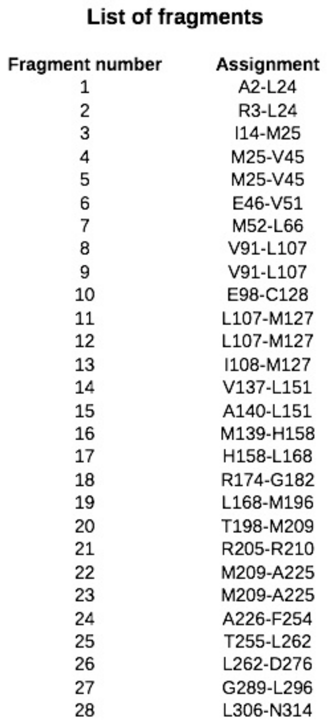


***Figure A - 10**: Fitting of the HDX model with only one parameter.*
For each protein, protection factors were calculated for varying value of B_c (with B_h fixed to 0) and averaged over the simulation in CHRMM36. The mean square deviation was averaged over the 5 proteins. With the optimal value of B_c=0.29, the overall agreement with experiment was as good as for the fitting with two parameters (MSD~4.3).

List of fragments

Fragment number	Assignment
1	H33-W29
2	L37-V46
3	A56-V67
4	A56-V70
5	Y69-V83
6	Q93-S110
7	K131-K152
8	V126-G164
9	V126-G164
10	V156-F169
11	N170-L191
12	F172-A211
13	I176-F212
14	L215-S224
15	I230-E244
16	S292-S308
16*	I230-L245
17	D239-V246
18	Q268-L294
19	L270-V293
20	L284-V293
21	Y288-I316
22	L294-E320
23	E290-S330
24	I316-V324

Table A - 1: *List of the fragments of φ12 P4 experimentally probed.*
The assignment of the fragment 16 reported in Ref (131) is indicated with an asterisk.



List of fragments

Fragment number	Assignment
1	A2-L24
2	R3-L24
3	I14-M25
4	M25-V45
5	M25-V45
6	E46-V51
7	M52-L66
8	V91-L107
9	V91-L107
10	E98-C128
11	L107-M127
12	L107-M127
13	I108-M127
14	V137-L151
15	A140-L151
16	M139-H158
17	H158-L168
18	R174-G182
19	L168-M196
20	T198-M209
21	R205-R210
22	M209-A225
23	M209-A225
24	A226-F254
25	T255-L262
26	L262-D276
27	G289-L296
28	L306-N314

Table A - 2: *List of the fragments of φ8 P4 experimentally probed.*
Fragments 1, 5, 8, 10, 11, 20, 21, 22 were ignored due to redundancy or poor quality of their experimental kinetics.

	Experimental data			Old assignment	New assignment		
Id	**m/z**	**z**	**exp-mass**	**sequence**	**sequence**	**mass**	**Δm**
1	870.3828	1	869.37498	22-29	22-29	869.36676	0.00822
2	1045.5247	1	1044.51688	37-46	36-45	1044.49749	0.01939
					37-46	1044.49749	0.01939
					38-47	1044.49749	0.01939
					98-106	1044.54913	0.03225
					128-138	1044.55635	0.03947
3	687.4032	2	1372.79076	56-67	56-67	1372.78265	0.00811
					57-68	1372.78265	0.00811
4	868.5086	2	1735.00156	56-70	56-70	1734.97805	0.02351
					173-187	1734.95627	0.04529
5	813.4265	2	1624.83736	69-83	69-83	1624.83078	0.00658
					168-181	1624.79189	0.04547
6	963.9651	2	1925.91456	93-110	93-110	1925.94174	0.02718
7	726.413	3	2176.21554	131-152	131-152	2176.19635	0.01919
8	785.6373	5	3923.1474	126-164	126-164	3923.12678	0.02062
9	981.7956	4	3923.15112	126-164	126-164	3923.12678	0.02434
10	801.3862	2	1600.75676	156-169	156-169	1600.75214	0.00462
					160-173	1600.74091	0.01585
					38-53	1600.78315	0.02639
11	842.1326	3	2523.37434	170-191	170-191	2523.36311	0.01123
12	1020.807	4	4079.19672	172-211	172-211	4079.16974	0.02698
					173-212	4079.16974	0.02698
13	938.525	4	3750.06872	176-212	176-212	3750.04745	0.02127
14	935.4548	1	934.44698	215-224	215-224	934.44296	0.00402
					117-125	934.40792	0.03906
					206-214	934.4872	0.04022
					290-297	934.48721	0.04023
15	808.4156	2	1614.81556	230-244	230-244	1614.81005	0.00551
					231-245	1614.81005	0.00551
					238-251	1614.85766	0.0421
16	864.9627	2	1727.90976	230-245	292-308	1727.91659	0.00683
					140-156	1727.92058	0.01082
					230-245	1727.89411	0.01565
17	930.5123	1	929.50448	239-246	239-246	929.50693	0.00245
					312-319	929.51818	0.0137
					81-89	929.48179	0.02269
					142-151	929.48177	0.02271
					305-313	929.52941	0.02493
18	756.905	4	3023.58872	268-294	268-294	3023.57881	0.00991
					257-283	3023.56756	0.02116
					258-284	3023.56756	0.02116
19	671.8521	4	2683.37712	270-293	269-292	2683.36776	0.00936
					270-293	2683.36776	0.00936
					271-294	2683.36776	0.00936
					272-295	2683.36776	0.00936
					42-67	2683.41802	0.0409
20	1120.5242	1	1119.51638	284-293	284-293	1119.51964	0.00326
					285-294	1119.51964	0.00326
					258-267	1119.50839	0.00799
					17-26	1119.4985	0.01788
					55-64	1119.55601	0.03963
					280-289	1119.55603	0.03965
					106-116	1119.5634	0.04702
21	1009.2095	3	3024.60504	288-316	288-316	3024.5992	0.00584
21					292-320	3024.62033	0.01529
22	710.6451	4	2838.54912	294-320	294-320	2838.51989	0.02923
22					295-321	2838.51989	0.02923
23	1112.3603	4	4445.40992	290-330	290-330	4445.39897	0.01095
24	1046.5303	1	1045.52248	316-324	315-323	1045.52913	0.00665
					316-324	1045.52913	0.00665
					317-325	1045.52913	0.00665
					263-271	1045.508	0.01448

Table A - 3: *Assignment for each fragment.*
In each column is indicated the index of the fragment, the experimental monoisotopic mass of the fragment, the number of charge, the previous assignment, the first residue of the new assignment, the last residue of the new assignment, the monoisotopic mass of the new fragment, the absolute difference of mass between the experimental and predicted mass of the new fragment, respectively.