Reconciling NMR Structures of the HIV-1 Nucleocapsid Protein (NCp7) using Extensive Polarizable Force Field Free-Energy Simulations

Léa El Khoury,[†,‡,¶,△] Frédéric Célerse,[†,§,△] Louis Lagardère,[‖,⊥] Luc-Henri Jolly,[‖] Etienne Derat,[§] Zeina Hobaika,[‡] Richard Maroun,[‡] Pengyu Ren,[#] Serge Bouaziz,[@] Nohad Gresh,[†] and Jean-Philip Piquemal[*,†,#]

†*Sorbonne Université, LCT, UMR 7616 CNRS, F-75005, Paris, France*

‡*UR EGP, Centre d'Analyses et de Recherche, Faculté des Sciences, Université Saint-Joseph de Beyrouth, Beirut, Lebanon*

¶*Current address: Department of Pharmaceutical Sciences, University of California, Irvine, California 92697, United States*

§*Sorbonne Université, CNRS, IPCM, Paris, France.*

‖*Sorbonne Université, IP2CT, FR2622 CNRS, F-75005, Paris, France*

⊥*Sorbonne Université, ISCD, F-75005, Paris, France*

#*Department of Biomedical Engineering, The University of Texas at Austin, USA*

@*Université Paris Descartes, CNRS, Laboratoire de Cristallographie et RMN Biologiques, Paris, France*

△*Contributed equally to this work*

E-mail: jean-philip.piquemal@sorbonne-universite.fr

Abstract

Using polarizable (AMOEBA) and non-polarizable (CHARMM) force fields, we compare the relative free energy stability of two extreme conformations of the HIV-1 NCp7 nucleocapsid that had been previously experimentally advocated to prevail in solution. Using accelerated sampling techniques, we show that they differ in stability by no more than 0.75-1.9 kcal/mol depending on the reference protein sequence. While the extended form appears to be the most probable structure, both forms should thus coexist in water explaining the differing NMR findings.

Introduction

The treatment of AIDS, caused by the Human Immunodeficiency Virus type-1 (HIV-1), has significantly improved by the association of reverse transcriptase and protease inhibitors.[1–3] Toward the development of novel antiviral agents, recalling that multitherapy still generates drug resistance, an alternative strategy is the targeting of viral proteins or nucleic acid sequences, the biological functions of which are ensured by domains which could not be mutated without a complete loss of activity.[4] The HIV-1 nucleocapsid protein NCp7, which plays a critical role at different steps of the retrovirus life cycle, is such a target.[5–7] NCp7 possesses a three-dimensional structure centered around two highly conserved zinc fingers. Site-directed mutagenesis have shown that any modification of this structure leads to a complete loss of HIV-1 infectivity.[8–10]

HIV-1 has several viral protein-encoding genes. The Gag gene is the precursor of four structural proteins of the virions.[12] Among these is p15 which is cleaved by the protease (PR) to give two smaller products: p6 and NCp7.[13] The latter depicted on Figure 1 is a small single-stranded nucleic binding protein of 55 amino acid residues. NCp7 was shown to play a crucial role not only in viral replication, but also during various steps of the virus life cycle. It is involved in viral RNA (RiboNucleic Acid) dimerization and packaging[14] as well as in the initiation of reverse transcription upon interacting with TAR (Trans-Activation

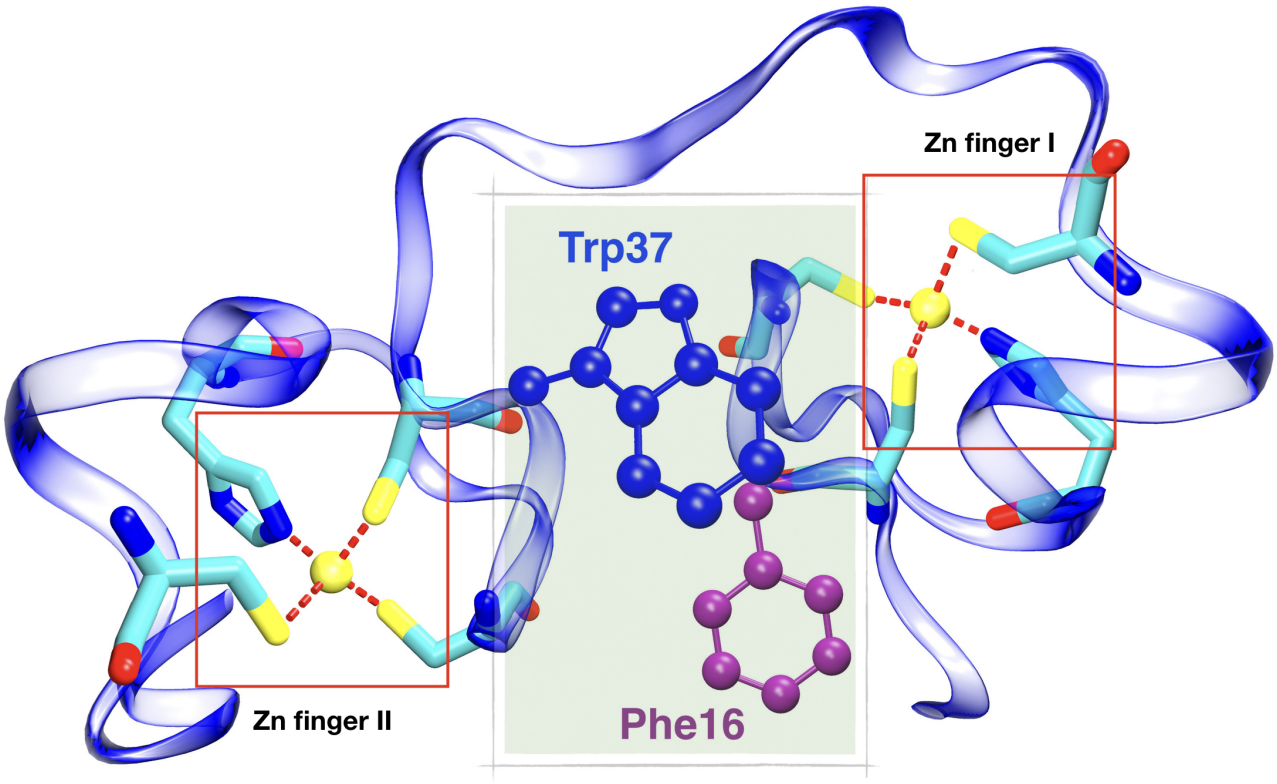


Figure 1: Representation of the NCp7 protein highlighting the two zinc finger domains and the stacking between residues Phe16 and Trp37. Secondary structure in blue is depicted in newribbons. The two Zn(II) are represented by yellow spheres with their respective amino–acids. The structure was extracted from the NMR structure of NCp7 determined in the Roques Laboratory, PDB code: 1ESK.[11]

Response element) and PBS (Primer Binding Site) sequences of the viral DNA.[15,16] Moreover, it protects the retroviral genome against cellular nucleases[17] and interacts with viral proteins such as Vpr[18] (Viral Protein R) and integrase.[19] Most of these functions are related to its higher affinity for single-stranded nucleic acids and to its structural features. In fact, NCp7 contains flexible N-terminal and C-terminal domains surrounding two highly conserved zinc finger motives CX2CX4HX4C, where X represents any amino acid residue. Each zinc finger motif is able to tightly coordinate one zinc cation with a very high affinity (dissociation constant around 10^{-13}). Each Zn^{2+} cation is tetrahedrally bound by the Cys thiolate groups and the His sp nitrogen.[20–22] This drives the transition from random to domain-ordered

conformations around each cation, which could be essential for biological activity.

Site-directed mutagenesis showed two aromatic residues, Phe16 on the first Zn-finger, and Trp37 on the second, to be essential for the activity of NCp7.[8,10,23–27] NMR chemical shift analyses and Nuclear Overhauser effects (NOEs) in the Roques Laboratory had indicated a spatial proximity between the two zinc fingers.[11,23,28–32 32] and led to the conclusion that Phe16 and Trp37 were spatially close (see Figure 1[32]). However, an opposite conclusion was obtained by the Summers Laboratory.[33]

Owing to the key involvement of both residues in the NCp7-RNA complex domain,[32] we deemed it essential to try and resolve this apparent structural difference. For that purpose, we undertook a detailed computational study on the fully solvated protein (18515 atoms) using both the non-polarizable CHARMM[34,35] and the higher accuracy[36,37] new generation polarizable AMOEBA force field.[38–40] To better focus on the Phe16-Trp37 interactions, enhanced sampling methods were considered with both Steered Molecular Dynamics (SMD) and Umbrella Sampling (US) methods.

Methods

PDB entries

The initial structures used in this work are based on the available NMR solution structures of $NCp7_{12-53}$ (PDB ID: 1ESK).[11] All simulations thus start from the stacked A form. All Zn-binding cysteine residues (Cys15, Cys18, Cys28, Cys36, Cys39, and Cys49) were considered to be in the anionic form and both Zn-binding histidines (His23 and His44) were neutral. The Tinker 8 package was used to prepare the initial system and to analyze the sequences[41]

Conventional MD simulations

All conventional MD (cMD) simulations (100 ns) were performed using the massively parallel software Tinker-HP[42] that allows for large-scale simulations using all types of force

fields. They were carried out with the polarizable AMOEBA force field[39,43–45] and the non-polarizable CHARMM22/CMAP force field.[34,35,46] The minimum distance from NCp7 to the nearest box edge was set to 10 Å. Each simulation cell has 60 Å each side, and contains a total of 18515 atoms. Chlorine ions were added at the box limits to neutralize the total charge of the system. Before performing cMD simulations, the energy of each starting system was minimized then heated up gradually to 298K. For every 20 K interval, 100 ps of NPT ensemble were generated. The systems were then equilibrated for 250 ps under the NPT ensemble. During these steps, only water molecules were relaxed. Subsequently, energy minimization of the whole system was done with the steepest descent algorithm. Production cMD simulations was then performed using the RESPA multitimestep integrator with an inner time step of 0.25 fs and an outer time step of 2 fs. Temperature and pressure were kept constant using Bussi's thermostat[47] coupled to Berendsen's barostat.[48] The Smooth Particle Mesh Ewald (SPME) method was employed to treat the long-range electrostatic interactions with a grid size of 60×60×60 Å. The real space cutoff was set to 7 Å, while the van der Waals (vdW) cutoff was set to 9 Å with a long-tail correction. A convergence criterion of 10^{-5} kcal/mol coupled to the Always Stable Predictor Corrector (ASPC)[49]was used for the polarization contribution.

To prevent unphysical Zn(II) decoordination when using non-polarizable force field, hard restraints (500 kcal/mol.Å^2) were added between cysteines (Cys15, Cys18, Cys28, Cys36, Cys39, and Cys49) / histidines (His23 and His44) and their bound zinc cations. For consistency they were retained for the AMOEBA simulations despite the better stability of the simulations.[44] Such constraints do not affect the studied mechanisms but should prevent structural denaturation during the cMD simulations. At their outcome, two distinct conformations were characterized: with a partial stacking between Phe16 and Trp37, denoted as A; and unstacked, denoted as B. Free energy computations were subsequently undertaken to study the respective stability of the two conformations.

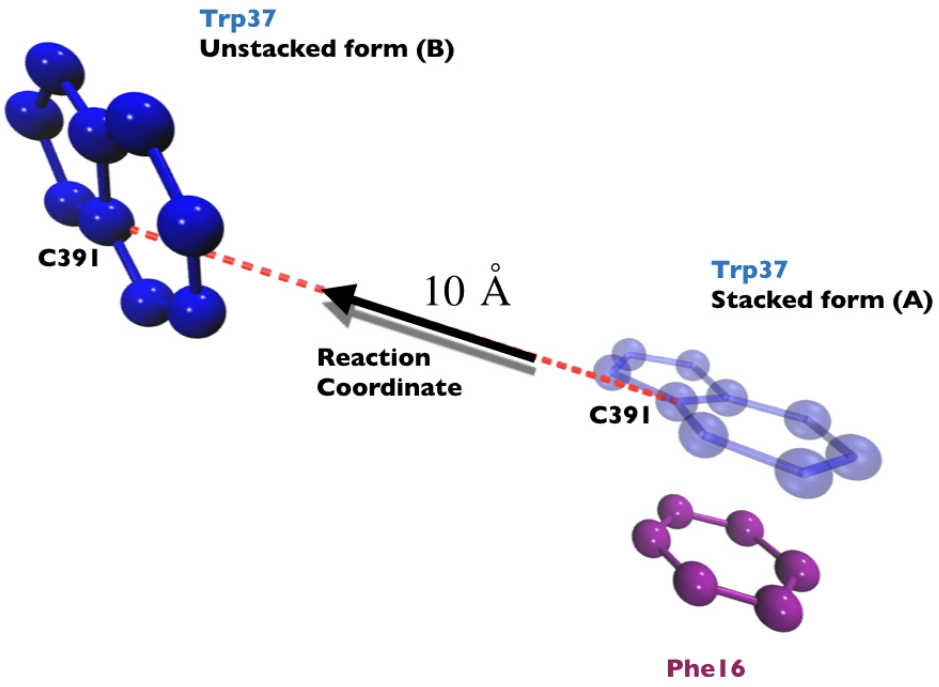


Figure 2: Representation of the chosen reaction coordinate (RC) for all SMD simulations. Both structures (A) and (B) obtained from cMD simulations have been aligned along the Phe16 residue, and the RC was then chosen as the distance between the C391(Trp37) from the stacked form and the C391(Trp37) from the unstacked form.

Steered MD simulations

SMD[50,51] is a simple free energy method which enhances the sampling of rare events, namely those occurring in a microsecond or millisecond frame. It can drive a conformational transition along a specific reaction coordinate (RC) in only a few nanosecond simulation. All SMD computations were carried out using Tinker-HP[42,52] using the NPT ensemble, a constant temperature of 298 K and a constant pressure of 1 bar enforced by the Bussi thermostat coupled to the Berendsen barostat. Periodic boundary conditions (PBC) were applied with use of the SPME method. A grid size of 60×60×60 Å, with an Ewald cutoff of 7 Å and a vdW cutoff of 9 Å and a long-tail correction were used. The AMOEBA force field and a Velocity Verlet integrator with a timestep of 1 fs were used for all simulations.

Starting points for each SMD simulation were taken from a snapshot extracted from the preceding cMD simulation. The chosen reaction coordinate (RC) should enable to pass from structure A to an unstacked form, B. It was illustrated on Figure 2. During the SMD simulations the aromatic carbons of Phe16 were kept fixed to prevent fictitious motions of the protein, and a harmonic potential was applied on the C391 of Trp37:

$$h(r, \lambda(v,t), t) = \frac{k}{2}(\epsilon(r) - \lambda(v,t))^2 \quad (1)$$

with k the spring constant chosen equal to 7.00 kcal/mol.Å^2, $\epsilon(r)$ the RC related to the steered atom position r at time t, and $\lambda(v,t)$ an external parameter corresponding to the external perturbation undergone by C391(Trp37) such as $\lambda(v,t) = \lambda_0 + \text{v}\times t$ with v the SMD velocity applied on C391(Trp37) chosen to be equal to 0.01 Å/ps and t the time of the simulation. A second spring k_2 equal to 7.00 kcal/mol.Å^2 (transverse to k_1) was chosen to avoid large fluctuations far from the RC. Avoiding large fluctuations far from the RC enables us to consider free energy profiles as the Potential of Mean Force (PMF) of the system along the RC according to the stiff spring approximation. 20 independent SMD trajectories were performed and each was randomized by a different initial set of velocities. Each simulation was computed during 1 ns, corresponding to a total simulation time of 20 ns. The pulling work was then calculated for each trajectory and the free energy difference between the initial and the final states was estimated by using the Jarzynski Equality (JE). Energy results with PMF are discussed in the results part of this study. Figures related to the SMD trajectories, namely force profiles in function of the simulation time and all the individual pulling works in function of the end-to-end distance are available in the Supplementary Information (SI).

Umbrella Sampling

The SMD methodology relates to the convergence of the exponential average of JE, and is strongly dominated by rare trajectories characterized by low pulling work values. Therefore

it is impossible to be sure that we obtain a sufficient number of trajectories to claim that the SMD results are converged. Keeping this reservation in mind, we decided to complete our SMD interpretations and provide more accurate insight in terms of free energy difference by considering the Umbrella Sampling (US) method.[53] The RC was chosen to be the same as the SMD one and was divided into 20 windows, ranging from 3 to 13 Å, successive windows being separated by a width of 0.5 Å to ensure for an efficient sampling along the RC. Figures related to the US convergence and showing all the umbrellas from each window are available in SI. In each window, a harmonic restrain of 10 kcal/mol Å^2 was applied, according to the US methodology. Both polarizable AMOEBA[39,43–45] and non-polarizable CHARMM22/CMAP force-fields[34,35,46] were considered.

For each window and with each force-field, the initial simulation parameters were the same as for the SMD. The only difference is that the Phe16 aromatic carbons were released and free to move while they were kept fixed in the previous SMD simulations. This provides more flexibility to the system during the simulations. Each window was computed during 3 ns, which corresponds to a simulation time of 60 ns for each force-field, whence a total simulation time of 120 ns. The first 100 ps of each window were not taken into account to reconstruct the PMF as they correspond to the equilibrium step of the system within its window. Finally, the free energy profile was reconstructed by combining all the windows using the Weighted Histogram Analysis Method (WHAM)[54] and the error in the free energy barrier was estimated using the Monte Carlo Bootstrap Analysis method directly implemented in Alan Grossfield's WHAM implementation.[54] While our SMD simulations took only 20 ns in AMOEBA, US needed a more important simulation time, three times larger. This confers robustness in terms of quantitative free energy analysis and could be useful to check the SMD convergence.

Results and discussions

1. Free NCp7 model: a conventional Molecular Dynamics approach in water

This section reports the main results from cMD upon preparing the accelerated sampling simulations. We studied the conformational behavior of NCp7 in solution, particularly the stability of the stacking occurring between Phe16 and Trp37 and its impact on the whole protein conformation. We started from the NMR structure resolved in the Roques Laboratory, where this stacking could be observed (cf Figure 1). In the context of both AMOEBA and CHARMM, it was lost after the first 5 ns. The distance between Phe16 and Trp37 finally fluctuated between 8 and 12 Å.

At this stage these results suggest that NCp7 does not have a single conformation, some of the most stable ones being found after 50 ns simulation. In viral cells, NCp7 is bound by other viral elements (TAR, PBS sequence, Integrase, Vpr), any of which could chaperon an extended-to-folded conformation change. Experimentally, NCp7 in its complex with viral DNA (vDNA) was found to adopt a folded conformation with stacked Phe16 and Trp37 rings[55] Similar observations were reported with the SL2 or SL3 loops of viral RNA.[56,57] In this connection, preliminary (40 ns) AMOEBA simulations of a vDNA-NCp7 complex showed such a conformation to remain stable during the cMD procedure.

The flexibility of unchaperoned NCp7, put forth by NMR and by the present calculations, would enable it to evade simultaneous complexation of both fingers by inhibitors. By contrast, the stabilization of a closed NCp7 conformation upon vDNA binding could have an additional implication. Thus the search for novel anti-NCp7 inhibitors could target a structured NCp7-vDNA complex with recognition sites on both fingers as well as on v-DNA rather than on the sole C-terminal Zn-finger domain as reported previously.[58] It is worth recalling in this connection that clinically active anti-HIV-1 integrase (IN) inhibitors do form a ternary complex with vDNA-bound IN rather than with IN or vDNA alone.[59]

2. Reconciling NMR structures using accelerated sampling Molecular Dynamics methods

A first study of the stacking free energy between Phe16-Trp37 using Steered Molecular Dynamics

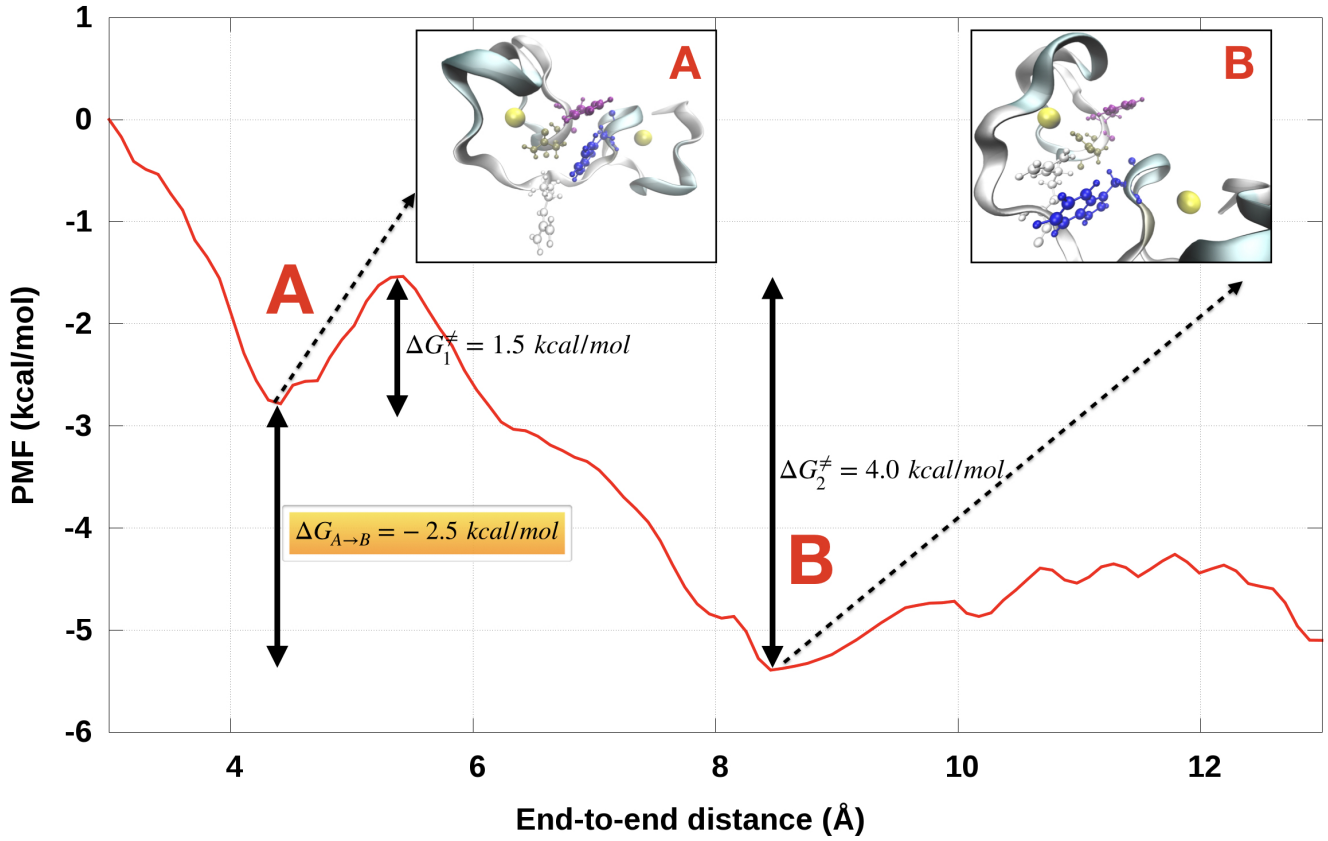


Figure 3: Free energy landscape along the C72(Phe16) – C391(Trp37) RC obtained with Steered Molecular Dynamics with the polarizable AMOEBA force field. ΔG, $\Delta G_1^{\neq}$ and $\Delta G_2^{\neq}$ are in kcal/mol. A represents the stacked form of NCp7 and B its unstacked form.

Since stacking was only observed during the first 5 ns in our cMD simulation, and since it governs the proximity between the two Zn-fingers, it is essential to monitor its evolution in the course of a long-time simulation, and the magnitude of the free energy barrier between the stacked (A) and the unstacked (B) conformers. This is enabled by the SMD method.[50,51] SMD was recently implemented in Tinker-HP by Célerse et al.,[52] and has the

Table 1: Free energy difference calculated with the Jarzynski Equality (JE) and associated Standard Deviation Work (SDW) for a set of 5/10/15/20 SMD trajectories using the AMOEBA force field. All values are calculated between states A and B obtained on Figure 3.

# of trajectories	Jarzynski's Equality (in kcal/mol)	Standard Deviation Work (in kcal/mol)
5	-3.40	4.44
10	-3.13	3.19
15	-2.69	3.14
20	-2.50	3.11

asset of smaller simulation time costs. In the context of the polarizable AMOEBA potential, upon choosing a straight line as RC (see Figure 2), it enables to induce rare events in a nanosecond MD by applying an external harmonic potential on a set of atoms to pull them along a specific spatial direction. Thus, the pulling work (cf Figure 2 in SI) is calculated and the free energy contribution necessary to pass from the initial to the final state is determined by integrating the work along the simulation time (cf Figure 3). Together with the choice of a suitable RC, the external perturbation applied on the system is the other critical input data. The empirically preset SMD velocity has to be chosen as slow as possible but has to remain in an accessible simulation time. If it is chosen too fast, non-equilibrium effects may contaminate the pulling work and therefore dramatically bias the calculated final free energy. Nevertheless, there is a means to overcome this theoretical problem upon resorting to the JE.[60,61] However, the difficulty to converge the exponential average requires to use an important number of trajectories to ensure for accuracy, whence a significant increase in terms of simulation time and tasking of computer resources.

We have thus considered a set of 20 independent SMD trajectories with a SMD velocity set to 0.01 Å/ps to induce the transition from A to B. The results are reported in Table 1. PMF provided a first free energy minimum of -2.5 kcal/mol followed by a smooth barrier of +1.5 kcal/mol to partially unstack Phe16 and Trp37 (See Figure 3). This is followed by a continuous energy decrease to reach state B, which in this context, and using AMOEBA, lies

2.5 kcal/mol lower than state A. Furthermore Figure 2 in SI shows that 8 trajectories over 20 have a negative pulling work: this implies that Phe16 and Trp37 could be unstacked without the need for additional external pulling work. Accordingly, NCp7 would display a modest preference for the unstacked form (B). This appears consistent with the previous cMD results showing the stacking to be lost after 5 ns of simulation and not recovered after 100 ns. As depicted on Figure 3 it could be observed that Arg32 does not initially interact with Trp37. It translates and rotates during the SMD simulation to finally interact with it (see movie in Supplementary Information, Arg32 being in white). Asn17 which first stabilizes the stacking by interacting with Trp37, rotates to interact with Phe16. A Phe16/Asn17/Arg32/Trp37 arrangement (see movie in SI) is eventually formed. This arrangement enhances the stabilization of B due to dispersion and long-range polarization effects occurring between these four residues, and could be accountable for the preferential stabilization of B compared to A.

However due to the slow JE convergence, as depicted in Table 1, care must be exercised against a strict quantitative interpretation of such a free energy difference. This has led us to calculate the Standard Deviation Work (SDW) on the 20 independent SMD trajectories to gain insight on the convergence of the JE result, considering that only small SDW values could ensure for reversibility and acceptable convergence of the free energy difference. As depicted in Table 1 SDW decreases upon passing from 5 to 20 trajectories, but at an increasingly slower rate. For 20 trajectories SDW still retains a rather high value of 3.1 kcal/mol. This implies that convergence was not reached and that more than 20 trajectories should be considered in this case at the cost of much more important simulation times.

SMD therefore shows A and B to be separated by a free energy difference of 2.5 kcal/mol favoring B. However the high calculated SDW implies that such a difference could actually be lower. Indeed, JE is strongly dependent on the presence of low pulling work trajectories. Thus, the quality of the simulation could be improved by increasing the number of such trajectories, but this would come at the price of a very significantly increased overall simulation

time while the optimal number of these trajectories remains yet to be determined. Therefore, we decided to consider an alternative approach toward a more accurate evaluation of this energy difference since it would not be more than three times as costly as the present SMD calculations.

A more accurate and quantitative procedure to evaluate the free energy difference between A and B forms: Umbrella Sampling

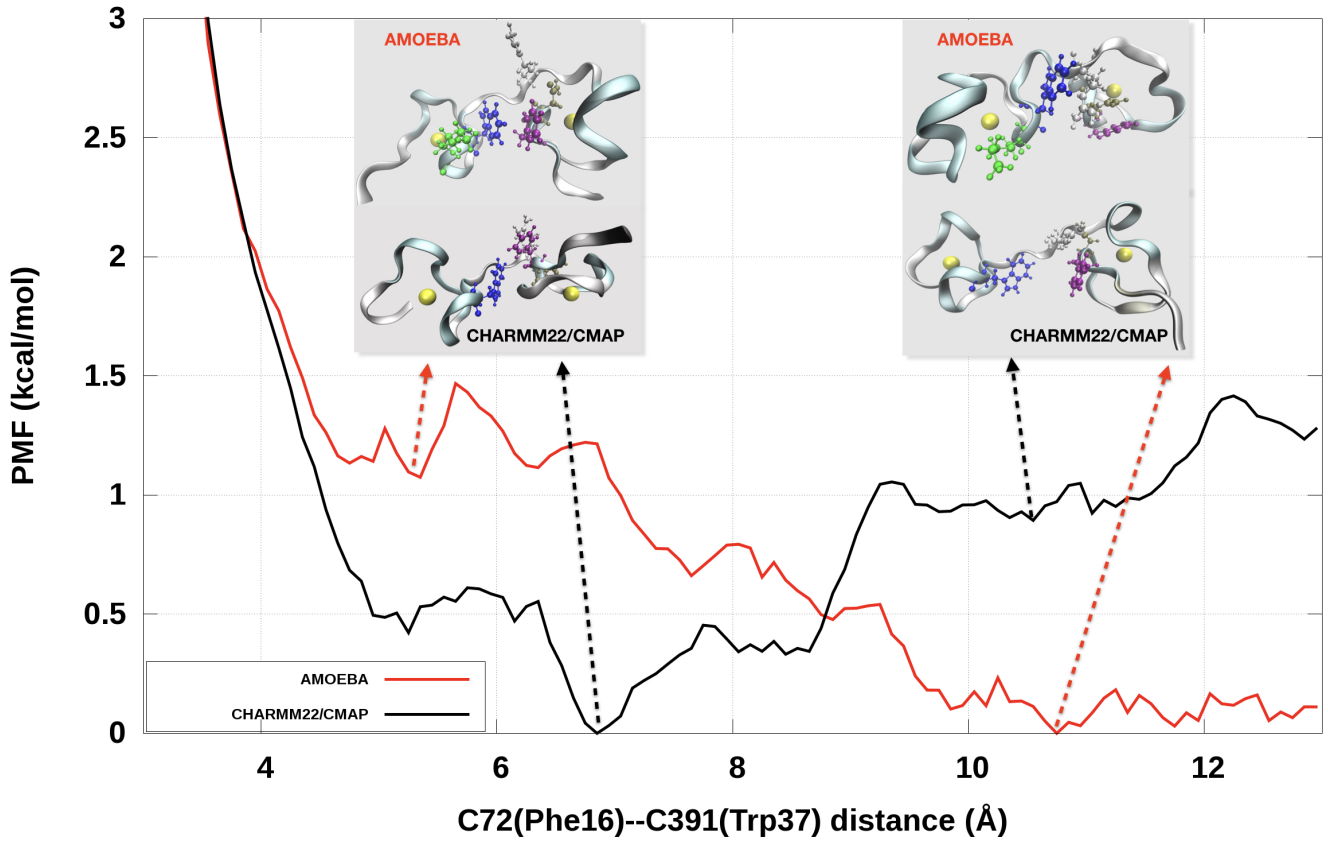


Figure 4: Free energy landscape along the C72(Phe16) – C391(Trp37) reaction coordinate obtained with Umbrella Sampling in CHARMM22/CMAP and AMOEBA. For each PMF, the minima corresponding to the stacked and unstacked forms are assigned with red (AMOEBA) and black (CHARMM) arrows. The corresponding structures are depicted in new ribbon style, zinc cations are in yellow, Phe16 is in purple, Trp37 in dark blue, Arg32 in grey and Asn17 in khaki, and, in the AMOEBA minima, Lys38 in green.

To overcome the slow SMD convergence we decided to resort to the Umbrella Sampling (US) procedure, known to be more robust. It is widely used for enhanced sampling simu-

lations[53] as it can ensure a better convergence to the right free energy value. It is however more expensive in terms of computational time compared to SMD but enables to accurately reproduce a PMF along one or more chosen RC(s). With both the AMOEBA[39,44,45] and non-polarizable CHARMM22/CMAP[34,35,46] force fields, we have recomputed the PMF with the same RC as for SMD using the procedure described in the Methods section. The results are reported on Figure 4.

Both force fields favor a non-stacked conformation. CHARMM displays a smooth free energy well at 6.5 Å with an unstacked conformation stabilized by a cation-π interaction between Phe16 and Arg32. This interaction is not observed in AMOEBA with which Arg32 interacts with Trp37 in the AMOEBA B minimum at 10.5 Å. State A with AMOEBA has a RC coordinate of 5.5 Å which is strongly similar to the corresponding SMD one (found at 4.5-5 Å). There is however a very small barrier ($<$ 0.5 kcal/mol) separating it from final state B, while by contrast it was separated by 1.5 kcal/mol in the SMD simulation. In fact, Figure 3 in SI shows a 1.5 kcal/mole barrier to be present in AMOEBA US simulations with 1 ns/window, but this magnitude is reduced upon increasing the simulation time, to 2 or 3 ns/window. Therefore it could be concluded that the energy barrier found in SMD could be due to limited sampling along the 4.5 - 8 Å interval. It also demonstrates that the SMD computations well localized both A and B minima but provided a slightly overestimated free energy barrier. It is consistent with the use of irreversible SMD simulations which are known to overestimate such quantities.[52] In AMOEBA state A, Lys38 has a cation-π interaction with Trp37. This is turn enhances the Trp37-Phe16 interaction. Such a Lys38-Trp37-Phe16 complex was not observed with CHARMM. AMOEBA state B, found with an RC of 10.5 Å, is 1.5$\pm$0.5 kcal/mol more stable than state A. It is stabilized by the same Phe16/Asn17/Arg32/Trp37 arrangement as in the previous SMD simulations. Such an arrangement is not found with CHARMM. A detailed study of the physical origin of the global differences (i.e. including also the solvation effect) between AMOEBA and CHARMM can be found in SI (see Table 1 and 2). Clearly, the A form is mainly favored by point-charge

electrostatics in CHARMM whereas the physics at play is different in AMOEBA. Indeed, in this case, The B form prevails through a cumulative stabilization of multipolar electrostatics and polarization contributions. It is important to note that beside the many-body effects included in the polarization energy present in both the protein and polarizable solvent, it is striking that the electrostatics contribution that includes permanent charge, dipole and quadrupole tends to give a very different picture for the global interaction compared to CHARMM.

According to Figure 4, the free energy difference between A and B is in between only 1 and 2 kcal/mol with both force fields, fully consistent with the AMOEBA SMD results giving a corresponding 1.2±0.5 kcal/mol difference (see the Methods section for the error bar description). However it has to be stressed that while AMOEBA and CHARMM agree about the magnitude of the free energy difference between A and B, they do not give the same structural interpretation. Indeed CHARMM favors A with stacked Phe16 and Trp37 residues, while AMOEBA shows A to be a higher free energy local minimum (see Figure 3). The AMOEBA result clearly validates Summers'[33] conclusion that a Phe16 and Trp37 proximity could exist as some NMR interknuckle interactions were detected but remained only transitory (lower population) whereas the linear B form dominates (higher population). These free energy results can be used to compute the ratio of the probability of the two states with the following formula:

$$\frac{p(A)}{p(B)} = \exp(-\beta \Delta G)$$

with $\beta = (k_b T)^{-1}$ and $\Delta G = G_A - G_B$ the free energy difference between state A and state B. In our case, it gives for AMOEBA a ratio of $\frac{p(A)}{p(B)}$ of 13.4% for a free energy difference of 1.2 kcal/mol, in the range of the findings of Casas-Finet et al. proposing a 12-17% ratio.[62] Our computations show that overall, switching from A to B appears energetically easy for free NCp7 in water on a biologically short time scale. NMR studies are based on observations

averaged in timescales of about a millisecond. Our results show that since both A and B are energetically close, they could both coexist in solution and be detected by NMR. Another possible origin of the experimental discrepancies is the fact that the structures used by the Roques' and Summers' groups have slightly different sequences, Summers' having 13 additional amino acids (11 on the N-terminal portion and 2 on the C-terminal part, including two lysines and two arginines). To ensure that these differences do not impact strongly our final result, we performed an additional US simulation using both AMOEBA and CHARMM on Summers' longer sequence (PDB: 1MFS) with the same protocol. The results depicted in Figure 4 and 5 in SI show that the same A and B minima are found for both force fields. AMOEBA exhibits a slightly higher 1.4±0.5 kcal/mol free energy barrier compared to the shorter sequence confirming the stabilization of an extended B form. CHARMM switches from a A form to a B one with a free energy difference of 1.75±0.5 kcal/mol. If we recompute the previously discussed ratio of the probability of the two states it decreases for AMOEBA from 13.4 % to 9.6 % as the free energy difference increases by 0.2 kcal/mol. For CHARMM, a barrier of 1.75 kcal/mol leads to a 28% ratio which appears less in line with experiment confirming the tendency for CHARMM to over-stabilize the folded conformation. Possibly, such an effect is linked to the presence of up to four cationic residues in the largest 1MFS sequence that have a direct impact on solvation through electrostatics and polarization. Stronger protein-solvent interactions could be the reason for the global prevalence of the B form with AMOEBA. Indeed, as we discussed previously, the AMOEBA force field proposes a different physical picture compared to CHARMM and embodies a net polarization preference for the elongated forms that can be tracked back to the presence of a stronger polarizable solvent environment able to better stabilize the more solvent accessible B forms. In this connection, it is important to point out that any effect modifying the solvation process could be important. Indeed, small differences in the experimental protocols, especially in the choice of the ionic strength, could also be partially accountable for the different experimental results and could also affect the computations. For example, Roques et al. considered a pH equal to

5.5 with 1M of NaOD or DCl with $ZnCl_2$ in excess with respect of the concentration of zinc fingers.[31] However, Summers et al. considered different experimental conditions with a pH equal to 6.5 and 25 mM of NaCl with 0.1 mM of $ZnCl_2$.[33] Different experimental conditions could easily shift the balance between A or B. The different experimental results could be then due to a different combination of protein length and ionic strength. To conclude, our theoretical results show that: i) A and B differ in stability by no more than 0.75-1.9 kcal/mol depending on the protein sequence; ii) their relative stability is sensitive to the length of the protein sequence which favors the unfolded structure. Consequently, NCp7 is a really flexible protein (as expected from its biological function) and its two forms clearly coexist in water solution explaining both the observation of NMR close contacts by Roques et al. but also confirming a probable prevalence of the elongated B form in solution as proposed by the Summers Laboratory.

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 810367), project EMC2. LEK acknowledges funding from the Research Council of Saint-Joseph University of Beirut, Lebanon (Project FS71), the Lebanese National Council for Scientific Research, CNRS-L (Project CNRS-FS80), the French Institute- Lebanon, and the French-Lebanese Program CEDRE (Project 35327UJ). FC acknowledges funding from the French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR11-IDEX-0004-02) and support of the Direction Génerale de l'Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. Computations have been performed at GENCI on the Occigen machine (CINES, Montpellier, France) on grant no A0070707671 and on the MESU machine at ISCD, Sorbonne Univerisité. PR is grateful for support by the Robert A. Welch Foundation (F-1691) and National Institutes of

Health (R01GM114237).

Competing interest

The authors declare no competing interests.

Additional information

Figures and Tables in the Supporting Information

- Figure S1: Instantaneous work of the pulling force for the 20 independent SMD simulations as a function of simulation time.
- Figure S2: Pulling work for the 20 independent SMD simulations as a function of the end–to–end distance. 8 trajectories are below 0 kcal/mol and 12 trajectories are above 0 kcal/mol.
- Figure S3: PMF obtained in US with the CHARMM and AMOEBA force fields for 1/2/3 ns simulation windows. Corresponding umbrellas are depicted under each PMF graph.
- Figure S4: PMF obtained in US using the polarizable AMOEBA force field with the Roques' (PDB: 1ESK) and Summers' (PDB: 1MFS) NCp7 structures. The same minima are found at 5 and 12 Å for both curves, although the free energy barriers differ by 0.2 kcal/mol.
- Figure S5: PMF obtained in US using the non–polarizable CHARMM force field with the Roques' (PDB: 1ESK) and Summers' (PDB: 1MFS) structures. Neither the same energy minima nor the same PMF curvatures are found between the two structures.

- Table 1: Energy decomposition of structures A and B using the CHARMM non–polarizable force field. The charge–charge term provides a better energetical stabilization for structure A compared to structure B.

- Table 2: Energy decomposition of structures A and B using the AMOEBA polarizable force field. The polarization and electrostatic energies provide a better stabilization for structure B compared to structure A. It shows that polarization exerts an essential role in stabilizing structure B compared to A.

This information is available free of charge via the Internet at `http://pubs.acs.org`

List of Figures

List of Tables

References

(1) Nachega, J. B.; Hislop, M.; Dowdy, D. W.; Chaisson, R. E.; Regensberg, L.; Maartens, G. Adherence to nonnucleoside reverse transcriptase inhibitor–based HIV therapy and virologic outcomes. *Ann. Intern. Med.* **2007**, *146*, 564–573.

(2) Group, D. A. D. S., et al. Use of nucleoside reverse transcriptase inhibitors and risk of myocardial infarction in HIV-infected patients enrolled in the D: A: D study: a multi-cohort collaboration. *Lancet* **2008**, *371*, 1417–1426.

(3) Mallal, S.; Nolan, D.; Witt, C.; Masel, G.; Martin, A.; Moore, C.; Sayer, D.; Castley, A.; Mamotte, C.; Maxwell, D., et al. Association between presence of HLA-B* 5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **2002**, *359*, 727–732.

(4) de Rocquigny, H.; Shvadchak, V.; Avilov, S.; Dong, C. Z.; Dietrich, U.; Darlix, J.; Mély, Y. Targeting the viral nucleocapsid protein in anti-HIV-1 therapy. *Mini–Rev. Med. Chem.* **2008**, *8*, 24.

(5) Rice, W. G.; Turpin, J. A.; Huang, M.; Clanton, D.; Buckheit, R. W.; Covell, D. G.; Wallqvist, A.; McDonnell, N. B.; DeGuzman, R. N.; Summers, M. F., et al. Azodicarbonamide inhibits HIV-1 replication by targeting the nucleocapsid protein. *Nat. Med.* **1997**, *3*, 341.

(6) Turpin, J. A.; Song, Y.; Inman, J. K.; Huang, M.; Wallqvist, A.; Maynard, A.; Covell, D. G.; Rice, W. G.; Appella, E. Synthesis and biological properties of novel pyridinioalkanoyl thiolesters (PATE) as anti-HIV-1 agents that target the viral nucleocapsid protein zinc fingers. *J. Med. Chem.* **1999**, *42*, 67–86.

(7) Darlix, J.-L.; Cristofari, G.; Rau, M.; Péchoux, C.; Berthoux, L.; Roques, B. *Adv. Pharmacol.*; Elsevier, 2000; Vol. 48; pp 345–372.

(8) Morellet, N.; De Rocquigny, H.; Mely, Y.; Jullian, N.; Demene, H.; Ottmann, M.; Gerard, D.; Darlix, J.; Fournie-Zaluski, M.; Roques, B. Conformational behaviour of the active and inactive forms of the nucleocapsid NCp7 of HIV-1 studied by 1H NMR. *J. Mol. Biol.* **1994**, *235*, 287–301.

(9) Schmalzbauer, E.; Strack, B.; Dannull, J.; Guehmann, S.; Moelling, K. Mutations of basic amino acids of NCp7 of human immunodeficiency virus type 1 affect RNA binding in vitro. *J. Virol.* **1996**, *70*, 771–777.

(10) Demene, H.; Dong, C.; Ottmann, M.; Rouyez, M.; Jullian, N.; Morellet, N.; Mely, Y.; Darlix, J.; Fournie-Zaluski, M. 1H NMR structure and biological studies of the His23. fwdarw. Cys mutant nucleocapsid protein of HIV-1 indicate that the conformation of the first zinc finger is critical for virus infectivity. *Biochemistry.* **1994**, *33*, 11707–11716.

(11) Roques, B.; Morellet, N.; de Rocquigny, H.; Déméné, H.; Schueler, W.; Jullian, N. Structure, biological functions and inhibition of the HIV-1 proteins Vpr and NCp7. *Biochemie.* **1997**, *79*, 673 – 680.

(12) Barre-Sinoussi, F.; Chermann, J.; Rey, F.; Nugeyre, M.; Chamaret, S.; Gruest, J.; Dauguet, C.; Axler-Blin, C.; Vezinet-Brun, F.; Rouzioux, C.; Rozenbaum, W.; Montagnier, L. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome. *Science.* **1983**, *220*, 868–871.

(13) Di Marzo Veronese, F.; Rahman, R.; Copeland, T. D.; Oroszlan, S.; Gallo, R.; Sarngadharan, M. Immunological and Chemical Analysis of P6, the Carboxyl-Terminal Fragment of HIV P15. *AIDS. Res. Hum. Retrov.* **1987**, *3*, 253–264.

(14) Cis elements and Trans-acting factors involved in the RNA dimerization of the human immunodeficiency virus HIV-1. *J. Mol. Biol.* **1990**, *216*, 689 – 699.

(15) Prats, A. C.; Sarih, L.; Gabus, C.; Litvak, S.; Keith, G.; Darlix, J. L. Small finger protein

of avian and murine retroviruses has nucleic acid annealing activity and positions the replication primer tRNA onto genomic RNA. *EMBO J. 7*, 1777–1783.

(16) Prats, A.; Roy, C.; Wang, P.; Erard, M.; Housset, V.; Gabus, C.; Paoletti, C.; Darlix, J. Cis elements and trans-acting factors involved in dimer formation of murine leukemia virus RNA. *J. Virol.* **1990**, *64*, 774–783.

(17) Aronoff, R.; Hajjar, A. M.; Linial, M. L. Avian retroviral RNA encapsidation: reexamination of functional 5' RNA sequences and the role of nucleocapsid Cys-His motifs. *J. Virol.* **1993**, *67*, 178–188.

(18) de Rocquigny, H.; Petitjean, P.; Tanchou, V.; Decimo, D.; Drouot, L.; Delaunay, T.; Darlix, J.-L.; Roques, B. P. The Zinc Fingers of HIV Nucleocapsid Protein NCp7 Direct Interactions with the Viral Regulatory Protein Vpr. *J. Biol. Chem.* **1997**, *272*, 30753–30759.

(19) Carteau, S.; Batson, S. C.; Poljak, L.; Mouscadet, J. F.; de Rocquigny, H.; Darlix, J. L.; Roques, B. P.; Käs, E.; Auclair, C. Human immunodeficiency virus type 1 nucleocapsid protein specifically stimulates Mg2+-dependent DNA integration in vitro. *J. Virol.* **1997**, *71*, 6225–6229.

(20) Cornille, F.; Mely, Y.; Ficheux, D.; Savignol, I.; Gerard, D.; Darlix, J.-L.; Fournie-Zaluski, M.-C.; Roques, B. P. Solid phase synthesis of the retro viral nucleocapsid protein NCp10 of Moloney Murine Leukaemia virus and related "zinc-fingers" in free SH forms. *Int. J. Pept. Prot. Res.* **1990**, *36*, 551–558.

(21) Morellet, N.; Jullian, N.; De Rocquigny, H.; Maigret, B.; Darlix, J.; Roques, B. Determination of the structure of the nucleocapsid protein NCp7 from the human immunodeficiency virus type 1 by 1H NMR. *EMBO J.* **1992**, *11*, 3059–3065.

(22) Mely, Y.; Cornllle, F.; Fournié-Zaluski, M.-C.; Darlix, J.-L.; Roques, B. F.; Gérard, D. Investigation of zinc-binding affinities of moloney murine leukemia virus nucleocapsid

protein and its related zinc finger and modified peptides. *Biopolymers.* **1991**, *31*, 899–906.

(23) Morellet, N.; de Roquigny, H.; Mely, Y.; Jullian, N.; Demene, H.; Ottmann, N.; Gérard, D.; Darlix, J.-L.; Fournie-Zaluski, M.-C.; Roques, B.-P. Conformational behaviour of the active and inactive forms of the nucleocapsid NCp7 of HIV-1 studied by 1H NMR. *J. Mol. Biol.* **1994**, *235*, 287–301.

(24) Aldovini, A.; Young, R. A. Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus. *J. Virol.* **1990**, *64*, 1920–1926.

(25) Dorfman, T.; Luban, J.; Goff, S.; Haseltine, W.; Göttlinger, H. Mapping of functionally important residues of a cysteine-histidine box in the human immunodeficiency virus type 1 nucleocapsid protein. *J. Virol.* **1993**, *67*, 6159–6169.

(26) Gorelick, R. J.; Nigida, S.; Bess, J.; Arthur, L.; Henderson, L.; Rein, A. Noninfectious human immunodeficiency virus type 1 mutants deficient in genomic RNA. *J. Virol.* **1990**, *64*, 3207–3211.

(27) Ottmann, M.; Gabus, C.; Darlix, J.-L. The central globular domain of the nucleocapsid protein of human immunodeficiency virus type 1 is critical for virion structure and infectivity. *J. Virol.* **1995**, *69*, 1778–1784.

(28) South, T. L.; Blake, P. R.; Sowder III, R. C.; Arthur, L. O.; Henderson, L. E.; Summers, M. F. The nucleocapsid protein isolated from HIV-1 particles binds zinc and forms retroviral-type zinc fingers. *Biochemistry.* **1990**, *29*, 7786–7789.

(29) Summers, M. F.; South, T. L.; Kim, B.; Hare, D. R. High-resolution structure of an HIV zinc fingerlike domain via a new NMR-based distance geometry approach. *Biochemistry.* **1990**, *29*, 329–340.

(30) South, T. L.; Blake, P. R.; Hare, D. R.; Summers, M. F. C-terminal retroviral-type zinc finger domain from the HIV-1 nucleocapsid protein is structurally similar to the N-terminal zinc finger domain. *Biochemistry.* **1991**, *30*, 6342–6349.

(31) Morellet, N.; Jullian, N.; De Rocquigny, H.; Maigret, B.; Darlix, J.; Roques, B. Determination of the structure of the nucleocapsid protein NCp7 from the human immunodeficiency virus type 1 by 1H NMR. *EMBO J.* **1992**, *11*, 3059–3065 (1992).

(32) Morellet, N.; Demene, H.; Teilleux, V.; Huynh-Dinh, H., T. de Roquigny; Fournie-Zaluski, M.-C.; Roques, B.-P. Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d (ACGCC). *J. Mol. Biol.* **1998**, *283*, 419–434.

(33) Lee, B. M.; Guzman, R. N. D.; Turner, B. G.; Tjandra, N.; Summers, M. F. Dynamical Behavior of the HIV–1 Nucleocapsid Protein. *J. Mol. Biol.* **1998**, *279*, 633–649.

(34) MacKerrell, A. D. J.; et al., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(35) MacKerell, A. D. J.; Feig, M.; Brooks, C. L. 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.* **2004**, *25*, 1400–1415.

(36) Melcr, J.; Piquemal, J.-P. Accurate biomolecular simulations account for electronic polarization. *Front. Biosci.* **2019**, *6*, 143.

(37) Shi, Y.; Ren, P.; Schnieders, M.; Piquemal, J.-P. *Reviews in Computational Chemistry Volume 28*; John Wiley Sons, Ltd, 2015; Chapter 2, pp 51–86.

(38) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(39) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7*, 3143–3161.

(40) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory. Comput.* **2018**, *14*, 2084–2108.

(41) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardere, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: software tools for molecular design. *J. Chem. Theory. Comput.* **2018**, *14*, 5273–5289.

(42) Lagardère, L.; Jolly, L. H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **2018**, *9*, 956–972.

(43) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B.* **2010**, *114*, 2549–2564.

(44) Wu, J. C.; Piquemal, J.-P.; Chaudret, R.; Reinhardt, P.; Ren, P. Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 2059–2070.

(45) Grossfield, A.; Ren, P.; Ponder, J. W. Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.

(46) Foloppe, N.; A. D. MacKerell, J. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86–104.

(47) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(48) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(49) Kolafa, J. Time-reversible always stable predictor–corrector method for molecular dynamics of polarizable molecules. *J. Comp. Chem.* **2004**, *25*, 335–342.

(50) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **2003**, *119*, 3559–3566.

(51) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946–5961.

(52) Célerse, F.; Lagardère, L.; Derat, E.; Piquemal, J.-P. Massively parallel implementation of Steered Molecular Dynamics in Tinker-HP:comparisons of polarizable and non-polarizable simulations of realistic systems. *J. Chem. Theory. Comput.* **2019**, *15*, 3694–3709.

(53) Torrie, G. M.; Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(54) *Alan Grossfield, "WHAM: the weighted histogram analysis method", version 2.09, http://membrane.urmc.rochester.edu/content/wham*.

(55) How the HIV-1 Nucleocapsid Protein Binds and Destabilises the ()Primer Binding Site During Reverse Transcription. *J. Mol. Biol.* **2008**, *383*, 1112 – 1128.

(56) Amarasinghe, G. K.; Guzman, R. N. D.; Turner, R. B.; Chancellor, K. J.; Wu, Z. R.; Summers, M. F. NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the RNA packaging signal. implications for genome recognition. *J. Mol. Biol.* **2000**, *301*, 491 – 511.

(57) De Guzman, R. N.; Wu, Z. R.; Stalling, C. C.; Pappalardo, L.; Borer, P. N.; Summers, M. F. Structure of the HIV-1 Nucleocapsid Protein Bound to the SL3 -RNA-RNA Recognition Element. *Science* **1998**, *279*, 384–388.

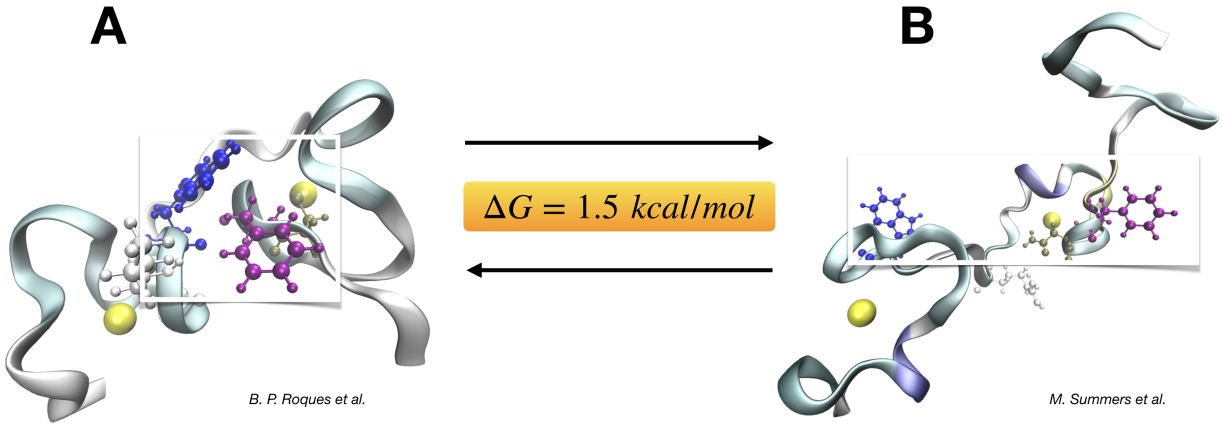
(58) Jenkins, L. M. M.; Hara, T.; Durell, S. R.; Hayashi, R.; Inman, J. K.; Piquemal, J.-P.; Gresh, N.; Appella, E. Specificity of acyl transfer from 2-mercaptobenzamide thioesters to the HIV-1 nucleocapsid protein. *J. Am. Chem. Soc.* **2007**, *129*, 11067–11078.

(59) Hare, S.; Gupta, S. S.; Valkov, E.; Engelman, A.; Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **2010**, *464*, 232.

(60) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. E.* **1997**, *78*, 2690.

(61) Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E.* **1997**, *56*, 5018.

(62) CasasFinet, J.; Urbaneja, M.; McGrath, G.; Gorelick, R.; Bosche, W.; Kane, B.; Arthur, I.; Henderson, L.; Copeland, T.; Lee, B.; DeGuzman, R.; Summers, M. Conformational heterogeneity in HIV-1 nucleocapsid protein p7. *Biophys. J.* **1997**, *72*, TUPM2.


A
B
$\Delta G = 1.5\ kcal/mol$
B. P. Roques et al.
M. Summers et al.