Kinetic characterization of the critical step in HIV-1 protease maturation

S. Kashif Sadiq[a,1], Frank Noé[b], and Gianni De Fabritiis[a,1]

[a]Computational Biophysics Laboratory, GRIB-IMIM, Universitat Pompeu Fabra, 08003 Barcelona, Spain; and [b]Research Center Matheon, Freie Universität Berlin, 14195 Berlin, Germany

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved October 16, 2012 (received for review June 29, 2012)

HIV maturation requires multiple cleavage of long polyprotein chains into functional proteins that include the viral protease itself. Initial cleavage by the protease dimer occurs from within these precursors, and yet only a single protease monomer is embedded in each polyprotein chain. Self-activation has been proposed to start from a partially dimerized protease formed from monomers of different chains binding its own N termini by self-association to the active site, but a complete structural understanding of this critical step in HIV maturation is missing. Here, we captured the critical self-association of immature HIV-1 protease to its extended amino-terminal recognition motif using large-scale molecular dynamics simulations, thus confirming the postulated intramolecular mechanism in atomic detail. We show that self-association to a catalytically viable state requires structural cooperativity of the flexible β-hairpin "flap" regions of the enzyme and that the major transition pathway is first via self-association in the semiopen/open enzyme states, followed by enzyme conformational transition into a catalytically viable closed state. Furthermore, partial N-terminal threading can play a role in self-association, whereas wide opening of the flaps in concert with self-association is not observed. We estimate the association rate constant (k_{on}) to be on the order of $\sim 1 \times 10^4\ s^{-1}$, suggesting that N-terminal self-association is not the rate-limiting step in the process. The shown mechanism also provides an interesting example of molecular conformational transitions along the association pathway.

conformational kinetics | Markov state model | high-throughput molecular dynamics

HIV, along with all retroviruses, achieves infectious maturation of nascent virus particles through cleavage of polyprotein precursors by the viral protease. In particular, the GagPol chains contain several covalently linked proteins including the protease itself (Fig. 1*A*). Thus, maturation of the virus is initiated by autocatalysis of viral protease initially embedded in GagPol precursors. HIV-1 protease is functional only in dimeric form (1, 2) because activity of monomeric protease precursor is three orders of magnitude less than the mature dimer (3), and yet only a single monomer is embedded within each precursor. Two individual monomers in different GagPol chains must, therefore, come together to form an embedded dimeric protease, which ultimately cleaves itself into a mature form (Fig. 1*B*).

Experiments indicate initial cleavage by a precursor protease still embedded in the GagPol chain occurs through an intramolecular, concentration-independent mechanism with the precursor protease cleaving its own terminus (4) and critically modulated by the N-terminal region (5–8). Mutations that block N-terminal cleavage result in severe loss of efficiency in catalytic activity. Cleavage at the protease (PR), reverse transcriptase (RT) junction at the C-terminal end of the protease by the precursor protease occurs via an intermolecular, concentration-dependent mechanism (9); mutations that block C-terminal cleavage, do not significantly affect either in vitro enzymatic activity and protease dimerization (9) or in vivo GagPol cleavage and virus maturation, suggesting that the precursor protease is well formed once the N terminus is cleaved. Therefore, experimentally, the only absolute prerequisite for mature-like catalytic activity and completion of viral precursor processing is autoprocessing at the N terminus of protease.

Structurally, this process of intramolecular cleavage requires that HIV-1 protease self-associate the precleaved N terminus to its own active site. Access to this site is modulated by a pair of flexible β-hairpin flaps (Fig. 1*C*) that must first open to allow entry and then close (10–15) to make substrate cleavage viable through a structurally conserved recognition pattern (16–19). Recent advances in paramagnetic relaxation enhancement (PRE) (20) reveal transient N-terminal contacts within the active site, but a complete structural characterization of this transient process in full atomic detail remains an outstanding challenge and is the aim of this work.

All-atom molecular dynamics simulation is a powerful computational tool to investigate such transient events (21) at the atomic level. Here, we present a computational study of the process of N-terminal self-association of HIV-1 protease. Using high-throughput ensembles of unbiased all-atom explicit solvent molecular dynamics simulations using ACEMD (22) on a distributed computing network (23), we investigate at full-atomic resolution the complete self-association process of the cleavage recognition site (VSFNF-PQIT) at the N terminus of HIV-1 protease (Fig. 1*C*), representing immature protease in GagPol precursor.

Although the above treatment provides evidence of the intramolecular cleavage mechanism, it is further interesting to understand more quantitatively the role of molecular conformational transitions along the self-association pathway. The existence of conformational ensembles in free enzyme is well established (24, 25). We have shown previously that several conformations in HIV-1 protease preexist in apo form (15).

Our investigation is partitioned as follows. Firstly, we perform 417×400 ns simulations (run set E1) of an immature protease in an initially N-terminal disassociated state with flaps in a semiopen conformation. These demonstrate that N-terminal active site entry is possible. However, further flap rearrangements are necessary to complete self-association to a catalytically viable closed-conformation within the structural envelope of existing peptidic ligand complexes (16–19). We, therefore, also perform 416×400 ns simulations (run set E2) of an immature protease starting from an initially self-associated N terminus and with flaps in a semiopen conformation, exploring the transition into a closed conformation.

Author contributions: S.K.S. and G.D.F. designed research; S.K.S. and G.D.F. performed research; S.K.S., F.N., and G.D.F. analyzed data; and S.K.S., F.N., and G.D.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: kashif.sadiq@upf.edu or gianni.defabritiis@upf.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210983109/-/DCSupplemental.

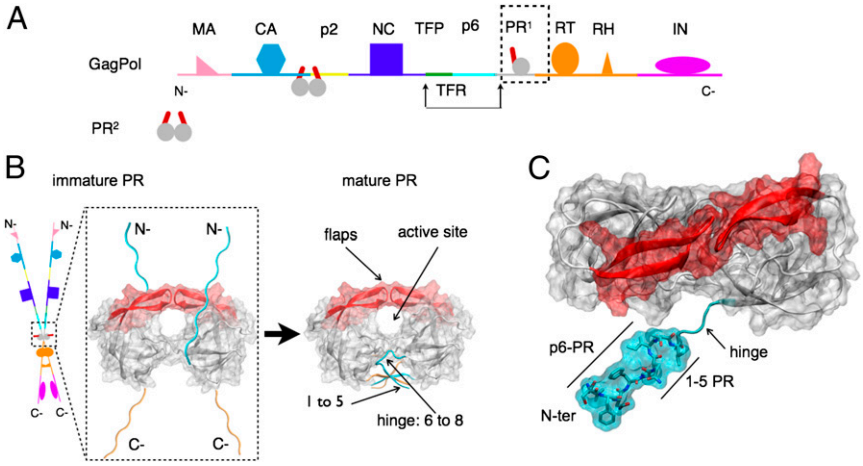


Fig. 1. (*A*) Schematic representation of a GagPol polyprotein precursor of HIV. GagPol is composed of a linear heteropolymer chain. Only a single monomer of protease (PR) is present on each GagPol chain. Mature protease cleaves GagPol at recognized cleavage sites between proteins. (*B*) Schematic representation of transient dimerization of two GagPol chains to form an embedded protease dimer. This precursor autocatalyzes its own liberation into the mature form of the dimeric protease that consists of natively folded N-terminal (cyan) and C-terminal (orange) chains in an interdigitated four-stranded β-sheet. A pair of β-hairpin structures termed the "flaps" (red) mediate active site access. (*C*) Structural representation of N-terminal (N-ter) construct (cyan) of immature HIV-1 protease, with 5-aa (VFSNF) extension corresponding to the p6-PR cleavage site. An unstructured hinge region connects the tertiary structure to the N-ter region.

Results and Discussion

Observation of N-Terminal Self-Association and Closed-Flap Catalytic Viability. No atomic resolution structure exists of an N-terminal self-associated HIV-1 protease, nor of the p6-PR octapeptide complex, which shares sequence identity to the N-terminal region. Therefore, we derived a putative N-terminal self-associated HIV-1 protease with closed-flap conformation (R1) from an existing crystal structure of MA-CA cleavage region peptide-bound HIV-1 protease, which shares partial sequence and structural overlap to the p6-PR cleavage region (*SI Materials and Methods*). The structure was subsequently relaxed in a molecular dynamics simulation for 1 μs. We tested the validity of this N-terminal self-association construct R1 by comparing against two control sets of molecular simulations: (*i*) MA-CA–bound (control C1) and (*ii*) p6-PR–bound (control C2) HIV-1 protease complexes (*SI Materials and Methods* and Figs. S1 and S2). Analysis showed that the cleavage peptide region of the construct maintains a stable conformation with cleavable geometry. Based on these analyses, the final structure of the R1 simulation was selected as a reference structure to compare unbiased self-association simulations of systems E1 and E2 (*SI Materials and Methods*).

The time evolution of a sample of trajectories for each of the two run sets is shown in Fig. 2 and highlights, in particular, a single trajectory from each set (E1, E2) with corresponding structural representations at various time points. The root-mean-square deviation (rmsd) of the N-terminal region with respect to the reference closed-flap N-terminal self-associated structure (R1) was measured along the time course for each run set. The flap conformation was measured using a 1D metric λ_x, which sharply separates open ($\lambda_x \approx -10$ Å), semiopen ($\lambda_x \approx -5$ Å), and closed ($\lambda_x \approx 5$ Å) flap conformations (Fig. 2).

For run set E1, from 417 simulations each run for 400 ns from a HIV-1 protease with an initially disassociated N terminus [Fig. 2 (A)] and with flaps in a semiopen conformation, we captured 18 events (4.5%) of self-association to the active site within 5-Å rmsd to the reference structure R1 and two events (0.5%) within 3-Å rmsd. However, the latter were not accompanied by a flap transition to the closed state. The majority of trajectories did not exhibit N-terminal entry, although a significant number of all trajectories (38%) formed an encounter complex (15 Å > rmsd > 10 Å) in which the N-terminal region was associated to the side of the active site but did not enter [Fig. 2 (B)]. N-terminal entry was observed via a mixture of modes, making use of an open-flap conformation [Fig. 2 (C) and Movie S1], as well as lateral threading to a self-associated state [Fig. 2 (D) and Movie S2]. The N terminus was also observed to adopt a hairpin conformation that initiated self-association, followed by flap opening. In vivo, lateral threading is not possible because the N-terminal region continues into the much longer upstream GagPol chain and is an artifact of the system construct. By contrast, a combined hairpin and flap-opening mechanism is physiologically permissible. For run set E2, from 416 simulations each for 400 ns starting from an initially tightly (rmsd < 3 Å) self-associated [Fig. 2 (E)] N-terminal region with a semiopen-flap conformation, around 70% of the trajectories stayed within 3-Å rmsd. Eight trajectories (2%) exhibited greater flexibility, sampling regions of rmsd > 5 Å.

For E1, only 10 trajectories (2.5%) remained exclusively within the semiopen conformation (-6 Å > λ_x > -1 Å); indeed, reversible flap transitions between semiopen and open ($\lambda_x < -8$ Å) occurred in 306 trajectories (73%); this is expected given the <10-ns timescale measured previously for the transition (10). Transitions between semiopen and closed ($\lambda_x \approx 5$ Å) were also observed in 158 trajectories (38%). For E2, 392 trajectories (94%) remained within the semiopen conformation, 22 (5.3%) made transitions to an open conformation and only two made (0.5%) a sharp transition [Fig. 2 (F)] into a catalytically viable closed conformation [Fig. 2 (G) and Movie S3]. The fact that both self-association from a disassociated state and conformational reversal from semiopen associated state to a catalytically viable closed-flap state occur provides compelling evidence that autocatalytic maturation of HIV-1 protease occurs via an intramolecular mechanism.

Analysis of Conformational Transitions. In this study, the entire ensemble of trajectories covers all steps of the complete association and dissociation pathway, thus permitting to assemble a kinetic model of the entire process using appropriate statistical methods. Recently, Markov (state) models (MSMs), also termed kinetic networks (or transition networks), have received a surge of interest (26–29) and have been used successfully to calculate several slow processes from ensemble molecular dynamics data (30–32) (*SI Text*). Here, we chose to build an MSM based on the two informative order parameters λ_x and N-ter rmsd. The 2D

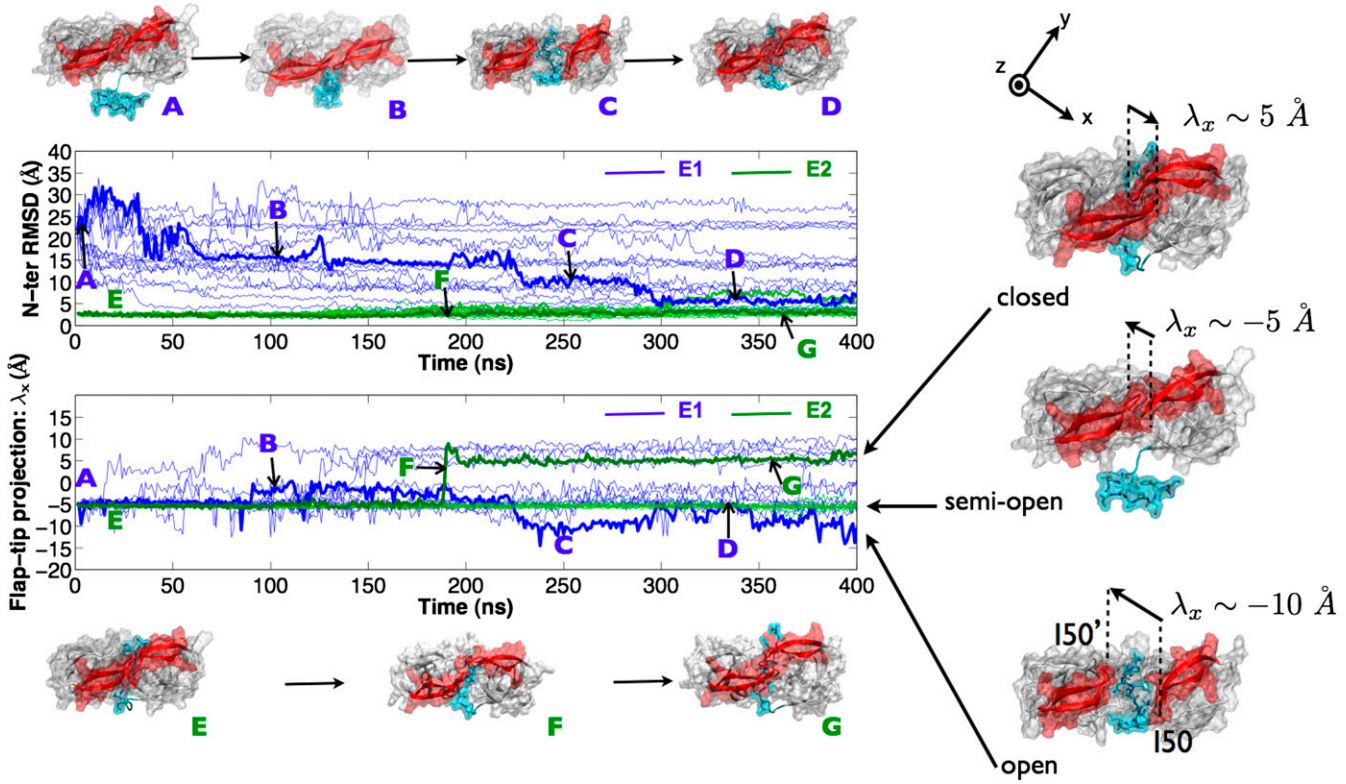


Fig. 2. N-terminal rmsd and flap-tip λ_x values for several of the 400 trajectories from each of the E1 (blue) and E2 (green) run sets. Representative trajectories showing capture of N-terminal self-association from unbound state (A through D) and enzyme-conformational flap-closure from an N-ter self-associated state (E through G) are in bold. Open, semiopen, and closed flap conformations can be sharply distinguished by a 1D metric based on the I50-I50′ flap-tip separation vector projected on the frame-aligned *x* axis (λ_x).

parameter space was discretized into 1,400 states based on a 1.0-Å^2 bin size. Each data point from the aggregate of ~335-μs simulation trajectories was assigned to a corresponding state. A transition matrix T(τ) was then constructed using a reversible estimator from a count matrix of transitions between these states at a lag time of τ = 50 ns (Fig. S3) and validated using a Chapman–Kolmogorov test (Fig. S4) as described previously (28).

Energetics and conformations. Fig. 3*A* shows the potential of mean force (PMF) on the two selected order parameters. The PMF energy is generated as $F(x, y) = -\log\pi_i(x, y)/kT$, where π_i is the stationary probability of state *x*, *y* estimated from the MSM transition matrix. Seven metastable energy wells are identifiable from the energetic map. Kinetic clustering to seven metastable states using Perron cluster cluster analysis (PCCA$^+$) results in state partitioning consistent with the energetic wells described by the PMF, and the corresponding states, S1 to S7, are labeled (Fig. 3*B*). Conformations A through G, described before, correspond to locations within this discretization space. Conformations A, B, E, and G lie within the energy wells of metastable states S7, S3, S2, and S1, respectively, conformations C and D are grouped together in S3, whereas conformation F is a transition state between S2 and S1.

The free energy of the most popular state was arbitrarily set to zero, and all other states were compared relative to it. Structurally, states S7 (1.3 kcal/mol) and S6 (1.7 kcal/mol) correspond to N-ter–disassociated semiopen/open and closed conformations, respectively. States S5 (1.3 kcal/mol) and S4 (1.7 kcal/mol) to closed-flap intermediate complexes, whereas S3 (0.6 kcal/mol) corresponds to a semiopen/open flap intermediate complex. Finally, states S2 (1.3 kcal/mol) and S1 (0 kcal/mol) correspond to the closed and semiopen N-ter–associated conformations. Our analysis shows that S1 is the energetically favored state, whereas S2 is by comparison an excited state with ΔG = 1.3 kcal/mol. The free energy of all states is within a range of 2 kcal/mol of each other. The equilibrium favored flap conformation in the disassociated state is semiopen/open (S7). We validated the convergence of the PMF calculation by applying a bootstrapping method that allowed determination of the error in the calculation (*SI Text*). The mean PMF was almost identical to that calculated from the entire dataset, and the SD was less than 0.4 kcal/mol for all relevant microstates within any given macrostate (Fig. S5), except for S2, which reached an error of 0.8 kcal/mol. The highest regions of error (1.6 kcal/mol) corresponded to poorly sampled regions at the extremities of the conformational landscape (Fig S6).

Kinetics and mechanism. The kinetics of interconversion between the disassociated state S7 and the catalytically viable state S2 was determined by computing the total net flux between the states [$(1.07 \pm 0.44) \times 10^4$ s^{-1}], from which the self-association constant could be calculated (30) to be $k_{on} = (1.12 \pm 0.56) \times 10^4$ s^{-1}. Furthermore, the total flux was decomposed into pathways along pairs of states from which the flux for all pathways along sets of states between S7 to S2 could be identified and sorted (Table S1). The flux network from the 90% most relevant transition pathways is shown in Fig. 3*C*. The remaining 10% of the flux is in minor pathways between the states shown in Fig. 3*C* and is omitted for clarity of visualization. Most of the flux passes through state S3 (8.3×10^3 s^{-1}); the state, therefore, acts as a major transition hub for the process. State S6, on the other hand, is a "dead-end" state, with insignificant flux passing through it.

The most probable transition pathway is S7→S3→S1→S2 and accounts for 53.6 % of the total flux, with a flux of $(5.73 \pm 2.78) \times 10^3$ s^{-1}. This pathway corresponds to the process of prior self-association in the semiopen/open conformation, followed by an

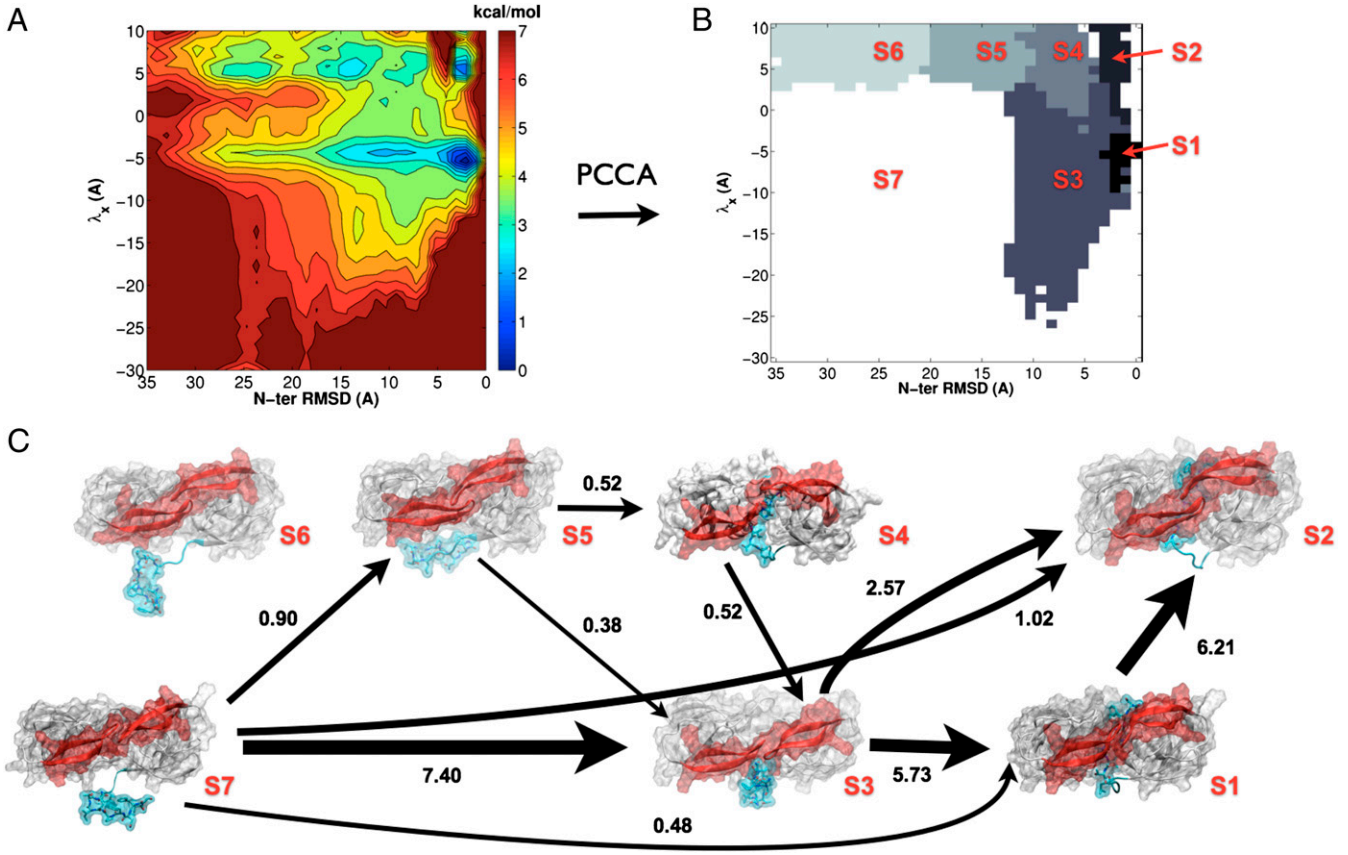


Fig. 3. (*A*) PMF from a MSM in the discretized λ_x N-terminal rmsd projection. (*B*) Kinetic clustering of the discretized λ_x N-terminal rmsd space into seven distinct metastable states, S1 through S7, using the PCCA method (44). (*C*) Network of the 90 % most relevant transition pathways from the disassociated semiopen state S7 to the self-associated catalytically viable state S2. The thickness of the arrows indicates the flux of folding trajectories between each pair of states. For each state, S1 to S7, a representative structure is shown. The numbers next to the arrows give the normalized net flux (in $\times 10^3$ s^{-1}).

induced-fit mechanism of flap closure into a catalytically viable state. The other most relevant pathways are in order: S7→S3→S2 (15.5% of total flux), S7→S2 (9.6% of total flux), S7→S1→S2 (4.5% of total flux), S7→S5→S4→S3→S2 (4.2% of total flux), and S7→S5→S3→S2 (3.6% of total flux).

The pathways corresponding to self-association via prior selection of a closed-flap conformation by a self-associating N terminus, that is, those that do not pass through S3 or S1 (S7→S6→S5→S4→S2, S7→S6→S4→S2, S7→S5→S4→S2, S7→S6→S2, S7→S5→S2, and S7→S4→S2) are far weaker (lying within the 10% residual flux not displayed), and combined, carry 7.6% of the total flux. The statistical error in the flux calculations was estimated by computing the SD of the fluxes obtained from 10 subsets of 400 trajectories taken from the entire dataset (*Materials and Methods*). The fluxes between each pair of states and from all 10 subsets are listed in full in Tables S2 and S3.

Interestingly, a seven-state model does not kinetically separate semiopen and open conformations in either the disassociated (S7) or intermediate states (S3). This agrees well with previous calculations and experimental measurements of the relaxation rate of fast-flap conformational opening, which exhibits a small kinetic barrier ($k_{\text{flap}} \approx 0.1$ ns^{-1}) (10, 15, 33, 34). Even though self-association via both flap opening and lateral threading was observed, it is difficult to attribute any event solely to either one or the other mechanism. In our data, self-association events occur via a mixture of the two modes on a nanosecond timescale. The intermediate state, therefore, permits both threading and partial flap opening concurrently and is reflected in the results of the kinetic clustering procedure that assign state S3 to both conformations. Although, in vivo, the flaps must open for GagPol to reach the linear peptide formation required for cleavage, our study shows that a threading mechanism may also partially contribute to traversal across the active site. This is further supported by the previous observation of lateral substrate binding (35). In vivo, threading could be achieved by a hairpin peptide formation, followed by flap opening and reclosing in conjunction with unfolding of the hairpin. The fact that there is a small flux from S4 to S2 suggests that a minor contribution to the hairpin-threading mechanism may also come from a partially closed flap form.

A wide-open flap conformation in the free enzyme has been characterized previously (14, 15). Here, a wide-open conformation corresponds to $\lambda_x < -20$ Å. Only a few trajectories transiently exhibited wide-opening, consistent with the experimentally observed 100-μs relaxation timescale (33) and with previous simulations of the mature protease (15). This also places a lower limit of ∼100 μs on the wide-opening relaxation time for an immature protease. Interestingly, we did not observe any wide-opening events in conjunction with N-terminal self-association, as reflected in the PMF (Fig. 3*A*), demonstrating that flap opening not wide opening is sufficient for N-terminal self-association and suggesting that the kinetic behavior of the flaps is connected to the state of the N terminus. Many events of flap wide opening do, however, show simultaneous association of the N terminus to the dimer interface, resembling the exterior strand of the mature dimer interface. Therefore, wide opening might predominantly occur in mature-like protease dimers; however, to determine

Table 1. HIV-1 protease maturation system constructs

System	Definition	Flaps	Protease	Bound ligand	Production runs, ns
E1	Disassociated N-ter	Semiopen	PR^+:PR	no	417 × 400
E2	Self-associated N-ter	Semiopen	PR^+:PR	no	416 × 400
R1	Self-associated N-ter	Closed	PR^+:PR	no	1 × 1,000
C1	MA-CA bound complex	Closed	PR:PR	SQNY-PIVQ	100 × 100
C2	p6-PR bound complex	Closed	PR:PR	SFNF-PQIT	100 × 100

PR^+ is PR with VSFNF N-terminal extension.

accurate kinetics for it in either the mature or immature protease would require at least an order of magnitude greater sampling.

Of the 18 trajectories that exhibited self-association, most occur via a mixture of lateral threading, N-terminal hairpin formation, and flap opening. Therefore, we recalculate the association constant while excluding those trajectories for which the initial entry event was lateral threading. This does not qualitatively change the results; the same energetic minima appear in the PMF, and a seven-state kinetic model yields a similar separation of states (Fig. S4). The calculated association rate constant is $k_{on} = 1.18 \times 10^4\ s^{-1}$, which is similar to the former result. Therefore, because flap-opening kinetics are fast, the kinetics of self-association are not rate-limited by the requirement that the flaps open in vivo.

The self-associated open state is also kinetically indistinguishable from the semiopen, S1. The tapering of the PMF in the low N-ter rmsd region shows that from the intermediate state, increased N-ter self-association sharply reduces the probability of maintaining an open state. This is consistent with a picture in which any instance of an open self-associated state converges rapidly to a semiopen state and is supported by nanosecond-timescale observations of flap closure after manual placement of ligands into a flap-open protease active site (36, 37).

Conclusions

We have investigated the proposed intramolecular mechanism by which immature HIV-1 protease autocatalyzes its own release from nascent GagPol polyprotein precursors, using all-atom explicit solvent molecular dynamics simulations of an immature dimeric HIV-1 protease construct. The observed overlap between forward and reverse processes, together with the completion of closed-form self-association, is compatible with N-terminal intramolecular autocatalysis and proceeds via flexible rearrangement of the N terminus, structurally coordinated opening of the flaps to allow entry, followed by closure of the flaps to form a catalytically viable complex. Furthermore, partial N-terminal threading plays a role in self-association, whereas wide opening of the flaps in concert with self-association is not observed. From these data, we construct an MSM process that permits a quantitative kinetic analysis. The most probable transition pathway from a disassociated to a catalytically viable state is first via self-association in the semiopen/open state of the protease, followed by a flap-conformational change into closed form.

The on rate (k_{on}) of self-association to a closed state is estimated from our simulations to be of the order of $\sim 1 \times 10^4\ s^{-1}$, whereas the rate constant of intramolecular autocatalytic cleavage, k_{first}, is several orders of magnitude smaller ($k_{first} \approx 7 \times 10^{-4}\ s^{-1}$) (5). Therefore, not only is N-ter self-association from a dimeric precursor possible, but it is also comparatively fast, not being the rate-limiting step in the overall autocatalytic process.

This indicates that either initial dimerization or postcleavage formation of the native fold is rate-limiting. However, because cleavage of the N terminus results in mature-like activity, formation of the postcleaved native fold is likely not to play a limiting role either. Structurally, this can be explained because the N-terminal junction serves as a tether that restricts the diffusion and, thus, function of the protease, but once cleaved, the protease is free to move around and process other substrates more easily. Thus, by elimination, the limiting step in the autocatalytic process is likely to be the initial dimerization of the GagPol chains to form an immature dimer capable of N-terminal cleavage, as seen also by Tang et al. (20).

The necessary function of the N terminus in autocatalysis makes targeting the preliberated structure a potentially resistance-proof strategy for a class of antiretroviral allosteric inhibitors of the immature protease.

Materials and Methods

System Preparation. Five different HIV-1 protease systems were constructed for this study. These were the E1, E2, R1, C1, and C2 constructs (Table 1). E1 was constructed to investigate initial N-ter self-association, and E2 was constructed to investigate subsequent flap closure to a catalytically viable state. R1 served as a reference closed-flap self-associated state, whereas C1 and C2 were MA-CA and p6-PR ligand–bound octapeptide control complexes against which R1 was validated. The exact protocol for preparation depended on the specific requirements of each system and are described in *SI Materials and Methods*. Corresponding crystal structures were taken from the Protein Data Bank (38). The standard AMBER force field (ff03) with standard ions (39) was used to describe all parameters. Each system was solvated using TIP3P water (40) and electrically neutralized (ionic concentration, 0.15 M NaCl) with long-range Coulomb interactions handled using the particle mesh Ewald summation method (PME) (41, 42). See *SI Materials and Methods* for the simulation settings and protocols. Experimental accuracy of the molecular simulation protocol for the HIV-1 protease has been validated previously using NMR S^2 order parameters (15). All production simulations were carried out using ACEMD (22). Production-ensemble simulations were deployed for systems E1, E2, C1, and C2 on the GPUGRID compute infrastructure (23). A set of 417 × 400 ns and 416 × 400 ns was used for the analysis of E1 and E2 sets, respectively. Similarly, a subset of 100 × 100 ns each was used for the analysis of the C1 and C2 control sets.

Analysis. N-terminal self-association proximity was measured using rmsd of the C_α atoms of residues −4 to +4 of the N terminus relative to the closed-bound reference structure (R1). Visual inspection confirms N-terminal active-site entry corresponds to an rmsd of within 5 Å to R1 and is, thus, used to define self-association. Catalytic viability of the N-terminal region requires closer self-association, to within 3-Å rmsd (*SI Materials and Methods*), as well as a conformational change in the flap region from the equilibrium semiopen to the closed conformation. We devised a 1D metric based on the I50-I50′ flap-tip separation vector that can distinguish these conformations sharply. The λ_x metric is the projection of this vector on the frame-aligned x axis. The x axis was chosen to be the vector between the center of mass (COM) of the backbone atoms of residues 23, 24, and 85 of the second and first monomers, respectively, each frame. These residues are within the β-sheeted region that supports each side of the active site and exhibit very small root-mean-square fluctuations (rmsfs) compared with other residues (15). The use of frame-aligned vector projection allows for opposite flap-handedness between semiopen and closed conformations to be represented by opposite signs (Fig. 2).

To describe the dynamics of the molecular system, we analyzed the sequence of transitions between discretized states by a MSM. The MSM construction was performed using a combination of the ACEMD toolkit (ACEMDTK) (http://multiscalelab.org/acemd/protocols) and the EMMA package (43). See *SI Materials and Methods* for details of the MSM construction. The aggregate of ∼335 μs of simulation data from ensemble sets E1 and E2 was used to construct a discretized MSM using a 2D projection in the λ_x N-ter rmsd space from which a PMF was calculated (Fig. 3).

Other reaction coordinates were also considered but the chosen 2D projection can clearly decompose flaps dynamics with N-terminal binding. N-terminal binding could be defined in terms of rmsd because there is a well-defined bound structure for comparison, but flap conformations are not sharply separated using rmsd in a single dimension. For example, relative to a closed conformation, several conformations yield similar (within 2 Å) rmsds to each other and, thus, degenerate states in the MSM. The λ_x metric exploits the symmetry of the protease and distinguishes the conformations of interest, and yet only uses a single dimension that is independent of any reference structure.

The discretized model was further clustered into a smaller set of coarse-grained metastable states using the PCCA method (44). Kinetic fluxes between these states was computed using the methodology presented previously (30). The error in the PMF and the fluxes was calculated by applying a bootstrapping method (Fig. S5). The flux was calculated for each subset based on the metastable state definitions obtained in the PCCA$^+$ clustering from the entire dataset (Tables S1 and S2).

ACKNOWLEDGMENTS. We thank the volunteers of GPUGRID who donated GPU computing time to the project. S.K.S. was supported by a European Commission Seventh Framework Programme Marie Curie Intra-European Fellowship. G.D.F. was supported by the Ramón y Cajal Scheme and Spanish Ministry of Science and Innovation Grant BIO2011-27450. F.N. acknowledges funding through Deutsche Forschungsgemeinschaft Program NO 825/3.

1. Darke PL, et al. (1994) Dissociation and association of the HIV-1 protease dimer subunits: Equilibria and rates. *Biochemistry* 33(1):98–105.
2. Darke PL (1994) Stability of dimeric retroviral proteases. *Methods Enzymol* 241: 104–127.
3. Ishima R, Torchia DA, Lynch SM, Gronenborn AM, Louis JM (2003) Solution structure of the mature HIV-1 protease monomer: Insight into the tertiary fold and stability of a precursor. *J Biol Chem* 278(44):43311–43319.
4. Pettit SC, Everitt LE, Choudhury S, Dunn BM, Kaplan AH (2004) Initial cleavage of the human immunodeficiency virus type 1 GagPol precursor by its activated protease occurs by an intramolecular mechanism. *J Virol* 78(16):8477–8485.
5. Louis JM, Nashed NT, Parris KD, Kimmel AR, Jerina DM (1994) Kinetics and mechanism of autoprocessing of human immunodeficiency virus type 1 protease from an analog of the Gag-Pol polyprotein. *Proc Natl Acad Sci USA* 91(17):7970–7974.
6. Wondrak EM, Louis JM (1996) Influence of flanking sequences on the dimer stability of human immunodeficiency virus type 1 protease. *Biochemistry* 35(39):12957–12962.
7. Louis JM, Clore GM, Gronenborn AM (1999) Autoprocessing of HIV-1 protease is tightly coupled to protein folding. *Nat Struct Biol* 6(9):868–875.
8. Louis JM, Wondrak EM, Kimmel AR, Wingfield PT, Nashed NT (1999) Proteolytic processing of HIV-1 protease precursor, kinetics and mechanism. *J Biol Chem* 274(33): 23437–23442.
9. Wondrak EM, Nashed NT, Haber MT, Jerina DM, Louis JM (1996) A transient precursor of the HIV-1 protease. Isolation, characterization, and kinetics of maturation. *J Biol Chem* 271(8):4477–4481.
10. Freedberg DI, et al. (2002) Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci* 11(2):221–232.
11. Scott WRP, Schiffer CA (2000) Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. *Structure* 8(12):1259–1265.
12. Chang CE, Shen T, Trylska J, Tozzini V, McCammon JA (2006) Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophys J* 90(11):3880–3885.
13. Tóth G, Borics A (2006) Flap opening mechanism of HIV-1 protease. *J Mol Graph Model* 24(6):465–474.
14. Hornak V, Okur A, Rizzo RC, Simmerling C (2006) HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc Natl Acad Sci USA* 103(4): 915–920.
15. Sadiq SK, De Fabritiis G (2010) Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins* 78(14):2873–2885.
16. Prabu-Jeyabalan M, Nalivaika EA, Romano K, Schiffer CA (2006) Mechanism of substrate recognition by drug-resistant human immunodeficiency virus type 1 protease variants revealed by a novel structural intermediate. *J Virol* 80(7):3607–3616.
17. Prabu-Jeyabalan M, Nalivaika E, Schiffer CA (2002) Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure* 10(3):369–381.
18. Prabu-Jeyabalan M, Nalivaika E, Schiffer CA (2000) How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *J Mol Biol* 301(5):1207–1220.
19. Prabu-Jeyabalan M, et al. (2006) Substrate envelope and drug resistance: Crystal structure of RO1 in complex with wild-type human immunodeficiency virus type 1 protease. *Antimicrob Agents Chemother* 50(4):1518–1521.
20. Tang C, Louis JM, Aniana A, Suh JY, Clore GM (2008) Visualizing transient events in amino-terminal autoprocessing of HIV-1 protease. *Nature* 455(7213):693–696.
21. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9(9):646–652.
22. Harvey MJ, Giupponi G, Fabritiis GD (2009) Acemd: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* 5(6):1632–1639.
23. Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G (2010) High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model* 50(3):397–403.
24. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5(11):789–796.
25. Henzler-Wildman KA, et al. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450(7171):913–916.
26. Bowman GR, Voelz VA, Pande VS (2011) Taming the complexity of protein folding. *Curr Opin Struct Biol* 21(1):4–11.
27. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 126(15):155101.
28. Prinz JH, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys* 134(17):174105.
29. Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B* 112(19):6057–6069.
30. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106(45):19011–19016.
31. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 132(5): 1526–1528.
32. Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc Natl Acad Sci USA* 108(25):10184–10189.
33. Ishima R, Freedberg DI, Wang YX, Louis JM, Torchia DA (1999) Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure* 7(9):1047–1055.
34. Katoh E, et al. (2003) A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex. *Protein Sci* 12(7):1376–1385.
35. Pietrucci F, Marinelli F, Carloni P, Laio A (2009) Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J Am Chem Soc* 131(33): 11811–11818.
36. Tóth G, Borics A (2006) Closing of the flaps of HIV-1 protease induced by substrate binding: A model of a flap closing mechanism in retroviral aspartic proteases. *Biochemistry* 45(21):6606–6614.
37. Hornak V, Okur A, Rizzo RC, Simmerling C (2006) HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J Am Chem Soc* 128(9):2812–2813.
38. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242.
39. Duan Y, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24(16):1999–2012.
40. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
41. Essmann U, Perera L, Berkowitz ML, Darden T (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–9593.
42. Harvey M, De Fabritiis G (2009) An implementation of the smooth particle mesh Ewald method on GPU hardware. *J Chem Theory Comput* 5(9):2371–2377.
43. Senne M, Trendelkamp-Schroer B, Mey ASJS, Schütte C, Noé F (2012) Emma - a software package for Markov model building and analysis. *J Chem Theory Comput* 8(7): 2223–2238.
44. Noé F, Horenko I, Schütte C, Smith JC (2007) Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J Chem Phys* 126(15):155102.

Supporting Information

Sadiq et al. 10.1073/pnas.1210983109

SI Materials and Methods

Molecular Simulation Protocol. The exact protocol for preparation depended on the specific requirements of each system and are described in the following sections. The general methodology is described here. Initial atomic coordinates were extracted from corresponding crystal structures in the Protein Data Bank (1). The standard AMBER force field (ff03) with standard ions (2) was used to describe all protease parameters. Each system was solvated using atomistic TIP3P water (3) and then electrically neutralized with an ionic concentration of 0.15 M NaCl, resulting in fully atomistic, explicit solvent systems each of ∼40,000 atoms. Minimization and equilibration simulations were carried out on a local cluster. The SHAKE algorithm (4) was used on all atoms covalently bonded to a hydrogen atom for stages of each simulation whose time step was greater than 1 fs. The long-range Coulomb interaction was handled using a GPU implementation (5) of the particle mesh Ewald summation method (PME) (6). A nonbonded cutoff distance of 9 Å was used with a switching distance of 7.5 Å for Van der Waals (VdW) interactions. A time step of 4 fs was made possible in all production simulations via the use of the hydrogen mass repartitioning scheme (7) implemented in ACEMD. This scheme takes advantage of the fact that individual atom masses do not appear explicitly in the equilibrium distribution; therefore, changing them only affects the dynamic properties of a system marginally (7) but not the equilibrium distribution. The change in the diffusion coefficient is minimal (10%) and small relative to the approximation that the TIP3 water model makes compared with real water (7).

Experimental accuracy of the molecular simulation protocol for the HIV-1 protease has been validated previously using NMR S^2 order parameters (8). All production simulations were carried out using ACEMD (9). Production-ensemble simulations were deployed for systems E1, E2, C1, and C2 on the GPUGRID compute infrastructure (10). Coordinate snapshots from all production simulations were generated every 100 ps. Initially, 500 runs each for systems E1 and E2 were submitted with a 500-ns limit per trajectory. Because several trajectories were not returned by the server, and some were more advanced than others at the time of analysis, a subset of 417 × 400 ns and 416 × 400 ns was used for the analysis of E1 and E2 sets, respectively. Similarly, a subset of 100 × 100 ns each was used for the analysis of the C1 and C2 control sets.

N-Terminal Extension of HIV-1 Protease. Initial structures were prepared corresponding to a dimeric immature protease with a single 5-aa (residues −5 to −1) N-terminal extension, with a sequence corresponding to the wild-type GagPol p6-PR cleavage site (VSFNF-PQITL), termed N-ter. The following structures were prepared: (*i*) N-ter disassociated with semiopen protease flap conformation (E1); (*ii*) N-ter self-associated in the active site with semiopen-flap conformation (E2); and (*iii*) a self-associated N-ter in the active site with closed-flap conformation (R1).

E1 was prepared as follows: atomic coordinates for wild-type dimeric HIV-1 protease were extracted from the crystal structure of Protein Data Bank (PDB) ID code 1HHP (1). For N-terminal self-association to occur, the immature protease cannot have the native-state conformation, as in the mature protease the N-terminal strand is distal from the active site. Furthermore, N-terminal association implies substantial conformational flexibility of a hinge region (residues 6–8) that connects the N-ter to the first region of downstream secondary structure, beginning at residue P9, which would be present in a folded precursor. Therefore, to allow conformational sampling of the putative N-terminal disassociated conformation, residues 1–8 in the first monomer of the 1HHP crystal structure were deleted, the residue builder in VMD (11) was used to add a preliminary chain from residue 5–8 with the corresponding sequence (VSFNF-PQITLWQR).

E2 and R1 were prepared as follows: The 1HHP structure was aligned (by backbone atoms of the protease excluding the flap residues 43–58 of each monomer) to the 1KJ4 structure containing the MA-CA cleavage site peptide bound to HIV-1 protease in the closed flap conformation. Using VMD, the first monomer N terminus chain (PDB ID code 1HHP for E2; PDB ID code 1KJ4 for R1) was dihedrally rotated by π and −0.1 π radians around the 8–9 and 6–7 C-N peptide bond, respectively, and residues 1–2 were deleted. This allowed the remaining N terminus to be rotated into the active site while avoiding steric clashes. Atomic coordinates for residues 1–7 (relabeled 5–2) of the MA-CA (VSQNY-PIVQ) peptide were connected to residues 3–99 of the first monomer of the respective systems and the MA-CA mutated into the p6-PR cleavage site. A monoprotonated (D25) state was assigned to the catalytic dyad for all three systems (12, 13). Crystallographic water molecules in 1KJ4 were preserved for R1. Water molecules are not present in the 1HHP structure. The final size of the E1, E2, and R1 systems was 39,369, 39,935, and 37,653 atoms, respectively.

Conjugate-gradient minimization was performed for 2,000 steps. During equilibration the position of all heavy protein atoms were restrained by a 10 $kcal \cdot mol^{-1} \cdot Å^{-2}$ spring constant. For systems E2 and R1 in which the connecting residues between the bound peptide and the protease were arbitrarily constructed, no restraints were applied to residues 3–10 from the outset of equilibration to allow for correct conformational reorientation of the linkage region. For all three systems, the hydrogen atoms and water molecules were then allowed to evolve for a total of 500 ps at 300 K to ensure thorough solvation of the system (14). The magnitude of the restraining spring constant was then set to 1 $kcal \cdot mol^{-1} \cdot Å^{-2}$ for 500 ps, then to 0.05 $kcal \cdot mol^{-1} \cdot Å^{-2}$ for another 500 ps, and then, finally, to zero for 500 ps. The temperature was maintained at 300 K using a Langevin thermostat with a low damping constant of 0.1/ps and the pressure maintained at 1 atm. An integration time step of 1 fs was used. The system was finally equilibrated for 6 ns of unrestrained simulation in the isothermal–isobaric ensemble (NPT) with an integration time step of 2 fs. All subsequent simulations were carried out in the canonical ensemble (NVT) with an integration time step of 4 fs. The final coordinates of E1 were used as input for production simulations. Because systems E2 and R1 were deliberately and arbitrarily built into self-associated conformations, further conformational relaxation was necessary. E2 was extended to 100 ns and R1 to 1 μs on a local GPU cluster; the output of these served as the input of subsequent production simulations; the R1 output additionally was the reference structure for subsequent analyses.

Production ensemble simulations were deployed for systems E1 and E2 on the GPUGRID compute infrastructure (10). Initially, 500 runs each for systems E1 and E2 were submitted with a 500-ns limit per trajectory. Because several trajectories were not returned by the server, and some were more advanced than others at the time of analysis, a subset of 417 × 400 ns and 416 × 400 ns was used for the analysis of E1 and E2 sets, respectively.

Control Octapeptide Systems. Initial structures were prepared for octapeptide ligand bound HIV-1 protease complexes for the (*i*) MA-CA (termed C1) and (*ii*) p6-PR (termed C2) cleavage sites.

Atomic coordinates for wild-type dimeric HIV-1 protease were extracted from the crystal structure of PDB ID code 1KJ4 (1). The first residue of the ligand was deleted to produce an octapeptide MA-CA sequence centered on the lytic peptide bond (SQNY-PIVQ) for C1. For C2, the corresponding residues were mutated to produce the p6-PR sequence (SFNF-PQIT). The inactive catalytic dyad D25N was converted into catalytically active D25 form with a monoprotonated state for both controls (12, 13). Crystallographic water molecules in PDB ID code 1KJ4 were preserved for C1 and C2. An additional water molecule was inserted between the lytic peptide bond and the catalytic dyad, as is expected in the general acid/general base (GA/GB) cleavage mechanism (15). The final size of the C1 and C2 was 37,647 and 37,695 atoms, respectively.

The equilibration procedure for C1 and C2 required more stringent considerations because these structures were derived closely from the crystal structure. Conjugate-gradient minimization was performed for 2,000 steps. During equilibration the position of all heavy protein atoms were restrained by a 1 kcal·mol^{-1}·Å^{-2} spring constant. All protein hydrogen atoms and all water molecules (except the catalytic one) were allowed to evolve for 1 ns at 300 K. Then, the restraining constant was set to zero for all atoms except the C_γ atoms of the catalytic dyad, the lytic bond atoms, and the catalytic water oxygen to preserve the geometry of the cleavage region, and the systems evolved for a further 1 ns. The restraining constant was then set to 0.1 kcal·mol^{-1}·Å^{-2} for the catalytic water oxygen and set to zero for all other atoms, and the systems evolved for a further 1 ns. The temperature was maintained at 300 K using a Langevin thermostat with a low damping constant of 0.1/ps, and the pressure was maintained at 1 atm. An integration time step of 2 fs was used. The systems were finally equilibrated for 10 ns of unrestrained simulation in the canonical ensemble (NVT) with an integration time step of 4 fs. The final coordinates of C1 and C2 were used as input for production simulations. All subsequent simulations were carried out in the NVT ensemble.

Production-ensemble simulations were deployed for systems C1 and C2 on the GPUGRID compute infrastructure (10). Initially, 150 runs each for systems C1 and C2 were submitted with a 100-ns limit per trajectory. Coordinate snapshots were generated every 25,000 time steps (100 ps). Because several trajectories were not returned by the server and some were more advanced than others at the time of analysis, a subset of 100 × 100 ns each was used for the analysis of C1 and C2 sets.

SI Conformational Relaxation of Preconstructed N-Terminal Self-Associated Protease

In addition to investigating the self-association process, we also tested the null hypothesis. That is, if autocatalysis of HIV-1 protease does not occur through intramolecular N-terminal self-association, then it follows that a constructed self-associated state with the flaps in a closed conformation and derived from existing crystal structures of octapeptide-bound cleavage complexes should be stoichiometrically unstable. This motivated the construction and relaxation of a closed-flap N-terminal self-associated structure, used as a reference system (termed R1).

Cross-Comparison of Octapeptide-Complexed Control Systems. No atomic resolution structure exists of an N-terminal self-associated HIV-1 protease, nor of the p6-PR octapeptide complex, which shares sequence identity to the N-terminal region. Therefore, R1 was derived from an existing crystal structure of MA-CA cleavage region peptide-bound HIV-1 protease, because of its partial sequence and structural overlap to the p6-PR cleavage region (Fig. S1*A*). It was then necessary to compare the flexibility of the R1 construct with respect to a p6-PR complexed HIV-1 protease which (as it was itself derived from MA-CA), in turn, needed to be compared against an MA-CA complexed system.

Therefore, two sets of control simulations were performed. The first control (C1) was a set of 100 × 100 ns explicit solvent simulations of the MA-CA octapeptide substrate complexed to HIV-1 protease and prepared from the 1KJ4 crystal structure. The second control (C2) was a set of 100 × 100 ns explicit solvent simulations of the p6-PR octapeptide HIV-1 protease complex derived from the crystal structure of the MA-CA complex.

The normalized frequency distribution of the rmsd relative to the C_α atoms of the equilibrated MA-CA ligand is shown for both C1 and C2 (Fig. S1*B*). Both systems exhibit a small rmsd for both the flaps (red) and the octapeptide ligands (cyan). The flap rmsd distribution peaks at 2 and 1.5 Å for the MA-CA (Fig. S1 *B, i*) and p6-PR systems (Fig. S1 *B, ii*), respectively, the latter being slightly sharper. Similarly, ligand rmsd peaks at 2 and 1.8 Å, respectively (Fig. S1 *B, iii* and *iv*), for the two systems, and, again, the p6-PR distribution is sharper.

Analysis of individual residue rmsds relative to the MA-CA structure for both octapeptide systems (Fig. S1*C*) was also performed. Sequence differences in p6-PR compared with MA-CA are highlighted in blue. The peak of all distributions was less than 3 Å, indicating stable fluctuations of each ligand amino acid within the active site. The distribution was not always identical for corresponding amino acids at a given sequence position (P_{-4} and P_1) and, indeed, was sometimes similar for nonidentical amino acids (P_2). Because the basis of differential enzymatic specificity for the different cleavage regions has partial mechanistic roots in the different flexibility of each ligand, it is not surprising that there is both overlap between nonidentical residues and heterogeneity between identical residues. The lack of large scale conformational fluctuations indicate that the p6-PR system sampled an equilibrium distribution over the 10-μs aggregate simulation time.

Comparison of N-Terminal Self-Associated Construct Against Octapeptide-Complexed Control Systems. A single production simulation of R1 was performed on a local GPU cluster for 1 μs to validate the stoichiometric stability of the closed bound state, hypothesized in our study. The flexibility of R1 was then compared against the intrinsic flexibility of the control system C2, consisting of the p6-PR octapeptide, which shares sequence identity with the N-terminal region (Fig. S2*A*).

The normalized frequency distribution of the rmsd relative to the C_α atoms of the equilibrated p6-PR ligand is shown for both C2 and S3 (Fig. S2*B*). The C2 system exhibits peak flap (red) and ligand (cyan) rmsds compared with its own equilibrated structure of 1.5 and 2 Å, respectively. The R1 system exhibits marginally increased rmsds relative to p6-PR, with peak values of 2 and 2.2 Å, respectively, for flaps and the octapeptide cleavage region. Thus, on average, the R1 construct exhibits very similar flexibility to C2.

Analysis of individual residue rmsds relative to the p6-PR equilibrated structure (Fig. S2*C*) reveals almost identical rmsd distributions near the N-terminal positions of the cleavage region (P_{-4} to P_{-2}). The rmsd increases by 2 Å at the P_{-1} and P_1 positions, is very similar again at P_2 and P_3, and differs substantially at P_4. These results are explained by the fact that R1 consists of a cleavage peptide region that is effectively tethered to the frame of the protease via the hinge region, whereas C2 is free. Thus, residues further from the hinge region should exhibit more similar distributions than those that are nearer.

Overall, the analysis of the N-terminal self-association construct (R1) with respect to control systems (C1 and C2) shows that the cleavage peptide region of the construct maintains a stable conformation with cleavable geometry. Based on our analyses, the final structure of the R1 simulation was selected as a reference structure to compare unbiased self-association simulations of systems E1 and E2.

SI MSMs

A discretized MSM of the complete N-ter self-association process was built. The 2D λ_x N-ter rmsd space within the range of access of the simulation data were discretized into 35 × 40 states based on a 1.0-Å^2 bin size. Variation of the bin size did not qualitatively change the outcome of the model. Each data point from the aggregate of ∼335-μs simulation trajectories was assigned to a corresponding state. A reversible transition matrix $T(\tau)$ was then constructed by counting transitions between these states at varying lag times, from which a plot of the implied timescale as a function of the lag time is obtained (Fig. S3). Convergence of the slowest mode of motion occurs at $\tau = 50$ ns and the corresponding transition matrix is used in the subsequent energetic and kinetic analyses. For example, the PMF of the conformational space is obtained from the first eigenvector of the transition matrix $T(\tau)$. This is further validated by performing a Chapman–Kolmogorov test as described in ref. 16 (Fig. S4).

The discretized model is further clustered into a smaller set of coarse-grained metastable states using the PCCA$^+$ method (17). This procedure clusters states into coarse-grained sets for which all constitutive states are more kinetically similar than other sets (i.e., have faster timescales for interconversion). The number of coarse-grained sets was set to seven based on the number of observable metastable states. The relative free energy of each metastable state was determined by integrating the probability distribution across all of the discretized microstates that compose it; the minimum free energy was arbitrarily set to zero.

The PCCA$^+$ method using a seven-state definition does not kinetically distinguish between semiopen and open conformations in the self-associating intermediate structure and, therefore, treats the observed modes of flap-opened association, N-terminal hairpin formation, and lateral threading as the same. This suggests fast interconversion between the two flap states; however, in vivo, the flaps need to open to permit traversal of the GagPol chain into the active site. Lateral threading in our simulations is observed because of the size of the much smaller size of the N-terminal system construct in comparison with a full-length GagPol. Therefore, to compute the error associated with the system setup, we exclude from the analysis all trajectories for which the initial mode of self-association is lateral threading and reconstruct the MSM using the reduced set of trajectories (Fig. S5). Excluding these trajectories makes very little difference to the PMF or to the segregation of states, both properties being qualitatively similar. The total flux computed for the S7→S2 transition is $k_{on} = 11.78 \times 10^{-6}$ ns^{-1}, which is a negligible difference to the original result. This confirms that interconversion between semiopen and open in the intermediate occurs rapidly and suggests that, in vivo, flap opening from an intermediate HIV-1 protease-GagPol structure is rapidly followed by GagPol entry and reformation of the semiopen state.

The total flux was decomposed into pathways along sets of states between S7 to S2. These, together with the percentage contribution of each pathway to the total flux, is listed in Table S1.

We applied a bootstrapping method to analyze the error in both the PMF and the flux calculations. Ten subsets, each containing 400 randomly selected trajectories from of the overall set of 833 trajectories in the E1 and E2 ensemble, were selected. The transition matrix and subsequently the PMF was calculated for each subset. The mean and SD of the PMF is shown in Fig. S6.

We applied the same bootstrapping method to calculate the mean and SD for each of the fluxes from S7 to S2, the total flux and the forward self-association constant. The transition matrices from each of the 10 subsets were used in conjunction with kinetic clustering that had been applied to the entire dataset. The fluxes for the overall dataset, each subset, together with the mean and SD are shown in Tables S2 and S3.

SI Movies

Supporting movies are provided that display various aspects of the HIV-1 protease N-terminal self-association process. Each movie displays a single trajectory from the E1 or E2 run sets and has a duration of between 400 and 500 ns of simulation time. The tertiary structure of the dimeric HIV-1 protease is depicted in white ribbon and white surface, the β-hairpin flaps are in red ribbon, the N-terminal extended region is in cyan surface, and the molecular structure representation and the hinge region are in cyan ribbon.

1. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242.
2. Duan Y, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24(16):1999–2012.
3. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
4. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23:327–341.
5. Harvey M, De Fabritiis G (2009) An implementation of the smooth particle mesh Ewald method on gpu hardware. *J Chem Theory Comput* 5(9):2371–2377.
6. Essmann U, Perera L, Berkowitz ML, Darden T (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–9593.
7. Feenstra K, Hess B, Berendsen H (1999) Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* 20: 786–798.
8. Sadiq SK, De Fabritiis G (2010) Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins* 78(14):2873–2885.
9. Harvey MJ, Giupponi G, Fabritiis GD (2009) Acemd: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* 5(6):1632–1639.
10. Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G (2010) High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model* 50(3):397–403.
11. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38, 27–28.
12. Kovalskyy D, Dubyna V, Mark AE, Kornelyuk A (2005) A molecular dynamics study of the structural stability of HIV-1 protease under physiological conditions: The role of Na+ ions in stabilizing the active site. *Proteins* 58(2):450–458.
13. Wittayanarakul K, Hannongbua S, Feig M (2008) Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J Comput Chem* 29(5):673–685.
14. Meagher KL, Carlson HA (2005) Solvation influences flap collapse in HIV-1 protease. *Proteins* 58(1):119–125.
15. Park H, Suh J, Lee S (2000) Ab initio studies on the catalytic mechanism of aspartic proteinases: Nucleophilic versus general acid/general base mechanism. *J Am Chem Soc* 122(16):3901–3908.
16. Prinz JH, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys* 134(17):174105.
17. Noé F, Horenko I, Schütte C, Smith JC (2007) Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J Chem Phys* 126 (15):155102.

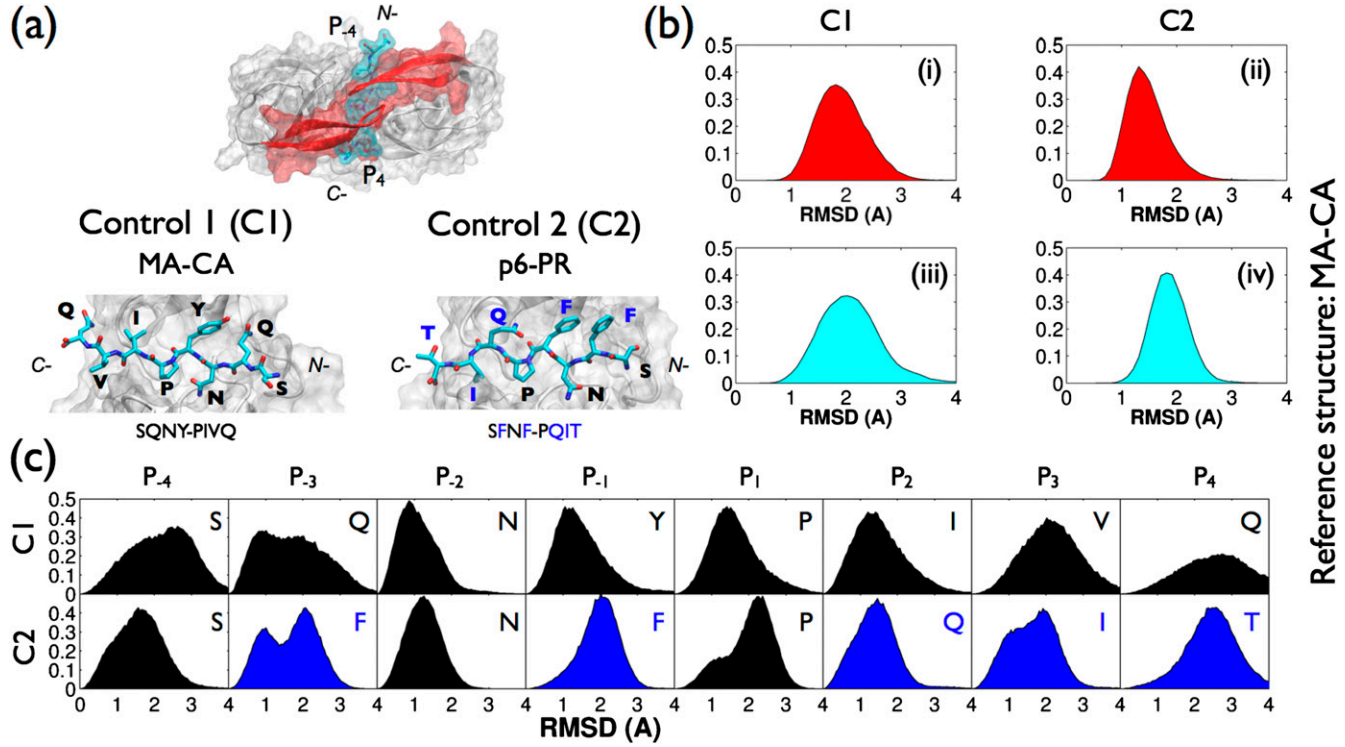


Fig. S1. Comparison of the flexibility of ligand-bound HIV-1 protease for the case of the MA-CA and p6-PR octapeptides. (*A*) Surface representation of ligand-bound HIV-1 protease from an equilibrated structure. The flaps (red) are in a closed conformation over the peptide (cyan). The p6-PR peptide exhibits partial sequence and structural similarity with MA-CA. (*B*) Normalized frequency distribution of C_α rmsd of C1, *i* and *iii*, and C2, *ii* and *iv*, relative to the equilibrated MA-CA–complexed structure for the flaps of the protease (red) and the ligand (cyan). (*C*) Normalized frequency distribution of C_α rmsd for individual residues in each octapeptide ligand in C1 and C2 relative to corresponding residues in the MA-CA–equilibrated structure.

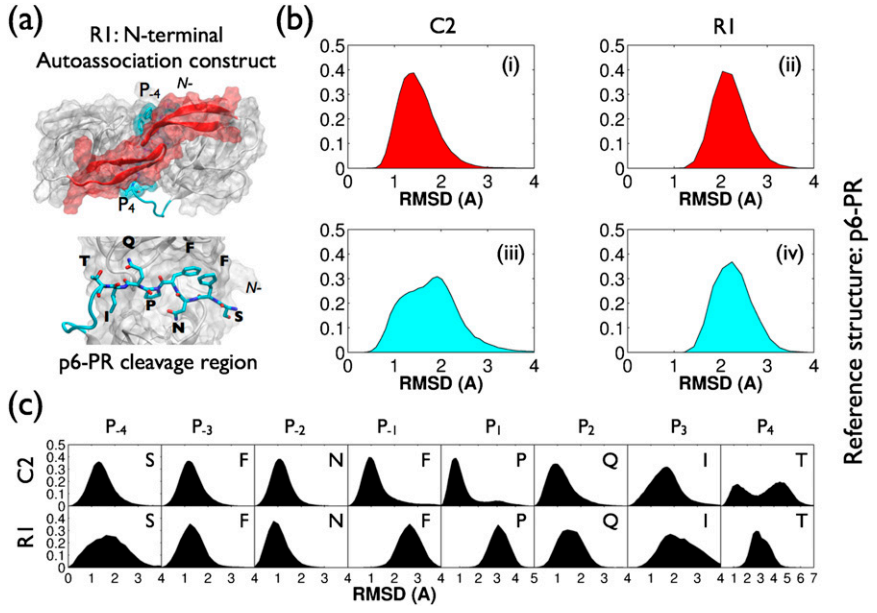


Fig. S2. Comparison of flexibility of N-terminal self-association construct with p6-PR cleavage peptide-bound HIV-1 protease. (*A*) Surface representation of S3. The flaps (red) are in a closed conformation over the N-terminal region (cyan). The N-terminal cleavage region sequence is by definition identical to p6-PR. (*B*) Normalized frequency distribution of C_α rmsd of C2, *i* and *iii*, and S3, *ii* and *iv*, relative to the equilibrated p6-PR–complexed structure for the flaps of the protease (red) and the cleavage region (cyan). (*C*) Normalized frequency distribution of C_α rmsd for individual residues in the cleavage peptide region for C2 and R1 relative to corresponding residues in the p6-PR–equilibrated structure.

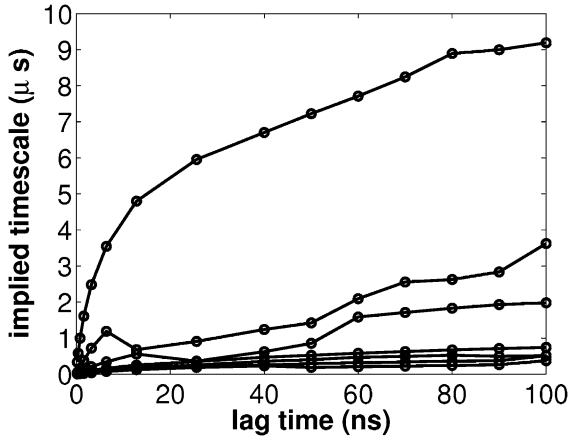


Fig. S3. Plot of implied timescale vs. lag time (τ) for 1,400-state λ_x N-ter rmsd MSM. The implied timescale of the slowest mode shows convergence by 50 ns.

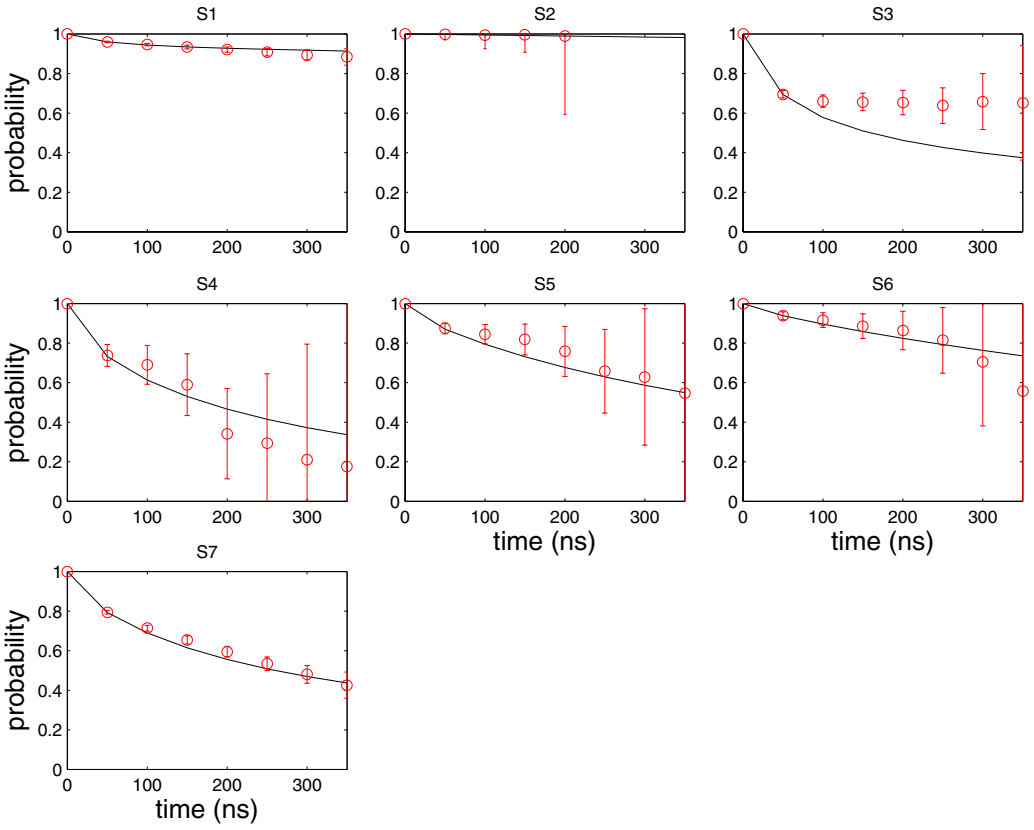


Fig. S4. Chapman–Kolmogorov test validating the Markov model. Each graph shows the evolution of the population of a given metastable state (S1 through S7) when initializing the population in that state. The evolution predicted by the Markov model estimated at a lag time of 50 ns (black solid lines) are compared with the directly observed in trajectory data (red bullets with error bars) over the accessible range of 0–350 ns. Error bars represent the 1-σ statistical uncertainty (68% confidence interval) of the simulation data. The curves are generally in good agreement within 2 σ, indicating that the Markov model is a good approximation of the underlying simulations.

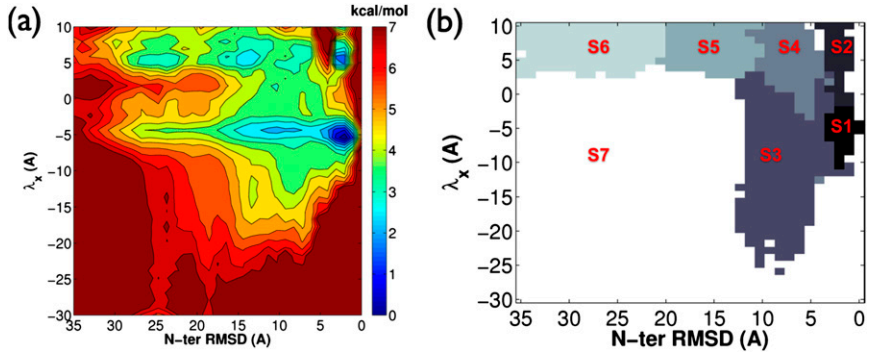


Fig. S5. (*A*) PMF from a MSM in the discretized λ_x N-terminal rmsd projection, excluding trajectories that exhibit threading as the predominant mode of self-association. (*B*) Kinetic clustering of the corresponding discretized λ_x N-terminal rmsd space into seven distinct metastable states, S1 to S7.

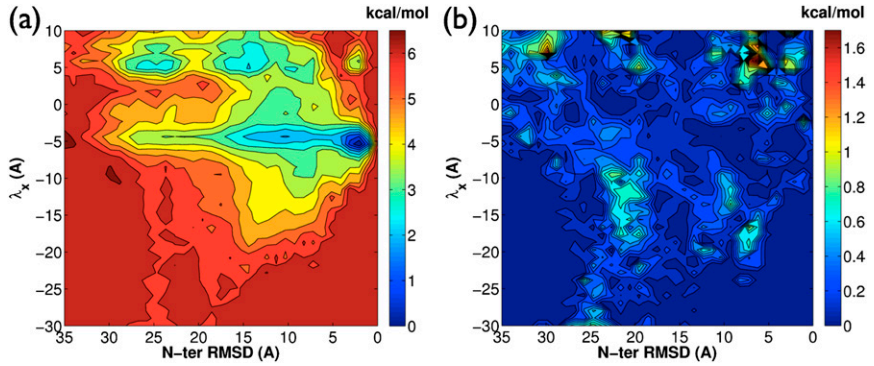


Fig. S6. (*A*) Mean of 10 PMF calculations using randomly selected subsets of the entire dataset. (*B*) SD of the 10 PMF calculations.

Table S1. Flux pathway decomposition for sets of states between S7 and S2

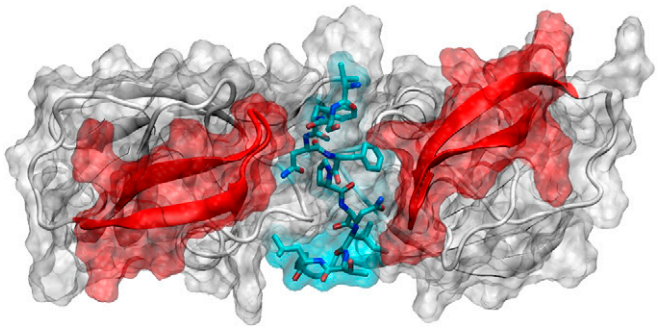
Pathway	Flux, ms^{-1}	Cumulative flux, ms^{-1}	Percentage of total flux	Cumulative percentage of total flux
S7→S3→S1→S2	5.734	5.734	53.59	53.59
S7→S3→S2	1.664	7.398	15.55	69.14
S7→S2	1.022	8.420	9.55	78.69
S7→S1→S2	0.478	8.898	4.47	83.16
S7→S5→S4→S3→S2	0.454	9.352	4.24	87.40
S7→S5→S3→S2	0.382	9.734	3.57	90.97
S7→S4→S2	0.294	10.028	2.75	93.72
S7→S5→S2	0.252	10.280	2.36	96.07
S7→S6→S2	0.228	10.508	2.13	98.21
S7→S6→S3→S2	0.098	10.606	0.92	99.12
S7→S5→S4→S1→S2	0.036	10.642	0.34	99.46
S7→S6→S4→S2	0.022	10.664	0.21	99.66
S7→S6→S5→S4→S2	0.012	10.676	0.11	99.78
S7→S6→S1→S2	0.010	10.686	0.09	99.87
S7→S5→S1→S2	0.008	10.694	0.07	99.94
S7→S5→S4→S2	0.006	10.700	0.06	100.00

Table S2. Total flux decomposition from S7 to S2 for all data and subsets 1–5 used in bootstrapping method

	Flux, ms^{-1}					
Pathway	All data	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
S3→S1	5.728	9.448	5.058	0.489	6.066	5.782
S4→S1	0.036	0.080	0.046	0.005	0.046	0.048
S5→S1	0.008	0.030	0.012	0.003	0.020	0.022
S6→S1	0.010	0.032	0.016	0.003	0.028	0.028
S7→S1	0.478	0.838	0.434	0.071	0.634	0.698
S1→S2	6.256	10.422	5.564	1.480	6.734	6.572
S3→S2	2.598	3.272	2.438	0.269	3.432	3.594
S4→S2	0.334	0.494	0.538	0.317	0.360	0.352
S5→S2	0.252	0.540	0.458	0.240	0.332	0.440
S6→S2	0.238	0.416	0.566	1.753	0.464	0.500
S7→S2	1.022	1.852	2.756	0.478	2.040	2.388
S4→S3	1.018	0.886	0.666	0.020	1.130	0.848
S5→S3	0.382	0.710	0.376	0.077	0.454	0.524
S6→S3	0.098	0.240	0.144	1.693	0.232	0.214
S7→S3	7.398	11.334	6.890	0.314	8.354	8.386
S6→S5	0.012	0.032	1.174	0.204	1.124	1.250

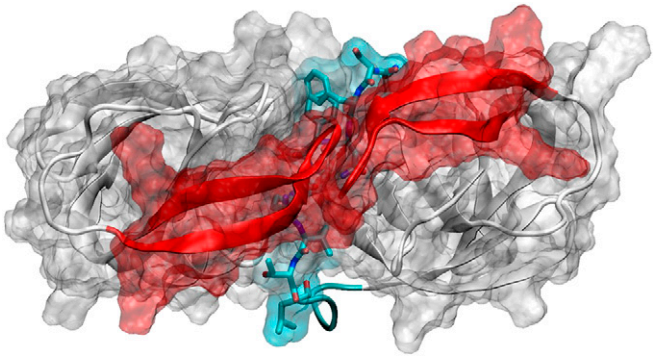
Table S3. Total flux decomposition from S7 to S2 for subsets 6–10 together with mean and SDs of all 10 subsets obtained from bootstrapping method

	Flux, ms^{-1}					
Pathway	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10	Mean (SD)
S3→S1	7.620	6.232	5.968	0.651	5.034	5.235 (2.780)
S4→S1	0.046	0.060	0.042	0.006	0.042	0.042 (0.022)
S5→S1	0.016	0.020	0.016	0.004	0.014	0.016 (0.008)
S6→S1	0.020	0.030	0.018	0.005	0.016	0.020 (0.010)
S7→S1	0.412	0.666	0.230	0.114	0.460	0.456 (0.258)
S1→S2	8.114	7.006	6.264	0.541	5.562	5.826 (2.911)
S3→S2	2.598	3.198	2.822	1.607	2.294	2.552 (1.003)
S4→S2	0.784	0.624	0.442	0.194	0.414	0.452 (0.168)
S5→S2	0.550	0.664	0.396	0.247	0.436	0.430 (0.134)
S6→S2	0.610	0.922	0.394	0.323	0.452	0.640 (0.424)
S7→S2	2.758	3.344	1.726	1.512	2.112	2.097 (0.794)
S4→S3	1.054	1.000	0.982	0.160	0.676	0.742 (0.376)
S5→S3	0.434	0.626	0.280	0.059	0.422	0.396 (0.211)
S6→S3	0.194	0.252	0.158	0.044	0.112	0.328 (0.484)
S7→S3	9.266	8.386	7.870	1.904	6.542	6.925 (3.353)
S6→S5	0.030	0.040	0.042	0.010	0.016	0.066 (0.124)
S7→S5	0.556	0.442	0.446	0.117	0.312	0.333 (0.179)
S5→S4	0.054	0.122	0.014	0.056	0.100	0.119 (0.163)
S6→S4	1.622	1.800	1.166	0.451	1.354	1.018 (0.592)
S7→S4	0.622	0.490	0.502	0.107	0.410	0.473 (0.449)
S7→S6	0.798	1.122	0.608	0.325	0.490	0.644 (0.246)
Total flux	15.414	15.759	12.045	4.423	11.270	11.981 (4.375)
$k_{S7\to S2}$	17.568	17.915	13.767	4.555	12.303	13.787 (5.568)



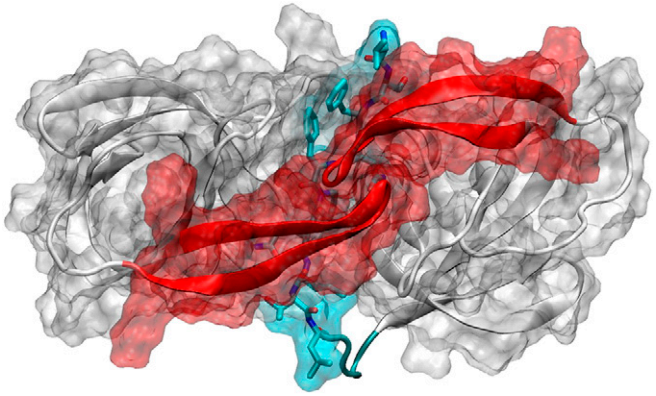
Movie S1. N-terminal self-association to the HIV-1 protease active site via an open-flap mechanism (E1).

Movie S1



Movie S2. N-terminal self-association to the HIV-1 protease active site via a threading mechanism (E1).

Movie S2



Movie S3. Conformational change of an N-terminal self-associated HIV-1 protease into a catalytically viable closed form (E2).

Movie S3