Insights into the Dynamics of HIV-1 Protease: a Kinetic Network Model Constructed from Atomistic Simulations

5/11/2011

*Nan-jie Deng, Weihua Zheng, Emillio Gallicchio and Ronald M. Levy**

BioMaPS Institute for Quantitative Biology and

Department of Chemistry and Chemical Biology,

Rutgers, the State University of New Jersey, Piscataway, NJ 08854

Email: ronlevy@lutece.rutgers.edu

Supporting Information

Kinetic network model and transition path theory

The time evolution of probabilities $\mathbf{P}(t)$ in a discretized state space is governed by the master equation

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{K}\mathbf{P}(t) \quad (1)$$

where $\mathbf{K}$ is the transition rate matrix consisting of $M \times M$ microscopic rate constants K_{ij} for transition from state j to state i. To assign microscopic rate constants K_{ij}, we employ a criteria based on structural similarity,[1-3] i.e. a non-zero transition rate between two given nodes exists if (1) their Cα RMSD is smaller than the cutoff distance of 1.5 Å, or (2) any cluster member (see below) of one node is within the cutoff distance from the other node. The cutoff distance of 1.5 Å corresponds to the average conformational change between sequential pairs of snapshots collected during a time interval of 1.5 ps along continuous REMD trajectories.

To also satisfy the detailed balance condition $K_{ij}p_{eq}(j) = K_{ji}p_{eq}(i)$, we choose

$$K_{ij} = k_{ij}, \; p_{eq}(j) \le p_{eq}(i) \quad (2)$$

$$K_{ij} = \frac{p_{eq}(i)}{p_{eq}(j)} k_{ij}, \; p_{eq}(j) > p_{eq}(i) \quad (3)$$

where k_{ij} is the base rate for transitions between node i and j. Although k_{ij} is generally specific for each pair of nodes and it is possible to determine their values by calibrating against conventional MD simulations, in this study we choose to treat the k_{ij}'s as a single scaling constant k_c. The value of k_c sets the time scale of the network. As mentioned above, the cutoff distance for connecting two structurally similar nodes corresponds to the conformational changes occurred during 1.5 ps time interval. It is therefore reasonable to set the base rate k_c between two connected conformations to $(1.5 \text{ ps})^{-1}$. This is a rough estimate of the basic time scale and

ignores the temperature dependence of the diffusional atomic motions over the replica exchange temperature range (285 K to 340 K).

To study the transitions between the reactant and the product, the nodes on the network are grouped into three subsets, i.e. *A* (reactant), *B* (product) and *I* (intermediate) states. The reactive flux $J_{i\to j}$ for transitions from node *i* to node *j*, can be expressed in terms of the committor probabilities p_{fold} for node *i* and *j*[4,5]

$$J_{i\to j} = K_{ji}\, p_{eq}(i)[p_{fold}(j) - p_{fold}(i)],\ p_{fold}(j) > p_{fold}(i) \tag{4}$$

The $p_{fold}(i)$, a key concept in transition pathways analysis, is defined as the probability for the molecule at node *i* to reach *B* before it reaches *A*. By definition $p_{fold}(i) = 0, i \in A$ and $p_{fold}(i) = 1, i \in B$. For the intermediate states, it can be shown that[4-6]

$$p_{fold}(i) = \sum_{j\neq i} p_{fold}(j) \frac{K_{ji}}{\sum_{k\neq i} K_{ki}},\ i \in I \tag{5}$$

The set of linear equations Eq. (5) is solved numerically to obtain $p_{fold}(i)$, which are substituted into Eq. (4) to yield the reactive flux $J_{i\to j}$ between all pairs of nodes. The total macroscopic reactive flux across an arbitrarily chosen interface that divides *A* and *B* is therefore[4,5]

$$J = \sum_{i\in A^*,\, j\in B^*} J_{i\to j} \tag{6}$$

where *A** and *B** correspond to the states located on the A- and B- sides of the interface, respectively.

The total flux *J* can be decomposed into a set of unidirectional transition pathways P_i, with each pathway contributing J_i to total flux *J*.[4,5] We use the pathway decomposition algorithm[5] in TPT to generate the pathways in descending order of J_i. The algorithm successively identifies the strongest pathway and removes it from the network along with its flux J_i. The key step in the process is to find the bottleneck edge, which is the one associated with the smallest flux $J_{i\to j}$

(min-current) along a reactive pathway. The pathway carrying the largest flux is the one with the largest min-current.

In addition to the calculation of pathways, the computed flux and commitor probabilities allow the estimation of the macroscopic rate constant of transition using the formula[7]

$$k = \frac{J}{\sum_i p_{eq}(i)[1 - p_{fold}(i)]} = \frac{J}{\sum_{i \in A} p_{eq}(i) + \sum_{j \in I} p_{eq}(j)[1 - p_{fold}(j)]} \quad (7)$$

In the two-state limit where the *I*-state population is negligible, Eq. (7) reduces to the familiar formula $k = J/P_{eq}(A)$.

We are interested in computing the equilibrium and kinetic properties at some target temperature T_0. In order to use the data sampled at temperatures other than T_0, each sampled conformation i needs to be reweighted by a weight factor $w_i(T_0)$ given by[8,9]

$$w_i(T_0) = \left\{ \sum_{k=1}^{S} N_k f_k \exp\left[\left(\frac{1}{k_B T_0} - \frac{1}{k_B T_k} \right) E_i \right] \right\}^{-1} \quad (8)$$

where N_k is the number of samples at each of the S replica exchange temperatures T_k, k_B is the Boltzmann constant and E_i the potential energy of sample i. The constants f_k's are related to the partition function of the temperature replica Q_k such that $f_k / f_{k'} = Q_{k'} / Q_k$. The f_k's are calculated from T-WHAM[8] by iteratively solving the WHAM equations. The weight factor $w_i(T_0)$ obtained by Eq. (8) are used to calculate the equilibrium probability $p_{eq}(i)$ for each node in the network.

To build the network, we cluster a total of 96000 conformational snapshots, obtained from REMD at a range of temperatures, into a smaller set of conformational states. Reducing the number of nodes by clustering simplified the tasks of numerically solving large matrix equations Eq. (5) for computing p_{fold}. The clustering is performed based on the Cα-RMSD of the flaps between any pair of replica exchange snapshots, using a cutoff radius of 0.8 Å. This value yields

a manageable number of clustered nodes for further analysis. The neighboring nodes found within the cutoff RMSD from a selected central node are merged to create a large composite node, whose equilibrium probability is the sum of the weights of the neighboring nodes. The resulting composite nodes typically consist of contributions from several original snapshots observed at different temperatures. The resulting kinetic network contains ≈ 32000 nodes and 7.8 $\times 10^6$ edges.

Depending on the number of nodes on the network, the number of pathways calculated using the path decomposition algorithm can be large, e.g. on the order of 10^3-10^4. To obtain mechanistic insights into the transition, it is necessary to group the many pathways into a much smaller set of clustered pathways, with each corresponding to many (e.g. 10^1 - 10^3) similar unclustered pathways. To perform clustering in the pathway space, we use average RMSD between nodes on two pathways as a measure for their dissimilarity, i.e.

$$dist(path_1, path_2) = \frac{1}{2}\left[\frac{1}{N_1}\sum_{i=1}^{N_1} r(i, path_2) + \frac{1}{N_2}\sum_{j=1}^{N_2} r(j, path_1)\right] \quad (9)$$

with

$$r(i, path_2) = \min(RMSD_{il}), i \in path_1, \forall l \in path_2$$

$$r(j, path_1) = \min(RMSD_{jk}), j \in path_2, \forall k \in path_1$$

Here *r(i, path₂)* is the minimum Cα-RMSD between node *i* on pathway 1 and the nodes on pathway 2. We calculate the pair-wise dissimilarities defined by Eq. (9) for all pairs of unclustered pathways and then use hierarchical clustering algorithm to compute the clustered pathways.

Stochastic simulation

Stochastic simulations were carried out on the network using the Gillespie algorithm.[10] In this algorithm, the waiting time at a given node *i* is an exponential random variable whose mean equals the inverse of the sum of the exiting rates from that node. The probability that the system subsequently lands on a connected node *j* is proportional to the microscopic rate from node *i* to node *j*. It can be shown that the algorithm generates realizations of random walks that satisfy the master equation, Eq. (1).[10]

To obtain the transition flux between two macrostates A and B, we start from a randomly chosen node and run a long equilibrium trajectory which contains numerous transition events. The total flux is the number of transition events divided by the total simulation time. We can also record the first passage time (FPT) for each individual transitions and compute the mean first passage time (MFPT). Both total flux and MFPT can be compared with the corresponding TPT calculations using Eq. (6) for total flux and Eq. (7) for transition rate constant, respectively. A more detailed comparison between the stochastic simulation and TPT pathway calculation can be made by projecting each reactive trajectory onto a TPT pathway and examine whether the flux distributions obtained from the two methods agree with each other. First, many reactive trajectories are collected, each containing a distinct transition event from reactant to product. A procedure is then used to associate a reactive trajectory with a structurally similar pathway, using a definition of trajectory-pathway distance similar to Eq. (9)

$$dist(traj_x, path_y) = \frac{1}{N_x}\sum_{i=1}^{N_x} r(i, path_y) \qquad (10)$$

where N_x is the number of nodes on trajectory x, $r(i, path_y)$ is the minimum Cα-RMSD between node i on trajectory x and the nodes on pathway y. A reactive trajectory x is considered to belong to pathway y if $dist(traj_x, path_y) = min\{dist(traj_x, path_z)\}$, $z = 1,\ldots, N_{path}$ and $dist(traj_x, path_y) <$

cutoff. Here the value of *cutoff* is chosen to be 1.6 Å, which corresponds to the lower limit of the distance between any two clustered pathways. Note that recrossing in trajectory is not included in the calculation of trajectory-pathway distance, i.e. if a same node is visited twice along a trajectory, then all the nodes in between are excluded from the distance calculation using Eq. (10).

Replica exchange simulation

REMD simulations were performed on HIV-1 PR using the molecular simulation package IMPACT.[11] Twelve replicas were run in parallel at temperatures between 285 and 340 K for a total of 13 ns for each replica with a time step of 1.5 fs. The last 12 ns were collected for analysis. Exchanges between adjacent temperatures were attempted every 1000 MD steps. The MD trajectories were saved every 1.5 ps. The unliganded semi-open crystal structure 1HHP was used as the starting conformation for all the replicas. The protein in aqueous solution was modeled by the OPLS-AA force field[12] version 2005 and the AGBNP implicit solvent model[13] which includes a novel nonpolar hydration free energy estimator. The AGBNP implicit solvent model has been used in a wide range of studies from protein and peptide folding,[3,14] protein conformational transitions,[15] and protein-ligand interactions.[16] For the current study, we used a modified nonpolar surface tension coefficient of 0.03 kcal mol^{-1} $Å^{-2}$ (reduced from the original value of 0.08 kcal mol^{-1} $Å^{-2}$) in the AGBNP solvent model, which yields physically correct dissociation/unfolding temperatures for the HIV-1 PR dimer. Because the melting temperature of HIV-1 PR ($T_m \approx 333$ K[17]) is within the REMD temperature range, weak structural restraints were used to prevent partial unfolding at the highest temperatures. To choose an appropriate set of restraints, an unrestrained REMD simulation was performed in the same temperature range. From this trial run, we identified the regions that showed the largest displacement from the

native structure. These involved the inter-strand β-sheet residues V11-G16 and I62-I66. Flat-bottomed distance restraints were then applied on 6 native H-bond donor/acceptor pairs to stabilize this β sheet. The distance-restrained atom pairs are: T12O-C67N, K14N-E65O, K14O-E65N, T12'O-C67'N, K14'N-E65'O and K14'O-E65'N. Note that none of the distance restraints are across the dimer interface, and the atoms involved are located at > 28 Å away from the flaps and > 12.5 Å away from the elbow region (residues 39-41). Furthermore, the distance restraint potential used in the simulation is flat-bottomed, with a relatively large width of 6 Å in the flat bottom region. Thus these restraints are only activated in high temperature replicas to prevent unfolding; in fact, they have essentially zero energy contribution at low to medium temperatures. The restraints are therefore not expected to affect the dynamics in the flaps region or the stability of the homodimer except at the highest temperature.

We have examined the 12 continuous walker trajectories spanning the temperature range 285 K to 340 K and found a total of 41 transition events among the three macrostates (semi-open, closed, fully-open) in these trajectories. Among these, four are between semi-open and closed states, and the rest of the transitions are for the semi-open ↔ fully-open and closed ↔ fully-open states. The number of transitions is sufficient to ensure that the reversible conformational transition can be modeled using the kinetic network model constructed from these trajectory data. We have also examined the convergence of the conformational equilibrium between different states by plotting the ratio of the closed and semi-open populations, and the ratio between the fully-open and semi-open states as functions of trajectory segments: see Fig. S6. The result indicates reasonable convergence of the equilibrium populations between the semi-open and closed states, especially after the first three nanoseconds. The relative populations between the semi-open and fully open states show larger fluctuation with simulation time. This is mainly

because the population of the fully open state is small (≈3.5% at 314 K), and thus shows relatively large fluctuations. Taken together, we think that the number of transition events contained in the continuous trajectories and the reasonably convergence in the equilibrium populations provided a good basis for the kinetics calculations.

Figure S1

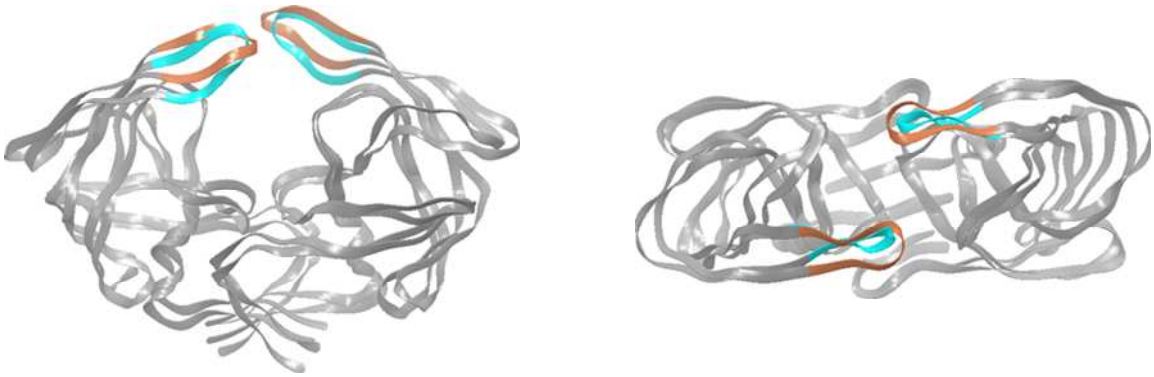

Fig. S1 Crystal structure 1TW7 (blue flaps) superimposed onto an intermediate structure (orange flaps) along the second dominant pathway at T = 285 K. Left: side view. Right: top view.

Figure S2

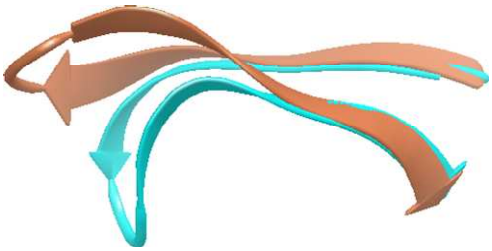

Fig. S2 A bent (blue) flap tip is superimposed onto an uncurled (brown) flap tip.

Fig. S3

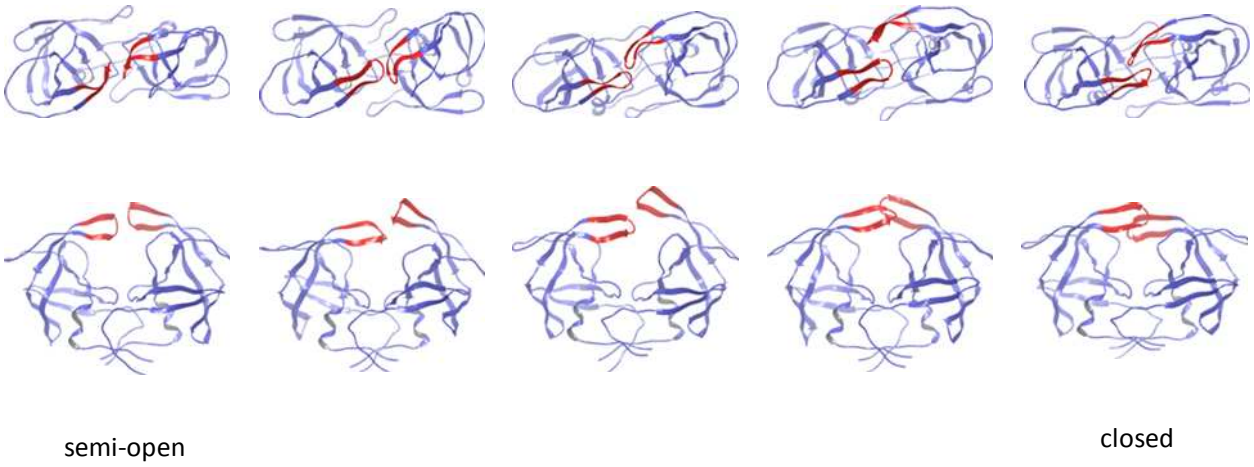

Fig. S3 Intermediate conformations along the pathway for the semi-open ↔ closed transition at T = 285 K. Upper: top view; Lower: side view.

Figure S4(a)

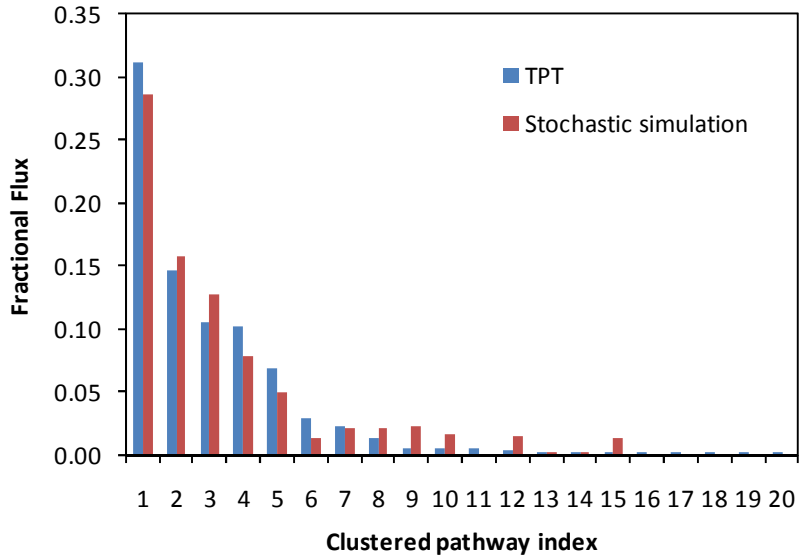

Figure S4(b)

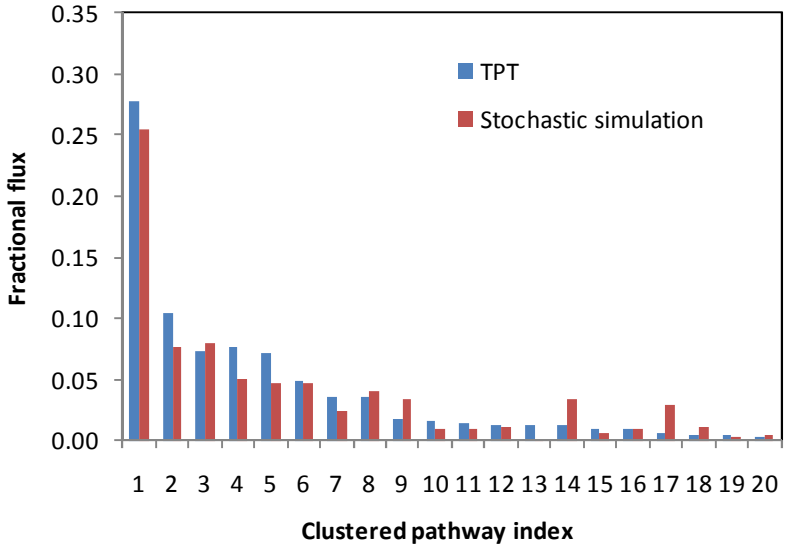

Fig. S4 Flux distributions for the semi-open → closed transition calculated using TPT formula and stochastic simulations. (a) T = 314 K; (b) T = 334 K.

Figure S5(a)

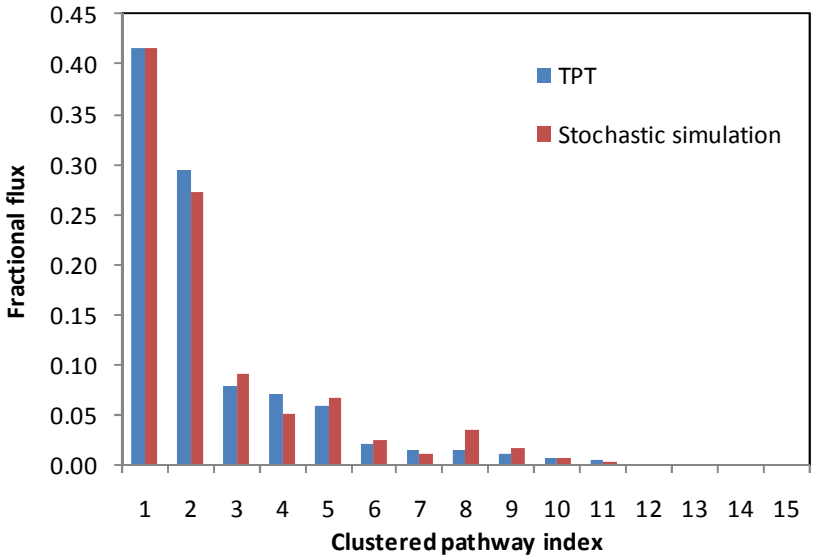

Figure S5(b)

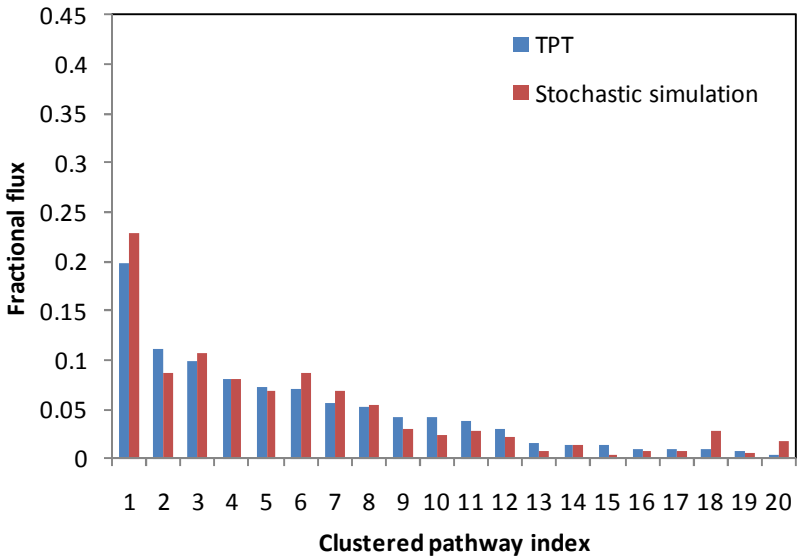

Fig. S5 Flux distributions for the semi-open → fully open transition calculated using TPT formula and stochastic simulations. (a) T = 285 K; (b) T = 334 K.

Fig. S6(a)

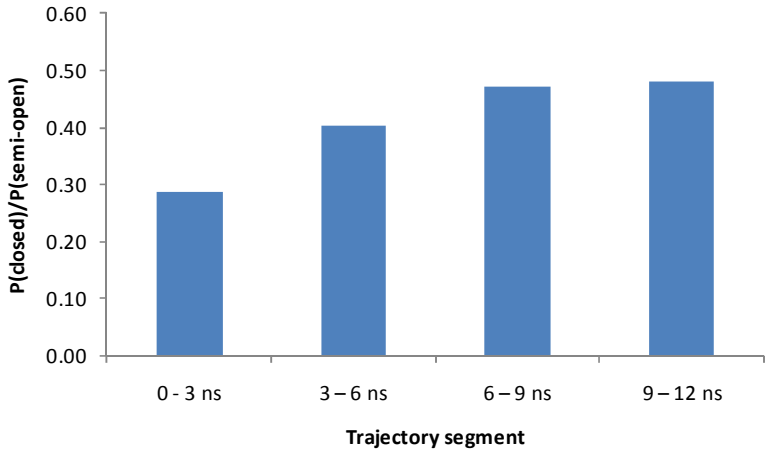

Fig. S6(b)

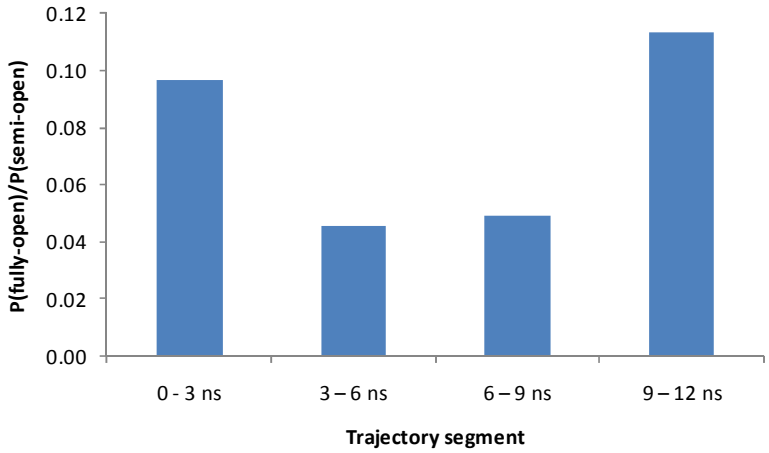

Fig. S6 (a) Ratio of the populations of the closed and semi-open states obtained using different trajectory segments. (b) Ratio of the populations of the fully-open and semi-open states. T = 314 K.

References

1. Ozkan S. B.; Dill K. A.; Bahar I. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model, *Protein Sci*., **2002**, *11*, 1958-1970.

2. Andrec M.; Felts A. K.; Gallicchio E.; Levy R. M. Protein Folding Pathways from Replica Exchange Simulations and a Kinetic Network Model. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 6801-6806.

3. Zheng W; Gallicchio E.; Deng N; Andrec M.; Levy, R. M. Kinetic network study of diversity and temperature dependence of Trp-Cage folding pathways: combining transition path theory and stochastic simulations. *J. Phys. Chem. B*, **2011**, *115*, 1512-1523.

4. Berezhkovskii A.; Hummer G.; Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics, *J. Chem. Phys*., **2009**, *130*, 205102.

5. Metzner P.; Schütte C.; Vanden-Eijnden E. Transition Path Theory for Markov Jump Processes, *Multiscale Model. Simul*. **2009**, *7*, 1192-1219.

6. Singhal N.; Snow C. D.; Pande V. S. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J Chem Phys.* **2004**, *121*, 415-425.

7. Noé F.; Schütte C.; Vanden-Eijndenb E.; Reichc L.; Weikl T. R.; Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations, *Proc. Natl. Acad. Sci. USA*, **2009**, *106*, 19011-19016.

8. Gallicchio E.; Andrec M.; Felts A. K.; Levy, R. M. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B*, **2005**, *109*, 6722-6731.

9. Zheng W.; Andrec M.; Gallicchio E.; Levy R. M. Recovering Kinetics from a Simplified Protein Folding Model Using Replica Exchange Simulations: A Kinetic Network and Effective Stochastic Dynamics, *J. Phys. Chem. B*, **2009**, *113*, 11702-11709.

10. Gillespie D. T., Markov processes: An introduction for physical scientists. Academic Press, Boston, **1992**.

11. Banks, J. L.; Beard H.S.; Cao Y.; Cho A. E.; Damm W.; Farid R.; Felts A. K.; Halgren T. A.; Mainz D. T.; Maple J. R.; Murphy R.; Philipp D. M.; Repasky M. P.; Zhang L. Y.; Berne B. J.; Friesner R. A.; Gallicchio E.; Levy R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem*., **2005,** *26*, 1752-1780.

12. Jorgensen W. L.; Maxwell D. S.; Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids, *J. Am. Chem. Soc*., **1996**, *118*, 11225-11236.

13. Gallicchio E.; Levy R. M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem*., **2004**, *25*, 479-499.

14. Felts A. K.; Harano Y.; Gallicchio E.; Levy R. M. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP

implicit solvent model. *Proteins: Structure, Function, and Bioinformatics*, **2004**, *56*, 310-321.

15. Ravindranathan K. P.; Gallicchio E.; Levy R. M. Conformational Equilibria and Free Energy Profiles for the Allosteric Transition of the Ribose Binding Protein. *J. Mol. Biol.*, **2005**, *353*, 196-210.

16. Ravindranathan K.P.; Gallicchio E.; Friesner R. A.; McDermott A. E.; Levy R. M. Conformational equilibrium of cytochrome P450 BM-3 complexed with N-Palmitoylglycine: A replica exchange molecular dynamics study. *J. Am. Chem. Soc.*, **2006**, *128*, 5786-5791.

17. Todd M. J.; Semo N.; Freire E. The structural stability of the HIV-1 protease, *J. Mol. Biol.*, **1998**, *283*, 475-488.