Monte Carlo Simulations of HIV Capsid Protein Homodimer

Fangqiang Zhu[1*] and Bo Chen[2]

1. Department of Physics, Indiana University - Purdue University Indianapolis, IN, USA

2. Department of Physics, University of Central Florida, Orlando, FL, USA

ABSTRACT

Capsid protein (CA) is the building block of virus coats. To help understand how the HIV CA proteins self-organize into large assemblies of various shapes, here we aim to computationally evaluate the binding affinity and interfaces in a CA homodimer. We model the N- and C-terminal domains (NTD and CTD) of the CA as rigid bodies, and treat the five-residue loop between the two domains as a flexible linker. We adopt a transferrable residue-level coarse-grained energy function to describe the interactions between the protein domains. In seven extensive Monte Carlo simulations with different volumes, a large number of binding / unbinding transitions between the two CA proteins are observed, thus allowing a reliable estimation of the equilibrium probabilities for the dimeric vs. monomeric forms. The obtained dissociation constant for the CA homodimer

* Correspondence: fzhu0@iupui.edu

This is the author's manuscript of the article published in final edited form as:
Zhu, F., & Chen, B. (2015). Monte Carlo Simulations of HIV Capsid Protein Homodimer. Journal of Chemical Information and Modeling, 55(7), 1361–1368. http://doi.org/10.1021/acs.jcim.5b00126

from our simulations, 20-25 μM, is in reasonable agreement with experimental measurement. A wide range of binding interfaces, primarily between the NTDs, are identified in the simulations. Although some observed bound structures here closely resemble the major binding interfaces in the capsid assembly, they are statistically insignificant in our simulation trajectories. Our results suggest that although the general-purpose energy functions adopted here could reasonably reproduce the overall binding affinity for the CA homodimer, further adjustment would be needed to accurately represent the relative strength of individual binding interfaces.

INTRODUCTION

Viruses (such as HIV) enclose their genomes by a protein coat called capsid.[1] The capsid is essential for the life cycle of the virus, and disruption of the capsid by inhibitors could potentially serve to treat virus infections.[2, 3] The mature HIV capsid is formed by multiple copies of a single capsid protein (CA). The structure of CA is largely alpha-helical, with two relatively rigid domains, respectively termed N-terminal domain (NTD) and C-terminal domain (CTD), connected by a short flexible linker. In native HIV particles, the capsid is normally in a cone shape and formed by ~1,500 CA proteins.[4, 5] In in-vitro experiments without the virus genome and other components, the CA proteins may self-assemble into different shapes such as tubes, cones and spheres,[6-9] with variable sizes. The self-assembly of the CA proteins has been extensively studied using X-ray, NMR, and cryo-electron microscopy (Cryo-EM).[10-26] Remarkably, a complete atomic model of the capsid was recently obtained by Cryo-EM combined with all-atom simulations.[17] Other theoretical and computational studies[27-30] also provided important insight into the structure and the assembling process for the HIV capsid.

A multi-protein complex features a number of contacts between pairs of adjacent proteins. The relative stability of a given assembly is primarily determined by the aggregated affinities of these pairwise bindings. To better understand the diverse morphology for the HIV CA assembly, it is thus highly relevant to elucidate the binding interfaces and affinities between two CA proteins. A comparison of the dimeric binding interfaces with those found in the intact capsid could shed important light on the energetics of the assembled structure. In principle, large protein assemblies could be formed either by employing the strongest dimeric binding interfaces or, alternatively, by making a large number of relatively weak protein-protein contacts, especially in a closed structure.

To study the pairwise interactions for CA, the equilibrium between CA monomers and homodimers was previously measured by sedimentation[20] with reported dissociation constant K_d ~18 μM and more recently by NMR experiments[18]. In this study, we employ a computational approach to investigate the binding between two copies of the CA protein.

A complete sampling of protein-protein interaction should allow the two proteins to fully explore their relative positions and orientations, so as to generate a statistically significant number of transitions between the bound and unbound states. Although all-atom simulations have been successfully applied to validate the stability of capsid structures,[17] they would be very inefficient to reveal the spontaneous binding / unbinding of the proteins. In principle, when the bound structure is known, a variety of enhanced sampling methods[31, 32] can be used to compute the free energy difference between the bound and unbound states, as similarly applied in the studies of ligand binding and conformational changes of small peptides[33], as well as in QM/MM calculations[34, 35]. Because the structures of free full-length CA dimers are unknown, however, these free-energy methods are not immediately applicable here, although they could be possibly used to examine the affinity of a given binding interface.

Alternative to all-atom simulations, coarse-grained computational techniques[36, 37] have found a broad range of applications in recent years. Energetic parameters for coarse-grained protein[38-40] and lipids[41] have been developed and refined, and models for protein-protein interaction[42] at the coarse-grained level have been successfully applied in the studies of protein docking[43, 44] and assemblies[45]. Because the NTD and CTD of the CA protein here are relatively rigid, treating them as rigid bodies and ignoring their internal flexibility could significantly simplify the energy model and expedite the sampling. Indeed, to simulate the assembling process for virus capsid, the protein domains have been described by a collection of spheres[27, 29],

cylinders[28] or beads[46]. These coarse-grained models typically incorporate specific experimental information for the assembly, such as the preferred binding interfaces, to design or optimize ad-hoc interaction energy functions and to better mimic the desired protein binding. Their main purpose is therefore to simulate the capsid assembly, rather than to elucidate the thermodynamics of free CA dimers in solution as the focus of this study.

Alternative to the ad-hoc models, transferrable energy functions for general protein-protein interactions are also available. In particular, Kim and Hummer developed and calibrated a residue-level coarse-grained model,[47] which employs effective energy functions that combine physics-based electrostatic interactions and knowledge-based contact potential[48]. The model was shown to well reproduce the experimental binding affinities and interfaces for a variety of weakly bound protein dimers[47] (with K_d >1 μM), and was successfully applied to simulate protein complexes and proteins with disordered segments.[49, 50] The approach thus appears to be also appropriate for calculating the thermodynamics of the CA dimer here. In this study, we perform Monte Carlo (MC) simulations based on this coarse-grained model to identify the binding affinity and interfaces between two CA proteins. As mentioned above, our adopted energy functions are general and transferrable, without any ad-hoc adjustment for the CA. Our calculation thus provides an unbiased assessment of the CA dimer conformations, without incorporating any prior knowledge other than the known structure of CA monomer.

METHODS

In this section, we describe our protein model, energy functions, and simulation protocols.

Protein structure

We adopted the crystal structure[13] 3H47 for the CA protein, which includes residues 1 to 219, with two short loops (6-8 and 88-90) and a 12-residue segment (176-187) missing. We used the program PLOP[51] to add the two missing loops, and built the missing segment (176-187) by taking coordinates from a Cyro-EM structure[17] followed by extensive minimization. In these modeling steps, the coordinates of all existing atoms in the crystal structure[13] were not altered. We then took the coordinates of the 219 C_α atoms (Fig. 1) as the input of our coarse-grained model. To reverse the mutations A14C and E45C introduced in the crystal structure[13], we changed residues 14 and 45 back to Ala and Glu, respectively, with the positions of their C_α atoms unchanged. In our coarse-grained model (Fig. 1), we treat the NTD (1-145) and the CTD (151-219) as two rigid domains, and the linker (146-150) as a flexible chain.

Energy functions

The energy functions adopted in this study were developed and described in details in Ref. 47. For the sake of completeness, we provide a brief description here. The total energy U_{total} of the system (with two or more proteins), as a function of the coordinates of all protein residues (represented by the C_α coordinates), consists of two components:

$$U_{\text{total}} = U_{\text{nonbond}} + U_{\text{bonded}}, \qquad (1)$$

where U_{nonbond} represents the non-bonded interactions (e.g., between residues from different proteins), and U_{bonded} accounts for the conformational energy of the flexible linker.

The non-bonded energy term is the sum of pairwise interactions:[47]

$$U_{\text{nonbond}} = \sum_{i,j} f_i f_j \varphi_{ij}(r_{ij}), \qquad (2)$$

in which r_{ij} is the distance between residues i and j. The summation above goes over all pairs of residues that are not in the same rigid domain and are separated by more than 3 residues if in the same protein, thus excluding the 1-2, 1-3, and 1-4 non-bonded interactions. The f_i and f_j are factors between 0 and 1 that measure the extent of exposure to the solvent for the residue, and are determined by[47]

$$f(s) = \tanh[5\tan(\pi s/2)], \qquad (3)$$

in which s (also between 0 and 1) is the relative solvent-accessible surface area. The value of s (and thus f) for each residue was obtained from the GETAREA online server[52].

The interaction potential $\varphi_{ij}(r)$ in Eq. 2 consists of a Lennard-Jones-type term $u_{ij}^{LJ}(r)$ and an electrostatic energy $u_{ij}^{EL}(r)$:

$$\varphi_{ij}(r) = u_{ij}^{LJ}(r) + u_{ij}^{EL}(r). \qquad (4)$$

The Lennard-Jones energy is in the form of[47]

$$u_{ij}^{LJ}(r) = \begin{cases} 4\varepsilon_{ij}[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^{6}] + 2\varepsilon_{ij} & \text{if } \varepsilon_{ij} > 0 \text{ and } r < 2^{1/6}\sigma_{ij} \\ -4\varepsilon_{ij}[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^{6}] & \text{otherwise} \end{cases}, \qquad (5)$$

where ε_{ij} and σ_{ij} represent the interaction strength and the characteristic distance, respectively. $u_{ij}^{LJ}(r)$ is always repulsive at short ranges, and can be repulsive or attractive at long ranges (Fig. 2A), depending on the sign of ε_{ij}. The interaction distance σ_{ij} is determined by $\sigma_{ij} = (\sigma_i + \sigma_j)/2$

, in which σ_i and σ_j are the van der Waals diameters for the two residues, and are taken from Ref. 47 for each residue type. The interaction strength ε_{ij} is determined by[47] $\varepsilon_{ij} = \lambda(e_{ij} - e_0)$, in which e_{ij} is the Miyazawa-Jernigan contact potential taken from Ref. 48 for each pair of residue types, and the optimal values $\lambda = 0.192$ and $e_0 = -1.85\ k_BT$ (with k_B the Boltzmann constant and T the temperature) were determined in Ref. 47. The electrostatic energy (Fig. 2A) is in a Debye-Hückel form:[47]

$$u_{ij}^{EL}(r) = \frac{q_i q_j}{4\pi\varepsilon r} e^{-r/\zeta}, \qquad (6)$$

in which q_i and q_j are the net charges of the two residues, respectively, and ζ is the Debye screening length. Our adopted ε corresponds to a dielectric constant of 80 for water.[47]

The bonded energy U_{bonded} in Eq. 1 consists of bond, angle, and torsion terms[47] for two, three, and four consecutive residues, respectively. The bond term for two adjacent residues applies a strong harmonic restraint on their distance.[47] In this study, instead, we always fix this distance to the ideal value[47] of 3.81 Å. The rigid bond length adopted here thus eliminated the need for the distance restraint, and improved the acceptance rate in the MC sampling. The bonded energy in this study therefore has two components only:

$$U_{\text{bonded}} = E_{\text{angle}} + E_{\text{torsion}}. \qquad (7)$$

Here E_{angle} is a function of the angle θ formed by three consecutive residues:

$$E_{\text{angle}}(\theta) = -\frac{1}{\gamma}\ln\left(\exp\left\{-\gamma[k_\alpha(\theta-\theta_\alpha)^2 + \varepsilon_\alpha]\right\} + \exp\left[-\gamma k_\beta(\theta-\theta_\beta)^2\right]\right), \qquad (8)$$

with the parameters taken from previous publications[47, 53]: γ = 0.1 mol/kcal; ε_α = 4.3 kcal/mol; θ_α = 1.60 rad; θ_β = 2.27 rad; k_α = 106.4 kcal/(mol·rad^2); k_β = 26.3 kcal/(mol·rad^2). The function $E_{\text{angle}}(\theta)$ features two minima (Fig. 2B), corresponding to the helical and extended secondary structures, respectively.[53] E_{torsion} is a function of the torsional angle φ formed by four consecutive residues:

$$E_{\text{torsion}}(\varphi) = \sum_{n=1}^{4} V_n [1 + \cos(n\varphi - \delta_n)], \tag{9}$$

where the parameters V_n and δ_n depend on the types of the middle two residues, and were taken from previous publications[47, 54]. Only the angle and torsion terms that involve at least one atom in the flexible linker need to be computed.

Monte Carlo moves

We employ four types of trial moves in our MC simulations, as described below.

1. Rigid-body translation of an entire protein. We randomly select one of the proteins, and apply a random translation with the x, y, and z displacements each chosen from a random number in the range of [-0.25 Å, 0.25 Å]. This trial move does not change the internal conformation or the orientation of the protein.

2. Rigid-body rotation of an entire protein. We randomly select one of the proteins, and rotate it as a rigid body around a random axis through the protein center and by a random angle smaller than 0.2 rad. This trial move does not change the internal conformation or the center of the protein.

3. Domain rotation. We randomly select one rigid domain (NTD or CTD) to apply a small random rotation, with the rotation center at the atom in the flexible linker immediately adjacent to the chosen domain. This trial move thus only changes the coordinates of the selected domain. In addition, all of the bond lengths remain unchanged in this operation.

4. Flexible linker move. We randomly select one atom (residue) on the flexible linker and change its coordinates. Specifically, we rotate the chosen atom around the line connecting its two nearest neighbors, by an angle in the range of [-0.1 rad, 0.1 rad]. This trial move thus will not change the distance between the atom and its nearest neighbors, ensuring that all bond lengths involving the atoms in the flexible linker remain strictly fixed at the ideal value[47] of 3.81 Å in our simulations, as mentioned earlier.

All possible system configurations of the proteins can in principle be accessed by a combination of the four MC moves above.

Simulation details

Our simulation system contains two copies of the CA protein. We assigned the net charge of each protein residue according to the standard protonation state at pH 7, i.e., +1 for Arg and Lys, and -1 for Asp and Glu. We also assigned a charge of +0.5 for the His residues.[47] We took ζ = 10 Å for the Debye screening length in Eq. 6, corresponding to salt concentrations of ~100 mM.[47] We carried out seven MC simulations at a constant temperature of 300 K and under the periodic boundary conditions. The periodic system has a cubic unit cell with the length ranging from 160 Å to 640 Å in the seven simulations (Table 1). When calculating the pairwise non-bonded interaction (Eq. 2) between two residues i and j, we took the closest distance among all periodic images as the inter-residue distance r_{ij}. Each MC simulation started with the two proteins in

random positions and orientations, and was run for 10^9 steps. In each MC step, we randomly choose (with equal probability) one of the four types of trial moves described earlier. We then calculate the change of the total energy due to the attempted trial move, and either accept or reject the move according to the Metropolis criterion. Each simulation was run on a single AMD Opteron processor at 2.6 GHz, and took ~65 days to finish 10^9 MC steps.

RESULTS

In each MC simulation, the two CA proteins explored a wide range of configurations. In particular, we observed a large number of transitions between the bound and unbound states, thus enabling a statistically reliable estimate of the thermodynamics for the protein-protein binding. In addition, due to different conformations of the flexible linker, the NTD and CTD of the same protein explored a diverse ensemble of relative positions and orientations, as also observed from recent NMR experiments.[18]

To analyze the relative positions of the two proteins, we introduce a contact strength based on the distance r_{ij} between two residues from different proteins. The contact strength for the residue pair is assigned 1 if $r_{ij} \leq 5$ Å or 0 if $r_{ij} \geq 8$ Å; when 5 Å < r_{ij} < 8 Å, the contact strength takes a value between 0 and 1 calculated by integrating a truncated Gaussian function. The list of contact strengths for all pairs of residues thus characterizes the protein-protein interaction pattern. For a given snapshot from the simulation, the sum of these contact strengths represents the effective number of contacting residue pairs and quantifies the extent of the contact between the two proteins. These inter-protein contacts can be further classified into domain contacts, i.e., those

between the two NTDs, between an NTD and a CTD, and between the two CTDs. We performed this analysis for all simulation trajectories and obtained the average proportions for each class of contacts (Table 1). In each of the seven MC simulations, the NTD-NTD, NTD-CTD, and CTD-CTD interactions account for ~60%, ~35%, and ~5% of the identified residue contacts, respectively (Table 1). Therefore, the bound state in our simulations primarily arises from the interactions between the NTDs in the two proteins.

The statistics of the domain-domain contacts (Table 1) also offers an evaluation for the convergence of the sampling. Although the seven MC simulations were performed under different volumes, the proportion for each type of domain contacts should ideally be identical across the simulations. Indeed, the percentages (in the parentheses of Table 1) for each individual type are roughly similar in our seven simulations, with relative variations all below 50%. Furthermore, because the two copies of the protein are identical, the contact numbers for NTD1-CTD2 and for CTD1-NTD2 should be asymptotically identical in each simulation. The values from our simulations (Table 1) indeed agree reasonably with this expectation as well. Overall, we thus consider the convergence in our simulations satisfactory albeit not perfect.

Binding affinity

As described in Supporting Information, the equilibrium binding probability for the two proteins in the simulations is directly related to the dissociation constant K_d in macroscopic systems. Our simulations can thus be used to estimate the binding affinity of the CA homodimer. Specifically, we define the bound and unbound states based on the contact strength introduced earlier. For each frame in the simulation trajectory, we sum up the contact strengths of all residue pairs between the two proteins, and classify it as a bound state if the total strength is larger than 1 or otherwise an

unbound state. The equilibrium probability p_b for the bound state is then taken as the proportion of the simulation frames assigned to this state in the trajectory, with $p_u = 1 - p_b$ the probability for the unbound state.

We applied two numerical methods to estimate the dissociation constant K_d from the bound / unbound probabilities. The first method is based on Eq. S17, which indicates that the ratio p_u / p_b is linearly proportional to the volume V_0 of the simulation system, with $K_d / 2$ the linear coefficient. By doing a linear fit (Fig. 3A) through data points for the obtained p_u / p_b ratios and the corresponding volumes from the seven MC simulations, we obtained a dissociation constant $K_d \approx 25$ μM. Our second method is based on Eq. S18 (an equivalent form of Eq. S17), which provides the functional dependence of the binding probability p_b on the effective protein concentration ($2/V_0$) in the simulation.[47] We applied a nonlinear fit (Fig. 3B) for the same set of probability data and obtained $K_d \approx 20$ μM. Given the statistical uncertainty of the data (as indicated by the error bars), the K_d values from the two methods are close to each other. In comparison, the experimental value for K_d is 18 μM from sedimentation measurement.[20] Recent NMR experiments[18] reported K_d values of 20 μM at 20 °C and 40 μM at 25 °C (which is close to 300 K in our simulations) for the CA dimer. Based on these experimental data, our calculated dissociation constant is in the correct order of magnitude.

Binding modes

We observed a very diverse set of binding modes in the simulations. The two CA proteins were found to make contacts through a large variety of interfaces, although the majority of the binding

interfaces involve contacts between the two NTDs, as mentioned earlier. To cluster the configurations of the two proteins in the simulation trajectory into different binding modes,[47] we applied a new method here to describe the binding poses, based on the relative orientation between the protein domains.

For any two rigid domains A and B from different proteins, we perform a rigid-body rotation such that A is in its reference orientation, and then use a unit vector $\hat{r}_{AB}$ to represent the direction from the center of A to the center of B. $\hat{r}_{AB}$ thus denotes the direction of domain B's center in domain A's reference frame. We similarly use a unit vector $\hat{r}_{BA}$ to represent A's center in B's reference frame. We then join the two vectors into an orientation vector $\begin{pmatrix} \hat{r}_{AB} \\ \hat{r}_{BA} \end{pmatrix}$, which represents the relative orientations between domains A and B. Moreover, we also calculate the total contact strength (defined earlier), n, between the two domains, and scale the orientation vector by n when $n < 1$. Under this treatment, the orientation vector will have a reduced length when the two domains are only in weak contact, and will be zero when the two are not in contact. We note that in this representation, a torsional rotation around the axis passing the two domain centers will not change the orientation vector, and this degree of freedom is thus ignored. We have four such orientation vectors to describe the NTD-NTD, NTD-CTD, CTD-NTD, and CTD-CTD contacts between the two proteins, respectively. The combined four vectors, with a total of 24 elements, thus represent the relative pose of the two proteins. In this description, the two proteins are treated in a symmetric manner.

For each simulation frame, we calculated the 24-element vector as described above. We then applied the k-means algorithm to partition all vectors from the simulation trajectory into

clusters in this 24-dimensional space. Because the two proteins are identical, when calculating the distance of a vector to a cluster center in the k-means algorithm, we swap the coordinates of the two proteins if the distance becomes shorter after the swap. All unbound structures correspond to a zero vector and thus naturally form a cluster at the origin. The k-means algorithm requires the number of clusters, k, as an input parameter. We experimented with a range of k values, but could not identify a perfect case in which all clusters are well separated from each other. This is consistent with our observation (through visual inspection) of a wide and continuous spectrum of binding interfaces from the simulation trajectories.

Figure 4A shows representative structures for the six bound-state clusters (with one additional cluster corresponding to the unbound state) when $k = 7$ is used in the k-means algorithm, which appears to yield better separations between the clusters in comparison to other k values. It is clear from the figure that the clusters are quite distinct from each other in terms of the contacts and the relative orientations between the two proteins, thus demonstrating the diverse binding patterns in the simulation. Furthermore, the energy for each member in the cluster and its root-mean-square-deviation (RMSD) to the representative structure are shown in the scatter plots (Fig. 4B). Due to the presence of intra-protein interactions, the system energy is generally not zero even when the two proteins are completely unbound. As expected, the energies of most bound structures are significantly lower than the average energy (Fig. 4B, *dashed lines*) of the unbound state. We also note that the majority of the members in each cluster have quite large RMSDs to the representative structure, and that the RMSDs do not appear to correlate well with the energies of the structures. This is mainly because the clusters are identified based on the relative orientation of the contacting domains as described earlier, and not on the RMSDs. For example, when the two NTDs are in close contact, the dangling CTDs may still adopt a variety of positions and

orientations, which would give rise to a wide range of dimer RMSDs without significantly affecting the interaction energy or the cluster assignment.

Comparison to assembled structure

Our simulation system consists of only two CA proteins, whereas in functional virus particles and in-vitro experiments, many copies of the protein form a large assembly with well-defined geometry. It is thus of interest to compare the binding interfaces in the isolated CA dimer vis-à-vis in the assembled structures. When the CA proteins assemble into viral capsid or helical tubes, there are four major interfaces between neighboring proteins:[17] (1) an interface between two NTDs that allows six CAs to form a hexameric ring;[25] (2) an interface between an NTD and a CTD of the neighboring CAs in the hexamer;[25] (3) an interface between two CTDs that connects adjacent CA hexamers;[25] and (4) a trimer interface where the CTDs from three different CA hexamers meet.[14] Although our simulations all started with the two CA proteins in random configurations and far apart from each other, we observed all four types of binding interfaces during the simulations. Figure 5 shows that some dimer structures from our simulations agree well with the average NTD-NTD, NTD-CTD and CTD-CTD contacts resolved in the tubular CA assembly[17], with the dimer RMSDs all below 1.8 Å. Although the full CTD trimer interface cannot be reproduced with only two CA proteins in our simulations, some snapshots nonetheless superimpose well (RMSD 1.4 Å) with two of the CTDs in the trimer (Fig. 5D). However, although all major binding interfaces in the capsid assembly were indeed sampled in our simulations, they only represent a very small fraction of the observed bound structures. The vast majority of the binding interfaces (Fig. 4) in our simulations do not resemble the protein contacts found in the assembled structures.

DISCUSSION

In this study, we employ a general transferrable energy model[47] along with MC sampling to characterize the interactions between two CA proteins. We model each CA as two rigid domains (NTD and CTD) connected by a flexible linker. Our simulations are sufficiently long to sample large numbers of transitions between the bound and unbound states for the protein pair, and the statistics from the simulations are generally consistent and satisfactory. In particular, the calculated dissociation constant K_d for the CA dimer from our simulations is in the correct order of magnitude in comparison with the experimental measurements[18, 20].

However, the binding modes observed in our simulations are surprisingly diverse. As shown in Fig. 4, the two proteins may bind in very different orientations, and none of the binding modes is predominant. Recent experiments[18] also indicated that CA dimers in solution are much more flexible and dynamic than found in the assembled structures. It is thus plausible that the individual binding interfaces in the CA capsid are not particularly strong and not necessarily favored in isolated CA dimers; instead, the protein contacts in the closed assembly may be stabilized by cooperative binding, as they are geometrically compatible with the packing of multiple proteins in a repetitive pattern. The relatively weak binding interfaces would also give rise to substantial flexibility and consequently the observed morphology in the CA assemblies.

Although the high diversity of CA dimer conformations in our simulations is consistent with experimental findings, the details of the binding modes are not in agreement with experiments. Most notably, the majority of experiments strongly indicated that the major contact in the CA dimer is at the CTD-CTD interface.[18, 20] In contrast, only a very small population of our

bound structures features the expected CTD-CTD binding site, and most of the observed protein contacts in our simulations are instead between the NTDs. We note that the CTD-CTD interface was indeed sampled in our simulations (Fig. 5C) but did not exhibit strong affinity. Therefore, the discrepancy is likely due to the protein interaction model here rather than insufficient sampling. One possible cause may be the crystal structure adopted in this study, in which some mutations were introduced to the CA protein to abolish the CTD-CTD contact and to induce the formation of the NTD hexameric ring.[13] Although all of the involved residues were mutated back to the wild types and the critical missing loop in the CTD was rebuilt in our simulation system (see Methods), the CTD conformation here might nonetheless be less optimal for forming the desired binding interface. In addition, our calculated binding affinity, with a dissociation constant of 20-25 μM, is somewhat stronger than the experimental measurement[18] (40 μM at 25 °C). It is thus likely that the NTD-NTD interaction in our model is over-estimated, and the excessive tendency to form the NTD-NTD interfaces further interferes with the potential binding between the CTDs. In light of these problems, although the model could in principle reflect the energetic variations by point mutations, we do not expect it to correctly predict the experimental mutational effect for the CA dimer at this stage, when the involved binding interface has not been reproduced with the correct affinity yet.

Given the general energy model adopted here without any ad-hoc optimization specific to the CA protein, it is probably not surprising that some subtle energetic balance between different binding modes is not accurately captured in the coarse-grained energy functions and in our simulations. Earlier studies[29] demonstrated that incorporation of known experimental information could help improve the energy function for CA. Based on the findings here, our energy model

could be refined, e.g., by making the NTD-NTD interaction weaker and the CTD-CTD interaction stronger at the conserved W184 / M185 site.

CONCLUSION

In this study, a general transferrable coarse-grained model[47] was used in MC simulations to characterize the binding for a CA homodimer. The statistical errors in our sampling appear to be modest, indicating a satisfactory convergence. The overall binding affinity for the homodimer calculated from our simulations is also in reasonable agreement with experimental measurements. In addition, major binding interfaces in the intact CA capsid are observed in the sampled structures in the simulations. The most frequent binding modes emerging from the simulations, however, do not agree with experiments, and we attribute the discrepancy primarily to the underlying energy model adopted here. We propose that incorporating CA-specific modifications into the general-purpose energy functions could help reproduce the desired binding mode. When the thermodynamics and the binding interfaces for the CA homodimer can be faithfully reproduced, the refined model may then be used to simulate the assembling of larger capsid complex in future studies.

ACKNOWLEDGEMENT

This work was supported by the Air Force Office of Scientific Research with grant number FA9550-13-1-0150. The authors declare no competing financial interest.

Supporting Information Available: a theoretical derivation for the relations between the binding probabilities in the simulations and the binding affinity measured in experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P., *Molecular Biology of the Cell*. 5th ed.; Garland Science: New York, 2008.

2. Prevelige, P. E., Jr., New Approaches for Antiviral Targeting of Hiv Assembly. *J Mol Biol* **2011,** 410, 634-640.

3. Engelman, A.; Cherepanov, P., The Structural Biology of Hiv-1: Mechanistic and Therapeutic Insights. *Nat Rev Microbiol* **2012,** 10, 279-290.

4. Ganser, B. K.; Li, S.; Klishko, V. Y.; Finch, J. T.; Sundquist, W. I., Assembly and Analysis of Conical Models for the Hiv-1 Core. *Science* **1999,** 283, 80-83.

5. Briggs, J. A.; Wilk, T.; Welker, R.; Krausslich, H. G.; Fuller, S. D., Structural Organization of Authentic, Mature Hiv-1 Virions and Cores. *EMBO J* **2003,** 22, 1707-1715.

6. Ehrlich, L. S.; Agresta, B. E.; Carter, C. A., Assembly of Recombinant Human Immunodeficiency Virus Type 1 Capsid Protein in Vitro. *J Virol* **1992,** 66, 4874-4883.

7. Li, S.; Hill, C. P.; Sundquist, W. I.; Finch, J. T., Image Reconstructions of Helical Assemblies of the Hiv-1 Ca Protein. *Nature* **2000,** 407, 409-413.

8. Ehrlich, L. S.; Liu, T.; Scarlata, S.; Chu, B.; Carter, C. A., Hiv-1 Capsid Protein Forms Spherical (Immature-Like) and Tubular (Mature-Like) Particles in Vitro: Structure Switching by Ph-Induced Conformational Changes. *Biophys J* **2001,** 81, 586-594.

9. Ganser-Pornillos, B. K.; von Schwedler, U. K.; Stray, K. M.; Aiken, C.; Sundquist, W. I., Assembly Properties of the Human Immunodeficiency Virus Type 1 Ca Protein. *J Virol* **2004,** 78, 2545-2552.

10. Gitti, R. K.; Lee, B. M.; Walker, J.; Summers, M. F.; Yoo, S.; Sundquist, W. I., Structure of the Amino-Terminal Core Domain of the Hiv-1 Capsid Protein. *Science* **1996,** 273, 231-235.

11. Momany, C.; Kovari, L. C.; Prongay, A. J.; Keller, W.; Gitti, R. K.; Lee, B. M.; Gorbalenya, A. E.; Tong, L.; McClure, J.; Ehrlich, L. S.; Summers, M. F.; Carter, C.; Rossmann, M. G., Crystal Structure of Dimeric Hiv-1 Capsid Protein. *Nat Struct Biol* **1996,** 3, 763-770.

12. Gamble, T. R.; Vajdos, F. F.; Yoo, S.; Worthylake, D. K.; Houseweart, M.; Sundquist, W. I.; Hill, C. P., Crystal Structure of Human Cyclophilin a Bound to the Amino-Terminal Domain of Hiv-1 Capsid. *Cell* **1996,** 87, 1285-1294.

13. Pornillos, O.; Ganser-Pornillos, B. K.; Kelly, B. N.; Hua, Y.; Whitby, F. G.; Stout, C. D.; Sundquist, W. I.; Hill, C. P.; Yeager, M., X-Ray Structures of the Hexameric Building Block of the Hiv Capsid. *Cell* **2009,** 137, 1282-1292.

14. Byeon, I. J.; Meng, X.; Jung, J.; Zhao, G.; Yang, R.; Ahn, J.; Shi, J.; Concel, J.; Aiken, C.; Zhang, P.; Gronenborn, A. M., Structural Convergence between Cryo-Em and Nmr Reveals Intersubunit Interactions Critical for Hiv-1 Capsid Function. *Cell* **2009,** 139, 780-790.

15. Chen, B.; Tycko, R., Structural and Dynamical Characterization of Tubular Hiv-1 Capsid Protein Assemblies by Solid State Nuclear Magnetic Resonance and Electron Microscopy. *Protein Sci* **2010,** 19, 716-730.

16. Byeon, I. J.; Hou, G.; Han, Y.; Suiter, C. L.; Ahn, J.; Jung, J.; Byeon, C. H.; Gronenborn, A. M.; Polenova, T., Motions on the Millisecond Time Scale and Multiple Conformations of Hiv-1 Capsid Protein: Implications for Structural Polymorphism of Ca Assemblies. *J Am Chem Soc* **2012,** 134, 6455-6466.

17. Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P., Mature Hiv-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics. *Nature* **2013,** 497, 643-646.

18. Deshmukh, L.; Schwieters, C. D.; Grishaev, A.; Ghirlando, R.; Baber, J. L.; Clore, G. M., Structure and Dynamics of Full-Length Hiv-1 Capsid Protein in Solution. *J Am Chem Soc* **2013,** 135, 16133-16147.

19. Bayro, M. J.; Chen, B.; Yau, W. M.; Tycko, R., Site-Specific Structural Variations Accompanying Tubular Assembly of the Hiv-1 Capsid Protein. *J Mol Biol* **2014,** 426, 1109-1127.

20. Gamble, T. R.; Yoo, S.; Vajdos, F. F.; von Schwedler, U. K.; Worthylake, D. K.; Wang, H.; McCutcheon, J. P.; Sundquist, W. I.; Hill, C. P., Structure of the Carboxyl-Terminal Dimerization Domain of the Hiv-1 Capsid Protein. *Science* **1997,** 278, 849-853.

21. Worthylake, D. K.; Wang, H.; Yoo, S.; Sundquist, W. I.; Hill, C. P., Structures of the Hiv-1 Capsid Protein Dimerization Domain at 2.6 a Resolution. *Acta Crystallogr D Biol Crystallogr* **1999,** 55, 85-92.

22. Pornillos, O.; Ganser-Pornillos, B. K.; Yeager, M., Atomic-Level Modelling of the Hiv Capsid. *Nature* **2011,** 469, 424-427.

23. Alcaraz, L. A.; del Alamo, M.; Barrera, F. N.; Mateu, M. G.; Neira, J. L., Flexibility in Hiv-1 Assembly Subunits: Solution Structure of the Monomeric C-Terminal Domain of the Capsid Protein. *Biophys J* **2007,** 93, 1264-1276.

24. Alcaraz, L. A.; Del Alamo, M.; Mateu, M. G.; Neira, J. L., Structural Mobility of the Monomeric C-Terminal Domain of the Hiv-1 Capsid Protein. *FEBS J* **2008,** 275, 3299-3311.

25. Ganser-Pornillos, B. K.; Cheng, A.; Yeager, M., Structure of Full-Length Hiv-1 Ca: A Model for the Mature Capsid Lattice. *Cell* **2007,** 131, 70-79.

26. Schur, F. K.; Hagen, W. J.; Rumlova, M.; Ruml, T.; Muller, B.; Krausslich, H. G.; Briggs, J. A., Structure of the Immature Hiv-1 Capsid in Intact Virus Particles at 8.8 a Resolution. *Nature* **2015,** 517, 505-508.

27. Krishna, V.; Ayton, G. S.; Voth, G. A., Role of Protein Interactions in Defining Hiv-1 Viral Capsid Shape and Stability: A Coarse-Grained Analysis. *Biophys J* **2010,** 98, 18-26.

28. Chen, B.; Tycko, R., Simulated Self-Assembly of the Hiv-1 Capsid: Protein Shape and Native Contacts Are Sufficient for Two-Dimensional Lattice Formation. *Biophys J* **2011,** 100, 3035-3044.

29. Grime, J. M.; Voth, G. A., Early Stages of the Hiv-1 Capsid Protein Lattice Formation. *Biophys J* **2012,** 103, 1774-1783.

30. Tsiang, M.; Niedziela-Majka, A.; Hung, M.; Jin, D.; Hu, E.; Yant, S.; Samuel, D.; Liu, X.; Sakowicz, R., A Trimer of Dimers Is the Basic Building Block for Human Immunodeficiency Virus-1 Capsid Assembly. *Biochemistry* **2012,** 51, 4416-4428.

31. Christ, C. D.; van Gunsteren, W. F., Multiple Free Energies from a Single Simulation: Extending Enveloping Distribution Sampling to Nonoverlapping Phase-Space Distributions. *J Chem Phys* **2008,** 128, 174112.

32. Riniker, S.; Christ, C. D.; Hansen, H. S.; Hunenberger, P. H.; Oostenbrink, C.; Steiner, D.; van Gunsteren, W. F., Calculation of Relative Free Energies for Ligand-Protein Binding, Solvation, and Conformational Transitions Using the Gromos Software. *J Phys Chem B* **2011,** 115, 13570-13577.

33. Lin, Z.; Timmerscheidt, T. A.; van Gunsteren, W. F., Using Enveloping Distribution Sampling to Compute the Free Enthalpy Difference between Right- and Left-Handed Helices of a Beta-Peptide in Solution. *J Chem Phys* **2012,** 137, 064108.

34. Warshel, A., *Computer Modeling of Chemical Reactions in Enzymes and Solutions*. Wiley: New York, 1997.

35. Repic, M.; Vianello, R.; Purg, M.; Duarte, F.; Bauer, P.; Kamerlin, S. C.; Mavri, J., Empirical Valence Bond Simulations of the Hydride Transfer Step in the Monoamine Oxidase B Catalyzed Metabolism of Dopamine. *Proteins* **2014,** 82, 3347-3355.

36. Noid, W. G., Perspective: Coarse-Grained Models for Biomolecular Systems. *J Chem Phys* **2013,** 139, 090901.

37. Noid, W. G., Systematic Methods for Structurally Consistent Coarse-Grained Models. *Methods Mol Biol* **2013,** 924, 487-531.

38. Messer, B. M.; Roca, M.; Chu, Z. T.; Vicatos, S.; Kilshtain, A. V.; Warshel, A., Multiscale Simulations of Protein Landscapes: Using Coarse-Grained Models as Reference Potentials to Full Explicit Models. *Proteins* **2010,** 78, 1212-1227.

39. Vicatos, S.; Rychkova, A.; Mukherjee, S.; Warshel, A., An Effective Coarse-Grained Model for Biological Simulations: Recent Refinements and Validations. *Proteins* **2014,** 82, 1168-1185.

40. Andrews, C. T.; Elcock, A. H., Coffdrop: A Coarse-Grained Nonbonded Force Field for Proteins Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of Amino Acids. *J Chem Theory Comput* **2014,** 10, 5178-5194.

41. Srivastava, A.; Voth, G. A., Solvent-Free, Highly Coarse-Grained Models for Charged Lipid Systems. *J Chem Theory Comput* **2014,** 10, 4730-4744.

42. Szilagyi, A.; Grimm, V.; Arakaki, A. K.; Skolnick, J., Prediction of Physical Protein-Protein Interactions. *Phys Biol* **2005,** 2, S1-16.

43. Janin, J., Protein-Protein Docking Tested in Blind Predictions: The Capri Experiment. *Mol Biosyst* **2010,** 6, 2351-2362.

44. Moal, I. H.; Torchala, M.; Bates, P. A.; Fernandez-Recio, J., The Scoring of Poses in Protein-Protein Docking: Current Capabilities and Future Directions. *BMC Bioinformatics* **2013,** 14, 286.

45. Saunders, M. G.; Voth, G. A., Coarse-Graining of Multiprotein Assemblies. *Curr Opin Struct Biol* **2012,** 22, 144-150.

46. May, E. R.; Feng, J.; Brooks, C. L., 3rd, Exploring the Symmetry and Mechanism of Virus Capsid Maturation Via an Ensemble of Pathways. *Biophys J* **2012,** 102, 606-612.

47. Kim, Y. C.; Hummer, G., Coarse-Grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J Mol Biol* **2008,** 375, 1416-1433.

48. Miyazawa, S.; Jernigan, R. L., Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J Mol Biol* **1996,** 256, 623-644.

49. Rozycki, B.; Kim, Y. C.; Hummer, G., Saxs Ensemble Refinement of Escrt-Iii Chmp3 Conformational Transitions. *Structure* **2011,** 19, 109-116.

50. Francis, D. M.; Rozycki, B.; Tortajada, A.; Hummer, G.; Peti, W.; Page, R., Resting and Active States of the Erk2:Heptp Complex. *J Am Chem Soc* **2011,** 133, 17138-17141.

51. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A., A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins* **2004,** 55, 351-367.

52. Fraczkiewicz, R.; Braun, W., Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *Journal of Computational Chemistry* **1998,** 19, 319-333.

53. Best, R. B.; Chen, Y. G.; Hummer, G., Slow Protein Conformational Dynamics from Multiple Experimental Structures: The Helix/Sheet Transition of Arc Repressor. *Structure* **2005,** 13, 1755-1763.

54. Karanicolas, J.; Brooks, C. L., 3rd, The Origins of Asymmetry in the Folding Transition States of Protein L and Protein G. *Protein Sci* **2002,** 11, 2351-2361.

55. Humphrey, W.; Dalke, A.; Schulten, K., Vmd: Visual Molecular Dynamics. *J Mol Graph* **1996,** 14, 33-38.

Table 1. Domain-domain contacts in each of the seven MC simulations.

	Volume	NTD1-NTD2	NTD1-CTD2	CTD1-NTD2	CTD1-CTD2
1	$(160\ Å)^3$	9.06 (59.9%)	2.82 (18.6%)	2.64 (17.5%)	0.60 (4.0%)
2	$(240\ Å)^3$	8.82 (60.3%)	2.63 (18.0%)	2.57 (17.6%)	0.62 (4.2%)
3	$(320\ Å)^3$	8.00 (59.1%)	2.71 (20.0%)	2.34 (17.3%)	0.49 (3.6%)
4	$(400\ Å)^3$	7.21 (59.3%)	2.22 (18.3%)	2.24 (18.4%)	0.48 (3.9%)
5	$(480\ Å)^3$	5.65 (60.7%)	1.58 (17.0%)	1.67 (18.0%)	0.41 (4.4%)
6	$(560\ Å)^3$	4.86 (61.0%)	1.37 (17.2%)	1.34 (16.8%)	0.40 (5.0%)
7	$(640\ Å)^3$	2.71 (63.1%)	0.61 (14.2%)	0.75 (17.4%)	0.23 (5.3%)

For each simulation (with the given periodic length), the sum of the contact strengths (defined in the text) between every pair of domains in different proteins is calculated for each frame in the trajectory, representing the effective number of contacting residue pairs between the two domains. The average contact numbers over all the frames in each simulation are given in the table. There are four types of domain-domain contacts, between NTD1/CTD1 in the first protein and NTD2/CTD2 in the second protein. The percentages of each type among the total contact numbers are provided in the parenthesis.

FIGURE LEGENDS

Figure 1. The structure (C_α atoms only) of the CA protein, which consists of two rigid domains (NTD and CTD) connected by a 5-residue flexible linker. Three segments (6-8, 88-90, and 176 to 187) that were missing in the crystal structure[13] and rebuilt through modeling (described in the text) are indicated by stars. All images of protein structures in this article were rendered using VMD.[55]

Figure 2. Some energy functions[47] adopted in this study. **(A)** Representative pairwise non-bonded energy terms as a function of the inter-residue distance r, including the Lennard-Jones-type energy (Eq. 5) $u_{ij}^{LJ}(r)$ for the Met/Trp pair (*solid line*, with $\varepsilon_{ij} < 0$) and for the Pro/Lys pair (*dashed line*, with $\varepsilon_{ij} > 0$), and the electrostatic energy (Eq. 6) $u_{ij}^{EL}(r)$ between a pair of residues with charges $\pm e$ under a dielectric constant of 80 and a Debye screening length ζ = 10 Å (*dash-dot line*). **(B)** The angle energy (Eq. 8) E_{angle} for the flexible linker as a function of the C_α-C_α-C_α angle.

Figure 3. Two methods for estimating dissociation constant K_d from the binding probabilities obtained in the MC simulations. **(A)** The ratio p_u / p_b for the probabilities of the unbound and bound states in each simulation, as a function of the system volume V_0. The *solid line* represents a linear fit. The dissociation constant determined from the slope of this line (according to Eq. S17) is $K_d \approx 25$ μM. **(B)** The probability p_b for the bound state as a function of the effective concentration ($2/V_0$ multiplied by Avogadro's number) in each simulation. According to Eq. S18, the best-fit curve (*solid line*) corresponds to a dissociation constant $K_d \approx 20$ μM.

Figure 4. Representative binding modes. As described in the text, the k-means algorithm was used to partition the protein poses into 7 clusters (with one cluster corresponding to the unbound state). **(A)** Simulation snapshots representing the 6 clusters of bound states. **(B)** Scatter plots for the bound structures in each cluster, displaying the dimer RMSD to the representative structure in (A) *vs.* the energy. The *dashed lines* indicate the average energy of the unbound structures.

Figure 5. A comparison of the binding interfaces in the assembled CA structure and the closest matches from our simulation frames. **(A)** The average NTD-NTD dimer conformation in the CA assembly from a Cryo-EM structure[17] is used as the reference and shown in the Cartoon representation. The simulation snapshot with the smallest RMSD to this reference is shown in C_α-atom trace. The RMSD (NTD atoms only) between this snapshot and the reference is 1.2 Å. **(B)** Similar to A, but with the average NTD-CTD interface from the Cryo-EM structure[17] as the reference. The RMSD (for the atoms in the contacting NTD/CTD) between the displayed simulation snapshot (in C_α-atom trace) and the reference is 1.7 Å. **(C)** Similar to A, but with the average CTD-CTD dimer from the Cryo-EM structure[17] as the reference. The RMSD (CTD atoms only) between this snapshot and the reference is 1.5 Å. **(D)** The average CTD trimer interface in the Cryo-EM assembly structure[17] is shown in the Cartoon representation, and the CA dimer (shown in C_α-atom trace) in a simulation frame is superimposed to two of the CTDs in the trimer, with an RMSD (CTD atoms only) of 1.4 Å.

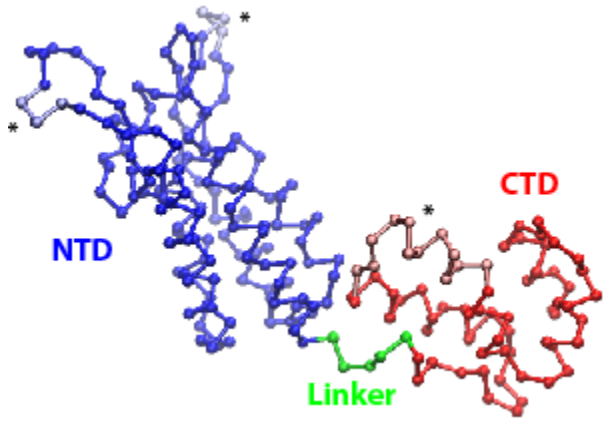


Figure 1

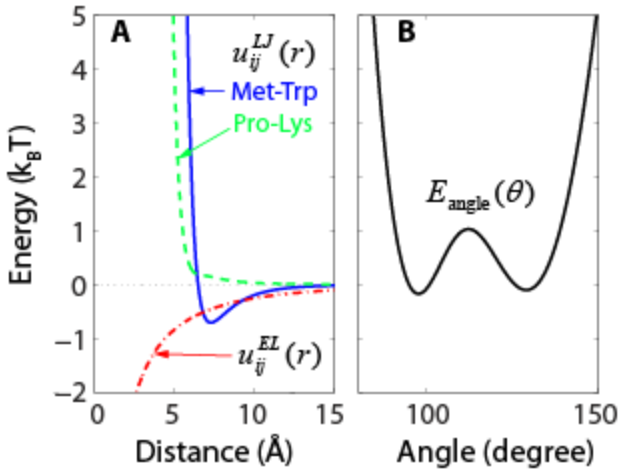

A
B
$u_{ij}^{LJ}(r)$
Met-Trp
Pro-Lys
$u_{ij}^{EL}(r)$
$E_{angle}(\theta)$
Energy (k_BT)
Distance (Å)
Angle (degree)
5
4
3
2
1
0
−1
−2
0
5
10
15
100
150

Figure 2

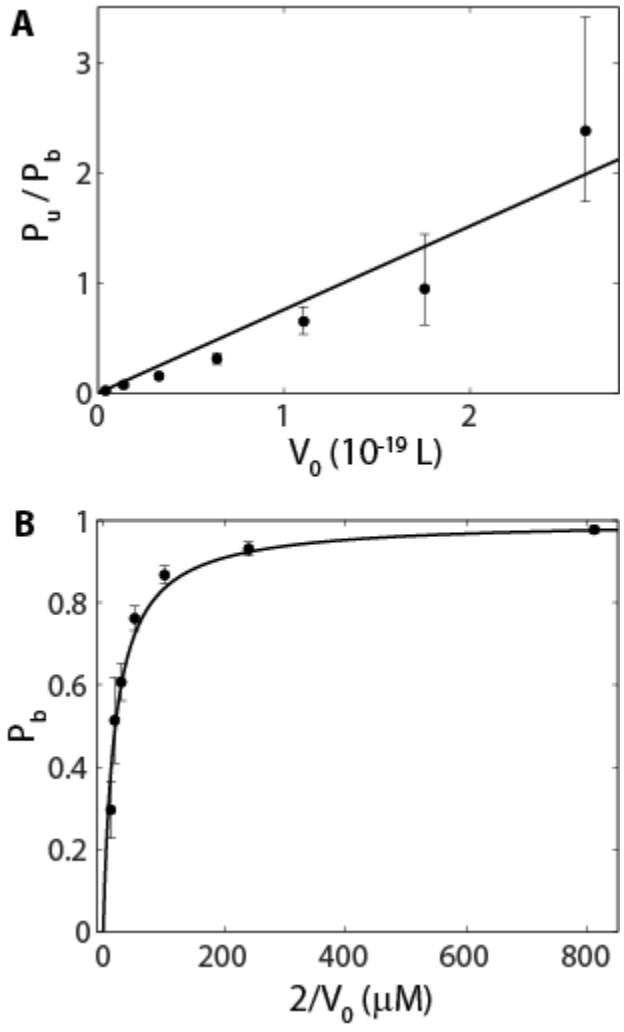


Figure 3

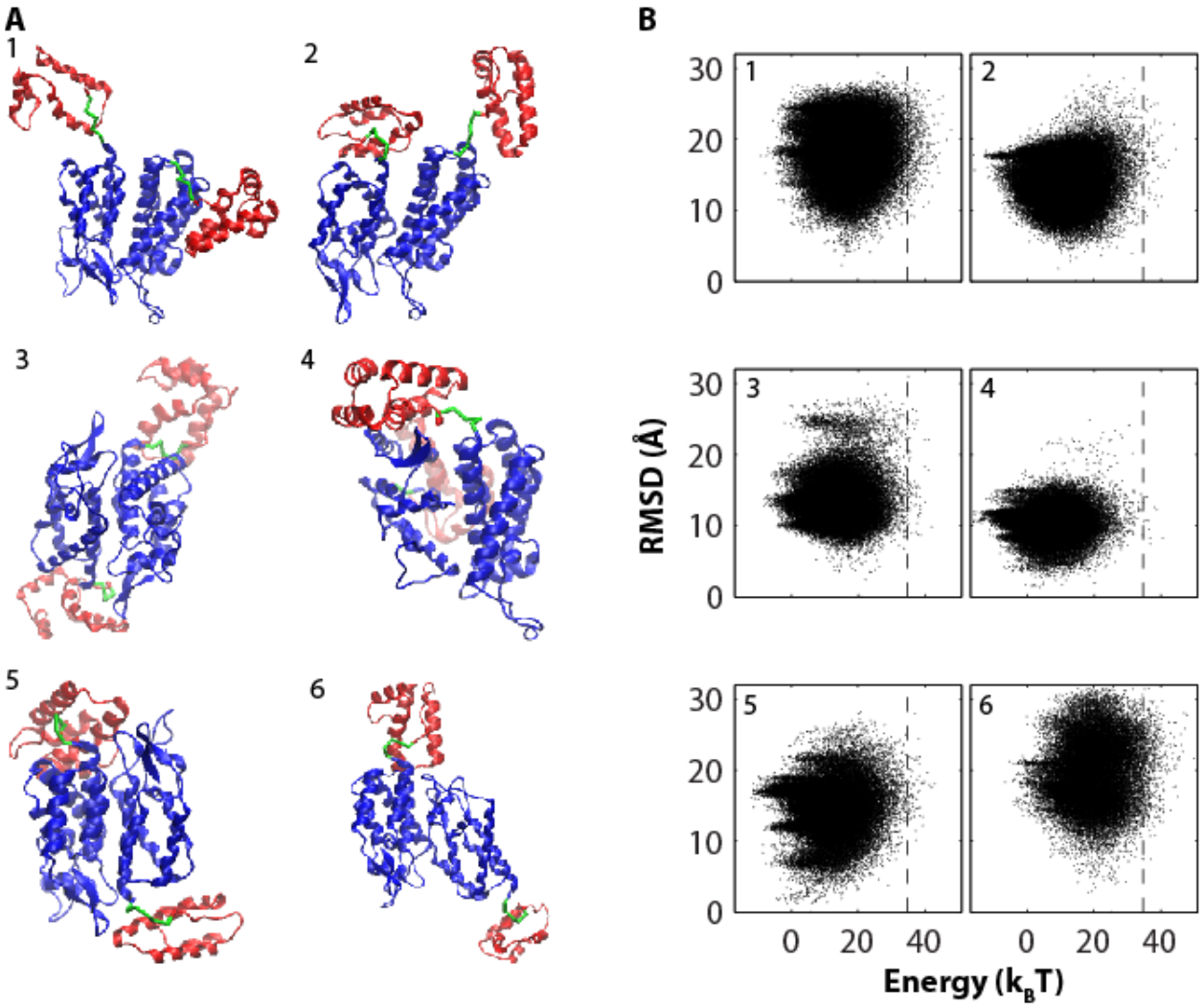


Figure 4

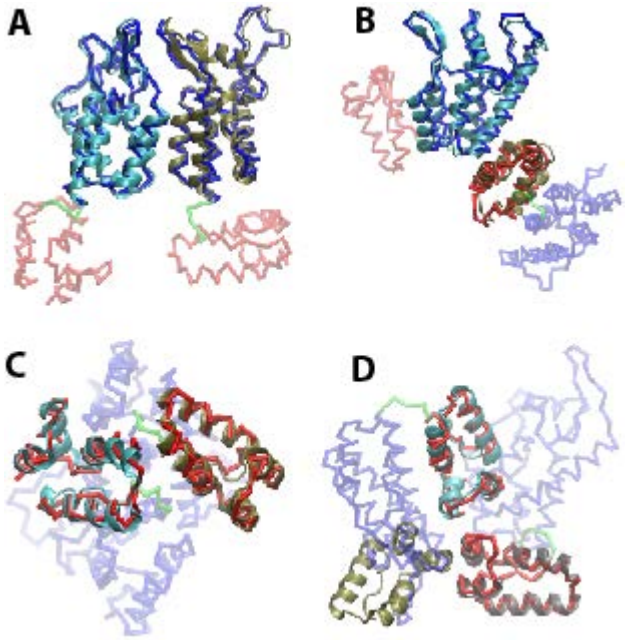
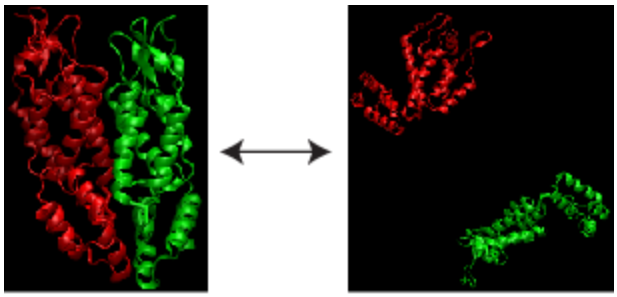


Figure 5



Bound Unbound