Molecular Simulation Study of Protein Conformation Change, Binding Mechanisms, and Allosteric Communication

Jason Pattis  
*University of Connecticut - Storrs, jason.pattis@uconn.edu*

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

**Recommended Citation**

https://opencommons.uconn.edu/dissertations/2377
Molecular Simulation Study of Protein Conformation Change, Binding Mechanisms, and Allosteric Communication
Jason Pattis, PhD
University of Connecticut, 2019

Many processes in biology involve conformational changes or binding events which can be described by a pathway or ensemble of pathways. These processes are challenging to study experimentally as obtaining the temporal and spatial resolution sufficient to understand the underlying physical mechanisms can be challenging. Molecular dynamics (MD) simulations is a powerful tool that can provide atomic resolution on these processes and aid in the design and interpretation of experiments. In this thesis, I will describe MD simulations using enhanced sampling methods to investigate several biomolecular systems, including the Lassa virus nucleoprotein, phosphodiesterase enzymes and a peptide from the Flock House virus (FHV).

The Lassa virus nucleoprotein (NP) has two domains and the N-terminal domain binds the single-stranded RNA genome. We initially focused on the N-terminal domain where PCA as well as metadynamics revealed a large energy barrier for NP opening the RNA binding pocket without RNA and a small barrier when bound. Anti-correlated motions in the transition state suggests NP may partially open and make initial contact with RNA, which then facilitates full opening and binding.

Further studies on the full-length NP were motivated by Hydrogen/Deuterium exchange data, which suggested disruption of an NP trimer may generate opening of the RNA binding pocket. From long timescale simulations and a two-stage adaptive sampling scheme, we constructed a Markov-State Model to describes the dynamics of the full-length NP in a trimer dissociated state. The model revealed domain level conformational changes as well as changes
near the RNA-binding pocket including shifting out of helix 8 and 9 which may allow room for RNA to contact the binding pocket.

Phosphodiesterase 6 (PDE6) is an enzyme in the vision signaling pathway and has high sequence similarity to PDE5 but a large difference in its catalytic rate of cGMP. Simulations revealed correlated motions between helix 12, which is far from the binding pocket, and H- and M-loops in PDE6 but not PDE5, which may explain difference in substrate access and binding. Finally we use TICA to evaluate the confidence in umbrella sampling calculations of folding of FHV gamma peptide on membranes of different lipid compositions.
Molecular Simulation Study of Protein Conformation Change, Binding Mechanisms, and Allosteric Communication

Jason Pattis

B.A., Quinnipiac University 2012

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy
at the
University of Connecticut

2019
Doctor of Philosophy Dissertation

Molecular Simulation Study of Protein Conformation Change, Binding Mechanisms, and Allosteric Communication

Presented by
Jason Pattis, B.A.

Major Advisor
Eric May

Associate Advisor
James Cole

Associate Advisor
Victoria Robinson

Associate Advisor
José Gascón

Associate Advisor
Carolyn Teschke

University of Connecticut
2019
Acknowledgments

I would like to first thank my advisor Dr. Eric May. Thank you for taking me into your lab. Thank you for encouraging me to peruse my ideas. Thank you for always being there so I could come to you for help. I am grateful for the exciting challenges you have thrown my way. I would like to thank all my committee members: Dr. James Cole, Dr. Victoria Robinson, Dr. José Gascón, and Dr. Carolyn Teschke. Thank you for helping me with my journey.

I would like to thank all the past and present members of the May Lab: Dr. Shivangi Nangia, Dr. Asis Jana, Dr. Kevin Boyd, Searle (IC) Duay, Aaron Feinstein, Jeffrey Tamucci, Dr. Allyn Brice, Mickey Ward, Shaan Kamal, Deaneira Lakheram, and Prakhar Bansal. Thank you to JD, for being a great climbing buddy. Thank you to Mickey for being a great friend. Thank you to Kevin for always being there for me for everything.

Thank you to all the great roommates I’ve had especially Drew Clearfield, Kevin Boyd, and Richard Whitehead. Thank you for the amazing memories. Thank you to everyone in the MCB department who helped make it a community, you make it possible to get through grad school. Thank you to all the friends I have made at other departments who have helped make my time here memorable.

Finally thank you to my family for supporting me. My mom and dad, Kathy and Jim, thank you for helping me through. Thank you to my brother, Ryan.
# Table of Contents

## Chapter 1: Introduction

- **1.1 Molecular Dynamics Simulations** ................................................................. 6
- **1.2 Path Sampling Strategies and Challenges** .................................................... 10
  - **1.2.1 Umbrella Sampling** ............................................................................ 10
  - **1.2.2 Metadynamics** ............................................................................... 13
  - **1.2.3 Markov State Modeling (MSM)** .......................................................... 15
- **1.3 Dimensionality Reduction** ......................................................................... 20
  - **1.3.1 Principal Component Analysis (PCA)** ............................................... 20
  - **1.3.2 Time-Lagged Independent Component Analysis (TICA)** .................. 21
- **1.4 Analysis of Allosteric Communication Between Residues** ....................... 21
  - **1.4.1 Root Mean Square Fluctuation (RMSF)** ............................................. 22
  - **1.4.2 Dynamic Cross Correlation Matrix (DCCM)** ..................................... 22
  - **1.4.3 Mutual Information** .......................................................................... 22
- **1.5 References** ............................................................................................... 23

## Chapter 2: Influence of RNA binding on the structure and dynamics of the Lassa Virus nucleoprotein

- **Abstract** ......................................................................................................... 29
- **Introduction** .................................................................................................... 30
- **Materials and Methods** .................................................................................. 34
- **Results and Discussion** .................................................................................. 39
- **Conclusions** ................................................................................................... 51
- **References** .................................................................................................... 53
Chapter 3: Markov State Modeling Reveals Large Structural Changes During the Lassa Virus Nucleoprotein Trimer to Monomer Transition

Abstract

Introduction

Methods

Results and Discussion

Conclusions

References

Supporting Material

Chapter 4: Catalytic Domains of Phosphodiesterase 5, 6, and 5/6 Chimera

Display Differential Dynamics and Ligand Dissociation Energy Barriers

Abstract

Introduction

Methods

Results

Discussion

Conclusions

References

Supporting Material

Chapter 5 Part I: Folding a viral peptide in different membrane environments: pathway and sampling analyses

Abstract
Chapter 5 Part II: Challenges with addressing the reliability of umbrella sampling: Using dimensionality reduction to gain insight into the sampling error and effectiveness of adaptively extending umbrella sampling windows

Introduction........................................................................................................ 167
Methods.................................................................................................................. 172
Methodology and Results...................................................................................... 172
Discussion.............................................................................................................. 183
References.............................................................................................................. 185

Chapter 6: Conclusions and Future Directions...................................................... 188
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Typical timescales of biophysical techniques to investigate protein motions</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>Image of Lassa NP N-terminal domain solvated</td>
<td>9</td>
</tr>
<tr>
<td>1.3</td>
<td>Example of the methodology of umbrella sampling</td>
<td>12</td>
</tr>
<tr>
<td>1.4</td>
<td>Example of the methodology of metadynamics</td>
<td>14</td>
</tr>
<tr>
<td>1.5</td>
<td>Example of the methodology of Markov State Modeling</td>
<td>16</td>
</tr>
<tr>
<td>1.6</td>
<td>Analytical example of the discretion error Markov State Modeling</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>Structure of the Lassa N-terminal domain in the closed and open state</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Equilibrium dynamics</td>
<td>40</td>
</tr>
<tr>
<td>2.3</td>
<td>Lassa NP and RNA structural dynamics in the open-RNA-bound state</td>
<td>42</td>
</tr>
<tr>
<td>2.4</td>
<td>PCA of open-RNA dynamics</td>
<td>44</td>
</tr>
<tr>
<td>2.5</td>
<td>FES of Lassa NP helix 6 opening</td>
<td>48</td>
</tr>
<tr>
<td>2.6</td>
<td>Contacts between helix 6 of monomer I and neighboring subunits in the crystal structure</td>
<td>50</td>
</tr>
<tr>
<td>2.S1</td>
<td>Comparison of open RNA bound dynamics of different subunits</td>
<td>57</td>
</tr>
<tr>
<td>2.S2</td>
<td>PCA of Close-apo Dynamics</td>
<td>58</td>
</tr>
<tr>
<td>2.S3</td>
<td>Convergence of apo FES by metadynamics</td>
<td>59</td>
</tr>
<tr>
<td>2.S4</td>
<td>Convergence of RNA-bound FES by metadynamics</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Lassa NP structure</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td>Mutational frustration of Lassa NP monomer in the trimeric state</td>
<td>70</td>
</tr>
<tr>
<td>3.3</td>
<td>Overlay of initial and final structures of the 4 µs Anton simulation</td>
<td>71</td>
</tr>
<tr>
<td>3.4</td>
<td>Implied timescale plot of top 7 processes</td>
<td>73</td>
</tr>
<tr>
<td>3.5</td>
<td>Free energy surface of the final model</td>
<td>74</td>
</tr>
</tbody>
</table>
3.6 Starting structure then structures from the five macrostates .......................... 76
3.7 Mutational frustration of the most favorable microstate .............................. 77
3.8 Helix 10 displacement ........................................................................ 78
3.S1 MSM protocol flow chart ..................................................................... 84
3.S2 Visualization of 3290 CA distances .................................................... 85
3.S3 Visualization of round 8 of counts based adaptive sampling .................. 86
3.S4 Visualization of round 5 of population uncertainty based adaptive sampling... 87
3.S5 GMRQ score for training set and validation set ..................................... 88
3.S6 Chapman-Kolmogorov test ................................................................. 89
3.S7 Free energy surface of final MSM ........................................................ 90
4.1 Structure and sequence of PDE catalytic subunits .................................... 96
4.2 RMSD comparison among PDE5, PDE6, and PDE5/6 .............................. 103
4.3 PDE5, PDE5/6, and PDE6 RMSF comparisons ...................................... 106
4.4 Collective motions in PDE6 and PDE5 .................................................. 109
4.5 Mutual information on apo systems ....................................................... 110
4.6 Change in mutual information upon ligand binding .............................. 111
4.7 Energetics and pathways of sildenafil unbinding from metadynamics ......... 113
4.8 sildenafil contacts with PDE during unbinding pathways ....................... 114
4.S1 Ramachandran plot for PDE6 homology model .................................... 126
4.S2 Ligand stability ........................................................................ 127
4.S3 Starting structures for metadynamics trials ......................................... 127
4.S4 Metadynamics collective variable trajectories ..................................... 128
4.S5 RMSD in equilibrium simulations ....................................................... 128
4.6 PMF of sildenafil unbinding trials ................................................................. 129
4.7 PMF of GMP unbinding by metadynamics .................................................. 129
4.8 PMF of GMP unbinding trials ................................................................. 130
5.1.1 PMF of folding FHV gamma1 peptide in different membrane environments .. 158
5.1.2 Two-dimensional free energy surface of gamma1 folding on PC, PG, and mixed bilayers ................................................................. 158
5.1.3 Segmental folding behavior of gamma1 in PC, PG, and mixed bilayers ...... 159
5.1.4 Ratio of PG:PC lipid contacts to the gamma1 peptide in the mixed bilayer... 159
5.1.5 Alternative PMF reconstructions ............................................................. 160
5.1.6 Projection of US data onto global tICs in the tIC1-tIC2 and tIC1-tIC3 subspaces for PC, PG, and mixed bilayer systems ................................................. 161
5.1.7 Residual helicity in US windows ............................................................ 162
5.1.8 Segmental membrane insertion depth of gamma1 in PC, PG and mixed... 163
5.1.9 Projection of the US trajectories onto global tIC1-S_{alpha} coordinate space... 163
5.1.10 Projection of US data onto systems tICs in the tIC1-tIC2 and tIC2 and tIC3 subspaces for PC, PG, and mixed bilayer systems ................................................. 164
5.1.11 Projection of US data onto global tIC1-tIC2 and tIC1-tIC3 subspaces....... 165
5.1.12 Projection of US data onto global tIC1-tIC2 and tIC1-tIC3 subspaces....... 165
5.2.1 The difference between the final PMF and the PMF when shorter umbrella windows are used ................................................................. 173
5.2.2 Example auto-correlation function ........................................................ 175
5.2.3 Independent sample size of different windows ........................................ 176
5.2.4 Jensen-Shannon divergence between consensus and observed probability
distribution along the collective variable.................................................. 177
5.2.5 Observed and consensus probability distributions of the windows with the highest
JS divergence in each system................................................................. 177
5.2.6 3 dimensional PMF in the CV, IC2, and IC4 variables............................ 178
5.2.7 Divergence of the 3-dimensional PMF for the PC, PC, and mixed system.... 179
5.2.8 Observed and consensus probability distribution for the highest divergence
window from PC...................................................................................... 180
5.2.9 Observed and consensus probability distribution for the highest divergence
window from PG.................................................................................... 181
5.2.10 Observed and consensus probability distribution for the highest divergence
window from MIX .................................................................................. 181
5.2.11 Comparison of different adaptive metrics vs extending evenly.............. 183
Chapter 1: Introduction
Biological systems function through a series of complex signaling pathways, where these signals are carried out by the constituent proteins. The proteins themselves are subject to structural dynamics where transitions between states can toggle the activity of the pathway. To gain insight into the nature of these protein structural dynamics requires an understanding of the pathway or ensemble of pathways that describe the transition. Examples of such transitions include: protein activation, conformational changes which can change binding partners, protein-protein interactions, and protein ligand interactions. There are many recent examples where molecular dynamics simulations have been instrumental in uncovering critical details about processes like these. Some recent examples include simulations of Abl tyrosine kinase that were able to discover an intermediate in the activation process that had not been seen in public structures but that Eli Lilly had been able to trap with a small molecule (1). Specific kinase inhibitors are needed for cancer therapeutics, but this has proven difficult with one of the greatest challenges being target flexibility. The structures of the 16 kinetic macrostates found can be used in drug discovery efforts by both individual docking screens as well as ensemble docking(2). The study also provides the mechanistic detail of changes between each macrostate is in the activation process. Simulations of the μ-Opioid receptor bound to either morphine or TRV-130 reveal differential conformational changes where both activate the G-protein pathway, which reduces pain, whereas only morphine activates the β-arrestin pathway, which leads to harmful side effects(3). Full characterization of the multiple activation and deactivation pathways as well as the typical flux along each pathway in the different ligand bound systems can be greatly useful in drug discovery efforts. Simulations can be used in studying assembly processes as well and have recently been employed in the discovery of the assembly process of the homo-trimer fibritin(4). These simulations aided experiments by identifying shot lived, flexible, non-specific contacts that initiated binding. Thermodynamic and
kinetic parameters can be calculated from simulations, a recent example was in the study of trypsin–benzamidine unbinding(5, 6). Here multiple unbinding pathways were discovered while reproducing experimental unbinding rates. These processes can be challenging to study experimentally, where different techniques may be needed to be used to observe different motions on different timescales (Figure 1.1)(7). Experimentally obtaining the time and spatial resolution sufficient to understand the underlying physical mechanisms can be extremely difficult. Detailed information from an ensemble of pathways, would reveal intermediates, barriers, and thermodynamic information which would be greatly useful in understanding these processes. Molecular dynamics simulations is a powerful tool in gaining atomic resolution on these processes and can greatly aid in informing the design and interpretation of experiments.

In this thesis, several long timescale biological processes will be investigated using advanced molecular simulation techniques. The first biological process studied will be the conformational changes that occurs in the N-terminal domain of Lassa virus nucleoprotein that allows gating of the binding pocket to accommodate RNA binding. Lassa’s single stranded genome is encapsulated by the nucleoprotein and the gating mechanism is needed to prevent off-target RNA binding. Second, the conformational change that occurs in the full-length nucleoprotein when it transitions from a trimeric to monomeric state is investigated. Hydrogen/Deuterium exchange data(8) suggests a coupling between oligomeric state and dynamics near the RNA binding pocket. Disruption of this unique gating mechanism may be an excellent target for an anti-viral drug and a full understanding would greatly aid in structure based drug design. Third, phosphodiesterase 5 and 6 are investigated. They have very similar sequences while having very different catalytic activity. Simulations can detect differential dynamics and different ligand unbinding barriers. Lastly, Flock house virus gamma peptide folding on
membranes of different lipid compositions is investigated. Gamma peptide disrupts the host membrane to allow viral entry. Understanding the thermodynamics of the gamma folding pathway will help understand the biological activity. These processes are expected to have large energy barriers and would take microseconds to milliseconds of simulation for these processes to occur one time (Figure 1.1). Because of this, more advanced simulation techniques are required so that these processes can be studied using tractable computational resources. These methods can reduce the computational cost of accumulating data for these transitions. Rather than blindly exploring the conformational space we can focus (bias) our simulations to explore those (collective) variables which appropriately describe the transition. A major theme in this work is using these biasing and dimensionality reduction techniques to enhance our simulation approaches and I will describe the basis of these methods in the following sections.
Figure 1.1 Typical timescales of biophysical techniques used to investigate protein motions. Adapted from (7)
1.1 Molecular Dynamics Simulations

Molecular dynamics simulations use a Newtonian model of physics to simulate the movement of biomolecules. These simulations create a timecourse or trajectory of the jiggling and wiggling of atoms. Data from MD can be used to calculate ensemble averages, examine pathways between states, and calculate the thermodynamics and kinetics of conformational changes. The first molecular dynamics simulation of a protein was published in 1977 (9) and the method has successfully been used to tackle a variety of different problems since then. Molecular dynamics simulations can complement experiments and provide additional atomistic detail on a process of interest, which can then lead to new hypotheses moving forward which can be experimentally tested.

Molecular dynamics simulations use potential energy functions to describe the intramolecular and intermolecular interactions. These potential energy functions have been parameterized by physical experiments and quantum mechanical simulations and are collectively referred to as molecular mechanics force fields. The current investigation will make use of one such energy function, the CHARMM27 additive force field (equation 1) (10-12).

\[
U(R) = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\phi (\phi - \phi_0)^2 + \sum_{\text{dihedrals}} K_\chi (1 + \cos(n\chi - \delta)) + \sum_{\text{impropers}} K_{\text{imp}} (\eta - \eta_0)^2 + \sum_{\text{atombonds}} \epsilon \left[ \left( \frac{R_{\min}}{r_{ij}} \right)^{12} - \left( \frac{R_{\min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}}
\]

Equation 1

The force field breaks down the potential energy into different types of interactions. The CHARMM27 force field has several bonded terms describing the energy of bonds and bond angles using Hooke’s law, dihedral angles using a sinusoidal function, impropers, which define out of plane bending, using a spring, and Urey-Bradley which is a 1,3 interaction, used to separate
symmetric and asymmetric bond stretching. CHARMM27 also contains terms for non-bonded interactions approximating van der Waals interactions with the Lennard-Jones potential and electrostatic interactions using Coulomb’s law. There is also a correction map (CMAP) potential. This grid-based energy brings the backbone phi and psi dihedral angles closer to experimental results. Different atom types have different parameters to define their behavior.

The research in this thesis primarily uses the GROMACS MD simulation package. (13) The MD procedure requires an initial atomic model from which forces are calculated as the negative gradient of the force field. A stochastic leapfrog integrator(14) is then used to integrate the forces over a short time step. A leapfrog integrator calculates velocities at half steps and positions and forces at full steps. This allows the determination of new positions and advances the simulation by a small time step of 2 femtoseconds. This cycle is then repeated millions of times to generate trajectories in the nanosecond to microsecond timescale range.

In the standard procedure in this study the protein is placed inside a box and solvated explicitly with TIP3P (transferable intermolecular potential with 3 points) water and 150 mM NaCl is added. Periodic boundary conditions are used so atoms and forces go through the boundaries and reenter the opposite side of the simulation box. Box sizes are generated with a 10 Å buffer such that sufficient space is left so that the protein does not interact with its periodic images. A representative image of a solvated system can be seen in Figure 1.2. Simulations are conducted in the isobaric-isothermal ensemble (NPT), kept at 300K and 1 ATM using the Parrinello-Rahman barostat(15) and a velocity-rescale thermostat(16). This keeps a constant number of atoms, constant pressure, and constant temperature giving conditions similar to the interior of a cell. These simulations will use cutoffs for non-bonded terms at 12 Å. A switch potential is imposed at 10 Å on the van der Waals forces to prevent a jump in force at the cutoff distance. Long range
electrostatics are handled by the particle mesh Ewald method (17). Here the long range electrostatic interactions are summed in Fourier space which can be calculated faster.
Figure 1.2. Image of Lassa N-terminal domain (PDBID: 3T5Q.K) solvated in a cubic water box with Na\(^+\) and Cl\(^-\) ions and a 10 Å buffer.
1.2 Path Sampling strategies and challenges

One problem where molecular dynamics can be powerfully insightful is studying a biological process that involves a large conformational change. Structural techniques such as X-ray crystallography or NMR will typically get structures of one or two endpoints of a process of interest. Molecular dynamics can provide a full pathway connecting these end states, find intermediates along the pathway, determine the thermodynamics of the process of interest, and some techniques allow for the calculation of kinetics of the process of interest. The timescales of these processes are typically out of reach for a standard simulation. As such several enhanced sampling techniques have been developed to allow these calculations to be done accurately with less data.

1.2.1 Umbrella Sampling

One method to sample along a pathway is to first have a reaction coordinate or collective variable (CV) which defines the transition along the process of interest. Then a harmonic potential is used to restrain sampling of one simulation (window) to a local environment using the functional form in equation 2.

\[ U = \frac{K}{2} (CV - CV_{\text{ref}})^2 \]  

Equation 2

Windows are placed along the reaction coordinate so separate simulations sample different regions. This ensures that all areas of the reaction coordinate are well sampled whereas without the restraint potential, sampling would be concentrated to the local minima and transition states would be under-sampled (Figure 1.3A). Typically a fairly strong spring constant (K) is used to prevent the sampling from drifting too far in high energy regions. Windows need to be spaced so that there is overlap between neighboring windows so that their probability distributions can be
aligned. The windows can be combined into a potential of mean force (PMF) which describes the free energy profile along the reaction coordinate using equation 3.

\[ A(x) = -K_B T \ln P'(x) - U' + F \]  

**equation 3**

Here the free energy \( A(x) \) is calculated by first subtracting out the biased potential \( U' \) from the biased probability distribution \( P'(x) \) then converting the unbiased probability distribution into a free energy using \(-K_B T \ln P(x)\). The different windows are aligned using a constant \( F \) (Figure 1.3B). This is known as the weighted histogram analysis method (WHAM)\(^{(18, 19)}\). The reaction coordinate or coordinates should capture all the slow processes in the system and all other motions should be fast enough that they are well sampled within the length of one window. If large energy barriers are not captured by the collective variable they are referred to as orthogonal energy barriers to the collective variable. These can trap the simulation in one small area of the orthogonal landscape or cause it to spend a disproportionately large amount of time close to the starting structure. This can lead to large problems as rather than getting a true average energy, the energy is skewed toward representing a small subset of structures. Furthermore this can be hard to detect as a consistent energy by be interpreted as consistent converged results when it actually means the system is stuck in a local minima. Other enhanced sampling techniques can be stacked on top of umbrella sampling to better sample motions not captured by the collective variable\(^{(20, 21)}\).
Figure 1.3. A) Two dimensional energy landscape with two basins; A and B. A reaction coordinate, $\xi$, (dotted line) defines the transition between basin A and basin B. Simulations started from the dots freely sample the orthogonal space but are restrained along the reaction coordinate by a harmonic potential. B) the biased probabilities ($P^B$) (bottom) can have the bias subtracted out, then have the individual windows stitched together to form a full PMF (Top solid line). Adapted from (43)
1.2.2 Metadynamics

Metadynamics again requires one or a small number of collective variables which define the transition along the process of interest. Metadynamics uses a history dependent bias which get deposited in the collective variable space, building up the bias potential to push the system towards rarely sampled conformations(22). The bias builds until all local and global minima are filled leaving a flat energy landscape where the system is freely diffusing (Figure 1.4). Once this happens the inverse of the bias can be used to obtain the underlying free energy surface of the system. In well-tempered metadynamics the height of the Gaussian hill biases being added is reduced over time causing the simulation to converge more smoothly(23). The bias (V) given at a particular time is given by equation 4.

\[
V(s, t) = \sum_{t'=0, \tau_G, 2\tau_G, \ldots}^{t'<t} W e^{-V(s(q(t'),t'))/\Delta T} \exp \left( -\sum_{i=1}^{d} \frac{(s_i(q) - s_i(q(t'))^2)}{2\sigma_i^2} \right)
\]

The height of a single Gaussian is defined by W and the width by \( \sigma \). The bias factor \( \Delta T \) controls the rescaling of the Gaussian height over time. \( \tau_G \) is the time between Gaussian depositions. It has been shown that well-tempered metadynamics converges asymptotically in one single simulation(24). Just like umbrella sampling, it is assumed that the reaction coordinates capture all the slow processes in the system with all other motions fast enough that they can be well sampled as the system traverses the reaction coordinate. If this is not true the starting structure can have a large bias on the end result free energy surface as the system can fail to cross an orthogonal energy barrier for the entirety of the simulation causing only part of the conformational landscape to be explored. Metadynamics can more easily be extended to a higher number of collective variables, sometimes up to three, and a variant of metadynamics called bias-exchange metadynamics(25, 26) allows an even higher number of collective variables to be chosen. Again other enhanced sampling
techniques can be stacked on top of metadynamics to enhance the sampling of motions not captured by the collective variable (4, 27).

Figure 1.4. Top) Trajectory of a metadynamics simulation along a collective variable $S$ hopping to different basins over time. Bottom) Gaussian hills are built up to push the system out of local minima and can be inverted to determine the free energy along the collective variable. Adapted from (44)
1.2.3 Markov State Modeling

In both umbrella sampling and metadynamics, a predetermined reaction coordinate/collective variable must be defined. It can often be challenging to choose (a) collective variable(s) which can capture the slow degrees of freedom, and the identification of optimal CVs is an active area of research. An alternative approach for studying pathways, which does not require a predefined CV, are Markov State Models. A Markov state model (MSM) can describe the dynamics of a system as a kinetic network with transition probabilities and rates between discrete states\(^{(28)}\). To construct an MSM the number of transitions between discrete states are counted in some time interval (lagtime) \(\tau\) from all trajectories to form a count matrix. From this a transition matrix is constructed which describes the conditional probability of transitioning from state \(i\) to state \(j\) at lagtime \(\tau\) \(^{(29)}\). Since only the conditional probability of transitioning is needed, many short simulations can be combined to build a model \(^{(28)}\). It has been found that many short simulations are more efficient at exploring a landscape then one long simulation\(^{(30)}\). MSMs do not require the \textit{a priori} knowledge of a collective variable which describes the process of interest but instead learns the slowest motions after sampling. A downside to this approach is that it typically requires more sampling than a biased simulation along a predefined collective variable.
Figure 1.5. A) four well potential energy surface and stationary probability distribution. B) The top eigenvectors of the MSM. The first gives the population and the next give the slowest processes in the system. C) the eigenvalue can be converted to the implied timescales of the slow processes. Adapted from (29)
From the transition matrix an eigenvector and eigenvalue decomposition is performed. The first eigenvector represents the stationary population of the system. The next eigenvectors represent in order the slowest processes in the system (Figure 1.5). The corresponding eigenvalues can be converted into the implied timescale ($t_i$) of those processes using $t_i = - \tau / \ln |\lambda_i|$ where $\lambda_i$ is the $i$-th eigenvalue of the MSM and $\tau$ is the lagtime. The first eigenvalue is 1 corresponding to the stationary process which has a timescale of infinity, representing the timescale of leaving the system. Unlike metadynamics and umbrella sampling MSMs can give direct kinetic information like mean first passage times, whereas in umbrella sampling and metadynamics timescales can only be inferred from the energy landscape by applying transition-state-theory or related theories (e.g. Kramer’s theory) with the assumption that the collective variable captures all the relevant kinetic barriers.
Figure 1.6. Two well potential and the slowest dynamical eigenfunction describing the process of transition from the left basin to the right basin. The discretion error is shown in red as the difference between a model with discrete states (green) and the true process (black). The smallest discretion error occurs when the transition state is binned more finely. Adapted from (29)
A typical procedure for construction of an MSM is to first find some internal set of coordinates (features) which can describe the dynamics of the biomolecule. Next these features are converted into a smaller number of kinetic coordinates using TICA (described in next section). Next the system configurations are clustered into discrete states in the lower dimensional space of a few to several TICA components. The number of clusters should balance the discretion error and statistical error. Discretion is caused because the continuous motions or propagator of the system is being represented as a stepwise function when the configurations are split into discrete states (Figure 1.6). Structures within these states may have slightly different properties which are being averaged together. When these properties are too different, resolution is lost in the model. Statistical error is caused when a state is under-sampled, and the small number of transitions may not well represent the true dynamics of the system. Discretion error can be minimized by having the transition states clustered more finely than the metastable states. A benefit of TICA is that it makes this easier by transforming the features into kinetic coordinates. Identifying kinetic transition states becomes easier and a diverse metastable region that encompasses many conformations will be compacted to a small region due to these conformations small kinetic distances from each other allowing them to be represented by a single cluster. A ergodic cutoff is used to ensure that only strongly connected states are included in the final model. It requires a minimum number of transitions between states ensuring the statistical error isn’t too high.

Parameters, such as: choice of features, number of features, TICA lagtime, number of TICA components, clustering scheme, and number of clusters can be chosen using a variational approach. The variational principal of conformational dynamics(31, 32) shows that when using cross validation to avoid over fitting there is an upper bound to the eigenvalues. Because of this upper limit a higher implied timescale means the slowest process has been better described. The
GMRQ score sums up the first $m$ eigenvalues and can be used as a score to select parameters which best describe the slowest process or processes in the system (33). In the work presented here the MSMs were estimated using a maximum likelihood approach and a sliding window. All MSMs are reversible and enforce detail balance (29).

1.3 Dimensionality Reduction

Biomolecules of interest typically have thousands of atoms making analysis of the biologically relevant motions extremely difficult. Picking a handful of distances or angles to look at can be influenced by the preconceived bias that a researcher goes into a calculation with. It is also easy to miss a relevant motion. Conversely, examining every possible distance and angle is time consuming and impractical. In addition, motions of biomolecules are often cooperative between different regions and it is beneficial to study them as such. Dimensionally reduction provides an automated way of identifying the largest (PCA) or slowest (TICA) motions in a system and treating them as a collective motion.

1.3.1 Principal Component Analysis

Principal component analysis (PCA) (34, 35) takes a set of input features, typically Cartesian coordinates, backbone dihedral angles, or alpha carbon distances and constructs a covariance matrix from the motions observed in the trajectory using equation 5.

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$$

An eigenvalue and eigenvector decomposition of this matrix allow the motions with the largest variance to be identified, which are the eigenvectors with the largest associated eigenvalues. The eigenvalues can be used to calculate the total variance that a specific principal component captures. Typically, a high percentage of the total variance can be captured in a handful of principal components.
1.3.2 Time-lagged Independent Component Analysis

Time-lagged Independent Component Analysis (TICA) (36-38) again like PCA transforms a set of input features, typically Cartesian coordinates, backbone dihedrals, or alpha Carbon distances. A covariance matrix and a time-lagged covariance matrix is constructed after some time interval (lag time \( \tau \)) (equation 6).

\[ c_{ij}(\tau) = \langle r_i(t)r_j(t+\tau) \rangle \]  

\text{equation 6}

The eigenvalue and eigenvector decomposition can reveal the slowest decorrelating motions. The eigenvalues can be transformed into an implied timescale of the process with equation 7.

\[ t_i = -\frac{\tau}{\ln|\lambda_i|} \]  

\text{equation 7}

Here \( \lambda_i \) is the i-th eigenvalue from TICA. The eigenvalues also represent the percentage of kinetic variance used. TICA is much better at capturing rare event processes and the top independent components are very good at describing the kinetics of a system. It has been seen that it takes a higher number of independent components to capture the majority of the kinetic variance. 13 independent components were needed to capture 95 % of the kinetic variance in bovine pancreatic trypsin inhibitor motions and 50 independent components were needed to capture 95 % of the kinetic variance in trypsin–benzamidine dynamics(39).

1.4 Analysis of allosteric communication between residues

Mechanistic information is often hard to pull out from an ensemble of paths and structures. Identifying which specific residues contribute most to a conformational change or allosteric network provides great detail and allows these hypotheses to be tested by mutations. Doing this in a quantitative way helps prevent overinterpretation by the researcher. Molecular dynamics simulations offer an excellent opportunity to identify correlated motions in biomolecules.
1.4.1 Root Mean Square Fluctuation

Root mean square fluctuation (RMSF) (40) is calculated with equation 8.

\[
\text{RMSF}_i = \left( \frac{1}{T} \sum_{t_j=1}^{T} |r_i(t_j) - r_i^{\text{ref}}|^2 \right)^{1/2}
\]

RMSF can be thought of as the standard deviation of the fluctuations over time (T) of each different residue from their mean \(r_i^{\text{ref}}\). This can identify which residues are contributing to the high variance motions. And can easily pinpoint large differences in dynamics. It does not however identify correlated motions between residues.

1.4.2 Dynamic cross correlation matrix

Dynamic cross correlation matrix (DCCM) (41, 42) is able to identify residues that move (are correlated) along the same vector or direction as well as residues that move along opposite vectors (are anti-correlated) using equation 9.

\[
C_{ij} = \frac{\langle r_i \times r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\left( \left( \langle r_i^2 \rangle - \langle r_i \rangle^2 \right) \left( \langle r_j^2 \rangle - \langle r_j \rangle^2 \right) \right)^{1/2}}
\]

The anti-correlated motions often look like breathing motions in a protein. This analysis is often performed using Cartesian coordinates as inputs. The potential problem in this approach is the alignment of structures can influence the correlation, which can be problematic when there are both local and global rearrangements occurring.

1.4.3 Mutual Information

Mutual information calculates the conditional probability that residue i and residue j change states at the same time. It is typically done in dihedral space. This takes away the directional component and the reliance on a good alignment. Mutual information \((I(X,Y))\) measures the extent of the uncertainty in a given variable \((Y)\) changes when the state of another variable \((X)\) is known.
The mutual information is defined as the difference between the Shannon entropies (S) of the marginal distributions of variables X and Y and the joint Shannon entropy (equation 10).

\[ I(X, Y) = S(X) + S(Y) - S(X, Y) \]  
\hspace{1.5cm} \text{equation 10}

I(X,Y) can be calculated from the marginal and joint probabilities with equation 11

\[ I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]  
\hspace{1.5cm} \text{equation 11}

References


Chapter 2: Influence of RNA binding on the structure and dynamics of the Lassa Virus nucleoprotein

Reproduced with permission from:


The following changes have been made:

- The manuscript has been reformatted to meet University Standards
- Figures numbers have the chapter number preprinted
Title: Influence of RNA Binding on the Conformational Dynamics of the Lassa Virus Nucleoprotein
Author: Jason Pattis, Eric R. May
Publication: Biophysical Journal
Publisher: Elsevier
Date: 27 January 2015
Copyright © 2015 Biophysical Society. Published by Elsevier Inc.

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions. Comments? We would like to hear from you. E-mail us at customercare@copyright.com
Abstract
Lassa virus protects its viral genome through formation of a ribonucleoprotein complex, in which the nucleoprotein (NP) encapsidates the single-stranded RNA genome. Crystal structures provide evidence that a conformational change must occur to allow for RNA binding. In this study, the mechanism by which NP binds to RNA and how the conformational changes in NP is achieved is investigated with molecular dynamics (MD) simulations. NP has been structurally characterized in an open configuration when bound to RNA and in a closed form in the absence of RNA. Our results show that when NP is bound to RNA, the protein is highly dynamic and the system undergoes spontaneous deviations away from the open state configuration. The equilibrium simulations are supported by free energy calculations, which quantify the influence of RNA on the free energy surface, which governs NP dynamics. We predict the globally stable states to be qualitatively in agreement with the observed crystal structures, but that both open and closed conformations are thermally accessible in the presence of RNA. The free energy calculations also provide a prediction of the location of the transition state for RNA binding and identify an intermediate metastable state which exhibits correlated motions that could promote RNA binding.
Introduction

Lassa virus is a member of the Old World arenavirus family, which causes hemorrhagic fever (Lassa fever) in infected humans. Lassa is predominant in Western Africa where it infects hundreds of thousands of people, resulting in 5,000-10,000 human deaths annually (1, 2). There is no effective vaccine against Lassa (2) and it is the most frequently transported hemorrhagic fever out of Africa to the United States and Europe (1, 3, 4). As Lassa is a major threat to public health, a critical step toward developing antiviral therapeutics and vaccines is a detailed understanding of the structure-function relationships in the Lassa virion.

Lassa has a simple genome, containing four genes, but employs a complex ambisense coding strategy. The single-stranded (ss) negative sense RNA genome consists of large (~7 kb) and small segments (~3.4 kb) (5). The large subunit encodes the viral RNA-dependent RNA polymerase (L) and a zinc-binding protein (Z), while the small segment encodes for a glycoprotein (GP), which is posttranslationally cleaved to form GP1 and GP2 glycoproteins, and an RNA binding nucleoprotein (NP). Prior to 2010 no high-resolution structures of any Lassa protein had been determined. In recent years significant progress has been made in the structural characterization of the Lassa proteins including structures of the Z-protein (6), the N-terminal endonuclease domain of L (7), GP1 (8) and numerous structures of NP (9-14).

NP contains 569 residues and is composed of N- and C-terminal domains connected by a flexible linker. The N-terminal domain is the RNA binding domain (9) and the C-terminal domain has double-stranded (ds)RNA nuclease functionality (13). The full-length structure has been...
determined (11, 12), as have structures of just the N-terminal domain (13) and the C-terminal domain (9, 10). The domain level structures are highly similar and topologically consistent with the full length structures. NP has multiple roles in the Lassa life cycle, including binding the viral RNA to form ribonucleoprotein (RNP) complexes, which is required for transcription and replication (15). NP also has immunosuppressive activity related to it’s ability to degrade dsRNA in C-terminal domain (13, 16).

Lassa NP binds RNA and contains a positively charged RNA binding cavity in the N-terminal domain. The first structure determined for Lassa NP was for the full-length protein, however the RNA binding pocket could not be determined (11). In a subsequent study, utilizing a C-terminal domain deletion mutant, an RNA bound form of Lassa NP N-terminal domain was solved (9). The overall topology of the unbound (apo) and RNA bound N-terminal domains were consistent, showing mainly α-helical secondary structural elements (Fig. 1). The major structural differences between the apo and RNA-bound forms were in the two helices close to the RNA binding site. In the apo form, α helix 5 (α5, residue 97 to 122) is extended compared to the RNA bound structure. The other major difference between the apo and RNA bound structures was in the positioning of α helix 6 (α6, residues 131 to 145) relative to the RNA binding groove. In the apo form (Fig. 1A), α6 is aligned perpendicular to the binding groove and in close proximity to the groove, potentially blocking RNA entry to the binding pocket. In the RNA-bound form (Fig. 1B) α6 is swung away from the binding groove, allowing RNA to enter and bind. The RNA-bound crystal structure asymmetric unit contained six subunits and α6 adopted different orientations in all subunits. In
all of the subunit structures α6 was spatially separated from the RNA binding site in an open configuration.

One potential explanation for why full-length NP could not be crystalized in an RNA-bound state is associated with the oligomeric state of NP in those studies. Multiple oligomeric states of the full-length protein have been detected (9) but based upon X-ray crystal structures (11, 12), small angle X-ray scattering (SAXS) (12), sucrose gradient ultracentrifugation (17) and electron microscopy (EM) (12) experiments a symmetric head-to-tail trimer of NPs appears to be the dominant oligomeric state in solution. It has been suggested that this trimeric self-association serves a similar function to the phosphoprotein in non-segmented, negative strand viruses, in that it allows a buildup of NP without non-specific RNA binding (9, 18). In the trimeric form the NP RNA-binding site is presumed to be occluded, due to positioning of α6 and a “gating loop” (residues 232-243) in the N-terminal domain (9). It is speculated that there may be coupling between the interdomain separation (or angle) and accessibility of the RNA binding site, but direct evidence for this coupling is currently lacking. The trimer form may represent an early stage RNP assembly intermediate that is disrupted as viral RNA-replication proceeds and the concentration of viral RNA increases. Formation of the RNP is essential for viral transcription and RNA-replication (19), and therefore understanding the initial RNA-NP interactions may provide new targets for developing anti-viral therapies (8, 20).
Figure 2.1. Structure of Lassa NP N-terminal domain in the closed (A) and open (B) states. These structures are derived from the PDB deposited structure 3MWP (A) and 3T5Q.K (B). The open structure is bound by a 6 base long strand of RNA (purple). The major structural differences between the bound and unbound forms are in the positioning of α helix 6 (red) and the disordering of a region of α helix 5 (orange).

Molecular modeling studies on Lassa have been limited by the lack of structural information, but recent studies have examined the Z protein (21) and NP (22, 23). The studies on NP involved binding calculations between NP and m7GpppN cap (23) and RNA (22). The earlier study (23) examined dTTP and m7GpppN binding to the closed conformation of NP to evaluate a proposed cap-snatching mechanism to prime mRNA synthesis (11). Using Molecular mechanics/generalized Born surface area (MM/GBSA) calculations (24), Han et al. predicted strong binding (-127 kcal/mol) of m7GpppN to NP. A strong binding interaction between NP and m7GpppN is not experimentally well established, and in fact there is evidence against this mechanism. A series of mutants in the RNA binding pocket failed to generate a significant reduction in mRNA levels without concomitant reduction in viral genome replication (12) and pull-down assays failed to show binding of NP to m7GTP (9). In a subsequent study by Han et al. MM/GBSA was again used to evaluate RNA binding to NP. In that study the open state was
modelled by using a pepsin-digested structure which lacked coordinates for α6, and used the closed state structure (PDBID:3MWP) as a homology modeling template. This approach may have generated a conformation of NP, in which the positioning of α6 was not as far removed from the RNA binding pocket as is observed in the open-state crystal structure (PDBID:3T5Q). Therefore, what is presented as an open conformation, may be quite similar to the closed state structure.

In that study, the RNA binding energy was evaluated between wild-type and two mutant NP variants, which are known to knock down transcriptional activity. The mutants are predicted to bind RNA slightly stronger than the wild type NP, which would imply a more complicated mechanism than simply weakening the binding. While the two studies may not be directly comparable, the binding energy to m7GpppN (-127 kcal/mol) was predicted to be stronger than binding RNA(-101 kcal/mol), which does not appear to be consistent with experimental studies.

This investigation examines the gating structural rearrangements in the N-terminal domain of NP and the mechanism of RNA binding. We have employed MD simulations to explore the influence of RNA binding on the NP dynamics and energy landscape in both the open and closed configurations. The current study advances upon previous studies by employing pathway based (metadynamics) methods, over end-point based (MM/GBSA) methods, which can provide mechanistic insights. This study serves to enhance our knowledge of Lassa NP behavior through the application of multiple MD techniques and analyses to resolve the dynamic structure and interactions that arise when NP undergoes binding of viral RNA.

**Materials and Methods**
**Model Building**

To generate complete models for the Open-RNA bound structure we used the pepsin digested structure (NPpep, PDBID: 3T5N) and modeled in missing segments. The coordinates of α6 (residues 128 to 145) were copied from the k monomer structure in PDBID: 3T5Q. The loops adjacent to α6 (residues 113 to 127 and residues 146 to 163) were modeled into the structure using MODELLER (25, 26). The structure for the Closed-apo state was based upon the full-length structure (PDBID: 3MWP). Residues 148 to 157 were modeled in using MODELLER. The Open-apo structure was created by taking this Open-RNA bound structure and removing the RNA atoms. The Closed-RNA structure was built by aligning the Closed-apo and Open-RNA bound structures along the RNA contacting residues (9) and then copying in the RNA coordinates. Additional models for the Open-RNA conformation were constructed in the same manner, except the coordinates of α6 were extracted from the A or G monomers in 3T5Q. The sequences of RNA used in these simulations is consistent with the RNA sequence in the crystal structure (PDBID:3T5Q), which is UAUCUC. This RNA sequence was used in all RNA simulation except for the simulation using the α6 position of the A monomer, which had more bases resolved in the crystal structure. The sequence in the A monomer simulation is UUAUCUCA.

**Equilibrium MD**

MD simulations were preformed using the GROMACS 4.6.1 simulation package (27) using the CHARMM27 force field (28, 29). Systems were solvated in a cubic box with CHARMM TIP3P water model, with Lennard-Jones parameters for the hydrogen atoms, and 150 mM NaCl. System sizes were approximately 68,000 atoms for the open NP systems and approximately 61,000 atoms
from the closed NP systems. The systems were energy minimized for 5,000 steps using the steepest decent (SD) algorithm while position restraining the non-hydrogen (heavy) atoms of the protein and RNA with a force-constant of 1,000 kJ/mol*nm². The position restraints were released and the system was minimized for another 5,000 steps of SD. Following energy minimization the systems were restrained in the same manner as energy minimization and 100 ps of NVT (300 K), followed by 100 ps of NPT (300 K, 1 atm) equilibration were run. Production MD simulations were performed in the NPT ensemble at 300 K and 1 atm using the Parrinello-Rahman barostat and the velocity-rescale thermostat (30). Coupling time constants of 1 ps were used for both the temperature and pressure coupling. A switching function was applied to the Lenard-Jones forces at 10 Å and were cutoff at 12 Å. Long range electrostatics (greater than 12 Å) were computed by the Particle Mesh Ewald method with a Fourier spacing of 1 Å. The trajectories were computed using the leap-frog stochastic dynamics integrator with a time step of 2 fs. Waters were keep rigid using the SETTLE algorithm (31) and non-water bonds involving hydrogen atoms were constrained with the LINCS algorithm (32). The Open-RNA, Open-apo and Closed-apo systems were simulated for 250 ns, while the Closed-RNA system was simulated for 150 ns.

Principle component analysis (PCA) was performed on the equilibrium simulations to identify dominant motions, which allow for the dynamics to be analyzed in a reduced basis set. PCA is performed by constructing a covariance matrix \( C_{ij} \) of the protein backbone atom positions using the GROMACS g_covar tool using simulation snapshots taken every 1 ps. The covariance matrices are given by
\[ C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \]  
(Eq. 1)

where \( x_{1}, ..., x_{3N} \), are the mass weighted Cartesian coordinates of the N backbone atoms, and \( \langle \rangle \) indicates averaging over the snapshots extracted from the simulations (e.g. time averaging).

The eigenvectors and eigenvalues of \( C_{ij} \) can be determined by standard linear algebra methods, and each eigenvector represents a direction of motion of the system and is termed a principal component (PC). The PCs with the largest eigenvalues capture the most variance (motion), and are therefore considered the global motions. It is often observed that analyzing just a few PCs will capture the majority of motion of the system, as well as functionally relevant motions, and therefore is an effective technique for dimensionality reduction (33, 34). The eigenvectors of our covariance matrices were determined using \( \text{g}_\text{aneig} \) tool. PCA free-energy surfaces were computed by projecting the equilibrium trajectories at 1 ps frequency onto the first two principle components and then Boltzmann inverting the probability density into a free energy \( (F = -kBT\ln(P)) \).

**Free Energy Calculations**

The Plumed 1.3 plugin (35) with Gromacs 4.6.2 was used to perform well-tempered metadynamics (36). The collective variable (CV) that was biased in the metadynamics simulations was the mean-squared-displacement (MSD) of the \( \alpha6 \ C_\alpha \) atoms measured with reference to the Closed-apo structure. A Gaussian width of 0.05 nm\(^2\) and a height of 0.3 kcal/mol was used. A harmonic upper wall was placed at 4.0 nm\(^2\) from the closed state, with a spring constant of 5000 kJ/mol*nm\(^4\).
Correlated dynamics were analyzed from the metadynamics trajectories by splitting the trajectory into segments based upon the value of the CV. The correlated dynamics were analyzed by constructing dynamic cross-correlation matrices (DCCM) on the sub-trajectories using the Bio3d package (37).

MM/GBSA calculations were performed to analyze the energetics of crystal contacts. The unit cell tool in Chimera (38) was used to display all the NPs in the unit cell using the P6 symmetry of the crystal lattice. A tetramer was selected where three NPs contact 6 of a fourth NP. A 35 ns simulation of the tetramer in explicit solvent was performed in GROMACS. The MM/GBSA analysis was performed in the CHARMM simulation package (version c38)(39) using the CHARMM27 force field on the last 30 ns of the tetramer. Snapshots were extracted every 100 ps, for a total of 300 analysis frames. The full equilibrium simulations were used for the MM/GBSA calculation of monomer opening and RNA binding, frames were extracted every 100 ps for these calculations as well. The binding energy was calculated from Eq. 2

\[ G_{\text{bind}} = \langle E_{\text{MM}} \rangle + \langle G_{\text{GB}} \rangle + \langle G_{\text{NP}} \rangle \quad (\text{Eq. 2}) \]

where \( E_{\text{MM}} \) is the molecular mechanics gas-phase energy, \( G_{\text{GB}} \) is the polar solvation free energy calculated by the generalized-Born formalism, and \( G_{\text{NP}} \) is the non-polar solvation free energy. \( G_{\text{GB}} \) is calculated using the GBSW implicit solvent model (40), with a smoothing length of 0.3 Å, a nonpolar surface tension coefficient of 0.03 kcal/mol*Å², and a salt concentration of 150 mM. The non-polar surface area was calculated as \( G_{\text{NP}} = \gamma(\text{surface area}) + b \), where we take \( \gamma=0.00542 \) and \( b=0.92 \) (41). The surface area is determined from the solvent-accessible surface area using a probe radius of 1.4 Å.
Results and Discussion

NP Equilibrium Simulations

We initiated our study by evaluating the equilibrium dynamics of the open and closed conformations in both the presence and absence (apo) of RNA. Therefore, four systems were examined: Open-RNA (Fig. 2.1B), Open-apo, Closed-RNA, Closed-apo (Fig. 2.1A). MD simulations for all four systems were conducted for at least 150 ns, and the distributions of backbone RMSD values for the four simulations are shown in Fig. 2.2A. The Closed-apo simulation was the most stable, undergoing a maximum 3.1 Å deviation. Somewhat surprisingly, the Closed-RNA simulation was the second most stable displaying a maximum RMSD of 3.5 Å. Given that the RNA-bound form is in an open configuration in the crystal structure, we expected to see larger structural changes, and possibility a transition to the open state when α6 is started in close proximity to the RNA.

The open state simulations showed NP to be considerably more dynamic than what was observed in the closed state simulations. The increased structural dynamics are reflected by a maximum RMSD of 6.2 Å in the Open-apo and 7.1 Å in the Open-RNA simulations. The observation that the open structures are more dynamic is consistent with the X-ray structure, where α6 adopts significantly different orientations in each of the 6 subunit structures within the asymmetric unit. However, the nature of the dynamics was unanticipated, especially in the Open-RNA case. We monitored if the open simulations made a transition toward the closed state, and found that the
Open-RNA did transition toward the closed structure at around 120 ns, but the Open-apo remained in the open configuration (Fig. 2.2B). The overlay of the closed state with the initial Open-RNA and final Open-RNA structures are shown in Fig. 2.2C and 2.2D, respectively. It can be seen that α6 rotates by ~90° and the C-terminal end of α6 adopts a position in close proximity to the RNA binding groove and in a manner highly similar to the closed state positioning.

Figure 2.2. Equilibrium Dynamics. A) The probability distributions in RMSD, calculated over the backbone atoms after least squares alignment to each simulation’s respective starting structure. B) The Cα RMSD of α6 for the open state simulations calculated after a global alignment to the Closed-apo state. C-D) Comparison of the Open-RNA (red) and Closed-apo (cyan) α6 position at the beginning (C) and end (D) of the 250 ns Open-RNA simulation.
There are six subunits in the RNA bound crystal structure (PDBID: 3T5Q) (9). The simulations discussed above were initiated from the K-subunit structure. In order to assess if the closing motion we observed was dependent on the initial conditions of the simulation (or a spurious fluctuation), we performed two additional simulations in which the RNA bound system was initiated in a different open configuration. We ran simulations of the A-subunit for 140 ns and the G-subunit for 195 ns. In both cases, α6 moves closer to the closed conformation (Fig. 2.3A), changing from greater than 20 Å, to around 15 Å. The degree of closing is not as drastic as observed for the K-subunit (Fig. 2.2B), however in all three simulation the system adopts a more closely packed structure than observed in the crystal structure. We have compared the A, G and K open state simulations on the same phase space of the first two principle components of the K subunit (Fig. 2.S1A). We find that the A and G subunits sample the same phase space as K does during the later part of the K simulations, indicating all three subunits end up with similar dynamics. However the K subunit starts in a different region of of the PC space, which is not sampled by the A and G subunits, which is likely do to the orientations differences of α6 in the different subunits (Fig. 2.S1B). We will further discuss this point in light of the free energy calculations present below.

**RNA Equilibrium Dynamics**

The binding of NP to RNA is thought to be nonspecific, given the RNA bound structure was crystalized with random, cellular RNA. However, in all six subunits of the crystal structure the third RNA position was always a purine residue (9). In the other RNA positions, except for position 8, the nucleotide could not be unambiguously determined. The eighth position was also
purine, but not all subunits contained the eighth nucleotide. This observation led to the speculation that there may be a partial sequence specificity governing the NP-RNA binding process. Our simulations are supportive of a partial specificity as we observe variable structural fluctuations of RNA positions in our simulation of the A, G and K-subunits. The most stable position is position 3 (Fig. 2.3B) consistent with the experimental observation that position 3 may be a high affinity site for binding purine residues. However, this result maybe a consequence of the structure determination process. By having a better resolved electron density at the third position, the atomic positions of that base could be determined at higher resolution resulting in increased stabilization at that position during the MD simulations.

Figure 2.3. Lassa NP and RNA structural dynamics in the Open-RNA bound state. A) The $\alpha6$ RMSD after global alignment to the closed state (same as Fig. 2.2B) for simulations started from the A and G subunit structures from PDBID: 3T5Q. B) RNA RMSF for the subunit simulations and also averaged over the three RNA bound simulations (subunits, A, G and K).

**Principle Component Analysis**
To analyze the global dynamics of the open and closed states, principal component analysis (PCA) was performed. The two principal component modes which capture the most variance (PC1 and PC2) were determined for the simulations initiated in the known crystallographic states: Open-RNA and Closed-apo. Mode projections were performed for all four simulations, and then the RNA bound simulations were combined to generate a pseudo-free energy surface in the PC degrees of freedom. Similarly, the apo simulations were combined to generate the apo free energy surface. The Open-RNA modes are presented in Fig. 2.4A-B. The dominant motions in both PC1 and PC2 involve twisting of $\alpha$6. PC1 is a lateral twisting of $\alpha$6 and its adjacent loops, while PC2 is a rotation of the helix where the N-terminal loop is moving upward while the C-terminal loop is moving downward and inward. The energy surface when RNA is bound (Fig. 2.4C) shows a contiguous energy surface indicating large energy barriers do not separate the open and closed states. A transition pathway of closing can be inferred from the energy surface (as indicated by the arrow). To move along this transition path, the system would move positively along PC2 followed by positive movement along PC1 followed by negative movement along PC2 to reach the closed configuration. In contrast, the apo surface is discontinuous (Fig. 2.4D) indicating an energy barrier separate the open and closed states; therefore, this transition was not sampled in the equilibrium simulations.
**Figure 2.4.** PCA of Open-RNA Dynamics. A-B) Mode 1 and 2 are imposed on the open state structure, where the arrows represent positive motion in the mode direction. C-D) The pseudo-free surfaces of the RNA-bound and apo simulations projected on the modes shown in (A) and (B). The starting and end points of the simulations are shown with the white shapes, where the Open state simulation begins at the circle and ends at the square and the Closed state simulation begins at the triangle and ends at the diamond.
An analogous analysis was performed using the principle components from the Closed-apo simulations, which are presented in Fig. 2.S2. While in that case the open and closed state simulations overlap for both the RNA bound and apo simulations, there is a lower barrier separating the open and closed state when RNA is bound, qualitatively consistent with the Open-RNA PCA.

**Metadynamics**

We have employed well-tempered metadynamics (36) to evaluate the free energy landscape governing the mobility of α6. The collective variable (CV) we bias in the metadynamics simulations is the mean-squared-displacement (MSD) of α6 from the closed conformation. Metadynamics was performed for 300 ns for the apo system and 370 ns for the RNA-bound system, which produce well-converged free energy surfaces (FESs, see Fig. 2.S3-S4). The converged FESs are presented in Fig. 2.5A as well as structures along the binding pathways (Fig. 2.5B). The apo energy surfaces shows the closed state to be separated from the open state by ~15 kcal/mol. The presence of RNA has a drastic effect to flatten the landscape. The open state is still the global minimum, but the energy separation between open and closed is less than 1 kcal/mol, which would allow for a subpopulation of closed states to exist at physiological temperatures. There is a crossing in the energy surfaces around 9 Å, which provides a prediction of the transition state structure that would bind RNA. In the apo metastable state just before the crossing (Fig. 2.5B structure II), the gating looping (residues 232-243) is in an up position,
occluding the binding site. As α6 moves further toward the open state, we observe the gating loop moves down to allow accessibility to the binding groove.

The metadynamics FES provides explanations for several observed behavior of the various systems in equilibrium MD. The flattened energy surface in the presence of RNA explains why spontaneous closing motions are observed in the equilibrium simulations (Fig. 2.2B, 3A). Furthermore the Open-apo simulation did not transition to the closed configuration and this may be due to the energy barrier around 15 Å. The large RMSDs (Fig. 2.2A) in the open state equilibrium MD simulations are consistent with the flat energy landscape in the regions beyond ~15 Å on FES. Whereas the FES near the closed conformation (<5 Å) display well-defined energy minima which restrict the conformational dynamics resulting in low RMSDs in the equilibrium simulations.

We have examined the correlated motions for several states of the system along the RNA binding pathway. We consider the apo global minimum state (state I, Fig. 2.5B), the apo metastable state prior to RNA binding (state II, Fig. 2.5B), the transition-state with RNA bound (state III, Fig. 2.5B) and the RNA global minimum (state IV, Fig. 2.5B). The dynamic cross-correlation matrices (DCCM) for these four states are presented in Fig. 2.5B. In the apo minimum energy state (State I), the system does not show significant correlations, but as it moves to the metastable state (State II) significant correlated motions become present. In particular α6 develops strong anti-correlation with regions surrounding the RNA-binding groove, which can be interpreted as expansion and contraction motions. Anti-correlated motions surrounding the RNA binding site is not a unique
feature to the Lassa NP; this behavior has also been observed in several viral RNA dependent RNA polymerases (RdRp) (42, 43). Once the RNA strand binds NP (State III) the correlations remain strong throughout the protein, but the correlations between \( \alpha 6 \) and the rest of the protein are diminished and display positive correlations as well. When the system reaches the stable RNA bound state (State IV), the correlations are diminished throughout the protein and the protein exhibits correlations most similar to the apo minimum energy state. This presents a mechanism in which the apo protein transitions to the excited metastable state where \( \alpha 6 \) becomes strongly (anti)correlated with regions of the protein, promoting fluctuations that increase access to the binding groove to allow for RNA to bind.

Another change in the DCCM between State III and State IV is in the correlations between residues 50-70 which are part of helices 3 and 4 that lie below the RNA binding groove. There are anti-correlated motions between this region and residues 150-160 which are on strand 11 which connects to \( \alpha 6 \). There are contacts between strand 11 and the the tips of \( \alpha 3 \) and \( \alpha 4 \) in State III, which become disrupted in State IV. These contacts act to restrict the mobility of \( \alpha 6 \) and excite anticorrelated motions. The positioning of \( \alpha 6 \) and the correlated motions between this region and \( \alpha 3 \) and \( \alpha 4 \) maybe important for facilitating RNA into the binding groove.
Figure 2.5. Free energy surface of Lassa NP α6 opening. A) FESs obtained from well-tempered metadynamics in the absence (apo) and presence of RNA. B) Stable and transition state structures along the transition pathways. The structures were determined by partitioning the metadynamics simulations into subspaces in the CV-space. The average structure in each subspace was determined and the presented structures are those structures with the smallest RMSD to the average structures. α6 is blue, RNA is yellow and the gate loop (resid 232-243) is red. Below each of the structures are dynamic cross-correlation matrices (DCCM) from the same regions of the metadynamics CV-space. State I = Apo RMSD 0-5.5 Å; State II = Apo RMSD 7-8.6 Å; State III = RNA Bound RMSD 8-9.2 Å; State IV = RNA Bound RMSD 14-15.3 Å. The vertical black lines in the DCCMs outline the α6 residues. X-Y axes are in the DCCMs are in units of residue number and the scale bar in State IV applies to all plots.

Crystal Contact Analysis

The minimum energy states in the apo and RNA bound metadynamics simulations roughly correspond to the known crystal structures. However, we do observe that the RNA bound state to be more closed in the presence of RNA than observed in the crystal structure. The RMSD values
of the α6 configurations (after global alignment) in the crystal structure compared to the closed configuration range from 23.4-26.1 Å, whereas our global minimum is significantly shifted inward to ~15 Å. Furthermore, given the propensity of α6 to undergo closing motions in the equilibrium simulations, we have examined the energetics of non-native contacts in the crystal structure to understand their role in stabilizing the open conformation.

Based upon inspection of the open state crystal structure (PDBID: 3T5Q), significant contacts between α6 from various subunits and neighboring proteins can be observed. To quantify the strength and favorability of these interactions we performed an MM/GBSA calculation to measure the binding affinity of α6 from subunit I to the interface created by neighboring subunit A within the same asymmetric unit and A and G subunits from a neighboring asymmetric unit (Fig. 2.6). Based upon the MM/GBSA calculation a binding energy of -245.51 kcal/mol was found. This protein-protein binding energy is certainly an overestimate given the entropy loss is not considered here, as has been done in another crystal contact study (44). To provide a basis for comparison we also performed MM/GBSA on the RNA binding to the open and closed states as well as the opening transition in the absence and presence of RNA. The results are presented in Table I. These calculations support that the crystal contact energies are significant and have the potential to alter the conformational state of the protein. Furthermore, these calculations show that spontaneous opening of the apo protein is unfavorable (ΔG = 164.21 kcal/mol), which is consistent with our metadynamics FES but that RNA binding (ΔG = -143.39 kcal/mol) alone is not sufficient to overcome the unfavorable opening energy. The additional favorable interaction of
the open state with the neighboring subunits could provide the additional energy required to stabilize the open-RNA bound state.

**Figure 2.6.** Monomer I (green) contacts between α6 and neighboring subunits in the crystal structure.
Table 1. MM-GBSA calculations for RNA Binding to open and closed conformations of the NP, Opening of the NP in the presence and absence (apo) of RNA and binding of Monomer I to neighbor subunits in the crystal structure. All units are in kcal/mol.

<table>
<thead>
<tr>
<th></th>
<th>RNA Binding</th>
<th>NP Opening</th>
<th>Protein-Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>APO</td>
<td>RNA Bound</td>
</tr>
<tr>
<td>NP Open</td>
<td>&lt;ΔEmm&gt;</td>
<td>-1290.93</td>
<td>252.92</td>
</tr>
<tr>
<td></td>
<td>&lt;ΔGgb&gt;</td>
<td>1157.50</td>
<td>-91.13</td>
</tr>
<tr>
<td></td>
<td>&lt;ΔGnp&gt;</td>
<td>-9.96</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>&lt;ΔGbind&gt;</td>
<td>-143.39</td>
<td>164.21</td>
</tr>
<tr>
<td>NP Closed</td>
<td>-1158.28</td>
<td>120.27</td>
<td>522.17</td>
</tr>
<tr>
<td></td>
<td>-91.13</td>
<td>-72.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.42</td>
<td>3.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-143.39</td>
<td>164.21</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

Using extensive MD simulations we have described the influence of RNA binding on the structure and dynamic properties of the Lassa NP. We have shown that the energy landscape in the presence of RNA is much flatter, with respect to the α6 position. This implies that α6 will sample a wide range of conformations that may be important in mediating interactions with other viral proteins, including interactions with other NPs. Indeed, other negative sense single-stranded RNA viruses are known to have small mobile helical elements which undergo conformational
changes to switch between RNA binding and non-binding modes (45). The dynamics of α6 may also be important in understanding the interaction between NP and the L protein, which is part of the RNP. Transient exposure the RNA due to fluctuations in the α6 position may provide an opportunity for L to engage with RNA and induce further conformational changes in the NP to allow for replication or transcription to proceed.

In addition to gaining insight into the influence of RNA on the stability of the known conformational states we are able to detect an intermediate structure along the RNA binding pathway. A metastable state exists when α6 is displaced from the closed configuration by ~8 Å in RMSD space. In the metastable configuration NP develops strong correlated motions between various regions of the protein and importantly, anticorrelated motions develop between α6 and several helices surrounding the RNA binding groove. These anticorrelated motions allow for the binding crevice to expand, pushing NP toward the transition-state where it would bind to the RNA strand. Binding to RNA would then induce further movement of α6 out to ~15 Å where the system stabilizes and correlated motions are suppressed. The location of α6 when bound to RNA is expected to sample many conformations, but to primarily exists in a more closed form than experimentally observed. We rationalize this difference as due to the favorable interactions between α6 and neighboring NP proteins in the crystal lattice, which would likely be absent in solution.
References


Figure 2.51. Comparison of open RNA bound dynamics of different subunits from PDBID:3T5Q. A) The equilibrium simulations are projected onto the PC1-PC2 subspace of the K-subunit. All three simulations sample the region where the K subunit simulation ends (see Fig. 4). B) The three different starting configurations of the K, G and A subunits.
Figure 2.S2. PCA of Close-apo Dynamics. A-B) Mode 1 and 2 are imposed on the open state structure, where the arrows represent positive motion in the mode direction. C-D) The pseudo-free surfaces of the RNA-bound and apo simulations projected on the modes shown in (A) and (B). The starting and end points of the simulations are shown with the white shapes, where the Open state simulation begins at the circle and ends at the square and the Closed state simulation begins at the triangle and ends at the diamond.
**Figure 2.S3.** Convergence of apo FES by metadynamics. The FES was computed after every 5 ns. The difference between the current and initial FES is shown in (A), and the difference between the current and previous (5 ns before) is shown in (B). In both plots the Free energy change is averaged over the 256 grid points in the CV space. At the end of the simulation each point on the FES is changing by less than 0.05 kcal/mol.

**Figure 2.S4.** Convergence of RNA-bound FES by metadynamics. The FES was computed after every 5 ns. The difference between the current and initial FES is shown in (A), and the difference between the current and previous (5 ns before) is shown in (B). In both plots the Free energy change is averaged over the 256 grid points in the CV space. At the end of the simulation each point on the FES is changing by less than 0.11 kcal/mol.
Chapter 3: Markov State Modeling Reveals Large Structural Changes During the Lassa Virus Nucleoprotein Trimer to Monomer Transition
Markov State Modeling Reveals Large Structural Changes During the Lassa Virus Nucleoprotein Trimer to Monomer Transition

Jason G. Pattis and Eric R. May

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT

Abstract
Lassa virus is an enveloped negative strand RNA virus that causes a severe hemorrhagic fever, Lassa fever, which leads to about 5000 deaths per year. Lassa contains a nucleoprotein (NP) which encapsulates the viral genomic RNA forming the ribonucleoprotein (RNP). The first crystal structure solved of the Lassa virus NP, was a trimer structure with no exposed RNA binding site. A subsequent crystal structure of only the N-terminal domain was co-crystalized with RNA bound showing several large conformational changes, including the opening of helix 6, the loss of helicity of helix 5 and shifting down of a loop. The suggested model is that NP forms a trimer to keep the RNA gate closed, preventing off target binding. Previous work has investigated helix 6 opening and has suggested that RNA may make contact with a partially open NP, which may facilitate full binding. The current work investigates the scenario in which the trimer is disrupted to observe if a monomeric NP undergoes a conformational change that would allow for increased access to the RNA binding pocket. We have conducted molecular dynamics simulations using long time scale simulations on specialized hardware and a two stage adaptive sampling scheme to sample this transition. From the trajectories (totaling over 27 µs of data), a Markov State Model was constructed to describe the energetics of the transition, which reveals an energetically favorable conformational change. The most significant changes occur at the domain interface including the
shifting out of helix 8 and 9 which may allow room for RNA to contact the RNA binding pocket, supporting a model in which significant structural reorganization of the NP is required for RNP formation.

**Introduction**

Lassa virus is the causative agent of Lassa fever, a severe hemorrhagic fever which is estimated to infect between 100,000 and 300,000 people per year, primarily in western Africa, and leads to 5000 deaths per year(1, 2). Transmission of Lassa virus to humans occurs through contact with Lassa infected rodent’s urine and feces. Both the NIH and the WHO have classified Lassa as a category A priority pathogen due to its high risk to public health. Lassa is also a threat to other parts of the world as it is the most frequent hemorrhagic fever to spread to Europe and the United States(1, 3, 4) There have been an increasing number of large outbreaks in the last few years(5, 6).

Lassa currently has only one vaccine that recently entered stage I clinical trials but no vaccines that have progressed further than stage I(7). The other treatment options are limited to only preventative care, with Ribavirin being helpful if administered early after infection(7). There is a significant need for increased understanding of the Lassa protein structures and interactions to aid in the development of new therapeutics.

Lassa is an enveloped virus that has two strands of single stranded RNA. It has four genes and uses an ambisense coding strategy where the polymerase and nucleoprotein (NP) are transcribed first, then the Z protein and glycoprotein precursor protein are transcribed later in the life cycle after RNA replication(8). The single stranded RNA genome is encapsulated by the NP, forming the ribonucleoprotein (RNP) to protect the RNA from detection by the immune system. The RNP also acts as a scaffold for the polymerase. The RNA genome segments have reverse
complementary tails, which stick together and form the promoter for the polymerase and cause the RNA to make a hairpin(9). While the 3D structure of the Lassa RNP is not known, cryo-EM structures of similar negative strand RNA viruses show the RNP as a twisted helical structure(10, 11).

The NP contains 569 residues comprising two domains which are connected by an unstructured linker segment. The N-terminal domain is involved in RNA binding(12) while the C-terminal domain contains an exonuclease to digest double stranded RNA(13). The first crystal structure of Lassa NP showed NP as a trimer with no exposed RNA binding site, Figure 3.1(14). A subsequent study was able to co-crystalize the N-terminal domain with a short strand of RNA bound to the NP(12). In the RNA-bound structure some large conformational differences from the apo trimer structure were observed, namely the shifting out of helix 6, the loss of helicity of helix 5 and the shifting down of a loop to the left of helix 6. Hydrogen deuterium exchange was also performed on the full length NP as well as a double mutant near the NP-NP interface, far from the RNA binding site, which elutes as a dimer. This double mutant showed much higher solvent

![Figure 3.1](image-url) Lassa NP structure. A) Trimer structure from PDBID: 3MWP, each subunit is colored differently. B) Monomer structure, also from PDBID: 3MWP. In both panels helices 5, 6, and 17 are colored red, blue and cyan, respectively. Both the trimer and monomer consist of chain B subunit(s).
exposure of helix 5, 6, and the RNA gate suggesting that disruption of the trimer may allow easier opening of the RNA binding pocket. High concentrations of RNA were unable to cause RNA binding.

The model put forward by Hastie and coworkers (12) was that NP trimerizes during viral transcription to prevent off target RNA binding and to build up a large store of NP. Then during RNA-replication some signal will break up the NP trimer, causing the C-terminal domain to shift away from the N-terminal domain and expose the RNA binding pocket. It was proposed that the shifted conformation the NP can bind RNA allowing for organization of NPs onto the newly formed viral genomic RNA and formation of the RNP. There is still much unknown about the molecular mechanisms that Lassa uses to function and more work is needed to understand them. Furthermore, these protein conformational changes may expose novel targets for drug discovery which block the proteins function.

The dynamics of the N-terminal domain have been studied previously (15). It was found that the slow global motions primarily involve helix 6 motions. Apo NP was found to have a large barrier for helix 6 opening whereas RNA bound NP was found to have a fairly small barrier for opening. This along with correlated motions in intermediate states suggested a mechanism where the RNA binding pocket starts to open allowing RNA to form some contacts with the NP, which stimulate further conformational changes in NP to facilitate RNA becoming fully bound. It was also found that helix 6’s position in the open state is influenced by crystal contacts and when NP is free in solution the helix 6 position may be closer to the RNA binding pocket. This observation was also noted in another simulation study of Lassa NP (16).

In the current work we extend the study on the dynamics of Lassa NP from just the N-terminal domain to the full-length NP. The Hastie model and the hydrogen deuterium exchange
suggest that that disruption of the trimer would allow shifting away of the C-terminal domain and easier opening of the RNA binding pocket. In the present study, one monomer from the trimeric APO crystal structure is solvated and simulated on the multi-microsecond timescale. A large conformational change is witnessed which is resampled using a two stage adaptive sampling scheme allowing the construction of a Markov state model (MSM) describing this transition. Potential value in this study includes the identification of metastable intermediates along the transition, which could be targeted by small molecules to interrupt the viral life cycle.

Methods

Simulation details

Anton simulations were prepared starting from chain B of the full length trimeric nucleoprotein (PDB ID: 3MWP)(14) which included one zinc atom. Missing loops were modeled in using MODELER(17, 18). The protein was solvated in a rectangular box with a 10 Å buffer on all sides and NaCl was added to a concentration of 150 mM. This system totaled approximately 137,000 atoms. Energy minimization and equilibration were run using NAMD v2.9(19). 5000 steps of energy minimization were run followed by 5 ns of NPT equilibration with 1 kcal/mol restraints on all protein backbone atoms. This was followed by 5 ns of equilibration with 0.25 kcal/mol restraints on backbone atoms, and then 10 ns of unrestrained MD. The CHARMM27(20, 21) force field including CMAP corrections and CHARMM TIP3P water model were used. 12 Å cutoffs were used with a switching function on van der Waals interactions starting 8 Å. Full electrostatic interactions were computed using the PME method.

Anton simulations were run in the NPT (isothermal, isobaric) ensemble, a Nose-Hoover thermostat at 300 K, and a Martyna-Tobias-Klein (MTK) barostat with isotropic scaling at 1
The simulations used the CHARMM27 force field with CMAP correction and the CHARMM TIP3P water. Frames were saved every 240 ps. The multigrator (22) was used for the integrator. The multigrator calculated bonded and near non-bonded forces every 2 fs and far non-bonded forces, which are the long-range electrostatics in the Ewald decomposition, are calculated every 6 fs. Gaussian Split Ewald (23) was used for the far non-bonded term.

GROMACS simulations were run starting from the NAMD equilibrated starting structure or from solvated frames pulled out of the Anton simulation. They were run using GROMACS 5.0.1 with the CHARMM27 force field (20, 21, 24) with CMAP corrections and the CHARMM TIP3P water. These simulations were run using a leap-frog stochastic dynamics integrator using a timestep of 2 fs. Frames were saved every 240 ps. Short ranged non-bonded interactions were calculated with a cutoff of 12 Å and where smoothly shifted to zero at the cutoff. For the Lennard-Jones potential the shifting began at 10 Å, and the Coloumb potential was shifted over the whole range (starting from 0 Å). PME was used for long-range electrostatics. The temperature was maintained at 300 K using the v-rescale algorithm (25) and pressure was maintained at 1 Atm using the Parrinello-Rahman isotropic pressure coupling.

**Adaptive Sampling and Markov State Model Construction**

MSMBuilder 2.8 (26) was used for all steps in constructing the Markov state model (MSM) except for the Bayesian MSM and the Chapman-Kolmogorov test, which were done in pyEmma 2.4 (27). A flow chart of the steps used for adaptive sampling and Markov State Model Construction is shown in Figure 3.S1. To describe the conformational space the raw Cartesian coordinates were transformed into internal coordinates (featurized). We sought to identify pairwise residue contacts which changed during the Anton simulation. The specific set of coordinates we
choose were the pairwise alpha carbon distances that were less that 13 Å apart for at least 2 % of
the simulation and had a standard deviation of at least 1.5 Å during the 4 μs Anton simulation.
The linker loop between domains was excluded from consideration in identifying the feature set.
Our criteria led to 3290 pairwise distances. A visualization of these distances can be seen in Figure 3.S2.

Time-lagged Independent component analysis (TICA) was used to transform the data into
kinetic coordinates(28, 29). In TICA a covariance matrix and a time-lagged covariance matrix of
features (filtered C alpha distances) is put through a generalized eigenvalue problem. A set of
linear combinations of input features (TICA components) and corresponding eigenvalues are
returned.

In an MSM the number of transitions between discrete states are counted in some time
interval (lagtime) τ from all trajectories to form a count matrix. From this a transition matrix is
constructed which describes the conditional probability of transitioning from state i to state j at
lagtime τ. MSMs were estimated using a maximum likelihood approach and a sliding window(30).
All MSMs are reversible and obey detailed balance(30).

In order to improve sampling in undersampled regions of the NP transition phase space a
counts-based adaptive sampling protocol was performed. In this approach, the data is featurized
and TICA is performed with a lag time of 7.2 ns being fit to the Anton trajectory and using the
kinetic map algorithm(31) to scale the eigenvectors. K-centers clustering(32) is performed on the
top 5 independent components to separate the space into 100 evenly spaced clusters. An MSM is
then generated with a lag time of 3.84 ns. The sum of each row of the count matrix is used to see
how well each state is sampled. The top ten states with the lowest counts are then subjected to two
independent 30 ns simulations with different starting velocities. The initial structure for these
simulations is determined by identifying the frame with the smallest nearest neighbor distance in 5 dimensional TICA space to the average TICA values for that cluster. This protocol was repeated for 12 rounds, generating 7.2 μs of additional simulation data. An example visualization of the states selected for counts-based adaptive sampling can be seen in Figure 3.S3.

In order to reduce the uncertainty in the MSM a second stage of adaptive sampling was performed. In this stage the data is featurized and TICA is performed with a lag time of 3.42 ns using the commute map algorithm(33) . Again K-centers clustering is performed on the top 5 independent components to separate the space into 100 evenly spaced clusters. Next a Bayesian MSM using a sparse prior(34) is constructed with 100,000 possible transition matrices, which could have created the raw data. The standard deviation in the population is calculated for each state. The top five max flux pathways from the starting state to the highest population cluster are calculated from the net flux matrix(35) . The ten states, which are part of one of the top five maximum flux pathways and have the highest standard deviation in their population are subjected to two independent 30 ns simulations with different starting velocities. This stage of adaptive sampling was repeated for 16 rounds, generating 9.6 μs of additional simulation data. A sample visualization of the states selected for uncertainty-based adaptive sampling can be seen in Figure 3.S4.

The TICA lag time, number of TICA components, and number of clusters were chosen based on the generalized matrix Rayleigh quotient (GMRQ)(36). It was found that the variational theorem(37, 38) bounds the GMRQ from above as the sum of the first \( m \) eigenvalues when avoiding over fitting through cross-validation. This allows the GMRQ to be used as a score of how different parameter choices affect how well the MSM captures the slow subspace of the system’s propagator. Here trajectories over 80 ns were split into 40 ns chunks giving 860 trajectories. 30
iterations of shuffle split where 50 % of trajectories were placed in the training set and 50 % of trajectories were placed in the validation set using Scikit-learn version 0.18.2. The ergodic cutoff was turned off and a prior count of 0.00000001 was placed in each cluster to ensure the calculation was done on the entire state space. The top three eigenvalues were examined making the max possible GMRQ score 4.

The model was validated using the implied timescale test and the Chapman Kolmogorov test. The implied timescales of a model are \( t_i = -\frac{\tau}{\ln|\lambda_i|} \) where \( \lambda_i \) is the i-th eigenvalue of the MSM. In the implied timescale test the top 7 implied timescales are plotted for MSMs of increasing lagtimes (Figure 3.4). At a Markovian (memory-less) lag time, increasing the lag time should give the same implied timescales because the models are capturing the same processes. The Chapman Kolmogorov test compares the transition probabilities between the left and right side of equation (1).

\[
P(k\tau) = P^k(\tau) \quad \text{eq. (1)}
\]

Here the right side of the equation is the original transition matrix at lag time \( \tau \) propagated to the k-th power (prediction) and left side is a new transition matrix made at lag time \( k\tau \) (estimation). For a full discussion see Prinz et. al.(30) This test is performed using PCCA+(39, 40) to lump the 75 state MSM into 5 macrostates due to the MSM having 4 slow timescale processes (\( > 200 \) ns) and a large gap in eigenvalues between the 4\textsuperscript{th} and 5\textsuperscript{th} slowest timescale process.

The free energy surface was constructed using the plot free energy tool from MSMExplorer version 1.2.0dev0(41). Here 1,500,000 frames are randomly selected with the probability of picking a frame is equal to the population of the state it is in. The density of frames is calculated by a bivariate kernel density estimate. This probability is converted to a free energy using \( F = -\).
The mean first passage time was calculated using transition path theory (35). Clusters were lumped into macrostates using PCCA+ (39, 40).

**Results and Discussion**

*Long timescale equilibrium simulations reveal domain level reorganization*

The Lassa NP is presumed to go through a large conformational change to switch from a trimeric structure where NP is being stored to the RNP structure, where the NP encapsulates the single stranded genomic RNA (12). Understanding this transition is important to gain insights into how the molecular machinery in Lassa virus acts to carry out its function and regulate its lifecycle. This large conformational change may also reveal pockets that can be targeted in drug discovery to identify novel inhibitors. The Hastie model of RNP formation suggests that disruption of the trimer would allow the C-terminal domain to swing away from the N-terminal domain and increase opening of helix 6 and helix 5. This is supported by hydrogen/deuterium exchange data, which indicated increased solvent exposure in helix 5, 6, and 17 (see Figure 3.1) in a double mutant which disrupts trimer contacts. In order to provide additional evidence for this, a monomer from the trimer crystal structure with the missing loops modeled in was placed through the Frustratometer2

![Figure 3.2](image)

**Figure 3.2.** Mutational frustration of Lassa NP monomer in the trimeric state (PDBID 3MWP.B), shown from front and top views. Green lines represent residue pairs which are minimally frustrated and red lines represent highly frustrated residue pairs.
web server(42) to identify highly frustrated residue pairs, especially for pairs between the N and C-terminal domain, where a conformational change could relieve the frustration. The mutational frustration is calculate by comparing the energy of a native residue pair interaction with the average of many different mutations of that pair. Figure 3.2 shows that there are many highly frustrated residue pairs between helix 5 and helix 17. There are also highly frustrated residues between helix 6 and its surrounding loops and the C-terminal domain. This may suggest that the formation of the trimer is placing pressure on the domain interface and that without the trimer contacts these domains may reorganize to reduce frustration and adopt a more favorable configuration.

Previous simulations have shown that in an isolated N-terminal domain there is a large energy barrier for helix 6 to open(15). Another possibility is that the C-terminal domain shifting away could be coupled with a conformational change in helix 6 and the RNA binding pocket. To investigate this further one monomer from the trimeric structure was isolated, solvated and run on specialized hardware to generate long-time scale (multimicrosecond) simulations to observe if trimer disruption leads to structural relaxation to an alternate NP conformation.

A 4 µs and a 825 ns simulation were run on the Anton supercomputer. In the 4 µs simulation a large conformational change was observed (Figure 3.3). Here the C-terminal domain shifted to

Figure 3.3. Overlay of initial and final structures of the 4 µs Anton simulation. The initial structure is consistent with the trimeric crystal structure (PDBID 3MWP.B), and is colored in cyan (N-term) and brown (C-term). The final structure is colored in blue (N-term) and purple (C-term). Structures are shown from front (0°), side (90°) and back (180°) views. Structures were aligned along the α5 helix.
the back of the N-terminal domain, helix 6 shifted in, and helix 9 and 10 shifted out. Because this large conformational change only occurred once in the 4 μs trajectory more sampling was needed to provide statistical significance of the thermodynamics and the kinetics of this transition. We used the tools of adaptive sampling and Markov State Modeling to improve the sampling of this transition and for analysis. Additional trajectories were spawned from the beginning, end, and a few midpoints from each of the Anton trajectories. Next a counts based adaptive sampling protocol was run for 12 rounds. Here new simulations are spawned from areas with poor sampling. This has been shown to be an efficient method to discover new states (43). Next population uncertainty adaptive sampling was performed. Here the uncertainty is calculated using a Bayesian MSM then new simulations are spawned from clusters, which have a high standard deviation of possible stationary distributions. This second phase of adaptive sampling was run for 16 rounds, followed by construction of our final MSM.

Model Selection and Validation

An initial guess of parameters suggested an MSM lag time of 3.84 ns would generate a Markovian kinetic model. We next used the generalized matrix Rayleigh quotient (GMRQ) score to select the number of clusters, TICA lag time, and number of TICA components. K-means clustering was used for all models and commute map was used for TICA. The highest validation GMRQ score was from a TICA lag time of 18 ns, 3 TICA components, and 75 clusters (Figure 3.S5). These parameters were used for the final model which we analyze and present here. The implied timescales of the MSM processes are related to the transition probability matrix eigenvalues by eq 2

\[
t_i = -\frac{r_i}{\ln \lambda_i}
\]
where, \( \tau \) is the MSM lag time, \( \lambda_i \) are the eigenvalues and \( t_i \) are the implied timescales. An implied timescales test was run (Figure 3.4) by scanning lag time values. We identified the implied timescales were relatively insensitive to the lag time chosen (beyond \( \sim 2 \) ns), which indicated the model is Markovian. We selected a lag time of 3.84 ns, which was in the range of lag times where the implied timescales were not rapidly changing and produced the largest timescale for the slowest process. A Chapman–Kolmogorov test was run with a Bayesian MSM with 1000 transition matrices to construct a 95% confidence interval. Five macrostates were determined, and this number of macrostates was chosen due to the presence of four slow timescale processes. The Chapman–Kolmogorov test shows excellent agreement between the estimated and the predicted transition probabilities (Figure 3.S6).
MSM Reveals Multiple Intermediate States

The model has five kinetic macrostates which are shown on top of the free energy surface (Figure 3.5, and without macrostates in Figure 3.S7). The starting structure is shown with a green x. NP moves from its starting structure linearly through the different locally stable intermediates finally reaching the most favorable state(F). Representative structures are shown in (Figure 3.6).

The mean first passage time from the starting state to the most favorable state was 20.47 μs, whereas the mean first passage time from the most favorable state back to the starting structure was 279.94 μs.

The transition from the starting state to the first intermediate (I1) involves separation being created between contacts in the N-terminal and C-terminal domain primarily between helix 5 and helix 17. A salt bridge between ARG115 to ASP375 is broken, as well as contacts between ASP557 to LYS110 and TRP331 during this stage of the transition. The gating loop shifts up and the loop between helix 6 and helix 7 shifts away from the C-terminal domain. Contact is also broken between THR34 and THR216 connecting the RNA gating loop (yellow) and helix 9 (red) (Figure 3.6). Transition between intermediate state 1 and intermediate state 2 involves the bottom of the
C-terminal domain moving away from the bottom of the N-terminal domain while the top of the C-terminal domain maintains contact with the N-terminal domain.

From intermediate state 2 to intermediate state 3 helix 10 (red) shifts out, the left side of the RNA gating loop moves down while the right side moves up. In addition, helix 6 moves in closer to the RNA binding groove. Also, the C-terminal domain shifts back and a salt bridge is formed between ASP504 and LYS65 and contacts between ASN 168 and LEU 505 are formed which connects the bottom of the C-terminal domain to the back of the N-terminal domain. From intermediate state 3 to intermediate state 4 separation is created between THR210 of helix 10 and GLN14, while the domains start to compact together more. From intermediate state 4 to the most favorable state (F) the domains compact even more. This structure is very stable according to the MSM and also has much fewer frustrated contacts according to the Frustratometer (Figure 3.7).
Several of the residues we have identified as having shifting contacts during the transition have been implicated to be functionally important for transcription. This was measured through a mini-genome assay, and mutations to ARG115, LYS110 and TRP331 all had significant effects on the transcriptional activity compared to wild-type (12). In addition, the structural transition we
observe has some consistency with hydrogen-deuterium exchange mass spectrometry data. In a structure which the trimer has been disrupted through protein-protein interface mutations, increased exchange (compared to wt trimer) was observed in helices 5, 10 and 17, which are regions we also observe structural changes(12). Lastly, we want to emphasis the structural reorganization of helix 10 on the left side of the NP, may facilitate RNA binding (Fig 3.5, red motif). Helix 10 can be considered a cap on the left side of the RNA binding groove and in our most favorable state helix 10 is swung away from the binding channel provide a potential access point for RNA to enter. The orientation of helix 10 in the most favorable and initial structure are compared with the RNA inserted from the N-terminal RNA bound structure (PDB ID:3T5Q), in Figure 3.8. The distance between the top of helix 10 (THR216) and the start of helix 11 (Leu 248) is changed by almost 10 Å, from 9.6 Å in the starting state (Figure 3.8 blue) to 19.2 Å the most favorable state (Figure 3.8 cyan).

Figure 3.7. Mutational frustration of the most favorable microstate (F), shown from front and top views. Green lines represent residue pairs which are minimally frustrated and red lines represent highly frustrated residue pairs.
Our simulation study supports a model in which, when the Lassa NP trimeric structure is disrupted, the C-terminal domain moves away from the N-terminal domain, swings back, then compresses in making new contacts in the back of the N-terminal domain. This observation is consistent with the qualitative model put forth from Hastie et al, when they determined the RNA bound conformation of the Lassa NP N-terminal domain. The domain scale movements we observe are coupled with shifting in of helix 6, movement of the RNA gating loop, and shifting out of helix 9 and 10. This shows that loss of trimer contacts can cause global conformational changes in the RNA binding pocket. The shifting out of helix 9 and 10 exposes some of the RNA binding pocket. This may provide a surface or a gateway for RNA to make initial contact. The

Figure 3.8. Helix 10 displacement. Comparing the position of helix 10 in the initial (blue) and most favorable (cyan) conformations. RNA (orange) is copied in from PDB ID: 3T5Q.k. Structures are aligned on helix 11 in back of binding groove (residues 245-259). Ca of residues 216 and 248 are shown as green and red vdW spheres for the most favorable and initial conformations, respectively.

Conclusions

Our simulation study supports a model in which, when the Lassa NP trimeric structure is disrupted, the C-terminal domain moves away from the N-terminal domain, swings back, then compresses in making new contacts in the back of the N-terminal domain. This observation is consistent with the qualitative model put forth from Hastie et al, when they determined the RNA bound conformation of the Lassa NP N-terminal domain. The domain scale movements we observe are coupled with shifting in of helix 6, movement of the RNA gating loop, and shifting out of helix 9 and 10. This shows that loss of trimer contacts can cause global conformational changes in the RNA binding pocket. The shifting out of helix 9 and 10 exposes some of the RNA binding pocket. This may provide a surface or a gateway for RNA to make initial contact. The
model we observe has consistency with experimental data in that increased hydrogen exchange observed in helix 10 and helix 17 could be accounted for by the structural changes we observe. There are other experimental observations including the opening of helix 6 and the loss of helicity of helix 5 that are observed in the RNA bound structure that was not observed in our MSM. However, there are variety of factors which could explain these differences including i) the H/D exchange in the disrupted trimer was still oligomeric (dimers and tetramers) ii) the presence of RNA could induce those changes, where RNA was not present in these simulations while the double mutant NP in H/D exchange had an A260/A280 ratio of 1.3 vs 0.95 of the wildtype (12), indicating that some RNA was present iii) or, despite our extensive sampling (> 25 µs), we were unable to exhaustively explore all relevant conformations of Lassa NP. The H/D exchange was run with a 10 s and 1000 s mixing time (12). These timescales are not accessible with current MD simulations.

The meta-stable and stable conformations we observe may be helpful in structural studies on determination of the Lassa RNP and also provide potential drug targets for anti-viral therapies. An approach could target trapping one of the intermediates with a small molecule which may cause non-native oligomers to form that would be detrimental to the virus. Other negative strand viruses have shown that conformational changes in mobile elements such as flexible helices or loops can play a role in the switch in oligomeric states(44), and we observe helix 6 may play this role for Lassa as new contacts with in the C-terminal domain or a shift in position may allows contact with a neighboring NP and facilitate formation of the RNP.
References


Figure 3.S1. Flow chart of the adaptive sampling protocol used in this study
Figure 3.S2. Visualization from A) Front and B) Top of 3290 CA distances (shown in blue) used as inputs for TICA. Only distances that are less than 13 Å for at least 2% of the Anton simulation and have a standard deviation greater than 1.5 Å are included. The linker loop is also not included.
Figure 3.S3. Visualization of round 8 of counts based adaptive sampling. Centers of 100 clusters from k-centers clustering plotted on A. tic 1 vs 2 and B. tic 1 vs 3. Color represents the sum of the row of the count matrix for that cluster.
Figure 3.S4. Visualization of round 5 of population uncertainty based adaptive sampling. A. Centers of 100 k-centers clusters are shown on tic 1 vs 2. The color corresponds to the standard deviation in their population from 100,000 transition matrices made using a Bayesian MSM. Top 6 maximum flux pathways from the starting structure (green x) to the most populated state.
Figure 3.S5. GMRQ score for training set (red) and validation set (blue) for different tica lag times, number of tica components, and number of clusters.
Figure 3.56. Chapman–Kolmogorov test, showing the transition probabilities between the 5 macrostates for predictions made by propagating the original model (blue) and estimation made by a new transition matrix made at lag times $k\tau$ (black).
Figure 3.S7. free energy surface of final MSM
Chapter 4: Catalytic Domains of Phosphodiesterase 5, 6, and 5/6 Chimera Display Differential Dynamics and Ligand Dissociation Energy

Reproduced with permission from:

The following changes have been made:

- The manuscript has been reformatted to meet University Standards
- Figures numbers have the chapter number prepended

For this work JGP performed the mutual information analysis, ran ligand unbinding simulations, performed analysis of ligand binding simulations, supervised SK and BL, and wrote the manuscript.

SK performed the equilibrium simulations, analyzed the equilibrium simulations, and wrote the manuscript.

BL ran ligand unbinding simulations
Catalytic Domains of Phosphodiesterase 5, 6, and 5/6 Chimera Display Differential Dynamics and Ligand Dissociation Energy Barriers

Title: Catalytic Domains of Phosphodiesterase 5, 6, and 5/6 Chimera Display Differential Dynamics and Ligand Dissociation Energy Barriers

Author: Jason G. Pattis, Shaan Kamal, Boyang Li, et al

Publication: The Journal of Physical Chemistry B

Publisher: American Chemical Society

Date: Jan 1, 2019

Copyright © 2019, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.
Catalytic Domains of Phosphodiesterase 5, 6, and 5/6 Chimera Display Differential Dynamics and Ligand Dissociation Energy Barriers

Jason G. Pattis, Shaan Kamal, Boyang Li, Eric R. May

Abstract
The enzyme phosphodiesterase 6 (PDE6) is a critical component of the visual signaling pathway, and functions to convert cGMP to GMP. The ability of PDE6 to affect cellular cGMP levels leads to deactivation of cGMP-gated ion channels in both rod and cone cells. PDE6 has been difficult to structurally characterize experimentally, though the structures of the closely related PDE5 and a PDE5/6 chimera have been determined by X-ray crystallography. In this work, we employ a computational approach to study the dynamics of the catalytic domains of PDE6, PDE5 and the PDE5/6 chimera. Through equilibrium molecular dynamics (MD) simulations we identify a region of PDE6 (α12) to be correlated to distal regions of the enzyme (H- and M-loops) which surround the catalytic center. These correlations are not observed for PDE5 and we speculate that these unique motions in PDE6 may relate to the high catalytic efficiency of PDE6 and the requirement for an endogenous inhibitory subunit (Pγ). We further investigate the ligand binding pathways and energetics by enhanced sampling simulations (metadynamics) using the inhibitor sildenafil and GMP. The energetics and pathways of ligand binding are consistent with the high efficiency of PDE6 and further implicate the α12 region as an important regulatory element for PDE6 functional dynamics.

Introduction
The enzyme phosphodiesterase 6 (PDE6) is a critical component of the pathway that converts light into the electrical signals that result in vision. PDE6 is a member of the class I cyclic nucleotide
phosphodiesterase superfamily, which contains 11 different PDE enzymes. PDE6 converts cyclic guanosine 3',5'-monophosphate (cGMP) to GMP, a critical and highly regulated step in the visual signaling pathway, and is found in both the rod and cone cells of the eye.⁵

This molecular signaling pathway that results in vision has several major steps. G-protein coupled receptor (GPCR) rhodopsin is activated by light and in turn activates G-protein transducin.⁴ The α-subunit of transducin then displaces the inhibitory PDE6 subunit γ (Pγ) from its position blocking the active site of PDE6.⁴ This allows PDE6 to rapidly hydrolyze cGMP to GMP.⁴ The drop in cGMP concentration deactivates cGMP-gated ion channels, causing hyperpolarization of the cell and then activation of the sensory neurons responsible for vision.⁴,⁵ One of the properties of PDE6 that makes it unique amongst PDEs is that it has a naturally occurring inhibitory subunit, Pγ.⁵,⁶ This relationship allows Pγ to regulate the activity of PDE6 and Pγ is known to selectively inhibit PDE6 and not other PDEs.⁵

PDE5 and PDE6 exist as dimers, with each monomer containing a catalytic domain and tandem GAF domains.⁷ The PDE5 catalytic domain can be expressed and purified as an active monomer,⁸ however there are no such reports for PDE6. In cone cells PDE6 is a homodimer (αα), and this work will focus on the catalytic domains in a monomeric state, as a minimalist approach to its dynamics. PDE6 has not been as well characterized as other PDEs because it is difficult to express in bacteria and purify. As a consequence, a high resolution structure of PDE6 has yet to be determined, although recent work characterized the overall topology of PDE6 through an 11 Å resolution cryo-EM reconstruction.⁵,⁹ A model construct was developed by Barren et. al to gain information of PDE6 structure through the creation of a chimera of PDE5 and PDE6 referred to as PDE5/6.¹⁰ PDE5 and PDE6 have high sequence homology, so substituting sequences unique to PDE6 into the corresponding areas on the PDE5 gene allowed the expression and isolation of
The PDE5/6 chimera is functional, as it can hydrolyze cGMP while also being effectively inhibited by Pγ, making it a reasonable experimental model to study PDE6. In comparing the structure of PDE5 and PDE5/6, it was found that PDE5/6’s H-loop (residues 660 to 683) is 26 Å closer to the M-loop (residues 788 to 811) than in PDE5 (Fig. 4.1), despite the H-loop amino acid sequence being exactly the same in PDE5 and PDE5/6. The H- and M-loops in the PDE5/6 also display helical structure, whereas in PDE5 the loops are unstructured. In PDE6 the regions of the H- and M-loops are believed to be involved in Pγ binding, as observed in the crystal structure of PDE5/6 bound to Pγ (PDBID: 3JWR, Fig. 4.1D). Given that Pγ is required for the regulation of PDE6 it is intriguing that structural differences are observed in between PDE5 and PDE5/6.

The efficiency of cGMP hydrolysis also differs between PDE5, PDE5/6, and PDE6. PDE5 and PDE5/6 hydrolyze cGMP to GMP with low efficiency (~0.55 cGMP/second) while PDE6 is the most efficient member of the PDE family, hydrolyzing cGMP at a rate of ~6,000-8,000 cGMP/second. PDE6’s high efficiency is needed for the near instantaneous process of vision to occur properly. Inhibitors of PDE5 have been developed for the treatment of erectile disfunction, the most notable being sildenafil (Viagra). However, given the high degree of structural similarity between PDE5 and PDE6, sildenafil can inhibit PDE6, which can cause vision impairment side effects.
Figure 4.1. Structure and sequence of PDE catalytic subunits. The structures of PDE5 (A), PDE5/6 (B) and PDE6 (C) are shown with coloring of the H-loop (blue), M-loop (red) and α-helix 12 (yellow). Mg²⁺ (pink) and Zn²⁺ (grey) ions in the catalytic site are shown as spheres. The PDE5 and PDE5/6 structures are taken from PDB IDs: 2H40 and 3JWQ, respectively, while the PDE6 structure is a homology model. D) The interactions of the Pγ (orange) with PDE5/6 is displayed from three perspectives to highlight interactions with H- and M-loops and α12. (PDBID: 3JWR) E) The sequence alignment of all three sequences are presented with colored overbars matching the structural regions highlighted in panels A-C. In addition, the green overbar region is the region in which the PDE6 sequence was inserted into the PDE5 sequence.
The aims of this study are to apply molecular dynamics (MD) simulations to the PDE5, PDE5/6, and PDE6 catalytic domains in order to develop a structure-function understanding of the enzyme. We seek to rationalize why slight sequence differences give rise to the structural changes in the H- and M-loops, generate hypotheses on the requirement of an inhibitory subunit for PDE6 and understand the origins of the vastly different catalytic efficiencies between PDE5 and PDE6. In this study, we perform equilibrium MD simulations on several systems comprising combinations of apo, sildenafil bound, GMP bound, and $P_\gamma$ bound for PDE5, PDE5/6 and PDE6. We analyze these simulations using standard structural metrics as well as principal component analysis (PCA) and mutual information analysis. From the equilibrium simulations we develop a hypothesis that the $\alpha$-helix 12 ($\alpha_{12}$, residues 674-696 in PDE6) is allosterically connected to the catalytic site in PDE6 based upon differences in the correlated motions between PDE5 and PDE6. We believe this connection may relate to the different catalytic rates of PDE5 and PDE6 and we conclude the study by exploring this hypothesis through the use of ligand unbinding metadynamics simulations, which allow us to estimate the free energy barriers for ligand (un)binding.

**Methods**

Coordinates for PDE5 and PDE5/6 were taken from X-ray crystal structures to initiate simulations of apo PDE5 (PDBID: 2H40), sildenafil bound PDE5 (PDBID: 2H42)\(^{14}\), apo PDE5/6 (PDBID: 3JWQ) and $P_\gamma$ bound PDE5/6 (PDBID: 3JWR).\(^{10}\) Zn\(^{+2}\) and Mg\(^{+2}\) ions bound in the active site were retained in all simulations, while waters observed in the crystal structures were removed. To study PDE6 despite the lack of an existing crystal structures in the PDB, we generated an apo homology model for PDE6 using I-TASSER\(^{15}\) and then modeled in $P_\gamma$ and sildenafil through structural alignments with PDE5/6 structures, which include a $P_\gamma$ bound structure (PDBID:
3JWR). Simulations containing GMP were based upon the PDE5 GMP bound structure (PDBID: 1T9S), but due to mutations in that PDE5 structure we aligned the GMP bound structure with the sildenafil bound PDE5 structure (PDBID: 2H42). Sildenafil was deleted and GMP was copied into the sildenafil bound structure. A PDE6 bound to GMP model was created by an alignment of our I-TASSER model with PDBID: 1T9S and copying over the GMP coordinates. 250 ns equilibrium simulations were performed for both PDE5 and PDE6 bound to GMP. The I-TASSER generated PDE6 model was validated by analyzing the backbone dihedral angles using PROCHECK (Fig. 4.S1). Only three residues (SER143, GLN273 and LEU331) fell outside of allowed regions, and those violating residues were either in unstructured loops or near the terminus. The different systems and simulations times are listed in Table 1. All simulations were run using GROMACS 4.6.5 with the CHARMM27 force field in the NPT ensemble. System sizes were approximately 9 nm X 9 nm X 9 nm and consisted of approximately 20,000 TIP3P waters and NaCl at 150 mM concentration. Langevin dynamics were performed with a 2 fs timestep and a friction factor of 1 ps⁻¹. Temperature was maintained at 300 K by the Langevin algorithm and the system pressure was isotropically coupled to a 1 atm pressure bath with the Parrinello-Rahman barostat. Nonbonded Lennard-Jones interactions were unmodified out to 1.0 nm and then smoothly shifted to zero between 1.0 and 1.2 nm. The electrostatic interactions were computed by the PME method where the direct interactions were smoothly switched off between 0 and 1.2 nm. Force field parameters for sildenafil were generated using SwissParam. Prior to the production simulations, all systems underwent an equilibration phase during which the protein heavy atoms, Zn²⁺ and Mg²⁺ ions and Pγ (if present) were restrained with a 1000 kJ/mol*nm² harmonic positional restraint. The equilibration involved three steps: i) energy minimization for
up to 5,000,000 steps using the steepest descent algorithm, ii) NVT equilibration for 100 ps, and iii) NPT equilibration for 100 ps.

For each simulation, the following analyses were performed: root mean square deviation (RMSD), root mean square fluctuations (RMSF), calculation of H- and M-loop distances, principal component analysis (PCA), and mutual information analysis. For PDE5 and PDE6 apo systems, PCA was performed on a master trajectory combining the five 300 ns simulations, excluding the first 50 ns of each simulation to account for equilibration. These master trajectories contained 1.25 µs of data. Mutual information analysis was performed on the five individual 300 ns trajectories for each system and the results were then averaged.

Ligand stability was checked using a protein aligned RMSD measurement, which captures both rigid body movements and internal structural changes. Small molecules (sildenafil and GMP) displayed low (≤ ~3 Å) RMSDs for both PDE5 and PDE6 (Fig. 4.S2A-D), indicating stable interactions and limiting concerns about docking of the ligands into the PDE6 homology model. For the PDE6-Pγ simulation the RMSD of Pγ is considerably higher reaching a steady value of ~8 Å in the last 300 ns of the 900 ns simulation (Fig. 4.S2E). However, the orientation of Pγ with respect to PDE6, namely the C-terminal region of Pγ occluding entry to the active site of PDE6 is maintained in the simulation, as shown by an overlay of the initial and final orientation of Pγ (Fig. 4.S2F).

Mutual information was calculated using the MDEntropy package version 0.3. Mutual information \( I(X,Y) \), measures the extent the uncertainty in a given variable \( Y \) changes when the state of another variable \( X \) is known. The mutual information is defined as the difference between the Shannon entropies \( S \) of the marginal distributions of variables \( X \) and \( Y \), and the joint Shannon entropy.
\[ I(X,Y) = S(X) + S(Y) - S(X,Y) \]  

\[ I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

. The analysis is performed in internal coordinate space to avoid alignment issue artifacts. The dihedrals angles of the backbone (phi and psi) as well as sidechain dihedrals are used in the analysis and the mutual information between a pair of residues is calculated by summing over all dihedral pairs between the residues.\(^{25}\) Mutual information has been successfully used to find co-varying regions of proteins that may be involved in allosteric conformational change.\(^{25}\) Dihedral information was calculated every 10 ps for each 300 ns apo trajectory then the five trials were averaged. In the mutual information figures, the data was filtered to provide clarity; in the figure presenting mutual information of the apo simulations the self \(I\) along the diagonal as well as all low \(I\) pairs \((I < 0.2)\) were set to zero. Differences in mutual information between ligand bound and unbound systems was normalized by diving through by the average value along the diagonal, effectively setting both bound and unbound states to have \(I=1\) along the diagonal.

Metadynamics\(^{26}\) simulations of sildenafil and GMP unbinding from the binding pockets of PDE5 and PDE6 were carried out to evaluate the free energy profile of binding. Initial conformations for sildenafil unbinding metadynamics were selected by first projecting the equilibrium sildenafil bound simulations onto the first two principal components of the respective apo enzyme. Structures were selected from the center and extreme positions in the PC projection map and were used as starting configurations for the metadynamics simulation trials (Fig. 4.S3). For GMP metadynamics unbinding, PCA was performed on the backbone including the carbonyl oxygen of the final 100 ns for each simulations. The starting structures for the metadynamics were chosen one from the center and four extremes of the first two principal components, analogous to
the sildenafil procedure.

In metadynamics, Gaussian potentials are deposited along a collective variable (CV) in a history dependent manner to bias the system towards more rarely sampled conformations. For sildenafil unbinding we define the distance between the center of mass (COM) of the protein and the COM of sildenafil as our reaction coordinate. In each trial a single unbinding event was observed (without rebinding) and the simulation was stopped when sildenafil reached more than 30 Å away from its initial binding pose. The PMFs were shifted to have the ligand-bound state be the reference free energy value ($\Delta G=0$). For PDE5 complete unbinding took between 54 ns and 84 ns with an average of 67.4 ns, and for PDE6 unbinding took between 21 ns and 65 ns with an average of 49.4 ns. Single pass ligand unbinding metadynamics has been used previously and shown to successfully determine unbinding pathways and associated kinetics$^{27,28}$ and to score the stability of different docked drug poses.$^{29}$ Metadynamics was performed with a hill height of 0.01 kJ/mol, a Gaussian width of 0.03 nm with a Gaussian potential deposited every 500 steps. For GMP unbinding we found the strong interaction between GMP and the coordinating metal ions lead to large protein distortions during our unbinding events, and therefore we removed the Zn$^{+2}$ and Mg$^{+2}$ ions in these simulations. The CV for GMP metadynamics was the mean square displacement (MSD) of GMP heavy atoms from the starting structure after an alignment on the protein $\alpha$ carbons. A Gaussian width of 0.1 nm$^2$, a hill height of 0.05 kJ/mol and a deposition rate of 1/500 steps were used. The simulations were stopped when the MSD reached 45 nm$^2$. Traces of the collective variable versus time for both sildenafil and GMP unbinding trials are shown in Fig. 4.S4. All metadynamics simulations were performed with GROMACS 4.6.6 patched with Plumed 2.1.$^{30}$
Sildenafil and GMP contacts with the enzyme during the transition path were analyzed using MDTraj v1.9. A residue was considered in contact with the ligand if any heavy atom on that residue was within 3.5 Å of any heavy atom on the ligand. In order to exclude residues from the bound state, any residue with a heavy atom within 4 Å of a ligand heavy atom at the start of the simulation was not considered.

Table 1. Systems and simulation runs and lengths

<table>
<thead>
<tr>
<th>System</th>
<th>Type</th>
<th>Runs</th>
<th>Run length</th>
<th>Total Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDE5 apo</td>
<td>Equilibrium</td>
<td>5</td>
<td>300 ns</td>
<td>1.5 µs</td>
</tr>
<tr>
<td>PDE6 apo</td>
<td>Equilibrium</td>
<td>5</td>
<td>300 ns</td>
<td>1.5 µs</td>
</tr>
<tr>
<td>PDE5/6 apo</td>
<td>Equilibrium</td>
<td>1</td>
<td>900 ns</td>
<td>900 ns</td>
</tr>
<tr>
<td>PDE6-Pγ</td>
<td>Equilibrium</td>
<td>1</td>
<td>900 ns</td>
<td>900 ns</td>
</tr>
<tr>
<td>PDE5/6-Pγ</td>
<td>Equilibrium</td>
<td>1</td>
<td>100 ns</td>
<td>100 ns</td>
</tr>
<tr>
<td>PDE5-Sildenafil</td>
<td>Equilibrium</td>
<td>1</td>
<td>200 ns</td>
<td>200 ns</td>
</tr>
<tr>
<td>PDE6-Sildenafil</td>
<td>Equilibrium</td>
<td>1</td>
<td>200 ns</td>
<td>200 ns</td>
</tr>
<tr>
<td>PDE5-GMP</td>
<td>Equilibrium</td>
<td>1</td>
<td>250 ns</td>
<td>250 ns</td>
</tr>
<tr>
<td>PDE6-GMP</td>
<td>Equilibrium</td>
<td>1</td>
<td>250 ns</td>
<td>250 ns</td>
</tr>
<tr>
<td>PDE5-Sildenafil</td>
<td>Metadynamics</td>
<td>5</td>
<td>54-84 ns</td>
<td>337 ns</td>
</tr>
<tr>
<td>PDE6-Sildenafil</td>
<td>Metadynamics</td>
<td>5</td>
<td>21-65 ns</td>
<td>247 ns</td>
</tr>
<tr>
<td>PDE5-GMP</td>
<td>Metadynamics</td>
<td>5</td>
<td>27-58 ns</td>
<td>220 ns</td>
</tr>
<tr>
<td>PDE6-GMP</td>
<td>Metadynamics</td>
<td>5</td>
<td>10-40 ns</td>
<td>93 ns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.7 µs</td>
</tr>
</tbody>
</table>

Results

Equilibrium Simulations

We begin by comparing the apo PDE5, PDE6 and PDE5/6 simulations. Since the PDE6 system is based upon a homology model, we wanted to evaluate if the model displayed structural
stability in a range consistent with the experimentally determined PDE5 and PDE5/6 structures. The RMSD probability distributions are presented in Fig. 4.2 and the individual simulation traces in Fig. 4.5. It can be observed that PDE5/6 is highly stable with the structure staying within 2 Å of the initial structure, though this is based upon a single long trajectory, whereas PDE5 and PDE6 data came from five shorter (300 ns) simulations. The high degree of stability of PDE5/6 may explain why the catalytic domain of the chimera could be crystallized. PDE5 and PDE6 both display good stability and have a consistent range between of RMSD values between the two systems, with all structures of PDE6 remaining within 4 Å of the initial configuration. PDE5 is slightly less stable than PDE6 and in one simulation the RMSD deviated to beyond 6 Å. Based upon the RMSD distribution we can conclude the homology model of PDE6 has the same range of structural stability as the experimentally determined structures for PDE5 and PDE5/6 and we believe it can serve as an informative model for PDE6.

We next examined the root-mean-squared-fluctuations per residue (RMSF) during the equilibrium simulations to evaluate if the different systems had different regions of flexibility. Fig. 4.3A presents the RMSF comparison between apo PDE5, PDE6 and PDE5/6 and we observe
significant differences. The major regions of flexibility are in the H- and M-loop regions, though PDE6 has considerably lower fluctuations in these regions than PDE5. Interestingly, PDE6 displays a region of high flexibility peaking around residue 699 (shown as residue 215 in Fig. 4.3A), which is towards the base of α-helix 12 (α12). PDE5/6 has low fluctuations throughout the structure, including in the H- and M-loops, which is consistent with the low RMSD of this structure (Fig. 4.2). We also examined the effect of the inhibitor (sildenafil) on both PDE5 and PDE6 and the effect of the inhibitory subunit (Pγ) on PDE6 and PDE5/6, the bound location of sildenafil and Pγ are shown in Fig. 4.3E and Fig. 4.3F, respectively. Sildenafil has a significant effect on PDE5 by reducing the flexibility in the M-loop (Fig. 4.3B). However, PDE6 does not display much variation in the RMSFs between apo and inhibitor bound simulations (Fig. 4.3C). Sildenafil is known to be a 10-fold more potent inhibitor of PDE5 than PDE6 (IC_{50} of 2.9-6 nM and 74 nM, respectively), so if PDE5 relies on enzyme flexibility for catalysis, then the more dramatic effect seen in PDE5 may correlate with the enhanced inhibitory efficacy against PDE5. The small effect Pγ has on PDE6 fluctuations is somewhat unexpected, though as a control the effect of Pγ on PDE5/6 was also examined. Pγ does function to inhibit PDE5/6 but similarly to PDE6 the effect of Pγ on the fluctuations of PDE5/6 is relatively minor (Fig. 4.3D)

**Correlated Motions**

While the atomic fluctuations show little variation between the apo and the inhibitor bound simulations of PDE6 (Fig. 4.3C), this does not preclude the possibility that the direction of motion could be different between these systems. We have performed principal component analysis (PCA) on the equilibrium simulations of PDE6 in apo and inhibitor bound states, as well as for PDE5 apo. Both PDE5 and PDE6 apo simulations are well characterized by a few low frequency modes,
as the first 10 modes account for 78% of the total variance for PDE6 and 84% in the PDE5 system. We examine the projection of the apo and inhibitor bound simulations onto the space of the PDE6 apo first two modes in Fig. 4.4A. We observe significant restriction of the motion of PDE6 in the
Figure 4.3. PDE5, PDE5/6 and PDE6 RMSF comparisons. A) RMSF comparison between apo PDE5, PDE6 and PDE5/6, residue numbering is 1-based since the constructs start at different residue numbers. B) RMSF comparison of PDE5 apo and inhibitor bound (sildenafil). C) RMSF comparison of PDE6 apo and inhibitor bound (sildenafil or Pγ). D) RMSF comparison between PDE5/6 apo and Pγ bound. The underbars in A-D are the locations of the H-loop (blue), α12 (green) and M-loop (red). The binding location of the inhibitor sildenafil is shown in (E) in green, and the inhibitory subunit (Pγ) is shown in (F) in orange. The H- and M-loop PC1-PC2 subspace when PDE6 is bound by either sildenafil or Pγ. This constriction of sampling in the PC1-PC2 subspace can be interpreted as the inhibitor bound systems having different
collective dynamics than the unbound native enzyme.

We also examined the dynamics of the H- and M-loops as these motifs form a surface which is adjacent to enzyme active site. We measured the distance between a Ser-Leu pair at the tips of the H- and M-loops (Ser617-Leu756 in PDE6 and Ser667-Leu796 in PDE5) to examine the loop separation. These residue pairs were chosen based upon the close proximity of the corresponding residues in the PDE5/6 crystal structure (PDBID: 3JWQ). The distance distribution (Fig. 4.4B) for PDE6 shows a sharply peaked distribution with a peak at less than 1 nm, indicating the H-M loops are in contact and either are relatively static or moving in a positively correlated fashion. In contrast, PDE5 shows a broad distribution with peaks in the range of 2-3 nm. The observed flexible separation in the H-M loops for PDE5 is consistent with a previous modeling study, though the simulation length was 100 times shorter in the previous work. This difference in the H-M distance distributions is informed by examination of the first PC for each system. In PDE6 (Fig. 4.4C), the first PC accounts for 27% of the variance and shows the most significant motion is in the H-loop, M-loop, α12 and the loop at the base of α12. The motion can be described as a clam shell motion where the H- and M-loops are moving down together and α12 and is moving up to close (or open) access to the binding site. In PDE5 (Fig. 4.4D) the first PC is very robust and accounts for 40% of the variance. The motion in the PDE5 PC1 in concentrated in the H- and M-loops, but shows the loops move in opposite directions causing separation between the loops, consistent with the separation distance distribution in Fig. 4.4B.

To further examine the correlations within PDE5 and PDE6, mutual information analysis was performed. Mutual information shows how knowledge about the state of one residue (X) reduces the uncertainty in the state of another residue (Y). The reduction in uncertainty indicates the conformational distribution of the two residues are correlated and therefore is an effective
approach to identifying allosteric sites within proteins. In apo PDE5 (Fig. 4.5A) significant mutual information is observed for the intersection of the H- and M-loop as well as the M-loop with itself. While the regions of significant mutual information are well defined in PDE5, there is wide-spread correlations within PDE6 (Fig. 4.5B). The strong correlation between the H- and M-loops present in PDE5 is greatly reduced in PDE6. The interpretation of this difference may be
that the H- and M-loops are cooperatively folding in PDE5, while they are rigid in PDE6. In PDE6 α12 shows high mutual information across many different regions of the protein. The rigidity of the structured H-loop allows global bending and hinging motions consistent with the motion of PDE6 PC1, which we suspect may allow global communication through the protein. This indicates that the conformation of α12 is highly connected to the conformation of the rest of the protein and

**Figure 4.4. Collective motions in PDE6 and PDE5.** A) The simulations of PDE5 apo, sildenafil bound and Pγ bound are projected on the PDE6 apo principal components 1 (PC1) and 2 (PC2) subspace. B) The H-M loop distance probability distribution for apo PDE5 and PDE6. The distance is defined by the center of mass distance between residues SER617 and LEU756 in PDE6 and the corresponding residues (SER667 and LEU796) in PDE5. C) The first PC of PDE6 apo accounts for 27% of the total variance. D) The first PC of PDE5 apo accounts for 40% of the total variance.
therefore may serve as a critical regulatory element of the enzyme. There is significant mutual information between PDE5 $\alpha_{12}$ and other regions of the protein, but this communication is more localized. The increased flexibility and disorder of the H-loop allows the sidechains of the H-loop to interact with $\alpha_{12}$ without transferring information through the rigid core of the protein. The correlations of $\alpha_{12}$ with spatially distant regions of both PDE5 and PDE6 is not intuitively obvious based upon structural considerations and it implies an allosteric regulation mechanism is active.

The effect of ligand (sildenafil and P$\gamma$) binding on the enzyme mutual information was also analyzed. The difference between the apo and ligand bound systems are presented in Fig. 4.6. The effect of sildenafil on PDE5 shows both increases and decreases in mutual information throughout the protein (Fig. 4.6A). The most significant change in mutual information occurs between the H- and M-loops where the mutual information is reduced upon ligand binding. The effect of sildenafil on the mutual information of PDE6 was less pronounced, showing slight reductions through the protein (Fig. 4.6B). The effect of P$\gamma$ binding on PDE6 was widespread, with more significant reductions in mutual information than the reductions in mutual information caused by sildenafil binding and displaying a consistent reduction in the $\alpha_{12}$ region (Fig. 4.6C). Overall the results

![Figure 4.5. Mutual information of apo systems.](image)

The mutual information for PDE5 (A) and PDE6 (B) is presented. In both figures the H-loop, M-loop and $\alpha_{12}$ regions are denoted by horizontal and vertical boxes.
point toward a consistent picture of sildenafil being more effective against PDE5 than PDE6 and that $P_\gamma$ is an effective regulator of PDE6 dynamics.

**Ligand (un)Binding**

Based upon the analysis of the principal components and mutual information of PDE5 and PDE6, we speculated that the collective motions of each enzyme could have an effect on ligand access to the active site. The correlated motions in PDE6 of the H-loop, M-loop and $\alpha_{12}$ region appear to generate an opening to the binding site, whereas the motion of the H- and M-loops in PDE5 are not concerted with regions below the binding pocket. We hypothesized that the high catalytic efficiency of PDE6 would be reflected by a relatively flat energy surface for ligand binding and we have utilized metadynamics\textsuperscript{26} to estimate the free energy profiles. By performing unbinding simulations of sildenafil from PDE5 and PDE6 we can evaluate the energetics of ligand unbinding by detecting barriers in the binding pathway that could affect kinetic aspects of catalysis.

Five unbinding trials were performed for both PDE5 and PDE6 and the averaged free energy profiles of sildenafil binding are presented in Fig. 4.7A and the individual trial profiles are presented in Fig. 4.S6. The profile is generated from averaging five metadynamics trials, in which each trial was initialized from a different starting configuration, extracted from the equilibrium

![Figure 4.6](image)

**Figure 4.6. Change in mutual information upon ligand binding.** A) The difference in mutual information between PDE5 bound to sildenafil and PDE5 apo. B) The difference in mutual information between PDE6 bound to sildenafil and PDE6 apo. C) The difference in mutual information between PDE6 bound to $P_\gamma$ and PDE6 apo.
sildenafil bound simulations. The free energy profiles show that there is a more favorable $\Delta G$ of binding for PDE5 (-13 kcal/mol) than PDE6 (-9 kcal/mol), which is not only qualitatively consistent with the lower IC$_{50}$ of sildenafil against PDE5 than PDE6, the calculations are also quantitively consistent with the experimental $\Delta G_{binding}$ based upon IC$_{50}$ values (PDE5: -11.9 kcal/mol; PDE6: -9.8 kcal/mol).$^{32,34}$ While the free energy change of binding is smaller for PDE6, the barriers along the binding pathway are also smaller, which would lead to faster binding kinetics ($k_{on}$). On the other hand, the PDE5 profile displays a large barrier (~4 kcal/mol) at around 1 nm. Also, the unbinding barrier is much smaller for PDE6 (~4.5 kcal/mol) than PDE5 (~10 kcal/mol), which may reflect shorter residence times (faster $k_{off}$), which would also indicate less effectiveness of sildenafil against PDE6.

The pathways of sildenafil unbinding are visualized in Fig. 4.7B-C, by tracing the ligand coordinates during the metadynamics runs. In PDE5, sildenafil interacts with several parts of the protein during unbinding. In some trials sildenafil interacts with $\alpha$-helix ($\alpha_{15}$, residues 809-836 in PDE5), on the right side of the binding pocket, briefly and then moves closer to the $\alpha_{12}$ region where it exits. Other PDE5 trials move toward the H-loop and exit on the left side of the binding pocket. For PDE6 all trials move toward $\alpha_{15}$ and wrap around this helix before exiting the binding pocket. Sildenafil does not make contact with the H-loop in any of our PDE6 trials and only contacts the $\alpha_{12}$ region in one trial. The additional rigidity and helical stability of the H-loop, M-loop and $\alpha_{12}$ region seen in both the metadynamics simulations as well as the apo equilibrium simulations, appears to create a larger binding pocket with lower barriers to unbinding.
The contacts sildenafil makes with PDE5 and PDE6 as it leaves the binding pocket were quantified using the MDTraj compute neighbors utility. This data is presented in Fig. 4.8 and is based upon the metadynamics unbinding trajectories. The number of frames in which sildenafil is in contact with a PDE5 or PDE6 residue is tabulated over all five trajectories and converted into a contact percentage. In PDE5 simulations, sildenafil makes contact with the H-loop and α12 and also to a lesser extent is in contact with the M-loop and α15. In PDE6 simulations, sildenafil makes significant contacts with α12 and α15, while it does not significantly contact the H- or M-loops. The contact analysis is consistent with a model in which PDE6 provides a focused binding pathway, while PDE5 appears to allow a broader range of pathway directions.

**Figure 4.7. Energetics and pathways of sildenafil unbinding from metadynamics.** A) The free energy as a function the distance between the center of mass of the drug and the center of mass of the protein. The average of 5 PMFs for each system are represented by the lines and the shaded regions represent the standard error over the five runs. Externalization pathways are shown from two vantage points for PDE5 (B) and PDE6 (C). Each pathway is traced by a different color and α12 is colored yellow in all structures to provide a common reference point.
Comparing the free energy profiles and unbinding pathways, we are able to correlate some of the energetic features with aspects of the unbinding pathways. In PDE5, sildenafil has to cross a large barrier about 5 Å from its starting location, which appears to arise as the ligand becomes constricted between the flexible H-loop, M-loop and α15. Sildenafil then encounters a free energy well as it gets past α15, but then faces another large free energy barrier as it crosses over the α12

Figure 4.8. Sildenafil contacts with PDE during unbinding pathways. Contacting residues are indicated on the PDE5 (A) and PDE6 (B) structures. Color scale goes from blue (no contacts) to red (highest percentage of contacts). The contacting percentage, averaged over 5 metadynamics trajectories, is shown for PDE5 (C) and PDE6 (D). In C-D the residues which are contacting sildenafil in the bound state are given a contact percentage of zero, to highlight the residues which make contact in the transition out of the binding pocket. The underbars in C-D are the locations of the H-loop (blue), α12 (green) and M-loop (red).

Comparing the free energy profiles and unbinding pathways, we are able to correlate some of the energetic features with aspects of the unbinding pathways. In PDE5, sildenafil has to cross a large barrier about 5 Å from its starting location, which appears to arise as the ligand becomes constricted between the flexible H-loop, M-loop and α15. Sildenafil then encounters a free energy well as it gets past α15, but then faces another large free energy barrier as it crosses over the α12
region while the H-loop sterically hinders its exit path. In PDE6 the H-loop is further away from sildenafil than in PDE5 making it easier for the ligand to leave its initial binding pose. The unbinding paths in PDE6 are characterized by a pathway contacting α15, α12, or in between these two helices. Sildenafil also stays in contact with the enzyme surface longer in PDE6 as it is still in contact at a COM distance of 29 Å, whereas in PDE5 it breaks contact at about 22 Å. During binding this feature may correspond to a larger area of first encounter that can “catch” sildenafil, then the correlated motions (PC1) may serve to shuttle the ligand into the binding pocket and/or provide accessibility to the binding pocket.

While understanding the ligand binding pathway of sildenafil to PDE5 and PDE6 may provide some insights into to general ligand binding and differences in catalytic rates, to probe the biological function more directly we performed metadynamics unbinding simulations of GMP from PDE5 and PDE6. We observe that PDE6 has a flat landscape with no significant barriers (within the uncertainty of calculation) along the pathway in going from unbound to bound (Fig. 4.S7-S8). This observation is consistent with a diffusion-limited catalytic rate which PDE6 is known to have. In contrast the PDE5 pathway does have a deep well at an intermediate state around 5.5 nm² in MSD space (~ 23 Å in RMSD).

Discussion

We believe that this work has provided evidence of a region of structural and functional significance in PDE6, the α12 helix. The RMSFs of all of the PDE6 systems highlight that in addition to the H- and M-loops showing high flexibility, the α12 region also shows high flexibility. This flexibility is absent from the corresponding residues in PDE5. The motions of the first
principal component support this, highlighting that the α12 region and the H- and M-loops cohesive ‘breathing’ movements are representative of PDE6’s overall motions. The importance of α12 and the α12 basal loop to PDE6 function has also been indicated experimentally. In a chimera in which residues 676-741 in PDE6 were substituted by the corresponding residues in PDE5 (residue 723-741) the resulting chimera was inactive and had poor solubility.37

Our hypothesis is that this concerted motion helps provide access and possibly recruit cGMP to the catalytic pocket to allow hydrolysis, which results in the high activity observed in PDE6. The lack of fluctuations and correlations in the corresponding residues in PDE5 and PDE5/6 along with the lack of the breathing motion in the first principal component indicates a difference in collective motions between PDE5 and PDE6. The correlated motions in PDE6 which couples the H-loop, M-loop and α12 regions may influence the catalytic rates in several ways, e.g. i) providing a binding surface for initial ligand interactions, ii) providing an opening motion that facilitates access to the binding site and iii) providing motions which could shuttle the ligand towards the binding site. It should be noted that the relationship between enzyme catalysis and protein conformational dynamics is a debated one. Theoretical studies have argued that the enzyme dynamics do not couple to the chemical reaction rate in enzymes.38,39 However, the overall catalytic rate can be effected by protein dynamics, as it has been shown the conformational changes associated with product release can be rate-limiting in some systems. Other studies have supported a link between dynamics and catalysis, in particular for dihydrofolate reductase.40

Determination of a PDE6 structure by X-ray crystallography has yet to be achieved, though a recent study generated a full-length rod PDE6 dimer structure using cryo-EM density and crosslinking restraints.11 In the present study we have generated a homology model to study the cone PDE6 catalytic domain, which is an approach that has been used in previous modeling studies
to examine small molecule interactions with PDE6.\textsuperscript{33,41,42} In all studies, including the current one, the PDE6 catalytic domain was studied as a monomer despite the full-length protein being known to exist and function as dimer.\textsuperscript{3} In the first such study (Cahill et al.)\textsuperscript{42}, the PDE6 modeling consisted of small molecule docking and protein dynamics and stability were not evaluated. In a subsequent study (Huang et al.)\textsuperscript{33} the PDE6 model was unstable in the absence of bound ligand, but through MM-PB(GB)SA calculations they were able to show different affinities for sildenafil and tadalafil (another PDE5 inhibitor) towards PDE5 and PDE6. They employed three different variants of the MM-PB(GB)SA calculations and the results were not always consistent across the methods, for example the IC\textsubscript{50} for tadalafil is two orders of magnitude higher than sildenafil against PDE6, but only two of the three calculations predicted a more favorable $\Delta G_{\text{binding}}$ for sildenafil. Simulations timescales in that study ranged from 30 – 150 ns. The most recent study (Kayık et al.)\textsuperscript{41} did produce a stable apo PDE6 model ($< 5 \text{ Å RMSD}$) during a 50 ns simulation, and they also observed increased fluctuations in the $\alpha_{12}$ region consistent with our findings. That study was focused on identifying compounds that could have selectivity for PDE5 over PDE6 or PDE11, based upon MM-PB(GB)SA. Our study is different to previous studies in that we have put more emphasis on understanding the protein dynamics and how PDE5 and PDE6 display different motions. Furthermore, we employ a different approach to understanding the ligand-protein interactions by using metadynamics which allows us to observe pathways and barriers, rather than just free energy differences.

This investigation has only examined monomers of PDE5 and PDE6 and this is a limitation to the current work. Both PDE5 and PDE6 exists in vivo as dimers and there are allosteric interactions between the subunits.\textsuperscript{43,44} While the computational cost will rise significantly in moving from monomer to dimer simulations, especially if one is to consider a full-length structure containing
the GAF domains, this is an important endeavor that is becoming more feasible with rising computational power.

Conclusions

When PDE6 is active, it is necessary for the enzyme to hydrolyze cGMP rapidly so that the signal can be passed on quickly and vision can occur nearly instantly. Our hypothesis is that $\alpha_{12}$ works in concert with the H- and M-loops to cause active transport of cGMP to the catalytic pocket for hydrolysis. When there is no light, there should be no signal propagation and no vision, so PDE6 should be completely off. The $\gamma$-subunit not only blocks access to the catalytic pocket, it also disrupts any allosteric interactions between the H- and M-loops and $\alpha_{12}$ by physically blocking direction connections between the regions (see Fig. 4.1D). The results of the equilibrium simulations have been further explored through additional analyses on the mutual information shared though the enzymes and through enhanced sampling (metadynamics) simulations which allow for the ligand binding mechanism and energetics to be estimated. We hope this study may inform future experimental efforts to study PDE6 possibly through creation of a new chimera that accounts for importance of $\alpha_{12}$ for PDE6 and its functionality. Additionally, $\alpha_{12}$ may present a possible allosteric therapeutic target for retinitis pigmentosa or other diseases in which PDE6 is implicated.

References

(1) Cote, R. H. Structure, Function, and Regulation of Photoreceptor Phosphodiesterase (PDE6). In Handbook of Cell Signaling; Bradshaw, R. A., Dennis, E., Eds.; Academic


Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in


https://doi.org/10.1063/1.4947037.

https://doi.org/10.1126/science.1198542.


https://doi.org/10.1074/jbc.M112.389189.

(43) Blount, M. A.; Zoraghi, R.; Bessay, E. P.; Beasley, A.; Francis, S. H.; Corbin, J. D.

https://doi.org/10.1124/jpet.107.126540.Beavo.


https://doi.org/10.1074/jbc.M803948200.
Figure 4.S1. Ramachandran plot for PDE6 homology model.
**Figure 4.S2. Ligand Stability.** The protein aligned RMSD for sildenafil in PDE6 (A) and PDE5 (B); for GMP in PDE6 (C) and PDE5 (D); and for Pγ in PDE6. The blue lines are the raw data and the black lines are 1 ns running averages. In (F) the orientation of Pγ, relative to PDE6, is shown at the beginning (red) and end (blue, t=900 ns) of the PDE6-Pγ bound simulation.

**Figure 4.S3. Starting structures for metadynamics trials.** The equilibrium simulations of Sildenafil bound to PDE5 (A) and PDE6 (B) were projected onto the principal components of apo simulations of PDE5 and PDE6, respectively. Starting structures for metadynamics were selected by hand (shown by red circles), by finding structures which corresponded to central and extreme positions in the PC projection plots.
Figure 4.4. Metadynamics collective variable (CV) trajectories. The CV traces for PDE5-sildenafil (A), PDE6-sildenafil (B), PDE5-GMP (C) and PDE6-GMP (D), are shown for all trials.

Figure 4.5. RMSD in equilibrium simulations. RMSD traces are shown for PDE5 (A), PDE6 (B) and PDE5/6 (C). Each trend line represents an independent simulation.
Figure 4.56. PMF of sildenafil unbinding trials. Sildenafil unbinding PMFs from metadynamics are shown for PDE5 (A) and PDE6 (B).

Figure 4.57. PMF of GMP unbinding by metadynamics. The free energy as a function the ligand mean-squared-displacement (MSD) from the initial bound conformation. The average of 5 PMFs for each system are represented by the lines and the shaded regions represent the standard error over the five runs.
Figure 4.S8. PMF of GMP unbinding trials. GMP unbinding PMFs from metadynamics are shown for PDE5 (A) and PDE6 (B).
Chapter 5. Part I: Folding a Viral Peptide in Different Membrane Environments: Pathway and Sampling Analyses

Reproduced with permission from:

The following changes have been made:

- The manuscript has been reformatted to meet University Standards
- Figures numbers have the chapter numbers prepended

For this work SN set up the systems, ran umbrella sampling simulations, analyzed segment helicity, analyzed depth, and wrote the manuscript.
JGP ran the tica analysis, tram analysis, sampling analysis, and wrote the manuscript
This Agreement between Jason Pattis ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number: 4663210078486
License date: Sep 06, 2019
Licensed Content Publisher: Springer Nature
Licensed Content Publication: Journal of Biological Physics
Licensed Content Title: Folding a viral peptide in different membrane environments: pathway and sampling analyses
Licensed Content Author: Shivangi Nangia, Jason G. Pattis, Eric R. May
Licensed Content Date: Jan 1, 2018
Licensed Content Volume: 44
Licensed Content Issue: 2
Type of Use: Thesis/Dissertation
Requestor type: academic/university or research institute
Format: print and electronic
Portion: full article/chapter
Will you be translating? no
Circulation/distribution: 1 - 29
Author of this Springer Nature content: yes
Title: Molecular Simulation study of protein conformation change, binding mechanisms, and allosteric communication
Institution name: University of Connecticut
Expected presentation date: Oct 2019
Order reference number: 555
Requestor Location: Jason Pattis
120 Jordan Road
ROCKAWAY, NJ 07866
United States
Attn: Jason Pattis

Total: 0.00 USD

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the Licence) between you and Springer Nature Customer Service Centre GmbH (the Licensor). By clicking 'accept' and completing the transaction for the material (Licensed Material), you also confirm your acceptance of these terms and conditions.

https://s100.copyright.com/AppDispatchServlet
Flocking a Viral Peptide in Different Membrane Environments: 
Pathway and Sampling Analyses

Shivangi Nangia, Jason G. Pattis, Eric R. May

Abstract
Flock House virus (FHV) is a well-characterized model system to study infection mechanisms in non-enveloped viruses. A key stage of the infection cycle is the disruption of the endosomal membrane by a component of the FHV capsid, the membrane active γ peptide. In this study, we perform all-atom molecular dynamics simulations of the 21 N-terminal residues of the γ peptide interacting with membranes of differing compositions. We carry out umbrella sampling calculations to study the folding of the peptide to a helical state in homogenous and heterogeneous membranes consisting of neutral and anionic lipids. From the trajectory data, we evaluate folding energetics and dissect the mechanism of folding in the different membrane environments. We conclude the study by analyzing the extent of configurational sampling by performing time-lagged independent component analysis.

1. Introduction
Understanding the thermodynamics of peptide association and folding in a membrane environment is critical to deciphering the underlying mechanisms of membrane disruption by membrane active peptides. The transition from an unstructured solution state to an α-helical membrane bound state is a common trait of small amphipathic membrane proteins that have been researched extensively both experimentally and computationally. The thermodynamic driving forces for protein-membrane interactions and stabilization of the folded state are a delicate balance
between enthalpic and entropic factors. Consequently, analyzing the peptide folding pathway and energetics in lipid bilayers can provide detailed insights into the biological activity of the peptide. A clear picture of the mode of membrane disruption employed by membrane active peptides has, in general, remained elusive.\textsuperscript{30,31}

Over the last two decades computational studies have been instrumental in providing insights into lipid and protein dynamics by utilizing both equilibrium and biased sampling methodologies to study the energetics, thermodynamics and structural dynamics of amphipathic membrane active proteins.\textsuperscript{18,20,23–27,32–36} A major area of emphasis in understanding the mechanism of action of membrane active peptides is characterizing the initial stage of membrane association and peptide folding. Previous studies in this area include that by Brooks and co-workers who examined the folding dynamics of the designed WALP peptides in an implicit membrane model using Temperature Replica Exchange Molecular Dynamics (T-REMD).\textsuperscript{25} Their studies showed that all variants of the WALP peptides penetrated the membrane with the N-terminal regions initiating the insertion and ultimately transitioning to an $\alpha$-helix trans-membrane configuration consistent with experimental observations. More recently, molecular dynamics (MD) simulations pertaining to the folding and penetration of a single transmembrane WALP peptide were carried out in both all-atom and coarse-grained models.\textsuperscript{24} The folding free-energy was determined in the coarse-grained representation as a function of helicity of the peptide using Hamiltonian REMD. Unbiased MD simulations have also been performed to study the folding dynamics of another widely studied membrane active peptide, the antimicrobial peptide melittin. These simulations have revealed that the peptide robustly associates with the membrane in a disordered state and attains helicity parallel to the surface of the membrane causing deformation of the bilayer as it folds.\textsuperscript{18} Long time scale
(17 μs) unbiased MD simulations have shown that there is a narrow distribution of folded melittin conformers that partition into the membrane interface.\textsuperscript{23}

In our previous work,\textsuperscript{37} we investigated the thermodynamic aspects of binding and the structural dynamics of the FHV 21 N-terminal residues of the γ peptide (known as γ\textsubscript{1}) using a multi-scale approach. We examined the binding and folding characteristics of γ\textsubscript{1} on pure phosphatidylcholine (PC), pure phosphatidylglycerol (PG) and a 50:50 mixed PC:PG membrane. Our findings from 1 μs equilibrium all-atom simulations were in agreement with experimental measurements of the configuration of γ\textsubscript{1} on a PC bilayer, where we observed ~70% helicity of γ\textsubscript{1}.\textsuperscript{38} On PG membranes, we observed low α-helical content ranging from 0-23%. The strong electrostatic interactions between cationic γ\textsubscript{1} and negatively charged PG may result in a higher entropy, less ordered bound state, which is consistent with ITC measurements.\textsuperscript{38} The folding propensity of γ\textsubscript{1} on the 50:50 PC:PG bilayer could not be inferred from the behavior on the homogeneous membranes. γ\textsubscript{1} displayed low helical content (16%) on the mixed bilayer, leading us to conclude the correlation between the amount of charge present in the membrane and the folding propensity of the peptide is not linearly related. We also performed simulations starting from the folded state on PC, PG and the mixed bilayer system, to probe the stability of the γ\textsubscript{1} helical conformation. We found that γ\textsubscript{1} does not unfold and remains embedded in the membranes throughout 1 μs simulations. From these observations, we proposed that an energy barrier separates the folded state from the low helicity state, and the barrier height has a dependence on the amount of charge in the membrane.\textsuperscript{37} Our equilibrium simulations were likely trapped in a metastable state, due to the rough energy landscape, which was insurmountable by conventional simulation approaches. To explore our
proposition further we appeal to enhanced sampling methods to overcome the limited sampling in equilibrium MD to produce barrier crossing events.

Different enhanced sampling methods offer varied advantages and disadvantages and one should make a well-informed selection of a method most applicable to addressing the scientific scenario being explored. A popular choice of method for studying protein-membrane systems is T-REMD. T-REMD works on the principle that molecules can sample through a rugged energy surface by making repeated swaps among its replicas that are simulated simultaneously at different temperatures. Although replica-exchange is a widely used method there are some disadvantages to this approach. The number of replicas required for efficient sampling scale as $f^{1/2}$ (where $f$ is the degrees of freedom), which for large biological systems with explicit solvent can lead to very high computational costs. Moreover, T-REMD is not an effective method to overcome entropic barriers, which are present in folding transitions. Path sampling techniques such as milestoning, forward flux sampling and transition path sampling also offer a non-biased simulation approach which can be employed to study activated processes by exploiting transition path theory and calculating the key transitions in the trajectory space rather than focusing on the stable states. The basis of these methods is to sample the fast occurring infrequent rare events involving a transition. Other enhanced sampling methods involve application of a bias potentials to accelerate the sampling in a desired region of configurational space are computational flooding, Metadynamics (MetaD) and Umbrella Sampling (US). These methods require the user to select an order parameter (collective variable, CV) along which a biasing potential can be applied to surmount free-energy barriers in the landscape. Both computational flooding and MetaD methods
rely on biasing potentials being added “on the fly” to the energy landscape of the system with the objective to sample all energy minima, but avoid excessive and re-sampling of local minima.

US is a mature and heavily utilized method in different biophysics studies including protein folding,\textsuperscript{49,50} peptide-peptide interactions,\textsuperscript{51,52} protein-DNA interactions,\textsuperscript{53} binding energies and interactions with lipid membranes,\textsuperscript{20,37,54,55} and conformational sampling of small molecules,\textsuperscript{56,57} among others. US relies on a stratification strategy, intermediate states along an order parameter are simulated with a restraint potential that keeps the system localized to a chosen point along the order parameter. A series of restrained simulations spanning the entire range of interest along the order parameter are simulated, and provided there are overlapping distributions between the umbrella windows, the probabilities can be unbiased and the potential of mean force (PMF) can be determined. Convergence in US is non-trivial to achieve or to evaluate and there are also choices regarding the restraint spacing and restraint force constant, though these are relatively easy to evaluate and there is a large literature to inform these choices. One important requirement of US is an initial path needs to be defined. In the limit of infinite sampling, the initial pathway would be irrelevant. Although, in practice it can have a significant effect, especially if there are slow degrees of freedom or other energy minima states separated by large energy barriers in orthogonal degrees of freedom to the US coordinate. Factors that make US a more robust and advantageous technique are that additional sampling can be carried out where sampling is sparse or in windows which display slow transitions and it is an appropriate method to maximally utilize parallel computing.

Many non-enveloped viruses\textsuperscript{58,59} contain a membrane active component of their capsid that in some systems is an amphipathic peptide, which is disconnected from the capsid. We have been
investigating the membrane lytic peptide of the non-enveloped Flock House Virus (FHV), which displays characteristics similar to antimicrobial peptides.\textsuperscript{37,60,61} In our previous study, we employed microsecond equilibrium simulations to examine the folding characteristics of $\gamma_1$, on membranes of different compositions.\textsuperscript{37} For the present study we aim to calculate the free energy profile of $\gamma_1$ folding in the presence of a membrane. We chose to employ US and have identified helicity of $\gamma_1$ as the order parameter for these calculations. To study the folding process one needs to consider two main aspects, the starting conformational state of the peptide and also its orientational features, i.e. the depth and angle of the peptide with respect to the bilayer. Our approach to addressing these initiation concerns was to initiate our simulations from the last snapshot of our previous work containing the bound helical state of $\gamma_1$ on different membrane compositions.\textsuperscript{37} The bound conformations are derived from 1 $\mu$s equilibrium simulations which have sampled the insertion depth of $\gamma_1$ at a depth consistent with experimental measurements, based upon Trp fluorescence.\textsuperscript{38} Initial unfolding pathways were constructed by applying steered molecular dynamics (SMD) to generate $\gamma_1$ conformations of varying helicity. In addition to evaluating energetic and structural aspects of the folding pathways we performed additional analyses to evaluate the convergence and uncertainty in the US data. We have utilized time-lagged independent components analysis (TICA) to identify the slowest decorrelating degrees of freedom. By performing projections of the US data into TICA subspaces, we can evaluate the connectivity of the US data and observe if regions in the phase space are undersampled leading to poor estimates of the free energy surfaces (FES).
2. Methods

2.1. System Setup

The \( \gamma_1 \) peptide is a 21 residue peptide with sequence ASMWERVKSIIKSSLAAASNI. Initial configurations used to generate unfolding pathways were obtained from the last snapshot of previous 1 \( \mu \)s equilibrium simulations.\(^{37}\) The three systems are helical \( \gamma_1 \) bound to PC, PG and 50:50 mix of PC:PG bilayers (Table 1). The bilayers consisted of 270 lipids for the PC bilayer, 288 lipids for the PG bilayer and a total of 200 lipid molecules for the mixed bilayer. The peptide carries a net +2 charge and both termini are charged, which was done to be consistent with deletion construct experiments of FHV virus like particles (VLP) which contained \( \gamma_1 \).\(^{62,63}\) The simulations were performed with the GROMACS package (version 5.0.1)\(^{64}\) using the CHARMM36 force field.\(^{65,66}\) The protein and membrane coordinates were extracted from the equilibrium simulations, but the systems were resolvated and reionized to reduce the system sizes. All systems were solvated with the TIP3P water model and were neutralized with 0.15 M NaCl. The systems were minimized using the steepest descent algorithm followed by an NVT equilibration for 5 ns, followed by 10 ns in the NPT ensemble. Pressure coupling was maintained at 1.0 bar using semi-isotropic coupling using the Parrinello-Rahman barostat. The temperature was maintained at 303.15 K using the velocity (v)-rescale coupling method. The temperature and pressure coupling constants were 1.0 ps and the compressibility value was \( 4.5 \times 10^{-5} \) bar\(^{-1} \). The long-range electrostatics were calculated by Particle Mesh Ewald (PME)\(^{67}\) method and non-bonded interactions were cutoff at 1.2 nm. The bonds lengths were constrained with the LINCS constraint algorithm and periodic boundary conditions were applied in X, Y and Z-directions, a time step of 2 fs was used for the equations of motion integration. All system configurations and snapshots were visualized using VMD software.\(^{68}\)
2.2. Umbrella Sampling

The helical content of $\gamma_1$ was chosen as the collective variable (CV) to perform US calculations. We have performed all-atom US simulations using GROMACS 5.0 with the PLUMED 2.2 plugin\textsuperscript{69,70} to quantify the free energy associated with the transition of $\gamma_1$ from a helical (H) to a random coil (C) conformation on different membrane compositions. The CV ALPHARMSD\textsuperscript{71} from the PLUMED 2.2 plugin was applied to generate intermediate configurations from H to C. This CV reports a sum, $S_\alpha$, computed from the backbone Root Mean Square Deviation (RMSD) of six residue segments with respect to an ideal $\alpha$-helix. The $S_\alpha$ value is calculated based on the following switching function

$$S_\alpha = \sum_{i} \frac{1 - \left(\frac{\text{RMSD}_i}{0.1}\right)^8}{1 - \left(\frac{\text{RMSD}_i}{0.1}\right)^{12}}$$

where RMSD is in units of nm and the sum runs over all possible consecutive six residue segments, yielding a theoretical maximum of $S_\alpha = 16$ for the 21 residue $\gamma_1$ peptide. Conversions between $S_\alpha$ and percent helicity are calculated by simply dividing the $S_\alpha$ value by the theoretical maximum $S_\alpha$ for the peptide (16) or segment (5).

In each of the systems the peptide was unfolded using steered molecular dynamics (SMD) with a restraint of 500 kJ/mol and a velocity of 0.001 $S_\alpha$ units/ps. Intermediate configurations spaced in 0.1 $S_\alpha$ increments were selected for US. An umbrella potential of 500 kJ/mol was applied in each umbrella window. The number of intermediate configurations (windows) for each system and the simulation time associated with each window is reported in Table 1. A total of 30.7 $\mu$s of data was collected.
2.3. Trajectory Analysis

The data excluding the first 20 ns from each umbrella window was used for both 1D and 2D Weighted Histogram Analysis Method (WHAM)\textsuperscript{72} to calculate the PMFs. Both 1D and 2D WHAM analysis was performed using Alan Grossfield’s code and the PMFs were converged to $10^{-3}$ kcal/mol.\textsuperscript{73} The GROMACS tool g_mindist was used to calculate the number of contacts between $\gamma_1$ and lipid molecules within 5 Å radius. To compute the segmental $S_\alpha$ in a given conformation, three overlapping regions of $\gamma_1$ were defined: residues 1-10, residues 6-15 and residues 12-21, described as the N-term, middle and C-term segments, respectively. The PLUMED plugin was also used for calculating the insertion depth of $\gamma_1$. For this parameter, the three non-overlapping regions of $\gamma_1$ are defined as, residues 1-7 as N-term, 8-14 as middle and 15-21 as C-term. An evaluation of the flexibility of the $S_\alpha$ restraint to allow the peptide to sample helicity in different regions of the peptide was performed using the compute_dssp utility of MDTraj.\textsuperscript{74} For each system the helicity was calculated in a low ($S_\alpha = 2$), medium ($S_\alpha = 7$) and high ($S_\alpha = 12$) window. The helicity was averaged in 10 ns segments and is presented in Fig. 5.1.S1, which shows the restraint does not rigidly fix the helical segments, but allows for some shifting of the conformation within an US window.

2.4. TICA and dTRAM Analyses

The US trajectories were analyzed using pyEMMA 2.4.\textsuperscript{74} As was done for WHAM, the first 20 ns were excluded from each US window and the frames in the input trajectory were separated by 10 ps. The system was featurized by using the distances between all the $C_\alpha$ atoms of the peptide in order to reduce the degrees of freedom for analysis. The choice of $C_\alpha$ distance pairs has been
shown to be a good metric to describe folding transitions and overcomes limitations inherent in position-based metrics.\textsuperscript{75} The time-lagged covariance matrix, \(\mathbf{C}(\tau)\), with components

\[
c_{ij}(\tau) = \langle r_i(t) r_j(t + \tau) \rangle
\]

(where \(r_i(t)\) represents different \(C_\alpha\) atom pair distances at time \(t\)) was constructed. Time-lagged independent component analysis (TICA)\textsuperscript{76,77} was performed with several different lag times all showing similar landscapes. A TICA lag time of 8 ns was chosen for further analysis because it gave the highest cumulative variance captured by the first few time-lagged independent components (tICs) and the slowest TICA timescales. TICA was performed on an individual system basis (system tICs) and also by combing all the systems into a master trajectory (global tICs). 2D projections onto tIC subspaces were performed and representative structures were chosen to visualize the transition pathways. To select the structures the trajectory data was clustered into 60 clusters based on the global tIC 1, 2, and 3 values using mini batch k-means clustering (30 clusters were used when using system tICs). A representative structure was pulled from each of the clusters we chose to highlight. Folding PMFs were also estimated using a transition based approach, discrete transition-based reweighting analysis method (dTRAM).\textsuperscript{78} Regular space clustering\textsuperscript{79} was used to generate microstates using a minimal distance separation of 0.015 \(S_\alpha\) units, from which the transition matrix was computed. A 2 ns lag time was used in the dTRAM analysis, and all PMFs were converged to a tolerance of at least \(10^{-7}\) kcal/mol.
Table 1. Simulation system details.

<table>
<thead>
<tr>
<th>System Description</th>
<th># of PC and/or PG lipids</th>
<th># US windows</th>
<th>Time of simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Per Window (ns)</td>
</tr>
<tr>
<td>Pure PC</td>
<td>270</td>
<td>125</td>
<td>80</td>
</tr>
<tr>
<td>Pure PG</td>
<td>288</td>
<td>148</td>
<td>60</td>
</tr>
<tr>
<td>Mixed</td>
<td>100:100</td>
<td>148</td>
<td>80</td>
</tr>
</tbody>
</table>

3. Results and Discussion

3.1 Folding Energetics

We assessed the folding of $\gamma_1$ as function of helicity ($S_\alpha$ value) on membranes of different lipid compositions. The three systems and the simulation details are described in Table 1. The three membrane compositions are PC (zwitterionic), PG (anionic) and a mixed bilayer composed of equal number of PC and PG lipid molecules. The energetics of $\gamma_1$ folding on the membrane were calculated from the US data, using the WHAM algorithm. Fig. 5.1.1 represents the 1D PMF for the three systems. We evaluated the total free energy change of folding of $\gamma_1$ to be similar for PC (-15.0 kcal/mol) and PG (-14.4 kcal/mol), while the mixed system had a smaller free energy change (-12.3 kcal/mol). The PC and mixed system both have a global minimum in a highly folded state (~92% helicity, state $H$), while for PG the highly folded state is metastable and the global minimum is at an intermediately helical configuration (~68% helicity, state $I$). The PC system also shows (meta)stability for $I$ state and there are barriers for both PC and PG separating the $H$ and $I$ basins. The barrier in the folding direction for the PG system is 3.6 kcal/mol and 3.0 kcal/mol for the PC
system. Interestingly the mix system PMF is qualitatively different than the homogenous systems as it does not display stability at the I state. The barrier in the folding direction for the mixed system is small (1.8 kcal/mol), while the unfolding barrier is 4.2 kcal/mol.

Some features of the free energy profiles are not consistent with our expectations based from our previous work. In equilibrium simulations we observed rapid folding on the microsecond timescale for γ₁ in a PC systems, but did not observe significant folding in PG or mixed bilayer systems. Therefore, our prediction was that PG and mixed bilayers would have higher energy barriers in the folding direction which is in contrast with the PMFs. However, the PMFs are consistent with our observation that the folded state of γ₁ is stable in all membrane compositions and must overcome an energy barrier to significantly unfold. In our equilibrium simulations, we observed differing degrees of membrane penetration in the different bilayers, so to examine this feature in the current study we calculated 2D PMFs using peptide insertion depth as the second coordinate. Fig. 5.1.2 presents the 2D PMFs and it can be seen that PC (Fig. 5.1.2A) system has several local minima along the folding pathway and samples the most deeply inserted conformations, which correlates with our equilibrium simulation observations. The PG FES (Fig. 5.1.2B) shows the least insertion and the mixed system (Fig. 5.1.2C) displays a smooth landscape without significant local minima.

3.2 Folding Mechanisms

Given the different thermodynamic stability of states along the folding pathways in different membranes environments, we wanted to understand if this reflected different folding pathways. To characterize the folding mechanisms we performed a segmental analysis of the peptide structure
and insertion depth. We calculated the $S_\alpha$ value of $\gamma_1$ over three different segments of the peptide (based on the definitions described in the Methods section). **Fig. 5.1.3** shows the segmental $S_\alpha$ value (averaged over the last 10 ns of the US simulations in each US window) with respect to the overall $S_\alpha$ value in each US window. We observe that for all three systems the folding is initiated in the N-term, followed by the middle region, and the C-term folds last. While the PC and mixed system both show the N-term folds to near completion before the middle region begins to fold, the PG system shows there is concomitant (cooperative) folding between the N-term and middle regions. There are three basic and one acidic residue in $\gamma_1$, and all of these charged residues are within the first 12 N-terminal residues of the peptide. The concentration of charges in the N-term and middle regions, maybe a driving force for the cooperative folding to align the charges in an energetically favorable manner when folding on the anionic PG membrane.

The mixed system also has an interesting feature not present in the homogeneous systems, which is that the C-term segment undergoes backtracking. The mixed system has a slightly metastable state around 40% helicity (**Fig. 5.1.1, Fig. 5.1.2C**), which is when the C-term segment begins to fold. Although, the C-term segment initiates folding, it returns to an unfolded state at around 60% total helicity. This backtracking feature may be indicative of a more frustrated folding landscape and barriers in orthogonal directions to the folding coordinate.\(^8^0\) We examined two orthogonal coordinates along the folding pathway, the segmental insertion depth for all systems (**Fig. 5.1.S2** and the ratio of PG to PC contacts in the mixed system (**Fig. 5.1.4**). The segmental insertion depths qualitatively showed similar features for all systems, namely the N-term and middle regions remained inserted through the pathway and the C-term segment transitions from solvent exposed to embedded during the folding. The deeper penetration of the N-term and middle regions of $\gamma_1$ as
opposed to the C-term region is in agreement with experimental\textsuperscript{38} and our previous computational results.\textsuperscript{37} The contacts formed by γ\textsubscript{1} with the lipids in the mixed system did not show a significant dependence on the folding coordinate, and did not shift dramatically from the initial pathway (Fig. 5.1.4). In general, the peptide maintains a higher ratio of PG to PC contacts, forming \( \sim 1.5 \) PG contacts for every PC contact. This PG:PC contact ratio was consistent with the previous equilibrium simulations.\textsuperscript{37}

Overall our ability to compare the US results with our previous equilibrium simulations, is likely hampered by differences in the pathways being sampled. The equilibrium study was initiated from a surface bound low helicity state in which the middle segment was folded while the N-term and C-term segments were unfolded. In the current study, the pathways were generated from SMD unfolding from a well equilibrated high helicity state. The US pathway does not pass through a conformation where the middle segment is folded in the absence of the N-term being folded and therefore the peptide is sampling different regions of the helicity-insertion phase space in comparing the equilibrium and US pathways.

3.3 Sampling Analyses

Our abilities to interpret the US results are dependent on several factors, including the pathway which is being sampled, but also the convergence of sampling along the generated path. We have performed several analyses to evaluate the quality of the sampling in this study. We have reconstructed the 1D free energy profile, using different segments of the US data with WHAM (Fig. 5.1.5A) and using a method based upon transition probabilities, dTRAM\textsuperscript{78} (Fig. 5.1.5B). In Fig. 5.1.5A the block averaged profiles for WHAM indicate the PG profile has low uncertainty,
the PC profile has moderate uncertainty and the mix profile has high uncertainty. Not surprisingly
the higher uncertainties are observed for the low helicity region of the PMF, which has a high
degeneracy of conformational states. TRAM methods have been shown to have a lower error than
WHAM, especially for shorter sampling times.\textsuperscript{78,81} It is seen that the results of WHAM (Fig. 5.1.1,
Fig. 5.1.5A) and dTRAM (Fig. 5.1.5B) are in qualitative agreement, and show good quantitative
agreement in the higher helicity region (> 70%). However, the WHAM profiles underestimate the
total free energy change of folding, as the dTRAM profiles have higher free energy values in the
unfolded states.

In addition to assessing the reliability of the reconstructed free energy profile, we utilized TICA
analysis to gain information on the mechanism of folding and to evaluate the landscape in the
slowest decorrelating degrees of freedom (slowest tICs). TICA has been used on US data
previously to analyze protein dynamics and as the basis for adaptive sampling.\textsuperscript{81,82} Our systems
were featurized using the distances between all C\textsubscript{α} atoms in the peptide, which provides a reduced
data set for performing TICA. Therefore, the TICA space is related to but not identical to the US
coordinate, which is the S\textsubscript{α} parameter. We performed a global TICA analysis by combining the
systems into a single-master system, which allows us to obtain global tICs, and compare the
systems on a consistent variable basis. The slowest global tIC (tIC1) describes the US coordinate
very well as the two are strongly correlated in all three systems. By projecting each US window
onto the tIC1-helicity space, it can be observed that the windows are well connected and evenly
spaced in the tIC1 coordinate (Fig. 5.1.S3). Of the three systems, only the mixed system (Fig.
5.1.S3C), has a substantial gap which occurs around 50 % helicity and may be an indication of
inadequate sampling and/or energy barriers in this region.
We also performed two-dimensional projections of the US data onto the global tIC1-tIC2 and global tIC1-tIC3 subspaces to observe the sampling in the slowest motions orthogonal to tIC1 (and by proxy orthogonal to the US coordinate). **Fig. 5.1.6** presents the 2D projections, where the surface is colored based on the biased probabilities, though relative probabilities in the 2nd and 3rd tICs, at fixed tIC1, should not be affected by the US restraints. We also performed the projections against the system specific tICs in both the tIC1-tIC2 and tIC1-tIC3 subspaces (**Fig. 5.1.S4**). The same projections onto the global tICs are presented in **Fig. 5.1.S5** and **5.1.S6**, but colored by $S_\alpha$ and US time, respectively. **Fig. 5.1.S5** shows that the folded state is occurring at high (positive) tIC1 values, while **Fig. 5.1.S6**, indicates there is no detectable drift occurring in these slow degrees of freedom during US within a window. In general, the 2D projections show a narrowing of sampling at high tIC1 values, which is consistent with the folded state having a low structural degeneracy. Four peptide configurations are shown for each system (**Fig. 5.1.6, Fig 5.1.S4**) to indicate the structural reorganization the peptide undergoes as it approaches the folded state. The PC and PG systems show relatively similar projections in **Fig. 5.1.6**, as the data is well connected in both subspaces and the sampling broadens at negative tIC1 values (low helicity states). In contrast, the mixed system displays a narrow path in the tIC1-tIC2 subspace and is disconnected in the tIC1-tIC3 subspace. While the PC and PG systems appear to follow a single pathway, the mixed system appears forked in the tIC1-tIC3 space and that a barrier located around tIC1=0 may separate these paths. A potential connection between the TICA projections and the PMF for the mixed system is that the lack of a minima in the PMF around 70% helicity could be due to the system be trapped in a higher energy pathway leading to higher energy states in the middle region (50-70% helicity) of the PMFs.
4. Conclusions

We have performed US sampling calculations to analyze the folding energetics and mechanisms of the FHV $\gamma_1$ peptide in homogeneous and mixed bilayer systems. We have estimated the free energy profile and observe that the different bilayers systems influence the equilibrium probabilities of the highly folded and intermediately folded states. The heterogeneous bilayer system displays several notable differences from the homogeneous systems, including the lack of a stable intermediately folded state, and folding backtracking in the C-term segment. Additional kinetic-based analyses show differences in sampling of the slow degrees of freedom between the homogeneous and mixed systems. In particular, a much narrower pathway is observed for the mixed system in the tIC2 direction and the appearance of multiple pathways and a barrier in the tIC3 direction. While the PMFs appear to be well converged, the TICA analysis could direct a future adaptive sampling approach which could enhance the exploration of configurations away from the initial pathway.

References


Figure 5.1. PMF of folding FHV $\gamma_1$ peptide in different membrane environments.

Figure 5.1.1. PMF of folding FHV $\gamma_1$ peptide in different membrane environments.

Figure 5.1.2. Two-dimensional free energy surface of $\gamma_1$ folding on PC (A), PG (B), and mixed (C) bilayers. The coordinates of the FES are the helicity and the peptide membrane insertion depth, measured relative to the upper leaflet mean phosphate positions.
Figure 5.1.3. Segmental folding behavior of γ₁ in PC (A), PG(B) and mixed (C) bilayers. The total peptide folding was dissected into three overlapping 10 residue segments: residues 1-10 (N-term), residues 6-15 (Middle) and residues 12-21 (C-term). The helicity values are obtained by averaging the $S_a$ values of the segments over the final 10 ns in each US window. The absolute $S_a$ values are divided by the theoretical limit which is 5 for a 10-residue segment. Curves are smoothed by performing a 5-point moving average.

Figure 5.1.4. Ratio of PG:PC lipid contacts to the γ₁ peptide in the mixed bilayer system. Data is shown for the initial pathway (blue) and averaged over the final 10 ns in each US window (red).
Figure 5.1.5. Alternative PMF reconstructions. A) WHAM was performed using 20 ns blocks and averaged over the blocks. For the PC and mixed system there were three blocks (20-40 ns, 40-60 ns and 60-80 ns) and for the PG system there were two blocks (20-40 ns and 40-60 ns). The shaded regions are the standard error ($\sigma\sqrt{n_{\text{blocks}}}$) of the block averaged PMFs. B) dTRAM analysis was performed using all data after excluding the first 20 ns for equilibration.
Figure 5.1.6. Projection of US data onto global tICs in the tIC1-tIC2 and tIC1-tIC3 subspaces for PC (A), PG (B) and mixed (C) bilayer systems. Representative structures are shown and denoted by red 1, 2, 3, 4 markers on the surfaces. The coloring of the surfaces are based on a pseudo free energy ($F = -k_B T \ln P_b$) in kcal/mol, where $P_b$ are the non-reweighted probabilities from US. In all the structures the N-terminus is oriented to the right side of the molecule and the molecules are orientated in a manner consistent with the solvent phase being in the up direction and the membrane phase being in the down
Figure 5.1.S1. Residual helicity in US windows. The DSSP helicity is shown for PC (A-C), PG (D-F) and mixed (G-I) systems. For each system the helicity is calculated in a low ($S_\alpha = 2$), medium ($S_\alpha = 7$) and high ($S_\alpha = 12$) helicity window. In each row of the figure the helicity progresses from low to high in going from left to right.
Figure 5.1.S2. Segmental membrane insertion depth of γ1 in PC (A), PG (B) and Mixed (C) bilayers. Segments are defined by 7 non-overlapping residue sequences: residues 1-7 (N-term), residues 8-14 (Middle) and residues 15-21 (C-term). Curves are smoothed by performing a 5-point moving average.

Figure 5.1.S3. Projection of the US trajectories onto the global tIC1- $S_\alpha$ coordinate space. Each US window is colored differently to distinguish the different trajectories. The data is shown for the PC (A), PG (B) and mixed (C) systems.
Figure 5.1_S4. Projection of US data onto system tICs in the tIC1-tIC2 and tIC1-tIC3 subspaces for PC (A), PG (B) and mixed (C) bilayer systems. Representative structures are shown and denoted by red 1,2,3,4 markers on the surfaces. The coloring of the surfaces are based on a pseudo free energy ($F = -k_BT \ln P_b$) in kcal/mol, where $P_b$ are the non-reweighted probabilities from US. In all the structures the N-terminus is oriented to the right side of the molecule and the molecules are orientated in a manner consistent with the solvent phase being in the up direction and the membrane phase being in the down direction.
Figure 5.1.S5. Projection of US data onto global tIC1-tIC2 (upper row) and global tIC1-tIC3 (lower row) subspaces. Data is colored based on the % helicity value of that frame.

Figure 5.1.S6. Projection of US data onto global tIC1-tIC2 (upper row) and global tIC1-tIC3 (lower row) subspaces. Data is colored based on the simulation time within each US window.
Chapter 5. Part II: Challenges with addressing the reliability of umbrella sampling: Using dimensionality reduction to gain insight into the sampling error and effectiveness of adaptively extending umbrella sampling windows
Challenges with addressing the reliability of umbrella sampling: Using dimensionality reduction to gain insight into the sampling error and effectiveness of adaptively extending umbrella sampling windows

Jason G. Pattis and Eric R. May

Introduction

Umbrella sampling is a molecular simulation technique that provides a method for studying the pathway and mechanism of a process of interest. The process of interest can be defined by one or more collective variables (CV) and multiple independent simulations are carried out by restraining each simulation to a certain value along the CV (typically referred to as windows). The probability distributions from these windows can be combined into one united energy landscape or potential of mean force (PMF) using techniques such as the Weighted Histogram Analysis Method (WHAM) (1, 2) or Transition(-based) Reweighting Analysis Method (TRAM) (3, 4) or other methods(5, 6). For PMF estimates, dTRAM(3) has several benefits over WHAM; it only requires that the system is in local equilibrium instead of global equilibrium like WHAM, and counting transitions has been shown to converge faster and be less sensitive to the choice of bin size.

There are a number of challenges to consider when performing umbrella sampling. First among these is the choice of collective variable. Umbrella sampling is typically done using a single reaction coordinate. This reduces the complexity of a system with 3N degrees of freedom down to only one dimension. A difficulty with this approach is that the collective variable needs to
capture the relevant dynamics of the system. There is no validation test currently available to
decide if this is appropriate or not. It has been found that even using a high quality reaction
coordinate like the time-lagged independent component analysis (TICA) components in a well
sampled system, many such components are required to describe the dynamics. For example, in
bovine pancreatic trypsin inhibitor, a small 58 residue protein, 18 TICA components were
required to capture 95 percent of the total kinetic variance and the top two TICA components
capture less then 10 percent of the total kinetic variance(7). Furthermore, the implied timescale
of the first five TICA components were all over a microsecond and the implied timescale
estimated from TICA of the first 10 TICA components were all over 200 nanoseconds(8). An
attempt to use umbrella sampling with only one reaction coordinate in bovine pancreatic trypsin
inhibitor, even with 100 ns windows would under-sample many processes in this system. In
addition, the PMF would only describe a small percentage of the total dynamics of the system,
because the slow motions in orthogonal degrees of freedom would be lumped into the 1-D PMF
and not individually resolved. It is difficult to discern the impact of these problems, even after
running the simulations, as the currently available error estimates would not tell you that you
have missed critical dynamics.

Another problem in reducing the description of a system to one dimension is that one
assumes the system is fully ergodic (samples from all possible conformations) and connected
without testing these. If another slow degree of freedom exists, neighboring windows may drift
in opposite directions, in motions orthogonal to the CV. Even though there is overlap along the
collective variable of interest these windows are sampling different non-overlapping areas of
phase space and should not be considered connected and whose conformations may not have similar properties that will not be properly averaged.

The WHAM equation assumes that all slow degrees of freedom are captured by the collective variable \(9, 10\). There is no way to statistically verify this assumption. If there is a degree of freedom slower than the collective variable it can cause both large errors in the PMF as well as an underestimation of the error when calculated by block averaging or bootstrapping. This is extremely problematic as it can appear as though a calculation is working properly and has low error when that’s not the case.

The most common ways to make error bars for umbrella sampling have severe limitations. Block averaging\(11\) is a common way to calculate error. Block averaging is a way to find out how many independent samples are contained within a time correlated sample. In block averaging a trajectory is cut up into many un-correlated blocks. To test for the proper block length that leads to un-correlated blocks the average value of interest and standard error is examined over different block lengths to look for a plateau. Then a measurement can be made on the average from individual blocks, and an average and standard error can be calculated between blocks. This procedure assumes the coordinates being measured is the slowest degree of freedom in the system. If there is a slower degree of freedom, individual blocks may be correlated with each other, leading to an overestimation of the effective sample count and underestimation of the uncertainty.

Bootstrapping is another common way to calculate error\(12\). Bootstrapping re-samples the already acquired data. First an auto-correlation time is calculated (equation 7). Next, a number of data points are selected with replacement equal to some percentage of the total data
(typically 50% to 80% of the total data). Sampling with replacement means the same data point can be chosen multiple times. Data points are also selected using the auto-correlation time to ensure the new data sets are not strongly correlated. This is repeated many times with each time can now be counted as one sample giving a large number of data sets where an average and standard error can be calculated. As with block averaging, if there is another degree of freedom slower that the collective variable the auto-correlation time and the standard error will be underestimated.

The sampling error can be represented by calculating the number of independent samples on the full landscape. The sampling error can be hard to quantify without a priori knowledge, as even the longest of simulations can stay stuck in local minima without discovering biologically relevant states. One strategy to quantify the sampling error is to use an automated method to detect metastable states, then use the population of metastable states as the slowest degree of freedom in the system, calculating the number of independent samples(13). The larger the number of independent samples the more likely the calculation is reproducible. Conversely, with fewer independent samples there is a larger probability a simulation has missed important structures, dynamics, and rare events. Calculating the number of independent samples of each motion in a system may provide some quantitative evidence of which motions are well sampled and which motions may need to be biased in order to be well sampled.

There is a great need for more advanced error metrics that can address the drawbacks discussed above: examining all degrees of freedom in a system, ergodicity, and the sampling error. One attempt by Zhu and Hummer uses Kullback-Leibler (KL) divergence to compare the consensus probability distribution with the observed probability distribution along the collective
variable(9). This was shown to be much better at identifying problematic PMFs than block averaging. For convergence the authors suggest all windows have a low KL divergence. Another study(14) extended this approach from WHAM to TRAM. Here the observed probability distribution is the probability distribution from one specific window (window k) and is defined in equation 1.

$$p_i^{(k), \text{observed}} = \frac{\sum_j c_{ij}^{(k)}}{\sum_j c_{j}^{(k)}}$$  \hspace{1cm} \text{equation 1}$$

$C_{ij}^k$ is the transition count from bin $i$ to bin $j$ in window $k$. The consensus distribution takes into account sampling in neighboring windows and is the final probability distribution of bin $i$ ($\pi_i$) reweighted as if it had the bias of window $k$ using the bias applied to widow $k$ ($\gamma_i^k$) and a normalization constant ($f$) (equation 2).

$$\pi_i^{(k), \text{consensus}} = f^{(k)} \gamma_i^{(k)} \pi_i$$  \hspace{1cm} \text{equation 2}$$

The authors go on to use Jensen-Shannon (JS) divergence between the consensus probability distribution and the observed probability distribution to identify problematic windows and adaptively extend those windows. JS divergence is defined in window $k$ by equation 3:

$$\text{Div}(k) = \frac{1}{2} D(\pi^{(k), \text{consensus}} \| M) + \frac{1}{2} D(\pi^{(k), \text{observed}} \| M)$$  \hspace{1cm} \text{equation 3}$$

Where $D(\pi | M)$ is the KL divergence defined as:

$$D(\pi \| M) = \sum_i \pi_i \log \frac{\pi_i}{M_i}$$  \hspace{1cm} \text{equation 4}$$

, and $M$ is the average between the consensus and observed probability distribution defined as:

$$M_i = \frac{1}{2} (\pi_i^{\text{consensus}} + \pi_i^{\text{observed}})$$  \hspace{1cm} \text{equation 5}$$
Identifying slow degrees of freedom not captured by the collective variable can be difficult. Two recent studies (14, 15) use TICA on umbrella sampling data to identify other slow motions in the system besides the collective variable. TICA and divergence can be used together to get a better picture of how well sampled an umbrella sampling calculation is. It can also be used in hindsight on a well-sampled system to compare different adaptive sampling protocols to see what a more efficient protocol would be.

These convergence and error metrics are typically tested on small test systems (3, 15) and greater investigation into how well they work on more complex biologically relevant systems is needed. The large amount of umbrella sampling data collected for FHV gamma1 peptide folding on membranes of different lipid compositions (16) provides an opportunity to test different error metrics as well as retrospectively ask if we could have gotten the same accuracy with less data to save time and computational costs.

**Methods:**

Data was binned using regular space clustering separating bins by \(0.025 \, \text{S}_{\text{alpha}}\) units. dTRAM and WHAM were run until the PMF was under the maximum error threshold of \(1 \times 10^{-4}\). Tram was performed with a lag time of 2 ns. PyEMMA (17) version 2.4 was used for both WHAM and dTRAM. The number of independent samples was calculated using the pymbar (6) package version 3.0.3.

**Methodology and Results:**

*Comparison of DTRAM vs WHAM*
It was tested to see if dTRAM converged with less data than WHAM on the umbrella sampling data from the complex system of FHV gamma1 peptide folding on membranes of different lipid compositions. The final PMF was assumed to be correct. Since the PMF is not a two state system the difference was calculated from the entire PMF. The PMFs at different timepoints were aligned by minimizing the root mean square error (RMSE) then summing the difference for each bin. The results can be seen in Figure 5.2.1. Only small differences were seen between PMFs calculated from each method, indicating that in this system dTRAM did not have a benefit over WHAM. In all three systems the difference steadily decreases over time but never plateaus as one might expect for a converged simulation.

**Comparison of Independent Samples**

To identify slow degrees of freedom in the system we performed TICA. It was determined that independent component 1 (IC 1) and the collective variable were strongly correlated, indicating that the motions in the other ICs should not be influenced by the bias(16). Next to get a quantitative measure of how well sampled these motions are, the number of independent
samples for the collective variable (CV), independent component 2 (IC 2), and IC 3 were calculated for each window. The number of independent samples was calculated by first using an autocorrelation function (equation 6).

\[
C_t \equiv \frac{\langle a_n a_{n+t} \rangle - \langle a_n \rangle^2}{\langle a_n^2 \rangle - \langle a_n \rangle^2}
\]

An autocorrelation function for the timeseries of the collective variable in PC window 120 (Figure 5.2.2A) is shown in Figure 5.2.2B. From the autocorrelation function an autocorrelation time (\(\tau_{ac}\)) can be found by calculating the area under the curve and stopping when the curve hits zero (equation 7).

\[
\tau_{ac} \equiv \sum_{t=1}^{T-1} \left( 1 - \frac{t}{T} \right) C_t
\]

Next the statistical inefficiency (g) is calculated with equation 8.

\[g = 1 + 2\tau_{ac}\]

Finally, the number of independent samples (\(N_{eff}\)) is calculated as the total simulation time (T) divided by the statistical inefficiency (equation 9). For a full discussion on calculating independent sample size see references (6, 18, 19).

\[N_{eff} = \frac{T}{g}\]
The top TICA eigenvectors give the slowest motions in the system. The number of independent samples for each TICA eigenvector (motion) for each window can be calculated which then, could be used to draw a line between what we consider slow motions from fast motions based upon the number of independent samples. The results (Figure 5.2.3) show that different windows and even neighboring windows have dramatically different numbers of independent samples. Zooming in on the low end all three systems (Figure 5.2.3D, E, and F) shows that there are windows with as few as 5 independent samples for IC 2 and IC 3 which is much lower than one would like it to be. This is also lower than the independent samples for the collective variable indicating that IC 2 and IC 3 are more poorly sampled than the collective variable. Iteratively optimizing collective variables has started to become popular(20-22). Unfortunately, the large differences in independent samples between windows makes it difficult to use this as a metric to decide how slow a specific independent component is or as a metric to

Figure 5.5.2. A) Time course of the collective variable from one example umbrella window with 20 ns thrown out for equilibration (blue). B) Auto-correlation function of the time series in A showing how correlated a data point is with another data point x ns later.
decide how many independent components should be biased in future simulations. However, it is still possible to run future simulations biasing the initial collective variable as well as one or more independent components.

Finding Problematic Windows

The next goal was to determine if one dimensional or multi-dimensional divergence could identify problematic windows. The JS divergence was calculated along the collective variable (Figure 5.2.4). Overall the divergence is very low for all three systems with PG having two slightly high divergence windows and MIX having one high divergence window all at very high percent helicity. The observed and consensus probability distributions were plotted for the highest

Figure 5.5.3. Independent sample size of different windows along collective variable of A and D), the CV B and E), IC 2 and C and F) IC 3. The bottom panels are zoomed in on the low independent samples.
divergence window of all three systems (Figure 5.2.5). The distributions in PC are fairly similar whereas the distributions in PG and MIX are noticeably different.

Figure 5.2.4. Jensen-Shannon divergence between the consensus and observed probability distribution along the collective variable

3 dimensional PMFs were created along the collective variable as well as with respect to IC 2 and IC 3 (Figure 5.2.6). Simulation data was projected onto IC 2 and IC 3 space and then was clustered into 10 clusters using k-means clustering. Each bin along the CV was then split into 10 separate bins. This process is known as the cartesian product of the two binning schemes (14). The JS divergence is calculated for the observed and consensus distribution for this 3 dimensional PMF (Figure 5.2.7). All three systems show much higher divergence and do not show higher divergence in the same windows as the 1D divergence. Several windows are quite close to the

Figure 5.2.5. Observed and consensus probability distribution of the windows with the highest JS divergence in each system
theoretical max of ln 2 (~0.693). The high divergence windows are not concentrated to one specific region of the PMF but instead spread out across the PMF. Again, the highest divergence window from each system had its observed and consensus probability distribution plotted. PC observed distribution samples from only one cluster in IC 2 and IC3 space (Figure 5.2.8). This means that in the PC system, window 72 is stuck in a local minima. The consensus distribution is exploring three clusters in IC 2 and IC 3 space, none of which are the cluster window 72 is exploring. This means that the neighboring windows around window 72 drifted in a different direction in IC 2 and IC 3 space and explored a different region of conformational space. Even though in the three dimensional space these are almost completely non overlapping in only percent helicity space these probability distributions are very similar, indicated by the low 1D divergence.

Figure 5.2.6. 3 dimensional PMF in the CV, IC2 and IC3 variables.
The PG system displayed similar characteristics to the PC system (Figure 5.2.9). The observed distribution from window 57 is stuck in a local minima sampling from only one cluster in IC 2 and IC 3 space. In the consensus distribution the neighboring windows sample from that same local minima as well as three other clusters.

![Figure 5.2.7. Divergence of the 3 dimensional PMF for the A) PC B) PG and C) MIX systems](image)

In the MIX system window 72 and the neighboring window both explored three different clusters in IC 2 and IC3 space, but they are different clusters. Window 72 moved in a positive IC 2 direction whereas the neighboring windows in the consensus distribution moved in a negative IC 2 direction.
Figure 5.2.8. Observed and consensus probability distribution for the highest divergence window for PC
Figure 5.2.9. Observed and consensus probability distributions for the highest divergence window from PG.

Figure 5.2.10. Observed and consensus probability distributions for the highest divergence window from MIX.
Comparison of Metrics for Adaptively Extending Windows

The divergence along the 1D and 3D PMFs were effective at identifying problematic windows. Next, we used and compared these metrics for adaptively extending windows with high divergence versus evenly extending window. We also considered using the slowest timescale from the biased MSM of an individual window, as a metric to extend a window. Because of the bias this timescale is hard to physically interpret, however it has been shown to be effective at identifying windows with large orthogonal barriers(5). Again in this analysis the final PMF was assumed to be correct and the difference was calculated between that and the current PMF in the adaptive scheme. All metrics started with all windows having 15 ns of data. For 1D divergence, 3D divergence, and MSM timescale, the top 30 percent of windows scoring the highest on the given metric were extended by 5 ns. For the 3 dimensional PMF TICA was recalculated each round with the current data. This method was compared against extending all windows evenly totaling the same amount of sampling. Windows were extended adaptively until at least one of the windows reached the length of the original simulation, since this was the maximal amount of data we had.
The results do not show good consistency between the different systems. There are small improvements using the adaptive strategies in the MIX system (Figure 5.2.11 C). In the PG system, CV plus 2 ICs and timescales do better than extending evenly. In PC evenly, CV and CV plus 2 ICs do about even while timescale does worse. Overall it is difficult to say that one does better than another. Even though these metrics seem efficient at finding problematic windows, these windows may stay stuck in local minima for the entirety of their simulation. Extending windows that stay stuck in local minima does not help the umbrella sampling calculation converge quicker, and points toward the value in methods which consider different structures in a given window, such as window-exchange umbrella sampling (23, 24).

Discussion

Addressing the reproducibility of umbrella sampling is an important and challenging topic in molecular simulations and we are seeing continual development of new methodologies which improve the accuracy and reproducibility of umbrella sampling. dTRAM is a recent advancement which allows better accuracy over WHAM in systems which are only sampling from local equilibrium instead of global equilibrium. When examining gamma1 folding it appears that dTRAM does not have an advantage over WHAM. This could be because the 20 ns of equilibration...
has ensured that no non-equilibrium motions along the CV are sampled during the production runs and the noise observed is due to other reasons.

Examining divergence along the CV will not always identify multiple pathways. If two neighboring windows are exploring different areas in orthogonal space that have the same energetics along the CV this area will show low divergence in one-dimension. This may not directly cause lower accuracy of the PMF but may lead to a mechanistic misinterpretation as this orthogonal conformational change may be interpreted as a change occurring as the process of interest occurs rather than orthogonal to the process of interest. This also puts strong confidence in the initial pathway as which of the two pathways are more likely to occur cannot be interpreted from the one dimensional PMF.

The divergence along the 3D PMF was able to reveal a greater level of detail. It showed that many windows were not sampling from the same areas in orthogonal space. This can be interrogated further by pulling out structures to examine how different these structures are. The 3D landscape showed a rough high dimensional landscape where several windows were stuck in local minima for the full 60 or 80 ns simulations. Many windows are not exploring the orthogonal space efficiently. This means that the accuracy of this calculation depends on the accuracy of the initial path. Often very little effort is put into generating an initial path and that can be a source of error. The high divergence was scattered across the PMF rather than concentrated to one specific region. This means that the divergence identified individual windows that were stuck in local minima rather than identifying a region of the PMF that was particularly rough and hard to converge.
Extending windows that were stuck in local minima did not speed up convergence. Adding additional enhanced sampling to these windows, even as simple as doing many short simulations instead of one long simulation in each window may have been a more successful strategy. Conversely, having additional enhanced sampling from the start stacked on top of all windows may have allowed for fast convergence of the majority of the landscape leaving only a small region that needs additional sampling. Further examination of these issues, especially in model systems where the target results are known a priori can help in evaluating the potential of new adaptive strategies.

References


Chapter 6: Conclusions and Future directions
Molecular simulations aimed at studying dynamic processes can be greatly aided by the use of dimensionality reduction and enhanced sampling to both transform these complex structural changes into a small number of easy to interpret dimensions and to accelerate the sampling of rare events. These have been used in this thesis to investigate the functions of several biological molecules.

The Lassa virus nucleoprotein (NP) is an RNA binding protein which is able to bind to RNA only at specific points in the viral life cycle. The NP is stored as a trimer to build a high concentration, then binds to and encapsulates newly formed genomic RNA during viral RNA replication. Our current understanding of this process is limited, and a better understanding of the structural transitions associated with gating to allow RNA binding, could serve as drug targets to disrupt this process. The N-terminal domain of the Lassa virus NP was found to have its dynamics dominated by helix 6 motions, which in its closed conformation, abuts the RNA binding site. The RNA bound equilibrium simulations showed a relatively flat landscape whereas the APO systems showed much smaller motions along the opening/closing direction.

Enhanced sampling of helix 6 opening shows that with RNA-bound, helix 6 has a small energy barrier to open with the open state being more energetically favorable. The APO system showed an ~17 kcal/mol energy barrier to open. Strong anti-correlated motions in the APO partially open state and the RNA-bound transition state suggest the RNA-binding pocket may come partially open, make contact with RNA, which enhances the breathing motions allowing the full conformational change to occur. The open RNA bound state observed was less open than the crystal structure which is explained by strong interactions with non-native crystal contacts.

The Hastie model suggests that disruption of the trimer of NP could allow the C-terminal domain to shift away and allow opening of the RNA-binding pocket. This is supported by
hydrogen/deuterium exchange of a trimer disrupting double mutant. A long timescale simulation followed by a two stage adaptive sampling scheme allowed creation of a Markov state model. This model of an isolated NP’s dynamics shows a large conformational change at both the domain level and within the N-terminal domain. This conformational change was favorable and can occur fairly quickly (~20 µs) whereas the transition back to a trimer like conformation takes much longer (~300 µs). During this change from trimer to monomer, the C-terminal domain shifts back behind the N-terminal domain, forming new contacts, helix 5 shifts down and in, and helix 9 and 10 shift away from the protein exposing the left side of the RNA-binding pocket.

In our simulated pathway the full opening of helix 6 as predicted by the hydrogen/deuterium exchange was not observed. This may be because the 27 µs of sampling in the simulations was much shorter than the 10 s and 1000 s experimental time for deuterium exchange measurements. It is also possible that contacts formed in the dimers and tetramers in the hydrogen/deuterium exchange cause different dynamics than that of a monomer. Adaptive sampling aimed at spawning new trajectories at large helix 6 distances from the trimer crystal structure distance could improve sampling, and potentially reveal greater solvent exposure of helix 6. However, the main results from the Markov state model support the proposed Hastie mechanism. In the monomer the C-terminal domain shifts toward the back of the N-terminal domain which would make room for an RNA strand to exit the N-terminal domain. The shifting out of helix 9 and 10 may allow initial contact of the binding pocket with RNA which may allow further conformational change toward an open bound state.

From the predictions made in this study there are many future possible experiments and calculations which could be performed. This study has suggested new contacts between the N- and C-terminal domain which could be mutated to see if they disrupt viral transcription or RNA
replication. Small molecule *in-silico* docking screens could be run on the metastable states to identify novel druggable pockets. The potential energy of the ligand bound to the metastable states could be used to estimate how perturbed the protein dynamics would be, with the inference that ligands causing more perturbed NP dynamics would be more likely to disrupt NP function. Additional studies could inform on the ribonucleoprotein complex organization through protein-protein docking of two of the relaxed monomers to each other to predict new NP-NP contacts.

It is possible that an RNA secondary structural element could act as a nucleation factor helping to polymerize NP. A structure-based (Gō) model with a potential pulling unbound RNA toward the bound state could be used to investigate the ability of RNA secondary structure elements to bind to NP. This more coarse grained approach would be computationally less demanding, so that many RNA secondary structural elements could be tested.

The importance of the different trimer contacts could be investigation by alanine-scanning mutations. The residual contribution to the binding energy could be calculated by an endpoint method such as molecular mechanics/generalized Born surface area (MM/GBSA) energy or free energy perturbation (FEP). It could also be done using a pathway technique such as umbrella sampling or metadynamics, by pulling apart the trimer.

There is still much unknown about how Lassa regulates its life cycle. The work in this thesis has supported and provided additional detail for the Hastie model of regulation between trimeric NP and the ribonucleoprotein (RNP).

Phosphodiesterase 6 (PDE6) is an enzyme which converts cyclic GMP to GMP. It is critical to the vision signaling pathway and has a very similar sequence to PDE5. Even though the sequences are very similar, PDE6 has a much faster catalytic rate. We found that the rigidity of the structured M-loop in PDE6 caused global hinging motions, shown by the first principal
component and the mutual information describing a far-reaching allostERIC network. This hinging motion may help pull substrate toward the catalytic site and help expel product. This is supported by a lower energy barrier for an inhibitor to unbind. Differences in unbinding paths were found, notably in PDE5 more contact was made with helix 15. Helix 12 was found to have strong allostERIC connections with the H- and M-loop in PDE5 and a large portion of the protein in PDE6. This may suggest that helix 12 has functional importance as an allostERIC regulatory element.

This study has provided a mechanism for how mutations far from the catalytic site can affect the catalytic rate. In the future one could build upward in system size to see if these observations hold true. A dimer of catalytic domains could be investigated looking at both apo vs one apo one substrate bound to see if correlated motions would allow allostERIC communication between dimers. The PDE6 full complex cryo-EM structure could be interrogated for parts of the PDE6 complex or other binding partners which could come into contact with helix 12 as we predict this could impact the allostERIC network and therefore the catalytic rate.

These results could serve as the basis for a computational screen for novel inhibitors. A functional screen could be performed examining the ability of novel ligands to perturb the allostERIC network. Binding affinity and off rate screens could also be performed. Absolute free energy perturbation could be used to rank the binding affinity of novel ligands and relative free energy perturbation could be used to determine if modifications to a ligand increase binding strength. One would aim for a ligand with high affinity for PDE6 and a large difference in affinity between PDE6 and PDE5 to prevent off target binding. A ligand unbinding metadynamics protocol similar to the one used here could be used to predict off rates. One would
aim for a ligand with a slow off rate for PDE6 and a large difference in off rates between PDE6 and PDE5.

Gamma peptide from Flock House virus (FHV) is responsible for lysing the endosomal cellular membrane to allow the viral genome entry into the cytoplasm. Many details remain unknown about this process. We examined gamma peptide folding on membranes of different lipid compositions. We found a strong preference for a partially folded structure ~68% helical in a PG bilayer. This conformation was less present in a PC bilayer and even less favorable in a mixed bilayer. The folding of the mixed bilayer system followed a different pathway in as determined by examining higher order TICA components. This causes the intermediate conformations and, barriers to be different on the different membranes. The folded state was also less favorable in the mixed bilayer system.

The difference in folding intermediates could lead to different oligomeric states and pores. This bias for certain membranes could be a selection for certain cell types. Possible oligomeric states could be compared by docking in an hydrophobic environment or implicit membrane.