Effects of Primary Structural Changes on HIV-1 Vif Complex Conformational Sampling

Lieza Chan
Skidmore College, lchan@skidmore.edu

Follow this and additional works at: https://creativematter.skidmore.edu/chem_stu_schol

Part of the Chemistry Commons

Recommended Citation
Chan, Lieza, "Effects of Primary Structural Changes on HIV-1 Vif Complex Conformational Sampling" (2018). Chemistry Senior Theses. 7.
https://creativematter.skidmore.edu/chem_stu_schol/7

This Thesis is brought to you for free and open access by the Chemistry at Creative Matter. It has been accepted for inclusion in Chemistry Senior Theses by an authorized administrator of Creative Matter. For more information, please contact jluo@skidmore.edu.
Effects of Primary Structural Changes on HIV-1 Vif Complex Conformational Sampling

Lieza M. Chan

Skidmore College
Chemistry Department
Katherine Aurelia Ball Lab
May 10, 2018
Table of Contents

Chapter 1: Background Information on HIV-Vif ..................................................... 3
Chapter 2: Experimental NMR Analysis................................................................. 9
Chapter 3: VCBC inclusion of full length EloB C-terminus.................................. 19
Chapter 4: Simulated Ensemble Completeness and Statistical Analysis ................. 31
Chapter 5: VCBC with mutation in Vif N-terminus............................................... 42
Chapter 6: Conclusions.......................................................................................... 50
Acknowledgments and References....................................................................... 53
Supplemental Section............................................................................................ 57
Chapter 1: Background Information on HIV-Vif

Intrinsically disordered proteins (IDP) are proteins that do not follow the traditional protein structure paradigm, which states that the function of the protein is dependent on the stability of its structure thermodynamically. IDPs are structurally dynamic and do not contain any tertiary structure yet they are responsible for many biological functions like cell signaling and regulation (Wright and Dyson, 2015). Alone, they do not fold spontaneously into a stable structure and instead fluctuate through more conformations than folded proteins (Dyson, 2016). IDPs have specific characteristics related to their flexibility that allow them to perform specific functions. For example, they contain small recognition elements that fold upon partner binding, their flexibility allows them to interact with different targets on different occasion, and they efficiently utilize conserved sequence motifs to mediate binding interactions (Wright and Dyson, 2015). The folding of many IDP proteins revolves around their ability to create protein-protein interactions, and therefore, studying their interactions with other proteins is important for understanding their biological functions. Since they mediate many cellular functions and participate in human diseases and infections, there is a necessity to find ways to characterize their dynamics and intermolecular interactions.

Human Immunodeficiency Virus Viral infectivity factor (HIV Vif) is an intrinsically disordered protein that aids the survival of HIV in the human body. HIV attacks the immune system’s T cells, decreasing the body’s ability to fight off infections and other diseases and causing patients to become more susceptible to other diseases. Currently there is no cure but there are medical treatments to control the disease. HIV uses the host cell’s ability to ubiquinate and degrade proteins on antiviral proteins. HIV Vif hijacks the ubiquitination mechanism to aid HIV spreading through the body. Understanding how the Vif protein functions could aid in understanding how IDPs interact with other proteins and lead to a potential drug target to stop the spread of HIV.

The Vif protein mediates the breakdown of antivirals by binding the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) antiviral enzyme to Cullin 5 (Cul 5) (Figure 1). Without Vif present, APOBEC3 can interact with HIV virions and inhibit viral replication with its deaminase activity (Richards et al., 2015). Deamination is the process of removing an amine group from a molecule and so it turns cytidines to uridines in the viral DNA. This causes a hypermutation that makes the viral infection nonproductive (Richards et al., 2015)
APOBEC proteins help to protect the human body from foreign infections; however, the process is prevented by Vif, which tags APOBEC proteins with ubiquitin, leading to their degradation and therefore the propagation of HIV through the body. Vif affects APOBEC3D, E, F, G, and H members of the APOBEC3 family making it a protein necessary for HIV to thrive in the body (Lu et al., 2013).

![Figure 1. Role of APOBEC proteins in propagation of HIV through host cells when Vif is and is not present. Figure reproduced from Cullen, 2003.](image)

Vif is an intrinsically disordered protein and so it needs to interact with other proteins to adopt a folded structure and to breakdown APOBEC proteins. The structure is important to the function of the protein complex because it determines the interface of protein-protein binding interactions, the stability of the complex, the conformation, and the overall surface charge (Alberts et al., 2002). Theses structural characteristics will define how a protein interacts with its environment. When Vif is inside host cells, it will bind to the proteins Elongin B (EloB) and Elongin C (EloC) at its C-terminus and to Core binding factor β (CBF-β) at its N-terminus (Evens et al., 2014). CBF-β primarily stabilizes Vif to allow for the binding of APOBEC proteins (Guo et al., 2014). EloB, EloC, and Cullin 5 (Cul5) bind to Vif to create an E3 ubiquitin ligase complex. E3 ubiquitin ligase complexes exist in the body naturally and are responsible for degrading proteins that are no longer needed (Lui and Nussinov, 2010) (Figure 2). The Vif complex is only active once all proteins are bound together and an APOBEC protein substrate
can bind to Vif.

Figure 2. a) Normal E3 Ubiquitin ligase complex in healthy host cells. RBX2 is another protein that binds the ubiquitin proteins to Cul5 via a E2 enzyme. The suppressors of cytokine signaling (SOCS)2 protein is adapter protein that connects the substrate protein to Cullin 5 (Lui and Nussinov, 2010). b) Vif hijacked ligase complex in HIV Vif+ host cells.

Understanding how this complex behaves in the body can help give better understanding of how to treat HIV since Vif is essential for HIV survival in cells (Guo et al., 2014). Protein complexes such as the VCBC complex sample different conformations some more favorable than other conformations. Looking at conformations and degree of conformational sampling reveals structure and therefore functional properties of the VCBC complex. Conformations more favorable for Vif binding to APOBEC that allow Cul5 to wrap around and attach the ubiquitin to the APOBEC antiviral could be important for developing a therapeutic treatment.

The binding of Vif to host cell proteins allows Vif to function. Vif interacts with four different host cell proteins in order to fold: CBF-β, Cul5, EloB, and EloC. The crystal structure for the Vif-CBF-β-CUL5-ELOB-ELOC (VCBC-Cul5) complex has previously been solved (Guo et al., 2014). Vif can be split up into a larger (α/β) domain and a smaller (α) domain that are connected through a linker region containing a coordinated zinc ion (Guo et al., 2014). The crystal structure found experimentally shows that Vif’s zinc-finger motif, which forms a tetrahedral interaction with a zinc ion, is important for adding stability to the domains (Guo et al., 2014). The zinc ion interacts with the HCCH arrangement of His and Cys residues on the Vif protein (Costa et al., 2014). This coordinates the Vif linker region and mutations to this region affect the binding of Cul5 (Evans et al., 2014). CBF-β binding is necessary to facilitate Vif folding and nucleation of the rest of the E3 ubiquitin ligase complex (Jager et al., 2011). There are many motifs in Vif that aid in its ability to bind to other proteins. The N-terminal peptide of
Vif forms an antiparallel β-sheet with β-strand S3 from CBF-β. CBF-β stabilizes the two loosely packing helices that make up the α-domain of Vif through binding. The α-domain of Vif contains the BC-box motif which is required for the interactions with EloB-EloC and Cul5. This interaction is similar to that seen of the SOCS-box motif in SOCS-box proteins (Guo et al., 2014). The similarity between the Vif BC-box motif and SOCS protein allow for Vif to hijack the E3 ubiquitin ligase complex since it can take on similar conformations to function as a ligase. Cul5 is responsible for attaching ubiquitin to APOBEC and so if it is not bound to the VCBC complex, there will be no degradation of APOBEC. Cul5 interacts with Vif through the helix α3 and the following loop of residues 116-131(Guo et al., 2014). Cul5 also interacts with EloC but not EloB. These specific interactions keep Cul5 and CBF-β bound to Vif giving this IDP more stability and structure.

Previous to the VCBC-Cul5 crystal structure, a slightly different model for Vif binding to EloB and EloC was proposed (Bergerson et al., 2010). It was observed that as EloBC binds with the Vif SOCS-box, the region folds (Marciszin and Engen, 2010). Through multiple NMR structures of 36 residues of Vif, EloB and EloC by Bergerson et al., two motifs were found to facilitate binding between these proteins. The Vif SOCS-box goes through a conformational change to form an α-helix which binds EloC through hydrophobic interactions (Bergerson et al., 2010). Another motif is the Pro-Pro-Leu-Pro (PPLP) motif of Vif which is seen to contribute to a hydrophobic interface for Cul5 and EloC binding (Bergerson et al., 2010). This PPLP has also been seen to interacts with residues 101-104 of the C-terminus domain of EloB (Lu et al., 2013). The interaction between the α-helix of EloC and Vif drives the folding of Vif causing the PPLP motif of Vif to interact with EloB. This weak interaction is driven by van der Waals forces, but it induces a conformational change from an unstructured to a structured state. This is important for the recruitment of CBF-β binding and degradation of A3G (Lu et al., 2013). Lu et al. suggests that this interaction could be used to make a new class of anti-HIV drugs. The other crystal structure that has been solved contains all of Vif, CBF-β, EloC, a truncated EloB, and part of Cul5 (Guo et al., 2014). This article does mention the SOCS box interactions with EloB and EloC however it does not state any interactions occurring between EloB and the PPLP motif (Guo et al., 2014) We used the Guo et al. crystal structure since it is more complete for our research. We have shown that the PPLP motif and EloB interaction did not appear in our simulations and therefore the Lu et al. article may have only seen this interaction due to the use
of an incomplete Vif structure.

The structure of the VCBC-Cul5 complex has been solved by x-ray crystallography (Guo et al., 2014). Methyl-TROSY NMR was completed by our lab through labeling the methyl peaks of specific amino acids—valine, isoleucine, leucine, alanine, methionine—of the backbone carbon (Cα) and the hydrogen attached (Hα). Specific amino acids, IVLMA, were labeled and compared to learn about conformational sampling on the ms-sec timescale causing peak broadening from the intermediate conformational exchange. NMR of the VCBC-Cul5 protein complex was attempted but the data was not well resolved which could be due to the complex being too large. VCBC without Cul5 was expressed and a spectrum was found; however, there were very little methyl peaks that appeared on the spectrum. The crystal structure of VCBC-Cul5 is missing 16 residues on the C-terminus of EloB (Guo et al., 2014). A construct containing these 16 residues was expressed and the new protein complex produced a better signal with more methyl peaks in the NMR spectra. This led to a more in-depth comparison of the VCBC Guo et al. crystal structure (VCBC truncated) compare to VCBC with the full-length EloB structure (VCBC) to understand why the spectra were so different. We predicted that with the 16 residues the structure was more stable and therefore the NMR showed more methyl peaks.

Characterizing IDPs in different conformational states more challenging since most experimental methods used, like NMR and crystallography, take an average of the overall structure. However, using computational methods to model the protein motions and interactions atomic level detail of any VCBC complex can be studied. When taking the NMR signal from a protein, this gives an average over time of a protein conformation. Using Molecular Dynamics (MD) simulations, the individual conformations of the protein can be modeled. All computational data must be compared to experimental results to conclude a certain model is realistic and reasonable to use for the protein complex.

MD is useful for determining different conformations of proteins and the transitions between the conformations. This is helpful for understanding flexible protein complexes that change conformations over time because it can capture different conformations in free energy landscapes and see how flexible a protein is (Pecora de Barros et al., 2017). Combining experimental data and MD simulations is helpful for creating a model of a protein at the atomic level that would be a realistic model in vivo (Perilla et al., 2017).
In Chapter 2, NMR spectra are analyzed to compare the number of available peaks that differ between spectrums of VCBC, truncated VCBC, and VCBC bound to APOBEC3F (A3F) and our attempt to assign all the peaks. Our NMR experimental data was collected by Prof. Aurelia Ball. Looking at the spectra of the VCBC complex with and without A3F, shows us that the VCBC complex with A3F (VCBC-A3) gave a clear signal with more methyl peaks appearing. This could be due to the decrease in exchange occurring on the intermediate time scale. We also try to identify the different peaks that appeared in our spectra and what type of residue they were. Using a python package called NMRglue, I labeled peaks by protein and residues type after generalizing what type of residue would appear in which area of the spectra. I also compared the VCBC truncated NMR spectrum to the VCBC NMR spectrum and saw that the VCBC spectrum has a clearer signal and produced better NMR results.

Our NMR results led us to run MD simulations of the VCBC truncated structure, and VCBC in Chapter 3. We ran simulations of VCBC truncated and VCBC to analyze why the NMR experimental data showed differences when a small primary structural change was made to the protein complex through the addition of 16 residues to the C-terminus of EloB. All these proteins are necessary for the E3 ubiquitin ligase complex to be activated so it is important to characterize the conformational sampling. Each of these simulations were analyzed in our lab using different analysis techniques. The primary method for characterizing global conformational sampling was Principle Component Analysis (PCA). PCA separates the motions of the protein complex into correlated global motions and unessential motions. PCA allows us to represent protein conformations as vectors that cover the motions of a molecular structure over time (David and Jacobs 2014). Each principle component represents one global motion and by projecting the protein complex trajectories onto them, we can visualize how the protein complex is moving. This gives us a qualitative way to characterize the global conformational sampling. Between VCBC truncated and VCBC there was about the same degree of conformational sampling along PC1 and PC2.

Chapter 4 discusses how we showed that the results we found were statistically significant. To test if our simulations of VCBC are run for enough time for the constructs to sample completely through their free energy landscape, the root-mean-squared deviation (RMSD) and radius of gyration are measured over the simulations. The running average of both RMSD and radius of gyration was calculated. As the average of each simulation converged over
time, the distance between the atoms and the movement of the protein relative to its center of mass show when they are no longer sampling new local conformations. This indicates that the VCBC simulations have been run long enough to sample most conformational state that it would exhibit. We used PCA on VCBC truncated, VCBC, VCBC-A3F, and VCBC-Cul5 construct 100-ns blocks to observe how the sum of the top 5 PCs changes as the simulations run. This represents how the variation of how much global correlated motions is within the top 5 eigenvectors. From this we say that the last 300 ns of all simulations gave similar variance of the global motions and therefore the last 300ns will be used for all analysis methods. Sigma-R plots were to show equilibration of independent simulations over time through viewing that the simulations 100ns blocks are the same for each construct. To test if our PCA values were significant, we used the Kolmogorov-Smirnov p-values to show that the projection of VCBC and VCBC-Cul5 on their combined eigenvectors is significantly different. Other constructs like VCBC truncated and VCBC-A3F are plotted on the combined eigenvectors but it is observed that the p-values found are not as useful to us since the eigenvectors are not completed using these construct trajectories.

In the final chapter, Chapter 5, VCBC is mutated (mutant VCBC) to observe any changes in global or local conformations that could lead to a potential drug target. In Vif, residue 25, which is a valine, was mutated to an alanine. Simulations were run with this mutation and through running averages of radius of gyration, we conclude that the simulations need to be run for longer since not all possible conformations had been sampled yet form the free energy landscape. Preliminary results show through PCA and atomic fluctuations that the VCBC mutant sampled similar degrees of global correlated motions and local atomic flexibility to VCBC-Cul5. However, more independent simulations and continuous simulations need to be completed before any definitive results can be concluded.
Chapter 2: Experimental NMR Analysis

Nuclear Magnetic Resonance (NMR) is a technique that can be used to discover the structure of compounds. Most protein structures are found using NMR or x-ray crystallography. NMR depends on the intrinsic quantum property of the nucleus, the spin. When the nucleus is in an applied magnetic field, the spin will either aligned or spin opposite direction. The electromagnetic radiation can flip the spin orientation compared to the magnetic field and this spin-flip has a change in energy. This change of energy is related to the Larmor frequency, which is the frequency at which the magnetic dipole precesses around the x-y plane due to the applied magnetic field. An NMR spectrum is a plot of the radio frequency applied compared to the absorption. The chemical shift is the difference between the Larmor frequency and the measured frequency. Using regular NMR to find structural and dynamic information on macromolecules is difficult because of poor peak dispersion, spectral crowding, and especially line broadening because of rapid transverse relaxation of nuclear magnetization (Clark et al., 2015).

NMR can be used to study the structure of proteins and their dynamics using a specific type method called methyl transverse relaxation optimized spectroscopy (TROSY). Methyl TROSY NMR used destructive interference between relaxation mechanisms to create spectra with improved resolution (Ollerenshaw et al., 2003). Labeling the $^1$H and $^{13}$C of methyl groups in selected amino acids allows for producing NMR spectra of protein samples (Ollerenshaw et al., 2003). Methyl groups are useful spectroscopic probes because they are found throughout a folded protein core and the resonance of these groups are intense and well dispersed (Clark et al., 2015). Relaxation due to $^1$H-$^1$H and $^1$H-$^{13}$C dipolar interactions are slow for macromolecules which is why TROSY experiments work well for proteins. In methyl systems, there are low relaxation rates because of destructive interference between dipolar relaxation interactions. This can be used to create $^1$H-$^{13}$C correlation maps to identify specific amino acids in a protein (Ollerenshaw JE et al., 2003). This method has been used to label eukaryotic protein actin successfully through $^{13}$C isoleucine δ1-methy labeling (Clark et al., 2015). Another article observed the catalytic domain of ABI kinase through $^1$H-$^{15}$N labelling in inset cells (Strauss et al., 2005). This method of analyzing the dynamics of proteins has been used in many experiments to identify the structure and dynamics of protein domains and the therefore was used by our lab to compare the dynamics of the VCBC complex.

Methyl-TROSY NMR was completed on three protein complexes. The crystal structure
for VCBC-Cul5 was solved for but 16 residues on the C-terminus of EloB were not included (Guo et al., 2014). The structure of the full-length EloB, EloC and half of the Vif protein (36 residues) was solved for (Lu et al., 2013). NMR experiments were completed on VCBC with the full length EloB and with truncated EloB C-terminus. NMR of the VCBC-A3F complex was also completed to compare the dynamics of when A3F is bound to the VCBC complex verses when it is not. NMR data was not collected of the VCBC-Cul5 complex due to the Cul5 protein being too large.

Methods

Methyl-TROSY NMR of the $^{13}$C-ILVMA-labeled VCBC complex was collected with the VCBC protein complex bound to single-stranded DNA. This selective labeling on specific amino acids allows for a reduction of peaks on the spectrum to decrease overlapping peaks. The VCBC complex tends to aggregate and precipitate out at high concentrations which are needed for NMR. To overcome this issue, the complex was bound to 14-T repeat of single-stranded DNA (dT14) to allow the experiment to be done under low-salt conditions (20 mM NaCl) and stopped the broadening of individual methyl peaks. The DNA helped to shield the positive surface charge of Vif from electrostatic interactions with other Vif complexes in solution.

The NMR spectroscopy was completed in 20 mM HEPES buffer at pH 7.5 with 20 mM NaCl and 0.5 TCEP. An 800 MHZ Bruker spectrometer with a cryogenically cooled probe was used to perform NMR at 300K (Tugarnio, 2006). In the VCBC-dT14 experiments, the concentration of VCBC was ~160 µM VCBC and completed with 256 scans in the $^1$H dimension. For the VCBC-dT14-A3Fctd experiments 816 scans were done in the $^1$H dimension, and the concentration of A3Fctd was ~108 µM and for VCBC was ~90 µM. Each of the spectra were processed using NMRPipe program (Delaglio 1995).

The figures and chemical shift assignments were done using python module NMRglue (Helmus et al., 2013). Picking peaks was completed using the NMRglue package. Many peaks that overlapped were taken out and some were added in that were not picked through NMRglue for the NMR spectra with A3F in solution. The peaks were also adjusted to fit the spectra more easily which was applied to all peaks evenly to ensure the relationship between peaks remains the same.
Results

Using Methyl-TROSY NMR the VCBC complex dynamics can be analyzed when adding other residues or proteins. The spectrum shows specific methyl peaks while other peaks disappear due to broadening of peaks. These broadened peaks are due to global conformational changes that occur on the intermediate NMR exchange timescale which is between milliseconds and seconds for macromolecules like proteins (Mittermaier and Kay, 2006).

When comparing the VCBC with full length EloB complex and the VCBC-A3F complex NMR spectra, more individual methyl peaks were seen when A3F was present (Figure 3). In the VCBC-A3F complex, A3F was not labeled; therefore, all methyl peaks were from VCBC for both spectra. The disappearance of peaks is due to less exchange on the intermediate timescale and since there is less conformational sampling, more peaks will appear. We see that VCBC-A3F is a more stable protein complex due to more peaks being visible on the NMR spectra.

![Figure 3. Overlayed NMR spectra where the red is with A3F and the black was VCBC without A3F.](image)

Since the VCBC-A3F spectrum contained more peaks, each peak is labeled with a dot to identify singular peaks that can possibly be assigned to a particular amino acid in a protein of the VCBC-A3F complex (Fig 4). A total of 152 peaks were picked out of the VCBC-A3F spectrum that could each correspond to an ILVMA amino acid.
In the VCBC complex, there are 194 ILVMA-type of amino acids that could possibly appear on the VCBC NMR Spectra. However, since there were not that many peaks picked on the VCBC-A3F spectra, it is likely some of the peaks were overlapped especially in the central
section of the spectra. There are likely many peaks in this region that cannot be individually identified. The loss of peaks could also be due to exchange on the intermediate timescale causing some amino acids in more flexible regions of the protein to disappear. The average signal of a flexible region of the protein will flatten out and not appear at this contour level height.

With the peaks that we were able to pick out of the spectra, we tried to identify which of the peaks were from specific proteins in the protein complex. Looking through the BMRB database, most of the ILVMA methyl groups have published resonances; however, no peaks of these specific types of amino acids were published of Vif (Table 1). The EloB and EloC peaks came the Lu et al. NMR structures; however, none of the Vif peaks that were labeled in our spectra had respective chemical shifts published in the article (Lu et al., 2013). The CBF-β peaks came from a separate NMR structure that was independent of other proteins (Huang et al., 1998). Some of the peaks correspond with the picked peaks on the VCBC and VCBC-A3F spectra (Table 1). More peaks were present that corresponded with ILVMA peaks in the VCBC-A3F spectrum compare to the VCBC spectrum.

Table 1. Methyl groups in the VCBC spectra that match published resonances.

<table>
<thead>
<tr>
<th>Protein</th>
<th># of ILVMA methyl groups</th>
<th># of ILVMA methyl groups with published resonances</th>
<th># of the published ILVMA resonances in VCBC-dT14 spectrum</th>
<th># of the published ILVMA resonances in VCBC-dT14-A3Fctd spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vif</td>
<td>88</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CBF-β</td>
<td>60</td>
<td>56</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>EloB</td>
<td>24</td>
<td>24</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>EloC</td>
<td>22</td>
<td>22</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

We plotted the published peaks of EloB, EloC, and CBF-β onto our VCBC-A3F spectrum to identify which peaks on the spectra corresponded with the specific VCBC proteins (Figure 6). When placing where peak locations were of each protein from published NMR chemical shifts, only a small number of EloB and EloC peaks corresponded to peaks on the VCBC-A3F spectrum.
Figure 6. Labeled amino acids from published NMR peaks for CBF-β, EloB, and EloC placed on top of VCBC-A3F spectra.

When comparing the database peaks to the picked peaks of the VCBC-A3F spectrum, most peaks could not be identified for which protein it belongs to. There is no specific region of the spectrum that each protein contains amino acids independently from the others. Amino acids from each protein are mixed throughout the spectra. Then, we colored the dots by their ILVMA-type of amino acid to try to specify the amino acid type for each peak.

Using the BMRB database, the chemical shifts for the labeled carbons and hydrogens for ILVMA amino acid regions were highlighted to identify the amino acid that corresponded with each peak (Figure 4). However, due to the major overlap of chemical shifts for the amino acids it is difficult to define each peak to a specific amino acid type. Instead, the different regions of where ILVMA give us an idea of which residues overlapped and which would be more readily countable.
Figure 7. VCBC-A3F NMR spectrum with picked peaks of NMRglue package in blue with the regions of peaks for specific amino acids would appear were colored accordingly.

From the spectras, it was seen that the Isoluecines were probably the easiest to count compare to the others since there was not as much overlap with other residue types. There were 18 peaks in the Ile-Cα region of the VCBC-A3F spectrum. However, in the VCBC spectrum only 2-3 peaks appeared in the same region. This supports that more broadening of peaks occurred in the VCBC spectrum like due to the reduced conformational exchange on the intermediate timescale for the VCBC-A3F protein complex.

The NMR spectrum with the truncated EloB in the VCBC complex contained fewer peaks. This was likely due to conformational sampling occurring at a intermediate timescale that could not be captured by NMR. However, by including the full-length tail of EloB, a small primary structure change of an additional 16 more residues, the NMR spectra improved and more methyl peaks were present. This is possibly due to the added stability to the complex using the full 118 residues in EloB.
The extra 16 residues on EloB affected the NMR signal so that more methyl peaks, each representing a single amino acid, could be identified. The full-length tail quenches more dynamics and less global conformational sampling to decrease the broadening and disappearance of methyl peaks. We explore this primary structural change through the use of computations MD simulations to understand how a small primary structural change could possibly affect global conformational sampling. Simulations of both VCBC complex structures including the full length EloB C-terminus and truncating EloB were compared using different analysis methods.

Discussion

Between the VCBC and VCBC-A3F spectra, more peaks were observed when A3F was bound to the VCBC complex. More singular peaks were picked from the VCBC-A3F spectrum that corresponded to ILVMA published peaks. Through trying to identify and label each peak to specific proteins and amino acid types, we learned it was difficult to pick out each peak through the methods of comparing our spectra to other known chemical shifts for specific residues in
published protein databases for ELoB, ELoC, and CBF-β. Sprangers and Kay in 2007 were able to identify and quantify specific domains of a 20S proteosome that might facilitate substrate localization to the catalytic chamber and modulate interactions of the nuclear localization-type signal with receptors. They were also able to gain insight into the function of the proteasome due to the mapping of one of the binding sites (Spranger and Kay, 2007). They were able to use Methyl-TROSY NMR with nuclear Overhaouse effect (NOE) to observe the dynamics of the proteosome and label each methyl peak on the spectra. Being able to do this for the VCBC complex would be helpful to also understand how VCBC modulates interactions with APOBEC proteins and give insight to how VCBC functions. Finding new methods for assigning all the peaks in the VCBC-3F spectra will be necessary for us to assign our spectra and give more quantitative differences on conformational differences between VCBC and VCBC-A3F.

Comparing VCBC with the extra 16 residues on the C-terminus of EloB and without made it necessary to gather more atomic level detail to understand why there was an increase in methyl peaks when the only change was a small primary structural difference. If the change caused a decrease in conformational sampling on the intermediate timescale, it leads us to believe that VCBC with the full-length EloB is a more stable structure and therefore this VCBC complex could be used to compare when other proteins like A3F or Cul5 are added to the structure instead of the crystal structure version of EloB. In the next chapter, MD simulations were run for both starting structures to look at global and local conformational sampling that distinguish the two structures.
Chapter 3: VCBC inclusion of full length C-terminus in EloB

From, previously run NMR experiments (chapter 2), we observe that there is a difference in exchange occurring on the intermediate timescale due to a small primary structure change in EloB. Experimentally, the NMR spectra of the VCBC complex with the truncated EloB contained fewer individual methyl peaks. However, when EloB contained its full C-terminus tail, a better spectrum was produced with more peaks appearing. From the crystal structure of VCBC-Cul5 that was solved, EloB is missing 16 residues off the C-terminus end of the protein (Guo et al., 2014). However, other scientists have solved the structure of the full EloB with EloC and part of Vif through NMR spectroscopy (Lu et al., 2013). Adding in the 16 residues into our VCBC simulations could affect the VCBC complex.

EloB interactions with EloC and Vif allow for Vif to gain structure and bind to APOBEC proteins through the necessary interactions. The N-terminus (UBL domain) of EloB binds to EloC and the COOH-terminal tail of EloB binds to the PPLP motif of Vif in a crystal structure of part of Vif, EloB, and EloC (Lu et al., 2013). Also seen through cell assays, the PPLP motif binds to EloB for the formation of functional complex (Bergeron et al., 2010). These interactions can induce a structural change in Vif that facilitates Vif's interactions with both Cul 5 and CBF-β (Lu et al., 2013). The EloB C-terminus tail has been proposed to improve Vif function and could be a possible drug target (Wang et al., 2013). Through mutating EloB to be missing 34 residues of the C-terminus tail, the interaction between Vif and CBF-β was impaired (Wang et al., 2013). The C-terminus of Vif is responsible for interacting, stabilizing, and assisting in folding of proteins of the SOCS family (Bullock et al., 2006 and 2007). Vif contains a SOCS-box that is responsible for binding to EloB on the C-terminus (Kamura et al., 2004). The binding of EloB and EloC to the SOCS-box could enhance the folding of full-length Vif in Escherichia coli (Wang et al., 2013). The C-terminus of EloB plays key roles in the folding and function of Vif, and therefore, if even part of those proteins are missing this could affect the VCBC-Cul5 complex from breaking down APOBEC proteins.

To investigate it further, MD simulations were run to compare the different conformations each exhibits and the amount of conformational sampling. To run MD simulations, the crystal structure was used for the starting structure with some alterations depending on the simulations need to be run(Guo et al., 2014). We decided to run simulations of VCBC with the 16 residues on the C-terminus of EloB (VCBC) and without the 16 residues
Running simulations on both structures also allows us to more directly compare our experimental results to our computational analysis since the structures will be exactly the same. Our primary analysis of global motions was done using Principle Component Analysis (PCA). This analysis allows us to find relationships and patterns in our protein complex movements that would be invisible from a pure visual examination. When applied to MD simulations, it detects global correlated motions of the systems which are called the principle components. PCA splits the configurational space into 2 sub-spaces; essential and irrelevant. The essential subspace consists of the correlated motions and are functionally important. The irrelevant subspace is all the independent, constrained fluctuations that act locally. By using PCA, the overall conformation changes and correlated motions of the VCBC complex with the full-length EloB C-terminus can be compared to the conformational changes of VCBC with the shorter EloB tail.

First, a covariance matrix is made to separate the essential motions to the irrelevant motions that define the subspaces. Covariance is the measurement of the degree that two random variables change together in a correlated or uncorrelated manner. The following equation is used to build the matrix to compare all α-Carbons to subsequent α-Carbons through the simulations.

\[
\text{Cov}(A, B) = \frac{\sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})}{n}
\]

where \(A_i - \bar{A}\) is the change in x, y, or z coordinates of α-Carbon from initial position to time averaged position and \(B_i - \bar{B}\) is the change in x, y, or z coordinates of another α-Carbon from initial position to time averaged position. The correlated degrees of freedom movement can be either positive or negative. If the value found is zero, then the movements are independent and placed into the irrelevant subspace. Then we diagonalize the Covariance matrix \((M)\) and using the following equation.

\[
M \nu = \lambda \nu
\]

where \(M\) represents the covariance matrix, find \(\nu\), the eigenvector, and \(\lambda\), the eigenvalue. The eigenvector is a metric of the correlated displacement of groups of \(C_\alpha\) through space. The eigenvalues, which are also called the principle components (PCs), represent the relationship between the dimensions of eigenvectors. The larger the eigenvalue, the more simulation time the eigenvector motion is capturing. The essential dynamics (ED) are the small number of principal components that describe most of the total changes in the atoms of the proteins. They are usually
the top three or five eigenvalues. The ED are important for studying large conformational changes in a protein complex (Skjaeven et al., 2010).

Methods

**MD Simulations**

The sequence of Vif-CBF-β- EloB- EloC with the full-length Elongin B (residues 1 to 118) was used to run 8 simulations. The starting structure came from two crystal structures. The main structure for Vif, CBF-β, EloC and residues 1-78 of EloB was taken from the crystal structure (4N9F) with Cullin 5 removed (Guo et al., 2014). The structures of residue 79-118 came from a crystal structure of the HIV Vif SOCS-box of EloBC, 2MA9 (Lu et al., 2013). To combine the two structures, residues 1-78 of EloB were aligned in the two structures using Visual Molecular Dynamics (VMD) (Case et al., 2016), and residues 79-102 of EloB from 4N9F were replaced with residues 79-118 of EloB from 2MA9. We then run simulations on this combined VCBC structure with the full-length EloB C-terminus (VCBC).

To set up the simulation, the new combined PDB to of VCBC was read into AMBER, and the force field and water models were chosen. AMBER99sb is a force field that is commonly used for IDPs as it has been consistent with experimental and the water model TIP3P works well with this force field (Rauscher et al., 2015). The Amber99sb force field (Case et al., 2016) and the TIP3P water model (Hornak et al., 2006) were used to perform simulations to stay consistent with previous simulations run in the lab as well. Amber was used to complete MD simulations and some analysis (Case et al., 2016). The force field defines the potential energy for all the atoms. When each atom moves, there will be a change of forces acting on the atoms and therefore the calculations needed were completed every femtosecond. This determines how the atoms move in accordance with the atoms around it. Two sets of tleap were run to make sure the format of the PDB file was compatible to use for simulations first and then to set up the simulations. During the second tleap, a parameter file and an input coordinate file are made based on the PDB file. The parameter files contain all information about the protein complex that doesn’t change throughout the simulations, like the atom structure of each residue, the charge of each residue, the order of amino acids in the protein chain. The input coordinate file contains the starting Cartesian coordinates for all atoms in the protein structure. Then, a water box is made around the protein and filled with water. Two more parameter and input coordinate files are
made that include all water atoms. Next energy minimization is completed twice; first the protein is restrained to allow the solvent to minimize to the lowest potential energy state and a second minimization without the protein restrained so that the protein structure minimizes to its lowest potential energy state. The first time was with restraints on the protein set to 500 kcal/mol with a 1000 steps steep decent followed by a 1000 steps of gradient minimization. The second time was without the restrained protein set to 10 kcal/mol to allow the protein to adjust to its minimization state without having the water molecules disturb the protein movements the same steps of steep decent and gradient minimization. Next the system is heated up mimic host cell body conditions. The box was heated up from 0 to 300 K completed in 20 ps with the protein restraints set to 10.0 kcal/mol. Two steps of equilibration are done first with the protein restrained and then without. The first equilibration is kept at a constant density with random velocity set with the protein restrained to allow the solvent to equilibrate to the pressure and temperatures. The first equilibration is done with the protein restrained at 1.0 kcal/mol over 20ps. Then the next equilibration step allows the protein conformation to equilibrate to the constant pressure before running the simulations. Then the simulations are begun and multiple independent simulations can be run at once. Traditional MD was used and a total of 600 ns was achieved for each of the 8 independent simulations for a total of 4800 ns of simulation. These simulations with the full length EloB tail were compared to previous simulations with the truncated EloB tail to see if there is a difference in the overall conformational ensemble.

Structural and Trajectory analysis

Analysis of the molecular dynamics simulations were performed using the cpptraj module of Amber (Case et al., 2016), and in-house python scripts. Interatomic distances, root-mean-squared-deviations (RMSD) of protein structures, root-mean-squared fluctuations, and cpptraj (Case et al., 2016). Analysis was completed on all 8 independent simulations and the last 300 ns.

Atomic Fluctuations

To look at more local conformational sampling, we measured the atomic fluctuations of each protein. The atomic fluctuations were calculated using the cpptraj module of Amber (Case et al., 2016). Before calculating the fluctuations, each protein was aligned independently using
RMSD so that the atomic fluctuations measured only represented local motions of each protein. The root mean squared fluctuations (RMSF) measured the flexibility of each residue throughout the simulations and shows which areas are more stable possibly due to interactions with other proteins of confinement of movement due to the placement of other proteins.

\[
\sqrt{\langle \Delta r^2 \rangle}
\]

where \( r \) is a three-dimensional position vector such that

\[
\langle \Delta r^2 \rangle = \langle \Delta x^2 + \Delta y^2 + \Delta z^2 \rangle
\]

The atomic fluctuations were calculated for VCBC with and without the full-length EloB and graphed using python onto separate plots.

**Principle Component Analysis**

PCA was completed using the *cpptraj* module of Amber (Case et al., 2016).

To prepare for PCA, all flexible loops and tail were removed from the structure so that they would not be considered in the grouping of correlated motions. So that the rotation of the entire protein was not considered and only movement of the protein relative to itself were used, root-mean-squared-deviation (RMSD) was used to align all the atoms on top of each other for every frame. The average of all structures in the trajectory was used as the reference. Cartesian coordinates of the C\(_\alpha\) position for each residue in the structured regions of the VCBC and VCBC with the truncated tail complexes (excluding flexible tails and loops). Three covariance matrixes were made to find different eigenvectors and eigenvalues. The first covariance matrix was from the VCBC truncated, the second from VCBC, and the third from both VCBC truncated and VCBC. Both sets of simulations were projected onto the different PCs to compare the global motions of the VCBC complex.

Histogram plots were made to view the spread of conformations sampled along each principle component through in-house python scripts. The resulting three sets of PCs with VCBC truncated and VCBC were compared to one another.

To compare the principal components from the VCBC truncated to the newly calculated principal components of VCBC, the overlap between the eigenvectors for the different proteins were compared. First the coordinates were normalized
where $a$ is each x, y, and z coordinate for the PC. This puts each length of the PC ranked from 0 to 1. Then, the overlap between two PC were done through the dot product of the normalized coordinate values

$$a \cdot b = \sum_{i}^{n} a_i \ast b_i$$

where $a$ are the lengths of one PC and b is the lengths of the other PC. The overlap quantified the similarity between the principal components of VCB118C and VCBC (Ozcan et al., 2013).

**Results**

Eight simulations were run on VCBC, each for 600 ns for a total of 4800ns of data. The starting structure for the VCBC simulations was constructed from the combination of two PDB files. Looking at the structure in VMD, the extra residue formed a beta-hairpin loop and interacted with EloC but it was not seen to interact with Vif (Figure 9).

![Figure 9. VCBC complex with full length tail including all 118 residues of EloB. The EloB section in lighter blue were residues that came from a different PDB structure than the other residues.](image)

This additional 16 residues contained more secondary structure in the C-terminus tail. This additional structure is thought to be the reason for the difference in NMR signals because the primary structural change lead to an increase in secondary structure which can affect the overall stability of the protein complex.

We looked at the atomic fluctuations of the backbone in EloB and EloC. We decided to look at EloC because the beta-hairpin loop of the C-terminus tail looked as if it was interacting
with EloC. The backbone atomic fluctuations of 4 independent simulations of VCBC and VCBC truncated were compared to see if there were any changes in flexibility with the addition of the 16 residues (Figure 10). The extra residues in EloB causes a decrease in flexibility of the C-terminus tail. With the truncated tail, the atomic fluctuations were much higher than any other residues in EloB.

![Figure 10. Atomic fluctuations of (A) EloB residues and (B) EloC residues. The atomic fluctuations of t VCBC complex (blue) and with VCBC truncated complex (pink). Analysis completed on the last 300ns of 4 independent simulations.](image)

This decrease in atomic fluctuations in the C-terminus of EloB was predicted because of the increase in secondary structure seen in the PDB made. The increase in structure decreases the flexibility and fluctuation of the tail. However, we expected to see a difference in the atomic fluctuations in EloC because in the PDB structure, there was a β-hairpin loop that interacted with EloC. We assumed that there would be a decrease in fluctuations where EloB was interacting with EloC; however, there is no difference observed between the two structures of EloC. Afterward we looked at the simulations in Visual Molecular Dynamics (VMD) a molecular visualization program for displaying large biomolecular system using 3-D graphics, we saw that the β-hairpin loop was not conserved during the simulation. It only stayed for the first couple of snapshots before the tail moved further away from EloC to a distance where EloB and EloC were no longer interacting along the EloB C-terminus tail.

PCA was completed on each of the simulations and the first eigenvalues captured more of the global correlated motions. The histograms for the VCBC and VCBC truncated individual eigenvectors did not give good results (Supplemental Section: Figure A). We expected to see more variation along the eigenvector for the simulations that were used to...
create the eigenvectors. For example, we thought that the VCBC truncated would have larger range of correlated motions along its eigenvector compare to VCBC. However, this was not the case for any of the projections. We also tried to split them up by individual simulation, but still did not see what we expected. Then we compared the overlap between the two eigenvalues. The overlap between VCBC and VCBC truncated PCs were calculated to see how similar or different their global motions were. The higher values represented more overlap between the PCs (Table 2).

Table 2. Overlap of VCBC with full length tail (VCB118C) and VCBC PCs which are not the same PC for each construct.

<table>
<thead>
<tr>
<th>VCBC truncated PCs</th>
<th>VCBC PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>49.97</td>
</tr>
<tr>
<td>2</td>
<td>52.69</td>
</tr>
<tr>
<td>4</td>
<td>1.73</td>
</tr>
</tbody>
</table>

The highest overlap is VCBC PC1 compared to VCBC truncated PC1 and PC2. The overlap can tell us whether the ED are similar. PC1 of VCBC truncated and PC1 of VCBC overlap can show us that the ED of VCBC and VCBC truncated are similar. This significant overlap shows that both constructs are sampling similar ED and degrees of global correlated motions.

We also found the eigenvalues based on both the full length and truncated VCBC eigenvectors and projected the snapshots of the simulations on to the eigenvectors. PC1 represents a clamshell opening and closing (Figure 10). PC2 represents an inward and outward twist of EloB and EloC compare to Vif and CBF-β. These motions hinge from Vif linker region, connects the α/β domain and the α domain. Overall conformational sampling between the two structures were relatively similar. The conformations that both complexes exhibited were far from the crystal structure along both PC1 and PC2.
Figure 10. PCA of VCBC with and without the full-length EloB tail ensembles. PC1 (43%) shows a clamshell opening and closing motion. PC2 (17.6%) represents an inward and outward twist of EloBC and Vif and CBF-β.

Both structures sampled about same degree of conformations along PC1 and PC2. Along PC1, the VCBC truncated complex has slightly wider open clamshell conformations than VCBC. From the PCA, it is seen that there are no major changes in global conformations when adding in the 16 residues. Explaining and viewing PCA can be difficult for students who are not familiar with how they are formed and what form of measurement it is. It can be beneficial for readers to have another variable related to PCA to grasp what is being compared.

Along PC1, the clamshell motion, when the protein complex is in a more closed conformation
some residues are a lot closer together than when the protein complex is in the open clamshell conformations. Using this residue to residue distance is used as another metric to distinguish different conformations.

We found two residues of interest that seemed to be getting closer and interacting when the clamshell motion was closed and not interacting when the clamshell motion was more open. We used contact maps to find which residues were the closest between EloC and CBF-β. From, the contact map, it seemed like EloC residue 68 and CBF-β residue 15 were more in contact in the open conformation (Supplemental Section: Figure B). There is a difference in the distance between EloC68 and CBF-β15 that is sampled by the different VCBC constructs when comparing 4 sindependent simulations from each(Figure 11).The residues that change distance with PC1, also change their orientations toward each other throughout the simulations.

![Figure 11](image)

Figure 11. Interatomic Distances between residues 310 and 528 which correspond to residue 68 of EloC and residue 15 of CBF-β. (A) Figure of when the two residues are the closer than 4 angstroms and interacting. (B) Figure when the two residues are apart and no longer interacting. Distances between EloC and CDB-b in VCBC truncated (C) and VCBC complex (D).

From these two graphs, we were looking for some correlation between the distance between these residues and the PC1 for VCBC truncated and VCBC. VCBC truncated samples a distance around 7-15 angstroms for a majority of the simulations and the distance can be as large
as 28 angstroms. For VCBC, the favorable distance is at 5 angstroms and the largest distance it get to is around 18 angstroms. The other four simulations of VCBC showed even more of a preference for 5 angstroms and less sampling of simulation time of larger distances (Supplemental Section: Figure C). We were able to observe that without the full-length EloB, the complex sampled a larger distance which agreed with what was seen on the PCA graph where the VCBC truncated structure extended further right along the PC1 eigenvector. This distance calculation can help to simplify how to understand the change and extremes of conformational sampling of both VCBC complexes along PC1. This additional method helps to compare the two VCBC complexes and the differences and similarities between local and global conformational changes.

Discussion

One of the interactions between the EloB tail and Vif that was proposed in multiple articles was not observed in our simulations (Lu et al., 2013 and Bergenson et al., 2010). Binding between the PPLP motif in Vif and EloB was not observed in our simulations. The two regions were very far apart during all of the independent simulations. Even though there were no major changes in global conformational sampling when missing 16 residues, not including the C-terminal end residues does affect local conformation sampling and the increased structure in the tail could possibly allow for more interactions with other proteins if it is more stable for a favorable interaction.

There were small local conformational differences in the atomic fluctuations of the EloB full length tail. This decrease in atomic fluctuations and increase in structure in the tail may have helped to produce a better NMR signal since less flexibility means that there will be less variation in signal so the average would be more accurate and therefore more peaks appeared instead of getting lost in the noise. The VCBC complex and the VCBC truncated complex have overall the same degree of conformational sampling on the global scale. Since both the PCA and distance calculations show that the VCBC truncated has a slightly more open conformation, this lead to a difference in NMR spectra. The larger more open conformation causes an increase in this degree of global conformational sampling that caused more disappearance of peaks due to motions on the intermediate timescale. Therefore, these 16 residues at the end of EloB can be included all future NMR experiments and computational simulations.
Both the decrease in atomic fluctuations and a slightly more open truncated complex could be possible explanations for the improvement in NMR signal from a small primary structural change of EloB in the VCBC structures. This can be done to improve the NMR signal that is produced of differing VCBC complex and the NMR spectra of VCBC-A3F included the full length EloB. So that our simulations can more directly be compared to the experimental data, our lab used the full length EloB to run simulations of the VCBC complex with Cul5 and with APOBEC.
Chapter 4: Simulated Ensemble Completeness and Statistical Analysis

To confidently state the results we saw when comparing the VCBC complexes, we need to know how fully the VCBC state is sampled. If the complex was not able to sample most of the possible conformations during the amount of time the simulations were run for, then the results from the analysis could only be preliminary results. The results would only represent a small portion of the conformational sampling that the VCBC complex could possibly exhibit. Testing if the simulations were run long enough is difficult because there is not definite answer in the scientific community for how long is long enough to run computational simulations. There is no definite way of knowing if your complex has sampled all possible conformations since we do not know every conformational state of all proteins. The way that we can suggest that we have sampled enough time is through looking to see if any new conformational states are being sampled. Looking at the root-mean-squared deviation (RMSD) of a protein complex compares the movement of the protein compare to its starting structure. If there are no large changes in the RMSD, then it is likely that the protein is no longer sampling any new conformations that have not already been sampled.

Also, analysis of the VCBC complex was not completed on all of the simulation time that was run. It was however completed on the same amount of time from each simulation of all the VCBC complex structure. To determine how many trajectories our analysis was to be on, we need to see how long it took any of our simulations to give equilibrated analysis results. Stating that there are differences between each simulation of VCBC alone, with Cul5, and with A3F confidently is only possible with the use of statistical significance tests.

In PCA, the PC eigenvalues represent how much of the correlated motion is capture by the eigenvector. By summing the first 5 PCs, we are calculating a majority of the global correlated motions that are being sampled by the complex. The magnitude of the eigenvalue is related to the total amount of variance caputred in the correlated motions. This can give us an idea of the amount of global motion sampling each protein complex is exhibiting and if any new correlated motions are being introduced later in the simulation. The higher the sum, the greater amount of global correlated motions are occurring during the simulations. These numbers can help identify the total amount of global motions and compare this to other simulations and separated blocks of the same simulation to see if conformational sampling is increasing as the simulation continues.
To test if our simulations of different VCBC structures were significantly different from each other and similar between independent simulations, a Sigma-R plot was applied. The Sigma-r plot is of the standard deviation of intermolecular distances as a function of the distances (Zhou et al., 2015). The standard deviations ($\sigma$) were calculated for all interatomic distances ($r$) and then averaged within intervals with step size $\Delta r$, to determine the average $\sigma$ for the inter-atomic distance interval. Using the following equation

$$\sigma_r = \frac{1}{N} \sum \sigma_{i,j}$$

Where $N$ is the number of inter atomic distances within the interval $r - \frac{\Delta r}{2} < r_{i,j} \leq r + \frac{\Delta r}{2}$. The symbols $r_{i,j}$ stands for the mean interatomic distance and $\sigma_{i,j}$ stands for the mean standard deviation of the distance between atoms I and j over the entire trajectory (Zou et al., 2015). The sigma-r plots illustrate the average range of motion of all the interatomic distances. This measures the average range of motion between all atoms within a protein complex. It is able to differentiate between macromolecules that have different domain structures, molecular weight, and single amino acid replacements. This is a way to differentiate global dynamics in a simplistic manner that is easy to understand through a one-dimensional plot.

To test to see if the Principle Component Analysis of our differing structures were significantly different, we applied the Kolomogorov-Sminoff (KS) test. This test was applied since our data does not have normal distribution and so a regular t-test could not be applied (Wasserman, 2003). This test was used in python after the PCA were split by independent simulations.

**Methods**

*RMSD and Radius of gyration*

Sampling completeness was determined by comparing the four different simulations. RMSD is the measure of the deviation of the molecular internal coordinates deviation from the starting structure which was the crystal structure for some simulations. Analysis of sampling completeness was done using the ptraj module in the Amber package and graphs were made in python (Case et al., 2016). The RMSD of each simulation was compared

$$RMSD = \sqrt{\frac{\sum d_i^2}{n}}$$
where \( d \) was the distance between each of the \( n \) pairs of equivalent atoms from two structures. The RMSD of all the simulations should be similar overtime to show that it is likely that most of the conformational sampling of the protein complex has occurred during the elapsed time. The radius of gyration in each simulation was calculated as well

\[
R_{\text{gyr}} = \sqrt{\frac{\sum m_i r_i^2}{\sum m_i}}
\]

where \( r_i \) was the distance of an atom from the center of mass of the protein and \( m_i \) is the mass of the protein. The running average of both RMSD and radius of gyration was taken. As the average of each simulation converged over time, the distance between the atoms and the movement of the protein relative to its center of mass has no conformational changes. If all the independent simulations average RMSD and radius of gyration converge, this would indicate that enough simulations had been run. This shows that the protein is no longer sampling new conformations that are very different from conformations that have already been sampled.

**Blocks of PCA**

To compare the amount of global conformational sampling between simulations and within a simulation, PCA analysis was completed on every 100-ns blocks of every simulation for VCBC runcated, VCBC, VCBC-Cul5, and VCBC-A3F. The PCs total of the first 5 PCs was taken for every 100-ns block of simulation. The sum of the top five PCs were graphed to show how the sum equilibrated as the simulation time increased using python in-house scripts.

**Sigma-R**

Sigma-r plots were based off of Zhou et al. methods. The distance between each pair of residues was calculated for all trajectories in 100-ns blocks for VCBC, VCBC-Cul5, and VCBC-A3F simulations. A python script was used to calculate the standard deviation of all the distances for each pair of electrons. The distances were binned into integers and the standard deviation for each distance bin was taken. Then the binned distance and average standard deviation was plotted.
**PCA of VCBC and VCBC-Cul5**

PCA was completed using *cpptraj* module of Amber (Case *et al.*, 2016). Cartesian coordinates of the Cα position for each residue in the structured regions of the VCBC and VCBC-Cul5 (excluding flexible tails and loops). The covariance matrix was made from VCBC and VCBC-Cul5 trajectories. Both sets of simulations were projected onto PCs to compare the global motions of the VCBC complex with and without Cul5. The first two PCs were viewed in VMD and characterized with snapshots of the extreme values. They were also plotted together to examine the global conformational sampling with and without Cul5.

**Kolmogorov-Smirnoff test**

For PCA, the trajectories were projected onto the eigenvectors and were averaged for each independent simulation. We compared the 8 simulations of one construct to the 8 simulations of another construct. The p-value between the constructs VCBC, VCBC truncated, VCBC-Cul5, VCBC-A3F PCA were calculated.

**Results**

The RMSD values of the four VCBC with the full-length EloB tail simulations were similar for all simulations (Figure 2). Since this value measures the deviations from the average structure and the RMSD is within the same range, this shows that it is likely each protein complex in the simulations are exhibiting the same range of conformational sampling.

![RMSD plot](image1.png)

![RMSD plot](image2.png)

Figure 12. RMSD of the eight simulations represented by different colors showing how similar the structure is to the crystal structure and the other simulations.
The running average of RMSD and the radius of gyration were used to see if the simulation has been run long enough. The running average of RMSD measures the average distance between the atoms and of the four simulations, it shows that there were some major changes in that over time, as new conformations are being sampled (Figure 13A). In all simulations, there seems to be convergence to a similar value. The running average of the radius of gyration between the four simulations shows convergence as the simulations time increases (Figure 13B). This shows that the simulations have likely been run for adequate amount of time.

![Running Average of RMSD](image1)

![Running Average of Radius of Gyration](image2)

Figure 13. (A) Running Average of RMSD measurements of the 8 VCBC simulations (B) Running Average of the Radius of Gyration of 4 VCBC simulations. Each colored line represents an independent simulation.

To look at the global conformational motions for the construct simulations, the sum of the first five PCs of the PCA completed on 100-ns blocks was graphed to observe any change in this sum as the simulation time increased. For the VCBC simulations, both sets of 4 simulation showed a gradual increase in total eigenvalues in the beginning but leveled off as the simulation time continued. The VCBC-Cul5 simulations had lower total eigenvalues than VCBC simulations and remained relatively the same as the simulations time increased. The VCBC truncated began with very high eigenvalue totals but the last 300 ns of simulation, the PC totals seems to level off. The last 300 ns for most simulations have similar 100-ns block total PC values except for the VCBC-A3F simulations (Figure 14). More simulation time is necessary to see if the total eigenvalues equilibrate as the simulations are continued for the VCBC-A3F construct.
Figure 14. Comparison of summation of first 5 eigenvalues where each line represents 4 independent simulations.

Testing that the simulated ensemble has been sufficiently sampled is important for simulations because you want to be able to sample as many if not all possible conformations that the protein would exhibit experimentally. Therefore, if the simulations were not run long enough, then not all conformations would be seen in the simulation. As the eigenvalues level out, this indicates that there are no new major conformations appearing that would change what the total sum of eigenvalues would be. It was decided that all analysis would be on the last 300 ns of the simulation time run for all the simulations because the sum of the eigenvalues were similar for the last three 100-ns blocks showing there is the same amount of variance captured in the correlated motions.

Sigma-R was completed for each 100-ns block and compared between the different simulated constructs. Sigma-R was run on the last 300 ns for VCBC, VCBC-A3F, and VCBC-Cul5 and graphed for 100 ns of one simulation of each construct. Twelve plots in total were made (Supplemental Section Figure D). We also graphed 100 ns of all simulations for one construct on one plot to compare the simulations to check consistency of simulations each construct (Figure 15). When comparing similar between constructs, VCBC Sigma-R plots show a relatively gradual increase in mean standard deviation as the distance (r) increases. For the VCBC-Cul5 plot there is a large increase and fluctuation of mean standard deviation for all
simulation time beginning around 70 angstroms. Before that the standard deviation is much lower overall compare to VCBC and VCBC-A3F. There is less of a linear relationship between standard deviation and the distance in the VCBC-Cul5 simulations. In the VCBC-A3F simulations, smaller standard deviations compare to VCBC, and there is more of a linear slope for most of the data. From these plots, different trends seem to appear between constructs.

![Sigma-R plots](image)

Figure 15. Sigma-R plot of the last 300ns of simulations split into 100-ns blocks. (A) VCBC 100-ns blocks of simulations from the 300ns into the simulation (B) VCBC-Cul5 100-ns blocks of simulations beginning from 100ns into the simulation (C) VCBC-A3F 100-ns blocks of simulation beginning from 100ns into the simulation

From the Sigma-R plots, simulation time can be evaluated to test if each construct has had enough tie to sample through all of its corresponding free energy landscape and sample all possible conformations. All simulations of VCBC and VCBC-Cul5 have been run long enough that the constructs are behaving similar in each block. The sigma-r plot of last 100-ns block of VCBC-A3F has larger variation from the normal trends seen showing that there are differences.
within the independent simulations and as simulation time increase. There is more variation of the VCBC-A3F construct between the 100-ns blocks and therefore it is likely the simulations need to be run longer.

We performed PCA on VCBC and VCBC-Cul5 combined to form the eigenvectors to compare the global conformational sampling changes when Cul5 is bound to the VCBC complex and the VCBC truncated complex. The PCs were visualized in VMD to show the primary motions that are sampled by the VCBC-Cul5 and VCBC complex (Supplemental Section Figure E). The first PC shows an opposing twist of Vif and CBF-β in relation to EloB and EloC. The second PC illustrates a clamshell opening motion. Both movements depend on the hinge motion of the Vif linker region. On the PC1 vs PC2 plots, there is variation of the constructs global conformational sampling (Figure 17).

![VMD visualizations of PC1 vs PC2](image_url)

Figure 17. Plots of PC1 vs PC2 made through projecting VCBC truncated, VCBC, VCBC-A3F, and VCBC-Cul5 trajectories onto the PCs constructed from VCBC and VCBC-Cul5.
The VCBC truncated and VCBC graphs show similar global conformations being sampling along both PC1 and PC2. However, VCBC-Cul5 samples conformations that are more similar to the crystal structure. Both VCBC constructs sample alternate conformations that are not exhibited by VCBC-Cul5. This also shows that VCBC-Cul5 is less flexible as a complex compare to VCBC since less conformations along the extreme ends of the PCs are being sampled.

Even though visually the difference between VCBC and VCBC-Cul5 PCA are different, we need to test if the difference is statically significant. P-values found for KS test showed that the VCBC and VCBC-Cul5 PCA were statistically significant (Table 3). VCBC and VCBC-Cul5 have significantly different global conformational sampling for both PC1 and PC2 (p-value <0.05). The same is seen for VCBC compare to VCBC-A3F (p-value <0.05). However, between VCBC-Cul5 and VCBC-A3F, only PC1 is significantly different from each other. Also on the VCBC/VCBC-Cul5 eigenvectors, VCBC and VCBC truncated are significantly different (p-value < 0.5). When we looked at the PC plots, visually VCBC truncated and VCBC looks similar. But by looking at the PCs from when VCBC truncated and VCBC form the covariance matrix, PC2 does show that two were not significantly different (P-value >0.05).

Table 3. P-values from Kolmogorov-Smirnoff test on PCA from the VCBC and VCBC-Cul5 combined eigenvectors and from the VCBC truncated and VCBC combined eigenvectors. PC1 and PC2 are different from the two different constructions of eigenvectors and therefore cannot be compared directly.

<table>
<thead>
<tr>
<th>Construction of eigenvectors</th>
<th>VCBC/VCBC-Cul5 eigenvectors</th>
<th>VCBC/VCBC truncated eigenvectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trajectories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCBC/VCBC-Cul5</td>
<td>0.0496</td>
<td>0.0026</td>
</tr>
<tr>
<td>VCBC/VCBC-A3F</td>
<td>0.0378</td>
<td>0.0481</td>
</tr>
<tr>
<td>VCBC-Cul5/VCBC-A3F</td>
<td>0.0125</td>
<td>0.0125</td>
</tr>
<tr>
<td>VCBC/VCBC-truncated</td>
<td>0.1502</td>
<td>0.3788</td>
</tr>
</tbody>
</table>

Overall PC1 captures more of the overall variance of the conformations of VCBC with and without Cul5 but PC2 is more informative of which simulations are similar or different. The p-values for both PCs of VCBC compared to VCBC-Cul5 are lower than 0.05, as expected. There were differences between PC1 and PC2 p-values that were not expected. It is hard to make conclusions about the PCA between the other structures for the eigenvectors were found from
VCBC full length EloB and VCBC-Cul5 simulations together and not from VCBC-A3F or VCBC simulations.

This led us to compare the VCBC and VCBC truncated PCA when the eigenvectors were made from these two constructs. Along PC1 the two constructs are significantly different but PC2 is the same for both. The magnitude of the p-value for PC2 was very high and not close to 0.05 showing that PCA of VCBC truncated and VCBC were not significantly different. This is what we expected to see through the KS test of the PC for PC2 however, for PC1 this was unexpected. The magnitude of the difference in PC1 showed that difference between the two constructs was minor. Overall this shows that both these constructs sample the global motions widely and they have similar global conformational sampling.

Discussion

We can only draw strong conclusions about differences between complexes if there is a statistically reproducible trend. To test if the simulations were run for a long enough time that the constructs had time to sample most if not all possible conformations, RMSD and radius of gyration were recorded for the VCBC simulations. By observing the running average of the two variables, we can conclude that our simulations have been run for long enough for most conformational sampling to occur within our protein complex.

To measure how much of the simulations should be used in our analysis, splitting up the PCA of each construct into 100-ns blocks, we were able to see how long it took for the total PCs to equilibrate. Since the magnitude of the eigenvalue is related to the total amount of variance captured in the correlated motions, the total of the PCs represent most of the global correlated motions for the protein complex. Overall the last 300 ns of each construct’s simulations except for the VCBC-A3F contain relatively the same sum of the first 5 PC values and so this set of simulations were used in all analysis methods. This also suggests that VCBC-A3F simulations need to be run longer to hopefully observe similar PC variance within 100-ns blocks.

From the Sigma-r plots, each of the constructs shows different trends of the mean standard deviation as a function of interatomic distances between the VCBC, VCBC-Cul5, and VCBC-A3F constructs. Overall the plots for simulations of VCBC and VCBC-Cul5 showed similar trend among their independent simulations. This test indicates a difference between VCBC and VCBC-Cul5 independent simulations and that these simulations have been run for
enough time. They also show that the simulations have been However, for VCBC-A3F we observed that more simulation time needs to be run.

Looking at the KS p-values of the first two PCs comparison between VCBC truncated, VCBC, VCBC-Cul5 and VCBC-A3F demonstrated that VCBC and VCBC-Cul5 PCs from the VCBC and VCBC-Cul5 combined eigenvectors were significantly different. When comparing VCBC truncated and VCBC PCs from the corresponding combined eigenvectors, the PC1 was significantly different between the two constructs and PC2 were not significantly different. This method can be best used to show the differences between two constructs PCs but not how similar. The difference between VCBC and VCBC-truncated for PC1 is not a large difference as the magnitude of the PC1 difference is small. This supports what we visually saw with the PC plots that they both sample these global motions widely, although there may be a slight difference in the way they are sampled.
Chapter 5: VCBC with mutation in Vif N-terminus

Once we have a general understanding of the VCBC complex, the next step is to try mutating the VCBC complex to see what residues are crucial for protein-protein interactions or specific conformational states. By making mutations that affect binding of proteins or change the favored conformational state for VCBC-Cul5 binding to APOBEC, we can make progress toward a possible therapeutic target. Cul5 is a protein of the VCBC-Cul5 complex that is crucial for the breakdown of APOBEC. Cul5 is a host cell protein that is part of host cell E3 ubiquitin ligase complexes when the body is not infected with HIV. However, Cul5 is not necessary for Vif to fold and bind to CBF-β, EloB, and EloC. When Cul5 binds to the VCBC complex, it may be able to stabilize it into a conformation that is favorable for binding to APOBEC. Cul5 is also responsible for attaching the ubiquitin protein to APOBEC, which tags the APOBEC for degradation. When Cul5 is not bound to the VCBC complex, the complex is more flexible and samples more conformational states (Nakashima et al., 2015). Other research completed by our lab has been on VCBC with and without Cul5. Simulations have been run and when comparing the PCA, VCBC-Cul5 does not exhibit alternate conformational that differ from the crystal structure like VCBC does. The addition of Cul5 decreases global conformational sampling and understanding what can cause a change in Cul5 interactions with the VCBC complex could affect how we can stop the degradation of APOBEC. Cul5 binds to Vif at helix α3, the following flexible loop of residues 116-131, and the HCCH motif which coordinates the zinc ion (Guo et al., 2014). The conformation of the VCBC complex with Cul5 and without are studied in our lab for insight into the structural stability which will affect how the complex functions. Making a mutation to Vif could perturb the conformations of the VCBC-Cul5 complex which could lead to a method of halting the breakdown of APOBEC.

Characterizing conformational sampling of the VCBC complex is crucial to understanding how Vif recruits substrate proteins and targets APOBEC and can lead toward development of therapeutics to prevent antiviral degradation. A way to test areas of Vif’s structure that are functionally important is to mutate residues in motifs of the proteins. Evans et al. stated that the N-terminal motif of Vif is important for the binding of Cul5 and therefore degradation of APOBEC. In their experiments they performed single alanine replacement of residues 25 through 30 of Vif. The mutations in the 25-30 region reduced the ability for Vif to bind to Cul5 but didn’t affect CBF-β or EloBC binding and the mutations did not affect the structure of the
protein complex according to their circular dichroism spectroscopy and size-exclusion chromatography (Evans et al., 2014). This could lead to a possible target for HIV pharmaceutical designs because if Vif and Cul5 do not bind then the APOBEC proteins cannot be ubiquitinated. The Vif mutants L24A and V25A both inhibit Vif’s ability to degrade A3G and A3F (Evans et al., 2014). The V25A mutation results in a 90-fold lower binding affinity to Cul5. Lowering the binding affinity of Vif and Cul5 could help decrease the probability of the whole VCBC-Cul5 complex binding together and therefore APOBEC proteins would not be tagged with ubiquitin to be degraded. To look at the effects of the V25A mutation on the VCBC complex (mutant VCBC), we ran MD to test the effect of a single point mutation at the N-terminus of Vif on the C-terminus of Vif, which is where Cul5 and EloBC bind to Vif.

Figure 18. Vif with mutation of Valine to an alanine at residue 25 which is highlight in magenta

Methods

MD Simulations

MD simulations were run in similar manner as the VCB118C complex simulations. Four independent simulations were run for 100ns each to allow for preliminary analysis. Residue 25 in Vif, a valine, was mutated to an alanine by editing the PDB file. Two carbons and the bonded hydrogens were removed from the side chain of the valine to make it into an alanine. Tleap was run to check the PDB file format, and optimize the location of the atoms of the new alanine residue. The two TER were taken out of Vif so there were no breaks in the Vif sequence. A second tleap was run subsequently for Amber to reformat the PDB file, add water and ions to the simulations, put in the water box size for the protein, and make the parameter topology and input coordinate files.

Minimization was completed twice; once, with restraints on the protein set to 500kcal/mol with a 1000-steps steep decent followed by a 1000-steps of gradient minimization
and the second time was without the restrained protein with restraints only set to 10 kcal/mol to allow the protein to adjust to its minimization state without having the water molecules disturb the protein movements. The box temp increases from 0 to 300 K completed in 20 ps with the protein restraints set to 10.0 kcal/mol. Then, two equilibration steps are completed to equilibrate the density of the simulations. The first equilibration is done with the protein restrained at 1.0 kcal/mol over 20ps. The simulations were run at 300 K. Traditional MD was used and a total of 300 ns was achieved for each of the 4 independent simulations for a total of 1200 ns of simulation.

**Analysis Methods**

Analysis was completed on 300 ns each for 4 independent simulations. RMSD and radius of gyration were used to determine sample completeness (Chapter 4 methods). Atomic fluctuations were calculated to look at local conformational perturbation to residue flexibility (Chapter 3 methods). Principle Component Analysis was also used to look at any changes in global conformational sampling (Chapter 3 methods).

**PCA of VCBC and VCBC-Cul5 with Mutant VCBC**

PCA was completed using *cpptraj* module of Amber (Case et al., 2016). Cartesian coordinates of the Cα position for each residue in the structured regions of the VCBC and VCBC-Cul5 (excluding flexible tails and loops) were used to make the covariance matrix. Mutant VCBC trajectories were projected onto PCs to compare the global motions of the VCBC complex with and without Cul5. The first two PCs were viewed in VMD and characterized with snapshots of the extreme values. They were also plotted together to examine the global conformational sampling with and without Cul5.

**Results and Discussion**

Since each of the four independent simulations have run for only 300 ns each, all of the data is preliminary results. RMSD is the deviation of the molecular internal coordinates from the starting structure. The RMSD and Radius of gyration of all the simulations should be similar overtime to show that it is likely that most of the conformational sampling of the protein complex has occurred during the elapsed time. The RMSD and radius of gyration values of the
four mutant VCBC simulations were similar for all simulations (Figure 19 and 20). Since RMSD measures the deviations from the average structure and appears within the same range between simulations, it is likely each protein complex in the simulations are exhibiting the same range of conformational sampling.

Plotting the running average of these two variables is another way to test if the simulations have been run to allow for complete conformational sampling. The running average of RMSD of the four simulations converging to a similar value illustrates that there were some major changes but over time less new conformations are being sampled (Figure 21A). The running average of the radius of gyration between the four simulations shows less convergence than the running average of the RMSD (Figure 21B).

More simulations need to be run as the running average of the radius of gyration from each independent simulation has not converged. Eventually, all the simulations will be run for
600 ns as this amount of time is seen to allow the VCBC complex to completely sample its free energy landscape (Chapter 4). Also, we saw that 300 ns was used to complete all methods of analysis in Chapter 4. A total of 600 ns was necessary for the VCBC complex to have leveling off of running average of RMSD and radius of gyration, so to compare the results of mutant VCBC to VCBC directly, a total of 600 ns for each mutant VCBC independent simulations needs to be run and then of the 600 ns, analysis will need to be run on the last 300 ns of simulations.

Preliminary PCA analysis of VCBC mutant was projected onto a plot with VCBC and VCBC-Cul5 PC1 and PC2 plot (Figure 22). From the plot, the VCBC mutant appears to be sampling conformations more like VCBC-Cul5 conformations than VCBC.

![Figure 22](image.png)

Figure 22. Mutant VCBC, VCBC, and VCBC-Cul5 construct trajectories projected onto PC1 and PC2 from VCBC and VCBC-Cul5 combined eigenvectors. Mutant VCBC is sampling a small range of conformations along both PC1 and PC2 compared to VCBC.

Over time, the trajectories may move more and sample more alternate conformations more like VCBC but this was seen very little in the first 300 ns. Because this is only analysis on the first 300 ns, more simulation time is needed to make sure all conformational sampling can occur to see if as the simulations continue, a wider range of conformational sampling is exhibited. The trajectory of mutant VCBC may continue to sample closer to the crystal structure.
as the first 300 ns from 4 independent simulations appear to be doing which would imply that mutant VCBC is sampling conformations more similar to VCBC-Cul5.

We compare atomic fluctuations between mutant VCBC, VCBC, and VCBC-Cul5 to observe local conformational changes in different structures. Overall, the mutant VCBC Vif protein had similar atomic fluctuations to the Vif protein in the VCBC-Cul5 simulations (Figure 22). The mutation was at residue 25 and the atomic fluctuations in the 20-30 region did not differ very much from the VCBC atomic fluctuations. This indicates that this mutation did not have a local conformational effect on the Vif protein in the initial simulations.

![Figure 23. Atomic Fluctuations of Vif comparing between three constructs.](image)

![Figure 24. Atomic Fluctuations of EloC comparing three constructs](image)

Cul5 binds to EloC, and as we have seen previously, when Cul5 binds to EloC there is a large drop in atomic fluctuations of residues 30-40. There is a smaller peak of atomic fluctuations were in this region for the mutant VCBC simulations compare to the wild-type
VCBC simulations (Figure 24). All other regions of EloC remain relatively similar to both VCBC and VCBC-Cul5 fluctuations. However, since the mutant VCBC simulations have only been run for 200ns, this could be due to the atoms not fluctuating as much as possible due to the time limit. As more simulations are run, it is likely the atomic fluctuations overall will increase as the atoms have more time to fluctuate more. If there is no change in Vif atomic fluctuations, then the atomic fluctuations are more similar to the VCBC-Cul5 atomic fluctuations. If in EloC the atomic fluctuations do not increase, then in the mutant VCBC atomic fluctuations, the region where Cul5 binds to EloC could potentially be affected by a single point mutation in Vif.

**Discussion**

Overall, more simulation time needs to be run to confirm all the preliminary results between VCBC and mutant VCBC for running average of RMSD and radius of gyration, PCA, and atomic fluctuations. From the running average of radius of gyration, more simulations time needs to be run to allow for complete sampling of the free energy landscape. Since the VCBC simulations were all run for 600 ns, this amount of simulation time will likely be a valid amount of time to run simulations (Chapter 4 results). Looking at the preliminary PCA, the mutant VCBC simulations appear to be sampling conformations close to the crystal structure and similar to VCBC-Cul5 simulations. The experimental data on this mutant showed that the N-terminus of Vif was important for Cul5 binding to the VCBC complex (Evans et al., 2014) Preliminary data does indicate that the mutation we made of changing residues 25 from a valine to an alanine may change the global conformational sampling of the VCBC and affects atomic fluctuations in EloC. This mutation lowers the experimental binding affinity of Cul5 for EloC, and our preliminary results show that the atomic fluctuations of residues 30-40 of EloC, where Cul5 binds, is lowered when VCBC is mutated. This decrease in atomic fluctuations is similar to what occurs when VCBC is bound to Cul5 in the VCBC-Cul5 atomic fluctuations.

However, again, this is all preliminary data and the mutant VCBC complex has not been able to completely sample its free energy landscape, and therefore both the PCA and atomic fluctuation results are likely to change. We plan to continue running our four independent simulations as well as run more independent simulations. Future studies will look at dihedral angles and hydrogen bond analysis to identify local conformational changes. We would also like to look at other single point mutations in the N-terminus of Vif in the VCBC complex to learn
more about the allosteric effect the N-terminus of Vif has on the C-terminus of Vif and other bound proteins like EloC. Mutations help us identify which residues are important for the conformational changes that we observe in the VCBC and VCBC-Cul5 constructs. This could help us find a Vif point mutation that could disrupt Cul5 binding so that the E3 ubiquitin ligase cannot be completed and therefore cannot breakdown APOBEC proteins.
Chapter 6: Conclusion

One goal of our research and analysis of simulations was to compare the VCBC and VCBC-Cul5 constructs. From my work and the work of the rest of the lab, we conclude that VCBC samples alternate conformations and it samples less with Cul5 bound. To support this, we first looked at experimental data to find differences between VCBC and VCBC-A3F since NMR spectroscopy could not be completed on VCBC-Cul5. From those results simulations were run on VCBC to compare analysis to VCBC-truncated and saw that there were no changes in degree of global motion sampling and little local conformational changes. This allows us confidently to use the full EloB sequence with the Guo et al. VCBC truncated structure even though it was not part of the same crystal structure because it did not appear to drastically affect the global correlated motions. We used VCBC to compare conformational sampling of VCBC-Cul5 instead of the VCBC truncated construct. To confidently draw conclusions from our results, we ran statistical analysis tests and tests for complete conformational sampling for each construct. Based on these results, both VCBC and VCBC-Cul5 simulations have been run for valid lengths of time but VCBC-A3F simulations need to be run longer. Since we know that VCBC-Cul5 samples fewer alternate conformations than VCBC, we next tried to mutate Vif to alter the conformational sampling of VCBC. Preliminary results from PCA show that mutant VCBC exhibits similar conformations to VCBC-Cul5 and the atomic fluctuations of VCBC-truncated are lower than VCBC. However, this all could be due to the mutant VCBC complex not having enough simulation time to sample all conformations in its free energy landscape. Overall we see that VCBC without Cul5 is more flexible and Cul5 helps to stabilize the VCBC complex to allow for the binding and breakdown of APOBEC proteins.

NMR spectra of VCBC truncated, VCBC, and VCBC-A3F were compared. VCBC-A3F produces a clear signal with more methyl peaks appearing due to the decrease in exchange occurring on the intermediate time scale. We also tried to identify the different peaks that appeared in our spectra and what type of residue they were. Using a python package called NMRglue, I labeled peaks by protein and residues type after generalizing what type of residue would appear in which area of the spectra. I also compared the VCBC truncated NMR spectrum to the VCBC NMR spectrum and saw that the VCBC spectrum has a clearer signal and produced better NMR results. Since VCBC-A3F compared to VCBC and VCBC compared to VCBC truncated each exhibited less conformational sampling on the intermediate time scale than the
latter, it could be that the additional protein A3F and additional residues to EloB both helped stabilized the protein complex.

Our NMR results lead us to run MD simulations of the VCBC truncated structure, and VCBC. We ran VCBC simulations to analyze why the NMR experimental data indicate that small primary structural change could lead to changes in exchange on the intermediate timescale. The atomic fluctuations showed that the full-length C-terminus tail has lower flexibility and an increase in secondary structure making it more stable. This added stability could have helped the NMR. PCA was used to observe global conformational sampling and does so through separating the motions of the protein complex into correlated global motions (essential motions) and unessential motions. Each principle component represents one global motion and PC1 showed a clamshell opening and PCC2 showed an opposing twist. By projecting the protein complex trajectories onto the eigenvectors, we can visualize how the protein complex is moving and the extremes of each PC. Comparing VCBC truncated and VCBC, they exhibit about the same degree of conformational sampling along PC1 and PC2. VCBC truncated sampled a slightly more open clamshell motion than VCBC. Then when viewing the distance that changed with the clamshell motion, we see that VCBC truncated samples larger distances during more of the simulation than VCBC. This open clamshell conformation increases conformational sampling occurs when adding 16 residues to the end of EloB in the VCBC complex and causes an increase in exchange on the intermediate timescale causing a decrease in peaks on the NMR spectra. From these results, the full-length EloB was included within the VCBC-Cul5 and VCBC-A3F constructs allowing us to have a more direct comparison to the NMR spectroscopy results since the constructs contain the same amino acid sequences.

We used multiple statistical analysis tests to see if our results were significant and if our simulations were run for enough time. The running average of both RMSD and radius of gyration are used to test if our simulations of VCBC are run for enough time for the constructs to sample completely through their free energy landscape. As the average of each simulation converged over time for the VCBC complex, the distance between the atoms and the movement of the protein relative to its center of mass show when they are no longer sampling new local conformations. We used PCA on VCBC truncated, VCBC, VCBC-A3F, and VCBC-Cul5 100-ns blocks to observe how the sum of the top 5 PCs changes as the simulations run. The last 300 ns of all simulations gave similar variance of the global motions for VCBC, and VCBC-Cul5 and
therefore the last 300 ns will be used for all analysis methods. However, the trends of VCBC-A3F 100-ns PCs totals did not remain level showing that longer simulations would need to be run since there was still an increase in variance of global motions occurring. Sigma-R plots made to show that the VCBC and VCBC-Cul5 were run for enough time as well; however, the VCBC-A3F simulations need to be run longer as the trends between each 100-ns block were not that similar to each other. Each construct shows different trends of mean standard deviations of distances as the interatomic distances increase. To test if our PCA values were significant, we used the Kolmogorov-Smirnoff p-values to show that the projection of VCBC and VCBC-Cul5 on their combined eigenvectors is significantly different. When VCBC and VCBC truncated trajectories are projected on their combined eigenvectors it was seen that they both exhibit a wide range of global conformations but likely not the exact same conformations.

Mutant VCBC contains a single point mutation in Vif at residue 25. Simulations of the valine residue mutated to an alanine are run and through running averages of radius of gyration, we conclude that the simulations need to be run for longer since not all possible conformations had been sampled yet from the free energy landscape. Preliminary results show through PCA and atomic fluctuations that the VCBC mutant sampled similar degrees of global correlated motions and local atomic flexibility to VCBC-Cul5. However, more independent simulations and continuous simulations need to be completed before any definitive results can be concluded. This mutation location has been seen to affect Cul5 binding and primary results do suggest that this mutation is affecting the global and local conformations of the VCBC complex.

To continue our comparison between VCBC and VCBC-Cul5, our next steps are to continue creating new constructs with mutations to see if the stable VCBC-Cul5 could be made to sample more alternate conformations like VCBC or confine the complex to an unfavorable conformation that disallows APOBEC protein from binding to Vif. Simulations are being run of VCBC-Cul5 with the V25A point mutation and other mutations in this C-terminus region of Vif will be run. Keeping Cul5 from binding to VCBC will keep the complex inactive and stop the binding of APOBEC that leads to its degradation and ubiquitination. This can help lead to a possible therapeutic target to stop the breakdown of APOBEC, which in turn would prevent HIV from spreading through the body.
Acknowledgements

I would like to thank my PI Prof. Ball for her guidance through my research and helping me write and edit this thesis. I would also like to thank my lab mates on the Vif project, Acadia Connor, Elise Tierney, and Sampriti Thapa for the research they completed in this paper. I would also like to thank everyone else in the lab for their support and thoughtful comments.

References


Supplemental Information

Figure A. Evecs of individual VCBC and VCBC full length
VCBC eigenvectors:

VCB118C eigenvectors:

VCBC and VCBC118C eigenvectors:

Py script for creating histograms of the PC for VCBC, VCBC-full length eigenvectors

def histogram1(pdb, pdb_name, snapshots, snapshots_name, pc, savepath):  # give where the file is located and where to save the files
    pdb_pca = pd.read_table(pdb, sep = '\s+')  # read in VCBC simulation onto VCB118C evecs
    snapshots_pca = pd.read_table(snapshots, sep = '\s+')  # read in VCB118C simulation onto VCB118C evecs
    binedges = np.arange(snapshots_pca['Mode'+pc].min(), snapshots_pca['Mode'+pc].max(), 10)
    hist1 = plt.figure()
    plt.hist(snapshots_pca['Mode'+pc], bins = binedges, normed = True, color = "red", alpha = 0.5, label = snapshots_name)  # plot histogram
    plt.axvline(x= pdb_pca['Mode'+pc][0], color='g', linewidth=2, label = pdb_name)  # plot line onto the histogram
    plt.ylabel ('Fraction of structure from simulation')
    plt.xlabel ('PC'+pc)
    plt.legend (loc = 'upper right')
    plt.savefig(savepath)

def histogram2(pdb, pdb_name, snapshots1, snapshots1_name, snapshots2, snapshots2_name, pc, savepath):  # give where the file is located and where to save the files
    pdb_pca = pd.read_table(pdb, sep = '\s+')  # read in file name and VCBC_pca becomes the dataframe
snaps1_pca = pd.read_table(snapshots1, sep = "\s+")
snaps2_pca = pd.read_table(snapshots2, sep = "\s+")
if snaps1_pca["Mode"+pc].min() < snaps2_pca["Mode"+pc].min():
    min_val = snaps1_pca["Mode"+pc].min()
else:
    min_val = snaps2_pca["Mode"+pc].min()
if snaps1_pca["Mode"+pc].max() > snaps2_pca["Mode"+pc].max():
    max_val = snaps1_pca["Mode"+pc].max()
else:
    max_val = snaps2_pca["Mode"+pc].max()
binedges = np.arange(min_val, max_val, 10)
hist2 = plt.figure()
plt.hist(snapshots1_pca["Mode"+pc], bins = binedges, normed = True, color = "blue", alpha = 0.5, label = snapshots1_name)
plt.hist(snapshots2_pca["Mode"+pc], bins = binedges, normed = True, color = "red", alpha = 0.5, label = snapshots2_name)
plt.axvline(x = pdb_pca["Mode"+pc][0], color = 'g', linewidth = 2, label = pdb_name)
plt.ylabel ('Fraction of total eigenvalues')
plt.xlabel ('PC2')
plt.legend (loc = 'upper left')
plt.savefig(savepath)
def histogram3(pdb1, pdb1_name, pdb2, pdb2_name, snapshots1, snapshots1_name, snapshots2, snapshots2_name, pc, savepath):
    # give where the file is located and where to save the files
    pdb1_pca = pd.read_table(pdb1, sep = "\s+")  # read in file name and VCBC_pca becomes the dataframe
    pdb2_pca = pd.read_table(pdb2, sep = "\s+")
    snaps1_pca = pd.read_table(snapshots1, sep = "\s+")
    snaps2_pca = pd.read_table(snapshots2, sep = "\s+")
    if snaps1_pca["Mode"+pc].min() < snaps2_pca["Mode"+pc].min():
        min_val = snaps1_pca["Mode"+pc].min()
    else:
        min_val = snaps2_pca["Mode"+pc].min()
    if snaps1_pca["Mode"+pc].max() > snaps2_pca["Mode"+pc].max():
        max_val = snaps1_pca["Mode"+pc].max()
    else:
        max_val = snaps2_pca["Mode"+pc].max()
    binedges = np.arange(min_val, max_val, 10)
    hist2 = plt.figure()
    plt.hist(snapshots1_pca["Mode"+pc], bins = binedges, normed = True, color = "blue", alpha = 0.5, label = snapshots1_name)
    plt.hist(snapshots2_pca["Mode"+pc], bins = binedges, normed = True, color = "red", alpha = 0.5, label = snapshots2_name)
    plt.axvline(x = pdb1_pca["Mode"+pc][0], color = 'g', linewidth = 2, label = pdb1_name)
    plt.axvline(x = pdb2_pca["Mode"+pc][0], color = 'g', linewidth = 2, label = pdb2_name)
    plt.ylabel ('Fraction of total eigenvalues')
    plt.xlabel ('PC2')
    plt.legend (loc = 'upper left')
    plt.savefig(savepath)
if __name__ == '__main__':  # need to run from command line
    import pandas as pd  # import needed modules
    import sys
    import numpy as np
    import matplotlib.pyplot as plt
    plt.switch_backend('agg')  # something about not displaying in command line
    if len(sys.argv) == 7:
        histogram1(sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4], sys.argv[5], sys.argv[6])  # set up the parameters for what will be inputed in command line
elif len(sys.argv) == 9:
    histogram2(sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4], sys.argv[5], sys.argv[6], sys.argv[7], sys.argv[8])
elif len(sys.argv) == 11:
    histogram2(sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4], sys.argv[5], sys.argv[6], sys.argv[7], sys.argv[8], sys.argv[9], sys.argv[10])

Figure B. Contact map between CBF-β residues and EloC residues

Figure C. VCBC Simulations 5-8 distance between residues

Py script for Sigma-R

def sigma_r(filenameA, filenameB, filenameC, savepath):
    maximum = 555
    minimum = 1
    r_matrixA = [[0 for i in range(int(maximum)-int(minimum)+1)] for i in range(int(maximum) - int(minimum)+1)]
    #makes the initial matrix of zeros for all the residue-residue contacts
    sigma_matrixA = [[0 for i in range(int(maximum)-int(minimum))+1)] for i in range(int(maximum) - int(minimum)+1)]
    #makes the initial matrix of zeros for all the residue-residue contacts
    r_matrixB = [[0 for i in range(int(maximum)-int(minimum))+1)] for i in range(int(maximum) - int(minimum)+1)]
    sigma_matrixB = [[0 for i in range(int(maximum)-int(minimum))+1)] for i in range(int(maximum) - int(minimum)+1)]
    r_matrixC = [[0 for i in range(int(maximum)-int(minimum))+1)] for i in range(int(maximum) - int(minimum)+1)]
    sigma_matrixC = [[0 for i in range(int(maximum)-int(minimum))+1)] for i in range(int(maximum) - int(minimum)+1)]

    for x in range(int(minimum),int(maximum)):
        for x2 in range(x+1, int(maximum)+1):
            A = pd.read_table(filenameA + str(x) + '_' + str(x2) + '.dat', sep='\s+') #reads in filenameA that
A.columns = ['#Frame', 'Dis']  #renames the column names
r_matrixA[r: int](minimum)][x-int](minimum)] = A['Dis'].mean()
sigma_matrixA[r: int](minimum)][x-int](minimum)] = A['Dis'].std()

#make the matrix into a dataframe; make matrix into data series first; make 2D data into 1D
r_dataframeA= pd.DataFrame(r_matrixA)
sigma_dataframeA= pd.DataFrame(sigma_matrixA)
sig_rA = pd.Series(r_matrixA['Dis'])

for x in range(int(minimum), int(maximum)):
    B = pd.read_table(filenameB + str(x) + '_' + str(x2) + '.dat', sep='\s+')
    B.columns = ['#Frame', 'Dis']
    r_matrixB[r: int](minimum)][x-int](minimum)] = B['Dis'].mean()
sigma_matrixB[r: int](minimum)][x-int](minimum)] = B['Dis'].std()

r_dataframeB= pd.DataFrame(r_matrixB)
sigma_dataframeB= pd.DataFrame(sigma_matrixB)
sig_rB = pd.Series(r_matrixB['Dis'])

for x in range(int(minimum), int(maximum)):
    C = pd.read_table(filenameC + str(x) + '_' + str(x2) + '.dat', sep='\s+')
    C.columns = ['#Frame', 'Dis']
    r_matrixC[r: int](minimum)][x-int](minimum)] = C['Dis'].mean()
sigma_matrixC[r: int](minimum)][x-int](minimum)] = C['Dis'].std()

r_dataframeC= pd.DataFrame(r_matrixC)
sigma_dataframeC= pd.DataFrame(sigma_matrixC)
sig_rC = pd.Series(r_matrixC['Dis'])

plt.plot(graphdataA, label = 'VCBC')
plt.plot(graphdataB, label = 'VCBC-Cul5')
plt.plot(graphdataC, label = 'VCBC-A3F')
plt.ylabel('r', fontsize=14)
plt.xlabel('Standard deviation', fontsize = 14)
plt.legend(loc = 'upper left', fontsize = 14)
plt.savefig(savepath)

if __name__ == '__main__':
    import matplotlib.pyplot as plt
import pandas as pd
import sys
import numpy as np
plt.switch_backend('agg')
sigma_r(sys.argv[1],sys.argv[2],sys.argv[3],sys.argv[4])

Figure D. Plots for Sigma-R of individual constructs compare to each other

Sim1

Sim2

Sim 3

Sim 4
Figure E: PCA from the Covariance Matrix made of VCBC and VCBC-Cul5 trajectories. Illustrations of both PC1 and PC2 motions.

PC1:
Swivel of Vif and CBF-β with respect to EloBC

PC2:
Clamshell opening with slight opposing twist