UNIVERSITY OF CALIFORNIA SAN DIEGO

Markov State Analyses of Influenza Neuraminidase

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Chemistry

by

Simon Park

Committee in charge:

Professor Romme Amaro, Chair
Professor Kamil Godula
Professor Andy McCammon

2022

The Thesis of Simon Park is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This thesis is dedicated to my Father and Mother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CK: Chapman-Kolmogorov

BMM: Bayesian Markov Model

BHMM: Bayesian Hidden Markov Model

HA: Hemmagglutinin

HMM: Hidden Markov Model

ITS: Implied Timescale

MSM: Markov State Models

MFPT: Mean First Passage Timescale

NA: Neuraminidase

MD: Molecular Dynamics

VMD: Visual Molecular Dynamics

ACKNOWLEDGEMENTS

I would like to acknowledge the wonderful people who have guided me through my master’s project including Professor Rommie Amaro who graciously accepted me into this lab, as well as my mentors Dr. Lorenzo Casalino and Christian Seitz who walked me though computation chemistry techniques and were extremely patient with me and walked me through in detail with the tools necessary to delve me into the world of computational chemistry.

When I first joined, I did have a rudimentary background with basic flux balance techniques but as someone who did undergraduate studies in neuroscience physiology, a lot of the advanced techniques I learned later on were completely new to me. There was a rigorous learning process and it took a lot of patience and explaining for me to grasp the theory behind the techniques used.

As I learned more I was greatly aided by the techniques I have learned through my courses as well. Though I did not utilize these techniques directly, the methods that I employed were very useful in understanding the theory behind the work I ended up doing as part of my research. All of the members of my lab were very helpful and welcoming, and learning so much would have not been possible without such a great support group.

ABSTRACT OF THE THESIS

Markov State Analyses of influenza Neuraminidase

by

Simon Park

Master of Science in Chemistry

University of California San Diego, 2022

Professor Rommie Amaro, Chair

Influenza, despite recently being pushed off to the wayside in the scope of public attention, has placed itself front and center as a recurring potential epidemic. Influenza kills a significant amount of people yearly, and in certain historic cases caused a worldwide epidemic.

Previous studies of influenza in the Amaro lab brought light to the function of neuraminidase especially in certain binding sites such as the 450 and 150-loops. The full scale molecular dynamic simulations of mesoscale viral models provided an enormous amount of data that could be exploited to analyze the biological and structural properties of influenza

glycoproteins, namely neuraminidase and hemagglutinin. The multiple copies of neuraminidase contained within these whole-virion models allowed for the application of Markov state models to study the dynamics of the neuraminidase 150 and 430-loops. Moreover, our results show that small changes in glycosylation lead to significant differences in the kinetics and morphology of the open and closed conformation of the neuraminidase 150-loop. However, the number of states between the models did not change, which consolidated our current understanding of the 150 – loop dynamics having a clear binary open-closed state.

In this thesis, Chapter 1 provides background information about influenza and presents the principles of Markov state model theory. Chapter 2 focuses on the workflow pertinent to Markov State Model building, detailing the process that led to choosing, tweaking and assessing the optimal MSM parameters and settings. Chapter 3 is comprised of the results of the MSM analysis of the neuraminidase 150-loop dynamics. Chapter 4 is an in-depth discussion of the results including future goals and directions.

Chapter 1

Introduction: Influenza and Neuraminidase

1.1 Influenza

Influenza remains one of the most threatening viruses present to this day. Despite the advent of the SARS-CoV-2 pandemic as the prototypical dangerous virus to plague the world population, influenza has continued to cause great amounts of deaths every year. Influenza peaked with particularly virulent HINI strains in the 1918, 1957, 1968, and 2009 pandemics, which resulted in millions of deaths worldwide, with the 1918 flu pandemic taking an estimated 50 million people worldwide.[1] Influenza is estimated by the WHO to cause 290,000-650,000 deaths every year due to respiratory complications alone.[2]

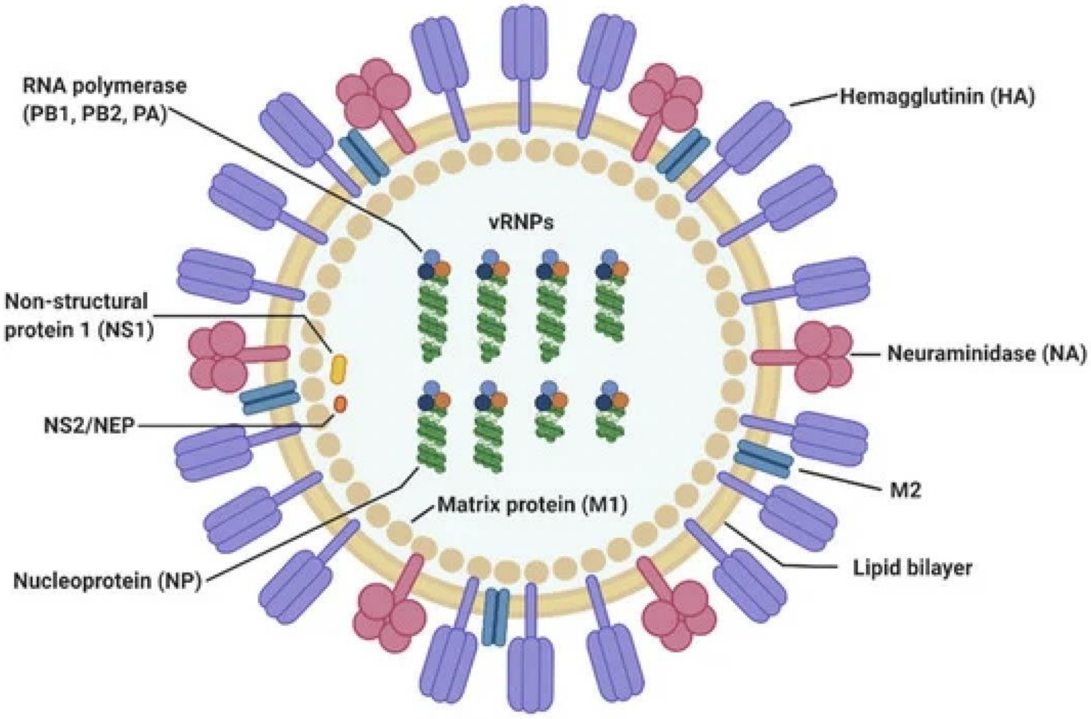

Figure 1.1: Influenza structure. Image from Jung H.E. et al., *Viruses* 2020,*12*(5), 504.[3]

Influenza has three distinct major types, A, B, and C, which can be differentiated both structurally and based on their biological capabilities. Influenza A can be transmitted from animals to humans and can cause pandemics, whereas influenza B is transmitted exclusively from human to human. Type A influenza has HA, NA, and M2 as the membrane proteins, M1 Matrix protein situated beneath the membrane, a ribonucleoprotein core, and the NEP/NS2 protein. Focusing mostly on the external proteins, type B influenza has HA, NA, NB, and BM2 as membrane proteins. Type C has a largely different structure and is not the cause of major epidemics like type A and B.[4] Type A is of great interest due to the fact that its zoonotic behavior and propensity to cause major pandemics. Zoonosis essentially highlights the viral ability to infect multiple different species of animals, and this biological ability increases the potency of the virus in several manners. Being able to be spread by animals increases simple infectivity as it can move through more vectors in a general area. It also means that vaccines have less efficacy in the grand scheme of things, because depending on the possible host animals, vaccinating host organisms can be extremely difficult logistically or downright impossible. On top of this, genetic shift is expedited by concurrent infections in carrier organisms, greatly enhancing the potency of viral evolution.

1.2 Neuraminidase

Of the influenza membrane proteins listed above, my research was focused primarily on the neuraminidase glycoprotein. Neuraminidase is particularly involved during the final state of the viral cycle, modulating release of the virus from the infected cell by cleaving the terminal sialic acid moieties presented by host cell-linked glycan. Tamiflu, or Oseltamivir, is one of the only non-vaccine treatments for influenza and their mechanism of action involves targeting a catalytic pocket present in the neuraminidase head and lined by the 150-loop and the 430-loop.[6]

Being able to understand the dynamics of neuraminidase 150-loop in particular can lead to more solutions to inhibit influenza activity without having to participate in the ongoing arms race of viral forecasting, although key mutations do occur within neuraminidase, the overall function and behavior remain somewhat conserved, as will be demonstrated throughout this thesis.

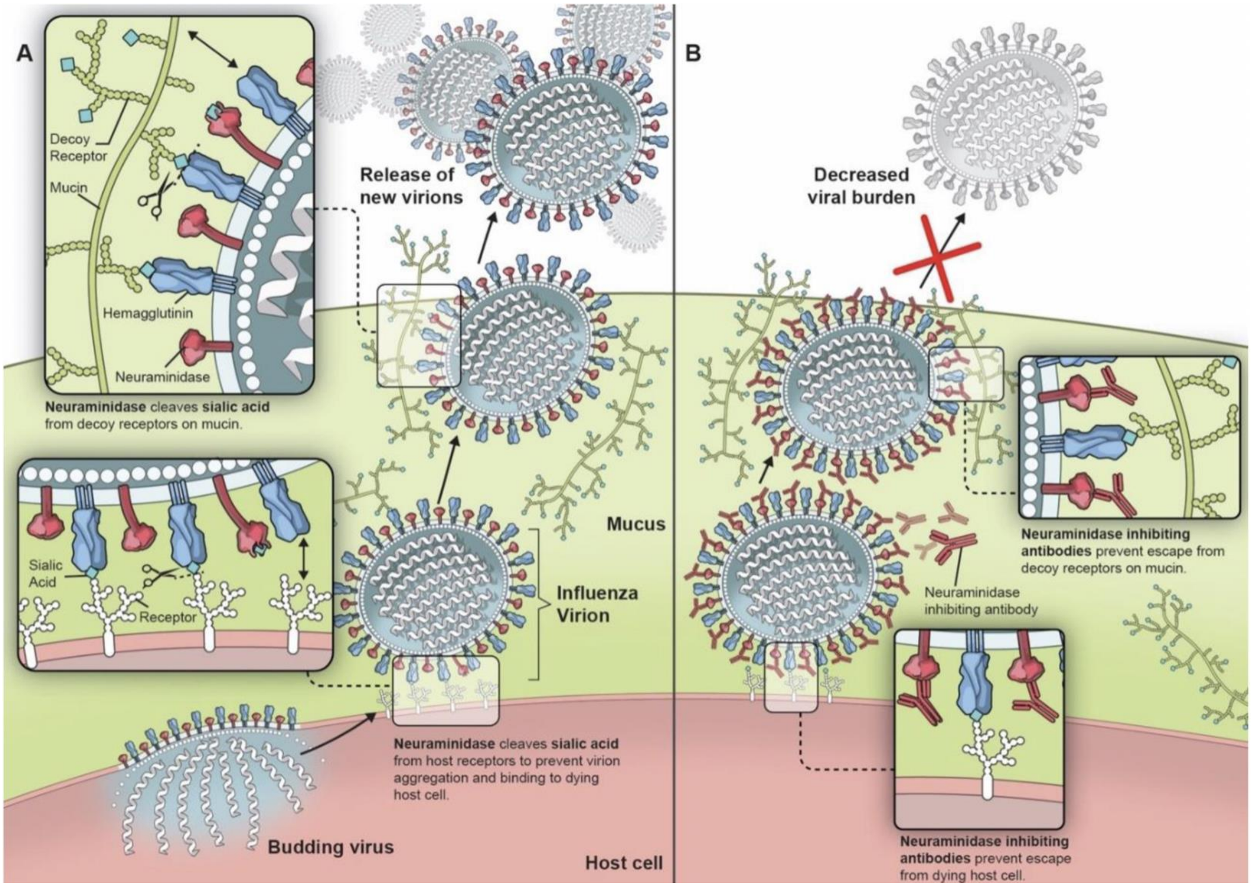
A
Decoy Receptor
Mucin
Hemagglutinin
Neuraminidase
Neuraminidase cleaves sialic acid from decoy receptors on mucin.
Release of new virions
Sialic Acid
Receptor
Mucus
Influenza Virion
Neuraminidase cleaves sialic acid from host receptors to prevent virion aggregation and binding to dying host cell.
Budding virus
Host cell
B
Decreased viral burden
Neuraminidase inhibiting antibodies prevent escape from decoy receptors on mucin.
Neuraminidase inhibiting antibody
Neuraminidase inhibiting antibodies prevent escape from dying host cell.

Figure 1.2: Influenza and neuraminidase activity in viral function. Image from Giurgea L.T. et al., Vaccines 2020, 8(3),409.[7]

In this thesis, we focused on the A/swine/Shandong /N1/2009(H1N1) strain and on the more recently emerged A/45/Michigan/2015(H1N1) strain. Key mutations at certain sites within the neuraminidase head set these two strains apart. Of note, a mutation appears on residue 432, where lysine in the 2009 strain becomes glutamic acid. This reverses a positive charge into negative within the 430-loop, one of the two major functional loops lining the active site in the neuramindase head. In addition, for the 2009 influenza strain, analyzed trajectories were based both on the glycosylated and unglycosylated whole virion models.[8]

1.3 150 and 430-loops

Neuraminidase has two loops associated with infectivity that have been the focus of many studies in order to generate treatments. Since the current approach to flu vaccines has been to use predictive algorithms in order to assess what future strains may look like, understanding the kinetics and functions of these loops of interest offer much in the realm of progress in combating influenza pandemics.[9] The 150-loop spans from residues 147 to 152 whereas the 430-loop spans residues 430 to 433. The 430-loop has been shown to be involved in intermittent salt bridge formation with the 150-loop. Both loops line the active site pocket cavity and therefore are implicated in the binding process of sialic acid substrate and drug inhibitors.[10] These two loops can be very close to each other, defining a closed conformation, or move far away from each other, leading to an open conformation. This makes the data sets simpler to process with two states to analyze. That being said, it is important to check for other states to any system to build the best model. Understanding the dynamics of these loops is crucial to understanding how to combat influenza.

1.4 Data Set

The data set of the system is based on mesoscale molecular dynamics(MD) simulations of the whole influenza viral envelope, including 30 neuraminidase tetramers, 236 hemagglutinin trimers, and 11 M2 proton channels. The first data set is based on ~120 ns of MD simulations of the unglycosylated 2009 strain performed by Durrant et al.[11] The second and third data sets are based on ~442 ns MD simulations of the glycosylated 2009 strain and ~425 ns MD simulation of the glycosylated 2015 strain respectively.[12] Considering the 30 neuraminidase tetramers present in the full-scale virion modes, an aggregate monomeric sampling of 14.4 µs, 53.0 µs, and 51.0 µs is attained for the three data sets listed above respectively.[13] The accuracy of the atomic detailed whole-virion models and the amount of simulation time are promising in regards to how faithful the simulations are to a real-world scenario. Having a long simulation is importance on validating a system because it offers more data points to analyze. One of the challenges that arose because of the system, however, was due to the fact that in the process of glycosylating the system, protein folding caused a challenge where the N146 site was glycosylated in some monomers and not in others. Because of this discrepancy, a comparison between monomers with N146 glycosylated and monomers without N146 glycosylated was performed for the 2009 and 2015 glycosylated models, leading to two further subsets for each system. Another challenge that arose from creating the subsets was that the data sets of having the glycosylation at the N146 site were much smaller in size, resulting in lesser validation because of the limited simulation time. The N146 data set for the Shandong 2009 model had 16 monomers with no-N146 glycosylation had the 104. The Michigan 2015 N146 data set had a population of 18 monomers with the no-N146 glycosylated data set had 102 monomers. Though some degree of validation was reached, this itself showed the importance of having sufficient sampling for the MSM analysis to work optimally. It becomes quite apparent that a smaller data set results in poor validation of the MSM. The trajectories of the glycosylated models were generated with 0.06 ns steps whereas the trajectory of the unglycosylated model was generated with 0.02 ns in between each step.

Chapter 2

Markov State Models

2.1 Markov Models

Markov state models(MSMs) are one of the most powerful tools to analyze long timescale dynamics of systems with distinct states. A classic MSM is a model made by characterizing a system into different states and generating a probability matrix to assess the dynamics of transitions of each state. Unlike a Bayesian model, MSMs assume each state is independent from its "past" in terms of states.[14] All MSMs and subsequent computation analyses were run with Pyemma.[5] By running the data set trajectories with different Markov models, a lot of information could be gathered from kinetics as well as structure.

Traditional MSMs were not the only model I attempted to build. To get a higher quality fit with data, multiple Markov models were built with differing results. Regardless of which of the models was best for each strain of virus, what was important was that the parameters, such as the lag time or number of states. The type of MSM models I decided to use were consistent across all the whole-virion models for the most logical comparisons. Some of the MSMs I built include hidden Markov models(HMM), Bayesian hidden Markov models(BHMM), and regular Markov state models(MSMs) as well as Bayesian Markov models(BMM).

Unlike MSM, HMM introduce several key factors that give distinct nature to building the model. As stated before, MSMs function by building a model by determining the kinetics of transitions between different observed states. However, HMM assumes there are hidden theoretical states that may be unobserved, so the model itself builds upon creating these theoretical states and testing the models for its Markovian nature. The process of building a hidden Markov model involves testing theoretical states to see if hidden intermediates would work well with the system in making a set of transition probabilities.[16]

Creating a BMM entails having a prior understanding of the probability matrix. It creates a model based on causative probabilities. Because Bayesians assume previous knowledge of the probability matrix, building the BHMM and BMM involves extrapolating parameters drawn from creating a model prior.[17]

2.2 Feature Selection

In order to build an MSM, it is necessary to select a feature that informs on the dynamic behavior of the studied biological system. Previous work done on the influenza neuraminidase with Markov state models showed that using the distance between residues 149 and 431 of neuraminidase head informed on the dynamics of the 150-loop. When selecting for a single feature, it simplifies the Markov model greatly because the transition matrix has only one dimension. Once the data set has been generated, further analysis can be done with all the limitations of having only one dimension.[18]

2.3 Implied Timescales

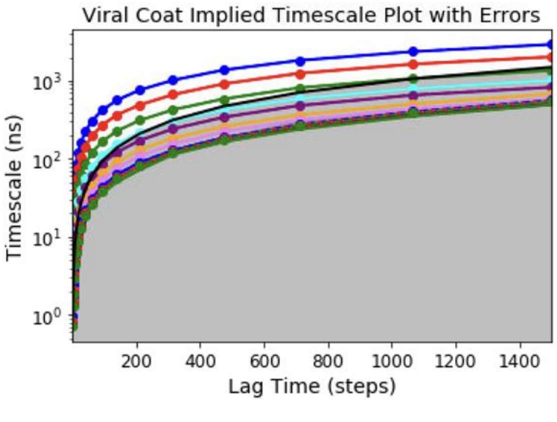

Figure 2.1: Timescale plot for the 2009 glycosylated model

The implied timescales plot greatly facilitates deciding parameters for building a rigorous MSM. The implied timescale plot is used to determine the relaxation timescale of the dynamics of the investigated process. This implied timescale plot effectively informs when the dynamics in question is affected by other states in the system, and from this plot, a proper lag time can be deduced to create additional Markov models. This chosen lag time must not be too small or too large. Extremely small lag times will increase the error of the approximations for the transition matrix probabilities, but extremely large timescales will cause numerical errors in estimations as well. An ideal lag time would be a point where the implied timescale(ITS) curve first tapers out and becomes flat. The $t_i = \frac{-\tau}{\ln|\lambda_i(\tau)|}$ timescale plot is defined by: where t_i is the estimation is the timescale, □ is the lag time, and □$_i$ is the eigenvector of the Markov transition matrix.[19]

Based on the implied timescale plot presented above in figure 2.1, a lag time of 500 steps was chosen as a start for the 2009 and 2015 glycosylated models, but a lag time of 1500 steps was used for the unglycosylated 2009 system. This is because for the glycosylated models, the trajectories were generated with a timestep of 0.06 ns while the unglycosylated models were generated with a timestep of 0.02 ns, which means we had to triple the steps to have equivalent lag time. I decided to run iterations of the models from lag time 250 to 750 with increments of 50 in the unglycosylated models just to improve potential quality of the results given that much of the process was qualitative judgements. In the end, running different MSMs led me to stick with my initial judgement of using 500 for the glycosylated models, as they were better validated as shown in later steps.

With this lag time, I built a traditional Markov state model for further data analysis. Building a traditional Markov model is necessary to make estimates on the states of other

Markov models, and given that we were running multiple models, it was important to run the different Markov models using the same parameters for the sake of consistency. Though feature selection was not selected upon previous work due to previous work being done on the neuraminidase 150-loop, it was still a crucial decision to run this analysis on all the influenza models to make sure the glycosylation, as well as the number of states were not affected by the differences in strain. That being said, even if differences were found, further analysis would be needed before adjusting for the number of states in the final calculations to see if the visualized states offer any pertinent insight into neuraminidase 150-loop behavior.

2.4 Timescale Separation Plot and Bayesian Markov Model

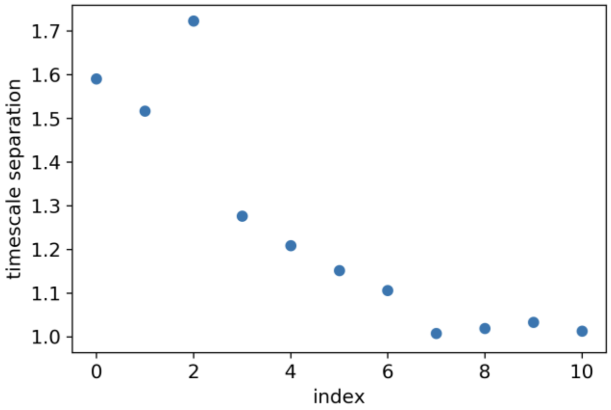

Figure 2.2: Timescale separation plot for 2009 glycosylated data set

Plotting the timescale separation allows assessment of how many stable states there are for further analysis. This is important because more accurate estimations can be produced after determining the amount of statistically significant stable states that comprise the data set. This parameter, called “nstates”, is used as an input in building further Markov models such as the BMMs that need further input to start. Each point in this figure denotes a relaxation timescale. Therefore, based on the number of peaks we can infer the number of metastable states. We wanted to keep the metastable states set to two ultimately because of what is known about neuraminidase 150-loop, but it was still helpful to run iterations with multiple metastable states to see if there was significant behavioral changes regarding the dynamics of the protein. Though individual timescale separation plots did raise the question of additional nstates, upon running the Markov state models with these states, it was clear that those states did not have much structural significance, and further analysis shows that those states also had very little metastable densities. All in all, the two state model seemed the most logical when analyzing the 150-loop. Running different timescale separation plots was also helpful in determining whether the Markov models that were being built were comparable across influenza strains to see if perhaps the differences themselves played a role in affected the number of states. Next, I generated the HMM, BMM, BHMM using the lag time (500 steps) for the glycosylated models and 1500 for unglycosylated system to see which Markov model validated the best. As shown below, the BMM had consistently good validation, so we opted to keep using the BMM to analyze all the three data sets.

2.5 Chapman-Kologromov Test

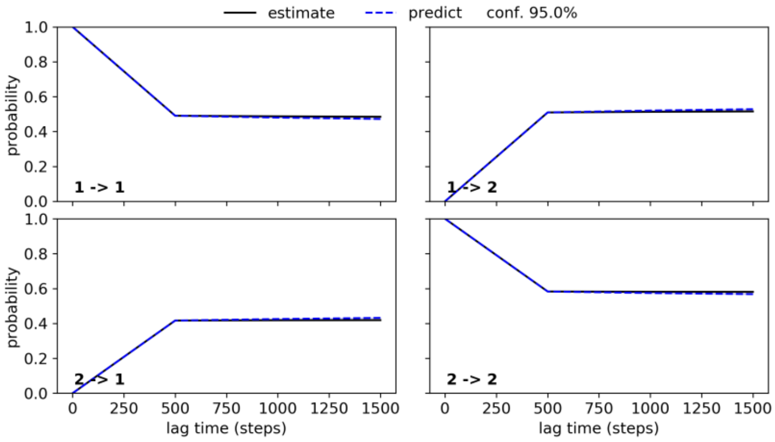

Figure 2.3: Chapman-Kologromov test for 2009 glycosylated Bayesian Markov Model

The Chapman-Kolmogorov, or CK equation is defined by: $P_{ij}^{(n)} \equiv \Pr(X_n = j | X_0 = i)$ where P_{ij} is the probability of the system moving from state *i* to *j*. Good validation would mean that the value of the CK equation would match that of the transition matrix.[20] We determined that a good CK test validation has the CK equation and transition matrix match up for around triple the lag time of the MSM being validated. This is the metric by which a Markov model is assessed to have Markovian behavior. That is to say when the graphs denote that the CK tests match the estimated result given by the calculated transition matrix, we can conclude that the parameters by which the Markov state model were made would give accurate results on the dynamics of the system.

2.6 Protein Rendering

The next steps would be to visually inspect the representative 150-loop conformation for each state extracted from the MSM analysis to see if the structures matched with the expected conformation. If the resulting structures clearly show the expected clear open and closed conformations, we went forward to calculate the metastable distributions and mean first passage times for the transition of the states. Visualizing the protein in order to assess the conformation was also necessary to understand the morphological and geometrical properties of the open and closed conformations of the 150-loop. Given that the system clustered around two main states given my input of nstates, I set a very high cutoff probability of 0.99 for all the Markov models to generate structures that were most likely to be the visual representatives of the metastable states to a very high degree of certainty.

2.7 Stationary Distributions and Mean First Passage Times

Next, we calculated the mean first passage time which is the average timespan required for the system to first go to a certain state to another for the first time. We also calculated the standard deviations for the mean first passage times to get a sense of the statistical error of the mean. We also generated the stationary distribution data, which shows us how much of each state is populated. A high population would mean that it is more probable for the system to be in that state at any given time. All this was done using computational scripts built into Pyemma.

Chapter 3

Results

3.1 CK Test Comparisons

In order to assess the reliability of the results, the first step was to see whether the parameters were sufficient in generating reproducibly Markovian models. As stated before, the way to do this was the run the various CK tests in order to find whether the transition matrices match the estimated model. To preserve simplicity, parameters were chosen such that all the data sets could be run with the same parameters with good results in validation. The pertinent number was two states with a lag time of 500 steps for all glycosylated models and 1500 steps for the unglycosylated model.

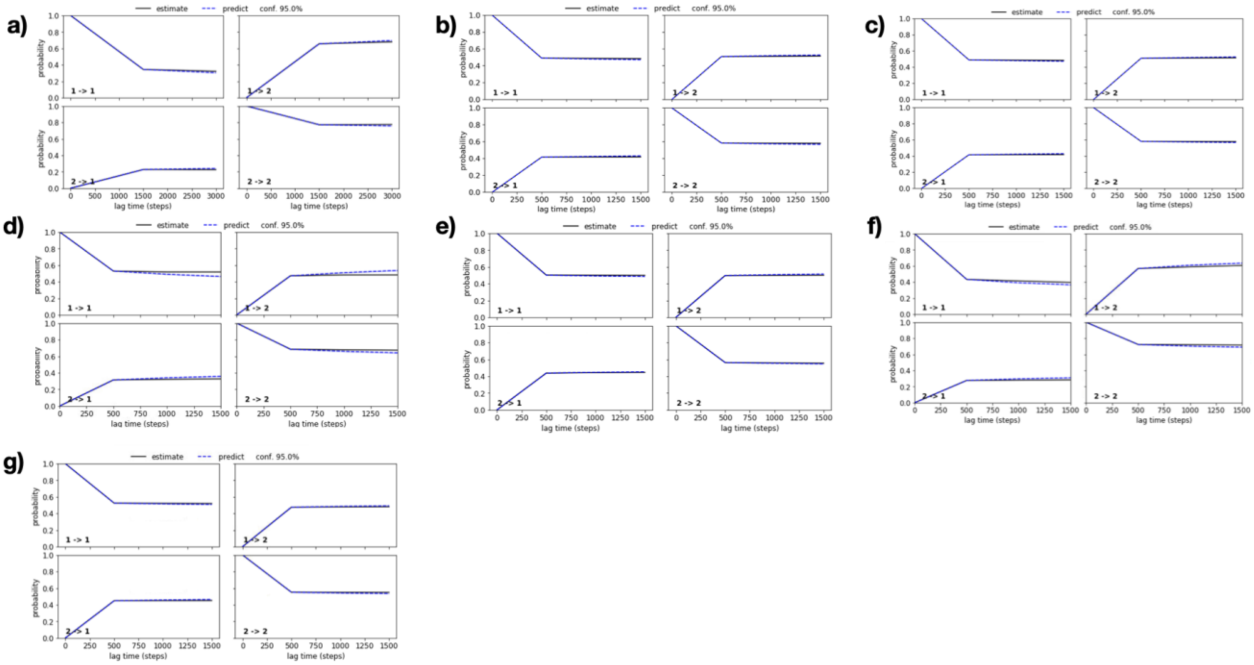

Figure 3.1: CK tests results. (a) 2009 unglycosylated. (b) 2009 glycosylated. (c) 2015 glycosylated. (d) N146 2009. (e) No-N146 2009. (f) N146 2015. (g) No-N146 2015.

The results as shown in figure 3.1 demonstrate that the MSMs built for all three datasets validated relatively well in Markovian fashion. The estimates and transition matrices matched well into a lag time of 1500 or more steps as desired for the three models. However, given that the N146 data pools were much smaller, there was some error involved. To address this, further iterations of the Markov models were done with different lag times, but the lag time of 500 steps was kept for the glycosylated models since only the limited size of the data was the source of the issues of validation for the N146 subset. Changing the lag time did little to alleviate such error of the CK test for the N146 subset, and only affected the validation of the other data sets negatively.

3.2 Rendering Protein Results and Results for Statistical Analyses

Using the high cutoff probability, I was able to extract representative conformation of the open and closed states for all the data sets. Even in the case of poor validation for the 146 subsets, the molecular representations show a net difference between the open-closed conformations. All image renderings done with VMD.[21] Surface representations were used to detail the shape of the neuraminidase active site, highlighting how the cavity becomes larger in the case of an open 150-loop. The distance between residue 149 and 431 was also highlighted to show the differences between the two states.

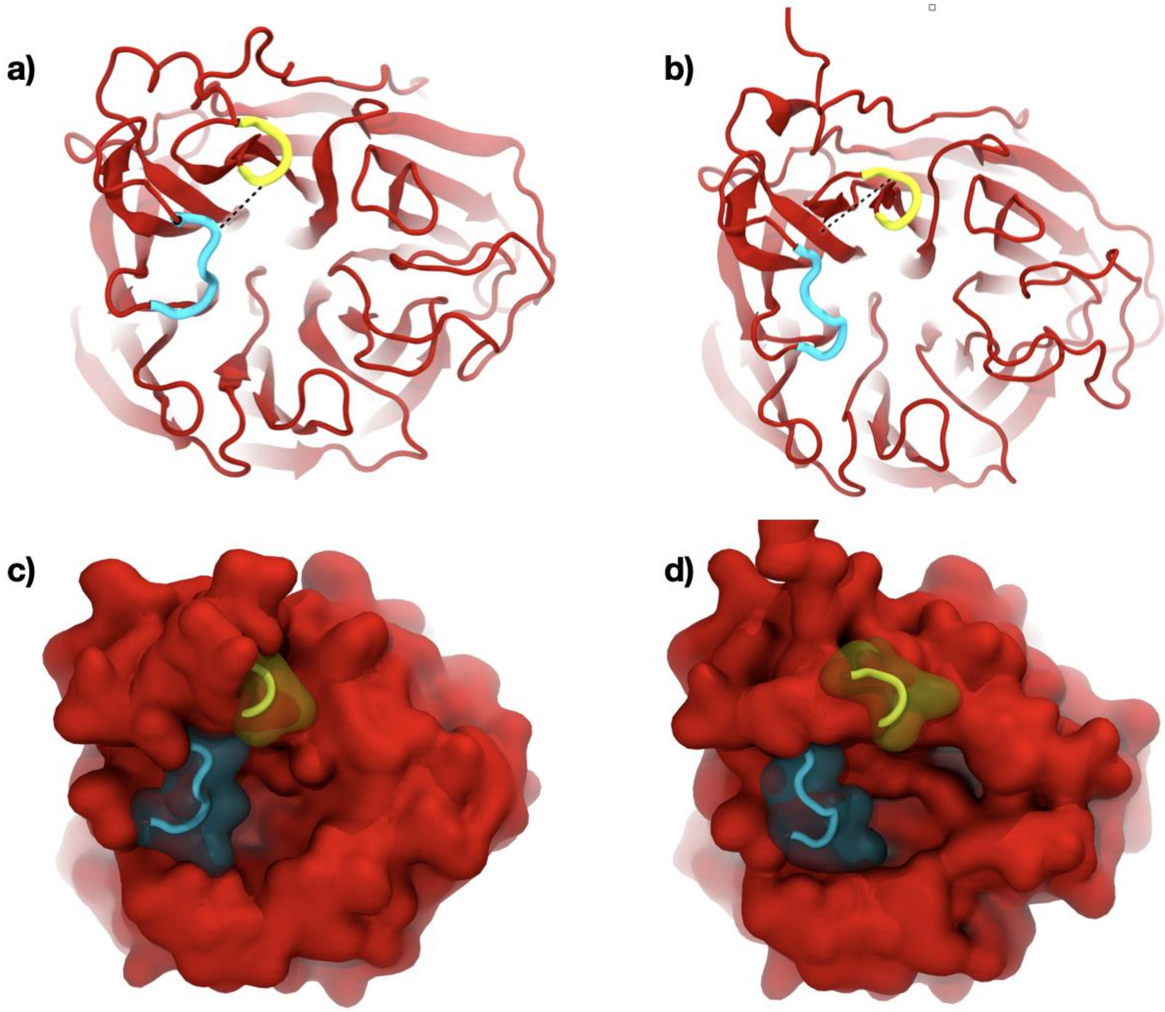

Figure 3.2: Rendered macrostates for the 2009 unglycosylated data set. Cyan denotes 150-loop, whereas yellow denotes 430-loop. (a) closed conformation. (b) open conformation. (c) Surface representation of closed conformation. (d) Surface representation of open conformation.

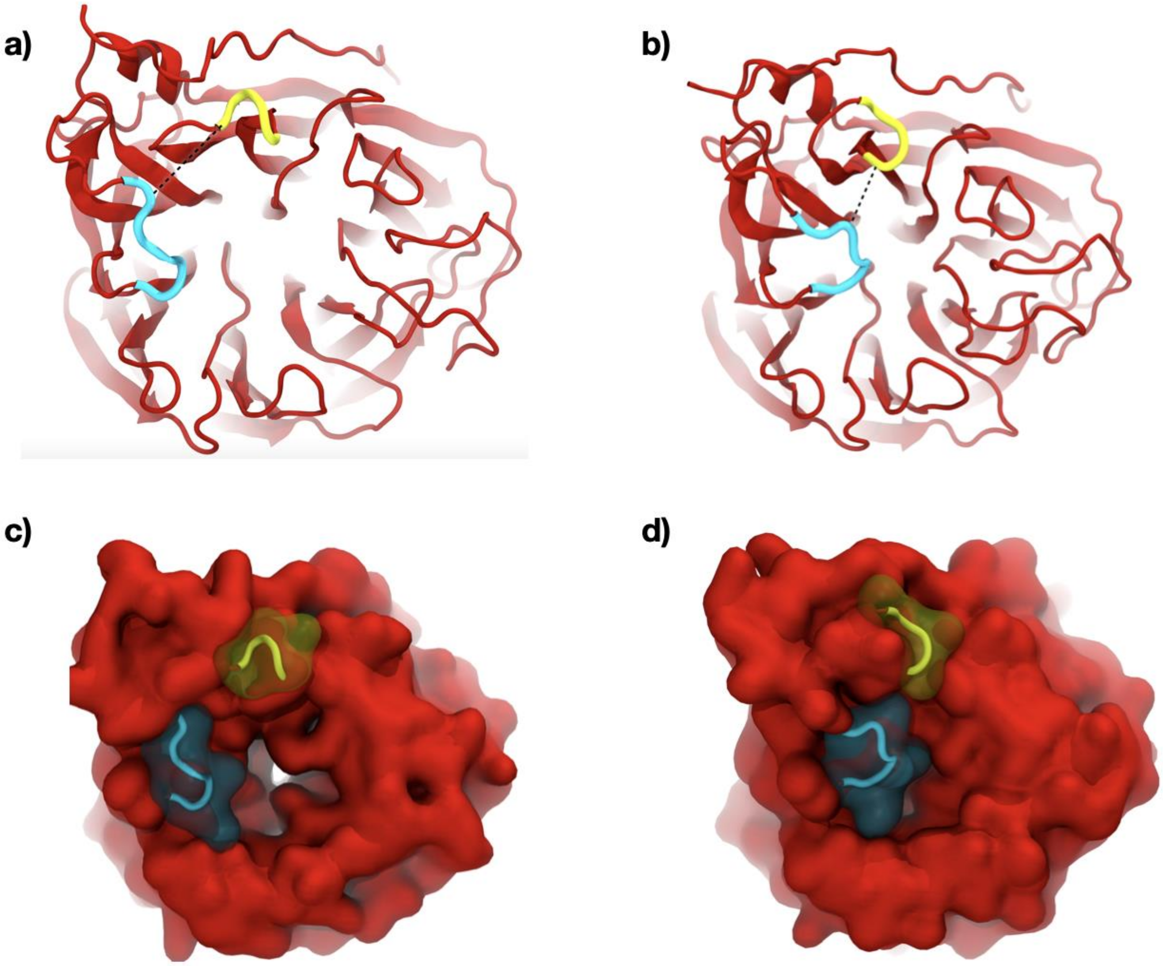

Figure 3.3: Rendered macrostates for the 2009 glycosylated data set. Cyan denotes 150-loop, whereas yellow denotes 430-loop. (a) open conformation. (b) closed conformation. (c) Surface representation of open conformation. (d) Surface representation of closed conformation.

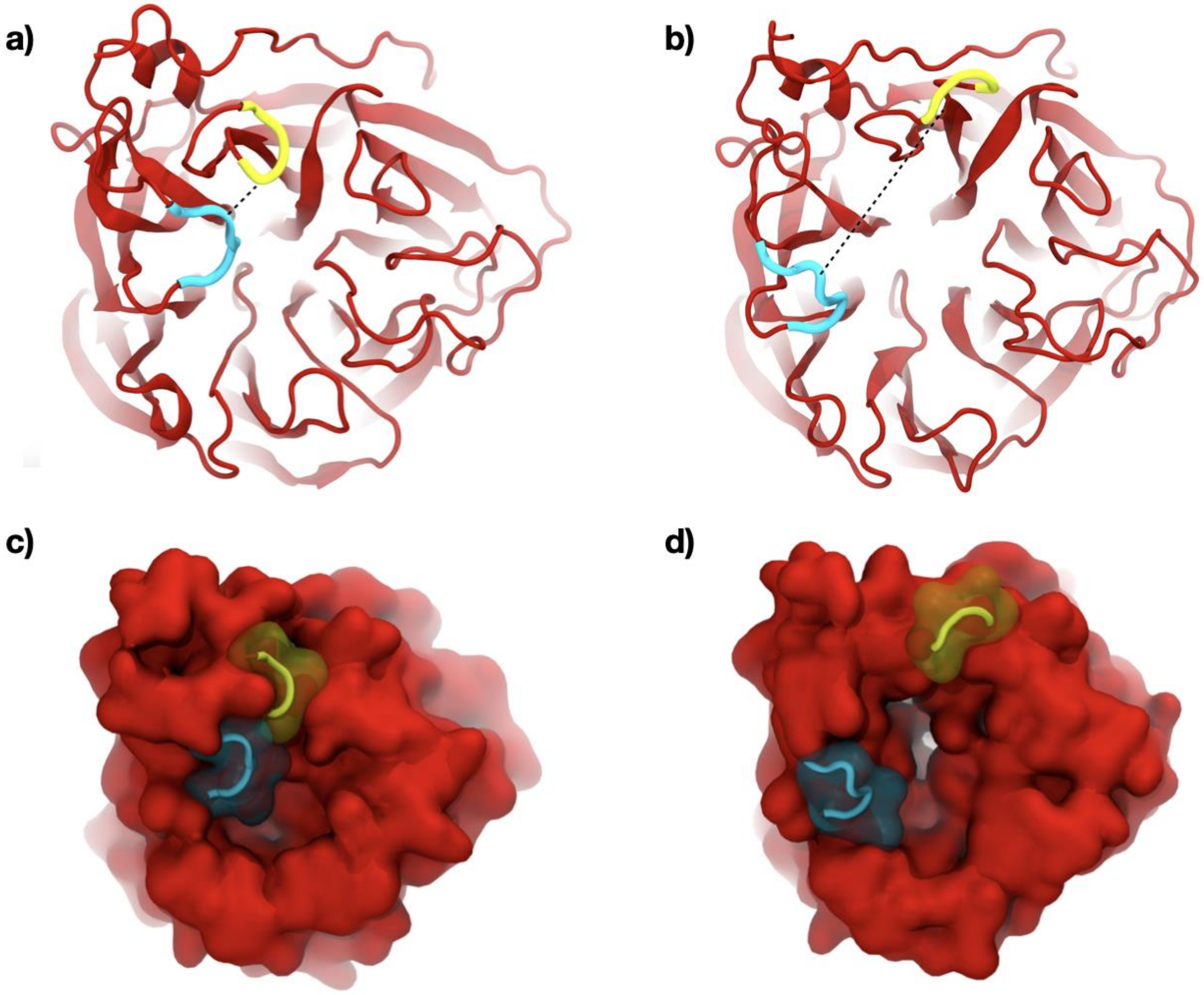

Figure 3.4: Rendered macrostates for the 2015 glycosylated data set. Cyan denotes 150-loop, whereas yellow denotes 430-loop. (a) closed conformation. (b) open conformation. (c) Surface representation of closed conformation. (d) Surface representation of open conformation.

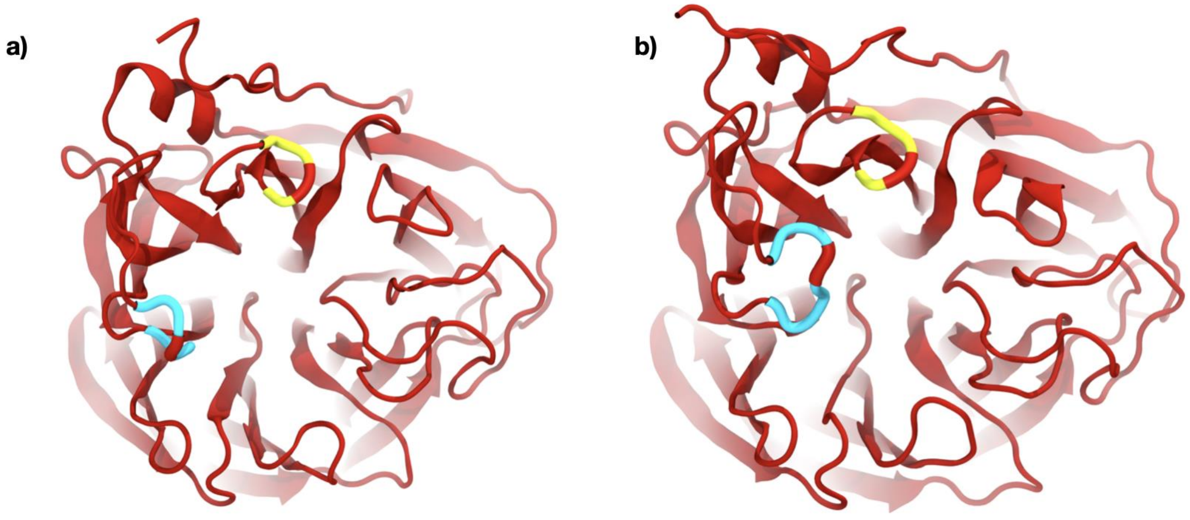

Figure 3.5: Rendered macrostates for the 2009 N146 data set. Cyan denotes 150-loop, whereas yellow denotes 430-loop. The crimson portion within the loop denote the 149 and 431 residues (a) open conformation. (b) closed conformation.

Despite having relatively poor validation in the CK test, the molecular representation shows that the chosen feature was able to discretize the data set properly. Moreover the noticeable differences between the two conformations, as also shown by the highlighted distance between residue between 431 and 149, pinpoints that the Markov state model analysis was able to characterize two relevant states from the 150-loop dynamics. This suggests that the adopted workflow was successful in studying the 150-loop dynamics.

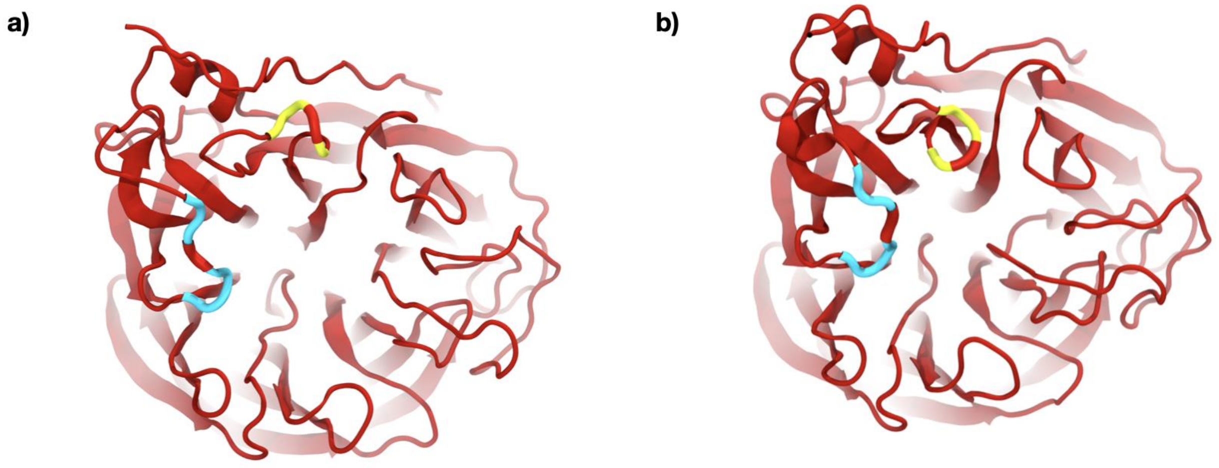

Figure 3.6: Rendered macrostates for the 2009 No-N146 data set. Cyan denotes 150-loop yellow is 430-loop while the crimson portion within the loop denote the 149 and 431 residues (a) open conformation. (b) closed conformation.

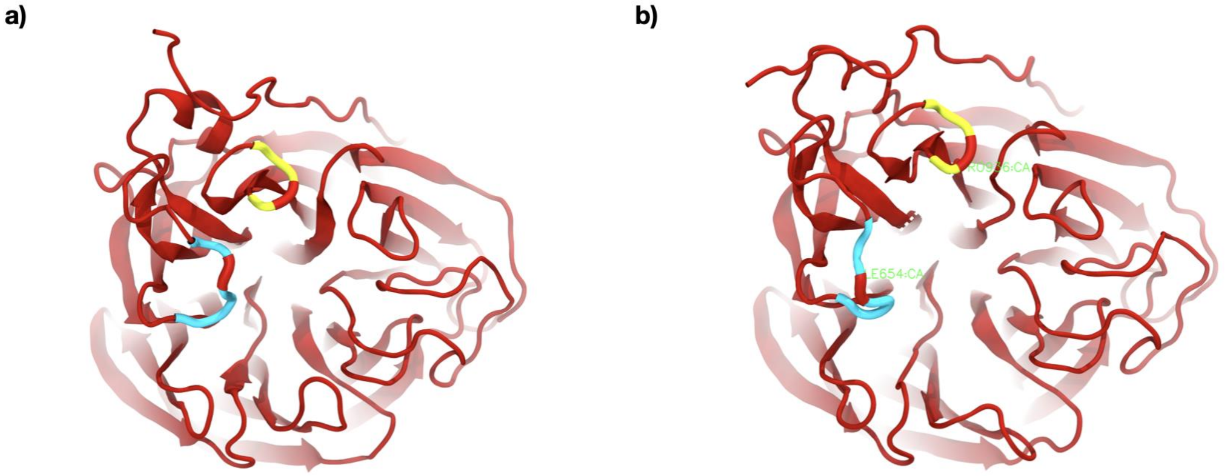

Figure 3.7: Rendered macrostates for the 2015 N146 data set. Cyan denotes 150-loop yellow is 430-loop while the crimson portion within the loop denote the 149 and 431 residues (a) closed conformation. (b) open conformation.

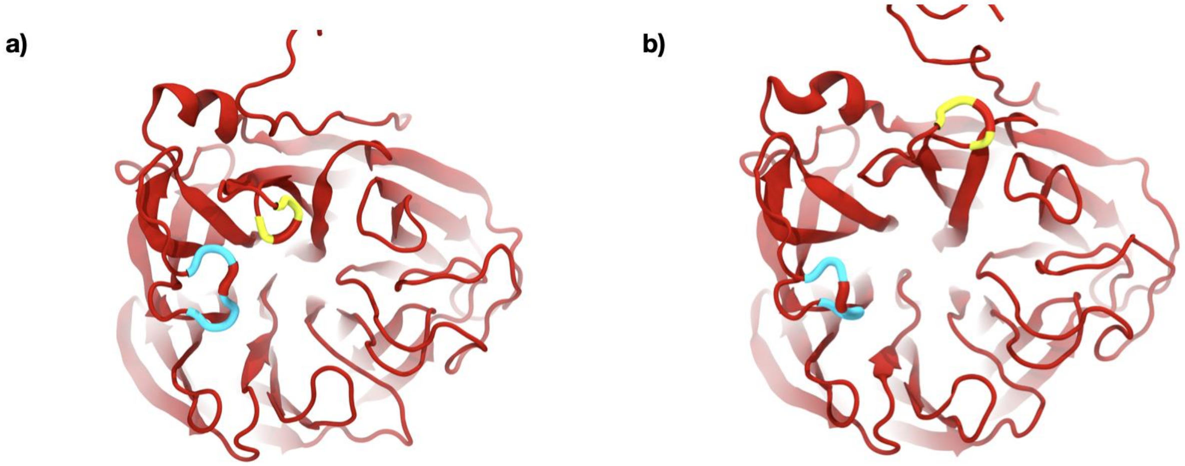

Figure 3.8: Rendered macrostates for the 2015 No-N146 data set. Cyan denotes 150-loop yellow is 430-loop while the crimson portion within the loop denote the 149 and 431 residues (a) closed conformation. (b) open conformation.

Table 3.1: All mean first passage times, distance between 149 and 431 residues, as well as stationary distributions. The distances were measured based off the rendered neuraminidases, and the mean first passage times and metastable distributions were calculated using statistical analyses using Pyemma.[22]

System	Closed to Open(ns) +/- Standard deviation	Open to Closed(ns) +/- Standard deviation	Distance(open) Angstroms	Distance(closed) Angstroms	Stationary distribution A:B	Lag time(steps)
2009 UNGLY	8149 +/- 247	7386 +/- 184	14.7(Ile2061-Pro2343)B	8.9(Ile2061-Pro2343)A	.11:.89 closed:open	1500
2009	34810 +/- 1410	4980 +/- 71.2	18.9(Ile654-Pro936)A	6.7(Ile654-Pro936)B	.21:.79 open:closed	500
2015	7573 +/- 121	5131 +/- 78.1	26.2(Ile654-Pro936)B	7.1(Ile654-Pro936)A	.37:.63 closed:open	500
N146 2009	7360 +/- 352	8503 +/- 373	19.3(Ile654-Pro936)A	9.8(Ile654-Pro936)B	.37:.63 open:closed	500
NO146 2009	6369 +/- 81.8	9147 +/- 217	18.7(Ile654-Pro936)A	7.4(Ile654-Pro936)B	.32:.68 open:closed	500
N146 2015	7821 +/- 360	21324 +/- 2360	16.2(Ile654-Pro936)B	7.4(Ile654-Pro936)A	.24:.76 closed:open	500
NO146 2015	1803 +/- 51.7	28727 +/- 742	24.9(Ile654-Pro936)B	5.1(Ile654-Pro936)A	.46:.54 closed:open	500

Chapter 4

Analysis

4.1 Effect of Glycosylation on the 150-loop dynamics

Comparing the 2009 unglycosylated model to the 2009 glycosylated model gave a clear view of the differences of the impact that Glycosylation had on the 150-loop dynamics. As seen from figures 3.2 to 3.4, as well as table 3.1, all of the models exhibit the clear open and closed conformation. In the case of the open state, the presence of glycans in the glycosylated 2009 and 2015 systems led to an increased distance between the 150 and 430-loop, as evidenced by the measured distance between 149 and the 150-loop and residue 431 located on the 430-loop(see table 3.1). In particular, the closed conformation in the unglycosylated 2009 model had a distance of 8.9 Å whereas the open had a distance of 14.7 Å. For the glycosylated 2009 model, the closed distance was 6.7 Å and the open was 18.9 Å. A larger distance (26.2 Å) between the two loops in the case of the open state was also observed for the 2015 model, confirming the trend observed for the 2009 glycosylated model.

Interestingly, the glycosylated 2009 model also showed a greater net difference between the 150 and 430-loop distance observed in the closed state and the open state, owing to the closed state being more closed than in the unglycosylated2009 model and the open state being more open. Another interesting note to be had was the metastable distributions. The unglycosylated model had a larger population for the open conformation whereas the glycosylated model had a larger population for the closed conformation, highlighting a shift of the predominant state towards the closed conformation in the case of the 2009 glycosylated system. The mean first passage times were also greatly affected by the Glycosylation. The glycosylated model had a whopping 34,800 ns closed-to-open and a shorter 4980 ns open-to-closed mean first passage time whereas the unglycosylated model had an 8,100 ns closed-to-open and 7,400 ns open-to-closed mean first passage times. However, mean first passage times may not be the best assessment of the system. When I generated multiple MSM types in order to assess which of them had the best consistent validation, I also formed mean first passage times for all types of Markov Models. Of these depending on the type of MSM the mean first passage times would be very different, though the metastable distributions would remain more or less the same throughout iterations as long as the Markov state models validated correctly. The shapes from the renders and metastable distributions were much more useful for this comparison, and make more logical sense. Of note, since the 2015 and 2009 glycosylated data sets are comprised of the N146 and no-N146 subsets, the expectation is that the mean first passage times and metastable distributions would reflect that tendency. However, for some reason, this turned out not to be the case for mean first passage times. The metastable distributions, however, were very near the expected amount where the metastable distribution of the main population was somewhere in vicinity of the subpopulations.

4.2 Effect of Strain

Because of the mutations that caused differences between the 2009 and 2015 influenza strains, especially the K432E mutation right within the 430-loops, I expected some differences in the dynamics of the 150-loop between the 2009 and 2015 strain. This was indeed the case. Between the 2009 and 2015 glycosylated strains, the 2009 had the 150 loop-430-loop distances in the closed and open state of 6.7 Å and 18.9 Å respectively, whereas in the 2015 strain the distances were 7.1 Å and 26.2 Å. The mean first passage times were also greatly affected, as the 2009 had a MFPT of 34,800 ns closed-to-open and a 4,980 ns open-to-closed transition. The 2015 had a MFPT of 7,570 ns for the closed-to-open transition, and a 5,130 ns MFPT for the open-to-closed transition. The stationary distribution trends were also reversed, with the 2009 having a 21%/79% distribution in favor of the closed conformation whereas the 2015 exhibited a distribution of 63%/37% in favor of the open conformation.

It is plausible that the mutations occurred within the 2015 strain might have had an impact on the 150-loop dynamics. Since both strains were involved in functional influenza, the change in neuraminidase 150-loop dynamics did not affect the capacity of the virus to infect on the whole, though it could be true that the dynamics could change the virulence and speed in which the virus operates. It is very plausible this is the case, though it would be hard to prove just looking at the historical spread of a virus given the multitude of factors involved. Nevertheless, creating these Markov state models and assessing the kinetics of the system could give a strong hypothesis as to how effective these components are with infectivity.

4.3 Effect of N146 Glycosylation

Compared to the no-N146 subsets for both groups, the N146 subsets had consistently less population in the closed conformation. It appears that having an N-linked glycan linked to N146 site skews the neuraminidase 150-loop towards being in the open state. For the 2009 strain, the population of open conformation goes from 32% to 37% in the case of the presence of glycan at the N146. For the 2015 strain, the open conformation population goes from 54% to 76% in the case of the presence of the glycan at the N146. The fact that the difference in effect of the Glycosylation is greater in the 2015 is peculiar, albeit the trends are the same. The behavior of the 150-430-loop distances given by the feature is also very interesting. The lack of the N146 glycan in the no-N146 subset increased the range of the distance for the 2015 influenza and the 2009 strain. For the 2009 N146 subset, the open-closed distances were 19.3 Å and 9.8 Å respectively. The no-N146 set for the 2009 yielded open-closed distances of 18.7 Å and 7.4 Å respectively. Looking at the distances, that is a range of 9.5 Å in the N146 and an 11.3 Å range in the no-N146. For the 2015, no-N146 gives us open-closed distances of 24.9 Å and 5.1 Å respectively, while the N146 gives us an open-closed distance of 16.2 Å and 7.4 Å respectively. This yields a range of 8.8 Å for the N146 for the 2015, and a range of 19.8 Å for no-N146. It seems based on these two strains of influenza, the Glycosylation at the 146 residue of neuramindase causes both an increase in the open conformation as well as a decrease in the extent by which the neuraminidase opens and closes.

4.4 Further Research Goals

Running Markov State models and doing analysis to get an idea on the kinetics is a start to a plethora of potentially epidemically and pharmaceutically significant findings. Given that neuraminidase is such a biologically relevant actor in infectiousness, it would be potentially groundbreaking to study more viral strains on the grounds of infectivity and virulence on the basis of neuraminidase behavior. Controlling for other factors, if there could, for example, be a link between the metastable states skewing towards a certain conformation, the variances or sheer distances of the open-close conformations, it could offer great insight into viral forecasting and preparing for future outbreaks of influenza, not to mention given the power of Markov state analyses, it could also be applied to other systems analyses involving drug binding to neuraminidase since it is the hot spot for many antivirals. In fact, Giugea et al proposes that adding neuraminidase into a vaccine cocktail would be beneficial and cites how effective anti-neuraminidase antibodies are in controlling the disease in the body.[23] Using MSM analysis allows us to not only study neuraminidase but also study the interactions with antibodies could also give insight on the exact mechanism of actions certain natural defenses against disease, which can provide insights on drug discoveries to mimic these natural defenses.

REFERENCES

(1) Gutenberg JK, Morens DM. 1918 Influenza: the mother of all pandemics. Emerg Infect Dis. 2006 Jan;12(1):15-22. doi: 10.3201/eid1201.050979. PMID: 16494711; PMCID: PMC3291398.

(2) World Health Organization. (2018, November 6). "Influenza (Seasonal)."

(3) Jung HE, Lee HK. Host Protective Immune Responses against Influenza A Virus Infection. V*iruses*. 2020; 12(5):504.

(4) Bouvier, Nicole M, and Peter Palese. "The biology of influenza viruses." *Vaccine* vol. 26 Suppl 4,Suppl 4 (2008): D49-53. doi:10.1016/j.vaccine.2008.07.039

(5) Mcauley, Julie & Gilbertson, Brad & Trifkovic, Sanja & Brown, Lorena & McKimm-Breschkin, Jennifer. (2019). "Influenza Virus Neuraminidase Structure and Functions." *Frontiers in Microbiology*. 10. 10.3389/fmicb.2019.00039.

(6) Tao J, Wang H, Wang W, Mi N, Zhang W, Wen Q, Ouyang J, Liang X, Chen M, Guo W, Li G, Liu J, Zhao H, Wang X, Li X, Feng S, Liu X, He Z, Zhao z. (2022) Binding mechanism of oseltamivir and influenza neuraminidase suggests perspectives for the design of new anti-influenza drugs. PLOS Computational Biology 18(7): e1010343.

(7) Giurgea LT, Morens DM, Taubenberger JK, Memoli MJ. Influenza Neuraminidase: A Neglected Protein and Its Potential for a Better Influenza Vaccine. Vaccines (Basel). 2020 Jul 23;8(3):409.

(8) Durrant JD, Kochanek SE, Casalino C, Leong PU, Dommer AC, and Amaro RE. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. ACS Central Science **2020** 6 (2), 189-196

(9) Lu, J., & Meyer, S. (2020). Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model. *International journal of environmental research and public health*, *17*(4), 1381.

(10) Amaro, R. E., Swift, R. V., Votapka, L., Li, W. W., Walker, R. C., & Bush, R. M. (2011). Mechanism of 150-cavity formation in influenza neuraminidase. *Nature communications*, *2*, 388.

(11) Durrant JD, Kochanek SE, Casalino C, Leong PU, Dommer AC, and Amaro RE. ACS Central Science **2020** 6 (2), 189-196

(12) Casalino L, Seitz C, Lederhofer J, Tsybovsky Y, Wilson I, Kanekiyo M, Rommie A. Breathing and tilting: Mesoscale simulations illuminate influenza glycoprotein vulnerabilities.

bioRxiv. Cold Spring Harbor Laboratory; 2022
https://www.biorxiv.org/content/10.1101/2022.08.02.502576v2

(13) Ibid

(14) Brooke E. Husic and Vijay S. Pande. Markov State Models: From an Art to a Science.Journal of the American Chemical Society **2018** 140 (7), 2386-2396

(15) Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JK, and Noé F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.Journal of Chemical Theory and Computation **2015** 11 (11), 5525-5542. DOI: 10.1021/acs.jctc.5b00743

(16) Yoon B. J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current genomics*, *10*(6), 402–415.

(17) Siebert, M., & Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic acids research*, *44*(13), 6055–6069.

(18) Konovalov, K. A., Unarta, I. C., Cao, S., Goonetilleke, E. C., & Huang, X. (2021). Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au*, *1*(9), 1330–1341.

(19) Swope, W. C., Pitera, J. W., Suits, F., Pitman, M. and Eleftheriou, M.: Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and beta-hairpin peptide. *Journal of Physical Chemistry B* 108, 6582-6594 (2004).

(20) Miroshin, R.. (2016). Special solutions of the Chapman–Kolmogorov equation for multidimensional-state Markov processes with continuous time. Vestnik St. Petersburg University: Mathematics. 49.

(21) Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

(22) Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JK, and Noé F. **2015** 11 (11), 5525-5542.

(23) Giurgea LT, Morens DM, Taubenberger JK, Memoli MJ. Vaccines (Basel). 2020 Jul