Markov State Model Analysis of LB6 Peptidomimetic Dynamics

Mariam Hergnyan
mariam.hergnyan@studenti.unipd.it

1. Introduction

Gelsolin amyloidosis (AGel) is a condition where the gelsolin (GSN) protein forms abnormal deposits in tissues, causing various health problems. Currently, there is no cure for AGel. More than ten genetic mutations in the GSN gene can lead to AGel. The D187N/Y mutation is the most common. These mutations cause the protein to break down in a harmful way, forming clumps of protein pieces: amyloids. These pieces, particularly those containing a region called Gelsolin Amyloidogenic Core(GAC$_{182-192}$), are responsible for the disease symptoms. Bollati et al. [1] propose a novel approach using selective peptidomimetics(lb5, lb6, lb7) to disrupt amyloidogenic GSN aggregates. Containing a segment of the amyloidogenic core, these peptidomimetcs possess the ability to engage with both monomeric and/or oligomeric forms of GAC182–192, resulting in the prevention of amyloid filament formation.

Peptidomimetics are synthetic molecules designed to mimic the structure and function of peptides, but compared to peptides there have more stability, which makes them valuable tools in drug discovery and development.

We will focus our work on one of these peptidomimetics: lb6. To understand the dynamics of amyloid formation and assess intervention efficacy, we employ Markov State Modeling (MSM). MSM offers a mathematical framework for analyzing time-series data from molecular dynamics simulations, facilitating the exploration of complex state dynamics.

1.1. Biological problem

Drug potency is related to the free energy dynamics of biomolecular systems[2]. To measure relative free energy, we utilize Markov State Models (MSMs) to derive the stationary distribution, which reflects equilibrium probabilities within the system. These probabilities, indicative of the system's free energy landscape, are pivotal in determining drug potency. Our investigation centers on unraveling these dynamics to inform the rational design of therapeutic agents.

1.2. Defining the question

The main question of the project is which are the states visited by the lb6 and what are the equilibrium probabilities of those state which will eventually lead us to the free energy of binding.

1.3. Plan of Work

Since the analysis is done using the PyEMMA package[7], the steps followed in this work are similar to the ones suggested in the package tutorial:

1. extracting molecular features from the raw data,
2. transforming those features into a suitable, low dimensional subspace,
3. discretizing the low dimensional subsets into a state decomposition,
4. estimating Bayesian MSM from the discrete trajectories and performing validation tests,
5. analyzing the stationary and kinetic properties of the MSM,
6. finding metastable macrostates,
7. computing the stationary distribution, the free energy and the MFPT of the states.

2. Method

All the code developed and data used during the analysis can be found on Github [3].

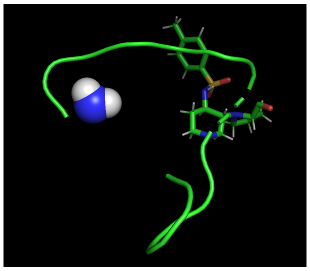

Figure 1. The structure of lb6 visualized in PyMol[8]
.

2.1. The Data

As mentioned before, the analysis is done on the lb6 peptidomimetic. The primary data sources include the initial structural coordinates in Protein Data Bank (PDB) format and the corresponding trajectory file in .xtc format, both obtained through MD methods. The trajectory file has undergone preprocessing, with water molecules filtered out. The trajectory contains 10,837 frames. The lb6 consists of 255 atoms (Figure 1). The sequence is ASNGLYASPCYSPHEILELEUMOLHISGLNTRPCYSGLYSERASNNHE.

The data utilized to derive the features for the analysis comprised five files, each representing a snapshot of the trajectory at various points (with an approximate interval of 2000 frames between each snapshot). For this task VMD [4] was used. Initially, we attempted to use a uniform step size of 252 frames. However, due to memory constraints, this approach was not feasible.

2.2. Methods

We begin with the application of the VAMP-2 [5]scoring method for feature selection. We will focus only on the backbone features and not the sidechains, so the feature selection will be done from backbone torsion angles, backbone atom positions and backbone atom distances. The VAMP-2 quantifies the kinetic variance captured by a given model.

The VAMP-2 score (C_{VAMP2}) is computed using the formula:

$$C_{\text{VAMP2}} = \frac{\sum_{i=2}^{n} \lambda_i}{\sum_{i=1}^{n} \lambda_i}, \tag{1}$$

where λ_i represents the eigenvalues of the cross-covariance matrix. A higher C_{VAMP2} signifies better feature representation of the molecular dynamics.

Higher VAMP-2 scores indicate a more comprehensive capture of relevant information regarding the system's kinetics. Selecting the model with the highest VAMP-2 score implies a superior representation of the kinetic variance compared to models with lower scores. The minimum attainable VAMP-2 score is 1, with higher scores indicating a better description of the molecule.

To manage the high dimensionality of our data, we employ dimensionality reduction techniques. While Principal Component Analysis (PCA) is a common choice, it may not adequately capture slow dynamics such as conformational changes or ligand binding. Time-lagged Independent Component Analysis (TICA) [9] surpasses PCA in this regard.

TICA is specifically tailored for analyzing time-series data, such as molecular dynamics trajectories. By incorporating time-lagged information, TICA identifies linear combinations of features that effectively capture slow processes within the data. Its application is often coupled with VAMP-2 analysis to evaluate the quality of dimensionality reduction and identify relevant dynamical processes.

The TICA transformation is defined as:

$$X_{\text{TICA}} = U^T X, \tag{2}$$

where X represents the input data matrix, U denotes the matrix of TICA eigenvectors, and X_{TICA} is the transformed data matrix.

The next step involves performing cluster analysis on the reduced-dimensional data to identify discrete states. While various clustering methods exist, we explore two: k-means and regspace.

K-means centers align with the density of data points, whereas regspace centers are uniformly spread across the dataset. Given our objective of discretization, k-means emerges as the preferred choice. However, if discovering new states is a goal, placing states in rarely observed regions may prove advantageous. Yet, in areas of low density, we will encounter poor statistical representation of these states. Thus, for our case, k-means is a better choice.

After obtaining the discrete trajectories, the next step involves determining the implied timescales (ITS) for a Bayesian MSM. For estimating a MSM we need to select a lag time, τ. This lag time should be long enough to ensure Markovian dynamics within our state space while remaining short enough to resolve the dynamics of interest.

Plotting the implied timescales (ITS) as a function of τ serves as a diagnostic tool for selecting the MSM lag time. The ITS, t_i, approximates the decorrelation time of the ith process and is computed from the eigenvalues λ_i of the MSM transition matrix using the formula:

$$t_i = -\frac{\tau}{\ln|\lambda_i(\tau)|} \tag{3}$$

When the ITS exhibit approximate constancy with the lag time, it indicates that our timescales have converged. Therefor, we select the smallest lag time with converged timescales to maximize the model's temporal resolution.

2.3. Statistical Analysis

After obtaining the ITS of the Bayesian MSM, we need to validate the markovianity of our probability matrix. The Chapman-Kolmogorov (CK) test [6] provides a robust means of validation. The CK propertyis defined as follows:

$$P(k\tau) = P^k(\tau), \tag{4}$$

where the left-hand side of the equation corresponds to an MSM estimated at lag time $k\tau$, and k is an integer greater than 1. The right-hand side of the equation represents our estimated MSM transition probability matrix raised to the power of k. (Figure 2).

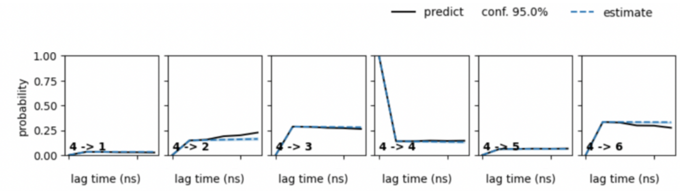

Figure 2. One example of the CK test for the 1st discrete state. We can see that there is not too much deviation between the estimate and the prediction. Therefore we can say that we have a good working model.

3. Results

As discussed in subsection 2.2, we conducted a VAMP2 calculation to identify optimal features and lag times. Despite VAMP2's limitations in lag time selection, we utilized it as a starting point for our analysis. Initially, we explored three lag times: $\tau = 0.5$ ns, $\tau = 1$ ns, and $\tau = 2$ ns. As illustrated in Figure 3, the highest VAMP2 score was obtained at $\tau = 0.5$ ns.

Given that the highest VAMP2 score indicates the feature that best describes our data, we determined that the backbone torsion angle property received the highest score. Therefore, we will base our feature selection on this information.

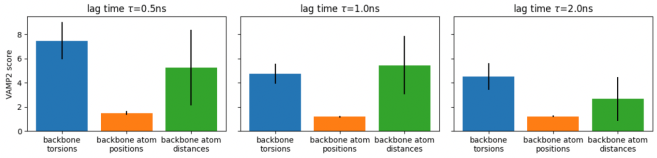

Figure 3. VAMP2 Score with Different Lag Times and Featurs

We further evaluated the dimension with the highest VAMP2 score while iterating through additional lag times τ. Our analysis revealed that at dimension = 10 and $\tau = 0.1$ ns, we achieved the optimal results.4

Using these parameters, we conducted TICA and initially obtained 10 independent components(IC). Specifically focusing on torsion angles, IC1 and IC2 capture dominant motions or conformational changes within the molecular system over time. IC1 typically represents the slowest process or primary mode of variation, while IC2 captures a secondary mode of motion orthogonal to IC1.

Distinct clusters in the density plot (Figure 5) correspond to various conformational states or basins, with transitions between states depicted by regions of lower density or overlapping densities. The projection reveals well-defined clusters of high density, representing metastable basins.

During k-means clustering, we observed saturation in the VAMP2 score after a certain number of cluster centers, leading us to infer the presence of six major clusters from the density plot. Consequently, we constructed our Markov State Model (MSM) using this insight. We computed the Implied Timescales (ITS) considering the cluster informa-

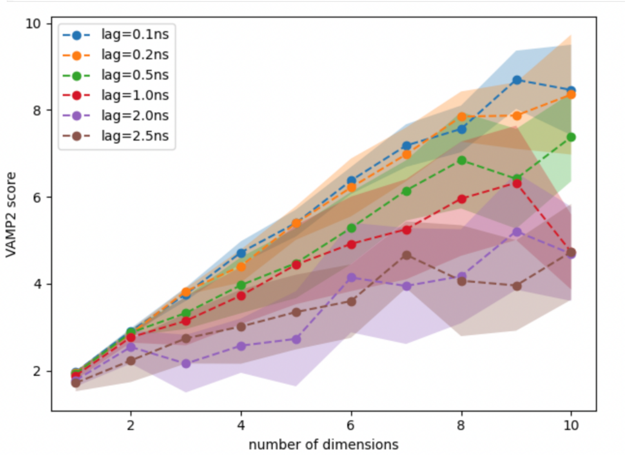

Figure 4. VAMP2 score with different lag times and dimensions

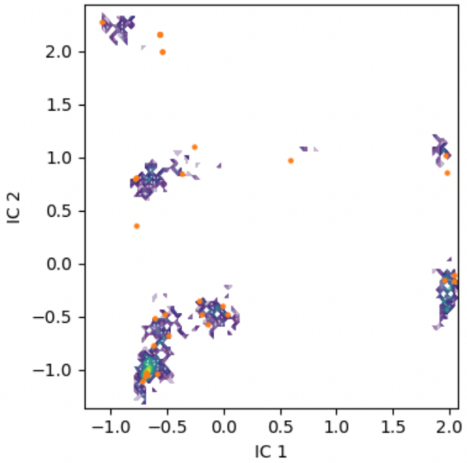

Figure 5. Density plot for the TICA projection IC1 and IC2 with cluster centers

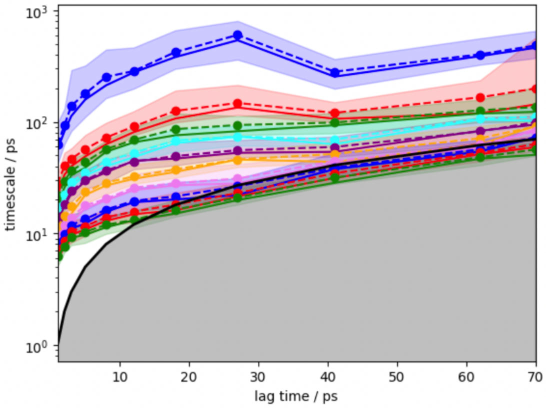

Figure 6. The Implied Timescales

tion. Figure 6 shows processes within the ITS above the gray shaded area, indicating slower dynamics than the lag time. Conversely, processes below the gray area are faster and cannot be resolved. Our analysis reveals nearly six resolved processes at a 10 ps timescale, with the ITS appearing to converge.

We proceed to compute the stationary distribution(Figure 7). We use the stationary distribution to visualize the weight of discrete states, showing the most probable areas in our feature space. The coefficients of the eigenvectors indicate the flow into and out of the Markov states, defining each process.

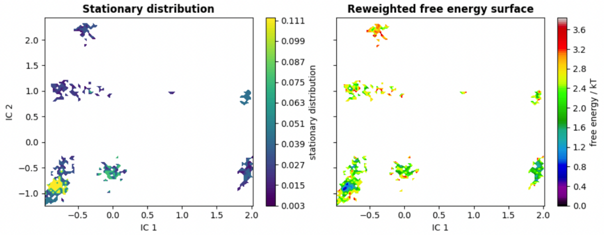

Figure 7. The stationary distribution of a MSM on a two-dimensional space defined by the first two ICs of a TICA projection. The color bar indicates the probability of the system being in that state in the long run

We continue with the PCCA++ algorithm which computes so called memberships, i.e., the probability of each microstate to belong to a given macrostate. Essentially, PCCA++ provides a fuzzy assignment of microstates to macrostates, with the resulting memberships capturing the degree of association between each microstate and macrostate. The first right eigenvector projected from TICA corresponds to the stationary process (equilibrium), and as such, it remains constant at 1. With the knowledge of PCCA++ metastable states, we can compute mean first passage times (MFPTs) between them. Additionally, we can compute MFPTs and equilibrium probabilities on the metastable sets and extract representative structures. In Figure 8, we observe the deduced states. It's evident that State 4 is not concentrated in one specific region. This observation aligns with our expectations, especially considering the results of the CK test. State 4 exhibited the highest deviation between the predictions of the Bayesian MSM and the actual estimates. Therefore, the dispersed nature of State 4 in the observed states was somewhat expected.

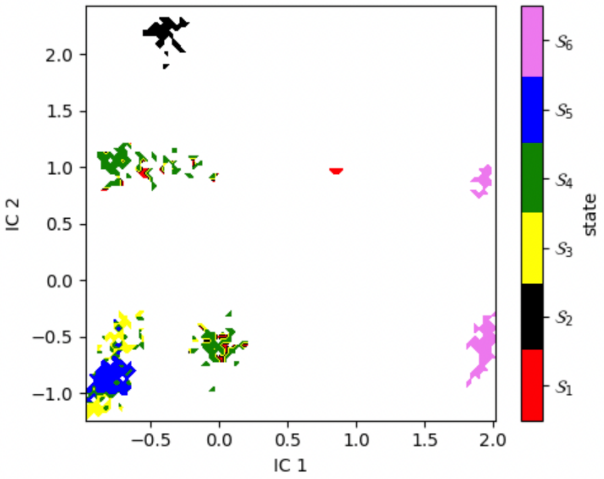

Figure 8. Metastable State Map of Microstates

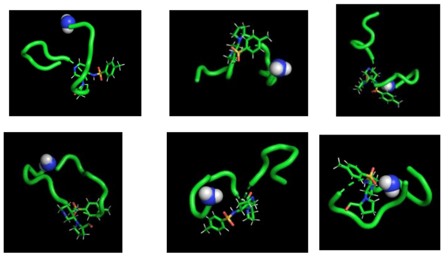

Figure 9. Visualization of Sampled Molecular Dynamics Trajectories from Metastable States with PyMol

4. Discussion

In our analysis, we conducted Bayesian MSM estimation and identified 6 discrete states. We derived the stationary distribution and computed the relative free energy for each S_i state:

$$G_{Si} = -k_B T \ln\left(\sum_j \pi_j\right) \tag{5}$$

As anticipated, visual inspection of the density plot suggested that the 4th state might be unstable. With quantitative results, our findings confirm this observation. Referring to Table 1, we observe that State 4 exhibits the highest stationary probability, indicating that the system predominantly occupies this state. In addition, state 4 has the lowest

State	π	G/k_bT
1	0.109854	2.208604
2	0.055085	2.898876
3	0.170944	1.766418
4	0.314029	1.158271
5	0.154797	1.865644
6	0.195292	1.633262

Table 1. Stationary distribution and free energy values for each state

relative free energy Higher probabilities and lower free energies typically imply easier transitions to and from a given state.

From the obtained MSM, we quantified Mean First Passage Times (MFPTs), representing the average duration for a system to transition between states. Lower MFPT values indicate more frequent transitions and potentially more biologically significant pathways. Reviewing table 2, we observe that diagonal values are zero, indicating no time is required for a state to transition to itself, as expected. Furthermore, the matrix lacks symmetry around the diagonal. For example, the MFPT from State 1 to State 2 (99.60 ns) significantly exceeds that from State 2 to State 1 (20.83 ns), highlighting an asymmetry in transition kinetics between these states.State 4 exhibits relatively low MFPT values to other states, suggesting it might represent a metastable state with multiple transition pathways.

States	1	2	3	4	5	6
1	0.00	99.60	8.27	8.89	15.03	27.28
2	20.83	0.00	20.56	13.48	25.97	22.27
3	12.64	104.50	0.00	9.16	15.73	28.00
4	15.30	101.39	12.49	0.00	13.23	34.45
5	15.69	104.77	12.01	6.44	0.00	31.84
6	17.78	83.06	13.89	18.50	23.21	0.00

Table 2. Mean First Passage Times (MFPT) in nanoseconds

5. Future Work

In future work, there are opportunities for significant improvements based on the insights gained from this analysis. Two primary challenges were encountered during the study: firstly, the computational calculations were time-consuming, and secondly, selecting the optimal parameters for analysis posed a challenge. With additional time and resources, exploring various techniques for dimensionality reduction, clustering methods, and conducting hyperparameter tuning could enhance the accuracy and efficiency of the analysis.

Furthermore, delving deeper into the dynamics of State 4, identified as the least stable, could yield valuable insights. Understanding the factors contributing to its instability could provide critical information for refining the model and improving overall system understanding.

References

[1] M Bollati, K Peqini, L Barone, C Natale, M Beeg, M Gobbi, L Diomede, M Trucchi, M de Rosa, and S Pellegrino. Rational design of a peptidomimetic inhibitor of gelsolin amyloid aggregation. *International journal of molecular sciences*, 23(22):13973, 2022.

[2] H Hata, D Phuoc Tran, M Marzouk Sobeh, and A Kitao. Binding free energy of protein/ligand complexes calculated using dissociation parallel cascade selection molecular dynamics and markov state model. *Biophys Physicobiol*, 18:305–316, Dec 2021.

[3] Mariam Hergnyan. Molecular simulations exam repository, 2024.

[4] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd - visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[5] Andreas Mardt, Luca Pasquali, Hao Wu, and et al. Vampnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5, 2018.

[6] Jan-Hendrik Prinz, Bettina G. Keller, and Frank Noé. Probing molecular kinetics with markov models: Metastable states, transition pathways and spectroscopic observables. *Physical Chemistry Chemical Physics*, 13(38):16912–27, 2011.

[7] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11:5525–5542, Oct. 2015.

[8] Schrodinger, LLC. The pymol molecular graphics system, version x.x, 2010.

[9] Steffen Schultze and Helmut Grubmüller. Time-lagged independent component analysis of random walks and protein dynamics. *Journal of Chemical Theory and Computation*, 17(9):5766–5776, 2021.