Low Dimensional Approximation of Protein Dynamics

Julius Neumann

Stanford University, njulius@stanford.edu

March 27, 2025

1 Introduction

"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

-Richard Feynman [1]

Understanding the motion of proteins would help in understanding a wide variety of biological processes on the molecular level. Some examples include the misfolding and aggregation of proteins in Alzheimer's disease and the catalysis of chemical reactions by enzymes. Experimental instruments cannot yet take a detailed video of how a protein moves over time because of its incredibly small size, but its average structure over a longer period of time can be determined using techniques such as X-ray crystallography or cryo-electron microscopy (cryo-EM). Then, starting from this average structure, the motion of the protein over time can be simulated based on the physics of electric attractions and repulsions between atoms. This technique is called molecular dynamics (MD) simulation. [1]

A typical MD simulation works in time steps on the order of femtoseconds. In each time step, the forces acting on each atom are calculated and used to update the velocity and position of each atom. This loop is usually repeated several billion times for a total simulation time on the order of microseconds. The 3D coordinates of each atom are saved every hundred to thousand time steps (approx. every picosecond) as a simulation frame. These frames can be loaded into a visualization software such as VMD (Visual Molecular Dynamics) and played as a video showing the motion of molecules over time.

2 Data

D.E. Shaw Research, a molecular dynamics research company, has made several of their simulation trajectories involving proteins from SARS-CoV-2 publicly available on their website. [2] From here I obtained the PSF, PDB, and DCD files for an MD simulation of the PLPro domain of the viral protein nsp3 (DESRES-ANTON-11730054). The PLPro domain is essential for the replication of the coronavirus.

I used the first 5000 frames of simulation data. The time step was 0.04888821 ps, and hence the total simulation time was about 244.4 ps.

In order to reduce the computation time for the SVD, I worked with a subset of the total number of atoms in the protein. In particular, I extracted the coordinates for the alpha carbon atoms, which form the backbone of the protein. While the entire protein has 4922 atoms, there are only 317 alpha carbon atoms.

3 Methods

3.1 Theory

Suppose that an MD simulation has n time steps and N atoms, and hence $m = 3N$ different atomic coordinates, since we have x, y, and z coordinates for each atom. Then the simulation data can be represented by an $m \times n$ matrix X:

$$X = \begin{bmatrix} x_1(t_1) & x_1(t_2) & \dots & x_1(t_n) \\ y_1(t_1) & y_1(t_2) & & y_1(t_n) \\ z_1(t_1) & z_1(t_2) & & z_1(t_n) \\ \dots & \dots & & \dots \\ x_N(t_1) & x_N(t_2) & \dots & x_N(t_n) \\ y_N(t_1) & y_N(t_2) & & y_N(t_n) \\ z_N(t_1) & z_N(t_2) & & z_N(t_n) \end{bmatrix} = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_n} \end{bmatrix}$$

Each column $\mathbf{x_i}$ of X describes the structure of the entire protein at a particular point in time.

The time-averaged structure of the protein (the average position of each atom over the course of the simulation) can be calculated to taking the

mean of each row of X, and using these row means to build an m-vector $\bar{\mathbf{x}}$.

We can subtract the the mean of each row from X to obtain a matrix A that describes the movement of each atom around its average position over the course of the simulation:

$$A = \begin{bmatrix} x_1(t_1) - \bar{x}_1 & \dots & x_1(t_n) - \bar{x}_1 \\ y_1(t_1) - \bar{y}_1 & & y_1(t_n) - \bar{y}_1 \\ z_1(t_1) - \bar{z}_1 & & z_1(t_n) - \bar{z}_1 \\ \dots & & \dots \\ x_N(t_1) - \bar{x}_N & \dots & x_N(t_n) - \bar{x}_N \\ y_N(t_1) - \bar{y}_N & & y_N(t_n) - \bar{y}_N \\ z_N(t_1) - \bar{z}_N & & z_N(t_n) - \bar{z}_N \end{bmatrix}$$

$$= X - \bar{\mathbf{x}}\ \mathbf{1}^T = \begin{bmatrix} \mathbf{x_1} - \bar{\mathbf{x}} & \mathbf{x_2} - \bar{\mathbf{x}} & \dots & \mathbf{x_n} - \bar{\mathbf{x}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{bmatrix}$$

Each column $\mathbf{a_i}$ of A represents the displacement of each atom from its average position at a particular time point during the simulation.

We proved in lecture that every real matrix has a singular value decomposition (SVD). Hence we can write A in the following form:

$$A = U\ \Sigma\ V^T$$

where U is an $m \times m$ matrix with orthonormal columns, Σ is an $m \times n$ matrix with zero entries everywhere except possibly the diagonal, and V_T is an $n \times n$ matrix with orthonormal columns.

For our purposes we assume that $n > m$, meaning that we have more time steps than atomic coordinates, which we can achieve by running the simulation long enough. In this case we can write the SVD in a shorter form as follows (this is the reduced SVD if $\text{rank}(A) = m$, which it most likely will be, since there is no linear dependence relationship between the rows of A):

$$A = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U\ \Sigma_1\ V_1^T$$

where V_1^T is $m \times n$, V_2^T is $m \times (m-n)$, and Σ_1 is $m \times m$. Since they get multiplied by zero, we can ignore the last $(m-n)$ rows of V^T.

For each of the m singular values σ_i, we have a corresponding column vector $\mathbf{u_i}$ in U (left singular vector of A) and a corresponding row vector $\mathbf{v_i^T}$ in V^T (right singular vector of A).

Each $\mathbf{u_i}$ consists of a list of atomic coordinates and specifies a 3-vector for each atom in the protein. Collectively, all these 3-vectors form a "mode of motion" of the protein. In one specific mode, each atom is only allowed to move along a single direction in 3-D space. The modes with the highest singular values are most important to the overall motion of the protein.

Each $\mathbf{v_i^T}$ consists of a scalar function of time and represents the projection of the protein's motion onto the mode given by $\mathbf{u_i}$. Since the mode of motion specifies a direction of motion for each atom, $\mathbf{v_i^T}$ gives how far along its direction each atom has moved at each time point during the simulation.

We can write A as a sum of m rank 1 matrices, each involving one mode of motion of the protein:

$$A = \sum_{i=1}^{m} \sigma_i \mathbf{u_i} \mathbf{v_i^T}$$

Since the singular values are ordered from highest to lowest, the top singular values matter the most, and the matrix A can be approximated by considering only the first k singular values, and setting the rest to 0. This produces a rank k approximation for A.

$$A \approx \hat{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i^T} \qquad k < m$$

Carl Eckart and Gale Young proved in a 1936 paper [3] that $\hat{A}_k$ is the best rank k approximation for A, i.e. it minimizes the distance of the approximation to A with respect to the Frobenius norm. For any $m \times n$ matrix B with $\text{rank}(B) = k$, we have

$$\|\hat{A}_k - A\|_F \leq \|B - A\|_F$$

Furthermore, Eckart and Young showed that the distance of $\hat{A}_k$ to A can be computed in terms of the singular values of A:

$$\|\hat{A}_k - A\|_F^2 = \sum_{i=k+1}^{m} \sigma_i^2 = \sum_{i=1}^{m} \sigma_i^2 \ - \ \sum_{i=1}^{k} \sigma_i^2$$

$$= \|A\|_F^2 \ - \ \|\hat{A}_k\|_F^2$$

$\|A\|_F^2$ can be interpreted as the sum of the squared deviations of each atom's exact position from its average position. For a simulation trajectory in which the protein is completely still $A = 0$ and $\|A\|_F^2 = 0$. In fact, the rank 0 approximation to A would be the zero matrix, and the rank 0 approximation to the protein's motion is simply a still frame of its average structure.

The rank 1 approximation to the protein's motion would consist of the average structure plus the motion of each atom along a single direction. The rank 2 approximation would include 2 directions, and the rank k approximation would include k directions, each with an independent path of motion along that direction. As we increase k, we increase the number of one-dimensional paths that we add up, producing better and better approximations to the protein's motion.

$\|\hat{A}_k - A\|_F^2$ describes how far off the rank k approximation is and can be interpreted as the

sum of the squared deviations of each atom's approximated position from its exact position. Furthermore, we can calculate the proportion of the squared deviations in A that is *not* explained by $\hat{A}_k$ as:

$$\frac{\|\hat{A}_k - A\|_F^2}{\|A\|_F^2} = \frac{\|A\|_F^2 - \|\hat{A}_k\|_F^2}{\|A\|_F^2} = 1 - \frac{\|\hat{A}_k\|_F^2}{\|A\|_F^2}$$

Hence the proportion of the squared in A that *is* explained by $\hat{A}_k$ is:

$$\frac{\|\hat{A}_k\|_F^2}{\|A\|_F^2} = \frac{\sum_{i \le k} \sigma_i^2}{\sum_i \sigma_i^2}$$

This can be interpreted as the proportion of the protein's motion that is accounted for by the rank k approximation.

The application of SVD to protein dynamics is described in more detail by Romo et. al. in their 1995 publication titled "Automatic Identification of Discrete Substrates in Proteins: Singular Value Decomposition Analysis of Time-Averaged Crystallographic Refinements." [4]

3.2 Computational Tools

I used the following Python libraries:

- MDAnalysis to read from and write to molecular dynamics file formats such as PSF, PDB, and DCD
- NumPy to manipulate matrices and calculate the SVD
- MatPlotLib to make simple 2-D plots

I also used VMD (Visual Molecular Dynamics) to produce 3-D visualizations of the simulation trajectory and its low rank approximation.

4 Results and Discussion

I used SVD to analyze the motion of all 317 alpha carbon atoms in the protein PLPro over 5000 simulation frames. Since there are 3 coordinates (x, y, and z) for each atom, each frame had a total of 951 coordinates. The matrix A was a 951×5000 matrix.

After calculating the SVD, I found that most of the 951 singular values are relatively low (see Figure 1). This means that the first few modes of motion are the most important to the motion of the protein, and that the lower modes of motion are just noise.

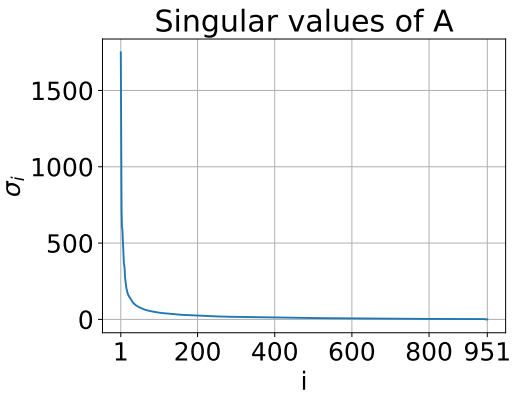

Figure 1: The singular values $\sigma_1, \sigma_2, ..., \sigma_m$ of A where $m = 951$.

I also calculated the proportion of the squared deviations in A that are explained by the rank k approximation $\hat{A}_k$, which corresponds to the proportion of the variation of the protein structure over the course of the simulation that is explained by the rank k approximation to the simulation trajectory. I plotted this quantity as a function of k in Figure 2 and the numerical values for the first 10 k's are shown in Table 1. One can see that higher rank approximations are better, and that the most improvement comes with the first few singular values.

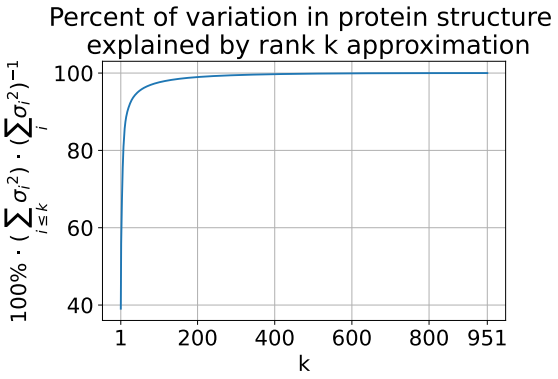

Figure 2: The percent of squared deviations in A that are explained by the rank k approximation $\hat{A}_k$. This is a measure for how well $\hat{A}_k$ approximates A.

k	$100\% \cdot (\sum_{i \le k} \sigma_i^2) \cdot (\sum_i \sigma_i^2)^{-1}$
1	39.06 %
2	55.88 %
3	62.55 %
4	67.25 %
5	71.73 %
6	75.48 %
7	78.60 %
8	80.94 %
9	82.63 %
10	84.24 %

Table 1: The percent of variation in protein structure explained by the first 10 low rank approximations.

Finally, I took a closer look at the rank 1 approximation of A:

$$\hat{A}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$$

As described in the Theory section under Methods, $\mathbf{u}_1$ is a vector in $\mathbb{R}^m = \mathbb{R}^{951}$ and specifies a direction of motion for each atom. $\mathbf{v}_1^T$ is a vector in $\mathbb{R}^n = \mathbb{R}^{5000}$ and describes the path of each atom along its direction of motion, i.e. the projection of the protein's motion onto the directions specified by the top mode of motion $\mathbf{u}_1$. I plotted this path as a function of time in Figure 3.

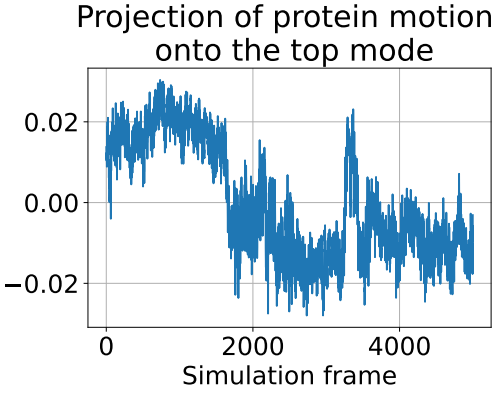

Figure 3: The motion of each atom along the single direction specified by the top mode of motion (the mode with the highest singular value).

To obtain the matrix A, I had to subtract off the average structure of the protein. Therefore, to obtain a rank 1 approximation $\hat{X}_1$ for the trajectory of the protein, I need to add the average structure of the protein to the rank 1 approximation of A:

$$\hat{X}_1 = \hat{A}_1 + \bar{\mathbf{x}}\ \mathbf{1}^T$$

I used VMD (Visual Molecular Dynamics) to create a video of the exact motion as well as the rank-1-approximated motion of the 317 alpha carbons in the protein. The links to the videos are below:

- Exact Trajectory
- Rank 1 Approximation
- Comparison of Exact Trajectory and Rank 1 Approximation

Figure 4 shows 3 representative frames comparing the structure of the protein from the exact trajectory and from the rank 1 approximation.

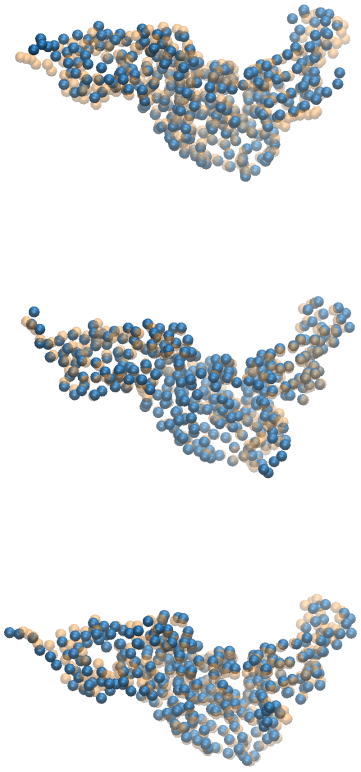

Figure 4: From top to bottom: frame 1, frame 2500, frame 5000. The alpha carbon atoms from the exact trajectory are shown as blue spheres (darker), and the alpha carbon atoms from the rank 1 approximation are shown as partially transparent orange spheres (lighter).

In the video for the rank 1 approximation to the protein's motion, we can see that each atom is only moving along a single direction. Interestingly, atoms that are close together in the protein move in similar directions. The protein could be

divided into several groups of atoms that tend to move as a group.

Even though the rank 1 approximation only captures 40% of the protein's motions (Table 1), it still leads to a decent approximation of the protein's trajectory, and higher rank approximations would be even better. With just the top 10 out of 951 modes of motion, we can describe 84% of the protein's motions (Table 1). This type of low dimensional approximation could potentially be used to compress the amount of storage space needed for a molecular dynamics trajectory. We do not need to store every single coordinate of every single atom at each time point to describe the most important features of a protein's motion.

References

[1] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nat. Struct. Mol. Biol.*, vol. 9, pp. 646–652, Sept. 2002. Number: 9 Publisher: Nature Publishing Group.

[2] D. E. S. Research, "Molecular dynamics simulations related to sars-cov-2," pp. 646–652, 2020. D. E. Shaw Research Technical Data.

[3] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, pp. 211–218, Sept. 1936.

[4] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Phillips Jr, "Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements," *Proteins: Structure, Function, and Bioinformatics*, vol. 22, no. 4, pp. 311–321, 1995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340220403.