Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

Constructing Ubiquitin's backbone by Linear Algebra technique

Introduction:

Protein structure is a fundamental of modern biology because its structure defines its biological function. Proteins must fold into specific three-dimensional structures in order to carry out essential functions such as catalyzing biochemical reactions, cell signal transmission, or providing structural stability within cells (Alberts et al., 2002). Structural prediction allows scientists to investigate such functions, especially for those proteins that are not experimentally defined. Traditional methods for predicting protein structures, such as X-ray crystallography and cryo-electron microscopy, are expensive and not feasible in all cases. Computer prediction is a scalable solution. Developments over the past few years, like AlphaFold, have come near to experimental precision in predicting structure from amino acid sequence alone and changed the way researchers tackle structural biology (Jumper et al., 2021). Structural accuracy guides drug discovery, disease mechanism, and synthetic biology design by providing atomistic insight into protein function. Protein structure prediction is thus a computational challenge and the gateway to understanding life at the molecular level.

This project aims to analyze the confrontation of ubiquitin by applying linear algebra techniques to its atomic coordinate data. Ubiquitin is a small protein that plays a central coordinating role in an extensive range of cellular processes, primarily protein breakdown. It covalently attaches to substrate proteins as tags via a mechanism known as ubiquitination, marking them for degradation by the 26S proteasome, a protein complex that degrades damaged or misfolded proteins. This tagging is vital in the maintenance of protein homeostasis, control of cell cycle progression, DNA repair, and signal transduction (Hershko & Ciechanover, 1998). Because of its central role in cellular quality control and signal transduction, ubiquitin is highly conserved across eukaryotic species, indicating an evolutionary importance.

Methods:

Proteins like ubiquitin undergo subtle but functionally important conformational changes, and quantifying these changes is a significant challenge to analysis. Traditional visualization methods can display protein structures but have

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346
limited ability for precisely defining motion, rotation, and flexibility in mathematical terms. This project addresses that limitation by applying linear algebra techniques to describe and interpret the backbone structure of ubiquitin as vectors and matrices in higher dimensional space. A fundamental idea is the transformation matrix, which through which structures can be systematically rotated, translated and scaled. In protein structure, rotation matrices are used to simulate how a structure moves when rotated around different axes without changing the distances and angles—something crucial when simulating conformational change. Singular Value Decomposition (SVD) is another crucial technique that breaks a matrix down into three components and indicates the most important directions of variance in the data. This is quite analogous to Principal Component Analysis (PCA), a method of dimensionality reduction of high-dimensional data to lower dimensions by finding the axes (principal components) upon which the data shifts most. PCA is particularly useful for structural variation in proteins since it highlights important conformational trends. Finally, analysis of eigenvalue and eigenvector typically done on a covariance matrix displays the directions (eigenvectors) and amplitude (eigenvalues) of shifting or variation. In structural biology, such vectors most often describe biologically meaningful movements such as domain motion or backbone motion.

Result:

We first started the analysis by using the Biopython Bio.PDB to load the 3D structure of the ubiquitin protein from a PDB file (1ubq.pdb). Then the structure was parsed, and backbone atoms, specifically the nitrogen (N), alpha carbon (CA), and carbon (C) atoms, were extracted from each amino acid residue in chain A by using the PDBParser. These atoms define the protein backbone and form the structural framework of the entire polypeptide chain. The backbone determines the overall 3D shape of the protein and serves as the anchor for side chains that influence function (Branden & Tooze, 1999). This approach ensures that the data used for geometric and linear algebra analysis reflect the true spatial conformation of the protein’s backbone without interference from variable side chains.

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

N,CA,C
[[27.34 24.43 2.614] [... [40.031 39.992 35.432]]

Fig1: 3D coordinates of the backbone atoms (N, CA, C) from the ubiquitin protein. Each row represents one atom's position in space

In Figure 1, we successfully extracted structural information, with the matrix containing the 3D coordinates of the backbone atoms (N, CA, C) from the ubiquitin protein. Each row corresponds to one atom, detailing its spatial coordinates in three-dimensional space. Before proceeding with downstream analysis, we chose to mean-center the data. It involves subtracting the mean position vector, calculated across all atoms, from each coordinate. Mathematically, this operation repositions the entire point cloud, aligning its geometric center with the origin.

Let X$\in$ $R^{n\,x\,3}$ represent the original coordinate matrix, where each row corresponds to a 3D position (x,y,z) of a backbone atom in the protein. The number of rows nnn reflects the total number of atoms considered. Let $\mu \in R^{1x3}$ be the mean vector, which contains the average $\bar{x}$, $\bar{y}$, $\bar{z}$ values across all atoms. To perform mean-centering, we subtract this mean vector μ from every row of the matrix X. We created a column vector of $1_n \in R^{nx1}$and multiplying it by μ, resulting in a full matrix where the mean vector is repeated for each atom.

The Mean-centering: $X_{center} = X - 1_n . \mu$

The resulting matrix then reflects only the relative variations among the atoms, eliminating any translational offsets. This step helps to eliminate translational bias and ensures that the analysis focuses solely on the relative positions and variations between atoms. Techniques such as PCA, SVD, and covariance analysis are sensitive to the coordinate system's origin. If the data is not centered, these methods might highlight directions of variance based on the protein's overall spatial position rather than its internal structural variations. By shifting the data so that its mean is at the origin, we ensure that the results reflect how the structure varies, rather than its location in space.

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

Then we applied singular Value Decomposition (SVD) to uncover the principle directions of structural variation within the ubiquitin backbone. In linear algebra, SVD factorizes the centered matrix $X \in R^{nx3}$into three component such that

X= $U\Sigma V^T$	U: left singular vectors associated with the atoms Σ: diagonal matrix - quantity captured how much variances is captured along each axis V^T: right singular vector revealing the dominant geometric directions in which the protein varies

SVD - Left Singular Vectors (U):
```
[[-0.09656212  0.06173004  0.06046915 ...  0.11789494  0.11192171
  0.10960832]
 [-0.09587613  0.06865751  0.0437778  ...  0.07675056  0.09253601
  0.09700049]
 [-0.08397631  0.07325414  0.04106664 ...  0.13227403  0.13284059
  0.14931743]
 ...
 [ 0.17832541 -0.06541389 -0.03623004 ...  0.96870816 -0.03167932
 -0.03360358]
 [ 0.18277311 -0.05545476 -0.04917058 ... -0.03180533  0.96752185
 -0.03449582]
 [ 0.18907605 -0.06511924 -0.06174331 ... -0.03351612 -0.03426667
  0.96344775]]
```

SVD - Singular Values (Σ):
```
[124.14365   88.883446  80.05192 ]
```

SVD - Right Singular Vectors (V^T):
```
[[ 0.57042986  0.57349575  0.5879731 ]
 [-0.04326601  0.7358521  -0.6757586 ]
 [ 0.820206   -0.36003363 -0.4445649 ]]
```

Figure 2: Formula for singular decomposition value and the result for $U\Sigma V^T$

Singular Value Decomposition (SVD) was performed on the mean-centered 3D coordinate matrix of the ubiquitin backbone to identify its principal modes of structural variation(fig2). The decomposition yielded three singular values: 124.14, 88.88, and 80.05, corresponding to the amount of variance captured along the first, second, and

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346
third principal directions, respectively. These values indicate that the majority of structural variability is concentrated along the first principal axis. The right singular vectors (V^T) define the orientation of these principal directions in 3D space, with the first vector [0.570,0.573,0.588] representing the dominant direction of motion. This suggests a consistent structural trend along a diagonal trajectory across the protein's spatial configuration. The left singular vectors (U) describe the contribution of each backbone atom to these principal components, allowing for the identification of residues involved in significant conformational shifts.

To validate and complement the findings from SVD, Principal Component Analysis (PCA) was also applied to the same mean-centered atomic coordinates using scikit-learn's PCA implementation. Limiting the analysis to three components, PCA produced a set of orthogonal vectors stored in pca.components_, each representing a distinct direction of structural variation in 3D space. The first principal component captured the largest proportion of the variance, confirming its role as the dominant axis of motion. The second and third components captured orthogonal directions of lesser, yet meaningful, variability. Notably, the directions identified by PCA aligned closely with those from SVD, reinforcing the robustness of the structural patterns observed.

In this activity, we use a matrix X ∈ $R^{n\times 3}$ as an input, where each row represents the Cartesian coordinates (x,y,z) of a backbone atom (N,CA, or C) along chain A of ubiquitin. As previously mentioned, we already already performed the mean-center to remove the translation bias, we computed the covariance matrix:

$$C = \frac{1}{n-1} X^T_{centered} X_{centered}$$

This 3 x 3 matrix summarizes how atomic displacements along the X,Y and Z dimensions co-vary across the backbone. Performing eigen-decomposition on the covariance matrix yields three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ and corresponding eigenvectors v_1, v_2, v_3, where: $Cv_i = \lambda_i v_i$

Covariance Matrix Eigenvalues: `[67.89271147 28.23044046 34.8029397 ]`
Covariance Matrix Eigenvectors:

```
[[ 0.57042987  0.82020596  0.04326601]
```

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

```
 [ 0.57349577 -0.36003363 -0.73585215]
 [ 0.5879731  -0.44456492  0.67575859]]
```

Each eigenvector $v_i \in R^3$ defines a principal axis of structural variation, and its associated eigenvalue λ_i quantifies the variances along the

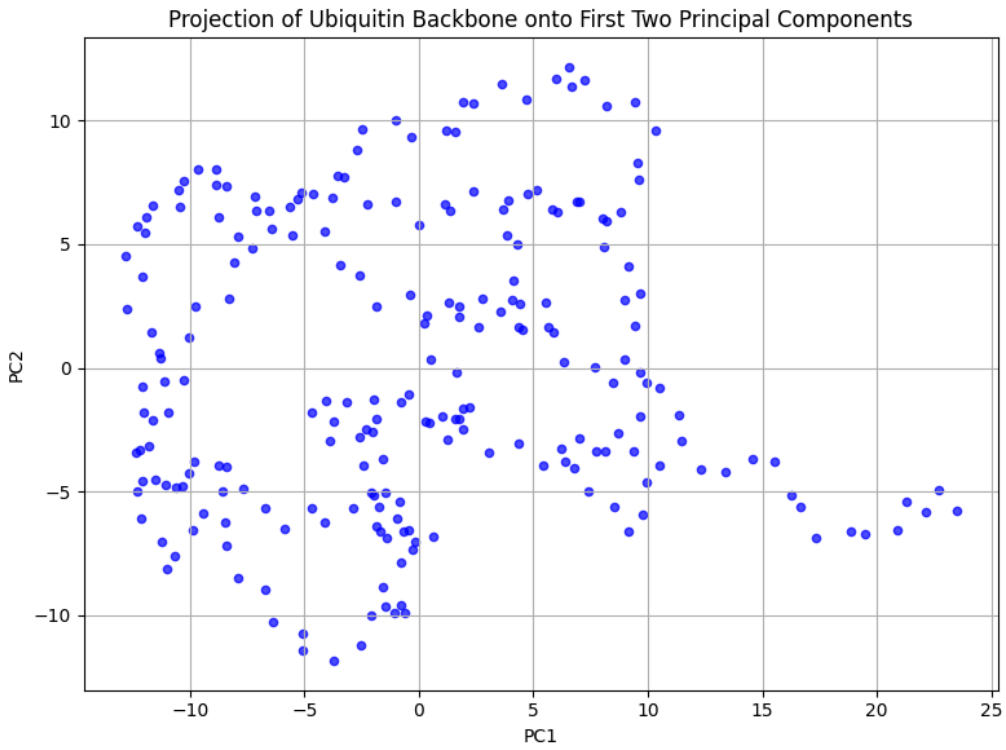


Figure 3: Projection of Ubiquitin Backbone onto First Two Principal Components: Each point in the scatter plot represents a backbone atom (N, CA, or C), with its position determined by the atom's loading onto the first (PC1) and second (PC2) principal axes.

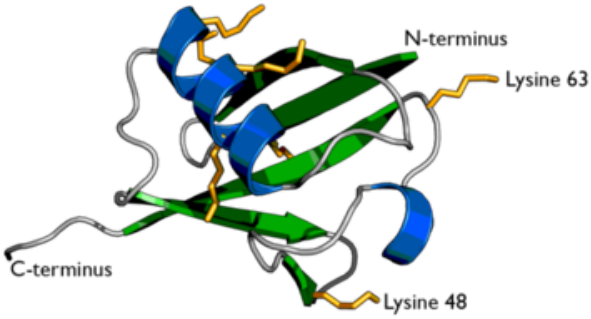
The structural distribution along the first two principal components was visualized in a 2D scatter plot (Figure 3). Each point represents a backbone atom (N, CA, or C), projected onto the PC1–PC2 plane. The plot reveals a smooth, continuous trajectory, reflecting a gradual spatial progression along the backbone. The spread along PC1 indicates the primary conformational trend, while variation along PC2 highlights additional flexibility orthogonal to this motion. This visualization confirms that the

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346
backbone variability in ubiquitin is structured and biologically relevant, reflecting underlying folding mechanisms and dynamic behavior.

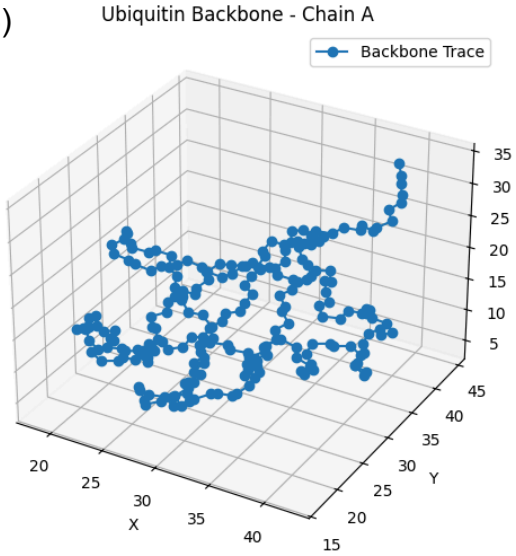
To further examine the structural variability of the ubiquitin backbone, we computed the covariance matrix from the mean-centered 3D coordinate data. This matrix quantifies how variations along the x, y, and z axes are correlated across all atoms. Eigen-decomposition was then performed on the covariance matrix to extract the eigenvalues and eigenvectors, which describe the magnitude and direction of structural variation, respectively. The eigenvectors represent the principal axes of flexibility, while the eigenvalues indicate how much variance each axis explains. The direction with the largest eigenvalue corresponds to the dominant mode of structural motion, aligning with the first principal component from PCA and the first right singular vector from SVD, thereby validating the consistency of the findings across all linear algebra methods.

To visualize the raw spatial configuration of the ubiquitin backbone, a 3D trace of the atomic positions was plotted (Figure 4). The trace reflects the ordered progression of N, CA, and C atoms along the chain, revealing the overall fold and compactness of the structure. When combined with the eigen-decomposition results, this visualization provides a spatial reference to interpret how the protein's shape varies along key motion axes.

A)



B)



Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

Figure 4. Ubiquitin Structural Visualization.

(A) Cartoon representation of the 3D structure of ubiquitin, highlighting key residues such as Lysine 48 and Lysine 63, which play crucial roles in polyubiquitination. (B) Computational reconstruction of the ubiquitin backbone using atomic coordinate data (N, CA, and C atoms) from chain A, plotted in 3D Cartesian space. This backbone trace reflects the spatial arrangement of residues and forms the basis for subsequent linear algebra analyses such as PCA, SVD, and eigen-decomposition.

Figure 4A and Figure 4B are two complementary representations of the exact same ubiquitin backbone—one viewed through a biological lens, the other through a mathematical one. In Figure 4A, the full cartoon rendering displays secondary structure elements, side chains, and key residues (e.g., Lys48 and Lys63), providing rich context for understanding ubiquitin's functional sites and fold. Figure 4B, by contrast, distills this information down to the raw Cartesian trace of just the backbone atoms (N, CA, C), isolating the protein's geometric skeleton. Although stripped of side-chain detail and ribbon diagrams, the backbone trace in 4B still faithfully mirrors the overall fold and compactness seen in 4A—confirming that the abstracted point cloud preserves the true spatial conformation. This streamlined view is precisely what enables rigorous linear algebraic analyses (SVD, PCA, covariance eigen-decomposition), since it removes extraneous annotation and focuses solely on the geometry. In essence, 4B is a mathematical subset of 4A: both derive from the same coordinate data, but each is tailored to a distinct purpose—biological interpretation versus quantitative structural analysis.

Conclusion

In this project, we successfully applied linear algebra techniques, including Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and eigen-decomposition of the covariance matrix, to analyze the structural variability of the ubiquitin protein backbone. By mean-centering the atomic coordinate data and systematically examining the principal directions of variance, we identified dominant modes of structural flexibility that align with known biological behavior. The consistency

between SVD, PCA, and covariance analysis validates the robustness of the findings, highlighting how mathematical tools can meaningfully describe complex biomolecular structures. Visualizations such as the 3D backbone trace and principal component projections provided intuitive insights into the geometric organization and dynamic potential of the protein.

Importantly, we successfully reconstructed the backbone of ubiquitin from atomic coordinate data, allowing a clear and structured analysis of its spatial conformation using linear algebraic methods. Overall, this study demonstrates that linear algebra provides a robust mathematical foundation for quantitatively analyzing protein structure and dynamics, supporting advancements in computational structural biology.

Limitations

While linear algebra methods have successfully been developed to analyze divergences in protein structures, there are a number of limitations. Firstly, we only analyzed the backbone atoms (N, CA, and C) and did not include side chains in the analysis; side chains can play an important role in folding and functional dynamics even though they do not affect protein diversity. Because side chains can truly impact a dynamic ensemble of conformations providing additional anchors and instabilities to the backbone, side-chain atoms would inform the analysis. Secondly, we analyzed a static structure using a single PDB file, rather than analyzing proteins in nature as dynamic ensembles. If we had utilized multiple conformations, like from molecular dynamics simulations, we could have provided a more reasonable portrayal of flexibility in proteins. As with a number of linear methods, like PCA and SVD, assuming that if there is a main mode of linear variation, the variation follows the linear relation of various data points, potentially excluding important nonlinear representation of protein's motion. Finally, if not uniformly controlled across each dataset, the common technical decisions we made, such as the mean-centering and normalizing the mean separately, could introduce biases to our conclusions about PCA or other scores in comparison. Future work could tackle these issues by using dynamic datasets of ensemble relationships, applying non-linear dimensionality reduction approaches, and chemistry of a wider atomic range.

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

Work References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Molecular Biology of the Cell (4th ed.). Garland Science.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

Code/ Python

```
import numpy as np
import Bio.PDB
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Load a sample PDB file (1ubq is a common small protein)
pdb_parser = Bio.PDB.PDBParser(QUIET=True)
structure = pdb_parser.get_structure("protein", "/Users/nguyenkim/Desktop/Applied
linear algebra/1ubq.pdb")

import numpy as np
import Bio.PDB
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Load PDB structure
pdb_parser = Bio.PDB.PDBParser(QUIET=True)
structure = pdb_parser.get_structure("protein", "/Users/nguyenkim/Desktop/Applied
linear algebra/1ubq.pdb")

# Extract backbone atom coordinates (N, CA, C from each residue in Chain A)
backbone_atoms = []
for model in structure:
        for chain in model:
        if chain.id == "A":
        for residue in chain:
                if Bio.PDB.is_aa(residue):
                for atom_name in ["N", "CA", "C"]:
                if atom_name in residue:
                        atom = residue[atom_name]
                        backbone_atoms.append(atom.coord)

# Convert to NumPy array
coords = np.array(backbone_atoms)
print(coords)
```

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

```
# Center coordinates
mean_centered = coords - np.mean(coords, axis=0)
print(mean_centered)
# --- Linear Algebra Analyses ---
# SVD
U, S, Vt = np.linalg.svd(mean_centered)
print("SVD - Left Singular Vectors (U):\n", U)
print("\nSVD - Singular Values (Σ):\n", S)
print("\nSVD - Right Singular Vectors (V^T):\n", Vt)

# PCA
from sklearn.decomposition import PCA
pca = PCA(n_components=3)
pca.fit(mean_centered)
pca_components = pca.components_
import matplotlib.pyplot as plt

# Project the mean-centered data onto the first two principal components
projected = pca.transform(mean_centered)

# Plot the 2D projection (PC1 vs PC2)
plt.figure(figsize=(8, 6))
plt.scatter(projected[:, 0], projected[:, 1], c='blue', s=20, alpha=0.7)
plt.title("Projection of Ubiquitin Backbone onto First Two Principal Components")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.grid(True)
plt.tight_layout()
plt.show()

# Covariance + Eigen-decomposition
cov_matrix = np.cov(mean_centered.T)
eigvals, eigvecs = np.linalg.eig(cov_matrix)

# --- Visualization ---
fig = plt.figure(figsize=(10, 6))
```

Name: *Kimin Wu Nguyen*
EID: VMN377
Class: M 346

```
ax = fig.add_subplot(111, projection='3d')
ax.plot(coords[:, 0], coords[:, 1], coords[:, 2], marker='o', label='Backbone Trace')
ax.set_title("Ubiquitin Backbone - Chain A")
ax.set_xlabel("X")
ax.set_ylabel("Y")
ax.set_zlabel("Z")
ax.legend()
plt.show()

# --- Print Results ---
print("SVD - Singular Values:", S)
print("PCA - Principal Components:\n", pca_components)
print("Covariance Matrix Eigenvalues:", eigvals)
print("Covariance Matrix Eigenvectors:\n", eigvecs)
```