Performance Grading of Clustering Algorithms on Molecular Dynamics Simulations of Proteins

Ephraim Kim, Joshua L. Phillips

School of Science and Technology, Georgia Gwinnett College, Lawrenceville, GA 30043
Department of Computer Science, Middle Tennessee State University, Murfreesboro, TN 37132

ABSTRACT

Computational studies continue to serve an important role in modeling and understanding protein dynamics in biology. Molecular Dynamics (MD) can model the molecular structures of proteins and simulate their motion over nanosecond-microsecond time scales using classical mechanics. MD simulations can reveal insights into the folding process that are beyond present laboratory means. When trajectories of the proteins' motion are generated by MD simulation, machine learning algorithms like k-means, spectral, and subspace clustering help identify the structures and processes that are integral to the folding process, which is challenging to do by eye. We aimed to evaluate the performance of these various algorithms with a special focus on the recent hybrid spectral/subspace method by comparing their normalized mutual information (NMI) scores over cumulative simulation time. Principal Component Analysis (PCA) was performed to visualize the trajectories and their clustering results. The theory of protein dynamics suggests that given an infinite amount of time the sampling space should become increasingly mixed. Algorithms that can still identify distinct structures are better suited for clustering MD data. We found that the hybrid spectral/subspace method delivered the best performance overall, and provided the most conservative estimate of the sampling adequacy.

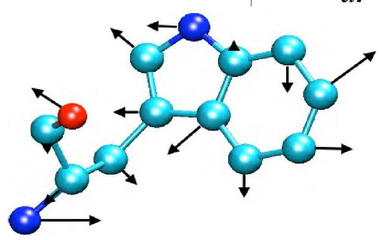
INTRODUCTION and BACKGROUND

MD simulations are a common tool in computational biology for simulating the dynamic behavior of biomolecular structures like proteins. Classical mechanics model the forces acting on the proteins on a molecular level and computers simulate the effects, allowing for exploration of the energy landscape of the protein, and consequently the conformation states of the protein. Of special interest is the protein's native conformation state, which determines the protein's function. This can be important in a myriad of scientific interests including supporting the design of safe and effective drugs, vaccine production, study of neurodegenerative diseases and other biomedical research.

However, while MD simulation is an effective way of studying protein folding, it is computationally expensive. For protein folding that happens beyond the millisecond range, healthy exploration of the possible conformation states can take an unfeasible amount of time (months or years). Protein also get stuck negotiating energy barriers on the way to their native conformations. Consequently, automated methods that can accurately organize protein conformations are especially useful to the computational biologist studying proteins and protein folding. We investigate four clustering algorithms and assess their impact through comparative analysis.

Simulate molecular dynamics using classical mechanics

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t \qquad \boldsymbol{a}(t) = \boldsymbol{F}(t)/m$$
$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\delta t \qquad \boldsymbol{F} = -\frac{d}{d\boldsymbol{r}}U(\boldsymbol{r})$$

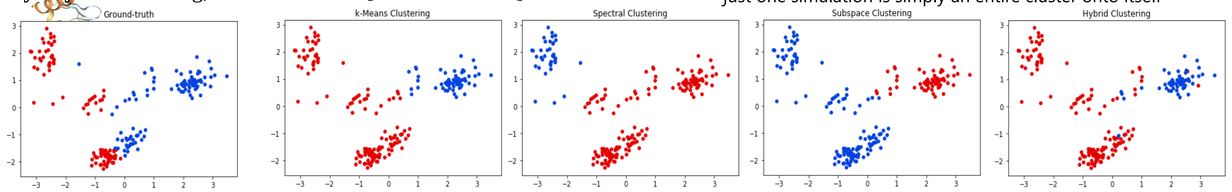


RESULTS


[4]

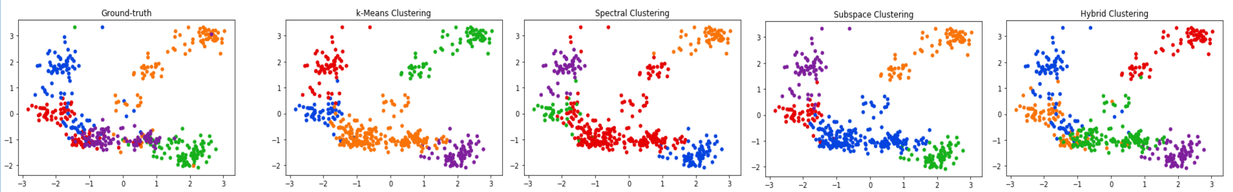
Lysozyme Clustering, first 200 Frames [Simulations 1-2]:

*note: we must start with at least two simulations, since just one simulation is simply an entire cluster onto itself



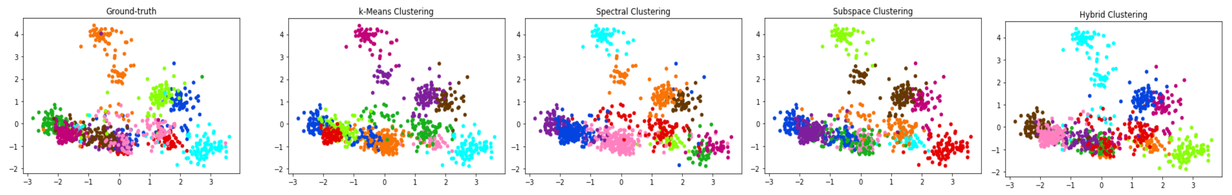
NMI: 0.4467 NMI: 0.2155 NMI: 0.4641 NMI: 0.4056

Lysozyme Clustering, first 500 Frames [Simulations 1-5]:



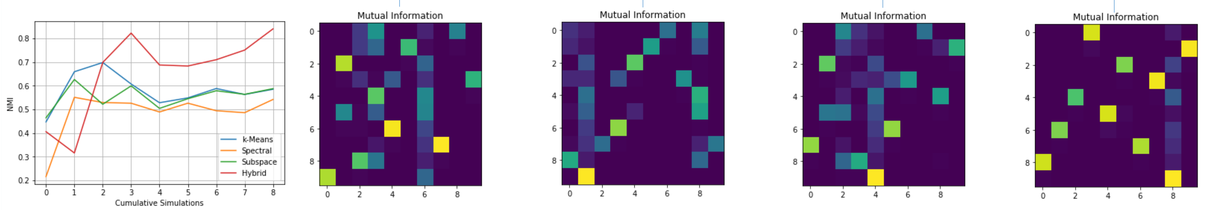
NMI: 0.6072 NMI: 0.4888 NMI: 0.5040 NMI: 0.6877

Lysozyme Clustering, 1000 Frames [Simulations 1-10]:



NMI: 0.5849 NMI: 0.5418 NMI: 0.5876 NMI: 0.8405

Lysozyme Overall NMI distribution + Joint PDFs of MI score [Simulations 1-10]:

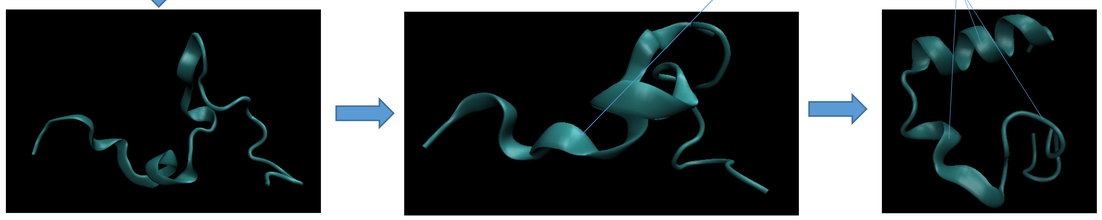


AMBER forcefield

$$V(r^N) = \sum_{\text{bonds}} k_b(l-l_0)^2 + \sum_{\text{angles}} k_a(\theta-\theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1+\cos(n\omega-\gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

VMD VIL Simulation using AMBER Frame 1: Unfolded



VMD VIL Simulation using AMBER Frame 10: Alpha helix forming

VMD VIL Simulation using AMBER Frame 242: Folding complete

METHODS

A total of 10 individual simulations of lysozyme were run. 10010 frames were subsampled down to 100 per 1001 frames of a simulation and then clustered cumulatively using k-means, spectral, subspace, and hybrid spectral/subspace over 10 total runs. PCA on two components enabled plotting 512 dimensional data in 2D. Lemkul's GROMACS[5] tutorial for lysozyme produced the basis for performing the MD runs and generating the trajectory data. VMD[6] made visualizing the trajectories possible. Calculated NMI scores for the four algorithms were the critical basis for grading and comparing algorithm performance. Joint probability distributions assisted in the visualization of mutual information. K-means was performed through sci-kit learn's machine learning library. Hyperparameters for spectral, subspace, and hybrid were manually tuned for performance.

Algorithm 1 Spectral clustering algorithm [1]

procedure SPECTRAL CLUSTERING(S, k) ▷ Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct
Construct a similarity graph by one of the ways described above.
Compute the normalized Laplacians L using equation 1.
Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of L.
Let $V' \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \ldots, v_k$ as columns. Construct matrix $Y' \in \mathbb{R}^{n \times k}$ from V' by normalizing the row sums to have norm 1, that is $y_{ij} = v_{ij}/(\sum_k v_{ik}^2)^{1/2}$.
Cluster the points $(\mathbf{u}_i)$ into clusters $F_1, \ldots, F_k$ with k-means algorithm.
return Clusters $A_1, \ldots, A_k$ with $A_j = \{j | \mathbf{y}_i \in F_j\}$.
end procedure

Algorithm 3 Hybrid Spectral/Subspace clustering algorithm [3]

(1) Compute optimization coefficients C.
(2) Compute affinity matrix S.
(3) Construct matrix $M = S \cdot C$ (SDS) or $M = S * C$ (SES)
(4) Construct graph Laplacians.
(5) Perform singular vector decomposition.
(6) Run k-means algorithm.

Algorithm 2 Sparse Subspace Clustering [2]

procedure SUBSPACE CLUSTERING(S, k) ▷ A set of points $\{\mathbf{x}_i\}_{i=1}^N$ lying in a union of m linear subspaces $\{S_\ell\}_{\ell=1}^m$.
Solve the sparse optimization program.
Normalize the columns of $\boldsymbol{C}$ as $c_i \leftarrow \frac{c_i}{\|c_i\|_\infty}$.
Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of L.
Form a similarity graph with N nodes representing the data points. Set the weights on the edges between the nodes by $\boldsymbol{W'} = |\boldsymbol{C}| + |\boldsymbol{C}|^\top$.
Apply spectral clustering described in Algorithm 1 to the similarity graph $\boldsymbol{W'}$.
return SpectralClustering($\boldsymbol{W'}, k$). ▷ see Algorithm 1
end procedure

DISCUSSION

- For computational biologists, the more accurate the algorithm, the better. The hybrid spectral/subspace method performed quite well, boasting the highest NMI in the most difficult space.
- Sci-kit learn's ordinary k-means algorithm delivered very respectable performance and often beat out spectral and subspace. However, manual hyperparameter optimization may have hid the clustering power of the spectral and subspace algorithms.
- Different forcefields for the same protein may affect the robustness of the clustering algorithms. Further inquiry is required.

REFERENCES

1. Y. Ng, Andrew & Jordan, Michael & Weiss, Yair. (2002). On Spectral Clustering: Analysis and an algorithm. Adv. Neural Inf. Process. Syst. 14.
2. Elhamifar, Ehsan & Vidal, René. (2009). Sparse subspace clustering. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009. 2790 - 2797. 10.1109/CVPR.2009.5206547.
3. Syzonenko, Ivan & Phillips, Joshua. (2018). Hybrid Spectral/Subspace Clustering of Molecular Dynamics Simulations. 325-330. 10.1145/3233547.3233595.
4. Image from the RCSB PDB (rcsb.org) of PDB ID 1DPX (Weiss, M.S., Palm, G.J., Hilgenfeld, R.) (2000) Crystallization, structure solution and refinement of hen egg-white lysozyme at pH 8.0 in the presence of MPD Acta Crystallogr.,Sect.D 56: 952-958
5. http://www.gromacs.org/
6. https://www.ks.uiuc.edu/Research/vmd/

ACKNOWLEDGEMENTS and CONTACT

Contact: Ephraim Kim
email: ekim2@ggc.edu

This work was funded in part by the National Science Foundation
Award #: 1757493
Special thanks to Jeff Payne for the helpful discussions and support.