DEFINING CONFORMATIONAL STATES OF PROTEINS USING DIMENSIONALITY REDUCTION AND CLUSTERING ALGORITHMS

Eugene Klyshko[1,2], Sarah Rauscher[1,2,3]

[1]Department of Physics, University of Toronto, Toronto, ON, Canada; [2]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, ON, Canada; [3]Department of Chemistry, University of Toronto, Toronto, ON, Canada;

ABSTRACT: Molecular dynamics (MD) simulations of proteins produce large data sets - long trajectories of atomic coordinates - and provide a representation of the sampling of a given molecule's structural ensemble. A deep quantitative analysis using advanced machine learning techniques is a means to interpret MD trajectories. To visualize the conformational space of the molecule and properly identify conformational states, we suggest combining clustering methods and dimensionality reduction algorithms. We investigate different choices of features to represent individual structures, clustering algorithms, similarity metric, and methods to assign the number of clusters.

Motivation

Defining *conformational state*s of proteins is a difficult and increasingly common problem:

- Direct observation of transitions between different states is a necessary step for building Markov State Models of proteins.
- Intrinsically disordered proteins (IDPs) lack a unique native three-dimensional structure, and instead have a structural ensemble consisting of many interconverting conformational states. As such, large data sets are used to represent conformational ensembles of IDPs, which create major challenges for researchers in the field.

Unsupervised Machine Learning is a quantitative approach that is used to extract meaningful information from large data sets, however the quality of such analysis highly depends on the proper choice of algorithms and their parameters.

For example, conventional clustering methods used to analyze *molecular dynamics* (MD) trajectories require specification of the number of clusters. This parameter is often unknown in advance and is the quantity of interest itself. The number of clusters is an approximation to number of conformational states since geometrically and energetically similar structures share the same conformational states.

The goal of the project is to develop a methodology to obtain an accurate and easy to use description of the conformational space of a protein from its MD trajectory.

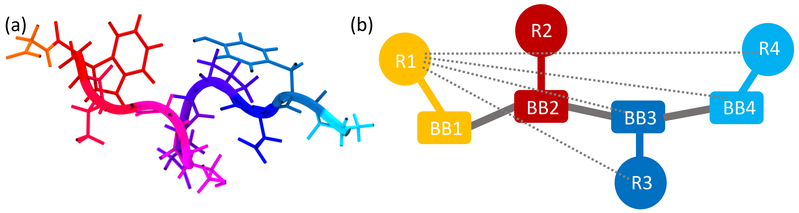

Fig.1: (a) Structure of chignolin used for analysis. Different residues are shown in different colors; **(b)** Construction of feature vector: pairwise distances between non-neighbouring amino acid side chain and backbone atoms are the components of 115-dimensional vector.

1. Choice of features

To validate the approach we performed a blind experiment, where procedure was applied to an MD simulation (with 785,612 frames) of the *chignolin* protein (**Fig. 1a**), for which conformational states are known.[1]

The first step is to investigate the choice of features for analysis. Features are numerical vectors describing each individual protein structure in the ensemble.

- Positions of protein atoms are represented as 3D vectors of x, y, z coordinates.
- Features must be invariant to translations and rotations.
- Features should account for the energy of conformations (i.e. include information about contacts).

The best choice – pairwise distances between backbone and side chain atoms of non-neighbouring amino acids (**Fig. 1b**). Distances were computed between the closest heavy atoms of the residues, which for this 10-residue protein is 115 components.

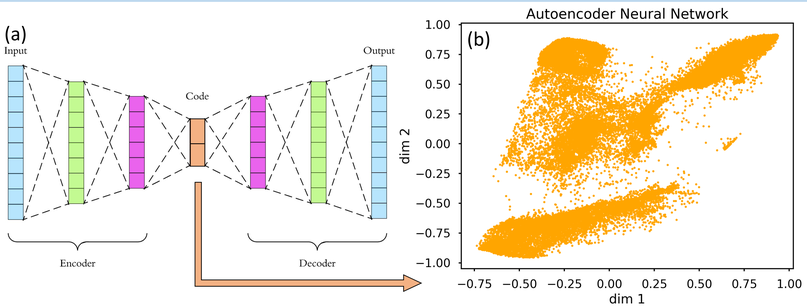

Fig. 2: (a) Architecture[6] of the autoencoder neural network used for dimensionality reduction; **(b)** Two-dimensional embedding of the 115-dimensional conformational space obtained with autoencoder.

2. Dimensionality reduction

To visualize conformational space in two dimensions we advocate the application of dimensionality reduction algorithms, such as:

- *Autoencoders*[2] – a special class of neural networks that take high dimensional vector as input and output. After training, the bottleneck layer is then used for encoding the original high dimensions into low dimensional representation (**Fig.2**).
- *Principal Component Analysis (PCA)* – extracting components with the largest variance of data. The first two principal components combined are responsible for 83% of the variance (**Fig. 3a**).
- *Multidimensional Scaling (MDS)*[3] – mapping objects from high dimensions into low dimensions by conserving their pairwise distances (**Fig. 3b**).
- *t-distributed Stochastic Neighbor Embedding (t-SNE)*[4] – embedding in 2D in such a way that similar objects are modelled by nearby points, and dissimilar by distant points (**Fig. 3c**).
- *Uniform Manifold Approximation and Projection (UMAP)*[5] – is a new manifold learning method built upon mathematical foundations of Laplacian eigenmaps (**Fig. 3d**).

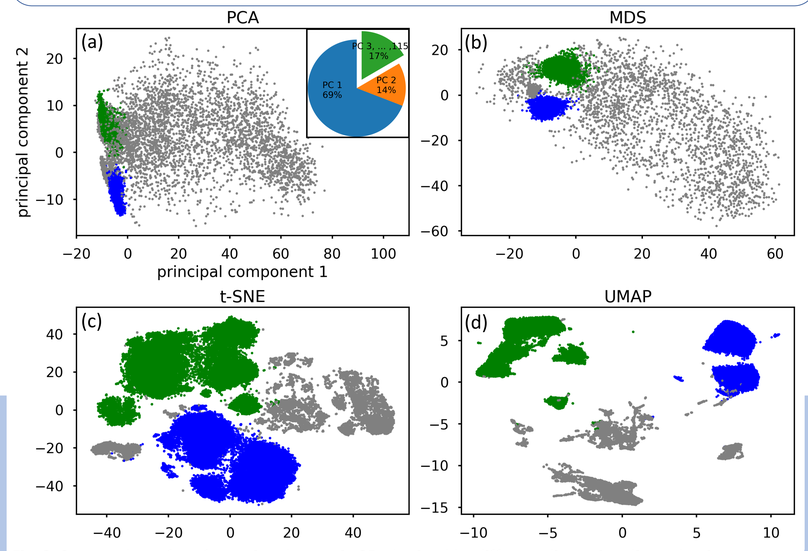

Fig. 3: Scatter plots of conformational space in 2D obtained by different dimensionality reduction techniques. Green and blue colors show two separate densely populated regions of conformational space. **(a)** PCA; *inset:* pie chart – variance by principal components; **(b)** MDS; **(c)** t-SNE; **(d)** UMAP

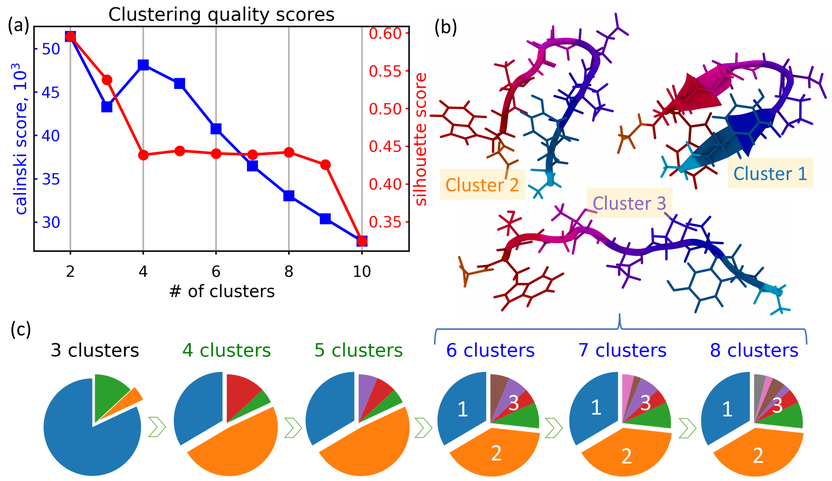

Fig. 4: Hierarchical clustering. (a) Clustering quality scores vs number of clusters: blue – calinski-harabasz score, red – silhouette score; **(b)** One representative structure for each of the three final conformational states; **(c)** The series of pie charts showing the relative population of the clusters. For 6, 7, 8 clusters the size of the two most populous clusters stays the same, suggesting conformational states shown in (b).

3. Clustering

Two clear dense regions in the two-dimensional embedding of the data set (**Fig. 3**) imply there are at least two highly populated conformational states. The next step is to perform clustering of the dataset on the original $\mathbf{R}^{115}$ feature space.

- Here we show results for *hierarchical clustering* with *'ward'* linkage and *'euclidean'* distance metric.
- It consistently shows that the two most populous clusters don't change when you divide conformational space into more clusters (**Fig. 4c**).
- Two most common clustering quality metrics: *'calinski-harabasz'* and '*silhouette'* scores (**Fig. 4a**) suggest the right number of clusters for the algorithm is in the range from 3 to 9.
- For the case of 6, 7 or 8 we define the first two conformational states corresponding to the two most populous clusters (accounting for 40% and 34% of population, respectively). The third conformational state represents all expanded structures in the ensemble and corresponds to clusters indicated by slice number '3' on the pie charts of **Fig. 4c.**
- One representative structure for each of three conformational states are shown in **Fig. 4b.** The difference between cluster 1 and cluster 2 conformations is in the position of Trp-9 below/above the hairpin, respectively.

Chemical & Physical Sciences
UNIVERSITY OF TORONTO MISSISSAUGA

NSERC CRSNG

computecanada

References

[1]. Kührová P et al.. *Biophys J.* 2012; 102(8):1897-906.
[2]. Baldi, P. *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37-49. 2012.
[3]. Kruskal, J *Psychometrika*, 29(1):1–27, 1964.
[4]. Van der Maaten, L and Hinton G. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
[5]. McInnes, L et al. *arXiv preprint arXiv:1802.03426* (2018).
[6]. https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Contacts: e.klyshko@mail.utoronto.ca, sarah.rauscher@utoronto.ca