Using Spectral Clustering to Sample Molecular States and Pathways
Shirley Ahn (안설희), Johannes Birgmeier

**Motivation**
- Molecular Dynamics (MD) simulations are quintessential tools for exploring the state space of biomolecules.
- However, exploring the state space is hindered by timescale barrier (i.e., temporal gap between simulation that requires a time step in fs and biological systems with timescales of ms).
- Hence, enhanced sampling techniques are used to explore the state space in a clever way.
- Concurrent Adaptive Sampling (CAS) algorithm uses a large number of short simulations or "walkers" with probabilities or "weights" to explore the state space.
- Reducing the number of walkers by clustering these walkers when there are too many walkers in the simulation can speed up the overall sampling.

**Methodology**
- CAS Algorithm
  - Walkers are binned to newly created circular microstates or "balls" and within each ball, walkers are split and/or merged so that each ball ends up with the same number of walkers. This is done to constantly observe all states irrespective of their energy barriers.
  - As the walkers explore the state space, the number of balls grows very quickly. When the total number of balls reaches a certain threshold, balls are clustered into a set number of macrostates.

**Spectral Clustering**
- Cluster balls based on normalized second eigenvector of the state transition matrix (i.e., probability changes in state space) and ball distance information using k-means.

**Results**
- Model of interest: penta-alanine, which has a 6-dimensional state space described by 3 pairs of dihedral angles $\phi$, $\psi$.
- First, we clustered just based on the normalized second eigenvector, which gave the following results:

![Fig 1. Penta-alanine and its 3 reference Ramachandran plots for each middle $\phi$, $\psi$ pair](image)

- Then, we decided to cluster based on the normalized second eigenvector and ball distance information:

- To further demonstrate that our clustering actually does a good job, we depict the feature space onto the first two principal components:

- We can see that the clusters are well-separated from PCA, but it is unclear which balls correspond to which physical state in the Ramachandran plot.

**Future Work**
- Increase time step of the simulation to get better sampled transition matrices. Run simulations further to collect mean first passage time statistics.

**Conclusion**: We can successfully cluster microstates into physically meaningful and well-separated macrostates and speed up the simulation by reducing “redundant” states and without sacrificing accuracy/precision.