Markov state models for molecular dynamics

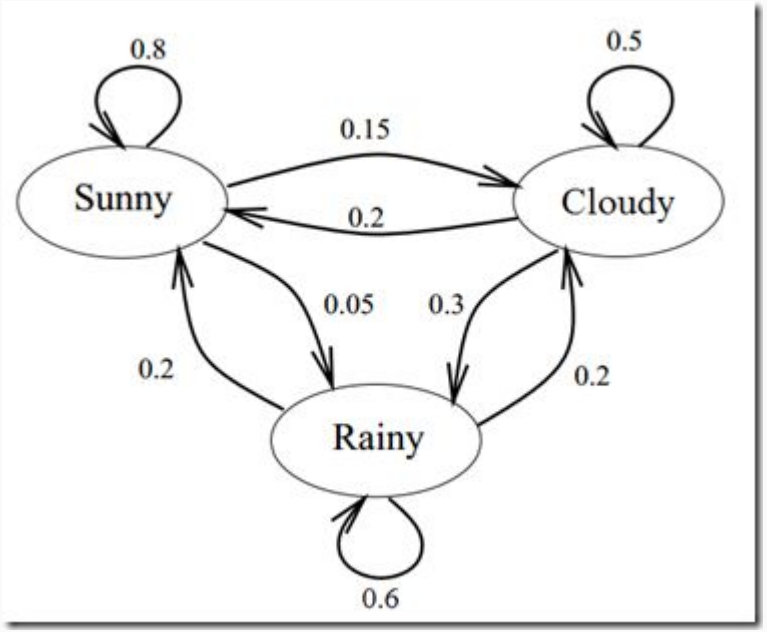
Michael Maduabum
Adithya Ganesh
Axel Sly

Everything you wanted to know about Markov State Models but were afraid to ask

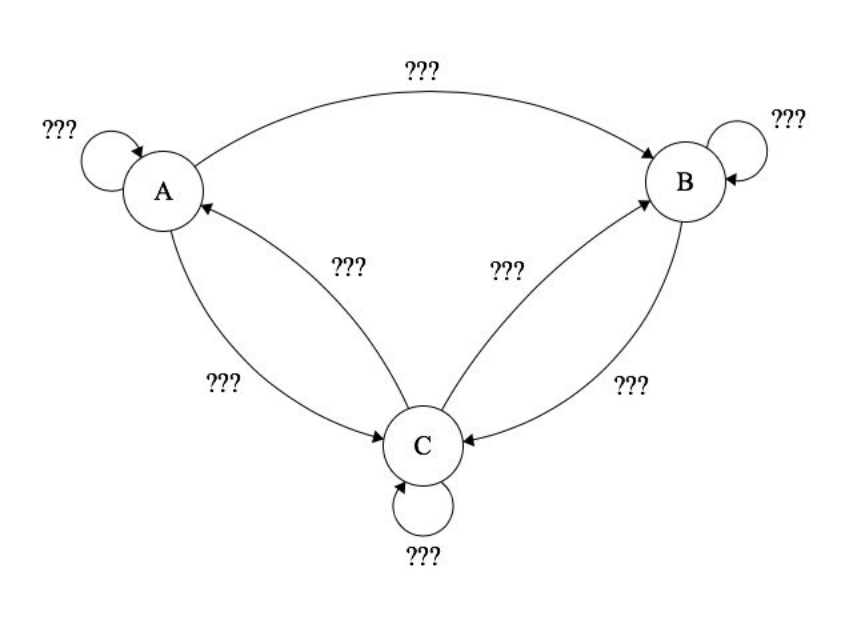
Vijay Pande, Kyle Beauchamp, Gregory Bowman

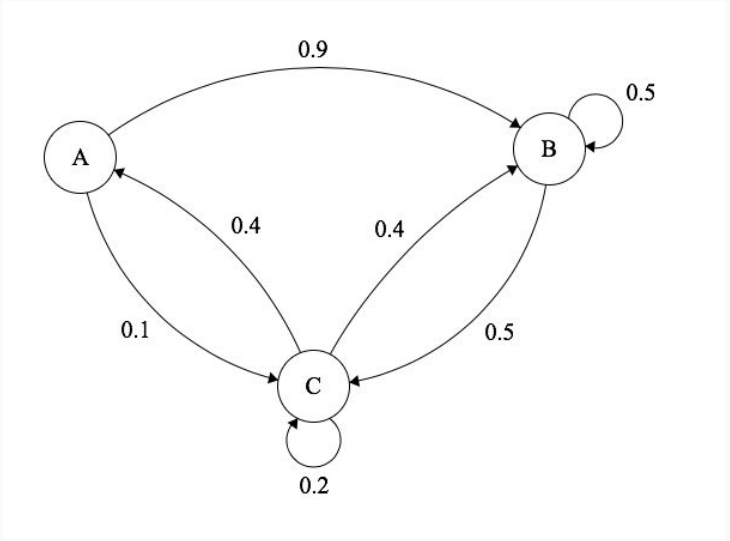
What is a Markov State Model?

- Way of modeling a random process
- Markov Assumption
  - Next state is determined only by current state



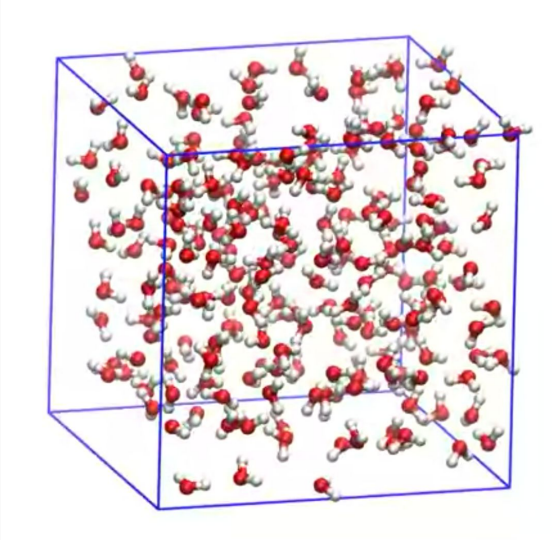
Transition probabilities





Applications

- Protein folding
- **Molecular Dynamics**
- Speech recognition
- Self-driving cars
- Gene prediction
- Machine translation
- Part-of-speech tagging



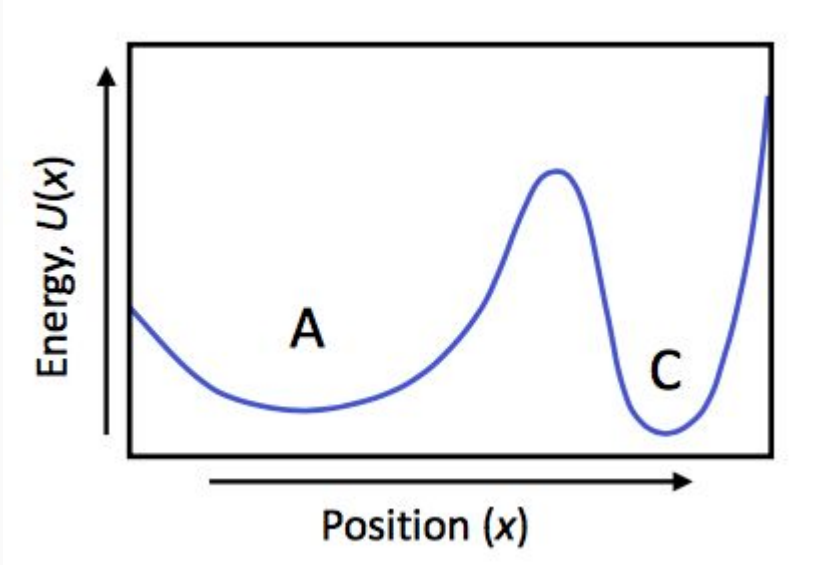
Challenges in Molecular Dynamics

- Computationally Intensive
  - Relevant simulation timescales milliseconds - seconds

- Force Fields are limited
- Analysis of simulation data
  - Trajectories

Building an MSM for MD

- Choose Markov states to correspond to kinetic macrostates
  - Similar conformations clustered to groups with a representative energy
- Build transitions matrix based on MD probabilities of state changes
  - Frequentist and Bayesian approaches



Molecular Dynamics with MSM’s

- N states with transition probabilities between them
  - N ~ thousands to millions
  - Large number of states gives more accurate representation
- Define states in a physically meaningful way
- Build the transition matrix efficiently
- Start with some initial data set
  - e.g. a molecular dynamics simulation

Adaptive Sampling

- Often there is not enough simulation data to immediately build an accurate Model
  - Need to perform more simulations to get novel data
- Use transition probability matrix to select macrostate to start simulation in
  - Statistical analysis to run simulations to generate the most data
- Significantly improves efficiency

Caveats

- Sampling is *still* a challenge
  - Long-timescale events are out of reach
- Limited by force fields
- State decomposition is difficult
  - Complex systems require more than just kinetic approaches
    - e.g. knots, topological effects
- Need new metrics for structural similarity

Questions?

Markov state models of biomolecular conformational dynamics

Chodera et al., 2014

Adithya Ganesh

Markov Chain Simulation

- Markov Chain simulation
  - http://setosa.io/blog/2014/07/26/markov-chains/index.html

Recent theoretical advances

- Old approach
  - Maximizing “metastability” -- quantity related to *lifetime* of states
- Recent advances
  - Formulate analysis as an eigenvalue problem
- Key intuition
  - Eigenvector with eigenvalue 1 → steady state!

$$\boldsymbol{\pi}^\top \mathbf{P} = \boldsymbol{\pi}^\top .$$

Markov Chain Eigenvector Simulation

- **Key intuition**
  - Eigenvector (with $\lambda = 1$)$\rightarrow$ steady state
- Simulation
  - http://setosa.io/ev/eigenvectors-and-eigenvalues/
- PageRank
  - *“The $25,000,000,000 Eigenvector - The Linear Algebra Behind Google”*
    - https://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf
  - A very similar Markov chain problem
    - Obtain transition probabilities by observing which pages link to others
    - Eigenvectors give you the steady state “page ranks”

EMMA Python package -- demo

Raw simulated trajectory

↓

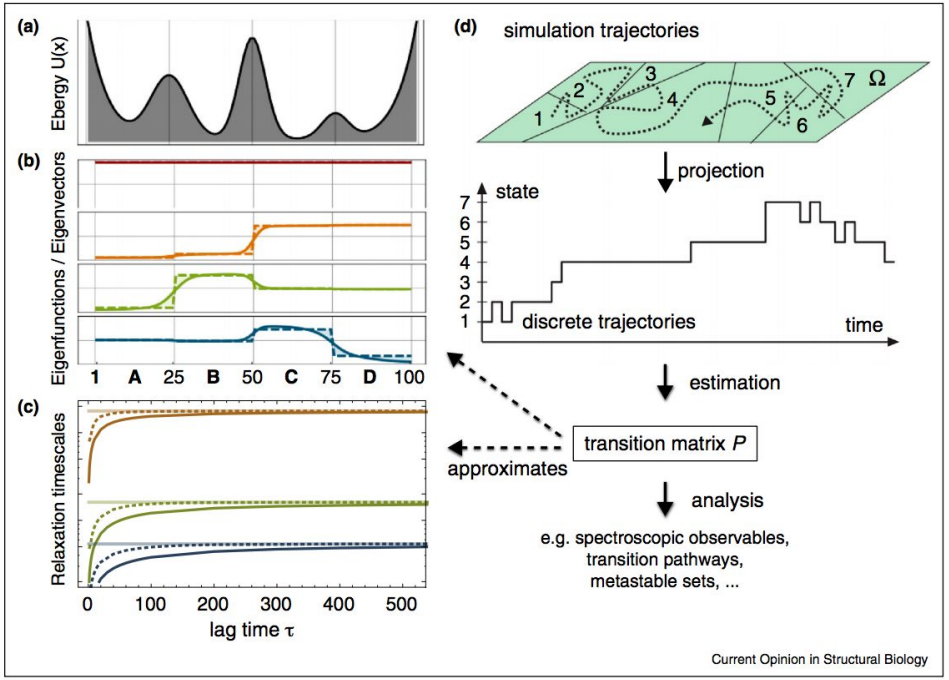
“Compressed” Trajectory (TICA projection)

↓

Clustering (K-means)

↓

Estimate MSM + visualize states



Questions?

HTMD: High-Throughput Molecular Dynamics for Molecular Discovery

Motivation

- Molecular dynamics simulations (MD) come with several limitations:
  - Data analysis
  - Reproducibility
  - Accuracy of force fields
  - Time sampling limitations
- Necessity of a more standardized methodology

High-Throughput Molecular Dynamics (HTMD)

- Programmable workspace for simulation based molecular discovery
- Python
- VMD

Integrated Platform

- Structure Manipulation
- System Building
- Molecular Simulation
- Adaptive Sampling
- Projecting and Clustering
- Markov State Models (MSMs)

2. METHODS

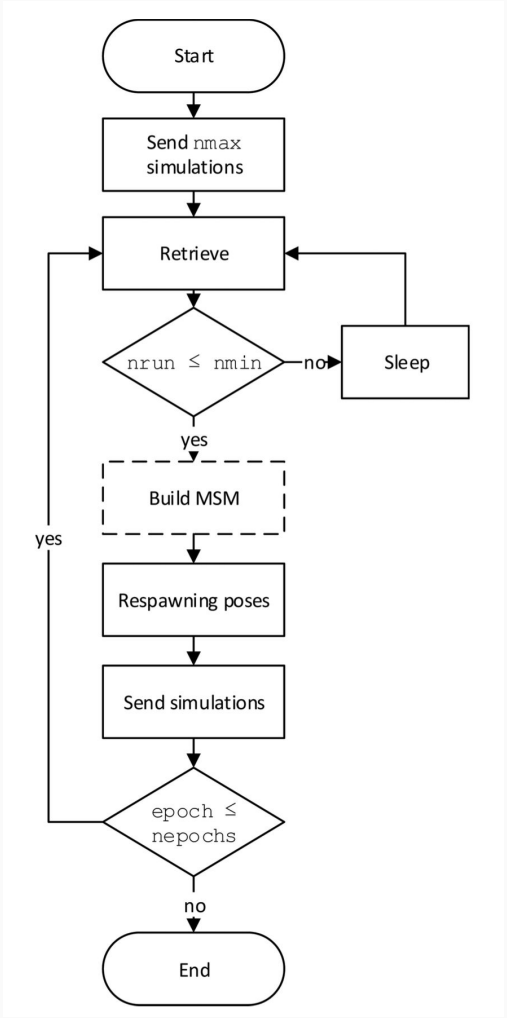
2.1. Integrated Platform for Molecular Modeling.

Listing 1: Manipulating molecular structures.

```
# Download the PDB structure of Barnase
mol = Molecule('2F4Y')
# Set the chain ID of Barnase to 'Y'
mol.set('chain', 'Y', sel='protein')
# Remove the carbon alpha of residue 15
mol.remove('residue 15 and name CA')
# Append Barstar to Barnase
mol.append(Molecule('2HXX'))
# Visualize in VMD or NGL
mol.view()
# Solvate in water box
solvate(mol, minmax=[[-50, -50, -50],
                     [50 ,  50,  50]])
```

Adaptive Sampling

- Efficient and intelligent sampling of conformational space
- From previous simulation knowledge, it identifies conformational subspaces that are under-sampled



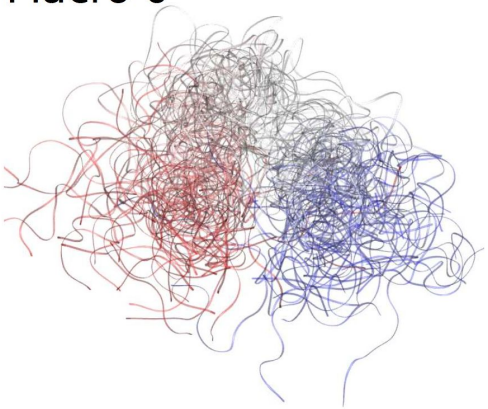
Protein Folding Analysis

```
# Collect and link simulation files
sims = simlist(glob('data/*/'),
    glob('input/*/structure.pdb'))
# Calculate protein contact maps
met = Metric(sims)
met.projection( MetricSelfDistance(
    'protein and name CA',
    metric='contacts') )
data = met.project()
# Project on slowest 10 TICA dimensions
tica = TICA(data, 20)
dataTica = tica.project(10)
# Cluster data with kmeans
dataTica.cluster(
    MiniBatchKMeans(n_clusters=1000))
# Link a model with the data
model = Model(dataTica)
# Calculate the Markov model
model.markovModel(300, 4)
# Macrostates equilibrium distribution
model.eqDistribution()
# Visualize the macrostates in VMD
model.viewStates()
# Calculate the kinetics
kin = Kinetics(model, temperature=360)
kin.getRates()
```

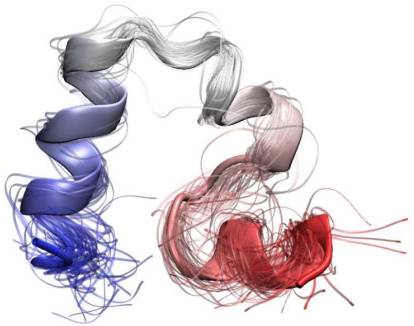
Protein Folding Analysis

The 4 macrostates produced by the Markov state model of Villin

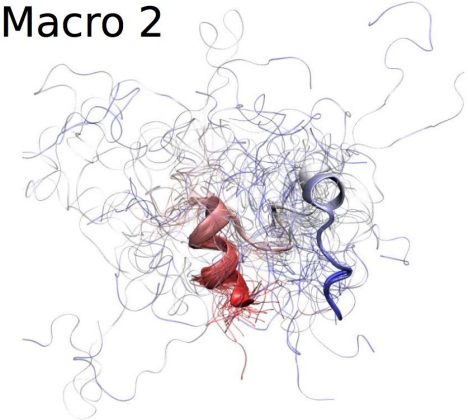
Macro 0



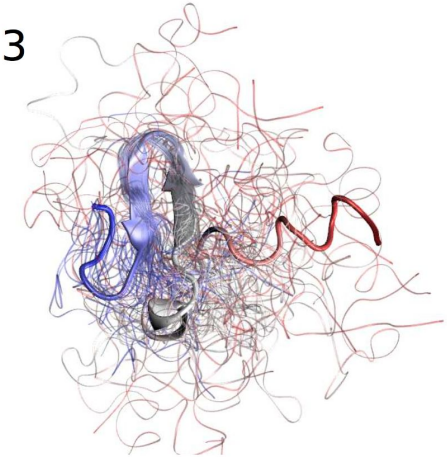
Macro 1



Macro 2

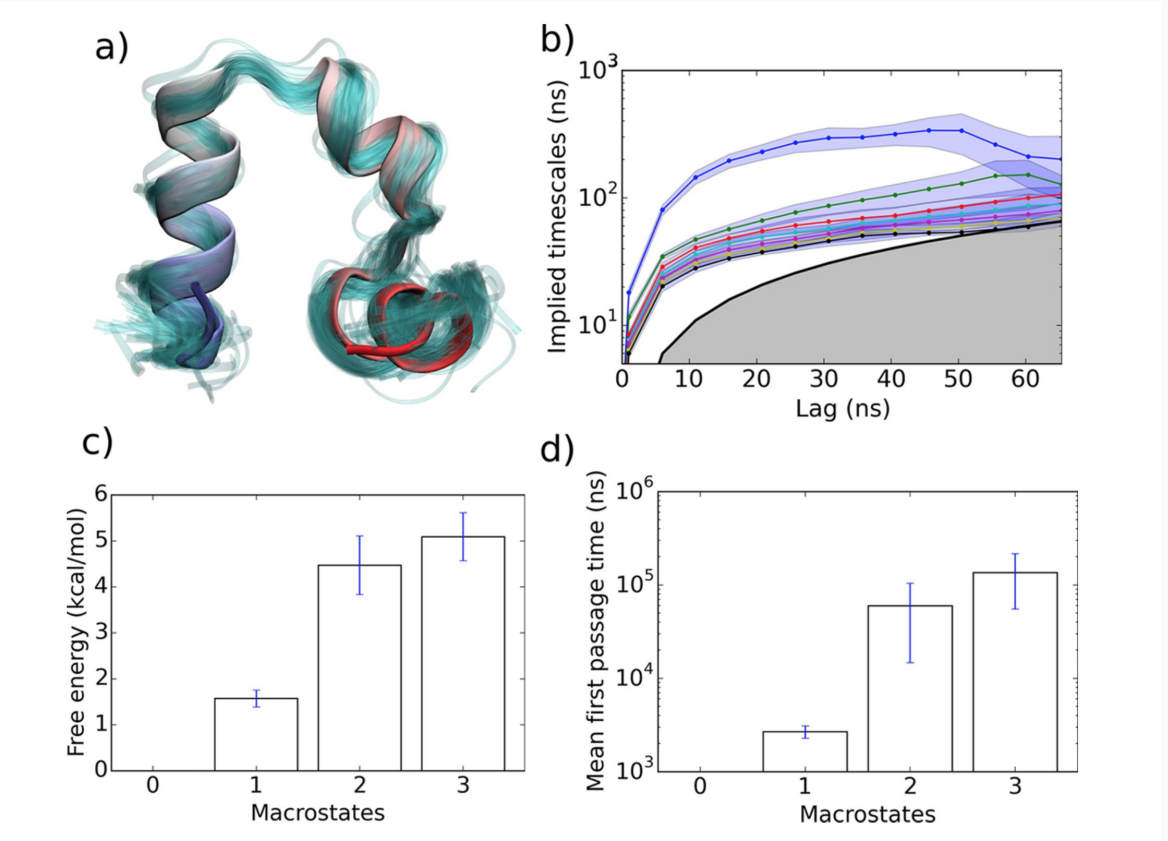


Macro 3



Questions?

Protein Folding Analysis



Adaptive Sampling Protein Folding