Markov state models

Nina Hinrichs

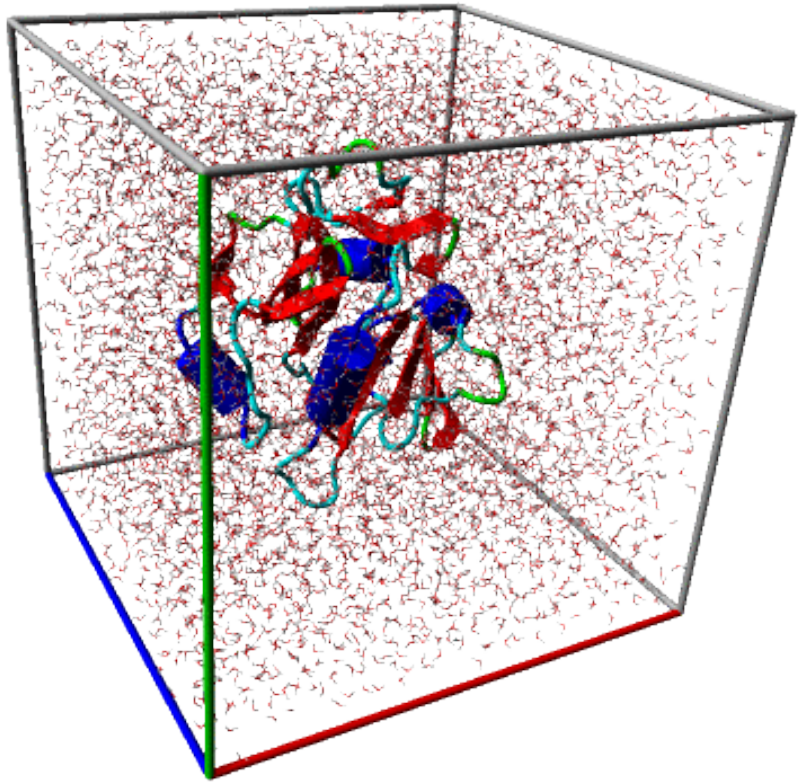
Departments of Computer Science and Statistics

University of Chicago

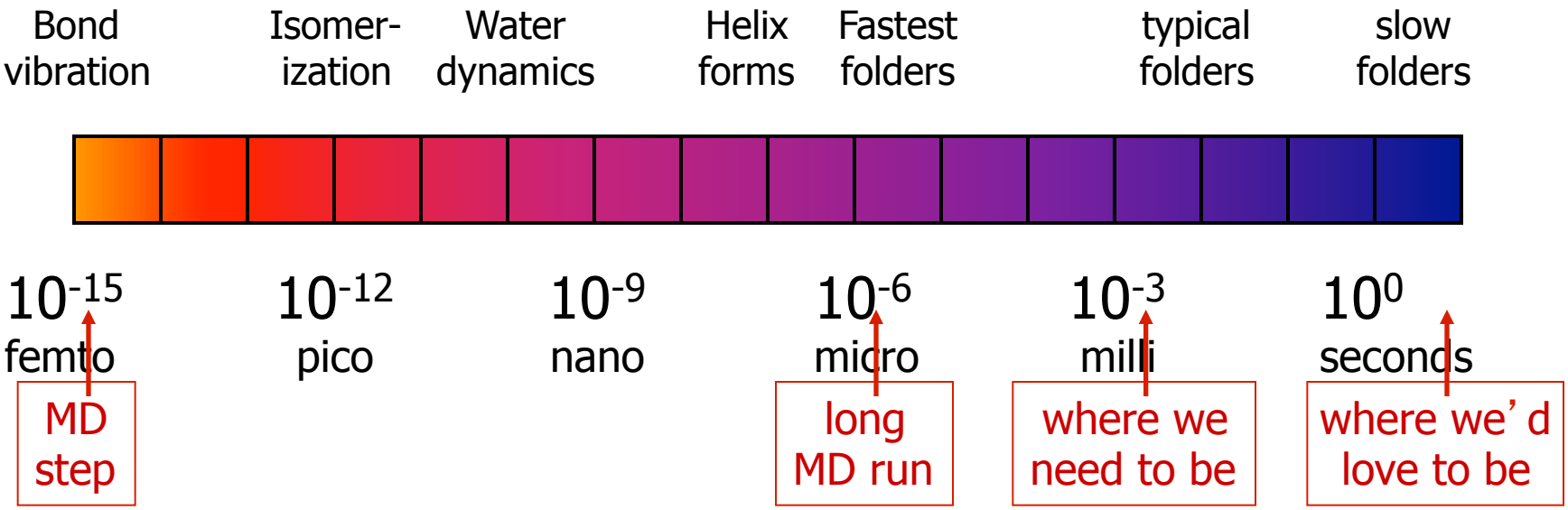
June 12, 2012

Molecular dynamics

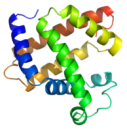
- Represent atoms of molecule and solvent
- Model forces on atoms
- Integrate laws of motion
- Small integration time step compared to motion timescales



Challenges

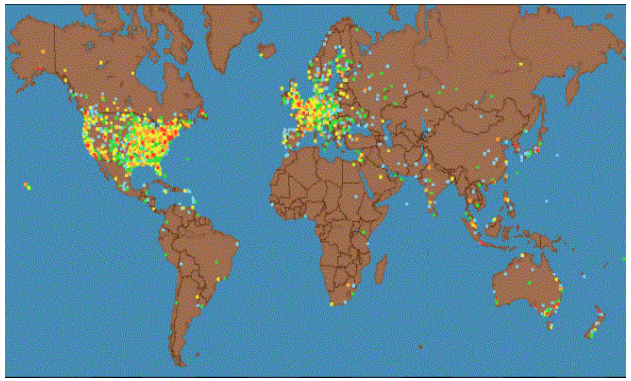


- Fundamental problem for simulation
  - Proteins fold in micro- to milliseconds
  - Computers can simulate nano- to microseconds
  - **How can we break this impasse?**



Large scale computing

Folding@home distributed computing

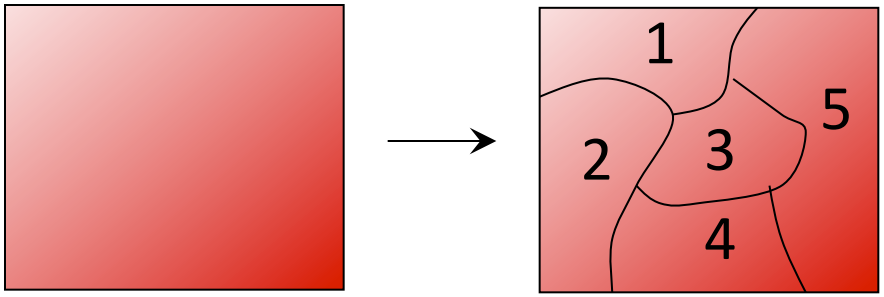


Outline

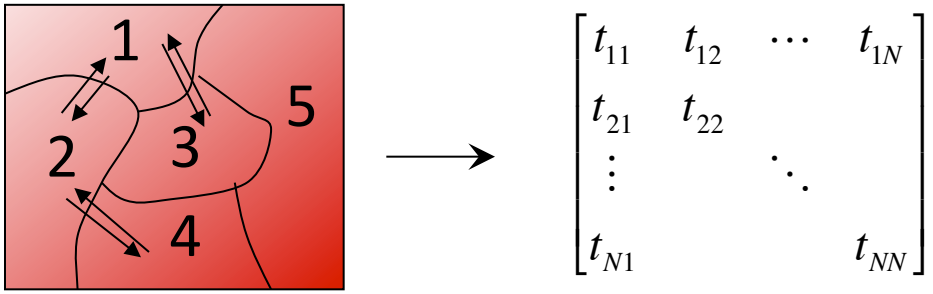
- Markovian state model for molecular motion
  - Model description, uses, examples
- Algorithms for building these models
  - Defining states and transition probabilities
- Methods for dealing with finite sampling
  - Uncertainty analysis, targeted sampling

Markovian state model

- Main idea: Partition conformation space of system into discrete states



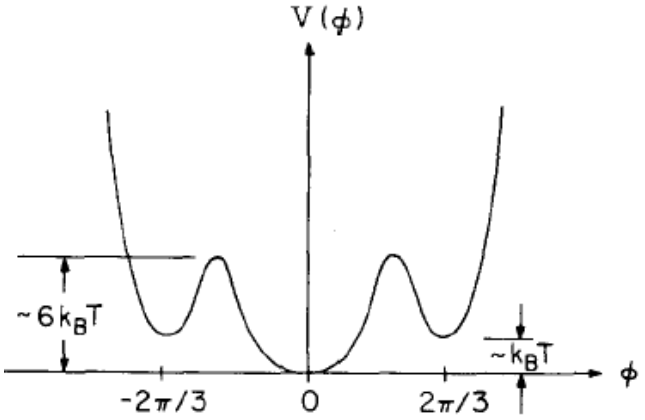
- Model dynamics as Markov jumps between states

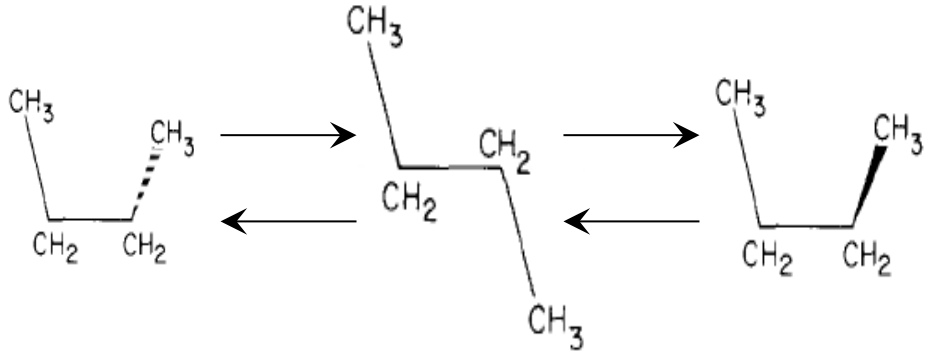


Chemical intuition

Chemical reactions often exhibit stochastic behavior

n-butane

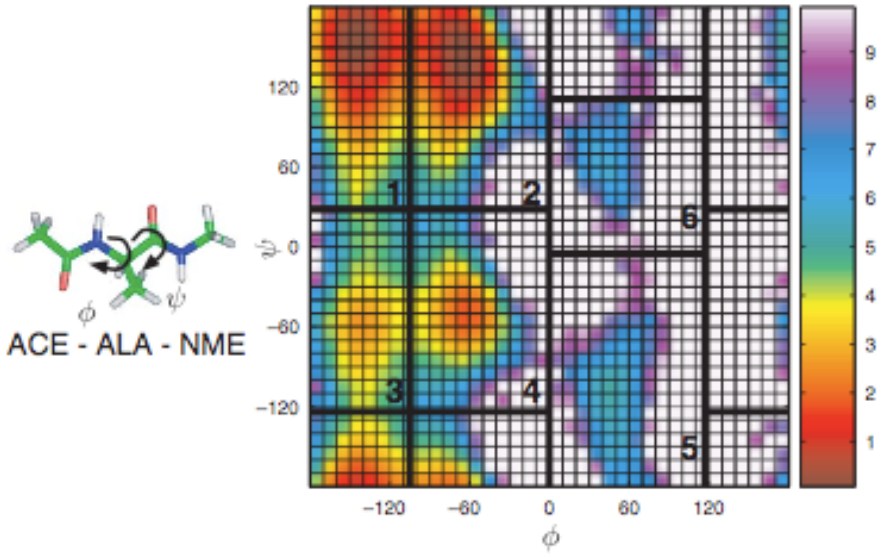




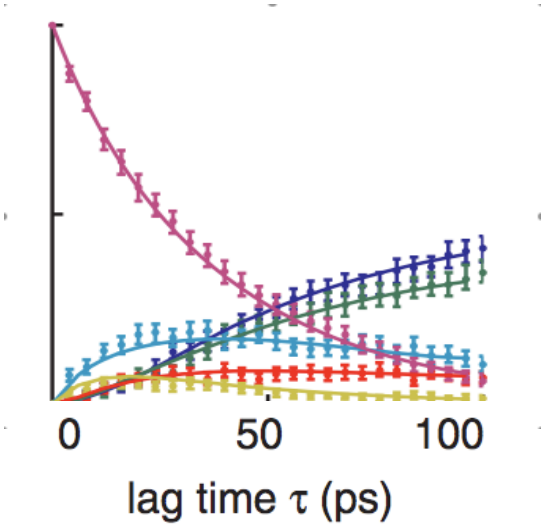
Chandler, Journal of Chemical Physics (1977)

Long-time behavior

$$\begin{bmatrix} p_1^{(1)} \\ p_2^{(1)} \\ \vdots \\ p_N^{(1)} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{N1} \\ t_{12} & t_{22} & & \\ \vdots & & \ddots & \\ t_{1N} & & & t_{NN} \end{bmatrix} \times \begin{bmatrix} p_1^{(0)} \\ p_2^{(0)} \\ \vdots \\ p_N^{(0)} \end{bmatrix}$$

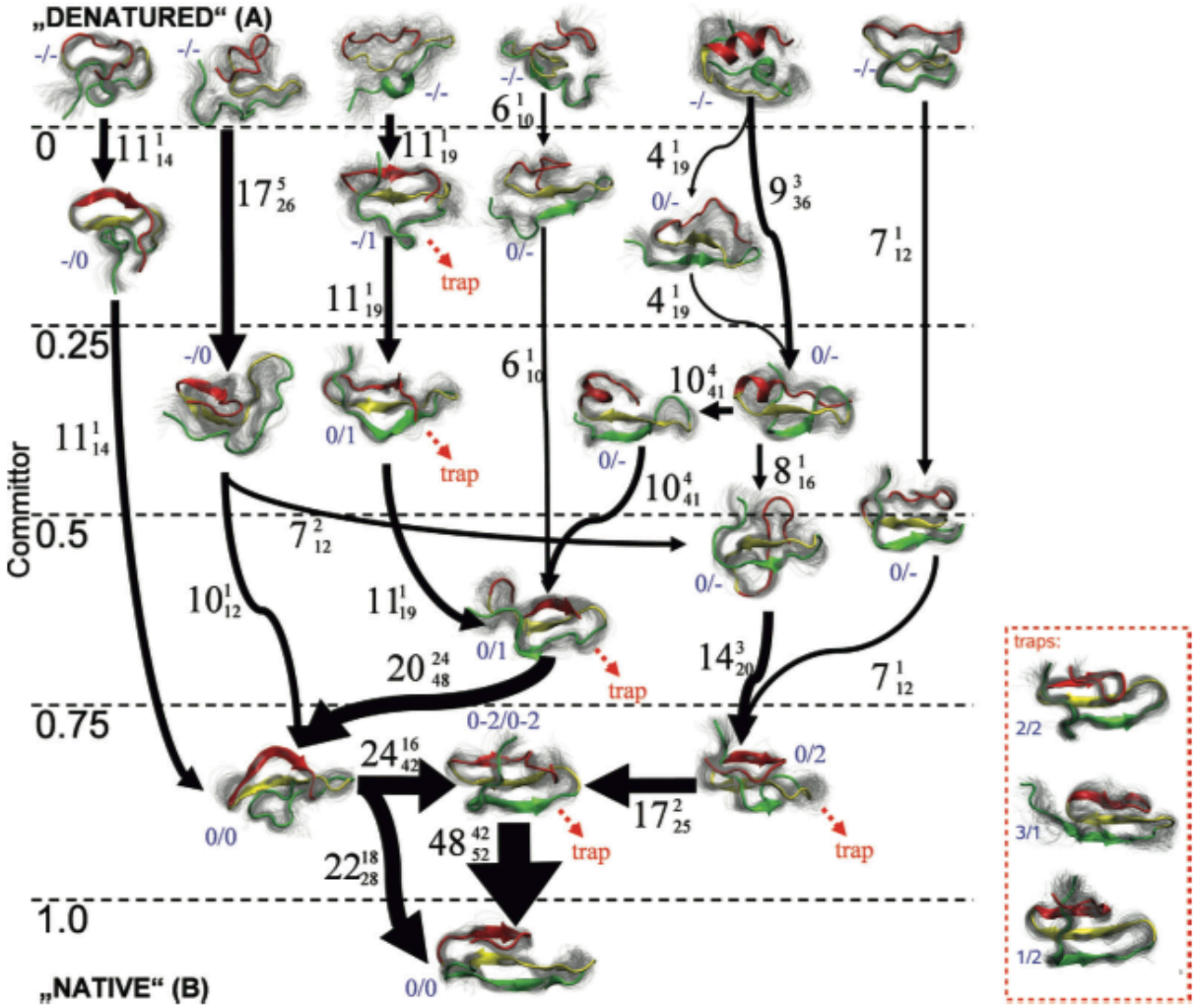


$$\begin{bmatrix} p_1^{(T)} \\ p_2^{(T)} \\ \vdots \\ p_N^{(T)} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{N1} \\ t_{12} & t_{22} & & \\ \vdots & & \ddots & \\ t_{1N} & & & t_{NN} \end{bmatrix}^T \times \begin{bmatrix} p_1^{(0)} \\ p_2^{(0)} \\ \vdots \\ p_N^{(0)} \end{bmatrix}$$



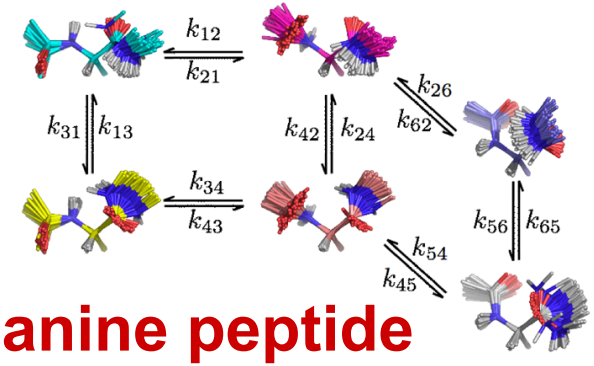
Chodera et al., *Multiscale Modeling and Simulation* (2006)

Mechanistic properties



Noe et al., *PNAS* (2009)

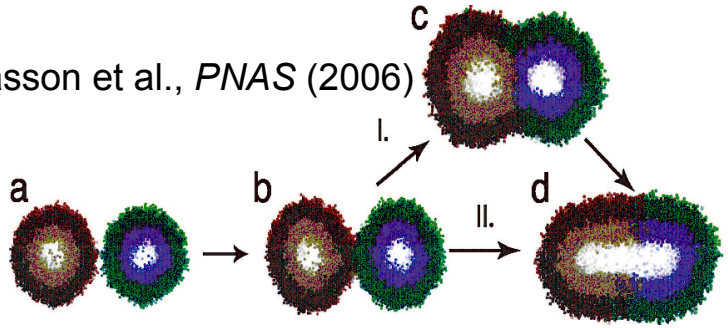
Example models



alanine peptide

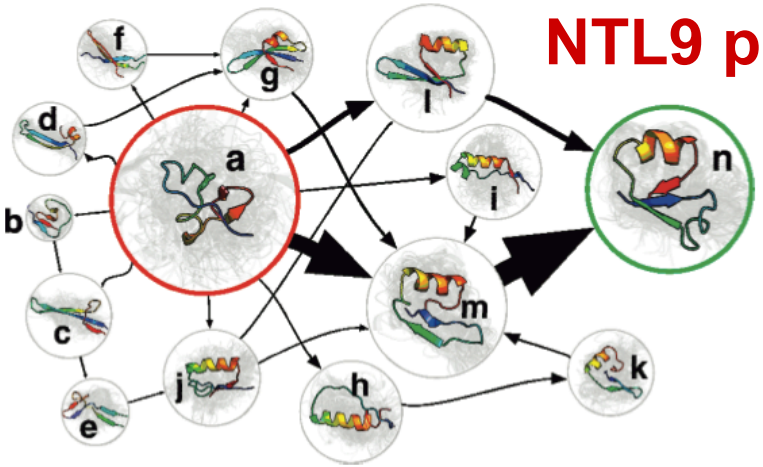
Chodera et al., *Multiscale Modeling and Simulation* (2006)

Kasson et al., *PNAS* (2006)

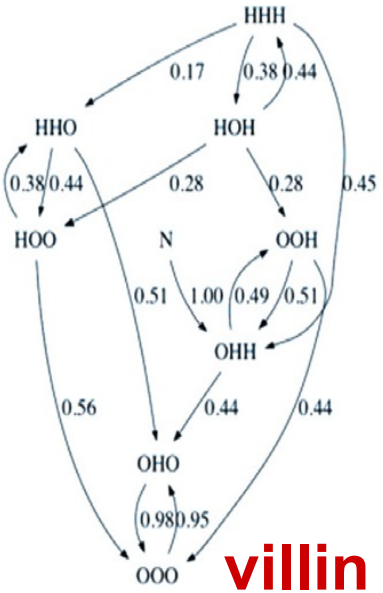


lipid vesicle fusion

NTL9 protein



Voelz et al., *JACS communications*, 2010



villin headpiece

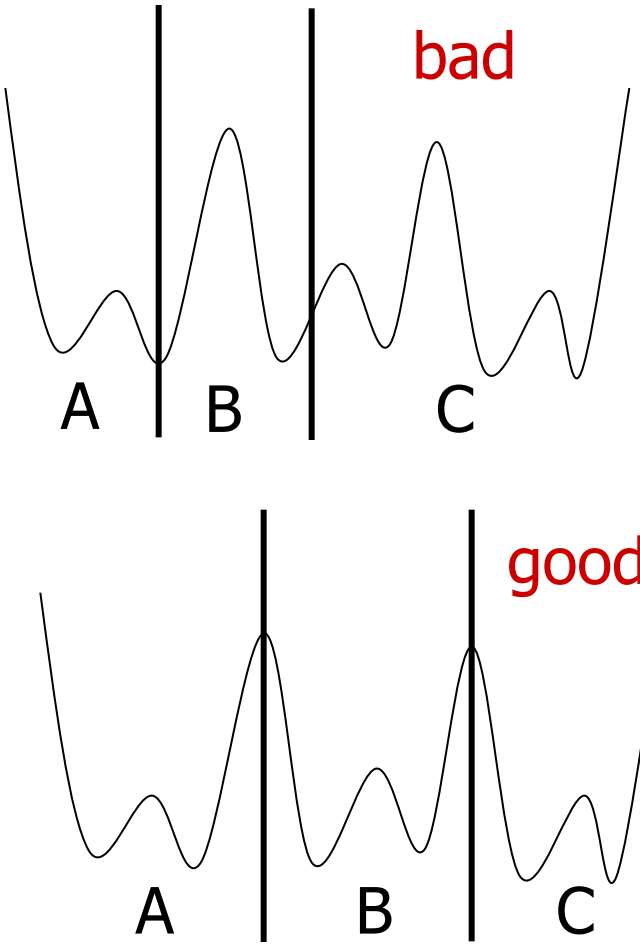
Jayachandran et al., *Journal of Structural Biology* (2006)

Outline

- Markovian state model for molecular motion
  - Model description, uses, examples
- Algorithms for building these models
  - Defining states and transition probabilities
- Methods for dealing with finite sampling
  - Uncertainty analysis, targeted sampling

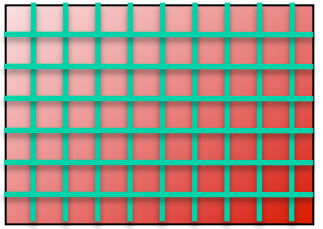
State definitions

- Challenge: find states which are Markovian
- Basic algorithm – combine *structural* and *kinetic* similarity
  - Subdivide into many structurally similar clusters
  - Regroup into kinetically similar states

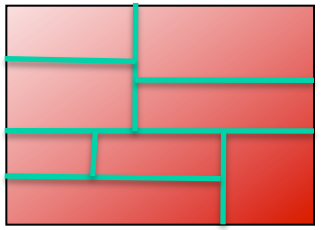


Partitioning space -- binning/boxing

- Binning
  - Subdivide each degree of freedom into finite number of bins
  - Each “cluster” is an assignment to a bin in each dimension

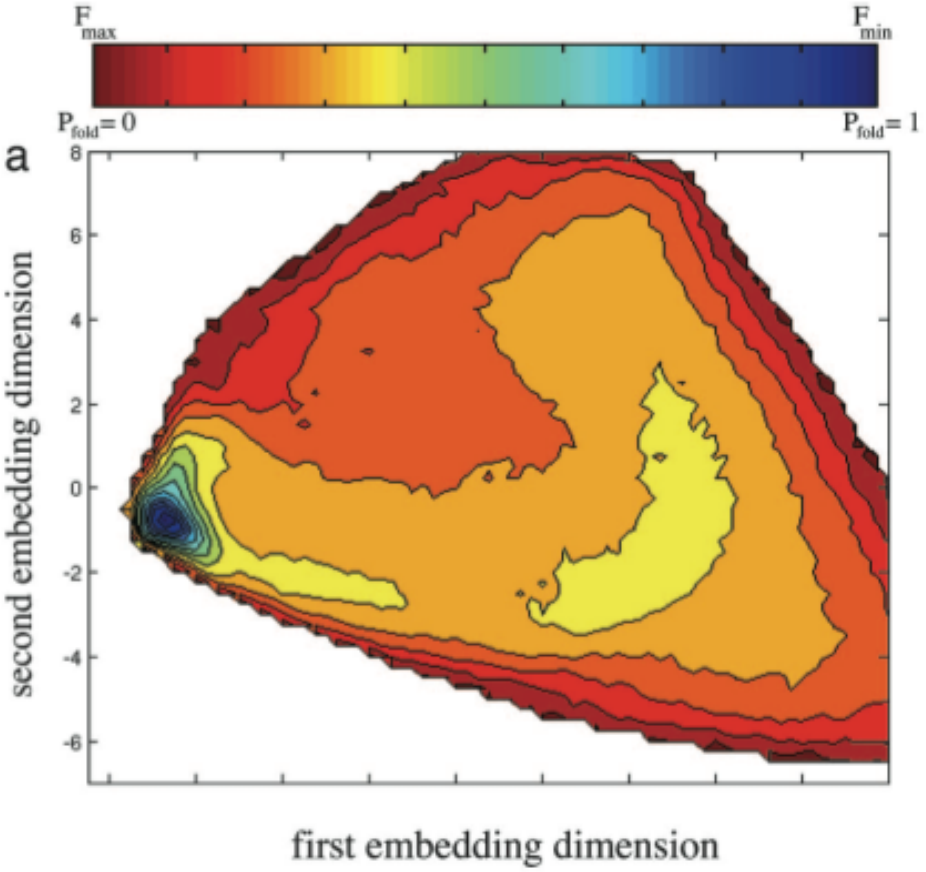


- Boxing
  - Iteratively subdivide by one dimension at a time
  - Results in fewer overall clusters



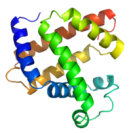
- Drawbacks
  - Number of clusters scales exponentially with number of dimensions

Partitioning space: Low dimensional projection



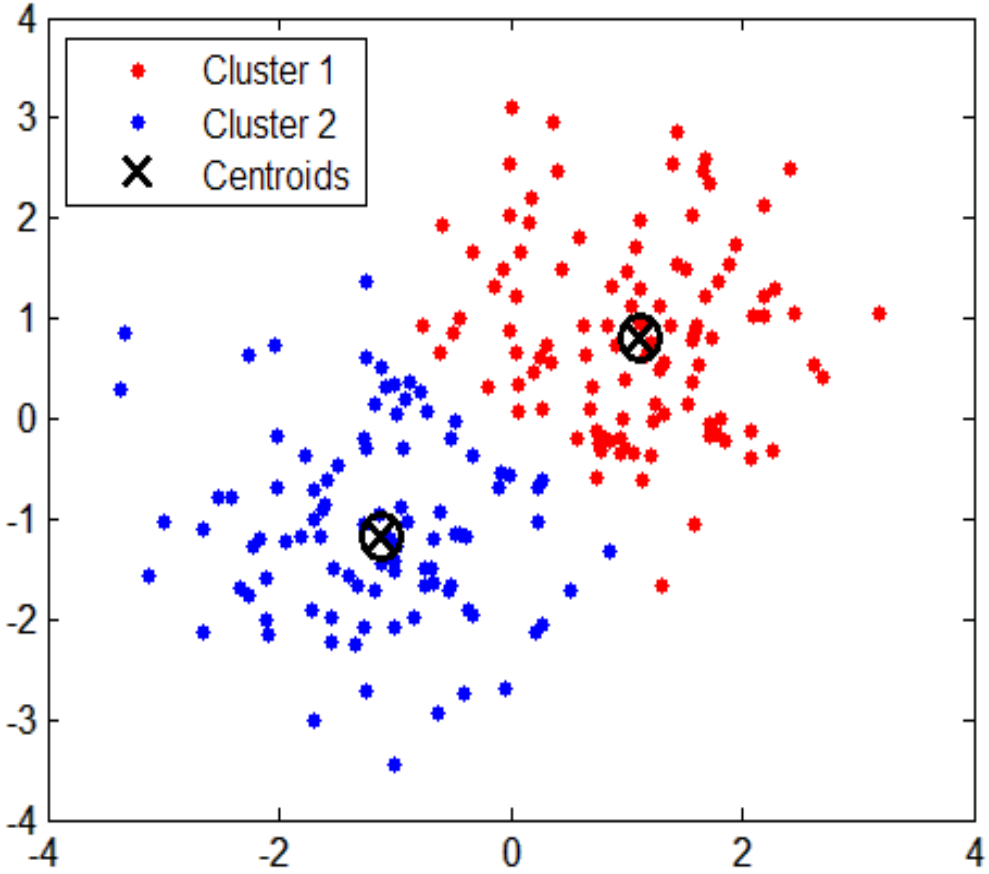
- Find low dimensional projection of data which preserves dynamics
- Various linear and non-linear techniques
- Bin data in this low-dimensional space

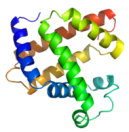
Das et al., *PNAS* (2006)



Partitioning space: High dimensional clustering

- Input: The *n* data points and number of desired clusters *k*
- Output: A set of *k* cluster centers that (approximately) minimize the sum of squared distances between each data point and the closest cluster center
- NP-complete for *k* > 1



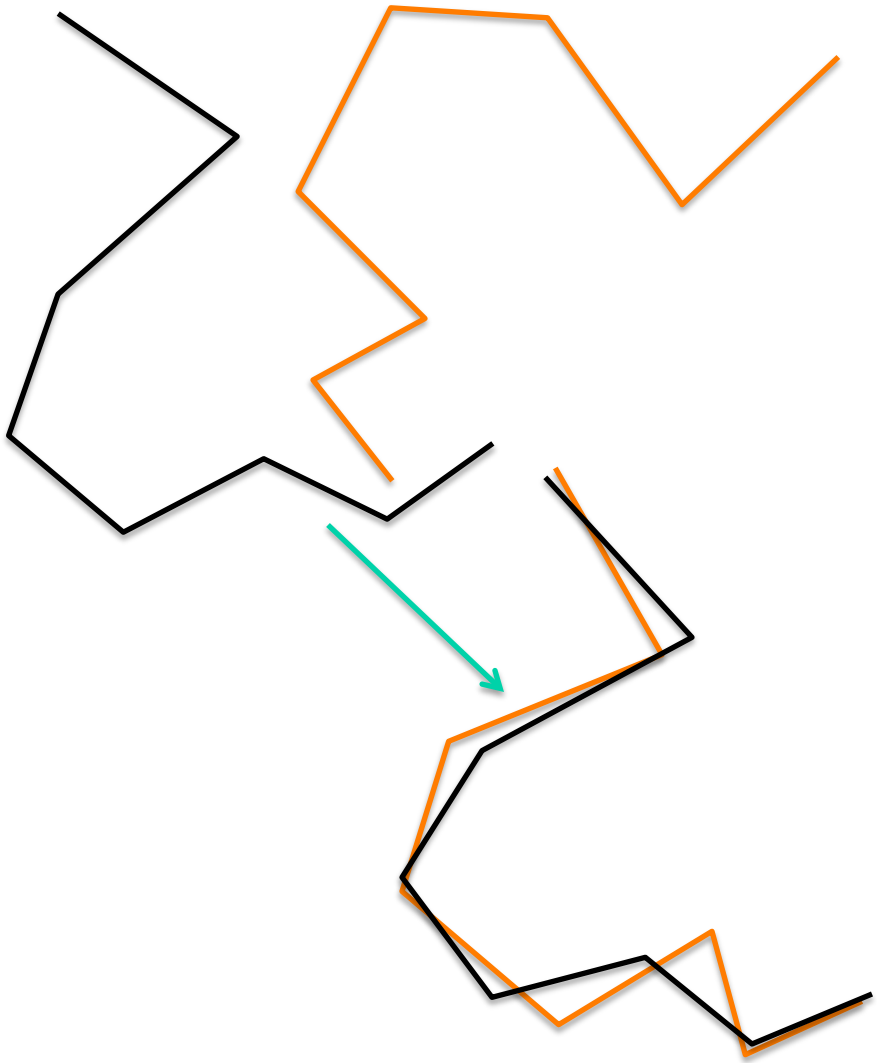


Distance metric for comparing molecular conformations

- Common distance metric between two structures is root mean square distance (RMSD)
  - N: number of atoms
  - d_i: distance between atom i positions

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{N}}$$

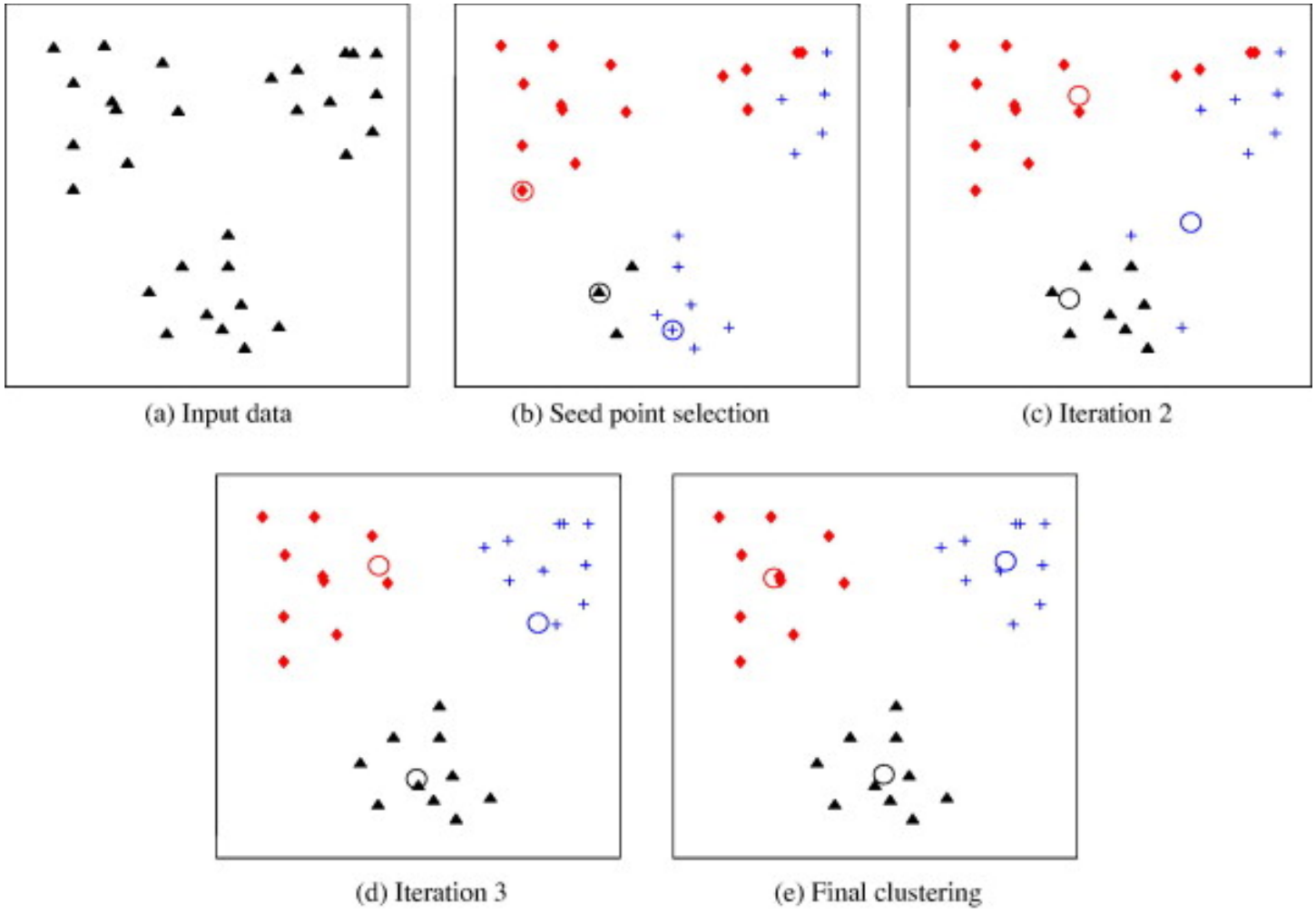
- We first translate and rotate one of the structures to compute the *minimum* RMSD (can be done efficiently using quarternions or singular values)



K-means clustering algorithm

- K-means algorithm:
  - Initialize centers (normally choose random points, either in the space or from the data)
  - Repeat:
    - Assign each data point to the closest cluster center
    - Compute new cluster center as average of points assigned to that cluster
  - Stop when cluster centers stop changing or when change in the sum of squared distances is small

K-means clustering example



(a) Input data

(b) Seed point selection

(c) Iteration 2

(d) Iteration 3

(e) Final clustering

Jain, *Pattern Recognition Letters*, 2009

Kinetic clustering motivation

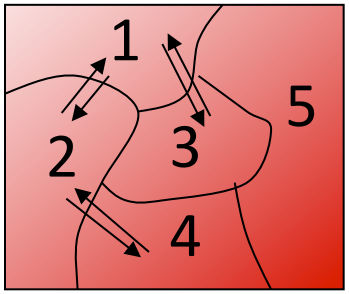
- Want compact model of the system
  - Hard to interpret model with 10000 states
  - Statistical uncertainty grows with number of states (more later)
- Main idea: Group together structural clusters into states such that transitions between states are rare

Kinetic similarity

- Want similarity measure between clusters that tells how likely a transition is between them
- Transition probability: Likelihood of being in cluster *j* at time τ if in cluster *i* at time 0
- Monte Carlo estimate of transition probabilities from molecular dynamics

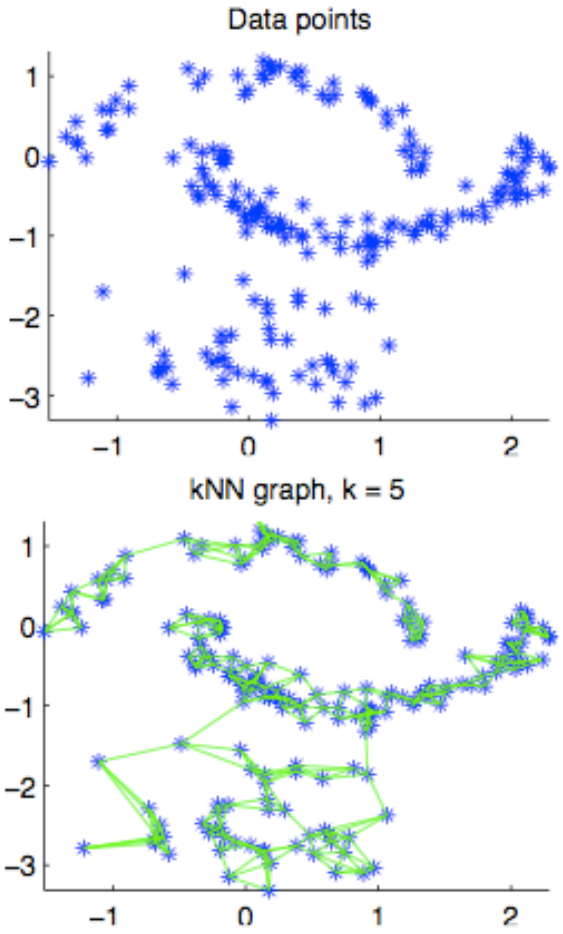
$$T_{ij}(\tau) = \frac{z_{ij}}{\sum_k z_{ik}}$$

- Assumes samples within each state follow the correct distribution



$$\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ t_{21} & t_{22} & & \\ \vdots & & \ddots & \\ t_{N1} & & & t_{NN} \end{bmatrix}$$

Spectral clustering

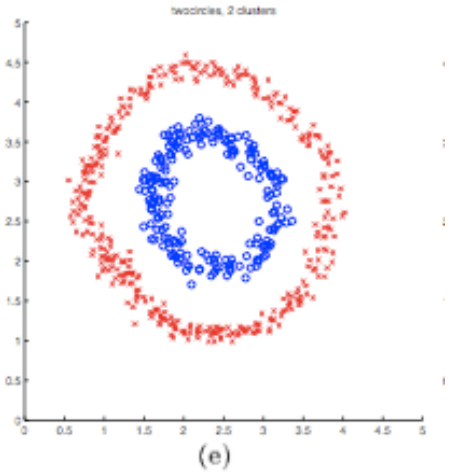


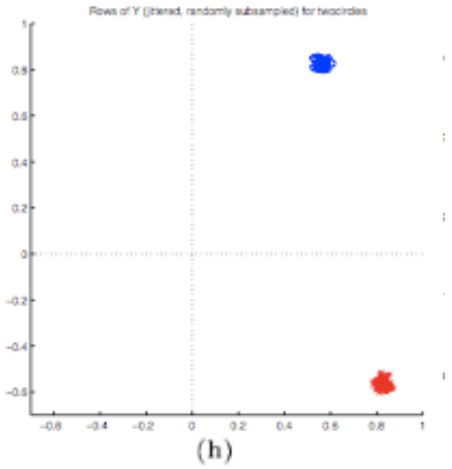
- General clustering method
- Define “similarity” between points
- Try to find “cut” of graph such that size of cut (sum of edges) is minimal (NP-hard)
- Can formulate problem as random walk over edges

Luxburg, *Statistics and Computing*, 2007

Spectral clustering algorithms

- K-means can only define convex regions in input space
- Project each data point onto top k-dimensional eigenvector space
- Similar points will have similar values for these longest processes
- Cluster in this k-dimensional space

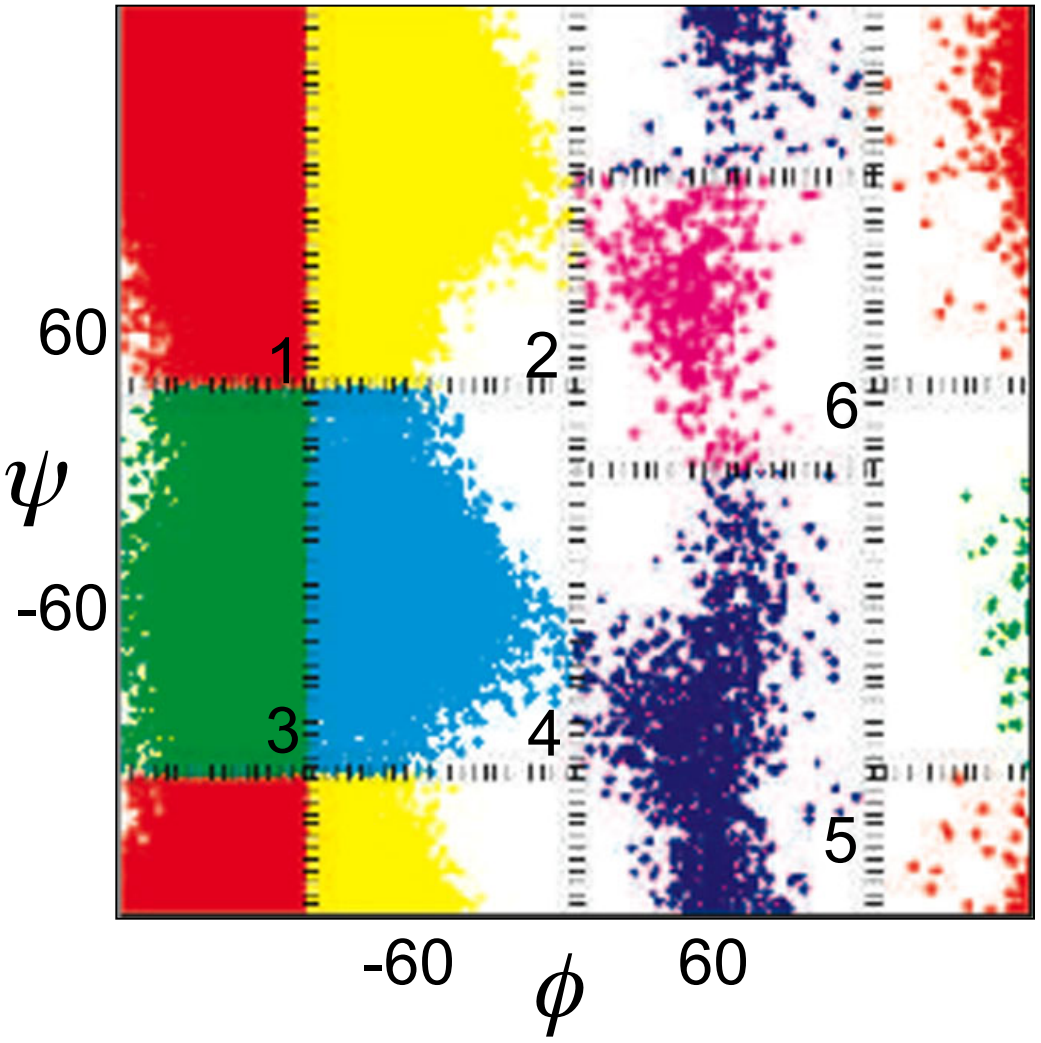


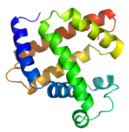


Ng et al., *NIPS*, 2001

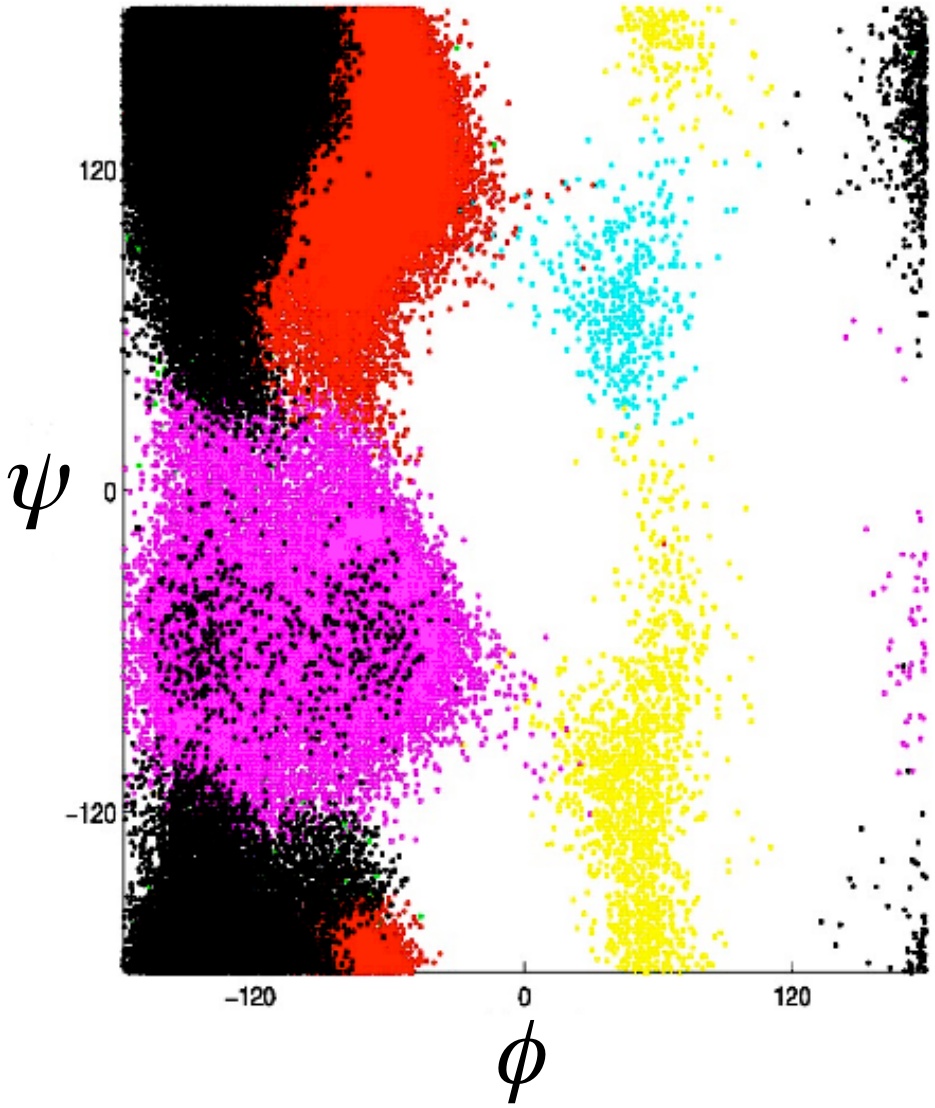
Blocked alanine peptide

Chodera et al., *Multiscale Modeling and Simulation* (2006)



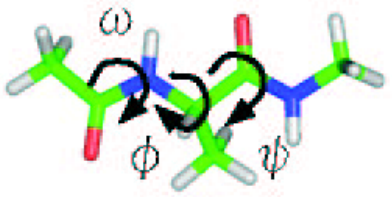


Automatic state decomposition of alanine peptide

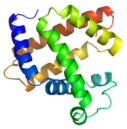


Black state sits on top of multiple other states!

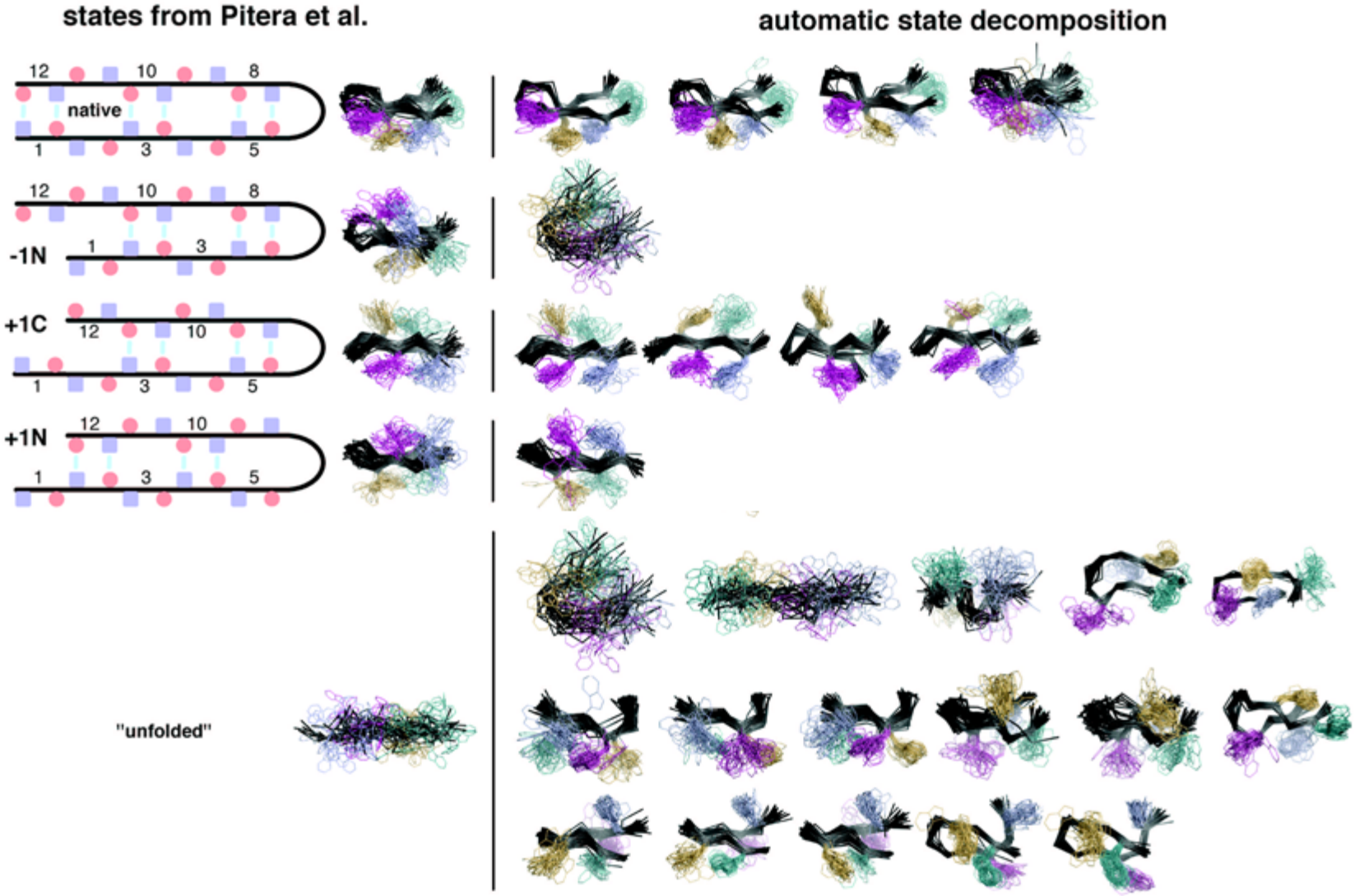
These conformations had an unusual peptide bond



Benefit of automatic algorithm



TrpZip peptide



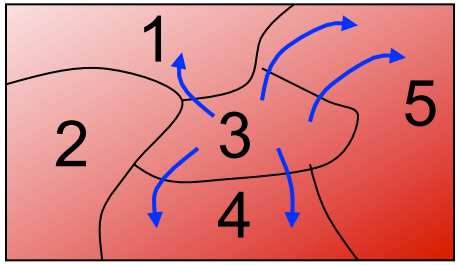
Outline

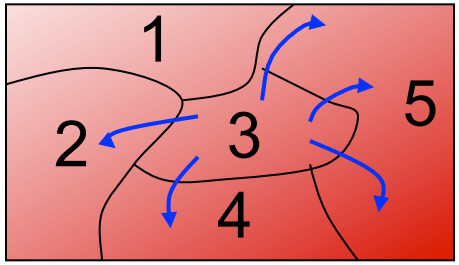
- Markovian state model for molecular motion
  - Model description, uses, examples
- Algorithms for building these models
  - Defining states and transition probabilities
- Methods for dealing with finite sampling
  - Uncertainty analysis, targeted sampling

Uncertainty analysis

Goal: Once we have the states, what is the uncertainty in the model?

Uncertainty caused by finite sampling





Both are reasonable but give different transition probabilities

→ Different MFPT, P_{fold}, eigenvalues, eigenvectors ...

Hinrichs and Pande, *Journal of Chemical Physics*, 2005/2007

Transition probabilities

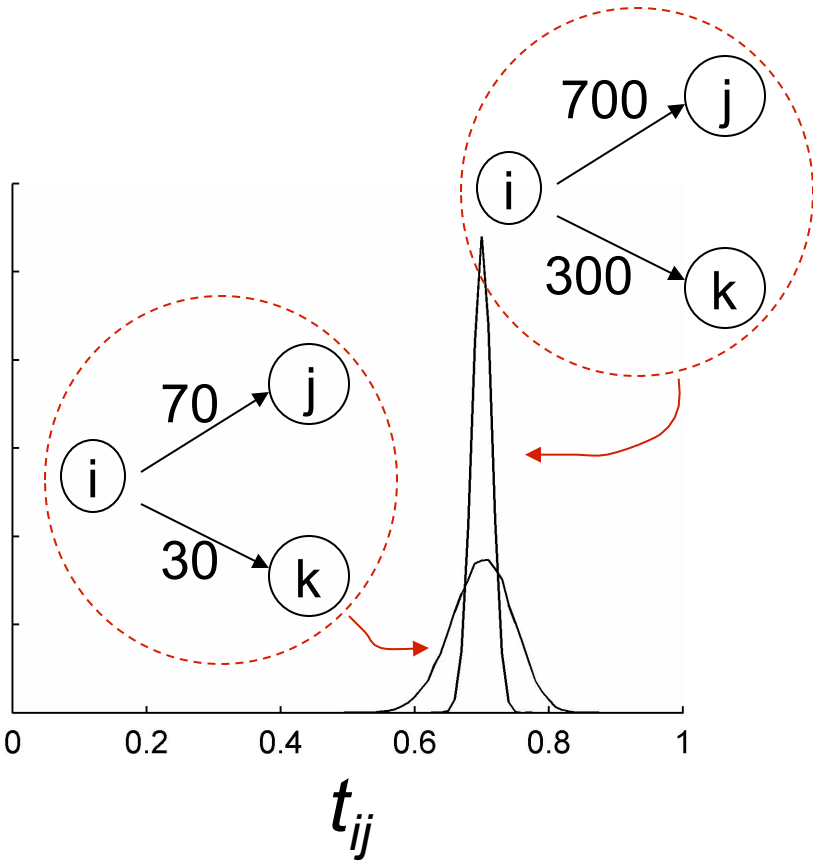
Recall that we calculate transition probabilities by counting:

$$t_{ij} = \frac{z_{ij}}{\sum_k z_{ik}}$$

Instead of getting a single value, we can talk about the *distribution* of transition probabilities

Bayes' Rule:

$$P(t_{i*} \mid counts) \propto P(counts \mid t_{i*}) P(t_{i*})$$



Closed-form solution

Idea: trade exact distribution for efficient approximation

Eigenvalue equation:

$$\det(\underbrace{\mathbf{T} - \lambda\mathbf{I}}_{\mathbf{A}}) = 0$$

Taylor series expansion:

efficient to calculate using adjoint systems

$$\lambda = \bar{\lambda} + \left.\frac{\partial\lambda}{\partial t_{11}}\right|_{\bar{\mathbf{A}}} \Delta t_{11} + \left.\frac{\partial\lambda}{\partial t_{12}}\right|_{\bar{\mathbf{A}}} \Delta t_{12} + \ldots + \left.\frac{\partial\lambda}{\partial t_{NN}}\right|_{\bar{\mathbf{A}}} \Delta t_{NN}$$

Multivariate normal approximation of $\Delta\mathbf{t}_{i*}$

→ Closed-form normal distribution for λ

Sampling strategies

Problem: Simulations are expensive. Even with distributed/ super computing simulations are run for months

How to intelligently allocate our resources?

Common approaches:

- equilibrium sampling – sample each conformation from its equilibrium distribution
- even sampling – sample equally from each state

New sequential approaches

Adaptive sampling

Goal: Reduce uncertainty of eigenvalue

Uncertainty analysis decomposes by transitions from each state

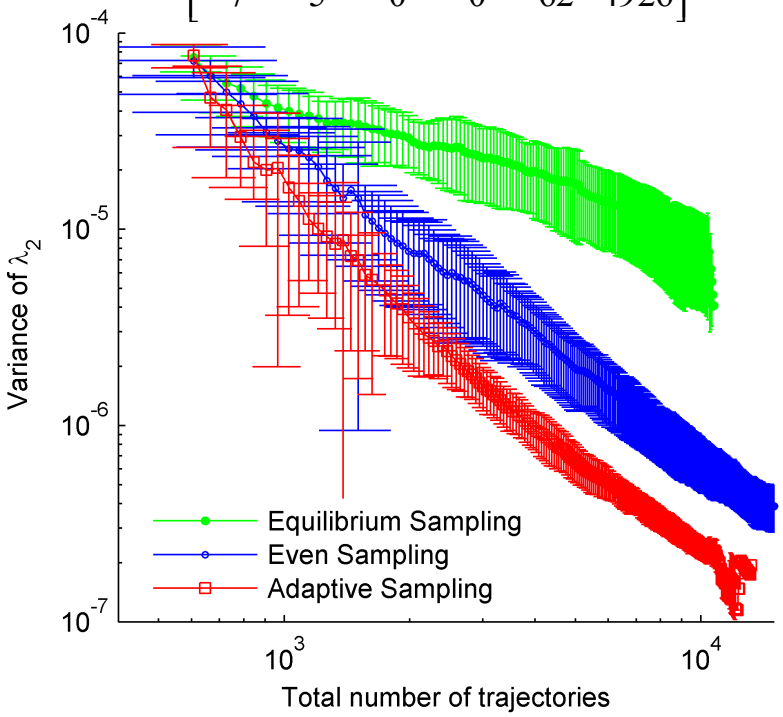
$$\lambda = \bar{\lambda} + \left(\left.\frac{\partial \lambda}{\partial t_{11}}\right|_{\bar{\mathbf{A}}} \Delta t_{11} + \ldots + \left.\frac{\partial \lambda}{\partial t_{1N}}\right|_{\bar{\mathbf{A}}} \Delta t_{1N} \right)$$
$$+ \left(\left.\frac{\partial \lambda}{\partial t_{21}}\right|_{\bar{\mathbf{A}}} \Delta t_{21} + \ldots + \left.\frac{\partial \lambda}{\partial t_{2N}}\right|_{\bar{\mathbf{A}}} \Delta t_{2N} \right)$$
$$+ \cdots$$
$$+ \left(\left.\frac{\partial \lambda}{\partial t_{N1}}\right|_{\bar{\mathbf{A}}} \Delta t_{N1} + \ldots + \left.\frac{\partial \lambda}{\partial t_{NN}}\right|_{\bar{\mathbf{A}}} \Delta t_{NN} \right)$$

Variance depends on both *uncertainty* of and *sensitivity* to transition probabilities

Adaptive sampling – alanine

Transition matrix of alanine peptide

$$\begin{bmatrix} 4380 & 153 & 15 & 2 & 0 & 0 \\ 211 & 4788 & 1 & 0 & 0 & 0 \\ 169 & 1 & 4604 & 226 & 0 & 0 \\ 3 & 13 & 158 & 4823 & 3 & 0 \\ 0 & 0 & 0 & 4 & 4978 & 18 \\ 7 & 5 & 0 & 0 & 62 & 4926 \end{bmatrix}$$



- Benefits
  - Very quickly reduce the variance
  - Reduce the total number of simulations
  - Need less computational power
  - Can study more complex systems

Conclusions

- Markov state models are tools to model long-scale dynamics
- Built from molecular dynamics simulations
  - Can be orders of magnitude shorter than system equilibration times
- Use both structural information and kinetic information to define states
- Have been applied successful to many biological systems
- Can calculate long-time behavior, folding pathways, error analysis, sampling, etc.